



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 18. OKTOBER 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	VIII
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	31
3.2.2 Unsupervised Learning	32
3.2.3 Semi-supervised Learning	32
3.2.4 Reinforcement Learning	32
3.3 Kryptowährung(en)	32
3.4 SaaS	33

Inhaltsverzeichnis

3.5	Microsoft Azure ML Studio	33
3.5.1	Allgemeine Beschreibung	33
3.5.2	Aufbau	33
3.5.3	Elemente	33
4	Einflüsse	35
5	Daten	36
5.1	Kurse	36
5.2	Überschriften (Keggle)	36
5.3	andere Kurse/börsen	36
6	Durchführung	37
7	Interpretation Fazit	38
8	Related Work	39
9	Ausblick	40
	Literaturverzeichnis	41

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

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

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

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 and

1 Hinführung zum Thema

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))(??). Mittlerweile 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).

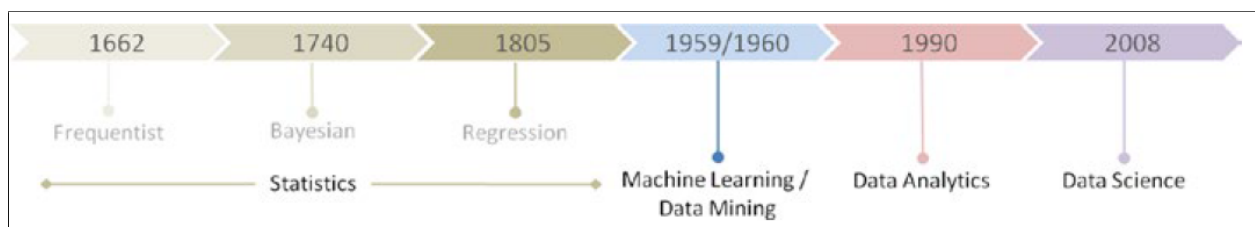


Abbildung 1.1: Learn from data evolution (Swamynathan, 2017, S. 66)

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 Hinführung 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

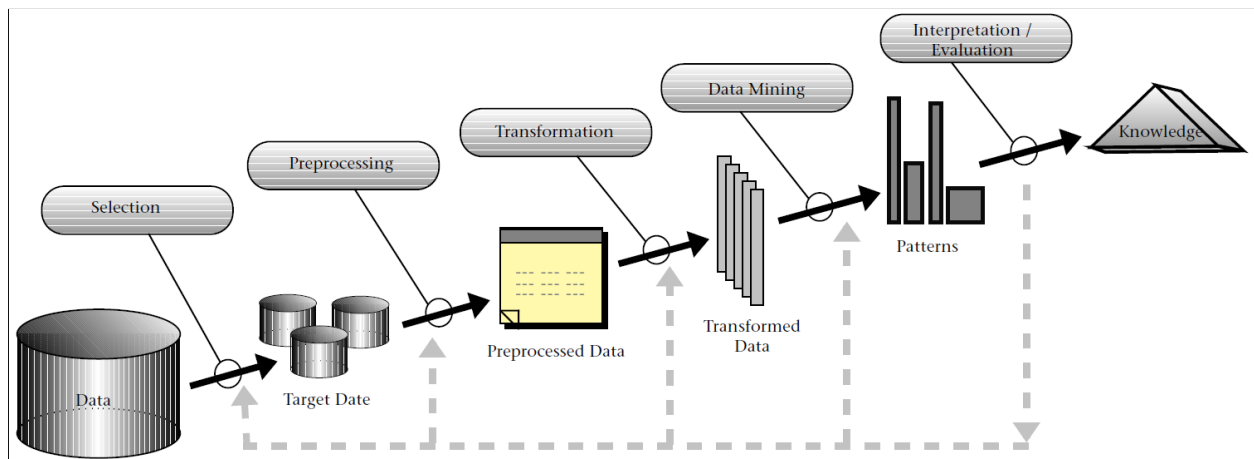


Abbildung 3.1: Ein Überblick über die Schritte des KDD Prozesses nach (Fayyad et al., 1996, S. 41)

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.

```
#Daten einlesen
data <- read.csv2("Beispiel_Berufserfahrung_Datensatz1.csv")

#Regression mit allen Faktoren
regression1 <- lm(Gehalt ~ Studium + ErfahrungExtern + ErfahrungIntern + ←
  Alter, data=data)
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).

```
#Regression mit signifikanten Faktoren
regression2 <- lm(Gehalt ~ Studium + ErfahrungExtern + ErfahrungIntern, ←
  data=data)
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).

```
# Transformation
data[, "ErfahrungGesamt"] <- data[, 3] + data[, 4]

# Regression
regression3 <- lm(Gehalt ~ Studium + ErfahrungGesamt, data=data)
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 Model:

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

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 Model, 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

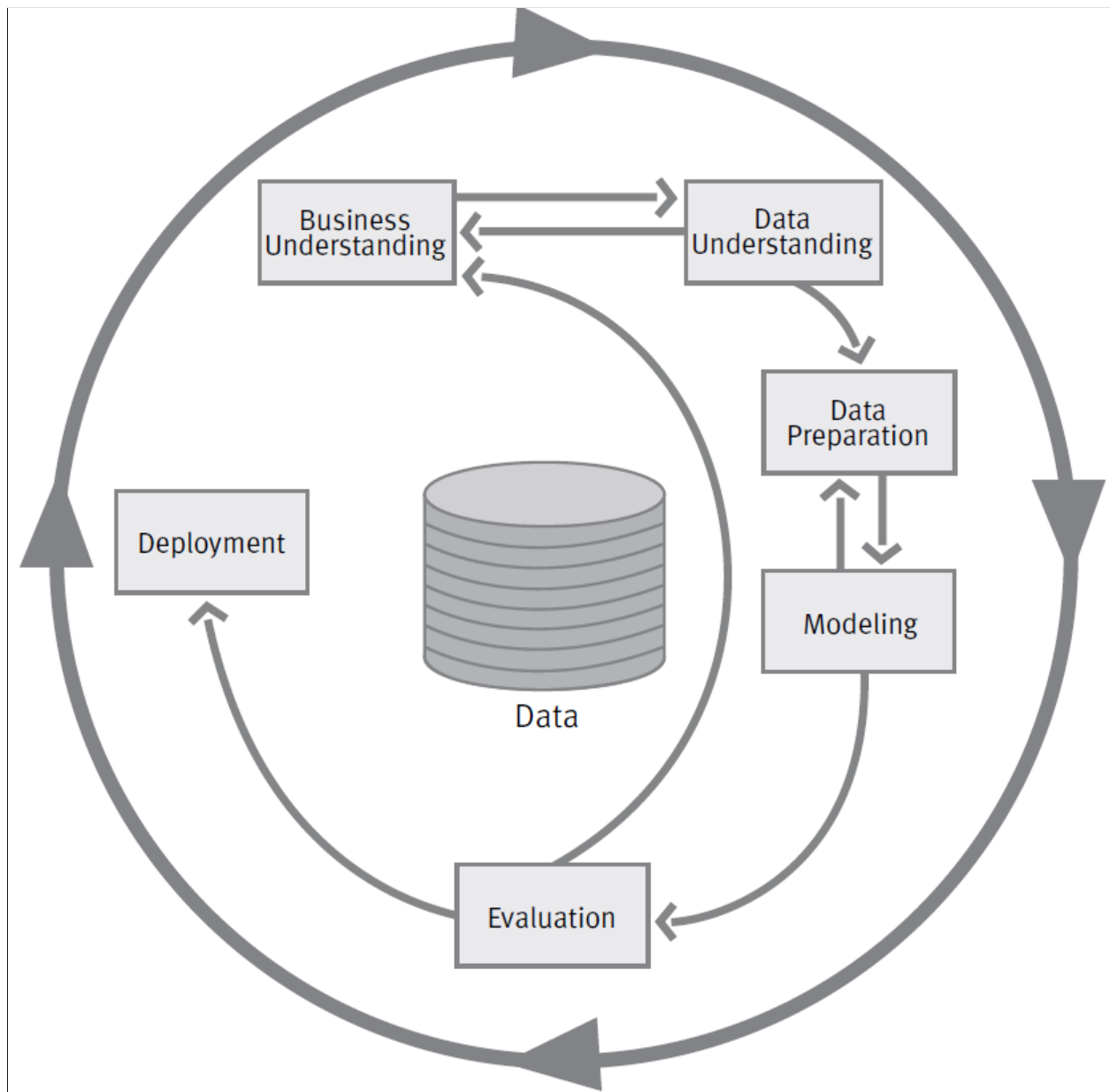


Abbildung 3.2: Phasen des CRISP-DM Referenzmodells nach (Chapman et al., 2000, S. 10)

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

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

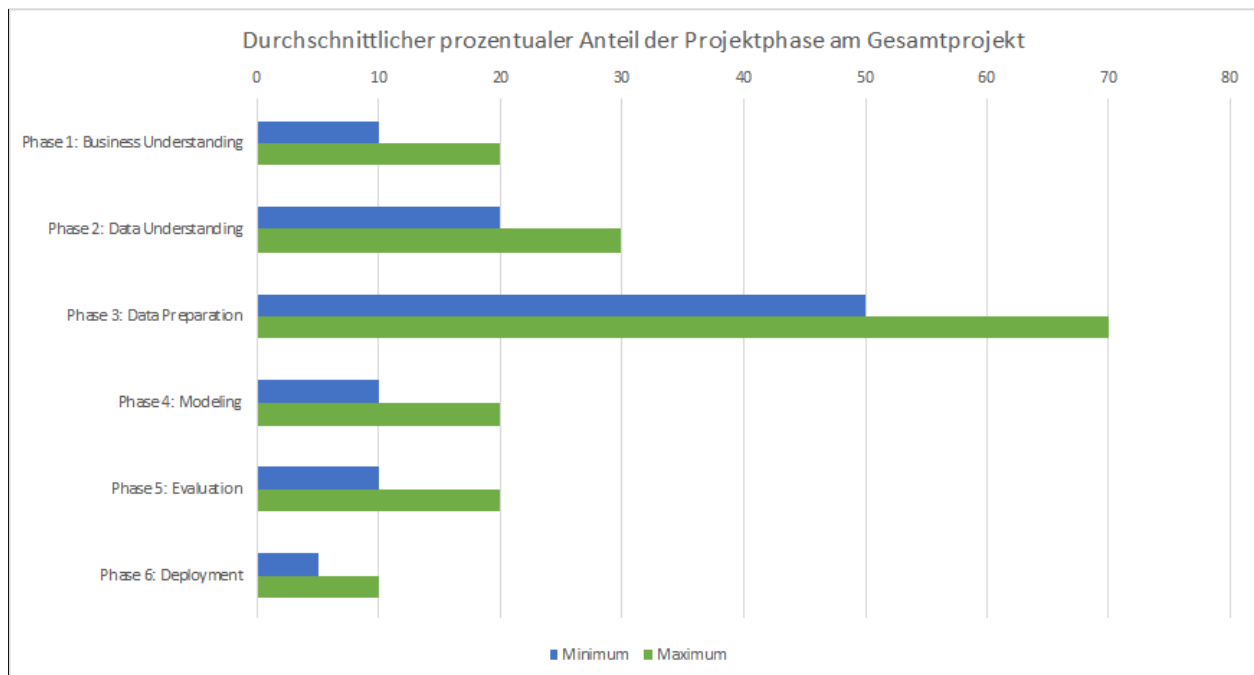


Abbildung 3.5: Durchschnittlicher prozentualer Anteil der CRISP-DM-Projektphase am Gesamtprojekt nach (Shearer, 2000, S. 15; eigene Darstellung)

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 3.6 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 Schlusstrich 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

3.2 Machine Learning

sehr sehr viele; verschieden ordnen: nach 1. nach 2. oder nach supervised, unsupervised... azure macht so, deswegen hier diese Kategorisierung; im Verlauf dieses Kapitels sehr viele Algorithmen in der Literatur für spezielle Fälle z.B. da Analyse mit Azure ML → nur die Algorithmen, die auch vorkommen **TODO: + honorable mentions??** englische Namen, da hauptsächlich genutzt

3.2.1 Supervised Learning

Man weiß, nach was man sucht... (training data mit input und korrektem output) • E.g. spam filtering, machine translation, face recognition, etc.

Classification

Two-class Classification auch „binomial classification“ (Ericson and Rohm, 2017) genannt

Multi-class Classification

Regression

Anomaly detection

3.2.2 Unsupervised Learning

Man sucht nur cluster/gruppen/etc Sometimes, we don't have access to any output value o , we simply have a collection of input examples i • In this case, what we try to do is to learn the underlying patterns of our data

Clustering

3.2.3 Semi-supervised Learning

nur wenig Labels

3.2.4 Reinforcement Learning

richtig oder falsch kommt erst noch

Sometimes we have no direct access to the «correct» output, but we can get some measure of the quality of an output o following input i

But we can get a measure of «how good/ bad» an output is

oft in robotic oder so, aber noch nicht in azure ml

3.3 Kryptowährung(en)

Bitcoin,ethereum, litecoin, dogecoin; auswahl hier nur 1/2

3.4 SaaS

3.5 Microsoft Azure ML Studio

3.5.1 Allgemeine Beschreibung

3.5.2 Aufbau

Projects

Experiments

Web Services

Notebooks

Datasets

Trained Models

Settings

3.5.3 Elemente

relevante auswählen

3 Grundlagen

Saved Datasets

Data Transformation Conversations

Data Transformation

Data Input and Output

Feature Selection

Machine Learning

OpenCV Library Models

Python Language Model

R Language Model

Statistical Functions

Text Analysis

Time Series Anomaly Detection

Web Service

4 Einflüsse

aus paper und mehr suchen

welchen einfluss hier; im nächsten teil dann: wie kann man das repräsentieren, welche daten gibt es da und kann man das abbilden?

beispiele: regierungen und regionen (usa, china, EU) → Gesetze

bitcoin-eigene dinge (volumen, umschlag, miner? etc.)

öffentlichkeit (twitter, zeitungen, blogs, domains im web)

natürliche Ressourcen (Öl, Gold, Silber, Diamanten w/e)

Financial Stress Index (FSI)

HIER PAPER NOCHMAL: * Economic Drivers * Transaction Drivers * Technical Drivers * Interest * Safe Haven * Influence of China

5 Daten

Welche Daten Brauche ich, wo kriege ich sie her, was steht drin, beschreibung, features etc.

5.1 Kurse

börse 1, 2, Währungen

5.2 Überschriften (Keggle)

5.3 andere Kurse/börsen

dax, china!, dow jones ...

6 Durchführung

von Ziele bis Interpretation

7 Interpretation Fazit

8 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.

9 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

- 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.
- 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.
- 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).

Literaturverzeichnis

- 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.
- Ericson, G. and Rohm, W. A. (2017). How to choose machine learning algorithms.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.
- Fraunhofer, I. (2010). Vorteile von SaaS-Angeboten | IT-Anbieter Umfrage.
- Fritsch, W. (2013). Salesforce.com überholt im CRM-Markt SAP.
- 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.
- Hertle, J. (2016). Datenanalyse - Vorlesung Master, Hochschule München, SS 2016.
- (IBM), I. B. M. C. (2017). IBM Watson.
- 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*.
- Larose, D. T. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. John Wiley & Sons. Google-Books-ID: UGu8AwAAQBAJ.
- Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System Bitcoin: A Peer-to-Peer Electronic Cash System.
- Ramasubramanian, K. and Singh, A. (2017). *Machine Learning Using R*. Apress, Berkeley, CA. DOI: 10.1007/978-1-4842-2334-5.

Literaturverzeichnis

- Reid, F. and Harrigan, M. (2013). An Analysis of Anonymity in the Bitcoin System. In Altshuler, 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.
- Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. *JOURNAL OF DATA WAREHOUSING*, 5(4):13–22.
- Swamynathan, M. (2017). *Mastering Machine Learning with Python in Six Steps*. Apress, Berkeley, CA. DOI: 10.1007/978-1-4842-2866-1.
- TSYS (2016). Kennen oder nutzen sie virtuelle Währungen wie Bitcoin?
- 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.