

www.pwc.de

Aktuarielle Modernisierung: Machine Learning im Pricing der Kompositversicherung

Frank Schönfelder



27.11.2017

qx-Club
Köln/Bonn/
Düsseldorf



pwc

Inhaltsverzeichnis

Kapitel

Seite

1	<i>Potenzial: Machine Learning im Pricing</i>	3
2	<i>Überblick Machine Learning</i>	7
3	<i>Smart Price</i>	12
4	<i>Projektbeschleuniger</i>	24
5	<i>Zusammenfassung & Ausblick</i>	34

Potenzial: Machine Learning im Pricing

1

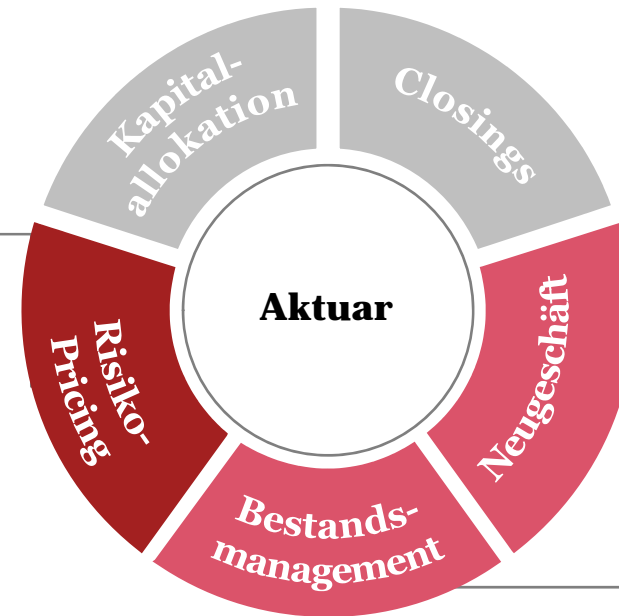
Themen heute

Scope heute

- Anwendung von Machine Learning im Risiko-Pricing im Retail-Segment der Kompositversicherung heute im Fokus

Übliche Charakteristika des Risiko-Pricings

- Modellierung erfolgt **einmal jährlich**, auf Stichtagsdaten mehrerer Jahre
- Ziel ist eine möglichst **feine Differenzierung** der Risiken
- Oft wird getrennt nach Schadenhäufigkeit und Schadendurchschnitt modelliert
- Methodisch werden **GLMs** oder einfachere Verfahren angewandt
- Dabei werden **Abhängigkeiten** meist individuell berücksichtigt, was bei hochdimensionalen Datensätzen eine Herausforderung ist
- Typisch sind systematische Schätzfehler in den „Rändern“ des Bestands



Deploybarkeit: Heute Randthema

- Deploybarkeit von Machine Learning Risiko-Pricing im Bestand und Neugeschäft wird heute nur gestreift

Typische Projektziele & Scope in der Praxis

Status Versicherer

- Verfügt bereits über ein fortschrittliches und ausdifferenziertes klassisches aktuarielle Risikomodell (GLMs), dass weiter genutzt werden soll
- Erste Anwendung von Machine Learning haben begonnen – insb. im Retail-Pricing
- Single Pricing Engine ist idR vorhanden
- Möchte Strukturierung, Know How und Delivery-Unterstützung von externem Berater mit praktischer Implementierungs- und Businesserfahrung

Status Quo PwC

- PwC verfügt über konkrete Projekterfahrung zum Implementierung von Machine Learning Techniken – auch im Retail Pricing
- U.a. auch Pricing Wettbewerb der französischen Aktuarsvereinigung mit Machine Learning Methoden gewonnen
- Ins. Smart Price Architektur zur Modernisierung der Wettbewerbsfähigkeit des Marktpreises von Retail-Versicherern zeigt hohe Praxis-Relevanz

Projektziele

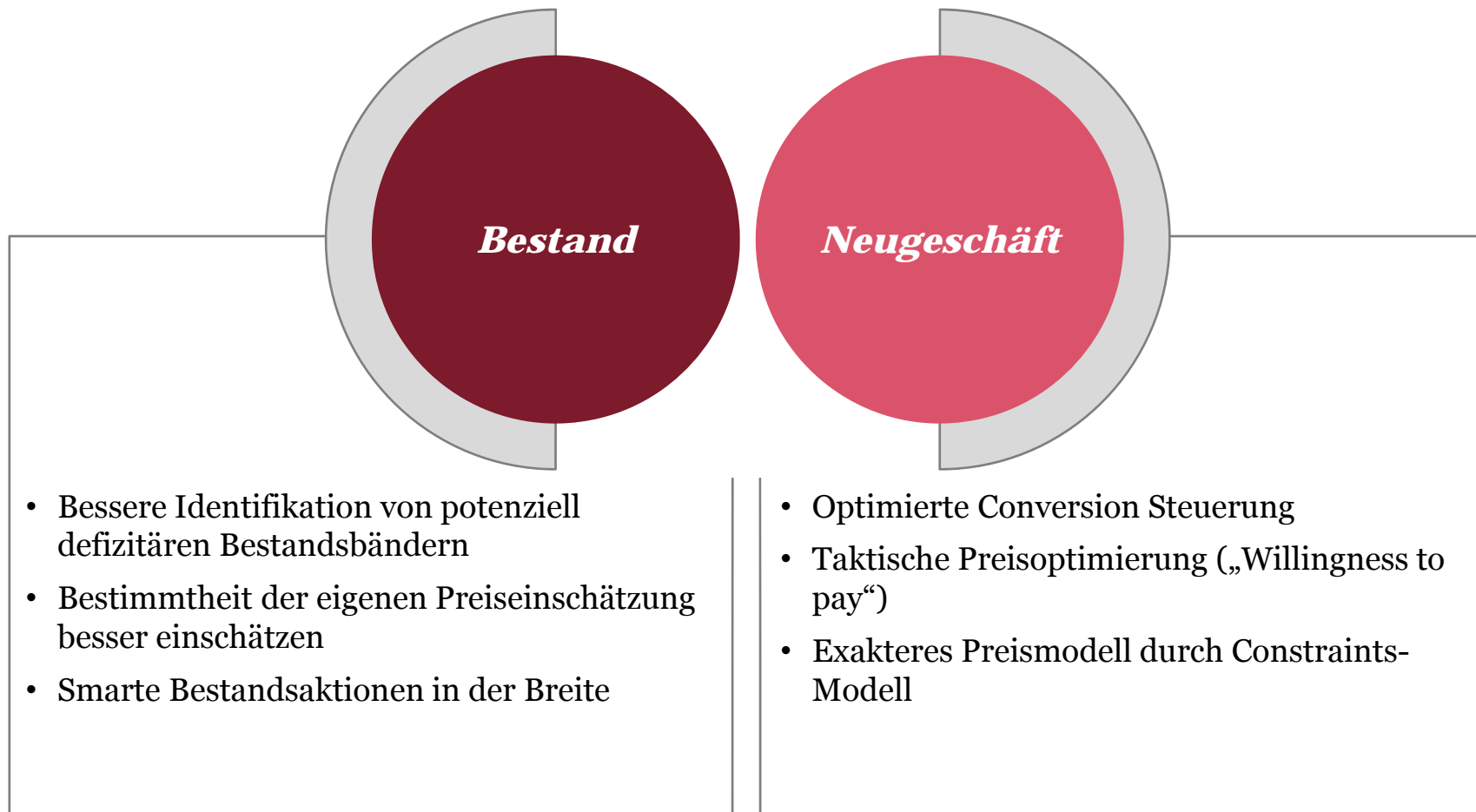
- Synergien und Wissenstransfer zwischen Versicherer und Berater
- Implementierbarkeit von Machine Learning als Ergänzung des klassischen Pricings nachweisen und ersten Prototypen möglichst weit aufbauen – mindestens auf Augenhöhe des GLMs pricen!

Scope

- Risikomodell eines Teil des Kraftfahrt-Retail-Buchs
- Ggf. eingeschränkt auf einen Teil der betrachteten Schadenarten
- Parallele Diskussion über Deploybarkeit in der Sparte – mit der Sparte
- Nutzbarmachung der eigenen Tools und der Tools des Beraters

Deploybarkeitsansätze

Anwenden von Machine Learning Risiko-Preisen im VU

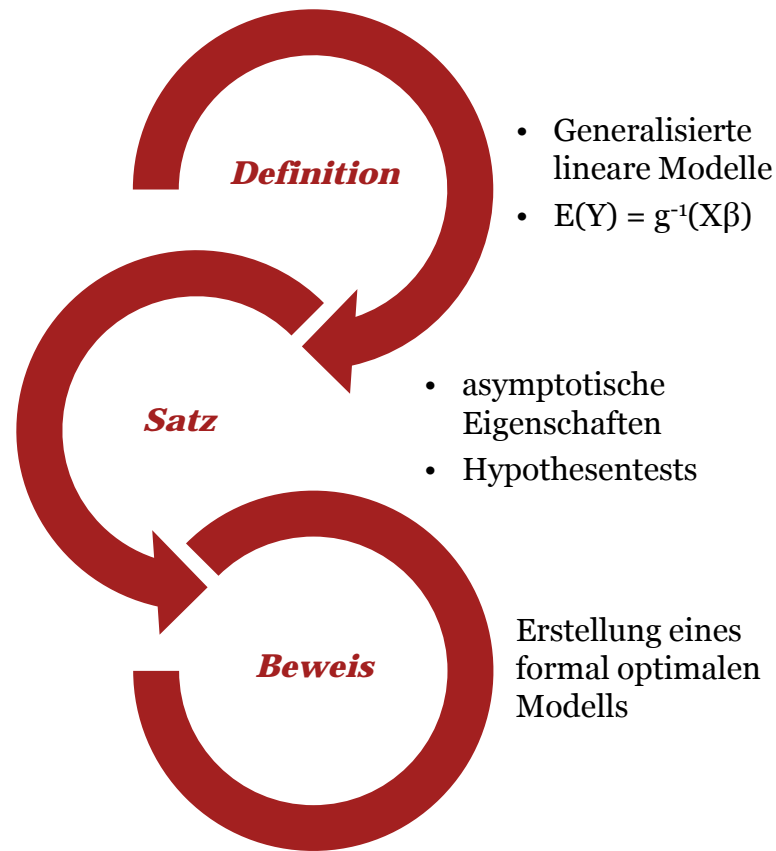


Überblick Machine Learning

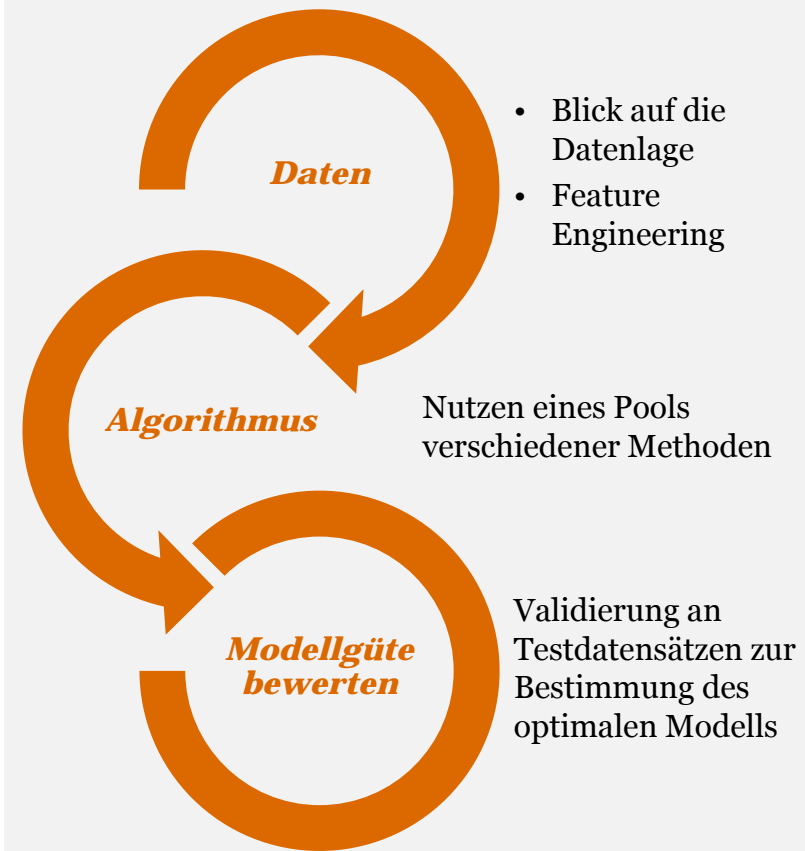
2

Modellierungsstrategien

Klassische Statistik



Machine Learning-Ansatz



Überblick Machine Learning Verfahren

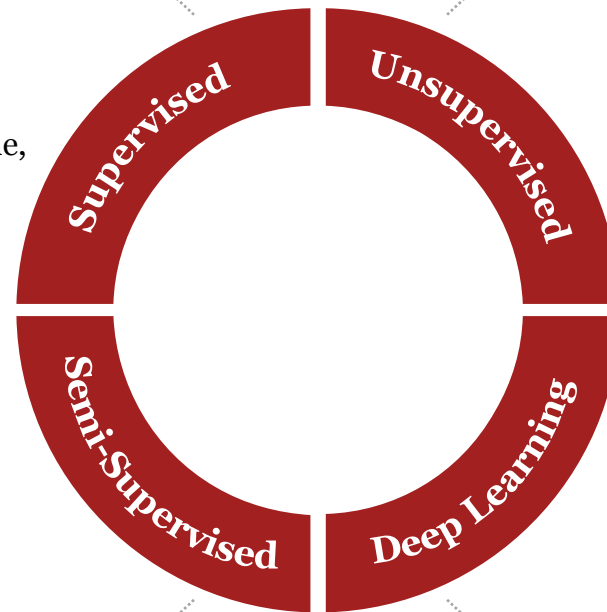
Machine Learning

- Oberbegriff für die künstliche **Generierung von Wissen** aus Erfahrung
- Erkennen von **Mustern und Gesetzmäßigkeiten** in den Lerndaten
- Beurteilung unbekannter Daten → **Lerntransfer**

Überblick Machine Learning-Verfahren

- Es existiert eine Zielvariable (z. B. Schadenaufwand).
- Geeignet zur Tarifierung
- Beispiele: Support Vector Machine, Decision Trees

- Es gibt für manche Daten Zielvariablen in der Beobachtung.
- Z.B. bei Modellierung von Net Promoter Scores
- Beispiele: Schnittmenge aus den obigen Verfahren



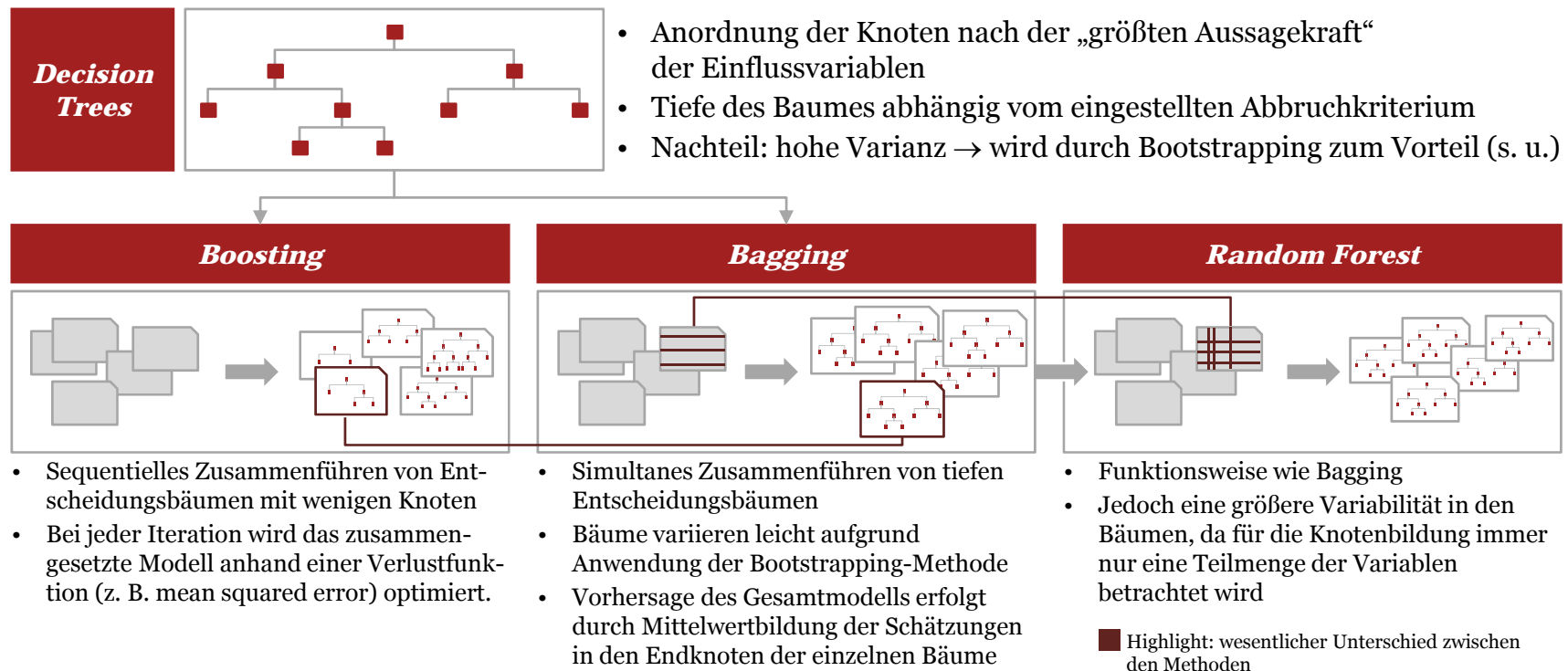
- Datensatz ohne im Voraus definierte Zielwerte
- Geeignet zum Clustern von Daten
- Beispiele: K-Means Clustering, Principal Component Analysis

- Komplexe innere Struktur in den Daten
- Modellierung abstrakter Zusammenhänge
- Beispiele: künstliche neuronale Netze

Eingesetzte Algorithmen

Angewandte Methoden

- Wir haben i. W. entscheidungsbaumbasierte Verfahren des Supervised Machine Learning angewandt.
- Dabei wird auf Basis eines Datensatzes mit bekannter Zielvariable (z. B. Schadenbedarf) ein Modell mithilfe der Einflussvariablen (Bestandsdaten) aufgestellt, um die Zielvariable für noch nicht bekannte Fälle vorherzusagen.



Evaluationsmetriken

Genauigkeit

Mean absolute error

$$\frac{1}{n} \sum_{i=1}^n |\text{Vorhersage}_i - \text{Istwert}_i|$$

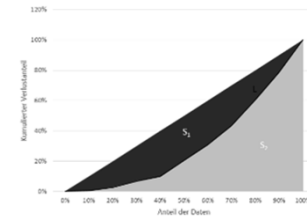
Mean squared error

$$\frac{1}{n} \sum_{i=1}^n (\text{Vorhersage}_i - \text{Istwert}_i)^2$$

Trennungseigenschaft

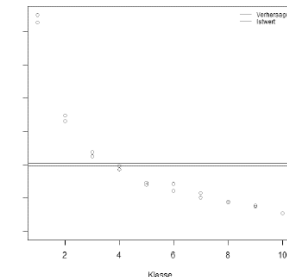
Gini-Koeffizient

- Sortiere die Ist-Werte anhand der jeweiligen Höhe der Vorhersagen.
- Nach diesem Schritt spielt die Höhe der Vorhersagen keine Rolle mehr.



Liftplot

- selbe Sortierung wie beim Gini-Koeffizienten mit anschließender Einteilung in Klassen
- Ein Modell ist gut, wenn:
 - die Abweichung innerhalb der Klassen gering ist
 - bei Betrachtung der Ist-Klassen ein Lift erkennbar ist



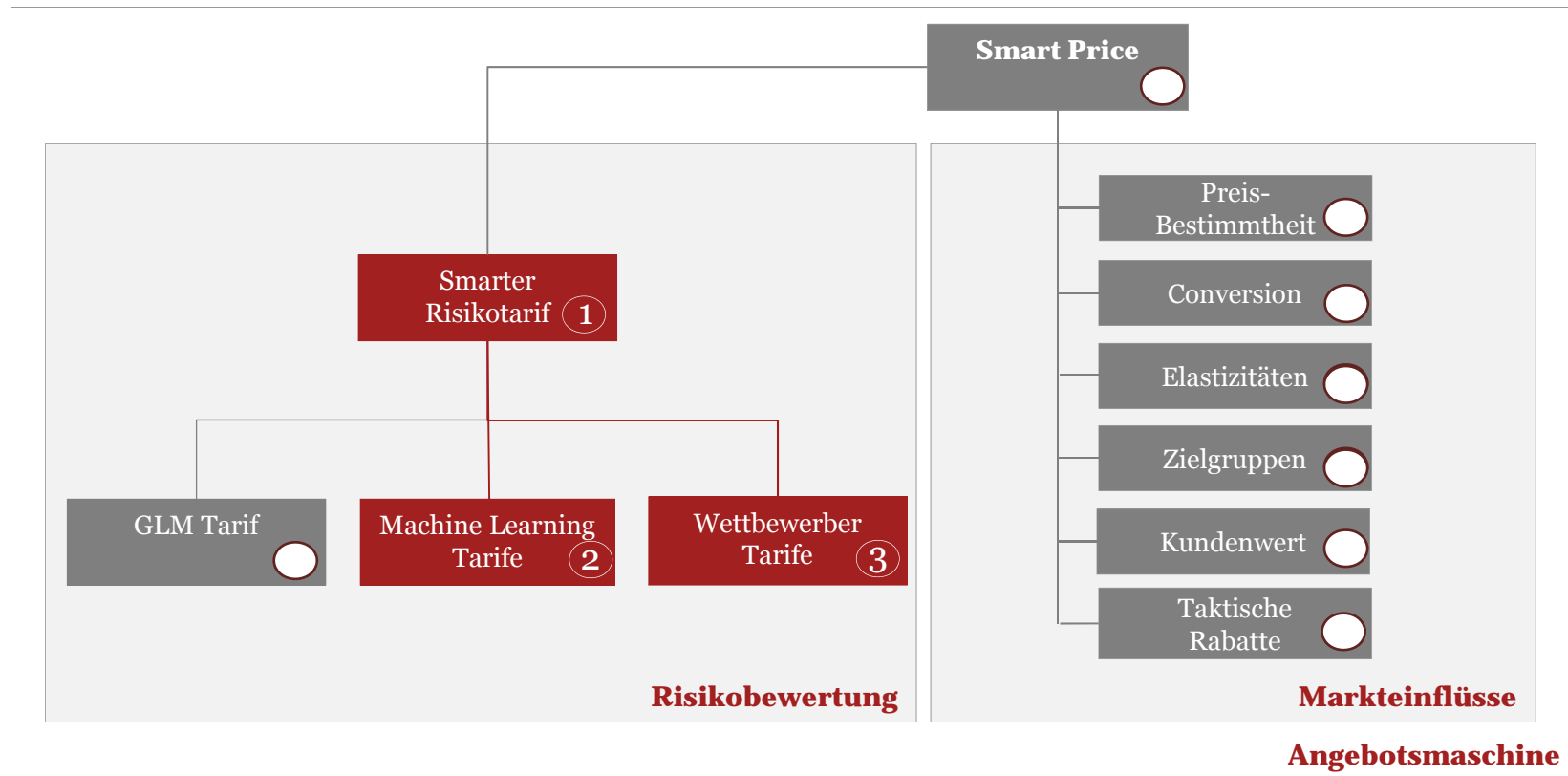
Smart Price

Anwendung von Machine Learning im Pricing

3

Smart Price Architektur

Architektur zur sukzessiven Anreicherung des Pricing Modells



- Module, die Gegenstand des heutigen Vortrags sind
- Weitere Module – werden in diesem Vortrag nicht behandelt

Mehrwerte im Überblick

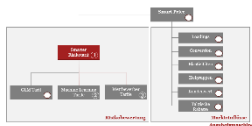
Modul Methode

Ziel

Vorteile

1

Smarter Risikotarif

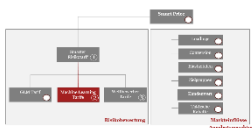


Automatisiertes Mischen der Preismodule zu einem finalen Preis

- Geschicktes **Bündeln aller Module** zu einem Gesamtpreis (Stacking)
- Offen für **variierende Mischung** im Zeitverlauf (z.B. induziert durch Backtesting)

2

Machine Learning Tarife

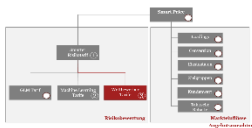


Ergänzende Schätzungen des Risikopreises, basierend auf Entscheidungsbäumen

- Effizientere **Handhabung von Abhängigkeiten** (im Vergleich zu GLM)
- **Automatisches Erkennen** wesentlicher neuer Strukturen
- Einfache Integration von **hochdimensionalen Datensätzen**

3

Wettbewerber Tarife



Verwenden der Preisinformationen der wesentlichen Wettbewerber für die eigene Preisfindung

- **Preise der Wettbewerber** für zusätzliche Erkenntnisse im Detail nutzbar (auch für Risikoeinschätzung)
- **Taktische Preiskomponente** technisch machbar
- **Effizient** machbar (Overfitted Machine Learning)

Smarter Risikotarif

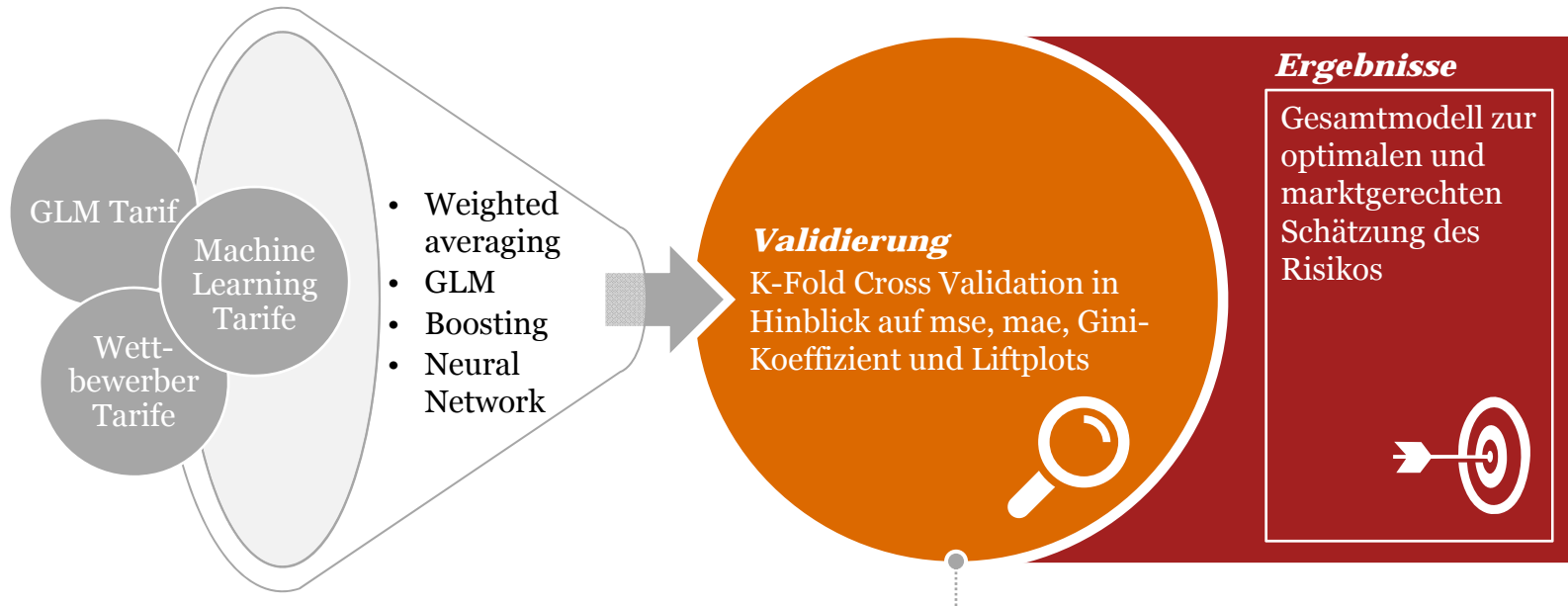
Modul Stacking – Aggregation zum smarten Risikotarif



Stacking



Mehrwert: Stacking ist ein wichtiges Instrument zum Bündeln der Informationen mehrerer Modelle. Mittels eines weiteren Testdatensatzes wird die Stärke und somit Einflussnahme einzelner Modelle erneut evaluiert.



Tipps und Tricks

Selbst vermeintlich schlecht ausdifferenzierte Modelle oder Modelle mit einer geringeren Prognosegüte können förderlich für die Gesamtprognose sein, da sie die Daten nochmal aus einem anderen Blickwinkel betrachten.

Stacking in der Praxis

Stacking am Beispiel GLM-Stacking

```
stacking_model <- glm( # GLM Funktionsaufruf
claims ~ # Zielvariable Schadenbedarf
pen_reg + # Penalized Regression-Schätzung
glm + # GLM-Schätzung
boosting + # Boosting-Schätzung
comp_4, # Wettbewerber Schätzung
data=models, # Datenbasis
family = „gaussian“) # Verteilungsannahme
summary(stacking_model)
```

```
> summary(stacking_model)
```

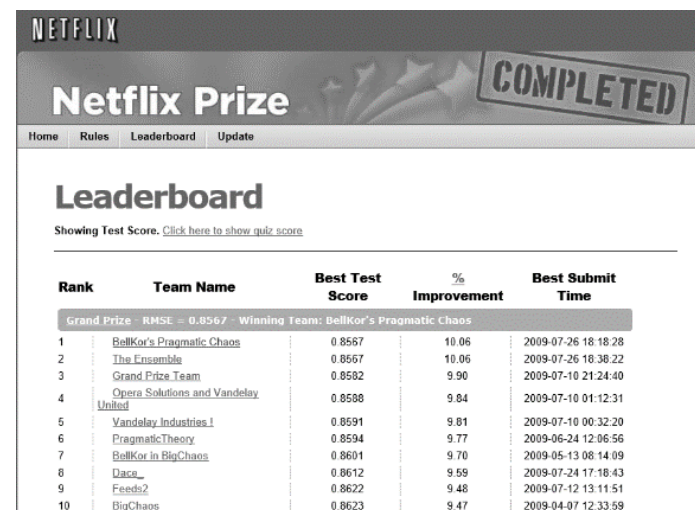
```
Call:
glm(formula = claims ~ pen_reg + glm + boosting + comp_4, family = "gaussian",
data = models)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1599.4   -92.8   -49.2   -26.7   9248.5
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.9352    12.6674   1.968  0.0490 *
pen_reg       0.2137     0.2027   1.054  0.2919
glm           0.4265     0.2044   2.087  0.0369 *
boosting     0.5003     0.2034   2.459  0.0139 *
comp_4      -0.5502     0.2738  -2.010  0.0445 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Netflix Prize

- Wettbewerb des Streaming Unternehmens Netflix
- Fragestellung: Wird ein Film einer bestimmten Person gefallen, basierend auf deren vergangene Filmbewertungen?
- Ziel: übertreffen des gegenwärtigen Algorithmus „Cinematch“ um mindestens 10%
- wichtigste Strategie dabei: **Stacking**



NETFLIX

Netflix Prize

Home Rules Leaderboard Update

Leaderboard

Showing Test Score. [Click here to show quiz score.](#)

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:36:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8598	9.84	2009-07-10 01:12:31
5	Vandalay Industries I	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59

Machine Learning Tarife

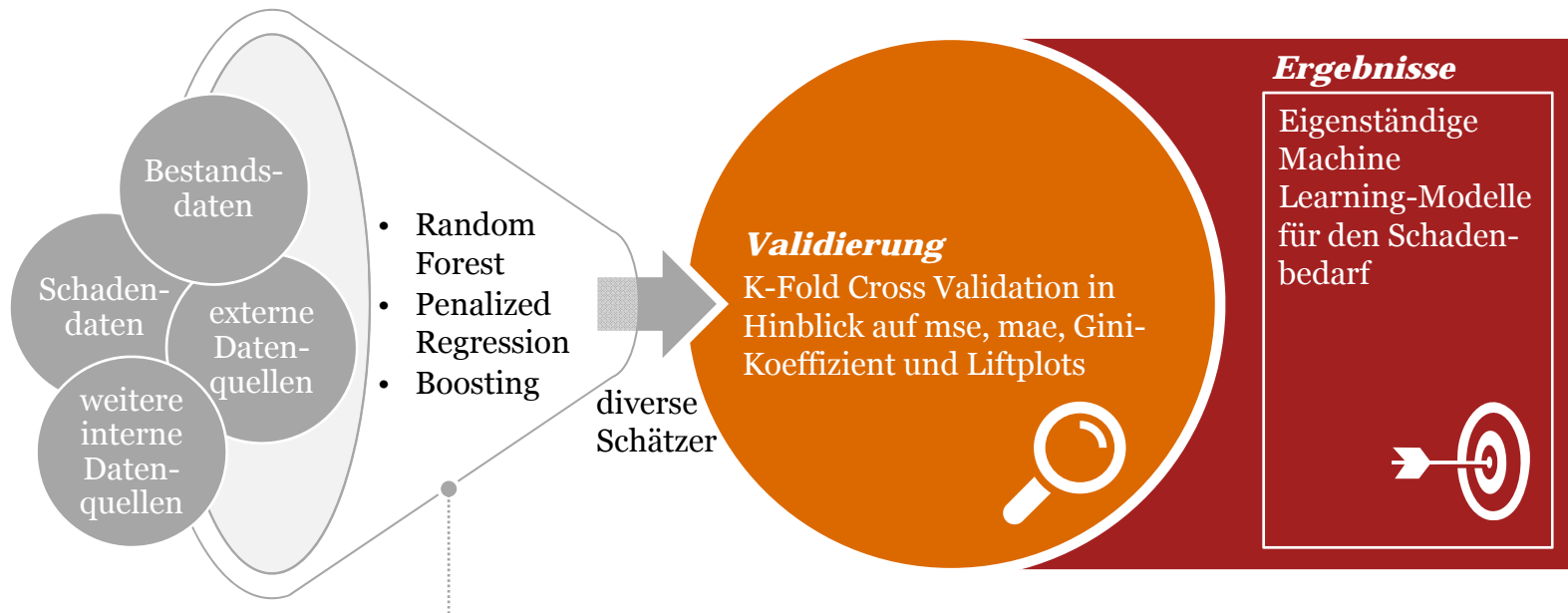
Mit Machine Learning Tarifmodell
vorhandene Daten nutzen



Machine Learning Tarifmodell



Mehrwert: Die Vorteile von Machine Learning-Methoden bei der Erstellung von Risikomodellen kommt besonders bei ihrem **intuitiven Umgang mit hochdimensionalen Daten** zu tragen, wo ein GLM schnell arbeitsintensiv wird.



Tipps und Tricks

Im Gegensatz zum hdtweedie-Paket kann glmnet bei der Berechnung das Exposure einzelner Verträge berücksichtigen. Mit Hdtweedie ist es dafür möglich, den Schadenbedarf direkt zu modellieren.

Random Forest – Einfache Umsetzung in R

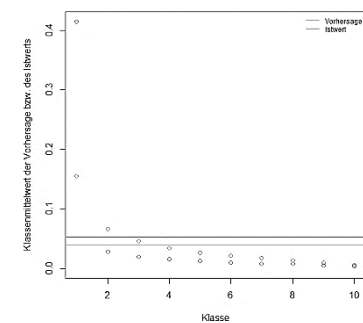
Tunen des Random Forest-Modells mithilfe von Parallelisierung

```
mtry <- seq(1,length(train_new),by=1) #  
Variablenvorselektion  
for (i in mtry){  
  rf <- foreach(ntree=rep(250,4), .combine=combine,  
  .packages=„randomForest“)  
  %dopar%                               # for-Schleife parallel  
  ausführen  
  randomForest                           # Funktionsaufruf und  
  (NB_claims ~.,                          Zielvariable  
  mtry=i,                                  # Variablenvorselektion  
  ntree=ntree                              # Anzahl der Bäume  
  data=train_new)}}                       # Datenbasis  
evaluation                                 # eigens erstellte Teststatistik
```

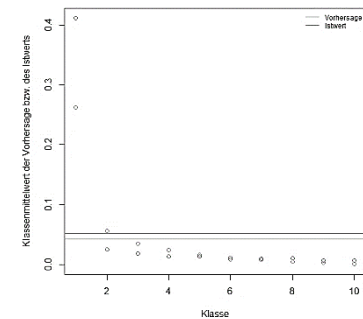
```
> evaluation  
  mtry      mae      mse      Lift      Gini  
1     1 0.08661544 0.05726947 0.1211274 0.3785918  
2     2 0.08659957 0.05728171 0.1235326 0.3790880  
3     3 0.08659989 0.05722510 0.1207841 0.3903983  
4     4 0.08657711 0.05720379 0.1231879 0.3839561  
5     5 0.08652277 0.05721452 0.1214686 0.3892301  
6     6 0.08655616 0.05722569 0.1262860 0.3909687  
7     7 0.08660438 0.05725256 0.1211267 0.3815071  
8     8 0.08657804 0.05724111 0.1255987 0.3868486  
9     9 0.08661319 0.05722993 0.1255951 0.3842008  
10    10 0.08660090 0.05723005 0.1225027 0.3826434
```

Beispieliftplots für verschiedene Modelle

```
rf_3 <- randomForest(  
  formula = NB_claims ~ .,  
  mtry = 3,  
  ntree = 1000,  
  data = train_new)  
predictions <- predict(  
  rf_3,test_new)  
liftplot(test_new$NB_claims,  
  predictions)
```



```
rf_10 <- randomForest(  
  formula = NB_claims ~ .,  
  mtry = 10,  
  ntree = 1000,  
  data = train_new)  
predictions <- predict(  
  rf_10,test_new)  
liftplot(test_new$NB_claims,  
  predictions)
```



Wettbewerber Tarife

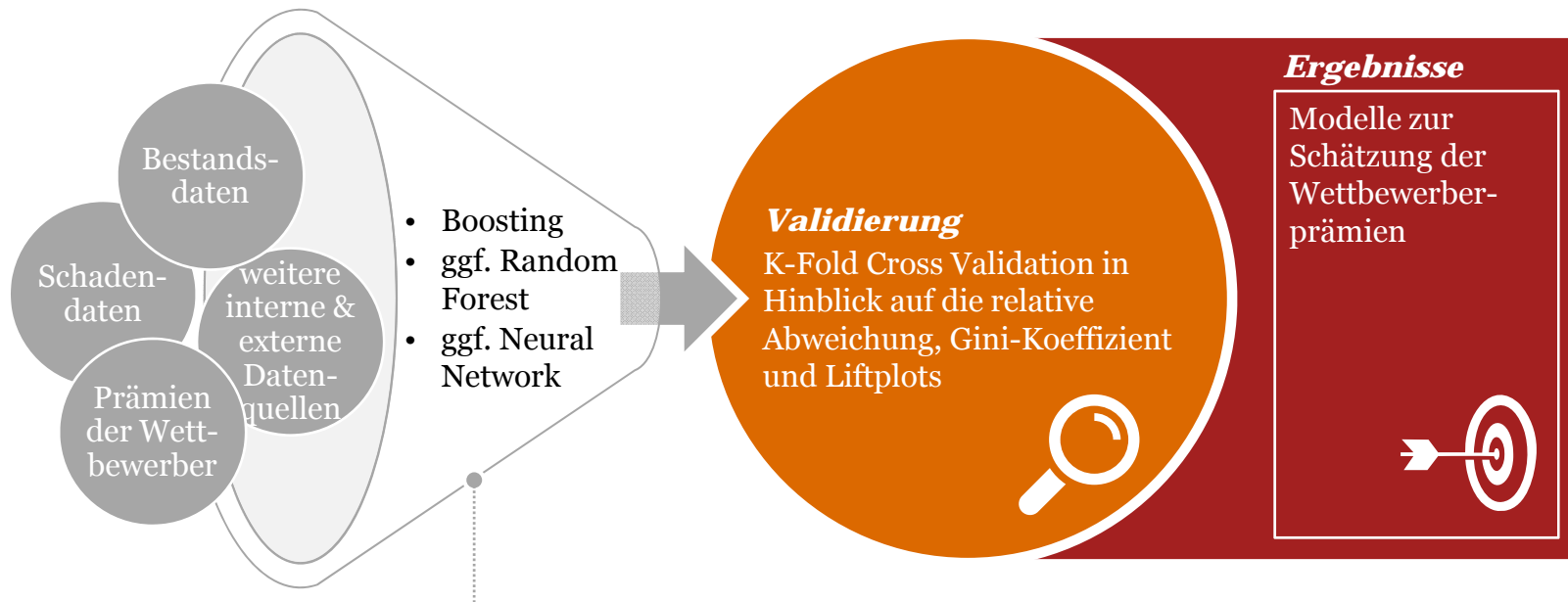
Preise von Wettbewerbern für eigene
Risikoeinschätzung nutzen



Reverse Engineering



Mehrwert: Mittels Reverse Engineering erlangt man ein tiefergehendes Verständnis über die Preisstruktur der Wettbewerber und kann diese effizient für die eigene Preispolitik nutzen.



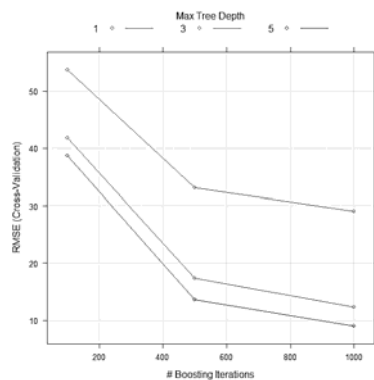
Tipps und Tricks

Bei den Prämien der Wettbewerber handelt es sich um geschlossene Formeln, die wenig Spielraum für Zufall lassen. Daher overfitten wir unser Boosting-Modell bewusst.

Codeauszüge Reverse Engineering

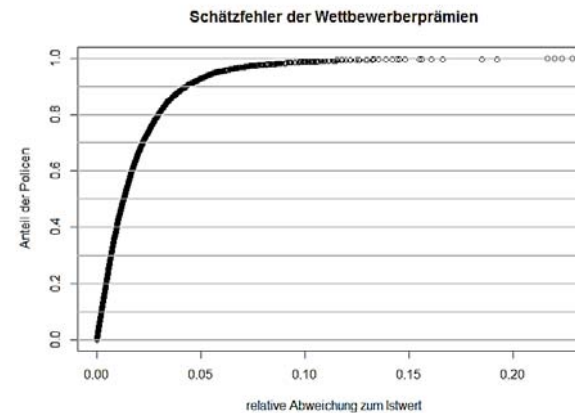
1 *Caret-Paket zum Tunen der Boosting-Parameter*

```
comp_prem_model <- train( # Tuning-Funktionsaufruf
formula = premium_4 ~., # Zielvariable
Wettbewerberprämie
method = „gbm“, # Boosting-Algorithmus
trControl = trcontrol, # Validierungsmethode, vorher
definiert
tuneGrid =tunegrid, # Tuningbereich, vorher definiert
metric = „RMSE“, # quadratische Abweichung
data = data_new) # Datenbasis
plot(comp_prem_model)
```



2 *Testen der Güte des Boosting-Modells*

```
comp_prem_model <- gbm( # Boosting-Funktionsaufruf
formula = premium_4 ~ ., #Ziel- und Einflussvariablen
n.trees = 12000, # Anzahl der Bäume
interaction.depth = 12, # Baumtiefe
shrinkage = 0.01, # Schrittgröße
data = data_new, # Datenbasis
family=„gaussian“) # Quadratfehler-Verlustfunktion
est_error(premiums,predict(comp_prem_model))
```



Projektbeschleuniger

Nutzbare Synergien & nötige Techniken

4

Quellen für die Einarbeitung Aktuariat

Literatur

1. Aktuarielle Methoden der Tarifgestaltung in der Schaden-/Unfallversicherung (2015)

Stellt die wesentlichen Verfahren zur statistischen Auswertung in der Tarifkalkulation sowie die in der Praxis verwendeten Ansätze zur konkreten Modellbildung dar.



frei zugängliches Inhaltsverzeichnis und Einleitung unter:
https://www.vvw.de/details.php?p_id=f55654a6f457c74f49162c4f58270811

2. Insurance Economics (2012)

Gibt einen Einblick in die Versicherungswirtschaft, u.a. in die Grundprinzipien der Versicherungswirtschaftslehre, Adversen Selektion und Moral Hazard.



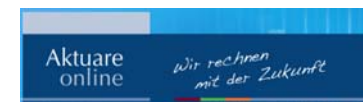
Weitere Literatur:

Farny, Dieter; *Versicherungsbetriebslehre* (2011) Verlag Versicherungswirtschaft GmbH, Karlsruhe

Weiteres

1. Deutsche Aktuarvereinigung (DAV)

Die Homepage der DAV bietet alle Informationen rund um das Thema „Aktuar“ sowie aktuelle Themen aus der Versicherungswirtschaft, u.a. dem Magazin „Aktuar Aktuell“.



Link:

<https://aktuar.de/politik-und-presse/aktuar-aktuell/Seiten/default.aspx>

2. Newsletter Actuarial Services

Bietet 2-3 mal im Jahr einen Einblick in die aktuellen aktuariellen Themen aus Sicht von PwC Europe.



Link:

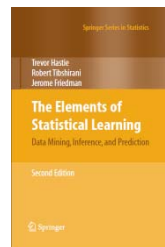
<https://www.pwc.de/de/newsletter/finanzdienstleistung/newsletter-actuarial-services.html>

Quellen für die Einarbeitung Machine Learning

Literatur

1. *The Elements of Statistical Learning*

Bietet einen umfassenden theoretischen und praktischen Einblick über sämtliche gängige statistische Verfahren

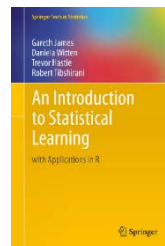


frei zugänglich unter:

<https://statweb.stanford.edu/~tibs/ElemStatLearn/>

2. *An Introduction to Statistical Learning*

Bietet einen praxisorientierteren Einstieg in statistische Verfahren mit vielen Anwendungsbeispielen und R-Programmcode



frei zugänglich unter:

<http://www-bcf.usc.edu/~gareth/ISL/>

Weiteres

1. *In-depth introduction to machine learning in 15 hours of expert videos*

Große Sammlung an Videomaterial für die Anwendung von Machine Learning in R, basierend auf den genannten Büchern und vorgetragen von den Autoren



Link:

<https://www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/>

2. *Stack Overflow*

Plattform zur Lösung von Code-bezogenen und fachlichen Fragestellungen aller Art durch eine Community von Programmierern und Data Scientists



Link:



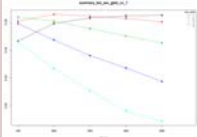
<http://stackoverflow.com/>

Pakete in R

Kategorie	Paketname	Inhalt
Algorithmen	<i>randomForest</i>	Random Forest (und Bagging)-Algorithmus
Algorithmen	<i>gbm</i>	Gradient Boosting-Algorithmus
Algorithmen	<i>xgboost</i>	Gradient Boosting-Algorithmus
Algorithmen	<i>H2O</i>	Paket, das bereits viele Algorithmen inkl. Tuning enthält
Parallelisierung	doSNOW	Registriert ein parallel backend für das foreach-Paket
Parallelisierung	foreach	Optimierung von for-Schleifen
Sonstiges	data.table	Praktische Handhabung großer Datenmengen auf SQL-Basis
Sonstiges	ggplot2	Bibliothek vieler graphischer Darstellungen von Daten
Sonstiges	<i>caret</i>	Optimierung von Methoden anhand verschiedener Merkmale (z. B. Cross Validation)

Pakete frei zugänglich unter <https://cran.r-project.org/>

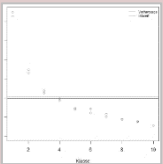
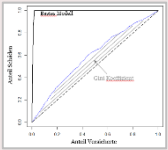

Konkrete Projektbeschleuniger PwC

Kategorie	R Code	Beschreibung	Anwendbarkeit
Architektur - Daten	Datenschnitte 	<ul style="list-style-type: none"> • Übliche Datenflussstrukturen für Multi-Layer Backtesting • Inkl. Unterstützung von Cross-Validation und mehreren Holdouts 	Hoch
Architektur – Machine Learning Modelle	Integrierte Modell-Library 	<ul style="list-style-type: none"> • Schablone für Modellarchitekturen • Inkl. Tuning, Cross Validation und Multi-Layer Stacking 	Hoch
Tuning	Tuningverfahren 	<ul style="list-style-type: none"> • Hyperparameteroptimierung/Tuning • Umsetzung für verschiedene ML-Methoden und Tuning-Verfahren (u.a. Gini) 	Hoch

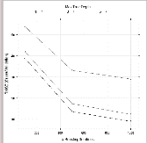
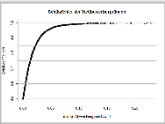
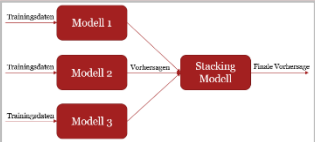
Konkrete Projektbeschleuniger PwC

Kategorie	R Code	Beschreibung	Anwendbarkeit
Algorithmen	Bagging 	<ul style="list-style-type: none"> • Innerhalb Package randomForest • Innerhalb Package rpart mit eigenen Bootstrap Stichproben für Poissonverteilung und Offset 	Hoch
Algorithmen	Random Forest 	<ul style="list-style-type: none"> • Innerhalb Package randomForest • Inklusive Möglichkeit zur Parallelisierung auf verschiedenen Kernen 	Hoch
Algorithmen	Boosting 	<ul style="list-style-type: none"> • Innerhalb Package gbm mit Poissonverteilung und Offset • Innerhalb Package xgboost (teilweise performanter) 	Hoch

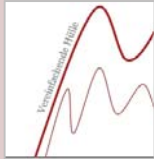
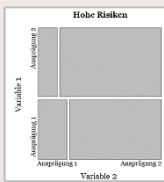

Konkrete Projektbeschleuniger PwC

Kategorie	R Code	Beschreibung	Anwendbarkeit
Backtesting	Liftplot 	<ul style="list-style-type: none"> Liftplot Funktion mit möglicher individueller Anpassung Sehr gut geeignet für eine graphische Analyse Veranschaulichung, in welchen Risikosegmenten das Modell gut bzw. schlecht abschneidet 	Hoch
Backtesting	Gini Koeffizient 	<ul style="list-style-type: none"> Funktion für den Graph der Lorenzkurve sowie für den daraus resultierenden Gini Koeffizient Sehr gute Möglichkeit für einen automatisierten Vergleich mehrerer Modelle und für den genauen Vergleich einzelner 	Hoch
Backtesting	Cross Validation 	<ul style="list-style-type: none"> Implementierung von Cross Validation für das Testen von Modellen Genauer als eine einfache Unterteilung in Trainings- und Testmenge 	Hoch

Konkrete Projektbeschleuniger PwC

Kategorie	R Code	Beschreibung	Anwendbarkeit
Methoden	Reverse Engineering 	<ul style="list-style-type: none"> Boosting Packages und Code für Reverse Engineering mit abgeändertem Tuning Vorgehen Caret Package und Code für noch schnelleres und übersichtlicheres Tuning 	Hoch
Backtesting	Reverse Engineering Statistiken 	<ul style="list-style-type: none"> Abgeänderte Evaluationskriterien zur Bestimmung der Güte des Reverse Engineering Modells Graphische Funktionen zur Veranschaulichung der Güte 	Hoch
Methoden	Stacking 	<ul style="list-style-type: none"> Code für den Stacking Prozess mit verschiedenen Modellen zur Auswahl Darunter auch Code für die spezielle Einteilung der Trainings- und Testdaten 	Hoch

Konkrete Projektbeschleuniger PwC

Kategorie	R Code	Beschreibung	Anwendbarkeit
Methoden	Constraint Modell 	<ul style="list-style-type: none"> Code für die Erstellung eines Constraint Modells Möglichkeit zur Berechnung einer TP/AP Ratio zur Bestimmung des Vorhersagenverlusts des einfachen Modells 	Mittel
Methoden	Mosaikplot 	<ul style="list-style-type: none"> Code für die Erstellung von Mosaikplots für definierte Variablen Für Interaktionen von zwei Variablen sowie für drei Variablen 	Niedrig
Methoden	Zoning Tool 	<ul style="list-style-type: none"> Selbsterstelltes Zoning Tool (etwa vergleichbar mit Radar) Möglichkeit zur individuellen Ausgestaltung von Zonenfaktoren Möglichkeit zum Vergleich für die Vorhersagegüte 	Niedrig

Zusammenfassung & Ausblick

5

Zusammenfassung & Ausblick

- Im Aktuariat ist **Risiko-Pricing** eine der vielversprechendsten Anwendungen von Machine Learning
 - Vorhandene Methoden sollten um Machine Learning ergänzt werden (insb. GLMs) – und werden so **messbar deutlich exakter als bisher**
 - **Machine Learning Pricing wird in der Praxis genutzt** – bei mehreren Versicherern
-
- **Machine Learning wird sich im Pricing durchsetzen**
 - Nicht nur genauere Preiseinschätzung, auch Bestimmtheit der Sicherheit der eigenen Preiseinschätzung erzeugen **wesentlichen wirtschaftlichen Mehrwert**

Vielen Dank für Ihre Aufmerksamkeit.

Frank Schönfelder
Senior Manager, Aktuar DAV
European Market Lead GI



Alsterufer 1
20354 Hamburg

Mobil: +49 175 2290268
frank.schoenfelder@pwc.com