Prepared By
Usman Tahir

Insurance Premium Default Propensity

26th April, 2020

McCombs School, University of Texas, Austin USA
Data Science & Business Analytics
By Great Learning

CAPSTONE PROJECT | Final Report

TABLE OF CONTENTS

Abstract

Purpose of this document is to present insights and analytical approach for building predictive model for insurance premium default. Detailed emphasize on important variables and statistical treatment through data preprocessing. Picking the analytical approach to solve the project problem for further analysis.

EXECUTIVE SUMMARY

Default in premium payments impact significantly on the profitability of the insurance company. Therefore, predicting defaults in advance cannot be overemphasized. For further details on objective please refer to section Introduction.

Exploring data through sequential data mining process of univariate, correlation, regression, bivariate and multivariate analysis helped us extracting hidden insights in the data. Dataset has imbalanced classes in default variable, having about 6% defaulted cases against 94% no-default cases, which pushes data models to keep high specificity, and that is against our requirement. We require maximum sensitivity on high confidence intervals. Therefore, oversampling technique is applied to make both classes balanced having 50% split.

Since after oversampling dataset crossed more than 100,000 observations therefore, we employed 2-Fold cross validation only and on top having 70-30 training vs testing data partitioning to prevent overfitting.

There are five dimensions of variables demographics, payment history, late payment, financial worth and outcome variables. As per our analysis we found that payment history and late payments are the key dimensions for predicting defaulted cases. Please refer to section "Importance of Dimensions" for further details.

Key finding is that customers paying through cash are more at risk of default. Secondly, 6 to 12 months late payments followed by other late payment variables also increase the propensity of default.

Other notable insight is that Risk Score between 93-98% is more at risk of default then 98-100% or 91-93%. One variable was created for analysis namely "Premium Rate", but surprisingly this variable has very weak correlation with default. Usually higher cost price reflect in higher churn rates as customer go with competition company but in this data certainly is not the case. This validates that Risk Score indeed is for determining the premium rates having same behavior against default.

In multivariate analysis it was noted that clusters of defaulted cases were in polygons engulfed within no default cases therefore tree-based algorithms were taken into consideration. The best model for prediction of defaulted cases, turns out to be bagging algorithm of Random Forest having 80.1% sensitivity with overall accuracy of 73% and area under the curve of 76.3%. Please refer to section "Random Forest" for further details.

Finally, it is recommended that late payment variables should be further divided into monthly aging slots, so that algorithm can detect patterns and make predictions well in advance. Secondly, creating promotions for encouraging customers to use alternate payment modes

instead of cash. Thirdly, there should be few variables showing customer satisfaction. Reason for this is that looking at the customer profile you can see that new customers are more than old customers showing that satisfaction could be strong predictor of default.

INTRODUCTION

In the age of fourth industrial revolution all businesses seek digital transformation. One of the key elements of digital transformation is your ability to manage DATA. Data Science and business analytics is the tool which is being employed on holy grail of DATA to extract hidden insights. Since, amount of data is exponentially increasing, therefore systematic process of data science is gaining popularity in recent times. Like any other industry 'INSURANCE' industry is no exception and in fact, it is one of the key areas where data science is being practiced at large scale. In our project we will be looking to build data model for Insurance company, which can predict their customer churn probability proactively.

Problem Statement

Premium paid by the customer is the major revenue source for insurance companies. Default in premium payments results in significant revenue losses and hence insurance companies would like to know upfront which type of customers would default premium payments. So that, they can prevent them from churning.

The objective of this project is to

- Build a model that can predict the likelihood of a customer defaulting on premium payments (Who is likely to default)
- Identify the factors that drive higher default rate (Are there any characteristics of the customers who are likely to default?)
- Propose a strategy for reducing default rates by using the model and other insights from the analysis (What should be done to reduce the default rates?)

Purpose

In order to achieve aforementioned objective, we are required to build predictive data model, which can consistently predict probability of customer churning out and defaulting on premium payments. Since this model needs to take into account many key indicators effecting DEFAULT, therefore data analysis with structured sequential procedures are applied on the data to come up with pragmatic and robust model. Mishandling of key predictors/variables can result in feeble data model with real consequences on business. Therefore, applying industry best practices is the safer approach as it is tried and test several times and proved their worth over time. Finally, at the end it is very important to conduct comprehensive statistical analysis on the data before coming up with any business rule to ensure that all the variations in the data are captured with appropriate confidence levels.

Business Opportunity

Due to operational predictive models, organizations are more informed about their customers and be able to manage churn rates better due to proactive alerts which enable them to extend promotions to concerned prospects. Owing to this, business enhance their profitability, because retaining customer is always cheaper than acquiring new one. Most importantly, since you are gauging customer behaviors through data and making effort to reduce frictions, therefore your understanding of customer demand raises. This in other words, mean that you understand your customer need better than before by building such models. Finally, customer experience is raised due to engaging interactions.

DATA REPORT

The dataset provided has customer information of some insurance company. It has several categorical and continuous variables.

Data Collection & Frequency

This dataset seems to be progressive record keeping, which usually updates over time to reflect updated status of particular customer. Which means that provided data is the snapshot of some particular date, where all the records show updated status of each customer. This dataset is updated on each late payment, payment of premium, change in risk score, change in circumstances of customer such as marital status, age, change in dependents, ownership of vehicles and house etc.

Visual Inspection of Data

There are total 79,853 rows in the dataset with total 17 columns. There are five categorical and twelve continuous variables in the dataset with no missing values and columns. All rows are complete 100%.

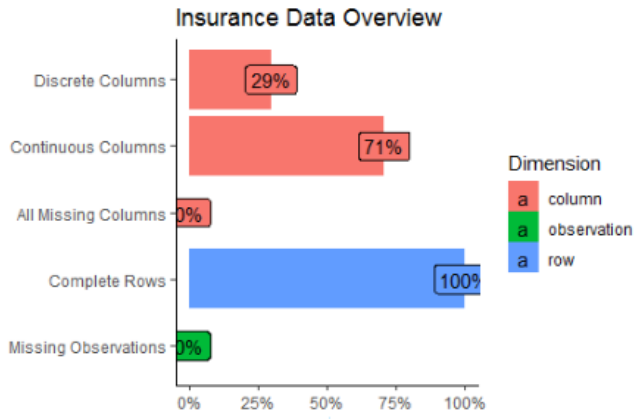


Figure 1: Dataset Introduction

There are 17 variables in the dataset and their names need to be changed so that it becomes easier for analysis as at the moment column names are not as per standard and they are very long. In further analysis we will be discussing new names rather than original dataset column names.

Variable Name	SCALE	Magnitude
CONTINUOUS VARIABLES		
Id	ORDINAL	Thousands
per_cash	INTERVAL	Percentage
age_in_days	INTERVAL	Thousands
Income	INTERVAL	Thousands
veh_owned	INTERVAL	Tens
3to6_Late	INTERVAL	Tens
6to12_Late	INTERVAL	Tens
more12_Late	INTERVAL	Tens
risk_score	INTERVAL	Tens
no_premiums	INTERVAL	Tens
Premium	INTERVAL	Thousands
dependants	INTERVAL	Tens
CATEGORICAL VARIABLES		
sourcing_channel	NOMINAL	Multiclass
area_type	NOMINAL	Binary Char.
Accom	NOMINAL	Binary Num.
Marital	NOMINAL	Binary Num.
Default	NOMINAL	Binary Num.

Table 1: Variables Scale & Magnitude

Understanding of Attributes

As discussed, purpose of the dataset is to represent updated customer profile viz a we risk score and defaults. Therefore, it must have all the dimensions required to maintain data usefulness. Therefore, we can easily classify all variables into five dimensions which are mentioned below:

Customer Demographics

This class represent customer profiles in the clienteleand the variables under this dimension are AGE, MARITAL, AREA_TYPE.

Financial Worth of the Customer

This class represent financial strength of the customer therefore variables such as DEPENDANTS, VEH_OWNED, ACCOM, INCOME are the factors which accomplishes the purpose.

Payment History

Purpose of this class is to show the credit history of the customer therefore variables like NO_PREMIUMS, PREMIUM, PER_CASH are the good indicators of how good history is of some particular customer.

Late Payment

Insurance industry tracks late payments such as any other industry tracks Days sales outstanding (Aging DSO). Therefore, variables providing aging information of late payments are included in this class such as 3TO6_LATE, 6TO12_LATE and MORE12_LATE.

Status Variables

These are potential response variables which results in some outcome basing on the above four dimensions. DEFAULT and RISK_SCORE are such variables. RISK_SCORE is connected to PREMIUM_RATE whereas DEFAULT shows churning of the customer.

Finally, it is important to note that our concerned response variable DEFAULT relies more on which attribute to predict “defaulted” outcome. Please note “defaulted” is shown as ‘0’ numeric value in the dataset and “no-default” as ‘1’. This is counter intuitive therefore we will reverse it so that our analysis is free from ambiguity.

EXPLORATORY DATA ANALYSIS

Let's begin our exploration journey of the insurance dataset, which is a five-step process as shown below in the figure 2.

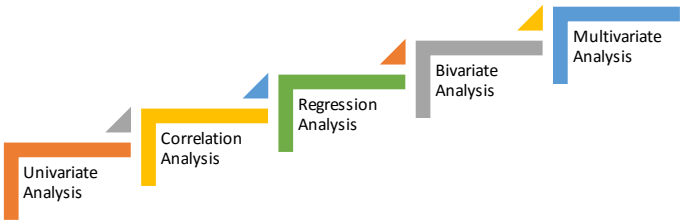


Figure 2: Exploratory Data Analysis Process

We will conduct univariate analysis to understand distribution of all variables and then perform correlation analysis to detect variables which are highly correlated with Default. In-order to filter out important variables out of correlated pairs we will perform basic level regression analysis (Logistic Regression). Performing bivariate & multivariate analysis of important variables.

Univariate Analysis

Before proceeding for univariate analysis lets remove "id" variable from the dataset as it is not required in the analysis:

PER_CASH

Min	1st Qtl.	Med	Mean
0.000	0.034	0.167	0.314
3rd Qtl.	**Max**	**Sd**	**Var**
0.538	1.000	0.335	0.112

Table 2: Statistical Summary - PER_CASH

This variable shows how much particular customer is paying their premiums through cash and this is part of Payment History dimension.

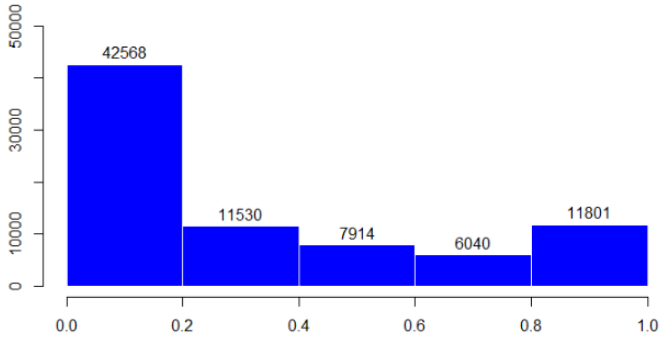


Figure 3: Univariate - Histogram of PER_CASH

Straightaway there are two observations 0 to 20 percent bin and 80 to 100 percent bins has the highest number of customers respectively. Then few are in the middle having varying share of cash payments.

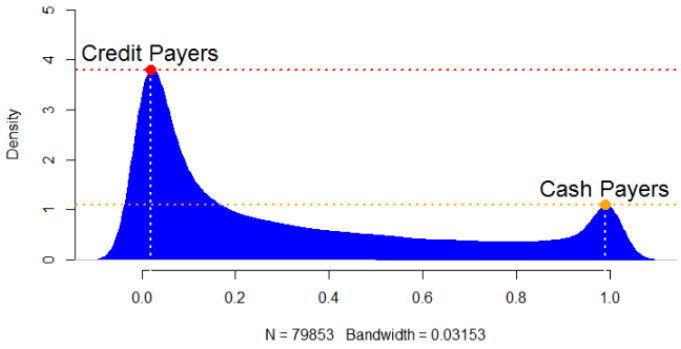


Figure 4: Univariate - Density Plot of PER_CASH

Therefore, we can say this behaviour of either paying all premiums by cash or through credit card is in majority of the customers. However, few customers vary as per the circumstances perhaps.

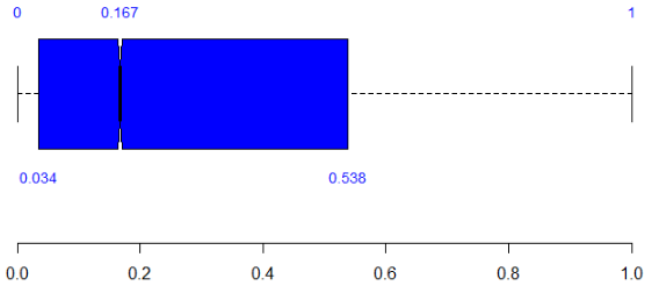


Figure 5: Univariate - Boxplot of PER_CASH

There are no outliers in this variable with slight skewedness to the right and confidence interval for median is very small.

AGE

This is the age of the customer in days at certain point of time as this data is a snapshot of one point of time. To make it easy to understand we convert this variable to years.

Min	1st Qtl.	Med	Mean
21.01	41.02	51.03	51.63
3rd Qtl.	**Max**	**Sd**	**Var**
62.02	103.02	14.27	203.64

Table 3: Statistical Summary - AGE

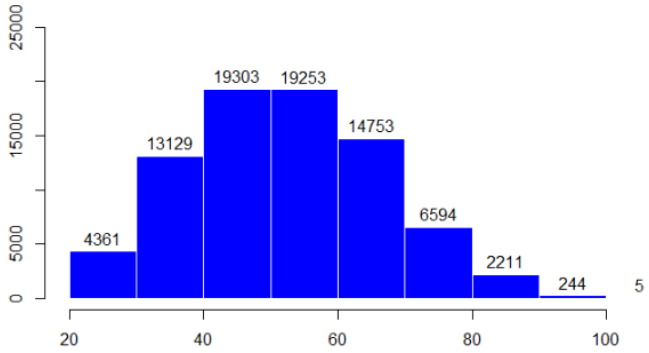


Figure 6: Univariate - Histogram of AGE

It is a normal distribution with slight skewedness to the right due to very old age individuals having more than 100 years of age.

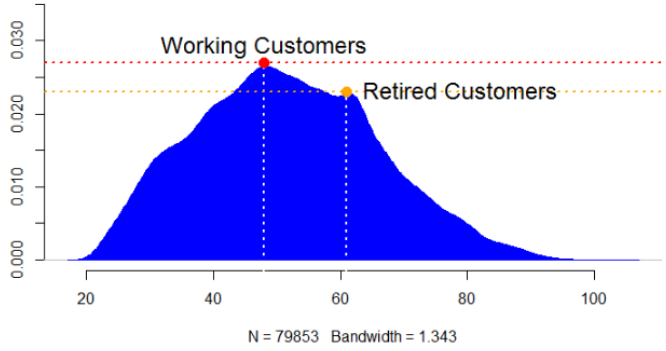


Figure 7: Univariate - Density Plot of AGE

We can see from density plot that working customers having insurance are at peak at age of 48 years, whereas customers after retirement age of 60 years again have a rising spike.

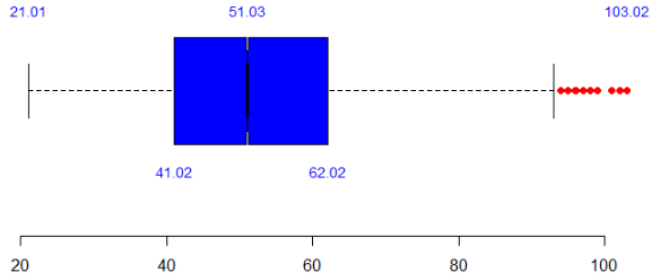


Figure 8: Univariate - Boxplot of AGE

There are outliers in this variable as anticipated due to very old age customers having age more than 100 years. 40 to 60 years of age bracket having half of the customers.

Income

Min	1st Qtl.	Med	Mean
4.381	5.033	5.222	5.213
3rd Qtl.	**Max**	**Sd**	**Var**
5.402	7.956	0.286	0.082

Table 4: Statistical Summary - Log 10 of Income

This variable has annual income of the customers. This variable is part of Financial worth of the customer. Problem with this variable is that it is highly skewed to upper tail due to the presence of very high-income customers. Therefore, to conduct univariate analysis we have to transform it to LOG10 so that we can extract few insights:

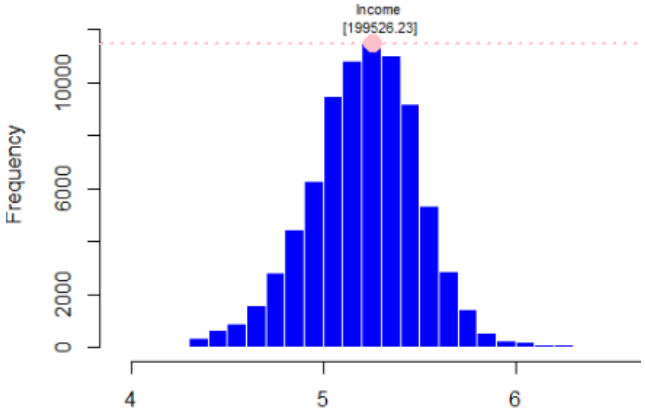


Figure 9: Univariate - Histogram of INCOME

The highest count of 11,446 is at income level of about two hundred thousand.

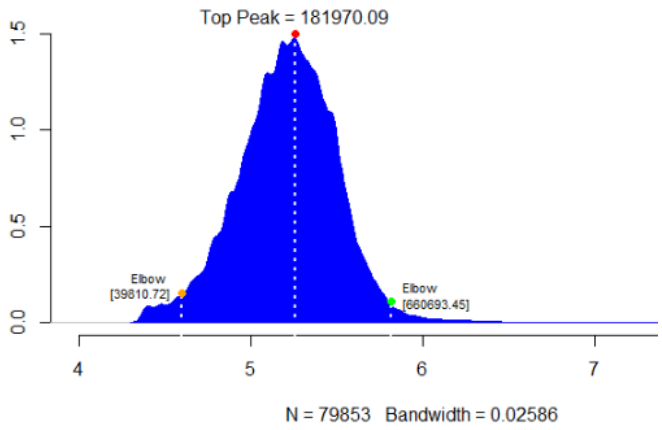


Figure 10: Univariate - Density Plot of INCOME

Despite being skewed variable, this variable has nice bell curve first 99 percentiles. We can see exactly three points where the distribution changes its direction and these income levels are 39,810 where density increases sharply, then at 181,970 have a peak from where again distribution started to nose dive till it reaches to elbow at 660,693.

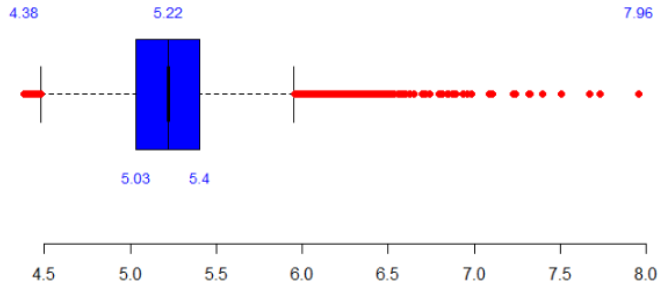


Figure 11: Univariate - Boxplot of INCOME

Just look at the boxplot of the variable and you can find outliers on both tails. However, number of outliers on upper tail by far outcounts lower tail outliers. Just visualize the spread of the outliers in income variable spread of the 99th percentile is about quarter of the spread between both ends of the outliers.

VEH_OWNED

Min	1st Qtl.	Med	Mean
1.00	1.00	2.00	1.99
3rd Qtl.	**Max**	**Sd**	**Var**
3.00	3.00	0.817	0.668

Table 5: Statistical Summary - VEH_OWNED

This variable has number of vehicles of particular customer owns. There are three values in the variable (1 to 3). This variable is part of financial worth dimension of the customer.

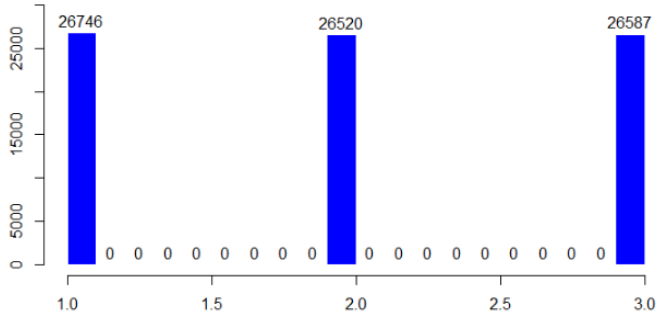


Figure 12: Univariate - Histogram of VEH_OWNED

Almost equal count of customers in all vehicle ownerships. Nothing to choose from any class of the variable.

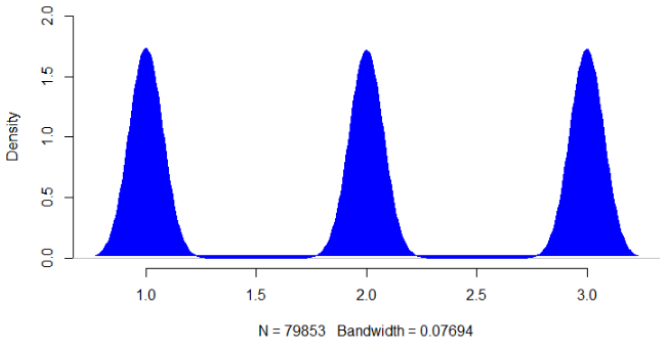


Figure 13: Univariate - Density of VEH_OWNED

Same density levels as evident in histogram as well. This variable is discrete in nature as per density plot having distribution with spikes.

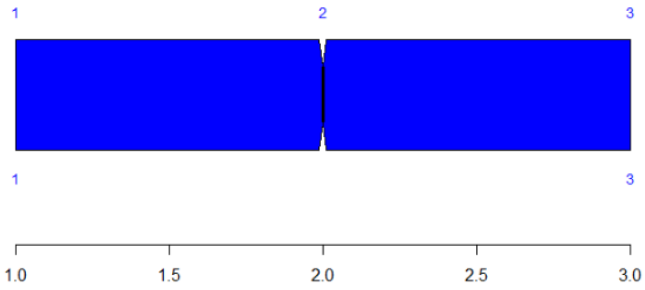


Figure 14: Univariate - Boxplot of VEH_OWNED

Overlapping minimum with 25th percentile and maximum with 75th percentile. Median is at two vehicles owned.

3To6 Late

Min	1st Qtl.	Med	Mean
0.00	0.00	0.00	0.24
3rd Qtl.	**Max**	**Sd**	**Var**
0.00	13.00	0.691	0.477

Table 6: Statistical Summary - 3To6 Months Late

This variable stores the count of the late payments if it is performed within 3 to 6 months after due date. As anticipated 95% of the customers are at zero late payments. 5% of the customers usually pays within 3 to 6 months after the due date.

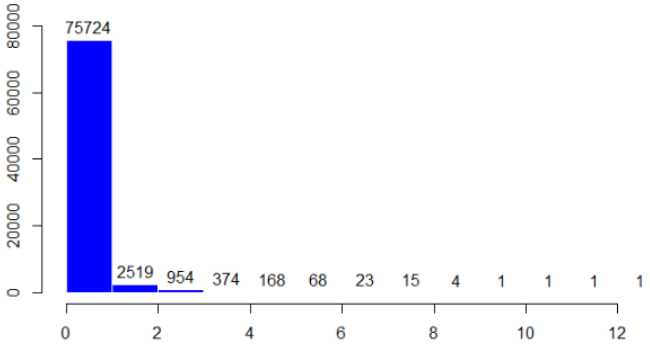


Figure 15: Univariate - Histogram of 3TO6_LATE

It is interesting to note that there are customers who did late payments more than 12 times without defaulting.

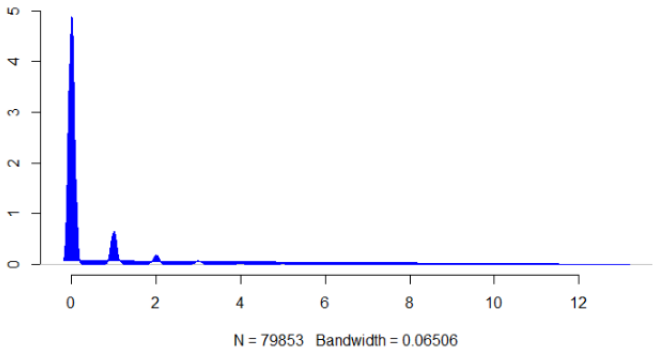


Figure 16: Univariate - Density of 3TO6_LATE

Discrete variable as per the density plot with highly skewed distribution. This shows that effectively this variable is any outlier because 95 percentiles are at 0 where this variable does not have any value.

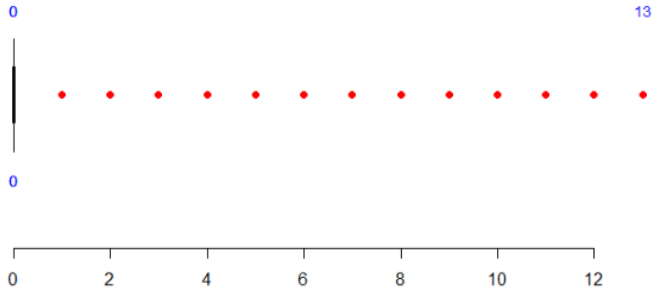


Figure 17: Univariate - Boxplot of 3TO6_LATE

After looking at the boxplot it becomes clear that this variable can only have outliers as it operates with the last 5 percentiles of the distribution. This is logical spread as most of the customers will not delay their premium payments thus always remain with zero late payment.

6TO12 LATE

Min	1st Qtl.	Med	Mean
0.00	0.00	0.00	0.07
3rd Qtl.	**Max**	**Sd**	**Var**
0.00	17.00	0.43	0.19

Table 7: Statistical Summary - 6To12 Months Late

Notice that this variable has less standard deviation than 3to6 months late due to less variance. This variable stores the count of the late payments if it is performed within 6 to 12 months after due date.

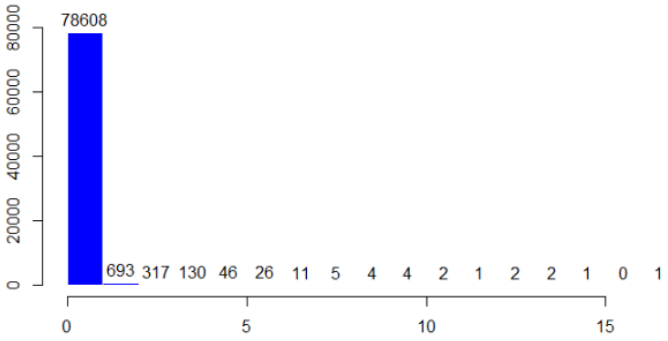


Figure 18: Univariate - Histogram of 6TO12_LATE

This variable has 98.4% of the customers at zero late payments. 1.6% of the customers usually pays within 6 to 12 months after the due date. It is interesting to note that there are customers who did late payments more than 15 times without defaulting.

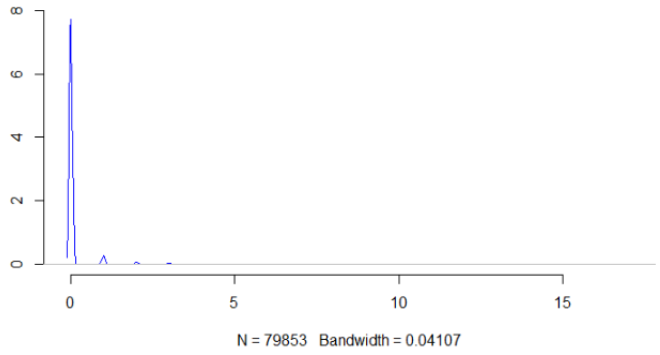


Figure 19: Univariate - Density of 6TO12_LATE

This variable again is discrete and having even more smaller minor class of having late payments more than zero.

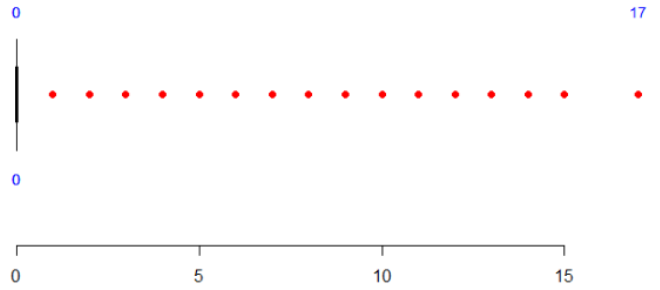


Figure 20: Univariate - Boxplot of 6TO12_LATE

As anticipated same behaviour as earlier late payment variable where variable more than zero can only apply to 1 percentile of the customer.

MORE12_LATE

Min	1st Qtl.	Med	Mean
0.00	0.00	0.00	0.05
3rd Qtl.	**Max**	**Sd**	**Var**
0.00	11.00	0.31	0.09

Table 8: Statistical Summary - More than 12 Months Late

Again, there is less standard deviation due to less variance than 6to12 months late. Thus, we can conclude that more than 12 months late premium payment has the least spread. Moreover, range is smaller as well if compared with other two late payment variables. This variable stores the count of the late payments if it is performed later than 12 months following due date.

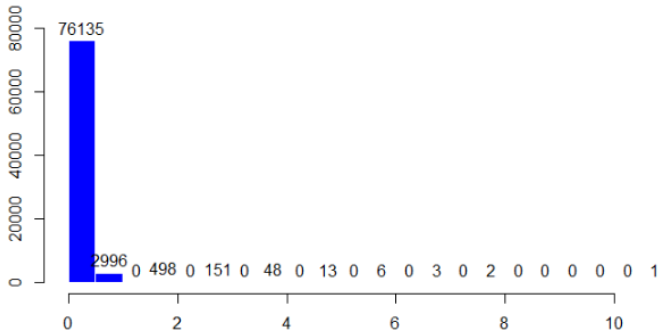


Figure 21: Univariate - Histogram of MORE12_LATE

This variable has 95% of the customers at zero late payments. 5% of the customers usually pays later than 12 months after the due date. It is interesting to note that there are customers who did late payments more than 10 times without defaulting.

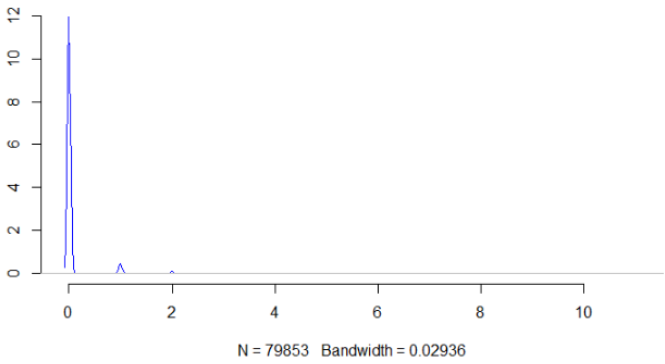


Figure 22: Univariate - Density of MORE12_LATE

This variable again is discrete and having even about the same size of minor class as of 3TO6_LATE.

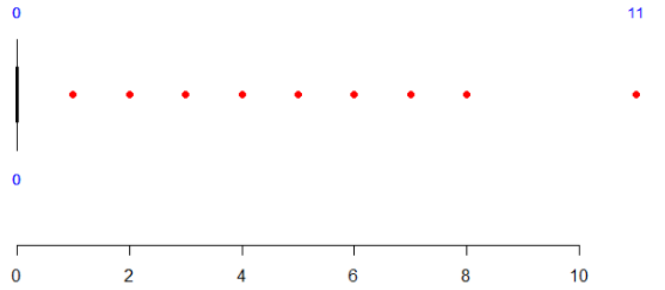


Figure 23: Univariate - Boxplot of MORE12_LATE

As anticipated same behaviour as earlier late payment variable where variable more than zero can only apply to 5% of the customer.

RISK_SCORE

Min	1st Qtl.	Med	Mean
91.90	98.83	99.18	99.07
3rd Qtl.	**Max**	**Sd**	**Var**
99.52	99.89	0.72	0.52

Table 9: Statistical Summary - Risk Score

We assume that risk score or insurance score is the metric or variable which determine the rate of premium to be charged to customer depending upon several predictors. This score keeps on changing with the time depending on the activities such as late payment or increasing payment history etc. It will be interesting to see what impact this variable has on default because usually with higher premium rate customer changes their service provider. We will see whether this holds true in this dataset or not.

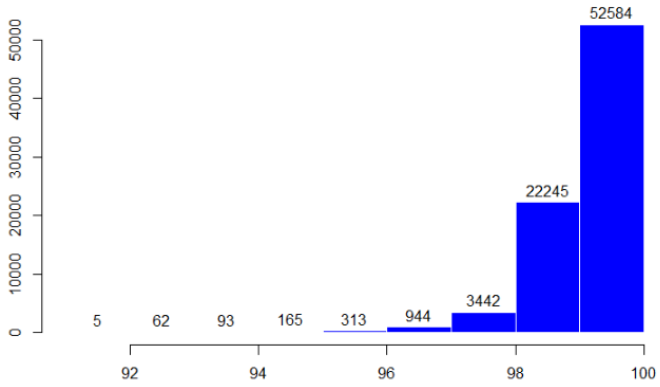


Figure 24: Univariate - Histogram of RISK_SCORE

Most of the customers are at the upper tail at maximum of score 100%.

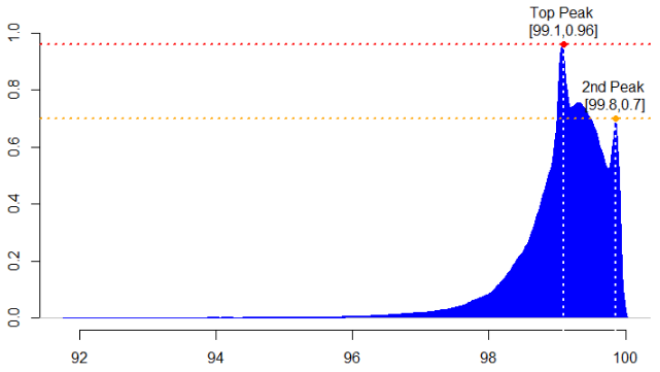


Figure 25: Univariate - Density of RISK_SCORE

We can find several peaks at 99.1% and 99.8% and inflex point at 98% where the distribution flattens with minimal count of customers.

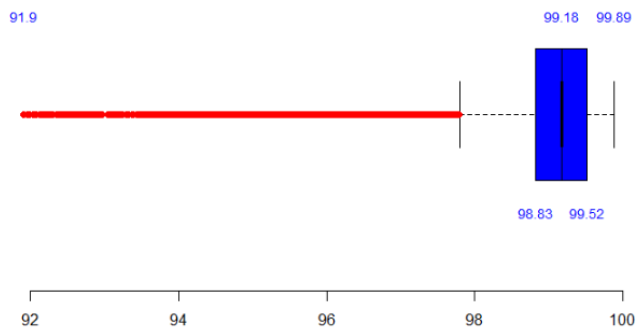


Figure 26: Univariate - Boxplot of RISK_SCORE

There are several outliers on the lower tail of the distribution. As shown in earlier plots as well that majority of the data is between 98 to 100 percentages. This would mean new customer ratio is more than old customers.

NO_PREMIUMS

Min	1st Qtl.	Med	Mean
2.00	7.00	10.00	10.86
3rd Qtl.	**Max**	**Sd**	**Var**
14.00	60.00	5.17	26.73

Table 10: Statistical Summary - Number of Premiums

It appears that premiums are being charged at the frequency of annual payments. 75% customers are less than 14 years old.

This variable stores the number of premiums paid and it updates on every payment. This can be classified in Payment History dimension.

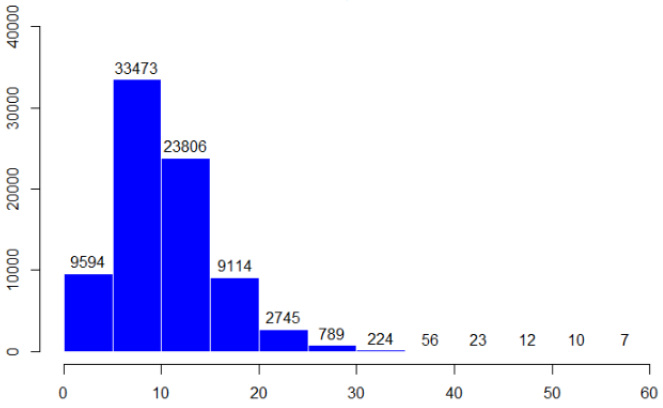


Figure 27: Univariate - Histogram of NO_PREMIUMS

Most of the customers have paid 5 to 10 premiums and their count is about 42% of the total customers.

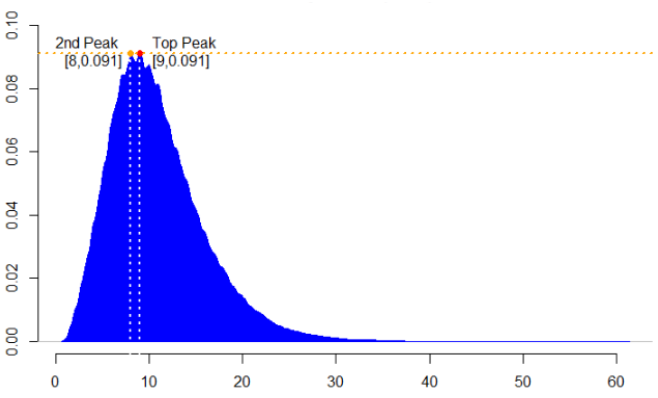


Figure 28: Univariate - Density of NO_PREMIUMS

Clearly, we can see that there are equal peaks at premium number 8 and 9. There is very long upper tail of the distribution.

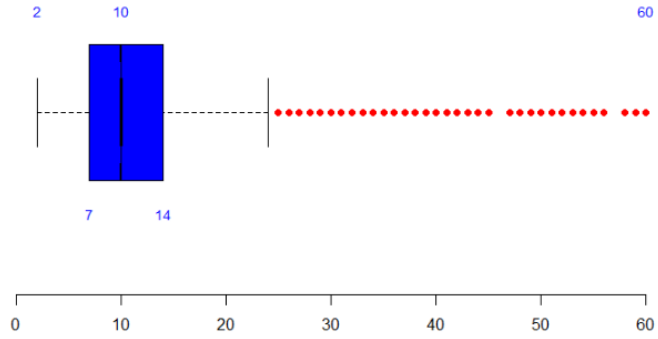


Figure 29: Univariate - Boxplot of NO_PREMIUMS

There are several outliers on the upper tail after 25 number of premiums.

PREMIUM

Min	1st Qtl.	Med	Mean
1,200	5,400	7,500	10,925
3rd Qtl.	**Max**	**Sd**	**Var**
13,800	60,000	9401.67	88,391,522

Table 11: Statistical Summary - Premium Amount

This variable stores the total premium amount paid so far. This variable updates on each and every payment. There are most of the customers who have paid somewhere in between 5,000 to 10,000 premium amounts.

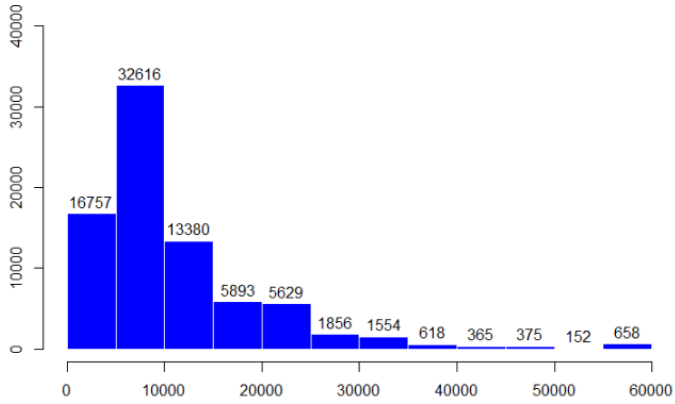


Figure 30: Univariate - Histogram of PREMIUM

Maximum number of customers are at precisely 5,500 premiums amount as shown in the density plot and histogram. Again this is logical as most of the customers are new in the clientele therefore their premium amount should have been less. It would have been interesting to guage customer satisfaction and its effect on default because it seems look at this data that most of the customer are not satisfied once crossed 14 years.

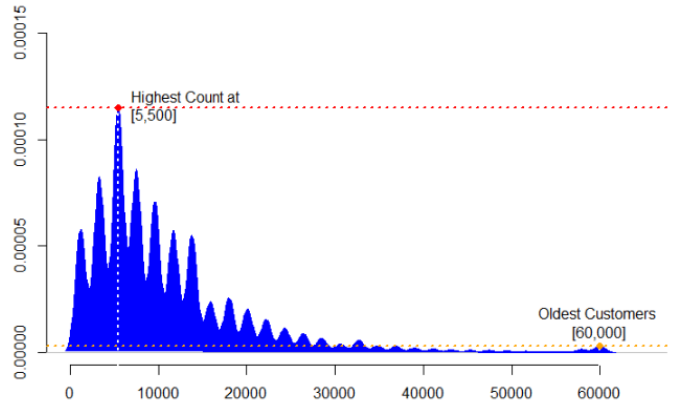


Figure 31: Univariate - Density of PREMIUM

There are oldest customers having paid 60,000 premiums.

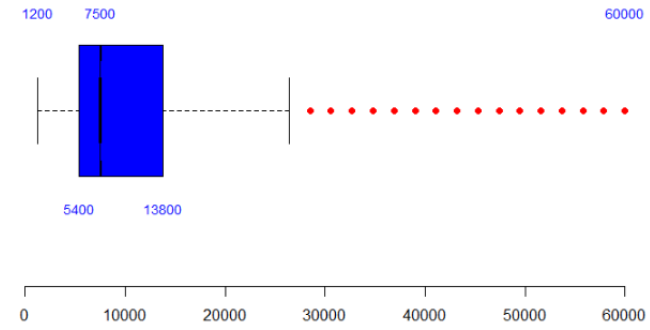


Figure 32: Univariate - Boxplot of PREMIUM

There are few outliers on the upper tail of the distribution. Most of the customers are within the range of 0 to 27,000 premium amount.

DEPENDANTS

Min	1st Qtl.	Med	Mean
1.00	2.00	3.00	2.5
3rd Qtl.	Max	Sd	Var
3.00	4.00	1.115	1.245

Table 12: Statistical Summary - Dependants

This variable shows total number of dependants on the customers. Again, this variable updates on change in circumstances of the customer.

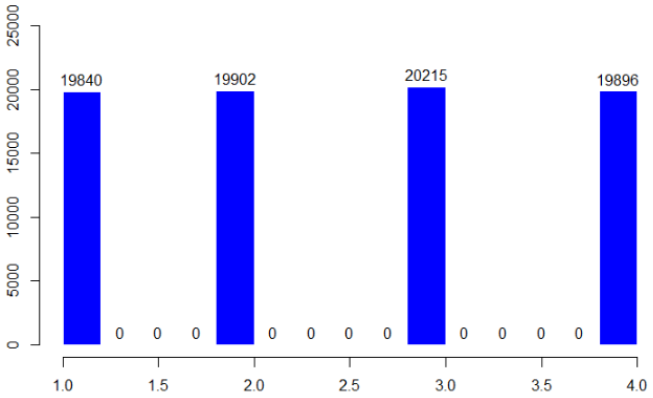


Figure 33: Univariate - Histogram of DEPENDANTS

There is almost a balance in all four values of the variable with minimal variation which can hardly be seen in the plot above. Once could argue that people with more dependants should go for insurance more than people with less dependants but here this is not the case certainly.

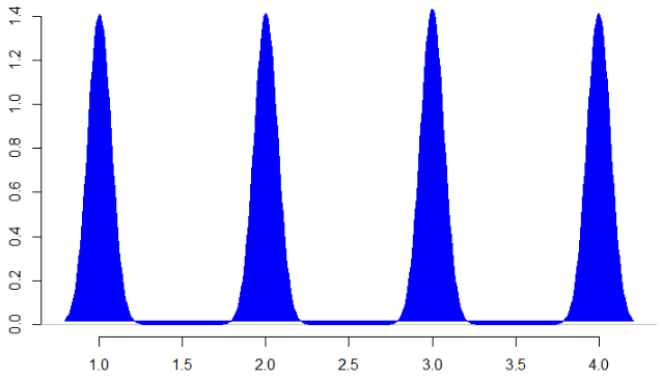


Figure 34: Univariate - Density of DEPENDANTS

Same densities of all values of dependants.

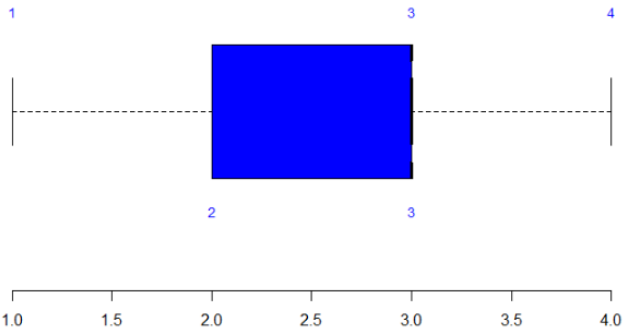
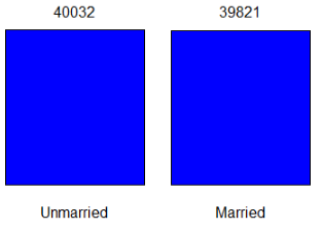


Figure 35: Univariate - Boxplot of DEPENDANTS

There are no outliers in this variable. There is skewness to the left implying that there are more customers with more than 2 dependants.

MARITAL

This variable shows marital status of the customer. We can see there is almost equal number of classes in the variable, which irrespective of their marital status customers are going for insurance coverage. Usually, married people are more concerned for security of their dependants but here we fail to notice that there is same concern in both classes of the customers. Would be interesting to see whether this variable has any impact on the default or not.

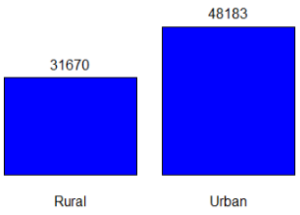


ACCOM

This variable shows whether the customer owns his residence or living in rented property. Again, startling similarity with previous variable where there is equal distribution in both classes. It was anticipated that customers with owned properties might be more concerned with the damage of their properties therefore going for insurance but certainly this is not the case here. May be this data is not of Home Insurance product.

AREA_TYPE

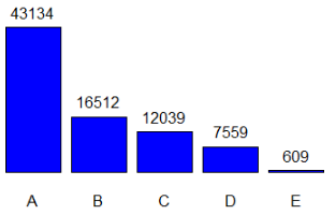
Most of the customer are from Urban areas and some are from Rural areas. There is a split of 60 to 40 percent.



This shows that urban customers prefer more going for insurance then rural customers. It may be due to more population density in the urban areas. May be education variable might be good variable to consider if we look at this variance.

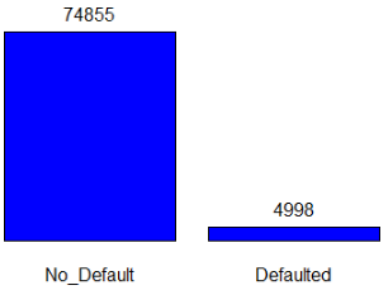
CHANNEL

The acquisition channels have variance with Channel A having most customers more than half and all other combined have less than this channel. Channel E has the least number of customers. It might be interesting to see impact of default on these channels. Although if proper screening process is employed then this should not affect on defaults.



DEFAULT

First impression of the distribution is imbalance classes of the variable. "Defaulted" is the minor class with 6% share of the total customers. Similarly, "No-Default" class is being the major class with 94% representation in the dataset. Since this variable is our RESPONSE variable therefore let's discuss it further some other aspects which could impact further analysis. As it is clear that this binary variable therefore, we will employ binary classification algorithms.



Binary classification algorithms are developed in a way that they look for overall accuracy metric for improvement in classification. Now if majority class represents more than 90% of the data then even the worst kind of classification model spit out classes with overall classification accuracy of 90%. But our objective is to detect MINOR CLASS more accurately then MAJOR class therefore, we are interested in SENSITIVITY more than ACCURACY. Secondly, we need to resample the dataset so that we force algorithms to identify both classes with equal importance.

Lastly, we have changed the value of "Defaulted" to "1" as previously "0" was assigned to minor class which was counter intuitive.

Correlation Analysis

Let's analyse all variable correlations for further analysis. First of all, we need to convert character based categorical variables to numeric variables in order to proceed.

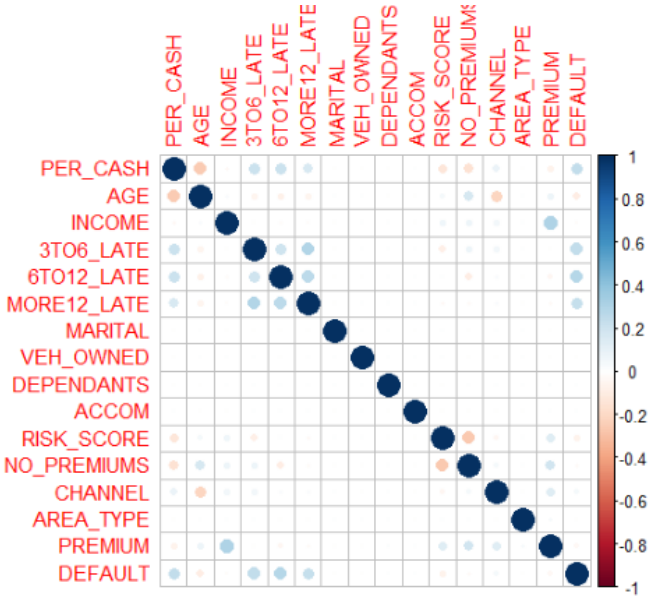


Figure 36: Correlation Matrix

We can see that in moderately significant correlations are in default variable. Correlation matrix reveal below pairs worth further analysis:

Important Predictors

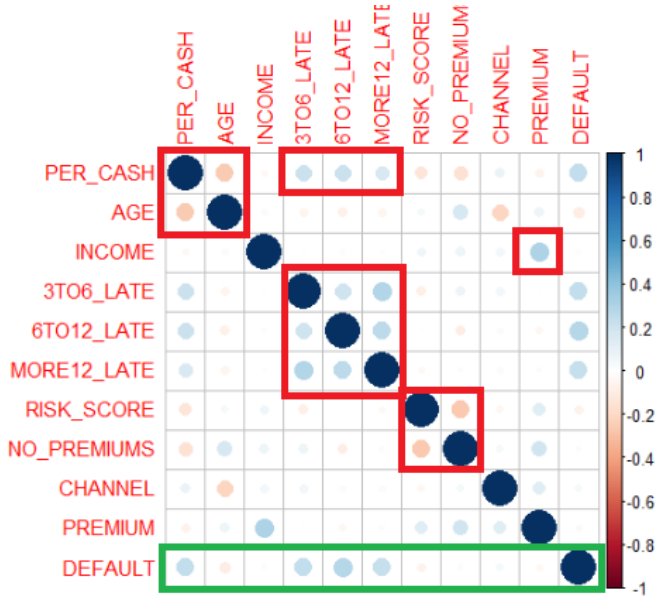
1. DEFAULT ~ 3TO6_LATE (Positive)
2. DEFAULT ~ 6TO12_LATE (Positive)
3. DEFAULT ~ MORE12_LATE (Positive)
4. DEFAULT ~ PER_CASH (Positive)
5. DEFAULT ~ AGE (Negative)
6. DEFAULT ~ RISK_SCORE (Negative)
7. DEFAULT ~ PREMIUM (Negative)
8. DEFAULT ~ CHANNEL (Negative)

These are the important predictors for default. Please note that all "Late Payment" variables with some "Payment History" variables are correlated with default more than any other dimensions identified earlier. RISK SCORE is also correlated with default at moderate level. Weak correlations can be noted in AGE, PREMIUM and CHANNEL.

Multicollinearity Prospects

- PREMIUM ~ INCOME
- AGE ~ PER_CASH
- PER_CASH ~ 3TO6_LATE
- PER_CASH ~ 6TO12_LATE
- PER_CASH ~ MORE12_LATE
- PREMIUM ~ NO_PREMIUMS
- RISK_SCORE ~ NO_PREMIUMS
- 3TO6_LATE ~ 6TO12_LATE
- 3TO6_LATE ~ MORE12_LATE
- 6TO12_LATE ~ MORE12_LATE

These variables are having correlations among themselves therefore there is a need to check for multicollinearity among them. Please note multicollinearity does not affect predictive ability but takes away descriptive power from the model. Therefore, it is important to detect multicollinearity and treat it if it is detecting above defined thresholds.



Red boxes are the area of concern for multicollinearity whereas, GREEN box shows the promising predictors for detecting default. Now proceed to next section to filter to validate important variables highlighted here in correlation analysis.

Regression Analysis

After correlation analysis where we have identified important variables which can predict default variation. Now there is a need to sort the list of variables as per their importance in determining defaulted cases. Moreover, being second step, it will validate the results of the correlation analysis as strength of correlation is not significant in many variables, therefore where to put threshold for selecting variable is difficult. Therefore, we will perform logistic regression on the dataset to determine what are the variable which are significant predictors of defaulted cases.

Our regression model is going to be binomial logistic regression due to the nature of the response variable. Please note that we have already switched the values from 0 to 1 for defaulted cases therefore if variable is predicting default cases as the value increases then there is going to be positive coefficient, and on the contrary it is going to be negative coefficient. We will pick only those predictors whose significance is more than 99%.

Please find below table of coefficients of logistic regression we conducted:

Table 13: Regression Analysis - Logit Coefficients

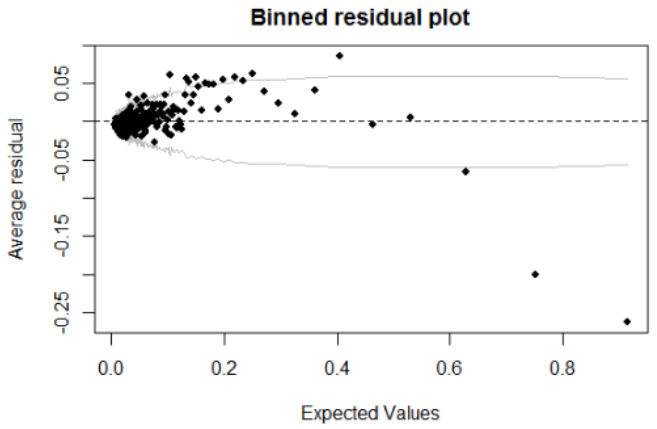
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.40E+01	2.03E+00	6.895	5.40E-12	***
PER_CASH	1.87E+00	4.87E-02	38.416	< 2e-16	***
AGE	-4.41E-05	3.65E-06	-12.08	< 2e-16	***
INCOME	-4.31E-07	1.23E-07	-3.498	0.000468	***
`3TO6_LATE`	4.22E-01	1.57E-02	26.902	< 2e-16	***
`6TO12_LATE	6.93E-01	2.47E-02	28.112	< 2e-16	***
MORE12_LATE	6.28E-01	3.30E-02	19.05	< 2e-16	***
RISK_SCORE	-1.76E-01	2.03E-02	-8.654	< 2e-16	***
NO_PREMIUMS	2.45E-02	3.53E-03	6.943	3.85E-12	***
CHANNEL	4.50E-02	1.50E-02	3.009	0.002618	**

It is important to note that as per correlation analysis NO_PREMIUM, was not correlated to DEFAULT variable but here in regression analysis it becomes significant in predicting response variable of DEFAULT. Similarly, there PREMIUM was showing more correlation but here its significance is negligible.

Residuals:

Min	1Q	Median	3Q	Max
-4.2109	-0.333	-0.2293	-0.182	3.6471

```
AIC of the model is 29818
```



Gray lines showing +/- 2 SE bands and is containing about 95% of the observations with in them. Therefore, model seems to be fitted reasonably good.

Now examine whether there is any multicollinearity in the model through VIF function of package CAR:

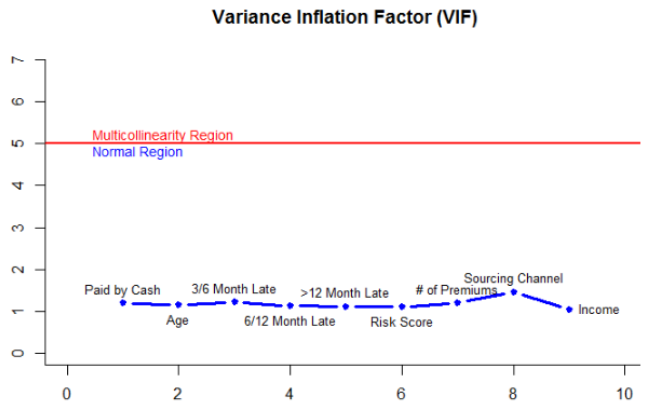


Figure 37: Regression Analysis - Variance Inflation Factor

We can clearly see that all predictors have less than 2 VIF which means there is no risk of multicollinearity in the regression model. We will not go to the level of accuracy, sensitivity and specificity as our intention at this stage is to pick important features for further analysis of bivariate.

Bivariate Analysis

Now let us examine variable pairs in the light of correlation and regression analysis to find insights in the data:

PER_CASH Vs. DEFAULT

As per correlation analysis PER_CASH has 24% positive correlation with DEFAULT. Let us evaluate binned interaction with DEFAULT to gauge what slots are more at risk.

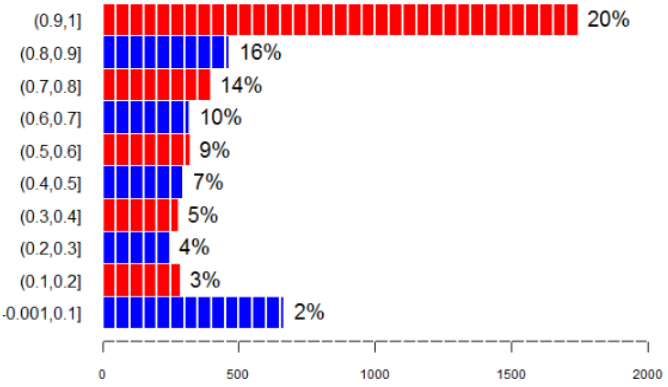


Figure 38: Bivariate - Bin Analysis of PER_CASH Vs DEFAULT

There is steady increase in the default ratio as the percentage cash increases. However, between 70% to 100% jump in ratio is sudden ranging from 14% to 20%. Therefore, we can say that customers paying premiums through cash more than 70% of times are more susceptible to default.

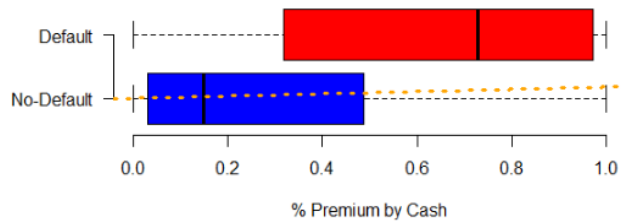


Figure 39: Bivariate: Boxplot of PER_CASH Vs DEFAULT

Bivariate boxplot shows positive correlation without outliers therefore correlation among this pair is certain. Looking at the plot we can see that CASH PAYERS default more often then CREDIT PAYERS. Now let us examine the statistical summary of the interaction to ascertain the findings of the correlation:

	Default	No Default
Minimum	0.000	0.000
1st Quantile	0.317	0.031
Median	0.728	0.148
3rd Quantile	0.971	0.487
Maximum	1.000	1.000
C.I.	0.713 - 0.742	0.145 - 0.150
Correlation	0.2409802	

Just see the variance in medians of both classes. In defaulted cases median is at 72% whereas in no default it sits at 14.8%.

AGE Vs. DEFAULT

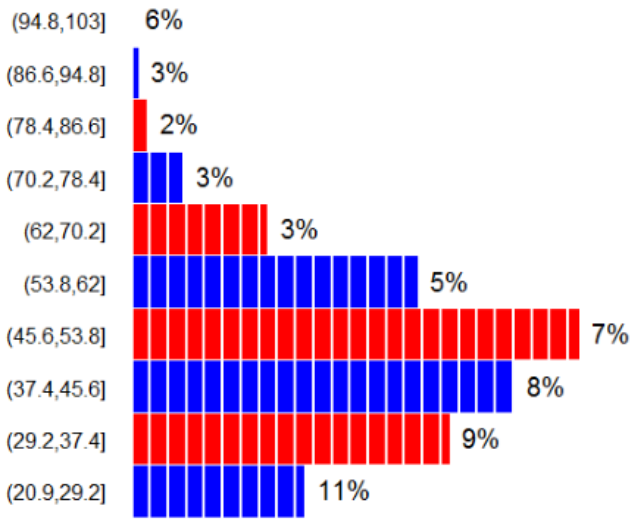


Figure 40: Bivariate - Bin Analysis of AGE Vs. DEFAULT

It is clear from the figure above that customers having age between 21 to 29 years tends to default more than any other slot. There is gradual decrease in default as the age increases.

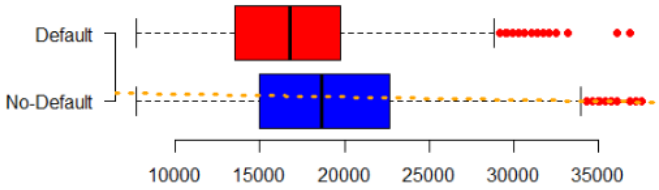


Figure 41: Bivariate - Boxplot of AGE Vs. DEFAULT

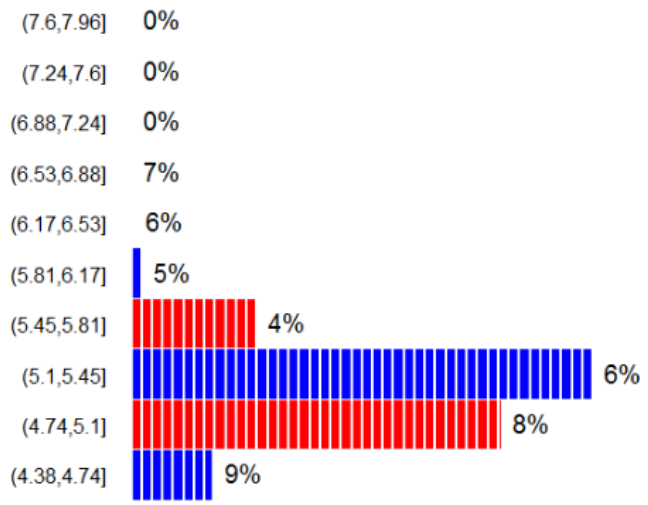
We can see there are outliers on upper tail of both classes. Range of defaulted classes reduces than no-default.

	Default	No Default
Minimum	7,671	7,670
1st Quantile	13,511	14,978
Median	16,800	18,630
3rd Quantile	19,725	22,640
Maximum	28,850	33,960
C.I.	16.6K-16.9K	18.5K-18.6K
Correlation	-0.09510293	

Please note in the above table age values are in days that is the reason they are in thousands. Just note the difference in the median of both the classes. Median decreased from 18,630 to 16,800 for defaulted class. Correlation strength of NEGATIVE 1%.

INCOME Vs. DEFAULT

Since INCOME variable is highly skewed therefore, we will use LOG10 transformation on this variable and then compare it with DEFAULT variable.



Again there is negative correlation implying that more income means less default cases therefore income levels above 6.88 (7.5 Million) has no default at all. Whereas, income band of 5.1 (125,892) to 6.88(7.5 Million) has acceptable level of default cases which is less than 6% (as default representation in the total dataset is 6%). However, it should be noted that high risk slot is between 4.38 (23,988) to 5.1 (125,892) having around 8.5% of default cases.

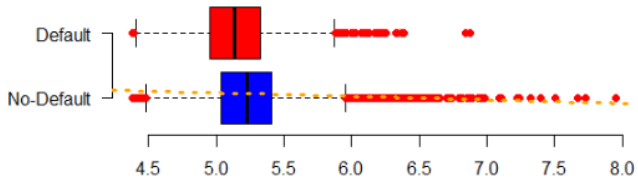


Figure 42: Bivariate - Boxplot of INCOME Vs. DEFAULT

Very weak negative correlation among this pair with lots of outliers. However, we can visually notice the movement of medians in both classes. Let us examine the statistical summary of the pair:

	No Default	Default
Minimum	4.485295	4.40841
1st Quantile	5.037944	4.954725
Median	5.227553	5.140194
3rd Quantile	5.406625	5.323726
Maximum	5.95918	5.875137
C.I.	5.225-5.229	5.131-5.148
Correlation	-0.06263611	

There is a negative correlation of about 6% with slight reduction of median for default cases.

3TO6_LATE Vs. DEFAULT

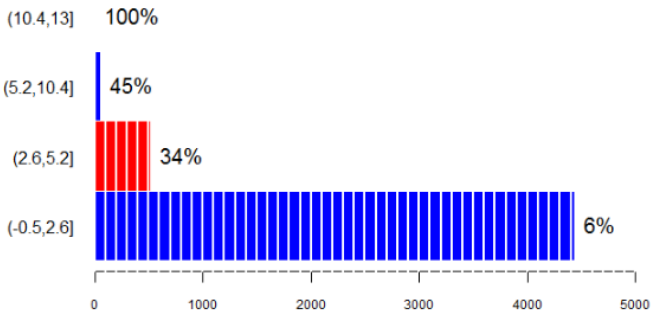
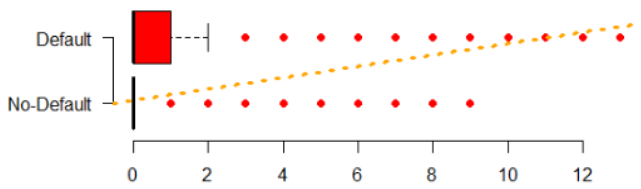


Figure 43: Bivariate - Bin Analysis of 3TO6_LATE Vs DEFAULT

We can classify the distribution into three categories more than 10 times late means certain default. If number of times of late payment is between 3 to 10 then there is

40% chance of default. However, delayed payment is less than 3 times then default rates are below acceptable level of 6%.

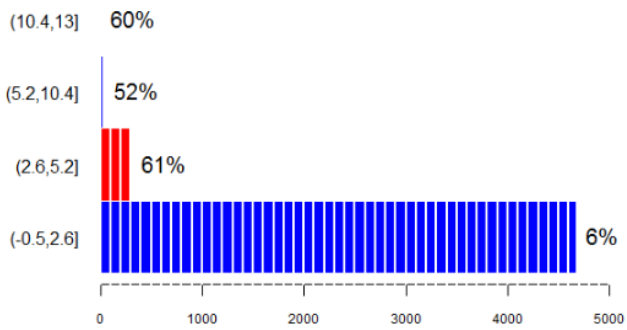


We can see there is significant positive trend in the plot shown here with many outliers. In this variable outliers are natural because delayed payment will always be fraction of the total population.

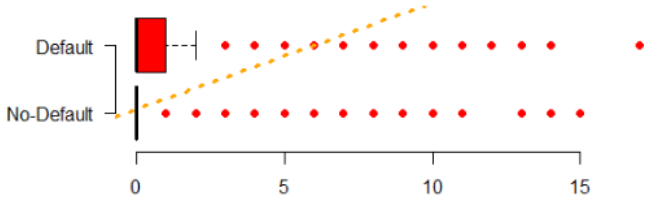
	Default	No Default
Minimum	0	0
1st Quantile	0	0
Median	0	0
3rd Quantile	1	0
Maximum	2	0
C.I.	-0.022-0.022	0
Correlation	0.2464685	

Looking at the five-number summary it is anticipated that most of the values will be zero due to the nature of the variable. Please note the strength of the correlation which is POSITIVE 24.6%.

6TO12_LATE Vs. DEFAULT



We can slice this variable distribution into two portions which is different from previous late payment variable. If number of late payments are less than 3 then default ratio is with-in acceptable range of 6%. However, if it is more than default ratio skyrocketed to around 60%. Due to this behaviour of the variable it is key variable especially for tree-based algorithms.

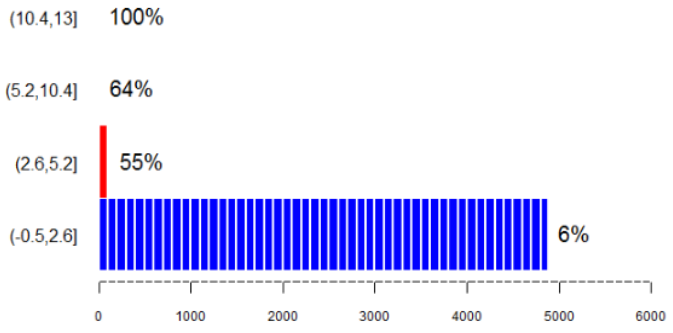


It can be clearly seen that trend line is more steep then 3to6 months late. Again as discussed before outliers are natural constituents of the variable.

	Default	No Default
Minimum	0	0
1st Quantile	0	0
Median	0	0
3rd Quantile	1	0
Maximum	2	0
C.I.	-0.022-0.022	0
Correlation	0.2840336	

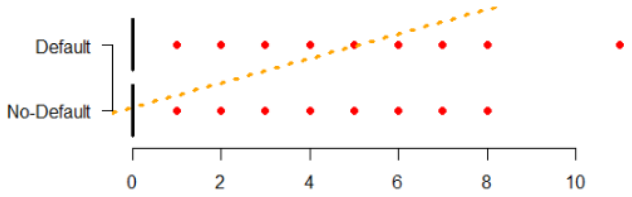
This variable has even stronger correlation with default than 3to6 months late variable.

MORE12_LATE Vs. DEFAULT



This variable can be categorised into three slots like 3to6 months late variable. If number of late payments are

more than 10 then 100% certain default case. However, if it is between 3 to 10 then around 60% default certainty. Finally, if number of late payments are less than 3 times then acceptable ratio of 6%. It seems that this variable has concatenated effect of previous two late payment variables if you notice it closely.

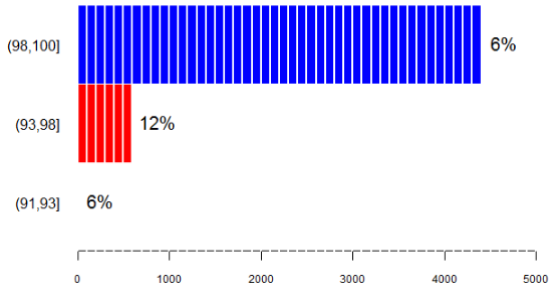


Steepness of the trend line seems to be in the middle of 3to6 months late and 6to12 months late. With all values to be outliers depicting even lesser observations in this variable than previous late payment variables.

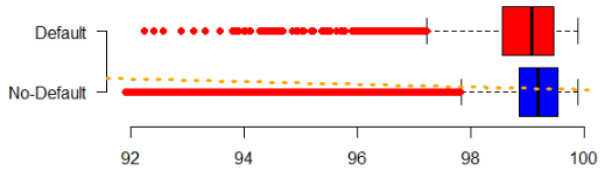
	Default	No Default
Minimum	0	0
1st Quantile	0	0
Median	0	0
3rd Quantile	0	0
Maximum	0	0
C.I.	0	0
Correlation	0.2384818	

Correlation is weaker than previous two late payment variables with all quartiles in zeros confirming that values of this variables are predominantly 0.

RISK_SCORE Vs. DEFAULT



This variable has somewhat peculiar distribution. We can divide this variable distribution into three classes. Classes at either tail of the distribution seems to have acceptable ratio of default cases which is 6%. However, in the middle values ranging from 93% to 98% default cases jump to 12%.

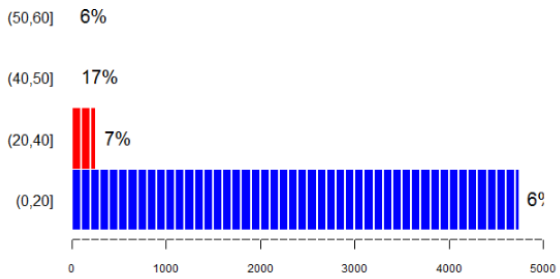


There are lot of outliers on the lower tail of the distribution. There is negative trend in the correlation. This variable shows that most of the customers in the clientele are new customers that is why 75 percentiles is on the upper tail between 98 to 100. Please note that more risk score means more risk and thus higher premium rate. As time passes risk profile gets low and premium rate decreases. Therefore, lower tail should have customers having long association with the insurance company.

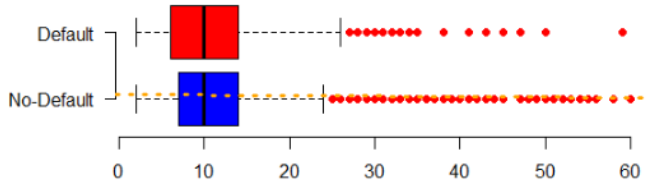
	Default	No Default
Minimum	97.22	97.83
1st Quantile	98.56	98.85
Median	99.066	99.19
3rd Quantile	99.46	99.53
Maximum	99.89	99.89
C.I.	99.04-99.08	99.18-99.19
Correlation	-0.0673339	

There is negative correlation of 6.7%. Range of the default is greater than no default.

NO_PREMIUMS Vs. DEFAULT



Number of premiums paid has the same bins as risk score where default ratio increases in the middle bin whereas it goes to acceptable ratio at upper and lower tails. Please note 40 to 50 has the highest ratio of 17% of default. It could be due to death of the customers because 40 annual payment means 45 years plus the years of age of customer remained minor around 20 years therefore average age of death could be 60 to 70 years.

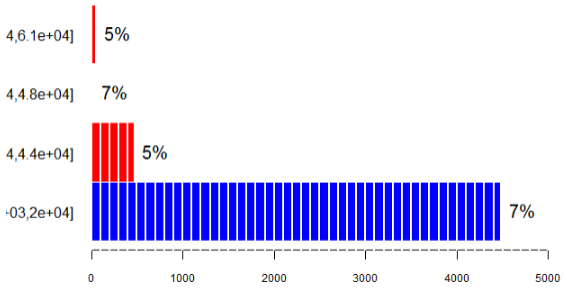


The trend line seems to be flat with slight tilt showing negative correlation.

	Defaulted	No Default
Minimum	2	2
1st Quantile	6	7
Median	10	10
3rd Quantile	14	14
Maximum	26	24
C.I.	9.82-10.17	9.95-10.04
Correlation	-0.02266427	

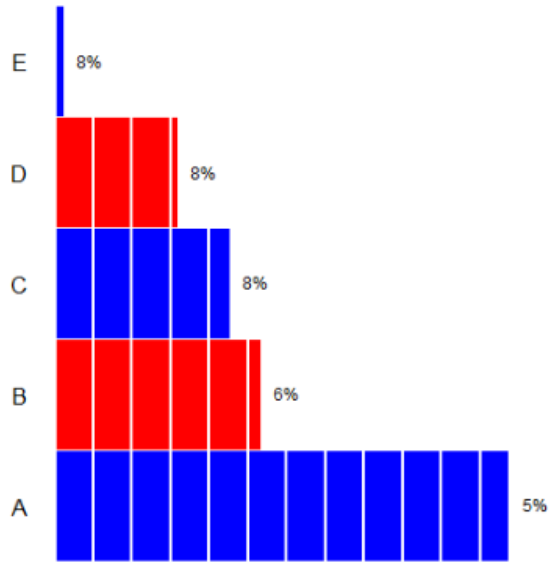
Looking at the statistical summary of the pair we can see nothing to choose from apart the correlation of negative 2%.

PREMIUM Vs. DEFAULT



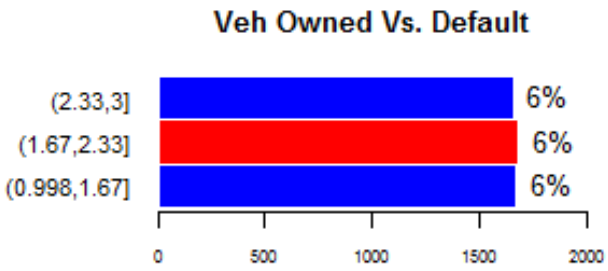
This variables seems to be very flat having having average 6% default rates all across with shade variations.

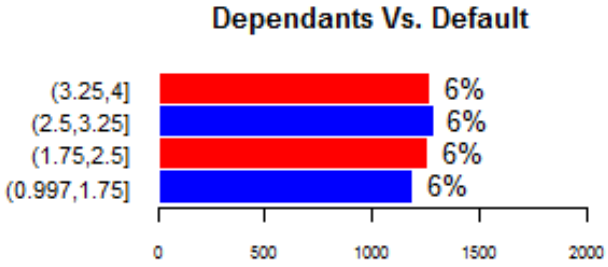
CHANNEL Vs. DEFAULT

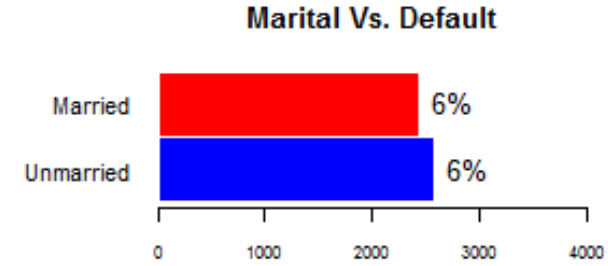


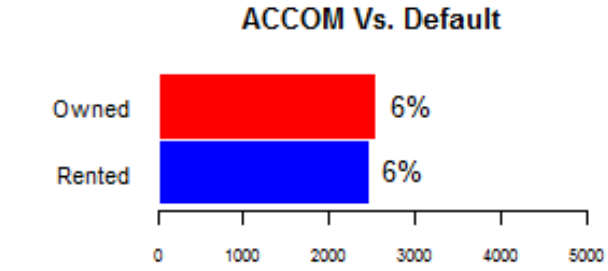
You can classify acquisition channels into two classes acceptable default rate (A, B) and higher default rates (C, D, E).

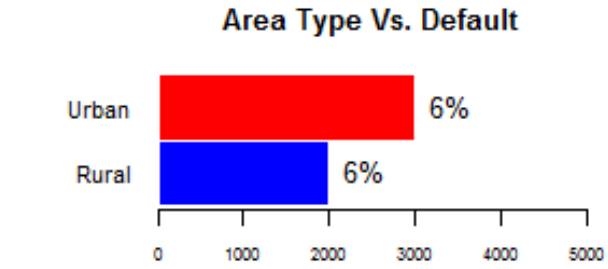
Now we have examined all important variables as identified in Correlation and Regression analysis. Just for once we will see weak variable interaction against default to see if there is important insight otherwise we will drop all following variables from further analysis:











You can see in all these plots that there is absolutely no variance in default due to variance in these variables which shows that interaction among these pairs are not significant thus these variables are left out of further analysis.

This concludes our bivariate analysis of the variables and we proceed for multivariate analysis in attempt to capture further insights.

Multivariate Analysis

Let us perform multivariate analysis on scatterplots with colour as a third dimension:

PREMIUM Vs. INCOME Vs. DEFAULT

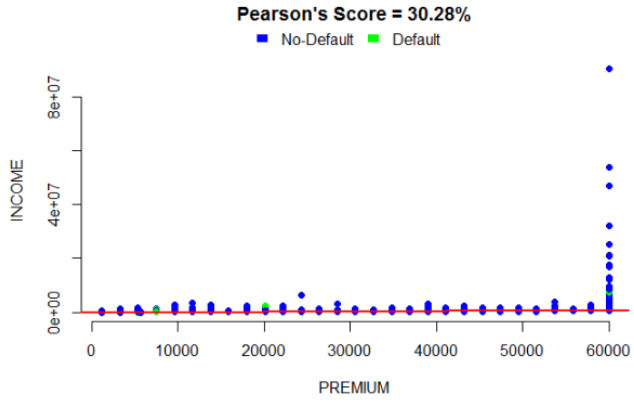


Figure 44: Multivariate PREMIUM Vs INCOME Vs DEFAULT

Correlation among INCOME and PREMIUM is 30.28%, which is significant in the dataset provided but we can see there is no effect of DEFAULT on them.

AGE Vs. PER_CASH Vs DEFAULT

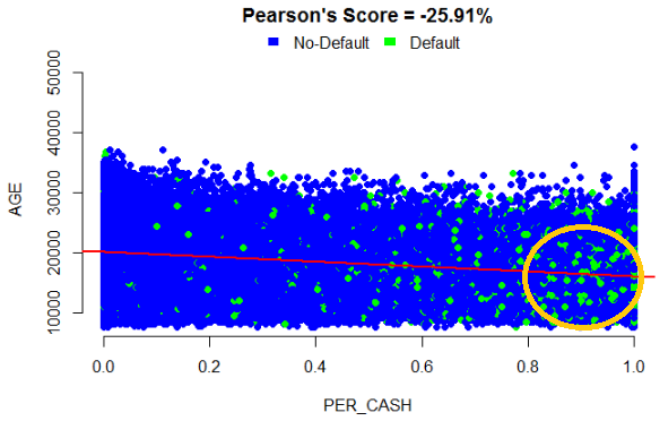


Figure 45: Multivariate AGE Vs DEFAULT Vs PER_CASH

This trio makes sense as you can see there is build-up of defaulted cases on bottom right quadrant of the plot, which is shown as brown circle. This shows that Lower age customer having maximum cash payment percentage forms default cases more than any other section of the distribution. Please note that correlation among AGE and PER_CASH is negative 25.9% but we have already established that this is not leading to multicollinearity in previous section.

NO_PREMIUMS Vs. PREMIUM Vs. DEFAULT

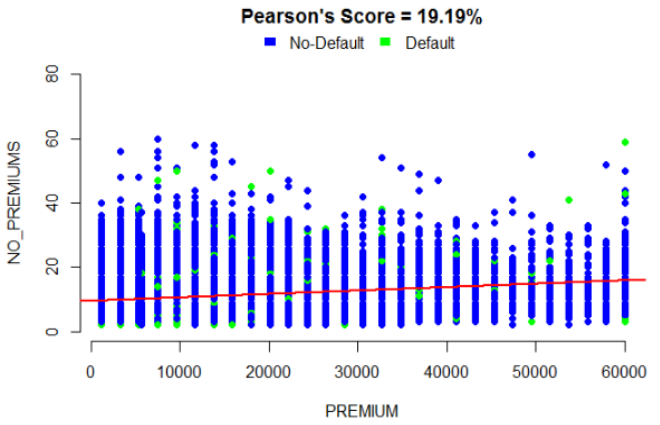


Figure 46: Multivariate PREMIUM Vs NO_PREMIUMS Vs DEFAULT

We can see although first two variables have correlation of about 20% but default seems to have no effect on them that is why green dots are evenly scattered across the plot.

RISK_SCORE Vs. NO_PREMIUMS Vs. DEFAULT

You can see in the plot that there is a build up of green dots at the center of the plot which is shown as brown square.

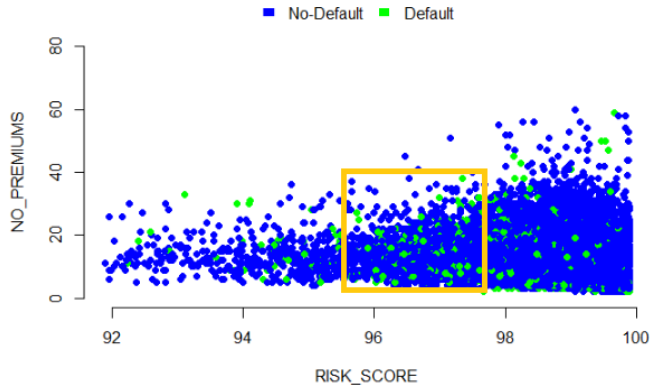


Figure 47: Multivariate RISK_SCORE Vs NO_PREMIUMS Vs DEFAULT

Again NO_PREMIUMS and RISK_SCORE has considerable correlation. As noted previously in bivariate analysis that interaction of Risk score and Number of Premiums paid is similar against default cases. This can be seen that right in the middle cluster of default cases build-up this implies that we may have to opt for tree based algorithm for predictive modelling.

PER_CASH Vs. 3TO6_LATE Vs. DEFAULT

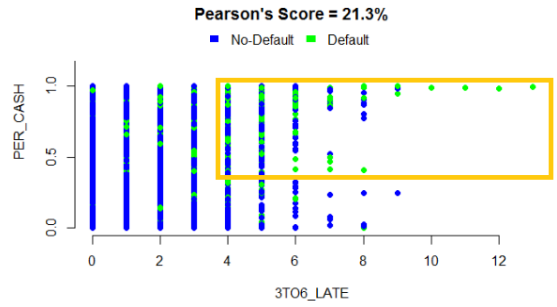


Figure 48: Multivariate PER_CASH Vs 3TO6_LATE Vs DEFAULT

We can see the cluster on the right top quadrant which is shown in the rectangle for majority defaulted cases. This pair justifies lower cash percent default as well.

PER_CASH Vs. 6TO12_LATE Vs. DEFAULT

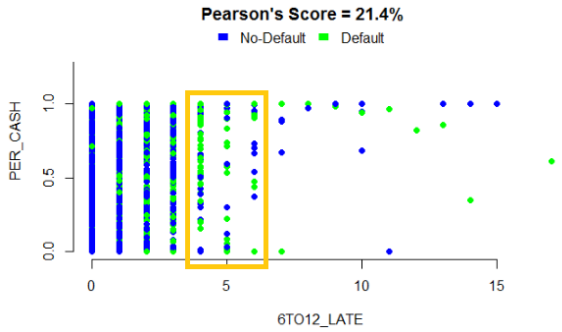


Figure 49: Multivariate PER_CASH Vs 6TO12_LATE Vs DEFAULT

We can see the build up of green dots in the rectangle shown above, where majority class seems to be defaulted. On higher count of late payment although green is at par with no-default but volume is too low to predict pattern. Secondly, this combination justifies the reason of defaulted cases even on low cash percentage so this pair is very important for the model building.

PER_CASH Vs. MORE12_LATE Vs. DEFAULT

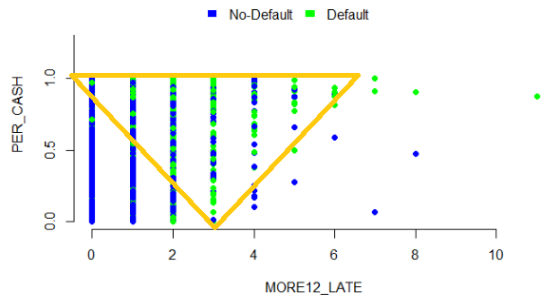


Figure 50: Multivariate PER_CASH Vs MORE12_LATE Vs DEFAULT

Green dots build up are in the shape of triangle. Explaining the pattern of late payment variables and cash percent variable. Interesting finding.

3TO6_LATE Vs. 6TO12_LATE Vs. DEFAULT

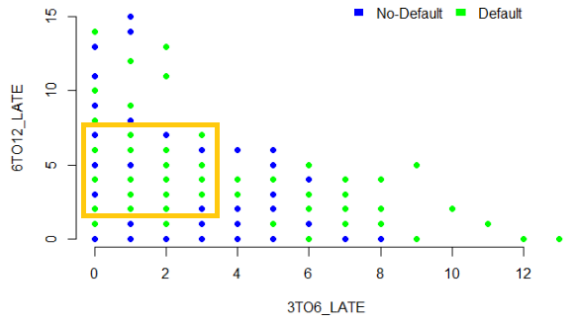


Figure 51:Multivariate 3TO6_LATE Vs 6TO12_LATE Vs DEFAULT

You can see majority of the cases in the square is default when late payments from both variables are more than 2. There are other sections as well where default dots can be seen in majority but their volume of data can be questioned.

MORE12_LATE Vs. 6TO12_LATE Vs. DEFAULT

In this plot you can see that once the count of more than 12 months late hits 2 then default rate shoots to significant levels of proportion and same trend continues on 3 count as well.

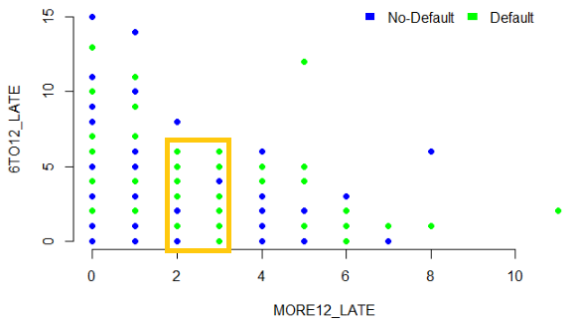


Figure 52: Multivariate 6TO12_LATE Vs MORE12_LATE Vs DEFAULT

However, on 4 and later randomness comes into the distribution.

MORE12_LATE Vs. 3TO6_LATE Vs. DEFAULT

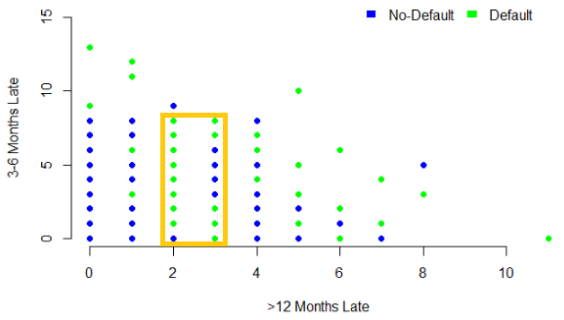


Figure 53: Multivariate 3TO6_LATE Vs MORE12_LATE Vs DEFAULT

Same pattern can be seen here as shown in previous variables.

CART (Multivariate Analysis Only)

Let us develop CART decision tree model for the purpose of multivariate analysis. Since, it is difficult to visualize multivariate having more than three dimensions therefore trees represent the interaction cleanly and in user friendly way.

Hyperparameter

Complexity parameter of the model to be selected is 0.0015 as shown in the figure below:

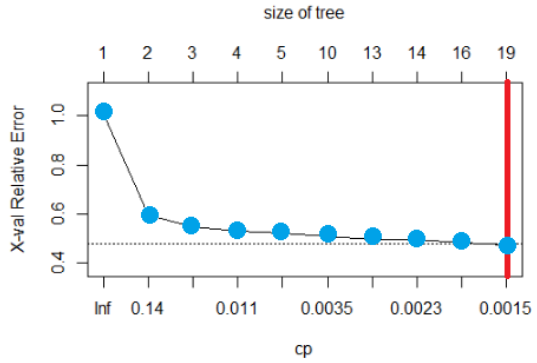


Figure 54: Complexity Parameter - CART Model

Path to DEFAULTED

Insight # 1: if payment is late more than 12 months once then it ends up in default even if cash payment is less than 40% and 3-6 Late and 6-12 Late is zero.

NODE # 17	
PER_CASH	T3TO6_LATE
< 40.45 %	< 0.5
T6TO12_LATE	MORE12_LATE
<0.5	>=0.5

Table 14: CART PATH DEFAULT # 1

Insight # 2: If payment is late 6-12 months then it ends up in default even if cash payment is less than 40% and 3-6 Late is zero.

NODE # 9	
PER_CASH	T3TO6_LATE
< 40.45 %	< 0.5
T6TO12_LATE	
>=0.5	

Table 15: CART PATH DEFAULT # 2

Insight # 3: If payment is late once between 3-6 months along with cash payment between 26.85-40.45% and RISK_SCORE more than 97.56% then customer defaults.

NODE # 161		
PER_CASH	T3TO6_LATE	T6TO12_LATE
< 40.45 %	>= 0.5	< 0.5
MORE12_LATE	T3TO6_LATE	RISK_SCORE
< 0.5	< 1.5	>= 97.56%
PER_CASH		
>26.85%		

Table 16: CART PATH DEFAULT # 3

Insight # 4: If payment is late once between 3-6 months along with cash payment less than 40.45% and RISK_SCORE less than 97.56% then customer defaults.

NODE # 81		
PER_CASH	T3TO6_LATE	T6TO12_LATE
< 40.45 %	>= 0.5	< 0.5
MORE12_LATE	T3TO6_LATE	RISK_SCORE
< 0.5	< 1.5	< 97.56%

Table 17: CART PATH DEFAULT # 4

Insight # 5: If payment is late more than once between 3-6 months along with cash payment less than 40.45% then customer defaults.

NODE # 41		
PER_CASH	T3TO6_LATE	T6TO12_LATE
< 40.45 %	>= 0.5	< 0.5
MORE12_LATE	T3TO6_LATE	
< 0.5	> 1.5	

Table 18: CART PATH DEFAULT # 5

Insight # 6: If payment is late in combination of 3-6 months late and more than 12 months late both along with cash payment less than 40.45% then customer defaults.

NODE # 21		
PER_CASH	T3TO6_LATE	T6TO12_LATE
< 40.45 %	>= 0.5	< 0.5
MORE12_LATE		

>= 0.5		

Table 19: CART PATH DEFAULT # 6

Insight # 7: If payment is late in combination of 3-6 months late and 6-12 months late, along with cash payment less than 40.45% then customer defaults.

NODE # 11		
PER_CASH	T3TO6_LATE	T6TO12_LATE
< 40.45 %	>= 0.5	>= 0.5

Table 20: CART PATH DEFAULT # 7

Insight # 8: When customer income is less than 187,000 and cash payment is between 52.65 and 82.35% having number of premiums paid more than 7.5 then customer defaults.

NODE # 775		
PER_CASH	T3TO6_LATE	T6TO12_LATE
>= 40.45 %	< 0.5	< 0.5
PER_CASH	MORE12_LATE	RISK_SCORE
< 82.35%	< 0.5	>= 98.6%
INCOME	PER_CASH	NO_PREMIUMS
< 10^5.272	>=52.65%	>=7.5

Table 21: CART PATH DEFAULT # 8

Insight # 9: When customer age is less than 57 years and cash payment is between 40.45% and 82.35% having Risk_Score less than 98.6% then customer defaults.

NODE # 195		
PER_CASH	T3TO6_LATE	T6TO12_LATE
>= 40.45 %	< 0.5	< 0.5
PER_CASH	MORE12_LATE	RISK_SCORE
< 82.35%	< 0.5	< 98.6%
AGE		
< 57.03		

Table 22: CART PATH DEFAULT # 9

Insight # 10: When customer cash payment is between 40.45% and 82.35% and late payment of more than 12 months once then customer defaults.

NODE # 49		
PER_CASH	T3TO6_LATE	T6TO12_LATE
>= 40.45 %	>= 0.5	< 0.5
PER_CASH	MORE12_LATE	
< 82.35%	>= 0.5	

Table 23: CART PATH DEFAULT # 10

Insight # 11: When customer cash payment is more than 82.35% then he defaults.

NODE # 25		
PER_CASH	T3TO6_LATE	T6TO12_LATE
>= 40.45 %	< 0.5	< 0.5
PER_CASH		
>= 82.35%		

Table 24: CART PATH DEFAULT # 11

Insight # 12: When customer cash payment is more than 40.45% and payment is late once or more 3-6 months then he defaults.

NODE # 13		
PER_CASH	T3TO6_LATE	T6TO12_LATE
>= 40.45 %	>= 0.5	< 0.5

Table 25: CART PATH DEFAULT # 12

Insight # 13: When customer cash payment is more than 40.45% and payment is late once or more 6-12 months then he defaults.

NODE # 7		
PER_CASH	T6TO12_LATE	
>= 40.45 %	>= 0.5	

Table 26: CART PATH DEFAULT # 13

We can conclude that there are three level of importance of variables:

S#	IMPORTANCE	VARIABLES
1	HIGH	PER_CASH, T6TO12_LATE, T3TO6_LATE
2	NORMAL	MORE12_LATE, RISK_SCORE
3	LOW	INCOME, AGE, NO_PREMIUMS

Table 27: CART - Variable Importance in Model

Model Performance

Let us examine the confusion matrix of the decision tree on out of sample data:

Acc. = 75.15 %	Actual 1	Actual 0
Prediction 1	1,172	5,626
Prediction 0	327	16,831

Table 28: Confusion Matrix CART

Sensitivity of the model is 78.18% whereas, specificity of the model is 74.9%. Area under the curve for the model is 76.6%.

DATA PRE-PROCESSING

After conducting exploratory analysis on the dataset provided finally, we will prepare our dataset for further model building stage. Our approach is shown below:

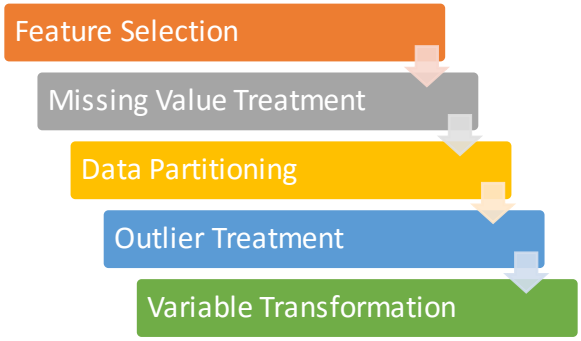


Figure 55: Data Preprocessing Process

Feature Selection

In the light of all the analysis we conducted on the dataset, there are following variables which need to be removed from the dataset as they don't have any significance on our response variables DEFAULT.

- MARITAL
- ACCOM
- AREA_TYPE
- DEPENDANTS
- VEH_OWNED
- ID

Now we have total 11 variables in the dataset with 79,853 rows. The names of the columns are mentioned below:

- PREDICTORS
  - PER_CASH
  - AGE
  - INCOME
  - 3TO6_LATE
  - 6TO12_LATE
  - MORE12_LATE
  - RISK_SCORE
  - NO_PREMIUMS
  - CHANNEL
  - PREMIUM
- RESPONSE
  - **DEFAULT**

Missing Value Treatment

There are no missing values in the dataset therefore there is no treatment required in this stage.

Data Partitioning

Now let's split our dataset through random sampling into two sets of 70-30 split. Our new two datasets are with following dimensions:

- **trainSample**
  - Columns = 11
  - Rows = 55,897
  - Defaulted = 3,499
  - No-Default=52,398
- **testSample**
  - Columns = 11
  - Rows = 23,956
  - Defaulted = 1,499
  - No-Default=22,457

Outlier Treatment

In late payment variables, most of the time customers do not delay payments that is why by nature these variables are skewed and their effective range becomes outlier because it is the area where bulk of the data is not present. Similarly, RISK_SCORE is having the same problem where most of the people will have normal score but few will go beyond normal ranges.

We will treat outliers of INCOME and AGE variables through Winsorizing them by capping the value with Q3+1.5*IQR and flooring with Q1-1.5*IQR.

INCOME Five-point summary after winsorize:

Min.	Q.1.	Med	Avg	Q.2.	Max.
24030	107620	166250	191520	252040	468670

Table 29: Winsorized INCOME Summary

AGE Five-point summary after winsorize:

Min.	Q.1.	Med.	Avg.	Q.3.	Max.
7670	14974	18625	18832	22636	34129

Table 30: Winsorised AGE Summary

Variable Transformation

We will transform three variables to raise comprehension of their values:

INCOME

Values ranging from thousands to tens of thousands and then into lakhs make it difficult to comprehend because of different scales. Therefore, we will do LOG10 transformation on this variable and after this below is the summary of the variable:

Min.	Q1	Med	Avg	Q3	Max.
4.381	5.032	5.221	5.206	5.40	5.671

Table 31: Summary after Transformation INCOME

AGE

We will transform age from number of days to number of years and for that we divide all values with 365 and resultant summary is mentioned below:

Min	Q1	Med	Avg	Q3	Max.
21.01	41.02	51.03	51.6	62.02	93.5

Table 32: Summary after Transformation AGE

RISK_SCORE

Since this variable is suppose to be percentage but its values range from 91 to 100 therefore it is better to transform them into decimal percentage values:

Min	Q1	Med	Avg	Q	Max.
0.919	0.9883	0.9918	0.9907	0.9952	0.9989

Table 33: Summary after Transformation RISK_SCORE

Variable Addition

There is one variable which needs to be created to boost analytical power of the dataset and it might generate some insights when combine with other variables. The suggested variable is mentioned below:

PREMIUM_RATE

We can create PREMIUM_RATE variable which can be calculated by dividing PREMIUM by NO_PREMIUMS. This variable has the potential to predict DEFAULT because when PREMIUM_RATE is high then customer might go for some other insurance company which are offering lower premium rates.

Min.	Q1	Med	Avg	Q3	Max.
30	475	872.7	1165	1486.4	26850

Table 34: Variable Addition - PREMIUM RATE

Univariate analysis of newly created variable:

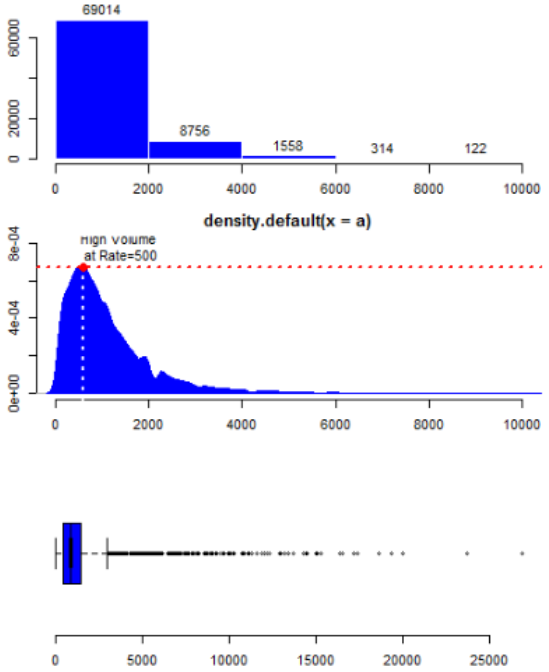


Figure 56: Univariate - PREMIUM_RATE

Most of the customers fall into the slot of less than 2000 premium rate. Highest peak is at 500 premium rate. There are several outliers on the upper tail of the distribution and IQR is narrow for the variable.

Let's conduct Bivariate Analyse first we analyse through bin analysis against DEFAULT and then we will do statistical analysis through boxplot and statistical summary of the pair variables.

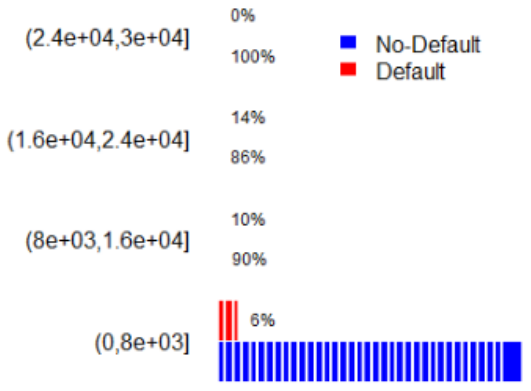


Figure 57: Bivariate Bin Analysis PREMIUM_RATE

We notice that default rate increases with the premium rate with the exception of highest slot of premium rate between 24,000 to 30,000.

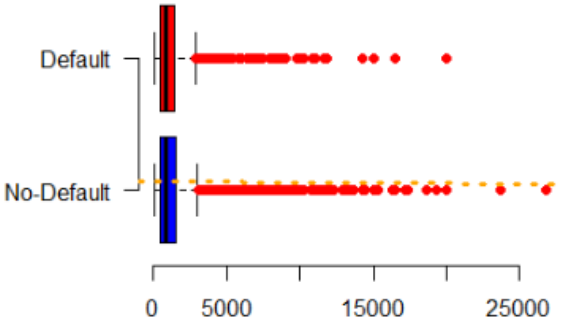


Figure 58: Bivariate Boxplot of PREMIUM_RATE

As you can see there are several outliers on the upper tail of the variable.

	No Default	Default
Min	30	38.70968
Q1	475	468.75
Med	872.7273	825
Q3	1500	1429.412
Max	3037.5	2850
Correlation	-0.0029	

Table 35: Bivariate Statical Summary of PREMIUM_RATE

In default range squeezes and median moves toward lower tail. Weak Correlation with default variable.

ANALYTICAL APPROACH

We will have three step modelling process to come up with final solution to the problem:

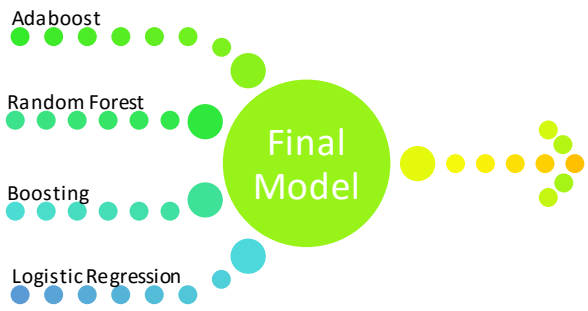


Figure 59: Analytical Approach Diagram

Reason for having three tier approach in building the final model is as follows:

Binomial Logistic Regression

We will develop binomial logistic regression model as an alternate model to challenge XGBoost Model with fulfilling all regression assumptions and analytical maturity.

There are standard practices which will be followed throughout the journey of Machine Learning Predictive Models such as K-Fold Cross Validation, Balancing Classes through Resampling and Model Performance Measure comparisons starting from Logistic Regression to Decision Tree to Random Forest and Finally to XGBoost.

Random Forest (Classification)

After conducting conclusive exploratory dive into the data with the help of Machine Learning now we will start the process of building the stage 1 model with the aim to reduce the variance due to the intrinsic nature of bagging by the random forest. We will evaluate outcomes of this model to take a final step of finish model.

Boosting (XGBoost)

We will develop the final production ready model through boosting and advantage of using this algorithm is not only reducing VARIANCE but also reducing the BIAS.

AdaBoost

Since XGBoost is prone to overfitting therefore we will develop model in AdaBoost to reduce the bias.

MODEL BUILDING

Let's start our model building exercise and first model we are going to develop is binomial logistic regression.

Binomial Logistic Regression

Although we developed same model earlier in our study as regression analysis to understand the relationships between dependent and independent variables.

Let's define assumptions of logistic regression before we start testing them:

Assumptions:

I. Appropriate Outcome Structure: Binomial Logistic regression requires dependent variable to be BINARY. As our dataset has binary DEFAULT dependant variable therefore it complies with the assumption.

II. Observation Independence: Observations in a dataset should be independent of each other. This assumption is true for our dataset as well.

III. Absence of Multicollinearity: Logistic regression requires less or no multicollinearity among independent variables. As we checked earlier in regression analysis there is minimal collinearity. Therefore, this assumption is valid as well.

IV. Linearity of Independent Variables & Log Odds: Although logistic regression not requires linear relation between dependent and independent variables. However, it requires independent variables to be linearly dependent on LOG ODDS.

V. Large Sample Size: Logistic Regression require large sample size. Our dataset is more than hundred thousand records which makes it more than enough to satisfy this assumption.

Resampling:

We will use ROSE package to oversample minority class in our dataset so that both classes represent equal proportion in the dataset. After applying ovun.sample method of the said package total number of observations become 104,796. Having equal proportion of 50% each for default.

2-Fold Cross-validation:

Developing model with two-fold cross validation as dataset is large in size. Logistic Regression coefficients are mentioned below in table:

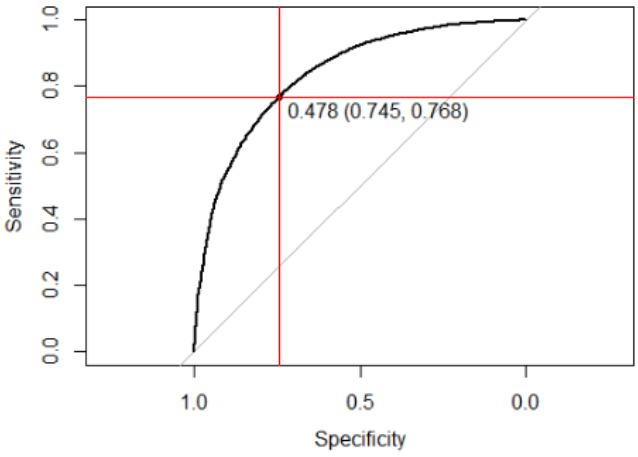
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	1.96E+01	1.03E+00	18.998	< 2e-16	***
PER_CASH	2.22E+00	8.11E-02	27.426	< 2e-16	***
AGE	-1.59E-02	5.98E-04	-26.589	< 2e-16	***
INCOME	-2.84E+00	2.00E-01	-14.152	< 2e-16	***
L3TO6	5.88E-01	1.05E-02	55.905	< 2e-16	***
L6TO12	1.01E+00	2.11E-02	47.742	< 2e-16	***
LMORE12	9.62E-01	2.48E-02	38.733	< 2e-16	***
RISK_SCORE	-1.64E+01	1.06E+00	-15.422	< 2e-16	***
NO_PREMIUMS	4.76E-02	2.03E-03	23.468	< 2e-16	***
CHANNEL	3.87E-02	9.52E-03	4.065	4.80E-05	***
PREMIUM	-8.58E-06	1.57E-06	-5.467	4.58E-08	***
PREMIUM_RATE	1.10E-04	1.20E-05	9.159	< 2e-16	***
CummLate	-1.15E-01	4.24E-03	-27.121	< 2e-16	***
Channel Behavior	-3.06E-01	8.09E-02	-3.783	0.000155	***

Table 36: Logistic Regression Coefficients

AIC	Accuracy	Kappa	Sensti.	Specif.
106,718	75.4%	50.08%	71.92%	78.88%

Acc. = 75.54 %	Actual 1	Actual 0
Prediction 1	37,686	11,066
Prediction 0	14,712	41,332

In sample accuracy of about 75.4% with true positive rates around 71.9%. Below is the ROC with 83.68% AUC.



Now let's evaluate the built model against out of sample dataset:

Acc. = 76.92 %	Actual 1	Actual 0
Prediction 1	1,126	5,155
Prediction 0	373	17,032

Table 37: Confusion Matrix - Logistic Regression

Accuracy of the model increases on out of sample data going to 76.92% with 75.1% sensitivity and 77.04% of Specificity. Please note oversampling was not done on test sample, But even then, sensitivity raise around 3%.

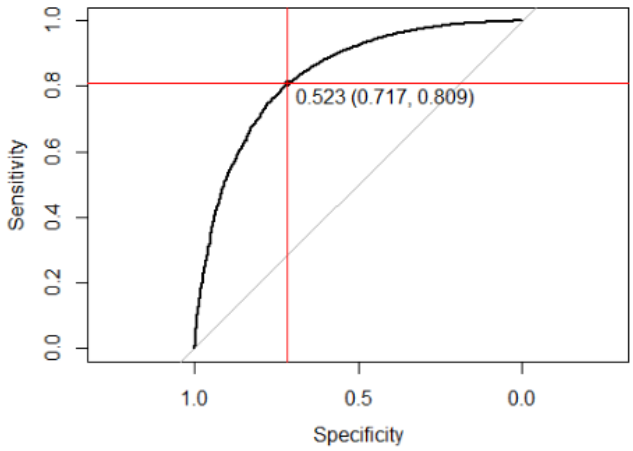


Figure 60: ROC Curve - Logistic Regression

Area under the curve of the model is 83.56%.

Random Forest

Now let's examine our dataset against ensemble method. Particularly bagging technique and the name of the model is Random Forest.

Hyperparameter

Following parameters are being selected after tuning the random forest model:

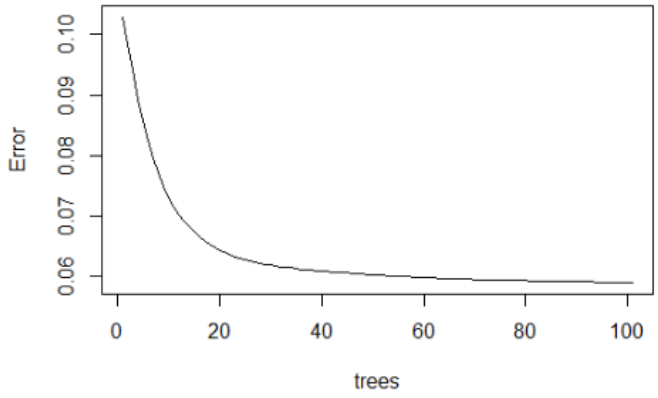


Figure 61: Mtry Selection Plot - Random Forest

1. We will set ntree =101 since our dataset is large therefore a greater number of trees should be minimum to achieve least OOB error.
2. We will set mtry=4, the difference between OOB error between 6 and 9 is minimal but as we have learned in previous model that there are 8 important variables therefore keeping mtry=6 would lead to same trees.
3. We will set NodeSize=100 this would prevent overfitting.

Model Performance

Looking at the out of bag chart we can notice that although uptill 20 trees majority variations are captured, but minimum error in True positives happened after 61 trees.

	%IncMSE	IncNodePurity
PER_CASH	1.82E-01	5.17E+03
AGE	1.04E-01	1.91E+03
INCOME	1.06E-01	1.77E+03
L3TO6	1.14E-01	2.16E+03
L6TO12	8.76E-02	1.66E+03
LMORE12	5.68E-02	9.19E+02

RISK_SCORE	1.03E-01	1.70E+03
NO_PREMIUMS	8.23E-02	1.03E+03
CHANNEL	3.87E-02	5.70E+02
PREMIUM	5.23E-02	6.20E+02
PREMIUM_RATE	8.91E-02	1.32E+03

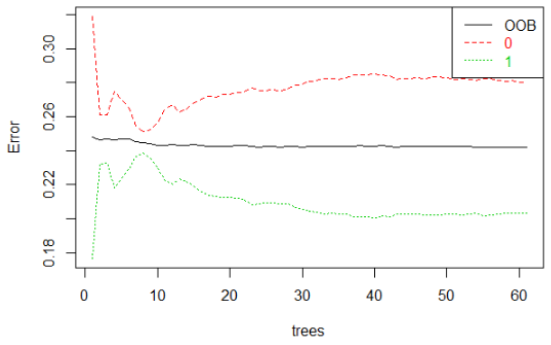
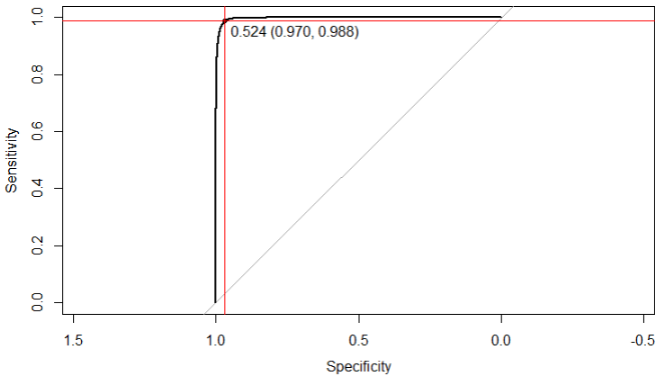


Figure 62: OOB Plot of Random Forest

Look at the confusion matrix you will notice that overall performance seems to be decreased but Sensitivity increased which is our desired state:

In-Sample	Actual 1	Actual 0
Prediction 1	51,936	1909
Prediction 0	462	50,489

Kappa = 95.48%

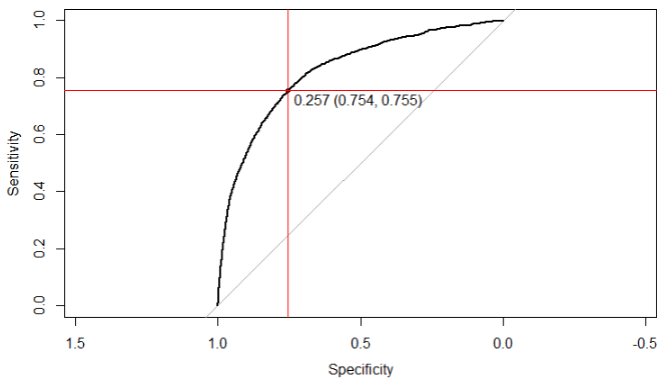


Confusion matrix of testing sample:

Out-Sample	Actual 1	Actual 0
Prediction 1	655	1,396
Prediction 0	844	21,061

Table 38: Confusion Matrix - Random Forest

Figure 63: ROC Curve - Random Forest



AUC reduced to 82.49 from 99.78%.

Kappa reduced to 31.98% from 95.48%. However, please note that above confusion matrix of out-sample is treated with 0.5 threshold which can be set to 0.257 to get 75.5% percent sensitivity.

Boosting – XGBoost

Please note that data is sensitive to overfitting and generally boosting is more prone to overfitting than bagging. Therefore, we will see how this model performs on the dataset provided.

Hyperparameter

After repeating XGBoost algorithm and comparing results of confusion matrix these are the optimum values of the hyperparameters:

- ETA=0.01 learning rate
- MAX_DEPTH=7 prevent overfitting by defining maximum depth of individual tree.
- MIN_CHILD_WEIGHT=27 stops splitting of the node if weight gets lower than benchmark mentioned here.
- NROUNDS=67 which is in classification is number of trees.

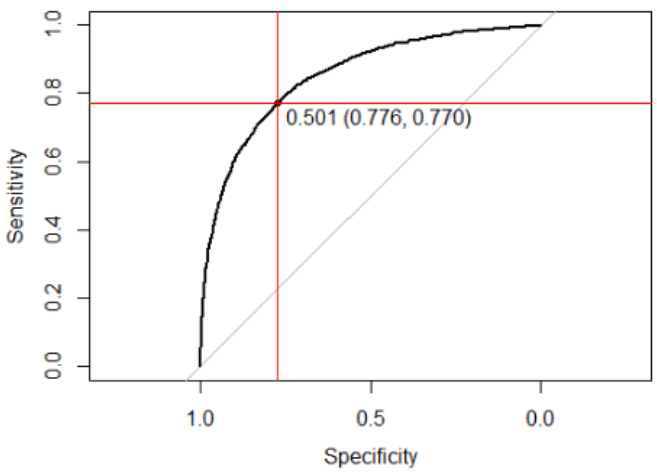
Model Performance

Let's evaluate confusion matrix of boosting model:

In-sample

Acc. = 77.5 %	Actual 1	Actual 0
Prediction 1	41,278	12,409
Prediction 0	11,120	39,989

Accuracy of 77.5% with sensitivity of 78.78% and kappa is 55.1%. The ROC is with AUC is 85.35%.



Out-sample

Acc. = 77.64 %	Actual 1	Actual 0
Prediction 1	1,157	5,245
Prediction 0	342	17,212

Table 39: Confusion Matrix - Boosting

Notice that overall performance is better than all other models but our major concern is true positive where this model lags behind. Sensitivity of the model is 77.1% and specificity is 76.6%. AUC is 83.89%. Kappa is 21.31%

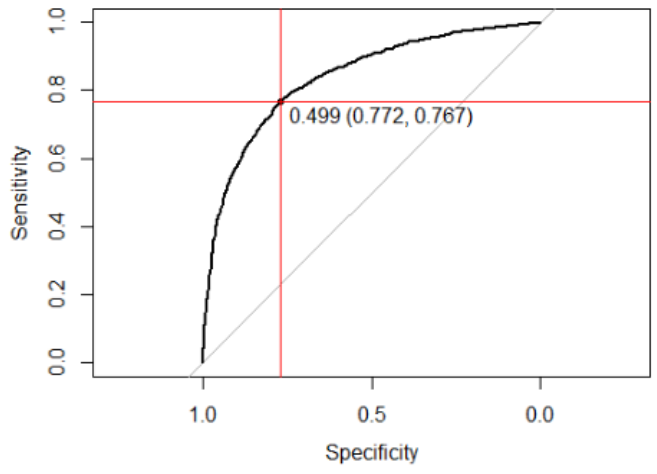


Figure 64: ROC Curve - Boosting

Please note area under the curve is slightly more than bagging but TRUE POSITIVE RATE is lesser than bagging which is our area of concern.

Adaptive Boosting – AdaBoost

Please note that data is sensitive to overfitting and generally boosting is more prone to overfitting than bagging. Therefore, we will see how this model performs on the dataset provided.

Hyperparameter

After repeating XGBoost algorithm and comparing results of confusion matrix these are the optimum values of the hyperparameters:

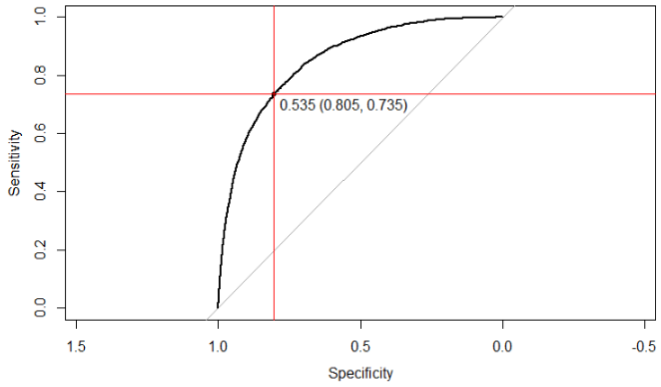
- TREE_DEPTH=3 prevent overfitting by defining maximum depth of individual tree.
- NROUNDS=200 which is in classification is number of trees.

Model Performance

Let's evaluate confusion matrix of boosting model in-sample:

Acc. = 76.5 %	Actual 1	Actual 0
Prediction 1	39,958	11,702
Prediction 0	12,440	40,696

Kappa is 53.93% whereas sensitivity is 76.59%.



AUC is 84.84%.

Out-Sample

Acc. = 76.36 %	Actual 1	Actual 0
Prediction 1	1,117	4,962
Prediction 0	382	17,495

Table 40: Confusion Matrix - Boosting

Notice that overall performance is better than all other models but our major concern is true positive where this

model lags behind. Sensitivity of the model is 74.5% and specificity is 77.9%. Kappa is 21.61%. AUC is 83.58%.

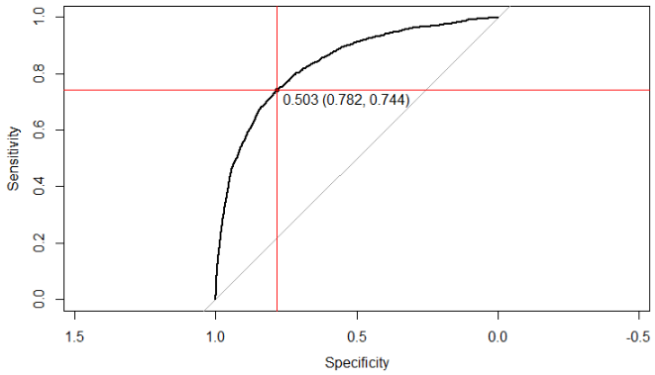


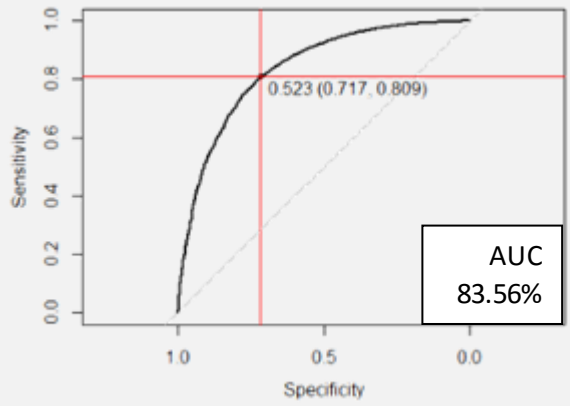
Figure 65: ROC Curve - Boosting

Best Model

Logistic Regression

Best ★

Kappa
23.5%

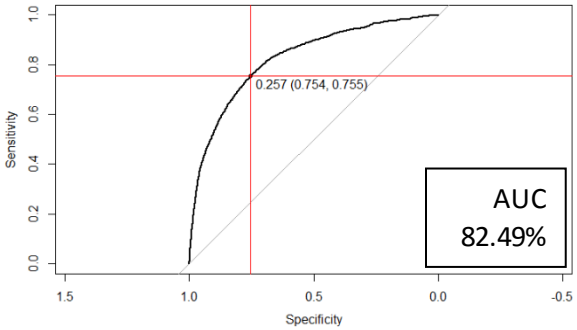


Confusion Matrix
Threshold = 0.5

Out of Sample	**Actual 1**	**Actual 0**
Prediction 1	1,126	5,155
Prediction 0	373	17,032

Random Forest

Kappa
31.98%

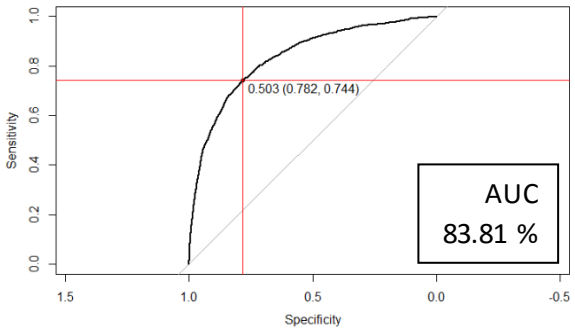


Confusion Matrix
Threshold = 0.5

Out of Sample	Actual 1	Actual 0
Prediction 1	655	1,396
Prediction 0	844	21,061

Adaptive Boosting

Kappa
21.61%

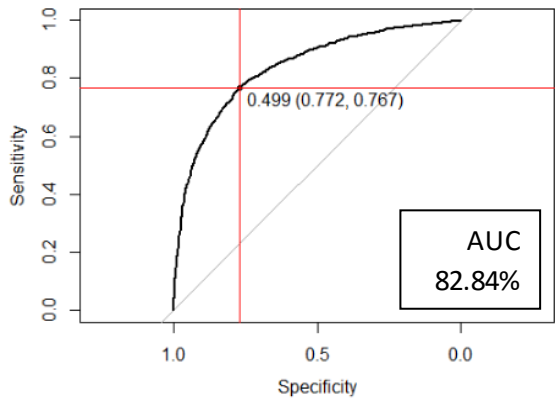


Confusion Matrix
Threshold = 0.5

Out of Sample	Actual 1	Actual 0
Prediction 1	1,117	4,962
Prediction 0	382	17,495

Extreme Gradient Boosting

Kappa
21.31%



Confusion Matrix
Threshold = 0.5

Out of Sample	Actual 1	Actual 0
Prediction 1	1,157	5,245
Prediction 0	342	17,212

Business Insights

After making all these models we learned that few trends which are very important for policy making of insurance.

Importance of Dimensions

As we classified customer profile into dimensions at the start of the document let us evaluate what is the importance of those dimensions against default:

1. PAYMENT HISTORY [HIGH]
2. LATE PAYMENTS [HIGH]
3. FINANCIAL WORTH [LOW]
4. DEMOGRAPHICS [LOW]

This clearly shows that it is very important to keep track of the customer payments in-order to predict whether the customer is going to default or not. There is minimal impact of demographics and financial worth of the customers.

Cash Payment & Late Payment Interaction

Another aspect is the interaction between CASH PAYMENT and LATE PAYMENT. It seems that high cash payment of more than 82% lead to default and forms a threshold of default, on the other hand late payments effect this threshold up and down. Therefore, even if cash payment is lower around 40-52% then 1 late payment decreases the threshold of default from 82% to 40% and customer lead to default.

Weak Predictors

Income, Age, Premium Rate and Risk Score are the weak predictors in the dataset.

CONCLUSION

Default is the predominantly circumstantial based event. Meaning that late payments and mode of payment effects the state of default which needs to be tracked on real-time basis. There is very weak correlation of other predictors such as Income, Age & Premium Rates. This is very logical because at the time of customer acquisition all the demographics related aspects of the customers are already screened therefore in default it does not play important role. However, against expectations premium rate is also weak predictor whereas usually churn rate significantly depend on service cost due to competition. The only threat remains change in circumstances of the customer which lead to default and thus it can be determined on real-time basis. It is critical to detect such

```
abline(h=0.767,col="red")

rocCurve$auc
```

ANNEXURE B - LIST OF FIGURES

ANNEXURE C - LIST OF TABLES