Machine Learning in Non-Life Pricing

Predicting Claims Frequencies using Tree-Based Models

SAV Mitgliederversammlung 2017

C. Buser (AXA)

Finanzielle Sicherheit
neu definiert

Motivation

Daten / Analyse

With data collection, 'the sooner the better' is always the best answer (Marissa Mayer)

The most valuable commodity I know of is information (Gordon Gekko)

Information is the oil of the 21st century, and analytics is the combustion engine (Peter Sondergaard)

Data is not information, Information is not knowledge, Knowledge is not understanding, Understanding is not wisdom (Cliff Stoll and Gary Schubert)

It's easy to lie with statistics. It's hard to tell the truth without statistics (Andrejs Dunkels)

He uses statistics as a drunken man uses lamp posts - for support rather than for illumination (Andrew Lang)

Motivation

Machine Learning

There is no reason and no way that a human mind can keep up with an artificial intelligence machine by 2035 (Gray Scott)

The sad thing about artificial intelligence is that it lacks artifice and therefore intelligence (Jean Baudrillard)

The progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximizing the societal benefit of AI
(Elon Musk, Stephen Hawking, Steve Wozniak)

Mit ihrer Tätigkeit tragen Aktuare eine **bedeutende soziale Verantwortung**. Die verwendeten Modelle wirken sich auf breite Bevölkerungsschichten aus. Wegen komplexen Zusammenhängen ist es oft nicht einfach, Resultate und Begründungen allgemein verständlich zu vermitteln. Es ist daher von zentraler Bedeutung, dass der Berufsstand in der Bevölkerung Vertrauen geniesst. Dieses Vertrauen muss sich der Berufsstand erarbeiten und erhalten (SAV Webseite)

Ausgangslage

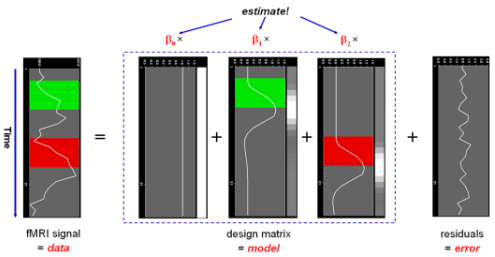
- Im Nichtleben Pricing hat sich Generalized Linear Model (GLM) als Methode zur Modellierung durchgesetzt
- Sie wird sowohl für Schadenfrequenzen als auch für Schadendurchschnitte verwendet
- GLM hat einige Vorteile und sich in der Praxis bewährt
- Es gibt aber noch weitere Methoden zur Modellierung, welche als Alternative verwendet werden können

In Zusammenarbeit mit der ETH haben wir in einer Masterarbeit Tree Based Models gegen die Resultate aus dem klassischen Pricing GLM verglichen:
Zöchbauer Patrick: Predicting Claims Frequencies using Tree-Based Models

- Benutzt wurden Kollisionsschäden (Schadenjahr 2010)
- Zielgrösse war die Schadenfrequenz
- Die Jahre 2011 bis 2014 wurden zur Validierung verwendet
- Daneben haben wir künstliche Daten zum Modellvergleich verwendet

Generalized Linear Model (GLM)

Idee



- GLM ist eine Verallgemeinerung vom linearen Regressionsmodell
- Es gibt eine Zielgrösse / abhängige Variable (z.B. Schadenhäufigkeit) und diverse erklärende Variablen (Alter, Geschlecht, Katalogpreis, etc.), deren Zusammenhang mit der Zielgrösse aufgezeigt werden soll
- Ziel ist die abhängige Variable möglichst gut vorherzusagen und Zusammenhänge zu verstehen (Hypothesen testen und Vorhersage)
- Durch das gemeinsame Berücksichtigen aller erklärenden Variablen wird verhindert, dass Rabatte / Zuschläge doppelt vergeben werden

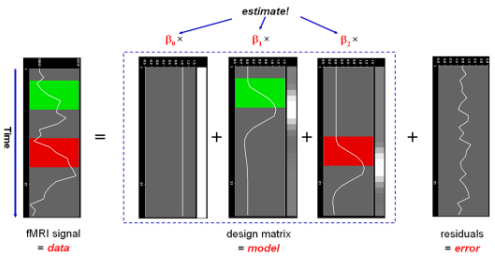
«Frauen fahren weniger Kilometer mit leichteren Fahrzeugen, die umweltfreundlicher sind, …»

Prämie	=	Basisprämie	x	Faktor Selbstbehalt	x	Faktor Leasing	x	Faktor Fahrzeugklasse	x	…

Faktoren aus dem GLM für Frequenz und Durchschnitt

Generalized Linear Model (GLM)

Vorteile



- Ein GLM ist einfach zu interpretieren:
  Ein Frequenz-Faktor 1.10 für Geschlecht bedeutet, dass Frauen 10% mehr Schäden haben als Männer

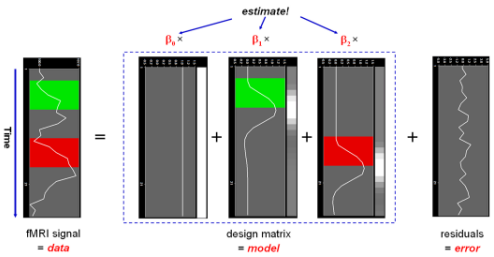
- Dank multiplikativer Struktur können die Resultate einfach in IT Systemen abgebildet werden (Grund-Rechenoperationen genügen)
  Wir übergeben heute den Tarif in Tabellenform an die IT

- Bei GLM wird eine Verteilungsannahme getroffen, oft «Anzahl Schäden ist Poisson verteilt» und «Schadenhöhe ist Gamma verteilt»
  Ist diese Annahme in der Praxis gut erfüllt, liefert GLM stabile Resultate und ist wenig anfällig auf «Ausreisser»

- In kommerzieller Pricing Software ist auch vorwiegend GLM implementiert

Generalized Linear Models (GLM)

Nachteile



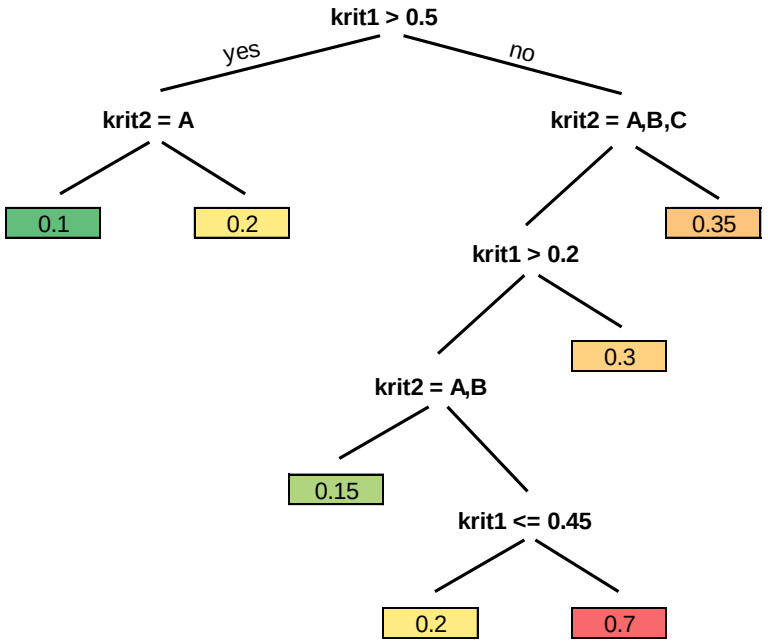
- GLM benötigen a priori Annahmen des Modellierers. Wird genügend Zeit und Knowhow in die Modelle gesteckt, sind die Resultate oft gut. Interaktionen müssen manuell getestet werden
  Es besteht die Gefahr, dass komplexe Zusammenhänge übersehen werden

- Ist die multiplikative Struktur in der Praxis in den Daten nicht gegeben wird das Modell in ein Korsett gezwungen, welches nicht passt. Dasselbe gilt für die Verteilungsannahmen

- Andere Methoden sind flexibler und können daher eine höhere Vorhersagekraft als GLM erreichen

Tree Based Models (Machine Learning)

Beispiel mit 2 Kriterien (erfundene Werte)



- Ein Baum wird von oben nach unten aufgebaut.
- Pro Knoten gibt es genau einen Split in einer einzigen Variablen
- In jedem Knoten und Blatt ist die erwartete Frequenz angegeben
- Ziel ist eine optimale Vorhersage für jedes Blatt (Risikogruppe)
- Ein Baum lässt sich auch als Tabelle zeichnen

Kriterium 2					
	D	0.35			0.2
	C	0.3	0.2	0.7	
	B		0.15		
	A				0.1
		0 bis 0.2	0.21 bis 0.45	0.46 bis 0.5	über 0.5
		Kriterium 1			

Tree Based Models (Machine Learning)

CART Algorithmus

$$R_1(j,s) = \{\boldsymbol{x} : x_j \leq s\} \quad \text{and} \quad R_2(j,s) = \{\boldsymbol{x} : x_j > s\}.$$

Find the splitting variable $\hat{j}$ and split-point $\hat{s}$ of these half-planes such that

$$\min_{j,s} \left[\min_{c_1} \sum_{\{i:x_i \in R_1(j,s)\}} (y_i - c_1)^2 + \min_{c_2} \sum_{\{i:x_i \in R_2(j,s)\}} (y_i - c_2)^2 \right].$$

Fortunately, the inner minimizations are simply given by

$$\hat{c}_1 = \frac{1}{|\{i : x_i \in R_1(j,s)\}|} \sum_{\{i:x_i \in R_1(j,s)\}} y_i \quad \text{and} \quad \hat{c}_2 = \frac{1}{|\{i : x_i \in R_2(j,s)\}|} \sum_{\{i:x_i \in R_2(j,s)\}} y_i.$$

Tree Based Models

Vorteile

- (Kleine) Trees sind einfach zu interpretieren:
  Sie können als Baum oder Tabelle aufgezeichnet werden

- Die Resultate können einfach in IT Systemen abgebildet werden (Grund-Rechenoperationen genügen)
  Eine Übergabe als Tabelle ist ebenfalls möglich

- Die Flexibilität von Trees ist grösser als beim GLM, d.h. Interaktionen werden automatisch abgebildet
  Grössere Flexibilität kann zu besseren Modellvorhersagen führen

- Es müssen keine Verteilungsannahmen getroffen werden (Kein Korsett)

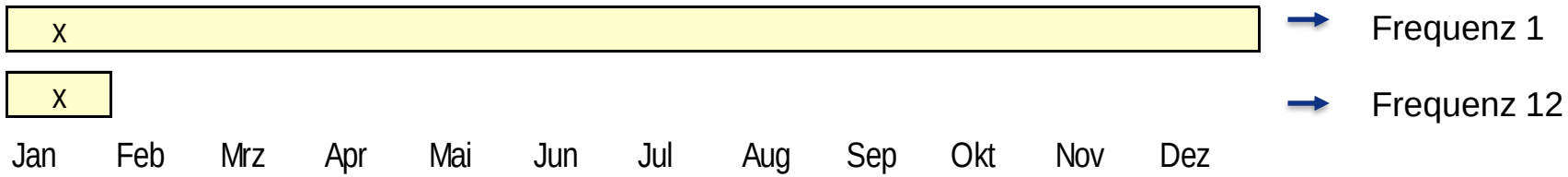
Tree Based Models

Nachteile

- Die Stabilität von Trees kann aufgrund zu hoher Flexibilität schlecht sein
  Die Stabilität kann mit Bagging oder Random Forest verbessert werden
  Es ist wichtig Tree Models auf unabhängigen Daten zu validieren

- Da keine Annahmen zu Verteilungen getroffen werden, besteht die Gefahr, dass Trees überall «gleich schlecht passen»

- Trees sind anfällig auf Ausreisser:



  - In einem GLM haben solche Ausreisser weniger Einfluss
  - Bei Trees besteht die Gefahr, dass nur solche Ausreisser modelliert werden

CART algorithm

Ausreisser – Lösung 1: Ersetze Loss by weighted Loss

$$R_1(j,s) = \{x : x_j \leq s\} \quad \text{and} \quad R_2(j,s) = \{x : x_j > s\}.$$

Find the splitting variable $\hat{j}$ and split-point $\hat{s}$ of these half-planes such that

$$\min_{j,s} \left[\min_{c_1} \sum_{\{i:x_i \in R_1(j,s)\}} (y_i - c_1)^2 + \min_{c_2} \sum_{\{i:x_i \in R_2(j,s)\}} (y_i - c_2)^2 \right].$$

Fortunately, the inner minimizations are simply given by

$$\hat{c}_1 = \frac{1}{|\{i : x_i \in R_1(j,s)\}|} \sum_{\{i:x_i \in R_1(j,s)\}} y_i \quad \text{and} \quad \hat{c}_2 = \frac{1}{|\{i : x_i \in R_2(j,s)\}|} \sum_{\{i:x_i \in R_2(j,s)\}} y_i.$$

Ersetze Standard Loss Funktion mit gewichteter Loss Funktion

$$\min_{j,s} \left[\min_{c_1} \sum_{\{i:x_i \in R_1(j,s)\}} w_i(y_i - c_1)^2 + \min_{c_2} \sum_{\{i:x_i \in R_2(j,s)\}} w_i(y_i - c_2)^2 \right].$$

Zöchbauer, P. (2016). *Data Science in Non-Life Pricing: Predicting Claims Frequencies using Tree-Based Models*. M.Sc. Thesis. Department of Mathematics, ETH Zurich

Poisson trees

Ausreisser – Lösung 2: Verteilungsannahme um die Vorhersage zu verbessern

- Benutze die Poisson Annahme und ersetze Loss Funktion durch die scaled deviance. Minimierung der deviance ist äquivalent zur Maximierung der Maximum Likelihood Funktion

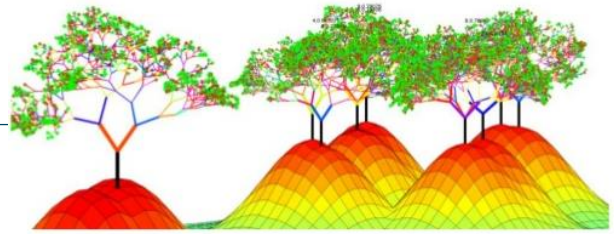
$$\min_{j,s}\left[\min_{\lambda_1}\sum_{\{i:x_i\in R_1(j,s)\}}\left(N_i\log\left(\frac{N_i}{\lambda_1 v_i}\right)-(N_i-\lambda_1 v_i)\right)\right.$$
$$\left.+\min_{\lambda_2}\sum_{\{i:x_i\in R_2(j,s)\}}\left(N_i\log\left(\frac{N_i}{\lambda_2 v_i}\right)-(N_i-\lambda_2 v_i)\right)\right],$$

where the inner minimizations are given by

$$\hat{\lambda}_1=\frac{1}{\sum_{\{i:x_i\in R_1(j,s)\}}v_i}\sum_{\{i:x_i\in R_1(j,s)\}}N_i,\quad\text{resp.}\quad\hat{\lambda}_2=\frac{1}{\sum_{\{i:x_i\in R_2(j,s)\}}v_i}\sum_{\{i:x_i\in R_2(j,s)\}}N_i.$$

Zöchbauer, P. (2016). *Data Science in Non-Life Pricing: Predicting Claims Frequencies using Tree-Based Models*. M.Sc. Thesis. Department of Mathematics, ETH Zurich

Bagging

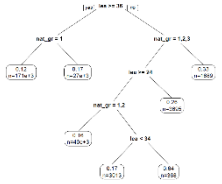


Idee in 2 Minuten

x (Ausreisser)

Daten

x



Ein Baum

Instabil, z.B. hat erster Split viel Gewicht: schlecht wenn dieser nur aufgrund eines Ausreisser entsteht

Viele Bäume

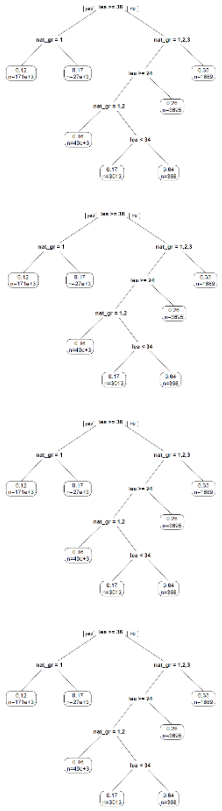
Datensample

Datensample

Datensample

x

Datensample



Mittelwert

Stabil, d.h. ein Split, der nur aufgrund eines Ausreissers passiert, kommt nur in wenigen Bäumen zum Tragen

Dafür ist die Interpretierbarkeit eines gemittelten Baumes schwierig

Random Forest

Idee in 2 Minuten

- Problem korrelierte erklärende Variablen:
  Bsp.: Marketing-Kampagne geschieht immer im Frühjahr wenn mehr Neu-Inverkehrsetzungen passieren
  Hat man nun mehr Neugeschäfte aufgrund der Kampagne oder ist der Grund die Saison, d.h. die zusätzlichen Neu-Inverkehrsetzungen?

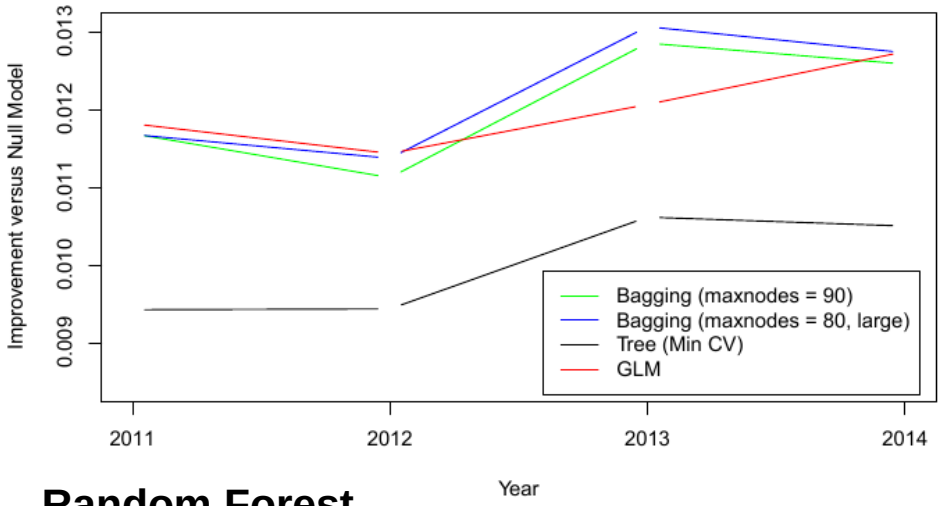
- Mit diesem Problem sind diverse statistische Verfahren konfrontiert und es gibt keine universelle Lösung

- Bei Random Forest (viele Bäume) werden nun nicht nur die Daten gesampelt sondern auch die verfügbaren erklärenden Variablen, d.h. z.B. für den ersten Split eines Baumes darf Lenkeralter nicht verwendet werden
  Damit können Korrelationen etwas durchbrochen werden mit dem Effekt von stabileren Resultaten. Wiederum verschlechtert sich die Interpretierbarkeit: «Man sieht den Wald vor lauter Bäumen nicht»

Resultate

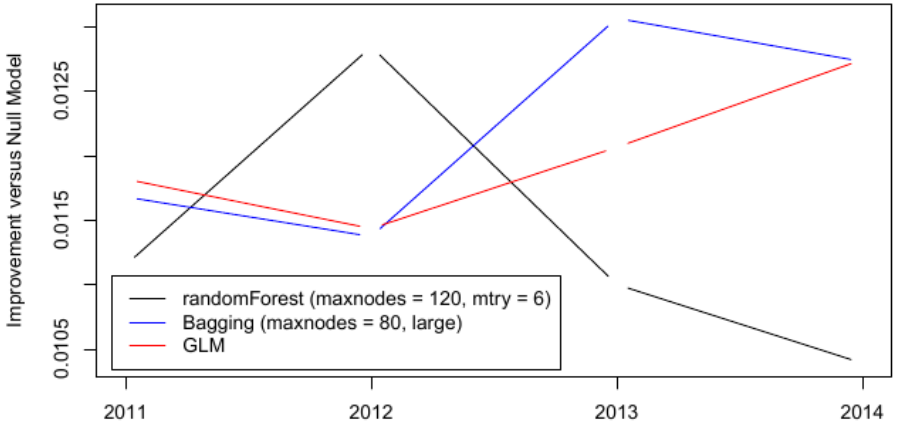
Out of sample deviance

Bagging



- Während GLM bessere Vorhersagen als einzelne Bäume brachte, war die Vorhersage bei Bagging und Random Forest vergleichbar

Random Forest



- Auf den Originaldaten hat Random Forest nicht funktioniert
- Nur mit der Einschränkung auf ganze Jahre haben wir mit Random Forest vergleichbare Resultate erzielt

Zöchbauer, P. (2016). *Data Science in Non-Life Pricing: Predicting Claims Frequencies using Tree-Based Models*. M.Sc. Thesis. Department of Mathematics, ETH Zurich

Poisson Deviance Tree Boosting Machine

Idee in 2 Minuten

Gradient Boosting Machines:
Verbessere ein schwaches Modell schrittweise, indem auf den Residuen ein Regressionsmodell gerechnet und mit diesem das Modellupdate gemacht wird

$$\widehat{f}_{m-1}(\mathbf{x}) \;\rightarrow\; \widehat{f}_m(\mathbf{x}) = \widehat{f}_{m-1}(\mathbf{x}) + \varrho_m \widehat{g}_m(\mathbf{x})$$

Poisson Deviance Tree Boosting Machine:
Beim Poisson Tree kann dies direkt gemacht werden, indem man in Schritt m neue Gewichte definiert

$$w_i^{(m)} = v_i e^{\widehat{f}_{m-1}(\mathbf{x}_i)}$$

Mit diesen Gewichten wird ein Poisson Tree Schätzer gerechnet und fürs Update verwendet

$$\widehat{f}_{m-1}(\mathbf{x}) \;\rightarrow\; \widehat{f}_m(\mathbf{x}) = \widehat{f}_{m-1}(\mathbf{x}) + \nu \sum_{t \in \mathcal{T}^{(m)}} \log(\bar{\mu}_t^{(m)})\, \mathbb{1}_{\{\mathbf{x} \in \mathcal{X}_t^{(m)}\}}.$$

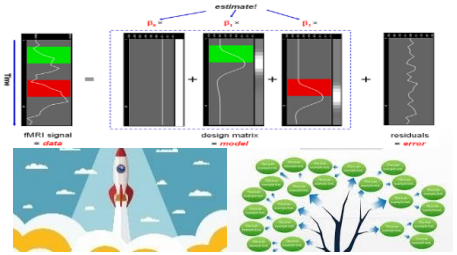
Resultate

Resultate auf unseren echten Daten liegen noch nicht vor

Für Details zur Methode und Resultaten auf synthetischen Daten verweise ich auf:

Wuthrich, Mario V. and Buser, Christoph, Data Analytics for Non-Life Insurance Pricing (March 28, 2017). Swiss Finance Institute Research Paper No. 16-68. Available at SSRN: https://ssrn.com/abstract=2870308

GLM und Machine Learning



Beispiel 1

1. Verwende bestehendes Tarifmodell (z.B. GLM)

2. Verwende einen einfachen Machine Learning Algorithmus (Single Tree oder einzelnen Boosting Schritt) auf den Residuen des Modells aus Schritt 1

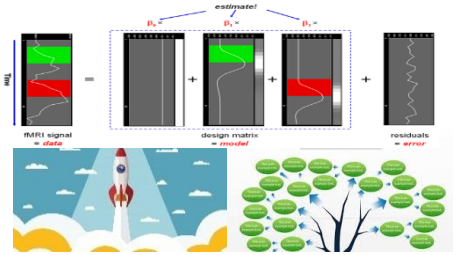
3. Interpretiere die Resultate, z.B. Identifikation einer fehlenden Interaktion zwischen Variablen

4. Füge neu identifizierte Variablen / Interaktionen / Transformationen dem Tarif Modell als Update hinzu

5. Iteriere Schritte 2 – 4 bis keine stabilen Modellverbesserungen mehr erreicht werden

GLM und Machine Learning

Beispiel 2: Gruppierung einer Variablen



- Um Nogagruppen zu bilden müssen ca. 700 Codes Gruppen zugeteilt werden, wobei bei vielen Codes wenige bis gar keine Daten existieren
- Klassisch für ein GLM wird eine Vorgruppierung gemacht und das Modell dann auf dieser Vorgruppierung gerechnet, um dann die finalen Gruppen zu bestimmen

- Alternativ kann die Nogacode Variable mit Hilfe eines Trees gruppiert werden. Dazu kann als Zielgrösse entweder die Original Zielgrösse oder die Residuen eines bestehenden Modells verwendet werden
- Das Resultat kann ins Tarifmodell integriert und gegen eine allfällig bestehende Gruppierung getestet werden

Achtung: ganz ohne Vorgruppierung / Transformation funktionieren auch Trees nicht stabil. Hier können z.B. Credibility Ansätze Abhilfe schaffen

Zusammenfassung

- Die Default Einstellungen von Tree Modellen passten schlecht zu unseren Frequenzmodellen (seltene Ereignisse und Ausreisser)
  Das Problem kann gelöst werden, indem man Poisson-Trees verwendet und/oder eine Gewichtung mit Vertragsdauer (Jahresrisiko) einführt
- Die Stabilität der Trees und damit auch die Vorhersagekraft verbessert sich mit Bagging oder Random Forest, dies allerdings auf Kosten der Interpretierbarkeit
  Diese Varianten erreichen vergleichbare Resultate zum GLM
- Auch bei Machine Learning Methoden funktioniert ein einfaches «Drücken auf den Knopf einer Black Box» nicht immer und kann zu falschen Aussagen führen
  Künstlich simulierte Daten helfen, um die Funktionalität neuer Methoden zu verstehen
- Die Wahl einer vernünftigen Baumgrösse ist ein nicht triviales Problem, wo keine Goldene Regel existiert
- Die Problemstellung war nicht ganz fair und hat GLM etwas bevorzugt
- Tree Modelle können eine Unterstützung bei der Modellierung mittels GLM bringen

Fragen

I keep saying that the sexy job in the next 10 years will be statisticians, and I'm not kidding.
(Hal Varian)

Das Leben steckt voller Überraschungen. Auch vor negativen Einflüssen sind wir nicht immer gefeit. Deshalb wird der Wunsch nach Schutz und Sicherheit immer ein Grundbedürfnis der Menschen sein (SAV Webseite)