MWSUG 2019 – Paper – IN- 083

Classifying Risk in Life Insurance using Predictive Analytics

Sai Gopi Krishna Govindarajula, Oklahoma State University

ABSTRACT

Ever wonder how many companies offer life insurance? There are more than 600 companies in the US alone offering life insurance policies. Insurance companies perform an underwriting process to assess the risk of life insurance applicants and then price the policies if approved. Those underwriters gather extensive information about applicants, which include extensive health histories, to classify risk profiles. The process of collecting existing data for the risk assessment, completing and obtaining any required patient health exams, and validating all the information often takes several weeks to months. In this fast-paced world, customers are prone to lose interest in finalizing policies from companies who take a prolonged time to evaluate an application.

With the advent of data analytics, the underwriting process can be streamlined and completed much faster. The intention of this project was to build predictive models based on past customer history and to recommend the most appropriate model to assess risk resulting in better underwriting practices and customer retention. A real-time data set having around 140 variables, which included a combination of categorical and continuous variables, was analyzed using SAS Enterprise Miner™ and Tableau® for predictive modeling and data visualization, respectively. Statistical models such as Logistic regression and machine learning algorithms such as Neural network, Decision Tree and Ensemble model were built and compared to assess risk. Results revealed that the Neural Network has shown the highest performance with a ROC index of 0.868.

INTRODUCTION

Insurance is often defined as the process of transferring risk. In the context of life insurance, where the claims usually tend to be in high amounts, it is important for an insurance firm to rightly assess the risk before issuing a policy keeping in mind the financial health of the company in long term. An insurer should take a judicious call while accepting applicants who may make huge claims as such customers could pose a threat to the profits made by the company. Hence, the underwriting process, in which an applicant's risk is evaluated, is a vital step of processing an insurance application.

Underwriting in insurance firms consumes great manual efforts to collect, process, and analyze data. Applicants must go through medical examinations and then submit corresponding tests' results. Underwriters assess this documentation along with other medical histories that have been gathered, to estimate the applicant's life expectancy, with help of mortality rate tables and actuarial formulae. This takes several weeks to months of time and are costly. Big data & analytics are starting to play a major role in data collection by leveraging the latest technology available in the market.

Auto insurance companies these days use telematics to track driving behaviors, thereby charging appropriate premiums based on past claim histories and how safe they drive. For example, the time taken to apply brakes or how sharp a driver takes a turn could give an insight into the driving behavior of the applicant. On the other hand, Life insurance companies are still in early

stages of leveraging analytics, but they took off in the right direction. For example, activity trackers such as Apple watch or Fitbit can give us some insight related to a person's lifestyle. Firms such as John Hancock offer discounts to their customers on the charged premiums based on their efforts to improve any unhealthy practices. After collecting the data needed, the next challenge would be to leverage this data and classify the risk level of the customer. The purpose of this research is to apply predictive modeling techniques and recommend an appropriate model to assess risk, streamline underwriting process and improve decision making.

EXPLORATORY DATA ANALYSIS

There are 59,381 observations in the data set obtained from Kaggle.com. The dataset was released by Prudential Life Insurance company and most of the variable names have been masked due to privacy concerns. The data involves 115 categorical variables and 11 continuous variables apart from a 'Response' variable which denotes the level of risk associated with a person's chances of claiming his/her life insurance, to get a life insurance quote. Below table gives a list of all variables and their descriptions:

Variable	Description
Id	A unique identifier associated with an application.
Product_Info_1-7	A set of normalized variables relating to the product applied for
Ins_Age	Normalized age of applicant
Ht	Normalized height of applicant
Wt	Normalized weight of applicant
BMI	Normalized BMI of applicant
Employment_Info_1-6	A set of normalized variables relating to the employment history of the applicant.
InsuredInfo_1-6	A set of normalized variables providing information about the applicant.
Insurance_History_1-9	A set of normalized variables relating to the insurance history of the applicant.
Family_Hist_1-5	A set of normalized variables relating to the family history of the applicant.
Medical_History_1-41	A set of normalized variables relating to the medical history of the applicant.
Medical_Keyword_1-48	A set of dummy variables relating to the presence of/absence of a medical keyword being associated with the application.
Response	This is the target variable, an ordinal variable relating to the final decision associated with an application

Table 1. Data set description

DATA PARTITIONING:

The data was divided into two parts using Data Partition node in SAS Enterprise Miner. 70% was used as training data and 30% was used for validation.

HANDLING MISSING VALUES:

Missing values often create a lot of problems in any analysis. In this research, variables which have more than 30% of missing values have been dropped using impute node. After trying several methods for rest of the variables, count method has been used to impute categorical and median has been used to impute continuous variables.

HANDLING VARIABLE TRANSFORMATIONS:

The variables in the data have already been standardized and look approximately normal. Hence there was no need to transform variables for normality.

DIMENSIONALITY REDUCTION:

The response variable is ordinal and has 8 levels of risk. To make it more interpretable, several levels have been combined. Risk levels 1 to 5 have been combined into one group, risk levels 6 and 7 into one group and risk level 8 has its own group. This step also helped to get an approximately balanced data set and so our modeling would be devoid of any bias that is introduced by unbalanced data.

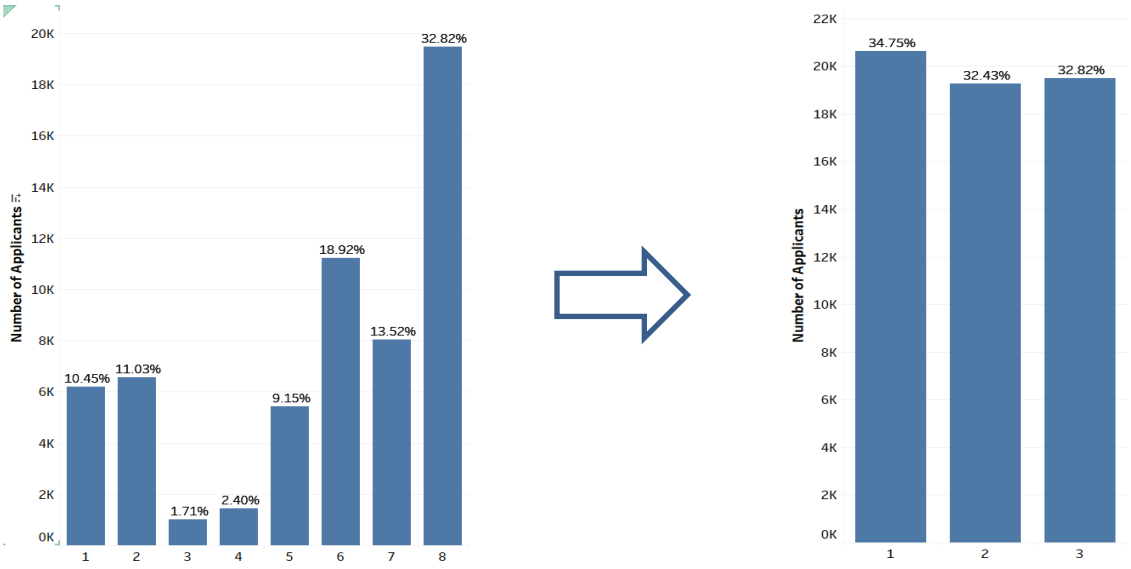

Figure 1. Response variable distribution before and after grouping

Efficient modeling requires a reduction in the number of variables by handling the curse of dimensionality. Feature extraction technique was used, to sum up values of all Medical_Keyword 1-48 variables to get a score and this new variable was used in modeling.

Another Feature extraction technique, PCA (Principal Components Analysis) derives new features from existing ones to create better attributes whereas CFS (Correlation Based Feature Extraction) selects the best attributes as they are. The resulting new features are difficult to

explain in PCA, as it's difficult to interpret each principal component that was created, whereas in CFS it's easier to understand as the features are not modified. The data set has several variable names masked and hence it would be difficult to interpret the important variables anyway, hence PCA was used here for analysis. Performing this also helps in reducing the dimensionality by reducing the number of variables. A total of 50 principal components were used for this research and they captured around 52% of total variance in prior variables.

MODELING:

After pre-processing data using the techniques discussed above, predictive modeling was performed on the data. Different machine learning algorithms were implemented namely Ordinal Logistic Regression, Neural Network, Decision tree, and Ensemble models. Best model selection is a crucial part of any statistical analysis and it should bring coherence to the central question in the discussion. There are several evaluation metrics to assess models like Misclassification Rate, Mean Squared Error (MSE), Akaike Information Criteria (AIC), Area Under Curve - Receiver Operating Characteristic (AUC-ROC Index) etc. The Validation data's ROC index was used as the selection criteria to choose the best model in this analysis. The reasons for the same are discussed below.

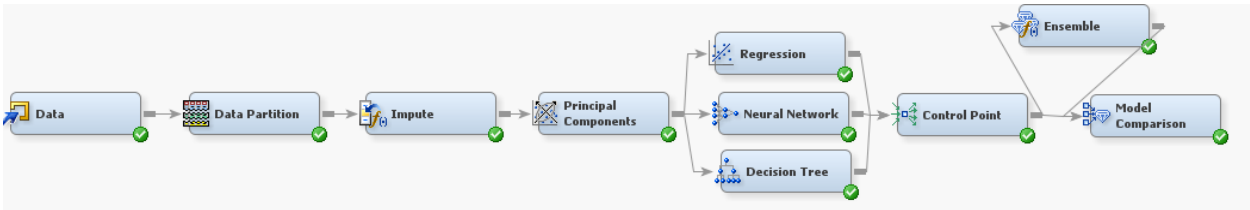

Figure 2. Node diagram of predictive modeling techniques

MSE is the average squared difference between estimated values and what is estimated. Hence it is more suitable for a continuous response variable. AIC is a relative measure of different models and compares how well the models fit the data by picking the one with maximum likelihood and also by penalizing extra variables. Although it provides a relative criterion to choose the best model, it cannot suggest whether the model is useful in an absolute sense. Misclassification rate tells how often the model is wrong and as we would want to classify risk as accurately as possible for the research question here, it is a valuable metric. Similarly, the AUC-ROC curve is a performance measurement that distinguishes between classes at different threshold settings. As the AUC score gets higher, the model gets better at rightly predicting 0s as 0s and 1s as 1s. For this research problem, the AUC-ROC index can give more information as we would like to choose the classifier case by case. For example, the risk classifier may need more focus on the probability of predicting a risky applicant as risky, rather than the probability of predicting a non-risky applicant as non-risky. The insurance firm cannot afford to classify a risky customer as non-risky and bear the brunt for the losses incurred. Hence the selection criterion for the best model has been chosen as the AUC-ROC index. Neural Network has the best AUC among all models with a value of 0.868. It is followed by the Ensemble model with the AUC-ROX index of 0.858. The table below compares all the other evaluation metrics from different models:

Model Node	Model Description	Target Variable	Selection Criterion: Valid: Roc Index	Valid: Average Squared Error	Valid: Root Average Squared Error	Valid: Misclassification Rate	Train: Akaike's Information Criterion	Train: Total Degrees of Freedom
Neural	Neural Network	Response	0.868	0.167576	0.409361	0.398496	70813.44	83128
Ensmbl	Ensemble	Response	0.858	0.175169	0.418531	0.413145	.	.
Reg	Regression	Response	0.849	0.180796	0.425201	0.432845	75384.81	83128
Tree	Decision Tree	Response	0.724	0.20225	0.449723	0.501768	.	83128

Figure 3. Output comparing evaluation metrics from different models

CONCLUSION

This research paper provides useful implications by building predictive models to assess risk in life insurance. This helps insurance firms to avoid strenuous manual efforts and long waiting periods to assess risk and offer policies to customers. Data analytics can help enhance the insurance business with better accuracy and in less time, which means increased customer satisfaction and loyalty.

The research demonstrated imputation methods used for missing values, and dimensionality reduction techniques to help retain only the variables that explain the target variable. Machine learning algorithms like ANN, Decision Tree, Ordinal Logistic Regression and Ensemble models were implemented. AUC-ROC index was chosen as the selection criterion for choosing the best model and ANN has the best index with a value of 0.868.

FUTURE IMPLICATIONS

Extensive research has not yet been done to identify factors such as genetic profiling and if it affects the health of a person. This could really play an important role and alter how the risk is assessed in the Life Insurance industry. If a person has unhealthy habits like smoking or drinking, it is agreeable that he will be a risky customer to the insurance company and so will be charged a high premium. But what if the ill-health someone is facing is due to his/her genetic make-up and not their lifestyle itself. Those with such unfortunate genetic drawbacks end up paying a fortune for health care for no mistake of theirs. But is it fair? Big data can help us identify genetic profiling and help make a distinction between unhealthy lifestyles by choice and genetic factors beyond our control.

Segmentation techniques can be explored to group applicants with a similar medical history or similar employment history. This can contribute to customized insurance plans for different groups and also can increase transparency to customers in the coverages being offered and the premium they pay.

REFERENCES

1) Bernard Marr. "How Big Data Is Changing Insurance Forever." Accessed December 16, 2015. https://www.forbes.com/sites/bernardmarr/2015/12/16/how-big-data-is-changing-the-insurance-industry-forever/#3bc15c0f289b
2) Stephen Abrokwah. "Predictive Analytics in the Life Insurance Process". Accessed December 2015. https://theactuarymagazine.org/predictive-analytics-in-the-life-insurance-process/
3) Mila Araujo. "What is an Insurance Premium (and How Does It Work)?". Accessed March 2019. https://www.thebalance.com/understanding-what-is-an-insurance-premium-4155239
4) Boodhun, N. & Jayabalan, M. Complex Intell. Syst. (2018) 4: 145. https://doi.org/10.1007/s40747-018-0072-1
5) Deloitte Consulting LLP. "Predictive Modeling for Life Insurance." Accessed April 2010. https://www.soa.org/globalassets/assets/files/research/projects/research-pred-mod-life-batty.pdf

ACKNOWLEDGMENTS

I am thankful to my professors Dr. Goutam Chakraborty and Dr. Miriam McGaugh for their support and guidance throughout the time of writing this research paper. I am also thankful to my mentor Praveen Kotekal for his inputs in this research.

CONTACT INFORMATION

Sai Gopi Krishna Govindarajula

Business Analytics Graduate Student (Class of 2020)

Oklahoma State University, Stillwater

Mobile: +1 (405) 780 - 2089

Email: sagovin@okstate.edu

LinkedIn: https://www.linkedin.com/in/sgopikg/