What is your life worth: A predictive Analytics Approach to Assessing life Insurance Risks.

Authors

Modupe Ajala

University of Massachusetts Lowell

Modupe_ajala@student.uml.edu
Masters of science in Data Analytics

Faculty Advisor

***Faculty Advisor Name**: Dr. Asil Oztekin*

***Affiliation:** University of Massachusetts Lowell*

***Department:** Operations & Information Systems, Manning School of Business*

***Email address:** Asil_Oztekin@uml.edu*

Summary

Getting life insurance is a long and tiring process which requires customers to take numerous medical exams that can take up to 30 days. The process turns people off especially in this one-click shopping world with on-demand everything. This study will be using machine learning-based data analytics to accurately predict risk classifications of customers, thereby making the process of owning life insurance shorter and less labor intensive while maintaining the customers privacy boundaries.

1) Problem and Motivation

Studies have shown that only 40% of US citizens own life insurance, millennials account for only 1 out of 5 of the 40%. This has to do with the long process of owning life insurance. This data analysis is going to look at how to correctly identify and classify customer's risk levels with the most important unique identifiers (such as: age, height, medical history) thereby speeding up the process of obtaining life insurance. This will greatly impact public perception of the industry and entice more people to get life insurance.

2) Approach

This study will adopt the CRISP-DM (Cross-Industry Standard Process for Data Mining) approach, which is used to accurately classify risks associated with life insurance quotes. CRISP-DM process consists of 6 phases, including: Business Understanding, Data Understanding, Data Preparation, Model building, Testing & Evaluation and finally Deployment. This study analyzes each step in a consecutive order. A strategic approach to understand the business motivation, data exploration and preprocessing, build several models, test the accuracy and get useful information from it.

3) Datasets

The dataset chosen to be analyzed is called "Life insurance assessment" which consists of information about life insurance applicants. It was compiled by Prudential, one of the largest issuers of life insurance in the USA, for a study on how to accurately classify risk using a more automated approach. The dataset consists of 59,400 observations and 128 variables. It has a combination of normalized, continuous and discrete variables. After pre-processing 59,381 rows and 66 columns were left. It was downloaded from Kaggle.com and it's an open dataset.

4) Tools & Analytics

A combination of tools and methodologies is used in this study. Excel is used for data preparation, it treated the missing values by filling in variables with less than 10% of missing values and removing variables with over 10% missing values.

Methods & algorithms: This study uses 3 algorithms for analysis which includes: Classification and Regression tree (CART), C5.0 Decision Tree, and k-Nearest Neighbor (k-NN).

IBM SPSS Modeler 18.1 is user-friendly and very popular, so it is used to build the models and perform testing.

5) Results

The dataset is used to train 3 different models, each model was subject to a 10-fold cross validation routine and then evaluated using a confusion matrix, other data such as predictor variable sensitivity was also extracted from the models. The C5 gave the best performance with the highest accuracy rate of 71.8%, The overall accuracy of the k-NN model was 54.02% and finally the overall accuracy of the CART model in predicting the risk classification was calculated to be 48.07%.

Risk Classification:	1	2	3	4	5	6	7	8
Accuracy:	0.718024284							
Precision:	0.637023	0.625611	0.429418	0.478291	0.626473	0.697231	0.547776	0.91508
Sensitivity:	0.705567	0.702846	0.64636	0.687815	0.711776	0.699348	0.718934	0.73746
Specificity:	0.958105	0.954191	0.990155	0.987241	0.962839	0.929413	0.93185	0.95298
T-n	51524	51096	58130	57643	52571	44781	49635	33543

Results Evaluation of the C5 Model.

Risk Classification:	1	2	3	4	5	6	7	8
Accuracy:	0.480743							
Precision:	0.037699	0.139499	0	0	0.305781	0.531292	0.399527	0.849864
Sensitivity:	0.356707	0.50976	0	0	0.477299	0.354415	0.342921	0.607571
Specificity:	0.898289	0.902098	0.982941	0.975952	0.932541	0.87624	0.903656	0.908904
T-n	52752	51950	58368	57953	52130	37277	45209	29194

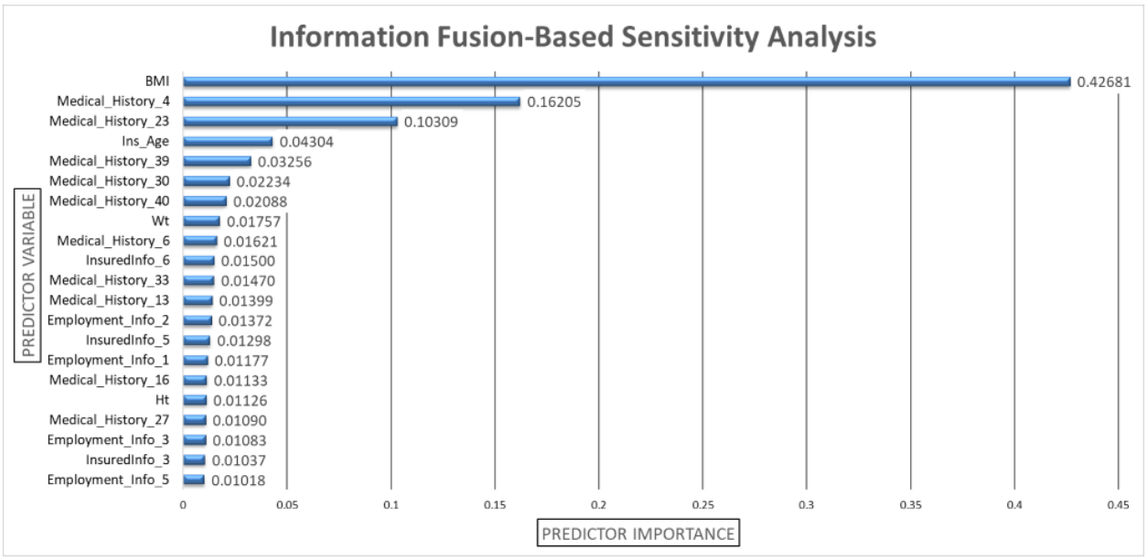
CART Model: Results Evaluation

Risk Classification:	1	2	3	4	5	6	7	8
Accuracy:	0.540172783							
Precision:	0.255679	0.271368	0.113524	0.161765	0.235272	0.549809	0.375109	0.918467
Sensitivity:	0.656599	0.588742	0.435606	0.473361	0.553966	0.482953	0.483229	0.561727
Specificity:	0.918896	0.915296	0.98481	0.979675	0.927217	0.891464	0.905626	0.94225
T-n	52344	51587	58219	57696	52920	41536	48134	25926

k-NN Model: Result Evaluation

Sensitivity analysis

Information fusion-based sensitivity analysis is a unbiased way to rank the predictor variables across multiple models by a weighting method. For this study the information fusion-based sensitivity analysis consisted of gathering the individual model predictor variable importance and applying a weighting scheme of using the individual model accuracy that was calculated from the confusion matrix.



6) Contributions and Uniqueness

The models adopted in this study can predict risk classification and identify the most significant risk factors in classifying life insurance risk. With this knowledge, life insurance companies can correctly classify customers into their appropriate risk status. This will help them quote the right premium and would reduce “Risk pooling” (when life insurance companies do not charge different premium prices to their customers based upon the risk classification). It will also help reducing the lengthy process of getting life insurance policy. This will in turn, impact public perception of the industry and entice more people to get life insurance.

7) Appendix: Visualization / analytics summary

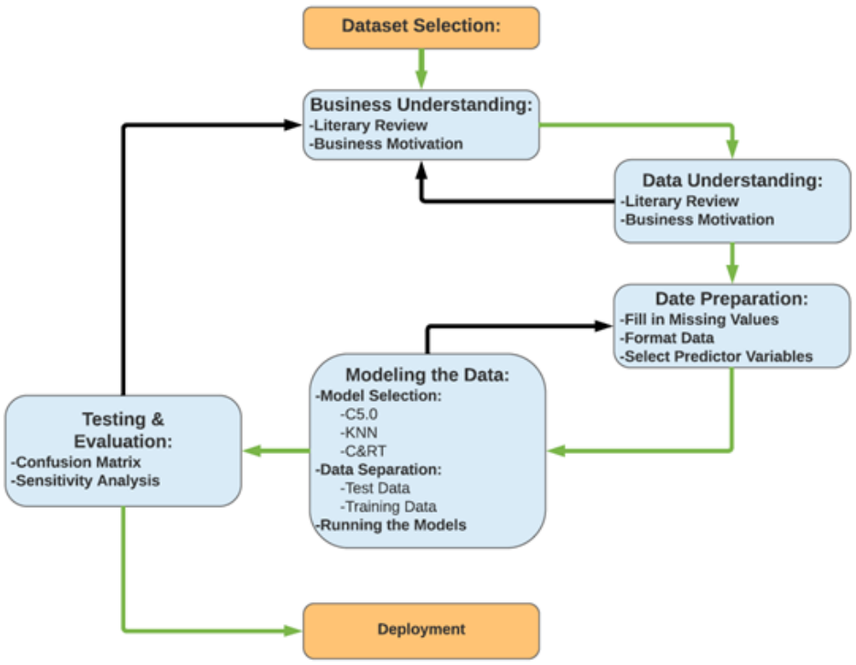


Figure 1. CRISP-DM workflow used for predictive analysis of life insurance risk

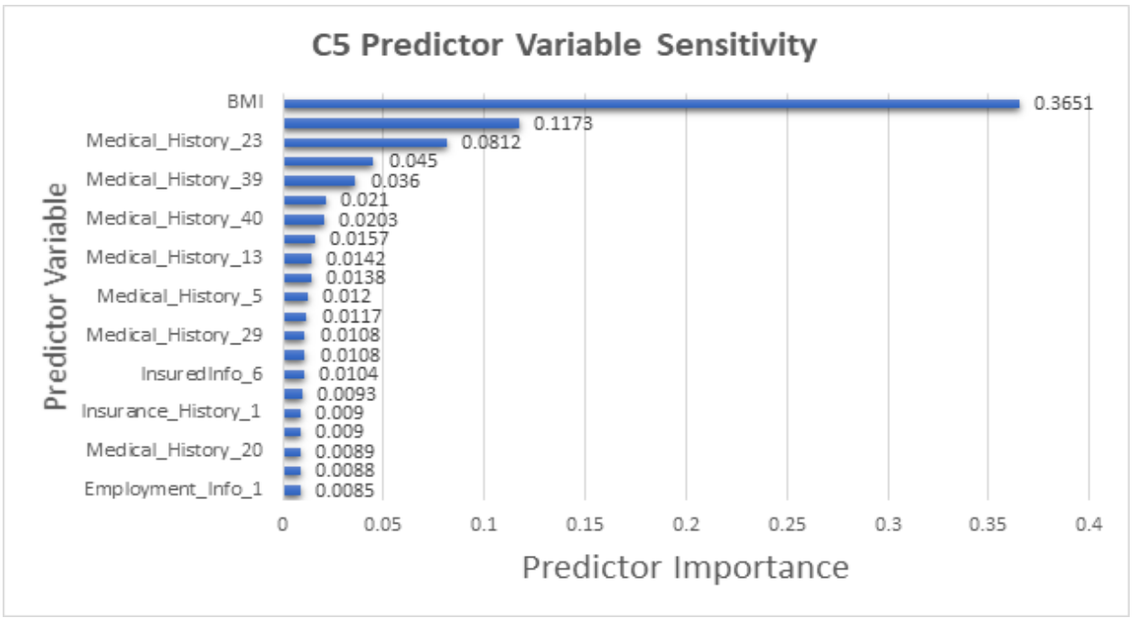


Figure 2. C5 Model: Predictor Variable Sensitivity

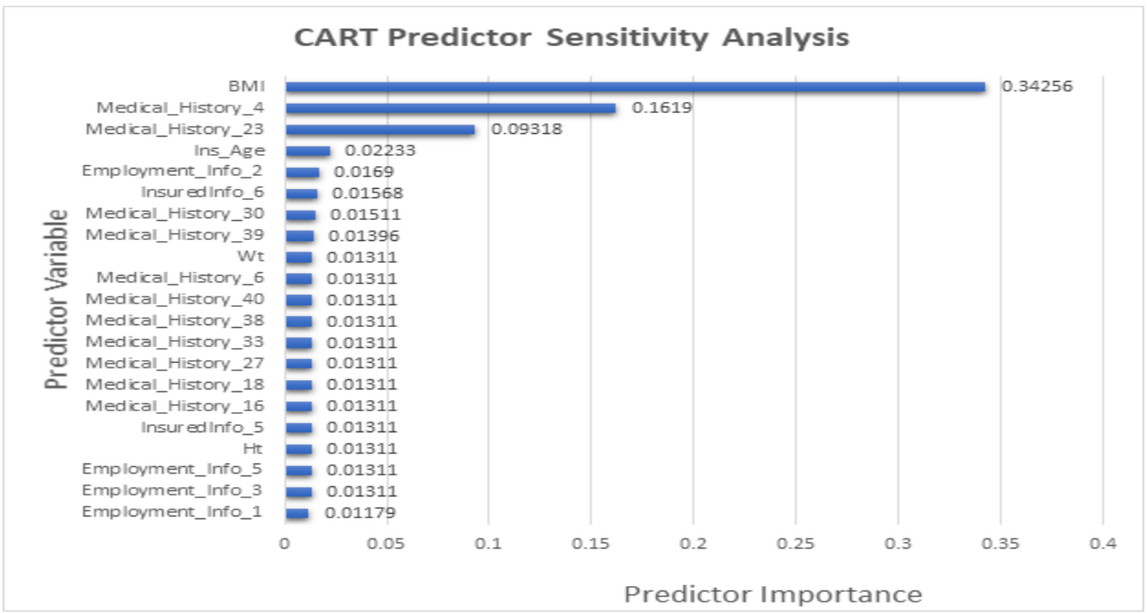


Figure 3. CART Model: Predictor Variable Sensitivity

8) References:

1. Hedengren, David & Stratmann, Thomas (2016). Is there Adverse Selection in Life Insurance Markets? *Economic Inquiry, Vol. 54, No. 1, pp. 450-463*
2. Blackman, Irv (2017). You maybe insurable after all… *Contractor Magazine, July 2017, pp. 24*
3. Hao, Minglie, Macdonald, Angus S., Tapadar, Pradip, & Thomas, R. Guy (2017) Insurance loss coverage and demand elasticities. *Insurance: Mathematics and Economics, 2018, Vol. 79, pp. 15-25*
4. Bajtelsmit, Vickie L., Villupuram, Sriram V., & Wang, Tianyang (2015) Life Insurer Cost of Equity with Asymmetric Risk Factors. *The Financial Review, 2015, Vol. 50, pp. 435-457*
5. Thomas, R. Guy (2008). Loss Coverage as a Public Policy Objective for Risk Classification Schemes. *The Journal of Risk and Insurance, 2008, Vol. 75, No. 4, pp. 997-1018*
6. Wuppermann, Amelie C. (2017). Private Information in Life Insurance, Annuity, and Health Insurance Markets*. *The Scandinavian Journal of Economics, 2017, Vol. 119, No. 4, pp. 855-881*
7. Rishel, Rod (2015). Life Insurance Outlook: New approaches for changing realities. *National Underwriter Life & Health, December 2015, pp. 38-42*

8. Wang, Hsin Chung, Yue, Jack C., Tsai, Ti-Hsuan (2016). Martial Status as a Risk Factor in Life Insurance: An Empirical Study in Taiwan. *Astin Bulletin, 2016, Vol. 46, No. 2, pp. 487-505*
9. Hackert, A., Brookman, J (2013). My life in numbers: an insurance decision. *Journal of Critical Incidents. Oct2013, Vol. 6, p71-74. 4p*
10. Schuman, G (2015). The devil is in the details: establishing an insured's intent to deceive in life and health insurance recession cases *FDCC Quarterly.Winter2015, Vol. 64 Issue 2, p84-113. 30p.*
11. Bajtelsmit, V., Villupuram S, Wang T (2015). Life insurer cost of equity with asymmetric risk factors. *Financial Review. Aug2015, Vol. 50 Issue 3, p435-457. 23p.*
12. Spierdijk, L., Koning, R (2014). Estimating outstanding claim liabilities: the role of unobserved risk factors. *Journal of Risk & Insurance. Dec2014, Vol. 81 Issue 4, p803-830. 28p.*
13. Gatzert, N., Wesker, H (2014). Mortality risk and its effect on shortfall and risk management in life insurance. *Journal of Risk & Insurance. Mar2014, Vol. 81 Issue 1, p57-90. 34p.*
14. Kaggle.com. (2015). Prudential life insurance assessment. Retrieved from https://www.kaggle.com
15. Nerdwallet. (2017) How life insurance works. Retrieved from https://www.nerdwallet.com/blog/insurance/how-does-life-insurance-work/
16. Sharda, Ramesh, Delen, Dursun, Turban, Efraim (2015). Business Intelligence and Analytics: Systems for Decision Support. Tenth Edition. *Pearson Eductation, Inc. Copyright 2015, 2011, 2007*
17. Loh, Wei-Yin (2014). Fifty Years of Classification and Regression Trees. *International Statistical Review ,2014, Vol. 82, No. 3, pp. 329-348*
18. Oztekin, Asil, Khan, M. Raiz (2014). A Business-Analytic Approach to Identify Critical Factors in Quantitative Disciplines. *Journal of Computer Information Systems, 2014, Vol. 54, No. 4, pp. 60-70*