Milliman

IT TAKES VISION

# Ein Anwedungsbeispiel für Machine Learning

Bernhard König
Zürich, Schweiz

31. September 2018

# Segmentieren von Schadendaten mit Entscheidungsbäumen

In der Reservierung in der Nichtlebensversicherung schätzt man zukünftige Cashflows von Schäden um damit angemessene Reserven zu bilden.

**Fragestellung: Können wir Teilportfolien identifizieren, welche ein unterschiedliches Abwicklungsverhalten zeigen?**

Wir verwenden einen Machine Learning Ansatz um ein Reservierungsmodelle auf Einzelschadendaten zu erstellen.

Die verwendeten Methoden eigenen sich gut, um ‘unterschiedliche’ Teilportfolios (Segmente) zu identifizieren.

# Chain-Ladder Illustration

Fragestellung: Wie hoch ist der Endschaden pro Schadenjahr?

$$\hat{C}_{i,j} = \hat{f}_{j-1}\, C_{i,j-1}$$

| Cumulative claims loss settlements | | Development year | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Claims occurrence year | 2005 | 1232 | 2178 | 2698 | 3420 | 3736 | 3901 | 3949 | 3963 |
| | 2006 | 1469 | 2670 | 3378 | 4223 | 4684 | 4919 | 4975 | |
| | 2007 | 1652 | 3068 | 4027 | 4981 | 5586 | 5873 | | |
| | 2008 | 1831 | 3465 | 4589 | 5676 | 6401 | | | |
| | 2009 | 2074 | 3993 | 5323 | 6563 | | | | |
| | 2010 | 2434 | 4697 | 6358 | | | | | |
| | 2011 | 2810 | | | | | | | |
| | 2012 | 3072 | | | | | | | |
| CLM estimator for claims loss settlement factor | | | 1,8508 | 1,3140 | 1,2422 | 1,1151 | 1,0491 | 1,0118 | 1,0035 |

$C_{5,2}$

3736+4684+5586+6401 = 20407

3420+4223+4981+5676 = 18300

$f_3 =$ 20407/18300 = 1,1151

Illustrative Beispieldaten

**Source:** A practical guide to the use of the chain-ladder method for determining technical provisions for outstanding reported claims in non-life insurance (Björn Weindorfer, University of Applied Sciences bfi Vienna)

# Hintergrund und Daten

- Der diesem Foliensatz zugrundeliegende Code und die Daten sind öffentlich verfügbar.
- Die Daten wurden mit der ‘Individual Claims History Simulation Machine’ von Andrea Gabrielli und Mario Wüthrich simuliert.
- 12 x 12 Abwicklungsdreieck mit Zahlungen

Jeder Schaden hat folgende Variablen (Erklärende in unseren Modellen):

- LoB: Line of Business (1, 2, 3, 4)
- AY, AQ: Schadenjahr und Schadenquartal
- age: Alter (15 bis 70)
- cc: kategorieller Schadencode mit Werten 1, …. 53 (nicht alle Werte kommen vor)
- inj_part: kategorielle Variable mit Werten 1, … 99 (nicht alle Werte kommen vor)

“Neural Networks Applied to Chain-Ladder Reserving” (M. Wüthrich) modelliert diese Daten mittels Neural Networks.



[https://people.math.ethz.ch/~wmario/simulation.html](https://people.math.ethz.ch/~wmario/simulation.html)

# Motivation

Chain-Ladder liefert oftmals gute Schätzwerte auf für das gesamte Portfolio.

Worin liegt der Mehrwert der Verwendung von Einzelschadendaten?

- Womöglich genauere Schätzungen vom Endschaden
- Detaillierte Schätzwerte für verschiedene Schadenarten (bzw. Subportfolios)
- Schätzung ändert sich wenn sich der Business Mix des Exposures ändert: Wenn im neuesten Jahr prozentual mehr Schäden mit ‘inj_part = 83’ auftreten, dann wird dies durch ein detailliertes Model ‘automatisch berücksichtigt’
- Kann das Verständnis der Schadendaten verbessern
- Erkenntnisse können womöglich andernorts verwendet werden (Claim management, Pricing, …)

# Ansatz

- Gibt es ‘Segmente’ mit unterschiedlichen Chain-Ladder Faktoren?

$$\hat{C}_{i,j} = \hat{f}_{j-1}\, C_{i,j-1}$$

$$\hat{f}_{j-1} = \frac{\sum_{i=1}^{I-j} C_{i,j}}{\sum_{i=1}^{I-j} C_{i,j-1}}$$

| Cumulative claims loss settlements | Claims occurrence year | Development year 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|---|
| | 2005 | 1232 | 2178 | 2698 | 3420 | 3736 | 3901 | 3949 | 3963 |
| | 2006 | 1469 | 2670 | 3378 | 4223 | 4684 | 4919 | 4975 | |
| | 2007 | 1652 | 3068 | 4027 | 4981 | 5586 | 5873 | | |
| | 2008 | 1831 | 3465 | 4589 | 5676 | 6401 | | | |
| | 2009 | 2074 | 3993 | 5323 | 6563 | | | | |
| | 2010 | 2434 | 4697 | 6358 | | | | | |
| | 2011 | 2810 | 4918 | | | | | | |
| | 2012 | 3072 | | | | | | | |
| CLM estimator for claims loss settlement factor | | | 1,8508 | 1,3140 | 1,2422 | 1,1151 | 1,0491 | 1,0118 | 1,0035 |

3736+4684+5586+6401 = 20407

3420+4223+4981+5676 = 18300

20407/18300 = 1,1151

-> finde Teilportfolios mit unterschiedlichen $\hat{f}_{j-1}$

Illustrative Beispieldaten

**Source:** A practical guide to the use of the chain-ladder method for determining technical provisions for outstanding reported claims in non-life insurance (Björn Weindorfer, University of Applied Sciences bfi Vienna)

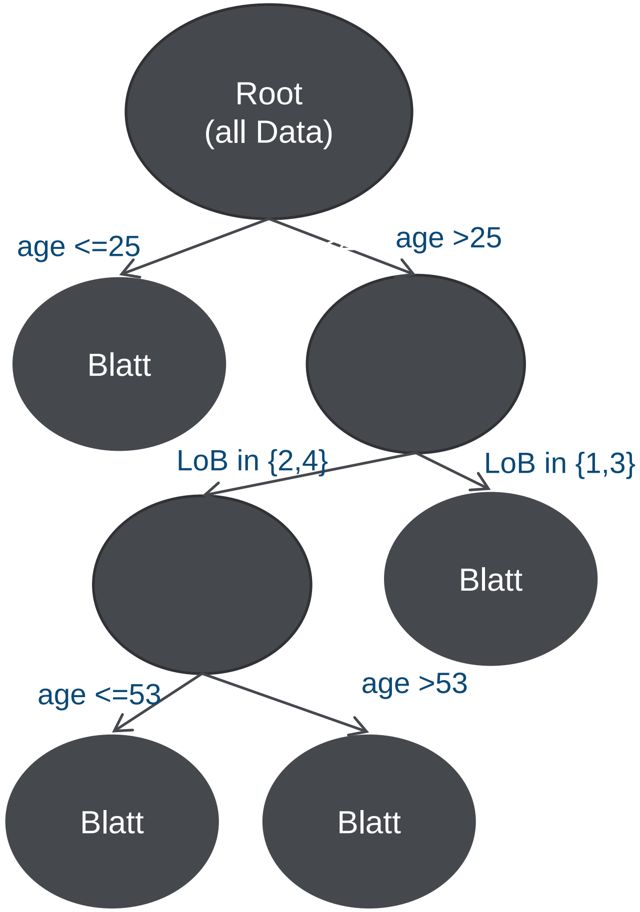
# Entscheidungsbäume

Wir teilen die Daten schrittweise in zwei Teilportfolien auf.

In jedem Schritt werden alle möglichen Splits betrachtet.

- numerische Variable: z.B. ‘age<=25’
- kategorielle Variable: ‘LoB in {2,4}’

Wir betrachten alle möglichen Splits und wählen den ‘besten’ aus.




Welcher Split der beste ist, wird durch eine vom Benutzer gewählten Gütefunktion bestimmt (purity funciton)

Die einzelnen Blätter sollten in sich homogen sein.

Verschiedene Blätter sollte sich unterscheiden.

# Wahl vom Splitting Kriterium

$C_{i,j}$ ist die Summe der einzelnen Zahlungen $P_{j,k}$ k=1:N
wobei N die Anzahl Schäden ist. $P_{j,k}$ ist also die kumulative Zahlung vom Schaden k (zum Stand vom Abwicklungsjahr j)

$$\sum_{i=1}^{I} C_{i,j} = \sum_{k=1}^{N} P_{j,k}$$

Für jeden Split minimiere: $weighted\_sse_{links} + weighted\_sse_{rechts}$

$$weighted\_sse_{links} = \sum_{k \in links} P_{j-1,k} \left[ \frac{P_{j,k}}{P_{j-1,k}} - \hat{f}_{j-1,links} \right]^2$$

Dabei ist ‘links’ und ‘rechts’ durch den Split definiert (z.B. ‘links ≙ ‘age<=25’).
Wobei $\hat{f}_{j-1,links}$ gefittet wird auf allen Beobachtungen im linken Kind.

Wohlgemerkt: Wir betrachten hier keinen ‘information gain’ (sse von der parent node ist nicht relevant). Wir stoppen mit dem Wachstum des Baumes, wenn ein minimum exposure (#Schäden) erreicht ist.

# Verwendete Daten

Wir verwenden 5 Millionen simulierte Schäden*

| Values in CHF Mio - Evaluated as of Dezember 31, 2005<br>Paid Loss - Cumulative | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Accident Year | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1994 | 342.4 | 525.5 | 589.5 | 623.3 | 643.7 | 658.1 | 668.9 | 677.1 | 683.7 | 689.3 | 694.2 | 697.1 |
| 1995 | 336.0 | 524.7 | 593.5 | 627.8 | 649.3 | 663.9 | 675.3 | 683.9 | 690.9 | 696.6 | 701.0 | |
| 1996 | 335.6 | 526.5 | 596.0 | 632.4 | 654.7 | 669.9 | 681.2 | 690.2 | 696.9 | 703.3 | | |
| 1997 | 326.7 | 511.5 | 578.5 | 614.1 | 636.4 | 651.9 | 662.8 | 671.3 | 677.9 | | | |
| 1998 | 324.4 | 516.3 | 589.0 | 626.5 | 649.5 | 665.0 | 676.5 | 685.4 | | | | |
| 1999 | 330.9 | 527.4 | 602.8 | 641.6 | 665.1 | 681.1 | 692.7 | | | | | |
| 2000 | 332.1 | 534.3 | 613.3 | 652.9 | 676.8 | 692.7 | | | | | | |
| 2001 | 333.5 | 541.7 | 623.0 | 663.6 | 687.7 | | | | | | | |
| 2002 | 349.7 | 567.0 | 653.4 | 696.3 | | | | | | | | |
| 2003 | 371.2 | 599.9 | 690.0 | | | | | | | | | |
| 2004 | 381.6 | 620.9 | | | | | | | | | | |
| 2005 | 400.6 | | | | | | | | | | | |

*die Modelle funktionieren auch auf weit kleineren Datensätzen.

# Chain-Ladder

- Offensichtlich gibt es hier zeitliche Effekte.
- Für dieses Anwendungsbeispiel ignorieren wir diese.
- Ebenso wurde ein willkürlicher Tail Faktor von 1 gewählt.

Paid Loss Development

| Accident Year | 12-24 | 24-36 | 36-48 | 48-60 | 60-72 | 72-84 | 84-96 | 96-108 | 108-120 | 120-132 | 132-144 | 144-Ult |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12-1994 | 1.535 | 1.122 | 1.057 | 1.033 | 1.022 | 1.016 | 1.012 | 1.010 | 1.008 | 1.007 | 1.004 | |
| 12-1995 | 1.562 | 1.131 | 1.058 | 1.034 | 1.023 | 1.017 | 1.013 | 1.010 | 1.008 | 1.006 | | |
| 12-1996 | 1.569 | 1.132 | 1.061 | 1.035 | 1.023 | 1.017 | 1.013 | 1.010 | 1.009 | | | |
| 12-1997 | 1.566 | 1.131 | 1.062 | 1.036 | 1.024 | 1.017 | 1.013 | 1.010 | | | | |
| 12-1998 | 1.591 | 1.141 | 1.064 | 1.037 | 1.024 | 1.017 | 1.013 | | | | | |
| 12-1999 | 1.594 | 1.143 | 1.064 | 1.037 | 1.024 | 1.017 | | | | | | |
| 12-2000 | 1.609 | 1.148 | 1.065 | 1.037 | 1.023 | | | | | | | |
| 12-2001 | 1.624 | 1.150 | 1.065 | 1.036 | | | | | | | | |
| 12-2002 | 1.621 | 1.152 | 1.066 | | | | | | | | | |
| 12-2003 | 1.616 | 1.150 | | | | | | | | | | |
| 12-2004 | 1.627 | | | | | | | | | | | |
| 12-2005 | | | | | | | | | | | | |
| Vol Wtd Avg | 1.593 | 1.140 | 1.062 | 1.036 | 1.023 | 1.017 | 1.013 | 1.010 | 1.009 | 1.007 | 1.004 | |
| Vol Wtd Avg Exc Hi/Lo | 1.595 | 1.141 | 1.063 | 1.036 | 1.023 | 1.017 | 1.013 | 1.010 | 1.008 | | | |
| Default | 1.593 | 1.140 | 1.062 | 1.036 | 1.023 | 1.017 | 1.013 | 1.010 | 1.009 | 1.007 | 1.004 | |
| Manual Selected | | | | | | | | | | | | 1.000 |
| Selected | 1.593 | 1.140 | 1.062 | 1.036 | 1.023 | 1.017 | 1.013 | 1.010 | 1.009 | 1.007 | 1.004 | 1.000 |
| Cumulative | 2.169 | 1.362 | 1.194 | 1.124 | 1.085 | 1.061 | 1.043 | 1.030 | 1.020 | 1.011 | 1.004 | 1.000 |
| Ratio to Ultimate | 0.461 | 0.734 | 0.837 | 0.890 | 0.921 | 0.943 | 0.959 | 0.971 | 0.981 | 0.989 | 0.996 | 1.000 |

# Modellansatz

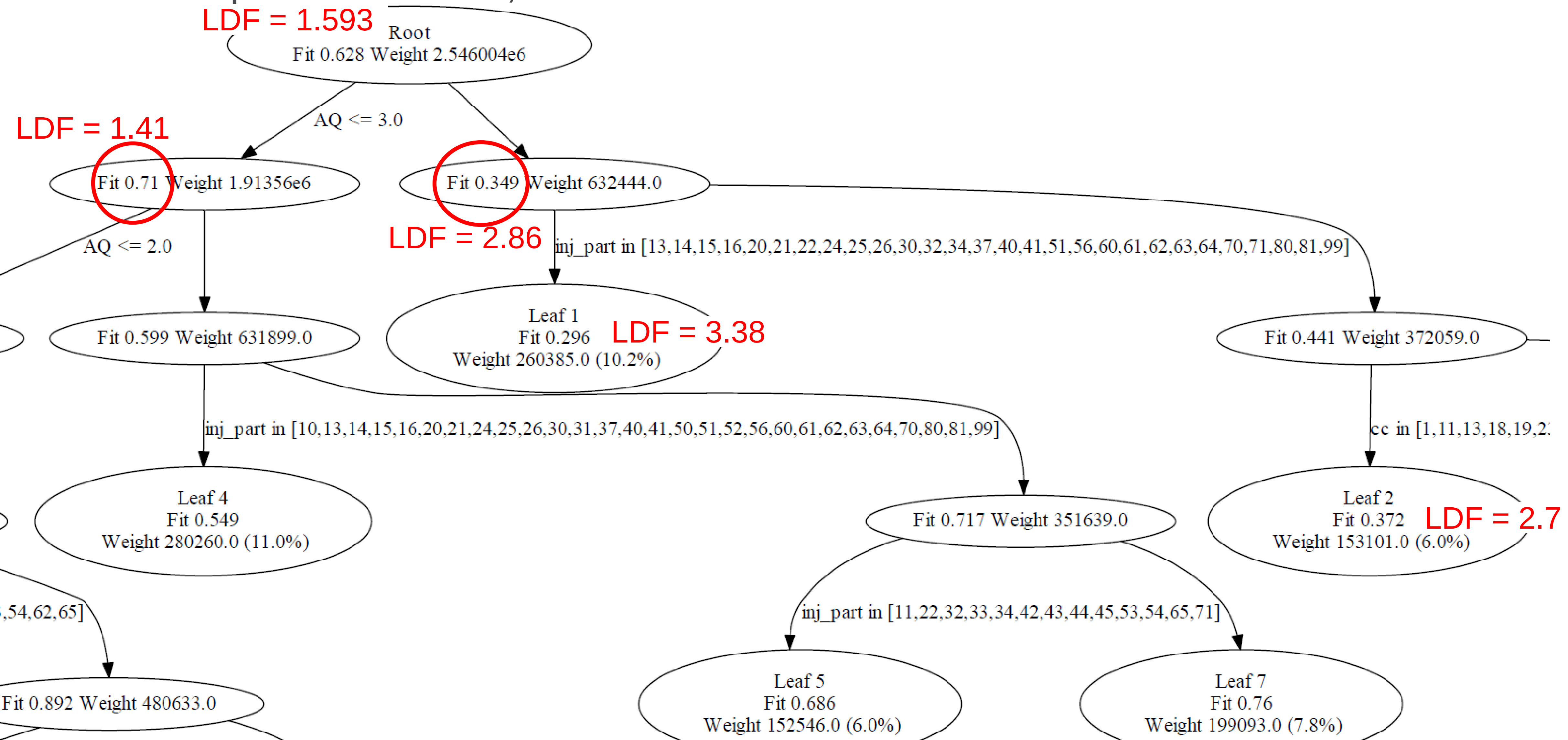
- Wir erstellen einen Entscheidungsbaum für jeden Abwicklungsfaktor (LDF)
- LDF Y1 = 1.593
- Um Division durch 0 zu vermeiden, werden jene Schäden ausgeschlossen für welche $P_{i,j,k} = 0$
- Wir modellieren das Inverse vom Chain-Ladder Faktor um sowenige Schäden wie möglich auszuschliessen.
- Wir teilen die Daten zufällig in 30% Validation- (hold-out) und 70% Trainingsdaten

# Resultat für Jahr 1

LDF Y1 = 1.593 = 1/0.628

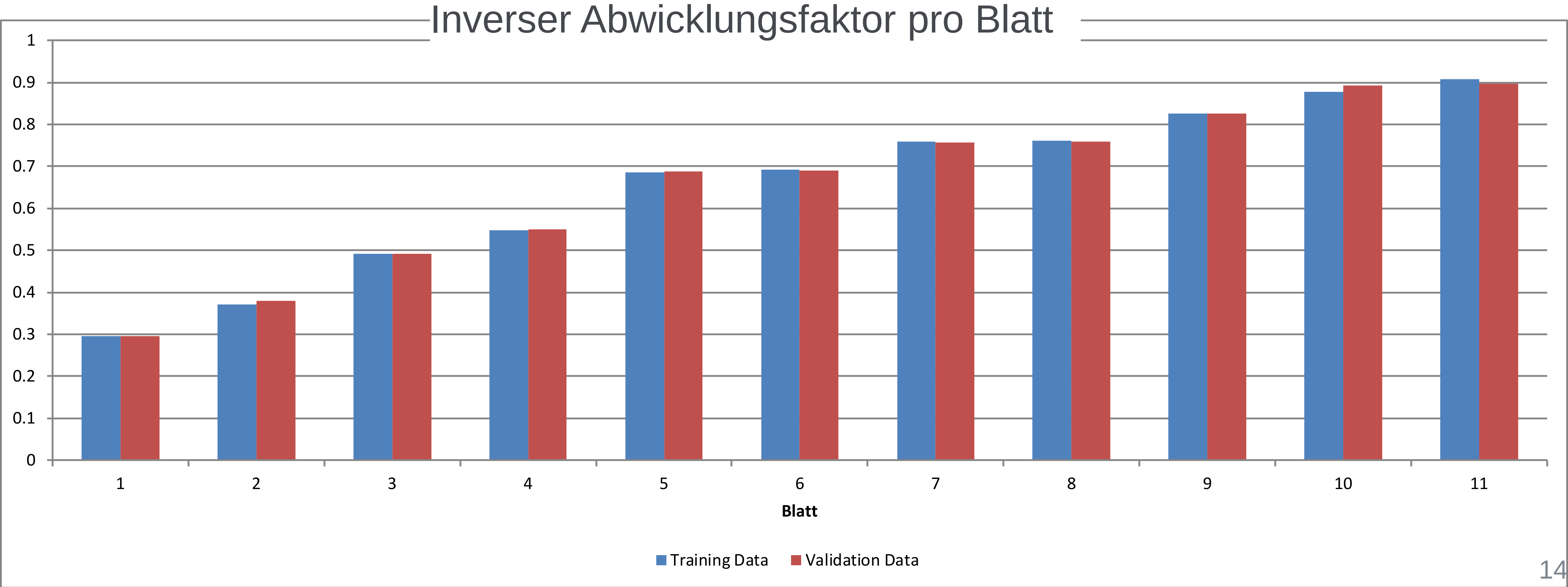
Weight = 70% der Anzahl Zeilen für LDF Y1 (für welche $P_{i,j,k}$ ungleich 0)

Der erste Split findet LDFs, welche sich um Faktor 2 unterscheiden.



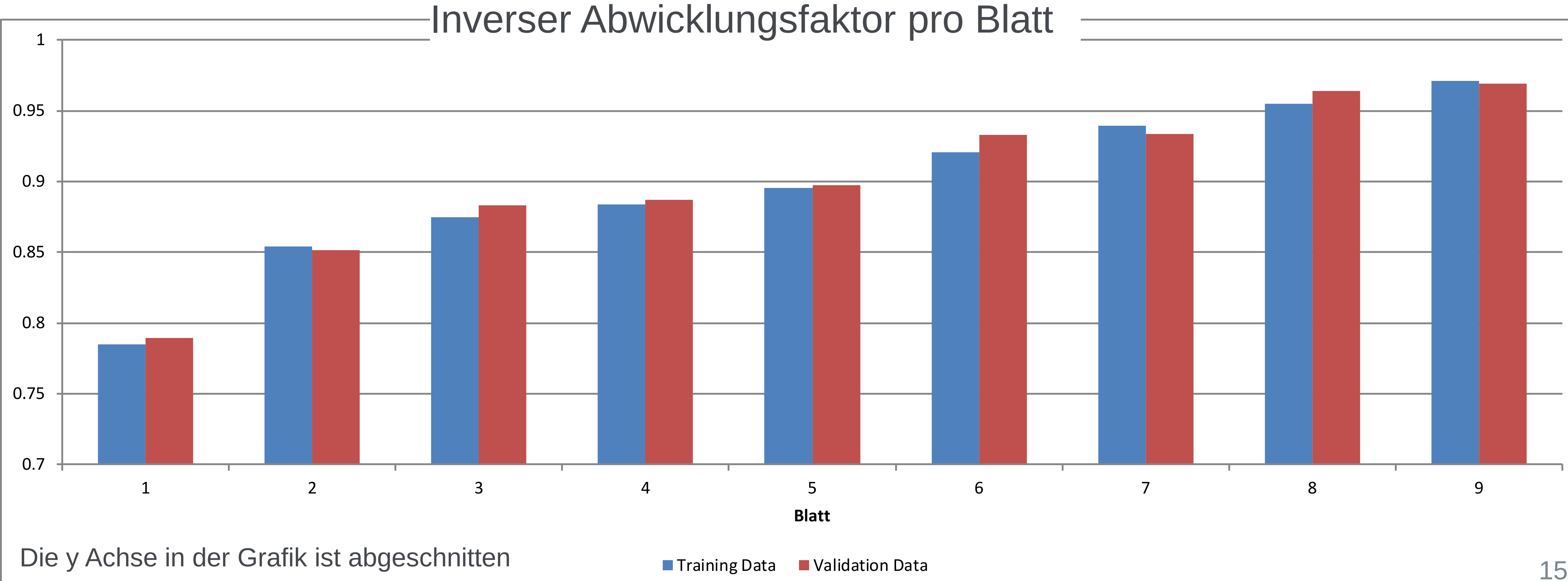
# Resultat für Abwicklungsfaktor 1

- Der Baum hat 11 Blätter (d.h. 11 Segmente)
- Die Mindestgrösse der Blätter wurde als 150’000 gewählt.
- Die Validierungsdaten zeigen Werte, welche nahe bei den Trainingsdaten liegen.
- Dieser Baum würde womöglich von Pruning/Zusammenfassung profitieren (z.B. Blätter 5-6 unterscheiden sich kaum)



# Resultat für Abwicklungsfaktor 2

- Der Baum hat 9 Blätter
- Die Mindestgrösse der Blätter wurde als 200’000 gewählt.




Die y Achse in der Grafik ist abgeschnitten

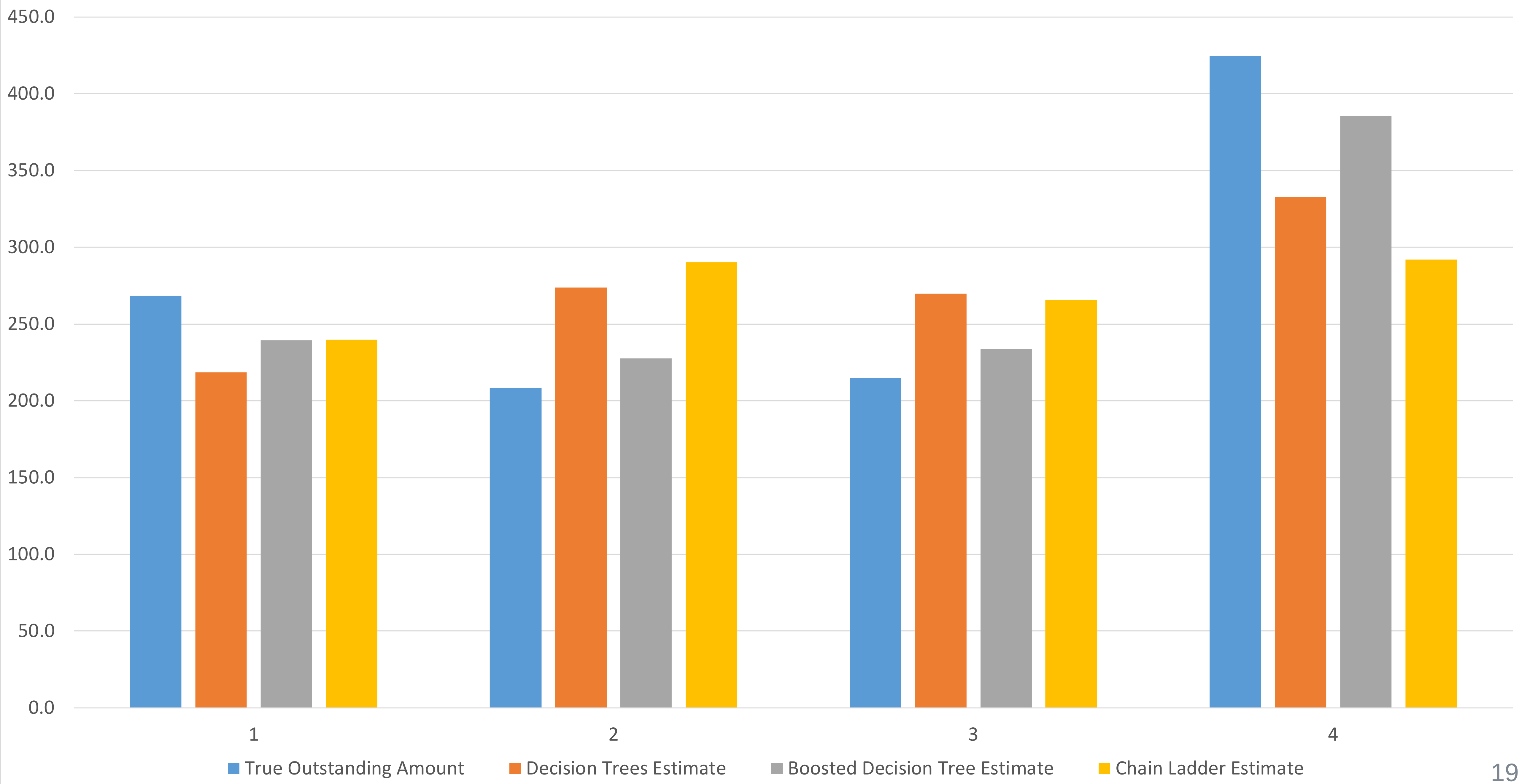
# Vergleich der Endschadenschätzung

| LoB | True Outstanding Amount | Decision Trees Estimate | Boosted Decision Tree Estimate | Chain Ladder Estimate | Decision Trees Error | Boosted Decision Tree Error | Chain Ladder Error |
|---|---|---|---|---|---|---|---|
| 1 | 268.3 | 218.5 | 239.5 | 239.8 | -49.8 | -28.8 | -28.5 |
| 2 | 208.4 | 273.6 | 227.6 | 290.4 | 65.3 | 19.2 | 82.1 |
| 3 | 214.8 | 269.8 | 233.8 | 265.8 | 55.0 | 19.0 | 51.0 |
| 4 | 424.7 | 332.7 | 385.8 | 291.9 | -92.1 | -39.0 | -132.9 |
| **Total** | **1'116.2** | **1'094.6** | **1'086.7** | **1'087.9** | **-21.6** | **-29.6** | **-28.3** |
| **In Prozent** | | | | | -1.9% | -2.6% | -2.5% |

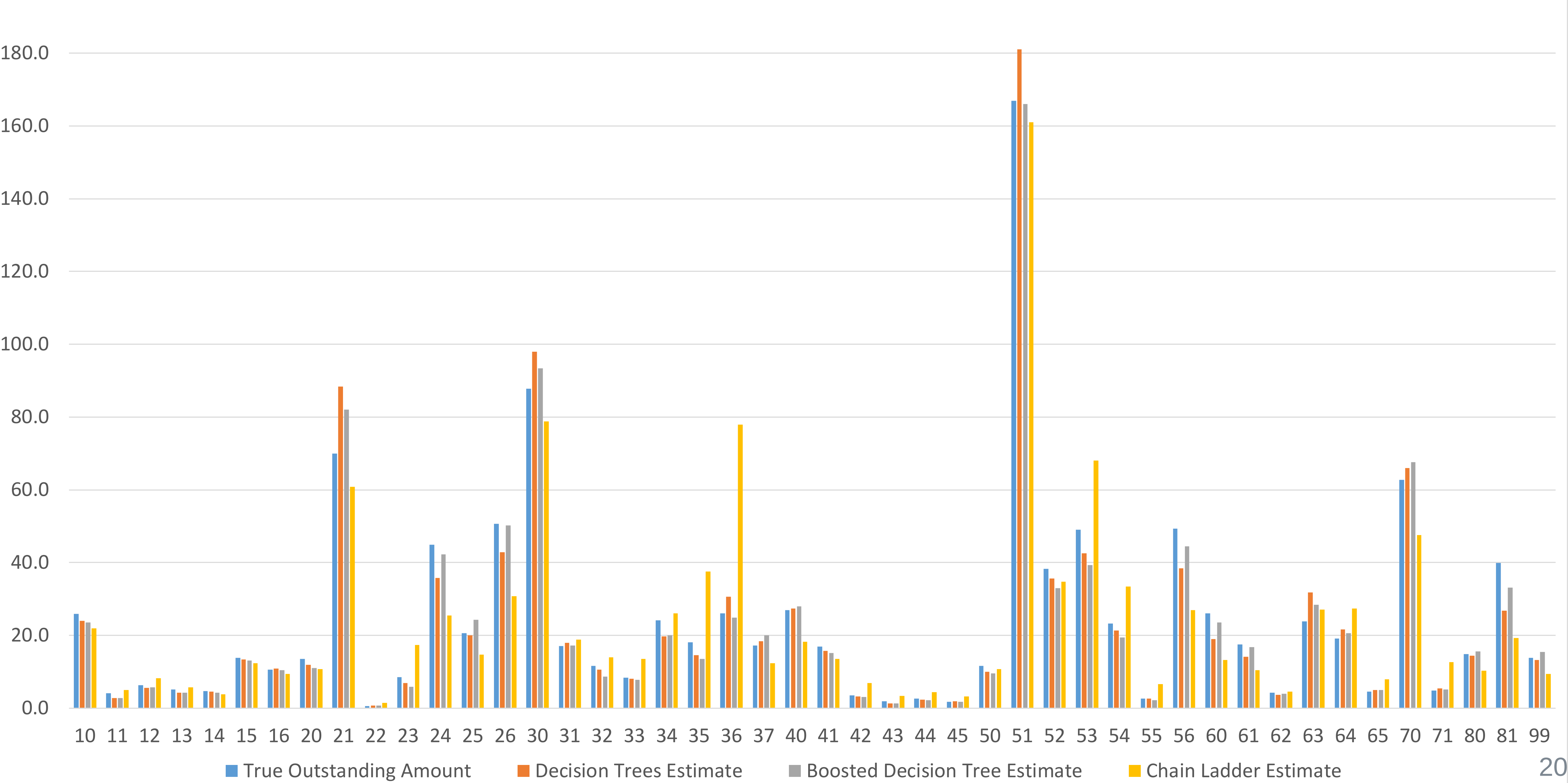
Zusätzlich zu den einzelnen Entscheidungsbäumen wurde ein Boosting Modell erstellt. Ein Boosting ist eine Kombination (ensemble) von mehreren Entscheidungsbäumen. Somit ist das Modelle komplexer und hat häufig eine höhere Vorhersagekraft als ein einzelner Entscheidungsbaum

Gesamthaft gesehen, ist die Schätzgenauigkeit der drei Modelle vergleichbar, jedoch sind die Entscheidungsbäume (und das Boosting) genauer, wenn es darum geht den Schadenaufwand für einzelne Subportfolien zu schätzen (siehe nächste zwei Folien).

# Estimates - LoB



# Estimates - injured part (inj_part)



# Mehrwert von Individual Claims Reserving

Der Mehrwert von individual claims reserving ist bei kleineren, inhomogeneren und weniger robusteren/stabileren Portfolien womöglich grösser als bei diesem Portfolio

**Paid Loss Development**

| Accident Year | 12-24 | 24-36 | 36-48 | 48-60 | 60-72 | 72-84 | 84-96 | 96-108 | 108-120 | 120-132 | 132-144 | 144-Ult |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12-1994 | 1.535 | 1.122 | 1.057 | 1.033 | 1.022 | 1.016 | 1.012 | 1.010 | 1.008 | 1.007 | 1.004 | |
| 12-1995 | 1.562 | 1.131 | 1.058 | 1.034 | 1.023 | 1.017 | 1.013 | 1.010 | 1.008 | 1.006 | | |
| 12-1996 | 1.569 | 1.132 | 1.061 | 1.035 | 1.023 | 1.017 | 1.013 | 1.010 | 1.009 | | | |
| 12-1997 | 1.566 | 1.131 | 1.062 | 1.036 | 1.024 | 1.017 | 1.013 | 1.010 | | | | |
| 12-1998 | 1.591 | 1.141 | 1.064 | 1.037 | 1.024 | 1.017 | 1.013 | | | | | |
| 12-1999 | 1.594 | 1.143 | 1.064 | 1.037 | 1.024 | 1.017 | | | | | | |
| 12-2000 | 1.609 | 1.148 | 1.065 | 1.037 | 1.023 | | | | | | | |
| 12-2001 | 1.624 | 1.150 | 1.065 | 1.036 | | | | | | | | |
| 12-2002 | 1.621 | 1.152 | 1.066 | | | | | | | | | |
| 12-2003 | 1.616 | 1.150 | | | | | | | | | | |
| 12-2004 | 1.627 | | | | | | | | | | | |
| 12-2005 | | | | | | | | | | | | |
| Vol Wtd Avg | 1.593 | 1.140 | 1.062 | 1.036 | 1.023 | 1.017 | 1.013 | 1.010 | 1.009 | 1.007 | 1.004 | |
| Vol Wtd Avg Exc Hi/Lo | 1.595 | 1.141 | 1.063 | 1.036 | 1.023 | 1.017 | 1.013 | 1.010 | 1.008 | | | |
| Default | 1.593 | 1.140 | 1.062 | 1.036 | 1.023 | 1.017 | 1.013 | 1.010 | 1.009 | 1.007 | 1.004 | |
| Manual Selected | | | | | | | | | | | | 1.000 |
| Selected | 1.593 | 1.140 | 1.062 | 1.036 | 1.023 | 1.017 | 1.013 | 1.010 | 1.009 | 1.007 | 1.004 | 1.000 |
| Cumulative | 2.169 | 1.362 | 1.194 | 1.124 | 1.085 | 1.061 | 1.043 | 1.030 | 1.020 | 1.011 | 1.004 | 1.000 |
| Ratio to Ultimate | 0.461 | 0.734 | 0.837 | 0.890 | 0.921 | 0.943 | 0.959 | 0.971 | 0.981 | 0.989 | 0.996 | 1.000 |

# Best Practice und Verfeinerung der Modelle

Wie bei jedem Modell ist es essentiell:

- Dass die Daten so fehlerfrei und konsistent wie möglich sind
- Die Daten genau zu verstehen

Die gezeigten Modelle können an verschiedenen Orten verbessert werden:

- Alternatives Splitting Kriterium: maximiere die absolute Differenz zw. Links/Rechts (liefert vergleichbare Resultate)
- Fine Tuning der Parameter
- Data Pre-Processing: Eventuell gewisse Werte zusammenfassen (inj_part, cc)
- Pruning der Trees
- Cross Validation
- Ensembles: Boosting, Bagging, Random Forest
- Die gezeigten Modell verwenden nur 70% der Daten. Man könnte die Schätzer auf 100% der Daten fitten (bei einem fixierten Modell)
- Noch nicht gemeldete Schäden könnten eventuell gesondert behandelt werden (bei Modellen die auf Einzelschadendaten basieren)

# Alternative Ansätze

- Wir haben hier für jeden LDF eine neue Segmentierung erstellt. Die Modelle sind dabei unabhängig.
- Alternativ könnte man Segmente definieren, welche gleich sind für alle LDFs (klassische Reservierungssegmente). Diese könnten dann in verschiedenen aggregierten Reservierungsmethoden (CL, BF,…) verwendet werden
- Man könnte die Daten erst aufteilen in 4 LoBs und dann 4 verschiedene Modelle erstellen.

# Parallele zum Loss Ratio Modelling

Der modellierte CL Faktor war ein Quotient zweier positiver Grössen.

Eine Loss Ratio $LR = \frac{\sum_i Schaden_i}{\sum_i Prämie_i}$ ist konzeptionell dasselbe

Die hier verwendeten Entscheidungsbäume eignen sich deshalb äusserst gut um die Loss Ratio von einem Portfolio zu modellieren.

Damit lassen sich profitable und unprofitable Kundensegmente identifizieren, was Rückschlüsse auf Tarifierung, Underwriting, Verkauf, Marketing und Strategie erlaubt.

# Fachgruppe Data Science

www.actuarialdatascience.org

- Peter Blum
- Andrea Ferrario
- Frank Genheimer
- Roger Hämmerli
- Thomas Hull
- Bernhard König
- David Lüthi
- Alexander Noll
- Robert Salzmann
- Jürg Schelldorfer
- Frank Weber
- Mario Wüthrich

# Thank you

Bernhard  König, Aktuar SAV
+41 79 706 01 26
bernhard.koenig@miliman.com

# Thank you

Bernhard  König, Aktuar SAV
+41 79 706 01 26
bernhard.koenig@miliman.com