EEBDA

1: Einführung

1.1 – Big Data Analytics

Charakteristiker von Big Data

Big Data lässt sich anhand der 5-V's identifizieren.

- Volume: Große Mengen von Daten, die durch Endbenutzer oder Maschinen erzeugt werden
- Velocity: Schnelle Generierung von großen Datenmengen
  > Bedarf an Methodiken, um diese Datenmengen zu verarbeiten
- Variety: Daten kommen in diversen Datenformaten & Datenstrukturen vor
  > Structured (Datenbank), unstructured (Filme), semistructures (Websites)
- Veracity: Qualität, Granularität & Genauigkeit der Daten
- Value: Daten müssen für die Unternehmung von Wert sein (Veracity ist wichtiger Einflussfaktor)

Von Daten zur Weisheit

1. Data: stellen selbst kein Wissen dar
2. Information: verarbeitete Daten mit Nutzen *wer, was, wo, wann*
3. Knowledge: angewandte Daten und Informationen *wie*
4. Wisdom: *warum* passiert etwas, warum entstehen Korrelationen

Big Data Analytics

- Descriptive Analytics: Extraktion der Informationen aus Daten (*was*)
- Diagnostic Analytics: Schaffung eines tieferen Verständnisses von Sachverhalten mittels statistischer Methoden um Muster und Korrelationen zu erkennen
- Predictive Analytics: Voraussage von zukünftigen Ereignissen durch vergangene Daten
- Prescriptive Analytics: Finden von neuen Lösungswegen für komplexe Probleme

Datentypen

- Strukturiert: Daten, die gemäß eines festgelegten Datenmodells oder Schematas gespeichert werden
- Unstrukturiert: Daten, derren tatsächlicher Inhalt nicht strukturiert vorliegt
- Semistrukturiert: Semistrukturierte Daten besitzen eine auf höherem Level festgelegte Form, wobei deren Inhalt nicht zwingend strukturiert vorliegt.

Fähigkeiten für BDA

Business Understanding IT & Infrastruktur Anwendung von Methoden & Algorithmen

Digitale Tranformation

Bei digitaler Transformation geht es nicht nur um Technologie (70% der Projekte scheitern)!

1. Strategien sind für sinnvolle, ertragreiche Investments nötig
2. Mitarbeiter sollten externen Beratern vorgezogen werden (besser Kenntnis der Umgebung)+
3. Kunden sollten in die Entwicklung miteinbezogen werden
4. Sorgen der Mitarbeiter müssen beseitigt werden
5. Flachie Hierarchien, Agilität und Prototyping begünstigen die Tranformation

Statistical Modelling vs. Machine Learning

- Statistical Modelling: Problemstellungen werden durch mathematische Gleichungen formalisiert. Fokus liegt auf Modellschätzung und deren Inferenz.
- Machine Learning: Effiziente Rechenalgorithemen, die auf Basis von Trainingsdaten ein Modell erlernen. Die Algorithmen basieren aber auf mathematisch-statistischen Konzepten.

Machine Learning und Künstliche Intelligenz

Machine Learning ⊂ Künstliche Intelligenz

- Machine Learning: Maschinen bekommen Zugriff auf Daten und sollen damit lernen
- Künstliche Intelligenz: Fähigkeit von Maschinen, komplexe Aufgaben auf intelligent zu lösen
  > Spezifische KI: auf spezifisches Problem zugeschnitten
  > Allgemeine KI: ähnlich zur Intelligenz eines Menschen, für generische Probleme

---

1.2 – Big Data

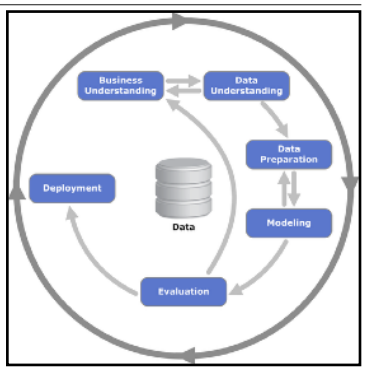
Digitale Wagenreihung der DB

Lesson Learned: In traditionellen und hochregulierten Branchen ist es nicht einfach, alte Strukturen und Technologien aufzubrechen und zu optimieren. Mit entsprechendem Domänenwissen lassen sich aber auch mit vorhandenen Daten neue Probleme lösen.

Data Engineering: Sammeln, Speichern & Verarbeiten großer Datenmengen

Lesson Learned: Die Massen von Daten, die in modernen Maschinen anfallen, können in zentralen Datenbanksystemen auf physisch verteilten Datenspeichern gesammelt werden. Alle Daten (Rohdaten) können in einem Data-Lake gespeichert und für ML herangezogen werden.

1.3 – Prozessmodelle



CRISP-DM

Das *CRoss-Indutry Standard Process for Data Mining* Modell unterscheidet zwischen sechs Prozessphasen, welche untereinander durch Feedbackschleifen in Wechselwirkung stehen.

① Business Understanding

Identifizierung der betriebswirtschaftlichen Anforderungen und Tranformation zu analytischen Zielen.

> Definieren d. Projektziels
> Bewertung des aktuellen Zustands
> Entwickeln der analytischen Ziele
> Aufstellen des Projektplans

② Data Understanding

Sammlung und Sichtung der Rohdaten zur Vorbereitung zur BDA Weiterverarbeitung; Identifizierung von unzureichender Datenqualität erfordert Anpassung der Projektziele.

> Sammeln der Daten
> Beschreiben der Daten
> Erste deskriptive Analyse der Daten (Extraktion der Informationen *was*)
> Verifizieren der Datenqualität

③ Data Preparation

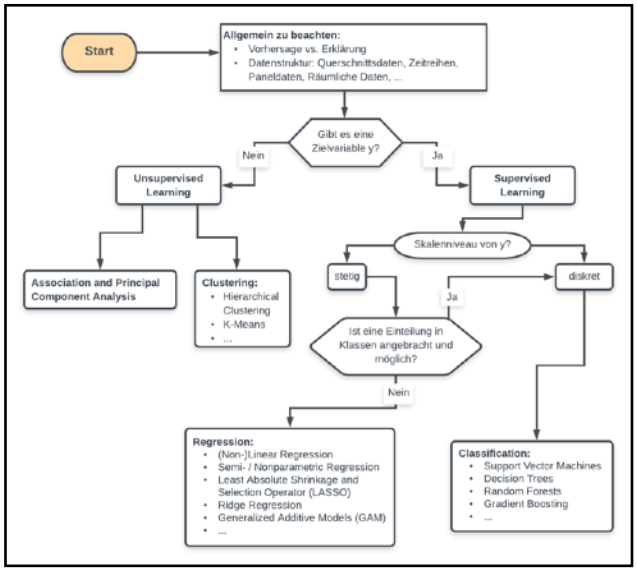
Aufbereitung der Daten (ML-kompatibel)

> Genaue Auswahl der Daten
> Säubern der Daten
> Formatieren der Daten
> Aufbauen eines neuen Datensatzes
> Integrieren der Daten in eine Modellierungsumgebung (Phython, R, etc.)

④ Modeling

Zur Analyse der aufbereiteten Daten muss ein geeignetes Modell gewählt werden.

> Auswählen der Modellierungsmethode
> Aufbauen des Test-Designs
> Aufbauen des Modells
> Bewerten des Modells



⑤ Evaluation

Prüfung, ob die definierten Projektziele durch die Ergebnisse des Modells erreicht werden.

> Evaluieren der Ergebnisse
> Überprüfen des Prozesses
> Bestimmen der zukünftigen Schritte

⑥ Deployment

Das Modell wird im Anschluss in die Unternehmensprozesse implementiert.

- Einbinden des Modells in das Unternehmen
- Monitoring und Wartung des Modells
- Erstellen eines Abschlussberichtes
- Abschlussbewertung des Projektes (Fazit, Lessons Learned, etc.)

2: Descriptive und Diagnostic Analytics

2.1 – Der χ^2-Test

Funktionsweise

Mit dem χ^2-Test kann man mehrer Variablen auf eine erwartete Verteilung prüfen.

- Voraussetzungen: $\forall i : np_i \geq 1 \quad$ für $80\,\%\ np_i \geq 5$

$$\chi^2 = \sum_{i=1}^{k} \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} \qquad df = k - 1$$

R-Code

```
chisq.test(<observed>, <expected>)
> returns χ², df, p-value
```

2.2 – Business Understanding

Theoretische Grundlagen der digitalen Abschlussprüfung

Der Wirtschaftsprüfer (WP) führt Abschlussprüfungen eigenverantwortlich und im öffentlichen Auftrag durch. Als Teil der Abschlussprüfung prüft der WP die angefallenden Kontobewegungen auf Unstimmigkeiten. Einfache Prüfverfahren sind

- Wochenendbuchungen/Sonntagsbuchungen: Für gewöhnlich sollten an Wochenenden keine Buchungen anfallen. Teilweise gibt es rechenintensive Buchungen, die an Wochenenden automatisiert durchgeführt werden.
- Kassenminusprüfung: Die Kasse (Barvermögen) darf niemals negativ sein.
- Prüfung von Duplikaten: Bei Einpflegen von Stamm- und Bewegungsdaten kann es zu Fehlern kommen. Der WP versucht eventuelle Duplikate in Stamm- und Bewegungsdaten anhand von Mustern zu identifizieren.

R

Import der Daten (in ein data.frame: unterschiedliche Datentypen möglich)

```
df.accounts_payable <- read.csv("case1_accountspayable.csv", sep=";",
↳header=TRUE, encoding="UTF-8")
head(df.accounts_payable, n=2) >Show head of data frame
summary(df.accounts_payable) >Show summary of statistical data of data frame
str(df.accounts_payable) >Show structure of data frame
dim(df.accounts_payable) >Show dimension of data frame
```

2.3 – Data Preparation

Konversion von Datentypen

Daten liegen häufig in ungeeigneter Struktur für die Analyse vor. Daher müssen BDAs die Daten konvertieren.

```
df.transactions$Date <- as.Date(df.transactions$Date, format="%d.%m.%Y")
Sys.setLocale("LC_TIME", locale = "en_US") >Sets time standard time format to US
df.transactions$DayOfTheWeek <- weekdays(df.transactions$Date) >Add column
levels(df.transactions$Reminders)[1] <- "0"
df.transactions$Reminders <- as.numeric(df.transactions$Reminders)
```

Filtern
Beim Filtern können Daten aus dem Datensatz entfernt werden, die für die Analyse nicht verwendet werden.
```
subset(<origin dataframe>, <boolsch value>)
```

Datensätze Zusammenfügen
Verschieden Datensätze können mithilfe des merge-Befehls und einem Primärschlüssel zusammengefügt werden.
```
merge(<dataframe_1>, <dataframe_2>, by="<primary_key>")
```

2.4 – Modelling I

Selektierung von Werten
Durch Filtern können bestimmte Werte (z.B. Sonntagsbuchungen) herausgefiltert werden.
```
df.sundays <- subset(df.transactions, df.transactions$DayOfTheWeek == „Sunday")
if(<bool>) { sth }
```

Iterative Prüfungen
Schleifen können zur iterativen Untersuchung von Daten verwendet werden.
```
for(i in <from>:<to>) { sth; }
```

Prüfung von Duplikaten
Mit Hilfe der R Funktion ‚duplicated' kann man eine Spalte auf Duplikate untersuchen.
```
<storeDuplicates> <- <column>[duplicated(<column>)]
```
Mit dem ‚%in%' Operator können Strings verglichen werden.
```
<subset> %in% <searchIn>
```

2.4 – Modelling II

Prüfung auf Statistische Abweichungen
Unter der Annahme, dass sich Menschen keine zufälligen Werte ausdenken können, kann auf systematische Abweichungen von der erwarteten Verteilung untersucht werden. Dazu könnte man mit Hilfe des χ^2-Tests die Abweichung von der Gleichverteilung einzelner Ziffern untersuchen.
```
Ints <- as.integer(<table>$<column>)
digitBeforeComma <- as.numeric(sapply(INT, function(x) {
↳substr(x, start = nchar(x), stop = nchar(x))
↳}))
distrDBC <- sapply(0:9, function(x) { sum(digitBeforeComma == x) })
```
Die Verteilung wird auf Abweichung von einer erwarteten Gleichverteilung untersucht
```
chisq.test(x = distrDBC, p = rep(0.1, 10))
```

Verteilung nach Benford
Führende Ziffern in Zahlen folgend abweichend von der Gleichverteilung überlicherweise der Verteilung nach Benford.

$$P(X = d) = \log_{10}(1 + 1/d)$$

```
Probs <- sapply(c(1:9),  function(x) {log10(1 + 1 / x)})
chisq.test(x = distrDBC, p = Probs)
```
Auch hier kann man mit Hilfe des χ^2-Tests Informationen über pot. Fehlbuchungen sammeln.

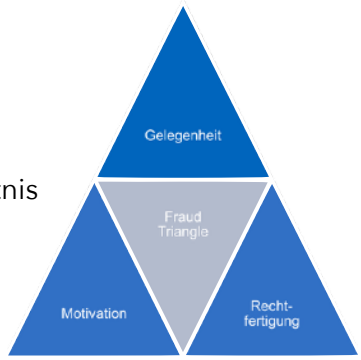
3: Predictive Analytics mit ML Algorithmen

3.1 – Vorkenntnisse

Buchführung
Siehe Zusammenfassung Bookkeeping.

3.2 – Business Understanding

Aufdeckung von Fraud
Es wird nach absichtlichen Verstößen durch Angestellte wie Untreue, Bilanzmanipulation oder Unterschlagung gesucht.



- Gelegenheit: Normalerweise durch Funktion des Mitarbeiters (Kenntnis des Kontrollsystems, hohes Vertrauen, Position) gegeben
- Motivation: Gestörtes Arbeitnehmer-Arbeitgeber-Verhältnis
- Rechtfertigung: Der Täter muss die Tat vor der Tatbegehung geg.ü. sich selbst rechtfertigen können, um sich berechtigt zu fühlen

Unternehmen geben Fraud als eines der größten Risiken an.
Entgegenwirken können Unternehmen mit:

- Vorbildfunktion d. Managements
- Operationalisierung von gesetzlichen & regulatorischen Vorgaben in Unternehmensleitlinien
  > Einprägung der Grundlagen in durch ethische Grundsätze geprägtes Selbstverständnis
- Anti-Fraud Management: Meldesystem, Compliance Officer

3.3 – Data Preparation I

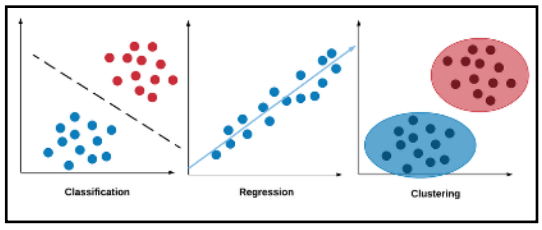
Datensätze Zusammenfügen
Mit der ‚rbind()' Function lassen sich mehrere Data-Frames (nach Prüfung der Spaltennamen) zusammenfügen.

```
<newFrame> <- rbind(<frame1>, ..., <frameN>)
```

3.5 – Modeling I

Machine Learning Arten

- Supervised Learning: Wird benutzt, um Erklärungen/ Prognosen einer Zielvariable zu erstellen.
  > Regression (*metric*), Klassifikation (*categorical*)
- Unsupervised Learning: Wird benutzt, wenn die Zielvariable nicht direkt beobachtbar ist.
  > Clusteranalyse (*categorical*)



Methodiken

- k-Nearest Neighbor Classification: neue Beobachtung anhand der k nächsten Nachbarn klassifizieren
  > Die ‚nächsten' Nachbarn können z.B. über Distanzmaße (euklidisch, Kosinus-Ähnlichkeit) bestimmt werden (auf Skalierung achten > ähnliche Skalierung schaffen, z.B. mittel min-max Normalisierung)

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

```
as.matrix(dist(trx.data[c("amount_scaled","tax_scaled")]))[1:6, 1:6]
library(class)
model <- knn(<training>, <test>, <training>$<var>, k = <number>)
```

- Entscheidungsbäume: hierarchische Abfolge dichotomer Entscheidungen, Algorithmen iterieren anhand der Entscheidungsregeln über den Baum
  1: alle Attribute auf Vorhersagegenauigkeit testen, mächtigstes als Wurzel wählen
  2: Datensatz wird gemäß des in 1. ermittelten Knotens in Untermengen aufgeteilt
  3: falls Beobachtung in Untermengen nicht dem Konsens entsprechen, mit der Untermenge wieder bei 1. beginnen.

```
library(rpart)  library(rpart.plot)
model <- rpart(class ~ <tb predicted> + <…>, data = <training>,
↳method = "class")
prp(<model>)   // plot
predict(<model>, <data>, type = "class")
```

- Support Vector Machines: Transformation der Daten in eine höhere Dimension, Bestimmung einer Hyperebene, die die Daten unterteilt, Hyperebene als Basis für zukünftige Klassifizierung

Methodenauswahl

1. Wie viel Daten sind vorhanden?
2. Welche Arten von Daten liegen vor ([un-]strukturiert)
3. Was soll geschätzt werden?
4. Wie wichtig ist die Visualisierung des Modellierungsprozesses?
5. Wie groß soll der Detailgrad sein (Daten aggregieren möglich)?
6. Beschränkungen durch Rechenleistung oder Speicherkapazitäten?

Trainings- und Testdaten

Daten werden üblicherweise zufällig in Trainings und Testdaten aufgeteilt. Das Model wird mit Hilfe der Trainingsdaten trainiert und im Anschluss daran mit den Testdaten validiert.

- Overfitting: Spezifizierung eines Modells, das zu viele erklärende Variablen enthält

Sampling

Mit Hilfe des samplings Befehl kann man aus einem Datensatz und einer gewünschten Größe eine zufälliges Sample generieren. Der Seed kann manuell festgelegt werden.

```
set.seed(123)  sample(1:nrow(data), size = 5)
```

3.6 – Data Preparation II

Große Datensätze

Da manche Operationen sehr rechen- und speicherintensiv sind, bietet es sich an, zufällige Testdatensätze zu samplen, mit denen man seinen Code testet.

Transformieren von Variablen

Zur Skalierung (z.B. für KNN) werden die Variablen transformiert. Teilweise müssen Datentypen angepasst werde (z.B. weil sie nicht numerisch sondern kategorisch sind)

```
<data>$<colT> <- <data>$<col> <…>
<data>$<colT> <- as.factor(<data>$<col>)
<data>$<col> <- NULL   // delete column
```

3.7 – Modelling II

Klassifizierungsaufgaben

Für Klassifizierungsaufgaben unter Verwendung von KNN und Entscheidungsbäumen werden eine Reihe von Libraries benötigt.

```
library(mlr)  library(rpart)  library(rpart.plot)
```

So kann eine Klassifizierungsaufgabe erstellt werden.

```
makeClassifTask(<data>, <target>)
```

Man kann sich eine Übersicht über die definierten Zielvariablen anzeigen lassen.

```
table(getTaskTargets(getTaskSize(<task>)))
```

Klassenungleichgewicht

Tritt eine Klasse deutlich öfter als eine andere auf, so spricht von von Klassenungleichgewicht. Dieses Ungleichgewicht kann z.B. durch Bildung eines Subsamples der häufiger auftauchenden Klasse.

```
undersample(<task>, rate = <p>)
```

Traininieren der KNN Modelle

Zunächst muss ein Lösungsalgorithmus gewählt werden.

```
learner <- makeLearner("classif.knn")
```

Ein Subset der Daten kann jetzt als Trainingsdaten verwendet werden.

```
model <- train(<learner>, task = <task>, subset = <subset>)
```

Das resultierende Model ist in einem Wrapper (wrappedModel).

Trainieren von Entscheidungsbäumen

Entscheidungsbäume lassen sich deutlich schneller ausführen, sind aber im Training rechenintensiver. Auch hier muss zunächst der Lösungsalgorithmus spezifiziert werden.

```
learner <- makeLearner("classif.rpart")
model <- train(<learner>, task = <task>, subset = <subset>)
```

Der Entscheidungsbaum kann geplottet werden.

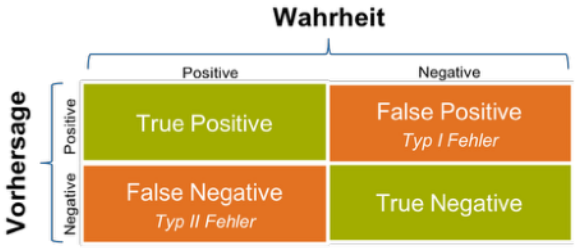
```
prp(<model>$learner.model, roundint = FALSE)
```

3.8 – Evaluation

Predictions

Mit der Predict Funktion können die Modelle zur Zuordnung von Daten genutzt werden.

```
prediction <- predict(<model>,
↳task = <task>, subset = <subset>)
```



- False-Positive-Rate (FPR), siehe Abbildung
- False-Negative-Rate (FNR), siehe Abbildung
- Missclassification Error (MMCE) gibt den gesamten Fehler an.
- Matthews Correlation Coefficient (MCC) ist ein balanziertes Evaluationsmaß $\in [-1; 1]$, wobei $+1$ das Beste und 0 nicht besser als eine Zufallsauswahl ist.

```
performance(<prediction>, measures = list(fpr, fnr, mmce, mcc))
calculateConfusionMatrix(<prediction>)
```

4: Predictive Analytics mit Regression

4.1 – Vorkenntnisse

Matrixalgebra

Addition:

$$A, B \in K^{m\times n} : A + B = c_{i,j} : c_{i,j} = a_{i,j} + b_{i,j} \in K^{m\times n}$$

Multiplikation:

$$A \in K^{m\times n}, B \in K^{n\times l} : A \cdot B = c_{i,j} : c_{i,j} = \sum_{k=1}^{n} a_{i,k} * b_{k,j} \in K^{m\times l}$$

Transponierte Matrix:

$$A = (a_{i,j}) \in K^{m\times n} \text{ ist } A^T := (a_{j,i}) \in K^{n\times m}$$

Inverse Matrix (nur wenn die Zeilen & Spalten voneinander linear unabhängig sind):

$$A \times A^{-1} = A^{-1} \times A = I_k : A \in K^{k\times k}$$

Determinante (für 2×2):

$$\det(B) = b_{1,1} \cdot b_{2,2} - b_{2,1} \cdot b_{1,2}$$

Einfaches Lineares Regressionsmodell

Das Ziel der linearen Regression ist ein Vorhersagemodell der Form

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i \qquad y = X \cdot \beta + \epsilon$$

Um die Regressoren β_i zu bestimmen, wird folgendes Minimierungsproblem gelöst

$$\min_{\widetilde{\beta}} (y - X \cdot \widetilde{\beta})' \cdot (y - X \cdot \widetilde{\beta}) = \min_{\widetilde{\beta}_0, \widetilde{\beta}_1} \sum_{i=1}^{n} (y_i - \widetilde{\beta}_0 - \widetilde{\beta}_1 \cdot x_i)^2$$

$$\hat{\beta}_{\text{OLS}} = (X' \cdot X)^{-1} \cdot X' \cdot y = \begin{pmatrix} \hat{\beta}_{0,\text{OLS}} \\ \hat{\beta}_{1,\text{OLS}} \end{pmatrix} \qquad \text{ordinary least squares (OLS) Schätzer}$$

BLUE:

1. Linearität der Parameter
2. Zufällig & unabhängige Ereignisse
3. Stichprobenvariation unter den Variablen
4. Bedingter Erwartungswert = 0
5. Homoskedastie (stabile Varianz)

Multiples Lineares Regressionsmodell

Das Ziel der multiplen linearen Regression ist ein Vorhersagemodell der Form

$$y_i = \beta_0 + \beta_1 \cdot x_{i,1} + \ldots + \beta_k \cdot x_{i,k} + \epsilon_i \qquad y = X \cdot \beta + \epsilon$$

$$\hat{\beta}_{\text{OLS}} = (X' \cdot X)^{-1} \cdot X' \cdot y = \begin{pmatrix} \hat{\beta}_{0,\text{OLS}} \\ \vdots \\ \hat{\beta}_{k,\text{OLS}} \end{pmatrix} \qquad \text{ordinary least squares (OLS) Schätzer}$$

BLUE:

1. Linearität der Parameter
2. Zufällig & unabhängige Ereignisse
3. Keine perfekte Multikollinearität (alle Regressoren fügen dem Model Informationen hinzu)
4. Bedingter Erwartungswert = 0
5. Homoskedastie (stabile Varianz)

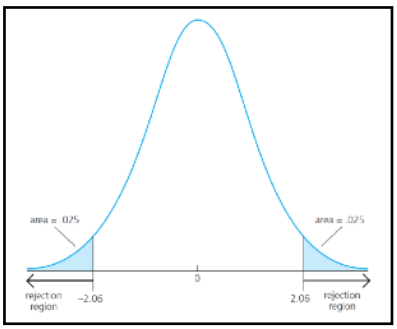
t-Tests

Zweiseitige Tests mit $H_0 : \beta_1 = 0$ und $H_A : \beta_1 \neq 0$, bei dem mit Signifikanzniveau von $\alpha \in [0; 1]$ abgelehnt wird.

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{\hat{\text{SE}}(\hat{\beta}_1)} \qquad t\text{-Teststatistik}$$

$$\hat{\text{SE}}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{(n-2) \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \text{Standardabweichung}$$

$$\hat{\epsilon} = y_i - \hat{y}_i \qquad \bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$



p-Wert

Der *p*-Wert entspricht dem Signifikanzniveau, zu dem die Nullhypothese gerade noch abgelehnt werden kann.

$$p^\star = \text{argmax}_\alpha \{\alpha : t_{\hat{\beta}_1} > |t(\frac{\alpha}{2}; n-2)|\}$$

Konfidenzintervalle

Mit Hilfe von Konfidenzintervallen lassen sich ebenfalls Aussagen zu Hypothesen treffen.

$$[\hat{\beta}_1 - |t(\alpha/2; n-2)| \cdot \widehat{\text{SE}}(\hat{\beta}_1); \hat{\beta}_1 + |t(\alpha/2; n-2)| \cdot \widehat{\text{SE}}(\hat{\beta}_1)]$$

F-Test

Der *t*-Test ermöglicht Aussagen darüber, ob einzelne Variablen signifikant von 0 abweichen. Der *F*-Test erlaubt das Testen von *r* unabhängigen Variablen und findet so vor allem in der multiplen linearen Regression Anwendung.

$$y_i = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4 + \epsilon$$

$$H_0 : \beta_2 = 0 \wedge \beta_3 = 0 \wedge \beta_4 = 0 \qquad H_A : \beta_2 \neq 0 \wedge \beta_3 \neq 0 \wedge \beta_4 \neq 0$$

$$r = 3 \qquad df = n - k - 1$$

$$F = \frac{(\text{SSR}_{H_0} - \text{SSR}_{H_A})/r}{\text{SSR}_{H_A}/(n-k-1)} = \frac{\Delta\text{SSR}/r}{\text{SSR}_{H_A}/(n-k-1)} \approx F(r; n-k-1) \qquad F\text{-Teststatistik}$$

SSR_{H_0} Summe der quadratieren Residuen im restringierten Modell

SSR_{H_A} Summe der quadratieren Residuen im *un*restringierten Modell

---

4.2 – Business Understanding

DCF Methode

Die Discounted Cash Flow Methode definiert den Unternehmenswert aus der zukünftigen Ertragskraft, also dem Barwert der zukünftigen Cashflows.

$$EV = \sum_{t=1}^{T} CF_t \cdot (1+i)^{-t} \qquad \lim_{T \to \infty} (EV) = CF \cdot \frac{1}{i-g}$$

i Zinssatz g (konstante) Wachstumsrate

Multiplikatoren

Da Cash Flows schwierig vorherzusagen sind, versucht man den Unternehmenswert als Vielfaches von Kennzahlen zu approximieren.

$$EV = \text{Werttreiber} \cdot \text{Multiplikator} = \text{Werttreiber} \cdot \frac{1}{i-g}$$

Der Multiplikator hat keine Allgemeingültigkeit und lässt sich nur begrenzt auf Unternehmen innerhalb einer Branche übertragen. Auch zeitlichen Veränderungen sollte Rechnung getragen werden.
Die Multiplikatoren können durch einfache lineare Regression approxmiert werden.

4.3 – Data Preparation

Zusammenführen von Datensätzen

Mit der Merge Funktion können Datensätze zusammengeführt werden.

```
merge(<data1>, <data2>, by = „<key>“)
```

Bereinigung des Datensatzes

Im Beispieldatensatz werden Finanzinstitute aufgrund von umgekehrter Bilanzierung aus dem Datensatz gefiltert.

```
data <- data[data$<col> != <sth>]
```

Auch Zeilen mit NA sollten herausgefilter werden.

```
data <- data[!is.na(data$<col>), ]
```

Nicht vorhandene Kategorien bei kategorischen Variablen sollten entfernt werden.

```
data <- droplevels(data)
```

Erzeugen zusätzlicher Variablen

Zusätzliche Variablen können implizit erzeugt werden.

```
data$<newCol> <- <calc>
```

4.4 – Modeling I

Lineares Regressionsmodell

```
regModel <- lm(<toPred> ~ <predBy>, data = <data>)
```

Die Koeffizienten lassen sich ausgeben. Dabei ist Intercept β_0 (die erklärenden Variablen folgen).

```
coefficients(<regModel>)
```

Zur Visualisierung kann die interne Plot Funktion verwendet werden.

```
plot(<yAxis> ~ <xAxis>, data = <data>)   // plot the scatterplot
abline(<regModel>)   // plot the model
```

Logarithmierung

Das Logarithmieren von nicht normalverteilten Variablen dient dazu, diese zu annähernd normalverteilten Variablen umzuformen.

```
<data>$<col> <- log(<data>$<col>)   // log for explanatory & explained variables
```

Ein lineares Regressionsmodell auf Basis logarithmierter Variablen nennt man log-log-Modell. Das Modell kann wie oben angegeben erstellt werden - es verändert sich lediglich die Interpretation der Koeffizienten.

Pooled Regression

Wird der Datensatz als Querschnittsdatensatz (also ohne Beachtung von anderen Effekten wie z.B. Zeit oder Sektoren) betrachtet, spricht man von einer gepoolten Regression.
Mit anderen Worten: Es wurde angenommen, dass Multiplikatoren über die Zeit und über die Sektoren stabil sind.

4.5 – Data Preparation II

Balancieren von Datensätzen

In einem unbalancierten Panel liegen für unterschiedliche Jahre unterschiedlich viele Beobachtungen vor. Zu Vereinfachungen kann man den Datensatz balancieren (z.B. nur Einträge, die in allen Jahren vorkommen).

```
balanced <- <data>[<key> %in% names(which(table(<key>) == <totalNumber>)), ]
cbind(<data>, <newColName> = table(
```

4.6 – Modelling II

Regressionsmodell 1

Heterogenitätsannahmen: Der Multiplikator ist weder zeit- noch sektorspezifisch.

$$y = X \cdot \beta + \epsilon \Rightarrow \hat{\beta}_{\text{OLS}} = (X' \cdot X)^{-1} \cdot X' \cdot y$$

Gepoolte Regression mit einer Regressionsgleichung.

Regressionsmodell 2

Heterogenitätsannahmen: Der Multiplikator ist zeit- aber nicht sektorspezifisch.

$$y_t = X_t \cdot \beta_t + \epsilon_t \Rightarrow \hat{\beta}_{t,\text{OLS}} = (X_t' \cdot X_t)^{-1} \cdot X_t' \cdot y_t \quad \text{für } t \in [1;T]$$

Es ergeben sich jeweils Regressionsgleichungen für die einzelnen Zeitperioden $t \in T$.

Regressionsmodell 3

Heterogenitätsannahmen: Der Multiplikator ist nicht zeit- aber sektorspezifisch.

$$y^s = X^s \cdot \beta^s + \epsilon^s \Rightarrow \hat{\beta}^s_{\text{OLS}} = (X^{s'} \cdot X^s)^{-1} \cdot X^{s'} \cdot y^s \quad \text{für } s \in [1;S]$$

Es ergeben sich jeweils Regressionsgleichungen für die einzelnen Sektoren $s \in S$.

Regressionsmodell 4

Heterogenitätsannahmen: Der Multiplikator ist zeit- und sektorspezifisch.

$$y_t^s = X_t^s \cdot \beta_t^s + \epsilon_t^s \Rightarrow \hat{\beta}^s_{t,\text{OLS}} = (X_t^{s'} \cdot X_t^s)^{-1} \cdot X_t^{s'} \cdot y_t^s \quad \text{für } s \in [1;S],\ t \in [1;T]$$

Es ergeben sich Regressionsgleichungen für alle Kombinationen aus $s \in S$, $t \in T$.

Seemingly Unrelated Regression (SUR)

Lineare Modelle blenden potentielle Interdependenzen zwischen den erklärenden Variablen aus.

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} X_1 & 0 & 0 \\ 0 & X_2 & 0 \\ 0 & 0 & X_3 \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix} = X \cdot \beta + \epsilon \qquad \text{Var}(\epsilon) = \Omega = \begin{pmatrix} \sigma_{1,1}I & \sigma_{1,2}I & \sigma_{1,3} \\ \sigma_{2,1}I & \sigma_{2,2}I & \sigma_{2,3}I \\ \sigma_{3,1} & \sigma_{3,2}I & \sigma_{3,3}I \end{pmatrix}$$

Die Varianz-Korvarianzmatrix dient der Überprüfung der Korrelation.

$$\hat{\beta}_{\text{FLGS}} = (X' \cdot \hat{\Omega}^{-1} \cdot X)^{-1} X' \cdot \hat{\Omega}^{-1} \cdot y \qquad \text{feasible generalized least squares}$$

SUR in R

Mit den Libraries ‚Systemfit' und ‚PLM' können Interdependenzen beachtet werden.

```
library(systemfit)   library(plm)
data <- pdata.frame(x = <data>, index = <indicesOfDependecies>)
model <- systemfit(formular = <explaining> ~ <explained>, method = "SUR",
↳data = <data>)
```

Die Koeffizienten können wie gewohnt ausgegeben werden.

4.7 – Evaluation

Chow-Test

Zunächst müssen die Hypothesen aufgestellt werden.

$$H_0 : \beta_1 = \beta_2 = \beta_3 \quad H_A : \beta_1 \neq \beta_2 \vee \beta_2 \neq \beta_3 \vee \beta_2 \neq \beta_3$$

Der Chow-Test ist ein Spezialfall des F-Tests. Es soll überprüft werden, ob einzelne disjunkte Teilgruppen der Stichprobe einem einzigen Regressionsmodell folgen (H_0) oder verschiedenen Regressionsmodellen, die sich in mindestens einem Koeffizienten voneinander unterscheiden (H_A).

$$F = \frac{(\mathrm{SSR}_{H_0} - \mathrm{SSR}_{H_A})/(G \cdot p)}{\mathrm{SSR}_{H_A}/(n - G \cdot p)} = \frac{\Delta\mathrm{SSR}/(G \cdot p)}{\mathrm{SSR}_{H_A}/(n - G \cdot p)} \approx F(G \cdot p; n - G \cdot p)$$

$$\mathrm{SSR}_{H_A} = \sum_{g=1}^{G} \mathrm{SSR}_g \qquad df = n - G \cdot p$$

p number of regressors G number of distinct groups

```
residuals <- sum(<data>$residuals^2)
chow <- ((<resH0> - sum(<resHA_1>, .. ., <resHA_G>))/(G*p))
↳ / (sum(<resHA_1>, .. ., <resHA_G>)/(n-G*p))
```

Das zugehörige Quantil kann aus der F-Statistik bestimmt bzw. abgeglichen werden.

```
qf(<1-alpha>, df1 = G*p, df2 = n-G*p)
```

Ablehnen der H_0 bedeutet, dass ein Multiplikator, der für die beiden Sektoren invariant ist, auf dem 5% Signifikanzniveau nicht ausgeschlossen werden kann.

Prognose

Um die Prognoseperformance zu evaluier, kann man auf RMSE (root mean squared errors) zurückgreifen.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2} \quad \text{RMSE} \qquad \frac{1}{n}\sum_{i=1}^{n}|\hat{\epsilon}_i| \quad \text{MAE (mean absolute error)}$$

```
forecast <- predict(object = <model>, newdata = <data>)
errors <- log(<forecast> - <actual>)
RMSE <- sqrt(mean(<errors>^2))
MAE <- mean(abs(<errors>))
```

5: Predictive Maintenance mit Statistical Process Control

5.1 – Vorkenntnisse

Ablaufbezogene Fertigungsverfahren: Werkstattfertigung
- Arbeitsplätze mit gleichartigen Arbeitsverrichtungen am selben Ort
- Werkstücke werden von Werkstatt zu Werkstatt transportiert
- Hohe innerbetriebliche Transportwege (Optimierungspotential)
- Universalmaschinen, die umgerüstet werden müssen (Optimierungspotential)

Wird verwendet für: Einzel- und Kleinserienfertigung

Ablaufbezogene Fertigungsverfahren: Fließfertigung
- Betriebsmittel und Arbeitsplätze nach Sequenz der erforderlichen Arbeitsschritte angeordnet
- Wenn Fertigungsschritt nicht notwendig ist, durchläuft Werkstück den Arbeitsschritt (passiv)
- Strenge Taktzeiten für kostengünstigen Weitertransport
- Arbeitsrhythmisierung mit monotonisierten Arbeitsabläufen
- Ausfall einer Arbeitsstation führt zu Störung der gesamten Produktion

Wird verwendet für: Großserien und Massenproduktion

5.2 – Business Understanding

Qualitätssicherung

Qualität ist ein wesentlicher Wettbewerbsfaktor für Unternehmen weswegen der Qualitätsicherung eine hervorgehobene Stellung im Unternehmen einnimmt.
Das Qualitätsmanagement umfasst alle technischen und organisatorischen Maßnahmen die sowohl präventiv, überwachend und steuernd zur Steigerung der Qualität beitragen.
- Einsatz von Sensoren > Echtzeitdaten
- Abweichungstoleranzen überschritten > Eingriff

Erfassung von Messdaten im Produktionsprozess

MES (Manufacturing Execution Systems) sammeln über Sensoren dauerhaft Daten über den Produktionsprozess. Durch Analyse dieser Daten können die MES Fehler frühzeitig erkennen und präventiv oder unmittelbar qualitätserhaltend oder terminierend in den Prozess einzugreifen.
Zur Analyse wird SPC-Software (Statistical Process Control) herangezogen, die eine Reihe von Informationen aus dem Fertigungsprozess erfassen.
Maße können nicht exakt eingehalten werden und unterliegen einer Schwankungen, die in begrenztem Rahmen zu akzeptieren sind (Toleranzgrenzen). Das ist besonders bei CTQ-Merkmalen (Cirtical to Quality) wichtig.
Schwankungen können verschiende Ursachen haben:
- Gewöhnliche Gründe für Schwankungen (common causes) > inhärent
- Spezielle Ursachen für Schwankungen (special/assignable causes [of variation]) > externe Einflüsse

5.3 – Data Understanding

Produktionsdaten

Mit der Library ‚Data.Table' können spezielle Methoden zum einlesen.

```
library(data.table)
data <- fread(file = <filename>, select = <range>, header = <bool>,
↳data.table = <bool>)
```

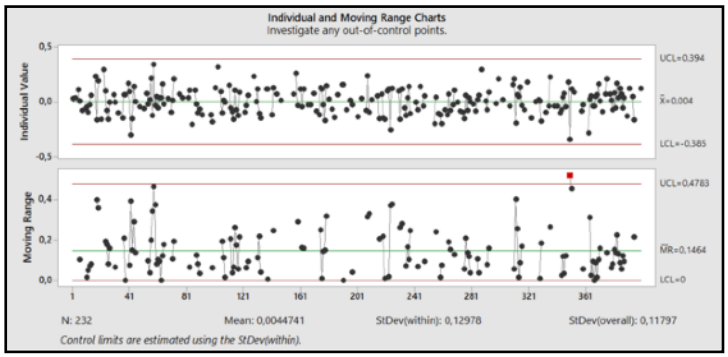
5.4 – Modelling

(X, R)-Charts

(X, R)-Charts, auch I-MR-Chart (Individual-Moving-Range-Chart) kann im Zusammenspiel mit Run-Charts (Nelson-Rules) zur präventiven Qualitätssicherung eingesetzt werden (predictive quality assurance oder predictive maintenance).

- Mittelwertlinie
- UCL (Upper Control Level)
- LCL (Lower Control Level)

Der Bereich zwischen UCL und LCL (spezifiziert durch Toleranzgrenzen oder technische Spezifikation) darf nicht verlassen werden. Beim übertreten dieser Grenzen muss die Produktion gestoppt werden.



Nelson-Rules (in R)

Die Stabilität eines Produktionsprozesses kann mit der Funktion *six_sigma_ctrl_chart_mod* beurteilt werden.

```
sic_sigma_ctrl_chart_mod(<dataFrame>,
↳lineColors = <vector with 7 elements for colors ±3,2,1;0 sd>
↳applyRules = <vector with 8 bools, one for each nelson rule>
↳rulesColor = <vector with 7 elements for colors of rules>
↳seg = <vector to split large data on multiple plots>
↳keepStats = <bool, analyse seg individually?>
↳verbose = <print additional information>, r1 = <from>, r2 = <to>)
```

Der Command gibt eine zweielementige List zurück (data.frame, das auf Basis des Individual Charts angibt, bei welcher ID welche Nelson-Regel erfüllt ist. Das zweite Listenelement ist ein data.frame, das auf Basis des Moving Range Charts angibt, bei welcher ID die erste Nelson-Regel erfüllt ist).

5.5 – Evaluation

Evaluation

Die Nelson-Rules dienen der statischen Prozesskontrolle und eignen sich zur Produktionsüberwachung. Durch frühzeitige Fehlererkennung und entsprechende Nebensteuerung kann SPC-Software nachhaltig zur Qualitätssicherung beitragen.

6: Ethik und Privatsphäre unter Big Data Analytics

6.1 – Deutscher Ethikrat

Kernaussagen
- Gesellschaftliche Veränderung durch Big Data Analytics (BDA)
- Einheitliche Regeln aufgrund globaler Datenströme schwierig
- Gefahr, dass auch anononymisierte Daten irgendwann Personenbezug erlauben
  > Damit auch zur Erhebungszeitraum unklar, ob/wie geschützt werden muss
- Detaillierte Personenprofile können missbraucht werden
- Selbstinduzierte Fremdbestimmung durch überzogene Selbstkontrolle
- Verfassungsrechtliche Maßstarbsnorm ist das Recht auf informationelle Selbstbestimmung
  > Datenschutzrechtliche Grundsätze: Personenbezug, Zweckbindung, Erforderlichkeit, Sparsamkeit, Einwilligung, Transparenz

1. Freiheit und Selbstbestimmung
2. Privatheit und Intimität
3. Souveränität und Macht
4. Schadensvermeidung und Wohltätigkeit
5. Gerechtigkeit, Solidarität und Verantwortung

- Gefährdungspotentiale durch BD: Privatheit, Manipulation des Denkens

6.2 – Ein ethisches Modell

Daten
- Empfehlungen aufgrund der breiten & abstrakten Perspektive schwierig
- Daten nicht nutzen, wenn: per Gesetz verboten, Qualität zu schlecht, Quelle nicht seriös

Die folgenden Werte sollten bei der Konzeption des ethischen Rahmens Beachtung finden.

Beneficial
Bei der Planung eines BDA Projektes müssen Nutzen und Risiko, sowohl gesamtgesellschaftlich, als auch individuell beachtet werden. Dabei sollte man auch potentielle Risikogruppen identifizieren.

Progressive
Einsatz dort, wo BDA besser performt, als andere Lösungsansätze. BDA nicht wegen des Trends.

Sustainable
Den Entwicklern muss klar sein, dass auch Algorithmen und Entscheidungsmodelle in ihrer Vorhersagegenauigkeit abnehmen und entsprechend eine vordefinierte Halbwertszeit besitzen.

Respectful
Nutzer, bzw. Produzenten der Daten sollten mit dem gebotenen Respekt behandelt werden. Die Daten sollten in ihrem Interesse und entsprechend ihrer Wünsche verarbeitet werden.

Fair
Entscheidungsmodelle sollten jeden Mensch gleich behandeln.

6.3 – Über Menschen urteilende Maschinen

Diskussionsthemen

- Mögliche Verzerrrung der Modelle durch ungenügende/fehlerhafte Trainingsdaten.
- Haftung für Entscheidungen durch Modelle beim Modellersteller?
- Können Menschen sich über Entscheidungsempfehlungen hinwegsetzen?
- Welche Fehlerquoten kann man akzeptieren?
- Transparenter Umgang mit den Optimierungszielen von Algorithmen

6.4 – EU: Trustworthy AI

Hauptkomponenten

Die Ethik-Kommision der EU entwickelte das Modell Trustworth AI mit folgenden Hauptkomp.:

- Lawful AI: im Einklang mit allen Gesetzen und Regularien
- Ethical AI: geltende ethische Prinzipien und Werte müssen berücksichtig werden
- Robust AI: Robustheit gegenüber Veränderung, sodass kein unvorhersehbarer Schaden entstehen kann

Prinzipien für Ethical AI

- Respect for Human Autonomy: Menschen müssen ihre volle und effektive Selbstbestimmung behalten und an demokratischen Prozessen teilhaben können.
- Prevention of Harm: KI-Systeme sollen keine Schäden verursachen/Menschen beeinträchtigen
- Fairness: gerechte Verteilung von Nutzen und Kosten, keine Diskriminierung und Stigmatisierung. Möglichkeit zur Anfechtung von Entscheidungen.
- Explicability: transparente Prozesse, ex-ante & ex-post Erklärung

Schlüsselvorrausetzungen für AI

- Human agency and oversight: Einschließlich der Grundrechte, des menschlichen Handelns und der menschlichen Aufsicht.
- Technical robustness and safety: Einschließlich Widerstandsfähigkeit gegen Angriffe und Sicherheit, Rückfallplan und allgemeine Sicherheit, Genauigkeit, Zuverlässigkeit und Reproduzierbarkeit.
- Privacy and data governance: Einschließlich der Wahrung der Privatsphäre, der Qualität und Integrität der Daten und des Zugangs zu den Daten.
- Transparency: Einschließlich Rückverfolgbarkeit, Erklärbarkeit und Kommunikation
- Diversity, non-discimination and fairness: Einschließlich der Vermeidung unlauterer Verzerrungen, der Zugänglichkeit und des universellen Designs sowie der Beteiligung der Interessengruppen.
- Societal and environmental wellbeing: Einschließlich Nachhaltigkeit und Umweltfreundlichkeit, soziale Auswirkungen, Gesellschaft und Demokratie.
- Accountability: Einschließlich Auditierbarkeit, Minimierung und Berichterstattung über negative Auswirkungen, Trade-offs und Rechtsbehelfe.