

A Study on The Probability That A Customer Will Renew Their Insurance

Abstract

In this paper, we will analyze the probability that a customer defaults on their insurance premium payment. Insurance is a mean of protection from any sudden unforeseen financial loss and risk profiling is an important aspect of it. Keeping customers satisfied and being able to custom target various demographics is crucial to maintain and increase customers. In the world of finance, the probability of default is the essential credit risk. It is used to give an estimate of the likelihood that a borrower will be unable to meet its debt obligations. The dataset consists of 17 parameters for 79853 customer observations with a combination of Indicator and continuous variables. Data mainly covers customers' demographic information, premium payment related behavior and risk profile information. Random tree and Logistic regression have been used for modelling. Their performance measures have then been compared to decide which is better suited for the dataset and hence used to draw meaningful conclusions which can be used by an insurance business to formulate new strategies. We are building a model to predict the probability that a customer will default the premium payment and hence in our analysis '*renewal*' would be the target or the response variable i.e. the dependent variable and other variables would be independent or the predictor variables. Through this study we will try to identify the key factors that influence the timely payment of premium by customers.

Introduction

Insurance is a form of risk management tool which allows the insured party to hedge the risk of an uncertain loss. Herein the entity offering the protection against the risk is called 'Insurer' and the customer taking the protection is called 'Insured'. For purchasing the protection ('Insurance') against the uncertain loss from the Insurer, the Insured has to pay a premium termed as 'Insurance premium' or simply 'premium' which is usually periodic in nature.

Insured has the right to claim compensation under the Insurance policy till the time the premiums are duly paid and the policy is therefore renewed on the likely date of renewal. However, at the discretion of Insured, if the premium is not paid ever after the due date or in other words the Insured default the premium payment, the Insurance policy gets lapsed and any further claim cannot be raised under the Insurance policy.

At present, Life Insurance, Health Insurance and General Insurance (Non-Life) are the commonly used risk management tools in Indian market. Insurers are governed by autonomous regulatory bodies (Insurance Regulatory and Development Authority in India) to protect the interest of the Insured and to prevent mis-selling and other unfair practices.

Insurance is an important risk management tool which is relevant to all sections of society to mitigate uncertain losses triggered due to various unforeseen events. Developed countries has higher insurance penetration and developing countries are catching up with increasing awareness and affordable cost of buying insurance.

From a commercial point of view, premium paid by the customer is the major revenue source for Insurer. Default in premium payments results in significant revenue losses and hence Insurer would like to know upfront which type of customers would default premium payments.

There are numerous type of factors that have an effect on insurance renewal like psychographic, socio-economic factors such as financial inclusion(taking a loan from a bank or getting a bank account), demographic factors such as gender or education levels of the household head etc.

Effect on insurance due to COVID-19

COVID has had both a positive and negative impact on the insurance market. Due to the outbreak of coronavirus, an increasing number of people have become more aware of insurance. Many of them consider insurance as a necessity to be prepared for any unforeseen circumstance in the future. Prior to the spread of COVID-19 in India, only about 10% of people showed an interest in buying insurance to cover medical emergencies including pandemics and infectious diseases. Now, however 71% of people consider it as a necessary safety tool.

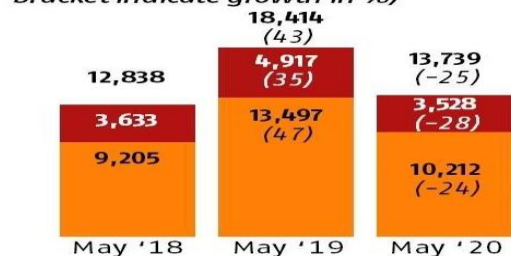
But, due to the pandemic most people have lost their jobs and many middleclass people were unable to pay their premium. As a result, they couldn't renew the premium and they had to discontinue their policies which had a large negative impact on companies as well.

"The disruption caused by the coronavirus' spread and the pandemic-induced lockdown resulted in Life insurance industry losing around four million policies and premiums of around Rs 45,000 crore", said Raj Kumar, managing director of LIC.

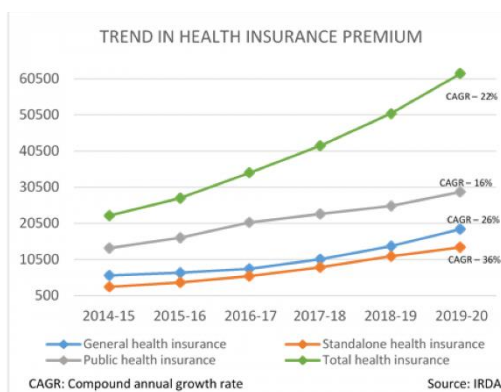
TAKING A COVID-HIT

New business premium of life insurance companies in May

■ LIC ■ Private life insurers (figures in bracket indicate growth in %)

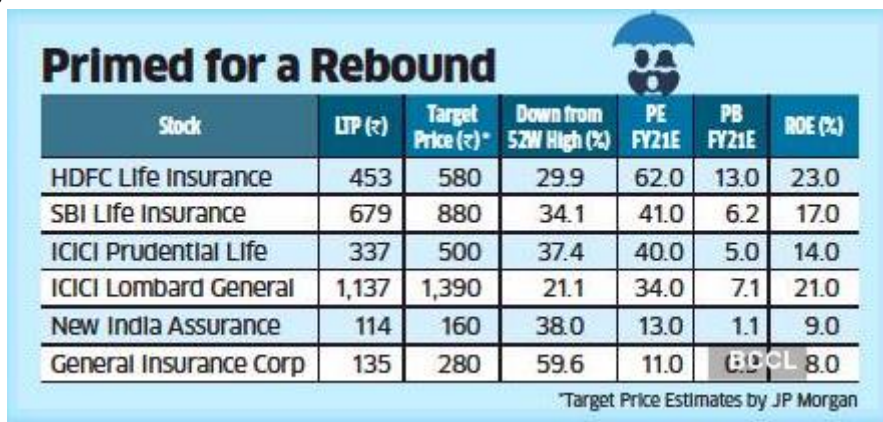


Source: Irdai



Life insurers faced high mortality claims due to COVID as a loss to insurance companies and also of low interest rates. But after the Lockdown phase, it seems to have got back on track from July as premium collection has turned positive, boosted by increased interest in insurance.

“As of August 31, LIC’s premium income is back to previous year’s level. So, we have covered the gap that was there in April, May when the premium went down by about 32 per cent,” said Kumar, managing director of LIC.



Stock	LTP (₹)	Target Price (₹) *	Down from 52W High (%)	PE FY21E	PB FY21E	ROE (%)
HDFC Life Insurance	453	580	29.9	62.0	13.0	23.0
SBI Life Insurance	679	880	34.1	41.0	6.2	17.0
ICICI Prudential Life	337	500	37.4	40.0	5.0	14.0
ICICI Lombard General	1,137	1,390	21.1	34.0	7.1	21.0
New India Assurance	114	160	38.0	13.0	1.1	9.0
General Insurance Corp	135	280	59.6	11.0	BDCL	8.0

*Target Price Estimates by JP Morgan

Literature Review

- Fuzzy-probabilistic multi agent system for breast cancer risk assessment and insurance premium assignment-(Farzaneh Tatari, Mohammad-R. Akbarzadeh-T, Ahmad Sabahi) (2012)**
 In their study, they have done risk assessment on the development of breast cancer. According to studies one in nine women will develop breast cancer at some point in her life. As it has been found that not all factors equally increase the chance of breast cancer development, the risk factors have been categorized into three groups of strong, moderate and minor risk factors. Factors such as age, family history, alcohol consumption etc. have been considered. Due to high imprecision and linguistic form of information it is difficult to analyze data and therefore they have used fuzzy-logic based analysis to handle the existing uncertainty. Based on this analysis, insurance companies can offer different premiums for different customer segments. They had concluded that carriers of certain kind of genes was a critical factor for deciding the risk of development of breast cancer.
- Risk Prediction in Life Insurance- (Noorhannah Boodhun & Manoj Jayabalan) (2018)**
 Their study was done to find and examine the factors that have an effect on insurance purchasing decision of customers. They developed a model from data of ~60,000 applicants having 21 features which describe the characteristics and nature of applicants like BMI, age, weight, family history, etc. They used algorithms like linear regression, random tree, REP Tree and Artificial Neural Network in their study to predict risk involved in life insurance policies. This model provided a very efficient method for life insurance business to classify the applicants for life insurance which is otherwise a very slow process that leads to people switching to different policies or not buying at all. Also, with the increase of data availability, automation of the process is necessary which is provided by the model for business advancement. It was found that REP Tree algorithm outperforms others with lowest MAE and RMSE of 1.5285 and 2.027 with CFS method, whereas linear regression with PCA method showed very satisfactory values of MAE and RMSE of 1.6396 and 2.0659 respectively.
- A Study on Factors Affecting Customers Investment Towards Life Insurance Policies- (Babita Yadav, Anshuja Tiwari) (2012)**
 They divided the applicants based on features like age, income, occupation etc. to analyze their behavior and its consequences on Insurance Sector. The main motive behind the study was that

almost 70% of people's lives are still uninsured and to give boost to the business development in this domain. They used various popular statistical algorithms like chi-square and correlation to analyze and to identify the most important factors for their hypothesis. Results of the study showed that people between the age of 30-40 are more likely to buy insurance. Also, company reputation, money back guarantee, low premium and kind of risk coverage attracted customers.

- **Life Insurance Industry of India – Past, Present & Future (A Study of LIC of India-Shilpa Agarwal and A.K Mishra)**

This study was based on examining the status of LIC in pre and post liberalised era (LPG- Liberalisation privatisation and globalization in year 1991) as well as estimating the future trend in LIC business. LIC was formed in 1956 and became a mammoth in insurance industry. The data used was from the LIC website. They used Method Of least squares for examining the future trend of their business. Based on the year 2009, they calculated the trend value for year 2020 and it showed the business of LIC is in increasing trend. Till 2013 there were 52 insurance companies operating in india of which 24 are in life insurance business and having a total share of 80.2%.

Some questions answered by previous researchers are:

- **Does age of the customer affect the chance of continual of premium payment or renewal of insurance?**
 - Previous works have shown and suggested that increase in age leads to an increase in the renewal of insurance.
- **Does gender of the customer have an effect on insurance renewal or purchase?**
 - Studies have suggested that females are less likely to buy or renew insurance than males.
- **Does type of area in which one lives affect the probability of insurance purchase?**
 - Due to higher awareness and understanding of financial tools and easier access people living in urban areas have a much higher chance of insurance renewal as compared to those who live in rural areas.

Objectives

The aim of this study is to understand the premium payment pattern of customers of an Insurance company. For the study, we have customer data available primarily covering:

- a) Customer demographic information e.g. Age, Income, Marital Status, residence area type etc.
- b) Insurance policy and premium payment related information e.g. premium, renewal, sourcing channel etc.
- c) Customer risk profile (risk score)

The objective is to predict the probability that a customer will default on premium payment, so that insurance agents can proactively reach out to the policy holders to follow up for the payment of premium. Simultaneously, it will also help understand customer demographics which are more likely to default and to price the premium amount in accordance to the same.

Data Dictionary

The dataset has 79853 records with 17 different variables. The target or the dependent variable in the given dataset is "renewal", which has values as 0 or 1. "0" indicates that customer has not renewed the premium and "1" indicates that customer has renewed the premium. The data is based on life insurance and has been collected between 2017 to 2019.

Below is the list of variables along with the description and categorization:

Variables	Description	Type
Id	Unique customer ID	Continuous
perc_premium_paid_by_cash_credit	% of the premium paid by cash payments	Continuous
age_in_days	Age of the customer in days	Continuous
Income	Income of the customer	Continuous
Count_3-6_months_late	Number of times premium was paid 3-6 months late	Continuous
Count_6-12_months_late	Number of times premium was paid 6-12 months late	Continuous
Count_more_than_12_months_late	Number of times premium was paid more than 12 months late	Continuous
Marital Status	0 indicates that customer is Unmarried and 1 indicates that customer is Married	Indicator
Veh_owned	Number of vehicles owned (1-3)	Indicator
No_of_dep	Number of dependents in the family on the customer(1-4)	Indicator
Accommodation:	0 indicates that customer has rented the accommodation and 1 indicates that customer has owned the accommodation	Indicator
Risk_score	Risk score of customer	Continuous
no_of_premiums_paid	Number of premiums paid till date	Continuous
sourcing_channel	Channel through which customer was sourced (A/B/C/D/E)	Indicator
residence_area_type	Residence type of the customer (Rural/Urban)	Indicator
premium	Premium amount	Continuous
renewal	0 indicates that customer has not renewed the premium and 1 indicates that customer has renewed the premium	Indicator

Data Source

The data used in this study has been collected from secondary sources. The data has been provided by The University of Texas at Austin (UT Austin) and Great lakes Institute of Management, Chennai.

Dataset

<https://drive.google.com/drive/folders/1sEeUzD8aFI3RRQvJvH5ntfp4C6ua8xVa?usp=sharing>

Methodology and Modelling

We will start with data pre-processing followed with exploratory data analysis, data normalization and outlier treatment. Data variables are normalized to avoid any one variable overshadowing the model and to make sure data remains uniform. This will help in getting a brief idea on the data we are working with and we will also be able to check for any missing values or exceptions. Also, we will be able to

check the presence of dependency and correlation among variables. After this we will use two different models logistic regression and random forest. We will then compare the model performance measures to find out which is better suited for this study to arrive at the right conclusions.

Hypothesis

From the knowledge of previous studies and market trends we observe following correlation between different variables and probability of insurance renewal.

- Age will have a positive effect and will be a significant factor - It is a common trend and has been observed by various studies that as a customer gets older, they are much more likely to pay their insurance premiums due to various reasons such as higher health risks, more financial stability or increased number of dependents in the family.
- Income will have a significant and positive impact - With higher incomes people are more likely to be able to pay their premiums. They are financially secure and are also looking for opportunities to use their money to protect their family and their own interests.
- Count of late payments will have a negative impact on the probability of renewal - A customer can be late on their premium payments for various reasons such as perceived or actual reduction in the need for insurance, sudden unavoidable expenditures or unemployment etc. Increase in the number of late payments is indicative of the fact that due to some reason the customer is unable to pay their premiums and is thus more likely to not renew their insurance.
- Marital status will have a positive impact but not a significant one - Along with the increased chance of higher number of dependents in the family there is a sense of commitment and want of protecting their loved ones. Other studies also have suggested that a married person is more likely to buy or renew insurance than an unmarried person.
- Vehicle ownership will not be a significant factor - The number of vehicles owned can be a very deceptive for companies trying to figure out potential customers who might renew. It is very possible that a customer with a stable and high income has only one vehicle due to personal preferences or availability of excellent public transport system in their locality. Such individuals are still as likely to renew their insurance as another customer with ownership of multiple vehicles. There is no general trend observed regarding effect of vehicles ownership and hence it is difficult and unreliable to use it as predictive measure.
- Number of dependents will have a positive effect but not a significant one as again it is a difficult to draw conclusions due to significantly varied possibilities in similar situations e.g. age and education of such dependents, income sufficiency to address other primary expenses like education, etc. will influence insurance buying behavior.
- Risk score generally has a positive impact on the purchase of insurance i.e. people at higher risks are more likely to buy insurance however the given data does not clarify the method of its calculation and its impact on credit worthiness of a customer
- Number of premiums paid and premium amount will have a positive and significant effect as they both generally indicate how committed a customer is in renewing their insurance.
- Percentage of premium paid in cash will have a negative effect - Paying in cash is generally indicative of the customer not having a long-term commitment towards insurance or not having sufficient savings and thus might not be able to have funds available in time to pay the premium.

- Residence area type will have a negative impact - As insurance awareness and financial stability is lesser in rural areas compared to urban areas it is generally observed that customers living in urban areas are more likely to purchase and renew insurance.
- Accommodation ownership will not have a significant impact – We did not observe a significant correlation between a customer living in rented or owned accommodation and insurance renewal decision as reasons for tenancy could be temporary posting, plans to relocate, etc and not necessarily a reflection of customer's financial stability.

Based on previous studies and observations , we are suggesting below correlation:

Renewal=f(Age(+ve), Income(+ve), Count of late payments(-ve), Marital Status(+ve), Number of dependents(+ve), Risk score(+ve), Percentage of premium paid in cash(-ve), Number of premiums paid (+ve), Premium(+ve), Residence area type(-ve)).

Data Overview

Variable	Initial Values
id	num 1 2 3 4 5 6 7 8 9 10 ...
perc_premium_paid_by_cash_credit	num 0.317 0 0.015 0 0.888 0.512 0 0.994 0.019 0.018 ...
age_in_days	num 11330 30309 16069 23733 19360 ...
Income	num 90050 156080 145020 187560 103050 ...
Count_3-6_months_late	num 0 0 1 0 7 0 0 0 0 0 ...
Count_6-12_months_late	num 0 0 0 0 3 0 0 0 0 0 ...
Count_more_than_12_months_late	num 0 0 0 0 4 0 0 0 0 0 ...
MaritalStatus	num 0 1 0 1 0 0 0 0 1 1 ...
Veh_Owned	num 3 3 1 1 2 1 3 3 2 3 ...
No_of_dep	num 3 1 1 1 1 4 4 2 4 3 ...
Accommodation	num 1 1 1 0 0 0 1 0 1 1 ...
risk_score	num 98.8 99.1 99.2 99.4 98.8 ...
no_of_premiums_paid	num 8 3 14 13 15 4 8 4 8 8 ...
sourcing_channel	chr "A" "A" "C" "A" ...
residence_area_type	chr "Rural" "Urban" "Urban" "Urban" ...
premium	num 5400 11700 18000 13800 7500 3300 20100 3300 5400 9600 ...
renewal	num 1 1 1 1 0 1 1 1 1 1 ...

By visual inspection we can see that income has significantly large values as compared to various variables and needs to be normalized. While sourcing channel and residence area type are character datatypes and need to be converted to integers for further analysis.

Data has a mix of Indicator and Continuous variables which mainly covers Customer's demographic information, premium payment related behavior and risk profiling.

Data limitations: Based on above and visual inspection of data, below are some of the limitations to the information that can be inferred:

- 'Veh_Owned' doesn't clarify the type of vehicles owned (2-wheeler or a 4-wheeler or both)
- 'No_of_dep' doesn't clarify the age group of dependents (kids, adults, elderly)
- 'risk_score' doesn't clearly clarify its relation to the credit worthiness of customer (is it directly proportional or inversely proportional). Also, there is no information provided on the calculation methodology of it.
- Sourcing Channel though of different values doesn't clearly mention what they represent i.e(banks, agents, third parties etc.

Data Preparation

- Data has no missing value
- Converting character type data to numeric:
 - Residence area type values to 1 and 0 (Rural =1, Urban =0)
 - Source Channel values to 1,2,3,4,5 (A, B, C, D, E)
- For better readability, we have added new columns:
 - 'cashPercent' to to display Cash premium payment in % terms.
 - 'age' to display customer's age in years for improved readability
 - 'countLatePayment' as a substitute for 'Count_3-6_months_late', 'Count_6-12_months_late', 'Count_more_than_12_months_late'

Exploratory Data Analysis

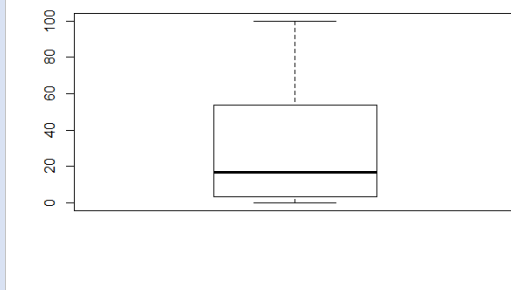
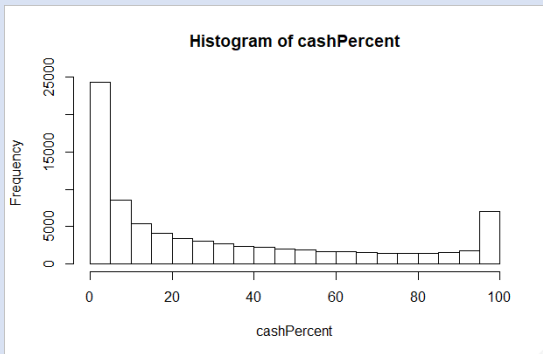
From the table below we get a brief idea about the distribution of variables.

Accomodation	risk_score	no_of_premiums_paid
Min. :0.0000	Min. :91.90	Min. : 2.00
1st Qu.:0.0000	1st Qu.:98.83	1st Qu.: 7.00
Median :1.0000	Median :99.18	Median :10.00
Mean :0.5013	Mean :99.07	Mean :10.86
3rd Qu.:1.0000	3rd Qu.:99.52	3rd Qu.:14.00
Max. :1.0000	Max. :99.89	Max. :60.00

Age	residence_area_type		
Min. : 21.00	Min. :0.0000		
1st Qu.: 41.00	1st Qu. :0000		
Median : 51.00	Median :0.0000		
Mean : 51.61	Mean :0.3966		
3rd Qu.: 62.00	3 rd Qu. :1.0000		
Max. : 103.00	Max:1.0000		
Income	Count_3-6_months_late	Count_6-12_months_late	
Min. : 24030	Min. : 0.0000	Min. : 0.00000	
1st Qu.: 108010	1st Qu.: 0.0000	1st Qu.: 0.00000	
Median : 166560	Median : 0.0000	Median : 0.00000	
Mean : 208847	Mean : 0.2484	Mean : 0.07809	
3rd Qu.: 252090	3rd Qu.: 0.0000	3rd Qu.: 0.00000	
Max. :90262600	Max. :13.0000	Max. :17.00000	
Count_more_than_12_months_late	Marital Status	Veh_Owned	
Min. : 0.00000	Min. :0.0000	Min. :1.000	
1st Qu.: 0.00000	1st Qu.:0.0000	1st Qu.:1.000	
Median : 0.00000	Median :0.0000	Median :2.000	
Mean : 0.05994	Mean :0.4987	Mean :1.998	
3rd Qu.: 0.00000	3rd Qu.:1.0000	3rd Qu.:3.000	
Max. :11.00000	Max. :1.0000	Max. :3.000	
No_of_dep	premium	renewal	cashPercent
Min. : 1.00	Min. :1200	Min. : 0.0000	Min. : 0.00
1st Qu.: 2.00	1st Qu.:5400	1st Qu.: 1.0000	1st Qu.: 3.40
Median : 3.00	Median :7500	Median : 1.0000	Median: 16.70
Mean : 2.50	Mean :10925	Mean : 0.9374	Mean : 31.43
3rd Qu.: 3.00	3rd Qu.:13800	3rd Qu.:1.0000	3rd Qu.: 53.80
Max. : 4.00	Max. :60000	Max. :1.0000	Max. : 100.00

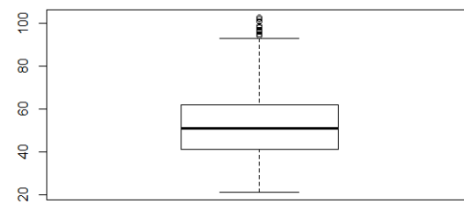
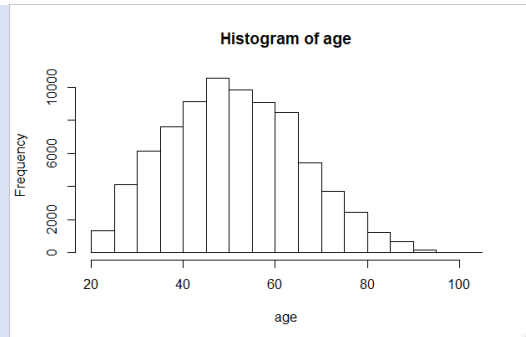
Univariate Data Analysis

Cash Percent



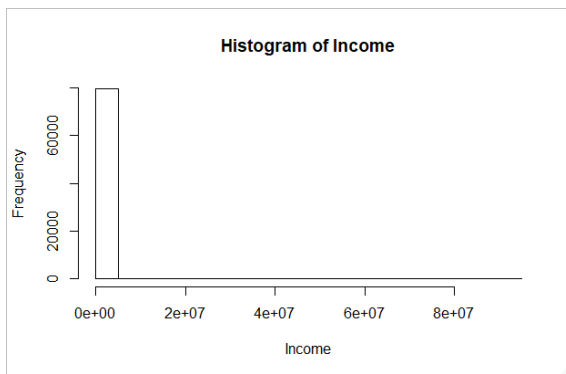
- Values range from 0 to 100 with majority of data points falling in the lower range of 0% to 5%
- Mean = 31.43%
- Data has outliers

Age

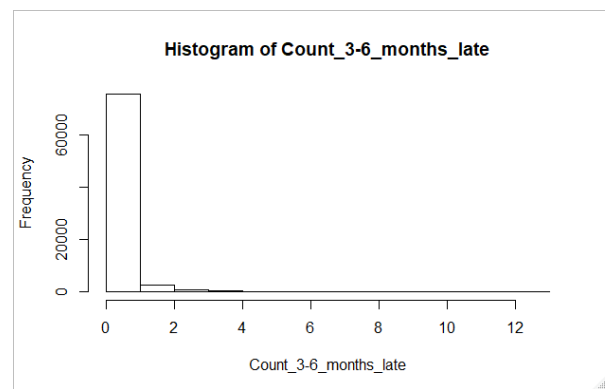


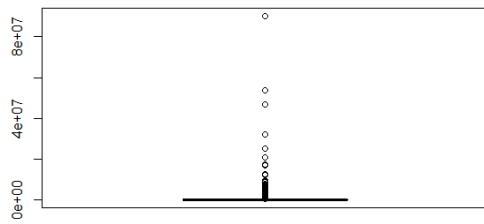
- Values range from 21 to 103 years with data appearing to be somewhat normally distributed
- Mean = ~51 years = Median
- Data has outliers

Income

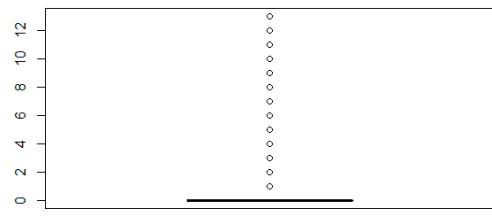


Count_3-6_months_late



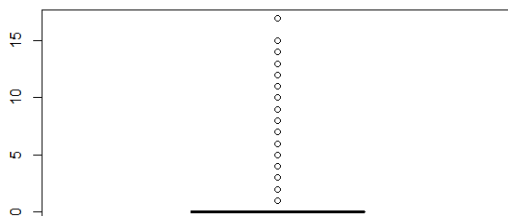
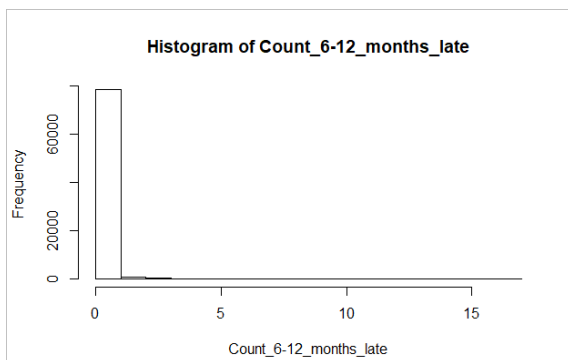


- Data has a wide range of 24,030 to 90,262,600 (right skew)
- Mean = 208847
- Data has too many outliers



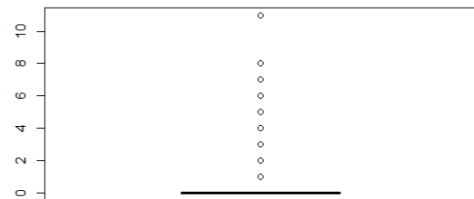
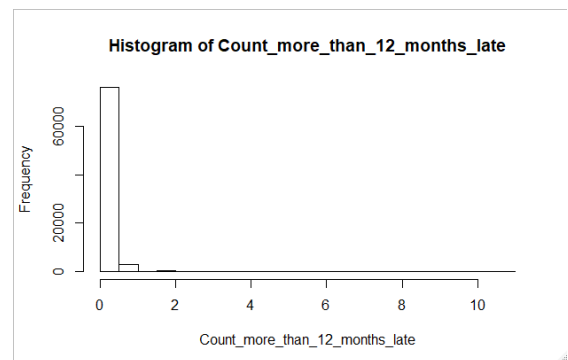
- Data varies from 0 to 13 with majority delay counts are 0 to 1 (right skew)
- Mean = 0.2484
- Data has too many outliers

Count_6-12_months_late



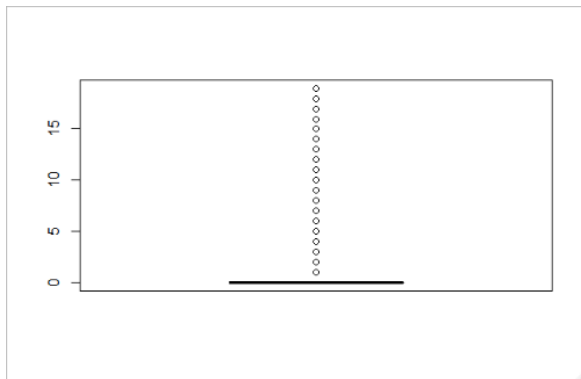
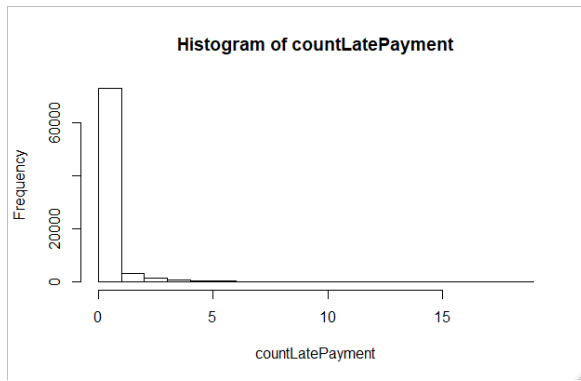
- Data varies from 0 to 17 with majority delay counts are skewed towards 0 to 2 (right skew)
- Mean = 0.07809
- Data has too many outliers

Count_more_than_12_months_late



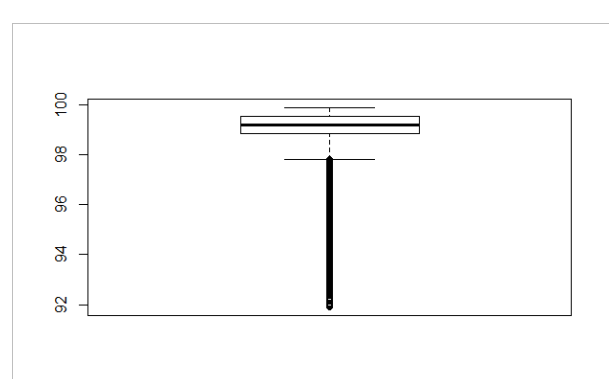
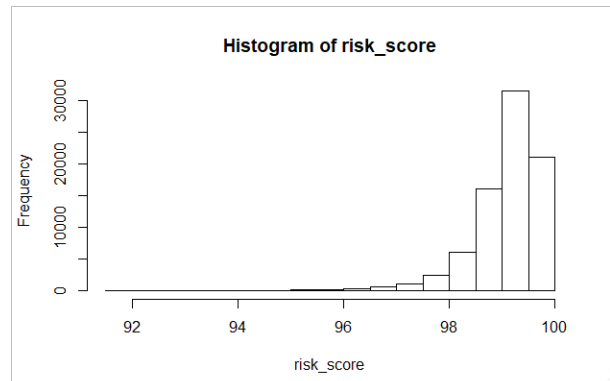
- Data varies from 0 to 11 with majority delay counts are skewed towards 0 to 1 (right skew)
- Mean = 0.05994
- Data has too many outliers

Count Late Payment



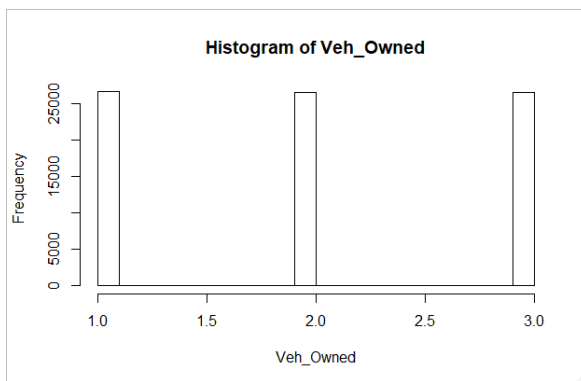
- Data varies from 0 to 19 with majority delay counts are skewed towards 0 to 1 (right skew)
- Mean = 0.3864
- Data has too many outliers

Risk Score

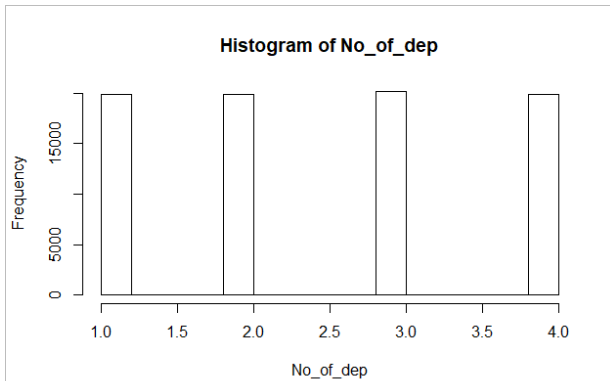


- Data varies from 91.90 to 99.89 with majority data skewed towards 99.0 to 99.5 (left skew)
- Mean = 99.07
- Data has too many outliers

Vehicles owned

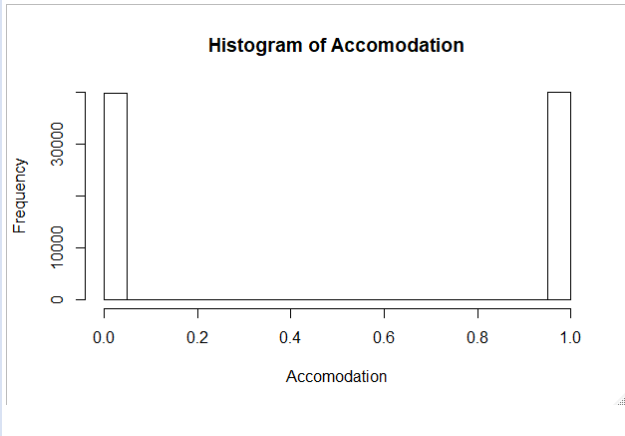


Number of dependents



- Data has 3 categories with almost equal no of cases for 1/2/3 vehicles owners

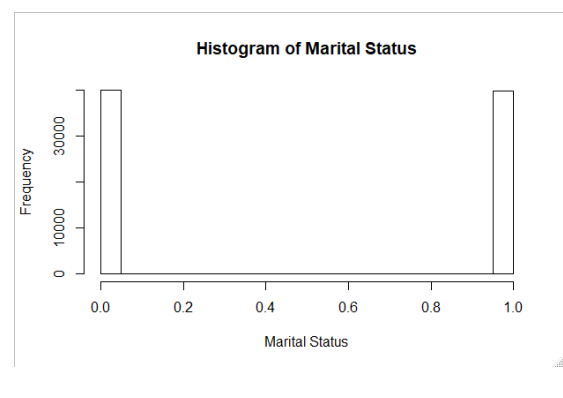
Accommodation



- Data has 2 categories with almost equal no of Owned and Rented cases

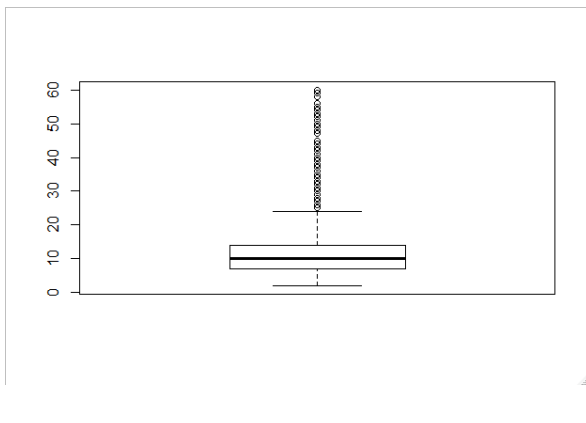
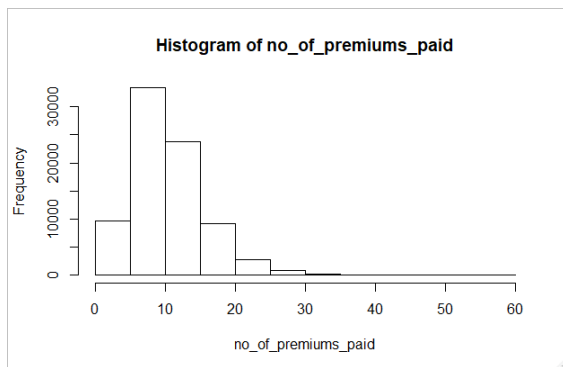
- Data has 4 categories with almost equal no of 1,2,3,4 dependent cases

Marital Status



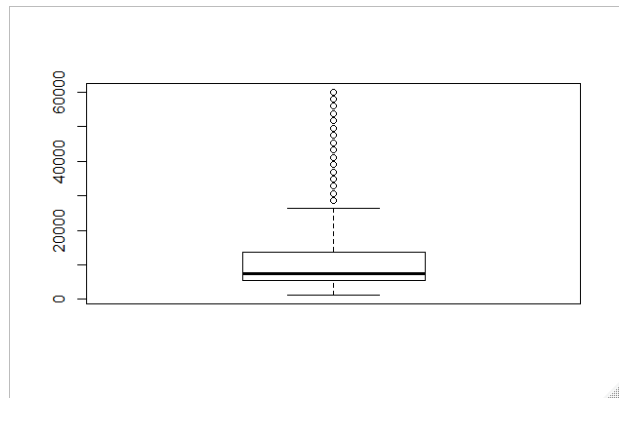
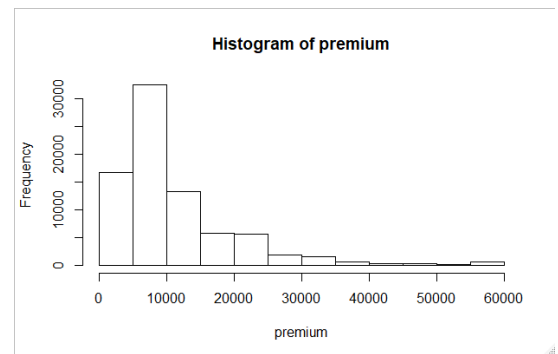
- Data has 2 categories with Unmarried customers count is slightly more than Married customers.

Number of premiums paid



- Data varies from 2 to 60 with majority data falling between 5 to 15 (right skew)

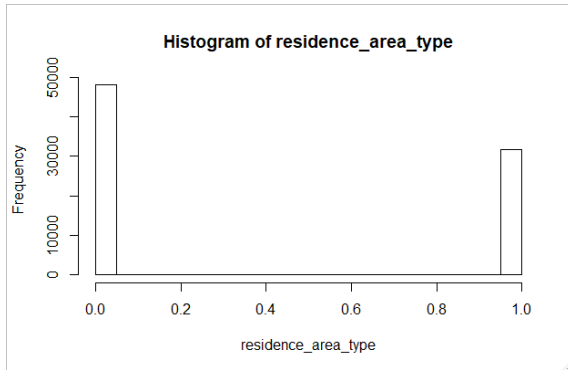
Premium



- Mean = 10.86
- Data has too many outliers

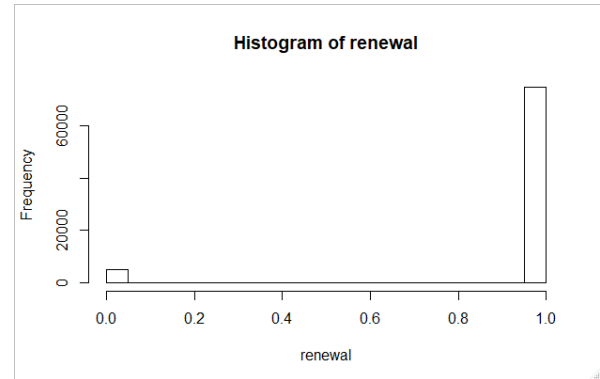
- Data varies from 1200 to 60000 with majority data falling between 5000 to 10000 (right skew)
- Mean = 10925
- Data has too many outliers

Residence area type



- Data has 2 categories with more number of Urban (0) cases than Rural (1)

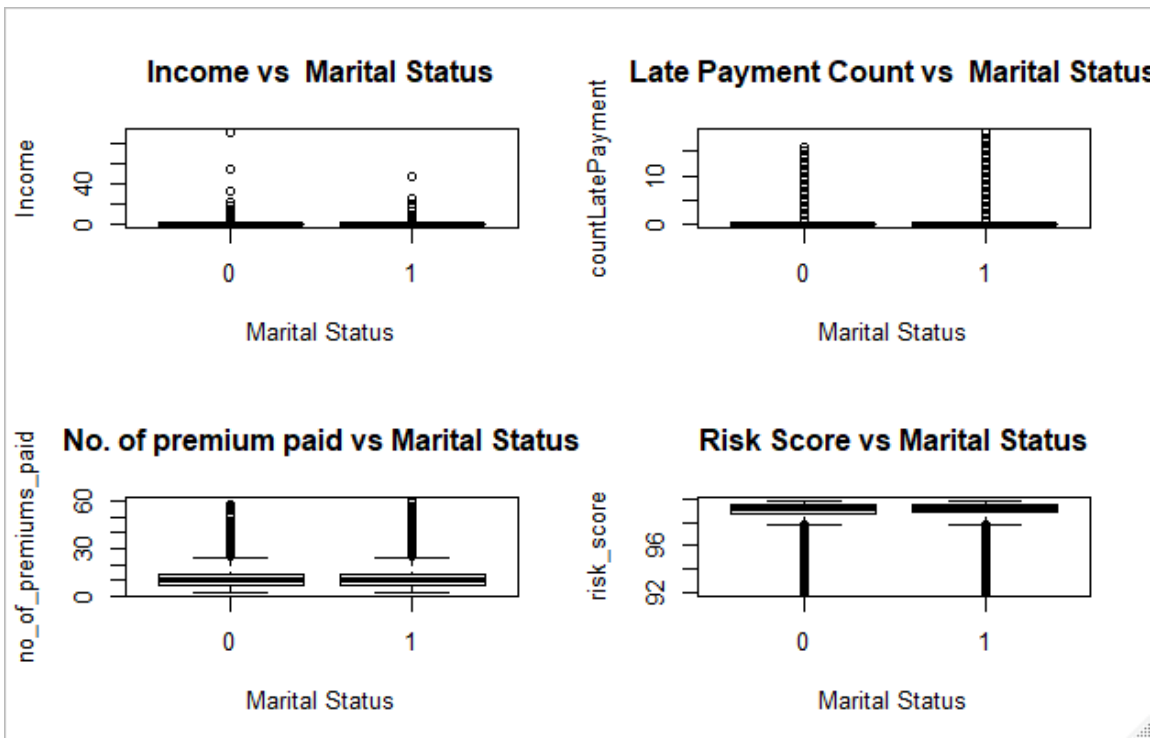
Renewal



- Data has 2 categories with more number of renewed cases than non-renewed cases. It may lead to data imbalance problem which needs to be properly handled

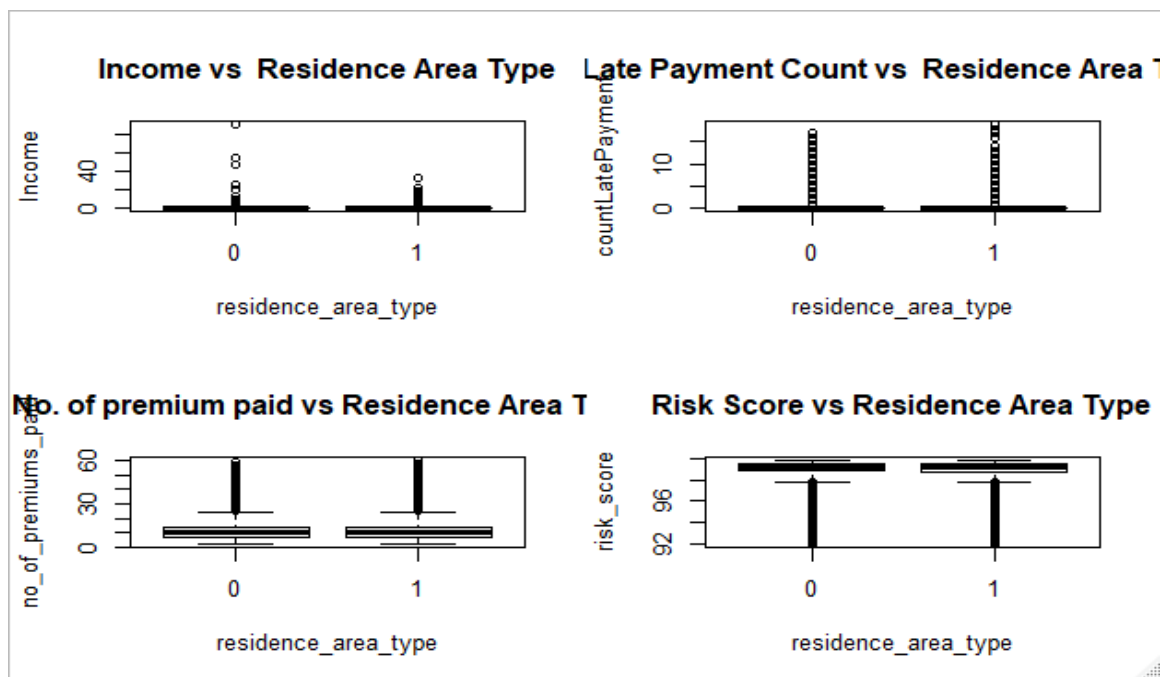
Bivariate Data Analysis

- Marital status vs Income, Late Payment, No of premium paid & Risk score



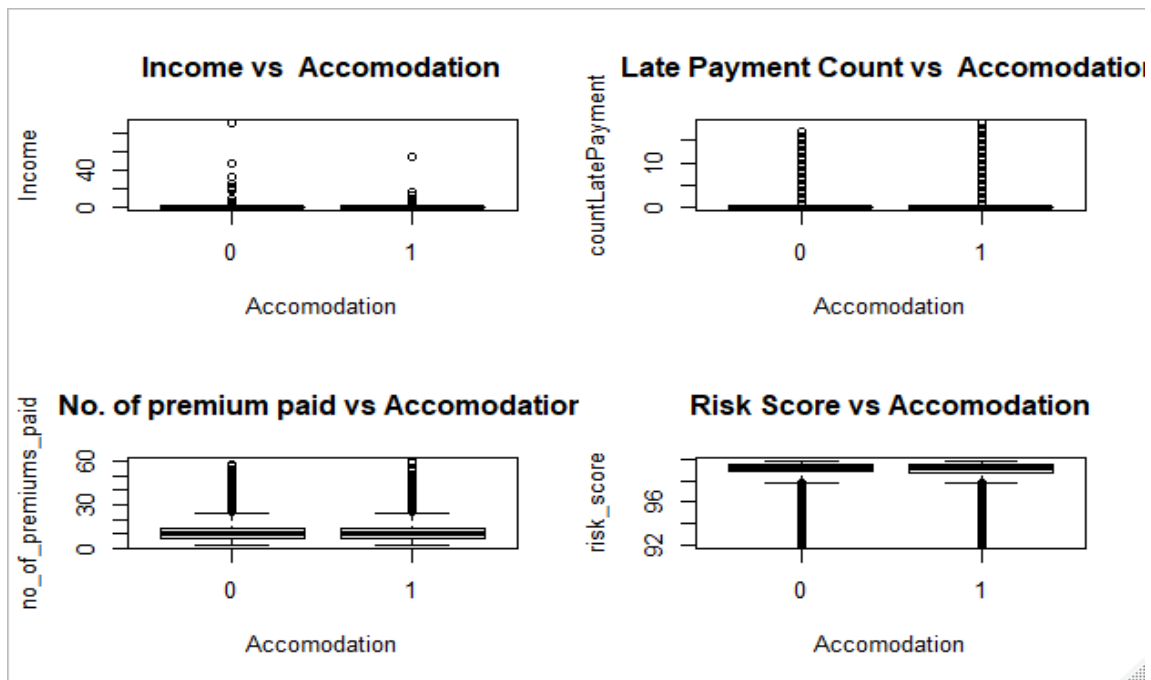
No significant difference across parameters between Married and Unmarried customers.

- Residence Area Type vs Income, Late Payment, No of premium paid & Risk score



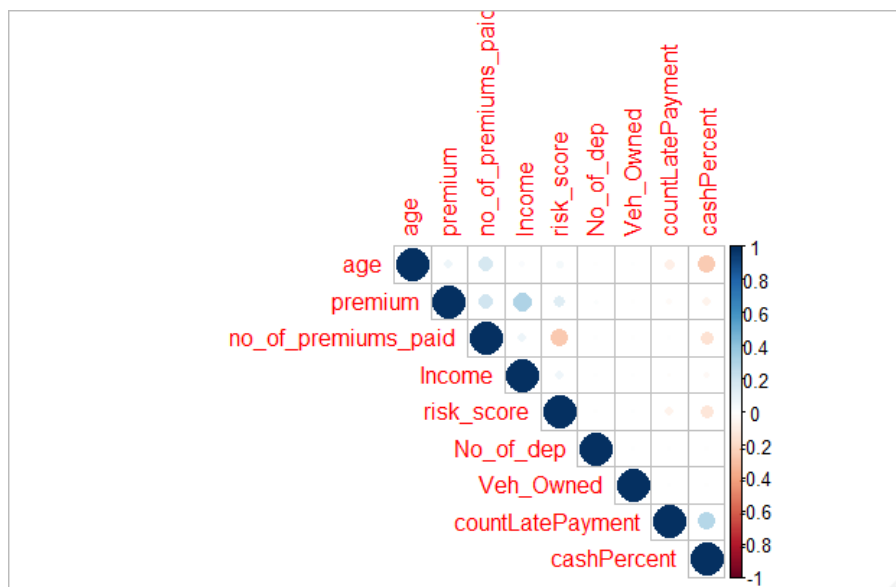
No significant difference across parameters between Urban and Rural resident customers.

- Accommodation vs Income, Late Payment, No of premium paid & Risk score



No significant difference across parameters between Rented and owned apartment customers.

Multivariate Data Analysis



- Positive correlation between Premium and Income.
- Low Positive correlation between Cash Premium Percent and Count of late payment.
- Negative correlation between Age and Cash Premium Percent.
- Negative correlation between Cash Premium Percent and No of premiums paid.
- Low Positive correlation between Premium and No of premiums paid.
- Low Positive correlation between Age and No of premiums paid.
- Negative correlation between Risk Score and No of premiums paid.

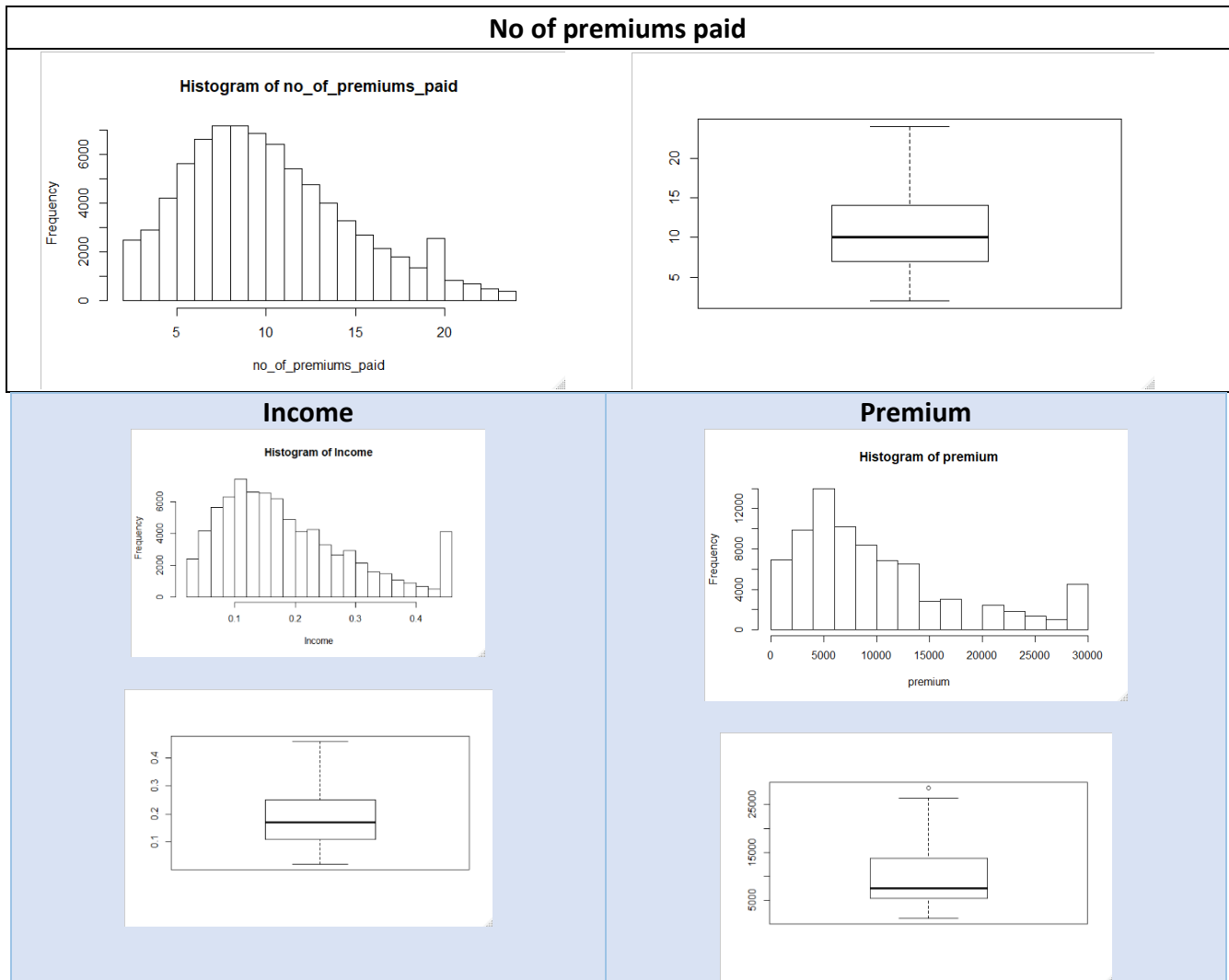
- Low Negative correlation between Risk Score and Cash Premium Percent.

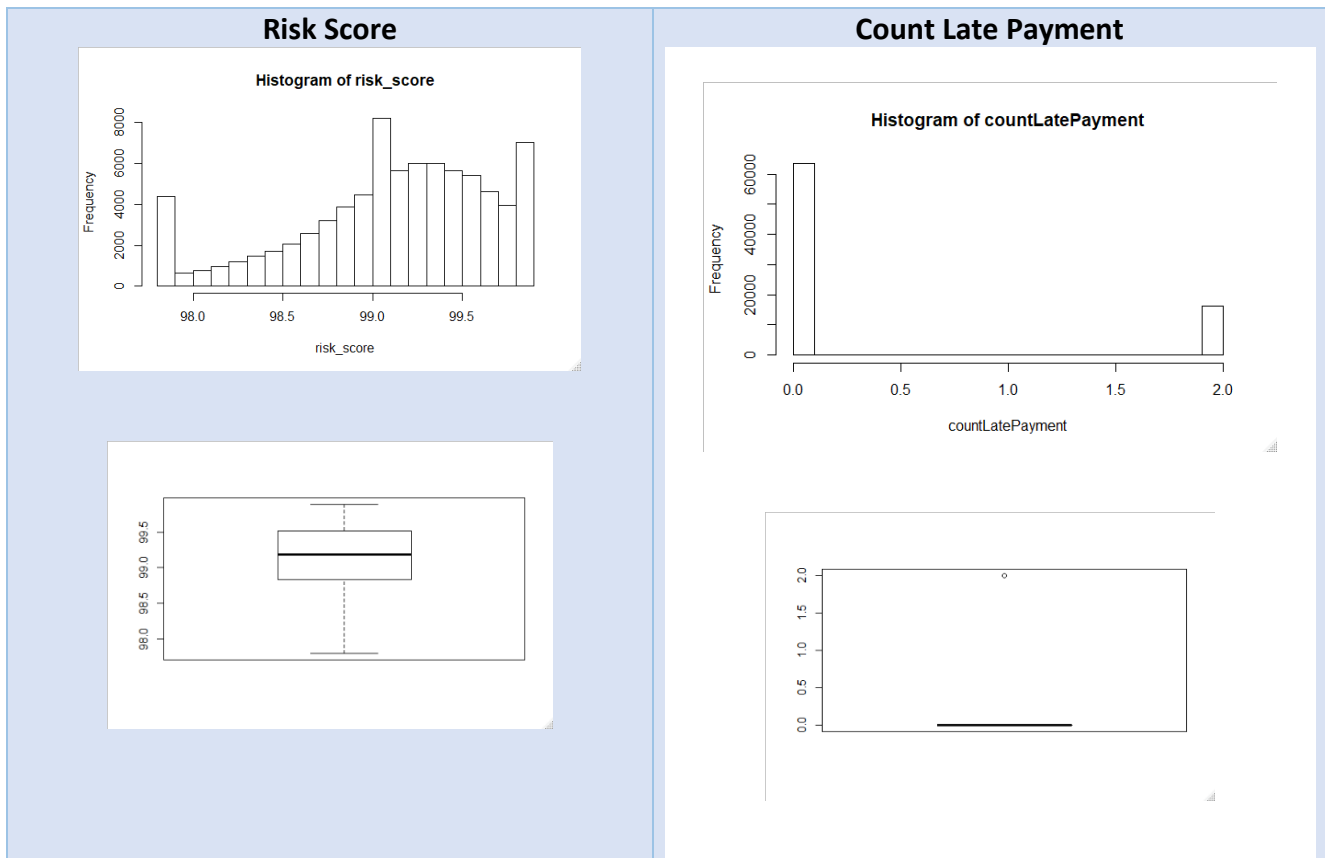
There is no high correlation among variables, in general. However, from the above we can infer that:

- Customers making higher % of cash payment are likely to make more delayed payments and are likely to have lower Risk Score.
- Higher age customers have paid more number of premiums but lesser premium amount in cash.
- Higher Income customers are likely to pay higher Premium.

Outlier treatment

Affected variables are treated for outliers using capping methodology . Distribution of treated variables is shown below.





Data normalization

Data variables are normalized to avoid any one variable overshadowing the model and data remains uniform. Function used to normalize any variable x is:

$$\text{Function}(x) = \frac{(x - \min(x))}{(\max(x) - \min(x))}$$

Variables normalized are **income**, **risk_score**, **premium**, **age**, **cash_Percent**, **no_of_premiums_paid**.

Synthetic Minority Over-sampling Technique (SMOTE)

It is a methodology to handle class imbalance problems. This is a statistical technique for increasing the number of cases in your dataset. The module works by generating new instances from existing minority cases.

In the dataset there is a clear class imbalance as renewal has only 6% of cases which has defaulted and remaining are No default cases.

We split the data such that we have 70% of the data is Train Data and 30% of the data is my Test Data and synthetically add entries to make it balanced.

Train data – SMOTE

Prior to smote operation no of entries for renewal

0	1
3475	52300

Post smote operation no of entries for renewal

0	1
20850	34750

Test data - SMOTE

Prior to smote operation no of entries for renewal

0	1
1523	22555

Post smote operation no of entries for renewal

0	1
16753	15230

Interpretation - With SMOTE operation new instances have been added to both test and train dataset hence addressing the imbalance problem.

Modelling

1) Logistic Regression

Logistic regression is a predictive analysis technique. It is used to predict a classification problem. It is easier to implement and makes no assumptions about the distribution but in this model multi-collinearity among variables can affect the outcome.

Some equations that are used are:

- Activation Function -Sigmoid function= $\text{sig}(x)=1/(1+e^{-x})$
- Cost/Error function= $-1*\sum y*\log y'+(1-y)*\log(1-y')$ where y and y' values are taken for limits i=1 to i=output size. Here y' is the scalar value in the model output and y is the corresponding target value for all i's.

Reason for the choice of this cost function is that it is differentiable which is needed for almost all the optimizers like gradient descent to optimize the weights which restricts us from using discrete values error function like $\sum (y_p - y_a)$ where y_p is the predicted class and y_a is the actual class as it can't be optimized. Moreover '-' sign is to maximize the probability by minimizing the los function .

Decreasing the cost will increase the maximum likelihood.

Steps:

- Analyze the Base data provided to us vis-à-vis the modified data and test if the modification is adding value to the model.
- Base data has individual columns for count for late payment (3_6 months , 6_12 months, >12 months) and modified data has single aggregated column for count of late payments.
 - Run Logistic Regression function on train data and observe the significant variables
 - Re-Run Logistic Regression function with significant variables
 - Build the prediction model
 - Use test data to analyze the model

Results:

- Based on Logistic regression with both Base and Modified data, “Vehicles Owned” , “Sourcing_channel” and “Residence area type” are insignificant variables. Intercept is significant in both models.
- However, as Modified data model is not adding any value to the Base data model, we will continue with the Base data model i.e. individual values for count of delay columns.
- Based on Logistic Regression, Income, `Count_3-6_months_late`, `Count_6-12_months_late`, Count_more_than_12_months_late, Marital Status, No_of_dep, Accommodation, risk score, number of premiums paid, premium, cash_Percent and age are significant variables.
- Regression equation is $\log \text{odds}(y) = 0.54873 + 28.58339 * \text{Income} - 7.57962 * \text{Count_3-6_months_late} - 20.56685 * \text{Count_6-12_months_late} - 10.71174 * \text{Count_more_than_12_months_late} + 0.05883 * \text{Marital Status} - 0.09284 * \text{No_of_dep} - 0.04515 * \text{Accommodation} + 1.23170 * \text{risk_score} - 1.96536 * \text{no_of_premiums_paid} + 0.50082 * \text{premium} - 1.89267 * \text{cashPercent} + 1.49127 * \text{age}$**
- Income, Marital Status, risk_score , premium , age have positive coefficients which means higher values of these variables will result in a likely renewal.
- `Count_3-6_months_late`, `Count_6-12_months_late`, Count_more_than_12_months_late, No_of_dep, Accommodation, number of premiums paid, cashPercent have negative coefficients which means higher values of these variables will **NOT** result in a likely renewal.

2)Random forest

Decision Trees generally tend to overfit data and are usually very sensitive to change in data. Random Forest is a technique used for better accuracy. It randomly selects observations and specific features to build multiple decision trees and then average the results (i.e. means for a continuous variable) or use the most popular prediction (i.e. modes for a classification variables) across all the trees for a robust prediction. The general procedure of using multiple trees to obtain better performance is called ensemble learning. The steps for this model are:

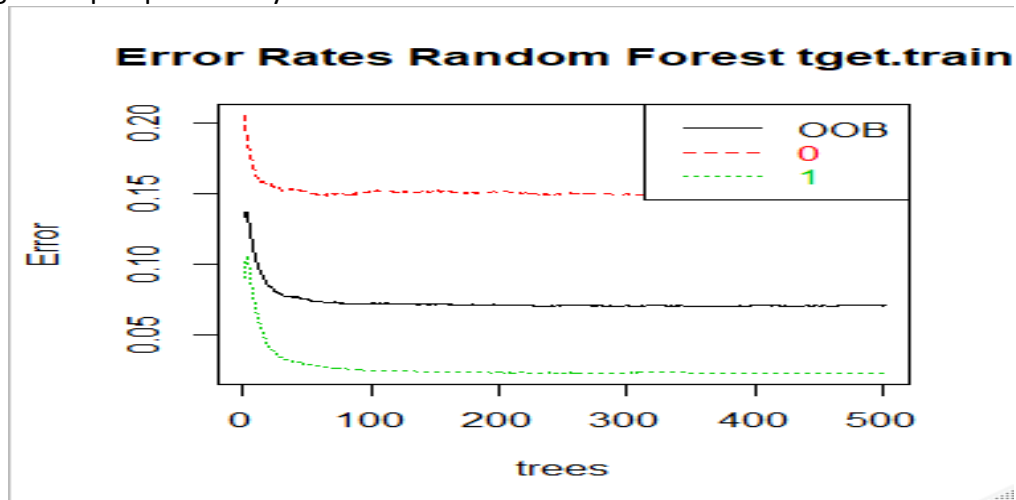
Impurity	Task	Formula	Description
Gini impurity	Classification	$\sum_{i=1}^C f_i(1 - f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.
Entropy	Classification	$\sum_{i=1}^C -f_i \log(f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.
Variance / Mean Square Error (MSE)	Regression	$\frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$	y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N} \sum_{i=1}^N y_i$
Variance / Mean Absolute Error (MAE) (Scikit-learn only)	Regression	$\frac{1}{N} \sum_{i=1}^N y_i - \mu $	y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N} \sum_{i=1}^N y_i$

Steps:

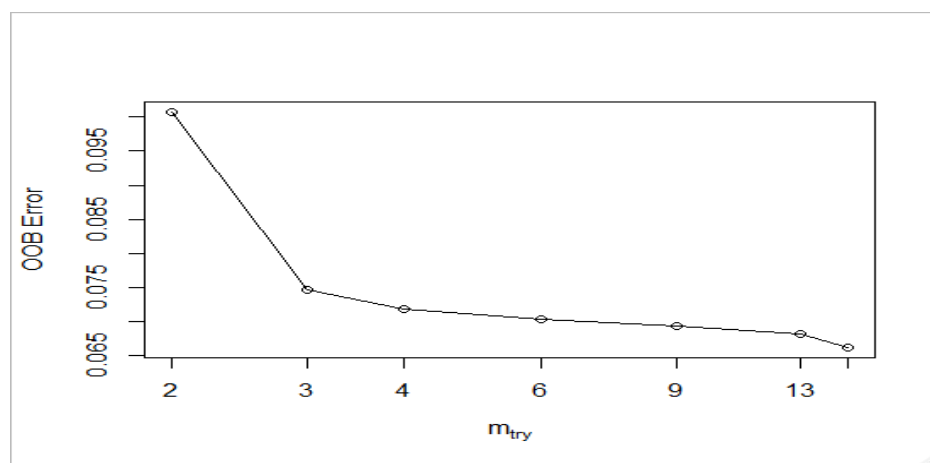
- Generate random forest using ‘renewal’ as dependent variable and others as independent variable.
- Predict values and assess model performance

Results:

- The error rate plot w.r.t number of trees reveals that 50 trees is a reasonably good assumption as error rate decrease is minimal or absent post that value. So, we can assume odd value of 51 trees. Odd number of trees are chosen to avoid the chance that the classifier gives equal probability for the two classes.



- Now we will “tune” the Random Forest by trying different m values. We need to consider the optimal number of variables at each internal node in the tree.
- mtry defines the number of variables randomly sampled as candidates at each split.
- From the graph below mtry value of 15 has the least OOB (Out of Bag) error.



- Based on Random Forest, Income, Count_3_6_months_late, Count_6_12_months_late, Count_more_than_12_months_late, No_of_dep, risk_score, no_of_premiums_paid, sourcing_channel, premium, cashPercent, age are significant variables.

Model Performance Measurements

We have compared the two models to see which is better and more reliable for drawing conclusions.

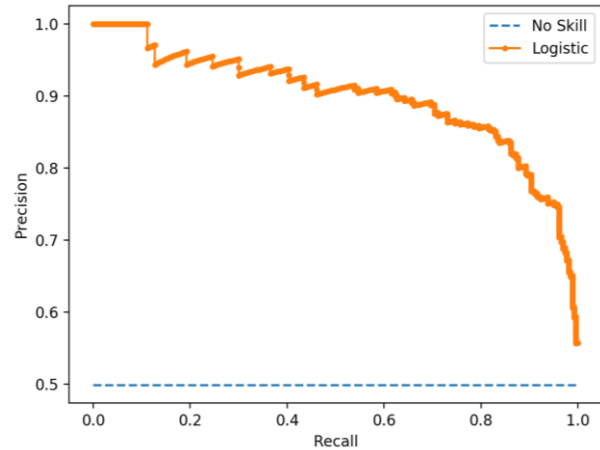
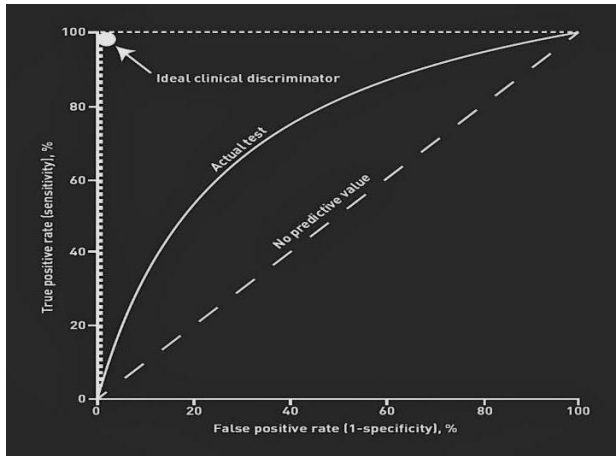
Parameter	Logistic regression	Random Forest
Classification Error Rate (CER)	0.24	0.10
Accuracy	0.76	0.90
Specificity (TN/(TN+FP))	0.65	0.85
Sensitivity (TP/(TP+FN))	0.89	0.96
Area Under Curve (AUC)	0.85	0.97
K-S	0.54	0.82

TN-True Negative FP-False Positive TP-True Positive FN-False Negative

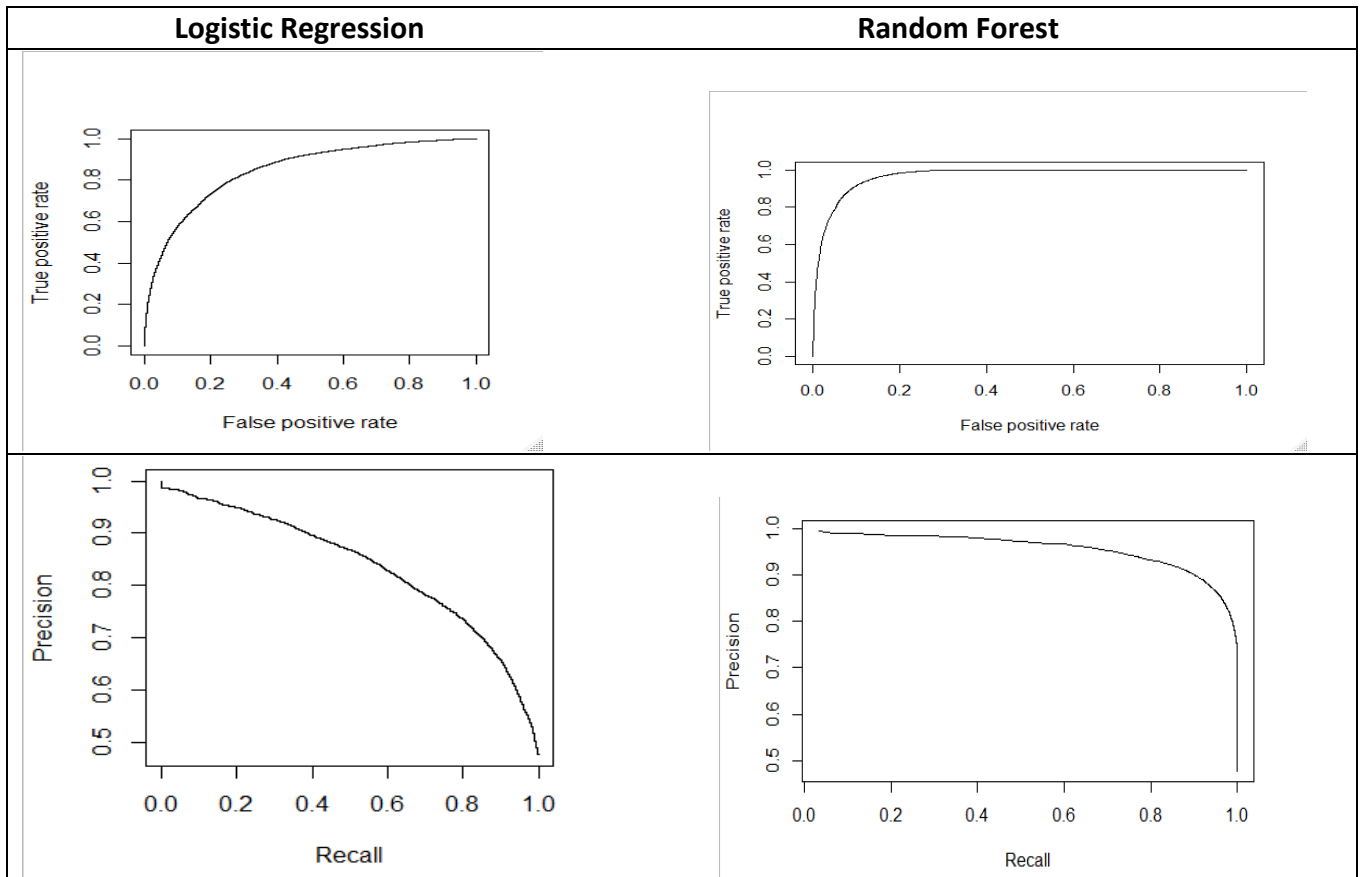
- The **Area Under the Curve (AUC)** is the measure of the ability of a classifier to distinguish between classes.
- The higher the AUC value for a classifier, the better its ability to distinguish between positive and negative classes.
- AUC near to the 1 which means it has a good measure of separability. A poor model has AUC near to the 0 which means it has the worst measure of separability.
- Sensitivity measures the proportion of positives that are correctly identified
- Specificity measures the proportion of negatives that are correctly identified
- K-S is a measure of the degree of separation between the positive and negative distributions. Therefore, higher K-S indicates better performance.

ROC curve and Precision-recall curve

- A ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR). The TPR is the proportion of observations that were correctly predicted to be positive out of all positive observations ($TP/(TP + FN)$). Similarly, the FPR is the proportion of observations that are incorrectly predicted to be positive out of all negative observations ($FP/(TN + FP)$).
- A ROC curve shows the trade-off between sensitivity (TPR) and specificity ($1 - FPR$). If the curve is closer to the top-left corner it indicates a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal ($FPR = TPR$). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.
- A precision-recall curve shows the relationship between precision (positive predictive value) and recall (sensitivity) for every possible cut-off.
- **Precision** = $TP / (TP + FP)$
- **Recall** = $TP / (TP + FN)$
- The closer a Precision-Recall Curve is to the upper right corner, the better the performance is.



Comparison of the curves for the two models



Interpretation of Model Measures:

- Random Forest has a lower CER
- Random Forest has a higher accuracy
- Logistic Regression has a lower specificity.
- Logistic Regression has a lower sensitivity
- Logistic Regression has a lower AUC
- Random Forest has a higher K-S

- As mentioned above from the curves it can be seen that the ROC curve for Random forest is closer to the left corner and the PRC curve is closer to the right corner.

Therefore, from above comparisons it can be seen that Random Forest has overall better performance indicators.

Summary

- The dataset consists of 17 variables and 79853 customer observations with a combination of Indicator and continuous variables.
- Data mainly covers Customer's demographic information, premium payment related behavior and Risk profile information.
- 'renewal' is the target or the response variable i.e. the Dependent variable and other variables would be independent or the predictor variables.
- There is no missing value in the data.
- Data has outliers present and is skewed on most of the numeric variables.
- We treated the variables for the outliers.
- Most of the categorical variables have equal representation of categories. 'renewal' has higher count of renewed cases than non-renewed cases. It may lead to data imbalance problem which needs to be handled. We addressed the same through methods like SMOTE.
- For better readability, we have converted Age in years and cash premium payment percentage in %
- As per bivariate analysis, there is no significant variation in the premium, count of late payment, risk score and No of premium paid vis-à-vis Marital status, Accommodation and Residence Type
- No significant correlation is present among variables. Few important inferences which we can draw from correlation plots are - Customers making higher % of cash payment are likely to make more delayed payments and are likely to have lower Risk Score. Higher age customers have paid more number of premiums but lesser premium amount in cash. Higher Income customers are likely to pay higher Premium.
- Based on Model performance measures, we observed that Random Forest has the best performance indicators.
- As per both Logistic Regression and Random Forest a mix of demographic and financial parameters are significant:
 - Out of the 15 variables, there are 12 significant variables as per both models, 11 being common.
 - Logistic Regression indicated that "Sourcing Channel" is not significant and Random Forest indicated that "Marital Status" is not significant variable for the study.
 - Both models confirmed that "Vehicle owned" and "Residence area type" are NOT significant variables.
- To further analyze whether there is a positive or negative impact:
 - Higher 'Income', 'Marital Status', 'Risk score', 'Premium' and 'Age' will result in a likely renewal but higher Count_3-6_months_late, Count_6-12_months_late, Count_more_than_12_months_late, 'No of dependent', 'Accommodation', 'No of premiums paid', and 'cash Percent' will result in a likely non-renewal.
- Businesses can use this information in their strategy formulation and can collect more data to find deeper correlations between renewal and different variables.

Conclusion

In this paper we have studied various aspects of the renewal of life insurance in India. We have analyzed dataset covering Customer's demographic information, premium payment related behavior and Risk profile information. After modelling for various determinants of life insurance renewal we found that most of the expected relationships between variables and renewal are supportive of our hypothesis. However, it is somewhat surprising to see that accommodation ownership, number of premiums paid and number of dependents have a negative effect. Possible reasons could be that as someone has paid a greater number of premiums, they start feeling that their current situation doesn't require renewal of life insurance or with the increase in number of dependents different kinds of other expenditures increase as well such as education, clothing, food etc.

Contrary to our initial expectation the model suggested that if a customer owns the accommodation, they are less likely to renew their insurance. To understand possible reasons for this further analysis is required to determine usage of this as a factor influencing their policies and strategy formulation. Insurance companies can target the younger population for new customers by redesigning their policies or products. Since the life expectancy of this group of people are high their survival/maturity can be given a thrust i.e it can be curtailed to look like an investment plan. Similarly, in rural areas women have a higher life expectancy and policy changes can be made to target them.

References

- Base paper- Ma, Y. (2020) Prediction of Default Probability of Credit-Card Bills. *Open Journal of Business and Management*, 8, 231-244
 - <https://www.scirp.org/journal/paperinformation.aspx?paperid=97459>
- International Journal of Marketing, Financial Services & Management Research Vol.1 Issue 7, July 2012, ISSN 2277 3622
 - <http://indianresearchjournals.com/pdf/ijmfsmr/2012/july/9.pdf>
- <https://www.sciencedirect.com/science/article/pii/S1532046412000883>
- Boodhun, N., Jayabalan, M. Risk prediction in life insurance industry using supervised learning algorithms. *Complex Intell. Syst.* 4, 145–154 (2018)
 - <https://link.springer.com/article/10.1007/s40747-018-0072-1#citeas>
- Johnson, Eric J., John Hershey, Jacqueline Meszaros, and Howard Kunreuther. "Framing, probability distortions, and insurance decisions." *Journal of risk and uncertainty* 7, no. 1 (1993): 35-51.
- Neter, John, and C. Arthur Williams Jr. "Acceptability of three normative methods in insurance decision making." *Journal of Risk and Insurance* (1971): 385-408.
- Beck, Thorsten, and Ian Webb. "Economic, demographic, and institutional determinants of life insurance consumption across countries." *The World Bank Economic Review* 17, no. 1 (2003): 51-88.