PROJECT REPORT

"Data Analysis for Car-Insurance Claim Prediction"

ISEN 613: ENGINEERING DATA ANALYSIS

Team:

Aashutosh Nema (UIN:928000800)

Mohit Kankariya (UIN: 527000946)

Ningyi Zhang (UIN 527005249)

Souvik Samanta (UIN: 427009736)

Vasu Kumar (UIN: 427007395)

EXECUTIVE SUMMARY

The risks of a car insurance depend on various factors such as the environment of the insured person, the make of the car, model of the car, the insured person etc. An understanding of these factors and their interactions can assist in determining the risk of a policy. This is essential for insurance companies while taking the financial decisions. A prediction model can help an insurance company in deciding the insurance claim. The company can even decide premiums based on the claim amount prediction. The project discussed here is to implement data analysis tools and techniques and predict whether an insured car will file a claim or not which is a classification prediction. Exploratory data analysis on the dataset infers that the company paid just above $150,000 over the calendar years of 2005-2006 in insurance claims. Accurately predicting the risk of claim on a car insurance can significantly change the way in which insurance policies are offered. Data mining techniques and data prediction modelling (which are a subset of machine learning) when applied to the features or predictors of a data set for an insurance firm can result in enormous gains, substantial savings and optimizing profit margins, risks and claims.

The data mining done for the project is the first stage in which the data correlation, impact and relation with claim amount is observed. The preliminary analysis is done on the various factors of households. The features of the data include vehicles in the household, year in which the vehicle was insured, make, model and sub model of the vehicle along with various categorical variables and numerical variables. Based on the importance of the variables, data modelling techniques are employed on the step selection wise and it was identified that 5 factors have a significant effect on the claims of a policy, namely: Var2, Var4, NVVar2, NVVar3, and Cat3. It is seen that most of the claims were made when the NVVar3 was equal to zero. Also, the Cat3 value being zero lead to a major percentage of the claims. The presence of these factors can significantly reduce the risk of claims and the policies can be curated accordingly. The other 3 factors do not have a direct correlation with the claims but can help in effectively predicting the claims. The predictors Var4 and Var6 are directly proportional to each other. So, using any one of these factors in our analysis will give similar results. Based on the relation of various parameters, it is identified that some categories have a positive impact for the claim amount and probability of making a claim. Specifically, Cat1 having the category B and category A of Cat 4 have high importance in deciding the claim prediction. The prediction model thus created was able to predict the insurance claims with a 76% accuracy. This was done while maintaining a healthy True-Positive rate of 35% to ensure the sensitivity since insurance firm can take decisions in obtaining the premium amount accordingly. This implies that the model has a high probability of correctly predicting a claim. The fact that this model identifies just 5 predictors along with specific categories of the model makes it very easy to implement. In conclusion the model performs exceptionally well in predicting the insurance claims. Making informed decisions about offering insurance policies will arm the company with a tool that is a potential gamechanger.

Teamwork delivers higher quality outcomes that are more thoughtfully constructed, effective and timely. Mr. Aashutosh performed data mining and concentrated on building linear and radial SVM models. Mr. Mohit contributed in analyzing the categorical variables and fitting decision tree model for the insurance firm. Ms. Ningyi effectively did the data imputation work and subset selection along with PCA for the for QDA model. Mr. Souvik concentrated on building LDA model with incorporation of data sampling techniques (down-sampling and up-sampling) to account for data imbalance. Mr. Vasu concentrated on data analysis-imputation work and using step-subset selection technique for logistic regression model. Each team member contributed efficiently and effectively on the data mining, analysis, report writing and in building an effective technique to predict the claim.

1. Technical report of Model 1- Logistic Regression

1.1 Introduction

The risk of car insurance depends on various factors such as the environment of the insured person, the make of the car, and the insured person. A good understanding of these factors and how they determine the risk of a policy is essential for insurance companies. The first model suggested, based on the performance on test data is Logistic regression. There is a substantial number of missing values (1.8%) in the data set. The data imputation/ feature engineering was done based on the following 4 main criteria:

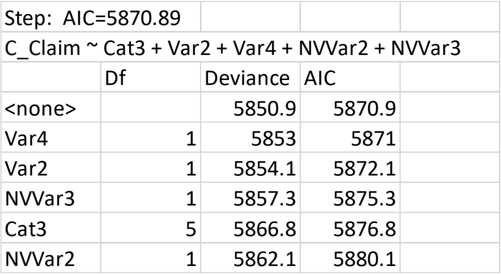
1) Assigning the values based on the distribution pattern
2) Assigning the values based on the correlation pattern
3) Assigning a new independent category (in case of too many missing values with no observable pattern)
4) Assigning the values after comparing the distribution pattern with C_Claim

1.2 Methodology

The insurance claim prediction is done on the new column – C_Claim which has response variable 0 for no claim and 1 for any positive claim. Due to high dimensionality (i.e. number of categories, predictors) and the size of data, we use dimension reduction techniques to reduce the computation time and improve the interpretability of the model. The techniques used for the modelling and feature selection are: correlation values between numerical variable, p value output (i.e. importance with respect to model), step selection AIC. The analysis and modelling were carried out with train and test split ratio of 70:30. Model building and selection was done on the training data set and its true positive rate and accuracy was compared in order to select the best logistic model.

1.3 Model Selection

This report discusses the logistic model built using step selection AIC that was in the criteria to effectively predict the logistic class for C_Claim. The step AIC value observed 5870.89 and the variables can be seen in the table 1.1.



Step: AIC=5870.89				
C_Claim ~ Cat3 + Var2 + Var4 + NVVar2 + NVVar3				
	Df	Deviance	AIC	
<none>		5850.9	5870.9	
Var4	1	5853	5871	
Var2	1	5854.1	5872.1	
NVVar3	1	5857.3	5875.3	
Cat3	5	5866.8	5876.8	
NVVar2	1	5862.1	5880.1	

Table 1.1. Step selection output

Based on the threshold values for the model, the pattern for classification was observed for true positive rates, true negative rates and accuracy.

1.4 Evaluation

This report discusses the output obtained with 35% true positive rate and the corresponding observed accuracy (i.e. 76%) for the model. The results of this regression were analyzed by comparing the confusion matrix for the training and test data.

After running the stepwise selection and the model on training data set, the prediction accuracy and true positive rate was consideration for the training data set. It performed at 73% Accuracy with 35% true positive rate based on our adjusted threshold value.

<table>
<tr><th colspan="5">Training Data Prediction</th></tr>
<tr><td></td><td colspan="2"><b>Actual</b></td><td rowspan="2"><b>True positive Rate</b></td><td rowspan="2">33%</td></tr>
<tr><td><b>Predicted</b></td><td><b>0</b></td><td><b>1</b></td></tr>
<tr><td><b>0</b></td><td>52978</td><td>334</td><td rowspan="2"><b>Accuracy</b></td><td rowspan="2">76%</td></tr>
<tr><td><b>1</b></td><td>16527</td><td>161</td></tr>
<tr><th colspan="5">Test Data Prediction</th></tr>
<tr><td></td><td colspan="2"><b>Actual</b></td><td rowspan="2"><b>True positive Rate</b></td><td rowspan="2">35%</td></tr>
<tr><td><b>Predicted</b></td><td><b>0</b></td><td><b>1</b></td></tr>
<tr><td><b>0</b></td><td>22608</td><td>147</td><td rowspan="2"><b>Accuracy</b></td><td rowspan="2">76%</td></tr>
<tr><td><b>1</b></td><td>7167</td><td>78</td></tr>
</table>

Table 1.2 Confusion Matrix for Logistic Regression using Step Selection

The threshold was adjusted to optimize the true positive rate and accuracy, since we observed a considerable amount of increase in true negative rate further we kept 35~33% as our true positive rate to obtain accuracy. Also, a very high accuracy (with low true positive rate) in fitting the training data might lead to overfitting in the test data. Table 1.2 shows the confusion matrix obtained from the training data set prediction and test data set prediction. Figure 1.1 and figure 1.2 shows the ROC curves for the true positive and true negative rates and kernel density curve for the probability values obtained.

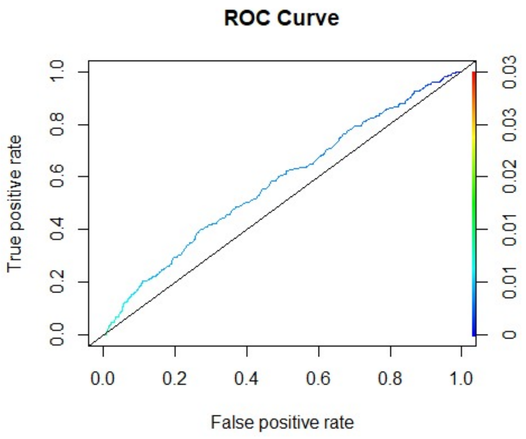


Figure 1.1. ROC Curve

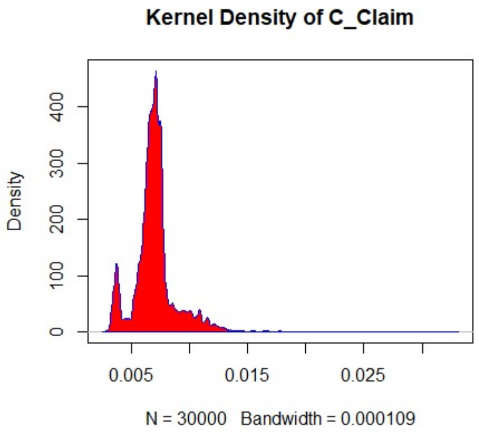


Figure 1.2. Probability density curve

1.5 Conclusion

The model obtained for logistic regression after the stepwise selection identifies the 5 most significant variable contributing to the C_Claim value. The relative test accuracy and true positive rate is similar to the training data set which is a key parameter in ensuring the performance of the model in order to judge its interpretability and repeatability. This reduction in the number of variables reduces the computation time and also improves the interpretability of the model. The initial model contained 0.7% of the C_Claim values. The test model accuracy of 76% is the highest achieved result with 35% true positive rate is selected as best output for this model.

2. Technical report of Model 2- Linear Discriminant Analysis (using Undersampling)

2.1 Methodology

The insurance claim prediction is done on the new column – C_Claim which has response variable 0 for no claim and 1 for any positive claim. Before fitting a model, the dimensionality (i.e. number of categories, predictors) and data imbalance need to be adjusted. Data imbalance was handled using sampling methods – over-sampling and under-sampling, and the dimensionality was reduced by observing the correlation matrix of predictors. The final model was selected based on accuracy, true positive rate, computation time and interpretability.

The data was initially split into training and test data with a 70:30 ratio for modelling on 70% of the data set, predicting the model on it and assessing its performance on test data set. Due to high data imbalance (0.72:99.28), sampling of data was done for accurate results. Over-sampling and under-sampling were performed on the training set using the ROSE package.

Oversampling – The training data of the C_Claim predictor was sampled and increased the minority class(1), from 506 to 15506 values. This increased the class ratio to 18.2:81.8, reducing the imbalance.

Undersampling - The training data of the C_Claim predictor was sampled and decreased the majority class(0), from 69494 to 3494 and then further to 1494 values. We created two undersampled models of total length 4000 and 2000, with a class ratio of 12.65:87.35 and 25.3:74.6 respectively.

2.2 Model selection

Due to the high dimensionality of the dataset, we created a correlation matrix for the data. Using these values, the following predictors were found to be significantly correlated: Vehicle, Calendar_Year, Model_Year, Cat1, Cat3, OrdCat, Var1, Var2, Var3, Var7, Var8, NVVar1, NVVar2, NVVar3 and NVVar4.

LDA was performed using the datasets and predictors mentioned above. We used threshold optimization for both predicting the training and test data. The best result was obtained using the down sampled training data with 4000 parameters. Below are the parameters observed for the following model: C_Claim~Vehicle+Calendar_Year+Model_Year+Cat1+Cat3+OrdCat+Var1+Var2+Var3+Var7+Var8+NVVar1+NVVar2+NVVar3+NVVar4

2.3 Evaluation

Comparing the confusion matrix for the training data set and under-sampled data set with the optimized threshold values., the under-sampled data set(4000) gives better accuracy and true positive rate as it can be seen from Table 2.1 and table 2.2.

LDA : variable reduction using Undersampling				
on Training	**Actual**		TP Rate	35%
Predicted	0	1		
0	2704	328	Accuracy	72%
1	792	176		

Table 2.1 Confusion matrix for Variable reduction LDA Undersampling on training data

The Figure 2.1 and figure 2.2 shows the trade off true positive rate and the accuracy for the LDA model using undersampling to infer the optimum threshold value. It also shows the density of probability distribution and the ROC Curve.

LDA using variable reduction					**LDA : variable reduction using Undersampling**				
On Test Data	**Actual**		TP Rate	34%	**on Test Data**	**Actual**		TP Rate	35%
Predicted	0	1			**Predicted**	0	1		
0	21551	142	Accuracy	72%	0	22348	147	Accuracy	75%
1	8233	74			1	7427	78		

Table 2.2 Confusion matrix for LDA using variable reduction using p value and covariance and using Under-sampling technique

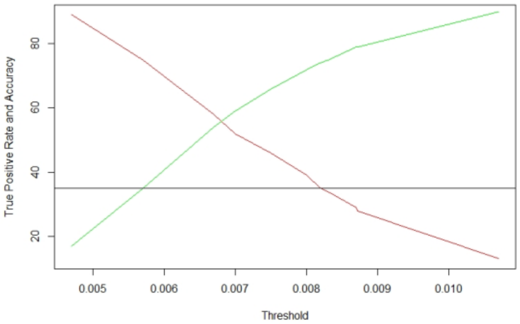


Figure 2.1 True positive rate and Accuracy plot

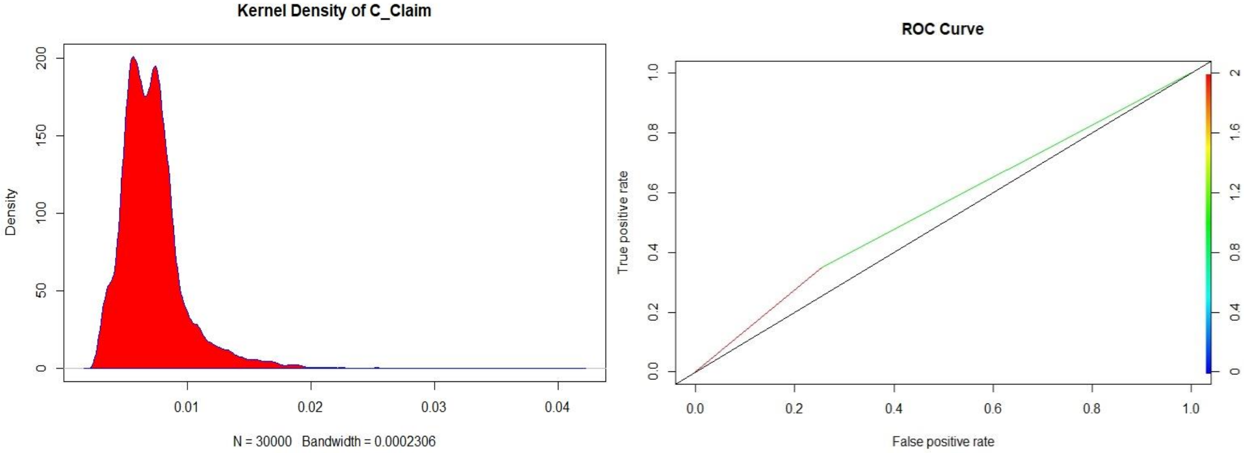


Figure 2.2 Probability density curve and ROC Curve

2.4 Conclusion

The final Linear Discriminant model is created using the down-sampled training data set and reduced dimensionality using correlation; 15 predictors out of the original 33 were used. The relative accuracy of the down-sampled dataset of 4000 size (75%) compared with the original training dataset (72%) and over-sampled dataset of 85000 size (74%) were higher for the same true positive rate (34% - 36%). Reducing the dimensionality and dataset size further improves the computation speed and interpretability, while removing the inherent data imbalance in the original data.

3. Technical report of Model 3- Quadratic Determinant Analysis

3.1 Methodology

The insurance claim prediction is done on the new column – C_Claim which has response variable 0 for no claim and 1 for any positive claim. We then used dimension reduction techniques to overcome the Curse of dimensionality due to large size of data, because it only the "Vital few" instead of "Trivial Many" that matter and significantly affect our prediction, rest all can be considered to increase noise and negatively affect our inference and prediction of data. After considering cross validation, best subset selection and PCA, the p-values and correlation between variables was used for the dimension reduction of the model.

3.2 Model Selection

The model selected is: *C_Claim ~ Cat1+Cat3+Cat8+Var2+Cat6+Cat11+Var1+Var3+Var7.* The steps involved in finalizing this model are described below.

The amputated data was initially split into Training and Test data with a 70:30 ratio. Given the high dimensionality of the data we eliminated certain noncontributing columns like Houshold_ID, Blind_Model, Blind_Submodel, Calender Year, X as these variables did not contribute much to our response and considering them would have led to increased computation time, addition of noise to the model and wasting the degrees of freedom.

For the remaining variables, we performed Quadratic Determinant Analysis with 70000 train observations and C_Claim as the response. We then further tried to improve the model based on the correlation between the variables and p-values generated from linear regression. On carefully weighing each factor and eliminating the variables producing noise we reach 7 important variables, which are Cat1, Cat3, Cat8, Var2, Cat6, Cat11, Var1, Var3, Var7.

3.3 Model Evaluation

Given the high skewness of our data, we reached a tradeoff between True positive rate and Accuracy of the model and these True positive and accuracy rates were considered acceptable. As trying to further increase the True Positive rates, the True Negative rates were found to increase substantially.

Quadratic Determinant Analysis				
on Train Data	**Actual**		**True positive Rate**	35%
Predicted	**0**	**1**		
0	58189	323	**Accuracy**	83%
1	11316	172		

Quadratic Determinant Analysis				
on Test Data	**Actual**		**True positive Rate**	36%
Predicted	**0**	**1**		
0	20180	319	**Accuracy**	76%
1	5979	176		

Table 3.1 Confusion matrix for model prediction on train and test data

Also, a very high accuracy in fitting the training data might lead to overfitting in the test data which might possibly lead to inaccurate predictions using the model. Hence we determine a cutoff value of 35% True Positive rate and try to best fit our model, trying to increase the accuracy while maintaining the True Positive rate.

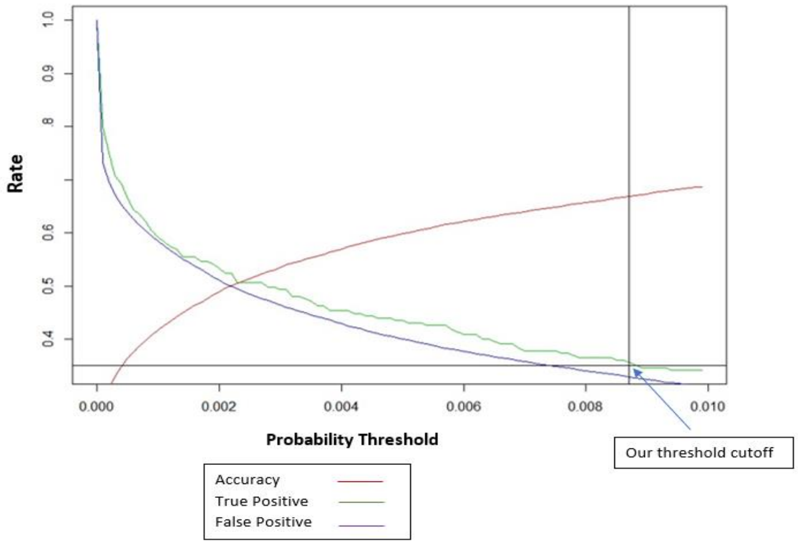


Figure 3.1 Accuracy, true positive and false positive rates

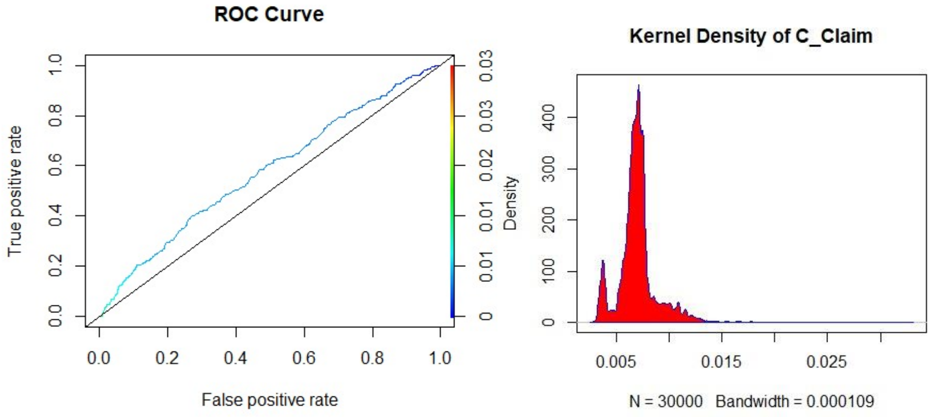


Figure 3.2 ROC Curve and Probability density

Figure 3.1 shows the accuracy with the threshold values. The plot is between 0-0.1 based on the understanding of distribution pattern of probability distribution. On selecting the threshold for 35% true positive rate will get a high considerable accuracy. The prediction on train data shows 83% accuracy while when it is performed on test data it results in 76% accuracy which infers the variability in prediction.

3.4 Conclusion

It is seen that the model obtained after elimination of predictors, based on correlation and p-values, consisted of 7 most significant variable contributing to the C_Claim value. We observe that after varying the threshold value our model Accuracy is 76% at 36% True positive rate, which is the best value of Accuracy we get given our True positive base cutoff and the Tradeoff between True positive and prediction accuracy. Hence our final QDA consists of 7 predictors as Cat1, Cat3, Cat8, Var2, Cat6, Cat11, Var1, Var3, Var7.

4. Best Model Selection

The data set for the C_Claim prediction on analysis showed skewedness towards zero because of the 0.72% of 1's present in it. The aim was to select the parameters of the model in such a way that it provides an acceptable level of accuracy without overfitting the data and predicting the true claim cases considering on a reduction of false positive predictions. The threshold was selected in a way that it maintains the balance between the True Positive rate, true Negative rates and accuracy to ensure its interpretability. On doing the analysis of the distribution pattern of predicted values, the true positive rate of 35% was seen to be optimum. The comparison for all models was done by observing the increase in the True Positive Rates, which showed a drastic increase in the True Negative Rates thereby consuming on the accuracy.

After finalizing on the True Positive Rate to be 35% Logistic regression, LDA and QDA were performed on various combinations of predictors in the models based on subset selection, p value, covariance and under sampling. Model 1 was selected which provides the best accuracy for the selected True Positive Rate of 35% for test data as well as training data to ensure its interpretability. The step selection for logistic regression technique was used in order to proceed for logistic regression prediction. Also, the analysis of covariance matrix and p value also bolstered the feature selection to explain the model. The step selection uses AIC values for the shortlist.

These models selected are as follows:

1) Logistic Regression- C*_Claim ~ Cat3 + Var2 + Var4 + NVVar2 + NVVar3*
   True Positive Rate- 35%, Accuracy- 76%
2) LDA-
   C_Claim~Vehicle+Calendar_Year+Model_Year+Cat1+Cat3+OrdCat+Var1+Var2+Var3+ Var7+Var8+NVVar1+NVVar2+NVVar3+NVVar4
   True Positive Rate- 35%, Accuracy-75%
3) QDA- C*_Claim ~ Cat1+Cat3+Cat8+Var2+Cat6+Cat11+Var1+Var3+Var7*
   True Positive Rate- 35%, Accuracy- 66%

For the selected true positive rate, the model selected after logistic regression was seen to achieve the best accuracy of the 3 methods. The QDA model was seen to have high accuracy (91%) on the training data however, its accuracy fell drastically when it was used on the test data. This gives us the perspective that the model is not interpretable. For the threshold set to compare the accuracy of the model, it was observed that logistic regression (used with step selection) performs slightly better than LDA on the test data set. Also, it is more interpretable and the relationships between the predictors and response variables.

Thus, it was concluded that a linear model would work well for this dataset. Logistic Regression tends to perform better in the presence of categorical values and is more robust to different type of data distributions. In addition to this Logistic Regression provides us with a highly interpretable model and is easier to compute. However, logistic regression models are prone to overfitting, it was ensured that the selected model would avoid that eventuality by optimizing the threshold value for prediction. Also, once the highly corelated variables were removed, the final model obtained from logistic regression can be expected to perform well for predicting whether an insured car will file a claim or not.

5. Evaluation of test results

To evaluate the Test data and to compare it with our training model, C_Claim variable was imputed corresponding to the Claim amount value, like the training data. The test data was then predicted using the logistic model selected after the initial analysis and the predicted C_Claim values were compared against the actual data values.

Test Data Prediction with original model				
	Actual		TP Rate	37.057%
Predicted	0	1		
0	37751	231	Accuracy	75.774%
	11882	136		

Test Data Prediction with improved accuracy				
	Actual		TP Rate	0.82%
Predicted	0	1		
0	49245	364	Accuracy	98.5%
1	388	3		

Table 5.1 Confusion Matrix with original model and maximum possible accuracy model

After comparison of the predicted and actual C_Claim values for the test data, it is observed that the model prediction achieves an accuracy of 75.77% with a true positive rate of 37.06%, while the accuracy and true positive rates for the training data were 76% and 35% respectively. For these values it is evident that the selected model performs equally well on the test as well as the training data. A "0.00754" probability threshold value was selected as this ensures a good accuracy while maintaining a considerable True Positive rate. Our model has an error rate of 24.22% with false positive rate as 23.93% and false negative rate as 62.94%. Our model relates to the realistic scenario as a significant True positive rate is critical for an insurance firm because it ensures that the positive claims are identified accurately, up to a certain extent, to frame the policies and premium accordingly. Also, presence of higher false positive rate in our error percentage is desirable as compare to false negative rate.

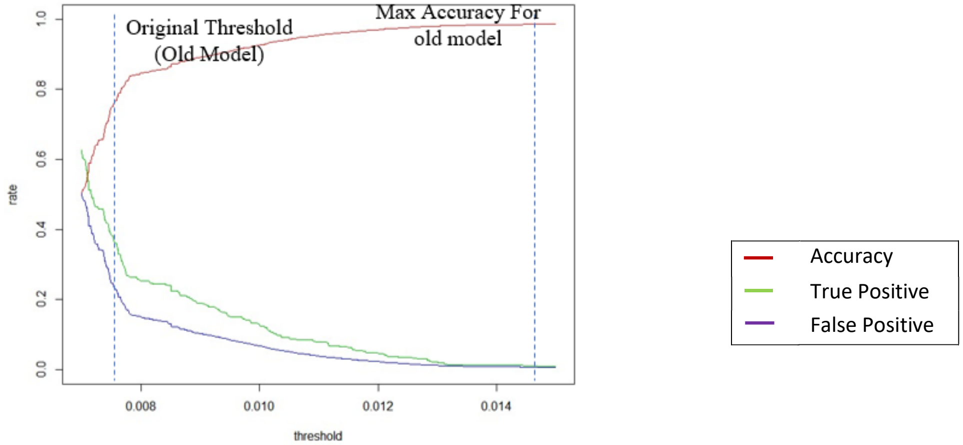


Figure 5.1 Interaction curve – Threshold – Accuracy – False Positive – True positive

As expecting a policy to be claimed while it goes unclaimed is a safe bet, while the contrary can lead to losses for the insurance firm, which is not desirable. The model accuracy can be improved from 76% percent to 98.5% at the expense of reduced True Positive rate, but this would also increase false negative rate to 99%, which is not desirable for realistic situation as while doing so we would be misleading the firm and the whole point for this analysis would be invalidated.

6.0 Improving the Test Error

6.1 Approach

After testing our current model on the test set, the objective is to modify it or build a better model for the prediction of given insurance data. Opting to improve the current model, the shortcomings of the current model and the data manipulation methods were identified. The analysis was based on the following parameters:

- Test error – The test error was aimed to be reduced
- Data distribution – The correlation between each variable and the Claim amount was better identified
- Correlation patterns - the distribution of categorical variables we modified based on their observed correlation in the training set and the test set

The above-mentioned factors were focused on while improving the current model. Analysis of the categorical variables showed that the distribution of several variables, namely Cat1, Cat4 and OrdCat, were changed to improve our model. The distribution of the above-mentioned variables was observed in the test dataset and depending on their occurrence and distribution the training dataset was modified accordingly. These changes improved the fit of our model on the test and training data, which in turn reduced the test error. The following changes were made,

- Cat1 – It contained 10 categories in the training set, namely A, B, C, D, E, F, G, H, I and J, whereas the test set contained only 3 categories A, B, C,.E, F and G The category H in the training set was changed as B, as occurrences of H were sparse.
- Cat 4 – It contained 3 categories in the training set, namely A, B and C, and the test set contained A and C. The category B was changed to A, as the value B was missing in test set.
- OrdCat – It contained 7 categories, namely 1, 2, 3, 4, 5, 6 and 7 in the training set, and the test set contained 6 categories namely 2, 3, 4, 5, 6 and 7. The category 1 of the training set was changed to 4 because of its distribution in the test set.

These changes created a boosting effect and improved the prediction of our model on the test data.

6.2 New model selection and Threshold decision point

The objective was to improve the accuracy and reducing the test error rate. Working on this objective, we optimized the threshold value to incorporate both conditions while also giving importance to the Sensitivity or True Positive Rate, as it is important for an insurance company to focus on the accurate number of claims predicted. Keeping a balance, the TP rate and Accuracy for test set were measured as 25 and 84 percent respectively. By sacrificing on the True Positive rate, we were able to achieve more than 98.4 percent accuracy on both the training and test set with a True Positive rate of 0.8 percent, but this is only observed due to the high data imbalance in the dataset provided. Such a high accuracy obtained at the cost of true positive rate has logical fallacies as essentially, we are overfitting the data and it is also an ineffective model to the insurance company.

In order to effectively perform logistic regression, we need to analyze the predictors majorly responsible for the Claim amount and only include them in the model to eliminate noise and improve prediction. The correlation between various predictors and the Claim Amount was observed for both the training and test data set. Based on the obtained results only the following predictors were included in the regression model: *Cat1, Cat4, OrdCat, Cat3, Var2, Var4, NVVar2 and NVVar3.*

6.3 Model Evaluation

The following table displays the confusion matrix for the test dataset and validation data set for both optimized and the maximum accuracy model:

Logistic Regression (Optimized) (Thresh-0.00815)					**Logistic Regression (Optimized) (Thresh-0.00815)**				
On Test Data	**Actual**		TP Rate	25.7%	**on Validation Data**	**Actual**		TP Rate	24.5%
Predicted	0	1			**Predicted**	0	1		
0	24162	167	Accuracy	80.7%	0	42107	277	Accuracy	84.4%%
1	5613	58			1	7526	90		
Logistic Regression (Max Accuracy) (T-0.015)					**Logistic Regression (Max Accuracy) (T-0.015)**				
On Test Data	**Actual**		TP Rate	0.4%	**on Validation Data**	**Actual**		TP Rate	0.8%
Predicted	0	1			**Predicted**	0	1		
0	29536	224	Accuracy	98.4%	0	49245	364	Accuracy	98.5%
1	239	1			1	388	3		

Table 6.1 Confusion matrix of logistic regression with optimized and max accuracy model

The above variation is obtained by varying the threshold value. The variation of Sensitivity and Accuracy with the threshold values are plotted on the following graph.

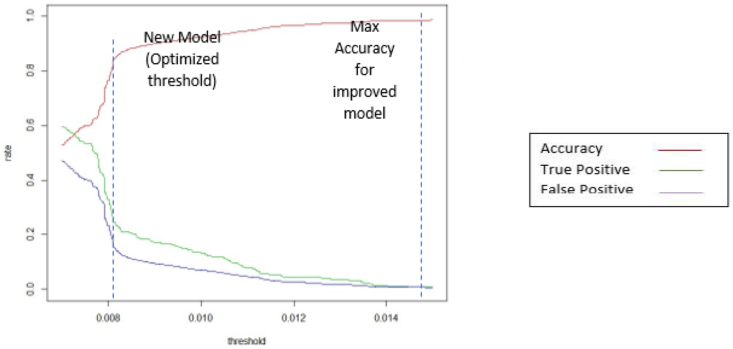


Figure 6.1: (a) Variation of Accuracy and Sensitivity with threshold for improved model on new validation set data

6.4 Conclusion

The final model has improved the accuracy significantly and can result in even high accuracy at the expense of true positive rate or sensitivity. The model selected was linear with step selection and categorical boosting. Since the objective of the project was set to be based on test error, the result was optimized for 84% test accuracy and 24.5% true positive rate. However, on adjustment of threshold the accuracy can be varied to a level of 98.5 %which can be viewed from figure 6.1 as well. For an insurance company, true positive rate/ sensitivity is also a critical factor in deciding the claim status based on parameters. That is why we are still considering some true positive rate in order to get the model predict actual claim status with an accuracy. The new model is improved by around 10% from the previous model and can be further improved to around 24~25% based on the client requirement compensating on the sensitivity.