Segmentation of Insurance Data with Decision Trees

Bernhard König*

Prepared for:
Fachgruppe "Data Science"
Swiss Association of Actuaries SAV

August 21, 2019

Abstract

We provide a tutorial to illustrate the usage of decision trees for the segmentation of insurance data. We consider a claims data set, which we partition into segments according to the loss development factors of the individual claims.

1 Introduction and Motivation

This data analytics tutorial has been written for the working group "Data Science" of the Swiss Association of Actuaries SAV, see

https://www.actuarialdatascience.org

The main purpose of the tutorial is to illustrate the usage of decision tree models for portfolio segmentation. In general insurance there are various examples which involve the segmentation of a portfolio (or more generally a data set), including but not limited to

- Pricing: When developing burning cost models (i.e. the expected loss for a given risk), we typically define multiple segments (tariff cells) which differ in their expected loss and thus receive a different premium.
- Claim frequency modelling: When pricing a portfolio, a frequent choice is to model the claim frequency separately from the claim severity. In that case, we are modelling the number of claims per exposure.
- Profitability analysis: For an existing book of business, it is a common exercise (e.g. for renewal pricing) to determine the profitability of different customer segments. Given the historical losses and earned premiums, one can divide the portfolio in different segments which have a varying profitability margin. This corresponds to a loss ratio model.
- Loss Reserving: Many reserving approaches, such as Chain-Ladder or Bornhuetter-Ferguson, work on aggregated loss development triangles. By dividing the data into homogeneous segments, we can generally derive more accurate estimates of the outstanding loss amount. This can be particularly

While the above problem settings can be solved with various different approaches, for this paper, we will limit our work to regression decision trees. One common property of the four mentioned problem settings, is the fact that the target variable is the ratio of two real valued variables (both of which are generally non-negative):

- Pricing: the target variable (burning cost) is the loss amount divided by the exposure
- Claim frequency modelling: the target variable (claim frequency) is the number of claims divided by the exposure
- Profitability analysis: the target variable (loss ratio) is the loss amount divided by the earned premium
- Loss Reserving: the development of a loss triangle considers the paid (or incurred) amount at time $t+1$ divided by the paid (or incurred) amount at time t

For this paper we consider the problem setting of loss reserving. However, we emphasize that our segmentation approach would work similarly for the other problem settings mentioned above. The primary reason why we consider a loss reserving example, is the fact that we have a large and rich claims data set available as described in the next Section.

*Milliman, bernhard.koenig@milliman.com

Values in CHF Mio - Evaluated as of Dezember 31, 2005 Paid Loss - Cumulative												
AY	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											

Table 1: *Cumulative Paid Triangle*

2 Data

On goal of this tutorial was to allow readers to reproduce the results of this work, such that it could serve as a hands-on learning tutorial. We have chosen simulated data by using the 'Individual Claims History Simulation Machine' which is described in detail on this website: https://people.math.ethz.ch/ wmario/simulation.html
The simulated data is based upon a large data set of Swiss occupational accident insurance claims. For the details of the simulated data, we refer to [?]. In summary, we have the following variables:

- `ClNr` is the claim number, this is a unique numerical identifier for each individual claim.
- `LoB` is the line of business, which is a categorical variable with labels 1, 2, 3, 4. We have no information on the meaning of the four labels.
- `cc` is the claims code, which is a categorical variable with labels in $1, ..., 53$ (with possible gaps).
- `AY` is the year of claims occurrence (accident year). It is integer valued with values between 1994 and 2005.
- `AQ` is the integer valued quarter of claims occurrence
- `age` is the integer valued age of the injured with values between 15 and 70.
- `inj_part` is the injured body part. This variable is categorical with labels in $1, ..., 99$ (with gaps). We have no information on the specific meaning of the labels.
- `RepDel` is the reporting delay in years with values between 0 and 11. A value of 0 indicates that the claim was reported during the accident year.

While the data contains the accident quarter for each individual claim, the development of the losses is only available by year. The claims data has 12 development years, meaning that for each claim, we have the paid amount for the accident year of that claim and the 11 years following that accident year. In Figure ?? we show the distribution of the number of claims for the most important variables in the data. We note that Wüthrich has applied a neural network to this data (see [?]). In this paper we consider a similar approach which is based on decision trees instead of neural networks.

3 The Mack chain-ladder Model

As a basis for our analysis we will consider the Mack chain-ladder (CL) model ([?]). Similar as other triangle based reserving approaches, Mack' chain-ladder uses claims data which is aggregated into a triangular form. Let $C_{i,j}$ be the cumulative payment of the accident year (`AY`) i and the development year j where $1994 \leq i \leq 2005$ and $0 \leq j \leq 11$. We assume that $C_{i,J}$ corresponds to the ultimate loss for accident year i, i.e. that there is no further development in the tail of the triangle. If we evaluate the data as of December 31, 2005 we have the triangle shown in Table ??. As of year-end 2005, the values of $C_{i,j}$ for $i + j > 2005$ are unknown.
The CL model assumes that there are non-negative CL factors f_j for $0 \leq j \leq J-1$ such that we can estimate the future cumulative payments as

$$\widehat{C}_{i,j} = \widehat{f}_{j-1} C_{i,j-1}, \qquad \text{for } j = 1, ..., J. \tag{1}$$

The CL estimators are defined by

$$\widehat{f}_{j-1} = \frac{\sum_{i=1}^{I-j} C_{i,j}}{\sum_{i=1}^{I-j} C_{i,j-1}}, \qquad \text{for } j = 1, ..., J. \tag{2}$$

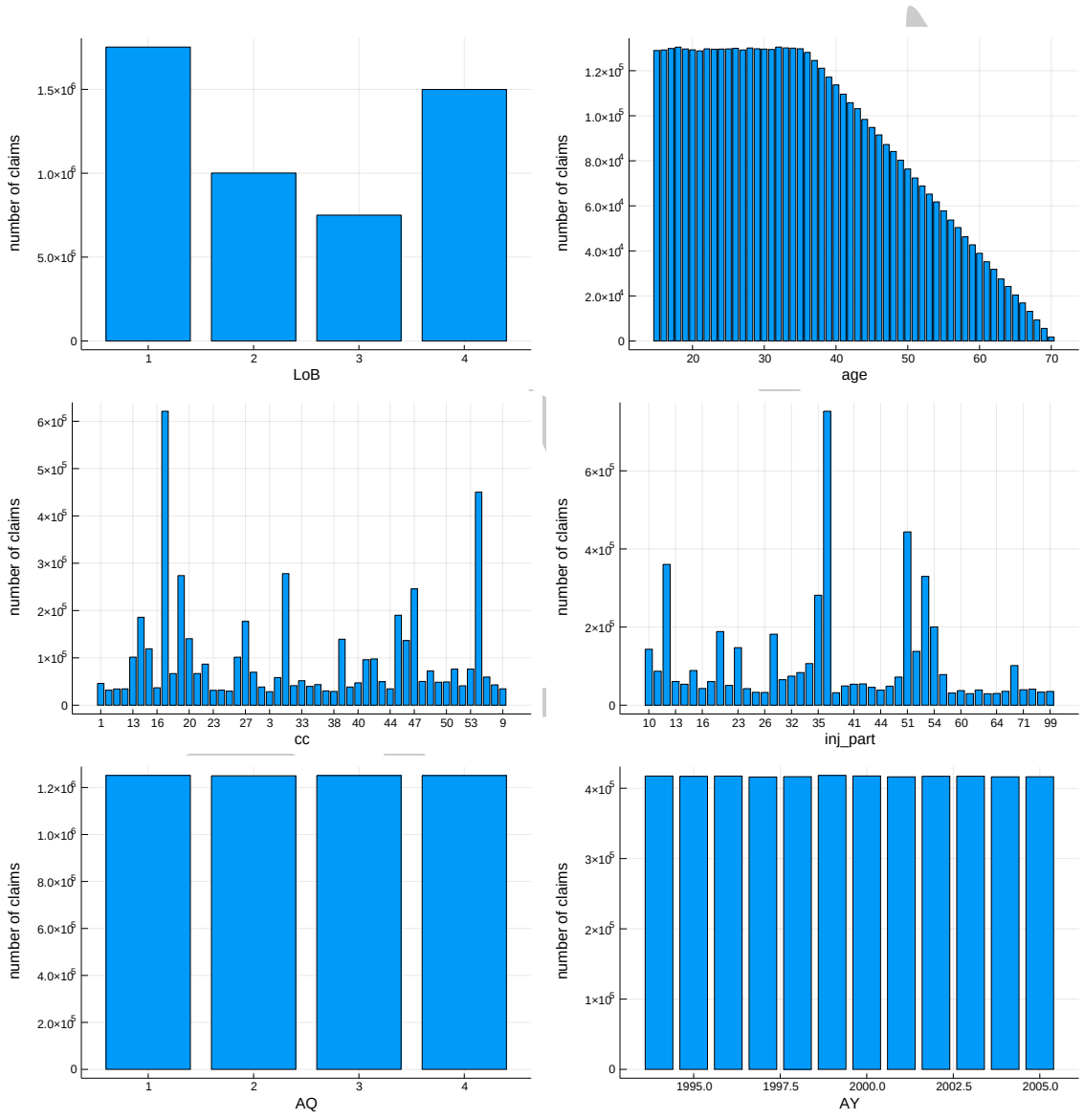

Figure 1: *Distribution of the number of claims for the main variables*

Evaluated as of Dezember 31, 2005 Paid Loss Development												
AY	0	1	2	3	4	5	6	7	8	9	10	11
1994	1.535	1.122	1.057	1.033	1.022	1.016	1.012	1.010	1.008	1.007	1.004	1.000
1995	1.562	1.131	1.058	1.034	1.023	1.017	1.013	1.010	1.008	1.006		
1996	1.569	1.132	1.061	1.035	1.023	1.017	1.013	1.010	1.009			
1997	1.566	1.131	1.062	1.036	1.024	1.017	1.013	1.010				
1998	1.591	1.141	1.064	1.037	1.024	1.017	1.013					
1999	1.594	1.143	1.064	1.037	1.024	1.017						
2000	1.609	1.148	1.065	1.037	1.023							
2001	1.624	1.150	1.065	1.036								
2002	1.621	1.152	1.066									
2003	1.616	1.150										
2004	1.627											
2005												
Selected CL factor	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

Table 2: *Chain-Ladder factors*

Values in CHF Mio												
AY	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	704.0
1996	335.6	526.5	596.0	632.4	654.7	669.9	681.2	690.2	696.9	703.3	708.0	711.0
1997	326.7	511.5	578.5	614.1	636.4	651.9	662.8	671.3	677.9	683.7	688.3	691.2
1998	324.4	516.3	589.0	626.5	649.5	665.0	676.5	685.4	692.1	698.1	702.8	705.7
1999	330.9	527.4	602.8	641.6	665.1	681.1	692.7	701.6	708.5	714.6	719.4	722.4
2000	332.1	534.3	613.3	652.9	676.8	692.7	704.4	713.5	720.5	726.7	731.6	734.7
2001	333.5	541.7	623.0	663.6	687.7	703.8	715.7	724.8	732.0	738.3	743.2	746.4
2002	349.7	567.0	653.4	696.3	721.1	738.0	750.5	760.1	767.6	774.2	779.4	782.7
2003	371.2	599.9	690.0	733.1	759.2	777.0	790.1	800.3	808.2	815.1	820.5	824.0
2004	381.6	620.9	708.0	752.2	779.0	797.2	810.7	821.1	829.2	836.3	841.9	845.5
2005	400.6	638.1	727.6	773.0	800.5	819.3	833.1	843.8	852.1	859.4	865.2	868.9

Table 3: *Chain-Ladder Estimates - The estimated ultimate loss is shown in the rightmost column*

In Table ?? we show the CL factors for our data, as well as the corresponding cumulative development factors and the ratio to ultimate. Table ?? eventually shows the corresponding estimated ultimate loss in the rightmost column. As we are working with simulated data, we can actually compare the CL estimates $\widehat{C}_{i,J}$ against the true future payments $C_{i,J}$, for $i = 1994, ..., 2004$ as shown in Table ??. We can see that the overall reserves are underestimated by 28.3 million. One reason might be the trend in the CL factors, as we observe higher factors for more recent accident years (especially for the development years 0 to 3) as shown in Table ??. We also use the term loss development factor to refer to the CL factors.

4 Decision Tree Approach

4.1. Weighted Squared Error as Splitting Criterion

As noted above, the aggregated claim triangle (Table ??) was constructed from individual claim payments. Let P_j, k denote the cumulative payments (at development year $1 \leq j \leq J$) of claim k for $k = 1, ..., N$ where N is the number of claims in the data set. By construction we have

$$\sum_{i=1}^{I} C_{i,j} = \sum_{k=1}^{N} P_{j,k} \qquad \text{for } j = 1, ..., J \tag{3}$$

In the following we will consider a decision tree for each development year $1 \leq j \leq J$. Let $\mathcal{X}$ be the feature space spanned by the explanatory variables. Our proposed tree algorithm considers all possible binary splits in $\mathcal{X}$. Each split will divide a node into a left and right child node, which we denote by L and R respectively. Let $P_{j,k} = P_{j,k}(x)$ for $x \in \mathcal{X}$.

Values in CHF Mio							
AY	Cumulative Paid Loss (1)	CL Factors (2)	Cumulative CL Factors (3)	Ultimate Loss (2) x (3) (4)	Outstanding Loss Reserves (4) - (1) (5)	True Reserves (6)	Error (5) - (6) (7)
1994	697.1	1.000	1.000	697.1	0.0	0.0	0.0
1995	701.0	1.004	1.004	704.0	3.0	3.2	-0.3
1996	703.3	1.007	1.011	711.0	7.7	8.3	-0.6
1997	677.9	1.009	1.020	691.2	13.3	13.6	-0.3
1998	685.4	1.010	1.030	705.7	20.4	20.6	-0.3
1999	692.7	1.013	1.043	722.4	29.7	30.2	-0.5
2000	692.7	1.017	1.061	734.7	41.9	42.1	-0.1
2001	687.7	1.023	1.085	746.4	58.7	56.6	2.1
2002	696.3	1.036	1.124	782.7	86.3	82.0	4.3
2003	690.0	1.062	1.194	824.0	134.0	139.3	-5.4
2004	620.9	1.140	1.362	845.5	224.6	228.4	-3.8
2005	400.6	1.593	2.169	868.9	468.3	491.7	-23.4
Total	7'945.5			9'033.5	1'087.9	1'116.2	-28.3

Table 4: *Chain-Ladder Estimates compared to true outstanding payments*

Consider the following splitting criterion

$$wSSE_l = \sum_{k \in L} P_{j-1,k} \left[\frac{P_{j-1,k}}{P_{j,k}} - \frac{1}{\widehat{f}_{j-1,l}} \right]^2, \tag{4}$$

where $\widehat{f}_{j-1,l}$ is the CL factor evaluated on the data in the left child node only, i.e.

$$\widehat{f}_{j-1,l} = \frac{\sum_{k \in L} P_{j,k}}{\sum_{k \in L} P_{j-1,k}}, \qquad \text{for } j = 1, ..., J \tag{5}$$

Equation ?? defines a weighted sum of squared errors. If we minimize this amount in each split, we obtain homogeneous nodes. We define $\widehat{f}_{j-1,r}$ and $wSSE_r$ analogously for the right child node. As, we are considering individual claims in Equation ??, the values of $P_{j-1,k}$ could be zero. To avoid division by zero, we are excluding all claims, which are zero at time j. In most cases, but not all, this implies that the cumulative payment was also zero at time $j-1$. We have chosen to consider the inverted CL factor in Equation ?? as the restriction for the payments to be strictly positive, generally excludes fewer claims when applied to the development year j instead of the development year $j-1$. We further note that claims, for which $P_{j,k}$ and $P_{j-1,k}$ are both zero, have no impact on the value of $\widehat{f}_{j-1,l}$ and $\widehat{f}_{j-1,r}$.
For any split, let

$$wSSE = wSSE_l + wSSE_r \tag{6}$$

be the sum of the error term in the left and right child node. We define the splitting criterion for the decision tree by minimizing this amount over all possible splits.

4.2. Stopping Criterion

We note that we do not consider any information gain, i.e. the value of the error in the parent node is not relevant for the choice of the best split. Our current approach considers a predefined absolute value as exposure threshold $T > 0$ to stop the growing of the tree. That is to say, the tree will never perform a split that results in a leaf with less than T exposure. Exposure is simply a weight assigned to each row in the data. A canonical choice is a constant exposure (of one) for each row. Another typical exposure measure in general insurance, specifically in pricing, is the coverage period measured in years (e.g. earned vehicle years).

4.3. Estimator

For a given tree we will consider the canonical estimator being the mean observed value for each leaf.

4.4. The absolute difference as an alternative splitting criterion

Let S be the set of all possible binary splits. Then, we can defined the difference splitting criterion as the split s, which maximizes the difference between the development factor in the left and right child as

$$s = \arg\max_{s \in S} \left| \frac{\sum_{k \in L_s} P_{j-1,k}}{\sum_{k \in L_s} P_{j,k}} - \frac{\sum_{k \in R_s} P_{j-1,k}}{\sum_{k \in R_s} P_{j,k}} \right|, \tag{7}$$

where L_s and R_s denote the left and right child node for the split s. We note that controlling the size of the child nodes is critical for this splitting criterion. By definition this splitting criterion will maximize the spread between different nodes (and leaves) of the tree. Even though we do not consider any measure of noise (such as the variance of the observations), this approach still leads to homogeneous leaves.
We note that the difference splitting criterion is very attractive from a computational point of view. Consider a numerical variable for which we want to find the optimal splitting threshold. Once the data is sorted by this explanatory variable, one can evaluate all splits by subsequently shifting one element ($P_j - k$ and $P_j - k$ respectively) from L_s to R_s. This is done by updating the sum in the numerator and denominator to account for the current element.

4.5. A multiplicative boosting approach

In the previous sections we have described a decision tree method to model the development of loss payments over time. More precisely we modelled the ratio $\sum_k P_{j,k} / \sum_k P_{j-1,k}$. Using a more general notation, let $q(x) = n(x)/d(x)$ be the ratio of a numerator $n(x)$ and a denominator $d(x)$ for a vector of predictor variables x. The boosting model derives an estimator $\hat{q}_M(x)$ for $q(x)$ where M is the number of iterations (i.e. the number of individual trees composing the boosting model). We note that many boosting models, such as Algorithm 10.2 in [?] are additive in the sense that they correspond to a sum of individual trees. For the modelling of ratios, we propose a different approach. Our boosting algorithm considers multiplicative residuals instead of additive residuals.

Multiplicative Boosting

1. Initialize $\hat{f}_0(x)$ as the average ratio of the whole data set. Compute $\hat{n}_0(x) = d(x)\hat{f}_0(x)$

2. For $m = 1..M$:
   (a) Let $r_m(x) = \frac{f(x)}{\hat{f}_{i-1}(x)} = \frac{n(x)}{\hat{n}_{i-1}(x)}$

   (b) Fit a decision tree (weak learner) to the ratio $r_m(x) = \frac{n(x)}{\hat{n}_{i-1}(x)}$
   Let $\hat{r}_m(x)$ be the estimated ratio of the tree

   (c) Set $\hat{f}_m(x) = \hat{f}_{m-1}(x) moderate(\hat{r}_m(x), \lambda)$
   where λ is a user specified learning rate and

$$moderate(r, \lambda) = \begin{cases} (r-1)\lambda + 1, & \text{if } r \geq 1 \\ (1-r)\lambda + 1, & \text{otherwise} \end{cases}$$

Similarly as other ensemble methods, we introduce randomization into the boosting algorithm by either sampling the predictor variables which are available to each individual tree or by bootstrapping (i.e. sampling the data).

4.5.1 Negative observed ratios

We note that the multiplicative boosting algorithm is not suitable for data, which contains negative ratios. This is because the concept of the moderation function is no meaningful for $r < 0$. For insurance data we typically do not observe loss development factors which are negative which is why this is generally not a limitation. Similarly loss ratios and claim frequency are generally non-negative.
Notably, the boosting algorithm will work fine, if only a few selected observations in the data have negative ratios. More precisely, the algorithm will work as intended if it will never find a node where the observed ratio is negative.
For data where negative observed ratios are common, appropriate pre-processing of the data might allow the algorithm to be applicable nonetheless.

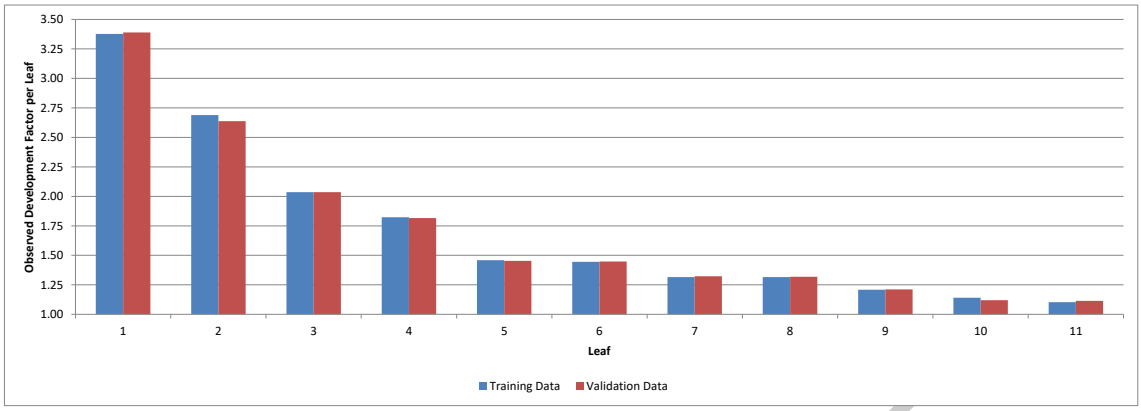

Figure 2: *Loss development factors for the leaves of the tree for development year 1*

5 Results

5.1. Decision Tree Approach

To validate the decision tree, we split the data randomly into a training (70% of the data) and validation set (30% of the data). Let us consider the first development year $j = 1$. From Table ?? we know that the overall CL factor is $1.593 = 1/0.628$. The resulting decision tree for is shown in Figure ??. The top node represents the root node with all the data. The 'Fit' value is the inverse of the overall CL factor for development year 1. The weight indicates that there are 2.546 million claims in the training data. Claims where both $P_{1,k}$ and $P_{2,k}$ are zero and claims with $P_{2,k} = 0$ where discarded. The condition for each split is shown between the edges. We note that the condition applies to the left child node. For the first split, all claims with accident quarter (`AQ`) go left. We can see that claims with $\texttt{AQ} = 4$ have a loss development factor of $1/0.349 = 2.86$ compared to $1/0.71 = 1.41$ for claims with $\texttt{AQ} \leq 4$. This is not too surprising, as claims which are notified late in the year have little time left for payments to be made in the same year. The nodes are further split and the whole tree has 11 leaves. Note that the leaves are sorted by fitted value (i.e. the inverse of the loss development factor).

Figure ?? shows the loss development factor for each leaf. The graph shows that the validation data behaves very similarly as the training data. This is the main validation criterion for the model. The stopping criterion for the growth of the tree was set to a minimum exposure per leaf of 0.15 million claims (which is roughly 6% of the data). If we would have chosen considerably smaller threshold, we would have overfitted the data. Figure ?? shows that there is considerably more differentiation for the leaves $1 - 4$ compared to the leaves $8 - 10$ where the leaves have development factors, which are closer together. The maximal value is 3.38 and the minimum is 1.10 for the training data. This is a considerable spread, bearing in mind that the overall CL factor for the training data is at 1.593. It shows that there are sizeable sub-segments which develop differently. We note that the ratio between the maximal and minimal observed value $3.06 = 3.38/1.10$ (for the training data) is also known as *lift*. Thus the lift is 3.06 for the training data and 3.04 for the validation data.

We have eventually constructed a decision tree for each individual development year. As we are considering a claims triangle, the number of claims are fewer for later accident years. Because of that, the threshold to stop growing the tree needs to be adjusted to account for the number of claims available and the signal strength. In Table ?? we show the number claims per development year and the threshold to stop growing the tree in the second and third column. The remaining columns show the number of leaves as well as the training and validation lift for the resulting decision trees. As the CL factors are converging to one (see ??), there is also less signal available for later development years. Therefore the lift is decreasing, i.e. we have little differentiation between the different leaves for later development years. Note that the last tree consists of the root node and no splits were performed due to lack of signal. In Figure ?? we show the CL Factors of each leaf for the development year 3. We can see that there is still some differentiation between the leaves and that the model validates reasonably. In contrast, for year 4, as shown in Figure ??, we can see that the individual leaves have a relatively similar development factor, especially for the leaves 1 to 3, where the model does not validate well. In fact, the tree could possibly be improved by pruning of some of the leaves, or by combining the leaves 2 and 3.

5.2. Combining the trees to predict an ultimate loss

In the previous sections, we have shown how we have constructed a decision tree for each development year. Each tree uses individual claims data, specifically the variables `age`, `AY`, `cc` and `inj_part`. The 11 trees have been constructed independently of each other. Each tree predicts a development factor for any given claim (for the respective development year). By multiplying these predicted development factors, we receive an age-to-ultimate factor for any given claim and development year. This age-to-ultimate factor is comparable to the second last row (titled Cumulative) in Figure ??.

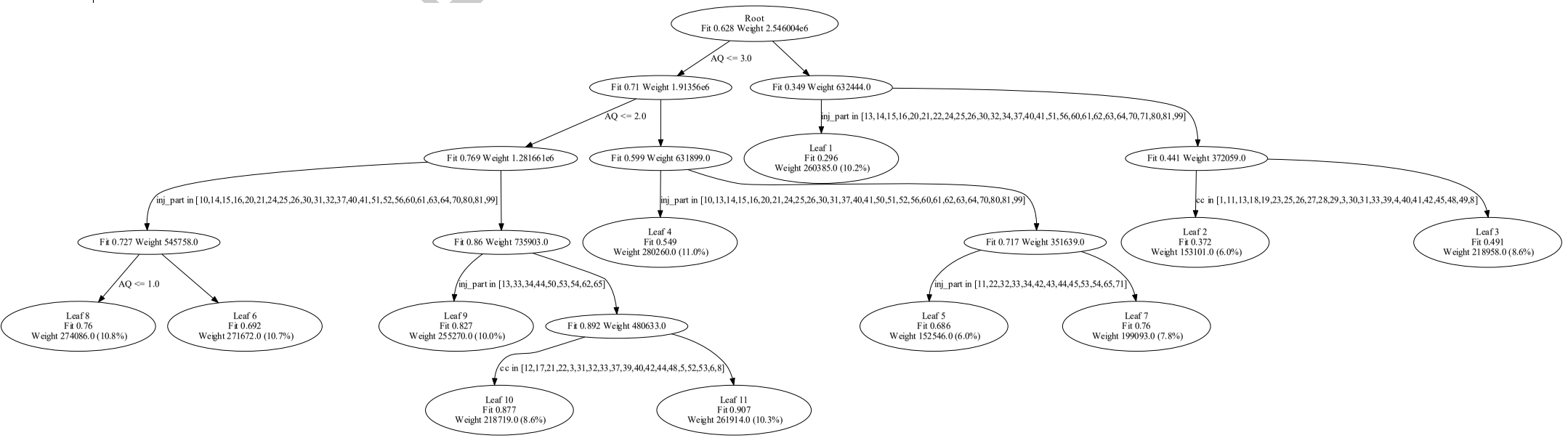

Figure 3: *Tree for development year 1. The split condition is true for the left child node.*

Development Year	Number of Claims in milllion (Training)	Threshold per leaf in million	Threshold per leaf in percentage of training data	Number of leaves	Lift Training	Lift Validation
1	2.54	0.15	6%	11	3.06	3.04
2	2.33	0.20	9%	9	1.24	1.23
3	2.10	0.27	13%	7	1.07	1.07
4	1.87	0.25	13%	6	1.03	1.03
5	1.64	0.24	14%	6	1.04	1.04
6	1.40	0.21	15%	6	1.02	1.02
7	1.17	0.20	17%	4	1.02	1.02
8	0.94	0.24	26%	3	1.01	1.01
9	0.70	0.18	26%	4	1.01	1.01
10	0.47	0.13	28%	3	1.01	1.01
11	0.23	0.13	57%	1	1.00	1.00

Table 5: *Size of the training data and threshold (stopping criterion) for tree growth by development year. The three rightmost columns show the number of leaves and the lift of the resulting trees*

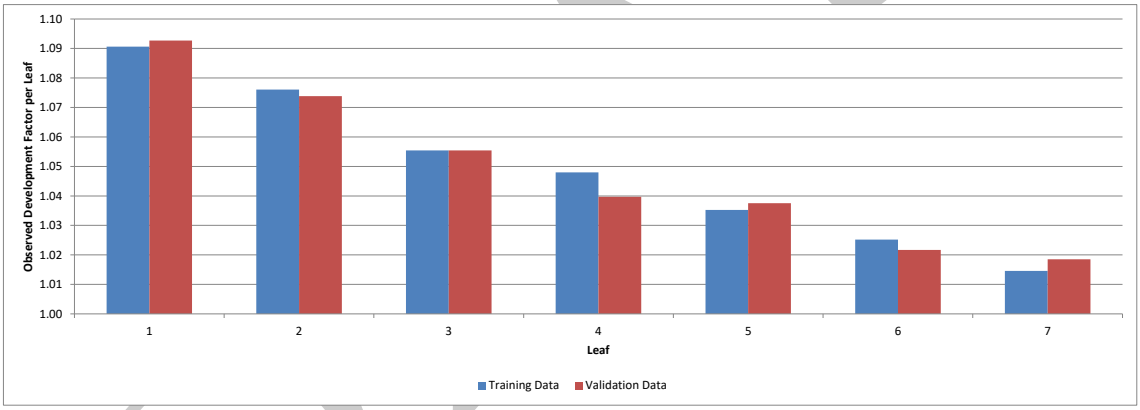

Figure 4: *Loss development factors for the leaves of the tree for development year 3*

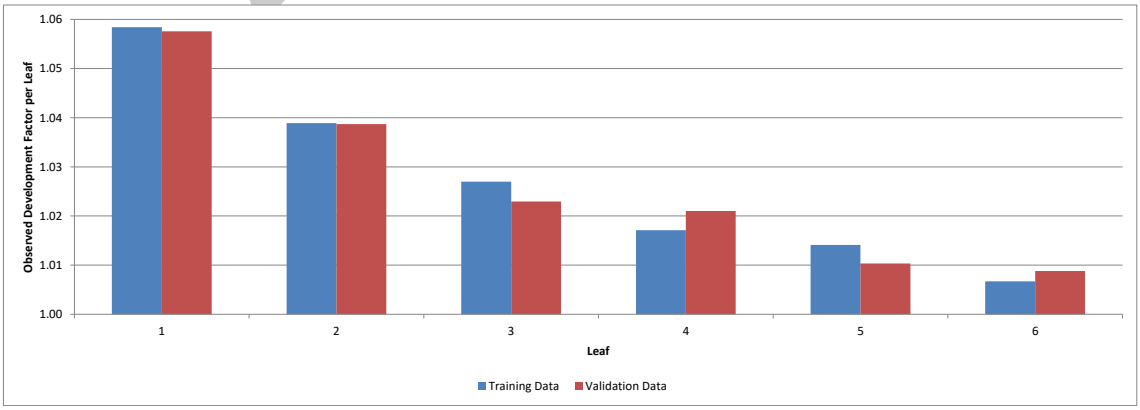

Figure 5: *Loss development factors for the leaves of the tree for development year 4*

Figure 6: *Loss development factors for the leaves of the tree for development year 4*

LOB1
Evaluated as of Dezember 31, 2005
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
1994	1.505	1.113	1.058	1.036	1.027	1.020	1.015	1.012	1.010	1.009	1.007	
1995	1.528	1.121	1.061	1.039	1.027	1.020	1.015	1.012	1.010	1.009		
1996	1.536	1.120	1.061	1.038	1.027	1.020	1.016	1.012	1.010			
1997	1.539	1.125	1.065	1.040	1.027	1.020	1.015	1.012				
1998	1.536	1.127	1.065	1.041	1.028	1.021	1.016					
1999	1.559	1.133	1.067	1.041	1.028	1.021						
2000	1.567	1.136	1.068	1.042	1.029							
2001	1.591	1.143	1.071	1.044								
2002	1.577	1.138	1.068									
2003	1.575	1.137										
2004	1.618											
2005												
Vol Wtd Avg	1.559	1.130	1.065	1.040	1.028	1.020	1.016	1.012	1.010	1.009	1.007	
Cumulative	2.158	1.384	1.225	1.150	1.106	1.076	1.055	1.039	1.026	1.016	1.007	1.000
Ratio to Ultimate	0.463	0.722	0.816	0.869	0.904	0.929	0.948	0.963	0.974	0.984	0.993	1.000

LOB2
Evaluated as of Dezember 31, 2005
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	1.454	1.104	1.043	1.021	1.013	1.009	1.005	1.004	1.004	1.003	0.998	
12-1995	1.478	1.121	1.047	1.024	1.015	1.012	1.008	1.007	1.004	1.002		
12-1996	1.485	1.121	1.048	1.024	1.013	1.010	1.006	1.003	1.004			
12-1997	1.462	1.113	1.049	1.022	1.017	1.010	1.007	1.005				
12-1998	1.523	1.132	1.051	1.024	1.011	1.009	1.006					
12-1999	1.533	1.134	1.050	1.023	1.013	1.007						
12-2000	1.562	1.140	1.050	1.023	1.014							
12-2001	1.560	1.138	1.052	1.025								
12-2002	1.561	1.142	1.051									
12-2003	1.560	1.142										
12-2004	1.583											
12-2005												
Vol Wtd Avg	1.524	1.129	1.049	1.024	1.014	1.009	1.007	1.005	1.004	1.003	0.998	
Cumulative	1.922	1.261	1.117	1.065	1.040	1.026	1.017	1.010	1.005	1.001	0.998	1.000
Ratio to Ultimate	0.520	0.793	0.895	0.939	0.961	0.974	0.983	0.990	0.995	0.999	1.002	1.000

LOB3
Evaluated as of Dezember 31, 2005
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	1.586	1.127	1.056	1.031	1.019	1.013	1.010	1.009	1.006	1.006	1.003	
12-1995	1.626	1.137	1.051	1.029	1.017	1.013	1.009	1.007	1.006	1.005		
12-1996	1.617	1.131	1.056	1.031	1.020	1.012	1.011	1.007	1.008			
12-1997	1.623	1.129	1.053	1.028	1.018	1.012	1.008	1.006				
12-1998	1.637	1.140	1.057	1.030	1.019	1.014	1.009					
12-1999	1.621	1.137	1.057	1.030	1.019	1.014						
12-2000	1.619	1.142	1.056	1.029	1.015							
12-2001	1.681	1.152	1.055	1.026								
12-2002	1.661	1.150	1.055									
12-2003	1.630	1.139										
12-2004	1.618											
12-2005												
Vol Wtd Avg	1.628	1.138	1.055	1.029	1.018	1.013	1.009	1.007	1.006	1.005	1.003	
Cumulative	2.143	1.316	1.156	1.096	1.065	1.046	1.032	1.022	1.015	1.008	1.003	1.000
Ratio to Ultimate	0.467	0.760	0.865	0.913	0.939	0.956	0.969	0.978	0.985	0.992	0.997	1.000

LOB4
Evaluated as of Dezember 31, 2005
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	1.606	1.143	1.074	1.044	1.032	1.025	1.019	1.015	1.013	1.011	1.009	
12-1995	1.624	1.145	1.074	1.047	1.032	1.024	1.019	1.015	1.013	1.011		
12-1996	1.647	1.154	1.079	1.049	1.034	1.025	1.020	1.016	1.014			
12-1997	1.646	1.156	1.080	1.050	1.035	1.025	1.021	1.016				
12-1998	1.672	1.163	1.081	1.052	1.036	1.026	1.021					
12-1999	1.665	1.165	1.083	1.052	1.035	1.026						
12-2000	1.689	1.171	1.085	1.052	1.036							
12-2001	1.674	1.168	1.083	1.049								
12-2002	1.690	1.178	1.088									
12-2003	1.698	1.179										
12-2004	1.690											
12-2005												
Vol Wtd Avg	1.665	1.163	1.081	1.050	1.034	1.025	1.020	1.015	1.013	1.011	1.009	
Cumulative	2.493	1.497	1.287	1.191	1.135	1.097	1.070	1.049	1.033	1.020	1.009	1.000
Ratio to Ultimate	0.401	0.668	0.777	0.840	0.881	0.911	0.935	0.953	0.968	0.981	0.991	1.000

Figure 7: *Development factors by accident year and development year for the four individual lines of business*

5.3. Development Factors per Line of Business compared to Development Factors for different Leaves

In Figure ?? we can see the development factors for each of the four lines of business. We can see that the factor for the first development year varies between 1.4 and 1.7.

For comparison purposes, we can consider a triangle for each of the leaves of the decision tree. In Figure ?? we show these triangles for the leaves 1,2,10 and 11. By design the first development year differs materially for these four segments. However, we can see that also the development factors 24-36 (and following) have show considerable differences for the four leaves. This shows that the decision tree has successfully identified homoegenous reserving segments. We note that it would be straightforward to amend the decision tree hyper parameters in order to produce less than 11 segments, which would result in segments which more exposure.

5.4. Boosting

For the boosting model, we have considered 8 iterations with a learning rate of $\lambda = 0.15$ and a sampling factor of 0.6 meaning that for each individual tree we have constructed a bootstrap sample (by sampling with replacement) of the size 60% of the total data. In Figure ?? we show the development of the lift (by iteration) for the development year 1. Although the lift is initially decreasing (for iterations 2 to 4), it eventually reaches a value of 3.09 for the training data and 3.13 for the validation data. This is only marginally better than the lift of the single decision tree for development year 1 (as shown in Table ??). In our experience, boosting models (or alternative ensemble approaches), generally outperform individual decision trees by a higher margin. However, due to the limited number of covariates available for this data, the boosting model does not seem to introduce much added value for the data considered for this paper.

5.5. Comparison of estimated Outstanding Amounts

In the following we compare the estimated outstanding amounts of five different models:

Leaf 1
Evaluated as of Dezember 31, 2005
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	3.249	1.219	1.077	1.041	1.026	1.018	1.015	1.011	1.011	1.008	1.005	
12-1995	3.275	1.251	1.084	1.045	1.028	1.020	1.016	1.012	1.010	1.007		
12-1996	3.379	1.242	1.086	1.047	1.029	1.022	1.016	1.012	1.011			
12-1997	3.284	1.234	1.086	1.049	1.033	1.021	1.017	1.011				
12-1998	3.436	1.271	1.097	1.053	1.033	1.023	1.018					
12-1999	3.380	1.284	1.095	1.049	1.030	1.022						
12-2000	3.458	1.293	1.099	1.054	1.033							
12-2001	3.481	1.286	1.098	1.052								
12-2002	3.447	1.289	1.102									
12-2003	3.310	1.278										
12-2004	3.467											
12-2005												
Vol Wtd Avg	3.381	1.266	1.092	1.049	1.030	1.021	1.016	1.011	1.011	1.008	1.005	
Cumulative	5.428	1.605	1.268	1.161	1.107	1.074	1.052	1.035	1.023	1.013	1.005	1.000
Ratio to Ultimate	0.184	0.623	0.789	0.861	0.904	0.931	0.951	0.966	0.977	0.988	0.995	1.000

Leaf 2
Evaluated as of Dezember 31, 2005
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	2.773	1.175	1.061	1.032	1.020	1.015	1.010	1.007	1.004	1.003	1.002	
12-1995	2.603	1.126	1.041	1.035	1.014	1.010	1.009	1.007	1.004	1.004		
12-1996	2.579	1.124	1.043	1.021	1.015	1.009	1.008	1.006	1.004			
12-1997	2.710	1.128	1.048	1.027	1.019	1.015	1.011	1.006				
12-1998	2.591	1.127	1.043	1.024	1.011	1.010	1.009					
12-1999	2.690	1.135	1.050	1.032	1.020	1.013						
12-2000	2.705	1.150	1.055	1.028	1.014							
12-2001	2.817	1.218	1.076	1.039								
12-2002	2.681	1.187	1.069									
12-2003	2.687	1.185										
12-2004	2.577											
12-2005												
Vol Wtd Avg	2.673	1.157	1.055	1.030	1.016	1.012	1.010	1.006	1.004	1.003	1.002	
Cumulative	3.544	1.326	1.146	1.087	1.055	1.038	1.025	1.016	1.009	1.005	1.002	1.000
Ratio to Ultimate	0.282	0.754	0.872	0.920	0.948	0.963	0.975	0.985	0.991	0.995	0.998	1.000

Leaf 10
Evaluated as of Dezember 31, 2005
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	1.107	1.028	1.014	1.010	1.007	1.004	1.003	1.001	1.001	1.002	1.001	
12-1995	1.131	1.044	1.019	1.011	1.007	1.005	1.003	1.002	1.002	1.004		
12-1996	1.118	1.034	1.020	1.013	1.008	1.006	1.003	1.003	1.002			
12-1997	1.106	1.030	1.016	1.008	1.002	1.003	1.003	1.001				
12-1998	1.128	1.038	1.022	1.012	1.007	1.007	1.005					
12-1999	1.123	1.037	1.015	1.008	1.005	1.003						
12-2000	1.144	1.045	1.023	1.012	1.007							
12-2001	1.153	1.054	1.031	1.016								
12-2002	1.138	1.038	1.016									
12-2003	1.141	1.040										
12-2004	1.178											
12-2005												
Vol Wtd Avg	1.134	1.039	1.020	1.011	1.006	1.005	1.003	1.002	1.002	1.003	1.001	
Cumulative	1.242	1.095	1.054	1.033	1.022	1.016	1.011	1.008	1.006	1.004	1.001	1.000
Ratio to Ultimate	0.805	0.913	0.949	0.968	0.979	0.985	0.989	0.992	0.994	0.996	0.999	1.000

Leaf 11
Evaluated as of Dezember 31, 2005
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	1.081	1.020	1.012	1.006	1.004	1.003	1.002	1.003	1.002	1.001	1.001	
12-1995	1.098	1.023	1.012	1.006	1.004	1.002	1.001	1.001	1.001	1.001		
12-1996	1.109	1.030	1.013	1.009	1.005	1.003	1.003	1.002	1.001			
12-1997	1.093	1.025	1.012	1.007	1.003	1.003	1.002	1.002				
12-1998	1.103	1.028	1.014	1.009	1.005	1.003	1.002					
12-1999	1.100	1.026	1.010	1.004	1.004	1.002						
12-2000	1.121	1.032	1.015	1.008	1.005							
12-2001	1.102	1.029	1.016	1.009								
12-2002	1.117	1.031	1.016									
12-2003	1.127	1.033										
12-2004	1.113											
12-2005												
Vol Wtd Avg	1.106	1.028	1.013	1.007	1.004	1.003	1.002	1.002	1.001	1.001	1.001	
Cumulative	1.177	1.065	1.036	1.022	1.015	1.010	1.008	1.006	1.004	1.002	1.001	1.000
Ratio to Ultimate	0.849	0.939	0.965	0.978	0.985	0.990	0.992	0.994	0.996	0.998	0.999	1.000

Figure 8: *Development factors by accident year and development year for selected leaves of the decision tree*

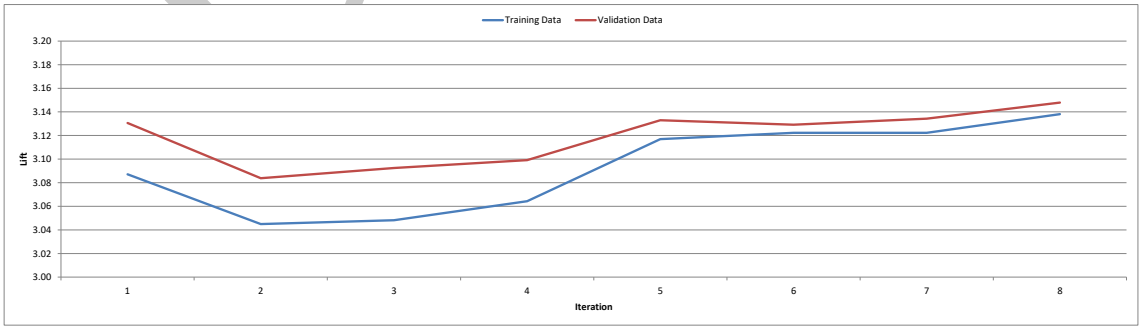

Figure 9: *Lift of boosting model for development year 1*

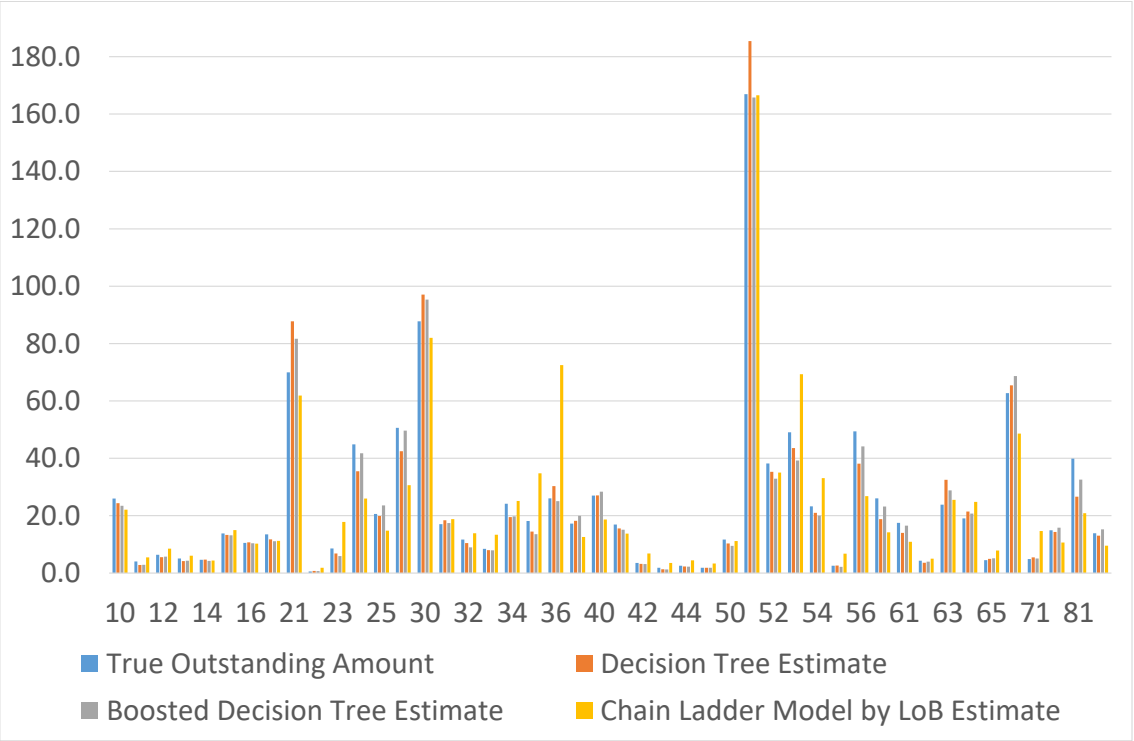

Figure 10: *Estimated outstanding amounts (in CHF Mio) versus true outstanding amount for the variable inj_part*

1. A decision tree model as described in Section ??
2. A boosted decision tree model as presented in Section ??
3. An estimate derived from 11 separate chain ladder models: one model for each leaf of the decision tree show in Figure ??
4. A single chain ladder model on the aggregate portfolio.
5. An estimate based on four separate chain ladder models: one model per line of business.

In Figure ?? we show the estimated outstanding amounts, as well as the error for each of these five models. We can see that overall the chain ladder model for each line of business, as well as the 11 chain ladder models for the 11 leaves, produce the lowest errors. We note that the main goal of our work was not to improve the overall estimate, but rather to derive a concept which accounts for individual claim properties. We will further elaborate on this point in the next section.

5.6. Estimated outstanding amounts for different cohorts

In Figure ?? we show the estimated outstanding loss amount for the variable inj_part. For better visibility we only show the results of the decision tree model from Section ??, the boosted decision tree model and the chain ladder model by LoB. We can see, that there are some values of inj_part for which the chain ladder model's projections deviate materially from the true outstanding amount (e.g. for class 23, 24, 35, 36, $53 - 56$ and 66). The decision tree based approaches generally fare considerably better for these classes. This indicates that these approaches may produce better results when a portfolio experiences a change of business mix over time.
In Figure ?? we show the corresponding graph for the variable *cc*.

6 Discussion

How to derive homogeneous reserving segments in practice? Generally this is, to a large extent, based on the experience of underwriters, reserving actuaries and other subject matter experts. For new (e.g. emerging) risks or for scenarios where a large amount of data (i.e. variables) is available, an algorithmic approach could support reserving actuaries when determining homogeneous segments.
If one had background information on the meaning of the values for `cc` and `inj_part`, one might be able to further improve the model, e.g. by pre processing the data or post processing the tree.

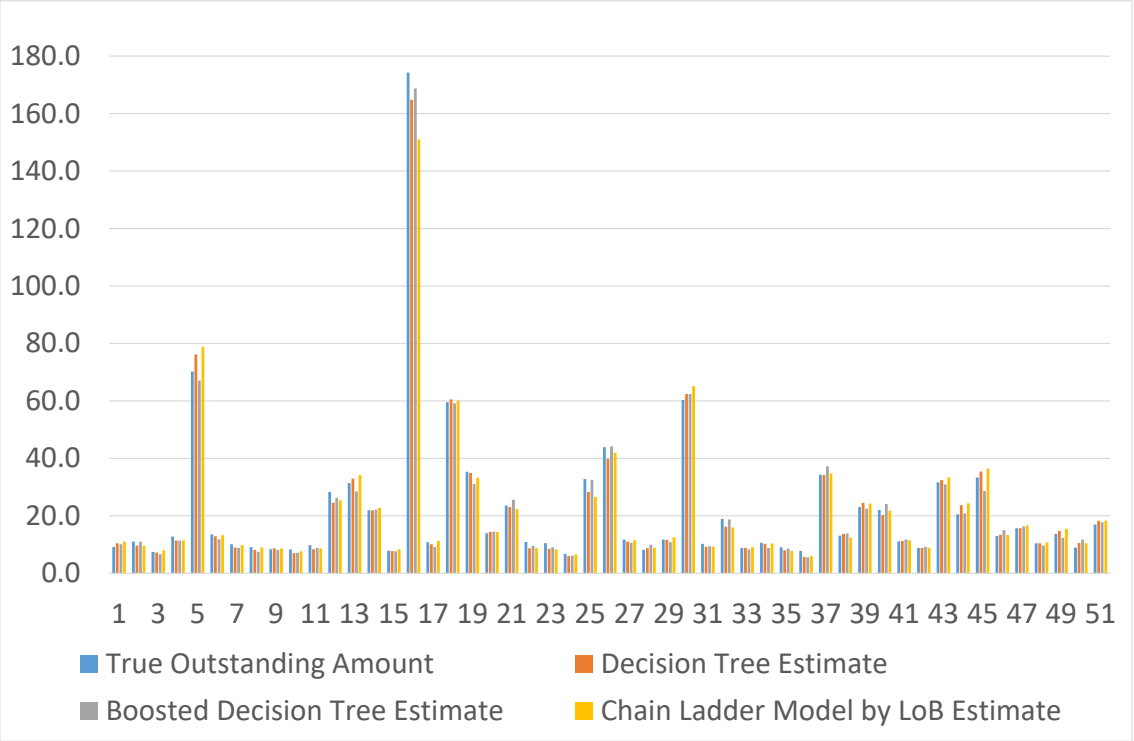

Figure 11: *Estimated outstanding amounts (in CHF Mio) versus true outstanding amount for the variable cc*

LoB	True Outstanding Amount	Decision Tree Estimate	Boosted Decision Tree Estimate	Chain Ladder Model for each Leaf of the 'Development Year 1 Decision Tree' Estimate	Aggregate Chain Ladder Model Estimate	Chain Ladder Model by LoB Estimate	Decision Tree Error	Boosted Decision Tree Error	Chain Ladder Model for each Leaf of the 'Development Year 1 Decision Tree' Error	Aggregate Chain Ladder Model Error	Chain Ladder Model by LoB Error
1	268.3	218.3	233.2	225.8	239.8	261.3	-50.0	-35.2	-42.5	-28.5	-7.0
2	208.4	273.5	230.2	283.6	290.4	194.8	65.1	21.8	75.2	82.1	-13.5
3	214.8	270.2	235.8	282.9	265.8	234.8	55.4	21.0	68.2	51.0	20.0
4	424.7	332.9	389.0	308.3	291.9	409.8	-91.9	-35.8	-116.4	-132.9	-14.9
Total	**1'116.2**	**1'094.8**	**1'088.1**	**1'100.7**	**1'087.9**	**1'100.8**	**-21.4**	**-28.1**	**-15.5**	**-28.3**	**-15.4**
Total Error in percentage							-1.9%	-2.5%	-1.4%	-2.5%	-1.4%

Figure 12: *Estimated Outstanding Loss Amounts for the different models*

We recognize that a data set with 5 million observations is larger than many claim data bases might be in practice. Therefore, we have performed models on smaller data sets, which has shown that the algorithms work perfectly fine with fewer observations.
In the presented use case, we have only made use of individual claim properties in order to derive a segmentation of the claims portfolio. Clearly, there are alternative concepts which make use of the individual claims data for projection purposes too. However, one advantage of our approach is the fact that once the segments are derived, we can applied any existing triangle based concepts. As such, the presented approach is relatively easy to adopt for an existing reserving process.
We note that, if there are exposure changes in an existing portfolio, a reserving model based on individual claims data might provide more accurate estimates than an aggregate approach, as it adjusts to the changes in business mix.

7 Outlook

Our presented approach of decision trees for (claims) portfolio segmentation could be further refined in various ways. Below are a few thoughts about possible future improvements.

1. Consider the information gain to determine when to stop growing the tree
2. Pruning of the individual trees
3. Cross Validation
4. Consider additional Ensemble Methods: Random Forest, Bagging and others
5. It is possible that the models (single tree as well as the boosting model)could be improved if one were to fit four models to each LOB.
6. Trees have been constructed independently. In practice one would likely further investigate the separate trees and possibly strive for a certain degree of consistency through the different development years

8 Appendix

8.1. R Code

The following R snippet shows the main parameters used to generate the claims data in R. We note that the code relies on the package developed by [?]

```
### Session --> Set Working Directory --> To Source File Location
setwd("~/Ressources/SimulationWuethrich/Simulation.Machine.V1")
source(file="./Functions.V1.R")

############################
### Generate the features ###
############################

V <- 5000000                          # totally expected number of claims (over 12 AY)
LoB.dist <- c(0.35,0.20,0.15,0.30)    # allocation of claims to 4 LoBs
growth <- c(0,0,0,0)                  # growth parameters (per LoB)
seed1 <- 100                          # setting seed for simulation
features <- Feature.Generation(V = V, LoB.dist = LoB.dist, inflation = growth, seed1 = seed1)

str(features)

############################################
### Simulate (and store) cash flow patterns ###
############################################

npb <- nrow(features)                 # blocks for parallel computing
seed1 <- 100                          # setting seed for simulation
std1 <- 0.85                          # standard deviation for total claim size simulation
std2 <- 0.85                          # standard deviation for recovery simulation
output <- Simulation.Machine(features = features,npb = npb,seed1 = seed1,std1 = std1,std2 = std2)
```

References

[1] [Gabrielli and Wüthrich, 2018] Andrea Gabrielli and Mario V. Wüthrich. An Individual Claims History Simulation Machine, Risks 2018, 6(2), 29, https://doi.org/10.3390/risks6020029

[2] [Wüthrich, 2017] Wüthrich, Mario V. Neural Networks Applied to Chain-Ladder Reserving (July 6, 2018). Available at SSRN: https://ssrn.com/abstract=2966126 or http://dx.doi.org/10.2139/ssrn.2966126

[3] [Mack, 1993] Mack, Thomas. Distribution-free calculation of the standard error of chain ladder reserve estimates (1993), ASTIN Bulletin 23/2, 213-225.

[4] [Breiman, Friedman, Olhsen, Stone, 1994] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees, Chapman & Hall, Boca Raton

[5] [hastie, Tibshirani, Friedman, 2009] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learnings (2009), https://web.stanford.edu/ hastie/ElemStatLearn/