Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelor-Thesis

Niklas Netz

Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen/Technologien

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Niklas Netz

Anomalie-Erkennung mit Hilfe von Machine-Learning-Algortihmen/Technologien

Abschlussarbeit eingereicht im Rahmen der Bachelorprüfung

im Studiengang Bachelor of Science Wirtschaftsinformatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer : Prof. Dr. Klaus-Peter Schoeneberg
Zweitgutachter : Andre Pietsch

Abgegeben am 03.08.2016

Niklas Netz

Thema der Arbeit

Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen/Technologien

Stichworte

Maschinelles Lernen, Data Mining, Big Data, Anomalie-Erkennung, überwachtes Lernen, Klassifikation, Log-Management, Maschinendaten, Logistische Regression, Support Vector Machine, RandomForest, Splunk Enterprise, Machine Learning Toolkit and Showcase

Kurzzusammenfassung

Die steigende Menge an Daten, von der besonders Maschinendaten in Unternehmen betroffen sind, zählt zu einer der am schnellsten wachsenden Bereiche der so genannten „Big Data“. Generiert werden die Maschinendaten von Webseiten, Applikationen, Servern, Netzwerkkomponenten und vielen weiteren Geräten. Diese werden anschließend unter anderem in Log-Dateien abgelegt. Zur Analyse von Log-Dateien wird in der Otto GmbH & Co. KG unter anderem die Software Splunk Enterprise der Firma Splunk Inc. eingesetzt. Die derzeit über 13 Milliarden indizierten Events werden in Splunk verarbeitet und bilden die Grundlage für fundierte Entscheidungen.

In der vorliegenden Arbeit wird untersucht, wie in diesen Datenmengen Anomalien anhand von unqualifizierten Eventmengen in der Backend-Architektur der Otto GmbH & Co. KG erkannt werden können. Dabei liegt der Fokus der Arbeit auf der Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen und -Technologien. Die Darlegung der wissenschaftlichen Grundlagen und Ergebnisse aus bereits getätigten Forschungsarbeiten unterschiedlicher Wissenschaftler bilden neben den Erkenntnissen aus Experteninterviews die Basis der Untersuchungen.

Niklas Netz

Title of the paper

Annomaly-Detection with Machine-Learning-Algorithm/Technology

Keywords

Machine-Learning, Data Mining, Big Data, Anomaly-Detection, Supervised-Learning, Classification, Log-Management, Machine-Data, Logistic Regression, Support Vector Machine, RandomForest, Splunk Enterprise, Machine Learning Toolkit and Showcase

Abstract

The increasing amount of data that are affected by the particular machine data in companies is one of the fastest growing areas of the so-called "Big Data". The machine data are produced by websites, applications, server, network components and many other devices. These produced data are stored in log files. The company Otto GmbH & Co. KG uses the software Splunk Enterprise form the company Splunk Inc. to analyse log files. Currently more than 13 billion indexed Events are processed in Splunk and form the basis for well-founded decisions.

This thesis examines how anomalies based on unqualified amounts of events in the backend-architecture of the Otto GmbH & Co. KG can be detected. At this point the focus is on Anomaly-Detection with the help of machine learning algorithms and -technologies. They are based on the knowledge gained from interviews with experts, which are carried out in this work. The explanation of scientific bases, results of research conducted by different scientists as well as gained knowledge from expert interviews build the basis of these investigations.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Die Datenmenge in Unternehmen steigt ständig. Besonders betroffen ist hierbei die steigende Anzahl von Maschinendaten in Unternehmen, die jede Sekunde, jede Minute, jeden Tag weiter wächst.[1] Laut einer Studie von EMC und dem Marktforschungsunternehmen IDC soll sich das weltweite Datenvolumen bis 2020 verzehnfachen - von 4,4 Zettabyte im Jahr 2014 auf 44 Zettabyte.[2] Das sind 44 Billionen Gigabyte. Die entstandene Studie berechnet und prognostiziert die jährlich produzierte Datenmenge.

Maschinendaten zählen dabei zu einer der am schnellsten wachsenden und am weitesten verbreiteten Bereiche der so genannten „Big Data“. Die generieten Maschinendaten werden von Webseiten, Applikationen, Servern, Netzwerkkomponenten und vielen weiteren Geräten im Unternehmen erzeugt und unter anderem in Log-Dateien abgelegt.[3]

Die so abgelegten Maschinendaten beinhalten wertvolle Informationen über Aktivitäten und Verhaltensweisen im Zusammenhang mit Kunden, Benutzern, Transaktionen, Anwendungen, Servern, Netzwerken usw. Die Semantik dieser generierten Maschinendaten ist sehr komplex, da diese in unzähligen, nicht vorhersagbaren Formaten vorliegen.[4]

Mit Hilfe spezieller Software, die Analyse und Monitoring von Log-Dateien aus dem Backend-Bereich ermöglicht, können unabhängig vom vorliegenden Format oder Speicherort die Maschinendaten indexiert werden. Durch die richtige Verarbeitung und Analyse aus diesen Daten können wichtige Erkenntnisse gewonnen und eine Vorstellung davon, was innerhalb der gesamten IT-Systeme und Technologieinfrastruktur geschieht, erhalten werden.[5]

1.1 Problemstellung und Motivation

In der Otto GmbH & Co. KG wird zur Analyse von Log-Dateien unter anderem die Software Splunk Enterprise der Firma Splunk Inc. eingesetzt. Es werden derzeit über 13 Milliarden indizierte Events in Splunk verarbeitet, die die Grundlage für fundierte Entscheidungen bilden. Die Retention-Time beträgt 36 Tage. Dadurch können die Daten

[1] Vgl. Splunk Inc., Unternehmensübersicht, S. 1
[2] Vgl. EMC und IDC, Executive summary
[3] Vgl. Splunk Inc., Unternehmensübersicht, S. 1
[4] Vgl. Splunk Inc., Maschinendaten
[5] Vgl. Splunk Inc., Operational Intelligence

der letzten fünf Tage des Vormonats mit aufbewahrt werden. Um die Richtigkeit der Datenmenge sicherzustellen, sind weitere Untersuchungen notwendig, da nicht davon auszugehen ist, dass die gewonnenen Daten die Grundgesamtheit repräsentieren. Vielmehr ist zu beachten, dass die Datensätze Werte enthalten können, die das Ergebnis verzerren.[6] Zu Verzerrungen kann es kommen, wenn in den Datensätzen Extremwerte, sogenannte Anomalien, aufkommen. Es ist zu betonen, dass es sich hierbei nicht um fehlerhafte Werte handeln muss.[7]

In der Backend-Architektur der Otto GmbH & Co. KG werden Anomalien derzeit weder durch statistische Verfahren, noch automatisiert mit Hilfe von Machine-Learning-Algorithmen und -Technologien erkannt. Auch ist ein Grundverständnis für die beschriebene Problematik nicht vorhanden. Derzeit wird zur Erkennung von Anomalien auf den Dashboards lediglich die Methode des „scharfen Hinsehens“ verwendet, die viele personelle Ressourcen bindet. Diese personellen Ressourcen könnten im operativen Bereich der Otto GmbH & Co. KG für andere Tätigkeiten wertvoller eingesetzt werden. Es ist anzunehmen, dass durch den Einsatz von Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen und -Technologien zur Erhöhung der Datenqualität, zur Erkennung von Systemstörungen, Verbesserung der Systemüberwachung sowie zur Reduzierung der Fehlerbehebungszeit beigetragen werden kann.

1.2 Zielsetzung der Arbeit

In der vorliegenden Arbeit wird untersucht, wie Anomalien anhand von unqualifizierten Eventmengen in der Backend-Architektur der Otto GmbH & Co. KG erkannt werden können. Bei unqualifizierten Eventmengen ist die Rede von einer Gesamtmenge an Events, die in Log-Dateien von Datenbanken, Netzwerkkomponenten und Servern erzeugt bzw. abgelegt werden. Hierbei soll ein Grundverständnis für das Erkennen von Anomalien mit statistischen Verfahren, sowie mit Hilfe von Machine-Learning-Algorithmen und -Technologien geschaffen werden, wobei der Fokus auf der Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen und -Technologien liegt.

Um anhand der Eventmengen Anomalien zu erkennen, ist vorab ein Verständnis für die Auf- bzw. Vorbereitung der Daten aufzubauen, um einen Transfer in das Betriebsumfeld der Otto GmbH & Co. KG zu erzielen. Wichtig ist dabei vor allem die Untersuchung, welche Technologie und welcher Algorithmus für welche Daten geeignet sind und wie

[6] Vgl. Rambold, A. (1999), S. 5
[7] Vgl. Streck, G. (2004), S. 158

die Daten für die entsprechende Technologie und den entsprechenden Algorithmus aufbereitet werden müssen.

1.3 Aufbau der Arbeit

Die vorliegende Arbeit ist in sechs Teile gegliedert, die sich an der beschriebenen Zielsetzung orientieren.

Abb. 1 Gang der Untersuchung
Quelle: Eigene Darstellung

Zunächst wird der Stand der Forschung dargelegt. Hier ist das Ziel, eine wissenschaftliche Basis der Bachelorarbeit aufzuzeigen, indem wichtige Forschungsarbeiten im Bereich der Anomalie-Erkennung und des Machine-Learnings vorgestellt werden.

In Kapitel drei wird die Grundlage für das Verständnis der weiteren Inhalte der Bachelor-Thesis gebildet. Es werden die Begriffe „Big Data“ und „Maschinendaten“ erläutert. Anschließend wird auf den Begriff der „Anomalie“, sowie auf die Gründe und Einsatzbereiche der Anomalie-Erkennung eingegangen.

Ein weiterer wichtiger Bereich, der im Kapitel der wissenschaftlichen Grundlagen behandelt wird, beleuchtet die statistischen Techniken zur Erkennung von Anomalien. Darüber hinaus wird der Begriff des „Machine-Learnings“ erläutert. Hier werden die Motivation und die Anwendung des maschinellen Lernens, die Begriffe „überwachtes und unüberwachtes Lernen“, sowie Beispiele für Anwendungsgebiete des Machine-

Learnings und Algorithmen des Machine-Learnings betrachtet. Darauffolgend wird die Ausgangssituation der Otto GmbH & Co. KG zum Thema der Anomalie-Erkennung aufgeführt. Hier werden unter anderem das Log-Management, die Verwendung und Hauptfunktionen der Software Splunk Enterprise und die Backend-Architektur vorgestellt.

Im Hauptteil der vorliegenden Arbeit wird die Empirische Analyse beschrieben, die aus Experteninterviews und der eigenen Untersuchung besteht. Aus dem Experteninterview und den bereits erfassten Grundlagen zur Anomalie-Erkennung und zum Machine-Learning, sowie der Ausgangssituation der Otto GmbH & Co. KG lassen sich möglicherweise wichtige Erkenntnisse für die eigene Durchführung und Umsetzung in die Praxis ableiten. In der eigenen Analyse wird die Beschreibung der vorliegenden Daten und die Datenvorbereitung aufgeführt. Anschließend wird mit Hilfe einer Technologie, die Machine-Learning-Algorithmen unterstützt, ein Algorithmus zur Anomalie-Erkennung ausgewählt und dieser auf unqualifizierte Eventmengen angewendet. Im Anschluss werden die erzielten Ergebnisse aufgeführt und interpretiert.

In Kapitel sechs wird die Arbeit mit einer Schlussbetrachtung, bestehend aus einer Zusammenfassung, einer kritischen Würdigung und einem Ausblick, abgerundet.

Der Aufbau der Arbeit sowie die Zusammenhänge zwischen den einzelnen Kapiteln sind im Bezugsrahmen in Abbildung 2 grafisch veranschaulicht.

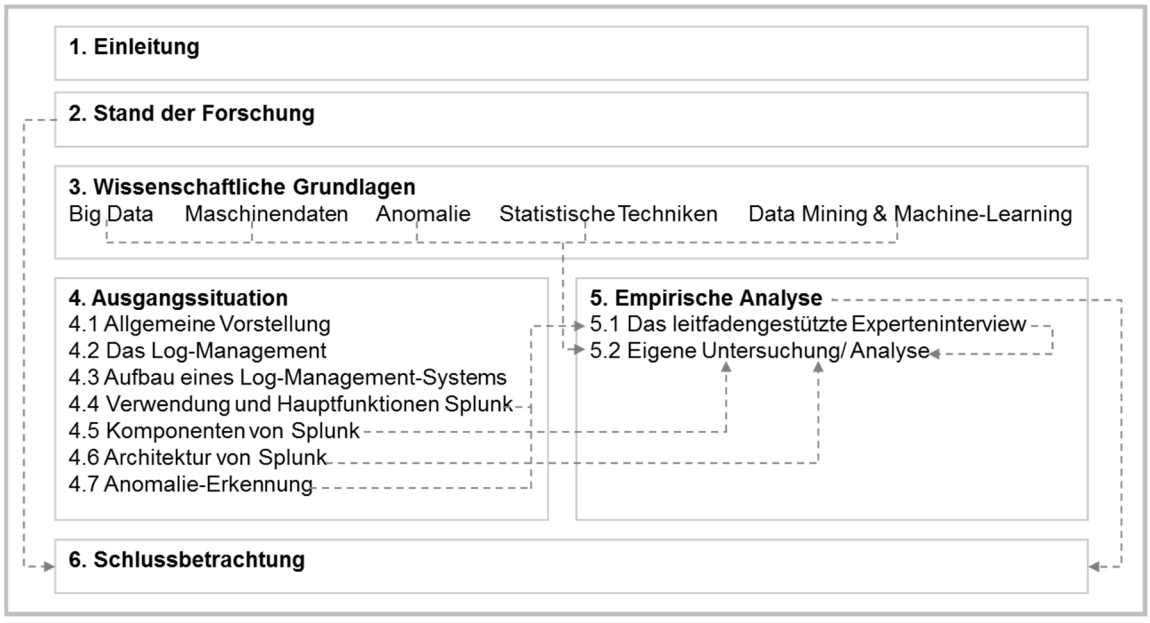


Abb. 2 Bezugsrahmen
Quelle: Eigene Darstellung

2 Stand der Forschung

Im 19. Jahrhundert hat die Forschung im Bereich der Ausreißer bzw. Anomalie-Erkennung im statistischen Bereich begonnen.[8] In der Statistik wurde hinsichtlich der Erkennung von Ausreißern Forschungsarbeit betrieben, die in der Literatur weiter aufgeführt wurde.[9] Seit dem sind weitere Anomalie-Erkennungstechniken in unterschiedlichen Forschungs- und Anwendungsgebieten entwickelt worden.[10]

Die folgende Abbildung zeigt die Forschungsbereiche der Anomalie-Erkennung mit den verbundenen Herausforderungen, sowie den Anwendungsgebieten der Anomalie-Erkennung.[11]

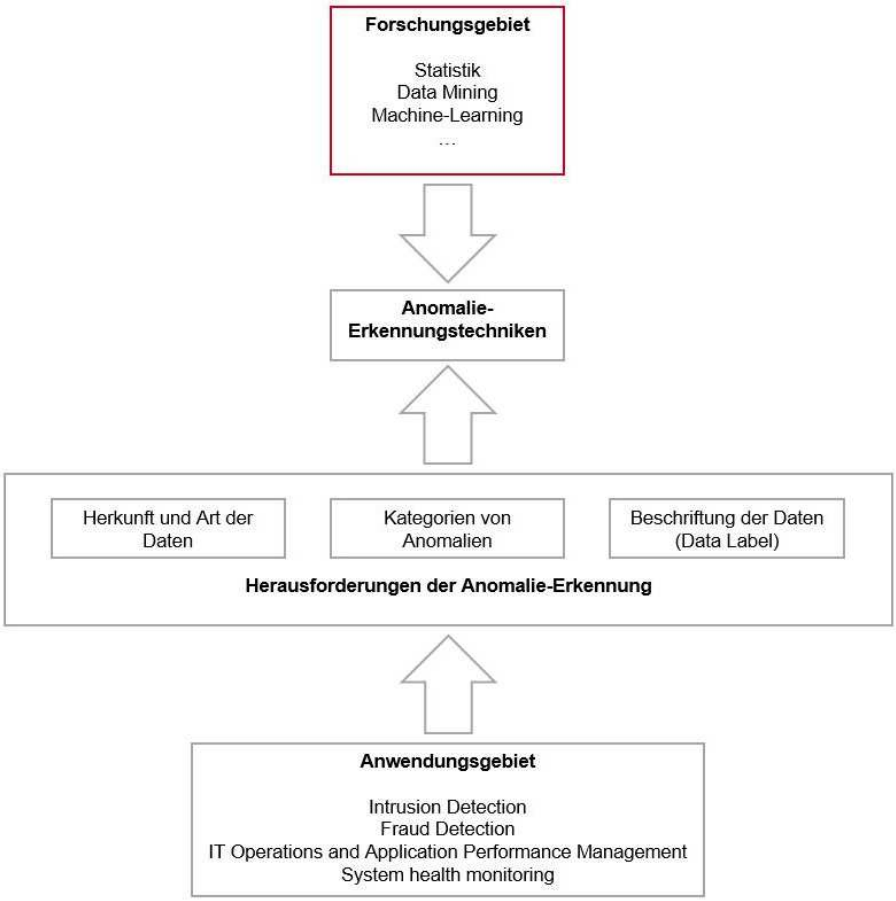


Abb. 3 Anomalie-Erkennungstechniken und deren Forschungsbereiche
Quelle: Eigene Darstellung in Anlehnung an Chandola, Banerjee, Kumar (2009), S. 4

[8] Vgl. Edgeworth, F., Y. (1887), S. 364 f.
[9] Vgl. Rousseeuw, P. and Leroy, A. (1987), Barnett, V. and Lewis, T. (1994), Hawkins, D. (1980)
[10] Vgl. Chandola, V., Banerjee, A., Kumar, V. (2009), S. 2
[11] In Kapitel drei wird der Begriff der Anomalie definiert und auf die Herausforderungen, sowie auf die Gründe und Einsatzbereiche der Anomalie-Erkennung eingegangen.

Die Abbildung stellt dar, dass die anzuwendende Anomalie-Erkennungstechnik vom Forschungsgebiet, den Herausforderungen der Anomalie-Erkennung und dem jeweiligen Anwendungsgebiet abhängt. Zu den Forschungsgebieten der Anomalie-Erkennung zählen die Statistik, das Data Mining und das Machine-Learning.

Im Bereich der Anomalie-Erkennung wurden einige Artikel und Bücher verfasst, die auf die verschiedenen Techniken zur Erkennung von Anomalien eingehen und diese beschreiben. Verfahren wie Klassifikation, Clustering, sowie Methoden aus der Statistik sind hier vor allem zu betrachten. Die Klassifikation und das Clustering sind Techniken, die dem Data Mining und dem Machine-Learning zuzuordnen sind. Auf diese und auch auf die statistischen Methoden der Anomalie-Erkennung wird im Kapitel der wissenschaftlichen Grundlagen näher eingegangen.

Einen Überblick über die Techniken zur Erkennung von Anomalien mit einem entsprechenden Verweis auf die jeweilige Literatur liefert die folgende Tabelle.

Autoren	**Techniken der Anomalie-Erkennung**		
	Klassifikation	Clustering	Statistik
Chandola, Banerjee, Kumar (2009)	✓	✓	✓
Hodge und Austin (2004)	✓	✓	✓
Agyemang (2006)	✓	✓	✓
Markou und Singh (2003a)	✓		
Markou und Singh (2003b)			✓
Patcha und Park (2007)	✓	✓	✓
Beckmann und Cook (1983)			✓
Bakar (2006)			✓

Tab. 1 Vergleich relevanter Literaturangaben
Quelle: Eigene Darstellung in Anlehnung an Chandola, Banerjee, Kumar (2009), S. 5

An dieser Übersicht ist zu erkennen, dass viele Autoren[12] die drei aufgeführten Techniken zur Anomalie-Erkennung nicht getrennt voneinander betrachten (siehe Chandola, Banerjee, Kumar (2009), Hodge und Austin (2004), Agyemang (2006), Patcha und Park (2007)). Hodge und Austin fokussieren sich in ihrer Arbeit auf Methoden des Machine-Learnings, sowie auf statistische Verfahren zu Anomalie-Erkennung.[13]

[12] Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung männlicher und weiblicher Sprachformen verzichtet. Sämtliche Personenbezeichnungen gelten gleichermaßen für beiderlei Geschlecht.
[13] Vgl. Hodge, V., Austin, J. (2004)

Einen umfassenden Überblick über Anomalie-Erkennungstechniken für numerische aber auch symbolische Daten bietet Agyemang.[14] Markou und Singh (2003a und 2003b) betrachten in ihren Arbeiten Ansätze der Klassifikation und der Statistik, wohingegen Patcha und Park zusätzlich das Clustering als Technik zur Anomalie-Erkennung thematisieren. Dabei betrachten Patcha und Park diese Techniken speziell im Bereich der Angriffserkennung.[15] Im Bereich der statistischen Anomalie-Erkennung haben, wie eingangs bereits erwähnt, viele Forschungsarbeiten stattgefunden. So sind auch Beckman, Cook und Bakar in diesem Zusammenhang aufzuführen.[16]

Besonders die oft zitierten Ausführungen von Chandola, Banerjee und Kumar sind grundlegend für die folgenden Kapitel. In ihrer Publikation beschreiben sie die Herausforderungen der Anomalie-Erkennung, sowie die verschiedenen Anwendungsgebiete der Anomalie-Erkennung und führen hierzu mögliche verwendbare Techniken auf.[17] Dabei wird hinsichtlich der Techniken auf weitere Literatur verwiesen. Die Anomalie-Erkennungstechniken werden in die verschiedenen Verfahrens- bzw. Methodenarten (Klassifikation, Clustering und Statistik) eingeteilt und hinsichtlich des Forschungsgebietes (Data Mining, Machine-Learning und der Statistik) näher beschrieben und diskutiert. Zudem werden die Vor- und Nachteile der jeweiligen Methode (Klassifikation, Clustering und Statistik) aufgeführt.[18]

[14] Vgl. Agyemang, M. (2006)
[15] Vgl. Markou, M., Singh, S. (2003a/b) und Patcha, A., Park, J. (2007)
[16] Vgl. Beckman, R., Cook, R. (1983) und Bakar, Z. (2006)
[17] Vgl. Chandola, V., Banerjee, A., Kumar, V. (2009), S. 6
[18] Vgl. Chandola, V., Banerjee, A., Kumar, V. (2009), S. 6 f.

3 Wissenschaftliche Grundlagen

In diesem Kapitel werden die für das Verständnis der gesamten Arbeit notwendigen Grundlagen beschrieben.

3.1 Big Data

Nicht nur das Internet erzeugt tagtäglich unzählige Mengen an Daten, sondern auch die in den Unternehmen zahlreich eingesetzten Informationssysteme.[19] Um aus diesen Daten wertvolle Informationen zu gewinnen, müssen aus der „Menge der (…) Daten diejenigen Informationen identifiziert werden mit denen die richtigen Schlussfolgerungen gezogen und Entscheidungen getroffen werden (können).“[20] Hier stoßen relationale Datenbanken bei der Verarbeitung solch großer und zum Teil unstrukturierter Datenmengen schnell an ihre Grenzen. Unzureichende Verarbeitung von unstrukturierten Daten, mangelnde Skalierbarkeit der Datenhaltung und Konsistenzsicherung der Daten gehören nach Schön zu den Nachteilen dieser Systeme. Um diese großen unstrukturierten, strukturierten und semi-strukturierten Datenmengen[21] zu erfassen, speichern und analysieren zu können werden Big Data Technologien eingesetzt. Mit Hilfe dieser Technologien sollen Abweichungen, Zusammenhänge oder Trends aus den Informationen erkannt werden, um Entscheidungen auf den Führungsebenen zu verbessern und zur Steigerung des Unternehmenswerts beizutragen.[22]

Zu den Einsatzgebieten von Big Data zählen unter anderem E-Commerce, Warenwirtschaft, Logistik und die IT. Letzteres nutzt Big Data z.B. für die Auswertung von Log-Dateien und weiteren IT-Daten. Dadurch können die Suche nach IT-Problemen, z.B. Sicherheitslücken, die IT-Nutzung- und Performance-Optimierung, sowie eine vorausschauende IT-Wartung und Instandhaltung verbessert werden.[23]

Durch die Vielschichtigkeit von Big Data ist zu erkennen, dass der Begriff als solches nicht eindeutig definiert werden kann.[24] Daher wird an dieser Stelle näher auf das von Doug Laney, Analyst des Marktforschungsanbieters Gartner, etablierte 3-V-Modell eingegangen.

---

[19] Vgl. Schön, D. (2016), S. 297
[20] Schön, D. (2016), S. 298
[21] Eine Definition dieser Begrifflichkeiten befindet sich in diesem Kapitel auf S. 9 unter dem Punkt Variety.
[22] Vgl. Schön, D. (2016), S. 298
[23] Vgl. Schön, D. (2016), S. 299 f.
[24] Vgl. Finlay, S. (2014), S. 13

Das 3-V-Modell, besser bekannt unter der Bezeichnung „the tree V's of Big Data", beinhaltet die drei Dimensionen Volume, Variety und Velocity.[25] Diese Kriterien sollen im Folgenden erläutert werden.

Volume (Datenmenge):

Das Volumen beschreibt die ständig steigende Datenmenge, die in sämtlichen Bereichen erhoben und gespeichert wird und Unternehmen zu Analysezwecken zur Verfügung steht. Beispielsweise generieren aktive Nutzer jede Minute auf Facebook über 650.000 verschiedene Inhalte oder verteilen ca. 35.000 Likes an Marken und Organisationen.[26] Diese großen Mengen an Daten stellen für relationale Datenbanken eine Herausforderung dar. Mit Hilfe von Big Data können die Herausforderungen von relationalen Datenbanken bewältigt und große Mengen an Daten gespeichert und verarbeitet werden.[27]

Variety (Datenvielfalt/Komplexität):

Ein weiterer wichtiger Aspekt der drei Dimensionen ist Variety. Dieser beschreibt die Vielfalt und Komplexität von Daten. Die riesigen Mengen an zu speichernden Daten weisen unterschiedliche Strukturen auf und stellen für traditionelle Datenbanksysteme ein weiteres Problem dar. Es wird hier zwischen strukturierten und unstrukturierten, sowie semi-strukturierten Daten unterschieden.

Bei strukturierten Daten handelt es sich zum Beispiel um eine relationale Tabelle, die Kundenstammdaten enthält.[28] Semi-strukturierte Daten hingegen bilden eine Mischform aus strukturierten und unstrukturierten Daten. Der Kopf einer E-Mail, der den Absender, Adressat und den Betreff beinhaltet, weißt eine klare Struktur auf und stellt den Teil der strukturierten Daten dar. Das Textfeld der E-Mail besteht hingegen häufig aus unstrukturierten Daten.[29] Unter Big Data werden alle vorhandenen Daten zusammengefasst und analysiert, egal, ob diese eine Struktur aufweisen oder nicht.[30]

Velocity (Datengeschwindigkeit):

Das Merkmal Velocity befasst sich mit der Datengeschwindigkeit. Dadurch, dass sich Daten z.B. im Internet und in den sozialen Netzwerken ständig verändern, muss nicht

---

[25] Vgl. Laney, D. (2001), S. 1
[26] Vgl. James, J. (2012)
[27] Vgl. Finlay, S. (2014), S. 13
[28] Vgl. Klein, D., Tran-Gia, P., Hartmann, M. (2013)
[29] Vgl. Schön, D. (2016), S. 303
[30] Vgl. Klein, D., Tran-Gia, P., Hartmann, M. (2013)

nur die Verarbeitungsmenge, sondern auch die Geschwindigkeit mit der die Daten verarbeitet werden, betrachtet werden.[31] Über Google werden jede Minute mehr als zwei Millionen Suchanfragen abgesetzt und auf Youtube 48 Stunden Videomaterial hochgeladen.[32] Die Auswertung dieser Daten soll möglichst in Echtzeit erfolgen, um schnellstmöglich wertvolle Informationen zu gewinnen.

Das Unternehmen IBM hat ein viertes Kriterium eingeführt und das 3-V-Modell auf ein 4-V-Modell erweitert. Dieses bezieht sich auf die Richtigkeit und Echtheit der Daten (Veracity).[33]

Veracity (Richtigkeit/Echtheit der Daten):

Daten liegen in unterschiedlichen Strukturen vor und stammen aus verschiedenen Quellen. Um aus diesen heterogenen Daten widerspruchsfreie Informationen zu gewinnen, ist es notwendig diese Daten zuvor zu bereinigen und zu harmonisieren. So ist bei der Datenauswertung und –bearbeitung die Sicherheit der Daten und Datenqualität zu beachten. „Falsche Daten sind zu löschen (und) unbestimmte Daten sind durch Transformation zu ergänzen (…).“[34] Nach Freiknecht sind unter falschen Daten beispielsweise grammatikalisch falsch übersetzte Texte oder auch irreführende Werbung zu verstehen.[35]

3.2 Maschinendaten

Maschinendaten zählen zu einem der am schnellsten wachsenden Bereiche von Big Data.[36] Maschinendaten sind Daten die von sämtlichen laufenden Systemen eines Rechenzentrums und von Geräten die an das Internet angebunden sind, generiert werden. Dazu zählen Netzwerkgeräte, Server und Applikationen.[37] Die generierten Maschinendaten stammen aus unterschiedlichsten Quellen und weisen eine komplexe Semantik auf.

Aufgezeichnet werden die Maschinendaten unter anderem in Log-Dateien, Konfigurationsdaten, Daten aus APIs und Sensordaten. Da aus ihnen verlässliche Aufzeichnungen zu Benutzertransaktionen, Kundenaktivitäten, Messwerten von Sensoren, Computerverhalten, Sicherheitsbedrohungen sowie betrügerischen

---

[31] Vgl. Schön, D. (2016), S. 303
[32] Vgl. James, J. (2012)
[33] Vgl. IBM (2012), S. 4
[34] Schön, D. (2016), S. 304
[35] Vgl. Freiknecht, J. (2914), S.13
[36] Vgl. Splunk Inc. Big Data und ihre versteckten Schätze
[37] Vgl. Splunk Inc. Maschinendaten

Aktivitäten und vielem mehr erhalten werden können, gehören Sie zu einer der wertvollsten Big Data Bereiche.[38] Durch das Analysieren solcher Maschinendaten besteht die Möglichkeit auftretende Anomalien zu entdecken, die interessante Informationen liefern können.

3.3 Anomalie

Der Begriff Anomalie ist nicht allgemeingültig definiert. Grundsätzlich sind ungewöhnliche Merkmale, die in den Daten auftauchen und ein vom Normalverhalten abweichendes Muster aufweisen, als eine Anomalie zu verstehen.[39] Hierzu zählen insbesondere Ausreißer, d.h. einzelne Daten, die stark von den übrigen Daten abweichen.[40]

Laut Barnett und Lewis ist unter einem Ausreißer folgendes zu verstehen: "We shall define an outlier in a set of data to be an observation (or subset of obervations) which appears to be inconsistent with the remainder of that set of data."[41] Besonders die Formulierung "appears to be inconsistent" zeigt auf der einen Seite, wie groß der Interpretationsspielraum ist, einen Wert als Ausreißer zu bezeichnen und auf der anderen Seite, „die Notwendigkeit, sich gerade dazu Gedanken machen zu müssen, was denn einen Ausreißer auszeichnet."[42]

Für die vorliegende Arbeit wird sich auf die oft zitierte Definition von Hawkins bezogen. Hawkins definiert einen Ausreißer als "an observation which deviates so much from other observations as to arouse suspicions that is was created by a different mechanism."[43]

Ausreißer sind also nicht nur falsch erfasste Werte, sondern können auch richtig und genau sein. Das Interesse Anomalien zu erkennen ist folglich groß, da Anomalien einerseits eventuell interessante Daten repräsentieren und andererseits großen Einfluss auf Statistiken haben können.[44]

Eine Anomalie in einem Computernetzwerk kann z.B. einen Hinweis auf einen gehackten Computer geben, der sensible Daten an ein unautorisiertes Ziel sendet.[45] Außerdem

---

[38] Vgl. Splunk Inc. Big Data und ihre versteckten Schätze
[39] Vgl. Haslett, J. (1992), S. 280
[40] Vgl. Runkler, T. (2015), S. 24
[41] Barnett, V., Lewis, T. (1994), S. 7
[42] Schendera, C. (2007), S. 165
[43] Hawkins, D. (1980), S.1
[44] Vgl. Bradley, R., Haslett, J. (1990), S. 1/5
[45] Vgl. Kumar, V. (2005), S. 4

können Anomalien, die während einer Kreditkartentransaktion auftreten, ein Anzeichen für einen Identitäts- oder Kreditkartendiebstahl sein.[46]

3.3.1 Herausforderungen der Anomalie-Erkennung

Um Anomalien zu erkennen, sind einige Faktoren zu beachten, die eine Herausforderung darstellen können. Diese werden im Folgenden genannt:[47]

- Die Einordnung von Daten in Normalverhalten und einer Anomalie ist nicht einfach, da die Grenze zwischen diesen beiden Verhaltensarten häufig nicht präzise erkennbar ist. Beispielsweise kann es sich bei einer Beobachtung einer Anomalie, die sich nahe an der Grenze zwischen den beiden Verhaltensarten befindet auch um Daten handeln, die ein Normalverhalten aufweisen.
- Zudem erhält eine Anomalie in jedem Anwendungsgebiet eine unterschiedliche Bedeutung. Z.B. kann im medizinischen Bereich die Schwankung der Körpertemperatur, die eine kleine Abweichung vom Normalverhalten aufweist, eine Anomalie darstellen. Hingegen kann im Anwendungsgebiet eines Aktienmarktes die gleiche Abweichung bei der Schwankung des Wertes einer Aktie als normal betrachtet werden. Folglich kann eine Technik zur Erkennung von Anomalien in einem Anwendungsgebiet nicht ohne Anpassungen auf einem anderen Gebiet angewendet werden.
- Des Weiteren liegen die Daten nicht beschriftet mit der Bezeichnung „Anomalie“ oder „Normal“ vor, um diese zum Trainieren/Validieren von Modellen zu verwenden. Dies stellt ein Problem bei der Nutzung von Anomalie-Erkennungstechniken dar.[48]
- Daten enthalten häufig ein Rauschen[49], auch bekannt unter der Bezeichnung „Noisy Data“, das Ähnlichkeiten zu einer Anomalie aufweist und daher schwer zu unterscheiden und zu entfernen ist.[50]

Die genannten Herausforderungen zeigen, dass die Erkennung von Anomalien nicht einfach ohne Weiteres zu bewältigen sind. Bei der Formulierung eines Problems spielen Faktoren, wie die Herkunft und Art der Eingangsdaten, die Verfügbarkeit von beschrifteten Daten (Data Label) sowie die Kategorien der Anomalie, die erkannt werden

---

[46] Vgl. Aleskerov, E., Freisleben, B., Rao, B. (1997), S. 220 f.
[47] Vgl. Chandola, V., Banerjee, A., Kumar, V. (2009), S. 3
[48] Im Kapitel 3.5.2.2 wird auf die dazu gehörenden Lernverfahren überwachtes / unüberwachtes Lernen näher eingegangen.
[49] Unter Rauschen wird eine ungewollte Anomalie in Daten verstanden. Vgl. Alpaydin, E. (2008), S. 28
[50] Vgl. Chandola, V., Banerjee, A., Kumar, V. (2009), S. 3

soll, eine Rolle. Zudem werden die genannten Faktoren für das jeweilige Anwendungsgebiet der Anomalie-Erkennung berücksichtigt.[51] Auf diese wird im folgenden Unterkapitel näher eingegangen.

3.3.1.1 Herkunft und Art der Eingangsdaten

Bei der Verwendung einer Technik zur Erkennung von Anomalien ist die Herkunft und Art der Eingangsdaten (input data) ein wichtiger Aspekt der zu berücksichtigen ist. Bei den Eingangsdaten handelt es sich um eine Sammlung von Daten, die auch als Objekte, Muster, Ereignisse, sowie Entitäten bezeichnet werden können. Die einzelnen Daten werden durch Attribute näher beschrieben. Als ein Attribut ist beispielsweise die Augenfarbe einer Person zu verstehen.[52] Ein Attribut ist eine Variable oder ein Feld, wobei der Wert eines Attributs die Attributausprägungen darstellt. Abbildung 2 zeigt den Zusammenhang dieser Begrifflichkeiten.

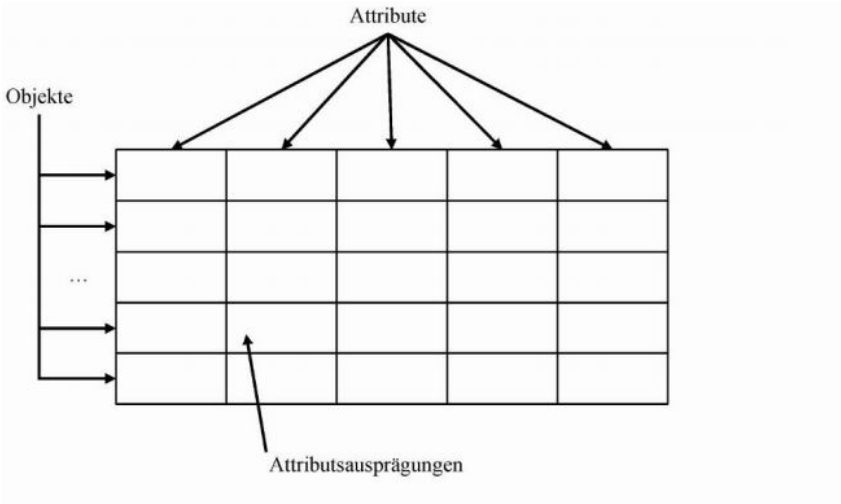


Abb. 4 Datenbasis des Data Minings
Quelle: Kantardzic, M. (2003), S. 11

Die beschriebenen Zusammenhänge zwischen Attribut, Objekt und Ausprägung können automatisiert durch Data Mining Methoden identifiziert werden.[53]

Die Attribute liegen für gewöhnlich nicht in einem einheitlichen Format vor und weisen verschiedene Wertebereiche auf. Nicht jede Data Mining Methode kann auf jedes Datenformat angewendet werden, so dass die Daten vorerst zu transformieren sind.[54]

---

[51] Vgl. Chandola, V., Banerjee, A., Kumar, V. (2009), S. 3
[52] Vgl. Tan, P.-N., Steinbach, M., Kumar, V. (2005), Chapter 2
[53] Eine detailliertere Ausführung zum Thema Data Mining ist in Kapitel 3.5.1 zu finden.
[54] Vgl. Piazza, F. (2010), S. 34

Das Format eines Attributs kann beispielsweise kategorial (categorical) oder stetig (continuous) sein.

Kategoriale (categorical) Daten sind diskrete oder qualitative Daten.[55] Dabei werden diese in verschiedene Kategorien bzw. Typen von Variablen eingeteilt. Dazu zählen nominale Variablen, binäre Variablen und ordinale Variablen.

Nominale Variablen können Objekte in Kategorien einteilen z.B. die Farbe eines Objektes in die Kategorie grün, blau oder rot. Zudem kann eine nominale Variable eine numerische Form aufweisen, wobei die numerischen Werte keinen mathematischen Aspekt betrachten. So können zehn Personen durch Nummern beschriftet werden, die dann allerdings nicht miteinander addiert werden können. Die nominale Variable findet im Bereich der Klassifikation[56] Anwendung.

Binäre Variablen sind ein Spezialfall der nominalen Variable. Diese können nur die Werte richtig oder falsch bzw. 1 oder 0 annehmen.[57]

Ordinale Variablen sind vergleichbar mit nominalen Variablen. Allerdings können ordinale Variablen eine Reihenfolge bzw. eine Rangordnung aufweisen, z.B. klein, mittel, und groß.[58]

Stetige (continuous) Daten sind auch als quantitative Daten zu verstehen.[59] Diese werden ebenfalls in verschiedene Kategorien bzw. Typen von Variablen eingeteilt. Hier sind die intervallskalierten Variablen und die verhältnisskalierten Variablen zu nennen.

Intervallskalierte Variablen stellen die Merkmalsausprägungen durch Zahlen dar. Dabei können deren Differenzen sinnvoll miteinander verglichen werden. Ein Beispiel stellt die Temperatur in Grad Celsius dar.[60]

Verhältnisskalierte Variablen besitzen im Vergleich zu den intervallskalierten Variablen einen natürlichen Nullpunkt und können somit in ein Verhältnis gesetzt werden. Beispiele der verhältnisskalierten Variable sind Einkommen und Geschwindigkeiten.[61]

---

[55] Vgl. Laerd Statistics (2013)
[56] Der Begriff Klassifikation wird in Kapitel 3.5.2.3 beschrieben.
[57] Vgl. Bramer, M. (2013), S. 10 f
[58] Vgl. ebd.
[59] Vgl. Laerd Statistics (2013)
[60] Vgl. Bramer, M. (2013), S. 10 f.
[61] Vgl. Bramer, M. (2013), S. 11 f.

Die Attribute legen fest, welche Anomalie-Erkennungstechnik anzuwenden ist. Für stetige (continuous) und kategoriale (categorical) Daten sind jeweils unterschiedliche statistische Techniken anzuwenden.[62]

3.3.1.2 Kategorien von Anomalien

Mit Hilfe verschiedener Techniken wird versucht, in einem Datensatz relevante Punkte (Anomalien) zu finden. Dabei werden die Anomalien in drei verschiedene Kategorien aufgeteilt: Point Anomalies, Contextual Anomalies und Collective Anomalies.[63]

Point Anomalies:

Unter „Point Anomalies" sind einzelne Datenpunkte zu verstehen, die sich von der restlichen Masse an Datenpunkten weit distanzieren. Ein Beispiel für eine „Point Anomaly" zeigt die folgende Abbildung.

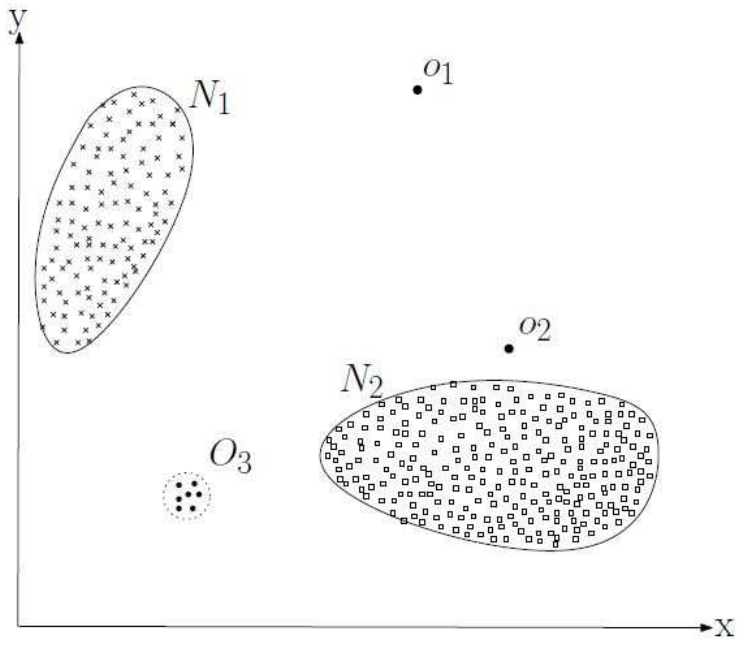


Abb. 5 Beispielhafte Darstellung von Point Anomalies
Quelle: Chandola, V., Banerjee, A., Kumar, V. (2009), S. 2

In den Regionen „N1" und „N2" sind Datenpunkte zu finden, die ein ähnliches Verhalten aufweisen. Im Gegensatz zu den Punkten „o1", „o2" und „O3", die sich von den eingekreisten Regionen „N1" und „N2" distanzieren und somit „Point Anomalies" darstellen.

---

[62] Vgl. Chandola, V., Banerjee, A., Kumar, V. (2009), S. 7
[63] Vgl. Chandola, V., Banerjee, A., Kumar, V. (2009), S. 7 f.

Ein vereinfachtes Praxisbeispiel, das eine „Point Anomaly“ darstellt, zeigt die Erkennung von Kreditkartenbetrug. Betrachtet man die durchgeführten Kreditkartentransaktionen einer Person und die damit verbundene Betragshöhe und stellt einen im Vergleich zu den normal getätigten Kreditkartentransaktionen höheren Transaktionsbetrag fest, so handelt es sich hierbei um eine „Point Anomaly“.[64]

Contextual Anomalies:

Bei einem Datensatz, in dem ein spezieller Kontext ein anomales Verhalten aufweist, ist die Rede von „Contextual Anomalies“. Eine Untersuchung dieser kontextbezogenen Anomalien treten häufig in Zeitreihendaten und Geodaten auf.[65] Die folgende Abbildung zeigt „Contextual Anomalies“ in Zeitreihendaten bezogen auf die monatliche Temperatur an.

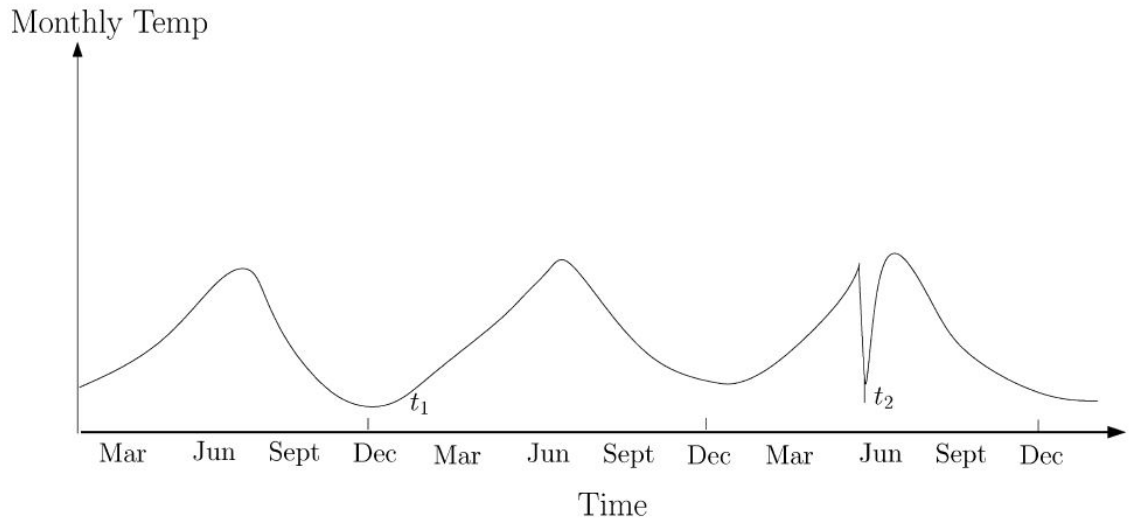


Abb. 6 Beispielhafte Darstellung von Contextual Anomalies
Quelle: Chandola, Banerjee, Kumar (2009), S. 8

„t1“ repräsentiert aufgrund der Jahreszeit ein valides Verhalten, so erreicht die Temperatur im Dezember einen Tiefpunkt. Dem gegenüber wurde im Juni des darauffolgenden Jahres ebenfalls ein Tiefpunkt „t2“ gemessen, der ähnliche Werte annimmt wie „t1“. Werden nun diese beiden Werte miteinander verglichen und in den Kontext eingeordnet, ist „t2“ aufgrund der Jahreszeit als nicht repräsentativ einzuordnen und somit als „Contextual Anomaly“ zu klassifizieren.[66]

---

[64] Vgl. Chandola, V., Banerjee, A., Kumar, V. (2009), S. 7
[65] Vgl. Weigend, A. S., Mangeas, M., Srivastava, A. N. (1995) S.373 f.
[66] Vgl. Chandola, V., Banerjee, A., Kumar, V. (2009), S. 8

Ein ähnliches Beispiel zeigt das Kaufverhalten mit einer Kreditkarte in unterschiedlichen Jahreszeiten. Beispielsweise weißt eine durchschnittliche wöchentliche Kreditkartenabrechnung 100 € auf, mit der Ausnahme der Wochenabrechnung vor Weihnachten. Diese beträgt 1.000 €. Zeigt die Kreditkartenabrechnung in einer Woche im Monat Juni ebenfalls einen Betrag von 1.000 € an, kann dieser wiederum als „Contextual Anomaly" definiert werden, da dies eine Abweichung vom Normalverhalten der kontextbezogenen Zeit wiederspiegelt.[67]

Collective Anomalies:

Bei einer Gruppe von Daten, die hinsichtlich des gesamten Datensatzes ein anormales Verhalten aufweisen, handelt es sich um eine „Collective Anomaly". Hier sind nicht die einzelnen Daten innerhalb der Gruppe als eine Anomalie anzusehen, sondern die gesamten Daten einer Gruppe. Eine beispielhafte Darstellung von „Collective Anomalies" zeigt die folgende Abbildung.

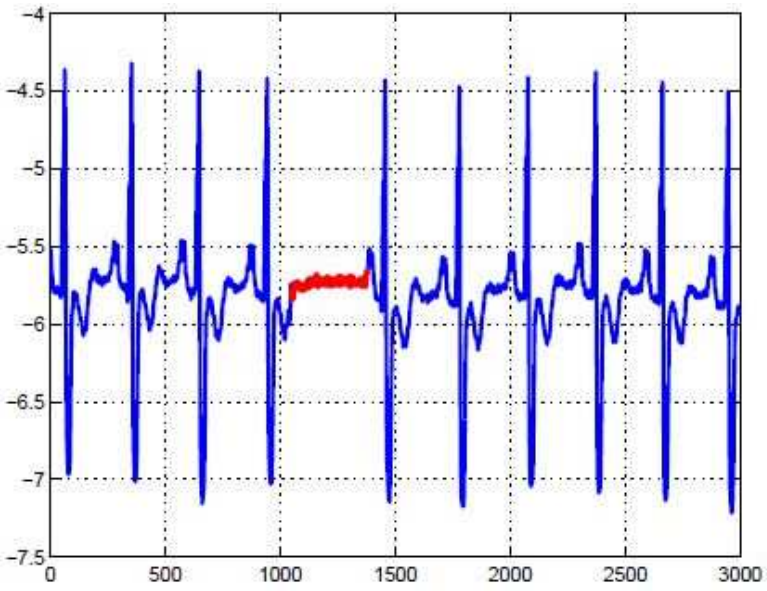


Abb. 7 Beispielhafte Darstellung von Collective Anomalies
Quelle: Chandola, Banerjee, Kumar (2009), S. 9

Um eine „Collective Anomaly" praxisnah darzustellen, wird eine Sequenz, die aufgetretene Aktionen eines Computers enthält, folgendermaßen aufgezeigt:

…http-web, buffer-overflow, http-web, http-web, smtp-mail, ftp, http-web, ssh, smtp-mail, http-web, **ssh, buffer-overflow, ftp,** http-web, ftp, smtp-mail, http-web…

[67] Vgl. Chandola, Banerjee, Kumar (2009), S. 8

Die markierte Sequenz mit den Ereignissen (**ssh, buffer-overflow, ftp**) zeigt eine typische webbasierte Attacke, bei der mit Hilfe eines Remotecomputers Daten von einem Host an ein entferntes Ziel per ftp[68] kopiert werden. Das Beispiel verdeutlicht, dass es sich bei der markierten Sequenz an Ereignissen um eine Anomalie handelt, da die Ereignisse zusammenhängend als Gruppe auftreten. Hingegen repräsentieren einzelne, zusammenhangslose Ereignisse keine Anomalie, da sie sich an verschiedenen Stellen in der Sequenz befinden.[69] „Collective Anomalies“ treten in Sequenzdaten, Graphdaten und Geodaten auf.[70]

3.3.1.3 Beschriftung der Daten (Data Label)

Ein weiterer Aspekt, der bei Verwendung einer bestimmten Technik zu Erkennung von Anomalien zu betrachten ist, ist die Beschriftung von Daten (Vergabe von Data Labels). Die Daten können hinsichtlich des Verhaltens z.B. mit der Beschriftung „Anomalie“ oder „Normal“ versehen werden. Allerdings ist das Beschriften von großen Datensätzen mit einem hohen Zeitaufwand verbunden, da die Beschriftung manuell von einem Experten durchzuführen ist.[71] Die Beschriftung der Daten kann zur Erstellung eines Trainingsdatensatzes dienen, auf den ein spezielles Lernverfahren zur Erkennung von Anomalien angewendet wird. Die verschiedenen Lernverfahren werden im Kapitel 3.5.2.2 näher beschrieben.

3.3.2 Gründe und Einsatzbereiche der Anomalie-Erkennung

Ziel ist es, ungewöhnliche Muster in den Daten zu erkennen, die vom Normalverhalten abweichen. Diese ungewöhnlichen Muster werden als Anomalie, Ausreißer oder Abweichung bezeichnet. Anomalien können in Daten auf einen kritischen Zustand hinweisen, bei dem ein Handlungsbedarf besteht. Aus diesem Grund wird die Anomalie-Erkennung in zahlreichen Bereichen angewendet. Dazu zählen: [72]

- „Intrusion Detection“ im Bereich der IT-Security, um beispielsweise Angriffe in einem Computernetzwerk zu erkennen,
- „Fraud Detection“ zur Erkennung von kriminellen Aktivitäten in kommerziellen Organisationen wie Banken, Kreditkartenunternehmen, Aktienmärkten und Versicherungsagenturen,

---

[68] „Das File Transfer Protocol (FTP) wird verwendet, um auf einen entfernten Computer zuzugreifen und Dateien von dort herunterzuladen.“ Laudon, C., Laudon, P., Schoder, D. (2010), S. 375
[69] Vgl. Chandola, V., Banerjee, A., Kumar, V. (2009), S. 9
[70] Vgl. Forrest, S., Warrender, C., Pearlmutter, B. (1999) S.133 f. und Noble, C. C., Cook, D. J. (2003) S. 631 f. und Shekhar, S., Lu, C.-T., Zhang, P. (2001) S. 371 f.
[71] Vgl. Chandola, V., Banerjee, A., Kumar, V. (2009), S. 10
[72] Vgl. Chandola, V., Banerjee, A., Kumar, V. (2009), S. 17

- „Medical and Public Health Anomaly Detection“, um im Bereich der Medizin anhand von Anomalien den Zustand eines Patienten oder eines technischen Gerätes zu überwachen,
- „Industrial Damage Detection“ zur Überwachung von industriellen Komponenten wie einem Motor oder Turbinen, sowie
- „System health monitoring“, um Störungen oder den Ausfall einer Komponente zu erkennen und zu verhindern.

Zudem wird die Anomalie-Erkennung im Bereich IT Operations and Application Performance Management eingesetzt. Dabei wird das Ziel verfolgt, Probleme aufzufinden und zu beheben bevor sie Auswirkungen auf den End-User haben.[73]

3.4 Statistische Techniken zur Erkennung von Anomalien

Statistische Techniken sind eine der ersten Ansätze die zur Erkennung von Anomalien angewandt wurden.[74] Im Bereich der Statistik können unter anderem Streuungsmaße und Box-Plots zur Identifikation von Ausreißern bzw. Anomalien dienen.[75]

3.4.1 Streuungsmaße

Zu den Streuungsmaßen gehören die Spannweite, der Quartilsabstand, die Varianz und die Standardabweichung.[76]

Spannweite

Die Spannweite R (englisch range) wird durch die Breite des Streubereichs, also durch den größten und kleinsten Messwert der Verteilung berechnet.

$$R = x_{max} - x_{min} \tag{3.1}$$

Dabei bezieht sich die Spannweite auf alle Werte einer Verteilung. Treten auffällig hohe Werte der Spannweite auf, geben diese einen Hinweis darauf, dass Ausreißer vorliegen.[77]

---

[73] Vgl. Prelert Inc. (2014), 3 Ways Anomaly Detection Improves IT Operations and Application Performance Management
[74] Vgl. Hodge, V., Austin, J. (2004), S. 85 f.
[75] Vgl. Schendera, C. (2007), S. 171 f. und Chandola, V., Banerjee, A., Kumar, V. (2009), S. 34 f., Dieser Abschnitt dient dem allgemeinen Verständnis und ist als Grundlage der Folgekapitel zu betrachten. Im Praxisteil wird nicht näher auf die statistischen Verfahren eingegangen, vielmehr werden die Algorithmen des Machine-Learnings betrachtet, die auf statistischen Verfahren aufbauen.
[76] Vgl. Schendera, C. (2007), S. 171
[77] Vgl. Schendera, C. (2007), S. 171

Quartilsabstand

Ein weiteres Maß der Streuung ist der Quartilsabstand (englisch interquartile range (IQR)), der durch die Differenz der Quartile Q3 (drittes Quartil) und Q2 (zweites Quartil) berechnet wird.

$$IQR= Q_3 – Q_1 \tag{3.2}$$

Das erste Quartil (Q_1) gibt die 25 %-Grenze und das dritte Quartil (Q_3) die 75 %-Grenze an. Dabei kann sich ein Quantil als Schwellenwert vorgestellt werden, bei dem ein Anteil der Werte kleiner ist als das Quantil und ein anderer Anteil größer ist. Beim 25 %-Quantil, bedeutet dies, dass 25 % der Werte unter dem 25 %- Quantil liegen oder gleich diesem Wert. Das Verhältnis zwischen diesen beiden Quartilen (Q1 und Q3) kann somit ebenfalls einen Hinweis auf einen Ausreißer liefern.[78]

Varianz

Die Varianz s^2 und die Standardabweichung s sind formal eng miteinander verbunden und werden häufig als Streuungsmaß eingesetzt.[79] „Die Varianz basiert auf der Abweichung der Messung (z.B. vom Mittelwert).“[80]

Bourier definiert die Varianz folgendermaßen: „Die Varianz ist die Summe der quadrierten Abweichungen der Merkmalswerte vom arithmetischen Mittel, dividiert durch die Anzahl der Merkmalsträger.“[81]

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.3}$$

Je größer die Variabilität um die Datenmenge des arithmetischen Mittels ist, desto größer ist auch die Varianz. Treten auffällig hohe Varianzen auf, sind diese durch Ausreißer verzerrt und sollten überprüft werden.[82]

---

[78] Vgl. Schendera, C. (2007), S. 172
[79] Vgl. Bourier, G. (2014), S. 96
[80] Schendera, C. (2007), S. 172
[81] Bourier, G. (2014), S. 97
[82] Vgl. Schendera, C. (2007), S. 172

Standardabweichung

Die Standardabweichung s wird aus der Varianz abgeleitet und ist die Quadratwurzel aus der Varianz.

$$s = \sqrt{s^2} \tag{3.4}$$

Liegt ein Wert außerhalb eines bestimmten Bereichs der 2,5- (bzw. 3-) fachen Standardabweichung, kann, bei der Annahme der Normalverteilung, dieser Wert als Ausreißer definiert werden.[83]

3.4.2 Boxplot

Eine weitere Technik zur Erkennung von Ausreißern bzw. Anomalien stellt im Bereich der Statistik der Boxplot dar.[84] Abbildung 8 zeigt den exemplarischen Aufbau eines Boxplots.

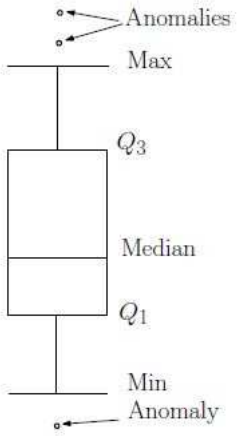


Abb. 8 Aufbau eines Boxplot
Quelle: Chandola, Banerjee, Kumar (2009), S. 34

Ein Boxplot ist eine grafische Darstellung, die als Streuungsdiagramm bezeichnet wird.[85] Die Darstellung eines Boxplot ermöglicht es in der Praxis sich einen ersten Überblick über die Verteilung der Daten und mögliche Abweichungen von einer „normalen“ Form zu verschaffen.[86] Ein Boxplot besteht aus einem „Kasten“. Dieser ist durch das erste Quartil (auch 25 %-Perzentil genannt) und dritte Quartil (75 %-Perzentil) der Verteilung begrenzt. Die Differenz des dritten und ersten Quartils wird als Interquartilsabstand bezeichnet. Dieser umfasst in der Regel 50 % der Werte. Die mittlere Linie eines Boxplot

[83] Vgl. Lohninger, H. (2012), Ausreißertests - Grundregeln
[84] Vgl. Laurikkala, J., Juhola, M., and Kentala, E. (2000), S. 20 f.
[85] Vgl. Schendera, C. (2007), S. 177
[86] Vgl. Cleff, T. (2011), S. 55

stellt den Median dar. Der Median wird auch als das zweite Quartil bzw. 50 % Perzentil bezeichnet.[87] Die von der Box nach oben bzw. nach unten ausgehenden Linien werden als „Whisker“ bzw. „Antenne“ bezeichnet. Diese Linien enden beim niedrigsten Wert (Minimum) bzw. höchsten Wert (Maximum) aller Verteilungen. Bei Werten, die mehr als das 1,5- fache der Boxlänge vom unteren Quartil nach unten oder vom oberen Quartil nach oben entfernt liegen, werden als Ausreißer bzw. Anomalien bezeichnet.[88]

Die beschriebenen statistischen Verfahren sind Grundlage für das Data Mining und Machine-Learning.[89] Im folgenden Kapitel wird der Fokus auf das Gebiet des Data Mining gelegt, das unter anderem die Anomalie-Erkennung als Aufgabe beinhaltet. Dabei werden die Methoden des Machine-Learnings zur Erkennung von Anomalien näher betrachtet.

3.5 Data Mining und Machine-Learning

Das maschinelle Lernen (Machine-Learning) ist als Forschungsgebiet eng mit dem Gebiet des Data Mining bzw. der Wissensentdeckung in Datenbanken (Knowledge Discovery in Databases) verbunden. Beide Gebiete beschäftigen sich mit der Extraktion von Wissen aus großen Datenbeständen, sowie der Modellierung und Realisierung von Lernphänomenen hinsichtlich der Unterstützung eines Computers.[90]

In den letzten Jahrzehnten finden Data Mining und Machine-Learning nicht nur Anwendung in den Bereichen Industrie und Wirtschaft, sondern gehen auch in den privaten Bereich über.[91] Lernalgorithmen dienen bei Google zur Optimierung der Suchmaschinenergebnisse[92], bei Netflix dem Lernen von Filminteressen, um hierfür dem Nutzer individuelle Vorschläge zu unterbreiten,[93] sowie zur Filterung von Spam-E-Mails.[94]

3.5.1 Data Mining

Der Begriff “Data Mining” wird im deutschen als Datenmustererkennung übersetzt und oft auf große und komplexe Datenmengen (Big Data) angewendet.[95] Fayyad et al. ordnen Data Mining als einen Analyseschritt im „Knowledge Discovery in Databases“-

---

[87] Schendera, C. (2007), S. 177
[88] Vgl. Cleff, T. (2011), S. 55
[89] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 2.6
[90] Vgl. Görz, G., Schneeberger, J., Schmid, U. (2014), S. 405
[91] Vgl. ebd.
[92] Vgl. Mathewson, J. (2012)
[93] Vgl. Bell, R. Koren, Y., Volinsky, C. (2010), S. 24 f.
[94] Vgl. Guzella, T., Caminhas, W. (2009), S. 10206 f.
[95] Vgl. Hagedorn, J., Bissantz, N., Mertens, P. (1997), S. 601

Prozesses (KDD-Prozesses) ein.[96] Dabei wird zwischen Data Mining und KDD nicht unterschieden, daher wird Data Mining als Synonym des Begriffs KDD verwendet.[97] KDD kann nach der oft zitierten Definition von Fayyad als „der nichttriviale Prozess der Identifizierung gültiger, neuer, potenziell nützlicher und letztendlich verständlicher Muster in Daten"[98] verstanden werden.

Der KDD-Prozess besteht aus den Phasen Selektion, Vorverarbeitung, Transformation, Data Mining, Interpretation und Evaluierung, sowie Integration und Visualisierung (siehe Abbildung 9):

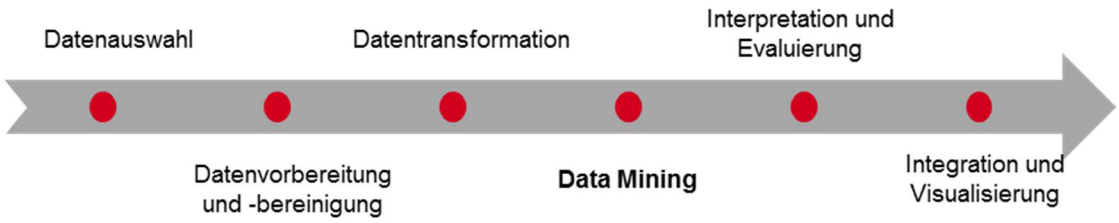


Abb. 9 Der KDD-Prozess
Quelle: Eigene Darstellung in Anlehnung an: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996), S. 10

Zunächst wird ein Verständnis für das Anwendungsgebiet entwickelt und ein KDD-Ziel festgelegt (diese Schritte sind in der Grafik nicht aufgeführt), bevor eine Analysebasis durch die Auswahl von Daten geschaffen wird und diese bereinigt und vorbereitet werden. Darauf folgt die Datenreduktion und Datenanreicherung durch Datentransformation, um eine geeignete Repräsentation zu erlangen. Anschließend werden durch die Auswahl einer geeigneten Data Mining-Methode (Klassifikation, Clustering-, Regressions- oder Assoziationsanalyse, siehe Kapitel 3.5.2.3) Algorithmen selektiert und nach Mustern in den Datensätzen gesucht. Die Muster werden im nächsten Schritt interpretiert, um die gewonnen Erkenntnisse zu visualisieren und schlussendlich in das Anwendungsgebiet zurück zu überführen.[99] Methoden des Data Mining nutzen dabei statistische Ansätze, sowie das Machine-Learning.

3.5.2 Machine-Learning

Große Datenmengen können durch das ständige Fortschreiten der Technologien gespeichert und verarbeitet werden. So erzeugt beispielsweise eine große Supermarktkette mit hunderten von Filialen, die in mehreren Ländern vertreten ist, durch

[96] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996), S. 6
[97] Vgl. Görz, G., Schneeberger, J., Schmid, U. (2014), S. 409
[98] Vgl. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996), S. 6
[99] Vgl. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996), S. 10

den Verkauf tausender Artikel an Millionen von Kunden jeden Tag mehrere Gigabytes an Daten. Dabei sichert bei jedem Verkauf ein Kassenterminal detaillierte Informationen. Zu diesen gehören das Datum, ein Kundenidentifikationscode, die Anzahl der gekauften Artikel, ein Verkaufspreis usw. Diese Daten können durch Analysen einen großen Informationsgewinn ermöglichen und beispielsweise für Vorhersagen genutzt werden.

Aus den Daten kann auch das Kundenverhalten analysiert und ein Muster hinsichtlich der gekauften Artikel erkannt werden. Ein Kunde der Bier kauft, kauft auch immer Chips. Genau in diesem Bereich ist die Anwendung von Machine-Learning sinnvoll und hilfreich.[100] Weitere Anwendungsgebiete des Machine-Learnings werden im Kapitel 3.5.2.3 aufgeführt.

Unter maschinellem Lernen ist das Programmieren eines Computers zu verstehen, sodass anhand von Beispieldaten oder Erfahrungswerten, die aus der Vergangenheit stammen, ein bestimmtes Leistungskriterium optimiert wird. Liegt ein Modell mit bestimmten Parametern vor, so bedeutet Lernen, ein Computerprogramm auszuführen, um mit Hilfe von Trainingsdaten oder Erfahrungswerten aus der Vergangenheit, die Parameter des Modells zu optimieren. Dabei kann das Modell für das Treffen von zukünftigen Vorhersagen verwendet werden (prädiktiv) oder zum Erlangen von Wissen aufgrund der vorliegenden Daten (deskriptiv).[101]

Mitchell definiert den Begriff maschinelles Lernen (Machine-Learning) folgendermaßen: „Machine Learning is the study of computer algorithms that improve automatically through experience.“[102]

Im Bereich des maschinellen Lernens lassen sich dabei die Bereiche Knowledge Discovery in Databases und Data Mining vom eigentlichen maschinellen Lernen (Machine-Learning) unterscheiden.

Beim KDD/Data Mining geht es vorwiegend um das Finden neuer Muster in den Daten, während es beim maschinellen Lernen primär um die Wiedererkennung von bereits bekannten Mustern in neuen Daten geht. Das KDD/Data Mining dient zudem zur Bereicherung des maschinellen Lernens durch die Vorverarbeitung und Erzeugung von Lerndaten, um zukünftige Vorhersagen treffen zu können. Beide Gebiete, sowohl das KDD/Data Mining als auch das Machine-Learning verfolgen das Ziel bereits gesammelte

---

[100] Vgl. Alpaydin, E. (2008), S. 1
[101] Vgl. Alpaydin, E. (2008), S. 3
[102] Mitchell, T. (1997)

Daten zu analysieren. Im Analyseschritt (Data Mining) werden Methoden (Algorithmen) des maschinellen Lernens eingesetzt. [103]

Die praktische Umsetzung des Machine-Learnings erfolgt mit Hilfe von Algorithmen. Die in Kapitel 3.5.2.4 aufgeführte Abbildung zeigt eine Übersicht von Algorithmen und der Einteilung der Algorithmen in verschiedene Kategorien.

3.5.2.1 Motivation und Anwendung des Machine-Learnings

Das Gebiet des maschinellen Lernens ist je nach Motivation der Forschenden aus unterschiedlichen Blickwinkeln zu betrachten.

In der Kognitionswissenschaft besteht das Interesse daran, das menschliche Lernen besser zu verstehen, um es mit Hilfe von Computermodellen abzubilden. Aus der Blickrichtung der Theorie besteht hingegen das Interesse, grundsätzlich ein Verständnis dafür aufzubauen, mit welchem Aufwand welche Lernaufgaben für welche lernenden Systeme zu bewältigen sind. Die anwendungsorientierte Sicht interessiert sich für die Entwicklung von Verfahren (Algorithmen), sowie für die Konstruktion von Systemen mit denen in der Praxis Lernaufgaben gelöst werden können und mit denen somit ein Nutzen und Mehrwert erbracht werden kann.[104]

Die verschiedenen Blickrichtungen können nicht getrennt voneinander betrachtet werden. Viele erfolgreich entwickelte Verfahren des maschinellen Lernens aus der Kognitionswissenschaft haben schließlich zu einem entsprechenden formalen Modell geführt, welches in verschiedenen praktischen Bereichen eingesetzt wird.[105]

3.5.2.2 Überwachtes und unüberwachtes Lernen

Beim maschinellen Lernen werden die Algorithmen in verschiedene Lernverfahren bzw. Lernmethoden eingeordnet. Hier sind hauptsächlich das überwachte Lernen (supervised learning) und unüberwachte Lernen (unsupervised learning) zu nennen, die im Folgenden näher beschrieben werden.[106]

Beim überwachten Lernen werden die vorliegenden Daten als Trainingsdatensatz bezeichnet. Zudem wird zwischen Ein- und Ausgabedaten unterschieden. Die Ausgabedaten sind durch eine Beschriftung (Data Label) zu versehen. Die Beschriftung

---

[103] Vgl. Görz, G., Schneeberger, J., Schmid, U. (2014), S. 409
[104] Vgl. Görz, G., Schneeberger, J., Schmid, U. (2014), S. 408
[105] Vgl. ebd.
[106] Vgl. Görz, G., Schneeberger, J., Schmid, U. (2014), S. 405

kann z.B. durch einen Experten erfolgen.[107] Anschließend verfolgt das überwachte Lernen das Ziel, die Abbildung von den Eingabedaten auf die Ausgabedaten zu erlernen, um das gelernte auf einen Testdatensatz anzuwenden.[108]

Zur Lernmethode des überwachten Lernens gehören sowohl die Regression als auch die Klassifikation.[109] Die Kategorisierung ist in der Abbildung 10 im Kapitel 3.5.2.4 aufgeführt. Dabei sind in der Unterteilung Regression und Klassifikation jeweils die möglichen verwendbaren Algorithmen dargestellt.

Das folgende Beispiel von Ertel soll in vereinfachter Form das überwachte Lernen demonstrieren.[110]

Geerntete Äpfel sollen automatisiert in Handelsklassen A und B eingeteilt werden. Dabei misst eine Sortieranlange mit Hilfe von Sensoren für jeden Apfel die zwei Merkmale „Größe“ und „Farbe“. Anschließend soll entschieden werden, in welche der beiden Handelsklassen der Apfel gehört. Dieses Vorgehen wird auch als klassische Klassifizierungsaufgabe bezeichnet. Die Klassifizierung erfolgt manuell durch einen Fachmann. D.h. der Fachmann definiert anhand der Merkmale „Größe“ und „Farbe“ welche Äpfel in welche Handelsklassen gehören. Diese Messwerte sind in der folgenden Tabelle dargestellt und stellen ein Satz an Trainingsdaten für das Sortieren der Äpfel dar.

Größe [cm]	8	8	6	3	…
Farbe	0,1	0,3	0,9	0,8	…
Handelsklasse	B	A	A	B	…

Tab. 2 Trainingsdaten für Klassifizierung
Quelle: Eigene Darstellung in Anlehnung an Ertel, W. (2013), S. 179

Das Ziel des überwachten Lernens besteht nun darin, aus den bereits vorliegenden Eingabedaten (Größe und Farbe) eine Abbildung auf die Ausgabedaten (Handelsklasse) zu erlernen, um für einen Testdatensatz (neue Äpfel) den Wert der Handelsklasse (A oder B) zu bestimmen.

Dieses Beispiel, das die Klassifikation als Methode zur Vorbereitung eines Trainingsdatensatzes für das überwachte Lernen beschreibt, kann ebenfalls auf die

[107] Vgl. Klose, O. (2015)
[108] Vgl. Alpaydin, E. (2008), S. 11
[109] Vgl. Alpaydin, E. (2008), S. 9, Die Methoden Regression und Klassifikation werden im Kapitel 3.5.2.3 beschrieben.
[110] Vgl. Ertel, W. (2013), S. 178 f.

Anomalie-Erkennung übertragen werden, indem die „Handelsklasse“ beispielsweise gegen „Verhalten“ ausgetauscht wird und die Daten mit den Labels „Anomalie“ oder „Normal“ beschriftet werden. Anschließend ist ein Algorithmus des maschinellen Lernens auszuwählen, der anhand der Trainingsdaten neue Merkmale eines Testdatensatzes in die jeweilige Verhaltensart einsortiert.[111]

Beim unüberwachten Lernen hingegen liegen nur die Eingabedaten vor. Die Ausgabedaten sind nicht bekannt, sodass beim unüberwachten Lernen das Erkennen von Regelmäßigkeiten in den Eingabedaten als Ziel deklariert ist.[112] Den unüberwachten Lernalgorithmen liegen also keine klassifizierten Beispiele wie beim überwachten Lernen vor.[113] Durch die existierende Struktur in den Eingabedaten versucht das unüberwachte Lernen bestimmte Muster zu erkennen, die häufiger auftreten als andere. In der Statistik ist dieser Vorgang unter dem Begriff der Dichteschätzung bekannt. Die Clusteranalyse ist eine Möglichkeit der Dichteschätzung.[114] Beim Clustering werden im Vergleich zum überwachten Lernen die Daten nicht klassifiziert,[115] es findet also keine Datenbeschriftung (Markierung von Data Labels) statt.[116] Das Ziel der Clusteranalyse besteht darin, Cluster (Häufungen) oder Gruppierungen von Eingabedaten zu finden.

Das folgende Beispiel von Alpaydin soll die Clusteranalyse verdeutlichen.[117]

Ein Unternehmen besitzt Kundendaten von ehemaligen Kunden. Darunter sind die vergangenen getätigten Transaktionen zwischen dem Kunden und dem Unternehmen zu finden. Das Interesse des Unternehmens besteht nun darin, das Kundenprofil zu untersuchen, um zu erkennen, welche Art Kunde häufig Transaktionen durchgeführt hat. In diesem Fall werden bei der Clusteranalyse alle Kunden mit einer ähnlichen Anzahl an durchgeführten Transaktionen derselben Gruppe zugewiesen. Diese Gruppierung kann das Unternehmen zum Treffen von gruppenspezifischen Strategieentscheidungen verwenden. So können beispielsweise Entscheidungen zielgenau hinsichtlich der angebotenen Dienstleistungen und Produkte getroffen werden. Außerdem wird durch die Gruppierung eine Identifikation von Ausreißern ermöglicht. Bei den Ausreißern handelt es sich um Objekte, die keinem der Cluster zugeordnet sind.

---

[111] Vgl. Dunning, T., Fiedman, E. (2014), S. 2
[112] Vgl. Alpaydin, E. (2008), S. 11
[113] Vgl. Görz, G., Schneeberger, J., Schmid, U. (2014), S. 405
[114] Vgl. Alpaydin, E. (2008), S. 11
[115] Vgl. Ertel, W. (2013), S. 230
[116] Vgl. Klose, O. (2015)
[117] Vgl. Alpaydin, E. (2008), S. 11

Beim unüberwachten Lernen kommen Verfahren der Segmentierung (Clusteranalyse) und der Assoziationsanalyse zum Einsatz. Der Unterschied dieser Verfahren liegt insbesondere in der Art der Struktur, nach der in den Daten gesucht wird. Häufig finden diese Verfahren Anwendung im Bereich der Wissensentdeckung in Datenbanken („Knowledge Discovery in Databases“) bzw. im Bereich des Data Mining. Dabei verfolgen diese Verfahren das Ziel Abhängigkeiten in den großen Mengen an Daten zu finden und diese kompakt zusammenzufassen.[118] Auch hier sind dem jeweiligen Verfahren verschiedene Algorithmen zugeordnet. Eine Übersicht ist in der Abbildung 10 in Kapitel 3.5.2.4 aufgeführt.

3.5.2.3 Beispiele für Anwendungsgebiete des Machine-Learnings

Durch die im Folgenden aufgeführten und beschriebenen Beispielanwendungen des maschinellen Lernens soll ein besseres Verständnis für die Verwendung des maschinellen Lernens gewonnen werden.

Dabei wird auf Beispiele aus den Bereichen der Assoziation, der Klassifikation und der Regression eingegangen. Die Segmentierung (Clusteranalyse) ist ebenfalls ein Anwendungsgebiet des maschinellen Lernens und wurde bereits im Kapitel 3.5.2.2 unter dem Lernverfahren des unüberwachten Lernens beschrieben.

Assoziation

Die Assoziationsanalyse ist eine der am meist verbreiteten Analyseverfahren im Data Mining und dient zur Entdeckung von Assoziationsregeln in Datenbanken.[119] An dieser Stelle kann das bereits in Kapitel 3.5.2 genannte Beispiel einer Supermarktkette erneut aufgegriffen und der Assoziationsanalyse zugeordnet werden.

Betrachtet wird die Warenkorbanalyse einer Supermarktkette. Dabei wird versucht Assoziationen zwischen den Produkten, die ein Kunde kauft, zu finden. Eine interessante Aussage kann z.B. sein: Wenn ein Kunde Bier und Pizza zusammen kauft, so ist es wahrscheinlich, dass der Kunde auch Chips kauft.

Das Finden solcher Assoziationsregeln ist von großem Interesse. So können diese in einem Supermarkt zur Planung der Warenanordnung genutzt werden. Der Supermarkt

---

[118] Vgl. Görz, G., Schneeberger, J., Schmid, U. (2014), S. 407
[119] Vgl. Görz, G., Schneeberger, J., Schmid, U. (2014), S. 442

kann somit die Chips in der Nähe von Pizza und Bier platzieren, um den Verkauf von Produkten zu steigern.[120]

Klassifikation

Im Bereich der Klassifikation beschreibt Alpaydin die Vorhersage des mit einem Kredit verbundenen Risikos auf Seiten der Kunden. Das Risiko beschreibt an dieser Stelle die Wahrscheinlichkeit dafür, dass der Betrag unter den vereinbarten Bedingungen rechtzeitig zurückgezahlt wird. Für die Bank ist die Vorhersage dieses Risikos von besonders hoher Bedeutung, um z.B. die Gewinnerzielung sicherzustellen.[121]

Anhand von Kundeninformationen z.B. Einkommen, Ersparnissen, Alter oder auch der finanziellen Vorgeschichte kann die Bank die Kreditwürdigkeit (das Risiko für den Betrag) kalkulieren. Zudem liegen der Bank vergangene Aufzeichnungen, die Informationen hinsichtlich der Darlehensabwicklung eines Kunden aufzeigen vor. Diese geben Auskunft darüber, ob ein Kunde ein Darlehen zurückgezahlt hat oder nicht. „Das Ziel besteht nun darin, anhand solcher Daten zu speziellen Darlehensanträgen auf eine allgemeingültige Regel rückzuschließen, welche die Assoziation zwischen den Attributen eines Kunden und dem zugehörigen Risiko codiert.“[122]

Mit Hilfe des maschinellen Lernens ist ein Modell auf die Vergangenheitswerte anzupassen, um das Risiko für die Vergabe eines neuen Kredits zu kalkulieren und daraufhin eine Entscheidungen zu treffen, ob ein Kredit vergeben wird oder nicht. Bei diesem Beispiel handelt es sich um ein Klassifikationsproblem, bei dem zwei Klassen zu betrachten sind: Kunden mit einem niedrigen Risiko und Kunden mit einem hohen Risiko.

Die Eingabedaten bestehen aus den Informationen der Kunden. Diese werden für die Klassifikation benötigt, um anhand der Eingabedaten zu entscheiden, ob ein Kunde der Klasse mit einem niedrigen Risiko oder der Klassen mit einem hohen Risiko zuzuordnen ist. Aus den Daten der Vergangenheit kann nach einem Training eine Klassifikationsregel die folgende Form aufweisen:[123]

IF Einkommen > X AND Ersparnisse > Y

THEN niedriges Risiko ELSE hohes Risiko.

---

[120] Vgl. Görz, G., Schneeberger, J., Schmid, U. (2014), S. 442
[121] Vgl. Alpaydin, E. (2008), S. 4 f.
[122] Alpaydin, E. (2008), S. 5
[123] Vgl. Alpaydin, E. (2008), S. 5

Die Trennung in unterschiedliche Klassen ist eine Funktion, die als Diskriminante bezeichnet wird. Sobald eine Regel definiert ist, die den Vergangenheitswerten und der Annahme, dass die Zukunft den Vergangenheitswerten ähnelt, entspricht, können Vorhersagen für neue Daten getroffen werden. Somit kann für einen neuen Darlehensantrag recht einfach entschieden werden, ob es sich um einen Kunden mit niedrigem oder hohem Risiko handelt.[124]

Bei der Klassifikation wird im Vergleich zur Regression, die einen Zahlenwert betrachtet, ein Klassencode z.B. 0 oder 1 verwendet, wobei die 0 für Kunden mit niedrigem Risiko und die 1 für Kunden mit hohen Risiko stehen.[125]

Regression

Als Beispiel der Regression führt Alpaydin ein System zur Vorhersage des Preises eines gebrauchten Autos auf. Dabei werden als Eingabedaten alle Attribute des Autos wie beispielsweise die Marke des Autos, das Baujahr, die Motorleistung und der Kilometerstand herangezogen, die den Wert des Autos beeinflussen. Die Ausgabe stellt den Preis des Autos dar. Diese Problemstellungen wird auch als Regressionsproblem bezeichnet, da die Ausgabedaten Zahlenwerte sind.

Bereits erfolgte Transaktionen, die den Verkaufspreis und den Typ des Autos mit seinen jeweiligen Attributen beinhalten, können als Trainingsdatensatz verwendet werden. Dabei bezeichnet X die Attribute des Autos und Y den Preis. Mit Hilfe des maschinellen Lernens ist eine Funktion an diese Daten anzupassen, um Y (den Preis) von X (den Attributen) zu erlernen.[126]

Den beschriebenen Beispielen für Anwendungsgebiete des Machine-Learnings, die auch als Data Mining Methoden bekannt sind, werden im folgenden Kapitel Algorithmen zugeordnet.

3.5.2.4 Machine-Learning-Algorithmen

In Abhängigkeit der Anwendungsgebiete des Machine-Learnings existiert eine Vielzahl an unterschiedlichen Algorithmen, die in einer Übersicht dargestellt werden.

Die folgende Abbildung zeigt eine Einteilung der verschiedensten Machine-Learning-Algorithmen hinsichtlich des Lernverfahrens (überwachtes und unüberwachtes Lernen),

---

[124] Vgl. ebd.
[125] Vgl. Alpaydin, E. (2008), S. 9
[126] Vgl. Alpaydin, E. (2008), S. 9

der jeweiligen Methode (Assoziationsanalyse, Klassifikation, Regression und Segmentierung bzw. Clusteranalyse) und der Art der Eingangsdaten (stetig oder kategorial).

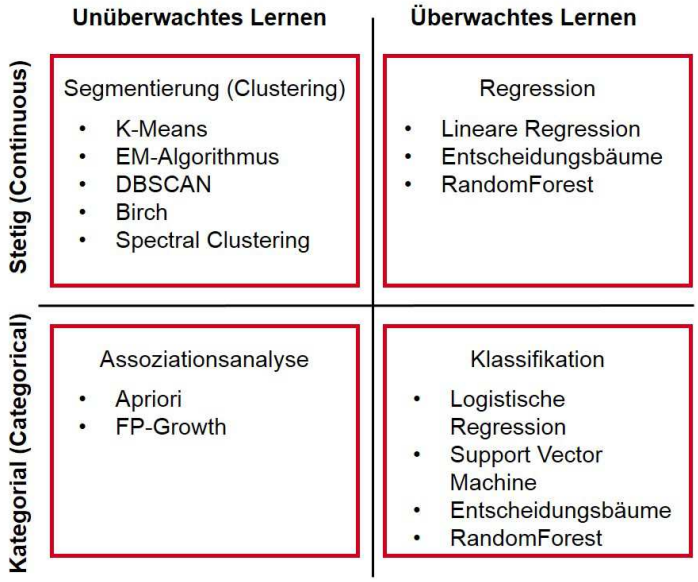


Abb. 10 Übersicht Machine-Learning-Algorithmen
Quelle: Eigene Darstellung in Anlehnung an Drieger, P. (2015)

Auf die Algorithmen wird hier nicht näher eingegangen. Eine detaillierte Betrachtung und Beschreibung eines bestimmten Algorithmus wird im Kapitel 5.2.6 (Algorithmen der Klassifikation) aufgeführt.

Um die hier beschriebenen theoretischen Grundlagen in ein praxisorientiertes Beispiel zu überführen und ein besseres Verständnis aufzubauen, werden im Folgenden die organisatorische Einordnung des Log-Managements in die Otto GmbH & Co. KG, sowie die Verwendung der Software Splunk Enterprise beschrieben.

4 Ausgangssituation im Thema Log-File-Analyse in der Otto GmbH & Co. KG

In diesem Kapitel wird das Unternehmen OTTO vorgestellt und der Bereich des Log-Managements organisatorisch in die Unternehmensstruktur eingeordnet. Weiter werden der Aufbau eines Log-Management-Systems, sowie die Hauptfunktionen der Software Splunk Enterprise, die unter anderem im Log-Management eingesetzt wird, beschrieben. Zudem folgt eine Erklärung der einzelnen Splunk Enterprise Komponenten. Abschließend wird auf die Architektur der Splunk-Umgebung und den derzeitigen Stand hinsichtlich der Anomalie-Erkennung eingegangen.

4.1 Allgemeine Vorstellung der Otto GmbH & Co. KG

Die Otto GmbH & Co. KG ist eines der erfolgreichsten E-Commerce-Unternehmen und bietet seit über sechs Jahrzehnten Mode- und Lifestyleprodukte auf Bestellung. In Deutschland ist OTTO mit rund sechs Millionen aktiven Kunden, 20 Millionen Bestellungen pro Jahr und über einer Millionen Visits pro Tag der größte Onlinehändler für Mode und Lifestyle. Im Geschäftsjahr 2015/2016 wurde mit otto.de 90 Prozent des Gesamtumsatzes erwirtschaftet.[127] Der Kunde steht bei den Aktivitäten und Entwicklungen des Unternehmens stets im Mittelpunkt. Durch den ständigen Austausch mit dem Kunden, ob über Telefon, Brief, E-Mail oder Social-Media wird optimal auf die Bedürfnisse eingegangen und neue Angebote entwickelt.

1949 wurde das Unternehmen OTTO von Werner Otto gegründet und ist noch heute im Familienbesitz. Der Sohn des Gründers, Dr. Michael Otto machte aus dem OTTO-Versand die international tätige Otto-Group mit 123 Unternehmen in 20 Ländern.[128]

Die Otto GmbH & Co. KG ist eine 100-prozentige Tochtergesellschaft der Otto Group mit Sitz in Hamburg und beschäftigt dort rund 4.350 Mitarbeiter. OTTO ist im Geschäftsjahr 2015/2016 weiter profitabel gewachsen, so konnte der Umsatz um zehn Prozent auf 2,561 Milliarden Euro gesteigert werden. Grund hierfür seien vor allem überdurchschnittlich viele Neukunden, die OTTO für sich gewinnen konnte.[129]

Zu den Geschäftsfeldern des Unternehmens zählen Fashion, Living und Technik. Besonders im Bereich Living wurde ein Umsatzplus von 20 Prozent erwirtschaftet, das OTTO den Vorsprung im Bereich Möbel und Heimtextilien weiter ausbauen lässt. Auch

---

[127] Vgl. Otto Basisinformation 03/2016
[128] Vgl. ebd.
[129] Vgl. Otto Pressemitteilung 21.03.2016

im Bereich Technik konnte auf ein positives Geschäftsjahr zurückgeblickt werden: Jeder dritte Fernseher, der in Deutschland online verkauft wurde, stammt von OTTO, bei den Waschmaschinen ist es jede zweite.[130]

Dass das Traditionsunternehmen sich ständig weiter entwickelt, zeigt ein Blick in die IT. 2013 wurde der Shop eigens runderneuert und 2015 auf ein Responsive Design umgestellt. Anfang 2016 konnte mit einem Großprojekt, in dem zentrale IT-Systeme gegen neue Entwicklungen ausgetauscht wurden ohne dabei den Kunden zu beinträchtigen, ein weiterer Meilenstein gesetzt werden.

Die Vorteile der digitalen Wandlung liegen auch bei den Kunden. Durch verstärkten Fokus auf „Verlässlichkeit und Planbarkeit in der Zustellung“[131] wird dem Kunden ermöglicht, sich stundengenau über die Lieferung der bestellten Ware zu informieren und sogar bei Großlieferungen Wunschlieferzeiten auszuwählen.

Im nächsten Geschäftsjahr soll diese digitale Transformation weiter ausgebaut und technologische Wachstumsbereiche vorangetrieben werden.[132] So wünscht sich auch der Bereichsvorstand Technology, Dr. Michael Müller-Wünsch, die IT bei OTTO noch stärker ins Tagesgeschäft zu integrieren: „IT funktioniert heute nicht mehr als Silo, als Maschinenraum im Hintergrund. IT muss dabei sein und mitmischen. Die Verhaltensänderungen der Konsumenten sind immer technologiegetrieben. Deshalb setzen wir uns frühzeitig mit neuen technischen Möglichkeiten auseinander und bewerten sie im Hinblick auf Potenzial und Mehrwert für den Kunden.“[133]

In diesem Zuge gewinnt auch die Verarbeitung und Analyse von Daten immer mehr an Bedeutung. Im Bereich der Datenanalyse in der Otto GmbH & Co.KG ist anzunehmen, dass die Technik der Anomalie-Erkennung mithilfe von Machine-Learning-Algorithmen dazu beitragen wird, den technologischen Fortschritt voranzutreiben und neue Potenziale zu entdecken.

4.2 Das Log-Management der Otto GmbH & Co. KG

In welchem Bereich das Log-Management der Otto GmbH & Co. KG organisatorisch eingeordnet ist, beschreibt das folgende Kapitel. Zudem werden die Aufgaben und Ziele des Log-Managements erläutert.

---

[130] Vgl. Otto Pressemitteilung 21.03.2016
[131] Vgl. ebd.
[132] Vgl. ebd.
[133] Otto GmbH & Co. KG, Dossier 10.11.2015

4.2.1 Organisatorische Einordnung des Log-Managements

Der Technologybereich der Otto GmbH & Co. KG ist ein eigener Vorstandsbereich und wird in die Hauptabteilungen Business Intelligence, Business Support Services, Category Management & Service Delivery, Design OTTO IT, Enterprise Architektur, Kunde & Service und Steuerung & Controlling gegliedert. Die Hauptabteilung Category Management & Service Delivery verantwortet neben der Abteilung Operation Competence, die Abteilungen Application Management, Einkaufssysteme, Kreativprozesse, Security & Architektur, Tool Competence, sowie die Abteilung Test & Release. Die Abteilung Operation Competence ist funktionsorientiert aufgebaut. Unterstützt wird die Abteilung Operation Competence unteranderem durch die Teams Application Services, Service Design, Projektmanagement und Service Management.

Das Team Service Design bildet das Bindeglied zwischen den Fachbereichen und der IT. Dabei wird das Business durch die Übersetzung von Anforderungen in passgenaue IT-Services unterstützt, sowie die Ausgestaltung und Realisierung bis hin zur Inbetriebnahme begleitet.

Das Team Projektmanagement verantwortet das Projektportfolio für die Domäne Category Management & Service Delivery. Im Fokus steht die Portfoliosteuerung über die Geschäftsjahre hinweg, sowie das stringente Management der Projekte durch die Projektleiter in der Domäne.

Das Team Service Management designt und managet die operativen ITSM Prozesse Incident Management, Request Fulfilment, Problem Management & Configuration Management. Außerdem wird hier die strategische Ausrichtung des Service Managements entwickelt.

Application Services ist in die Teilbereiche Prozessteuerung, Monitoring, sowie Reporting gegliedert. Die Prozesssteuerung ist für die Steuerung und Regelung technischer Prozesse zuständig. Dem Monitoring ist das Log-Management zugeordnet. Die managementgerechte Aufbereitung von Reports zur Darstellung erfolgt im Teilbereich Reporting.

4.2.2 Aufgaben und Ziele des Log-Managements

Die Aufgaben des Log-Managements können in zwei Bereiche gegliedert werden: Logfile-Analyse und Monitoring.

- Bei der Logfile-Analyse werden Log-Dateien hinsichtlich bestimmter Kriterien untersucht. Die Log-Dateien stellen eine wichtige Informationsquelle dar, da aus ihnen wichtige Informationen gewonnen werden können, die zur Fehlerbehebung beitragen. Die Logfile-Analyse erfolgt durch den Endanwender. Die Aufbereitung der Analyse kann durch die Erstellung von Dashboards und Reports erfolgen. Dabei dienen die Dashboards zur Visualisierung von verschiedenen Aspekten, um Schlussfolgerungen aus diesen zu ziehen. Die Schlussfolgerungen können zur Überwachung herangezogen werden. Die Reports dienen zur historisierten Aufbereitung der Daten.
- Beim Monitoring steht die Überwachung von Kennzahlen eines Prozesses im Fokus. Das Application Performance Monitoring (APM) überprüft die Leistungen von Anwendungen. Zudem dient das Monitoring zur Anbindung an das Eventmanagement, welches eintreffende Ereignisse bewerten soll, um eine entsprechende Eskalation, wie etwa die Erstellung eines Incidents, auszusteuern.

So zählt zu den Zielen des Log-Managements, das sich aus den Bereichen Logfile-Analyse und Monitoring zusammensetzt, die Sicherstellung, Probleme rechtzeitig zu erkennen und zu beseitigen, sowie wirtschaftliche Schäden zu vermeiden. Zudem soll eine Identifizierung der Ursache eines aufgetretenen Problems erfolgen und durch die Performance-Überwachung von Anwendungen die betriebliche Stabilität sichergestellt werden.

4.3 Aufbau eines Log-Management-Systems

Um Log-Dateien an einem zentralen Ort zu sammeln und zur strukturierten Analyse bereitzustellen, wird ein Log-Management-System benötigt. Dabei besteht ein Log-Management-System grundsätzlich aus einer Komponente zur Einsammlung der Daten (Einsammlungskomponente), einer Komponente zur Speicherung und Indizierung der Daten (Speicher- und Indizierungskomponente), einer Komponente zum Durchsuchen der Daten (Suchkomponente) und einer Komponente, die zur Verwaltung und Steuerung dient (Steuerungskomponente).[134] Der Aufbau eines Log-Management-Systems wird in der Abbildung 11 in vereinfachter Form dargestellt.

[134] Vgl. Dechert, M. (2015), Besser zentral: Professionelles Logging

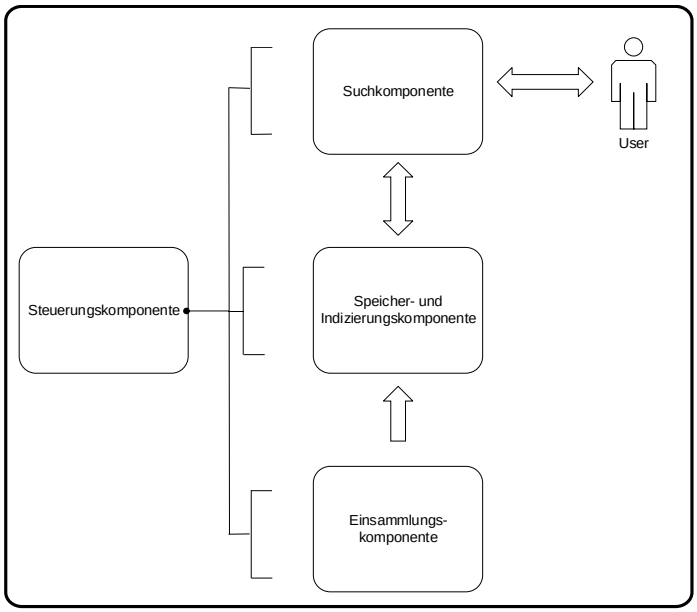


Abb. 11 Aufbau und Komponenten eines Log-Management-Systems

Die Einsammlungskomponente wird verwendet, um die Daten von den Quellsystemen einzusammeln und anschließend an die Speicher- und Indizierungskomponente weiterzuleiten. Die Speicher- und Indizierungskomponente dient zur Datenindizierung bzw. Datenhaltung. Hier werden die eingehenden Rohdaten in Ereignisse (Events) umgewandelt und anschließend in einem Index gespeichert. Nachdem die Ereignisse in einem Index gespeichert vorliegen, können über die Suchkomponente Suchanfragen auf diesen Daten durchgeführt werden. Eine Steuerungskomponente ist für die Verwaltung und Steuerung der zuvor beschriebenen Komponenten zuständig.

In der Otto GmbH & Co. KG wird das Log-Management unter anderem mit der Software Splunk Enterprise betrieben. Die Verwendung und Hauptfunktionen der Software Splunk Enterprise werden im nächsten Abschnitt genauer betrachtet. Anschließend wird auf die Komponenten der Software Splunk Enterprise eingegangen.

4.4 Verwendung und Hauptfunktionen der Software Splunk Enterprise

Die Software Splunk Enterprise ist eine Plattform für Operational Intelligence und ermöglicht es, den größtenteils ungenutzten Wert der Big Data, der aus Systemen der IT-Infrastruktur und Geschäftsanwendungen erzeugt wird, zu sammeln, zu analysieren und zielführend zu verwerten.[135]

[135] Vgl. Splunk Inc., Splunk® Enterprise – Die Plattform für Operational Intelligence

Die Hauptfunktionen des Systems Splunk Enterprise sind die Aufbereitung und Analyse von Log-Dateien aus dem Backend. Anhand mehrerer Komponenten (siehe Abbildung 12) werden die Maschinendaten unabhängig von Format und Speicherort erfasst, gesammelt, indexiert und den Usern zur Verfügung gestellt.

Die User können in diesen Daten Suchen ausführen, sich Dashboards erstellen und sich über das Erreichen von Schwellwerten via Alerting benachrichtigen lassen.

Ein Dashboard besteht in Splunk aus Teilfenstern. Diese können Module wie Suchfelder, Diagramme, Tabellen und Formulare enthalten. Die Teilfenster eines Dashboards sind mit der gespeicherten Suche verknüpft, sodass die Ergebnisse einer abgeschlossenen Suche, sowie die Daten aus einer im Hintergrund ausgeführten Echtzeitsuche in diesem angezeigt werden können.[136]

Das Durchsuchen der Daten erfolgt mit Hilfe der Splunk Search Processing Language (SPL). Mit Hilfe der Suche können z.B. statistische Suchen durchgeführt, Metriken berechnet, Trends, Spitzen und Muster erkannt werden.[137]

Splunk ermöglicht es, die Auswertung einer zuvor erstellten Suche in einem Dashboard zu visualisieren, sowie Berichte zu erstellen, um die Ergebnisse verständlicher und aussagekräftiger darzustellen.[138]

4.5 Komponenten von Splunk Enterprise

Splunk Enterprise kann auf einem einzelnen Client installiert und betrieben werden und stellt damit bereits eine vollständige Splunk-Umgebung dar. Es besteht die Möglichkeit Log-Daten vom eigenen System, sowie von anderen Systemen entgegenzunehmen, zu indizieren und für Suchanfragen bereit zu halten.[139] Der Zugriff und die Bedienung erfolgen über eine Weboberfläche.

Wird aus Performance oder infrastrukturellen Gründen dieser zentralisierte Einzelaufbau den Ansprüchen nicht gerecht, sind die einzelnen Splunk-Komponenten auf separate Systeme zu verteilen. Dabei übernimmt jede der Komponenten eine spezielle Aufgabe.[140]

---

[136] Vgl. Splunk Inc., Splunk Schnellreferenz
[137] Vgl. Splunk Inc., Splunk® Enterprise – Die Plattform für Operational Intelligence
[138] Vgl. ebd.
[139] Vgl. Splunk Inc., Splunk® Enterprise – Distributed Deployment Manual - Scale your deployment with Splunk Enterprise components
[140] Vgl. ebd.

In einer verteilten Splunk Enterprise Umgebung sind die folgenden Komponenten aufgeführt: Search Head, Indexer, Forwarder und Management Server.[141] Die Abbildung 12 stellt den Aufbau der einzelnen Splunk-Komponenten in einer verteilten Umgebung dar.

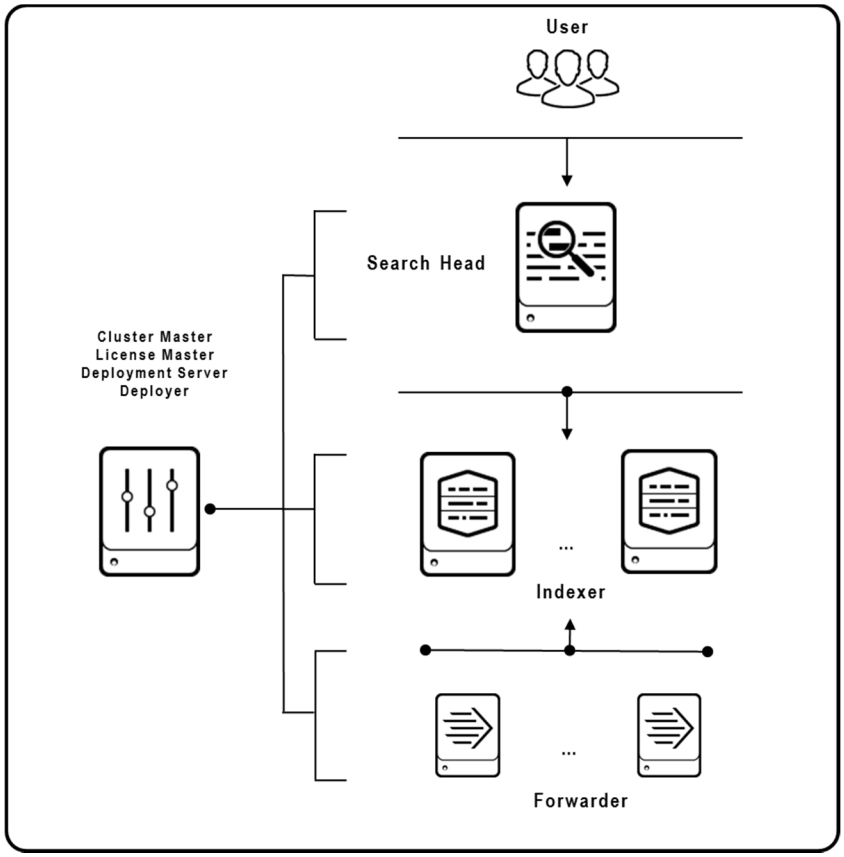


Abb. 12 Komponenten einer verteilten Umgebung
Quelle: Eigene Darstellung in Anlehnung an Splunk Inc., Splunk® Enterprise – Distributed Deployment Manual - Small enterprise deployment: Single search head with multiple indexers

Search Head (Suchkomponente)

Ein Search Head dient zur Bereitstellung der Daten für die User. Diese Komponente stellt das Frontend für den Benutzer dar. Der normale Benutzer hat nur diese Komponente als Einstiegspunkt und keine weiteren Splunk Komponenten. Der User greift über eine Suchfunktion auf die indizierten Daten zu und kann somit Suchen auf den Daten durchführen. Die Ergebnisse der durchgeführten Suche liefert der Indexer. Dabei nimmt der Search Head die Antworten vom Indexer entgegen und stellt sie dem

[141] Vgl. Splunk Inc., Splunk® Enterprise – Distributed Deployment Manual - Scale your deployment with Splunk Enterprise components

Benutzer dar. Die durchgeführten Suchen können im Search Head gespeichert werden, um zu einem späteren Zeitpunkt auf diese erneut zuzugreifen.[142]

Indexer (Speicher- und Indizierungskomponente)

Ein Indexer dient in der Splunk-Umgebung zur Datenindizierung bzw. Datenhaltung. Die eingehenden Rohdaten wandelt der Indexer in Ereignisse (Events) um und speichert die Ereignisse in einem Index. Außerdem werden hier die indizierten Daten bei Suchanfragen durchsucht.[143] Um aus den Rohdaten z.B. einer Log-Datei, die in Splunk eingeht, ein suchbares Event zu erzeugen, wird die Data Pipeline durchlaufen. Dabei durchlaufen die Log-Daten erst eine Parsing Pipeline und anschließend eine Indexing Pipeline. Die Parsing Pipeline befindet sich in der Regel auf einem Indexer, kann sich aber auch auf einem Heavy Forwarder befinden.[144] Die Data Pipeline besteht insgesamt aus den Segmenten Input, Parsing, Indexing und Search:[145]

Beim Segment Input wird zunächst der eingehende Rohdatenstrom einer angeschlossenen Quelle erfasst. Der erfasste Rohdatenstrom einer Quelle wird in 64K-Blöcke zerteilt und jeder dieser Blöcke mit Metadaten annotiert. Anschließend werden diese an das Segment Parsing weitergegeben.[146]

Parsing ist das zweite Segment der Data Pipeline. Hier werden die Daten analysiert und transformiert, bevor diese im Segment Indexing indexiert werden.[147] Durch den Vorgang des Indexing werden die zuvor analysierten und transformierten Daten in einen Index auf die Festplatte geschrieben.[148] Anschließend sind durch definierte Suchen eines Users, die Events eines Index abrufbar und es können Dashboards, sowie Reports aus den Ergebnissen der Suche erzeugt werden.[149]

Die folgende Abbildung stellt die zuvor beschriebenen Segmente in der Data Pipeline dar, um den Ablauf zu verdeutlichen.

---

[142] Vgl. Splunk Inc., Splunk - Schnellreferenz
[143] Vgl. ebd.
[144] Vgl. Splunk Inc., Splunk® Managing Indexers and Clusters of Indexers - Event processing and the data pipeline
[145] Vgl. Splunk Inc., Splunk® Enterprise – Distributed Deployment Manual - How data moves through Splunk Enterprise: the data pipeline
[146] Vgl. ebd.
[147] Vgl. ebd.
[148] Vgl. ebd.
[149] Vgl. Splunk Inc., Splunk® Enterprise – Distributed Deployment Manual - How data moves through Splunk Enterprise: the data pipeline

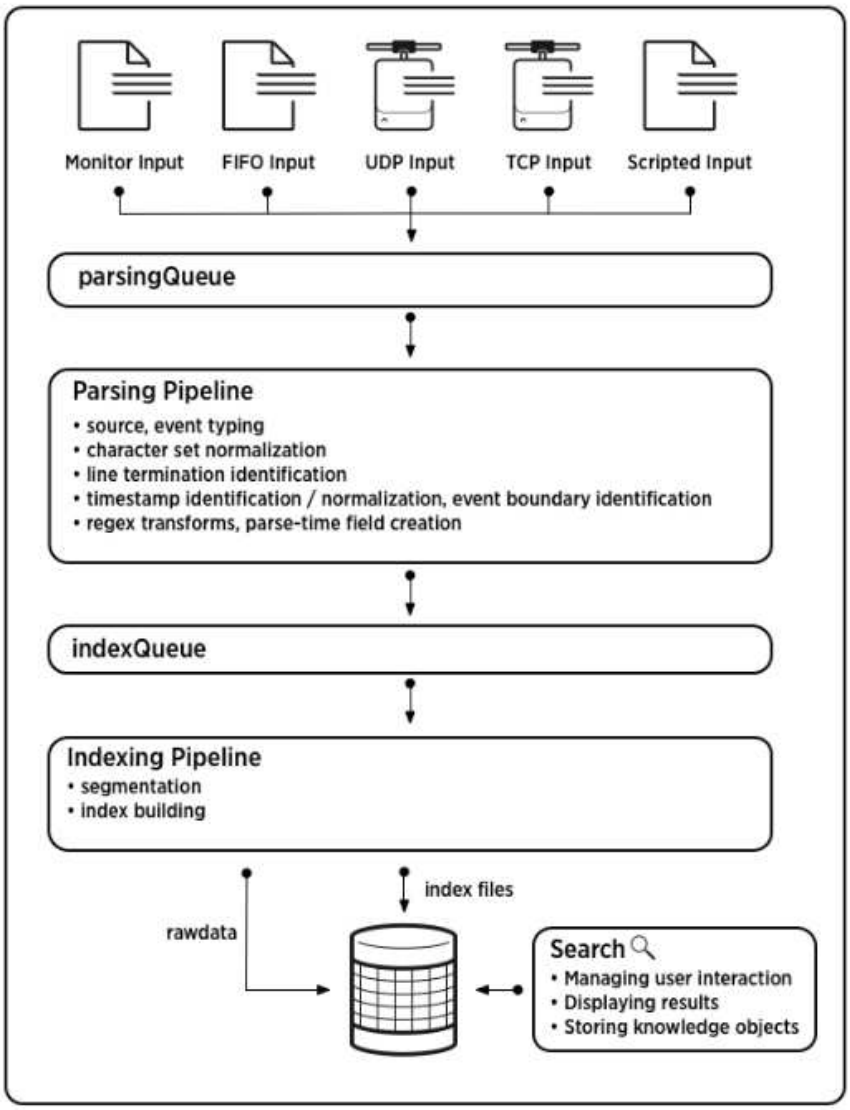


Abb. 13 Aufbau der Data Pipeline
Quelle: Splunk Inc., Splunk® Enterprise – Distributed Deployment Manual - How data moves through Splunk Enterprise: the data pipeline

Forwarder (Einsammlungskomponente)

Ein Forwarder ist eine Splunk Enterprise Instanz, die Daten z.B. an einen Indexer weiterleitet. Dabei wird zwischen Heavy Forwarder und Universal Forwarder unterschieden:[150]

Ein Heavy-Forwarder dient primär dazu, Daten von Quellsystemen aller Art einzusammeln und diese in einem gesammelten Datenstrom an die Indexer weiterzuleiten, um die Daten dort zu indizieren. Dabei leitet der Heavy-Forwarder die Rohdaten allerdings nicht einfach an die Indexer weiter, sondern führt bereits vorweg

[150] Vgl. Splunk Inc., Splunk® Enterprise – Distributed Deployment Manual - Forwarders

den Vorgang des Parsing durch. Dadurch entfällt das Parsing beim Indexer und der Indexer muss somit nur noch die Indexierung vornehmen.[151]

Universal Forwarder werden ebenfalls verwendet, um Daten von Quellsystemen aller Art einzusammeln und diese an eine weitere Splunk Enterprise Instanz weiterzuleiten bzw. zu verteilen. Ein Universal Forwarder beherrscht allerdings kein Parsing, sodass die Daten als Rohdaten weitergeleitet werden. [152]

Management Server (Steuerungskomponete)

Der Management Server ist die zentrale Komponente der Splunk-Umgebung und dient zur Verwaltung und Steuerung der zuvor genannten Komponenten. Dabei kann der Management Server verschiedene Aufgaben bzw. Rollen übernehmen. Zu diesen zählen:[153]

- Deployment Server
- Licence Master
- Cluster Master
- Deployer

Der Deployment Server dient zur Verteilung von Konfigurationen an die sich in der verteilten Umgebung befindlichen Komponenten.[154]

Die Verwaltung der Lizenzen der jeweiligen Komponente übernimmt der Licence Master.[155]

Der Cluster Master, auch genannt „Master Node“, dient zur Koordination eines Index Cluster. Zur Koordination gehört die Steuerung der Replikation[156] zwischen den Indexern, sowie die Mitteilung eines anfragenden Search Head, auf welchem Indexer die gesuchten Daten vorliegen. Außerdem unterstützt der Cluster Master bei einem Ausfall eines Indexer. Dabei leitet er notwendige Aktionen ein, um die Datenbestände, auf die nicht mehr zugegriffen werden kann, auf einem anderen Indexer verfügbar zu

---

[151] Vgl. Splunk Inc., Splunk® Enterprise – heavy forwarder
[152] Vgl. Splunk Inc., Splunk® Enterprise – Forwarding Data
[153] Vgl. Splunk Inc., Splunk® Enterprise – Distributed Deployment Manual - Components that help to manage your deployment
[154] Vgl. Splunk Inc., Splunk® Enterprise – deployment server
[155] Vgl. Splunk Inc., Splunk® Enterprise – license master
[156] Die Index-Replikation ist ein Prozess, bei dem die auf einem Indexer befindlichen und zu einem Index gehörenden Roh- und indizierten Daten durch einen festgelegten Replikationsfaktor als Kopien auf eine Anzahl Indexer, die sich im Cluster befinden, repliziert werden. Dadurch verfügt der Cluster über genau diese Anzahl an Kopien von Rohdaten und es wird eine Ausfallsicherheit hergestellt. Vgl. Splunk Inc., Splunk® Enterprise 6.0.8 - Managing Indexers and Clusters of Indexers, S. 70

machen.[157] Der Deployer übernimmt die Verteilung der Konfigurationen an einen Search Head Cluster.[158]

4.6 Architektur der Splunk-Umgebung in der Otto GmbH & Co. KG

Der zuvor beschriebene Aufbau einer verteilten Splunk-Umgebung ist auch in der Architektur der Splunk-Umgebung der Otto GmbH & Co. KG zu finden. Die Komponenten Search Head, Indexer, Forwarder und der Management Server werden in dieser Umgebung durch weitere Komponenten ergänzt. Die Abbildung 14 verdeutlicht den Architekturaufbau der Splunk-Umgebung im Log-Management der Otto GmbH & Co. KG.

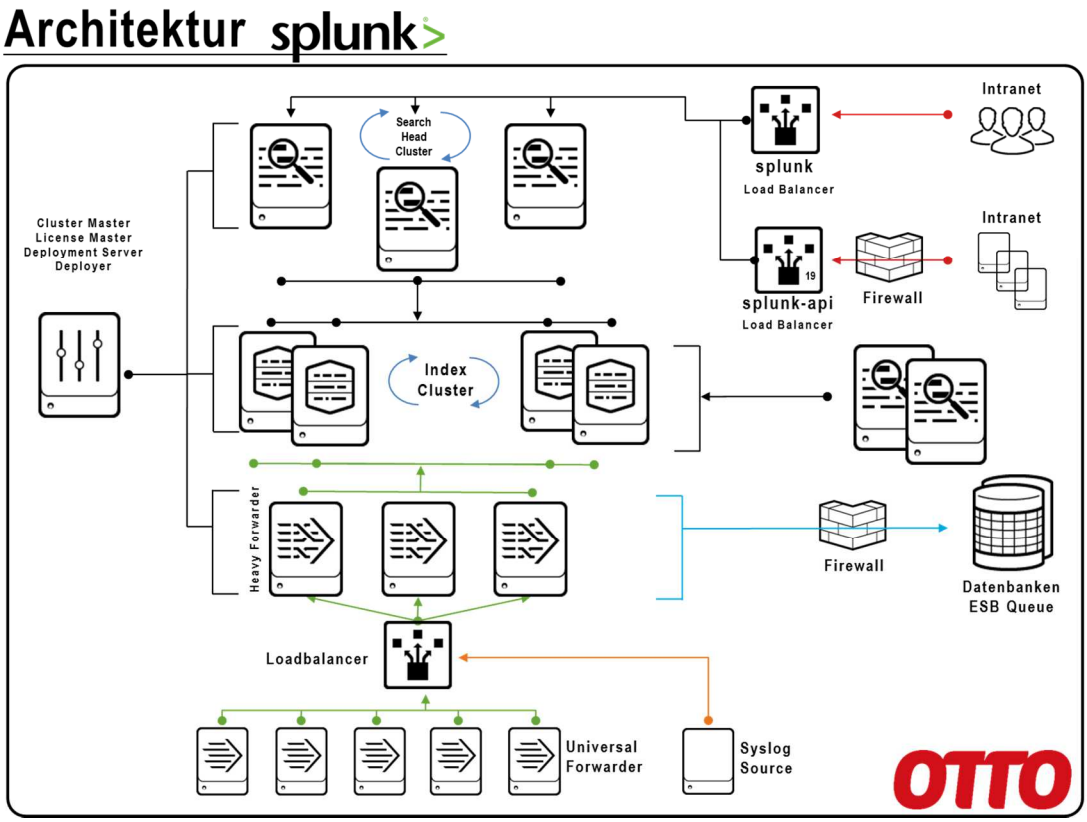


Abb. 14 Architektur der Splunk-Umgebung in der Otto GmbH & Co. KG
Quelle: Otto GmbH & Co. KG: Aktuelle Architektur Splunk, Stand 2016

Universal Forwarder

Die Universal Forwarder werden verwendet, um Daten von Quellsystemen aller Art einzusammeln und leiten diese unverarbeitet als Rohdaten über einen Loadbalancer an

[157] Vgl. Splunk Inc., Splunk® Enterprise 6.0.8 - Managing Indexers and Clusters of Indexers, S. 75
[158] Vgl. Splunk Inc., Splunk® Enterprise – deployer

die Heavy-Forwader weiter.[159] Zu den Quellsystemen zählen z.B. Endgeräte wie Server. Ein Universal Forwarder entspricht einem Agenten auf dem Quellsystem. In der Splunk-Umgebung der Otto GmbH & Co. KG sind derzeit ca. 300 Universal Forwarder im Einsatz.

Loadbalancer

Der Loadbalancer dient zur Lastverteilung eingehender Daten und Anfragen und bietet die Möglichkeit zur Skalierung der verarbeitenden Infrastruktur.

Heavy Forwarder

Die Heavy Forwarder werden primär verwendet, alle eingehenden Verbindungen von den Universal Forwardern, sowie Syslog[160] Verbindungen anzunehmen und diese in einem gesammelten Datenstrom an die Indexer weiterzuleiten, um die Daten dort zu indizieren. Zuvor führen die Heavy Forwarder, wie oben beschrieben noch den Vorgang des Parsings durch. Derzeit sind drei Heavy Forwarder im Einsatz. Einer der Heavy Forwarder ist dafür zuständig, die Daten aus den Datenbanken, die an Splunk angebunden sind, einzusammeln und diese ebenfalls an die Indexer zur Indizierung weiterzuleiten.

Indexer

In der Architektur existieren vier Indexer, die in einem Cluster (Index-Cluster) miteinander verbunden sind. Dabei wird die Index-Replikation mit einem Replikationsfaktor von drei betrieben. Durch einen Replikationsfaktor von drei wird sichergestellt, dass die Informationen in dem Cluster in einer Mindestzahl von drei Kopien vorhanden sind und eine Ausfallsicherheit hergestellt wird.[161]

Search Head

In der Splunk-Umgebung befinden sich drei Search Heads, die ebenfalls in einem Cluster betrieben werden. Durch einen vorgeschalteten Load Balancer wird eine Lastverteilung der Suchanfragen ermöglicht. Die Search Heads in einem Cluster werden als member bezeichnet, wobei ein Clustermember eine spezielle Rolle übernimmt und captain genannt wird. Der captain koordiniert unter anderem die Replikation und das

[159] Vgl. Splunk Inc., Splunk® Enterprise – Forwarding Data
[160] Syslog ist ein Protokoll, das Netzwerkgeräte wie z.B. Router oder Switches zur Übermittlung von Log-Meldungen an einen zentralen Server verwenden. Dies ermöglicht es die Komponenten zu überwachen. Vgl. Bernstein, H. (2015), S. 398
[161] Vgl. Splunk Inc., Splunk® Enterprise 6.0.8 - Managing Indexers and Clusters of Indexers, S.77

Job-Scheduling unter den anderen membern. Falls der captain ausfällt, übernimmt ein anderer member seine Rolle.[162]

Management Server

Der Management Server dient als zentrale Komponente der Splunk-Umgebung und verfügt über die schon im vorherigen Abschnitt beschriebenen Rollen: Deployment Server, Licence Master, Cluster Master, sowie Deployer.

Durch die Verwendung der Software Splunk Enterprise können im Backend der Otto GmbH & Co. KG die Log-Dateien unabhängig von der Herkunft der Daten und des Datenformats aus den verschiedensten angeschlossenen Datenquellen an einem zentralen Ort bereitgestellt werden, um anschließend aus diesen Daten wichtige Informationen zu erhalten.

So besteht zudem mit Hilfe der Software Splunk Enterprise die Möglichkeit in den Datenmengen (derzeit ca. 13 Milliarden Events in 36 Tagen) Anomalien zu erkennen, die auf ein vom Normalverhalten abweichendes Verhalten hinweisen. Der Stand der Anomalie-Erkennung in der Otto GmbH & Co. KG wird im nächsten Kapitel beschrieben.

4.7 Anomalie-Erkennung in der Otto GmbH & Co. KG

Im Log-Management der Otto GmbH & Co. KG werden Anomalien derzeit weder durch statistische Verfahren, noch automatisiert mit Hilfe von Machine-Learning-Algorithmen und Technologien erkannt. Zur Erkennung von Anomalien auf den Dashboards wird die Methode des „scharfen Hinsehens“ verwendet, die personelle Ressourcen bindet und sehr zeitaufwändig ist. Dabei werden zwar Schwellwerte anhand der Methode des „scharfen Hinsehens“ ermittelt und darauf Eskalationsstufen angesetzt, dies erfolgt allerdings statisch und kann zu Fehlalarmen führen. Zudem birgt eine manuelle Erkennung ein hohes Fehlerpotenzial, welches dem menschlichen Faktor geschuldet ist. Diese personellen Ressourcen können im operativen Bereich der Otto GmbH & Co. KG für andere Tätigkeiten wertvoller eingesetzt werden.

Um die Herangehensweise der Anomalie-Erkennung mit Hilfe der Methode des „scharfen Hinsehens“ zu beschreiben, wird die Analyse der unqualifizierten Eventmengen über einen bestimmten Zeitraum, beispielsweise der letzten 35 Tage näher betrachtet. Die Gesamtmenge an Events, die in Log-Dateien von Datenbanken, Netzwerkkomponenten und Servern erzeugt bzw. abgelegt werden, sind als

[162] Vgl. Splunk Inc., Splunk® Enterprise - Distributed Search - About search head clustering

unqualifizierten Eventmengen zu bezeichnen. Diese werden in der Splunk-Umgebung indiziert. Dabei werden bei unqualifizierten Eventmengen keine Details betrachtet, sondern lediglich die Menge der eintreffenden Events (Counts). Mit Hilfe einer Suche in Splunk besteht die Möglichkeit sich die Anzahl der Events über einen „timechart“[163] in einem Liniendiagramm darstellen zu lassen und somit mögliche Auffälligkeiten zu erkennen. Um einen detaillierteren Verlauf zu erhalten, sind die Anzahl der Events, die an einem Tag anfallen in einem Intervall von einer Stunde aufgeführt. Die folgende Abbildung stellt das beschriebene Szenario dar und soll verdeutlichen, dass an dieser Stelle mit Hilfe der Methode des „scharfen Hinsehens“ die Erkennung von Anomalien nicht ausreicht. Es kann auf den ersten Blick nicht genau erkannt und beschrieben werden, wo sich ein Ausreißer befindet und ob es sich dabei tatsächlich um einen handelt.

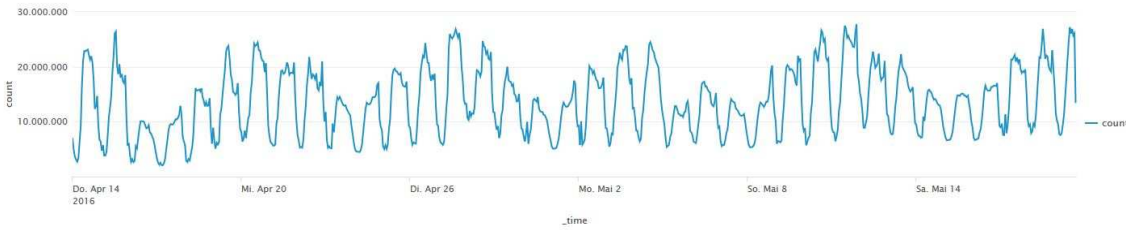


Abb. 15 Anzahl unqualifizierter Eventmengen der letzten 35 Tage

Aus diesem Grund erfolgt in Kapitel 5.2 eine Analyse, wie Anomalien unter Verwendung von Data Mining-Methoden und dem Einsatz von Machine-Learning-Algorithmen und -Technologien anhand von unqualifizierten Eventmengen in der Backend-Umgebung erkannt werden können. Für die Erkennung von Anomalien soll dabei ein Grundverständnis mit statistischen Verfahren, sowie mit Hilfe von Machine-Learning-Algorithmen und –Technologien geschaffen werden, wobei der Fokus deutlich auf der Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen und -Technologien liegt. Grundlage für die Untersuchung im Kapitel 5.2 sind Auswertungen von Experteninterviews, die im Rahmen der Bachelorarbeit durchgeführt wurden, sowie die im Kapitel 3 erarbeiteten wissenschaftlichen Grundlagen.

[163] timechart ist ein Suchkommando von Splunk, das tabellarische Ergebnisse in einer zeitbasierten Darstellung zurückliefert. Vgl. Splunk Inc., Splunk - Schnellreferenz

5 Empirische Analyse

Nachdem in den vorherigen Kapiteln der Stand der Forschung, die wissenschaftlichen Grundlagen und die Ausgangssituation der Otto GmbH & Co. KG aufgeführt und beschrieben wurden, findet im Folgenden die Empirische Analyse statt, die sich aus leitfadengestützten Experteninterviews, sowie der eigenen Analyse zusammensetzen.

5.1 Das leitfadengestützte Experteninterview

Für die Durchführung von Interviews mit Experten wurde die Methode eines leitfadengestützten Experteninterviews ausgewählt. Ein leitfadengestütztes Experteninterview ist eine Methode, die der qualitativen Forschung zugeordnet ist. Bei dem leitfadengestützten Interview handelt es sich um ein offenes, teilstrukturiertes und qualitatives Interview. Den Befragten liegt ein erarbeiteter Leitfaden mit offen formulierten Fragen zur Orientierung vor, auf die die Befragten frei antworten können.[164] Der Leitfaden dient zur Eingrenzung der Thematik und soll sicherstellen, dass bei großen Ausschweifungen der Experten hinsichtlich der Thematik, dem Interviewer eine Rückkehr zum Leitfaden ermöglicht wird.[165] Durch die Befragung von verschiedenen Experten und der Verwendung des gleichen Leitfadens können unterschiedliche Antworten und Meinungen gesammelt und diese anschließend miteinander verglichen werden. Zudem kann ein vielfältiges Ergebnis erzielt werden, da nicht zwingend mit widerspruchsfreien Aussagen zu rechnen ist.

Das Experteninterview stellt eine besondere Form des Leitfadeninterviews dar. Dabei zählen zu den Besonderheiten bei einem Experteninterview vor allem die Definition und Auswahl der Experten. Der Expertenstatus ist abhängig von der ausführenden Position bzw. Funktion des jeweiligen Experten.[166] Die Auswahl der Experten ist in Kapitel 5.1.1 aufgeführt.

Der für die Durchführung der Experteninterviews entwickelte Leitfaden setzt Kenntnisse in der Thematik voraus. Um den Leitfaden zu strukturieren, ist dieser in Themenbereiche aufgeteilt worden. Der Leitfaden setzt sich dabei aus vier Bereichen zusammen:

- Anomalie-Erkennung
- Technologien bzw. Applikationen zur Anomalie-Erkennung
- Statistische Anomalie-Erkennung

---

[164] Vgl. Mayer, H. (2013), S. 37
[165] Vgl. Flick, U. (1999), S. 112 f.
[166] Vgl. Scholl, A. (2015), S. 68 f.

- Anomalie-Erkennung mit Hilfe von Machine-Learning

Auf diesen vier Bereichen liegt im Interview der Fokus. Der Leitfaden ist im Anhang 1 und 2 zu finden.

5.1.1 Auswahl der Experten

In der vorliegenden Untersuchung soll Expertenwissen in den Bereichen „Machine-Learning“, „Anomalie-Erkennung“ und der „Datenanalyse“ erfasst werden. Deshalb wurde bei der Auswahl der zu befragenden Experten darauf geachtet, dass in den genannten Themengebieten Erfahrungen und Fachwissen vorliegen.

In der folgenden Tabelle sind die Namen der Experten mit denen ein Interview geführt wurde, die Position und Funktion der Experten, eine kurze Begründung für die Auswahl des jeweiligen Experten, sowie die gewählte Interviewform aufgeführt.

Name	Position/Funktion	Begründung der Auswahl	Interviewform
Philipp Drieger	Sales Engineer bei Splunk Inc., Experte für Machine-Learning	Erfahrungen mit der Machine-Learning App von Splunk, Projekte im Bereich Machine-Learning umgesetzt	Webkonferenz
Mika Borner	Management Consultant bei LC Systems, Experte für Anomalie-Erkennung	Use-Cases im Bereich der Anomalie-Erkennung umgesetzt	Webkonferenz
Christian Günther	Leiter Data Analytics bei LC Systems, Experte Datenanalyse	Langjährige Erfahrungen im Bereich Datenanalyse und Verwendung der Software Splunk	Persönlich

Tab. 3 Auswahl der Experten für die Experteninterviews

5.1.2 Durchführung des Experteninterviews

Bevor das Interview mit den Experten geführt wurde, sind diese über das Vorhaben informiert worden, um daraufhin der Teilnahme zur Durchführung des Interviews zuzustimmen. Anschließend wurden die Leitfäden mit den formulierten Fragen zur Vorbereitung auf das Interview den Experten zur Verfügung gestellt. Danach ist ein Termin zur Durchführung des Interviews vereinbart worden. Die Durchführung der Interviews fand im Zeitraum vom 05.04.2016 bis 20.04.2016 statt. Zu Beginn jedes

Interviews wurde den Befragten kurz eine Einleitung in das Thema gegeben. Außerdem fand eine kurze Vorstellung der eigenen Person statt, sowie die Frage nach der Erlaubnis zur Aufzeichnung des Gesprächs. Anschließend wurden die Befragten gebeten, sich selbst kurz vorzustellen. Das erste Interview wurde persönlich und vor Ort durchgeführt. Die beiden anderen Interviews sind über eine Webkonferenz (Online-Meeting) durchgeführt worden. Eine detaillierte Übersicht der geführten Interviews liefert die Tabelle 3. Die beiden Interviews, die über eine Webkonferenz durchgeführt worden sind, wurden nach Zustimmung der Befragten zur Auswertung aufgezeichnet. Dies stellt sicher, dass es zu keinem Informationsverlust kommt und der Interviewer sich vollständig auf das Interview konzentrieren kann.

Aufgrund eines vom Leitfaden abweichenden Gesprächsverlaufs beim dritten Experteninterview, ist im Folgenden lediglich ein Vergleich der anderen beiden geführten Interviews möglich. Diese erfolgten mit Unterstützung des Leitfadens. Die Erkenntnisse vom dritten Experteninterview werden in der abschließenden Zusammenfassung aufgegriffen.

5.1.3 Auswertung des Experteninterviews

Der erste Schritt der Datenauswertung besteht in der Transkription, der in der Webkonferenz getätigten Aufnahmen. Dabei stehen die Informationen und der Inhalt des gesprochenen im Mittelpunkt der Experteninterviews. Aus diesem Grund wurde auf eine aufwändige Notation verzichtet. Um die Richtigkeit der Transkriptionen sicherzustellen, wurden diese von den Experten überprüft und zur Auswertung freigegeben. Durch die im Vorfeld getätigte Strukturierung des Leitfadens in Themenbereiche lassen sich Gemeinsamkeiten und Unterschiede der Experteninterviews aufzeigen. Dabei werden die Antworten zu den im Experteninterview gestellten Fragen miteinander verglichen und eine textnahe Paraphrasierung vorgenommen. Dadurch werden die für die Untersuchung relevanten Aspekte herausgefiltert. Die Ergebnisse der Auswertung werden im Folgenden dargestellt.

5.1.4 Ergebnisse der Auswertung

Hinsichtlich der Begriffsdefinition der Anomalie herrscht bei beiden Experten Einigkeit. Für beide Experten ist unter einer Anomalie eine Abweichung von einem erwarteten

Verhalten zu verstehen. Außerdem betonen beide Experten, dass das erwartete Verhalten bekannt sein muss, um eine Anomalie zu erkennen.[167]

Für die Erkennung von Ausreißern bzw. Anomalien erwähnen beide, dass eine „Baseline“ bzw. das Setzen eines Schwellenwerts erfolgen muss. Dabei ist auf statistische Methoden zurückzugreifen.[168] Borner bezieht sich dabei auf ein konkretes Beispiel. Dabei betrachtet er die „Responsezeiten“ einer Webapplikation, um zu analysieren wie lange der Aufruf einer Webseite gedauert hat. „Dann versucht man eine Baseline zu erstellen und bei Abweichungen der Baseline zu alarmieren.“[169]

Bei der Erkennung von Anomalien können einige Herausforderungen auftreten. Hier wird von beiden Experten die Auswahl eines geeigneten Algorithmus genannt, bei dem die Fehlerrate (false/positives) gering und der Grad der Automatisierung hoch ist.[170] Grund hierfür ist laut Drieger, dass bei zu vielen false/positives das Verfahren unbrauchbar ist und das Problem der Erkennung von Anomalien deshalb sehr vielschichtig sein kann.[171]

Borner greift hier das Beispiel der Responsezeiten wieder auf. „Also bei einer Responsezeit die zwischen 10 und 100 Millisekunden liegt, erwarte ich, dass eine Anomalie vielleicht bei 120 Millisekunden liegt. Aber es kann durchaus sein, dass die Anomalie unter 10 Millisekunden auftritt, weil das System zu schnelle Antworten liefert. Dieses Beispiel zeigt, dass es immer unerwartete Dinge gibt und wenn man dann den falschen Algorithmus wählt, dann verpasst man solche Anomalien.“[172]

Hinsichtlich des Einsatzgebiets der Anomalie-Erkennung führen beide Experten den Security-Bereich, sowie den Bereich des IT-Betriebs (IT-Operations bzw. Application Performance Monitoring) auf. So spielen in diesen Bereichen die frühzeitige Erkennung von Systemausfällen und Beeinträchtigungen von Komponenten, sowie die Erkennung von ungewöhnlichen Backend-Zugriffen eine wesentliche Rolle, um schnellstmöglich reagieren zu können.[173]

---

[167] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 1.1 und Borner, M. (2016), Management Consultant Anhang 2, Frage 1.1
[168] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 1.2 und Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 1.6
[169] Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 1.6
[170] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 1.4 und Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 1.4
[171] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 1.4
[172] Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 1.4
[173] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 1.5 und Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 1.7

Drieger und Borner empfehlen zur Anomalie-Erkennung die bereits integrierten Befehle innerhalb der Software Splunk, sowie die Machine-Learning App von Splunk. Borner nennt zusätzlich noch eine App von Prelert, die er zur Erkennung von Anomalien favorisiert.[174]

In der Erkennung von Anomalien mit Hilfe von statistischen Mitteln, sowie der Erkennung von Anomalien mit Hilfe von Machine-Learning-Algorithmen sehen Drieger und Borner beide eine Gemeinsamkeit. So sind die statistischen Mittel als Basis der Machine-Learning-Algorithmen anzusehen, die stets weiter auf diesen aufbauen. [175]

Bei der Vorgehensweise der Anomalie-Erkennung mit Hilfe von statistischen Mitteln sprechen beide Experten an, dass zu Beginn der Datenanalyse geeignete Felder bzw. Feldwerte in den Daten gefunden werden müssen. Als hilfreiches Kommando nennen beide das „timechart"-Kommando von Splunk.[176]

Einen wichtigen Aspekt, um Anomalien in den Daten zu erkennen, spielt die Aufbereitung der Daten, um eine gute Grundlage zu schaffen. Sowohl bei der Erkennung von Anomalien mit Hilfe statistischer Methoden als auch mit Hilfe von Machine-Learning-Algorithmen. [177] Laut Drieger sind „Konsistenz und bestmögliche Qualität der Daten …eine gute Grundlage für die Anomalie-Erkennung". [178] Beispielweise sind mit dem Splunk-Kommando „analyzefields" geeignete Felder für die Vorhersage zu identifizieren.[179]

Die Extraktion und Identifizierung von Feldern wird hinsichtlich der Durchführung der Datenaufbereitung bzw. Datenvorbereitung von beiden Experten aufgeführt.[180] So können laut Drieger „mit Hilfe von Splunk, aus den unstrukturierten Daten zunächst diejenigen Felder extrahiert werden, die für die Analyse wichtig sind. Dabei gibt es bereits die Möglichkeit, die Qualität der Regex basierten Feldextraktion zu prüfen und ggf. zu korrigieren. Ferner gibt es weitere Splunk Befehle, um die Daten z.B. anzureichern. „lookup" wäre hier zu nennen oder zusätzliche Felder zu erzeugen. Da

---

[174] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 2.1 und Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 2.1
[175] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 2.6 und Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 2.6
[176] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 3.1 und Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 3.1
[177] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 3.2, 4.1, 4.4 und Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 3.2, 4.4
[178] Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 3.2
[179] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 4.1
[180] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 3.3, 4.4, 4.5 und Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 3.2, 4.4

wäre der Befehl „eval“ (für Evaluation von Feldern) zu nennen.“[181] Borner ergänzt, dass für die App von Prelert zur Erkennung von Anomalien die Konvertierung von Zeiteinheiten (z.B. bei einer Mischung aus Milli- und Nanosekunden) zu beachten ist. [182] Zudem müssen die Feldwerte korrekte Werte beinhalten, da nach der Extraktion von Feldern, Prelert anhand der Feldnamen die Anomalie berechnet. [183]

Zur Erkennung von Anomalien mit Hilfe von Machine-Learning-Algorithmen betrachten beide Experten die jeweilige App mit der Sie bereits Erfahrungen gesammelt haben. Bei Drieger ist dies die Machine-Learning-App von Splunk und bei Borner die Anomaly Detective App von Prelert.[184]

Lediglich bei der Machine-Learning-App von Splunk sind die unterstützten Algorithmen öffentlich bekannt.[185] So unterstützt laut Drieger die Machine-Learning-App von Splunk ab der Version 1.0 die Algorithmen „Birch, DBSCAN, KMeans, LinearRegression, LogisticRegression, PCA, SVM, SpectralClustering, BernoulliNB, ElasticNet, FieldSelector, GaussianNB, KernelPCA, Lasso, OneClassSVM, RandomForestClassifier, RandomForestRegressor, Ridge, Standardscaler, TFIDF.“ [186]

Unabhängig von der verwendeten App zur Anomalie-Erkennung raten beide Experten, zu Beginn auf die herkömmlichen Splunk-Befehle zurückzugreifen und die Komplexität gering zu halten.[187] So verweist Drieger nach einer ersten Datenexploration auf die Befehle „timechart“, „stats“, „cluster“, „kmeans“ und „anomal*“. Sollten durch die Verwendung der genannten Befehle keine zufriedenstellenden Ergebnisse erreicht werden, ist die Machine-Learning App einzusetzen. [188]

Borner empfiehlt innerhalb der Prelert App die Verwendung von integrierten Funktionen im Ad-hoc-Modus, um Anomalien zu erkennen.[189] Sind die Ergebnisse zufriedenstellend, ist eine Realtime-Suche einzurichten mit der fortlaufend eine Baseline erzeugt wird. [190]

---

[181] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 3.3
[182] Vgl. Borner, M. (2016), Management Consultant, Experteninterview Anhang 2, Frage/Antwort 3.2
[183] Vgl. Borner, M. (2016), Management Consultant, Experteninterview Anhang 2, Frage/Antwort 4.4
[184] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 4.2 und Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 4.2
[185] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 4.3 und Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 4.3
[186] Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 4.3
[187] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 4.6 und Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 4.8
[188] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 4.6
[189] Vgl. Borner, M. (2016), Management Consultant, Experteninterview Anhang 2, Frage/Antwort 4.8
[190] Vgl. Borner, M. (2016), Management Consultant, Experteninterview Anhang 2, Frage/Antwort 4.1

Das Trainieren eines Verhaltens mit einem Algorithmus in der jeweiligen App gestaltet sich bei der Machine-Learning-App von Splunk und der App von Prelert unterschiedlich.[191]

So beschreibt Drieger, dass das Training eines Modells und das Anwenden des gelernten Modells auf einen Testdatensatz mit Hilfe der Befehle „fit“ und „apply“ in der Machine-Learning-App durchgeführt wird. [192]

Borner beschreibt das in der Prelert App eine Baseline zu erstellen ist (z.B. der letzten 30 Tage), die sich konstant aktualisiert und aufgrund historischer Daten das Verhalten trainiert. [193]

Beide Experten erwähnen hinsichtlich der Durchführung des Trainings und der damit verbundenen Aufteilung in Trainings- und Testdaten, das Splitting in verschiedene Zeitabschnitte. [194]

So erfolgt das Training laut Drieger „durch eine Aufteilung des Datensets in Trainingsdaten und Testdaten in einem bestimmten Verhältnis, z.B. 80 zu 20. Dann muss entschieden werden, ob die Aufteilung durch zufälliges Sampling oder definierte Zeitabschnitte vorgenommen wird.“[195] Dabei kann „das Modell .. auf Testdaten oder aber auch auf neue Daten angewendet werden.“[196]

Laut Borner wird „ein Splitting in Trainings- und Testdaten .. bei der Prelert App indirekt vorgenommen, indem anhand der Durchsuchten Daten eine Baseline erzeugt wird und diese ein Verhalten mit Hilfe von Algorithmen lernt, welches auf einen anderen Satz an Daten angewendet wird. Mit einem anderen Satz an Daten ist einfach nur ein anderer Zeitraum gemeint.“[197] Dabei fügt Borner hinzu, dass „das gelernte .. nur auf den konkreten Fall angewendet werden [kann].“[198]

5.1.5 Zusammenfassung

Bei der Erkennung von Anomalien mit Hilfe von statistischen Methoden, sowie der Erkennung von Anomalien mit Hilfe von Machine-Learning-Algorithmen und -

---

[191] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 4.9 und Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 4.9
[192] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 4.9
[193] Vgl. Borner, M. (2016), Management Consultant, Experteninterview Anhang 2, Frage/Antwort 4.9
[194] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 4.10 und Borner, M. (2016), Management Consultant, Experteninterview Anhang 2, Frage/Antwort 4.10
[195] Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 4.10
[196] Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 4.12
[197] Borner, M. (2016), Management Consultant, Experteninterview Anhang 2, Frage/Antwort 4.10
[198] Borner, M. (2016), Management Consultant, Experteninterview Anhang 2, Frage/Antwort 4.11

Technologien spielt die Datenaufbereitung eine wichtige Rolle, um vorweg eine gute Datengrundlage zu schaffen. So besteht laut den Experten bei einigen Unternehmen großes Interesse Anomalien zu erkennen, nur sind diese zum Großteil damit beschäftigt die eigenen Daten erst einmal zu verstehen und aufzubereiten, um dann eine Anomalie-Erkennung mit Hilfe von Algorithmen durchzuführen.[199]

Der Anomalie-Erkennung kann sich zunächst mit Hilfe der Software Splunk Enterprise und den integrierten Kommandos, die vor allem einen statistischen Ansatz bieten, genähert werden. Um Anomalien in der Splunk-Umgebung mit Hilfe von Machine-Learning-Algorithmen zu erkennen, wird die Machine-Learning-App von Splunk oder die Anomaly Detective App von Prelert empfohlen.[200] Beide Apps können in die Software Splunk Enterprise integriert werden.

Als eine Herausforderung in der Anomalie-Erkennung sehen alle drei Experten die Wahl eines geeigneten Algorithmus. Dabei führen alle drei Experten auf, dass sowohl die Erkennung von Anomalien mit Hilfe von statistischen Mitteln, als auch die Erkennung von Anomalien mit Hilfe von Machine-Algorithmen ein komplexes Themengebiet sind.[201]

Wie bereits im Kapitel 2 aufgeführt, ist das Thema in der Forschung nicht unbekannt. Allerdings ist die Überführung in die Praxis nicht einfach realisierbar.[202]

5.2 Eigene Analyse anhand des KDD-Prozesses

In diesem Kapitel wird untersucht, wie Anomalien unter Verwendung von Data Mining-Methoden und dem Einsatz von Machine-Learning-Algorithmen anhand von unqualifizierten Eventmengen in der Splunk-Umgebung der Otto GmbH & Co. KG erkannt werden können. Zuerst wird beschrieben, welche Technologie bzw. Applikation, die in Splunk integriert werden kann und zudem Machine-Learning-Algorithmen unterstützt, ausgewählt wurde. Danach folgt die Durchführung der Analyse anhand des im Kapitel 3.5.1 aufgeführten KDD-Prozesses. Die Ergebnisse (Vorhersagegenauigkeiten) der angewendeten Algorithmen (Modelle) werden aufgezeigt und evaluiert. Abschließend wird der Algorithmus, der die beste Modellgüte aufweist auf einen unbekannten Datensatz angewendet.

---

[199] Vgl. Günther, C. (2016), Leiter Data Analytic, Experteninterview Anhang 3
[200] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 2.1 und Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 2.1
[201] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Borner, M. (2016), Management Consultant Anhang 2 und Günther, C. (2016), Leiter Data Analytic, Experteninterview Anhang 3
[202] Vgl. Günther, C. (2016), Leiter Data Analytic, Experteninterview Anhang 3

5.2.1 Auswahl einer Technologie zur Erkennung von Anomalien mit Hilfe von Machine-Learning-Algorithmen

Um in der Splunk-Umgebung die Erkennung von Anomalien mit Hilfe von Machine-Learning-Algorithmen zu untersuchen, ist eine Technologie bzw. Applikation in Splunk zu integrieren, die Machine-Learning-Algorithmen unterstützt. Hier sind in den durchgeführten Experteninterviews zwei Applikationen genannt worden: Die Machine-Learning App von Splunk (ML Toolkit and Showcase) und die App von Prelert (Prelert Anomaly Detective App).[203]

Die Entscheidung für das weitere Vorgehen ist auf die Machine-Learning App von Splunk gefallen, da diese im Vergleich zur Anomaly Detective App von Prelert ohne Lizenzkosten erhältlich ist und die integrierten Algorithmen öffentlich einsehbar sind. Gerade letzteres ist von hoher Bedeutung, da für die Algorithmen eine entsprechende Datenaufbereitung vorzunehmen ist.

Die Voraussetzung zur Installation der neusten Version der Machine-Learning App von Splunk (ML Toolkit and Showcase Version 1.1) sind eine Splunk Enterprise Version 6.4 und ein Python-Add-on (Python for Scientific Computing Add-on), das Machine-Learning Libraries enthält.[204]

Da zum Zeitpunkt der Arbeit in der Splunk-Umgebung der Otto GmbH & Co. KG auf den Komponenten die Splunk Enterprise Version 6.3.3 installiert ist, wurde auf einem lokalen Client die Splunk Enterprise Version 6.4 installiert, sodass anschließend die Machine-Learning App von Splunk und das benötigte Python-Add-on integriert werden konnten.

In den folgenden Abschnitten werden hinsichtlich der Untersuchung zur Erkennung von Anomalien die Prozessschritte des KDD-Prozesses durchlaufen und beschrieben.

5.2.2 Datenauswahl und -Beschreibung

Der erste Schritt des KDD-Prozesses betrifft die Auswahl und Beschreibung der Daten. In dieser Phase wird geprüft, welche Daten benötigt und verfügbar sind, um das Ziel (Anomalien zu erkennen) zu erreichen. Die Auswahl von Daten wird auch als Dataset bezeichnet.[205] Dabei werden aus den in der Splunk-Umgebung indizierten Daten die für

---

[203] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 2.1 und Borner, M. (2016), Management Consultant Anhang 2, Frage 2.1
[204] Vgl. Splunk Inc., Machine Learning Toolkit and Showcase
[205] Vgl. Sharafi, A. (2013), S. 62

die Analyse relevanten Daten (Datensätze) und Merkmale (Datenfelder) ausgewählt. Dieser Datenbestand wird anschließend in der Vorverarbeitungsphase bereinigt.

Für die Analyse und Erkennung von Anomalien wurden unqualifizierte Eventmengen ausgewählt, die in einem Zeitraum innerhalb von 36 Tagen angefallen sind. Bei der Auswahl der Daten wurden als Zeitraum die letzten 35 Tage mit einem Stundenintervall von einer Stunde ausgewählt, um eine möglichst hohe Auflösung (viele Datensätze) für die Analyse zu haben. Dabei handelt es sich bei den unqualifizierten Eventmengen lediglich um die Anzahl der Events (Counts), die über den genannten Zeitraum indiziert wurden. Mit Hilfe der folgenden Suche wird die Anzahl der Events aller Indizes über den Zeitraum der letzten 35 Tage unter Verwendung der Splunk Search Processing Language gezählt.

```
| tstats prestats=t count where index="*" OR index"_*" by _time span=1h
| timechart count span=1
```
(5.1)

Die folgende Tabelle stellt einen Ausschnitt des Ergebnisses dar. Das vollständige Ergebnis befindet sich auf dem anliegenden Datenträger unter der Bezeichnung „ba_count_letzte_35_tage_alle_indizes_als_tabelle.csv“.

_time	count
2016-04-26T00:00:00.000+0200	6378374
2016-04-26T01:00:00.000+0200	4421592
2016-04-26T02:00:00.000+0200	3237106
2016-04-26T03:00:00.000+0200	3694523
2016-04-26T04:00:00.000+0200	3492415

Tab. 4 Ausschnitt des Ergebnisses der Suche

Auf die durchgeführte Suche und das Ergebnis der Suche wird in der Datenvorbereitung und –Bereinigung, sowie im Prozessschritt der Transformation näher eingegangen.

5.2.3 Datenvorbereitung und –Bereinigung

Nachdem die Daten für die Analyse ausgewählt wurden, sind diese aufzubereiten und zu bereinigen (engl. Data Cleansing), um eine hohe Qualität des Datenbestandes für die weiteren Schritte zu gewährleisten.[206] Die Datenqualität ist für die Anwendbarkeit der Analyseverfahren von entscheidender Bedeutung. So sind in dieser Phase unter Umständen Inkonsistenzen zu beseitigen, beispielsweise durch die Entfernung

[206] Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 3.2

fehlender oder fehlerhafter Attributwerte oder durch die Anreicherung weiterer Attribute des bereits vorhandenen Datenbestands.[207]

Das Ergebnis der Suche (siehe Tabelle 4) stellt die Anzahl der Events (Counts), die in der Splunk-Umgebung indiziert wurden, der letzten 35 Tage in einem stündlichen Intervall dar. Dabei sind lediglich die Spalten „_time“ und „count“ erzeugt worden.

Diese Daten wurden zuerst auf Fehler und Auffälligkeiten untersucht. Die mit Hilfe der Suche erzeugten Daten sind von Beginn an fehlerfrei gewesen. Es war vorerst nicht notwendig, die unterschiedlichen Verfahren des „Data Cleansing“ anzuwenden, da keine doppelten Spalten, fehlerhaften Werte und auch keine fehlenden Werte in den zur Verfügung gestellten Daten vorliegen. Bei fehlenden Werten sind beispielsweise die folgenden Verfahren des Data Cleansing anzuwenden: [208]

- Verwendung einer globalen Konstante für die fehlenden Werte
- Verwendung des Mittelwerts oder des Medians, um die fehlenden Werte zu ersetzen
- Entfernung der fehlenden Werte, durch herausfiltern

Da in dieser Arbeit das überwachte Lernen und die Methode der Klassifikation zur Erkennung von Anomalien auf die Daten angewendet werden, sind diese Daten mit weiteren Attributen anzureichern. Auf die Auswahl der Data Mining-Methode wird im Kapitel 5.2.5 näher eingegangen. Beim überwachten Lernen ist für den Datensatz eine Spalte mit einem Data Label zu versehen, damit ein Algorithmus die Abbildung von den Eingabedaten auf die Ausgabedaten erlernen kann. So sind zur Erkennung von Anomalien mit Hilfe des Klassifikationsansatzes die Spalten „_time“ und „count“ mit weiteren Informationen anzureichern. Die Spalte für die Ausgabedaten wurde auf die Bezeichnung „verhalten“ festgelegt. Das „verhalten“ kann den Wert „Anomalie“ oder den Wert „Normal“ enthalten.

Um die Klassifikation für die Spalte „verhalten“ durchzuführen, ist eine untere und eine obere Grenze definiert worden. Diese legt fest, wann es sich um eine Anomalie handelt oder nicht.

---

[207] Vgl. Sharafi, A. (2013), S. 62
[208] Vgl. Han, J., Kamber, M., Pei, J. (2012), S. 88

Im Folgenden ist die Klassifikationsregel der Übersicht halber vereinfacht dargestellt und wird nachfolgend im Detail betrachtet:

IF percentageDifftag > upperBound OR percentageDifftag < lowerBound

THEN Anomalie ELSE Normal.

Das Attribut „percentageDifftag“ stellt die Differenz des aktuellen Counts zum Count vor einem Tag prozentual dar. Wie dieses Attribut in Splunk erzeugt worden ist, wird im Folgenden beschrieben.

Um die untere (lowerBound) und obere Grenze (upperBound) zu bestimmen, wurde jeweils ein konstanter Wert ausgewählt. Für die obere Grenze (upperBound) 45.0 und für die untere Grenze (lowerBound) -45.0. Die konstanten Werte von 45.0 und -45.0 wurden festgelegt, um bei der Klassifizierung eine angemessene Verteilung hinsichtlich des Verhaltens der Anomalie zu erreichen.

Die Attribute „upperBound“ und „lowerBound“ wurden in Splunk durch Verwendung des „eval“-Kommandos[209] erzeugt.

```
| eval upperBound= 45.0                                    (5.2)
| eval lowerBound= -45.0
```

Anschließend wurde mit Hilfe des „eval“-Kommandos das Feld „verhalten“ erzeugt und die Klassifikationsregel angewendet.

```
| eval verhalten=if(percentageDifftag > upperBound OR      (5.3)
percentageDifftag < lowerBound, “Anomalie”, “Normal”)
```

So handelt es sich bei dem Datensatz um eine Anomalie, sobald die prozentuale Differenz zum Count des Vortags die obere Grenze (45.0) oder die untere Grenze (-45.0) über- bzw. unterschreitet.

Anschließend ist der Datensatz mit weiteren Attributen (Spalten) angereichert worden, um den Algorithmen der Klassifikation mehr Input (Informationen) zu geben und ein besseres Ergebnis hinsichtlich des Lernvorgangs von den Eingabedaten auf die Ausgabedaten, sowie der Vorhersagegenauigkeit zu erreichen.

---

[209] Das eval Kommando dient zur Erzeugung eines neuen Feldes und zur Berechnung eines Ausdrucks. Der Ergebniswert wird anschließend dem neu erzeugten Feld zugewiesen. Vgl. Splunk Inc., Splunk Schnellreferenz

Zuerst sind Attribute erzeugt worden, die eine Differenz vom aktuellen Count zum vorherigen Count liefern. Dabei wurden die Differenzen von einem aktuell auftretenden Count zu den Counts der letzten Stunde, des letzten Tages, sowie der letzten Woche betrachtet. Um dies in Splunk umzusetzen, wurde das Kommando „delta“[210] der Splunk Search Processing Language verwendet. Die Folgenden Attribute wurden erzeugt:

| delta count p=1 as diff_1stunde (5.4)

| delta count p=24 as diff_1tag (5.5)

| delta count p=168 as diff_1woche (5.6)

Anschließend wurden Attribute, die den Count der letzten Stunde, den Count des letzten Tages und den Count der letzten Woche repräsentieren, erzeugt. Hierzu konnte ebenfalls das Splunk-Kommando „eval“ genutzt werden:

| eval countletztestunde= count – diff_1stunde (5.7)

| eval countletztertag= count – diff_1tag (5.8)

| eval countletztewoche= count – diff_1woche (5.9)

Zudem wurden die zuvor ermittelten Differenzen „diff_1stunde“, „diff_1tag“ und „diff_1woche“ in Prozent umgerechnet und in einem entsprechenden Feld gespeichert.

| eval percentageDiffstunde= 100/count*diff_1stunde (5.10)

| eval percentageDifftag= 100/count*diff_1tag (5.11)

| eval percentageDiffwoche= 100/count*diff_1woche (5.12)

Die vollständige Suche zur Anreicherung des Datensatzes ist im Folgenden aufgeführt.

| tstats prestats=t count where index="*" OR index"_*" by _time span=1h (5.13)
| timechart count span=1h
| delta count p=1 as diff_1stunde
| eval countletzterstunde=count-diff_1stunde
| eval percentageDiffstunde=100/count*diff_1stunde

[210] Das delta Kommando berechnet für jedes Ereignis in dem Count vorkommt, die Differenz zwischen Count und seinem vorherigen Count (Wert) und speichert diesen im Ergebnis nach Angabe des Kommandos „as“ in einem neuen Feld ab. Der Parameter „p“ dient zur Angabe, wie viele vorherige Werte betrachtet werden sollen. Da der Count in einem stündlichen Intervall erzeugt wird, bezeichnet p=1 den Count der letzten Stunde. Vgl. Splunk Inc., Splunk Schnellreferenz

```
| delta count p=24 as diff_1tag
| eval countletztertag=count-diff_1tag
| eval percentageDifftag=100/count*diff_1tag
| delta count p=168 as diff_1woche
| eval countletztewoche=count-diff_1woche
| eval percentageDiffwoche=100/count*diff_1woche
| eval upperBound=45.0, lowerBound=-45.0
| eval verhalten=if(percentageDifftag > upperBound OR percentageDifftag < lowerBound, "Anomalie", "Normal")
```

Einen Ausschnitt des Ergebnisses nach der Datenanreicherung stellt die Tabelle 5 dar. Das vollständige Ergebnis befindet sich auf dem anliegenden Datenträger unter der Bezeichnung „ba_count_letzte_35_tage_alle_indizes_inklusive_datenanreicherung_als_tabelle.csv“.

_time	count	countletzterstunde	countletztertag	countletztewoche
2016-04-26T00:00:00.000+0200	6378374			
2016-04-26T01:00:00.000+0200	4421592	6378374		
2016-04-26T02:00:00.000+0200	3237106	4421592		

Tab. 5 Ausschnitt des Ergebnisses nach der Datenanreicherung

Nun ist anhand der Tabelle 5 zu erkennen, dass die Spalten „countletzterstunde“, „countletztertag“, „countletztewoche“ fehlende Werte aufweisen. Aus diesem Grund ist an dieser Stelle auf das Verfahren der Filterung des Data Cleansing zurückzugreifen und dieses anzuwenden. Dabei wurde das Splunk-Kommando „search“[211] verwendet.

```
| tstats prestats=t count where index="*" OR index"_*" by _time span=1h          (5.14)
| timechart count span=1h
| delta count p=1 as diff_1stunde
| eval countletzterstunde=count-diff_1stunde
| eval percentageDiffstunde=100/count*diff_1stunde
| delta count p=24 as diff_1tag
| eval countletztertag=count-diff_1tag
| eval percentageDifftag=100/count*diff_1tag
| delta count p=168 as diff_1woche
```

[211] Das search Kommando dient zur Filterung von Ergebnissen, die auf den Suchausdruck passen. Vgl. Splunk Inc., Splunk Schnellreferenz

```
| eval countletztewoche=count-diff_1woche
| eval percentageDiffwoche=100/count*diff_1woche
| eval upperBound=45.0, lowerBound=-45.0
| eval verhalten=if(percentageDifftag > upperBound OR percentageDifftag < lowerBound, "Anomalie", "Normal")
| search diff_1woche=*
```

Durch die Filterung des Ergebnisses mit Hilfe des Kommandos „| search diff_1woche=*“ sind zwar die Datensätze vom 26.04.2016 bis 02.05.2016 aus dem Datensatz für die Analyse ausgeschlossen worden, allerdings ist hierdurch ein vollständiger Datensatz ohne fehlende Werte entstanden. Die herausgefilterten unvollständigen Datensätze werden nicht weiter verarbeitet. Einen Ausschnitt des Ergebnisses der Filterung zeigt die Tabelle 6.

_time	count	countletzterstunde	countletztertag	countletztewoche
2016-05-03T00:00:00.000+0200	8732086	8864780	9280102	6378374
2016-05-03T01:00:00.000+0200	7239469	8732086	5732160	4421592
2016-05-03T02:00:00.000+0200	5452099	7239469	10096731	3237106
2016-05-03T03:00:00.000+0200	5970306	5452099	5827948	3694523

Tab. 6 Ergebnis nach Anwendung des Data Cleansing

Das vollständige Ergebnis nach Anwendung des Data Cleansing befindet sich auf dem anliegenden Datenträger unter der Bezeichnung „ba_count_letzte_35_tage_alle_indizes_inklusive_datenanreicherung_und_cleansing.csv“.

In der folgenden Phase (Transformation) wird für eine geeignete Repräsentation der Daten, der Datenbestand in die vom gewählten Analyseverfahren benötigte Form gebracht.

5.2.4 Datentransformation

In der Phase der Datentransformation werden die vorverarbeiteten und bereinigten Daten in eine für das Ziel des KDD-Prozesses geeignete Repräsentation (Darstellungsform und Format) transformiert. Beispielhafte Schritte der Transformation sind die Diskretisierung von Attributen und die Selektion von Attributen. [212] Ein weiterer Schritt der Transformation, der an dieser Stelle hinzuzufügen ist, betrifft die Aufbereitung

[212] Vgl. Ester, M. (2013), S. 3

der Rohevents, die im Kapitel 5.2.2 durch die aufgeführte Suche und unter Verwendung des „timechart“-Kommandos erzeugt wurden und dadurch über einen Zeitverlauf betrachtet werden können. Dabei ist durch Verwendung der zusätzlichen Option „span“[213] und der Parametereinstellung „span=1h“, die beim „timechart“-Kommando genutzt werden kann, die Anzahl der Events eines Tages auf ein stündliches Intervall festgelegt worden.

Die Diskretisierung von Attributen ist zu beachten und ggf. anzuwenden, falls ein bestimmter Data Mining-Algorithmus verwendet werden soll, der nur kategorische Attribute oder nur numerische Attribute verarbeiten kann. Folglich ist entweder die Diskretisierung numerischer Attributwerte oder die Transformation kategorischer Attributwerte in numerische Attributwerte durchzuführen.[214]

Die Selektion von Attributen wird vorgenommen, da eine zu große Anzahl von Attributen die Effizienz von Methoden des Data Minings, sowie die Qualität des Ergebnisses deutlich verschlechtern können.[215] Um eine entsprechende Auswahl für die Vorhersage des Feldes „verhalten“ durchzuführen, ist der Splunk Befehl „analyzefields“, der bereits im Experteninterview angesprochen wurde, verwendet worden. Dieser ermöglicht es, geeignete Felder für die Vorhersage zu identifizieren.[216] Dabei werden für alle aufgeführten Felder berechnet, wie gut diese zur Vorhersage der Werte des in „classfield“ angegeben Feldes geeignet sind.[217] Der Befehl „analyzefields“ ist auf dem zuvor in der CSV-Datei gespeicherten Datensatz „ba_count_letzte_35_tage_ alle_indizes_inklusive_datenanreicherung_und_cleansing.csv“ angewendet worden. Um den Datensatz in Splunk einzulesen wurde das Kommando „inputlookup“ verwendet.[218] Im Folgenden ist die vollständige Suche aufgeführt.

```
| inputlookup ba_count_letzte_35_tage_alle_indizes_inklusive_ (5.15)
datenanreicherung_und_cleansing.csv
| analyzefields classfield=verhalten
```

[213] Span ist eine zusätzliche Option, die beim timechart-kommando angegeben werden kann, um einen Zeitbereich fest zu definieren. Vgl. Splunk Inc., Splunk® Enterprise – Search Reference - timechart
[214] Vgl. Ester, M. (2013), S. 4
[215] Vgl. Ester, M. (2013), S. 4
[216] Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Frage/Antwort 4.1
[217] Vgl. Splunk Inc., Splunk® Enterprise – Search Reference - analyzefields
[218] Vgl. Splunk Inc., Splunk Schnellreferenz

Tabelle 7 zeigt das Ergebnis nach Anwendung des Befehls „analyzefields“. Das Ergebnis ist ebenfalls als CSV-Datei auf dem Datenträger unter der Bezeichnung „analyzefields.csv“ abgelegt.

field	count	cocur	acc	balacc
count	504	1.000.000	0.714286	0.628205
countletzterstunde	504	1.000.000	0.714286	0.615385
countletztertag	504	1.000.000	0.658730	0.662393
countletztewoche	504	1.000.000	0.551587	0.553419
diff_1stunde	504	1.000.000	0.750000	0.506410
diff_1tag	504	1.000.000	0.908730	0.925214
diff_1woche	504	1.000.000	0.742063	0.617521
lowerBound	504	1.000.000	0.071429	0.500000
percentageDiffstunde	504	1.000.000	0.579365	0.504274
percentageDifftag	504	1.000.000	0.956349	0.976496
percentageDiffwoche	504	1.000.000	0.720238	0.605769
upperBound	504	1.000.000	0.071429	0.500000

Tab. 7 Identifikation geeigneter Attribute für die Vorhersage

Die Spalte „field“ führt die im Datensatz vorhandenen Attribute auf. Die Spalte „count“ stellt die Größe des Datensatzes dar. Bei der Spalte „cocur“ erfolgt die Angabe des Auftretens des jeweiligen Feldes im Verhältnis zum Feld „verhalten“. Ein Wert von 1 bedeutet, dass das jeweilige Feld in jedem Event auftritt, indem auch das Feld „verhalten“ vorkommt. Die Spalte „acc“ (accuracy) gibt die Genauigkeit zur Vorhersage des Feldes „verhalten“ an. „balacc“ ist der Durchschnitt (arithmetisches Mittel) der Genauigkeiten (accuracy) dieser Vorhersagen.[219]

Für die weitere Analyse ist die Auswahl zur Vorhersage der Spalte „verhalten“ auf die Attribute „count“ und „percentageDifftag“ gefallen, da das Attribut „.percentageDifftag“ eine hohe Genauigkeit (accuracy) aufweist. Das Feld „count“ ist zu berücksichtigen, da dieses ein wesentliches Kriterium bei der Untersuchung zur Erkennung von Anomalien darstellt.

[219] Vgl. Splunk Inc., Splunk® Enterprise – Search Reference - analyzefields

Nachdem die Datenselektion, -vorverarbeitung, -bereinigung und -transformation durchgeführt wurden, folgt die Auswahl einer Data Mining-Methode zur Erkennung von Anomalien.

5.2.5 Auswahl einer Data Mining-Methode zur Anomalie-Erkennung

In diesem Abschnitt wird für das festgelegte Ziel eine geeignete Data Mining-Methode (z. B. Assoziation, Clustering, Regression oder Klassifikation, siehe Kapitel 3.5.2.3) ausgewählt. Anschließend wird ein passender Algorithmus selektiert. Wie bereits in Kapitel 5.2.3 erwähnt, ist die Auswahl der Data Mining-Methode zur Erkennung von Anomalien auf die Klassifikation, die zu den Methoden des überwachten Lernens gehört, gefallen.

Die Klassifikation wurde ausgewählt, da diese Methode dazu dient, anhand einer vorgegebenen Datenmenge ein Modell (Algorithmus) aufzubauen, mit dem sich anschließend unbekannte Daten aufgrund ihrer Eigenschaften (Attributwerte) in Klassen einteilen lassen. Bei der vorgegebenen Datenmenge sind sowohl die Klasse (verhalten) als auch die Zuordnung der Attribute zu der jeweiligen Klasse vorher bekannt. Die Zuordnung der jeweiligen Klasse ist bereits im Kapitel 5.2.3 durch Anwendung einer Klassifikationsregel beschrieben worden. Ein weiteres Argument, das an dieser Stelle für die Klassifikation spricht, betrifft den Typ der vorliegenden Daten, da es sich bei dem Feld „verhalten“ um kategoriale Daten handelt.

Bei Anwendung eines Klassifikationsverfahrens ist die vorgegebene Datenmenge (enthält die Zuordnung der Klassen) in einen Trainingsdatensatz und einen Testdatensatz zu teilen. In der Praxis werden 2/3 der Daten als Trainingsdaten und 1/3 der Daten als Testdaten verwendet.[220] Die Zugehörigkeiten der Datensätze zu der jeweiligen Partition (Trainingsdatensatz oder Testdatensatz) werden zufällig bestimmt, sodass bei mehreren Ausführungen ein ähnlicher Effekt wie beim Einsatz der Kreuzvalidierung erreicht wird. Eine Kreuzvalidierung konnte aus technischen Gründen nicht eingesetzt werden, da zum Zeitpunkt der Arbeit die Machine-Learning App von Splunk diese nicht unterstützt.

Entgegen der Empfehlung aus der Praxis wurde die Trainingspartition auf 30 % und die Testpartition auf 70 % festgelegt, um in der Analyse einer Überanpassung (engl. Overfitting) des Modells entgegenzuwirken. Es kann beim Bilden eines Vorhersagemodells die Gefahr entstehen, dass es sich zu gut an die Daten anpasst mit

[220] Vgl. Han, J., Kamber, M., Pei, J. (2012), S. 370

denen es berechnet worden ist, sodass es für die Vorhersagen auf neue Daten ungeeignet ist. Dies bezeichnet eine Überanpassung. [221]

Auf das Ergebnis nach Anwendung des Data Cleansing sind für das Splitting der Daten in Trainings- und Testdaten die folgenden Kommandos innerhalb der Splunk Machine-Learning App angewendet worden.

```
| eval random_key=random()                                                    (5.16)
| eval split_key=if(random_key / 2147483647 > 0.3, “test”, “train”)
```

Hierbei wird der Variablen “random_key” die Funktion “random()” zugewiesen. Diese erzeugt einen zufälligen Integer-Wert zwischen 0 und 2147483647.[222] Anschließend wird der Variablen „split_key“ entweder der Wert „test“ oder „train“ zugewiesen. Dabei erfolgt ein Splitting der Daten in 30 % Trainingsdaten und 70 % Testdaten. Da die Kreuzvalidierung nicht unterstützt wird, ist das Verfahren des Splittings fünf Mal durchgeführt und die Ergebnisse anschließend jeweils in einer CSV-Datei gespeichert und exportiert worden.

Die Bezeichnungen der CSV-Dateien sind im Folgenden aufgeführt und bilden die Grundlage für die weitere Analyse.

- „count_letzte_35_tage_alle_indizes_inklusive_datenanreicherung_und_cleansing_30training_70test.csv“
- „count_letzte_35_tage_alle_indizes_inklusive_datenanreicherung_und_cleansing_final_30_Training_70_Test_2.csv“
- „count_letzte_35_tage_alle_indizes_inklusive_datenanreicherung_und_cleansing_final_30_Training_70_Test_3.csv“
- „count_letzte_35_tage_alle_indizes_inklusive_datenanreicherung_und_cleansing_final_30_Training_70_Test_4.csv“
- „count_letzte_35_tage_alle_indizes_inklusive_datenanreicherung_und_cleansing_final_30_Training_70_Test_5.csv“

[221] Vgl. Janssen, J., Laatz, W. (2013), S. 437
[222] Vgl. Splunk Inc., Splunk Schnellreferenz

5.2.6 Algorithmen der Klassifikation

Nach Auswahl der Data Mining-Methode ist ein geeigneter Klassifikationsalgorithmus auszuwählen. Dieser Algorithmus (Klassifikator) dient zum Lernen der Zuordnung anhand der bereits klassifizierten Trainingsdaten.

Im Rahmen dieser Arbeit werden die drei folgenden Algorithmen der Klassifikation näher betrachtet.

- Logistische Regression (LR)
- Support Vector Machine (SVM)
- RandomForestClassifier (RFC)

Logistische Regression (LR)

Die Logistische Regression, auch bezeichnet als Logit-Modell, ist ein Verfahren der Klassifikation, mit der anhand unabhängiger Variablen (Prädiktoren) eine Vorhersage auf eine kategoriale abhängige Variable (Zielvariable) getroffen werden kann.[223] Dabei wird bei der Logistischen Regression zwischen der binären (dichotomen) logistischen Regression und der multinominalen logistischen Regression unterschieden. So werden bei der binären logistischen Regression nur zwei Ausprägungen (z.B. 0 oder 1 bzw. Normal oder Anomalie) betrachtet, wohingegen die multinominale logistische Regression drei oder mehrere Ausprägungen verwendet.[224]

Die logistische Regression erweitert den Ansatz der linearen Regression. Bei der linearen Regression ist die Vorhersage einer kategorialen Zielvariable nicht möglich. Grund hierfür ist, dass bei dem Versuch ein Regressionsmodell mit Hilfe einer linearen Regression und vorliegenden unabhängigen, metrisch skalierten (stetigen) Variablen, die geschätzten Werte der Zielvariable mehr als zwei diskrete Werte (0 und 1) annehmen würden.[225]

Dieses Problem wird bei der logistischen Regression durch eine Transformation der Zielvariablen zu einem Logit gelöst.

So führt Lohninger auf, dass „in solchen Fällen … man versuchen [kann], statt der Zielvariablen die Chance (engl. Odds), dass eine bestimmte Ausprägung der Zielvariable

---

[223] Vgl. Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (2016), S. 284
[224] Vgl. Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (2016), S. 284, In dieser Arbeit liegt der Fokus auf der binären logistischen Regression, da nur die zwei Ausprägungen Normal und Anomalie bei der Zielvariable relevant sind.
[225] Vgl. Lohninger, H. (2012), Logistische Regression

eintritt, zu schätzen.“[226] Die Transformation der Zielvariablen zu einem Logit wird im Folgenden in Anlehnung an Lohninger beschrieben.

Angenommen, die Zielvariable y ist eine dichotome Variable mit den Ausprägungen 0 und 1. p ist die Wahrscheinlichkeit, dass y eintritt (y = 1). So ergibt sich, dass (1 - p) die Wahrscheinlichkeit ist, dass y nicht eintritt (y = 0). Dann ist $\frac{p}{(1-p)}$ die Chance, dass y eintritt.[227]

An dieser Stelle kommt die Funktion, die als Logit bezeichnet wird zum Einsatz. Der Logit ist der Logarithmus (ln) der Chance und linear abhängig von x. Dabei wird bei der Logit-Funktion der Logarithmus auf die Chance angewendet und ähnlich einer linearen Regression ausgedrückt.[228]

$$\text{logit}(y)=\ln\left[\frac{p}{1-p}\right] \tag{5.17}$$

Die Wahrscheinlichkeit des Auftretens einer bestimmten Ausprägung wird durch die Logit-Funktion so transformiert, dass sich ein Wert von minus unendlich für das sichere Auftreten eines Zustands und ein Wert von plus unendlich für das sichere Auftreten des anderen Zustands ergibt. Sind die Chancen gleich, ergibt sich ein Wert von null.[229]

Support Vector Machine (SVM)

Die Stützvektormaschine (engl. Support Vector Machine) ist ein Klassifikator und gehört ebenfalls zu den Methoden des überwachten Lernens. Die Grundlage für eine SVM bildet eine Menge von Trainingsobjekten. Für diese Menge ist jeweils bekannt, welcher Klasse sie zugehören. Das Ziel der SVM besteht nun darin, aus der Klassenzuordnung eine Klassifikationsregel abzuleiten, sodass auf einem neuen Datensatz mit einer hohen Genauigkeit ebenfalls die Klassifizierung erfolgen kann.[230] Dabei verfolgt die SVM den Ansatz, lineare und nichtlineare Modelle zu vereinen. Dies wird anhand eines Zweiklassenproblems im Folgenden grob skizziert.[231]

Liegen zwei linear seperable Klassen vor, ist es einfach eine Hyperebene zu finden, die als Trennfläche dient und die Trainingsobjekte in zwei Klassen teilt. Unter Linearer Separierbarkeit ist in einem zweidimensionalen Raum eine Trennung von zwei Klassen

---

[226] Lohninger, H. (2012), Logistische Regression
[227] Vgl. Lohninger, H. (2012), Logistische Regression
[228] Vgl. ebd.
[229] Vgl. Lohninger, H. (2012), Logistische Regression
[230] Vgl. Görz, G., Schneeberger, J., Schmid, U. (2014), S. 427 f.
[231] Vgl. Ertel, W. (2013), S. 281 f.

durch eine Gerade (Hyperebene) zu verstehen.[232] Allerdings existieren meist sehr viele solcher Ebenen, sodass eine Ebene gesucht wird, die einen möglichst großen minimalen Abstand zu beiden Klassen hat. Diese Hyperebene wird durch vereinzelte Punkte, die sich im Grenzbereich, also am nächsten an der Trennebene befinden, definiert. Diese Punkte, bezeichnet als Support-Vektoren (Stützvektoren) verleihen der Support Vector Machine auch ihren Namen. Die Support-Vektoren haben alle den gleichen Abstand zur Hyperebene. In Abbildung 16 ist dieses Zweiklassenproblem mit einer linear trennenden Hyperebene dargestellt. Bei den umrandeten Punkten handelt es sich um die Support-Vektoren.[233]

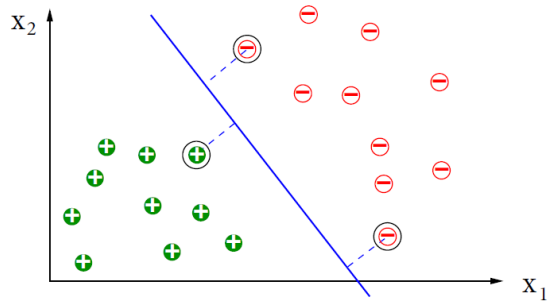


Abb. 16 Linear trennende Hyperebene der SVM bei einem Zweiklassenproblem
Quelle: Ertel, W. (2013), S. 282

Dieses Verfahren wird nun auch versucht mit Hilfe der Support Vector Machine auf nichtlinear seperable Probleme anzuwenden. Abbildung 17 zeigt das Problem bei nichtlinearer Trennung der Hyperebene. Hier kann nicht ohne weiteres Vorgehen eine Hyperebene definiert werden, da die Instanzen nicht linear trennbar sind.[234]

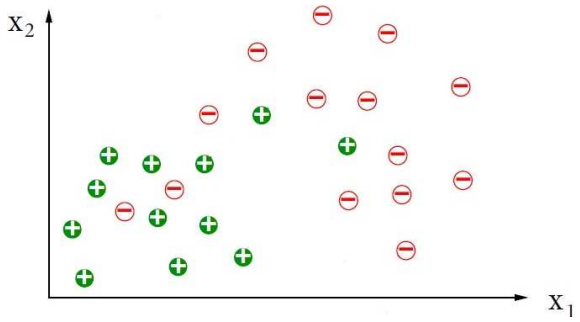


Abb. 17 Nicht linear trennbare Instanzen
Quelle: Eigene Darstellung in Anlehnung an Ertel, W. (2013), S. 282

So ist ein zweistufiges Vorgehen durchzuführen: Zuerst wird eine nichtlineare Transformation auf die Daten angewendet, um die transformierten Daten linear separabel zu gestalten. Anschließend werden die Support-Vektoren in dem transformierten Raum bestimmt. Hinsichtlich des ersten Schrittes ist zu erwähnen, dass

[232] Vgl. Cleve, J., Lämmel, U. (2014), S. 131
[233] Vgl. ebd.
[234] Vgl. Ertel, W. (2013), S. 281 f.

bei Daten, die widerspruchsfrei sind, durch Transformation des Vektorraums immer die Möglichkeit besteht die Klassen linear separabel zu machen. Um einen widersprüchlichen Datenpunkt handelt es sich, wenn dieser zu beiden Klassen gehört. Die Trennung wird z.B. erreicht durch die Einführung einer neuen (n+1)-ten Dimension und der folgenden Festlegung:

$$x_{n+1} \begin{cases} 1 \; falls \; x \; \in Klasse \; 1 \\ 0 \; falls \; x \; \in Klasse \; 0 \end{cases} \tag{5.18}$$

Da diese Formel nur auf neue zu klassifizierende Punkte mit unbekannter Klasse hilfreich anzuwenden ist, wird eine allgemeine Transformation benötigt, die unabhängig von den aktuellen Daten sein sollte. Allerdings kann durch einen sogenannten Kernel-Trick gezeigt werden, dass für beliebig geformte Klassentrennlinien im ursprünglichen Vektorraum eine solche generische Transformation existiert, sodass die Daten dann linear separabel im transformierten Raum vorliegen. Dabei wächst die Zahl der Dimensionen des neuen Vektorraums exponentiell mit der Zahl der Dimensionen des ursprünglichen Vektorraums, aber die höhere Anzahl der Dimensionen ist bei Verwendung der Support-Vektoren nicht problematisch, da die trennende Ebene nur durch wenige Parameter bestimmt wird. Bei dieser zentralen nichtlinearen Transformation des Vektorraums ist die Rede vom Kernel. Aus diesem Grund sind Support-Vektor-Maschinen auch unter dem Namen Kernel-Methoden bekannt.

Auf die mathematische Herleitung wurde an dieser Stelle verzichtet, sodass für einen tieferen Einblick auf weiterführende Literatur verwiesen wird.[235]

RandomForestClassifier (RFC)

Der RandomForestClassifier ist ein Klassifikationsverfahren des überwachten Lernens. Dabei handelt es sich bei diesem Algorithmus um eine Ensemble-Methode, bei der die Idee darin besteht mehrere Modelle zu kombinieren, um eine höhere Vorhersagegenauigkeit zu erzielen. Das Ensemble des Random Forest setzt sich dabei aus mehreren Entscheidungsbaum-Modellen (Wald bzw. Forest) zusammen.[236] Hierbei greift der Random Forest auf zwei Zufallsmechanismen (*Bagging* und *Random Feature*

[235] S. Schölkopf und A. Smola. Learning with Kernels: Support VectorMachines, Regularization, Optimization, and Beyond. MIT Press (2002), E. Alpaydin. Introduction to Machine Learning.MIT Press (2004), C. J. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. DataMin. Knowl. Discov. 2 (1998) 2, 121–167.
[236] Vgl. Breiman, L. (2001), S. 5 f.

Selection) zurück, da die Zufälligkeit zur Erhöhung der Vielfältigkeit der einzelnen Entscheidungsbäume beiträgt.

Bagging steht für „Bootstrap Aggregating“ und ist ein Algorithmus, der eine große Anzahl von Entscheidungsbäumen generiert. Die Vorhersagen der generierten Entscheidungsbäume werden per Mehrheitsentscheidung kombiniert. Jeder Baum darf für die Klassifikation in diesem Wald eine Entscheidung treffen. Die Klasse mit den meisten Stimmen (auch Votes genannt) entscheidet die endgültige Klassifikation bzw. legt die vorhergesagte Klasse fest. Dabei wird jeder einzelne Entscheidungsbaum auf einer Bootstrap-Stichprobe der Trainingsmenge gelernt. Erzeugt werden die Stichproben, indem *n* Datensätze aus der ursprünglichen Trainingsmenge *E* mit Zurücklegen gezogen werden. So enthält jede Bootstrap-Stichprobe die gleiche Anzahl von Datensätzen wie *E* und ein bestimmter Trainingsdatensatz kann mehrfach oder gar nicht enthalten sein. Das bedeutet, dass eine zufällige Anzahl von Daten aus dem Trainingsdatensatz genommen wird. Durch den Einsatz und der Kombination vieler Entscheidungsbäume wird die durch den Bagging-Algorithmus gelernte Klassifikationsregel stabiler und weißt geringere Vorhersagefehler auf. [237]

Der zweite Zufallsmechanismus der beim Random Forest verwendet wird, ist *Random Feature Selection (*auch *Random Attribute Selection* genannt*)*. Dieser dient dazu das Klassifikationsmodell variabler zu gestalten. Dabei werden durch diesen Ansatz nicht alle Attribute |A| auf jeder Knotenebene zur Festlegung des Splits herangezogen, sondern nur eine Teilmenge. Diese Teilmenge definiert Breiman mit dem $\log_2|A| + 1$. So ergibt sich beispielweise bei 20 Attributen $\log_2 20 + 1 = 5{,}321$. Das bedeutet, dass auf jeder Knotenebene fünf Attribute aus den noch vorhandenen Attributen ausgewählt werden.[238]

Beim Random Forest werden die Bäume nicht zurückgeschnitten (engl. Pruning). Dadurch werden die Bäume auf Basis der vorliegenden Trainingsdatenmenge vollständig ausgelernt.[239]

5.2.7 Anwendung von Algorithmen der Klassifikation

In diesem Abschnitt wird beschrieben, wie die zuvor erläuterten Algorithmen (Logistische Regression, Support Vector Machine und RandomForestClassifier) angewendet

[237] Vgl. Görz, G., Schneeberger, J., Schmid, U. (2014), S. 419
[238] Vgl. Breiman, L. (2001), S. 12
[239] Vgl. Han, J., Kamber, M., Pei, J. (2012), S. 383

werden. Dabei werden die drei Algorithmen zuerst auf dem Trainingsdatensatz trainiert und anschließend auf dem Testdatensatz angewendet. Daraufhin folgt die Evaluierung der Modelle anhand der Konfusionsmatrix und der Bewertungsmaße, die im Kapitel 5.2.8 aufgeführt und beschrieben werden.

Zum Trainieren eines Algorithmus (Modells) wird innerhalb der App das Kommando „| fit“ verwendet. Das trainierte Modell ist zudem unter einer Bezeichnung abzuspeichern. Dies wird über das Schlüsselwort „into“ erreicht, dem anschließend eine entsprechende Bezeichnung für das Modell folgt. Die Anwendung des trainierten Modells erfolgt mit dem Kommando „| apply“.[240]

Weiterhin ist in der Machine-Learning App festzulegen, welche Merkmale (Selektion der Attribute) zum Trainieren verwendet werden und welches Attribut (Klasse) hinsichtlich der Vorhersage zu berücksichtigen ist. Dabei ist zuerst das Feld der Vorhersage zu definieren. Darauf folgt das Schlüsselwort „from“ und die ausgewählten Merkmale.

Um eine Vergleichbarkeit für die Evaluierung der Algorithmen herzustellen, wurden alle drei Algorithmen anhand der Merkmale „count“ und „percentageDifftag“ trainiert. Die unter (5.19) aufgeführten Befehlszeilen stellen das Training der Logistischen Regression dar. Diese Vorgehensweise ist ebenfalls für die beiden anderen Algorithmen (Support Vector Machine und RandomForestClassifier) durchgeführt worden. Die Abbildungen des Trainings, sowie der Anwendung auf die Testdaten der Algorithmen Logistische Regression, Support Vector Machine und RandomForestClassifier befinden sich in den Anhängen 4, 5 und 6. Auf dem anliegenden Datenträger befinden sich zusätzlich Abbildungen, die die Anwendung der drei Modelle auf die jeweils weiteren vier Datensätze darstellen.

```
| inputlookup count_letzte_35_tage_alle_indizes_inklusive_datenanreicherung (5.19)
_und_cleansing_30training_70test.csv
| search split_key=train
| fit LogisticRegression into “BA_Modell_Logistic_Regression_Split_30_Train_70_Test”
“verhalten” from count, percentageDifftag
```

[240] Vgl. Splunk Inc., Splunk® Enterprise – Machine Learning Toolkit and Showcase App – User Guide – Custom search commands

Das Modell wurde unter dem Namen „BA_Modell_Logistic_Regression_ _Split_30_Train_70_Test“ abgespeichert und auf die Testdaten angewendet (siehe (5.20)).

```
| inputlookup count_letzte_35_tage_alle_indizes_inklusive_datenanreicherung (5.20)
_und_cleansing_30training_70test.csv
| search split_key=test
| apply “BA_Modell_Logistic_Regression_Split_30_Train_70_Test”
```

Zudem wurde die Konfusionsmatrix mit Hilfe des Kommandos „| `confusionmatrix("verhalten","predicted(verhalten)")`“ für das angewandte Modell auf die Testdaten angewendet (siehe (5.21)).

```
| inputlookup count_letzte_35_tage_alle_indizes_inklusive_datenanreicherung (5.21)
_und_cleansing_30training_70test.csv
| search split_key=test
| apply “BA_Modell_Logistic_Regression_Split_30_Train_70_Test”
| `confusionmatrix("verhalten","predicted(verhalten)")`
```

Die folgende Tabelle stellt die erzeugte Konfusionsmatrix nach Anwendung des Modells auf die Testdaten dar.

Predicted actual	**Predicted Anomalie**	**Predicted Normal**
Anomalie	23	0
Normal	16	297

Tab. 8 Konfusionsmatrix des angewendeten Modells auf die Testdaten

Die Konfusionsmatrix des angewendeten Modells (Logistische Regression) auf die Trainingsdaten, sowie die Anwendung der trainierten Modelle (Support Vector Machine und RandomForestClassifier) auf die Trainings- bzw. Testdaten und deren dazugehörigen Konfusionsmatrizen sind im Anhang 4, 5 und 6 aufgeführt. Die Konfusionsmatrizen der angewendeten Modelle auf die jeweils vier weiteren Datensätze sind auf dem anliegenden Datenträger zu finden.

5.2.8 Durchführung der Analyse und Auswertung der erzielten Ergebnisse

Nachdem die trainierten Data Mining-Verfahren (Modelle) auf die Testdaten (Testsplit) angewendet wurden, ist die Evaluierung der prognostizierten Annotation (Klassenzuordnung) mit der tatsächlichen Annotation erforderlich. Da im Rahmen dieser Arbeit ausschließlich Verfahren der Klassifikation verwendet wurden, werden folglich

auch lediglich Bewertungsmaße für die verwendeten Verfahren vorgestellt und angewendet.

Zur Ermittlung der Klassifikationsgüte bzw. der Vorhersagegenauigkeit eines Modells wird die Konfusionsmatrix verwendet. Diese beschreibt grundsätzlich, wie gut ein Modell die einzelnen Klassen prognostiziert. Bei zwei Klassifikationsmöglichkeiten (Anomalie oder Normal) besteht die Konfusionsmatrix aus einer 2x2 Matrix.[241] Den Aufbau der Konfusionsmatrix zeigt die Tabelle 9.

		Vorhersage des Klassifikators	
		Anomalie	Normal
Tatsächliche Klasse	Anomalie	(TP) true positive	(FN) false negative
	Normal	(FP) false positive	(TN) true negative

Tab. 9 Konfusionsmatrix zur Evaluierung der Klassifikationsgüte
Quelle: Eigene Darstellung in Anlehnung an Witten, I., Frank, E., Hall, M. (2011), S. 164

Das Vorhersageergebnis des Klassifikators ist anschließend mit der tatsächlichen Klasse, die als Wert im zugehörigen Feld der Testdaten enthalten ist, zu vergleichen. Dabei wird gezählt, wie häufig jede der vier möglichen Kombinationen von „Vorhersage des Klassifikators“ und „Tatsächliche Klasse“ vorgekommen ist. Diese Häufigkeiten werden in der Konfusionsmatrix eingetragen, sodass sich vier Möglichkeiten für die Klassifikationen ergeben. Die vom Klassifikator korrekt erzeugten Ergebnisse stehen in der Tabelle in der Hauptdiagonale (true positive und true negative) und die falsch vorhergesagten Ergebnisse in der Nebendiagonale (false positive und false negative).[242] Im Folgenden werden die einzelnen Felder der Konfusionsmatrix erläutert.

(TP) true positive: Das Auftreten des „Counts“ wurde als Anomalie klassifiziert und die Vorhersage des Klassifikators ist korrekt.

(FN) false negative: Das Auftreten des „Counts“ wurde als Anomalie klassifiziert, aber die Vorhersage des Klassifikators ist falsch.

(FP) false positive: Das Auftreten des „Counts“ wurde als Normal klassifiziert, aber die Vorhersage des Klassifikators ist falsch.

(TN) true negative: Das Auftreten des „Counts“ wurde als Normal klassifiziert und die Vorhersage des Klassifikators ist korrekt.

---

[241] Vgl. Müller, R., Lenz, H. (2013), S.97
[242] Vgl. Müller, R., Lenz, H. (2013), S.98

Aus der Konfusionsmatrix werden anschließend die Bewertungsmaße gebildet. Die folgenden Bewertungsmaße werden zur Evaluation der Klassifikationsgüte eines Modells betrachtet.[243]

Accuracy (Klassifikationsgenauigkeit) gibt an, wie viel Prozent der Objekte in der Testmenge korrekt der richtigen Klasse zugeordnet wurden.

$$\textbf{Accuracy: } \frac{TP+TN}{TP+TN+FP+FN} \tag{5.22}$$

Analog zur Klassifikationsgenauigkeit gibt der tatsächliche **Klassifikationsfehler** (auch Falschklassifikationsrate genannt) an, wie viel Prozent der Datenobjekte in der Testmenge einer falschen Klasse zugeordnet wurden.

$$\textbf{Klassifikationsfehler: } \frac{FP+FN}{TP+TN+FP+FN} \tag{5.23}$$

Die Klassifikationsgenauigkeit- und der Klassifikationsfehler addieren sich entsprechend zu 1 oder 100 %.

Precision (Präzision) gibt den Anteil der richtig positiv klassifizierten Objekte (TP) an der Gesamtanzahl der positiv klassifizierten Objekten (TP+FP) an.

$$\textbf{Precision: } \frac{TP}{TP+FP} \tag{5.24}$$

Recall (Sensitivität bzw. Trefferquote) gibt den Anteil der richtig als positiv klassifizierten Objekte (TP) an der Gesamtheit der tatsächlichen positiven Objekte an.

$$\textbf{Recall: } \frac{TP}{TP+FN} \tag{5.25}$$

F-Measure (F-Maß) oder auch als F_1 bezeichnet, dient dazu Klassifikatoren anhand einer einzigen Kennzahl zu vergleichen bzw. zu beurteilen. Das F-Maß kombiniert die Präzision (Precision) und die Trefferquote (Recall) durch das harmonische Mittel. Die true negatives (TN) werden hingegen nicht berücksichtigt, sodass der Fokus maßgeblich auf die Prognosegüte der positiven, also der hier kleineren Klasse (Anomalie) gerichtet wird und die Betrachtung der größeren Klasse (Normal) eher vernachlässigt wird. Ein Wert nahe eins liefert dabei ein gutes Ergebnis.[244]

---

[243] Vgl. Müller, R., Lenz, H. (2013), S.98
[244] Vgl. Müller, R., Lenz, H. (2013), S.99

$$\textbf{F-Measure } (F_1): \frac{2*TP}{2*TP+FP+FN} = \frac{2*Recall*Precision}{Recall+Precision} \quad (5.26)$$

Im Folgenden werden die Ergebnisse (Vorhersagegenauigkeiten) der drei angewendeten Modelle auf die Trainings- und Testdaten zur Erkennung von Anomalien aufgeführt. Bei den Ergebnissen handelt es sich um die durchschnittlichen Genauigkeiten. Diese ergeben sich aus den Ausführungen der jeweiligen Modelle auf den fünf erzeugten Trainings- und Testdatensätze. Im Anhang 9 befindet sich die Berechnung der einzelnen Ergebnisse der jeweiligen Modelle.

	Logistische Regression	**Support Vector Machine**	**RandomForest Classifier**
Accuracy	(94 %)	(100 %)	(100 %)
Klassifikationsfehler	(6 %)	(0 %)	(0 %)
Precision	(55 %)	(100 %)	(100 %)
Recall	(100 %)	(100 %)	(100 %)
F-1	(71%)	(100 %)	(100 %)

Tab. 10 Durchschnittliche Ergebnisse der Vorhersagegenauigkeit (Trainingsdaten)

	Logistische Regression	**Support Vector Machine**	**RandomForest Classifier**
Accuracy	(95 %)	(93 %)	(99 %)
Klassifikationsfehler	(5 %)	(7 %)	(1 %)
Precision	(62 %)	undefinierbar	(98 %)
Recall	(98 %)	(0 %)	(95 %)
F-1	(75 %)	(0 %)	(97 %)

Tab. 11 Durchschnittliche Ergebnisse der Vorhersagegenauigkeit (Testdaten)

Die Ergebnisse zeigen, dass das Modell der Support Vector Machine auf den Trainingsdaten sehr gute Ergebnisse erzielt hat, diese auf den Testdaten allerdings eine Verschlechterung aufweisen und somit auf eine Überanpassung (Overfitting) des Modells hindeuten. Die Logistische Regression erzielt sowohl auf den Trainings- und Testdaten eine hohe Klassifikationsgenauigkeit (Accuracy), allerdings eine niedrige Präzision (Precision), wodurch ein erhöhtes Aufkommen von Fehlalarmen möglich ist. Auch die Kombination aus Precision und Recall weisen einen geringeren Wert auf. Die besten Ergebnisse erzielt das Modell des RandomForestClassifier. Sowohl auf den fünf unterschiedlichen Trainings- als auch Testdaten wird eine hohe Modellgüte erreicht. Neben der Genauigkeit für die Modellauswahl sind die Performance und die Stabilität als

weitere Kriterien zu betrachten. Die Performance ist über die Anzahl der verwendeten Felder zu bewerten. Je niedriger die Anzahl der verwendeten Felder, desto schneller die Bearbeitungszeit. Da lediglich zwei Merkmale für das Training des Modells herangezogen wurden, ist die Performance als sehr gut zu bewerten. Ebenfalls ist die Stabilität des Modells als positiv anzusehen, da sich bei mehreren Ausführungen die Gesamtgenauigkeit kaum verändert hat (siehe Berechnung der Bewertungsmaße der einzelnen RandomForestClassifier Modelle im Anhang 9). Aufgrund dieser Faktoren ist die Entscheidung zur Anwendung eines Modells auf einen unbekannten Datensatz auf den Algorithmus (Modell) RandomForestClassifier gefallen.

In einem nächsten Schritt sind die fünf erzeugten Modelle des RandomForestClassifier auf den Datensatz „Datensatz_Ohne_Spalte_verhalten“ (Zeitraum vom 25. Mai bis 06. Juni 2016) angewendet worden. Dieser Datensatz enthält lediglich die Spalten „_time“, „count“, „percentageDifftag“, „diff_1tag“ und „countletztertag“. Das Verhalten (Anomalie oder Normal) ist durch das trainierte Modell vorherzusagen. Die jeweiligen Ergebnisse, der Modelle des RandomForestClassifier sind im Anhang 7 aufgeführt.

Die Tabelle 12 zeigt einen Ausschnitt der Ergebnisse, der die vom jeweiligen Modell des RandomForestClassifier identifizierten Anomalien bei Anwendung auf dem unbekannten Datensatz darstellt.

Bezeichnung des Modells	**Anzahl erkannter Anomalien**
BA_Modell_RandomForest_Split_30_Train_70_Test	21
BA_Modell_RandomForest_Split_30_Training_70_Test_2_	17
BA_Modell_RandomForest_Split_30_Training_70_Test_3	21
BA_Modell_RandomForest_Split_30_Training_70_Test_4	19
BA_Modell_RandomForest_Split_30_Training_70_Test_5	23

Tab. 12 Anzahl der erkannten Anomalien von RFC Modellen

In der obigen Tabelle ist zu erkennen, das bei der Anwendung des RandomForestClassifier Algorithmus auf einem Datensatz, bei dem die Spalte „verhalten“ nicht existiert, Anomalien erkannt werden. Die Anzahl der erkannten Anomalien ist bei allen fünf Modellen ähnlich. Dies untermauert die hohe Modellgüte des Algorithmus.

Der folgende Ausschnitt stellt beispielhaft dar, wie ein zuvor trainiertes Modell (RandomForestClassifier) Anomalien auf einem unbekannten Datensatz identifiziert. Die

Ergebnisse (Darstellung in Diagrammen) der einzelnen RandomForestClassifier Modelle befinden sich im Anhang 8.

Abb. 18 Anwendung des RFC zur Erkennung von Anomalien auf unbekannte Daten

Die blaue Linie stellt den Count über den Zeitraum vom 25. Mai bis 06. Juni 2016 dar. Die vom Algorithmus erkannten Anomalien werden mit gelben Punkten abgebildet. In diesem Fall wurde ein Wert als Anomalie gekennzeichnet, bei dem das Verhalten bezogen auf den Vortag abweicht. Es ist darauf hinzuweisen, dass ein Peak in der Abbildung nicht unbedingt auf eine Anomalie hindeutet. Daher sind die hier als Anomalie klassifizierten Werte in einem weiteren Schritt auf Validität zu überprüfen.

6 Schlussbetrachtung

Große Datenmengen können durch das ständige Fortschreiten von Technologien gespeichert und verarbeitet werden. An dieser Stelle ist die Anwendung von Machine-Learning sinnvoll und hilfreich und findet in sämtlichen Bereichen von Unternehmen und Organisationen Anwendung. Durch die Verwendung von Machine-Learning kann ein bereits bekanntes Muster in neuen Daten wiedererkannt bzw. anhand von Erfahrungswerten, die aus der Vergangenheit stammen, ein bestimmtes Leistungskriterium optimiert werden.

6.1 Zusammenfassung

In der Otto GmbH & Co. KG werden große Datenmengen gespeichert und verarbeitet. In der Backend-Umgebung werden 13 Milliarden Events im Zeitraum der letzten 36 Tage indiziert, die die Grundlage für fundierte Entscheidungen bilden. Diese Daten können durch Analysen einen großen Informationsgewinn ermöglichen und beispielsweise für Vorhersagen genutzt werden, bei denen die Anwendung von Machine-Learning-Algorithmen einen wesentlichen Beitrag leisten kann. Durch Machine-Learning-Algorithmen besteht die Möglichkeit auftretende Anomalien in den Daten zu entdecken, die auf einen kritischen Zustand hinweisen, bei dem ein Handlungsbedarf besteht.

In der Backend-Architektur der Otto GmbH & Co. KG werden Anomalien derzeit weder durch statistische Verfahren, noch automatisiert mit Hilfe von Machine-Learning-Algorithmen und Technologien erkannt. Grund hierfür ist vor allem das Fehlen personeller Ressourcen, die aufgrund der Wachstumsausrichtung des Konzerns an ihre Aufgaben gebunden sind und somit kein Verständnis für die beschriebene Thematik aufbauen konnten. Daher wurde in der vorliegenden Arbeit stets das Ziel verfolgt, ein grundlegendes Verständnis aufzubauen, wie in den Datenmengen Anomalien, die ein vom Normalverhalten abweichendes Muster aufweisen, mit Hilfe von Machine-Learning-Algorithmen erkannt werden können.

Mit Hilfe dieser Arbeit konnte Wissen für die Auf- und Vorverarbeitung der Daten aufgebaut werden. Diese Phase ist als besonders bedeutsame Phase des KDD-Prozesses anzusehen. Um einen Transfer in das Betriebsumfeld der Otto GmbH & Co. KG zu gewährleisten, wurde weiter untersucht welche Technologie und welcher Algorithmus für welche Daten geeignet sind und wie diese Daten entsprechend aufbereitet werden müssen, um Anomalien zu erkennen. Zur systemseitigen

Unterstützung der Untersuchungen ist als Technologie die Machine-Learning-App von Splunk gewählt worden, die zahlreiche Algorithmen zur Analyse der Daten beinhaltet.

Aufgrund der begrenzten Zeit dieser Thesis lag der Fokus auf dem Ansatz des überwachten Lernens. Hierbei fand das Verfahren der Klassifikation Anwendung, wobei drei Algorithmen (Logistische Regression, Support Vector Machine (SVM) und RandomForestClassifier) näher untersucht und die Ergebnisse anhand von Bewertungsmethoden miteinander verglichen wurden. Die Analyse hat sich stets an den Phasen des KDD-Prozesses orientiert. Hierbei sind die Algorithmen anhand eines erzeugten Trainingsdatensatzes trainiert und anschließend auf einem Testdatensatz angewendet worden. Dabei hat sich herauskristallisiert, dass der RandomForestClassifier die beste Modellgüte aufweist und als Modell zur Erkennung von Anomalien auf einem unbekannten Datensatz angewendet wurde.

6.2 Kritische Würdigung

Basis dieser Arbeit sind nicht nur die wissenschaftlichen Grundlagen und der Stand der Forschung, sondern auch Experteninterviews, die im Rahmen dieser Arbeit durchgeführt worden. Durch das Zusammenspiel von Ergebnissen aus der Forschung und Erkenntnissen aus den Interviews konnte die Theorie in die eigene Analyse Einfluss finden.

Dennoch ist hervorzuheben, dass von den vier beschriebenen Data Mining-Verfahren (siehe Abbildung 10) lediglich ein Verfahren (Klassifikation) für die eigene Analyse herangezogen werden konnte. Die anderen drei Verfahren konnten nicht berücksichtigt werden, da allein die Datenaufbereitung innerhalb des gewählten Verfahrens (Klassifikation) sehr viel Zeit in Anspruch genommen hat. Die Datenaufbereitung und Transformation stellt somit die zeitaufwendigste Phase im KDD-Prozess dar.

Positiv zu bewerten ist, dass durch die Anwendung eines Data Mining-Verfahrens ein Algorithmus bestimmt werden konnte, mit dem innerhalb der Machine-Learning App von Splunk Anomalien in unqualifizierten Eventmengen erkannt werden konnten.

Solange die drei anderen Verfahren des Data Mining nicht untersucht wurden, kann aus diesem Ergebnis allerdings keine Allgemeingültigkeit abgeleitet werden.

6.3 Ausblick

Nachdem im Rahmen dieser Arbeit der Ansatz des überwachten Lernens mit Hilfe des Klassifikationsverfahrens zur Erkennung von Anomalien betrachtet wurde, ist in einem weiteren Schritt der Ansatz des unüberwachten Lernens zur Erkennung von Anomalien zu untersuchen. Zur systemseitigen Unterstützung könnte hierfür die in den Experteninterviews erwähnte Anomaliy-Detective-App von Prelert herangezogen werden, da diese zu 100 % Algorithmen des unüberwachten Lernens beinhaltet.[245] Anschließend wird empfohlen die Erkenntnisse beider Untersuchungen gegenüberzustellen, um die für die Praxis besser geeignete Methode festzustellen und diese entsprechend in das Betriebsumfeld zu integrieren.

In einem nächsten Schritt sind die gewonnen Erkenntnisse auf einen konkreten Anwendungsbereich, bei dem es sich um qualifizierte Eventmengen der Backend-Architektur handelt, anzuwenden, um mögliche Anomalien zu erkennen und zur Überwachung der Systeme beizutragen. Die Überwachung der Verfügbarkeit einer bestimmten Datenbankverbindung, des Ausfalls einer konkreten Systemkomponente oder des Fehlschlagens von Benutzerlogins sind an dieser Stelle als Beispiele zu nennen. Hierbei könnte die Möglichkeit geprüft werden, ob nach Erkennung von Anomalien die in den Eventmengen auftreten, eine Handlungsempfehlung in Form einer E-Mail, Incident oder einer anderen Art der Eskalation erfolgen kann.

Bereits in der Vorrecherche hat sich gezeigt, dass Machine-Learning im Zusammenhang mit der Anomalie-Erkennung ein komplexes Themengebiet mit verbundenen Herausforderungen darstellt. Dies konnte durch Experten bestätigt werden.[246] Daher ist abschließend festzuhalten, dass in der Praxis und auch in der Otto GmbH & Co. KG noch mehr Forschung betrieben werden sollte, um die Zuverlässigkeit der Ergebnisse zu steigern. Schlussendlich wird die Technik der Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen dazu beitragen, den technologischen Fortschritt voranzutreiben und neue Potenziale zu entdecken.

---

[245] Vgl. Borner, M. (2016), Management Consultant Anhang 2, Frage/Antwort 2.2 und Prelert Inc., Anomaly Detective API Engine: Put Machine Learning to Work - Benefits

[246] Vgl. Drieger, P. (2016), Sales Engineer, Experteninterview Anhang 1, Borner, M. (2016), Management Consultant Anhang 2 und Günther, C. (2016), Leiter Data Analytic, Experteninterview Anhang 3

Anhang

Anhang 1: Leitfaden für Experteninterview: Unternehmen Splunk Inc.

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Experteninterview

„Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen/Technologien“

Niklas Netz
Telefon: +49 (0) 40 64611744
Mobil: +49 (0) 171530512
Mail: niklas.netz@haw-hamburg.de oder niklas.netz@otto.de

Fakultät: Technik und Informatik
Department: Informatik
Hochschule für Angewandte Wissenschaften
Studiengang: Bachelor of Science Wirtschaftsinformatik

Leitfaden für Experteninterviews

Einleitung

Vielen Dank, dass Sie mir im Rahmen meiner Bachelor-Thesis, die Gelegenheit für dieses Experteninterview geben. Mein Name ist Niklas Netz und ich studiere Wirtschaftsinformatik an der Hochschule für Angewandte Wissenschaften in Hamburg und bin nebenbei als Werkstudent (Data Analyst) bei der Otto GmbH & Co.KG tätig. Die Bachelorarbeit wird in Kooperation mit dem Unternehmen Otto GmbH & Co. KG durchgeführt. Das Thema meiner Bachelor-Thesis lautet: „Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen/Technologien."

In der Otto GmbH & Co. KG wird zur Analyse von Log-Dateien unter anderem die Software „Splunk" eingesetzt. Es werden derzeit über 13 Milliarden indizierte Events in der Splunk verarbeitet, die die Grundlage für fundierte Entscheidungen bilden. Durch eine Retention-Time von 36 Tagen, können die letzten fünf Tage des Vormonats aufbewahrt werden.

In der Backend-Architektur der Otto GmbH & Co. KG werden Anomalien derzeit weder durch statische Verfahren, noch automatisiert mit Hilfe von Machine-Learning-Algorithmen und -Technologien erkannt. Auch ist ein Grundverständnis für die beschriebene Problematik nicht vorhanden. Derzeit wird zur Erkennung von Anomalien auf den Dashboards lediglich die Methode des „scharfen Hinsehens" verwendet, die viele personelle Ressourcen bindet. Diese personellen Ressourcen könnten im operativen Bereich der Otto GmbH & Co. KG daher für andere Tätigkeiten wertvoller eingesetzt werden.

In der vorliegenden Arbeit wird untersucht, wie Anomalien anhand von unqualifizierten Eventmengen in der Backend-Architektur der Otto GmbH & Co. KG erkannt werden können. Hierbei soll ein Grundverständnis für das Erkennen von Anomalien mit statistischen Verfahren, sowie mit Hilfe von Machine-Learning-Algorithmen und -Technologien geschaffen werden, wobei der Fokus auf der Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen und -Technologien liegt. Um anhand der Eventmengen Anomalien zu erkennen, ist vorab ein Verständnis für die Auf- bzw. Vorbereitung der Daten aufzubauen, um einen Transfer in das Betriebsumfeld der Otto GmbH & Co. KG zu erzielen. Wichtig ist dabei vor allem die Untersuchung, welche Technologie und welcher Algorithmus für welche Daten geeignet sind und wie die Daten

für die entsprechende Technologie und den entsprechenden Algorithmus aufbereitet werden müssen.

Das Interview wird ca. 90 Minuten in Anspruch nehmen.

Zur Durchführung des Interviews möchte ich noch kurz einige Erläuterungen geben:

Da alle Befragten dieselben Fragen während des Interviews gestellt bekommen, kann es vorkommen, dass Ihnen einige Fragen zu Ihrer speziellen Situation nicht passend erscheinen. Deshalb bitte ich Sie dies im Voraus zu entschuldigen.

Zudem werde ich das Gespräch aufzeichnen. Dies dient rein zur Kontrolle meiner Mitschrift. Sind Sie damit einverstanden?

Sollte etwas veröffentlicht werden, so wird es zur Autorisierung vorgelegt.

Das Gespräch ist wie folgt gegliedert:

1. Anomalie-Erkennung
2. Technologien bzw. Applikationen zur Anomalie-Erkennung
3. Statistische Anomalie-Erkennung
4. Anomalie-Erkennung mit Hilfe von Machine-Learning

Haben Sie noch Fragen bevor es losgeht?

Interview-Leitfaden

Datum: 19.04.2016, 09.30 Uhr **Ort:** Hamburg

Angaben zur Person: Philipp Drieger

Position/Funktion: Sales Engineer

Akademische Laufbahn (kurz):

Magister Artium Philosophie, Erziehungswissenschaften und Informatik. M.A. Arbeit: “Virtuelle Welten. Eine neue Dimension unseres Wirklichkeitsverständnisses.”

Wissenschaftliche Veröffentlichungen:

Drieger, P.: Semantic Network Analysis as a Method for Visual Text Analytics. In: Procedia – Social and Behavioral Sciences, Volume 79, 6 June 2013, Manuel Fischer (Ed.), Elsevier, pp. 4–17. 9th Conference on Applications of Social Network Analysis (ASNA) 2012.

Drieger, P.: Visual text analytics using semantic networks and interactive 3d visualization. In: EuroVA 2012: International Workshop on Visual Analytics (Vienna, Austria, 2012), K. Matkovic and G. Santucci (Eds.), Eurographics Association, pp. 43–47.

Dissertation (nicht abgeschlossen), Arbeitstitel “Visual Text Analytics”

Berufliche Laufbahn (kurz):

Seit 2002 tätig als Softwareentwickler und IT Consultant in verschiedenen Branchen (Automotive, Logistik, Software). Weitere Infos und Details siehe http://www.noumentalia.de/

Seit 2015 als Sales Engineer bei Splunk (http://www.splunk.com/) in verschiedensten Projekten (IT Operations, Security, Business Analytics und IoT) in den Regionen DACH und EE tätig.

Seit 2016 Subject Matter Expert für IoT und Machine Learning.

Anomalie-Erkennung

1.1 Zuerst würde ich gerne von Ihnen wissen, was Sie unter einer Anomalie verstehen?

Eine Anomalie ist für mich eine Abweichung von einem erwarteten Verhalten, Muster oder Struktur. Das bedeutet, dass eine Anomalie nur dann als solche erkannt werden kann, wenn das zu erwartende Verhalten bekannt ist. Das heißt, je nach Art des Datensets können sich diese Strukturen stark unterscheiden, z.B. eine univariate Zeitreihe im Vergleich zu einer multivariaten Struktur mit numerischen und kategorialen Feldern. Je nach Datenset machen deshalb unter Umständen nur bestimmte Ansätze Sinn.

1.2 Wie lassen sich Ausreißer bzw. Anomalien erkennen?

Übliche Ansätze zur Erkennung gehen zuerst von einem „Baselining“ aus, indem eine passende Metrik aus historischen Daten ermittelt wird. Anhand des Baselinings können Abweichungen, also Anomalien dann erkannt werden. Diese Metrik kann mit Hilfe verschiedener statistischer Methoden, z.B. (gleitender) Durchschnitt, Median, Quartile, etc. ermittelt werden.

1.3 Welche möglichen Ursachen für das Auftreten von Anomalien gibt es?

Mögliche Ursachen können entweder durch tatsächliche Abweichungen in den Daten gegeben sein oder aber auch durch fehlende Daten oder Fehler in den Daten, z.B. verursacht durch Messfehler. Wenn man solche Fehler ausschließen kann und es sich um eine Anomalie handelt, dann kann anhand der konkreten Datenlage auch verstanden werden, was die Ursache für die Anomalie ist. Darüber hinaus können durch die Korrelation mit anderen Daten mögliche Zusammenhänge oder Ursachen weitergehend untersucht werden.

1.4 Welche Herausforderungen bzw. Probleme sehen Sie in der Anomalie-Erkennung?

In meinen Augen liegt die größte Herausforderung in der Findung eines geeigneten Verfahrens bzw. Algorithmus, der Anomalien in einem Datenset bestmöglich, d.h. mit geringster Fehlerrate (false positives) und höchstem Automatisierungsgrad identifiziert. Im anderen Fall muss unter Umständen manuell jede einzelne gefundene Anomalie überprüft werden oder es gibt zu viele false positives, so dass das Verfahren unbrauchbar ist. Das Problem der Erkennung kann deshalb sehr vielschichtig sein: Ist das Datenset für die gewünschte Anomalieerkennung geeignet? Besitzt es schon

entsprechend brauchbare Attribute oder muss es ggf. vorverarbeitet oder angereichert werden? Hier ist das Stichwort Feature Engineering zu nennen. Welcher Algorithmus und welche Methode zur Erkennung, z.B. Klassifizierung, Clustering, Next Neighbor, statistisch, informationstheoretisch oder spectral angewendet werden.

1.5 Können Sie mir einen konkreten Use-Case bezüglich der Anomalie-Erkennung nennen?

Im Security Bereich spielt Anomalie-Erkennung historisch schon eine große Rolle z.B. bei Intrusion Detection Systemen, also bei der Erkennung eines Einbruchs in ein Computersystem oder –Netzwerk. Wenn man an Betrugserkennung z.B. im Bereich von Finanztransaktionen denkt, ist Anomalie-Erkennung auch ein prominenter Ansatzpunkt. Auch bei der Bildverarbeitung, z.B. im medizinischen Bereich oder in der Textanalyse spielen solche Verfahren eine Rolle. Und schließlich auch im Bereich des IT-Betriebs „IT Operations“, wenn es darum geht, Systemausfälle oder Beeinträchtigungen von Komponenten in einer Infrastruktur frühzeitig zu erkennen, um schnellstmöglich handeln zu können.

1.6 Haben Sie bereits einen Use-Case im Bereich der Anomalie-Erkennung umgesetzt?

Im Security Bereich habe ich in einem Projekt mit Hilfe eines Clustering-Verfahrens eine Anomalie-Erkennung umgesetzt. In einem anderen Projekt ging es darum, Anomalien im Bereich von Sensordaten aus der industriellen Fertigung zu erkennen.

1.7 Wenn ja, wie haben Sie die Anomalie-Erkennung umgesetzt und in welchen Anwendungsgebiet?

Die genannte Anomalie-Erkennung im Security Bereich beinhaltete kein „Baselining“, wie ich es oben genannt habe, sondern eine Clusteranalyse, welche dann Ausreißer, also Events, die sich nicht mit mehreren anderen Events in einem Cluster zusammen gruppiert haben, identifizieren konnte. Technisch wurde dies mit dem Splunk Befehl „cluster“ umgesetzt.

Das andere Projekt mit den Sensordaten wurde mit einem technisch aufwändigeren Verfahren umgesetzt bei dem mehrere Algorithmen kombiniert zum Einsatz kamen.

Technologien bzw. Applikationen zur Anomalie-Erkennung

2.1 Welche Technologie bzw. App bezüglich der eingesetzten Applikation „Splunk“ empfehlen Sie, um Anomalien zu erkennen?

Mit Hilfe von Splunk kann sich dem Thema der Anomalie-Erkennung auf verschiedenen Wegen angenähert werden. Zunächst gibt es bereits einige Befehle, die in der Splunk Suchsprache SPL (Search Processing Language) verfügbar sind:

1. „anomalies“: Für jedes Event wird ein „unexpectedness value“ errechnet, welcher dann mit einem Schwellwertverfahren Anomalien identifiziert.

2. „anomalousvalue“: für alle Felder in allen Events wird ein anomalie score errechnet und das Ergebnis in Form neuer Felder zu den Events hinzugefügt, welche dann für die weitere Analyse verwendet werden können.

3. „anomalydetection“: errechnet für alle Events eine Wahrscheinlichkeit und identifiziert so Events, die mit geringer Wahrscheinlichkeit auftreten und damit als Anomalien gekennzeichnet werden. (seit der Splunk Version 6.3)

4. „cluster“: bestimmt mit Hilfe eines definierbaren Parameters Events die anhand eines Vergleichs der Textvektoren zueinander ähnlich sind und ordnet so die Events Clustern zu. Ein Ansatz, um hier Anomalien zu erkennen, wäre z.B. die Cluster mit den wenigsten Events zu untersuchen.

5. „kmeans“: gruppiert ähnliche Events in eine durch einen Parameter definierbare Anzahl von Clustern.

Über diese Befehle hinaus lässt sich z.B. mit Hilfe der Befehle streamstats, x11 oder trendline ein (adaptives) Baselining definieren, welches dann zur Identifikation von Anomalien verwendet werden kann.

Der Befehlssatz der Suchsprache in Splunk kann aber, wie Sie sagen, mit Apps erweitert werden. Die aktuell wichtigste App in diesem Umfeld ist die Python basierte Machine Learning App von Splunk, wo sich eine ganze Reihe zusätzlicher Algorithmen findet.

2.2 Haben Sie mit der genannten App bereits Erfahrungen im Umgang der Anomalie-Erkennung gesammelt?

Ja, die Machine-Learning App bietet hier Methoden zur Erkennung numerischer und kategorialer Outlier, welche mit Hilfe eines graphischen Interfaces auf verschiedene Beispieldatensets direkt angewendet und getestet werden kann.

2.3 Was für einen Use-Case haben Sie mit der App hinsichtlich der Anomalie-Erkennung behandelt?

Da die Machine-Learning App aktuell noch im beta Stadium ist, wurde hiermit noch kein produktives Projekt umgesetzt, allerdings bin ich in einigen Projekten involviert, die diese App evaluieren und verschiedene Use-Cases erforschen. Ein aktuell laufendes Projekt im Bereich IT Operations ist die Kategorisierung der Fehlercodes von Alerts anhand historischer Daten mit dem Ziel die Kritikalität der Alerts zu bestimmen und die Rate an false positives zu reduzieren.

2.4 Mit welchen Mitteln haben Sie die Anomalie-Erkennung durchgeführt?

In diesem Fall erfolgt die Bestimmung der Anomalien mit Hilfe kategorialer Verfahren wie z.B. logistische Regression. Eine Anomalie entspricht in diesem Fall der Erkennung tatsächlich kritischer Events.

2.5 Statistische Mittel oder mit Hilfe von Machine-Learning-Algorithmen?

Siehe 2.4. Logistische Regression. Wie schon angesprochen, wäre die Logistische Regression ein Ansatzpunkt, wobei die Machine-Learning App mit einer ganzen Reihe an bereits implementierten Algorithmen ausgestattet ist und diese dann entsprechend verwendet werden können. Je nachdem welcher Algorithmus für die Problemstellung der richtige ist.

2.6 Wie würden Sie die Abgrenzung der Anomalie-Erkennung mit Hilfe von statistischen Mitteln von der Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen beschreiben?

Ich würde zunächst eine Gemeinsamkeit beider Ansätze erwähnen, nämlich dass beide Verfahren auf historischen Daten beruhen, welche zur Anomalie-Erkennung herangezogen werden. Der Unterschied liegt meines Erachtens darin, dass im Falle von Machine-Learning ein Modell aus den Daten heraus generalisiert wird und zur Identifikation von Anomalien herangezogen wird wohingegen bei statistischen Verfahren „nur“ ein „Baselining“ stattfindet und darauf aufbauend eine ggf. sogar adaptive, dynamische aber immer irgendwie schwellwertbasierte Identifikation der Anomalien erfolgt. Es liegt aber auch an der Definition, was Sie unter „statistischen Mitteln“ verstehen, da viele Machine-Learning Algorithmen ja natrülich auf statistischen Ansätzen aufbauen.

3.1 Wie sind Sie zu Beginn der Datenanalyse hinsichtlich der Anomalie-Erkennung mit statistischen Mitteln vorgegangen?

Zunächst gilt es ein geeignetes Feld entweder in den Daten zu finden oder es entsprechend zu erzeugen, z.B. mit den Splunk Befehlen timechart oder stats count by. Dann gilt es diese Ausprägungen dahingehend zu analysieren, ob Sie zur Erkennung der Anomalien geeignet sind.

3.2 Mussten die Daten zuvor aufbereitet werden?

In der Regel ja, da man nicht immer davon ausgehen kann, dass die Daten bereits gut geeignet sind. Konsistenz und bestmögliche Qualität der Daten sind eine gute Grundlage für die Anomalie-Erkennung.

3.3 Wenn ja, wie haben Sie diese Datenaufbereitung bzw. Datenvorbereitung durchgeführt?

Wenn man komplett von den Rohdaten her beginnt, können mit Hilfe von Splunk, aus den unstrukturierten Daten zunächst diejenigen Felder extrahiert werden, die für die Analyse wichtig sind. Dabei gibt es bereits die Möglichkeit, die Qualität der Regex basierten Feldextraktion zu prüfen und ggf. zu korrigieren. Wenn die Daten dann in der Splunk Suche verfügbar sind, kann man sich einen ersten Überblick verschaffen und das Datenset z.B. auf offensichtliche Ausreißer oder Inkonsistenzen überprüfen und ggf. auch Bereinigen (Data Cleansing), z.B. mit dem Splunk Befehl „outlier“. Der große Ausreißer oder fehlerhafte Werte entfernt. Ferner gibt es weitere Splunk Befehle, um die Daten z.B. anzureichern. „lookup“ wäre hier zu nennen oder zusätzliche Felder zu erzeugen. Da wäre hier der Befehl „eval“ (für Evaluation von Feldern) zu nennen. Darüber hinaus können die Daten ggf. auch weiter gefiltert, aggregiert oder verdichtet und in einem neuen Index, einem sogenannten „summary index“ gespeichert werden.

3.4 Welche Statistischen Mittel bzw. Kommandos haben Sie zur Erkennung von Anomalien verwendet?

Bei den rein statistischen Ansätzen habe ich bisher mit den Befehlen „streamstats“, „eventstats“, sowie „timechart“ und „stats“ mit verschiedenen Funktionen wie count, average, min, max, median, perc, stdev, etc. gearbeitet.

3.5 Können Sie mir evtl. eine Möglichkeit nennen, mit der ich ein vereinfachtes Testszenario zur Erkennung von Anomalien mit Hilfe von statistischen Mitteln innerhalb der Splunk-Umgebung darstellen kann?

Das Beispieldashboard „Detect Numeric Outliers“ in der Machine-Learning App zeigt genau solch einen rein statistischen Ansatz und kommt allein mit den Befehlen „eval“ und „streamstats“ aus. Als verfügbares Beispieldatenset sind dort „Antwortzeiten eines Servers“ aufgeführt.

4.1 Wie sind Sie zu Beginn der Datenanalyse hinsichtlich der Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen grundsätzlich vorgegangen?

Eine gute Datengrundlage ist auch hier erforderlich und deckt sich demnach mit dem Vorgehen wie in Frage 3.3 beschrieben. Ggf. prüft man die Werte in den Feldern hier etwas genauer, um gute Konsistenz für Machine-Learning-Algorithmen zu gewährleisten. Erwähnenswert ist an dieser Stelle der Splunk Befehl „analyzefields“, welcher es erlaubt, geeignete Felder für die Vorhersage zu identifizieren.

4.2 Mit welcher Applikation haben Sie innerhalb von Splunk die Anomalie-Erkennung durchgeführt? ML-App? Prelert?

Bisher mit den oben beschriebenen Splunk Suchbefehlen sowie der genannten Machine-Learning App und den darin enthaltenen Algorithmen. Mit Prelert bisher noch nicht.

4.3 Welche Algorithmen unterstützt die jeweilige Applikation?

Splunk Enterprise unterstützt in der Suchsprache folgende Algorithmen. Für Clustering gibt es „kmeans“ und „cluster“ im Befehlssatz. Direkt zur Anomalie Detektion die bereits angesprochenen „anomal*“ Befehle. Zur Erzeugung statistischer Metriken die oben genannten Befehle „stats“, „streamstats“, „eventstats“, „eval“.

In der Machine-Learning App ab der Version 1.0 sind folgende Algorithmen verfügbar: Birch, DBSCAN, KMeans, LinearRegression, LogisticRegression, PCA, SVM, SpectralClustering, BernoulliNB, ElasticNet, FieldSelector, GaussianNB, KernelPCA, Lasso, OneClassSVM, RandomForestClassifier, RandomForestRegressor, Ridge, StandardScaler, TFIDF.

Wichtig ist hierbei anzumerken, dass alle genannten Algorithmen in Python implementiert sind und ggf. modifiziert werden können. Ebenso kann man sehr einfach neue Algorithmen in das bestehende Framework mit geringem Aufwand hinzufügen.

4.4 Mussten die Daten zuvor aufbereitet werden?

Ja, wie oben beschrieben.

4.5 Wenn ja, wie haben Sie diese Datenaufbereitung bzw. Datenvorbereitung durchgeführt?

Auch wie schon oben beschrieben. Entsprechende Datenqualität sichern und die entsprechenden Felder identifizieren.

4.6 Welche Algorithmen empfehlen Sie zur Erkennung von Anomalien?

Ich empfehle grundsätzlich einfach anzufangen. In der Suchsprache gibt es ja bereits einige Befehle, die sehr schnell und einfach auf Daten in Splunk angewendet werden können. So kann man schnell verschiedene Ansätze testen und eine Einschätzung bekommen, was passt. Ich würde nach einer ersten Datenexploration z.B. mit Hilfe von „timechart“, „stats“ etc. zunächst mit „cluster“, ggf. auch mit „kmeans“ und den „anomal*“ Befehlen beginnen. Wenn sich damit erste gute Ergebnisse zeigen, kann man die Parameter der Befehle verfeinern oder mit Hilfe von „eval“ weitere Transformationen auf den Daten einführen.

Wenn sich mit den verfügbaren Befehlen keine guten Ergebnisse erzielen lassen, würde ich gezielter mit der Machine-Learning App arbeiten. Unabhängig davon kann natürlich auch gleich mit der Machine-Learning App begonnen werden, wenn z.B. schon im Vorfeld ein konkreter Algorithmus identifiziert wurde, mit dem gearbeitet werden soll.

4.7 Welche Machine-Learning-Algorithmen haben Sie zur Erkennung von Anomalien verwendet?

In der Splunk Suchsprache: „cluster“, „kmeans“, „anomalydetection“ bzw. „anomaluesvalue“, sowie Verfahren mit „streamstats“. In der Machine-Learning App habe ich bisher mit PCA, DBSCAN, KMeans gearbeitet.

4.8 Anhand welcher Kriterien haben Sie den Algorithmus ausgewählt?

Die Auswahlkriterien sind primär anhand der Begebenheiten des Datensets zu klären. Sekundär kann man nichtfunktionale Merkmale wie z.B. die Laufzeit des gewählten Algorithmus in Erwägung ziehen. Ich möchte ein paar Beispiele zur Verdeutlichung anführen. Da der SPL Befehl „cluster“ intern die Events tokenisiert und auf dieser Basis die Ähnlichkeit der Events bestimmt, ist dieser gut geeignet für unstrukturierte, textlastige Events.

DBSCAN bietet im Vergleich zu anderen Clusteringalgorithmen meines Erachtens eine gute Performance zur Erkennung von Ausreißern, sowie eine gute Genauigkeit. Im Falle einer Zeitreihe, also mit „timechart“ aufbereitbare Daten, bietet sich ein „Baselining“ mit streamstats an.

Bei komplexen mehrdimensionalen Daten kann eine Reduktion der Dimensionen durch PCA (Principal Component Analysis) geeignet sein, um Outlier und Cluster zu erkennen.

4.9 Haben Sie mit einem der Algorithmen das Verhalten trainiert und dieses dann auf einen Testdatensatz angewendet?

Ja, in der Machine-Learning App gibt es hierfür die Befehle „fit“ und „apply“. Wenn ein Algorithmus ausgewählt wurde, kann mit Hilfe von „fit“ ein Modell trainiert werden. Mit dem Befehl „apply“ kann dann das trainierte Modell auf neue Daten oder wie Sie sagen auf einen Testdatensatz angewendet werden. In den oben genannten Beispielen lassen sich aber nicht alle Algorithmen in einem Modell persistieren. Z.B. verschiedene Clustering-Algorithmen.

4.10 Wie haben Sie das Trainieren der Daten durchgeführt?

Das Training erfolgt durch eine Aufteilung des Datensets in Trainingsdaten und Testdaten in einem bestimmten Verhältnis, z.B. 80 zu 20. Dann muss entschieden werden ob die Aufteilung durch zufälliges Sampling oder definierte Zeitabschnitte vorgenommen wird. Idealerweise werden verschiedene Ansätze der Kreuzvalidierung wiederholt und variiert, um dadurch zu gewährleisten, dass ein Modell weder overfitted noch underfitted ist.

4.11 Wie kann das durchs Trainieren gelernte Verhalten angewendet werden?

Im Falle der Machine-Learning App kann, wie gesagt, durch den Befehl „apply“ ein trainiertes Modell auf die Daten angewendet werden.

4.12 Auf welche Bereiche ist das gelernte anwendbar?

Das Modell kann auf Testdaten oder aber auch auf neue Daten angewendet werden. Man könnte durch sogenannte „scheduled searches“ in Splunk das Training eines Modells in einem gewünschten Zeitintervall automatisch wiederholen lassen, so dass das Modell immer auf dem aktuellen Datenbestand trainiert ist. Dieses kann dann z.B. zur Klassifizierung neuer Events verwendet werden.

Bevor wir das Interview abschließen, möchte Ich von Ihnen gerne wissen, ob aus Ihrer Sicht eine wichtige Frage ungestellt blieb? Ist Ihnen während des Interviews z.B. irgendein offener Punkt aufgefallen, den ich noch mit beachten sollte?

Ich finde Sie haben mit Ihren Fragen die wichtigsten Aspekte abgedeckt. Wenn man die Forschung und auch die Projekte in der Industrie in diesem Gebiet betrachtet, wird schnell klar, dass sowohl das Thema Anomalie-Erkennung als auch Machine-Learning nicht unkomplex ist. Viele Projekte haben das gezeigt und es bedarf auch einem gewissen Ehrgeiz, für ein konkretes Datenset eine gute Lösung zu entwickeln. Wenn dieses gefunden wird, ist die Aussicht allerdings sehr gut: viele Projekte und bekannte Produkte, die auf Machine Learning Technologien aufbauen, sind sehr erfolgreich und dann in der Regel auch wirtschaftlich von großem Wert für Unternehmen, die diese Technologien einsetzen. Genau das wünsche ich Ihnen bei Ihrer Arbeit zu diesem Thema bei OTTO und hoffe, dass Sie einen oder mehrere erfolgreiche Ansätze entwickeln. Für Fragen stehe ich Ihnen natürlich auch nach diesem Interview noch zur Verfügung, wünsche Ihnen viel Erfolg und bedanke mich, dass Sie mich zu diesem Thema interviewt haben.

Zusätzliche Fragen:

Können Sie mir evtl. noch Referenzkunden nennen, die im Bereich der Anomalie-Erkennung mit Hilfe von Machine-Learning und Splunk schon Erfahrungen gesammelt haben, um evtl. einen weiteren Experten zu interviewen?

Das würde ich sehr gerne. Es gibt in der Tat einige Kunden, die sich mit dem Thema der Anomalie-Erkennung beschäftigen, allerdings unterliegen wir da leider entsprechenden Vertragsregelungen, sodass ich Ihnen leider keine konkreten Kunden nennen kann. Insofern kann ich Ihnen an dieser Stelle nur sagen, dass das Thema heiß begehrt ist. Dies wird deutlich, da das Thema in sehr vielen Fragestellungen genau mit diesen Thematiken betrachtet wird.

Dank: Zum Abschluss möchte ich mich ganz herzlich für die Zeit, die Sie sich genommen haben bedanken!

Anhang 2: Leitfaden für Experteninterview: Unternehmen LC Systems

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Experteninterview

„Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen/Technologien“

Niklas Netz
Telefon: +49 (0) 40 64611744
Mobil: +49 (0) 171530512
Mail: niklas.netz@haw-hamburg.de oder niklas.netz@otto.de

Fakultät: Technik und Informatik
Department: Informatik
Hochschule für Angewandte Wissenschaften
Studiengang: Bachelor of Science Wirtschaftsinformatik

Leitfaden für Experteninterviews

Einleitung

Vielen Dank, dass Sie mir im Rahmen meiner Bachelor-Thesis, die Gelegenheit für dieses Experteninterview geben. Mein Name ist Niklas Netz und ich studiere Wirtschaftsinformatik an der Hochschule für Angewandte Wissenschaften in Hamburg und bin nebenbei als Werkstudent (Data Analyst) bei der Otto GmbH & Co.KG tätig. Die Bachelorarbeit wird in Kooperation mit dem Unternehmen Otto GmbH & Co. KG durchgeführt. Das Thema meiner Bachelor-Thesis lautet: „Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen/Technologien."

In der Otto GmbH & Co. KG wird zur Analyse von Log-Dateien unter anderem die Software „Splunk" eingesetzt. Es werden derzeit über 13 Milliarden indizierte Events in der Splunk-Umgebung verarbeitet, die die Grundlage für fundierte Entscheidungen bilden. Durch eine Retention-Time von 36 Tagen, können die letzten fünf Tage des Vormonats aufbewahrt werden.

In der Backend-Architektur der Otto GmbH & Co. KG werden Anomalien derzeit weder durch statische Verfahren, noch automatisiert mit Hilfe von Machine-Learning-Algorithmen und -Technologien erkannt. Auch ist ein Grundverständnis für die beschriebene Problematik nicht vorhanden. Derzeit wird zur Erkennung von Anomalien auf den Dashboards lediglich die Methode des „scharfen Hinsehens" verwendet, die viele personelle Ressourcen bindet. Diese personellen Ressourcen könnten im operativen Bereich der Otto GmbH & Co. KG daher für andere Tätigkeiten wertvoller eingesetzt werden.

In der vorliegenden Arbeit wird untersucht, wie Anomalien anhand von unqualifizierten Eventmengen in der Backend-Architektur der Otto GmbH & Co. KG erkannt werden können. Hierbei soll ein Grundverständnis für das Erkennen von Anomalien mit statistischen Verfahren, sowie mit Hilfe von Machine-Learning-Algorithmen und -Technologien geschaffen werden, wobei der Fokus auf der Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen und -Technologien liegt. Um anhand der Eventmengen Anomalien zu erkennen, ist vorab ein Verständnis für die Auf- bzw. Vorbereitung der Daten aufzubauen, um einen Transfer in das Betriebsumfeld der Otto GmbH & Co. KG zu erzielen. Wichtig ist dabei vor allem die Untersuchung, welche Technologie und welcher Algorithmus für welche Daten geeignet sind und wie die Daten

für die entsprechende Technologie und den entsprechenden Algorithmus aufbereitet werden müssen.

Das Interview wird ca. 90 Minuten in Anspruch nehmen.

Zur Durchführung des Interviews möchte ich noch kurz einige Erläuterungen geben:

Da alle Befragten dieselben Fragen während des Interviews gestellt bekommen, kann es vorkommen, dass Ihnen einige Fragen zu Ihrer speziellen Situation nicht passend erscheinen. Deshalb bitte ich Sie dies im Voraus zu entschuldigen.

Zudem werde ich das Gespräch aufzeichnen. Dies dient rein zur Kontrolle meiner Mitschrift. Sind Sie damit einverstanden?

Sollte etwas veröffentlicht werden, so wird es zur Autorisierung vorgelegt.

Das Gespräch ist wie folgt gegliedert:

1. Anomalie-Erkennung
2. Technologien bzw. Applikationen zur Anomalie-Erkennung
3. Statistische Anomalie-Erkennung
4. Anomalie-Erkennung mit Hilfe von Machine-Learning

Haben Sie noch Fragen bevor es losgeht?

Interview-Leitfaden

Datum: 20.04.2016, 09.30 Uhr **Ort:** Hamburg

Angaben zur Person: Mika Borner

Position/Funktion: Management Consultant

Akademische Laufbahn (kurz):

Bachelor of Science IT & Telecommunications Häme University of Applied Sciences, Finland

Berufliche Laufbahn (kurz):

1999 – 2001 System Engineer Telia Mobile AB

2002 – 2005 System Engineer IBM Global Services

2005 – 2006 Senior System Engineer Clariden Bank

2006 – 2007 Senior System Engineer RBS Coutts

2007 – 2008 Senior System Engineer Credit Suisse

2008 Senior System Engineer Uplink AG

2008 – 2011 Senior System Engineer Swisscom Schweiz AG

2012 – heute Management Consultant LC Systems-Engineering

Anomalie-Erkennung

1.1 Zuerst würde ich gerne von Ihnen wissen, was Sie unter einer Anomalie verstehen?

Also aus meiner Sicht ist eine Anomalie ein Ereignis, bei dem ein unerwarteter Wert zurückkommt. Also wenn ich etwas in einem Bereich erwarte und ich erhalte einen Wert zurück, der außerhalb dieses Bereichs liegt. Den Bereich kann ich entweder selber statisch definieren oder dieser kann gelernt sein.

1.2 Wie lassen sich Ausreißer bzw. Anomalien erkennen?

Also ich habe einen Referenzwert oder Referenzwerte, bei denen ich weiß, dass ich diese schon gesehen habe. Wenn ich mich dann außerhalb dieses Bereichs befinde, dann triggere ich etwas, also ich vergleiche eigentlich was ich statisch definiert habe, mit dem was ich anschließend erhalte.

1.3 Welche möglichen Ursachen für das Auftreten von Anomalien gibt es?

Da gibt es ganz unterschiedliche Ursachen. Z.B. Softwarebugs, also Anwendungen, bei denen etwas schief läuft aufgrund eines Bugs, oder durch externe Einflüsse. Hier kann es sich um irgendwelche Systemausfälle, Netzwerkausfälle, Kapazitätsengpässe handeln.

1.4 Welche Herausforderungen bzw. Probleme sehen Sie in der Anomalie-Erkennung?

Ein Problem bzw. eine Herausforderung, die ich bei der Anomalie-Erkennung sehe, ist es den richtigen Algorithmus zu finden. Betrachtet man beispielsweise eine „Responsezeit“ und man erwartet hierbei eigentlich das die Anomalie nach oben geht, dass also eine hohe „Responsezeit“ vorliegt, wobei die Anomalie ja auch nach unten gehen kann, da ich zu schnelle Ergebnisse erhalte, muss dies bei der Wahl des Algorithmus beachtet werden.

Also man muss immer ein wenig aufpassen, wie man die Werte betrachtet. Also bei einer „Responsezeit“, die zwischen 10 und 100 Millisekunden liegt, erwarte ich, dass eine Anomalie vielleicht bei 120 Millisekunden liegt. Aber es kann durchaus sein, dass die Anomalie unter 10 Millisekunden auftritt, weil das System zu schnelle Antworten liefert. Dieses Beispiel zeigt, dass es immer unerwartete Dinge gibt und wenn man dann den falschen Algorithmus wählt, dann verpasst man solche Anomalien.

1.5 Können Sie mir einen konkreten Use-Case bezüglich der Anomalie-Erkennung nennen?

Ja. Wir bewegen uns häufig bei allgemeinen Metriken wie einer Anzahl, also dem zählen von Requests z.B. bei Logins im Security-Umfeld oder bei Metriken wie den „Responsezeiten“.

1.6 Haben Sie bereits einen Use-Case im Bereich der Anomalie-Erkennung umgesetzt?

Ja, also dies kommt sehr oft bei „Responsezeiten“ vor. Das kann z.B. eine Webapplikation sein, bei der man schaut, wie schnell wird die Seite geladen oder wie lange dauert der Backend Call. Dann versucht man eine Baseline zu erstellen und bei Abweichungen der Baseline zu alarmieren. Ein weiterer typischer Fall ist die Anzahl der requests. Ich weiß z.B. das pro Stunde eine bestimmte Anzahl an Seiten aufgerufen wurde und wenn dies unterschritten wird, dann stimmt etwas nicht.

1.7 Wenn ja, wie haben Sie die Anomalie-Erkennung umgesetzt und in welchen Anwendungsgebiet?

Also im Bereich Security haben wir einige Use-Cases durchgeführt, sowie im Bereich Application Perfomance Monitoring. Im Bereich Security ging es beispielsweise darum, Logins, privilegierte Zugriffe, Brute-Force Attacken oder ungewöhnliche Login-Zeiten zu Monitoren. Im Bereich Application Perfomance Monitoring ging es um Applikationen, Backend-Zugriffe, wie verhalten sich Applikationen im Frontend-Bereich z.B. wie schnell erhält ein Benutzer Antworten. Im Systembereich wird die Disk-Performance, Storage-Performance, IOPS und „Responsezeiten“ von Festplatten betrachtet. Und hier haben wir unterschiedliche Lösungen verwendet. Also man kann gewisse Dinge mit „lookup tables“ erledigen. Z.B. statische „lookup tables“ bei denen ich feste Werte eintrage und diese für einen Host, oder für eine Applikation, oder für eine URL als Schwellenwerte verwende. Zudem kann man sich auch mit einer weiteren Suche, die auf historische Werte zurückgeht, dynamische „lookup tables“ erstellen. Z.B. eine lookup table die den Durchschnitt der letzten Woche oder eines bestimmten Tages beinhaltet. Dann kann man diese lookuptable nehmen und mit aktuellen Werten vergleichen.

Das ist die einfachste Methode, um noch ein bisschen Dynamik in die Baselines oder in die Schwellenwerte zu bekommen. Eine weitere Möglichkeit bietet die Prelert App, die wir verwenden. Diese beinhaltet noch mehr Algorithmen. Man kann z.B. mehrere Metriken gleichzeitig anschauen. So kann man einen „count“ und eine „Responsezeit“ gleichzeitig mit der Baseline vergleichen. Also wenn meine „Responsezeit“ für eine

„request“ extrem hochgeht, dann kann ich unter Umständen auch weniger Anfragen beatworten. Zudem kann ich schauen geht mein „count“ runter, wenn meine „Responsezeit“ hochgeht und wenn diese beiden Metriken zusammen große Abweichungen aufzeigen, dann erhalte ich, einen Anomalyscore oder eine kleine Probabilität, die sehr klein ist. (Anomalyscore ist eine Bezeichnung von Prelert). Prelert unterstützt Algorithmen, die über mehrere Metriken schauen und überprüfen, ob ich insgesamt eine hohe Anomalie habe. Also es wird auch noch klassifiziert.

Technologien bzw. Applikationen zur Anomalie-Erkennung

2.1 Welche Technologie bzw. App bezüglich der eingesetzten Applikation „Splunk“ empfehlen Sie, um Anomalien zu erkennen?

Also es kommt immer drauf an. Splunk beinhaltet auch schon integrierte Anomalie-Kommandos. Mit der Machine-Learning-App von Splunk kommt noch mehr Funktionalität hinzu, um Anomalien zu erkennen (Algorithmen). Bei Prelert allerdings, sind die Funktionen schon breit vorhanden und die Algorithmen betrachten nicht nur eine Dimension, wie dies zum Teil bei der ML-App oder bei Splunk „out oft he box“ der Fall ist. Sondern hier hat man mehrere Dimensionen. Prelert enthält Funktionen, die Lernen, wie das Verhalten über Wochen, Tage, Stunden ist.

2.2 Haben Sie mit der genannten App bereits Erfahrungen im Umgang der Anomalie-Erkennung gesammelt?

Ja. Außerdem denke ich, dass von den Algorithmen her, die Prelert App bisher am meisten kann und wahrscheinlich die besten Algorithmen liefert.

2.3 Was für einen Use-Case haben Sie mit der App hinsichtlich der Anomalie-Erkennung behandelt?

Z.B. Security Use-Cases, bei denen wir geschaut haben, wie loggen sich User ein. Es wurde z.B. gezählt, wie oft hat sich ein User pro Tag eingeloggt. Zudem wurden logins über alle Systeme und ein zusätzliches Baselining über die Hosts betrachtet. Also ich betrachte dann die User pro Host. Wenn ein User sich an einem Host zehn Mal einloggt, dann habe ich für diesen Host eine Baseline und wenn der User sich an einem anderen Host vllt. 50-mal einloggt, dann erzeugt Prelert auch für diesen Host eine Baseline. Also pro Host eine Baseline. Dies kann dann stündlich oder täglich erfolgen. Wenn man z.B. weiß, dass am Montag der Kollege um 8 Uhr anfängt zu arbeiten, dann loggt er sich relativ viel ein und am Samstag in der Nacht hat er sich noch nie eingeloggt.

2.4 Mit welchen Mitteln haben Sie die Anomalie-Erkennung durchgeführt? (Bezogen auf die Prelert App)

Also es ist so, dass die Algorithmen proprietär sind, aber wie diese zu verwenden sind, ist gesteuert über Funktionen. Ich tue mal einen Link in den Chat-Verlauf (http://www.prelert.com/docs/splunk_app/latest/contents.html). Hier gibt es unter dem Punkt „Analysis Functions“, Funktionen, die ich verwenden kann.

2.5 Statistische Mittel oder mit Hilfe von Machine-Learning-Algorithmen?

Mit beiden Mitteln.

2.6 Wie würden Sie die Abgrenzung der Anomalie-Erkennung mit Hilfe von statistischen Mitteln von der Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen beschreiben?

Also ich kann mit Hilfe von statistischen Funktionen für Algorithmen schon sehr weit kommen. Ich kann viel „out of the box“ mit Splunk machen. Die Frage ist nur, wie kompliziert oder wie komplex wird es. Wenn ich auf das Beispiel mit der „lookup table“ eingehe, also ich generiere mir eine „lookup table“, die auch ein wenig Dynamik beinhaltet und die erstelle ich für eine Dimension und dann möchte ich noch eine zweite Dimension hinzunehmen und schön wäre auch noch einen Zeitfaktor zu verwenden. Dann wird die Suche innerhalb Splunks schon extrem komplex und da denke ich, dass wenn man eine App wie Prelert oder eine ML-App nimmt, kann das die Suche schon ziemlich vereinfachen, weil die Algorithmen generisch integriert sind, sodass man diese nur noch parametrisieren muss. Wenn ich das selbst schreibe, dann muss ich die Algorithmen selber irgendwie implementieren und die mathematischen Modelle sind dann je nachdem auch nicht ganz trivial.

Statistische Anomalie-Erkennung

3.1 Wie sind Sie zu Beginn der Datenanalyse hinsichtlich der Anomalie-Erkennung mit statistischen Mitteln vorgegangen?

Hier ist es eigentlich so, dass man bei den statistischen Mitteln sich anschaut, was für Feldwerte man hat. Außerdem schaut man mit einem „timechart“ über die Werte hinweg, um zu sehen in welchem Bereich die Metriken liegen und modelliert dann eigentlich um das Ergebnis herum. Problematisch ist es dann, wenn man einen „timechart“ mit einem „average“ über eine „Responsezeit“ pro host macht und sich denkt, dass man schon irgendeinen Wertebereich erhalten wird. Hier besteht die Gefahr, dass ich bei einem „average“ die Outlier nicht sehe. Da muss man dann vielleicht noch ein „min“ und „max“ hinzunehmen und dann den Zeitbereich ein wenig einschränken, um die Ausreißer nicht aus der Suche zu entfernen. Man nähert sich also an. Zudem kann die Wahl einer Auflösung bei einem „timechart“ problematisch sein. Wenn ich bei einem „timechart“ die Auflösung von einem Tag wähle und habe dort irgendwo einen kleinen Ausreißer in den Daten, dann sehe ich diesen nicht.

3.2 Mussten die Daten zuvor aufbereitet werden?

Ja. Grundsätzlich muss man bei Splunk immer schauen, dass man die Metriken sauber als Felder extrahiert. Also mit Regulären Ausdrücken, die richtigen Zahlen in das richtige Feld extrahiert. Zudem muss man evtl. noch Einheiten konvertieren, z.B. bei einer Mischung aus Milli- und Nanosekunden.

3.3 Wenn ja, wie haben Sie diese Datenaufbereitung bzw. Datenvorbereitung durchgeführt?

Siehe Begründung von der vorherigen Antwort!

3.4 Welche Statistischen Mittel bzw. Kommandos haben Sie zur Erkennung von Anomalien verwendet?

„Avg“, „min“, „max“.

3.5 Können Sie mir evtl. eine Möglichkeit nennen, mit der ich ein vereinfachtes Testszenario zur Erkennung von Anomalien mit Hilfe von statistischen Mitteln innerhalb der Splunk-Umgebung darstellen kann?

Ich habe in einer „lookup table“ meinen Referenzwert oder vielleicht mehrere Referenzwerte, vielleicht „min“ und „max“ und dann die Werte von weiteren Events, die ich jetzt mit den Werten der „lookup table“ und einem Eval-Kommando vergleiche. Das Eval-Kommando zeigt mir dann an, ob ich drüber oder drunter oder drinne liege. Mit dem

Eval-Kommando wird also eine Variable erzeugt, die eine Untere und eine obere Grenze darstellt.

Anomalie-Erkennung mit Hilfe von Machine-Learning

4.1 Wie sind Sie zu Beginn der Datenanalyse hinsichtlich der Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen grundsätzlich vorgegangen?

Also wenn man z.B. Prelert verwendet, dann ist es so, das Prelert verschiedene Funktionsmodi enthält. Da gibt es eine Ad-hoc Suchmöglichkeit. Also ich lade mir z.B. den letzten Tag in Prelert rein und wende meinen Algorithmus auf diesen Tag an, dann kann ich schon Ad-hoc sehen, was für Anomalien evtl. auftreten. Das ist eine Variante.

Wenn ich dann der Meinung bin, ich habe die richtige Funktion verwendet, dann muss man in Prelert eine Realtime-Suche einrichten, die z.B. fortlaufend die Baseline erstellt. Also ich konfiguriere Prelert so, dass ich alle 5 Minuten oder jede Stunde die aktuellen Werte in einer Baseline hinterlege und dann werden die aktuellen Werte mit der Baseline verglichen.

Also es gibt verschiedene Betriebsmodi in Prelert, das Ad-hoc, mit dem ich schnell etwas reinladen und auch eine Baseline einrichten kann. Diese Baseline wird dann auch konstant angepasst. Das ist vor allem dann wichtig, wenn ich zeitliche Baselines habe. Also wenn wir von Montag bis Sonntag an jedem einzelnen Tag ein bestimmtes Verhalten haben, dann benötigt Prelert eine Learningphase. Da ist es am besten, wenn man zwei bis drei Wochen an Daten hat, dann kann Prelert daraus eine Baseline erstellen, die dann wesentlich exakter ist als wenn ich nur eine Woche an Daten verwende.

4.2 Mit welcher Applikation haben Sie innerhalb von Splunk die Anomalie-Erkennung durchgeführt? ML-App? Prelert?

Prelert. An dieser Stelle aber noch eine kleine Ergänzung. Splunk bietet zudem die Premium-Lösung IT-SI (IT-Service-Intelligence) an. Diese beinhaltet auch das Baselining von KPI‘s und beherrscht ein wenig die Anomalie-Erkennung. Diese habe ich aber noch nicht ausprobiert, daher kann ich nicht beurteilen, wie gut die Algorithmen sind.

4.3 Welche Algorithmen unterstützt die jeweilige Applikation?

Die Algorithmen von Prelert sind nicht öffentlich bekannt. Aber die Algorithmen sind vor allem für IT Operations Use-Cases oder Security Use-Cases geeignet.

4.4 Mussten die Daten zuvor aufbereitet werden?

Ja. Analog zu der statistischen Anomalie-Erkennung. Also ich habe Feldextraktionen und die Feldwerte müssen korrekte Werte beinhalten, weil ich Prelert nur diese Feldnamen mitteile und aufgrund des Feldnamens berechnet Prelert dann die Anomalie.

4.5 Wenn ja, wie haben Sie diese Datenaufbereitung bzw. Datenvorbereitung durchgeführt?

Begründung siehe eine Frage vorher (4.4)

4.6 Welche Algorithmen empfehlen Sie zur Erkennung von Anomalien?

Ausgelassen, da hinsichtlich der Prelert App die genauen Algorithmen nicht bekannt sind.

4.7 Welche Machine-Learning-Algorithmen haben Sie zur Erkennung von Anomalien verwendet?

Ausgelassen, da diese in Prelert nicht veröffentlicht sind.

4.8 Anhand welcher Kriterien haben Sie den Algorithmus ausgewählt?

Eigentlich würde ich ohne sofort mit Prelert zu beginnen, erstmal grundsätzlich über die vorliegenden Werte schauen, um ein Gefühl zu bekommen, was diese überhaupt Aussagen. Danach würde ich dann den Ad-hoc-Modus von Prelert verwenden. Also ich würde ein paar der Funktionen die Prelert bietet anwenden und schauen, ob ich da ggf. schon Anomalien erkenne. Bei Prelert gibt es zudem auf der Seite Use-Cases, bei denen aufgezeigt wird, für welche Felder man welche Funktion verwenden sollte oder könnte. Daran kann man sich ein wenig orientieren, weil häufig betrachtet man bei sich keinen Sonderfall. Also wenn es um Metriken wie „requests“ geht.

4.9 Haben Sie mit einem der Algorithmen das Verhalten trainiert und dieses dann auf einen Testdatensatz angewendet? (ist dies bei Prelert auch möglich)?

Also bei Prelert ist das Vorgehen so, dass ich die bestehenden Daten, also die die ich schon indiziert habe, eigentlich auch als Trainingsdaten verwende. Ich möchte ja mit den Produktionsdaten arbeiten und trainieren. Also ich kann bei Prelert mit der Realtime Funktion die Baseline ja konstant aktualisieren. Also im Prinzip erstelle ich mir eine Baseline aus den letzten 30 Tagen, dann schaut sich Prelert die historischen Daten an und trainiert mit Hilfe dieser Daten. Ich denke das macht auch Sinn. Ich würde jetzt keine Daten von einem Testsystem verwenden und mit diesen Trainieren, weil diese ja nicht der Produktion entsprechen.

4.10 Wie haben Sie das Trainieren der Daten durchgeführt?

Ein Spliting in Trainings- und Testdaten wird bei der Prelert App indirekt vorgenommen, indem anhand der Durchsuchten Daten eine Baseline erzeugt wird und diese ein Verhalten mit Hilfe von Algorithmen lernt, welches auf einen anderen Satz an Daten angewendet wird. Mit einem anderen Satz an Daten ist einfach nur ein anderer Zeitraum gemeint.

4.11 Wie kann das durchs trainieren gelernte Verhalten angewendet werden?

Das gelernte kann nur auf den konkreten Fall angewendet werden. Also man muss sich das so vorstellen: Wenn ich jetzt das Beispiel von den Webservern mit den „Responsezeiten“ oder der Anzahl an „requests“ nehme, dann sage ich hier eigentlich schon für das Baselining, ich suche nur in einem bestimmten Index über die hosts und Metriken und dann wird nur für diese Hosts das Training durchgeführt. Also er lernt Host x, y oder z hat sich so verhalten. Wenn jetzt ein neuer Host dazu kommt, dann wird für diesen eine neue Baseline erzeugt. Also es wird nicht zwingend betrachtet, was die anderen Hosts gemacht haben, außer wenn ich einen populationsbasierten Algorithmus gewählt habe. Dann würde es so sein, dass ich den neuen Host mit anderen Hosts vergleiche.

4.12 Auf welche Bereiche ist das gelernte anwendbar?

Siehe vorherige Antwort. (Nur auf den konkreten Fall anwendbar)

Bevor wir das Interview abschließen, möchte Ich von Ihnen gerne wissen, ob aus Ihrer Sicht eine wichtige Frage ungestellt blieb? Ist Ihnen während des Interviews z.B. irgendein offener Punkt aufgefallen, den ich noch mit beachten sollte?

Ja was ich noch hinzufügen möchte ist, das ich mit der Machine-Learning App oder mit der Prelert App jetzt sehr mächtige Algorithmen habe, aber das Problem ist, auch die Algorithmen sind nicht perfekt. Sobald ich es irgendwie mit Usern zu tun habe (vllt. im Security-Bereich), da habe ich immer etwas Unerwartetes. Oder ich habe einen User der 3 Wochen Urlaub hat oder für ein halbes Jahr ins Sabbatical geht und dann zurück kommt, dann ist meine Baseline für diesen User nicht mehr brauchbar. Oder man betrachtet Feiertage wie Weihnachten und es finden Sonderaktionen wie Rabattaktionen statt. Das sind Ausnahmesituationen, bei denen z.B. auch auf einer Webseite mehr Aufrufe erfolgen. Dies ist eigentlich eine Anomalie, aber keine richtige. Das ist eine gewollte Anomalie. Solche Ausnahmen sind extrem schwierig zu modellieren. Also die Anomalydetection und das Machine-Learning verringern womöglich die Anzahl an

false/positives oder false/negatives, aber man eliminiert sie nicht. Also man wird immer mal wieder einen Ausreißer haben der trotzdem normal ist.

Und noch etwas was ich erwähnen möchte, ich beschäftige mich ja auch nicht jeden Tag mit der Anomalie-Erkennung. Man muss sich mit der Thematik schon sehr auseinandersetzen, da man nicht einfach einen „roten"-Knopf drücken kann und dann läuft die Anomalydetection. Man muss sich konstant damit beschäftigen, also mit den Algorithmen.

Zusätzliche Fragen:

Und können Sie mir evtl. noch Referenzkunden nennen, die im Bereich der Anomalie-Erkennung Erfahrungen gesammelt haben, um ein weiteres Experteninterview zu führen?

Das müsste ich abklären. Da würde ich dann auf Dich zu kommen. Aber allzu viele gibt es eigentlich nicht, dass muss man auch sagen. Ich denke das ganze Thema ist doch noch relativ frisch.

Dank: Zum Abschluss möchte ich mich ganz herzlich für die Zeit, die Sie sich genommen haben bedanken!

Anhang 3: Experteninterview: Unternehmen LC Systems

Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Zusammenfassung des Experteninterviews

„Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen/Technologien“

Niklas Netz
Telefon: +49 (0) 40 64611744
Mobil: +49 (0) 171530512
Mail: niklas.netz@haw-hamburg.de oder niklas.netz@otto.de

Fakultät: Technik und Informatik
Department: Informatik
Hochschule für Angewandte Wissenschaften
Studiengang: Bachelor of Science Wirtschaftsinformatik

Datum: 05.04.2016 Ort: Hamburg

Angaben zur Person: Christian Günther

Position/Funktion: Leiter Data Analytic

Akademische Laufbahn (kurz):

Berufliche Laufbahn (kurz):

Christian Günther ist seit 2000 bei LC Systems verantwortlich für den Dienstleistungsbereich und dessen Leistungserbringung. In seiner Funktion erfüllt er zusätzlich Management Consulting-Aufträge und ist in Großprojekten beratend und begleitend aktiv und hat den Data Analytics Bereich bei LC Systems erfolgreich aufgebaut. Er bewegt sich seit 25 Jahren auf der Dienstleistungsseite im Data Center-Umfeld und betreut Großkunden aus den Bereichen Pharma, Banken und Versicherungen, Bund- und Militär sowie Telco-Unternehmen.

Die folgende Zusammenfassung, fasst die gewonnenen Erkenntnisse des Gesprächs vom 05.04.2016 zusammen.

Zusammenfassung der gewonnenen Erkenntnisse

Aus dem Gespräch ist hervorgegangen, dass der Begriff der Anomalie in der Wissenschaft häufig eine andere Bedeutung im Vergleich zur Praxis aufweist.

Dies ist deutlich geworden, da wir im geführten Gespräch immer wieder auf die Bedeutung bzw. die Definition der Anomalie eingegangen sind, um ein Verständnis dafür zu entwickeln „was eigentlich genau unter einer Anomalie zu verstehen ist“.

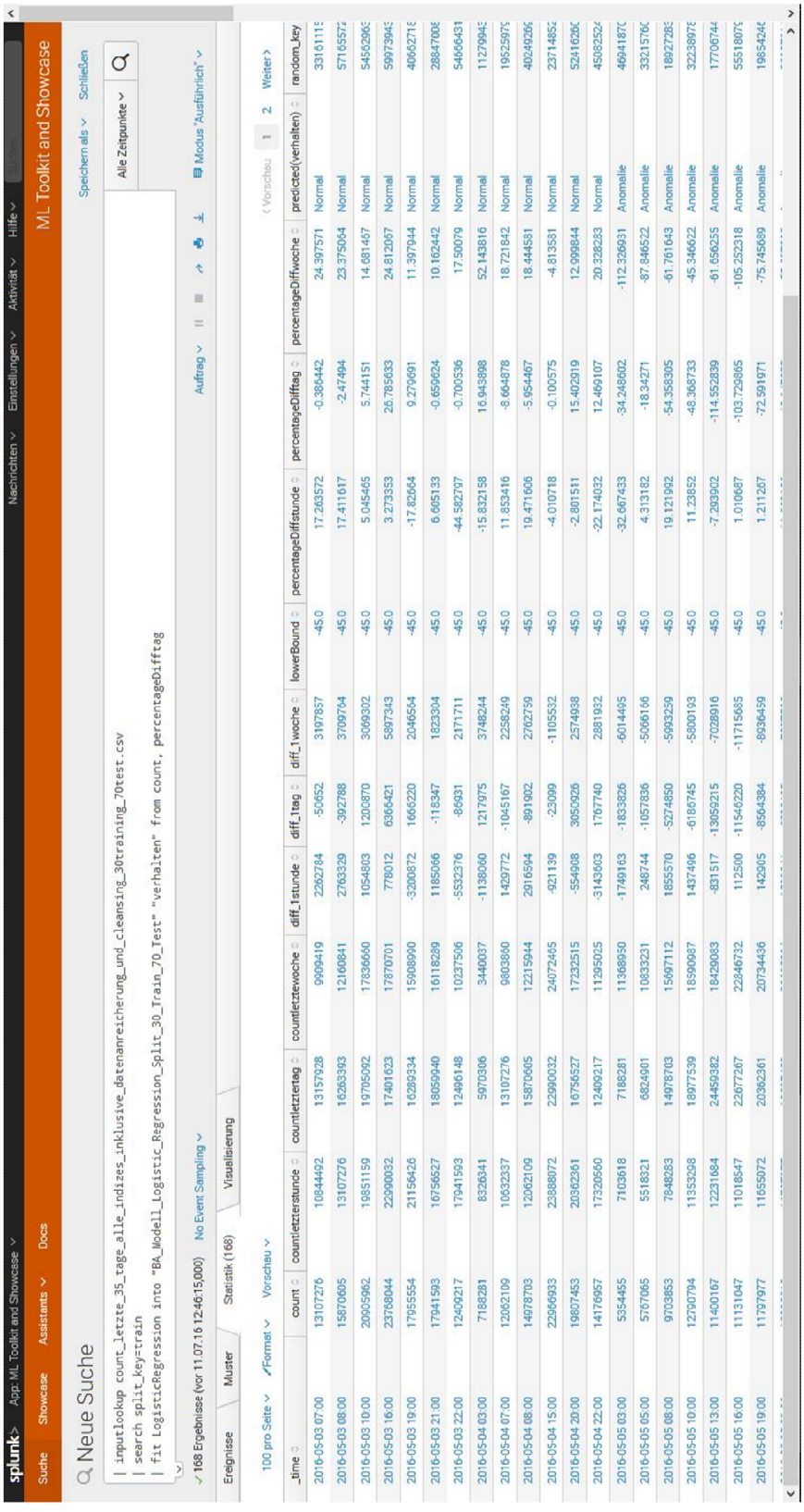
Zudem ist klar geworden, dass in der Praxis nur sehr wenige Kunden (die Rede war von fünf Kunden), die Anomalie-Erkennung mit Hilfe von Algorithmen betreiben und falls dies der Fall ist, dann derzeit lediglich im Testbetrieb.

Das Interesse der Anomalie-Erkennung besteht zwar bei vielen Unternehmen, allerdings sind viele Unternehmen gerade erst dabei, die eigenen Daten zu verstehen, die Daten aufzubereiten, um dann eine Anomalie-Erkennung mit Hilfe von Algorithmen durchführen zu können.

Die Anomalie-Erkennung befindet sich im Praxisbereich also noch in den „Kinderschuhen“. Im Bereich der Forschung tut sich zwar einiges hinsichtlich der Anomalie-Erkennung, aber die Transformation in die Praxis ist noch nicht einfach realisierbar.

Außerdem hat sich aus dem Gespräch herauskristallisiert, dass die Erkennung von Anomalien in den großen Mengen an Daten ein komplexes Themengebiet aufweist.

Anhang 4: Anwendung des Algorithmus der Logistischen Regression (LR)



splunk> App: ML Toolkit and Showcase

Suche | Showcase | Assistants | Docs | Nachrichten | Einstellungen | Aktivität | Hilfe | Suchen

ML Toolkit and Showcase

Neue Suche

Speichern als | Schließen

```
| inputlookup count_letzte_35_tage_alle_indizes_inklusive_datenanreicherung_und_cleansing_30training_70test.csv
| search split_key=train
| fit LogisticRegression into "BA_Modell_Logistic_Regression_Split_30_Train_70_Test" "verhalten" from count, percentageDiffTag
```

Alle Zeitpunkte

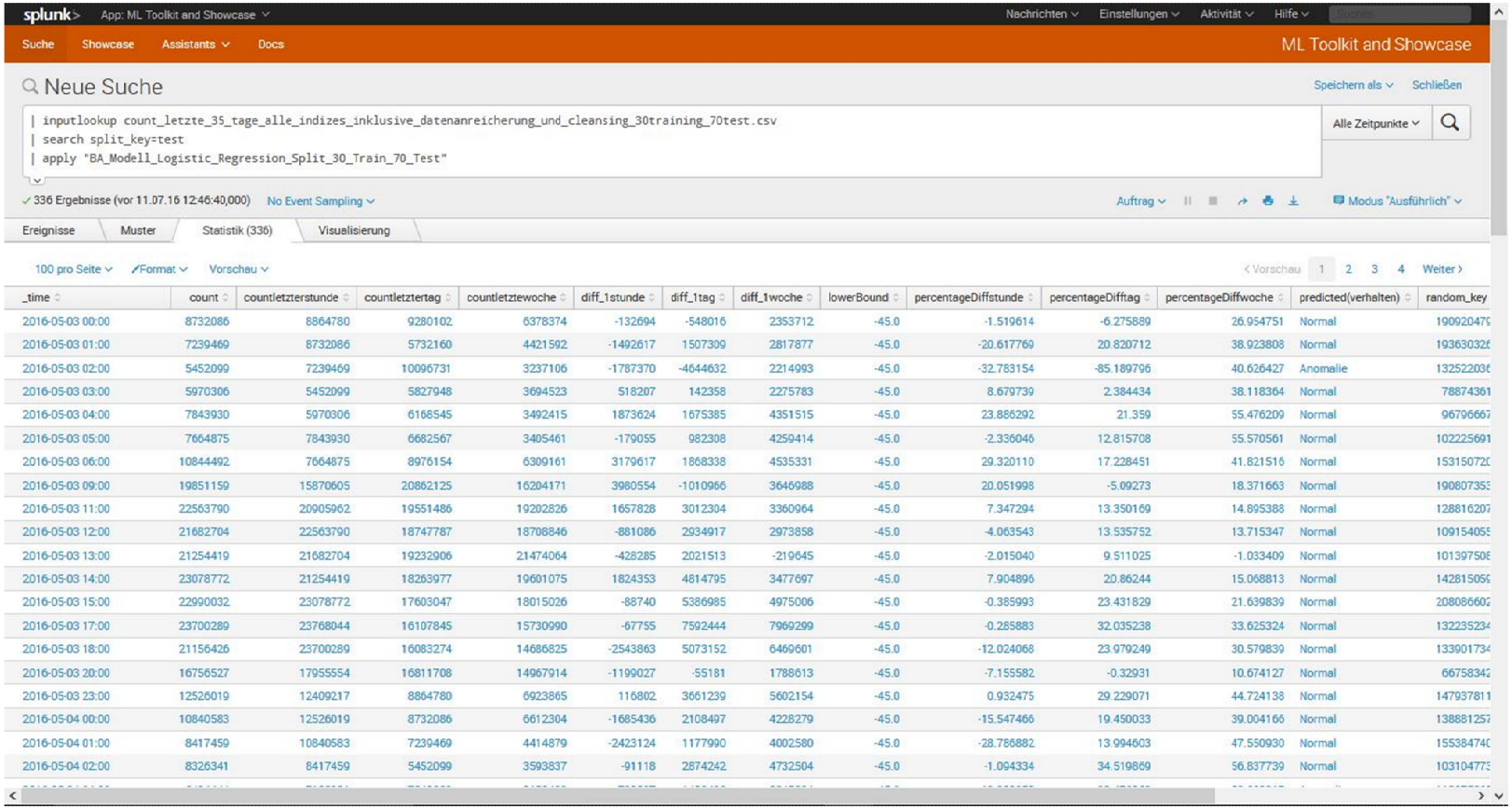
✓ 168 Ergebnisse (vor 11.07.16 12:46:15,000) | No Event Sampling | Auftrag | Modus: Ausführlich

Ereignisse | Muster | Statistik (168) | Visualisierung

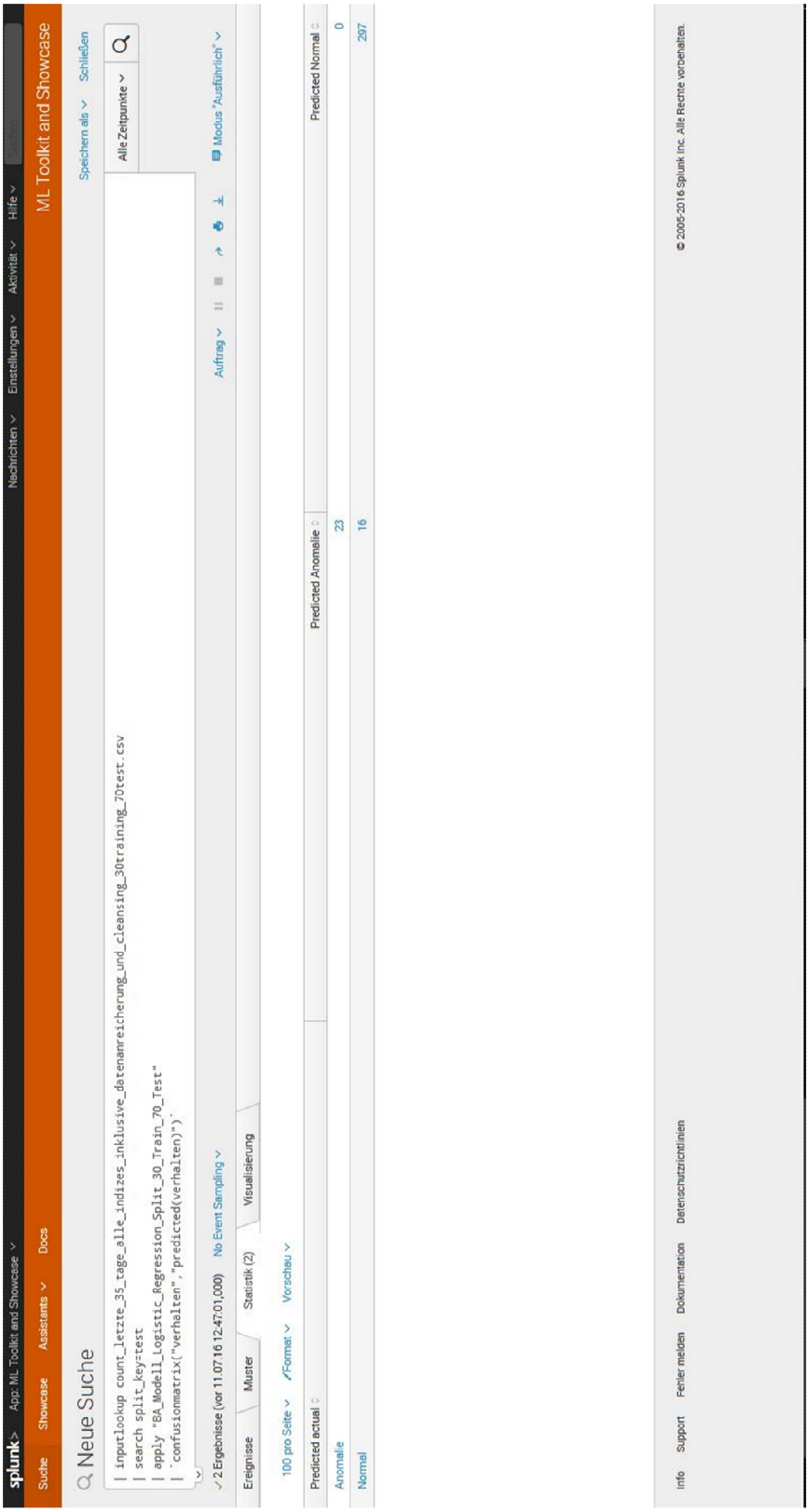
100 pro Seite | Format | Vorschau | ‹ Vorschau 1 2 Weiter ›

_time	count	countletzterstunde	countletztertag	countletzterwoche	diff_1sturde	diff_1tag	diff_1woche	lowerBound	percentageDiffstunde	percentageDiffTag	percentageDiffwoche	predicted(verhalten)	random_key
2016-05-03 07:00	13107276	10844492	13157928	9909419	2262784	-50652	3197857	-45.0	17.263572	-0.386442	24.397571	Normal	3316111
2016-05-03 08:00	15870605	13107276	16263393	12160841	2763329	-392788	3709764	-45.0	17.411617	-2.47494	23.375064	Normal	5716557
2016-05-03 09:00	20905962	19851159	19705092	17836660	1054803	1200870	3069302	-45.0	5.045465	5.744151	14.681467	Normal	5456206
2016-05-03 10:00	23768044	22990032	17401623	17870701	778012	6366421	5897343	-45.0	3.273353	26.785633	24.812067	Normal	5997394
2016-05-03 19:00	17955554	21156426	16289334	15908990	-3200672	1666220	2046564	-45.0	-17.82664	9.279691	11.397944	Normal	4066271
2016-05-03 21:00	17941593	16756527	18059040	16118289	1185066	-118347	1823304	-45.0	6.605133	-0.659624	10.162442	Normal	2884700
2016-05-03 22:00	12409217	17941593	12496148	10237506	-5532376	-86931	2171711	-45.0	-44.582797	-0.700536	17.500079	Normal	5466643
2016-05-04 03:00	7188281	8326341	5970306	3440037	-1138060	1217975	3748244	-45.0	-15.832158	16.943898	52.143816	Normal	1127994
2016-05-04 07:00	12062109	10632337	13107276	9803860	1429772	-1045167	2258249	-45.0	11.853416	-8.664878	18.721842	Normal	1952597
2016-05-04 08:00	14978703	12062109	15870605	12215944	2916594	-891902	2762759	-45.0	19.471606	-5.954467	18.444581	Normal	4024926
2016-05-04 15:00	22966933	23888072	22990032	24072465	-921139	-23099	-1105532	-45.0	-4.010718	-0.100575	-4.813581	Normal	2371485
2016-05-04 20:00	19807453	20362361	16756527	17232515	-554908	3050926	2574938	-45.0	-2.801511	15.402919	12.999844	Normal	5241626
2016-05-04 22:00	14176957	17320560	12409217	11295025	-3143603	1767740	2881932	-45.0	-22.174032	12.469107	20.328283	Normal	4508252
2016-05-05 03:00	5354455	7103618	7188281	11368950	-1749163	-1833826	-6014495	-45.0	-32.667433	-34.248602	-112.326931	Anomalie	4694187
2016-05-05 05:00	5767065	5518321	6824901	10833231	248744	-1057836	-5066166	-45.0	4.313182	-18.34271	-87.846522	Anomalie	3321576
2016-05-05 08:00	9703853	7848283	14978703	15697112	1855570	-5274850	-5993259	-45.0	19.121992	-54.358305	-61.761643	Anomalie	1892728
2016-05-05 10:00	12790794	11353298	18977539	18590987	1437496	-6186745	-5800193	-45.0	11.23852	-48.368733	-45.346622	Anomalie	3223897
2016-05-05 13:00	11400167	12231684	24459382	18429083	-831517	-13059215	-7028916	-45.0	-7.293902	-114.552839	-61.656255	Anomalie	1770674
2016-05-05 16:00	11131047	11018547	22677267	22846732	112500	-11546220	-11715685	-45.0	1.010687	-103.729865	-105.252318	Anomalie	5551807
2016-05-05 19:00	11797977	11655072	20362361	20734436	142905	-8564384	-8936459	-45.0	1.211267	-72.591971	-75.745689	Anomalie	1985424

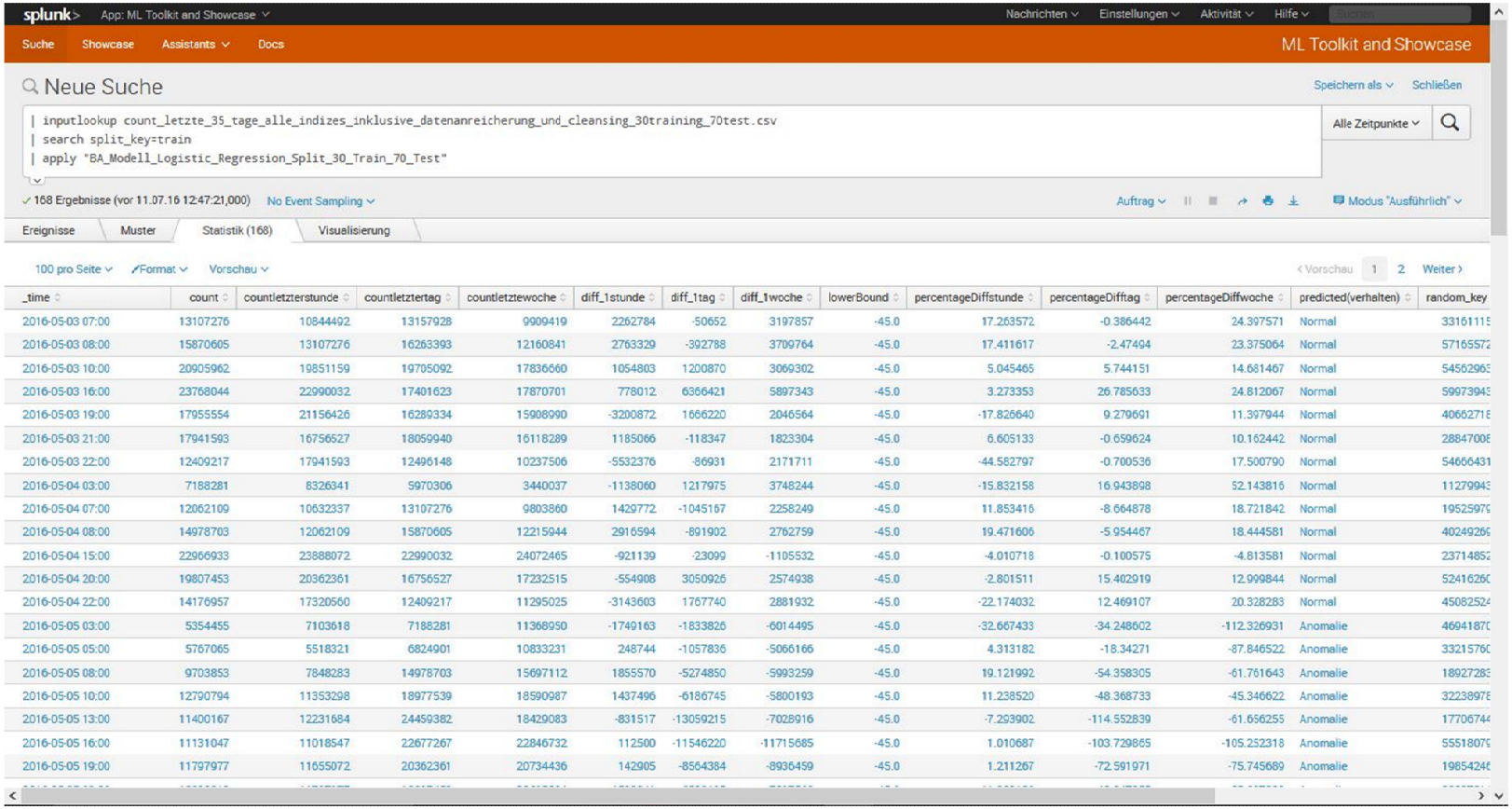
A1 Training des Algorithmus der Logistischen Regression



A2 Anwendung des trainierten Models (LR) auf die Testdaten

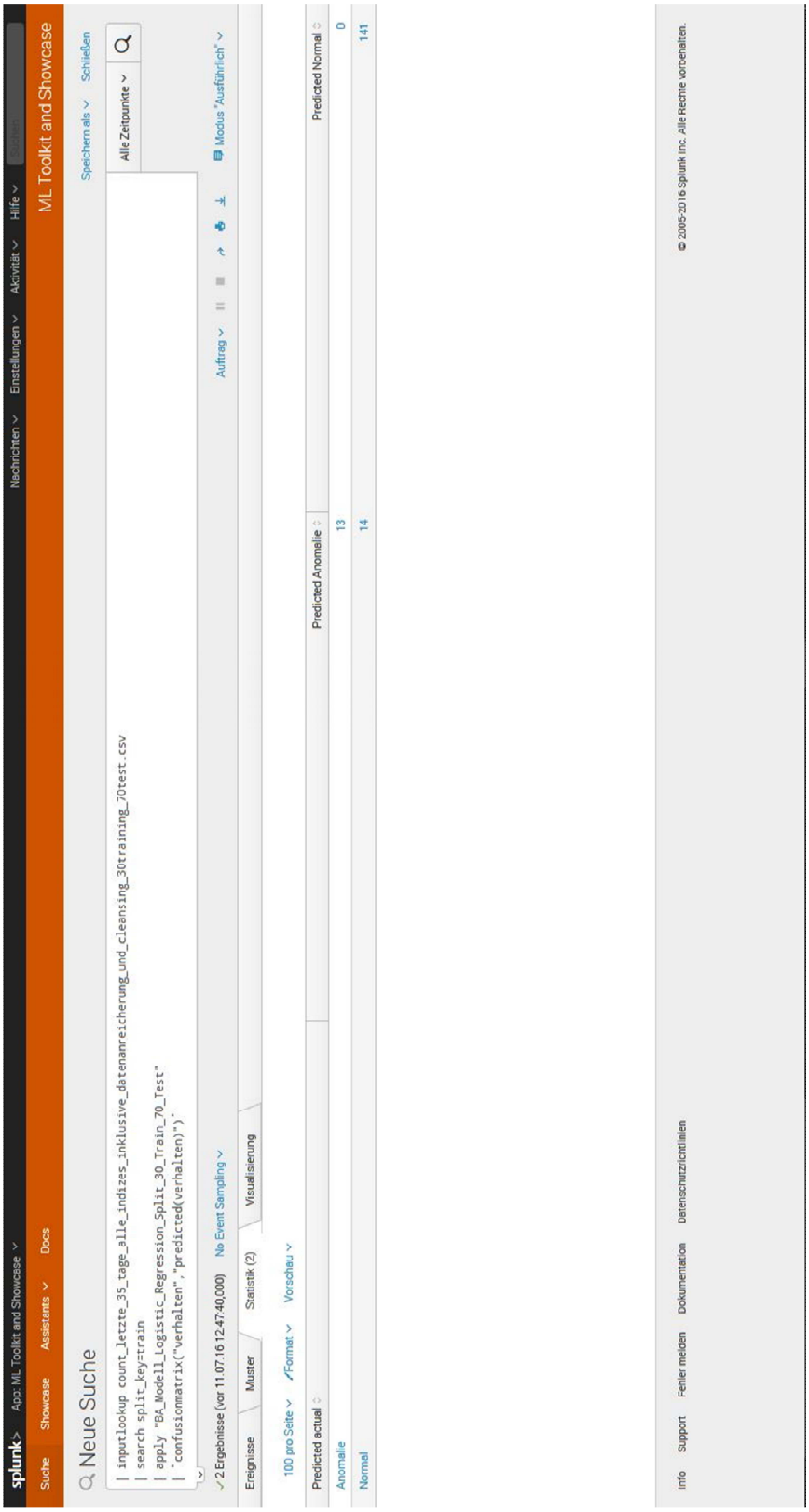


A3 Konfusionsmatrix des angewendeten Modells (LR) auf die Testdaten



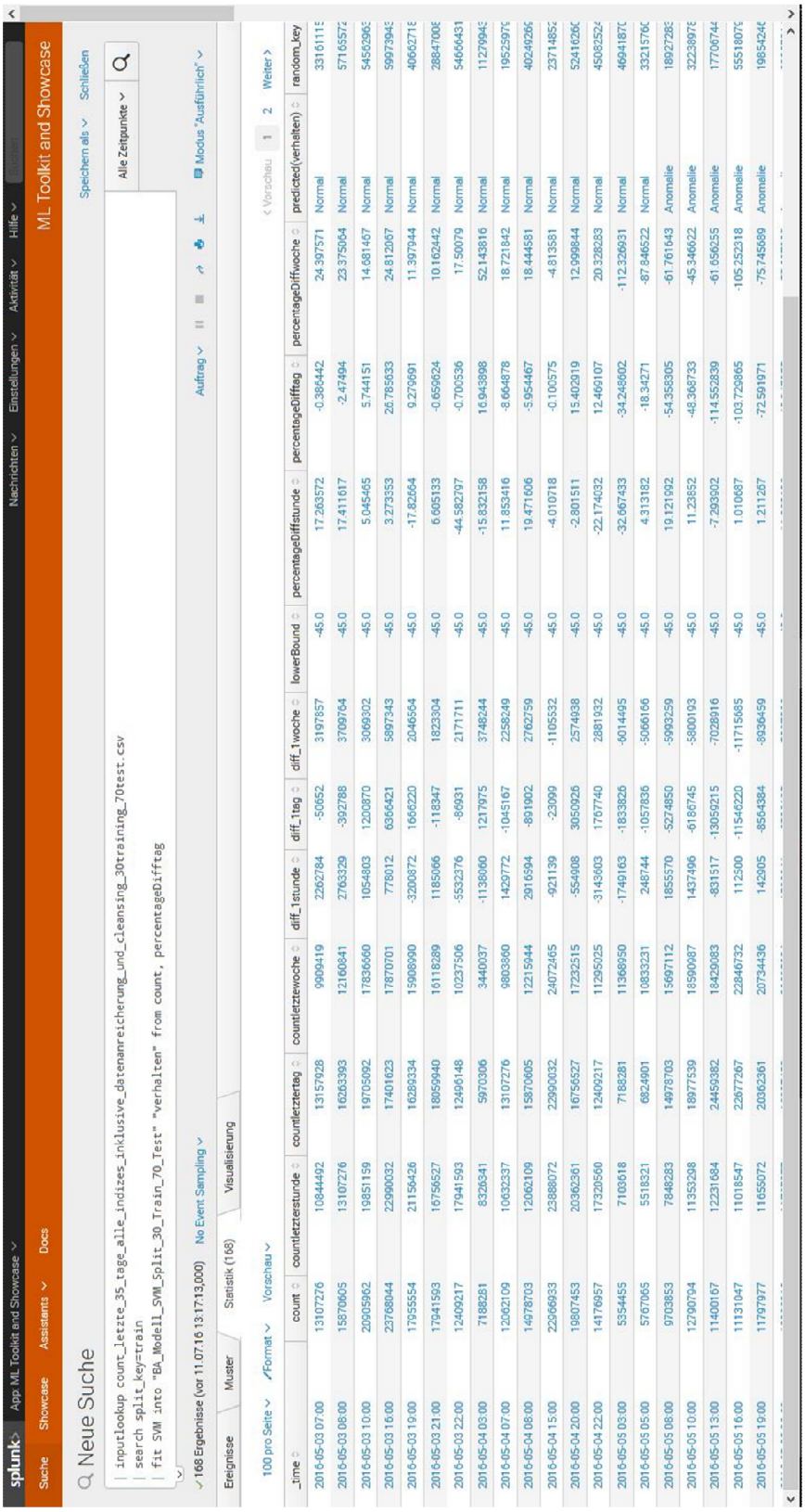
_time	count	countletzterstunde	countletztertag	countletztewoche	diff_1stunde	diff_1tag	diff_1woche	lowerBound	percentageDiffstunde	percentageDifftag	percentageDiffwoche	predicted(verhalten)	random_key
2016-05-03 07:00	13107276	10844492	13157928	9909419	2262784	-50652	3197857	-45.0	17.263572	-0.386442	24.397571	Normal	3316111
2016-05-03 08:00	15870605	13107276	16263393	12160841	2763329	-392788	3709764	-45.0	17.411617	-2.47494	23.375064	Normal	5716557
2016-05-03 10:00	20905962	19851159	19705092	17836660	1054803	1200870	3069302	-45.0	5.045465	5.744151	14.681467	Normal	5456296
2016-05-03 16:00	23768044	22990032	17401623	17870701	778012	6366421	5897343	-45.0	3.273353	26.785633	24.812067	Normal	5997394
2016-05-03 19:00	17955554	21156426	16289334	15908990	-3200872	1666220	2046564	-45.0	-17.826640	9.279691	11.397944	Normal	4066271
2016-05-03 21:00	17941593	16756527	18059940	16118289	1185066	-118347	1823304	-45.0	6.605133	-0.659624	10.162442	Normal	2884700
2016-05-03 22:00	12409217	17941593	12496148	10237506	-5532376	-86931	2171711	-45.0	-44.582797	-0.700536	17.500790	Normal	5466643
2016-05-04 03:00	7188281	8326341	5970306	3440037	-1138060	1217975	3748244	-45.0	-15.832158	16.943898	52.143816	Normal	1127994
2016-05-04 07:00	12062109	10632337	13107276	9803860	1429772	-1045167	2258249	-45.0	11.853416	-8.664878	18.721842	Normal	1952597
2016-05-04 08:00	14978703	12062109	15870605	12215944	2916594	-891902	2762759	-45.0	19.471606	-5.954467	18.444581	Normal	4024926
2016-05-04 15:00	22966933	23888072	22990032	24072465	-921139	-23099	-1105532	-45.0	-4.010718	-0.100575	-4.813581	Normal	2371485
2016-05-04 20:00	19807453	20362361	16756527	17232515	-554908	3050926	2574938	-45.0	-2.801511	15.402919	12.999844	Normal	5241626
2016-05-04 22:00	14176957	17320560	12409217	11295025	-3143603	1767740	2881932	-45.0	-22.174032	12.469107	20.328283	Normal	4508252
2016-05-05 03:00	5354455	7103618	7188281	11368950	-1749163	-1833826	-6014495	-45.0	-32.667433	-34.248602	-112.326931	Anomalie	4694187
2016-05-05 05:00	5767065	5518321	6824901	10833231	248744	-1057836	-5066166	-45.0	4.313182	-18.34271	-87.846522	Anomalie	3321576
2016-05-05 08:00	9703853	7848283	14978703	15697112	1855570	-5274850	-5993259	-45.0	19.121992	-54.358305	-61.761643	Anomalie	1892728
2016-05-05 10:00	12790794	11353298	18977539	18590987	1437496	-6186745	-5800193	-45.0	11.238520	-48.368733	-45.346622	Anomalie	3223897
2016-05-05 13:00	11400167	12231684	24459382	18429083	-831517	-13059215	-7028916	-45.0	-7.293902	-114.552839	-61.656255	Anomalie	1770674
2016-05-05 16:00	11131047	11018547	22677267	22846732	112500	-11546220	-11715685	-45.0	1.010687	-103.729865	-105.252318	Anomalie	5551807
2016-05-05 19:00	11797977	11655072	20362361	20734436	142905	-8564384	-8936459	-45.0	1.211267	-72.591971	-75.745689	Anomalie	1985424

A4 Anwendung des trainierten Modells (LR) auf die Trainingsdaten



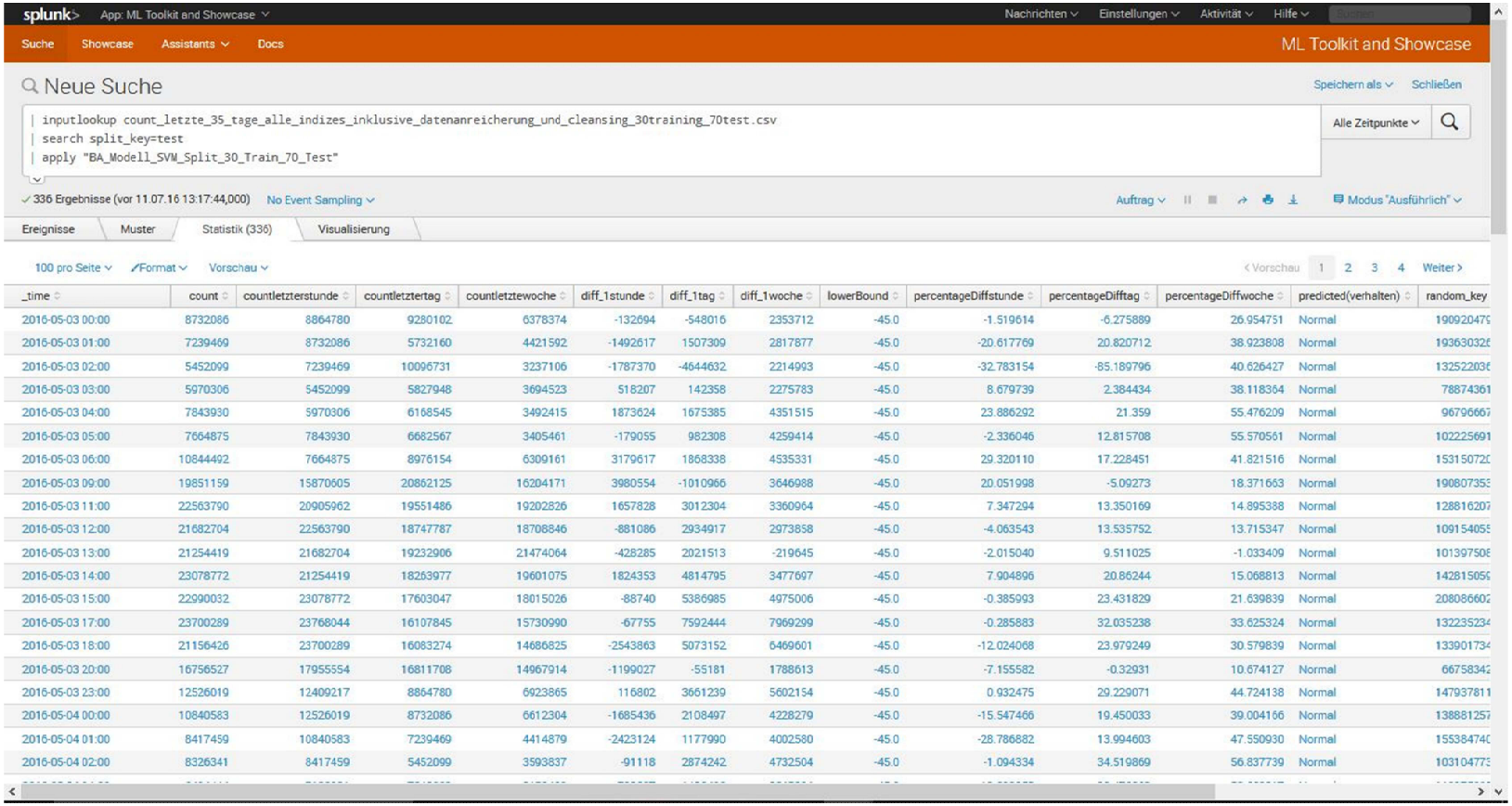
A5 Konfusionsmatrix des angewendeten Modells (LR) auf die Trainingsdaten

Anhang 5: Anwendung des Algorithmus Support Vector Machine (SVM)



A6 Training des Algorithmus Support Vector Machine

A7 Anwendung des trainierten Modells (SVM) auf die Testdaten



splunk> App: ML Toolkit and Showcase — Nachrichten — Einstellungen — Aktivität — Hilfe

Suche | Showcase | Assistants | Docs — ML Toolkit and Showcase

Neue Suche — Speichern als — Schließen

```
| inputlookup count_letzte_35_tage_alle_indizes_inklusive_datenanreicherung_und_cleansing_30training_70test.csv
| search split_key=test
| apply "BA_Modell_SVM_Split_30_Train_70_Test"
```

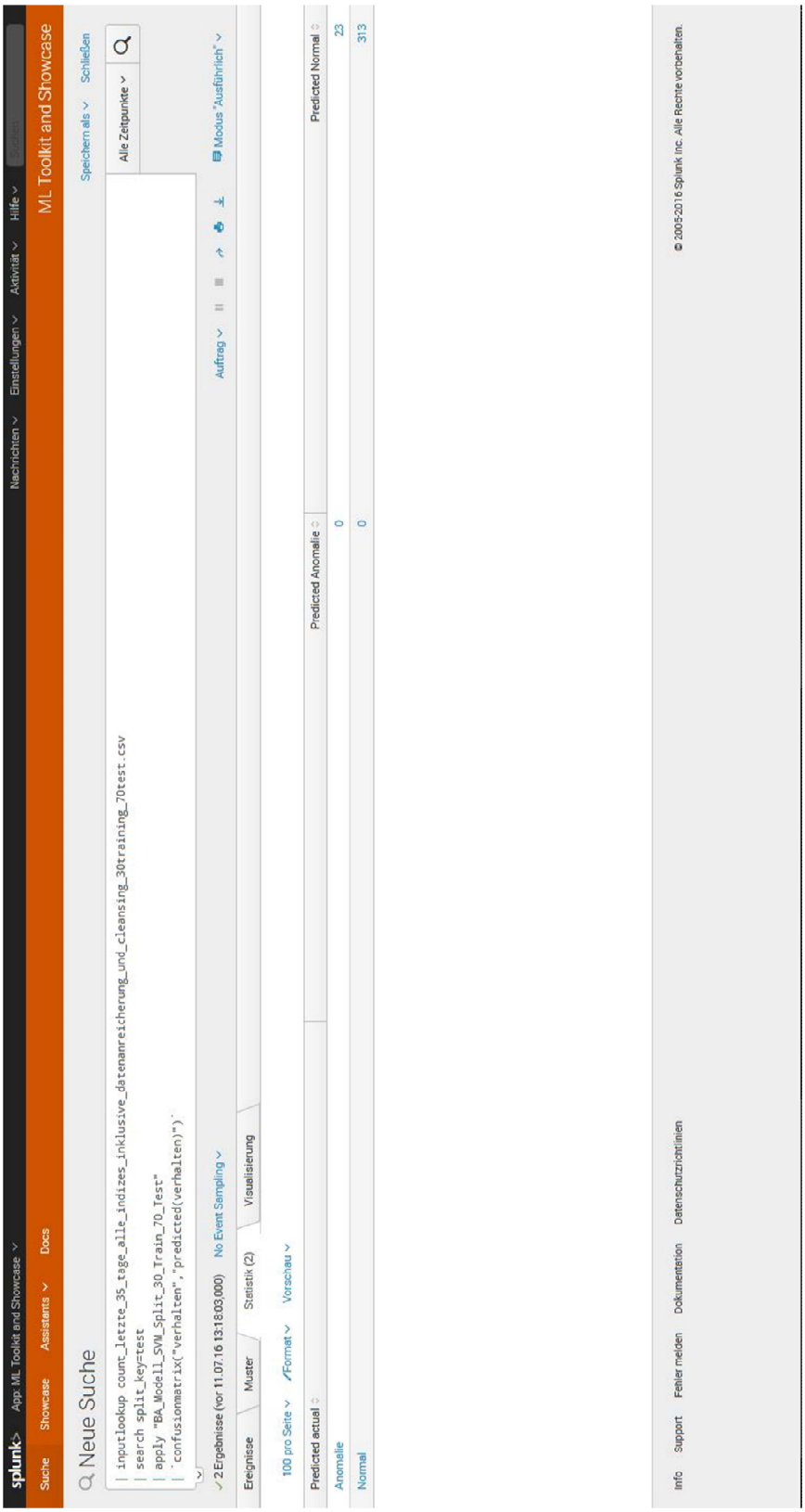
Alle Zeitpunkte

✓ 336 Ergebnisse (vor 11.07.16 13:17:44,000) — No Event Sampling — Auftrag — Modus "Ausführlich"

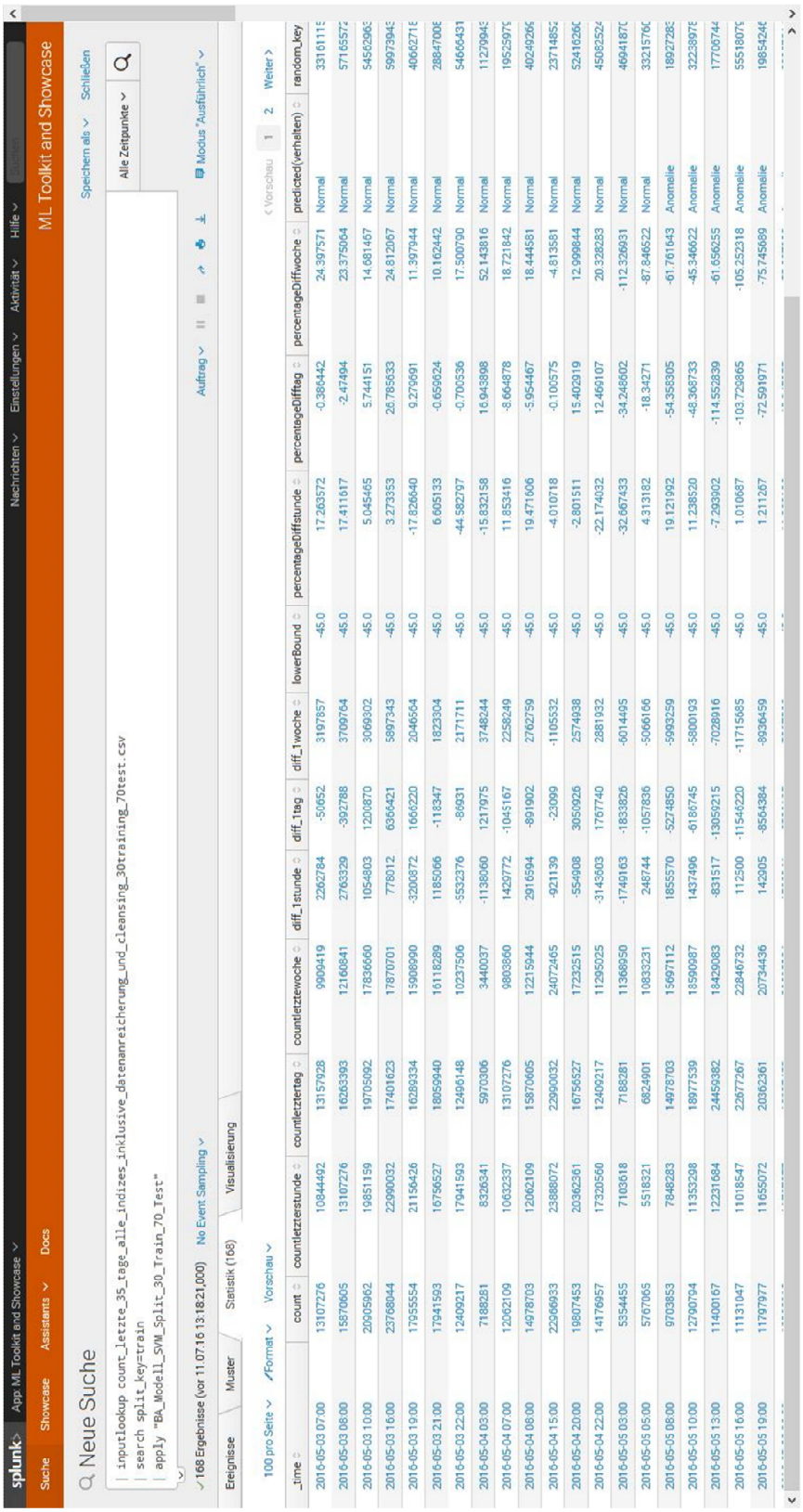
Ereignisse | Muster | Statistik (336) | Visualisierung

100 pro Seite — Format — Vorschau — < Vorschau 1 2 3 4 Weiter >

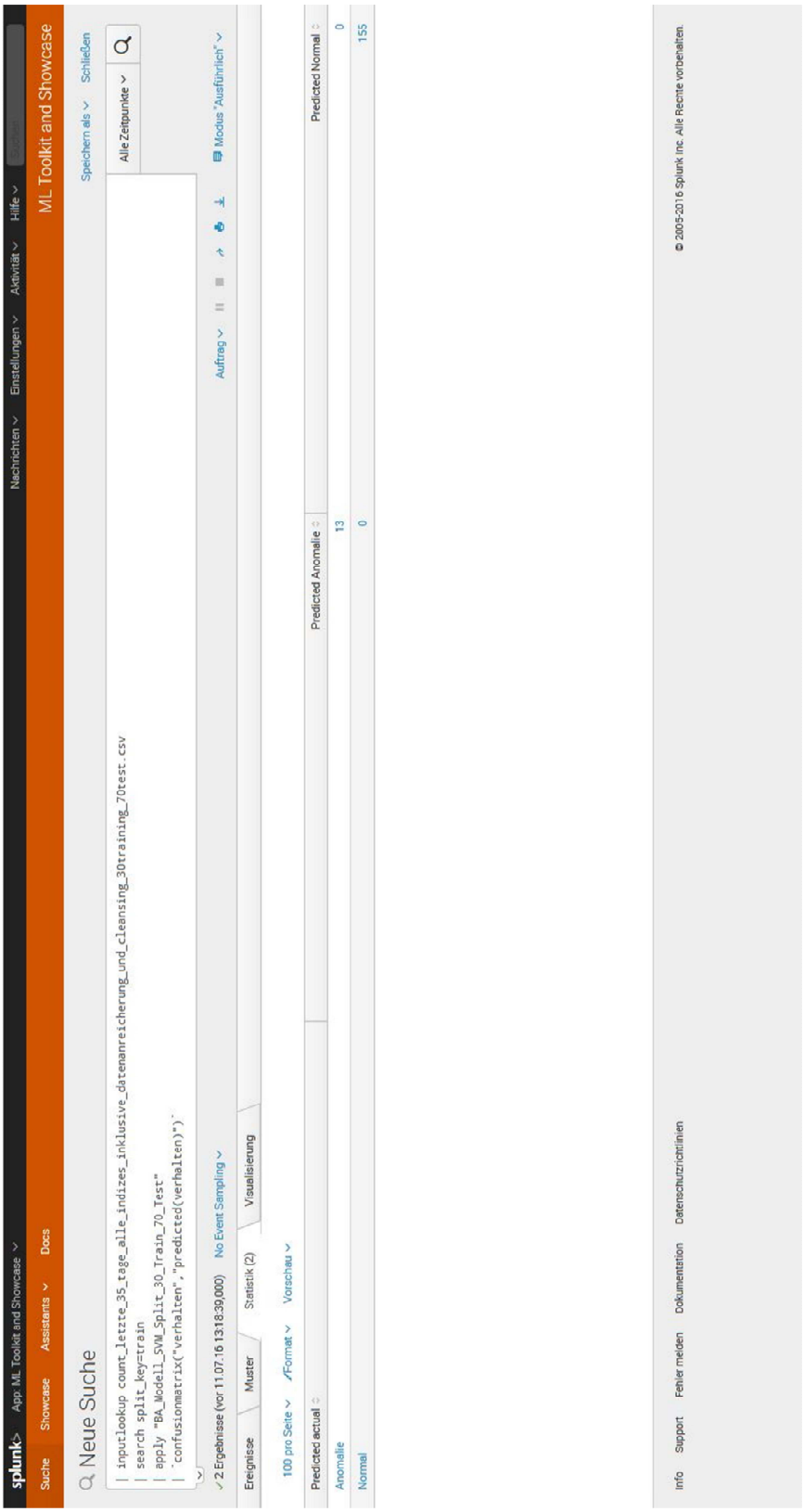
_time	count	countletzterstunde	countletztertag	countletztewoche	diff_1stunde	diff_1tag	diff_1woche	lowerBound	percentageDiffstunde	percentageDifftag	percentageDiffwoche	predicted(verhalten)	random_key
2016-05-03 00:00	8732086	8864780	9280102	6378374	-132694	-548016	2353712	-45.0	-1.519614	-6.275889	26.954751	Normal	19092047
2016-05-03 01:00	7239469	8732086	5732160	4421592	-1492617	1507309	2817877	-45.0	-20.617769	20.820712	38.923808	Normal	19363032
2016-05-03 02:00	5452099	7239469	10096731	3237106	-1787370	-4644632	2214993	-45.0	-32.783154	-85.189796	40.626427	Normal	13252203
2016-05-03 03:00	5970306	5452099	5827948	3694523	518207	142358	2275783	-45.0	8.679739	2.384434	38.118364	Normal	78874361
2016-05-03 04:00	7843930	5970306	6168545	3492415	1873624	1675385	4351515	-45.0	23.886292	21.359	55.476209	Normal	96796667
2016-05-03 05:00	7664875	7843930	6682567	3405461	-179055	982308	4259414	-45.0	-2.336046	12.815708	55.570561	Normal	102225691
2016-05-03 06:00	10844492	7664875	8976154	6309161	3179617	1868338	4535331	-45.0	29.320110	17.228451	41.821516	Normal	15315072
2016-05-03 09:00	19851159	15870605	20862125	16204171	3980554	-1010966	3646988	-45.0	20.051998	-5.09273	18.371663	Normal	19080735
2016-05-03 11:00	22563790	20905962	19551486	19202826	1657828	3012304	3360964	-45.0	7.347294	13.350169	14.895388	Normal	12881620
2016-05-03 12:00	21682704	22563790	18747787	18708846	-881086	2934917	2973858	-45.0	-4.063543	13.535752	13.715347	Normal	10915405
2016-05-03 13:00	21254419	21682704	19232906	21474064	-428285	2021513	-219645	-45.0	-2.015040	9.511025	-1.033409	Normal	10139750
2016-05-03 14:00	23078772	21254419	18263977	19601075	1824353	4814795	3477697	-45.0	7.904896	20.86244	15.068813	Normal	14281505
2016-05-03 15:00	22990032	23078772	17603047	18015026	-88740	5386985	4975006	-45.0	-0.385993	23.431829	21.639839	Normal	20808660
2016-05-03 17:00	23700289	23768044	16107845	15730990	-67755	7592444	7969299	-45.0	-0.285883	32.035238	33.625324	Normal	13223523
2016-05-03 18:00	21156426	23700289	16083274	14686825	-2543863	5073152	6469601	-45.0	-12.024068	23.979249	30.579839	Normal	13390173
2016-05-03 20:00	16756527	17955554	16811708	14967914	-1199027	-55181	1788613	-45.0	-7.155582	-0.32931	10.674127	Normal	6675834
2016-05-03 23:00	12526019	12409217	8864780	6923865	116802	3661239	5602154	-45.0	0.932475	29.229071	44.724138	Normal	14793781
2016-05-04 00:00	10840583	12526019	8732086	6612304	-1685436	2108497	4228279	-45.0	-15.547466	19.450033	39.004166	Normal	13888125
2016-05-04 01:00	8417459	10840583	7239469	4414879	-2423124	1177990	4002580	-45.0	-28.786882	13.994603	47.550930	Normal	15538474
2016-05-04 02:00	8326341	8417459	5452099	3593837	-91118	2874242	4732504	-45.0	-1.094334	34.519869	56.837739	Normal	10310477



A8 Konfusionsmatrix des angewendeten Modells (SVM) auf die Testdaten

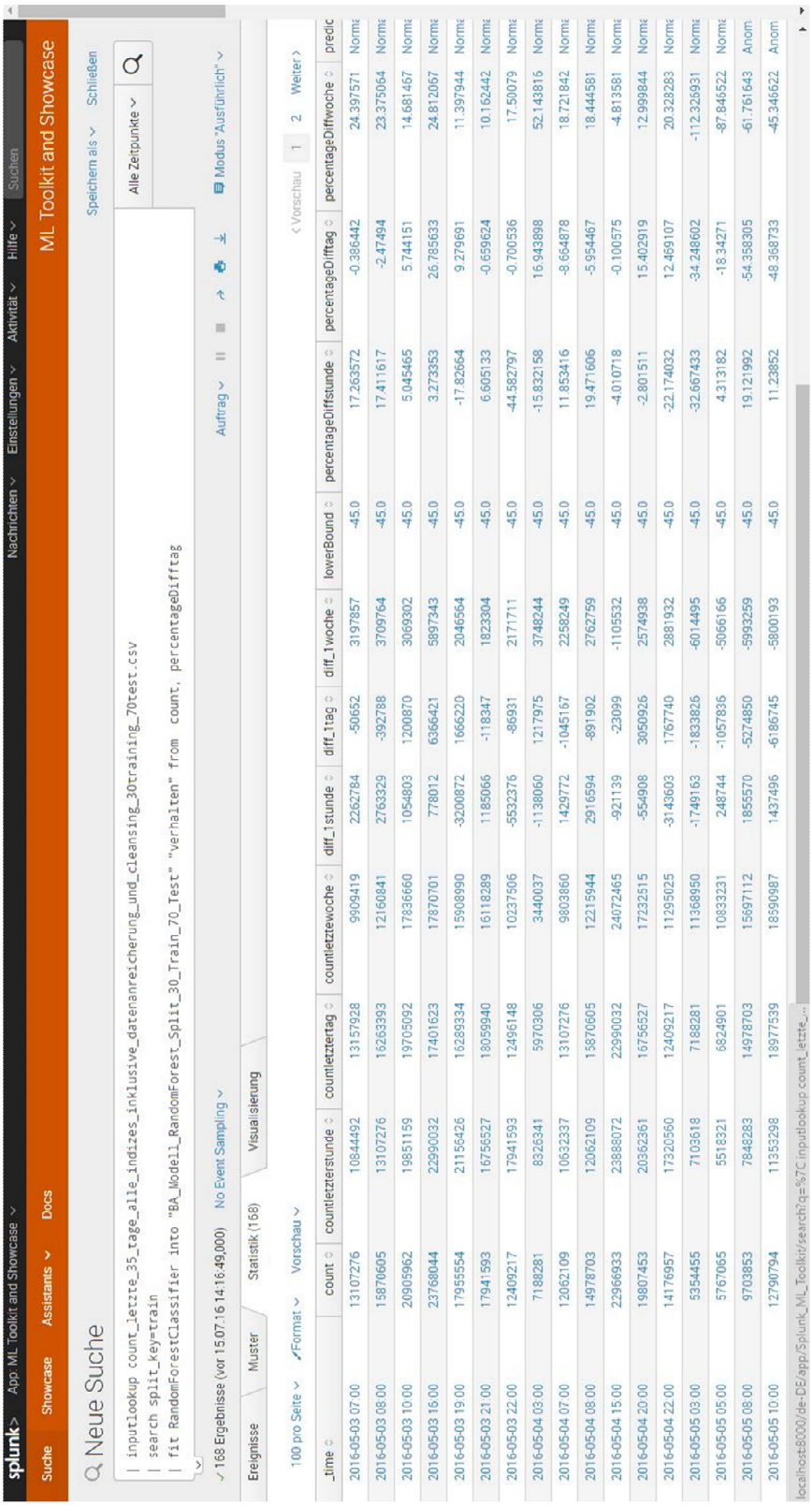


A9 Anwendung des trainierten Modells (SVM) auf die Trainingsdaten

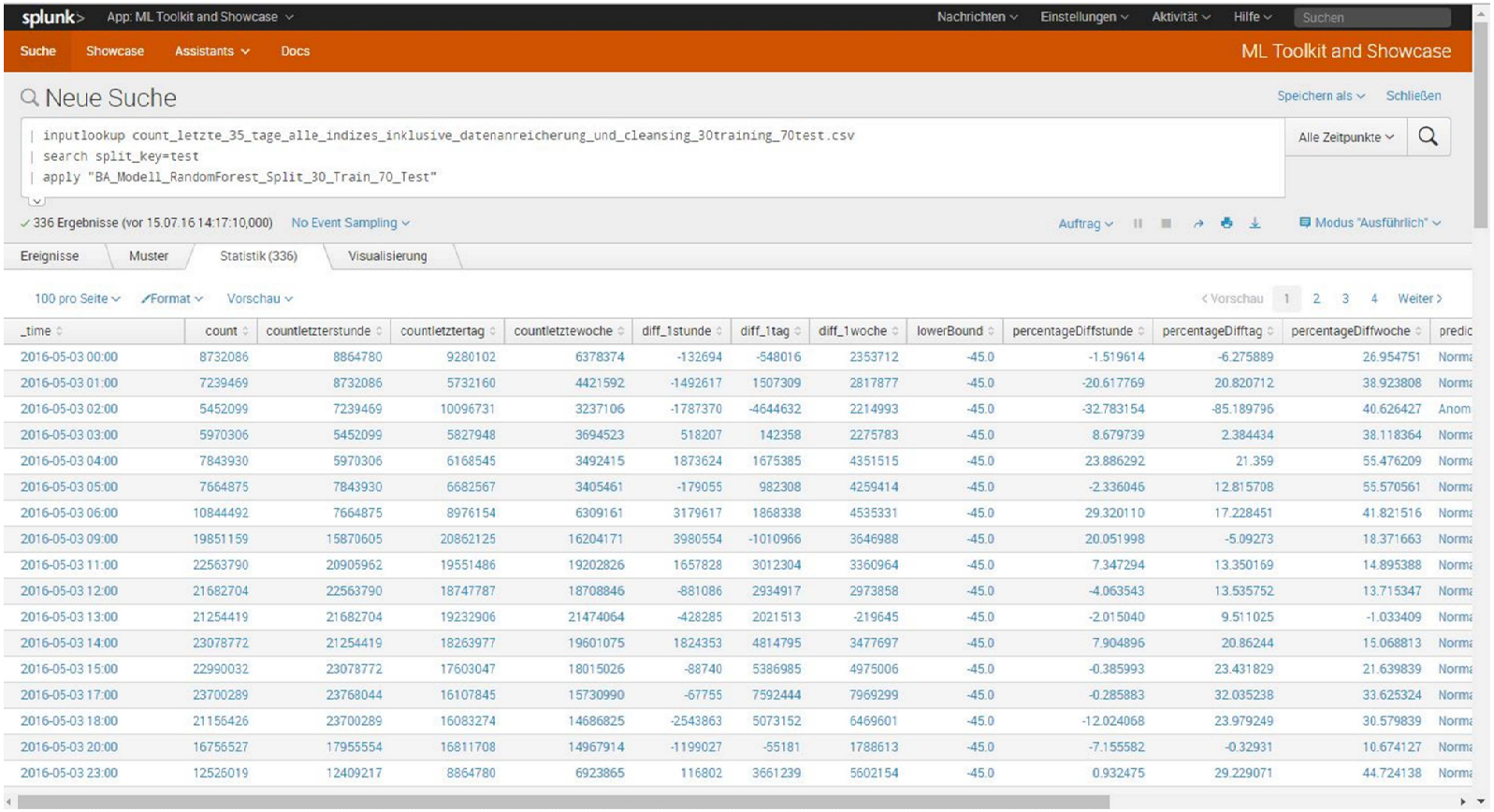


A10 Konfusionsmatrix des angewendeten Modells (SVM) auf die Trainingsdaten

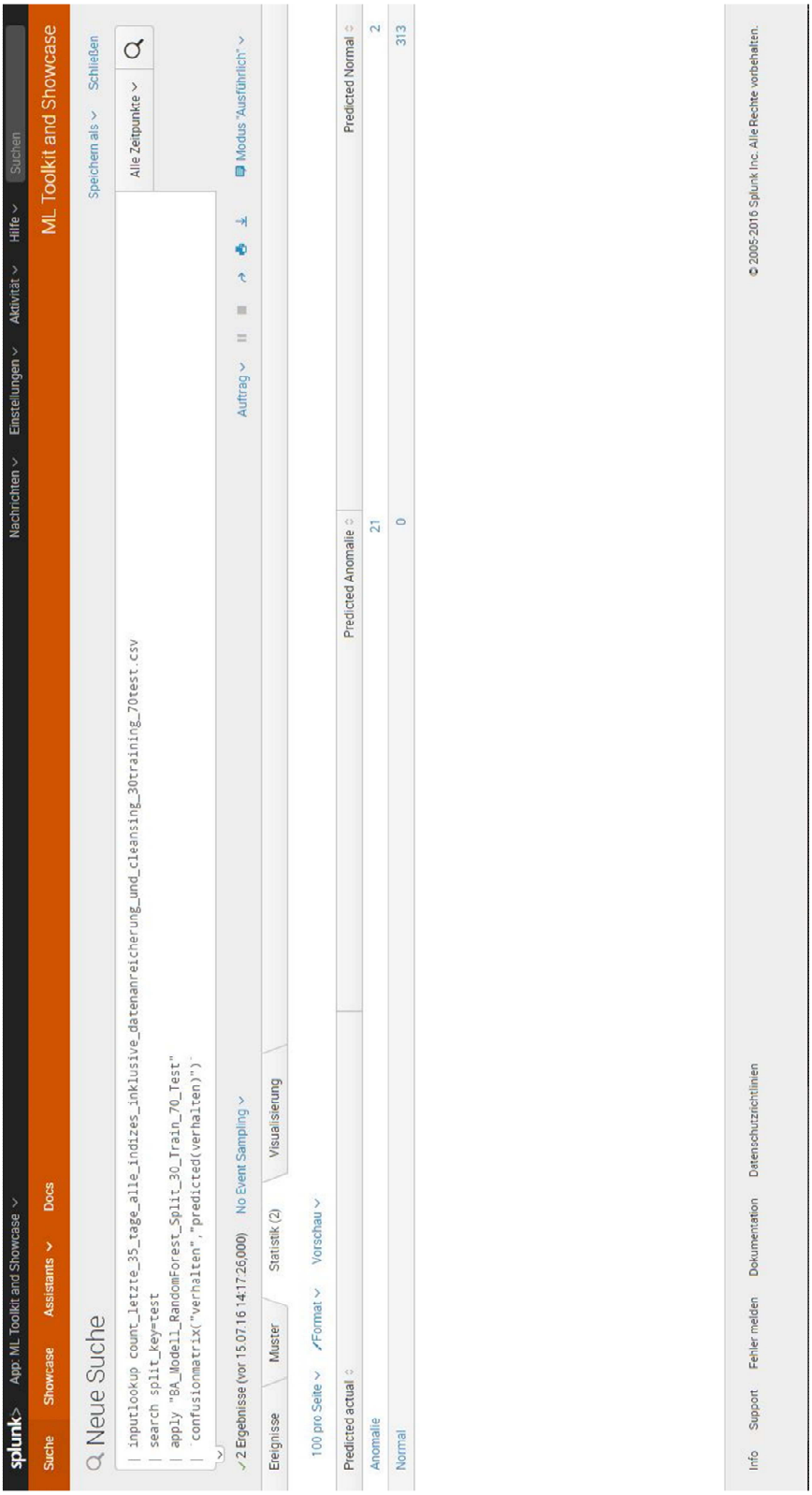
Anhang 6: Anwendung des Algorithmus RandomForestClassifier (RFC)



A11 Training des Algorithmus RandomForestClassifier

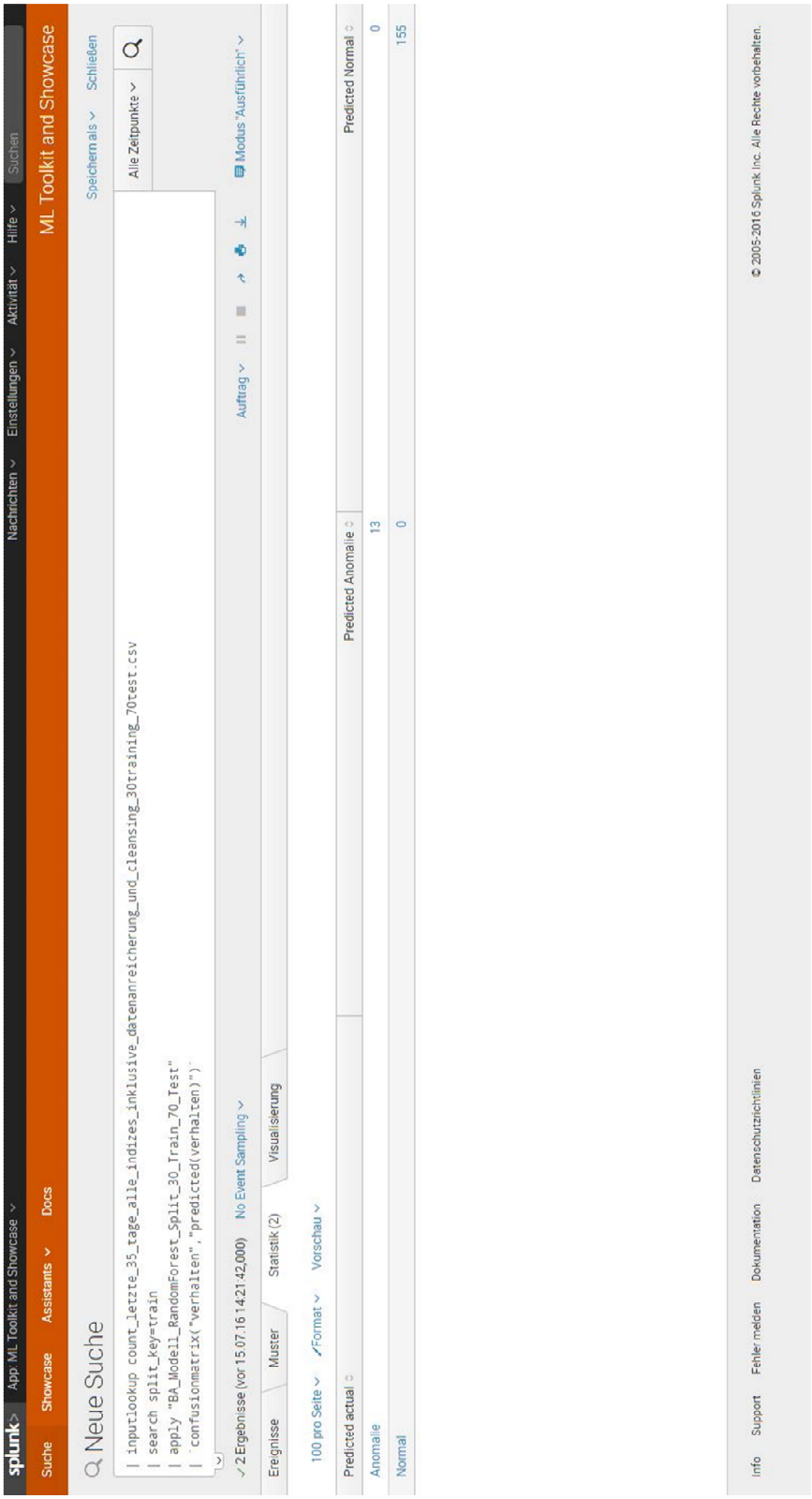


A12 Anwendung des trainierten Modells (RFC) auf die Testdaten



A13 Konfusionsmatrix des angewendeten Modells (RFC) auf die Testdaten

A14 Anwendung des trainierten Modells (RFC) auf die Trainingsdaten



A15 Konfusionsmatrix des angewendeten Modells (RFC) auf die Trainingsdaten

Anhang 7: Anwendung der RFC Modelle auf den Datensatz „Datensatz_Ohne_Spalte_verhalten.csv“

Das Modell „BA_Modell_RandomForest_Split_30_Train_70_Test“ hat 21 Anomalien auf dem Datensatz „Datensatz_Ohne_Spalte_verhalten.csv“ erkannt:

_time	count	countletztertag	diff_1tag	percentageDifftag	predicted(verhalten)
2016-05-26T17:00:00.000+0200	14424948	25427241	-11002293	-7.627.267	Anomalie
2016-05-26T22:00:00.000+0200	12691674	19907531	-7215857	-56.855.045	Anomalie
2016-05-27T18:00:00.000+0200	13775238	27822922	-14047684	-101.977.795	Anomalie
2016-05-27T19:00:00.000+0200	13421626	28605016	-15183390	-113.126.308	Anomalie
2016-05-28T08:00:00.000+0200	10394838	15082683	-4687845	-45.097.817	Anomalie
2016-05-28T15:00:00.000+0200	12038158	17560285	-5522127	-4.587.186	Anomalie
2016-05-28T21:00:00.000+0200	11098746	16930482	-5831736	-52.544.098	Anomalie
2016-05-29T01:00:00.000+0200	5005141	7724084	-2718943	-54.323.005	Anomalie
2016-05-29T06:00:00.000+0200	5175139	10332184	-5157045	-99.650.367	Anomalie
2016-06-01T05:00:00.000+0200	7325410	11803411	-4478001	-61.129.698	Anomalie
2016-06-02T02:00:00.000+0200	5808819	9460986	-3652167	-62.872.797	Anomalie
2016-06-02T03:00:00.000+0200	5996202	9028756	-3032554	-5.057.458	Anomalie
2016-06-02T22:00:00.000+0200	13666784	22507892	-8841108	-64.690.479	Anomalie
2016-06-04T03:00:00.000+0200	5889557	8544229	-2654672	-45.074.222	Anomalie
2016-06-04T05:00:00.000+0200	6365971	9234632	-2868661	-45.062.426	Anomalie
2016-06-04T07:00:00.000+0200	9724621	15514100	-5789479	-59.534.238	Anomalie
2016-06-04T08:00:00.000+0200	11838832	17720013	-5881181	-49.677.037	Anomalie
2016-06-04T09:00:00.000+0200	13623685	20259351	-6635666	-48.706.837	Anomalie
2016-06-04T10:00:00.000+0200	14401558	22042092	-7640534	-53.053.524	Anomalie
2016-06-04T15:00:00.000+0200	12977615	19783497	-6805882	-52.443.242	Anomalie
2016-06-05T06:00:00.000+0200	6219614	10009751	-3790137	-6.093.846	Anomalie

A16 BA_Modell_RandomForest_Split_30_Train_70_Test: predicted

Vollständiges Ergebnis siehe Datei „predicted_verhalten_BA_Modell_RandomForest_Split_30_Train_70_Test.csv“ auf dem Datenträger.

Das Modell „BA_Modell_RandomForest_Split_30_Training_70_Test_2_“ hat 17 Anomalien auf dem Datensatz „Datensatz_Ohne_Spalte_verhalten.csv“ erkannt:

_time	count	countletztertag	diff_1tag	percentageDifftag	predicted(verhalten)
2016-05-26T17:00:00.000+0200	14424948	25427241	-11002293	-7.627.267	Anomalie
2016-05-26T22:00:00.000+0200	12691674	19907531	-7215857	-56.855.045	Anomalie
2016-05-27T18:00:00.000+0200	13775238	27822922	-14047684	-101.977.795	Anomalie
2016-05-27T19:00:00.000+0200	13421626	28605016	-15183390	-113.126.308	Anomalie
2016-05-28T21:00:00.000+0200	11098746	16930482	-5831736	-52.544.098	Anomalie
2016-05-29T01:00:00.000+0200	5005141	7724084	-2718943	-54.323.005	Anomalie
2016-05-29T06:00:00.000+0200	5175139	10332184	-5157045	-99.650.367	Anomalie
2016-06-01T05:00:00.000+0200	7325410	11803411	-4478001	-61.129.698	Anomalie
2016-06-02T02:00:00.000+0200	5808819	9460986	-3652167	-62.872.797	Anomalie
2016-06-02T03:00:00.000+0200	5996202	9028756	-3032554	-5.057.458	Anomalie
2016-06-02T22:00:00.000+0200	13666784	22507892	-8841108	-64.690.479	Anomalie
2016-06-04T07:00:00.000+0200	9724621	15514100	-5789479	-59.534.238	Anomalie
2016-06-04T08:00:00.000+0200	11838832	17720013	-5881181	-49.677.037	Anomalie
2016-06-04T09:00:00.000+0200	13623685	20259351	-6635666	-48.706.837	Anomalie
2016-06-04T10:00:00.000+0200	14401558	22042092	-7640534	-53.053.524	Anomalie
2016-06-04T15:00:00.000+0200	12977615	19783497	-6805882	-52.443.242	Anomalie
2016-06-05T06:00:00.000+0200	6219614	10009751	-3790137	-6.093.846	Anomalie

A17 BA_Modell_RandomForest_Split_30_Training_70_Test_2_: predicted

Vollständiges Ergebnis siehe Datei „predicted_verhalten_BA_Modell_RandomForest_Split_30_Training_70_Test_2_.csv“ auf dem Datenträger.

Das Modell „BA_Modell_RandomForest_Split_30_Training_70_Test_3“ hat 21 Anomalien auf dem Datensatz „Datensatz_Ohne_Spalte_verhalten.csv“ erkannt:

_time	count	countletztertag	diff_1tag	percentageDifftag	predicted(verhalten)
2016-05-26T17:00:00.000+0200	14424948	25427241	-11002293	-7.627.267	Anomalie
2016-05-26T22:00:00.000+0200	12691674	19907531	-7215857	-56.855.045	Anomalie
2016-05-27T18:00:00.000+0200	13775238	27822922	-14047684	-101.977.795	Anomalie
2016-05-27T19:00:00.000+0200	13421626	28605016	-15183390	-113.126.308	Anomalie
2016-05-28T08:00:00.000+0200	10394838	15082683	-4687845	-45.097.817	Anomalie
2016-05-28T15:00:00.000+0200	12038158	17560285	-5522127	-4.587.186	Anomalie
2016-05-28T21:00:00.000+0200	11098746	16930482	-5831736	-52.544.098	Anomalie
2016-05-29T01:00:00.000+0200	5005141	7724084	-2718943	-54.323.005	Anomalie
2016-05-29T06:00:00.000+0200	5175139	10332184	-5157045	-99.650.367	Anomalie
2016-06-01T05:00:00.000+0200	7325410	11803411	-4478001	-61.129.698	Anomalie
2016-06-02T02:00:00.000+0200	5808819	9460986	-3652167	-62.872.797	Anomalie
2016-06-02T03:00:00.000+0200	5996202	9028756	-3032554	-5.057.458	Anomalie
2016-06-02T22:00:00.000+0200	13666784	22507892	-8841108	-64.690.479	Anomalie
2016-06-04T03:00:00.000+0200	5889557	8544229	-2654672	-45.074.222	Anomalie
2016-06-04T05:00:00.000+0200	6365971	9234632	-2868661	-45.062.426	Anomalie
2016-06-04T07:00:00.000+0200	9724621	15514100	-5789479	-59.534.238	Anomalie
2016-06-04T08:00:00.000+0200	11838832	17720013	-5881181	-49.677.037	Anomalie
2016-06-04T09:00:00.000+0200	13623685	20259351	-6635666	-48.706.837	Anomalie
2016-06-04T10:00:00.000+0200	14401558	22042092	-7640534	-53.053.524	Anomalie
2016-06-04T15:00:00.000+0200	12977615	19783497	-6805882	-52.443.242	Anomalie
2016-06-05T06:00:00.000+0200	6219614	10009751	-3790137	-6.093.846	Anomalie

A18 BA_Modell_RandomForest_Split_30_Training_70_Test_3: predicted

Vollständiges Ergebnis siehe Datei „predicted_verhalten_BA_Modell_RandomForest_Split_30_Training_70_Test_3.csv“ auf dem Datenträger.

Das Modell „BA_Modell_RandomForest_Split_30_Training_70_Test_4“ hat 19 Anomalien auf dem Datensatz „Datensatz_Ohne_Spalte_verhalten.csv“ erkannt:

_time	count	countletztertag	diff_1tag	percentageDifftag	predicted(verhalten)
2016-05-26T22:00:00.000+0200	12691674	19907531	-7215857	-56.855.045	Anomalie
2016-05-27T18:00:00.000+0200	13775238	27822922	-14047684	-101.977.795	Anomalie
2016-05-27T19:00:00.000+0200	13421626	28605016	-15183390	-113.126.308	Anomalie
2016-05-28T08:00:00.000+0200	10394838	15082683	-4687845	-45.097.817	Anomalie
2016-05-28T15:00:00.000+0200	12038158	17560285	-5522127	-4.587.186	Anomalie
2016-05-28T21:00:00.000+0200	11098746	16930482	-5831736	-52.544.098	Anomalie
2016-05-29T01:00:00.000+0200	5005141	7724084	-2718943	-54.323.005	Anomalie
2016-05-29T06:00:00.000+0200	5175139	10332184	-5157045	-99.650.367	Anomalie
2016-06-01T05:00:00.000+0200	7325410	11803411	-4478001	-61.129.698	Anomalie
2016-06-02T02:00:00.000+0200	5808819	9460986	-3652167	-62.872.797	Anomalie
2016-06-02T03:00:00.000+0200	5996202	9028756	-3032554	-5.057.458	Anomalie
2016-06-02T22:00:00.000+0200	13666784	22507892	-8841108	-64.690.479	Anomalie
2016-06-04T03:00:00.000+0200	5889557	8544229	-2654672	-45.074.222	Anomalie
2016-06-04T05:00:00.000+0200	6365971	9234632	-2868661	-45.062.426	Anomalie
2016-06-04T07:00:00.000+0200	9724621	15514100	-5789479	-59.534.238	Anomalie
2016-06-04T08:00:00.000+0200	11838832	17720013	-5881181	-49.677.037	Anomalie
2016-06-04T09:00:00.000+0200	13623685	20259351	-6635666	-48.706.837	Anomalie
2016-06-04T15:00:00.000+0200	12977615	19783497	-6805882	-52.443.242	Anomalie
2016-06-05T06:00:00.000+0200	6219614	10009751	-3790137	-6.093.846	Anomalie

A19 BA_Modell_RandomForest_Split_30_Training_70_Test_4: predicted

Vollständiges Ergebnis siehe Datei „predicted_verhalten_BA_Modell_RandomForest_Split_30_Training_70_Test_4.csv“ auf dem Datenträger.

Das Modell „BA_Modell_RandomForest_Split_30_Training_70_Test_5“ hat 23 Anomalien auf dem Datensatz „Datensatz_Ohne_Spalte_verhalten.csv“ erkannt:

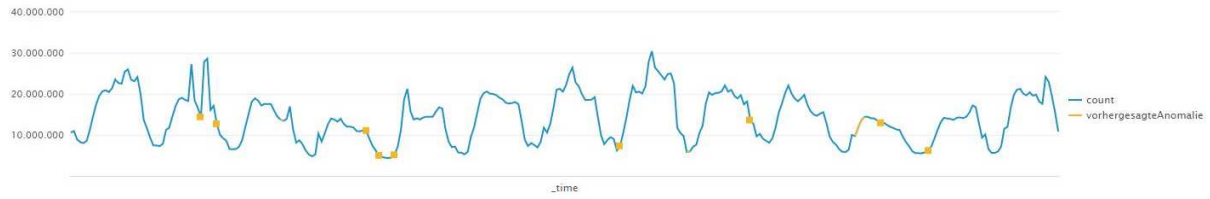
_time	count	countletztertag	diff_1tag	percentageDifftag	predicted(verhalten)
2016-05-26T17:00:00.000+0200	14424948	25427241	-11002293	-7.627.267	Anomalie
2016-05-26T22:00:00.000+0200	12691674	19907531	-7215857	-56.855.045	Anomalie
2016-05-27T14:00:00.000+0200	17559740	27297082	-9737342	-55.452.655	Anomalie
2016-05-27T18:00:00.000+0200	13775238	27822922	-14047684	-101.977.795	Anomalie
2016-05-27T19:00:00.000+0200	13421626	28605016	-15183390	-113.126.308	Anomalie
2016-05-28T08:00:00.000+0200	10394838	15082683	-4687845	-45.097.817	Anomalie
2016-05-28T15:00:00.000+0200	12038158	17560285	-5522127	-4.587.186	Anomalie
2016-05-28T21:00:00.000+0200	11098746	16930482	-5831736	-52.544.098	Anomalie
2016-05-29T01:00:00.000+0200	5005141	7724084	-2718943	-54.323.005	Anomalie
2016-05-29T06:00:00.000+0200	5175139	10332184	-5157045	-99.650.367	Anomalie
2016-06-01T05:00:00.000+0200	7325410	11803411	-4478001	-61.129.698	Anomalie
2016-06-02T02:00:00.000+0200	5808819	9460986	-3652167	-62.872.797	Anomalie
2016-06-02T03:00:00.000+0200	5996202	9028756	-3032554	-5.057.458	Anomalie
2016-06-02T15:00:00.000+0200	20567599	30429084	-9861485	-479.467	Anomalie
2016-06-02T22:00:00.000+0200	13666784	22507892	-8841108	-64.690.479	Anomalie
2016-06-04T03:00:00.000+0200	5889557	8544229	-2654672	-45.074.222	Anomalie
2016-06-04T05:00:00.000+0200	6365971	9234632	-2868661	-45.062.426	Anomalie
2016-06-04T07:00:00.000+0200	9724621	15514100	-5789479	-59.534.238	Anomalie
2016-06-04T08:00:00.000+0200	11838832	17720013	-5881181	-49.677.037	Anomalie
2016-06-04T09:00:00.000+0200	13623685	20259351	-6635666	-48.706.837	Anomalie
2016-06-04T10:00:00.000+0200	14401558	22042092	-7640534	-53.053.524	Anomalie
2016-06-04T15:00:00.000+0200	12977615	19783497	-6805882	-52.443.242	Anomalie
2016-06-05T06:00:00.000+0200	6219614	10009751	-3790137	-6.093.846	Anomalie

A20 BA_Modell_RandomForest_Split_30_Training_70_Test_5: predicted

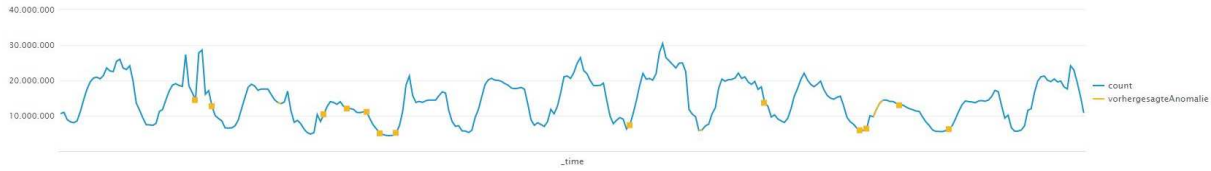
Vollständiges Ergebnis siehe Datei „predicted_verhalten_BA_Modell_RandomForest_Split_30_Training_70_Test_5.csv“ auf dem Datenträger.

Anhang 8: Diagramme der RFC Modelle nach Anwendung auf den Datensatz „Datensatz_Ohne_Spalte_verhalten.csv“

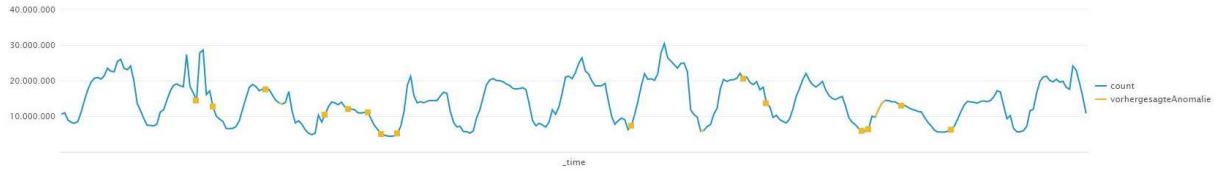
A21 Liniendiagramm: BA_Modell_RandomForest_Split_30_Train_70_Test



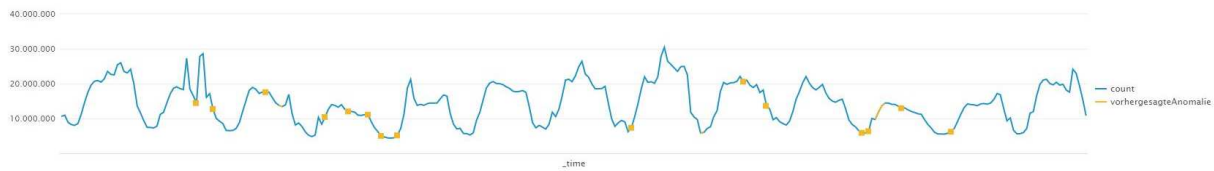
*A22 Liniendiagramm: BA_Modell_RandomForest_Split_30_Training_70_Test_2_*



A23 Liniendiagramm: BA_Modell_RandomForest_Split_30_Training_70_Test_3



A24 Liniendiagramm: BA_Modell_RandomForest_Split_30_Training_70_Test_4



A25 Liniendiagramm: BA_Modell_RandomForest_Split_30_Training_70_Test_5

Anhang 9: Berechnung der Bewertungsmaße anhand der Konfusionsmatrix

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

Logistische Regression:

BA_Modell_Logistic_Regression_Split_30_Train_70_Test

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{13+141}{13+141+14+0} = \mathbf{0,92}\ (\mathbf{92}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+14}{13+141+14+0} = \mathbf{0,08}\ (\mathbf{8}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{13}{13+14} = \mathbf{0,48}\ (\mathbf{48}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{13}{13+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*13}{2*13+14+0} = \mathbf{0,65}\ (\mathbf{65}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

Logistische Regression:

BA_Modell_Logistic_Regression_Split_30_Train_70_Test

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{23+297}{23+297+16+0} = \mathbf{0,95}\ (\mathbf{95}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{16+0}{23+297+0+16} = \mathbf{0,05}\ (\mathbf{5}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{23}{23+16} = \mathbf{0,59}\ (\mathbf{59}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{23}{23+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*23}{2*23+16+0} = \mathbf{0,74}\ (\mathbf{74}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

Logistische Regression:

BA_Modell_LogisticRegression_Split_30_Training_70_Test_2

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{10+119}{10+119+9+0} = \mathbf{0,93}\ (\mathbf{93}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{9+0}{10+119+9+0} = \mathbf{0,07}\ (\mathbf{7}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{10}{10+9} = \mathbf{0,53}\ (\mathbf{53}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{10}{10+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*10}{2*10+9+0} = \mathbf{0,69}\ (\mathbf{69}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

Logistische Regression:

BA_Modell_LogisticRegression_Split_30_Training_70_Test_2

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{24+333}{24+333+7+2} = \mathbf{0,98}\ (\mathbf{98}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{7+2}{24+333+7+2} = \mathbf{0,02}\ (\mathbf{2}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{24}{24+7} = \mathbf{0,77}\ (\mathbf{77}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{24}{24+2} = \mathbf{0,92}\ (\mathbf{92}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*24}{2*24+7+2} = \mathbf{0,84}\ (\mathbf{84}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

Logistische Regression:

BA_Modell_LogisticRegression_Split_30_Training_70_Test_3

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{10+134}{10+134+6+0} = \mathbf{0,96}\ (\mathbf{96}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{6+0}{10+134+6+0} = \mathbf{0,04}\ (\mathbf{4}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{10}{10+6} = \mathbf{0,63}\ (\mathbf{63}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{10}{10+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*10}{2*10+6+0} = \mathbf{0,77}\ (\mathbf{77}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

Logistische Regression:

BA_Modell_LogisticRegression_Split_30_Training_70_Test_3

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{25+317}{25+317+11+1} = \mathbf{0,97}\ (\mathbf{97}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{11+1}{25+317+11+1} = \mathbf{0,03}\ (\mathbf{3}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{25}{25+11} = \mathbf{0,69}\ (\mathbf{69}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{25}{25+1} = \mathbf{0,96}\ (\mathbf{96}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*25}{2*25+11+1} = \mathbf{0,81}\ (\mathbf{81}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

Logistische Regression:

BA_Modell_LogisticRegression_Split_30_Training_70_Test_4

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{11+131}{11+131+9+0} = \mathbf{0,94}\ (\mathbf{94}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{9+0}{11+131+9+0} = \mathbf{0,06}\ (\mathbf{6}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{11}{11+9} = \mathbf{0,55}\ (\mathbf{55}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{11}{11+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*11}{2*11+9+0} = \mathbf{0,71}\ (\mathbf{71}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

Logistische Regression:

BA_Modell_LogisticRegression_Split_30_Training_70_Test_4

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{25+297}{25+297+31+0} = \mathbf{0,91}\ (\mathbf{91}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{31+0}{25+297+31+0} = \mathbf{0,09}\ (\mathbf{9}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{25}{25+31} = \mathbf{0,45}\ (\mathbf{45}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{25}{25+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*25}{2*25+31+0} = \mathbf{0,62}\ (\mathbf{62}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

Logistische Regression:

BA_Modell_LogisticRegression_Split_30_Training_70_Test_5

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{14+142}{14+142+11+0} = \mathbf{0,93}\ (\mathbf{93}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{11+0}{14+142+11+0} = \mathbf{0,07}\ (\mathbf{7}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{14}{14+11} = \mathbf{0,56}\ (\mathbf{56}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{14}{14+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*14}{2*14+11+0} = \mathbf{0,72}\ (\mathbf{72}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

Logistische Regression:

BA_Modell_LogisticRegression_Split_30_Training_70_Test_5

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{22+299}{22+299+16+0} = \mathbf{0,95}\ (\mathbf{95}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{16+0}{22+299+16+0} = \mathbf{0,05}\ (\mathbf{5}\%)$

Precision: $\frac{TP}{TP+FP} = \frac{22}{22+16} = \mathbf{0,58}\ (\mathbf{58}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{22}{22+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*22}{2*22+16+0} = \mathbf{0,73}\ (\mathbf{73}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

Support Vector Machine:

BA_Modell_SVM_Split_30_Train_70_Test

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{13+155}{13+155+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+0}{13+155+0+0} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{13}{13+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{13}{13+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*13}{2*13+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

Support Vector Machine:

BA_Modell_SVM_Split_30_Train_70_Test

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{0+313}{0+313+0+23} = \mathbf{0,93}\ (\mathbf{93}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+23}{0+313+0+23} = \mathbf{0,07}\ (\mathbf{7}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{0}{0+0} = \boldsymbol{undefinierbar}$

Recall: $\frac{TP}{TP+FN} = \frac{0}{0+23} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*0}{2*0+0+23} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

Support Vector Machine:

BA_Modell_SupportVectorMachine_Split_30_Training_70_Test_2

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{10+128}{10+128+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+0}{10+128+0+0} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{10}{10+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{10}{10+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*10}{2*10+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

Support Vector Machine:

BA_Modell_SupportVectorMachine_Split_30_Training_70_Test_2

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{0+340}{0+340+0+26} = \mathbf{0,93}\ (\mathbf{93}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+26}{0+340+0+26} = \mathbf{0,07}\ (\mathbf{7}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{0}{0+0} = \boldsymbol{undefinierbar}$

Recall: $\frac{TP}{TP+FN} = \frac{0}{0+26} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*0}{2*0+0+26} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

Support Vector Machine:

BA_Modell_SupportVectorMachine_Split_30_Training_70_Test_3

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{10+140}{10+140+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+0}{10+140+0+0} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{10}{10+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{10}{10+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*10}{2*10+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

Support Vector Machine:

BA_Modell_SupportVectorMachine_Split_30_Training_70_Test_3

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{0+328}{0+328+0+26} = \mathbf{0,93}\ (\mathbf{93}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+26}{0+328+0+26} = \mathbf{0,07}\ (\mathbf{7}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{0}{0+0} = \boldsymbol{undefinierbar}$

Recall: $\frac{TP}{TP+FN} = \frac{0}{0+26} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*0}{2*0+0+26} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

Support Vector Machine:

BA_Modell_SupportVectorMachine_Split_30_Training_70_Test_4

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{11+140}{11+140+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+0}{11+140+0+0} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{11}{11+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{11}{11+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*11}{2*11+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

Support Vector Machine:

BA_Modell_SupportVectorMachine_Split_30_Training_70_Test_4

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{0+328}{0+328+0+25} = \mathbf{0,93}\ (\mathbf{93}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+25}{0+328+0+25} = \mathbf{0,07}\ (\mathbf{7}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{0}{0+0} = \boldsymbol{undefinierbar}$

Recall: $\frac{TP}{TP+FN} = \frac{0}{0+25} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*0}{2*0+0+25} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

Support Vector Machine:

BA_Modell_SupportVectorMachine_Split_30_Training_70_Test_5

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{14+153}{14+153+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+0}{14+153+0+0} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{14}{14+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{14}{14+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*14}{2*14+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

Support Vector Machine:

BA_Modell_SupportVectorMachine_Split_30_Training_70_Test_5

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{0+315}{0+315+0+22} = \mathbf{0,93}\ (\mathbf{93}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+22}{0+315+0+22} = \mathbf{0,07}\ (\mathbf{7}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{0}{0+0} = \boldsymbol{undefinierbar}$

Recall: $\frac{TP}{TP+FN} = \frac{0}{0+25} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*0}{2*0+0+22} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

RandomForestClassifier:

BA_Modell_RandomForest_Split_30_Train_70_Test

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{13+155}{13+155+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+0}{13+155+0+0} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{13}{13+13} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{13}{13+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*13}{2*13+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

RandomForestClassifier:

BA_Modell_RandomForest_Split_30_Train_70_Test

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{21+313}{21+313+0+2} = \mathbf{0,99}\ (\mathbf{99}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+2}{21+313+0+2} = \mathbf{0,01}\ (\mathbf{1}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{21}{21+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{21}{21+2} = \mathbf{0,91}\ (\mathbf{91}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*21}{2*21+0+2} = \mathbf{0,95}\ (\mathbf{95}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

RandomForestClassifier:

BA_Modell_RandomForest_Split_30_Training_70_Test_2_

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{10+128}{10+128+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+0}{10+128+0+0} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{10}{10+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{10}{10+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*10}{2*10+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

RandomForestClassifier:

BA_Modell_RandomForest_Split_30_Training_70_Test_2_

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{23+340}{23+340+0+3} = \mathbf{0,99}\ (\mathbf{99}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+3}{23+340+0+3} = \mathbf{0,01}\ (\mathbf{1}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{23}{23+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{23}{23+3} = \mathbf{0,88}\ (\mathbf{88}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*23}{2*23+0+3} = \mathbf{0,94}\ (\mathbf{94}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

RandomForestClassifier:

BA_Modell_RandomForest_Split_30_Training_70_Test_3

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{10+140}{10+140+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+0}{10+140+0+0} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{10}{10+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{10}{10+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*10}{2*10+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

RandomForestClassifier:

BA_Modell_RandomForest_Split_30_Training_70_Test_3

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{25+328}{25+328+0+1} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+1}{25+328+0+1} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{25}{25+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{25}{25+1} = \mathbf{0,96}\ (\mathbf{96}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*25}{2*25+0+1} = \mathbf{0,98}\ (\mathbf{98}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

RandomForestClassifier:

BA_Modell_RandomForest_Split_30_Training_70_Test_4

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{11+140}{11+140+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+0}{11+140+0+0} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{11}{11+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{11}{11+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*11}{2*11+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

RandomForestClassifier:

BA_Modell_RandomForest_Split_30_Training_70_Test_4

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{25+327}{25+327+1+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{1+0}{25+327+1+0} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{25}{25+1} = \mathbf{0,96}\ (\mathbf{96}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{25}{25+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*25}{2*25+1+0} = \mathbf{0,98}\ (\mathbf{98}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

RandomForestClassifier:

BA_Modell_RandomForest_Split_30_Training_70_Test_5

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{14+153}{14+153+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{0+0}{14+153+0+0} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{14}{14+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{14}{14+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*14}{2*14+0+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

RandomForestClassifier:

BA_Modell_RandomForest_Split_30_Training_70_Test_5

Accuracy: $\frac{TP+TN}{TP+TN+FP+FN} = \frac{22+314}{22+314+1+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

Klassifikationsfehler: $\frac{FP+FN}{TP+TN+FP+FN} = \frac{1+0}{22+314+1+0} = \mathbf{0,00}\ (\mathbf{0}\ \%)$

Precision: $\frac{TP}{TP+FP} = \frac{22}{22+1} = \mathbf{0,96}\ (\mathbf{96}\ \%)$

Recall: $\frac{TP}{TP+FN} = \frac{22}{22+0} = \mathbf{1,00}\ (\mathbf{100}\ \%)$

F-Measure (F_1): $\frac{2*TP}{2*TP+FP+FN} = \frac{2*22}{2*22+1+0} = \mathbf{0,98}\ (\mathbf{98}\ \%)$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

RandomForestClassifier: Durchschnitt aus allen fünf Splitting-Durchgängen

Accuracy: $\frac{500}{5} = 100\ \%$

Klassifikationsfehler: $0\ \%$

Precision: $\frac{500}{5} = 100\ \%$

Recall: $\frac{500}{5} = 100\ \%$

F-Measure (F_1): $\frac{500}{5} = 100\ \%$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

RandomForestClassifier: Durchschnitt aus allen fünf Splitting-Durchgängen

Accuracy: $\frac{497}{5} = 99,4\ \% \sim 99\ \%$

Klassifikationsfehler: $\frac{3}{5} = 0,6\ \% \sim 0,01 = 1\ \%$

Precision: $\frac{492}{5} = 98,4\ \% \sim 98$

Recall: $\frac{475}{5} = 95\ \%$

F-Measure (F_1): $\frac{483}{5} = 96,6\ \% \sim 97$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

Logistische Regression: Durchschnitt aus allen fünf Splitting-Durchgängen

Accuracy: $\frac{468}{5} = 93,6\ \% \sim 94\ \%$

Klassifikationsfehler: $\frac{32}{6} = 6,4\ \% \sim 6\ \%$

Precision: $\frac{275}{5} = 55\ \%$

Recall: $\frac{500}{5} = 100\ \%$

F-Measure (F_1): $\frac{354}{5} = 70,8\ \% \sim 71\ \%$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

Logistische Regression: Durchschnitt aus allen fünf Splitting-Durchgängen

Accuracy: $\frac{476}{5} = 95,2\ \% \sim\ 95\ \%$

Klassifikationsfehler: $\frac{24}{5} = 4,8\ \% \sim 5\ \%$

Precision: $\frac{308}{5} = 61,6\ \% \sim\ 62\ \%$

Recall: $\frac{488}{5} = 97,6\ \% \sim\ 98\ \%$

F-Measure (F_1): $\frac{374}{5} = 74,8\ \% \sim 75\ \%$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Trainingsdaten):

Support Vector Machine: Durchschnitt aus allen fünf Splitting-Durchgängen

Accuracy: $\frac{500}{5} = 100\ \%$

Klassifikationsfehler: $0\ \%$

Precision: $\frac{500}{5} = 100\ \%$

Recall: $\frac{500}{5} = 100\ \%$

F-Measure (F_1): $\frac{500}{5} = 100\ \%$

Berechnung der Bewertungsmaße anhand der Konfusionsmatrix (Testdaten):

Support Vector Machine: Durchschnitt aus allen fünf Splitting-Durchgängen

Accuracy: $\frac{465}{5} = 93\ \%$

Klassifikationsfehler: $\frac{35}{5} = 7\ \%$

Precision: $undefinierbar$

Recall: $0\ \%$

F-Measure (F_1): $0\ \%$

Literaturverzeichnis

Literatur

Alpaydin, E. (2008): *Maschinelles Lernen*, München, Oldenbourg Wissenschaftsverlag

Backhaus, K., u.a. (2016): *Multivariate Analysemethoden – Eine anwendungsorientierte Einführung*, 14. Auflage, Berlin und Heidelberg, Springer-Verlag

Barnett, V. & Lewis, T. (1994): *Outliers in Statistical Data*, 3. Auflage, New York, John Wiley & Sons

Bernstein, H. (2015): *Informations- und Kommunikationselektronik*, Berlin/Boston, De Gruyter Oldenbourg

Bourier, G. (2014): *Beschreibende Statistik: Praxisorientierte Einführung – Mit Aufgaben und Lösungen*, 12. Auflage, Wiesbaden, Springer Fachmedien

Bramer, M. (2013): *Principles of Data Mining*, 2. Auflage, London, Springer-Verlag

Cleff, C. (2011): *Deskriptive Statistik und moderne Datenanalyse: Eine computergestützte Einführung mit Excel, PASW (SPSS) STATA*, 2. Auflage, Wiesbaden, Springer Fachmedien

Cleve, J., Lämmel, U. (2014): *Data Mining*, München, Oldenbourg Wissenschaftsverlag

Dunning, T., Friedman, E. (2014): *Practical Machine Learning: A New Look At Anomaly Detection*, Sebastopol, O'Reilly Media, Inc

Ertel, W. (2013): *Grundkurs künstliche Intelligenz: Eine praxisorientierte Einführung*, 3. Auflage, Wiesbaden, Springer Fachmedien

Ester, M. (2013): *Knowledge Discovery in Databases - Techniken und Anwendungen*, Berlin und Heidelberg, Springer-Verlag

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P (1996): *Advances in Knowledge Discovery and Data Mining*, Cambridge, USA, AAAI/MIT Press

Finlay, S. (2014): *Predictive Analytics, Data Mining and Big Data - Myths, Misconceptions and Methods*, London, palgrave macmillan

Flick, U. (1999): *Qualitative Forschung - Theorie, Methoden, Anwendung in Psychologie und Sozialwissenschaften*, Reinbek, Rowohlt Tb

Freiknecht, J. (2014*): Big Data in der Praxis: Lösungen mit Hadoop, HBase und Hive – Daten speichern, aufbereiten, visualisieren*, München, Carl Hanser Verlag

Görz, G., Schneeberger, J., Schmid, U. (2014): *Handbuch der Künstlichen Intelligenz*, 5. Auflage, München, Oldenbourg Wissenschaftsverlag

Han, J., Kamber, M., Pei, J. (2012): *Data Mining: Concepts and Techniques*, Third Edition, Waltham, Morgan Kaufmann Publishers

Hawkins, D. M. (1980): *Identification of Outliers*, London, Chapman & Hall

Janssen, J., Laatz, W. (2013): *Statistische Datenanalyse mit SPSS – Eine anwendungsorientierte Einführung in das Basissystem und das Modul exakte Tests*, 8. Auflage, Berlin und Heidelberg, Springer-Verlag

Kantardzic, M. (2003): *Data Mining - Concepts, Models, Methods, and Algorithms*, 2. Auflage, New Jersey, IEEE Press

Laudon, C., Laudon, P., Schoder, D. (2010): *Wirtschaftsinformatik: Eine Einführung*, 2. Auflage, München, Pearson Deutschland GmbH

Mayer, H. (2013): *Interview und schriftliche Befragung – Grundlagen und Methoden empirischer Sozialforschung*, 6. Auflage, München, Oldenbourg Wissenschaftsverlag

Müller, R., Lenz, H. (2013): *Business Intelligence*, Berlin und Heidelberg, Springer-Verlag

Piazza, F. (2010): *Data Mining im Personalmanagement: Eine Analyse des Einsatzpotenzials zur Entscheidungsunterstützung*, Wiesbaden, Springer Fachmedien

Rambold, A. (1999): *Ausgewählte Verfahren zur Identifikation von Ausreißern und einflußreichen Beobachtungen in multivariaten Daten und Verfahren*, München, Herbert Utz Verlag

Rousseeuw, P. J. and Leroy, A. M. (1987): *Robust regression and outlier detection*., New York, John Wiley & Sons, Inc.

Runkler, T. A. (2015): *Data Mining: Modelle und Algorithmen intelligenter Datenanalyse*, 2. Auflage, Wiesbaden, Springer Fachmedien

Schendera, C. FG (2007): *Datenqualität mit SPSS*, München, Oldenbourg Wissenschaftsverlag

Scholl, A. (2015): *Die Befragung*, 3. Auflage, Konstanz und München, UVK Verlagsgesellschaft

Schön, D. (2016): *Planung und Reporting: Grundlagen, Business Intelligence, Mobile BI und Big-Data-Analytics*, 2.Auflage, Wiesbaden, Springer Gabler

Sharafi, A. (2013): *Knowledge Discovery in Databases - Eine Analyse des Änderungsmanagements in der Produktentwicklung*, Wiesbaden, Springer Fachmedien

Streck, G. (2004): *Einführung in die Statistik*, Norderstedt, Books on Demand GmbH

Tan, P.-N., Steinbach, M., Kumar, V. (2005): *Introduction to Data Mining*, 1. Auflage, Harlow, Pearson

Witten, I., Frank, E., Hall, M. (2011): *Data Mining – Practical Machine Learning Tools and Techniques*, Third Edition, Burlington, Morgan Kaufmann Publishers

Zeitschriften

Agyemang, M., Barker, K., and Alhajj, R. (2006): *A comprehensive survey of numeric and symbolic outlier mining techniques*, in Intelligent Data Analysis 10, 6, S. 521-538.

Aleskerov, E., Freisleben, B., Rao, B. (1997): *Cardwatch: A neural network based database mining system for credit card fraud detection.*, in IEEE Computational Intelligence for Financial Engineering. S. 220-226

Bakar, Z., u.a. (2006): *A comparative study for outlier detection techniques in data mining*., in Cybernetics and Intelligent Systems, IEEE Conference on, S. 1-6

Beckman, R. J. and Cook, R. D. (1983): *Outlier...s.*, in Technometrics 25, 2, S. 119-149

Bell, R., Koren, Y., Volinsky, C. (2010): *All togehter now: A perspective on the Netflix prize*, in Chance, Vol. 23/1, S. 24-29

Bradley, R., Haslett, J. (1990): *Interactive graphics for the exploratory analysis of spatial data – the interactive variogram cloud.* Vortrag bei CODATA Geostatistics meeting in Leeds, S. 1/5

Breiman, L. (2001): *Random Forests.*, in Machine Learning, Vol. 45, S. 5–32

Chandola, V., Banerjee, A., Kumar, V. (2009): *Anomaly Detection: A Survey*, in: ACM Computing Surveys, Vol. 41, No. 3, Article 15

Edgeworth, F. Y. (1887): *On discordant observations.* Philosophical Magazine 23, 5, S. 364-375.

Forrest, S., Warrender, C., Pearlmutter, B. (1999): *Detecting intrusions using system calls: Alternate data models*, in IEEE ISRSP. IEEE Computer Society, Washington, S.133-145

Guzella, T., Caminhas, W. (2009): *A review of machine learning approaches to spam filtering*, in: Expert Systems with Applications, Vol. 36/7, S. 10206-10222

Hagedorn, J. Bissantz, N., Mertens, P. (1997): *Data Mining (Datenmustererkennung): Stand der Forschung und Entwicklung*, in Wirtschaftsinformatik, 39, 6, S. 601-612.

Haslett, J. (1992): *Spatial Data Analysis-Challenges*, in: Journal of the Royal Statistical Society. Series D (The Statistician), Vol. 41, No. 3, S. 271 – 284

Hodge, V., Austin, J. (2004): *A Survey of Outlier Detection Methodologies.*, in Artificial Intelligence Review, 22, 2, S. 85-126.

Kumar, V. (2005): *Parallel and Distributed Computing for Cybersecurity. Distributed Systems Online*, in IEEE Computer Society, Vol. 6, No. 10, S. 4

Laurikkala, J., Juhola, M., and Kentala, E. (2000): *Informal identification of outliers in medical data.*, in Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology. S. 20-24

Markou, M. und Singh, S. (2003a): *Novelty detection: a review-part 1: statistical approaches.*, in Signal Processing 83, 12, S. 2481-2497

Markou, M. und Singh, S. (2003b): *Novelty detection: a review-part 2: neural network based approaches.*, in Signal Processing 83, 12, S. 2499-2521

Noble, C. C., Cook, D. J. (2003): *Graph-based anomaly detection*, in 9th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, S.631-636

Patcha, A. und Park, J. (2007): *An overview of anomaly detection techniques: Existing solutions and latest technological trends.*, in Comput. Networks 51, 12, S. 3448-3470.

Shekhar, S., Lu, C.-T., Zhang, P. (2001): *Detecting graph-based spatial outliers: algorithms and applications (a summary of results)*, in 7th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM Press, New York, S. 371-376

Weigend, A. S., Mangeas, M., Srivastava, A. N. (1995): *Nonlinear gated experts for time-series - discovering regimes and avoiding overfitting.*, in International Journal of Neural Systems 6, S. 373-399

Internet

Dechert, M. (2015): *Besser zentral: Professionelles Logging*, URL: http://www.heise.de/developer/artikel/Besser-zentral-Professionelles-Logging-2532864.html?artikelseite=2 [eingesehen am 19.05.2016]

EMC und IDC (2014): *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things – Executive Summary*, URL: http://germany.emc.com/leadership/digital-universe/2014iview/executive-summary.htm [eingesehen am 27.02.2016]

IBM Global Business Services (2012): Analytics: *The real-world use of big data*, URL: https://www.ibm.com/smarterplanet/global/files/se__sv_se__intelligence-__Analytics_-_The_real-world_use_of_big_data.pdf [eingesehen am 06.03.2016]

James, J. (2012): *How Much Data is Created Every Minute?*, URL: https://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/ [eingesehen am 11.03.2016]

Klein, D., Tran-Gia, P., Hartmann, M. (2013): *Big Data*, URL: http://www.gi.de/nc/service/informatiklexikon/detailansicht/article/big-data.html [eingesehen am 13.03.2016]

Klose, O. (2015): *Machine Learning (2) - Supervised versus Unsupervised Learning*, URL: http://oliviaklose.com/machine-learning-2-supervised-versus-unsupervised-learning/ [eingesehen am 31.03.2016]

Laerd Statistics (2013): *Types of Variable*, URL: https://statistics.laerd.com/statistical-guides/types-of-variable.php [eingesehen am 03.04.2016]

Laney, D. (2001): *3D Data Management: Controlling Data Volume, Velocity, and Variety, Application Delivery Strategies*, META Group Inc. (Hrsg.), URL: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf [eingesehen am 05.03.2016]

Lohninger, H. (2012): *Ausreißertests - Grundregeln*, URL: http://www.statistics4u.info/fundstat_germ/cc_outlier_tests_4sigma.html [eingesehen am 14.05.2016]

Lohninger, H. (2012): *Logistische Regression*, URL: http://www.statistics4u.info/fundstat_germ/ee_logistic_regression.html [eingesehen am 04.06.2016]

Mathewson, J. (2012): *Three Strategies for SEO Post Google Panda*, IBM Press, URL: http://www.ibmpressbooks.com/articles/article.asp?p=1829428 [eingesehen am 28.03.2016]

Mitchell, T. (1997): *Machine Learning*, McGraw Hill, URL: http://www.cs.cmu.edu/~tom/mlbook.html [eingesehen am 29.03.2016]

Otto GmbH & Co. KG (2015): Dossier: *Wir machen digitale Zukunft - die Transformation bei OTTO*, URL: https://www.otto.de/unternehmen/de/newsroom/dossiers/digitale-transformation.php [eingesehen am 18.04.2016]

Otto GmbH & Co. KG (2016): *Basisinformation – Über die Otto-Einzelgesellschaft*, März 2016, URL: https://www.otto.de/unternehmen/media-oc/docs/newsroom/-basismaterial/BM_Abbinder.pdf [eingesehen am 17.04.2016]

Otto GmbH & Co. KG (2016): *Pressemitteilung 21.03.2016: OTTO steigert Umsatz um rund zehn Prozent bei überproportionaler Ergebnisentwicklung*, URL: https://www.otto.de/unternehmen/de/newsroom/news/2016/PM_WPK_2016.php [eingesehen am 17.04.2016]

Prelert Inc. (2016): *Anomaly Detective API Engine: Put Machine Learning to Work – Benefits*, URL: http://info.prelert.com/products/anomaly-detective-engine [eingesehen am 17.07.2016]

Splunk Inc. (2015): *Big Data und ihre versteckten Schätze*, URL: http://www.splunk.com/de_de/solutions/solution-areas/big-data.html [eingesehen am 03.04.2016]

Splunk Inc. (2015): *Machine Learning Toolkit and Showcase App – User Guide – Custom search commands*, URL: http://docs.splunk.com/Documentation/MLApp/1.1.0/User/Customsearchcommands [eingesehen am 09.07.2016]

Splunk Inc. (2015): *Machine Learning Toolkit and Showcase*, URL: https://splunkbase.splunk.com/app/2890/#/overview [eingesehen am 26.05.2016]

Splunk Inc. (2015): *Maschinendaten - Riesige Datenströme, ständig wachsende Quellen, höchst wertvoll*, URL: https://www.splunk.com/de_de/view/what-is-it-data/SP-CAAACDC [eingesehen am 03.04.2016]

Splunk Inc. (2015): *Maschinendaten – Was sind Maschinendaten?*, URL: http://www.splunk.com/de_de/resources/machine-data.html [eingesehen am 29.03.2016]

Splunk Inc. (2015): *Operational Intelligence – Gewinnen Sie Einblicke aus Maschinendaten*, URL: http://www.splunk.com/de_de/resources/operational-intelligence.html [eingesehen am 29.03.2016]

Splunk Inc. (2015): *Splunk® Enterprise - deployer*, URL: http://docs.splunk.com/Splexicon:Deployer [eingesehen am 13.05.2016]

Splunk Inc. (2015): *Splunk® Enterprise - deployment server*, URL: http://docs.splunk.com/Splexicon:Deploymentserver [eingesehen am 13.05.2016]

Splunk Inc. (2015): *Splunk® Enterprise – Distributed Deployment Manual - Forwarders*, URL: http://docs.splunk.com/Documentation/Splunk/6.0/Deploy/Scaleyourdeployment [eingesehen am 09.05.2016]

Splunk Inc. (2015): *Splunk® Enterprise – Distributed Deployment Manual - Small enterprise deployment: Single search head with multiple indexers*, URL: http://docs.splunk.com/Documentation/Splunk/6.3.3/Deploy/Searchheadwithindexers [eingesehen am 15.05.2016]

Splunk Inc. (2015): *Splunk® Enterprise - Distributed Search - About search head clustering*, URL: http://docs.splunk.com/Documentation/Splunk/6.2.3/DistSearch/AboutSHC [eingesehen am 13.05.2016]

Splunk Inc. (2015): *Splunk® Enterprise - licence master*, URL: http://docs.splunk.com/Splexicon:Licensemaster [eingesehen am 13.05.2016]

Splunk Inc. (2015): *Splunk® Enterprise - Managing Indexers and Clusters of Indexers - Event processing and the data pipeline*, URL: http://docs.splunk.com/Documentation/Splunk/6.4.0/Indexer/Howindexingworks [eingesehen am 09.05.2016]

Splunk Inc. (2015): *Splunk® Enterprise – Search Reference - analyzefields*, URL: http://docs.splunk.com/Documentation/Splunk/6.4.1/SearchReference/Analyzefields [eingesehen am 15.07.2016]

Splunk Inc. (2015): *Splunk® Enterprise – Search Reference - timechart*, URL: http://docs.splunk.com/Documentation/Splunk/6.4.1/SearchReference/Timechart [eingesehen am 29.05.2016]

Splunk Inc. (2015): *Splunk® Enterprise 6.0.8 - Managing Indexers and Clusters of Indexers*, URL: http://docs.splunk.com/index.php?title=Documentation:Splunk:Indexer:Aboutindexesandindexers:6.0&action=pdfbook [eingesehen am 12.05.2016]

Splunk Inc. (2015): *Unternehmensübersicht – Splunk: Von Maschinendaten zu Opterational Intelligence*, URL: https://www.splunk.com/web_assets/pdfs/secure-/Splunk_Company_Overview_de.pdf [eingesehen am 29.03.2016]

Splunk Inc. (2016): *Splunk Schnellreferenz*, URL: https://www.splunk.com/web_assets/pdfs/secure/Splunk_Quick_Reference_Guide_de.pdf [eingesehen am 18.04.2016]

Splunk Inc. (2016): *Splunk® Enteprise – Die Plattform für Operational Intelligence*, URL: http://www.splunk.com/de_de/products/splunk-enterprise.html [eingesehen am 18.04.2016]

Splunk Inc. (2016): *Splunk® Enteprise – Distributed Deployment Manual - Scale your deployment with Splunk Enterprise components*, URL: http://docs.splunk.com/Documentation/Splunk/latest/Deploy/Distributedoverview [eingesehen am 12.05.2016]

Splunk Inc. (2016): *Splunk® Enteprise – Forwarding Data*, URL: http://docs.splunk.com/Documentation/Splunk/6.2.2/Forwarding/Introducingtheuniversalforwarder [eingesehen am 18.04.2016]

Splunk Inc. (2016): *Splunk® Enterprise – Distributed Deployment Manual - How data moves through Splunk Enterprise: the data pipeline*, URL: http://docs.splunk.com/Documentation/Splunk/6.2.9/Deploy/Datapipeline [eingesehen am 12.05.2016]

Splunk Inc. (2016): *Splunk® Enterprise – Distributed Deployment Manual - Components that help to manage your deployment*, URL: http://docs.splunk.com/Documentation/Splunk/latest/Deploy/Manageyourdeployment [eingesehen am 12.05.2016]

Splunk Inc. (2016): *Splunk® Enterprise – heavy forwarder*, URL: http://docs.splunk.com/Splexicon:Heavyforwarder [eingesehen am 12.05.2016]

Sonstiges

Borner, M. (2016): *Experteninterview: „Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen/Technologien“*, im Gespräch mit N.Netz, Hamburg 20.04.2016, Transkription siehe Anhang 2

Drieger, P. (2015): *Splunk_Machine_Learning_OneMoreSlide*, betriebsinterne Unterlage der Splunk Inc.

Drieger, P. (2016): *Experteninterview: „Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen/Technologien“*, im Gespräch mit N.Netz, Hamburg 19.04.2016, Transkription siehe Anhang 1

Günther, C. (2016): *Experteninterview: „Anomalie-Erkennung mit Hilfe von Machine-Learning-Algorithmen/Technologien“*, im Gespräch mit N.Netz, Hamburg 05.04.2016, Transkription siehe Anhang 3

Otto GmbH & Co. KG (2016): *Aktuelle Architektur Splunk*, Stand 2016

Prelert Inc. (2014): *3 Ways Anomaly Detection Improves IT Operations and Application Performance Management*, Whitepaper online verfügbar nach Anmeldung [eingesehen am 03.03.2016]

Versicherung über Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, den _____ _____