Rein in den Dschungel

Wofür Big Data die Marktforschung braucht.

Dr. Hannes Schettler

IfaD Institut für angewandte Datenanalyse GmbH

Überblick

Vorstellung

Marktforschung oder Data Mining?

Database Enrichment

Machine-Learning-Verfahren

Zwei Fallstudien

Ausblick und Fazit

IfaD

Vorstellung

IfaD
Institut für angewandte Datenanalyse

- Seit fast 40 Jahren Dienstleister für Marktforscher in Instituten und Unternehmen
- Über 50 Mitarbeiter
- Umfassendes Service-Angebot
  - Datenerhebung
  - Datenverarbeitung
  - Präsentation
  - Analyse

Zu meiner Person

- Naturwissenschaftlicher Hintergrund
  - Promotion in der Elementarteilchen-Physik
- Seit 2 Jahren bei IfaD
- Schwerpunkte
  - Statistische Methoden
  - Data Mining
  - Software-Entwicklung

IfaD

Marktforschung vs. Data Mining?

Wozu im Big-Data-Zeitalter neue Daten erheben?

Wozu noch Marktforschung, wenn man sowieso schon alles weiß?

Marktforschung		Big Data
Einstellung, Meinung	**Inhalt**	Verhalten, Beschreibung
Erhebungsaufwand	**Verfügbarkeit**	Liegen ohnehin vor
Hoch	**Qualität**	Bearbeitungsaufwand
Subjektiv	**Validität**	Objektiv
Hoch	**Relevanz**	Analyseaufwand

- Will ein Unternehmen erfolgreich sein, muss es seine Kunden verstehen
- Muss es tatsächlich die Einstellung und Meinung kennen?
- Nicht unbedingt: Mit ausreichend Daten über Stimuli + Reaktionen lässt sich die Frage nach der Meinung umgehen
- Aber: Selten liegen diese Daten ausreichend vor
- Was ist mit echten Innovationen?

IfaD

Marktforschung und Data Mining!

Beide Welten für sich ergeben kein komplettes Bild

Mehrwert durch Verknüpfung von Big Data mit sorgfältig erhobenen Marktforschungsdaten

Quantität und Objektivität von Big Data

Qualität und Relevanz von Marktforschungsdaten

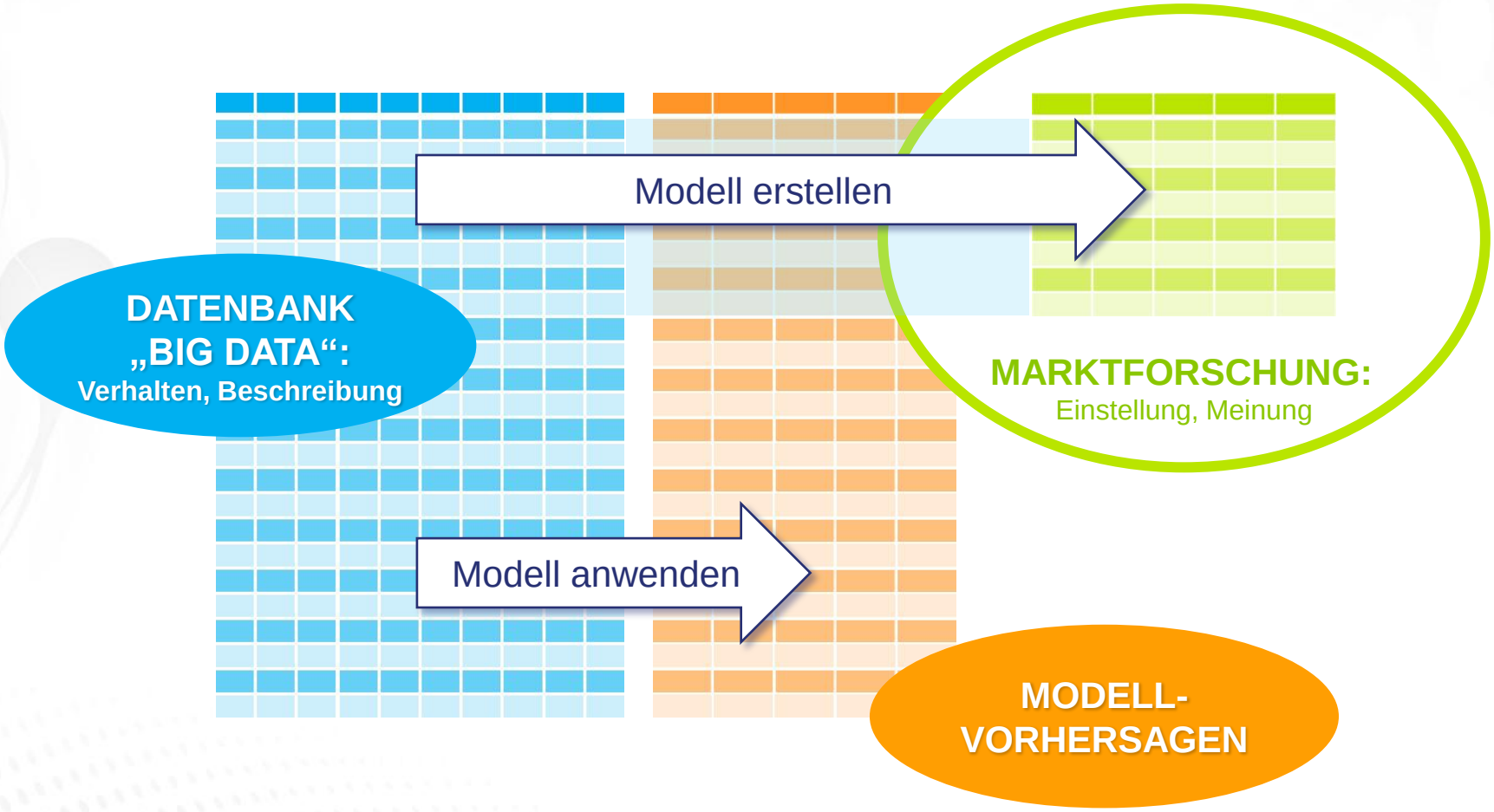
360-Grad-Sicht auf den Kunden

Herausforderung

- Daten-getriebener Ansatz
- Mustererkennung in den Zusammenhängen:
  - Welche Verhaltensweisen und welche Eigenschaften lassen auf welche Einstellung und welche Meinung schließen?
- Fallweise Prognosen der Einstellung und Meinung

IfaD

Database Enrichment



IfaD

Verfahren: Predictive Analytics

Klassische Methoden

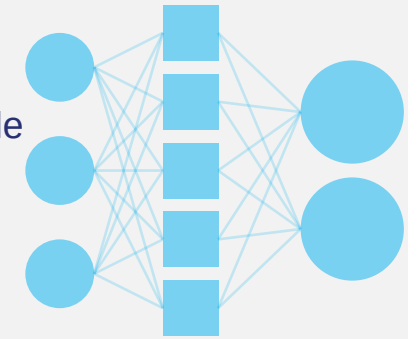
- Diskriminanzanalyse: lineare Klassen-Trennung
- Lineare Regression: linearer Fit der Zielvariable

Intuitive Methode: *k*-Nearest-Neighbors

- Nachbarschaft nach Ähnlichkeit der Prädiktoren
- Schätzung: Mittel/Mehrheit von *k* Nachbarn

Neuronale Netze

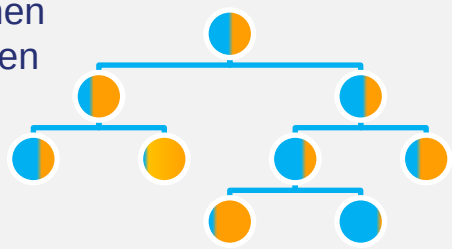
- Netzwerk aus Neuronenschichten
- 1. Schicht: Prädiktoren
  letzte Schicht: Zielvariable
- Ein Hidden Layer ausreichend für viele Anforderungen



- Lernprozess: Anpassung der Gewichte für die Signalweiterleitung
- Hohe Komplexität, aber schwierig zu optimieren

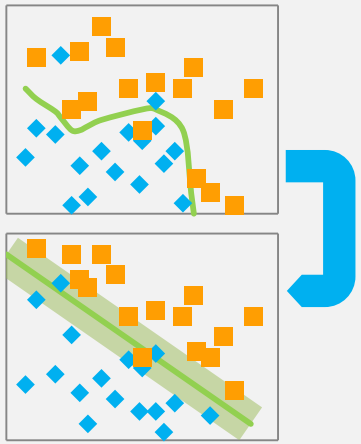
Entscheidungsbäume

- Hierarchische Struktur meist binärer Splits
- Einfach zu verstehen und zu interpretieren
- Mächtige Erweiterungen:
  Boosted Trees
  Random Forests



Support Vector Machines

- Klassen-Trennung mit Hyperebenen
- Nicht-Linearität von Vektor-Transformation absorbiert
- Projektion in hoch-dimensionalen Feature-Raum
- „Kernel-Trick“: Vektor-Transformation nur implizit



IfaD

Welches Verfahren ist das beste?

Keine allgemeingültige Antwort

Hängt von vielen Faktoren ab:

- Was ist die **Aufgabenstellung**?
- Welche **Anforderungen** gibt es?
- Wie sind die **Daten**?

Fehlende Werte

- Fehlende Werte in wichtigen Prädiktoren?
- Imputation sinnvoll?
- Manche Modelle können mit fehlenden Werten umgehen

Fallzahl

- Geringe Fallzahl kann zu großen Fluktuationen führen
- Einfache Modelle stabiler

Prädiktoren

- Anzahl und Skalenniveaus
- Wechselwirkungen zwischen den Prädiktoren
- Verteilung der Information auf die Prädiktoren

Komplexität

- Das Modell sollte nicht komplexer sein als das Problem
- Lineare Zusammenhänge werden am besten von linearen Modellen erklärt

Best Practice

Testen und vergleichen

IfaD

Fallbeispiele

Illustration anhand von zwei Fallbeispielen

- Reine Befragungen mit jeweils hohen Fallzahlen
- Aufteilung in „Datenbank-“ und „Marktforschungsdaten“ nachträglich
  - Fälle: zufällig, wiederholt
  - Variablen: typisch
- Die Ergebnisse der Verfahren können wirklich getestet werden
- Allerdings: keine echten Verhaltensdaten

1. Der Schuh-Händler

- 40 000 Fälle
- Kunden-Segmentierung
- 5 Typen: praktisch, modebewusst, experimentierfreudig, …
- Kategorische Zielvariable: Klassifikation

2. Der Event-Veranstalter

- 15 000 Fälle
- Zufriedenheits-Score
- Kontinuierliche Zielvariable: Regression

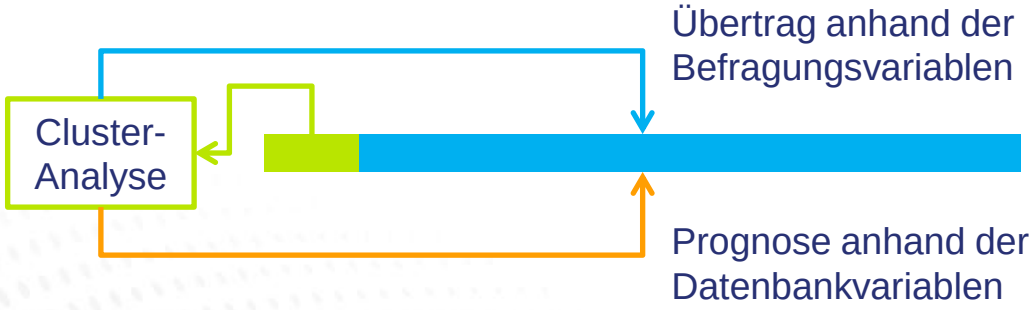
IfaD

Der Schuh-Händler: Test-Setup

Fallzahl: Testen aller Modelle mit verschiedenen Fallzahlen

Fluktuationen: Wiederholte Aufteilung der Fälle in „Befragungs-“ und „Datenbankdaten“

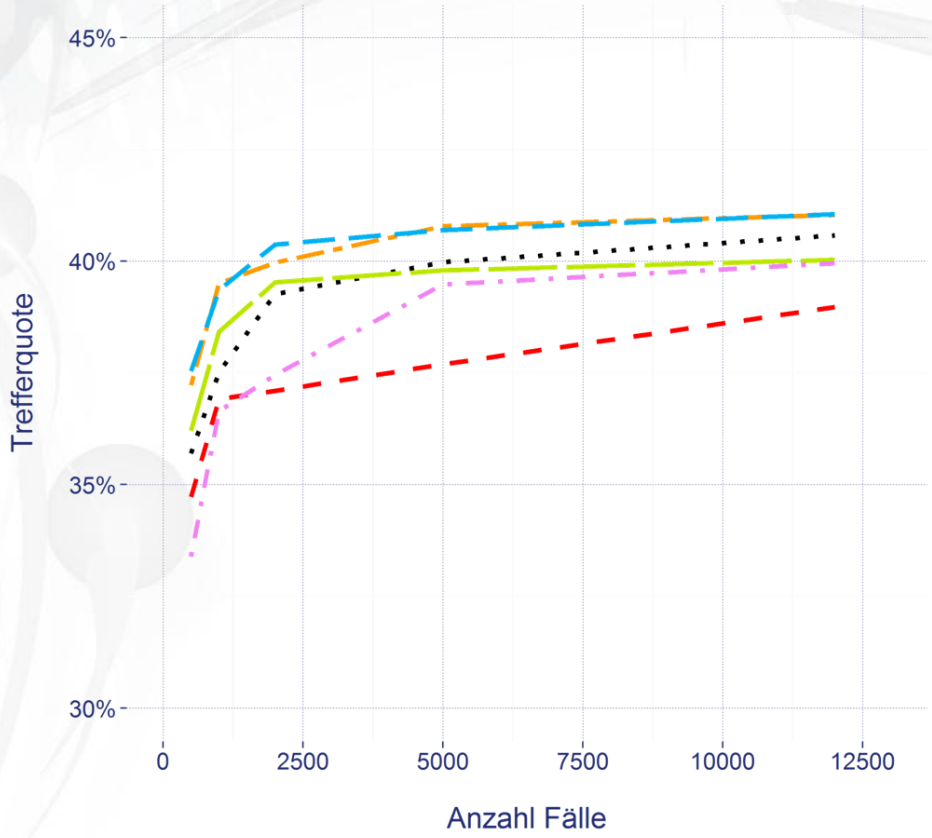
Test

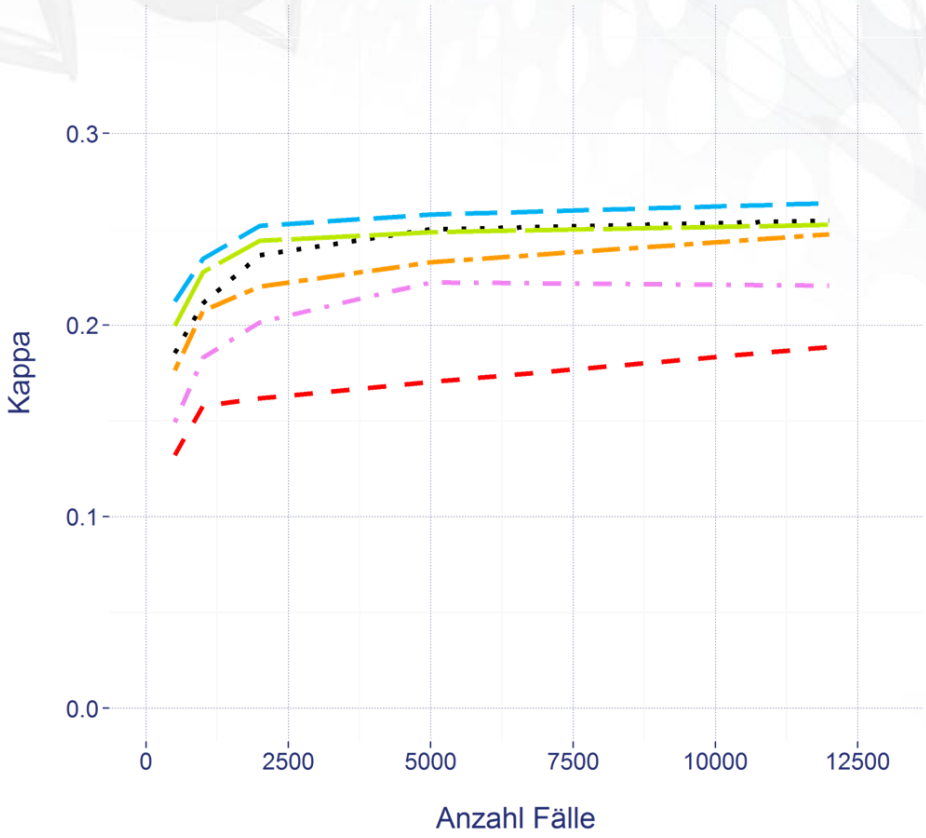


- Durch Übertrag ist der „wahre“ Kunden-Typ für alle Fälle bekannt
- Übertrag nur möglich, da alle Daten aus der Befragung
- Dadurch können Modelle tatsächlich getestet werden

IfaD

Der Schuh-Händler: Ergebnis





Diskriminanzanalyse · Boosted Tree · Neuronales Netz
k-Nearest-Neighbors · Random Forest · SVM

Trefferquote

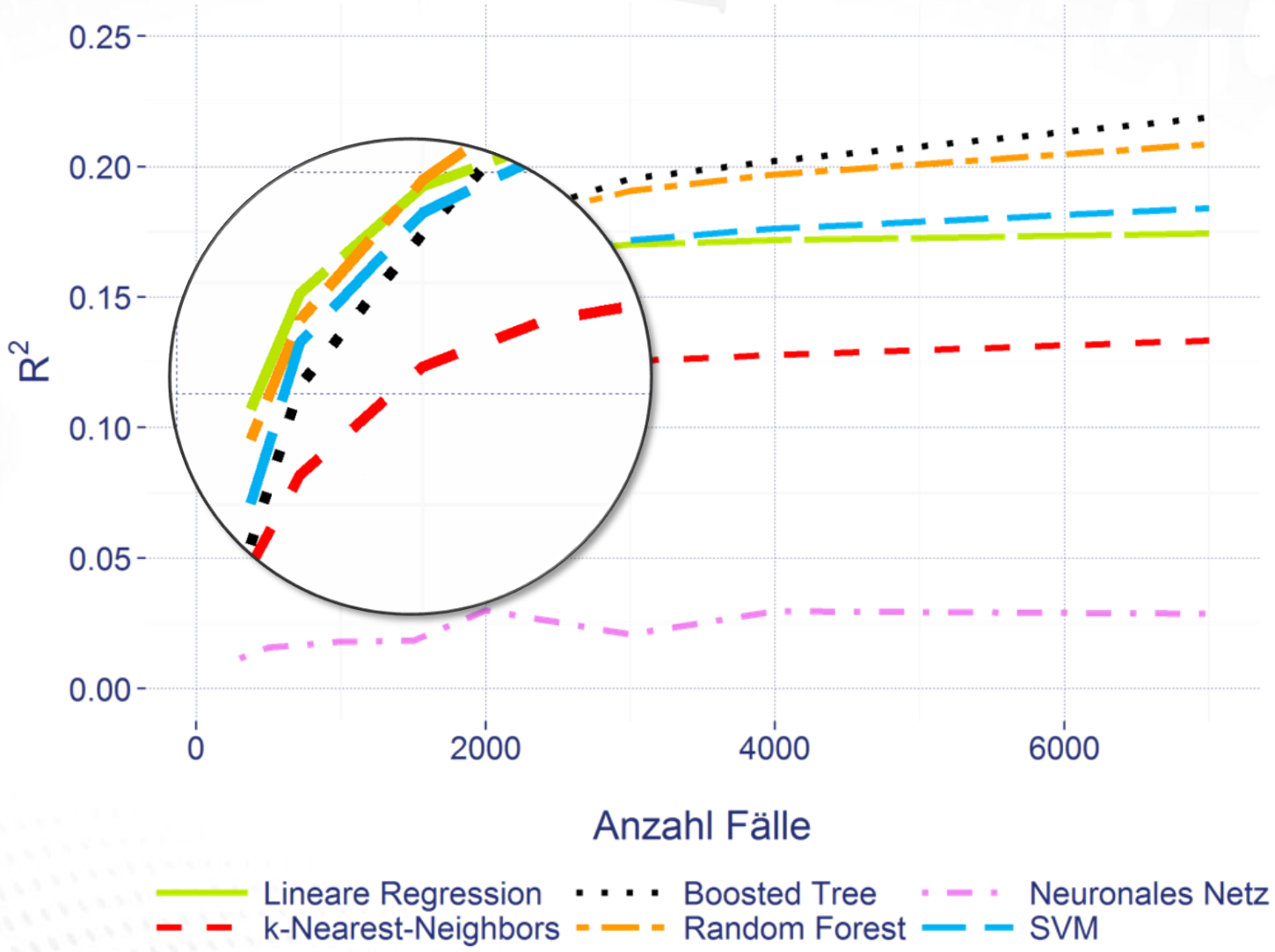
- Anteil richtig vorhergesagter Fälle
- Abhängig von der Anzahl an Klassen
- Problem mit unterschiedlichen Klassengrößen

Cohens Kappa

- Trefferquote relative zum Zufall
- Zufällige Zuordnung $\rightarrow \kappa = 0$
- Perfekte Zuordnung $\rightarrow \kappa = 1$

IfaD

Der Event-Veranstalter: Ergebnis



IfaD

Der Event-Veranstalter: Ergebnis-Einordnung

Bestes Modell:

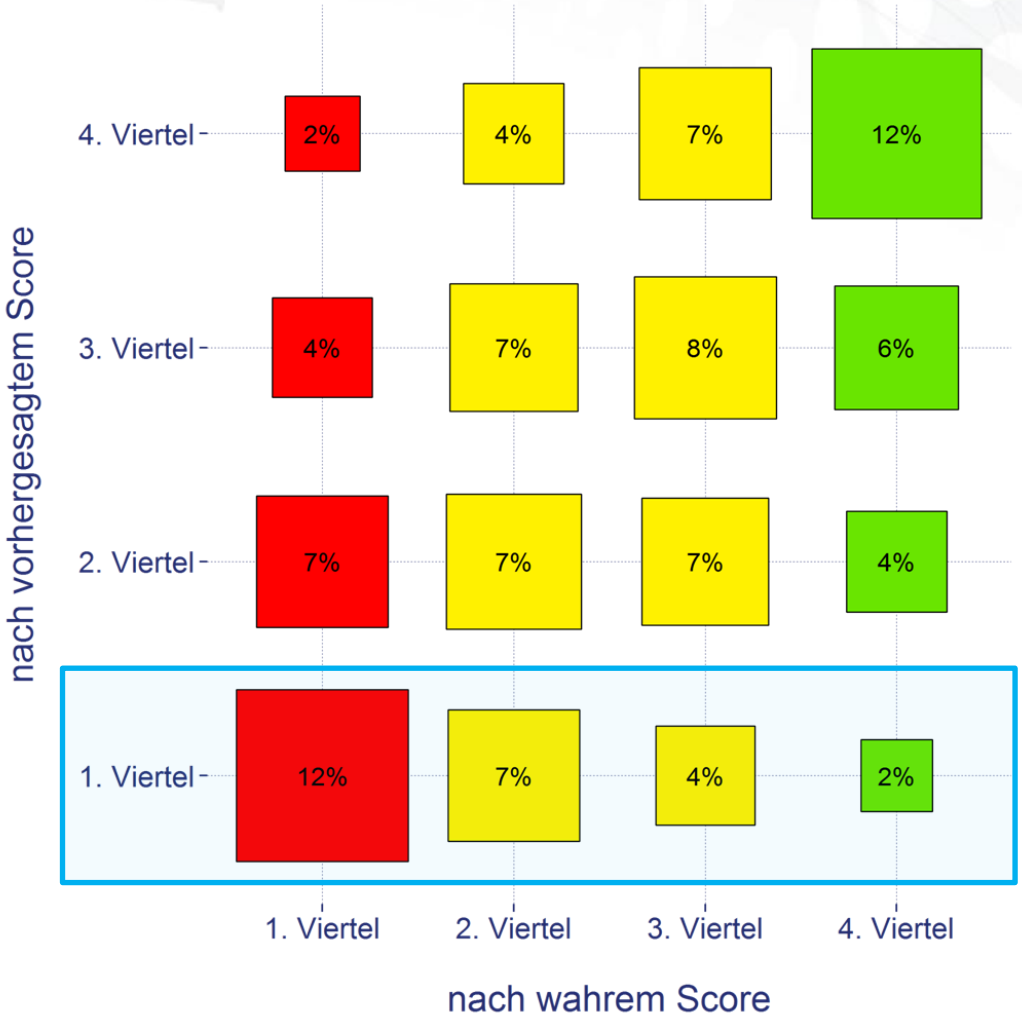
- $R^2 = 0.2$
- Ein Fünftel der Varianz des Zufriedenheits-Scores kann durch das Modell erklärt werden

Gut genug?

- Was steckt in den Daten?
  - Geht es besser?
- Hängt von Fragestellung und Anforderungen ab

Bsp.: spezielles Angebot um möglicher Unzufriedenheit entgegenzuwirken

- Angebot an ein Viertel der Partner:
  50% der Unzufriedenen können erreicht werden



IfaD

Fazit

- Marktforschung kann Big Data um eine neue Qualität ergänzen.
- Data Mining kann Marktforschung nicht ersetzen, wohl aber dazu beitragen, dass Marktforschungsdaten im Big-Data-Umfeld genutzt werden können.
- Mit Machine-Learning-Algorithmen können Erkenntnisse aus der Marktforschung z.B. mit CRM-Daten verknüpft werden.
- Die Modelle müssen sorgsam erstellt und validiert werden.
- Informationserhaltung:
  Was in den Daten nicht drinsteckt, kann man auch nicht rausholen.

Wofür braucht Big Data die Marktforschung?

Um zu lernen, wie von den Eigenschaften und dem Verhalten eines Kunden auf dessen Einstellung zu schließen ist.

IfaD

Raus aus dem Dschungel ...

Vielen Dank
für Ihre Aufmerksamkeit!

IfaD