Analyse von Transaktionsdaten im Online-Ticketing mit Data-Mining-Methoden

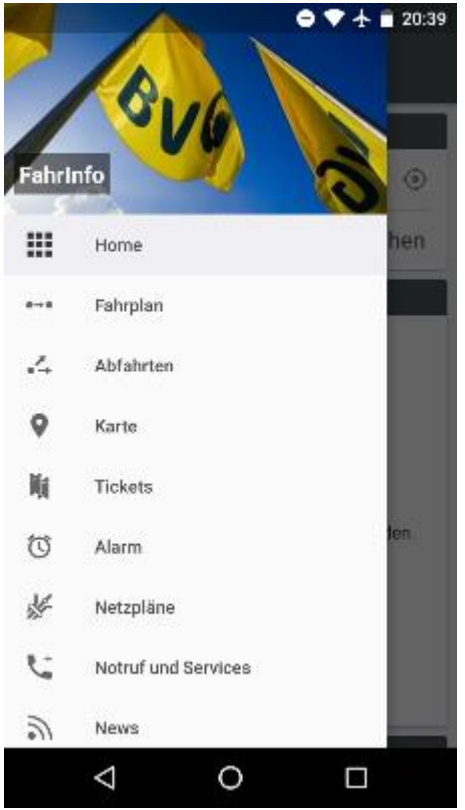
Marten Pfannenschmidt, Freie Universität Berlin
Prof. Dr. Jan Fabian Ehmke, Europa-Universität Viadrina
Frank Schreier, Berliner Verkehrsbetriebe (BVG)

Nahverkehrs-Tage 2017

Die neuen Daten des ÖPNV

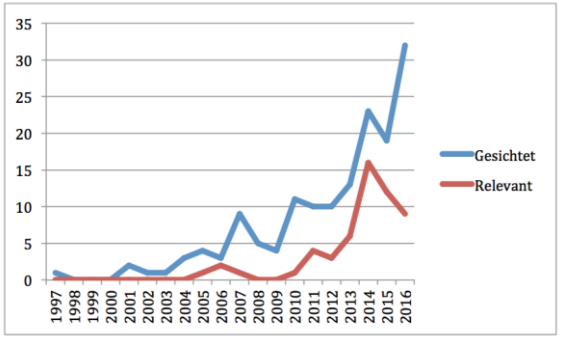
- Zunehmend digitale Vertriebskanäle im ÖPNV
- Ca. 30 Verkehrsunternehmen mit Handyticket
- Stamm- und Transaktionsdaten werden **automatisiert** erhoben

- Idee: Analyse umfangreicher **Stamm- und Transaktionsdaten** zur Erstellung von Vorhersagen des Kundenverhaltens

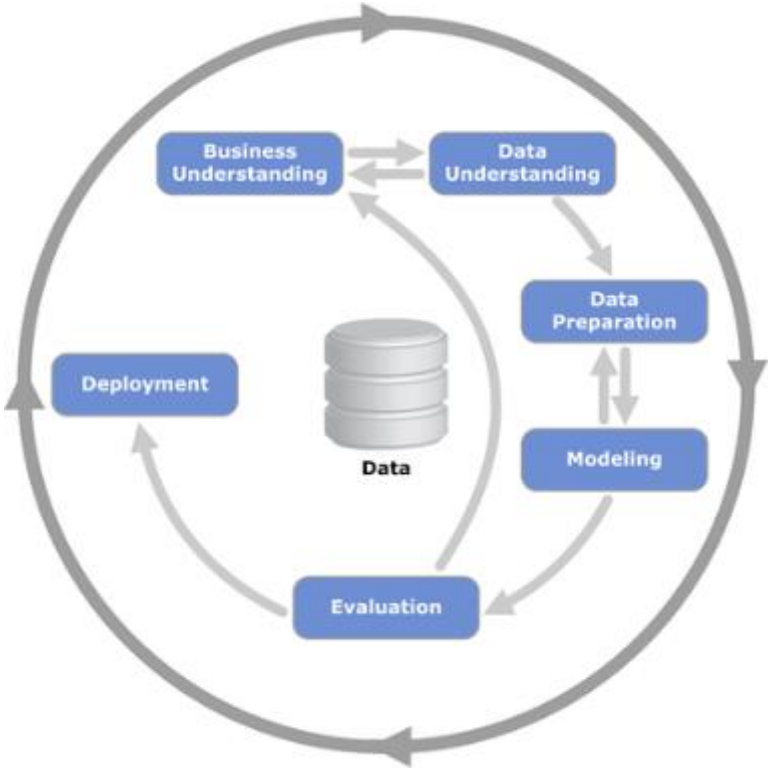


Ein Blick in die Literatur

- Massendatenanalyse im ÖPNV-Ticketing
- Zunehmende Relevanz des Themas in der Forschung in den letzten fünf Jahren
- Fokus auf **Automated-Fare-Collection-Systeme** und **Smart-Card-Anwendungen**
- Prognose von Fahrgastströmen zur verbesserten Angebotsplanung
- Marketingperspektive selten



Massendatenanalyse mit CRISP-DM 1/2

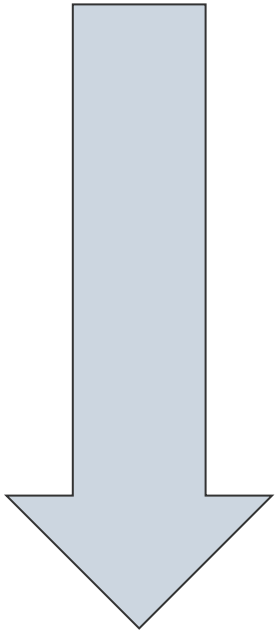


Analyse anonymisierter Transaktions- und Stammdaten aus der Fahrinfo-plus-Applikation der BVG

CRoss **I**ndu**S**try **P**rocess for **D**ata **M**ining

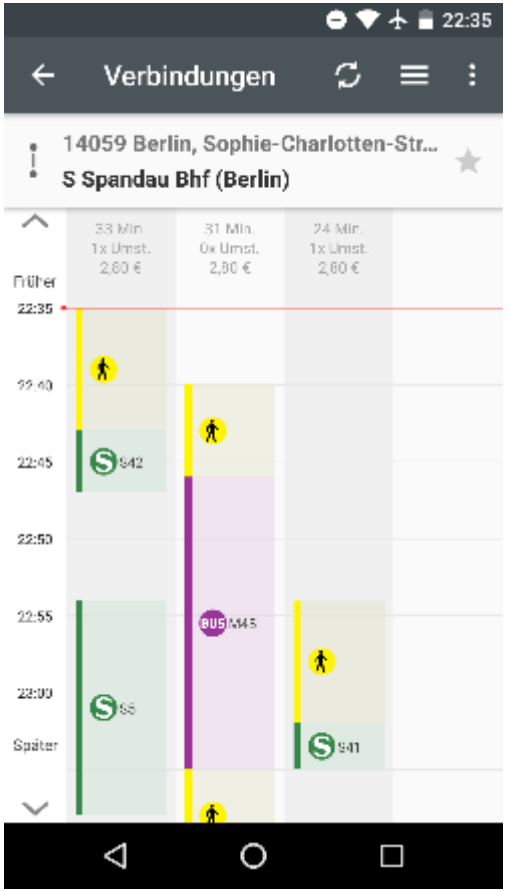
Massendatenanalyse mit CRISP-DM 2/2

Business Understanding	Fragestellungen aus Unternehmenskontext in ein Data-Mining-Problem übersetzen
Data Understanding	Bedeutung, Umfang und Qualität der Daten
Data Preparation	Bereinigen der Daten, Herausarbeiten weiterer Attribute
Modeling	Modellauswahl und Training des Modells
Evaluation	Wie gut ist das Modell?
Deployment	Wie kann das Modell in der Praxis eingesetzt werden?



Fahrinfo plus der BVG

- Ca. 1 Mrd. Fahrgäste jährlich
- 2,7 Mio. Downloads in
  Play-, App- und Windows-Store
- ca. 350.000 registrierte Kunden
- Produkte des klassischen Bartarifs
  (Einzelfahrausweis, Tagestickets, Touristik)
- 7% Umsatzanteil bezogen auf
  relevante Tarifprodukte
- Jährliche Verdoppelung des Umsatzes
  seit Start der App



Fahrinfo plus – Datenbestand

Stammdaten

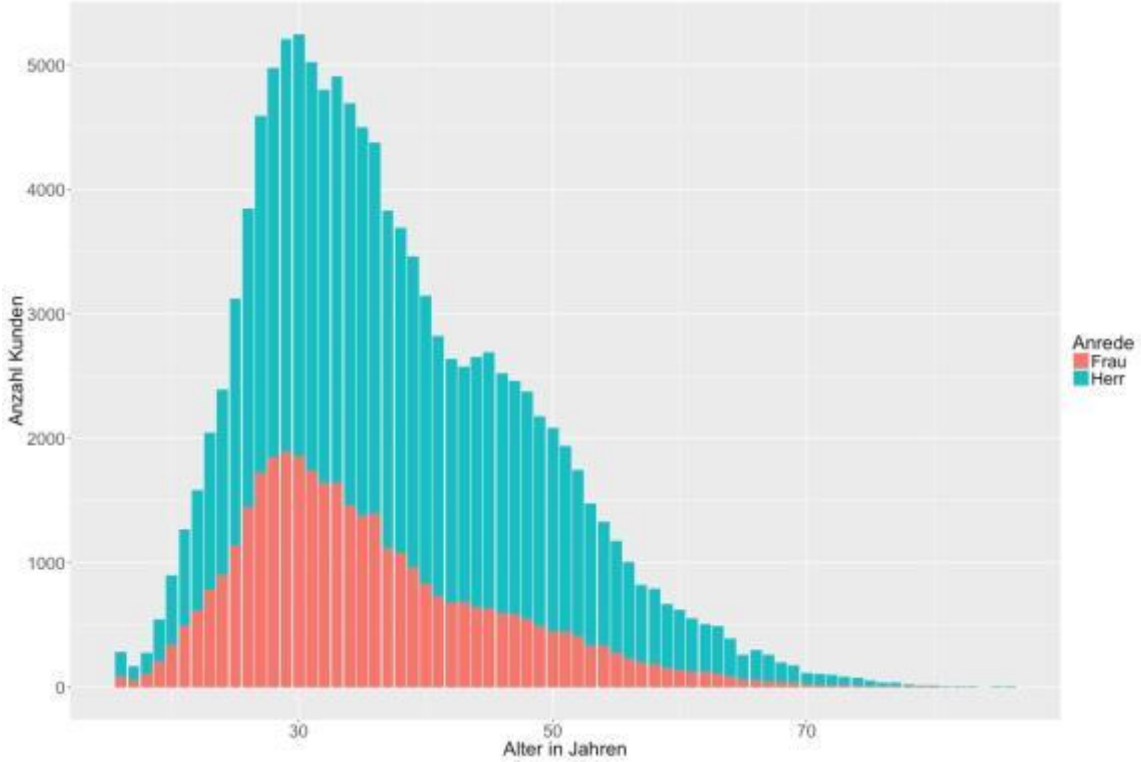
- Kundennummer
- Anrede
- Geburtsdatum
- Straße, Postleitzahl, Stadt, Land
- Registrierungsdatum
- Letzter Login
- Letzte Bestellung
- Sperrstatus

Transaktionsdaten

- IDs: Kundennummer, Bestellnummer, Ticketnummer
- Zeitliche Informationen: Kaufzeitpunkt, Gültig-ab- und Gültig-bis-Datum
- Ticketinformationen: Produktnummer, Produktbezeichnung, Preis
- Räumliche Daten: Starthaltestelle, Geltungsbereich

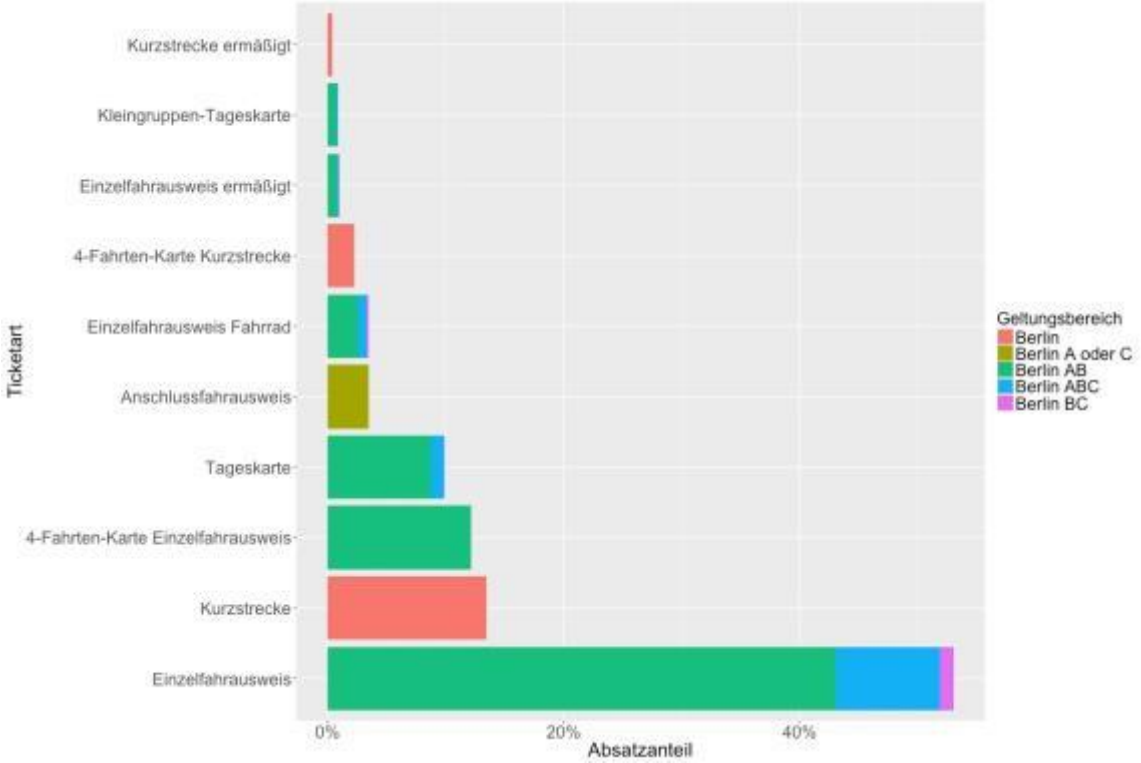
Explorative Analyse
Demografische Struktur der ÖPNV-Kunden

- 88% haben deutschen Wohnsitz
- 50% aller Kunden mit Wohnsitz in Berlin
- 58% der Kunden sind zwischen 25 und 40 Jahre alt
- Weniger weibliche als männliche Kunden



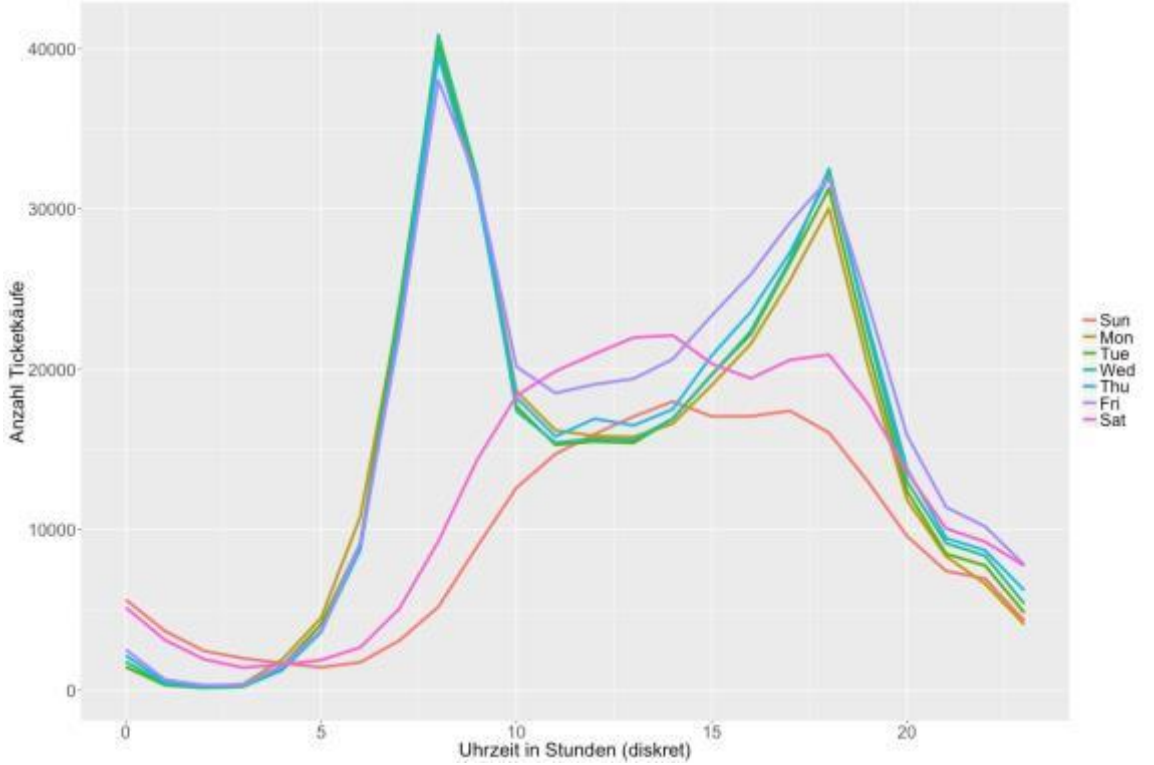
Explorative Analyse
Welche Tickets werden häufig gekauft?

- Top 10 Tickets nach Absatz
- 53% des Gesamtabsatzes sind Einzelfahrausweise
- Tarifbereich AB dominant



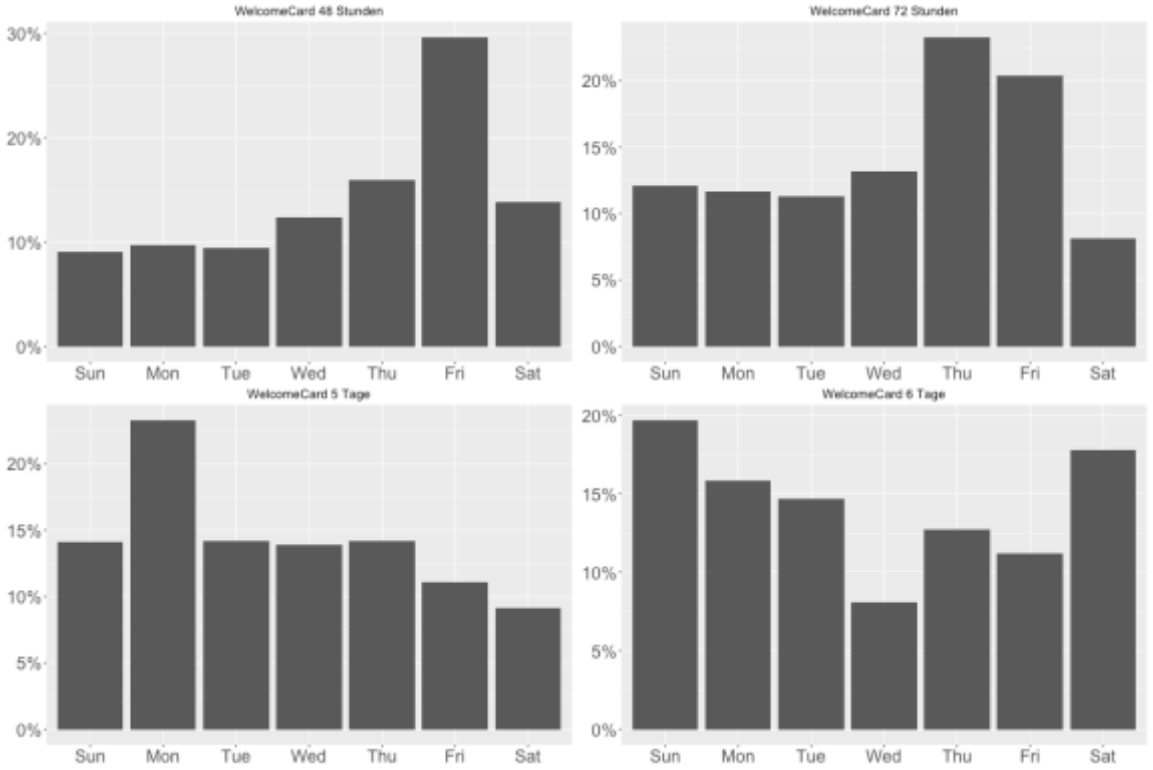
Explorative Analyse
Wann werden Tickets gekauft?

- Deutliche Unterschiede zwischen Werktagen und Wochenenden
- Werktags Verkaufsspitzen zur Rush-Hour
- An Wochenenden Verkäufe über den Tag verteilt



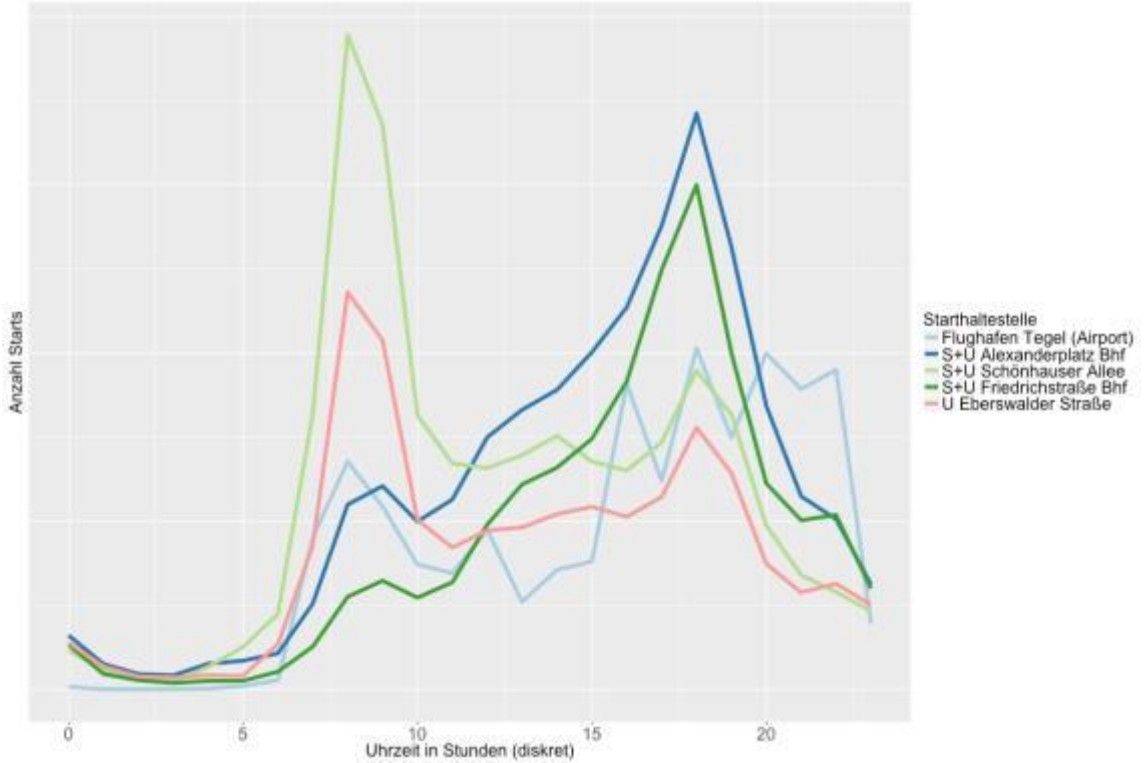
Explorative Analyse
Besonderheiten weniger gefragter Tickets

- Touristische Produkte folgen eindeutigen Mustern
- Kaufentscheidung wird abhängig vom Wochentag getroffen



Explorative Analyse
Startzahlen über den Tagesverlauf

- Top 5 Haltestellen nach Startzahlen
- Möglichkeit der Klassifizierung von Starthaltestellen aufgrund von Verkaufsmaxima



Data Mining – Welche Attribute definieren das Ticket?

- Ziel: **Zusammenhänge in Massendatensätzen erkennen** und daraus **Vorhersagen** ableiten (Klassifikation durch Generalisierung)
- Hier: Geeignete Attribute aus Stamm- und Transaktionsdaten **auswählen** und damit das **wahrscheinlichste Ticket** prognostizieren
- „Überwachtes“ Verfahren

Gesamter Datensatz

...	...	Einzelfahrausweis
...	...	Tageskarte
...	...	Einzelfahrausweis
...	...	Einzelfahrausweis

Trainings-Datensatz

...	...	Einzelfahrausweis
...	...	Tageskarte
...	...	Einzelfahrausweis

Test-Datensatz

...	...	Einzelfahrausweis

Vorhersagen mit Data Mining – Klassifizierende Verfahren

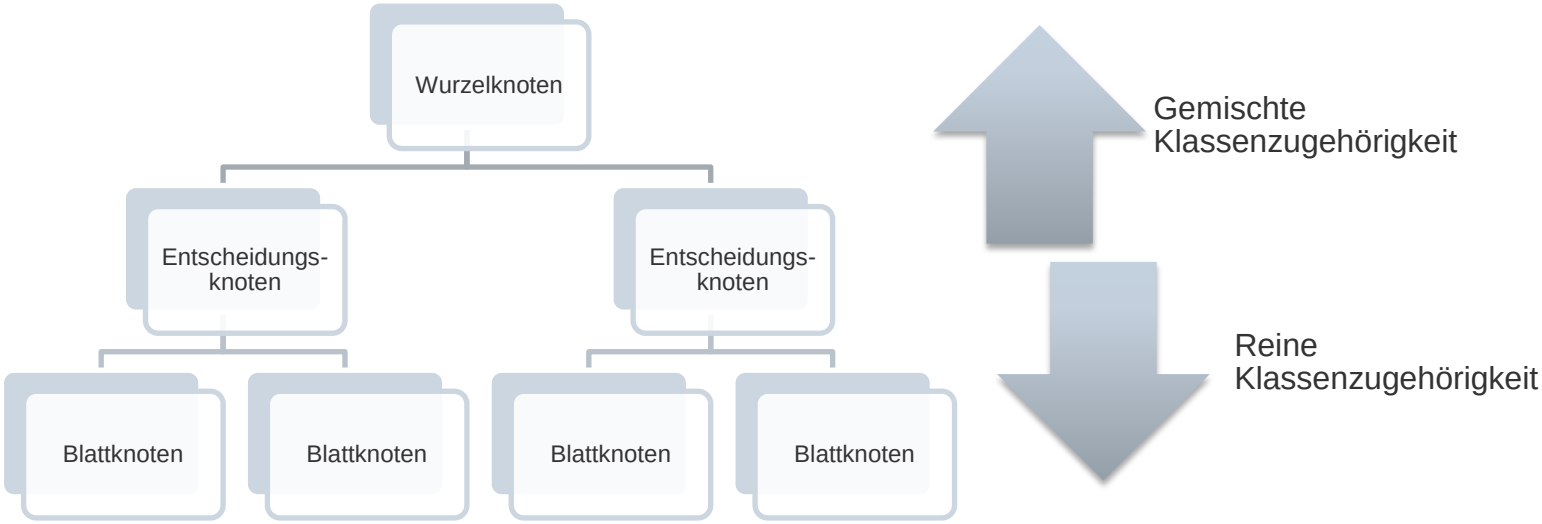
Kann das vom Kunden gewünschte Ticket mit zufriedenstellender Qualität prognostiziert werden?

Verschiedene Modellvarianten:

1. Immer die „dominante“ Klasse wählen → Einfaches Modell als Baseline
2. Decision Tree → Etabliert, gute Interpretierbarkeit
3. Random Forest → Komplexer, aber häufig bessere Prognoseergebnisse

Vorhersagen mit Decision Trees 1/2

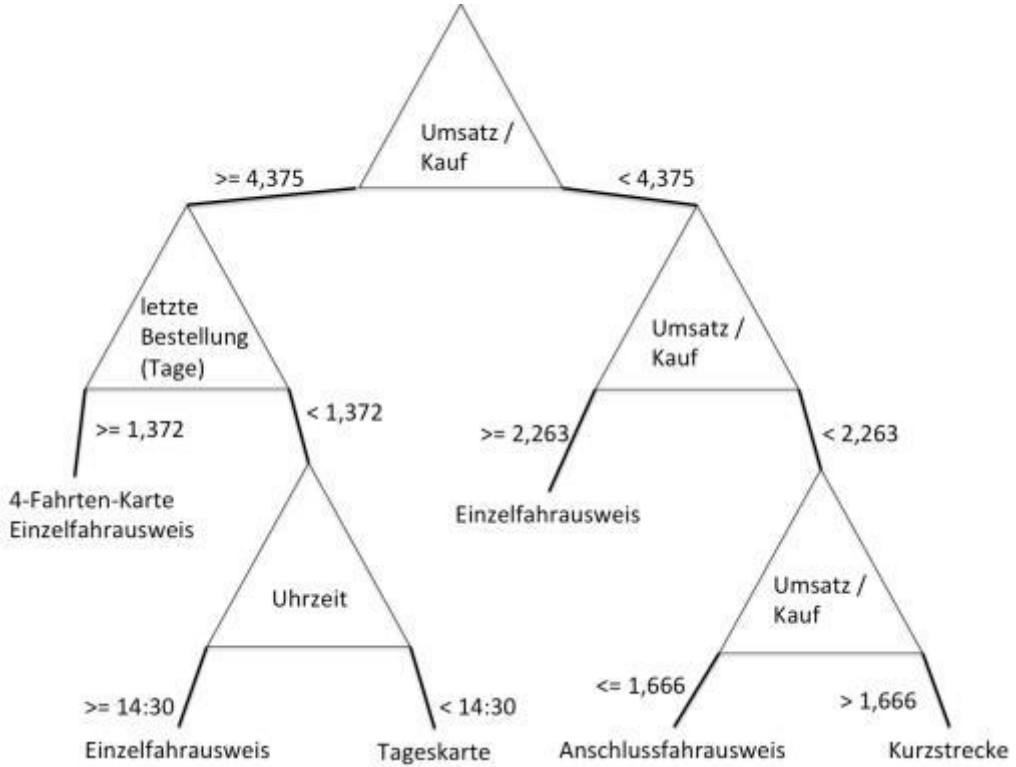
- Datensatz wird systematisch schrittweise aufgeteilt
- Aufteilung auf Grundlage von Variablenausprägungen (z.B. „m“/„w“)
- Ziel: Möglichst „reine“ Blattknoten



Vorhersagen mit Decision Trees 2/2

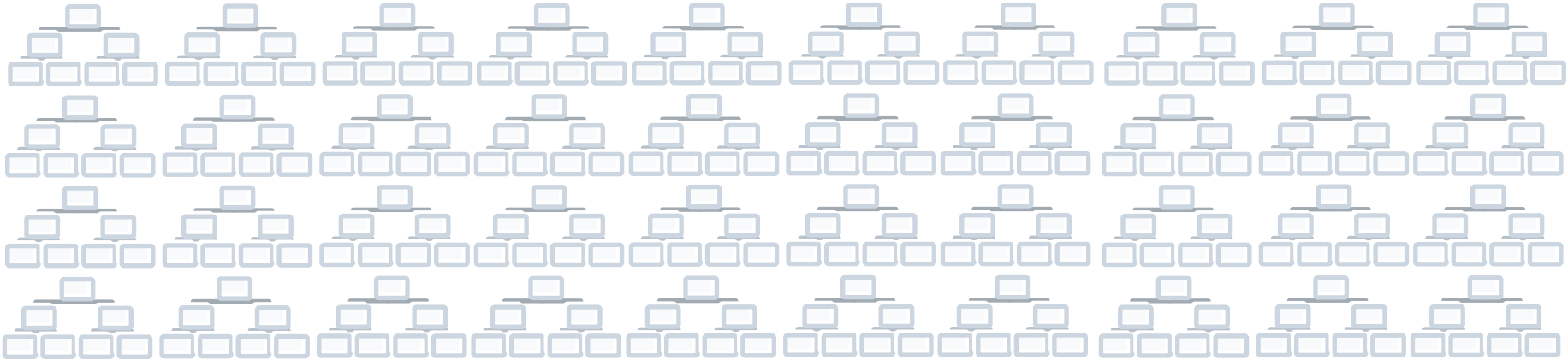
Input:
- Alter, Geschlecht, Herkunft
- Wochentag, Uhrzeit
- Produktvariation, letzte Bestellung, Beziehungsdauer, Kaufintervall, Umsatz pro Kauf

Ergebnis:
- Stark generalisierendes Modell
- Lediglich 5 von 26 Produkten werden berücksichtigt



Vorhersage mit Random Forests

- 1000 verschiedene Decision Trees
- Eingeschränkte Auswahl an Input-Variablen, um unterschiedliche Decision Trees zu erhalten
- Prognose erfolgt durch Mehrheitsentscheid



Vergleich der Vorhersagemodelle

$$\text{Vorhersagegenauigkeit} = \frac{\text{Korrekte Vorhersagen}}{\text{Gesamtanzahl Vorhersagen}}$$

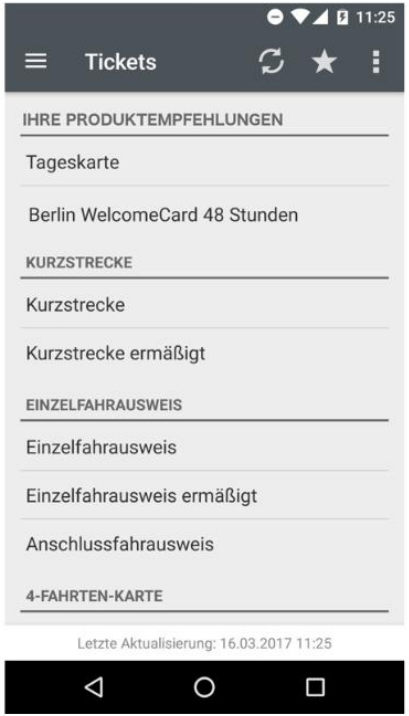
Kann das vom Kunden gewünschte Ticket prognostiziert werden?

Modell	Vorhersagegenauigkeit
Baseline	53%
Decision Tree	58%
Random Forest	64%

Prototypische Implementierung und Anwendbarkeit

Aktuelle Ticketauswahl

Mögliche zukünftige Ticketauswahl



2 Produktvorschläge

- Topseller-Genauigkeit: 62%
- Random F.-Genauigkeit: 83%

3 Produktvorschläge

- Topseller-Genauigkeit: 79%
- Random F.-Genauigkeit: 90%

Vielen Dank!

Marten Pfannenschmidt, Freie Universität Berlin
Prof. Dr. Jan Fabian Ehmke, Europa-Universität Viadrina
Frank Schreier, Berliner Verkehrsbetriebe (BVG)

Zur Bewertung des Vortrages
gehen Sie bitte auf die Webseite

www.menti.com

und geben Sie bitte folgenden

Code ein: **21 09 0**