Predictive Modelling on Insurance Claim

GAURAV SITAULA, MD REZWAN ISLAM, ALI SHAKIBA
GROUP-13

Contents

Executive Summary

The main purpose of this report is to build a predictive model for the insurance claim. We have around 8000 observations of the customer data which contains around 27% of the customers who claimed the insurance during their insurance period. For the insurance company, one of the major challenges is that they want to reduce such claims. This analyses is beneficial in two ways. On one hand, if the client is more likely to claim the insurance, the company might increase their premium rate. On the other hand, they can focus more on the customers who are less likely to claim in the future which in turn are the lucrative customers to target in future marketing endeavors.

We have used several models to predict the insurance claim. We have built decision tree model, regression model and neural network model and compared their performance based on the certain criteria. As we are focusing on the decision, we have selected the best model based on the validation misclassification rate. After analyzing all the models, we have decided the regression model as our final predictive model.

This regression model has the validation misclassification rate of 0.217613 which means that the proportion of the disagreement between the prediction and the outcome will be 21.7613%. The top five most important traits of this data are a) where do customers live in? b) What type of job the customers do? c) Whether the customers are married or not d) Record points e) Income of the customers. In a nutshell, this model would enable the insurance company to better predict its customers' traits which in turn would result in increased operational efficiency.

Introduction

The purpose of auto insurance is to protect us against financial loss if we have any vehicle related accident. This is a contract between the owner of the vehicle and the insurance company. The owner agrees to pay the premium and the insurance company agrees to pay his/her losses as defined in auto insurance policy. Typically, auto insurance provides property, liability, and medical coverage.

According to a January 2016 report from the National Association of Insurance Commissioners, "the countrywide average auto insurance expenditure rose 3.3 percent to $841.23 in 2013 from $814.63 in 2012" ("Auto Insurance", 2017). This data alone shows the severity of auto insurance expenditure in US. The following table shows the frequency of claims for different purposes under the auto insurance coverage.

Private Passenger Auto Insurance Losses, 2006-2015

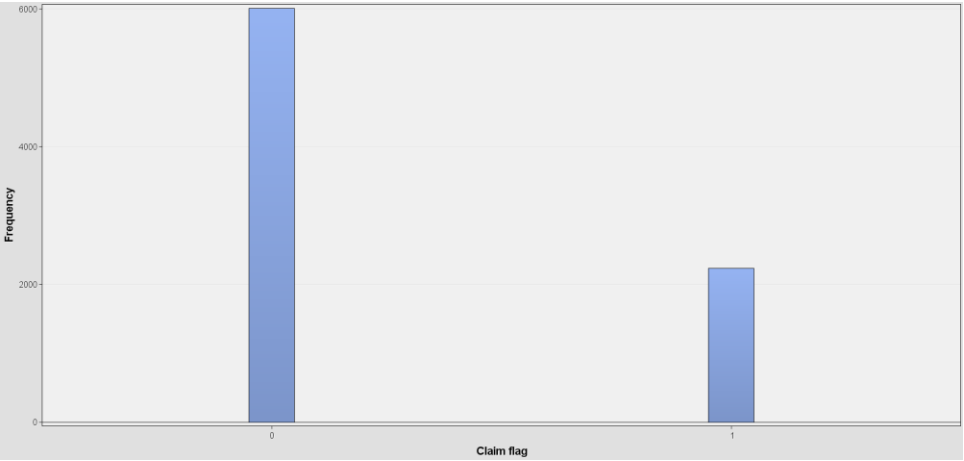
	Bodily injury	Property damage	Collision	Comprehensive
Year	Claim frequency[1]	Claim frequency[1]	Claim frequency[1]	Claim frequency[1]
2006	0.98	3.40	4.87	2.40
2007	0.90	3.46	5.20	2.48
2008	0.91	3.42	5.35	2.57
2009	0.89	3.49	5.48	2.75
2010	0.91	3.53	5.69	2.62
2011	0.92	3.56	5.75	2.79
2012	0.95	3.50	5.57	2.62
2013	0.95	3.55	5.71	2.57
2014	0.87	3.66	5.95	2.80
2015	0.91	3.73	6.05	2.73

For the auto insurance providers, insurance claim is the main business objective to tackle with. Keeping this context in mind, the purpose of this report is to develop a predictive modelling using Decision Tree, Regression Analysis, and Neural Network to understand the vital factors and thereby predict the insurance claim. Successful model would enable the insurance company to identify the factors of high number of insurance claim. The first step of building a data model is to understand the dataset and prepare the dataset for data modelling. In the following segment, we would discuss the data set and then the preparation of data.

Exploratory Data Analysis:

We have total 8240 number of observations, of which 6010 observations are with CLAIM_IND =0 and 2230 observations with CLAIM_IND =1 in the given dataset. Several factors affect the insurance claim such as Age of the driver, type of the car, age of the vehicle, Income of the owner, no. of previous claims, insurer's family, and financial backgrounds and so on. To examine these factors, total of 21 predictor variables are chosen and data are collected on these variables. Out of these 21 predictor variables, 11

are interval variables, 5 are nominal variables, 4 are binary variable and the final variable is ID. We have two target variables, CLAIM_AMOUNT and CLAIM_IND. The former target variable is of interval type and latter is Binary type. Since, we are concerned about whether a person claims the insurance or not regardless of the claim amount, we reject CLAIM_AMOUNT as our target variable. Hence, the response variable is CLAIM_IND which has two labels namely 0 indicating “No-claim” and 1 indicating “Claim”. To explore the datasets, we have used ‘StatExplore” node in the Predictive modelling diagram. The following graph shows the distribution the response variable data.

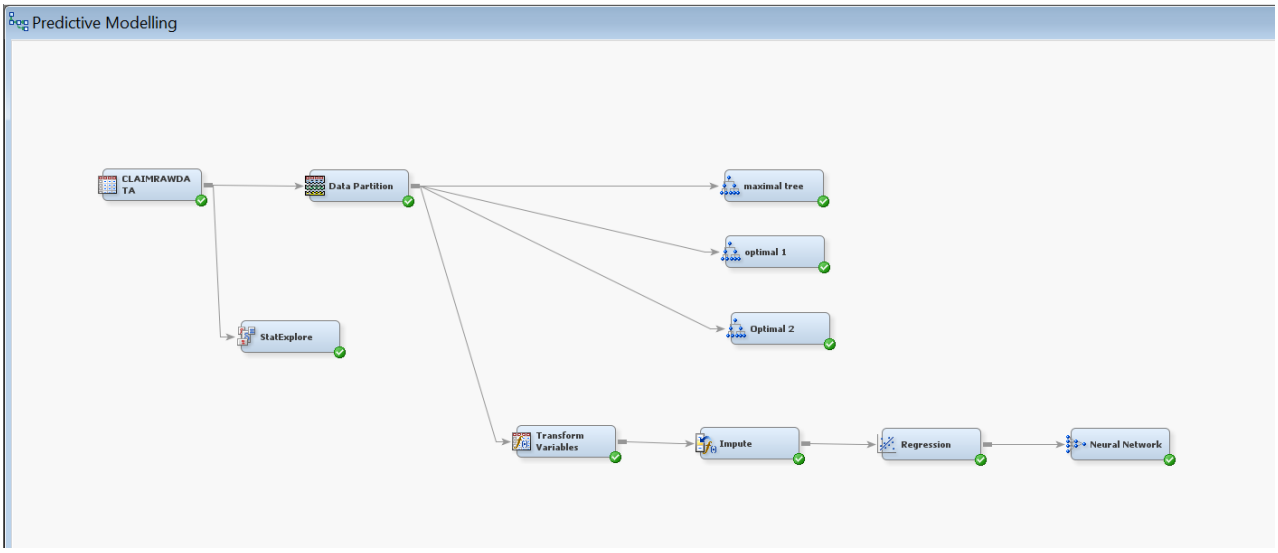


Data Preparation:

The next step of the analysis is to partition the total dataset into two separate datasets. One is the training dataset and another is the validation data set. We have followed 50:50 split criterion. Technically, we have created a data partition node, connected it to the data source and then run the program in the SAS Enterprise Miner. By doing so, the data set is ready to build any model upon it and test the result of the prediction on the validation data that is created by our prospective models. For decision tree models: data irrelevancy, redundancy, and missing values, all are taken care by the inherent algorithm of split search. Initially, we have found out the maximal tree. Then, we have pruned the maximal tree and assessed the optimal tree based on the validation data sets.

Thereafter, we have performed the Regression model and subsequently the Neural Network model. Both regression and neural network models can’t perform analysis with instances that have the missing values. To cope with this situation, at first, we have imputed the missing values using the “Impute” node in SAS. Secondly, the variables which have the skewed distribution are transformed through the “Transform Variables” node. While analyzing the variables, we have found some of the them have right skewed distribution and hence we have applied the log-transformation.

Next, our objective is to remove the irrelevant and redundant datasets. For this purpose, we have used the stepwise method to select the more important variables eliminating the irrelevant variables. We have set two threshold values for entry significance level (=25%) and stay significance level (=15%) to be equally flexible as well as selective. Then regression model is run and effective measures are noted. We have connected the Neural Network node with the regression node to see how it works on our dataset.



Performance Evaluations

Our main target is to identify whether a person claims the insurance or not. This is the decision so we evaluate the model performance based on the misclassification rate. The following table suggests the summary of the performance evaluation for each best model.

Performance Measure	Decision Tree	Regression Model	Neural Network
Misclassification Rate	0.247	0.217613	0.218341

Based on the Misclassification rate, the regression Model gives the lowest rate. So, we choose regression model as our final predictive model for predicting the insurance claim.

Model Interpretation

- Area: the odds ratio of **Area** (highly urban/urban vs z_highly rural/rural) is 12.070. This means that for highly urban/urban area, the odds of claiming the insurance is 1107% higher than the odds of claiming the insurance in highly rural or rural area (on average) keeping all the other variables constant.
- CLM_FREQ: the odds ratio of **CLM_FREQ** is 1.106. This means that for each number increase in the previous claim frequency, the odds the odds of claiming the insurance would increase by a factor of 10.6% (on average) keeping all the other variables constant and so on.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Validation Misclassification Rate
1	AREA		1	1	225.4060		<.0001	0.2707
2	IMP_JOB		10	2	257.6763		<.0001	0.2720
3	STATUS		1	3	91.4431		<.0001	0.2639
4	MVR_PTS		1	4	68.1933		<.0001	0.2562
5	IMP_LOG_INCOME		1	5	46.6689		<.0001	0.2489
6	CHILD_DRIV		1	6	45.6504		<.0001	0.2453
7	VEHICLE_TYPE		7	7	65.0863		<.0001	0.2392
8	EDU_LEVEL		4	8	47.7976		<.0001	0.2329
9	REVOKED		1	9	28.3167		<.0001	0.2259
10	CAR_USE		1	10	22.6786		<.0001	0.2234
11	DISTANCE		1	11	20.6836		<.0001	0.2198
12	LOG_VEHICLE_VAL		1	12	15.1654		<.0001	0.2222
13	CLM_FREQ		1	13	13.5351		0.0002	0.2186
14	M_YOJ		1	14	6.0577		0.0138	0.2176
15	M_LOG_HOUSE_VAL		1	15	4.0912		0.0431	0.2181
16	M_AGE		1	16	1.9642		0.1611	0.2181
17		M_AGE	1	15		1.7850	0.1815	0.2181

References

- *Auto Insurance*. (2017). *III*. Retrieved 27 February 2017, from http://www.iii.org/fact-statistic/auto-insurance
- *What is auto insurance?*. (2017). *III*. Retrieved 27 February 2017, from http://www.iii.org/article/what-auto-insurance