Kursvorhersage von Kryptowährungen mit Azure Machine Learning

Forecasting prices 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 14. August 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, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer 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

Tabellenverzeichnis

Listings

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 wird in Abschnitt 2 erläutert.

1.2 Bitcoin als Vorreiter der Krypotowährungen

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) 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) Ob das Bitcoinnetzwerk wirklich absolute Anonymität gewährt, wird stark kritisiert.(Reid and Harrigan; Androulaki et al.). 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.). Nicht zu vernachlässigen ist an dieser Stelle auch das Interesse der Industrie an „Smart Contracts“(Dannen, S. 10), die beispielsweise im Bereich des Internet of Things Anwedung finden.(Christidis and Devetsikiotis)

Neben Bitcoin hat sich deshalb zusätzlich eine Vielzahl an anderen sogenannten Kryptowährungen entwickelt. Die Währungen mit dem größten Marktvolumen sind Bitcoin(**??**) und Ethereum(**??**)(Wood).(Brandt; CoinMarketCap) Daneben gibt es noch sogenannte Altcoins (aus dem Englischen: alternative coin(Bajpai))(**??**). Mittlerweile umfassen diese 664 Bitcoin-Alternativen.(CoinDesk). Obgleich die tatsächliche Nutzung der Krypowährungen sehr gering ist (1% der Befragten in Deutschland(TSYS)), steigt das Interesse an Kryptowährungen(WikiTrends; GoogleTrends).
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, , S. 2; Swamynathan, , 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, S. 53; eigene Übersetzung) Bekannte Anwendungen aus dem Alltag sind Empfehlungssysteme oder Spamerkennungen.(Swamynathan, S. 53)
Data Mining beschreibt den Prozess, aus einer gewaltigen Menge an Daten die „richtigen Daten“, zur „richtigen Zeit“ für die „richtigen Entscheidungen“(Swamynathan, S. 61; eigene Übersetzung) zu gewinnen. Um diesen Prozess haben sich im Laufe Zeit drei Frameworks gebildet:(Swamynathan, 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, S. 4)

Eng verwandt mit dem Data Mining ist die Datenanalyse (engl. Data Analysis; in der Industrie auch Business Analytics(Swamynathan, S. 58)). Sie wird benutzt um(Hertle, 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 Datentenanalyse (engl. Explorative Data Analysis; EDA) und der Induktiven Statistik.(Hertle, 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) Er versucht die geänderten Anforderungen der heutigen Datenanalyse abzubilden (siehe 1.1).

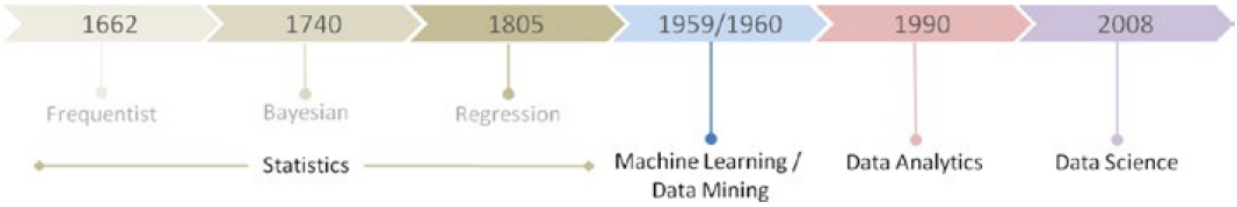

Abbildung 1.1: Learn from data evolution (Swamynathan, S. 66)

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.

Verdeutlicht wird dies durch Projekte wie Googles DeepMind(**?**), IBMs Watson((IBM)) 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, 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., S. 18). Innerhalb des Cloud Computing unterscheidet man weiterhin zwischen verschiedenen Cloud-Diensten (engl. cloud services). Nach (Appelrath et al., S. 20) differenziert man zwischen den Services in Abbildung 1.1. (Appelrath et al., 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) 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) erklärt werden, andererseits aber auch durch Tendenz der Softwarebranche hin zur serviceorientierten Architekturen (engl. service oriented atchitekture; SOA).(Appelrath et al., S. 22) Dieser Trend zu SaaS kann beobachtet werden, wenn reine Cloud-Anbieter wie Salesforce „klassische“ Anbieter wie SAP den Rang als „Spitze des Weltmarkts der Software für Customer Relationship Management (CRM)“(Fritsch) ablaufen.
Laut einer Studie von (Bitkom and KPMG) 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.

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

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 er 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 Formulier
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 vorauszusagen 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 Bezeichung Knowledge Discovery in Databases wurde hauptsächlich von (Fayyad et al.) 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., 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.

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., 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., S. 42) hergestellt werden. Dabei werden zuerst Daten aus unterschiedlichen - oft heterogenen - Quellen zusammengeführt und dann hinsichtlich des Ziels verdichtet.(Swamynathan, S. 70)

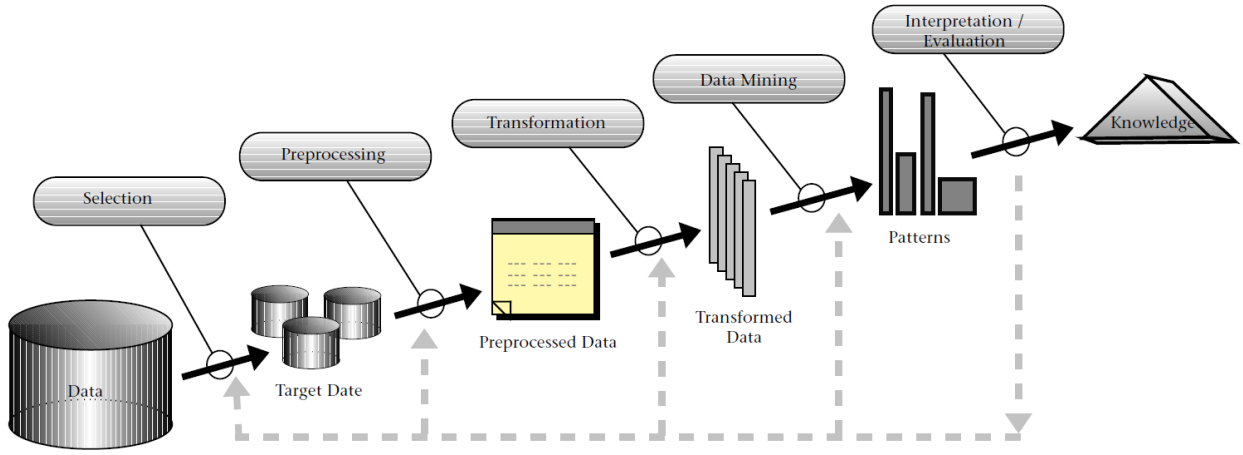

Abbildung 3.1: Ein Überblick über die Schritte des KDD Prozesses nach (Fayyad et al., S. 41)

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, S. 70) Um die Qualität der Daten und des Mining zu verbessern, werden unter anderem folgende Aspekte betrachtet:(Fayyad et al., , S. 42; Swamynathan, , S. 70)

Outliner treatment

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

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 Datenpunk 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 Genauigkei

Transformation

Der letzte Schritt vor dem eigentlichen Data Mining ist die Transformation. In diesem Prozessschritt geht es darum, „mit Dimensionsreduktions- oder -transformationsmethonden die effektive Anzahl an Variablen [...] zu reduzieren“(Fayyad et al., 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:

Tabelle 3.1 zeigt einen einfachen Datensatz, in dem die Mitarbeiter einer Firma und die

	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

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

Ohne weiter auf die genauen Beizeichnungen 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 Bestimmt-

	Estimate	Std. Error	t value	Pr(>||)
(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

heitsmaß (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

	Estimate	Std. Error	t value	Pr(>||)
(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(>||)
(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, S. 71; veränderte Version) gibt es zur Transformation folgende Möglichkeiten:

- 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 verletzten

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., , S. 42; Swamynathan, , 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 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., S. 42) Aus diesem Grund ist es auch jederzeit möglich, zu einem vorangegangen 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:

$$Gehalt = Studium \times 327,17 + ErfahrungGesamt \times 876,69 + 40107,10$$

Ein Muster (engl. pattern) beschreibt dagegen nur eine kleine „lokale Struktur“, die „nur über einen begrenzen Bereich“ Aussagen macht.(Swamynathan, 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). „Fayyad et al. benutzt patterns und models synonym“.(Kraker and Dennerlein)
Das Interpretieren der Ergebnisse beinhaltet ebenfalls das Zusammenfassen der Erkenntnisse und gegebenenfalls das Visualisieren.(Swamynathan, 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., 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 viele Iterationen geben, die auch „loops between any two steps“ beinhalten können.(Fayyad et al., 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 gegründetem Konsortium aus „Daimler-Benz (now DaimlerChrysler), Integral Solutions Ltd. (ISL) [jetzt SPSS], NCR, and OHRA“(Shearer, S. 13) erarbeitet. Die Version 1.0 wurde 2000 vorgestellt.(Shearer, S. 13) In Umfragen (1999, 2002, 2004, 2007) wird das Modell als führend in Bereich von „data mining/predictive analytics projects“(Swamynathan, S. 72) bezeichnet. Das Modell ist „nicht-properitär, dokumnetiert und frei verfügbar“(Shearer, 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, S. 13; eigene Übersetzung)
TODO: evtl. bessere Bezeichnung

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.

Business Understanding

Data Understanding

Data Preparation

Modeling

Evaluation

Deployment

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

3.1.4 Auswahl

3.2 Machine Learning

TODO: semi supervised etc.

3.2.1 Supervised...

Man weiß, nach was man sucht...

Decision Tree

Neares Neighbour

Random Forest

SVM

3.2.2 Unsupervised...

K means

Hierarchical clustering

Neuronal networks

...

Man sucht nur cluster/gruppen/etc

3.3 Krypotwährung(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

relevate auswählen

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

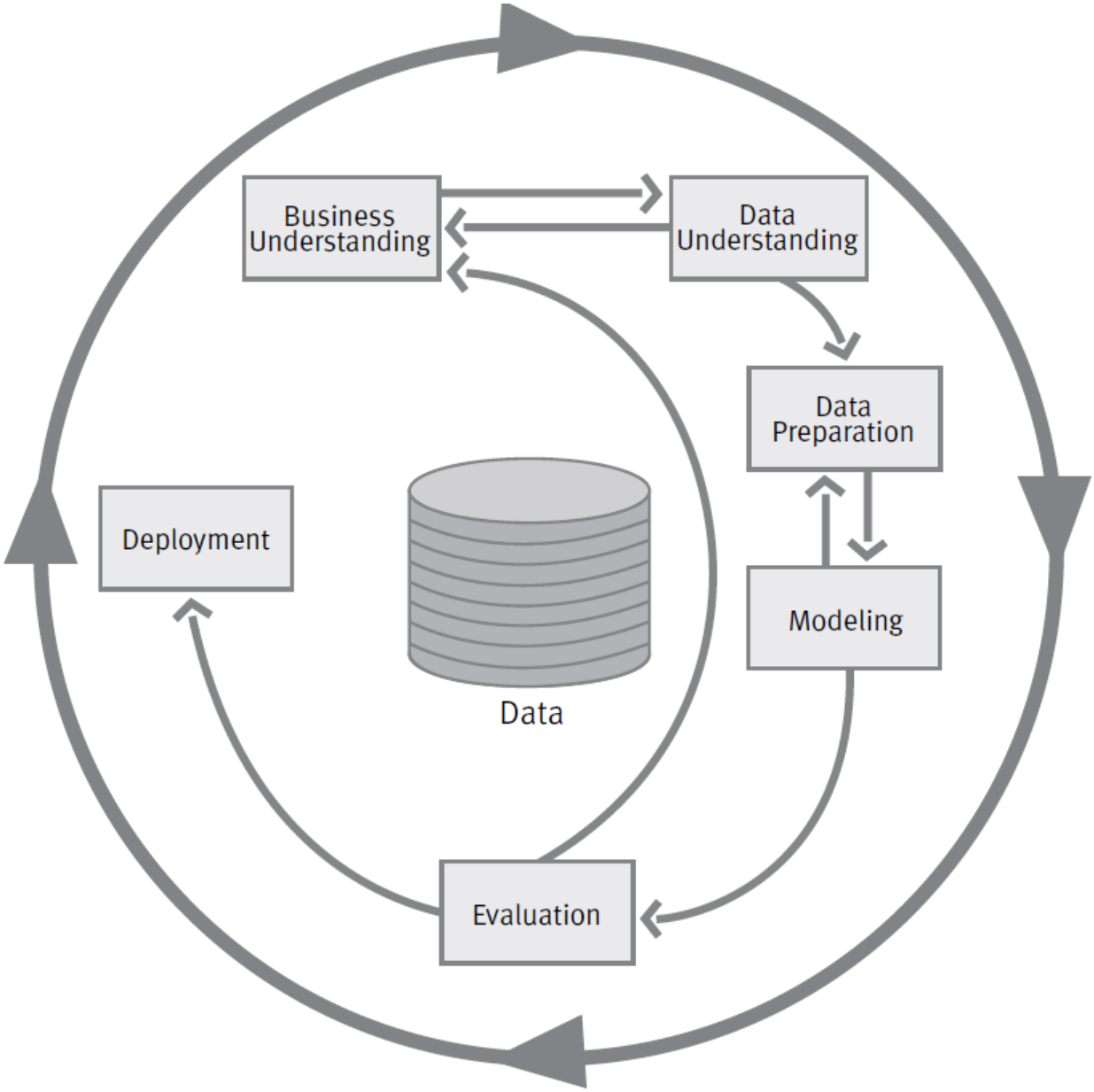

Abbildung 3.2: Phasen des CRISP-DM Referenzmodells nach (Chapman et al., S. 10)

4 Einflüsse

aus paper und mehr suchen

welchen einluss 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, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer 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, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer 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. 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. *Future Business Clouds: Ein Beitrag zum Zukunftsprojekt Internetbasierte Dienste für die Wirtschaft.* Herbert Utz Verlag.

Bajpai, P. Altcoin.

Baur, D. G., Lee, A. D., and Hong, K. Bitcoin: Currency or investment?

Bitkom and KPMG. Welche public-cloud-anwendungen als software-as-a-service nutzen sie?

Brandt, M. Infografik: Die top 10 der kryptowährungen.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. CRISP-DM 1.0 step-by-step data mining guide.

Christidis, K. and Devetsikiotis, M. Blockchains and smart contracts for the internet of things. 4:2292–2303.

CoinDesk. Anzahl der altcoins weltweit in ausgewählten monaten von dezember 2015 bis september 2016.

CoinMarketCap. Ranking der größten virtuellen währungen nach marktkapitalisierung im juli 2017 (in millionen US-dollar).

Dannen, C. *Introducing Ethereum and Solidity.* Apress. DOI: 10.1007/978-1-4842-2535-6.

Dhar, V. Data science and prediction. vol. 56 no. 12:10. doi:10.1145/2500499.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery in databases. 17(3):37.

Fraunhofer, I. Vorteile von SaaS-angeboten | IT-anbieter umfrage.

Fritsch, W. Salesforce.com überholt im CRM-markt SAP.

Gartner. Umsatz mit software-as-a-service (SaaS) weltweit von 2010 bis 2016 und prognose bis 2020 (in milliarden US-dollar).

GoogleTrends. GoogleTrends vergleich: Bitcoin, ethereum, cryptocurrency.

Hertle, J. Datenanalyse - vorlesung master, hochschule münchen, SS 2016.

(IBM), I. B. M. C. IBM watson.

Kim, P. *MATLAB Deep Learning*. Apress. DOI: 10.1007/978-1-4842-2845-6.

Kraker, P. and Dennerlein, S. Towards a model of interdisciplinary teamwork for web science: What can social theory contribute?

Nakamoto, S. Bitcoin: A peer-to-peer electronic cash system bitcoin: A peer-to-peer electronic cash system.

Ramasubramanian, K. and Singh, A. *Machine Learning Using R*. Apress. DOI: 10.1007/978-1-4842-2334-5.

Reid, F. and Harrigan, M. 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. The CRISP-DM model: The new blueprint for data mining. 5(4):13–22.

Swamynathan, M. *Mastering Machine Learning with Python in Six Steps*. Apress. DOI: 10.1007/978-1-4842-2866-1.

TSYS. Kennen oder nutzen sie virtuelle währungen wie bitcoin?

WikiTrends. Compare popularity of bitcoin, cryptocurrency, ethereum on wikipedia | wiki trends.

Wood, G. Ethereum: A secure decentralised generalised transaction ledger. 151.