Exploratory Analysis of Mortality Experience using Machine Learning

Annie Wang, ASA

October 31, 2018

1 Executive Summary

Machine learning has long transitioned from academic obscurity to mainstream hype. For life insurance companies, which have only recently begun exploring these methods, the emphasis has been on prediction – as evidenced by accelerated underwriting programs and the new Predictive Analytics exam requirement for the SOA.

Still, these techniques are not yet widely employed. To use the predictions from a model in decision-making, an actuary must be able to trust the model – and machine learning systems are almost as famous for being hard to interpret as they are for being able to predict the future. The coefficients employed in a Neural Network are completely opaque to human understanding. And what use is an unexplainable model? Black boxes belong in planes, not actuarial analyses.

This interpretation, though well-intentioned, is outdated. In fact, the same models used for predictive analytics can also be powerful tools for explaining data. This paper will demonstrate the use of two machine learning algorithms, Classification and Regression Trees (CART) and Random Forests, as tools for explaining data and actual to expected ratios (A/Es).

Exploratory data analysis (EDA) refers to the informal, yet crucial process of diving into and understanding the data. It is the first step in any data project. Traditional EDA involves a combination of univariate or bivariate methods: summary statistics, histograms, and one-way tables for the former, scatterplots and cross tables (like Table 1) for the latter. These allow interactions of at most two or three variables at a time. By taking advantage of powerful machine learning algorithms, an actuary can quickly understand higher order interactions between variables, and spot outliers that low-dimension analyses miss.

The data to be explored is preliminary industry mortality experience compiled by the SOA, which includes 3.4 million deaths and 352 million policy exposure years over 30 million rows, as well as basic policyholder characteristics.

Confusingly, the intersection of machine learning and EDA overlaps with many other data science concepts – including data mining, descriptive analytics, and anomaly detection

	Actual Claims		A/E	
	Count	Amount ($M)	Count	Amount
Perm	2,752,140	54,494	107%	99%
UL	336,297	38,827	121%	103%
Term	230,360	52,613	111%	89%
VL	58,081	9,442	108%	103%
ULSG	47,765	20,837	101%	87%
VLSG	15,164	2,878	110%	99%

Table 1: Actual to Expected experience compared to the 2015 VBT.

– but the essential goal is the same: to discover unexpected outliers, trends, or anomalies; to get a better understanding of the data; and to provoke interesting questions about the data. CART is well-suited for this task, with features that include:

- Intuitively appealing visualizations that are easily interpretable by non-technicians, yet
- Also enable the application of expert knowledge to extract substantial insight.
- Nonparametric: Does not require any distributional assumptions (normality, homogeneity, independence, etc).
- Automatically deals with complex, non-linear relationships, as well as outliers and missing data.
- Works "out-of-the-box". No need for data munging or pre-specification of non-linearities and interactions.
- Can take both numerical and categorical inputs.

CART is flexible, fast, and most of all, extremely interpretable – a combination distinct from other popular predictive models like Neural Networks and Generalized Linear Models. CART is able to pick out interesting outliers in A/Es, and identify the variables that are most important to explaining fluctuations in A/Es.

2 Classification and Regression Tree

Classification and Regression Trees (CART) are one of the most popular machine learning methods. They involve making recursive yes/no rules to partition the data into smaller, more homogenous groups – a process that is familiar to human decision-making. The set of splitting criteria is summarized in a binary decision tree.

Figure 1 gives an example of a simple regression tree model on cars where the variable of interest is miles per gallon (mpg). The top of the tree shows the overall average of 20 mpg followed by a binary split on weight. Weight less than 2.3k lbs is the strongest predictor of high mpg, followed by low number of cylinders, and finally low displacement (engine volume).The result is a partition into more homogenous groups of 10, 16, 21, and 30 mpg cars.

The CART algorithm automates the process of fitting the best decision tree structure to the data. Its goal is to use recursive binary splits to partition the data into interesting, non-overlapping regions. The general steps are as follows:

1. Start at the top node of the tree containing the complete data S. Search the universe of binary splits for every input variable and select the one that produces the two most "pure" nodes. For regression trees, where the variable of interest is numeric, this occurs with minimized residual sum of squared error:

$$SSE = \sum_{i \in S_1} (y_i - \bar{y}_{S_1})^2 + \sum_{j \in S_2} (y_j - \bar{y}_{S_2})^2 \tag{1}$$

   where S_1, S_2 are the two groups (nodes) of data resulting from the binary split of an input variable, and $\bar{y}_{S_1}$ and $\bar{y}_{S_2}$ are the averages of the response variable for the two groups produced by the binary split. That is, pick the partition into two nodes that produces the least sum squared distance to the mean for each node.

2. For each of the two new nodes, again look for the best binary split that maximizes node purity. Repeat until some stopping criterion[1] is reached. A stopping criterion prevents the tree from being grown until each node only has one observation.

[1] Eg. min required variance or min number of observations.

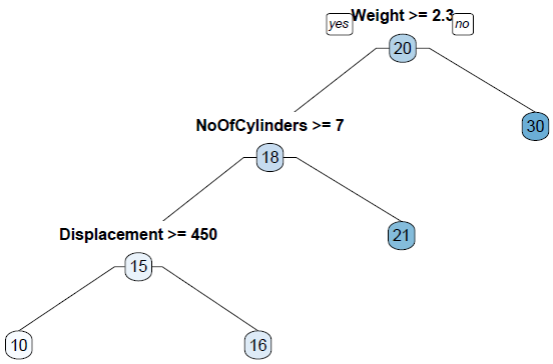


Figure 1: A regression tree on car mpg.

3. Apply 10-fold cross-validation to "prune" the tree to a more managable size and prevent overfitting. Divide the complete data into 10 non-overlapping folds each using a different 10% slice of the data as the test data and the remaining 90% as training data. Pick the subtree that minimizes a cost function that rewards accuracy and penalizes tree complexity.

For the life actuary, a regression tree can be used to automatically partition mortality experience into pockets of observations that have distinct A/Es. Groupings that have A/E much higher or lower than 100% will generate leads to investigate.

Note that Term, Perm, and Interest Sensitive Life (UL, ULSG, VL, VLSG) products are fundamentally different in

- mechanics: level premium vs. flexible; inforce by continuous payment, positive fund value, or secondary guarantee; and so on
- purpose: pure insurance vs. cash value
- policyholder characteristics: younger workers protecting dependents against breadwinner's death; children given insurance as a savings vehicle; retirees doing estate planning

and so on. Splitting the analysis by product type reduces the variance within each tree and simplifies analysis.

Claim frequency is far more stable than claim amounts, and regressing on A/E by count rather than amount eliminates variance due to both (i) the enormous range of face amounts in this data (1K-50M), and (ii) not having exact face amounts for each claim in the public data.

First, we fit a regression tree on the Term data. The actuary should expect any decent algorithm to split Term experience by post-level (typically at least 150%-200% A/E) and within-level-period; this can be used as a reasonableness check.

Figure 2 shows the result of the regression tree partition on Term experience. The top of the tree displays the overall Term A/E by count of 111% (same as in Table 1), where E is the Sex and Smoker distinct 2015 VBT mortality. Interestingly, while the tree does recognize PLT status as a high-A/E outlier, it splits on face amount first. That is, it ranks low face amount as an even better predictor of high mortality than PLT status.

This unintuitive result can be confirmed by manually calculating the A/Es that result from a first-split on PLT: 104% A/E for non-PLT and 166% A/E for PLT. This is less disparity than the A/Es resulting from the face amount partition chosen by the machine: Term policies with face amount >100K or <10K have 98% A/E, while those with 10k-99k face have 169% A/E. These correspond to 82% and 18% of expected count respectively.

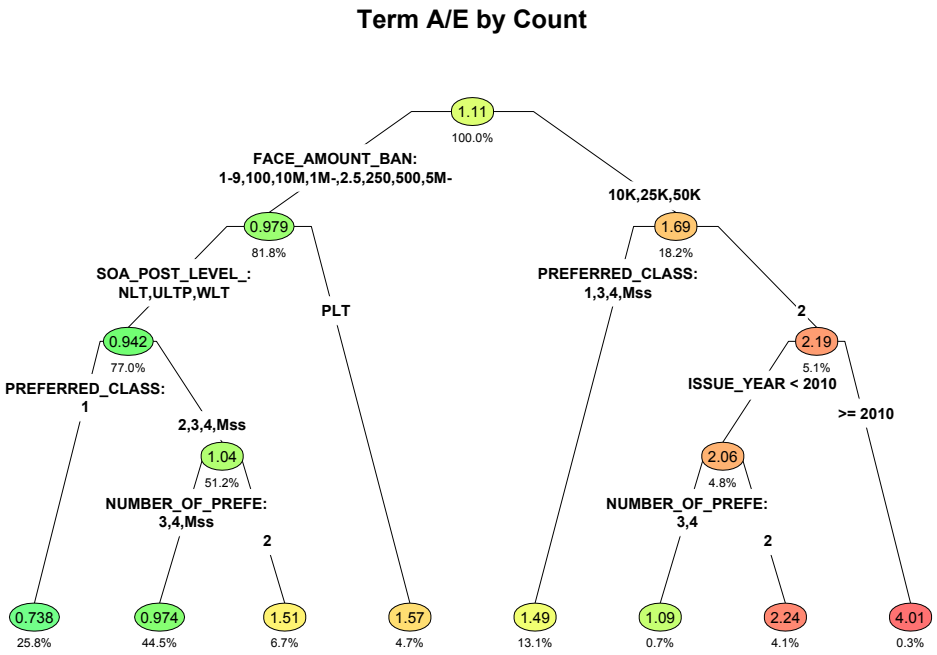


Figure 2: Regression decision tree using 10-fold cross-validation on Term data. At each node, the number in the circle is the A/E by count; below, the percentage of expected claims by count. Face amount bands are truncated here (eg. "10K" is "10K-24.9K").

An outlier may be due to random variation or bad data, or it may indicate something more theoretically interesting. The life actuary can come up with a number of different explanations of why low face amount is so strongly predictive of poor mortality:

- Low band claims and exposure are not reported consistently. (Data issue)
- Substandard mortality is associated with lower face amounts, but cannot be separated out; the data lacks a substandard rating field. It's unclear whether substandard lives are excluded entirely or included but without a label. (Data issue)
- A combination of mortality drivers are increasing the A/E associated with low face band. For example, smoker status, substandard mortality, earlier issue year, and non-preferred risk class are all linked to lower face amounts. (Correlation)
- Low face amount is a strong driver of high mortality independently of other risk factors. (Causation)

The prior belief that PLT status should have the strongest effect leads us to believe that a data issue is the most likely explanation.

At the bottom of Fig 2 are the terminal nodes (also called leaves) for which no further decisions are made. These are the final subgroupings with homogenous A/Es. A careful look at these reveals interesting trends. For 70% of the Term block, the pure 2015 VBT fits the experience quite well:

- The largest node contains 44% of Term by expected count and has A/E of 97%.
- The next largest terminal node is the best-preferred, non-PLT, non-low-face group. These make up of 26% of Term and have only 74% A/E. Policies underwritten with the best preferred class (`Preferred Class` = 1) are expected to have better mortality.

The remaining 30% of Term expectation in the next six leaves have high A/Es. They contain both expected and unexpected outliers:

- Within the non-low-face group, PLT is the strongest predictor of high mortality at 157% A/E and 4.7% of expectation by count. For low-face policies, other variables dominate PLT.
- Both sides of the tree identify `Number of Preferred Classes` value of "2", but not 1, 3, 4, or Missing (Not Preferred), as having much higher than average mortality.
  - One plausible explanation is that since nonsmoker can have 2, 3, or 4 prefered classes, while smokers can have only 2, the tree is picking up high smoker mortality (beyond the already higher expectation of the Smoker 2015 VBT).
- The last node of the tree is an extreme outlier: Low-face policies with preferred class of "2" issued after 2010 have an A/E by count of 401%. For this group, 670 claims were expected, but the actual count was 2,687.

In addition to the decision tree plot, the CART model also scores each variable on its importance to A/E. The variable importance chart in Figure 3 ranks `Face Amount Band` as the most important, followed by two preferred class variables. After is PLT status and then issue year, as well as three variables (Level Term period, duration and issue age) that didn't make it into the decision tree in Figure 2, but were close.

Many input variables were not important enough to even make it onto the variable importance plot, including: `Observation Year`, `Gender`, `Smoker Status`, and `Common Company Indicator 57` (if a company submitted data for at least 5 of 7 years). This highlights the tree model's intrinsic feature selection. Variables unimportant to splitting A/E are less likely to be far from 100% or to have severe data issues.

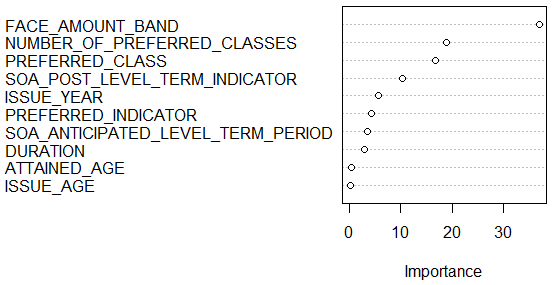


Figure 3: Relative variable importance in the Term CART model.

The variable importance plot also provides an example of CART's bias towards variables with many distinct values. `Preferred Indicator` has values of 0 or 1, and is described completely by `Preferred Class` (1-4 for preferred, Missing for non-preferred). Binary variables like `Gender` or `Smoker Status` have only one possible binary split. Compare this to `Issue Year`, which for Term ranges from 1926 to 2015, or Face Amount Band, a categorical variable with 11 values and 1,023 possible binary splits.

Next, we examine the regression trees fit on Perm and ISL. Note that The Perm and ISL models also identify low face amounts as strong predictors of high mortality:

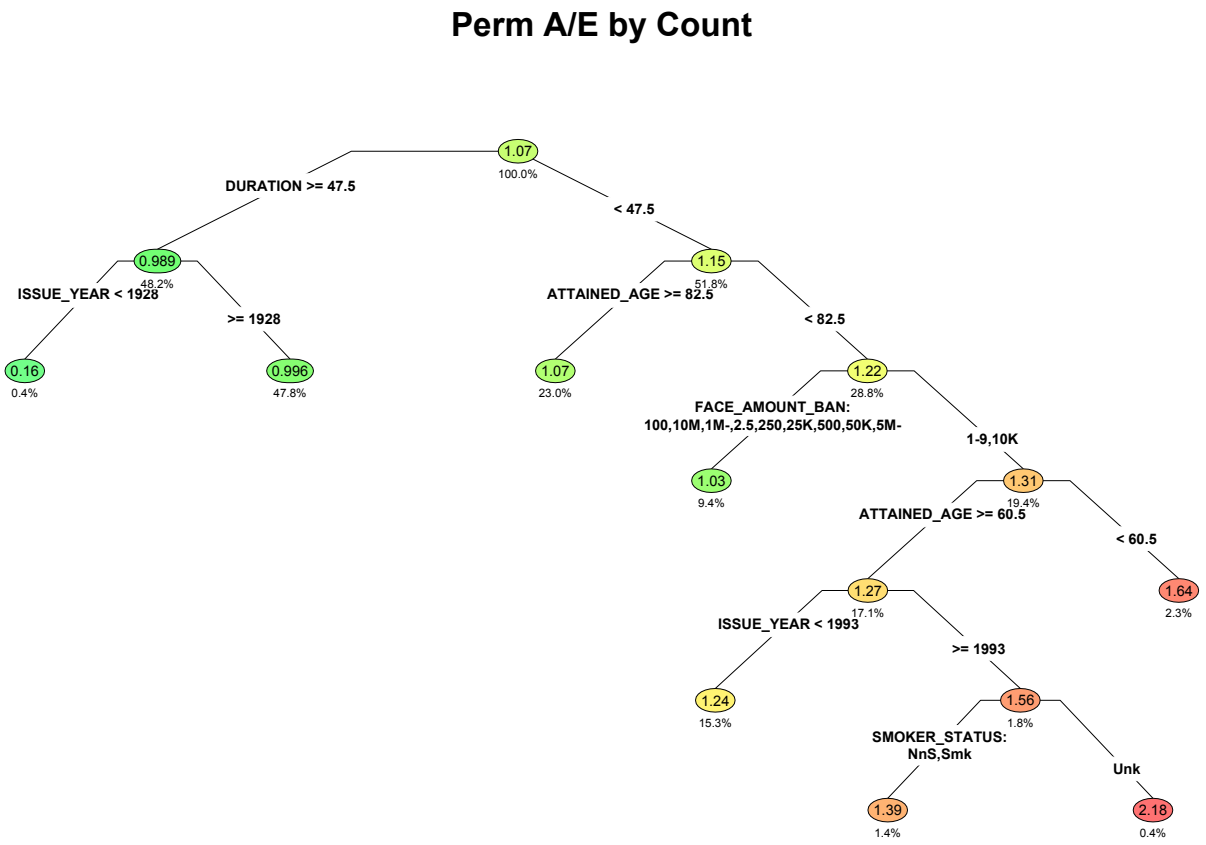
- For ISL a face amount of 1-99K is the most important predictor of high mortality.
- Perm places more weight on issue year, duration and attained age than low (1-24K) face amount.
- The effect of low band on increasing A/E is smaller for ISL and Perm (~20-30%) than for Term (~70%).
- While the ratio of non-low-band to low-band for Term in Fig 2 is 4:1, for ISL the ratio is 1:2. For Perm, low-band also outnumbers non-low-band.

For Perm, the immediate split divides Perm 50-50 into policies with duration greater than or less than 48, with A/Es of 99% and 115% respectively. Note

- In the duration $\geq$ 48 branch, those issued before 1928 make up less than 1% (0.4% / 48.2%) of expectation, but are unusual enough at 16% A/E to skew the results: Removing this outlier raises A/E from 99% to 100%.
- In the duration $<$ 48 branch, those with attained age $\geq$ 83 have 107% A/E compared to 122% for those that are younger. Within the younger 122% group the tree found a group with "normal" mortality of 103%, corresponding to those with face amounts greater than 25K.
  - Within the low band ($<$25k) group, those with attained age $<$ 61 form a high outlier at 164%.
  - For those with attained age $>$ 61, the 127% A/E is further split into terminal nodes of 124%, 139%, and 218% based on issue year $>$ 1993 and Unknown smoker status.

For the ISL decision tree,

- Face amount of 1-99K is the most important predictor of high mortality.



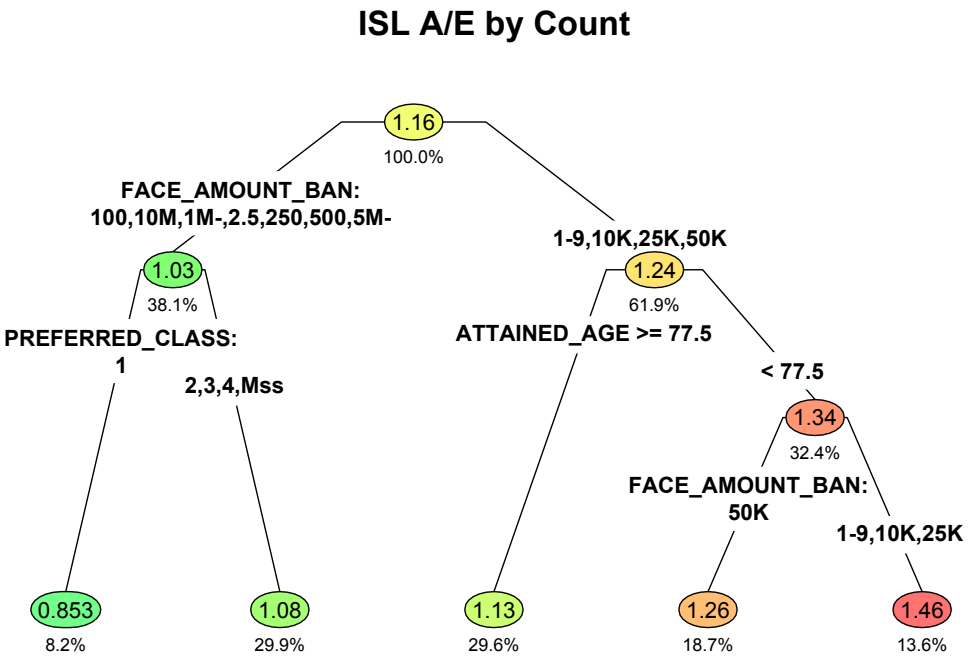


Figure 4: Regression decision trees on Perm and ISL.

- In the low-band group, those with attained age ≥ 78 have 113%, compared to 134% for those younger than 78. This echoes the attained age split at 83 in the Perm tree, where younger attained age is associated with higher mortality, suggesting a true mortality trend common between ISL and Perm.

- The tree is small with only 4 splits. A very slight increase to the complexity parameter increases the tree size from a depth of 3 to 14. To preserve interpretability (and printability), the small decision tree is chosen. Alternatives include fitting a tree for each of UL, ULSG, VL, VLSG separately, or SG versus non-SG.

Low face amount is an important predictor of high mortality in all Term, Perm, and ISL decision trees. This data issue affects all product types and must be dealt with before the final experience analysis.

To summarize, tree-based models are simple to use, fast, and graphically interpretable. The machine conducts an exhaustive search for the best binary splits, calculates A/Es for each partition, and conveniently packages them into a printable, richly interpretable plot as in Fig 2. In addition, the tree model is able to automatically spot and single out outliers as terminal nodes. By identifying homogenous subgroups of the original overall A/E, the algorithim automatically sifts out outliers – pockets of observations with A/Es that deviate markedly from 100%. These qualities make CART models ideal for actuarial EDA.

3 Random Forest

While CART models are highly interpretable and excellent exploratory tools, in the predictive analytics space they have several disadvantages.

- The simplicity of the small tree model also gives it sub-optimal predictive performance compared to other ML techniques. A single tree may need to be grown very deep to produce accurate predictions, which negates interpretability.

- Individual tree models tend to be unstable: a small change in the data can potentially produce a completely different set of splits.

These issues are addressed by extensions of CART such as Random Forest, an ensemble method that constructs many decision trees. By building and averaging the predictions of hundreds of trees, a random forest produces more granular predictions that rival more complicated techniques in accuracy. The general steps to the algorithm are as follows:

1. Select the number of models to build, eg. 500.

2. For each tree,

   (a) Generate a bootstrapped training sample by taking repeated samples from the data.

   (b) Randomly select k of the original predictors. For a model with p predictors this is usually $\sqrt{p}$ or $p/2$.

   (c) For each split, choose the best predictor among the k predictors and partition. Stop growing the tree when the stopping criterion is reached.

3. Take the average (or some other linear combination) of the predictions of the 500 trees.

The random forest greatly reduces bias and model noise compared to a single tree using all predictors. The curious practice of only using a subset of potential inputs for each tree reduces correlation between trees and potentially allows for globally optimal splits. The single tree CART is a top-down, greedy process that only optimizes one decision at a time. By randomly permutating the input variables, the random forest allows a less important variable to be split first and potentially produce better partitions later on.

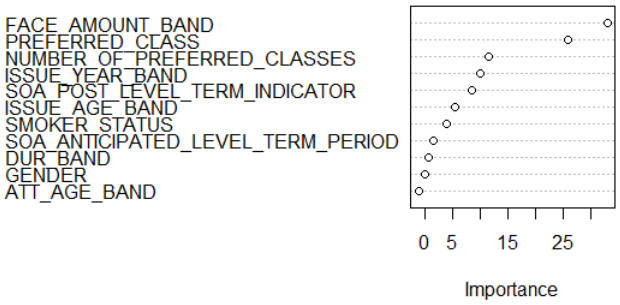


Figure 5: Relative variable importance in the random forest Term model.

At the same time, the ensemble approach of averaging hundreds of models results in the loss of the intuitive visualizations that made tree models so appealing in the first place. Relative to CART, random forest is a black box; inspecting the structure of a single tree in a random forest model tells the user almost nothing about the forest's predictions. Still, random forests greatly outclass the single decision tree in pure predictive power, as seen by their popularity in machine learning competitions. Do they have value in exploratory analysis as well?

Random forests are very memory intensive, and the typical PC does not have enough memory to fit a random forest on data with millions of rows[2]. The Term data can be compressed from 13 million rows to under 200,000 by banding numeric variables like issue age and duration, as well as summing experience over all observation years.

The two main types of visualization for random forests include variable importance plots (as seen in CART earlier), and partial dependence plots (for numeric variables). Figure 5 displays the variable importances for variables the Term random forest.

Figures 3 and 5 share many similarities. Both use mean decrease in impurity as the measure of variable importance. They also share nearly all of the 10 most important variables. Both select `Face Amount Band` as the highest ranked.

In the random forest, issue year has moved above PLT indicator in importance. Issue age has moved from #10 to #5. The `Preferred Indicator` variable was dropped from input variables as it is wholly described by `Preferred Class`, and in the random forest variable importance plot has been replaced by `Gender`. `Issue Year` has moved above PLT indicator in importance, which warrants investigation for data issues. A possible starting place is a comparison of pre- and post-2010 issue years, as suggested by the CART plot.

The drawback of variable importance is that it does not describe the nature of the relationships. Figure 5 by itself tells us nothing about whether A/E increases with face amount, or what specific splits of face amount are important; only that it is the "best" at decreasing node impurity. CART makes the binary splits and resulting A/Es explicit, giving it a numerical clarity that the variable importance plots lack. The variable importance plot also fails to explain any interactions between variables. It's difficult to translate the small differences in importance rankings between CART and RF on Term back into any concrete improvements to Figure 2. In this case, the random forest does not give additional insight over the simpler CART model.

[2]http://datascience.la/benchmarking-random-forest-implementations/

4 Conclusion

Tree based models like CART and Random Forest are powerful and versatile tools for non-parametric classification and regression. They naturally handle mixed data, large data, nonlinearity, and outliers without manual heavy lifting. CART in particular is exceptionally interpretable and extracts valuable information quickly and with minimal set up. These qualities make ideal for exploratory data analysis, where intuitive visualization, ease of use, and interpretability are key.

Random forests are a variant of tree-based machine learning models that solve the problems of coarse predictions, instability, and local greediness associated with CART. At the same time, they trade away interpretability and explanatory power. CART is better for the informal process of EDA as it is a simple, off-the-shelf procedure that produces immediate insights. It produces visualizations that are interpretable by either the domain expert or the non-technician.

In this paper we've seen how CART can be applied to mortality experience studies to discover the most important variables associated with unusual A/Es and identify unusual outliers in the data. CART is a worthy addition to the actuary's toolbox. Directions for future work include application of other data mining methods, such as PRIM and Bayesian networks for outlier detection. PRIM is similar to CART in taking a nonparametric, recursive splitting approach to decreasing the heterogeneity of the data, and may be better at detecting outliers in complex data sets. MARS is an technique that automatically fits piece-wise linear functions to model nonlinearities, and can be modified to generalize the CART splitting algorithm. Traditional parametric methods such as log-link GLM and logistic regression may be improved by combining with feature selection algorithms and MARS for nonlinearities.

A Appendix - Data preprocessing

The data is provided by the SOA for the 2018 Data Analysis Contest, which is sponsored by the Society of Actuaries' Individual Life Experience Committee. Note the Contest disclaimer applies: "DISCLAIMER: This dataset is preliminary and should not to be used for any professional purpose, including but not limited to, pricing, product development, valuation or other risk management purposes."

This section contains the data issues I found by the usual univariate/bivariate methods and applied before the analysis.

- Missing data. Of the total 32 fields in the data set, two contain missing data. `Number of Preferred Classes` and `Preferred Class` each contain 10,979,288 missing values. This encompasses over a third of the entire data set and likely corresponds to non-preferred underwriting. Every observation that is missing `Number of Preferred Classes` is also missing `Preferred Class`, and vice versa. Other fields explicitly label unknown or not applicable attributes.

- Level Term with unknown Level Term Period. While most Term policies have a Level Term Period in $5, 10, 15, 20, 25, 30$, a significant amount of Level Term exposure has "Unknown" level period. A split by observation year shows that all level periods from 2009-2011 are Unknown, while approximately 10% of data after 2012 has unknown level period.

Year	% Unknown LTP
2009	100%
2010	100%
2011	100%
2012	9%
2013	9%
2014	10%
2015	10%

Table 2: Percentage of missing level term period.

- Suspicious number of supercentenarians. A summary table of attained age over all observations reveals that the maximum in the data set is 120. This is immediately suspicious, since only one person to date has ever been verified to have lived to age 120. Table 3 summarizes exposure for the oldest observations. In 2015 the exposure for 120-year-olds jumped from 2.9 to 24.3 policy years, yet in 2014 the exposure for 119-year-olds was only 8.6. This cannot by explained by new issuance to 119-year-olds either – not only would that be bad business practice, but the maximum issue age in the data set is only 99.

	Attained Age						
Year	120	119	118	...	110	...	106
2013	3.3	7.3	6.8	...	40	...	218
2014	2.9	**8.6**	13.9	...	61	...	204
2015	**24.3**	39.3	33.9	...	91	...	207

Table 3: Policy exposure year count for the oldest attained ages.

- Records with zero exposure. The data set contains 232,828 records containing zero exposure (`Policies Exposed = 0` and `Amount Exposed = 0`) and zero expected claims. Three of these records have positive `Death Claim Amount` totaling $2.2MM.

- "Other" product type has enormous outlier of actual claims and exposure in 2014. See next Appendix.

Missing Preferred Class was marked with "Unknown". For Term, recognize that it won't be possible to separate level and post level mortality prior to 2012. Given the highly questionable biological validity of the extreme older ages, as well as the unexplained jump in 2015 exposure, remove all records with attained age 106 or greater from the rest of the analysis. All records with zero exposure were removed from later analysis as they do not contribute to expected. "Other" product type was removed.

B Appendix - Table selection

The data set gives us a choice of five different sets of mortality tables to use in calculating expected counts and amounts of claims. These include the following:

- 1975-80 Modified Basic Table with extension of issue ages through 99 and attained age to 120 (1975-80 MBT),
- 2001 Valuation Basic Table (2001 VBT),
- 2008 Valuation Basic Table (2008 VBT),
- 2008 Valuation Basic Table Limited Underwriting (2008 VBT LU), and
- 2015 Valuation Basic Table (2015 VBT).

Instead of inspecting tables of A/Es, we can plot the actual to expected experience by product type using 100% of each table mortality as the E.

Figure 6 and 7 show A/E by Count and Amount, respectively. Immediately, the huge Other outlier in 2014 stands out as a data issue.

In general, the older 1975-80 MBT and 2001 VBT tables greatly overestimate both count and amount. The 2008 VBT and 2015 VBT are far closer to actual. The 2015 VBT is a better fit by amount than 2008 VBT for every product, but tends to underestimate claim count. The unusually high count of low face band policies explains much of the by-count variance. Thus the 2015 VBT is an appropriate choice for this exploratory analysis.

C Appendix - Technical specifications

All analysis was done in R. The packages used include

- **rpart** - Recursive partitioning for classification, regression and survival trees.
- **rpart.plot** - Pretty plots for rpart models.
- **ranger** - A memory-efficient implementation of the random forests algorithm that also allows for weighting.
- **tidyverse** - The classic data manipulation package.

This document was prepared in LyX, an open source document processor that uses the LaTeX type setting engine. The dataset is over 30 million rows long and 10GB large, but 16GB of RAM is sufficient.

D Appendix - Perm and ISL CART Variable Importance

Figure 8 shows variable importance plots for the CART models in Figure 4.

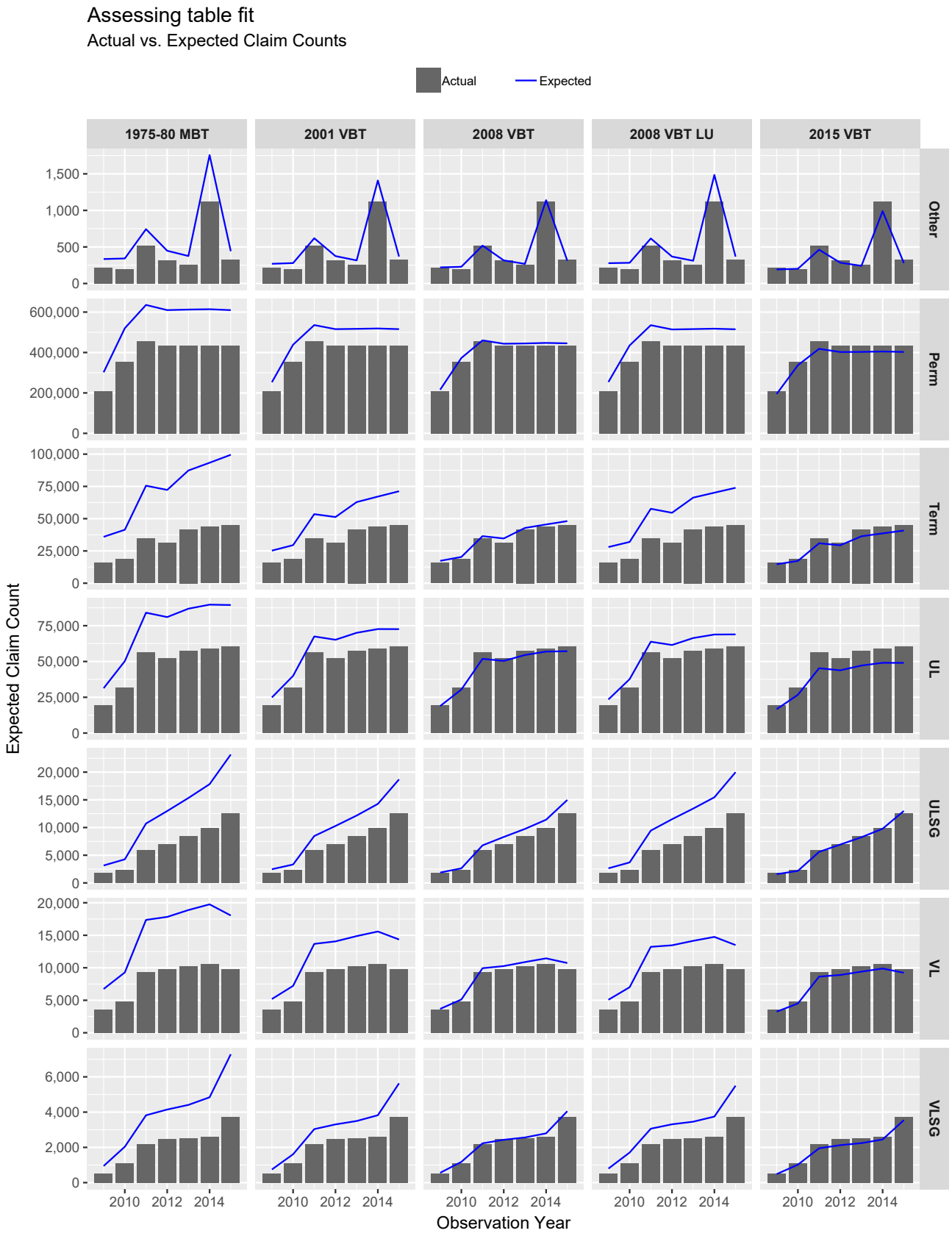


Figure 6: A/E by Count.

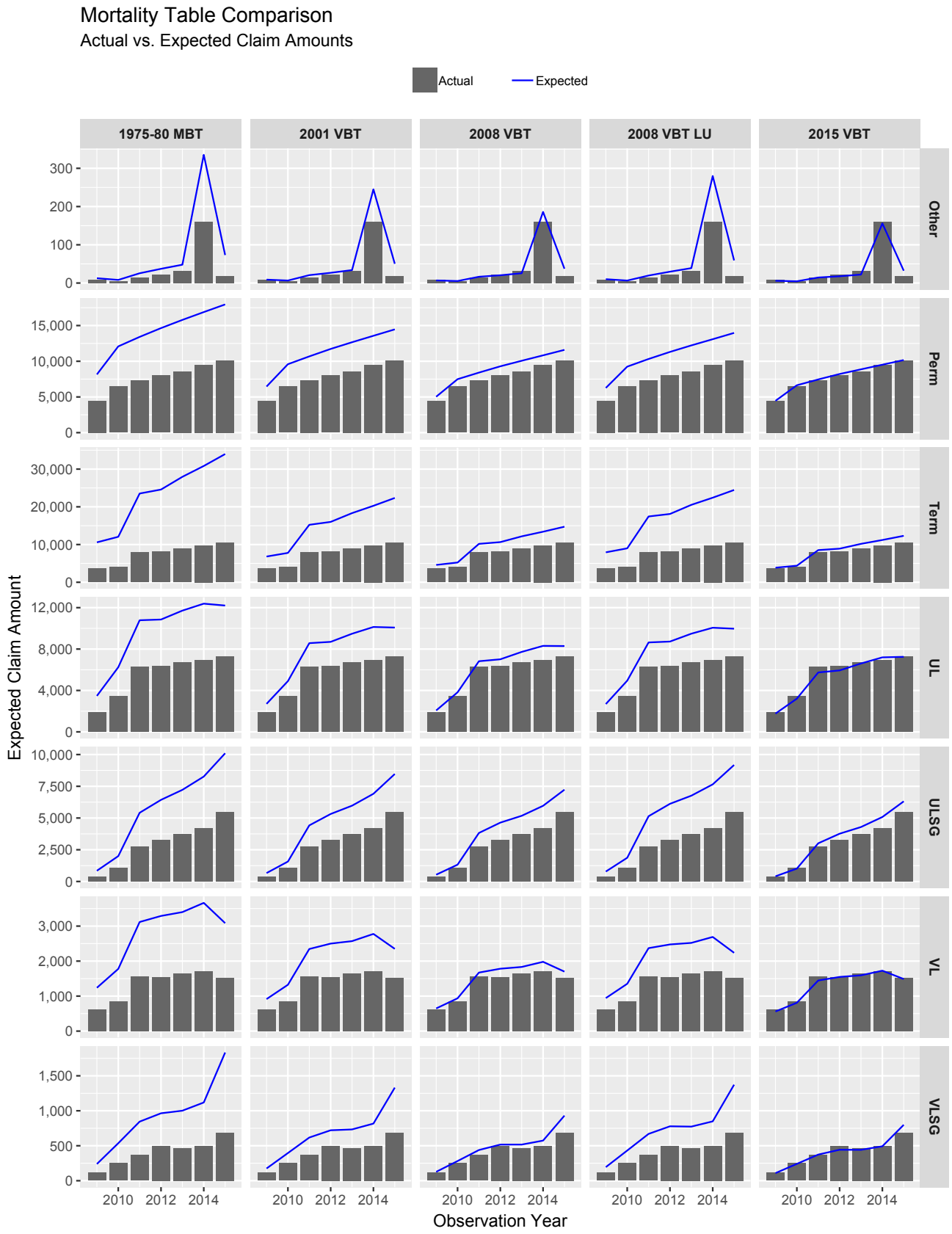
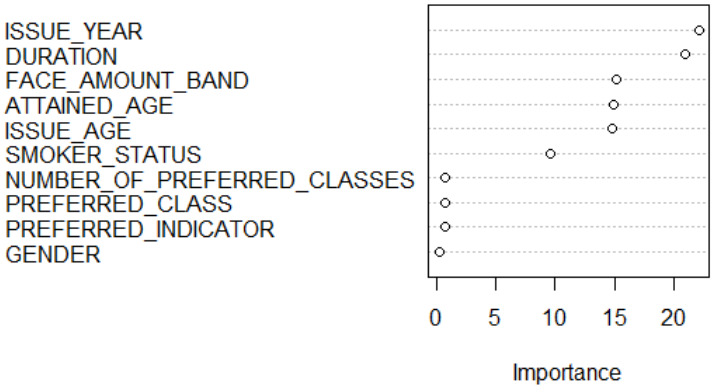


Figure 7: A/E by Amount.



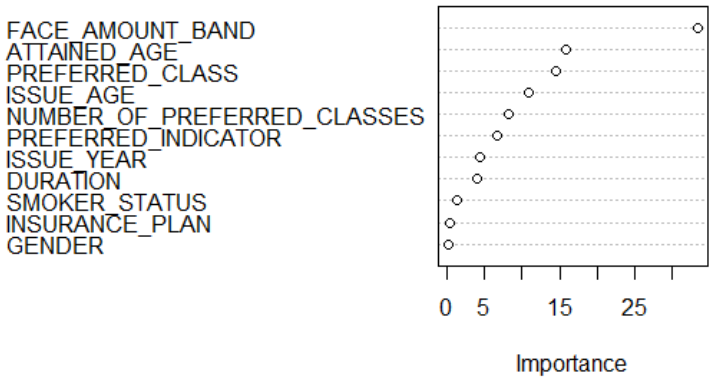


Figure 8: Variable Importance for Perm (top) and ISL (bottom).

References

1. James, G. & Witten, D. & Hastie, T. & Tibshirani, R. (2016). *Introduction to Statistical Learning with Applications in R*. Springer Series in Statistics.

2. Kuhn, M. & Johnson. K (2013). *Applied Predictive Modeling.* Springer Series in Statistics.

3. Hastie, T. & Tibshirani, R. & Friedman, J. (2013). *The Elements of Statistical Learning.* Springer Series in Statistics.

4. Jones, Z.M. & Linder, F. (2015). Exploratory Data Analysis using Random Forests. Prepared for the 73rd annual MPSA conference.

5. Holland, B. (2018). Beyond Actual to Table: Models in Experience Studies. Society of Actuaries.

6. Kolyshkina, I. & Brookes, R. (2002). Data mining approaches to modelling insurance risk. PricewaterhouseCoopers.

7. Andriyashin, A. (2005). Financial Applications of Classification and Regression Trees. A Master Thesis Presented to CASE - Center of Applied Statistics and Economics, Humboldt University, Berlin.

8. Ott, A. & Hapfelmeier, A. (2017). Nonparametric Subgroup Identification by PRIM and CART: A Simulation and Application Study. Computational and Mathematical Methods in Medicine.

9. Pollack, C. (2003). Using Non-Parametric Techniques to Understand Your Data. Presented to the Institute of Actuaries of Australia XIV General Insurance Seminar 2003.