HAW HAMBURG

Bachelorarbeit

Ronja Schöttler

Maschinelles Lernen zur Analyse von Einflussfaktoren auf den Preis und Preisentwicklungen im deutschen Glasfasermarkt

Fakultät Technik und Informatik
Department Informatik

Faculty of Computer Science and Engineering
Department Computer Science

Ronja Schöttler

Maschinelles Lernen zur Analyse von Einflussfaktoren auf den Preis und Preisentwicklungen im deutschen Glasfasermarkt

Bachelorarbeit 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. Ulrike Steffens
Zweitgutachter: Prof. Dr.-Ing. Olaf Zukunft

Eingereicht am: 05. September 2019

Ronja Schöttler

Thema der Arbeit

Maschinelles Lernen zur Analyse von Einflussfaktoren auf den Preis und Preisentwicklungen im deutschen Glasfasermarkt

Stichworte

Glasfaserpreise, Mietleitungen, maschinelles Lernen, Regression, Entscheidungsbäume

Kurzzusammenfassung

Diese Arbeit beschäftigt sich mit der Analyse von Preisen von Glasfaser Mietleitungen, zwecks Schaffung einer Grundlage für eine Applikation, welche Angebote im deutschen Glasfasermarkt evaluiert. Dazu wird analysiert, welche Faktoren die Preise von Mietleitungen beeinflussen und wie sich die Preise zeitlich entwickeln. Anschließend wird, unter anderem mit Hilfe von Methoden des maschinellen Lernens, ein Modell entworfen, welches einen marktkonformen Preis für ein Angebot vorhersagen kann.

Ronja Schöttler

Title of Thesis

Machine learning for the analysis of influencing factors on price and price developments in the German optical fiber market

Keywords

fiber price, leased lines, machine learning, regression, decision tree

Abstract

This thesis is about an analysis of fiber leased lines prices in order to create a basis for an application that evaluates offers in the German optical fiber market. It is analyzed which factors influence the prices and how is the price developed over time. Finally a model is designed that predicts a market price for an offer. For that machine learning algorithms were applied.

Inhaltsverzeichnis

Abbildungsverzeichnis **v**

Tabellenverzeichnis **vii**

1 Einleitung **1**

1.1 Motivation 1
1.2 Aufgaben und Ziele 2
1.3 Gliederung der Arbeit 2

2 Grundlagen **4**

2.1 Glasfaserleitungen 4
2.1.1 Verlegearten und Festlegung der Trassenführung 4
2.1.2 Kostenmodell einer Glasfaserleitung 8
2.2 Der Beschaffungsprozess 9
2.3 Maschinelles Lernen 11
2.3.1 Grundlagen des überwachten Lernens 11
2.3.2 Regressionsverfahren 12
2.3.3 Klassifikationsverfahren 13

3 Analyse der Preise **16**

3.1 Anforderungen an die Analyse 16
3.2 Aufstellung der Hypothesen 17
3.2.1 Hypothesen zu den Einflussfaktoren auf den Preis 17
3.2.2 Hypothesen zu der Preisentwicklung 21
3.3 Datenaufbereitung 22
3.3.1 Datenerhebung und Integration 22
3.3.2 Datenberechnung 26
3.3.3 Datenaggregation 27
3.3.4 Datenbereinigung 28

3.4 Durchführung der Analyse 29
3.4.1 Analyse der Einflussfaktoren des Preises 29
3.4.2 Analyse der Preisentwicklung 47

4 Entwicklung der Modelle **51**
4.1 Vorverarbeitung der Daten 52
4.2 Entwurf der Vorhersagemodelle 53
4.2.1 Vorhersagemodell für den monatlichen Preis 54
4.2.2 Vorhersagemodell für den einmaligen Preis 56
4.3 Entwurf eines Trendmodells 59
4.3.1 Trendmodell des monatlichen Preises pro Meter 59
4.3.2 Konkatenation der Modelle 61
4.4 Prognose der inflationären Entwicklung der Preise 62
4.4.1 Beachtung der inflationären Entwicklung des monatlichen Preises im Modell 62
4.4.2 Beachtung der inflationären Entwicklung des einmaligen Preises im Modell 64
4.5 Bewertung der Modelle 64
4.5.1 Bewertung des Modells für den monatlichen Preis 65
4.5.2 Bewertung des Modells für den einmaligen Preis 66

5 Fazit und Ausblick **68**

A Anhang **75**
A.1 Bodenübersicht von Deutschland 75
A.2 Zuordung der Bodentypen zu den Bodenklassen 78
A.3 Datensatz 80
A.4 Entscheidungsbaum 81

Selbstständigkeitserklärung **83**

Abbildungsverzeichnis

2.1 Kabelschutzrohr mit 7 Rohr-in-Rohr Mikrorohren. Quelle: Loiber (2018) S.37 Abb.27 5
2.2 Verbindung zweier Standorte unter Nutzung eines bestehenden Netzes. . . 7
2.3 Anbindung einer neuen Leitung an einem bestehenden Anschlusspunkt. Quelle: Telefónica Germany GmbH und Co. OHG (2009) 7
2.4 Bodenklassen nach DIN 18300. Quelle: Landesamt, für Bergbau, Energie und Geologie 9
2.5 Darstellung des "Resource Lifecycle ManagementProzesses. Quelle: Czarnecki und Dietze (2017), S.165 Abb. 4.56 modifiziert 10
2.6 Entscheidungsbaum zum Sterbealter einer Person. Quelle: van der Aalst (2016), S.95 Abb. 4.1 modifiziert 14

3.1 Erschwernisfaktoren der Bodenklassen. 24
3.2 "Übersetzungsmatrizen" für das Merkmal Area Class (links) und das Merkmal Besiedlungsklasse (rechts). 28
3.3 Monatliche Kosten der Leitung pro Meter. 33
3.4 Monatliche Kosten der Leitung pro Meter des Lieferanten A. 34
3.5 Monatliche Kosten der Leitung pro Meter des Lieferanten B. 34
3.6 Einmalige Kosten der Leitung pro Meter. 35
3.7 Darstellung der Trassen nach der Area Class und den monatlichen Kosten. 37
3.8 Darstellung der Trassen nach der Besiedlungsklasse und den monatlichen Kosten. 38
3.9 Darstellung der Trassen nach der Area Class und den einmaligen Kosten. . 39
3.10 Darstellung der Trassen nach der Besiedlungsklasse und den einmaligen Kosten. 40
3.11 Monatliche Kosten der Leitung je nach Höhe des Erschwernisfaktors der Bodenklasse. 41

3.12 Einmalige Kosten der Leitung je nach Höhe des Erschwernisfaktors der Bodenklasse. 42
3.13 Monatliche Kosten der Trasse nach Anzahl der möglichen Lieferanten. . . 43
3.14 Einmalige Kosten der Trasse nach Anzahl der möglichen Lieferanten. . . . 44
3.15 Monatliche Kosten der Leitung nach Länge der Vertragsdauer. 45
3.16 Einmalige Kosten der Leitung nach Länge der Vertragsdauer. 46
3.17 Zeitliche Entwicklung der monatlichen Preise pro Meter. 48
3.18 Zeitliche Entwicklung der einmaligen Preise. 49

4.1 Vorgang der One-Hot Kodierung am Beispiel des Merkmals Area Class. . 53
4.2 Ausschnitt des Trainingsdatensatzes für die Aufstellung eines Vorhersagemodells für den monatlichen Preis. 54
4.3 Ausschnitt des Trainingsdatensatzes für die Aufstellung eines Vorhersagemodells für den einmaligen Preis. 56
4.4 Darstellung der Preise pro Meter im zeitlichen Verlauf, sowie der erstellten exponentiellen Regressionsfunktion. 60

A.1 Bodenübersichtskarte von Deutschland 1 : 3 000 000. Quelle: Bundesanstalt für Geowissenschaften und Rohstoffe (2014) 77
A.2 Zuordnung der Bodentypen zu den Bodenklassen. 79
A.3 Ausschnitt des Datensatzes. 80
A.4 Entscheidungsbaum aus dem Random Forest zur Vorhersage eines einmaligen Preises. 82

Tabellenverzeichnis

4.1 Preiskategorien der einmaligen Kosten. 57
4.2 Bewertungsbogen für die Vorhersagemodelle. 65
4.3 Bewertung des Modells für den monatlichen Preis. 66
4.4 Bewertung des Modells für den einmaligen Preis. 67

1 Einleitung

1.1 Motivation

Das Nutzungsverhalten der Gesellschaft in Bezug auf das Internet hat sich in den letzten Jahren weiter entwickelt. Die Menschen verbringen immer mehr Zeit online. Dadurch steigen die Anforderungen an den Zugang zum Internet. Eine hohe Geschwindigkeit steht im Vordergrund der Anforderungen. Dieser Anspruch ist technisch mit einer Datenübertragung über Glasfaserleitungen möglich. Glasfaserkabel haben gegenüber anderen Übertragungskanälen den Vorteil, dass sie eine hohe Bandbreite bereitstellen und somit eine hohe Datenübertragung ermöglichen. Durch Glasfaserleitungen können die stetig steigenden Anforderungen an den Internetzugang realisiert werden.
Somit steigt die Bedeutung von Glasfaserleitungen auch für die Telekommunikationsunternehmen. Die Unternehmen möchten im Wettbewerb nicht verdrängt werden und den Kunden zufriedenstellen. Das beschriebene Nutzungsverhalten des Kunden hat zur Folge, dass der Fokus der Telekommunikationsunternehmen auf der Anbindung der Mobilfunkmasten an eine Glasfaserinfrastruktur liegt. Einerseits investieren die Unternehmen in Eigenrealisierungen von Glasfaserinfrastruktur. Andererseits besteht die Möglichkeit Glasfaserinfrastruktur bei anderen Glasfaser-Infrastruktur-Betreibern anzumieten.

Diese Arbeit wurde in einem großen Telekommunikationsunternehmen durchgeführt. Das Geschäftmodell des betrachteten Unternehmens sieht vor, dass benötigte Glasfaserverbindungen vorwiegend von anderen Glasfaser-Infrastruktur-Betreibern angemietet werden. Deshalb bezieht sich diese Arbeit ausschließlich auf angemietete Glasfaserleitungen.

Der Beschaffungsprozess von Glasfaserleitungen stellt die Einkaufsorganisation des Telekommunikationsunternehmen vor einige Schwierigkeiten. Es muss vor allem entschieden werden, ob ein erhaltenes Mietangebot für eine Glasfaserleitung angemessen ist. Aus diesem Grund wäre eine Applikation, die ein Angebot analysiert und ermittelt, ob es sich um einen marktkonformen Preis handelt, sinnvoll. Um jedoch ermitteln zu können, ob

ein Preis marktkonform ist, muss bekannt sein, welche Faktoren ausschlaggebend für den Preis sind. Diese Bachelorarbeit beschäftigt sich mit der Erfassung dieser Faktoren, um ein Modell zu entwickeln, welches einen marktkonformen Preis für ein Mietangebot einer Glasfaserleitung vorhersagt.

1.2 Aufgaben und Ziele

Ziel dieser Arbeit ist es zu analysieren, welche Faktoren Einfluss auf den Mietpreis einer Glasfaserleitung haben und wie sich die Preise in Zukunft entwickeln. Es soll ein Modell entworfen werden, welches anhand der Einflussfaktoren und der Preisentwicklung einen Preis für ein Mietangebot bestimmt. Dieses Ergebnis soll die Grundlage für eine Applikation bilden, welche bestimmt, ob der Preis eines Mietangebot marktkonform ist oder nicht. Die Applikation soll als Unterstützung des Beschaffungsprozesses von Glasfaserleitungen im Telekommunikationsbereich dienen.

Um herauszufinden, welche Einflussfaktoren auf den Preis existieren und wie sich die Preisentwicklung gestaltet, werden Daten von gemieteten Glasfaserleitungen eines großen Telekommunikationsunternehmens analysiert. Anschließend wird ein Modell entworfen, welches einen marktkonformen Preis für ein Angebot bestimmt. Dabei kommen Methoden des maschinellen Lernens zum Einsatz.

1.3 Gliederung der Arbeit

Die Arbeit ist folgendermaßen strukturiert:
In Kapitel 2 wird ein Überblick über Glasfaserleitungen gegeben. Die Informationen sind relevant für das Verständnis der Arbeit. Das Kapitel beinhaltet eine Erläuterung der Verlegearten von Glasfaser und die Festlegung der Trassenführung, sowie das Kostenmodell einer Glasfaserleitung. Zudem wird der Beschaffungsprozess beschrieben. Ferner werden die Algorithmen des maschinellen Lernens vorgestellt, welche zur Erstellung des Vorhersagemodells genutzt werden.

Das Kapitel 3 beinhaltet die Preisanalyse. Zuerst werden die Methoden der Datenerhebung dargelegt. Anschließend wird die Analyse der Daten hinsichtlich der Einflussfaktoren auf den Preis einer gemieteten Leitung und der Preisentwicklung erläutert.

Die Aufstellung eines Vorhersagemodells für den Preis ist Gegenstand von Kapitel 4. Dabei werden Methoden des maschinellen Lernens angewandt. Das Ergebnis besteht aus einem Modell, welches einen marktkonformen Preis für ein Mietangebot bestimmt.

Ein Fazit und Ausblick sind Bestandteil des fünften Kapitels.

2 Grundlagen

Dieses Kapitel beinhaltet eine Einführung in das Thema Glasfaserleitungen und eine Erläuterung der Methoden, die zur Aufstellung der Vorhersagemodelle verwendet wurden. In Abschnitt 2.1 werden die grundlegenden Eigenschaften von Glasfaserleitungen vorgestellt. Dies beinhaltet die Verlegearten von Glasfaserleitungen, die Festlegung der Trassenführung, sowie das Kostenmodell. Die Vorstellung des Kostenmodells einer Glasfaserleitung ist bedeutsam für den Rest der Arbeit, da angenommen wird, dass sich diese Faktoren in dem Mietpreis widerspiegeln, welchen es zu untersuchen gilt. Zudem wird in Abschnitt 2.2 der Beschaffungsprozess von Glasfaserleitungen vorgestellt. Im Anschluss wird in Abschnitt 2.3 ein Überblick über maschinelles Lernen gegeben und die spezifischen Verfahren, die zur Aufstellung des Vorhersagemodells genutzt wurden, werden vorgestellt. Zu den Verfahren zählen der Random Forest, sowie die Ridge Regression.

2.1 Glasfaserleitungen

Glasfaser ist ein Medium, welches zur Datenübertragung genutzt wird. Über Lichtwellen werden Daten von einem Ende bis ans andere Ende übertragen. Glasfaserkabel haben einige Vorteile. Sie bieten eine hohe Bandbreite, sowie eine hohe Reichweite. Bei der Verlegung von Glasfaserkabeln sind mehrere Verlegearten möglich. Die Verlegearten, sowie das Kostenmodell einer Glasfaserleitung, werden in dem folgenden Abschnitt dargestellt.

2.1.1 Verlegearten und Festlegung der Trassenführung

Glasfaserkabel können unter- und oberirdisch verlegt werden. Die oberirdische Verlegung, wie die Luftverkabelung, wird nicht genauer betrachtet, da sie in den meisten Fällen, aus Gründen des Stadtbildes, in Deutschland nicht zulässig ist. Im Folgenden wird auf die unterirdische Verlegung eingegangen.

Bei der unterirdischen Verlegung wird zwischen mehreren Verlegearten unterschieden:

- konventionelle Rohr- bzw. Kabelkanalverlegung (inkl. Rohrteiler)
- Rohr-in-Rohr Verlegung (Mikrorohr-in-Rohr)
- erdverlegbare Mikrorohrverbände
- erdverlegtes Kabel

Während bei einer Kupferkabelverlegung meistens die konventionelle Variante angewandt wird (konventionelle Rohr- bzw. Kabelkanalverlegung) sorgen die Eigenschaften von Glasfaser dafür, dass die Möglichkeit besteht Mikrorohre zu verwenden. Die Mikrorohre bieten einen großen Vorteil, da sie eine nachträgliche Verlegung von Glasfaser erlauben. Es existiert die Möglichkeit Glasfaser nachträglich in die Mikrorohre einzublasen.
Bei dieser Verlegeart unterscheidet man zwischen zwei unterschiedlichen Varianten, der Rohr-in-Rohr-Verlegung und den erdverlegbaren Mikrorohrverbänden. Bei einer Rohr-in-Rohr Verlegung werden die Mikrorohre in ein Kabelschutzrohr gelegt. Dies ist in Abbildung 2.1 dargestellt. Dagegen werden bei den erdverlegbare Mikrorohrverbänden mehrere Mikrorohre zu einem Verbund zusammengefasst.

Abbildung 2.1: Kabelschutzrohr mit 7 Rohr-in-Rohr Mikrorohren. Quelle: Loiber (2018) S.37 Abb.27

Ein Vergleich der beiden Varianten zeigt, dass die Mikrorohre, welche in ein Kabelschutzrohr gelegt werden eine dünnere Außenwand besitzen und somit einen größeren Innendurchmesser haben. Dies sorgt dafür, dass mehr Platz für das Einblasen von Glasfaser besteht. Im Gegensatz zu den erdverlegbaren Mikrorohrverbänden sind sie nicht für eine direkte Erdverlegung geeignet. Eine weitere Variante ist, Glasfaserkabel direkt in die Erde zu verlegen. Dafür wird jedoch eine entsprechende Schutzummantelung benötigt. [Kulenkampff u. a. (2019), S.36 ff.]

Sollen nun zwei Standorte durch eine Glasfaserleitung verbunden werden, muss die Trassenführung festgelegt werden. Dabei ist es nicht so, dass der kürzeste Weg, also die Luftlinie, zwischen den beiden Standorten gewählt wird. In der Praxis wäre dies aufgrund von geographischen und städtischen Hindernissen, wie zum Beispiel Flüssen oder existierenden Gebäuden oftmals nicht möglich. Zudem ist dieser Ansatz nicht kostensparend, da die meisten Kosten bei den Tiefbauarbeiten anfallen. Diese wären in dem Fall des Baus einer komplett neuen Strecke zwischen zwei Standorten sehr hoch. [Kulenkampff u. a. (2019), S.40]
Um also Kosten zu sparen, werden die Tiefbauarbeiten auf ein Minimum reduziert. Die Standorte werden an den nächstgelegenen Anschlusspunkt eines bestehenden Netzes angebunden. Durch die so entstehenden Synergien, kann ein Kosten reduzierender Ausbau des Breitbandnetzes ermöglicht werden. [Breitband Kompetenz Zentrum, Niedersachsen]
Die Verbindung zweier Standorte unter Nutzung eines bestehenden Netzes veranschaulicht die Abbildung 2.2. Auf der linken Seite ist das bestehende Netz dargestellt. Das Ziel ist es die lila eingefärbten Orte A und B miteinander zu verbinden. Auf der rechten Seite der Abbildung ist eine mögliche Lösung der Verbindung veranschaulicht. Beide Standorte werden an den nächstgelegenen Anschlusspunkt des bestehenden Netzes angebunden. Die gestrichelte Linie verdeutlicht die Strecke der Tiefbauarbeiten. Wie die Anbindung einer neuen Leitung an einen bestehenden Anschlusspunkt aussieht, ist in Abbildung 2.3 dargestellt. Bei dem gebündelten Strang Glasfaserleitungen handelt es sich um das bereits bestehende Netz. Die zwei Glasfaserleitungen daneben verdeutlichen die neue Anbindung. In diesem Beispiel wurde eine Rohr-in-Rohr Verlegung angewandt.

Ferner ist zu erwähnen, dass eine Leitung einen Anfangspunkt und einen Endpunkt besitzt. Dabei ist irrelevant, welches Ende als Startpunkt betitelt wird. Jedem dieser Standorte ist eine Standortnummer zugeordnet, welche durch das Unternehmen vergeben wurde. Zudem besitzt jeder Standort eine Adresse.

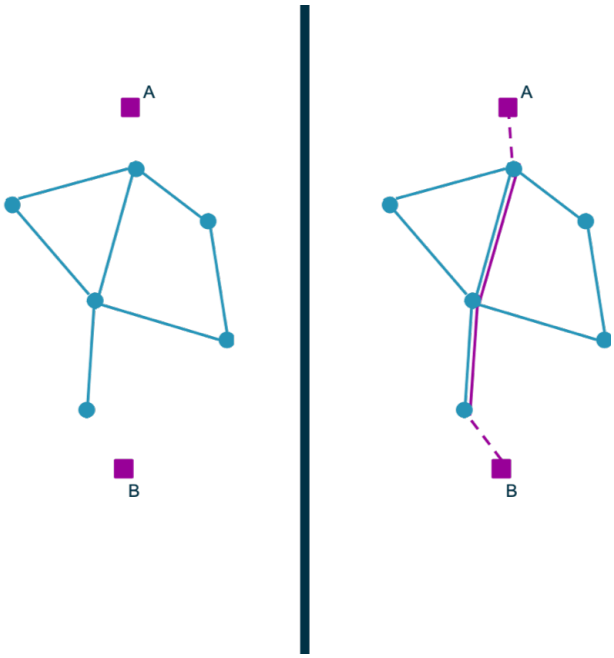


Abbildung 2.2: Verbindung zweier Standorte unter Nutzung eines bestehenden Netzes.

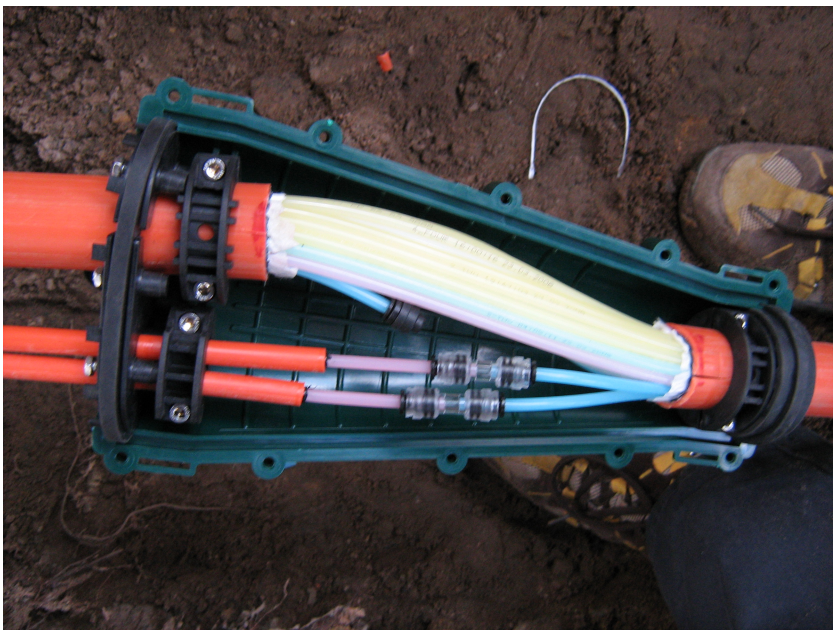


Abbildung 2.3: Anbindung einer neuen Leitung an einem bestehenden Anschlusspunkt. Quelle: Telefónica Germany GmbH und Co. OHG (2009)

2.1.2 Kostenmodell einer Glasfaserleitung

Der Preis des Baues eines neuen Glasfasersegmentes hängt von einigen Faktoren ab. Das wissenschaftliche Institut für Infrastruktur und Kommunikationsdienste ("WIK Consult") hat sich im Rahmen der Aufstellung eines analytischen Kostenmodelles im Anschlussnetz mit solchen Faktoren beschäftigt. In dem folgenden Teil wird darauf eingegangen, um welche Faktoren es sich dabei handelt. Der Fokus liegt dabei auf den wesentlichen Faktoren. Eine vollständige Erläuterung des Kostenmodelles einer Glasfaseranschlussleitung befindet sich in dem Referenzdokument "Analytisches Kostenmodell für das Anschlussnetz AKM-AN Version 3.0" [Kulenkampff u. a. (2019)].

Bei dem Bau einer Glasfasertrasse ist die Länge der Leitung ein ausschlaggebender Faktor. Je nach Länge der Leitung kommen Kosten für das Medium und für Kabelverbindungen, wie zum Beispiel der Verbindungsmuffen und das Spleißen, auf. [Kulenkampff u. a. (2019), S.34]

Allerdings ist "Der Tiefbau [...] für den weitaus größten Teil der Investitionen in eine Anschlussnetzinfrastruktur verantwortlich." [Kulenkampff u. a. (2019), S.40] Die Tiefbauarbeiten erfordern das Ausheben der Erdoberfläche und das anschließende Einfüllen eines Sandbettes zur Verlegung der Leitung. Anschließend findet der Rückbau des Grabens statt. Um den Graben wieder aufzufüllen kann das ausgehobene Material genutzt werden. Sollte dies nicht möglich sein, muss Ersatzmaterial besorgt werden. Anschließend muss die Oberfläche wiederhergestellt werden. Die Kosten der Oberflächenarbeit hängen von der Art der Oberfläche ab.
Es wird zwischen folgenden Oberflächentypen unterschieden:

- Grünfläche
- Asphalt (Gehweg)
- Asphalt (Fahrbahn)
- Pflaster
- Platten

Je nach Oberflächentyp ändern sich die Kosten. [Kulenkampff u. a. (2019), S.40]

Außerdem werden die Kosten der Tiefbauarbeiten durch einen weiteren Faktor beeinflusst. "Bei Aushubarbeiten beeinflusst die Bodenart die Preise der Tiefbauleistungen"

[Kulenkampff u. a. (2019), S.89]. Es existiert ein Unterschied, ob der Boden, der ausgehoben werden soll aus Fels oder Sand besteht. Um diesen Aspekt in dem Kostenmodell zu berücksichtigen werden beim "WIK" die Bodenklassen der DIN Norm 18300 herangezogen.
Die DIN Norm 18300 teilt die Böden in sieben Klassen ein, die sogenannten Bodenklassen. Eine Aufteilung geschieht nach der Lösbarkeit der Böden und nach erdbautechnischen Eigenschaften. Die Aufteilung der Bodenklassen ist in Abbildung 2.4 dargestellt. [Landesamt, für Bergbau, Energie und Geologie]
Die unterschiedlichen Bodenklassen beeinflussen die Kosten des Tiefbaus. [Kulenkampff u. a. (2019), S.89]

Bodenklassen nach DIN 18300	
Bodenklasse 1	Oberboden
Bodenklasse 2	Fließenden Bodenarten
Bodenklasse 3	Leicht lösbare Bodenarten
Bodenklasse 4	Mittelschwer lösbare Bodenarten
Bodenklasse 5	Schwer lösbare Bodenarten
Bodenklasse 6	Leicht lösbarer Fels und vergleichbare Bodenarten
Bodenklasse 7	Schwer lösbarer Fels

Abbildung 2.4: Bodenklassen nach DIN 18300. Quelle: Landesamt, für Bergbau, Energie und Geologie

2.2 Der Beschaffungsprozess

Der Beschaffungsprozess wird in dem "Resource Lifecycle Management"-Prozess von Christian Czarnecki beschrieben. Der Begriff "resource" entspricht physischen Elementen in der technischen Infrastruktur. Dabei handelt es sich um Netzkomponenten, wie zum Beispiel eine Glasfaserleitung.
Der Prozessablauf ist in dem "Business Process Model and Notation"-Diagramm (kurz: BPMN) in der Abbildung 2.5 dargestellt. Zunächst wird die Verbindung geplant ("Resource Planning"). Daraufhin wird ein Entwurf konstruiert ("Resource Development"), welcher anschließend umgesetzt wird ("Resource Implementation"). Danach kann die Leitung in Betrieb genommen werden ("Resource Operations").

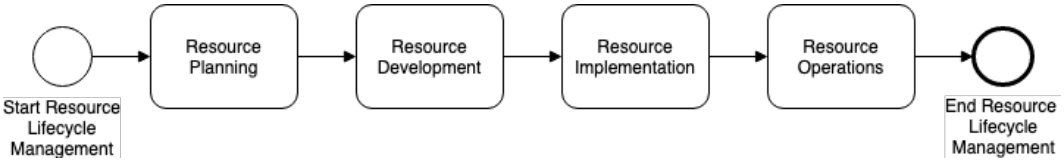


Abbildung 2.5: Darstellung des "Resource Lifecycle Management"-Prozesses. Quelle: Czarnecki und Dietze (2017), S.165 Abb. 4.56 modifiziert

Physische Elemente in der technischen Infrastruktur werden auch von externen Lieferanten, wie Glasfaser-Infrastruktur-Betreibern, angeboten. Somit besteht die Möglichkeit für Telekommunikationsunternehmen Glasfasersegmente anzumieten. Die in dem BPMN vorgestellten Vorgänge bleiben trotzdem weiterhin in dem Unternehmen bestehen. Denn bevor ein Telekommunikationsunternehmen ein entsprechendes Angebot annimmt wird die Verbindung geplant und ein Entwurf wird konstruiert ("Resource Planning and Devolpment"). Daraufhin muss das Telekommunikationsunternehmen ein entsprechenden Angebot finden, abschließen und die Leitung an das eigene Netz anbinden ("Resource Implementation"). Daran schließt der Vorgang "Resource Operations" an. [Czarnecki und Dietze (2017)]

Somit existieren mehrere Varianten für Telekommunikationsunternehmen um den Bedarf an Glasfaser zu erfüllen. Zum einen können die Leitungen selber verlegt werden, zum anderen können sie die Leitungen bei externen Lieferanten anmieten. Außerdem ist eine Kombination der beiden genannten Varianten möglich.
Das Geschäftsmodell des betrachteten Telekommunikationsunternehmens sieht vor, dass wenn eine Anbindung zweier Standorte benötigt wird, die benötigten Glasfasersegmente vorwiegend angemietet werden. Die externen Lieferanten werden im nachfolgenden Teil als Lieferanten oder Carrier betitelt.
Ein Vertrag über eine gemietete Leitung sieht meistens vor, dass zunächst ein einmaliger Preis gezahlt wird. Zudem wird über die gesamte Dauer der Mietzeit ein monatlicher Preis gezahlt. In beiden Preisen sind die Betriebskosten und Bereitstellungskosten der Leitung enthalten.

2.3 Maschinelles Lernen

Maschinelles Lernen ist eine Disziplin, bei der Modelle entworfen werden, die mit Daten trainiert werden, um anschließend für neue Daten Vorhersagen treffen zu können. Es werden also vorhandene Daten für die Entwicklung eines Modells genutzt, welches anschließend für neue Daten ein Ergebnis vorhersagen kann. [Grus (2016), S.150]
Im Kontext dieser Bachelorarbeit soll ein solches Modell für die Vorhersage eines marktkonformen Mietpreises für eine Glasfaserleitung entworfen werden.
Das maschinelle Lernen teilt sich grundsätzlich in zwei Ansätze. In das überwachte Lernen und in das unüberwachte Lernen. Der Unterschied besteht darin, dass beim überwachten Lernen der Trainingsdatensatz bereits den erwünschten Ausgabewert beinhaltet. Beim unüberwachten Lernen wird ein Datensatz zum Trainieren übergeben, welcher den erwünschten Ausgabewert nicht enthält. [Grus (2016), S.150] [Raschka und Mirjalili (2018)]
Im Folgenden wird auf den Ansatz des überwachten Lernens eingegangen.

2.3.1 Grundlagen des überwachten Lernens

Für das überwachte Lernen wird ein gekennzeichneter Trainingsdatensatz benötigt, ein Datensatz in dem der erwünschte Ausgabewert bereits enthalten ist. Dieser Ausgabewert wird als Antwortvariable betitelt. Die übrigen Variablen des Datensatzes werden als Prädiktorvariablen bezeichnet. Das Ziel ist es herauszufinden, wie der Zusammenhang zwischen den Prädiktorvariablen und der Antwortvariable ist. Die Prädiktorvariablen werden dabei auch oft als unabhängige Variablen bezeichnet, während die Antwortvariable als abhängige Variable betitelt wird. [van der Aalst (2016), S.92 f.]

Um den Zusammenhang der Variablen herauszufinden, können unterschiedliche Verfahren angewandt werden. Das überwachte Lernen lässt sich in zwei Verfahren unterteilen, die Klassifikation und die Regression. Die Wahl des Verfahrens entscheidet sich nach der Skalierung der Antwortvariable. Die Antwortvariable kann numerisch oder kategorial sein. Handelt es sich um einen numerischen Wert, werden Regressionsverfahren angewandt. In dem folgenden Abschnitt werden einige Regressionsverfahren vorgestellt.
Ist die Antwortvariable kategorial, werden Klassifikationsverfahren angewandt. Es existieren mehrere Klassifikationsverfahren, einige werden in dem folgendem Abschnitt vorgestellt. [van der Aalst (2016), S.93]

2.3.2 Regressionsverfahren

Bei Regressionsverfahren wird eine Funktion gesucht, die den Zusammenhang zwischen den Prädiktorvariablen und der Antwortvariable am Besten aufzeigt. Anhand dieser Funktion können dann Vorhersagen getroffen werden.

Lineare Regression

Die bekannteste Regression ist die lineare Regression. Voraussetzung für die lineare Regression ist, dass nur eine Prädiktorvariable existiert und diese in einem linearen Zusammenhang mit der Antwortvariable steht. Die Funktion der linearen Regression ist gegeben durch:

$$f(x) = w_0 + w_1 x \tag{2.1}$$

Diese Funktion ist in einem Koordinatensystem darstellbar. Die Konstante w_0 stellt den Achsenabschnitt dar und die Konstante w_1 beschreibt die Steigung der Funktion.
Die Bestimmung der Konstanten erfolgt über die Minimierung der Verlustfunktion:

$$\frac{1}{2n} \sum_{i=1}^{n} (y_\mathrm{i} - f(x))^2 \tag{2.2}$$

Dabei ist n die Anzahl der Instanzen im Datensatz und y der Wert der abhängigen Antwortvariable.
Sind alle Konstanten bestimmt kann für jedes x eine Vorhersage getroffen werden. Dafür wird der Wert x in die Formel 2.1 eingesetzt und *f(x)* berechnet. [Hand u. a. (2001), S.67 ff.] [Pedregosa u. a. (2011), S.2825 ff.]

Multiple lineare Regression

Die multiple Regression unterscheidet sich zu der linearen Regression insofern, dass eine Mehrzahl an Prädiktorvariablen vorhanden sein können. Aus diesem Grund wird die Funktion der linearen Regression erweitert und sieht dementsprechend folgendermaßen aus:

$$f(x_\mathrm{i}) = w_0 + \sum_{i=1}^{m-1} w_\mathrm{i} x_\mathrm{i} \tag{2.3}$$

Statt der Eingabe eines einzelnen x-Wertes ist es nun möglich beliebig viele x-Werte einzusetzen. Die Anzahl i der x-Werte ist bestimmt durch die Anzahl der Prädiktorvariablen.

Eine Darstellung der Funktion im zweidimensionalen Raum ist nicht mehr möglich. Die Punkte befinden sich im m-dimensionalen Raum. Die Zahl m korrespondiert mit der Anzahl der Variablen, wobei es sich bei m-1 um die Anzahl der Prädiktorvariablen handelt. Die Konstanten werden auch in diesem Fall mithilfe der Verlustfunktion bestimmt. Dies funktioniert nur, wenn die Prädiktorvariablen untereinander nicht korrelieren. Sollten die Prädiktorvariablen untereinander korrelieren, gibt es spezielle Regressionsverfahren, die mit solch einer Situation umgehen können. Ein Beispiel für ein solches Verfahren ist die Ridge Regression. [Hand u. a. (2001), S.367 ff.] [Pedregosa u. a. (2011), S.2825 ff.]

Ridge Regression

Die Ridge Regression ist ähnlich zur multiplen linearen Regression. Die Formel der Vorhersagefunktion bleibt dieselbe. Allerdings wird zur Bestimmung der Konstanten die Verlustfunktion um einen Strafterm (blau markiert) erweitert. Die gesamte Formel sieht folgendermaßen aus:

$$\frac{1}{2n}\sum_{i=1}^{n}(y_\mathrm{i} - f(x_\mathrm{i})^2 + \lambda \sum_{j=1}^{m} {w_\mathrm{j}}^2 \tag{2.4}$$

Der Parameter λ gibt die Größe der Bestrafung an. Je größer λ, desto größer die Bestrafung und desto stabiler werden die Koeffizienten gegenüber möglicher Korrelation. Diese Regularisierung ist auch bekannt als L_2 - Regularisierung.
Die Ridge Regression wird eingesetzt, wenn die Prädiktorvariablen untereinander korrelieren und eine Überanpassung (engl. Overfitting) an die Trainingsdaten vermieden werden soll. [Pedregosa u. a. (2011), S.825 ff.]

2.3.3 Klassifikationsverfahren

Das Ziel von Klassifikation ist es Instanzen zu klassifizieren, basierend auf den Prädiktorvariablen. [van der Aalst (2016), S.93]

Entscheidungsbaum

"Ein Entscheidungsbaum verwendet eine Baumstruktur, um mögliche Entscheidungspfade und ein Ergebnis für jeden Pfad zu speichern."[Grus (2016), S.213] Es werden hierarchisch aufeinanderfolgende Entscheidungen in den Knoten des Baumes dargestellt. Jeder Knoten fragt eine Entscheidung hinsichtlich der Ausprägung der Prädiktorvariablen ab. Durch eine Entscheidung teilt ein Knoten die Menge von Instanzen in beliebig

viele kleinere Untermengen. Die Blätter des Baumes beinhalten mögliche Werte der Antwortvariablen. [van der Aalst (2016), S.94]

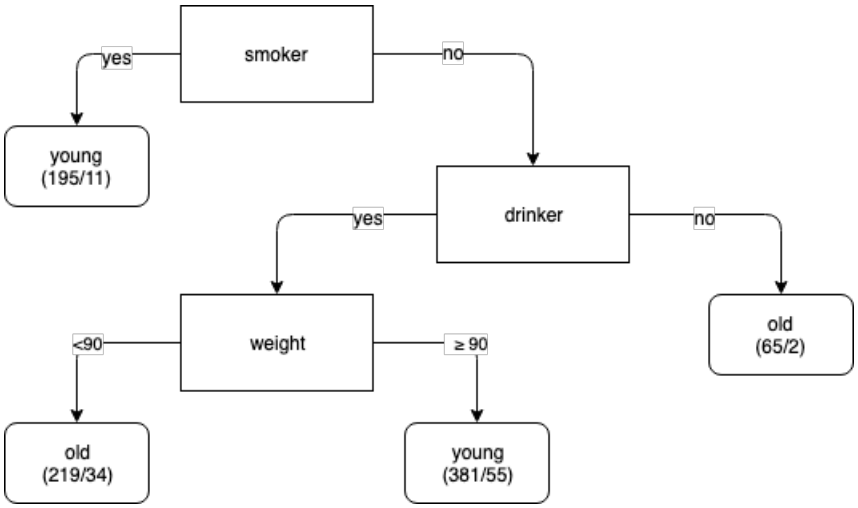


Abbildung 2.6: Entscheidungsbaum zum Sterbealter einer Person. Quelle: van der Aalst (2016), S.95 Abb. 4.1 modifiziert

Die Abbildung 2.6 zeigt einen beispielhaften Entscheidungsbaum. Der abgebildete Entscheidungsbaum trifft eine Entscheidung darüber, ob eine Person im jungen Alter stirbt oder nicht. Diese Entscheidung hängt von den Werten der Prädiktorvariablen ab, welche angeben, ob eine Person raucht, Alkohol trinkt und wie viel sie wiegt. Je nach Ausprägung dieser Prädiktorvariablen wird ein Entscheidungspfad gewählt. In dem Blattknoten in dem der Entscheidungspfad endet ist das Ergebnis hinterlegt. Die Blätter des Baumes beinhalten zwei Werte. Die erste Zahl gibt an, wie viele Instanzen als solche klassifiziert wurden. Der zweite Wert gibt an, wie viele Instanzen davon falsch klassifiziert wurden.
Eine Person, die nicht raucht, Alkohol trinkt und über 90kg wiegt, wird, entsprechend der Vorhersage des Entscheidungbaums, in jungem Alter sterben. 381 Personen wurden dieser Klasse zugeordnet, während 55 davon falsch klassifiziert wurden. [van der Aalst (2016), S.94 ff.]

Das Aufstellen eines Entscheidungsbaumes erfordert das Treffen vieler Entscheidungen. Zum Beispiel muss entschieden werden, in welcher Reihenfolge die Prädiktorvariablen abgefragt werden. Diese Wahl kann nach unterschiedlichen Methoden geschehen. Oftmals wird die Entropie genutzt.
Die Entropie ist ein Maß für die Unsicherheit in einem Datensatz mit Elementen. An-

genommen man hat einen Datensatz X mit n Elementen und k möglichen Werten mit v_1,v_2,...v_k. Der Wert c_i beschreibt wie oft der Wert v_i vorkommt. Wenn alle Elemente n aus dem Datensatz X den gleichen Wert annehmen, gibt es keinerlei Unsicherheit. Die Entropie ist gering.
Würden die Elemente n gleichmäßig über die k möglichen Werte verteilt sein, gäbe es eine Menge Unsicherheit und die Entropie wäre hoch.
Die Entropie für einen Datensatz X mit n Werten lässt sich folgendermaßen berechnen:

$$E = -\sum_{i=1}^{k} p_\mathrm{i} \cdot log_2 \cdot p_\mathrm{i} \tag{2.5}$$

Wobei p_i die Wahrscheinlichkeit ist mit der ein Element den Wert v_i annimmt. [van der Aalst (2016), S.97 ff.]

Der Entscheidungsbaum bietet einige Vorteile. Zum einen ist das Modell des Entscheidungsbaums leicht zu interpretieren und weist einen transparenten Vorhersageprozess auf. Zum anderen kann das Modell des Entscheidungsbaums mit numerischen, sowie kategorialen Daten arbeiten.
Der Nachteil ist, dass es schwierig ist, einen optimalen Entscheidungsbaum mit den Trainingsdaten zu finden, denn das Modell kann sich sehr gut an die Trainingsdaten anpassen und lässt sich dann schwer Generalisieren. Dieses Phänomen wird als Überanpassung (engl. Overfitting) bezeichnet. [Grus (2016), S.214]

Random Forest

Um das Problem der Überanpassung (engl. Overfitting) von Entscheidungsbäumen zu umgehen wird oftmals das Random Forest Verfahren verwendet. Ein Random Forest besteht aus mehreren Entscheidungsbäumen, die sich untereinander unterscheiden. Jeder der Entscheidungsbäume sagt eine Klasse vorher. Am Ende stimmen die Entscheidungsbäume die Klassifikation eines Elementes ab. Dabei wird für die Klasse entschieden die am häufigsten vorhergesagt wurde.

Der Random Forest ist eine Technik, die unter das Anwendungsfeld "Ensemble Learning" fällt. Beim "Ensemble Learning" werden unterschiedliche Lernverfahren miteinander kombiniert, um ein stabiles Modell zu erhalten. [Grus (2016), S.223 f.]

3 Analyse der Preise

In diesem Kapitel geht es um die Analyse der Mietpreise einer Glasfaserleitung. Als Erstes werden in Abschnitt 3.1 die Anforderungen an die Analyse vorgestellt. In Abschnitt 3.2 werden Hypothesen hinsichtlich der Einflussfaktoren und der Preisentwicklung aufgestellt, die später auf ihre Gültigkeit untersucht werden. Die Gültigkeit der Hypothesen kann nur anhand von Daten überprüft werden, die passende Informationen enthalten. Welche Daten für die Analyse genutzt werden und woher diese Daten stammen ist Bestandteil des Abschnittes 3.3.
Im nachstehenden Abschnitt 3.4 wird darauf eingegangen wie die Analyse durchgeführt und die Hypothesen überprüft wurden. Zudem wird das Ergebnis dargestellt.

3.1 Anforderungen an die Analyse

Bei der Preisanalyse sollen zwei Fragen hinsichtlich des Preises einer Glasfaserleitung untersucht werden. Folgende Fragestellungen sollen überprüft werden:

1. Welche Faktoren beeinflussen den Preis einer Leitung?
2. Existiert ein zeitlicher Trend der Preise?

Dabei soll der monatliche und der einmalige Preis berücksichtigt werden.
Eine weitere Anforderung ist, dass die Ergebnismenge der Einflussfaktoren ausschließlich Merkmale enthalten soll, die einen mittelmäßigen bis starken Einfluss auf den Preis besitzen. Merkmale, die einen zu geringen Einfluss haben, würden keinen relevanten Mehrwert für ein Preisvorhersagemodell bieten und nur die Komplexität des Modelles erhöhen.

3.2 Aufstellung der Hypothesen

Im Folgenden werden Hypothesen aufgestellt, die eine mögliche Antwort auf die Fragestellungen darstellen. Diese Hypothesen werden auf ihre Gültigkeit geprüft. Bei einem Hypothesentest gilt es die Nullhyphotese H_0 zu widerlegen um nachzuweisen, dass die Alternativhypothese H_1 gilt. H_0 und H_1 müssen sich deshalb widersprechen, wobei die nachzuweisende Behauptung in der Alternativhypothese H_1 formuliert wird. Zu den einzelnen Hypothesen wird außerdem eine Begründung dargelegt, warum diese Vermutung eintreffen sollte. [Hand u. a. (2001), S.112 ff.]

3.2.1 Hypothesen zu den Einflussfaktoren auf den Preis

Um zu untersuchen, welche Faktoren den Preis einer Leitung beeinflussen, werden Zusammenhangshypothesen aufgestellt. Dabei wird zwischen einen positiven, einen negativen und keinen Zusammenhang unterschieden. Ein positiver Zusammenhang beschreibt die Situation, dass höhere Werte des einen Merkmales mit höheren Werten des anderen Merkmales einhergehen. Ein negativer Zusammenhang dagegen beschreibt den Fall, dass höhere Werte des einen Merkmales mit niedrigeren Werten des anderen Merkmales einhergehen.

Hypothese 1.1

Es wird angenommen, dass die Länge einer Leitung den Mietpreis beeinflusst. Dies wird vermutet, da die Länge einen Einfluss auf die Kosten des Trassenbaus hat.
Außerdem beansprucht die Wartung einer längeren Leitung mehr Zeit und ist somit kostenintensiver für die Lieferanten. Die Länge hat aus diesem Grund vermutlich einen positiven Einfluss auf den monatlichen Preis, sowie auf den einmaligen Preis. Es werden zwei Hypothesen abgeleitet.

Hypothese 1.1.1:

H_0 = *Die Länge einer Leitung hat keinen positiven Einfluss auf den monatlichen Preis.*

H_1 = *Die Länge einer Leitung hat einen positiven Einfluss auf den monatlichen Preis.*

Hypothese 1.1.2:

H_0 = *Die Länge einer Leitung hat keinen positiven Einfluss auf den einmaligen Preis.*

H_1 = *Die Länge einer Leitung hat einen positiven Einfluss auf den einmaligen Preis.*

Hypothese 1.2

Ferner wird angenommen, dass die geographische Lage der Leitungen den Mietpreis bestimmt. Begründet ist dies dadurch, dass die Wiederherstellung der Oberfläche nach der Verlegung einen wesentlichen Kostenfaktor beim Bau der Trasse ausmacht. Es wird vermutet, dass eine Wiederherstellung der Oberfläche in der Innenstadt höhere Kosten erzeugt, da die innerstädtische Oberfläche eher aus Asphalt, Pflaster oder Platten besteht. Wohingegen bei einer ländlichen Gegend eher mit einer Grünfläche zu rechnen ist. Die Kosten zur Wiederherstellung einer Grünfläche werden vermutlich geringer sein.
Hinzu kommt, dass die Tiefbauarbeiten und somit die Oberflächenarbeiten in der Stadt vermutlich eine geringe Fläche abdecken, da der nächstgelegene Anschlusspunkt eines bestehenden Netzes in der Stadt vermutlich dichter ist, als es auf dem Land der Fall ist. Die anfallenden Kosten der Oberflächenarbeiten werden sich vermutlich in dem Mietpreis widerspiegeln, in dem einmaligen Preis, sowie in dem monatlichen Preis. Aus diesem Grund kommen zwei Hypothesen zustande.

Hypothese 1.2.1:

H_0 = *Die geographische Lage einer Leitung hat keinen Einfluss auf den monatlichen Preis.*

H_1 = *Die geographische Lage einer Leitung hat Einfluss auf den monatlichen Preis.*

Hypothese 1.2.2:

H_0 = *Die geographische Lage einer Leitung hat keinen Einfluss auf den einmaligen Preis.*

H_1 = *Die geographische Lage einer Leitung hat Einfluss auf den einmaligen Preis.*

Hypothese 1.3

Außerdem wird angenommen, dass die Bodenklasse einen Einfluss auf die Preise hat. Der Grund für diese Vermutung ist, dass die Kosten für den Leitungsbau durch die Kosten der Tiefbauarbeiten beeinflusst werden und die Kosten der Tiefbauarbeiten je nach Bodenklasse variieren. "WIK Consulting" zog zur Evaluierung der Böden, die Bodenklassen nach DIN 18300 heran. Eine Einteilung der Böden in die Bodenklassen ist in Abbildung 2.4 dargestellt. Es wurde bewiesen, dass die Bodenklassen nach DIN 18300 einen Einfluss auf die Kosten des Trassenbaus besitzen. Je höher eine Bodenklasse, umso steiniger und felsiger der Boden. Dies spiegelt sich positiv in den Kosten der Tiefbauarbeiten wider. [Kulenkampff u. a. (2019), S.89 f.]
Es wird angenommen, dass sich dies nicht nur in den Kosten der Leitung widerspiegelt, sondern auch in den monatlichen und einmaligen Mietpreisen einer Leitung.

Hypothese 1.3.1:

H_0 = *Die Höhe der Bodenklasse hat keinen positiven Einfluss auf den monatlichen Preis.*

H_1 = *Die Höhe Bodenklasse hat einen positiven Einfluss auf den monatlichen Preis.*

Hypothese 1.3.2:

H_0 = *Die Höhe der Bodenklasse hat keinen positiven Einfluss auf den einmaligen Preis.*

H_1 = *Die Höhe der Bodenklasse hat einen positiven Einfluss auf den einmaligen Preis.*

Hypothese 1.4

Eine weitere Vermutung ist, dass der Wettbewerb die Preise beeinflusst. Ein Lieferant, der in einem Gebiet alleine agiert und Leitungen anbietet, kann diese zu einem höheren Preis anbieten. Währenddessen er seine Produkte zu einem geringeren Preis anbieten muss, wenn der Wettbewerb höher ist, damit der Kunde das Produkt kauft. Dieses Prinzip wird als Marktgleichgewicht bezeichnet. Angebot und Nachfrage befinden sich immer im Gleichgewicht. Befindet sich auf dem Markt zum Beispiel eine höhere Anzahl an Wettbewerbern existieren mehrere Angebote und daraus resultiert ein geringerer Preis. [Pollert u. a. (2016)]
Daher lassen sich zwei Hypothesen herleiten.

Hypothese 1.4.1:

H_0 = *Die Anzahl der Wettbewerber hat keinen negativen Einfluss auf den monatlichen Preis.*

H_1 = *Die Anzahl der Wettbewerber hat einen negativen Einfluss auf den monatlichen Preis.*

Hypothese 1.4.2:

H_0 = *Die Anzahl der Wettbewerber hat keinen negativen Einfluss auf den einmaligen Preis.*

H_1 = *Die Anzahl der Wettbewerber hat einen negativen Einfluss auf den einmaligen Preis.*

Hypothese 1.5

Als weiterer Einflussfaktor wird die Vertragsdauer angenommen. Es wird vermutet, dass ein Zusammenhang zwischen dem Preis und der Vertragsdauer existiert. Je länger die Vertragsdauer ist, desto länger ist das Telekommunikationsunternehmen an den Lieferanten gebunden. Dies könnte zur Folge haben, dass der Preis geringer ist, da der Lieferant über einen längeren Zeitraum feste Einnahmen erhält. Es werden zwei Hypothesen aufgestellt:

Hypothese 1.5.1:

H_0 = *Die Länge der Vertragsdauer hat keinen negativen Einfluss auf den monatlichen Preis.*

H_1 = *Die Länge der Vertragsdauer hat einen negativen Einfluss auf den monatlichen Preis.*

Hypothese 1.5.2:

H_0 = *Die Länge der Vertragsdauer hat keinen negativen Einfluss auf den einmaligen Preis.*

H_1 = *Die Länge der Vertragsdauer hat einen negativen Einfluss auf den einmaligen Preis.*

3.2.2 Hypothesen zu der Preisentwicklung

Um zu untersuchen, wie sich der Preis zeitlich entwickelt und ob ein Trend existiert, werden Veränderungshypothesen aufgestellt.

Hypothese 2.1

Es wird vermutet, dass der Preis einer Leitung einem zeitlichen Trend folgt. Dies sei darin begründet, dass der Netzausbau die Nachfrage nach Glasfasertrassen steigen lässt. Steigt die Nachfrage nach einem Produkt, so steigt auch der Preis des Produktes. Dieser Grundsatz ist in dem Prinzip des Marktgleichgewichtes beschrieben. [Pollert u. a. (2016)]

Hypothese 2.1.1:

H_0 = *Der monatliche Preis folgt keinem zeitlichen Trend.*

H_1 = *Der monatliche Preis folgt einem zeitlichen Trend.*

Hypothese 2.1.2:

H_0 = *Der einmalige Preis folgt keinem zeitlichen Trend.*

H_1 = *Der einmalige Preis folgt einem zeitlichen Trend.*

3.3 Datenaufbereitung

Im folgenden Abschnitt wird erläutert, aus welcher Datenquelle die Daten, auf die sich die Analyse stützt, extrahiert wurden. Weiterhin wird darauf eingegangen, wie die Daten für die Analyse aufbereitet wurden und welche Methoden bei der Datenbereinigung Anwendung fanden.

3.3.1 Datenerhebung und Integration

Primär stammen die Daten aus einer internen Datenbank des betrachteten Telekommunikationsunternehmens, in der alle Leitungen des Unternehmens hinterlegt sind. Ein Datensatz beschreibt eine Leitung. Folgende Informationen zu einer Leitung wurden extrahiert:

- Der Lieferant

  Bei dem Lieferanten handelt es sich um den Carrier, von dem die Glasfaserleitung bezogen wird.

- Das Produkt

  Das Produkt gibt den Übertragungskanal an.

- Der einmalige Preis

  Der einmalige Preis ist der Preis, welcher einmalig am Anfang des Mietverhältnisses für die Leitung bezahlt wurde.

- Der monatliche Preis

  Der monatliche Preis ist der Preis, welcher monatlich, seit Beginn des Mietverhältnisses, für die Leitung bezahlt wird.

- Kostenpflichtig ab

  Das Datum an dem das Mietverhältnis begann.

- Mietzeitbindung

  Das Datum bis zu dem das Mietverhältnis besteht.

- Die Adressen

  Da eine Leitung zwei Enden besitzt gehören zu jeder gemieteten Leitung zwei Adressen. Die Kennzeichnung wird mit A und B erweitert.

- Die Standortnummern

  Da eine Leitung zwei Enden besitzt gehören zu jeder gemieteten Leitung zwei Standortnummern. Die Kennzeichnung wird mit A und B erweitert.

Anschließend wurde eine Filterung der Daten vorgenommen. Es wurde sich nur auf angemietete Leitungen des betrachteten Telekommunikationsunternehmens beschränkt. Außerdem wurden die Leitungen, bei denen das Telekommunikationsunternehmen selbst als Lieferant eingetragen ist, entfernt, da es sich bei diesen Leitungen um Eigentum des Unternehmens handelt. Ferner wurden nur Leitungen aus Glasfaser betrachtet, die sich innerhalb Deutschlands befinden.
Da diese Daten für die Überprüfung der Hypothesen nicht ausreichend waren, wurden weitere Informationen zu den Leitungen erhoben.

Zum einen wurde die Area Class zu einer Leitung hinzugefügt. Die Area Class kann folgende Werte annehmen: suburban, urban oder rural. Jedoch geht dabei der Wert nicht mit der Siedlungsgeographie eines Standortes einher. Die Area Class bezieht sich in diesem Fall auf die Anzahl der Personen, die sich während eines bestimmten Zeitraums an dem jeweiligen Standort eingewählt haben. So könnte es durchaus sein, dass es sich bei einem urbanen Standort nicht um einen städtischen Raum handelt, sondern zum Beispiel um eine eher ländliche Gegend in der eine Autobahn verläuft. Da an diesem Autobahnstreckenstück viel Verkehr ist und sich dementsprechend viele Personen an dem Punkt einwählen, ist dem Standort der Wert urban zugeordnet. Die Area Class wurde pro Standortnummer erhoben. Da jede Leitung zwei Standortnummern besitzt wurden jeder Leitung die zwei entsprechenden Area Classes zugeordnet.
Im Kontext dieser Arbeit wird zudem angenommen, dass eine Leitung mindestens 50m lang ist. Aus diesem Grund ist zu erwarten, dass eine Leitung durch unterschiedliche Area Classes verläuft. Dies wurde aber nicht weiter berücksichtigt, da dafür entsprechende Informationen nicht vorhanden sind.

Zusätzlich werden Informationen zu den Bodenklassen an den jeweiligen Standorten benötigt, um die Hypothese 1.3 überprüfen zu können. Da es intern keine Informationen zu

den Bodenklassen in Deutschland gibt und auch keine frei zugänglichen Quellen existieren, welche die Bodenklassen in Deutschland aufzeigen, erfolgte die Erhebung der Daten über einen alternativen Vorgang.
Eine Karte der Bodenübersicht von Deutschland von der Bundesanstalt für Geowissenschaften und Rohstoffe diente als Quelle. Diese Karte stellt die Bodentypen in Deutschland dar. Die Karte ist im Anhang A.1 ersichtlich. [Bundesanstalt für Geowissenschaften und Rohstoffe (2014)]
Im ersten Schritt wurde diese Karte in ein GIS System geladen und mit einer Karte der Standortnummern der Leitungen verknüpft. Hierzu wurde das Tool MapInfo [1] verwendet. In MapInfo wurde ein Verschnitt der Karten erstellt und anschließend über eine SQL Abfrage die Information, welche Bodentypen bei den Standortnummern vorherrschen, extrahiert. Die gewonnenen Daten wurden als Excel Datei exportiert. Daraufhin wurde von einem Experten eine Übersetzung der Bodentypen in die sieben Bodenklassen durchgeführt. Die Zuordnung ist im Anhang A.2 dargestellt.
Da die Bodenübersichtskarte die Oberfläche Deutschlands beschreibt und bodenkundliche Informationen bereitstellt, während die Bodenklassen nach bautechnischen Kriterien aufgeteilt sind, könnte eine Ungenauigkeit bei der Übersetzung entstanden sein.
Das "WIK Consulting" legt die Bestimmung eines absoluten Preises für die Bodenklasse drei aus. Eine Anpassung an die anderen Bodenklassen erfolgt über einen Erschwernisfaktor. Da das "WIK Consultig" von einer Veröffentlichung der Liste abgesehen hat, wurde eine Liste der Erschwernisfaktoren der Bodenklassen von einem Experten aus dem Unternehmen verwendet. Diese Liste in der Abbildung 3.1 ersichtlich. Jeder Bodenklassen wurde der entsprechenden Erschwernisfaktoren zugeordnet. Dies Informationen wurden schließlich in dem Datensatz integriert.

Bodenklasse	Erschwernisfaktor
Klasse 2	1
Klasse 3	1
Klasse 4	1
Klasse 5	1
Klasse 6	1,35
Klasse 7	1,6

Abbildung 3.1: Erschwernisfaktoren der Bodenklassen.

[1] Bei MapInfo handelt es sich um eine Geoinformationssystem-Software.

Ferner wurden noch Informationen zu der Besiedlung der einzelnen Orte, die Besiedlungsklasse, dem Datensatz hinzugefügt. Die Besiedlungsklasse gibt den Grad der Besiedlung an einem bestimmten Ort an. Die Daten wurden von einem Experten aus dem Unternehmen zusammengetragen und berechnet. Es wurden pro Postleitzahlengebiet die Einwohnerzahl, die Anzahl der Haushalte und der Gebäude zusammengetragen. Anhand dieser Informationen wurde berechnet, wie viele Personen in einem Haushalt leben und wie viele Haushalte auf ein Gebäude kommen.
Es wurde die Annahme getroffen, dass im innerstädtischen Raum die Anzahl der Personen pro Haushalt gering ist. Diese Annahmen beruht auf der Tatsache, dass die Anzahl der Personen pro Haushalt mit steigender Einwohnerzahl abnimmt und eine hohe Einwohnerzahl für ein städtisches Gebiet spricht. [Statistisches Bundesamt (2018)]
Zudem wurde die Annahme getroffen, dass eine relativ hohe Anzahl an Haushalten pro Gebäude vorwiegend in städtischen Gebieten vorkommt. Diese Annahme beruht auf einer Analyse hinsichtlich der Bevölkerungsdichte und den Haushalten pro Gebäude in Hamburg. Für die Analyse wurden Daten aus dem Telekommunikationsunternehmen verwendet, sowie Daten bezüglich der Einwohnerzahlen und Fläche der Stadtteile Hamburgs von dem statistischen Amt für Hamburg und Schleswig-Holstein. [Statistisches Amt für Hamburg und Schleswig-Holstein (2018)]
Das Ergebnis zeigte, dass eine hohe Bevölkerungsdichte eine relativ hohe Anzahl an Haushalten pro Gebäuden bedingt. Eine hohe Bevölkerungsdichte wiederum spricht für ein innerstädtisches Gebiet. Es wird angenommen, dass dieses Ergebnis auf alle Gebiete in Deutschland übertragbar ist.
Anhand dieser Annahmen wurden die Postleitzahlengebiete in drei unterschiedliche Besiedlungsklassen eingeteilt:

- **Klasse 1**: Weniger als 2 Einwohner pro Haushalt und 5 oder mehr Haushalte pro Gebäude, somit innerstädtischer Raum.
- **Klasse 2**: Weniger als 5 Haushalte pro Gebäude, aber mehr als 2.5, daher ein Gebiet zwischen der Kernstadt und dem Umland.
- **Klasse 3**: Weniger als 2.5 Haushalte pro Gebäude, daher eher Einfamilienhäuser und somit ein ländlicher Raum.

Die Informationen der Besiedlungsklasse wurden in den Datensatz integriert.

Da für die Analyse die Länge einer Leitung benötigt wird, wurden außerdem die Koordinaten der einzelnen Standortnummern dem Datensatz hinzugefügt. Die Koordinaten

sind in Form von Gauß-Krüger Koordinaten angegeben.
Das Gauß-Krüger Koordinatensystem teilt die Erde in 3% breite Meridiansteifen, welche zwischen dem Südpol und dem Nordpol verlaufen. Zu jedem Meridianstreifen gehört ein Mittelmeridian, welcher parallel zu dem jeweiligen Meridianstreifen verläuft. Das Gauß-Krüger Koordinatensystem ermöglicht es jeden Punkt der Erde durch eine Koordinate eindeutig zu verordnen. Eine Gauß-Krüger Koordinate besteht aus einem Hochwert und einem Rechtswert. Der Hochwert gibt an, wie weit der Punkt vom Äquator entfernt ist und der Rechtswert gibt die Entfernung des Punktes zum nächstgelegenen Mittelmeridian an. Beide Werte werden in der Einheit Meter angegeben. Zusätzlich wird dem Rechtswert die Kennziffer des nächstgelegenen Mittelmeridians vorangestellt. Der Vorteil der Gauß-Krüger Koordinaten liegt darin, dass die Krümmung der Erdoberfläche mit beachtet wird. Somit ist über den Satz des Pythagoras der Abstand zwischen zwei Punkten ermittelbar. [Seidel (2006)]

3.3.2 Datenberechnung

Um die Datensätze vergleichen und analysieren zu können, wurden weitere Werte aus den vorhandenen Daten errechnet. Im Folgenden wird erläutert, um welche Werte es sich dabei handelt und wie sie berechnet wurden.

Zum einen wurde die Länge einer Leitung berechnet. Dazu wurden die Gauß-Krüger Koordinaten der Standortnummern verwendet. Die Berechnung der Länge erfolgte über den Satz des Pythagoras. Das Ergebnis beinhaltet die Länge der Luftlinie in Metern zwischen den beiden Standorten. Es ist zu erwähnen, dass wie bereits erläutert, es sich bei der Luftline oftmals nicht um die reelle Trassenführung handelt.

Ferner wurde die Vertragsdauer der Leitungen berechnet, um später einen Vergleich durchführen zu können. Errechnet wurde dieser Wert in dem das Enddatum der Mietzeitbindung vom Datum des Beginns der Mietzeitbindung (Merkmal Kostenpflichtig ab) subtrahiert wurde. Dieser Wert wurde anschließend auf Monate umgerechnet, da eine Darstellung der Vertragsdauer in Tagen als zu detailliert und eine Darstellung in Jahren als zu grob angesehen wird.

Außerdem wurde anhand des Datensatzes die Anzahl der Wettbewerber in den jeweiligen Postleitzahlengebieten bestimmt. Die zweistelligen Postleitzahlengebiete wurden betrachtet. Dabei wurden nur die Lieferanten berücksichtigt bei denen das betrachte-

te Telekommunikationsunternehmen Leitungen gemietet hat. Lieferanten, bei denen das Unternehmen aufgrund eines zu hohen Preises keine Leitungen in dem Gebiet erworben hat oder mit denen das Unternehmen bewusst keine Geschäftsbeziehung eingegangen ist, wurden nicht berücksichtigt. Dieser Umstand könnte dafür Sorgen, dass ein höherer Wettbewerb in den Gebieten besteht als angenommen.
Da dieser Wert pro Postleitzahlengebiet erhoben wurde und eine Leitung zwei Adressen und somit zwei Postleitzahlen besitzt, hat dieses Merkmal zwei Ausprägungen pro Instanz.

Um die Leitungen hinsichtlich der einmaligen Preise vergleichen zu können, wurden die einmaligen Kosten aufgezinst. Hierzu wurde der reale Kapitalzinssatz von 4,87% genutzt. [Bundesnetzagentur (2019), S.98]
Es wurde der Endwert des Betrages zum heutigen Jahr 2019 bestimmt. Dazu wurde die Formel der diskreten Verzinsung angewandt: [Horsch (2018)]

$$K_n = K_0 \cdot (1+i)^n \tag{3.1}$$

wobei:

K_n = Endwert

K_0 = Barwert

i = Zinssatz

n = Anzahl der Jahre

Die Anzahl der Jahre n ist die Differenz zwischen dem Anfang der Vertragsdauer (Merkmal Kostenpflichtig ab) und dem heutigen Jahr 2019.

3.3.3 Datenaggregation

In diesem Abschnitt wird erläutert, wie mit den Merkmalen verfahren wurde, die zwei Ausprägungen pro Instanz besitzen. Dies ist bei den Merkmalen Gewichtungsfaktor der Bodenklasse, Besiedlungsklasse, Wettbewerb und Area Class der Fall. Um die Attribute auf einen Wert abzubilden wurden für die nominalskalierten Merkmale "Übersetzungsmatrizen" erstellt. Bei den Merkmalen Besiedlungsklasse und Area Class handelt es sich um nominal skalierte Daten. Anhand der "Übersetzungsmatrizen" wurden den zwei Ausprägungen ein eindeutiger Wert zugewiesen. Dieser wurde im Datensatz aufgenommen.

Die "Übersetzungsmatrizen" sind in Abbildung 3.2 dargestellt.
Bei den Merkmalen Wettbewerb und den Gewichtungsfaktoren der Bodenklasse handelt es sich um intervallskalierte Merkmale. Dies bietet den Vorteil, dass die Ausprägungen miteinander verrechenbar sind. Es wurde der jeweilige Durchschnitt gebildet.

Area Class	urban	suburban	rural
urban	urban	urban	suburban
suburban	urban	suburban	rural
rural	suburban	rural	rural

Besiedlung	1	2	3
1	1	1	2
2	1	2	3
3	2	3	3

Abbildung 3.2: "Übersetzungsmatrizen" für das Merkmal Area Class (links) und das Merkmal Besiedlungsklasse (rechts).

3.3.4 Datenbereinigung

Die Datenbereinigung bezeichnet den Prozess zum Korrigieren und Entfernen fehlerhafter Daten eines Datensatzes.
Als Erstes wurde der Datensatz auf Vollständigkeit geprüft. Dabei wurde überprüft, ob und wie viele leere Werte, sogenannte "missing values" in dem Datensatz auftraten. Es gibt verschiedene Methoden zum Umgang mit "missing values". Zum Beispiel können die fehlenden Werte durch den Durchschnittswert ersetzt werden. Da diese Methode zwar den Durchschnitt des Merkmales nicht verändert, aber einen Einfluss auf ein Vorhersagemodell hat und diese verfälscht, wurde diese Methode nicht weiter in Betracht gezogen.
Eine weitere Methode ist, dass der fehlende Wert durch einen zufälligen Wert ersetzt wird. Diese Methode hat auch einen negativen Einfluss auf die folgenden statistischen Analysen. Durch das Einsetzen von zufälligen Werten werden Maße, wie die Varianz, verfälscht. Deshalb wurde diese Methode nicht genutzt.
Ferner existiert die Möglichkeit die fehlenden Werte von einem Experten schätzen zu lassen. Da dies einen hohen Aufwand erzeugen wurde diese Methode nicht angewandt. [Han und Kamber (2011), S.88 ff.]
Da es sich um eine geringe Anzahl von fehlenden Werten handelte, bestand außerdem die Variante die unvollständigen Datensätze zu entfernen. Diese Methode fand Anwendung. Der Nachteil in dieser Methode liegt darin, dass durch das Löschen der unvollständigen Datensätze Informationen verloren gehen. Der Vorteil ist, dass diese Variante nicht die statistischen Analysen und das Vorhersagemodell verfälscht.

Als nächstes wurden die Ausreißer aus dem Datensatz entfernt. Bei Ausreißern handelt es sich um Datenwerte, die nicht den Erwartungen entsprechen, die Extremwerte. Um die Ausreißer besser identifizieren zu können, wurde zur Ausreißer-Analyse ein Experte hinzugezogen. Im Zuge dessen wurden alle monatlichen Kosten die 0 EUR oder über 5000 EUR betrugen entfernt, da diese Preise von dem Experten als nicht realistisch eingestuft wurden. Außerdem wurden alle Leitungen entfernt, deren Länge geringer als 50m ist, da es sich bei diesen Leitungen oftmals um Inhausverkabelungen handeln könnte. Zudem wurden alle Leitungen, die länger als 60000m sind, entfernt.
Ferner wurde im Zuge der Datenbereinigung der Preis pro Meter aus den monatlichen Kosten und der Länge berechnet. Es wurden alle Leitungen entfernt, deren Preis pro Meter unter 0,01 EUR/Meter liegt und über 6 EUR/Meter.
Die Datensammlung enthielt anschließend 3501 Zeilen. Ein Ausschnitt, der für die Analyse relevanten Daten, ist im Anhang unter A.3 zu finden.

3.4 Durchführung der Analyse

In diesem Abschnitt wird die Vorgehensweise der Analyse erläutert. Anschließend werden die Ergebnisse [2] vorgestellt.

3.4.1 Analyse der Einflussfaktoren des Preises

Erläuterung der Durchführung

Um die Gültigkeit der Hypothesen überprüfen zu können wurden als Erstes repräsentative Werte ausgewählt. Anschließend wurden Streudiagramme erstellt. Zudem wurden Korrelationsanalysen durchgeführt. Die Korrelationsanalyse ist eine Methode der mathematischen Statistik und wird genutzt um Zusammenhänge zwischen Merkmalen zu berechnen. Je nach Skalierung der Daten existieren unterschiedliche Korrelationskoeffizienten.

[2] Die in der Arbeit enthaltenen Diagramme bilden nicht die reellen Preise ab. Die Preise wurden mit einem zufälligen Faktor multipliziert, aus Gründen der Geheimhaltung.

In diesem Fall wurde für die nominal skalierten Daten im Zuge einer Kontingenzanalyse der Kontingenzkoeffizient berechnet. Bei der Kontingenzanalyse wird als Erstes eine Kontingenztabelle aufgestellt. Diese stellt die Anzahl der Kombinationen der Merkmalsausprägungen dar. Bei der Analyse werden die erwarteten Häufigkeiten mit den beobachteten Häufigkeiten verglichen. Die erwarteten Häufigkeiten errechnen sich aus den Summen der Zeilen und Spalten der Kontingenztabelle. Die Merkmale heißen unabhängig, wenn die erwartete Häufigkeit mit der beobachteten Häufigkeit übereinstimmt. Anschließend wird der Wert Chi-Quadrat durch folgende Formel berechnet:

$$\chi^2 = \sum_{j=1}^{m} \sum_{i=1}^{k} \frac{(h_{\mathrm{ij}} - h^{\mathrm{e}}{}_{\mathrm{ij}})^2}{h^{\mathrm{e}}{}_{\mathrm{ij}}}$$

mit:

h^{e}_{ij} = erwartete Häufigkeit

h_{ij} = beobachtete Häufigkeit

m = Anzahl der verschiedenen Ausprägungen des Merkmals X

k = Anzahl der verschiedenen Ausprägungen des Merkmals Y

Da Chi-Quadrat größer wird, wenn die Anzahl der Merkmalsträger steigt, wird der Kontingenzkoeffizient berechnet. Dieser kompensiert diese Abhängigkeit. Er berechnet sich folgendermaßen:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

wobei:

n = Anzahl der Merkmalsträger

Für diesen Koeffizienten gilt:

$$C \leq \sqrt{(l-1)/l} \tag{3.2}$$

mit:

l = min(m,k)

Um den Zusammenhang quantitativ beziffern zu können wird schließlich C^* berechnet. Dafür wird folgende Formel verwendet:

$$C^* = \frac{C}{\sqrt{(l-1)/l}}$$

Der Wert C^* gibt an, wie stark der statistische Zusammenhang zwischen zwei Merkmalen ist. Er liegt bei 0, wenn es sich um zusammenhangslose Merkmale handelt und bei 1, wenn die Werte stark korrelieren.
Sollte ein starker statistischer Zusammenhang zwischen den Merkmalen bestehen, wird die Signifikanz des Zusammenhanges überprüft. Dies kann anhand des Chi-Quadrat Tests erfolgen. Anhand des Signifikanzniveaus und der Freiheitsgerade *df* ist es möglich den kritischen Bereich in der Chi-Quadrat-Verteilung abzulesen. Der kritische Wert entspricht dem (1-α)-Quantil der Chi-Quadrat Verteilung mit *df = (m-1)(n-1)*. Ein signifikantes Ergebnis ist dann gegeben, wenn Chi-Quadrat über dem kritischen Bereich liegt.

Für die intervallskalierten Daten und die ordinal skalierten Daten wurde der Rangkorrelationskoeffizient nach Spearman berechnet. Für intervallskalierte Daten wäre es auch möglich den Korrelationskoeffizienten nach Pearson zu berechnen, davon wurde jedoch abgesehen, da er gegenüber dem Korrelationskoeffizienten nach Spearman anfälliger ist gegenüber Extremwerten. [Held (2010)]
Außerdem müssen für einen Signifikanztest nach Pearson die Daten in normalverteilter Form vorliegen. Dies ist bei den meisten erhobenen Daten nicht der Fall.
Der Rangkorrelationskoeffizient berechnet sich folgendermaßen:

$$r_{\mathrm{sp}} = \frac{\sum_{i=1}^{n}(rang(X_{\mathrm{i}}) - \overline{rang(X)})(rang(Y_{\mathrm{i}}) - \overline{rang(Y)})}{\sqrt{\sum_{i=1}^{n}(rang(X_{\mathrm{i}}) - \overline{rang(X)})^2}\sqrt{\sum_{i=1}^{n}(rang(Y_{\mathrm{i}}) - \overline{rang(Y)})^2}}$$

Jedem X-Wert und Y-Wert wird ein Rang zugewiesen, welcher in der Berechnung genutzt wird. Der Rang wird je nach Größe der Ausprägung vergeben. Der Rangkorrelationskoeffizient berechnet die Korrelation zwischen den Rängen von X und Y.
Der Rangkorrelationskoeffizient gibt die Stärke und Richtung des linearen Zusammenhanges an. Nimmt das Ergebnis den Wert -1 an, handelt es sich um einen stark negativen linearen Zusammenhang. Bei einen Rangkorrelationskoeffizienten von 1, um einen starken positiven linearen Zusammenhang und bei einen Ergebnis von 0, um keinen linearen Zusammenhang. [Moser und Schmidt (2011), S.167 ff.]
Sollte der Rangkorrelationskoeffizient zeigen, dass ein starker linearer Zusammenhang

existiert, wird die Signifikanz des Zusammenhangs überprüft. Dafür wird der p-Wert ermittelt. Ein signifikanter Zusammenhang besteht, wenn der p-Wert kleiner als das festgelegte Signifikanzniveau ist. Sobald dies der Fall ist kann die H_0 Hypothese abgelehnt werden.

In den meisten Fällen wird von einem schwachen linearen Zusammenhang zweier Merkmale geredet, wenn der Wert des Korrelationskoeffizienten bei 0,3, bzw. -0,3 liegt. [Held (2010)]
Da der betrachtete Markt allerdings sehr heterogen ist und somit auch die Datenbasis, wird für diese Analyse festgehalten, dass ein Zusammenhang vorhanden ist, wenn der Korrelationskoeffizient, sowie C^* größer als 0,25, beziehungsweise kleiner als -0,25, ist. Sollte der Korrelationskoeffizient sich zwischen -0,25 und 0,25 befinden, wird in diesem Fall von keinem Zusammenhang gesprochen. Der Grund dafür ist, dass Merkmale, die einen geringen Einfluss auf den Preis haben, keinen Mehrwert bieten. Für die anschließende Aufstellung eines Vorhersagemodells würde das bedeuten, dass diese Merkmale nur die Komplexität erhöhen und dies zur Überanpassung (engl. Overfitting) des Modells führen könnte. [Hand u. a. (2001), S.67]

Festlegen des Signifikanzniveau

Um die Signifikanz eines Zusammenhangs zwischen zwei Merkmalen bestimmen zu können, wird ein Signifikanzniveau festgelegt. Das Signifikanzniveau α gibt die maximale Wahrscheinlichkeit, mit der ein Fehler passieren darf, an. Deshalb wird es auch oft als Irrtumswahrscheinlichkeit bezeichnet. Da in den allermeisten Fällen $\alpha = 5\%$ als Signifikanzniveau gewählt wird, wird dies hier auch angewandt. [Hand u. a. (2001), S.112 ff.]

Ergebnis

Überprüfung der Hypothese 1.1.1

Um die Hypothese zu überprüfen werden die Merkmale Länge und monatliche Kosten der Leitung betrachtet. Als Erstes wurde ein Streudiagramm erstellt. In dem Streudiagramm Abbildung 3.3 ist keine ausgeprägte lineare Abhängigkeit zu erkennen. Da es sich bei beiden Merkmalen um intervallskalierte Daten handelt wurde der Rangkorrelationskoeffizient nach Spearman berechnet. Das Ergebnis liegt bei $r_{sp} = 0{,}407$. Es existiert ein mittelmäßiger, positiver linearer Zusammenhang zwischen den beiden Merkmalen. Die

Überprüfung des p-Wertes p = 4,736e^{-140} zeigt, dass ein signifikanter Zusammenhang vorliegt.

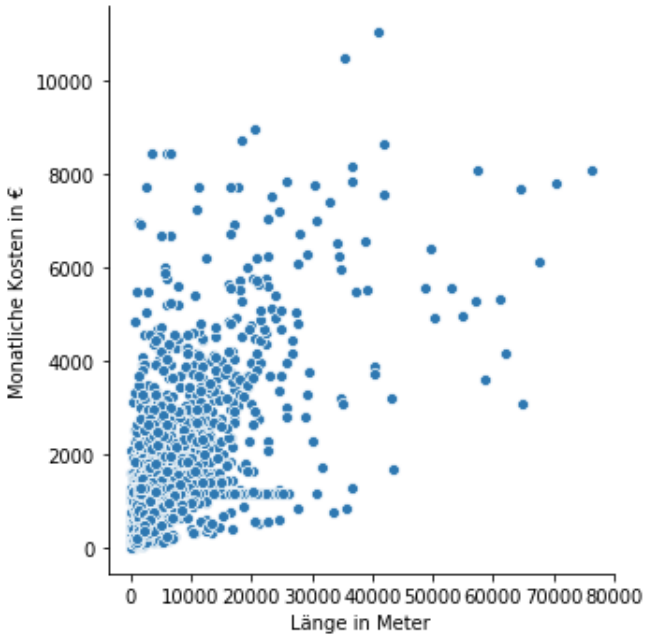


Abbildung 3.3: Monatliche Kosten der Leitung pro Meter.

Dieser lineare Zusammenhang wurde genauer untersucht. Dazu wurden die zehn, in dem Datensatz am häufigst vorkommenden Lieferanten von Mietleitungen, herausgefiltert. Anschließend wurde für jeden dieser Lieferanten ein Streudiagramm hinsichtlich der Länge und der monatlichen Kosten erstellt, sowie eine Korrelationsanalyse durchgeführt. Die Analyse zeigt zwei differenzierte Ergebnisse. Die zwei repräsentativsten Fälle werden im Folgenden dargestellt.
In dem in Abbildung 3.4 dargestellten Streudiagramm, welches die Trassen eines Lieferanten A darstellt, ist eine starke lineare Abhängigkeit erkennbar. Die Rangkorrelationsanalyse ergab einen Wert von r_{sp} = 0,94. Der p-Wert bestätigt, dass ein signifikanter Zusammenhang vorliegt p = 1,245e^{-41}.
Dagegen zeigt das Streudiagramm Abbildung 3.5 eines anderen Lieferanten B, dass keine lineare Abhängigkeit vorhanden ist. Die Rangkorrelationsanalyse ergibt r_{sp} = -0,097.

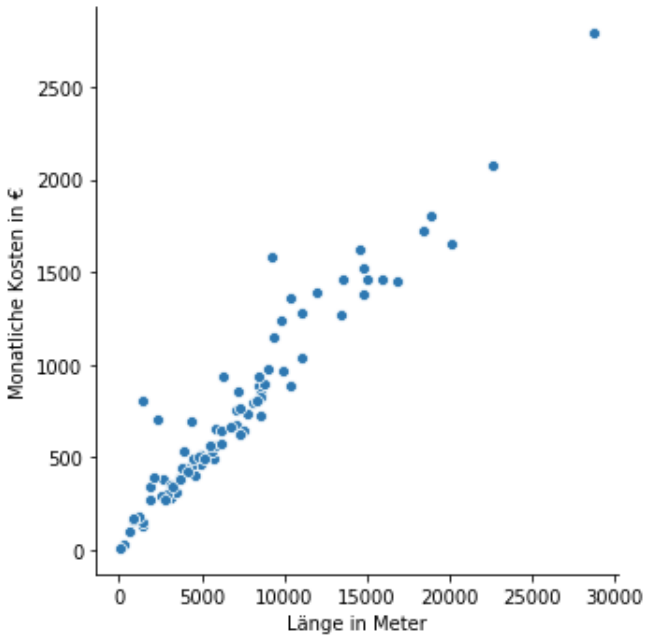


Abbildung 3.4: Monatliche Kosten der Leitung pro Meter des Lieferanten A.

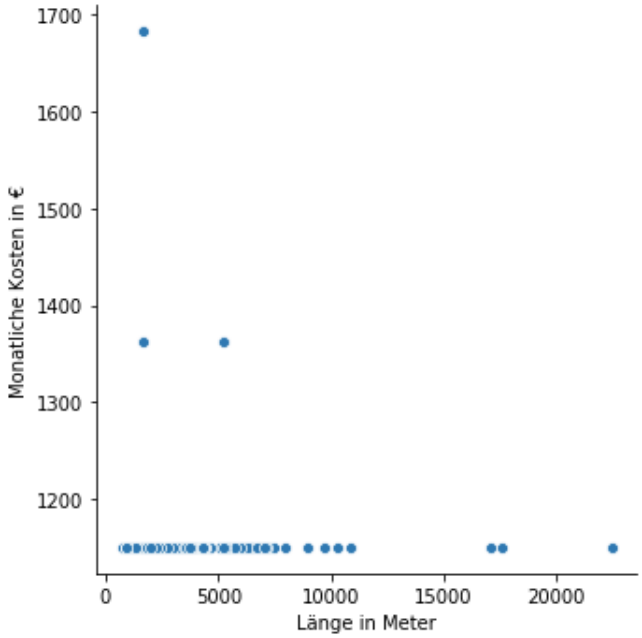


Abbildung 3.5: Monatliche Kosten der Leitung pro Meter des Lieferanten B.

Auswertung der Hypothese 1.1.1

In diesem Fall kann die H_0 Hypothese weder bestätigt noch angenommen werden, da das Ergebnis zeigt, dass beide Fälle eintreten können. Die Lieferanten verfolgen unterschiedliche Preismodelle. Bei einigen Lieferanten hat die Länge Einfluss auf die monatlichen Preise und das betrachtete Telekommunikationsunternehmen zahlt einen Preis pro Meter (erster Fall). Bei anderen Lieferanten ist es so, dass ein fester monatlicher Preis pro Leitung bezahlt wird (zweiter Fall). Der Lieferant verlangt einen Preis pro Stück. Die Länge der Leitung hat in diesem Fall keinen Einfluss auf die monatlichen Preise.

Überprüfung der Hypothese 1.1.2

Zur Überprüfung dieser Hypothese werden die Merkmale einmalige Kosten und die Länge betrachtet. Es wurde ein Streudiagramm hinsichtlich der einmaligen Kosten und der Länge erstellt. Das Streudiagramm Abbildung 3.6 lässt keine lineare Abhängigkeit der beiden Merkmale vermuten.

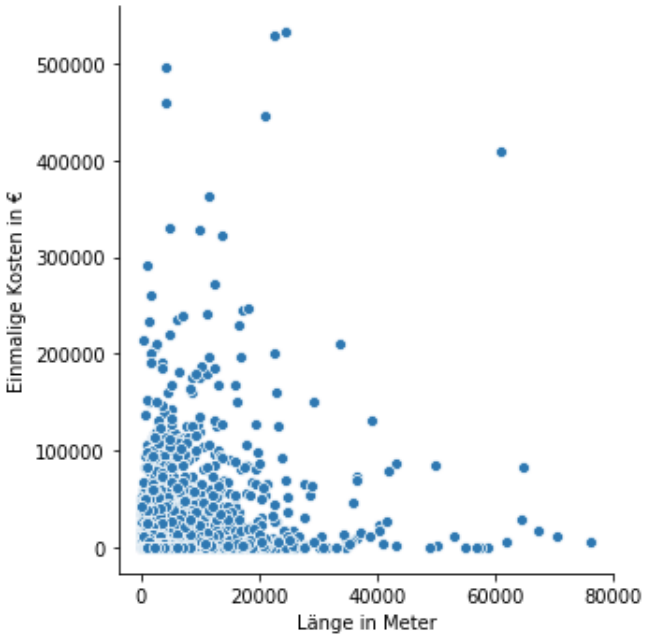


Abbildung 3.6: Einmalige Kosten der Leitung pro Meter.

Da es sich bei den Merkmalen um intervallskalierte Merkmale handelt, wurde der Rangkorrelationskoeffizient nach Spearman berechnet. Der Rangkorrelationskoeffizient nach Spearman beträgt $r_{sp} = 0{,}039$.

Auswertung der Hypothese 1.1.2

Da die Rangkorrelationsanalyse einen sehr geringen Wert liefert und zeigt, dass keine starke, positive Abhängigkeit zu erkennen ist, kann die H_0 Hypothese nicht widerlegt werden und wird angenommen. Die Länge hat keinen Einfluss auf den einmaligen Preis einer Leitung. Dieses Ergebnis lässt sich folgendermaßen erklären:
In den einmaligen Kosten spiegeln sich oftmals die Kosten des Tiefbaus wider. Da die meisten Trassenführungen jedoch nicht vollständig neu gebaut werden, sondern bis zu dem nächstgelegenden Anschlusspunkt, entspricht die erhobene Länge nicht der Länge des Tiefbaus. Es besteht dadurch keine Abhängigkeit zwischen den beiden Merkmalen.

Überprüfung der Hypothese 1.2.1

Um die Hypothese zu überprüfen werden Merkmale der Ausprägung der geographischen Lage einer Trasse benötigt. Zwei Merkmale des Datensatzes beziehen sich auf die geographische Lage einer Leitung, die Area Class und die Besiedlungsklasse. Es wird jeweils der statistische Zusammenhang zwischen den Merkmalen und den monatlichen Kosten untersucht.
Als Erstes wird der Zusammenhang zwischen der Area Class und den monatlichen Kosten überprüft. Da es sich bei der Area Class um ein nominalskaliertes Merkmal handelt, wird im ersten Schritt ein kategorisches Streudiagramm erstellt. Das Streudiagramm Abbildung 3.7 lässt vermuten, dass der monatliche Preis je nach Area Class variiert. Um die Stärke des statistischen Zusammenhangs festzustellen, wurde der Preis kategorisiert und eine Kontingenztafel aufgestellt. Anschließend wurde der C^* Koeffizient berechnet, $C^* = 0{,}448$. Es existiert ein statistischer Zusammenhang zwischen den beiden Merkmalen.
Um den Zusammenhang auf Signifikanz zu überprüfen, wurde der kritische Bereich in der Chi-Quadrat-Verteilung abgelesen,. Der kritische Bereich liegt bei 41,3371. Der ermittelte Chi-Quadrat Wert $X^2 = 540{,}161$ liegt deutlich über dem kritischen Bereich. Es existiert somit ein signifikanter Zusammenhang.

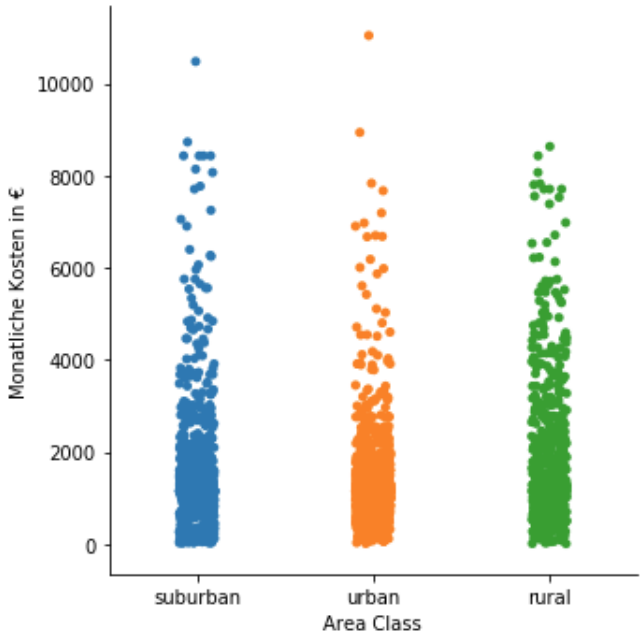


Abbildung 3.7: Darstellung der Trassen nach der Area Class und den monatlichen Kosten.

Anschließend wurde die Analyse für die Besiedlungsklassen durchgeführt. Da es sich bei der Besiedlungsklasse auch um nominalskalierte Daten handelt, wurde die gleiche Vorgehensweise gewählt. Das Streudiagramm Abbildung 3.8 lässt vermuten, dass der Preis bei unterschiedlichen Besiedlungsklassen variiert. Außerdem wurde im Zuge der Korrelationsanalyse der C^* Koeffizient berechnet, $C^* = 0{,}333$. Auch in diesem Fall liegt ein statistischer Zusammenhang vor.
Eine Überprüfung der Signifikanz anhand des Chi-Quadrat Testes ergab, dass der Zusammenhang signifikant ist, $279,57 \geq 41{,}3371$.

Auswertung der Hypothese 1.2.1

Beide berechneten Koeffizienten liegen über 0,25 und beide Merkmale stehen in einem signifikanten statistischen Zusammenhang mit dem monatlichen Preis. Somit kann die H_0 Hypothese abgelehnt werden und die H_1 Hypothese wird angenommen. Das bedeutet, dass die geographische Lage einen Einfluss auf den monatlichen Preis hat und die Vermutung, dass die Oberflächenarbeiten sich in dem Mietpreis widerspiegeln, zutrifft. Die geographische Lage wird in diesem Fall widergespiegelt durch die Merkmale Area Class und Besiedlungsklasse.

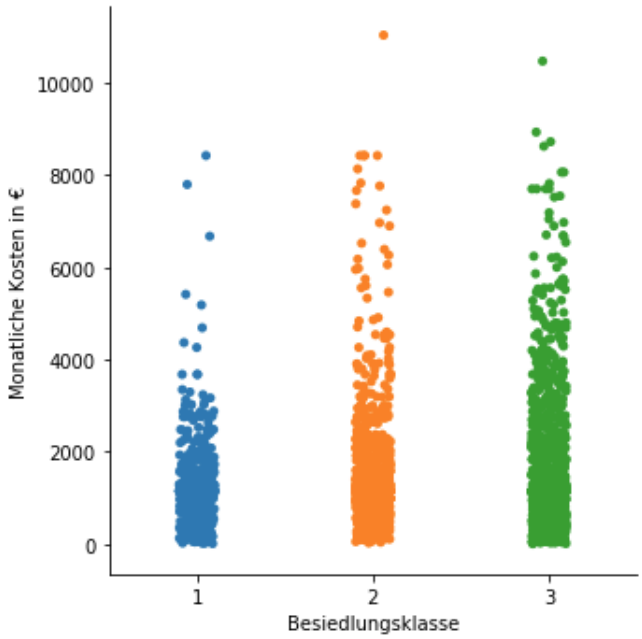


Abbildung 3.8: Darstellung der Trassen nach der Besiedlungsklasse und den monatlichen Kosten.

Überprüfung der Hypothese 1.2.2

Um zu überprüfen, ob eine Korrelation zwischen dem einmaligen Preis und der geographischen Lage einer Trasse besteht, wird der Zusammenhang zwischen den Merkmalen einmalige Kosten und der Besiedlungklasse, sowie zwischen den Merkmalen einmalige Kosten und der Area Class untersucht.
Als Erstes wird die Korrelation zwischen den einmaligen Kosten und der Area Class berechnet. Da es sich bei der Area Class um nominalskalierte Daten handelt, wurde im ersten Schritt ein kategorisches Streudiagramm erstellt. Dieses ist in Abbildung 3.9 abgebildet. Um die Stärke des statistischen Zusammenhangs zu quantifizieren, wurde der einmalige Preis kategorisiert und eine Kontingenztafel aufgestellt. Anschließend wurde der C^* Koeffizient berechnet, $C^* = 0{,}432$. Dieser Wert zeigt, dass ein relativ starker statistischer Zusammenhang vorliegt. Der Chi-Quadrat Test bestätigt, dass es sich um einen signifikanten Zusammenhang handelt. Chi-Quadrat liegt mit $X_2 = 496{,}973$ über dem kritischen Bereich von 46,1943.

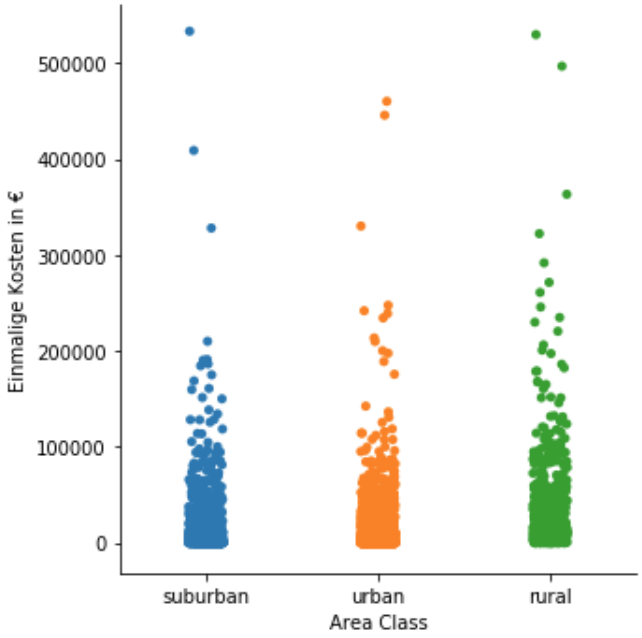


Abbildung 3.9: Darstellung der Trassen nach der Area Class und den einmaligen Kosten.

Ferner wurde überprüft, ob ein Zusammenhang zwischen der Besiedlungsklasse und den einmaligen Kosten existiert. Wie bei den vorherigen Analysen wurde auch in diesem Fall ein kategorisches Streudiagramm erstellt. Auch hier lässt sich anhand des Streudiagramms in Abbildung 3.10 vermuten, dass der Preis je nach Besiedlungsklasse variiert. Mehr Aufschluss über den statistischen Zusammenhang liefert der berechnete C^* Koeffizient, welcher $C^* = 0{,}264$ beträgt. Er zeigt, dass ein statistischer Zusammenhang existiert. Auch dieser ist Signifikant, denn der berechnete Chi-Quadrat Wert liegt mit $X_2 = 171{,}336$ über dem ermittelten kritischen Bereich.

Auswertung der Hypothese 1.2.2

Auch in diesem Fall liegen beide C^* Koeffizienten über der festgelegten Koeffizientengröße von 0,25 und beide Analysen zeigen einen signifikanten Zusammenhang auf. Dies bedeutet: Die H_0 Hypothese ist widerlegt. Die Hypothese H_1, welche besagt, dass die geographische Lage einer Leitung Einfluss auf den einmaligen Preis hat, wird angenommen.

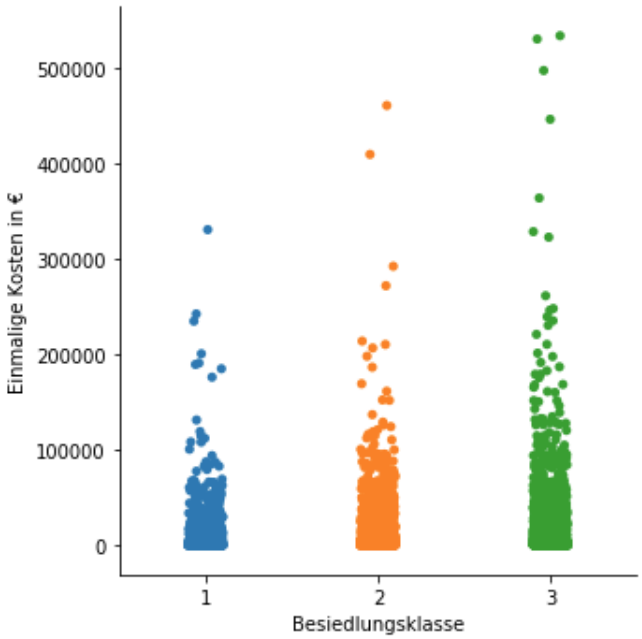


Abbildung 3.10: Darstellung der Trassen nach der Besiedlungsklasse und den einmaligen Kosten.

Überprüfung der Hypothese 1.3.1

Um die Gültigkeit dieser Hypothese zu testen wurden zum einen die monatlichen Kosten betrachtet, sowie die Erschwernisfaktoren der Bodenklasse. Da es sich bei den Erschwernisfaktoren der Bodenklasse um intervallskalierte Merkmale handelt, wurde die Rangkorrelationsanalyse nach Spearman durchgeführt. Die Erstellung des Streudiagramms, welches in Abbildung 3.11 dargestellt ist, und die Berechnung des Rangkorrelationskoeffizienten zeigen keine starke Korrelation der beiden Merkmale. Das Ergebnis des Koeffizienten r_{sp} liegt bei r_{sp} = -0,151.

Auswertung der Hypothese 1.3.1

Die Analyse zeigte, dass nur ein sehr geringer linearer Zusammenhang zwischen den beiden Merkmalen besteht. Die H_0 Hypothese kann nicht widerlegt werden.
Da bewiesen wurde, dass die Tiefbaukosten den stärksten Kostentreiber bei dem Bau der Anschlussnetzinfrastruktur ausmachen und diese durch die Bodenklassen beeinflusst werden ist dieses Ergebnis nicht erklärbar.

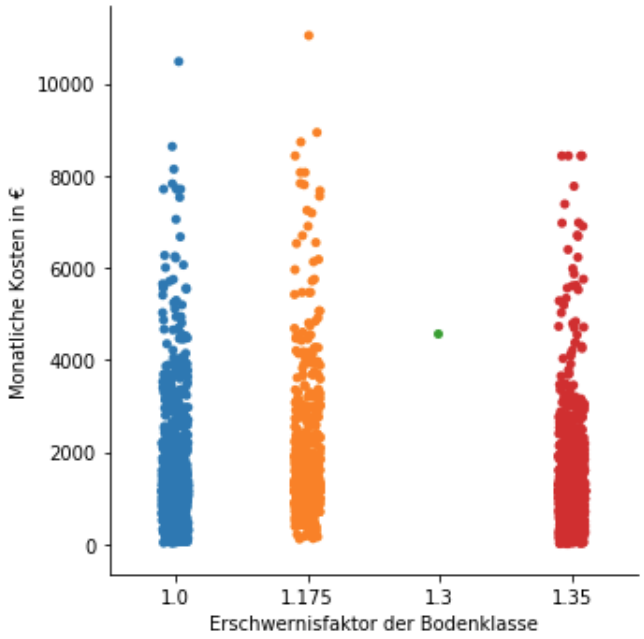


Abbildung 3.11: Monatliche Kosten der Leitung je nach Höhe des Erschwernisfaktors der Bodenklasse.

Es ist allerdings zu berücksichtigen, dass nur die Bodenklassen an den Standorten A und B erhoben wurde und die Möglichkeit besteht, dass die Tiefbauarbeiten eine längere Strecke ausmachen und dabei weitere Bodenklassen durchqueren. Es besteht zur Zeit keine Möglichkeit diese Informationen herauszufinden, da der genaue Verlauf der Strecke der Tiefbauarbeiten nicht bekannt ist.

Überprüfung der Hypothese 1.3.2

Wie bei der Überprüfung der Hypothese 1.3.1 wird hier das Merkmal der Erschwernisfaktoren der Bodenklassen herangezogen. Allerdings findet eine Analyse in Hinblick auf die einmaligen Kosten statt. Ein Streudiagramm und der Rangkorrelationskoeffizient nach Spearman finden hier Anwendung. Das kategorische Streudiagramm Abbildung 3.12 und der berechnete Rangkorrelationskoeffizient, welcher r_{sp} = -0,328 beträgt, zeigen, dass im Gegensatz zur Vermutung ein mittelmäßiger, negativer linearer Zusammenhang besteht.

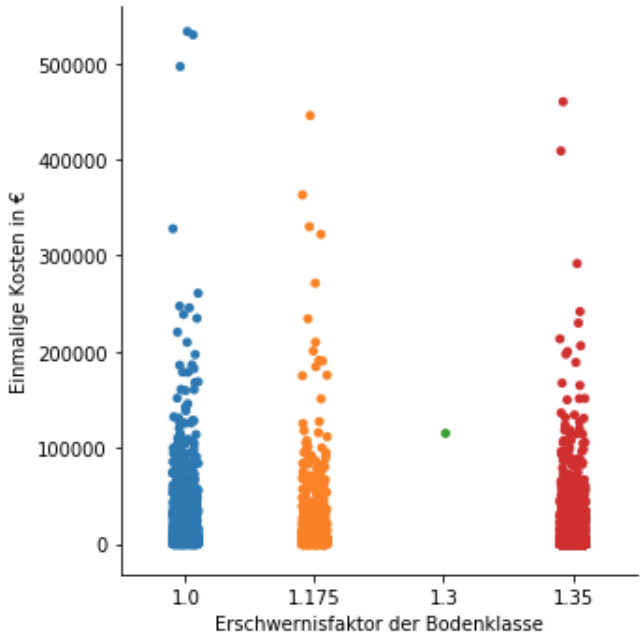


Abbildung 3.12: Einmalige Kosten der Leitung je nach Höhe des Erschwernisfaktors der Bodenklasse.

Auswertung der Hypothese 1.3.2

Da es sich um einen negativen linearen Zusammenhang handelt und der Korrelationkoeffizient geringer als 0,25 ist, wird die H_0 Hypothese beibehalten.
Für das Ergebnis existiert keine Erklärung, weshalb es keine weitere Beachtung findet. Es ist nicht plausibel, dass ein größerer Erschwernisfaktor einen geringeren Preis mit sich bringt. Hier tritt der Fall ein, dass eine Korrelation nicht immer Kausalität voraussetzt.

Überprüfung der Hypothese 1.4.1

Um diese Hypothese zu überprüfen wird der Zusammenhang der Merkmale Wettbewerb und monatliche Kosten analysiert. Es handelt sich in diesem Fall um intervallskalierte Daten. Das Streudiagramm Abbildung 3.13 lässt keinen linearen Zusammenhang der Merkmale vermuten. Die Rangkorrelationsanalyse nach Spearman bestätigt dieses Ergebnis. Der Rangkorrelationskoeffizient beträgt r_{sp} = -0,083. Es existiert kein linearer Zusammenhang zwischen den beiden Merkmalen.

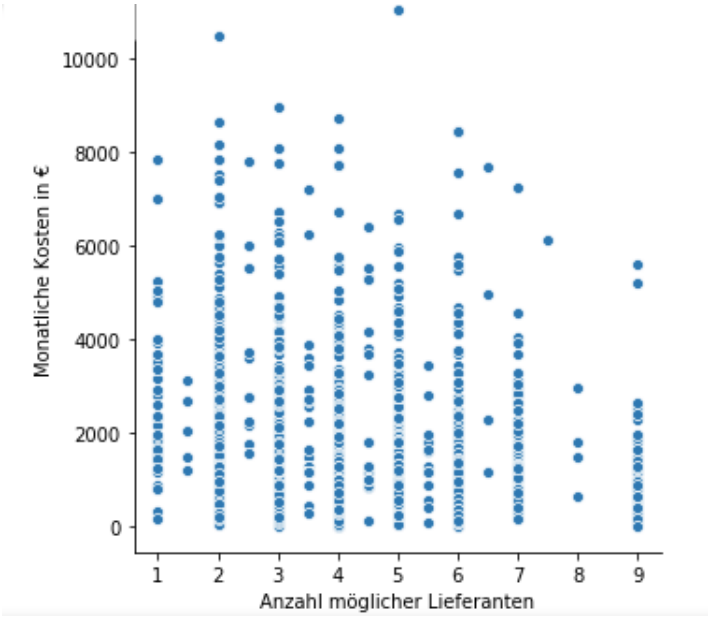


Abbildung 3.13: Monatliche Kosten der Trasse nach Anzahl der möglichen Lieferanten.

Auswertung der Hypothese 1.4.1

Es besteht keine Korrelation zwischen den beiden Merkmalen. Dies beweisen das Streudiagramm sowie der Rangkorrelationskoeffizient. Die H_0 Hypothese kann nicht widerlegt werden und wird angenommen. Der Grund für dieses Ergebnis kann darin liegen, dass das Merkmal nur die Anzahl der Lieferanten pro Postleitzahlengebiet angibt, bei denen das betrachtete Telekommunikationsunternehmen Leitungen gemietet hat. In den Postleitzahlengebieten könnte aber ein viel höherer Wettbewerb vorherrschen als angenommen, da eventuell noch weitere Lieferanten in dem Gebiet agieren. Diese Lieferanten wurden nicht berücksichtigt, da das Unternehmen bei denen keine Leitungen gemietet hat, weil diese zu teuer sind oder weil bewusst keine Geschäftsbeziehung besteht.

Überprüfung der Hypothese 1.4.2

Auch hier wurde das Merkmal Wettbewerb für die Analyse genutzt. Außerdem werden die einmaligen Kosten betrachtet. Es handelt sich um intervallskalierte Daten. Ein Streudiagramm hinsichtlich der beiden Merkmale wurde erstellt. In dem Streudiagramm Abbildung 3.14 ist kein linearer Zusammenhang erkennbar. Die Rangkorrelationsanalyse

nach Spearman zeigt, dass kein linearer Zusammenhäng zwischen den beiden Merkmalen vorhanden ist, $r_{sp} = 0{,}035$.

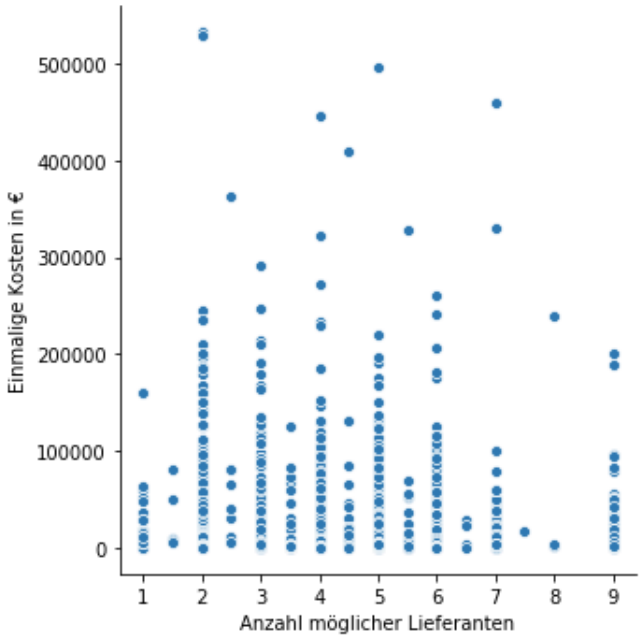


Abbildung 3.14: Einmalige Kosten der Trasse nach Anzahl der möglichen Lieferanten.

Auswertung der Hypothese 1.4.2

Die Analyse zeigt, dass der Wettbewerb keinen Einfluss auf die einmaligen Kosten hat. Die Hypothese H_0 wird angenommen.
Wie eben erwähnt könnte der Grund für dieses Ergebnisses in der Erfassung der Daten zu der Höhe des Wettbewerbes liegen.

Überprüfung der Hypothese 1.5.1

Die Hypothese H_0 besagt, dass kein negativer Zusammenhang zwischen der Vertragsdauer und den monatlichen Kosten existiert. Um dies zu widerlegen werden die Merkmale monatliche Kosten und die Vertragsdauer in Monaten in Hinblick auf einen linearen Zusammenhang analysiert. Bei den beiden Merkmalen handelt es sich um intervallskalierte Merkmale. Eine Rangkorrelationsanalyse nach Spearman wurde durchgeführt und ein

Streudiagramm wurde erstellt. Das Streudiagramm Abbildung 3.15, sowie die Korrelationsanalyse zeigen, dass kein linearer Zusammenhang zwischen den beiden Merkmalen besteht. Der Korrelationskoeffizient beträgt r_{sp} = -0,087.

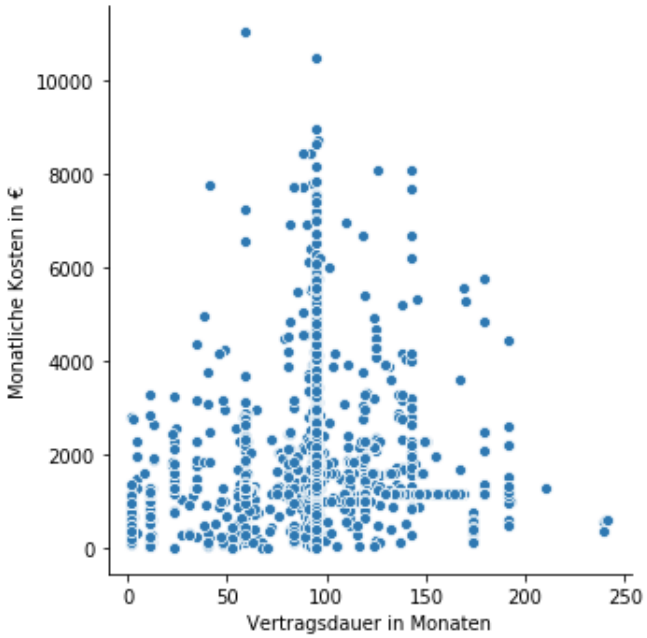


Abbildung 3.15: Monatliche Kosten der Leitung nach Länge der Vertragsdauer.

Auswertung der Hypothese 1.5.1

Der Rangkorrelationskoeffizient beträgt fast null, somit ist kein linearer Zusammenhang zwischen den Merkmalen vorhanden. Die H_0 Hypothese wird angenommen.
Da bei der Datenerhebung des Merkmals Vertragsdauer keine Ungenauigkeiten entstanden sein können, besteht kein Zweifel in der Gültigkeit der H_0 Hypothese, auch wenn eine andere Vermutung angenommen wurde.

Überprüfung der Hypothese 1.5.2

Die Überprüfung der Hypothese wurde gleichermaßen durchgeführt, wie die Überprüfung der Hypothese 1.5.1. Der einzige Unterschied bestand darin, dass statt der monatlichen Kosten die einmaligen Kosten verwendet wurden. Es handelt sich um intervallskalierte

Merkmale aus diesem Grund wurde ein Streudiagramm erstellt und der Rangkorrelationskoeffizient nach Spearman berechnet. Das Streudiagramm Abbildung 3.16 lässt keinen linearen Zusammenhang erkennen. Der Rangkorrelationskoeffizient nach Spearman weist folgendes Ergebnis auf: r_{sp} = -0,524.

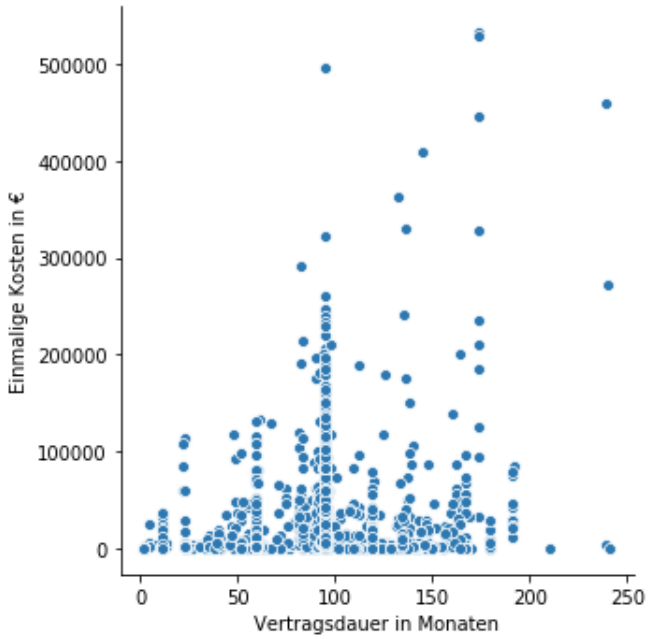


Abbildung 3.16: Einmalige Kosten der Leitung nach Länge der Vertragsdauer.

Auswertung der Hypothese 1.5.2

Das Ergebnis ist in sich widersprechend. Wo im Streudiagramm keine Abhängigkeit zu erkennen ist, zeigt der Rangkorrelationskoeffizient einen starken negativen, linearen Zusammenhang. Dies lässt sich folgendermaßen begründen:
Der Rangkorrelationskoeffizient reagiert nicht stark auf Extremwerte. In dem dargestellten Zusammenhang sind jedoch viele Extremwerte ersichtlich, weshalb augenscheinlich kein linearer Zusammenhang existiert. Diese Extremwerte treten aber häufiger auf und sind kein Datenerhebungsfehler. Deshalb sollten sie bei der Analyse mehr Beachtung finden. Aus diesem Grund wurde der Korrelationskoeffizient nach Pearson herangezogen, welcher sensibler gegenüber Extremwerten ist. Die Berechnung ergab r_{XY} = -0,133. Dies lässt darauf schließen, dass kein linearer Zusammenhang vorhanden ist. Die H_0 Hypothese wird nicht widerlegt, sondern angenommen.

Gesamtauswertung

Die Fragestellung "*Welche Faktoren beeinflussen den Preis einer Leitung?*" kann nun beantwortet werden. Und zwar beeinflusst die geographische Lage einer Trasse den Preis. Der einmalige Preis, sowie der monatliche Preis werden durch die geographische Lage der Leitung beeinflusst. Die geographische Lage einer Leitung wird durch die Merkmale Area Class und Besiedlungklasse dargestellt.
Außerdem besitzt die Länge einer Leitung einen Einfluss auf den monatlichen Preis. Jedoch nicht in jeden Fall, es hängt von den Preismodellen der Lieferanten ab. Werden die Trassen, wie bei dem Lieferanten im ersten Fall (Abbildung 3.4), nach Metern bezahlt, existiert eine starke Korrelation der Merkmale.

3.4.2 Analyse der Preisentwicklung

Erläuterung der Durchführung

Um festzustellen, ob eine zeitliche Entwicklung der Preise erfolgt ist, wurde zuerst das Datum, seit dem das betrachtete Telekommunikationsunternehmen die Leitungen gemietet hat (Merkmal: kostenpflichtig ab), auf das Anmietungsjahr begrenzt. Anschließend wurde für jedes Anmietungsjahr ein Durchschnittspreis, der in dem Jahr angemieteten Leitungen, berechnet. Diese Werte wurden in einem Liniendiagramm abgebildet und hinsichtlich eines Trends analysiert.

Ergebnis

Überprüfung der Hypothese 2.1.1

Um diese Hypothese zu überprüfen werden die monatlichen Kosten herangezogen. Da eben bewiesen wurde, dass bei einigen Lieferanten eine starke lineare Abhängigkeit zwischen dem monatlichen Preis und der Länge einer Leitung besteht, wird der monatliche Preis pro Meter herangezogen. Es wird davon ausgegangen, dass die Leitungen, die einen einheitlichen Stückpreis besitzen, gleichmäßig in Hinblick auf die Länge verteilt sind. Dadurch entsteht kein negativer Einfluss auf den berechneten Durchschnittswert pro Meter. Es ist nicht möglich den Einfluss der geographischen Lage aus den Kosten herauszurechnen, da die Merkmale der geographischen Lage nicht quantifizierbar sind.

Es wird schließlich der monatliche Preis pro Meter im zeitlichen Verlauf betrachtet. Das Liniendiagramm ist in der Abbildung 3.17 dargestellt. Es handelt sich um einen Durchschnittspreis pro Meter in den unterschiedlichen Jahren. Der schraffierte Bereich kennzeichnet die Jahre, in denen relativ wenig Leitungen angemietet wurden.

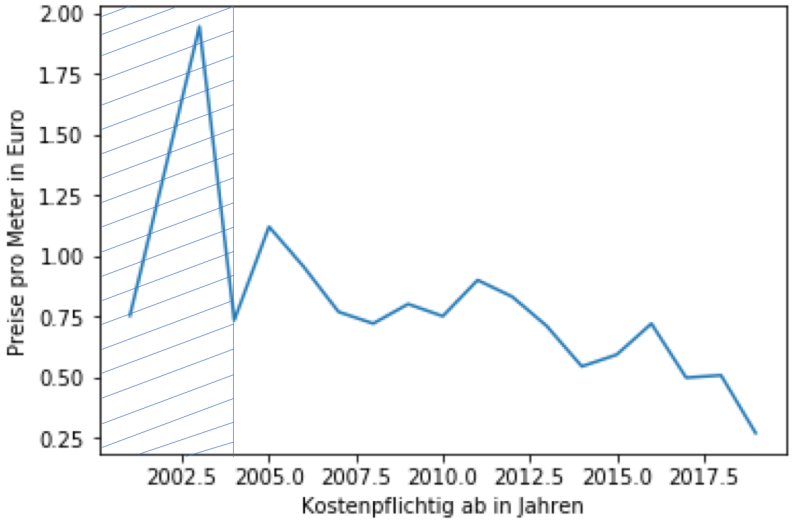


Abbildung 3.17: Zeitliche Entwicklung der monatlichen Preise pro Meter.

Auswertung der Hypothese 2.1.1

Es ist deutlich zu erkennen, dass ein negativer Trend existiert. Die monatlichen Preise pro Meter werden immer geringer. Die H_0 Hypothese wird abgelehnt und die H_1 Hypothese wird angenommen. Der monatliche Preis pro Meter folgt einem Trend.
Allerdings folgt im Gegensatz zu der aufgestellten Vermutung der Preis einem negativen Trend. Grund dafür könnte sein, dass in den ersten Jahren das initiale Netz erst aufgebaut werden musste und dies höhere Kosten verursacht hat. Es existierte noch keine große Infrastruktur an die angeknüpft werden konnte.

Überprüfung der Hypothese 2.1.2

Um zu untersuchen, ob diese Hypothese gilt werden die einmaligen Kosten analysiert. Da eben bewiesen wurde, dass nur die geographische Lage Einfluss auf die einmaligen Kosten hat und die Ausprägung dieser Merkmale nicht quantifizierbar ist, somit nicht verrechenbar mit den Kosten ist, werden die einmaligen Kosten nicht verrechnet. Es wurde für jedes Jahr ein durchschnittlicher einmaliger Preis gebildet. Dieser wurde in einem Liniendiagramm veranschaulicht, welches dargestellt ist in der Abbildung 3.18. Der schraffierte Bereich kennzeichnet die Jahre, in denen relativ wenig Leitungen angemietet wurden.

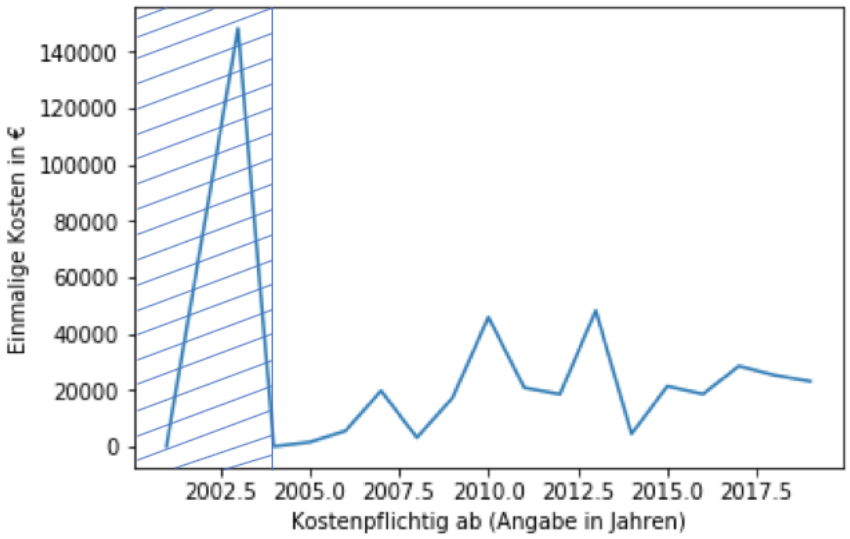


Abbildung 3.18: Zeitliche Entwicklung der einmaligen Preise.

Auswertung der Hypothese 2.1.2

Es ist keine eindeutige Tendenz des einmaligen Preises im zeitlichen Verlauf zu erkennen. Die Kurve schwankt. Weder ein positiver noch ein negativer Trend ist ersichtlich. Die H_0 Hypothese kann nicht widerlegt werden.

Gesamtauswertung

Die Fragestellung "*Existiert ein zeitlicher Trend der Preise?*" kann nun beantwortet werden. Es wird festgehalten, dass der monatliche Preis pro Meter einem negativen Trend folgt. Der einmalige Preis scheint keinen zeitlichen Trend zu besitzen. Dies ist damit begründet, dass der einmalige Preis meistens die Kosten widerspiegelt, die beim Bau der Trassenführung entstanden sind. Da diese je nach Lage und Länge des Tiefbaus variieren ist keine zeitliche Abhängigkeit ersichtlich.

4 Entwicklung der Modelle

Dieses Kapitel beschreibt die Entwicklung eines Modells zur Bestimmung eines marktkonformen Mietpreises für eine Glasfaserleitung. Das Modell soll als Grundlage für eine Applikation dienen. Es wird ein Modell entworfen, welches einen marktkonformen monatlichen Preis vorhersagt, sowie eins, welches einen marktkonformen einmaligen Preis bestimmt. Die Modelle berücksichtigen die Einflussfaktoren der Preise, sowie die zeitliche Entwicklung.

Um ein Vorhersagemodell zu erstellen, welches anhand der Prädiktorvariablen die Antwortvariable, in diesem Fall den Preis, prognostiziert werden Methoden des maschinellen Lernens angewandt. Zur Anwendung von Methoden des maschinellen Lernens ist eine Vorverarbeitung der Daten notwendig. Dieser Vorgang wird in Abschnitt 4.1 erläutert.
Das Vorhersagemodell, welches einen monatlichen Preise für ein Angebot vorhersagt, anhand der Ausprägungen der Einflussfaktoren wird in Abschnitt 4.2.1 vorgestellt. Das Modell, welches zur Vorhersage eines einmaligen Preise, anhand der Einflussfaktoren dient, ist Gegenstand von Abschnitt 4.2.2.
Zudem wird ein Trendmodell der Preise pro Meter entwickelt. Dieses wird in das Vorhersagemodell für die monatlichen Preise integriert. Der Entwurf des Modells, sowie die Verknüpfung der beiden Modelle ist Gegenstand von Abschnitt 4.3.
Anschließend wird in Abschnitt 4.4 die Darstellung der inflationären Entwicklung der Preise in den Modellen erläutert. Eine Bewertung der Modelle ist Gegenstand von Abschnitt 4.5.

4.1 Vorverarbeitung der Daten

Als Erstes wurde der ursprüngliche Datensatz in zwei neue Datensätze unterteilt.
Ein Datensatz der folgende Merkmale enthält: monatliche Kosten, Länge einer Leitung, Area Class und die Besiedlungsklasse. Dieser Datensatz dient als Grundlage für die Erstellung eines Vorhersagemodells für den monatlichen Preis.
Der zweite Datensatz enthält das Merkmal einmaligen Kosten, sowie die Merkmale, die Einfluss auf die einmaligen Preise haben. Dazu zählen die Merkmale Area Class und Besiedlungsklasse. Dieser Datensatz wird als Grundlage für die Erstellung eines Vorhersagemodells für den einmaligen Preis verwendet.

Bei beiden erstellten Datensätzen handelt es sich um gekennzeichnete Datensätze. Die Datensätze beinhalten den erwünschten Ausgabewert, nämlich die Preise der Mietleitung. Diese Variablen werden als Antwortvariable betitelt. Die übrigen Merkmale in den jeweiligen Datensätzen fungieren als Prädiktorvariablen.
Dadurch, dass es sich bei den Datensätzen um gekennzeichnete Datensätze handelt, finden zur Erstellung der Vorhersagemodelle Methoden des überwachten Lernens Anwendung.

Im ersten Schritt wurden die beiden Datensätze in zwei Teilmengen unterteilt, in einen Trainingsdatensatz und einen Testdatensatz. Der Trainingsdatensatz dient zum trainieren des Modells. Der Testdatensatz wird zum Testen des Modells genutzt. Die Teilung der Datensätze wurde im Verhältnis 70:30 vorgenommen, 30% Testdaten und 70% Trainingsdaten. [Vorgehensweise: Raschka und Mirjalili (2018)]

Anschließend wurden für die Analysen die numerischen Prädiktorvariablen umskaliert, sodass der Mittelwert der Merkmale bei 0 liegt und die Merkmale eine einheitliche Varianz aufweisen. Dies war notwendig, da sonst die unterschiedlichen Merkmale nicht miteinander vergleichbar sind. Ferner setzen einige Klassifikations- und Regressionsverfahren eine Umskalierung der Merkmale voraus. Für ein Merkmal mit der Ausprägung x wurde der neue Wert z wie folgt berechnet:

$$z = (x - u)/s \tag{4.1}$$

Dabei handelt es sich bei u um den Mittelwert und bei s um die Standardabweichung. Für die Berechnung des Mittelwerts und der Standardabweichung wurden die Merkmale aus dem Trainingsdatensatz genutzt. Die unterschiedlichen Merkmale wurden unabhän-

gig voneinander betrachtet. Eine Umskalierung der numerischen Merkmale erfolgt für den Trainingsdatensatz und den Testdatensatz. Beim Umskalieren der numerischen Merkmale aus dem Testdatensatz wurden die, für das jeweilige Merkmal aus dem Trainingsdatensatz, berechneten Werte genutzt.

Um die Informationen der kategorialen Merkmale für Verfahren des überwachten Lernens nutzen zu können, wurden die kategorialen Merkmale kodiert. Dafür wurde die "One-Hot"-Kodierung verwendet. Die "One-Hot"-Kodierung transformiert ein kategoriales Merkmal, das n unterschiedliche Werte annehmen kann, in n neue Merkmale. Das Merkmal, welches der vorherigen Ausprägung entsprach, bekommt den Wert 1 zugewiesen, während die weiteren neuen Merkmale die Ausprägung 0 erhalten.
Angewandt wurde die "One-Hot"-Kodierung unter anderem für das Merkmal Area Class. Das Merkmal Area Class kann drei unterschiedliche Werte annehmen. Dazu zählen die Werte urban, rural und suburban. Die "One-Hot"- Kodierung transformiert das Merkmal in drei neue Merkmale. In diesem Fall sind die neuen Merkmale urban, suburban und rural. Dieser Vorgang ist am Beispiel der Area Class in der Abbildung 4.1 dargestellt. Eine Instanz die vorher die Ausprägung urban aufwies erhält die Ausprägung 1 für das neue Merkmal urban. Die Merkmale suburban und rural erhalten die Ausprägung 0. [Pedregosa u. a. (2011), S.2825 ff.] [Grus (2016), S.129 ff.]

Abbildung 4.1: Vorgang der One-Hot Kodierung am Beispiel des Merkmals Area Class.

4.2 Entwurf der Vorhersagemodelle

In diesem Kapitel werden die Vorhersagemodelle vorgestellt. Es wurde ein Modell für den monatlichen Preis erstellt, welches einen marktkonformen monatlichen Preis vorhersagt anhand der Ausprägungen der Einflussfaktoren. Zu dem wurde ein Vorhersagemodell für

den einmaligen Preis entworfen, diese bestimmt anhand der Ausprägungen der Einflussfaktoren einen Preis.

4.2.1 Vorhersagemodell für den monatlichen Preis

Als Grundlage für die Erstellung des Modells dient der vorverarbeitete Datensatz. Es werden nur die Daten von Carriern betrachtet, deren monatlicher Preis kein Festpreis ist[1]. Der Datensatz besteht somit aus den Prädiktorvariablen Länge, rural, suburan, urban und den drei Besiedlungsklassen. Die monatlichen Kosten fungieren als Antwortvariable. Ein Ausschnitt der Daten ist in Abbildung 4.2 dargestellt.

rural	suburban	urban	Besiedlungsklasse 1	Besiedlungsklasse 2	Besiedlungsklasse 3	Länge	Monatliche Kosten
1.0	0.0	0.0	0.0	0.0	1.0	-0.210717	580.00
0.0	0.0	1.0	0.0	1.0	0.0	-0.615058	580.00
0.0	0.0	1.0	1.0	0.0	0.0	-0.526128	354.50
0.0	0.0	1.0	1.0	0.0	0.0	-0.638575	354.50
1.0	0.0	0.0	0.0	0.0	1.0	-0.458322	1250.00
1.0	0.0	0.0	0.0	0.0	1.0	-0.464984	1250.00
0.0	0.0	1.0	1.0	0.0	0.0	-0.758778	1750.00
0.0	1.0	0.0	1.0	0.0	0.0	-0.727343	561.71

Prädiktorvariable | Antwortvariable

Abbildung 4.2: Ausschnitt des Trainingsdatensatzes für die Aufstellung eines Vorhersagemodells für den monatlichen Preis.

Es wird ein Vorhersagemodell entworfen, welches anhand der Prädiktorvariablen die Antwortvariable prognostiziert.
Da es sich bei dem Datensatz um einen gekennzeichneten Datensatz handelt und die Antwortvariable numerisch ist, wurden unterschiedliche Regressionsverfahren angewandt. Die Regressionsmodelle wurden mittels des Trainingsdatensatzes trainiert und anschließend mit dem Testdatensatz, hinsichtlich der Korrektklassifizierungsrate, getestet. Die

[1] Da diese Information nicht eindeutig im Datensatz hinterlegt ist, wurde analysiert, wie oft ein Preis bei einem Lieferant mehrfach vorkommt. Es wird angenommen, dass sobald ein Preis mehrfach auftritt der Lieferant einen Preis pro Stück vertritt. Von der Einbeziehung der Daten von Carriern, die einen Preis pro Stück vertreten, wird abgesehen, da es in diesem Fall keine Einflussfaktoren gibt. Der Preis ist ein Festpreis.

Korrektklassifizierungsrate gibt an, wie oft die Werte der Antwortvariable aus dem Datensatz mit den vorhergesagten Werten übereinstimmen. [Raschka und Mirjalili (2018), S.36]
Ferner wurde die Höhe der Bestrafung der Regressionsfunktionen so gewählt, dass die Korrektklassifizierungsrate maximal ist. Dies wurde durch Testen unterschiedlicher Werte für den λ-Wert ermittelt.
Außerdem wurde getestet, ob die unterschiedlichen Prädiktorvariablen einen Mehrwert für das Modell bringen.
Schließlich wurde das Modell mit der höchsten Korrektklassifizierungsrate ausgewählt als Vorhersagemodell für den monatlichen Preis. [Raschka und Mirjalili (2018), S.36 f.]

Beim Testen der unterschiedlichen Modelle ist herausgekommen, dass das Merkmal Besiedlungsklasse keinen Mehrwert für die Modelle bringt und die Korrektklassifizierungsrate der Vorhersage verschlechtert. Aus diesem Grund wurde das Merkmal als Prädiktorvariable nicht weiter berücksichtigt.

Die höchste Korrektklassifizierungsrate wies das Modell auf, welches mit dem Ridge Regression Verfahren entworfen wurde. Dieses Modell erzielte beim Testen das beste Ergebnis. Mit einer Korrektklassifizierungsrate von 70,67% wird das korrekte Ergebnis vorhergesagt. Dieses Modell fungiert als Vorhersagemodell für die monatlichen Preise.
Der Bestrafungswert für das Modell wurde auf $\lambda = 34$ festgelegt.
Für die Vorhersagefunktion 2.3 der Ridge Regression ergibt sich folgendes [2]:

$$f_1(x) = -56,776x_1 + 36,543x_2 + 20,233x_3 + 447,223x_4 + a \tag{4.2}$$

- x_1 = Ausprägung des Merkmals rural
- x_2 = Ausprägung des Merkmals suburban
- x_3 = Ausprägung des Merkmals urban
- x_4 = Ausprägung des Merkmals Länge

Anhand der Höhe der einzelnen w-Werte ist zu erkennen, wie stark der Einfluss der Merkmale auf die monatlichen Preise ist. In diesem Fall hat das Merkmal Länge den stärksten Einfluss und somit die meiste Vorhersagekraft.

[2] Die w-Werte wurden durch Konstanten ersetzt, welche blau markiert sind. Für den w_0-Wert wurde die Konstante a eingesetzt. Diese dient zur Anonymisierung, da es sich um die Funktion handelt, die die konkreten Preise widerspiegelt.

4.2.2 Vorhersagemodell für den einmaligen Preis

Der Trainingsdatensatz, welcher die Grundlage für das Modell für den einmaligen Preis bildet, beinhaltet die Merkmale suburban, urban und rural, sowie die drei möglichen Besiedlungsklassen. Diese Merkmale bilden die Prädiktorvariablen. Zudem sind in dem Datensatz die einmalige Kosten enthalten. Dieses Merkmal fungiert als Antwortvariable. Ein Ausschnitt des Datensatzes ist in der Abbildung 4.3 dargestellt.

Rural	Suburban	Urban	Besiedlungsklasse 1	Besiedlungsklasse 2	Besiedlungsklasse 3	Einmalige Kosten
0.0	0.0	1.0	1.0	0.0	0.0	0.000000
0.0	0.0	1.0	1.0	0.0	0.0	460106.901900
1.0	0.0	0.0	0.0	0.0	1.0	271356.070500
0.0	1.0	0.0	1.0	0.0	0.0	5350.080254
0.0	0.0	1.0	0.0	1.0	0.0	5350.080254
0.0	0.0	1.0	1.0	0.0	0.0	0.000000
1.0	0.0	0.0	0.0	0.0	1.0	213385.214800
0.0	0.0	1.0	0.0	1.0	0.0	3945.948724

Prädiktorvariable | Antwortvariable

Abbildung 4.3: Ausschnitt des Trainingsdatensatzes für die Aufstellung eines Vorhersagemodells für den einmaligen Preis.

Da die Antwortvariable numerisch ist werden zuerst Regressionsmodelle aufgestellt. Die Modelle werden anhand des Trainingsdatensatzes trainiert und anschließend mittels des Testdatensatzes, hinsichtlich der Korrektklassifizierungsrate, getestet.

Da die aufgestellten Regressionsmodelle keine gute Korrektklassifizierungsrate aufwiesen, wurden, um ein besseres Modell zu erhalten, die einmaligen Kosten in Klassen aufgeteilt und Klassifikationsverfahren angewandt. [Vorgehensweise: van der Aalst (2016), S.93]
Die einmaligen Kosten wurden in siebzehn Kategorien eingeteilt. Dabei wurde versucht eine Gleichverteilung der Daten auf die Klassen zu erhalten. Die siebzehn Kategorien sind in der Tabelle 4.1 aufgeführt.

Anschließend wurden unterschiedliche Klassifikationsverfahren durchgeführt. Unter anderem wurde das K-Nearest-Neighbour Verfahren, Entscheidungsbäume, eine Stützvektormethode (engl. Support Vector Machine), ein Random Forest und das Naive Bayes Verfahren angewandt.

Kategorie	**Kosten von...**	**Kosten bis...**
Kategorie 1	0 EUR	1 EUR
Kategorie 2	1 EUR	1000 EUR
Kategorie 3	1.000 EUR	2.500 EUR
Kategorie 4	2.500 EUR	5.000 EUR
Kategorie 5	5.000 EUR	10.000 EUR
Kategorie 6	10.000 EUR	15.000 EUR
Kategorie 7	15.000 EUR	20.000 EUR
Kategorie 8	20.000 EUR	25.000 EUR
Kategorie 9	25.000 EUR	35.000 EUR
Kategorie 10	35.000 EUR	45.000 EUR
Kategorie 11	45.000 EUR	55.000 EUR
Kategorie 12	55.000 EUR	65.000 EUR
Kategorie 13	65.000 EUR	75.000 EUR
Kategorie 14	75.000 EUR	100.000 EUR
Kategorie 15	100.000 EUR	200.000 EUR
Kategorie 16	200.000 EUR	400.000 EUR
Kategorie 17	400.000 EUR	540.000 EUR

Tabelle 4.1: Preiskategorien der einmaligen Kosten.

Der Random Forest wies die höchste Korrektklassifizierungsrate beim Testen auf. Aus diesem Grund fungiert der Random Forest als Vorhersagemodell für den einmaligen Preis. Der entworfene Random Forest besteht aus zehn Entscheidungsbäumen, die nach dem Entropie Maß aufgestellt wurden. Er weist eine Korrektklassifizierungsrate von 42,91% auf. Ein Entscheidungsbaum aus dem Random Forest ist im Anhang A.4 dargestellt.
In der ersten Zeile der Knoten des Baumes wird eine Aussage geprüft. Zum Beispiel handelt es sich beim Wurzelknoten um die Frage, ob das Merkmal rural $<= 0,5$ ist. Das Merkmal rural kann nur die binären Ausprägungen 1 und 0 haben. Die Ausprägung 1, wenn es sich um eine Leitung handelt, die in einem ruralen Gebiet liegt und den Wert 0, wenn das Gegenteil zutrifft. Sollte das Merkmal die Ausprägung 1 besitzen und somit $>= 0,5$ sein, wird der ablehnende Pfad [3] gewählt. In der zweiten Zeile eines Knotens ist der Wert der Entropie abgebildet und die dritte Zeile gibt die Anzahl der Instanzen an, die an dem gegebenen Knoten vorherrschen. Dadrunter, unter der Bezeichnung "value", ist eine Liste angeben, aus der erkenntlich ist, wie viele Instanzen an dem gegebenen Knoten in die unterschiedlichen Preiskategorien fallen. Der letzte Wert "class" gibt an welche Vorhersage der gegebene Knoten erzeugt.
Die Farbgebung der Blätter resultiert aus dem Wert der Vorhersage der Klasse des Knotens, sowie der Stärke der Entropie.

In dem Wurzelknoten des Baumes wird examiniert, ob es sich bei der Leitung um eine Leitung handelt, die in einem ruralen Gebiet liegt. Sollte dies nicht der Fall sein wird der Pfad links gewählt. Es ist zu erkennen, dass alle weiteren Knoten links von dem Wurzelknoten eine sehr niedrige Preiskategorie vorhersagen. Liegt die Leitung dagegen in einem ruralen Gebiet und der rechte Pfad wird gewählt, ist erkennbar, dass die Knoten des rechten Unterbaums höhere Preiskategorien vorhersagen. Werden die Pfade des Baumes weiter nach unten verfolgt zu den unterschiedlichen Blättern des Baumes wird ersichtlich, dass Leitungen, die hinzukommend in einer hohen Besiedlungsklasse liegen, in eine hohe Preiskategorie fallen. Zum Beispiel wird für eine Leitung, die in einem ruralen Gebiet der Besiedlungsklasse drei liegt die Kategorie fünf vorhergesagt. Der einmalige Preis einer entsprechenden Leitung liegt dementsprechend zwischen 5.000 EUR und 10.000 EUR.

Die Prädiktormerkmale haben einen bestimmten Betrag an Wichtigkeit für das Modell.

[3] Bei dem Entscheidungsbaum im Anhang A.4 ist der ablehnende Pfad immer der rechte Pfad

Dieser Betrag ist für die unterschiedlichen Merkmale wie folgt:

- rural= 40,63%
- suburban = 8,51%
- urban = 15,88%
- Besiedlungsklasse 1 = 12,71%
- Besiedlungsklasse 2 = 8,27%
- Besiedlungsklasse 3 = 14,00%

In diesem Fall hat das Merkmal rural den meisten Einfluss auf das Vorhersagemodell, dementsprechend hat dieses Merkmal die meiste Vorhersagekraft.

Das Modell des Random Forest fungiert als Vorhersagemodell für die einmaligen Preise einer Leitung. Berücksichtigt in dem Modell wird die geographische Lage, ausgeprägt durch die Merkmale Area Class und die Besiedlungsklasse.

4.3 Entwurf eines Trendmodells

Im folgenden Abschnitt wird das entworfene Trendmodell der monatlichen Preise pro Meter vorgestellt. Anschließend wird die Verknüpfung des Trendmodells mit dem entworfenen Vorhersagemodell für den monatlichen Preis erläutert.

Ein Trendmodell für die einmaligen Preise wurde nicht entwickelt, da das Ergebnis der Analyse der Preise zeigte, dass kein zeitlicher Trend existiert.

4.3.1 Trendmodell des monatlichen Preises pro Meter

Das Trendmodell baut auf den Ergebnissen des Kapitels 3.4.2 auf. Um einen zeitlichen Trend zu bestimmen, wird eine Funktion gesucht, die den Trend, welcher in Abbildung 3.17 zu erkennen ist, am Besten beschreibt. Es existieren lineare Trendmodelle, sowie nicht lineare Trendmodelle. Ein typisches nicht lineares Trendmodell ist das Exponentialmodell. Bei einer exponentiellen Entwicklung einer Zeitreihe stimmt der Verlauf mit einer exponentiellen Funktion überein.

Um eine Funktion zu finden, die die Punkte am besten approximiert werden Regressionsverfahren angewandt. [Schlittgen und Streitberg (1999), S.12 ff.]
Bewertet wurde die Regressionsfunktionen nach dem Bestimmtheitsmaß. Das Bestimmtheitsmaß wird mit R^2 bezeichnet und bewertet die Anpassungsgüte einer Regressionsgerade an den Datensatz. R^2 nimmt einen Wert zwischen 0 und 1 an. Eine hohe Anpassungsgüte ist vorhanden, wenn $R^2 = 1$. [Rottmann u. a. (2018)] [Grus (2016), S.193]

Das Ergebnis zeigt, dass ein Exponentialmodell ein höheres Bestimmtheitsmaß, mit $R^2 = 0{,}541$ aufweist, als ein lineares Modell. Aus diesem Grund wurde für das Exponentialmodell entschieden. Ein Diagramm[4], in dem der Trend der Preise pro Meter dargestellt ist, sowie die exponentielle Regressionsfunktion ist in Abbildung 4.4 ersichtlich.

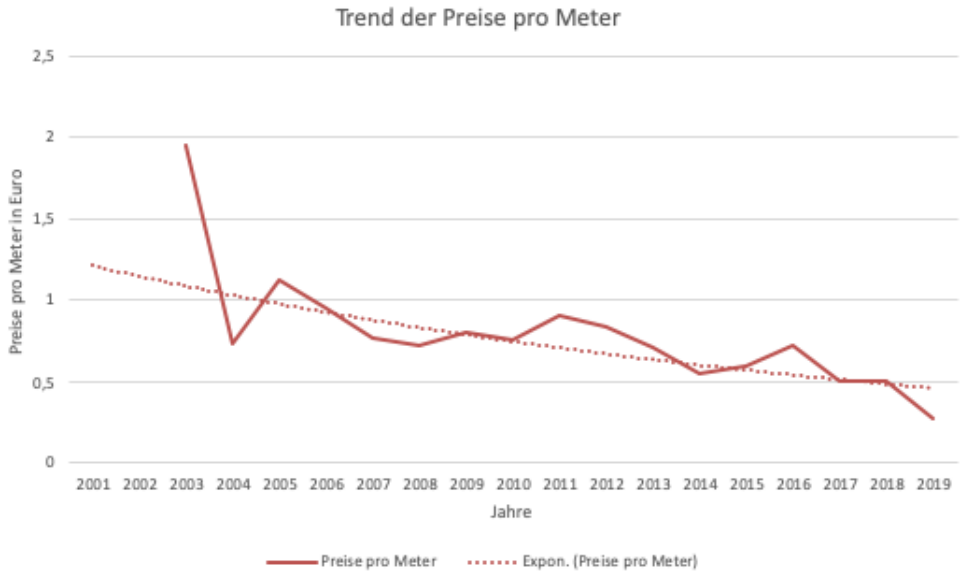


Abbildung 4.4: Darstellung der Preise pro Meter im zeitlichen Verlauf, sowie der erstellten exponentiellen Regressionsfunktion.

[4]Zur Darstellung wurden nicht die reellen Preis pro Meter verwendet.

Die Gleichung der Exponentialfunktion, die die Punkte am besten approximiert lautet[5]:

$$f_2(x) = b \cdot e^{-0{,}054x} \tag{4.3}$$

Bei dieser Exponentialfunktion handelt es sich um eine e Funktion. Diese e Funktion spiegelt die Funktion zur Berechnung einer stetigen Verzinsung wieder. Folgende Funktion wird für die stetigen Verzinsung genutzt:

$$K_{\mathrm{t}} = K_0 \cdot e^{r \cdot t} \tag{4.4}$$

- $\mathrm{K_t}$ = Endwert
- $\mathrm{K_0}$ = Anfangswert
- r = Zins
- t = Anzahl der Perioden

Es ist ersichtlich, dass beide Formeln von der Struktur übereinstimmen. Der Wert r ist der Zinswert, dieser nimmt den Wert r = 0,054 in der aufgestellten Funktion 4.3 an. [Schölisch (2018)]
Inhaltlich bedeutet dies, dass der Preis pro Meter Glasfaser pro Jahr stetig abnimmt um 5,4%.

4.3.2 Konkatenation der Modelle

Dieser Abschnitt beschäftigt sich mit der Verknüpfung des Vorhersagemodells für den monatlichen Preis und dem Trendmodell des monatlichen Preises pro Meter. Das verknüpfte Modell dient zur Vorhersage der monatlichen Preise eines Mietangebotes für eine Glasfaserleitung.

Es wird davon ausgegangen, dass der Preis pro Meter in der Zukunft weiterhin der ermittelten Funktion 4.3 folgen wird und dementsprechend pro Jahr um 5,4% abnimmt. Um den preislichen Trend in das entwickelte Vorhersagemodell zu integrieren, wird die Formel der stetigen Verzinsung 4.4 genutzt. Der vorhergesagte monatliche Preise, welcher

[5] Da es sich um die Formel handelt, die die konkreten Preise wider spiegelt, wurde zur Anonymisierung die Konstante b eingeführt. Er ist nicht relevant für den übrigen Kontext.
Zudem entspricht der x-Wert der Anzahl der vergangenen Jahre seit dem ersten, in dem Datensatz vorkommenden Jahr. Das Jahr 2001 ist dementsprechend Jahr Nummer 1

über die entworfene Funktion 4.2 ermittelt werden kann, fungiert als Wert für K_0 (blau markiert). Der Preis muss nicht auf einen Preis pro Meter berechnet werden, obwohl der Trend für den Preis pro Meter ermittelt wurde. Aus mathematischer Sicht ergibt sich dasselbe Ergebnis, wenn der Preis pro Meter berechnet wird, anschließend der neue Preis pro Meter und dies wieder auf einen Gesamtpreis berechnet wird, wie wenn der Trend auf den Gesamtpreis berechnet wird.
Der Zinswert r entspricht, den aus der Exponentialfunktion ermittelten Trend (lila markiert) und der Wert t entspricht der Anzahl der seit dem heutigen Jahr vergangenen Jahre in dem der Preis prognostiziert werden soll. Somit ergibt sich folgende Formel:

$$f_3(t) = f_1(x) \cdot e^{-0{,}054 \cdot t} \tag{4.5}$$

Anhand dieses Modells kann ein marktkonformer monatlicher Preis für eine Leitung bestimmt werden. Berücksichtigt dabei wird die Länge einer Leitung, die Area Class der Standorte, sowie der Trend der Preise.

4.4 Prognose der inflationären Entwicklung der Preise

Als Inflation wird der anhaltende Prozess der Geldentwertung bezeichnet. Dies macht sich durch eine allgemeine Preiserhöhung bemerkbar. Der prozentuale Preisanstieg des Preisniveaus wird als Inflationsrate bezeichnet. [Pollert u. a. (2016)]
Die europäische Zentralbank prognostiziert für die nächsten 5 Jahre eine Inflationsrate von 1,8%. [European Central Bank (2019)]
Da die Preiserhöhung durch die Inflation Einfluss auf die zukünftigen Preise eines Mietangebotes einer Glasfaserleitung hat, soll dies in den Modellen zur Vorhersage eines marktkonformen Preises berücksichtigt werden. Dafür wird die prognostizierte Inflationsrate für die nächsten 5 Jahre von der europäischen Zentralbank verwendet.

4.4.1 Beachtung der inflationären Entwicklung des monatlichen Preises im Modell

Um die inflationäre Entwicklung der monatlichen Preise zu berücksichtigen, wird das bereits entworfene Modell zur Vorhersage eines marktkonformen monatlichen Preises erweitert. Die Preissteigerung, die die Inflation verursachst wird mit ein berechnet. Es

wurde prognostiziert, dass die Inflationsrate der nächsten 5 Jahre 1,8% beträgt. Der monatliche Preis wird dementsprechend jedes Jahr um 1,8% zu nehmen.

Um den Effekt der Inflation zu berechnen, wird die Formel 3.1 zur diskreten Verzinsung verwendet. Der für den heutigen Zeitpunkt prognostizierte Preis, ermittelt über die Funktion 4.2, fungiert als K_0 (blau markiert). Der Zinssatz entspricht der Inflationsrate (rot markiert) und die Anzahl der Jahre n wird entsprechend der Anzahl der vergangenen Jahre seit dem heutigen Jahr gesetzt. Folgende Funktion ergibt sich:

$$K_{\mathrm{n}} = f_1(x) \cdot (1 + 0,018)^n \tag{4.6}$$

Der monatlichen Preis in n Jahren unter Berücksichtigung der Inflation ist mit dieser Funktion errechenbar.
Diese Funktion wird mit der Funktion 4.5 verknüpft, um ein Modell zur Vorhersage der monatlichen Preise unter Berücksichtigung der zeitlichen Entwicklung zu erhalten. Die zeitliche Entwicklung der Preise ist geprägt durch den Trend sowie die Inflation. Dafür wird der zweite Faktor der Funktion 4.6 in der Funktion 4.5 ergänzt (rot markiert). Die Variable n wird durch die Variable t ersetzt, da beide Variablen die Anzahl der Jahre angeben. Es ergibt sich folgende Formel:

$$f_4(x) = f_1(x) \cdot e^{-0,054 \cdot t} \cdot (1 + 0,018)^t \tag{4.7}$$

Aufgrund des Kommutativgesetzes ist es unerheblich, ob zuerst der Trend auf den Preis berechnet wird oder die Inflation.

Dieses Modell ermöglicht es einen monatlichen Preis vorherzusagen unter Berücksichtigung der Einflussfaktoren, sowie der zeitlichen Entwicklung. Allerdings ist zu beachten, dass das gesamte Modell nach einigen Jahren hinsichtlich der zeitlichen Komponenten überarbeitet werden muss. Der Trend der Preise wird sich wahrscheinlich ändern und die Inflation beruht auf einen für 5 Jahre voraus prognostizierten Wert. Dies Modell liefert, aufgrund der Änderungen der zeitlichen Komponenten, nach 5 Jahre keine Zuverlässigen Ergebnisse mehr.

4.4.2 Beachtung der inflationären Entwicklung des einmaligen Preises im Modell

Die Inflation hat einen Einfluss auf den einmaligen Preis eines Mietangebotes einer Glasfaserleitung. Es wird jedoch davon abgesehen die Inflation in dem Modell zur Vorhersage eines marktkonformen einmaligen Preises zu berücksichtigen. Der Grund dafür besteht darin, dass das Vorhersagemodell für den einmaligen Preis eine Preiskategorie vorhersagt und angenommen wird, dass die Wahrscheinlichkeit dafür, dass der vorhergesagte einmalige Preis in die gleiche Kategorie fällt, wie der vorhergesagt einmalige Preis unter Berücksichtigung der Inflation, gleich hoch ist.

4.5 Bewertung der Modelle

Die Bewertung des Modells für den einmaligen Preis, sowie des Modells für den monatlichen Preis wird anhand des Kriterienbogen Tabelle 4.2 vollzogen.
Das Kriterium *Zuverlässigkeit* beurteilt die Zuverlässigkeit der Vorhersagen des Modells. Zur Bewertung dieses Kriteriums wird der Wert der Korrektklassifizierungsrate, welcher beim Testen der Modelle berechnet wurde, herangezogen.
Das Kriterium *Komplexität* des Modells beurteilt die Anzahl der benötigten Elemente zur Berechnung einer Vorhersage und wie die unterschiedlichen Elemente in Verbindung miteinander stehen. [Feess (2018)]
Hinzu kommt das Kriterium *Effizienz*. Um die Effizienz der Modelle bewerten zu können wurden 100.000 zufällige Testdaten erzeugt und es wurde untersucht, wie lange es dauert, bis das Modell eine Vorhersage für alle Instanzen des Datensatzes berechnet hat.
Ferner wird das Modell hinsichtlich des Kriterium *Informationsgewinn* bewertet. Dafür wird der Informationsgehalt der Ausgabewerte beurteilt.

Bewertungsaspekt	**Bewertung**			
Zuverlässigkeit	(- -)	(-)	(+)	(++)
Komplexität	(- -)	(-)	(+)	(++)
Effizienz	(- -)	(-)	(+)	(++)
Informationsgewinn	(- -)	(-)	(+)	(++)
Gesamtbewertung	(- -)	(-)	(+)	(++)

Tabelle 4.2: Bewertungsbogen für die Vorhersagemodelle.

4.5.1 Bewertung des Modells für den monatlichen Preis

Die Bewertung des Modells für den monatlichen Preis ist in der Tabelle 4.3 dargestellt. Die gelb hervorgehobenen Felder kennzeichnen die Bewertung.
Das Kriterium *Zuverlässigkeit* des Modells wurde mit einem Plus bewertet. Der Grund dafür liegt darin, dass die Korrektklassifizierungsrate des Vorhersagemodells anhand der Ausprägungen der Einflussfaktoren bei 70,67% liegt und das Bestimmtheitsmaß, also die Anpassungsgüte der Funktion des Trendmodells bei R^2=0,541 liegt. Die Zuverlässigkeit des Modells ist dementsprechend gut.
Das Kriterium *Komplexität* erhielt eine niedrigere Bewertung. Die komplette Formel zur Vorhersage eines marktkonformen Preises unter Berücksichtigung der zeitlichen Entwicklung weist einen höheren Grad an Komplexität auf, da sie aus drei Termen zusammengesetzt ist. Allerdings handelt es sich nur um eine relativ geringe Anzahl an Merkmalen die zur Berechnung eines marktkonformen Preises notwendig sind.
Das Kriterium *Effizienz* des Modells erhielt eine sehr gute Bewertung. Das Modell benötigte für eine Vorhersage eines monatlichen Preises für 100.000 Angebote 0,012 sec. Dies ist ein sehr gutes Ergebnis, gerade weil das Modell für die Praxis nicht dafür ausgelegt sein muss 100.000 Angebote von Leitungen auf einmal zu evaluieren.
Ferner wurde das Kriterium *Informationsgewinn* mit sehr gut bewertet, da es möglich ist mit dem Modell konkrete Preise vorherzusagen und der zeitliche Einfluss mit berücksichtigt wird.
Die Gesamtbewertung ergibt ein Ergebnisse zwischen gut und sehr gut. Da die Wichtigkeit des Kriterium *Zuverlässigkeit* als am höchsten angesehen ist, resultiert daraus eine gute Gesamtbewertung.

Bewertungsaspekt	**Bewertung**			
Zuverlässigkeit	(- -)	(-)	(+)	(++)
Komplexität	(- -)	(-)	(+)	(++)
Effizienz	(- -)	(-)	(+)	(++)
Informationsgewinn	(- -)	(-)	(+)	(++)
Gesamtbewertung	(- -)	(-)	(+)	(++)

Tabelle 4.3: Bewertung des Modells für den monatlichen Preis.

4.5.2 Bewertung des Modells für den einmaligen Preis

Die Bewertung des Modells für den einmaligen Preis ist in der Tabelle 4.4 dargestellt. Die gelb hervorgehobenen Felder kennzeichnen die Bewertung.
Die Bewertung des Kriterium Zuverlässigkeit fiel nicht gut aus, da das Modell nur eine Korrektklassifizierungsrate von 42,91% bietet.
Im Gegensatz dazu erhielt das Kriterium *Komplexität* eine sehr gute Bewertung. Nur wenige Merkmale werden zur Berechnung einer Vorhersage der einmaligen Preise benötigt. Eine Interpretation der Zusammenhänge der Merkmale ist anhand der Entscheidungsbäume aus dem Random Forest einfach. Ein Entscheidungsbaum stellt ein leicht verständliches Modelle dar.
Das Kriterium *Effizienz* wurde gut bewertet. Eine Vorhersage der einmaligen Preise für einen Testdatensatz mit 100.000 Instanzen dauerte 0.22 sec. Dies ist ca. 18 mal länger als die Dauer, die das Modell für die monatlichen Preise für die Vorhersage benötigt, aber trotzdem immer noch schnell.
Das Kriterium *Informationsgewinn* wurde schlecht bewertet. Der Grund dafür liegt darin, dass das Modell nur Preisekategorien voraussagt. Außerdem entscheiden die Entscheidungsbäume des Random Forest meistens nur zwischen fünf Kategorien. Der Grund dafür liegt vermutlich darin, dass nur eine annähernde Gleichverteilung der Anzahl der Daten pro Klasse in dem Trainingsdatensatz gegeben war.
Die Gesamtbewertung des Modells findet sich in der Mitte des Bewertungsbogen wieder. Da wiederum die Wichtigkeit des Kriterium Zuverlässigkeit als am höchsten angesehen wird, erhält das Modell insgesamt eine geringere Gesamtbewertung.

Bewertungsaspekt	**Bewertung**			
Zuverlässigkeit	(- -)	(-)	(+)	(++)
Komplexität	(- -)	(-)	(+)	(++)
Effizienz	(- -)	(-)	(+)	(++)
Informationsgewinn	(–)	(-)	(+)	(++)
Gesamtbewertung	(- -)	(-)	(+)	(++)

Tabelle 4.4: Bewertung des Modells für den einmaligen Preis.

5 Fazit und Ausblick

Im Kontext dieser Bachelorarbeit wurden zwei Modelle entworfen. Eins zur Vorhersage eines marktkonformen monatlichen Preises, sowie ein weiteres zur Vorhersage eines marktkonformen einmaligen Preises für ein Mietangebot einer Glasfaserleitung. Das Modell für den monatlichen Preis weist im Gegensatz zu dem Modell für den einmaligen Preis eine gute Gesamtbewertung auf. Es kann einem Telekommunikationsunternehmen einen Mehrwert bieten und als Unterstützung im Beschaffungsprozess dienen. Das Modell für den einmaligen Preis hingegen weist eine insgesamt eher niedrigere Bewertung auf. Vor allem da die Korrektklassifizierungsrate des Modells sehr gering ist, wird dem Modell kein großer Mehrwert zugesprochen.

Dieses Ergebnis kann verbessert werden, wenn mehrere Informationen zu einer gemieteten Glasfaserleitung vorliegen würden und die Daten ordentlich gespeichert werden würden. Im Kontext dieser Arbeit konnte nachgewiesen werden, dass die geographische Lage der Leitung und die Länge der Leitung einen Einfluss auf die monatlichen Preise besitzen und die geographische Lage einen Einfluss auf den einmaligen Preis besitzt. Durch weitere Daten besteht die Möglichkeit nachzuweisen, ob die Bodenbeschaffenheit und die Länge des Tiefbaus einen Einfluss auf den Preis haben. So ein Ergebnis könnte die Möglichkeit bieten, neue Modelle zu entwickeln oder die bestehenden Modelle zu verbessern, um dadurch ein zuverlässigeres Modell zu erhalten.
Aus diesem Grund ergibt sich folgende Handlungsempfehlung für das betrachtete Telekommunikationsunternehmen:

1. Eine Speicherung der Daten über die Länge des Tiefbaus ist sinnvoll. Die Länge des Tiefbaus stellt wahrscheinlich einen Faktor dar, der Einfluss auf die Preise besitzt. Leider konnte dies im Zuge der Arbeit nicht nachgewiesen werden, da keine Informationen zu der Länge des Tiefbaus einer Leitung hinterlegt sind.

2. Zudem würde eine Erhebung der Daten der jeweiligen Bodenklasse sinnvoll sein. Da bereits bewiesen ist, dass die Bodenklasse einen Kosten bestimmenden Einfluss auf

den Bau einer Glasfaserleitung besitzt, wird sie zu einer hohen Wahrscheinlichkeit auch die Mietpreise beeinflussen. Dieser Einfluss konnte in dem Kontext dieser Arbeit nicht nachgewiesen werden. Angenommen wird, dass der Grund dafür in der Erhebung der Daten über die jeweiligen Bodenklassen liegt. Die Informationen über die Bodenklassen wurden über mehrere alternative Wege erhoben und weisen deshalb vermutlich einen höheren Grad an Ungenauigkeit auf.

3. Ferner würde eine Angaben darüber, ob der Lieferant einen festen monatlichen Preis für eine Leitung verlangt die Analyse verbessern. Da einige Lieferanten einen monatlichen Festpreis fordern, existieren in so einem Fall keine preislichen Einflussfaktoren. Eine Analyse der Einflussfaktoren auf die monatlichen Preise wäre in diesem Fall nicht nötig. Fordert der Lieferant jedoch keinen festen Preis ist eine Analyse der Einflussfaktoren sinnvoll.
   Würden Angaben über die Preismodelle der Lieferanten existieren, wäre eine Trennung der Daten für die Analyse problemloser und man würde vermutlich aussagekräftigere Ergebnisse erhalten.

4. Zudem würde eine Speicherung von Negativdaten einen Mehrwert für die Analyse bieten. Mit Negativdaten sind in diesem Fall Mietangebote gemeint, die das betrachtete Telekommunikationsunternehmen nicht angenommen hat. Wären solche Daten vorhanden, wäre es möglich zu analysieren aus welchen Gründen Angebote nicht angenommen wurden. Die Integration solcher Informationen in das Ergebnis könnte dieses verbessern. Zudem könnte man in eine Applikation integrieren, ob entsprechende Angebote in der Vergangenheit eher angenommen oder abgelehnt wurden.

5. Ferner würde eine ordentliche Speicherung der bereits vorhandenen Daten schon einen großen Mehrwert für die Analysen ergeben. Die vorgegebene Struktur der Daten ist sinnvoll, wurde jedoch teils nicht befolgt. Viele Daten lagen in einer Form vor die nicht optimal ist und wurden aus diesen Grund bei der Analyse und bei der Entwicklung der Modelle nicht berücksichtigt.

Es ist vorstellbar bessere Ergebnisse zu erzielen würden diese Daten erhoben und ordentlich abgespeichert werden.

Es besteht die Möglichkeit die entworfenen Modelle als Grundlage für eine Applikation zu nutzen, die ermittelt, ob es sich bei einem Angebot einer Glasfaserleitung um einen

marktkonformen Preis handelt. Ob die Verfolgung dieses Ansatzes sinnvoll ist, ist fragwürdig. Die entworfenen Modelle, die als Grundlage für die Applikation dienen sollen, bieten keine hohe Zuverlässigkeit der Vorhersagen. Die Aussagen der Applikation würden dementsprechend auch keine hohe Zuverlässigkeit bieten. Es wird jedoch eine Applikation gewünscht, die eine hohe Zuverlässigkeit bietet, damit sie einen Hilfestellung im Beschaffungsprozess darstellt. Somit existieren grundsätzlich zwei Varianten für die Zukunft.

1: Der Ansatz einer Applikation wird weiter verfolgt und das Ziel besteht darin eine möglichst generische Applikation zu entwickeln. Es soll die Möglichkeit bestehen Faktoren der Modelle problemlos zu ändern und die GUI zu erweitern, so dass es möglich ist ein verbessertes Modell ohne viele Änderungen der Applikation zu integrieren.

2: Der Fokus der Zukunft liegt darauf Modelle zu entwickeln, welche eine höhere Korrektklassifizierungsrate bieten. Entweder durch die Verbesserung der entworfenen Modelle oder durch das entwerfen neuer Modelle. Um dies zu erreichen müssten mehr Daten erhoben werden und ordentlich abgespeichert werden.
Der Aufwand der Implementierung einer Applikation, die keine hohe Zuverlässigkeit bietet wird als zu hoch angesehen. Erst wenn Modelle entstanden sind, die eine höhere Zuverlässigkeit bieten wird der Ansatz einer Applikation weiter verfolgt.

Welche Variante weiter verfolgt wird wird durch das betrachtete Telekommunikationsunternehmen entschieden.
Persönlich wird folgende Meinung vertreten:
Eine Verbesserung der Zuverlässigkeit der Modelle setzt vorerst voraus, dass die Daten ordentlich gespeichert werden und weitere Daten erhoben werde. Dies ist ein Prozess der lange dauern kann und nebenbei durchgeführt werden kann. Somit würde ich die Zukunftsvariante 1 empfehlen.

Literaturverzeichnis

[van der Aalst 2016] Aalst, W. van der: *Process Mining - Data Science in Action*. Springer-Verlag, 2016. – ISBN 978-3-662-49851-4

[Breitband Kompetenz Zentrum, Niedersachsen] Breitband Kompetenz Zentrum, Niedersachsen: *Infrastrukturatlas der Bundesnetzagentur*. – URL https://www.breitband-niedersachsen.de/index.php?id=382. – Zugriffsdatum: 18.07.2019

[Bundesanstalt für Geowissenschaften und Rohstoffe 2014] Bundesanstalt für Geowissenschaften und Rohstoffe: *Bodenübersicht von Deutschland 1 : 3 000 000*. 2014. – URL https://produktcenter.bgr.de/terraCatalog/Start.do. – Zugriffsdatum: 12.07.2019

[Bundesnetzagentur 2019] Bundesnetzagentur: *Konsulationsentwurf zur Genehmigung von Entgelten für Carrier-Festverbindungen (CFV Ethernet 2.0), die jeweils zugehörige Expressentstörung und weitere Leistungen*. 2019. – URL https://www.bundesnetzagentur.de/DE/Service-Funktionen/Beschlusskammern/1_GZ/BK2-GZ/2018/2018_001bis099/BK2-18-0003/BK2-18-0003_Konsultationsentwurf_Download_BA.pdf?__blob=publicationFile&v=3. – Zugriffsdatum: 12.07.2019

[Czarnecki und Dietze 2017] Czarnecki, C. ; Dietze, C.: *Reference Architecture for the Telecommunications Industry*. Springer Verlag, 2017. – ISBN 978-3-319-46755-9

[European Central Bank 2019] European Central Bank: *HICP Inflation forecasts*. 2019. – URL https://www.ecb.europa.eu/stats/ecb_surveys/survey_of_professional_forecasters/html/table_hist_hicp.en.html. – Zugriffsdatum: 22.07.2019

[Feess 2018] Feess, E.: *Gabler Wirtschaftslexikon - Komplexität*. 2018. – URL https://wirtschaftslexikon.gabler.de/definition/komplexitaet-39259/version-262672. – Zugriffsdatum: 19.07.2019

[Grus 2016] Grus, J.: *Einführung in Data Science - Grundprinzipien der Datenanalyse mit Python*. O'Reilly Media, 2016. – ISBN 978-3-96009-021-2

[Han und Kamber 2011] Han, J. ; Kamber, M.: *Data Mining - Concepts and Techniques*. 2011. – ISBN 978-0-12-381479-1

[Hand u. a. 2001] Hand, D. ; Mannila, H. ; Smyth, P.: *Principles of Data Mining*. 2001. – ISBN 978-0262082907

[Held 2010] Held, U.: *Tücken von Korrelation: die Korrelationskoeffizienten von Pearson und Spearman*. 2010. – URL https://medicalforum.ch/de/resource/jf/journal/file/view/article/smf/de/smf.2010.07285/fdd6ee849e3394e1313f533b5e722403cc6d5f44/2010_38_130.pdf/. – Zugriffsdatum: 15.07.2019

[Horsch 2018] Horsch, A.: *Gabler Wirtschaftslexikon - diskrete Verzinsung*. 2018. – URL https://www.gabler-banklexikon.de/definition/diskrete-verzinsung-99951/version-347058. – Zugriffsdatum: 15.08.2019

[Kulenkampff u. a. 2019] Kulenkampff, G. ; Plückebaum, T. ; Zoz, K.: *Analytisches Kostenmodell für das Anschlussnetz*. WIK-Consult, 2019

[Landesamt, für Bergbau, Energie und Geologie] Landesamt, für Bergbau, Energie und Geologie: *Bodenklassen für Erdbarbeiten nach DIN 18300 (IBOKLA 50)*. – URL https://www.lbeg.niedersachsen.de/karten_daten_publikationen/karten_daten/baugrund/bodenklassen_erdarbeiten_nach_din_18300/bodenklassen-fuer-erdarbeiten-nach-din-18300-ibokla-50-621.html. – Zugriffsdatum: 30.07.2019

[Loiber 2018] Loiber, H.: *Planungsleitfaden Breitband - Leitfaden zur Planung und Errichtung von Glasfaser-Zugangsnetzen*. BMVIT, 2018. – URL https://www.bmvit.gv.at/service/publikationen/telekommunikation/downloads/planungsleitfaden_outdoor_ua.pdf. – Zugriffsdatum: 18.07.2019

[Moser und Schmidt 2011] Moser, K. ; Schmidt, F.: *Beschreibende Statistik und Wirtschaftsstatistik*. Springer Verlag, 2011. – ISBN 978-3-540-37459-6

[Pedregosa u. a. 2011] Pedregosa, F. ; Varoquaux, G. ; Gramfort, A. ; Michel, V. ; Thirion, B. ; Grisel, O. ; Blondel, M. ; Prettenhofer, P. ; Weiss, R. ;

Dubourg, V. ; Vanderplas, J. ; Passos, A. ; Cournapeau, D. ; Brucher, M. ; Perrot, M. ; Duchesnay, E.: Scikit-learn: Machine Learning in Python. In: *Journal of Machine Learning Research* 12 (2011), S. 2825–2830

[Pollert u. a. 2016] Pollert, A. ; Kirchner, B. ; Polzin, J.M. ; Pollert, M.C.: *Duden Wirtschaft von A bis Z: Grundlagenwissen für Schule und Studium, Beruf und Alltag.* Bibliographisches Institut, 2016 (Duden Spezialwörterbücher). – ISBN 9783411912155

[Raschka und Mirjalili 2018] Raschka, S. ; Mirjalili, V.: *Machine Learning mit Python und Scikit-Learn und TensorFlow - Das umfassende Praxis-Handbuch für Data Science, Deep Learning und Predictive Analytics.* mitp Verlag, 2018. – ISBN 978-3-95845-735-5

[Rottmann u. a. 2018] Rottmann, H. ; Auer, B. ; Kamps, U.: *Gabler Wirtschaftslexikon - Bestimmtheitsmaß.* 2018. – URL https://wirtschaftslexikon.gabler.de/definition/bestimmtheitsmass-31758/version-255309. – Zugriffsdatum: 24.07.2019

[Schlittgen und Streitberg 1999] Schlittgen, R. ; Streitberg, B.: *Zeitreihenanalyse.* Oldenbourg Verlag, 1999. – ISBN 3-486-24982-7

[Schölisch 2018] Schölisch, D.: *Gabler Wirtschaftslexikon - stetige Verzinsung.* 2018. – URL https://www.gabler-banklexikon.de/definition/stetige-verzinsung-61583/version-344019. – Zugriffsdatum: 20.08.2019

[Seidel 2006] Seidel, H.: *Die Mathematik der Gauß-Krüger-Abbildung.* 2006. – URL http://henrik-seidel.gmxhome.de/gausskrueger.pdf. – Zugriffsdatum: 16.07.2019

[Statistisches Amt für Hamburg und Schleswig-Holstein 2018] Statistisches Amt für Hamburg und Schleswig-Holstein: *Datentabelle für die interaktive Karte der Hamburger Stadtteil-Profile: Berichtsjahr 2017.* 2018. – URL https://www.statistik-nord.de/zahlen-fakten/regionalstatistik-datenbanken-und-karten/hamburger-stadtteil-profile-und-interaktive-karten/. – Zugriffsdatum: 16.07.2019

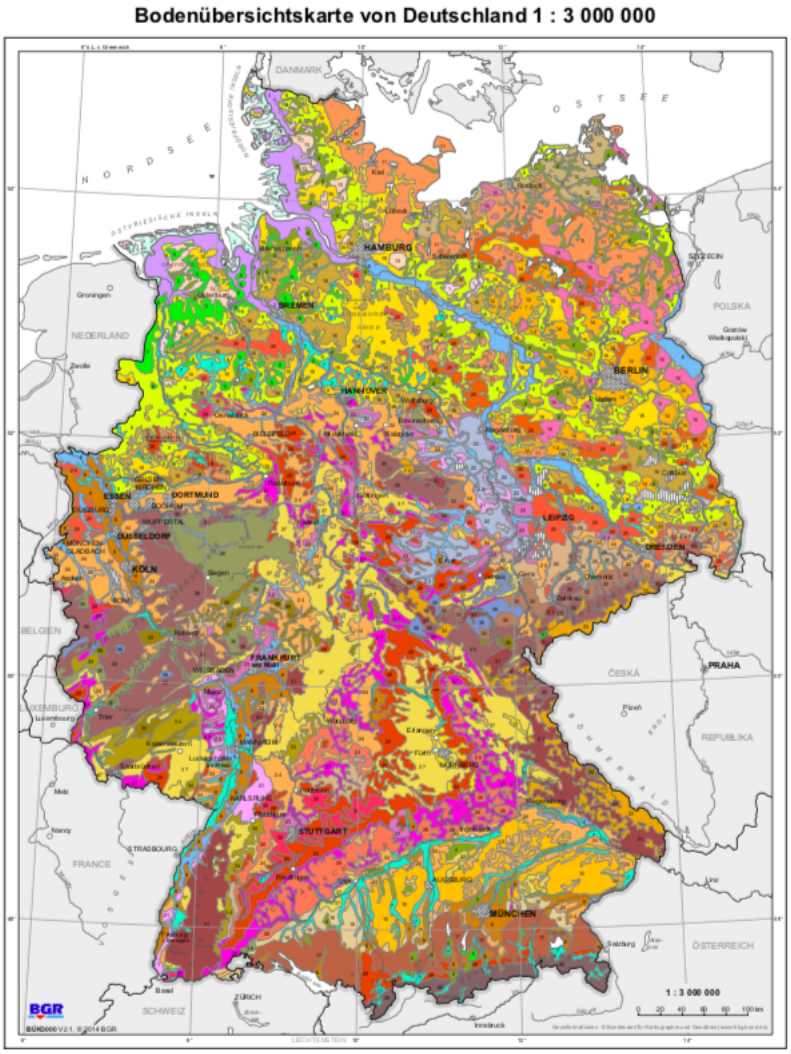
[Statistisches Bundesamt 2018] Statistisches Bundesamt: Bevölkerung, Familien, Lebensformen. In: *Statistisches Jahrbuch 2018* (2018), S. 59.

– URL https://www.destatis.de/DE/Themen/Querschnitt/Jahrbuch/jb-bevoelkerung.html. – Zugriffsdatum: 15.07.2019

[Telefónica Germany GmbH und Co. OHG 2009] Telefónica Germany GmbH und Co. OHG: *Anbindung einer neuen Leitung an einem bestehenden Anschlusspunkt.png*. 2009

A Anhang

A.1 Bodenübersicht von Deutschland




LEGENDE
Böden der Küstenregion und Moore
1 Podsol-Regosol und Regosol aus trockenen, nährstoffarmen Sanden
2 Wattboden im Gezeitenbereich der Nordsee
3 Marschboden, vorwiegend Kleimarsch aus brackischen oder fluviatilen Ablagerungen sowie Kalkmarsch aus marinen Ablagerungen
4 Niedermoorboden aus mächtigen Niedermoortorfen, teils mit mineralischen Zwischenschichten, Moorgley und Gley
5 Hochmoorboden aus mächtigen Hochmoortorfen über Niedermoortorf, Mudde oder Mineralboden
Böden der breiten Flusstäler einschließlich Terrassenflächen und Niederungen
6 Auenboden, in tieferen Lagen Gley aus lehmigen bis tonigen Auensedimenten, in Schwarzerdegebieten Gley-Tschernosem aus kalkhaltigen, tonig-schluffigen Ablagerungen
7 Auenboden, in tieferen Lagen Gley aus sandigen bis tonigen Flusssedimenten, häufig in kleinflächigem Wechsel
8 Parabraunerde aus schluffig-lehmigen Deckschichten und Auenpararendzina aus kalkhaltigen, sandig lehmigen Hochflut- und Auenablagerungen. Podsol-Baunerde und Braunerde der sandigen Terrassen
9 Podsol, Braunerde-Podsol und Gley-Podsol aus sandigen Flussablagerungen sowie Gley der sandigen Urstromtäler und Niederungen
Böden des wellig-hügeligen Flachlandes und der Hügelländer
10 Braunerde, Parabraunerde und Pararendzina aus lössvermischten Tertiärablagerungen
11 Parabraunerde, Fahlerde und Pseudogley aus Geschiebelehm über Geschiebemergel
12 Pseudogley-Gley und Pseudogley aus lehmig-sandigem Geschiebemergel
13 Podsol-Parabraunerde und Podsol-Fahlerde aus sandigen Deckschichten über Geschiebelehm
14 Fahlerde und Bänderparabraunerde sowie Braunerde und Podsol-Braunerde aus sandigen Deckschichten über Geschiebelehm
15 Bänderparabraunerde, Fahlerde und Braunerde sowie Pararendzina und Regosol im engräumigen Wechsel der sandigen bis lehmigen Endmoränen
16 Pseudogley-Braunerde und Pseudogley-Fahlerde sowie Pseudogley aus Geschiebedecksand über Geschiebelehm
17 Parabraunerde, Braunerde und Pararendzina aus lehmig-sandigen, kalkhaltigen Moränenablagerungen im Alpenvorland
18 Braunerde, Parabraunerde und Pseudogley aus kalkhaltigen, lehmig-sandig-kiesigen, lössvermischten Moränenablagerungen im Alpenvorland
19 Podsol-Braunerde, Podsol-Bänderparabraunerde und Pseudogley-Podsol aus trockenen, nährstoffarmen Sanden
20 Braunerde, Bänderparabraunerde und Podsol-Braunerde aus nährstoffreichen Sanden
Böden der Berg- und Hügelländer sowie der Mittelgebirge aus Festgestein, dessen Verwitterungsmaterial und Umlagerungsdecken
21 Tschernosem bzw. Pararendzina aus Löss im Wechsel mit Rendzina aus Mergel- und Kalkstein
22 Tschernosem der Mitteldeutschen Trockengebiete aus mächtigem Löss, Tschernosem und Pseudogley-Tschernosem aus Löss über Ton- und Mergelstein

Abbildung A.1: Bodenübersichtskarte von Deutschland 1 : 3 000 000. Quelle: Bundesanstalt für Geowissenschaften und Rohstoffe (2014)

A.2 Zuordung der Bodentypen zu den Bodenklassen

Legende	Bodenklasse
Böden der Küstenregion und Moore	
Podsol-Regosol und Regosol, aus trockenen, nährstoffarmen Sanden	2
Wattboden im Gezeitenbereich der Nordsee	2
Marschboden, vorwiegend Kleimarsch aus brackischen oder fluviatilen Ablagerungen sowie Kalkmarsch aus marinen Ablagerungen	2
Niedermoorboden aus mächtigen Niedermoortorfen, teils mit mineralischen Zwischenschichten, Moorgley und Gley	2
Hochmoorboden aus mächtigen Hochmoortorfen über Niedermoortorf, Mudde oder Mineralboden	2
Böden der breiten Flusstäler einschließlich Terrassenflächen und Niederungen	
Auenboden, in tieferen Lagen Gley aus lehmigen bis tonigen Auensedimenten, in Schwarzerdegebieten Gley-Tschernosem aus kalkhaltigen, tonig-schluffigen Ablagerungen	3
Auenboden, in tieferen Lagen Gley aus sandigen bis tonigen Flusssedimenten, häufig in klein-flächigem Wechsel	3
Parabraunerde aus schluffig-lehmigen Deckschichten und Auenpararendzina aus kalkhaltigen, sandig-lehmigen Hochflut- und Auenablagerungen. Podsol-Baunerde und Braunerde der sandigen Terrassen	3
Podsol, Braunerde-Podsol und Gley-Podsol aus sandigen Flussablagerungen sowie Gley der sandigen Urstromtäler und Niederungen	3
Böden des wellig-hügeligen Flachlandes und der Hügelländer	
Braunerde, Parabraunerde und Pararendzina aus lössvermischten Tertiärablagerungen	4
Parabraunerde, Fahlerde und Pseudogley aus Geschiebelehm über Geschiebemergel	4
Pseudogley-Gley und Pseudogley aus lehmig-sandigem Geschiebemergel	4
Podsol-Parabraunerde und Podsol-Fahlerde aus sandigen Deckschichten über Geschiebelehm	4
Fahlerde und Bänderparabraunerde sowie Braunerde und Podsol-Braunerde aus sandigen Deckschichten über Geschiebelehm	4
Bänderparabraunerde, Fahlerde und Braunerde sowie Pararendzina und Regosol im engräumigen Wechsel der sandigen bis lehmigen Endmoränen	4
Pseudogley-Braunerde und Pseudogley-Fahlerde sowie Pseudogley aus Geschiebedecksand über Geschiebelehm	4
Parabraunerde, Braunerde und Pararendzina aus lehmig-sandigen, kalkhaltigen Moränenablagerungen im Alpenvorland	4
Braunerde, Parabraunerde und Pseudogley aus kalkhaltigen, lehmig-sandig-kiesigen, lössvermischten Moränenablagerungen im Alpenvorland	4
Podsol-Braunerde, Podsol-Bänderparabraunerde und Pseudogley-Podsol aus trockenen, nährstoffarmen Sanden	4
Braunerde, Bänderparabraunerde und Podsol-Braunerde aus nährstoffreichen Sanden	4

Böden der Berg- und Hügelländer sowie der Mittelgebirge aus Festgestein, dessen Verwitterungsmaterial und Umlagerungsdecken	
Tschernosem bzw. Pararendzina aus Löss im Wechsel mit Rendzina aus Mergel- und Kalkstein	5
Tschernosem der Mitteldeutschen Trockengebiete aus mächtigem Löss, Tschernosem und Pseudogley-Tschernosem aus Löss über Ton- und Mergelgesteinen	5
Tschernosem-Parabraunerde und Parabraunerde-Tschernosem aus Löss, Lösslehm oder aus lössähnlichen Schluffablagerungen	5
Parabraunerde, Fahlerde, Pseudogley-Parabraunerde und Pseudogley aus Löss oder Lösslehm und lössvermischten Verwitterungsprodukten über verschiedenen Gesteinen	5
Parabraunerde, Fahlerde und Braunerde aus Sandlöss über Sand oder Lehm sowie aus sandvermischtem Löss oder Lösslehm	5
Lösslehm über verschiedenen Gesteinen	5
Böden der Berg- und Hügelländer sowie der Mittelgebirge	
Dolomitgesteinen im Wechsel mit schluffig-tonigen Umlagerungsprodukten der Kalkstein-verwitterung	6
Dolomitsteinverwitterung, Parabraunerde mit dünner Lössdecke sowie Rendzina aus Kalkstein	6
Tongesteinen sowie Braunerde aus Mergelgesteinen und kalkhaltigen Schottern	6
Braunerde aus basenreichen Tuffen	6
morpher Gesteine	6
Schiefern, Sandstein, Quarzit und sauren bis intermediären magmatischen Gesteinen	6
Podsolige Braunerde und Podsol-Braunerde aus Verwitterungsmaterial saurer magmatischer und metamorpher Gesteine	6
Braunerde, Podsol-Braunerde aus Verwitterungsmaterial von Schluff-, Sand- und Tonsteinen	6
Anteilen von Grauwacke, Sandstein, Quarzit und Phyllit	6
Schluffschiefer, Grauwacke und Phyllit	6
Podsolige Braunerde aus Verwitterungsmaterial basenarmer quarzitischer Sandsteine und Konglo	6
Sandstein und Quarzit	6
Podsol-Braunerde aus Sand- und Schluffstein sowie Pelosol-Braunerde aus Mergel- und Tonsteinen	6
Böden des Hochgebirges	
subalpinen Höhenstufen der Alpen sowie Ranker aus kalkfreien Silikatgesteinen und Rohböden der alpinen Fels- und	7
Antropogen veränderte Böden, Siedlungsgebiete und Gewässerfläche	
Versiegelte Flächen in größeren Städten	6
Technogen gestaltete Böden und große Abbauflächen	5
Binnengewässer	4

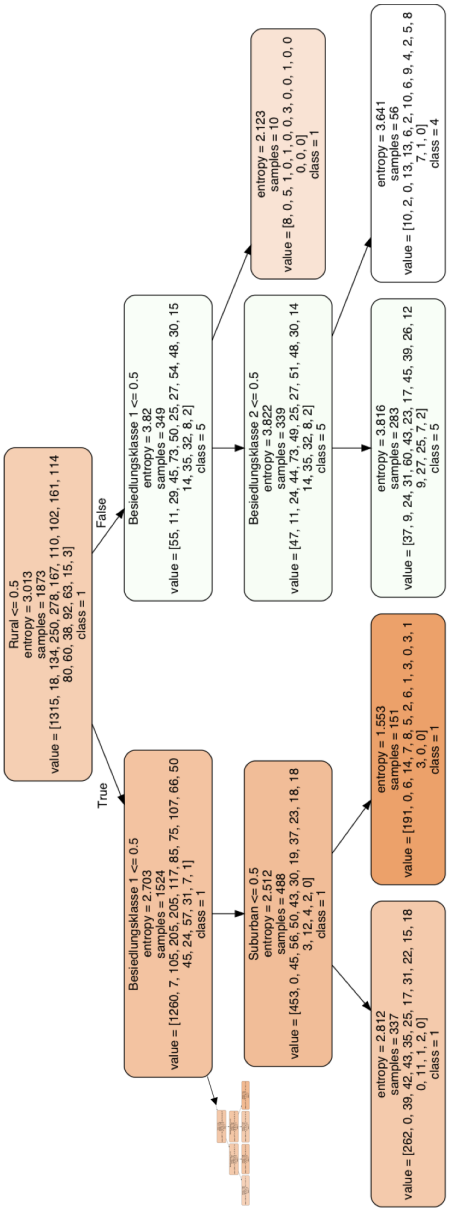
Abbildung A.2: Zuordnung der Bodentypen zu den Bodenklassen.

A.3 Datensatz

	Lieferant verschlüsselt	Monatliche Kosten	Einmalige Kosten	Länge	Preise pro Meter aus monatl. Kosten	Vertragsdauer in Monaten	Kostenpflichtig ab jahr	Besiedlungs Klasse	Area Class	Gewichtungsfaktor der Bodenklasse	Wettbewerb
0	1	1274.200	0.000000	1691.702693	0.753206	210	2001	2	suburban	1.350	6.0
1	2	553.288	460106.901900	4106.952398	0.134720	239	2003	2	urban	1.350	7.0
2	3	607.591	271356.070500	12536.175890	0.048467	240	2003	2	rural	1.175	4.0
3	4	3105.000	5350.080254	438.310392	7.084021	59	2003	3	urban	1.350	2.0
4	4	3105.000	5350.080254	1504.437769	2.063894	59	2003	3	urban	1.350	2.0

Abbildung A.3: Ausschnitt des Datensatzes.

A.4 Entscheidungsbaum



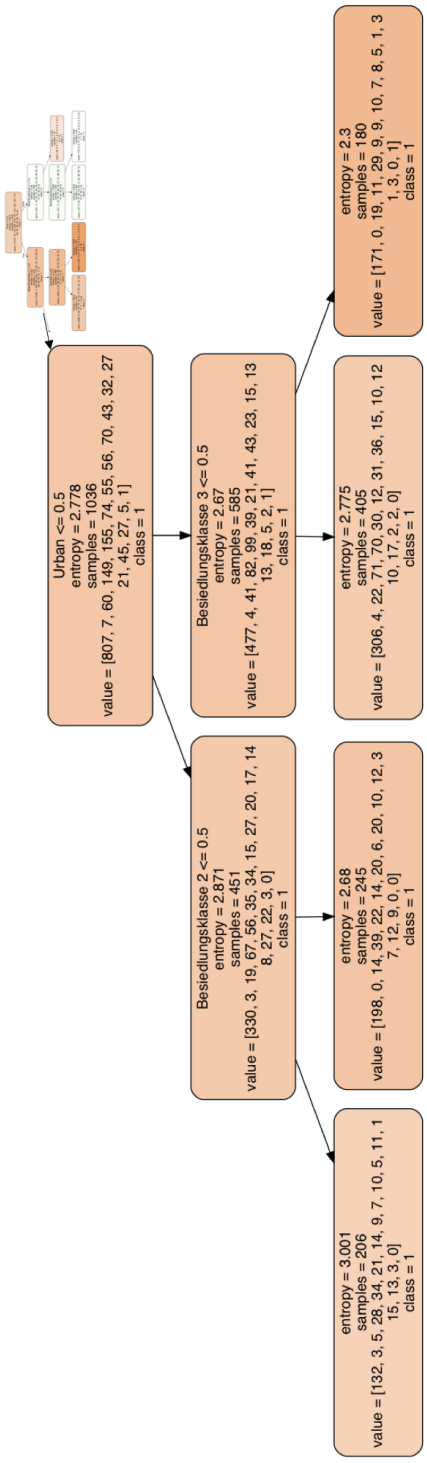


Abbildung A.4: Entscheidungsbaum aus dem Random Forest zur Vorhersage eines einmaligen Preises.

Erklärung zur selbstständigen Bearbeitung einer Abschlussarbeit

Gemäß der Allgemeinen Prüfungs- und Studienordnung ist zusammen mit der Abschlussarbeit eine schriftliche Erklärung abzugeben, in der der Studierende bestätigt, dass die Abschlussarbeit „– bei einer Gruppenarbeit die entsprechend gekennzeichneten Teile der Arbeit [(§ 18 Abs. 1 APSO-TI-BM bzw. § 21 Abs. 1 APSO-INGI)] – ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt wurden. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich zu machen."

Quelle: § 16 Abs. 5 APSO-TI-BM bzw. § 15 Abs. 6 APSO-INGI

Erklärung zur selbstständigen Bearbeitung der Arbeit

Hiermit versichere ich,

Name: _____

Vorname: _____

dass ich die vorliegende Bachelorarbeit – bzw. bei einer Gruppenarbeit die entsprechend gekennzeichneten Teile der Arbeit – mit dem Thema:

Maschinelles Lernen zur Analyse von Einflussfaktoren auf den Preis und Preisentwicklungen im deutschen Glasfasermarkt

ohne fremde Hilfe selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

_____	_____	_____
Ort	Datum	Unterschrift im Original