



KURSANALYSE VON KRYPTOWÄHRUNGEN MIT AZURE MACHINE LEARNING

PRICE ANALYSIS OF CRYPTOCURRENCIES USING AZURE MACHINE
LEARNING

ABSCHLUSSARBEIT

ZUR ERLANGUNG DES AKADEMISCHEN GRADES
MASTER OF SCIENCE

VORGELEGT VON

SEBASTIAN LISCHEWSKI

GEBOREN AM 08.08.1991 IN ROSENHEIM
MATRIKELNUMMER: 04326912

MÜNCHEN, DEN 27. DEZEMBER 2017

Prüfer: Prof. Dr. PATRICK MÖBERT, Hochschule München

Erklärung

Hiermit erkläre ich, dass ich die Bachelorarbeit selbständig verfasst, noch nicht anderweitig für Prüfungszwecke vorgelegt, keine anderen als die angegebenen Quellen oder Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

Ort, Datum

Unterschrift

Zusammenfassung

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Inhaltsverzeichnis

Abbildungsverzeichnis	VI
Tabellenverzeichnis	VII
Listings	IX
1 Einführung zum Thema	1
1.1 Thema der Arbeit	1
1.2 Bitcoin als Vorreiter der Kryptowährungen	1
1.3 Machine Learning, Data Mining, Data Analysis und Data Science	2
1.4 Cloud-Dienste und SaaS	4
2 Vorgehen und Ziele	7
3 Grundlagen	8
3.1 Data Mining Frameworks	8
3.1.1 Knowledge Discovery in Databases (KDD) process model	8
3.1.2 Cross Industrial Standard Process for Data Mining (CRISP – DM) .	16
3.1.3 Sample, Explore, Modify, Model and Assess (SEMMA)	31
3.1.4 Auswahl	31
3.2 Machine Learning	31
3.2.1 Supervised Learning	34
3.2.2 Unsupervised Learning	36
3.2.3 Semi-supervised Learning	38
3.2.4 Active Learning	39
3.2.5 Reinforcement Learning	39
3.3 Kryptowährung(en)	40

Inhaltsverzeichnis

3.4	Microsoft Azure ML Studio	41
3.4.1	Allgemeine Beschreibung	42
3.4.2	Aufbau	42
4	Durchführung der Analyse	45
4.1	Business Understanding	46
4.1.1	Determine the Business Objectives	46
4.1.2	Assess the Situation	47
4.1.3	Determine the Data Mining Goals	54
4.1.4	Produce a Project Plan	56
4.2	Data Understanding	59
4.2.1	Collect the Initial Data	59
4.2.2	Describe the Data	65
4.2.3	Explore the Data	78
4.2.4	Verify Data Quality	81
4.3	Data Preparation	83
4.3.1	Select Data	83
4.3.2	Clean, Construct, Integrate, Format Data	87
4.4	Modeling	97
4.4.1	Select the Modeling Technique	97
4.4.2	Generate Test Design	99
4.4.3	Build the Model	101
4.4.4	Assess the Model	101
4.5	Evaluation	101
4.5.1	Evaluate Results	101
4.5.2	Review Process	101
4.5.3	Determine Next Steps	101
5	Interpretation Fazit	102
6	Related Work	103
7	Ausblick	104
	Literaturverzeichnis	105

Abbildungsverzeichnis

1.1	Learn from data evolution (Swamynathan, 2017, S. 66)	3
3.1	Ein Überblick über die Schritte des KDD Prozesses nach (Fayyad et al., 1996, S. 41)	9
3.2	Phasen des CRISP-DM Referenzmodells nach (Chapman et al., 2000, S. 10)	17
3.3	Generische Aufgaben (bold) und Output (<i>italic</i>) des CRISP-DM Referenzmodells (Chapman et al., 2000, S. 12)	18
3.4	Betrachten der Datentypen des Datensatzes in RStudio	22
3.5	Durchschnittlicher prozentualer Anteil der CRISP-DM-Projektphase am Gesamtprojekt nach (Shearer, 2000, S. 15; eigene Darstellung)	24
3.6	Machine Learning Types nach (Ramasubramanian and Singh, 2017, S. 222) .	32
3.7	Beispiel für eine Klassifikation aus (Suthaharan, 2016, S. 8)	35
3.8	Beispiel für zwei Cluster (Suthaharan, 2016, S. 9)	38
3.9	Iterativer Reinforcement Learning Prozess aus (Lison, 2012, S. 25)	39
4.1	Google Websuchen und Newssuchen für „Bitcoin“ im zeitlichen Verlauf . . .	79
4.2	Google Websuchen und Newssuchen für „Bitcoin“ im zeitlichen Verlauf . . .	80
4.3	Google Websuchen und Newssuchen für „Ethereum“ im zeitlichen Verlauf . .	81
4.4	Wikipedia Seitenaufrufe „Bitcoin“ und der Bitcoinkurs im zeitlichen Verlauf	84
4.5	Wikipedia Seitenaufrufe „Ethereum“ und der Etherumkurs im zeitlichen Verlauf	85
4.6	Zusammenführen der Aktienindices in Azure Machine Learning Studio . . .	89
4.7	Nachbearbeiten des Ergebnisdatensatzes in Azure Machine Learning Studio .	90
4.8	Feature Hashing der abcnews in Azure Machine Learning Studio	91
4.9	Erzeugen des finalen Datensatzes für das Machine Learning	95
4.10	Auswahl des Analysezeitraums und bereinigen fehlender Daten in Azure Machine Learning Studio	96
4.11	Verhältnis der Aufteilen in Trainings-, Test- und Validationsdatensatz	100

Tabellenverzeichnis

1.1	Cloud-Diensttypen	5
3.1	Einfacher Datensatz mit Berufserfahrung und Gehalt	11
3.2	Output der Regression mit allen Variablen	12
3.3	Output der Regression ohne Alters-Variable	13
3.4	Output der Regression mit zusammengefassten Werten	13
3.5	Aufruf des head()-Befehls zum Betrachten der Daten	21
3.6	Die zwei Hauptaufgaben des Schrittes Integrate Data	26
3.7	Subjective Grouping nach (Ramasubramanian and Singh, 2017, S. 224-229) .	33
3.8	Hauptkomponenten des Azure Machine Learning Studios	44
4.1	Behandelte theoretische Abschnitte im Kontext der Arbeit	45
4.2	Output des Schrittes „Determine the Business Objectives“	46
4.3	Mögliche Einflussfaktoren auf den Kurs von Kryptowährungen	51
4.4	Output des Schrittes „Assess the Situation“	53
4.5	Output des Schrittes „Determine the Data Mining Goals“	56
4.6	Output „Project Plan“ des Schrittes „Produce a Project Plan“	58
4.7	Output „Initial assessment of tools and techniques“ des Schrittes „Produce a Project Plan“	58
4.8	Output „Initial data collection report“ des Schrittes „Collect the Initial Data“	64
4.9	Data description report für BTC _Total _Volume _Daily _Full	65
4.10	Data description report für BTC _Difficulty _Daily _Full	66
4.11	Data description report für BTC _Transaction _Number _Fully _Daily . .	66
4.12	Data description report für BTC _Price _Multiple _Daily	67
4.13	Data description report für ETH _Total _Volume _Daily _Full	67
4.14	Data description report für ETH _Difficulty _Daily _Full	68
4.15	Data description report für ETH _Transaction _Number _Fully _Daily . .	68
4.16	Data description report für google _Trends _BTC _Newssearch und google _Trends _BTC _Websearch	69

Tabellenverzeichnis

4.17	Data description report für google _Trends _ETH _Newssearch und google _Trends _ETH _Websearch	69
4.18	Data description report für Wiki _Page _Views _BTC	69
4.19	Data description report für Wiki _Page _Views _ETH	70
4.20	Data description report für abcnews _Date _Text	70
4.21	Data description report für alle Aktienindices	71
4.22	Data description report für STLFSI _history	72
4.23	Data description report für alle Währungen	72
4.24	Data description report für alle natürlichen Ressourcen	73
4.25	Data description report für ETH _Price _Volume _Full _Daily	74
4.26	Data description report für BTC _Price _Volume _Full _Daily	74
4.27	Data description report für bitcoinDataset	76
4.28	Data description report für ethereumDataset	77
4.29	Output des Schrittes „Explore the Data“	81
4.30	Data quality report des Schrittes „Verify data quality“	82
4.31	Inkludierte und exkludierte Datensätze für die Analyse	86
4.32	Data cleaning report	93
4.33	Output: Generated records	94
4.34	Output: Merged data	94

Listings

3.1	Regression mit allen Faktoren	12
3.2	Regression ohne Alter	12
3.3	Regression mit zusammengefassten Werten	13
3.4	Einlesen aller Daten und Betrachten des „Kopfes“	21
4.1	Google Websuchen und Newssuchen für „Bitcoin“ im zeitlichen Verlauf in R	78
4.2	Google Websuchen und Newssuchen für „Bitcoin“ und Bitcoinkurs im zeitlichen Verlauf in R	79
4.3	82
4.4	Aufbereiten der Indices-Datensätze	87
4.5	Entfernen der alten Überschriften	91
4.6	Aggregieren der Hashing-Ergebnisse	91
4.7	Konvertierung von pseudo-numerischen Werten	93
4.8	Hinzufügen und Popularisieren der Spalte „increase“ in Azure Machine Learning Studio	96

1 Hinführung zum Thema

1.1 Thema der Arbeit

In der vorliegenden Arbeit werden Einflussfaktoren auf den Kurs von ausgewählten Kryptowährungen gesucht und der Grad des Einflusses evaluiert. Dies geschieht mit dem Ziel herauszufinden, ob sich die Kursschwankungen der digitalen Währungen voraussagen lassen und wenn ja, in welchem Maße. Im nachfolgenden Kapitel wird auf die Motivation hinter der Analyse eingegangen. Das genaue Vorgehen und die Ziele werden in Abschnitt 2 erläutert.

1.2 Bitcoin als Vorreiter der Kryptowährungen

Geld online von einem Teilnehmer direkt zu einem Anderen senden, ohne dabei (Transaktions-)Gebühren für einen zwischengelagerten Finanz-Dienstleister zahlen zu müssen, ist der Gedanke hinter dem „Peer-To-Peer Electronic Cash System“ (Nakamoto, 2008) Bitcoin. Obwohl es Teilnehmern ohne Aufwand möglich ist, dem Netzwerk beizutreten oder es wieder zu verlassen, ist es solange unangreifbar, solange ein Angreifer nicht dauerhaft über mehr Rechenkapazität verfügt, als das komplette restliche Netzwerk. (Nakamoto, 2008) Ob das Bitcoinnetzwerk wirklich absolute Anonymität gewährt, wird stark kritisiert. (Reid and Harrigan, 2013; Androulaki et al., 2013). In der Tat werden beim Nutzen des Netzwerk jedoch keine persönlichen Informationen an ein Kreditinstitut (wie PayPal, Paydirekt, ApplePay oder Masterpass) weitergegeben. Diese Argumente (Kostenreduktion, Sicherheit und Anonymität) sorgen für Interesse an der digitalen Währung (auch hier gibt es Kritiker, die den Bitcoin als Investition und nicht als Währung bezeichnen) (Baur et al., 2015). Nicht zu vernachlässigen ist an dieser Stelle auch das Interesse der Industrie an „Smart Contracts“ (Dannen, 2017, S. 10), die beispielsweise im Bereich des Internet of Things Anwendung finden. (Christidis

1 Hinführung zum Thema

and Devetsikiotis, 2016)

Neben Bitcoin hat sich deshalb zusätzlich eine Vielzahl an anderen sogenannten Kryptowährungen entwickelt. Die Währungen mit dem größten Marktvolumen sind Bitcoin(??) und Ethereum(??)(Wood, 2014).(Brandt, 2017; CoinMarketCap, 2017) Daneben gibt es noch sogenannte Altcoins (aus dem Englischen: alternative coin(Bajpai, 2014))(??). Zum Zeitpunkt dieser Arbeit umfassen diese 664 Bitcoin-Alternativen.(CoinDesk, 2017). Obgleich die tatsächliche Nutzung der Kryptowährungen sehr gering ist (1% der Befragten in Deutschland(TSYS, 2016)), steigt das Interesse an Kryptowährungen(WikiTrends, 2017; GoogleTrends, 2017).

TODO: irgendwas zu Technik später oder so?

1.3 Machine Learning, Data Mining, Data Analysis und Data Science

Die Themen Machine Learning, Data Mining, Data Analysis und Data Science sind verwandte Begriffe aus dem interdisziplinären Bereich der Statistik und Informatik.

Der Begriff Machine Learning gehört in der Informatik und Mathematik zur Familie der Künstlichen Intelligenz.(Kim, 2017, S. 2; Swamynathan, 2017, S. 54). Es kann als „Sammlung von Algorithmen und Techniken“ verstanden werden, die „genutzt werden, um Computersysteme zu erstellen, die aus Daten lernen, um Vorhersagen zu erstellen“.(Swamynathan, 2017, S. 53; eigene Übersetzung) Bekannte Anwendungen aus dem Alltag sind Empfehlungssysteme oder Spamerkennungen.(Swamynathan, 2017, S. 53)

Data Mining beschreibt den Prozess, aus einer gewaltigen Menge an Daten die „richtigen Daten“, zur „richtigen Zeit“ für die „richtigen Entscheidungen“(Swamynathan, 2017, S. 61; eigene Übersetzung) zu gewinnen. Um diesen Prozess haben sich im Laufe Zeit drei Frameworks gebildet:(Swamynathan, 2017, S. 69):

- Knowledge Discovery Databases (KDD) process model
- Cross Industrial Standard Process for Data Mining (CRISP – DM)
- Sample, Explore, Modify, Model and Assess (SEMMA)

Neben Schnittmengen mit Künstlicher Intelligenz, Machine Learning und der Statistik, befasst Data Mining sich ebenfalls mit Datenbanksystemen.(Ramasubramanian and Singh, 2017, S. 4)

1 Hinführung zum Thema

Eng verwandt mit dem Data Mining ist die Datenanalyse (engl. Data Analysis; in der Industrie auch Business Analytics(Swamynathan, 2017, S. 58)). Sie wird benutzt um(Hertle, 2016, S. 2; Teil 1)

1. Messdaten zu verstehen,
2. Gesetzmäßigkeiten zu extrahieren und
3. die Zukunft vorherzusagen.

Dazu bedient sie sich der deskriptiven Statistik, der explorativen Datentenenalyse (engl. Explorative Data Analysis; EDA) und der Induktiven Statistik.(Hertle, 2016, S. 17)

Um

- den Anstieg der Datenmengen in der Datenanalyse,
- die Veränderung im Aussehen der Daten (unstrukturiert oder semi-strukturiert statt strukturiert) und
- die Wandlung Semantik der zugrundeliegenden Daten (Daten liegen in Markup-Sprachen vor und enthalten zusätzliche Informationen)

darzustellen, hat sich der Begriff Data Science entwickelt.(Dhar, 2013) Er versucht die geänderten Anforderungen der heutigen Datenanalyse abzubilden (siehe 1.1).

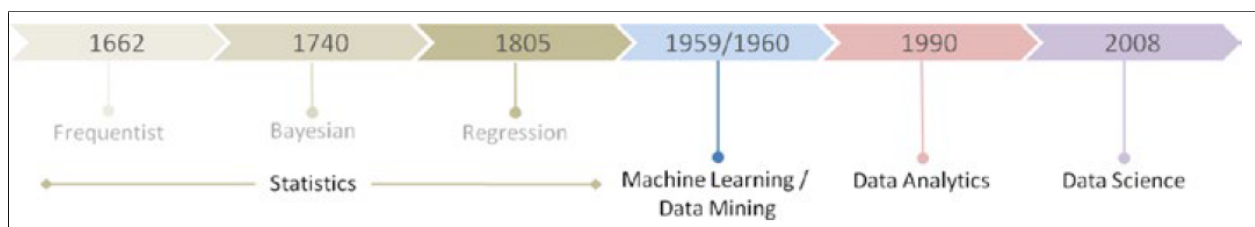


Abbildung 1.1: Learn from data evolution (Swamynathan, 2017, S. 66)

Wie anfänglich erwähnt, sind alle genannten Begriffe miteinander verwandt. Das Gewinnen von Erkenntnissen aus Daten, um beispielsweise die Zukunft vorherzusagen, nennt sich Data Analysis. Werden die Daten aus verschiedensten Datenbanken oder Datawarehouses gewonnen, spricht man von Data Mining. Handelt es sich dabei noch um Informationen unterschiedlicher Struktur und große Datensätze, so befindet man sich im Bereich der Data Science. Der inhärente Erkenntnisgewinn dieser Verfahren kann von von menschlicher Seite kommen oder durch Machine Learning geschehen.

1 Einführung zum Thema

Verdeutlicht wird dies durch Projekte wie Googles DeepMind(?), IBMs Watson((IBM), 2017) oder Sprachassistenten wie Siri, Alexa und Bixby. Sie zeigen, dass großes Interesse an Machine Learning und Data Science herrscht. Deshalb haben sich auch ganze Berufsfelder wie „machine learning engineer“, „data engineer“ oder „data scientist“(Ramasubramanian and Singh, 2017, S. 1) gebildet.

1.4 Cloud-Dienste und SaaS

Cloud Computing beschreibt „ein Modell, das es erlaubt bei Bedarf, jederzeit und überall bequem über ein Netz auf einen geteilten Pool von konfigurierbaren Rechnerressourcen (z. B. Netze, Server, Speichersysteme, Anwendungen und Dienste) zuzugreifen, die schnell und mit minimalem Managementaufwand oder geringer Serviceproviderinteraktion zur Verfügung gestellt werden können“(Appelrath et al., 2014, S. 18). Innerhalb des Cloud Computing unterscheidet man weiterhin zwischen verschiedenen Cloud-Diensten (engl. cloud services). Nach (Appelrath et al., 2014, S. 20) differenziert man zwischen den Services in Abbildung 1.1.

1 Hinführung zum Thema

Diensttyp	Beschreibung
Infrastructure as a Service (IaaS)	Virtuelle Hardware oder Infrastruktur, zum Beispiel Speicherplatz, Rechenleistung oder Netzwerkbandbreite
Platform as a Service (PaaS)	Programmierframeworks, Bibliotheken und Werkzeuge, um Anwendungen unter eigener Kontrolle auf Cloud-Infrastrukturen bereitstellen zu können, ohne die zugrunde liegende Infrastruktur wie Netzwerk, Server, Betriebssysteme oder Speicher managen oder kontrollieren zu müssen
Software as a Service (SaaS)	Vollständige Anwendungen, die auf Cloud-Infrastrukturen betrieben und beispielsweise über einen Webbrowser aufrufbar sind, wobei Nutzer weder die zugrunde liegende Cloud-Infrastruktur noch individuelle Anwendungseinstellungen (mit der möglichen Ausnahme der eingeschränkten Konfiguration von Nutzereinstellungen) kontrollieren müssen und können
Mashup as a Service (MaaS)	Verknüpfung einzelner Software-Komponenten (unter anderem auch Cloud-Dienste) zu einem aggregierten Cloud-Dienst
Business Process as a Service (BPaaS)	Konkrete Geschäftsanwendungen (beispielsweise CRM) als Verknüpfung einzelner Software-Komponenten (standardisierte MaaS)

Tabelle 1.1: Cloud-Diensttypen

(Appelrath et al., 2014, S. 23) sprechen generell von „Cloud Computing als disruptiver Innovationsfaktor“. An dieser Stelle wird besonders Software as a Service betrachtet. Dort stieg der Umsatz von 10,75 Mrd. USD im Jahr 2010 auf 38,57 Mrd. USD im Jahr 2016. Für die Zukunft (2020) wird sogar ein Umsatz von 75,73 Mrd. USD prognostiziert. (Gartner, 2017) Das ist eine Steigerung von über 700% in nur 10 Jahren. Dies kann einerseits durch offensichtliche Vorteile, wie „höhere Stabilität und Planungssicherheit“, der „Möglichkeit Anwender schnell ins System einzuführen“ und „Erschließung neuer Kundengruppen“ (Fraunhofer, 2010) erklärt werden, andererseits aber auch durch Tendenz der Softwarebranche hin zur serviceorientierten Architekturen (engl. service oriented architecture; SOA). (Appelrath et al., 2014, S. 22) Dieser Trend zu SaaS kann beobachtet werden, wenn reine Cloud-Anbieter wie Salesforce

1 Hinführung zum Thema

„klassische“ Anbieter wie SAP den Rang als „Spitze des Weltmarkts der Software für Customer Relationship Management (CRM)“ (Fritsch, 2013) ablaufen.

Laut einer Studie von (Bitkom and KPMG, 2017) greifen 23% der befragten Unternehmen in Deutschland neben „Office Anwendungen aus der Cloud“, „Security as a Service“ und „Groupware“ auf „Business Intelligence/Big Data“-Software aus der Cloud zurück. Zu dieser Kategorie gehört auch Azure Machine Learning (kurz: Azure ML) von Microsoft, welches zur Analyse in dieser Arbeit verwendet wird.

2 Vorgehen und Ziele

Nach der Einführung in das Thema und dem Einordnen in aktuelle Themenfelder, wird nun das Vorgehen und das Ziel der Arbeit erläutert.

Der anschließende Abschnitt xxx befasst sich mit den Grundlagen, die für das Verständnis der Ausarbeitung nötig sind. Dort wird beispielsweise auf die verschiedenen Kategorien des Machine Learning (in xxx) und die zugehörigen Algorithmen und Verfahren eingegangen. Der nachfolgende Teil xxx befasst sich damit, Einflussfaktoren auf die Kurse von Kryptowährungen zu isolieren. Sind die Einflüsse gefunden, wird dargelegt, wie diese als Daten(satz) abgebildet werden können (xxx) und was als Quelle der Daten dient (xxx). In Punkt xxx werden die Datensätze beschrieben. Anschließend (Gliederungspunkt xxx) wird gezeigt, wie die Analyse durchgeführt wird. Dabei wird der Prozess yyy (siehe xxx) **TODO: welcher Prozess? CRISP/KDD...; bereinigung etc..** durchlaufen. In Abschnitt xxx werden die Ergebnisse interpretiert und es werden Schlüsse gezogen. **TODO: bessere Formulierung**

Den Abschluss stellt der Ausblick (xxx) dar. Dieser Teil befasst sich damit, welchen Nutzen die Arbeit bringt (xxx) und wie die Erkenntnisse weiter verwendet werden können (xxx).

TODO: refs

TODO: related work

TODO: später genauer eingehen auf die Sachen Als Ziel steht über der Arbeit, ob es möglich ist, den Kurs oder Kursschwankungen von Kryptowährungen mit Hilfe von Machine Learning vorausszusagen oder nicht. **TODO: braucht man das „oder nicht“?**

TODO: grafische Darstellung anfügen

3 Grundlagen

3.1 Data Mining Frameworks

Wie in Abschnitt 1.3 bereits erwähnt, haben sich um das Data Mining drei bekannte Frameworks entwickelt. Diese werden im Nachfolgenden genauer betrachtet. Anschließend findet die Auswahl statt, welches Rahmenwerk Anwendung in dieser Arbeit findet.

3.1.1 Knowledge Discovery in Databases (KDD) process model

Die Bezeichnung Knowledge Discovery in Databases wurde hauptsächlich von (Fayyad et al., 1996) geprägt. Sie beschreiben in ihrer Arbeit ein Problem der 1990er Jahre. Wie auch heute noch, stieg damals die Masse der gespeicherten Daten exponentiell **TODO: wirklich „exponentiell“?** an. Die Manuelle Auswertung dieser Datensätze erforderte mehr Arbeitskraft als vorhanden war. (Fayyad et al., 1996, S. 38) beschreiben es als „data overload“. Deswegen versuchte man, die Prozesse zur Findung von Erkenntnissen zu automatisieren. Daraus hat sich ein Standardvorgehen entwickelt, dass das KDD-Prozessmodell darstellt.

3 Grundlagen

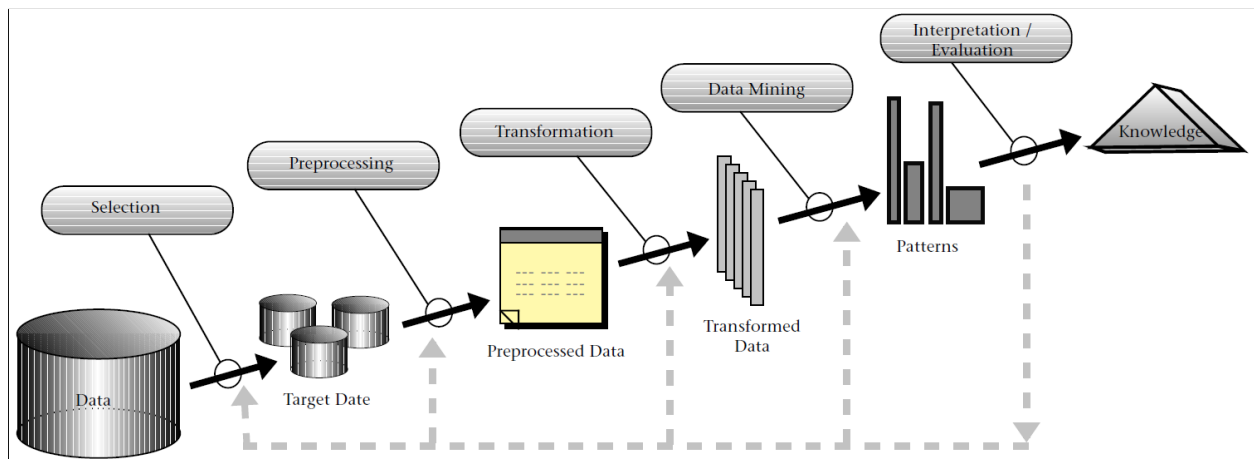


Abbildung 3.1: Ein Überblick über die Schritte des KDD Prozesses nach (Fayyad et al., 1996, S. 41)

Selection

Bevor der erste eigentliche Schritt, die Selektion der Daten, erfolgen kann, ist es unabdingbar, ein „Verständnis für das Anwendungsgebiet zu entwickeln“. (Fayyad et al., 1996, S. 42; eigene Übersetzung) Dies inkludiert auch, Ziele zu setzen und Fragen zu formulieren, die durch das spätere Data Mining (Schritt 3.1.1) beantwortet werden sollen. **TODO: wirklich „Ziele setzten“?** Ist das Verständnis hergestellt, kann ein „target data set“ (Fayyad et al., 1996, S. 42) hergestellt werden. Dabei werden zuerst Daten aus unterschiedlichen - oft heterogenen - Quellen zusammengeführt und dann hinsichtlich des Ziels verdichtet. (Swamynathan, 2017, S. 70)

Preprocessing

Die verbleibende Teilmenge der ursprünglichen Daten muss nun noch gesäubert und für die nächsten Schritte vorbereitet werden. Dies geschieht, da unbereinigte Daten sowohl den Data Mining-Prozess verschlechtern können (unverlässliche oder falsche Ergebnisse), als auch die Zeit für das Mining deutlich verlängern können. (Swamynathan, 2017, S. 70) Um die Qualität der Daten und des Mining zu verbessern, werden unter anderem folgende Aspekte betrachtet: (Fayyad et al., 1996, S. 42; Swamynathan, 2017, S. 70)

3 Grundlagen

Outliner treatment

Ein Ausreißer (engl. outlier) kann beispielsweise ein „Extremer Wert in einer Variablen“ oder ein „Extremer Wert des Residuums bei einer sinnvollen Regression“ (Hertle, 2016, S. 25; Teil 5b) sein. Ein Vorgehen für Ausreißer kann folgendermaßen aussehen (nach (Hertle, 2016, S. 25; Teil 5b)):

1. Identifizieren der Ausreißer (evtl. durch eine erste Regression)
2. Interpretation im Sachzusammenhang (Messfehler oder wichtiger Teil der Population)
3. Entscheidung, ob man eine Regression der Daten mit oder ohne diese Ausreißer haben möchte
4. In der Darstellung der Ergebnisse auf die Ausreißer explizit eingehen und Vorgehen erläutern

Noise removal

Auch in einem Datensatz, der auf Ausreißer untersucht wurde, befinden sich immer noch unbekannte, unvollständige, falsche und fehlende Werte („attribute noise“). Zusätzlich können Datenklassen falsch gekennzeichnet sein („class noise“). Ist ein Datensatz von diesen Problemen betroffen, spricht man von „noisy data“. Auf die Lösung dieses Problems wird an dieser Stelle nicht weiter eingegangen.

Identifying duplicated values

Wie oben angesprochen, wird der zu analysierende Datensatz aus mehreren Quellen zusammengeführt. Durch diesen Schritt können Datensätze doppelt (oder noch öfter) vorkommen. Das wird deutlich, wenn man folgendes Beispiel betrachtet:

Über eine Kundenkarte werden Daten von Kunden eines Supermarkets je Filiale gespeichert. Bei einer überregionalen Kundenanalyse tauchen Kunden mehrfach auf, die in verschiedenen Filialen eingekauft haben. Hier ist anzumerken, dass doppelte Werte nicht zwangsläufig gelöscht werden müssen, sie sollten jedoch bei der Analyse bedacht werden.

3 Grundlagen

Check for inconsistency

Je größer ein Datensatz ist, umso wahrscheinlicher enthält der auch Inkonsistenzen TODO: quelle hierfür?. Dies wird ebenfalls durch die Fusion von mehreren Quellen verstärkt (Beispiel: unterschiedliches Alter für einen Kundenstammsatz). Auch hier muss geprüft werden, wie mit diesen Werten umzugehen ist. Eventuell können Regeln festgelegt werden wie „immer der neuste Datenpunkt ist der richtige“.

Time series and changes

Der letzte Punkt, der beim Preprocessing betrachtet werden muss, ist der Zusammenhang der Daten und dem Erfassungszeitpunkt. So können sich im Laufe der Zeit die Messmethodik (z.B. andere Sensoren), die Messgenauigkeit (z.B. bessere Sensoren) oder die Abstände der Messungen verändern. TODO: warum ist das schlecht: ungleich verteilte Datensätze/inkonsistente Genauigkeit

Transformation

Der letzte Schritt vor dem eigentlichen Data Mining ist die Transformation. In diesem Prozessschritt geht es darum, „mit Dimensionsreduktions- oder -transformationsmethoden die effektive Anzahl an Variablen [...] zu reduzieren“ (Fayyad et al., 1996, S. 42; eigene Übersetzung). Dies geschieht beispielsweise durch das identifizieren und eliminieren invarianter Variablen. Ebenfalls wird versucht, solche Variablen zu finden, die mehrere Andere repräsentieren. Anschaulich dargestellt an einem Beispiel:

	Person	Studium	ErfahrungExtern	ErfahrungIntern	Alter	Gehalt
1	1	6	1	4	24	46450
2	2	18	30	15	55	85150
3	3	11	7	7	31	55900
4	4	11	15	8	36	63650
5	5	10	1	16	33	59050
6	6	6	25	6	38	68750
7	7	10	20	20	50	79000
8	8	7	0	1	23	43050

Tabelle 3.1: Einfacher Datensatz mit Berufserfahrung und Gehalt

3 Grundlagen

Tabelle 3.1 zeigt einen einfachen Datensatz, in dem die Mitarbeiter einer Firma und die zugehörigen Gehälter festgehalten sind. „Studium“ beschreibt die Anzahl der Halbjahre im Studium. Analog dazu „ErfahrungExtern“ und „ErfahrungIntern“ die Berufserfahrung in Halbjahren außerhalb und innerhalb der Firma. Zusätzlich ist das Alter der Personen gegeben. Führt man eine Regression (Listing 3.1) für den Datensatz durch (mit Studium, ErfahrungExtern, ErfahrungIntern, Alter als unabhängige und Gehalt als abhängige Variablen), ergibt sich das Ergebnis in Tabelle 3.2.

```

1 | #Daten einlesen
2 | data <- read.csv2("Beispiel_Berufserfahrung_Datensatz1.csv")
3 |
4 | #Regression mit allen Faktoren
5 | regression1 <- lm(Gehalt ~ Studium + ErfahrungExtern + ErfahrungIntern +
6 |                   Alter, data=data)
7 | summary(regression1)

```

Listing 3.1: Regression mit allen Faktoren

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	40000	1.415e-11	2.828e+15	<2e-16 ***
Studium	300	5.167e-13	5.807e+14	<2e-16 ***
ErfahrungExtern	850	4.821e-13	1.763e+15	<2e-16 ***
ErfahrungIntern	950	6.308e-13	1.506e+15	<2e-16 ***
Alter	2.010e-13	8.103e-13	2.480e-01	0.82

Tabelle 3.2: Output der Regression mit allen Variablen

Ohne weiter auf die genauen Bezeichnungen einzugehen, gibt die Sternnotation von R an, dass die unabhängigen Variablen Studium, ErfahrungIntern und ErfahrungExtern signifikant sind. Das Alter hingegen nicht. Die Regression hat ein adjustiertes Bestimmtheitsmaß (R^2 ; engl. adjusted R-squared) von 1. Das bedeutet, dass das Gehalt vollständig durch die gegebenen Variablen erklärt werden kann (dies wird in der Realität jedoch nie erreicht).

```

1 | #Regression mit signifikanten Faktoren
2 | regression2 <- lm(Gehalt ~ Studium + ErfahrungExtern + ErfahrungIntern,
3 |                   data=data)
4 | summary(regression2)

```

Listing 3.2: Regression ohne Alter

3 Grundlagen

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	40000	2.393e-12	1.671e+16	<2e-16 ***
Studium	300	2.948e-13	1.018e+15	<2e-16 ***
ErfahrungExtern	850	8.948e-14	9.500e+15	<2e-16 ***
ErfahrungIntern	950	1.642e-13	5.787e+15.75	<2e-16 ***

Tabelle 3.3: Output der Regression ohne Alters-Variable

Führt man die Regression nun ohne das Alter durch (Listing 3.2 und Tabelle 3.3) bleibt R^2 gleich. Der Datensatz wurde also bereits um eine Variable reduziert, ohne das Ergebnis der Regression zu verschlechtern.

Betrachtet man die Faktoren ErfahrungExtern und ErfahrungIntern, so fällt auf, dass sie einen ähnlichen Einfluss auf das Gehalte erzielen (850 und 950).

```

1 #Transformation
2 data[,"ErfahrungGesamt"] <- data[,3] + data[,4]
3
4 #Regression
5 regression3 <- lm(Gehalt ~ Studium + ErfahrungGesamt, data=data)
6 summary(regression3)

```

Listing 3.3: Regression mit zusammengefassten Werten

Fasst man beide Variablen zusammen (Listing 3.3), zeigt sich im Ergebnis (Tabelle 3.4), dass R^2 bei 0,9988 liegt. Die Güte der Regression hat sich also nur minimal verschlechtert.

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	40107.10	528.87	75.83	7.55e-09 ***
ErfahrungGesamt	876.69	15.86	55.26	3.67e-08 ***
Studium	327.17	64.27	5.09	0.0038 **

Tabelle 3.4: Output der Regression mit zusammengefassten Werten

Zusammenfassend lässt sich für dieses Beispiel sagen, dass die Variablen im Datensatz um die Hälfte reduziert wurden, ohne die Aussagekraft deutlich zu verschlechtern. In einem realen Datensatz ist diese Arbeit zwar nicht so trivial und offensichtlich, jedoch gelten die gleichen Prinzipien.

Nach (Swamynathan, 2017, S. 71; veränderte Version) gibt es zur Transformation folgende Möglichkeiten:

3 Grundlagen

- Smoothing (binning, clustering, regression, etc.)
- Aggregation (im Beispiel: das Zusammenfassen der Berufserfahrung)
- Generalization (Ersetzen von primitiven Datenobjekten durch höherstufige Konzepte)
- Normalization (min-max-scaling oder z-score)
- Feature construction aus bereits bestehenden Attributen durch Techniken wie die Hauptkomponentenanalyse (engl. principal components analysis; PCA), Multidimensional scaling (MDS) oder Locally-linear embedding (LLE)
- Compression (zum Beispiel wavelets, PCA, clustering etc.)
- andere Datenreduzierungstechniken bei denen das Datenvolumen sinkt, ohne die Integrität der Originaldaten zu verletzen

Data Mining

Ist der Datensatz präpariert, so findet das eigentliche Data Mining statt. Dabei muss sich der Anwender für eine oder auch mehrere Methoden für das Mining entscheiden, um die anfänglichen Ziele zu erreichen und die Fragestellungen zu beantworten. Zur Auswahl stehen beispielsweise (Fayyad et al., 1996, S. 42; Swamynathan, 2017, S. 71):

- zusammenfassende und beschreibende Methoden: Mittelwert (arithmetisches Mittel), Median, Modus, Standardabweichung, Klassen- und Konzeptbeschreibungen, grafische Plots,
- Vorhersagende Modelle (engl. predictive models): Klassifikationen und Regressionen und
- Cluster-Analysen.

Eine genauere Beschreibung der Methoden (und der zugehörigen Algorithmen) im Kontext des Machine Learning befindet sich in Abschnitt 3.2. Je nach Beschaffenheit der zugrundeliegenden Daten und der gewählten Methode, muss ein passender Algorithmus gewählt und dieser korrekt parametrisiert werden. Zum Data Mining gehört auch, Hypothesen zu formulieren und das Ergebnis im Auge zu behalten: Ist der Endnutzer der Analyse an einem vorhersagenden Model interessiert (zum Beispiel für Wartungsarbeiten) oder an einem Jetzt-bezogenen (zum

3 Grundlagen

Beispiel für eine strategische Ausrichtung nach den aktuellen Kundensegmenten)?

Anschließend erfolgt das (automatische) Mining der Daten. Je besser die vorhergehenden Schritte durchgeführt wurden, desto potenter ist das Ergebnis.(Fayyad et al., 1996, S. 42) Aus diesem Grund ist es auch jederzeit möglich, zu einem vorangegangenen Prozessschritt zu springen, um neu erlangte Einsichten einfließen zu lassen (siehe zurückspringender grauer Pfeil in Abbildung 3.1).

Interpretation/Evaluation

Zuletzt werden die gefundenen Muster und trainierten Modelle interpretiert. Ein Muster macht Aussagen über jeden Datenpunkt im betrachteten Raum. Ein Beispiel bei einem einfachen linearen Modell:

$$y = m \times x + t$$

Zu obigem Fall:

$$Gehalt = Studium \times 327,17 + Erfahrung_{Gesamt} \times 876,69 + 40107,10$$

Ein Muster (engl. pattern) beschreibt dagegen nur eine kleine „lokale Struktur“, die „nur über einen begrenzten Bereich“ Aussagen macht.(Swamynathan, 2017, S. 71; eigene Übersetzung) Im Fall des linearen Modells wäre es eine bestimmte Gleichung, zum Beispiel

$$y = 2 \times x + 5$$

oder

$$6 \times 327,17 + 5 \times 876,69 + 40107,10 = 46453,57$$

(Kraker and Dennerlein, 2013). „Fayyad et al. benutzt patterns und models synonym“. (Kraker and Dennerlein, 2013)

Das Interpretieren der Ergebnisse beinhaltet ebenfalls das Zusammenfassen der Erkenntnisse und gegebenenfalls das Visualisieren.(Swamynathan, 2017, S. 71) Als Evaluieren wird das Eingliedern der Resultate in andere Systeme (zur Weiterverarbeitung oder Verbreitung), das Prüfen auf (und Lösen von) Konflikten mit anderen Untersuchungen und nicht zuletzt das Dokumentieren der Befunde bezeichnet.(Fayyad et al., 1996, S. 42)

TODO: Befund nur medizinisch?; synonym) An dieser Stelle sei erneut angemerkt, dass das erste Ergebnis des KDD-Prozesses nicht das Endergebnis sein muss. Es kann durchaus

3 Grundlagen

viele Iterationen geben, die auch „loops between any two steps“ beinhalten können.(Fayyad et al., 1996, S. 42)

3.1.2 Cross Industrial Standard Process for Data Mining (CRISP – DM)

Bei Cross Industrial Standard Process for Data Mining handelt es sich - wie bei KDD - um ein Referenzmodell für Data Mining. Das Modell wurde von einem 1996 gegründeten Konsortium aus „Daimler-Benz (now DaimlerChrysler), Integral Solutions Ltd. (ISL) [jetzt SPSS], NCR, and OHRA“(Shearer, 2000, S. 13) erarbeitet. Die Version 1.0 wurde 2000 vorgestellt.(Shearer, 2000, S. 13) In Umfragen (1999, 2002, 2004, 2007) wird das Modell als führend in Bereich von „data mining/predictive analytics projects“(Swamynathan, 2017, S. 72) bezeichnet. Das Modell ist „nicht-properitär, dokumentiert und frei verfügbar“(Shearer, 2000, S. 13; eigene Übersetzung). Es ist ebenfalls in vielen Bereichen nutzbar, da es weder Industriesektor-, Werkzeugs- noch Anwendungsspezifisch ist. Grundsätzlich bekräftigt das Modell best practices und soll zu besseren und schnelleren Ergebnissen führen.(Shearer, 2000, S. 13; eigene Übersetzung)

TODO: evtl. bessere Bezeichnung

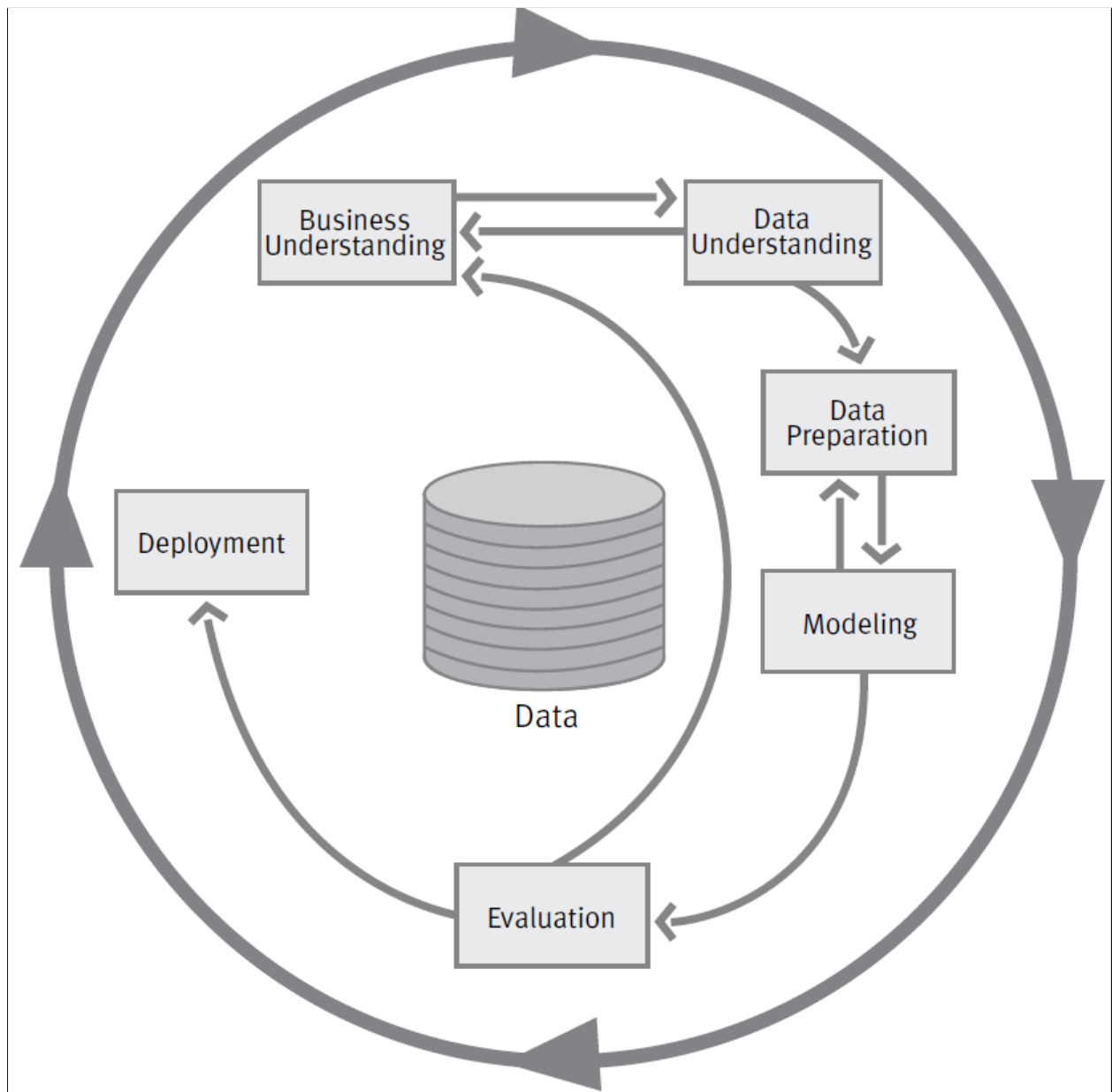


Abbildung 3.2: Phasen des CRISP-DM Referenzmodells nach (Chapman et al., 2000, S. 10)

Wie in Abbildung 3.2 zu sehen ist, umfasst das Referenzmodell sechs Phasen. Genau wie beim KDD-Prozessmodell handelt es sich nicht um ein lineares Modell, sondern um eines, das Rückschritte und Iterationen erlaubt. Im Nachfolgenden wird zuerst immer ein Prozessschritt kurz vorgestellt und darunter detaillierter betrachtet. Als Referenz dient unter anderem Abbildung 3.3, die zusätzlich den Output der einzelnen Schritte zeigt.

3 Grundlagen

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives <i>Background Business Objectives Business Success Criteria</i> Assess Situation <i>Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits</i> Determine Data Mining Goals <i>Data Mining Goals Data Mining Success Criteria</i> Produce Project Plan <i>Project Plan Initial Assessment of Tools and Techniques</i>	Collect Initial Data <i>Initial Data Collection Report</i> Describe Data <i>Data Description Report</i> Explore Data <i>Data Exploration Report</i> Verify Data Quality <i>Data Quality Report</i>	Select Data <i>Rationale for Inclusion/Exclusion</i> Clean Data <i>Data Cleaning Report</i> Construct Data <i>Derived Attributes Generated Records</i> Integrate Data <i>Merged Data</i> Format Data <i>Reformatted Data Dataset Dataset Description</i>	Select Modeling Techniques <i>Modeling Technique Modeling Assumptions</i> Generate Test Design <i>Test Design</i> Build Model <i>Parameter Settings Models Model Descriptions</i> Assess Model <i>Model Assessment Revised Parameter Settings</i>	Evaluate Results <i>Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models</i> Review Process <i>Review of Process</i> Determine Next Steps <i>List of Possible Actions Decision</i>	Plan Deployment <i>Deployment Plan</i> Plan Monitoring and Maintenance <i>Monitoring and Maintenance Plan</i> Produce Final Report <i>Final Report Final Presentation</i> Review Project <i>Experience Documentation</i>

Abbildung 3.3: Generische Aufgaben (bold) und Output (italic) des CRISP-DM Referenzmodells (Chapman et al., 2000, S. 12)

Business Understanding

Die erste und vielleicht wichtigste Phase (Shearer, 2000, S. 14) des CRISP-DM Prozesses ist das „Business Understanding“, oder auch „Research Understanding“ (Larose, 2014, Punkt 1.4.1.1). Die Aufgabe dieser Phase ist, die „Ziele und Erwartungen“ (Swamynathan, 2017, S. 73) des Projektes zu verstehen, dieses „Wissen in eine Machine Learning Problem Definition zu übersetzen“ und schließlich einen „Vorläufigen Plan“ (Shearer, 2000, S. 14) aufzustellen:

Determine the Business Objectives

Dieser Teilschritt soll hauptsächlich die Frage beantworten, warum die Analyse durchgeführt wird. Dies hat direkten Einfluss auf die Zeile des Projekts und soll verhindern, dass „viel Aufwand für das Finden von richtigen Antworten auf falsche Fragen“ (Chapman et al., 2000, S. 14) verschwendet wird.

3 Grundlagen

Assess the Situation

Um das Ziel der Analyse so genau wie möglich zu treffen, muss genau nachgeforscht werden, welche Ressourcen verfügbar sind, welchen Zwängen und Grenzen die Analyse unterlegen ist und unter welchen Annahmen sie stattfindet. (Chapman et al., 2000, S. 14) Vereinfacht lässt sich sagen, dass hier die Fragen aus dem vorhergehenden Schritt detaillierter betrachtet werden.

Determine the Data Mining Goals

Die gefundenen Ziele sind meist in Geschäftssprache formuliert. Für die Analyse müssen die Ziele jedoch im Terminus technicus des Data Mining formuliert sein. Ein Beispiel dazu ist die Übersetzung von „Increase catalog sales to existing customers.“ in „Predict how many widgets a customer will buy, given their purchases over the past three years, demographic information (age, salary, city, etc.), and the price of the item.“ (Chapman et al., 2000, S. 16)

Produce a Project Plan

Der Projektplan (engl. Project Plan) liefert einen konkreten Plan, wie die gesetzten Ziele zu erreichen sind. Nach (Shearer, 2000, S. 15) beinhaltet er:

- Die Schritte, die nacheinander durchzuführen sind.
- Eine Timeline **TODO: richtiges Wort?** für die Durchführung.
- Eine Auflistung potentieller Risiken im Projektverlauf.
- Eine Aufstellung der zu nutzenden Werkzeuge und Techniken (Chapman et al., 2000, S. 16)

Data Understanding

In der vorhergegangenen Phase wurde festgelegt, welche Daten zum Erreichen der Ziele benötigt werden. Nun werden diese Daten gesammelt und untersucht. Im Fokus der Untersuchung liegt dabei (Swamynathan, 2017, S. 73)

- Datenlücken finden,

3 Grundlagen

- die Relevanz der erfassten Daten (hinsichtlich der Ziele) zu klären,
- die allgemeine Datenqualität festzustellen und
- „erste Einblicke in Daten“ zu erhalten, um „geeignete Hypothesen“ (Swamynathan, 2017, S. 73; eigene Übersetzung) zu formulieren.

Im Zuge dessen können auch bereits „subsets“ isoliert werden, die „actionable patterns“ (Larose, 2014, Punkt 1.4.1.2.d) (etwa: verfolgbare Muster) enthalten könnten. Mit jedem Fortschritt in dieser Phase ist es eventuell nötig, das Ergebnis der Business Understanding-Phase zu adjustieren. Durch dieses Vorgehen, entsteht ein iterativer Prozess. (Swamynathan, 2017, S. 73) (Larose, 2014, Punkt 1.4.1.2.b) empfiehlt den Einsatz von explorativer Datenanalyse (siehe Abschnitt 1.3).

Collect the Initial Data

Die Hauptaufgabe dieses ersten Schrittes ist, die benötigten Daten zu beziehen. Dabei kann auch entweder der direkte Zugriff auf die Daten gemeint sein oder nur das Erhalten der Zugangsinformationen. Eventuell werden die Datensätze gleich in Systeme oder Werkzeuge geladen, die zur späteren Weiterverarbeitung genutzt werden. Die Integration inhomogener Daten (unterschiedliche Strukturen/Formate etc.) kann bereits hier erfolgen oder in der Data Preparation (siehe Punkt 3.1.2). (Chapman et al., 2000, S. 18)

Sollten in diesem Schritt Probleme auftauchen, sollten sie - wenn möglich mit Lösung - gut dokumentiert werden, um den Projektverlauf reproduzierbar zu gestalten. Ein Beispiel können lange Antwortzeiten einiger Datenquellen sein. (Shearer, 2000, S. 15)

Describe the Data

Im Schritt „Describe the Data“ werden die beschafften Daten dann oberflächlich beschrieben. (Chapman et al., 2000, S. 18) Dabei wird unter anderem auf

- das Format der Daten,
- die Größe des Datensatzes,
- die Anzahl der Beobachtungen und Einträge in den Daten und
- die Beschaffenheit der Einträge

3 Grundlagen

geachtet. Dabei soll einerseits die Frage geklärt werden, ob die vorhandenen Daten alle relevanten Daten für die Ziele des Data Mining enthalten. Andererseits wird das Verständnis der Daten geschärft. (Shearer, 2000, S. 15)

Anhand des nachfolgenden Beispiels (Datensatz von (Hertle, 2016, Case Lasagne Test.xlsx)) werden die Schritte „Collect the Initial Data“ und „Describe the Data“ kurz visualisiert. Zuerst werden die Daten eingelesen (Listing 3.4) und anschließend oberflächlich betrachtet.

```
1 data <- read.csv2("Case_Lasagne_Test.csv")
2 head(data)
```

Listing 3.4: Einlesen aller Daten und Betrachten des „Kopfes“

Aus Tabelle 3.5 kann abgelesen werden, dass es sich um einen Datensatz mit 12 Variablen handelt.

Person	Alter	Gewicht	Einkommen	Angestellt	Wert.Auto	Umsatz.Kreditkarte	Geschlecht
1	48	65	91700	nein	2190	3510	m
2	33	75	40740	nein	2110	740	w
3	51	70	45080	ja	5140	910	m
4	56	91	26600	nein	700	1620	w
5	28	81	113960	ja	26620	600	m
6	51	65	102200	ja	24520	950	w

alleinstehend	Wohnung	Supermarkt.besuche.pro.Monat	Lasagne.probiert
nein	Haus	7	nein
nein	Wohnung	4	ja
nein	Wohnung	1	nein
nein	Haus	3	nein
nein	Appartement	3	ja
nein	Wohnung	2	nein

Tabelle 3.5: Aufruf des head()-Befehls zum Betrachten der Daten

Ebenfalls entnommen werden kann, dass es sich um demographische Angaben über Personen handelt. Zusätzlich wurde zu jeder Person erfasst ob sie Lasagne probiert hat. In Abbildung 3.4 wird der Datensatz schließlich in RStudio betrachtet.

3 Grundlagen

Data	
data	856 obs. of 13 variables
i..Person	: int 1 2 3 4 5 6 7 8 9 10 ...
Alter	: int 48 33 51 56 28 51 44 29 28 29 ...
Gewicht	: int 65 75 70 91 81 65 68 70 75 78 ...
Einkommen	: int 91700 40740 45080 26600 113960 102200 92960 64680 85540 13720 ...
Angestellt	: Factor w/ 2 levels "ja","nein": 2 2 1 2 1 1 1 1 1 1 ...
wert.Auto	: int 2190 2110 5140 700 26620 24520 10130 10250 17210 2090 ...
Umsatz.Kreditkarte	: int 3510 740 910 1620 600 950 3500 2860 3180 1270 ...
Geschlecht	: Factor w/ 2 levels "m","w": 1 2 1 2 1 2 2 1 1 2 ...
alleinstehend	: Factor w/ 2 levels "ja","nein": 2 2 2 2 2 2 1 2 2 1 ...
wohnung	: Factor w/ 3 levels "Appartement",...: 2 3 3 2 1 3 3 3 3 1 ...
Supermarkt.besuche.pro.Monat	: int 7 4 1 3 3 2 6 5 10 7 ...
Lasagne.probiert	: Factor w/ 2 levels "ja","nein": 2 1 2 2 1 2 1 1 1 1 ...
X	: logi NA NA NA NA NA NA NA ...

Abbildung 3.4: Betrachten der Datentypen des Datensatzes in RStudio

Zu sehen ist hier, dass es sich um einen Datensatz mit 13 Variablen und 856 Beobachtungen handelt (hier scheint ein falscher Zeichensatz vorzuliegen oder Ähnliches, da eine zusätzliche leere Spalte „X“ angezeigt wird **TODO: was hier tun?**). Zusätzlich können die vorgeschlagenen Datentypen von R betrachtet werden. Bei den meisten Spalten handelt es sich um Ganzzahlen (int) oder Factoren (manchmal auch als Enums bezeichnet).

Explore the Data

Ist die grobe Sichtung der Daten abgeschlossen, wird näher an der Fragestellung des Data Mining gearbeitet. Dazu werden „Abfrage-, Visualisierungs- und Reporting[-Techniken]“ (Shearer, 2000, S. 16; eigene Übersetzung) eingesetzt. Um der Antwort auf die ursprüngliche Fragestellung näher zu kommen oder die Fragestellung zu verfeinern, werden beispielsweise folgende Eigenschaften betrachtet (Chapman et al., 2000, S. 18; eigene Übersetzung):

- Die Verteilung der Schlüsselattribute (zum Beispiel der Zielvariablen bei einer Vorhersage).
- Die Beziehungen zwischen Wertepaaren oder kleinen Attributgruppen.
- Die Ergebnisse einfacher Aggregationen.
- Die Beschaffenheit von aussagekräftigen Teilgruppen von Werten.
- Die Ergebnisse einfacher statistischer Analysen.

3 Grundlagen

Verify Data Quality

Der letzte Schritt der zweiten Phase evaluiert die Qualität der Daten. (Chapman et al., 2000, S.19) nutzen die Zielfragen:

- „Is the data complete (does it cover all the cases required)?“
- „Is it correct, or does it contain errors and, if there are errors, how common are they?“
- „Are there missing values in the data?“
- „If so, how are they represented, where do they occur, and how common are they?“

(Shearer, 2000, S. 16) empfiehlt zusätzlich noch, zu prüfen, ob die Werte plausibel sind, wie die Schreibweisen sind, ob Attribute mit unterschiedlichen Werten aber gleicher Bedeutung vorhanden sind und schließlich ob es einen „conflict with common sense“ ,wie „teenagers with high income“(Shearer, 2000, S. 16) gibt.

Data Preparation

In der aufwendigsten Phase des ganzen Prozesses (siehe Abbildung 3.5) wird das „final data set“(Larose, 2014, Punkt 1.4.1.3.a; Shearer, 2000, S. 16) erzeugt. Dies geschieht durch(Swamynathan, 2017, S. 73)

- generelle Transformationen,
- Füllen der Datenlücken, die in vorhergehenden Schritten aufgedeckt wurden,
- Befassen mit fehlenden Werten,
- Herausarbeiten, welche Features des Datensatzes die größte Relevanz haben und welche neuen Features sinnvoll wären.

Wie bereits erwähnt, handelt es sich nicht nur um die Phase, die den meisten Aufwand erfordert, sondern auch um die, von der die Genauigkeit des Endresultates zu großen Stücken abhängt.(Swamynathan, 2017, S. 73)

3 Grundlagen

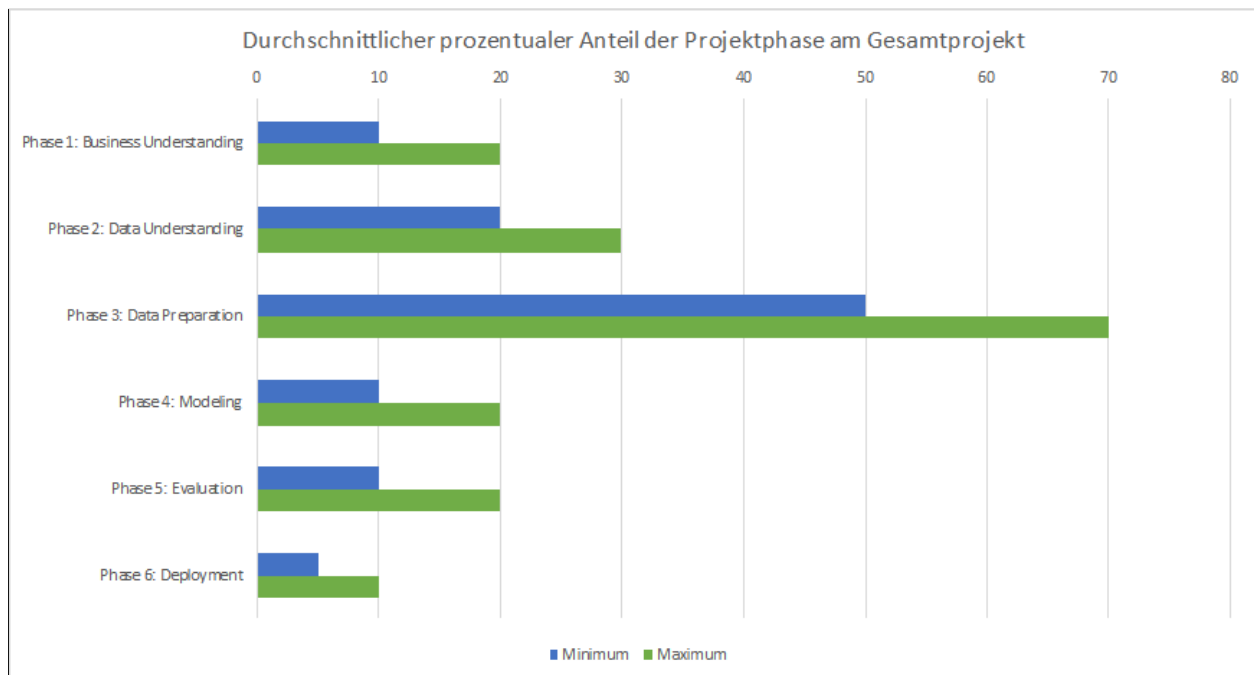


Abbildung 3.5: Durchschnittlicher prozentualer Anteil der CRISP-DM-Projektphase am Gesamtprojekt nach (Shearer, 2000, S. 15; eigene Darstellung)

Select Data

Genauer wird dabei im ersten Schritt ausgewählt, welche Daten Teil der Analyse bleiben und welche exkludiert werden. Kriterien sind dabei, die Relevanz hinsichtlich der Ziele, die Qualität der Daten und technische Grenzen (Shearer, 2000, S. 16) (wie „data volume or data types“ (Chapman et al., 2000, S. 21)). Zusätzlich kann überlegt werden, ob einige Attribute wichtiger sind als andere. So kann beispielsweise bei einer landesweiten Kundenanalyse die Postleitzahl der Kunden ausreichen und Straße und Hausnummer vernachlässigt werden. (Shearer, 2000, S. 16) (Chapman et al., 2000, S. 21) merken an, dass diese Phase sowohl „attributes (columns)“, als auch „records (rows)“ umfasst. Wie bereits in den vorhergehenden Schritten muss auch hier erklärt werden, warum Entscheidungen getroffen wurden und eine Dokumentation angefertigt werden. (Shearer, 2000, S. 16)

Clean Data

Im Schritt „Verify Data Quality“ wurde herausgearbeitet, wie die Qualität der Daten ist **TODO: bessere Formulierung** und wie mangelbehaftet sie sind. Jetzt werden Maßnahmen dagegen ergriffen. Neben trivialen Vorgehen wie „Auswahl von reinen Untermengen“ oder „Einfügen von passenden Standardwerten“ können „anspruchsvollere Techniken wie das

3 Grundlagen

Schätzen von fehlenden Werten“(Chapman et al., 2000, S.21; eigene Übersetzung) zum Zuge kommen.

Construct Data

Die gereinigten Daten sind noch nicht fertig für die Modeling-Phase. Manchmal ist es notwendig, einem Datensatz neue Zeilen hinzuzufügen. Betrachtet man wieder eine Kundenanalyse, so ist es vielleicht nötig, für einen Kunden, der in einem Quartal keinen Einkauf getätigt hat, einen leeren Einkauf (null Euro) anzulegen, falls der eingesetzte Algorithmus dies erfordert.(Chapman et al., 2000, S. 22) Er kann auch verlangen, dass abgeleitete statt der Ursprungswerte benötigt werden. Dabei gibt es nach (Shearer, 2000, S. 16) zwei Fälle:

1. Wenn zu einem Kunden ein Bewegungsprofil vorhanden ist (in welchem Geschäft er einkaufen war), so ist möglicherweise sinnvoll, nicht das gesamte Profil zu betrachten, sondern lediglich die Fläche zu betrachten, in der er sich bewegt hat.
2. Ebenfalls zielführend kann eine „single-attribute transformation“ sein. Dabei wird beispielsweise das genaue Alter der Kunden in Altersspannen umgewandelt oder sprechende Werte wie „(“definitely yes,,, “yes,,, “don’t know,,, “no,,)“ in numerische Werte übersetzt.

(Shearer, 2000, S. 16; eigene Übersetzung) merkt aber auch an, dass es nicht immer sinnvoll ist dies zu tun, auf jeden Fall „nicht nur um die Anzahl der Inputattribute zu reduzieren.“

Integrate Data

Der vorletzte Schritt der Data Preparation ist das Zusammenführen von mehreren Quellen oder Tabellen mit dem gleichen Thema. Dadurch können „neue Beobachtungen oder Werte“(Chapman et al., 2000, S. 22) gewonnen werden. In Tabelle 4.34 werden die zwei Hauptaufgaben(Shearer, 2000, S. 17) genauer erläutert.

3 Grundlagen

Aufgabe	Erläuterung	Beispiel
Join	Mehrere Tabellen zum gleichen Thema werden zusammengeführt.	<p>Die drei Ausgangstabellen</p> <ul style="list-style-type: none"> • „information about each store’s general characteristics (e.g., floor space, type of mall)“ • „summarized sales data (e.g., profit, percent change in sales from previous year)“ • „information about the demographics of the surrounding area“ <p>werden in</p> <ul style="list-style-type: none"> • „a new table with one record for each store, combining fields from the source tables“ (Shearer, 2000, S. 16) <p>zusammengeführt.</p>
Aggregation	Errechnen neuer Werte aus Informationen verschiedener Tabellen.	<p>Das Überführen von einer</p> <ul style="list-style-type: none"> • „table of customer purchases, where there is one record for each purchase“ <p>in eine</p> <ul style="list-style-type: none"> • „new table where there is one record for each customer“ <p>mit den Feldern</p> <ul style="list-style-type: none"> • „number of purchases, the average purchase amount, the percent of orders charged to credit cards, the percent of items under promotion, etc.“ (Shearer, 2000, S. 17)

Tabelle 3.6: Die zwei Hauptaufgaben des Schrittes Integrate Data

3 Grundlagen

Format Data

Die Datenformatierung umfasst „hauptsächlich syntaktische Abänderungen“ und „verändert nicht die Bedeutung“(Chapman et al., 2000, S. 22; eigene Übersetzung) der Daten. Das kann zum Beispiel das „Entfernen von unerlaubten Zeichen in Zeichenketten“(Shearer, 2000, S. 17; eigene Übersetzung) sein.

Mit diesem Schritt ist die Data Preperation abgeschlossen und es kann mit dem Modeling begonnen werden.

Modeling

Die Phase des Modeling umfasst die Auswahl einer oder mehrerer Data Mining-Algorithmen, die Optimierung ihrer Parameter und Settings und der Evaluierung des erzeugten Models.(Swamynathan, 2017, S. 73; Larose, 2014, Punkt 1.4.1.4) Eventuell muss der Datensatz noch angepasst werden, sodass die Data Preperation-Phase noch einmal durchlaufen werden muss.(Larose, 2014, Punkt 1.4.1.4)

Select the Modeling Technique

Wie der Name des ersten Schrittes bereits andeutet, wird eine Modellierungstechnik ausgewählt. Werden mehrere Techniken ausgewählt, so wird diese Phase mehrfach (parallel) durchlaufen. Wichtig ist hier, dass getroffene Annahmen (wie „alle Attribute sind stetig Gleichverteilt“ oder „fehlende Werte sind nicht zugelassen“(Chapman et al., 2000, S. 24)) dokumentiert werden.(Shearer, 2000, S. 17)

Generate Test Design

Vor der eigentlichen Modellierung wird festgelegt, wie die „Qualität und Validität“(Chapman et al., 2000, S. 24) festgestellt werden soll. Für „supervised data mining“ werden dabei meist „error rates“(Chapman et al., 2000, S. 24) herangezogen. Dazu wird das Modell mit einem Datensatz (train set) trainiert und mit einem anderen (test set) getestet.(Shearer, 2000)

Build the Model

Der kürzeste Schritt dieser Phase ist „Build the Model“. Hier wird das Model mit Hilfe eines - möglicherweise vorher bereits gewählten - Werkzeuges erzeugt.(Chapman et al., 2000, S. 24;

3 Grundlagen

Shearer, 2000, S. 17) Wurden zwei Schritte zuvor mehrere Modellierungstechniken ausgewählt, so liegen an dieser Stelle mehrere Models vor.

Assess the Model

Ist das Modellieren abgeschlossen, werden die Ergebnisse auf Basis

- des Verständnisses aus der ersten Phase (Business Understanding),
- der Data Mining-Ziele und
- des Test Designs aus dieser Phase

interpretiert. Der Analyst hat die Aufgabe, den Grad des Erfolgs des Data Mining zu bestimmen. Dazu kann er Experten heranziehen, um das Ergebnis beispielsweise auf Geschäftsebene zu diskutieren. Zusätzlich wird eine Rangliste aller Modelle aufgestellt, die den Erfolg hinsichtlich der „Business Ziele“ (aus Phase „Business Understanding“, Schritt „Determine the Business Objectives“) abbildet.(Chapman et al., 2000, S. 25; Shearer, 2000, S. 17)

In diesem Schritt werden die Modelle ein erstes Mal interpretiert. **TODO: bessere Wortwahl als „interpret**

Eine genauere Evaluation und zusätzliche Ergebnisse, Erkenntnisse und Dokumente aus den vorhergehenden Schritten werden in der nachfolgenden Evaluation-Phase bewertet.(Chapman et al., 2000, S.25)

Evaluation

Die Tatsache, dass es sich beim CRISP-DM-Referenzmodell um einen Prozess handelt und nicht um strikt getrennte Einzelschritte, wird besonders in der Evaluations-Phase deutlich. Erstens wird das Ranking aus dem vorherigen Schritt in in einem „Benchmarking“ über die „Models mit einer hohen Genauigkeit“(Swamynathan, 2017, S. 73; eigene Übersetzung) verfeinert. Zweitens werden die Models erneut mit frischen Daten (nicht aus Schritt „Generate Test Design“) verifiziert und gegen die Business-Anforderungen aus Phase 1 geprüft.(Swamynathan, 2017, S. 73) Ziel dieser Phase ist vor allem, dem Projektleiter genug Wissen an die Hand zu geben, um zu Entscheiden, wie mit den Ergebnissen des ganzen Prozesse weiter verfahren wird.

3 Grundlagen

Evaluate Results

Während sich bisher hauptsächlich um die „Genauigkeit und Allgemeingültigkeit“ der Modelle gekümmert wurde, wird jetzt auch betrachtet, ob es „irgendwelche Businessgründe gibt“, durch die das Model „mangelhaft“ (Shearer, 2000, S. 18; eigene Übersetzung) wird. Falls „time und budget“ (Chapman et al., 2000, S. 26) es erlauben, können die Ergebnisse bereits in echte Systeme in Testumgebungen implementiert werden. Wie bereits angemerkt, werden in dieser Phase auch andere „findings“ evaluiert, die beispielsweise auf zukünftige Herausforderungen hinweisen. (Shearer, 2000, S. 18) Ist dies geschehen, „fasst der Data Analyst die Bewertungen der Ergebnisse hinsichtlich der geschäftlichen Erfolgskriterien zusammen“ und gibt seine Wertung ab, „ob das Projekt bereits die initialen geschäftlichen Ziele erreicht“ (Shearer, 2000, S. 18; eigene Übersetzung).

Review Process

Im Review wird abgesichert, dass kein Faktor unbeachtet geblieben ist und keine Aufgabe vergessen wurde. Ebenfalls wird die Qualität gesichert (zum Beispiel Bugs in Softwarekomponenten gesucht) und rechtliche Überlegungen angestellt („Dürfen wir diese Kundendaten produktiv für diese Analyse benutzen?“). (Shearer, 2000, S. 18; Chapman et al., 2000, S. 27)

Determine Next Steps

Schließlich werden alle Bewertungen bis hierher genutzt, um zu entscheiden, ob eine weitere Prozessiteration durchlaufen wird, oder, ob in die Deployment-Phase übergegangen wird. Laut (Shearer, 2000, S. 18) trifft diese Entscheidung der Projektleiter. (Chapman et al., 2000, S. 17) sind der Meinung, dass das ganze Projektteam entscheiden sollte.

Deployment

Ist die Entscheidung für das Deployment gefallen, wird die letzte Phase initiiert. Zu Beginn des CRISP-DM-Prozesses wurden Ziele festgelegt, die begründen, weshalb das Data Mining durchgeführt werden soll. Eine einfache Implementierung wäre das Erstellen eines Reports, eine Komplexere dagegen, den Data Mining Prozess in eine andere Abteilung zu portieren (Larose, 2014, Punkte 1.4.1.6.b und c) oder „Echtzeit-Personalisierung von Webseiten“ (Shearer, 2000, S. 18; eigene Übersetzung) durchzuführen.

3 Grundlagen

Die Implementierung des Models in die produktiven Systeme befriedigt diese Ziele nicht alleine. Auch das Training jener Personen, die das Wissen im Geschäftsprozess anwenden, muss durchgeführt werden. Dies beinhaltet sowohl die Fähigkeit, die Ergebnisse zu interpretieren, als auch zu verstehen, wie sie die Entscheidungsfindung unterstützen können.(Swamynathan, 2017, S. 73)

Da die weiteren Aufgaben oft nicht vom Data Analyst durchgeführt werden (Larose, 2014, Punkt 1.4.1.6.d), muss der Anwender die Pflege eines Machine Learning-Models verstehen und übernehmen (z.B. in welchen Intervallen das Model trainiert wird).(Swamynathan, 2017, S. 74)

Plan Deployment

Der Erste Schritt der Deploymentphase ist die Auswahl und Dokumentation einer geeigneten Strategie für den Einsatz oder das Rollout in die Geschäftsumgebung.(Shearer, 2000, S. 18; Chapman et al., 2000, S. 28) **TODO: Formulierung** .

Plan Monitoring and Maintenance

Zusätzlich zum Rollout, muss die Überwachung und Wartung bedacht und geplant werden. Das soll der Fehlbenutzung der Data Mining-Ergebnisse vorbeugen.(Shearer, 2000, S. 18; Chapman et al., 2000, S. 29)

Produce Final Report

Ein nicht unbedingt Data-Mining-spezifischer Schritt, ist das erstellen eines Abschlussberichts. Dieser kann sich je nach Projekttyp unterscheiden. Er kann die Form einer Zusammenfassung haben oder eine ausgedehnte und detaillierte Präsentation sein.(Shearer, 2000, S. 18; Chapman et al., 2000, S. 29) (Larose, 2014, Punkt 1.4.1.1) merkt an, dass es sich auch um Forschungsprojekte handeln kann. In diesem Fall ist der Report möglicherweise **TODO: Wort** eine Veröffentlichung der Ergebnisse. Der Abschlussbericht „enthält alle bisher erzeugten Auslieferungsgegenstände und fasst [...] die Ergebnisse zusammen.“(Shearer, 2000, S 18; eigene Übersetzung)

Review Project

Den Schlussstrich zieht das Review des Projektes. Hier wird festgehalten, was im Projektverlauf gut und schlecht lief. Zusätzlich soll das Wissen konserviert werden, wie der Prozess verbessert werden könnte. (Shearer, 2000, S. 18; Chapman et al., 2000, S. 29) **TODO: Formulierung...** (Shearer, 2000, S. 18; eigene Übersetzung) empfiehlt „Interviews mit allen wichtigen Projektteilnehmern“. In „idealen Projekten“ umfasst das Review „alle Reports, die in vorhergehenden Projektphasen [...] verfasst wurden.“ (Chapman et al., 2000, S. 29; eigene Übersetzung)

3.1.3 Sample, Explore, Modify, Model and Assess (SEMMA)

3.1.4 Auswahl

An dieser Stelle fehlt noch, dass ich mich für CRISP-DM entscheide, weil es de facto standard ist (KDNugget Umfrage..). Deswegen im Praxisteil dann Orientierung am Prozessmodell und den Schritten aber auch am mitgelieferten User guide.

3.2 Machine Learning

Der nun folgende Abschnitt befasst sich mit Machine Learning. Zu Beginn der Arbeit wurde bereits erwähnt, dass maschinelles Lernen als „Sammlung von Algorithmen und Techniken“ verstanden werden kann, die „genutzt werden, um Computersysteme zu erstellen, die aus Daten lernen, um Vorhersagen zu erstellen“. (Swamynathan, 2017, S. 53; eigene Übersetzung) Um diese Algorithmensammlung genauer zu betrachten, ist es sinnvoll sie nach bestimmten Kategorien zu ordnen.

- (Kubat, 2017) unterscheidet in seinem Werk unter anderem nach verschiedenen Klassifikationen (Bayesianisch, Nearest-Neighbor, Linear und Polynomial), künstlichen neuronalen Netzen (engl. artificial neuronal network, kurz: ANN), Entscheidungsbäumen, Unsupervised Learning, Genetische Algorithmen und Reinforcement Learning.
- Einen anderen Ansatz wählt (Swamynathan, 2017). Er gliedert die Algorithmen nach Supervised Learning (mit Regressionen und Klassifikationen), Unsupervised Learning (mit Clusteranalyse, Dimensionsreduzierung und Anomalie-Erkennung) und Reinforcement

3 Grundlagen

Learning (Markow-Markow-Entscheidungsprozess, Q-Learning, Temporal Difference- und Monte-Carlo Methoden). (Kim, 2017) wählt die gleiche Kategorisierung in die drei Typen.

- (Paluszek and Thomas, 2017) sehen neben Supervised und Unsupervised Learning noch Semisupervised und Online Learning.

Diese unterschiedlichen Gliederungen erklären (Ramasubramanian and Singh, 2017, S. 222) damit, dass entweder nach „Learning types“ (siehe Abbildung 3.6) oder „Subjective grouping“ (siehe Tabelle 3.7) klassifiziert werden kann.

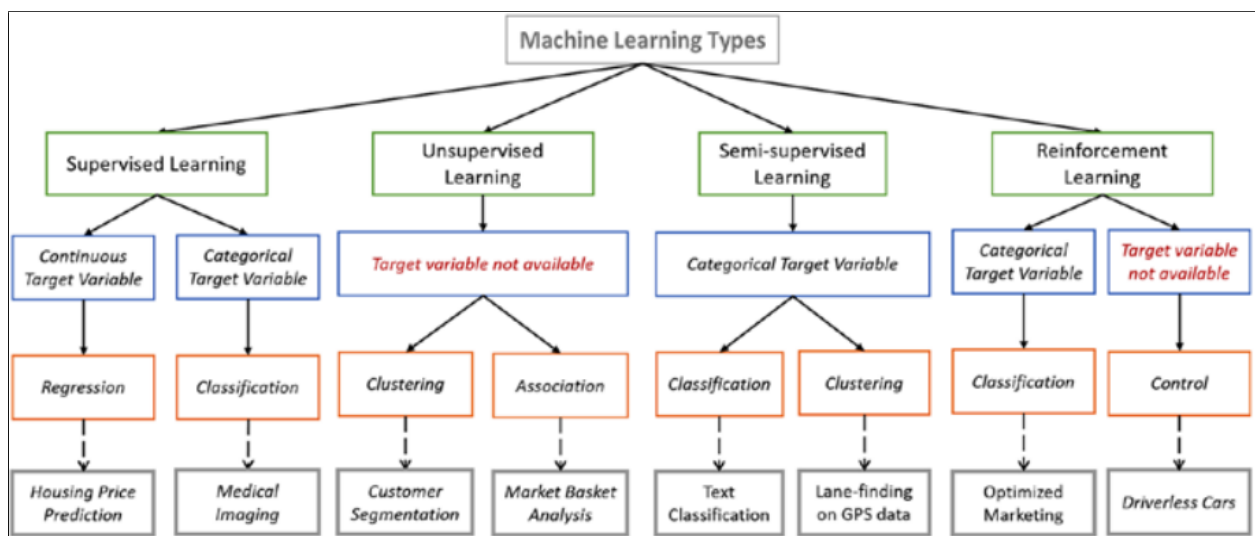


Abbildung 3.6: Machine Learning Types nach (Ramasubramanian and Singh, 2017, S. 222)

3 Grundlagen

Gruppe	Algorithmen
Regression Analysis	Ordinary Least Square Regression (OLSR) Linear Regression Logistic Regression Stepwise Regression Polynomial Regression Locally Estimated Scatterplot Smoothing (LOESS)
Distance-based algorithms	k-nearest Neighbor (kNN) Learning Vector Quantization (LVQ) Self-Organizing Map (SOM)
Regularization algorithms	Ridge Regression Least Absolute Shrinkage and Selector Operator (LASSO) Elastic Net Least-Angle Regression (LARS)
Decision tree algorithms	Classification and Regression Tree (CART) Iterative Dichotomiser 3 (ID3) C4.5 and C5.0 (different versions of a powerful approach) Chi-squared Automatic Interaction Detection (CHAID) Random Forest Conditional Decision Tree
Bayesian algorithms	Naive Bayes Gaussian Naive Bayes Multinomial Naive Bayes Nayesian Belief Network (BNN) Bayesian Network (BN)
Clustering algorithms	k-Means k-Medians Partitioning Around Medoids (PAM) Hierarchical Clustering
Association rule mining	Apriori algorithm Eclat algorithm FP-growth algorithm Context Based Rule Mining
Artificial neural networks	Perception Back-Propagation Hopfield Network Radial Basis Function Network (RBFN)
Deep learning algorithms	Beep Boltzmann Machine (DBM) Deep Belief Networks (DBN) Convolutional Neural Network (CNN) Stacked Auto-Encoders
Dimensionality reduction algorithms	Principam Component Analysis (PCA) Principam Component Regression (PCR) Partial Least Squares Regression (PLSR) Multidimensional Scaling (MDS) Linear Discriminant Analysis (LDA) Mixture Discriminant Analysis (MDA) Quadratic Discriminant Analysis (QDA)
Ensemble learning	Boosting Bagging AdaBoost Stacked Generalization (blending) Gradient Boost Machines (GBM)
Text mining algorithms	Automatic summarization Named entity recognition (NER) Optical character recognition (OCR) Part-of-speech tagging Sentiment analysis Spec recognition Topic Modeling

Tabelle 3.7: Sebjective Grouping nach (Ramasubramanian and Singh, 2017, S. 224-229)

Da der Artikel „How to choose algorithms for Microsoft Azure Machine Learning“ von (Ericson and Rohm, 2017b) nach „Supervised“, „Unsupervised“ und „Reinforcement learning“ zurückgreift, wird nachfolgend diese Gliederung genutzt und um „Semi-supervised Learning“

3 Grundlagen

und „Active Learning“ erweitert.

TODO: raus?: Anzumerken ist, dass nur Algorithmen beschrieben werden, die zum Zeitpunkt der Arbeit in „Microsoft Azure Machine Learning Studio“ verfügbar sind.

3.2.1 Supervised Learning

Die möglicherweise am einfachsten nachvollziehbare Gruppe des Machine Learning ist das Supervised Learning, da es „sehr ähnlich zu dem Prozess ist, in dem Menschen Dinge lernen“ (Kim, 2017, S. 13; eigene Übersetzung). Es existiert ein Datensatz, bei dem für jeden Input ein Output vorhanden ist. Ein Beispiel können Patientendaten sein, die als Output eine Variable besitzen, die angibt, ob ein Patient an Krebs erkrankt ist oder nicht. (Ramasubramanian and Singh, 2017, S. 222) Diese „response variable“ (Krebs oder nicht Krebs) wird als „label“ (Ramasubramanian and Singh, 2017, S. 222) bezeichnet. Die Aufgabe des Learning Prozesses ist es dann, einen Zusammenhang zwischen dem Input (Patientendaten) und dem Label (Krebs oder nicht Krebs) herzustellen. Dies geschieht mit sogenannten „training sets“ (Paluszek and Thomas, 2017, S. 5) Der zweite Schritt ist dann das Überprüfen des entstandenen Models. Dabei wird das Model auf ein zweites gelabeltes „test set“ (Paluszek and Thomas, 2017, S. 5) angewandt und das Ergebnis überprüft.

Die Algorithmen des Supervised Learning lassen sich erneut aufteilen:

Classification

Betrachtet man die Labels und stellt fest, dass sie die Datensätze in Kategorien unterteilen (Krebs oder nicht Krebs) oder eine Wahrscheinlichkeit angeben (Person zu 89% Max Mustermann bei einer Gesichtserkennung), handelt es sich um eine Klassifikation (engl. Classification). (Swamynathan, 2017, S. 67) (Kauchak, 2016, S. 5) nennt als Beispiele

- biometrische Erkennungen (Gesicht, Iris, Unterschrift etc.),
- Buchstabenerkennung,
- Spamfilter und
- medizinische Diagnosen.

3 Grundlagen

Weiterhin kann zwischen Klassifikationen mit nur zwei Labels („two-class or binomial classification“ (Ericson and Rohm, 2017a)) (siehe Abbildung 3.7) oder mehr als zwei Labels („multi-class classification“ (Ericson and Rohm, 2017a)) unterschieden werden.

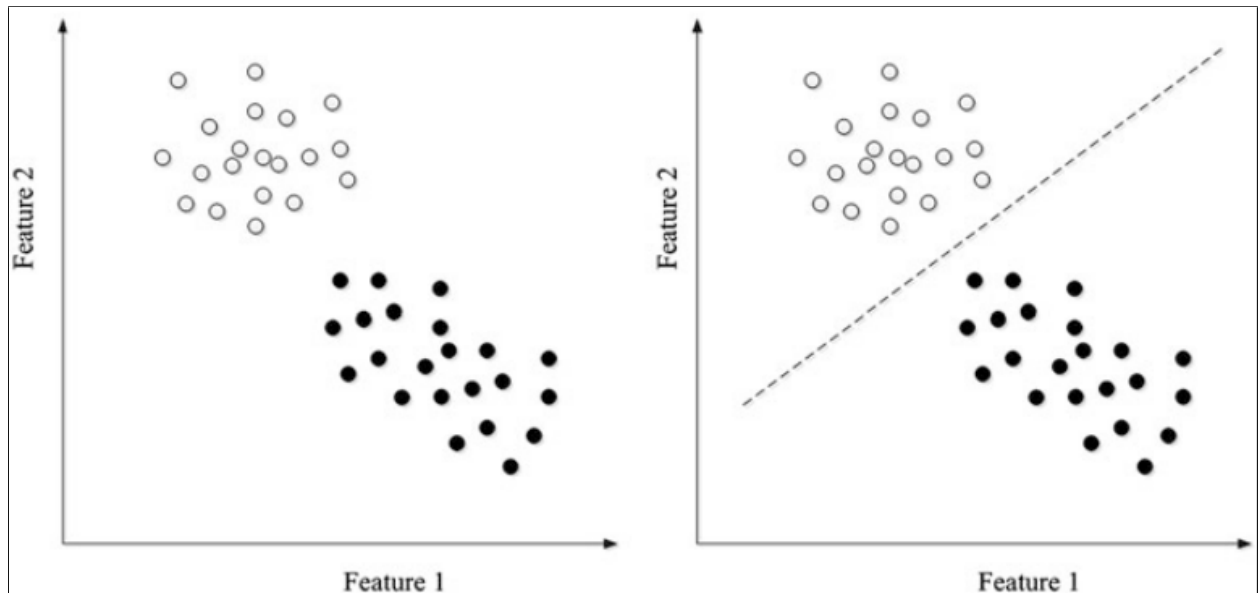


Abbildung 3.7: Beispiel für eine Klassifikation aus (Suthaharan, 2016, S. 8)

Regression

Wenn eine Unterteilung in Kategorien wie gerade genannt nicht möglich ist und die Output variable ein fortlaufender **TODO: Wort** Wert ist, werden Regressionen benutzt. Dabei liegt der „Hauptfokus [...] darin, einen Zusammenhang zwischen einer abhängigen Variablen und einer oder mehreren unabhängigen [...] Variablen herzustellen“ (Swamynathan, 2017, S. 60; eigene Übersetzung). Neben dem bekanntesten Beispiel

- einen Kurs an der Börse vorherzusagen (Kauchak, 2016, S. 5; Kubat, 2017, S. 207; Ericson and Rohm, 2017a), gibt es Anwendungsfälle in
- der Epidemiologie,
- der Auto- und Flugzeugnavigation und
- Analysen im zeitlichen Verlauf (Wetterveränderung im Verlauf der Zeit) (Kauchak, 2016, S. 5).

Anomaly detection

Im bereits mehrfach zitierten Artikel „How to choose algorithms for Microsoft Azure Machine Learning“ von (Ericson and Rohm, 2017a) wird noch eine zusätzliche Kategorie genannte: Anomaly detection. Bei (Swamynathan, 2017, S. 68) ist die Anomaly detection dem Unsupervised Learning zugeordnet. Dies rührt daher, dass es darauf ankommt, ob eine Outputvariable („Label“) vorhanden ist, oder nicht. Tatsächlich ist es so, dass es sowohl Szenarien für Supervised und Unsupervised Anomaly detection gibt, als auch für das später noch beschriebene Semi-supervised Learning.(Chandola et al., 2009, S. 15:10) Unabhängig davon beschreibt Anomaly detection das Finden von Mustern in Daten, die vom erwarteten Verhalten abweichen. Diese Pattern werden meist als Anomalität (engl. anomaly) oder Ausreißer (engl. outlier) bezeichnet.(Chandola et al., 2009, S. 15:1) Die Anomaly detection kann für folgende Szenarien genutzt werden(Chandola et al., 2009, S. 15:2):

- Kreditkartenbetrugserkennung
- Versicherungsbetrugserkennung
- Gesundheitsprüfungen
- Intrusion Detection
- Militärische Überwachungen
- Anwendungen in „der Welt des Internet der Dinge“[S. 68](Swamynathan, 2017)

3.2.2 Unsupervised Learning

Der „Glücksfall“ TODO: kann man das so sagen?, dass der vorhandene Datensatz Label besitzt, ist unter realen Umständen häufig nicht der Fall. Um aus diesen Daten trotzdem Schlüsse zu ziehen, werden Methoden des Unsupervised Learning herangezogen. Hier liegt der Fokus auf dem „Entdecken von aufschlussreichen Eigenschaften der verfügbaren Daten“(Kubat, 2017, S.277; eigene Übersetzung) und der „Untersuchung der Charakteristik der Daten“(Kim, 2017, S. 13; eigene Übersetzung). Dies kann das Ziel haben komplexe und vielschichtige Daten zu vereinfachen und zu strukturieren(Ericson and Rohm, 2017a) oder die Daten in Gruppen aufzuteilen(Lison, 2012, S. 22). Um diese Gruppen ähnlicher Daten -

3 Grundlagen

sogenannte Cluster - geht es im nächsten Abschnitt.

Supervised Learning findet man nach (Ramasubramanian and Singh, 2017, S. 223)

- bei der Aufteilung von Kundendaten in Segmente,
- in Analysen von sozialen Netzwerken,
- in der Klimatologie,
- bei der Bildkompression und
- in der Bioinformatik.(Kauchak, 2016, S. 6)

Clustering

Beim angesprochenen Clustering handelt es sich um das „Identifizieren von distinkten Gruppen [...] basierend auf irgendeiner Art der Ähnlichkeit innerhalb des vorliegenden Datensatzes“(Swamynathan, 2017, S. 195; eigene Übersetzung). Damit die Objekte der gebildeten Cluster „aussagekräftig und sinnvoll“ sind, sollen „die Objekte innerhalb eines Clusters [...] homogen sein“ und „zu Objekten anderer Cluster“(Ramasubramanian and Singh, 2017, S. 337; eigene Übersetzung) heterogen (siehe Abbildung 3.8). Es kann jedoch auch sein, dass Objekte zu mehreren Clustern gehören. Dies nennt sich „Soft Clustering“ - im Gegensatz zum „Hard Clustering“(Ramasubramanian and Singh, 2017, S. 339). Die Metrik für die Ähnlichkeit ist nicht festgelegt. Möglich sind

- die „Distanz [...] zwischen Beobachtungen“,
- die „Entfernung vom Mittelwert jeder Beobachtung/des Clusters“,
- die „Signifikanz [einer] statistischen Verteilung“ oder
- die „Dichte im Datenraum“(Ramasubramanian and Singh, 2017, S. 338; eigene Übersetzung).

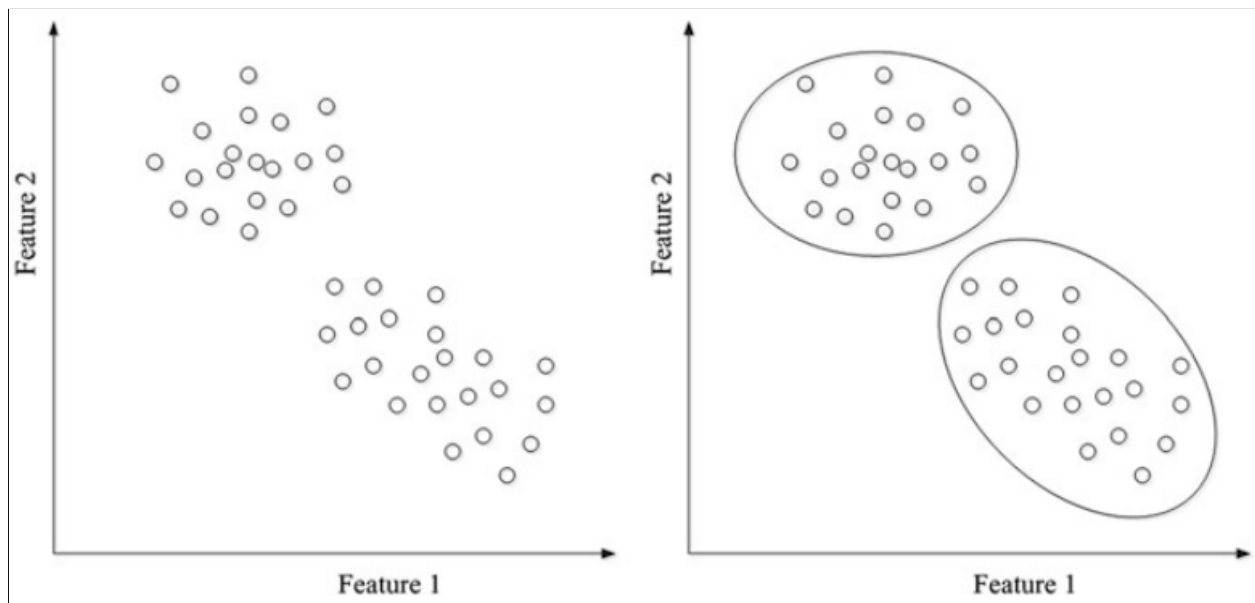


Abbildung 3.8: Beispiel für zwei Cluster (Suthaharan, 2016, S. 9)

3.2.3 Semi-supervised Learning

TODO: hier so oft „label...“:

Bis jetzt wurden zwei Extremfälle beschrieben: entweder es gibt keine Labels oder alle Daten sind gelabelt. Es existieren jedoch auch Fälle dazwischen. Besitzen die meisten Daten ein Label, so ist es eventuell möglich, die Datensätze ohne Label für das Learning zu entfernen und ein Model aus dem Supervised Learning heranzuziehen. Tritt aber ein Problem auf, bei dem nur sehr wenige Daten gelabelt sind, empfiehlt sich ein Vorgehen aus dem Semi-supervised Learning. Methoden dieser Familie des Machine Learning beruhen auf der Annahme, dass „die Daten wichtige Informationen über die Gruppenzugehörigkeit beinhalten“, „obwohl die Gruppenzugehörigkeit [...] unbekannt ist“ (Ramasubramanian and Singh, 2017, S. 223; eigene Übersetzung). Als umfassendes Werk für Semi-supervised Learning ist an dieser Stelle (Chapelle et al., 2006) zu empfehlen.

Zur Anwendung kommt Semi-supervised Learning seit den 1990er Jahre in „natural language problems“ und „text classification“ (Chapelle et al., 2006, S. 4).

3.2.4 Active Learning

Ein Sonderfall des Semi-Supervised Learning ist das Active Learning. Hier wird davon ausgegangen, dass nur wenige oder keine Labels vorhanden sind, diese jedoch durch einen „Menschen mit umfangreichem Wissen im Themengebiet“ (Olsson, 2009, S. i; eigene Übersetzung) hinzugefügt werden können. Alle Daten mit einem Label zu versehen wäre jedoch zu „schwer, zeitaufwendig oder zu teuer“ (Settles, 2010, Abstract; eigene Übersetzung). Mit dem Ziel an den Menschen nur so wenig Anfragen (engl. Queries) wie nötig zu stellen (Olsson, 2009, Abstract), werden Algorithmen entworfen, die selbst wählen können, welche Datensätze gelabelt werden. (Settles, 2010, Abstract)

Nach (Settles, 2010, S. 4) wird Active Learning beispielsweise in

- der Spracherkennung,
- der Informationsextraktion oder in
- der Klassifikation oder dem Filtern von Dokumenten oder Mediendateien eingesetzt

3.2.5 Reinforcement Learning

„Beim Reinforcement Learning interagiert der Lerner“ - also ein Programm - „mit seiner Umwelt“ (Settles, 2010, S. 45; eigene Übersetzung). Er „experimentiert“ dabei um eine Lösung auf das gestellte Problem zu finden und erhält je nach Ausgang seiner Aktion eine Belohnung oder eine Bestrafung. (Kubat, 2017, S. 331) Es wird also nicht direkt mit einem Output (Label) gearbeitet, sondern lediglich die Qualität des Outputs bewertet. Das Ziel ist, durch ein iteratives Vorgehen (siehe Abbildung 3.9), ein maximales Endergebnis (größer Reward) zu erreichen. (Swamynathan, 2017, S. 69)

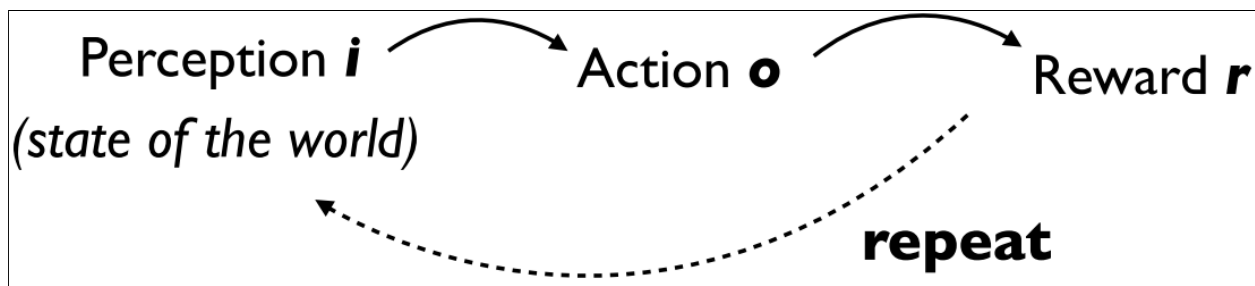


Abbildung 3.9: Iterativer Reinforcement Learning Prozess aus (Lison, 2012, S. 25)

3 Grundlagen

Reinforcement Learning kommt in Szenarien zum Einsatz, in denen noch keine Daten zum Lernen zur Verfügung stehen oder erst nach und nach aktualisiert werden. (Ramasubramanian and Singh, 2017, S. 223)

Ein prominentes Beispiel für ein System mit Reinforcement Learning ist Google DeepMinds AlphaGo. Es handelt sich dabei um einen Go (asiatisches Brettspiel) spielendes Computersystem, dass seine Spielstärke durch spielen gegen sich selbst erreichte. (Silver et al., 2017)

3.3 Kryptowährung(en)

In der Motivation (siehe Abschnitt 1.2) wurde bereits auf Kryptowährungen allgemein und auf Bitcoin im Speziellen eingegangen. Es wurde ebenfalls angemerkt, dass die zwei Währungen mit dem größten Marktvolumen Bitcoin (BTC) und Ethereum (ETH) sind. Aus diesem Grund werden nur die Kurse dieser beiden Währungen analysiert. In diesem Abschnitt wird darauf verzichtet, weiter auf die allgemeine Technik hinter BTC und ETH einzugehen. In Punkt 4.2.2 - vor allem in den Tabellen 4.27 und 4.28 - tauchen bei der Beschreibung der Datensätze jedoch Begriffe auf, die einer genaueren Erklärung bedürfen:

- Die **Marktkapitalisierung** ist der aktuelle Marktwert einer Kryptowährung. Er ergibt sich aus dem Produkt des aktuellen Handelskurses und der totalen Anzahl der geschürften Währungseinheiten. Bei einem Kurs von 500 USD pro Ether und einem Volumen von 1000 Ether, ist die Marktkapitalisierung 500000 USD.
- Das Handelsvolumen (engl. **trade volume**) ist die Anzahl der Kryptowährungseinheiten, die in einem bestimmten Zeitfenster - meist 24 Stunden - gehandelt wurden.
- Die **Blocksize** ist die Größe eines Blocks (meist in Megabyte).
- **Stale Blocks** sind Blöcke der Blockchain, die korrekt sind, aber ihren Weg in die Blockchain nicht gefunden haben. Dies kann passieren, wenn zwei Blöcke gleichzeitig fertiggestellt werden, jedoch nur einer davon in die Blockchain aufgenommen werden kann. (Project, 2017c) Im Ethereumnetzwerk heißen diese Blöcke '**Uncles**'. Entgegen dem Bitcoinnetzwerk werden dort Miner für Stale Blocks (Uncles) entlohnt. (Jdebunt, 2017) Es besteht auch die Möglichkeit eines '**Orphaned**' Blocks. Das ist ein Block, über

3 Grundlagen

dessen Elternblöcke ein berechnender Knoten des Netzwerks keine Informationen hat und ihn deswegen nicht validieren kann.(Project, 2017b)

- Die Bestätigungszeit (engl. **confirmation time**) ist die Zeit, die zwischen einer Transaktion und ihrer Aufnahme in die Blockchain vergeht.(Kumar, 2017)
- Bei der **Hashrate** handelt es sich um die Anzahl der Hashes eines Netzwerk pro Sekunde. Trifft man auf den Begriff im Kontext des Bitcoins, handelt es sich meist um Terahashes, bei Ethereum um Gigahashes (entspricht $\frac{1}{1000}$ Terahash).(Kumar, 2017)
- Miningschwierigkeit (**difficulty**) meint die Schwierigkeit, einen validen Block zu finden. Als Referenz dient die einfachste Möglichkeit einen Block zu finden mit der Schwierigkeit 1.(Project, 2017a)
- Aufgebrachte Rechenleistung wird im Ethereumnetzwerk in **Gas** (engl. Aussprache) bemessen. Wie in Teil 1.2 bereits genannt, können im Netzwerk beispielsweise Smart-contracts ausgeführt werden. Die Kosten für die dort verbrauchte Leistung wird in Gas gezahlt. Wenn die Leistung vorher nicht abgemessen werden kann, wird ein Limit (**gas limit**) gesetzt, nach dessen Verbrauch eine Aktion abgebrochen wird.(Wood, 2014, S. 4)

Zum tieferen Verständnis können folgende Werke dienen:

3.4 Microsoft Azure ML Studio

Im nun folgenden Abschnitt der Arbeit wird das Microsoft Azure ML Studio in drei Schritten erklärt:

1. Zuerst erfolgt eine allgemeine Beschreibung des Machine Learning Studios.
2. Dann wird der Aufbau eines Projektes beschrieben und
3. zuletzt die vorgefertigten Module vorgestellt.

Es sollte in diesem Teil beachtet werden, dass die Quellen hauptsächlich Microsoft-eigen sind, da außer einigen Blockartikeln wenig externe Literatur oder Meinungen verfügbar sind.

3.4.1 Allgemeine Beschreibung

Zum Ende des Abschnittes 1.4 wurde die **Software as a Service** (SaaS) Microsoft Azure Machine Learning Studio bereits angesprochen. Es handelt sich dabei um eine Cloud-basierte Web-IDE (**I**ntegrated **D**evelopment **E**nvironment). Der Kern der Anwendung ist ein „drag and-drop tool“, das genutzt wird um „predictive analytics solutions“ zu entwerfen und bereitzustellen. (Ericson and Rohm, 2017c) Es können sowohl vorgefertigte Bausteine genutzt werden, als auch selbst entworfene Skripte (in R und/oder Python).

3.4.2 Aufbau

TODO: evtl. visualisieren!

Die strukturellen Hauptkomponenten sind Tabelle 3.8 zu entnehmen. Die Tabelle basiert auf dem Artikel „What is Azure Machine Learning Studio?“ von (Ericson and Rohm, 2017c) und dem Machine Learning Studio selbst (<https://studio.azureml.net/Home>).

Komponente	Beschreibung
Settings	Hier finden sich Einstellmöglichkeiten zur Benutzerverwaltung (für kollaborative Arbeit im Studio), das Preismodell (10GB Workspace Storage sind in der kostenfreien Version enthalten) und zu den Authorisierungstoken. Die Settings sind in diesem Kontext nicht weiter interessant.

3 Grundlagen

Notebooks	<p>Die Notebooks sind Verknüpfungen auf Jupyter Notebooks. Diese wiederum sind „Webanwendungen“, die genutzt werden um „Dokumente, die live code, Gleichungen, Virtualisierungen und erzählenden Text“ enthalten „zu erstellen und zu teilen“. Genutzt werden sie unter anderem für</p> <ul style="list-style-type: none"> • „data cleaning“, • „transformation, • numerical simulation, • statistical modeling, • data visualization“ und • „machine learning“(noa, 2017b). <p>Project Jupyter ist non-profit und open-source.(noa, 2014)</p>
Experiments	<p>Experimente sind das Kernelemente des Azure ML Studios. Sie stellen den Rahmen für das Erproben verschiedener Problemlösungen. Experimente sollten dabei mindestens ein Datenset (dazu gleich mehr) und ein Modul enthalten. Ist ein Experiment ausgereift, kann der Status von „training“ to „predictive“ geändert werden und als Web Service bereit gestellt werden (dazu ebenfalls nachfolgend Mehr).</p>
Datasets	<p>Für das Experimentieren benötigte Datensätze (z.B. .csv-, .tsv-, .arff-, .txt-, .zip-, oder .RData-Dateien) können in die Web-IDE geladen und als Datasets abgespeichert werden. So können die Dateien ohne erneutes Hochladen wiederverwendet werden. Eine andere Möglichkeit, Daten im Studio zu benutzen, ist das Experiment Item „Import Data“, das unter anderem auf Daten aus Azure Blob Storage, Azure Table Storage, Azure DocumentDB, Azure SQL Datenbanken, Hive Queries oder HTTP-Quellen zugreifen kann.</p>
Experiment Items	<p>Die Sammlung aller drag-and-drop Elemente im Studio werden Experiment Items genannt. Die umfassen Elemente für alle Prozessschritte, wie Input, Transformation, Selection, Machine Learning Module, Visualisierung, Scoring etc.</p>

3 Grundlagen

Trained Models	Fertigtrainierte Machine Learning Module können als eigenständige Komponenten genutzt werden. Ein Beispiel dafür wäre die Wiederverwendung in einem anderen Experiment mit neuen Daten oder der Vergleich zweier fertiger Modelle.
Web Services	Ist ein Experiment auf dem Status „predictive“, können die Input- und Output-Felder in Web Services umgewandelt werden, um sie in andere Anwendungen oder Systeme zu integrieren.
Projects	Projects dienen als Organisatorisches Strukturelement. Durch sie können Experimente, Datasets, Notebooks etc. gruppiert werden.

Tabelle 3.8: Hauptkomponenten des Azure Machine Learning Studios

4 Durchführung der Analyse

Nachdem in den vorhergegangenen Abschnitten alle Aspekte des Themas „Kursanalyse von Kryptowährungen mit Azure Machine Learning“ betrachtet wurden, widmet sich dieser Abschnitt der praktischen Umsetzung. Jeder Themenblock hat seinen Teil zum Erstellen eines Kontextes beigetragen, in dem die Analyse durchgeführt werden kann (siehe Tabelle 4.1).

Themenblock	Inhalt	Ziel
Data Mining Frameworks (3.1)	Beschreibung der bekanntesten Frameworks des Data Mining Prozesses und Auswahl eines Frameworks für die vorliegende Arbeit	Detailliertes Beschreiben des nachfolgenden Prozessmodells
Machine Learning (3.2)	Vorstellung einer Möglichkeit zur Einordnung von Machine Learning Typen und Algorithmen	Verständnisaufbau für die Analyse in diesem Teil der Arbeit
Kryptowährungen (3.3)	Hintergrundwissen zu Kryptowährungen	Für eine Analyse ist Hintergrundwissen wichtig. Dieses sogenannte „Domain“-Wissen ist essentieller Bestandteil einer Analyse mit CRISP-DM
Microsoft Azure Machine Learning Studio (3.4)	Allgemeine Beschreibung und Aufbau des Werkzeugs	Das Studio soll als Werkzeug zur Analyse eingesetzt werden.

Tabelle 4.1: Behandelte theoretische Abschnitte im Kontext der Arbeit

Wie in Punkt 3.1.4 angesprochen, wird als Hilfe für das Prozessmodell CRISP-DM der zugehörige User Guide (Chapman et al., 2000, S. 30-56) herangezogen. Die, im Guide genannten, Outputs jedes Prozessschritts werden nachfolgend speziell hervorgehoben. Das CRISP-DM

Prozessmodell ist sehr generisch gehalten. Dies ist beispielsweise daran zu erkennen, dass das Modell in vier übereinander liegende Abstraktionsschichten gegliedert ist. (Chapman et al., 2000, S. 6) Dies dient der Anpassungsfähigkeit an viele heterogene Projekte. Diese Anpassungsfähigkeit wird gleich im ersten Schritt genutzt.

4.1 Business Understanding

4.1.1 Determine the Business Objectives

In gewöhnlichen Industrie- oder Forschungsprojekten ist es wichtig, die Stakeholder (vor allem Geldgeber) und den Reifegrad und die Akzeptanz des Data Mining im Projektumfeld zu analysieren. Dies rückt im vorliegenden Fall in den Hintergrund. Die Anderen Outputs sind jedoch ebenso wichtig.

Output	Beschreibung
Background	Die Analyse wird im Rahmen einer Masterarbeit durchgeführt. Nur eine Person ist daran beteiligt.
Business objectives	Die Untersuchung hat zwei Hauptziele. Das eine ist die Analyse der Cryptowährungen an sich. Es soll herausgefunden werden, ob Kursschwankungen mit Hilfe der Isolation von Einflussfaktoren und den Mitteln des Machine Learning vorausgesagt werden können oder anderweitige Auffälligkeiten zu beobachten sind. Das andere Ziel ist die Einarbeitung in das Werkzeug Azure Machine Learning.
Business success criteria	Die Erkenntnis, dass eine Vorhersage nicht möglich ist, oder dass wichtige Einflüsse nicht gefunden wurden, ist durchaus möglich und bedeutet keinesfalls ein Scheitern des Projekts. Hinsichtlich des Werkzeugs Azure ML, ist es beispielsweise interessant, welchen Restriktionen das Tool unterlegen ist. Das Betrifft sowohl Funktionen, die (noch) nicht vorhanden sind, oder technische Limitationen, wie Geschwindigkeit, Volumenbegrenzungen etc..

Tabelle 4.2: Output des Schrittes „Determine the Business Objectives“

4.1.2 Assess the Situation

Dieser Teil befasst sich vor allem damit, welche Ressourcen zur Verfügung stehen (Hardware, Software, personell) und welche sonstigen Bedingungen erfüllt sein müssen oder das Projekt begrenzen. Dazu zählt auch das Finden von Daten, die für die Modellierung genutzt werden können. Anzumerken ist hierbei, dass es noch nicht um das tatsächliche Laden der Daten im Sinne von Dateien geht, sondern um das Finden von potentiellen Quellen für Daten. Zusätzlich sollen noch eine Risiko- und eine Kosten-Nutzen-Analyse durchgeführt werden. Das Hauptaugenmerk liegt jedoch auf der Erschließung der Daten.

An Stelle einer vollständigen Risikoanalyse (Kontext herstellen, Risiken identifizieren, analysieren, evaluieren, managen(Sowa, 2017, S. 43)), tritt eine Aufzählung der 3 Hauptrisiken. Dies geschieht einerseits aus Gründen der Verhältnismäßigkeit, andererseits liegt der Fokus der Arbeit auf einem anderen Thema. Ähnlich verhält es sich mit der Kosten-Nutzen-Analyse. Sie wird zur Bewertung der Wirtschaftlichkeit herangezogen, was in diesem Kontext nicht relevant ist. Deswegen wird auf sie vollständig verzichtet.

Die schwerste Aufgabe dieses Teils ist das Finden von Daten, die Einfluss auf den Kurs von Kryptowährungen haben (könnten). Forschungen in diesem Bereich identifizierten einerseits öffentliches Interesse (soziale Medien, Google Suchanfragen, etc.)(Kristoufek, 2013; Garcia et al., 2014) und andererseits auch wirtschaftliche Faktoren („standard economic theory“, also Angebot und Nachfrage, Investoren)(Kristoufek, 2015) als Haupteinflussfaktoren. Anhand den Aussagen dieser Paper und zusätzlichen Überlegungen ergeben sich folgende Faktoren, die bei der Analyse bedacht werden (Tabelle 4.3).

Darüber hinaus wird die Quelle aufgeführt, über die die Daten bezogen werden. Bei monetären Strömen oder Kursen, wird immer der Kurs in USD herangezogen.

Damit eine Analyse der Kryptowährungen möglich ist, müssen zu diesen historische Kursdaten beschafft werden. Dieser, auf den ersten Blick simpel wirkende Schritt, ist in Wirklichkeit **TODO: formulierung** nicht trivial. Kryptowährungen werden an dutzenden Portalen gleichzeitig gehandelt. Dabei erscheinen genauso schnell neue Börsen, wie alte verschwinden. Auch das Handelsvolumen und die Handelswährung unterscheidet sich. Hinzu kommt, dass an Bitcoin- oder Ethereum-Handelsportalen meist 24 Stunden täglich gehandelt werden kann. Aus diesem Grund sind solche Datenquellen ausgewählt worden, die entweder ihre Daten direkt von der Blockchain erhalten oder einen gewichteten Mittelwert über die größten Handelsplattformen berechnen.

Die nächste Schwierigkeit ist die Auswahl der Aktienindizes, die für die Analyse herangezogen werden. Es existiert keine allgemein gültige oder anerkannte Liste mit 'den wichtigsten Ak-

4 Durchführung der Analyse

„Index“ als „major world indices“ deklariert. (Limited, 2017) Es liegt dabei eine große Überschneidung mit anderen Stock Market-Seiten vor. (liveindex.org, 2017; Finance, 2017a; Finance, 2017b) Zusätzlich zu den Aktienindices wird ein Financial Stress Index (FSI) herangezogen. Ein solcher „index misst die aktuelle Belastung in einem finanzwirtschaftlichen System“ (Vermeulen et al., 2014, S. 1; eigene Übersetzung) In diesem Fall ist es der St. Louis Fed Financial Stress Index, der 18 Einzelfaktoren aus drei Kategorien bündelt. (of St. Louis, 2017)

Ferner werden die Währungen der acht größten Volkswirtschaften (Fund, 2017) und die Kurse für Gold, Silber und Rohöl mit einbezogen. Es werden die Wechselkurse zum Dollar betrachtet, sprich Fremdwährung/USD. Bei den Ölkursen wird Brent, das wichtigste Rohöl für den europäischen Markt (noa, 2016), und West Texas Intermediate (WTI), das Pondon für den US-Markt, betrachtet (noa, 2017c).

Cryptowährungs-eigene Faktoren

Daten	für BTC	für ETH
Handelsvolumen	ja, von https://bitcoincharts.com/	ja, von https://coinmarketcap.com/
Coin Volumen (Gesamtanzahl der vorhandenen Bitcoins/des Ethers)	ja, von https://blockchain.info/	ja, von https://etherscan.io/
Mining-Schwierigkeit TODO: Erklärung	ja, von https://data.bitcoinity.org/	ja, von https://etherscan.io/
Anzahl der Transaktionen	ja, von https://blockchain.info/	ja, von https://etherscan.io/
Hashrate	ja, von https://www.kaggle.com/	ja, von https://www.kaggle.com/
Marktkapitalisierung	ja, von https://www.kaggle.com/	ja, von https://www.kaggle.com/

Öffentliches Interesse

Daten	für BTC	für ETH
Google Websuchen	ja, von https://trends.google.de/trends/	ja, von https://trends.google.de/trends/
Google News-Suchen	ja, von https://trends.google.de/trends/	ja, von https://trends.google.de/trends/

4 Durchführung der Analyse

Wikipedia Seitenaufrufe	ja, von https://wikimedia.org/api/rest_v1/	ja, von https://wikimedia.org/api/rest_v1/
Tweets (Twitter Nachrichten)	nein, nicht kostenlos verfügbar	
Zeitungsartikel/-Überschriften	ja, von https://www.kaggle.com/therohk/million-headlines/data/	
Blogartikel	TODO: ???	TODO: ???
(Web-)Domains	TODO: ???	TODO: ???

(Aktien)indizes

Daten	für BTC	für ETH
Dow 30	ja, von https://www.investing.com/indices/	
S&P 500	ja, von https://www.investing.com/indices/	
Nasdaq	ja, von https://www.investing.com/indices/	
SmallCap	ja, von https://www.investing.com/indices/	
S&P 500 VIX	ja, von https://www.investing.com/indices/	
S&P/TSX	ja, von https://www.investing.com/indices/	
TR Canada	ja, von https://www.investing.com/indices/	
Bovespa	ja, von https://www.investing.com/indices/	
IPC	ja, von https://www.investing.com/indices/	
DAX	ja, von https://www.investing.com/indices/	
FTSE 100	ja, von https://www.investing.com/indices/	
CAC 40	ja, von https://www.investing.com/indices/	
Euro Stoxx	ja, von https://www.investing.com/indices/	
AEX	ja, von https://www.investing.com/indices/	
IBEX	ja, von https://www.investing.com/indices/	
FTSE MIB	ja, von https://www.investing.com/indices/	
SMI	ja, von https://www.investing.com/indices/	
PSI	ja, von https://www.investing.com/indices/	
BEL	ja, von https://www.investing.com/indices/	
ATX	ja, von https://www.investing.com/indices/	
OMXS30	ja, von https://www.investing.com/indices/	
OMXC20	ja, von https://www.investing.com/indices/	

4 Durchführung der Analyse

MICEX	ja, von https://www.investing.com/indices/
RTSI	ja, von https://www.investing.com/indices/
WIG20	ja, von https://www.investing.com/indices/
Budapest SE	ja, von https://www.investing.com/indices/
BIST 100	ja, von https://www.investing.com/indices/
TA 35	ja, von https://www.investing.com/indices/
Tadawul All Share	ja, von https://www.investing.com/indices/
Nikkei 225	ja, von https://www.investing.com/indices/
S&P/ASX 200	ja, von https://www.investing.com/indices/
DJ New Zealand	ja, von https://www.investing.com/indices/
Shanghai	ja, von https://www.investing.com/indices/
SZSE Component	ja, von https://www.investing.com/indices/
China A50	ja, von https://www.investing.com/indices/
DJ Shanghai	ja, von https://www.investing.com/indices/
Hang Seng	ja, von https://www.investing.com/indices/
Taiwan Weighted	ja, von https://www.investing.com/indices/
SET	ja, von https://www.investing.com/indices/
KOSPI	ja, von https://www.investing.com/indices/
IDX Composite	ja, von https://www.investing.com/indices/
Nifty	ja, von https://www.investing.com/indices/
BSE Sensex	ja, von https://www.investing.com/indices/
PSEi Composite	ja, von https://www.investing.com/indices/
STI Index	ja, von https://www.investing.com/indices/
Karachi	ja, von https://www.investing.com/indices/
HNX 30	ja, von https://www.investing.com/indices/
CSE All-Share	ja, von https://www.investing.com/indices/
St. Louis Fed Financial Stress Index (STLFSI)	ja, von https://fred.stlouisfed.org/series/STLFSI

Währungen der größten Volkswirtschaften (nach BIP)

Daten	für BTC	für ETH
China (CNY)	ja, von https://www.investing.com/currencies/single-currency-crosses	
Japan (JPY)	ja, von https://www.investing.com/currencies/single-currency-crosses	
Deutschland (EUR)	ja, von https://www.investing.com/currencies/single-currency-crosses	

4 Durchführung der Analyse

Großbritannien (GBP)	ja, von https://www.investing.com/currencies/single-currency-crosses
Frankreich (EUR)	ja, von https://www.investing.com/currencies/single-currency-crosses
Indien (INR)	ja, von https://www.investing.com/currencies/single-currency-crosses
Brasilien (BRL)	ja, von https://www.investing.com/currencies/single-currency-crosses

natürliche Ressourcen

Daten	für BTC	für ETH
Goldpreis	ja, von https://www.investing.com/commodities/	
Silberpreis	ja, von https://www.investing.com/commodities/	
Brent (Rohöl Europa)	ja, von https://www.investing.com/commodities/	
WTI (Rohöl USA)	ja, von https://www.investing.com/commodities/	

Tabelle 4.3: Mögliche Einflussfaktoren auf den Kurs von Kryptowährungen

Output	Beschreibung
---------------	---------------------

4 Durchführung der Analyse

Inventory of resources	<p>Personal</p> <ul style="list-style-type: none">• 1 Person mit Zugang zu den Recherche Ressourcen der Hochschule München (OPAC, DBIS, ZDB etc.(noa, 2017a)) <p>Hardware</p> <ul style="list-style-type: none">• 1 PC (CPU: AMD Ryzen 5 1600 Sechskern; RAM: 8GB; GPU: NVIDIA GeForce GTX 1060 (6GB VRAM); Windows 10 Education Build 15063.674) <p>Software</p> <ul style="list-style-type: none">• 1 „Free“-Account Microsoft Azure Machine Learning Studio mit Workspace in „South Central US“• Version des Jupyter Notebooks 5.1.0• Auf dem PC: R Version 3.4.1• Auf dem PC: RStudio Version 1.0.153• Excel 2016 (Microsoft Office 365 ProPlus) Version 1710• Notepad++ Version 7.5.1 <p>Daten</p> <ul style="list-style-type: none">• in Tabelle 4.3 genannte Daten• Kurse BTC/USD und ETH/USD• Zusätzliche Eigenschaften von Bitcoin und Ethereum. (?) stellt unter der 'CC0: Public Domain'-Lizenz einen Datensatz zur Verfügung, der besondere Eigenschaften der Währungen enthält. Beispiele für Bitcoin sind die „Anzahl der einzigartigen Adressen“ in der „Bitcoin Blockchain“; oder die „Anzahl der uncles pro Tag“(?, eigene Übersetzung) für Ethereum.
------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

4 Durchführung der Analyse

Requirements, assumptions, and constraints	<p>Zu Bedenken ist, dass bei einer kostenlosen Subscription im Azure ML Studio nur 10GB Storage verfügbar sind. Zusätzlich müssen Daten, die analysiert werden sollen, in das Tool geladen werden. Bei einem Upload vom PC wird das durch die Upload-Bandbreite limitiert. Eventuell müssen Daten im Projektverlauf auch öfter Hochgeladen werden, was zu Verzögerungen führen könnte. Außerdem lässt sich in der freien Version nur jeweils ein Experiment gleichzeitig ausführen. Das parallele Trainieren von Modellen ist somit nicht möglich.</p> <p>Obwohl zur Hilfe neben der Web-IDE auch RStudio genutzt werden kann, soll die Untersuchung hauptsächlich mit den Azure ML Studio Bordmitteln durchgeführt werden.</p> <p>Schließlich sollen nur solche Daten genutzt werden, die entweder frei verfügbar sind oder mit dem Hochschulzugang zu beschaffen sind. Das schließt präparierte Datensätze von Bezahlseiten aus.</p>
Risks and contingencies	<ul style="list-style-type: none"> • keinen Zugriff auf Azure ML Studio (Server-seitige Probleme; Ablauf der Free-Subscription) • benötigte Daten nicht verfügbar • TODO: todo • TODO: todo
Terminology	Als Glossar dient das Abkürzungsverzeichnis zu Beginn der Arbeit.
Costs and benefits	—

Tabelle 4.4: Output des Schrittes „Assess the Situation“

4.1.3 Determine the Data Mining Goals

Output	Beschreibung
Data mining goals	<p>Es sollen mit Hilfe von Azure Machine Learning Studio mehrere Machine Learning Models erzeugt werden. Sie sollen folgende Sachverhalte TODO: Wort vorhersagen:</p> <ul style="list-style-type: none"> • Steigt der Kurs innerhalb eines 24-Stunden Zyklus über das Hoch des vorhergehenden 24-Stunden Zyklus? (Mögliche Antworten sind <i>ja</i> := 1 und <i>nein</i> := 0) • Fällt der Kurs innerhalb eines 24-Stunden Zyklus unter das Tief des vorhergehenden 24-Stunden Zyklus? (Mögliche Antworten sind <i>ja</i> := 1 und <i>nein</i> := 0) • Wie hoch ist das Hoch des nächsten 24-Stunden Zyklus? (Antwort: Preis eines Bitcoins/Ethers in USD) • Wie niedrig ist das Tief des nächsten 24-Stunden Zyklus? (Antwort: Preis eines Bitcoins/Ethers in USD) <p>Bei den ersten beiden Fällen handelt es sich um eine two-class classification (siehe 3.2.1). Die letzten beiden Fragen fallen in das Gebiet der Regressionen (siehe 3.2.1).</p>

4 Durchführung der Analyse

Data mining success criteria	<p>Die Klassifikationen werden bewertet anhand:</p> <ul style="list-style-type: none"> • Accuracy: $\frac{t_p+t_n}{t_p+t_n+f_p+f_n}$ (Anteil der insgesamt richtig vorhergesagten Werte) • Precision: $\frac{t_p}{t_p+f_p}$ (Anteil der richtig vorhergesagten true-Werte) • Recall: $\frac{t_p}{t_p+f_n}$ (Anteil der richtig vorhergesagten true-Werte an allen eigentlich richtigen true-Werten) • F1-score: $2 \times \frac{Precision \times Recall}{Precision + Recall}$ (Harmonisches Mittel aus Precision und Recall; Bester Wert = 1) • AUC: Fläche unter der Receiver Operating Characteristics (ROC) Kurve; „Je besser die Klassifizierungsfähigkeit des Klassifikators desto höher ist der AUC-Wert“ (Lohninger, 2013, ROC-Kurve) <p> t_p : true positive t_n : true negative f_p : false positive f_n : false negative </p>
------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

4 Durchführung der Analyse

	<p>Die Regressionen werden bewertet anhand:</p> <ul style="list-style-type: none">• Mean absolute error (MAE): Unterschied von vorhergesagten und tatsächlichen Werten; je kleiner desto besser(Mircosoft, 2017)• Root mean squared error (RMSE): Macht eine Aussage darüber, „wie gut eine Funktionskurve an vorliegende Daten angepasst ist“; „je größer der RMSE [...], desto schlechter“(Statista)• Relative absolute error (RAE), Relative squared error (RSE): Ähnlich dem RMSE, aber für den Vergleich von Regressionen mit unterschiedlichen Maßeinheiten geeignet.(Dr. Sayad, 2017) Im vorliegenden Fall können die anderen Größen herangezogen werden, da alle Modelle die gleiche Einheit für die vorhergesagten Werte nutzen.• Coefficient of determination: Gütemaß für die Aussagekraft der Regression; <i>Regression passt perfekt</i> := 1 <i>Regression erklärt nichts</i> := 0; kleine Werte sind normal, große sollten misstrauisch machen(Mircosoft, 2017)
--	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Tabelle 4.5: Output des Schrittes „Determine the Data Mining Goals“

TODO: welche werte bedeuten was

TODO: bessere beschreibungen

TODO: was zu den quellen sagen

4.1.4 Produce a Project Plan

Nach diesem sehr Daten-orientierten Teilschritt, befasst sich der nachfolgende Schritt einerseits mit der (groben) Planung des weiteren Projekts und andererseits mit einer anfänglichen Betrachtung der eingesetzten Werkzeuge. Grundsätzlich orientiert sich der Projektplan am

4 Durchführung der Analyse

Referenzmodell CRISP-DM. In einem komplexeren Projekt mit mehreren kollaborativ arbeitenden Personen, kommt diesem Schritt eine größere Rolle zu als im vorliegenden Fall. Wichtige Meilensteine, die zu bestimmten Zeitpunkten fertig vorliegen müssen, um einen unterbrechungsfreien Projektverlauf zu gewährleisten, stellen die Outputs jedes Prozessschrittes dar. Eine genauere zeitliche Einschätzung ist dem Output „Project Plan“ in Tabelle 4.7 zu entnehmen. Zu sehen ist dort, die geschätzte Arbeitszeit in Tagen für jeden Schritt und der Anteil am Gesamtprojekt. Anzumerken ist hier, dass die Angaben durch Rückschritte in frühere Phasen verzerrt werden können.

Bei der initialen Betrachtung der Werkzeuge werden laut Prozessmodell verschiedene alternativen gegeneinander abgewogen und ihr Zweck im Projekt festgelegt. Da das Werkzeug Azure Machine Learning Studio bereits vorgegeben ist, wird hier nur der Zweck der Werkzeuge im Projektverlauf betrachtet. Wird das Werkzeug R (bzw. RStudio) eingesetzt, so erhält es den Vorzug gegenüber seiner Konkurrenz (z.b. Python) aufgrund der Erfahrung des Entwicklers.

Prozessschritt	geschätzte Zeit in Tagen	Prozentualer Anteil
Collect initial data	2	20%
Describe data	2	20%
Explore data	2	20%
Verify data quality	2	20%
Data understanding gesamt	$\sum 2$	$\sum 2\%$
Select data	2	20%
Clean data	2	20%
Construct data	2	20%
Integrate data	2	20%
Format data	2	20%
Data preparation ge- samt	$\sum 2$	$\sum 2\%$
Select modeling techni- que	2	20%
Generate test design	2	20%
Build model	2	20%
Assess model	2	20%

4 Durchführung der Analyse

Modeling gesamt	$\Sigma 2$	$\Sigma 2\%$
Evaluate results	2	20%
Review process	2	20%
Determine next steps	2	20%
Evaluation gesamt	$\Sigma 2$	$\Sigma 2\%$
Plan deployment	2	20%
Plan monitoring and maintenance	2	20%
Produce final report	2	20%
Review project	2	20%

Tabelle 4.6: Output „Project Plan“ des Schrittes „Produce a Project Plan“

Werkzeug	Zweck im Projekt
Microsoft Azure Machine Learning Studio	Vorgabe im Projekt; soll in jeder Phase genutzt werden, wo es möglich ist
Jupyter Notebooks	Integriert in Azure ML Studio; für R Code, der in der Cloud ausgeführt werden soll
R	zur Nutzung in Situationen für die Azure ML keine vorgefertigten Experiment Items bereitstellt
RStudio	zum lokalen entwerfen von Scripten vor der Ausführung in der Cloud und zur Vorbereitung der Datensätze
Excel	zur Betrachtung, Vorbereitung und Konvertierung von Daten vor dem Upload in die Cloud
Notepad++	zur Betrachtung, Vorbereitung und Konvertierung von Daten vor dem Upload in die Cloud

Tabelle 4.7: Output „Initial assessment of tools and techniques“ des Schrittes „Produce a Project Plan“

4.2 Data Understanding

4.2.1 Collect the Initial Data

In Schritt 4.1.2 wurden Einflussfaktoren auf den Kurs festgehalten. Zusätzlich wurden bereits Quellen für Daten zu diesen Faktoren gesucht. In diesem Schritt werden nun die Daten tatsächlich (im Sinne von 'echten' Dateien) bezogen und bereits in Azure Machine Learning geladen. Ebenfalls Teil dieses Schrittes ist es, die Daten grob zu beschreiben. Im vorliegenden Fall wird die Datenspanne und weitere offensichtliche Merkmale beschrieben.

Die Herkunft der Daten ist nicht weiter beschrieben, da dies bereits in 4.1.2 abgehandelt wurde.

Alle gesammelten Daten enthalten Header-Informationen (Preis, Kurs, Datum etc.). Azure ML Studio kann diese Header auslesen und nutzt sie für einfache Select- oder Join-Befehle. Damit dies möglich ist, müssen die Daten als 'Generic CSV File with a header (.csv)' oder 'Generic TSV File with a header (.tsv)' vorliegen. In manchen Fällen ist es möglich, dass Microsoft Excel beim Speichern einer Datei als '.csv' statt Kommas, Strichpunkte zum Trennen nutzt. Hier ist darauf zu achten, dass tatsächlich Kommas als Trennzeichen genutzt werden, da Azure ML Studio diese sonst nicht interpretieren kann. Die Datei-Endung '.tsv' ist nicht zwingend erforderlich. Ein 'Tab separated values'-File kann auch mit der Endung '.txt' hochgeladen und richtig interpretiert werden. Dies ist wichtig, da Excel als Endung für diese Dateien nicht '.tsv' sondern '.txt' wählt.

Azure ML Studio bietet die Möglichkeit, mehrere Dateien gezippt in das Tool zu laden. Dies bietet sich an, wenn viele einzelne Dateien hochgeladen werden sollen. Leider kann das Zip-File pro Experiment nur einmal verwendet werden und nur eine Datei daraus kann entpackt werden. Dadurch ist diese Option im Falle dieser Arbeit nicht praktikabel. Es müssen also alle Dateien einzeln geladen werden.

Cryptowährungs-eigene Faktoren		
Datensatz	Reichweite	Besonderheiten

4 Durchführung der Analyse

BTC __Total __Volume __Daily __Full	3.1.2009 bis 5.11.2017	teilweise kein Anstieg des Volumens im Datensatz; große Lücken im Datensatz (2012-2016)
BTC __Difficulty __Daily __Full	6.11.2012 bis 5.11.2017	Zeit in UTC
BTC __Transaction __Num- ber __Fully __Daily	3.1.2009 bis 5.11.2017	
BTC __Price __Multiple __Daily	17.7.2010 bis 5.11.2017	Zeit in UTC; sehr Lückenhaft, da erst ab 2016 alle enthaltenen Börsen operieren
ETH __Total __Volume __Daily __Full	30.7.2015 bis 5.11.2017	Zeit in UTC und UnixTime-Stamp
ETH __Difficulty __Daily __Full	30.7.2015 bis 5.11.2017	Zeit in UTC und UnixTime-Stamp
ETH __Transaction __Num- ber __Fully __Daily	30.7.2015 bis 5.11.2017	Zeit in UTC und UnixTime-Stamp; ganz am Anfang einige 0-Werte
Öffentliches Interesse		
Datensatz	Reichweite	Besonderheiten
google __Trends __BTC __Websearch	01.2009 bis 11.2017	Daten im Abstand von 1 Monat; auf einer Scala von 0-100; erster nicht-null-Wert bei 05.2011
google __Trends __ETH __Websearch	01.2011 bis 11.2017	Daten im Abstand von 1 Monat; auf einer Scala von 0-100; erster nicht-null-Wert bei 08.2014
google __Trends __BTC __Newssearch	01.2009 bis 11.2017	Daten im Abstand von 1 Monat; auf einer Scala von 0-100; erster nicht-null-Wert bei 04.2011

4 Durchführung der Analyse

google __Trends __ETH __Newssearch	01.2011 bis 11.2017	Daten im Abstand von 1 Monat; auf einer Scala von 0-100; erster nicht-null-Wert bei 07.2014
Wiki __Page __Views __BTC	1.7.2015 bis 7.11.2017	Enthält Datum als zusammengesetzte Zahl im Format yyyyddmm00; enthält 7 Features, von denen 5 für jede Zeile gleich sind
Wiki __Page __Views __ETH	1.7.2015 bis 7.11.2017	Enthält Datum als zusammengesetzte Zahl im Format yyyyddmm00; enthält 7 Features, von denen 5 für jede Zeile gleich sind
abcnews __Date __Text	19.2.2003 bis 30.9.2017	großer Datensatz; viel Text; kann von Excel nicht komplett geöffnet werden
(Aktien)indizes		
Datensatz	Reichweite	Besonderheiten
AEX, BFX, XU100, BVSP, VIX, CSE, GDAXI, DJI, FTSE, FTMIB, HSI, IBEX, MXX, JKSE, KSE, KS11, MCX, IXIC, NSEI, N225, OMXC20, OMXS30, IRTS, SPX, AXJO, GSPTSE, SSEC, SSMI, TA35, TASI, TRX50CAP, US2000, WIG20	1.1.2009 bis 9.11.2017	alle Indices enthalten Lücken für Wochenenden und Feiertage; Datum in einem Format mit Text (!); gilt auch für nachfolgende Datensätze mit anderen Zeitspannen
ATX (ATX) __history.txt	23.3.2015 bis 9.11.2017	verhältnismäßig kurze Zeitspanne
BSE Sensex 30 (BSESN) __history.txt	24.2.2011 bis 9.11.2017	

4 Durchführung der Analyse

Budapest SE (BUX) __history.txt	7.3.2011 bis 9.11.2017	
Dow Jones New Zealand (NZDOW) __history.txt	25.8.2011 bis 9.11.2017	
Dow Jones Shanghai (DJSH) __history.txt	6.3.2011 bis 9.11.2017	
Euro Stoxx 50 (STOXX50E) __history.txt	15.8.2011 bis 9.11.2017	
FTSE China A50 (FTXIN9) __history.txt	19.3.2010 bis 9.11.2017	
FTSE Straits Times Singapore (STI) __history.txt	7.3.2011 bis 9.11.2017	
HNX 30 (HNX30) __history.txt	4.11.2014 bis 9.11.2017	verhältnismäßig kurze Zeitspanne
PSEi Composite (PSI) __history.txt	3.11.2011 bis 9.11.2017	
PSI 20 (PSI20) __history.txt	25.4.2010 bis 9.11.2017	
SET Index (SETI) __history.txt	18.3.2011 bis 9.11.2017	
SZSE Component (SZSC1) __history.txt	14.9.2012 bis 9.11.2017	
Taiwan Weighted (TWII) __history.txt	17.3.2011 bis 9.11.2017	
STLFSI __history.csv	2.1.2009 bis 27.10.2017	Zeilen im Abstand einer Woche; Index kann positive und negative Werte annehmen
Währungen der größten Volkswirtschaften		
Datensatz	Reichweite	Besonderheiten
CNY __USD __history	1.1.2009 bis 13.11.2017	Datum in einem Format mit Text (!); Lücken an Wochenenden und Feiertagen

4 Durchführung der Analyse

JPY _USD _history	12.5.2009 bis 13.11.2017	Datum in einem Format mit Text (!); Lücken an Wochenenden und Feiertagen
EUR _USD _history	1.1.2009 bis 13.11.2017	Datum in einem Format mit Text (!); Lücken an Wochenenden und Feiertagen
GBP _USD _history	1.1.2009 bis 13.11.2017	Datum in einem Format mit Text (!); Lücken an Wochenenden und Feiertagen
INR _USD _history	09.09.2014 bis 13.11.2017	Datum in einem Format mit Text (!); Fehlende Daten vor dem 09.09.2014 Lücken an Wochenenden und Feiertagen
BRL _USD _history	1.1.2009 bis 13.11.2017	Datum in einem Format mit Text (!); Lücken an Wochenenden und Feiertagen
natürliche Ressourcen		
Datensatz	Reichweite	Besonderheiten
gold _history	1.1.2009 bis 13.11.2017	Datum in einem Format mit Text (!); Lücken an Wochenenden und Feiertagen
silver _history	1.1.2009 bis 13.11.2017	Datum in einem Format mit Text (!); Lücken an Wochenenden und Feiertagen
oil _brent _history	1.1.2009 bis 13.11.2017	Datum in einem Format mit Text (!); Lücken an Wochenenden und Feiertagen
oil _wti _history	1.1.2009 bis 13.11.2017	Datum in einem Format mit Text (!); Lücken an Wochenenden und Feiertagen
ETH/USD-Kurs		
Datensatz	Reichweite	Besonderheiten

4 Durchführung der Analyse

ETH _Price _Volume _Full _Daily	7.8.2015 bis 5.11.2017	historische Kursdaten für ETC (Preis und Volumen); Datum in einem Format mit Text (!); chronologisch absteigend sortiert (ältestes Datum zum Schluss)
BTC/USD-Kurs		
Datensatz	Reichweite	Besonderheiten
BTC _Price _Volume _Full _Daily	13.9.2011 bis 6.11.2017	historische Kursdaten für BTC (Preis und Volumen); am Anfang einige Lücken
zusätzliche Eigenschaften		
Datensatz	Reichweite	Besonderheiten
bitcoinDataset	6.10.2009 bis 30.10.2017	sehr detaillierte Bitcoin Eigenschaften; 24 Features, z.B. Anzahl der einzigartigen Adressen im Netzwerk oder Kosten pro Transaktion
ethereumDataset	30.7.2015 bis 3.10.2017	sehr detaillierte Ethereum Eigenschaften; 18 Features, z.B. Anzahl der Adressen im Ethereum-Netzwerk oder Anzahl der Blocks und Uncles

Tabelle 4.8: Output „Initial data collection report“ des Schrittes „Collect the Initial Data“

Wie in Tabelle 4.8 zu sehen ist, umfassen die Daten verschiedene Zeitspannen. Einige Aktienindices, vor allem ATX und HNX30, existieren noch nicht so lange, wie die historischen Daten des Bitcoins. Es gibt mehrere Wege, um damit umzugehen. Da auch ohne diese noch über 30 andere Indices zur Verfügung stehen, fließen sie nicht weiter in die Untersuchung ein. Einige der Datensätze beinhalten nicht nur rohe Daten, sondern bereits Brechungen, wie den prozentualen Anstieg eines Kurses zum Vortag. Obwohl die Features des Datensatzes erst im nachfolgenden Schritt beschrieben werden, kann hier schon festgehalten werden,

4 Durchführung der Analyse

dass diese 'Zusatzinformationen' von der Analyse ausgeschlossen werden. Es handelt sich dabei lediglich um eine - für den menschlichen Betrachter einfach zu verstehende - andere Schreibweise für die Daten in einer Zeitreihe und nicht um zusätzliche Informationen.

4.2.2 Describe the Data

Eine detaillierte Beschreibung der oben genannten Datensätze erfolgt nun. Für jeden Satz, bzw. jede homogene Gruppe (wie die Aktienindices), werden die Anzahl der Spalten (Features), die Anzahl der Reihen (Observations) und die Dateigröße (in Kilobyte; KB) angegeben. Außerdem werden für die Features genauer erläutert. Beim Datentyp 'numerisch' handelt es sich um eine Ganzzahl, bei 'numerisch (5)' um eine Gleitkommazahl mit bis zu fünf Nachkommastellen.

Datensatz	BTC _Total _Volume _Daily _Full	
Observations	1615	
Features	2	
Dateigröße	32	
Feature	Datentyp	Besonderheit
Date	Datum im Format 'dd/mm/yyyy HH:MM'	HH:MM unbenutzt
Volume	numerisch (1)	extrem große Lücke zwischen 20.9.2012 und 9.7.2016; eventuell nicht zu gebrauchen; zwischen 50.0 und 16665662.5

Tabelle 4.9: Data description report für BTC _Total _Volume _Daily _Full

Datensatz	BTC _Difficulty _Daily _Full	
Observations	1826	
Features	2	
Dateigröße	78	
Feature	Datentyp	Besonderheit

4 Durchführung der Analyse

Date	Datum im Format 'yyyy-mm-dd HH:MM:SS UTC'	HH:MM:SS unbenutzt
Difficulty	numerisch (2)	Inkonsistenzen in der Difficulty, teilweise um Faktor 100 unterschiedliche Werte in aufeinanderfolgenden Zeilen; manche mit Nachkommastellen, manche ohne

Tabelle 4.10: Data description report für BTC _Difficulty _Daily _Full

Datensatz	BTC _Transaction _Number _Fully _Daily	
Observations	1615	
Features	2	
Dateigröße	46	
Feature	Datentyp	Besonderheit
Date	Datum im Format 'dd/mm/yy HH:MM'	HH:MM:SS unbenutzt
Transactions	numerisch (0)	

Tabelle 4.11: Data description report für BTC _Transaction _Number _Fully _Daily

Datensatz	BTC _Price _Multiple _Daily	
Observations	2669	
Features	11	
Dateigröße	279	
Feature	Datentyp	Besonderheit
Time	Datum im Format 'dd/mm/yy HH:MM'	HH:MM:SS unbenutzt
bit-x	numerisch (Gleitkommazahl)	Wert eines Bitcoins in USD an dieser Börse; mit Lücken
bitbay	numerisch (Gleitkommazahl)	Wert eines Bitcoins in USD an dieser Börse; mit Lücken

4 Durchführung der Analyse

cex.io	numerisch (Gleitkom- mazahl)	Wert eines Bitcoins in USD an dieser Börse; mit Lücken
coinbase	numerisch (Gleitkom- mazahl)	Wert eines Bitcoins in USD an dieser Börse; mit Lücken
exmo	numerisch (Gleitkom- mazahl)	Wert eines Bitcoins in USD an dieser Börse; mit Lücken
gemini	numerisch (Gleitkom- mazahl)	Wert eines Bitcoins in USD an dieser Börse; mit Lücken
hitbtc	numerisch (Gleitkom- mazahl)	Wert eines Bitcoins in USD an dieser Börse; mit Lücken
itbit	numerisch (Gleitkom- mazahl)	Wert eines Bitcoins in USD an dieser Börse; mit Lücken
kraken	numerisch (Gleitkom- mazahl)	Wert eines Bitcoins in USD an dieser Börse; mit Lücken
others	numerisch (Gleitkom- mazahl)	Wert eines Bitcoins in USD an anderen Börsen; ohne Lücken

Tabelle 4.12: Data description report für BTC _Price _Multiple _Daily

Datensatz	ETH _Total _Volume _Daily _Full	
Observations	830	
Features	3	
Dateigröße	33	
Feature	Datentyp	Besonderheit
Date(UTC)	Datum im Format 'm/dd/yyyy'	
UnixTimeStamp	numerisch (0)	zusätzliches Datum als Unix Timestamp
Value	numerisch (0)	von 7204930659375 (min) bis 9554710634375 (max)

Tabelle 4.13: Data description report für ETH _Total _Volume _Daily _Full

4 Durchführung der Analyse

Datensatz	ETH _Difficulty _Daily _Full	
Observations	830	
Features	3	
Dateigröße	32	
Feature	Datentyp	Besonderheit
Date(UTC)	Datum im Format 'm/dd/yyyy'	
UnixTimeStamp	numerisch (0)	zusätzliches Datum als Unix Timestamp
Value	numerisch (0)	

Tabelle 4.14: Data description report für ETH _Difficulty _Daily _Full

Datensatz	ETH _Transaction _Number _Fully _Daily	
Observations	830	
Features	3	
Dateigröße	31	
Feature	Datentyp	Besonderheit
Date(UTC)	Datum im Format 'm/dd/yyyy'	
UnixTimeStamp	numerisch (0)	zusätzliches Datum als Unix Timestamp
Value	numerisch (0)	

Tabelle 4.15: Data description report für ETH _Transaction _Number _Fully _Daily

Datensatz	google _Trends _BTC _Newssearch und google _Trends _BTC _Web	
Observations	107	
Features	2	
Dateigröße	2	
Feature	Datentyp	Besonderheit
Monat	Datum im Format 'yyyy-mm'	keine Stelle für 'Tag'

4 Durchführung der Analyse

bitcoin: (Weltweit)	numerisch (0)	Scala von 0 bis 100
---------------------	---------------	---------------------

Tabelle 4.16: Data description report für google __Trends __BTC __Newssearch und google __Trends __BTC __Websearch

Datensatz	google __Trends __ETH __Newssearch und google __Trends __ETH __Websearch	
Observations	83	
Features	2	
Dateigröße	1	
Feature	Datentyp	Besonderheit
Monat	Datum im Format 'yyyy-mm'	keine Stelle für 'Tag'
Ethereum: (Weltweit)	numerisch (0)	Scala von 0 bis 100

Tabelle 4.17: Data description report für google __Trends __ETH __Newssearch und google __Trends __ETH __Websearch

Datensatz	Wiki __Page __Views __BTC	
Observations	861	
Features	7	
Dateigröße	57	
Feature	Datentyp	Besonderheit
project	text	immer 'en.wikipedia'
article	text	immer 'Bitcoin'
granularity	text	immer 'daily'
timestamp	Datum im Format 'yyyymmddhh'	hh immer 00
access	text	immer 'all-access'
agents	text	immer 'all-agents'
views	numerisch (0)	

Tabelle 4.18: Data description report für Wiki __Page __Views __BTC

Datensatz	Wiki __Page __Views __ETH	
Observations	861	

4 Durchführung der Analyse

Features	7	
Dateigröße	57	
Feature	Datentyp	Besonderheit
project	text	immer 'en.wikipedia'
article	text	immer 'Ethereum'
granularity	text	immer 'daily'
timestamp	Datum im Format 'yyyymmdd00'	hh immer 00
access	text	immer 'all-access'
agents	text	immer 'all-agents'
views	numerisch (0)	

Tabelle 4.19: Data description report für Wiki _Page _Views _ETH

Datensatz	abcnews _Date _Text	
Observations	mehr als 1048576 (Excel Maximum)	
Features	2	
Dateigröße	53480	
Feature	Datentyp	Besonderheit
publish _date	Datum im Format 'yyyymmdd'	mehrere (hundert) Einträge für einen Tag
headline _text	Text	ohne Satzzeichen; alles in Kleinbuchstaben

Tabelle 4.20: Data description report für abcnews _Date _Text

Datensatz	alle Aktienindices	
Observations	unterschiedlich; bei vollständigen (1.1.2009 bis 5.11.2017) ca. 2230	
Features	7	
Dateigröße	bei vollständigen (1.1.2009 bis 5.11.2017) 253 bis 102	
Feature	Datentyp	Besonderheit

4 Durchführung der Analyse

Date	Datum im Format 'mmm dd, yyyy' (mmm als Text)	Monat als Text (z.B. 'Jan 02, 2009')
Price	numerisch (2)	Nachkommastellen durch Punkt abgetrennt; Tausen- der durch Komma
Open	numerisch (2)	Nachkommastellen durch Punkt abgetrennt; Tausen- der durch Komma
High	numerisch (2)	Nachkommastellen durch Punkt abgetrennt; Tausen- der durch Komma
Low	numerisch (2)	Nachkommastellen durch Punkt abgetrennt; Tausen- der durch Komma
Vol.	Mischform aus Zahl und Text	Werte mit 'K' für Tausend (Kilo), 'M' für Millionen (engl. Millions) und 'B' für Milliarden (engl. Billions); Fehlende Werte mit Strich (-) gekennzeichnet
Change %	Prozentzahl mit Proz- entzeichen (%)	negative und positive Werte

Tabelle 4.21: Data description report für alle Aktienindices

Datensatz	STLFSI _history	
Observations	461	
Features	2	
Dateigröße	9	
Feature	Datentyp	Besonderheit
DATE	Datum im Format 'dd/mm/yyyy'	in wöchentlichem Abstand

4 Durchführung der Analyse

STLFSI	numerisch (3)	im Intervall von [-1,586;3,246]
--------	---------------	---------------------------------

Tabelle 4.22: Data description report für STLFSI _history

Datensatz	alle Währungen	
Observations	unterschiedlich; INR: 996, BRL: 2314, JPY: 2615, CNY: 2314, GBP: 2339, EUR: 2323	
Features	6	
Dateigröße	INR: 51; BRL, CNY: 110; EUR, GBP: 111; JPY: 134	
Feature	Datentyp	Besonderheit
Date	Datum im Format 'mmm dd, yyyy' (mmm als Text)	Monat als Text (z.B. 'Jan 02, 2009')
Price	numerisch (4)	Nachkommastellen durch Punkt abgetrennt
Open	numerisch (4)	Nachkommastellen durch Punkt abgetrennt
High	numerisch (4)	Nachkommastellen durch Punkt abgetrennt
Low	numerisch (4)	Nachkommastellen durch Punkt abgetrennt
Change %	Prozentzahl mit Prozentzeichen (%)	negative und positive Werte

Tabelle 4.23: Data description report für alle Währungen

Datensatz	alle natürlichen Ressourcen	
Observations	unterschiedlich; Gold: 2289, Silber: 2692 Brent: 2290, WTI: 2283	
Features	7	
Dateigröße	unterschiedlich; Gold: 139, Silber: 142, Brent: 121, WTI: 118	
Feature	Datentyp	Besonderheit

4 Durchführung der Analyse

Date	Datum im Format 'mmm dd, yyyy' (mmm als Text)	Monat als Text (z.B. 'Jan 02, 2009')
Price	numerisch (2)	Nachkommastellen durch Punkt abgetrennt
Open	numerisch (2)	Nachkommastellen durch Punkt abgetrennt
High	numerisch (2)	Nachkommastellen durch Punkt abgetrennt
Low	numerisch (2)	Nachkommastellen durch Punkt abgetrennt
Change %	Prozentzahl mit Proz- entzeichen (%)	negative und positive Werte

Tabelle 4.24: Data description report für alle natürlichen Ressourcen

Datensatz	ETH _Price _Volume _Full _Daily	
Observations	822	
Features	7	
Dateigröße	51	
Feature	Datentyp	Besonderheit
Date	Datum im Format 'mmm dd, yyyy' (mmm als Text)	Monat als Text (z.B. 'Jan 02, 2009')
Open	numerisch (2)	Nachkommastellen durch Punkt abgetrennt
High	numerisch (2)	Nachkommastellen durch Punkt abgetrennt
Low	numerisch (2)	Nachkommastellen durch Punkt abgetrennt
Close	numerisch (2)	Nachkommastellen durch Punkt abgetrennt
Volume	numerisch (0)	Kommas zwischen Tausen- dern

4 Durchführung der Analyse

Market Cap	numerisch (0)	Kommas zwischen Tausendern
------------	---------------	----------------------------

Tabelle 4.25: Data description report für ETH _Price _Volume _Full _Daily

Datensatz	BTC _Price _Volume _Full _Daily	
Observations	2247	
Features	8	
Dateigröße	148	
Feature	Datentyp	Besonderheit
Date	Datum im Format 'dd/mm/yyyy HH:MM'	HH:MM ungenutzt
Open	numerisch (2)	Nachkommastellen durch Punkt abgetrennt
High	numerisch (2)	Nachkommastellen durch Punkt abgetrennt
Low	numerisch (2)	Nachkommastellen durch Punkt abgetrennt
Close	numerisch (2)	Nachkommastellen durch Punkt abgetrennt
Volume (BTC)	numerisch (0)	Kommas zwischen Tausendern
Volume (Currency)	numerisch (0)	Kommas zwischen Tausendern
Weighted Price	numerisch (0)	Kommas zwischen Tausendern

Tabelle 4.26: Data description report für BTC _Price _Volume _Full _Daily

Datensatz	bitcoinDataset
Observations	2920
Features	24
Dateigröße	728

4 Durchführung der Analyse

Feature	Datentyp	Besonderheit
Date	Datum im Format 'dd/mm/yyyy HH:MM'	HH:MM ungenutzt
btc __market __price	numerisch (9)	
btc __total __bitcoins	numerisch (1)	nur Werte mit '.0' oder '.5' am Ende
btc __market __cap	numerisch (5)	
btc __trade __volume	numerisch (4)	Zu Beginn immer 0; enthält Lücken
btc __blocks __size	numerisch (4)	Zu Beginn immer 0; Nachkommastellen erst ab 20.4.2016
btc __avg __block __size	numerisch (16?)	Bis auf einen Wert (2.10.2017) immer < 0
btc __n __orphaned __blocks	numerisch (0)	Nur Werte im Intervall [0;7]
btc __n __transactions __per __block	numerisch (8)	Nachkommastellen erst ab 20.4.2016
btc __median __confirmation __time	numerisch (11)	Manche Werte scheinen Perioden darzustellen (7,866666666667 oder 7,933333333333)
btc __hash __rate	numerisch (17)	Große Spanne von $2.04e-6$ bis $1.0e12$
btc __difficulty	numerisch (11)	maximal 12 Stellen (Vorkommastellen + Nachkommastellen = 12)
btc __miners __revenue	numerisch (4)	
btc __transaction __fees	numerisch (8)	
btc __cost __per __transaction __percent	numerisch (11)	Werte werden nach unten (zeitlich später) kleiner

4 Durchführung der Analyse

btc _cost _per _transacti- on	numerisch (11)	maximal 12 Stellen (Vorkom- mastellen + Nachkommas- tellen = 12)
btc _n _unique _addresses	numerisch (0)	
btc _n _transactions	numerisch (0)	
btc _n _transactions _total	numerisch (0)	Werte kumulativ (stetig stei- gend)
btc _n _transactions _ex- cluding _popular	numerisch (0)	
btc _n _transactions _ex- cluding _chains _longer _than _100	numerisch (0)	Bis zum 18.4.2010 identisch mit btc _n _transactions _excluding _popular
btc _output _volume	numerisch (7)	viele Anfangswerte sind glat- te Zehner
btc _estimated _transacti- on _volume	numerisch (6)	Nachkommastellen erst ab 20.4.2016
btc _estimated _transacti- on _volume _usd	numerisch (4)	Nachkommastellen erst ab 20.4.2016

Tabelle 4.27: Data description report für bitcoinDataset

Datensatz	ethereumDataset	
Observations	797	
Features	18	
Dateigröße	115	
Feature	Datentyp	Besonderheit
Date(UTC)	Datum im Format 'mm/dd/yyyy'	
UnixTimeStamp	Datum im Format Un- ixTimeStamp	
eth _etherprice	numerisch (2)	bbb
eth _tx	numerisch (0)	
eth _address	numerisch (0)	Werte stetig steigend
eth _supply	numerisch (4)	Werte stetig steigend

4 Durchführung der Analyse

eth __marketcap	numerisch (9)	
eth __hashrate	numerisch (4)	
eth __difficulty	numerisch (3)	die ersten drei Werte sind < 0 , was der Definition der Schwierigkeit widerspricht
eth __blocks	numerisch (0)	immer vierstellig
eth __uncles	numerisch (0)	
eth __blocksize	numerisch (0)	
eth __blocktime	numerisch (2)	Nur Werte im Intervall [4,46;30,31]
eth __gasprice	numerisch (0)	
eth __gaslimit	numerisch (0)	vermutlich stufenhafter Anstieg/Abfall
eth __gasused	numerisch (0)	
eth __ethersupply	numerisch (5)	
eth __ens __register	numerisch (0)	von 30.7.2015 bis 3.5.2017 keine Daten; von 4.5.2017 bis 8.5.2017 immer Nullwerte

Tabelle 4.28: Data description report für ethereumDataset

Obwohl die Qualität der Daten erst in einem nachfolgenden Schritt genauer betrachtet wird, können nach den Data description reports schon potentielle Stolperfallen **TODO: darf man das so schreiben** identifiziert werden. Diesen Herausforderungen muss sich angenommen werden:

- Die Formate des Datums sind unterschiedlich. Auch sind für einige Datensätze sieben Observations pro Woche verfügbar (z.B. für den Bitcoin-Kurs), für Andere nur 5 (z.B. für den Dow Jones Industrial Average). Daraus ergibt sich das Problem, wie die Daten am besten aneinander gesetzt (gejoint) werden.
- In den Daten befinden sich Lücken. Obwohl diese in der Regel im Verhältnis zum gesamten Datensatz klein sind, gibt es auch größere Lücken (z.B. die Anzahl der Registrierungen beim Ethereum Name Service pro Tag).

4 Durchführung der Analyse

- Es tauchen Attribute doppelt auf. So beinhaltet sowohl der Datensatz BTC _Difficulty _Daily _Full das Schwierigkeitsmaß für das Minen, als auch der Datensatz bitcoinDataset.
- Ein besonderen Fall stellt abcnews _Date _Text dar. Hierbei handelt es sich einerseits um einen Datensatz, der zu groß ist, um komplett in einem Programm geöffnet zu werden, andererseits beinhaltet er Informationen (Überschriften) als 'Klartext'.
TODO: Klartext
- Am Ende von Abschnitt 4.2.1 wurde bereits angesprochen, dass das Feature 'Change' (z.B. bei den Aktienindices) keinen Mehrwert für die Analyse bietet.

4.2.3 Explore the Data

Der User Guide empfiehlt, um die bekannten Daten weiter zu untersuchen, „Abfrage-, Visualisierungs- und Reportingtechniken anzuwenden“ (Chapman et al., 2000, S. 40; eigene Übersetzung). Es ist sicherlich aufschlussreich, alle Features auf einen Graphen mit dem Kryptowährungskurs zu plotten. Da dies jedoch sehr aufwendig ist, werden hier nur einige Daten betrachtet. Außerdem ist es wahrscheinlich, dass komplexere Zusammenhänge nicht trivial visuell erkennbar sind.

Betrachtet man das Interesse am Bitcoin (Google Websuchen und Google Newssuchen) mit dem R Statement in Listing 4.1, so zeigt sich, dass der erste Ausschlag Mitte 2011 zu erkennen ist und der Nächste erst Anfang 2013 (Abbildung 4.1).

```
1 # Read BTC datasets and convert dates
2 google_Trends_BTC_Websearch.csv <-
  read.csv("google_Trends_BTC_Websearch.csv");
3 google_Trends_BTC_Websearch.csv[,1] <-
  as.Date(google_Trends_BTC_Websearch.csv[,1])
4 google_Trends_BTC_Newssearch.csv <-
  read.csv("google_Trends_BTC_Newssearch.csv");
5 google_Trends_BTC_Newssearch.csv[,1] <-
  as.Date(google_Trends_BTC_Newssearch.csv[,1])
6 attach(google_Trends_BTC_Websearch.csv)
7 attach(google_Trends_BTC_Newssearch.csv)
8
9 #Plot the data in line graphs
10 ggplot(google_Trends_BTC_Websearch.csv) +
11   geom_line(data = google_Trends_BTC_Websearch.csv, aes(x = Monat, y =
     bitcoin...Weltweit., color = "BTC_Websearch")) +
```

4 Durchführung der Analyse

```
12 | geom_line(data = google_Trends_BTC_Newssearch.csv, aes(x = Monat, y =  
13 |   bitcoin...Weltweit., color = "BTC_Newssearch")) +  
14 | xlab('Date') +  
15 | ylab('GoogleSearchIndex')+  
16 | scale_colour_manual("",  
17 |   breaks = c("BTC_Websearch", "BTC_Newssearch"),  
   values = c("green", "blue"))
```

Listing 4.1: Google Websuchen und Newssuchen für „Bitcoin“ im zeitlichen Verlauf in R

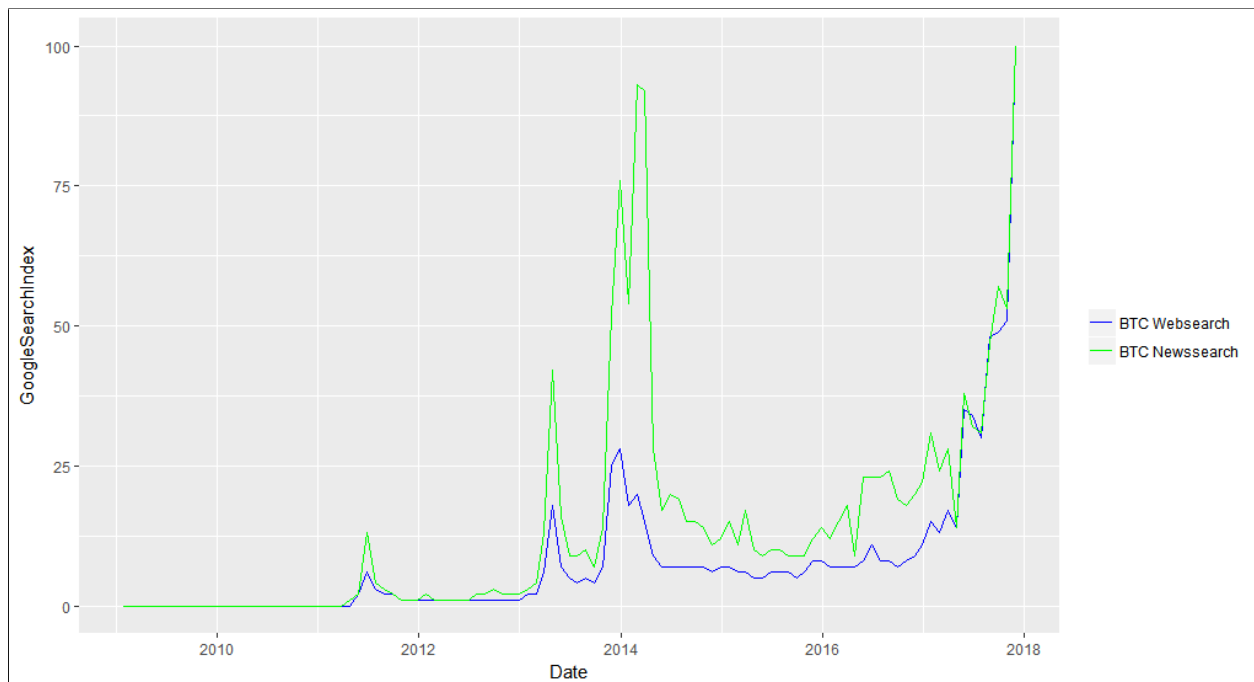


Abbildung 4.1: Google Websuchen und Newssuchen für „Bitcoin“ im zeitlichen Verlauf

Interpretiert man die Googlesuchen als Indikator für den Bekanntheitsgrad des Bitcoins, so lässt sich daraus schließen, dass die Währung bis 2011 sehr unbekannt und erst nach 2013 wirklich bekannt war. Zieht man nun den Bitcoinkurs heran (Listing 4.2), so lässt sich ein Zusammenhang erkennen (Abbildung 4.2).

```
1 | #Read historical BTC price data and convert date  
2 | btcPrice.csv <- read.csv2("btcPrice.csv");  
3 | btcPrice.csv[,1] <- as.Date(btcPrice.csv[,1], "%d/%m/%Y")  
4 | attach(btcPrice.csv)  
5 |  
6 | #Plot the data  
7 | ggplot(google_Trends_BTC_Websearch.csv) +  
8 |   geom_line(data = google_Trends_BTC_Websearch.csv, aes(x = Monat, y =  
   bitcoin...Weltweit., color = "BTC_Websearch")) +
```


4 Durchführung der Analyse

```
9 | geom_line(data = google_Trends_BTC_Newssearch.csv, aes(x = Monat, y =  
    bitcoin...Weltweit., color = "BTC_Newssearch")) +  
10 | geom_line(data = btcPrice.csv, aes(x = i..Date, y = btc_market_price/50,  
    color = "Bitcoinkurs")) +  
11 | xlab('Date') +  
12 | ylab('GoogleSearchIndex')+  
13 | scale_colour_manual("",  
14 |     breaks = c("BTC_Websearch", "BTC_Newssearch",  
    "Bitcoinkurs"),  
15 |     values = c("red", "green", "blue"))
```

Listing 4.2: Google Websuchen und Newssuchen für „Bitcoin“ und Bitcoinkurs im zeitlichen Verlauf in R

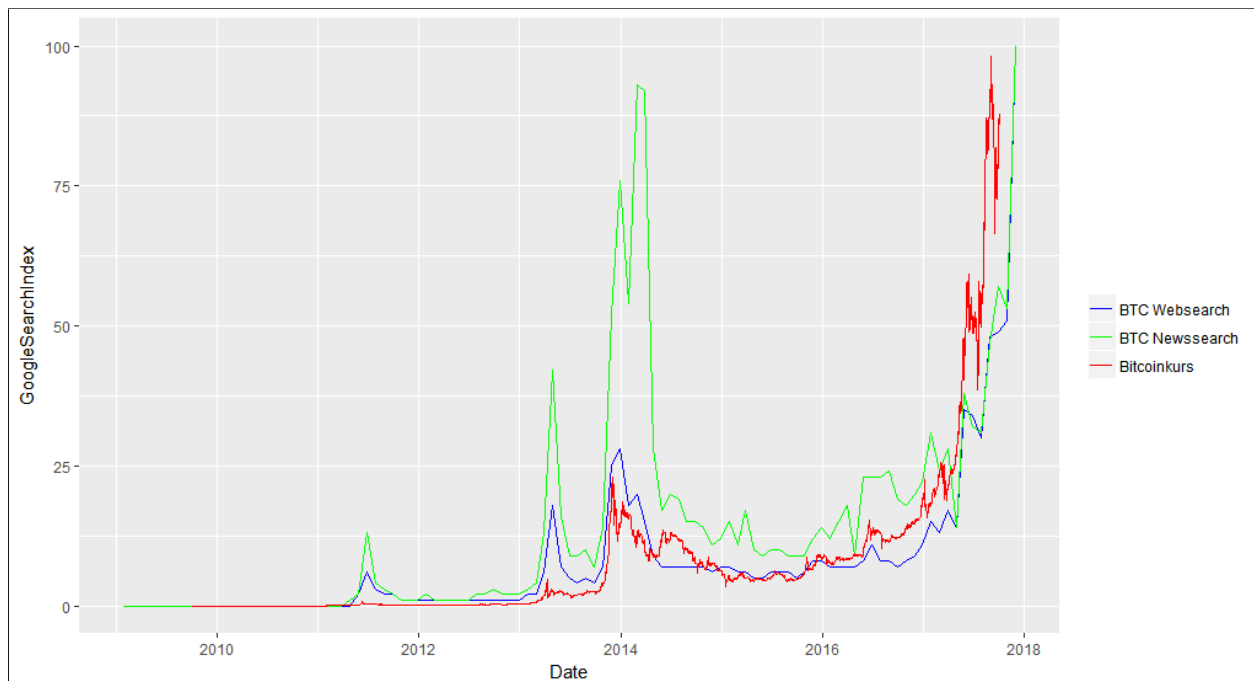


Abbildung 4.2: Google Websuchen und Newssuchen für „Bitcoin“ im zeitlichen Verlauf

Anzumerken ist, dass die historischen Kursdaten hier in keinem Verhältnis zur y-Achse stehen und nur der grafischen Darstellung dienen. Anders verhält es sich bei Ethereum (siehe Abbildung 4.3). Hier besteht schon vor dem offiziellen Start des Ethereumnetzwerks (Tual, 2015) ein gewisses Interesse.

4 Durchführung der Analyse

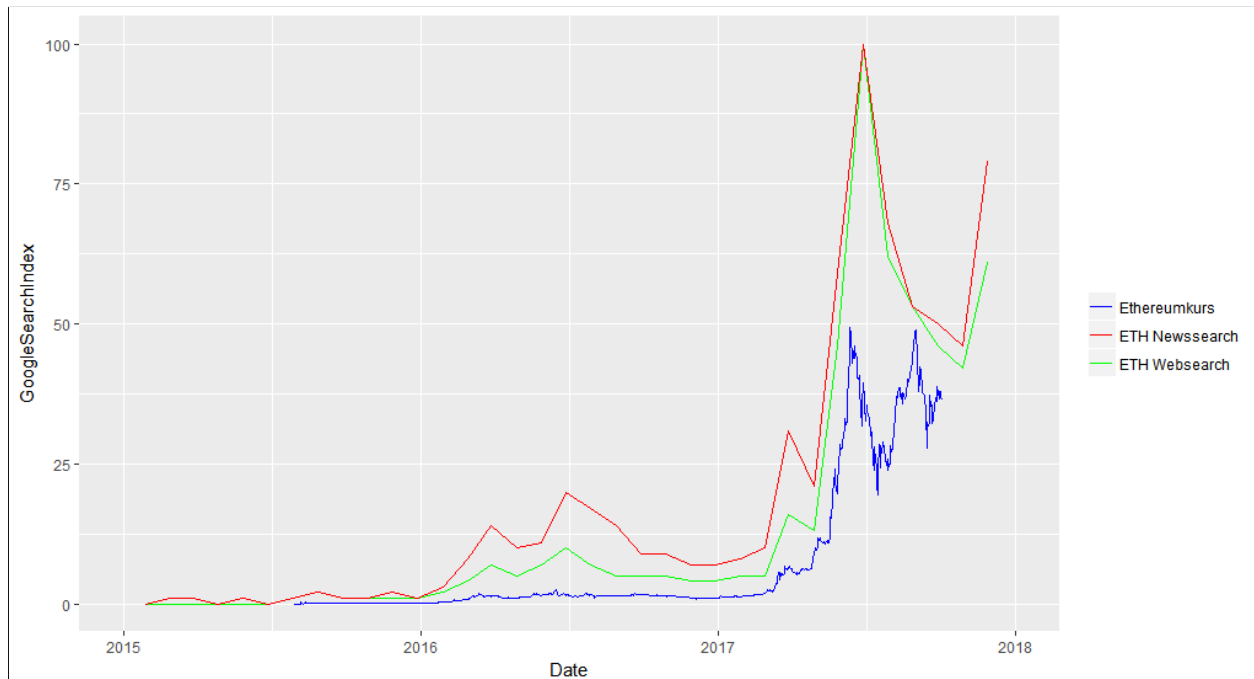


Abbildung 4.3: Google Websuchen und Newssuchen für „Ethereum“ im zeitlichen Verlauf

Als Schlussfolgerung lässt sich daraus ableiten, dass es sehr unwahrscheinlich ist, dass Indikatoren wie Aktienindices oder der Ölpreis den Kryptowährungskurs beeinflussen, wenn sie noch unbekannt ist. Deswegen empfiehlt sich für die Analyse der Zeitraum, seit die Währungen Bekanntheit erlangt haben.

Output	Beschreibung
Data exploration report	Analyse des Bitcoinurses erst ab 1.1.2011; des Etherumkurses ab 30.7.2015

Tabelle 4.29: Output des Schrittes „Explore the Data“

4.2.4 Verify Data Quality

Bei der Sicherung der Datenqualität fällt auf, dass die Daten oberflächlich eine hohe Qualität aufweisen. Beispielsweise sind alle Datensätze in USD angegeben und strukturiert. Sie enthalten pro Datei nicht viele Features und besitzen kaum lückenhafte Spalten. Jedoch sind einige Dinge zu beachten (siehe Tabelle 4.30).

4 Durchführung der Analyse

Problem	Lösung
Die verschiedenen Datensätze haben unterschiedliche Datumsformate. Dies birgt Risiken beim Zusammenführen der Daten.	Vor dem Joinen der Daten müssten die Formate uniformiert werden. Dies kann beispielsweise mit der R-Methode <pre> as.Date(data\$DateColumn, format='%Y%m%d')</pre> geschehen, solange das Datum in einem gültigen Format ist.
Beim Zusammensetzen der Datensätze entstehen Lücken, da einige Datensätze (z.B. Kryptowährungs-eigene Eigenschaften) sieben Observations pro Woche festhalten, Andere (z.B. Aktienindices) nur fünf. Außerdem gibt es Datensätze mit Feiertagen (z.B. Neujahr) oder solche mit nur einer Observation pro Woche (z.B. STLF-SI).	Azure ML Studio bietet das Experiment Item 'Clean Missing Data' an. Neben den Möglichkeiten, fehlende Daten mit dem Modus, Median oder Mittelwert auszufüllen, kann das Verfahren MICE(Azur et al., 2011) oder die Probabilistic PCA(Tipping and Bishop, 1999) genutzt werden.
Durch die unterschiedlichen Formatierungen gibt es inkonsistente Separatoren (Komma, Semikolon, Tab), Dezimaltrennzeichen (Punkt, Komma) und Tausendertrennzeichen (Punkte, ohne Trennzeichen).	Vor dem Zusammenführen müssen die Separatoren und Trennzeichen untersucht und eventuell umgeändert werden. Dies kann mit Excel ('Speichern als...' oder Notepad++ ('Find and Replace')) geschehen. Nach dem Joinen muss nachgeprüft werden.
Die Schwierigkeit für das Ethereuming wird im Datensatz ethereumDataset anfangs mit kleiner als 1 angegeben. Dies ist per Definition unmöglich.	Entweder es handelt sich hier um einen undokumentierte Sonderfall zu Beginn des Netzwerks oder es ist ein Fehler im Datensatz. Trotz dieser Unstimmigkeit, ist die Tatsache insignifikant und kann vernachlässigt werden.

Tabelle 4.30: Data quality report des Schrittes „Verify data quality“

hier aufzählen alles und bei select dann reduzieren und bei clean dann lücken etc. bei construct dann bis zu azure zusammenbauen

4.3 Data Preperation

4.3.1 Select Data

Die Datensätze sind nun gut beschrieben und es wurde auf Herausforderungen und Probleme hingewiesen. Die folgenden Punkte befassen sich damit, die richtigen Daten und Features weiter zu reduzieren („Select Data“), sie zu reinigen („Clean Data“) und schließlich zum endgültigen Analyse-Datensatz zusammenfassen („Construct Data“, „Integrate Data“ und „Format Data“). Ausgenommen der jetzt begründet ausgeschlossenen Daten, werden alle Vorgestellten zum bilden des Models genutzt. Tabelle 4.32 liefert einen Überblick.

In Punkt 4.2.1 wurde festgestellt, dass der Datensatz 'BTC _Total _Volume _Daily _Full' große Lücken enthält. Darüber hinaus enthält der Datensatz 'bitcoinDataset' ebenfalls die Anzahl der Bitcoins als Feature 'btc _total _bitcoins'. Ein stichprobenhafter Konsistenzcheck ergibt, dass die vorhanden Daten sich decken. Aus diesem Grund wird der Lückenhafte Datensatz exkludiert. 'BTC _Difficulty _Daily _Full' enthält wie 'bitcoinDataset' die Miningschwierigkeit (siehe Punkt 3.3). Obwohl der einzelne Datensatz minimal genauere Werte enthält, fällt die Wahl erneut auf das 'bitcoinDataset', da es Daten über eine längere Zeitspanne enthält. Genauso verhält sich bei 'BTC _Transaction _Number _Fully _Daily' und 'BTC _Price _Multiple _Daily'. Analog wird bei den Ethereumdatensätzen vorgegangen. 'ETH _Total _Volume _Daily _Full' enthält die gleichen Daten wie 'ethereumDataset'. Allerdings um Faktor 100000 erhöht. Vermutlich handelt es sich hierbei um eienn Konvertierungs- oder Kopierfehler, da das derzeit mögliche Maximum bei 100 Millionen Ether liegt.(Buterin, 2016; Hawksby-Robinson, 2017) Die Werte für die Schwierigkeit und Anzahl der Transaktionen stimmen bei 'ETH _Difficuly _Daily _Full' bzw. 'ETH _Transaction _Number _Fully' mit 'ethereumDataset' genau überein. Auch wenn die Datensätze an dieser Stelle aussortiert werden, war ihre Beschreibung keine Verschwendung. Sie hat dazu beigetragen, das Domainwissen (siehe ?? **TODO: anfügen**) zu vertiefen und die Plausibilität der Daten zu prüfen. Der Bitcoinkurs soll vom 1.1.2011 an analysiert werden. Für einige Aktienindices liegen für diese Zeit noch keine Daten vor. Um die über 50 Indices etwas auszudünnen, werden nur solche für das Machine Learning genutzt, für die Daten vorhanden sind. Dadurch wird

4 Durchführung der Analyse

versucht, Datenlücken zu vermeiden. Damit der Arbeitsaufwand reduziert wird, wird die Auswahl für die Analyse des Etherumpreises übernommen. Die Google Trends Daten werden beibehalten. Anders verhält es sich bei den Wikipedia-Seitenaufrufen. Für die Bitcoinanalyse ist die Zeitspanne des Datensatzes zu kurz (es fehlen 4 1/2 Jahre: von 1.1.2011 bis 1.7.2015). Eine mögliche Korrelation (siehe Abbildung 4.4) zwischen Seitenaufrufen und Bitcoinkurs müsste gesondert untersucht werden.

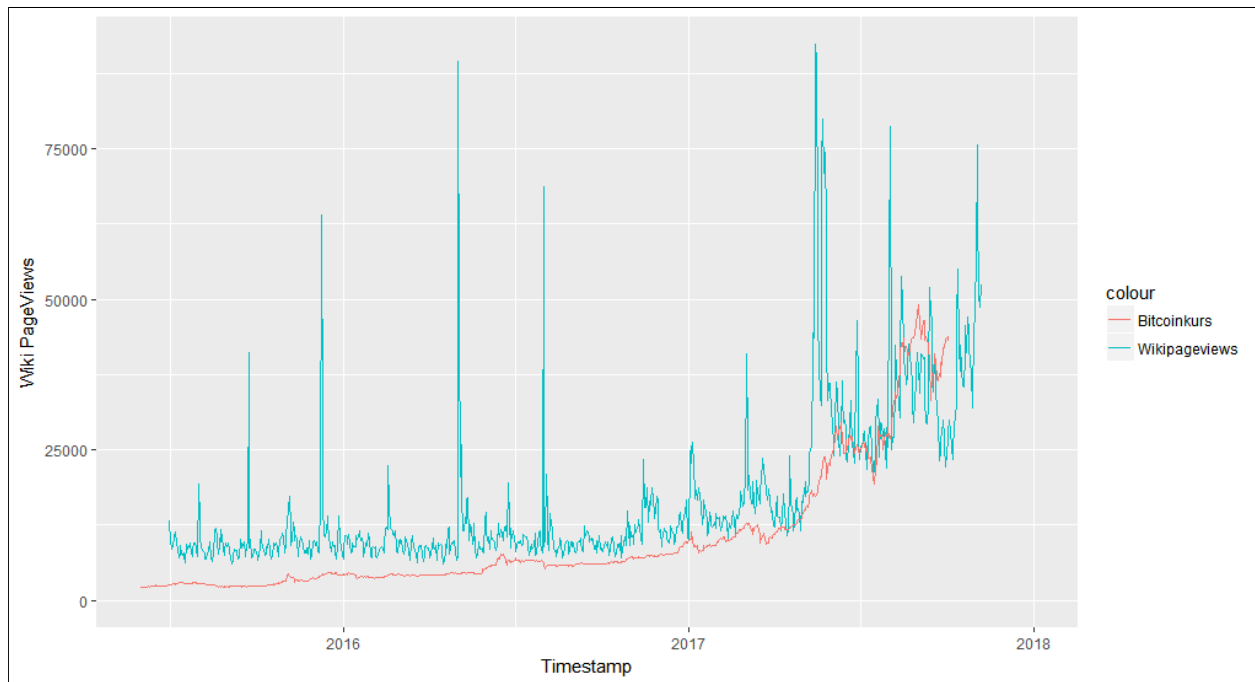


Abbildung 4.4: Wikipedia Seitenaufrufe „Bitcoin“ und der Bitcoinkurs im zeitlichen Verlauf

Für die Etherumkursanalyse werden die Seitenaufrufe herangezogen, da sich die Daten zeitlich decken (siehe Abbildung 4.5). Erneut ist bei beiden Diagrammen der Kryptowährungskurs nur relativ dargestellt und nicht in absoluten Zahlen. **TODO: formulierung**

4 Durchführung der Analyse

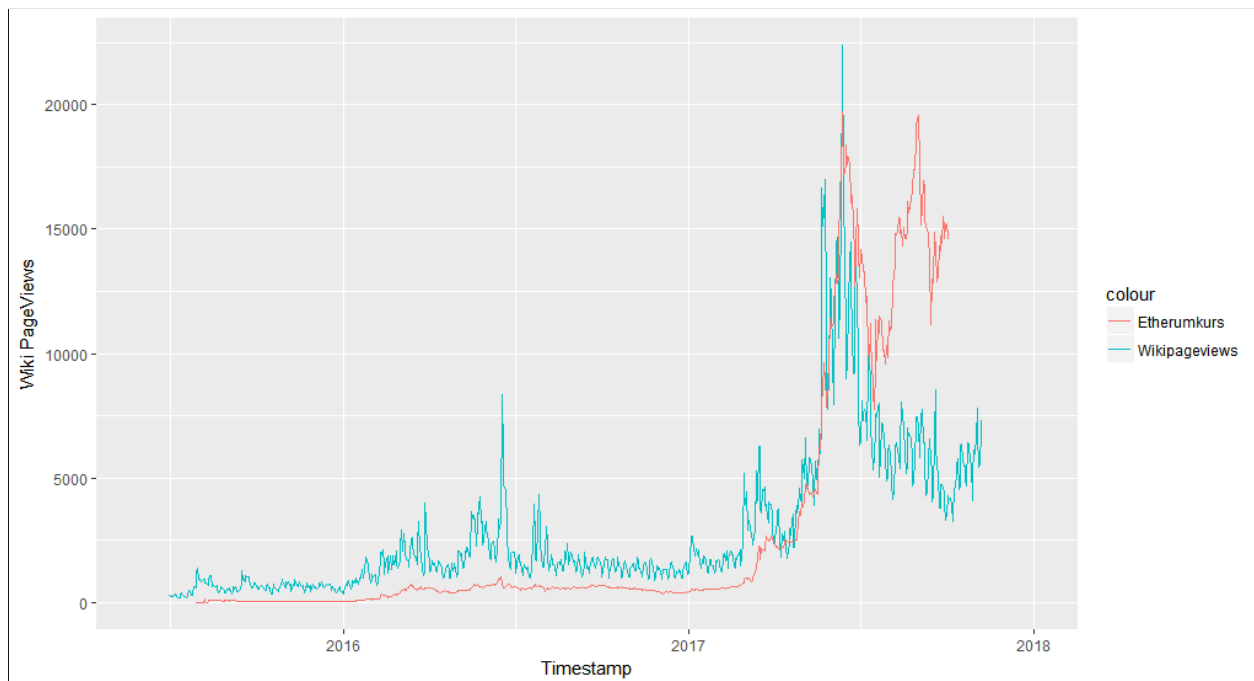


Abbildung 4.5: Wikipedia Seitenaufrufe „Ethereum“ und der Etherumkurs im zeitlichen Verlauf

Obwohl für den STLFSI nur wöchentlich - nicht täglich - berechnet wird, wird er beibehalten. Bei den Währungen wird nur der Kurs INR/USD exkludiert, da hier nur Daten ab dem 9.9.2014 vorliegen. Alle Daten zu natürlichen Ressourcen (Gold, Silber, Öl) werden inkludiert. Dem Datensatz 'ETH _Price _Volume _Full _Daily' fehlt die erste Woche an Daten, ist sonst aber reicher an Informationen über den Ethereum/USD-Kurs als das 'ethereumDataset', da es nicht nur den Durchschnittspreis eines Tages enthält, sondern sowohl den ersten und letzten Kurs in einem 24-Stunden-Fenster, als auch den Höchsten und den Niedrigsten. Dem 'BTC _Price _Volume _Full _Daily' hingegen fehlt das erste halbe Jahr an Daten und weist erst ab Ende (18.12.) 2011 keine Lücken mehr auf. In allen Datensätzen (Aktienindices, Währungen, natürliche Ressourcen) wird das Feature '%Change' gestrichen, da in 4.2.2 festgestellt wurde, dass es redundant ist. abcnews _Date _Text wird beinhalten.

Datensatz	Inkludiert	Exkludiert
BTC _Total _Volume _Daily _Full		X
BTC _Difficulty _Daily _Full		X
BTC _Transaction _Number _Fully _Daily		X
BTC _Price _Multiple _Daily		X

4 Durchführung der Analyse

ETH _Total _Volume _Daily _Full		X
ETH _Difficuly _Daily _Full		X
ETH _Transaction _Number _Fully _Daily		X
google _Trends _BTC _Websearch	X	
google _Trends _ETH _Websearch	X	
google _Trends _BTC _Newssearch	X	
google _Trends _ETH _Newssearch	X	
Wiki _Page _Views _BTC		X
Wiki _Page _Views _ETH	X	
abcnews _Date _Text	X	
FTXIN9, PSI20, AEX, BFX, XU100, BVSP, VIX, CSE, GDA- XI, DJI, FTSE, FTMIB, HSI, IBEX, MXX, JKSE, KSE, KS11, MCX, IXIC, NSEI, N225, OMXC20, OMXS30, IRTS, SPX, AXJO, GSPTSE, SSEC, SSMI, TA35, TASI, TRX50CAP, US2000, WIG20	X	
ATX, BSESN, BUX, NZDOW, DJSH, STOXX50E, STI, HNX30, PSI, SETI, SZSC1, TWII		X
STLFSI	X	
CNY _USD _history	X	
JPY _USD _history	X	
EUR _USD _history	X	
GBP _USD _history	X	
INR _USD _history		X
BRL _USD _history	X	
gold _history	X	
silver _history	X	
oil _brent _history	X	
oil _wti _history	X	
ETH _Price _Volume _Full _Daily	X	
BTC _Price _Volume _Full _Daily		X
bitcoinDataset	X	
ethereumDataset	X	

Tabelle 4.31: Inkludierte und exkludierte Datensätze für die Analyse

TODO: Textdingens: abc

4.3.2 Clean, Construct, Integrate, Format Data

Die Schritte 'Clean Data', 'Construct Data', 'Integrate Data' und 'Format Data' gehen Hand in Hand. Deswegen werden die Punkte zusammengefasst. Bei der tatsächlichen Durchführung wurden die Schritte iterativ und inkrementell durchlaufen.

Die Menge aller Daten für die Untersuchung ist heterogen. Es existieren jedoch Blöcke (z.B. alle Währungskurse oder alle natürlichen Ressourcen), die innerhalb eine homogene Struktur aufweisen. Aus diesem Grund ist es ratsam, zuerst die homogenen Daten zusammenzufassen und diese großen Blöcke dann zusammenzufügen. Durch dieses Bottom-Up-Zusammensetzen wird die Integration vereinfacht. Das Vorgehen ist dabei folgendes:

- Die einzelnen Datensätze werden so bearbeitet, dass sie für die Zusammenführung in ihren Block bereit sind.
- Die bearbeiteten Daten werden ihrem Datenblock hinzugefügt. Das Ergebnis wird untersucht und eventuelle Unreinheiten beseitigt.
- Alle Blöcke werden in einen abschließenden Datensatz integriert. Dieser wird wiederum inspiziert und bereinigt.

Nachfolgend wird das Bereinigen und Zusammensetzen beschrieben. Zuerst werden die Aktienindices bearbeitet (Listing 4.4) und gejoint (Abbildung 4.6).

```

1 | #clear all environment data
2 | rm(list=ls())
3 |
4 | #use anytime for date conversions
5 | library(anytime)
6 |
7 | #Function to format the data and rename columns
8 | processDataSet <- function(fileName){
9 |
10 |   #Read Data as csv
11 |   dataSet <- read.csv2(fileName)
12 |
13 |   #Name for columns
14 |   name <- basename(fileName)
15 |   dateName <- paste(name, "Date")
16 |   priceName <- paste(name, "Price")
17 |   openName <- paste(name, "Open")
18 |   highName <- paste(name, "High")

```


4 Durchführung der Analyse

```
19 lowName <- paste(name, "Low")
20 volumeName <- paste(name, "Volume")
21 colNames <- c(dateName, priceName, openName, highName, lowName, volumeName)
22
23 #attach and rename columns
24 colnames(dataSet)[c(1, 2, 3, 4, 5, 6)] <- colNames
25
26 #select what to keep --> drop column Change
27 keep <- colNames
28 dataSet <- dataSet[,keep]
29
30 #remove all thousand-seperator-commas
31 dataSet[,priceName] <- gsub(",", "", dataSet[,priceName])
32 dataSet[,openName] <- gsub(",", "", dataSet[,openName])
33 dataSet[,highName] <- gsub(",", "", dataSet[,highName])
34 dataSet[,lowName] <- gsub(",", "", dataSet[,lowName])
35
36 #change type Factor to correct one
37 dataSet[,dateName] <- anytime(dataSet[,dateName])
38 dataSet[,priceName] <- as.numeric(as.character(dataSet[,priceName]))
39 dataSet[,openName] <- as.numeric(as.character(dataSet[,openName]))
40 dataSet[,highName] <- as.numeric(as.character(dataSet[,highName]))
41 dataSet[,lowName] <- as.numeric(as.character(dataSet[,lowName]))
42 dataSet[,volumeName] <- pseudoNumeric.as.numeric(dataSet[,volumeName])
43
44 #missing values to NA (missing Data)
45 dataSet[, volumeName][dataSet[, volumeName] == "-"] <- NA
46
47 #save as csv; not csv2, because azure ML studio can't read semicolons as
seperators
48 write.csv(dataSet, file = paste(name, "_processed.csv"), append = FALSE,
49           quote = TRUE, sep = ",",
50           eol = "\n", na = "", dec = ".", row.names = TRUE,
51           col.names = TRUE, qmethod = c("escape", "double"),
52           fileEncoding = "")
53
54 #get all files to convert
55 files <- list.files(path=".", pattern="*.csv", full.names=T, recursive=FALSE)
56
57 #convert
58 lapply(files, processDataSet)
```

Listing 4.4: Aufbereiten der Indices-Datensätze

4 Durchführung der Analyse

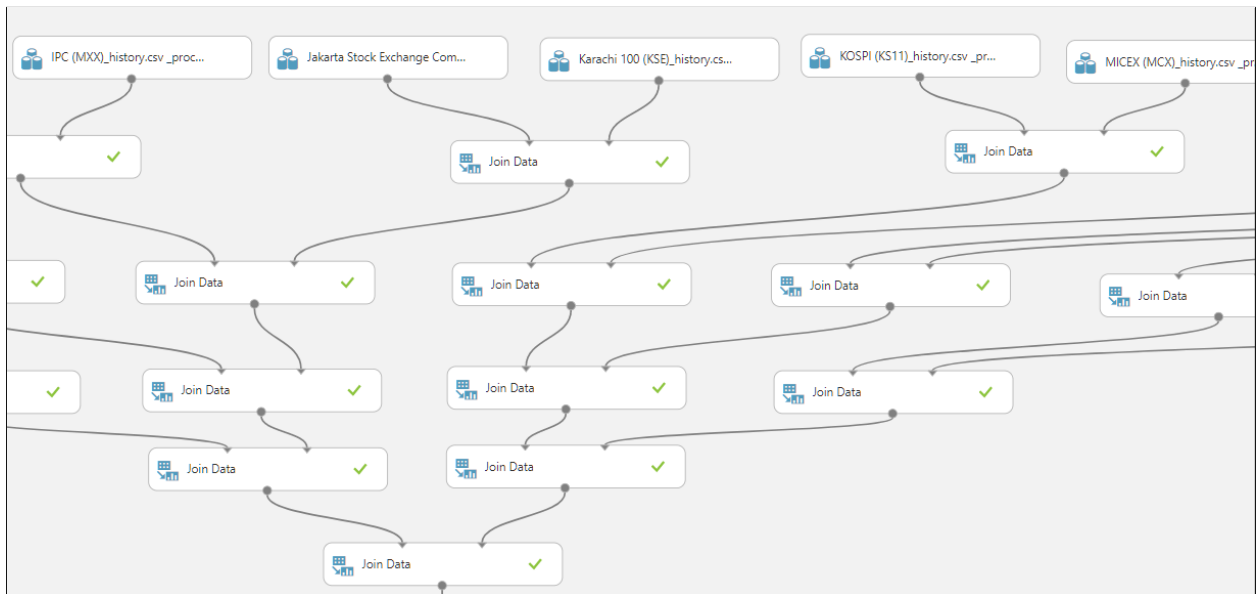


Abbildung 4.6: Zusammenführen der Aktienindices in Azure Machine Learning Studio

Daraufhin werden sie nachbearbeitet (die Beschreibung bezieht sich auf Abbildung 4.7): In jedem Ursprungsdatensatz existiert eine Spalte, die die Zeilen durchnummeriert. Dies diente zum beheben von Fehlern und wird nun entfernt (1. Select Columns in Dataset). Ebenso enthält das Ergebnis noch 35 Datums-Spalten (von jedem Ursprungsdatensatz eine). Diese werden bis auf Eine entfernt (2. Select Columns in Dataset und Edit Metadata).

4 Durchführung der Analyse

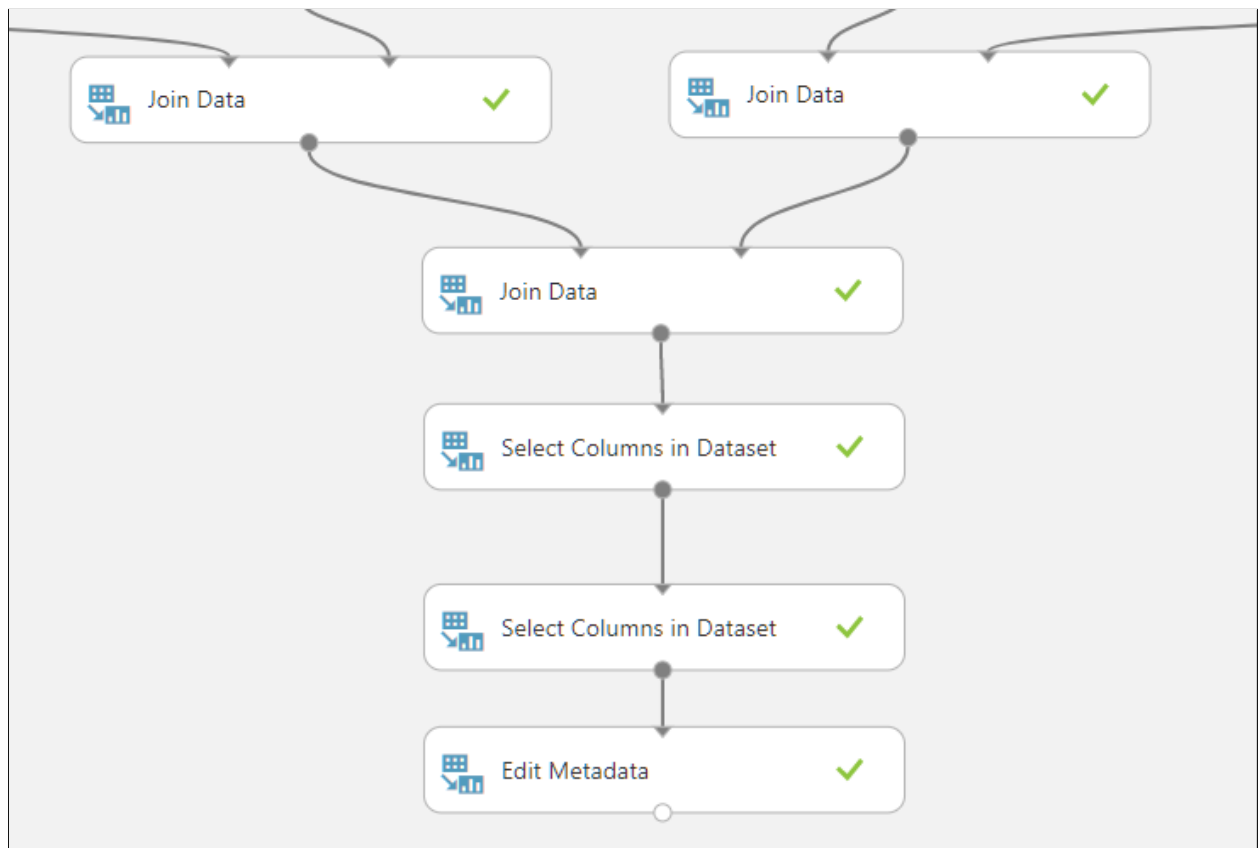


Abbildung 4.7: Nachbearbeiten des Ergebnisdatensatzes in Azure Machine Learning Studio

Analog werden die natürlichen Ressourcen zu einem Datenblock zusammengefasst. Ähnlich wird mit den Währungskursen vorgegangen. Die Verarbeitung in R ist jedoch einfacher, da sie keine Tausendertrennzeichen, keine pseudo-numerischen Werte und keine Lücken enthalten. Spezieller ist der Datensatz 'abcnews _Date _Text' (siehe Abbildung 4.8). Er wird mit dem Experiment Item 'Preprocess Text' vorbearbeitet, nachdem alte Überschriften (vor dem 1.1.2011) entfernt wurden (Execute R Script; siehe Listing 4.5 Zeile 6).

4 Durchführung der Analyse

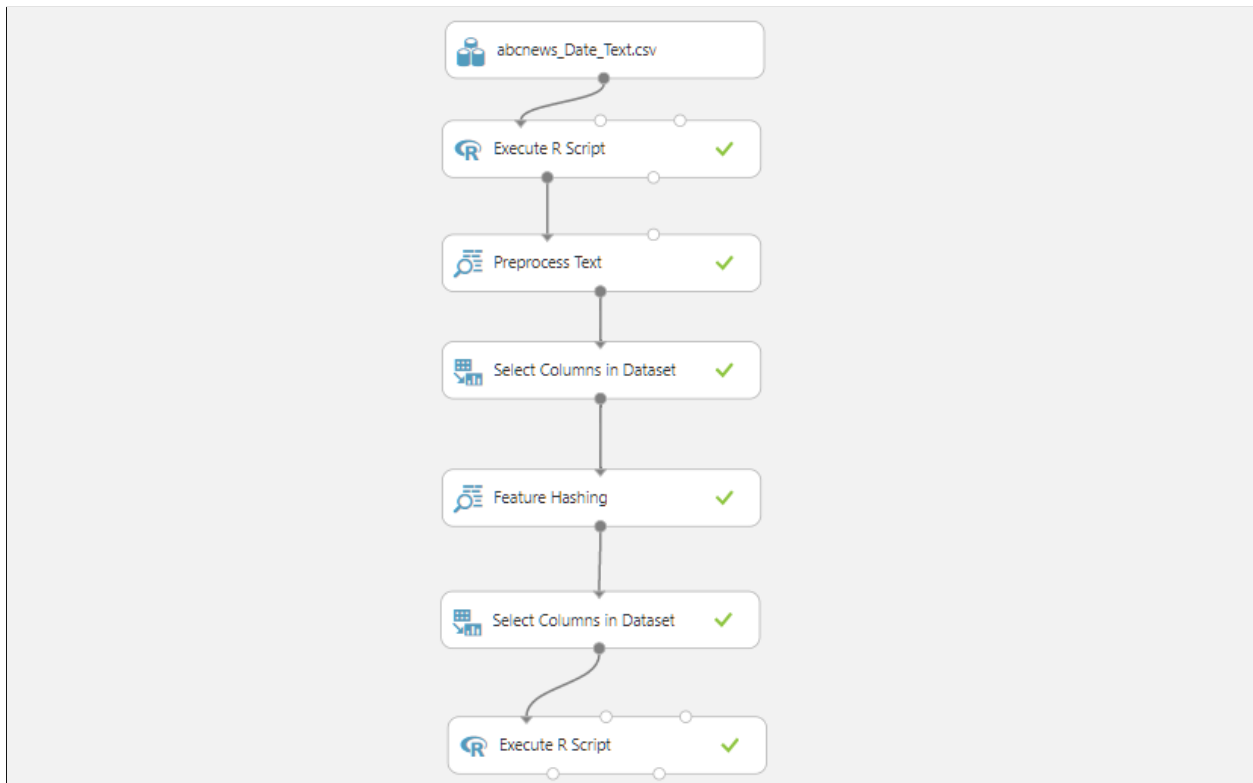


Abbildung 4.8: Feature Hashing der abcnews in Azure Machine Learning Studio

```
1 | # Map 1-based optional input ports to variables
2 | abcnews <- maml.mapInputPort(1) # class: data.frame
3 | attach(abcnews)
4 |
5 | #remove old news
6 | abcnews <- abcnews[abcnews$publish_date > 20110101,]
7 |
8 | #convert text to date
9 | abcnews$publish_date <- as.Date(as.character(abcnews$publish_date), "%Y%m%d")
10 |
11 | # Select data.frame to be sent to the output Dataset port
12 | maml.mapOutputPort("abcnews");
```

Listing 4.5: Entfernen der alten Überschriften

Das Preprocessing führt unter anderem eine Lemmatisierung durch und entfernt Zahlen und Sonderzeichen. Anschließend wird ein Feature Hashing durchgeführt. Dabei handelt es sich um die Überführung von Texttupeln in Zahlwerte, damit eine Verarbeitung durch Computer möglich wird. Für jede Überschrift existiert eine einzelne Zeile im Datensatz. Diese werden mit dem R-Script in Listing 4.6 in Azure aggregiert.

```
1 | # Map 1-based optional input ports to variables
```

4 Durchführung der Analyse

```

2 | abcnews <- maml.mapInputPort(1) # class: data.frame
3 | attach(abcnews)
4 |
5 | #aggregate by date
6 | abccompact <- aggregate(. ~ publish_date, abcnews, sum)
7 |
8 | # Select data.frame to be sent to the output Dataset port
9 | maml.mapOutputPort("abccompact");

```

Listing 4.6: Aggregieren der Hashing-Ergebnisse

Zuletzt werden noch die Google News- und Websuchen zusammengefasst. Der abschließende Data cleaning report ist in Tabelle 4.32 zu finden.

Problem	Lösung	Werkzeug
unterschiedliche Datumsformate	Zuerst werden die Daten in Text und anschließend mit dem Packet 'anytime' in R konvertiert.	R + Package 'anytime' (siehe Listing 4.4 Zeile 37)
Lücken durch joinen der Daten	Es handelt sich um 'Missing not at random'-Fehler (MNAR)(Graham, 2009, S. 553). Das bedeutet, dass die Lücken absichtlich sind (weil an diesen Tagen nicht gehandelt wird). Da es kein statistisches Verfahren zum schließen solcher Lücken gibt, werden sie beibehalten. (Leonhart, 2014, S. 1109)	
unterschiedliche Separatoren, Trennzeichen etc.	Die Daten werden mit Excel von '.tsv' in '.csv' konvertiert. Eine programmatische Änderung oder 'Find and Replace' ist schwer, da es einen Tab als Trennzeichen zwischen Tag, Monat und Jahr gibt. Kommas als Tausendertrennzeichen werden entfernt.	Excel 'Save As...' und R-Methode 'gsub' (siehe Listing 4.4 Zeilen 30-34)
pseudo-numerische Werte (z.B. „1.5M“ oder „0.07k“)	Bei betroffenen Werten wird der Buchstabe entfernt und die Zahl mit dem entsprechenden Wert multipliziert (z.B. 1.3×1000 für 1.3K)	R-Methoden 'gsup' und 'grepl', sowie mathematische Operationen (siehe Listing 4.7)

4 Durchführung der Analyse

Fehlende Werte	Fehlende Werte im Volumen werden als '-' angegeben, nicht als 'Nichts'. Durch ersetzen des '-' in R wird das Problem behoben.	R: Listing 4.4 Zeile 45
----------------	-------------------------------------------------------------------------------------------------------------------------------	-------------------------

Tabelle 4.32: Data cleaning report

```

1  #clear environment
2  rm(list = ls())
3
4  #converts a vector of pseudo numerics to a vector of numerics
5  pseudoNumeric.as.numeric <- function(initialVector) {
6
7      #the new vector to fill with values
8      newVector <- c()
9
10     #convert from factors to characters
11     initialVector <- as.character(initialVector)
12
13     #convert each value
14     for(x in initialVector){
15
16         #see which case
17         #(1) k or K for 1000
18         #(2) m or M for 1000000
19         #(3) b or B for 1000000000
20         if (grepl("k", x) || grepl("K", x)) {
21
22             #remove character
23             x <- gsub("K", "", x)
24             x <- gsub("k", "", x)
25
26             #convert to numeric
27             x <- as.numeric(x)
28
29             #multiply to get correct number
30             x <- x * 1000
31         }
32         if (grepl("m", x) || grepl("M", x)) {
33             #remove character
34             x <- gsub("m", "", x)
35             x <- gsub("M", "", x)
36
37             #convert to numeric
38             x <- as.numeric(x)
39
40             #multiply to get correct number
41             x <- x * 1000000
42         }
43     }
44     if (grepl("b", x) || grepl("B", x)) {

```

4 Durchführung der Analyse

```

45   #remove character
46   x <- gsub("b", "", x)
47   x <- gsub("B", "", x)
48
49   #convert to numeric
50   x <- as.numeric(x)
51
52   #multiply to ge correct number
53   x <- x * 10000000000
54 }
55
56 #add the newly converted value to the return vector
57 newVector <- c(newVector, as.character(x))
58 }
59
60 #return when finished
61 return(newVector)
62 }
```

Listing 4.7: Konvertierung von pseudo-numerischen Werten

Der Schritt 'Construct data' enthält die Möglichkeit, neue Features hinzuzufügen. Hier ist kein Bedarf bedarf dafür. Außerdem handelt es sich teilweise schon um stark aggregierte Features, wie den STFL (siehe 4.1.2). Damit sind die Konvertierungen, Formatierungen und Zusammenführungen der Datenblöcke abgeschlossen. Die Ergebnisse sind in den Tabellen 4.33 und 4.34 festgehalten.

Output	Beschreibung
Hashed __news	Die abcsnews nach dem Feature Hashing und der Aggregation.

Tabelle 4.33: Output: Generated records

Output	Beschreibung
Joined __Indices	Alle Aktienincides in einem Datensatz.
Joined __Currencies	Alle Währungen in einem Datensatz.
Joined __Resources	Alle natürlichen Ressourcen in einem Datensatz.
Joined __Public	Google Suchen und Google Newssuchen in einem Datensatz.

Tabelle 4.34: Output: Merged data

4 Durchführung der Analyse

Syntaktische Änderungen an den Datensätzen, also verändern der Spaltenreihenfolgen oder der Spaltenbezeichnungen sind Teil des Schrittes 'Format data'. Es kann beispielsweise in einem eingesetzten Werkzeug notwendig sein, dass die erste Spalte einer Tabelle ein eindeutiger Schlüssel sein muss. (Chapman et al., 2000, S. 46) Obgleich dieser Fall hier nicht eintritt, werden syntaktische Änderungen vorgenommen: Wie oben bereits erwähnt, entstehen beim Zusammenführen der Tabellen doppelte Spalten. Werden zwei Tabellen gejoint, so existieren zwei Datumsspalten, die die gleichen Werte beinhalten (rechte Datumsspalte und linke Datumsspalte). Diese Redundanzen werden entfernt. Außerdem wird die beibehaltene Spalte in 'Date' umbenannt.

Alle erzeugten Blöcke und einzelnen Datensätze werden nun im Azure Machine Learning Studio zu einem Datensatz zusammengefasst. Für das Machine Learning werden aus diesem die benötigten Spalten für BTC bzw. ETH ausgewählt. Da das Joinen wie bisher verläuft, wird auf eine extra Beschreibung von Abbildung 4.9 verzichtet.

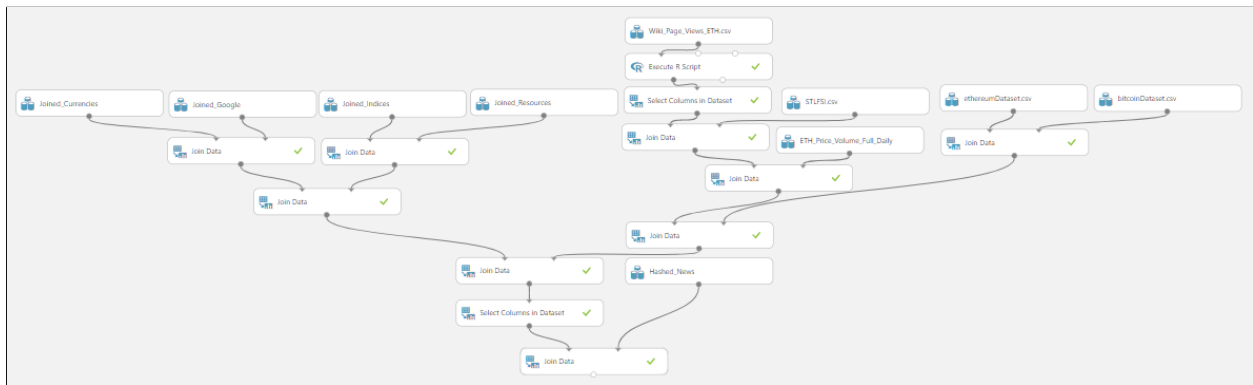


Abbildung 4.9: Erzeugen des finalen Datensatzes für das Machine Learning

Der letzte Schritt vor dem Machine Learning ist das Auswählen des Analysezeitraums und das bereinigen fehlender Daten (siehe Abbildung 4.10).

4 Durchführung der Analyse

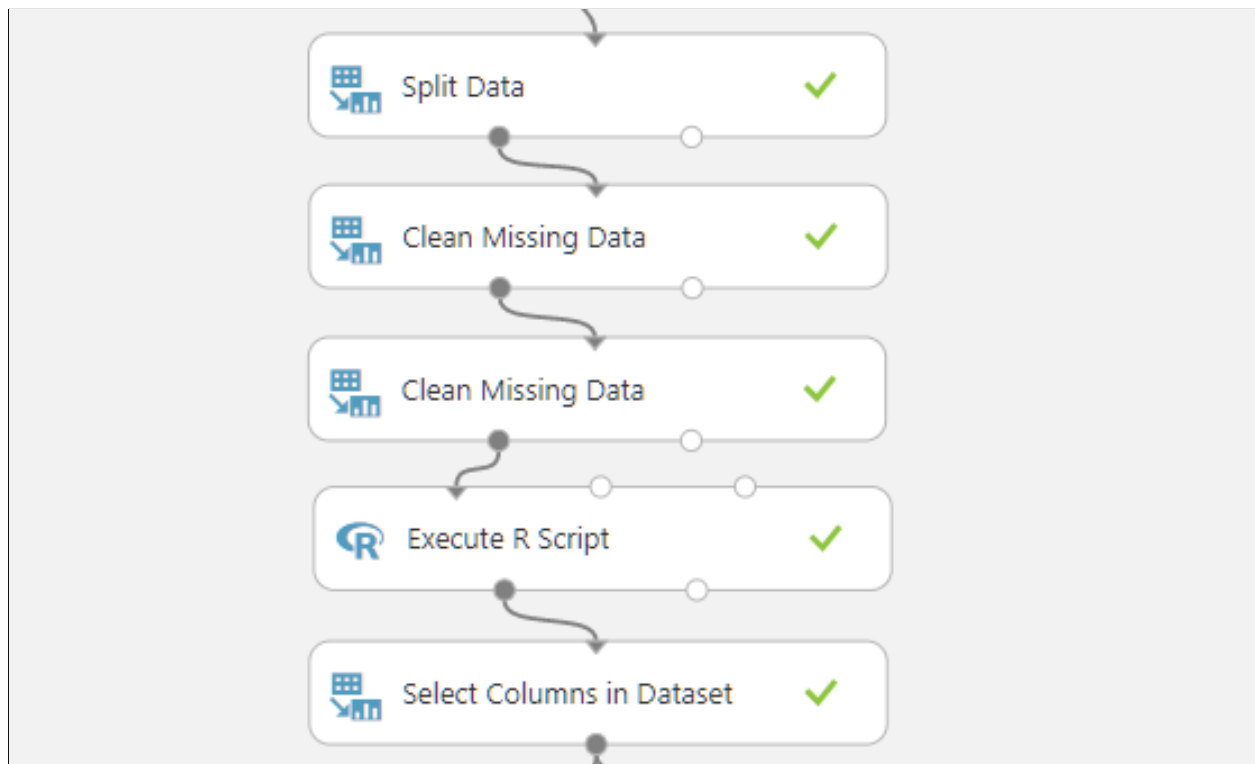


Abbildung 4.10: Auswahl des Analysezeitraums und bereinigen fehlender Daten in Azure Machine Learning Studio

Um nur die Zeilen zur Analyse zu benutzen, die im ausgewählten Zeitintervall (siehe Punkt 4.2.3) liegen, wird das Datensatz noch aufgesplittet (Experiment Item 'Split Data' mit Relative Expression \,„Date“ >= 2011-01-01T00:00:00 für BTC und \,„Date“ >= 2015-07-30T00:00:00 für ETH). Daraufhin werden mit dem Item 'Clean Missing Data (Probabilistic PCA)' fehlende Daten ersetzt, die nicht unter die Kategorie MNAR fallen. Bei neun Spalten ist dies nicht möglich (keine Fehlermeldung in Azure Machine Learning; vermutlich zu wenige Werte für Probabilistic PCA). Diese werden mit einem erneuten 'Clean Missing Data (Remove entire column)' entfernt. Als Abschluss wird mit dem R-Skript in Listing 4.8 eine neue Spalte angefügt, die angibt, ob der Kurs zum Vortag gestiegen ist (Wert: 1) oder nicht (Wert: 0). Dieser wird im Nachfolgenden als Vorhersagewert genutzt.

```
1 | # Map 1-based optional input ports to variables
2 | data <- maml.mapInputPort(1) # class: data.frame
3 |
4 | #add column increase and default to NA
5 | data[, "increase"] <- NA
6 |
7 | #get the column with the price
8 | priceVector <- data[, "btc_market_price"]
```

```

9
10 #returns 0 if there was no increase in value; else returns 1
11 checkForIncrease <- function(prev, cur) {
12   #skip na values
13   if ((is.na(prev)) || (is.na(cur))) {
14     return(0)
15   } else {
16     #set to 1
17     increase <- 1
18
19     #if there was no increase --> set to zero
20     if (prev >= cur) {
21       increase <- 0
22     }
23
24     #return increase
25     return(increase)
26   }
27 }
28
29 #for each value
30 for (i in 1:length(priceVector)) {
31   #get values
32   prev <- data[i, "btc_market_price"]
33   cure <- data[i + 1, "btc_market_price"]
34
35   #fill values
36   data[i + 1, "increase"] <-
37     checkForIncrease(prev, cure)
38 }
39
40 # Select data.frame to be sent to the output Dataset port
41 maml.mapOutputPort("data")

```

Listing 4.8: Hinzufügen und Popularisieren der Spalte „increase“ in Azure Machine Learning Studio

4.4 Modeling

4.4.1 Select the Modeling Technique

Wie in Punkt 4.1.3 gefordert, wird einerseits versucht, einen Wert für den Kurs vorherzusagen (mit Hilfe von Regressionen) und andererseits, ob der Kurs steigt oder nicht (Two-Class classification). In beiden Fällen handelt es sich um Supervised Learning, da für jeden Input

4 Durchführung der Analyse

ein Output vorhanden ist (siehe Punkt 3.2.1). In Azure verfügbare Two-Class classification Algorithmen sind:

- Two-Class Support Vector Machine
- Two-Class Neural Network
- Two-Class Logistic Regression
- Two-Class Locally-Deep Support Vector Machine
- Two-Class Decision Jungle
- Two-Class Decision Forest
- Two-Class Boosted Decision Tree
- Two-Class Bayes Point Machine
- Two-Class Averaged Perceptron

Bei den Regressionen sind es:

- Bayesian Linear Regression
- Boosted Decision Tree Regression
- Decision Forest Regression
- Fast Forest Quantile Regression
- Linear Regression
- Neural Network Regression
- Ordinal Regression
- Poisson Regression

4 Durchführung der Analyse

Ausgeschlossen werden Ordinal Regression (nur geeignet, um Rangfolgen aufzustellen), Poisson Regression (nur geeignet, um vorherzusagen, wie oft etwas passiert), Linear Regression (nur für sehr simple Modelle) und Ordinal Regression (zwar geeignet, nutzt aber andere Metriken und ist nicht vergleichbar). Die Algorithmen besitzen konfigurierbare Parameter. Die Güte der erzeugten Modelle hängt unter anderem von diesen einstellbaren Werten ab. Azure bietet die Möglichkeit, diese Parameter automatisch zu optimieren. Dazu wird das Experiment Item 'Tune Model Hyperparameter' eingesetzt. Es verbessert das Ergebnis des Models hinsichtlich eines bestimmten Wertes (Genauigkeit, Präzision etc.). Im vorliegenden Fall werden als Zielparameter die Metriken aus den Data Mining Goals (Punkt 4.1.3) gewählt: F1-Score für Klassifikationen und Bestimmtheitsmaß (Coefficient of determination) für die Regressionen.

4.4.2 Generate Test Design

Es muss überlegt werden, in welchem Verhältnis die zu analysierenden Daten aufgeteilt werden (Trainingsdaten:Testdaten). Es gibt kein allgemeingültiges, optimales Verhältnis. In der Praxis sind jedoch Verhältnisse wie 70:30 oder 80:20 üblich. Es existieren auch andere Ansätze, die wiederum auf Algorithmen basieren.(Crowther and Cox, 2005) Beachtet werden muss jedoch, dass nicht einfach obere und untere Hälfte des Datensatzes genutzt werden, sondern die Aufteilung nach dem Zufallsprinzip geschieht. Da das Experiment Item 'Tune Model Hyperparameter' eingesetzt wird, gibt es die Möglichkeit, ein zusätzliches Validation Dataset zu nutzen. Das für diese Analyse gewählte Verhältnis ist in Abbildung 4.11 zu sehen.

4 Durchführung der Analyse

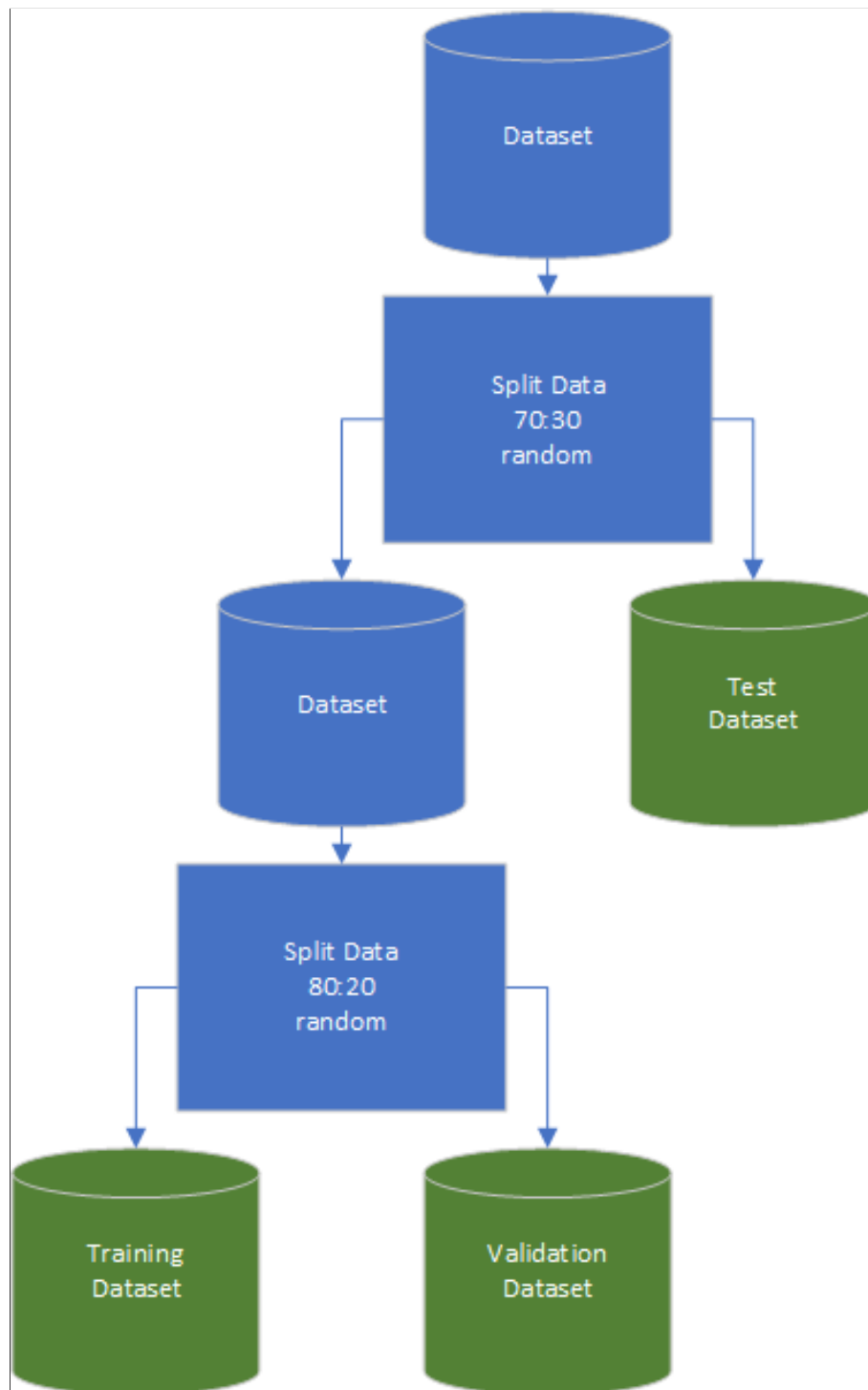


Abbildung 4.11: Verhältnis der Aufteilen in Trainings-, Test- und Validationsdatensatz

4.4.3 Build the Model

Ziel: Models bauen -> „azure ml run“

model aufschreiben und welche parameter „model description“

sind quasi 2: 2class beschreibung und regression beschreibung

2class: Algorithmus Accuracy Precision Recall F1-Score AUC Average Log Loss Two-Class Support Vector Machine Two-Class Neural Network Two-Class Logistic Regression Two-Class Locally-Deep Support Vector Machine Two-Class Decision Jungle Two-Class Decision Forest Two-Class Boosted Decision Tree Two-Class Bayes Point Machine Two-Class Averaged Perceptron

4.4.4 Assess the Model

hier r modul und add rows -> übersichtlich in tabelle schreiben ranking schwer, da unterschiedl.

ranking bei „nicht-treffen“ sinnlos, deswegen nur 2class regressionen

4.5 Evaluation

4.5.1 Evaluate Results

4.5.2 Review Process

4.5.3 Determine Next Steps

5 Interpretation Fazit

6 Related Work

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

7 Ausblick

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Literaturverzeichnis

(2014). About Project Jupyter.

(2016). Brent (Öl). Page Version ID: 156366434.

(2017a). Hochschule München - Bibliothek - Recherche - Übersicht.

(2017b). The Jupyter Notebook.

(2017c). West Texas Intermediate. Page Version ID: 169755263.

Androulaki, E., Karame, G. O., Roeschlin, M., Scherer, T., and Capkun, S. (2013). Evaluating User Privacy in Bitcoin. In *Financial Cryptography and Data Security*, Lecture Notes in Computer Science, pages 34–51. Springer, Berlin, Heidelberg.

Appelrath, H.-J., Kagermann, H., and Krcmar, H. (2014). *Future Business Clouds: Ein Beitrag zum Zukunftsprojekt Internetbasierte Dienste für die Wirtschaft*. Herbert Utz Verlag.

Azur, M. J., Stuart, E. A., Frangakis, C., and Leaf, P. J. (2011). Multiple Imputation by Chained Equations: What is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49.

Bajpai, P. (2014). Altcoin.

Baur, D. G., Lee, A. D., and Hong, K. (2015). Bitcoin: Currency or Investment? SSRN Scholarly Paper ID 2561183, Social Science Research Network, Rochester, NY.

Bitkom and KPMG (2017). Welche Public-Cloud-Anwendungen als Software-as-a-Service nutzen Sie?

Brandt, M. (2017). Infografik: Die Top 10 der Kryptowährungen.

Literaturverzeichnis

- Buterin, V. (2016). Let's talk about the projected coin supply over the coming years.. • r/ethereum.
- Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58.
- Chapelle, O., Schölkopf, B., and Zien, A., editors (2006). *Semi-supervised learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass. OCLC: ocm64898359.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.
- Christidis, K. and Devetsikiotis, M. (2016). Blockchains and Smart Contracts for the Internet of Things. *IEEE Access*, 4:2292–2303.
- CoinDesk (2017). Anzahl der Altcoins weltweit in ausgewählten Monaten von Dezember 2015 bis September 2016.
- CoinMarketCap (2017). Ranking der größten virtuellen Währungen nach Marktkapitalisierung im Juli 2017 (in Millionen US-Dollar).
- Crowther, P. S. and Cox, R. J. (2005). A Method for Optimal Division of Data Sets for Use in Neural Networks. In *Knowledge-Based Intelligent Information and Engineering Systems*, Lecture Notes in Computer Science, pages 1–7. Springer, Berlin, Heidelberg.
- Dannen, C. (2017). *Introducing Ethereum and Solidity*. Apress, Berkeley, CA. DOI: 10.1007/978-1-4842-2535-6.
- Dhar, V. (2013). Data Science and Prediction. *communications of the acm*, vol. 56 no. 12:10. doi:10.1145/2500499.
- Dr. Sayad, S. (2017). Model Evaluation.
- Ericson, G. and Rohm, W. A. (2017a). How to choose machine learning algorithms.
- Ericson, G. and Rohm, W. A. (2017b). Microsoft Azure Machine Learning: Algorithm Cheat Sheet.
- Ericson, G. and Rohm, W. A. (2017c). What is Azure Machine Learning Studio?
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.

Literaturverzeichnis

- Finance, Y. (2017a). Major world indices - Yahoo Finance.
- Finance, Y. (2017b). Major World Indices - Yahoo Finance.
- Fraunhofer, I. (2010). Vorteile von SaaS-Angeboten | IT-Anbieter Umfrage.
- Fritsch, W. (2013). Salesforce.com überholt im CRM-Markt SAP.
- Fund, T. I. M. (2017). World Economic Outlook Database.
- Garcia, D., Tessone, C. J., Mavrodiev, P., and Perony, N. (2014). The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy. *Journal of The Royal Society Interface*, 11(99):20140623–20140623.
- Gartner (2017). Umsatz mit Software-as-a-Service (SaaS) weltweit von 2010 bis 2016 und Prognose bis 2020 (in Milliarden US-Dollar).
- GoogleTrends (2017). GoogleTrends Vergleich: Bitcoin, Ethereum, Cryptocurrency.
- Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, 60(1):549–576.
- Hawksby-Robinson, S. (2017). What is Ethereum’s inflation rate? (how quickly will new ether be created) - Ethereum Stack Exchange.
- Hertle, J. (2016). Datenanalyse - Vorlesung Master, Hochschule München, SS 2016.
- (IBM), I. B. M. C. (2017). IBM Watson.
- Jdebunt (2017). What are Ethereum Uncles?
- Kauchak, D. (2016). zoteroVoll2.pdf.
- Kim, P. (2017). *MATLAB Deep Learning*. Apress, Berkeley, CA. DOI: 10.1007/978-1-4842-2845-6.
- Kraker, P. and Dennerlein, S. (2013). Towards a Model of Interdisciplinary Teamwork for Web Science: What can Social Theory Contribute? *Web Science 2013 Workshop: Harnessing the Power of Social Theory for Web Science*.
- Kristoufek, L. (2013). BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, 3:srep03415.

Literaturverzeichnis

- Kristoufek, L. (2015). What Are the Main Drivers of the Bitcoin Price? Evidence from Wavelet Coherence Analysis. *PLOS ONE*, 10(4):e0123923.
- Kubat, M. (2017). *An Introduction to Machine Learning*. Springer International Publishing, Cham. DOI: 10.1007/978-3-319-63913-0.
- Kumar, raj, s. (2017). Cryptocurrency Historical Prices.
- Larose, D. T. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons. Google-Books-ID: UGu8AwAAQBAJ.
- Leonhart, R. (2014). *Dorsch – Lexikon der Psychologie*. M. A. Wirtz (Hrsg.), Bern: Hogrefe Verlag, 18. edition.
- Limited, F. M. (2017). Major Stock Indices.
- Lison, P. (2012). An introduction to machine learning.
- liveindex.org (2017). Live Index.
- Lohninger, H. (2013). *Grundlagen der Statistik*. Web-version edition.
- Mircosoft (2017). Evaluate Model.
- Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System Bitcoin: A Peer-to-Peer Electronic Cash System.
- of St. Louis, F. R. B. (2017). St. Louis Fed Financial Stress Index [STLFSI].
- Olsson, F. (2009). A literature survey of active machine learning in the context of natural language processing.
- Paluszek, M. and Thomas, S. (2017). *MATLAB Machine Learning*. Apress, Berkeley, CA. DOI: 10.1007/978-1-4842-2250-8.
- Project, B. (2017a). Difficulty, Network Difficulty - Bitcoin Glossary.
- Project, B. (2017b). Orphan Block - Bitcoin Glossary.
- Project, B. (2017c). Stale Block - Bitcoin Glossary.
- Ramasubramanian, K. and Singh, A. (2017). *Machine Learning Using R*. Apress, Berkeley, CA. DOI: 10.1007/978-1-4842-2334-5.

- Reid, F. and Harrigan, M. (2013). An Analysis of Anonymity in the Bitcoin System. In Althshuler, Y., Elovici, Y., Cremers, A. B., Aharony, N., and Pentland, A., editors, *Security and Privacy in Social Networks*, pages 197–223. Springer New York. DOI: 10.1007/978-1-4614-4139-7_10.
- Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11.
- Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *JOURNAL OF DATA WAREHOUSING*, 5(4):13–22.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359.
- Sowa, A. (2017). *Management der Informationssicherheit*. Springer Fachmedien Wiesbaden, Wiesbaden. DOI: 10.1007/978-3-658-15627-5.
- Statista. Root Mean Square Error (RMSE; dt.: Wurzel der mittleren Fehlerquadratsumme) - Statista Definition.
- Suthaharan, S. (2016). *Machine Learning Models and Algorithms for Big Data Classification*, volume 36 of *Integrated Series in Information Systems*. Springer US, Boston, MA. DOI: 10.1007/978-1-4899-7641-3.
- Swamynathan, M. (2017). *Mastering Machine Learning with Python in Six Steps*. Apress, Berkeley, CA. DOI: 10.1007/978-1-4842-2866-1.
- Tipping, M. E. and Bishop, C. M. (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society*, 61(3):611–622.
- TSYS (2016). Kennen oder nutzen sie virtuelle Währungen wie Bitcoin?
- Tual, S. (2015). Ethereum Launches.
- Vermeulen, R., Hoeberichts, M., Vašíček, B., Hájková, D., Šmídková, K., and de Haan, J. (2014). Financial stress indexes and financial crises. *Journal of Economic Literature*, JEL-code: E5, G10.

Literaturverzeichnis

- WikiTrends (2017). Compare popularity of Bitcoin, Cryptocurrency, Ethereum on Wikipedia | Wiki Trends.
- Wood, G. (2014). Ethereum: A secure decentralised generalised transaction ledger. *Ethereum Project Yellow Paper*, 151.