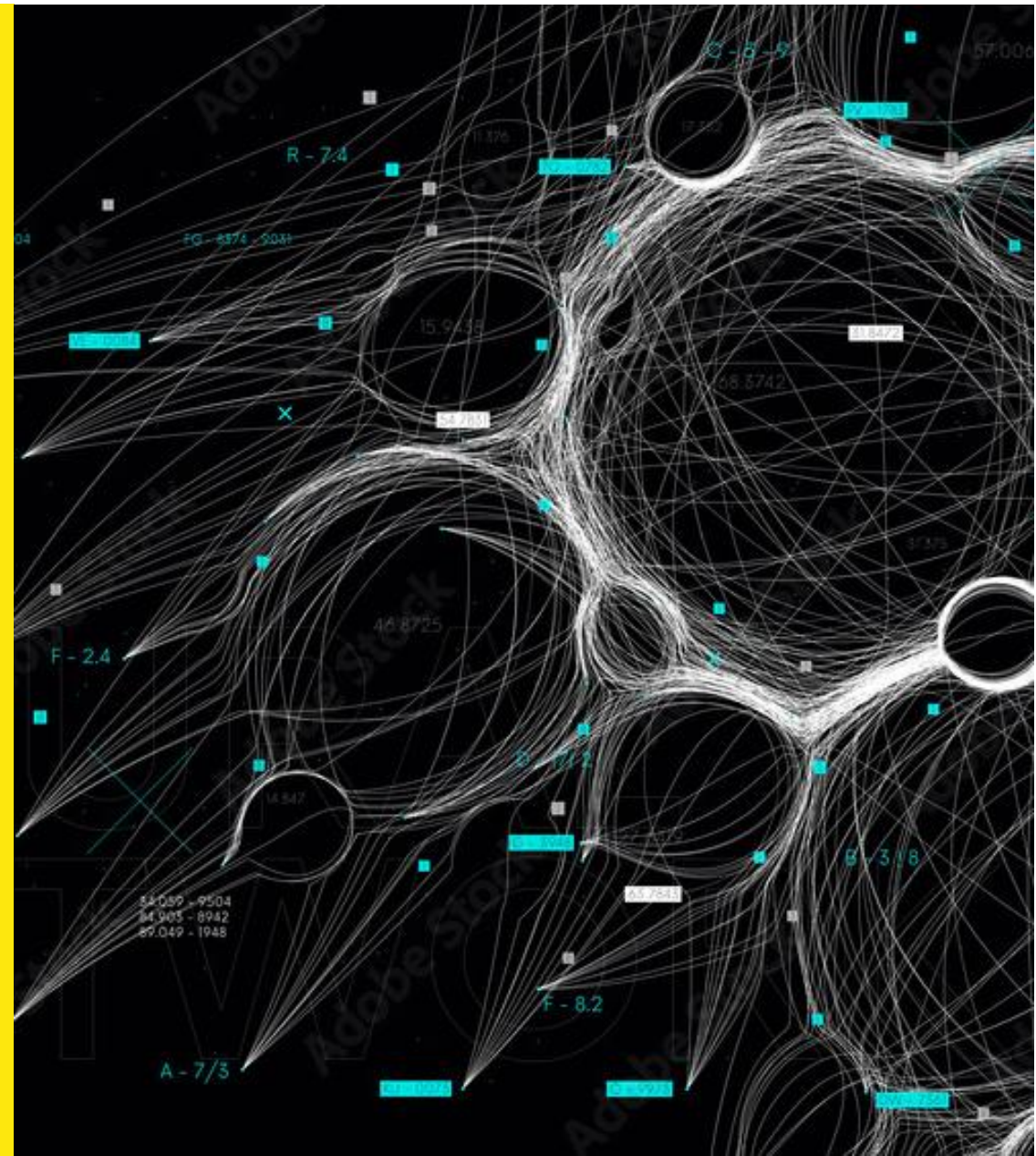


Die Digitalisierung des Controllings und Kompetenzfelder der Wirtschaftsinformatik

Prof. Dr. Maximilian Koch
Lars Fluri

4. Dezember 2023



Über mich



- Wissenschaftlicher Mitarbeiter
- Dozent (ab FS24)

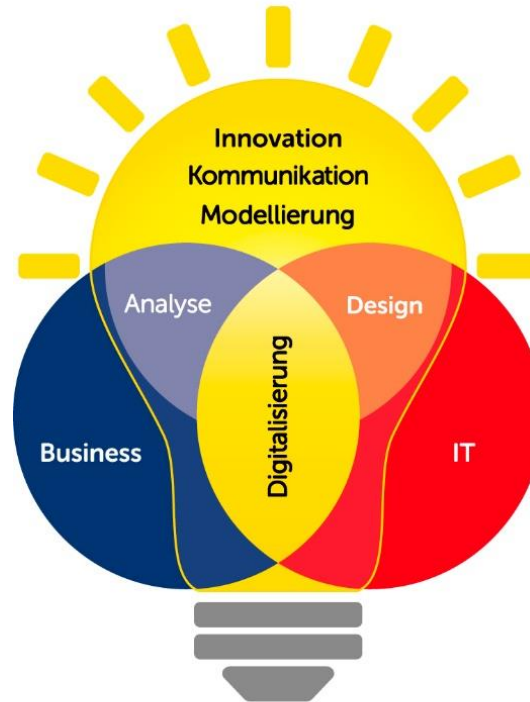


- Doktorand in Machine Learning
- Dozent



- Quantitative Researcher in Forschungsprojekt

Inspiration



© Hochschule für Wirtschaft FHNW
fhnw.ch/de/studium/wirtschaft/wi

Inspiration

Software Engineering Leadership

Requirements Engineer
IT-Projektmanager
Software Analyst

Business Analytics

Data Analyst
Data Scientist
Digital Marketing Manager
Business Development Manager

Digital Business Management

Programm- und Projektmanagement
IS- und IT-Businesskoordination
IS- und IT-Consulting
Cybersecurity Consulting
Business-Analyst
IS- und IT-Auditing

Was die Wirtschaftsinformatik so einzigartig macht: <https://www.fhnw.ch/de/studium/wirtschaft/wi>

Inspiration

Was muss bspw. ein Data Analyst können?

MIGROS

Data Analyst (w/m/d) 60-100%

Migros-Genossenschafts-Bund · Zurich, Zurich, Switzerland

1 week ago · 57 applicants



[See who Migros-Genossenschafts-Bund has hired for this role](#)

Apply ↗

Save

Was du bewegst

- Datenanalyse von un-/strukturierter Datensätzen sowie deren Aufbereitung für Präsentationen und Reports, um die strategische Entscheidungsfindung vorantreiben
- Identifizieren und Analysieren von relevanten Mustern und Zusammenhängern in Daten, Ableiten von Entscheidungsgrundlagen und Handlungsempfehlungen
- Entwerfen und Umsetzen automatischer Datenanalysen und Visualisierungen in Form von Datenprodukten
- Kontinuierliches Verbessern und Weiterentwickeln von quantitativen Auswertungen sowie Aufbereiten von Datenbeständen für die Ad-hoc Analysen

Was du mitbringst

- Bildung: Abgeschlossenes Studium (Uni/ETH, FH, HF)
- Berufserfahrung: Hands-on Berufserfahrung in der Datenmodellierung sowie im Aufbau, (Weiter-)Entwicklung und den Betrieb von Reports und Visualisierungen (Dashboards etc.) diverser Datenprodukte
- Dein Tech-Stack umfasst SQL, Python, Cloud Lösungen (z.B. GCP) und PowerBI
- Erfahrung in der Identifikation verschiedener Kennzahlen und KPIs sowie deren valide Messung
- Hohe Eigeninitiative und Kund*innen-Fokus
- Stark in der Vernetzung mit verschiedenen Business und IT Stakeholdern
- Deutsch (fließend)
- Englisch (sehr gute Kenntnisse)

Wichtige Themenfelder der (Wirtschafts-)Informatik

**Artificial Intelligence (Machine Learning,
Statistical Learning)**

Erstellung von produktionsreifen Modellen

Datenstrukturen

Aufbau und Organisation von Data Pipelines,
Data Lakes, Data Warehouses

Big Data

Verarbeitung von gemischten (strukturierten und
unstrukturierten) Daten
Verarbeitung und Instandhaltung von sehr
grossen Datenmengen

Ziel der Vorlesung

Anwendungsfelder / Business Cases aus der Praxis

- Wirtschaftliche Relevanz
- Bedarf nach digitalen Lösungen

Lösungen an der Schnittstelle von Wirtschaftswissenschaften und Informatik

- Technische Perspektive einnehmen
- Schnittstellen-Expertise einbringen

Kapitelübersicht

Predictive Forecasting für Budgetierung und Absatzplanung

Wie können wir unsere Budgetprognosen mithilfe von Predictive Forecasting verbessern?

Anwendung: Urban Connect

Fraud Detection in Controlling / Management Accounting

Wie können wir mögliche Betrugsversuche erkennen?

Clustering

Wie können wir Kundengruppen sinnvoll in verschiedene Cluster einteilen und daraus Marketing/Verkaufsmassnahmen ableiten?

Anwendung: Coop

Kapitelübersicht

Kapitel 1: Budgetierung

Was ist eine Budgetprognose?

Eine Schätzung der Absatzmenge in zukünftigen Geschäftsjahren, basierend auf heute verfügbaren Daten.

Weshalb benötigen wir Budgetierung und Budgetprognosen überhaupt?

- Schätzung von Liefermengen
- Schätzung des Personalbedarfs
- Gewinnschätzungen

Wie können wir digitale Tools einsetzen?

Grössere Datenmengen zwingen uns, digitale Kompetenzen zur Datenbewältigung und -analyse zu verwenden

4. Dezember 2023 Kompetenzfelder der Wirtschaftsinformatik www.fhnw.ch/wirtschaft 10

Fallstudie: Urban Connect

Urban Connect ist ein Provider für sog. «Corporate Mobility», also Mobilitätslösungen für Firmen.

4. Dezember 2023 Kompetenzfelder der Wirtschaftsinformatik www.fhnw.ch/wirtschaft 20

Kapitel 2: Fraud Detection

Betrugsversuche hinterlassen immer digitale Spuren. Wie finden wir diese?

SBF Made \$9 Billion Disappear. This Forensic Accountant Found It

A forensic financial expert broke down how \$2 billion vanished from the balance sheet of Wirecard, whose ex-CEO was just arrested

4. Dezember 2023 Kompetenzfelder der Wirtschaftsinformatik www.fhnw.ch/wirtschaft 30

Kapitel 3: Clustering

Senior Data Scientist MarTech

Digitales Marketing (MarTech) - Full-time - Associate

1,001-5,000 employees - Retail

6 company alumni work here - 14 school alumni work here

6 of 10 skills match your profile - you may be a good fit

View notifications related to this job post

About the job

The mission of the MarTech team is to develop the right content for the right channel at the right time. We work in a cross-functional and agile way. We have access to various marketing channels, including email, social media, and search engines. We use data to understand our audience and optimize our marketing efforts. We are looking for a Senior Data Scientist who can help us build a data-driven marketing strategy. The role involves working with large datasets, developing machine learning models, and collaborating with marketing teams to implement data-driven campaigns. The ideal candidate should have a strong background in statistics, machine learning, and data visualization. They should also have experience working in a fast-paced, agile environment. The role is based in Zurich, Switzerland, and requires a Master's degree in a relevant field. The salary is competitive and includes benefits. The role is open until the position is filled.

4. Dezember 2023 Kompetenzfelder der Wirtschaftsinformatik www.fhnw.ch/wirtschaft 42

Fallstudie: Coop

Coop möchte seine Kunden anhand ihres Konsumverhaltens in verschiedene Gruppen einteilen, um gezielter Werbung und Promotionen zu schaffen.

4. Dezember 2023 Kompetenzfelder der Wirtschaftsinformatik www.fhnw.ch/wirtschaft 51

Zusammenfassung

Predictive Forecasting

- Lineare / Nonlineare / ML-Modelle
- Overfitting / Underfitting: Generalisierbarkeit
- Vorgängige explorative Datenanalyse ist wichtig

Financial Fraud

- Strukturierte / unstrukturierte Daten
- Datenarchitektur
- Kosten / Nutzen-Abschätzungen

Clustering

- Agglomerative vs. Divisive models
- k-means
- Unsupervised Learning

4. Dezember 2023 Kompetenzfelder der Wirtschaftsinformatik www.fhnw.ch/wirtschaft 60

Kapitel 1: Budgetierung

Was ist eine Budgetprognose?

Eine Schätzung der Absatzmenge in zukünftigen Geschäftsjahren, basierend auf heute verfügbaren Daten.

Weshalb benötigen wir Budgetierung und Budgetprognosen überhaupt?

- Schätzung von Liefermengen
- Schätzung des Personalbedarfs
- Gewinnschätzungen

Wie können wir digitale Tools einsetzen?

Grössere Datenmengen zwingen uns, digitale Kompetenzen zur Datenbewältigung und -analyse zu verwenden

(Predictive) Forecasting

Predictive Forecasting

“Predictive Forecasting is an extension of classic forecasting. It considers a multitude of inputs, values, trends, cycles and fluctuations of the data in different business areas, to make predictions.”

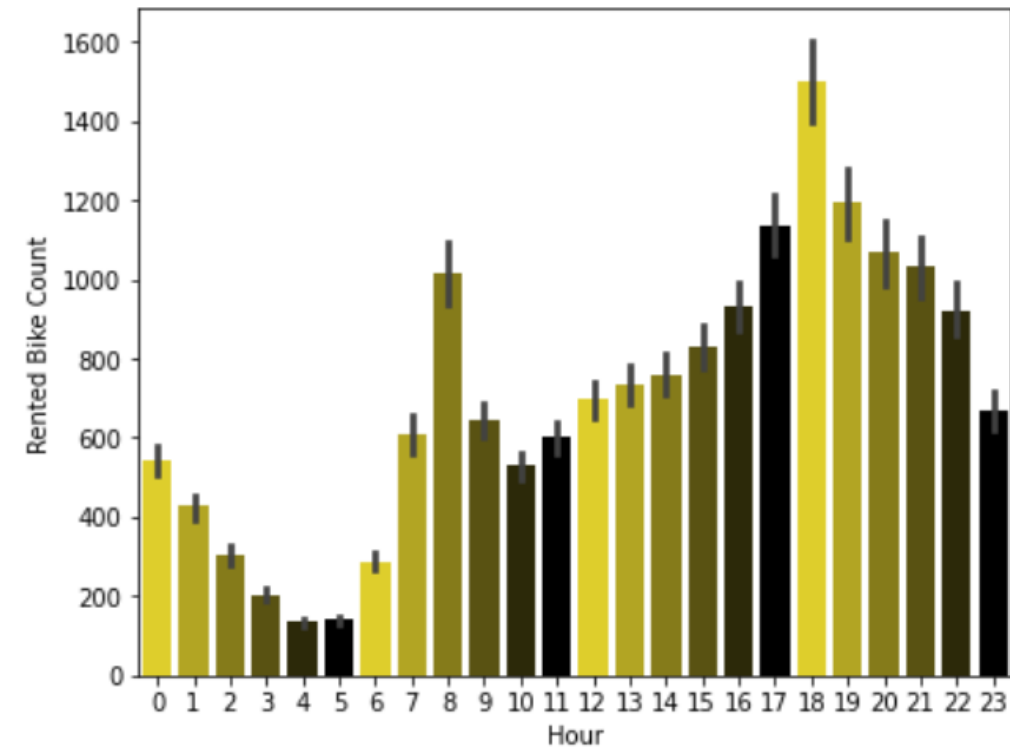
Ungewisse Zukunft

Wir wissen nicht, wie sich die Daten in der Zukunft entwickeln werden. Deshalb müssen wir Modelle mit vorhandenen (historischen) Daten schätzen.

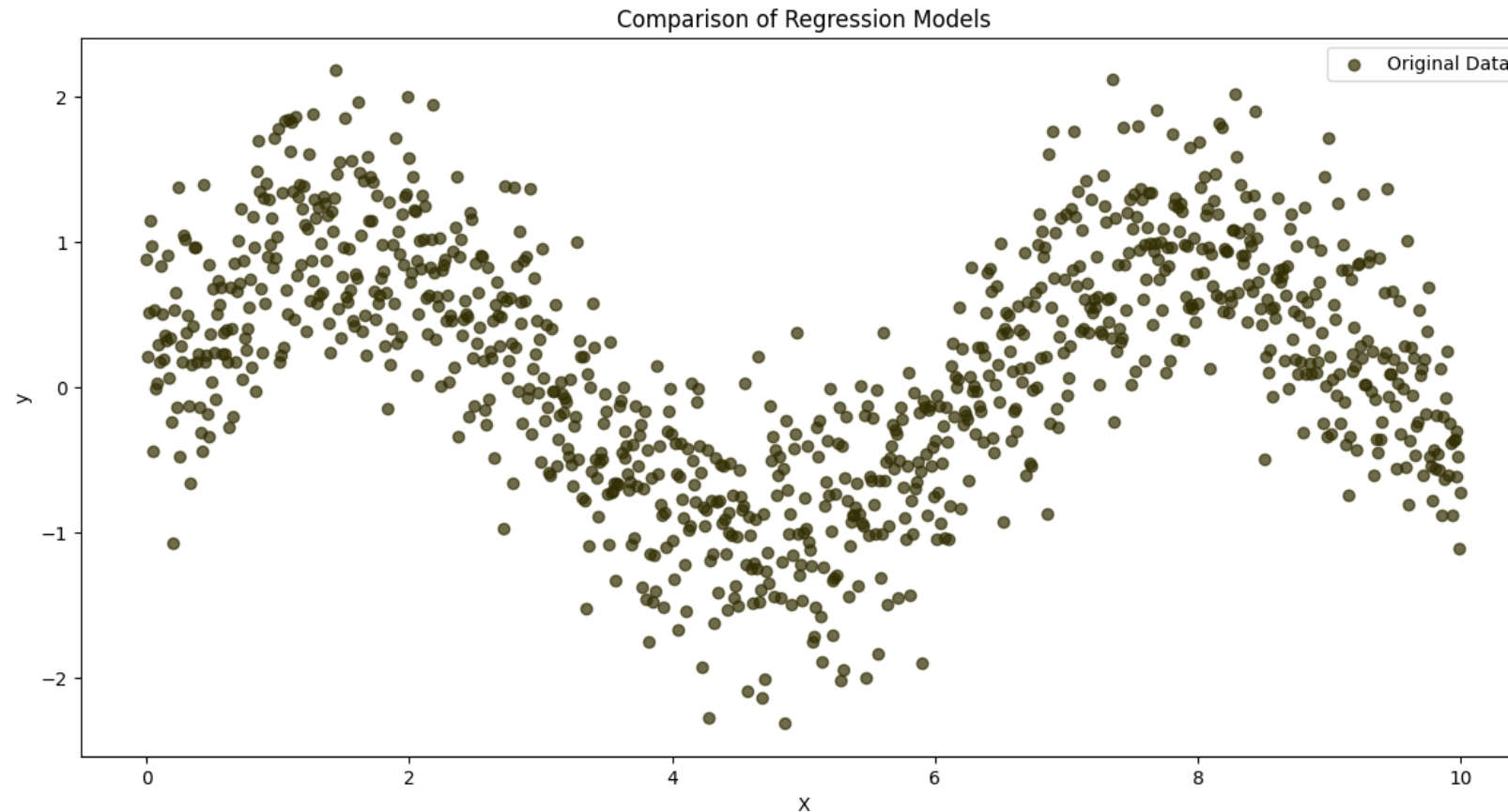
Komplexe Zusammenhänge

Zusammenhänge und Entwicklungen in der Zukunft folgen oft komplizierten Mustern. Einige Muster, welche wir beispielsweise in Zeitreihendaten erkennen können:

- Tageszeitabhängige Effekte
- Saisonale Effekte
- Jährliche Effekte
- «Zufällige» Effekte



Modellauswahl



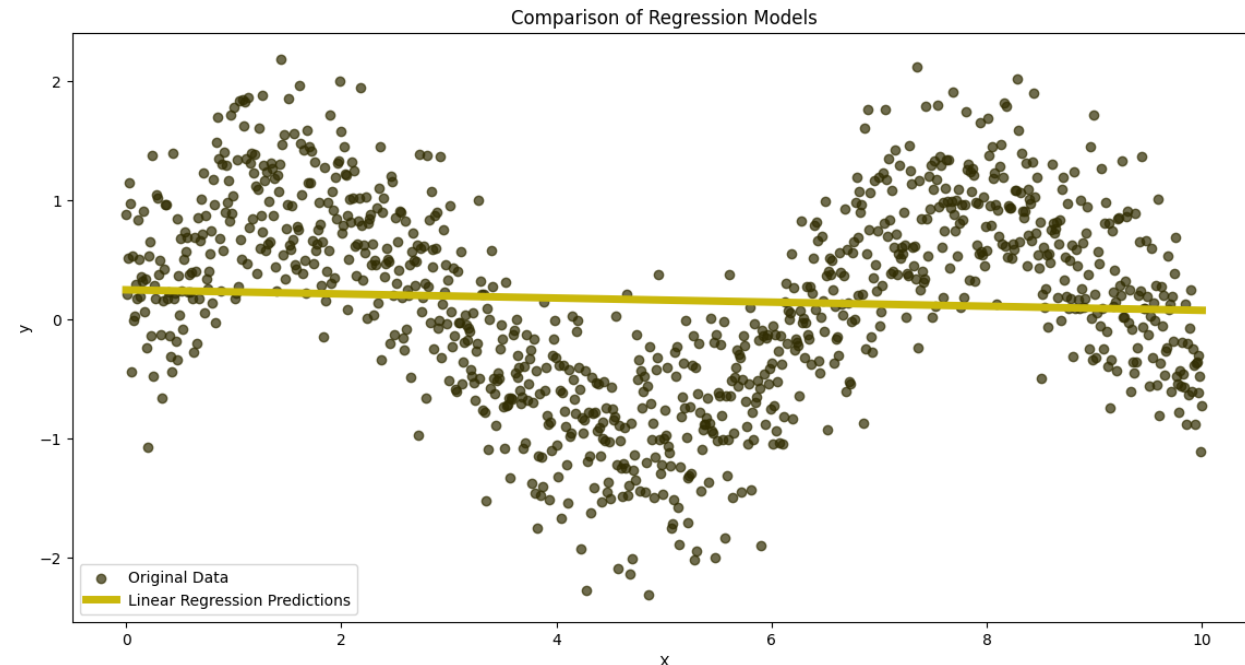
Modellauswahl

Lineare Modelle

Wir gehen davon aus, dass der Zusammenhang zwischen unseren Variablen linear ist.

Für die meisten Daten ist die eine sehr einschränkende Annahme.

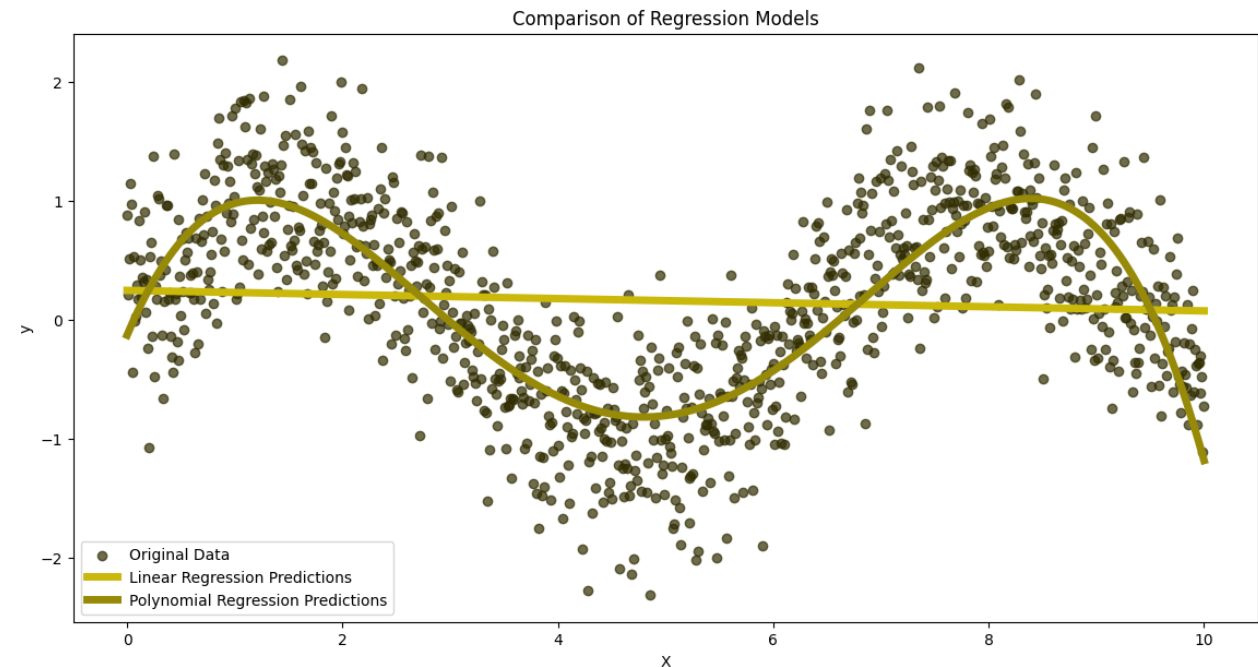
Wir brauchen ein flexibleres Modell.



Modellauswahl

Nichtlineare Modelle

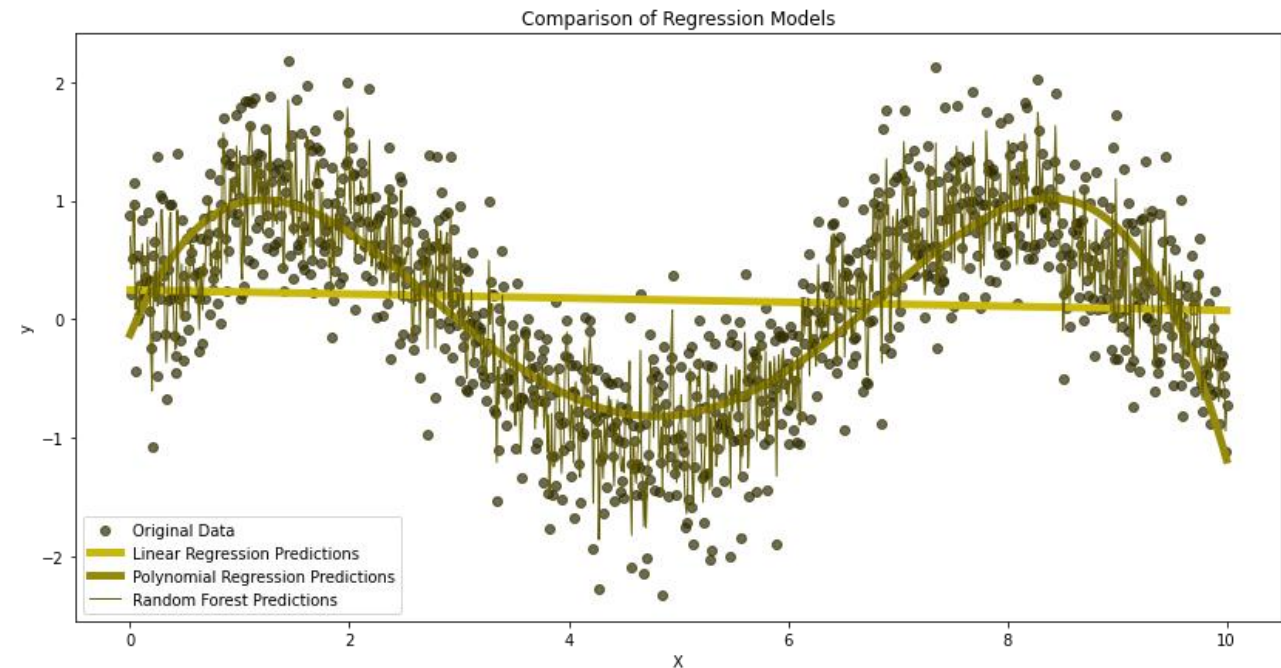
Nichtlineare (in diesem Fall parametrische) Modelle erlauben eine flexiblere Schätzung, da auch nichtlineare Terme berücksichtigt werden können.



Modellauswahl

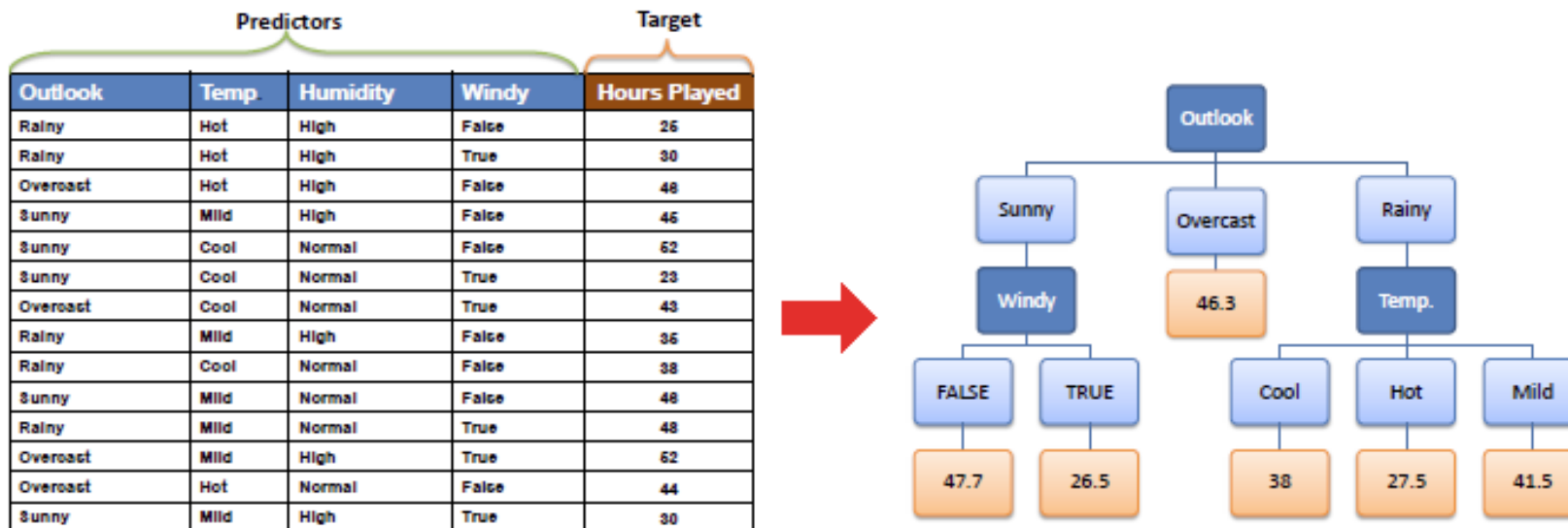
Nichtlineare Modelle

Nichtlineare (in diesem Fall parametrische) Modelle erlauben eine flexiblere Schätzung, da auch nichtlineare Terme berücksichtigt werden können.



Decision Trees

Wir teilen den Datensatz in Teildatensätze, um möglichst homogene Gruppen zu erhalten.



saedsayad.com/decision_tree_reg.htm#:~:text=Decision%20tree%20builds%20regression%20or,decision%20nodes%20and%20leaf%20nodes.

Decision Trees

Der Baum hat verschiedene Elemente:

- Decision Node, bspw. Outlook
- Leaf node: Hours played (Schätzung),
- Root node: Erster Decision Node, bester Schätzer

Die Schätzung für unsere Zielvariable ist der Mittelwert der Untergruppe des Datensatzes.

In diesem Beispiel: Bei sonnigem (sunny) Wetter und keinem Wind (windy == False) werden im Durchschnitt 47.7 Stunden gespielt.



saedsayad.com/decision_tree_reg.htm#:~:text=Decision%20tree%20builds%20regression%20or,decision%20nodes%20and%20leaf%20nodes

Modellzusammenfassung

Lineare Regression	Technik, die dazu dient, den Zusammenhang zwischen einer Zielgröße und einer oder mehreren anderen Größen durch eine gerade Linie zu beschreiben und Vorhersagen zu machen.
Polynomregression	Erweiterte Form der linearen Regression, die nicht-lineare Beziehungen zwischen einer Zielgröße und einer oder mehreren anderen Größen durch eine Kurve statt einer geraden Linie abbildet.
Decision Trees	Modell im maschinellen Lernen, das Daten in einer Baumstruktur klassifiziert oder vorhersagt, indem es schrittweise Entscheidungen aufgrund von Merkmalen der Daten trifft.
Random Forests	Random Forests sind ein Ensemble-Lernverfahren, das viele Entscheidungsbäume kombiniert, um robustere und genauere Vorhersagen und Klassifizierungen zu erzielen.

Fallstudie: Urban Connect

Urban Connect ist ein Provider für sog. «Corporate Mobility», also Mobilitätslösungen für Firmen.



Nachfrage nach Fahrrädern

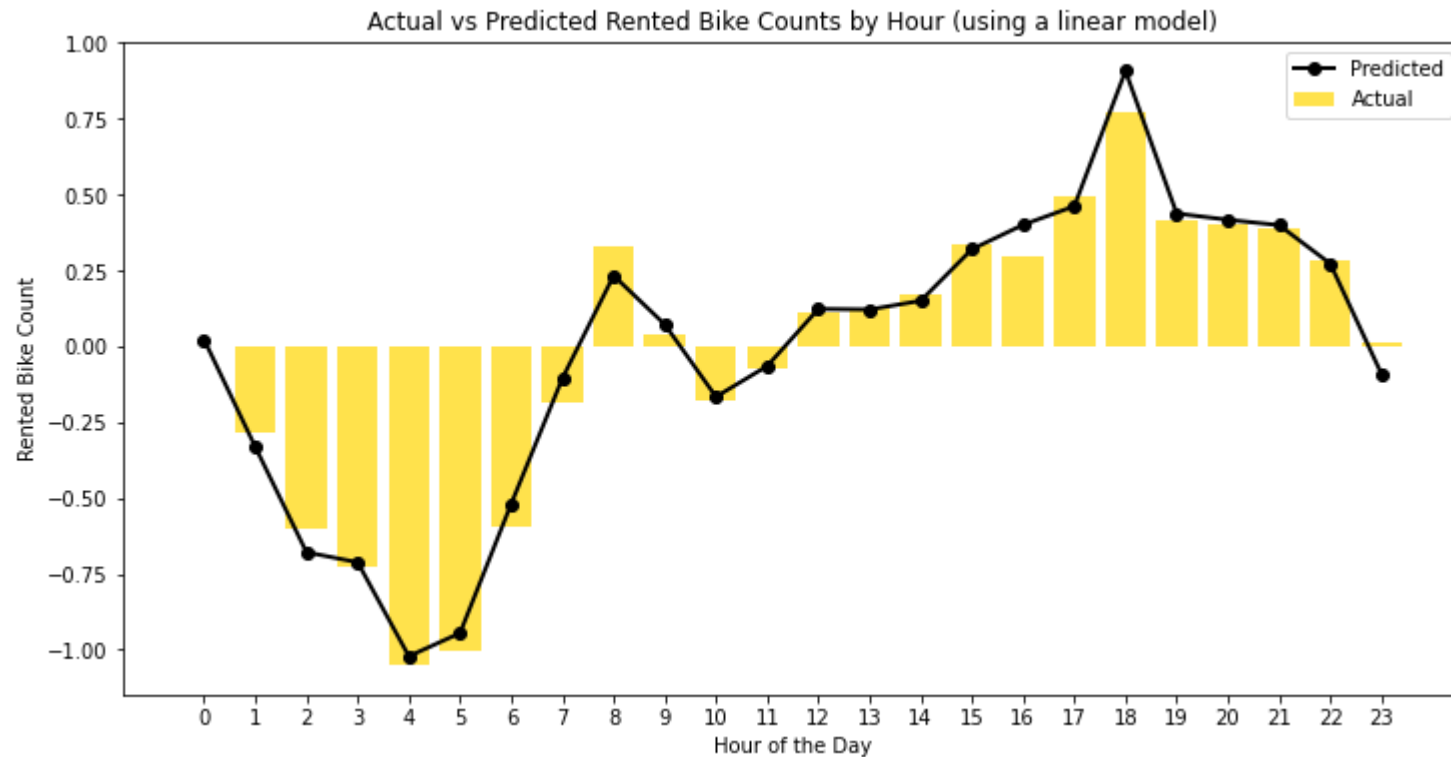
Fragestellung Wie gross ist die Nachfrage nach Fahrrädern an einem gegebenen Tag?

Relevante Merkmale

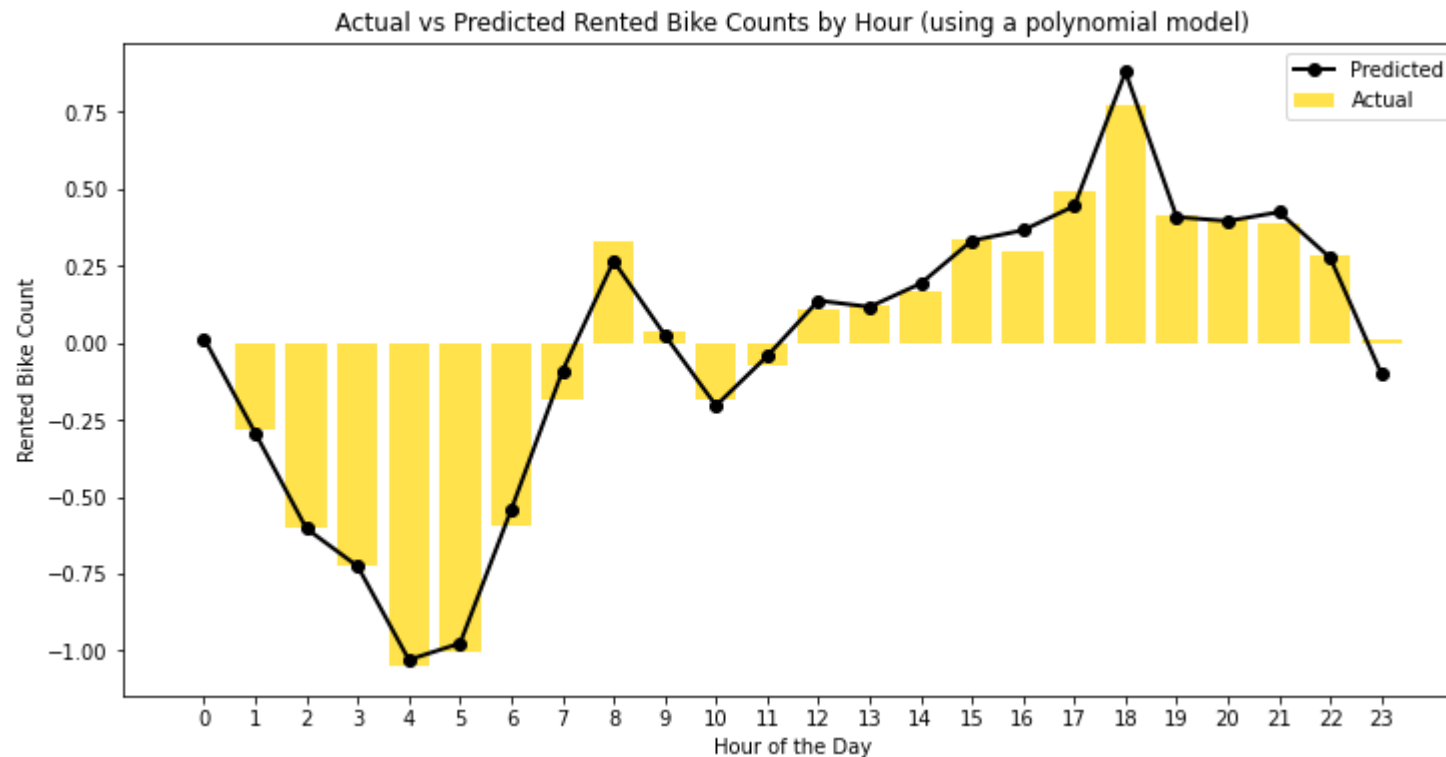
- Tageszeit
- Temperatur
- Sonneneinstrahlung
- Schnee / Regen

Generalisierbarkeit Das Modell sollte möglichst gut Out-of-Sample passen.

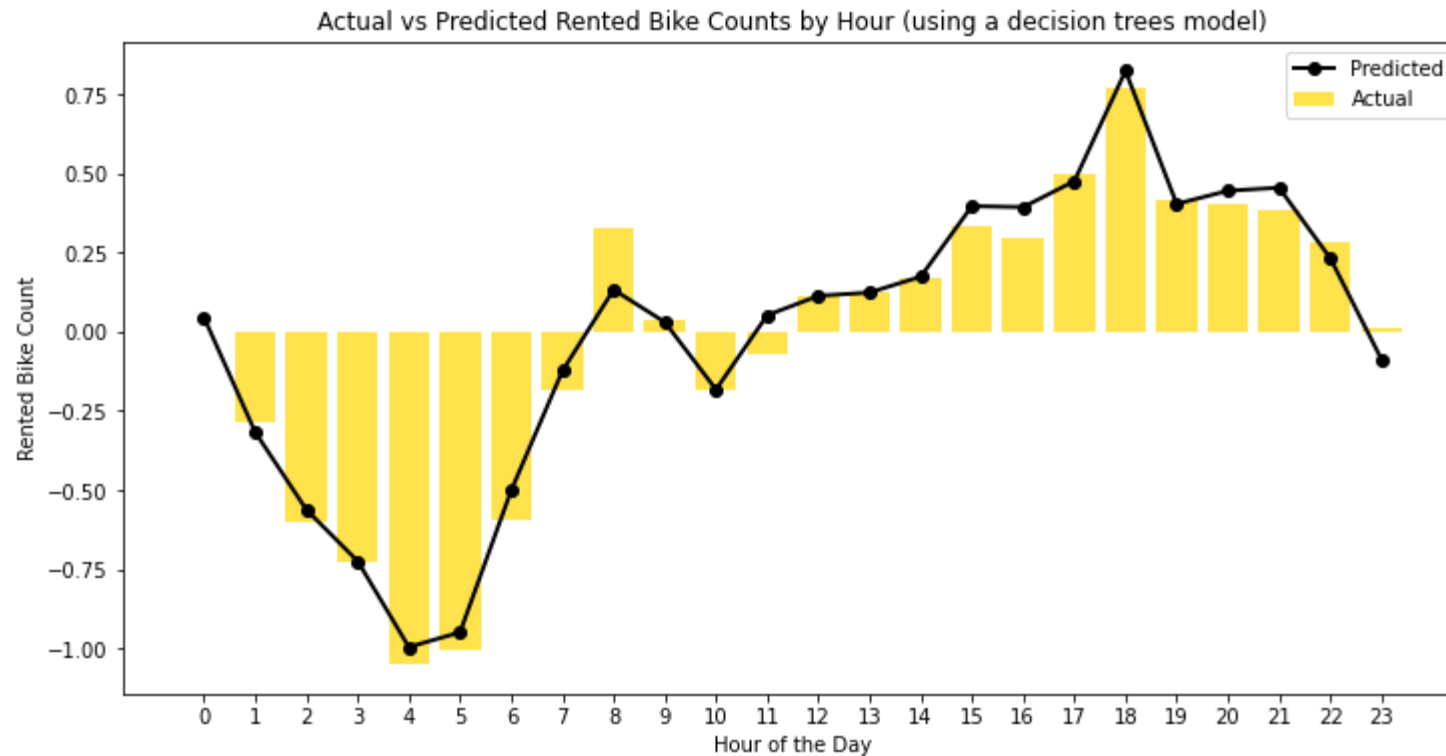
Prognose mithilfe eines linearen Modells



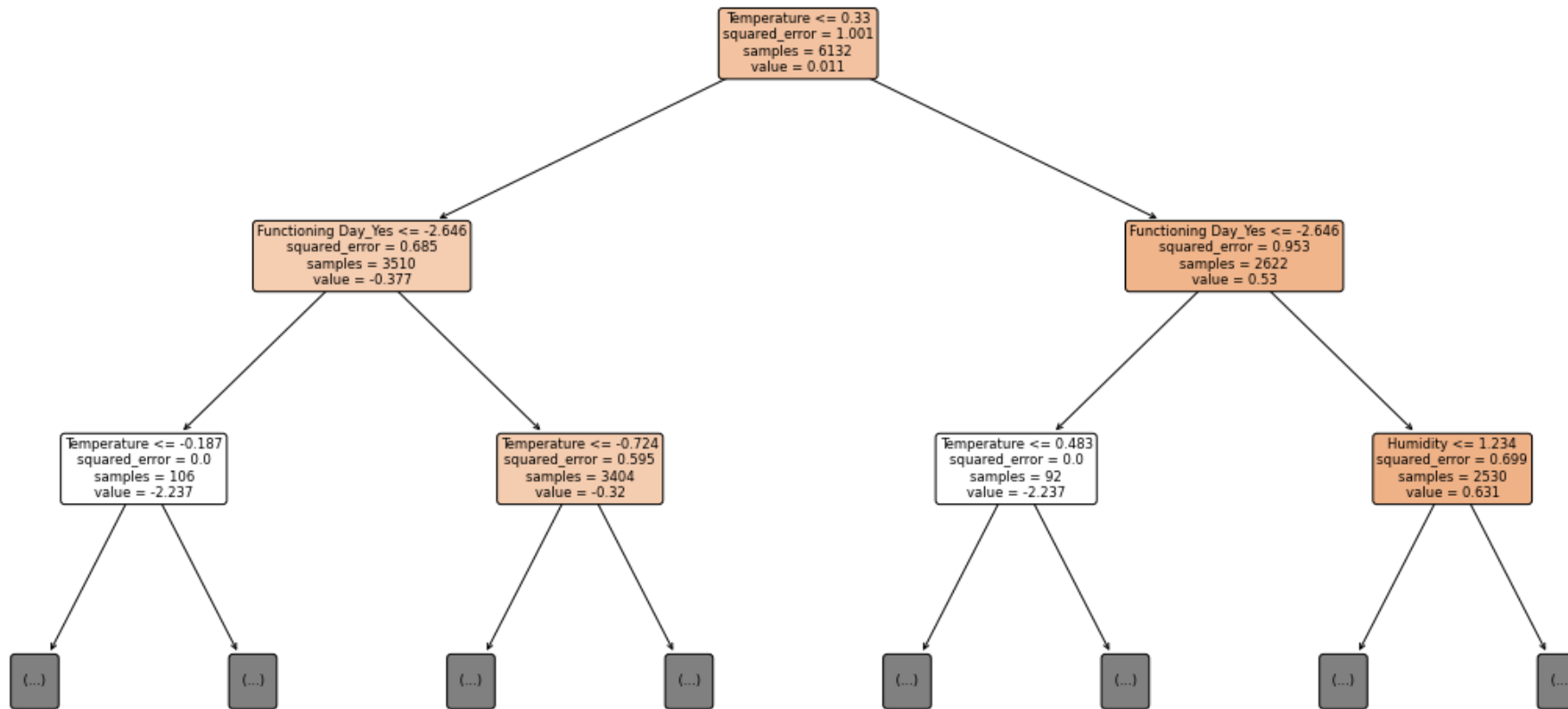
Prognose mithilfe eines Polynommodells



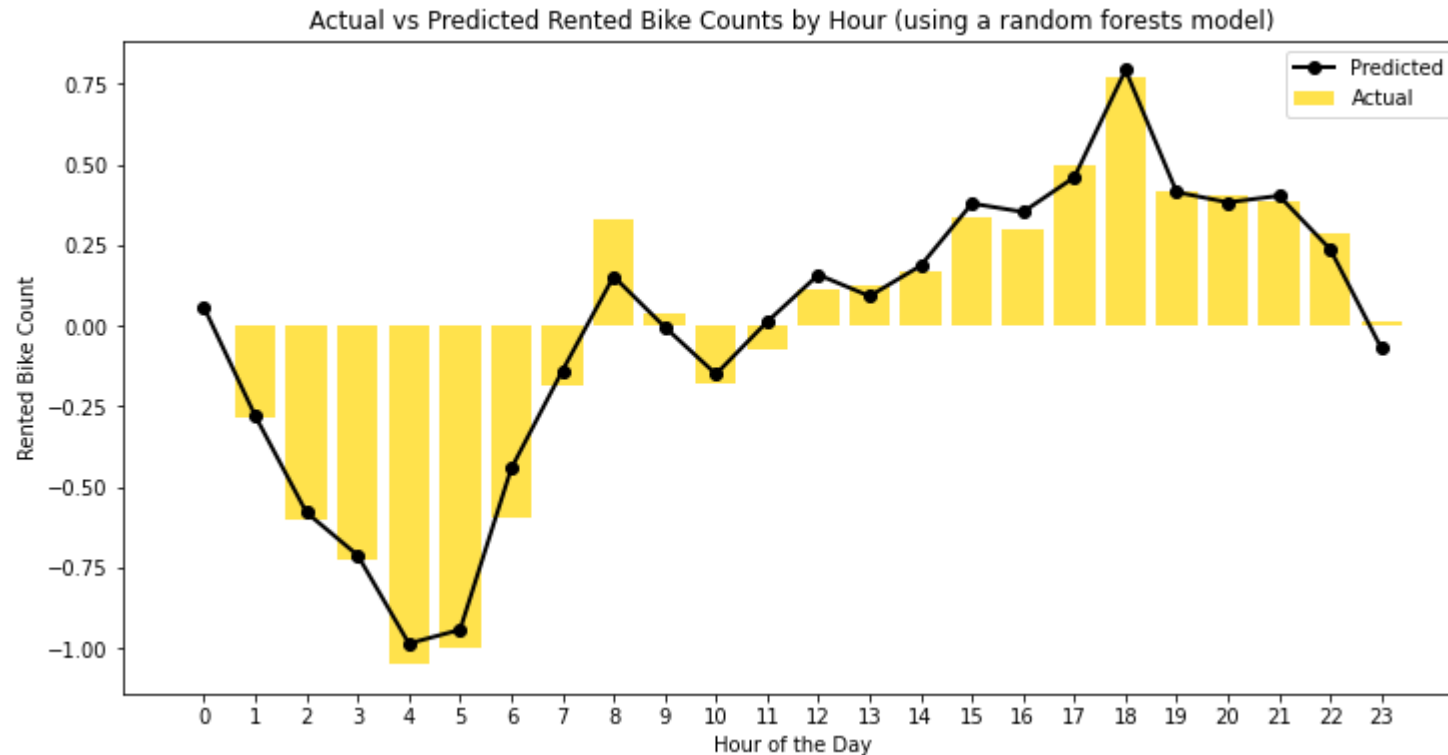
Prognose mithilfe eines Decision Tree-Models



Prognose mithilfe eines Decision Tree-Models



Prognose mithilfe eines Random Forest-Modells



Urban Connect: Resultate

Wir erreichen die beste Performance nicht unbedingt immer mit dem kompliziertesten Modell.

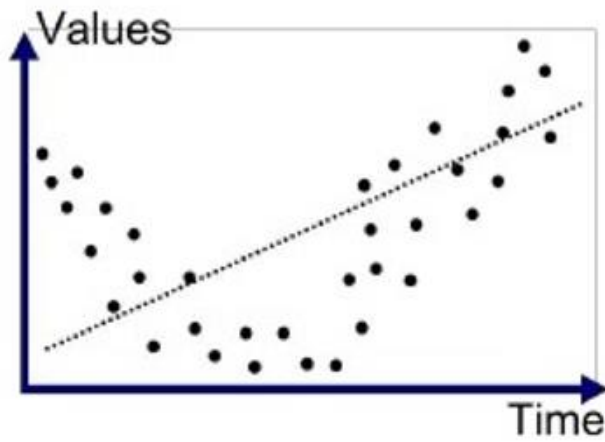
	R-Square	MSE	CV Accuracy	CV std
Polynomial Regression	88.88%	11.07%	80.38%	0.60%
Random forest Regression	87.40%	12.56%	88.31%	0.73%
Linear Regression	80.83%	19.10%	80.38%	0.60%
Decision Tree Regression	79.20%	20.72%	80.05%	1.57%

Maximum Score in each Column				
	R-Square	MSE	CV Accuracy	CV std
Polynomial Regression	88.88%	11.07%	80.38%	0.60%
Random forest Regression	87.40%	12.56%	88.31%	0.73%
Linear Regression	80.83%	19.10%	80.38%	0.60%
Decision Tree Regression	79.20%	20.72%	80.05%	1.57%

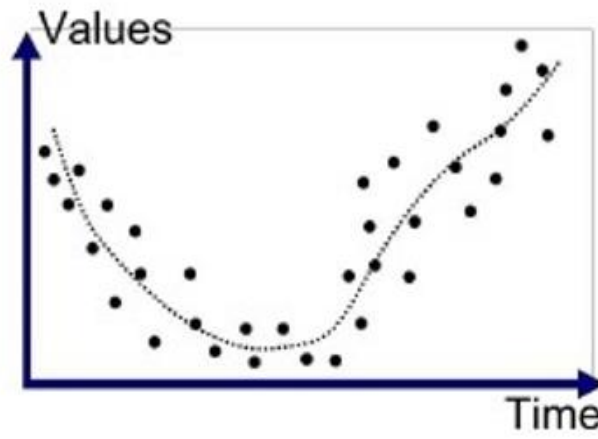
Weshalb ist dies der Fall?

Urban Connect: Resultate

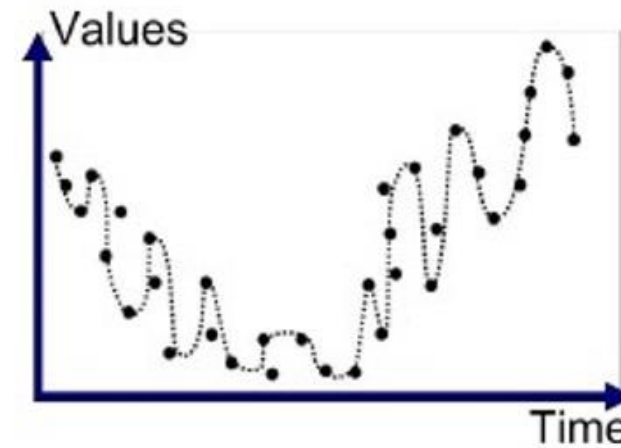
Sogenanntes Overfitting:



Underfitted



Good Fit/Robust



Overfitted

medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

Predictive Forecasting: Key Takeaways

Datenkomplexität

Zusammenhänge sind in der Praxis meistens nicht linear

Relevante Merkmale

- Komplexe Modelle: Höhere Rechenzeit
- Einfachere Modelle: Berücksichtigen nicht alle wichtigen Zusammenhänge

Generalisierbarkeit

Modelle sollten auch ausserhalb der Trainingsdaten gute Prognosen liefern können.
Overfitting vermeiden

Kapitel 2: Fraud Detection

Betrugsversuche hinterlassen immer digitale Spuren. Wie finden wir diese?

SBF Made \$9 Billion Disappear. This Forensic Accountant Found It

This guy is so good, FTX should've hired him years ago.

By **Maxwell Zeff** Published October 19, 2023 | Comments (60)



Photo: Michael M. Santiago (Getty Images)

gizmodo.com/sbf-made-billions-disappear-forensic-accountant-found-1850942819

HOME > NEWS > STOCKS

A forensic financial expert broke down how \$2 billion vanished from the balance sheet of Wirecard, whose ex-CEO was just arrested

Shalini Nagarajan Jun 23, 2020, 4:47 PM CEST



Reuters

Your Market View

NAME / PRICE	+ / -	%	DATE
▲ TSLA 241.49	-2.65	-1.09%	11/30/23 4:43 PM
▲ AAPL 188.66	-0.71	-0.38%	11/30/23 4:43 PM
▲ MSFT 377.53	-1.32	-0.35%	11/30/23 4:43 PM
▲ NFLX 473.58	-3.61	-0.76%	11/30/23 4:43 PM
▲ SPOT 181.24	-1.07	-0.58%	11/30/23 4:43 PM

markets.businessinsider.com/news/stocks/wirecard-scandal-numbers-financial-forensic-expert-breakdown-2020-6-1029332810

Was ist ein Forensic Accountant / Consultant?

The Forensic Technology & eDiscovery team is a group of technical specialists that leverages data and technology to investigate high-profile financial crime matters related to fraud, corruption, money-laundering, misconducts and support the enforcement of regulatory requirements.

The nature of our work requires the collection, processing and management of large sets of communications, documents and records from a wide array of information systems. We combine deep investigation expertise with Forensic and eDiscovery technology to accelerate the fact-finding process.

careers.ey.com/ey/job/Zurich-Consultant-Forensic-Technology-and-eDiscovery-8005/783598701/)

Fraud Detection

Sicherstellung Integrität

Integrität von Bilanzzahlen ist von entscheidender Bedeutung für das Vertrauen von Stakeholdern und für fundierte Geschäftsentscheidungen

Unterscheidung legitim vs. illegitim

Verlässliche Unterscheidung von legitimen Transaktionen / Handlungen und illegitimen

Verarbeitung von grossen Datenmengen

Aufbau von effizienten Hardware- und Softwarelösungen

Das Problem von Investigations und Fraud Detection

Big Data

Datenbanken werden immer grösser

Gemischte Datenstrukturen

gemischte Formen von Daten (strukturiert / unstrukturiert)

Datenaufbereitung

Aufbau von Pipelines und Systemen, um Daten effizient abzurufen und zu analysieren

Tools

Analoge Auswertung von Daten ist nicht umsetzbar. Verwendete Methoden: Lineare Regression, logistische Regression, Decision Trees, Neural Networks, NLP

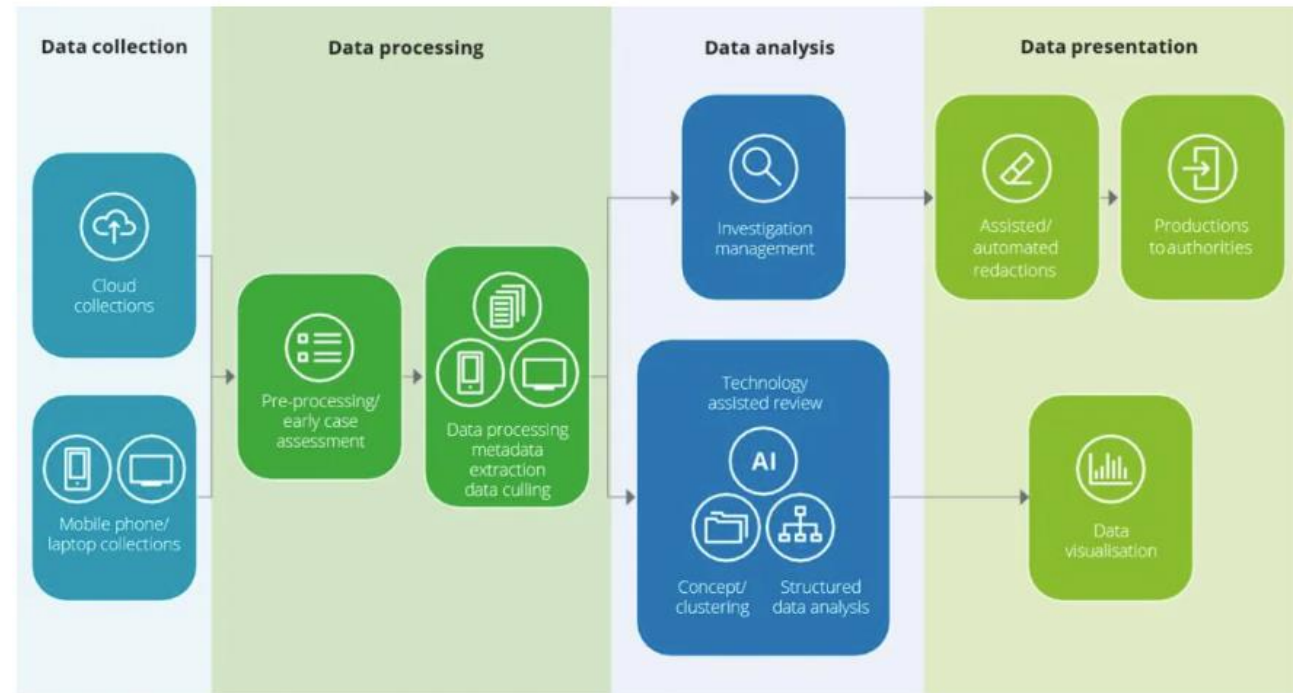
[kpmg.com/pl/en/home/insights/2021/02/comprehensive-fraud-detection-and-verification-process.html#:~:text=Benford's%20law%20considers%20the%20distribution,suspicious%20or%20possibly%20manipulated%20data.](https://www.kpmg.com/pl/en/home/insights/2021/02/comprehensive-fraud-detection-and-verification-process.html#:~:text=Benford's%20law%20considers%20the%20distribution,suspicious%20or%20possibly%20manipulated%20data.)

Crash Course eDiscovery

Digitale Untersuchung, um Beweise in E-Mails, Geschäftskommunikation und anderen Daten zu finden, die in Gerichtsverfahren oder strafrechtlichen Verfahren verwendet werden könnten.

Verwendete Methoden aus Ihrem Studium:

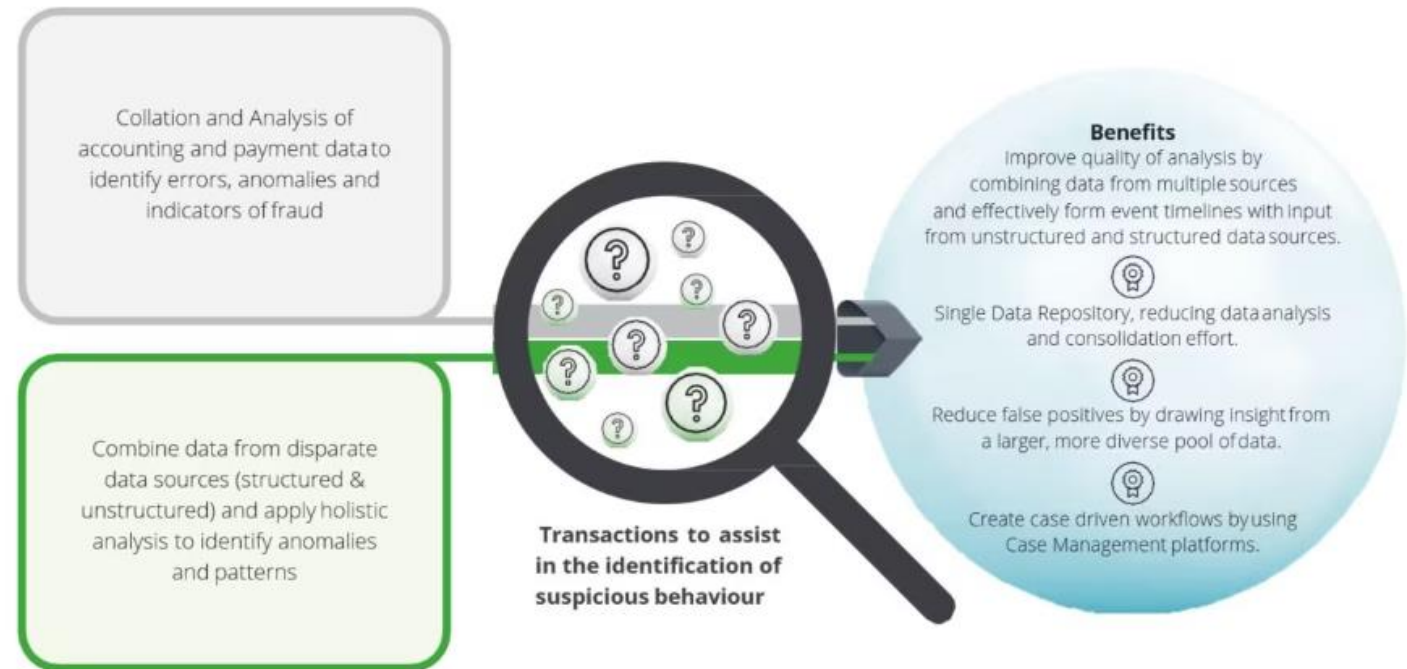
- Datenbanken: Data Warehousing, Data Lakes, Data Pipelines
- Cloud Engineering
- Data Science (Natural Language Processing, Data Visualisation)
- Machine Learning
- Data Visualisation



www2.deloitte.com/ch/en/pages/forensics/solutions/forensic-technology.html

Wie entdeckt man Financial Crime?

- Grosse Datenmengen sammeln, einlesen, und verarbeiten
- Holistische Analyse zur Erkennung von Mustern und Anomalien
- Verwendung von NLP und anderen ML-Methoden



www2.deloitte.com/ch/en/pages/forensics/solutions/forensic-technology.html

Ein kurioses Beispiel: Benfords Law

Verteilung von ersten Ziffern in einer Zahl.

Der Kanadisch-amerikanische Astronom Simon Newcomb entdeckte im Jahr 1881, dass frühere Seiten in Logarithmenbüchern (beginnend mit 1) viel häufiger benutzt wurden als andere.

Newcomb leitet daraus die Hypothese ab, dass die ersten Ziffern einer Zahl einer bestimmten Verteilung folgen sollten.



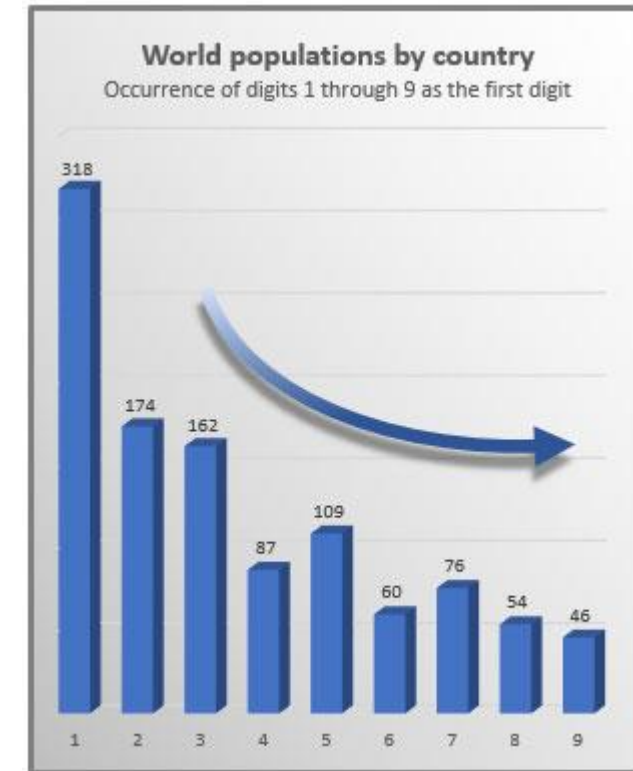
en.wikipedia.org/wiki/Napierian_logarithm

Welche Daten folgen Benfords Law?

Einige spannende (und überraschende) Anwendungen finden Sie [hier](#).

Die Daten müssen folgende Regeln erfüllen:

- Numerische Daten
- Zufällig generierte Zahlen (keine Beschränkung durch Minimum oder Maximum)
- Grosse Datensätze
- Zahlen steigen auf durch 10,100,1000, etc.



<https://www.journalofaccountancy.com/issues/2017/apr/excel-and-benfords-law-to-detect-fraud.html>

Beispiel: Benfords Law in Accounting and Finance

Einsatz im Accounting

Benfords Law kann auch eingesetzt werden, um Betrug oder Betrugsversuche in Accounting-Daten zu erkennen

Erkennung von Betrug

Wenn die Ziffern in der Jahresrechnung (oder anderen Dokumenten) nicht der Verteilung folgen, kann dies ein Indikator sein, dass Finanzaufgaben gefälscht oder bearbeitet wurden

Ausnahmen

Viele Zahlen sind nicht unbedingt zufällig und folgen deshalb nicht Benfords Law.

Beispiele

Preissetzung von Artikeln
Absatzmengen

Beispiel: Benfords Law in Accounting and Finance

- Bewerber für staatliche Unterstützungsgelder beantragen finanzielle Hilfe.
- Diese Hilfgelder sind häufig an Bedingungen gebunden, bspw. dass das Einkommen eine gewisse Schwelle nicht überschneidet.
- Hier entsteht der Anreiz, das eigene Einkommen zu fälschen. Dies kann somit genauer analysiert werden.



kpmg.com/pl/en/home/insights/2021/02/comprehensive-fraud-detection-and-verification-process.html#:~:text=Benford's%20law%20considers%20the%20distribution,suspicious%20or%20possibly%20manipulated%20data.

Grenzen

Kosten / Nutzen

Nutzen / Einsparung durch Entdeckung Betrug sollte grösser sein als Implementierungskosten

Ethical AI / Ethical Data Science

Sicherstellen, dass Modelle keinen unethischen Bias in den Daten aufnehmen

«What you see is all there is»

Nur analysierbar, was in den Daten vorhanden ist

Financial Fraud und eDiscovery: Key Takeaways

Vielseitige Applikation

Wir benötigen Erfahrung um Datenbanken, Data Science, Machine Learning

Nutzen / Kosten

Nutzen sollen Kosten überwiegen

Interdisziplinäre Kenntnisse

Es reicht nicht aus, Modelle zu implementieren. Wirtschaftliche Kenntnisse müssen auch eingearbeitet werden.

Kapitel 3: Clustering



Senior Data Scientist MarTech

Digitec Galaxus AG · Zurich, Switzerland **1 hour ago** · 6 applicants



Hybrid · Full-time · Associate



1,001-5,000 employees · Retail



6 company alumni work here · 19 school alumni work here



6 of 10 skills match your profile - you may be a good fit



View verifications related to this job post. [Show all](#)

About the job

The mission of the MarTech team is to display the right content for the right people on every marketing channel at all times. We work on a wide range of products and features that reach several million users every month. Additionally, we constantly challenge the status quo and explore new personalization approaches in order to get closer to our mission step by step. As part of the online shop's BI team, you will work closely with the MarTech team and other experts in the field of business intelligence and data science.

Tasks

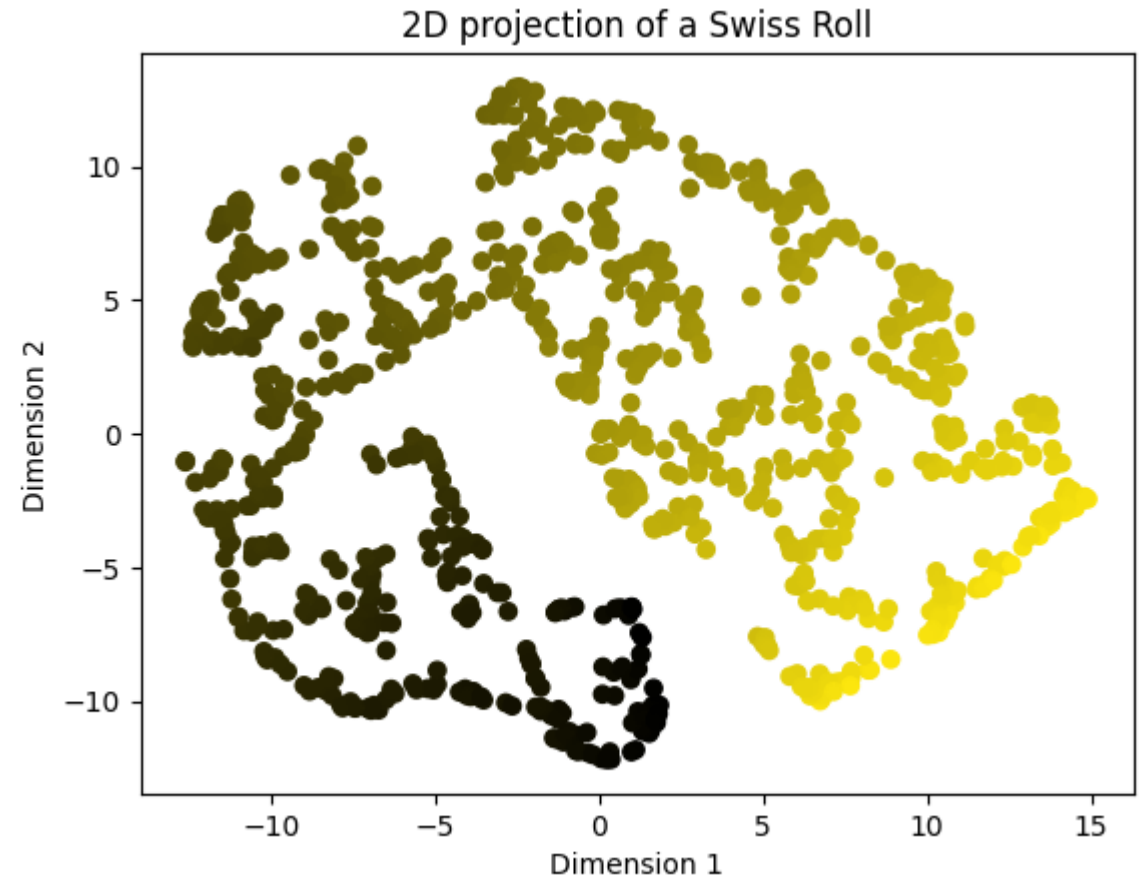
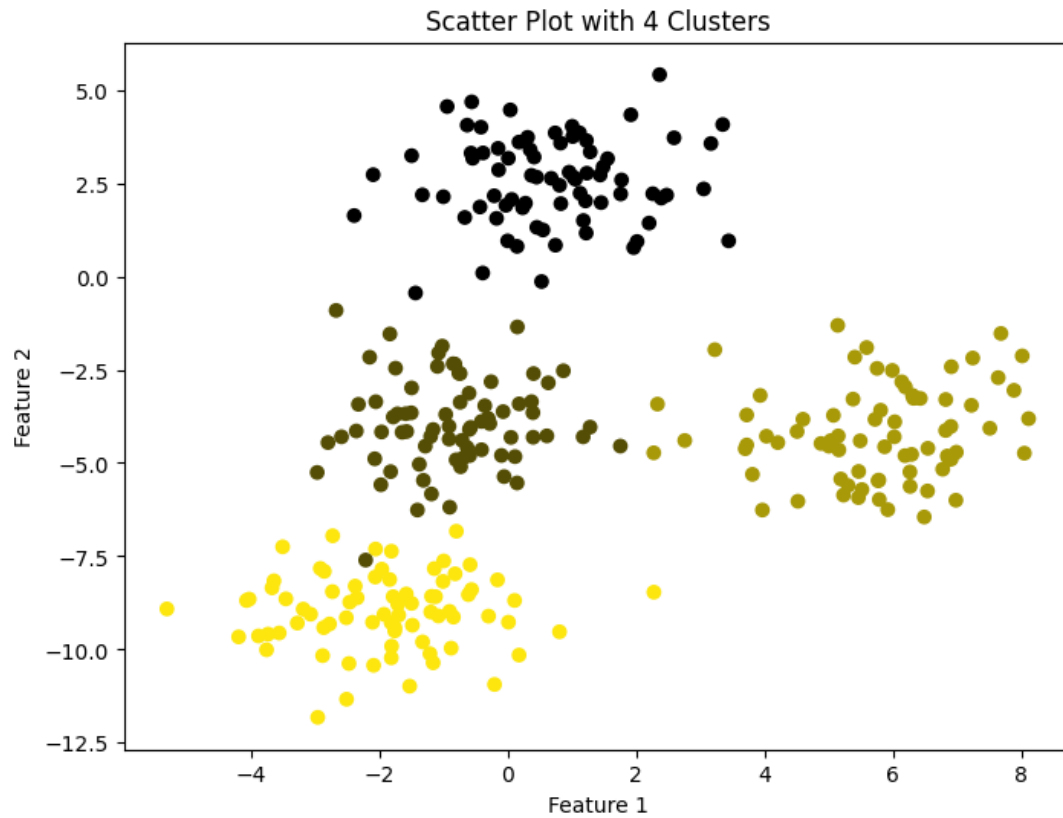
- Create and analyze large data sets to gain insights and develop new models.
- Develop existing models in the MarTech area and create prototypes for new models, which will be brought into production together with the product team.
- Use advanced statistical analyses to enable the product team to learn quickly from experiments and data.
- Drive significant marketing impact with control theory application and machine learning methods.
- Develop and manage success metrics in close collaboration with the Product Owner for MarTech.
- Act as an interface between the MarTech team and the online shop BI team.

Requirements

- You have a background in data science and/or control engineering and at least 2 years of practical experience in a relevant professional field
- You have some experience with controller design (MPC, LQR, DDP, etc.)
- You have basic knowledge of machine learning using libraries such as Scikit-Learn, TensorFlow, and PyTorch. You are proficient in at least one scripting language (e.g. Python, Scala, R)
- You can handle large datasets and have advanced knowledge of SQL and database systems (e.g. BigQuery).
- You are competent in statistics and well-versed in topics such as statistical modeling, A/B testing, significance tests, etc.
- You can articulate and communicate well, are open to working closely with an interdisciplinary Scrum team, and act proactively.

Requis

Kapitel 3: Clustering



Methodik und Relevanz

Controlling / Marketing

Wir versuchen, Beobachtungen (in unserem Fall Kunden) möglichst gut in Gruppen einzuteilen.

Goodness-of-Fit

Beobachtungen sollen in einer Gruppe möglichst ähnlich, zwischen den Gruppen möglichst unähnlich sein. Dies verstehen wir als internal cohesion (homogeneity), external isolation (separation)

Vielseitige Anwendungsfelder

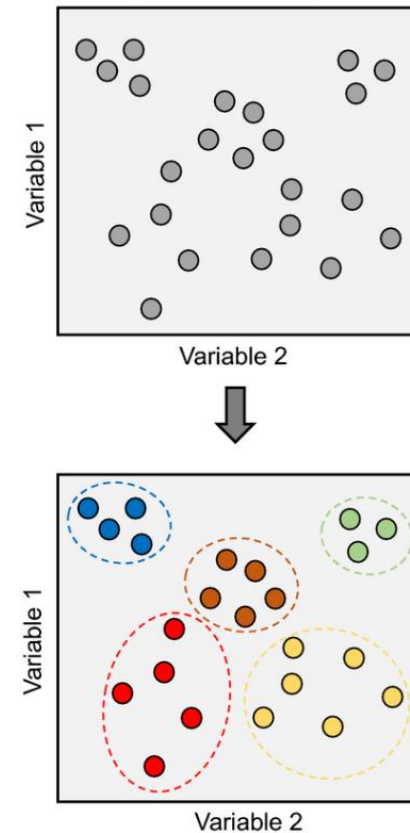
- Kreditwürdigkeitsschätzung
- Targeted Ads: Markt– und Kundensegmentierung
- Supply Chain Optimierung: Zulieferer in Risikogruppen unterteilen
- Konkurrenzanalyse
- Fraud Detection / Transaktionsanalyse (vgl. vorheriges Kapitel)

Unsupervised Learning

Es gibt keine richtige Antwort: sog. Unsupervised learning problem

Die Anzahl der Gruppen / die Zugehörigkeit zu den Gruppen ist nicht objektiv bestimmbar.

Clustering ist somit eher eine «Kunst», kein Handwerk



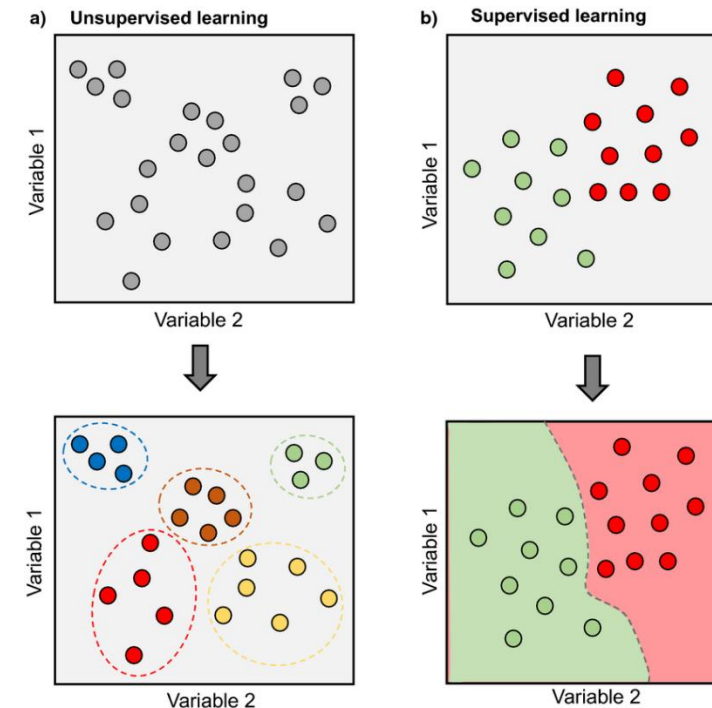
researchgate.net/figure/Supervised-and-unsupervised-machine-learning-a-Schematic-representation-of-an_fig3_351953193

Unsupervised Learning

Es gibt keine richtige Antwort: sog. Unsupervised learning problem

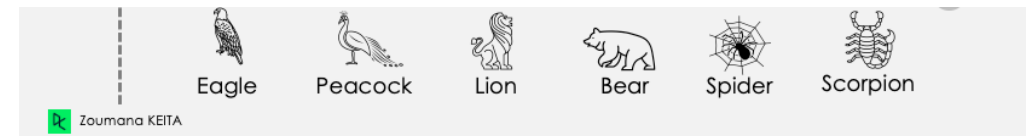
Die Anzahl der Gruppen / die Zugehörigkeit zu den Gruppen ist nicht objektiv bestimmbar.

Clustering ist somit eher eine «Kunst», kein Handwerk



researchgate.net/figure/Supervised-and-unsupervised-machine-learning-a-Schematic-representation-of-an_fig3_351953193

Hierarchisches Clustering

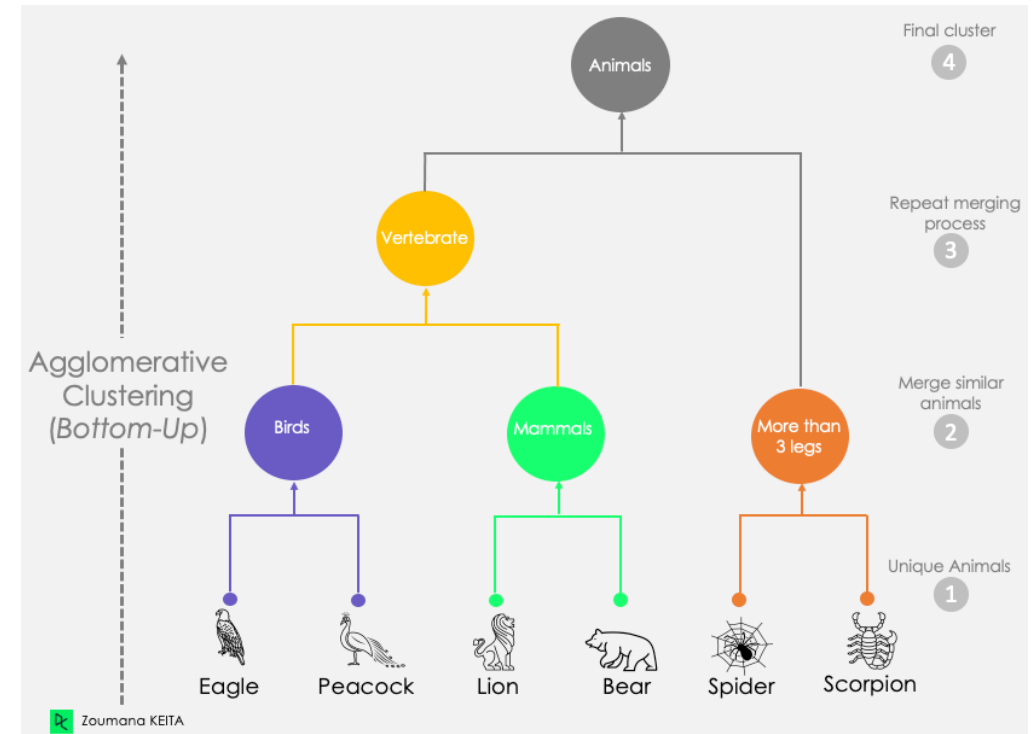


www.datacamp.com/tutorial/introduction-hierarchical-clustering-python

Hierarchisches Clustering

Wir können entweder:

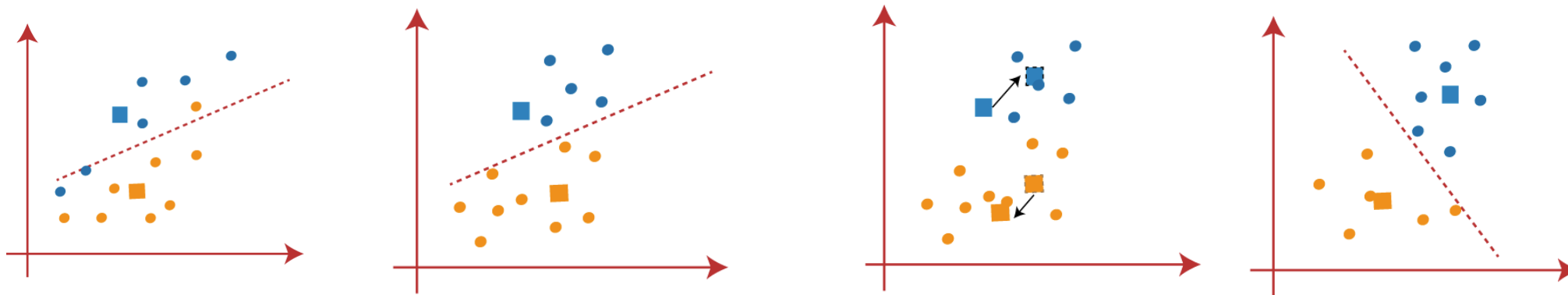
- mit einer grossen Gruppe starten und dann teilen (hierarchical divisive clustering)
- mit einer einzigen Beobachtung starten und dann Gruppen verbinden (agglomerative divisive clustering)



www.datacamp.com/tutorial/introduction-hierarchical-clustering-python

K-means Clustering

Teile die Datensets in K unterschiedliche, nicht-überlappende Cluster, indem die Varianz innerhalb der Cluster minimiert wird.



<https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/>

Methodenauswahl

AGNES (AGglomerative NESTing)

Wir starten mit n Clustern (bei n Beobachtungen) und fügen Beobachtungen in Gruppen zusammen.

DIANA (DIvisive ANALysis clustering)

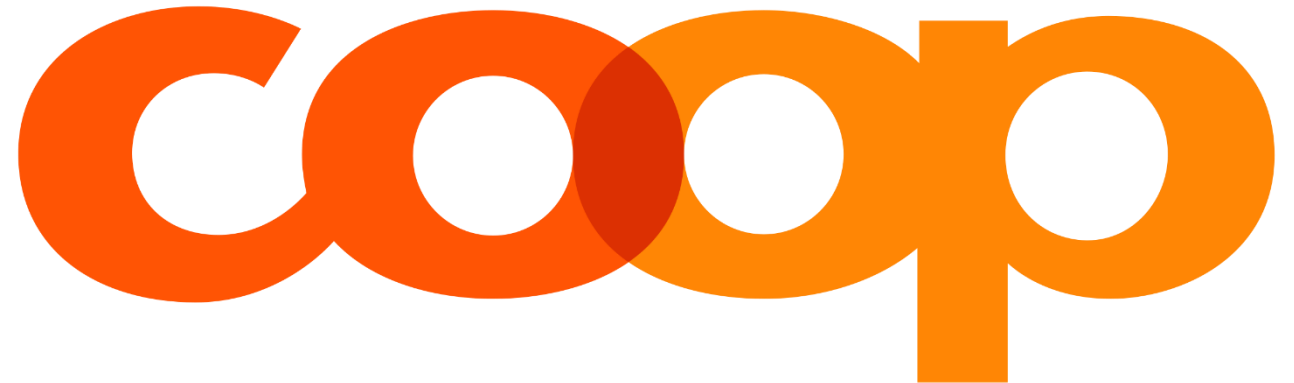
Wir starten mit einer Gruppe und teilen diese in einem iterativen Verfahren.

K-means

Teile die Datensets in K unterschiedliche, nicht-überlappende Cluster, indem die Varianz innerhalb der Cluster minimiert wird.

Fallstudie: Coop

Coop möchte seine Kunden anhand ihres Konsumverhaltens in verschiedene Gruppen einteilen, um gezielter Werbung und Promotionen zu schalten.



de.m.wikipedia.org/wiki/Datei:Coop.svg

Fallstudie: Coop

Fragestellung

Wie können wir unsere Kunden optimal in verschiedene Gruppen einteilen?

Relevante Merkmale

- Anzahl Kinder
- Einkommen
- Alter
- Ausgaben

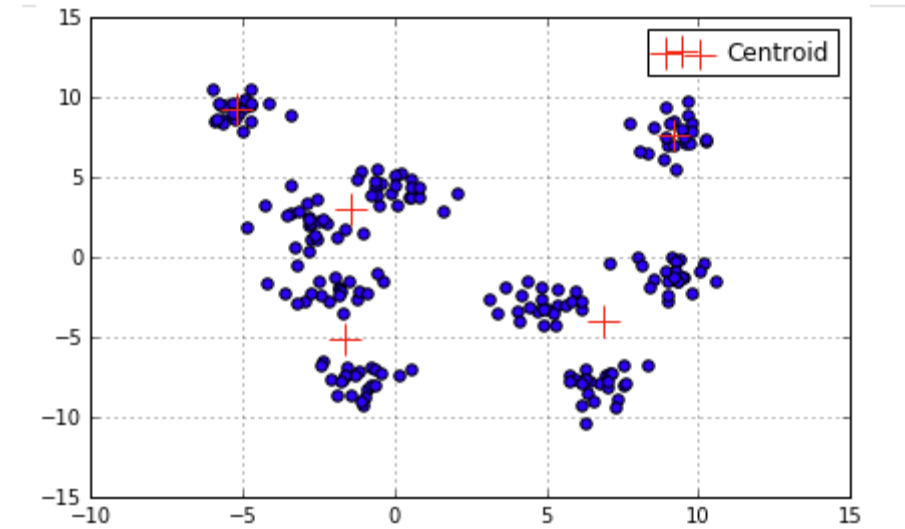
Kriterium

Homogene Gruppen

Auswertung

Distortion Score: Durchschnittliche Distanz zwischen Centroid (=Mittelpunkt) eines Clusters und den Beobachtungen.

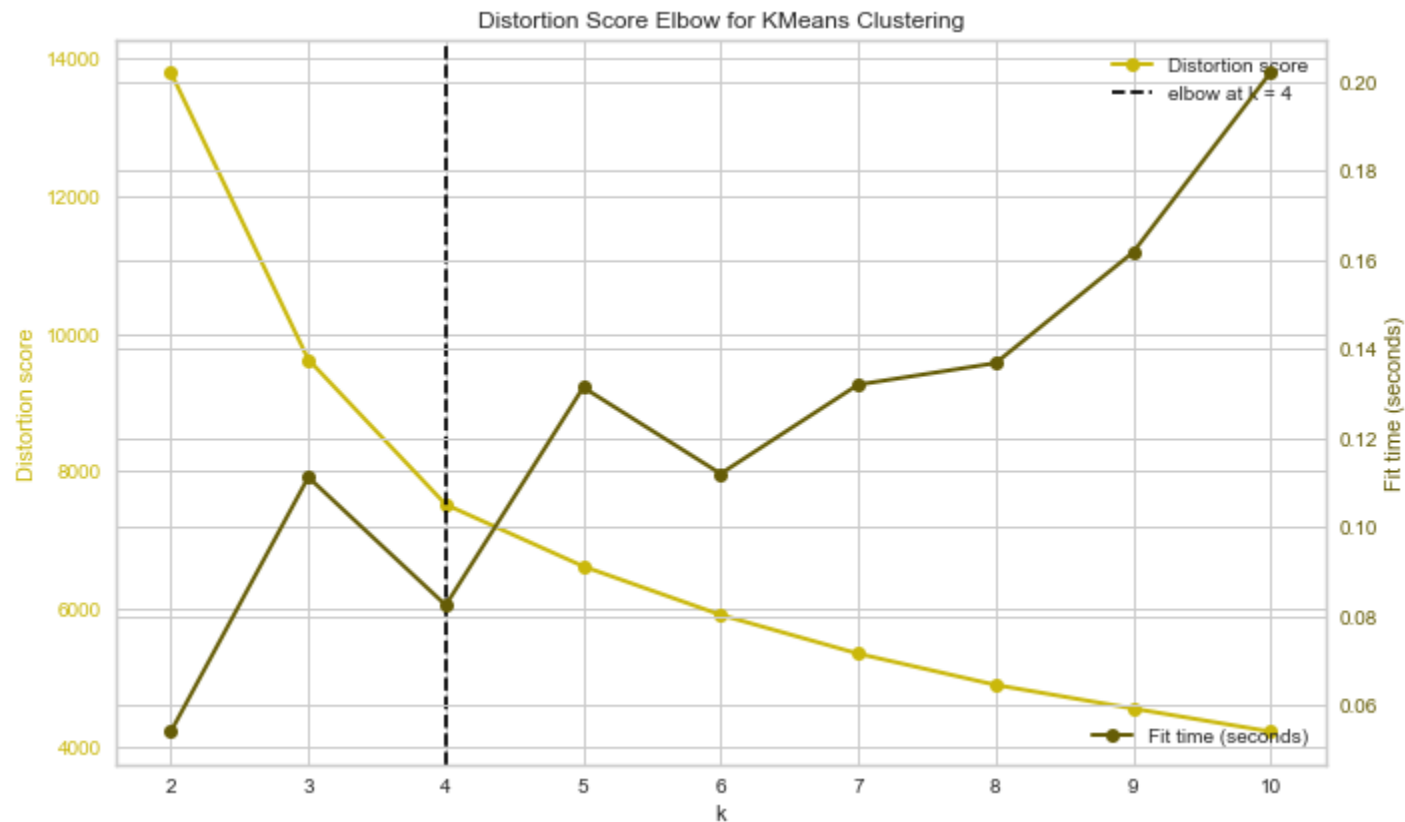
Punkte sollten möglichst nahe beim Centroid liegen.



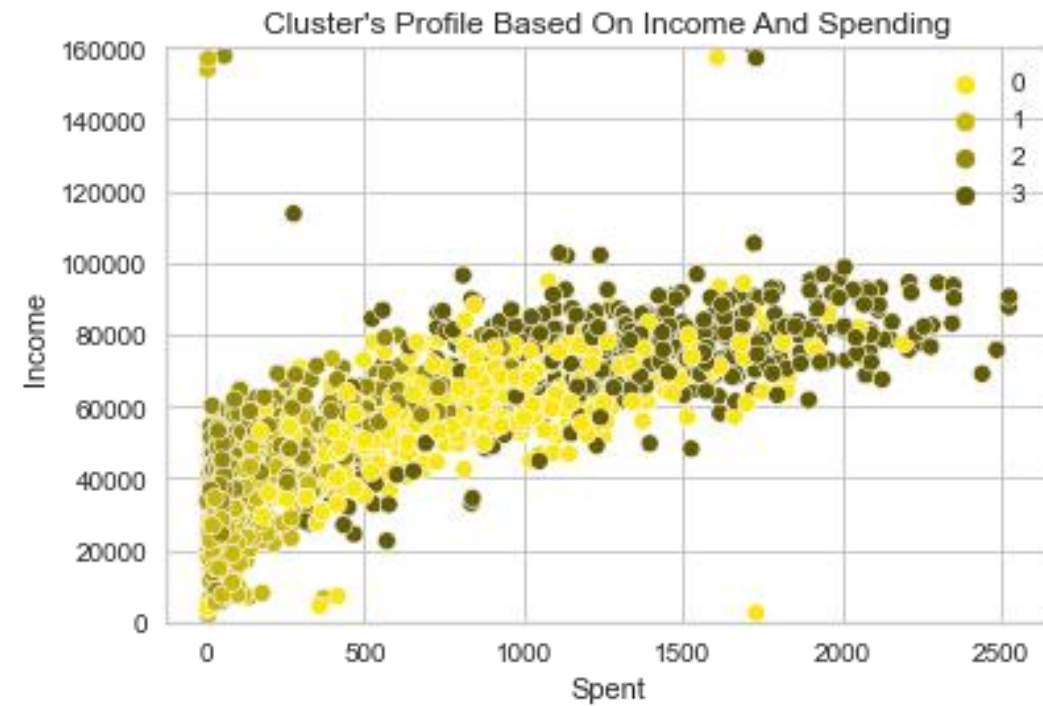
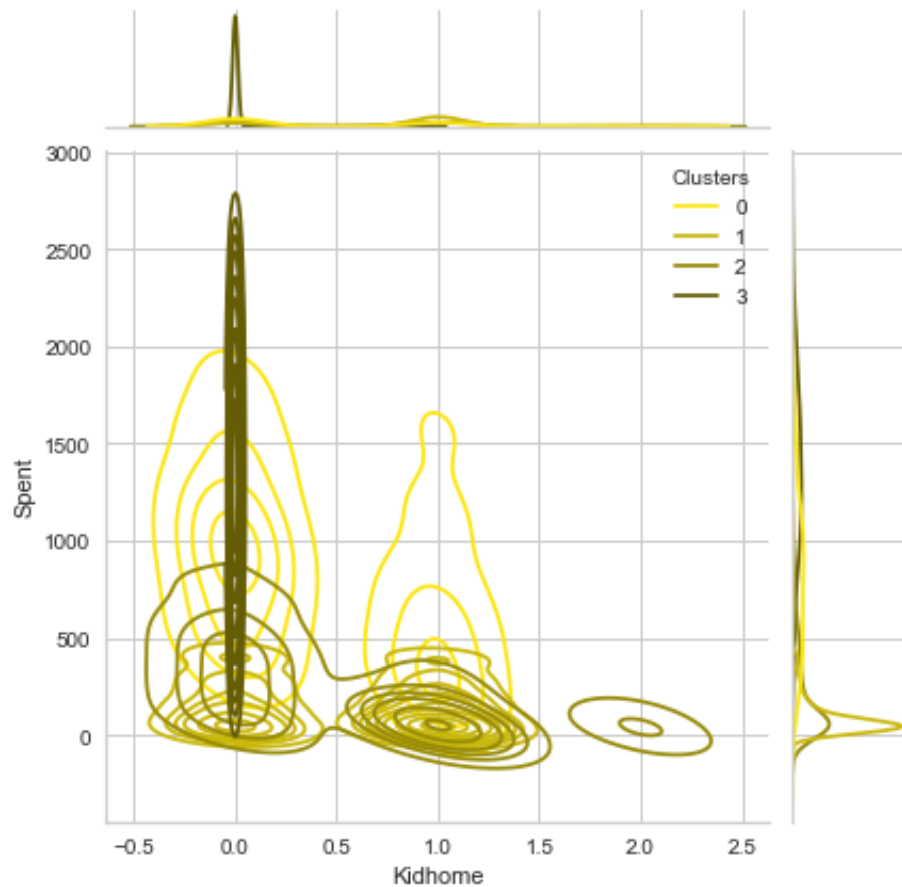
<https://avidml.wordpress.com/2016/10/29/easily-understand-k-means-clustering/>

Auswertung

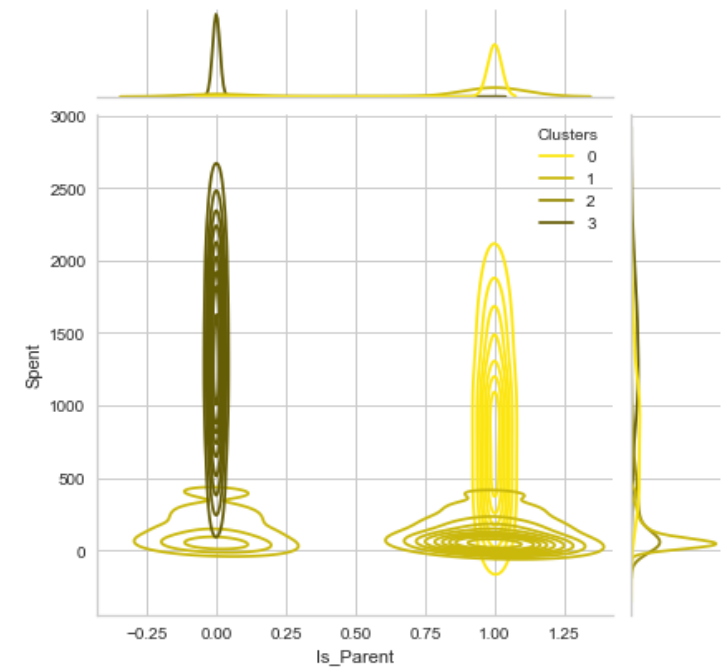
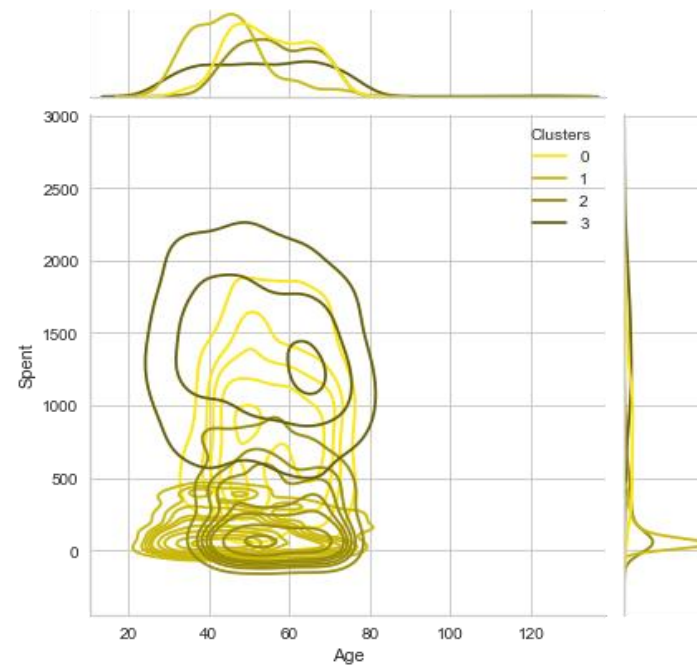
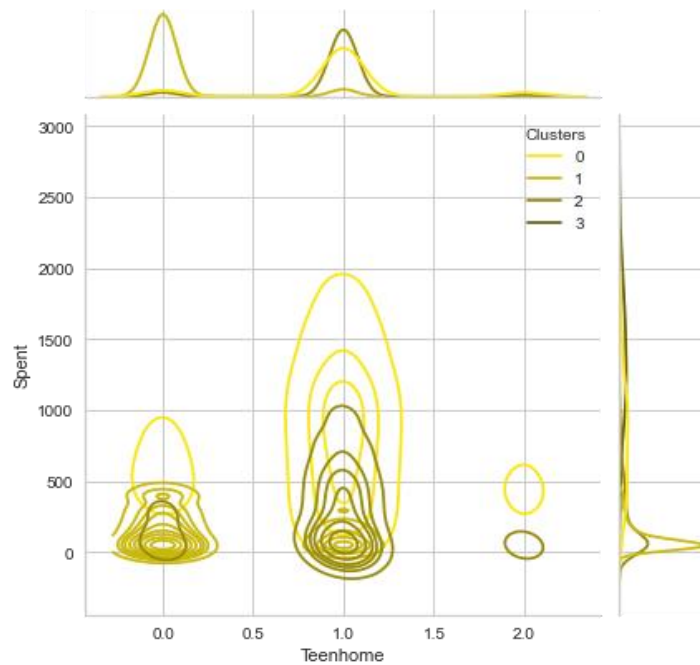
Jeder zusätzliche Cluster reduziert den Distortion Score, es gibt jedoch einen optimalen Punkt für die Anzahl Cluster.



Auswertung



Auswertung



Auswertung

Gruppe 0

- Sind sicherlich Eltern
- Mindestens 2 Familienmitglieder, maximal 5
- Grossteil hat Teenager
- Älter als der Durchschnitt
- Low Income-Gruppe

Gruppe 2

- Eltern
- Mindestens 2 Familienmitglieder, maximal 4
- Auch alleinerziehende Eltern sind enthalten
- Haben meistens Teenager zuhause
- Älter als der Durchschnitt der Kunden

Gruppe 1

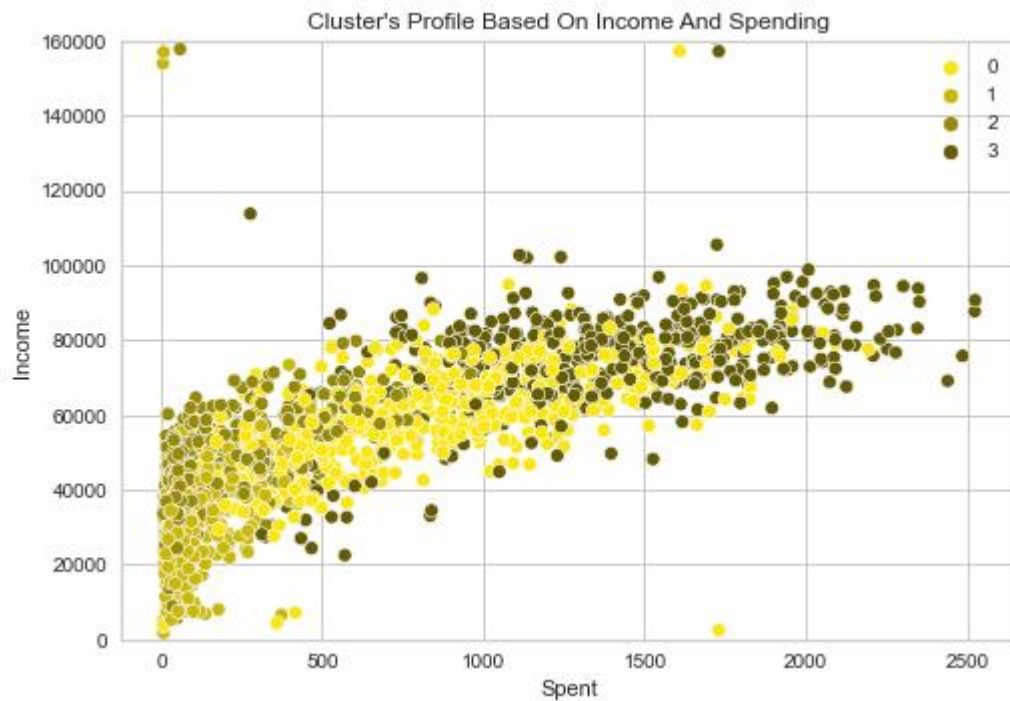
- Grossteil der Gruppe sind Eltern
- Maximal 3 Familienmitglieder
- Haben meistens nur ein Kind (keine Teenager üblicherweise)
- Jünger als der Durchschnitt
- Wenig Ausgaben

Gruppe 3

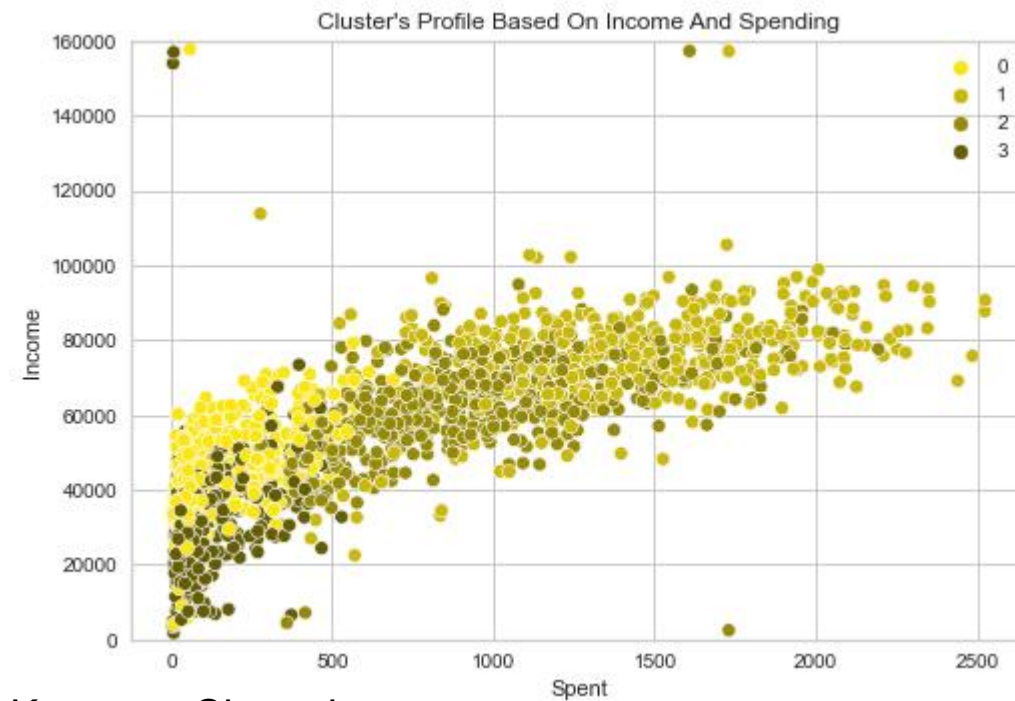
- Sicherlich keine Eltern
- Maximal zwei Familienmitglieder
- Paare bilden knappe Mehrheit
- Über alle Altersgruppen verteilt
- High Income-Gruppe (sog. DINK-Gruppe)

Auswertung

Wie können wir sicherstellen, dass unser Clustering robust ist?



Agglomeratives Clustering



K-means Clustering

Clustering: Key Takeaways

Was ist eine Budgetprognose?

Eine Schätzung der Absatzmenge in zukünftigen Geschäftsjahren, basierend auf heute verfügbaren Daten.

Weshalb benötigen wir Budgetierung und Budgetprognosen überhaupt?

- Schätzung von Liefermengen
- Schätzung des Personalbedarfs
- Gewinnschätzungen

Zusammenfassung

Predictive Forecasting

- Lineare / Nonlineare / ML-Modelle
- Overfitting / Underfitting: Generalisierbarkeit
- Vorgängige explorative Datenanalyse ist wichtig

Financial Fraud

- Strukturierte / unstrukturierte Daten
- Datenarchitektur
- Kosten / Nutzen-Abschätzungen

Clustering

- Agglomerative vs. Divisive models
- k-means
- Unsupervised Learning

Appendix

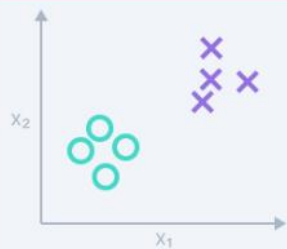
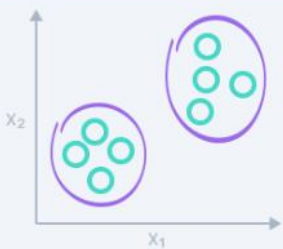
Data Science / Machine Learning 101

Wenn wir Datenanalyse betreiben, folgen wir meistens den folgenden Schritten

1. Data Loading: Laden der Daten
2. Data preprocessing: Eliminierung von irrelevanten Variablen
3. Exploratory Data Analysis (EDA): Erste Analyse und Visualisierung von Daten
4. Vorbereitung Daten & Modell: Aufbereitung der Daten, Initialisierung des Modells
5. Predictive Data Analysis: Schätzung des Modells und Analyse des «Goodness-of-Fit»

(Un-)supervised learning

- Supervised Learning: Ausprägung der Zielvariablen ist bekannt (bspw. Einkommenschätzung)
- Unsupervised Learning: Ausprägung der Zielvariablen ist nicht bekannt

Supervised learning	Unsupervised learning
Input data is labeled	Input data is unlabeled
Has a feedback mechanism	Has no feedback mechanism
Data is classified based on the training dataset	Assigns properties of given data to classify it
Divided into Regression & Classification	Divided into Clustering & Association
Used for prediction	Used for analysis
Algorithms include: decision trees, logistic regressions, support vector machine	Algorithms include: k-means clustering, hierarchical clustering, apriori algorithm
A known number of classes	A unknown number of classes
	

Gütemasse für Modelle

R^2

Anteil der Varianz der abhängigen Variable (Zielvariable), welche von den unabhängigen Variablen (Merkmalen) in einem Modell erklärt wird.

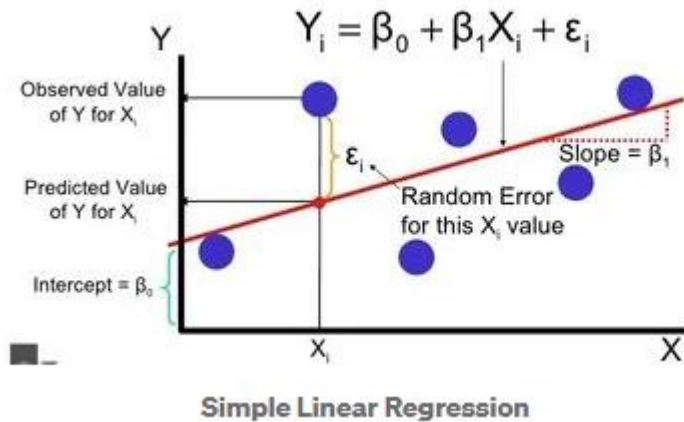
MSE (Mean Squared Error)

Durchschnittliche Fehlerquadratsumme. Zeigt auf, wie gut ein Modell an die Daten approximiert.

Cross Validation Accuracy (CVA)

Die Schätzgenauigkeit in der Cross Validation.

Regressionsmodelle (erweitert): Lineare Regression



$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Labels for the equation components:

- Dependent Variable: Y_i
- Population Y intercept: β_0
- Population Slope Coefficient: β_1
- Independent Variable: X_i
- Random Error term: ϵ_i

Groupings:

- Linear component: $\beta_0 + \beta_1 X_i$
- Random Error component: ϵ_i

<https://medium.com/@ram420/linear-regression-bebe2485415a>

Label und One Hot Encoding

Wann entscheiden wir uns für Label Encoding?

- Label Encoding: Wenn wir davon ausgehen, dass die Merkmale eine Ordnung haben. Bsp.: Letzter Bildungsabschluss
- One Hot Encoding: Wenn Merkmale keine Ordnung haben. Bsp.: Länderindikator

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50



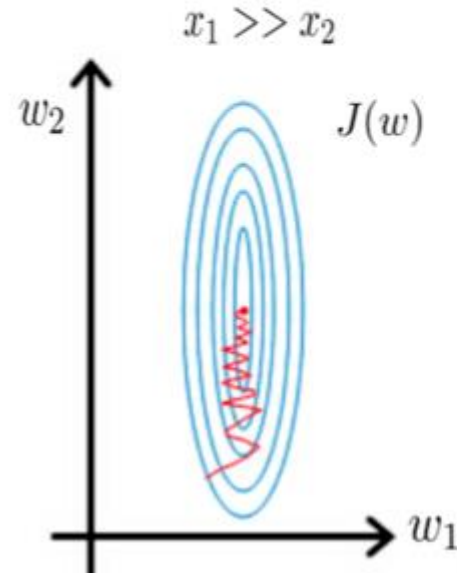
One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

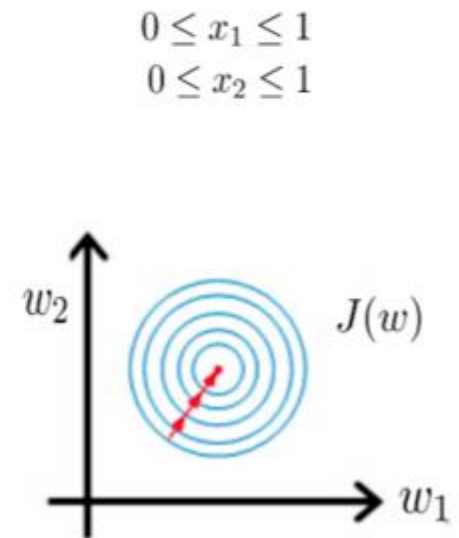
Datenskalierung

ML-Modelle minimieren einen Schätzfehler.
Unterschiedliche Skalen erschweren die
Optimierung.

Gradient descent
without scaling

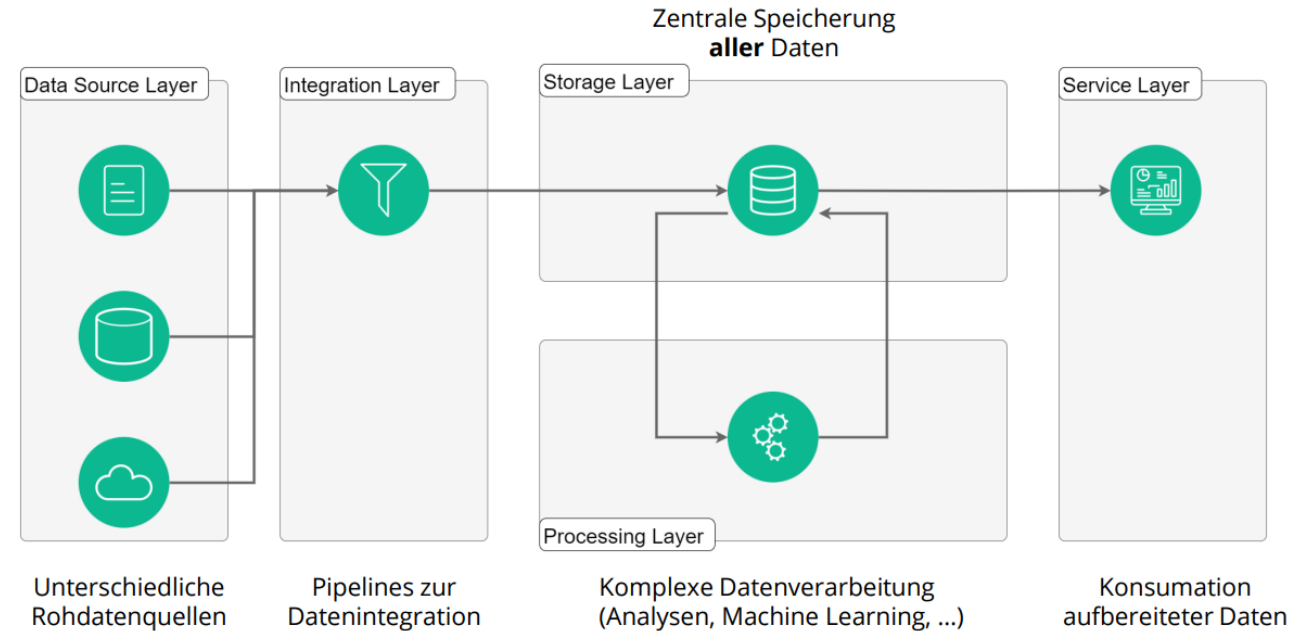


Gradient descent
after scaling variables



Dateninfrastruktur

Wir beschäftigen uns heute nicht mit einer kompletten Datenpipeline.
Wir sind hauptsächlich interessiert an der Processing und Service Layer.



Wichtige Begriffe und Konzepte

Training / Testing-Split

Teilen von Daten in zwei Teilssets:

- Training für die Schätzung der Modellparameter
- Testing für die Schätzung der Modellperformance (=Güte der Prognose)

Overfitting

Das Modell ist überspezifiziert und nimmt auch irrelevante Zusammenhänge auf

Goodness-of-Fit

Wie gut ein Modell Datenzusammenhänge beschreiben kann.