SOCIETY OF
ACTUARIES

Predictive Analytics

2016 Call for Essays

AUGUST 2016

SOCIETY OF ACTUARIES

I’ve been fascinated by the broad and important worldwide decisions in recent months that have leveraged the power of predictive analytics. Equity and currency exchange markets rose and fell as forward-looking models estimated the chances of the Brexit from the European Union. Business decision-makers in Canada began estimating the outlook for privately sponsored pension plans as finance ministers looked to expand the Canada Pension Plan. And in the United States, it wouldn’t be the 2016 election cycle without a wide variety of predictive models using every new piece of information to estimate probabilities on a variety of outcomes. I’m guilty myself of gleefully over-indulging on FiveThirtyEight podcasts to catch up on the most recent changes to general election forecasts.

It seems we’re to the point where not a day goes by that the topic of predictive analytics doesn’t come into our actuarial conversations. With the growing amount of data continuously being generated on consumer behavior, analytics tools and methods continue to grow in importance for our profession. These methods have become more engrained in how we determine prices for products, estimate future liabilities and help our businesses make better decisions. SOA meetings now routinely include more education and opportunity for actuaries to learn deeper concepts in predictive analytics.

As a continuing way to highlight how actuaries are using these concepts in practice, we invite you to read through the following collection of articles from our members. They provide additional insights into how actuaries are using predictive analytics to provide solutions in their work. Let us know your thoughts and share other examples of predictive analytics in action at the SOA Engage Research Community, our new online forum for discussing ideas.

R. Dale Hall

R. Dale Hall, FSA, CERA, CFA, MAAA
Managing Director of Research
Society of Actuaries

Contents

This publication is provided for informational and educational purposes only. Neither the Society of Actuaries nor the respective authors’ employers make any endorsement, representation or guarantee with regard to any content, and disclaim any liability in connection with the use or misuse of any information provided herein. This publication should not be construed as professional or financial advice. Statements of fact and opinions expressed herein are those of the individual authors and are not necessarily those of the Society of Actuaries or the respective authors’ employers.

Comparing Policyholder Efficiency in Variable Annuity Lapses

Jenny Jin, FSA, MAAA, and
Vincent Embser, ASA, CERA, MAAA

Introduction

People are living longer and healthier lives. The need for retirement products to supplement retirement income is also increasing. The life insurance industry has responded to this increasing demand by offering a wide array of annuity products. In particular, variable annuity products have grown to become an important part of retirement planning, due to the attractive benefit of guaranteed lifetime income.

There are many types of options embedded within a variable annuity contract. These range from product features such as income guarantees to implicit options such as the choice to keep or cancel the contract. Options are valuable to customers, and how customers use these options provide important information for insurance providers. In this article, we will study three common factors that actuaries use to formulate the lapse assumption for Guaranteed Lifetime Withdrawal Benefit (GLWB) policies and how predictive modeling can help actuaries establish more appropriate lapse functions. Lapse behavior, or conversely persistency behavior, is a key assumption in the pricing, valuation and risk management of variable annuity contracts. Given the embedded guarantees in the variable annuity product, the economic impact of lapses can vary based on how valuable the guarantees are. Companies that are better positioned to use data to understand how policyholders behave can ultimately gain an edge.

Model Forms

Under the traditional framework used by companies to set their assumptions, the lapse rate for a policy is assumed to follow a simple equation:

$$\text{Lapse Rate} = \text{Base Lapse Rate} \times \text{Dynamic Lapse Factor}$$

where the base lapse rate is a function of the policy's duration and the dynamic lapse factor is a function of moneyness.

Moneyness is defined as the ratio of guaranteed benefit value to account value. To set these assumptions, companies often rely on a tabular experience study approach or on their own judgment in situations where data are not available.

Under a logistic predictive model framework, a number of coefficients are jointly estimated for an underlying data set, and the model directly outputs lapse probabilities as a function of a number of explanatory variables. The basic functional form of these predictive models is as follows:

$$\text{Lapse Rate} = \text{Odds} / (1 + \text{Odds})$$

where

$$\text{Odds} = \exp(\text{Intercept} + B_1 \times \text{Variable}_1 + \ldots + B_n \times \text{Variable}_n).$$

Both traditional and predictive models account for the effects of duration, moneyness and product type. However, the predictive model framework provides a statistically grounded method for estimating these effects together and can be readily expanded to include additional variables, as well as interactions between variables. In addition to the duration and moneyness factors, a predictive model can include other policy, demographic and macroeconomic variables. In this article, we have focused on a simplified version of the predictive model using just duration and moneyness for comparison with the traditional model.

Data

The "traditional model" referenced in this article is a set of pricing assumptions for a GLWB product with a seven-year surrender charge (SC), taken from a recent survey of variable annuity writers. Note that this model is not representative of any single company but reflects the average assumptions for base lapse rates and dynamic lapse factors across a number of companies.

The "predictive model" referenced in this article is based on a Milliman study referred to in this paper as the VALUES[1] model. For building the predictive model, we looked at quarterly lapse experience of

[1] Variable Annuity Lapse Utilization Experience Study

GLWB products based on 21 million records from 12 major variable annuity writers. We used a 70% random sample of these records as a training data set to fit the predictive model. We evaluated the predictability of the model on the remaining 30% holdout data set.

Comparison of Lapse Models

DURATION EFFECT

Figure 1 shows lapse rate predictions from these models when applied to the seven-year SC policies in the holdout data set. The blue line gives the base lapse rates for a hypothetical *at-the-money* policy during the first 15 policy years according to our representative industry model. The orange line shows the average lapse rate predicted by the same model, but incorporating the dynamic effect from each record's individual level of moneyness. This line is lower than the base lapse rate, as policies issued in the last 10 years emerged from the 2008 global financial crisis and tended to be in-the-money over time, which reduced the predicted lapse rates. The red line shows the actual lapse experience observed between 2007 and 2013. Finally, the green line provide the lapse rates from our predictive model, aggregated across these same policies.

One key observation is that the traditional model produces higher aggregate lapse rates for the shock year (the year immediately after SCs expire) and the post-SC years. In particular, the post-SC aggregate lapse rates from the predictive model are approximately 3 to 3.5% lower per year than rates from the average industry assumption. This could have significant implications for pricing and valuation, because the difference in annual lapse rates would be compounded over years.

MONEYNESS EFFECT

If we hold all else equal and vary the moneyness variable in our predictive formula, we can construct a dynamic lapse curve for comparison with the industry assumption. Figure 2 shows this relationship.

The slope of the dynamic lapse factors reflects the efficiency of policyholder behavior. This graph indicates

Figure 1 Aggregate Lapse Rates by Duration

Predictions and Exposure Based on Holdout Data Set (30% of Total)

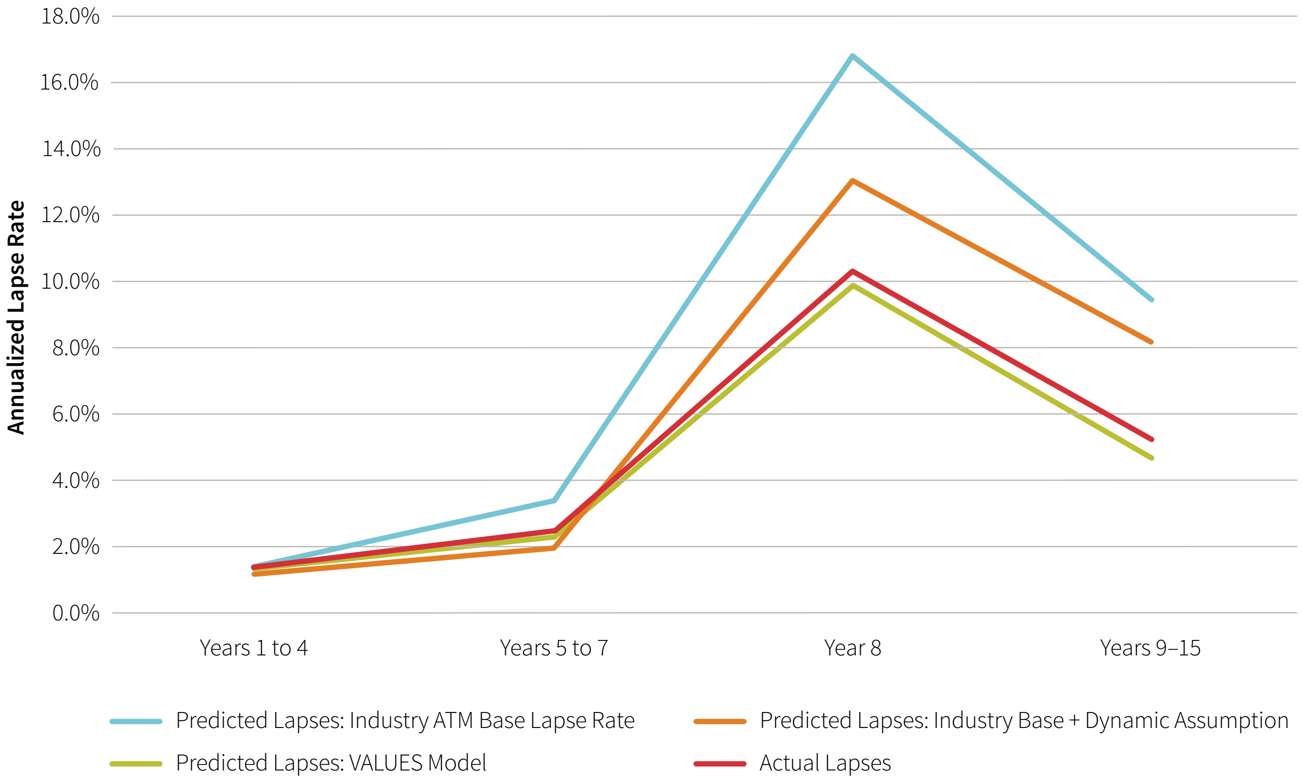


Figure 2 GLWB Dynamic Lapse Relative to Moneyness

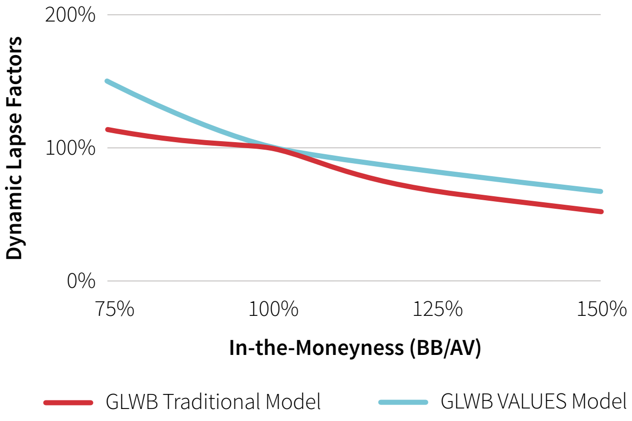
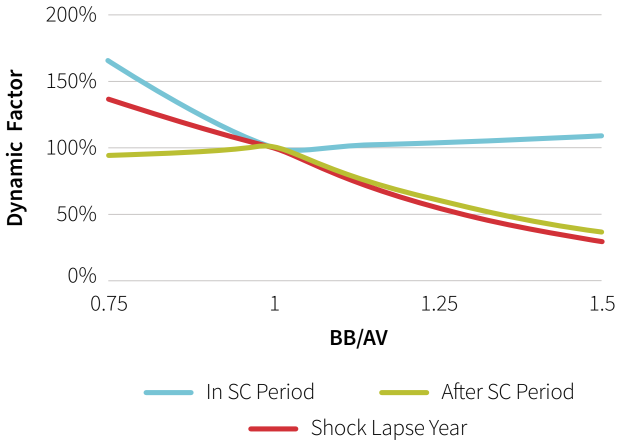


Figure 3 GLWB Dynamic Lapse Behavior Using Predictive Model



that the traditional model (red line) assumes lower dynamic lapse factors for policies both in- and out-of-the-money, compared with the predictive model. That is, typical industry assumptions underestimate sensitivity for out-of-the-money policies and overestimate sensitivity to in-the-money policies. On average, the overall lapse rates predicted from the industry assumption are higher than actual experience, as shown in Figure 1. This means that companies are likely overestimating lapses despite adjusting for dynamic behavior, and the true dynamics of how policies behave is potentially lost in the data.

Lapse Behavior During and After the SC Period

Some actuaries believe that policyholders do not behave the same way when a SC is levied. If the relationship between a target variable and an explanatory variable differs depending on the value of another explanatory variable, then there is an interaction between the two explanatory variables. One of the most attractive aspects of a predictive modeling approach, relative to the traditional model, is the ease with which this type of interaction relationship can be explored and added to a model, arriving at credible estimates. In a traditional framework, data insufficiency across multiple dimensions would preclude such interactions from being incorporated.

To illustrate this dynamic, we added interactions between the in-the-moneyness variable and the three distinct SC phases. Figure 3 uses this expanded model to derive dynamic lapse curves for each phase (during in blue, end in red, after in green) of a seven-year GLWB product.

This new model implies that policyholder behavior with respect to moneyness varies across the life of the policy. In particular, policyholders seem to be most efficient at the end of their SC periods (red line). Within the SC period, policies are observed to behave inefficiently, which results in an inverted slope (blue line).

In future work, we will continue to explore these and other interactions and evaluate the significance of their effects. By analyzing policy behavior, companies can gain further insights into how their customers are interacting with their products. Doing this well will empower companies to use these insights for better product development and in-force management.

Jenny Jin, FSA, MAAA, is a consulting actuary at Milliman Inc. in Chicago, IL. She can be reached at *jenny.jin@milliman.com*

Vincent Embser, ASA, CERA, MAAA, is an associate actuary at Milliman Inc. in Chicago, IL. He can be reached at *vincent.embser@milliman.com*

Insurance Product Recommendation System

Kailan Shang, FSA, CFA, PRM, SCJP

Recommending an appropriate insurance product is important for insurers to acquire new business from either a new client or an existing client. The client's demographic and financial information is useful for predicting the most wanted insurance product type. With the right recommendation, the probability of completing a sale will be higher and the length of time to make a sale can be shortened. Traditionally, insurance agents have used household financial planning to help clients choose insurance products. However, it requires significant time, skill and experience on the agent's part. Other distribution channels such as telemarketing may not provide the opportunity to conduct such a complicated analysis for customers. On the other hand, insurance companies hold relevant information that is very helpful for predicting the next likely sale to a client.

Business Case

A life insurance company wanted to improve the effectiveness of its selling efforts to reduce cost and increase sales volume. The company has millions of existing policyholders, and it had established a partnership with another financial institution that allows cross-selling. Demographic information, purchase history, financial information and claim information on existing customers are available. The company was interested in knowing the most likely product that a client would purchase and the probability that the sale would be completed.

Data

Five categories of data are used for the project:

1. **Demographic information** including age, gender, address, ZIP code, smoker/nonsmoker, health status, occupation, marital status and information about dependents
2. **Financial information** including assets, real estate, income, loans and spending
3. **Purchase history** including product type, product name, issue age, face amount, premium rate, face amount change, partial withdrawal, policy loan and product conversion
4. **Claim history** including time, amount, payment, etc.
5. **Communication history** including last contact time, reason, outcome, complaints, etc.

For categorical variables such as address, ZIP code and communication reasons, dummy variables are created to represent them. The data are not complete for all clients. For missing demographic or financial data, the value in the most similar record is used. The similarity is measured by the Euclidean distance given by

$$\sqrt{\sum_{i=1}^{n} (Y_i - X_i)^2 \quad i \neq l}$$

where

X is the data record with missing value for variable *l*.
Y is a complete data record in the data set.
n is the number of variables in the data set.

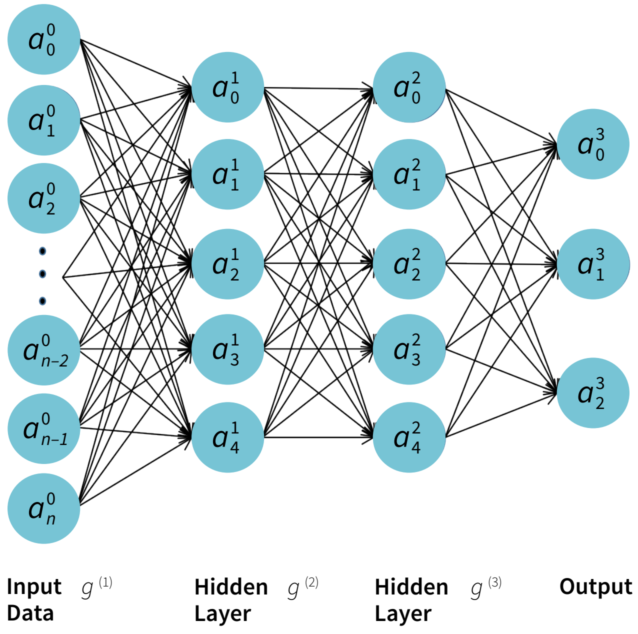
Models

Given the large amount of explanatory variables and the complicated relationships, traditional linear and nonlinear regression models that require exact model specification are not suitable. An artificial neural network (ANN) model was chosen to estimate the probability of new insurance purchases. ANN models mimic human neural networks, which are capable of making complicated decisions with layers of neurons. ANN models can approximate complicated relationships whose model specifications are unknown. A multilayer free-forward neural network model was used for the estimation. Figure 1 shows the model structure.

Notes:

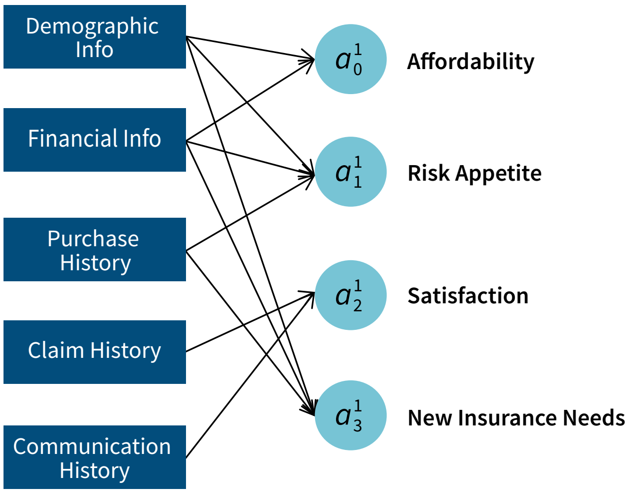
1. A sigmoid function is used for specifying the relations between layers, $g(x) = \frac{1}{1+e^{-x}}$.
2. Each node in the network is determined by the nodes in the previous layer, $a_j^i = g\,(\theta_j^{i-1} \text{x } a^{i-1})$, where a_j^i is node *j* in layer *I*, and $a^{\text{i}-1}$ is a column vector including all the nodes in the previous layer, where θ_j^{i-1} is a row vector including the weights for all the nodes in layer i−1 for estimating a_j^i.

Figure 1 ANN Model Structure



The first layer is the input data. The second and third layers are hidden layers. The fourth layer is the output layer, which comprises the probabilities of buying each of three insurance products. Back-propagation with random initialization of model parameters is used to train the ANN. For some nodes in the second layer, not all the input data (first layer) are used to determine their values. As shown in Figure 2, expert opinions are used to construct part of the second layer. Four nodes are used to represent the affordability, risk appetite, client satisfaction and new insurance needs determined by selected subsets of input data. Experts' inputs on the weights (θ_j^0 , j = 0 to 4) of input data for the four nodes are used for parameter initialization. The remaining node a_4^1 in the second layer is assumed to be affected by all input data to allow model flexibility.

Figure 2 Heuristic Training for the Second Layer



Some existing customers had already bought two or more products. They are used as the positive examples in the calibration and are the key to predicting the likelihood of buying a second product and its most likely type. The data set was divided randomly into training data and validation data. The calibrated model based on the training data was validated using several practical approaches. For example, the ANN outputs for a customer who has bought a universal life (UL) product are that the customer has a probability of 80%, 30% and 50% to buy a new term life (TL), new UL and long-term care (LTC) product, respectively. Therefore, the recommended product is TL. The 10 most similar customers with two or more products including at least one UL product are sought. If no fewer than 50% (80%/[80%+30%+50%]) of the 10 customers have bought a TL product, the ANN model is considered reasonable. The similarity is measured using Euclidean distance. Because of the large number of input variables, the impact of important variables could be diminished by unimportant variables if equal weights are applied. The four nodes in the second layer (affordability, risk appetite, satisfaction and new insurance needs) are used instead to determine the similarity of customers. Using the validation data, the model has a reasonable rate of 69%.

The model was also validated using a pilot sales project by contacting around 2,000 existing customers with only one insurance purchase in the past. These customers are those with a high chance of purchasing a second product as predicted by the ANN model. The success rate of selling a second product is 4.5% compared to a past average level of 1.3% when the selection of contacted customers was based on qualitative analysis targeting high-net-worth clients. The 4.5% success rate is lower than expected based on the model prediction, with possible reasons including changes in family and financial conditions and having made purchases with other insurers.

Results

The ANN model was used to estimate the most likely product to be bought and the probability of completing the new sale for each existing customer. The customers were then ordered by the probability. Table 1 shows the percentage of customers who will buy a product with a probability higher than a certain value based on the model result. For example, 3% of the existing customers will buy an LTC product with a probability of 50%.

Using these results, existing customers can be contacted with relevant product information. The cost can also be managed by limiting the selling efforts only to customers with a high probability of completing the sale. The model can be further enhanced by estimating the cost and face amount of a product that a customer is likely to accept.

Table 1 ANN Result Summary for a Sample Data Set

	Probability		
Product	70%	50%	30%
TL	5%	11%	17%
UL	1%	3%	4%
LTC	2%	3%	6%

Kailan Shang, FSA, CFA, PRM, SCJP, is co-founder of Swin Solutions Inc. He can be reached at *kailan.shang@swinsolutions.com.*

Machine Reserving: Integrating Machine Learning Into Your Reserve Estimates

Dale Cap, ASA, MAAA

Two hundred years ago a captain may have had only a sounding line and his experience to navigate through uncharted waters. Today a captain has access to many other data sources and tools to aid in his navigation, including paper charts, online charts, engineering surveys, a depth sounder, radar and GPS. These new tools don't make the old tools obsolete, but any mariner would be safer and more accurate in their piloting by employing all the tools at their disposal.

In the same vein, actuaries who solely use traditional reserving techniques, such as triangle-based methods, aren't capitalizing on new technologies. Actuaries should start adopting other techniques such as machine learning (ML). ML is a field of predictive analytics that focuses on ways to automatically learn from the data and improve with experience. It does so by uncovering insights in the data without being told exactly where to look.

ML is the GPS for actuaries. As GPS improved navigation, ML has the potential to greatly enhance our reserves. It is important to note though that ML is not just about running algorithms; it is a process. At a high level this process includes defining the problem, gathering data and engineering features from the data, and building and evaluating the models. As in the actuarial control cycle, it is important to continually monitor results.

Through our research, we have found significant improvements in the prediction of reserves by employing this ML process. Overall we have found a reduction in the standard and worst case errors by 10%. To assist actuaries in testing the value of ML for themselves, this paper will provide an outline of the ML process.

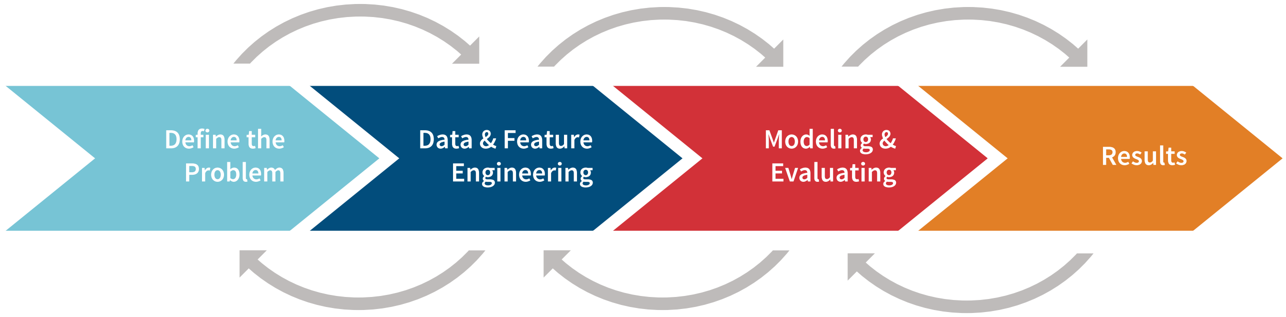
Define the Problem

Similar to the Actuarial Control Cycle, the first step is to define the problem. In our context, we are interested in efficiently calculating the unpaid claims liability (UCL). We want to calculate this quantity in an accurate manner that minimizes the potential variance in the error of our estimate.

Actuaries often use various triangle-based methods such as the Development and the Paid Per Member Per Month (Pd PMPM) to set reserves. These methods in principle attempt to perform pattern recognition on limited information contained within the triangles. Although these methods continue to serve actuaries well, information is being left out that could enhance the overall reserve estimate. To make up for the lack of information used to estimate the reserves, an actuary relies heavily on his or her judgment. Although judgment is invaluable, biases and other elements can come into play, leading to large variances and the need for higher reserve margins.

As described in our prior article[1], the range of reserve estimate error present in company statements pulled from the Washington State Office of the Insurance Commissioner website was −10% to 40%. This

Figure 1 Machine Learning Process



[1] Cap, Coulter, & McCoy, 2015

represents a wide range of error and has significant implications, including an impact to the insurer's rating status, future pricing and forecasting decisions, calculation of premium deficiency reserves, or even unnecessary payout of performance bonuses.

Data and Feature Engineering

Gathering data is something that actuaries are already good at. Leveraging off their expertise along with other subject matter experts will be helpful in identifying all available sources for use. There is often a saying with ML that more data often beat a sophisticated algorithm.

Once the data have been gathered, the actuary will need to engineer the data to improve the model's predictive power. This is referred to as *feature engineering* and can include the transformation, creation, conversion or other edits/additions to the data that will benefit the process. As an example, suppose we were estimating the cost of a house with only two fields: the length and the width of the house. We could help improve the model by feature engineering a new feature called square footage, where we would multiply the length and width.

The gathering and engineering of the data can be a difficult stage to get through, and without the right people on the team, it could lead to a wasted effort. Having domain knowledge on the team enables a more thoughtful consideration of what sources and features are important. In our research we have found many features that have predictive power for reserve setting. The following is a sample list of features that could provide value:

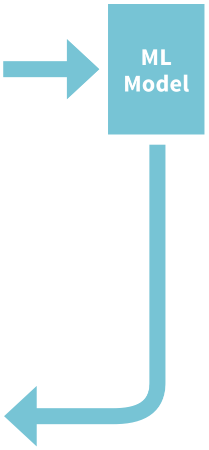
- Seasonality
- Number of workdays
- Check runs in a month
- Large claims
- Inventory
- Inpatient admits/days
- Membership mix by product
- Change in duration
- Cost-sharing components
- Demographics
- Place of service

Modeling and Evaluating

Once the data have been prepared, the user will apply various ML models to the data set. In general, there are two types of data: the training set and the testing set.

Figure 2 Supervised Machine Learning

Incurred Month	Area	Lag Month	Work-days	Check Runs	Inventory	Incurred & Paid	Age/ Gender	Other Features	Actual Incurred
Jan -10	x	x	x	x	x	x	x	x	250.00
Jan -10	x	x	x	x	x	x	x	x	240.00
Jan -10	x	x	x	x	x	x	x	x	265.00
Jan -10	x	x	x	x	x	x	x	x	280.00
Jan -10	x	x	x	x	x	x	x	x	275.00
—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—
—	—	—	—	—	—	—	—	—	—
Dec -11	x	x	x	x	x	x	x	x	335.00
Dec -11	x	x	x	x	x	x	x	x	305.00



Incmo	Area	Lag Month	Work-days	Check Runs	Inventory	Incurred & Paid	Demo-graphic	Other Features	Expected Incurred
Dec -12	x	x	x	x	x	x	x	x	250.00

The training set is the data used to train and cross-validate the model and comprises historical data (in the case of reserving, historical completed data). The testing data on the other hand include only the data from which you wish to derive predictions (for example, the current month's reserve data).

To evaluate the model, a portion of the training set is withheld in order to cross-validate the results. The models that are identified to perform well on the withhold set are then applied to the testing data to create the predictions.

There are many different machine learning models, each of which has its own strengths and weaknesses. Thus there is no one model that works best on all problems.

Results

For our research we used supervised learning techniques classified as regression. We ran various ML models and determined which ones were the most appropriate for the problem based on cross-validation techniques. We then used an ensemble method to blend the various model outputs for an overall prediction. An example of this type of technique can be found in our prior article[2].

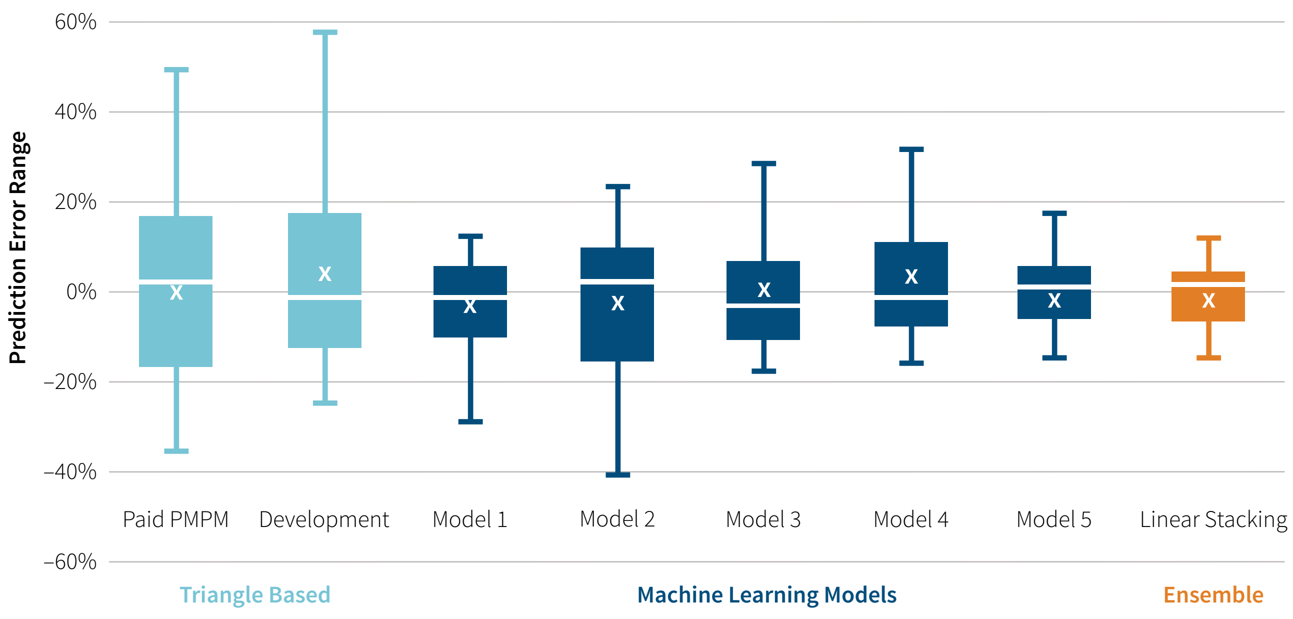
These results were then compared against typical triangle-based methods, where we tested the percentage range of UCL error over 24 different reserving months. Overall we found that ML added significant value in reserve setting, and we highly encourage reserving teams to explore this process for themselves.

Conclusion

Predictive analytics are not new to actuaries. Methods like these are fairly common on the casualty side and have recently become more popular within health care for predicting fraud, readmission and other aspects. However, those within health care are often being led by data science teams, who continue to fill a larger analytics role within the health space. It is only a matter of time before these techniques become standard to reserving. The question is: Who will fill this role? Will actuaries stay at the helm, or will we transfer some of our functions to data science teams?

We hope that the process outlined above will provide some guidance and at least prepare actuaries for their first steps in this space.

Figure 3 UCL Prediction Error Range (Box and Whisker Plot)



[2] Cap, Coulter, & McCoy, 2015

Appendix

	Triangle Based		Machine Learning Models					Ensemble
Statistics	Paid PMPM	Development	Model 1	Model 2	Model 3	Model 4	Model 5	Linear Stacking
Mean Error	1.1%	4.5%	–2.6%	–1.6%	–0.5%	2.5%	–1.1%	–0.9%
Standard Error	22.1%	21.7%	9.8%	16.7%	13.5%	14.0%	9.1%	8.8%
Kurtosis	–43.8%	34.0%	78.7%	–18.9%	–17.6%	–38.4%	181.5%	114.4%
Skew	24.1%	100.8%	–65.1%	–58.1%	87.0%	83.9%	–75.3%	–95.1%
Cumulative Error	431.2%	390.1%	184.7%	329.0%	264.6%	264.4%	162.6%	157.6%
Worst Error	49.1%	57.7%	29.0%	41.1%	28.5%	31.4%	27.0%	25.3%

Dale Cap, ASA, MAAA, is an actuary with a passion for analytics. He can be reached at *dale_j_cap@outlook.com.*

Variable Selection Using Parallel Random Forest for Mortality Prediction in Highly Imbalanced Data

Mahmoud Shehadeh,
Rebecca Kokes, FSA, MAAA and
Guizhou Hu, M.D., Ph.D.

In the last few years, the industry has started moving away from traditional actuarial methods toward more statistically sound methodologies including parametric and nonparametric approaches such as generalized linear models and machine learning algorithms to better assess risks. Using such techniques and utilizing the full potential of underwriting data can improve mortality prediction greatly. In the life insurance industry, actuaries and underwriters need to process a substantial amount of data in order to assess the mortality of applicants as quickly and accurately as possible. Examples of such data include, but are not limited to, demographic, paramedic and medical history. The data can be numerical (age, face amount), categorical (gender, smoking status) and even free text. However, using such data is not always straightforward.

In this article we present a practical example of the intersection of life insurance, machine learning and big data technology. The aim is to use a random forest (RF) algorithm to identify the most important predictors (from a set of hundreds of variables) that can be used in mortality prediction, that is, to reduce number of variables from hundreds to dozens while retaining the predictive power. In addition, the use of parallel computing to speed the process and stratified sampling to deal with highly imbalanced data is discussed.

The medical history of applicants includes hundreds, if not thousands, of unique keywords that have the potential to increase the accuracy of risk assessment. Typically, one can include this information as predictors in regression analysis following two approaches: first, by manually selecting the most important terms, and, second, by grouping medical terms by disease type. Note that both approaches aim to reduce the number of predictors to a manageable size, since building a closed-form regression equation using hundreds or thousands of predictors may not seem practical, especially when the number of observations is small. Furthermore, in both approaches one can lose information either by ignoring rare important variables or by averaging out the severity of these medical terms. In addition, medical and underwriting expertise are needed for selecting or grouping the medical terms, and thus the process can be time-consuming and expensive.

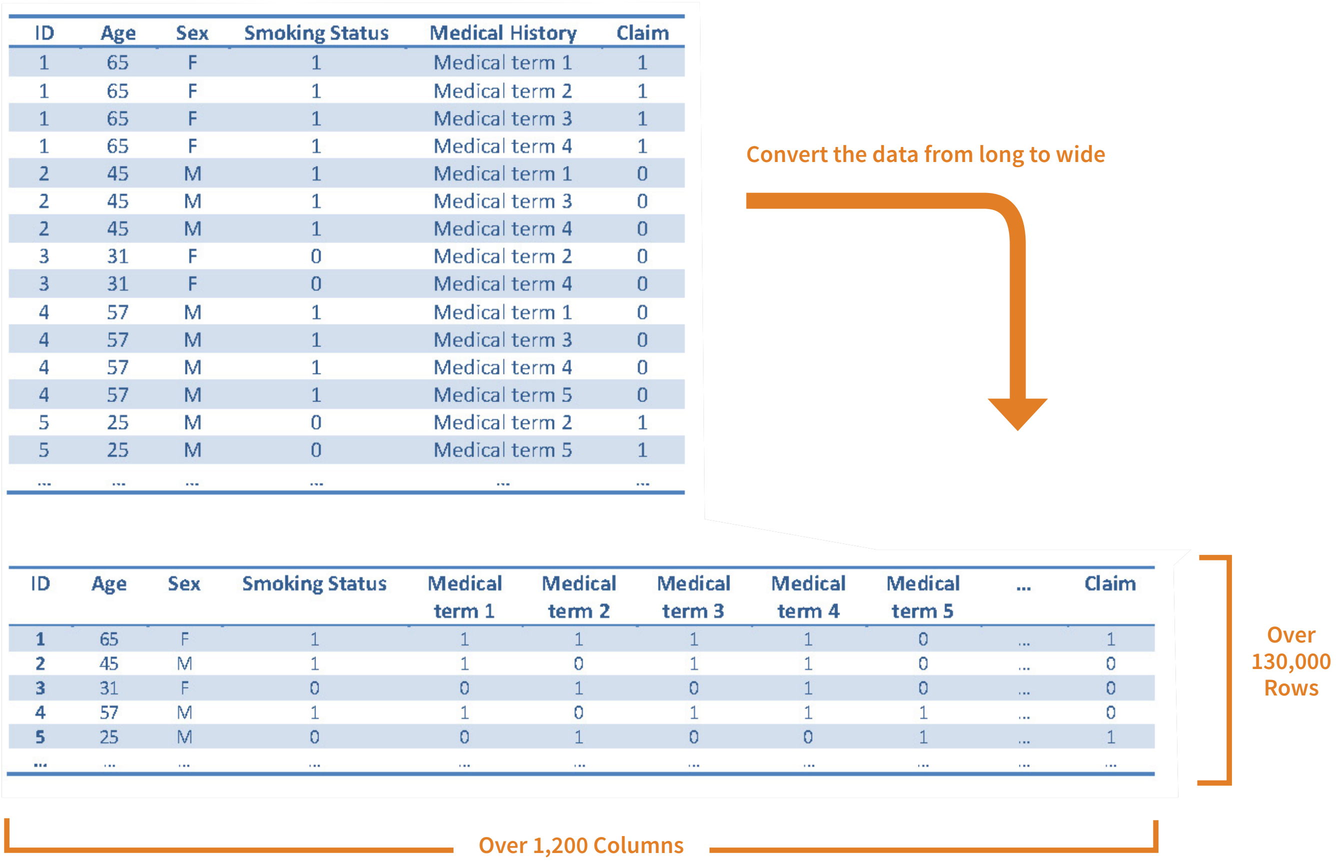
In this article we use historical underwriting data of more than 130,000 applications along with their demographic (age, sex and smoking status) and medical history information. Medical history includes more than 1,200 unique medical and disease terminologies (e.g., CAD, hypertension, alcohol, anemia, Parkinson's disease). Furthermore, the data set is considered highly imbalanced since the number of claims represents less than 5% of the total population. The first step in the process is to extract the historical underwriting data from different databases and then convert it from long to wide, binary and sparse matrix so that each medical terminology can be used as a predictor in the analysis. After converting, the final data have a dimension of more than 130,000 rows and more than 1,200 columns (see Figure 1).

Method

RF, introduced by Breiman in 2001[1], is a machine learning approach that can be used for regression and classification problems. RF is an ensemble algorithm of unpruned decision trees in which each tree is built from the learning data using a randomly selected sample with replacement (a bootstrap sample) and a number of randomly selected predictors from the set of all features. The RF technique has become popular in the machine learning literature because of its smaller prediction variance and ability to deal with a large number of predictors while capturing the interaction structure in the data. Breiman's RF algorithm can be implemented in R using the

[1] Leo Breiman, "Random Forests," *Machine Learning* 45, no. 1 (2001): 5–32.

Figure 1 Converting the Data from Long to Wide, Binary and Sparse Matrix Structure



ID	Age	Sex	Smoking Status	Medical History	Claim
1	65	F	1	Medical term 1	1
1	65	F	1	Medical term 2	1
1	65	F	1	Medical term 3	1
1	65	F	1	Medical term 4	1
2	45	M	1	Medical term 1	0
2	45	M	1	Medical term 3	0
2	45	M	1	Medical term 4	0
3	31	F	0	Medical term 2	0
3	31	F	0	Medical term 4	0
4	57	M	1	Medical term 1	0
4	57	M	1	Medical term 3	0
4	57	M	1	Medical term 4	0
4	57	M	1	Medical term 5	0
5	25	M	0	Medical term 2	1
5	25	M	0	Medical term 5	1
...	...	...	...	...	...

Convert the data from long to wide

ID	Age	Sex	Smoking Status	Medical term 1	Medical term 2	Medical term 3	Medical term 4	Medical term 5	...	Claim
1	65	F	1	1	1	1	1	0	...	1
2	45	M	1	1	0	1	1	0	...	0
3	31	F	0	0	1	0	1	0	...	0
4	57	M	1	1	0	1	1	1	...	0
5	25	M	0	0	1	0	0	1	...	1
...	...	...	...	...	...	...	...	...	...	...

Over 130,000 Rows

Over 1,200 Columns

randomForest[2] package, which is available from the CRAN website. Another great advantage of the supervised RF is that it can measure the importance of each predictor in classifying the response variable appropriately.

Having converted the data into an appropriate shape, we used a training set (75% of the data) to build a RF model. The claim variable (binary) was used as a response, while age, sex, smoking status and medical terminologies were used as predictors. During the process, we ran into two issues: slow run time and poor predictive performance when using the test set (the remaining 25% of the data).

In regard to the first issue, since the number of computations that need to be carried out is tremendous because of the large number of features and relatively large number of observations, the run time using a single-core CPU (the default option in R) was estimated to take more than a week to complete. To reduce the run time, a parallel RF using eight-core CPUs was built instead, reducing the run time to less than 48 hours. Utilizing parallel computing in R was possible by using packages such as snowfall[3] and rlecuyer.[4] In Figure 2 we provide a schematic comparison of sequential versus parallel RFs.

[2] Andy Liaw and Matthew Wiener, “Classification and Regression by randomForest,” *R News* 2, no. 3 (2000): 18–22.

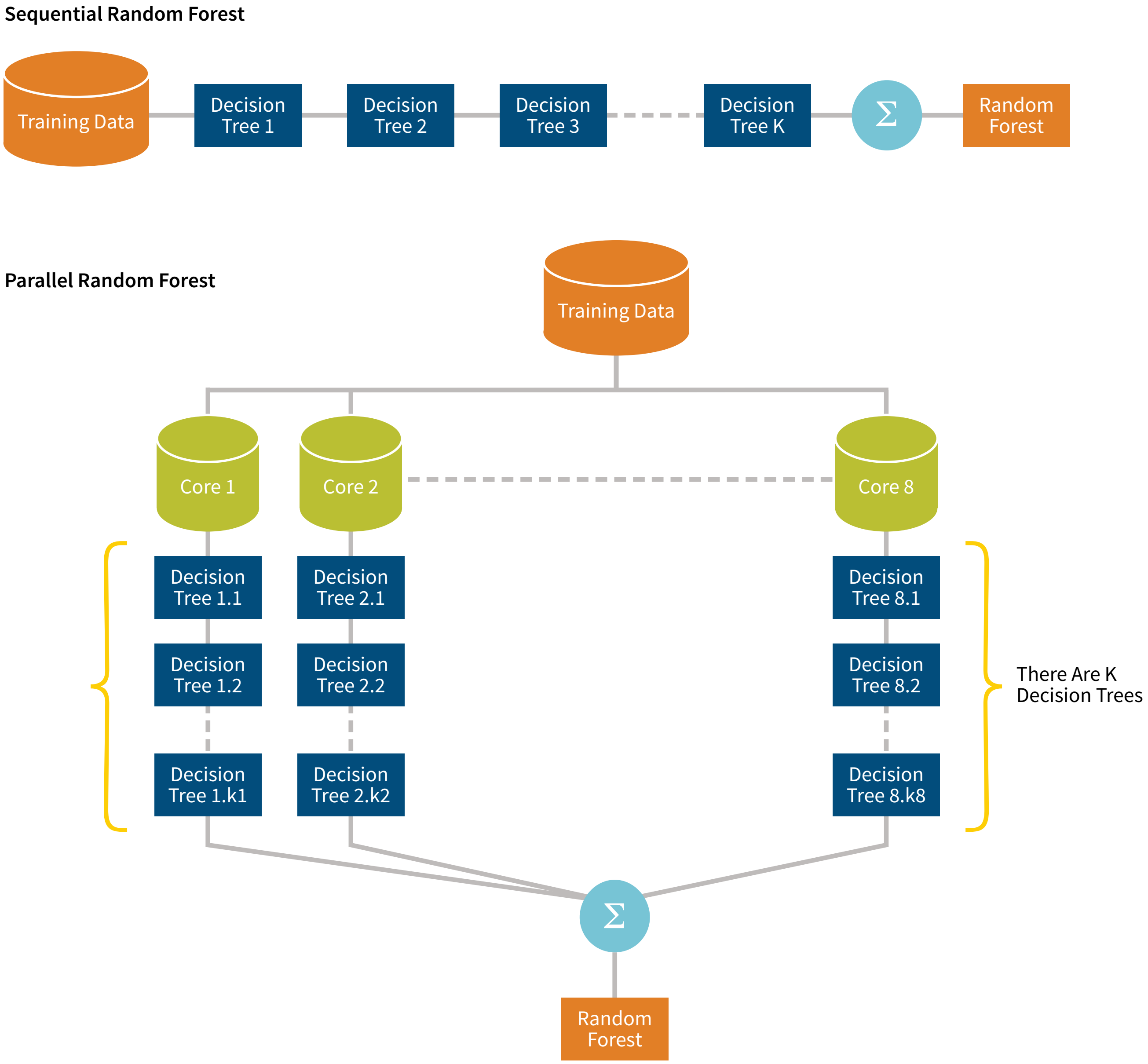
[3] Jochen Knaus, “Easier Cluster Computing (Based on Snow),” R Package Version 1.84-6.1 (2015), http://CRAN.R-project.org/package=snowfall.

[4] H. Sevcikova and T. Rossini, “Rlecuyer: R Interface to RNG with Multiple Streams,” R Package Version 0.3-4 (2015), http://CRAN.R-project.org/package=rlecuyer.

Results

Upon successfully building the model, we tested its predictive performance using the test data, and the results were poor. The model was able to identify fewer than 1% of the claims because of the imbalanced structure of the data, where the number of no-claim observations excessively exceeds the number of observations with claims. As we explained earlier, the RF model is built using bootstrap samples, and in the case of imbalanced data there is a high chance that the bootstrap samples contain few (or even none) of the claim observations. Fortunately, the randomForest package has the option of using stratified bootstrap samples (random samples with replacement from claim and nonclaim observations). Utilizing the stratified bootstrap samples option, we ran the model again, and we were able to correctly identify 67% of the claims. As expected, age, sex and smoking status, along with other major impairments, were found to be highly predictive.

Figure 2 Sequential vs. Parallel Random Forests Consist of *K* Decision Trees



The next goal in the process is to reduce number of predictors from more than 1,200 to a manageable size while retaining the predictive power. To do so, we relied on the results of the second model to arbitrarily select the top 34 most important predictors. Then, using only these predictors and the same training set as before, we built a new RF model that was able to correctly predict 65% of the claims.

Conclusion

In this article we introduced and applied the RF classifier to predict claim observations using a hold-out sample. Issues related to a slow run time and poor predictive performance were successfully avoided by utilizing the power of parallel computing and the stratified bootstrap samples. In addition, we were able to reduce the number of predictors from more than 1,200 to 34 while still retaining most of the predictive power. Ultimately the results can be used to support underwriting decisions. In future work we will try different ensemble approaches and other stratified sampling techniques.

Mahmoud Shehadeh is an actuarial student at Gen Re in Stamford, CT, with two years of actuarial experience. He can be reached at *mahmoud.shehadeh@genre.com*.

Rebecca Kokes, FSA, MAAA, is an actuary at Gen Re in Stamford, CT. She can be reached at *rebecca.kokes@genre.com*.

Guizhou Hu, M.D., Ph.D., is vice president, chief decision analytics at Gen Re in Stamford, CT. He can be reached at *Guizhou.Hu@genre.com*.