CAPSTONE PROJECT

Insurance Premium Default Propensity
Final Submission

Abstract

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.

KUDAKWASHE NYIKADZINO

Table of Contents

1) Executive Summary **2**

2) Introduction **3**

Defining Problem Statement 3

3) Data Report **4**

a) Data Collection 4

b) Dataset Inspection 4

c) Dataset Attributes 5

4) Initial Exploratory Analysis **7**

a) Univariate Analysis 7

b) Bivariate Analysis 18

5) Variable Relationships and Insights **24**

a) Variable relationships and importance 24

b) Insightful Visualisations 26

6) Data Pre-Processing **29**

a) Removal of unwanted variables 29

b) Missing values treatment 29

c) Outlier treatment 29

d) Variable transformation 30

7) Analytical Approach **32**

8) Modelling Process **33**

a) Modelling and validation 33

b) Algorithms and Models 33

9) Modelling Comparison **42**

10) Interpretation of the best model **44**

11) Insights, Recommendations and Conclusion **45**

12) Annexures **46**

1) Executive Summary

The Insurer in our study is concerned with the impact that defaulting customers have on its revenues. The Insurer would like to identify these customers and provide risk mitigation strategies that reduce the impact of these customers on revenue. The Insurer also needs a model that can identify customers that are likely to default and use the model as a tool to offer or decline insurance cover. Going through this study will highlight those factors that the Insurer can focus on to increase its revenues. There are additional benefits to society that may result from this study like potential reduction in premiums due to lower risk customers improving better participation in the financial system.

The dataset used in the study contains customer specific features like age, income, marital status for example. Our focus is whether any of these features can assist in determining whether a customer is likely to default on their premium. We start by looking at each of these features through exploratory data analysis. We then look at how these features relate to each other and most importantly how they relate to our target variable default. We then start preparation of the data for modelling by treating for outlier observations, scaling these different features as well as normalisation. These processes ensure we reduce bias and improve the performance of the models. We proceed to split our dataset into two to allow us to validate the results of our model. This problem is a classification problem, so we apply different classification algorithms and select the best model based on a range of model evaluation metrics.

We then proceed to the final stage of the study where we identify the most important features or characteristics. These characteristics help in determining whether they are likely to default on their premiums or not. We interpret the result of the chosen model and provide justification for the selection.

One important insight is that the percentage of premium paid by cash is the most important variable as identified by all models. Knowing this information provides an important feature which helps us determine whether a customer will likely default or not. The Insurer should make sure this data point should be provided by every customer. If not provided, it can be used as a qualifier to deny insurance or adjust the premium paid. The more cash that is being used, the more likely customers default. Providing alternative non-cash methods to pay for premiums can be used to reduce defaults.

Some information does not provide relevant information on whether a customer will default or not. An example is knowing whether a customer stays in rural or urban areas. This information can be removed from application forms which makes the process easier for customers and reduces the burden of capturing irrelevant information for the business.

Some distribution channels have been identified as unimportant and these can me removed from the marketing approach. The business can then focus on those that are likely to improve its revenues with lower likelihood of default.

Finally, the model selection required a balance between highly accurate model against a model that performed well in identifying customers likely to default. The chosen model performed well in identifying defaulting customers at the expense of overall accuracy. This means that will be instances when customers are incorrectly classified as defaulters when they are no which means reduced opportunities for the business. This is a trade-off the business may be willing to take knowing that the model correctly identifies defaulters hence reducing the risk on revenue which is the focus of the study.

2) Introduction

a. Defining problem statement

Defaulting clients have a negative impact on revenues of insurers. Understanding which clients are likely to default will help in reducing the risk associated with lower revenues. An insurer would be best served by knowing who these customers are upfront. This will allow them to either adjust premiums or decline insurance coverage. Building a model will assist in predicting which customers are likely to default and help insurers reduce the risk associated with defaults.

b. Need of the study/project

i. Identify customers likely to default and take precautionary action.
ii. Reduce risk exposure which come from defaults by identifying factors that increase defaults.
iii. Develop a business model that is based on the application of the resulting model instead of overly relying employees to make subjective risk assessments and better control risk of default.
iv. Identify factors that will help increase revenue.

c. Understanding business/social opportunity

i. Reducing default rates can assist in lowering risks and allow for reduction of insurance premiums which will help improve the competitiveness of the company.
ii. Reduction of default rates will allow for lower premiums and this in turn will make the insurance offering available to more customers improving participation in the financial systems by more people

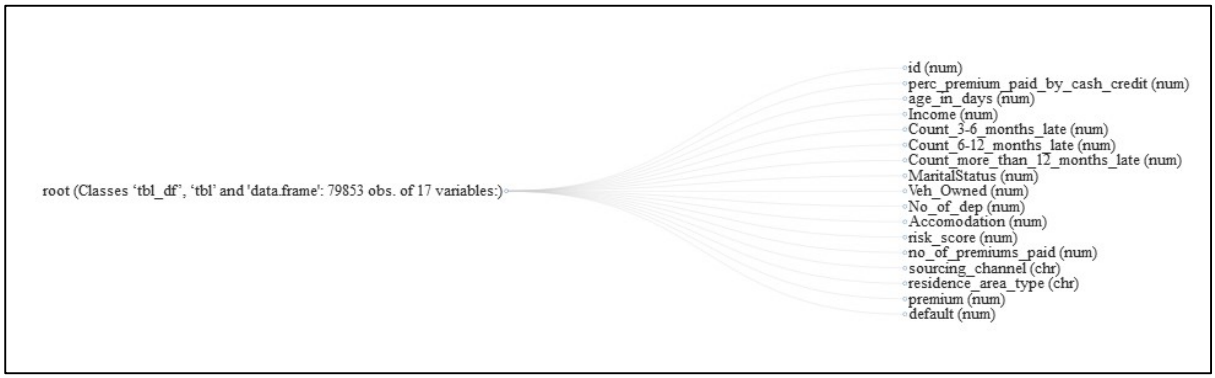
3) Data Report

a. Understanding how data was collected in terms of time, frequency and methodology

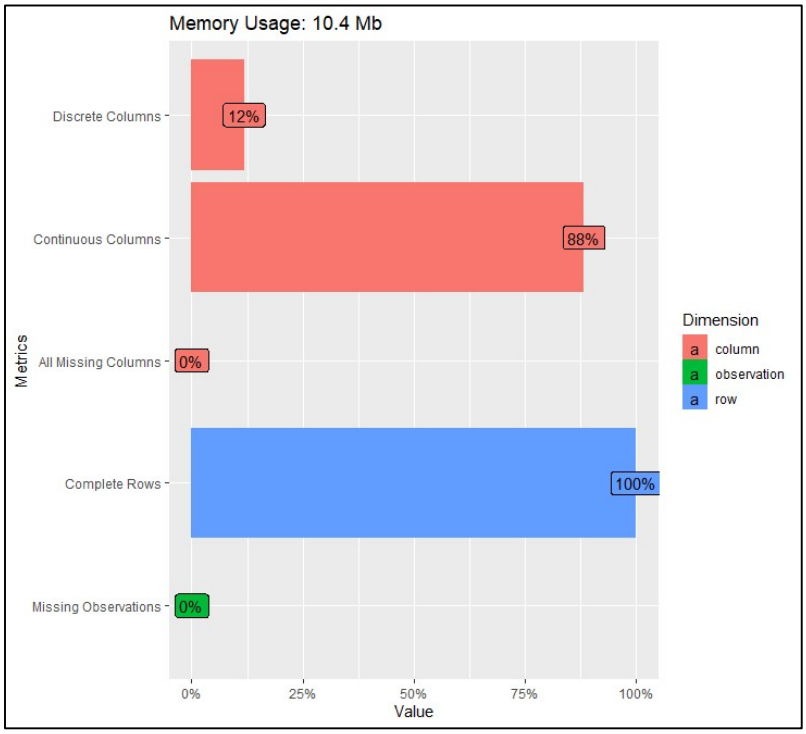
Based on the sequential nature of the unique customer ids, this does not seem to be a random sample of policy holders included in the dataset. The data was collected for the first customer in chronological order. There is no detail of whether the data represents the top N policy holders or whether it is the entire customer base for the company. This is cross sectional data which shows a snapshot of a customer and their characteristics at a point in time.

b. Visual inspection of data (rows, columns, descriptive details)

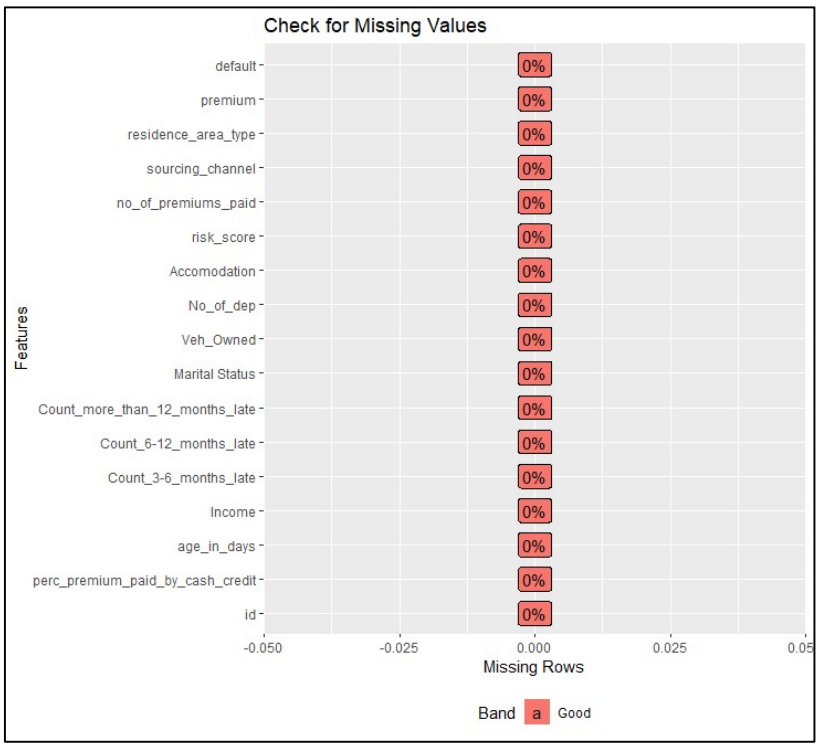
The dataset has 79853 observations and 17 variables (shown below). There are 15 numeric variables and two character variables.



Discrete columns make up 12% with 88% being continuous (below). None of the columns contain missing values and this means there are no missing observations in the dataset.



The chart below further confirms that there are no missing values in any of the columns in the dataset.



c. Understanding of attributes

Character variables will need to be changed to factor variables to reflect the different variable levels. The following variables will be changed to factors to reflect the correct variable type:

- id (unique customer id)
- MaritalStatus
- Accommodation
- sourcing_channel
- residence_area_type
- default (target variable)

Variable names will be changed to ensure naming consistency. Some variable have underscores and others have spaces or hyphens which is inconsistent.

The id variable after conversion to a factor variable shows that there are 79853 unique values which means there are no duplicate customer records based on the unique identifier. The variable will not be included in further analysis because it does not add additional value beyond confirming the absence of duplicate customer records.

The age variable is currently given in days. This will be converted to years without all age values rounded down to the last birthday which is consistent with one of the methods used by insurers. The variable is renamed to age_in_years.

The table below shows a description of each variable together with the old in the data set.

Old Variable Name	New Variable Name
id	id
perc_premium_paid_by_cash_credit	cash_premium_percent
age_in_days	age_in_years
Income	income
Marital Status	marital_status
Veh_owned	vehicles_owned
Count_3-6_months_late	late_3_6_months
Count_6-12_months_late	late_6_12_months
Count_more_than_12_months_late	late_over_12_months
Risk_score	risk_score
No_of_dep	dependents
Accommodation	accommodation
no_of_premiums_paid	no_of_premiums_paid
sourcing_channel	sourcing_channel
residence_area_type	residence_area_type
premium	premium
default	default

4) Initial Exploratory data analysis

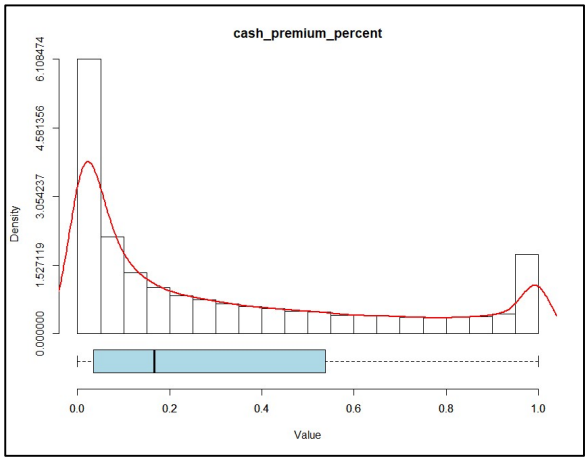
a. Uni-variate analysis

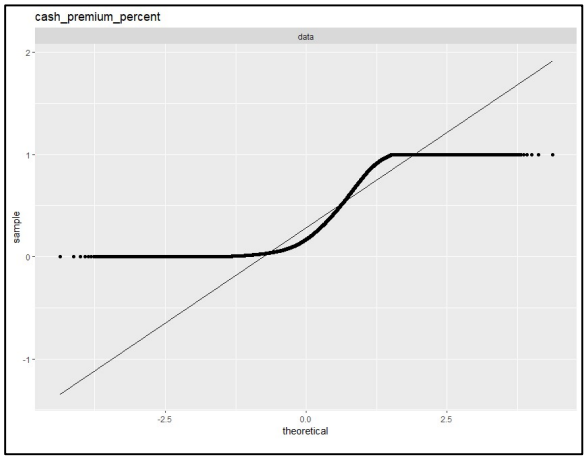
i. Percentage of premiums paid by cash

The percentage of premiums paid by cash is right skewed with a smaller peak on the right of the distribution (chart below). The second peak shows some customers are paying 100% of their premiums in cash. The right tail of the box and whisker is significantly larger than the left tail confirming the skew. The mean is almost double the median at 31.4% and 16.7% respectively. The QQ plot (below) shows the distribution is not normal with many points away from the normal distribution diagonal line.

The minimum percentage premium being paid by cash is 0% which is clients that are not paying a proportion of their premium as cash. The maximum is 100% and these are clients that are paying all cash for their premiums making the range 100% as well. The proportion of total customer not paying any cash is 7.2%. The standard deviation is 33.4% and the range is high which indicates high dispersion. The coefficient of variation is close to 1 indicating dispersion that is not too high.

There are no outliers in this distribution though it is siginifcantly right skewed.



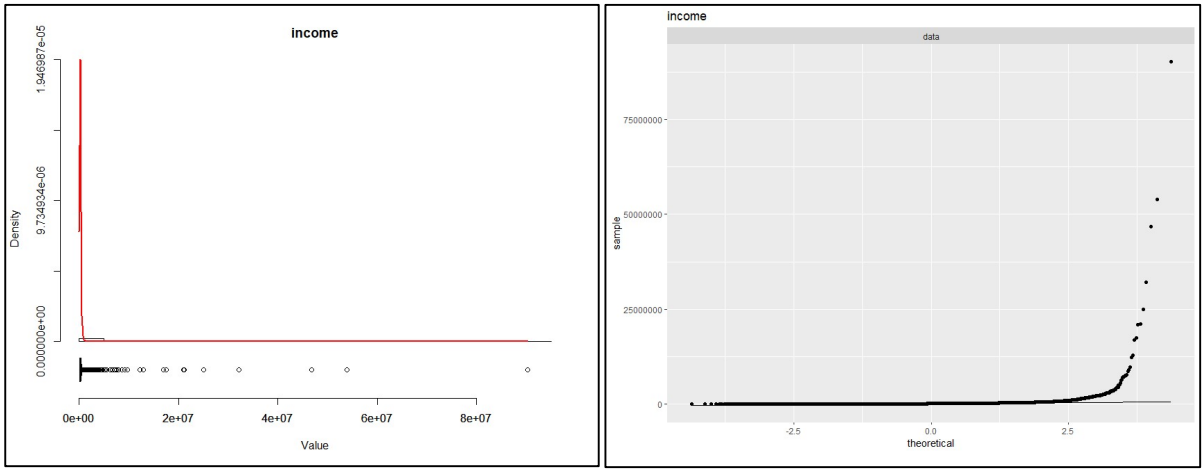


ii. Income

Income distribution is excessively right skewed (chart below). The right tail of the distribution is significantly larger than the left tail. The average customer income is 208 847.17 which is larger than the median income of 166 560.00. The QQ plot (below) shows there are points that are far from the diagonal line making this distribution different from a normal distribution.

All customers have an income with the lowest earning customer earning 24 030.00. The largest customer income is 90 262 600.00 which is more than 3 700 times larger than the lowest earning customer. This large income value significantly affects the mean which has an impact on other statistical measures which can cause unreliable results in the study. The range of the distribution as a result is significantly large at 90 238 570.00 indication a large dispersion of the distribution. The average dispersion from the mean of customer incomes is 496 582.60 as represented by the standard deviation. The coefficient of variation is 2.38 which also indicates a high degree of dispersion.

There is a significant number of outliers in this distribution compared to other variables. There are 3 428 outliers making 4.3% of all income values but the impact of these outliers on the distribution is large because they are very large values.

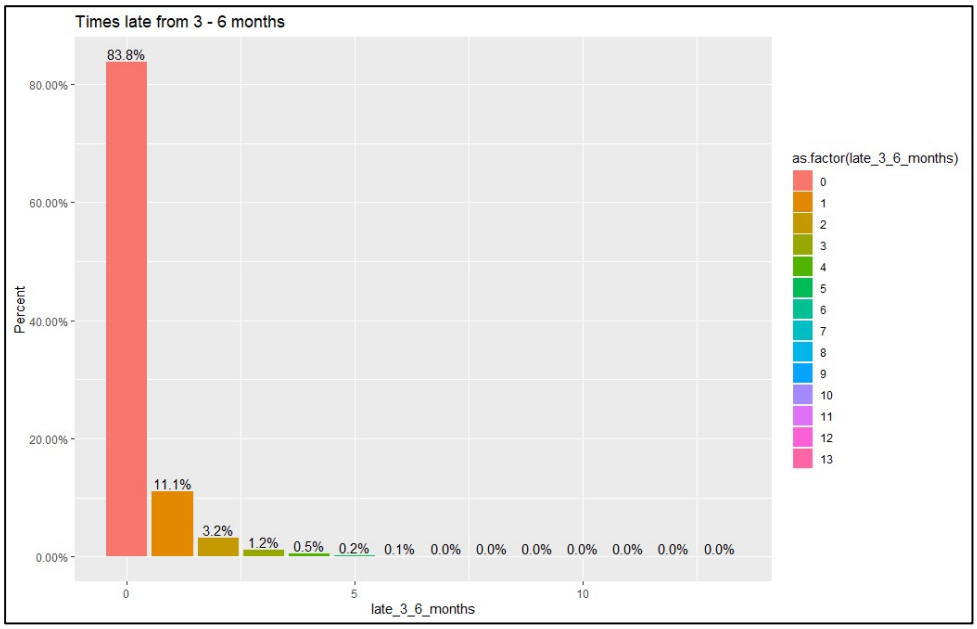


iii. Number of times premium was paid was paid late by 3-6 months

Most customers i.e. 83.8% have paid their premiums before the payments are between 3 and 6 months late. The distribution is right skewed and does not closely resemble a normal distribution. The average number of times payments have been made from 3 to 6 months is less than 1 at 0.25 which is close to 0 like the median. The standard deviation is 0.69 and the coefficient of variation is high at 2.78 reflecting large dispersion.

The maximum number of times someone has paid their premium late in these months is 13 times which also means the range is 13.

There are outlier customers that have paid their premiums late at least once. They are 12 955 clients who are outliers and they make up 16.2% of all customers and these may present a large default risk though this will be investigated later.

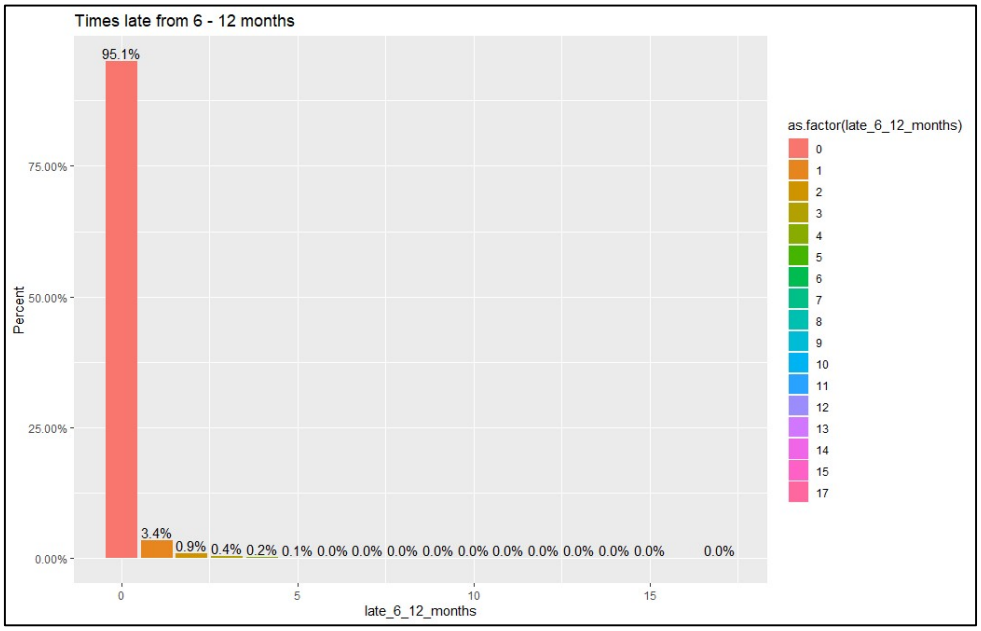


iv. Number of times premium was paid late by 6-12 months

The number of customers not paying their premiums between 6 and 12 months at least once has increased to 95.1% from 83.8% for those paying late at least once between 3 and 6 months. The distribution is right skewed with the average close to 0 at 0.08 which is like the median which is 0.

The range is 17 which is the same as the maximum number of times is an increase of 4 compared to the 3 – 6 months variable. One would expect the number of times customers pay their premium late to decrease with the increase of number of months they are late. This is not the case with the 6- 12 months' timeframe. The standard deviation is lower than the 3 – 6 months variable but the coefficient of variation is twice as large. This shows there is more dispersion in the number of times customers have been 6 – 12 months late compared to being late in the 3 – 6 months range.

There are 4.9% of customers who are outliers, and this is large decrease compared to the 3 – 6 months variable. This risk of likely default may however be higher since these customers take longer to make their premium payments.



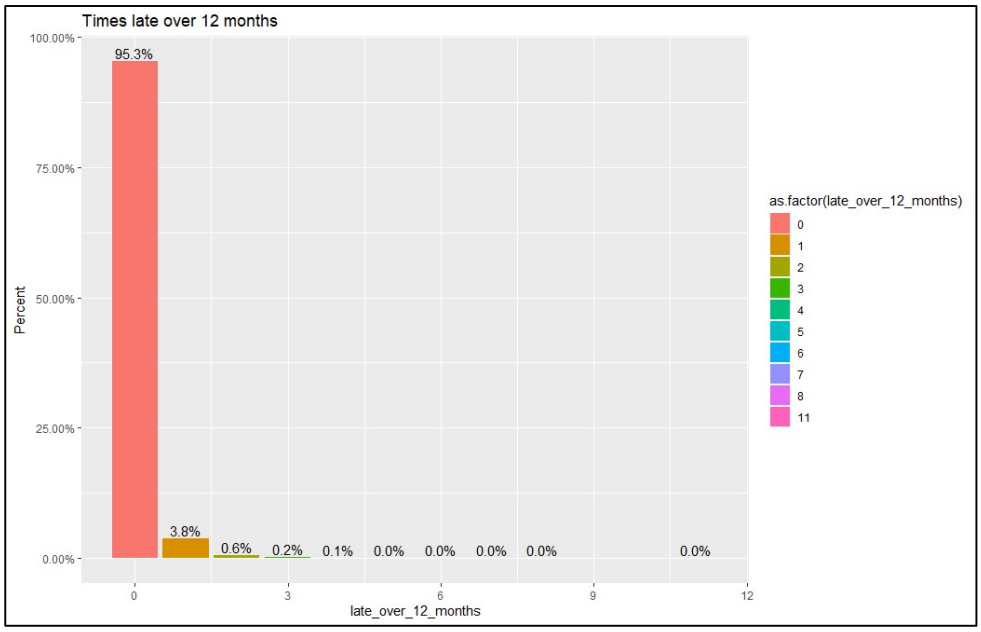
	cash_premium_percent	income	late_3_6_months	late_6_12_months	late_over_12_months
nbr.val	79853.00000	79853.00	79853.00000	79853.00000	79853.00000
nbr.null	5723.00000	0.00	66898.00000	75928.00000	76135.00000
nbr.na	0.00000	0.00	0.00000	0.00000	0.00000
min	0.00000	24030.00	0.00000	0.00000	0.00000
max	1.00000	90262600.00	13.00000	17.00000	11.00000
range	1.00000	90238570.00	13.00000	17.00000	11.00000
sum	25096.81900	16677073160.00	19833.00000	6236.00000	4786.00000
median	0.16700	166560.00	0.00000	0.00000	0.00000
mean	0.31429	208847.17	0.24837	0.07809	0.05994
SE.mean	0.00119	1757.30	0.00245	0.00154	0.00110
CI.mean.0.95	0.00232	3444.30	0.00479	0.00303	0.00216
var	0.11217	246594275898.97	0.47762	0.19031	0.09724
std.dev	0.33491	496582.60	0.69110	0.43625	0.31184
coef.var	1.06563	2.38	2.78256	5.58626	5.20296

v. Number of times premium was paid was paid late by over 12 months

Customers in this bucket pose a significant risk to the revenue the company receives since they have at least spent a year without making one of their premium payments. There is a slight increase to 95.3% of customers that have not missed any payment in a timeframe less than a year. This may indicate there is not much of a difference between customers making late premiums at least once between 6 – 12 months and from 12 months and above.

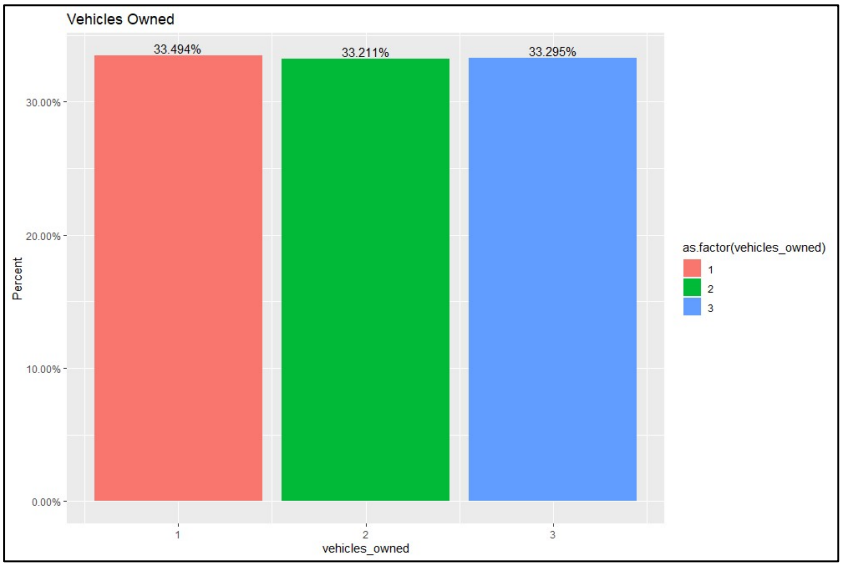
The distribution is also right skewed with a mean of 0.06 which is like the mode of 0 and does not resemble a normal distribution. The maximum number of times customers are late with payments is 11 which is lower than the 6 – 12 months' timeframe. The standard deviation is the lowest compared to the three time intervals at 0.31 but its coefficient of variation is the second largest of the three at 5.20. This is significantly higher than 1 which shows higher dispersion.

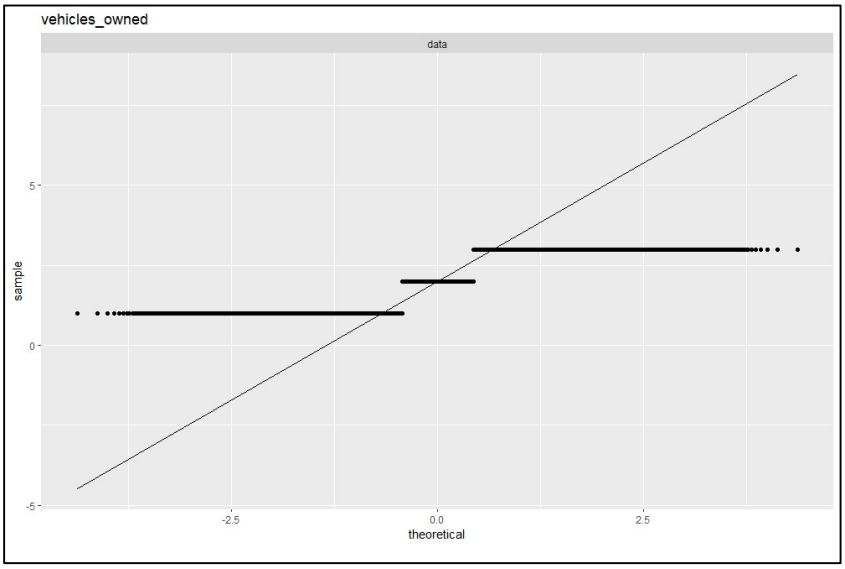
There are not outliers present in this distribution.



vi. Vehicles owned

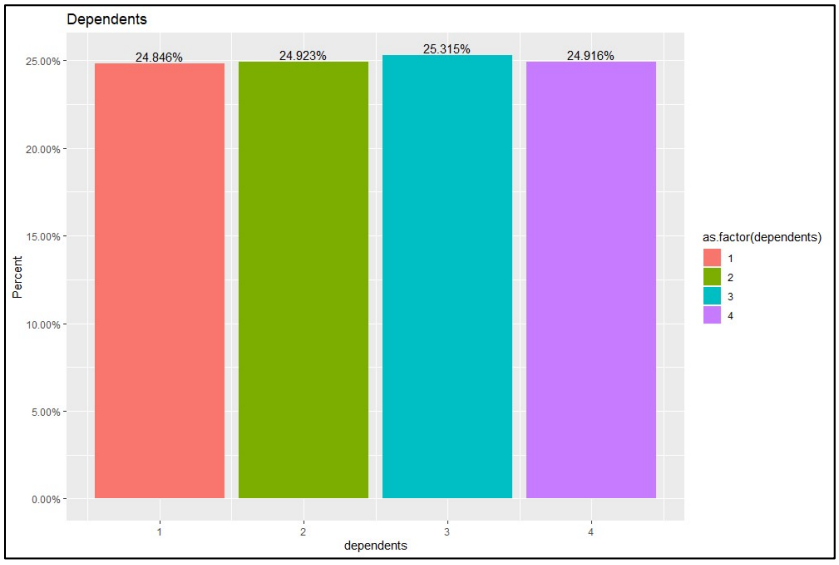
Customer own at least 1 and at most 3 vehicles and the distribution of ownership across the number of vehicles owned is similar at around 33% for each number owned. This strongly resembles a uniform distribution and confirmation is further provided by the QQ plot. The points are cutting across the normal distribution diagonal line which shows the distribution far from resembling a normal distribution.

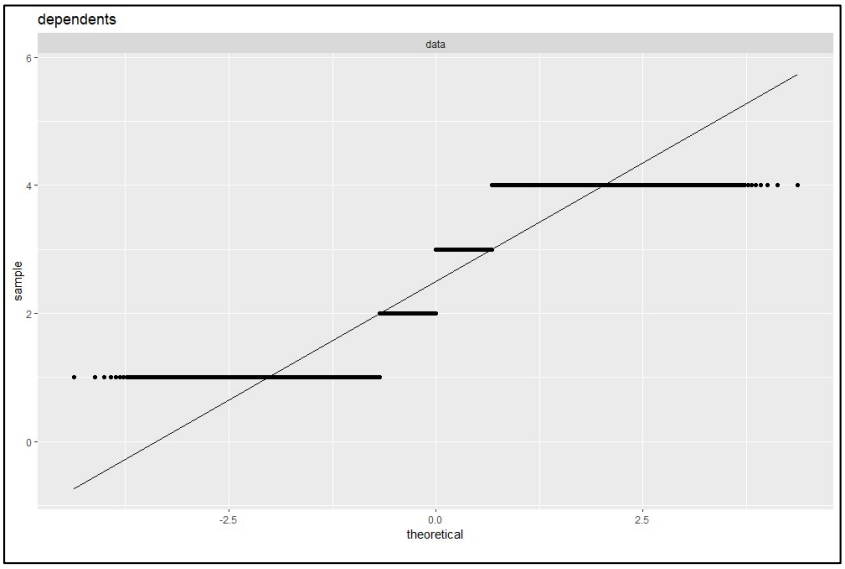




vii. Dependents

The number of dependents the customers have is evenly distributed across the different numbers at around 25%. The QQ plot shows that very few values are along the normal distribution line. The distribution closely resembles a uniform distribution than a normal distribution. The minimum number of dependents customers have is 1 and the maximum is 4 which gives a range of 3. The average number of dependents can be rounded up to 3 which is the same as the median number. The standard deviation is 1.1 and the coefficient of variation is 0.446 which shows low dispersion since its below 1.





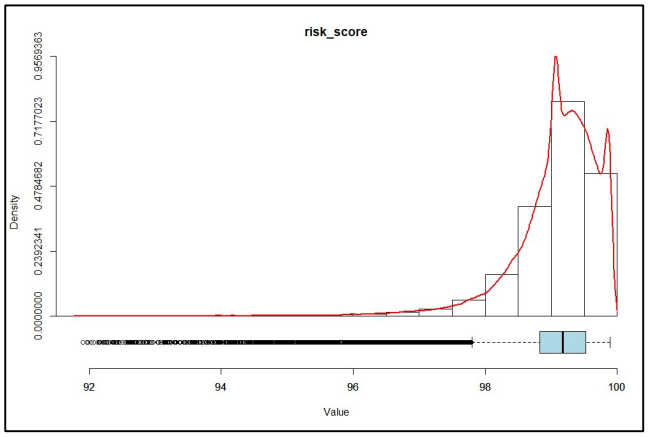
viii. Risk Score

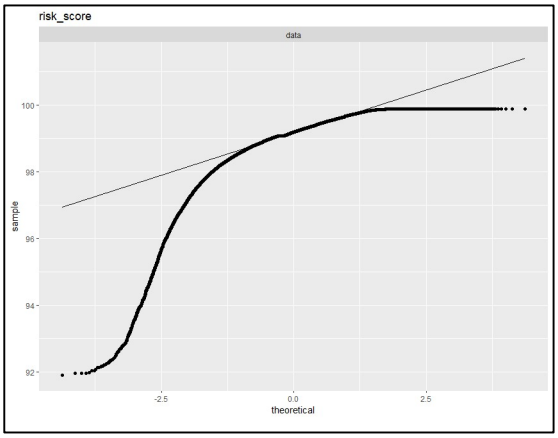
The risk score distribution is significantly left skewed. Since the score is like a credit score, the higher the value the better and from the distribution most customers have a score above 96 from a total of 100. The mean is slightly less than the median at 99.07 and 99.18 respectively which also confirms the skew. The lower tail of the box and whisker is also larger than the upper tail. The large skew indicates the distribution does not resemble a normal distribution and this is confirmed by the QQ plot. The QQ plot shows more values on the lower end are further from the normal distribution line with fewer on the upper end also away from the line.

Usually a lower risk score is associated with a higher probability of default. The lowest customer risk score is 91.9 and the highest is 99.89 which results in a small range of 7.99. the standard deviation is 0.73 which seems low. The coefficient of variation is 0.01 which is low and represents low dispersion which the range and standard deviation also show. This coefficient of variation is the lowest amongst all the numeric variables in the dataset.

There are outliers in the risk score variable and they are all below the lower tail of the box and whisker. These are customers with low risk scores which may indicate their likelihood to

default on premiums. There are 3 784 customers which is about 4.7% of all customers with a low risk scores which are outliers.



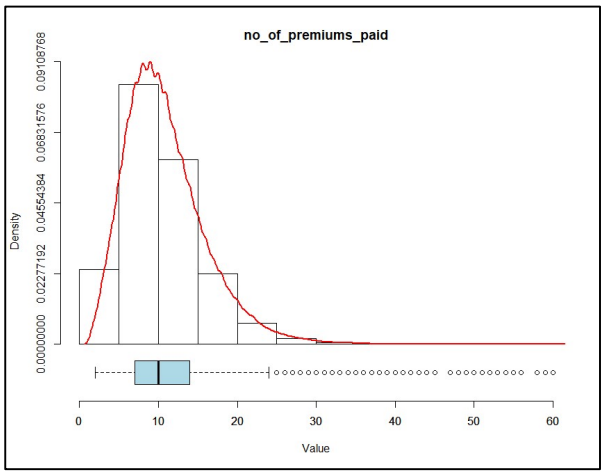


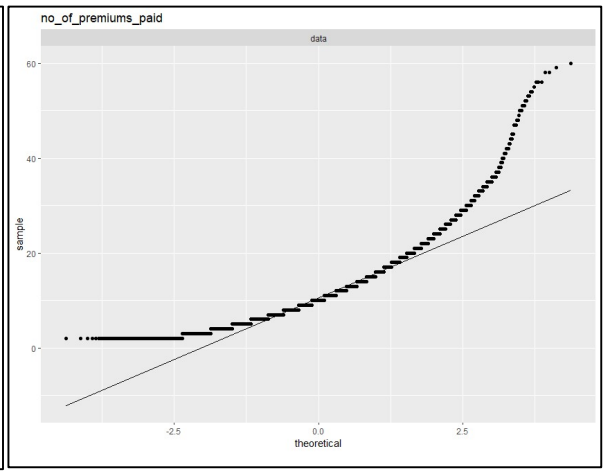
ix. Total premiums paid

The number of premiums paid by customers are right skewed with the average number of premiums at close to 11 and the median at 10. The number of premiums paid on the upper and lower ends of the QQ plot normal line moves away from the line. With the upper end deviating more from the normal line compared to the lower end. This visually shows that the distribution is different from a normal distribution.

The minimum number of premiums paid by customers is 2 and the customer that has paid the most premiums has paid 60 premiums which shows a range of 58. There no customers that have not yet paid premiums. This may mean that customers in the dataset have at least been with the company for 2 months if these are monthly paid premiums. The standard deviation is 5.2 and the coefficient of variation is 0.48. Though the range seems wide, the coefficient of variation shows low dispersion since it is less than 1.

There are outliers present in the upper end of the variable's distribution and these represent about 2% of customers.





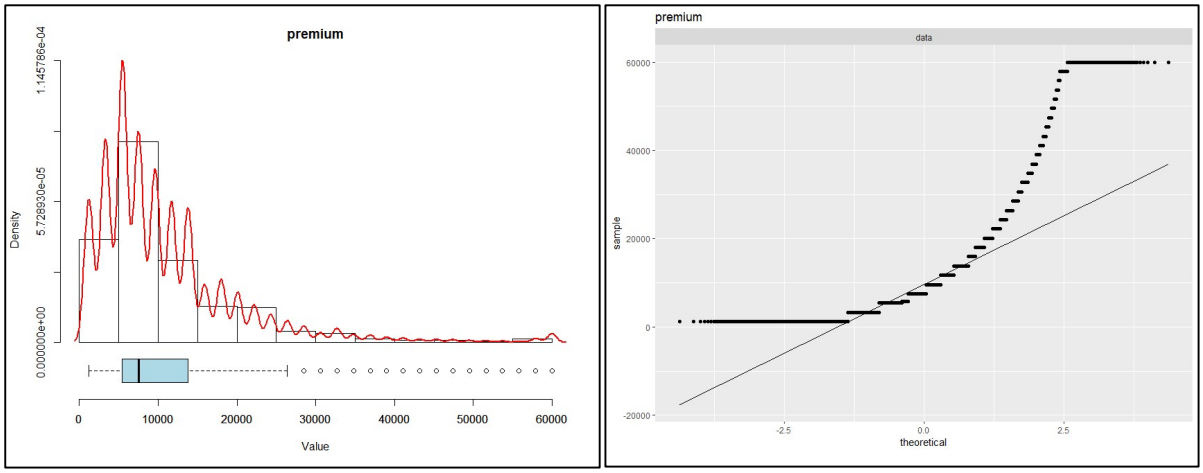
x. Total premiums paid to date

The total premiums paid to date are right skewed and resembling the total number of premiums paid to date. The right whisker of the box and whisker plot is also longer than the left whisker. The density plot is not a smooth line which indicates that total amounts paid to

date are more discrete than continuous. The average amount of premiums paid is 10 924.51 which is larger than the median of 7 500 which also confirms the presence of the right skew.

The customer with the least total amount of premiums paid so far is 1 200 and the customer with the most has paid a total of 60 000. This means the range is 58 800 and the standard deviation is 9 401.68. The dispersion as indicated by the coefficient of variation of 0.86 is low since its less than 1.

There are outliers on the upper tail of the distribution, and these represent 5.7% of customers.



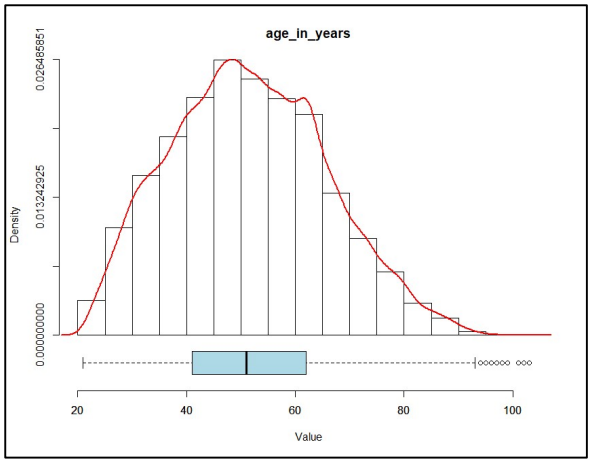
xi. Age

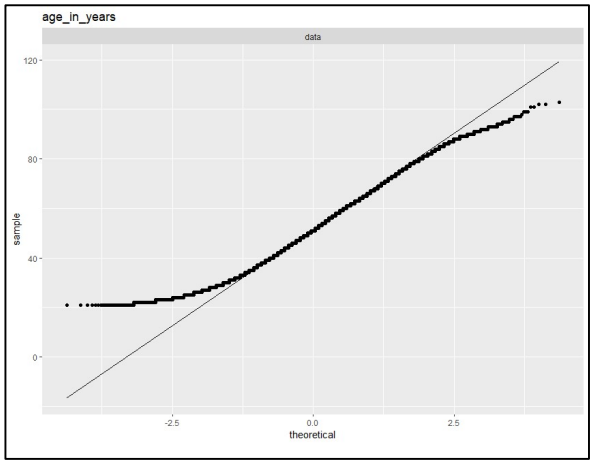
The age of customers in years is closer to a normal distribution than any of the variables in the data set. It shows a slight right skewness with right whisker of the box and whisker plot being longer than the left. The QQ plot shows most of the point lie on the normal distribution line which means that it is not too far from being a normal distribution.

The average customer age is 51.6 but based on their last birthday it is the same as the median at 51. The youngest customer is 21 years old and the oldest is 103 years old. Since the minimum age is 21 years, the distribution does not include those younger than 21 years potentially because of the insurance product requirement. This could explain why the right skew because the distribution excludes all possible ages due to the nature of the product.

The range is 82 years which seems large and the standard deviation is 14 years. The coefficient of variation indicates that the dispersion is low at 0.28 since its less than one.

There are customers that are outliers and are above the right whisker of the box and whisker plot. There 44 of these older clients and the make up only 0.06% of the total customers.

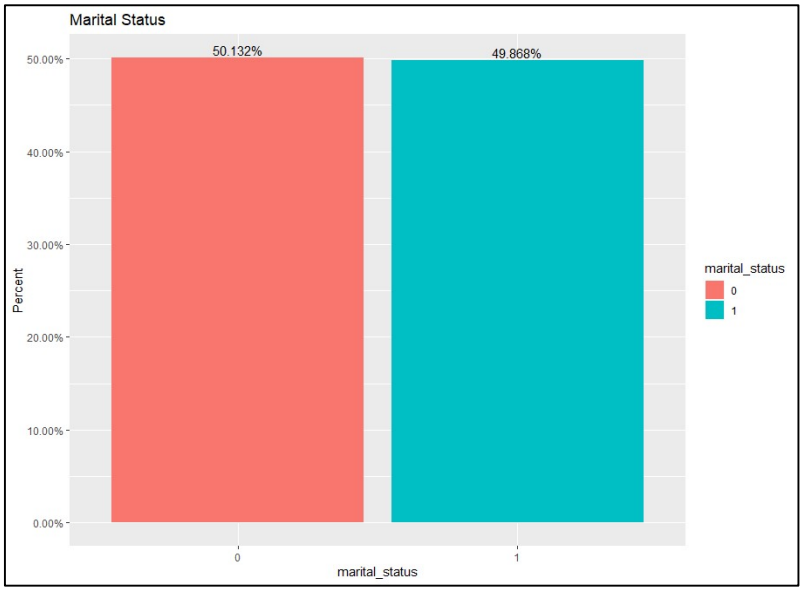




	vehicles_owned	dependents	risk_score	no_of_premiums_paid	premium	age_in_years
nbr.val	79853.00000	79853.00000	79853.00000	79853.0000	79853.000	79853.0000
nbr.null	0.00000	0.00000	0.00000	0.0000	0.000	0.0000
nbr.na	0.00000	0.00000	0.00000	0.0000	0.000	0.0000
min	1.00000	1.00000	91.90000	2.0000	1200.000	21.0000
max	3.00000	4.00000	99.89000	60.0000	60000.000	103.0000
range	2.00000	3.00000	7.99000	58.0000	58800.000	82.0000
sum	159547.00000	199873.00000	7910816.56400	867514.0000	872354700.000	4121006.0000
median	2.00000	3.00000	99.18000	10.0000	7500.000	51.0000
mean	1.99801	2.50301	99.06724	10.8639	10924.508	51.6074
SE.mean	0.00289	0.00395	0.00257	0.0183	33.271	0.0505
CI.mean.0.95	0.00567	0.00774	0.00503	0.0359	65.210	0.0990
var	0.66789	1.24524	0.52692	26.7360	88391521.800	203.6467
std.dev	0.81725	1.11590	0.72589	5.1707	9401.677	14.2705
coef.var	0.40903	0.44582	0.00733	0.4760	0.861	0.2765

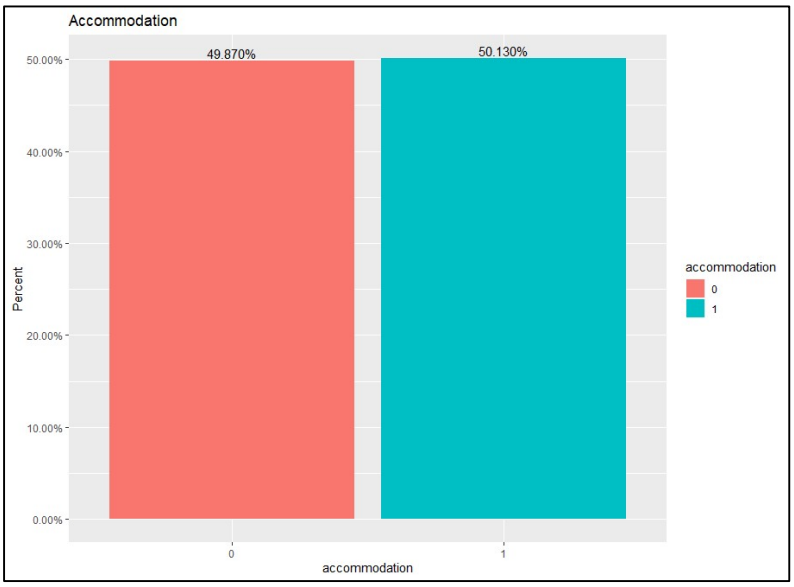
xii. Marital Status

The proportion of customers that are married i.e. 1 is like the proportion of those that are not married i.e. 0. These proportions are 50.1% and 49.9% respectively which means you are equally likely to pick a married customer as much as you are likely to pick an unmarried customer.



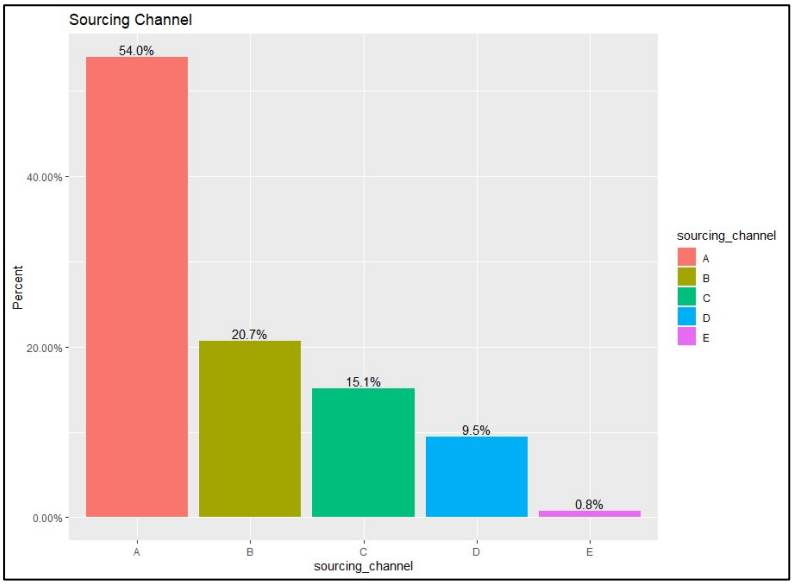
xiii. Accommodation

The proportion of customers who own their residence (1) is like those who rent (0) at 49.9% and 50.1% respectively (below).



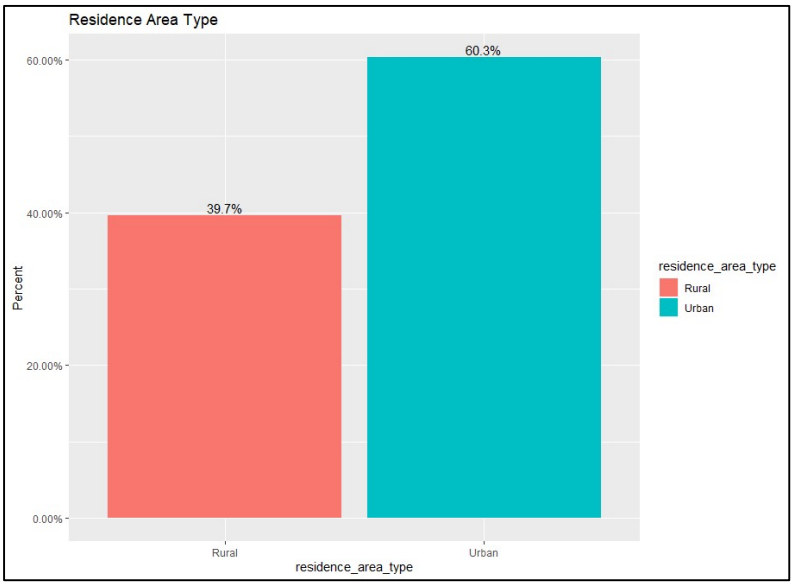
xiv. Sourcing Channel

More than half of the customers come through channel A which represents 54% of customers which is more than double the next popular channel B at 20.7%. Channel C and D are 15.1% and 9.5% respectively and the least popular channel is E at 0.8%.



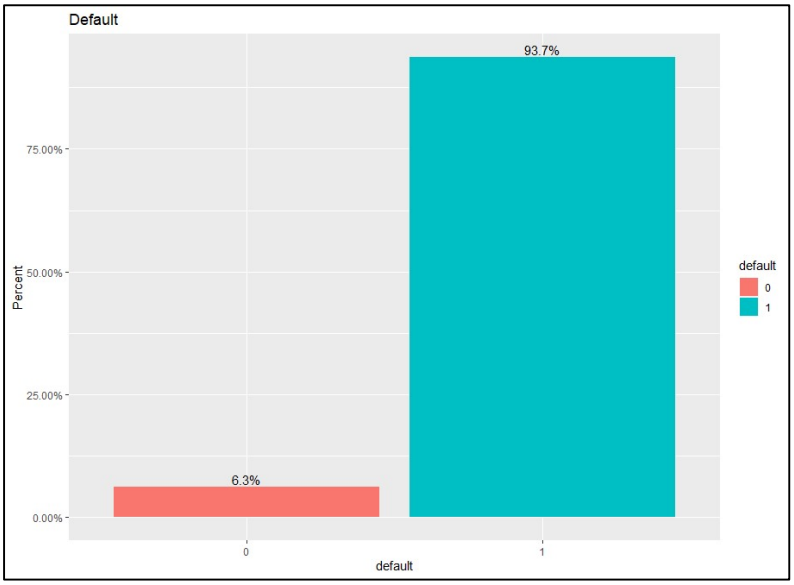
xv. Residence Area Type

Most of the customers are from urban areas and they make up 60.3% of all customers (below). Customers residing in rural areas make up the remaining 39.7% of customers. Typically, customers in urban areas pay higher premiums because of risks like crime which may not be as prevalent in rural areas. Do these higher potentially higher premiums for urban areas be related to higher defaults? This one of the questions that will be looked at later.



xvi. Default

The target variable default shows that 6.3% of the customers have defaulted on their premiums and the remaining 93.7% have not defaulted. This is the focus variable for the study as we aim to reduce the number of customers that default by looking at characteristics of customers that have defaulted and those of customers that have not defaulted. Knowing this information will allow the company to identify the customers that pause a risk and are likely to default. It also allows the company to determine the likelihood that a new customer defaults before they purchase the insurance product.



b. Bi-variate analysis

i. Default vs Marital Status

Knowing the customer's marital status may not be an important factor in determining whether a customer will default on their premium or not. This is because the pattern of the bar plots is similar for married customers and those are unmarried when considering default. Also noting that almost equal proportions of customers are either married or unmarried.

There is small difference with married customers having a default rate of 6.1% compared to 6.4% for unmarried customers. The significance of this slight difference will be determined through statistical methods to see if there is a true difference between married and unmarried customers when it comes to defaulting on premiums. The overall pattern of default is 6.3% which is like the two marital statuses of customers.

default / marital_status	0	1	Totals
0	6.4%	93.6%	100.0%
1	6.1%	93.9%	100.0%
Totals	6.3%	93.7%	100.0%

ii. Default vs Accommodation

The pattern of proportions of customers that own their accommodation compared to those that rent is similar. This also compares to the overall pattern of accommodation when taking default rates into account. For customers that rent, about 6.2% default and this compares to 6.4% of customers that own their accommodation and the overall 6.3% default rate. This

means it is not easy to distinguish whether customer that own their accommodation are likely to default than those that rent.

The proportion of customers that own and those that rent is similar making it more likely that accommodation may not be an important variable in identifying customers that default.

accommodation \ default	0	1	Totals
0	6.2%	93.8%	100.0%
1	6.4%	93.6%	100.0%
Totals	6.3%	93.7%	100.0%

iii. Default vs Sourcing Channel

There are differences in default rates based on the channel that the customer is sourced. Channel A at 5.4% default has the lowest default rate which is lower than the 6.3% for all customers. Customers sourced through channel D default the most at 8.4% with customer sourced through channel C and E with higher than average defaults at 7.5% and 7.6% respectively. Customers sourced through channel B have a default of 6.5% which is close to the overall default rate.

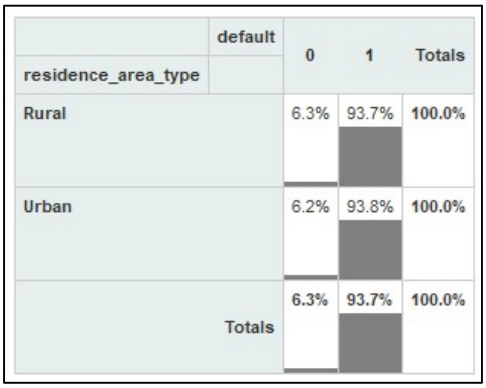
These differences in defaults across the channels may indicate knowing the channel the customer was source could likely assist in determining how likely they are to default. This makes knowing the sourcing channel are likely important factor in the study.

sourcing_channel \ default	0	1	Totals
A	5.4%	94.6%	100.0%
B	6.5%	93.5%	100.0%
C	7.5%	92.5%	100.0%
D	8.4%	91.6%	100.0%
E	7.6%	92.4%	100.0%
Totals	6.3%	93.7%	100.0%

iv. Default vs Residence Area Type

Knowing whether a customer resides in an urban area or rural area may not assist in determining whether they will default or not. The proportion of customers living in urban areas that have defaulted is 6.2% and those in rural areas is 6.3%. There is little difference between the two that it is unlikely one can easily distinguish whether a customer will default.

More customers stay in urban areas i.e. 60.3% which may mean the risk the impact of the defaults may not be potentially the same for urban and rural customers. Knowing the potential overall risk may be more important than just knowing the residence area type.



default / residence_area_type	0	1	Totals
Rural	6.3%	93.7%	100.0%
Urban	6.2%	93.8%	100.0%
Totals	6.3%	93.7%	100.0%

v. Default vs Percentage of premiums paid by cash

There is a difference between customers that default based on the proportion of premiums they pay by cash. The box and whisker plots show that customers that default i.e. 0 have longer left tails than right tails which contrasts with those that have not defaulted with longer right tails (charts below). The median value of the percentage of income paid cash is higher for those that default at close to 70% of premiums paid by cash. For customers that have not defaulted, it's just below 20%. The density plots show the opposite skews for customers that default and those that did not default to confirm the observations above. This means knowing the percentage of the premium paid by cash is likely important in determining whether a customer will default or not.

vi. Default vs number times premium was paid late

Customers that have not defaulted show that they are likely not to make a late payment in any of the three timeframes. The box and whisker plots for 3 – 6 months, 6 – 12 months and over 12 months for these customers are all at zero even though they have outliers. Customers in the 3 – 6 and 6 – 12 month buckets have similar box and whisker plots for customers that default. They are also similar for the same buckets for those that do not default. These similarities between these two buckets may mean that knowing the number of times a premium was paid late for a period of less than a year may not be helpful in determine whether a customer has defaults. The over 12 months shows a difference from the other 2 which may mean knowing whether a customer has paid their premium for over a year may help determine whether the customer defaults.

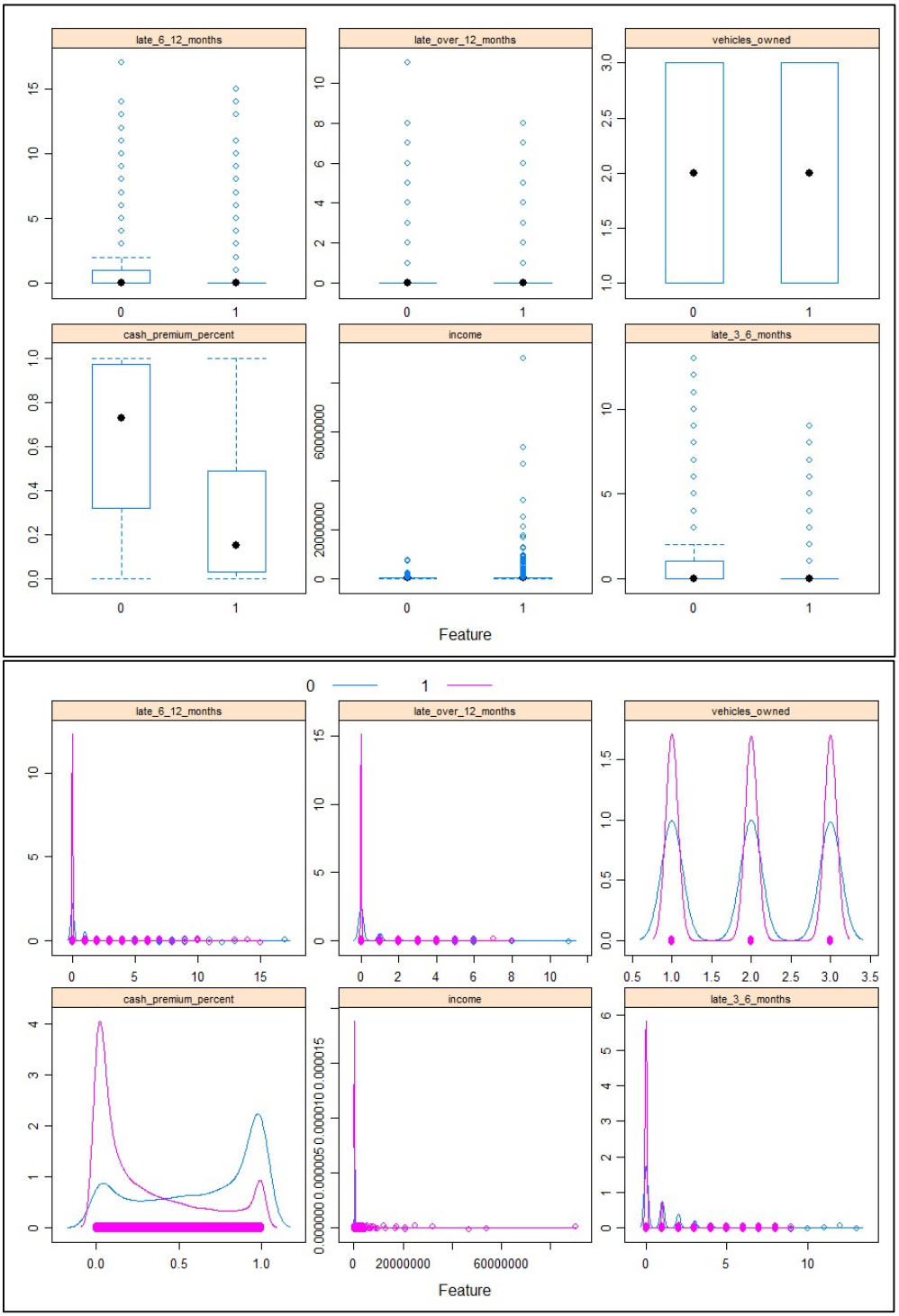
vii. Defaults vs Vehicles owned

The box plots are similar for customer that default vs those that do not default. The kurtosis is higher for customers that did not default compared to those that defaulted. Knowing the number of vehicles alone may not entirely determine whether a customer will default or not

viii. Default vs Income

Customers that have not defaulted show that they have a wider distribution of income. These customers have significantly large outliers on the right tail compared to those that defaulted.

This means potentially knowing a customer's income may help in determining whether a customer will default or not.



ix. Default vs Premium

The total premiums paid by customers are different for customers that have defaulted and those that have not defaulted. The box and whisker plot for those that have not defaulted is wider than that of those that have defaulted (charts below). This difference may be important in determining whether a customer will default.

x. Default and Age

Older customers have not defaulted as much as younger customers. The median for customers that have not defaulted is higher than those that have defaulted. The density plot also confirms this with those that defaulted having a plot more to the left of those that have not defaulted indicating that age is an important factor in determining whether a customer defaults.

xi. Default vs Dependents

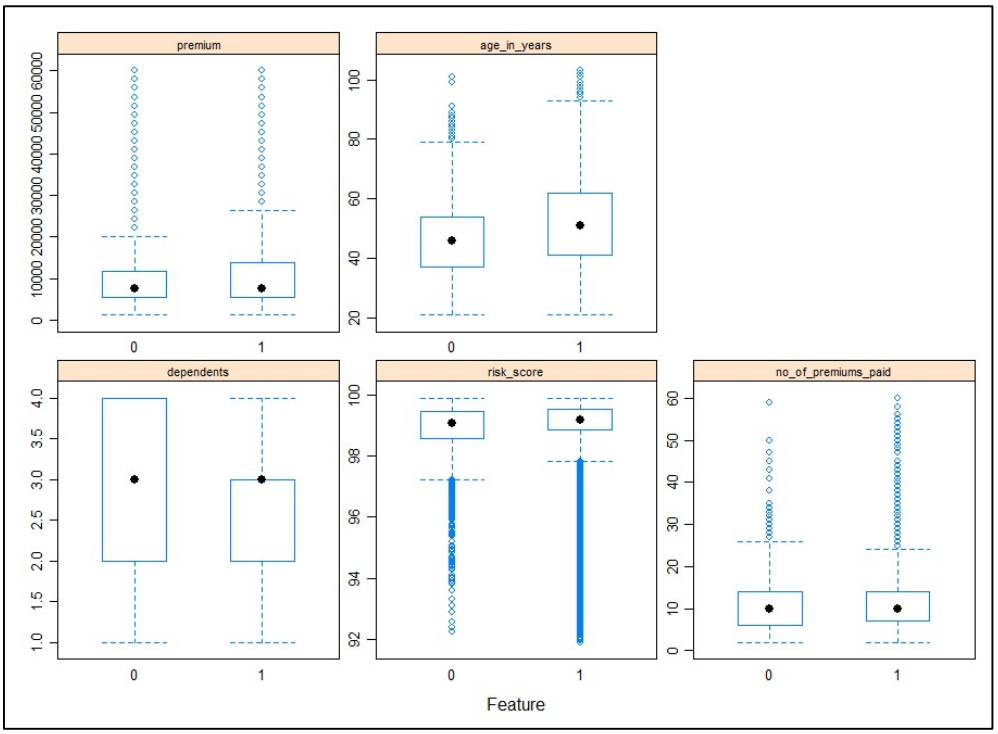
There is a difference in the box and whisker plots for customers that have defaulted and those that have not defaulted. The box plot for those that defaulted does not have a right whisker which shows a larger proportion of customers with more depended defaulted. Dependents seem important in determining whether a customer will default.

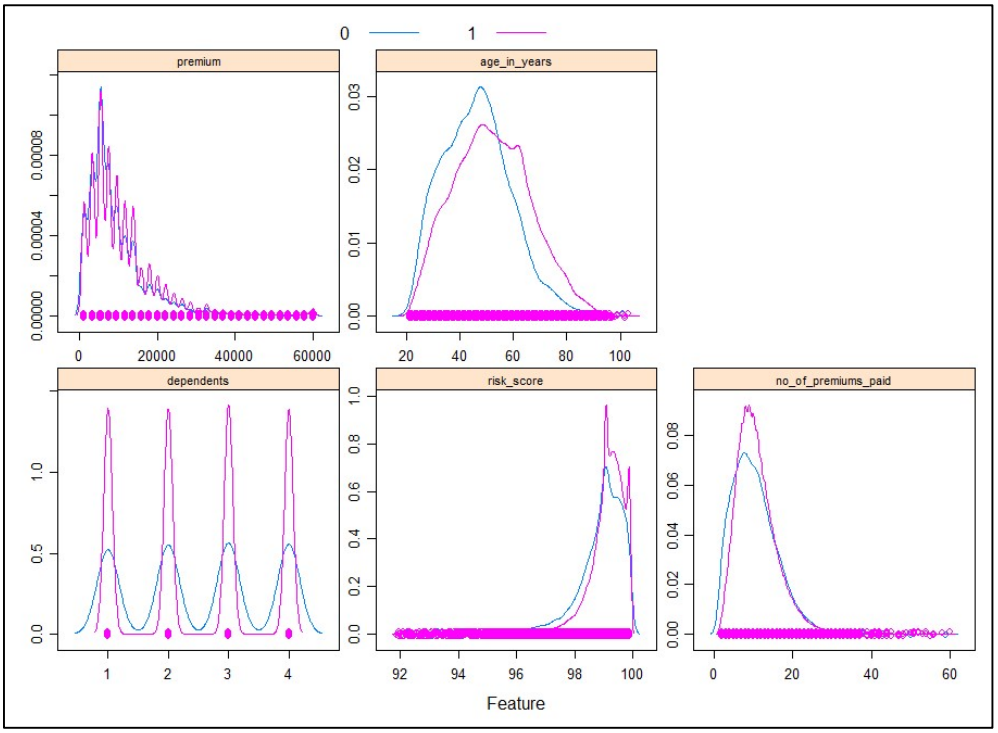
xii. Default vs Risk Score

Both distributions for customers that defaulted and those that did not default are left skewed but the tail of those that defaulted thicker which means knowing the score is important in identifying customers that are likely to default. The box and whisker plot for those that did not default is narrower than those that defaulted. This could also indicate there is a score when customers are more likely to default which make knowing the score an important in knowing which customers are likely to default.

xiii. Default vs Number of premiums paid

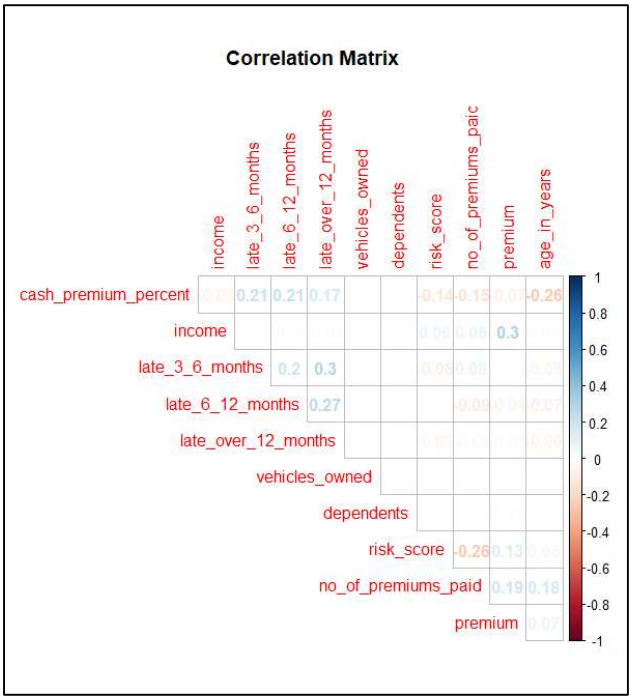
The density plot for customers that defaulted is more to the left of those that have not defaulted. Customer with lower number of premiums paid are likely to default than customers that have paid more premiums. This makes the variable important from this initial analysis in determining whether a customer defaults.





xiv. Correlation Matrix

The correlation matrix below does not show any significantly large positive or negative correlations between the variables. The largest positive correlation is 0.3 which is between late payments between 3 – 6 months and late payments over 12 months. The 0.3 positive correlation is also present between total premiums paid and the number of premiums. The lowest negative premium is between the risk score and number of premiums paid. It is also the same between percentage of premiums paid in cash and the customer's age. These low correlations are a first indication that we may not have an issue with multicollinearity though further tests will be required.



5) Variable Relationships

a. Relationship among variables, important variables

From our initial analysis, 6.3% of the customers had defaulted and determining their characteristics is important in building a model that will predict the likelihood of default. The tables below are a combination of all factor variables and the different variable combinations show the default rates for these combinations. From these tables we can deduce the following:

i. Relationship between the sourcing channel and other variables

The initial analysis showed that there were different relationships between the sourcing channel and the default rates. The tables below show the default rates based on all the combinations of factor variables. Sourcing channel A shows the lowest default rates which are all below the overall rate of 6.3%. for example, a customer that came through channel A whose residence area type is Rural, living in rented accommodation and unmarried is 5.1%.

Customers that came through channel C and D have default rates higher than the overall rate of 6.3%. The combination of variables that came through channel C shows a level of consistency between all combinations of default rates between 6.8% and 8.1%. There is also a level of consistency for all the combinations for channel A with rates ranging between 5.1% and 6.1%. This consistency for channel A and C if viewed in isolation within each channel, one may not be able to distinguish whether a combination of variables can help predict a default rate because of their similarities. However, if the combinations are viewed across channels, there is a clear difference in default rates across channels which makes knowing the channel and combination of all factor variables important in knowing default rates.

ii. The best and worst default rates

Based on the combination of factor variables, customers that are least likely to default are customers sourced through channel E, from an urban area using rental accommodation and married. Potential reasons why this may be the case is incomes in urban areas are usually higher than those in rural areas. There is a possibility that the married customers have two sources of income. Rentals may also have lower monthly payments compared to bond payments and there are usually lower maintenance costs paid for rental properties. All these factors mean that customers have potentially more disposable income which lowers the likelihood of defaulting on other payments like the insurance premium.

Interestingly, the worst default rate of 10.5% also comes from channel E. in fact, the three highest default rates of 10.0%, 10.2% and 10.5% come from channel E. The highest defaulters are customers from channel E, from urban areas, with owned property who are unmarried. Possible explanations are that urban areas are usually associated with a higher cost of living compared to rural areas. This means a single source of income from an unmarried customer will be put under strain with a high cost of living. Potential high bond costs compared to rentals and maintenance costs for owned property put further strain on income. All these factors reduce disposable income and may result in higher likelihood of customers not making other payments like the insurance premium.

We have determined that the channel is an important variable. This situation above where the best and worst default rate comes from the same channel shows that if we hold the channel constant, we still end up with a different outcome hinting the importance of other variables.

iii. Variable interaction and their importance

Earlier we saw comparing marital status to default and comparing accommodation to default in isolation did not provide much information in determining the likelihood of default. There were similar number of customers who did not default compared to those who defaulted. Viewing these variables in isolation did not provide much information. In combination with other variables, the default rates were different. This means the interaction of variables i.e. sourcing channel, residential area type, accommodation and status is important in determining the default rate. We will need to keep these variables until we are able to isolate the degree to which they cause this variation when combined.

	Default	
Variable	**0**	**1**
A		
Rural		
Accommodation 0		
Marital Status 0	5.1%	94.9%
Marital Status 1	5.5%	94.5%
Accommodation 1		
Marital Status 0	6.1%	93.9%
Marital Status 1	5.6%	94.4%
Urban		
Accommodation 0		
Marital Status 0	5.5%	94.5%
Marital Status 1	5.2%	94.8%
Accommodation 1		
Marital Status 0	5.3%	94.7%
Marital Status 1	5.5%	94.5%

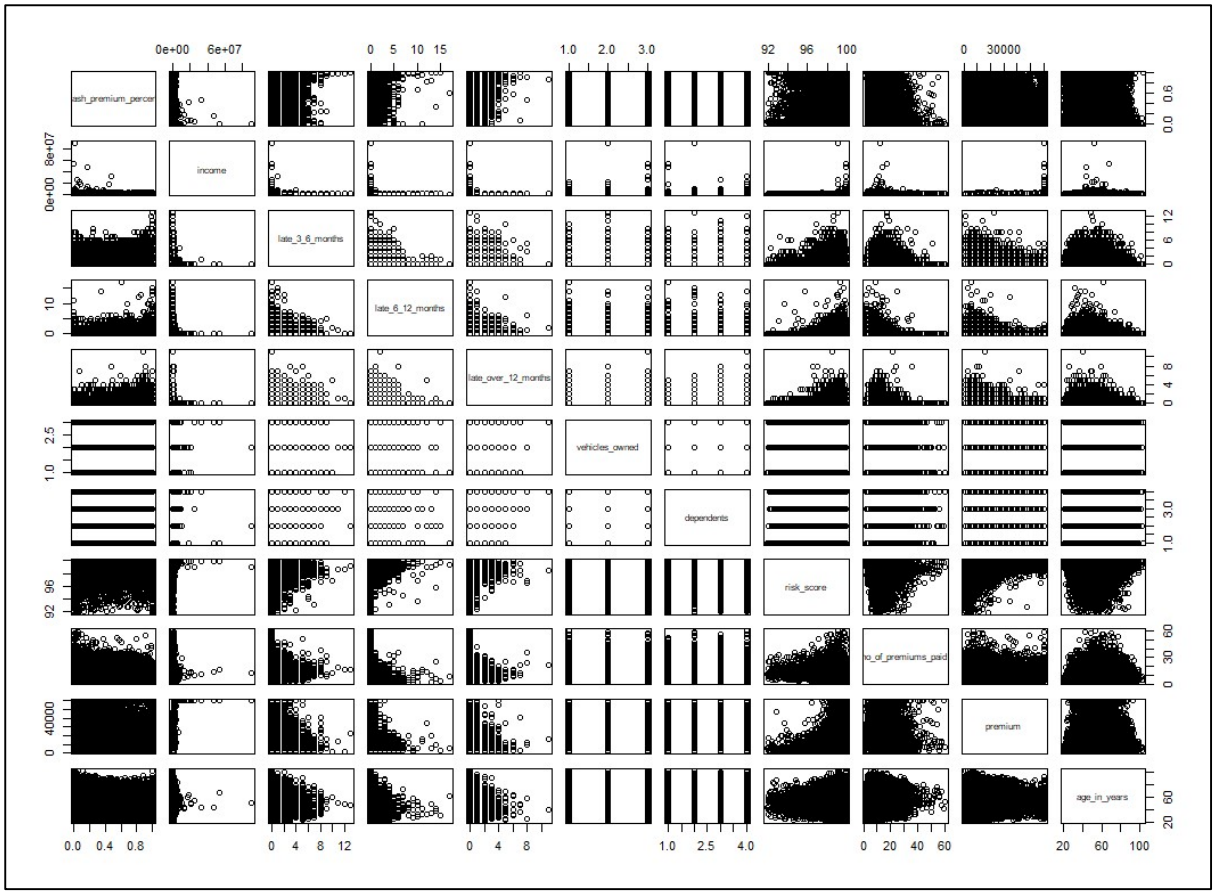
	Default	
Variable	**0**	**1**
B		
Rural		
Accommodation 0		
Marital Status 0	6.7%	93.3%
Marital Status 1	6.3%	93.7%
Accommodation 1		
Marital Status 0	7.1%	92.9%
Marital Status 1	6.1%	93.9%
Urban		
Accommodation 0		
Marital Status 0	6.1%	93.9%
Marital Status 1	6.0%	94.0%
Accommodation 1		
Marital Status 0	7.2%	92.8%
Marital Status 1	6.3%	93.7%

	Default	
Variable	**0**	**1**
C		
Rural		
Accommodation 0		
Marital Status 0	7.6%	92.4%
Marital Status 1	7.4%	92.6%
Accommodation 1		
Marital Status 0	7.4%	92.6%
Marital Status 1	6.8%	93.2%
Urban		
Accommodation 0		
Marital Status 0	7.7%	92.3%
Marital Status 1	8.1%	91.9%
Accommodation 1		
Marital Status 0	7.1%	92.9%
Marital Status 1	7.7%	92.3%

	Default	
Variable	**0**	**1**
D		
Rural		
Accommodation 0		
Marital Status 0	9.6%	90.4%
Marital Status 1	6.5%	93.5%
Accommodation 1		
Marital Status 0	9.0%	91.0%
Marital Status 1	8.3%	91.7%
Urban		
Accommodation 0		
Marital Status 0	9.0%	91.0%
Marital Status 1	7.7%	92.3%
Accommodation 1		
Marital Status 0	9.5%	90.5%
Marital Status 1	7.3%	92.7%

	Default	
Variable	**0**	**1**
E		
Rural		
Accommodation 0		
Marital Status 0	6.3%	93.8%
Marital Status 1	10.0%	90.0%
Accommodation 1		
Marital Status 0	6.3%	93.8%
Marital Status 1	5.4%	94.6%
Urban		
Accommodation 0		
Marital Status 0	10.2%	89.8%
Marital Status 1	4.1%	95.9%
Accommodation 1		
Marital Status 0	10.5%	89.5%
Marital Status 1	7.7%	92.3%

For numeric variables, the chart below shows a high-level view of how they relate to each other. This allows us to have a visual inspection of the type of relationships that exist between them for example, there seems to be a positive relationship between the risk score and premiums as well as positive relationship between the risk score and number of premiums paid. We can also see the relationships between the late payment variables i.e. 3 – 6 months, 6 – 12 months and over 12 months have similar patterns across other variables like the risk score, premiums, number of premiums, income and percentage of premiums paid as cash. For the chart below, we do not have additional detail of how these variables interact with our target variable in default which will be examined in the next section.



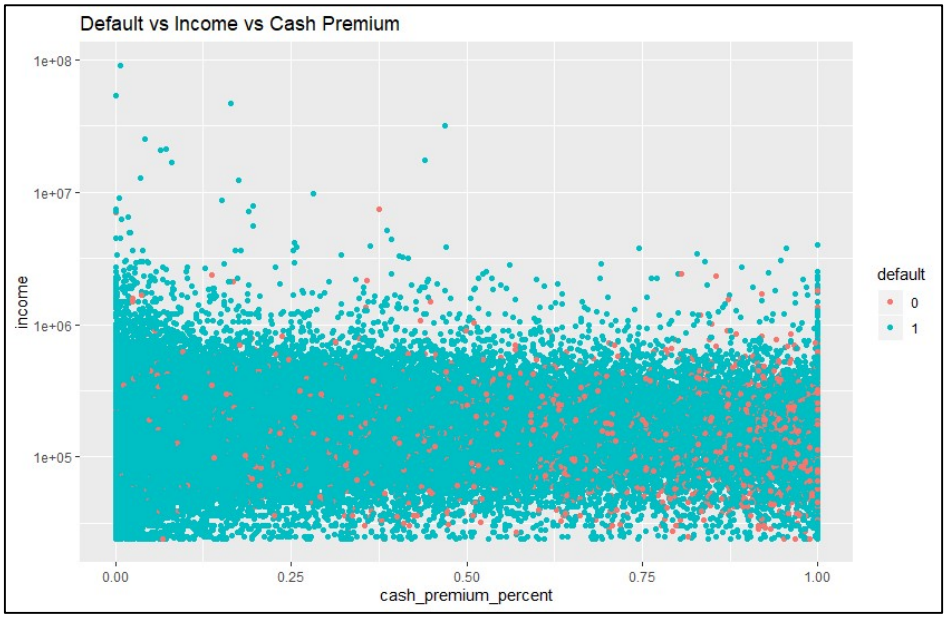
b. Insightful Visualizations

i. Defaults, Income and Cash Premiums

The chart below shows there is a weak negative relationship between income and the percentage of premiums paid in cash. Customers with higher incomes seems to pay a lower percentage of their premiums in cash. The concentration of customers that have defaulted is on the bottom right of the chart where incomes are lower, and percentages of income paid in cash are higher.

We can see from the chart that most customers earn an income below 1 600 000 which is 99.7% of all customers meaning those earning above 1 600 000 make up 0.3%. the default rate for these customers is 6.3% which is consistent with the overall default rate. The default rate for customers with an income above 1 600 000 is 4.7%. This confirms a negative relationship between the default rate and income. The default rate for customers paying less than or equal to 50% of their premium in cash is 3.0% which is less than half of the overall default rate of 6.3%. the default rate for customers paying more than 50% of their premiums in cash is more than double the overall default rate at 15.1%.

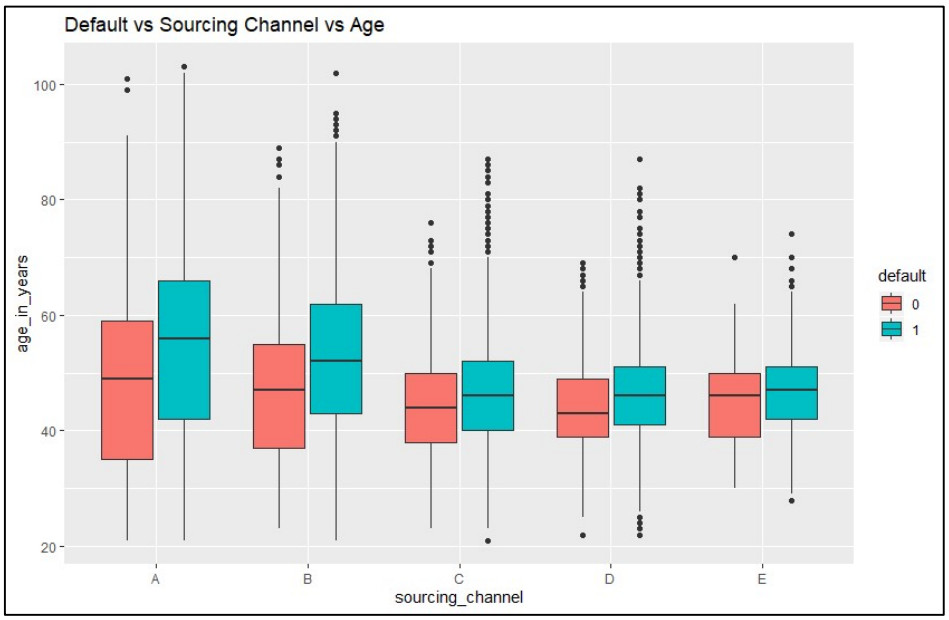
The conclusion from the below chart is that customers with low incomes and paying most of their premiums in cash have a greater likelihood of defaulting.



ii. Defaults, Sourcing Channel and Age

The sourcing channel in relation to age remains an important factor in determining the likelihood of default. This is confirmed by the differences in the distributions of age across sourcing channels as well as default. Channel A has the widest spread of age for both customers who defaulted and those that did not. The highest median age for customers that did not default is in channel A at 56 years. The median age for the different channels from A to E is decreasing for customers that did not default. This is the case as well for customers that defaulted as well. The lowest median for customers that defaulted is in channel D at 43 years.

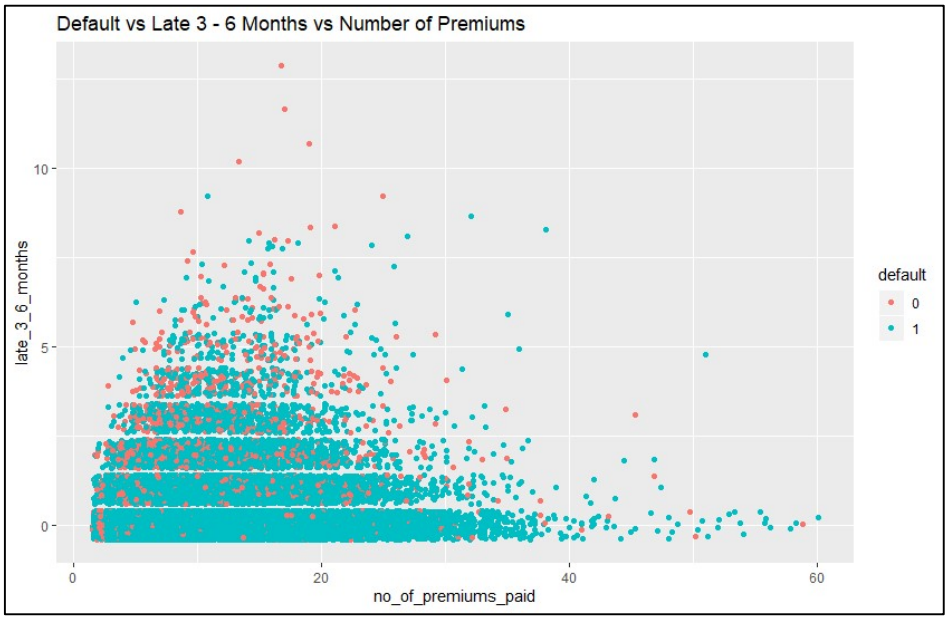
From this, we can conclude that knowing the channel and the customer's age will help in determining the likelihood of default.



iii. Default, Late Payments 3 – 6 Months and Premiums

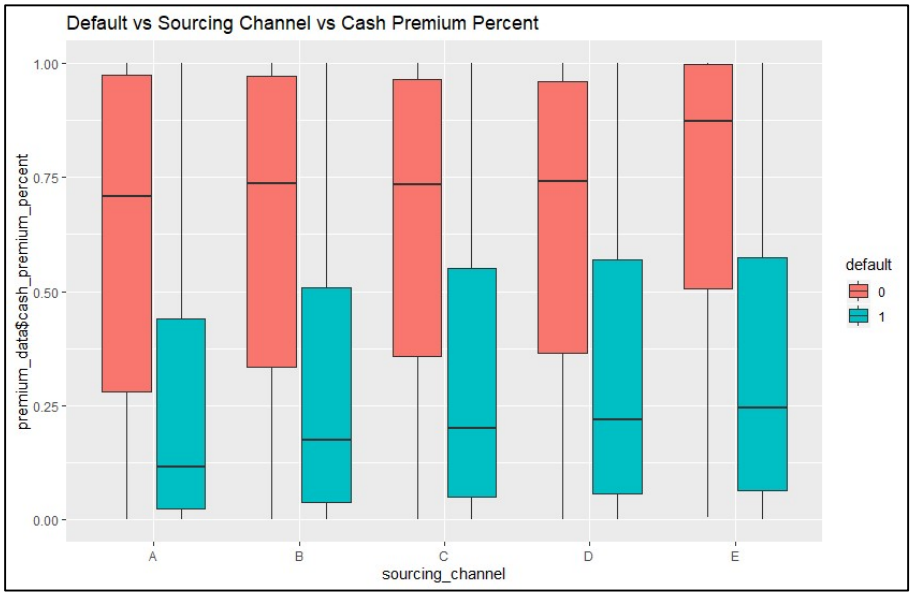
The concentration of defaults when late payments from 3 – 6 months and premiums are on the lower left corner. This is where customers have between 1 and 5 late payments and number of premiums are below 20. The default rate for the above parameters is 17.8% which is close to three times the overall default rate.

From this we can deduce that customers with a low number of premiums paid with at least a late payment are more likely to default. The similarity of pattern between all the late payment variables means a similar generalisation can be made for the 6 – 12 months and over 12 months late payments.



iv. Default, Sourcing Channel and Cash Premium Percent

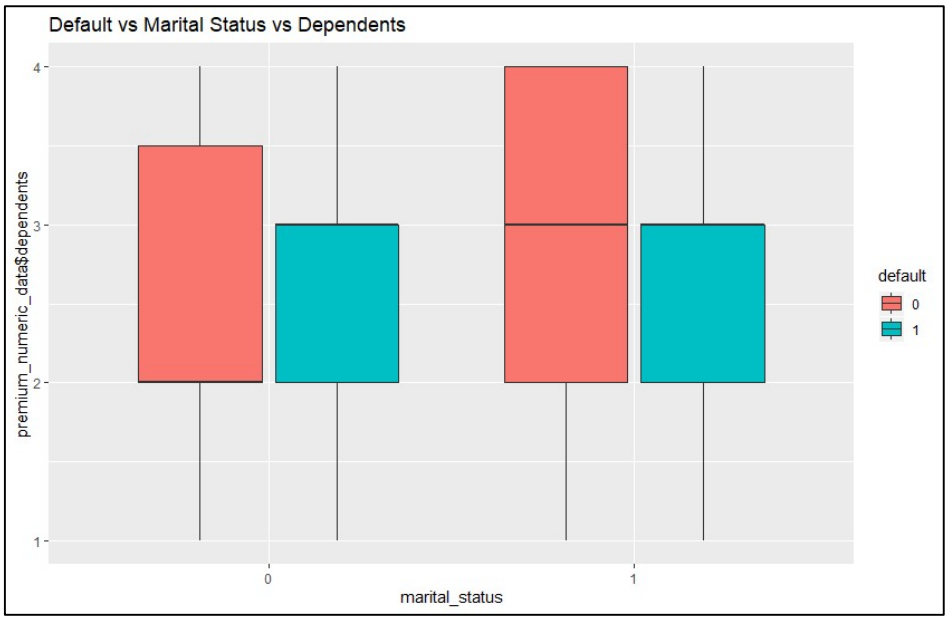
Here we observe that regardless of the channel, customers who pay most of their premium in cash are likely to default.



6) Data pre-processing

a. Removal of unwanted variables

All variables have been retained as they appear to contain information relevant to determining whether customers default or not. Variables like number of dependents and vehicles owned when viewed against the default seem not to provide much information that would assist in determining the likelihood of default. With this alone, they become candidates for removal. However, if they are viewed in combination with other variables the seem to provide additional information. The chart below shows the median number of dependents for unmarried customers are 2 and 3 for defaulters and non-defaulters respectively. Which is different for married customers at 3 for defaulters and non-defaulters.



b. Missing Value treatment

There are no missing values in the dataset.

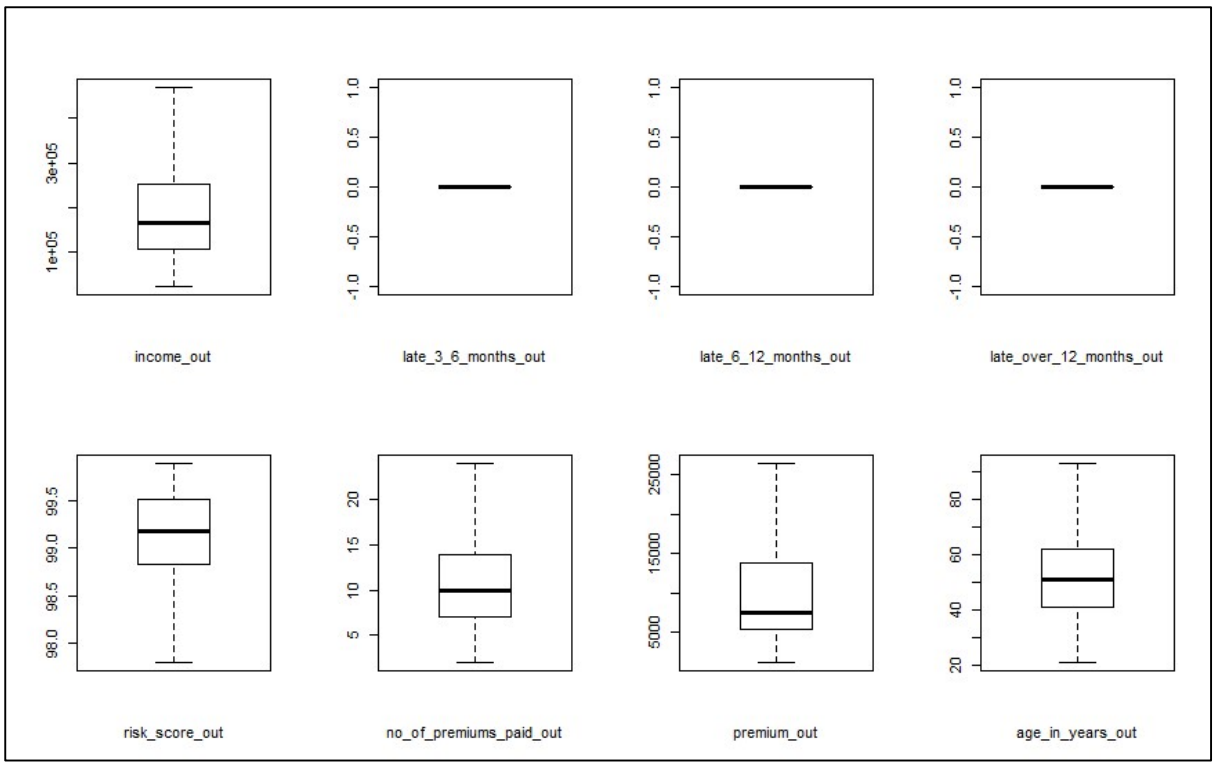
c. Outlier treatment

There are 8 numeric variables that contain outliers and they are listed below. The treatment of outliers in the variables uses the capping method. The method replaces outliers that are 1.5 times above or below the inter quartile range (IQR) limits. The replacements for these outliers are the 5th percentile for values below the lower IQR limit and the 95th percentile for values above the upper limit of the IQR. The late_3_6_months variable uses the 83rd percentile to handle all the outliers and the premium variable uses the 94th percentile to clear all outliers.

- income
- late_3_6_months
- late_6_12_months
- late_over_12_months
- risk_score

- no_of_premiums_paid
- premium
- age_in_years

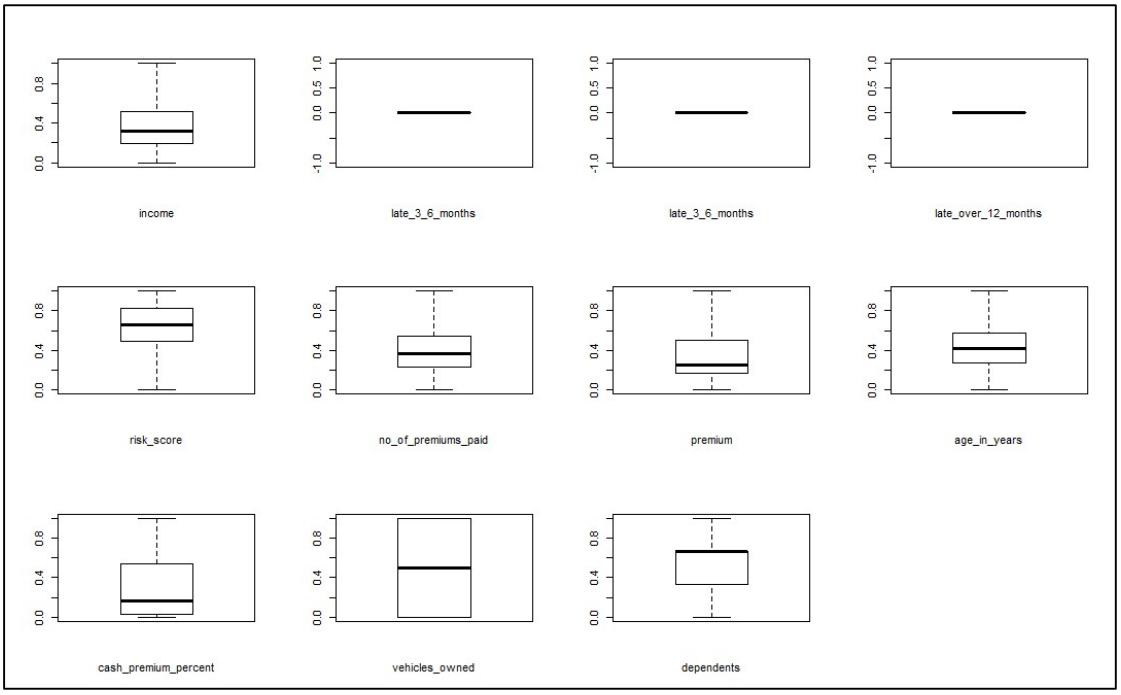
the variables below after treatment show all outliers have been removed. The removal of these outliers will aid in reducing the skewness and allow for more robust models that reduce overfitting.



d. Variable transformation

i. Variable scaling

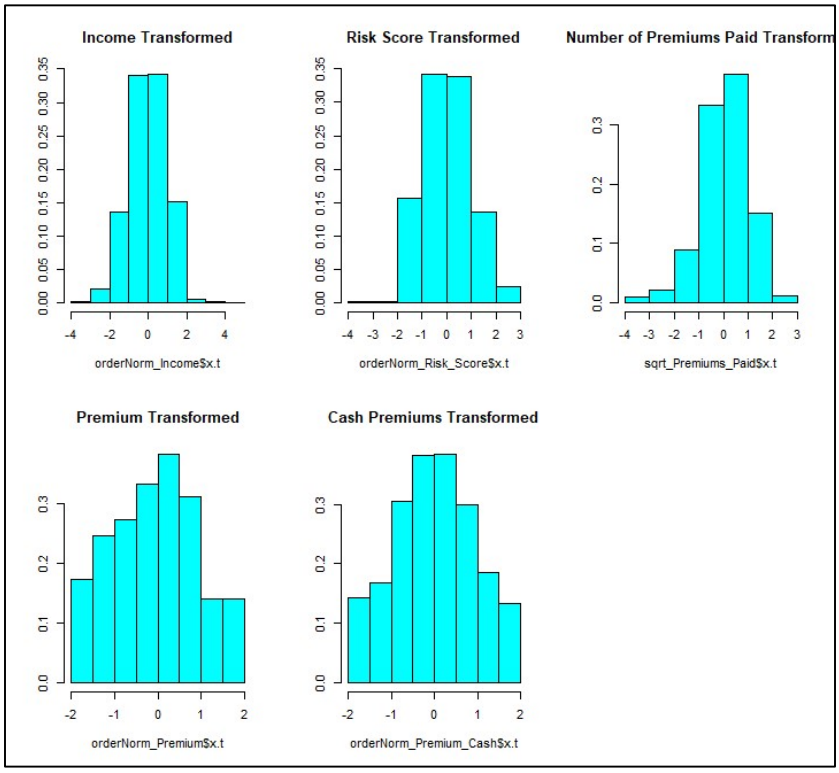
The different units of measure for the numeric variables require scaling because variables of larger magnitudes like income will have a disproportionate impact on determining its effect on defaults for example. Scaling allows variables to give information on equal standing negating the impact of some variables with large magnitudes. From the diagram below, all variable scales have a minimum of 0 and a maximum of 1.



ii. Variable normalisation

There are instances where assumptions of normality are required for example in linear regression and this means variables should follow normal distributions. We have already seen that apart from age, the other variables do not follow a normal distribution.

The bestNormalize package has been used to identify the most appropriate transformation for the variable. This transformation is the applied resulting in distributions that closely resemble normal compared to the initial variable distributions (chart below).



iii. Addition of new variable

A variable reflecting the average premium paid by the customer has been added. This is the total number of premiums paid divided by the number of premiums paid. This allows us to determine the profitability of groups of customers particularly in relation to the likelihood of default. We can see below the minimum average premium paid is 50 and the maximum is 13 200 with an average payment of 1 103.06 per payment. To improve revenue, customers with higher payments relative to the risk of default maybe be desirable.

	stat.desc(ave_premium_per_payment)
nbr.val	79853.000
nbr.null	0.000
nbr.na	0.000
min	50.000
max	13200.000
range	13150.000
sum	88082552.470
median	872.727
mean	1103.059
SE.mean	3.254
CI.mean.0.95	6.377
var	845343.385
std.dev	919.426
coef.var	0.834

7) Analytical approach

A key feature of our problem is that our target variable, default is known which lends itself to a supervised learning approach. With our problem, we want to know whether a customer will default or not which means our target is a categorical variable. This means we can approach the problem as a classification problem. In this case we can use approaches like naïve bayes, logistic regression, k nearest neighbours, decision trees, random forests or support vector machines.

Some aspects of the problem may require the use of unsupervised learning for example dimensionality reduction using principal component analysis.

8) Modelling Process

a. Modelling and Validation

Our problem requires identifying customers that will default on their insurance premiums. This means we have a classification problem where we will use classification algorithms to identify these customers.

iv. One hot encoding

Machine learning algorithms require numeric variables this means we convert all the factor variables like sourcing channel. Since there is usually no intrinsic ordering in the factor variables, binary numbers are used to identify the occurrence of the factor 1 is used. New columns representing the number of factor levels are created for example sourcing channel will be replaced by five new columns. The model used for encoding is created on the training dataset only and then applied to the test dataset. This caters for scenarios where new factor levels are created in new scoring data.

v. Training/Test split

Our dataset has been split into training and test datasets. The split is 75:25 which means 75% of all observations are allocated to the training dataset and the other 25% to the test dataset. The split is random however it still maintains the same proportions for the target variable i.e. default. This means that the default rates remain at 6.3% across the 2 datasets. The training dataset is used to learn the relationship between the customers that default and the rest of the variables. The resulting model is then used to evaluate how good it is in predicting on values on the test dataset.

vi. SMOTE - Handling imbalanced data

Only 6.3% of customers defaulted and this is a low resulting in an unbalanced distribution of the default variable. Application of algorithms particularly those that reply on sampling may lead to bias and inaccuracy. We use the Synthetic Minority Over-sampling Technique (SMOTE) to deal with this imbalance. SMOTE oversamples the infrequent event in this case customers that default. This synthetically creates observations of default and reduces the occurrence of customers that didn't default to balance the dataset. This is only applied to the training dataset and not the test dataset to avoid contaminating or introducing bias to the model.

vii. Cross validation and performance measure

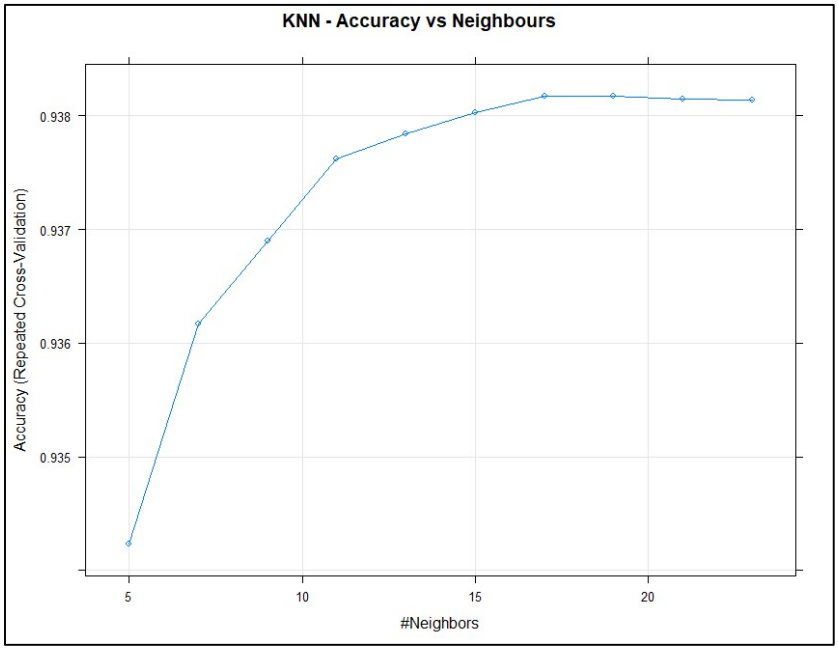
The Repeated k-fold cross validation method is used to assess the performance of the models. It uses subsets of the data to determine the performance of these algorithms. The primary measure of performance is accuracy which is the metric that is used to select the best model given the parameters used for each algorithm.

b. Algorithms and Models

i. K-Nearest Neighbour Classifier (KNN)

The KNN classifier is a supervised classification algorithm which is applicable to our problem of determining defaulting or non-defaulting customers. KNN aims to determine default based on the nearest class neighbours using Euclidean distance between the observations.

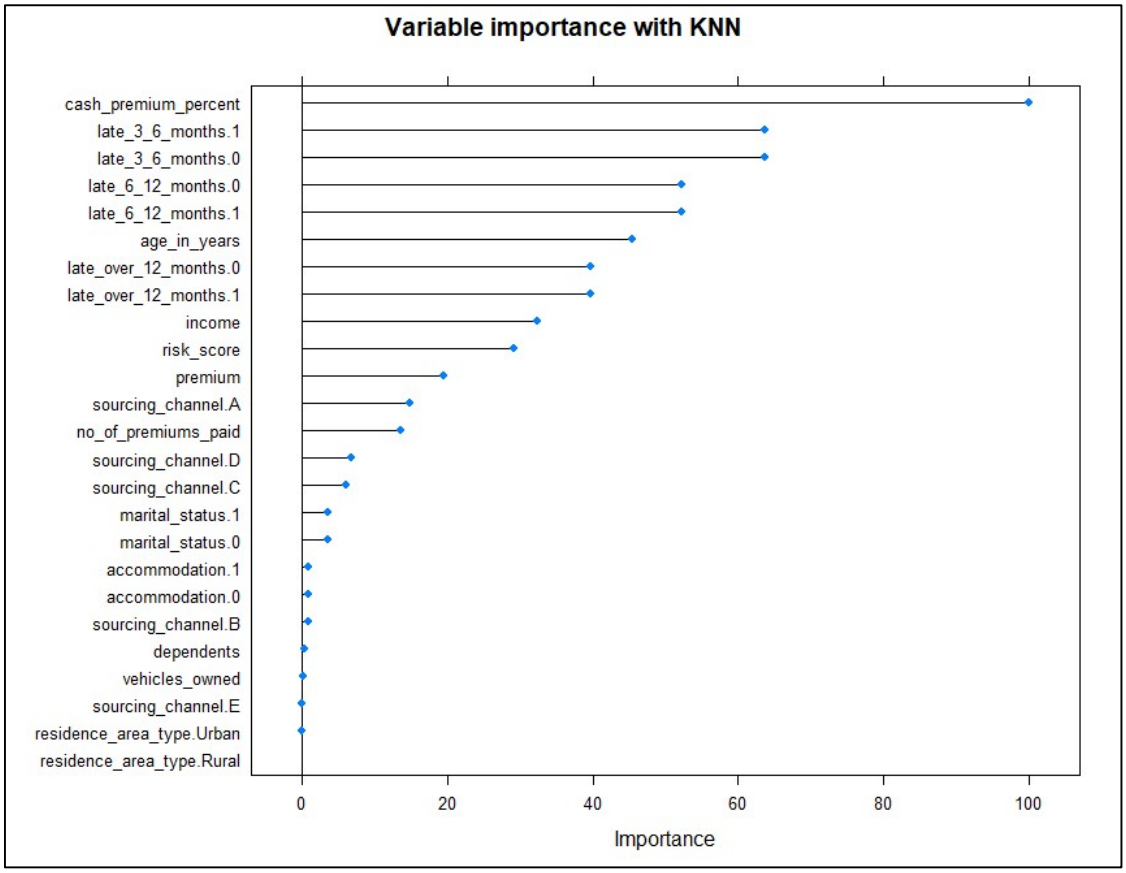
From our model, 17 neighbours have been identified to give the highest accuracy (below). This is what is used to determining the class into which an observation falls i.e. default or non-default.



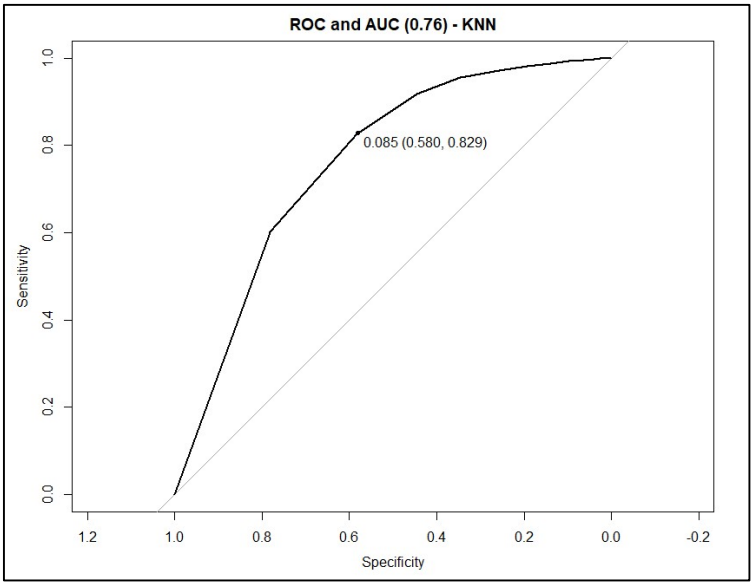
From the confusion matrix, the model shows a high accuracy of 93.8%. Our goal is to identify customers that are predicted to default that defaulted. The focus will be on sensitivity in this case as it shows the proportion of customers who defaulted who got predicted correctly as defaulters. Sensitivity in this case is low at 6.3% which represents the 1170 customers that were predicted not to default but defaulted compared to only 79 who were identified correctly (below).

```
Confusion Matrix and Statistics - KNN

          Reference
Prediction     0     1
         0    79    75
         1  1170 18638

               Accuracy : 0.938
                 95% CI : (0.934, 0.941)
    No Information Rate : 0.937
    P-Value [Acc > NIR] : 0.461

                  Kappa : 0.1

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.06325
            Specificity : 0.99599
         Pos Pred Value : 0.51299
         Neg Pred Value : 0.94093
             Prevalence : 0.06257
         Detection Rate : 0.00396
   Detection Prevalence : 0.00771
      Balanced Accuracy : 0.52962

       'Positive' Class : 0
```

The percentage of premium paid as cash is the most important variable based on KNN (below). Payments paid late in 3 – 6 months and between 6 – 12 months are the next most important variables. Residence area type is the least important variable in the model.



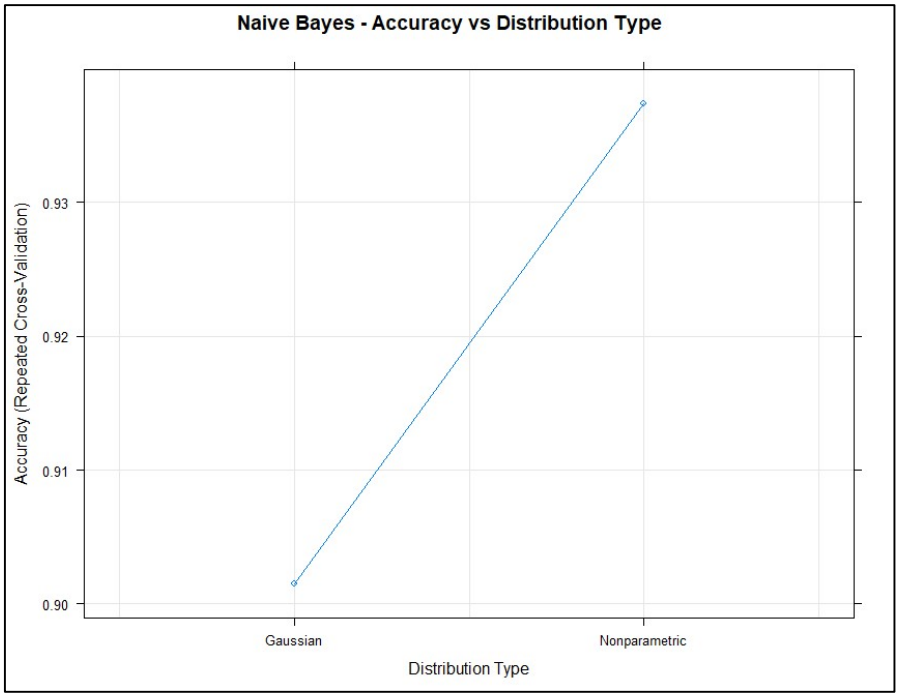
The receiver operating characteristic curve (ROC) shows the performance of the classification model at all classification thresholds. It is based on the sensitivity and specificity and the closer the curve is to the top left corner the better the performance. Based on the ROC curve if accuracy is not taken into consideration, the best performing model would have a sensitivity of 0.829 and a specificity of 0.580. Sensitivity is higher in this case but at the expense of accuracy. The Area Under the Curve (AUC) is used to determine which model is better at predicting the classes best. The higher the AUC the better the model and for KNN it is 0.76 (below).



ii. Naïve Bayes

Naïve Bayes is based on based theorem and has its foundations in conditional probability. In our case, it will be the probability that someone defaults given the other features or variables. Observations are then assigned to the class with the largest probability score.

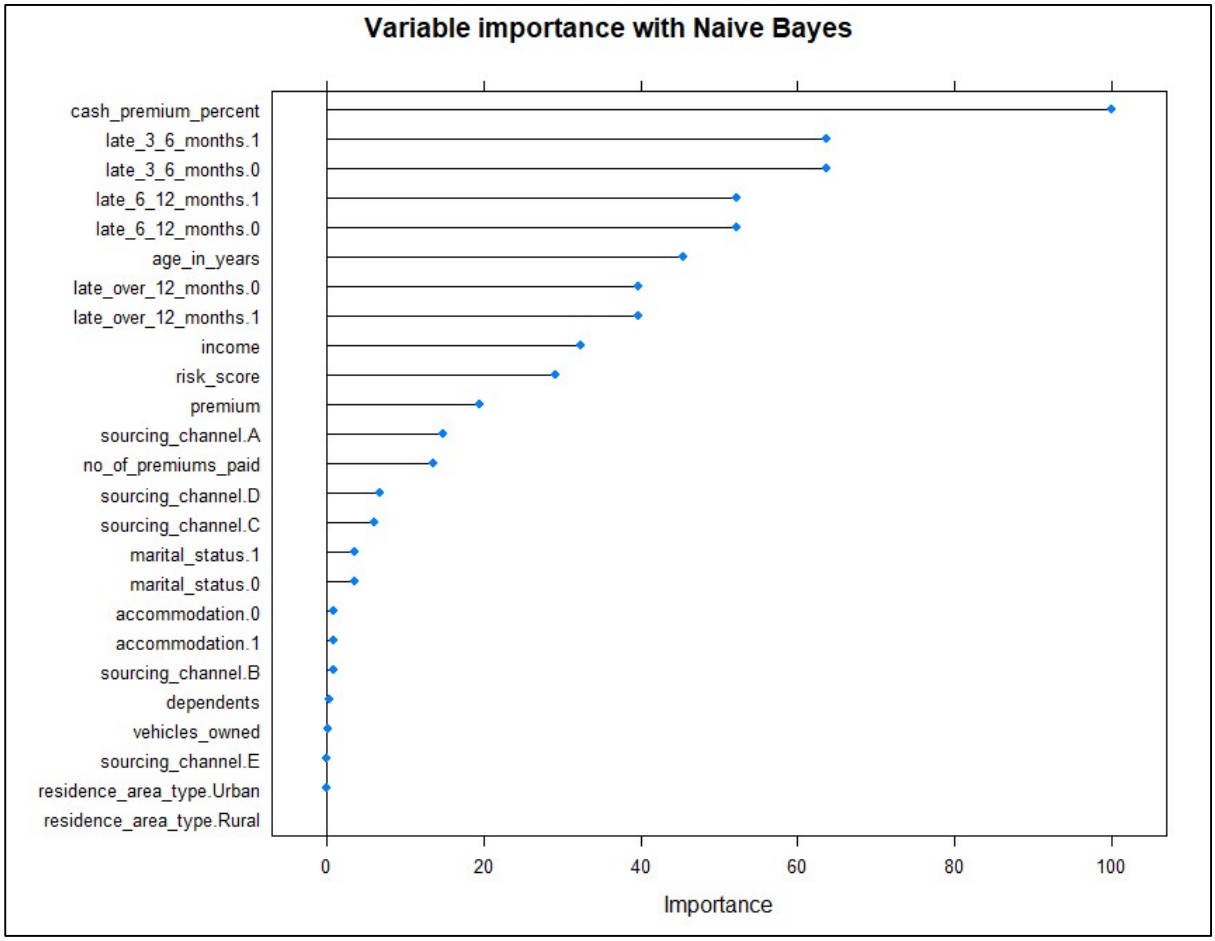
The Nonparametric distribution gives a higher accuracy (93.7%) compared to the Gaussian with an accuracy just above 90.0% (below).



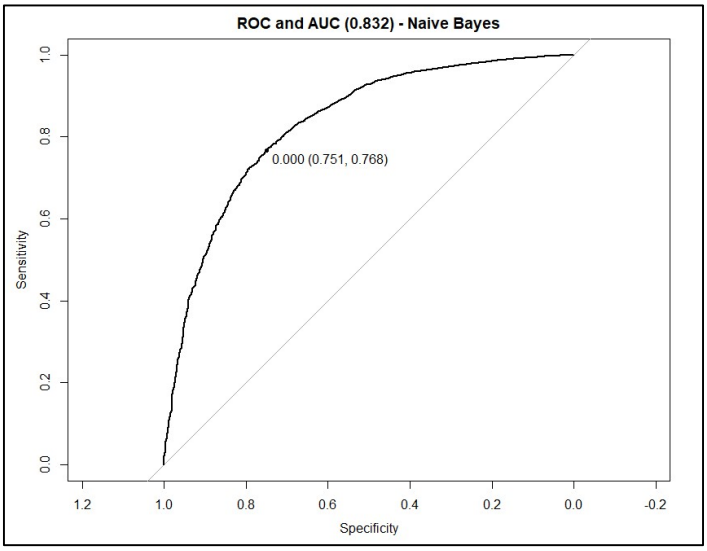
Sensitivity for the Naïve Bayes model is 0 as it fails to predict customers that default (below). It also identifies many customers 1249 that have defaulted as non-defaulters. This causes a problem because although the model has a high accuracy, our metric of focus i.e. specificity performs poorly.

```
Confusion Matrix and Statistics – Naive Bayes

          Reference
Prediction     0     1
         0     0     0
         1  1249 18713

               Accuracy : 0.937
                 95% CI : (0.934, 0.941)
    No Information Rate : 0.937
    P-Value [Acc > NIR] : 0.508

                  Kappa : 0

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.9374
             Prevalence : 0.0626
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 0
```

The most important variable based on Naïve Bayes is percentage of premium paid as cash which is the same as the KNN classifier (below). The next important variables are all related to time and these include all the late payments for the given months as well as age. This is the same as the KNN classifier. Three least important variable is also residence area type.



The ROC curve shows higher specificity compared to KNN at 0.751 but a lower sensitivity at 0.768. The AUC for Naïve Bayes is 0.832 which is higher than KNN.



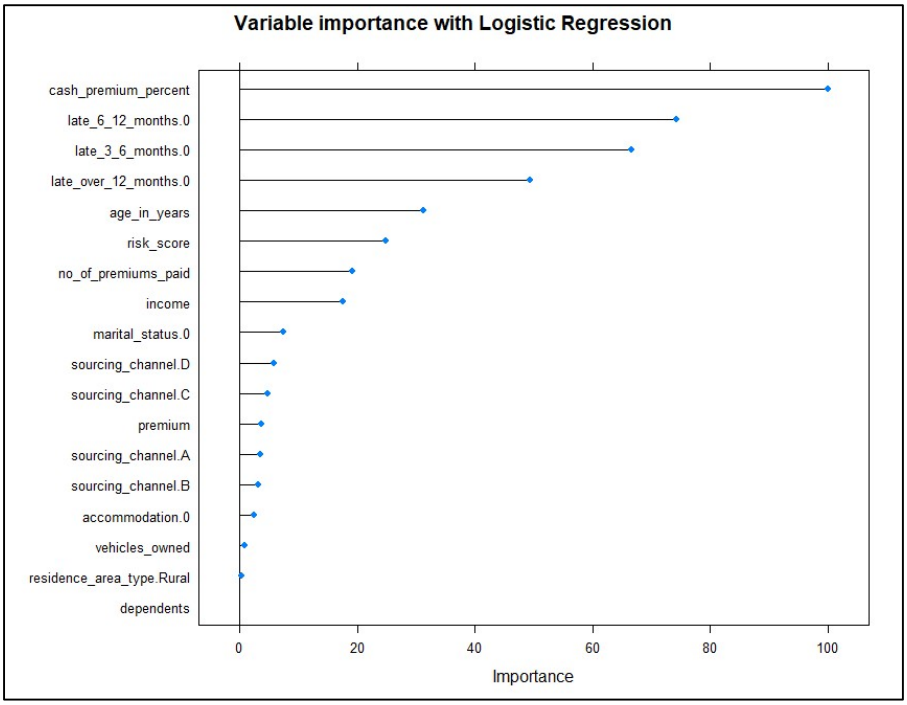
iii. Logistic Regression

Logistic regression is used to explain the relationship between a dependent binary variable (default) and independent variables. One assumption is that there should be no outliers which has been addressed by outlier treatment.

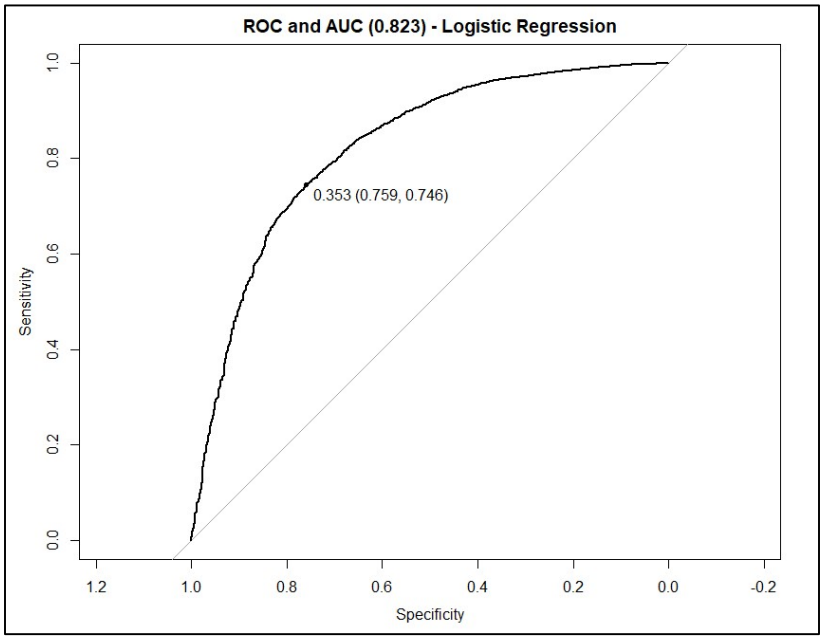
The logistic regression model shows a lower accuracy compared to KNN and Naïve Bayes at 85.4%. It does show a higher sensitivity value of 0.597 and is able to predict more customer that default when they have defaulted i.e. true positives. It does have a lower specificity compared to the other two models where it identifies more customers as defaulters when they are not.

```
Confusion Matrix and Statistics – Logistic Regression

          Reference
Prediction     0     1
         0   745  2403
         1   504 16310

               Accuracy : 0.854
                 95% CI : (0.849, 0.859)
    No Information Rate : 0.937
    P-Value [Acc > NIR] : 1

                  Kappa : 0.274

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5965
            Specificity : 0.8716
         Pos Pred Value : 0.2367
         Neg Pred Value : 0.9700
             Prevalence : 0.0626
         Detection Rate : 0.0373
   Detection Prevalence : 0.1577
      Balanced Accuracy : 0.7340

       'Positive' Class : 0
```

The percentage of premium paid as cash is also the most important variable with the time-based features like months when late payments are made following. The least important variable is number of dependents (below).



The ROC shows the point furthest from the diagonal with specificity of 0.759 and a sensitivity of 0.746. The AUC is at 0.823 which is lower than the Naïve Bayes model.



iv. Bagging

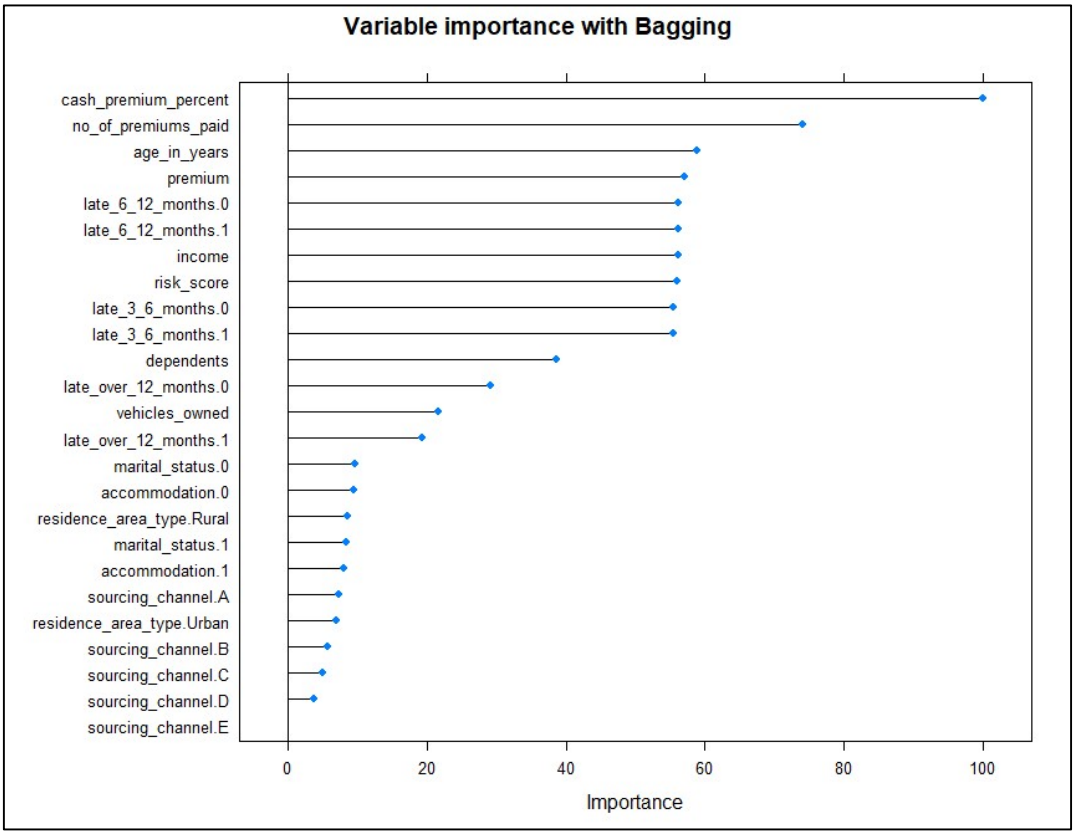
Bagging is an ensemble method combines multiple models to make accurate predictions which are better than any of each of those models.

The accuracy of the bagging model is 89.4% with a sensitivity of 0.428 which is lower than the logistic regression model. It has a higher specificity of 0.927 but it predicts 715 customers as non-defaulters when they are defaulters (below).

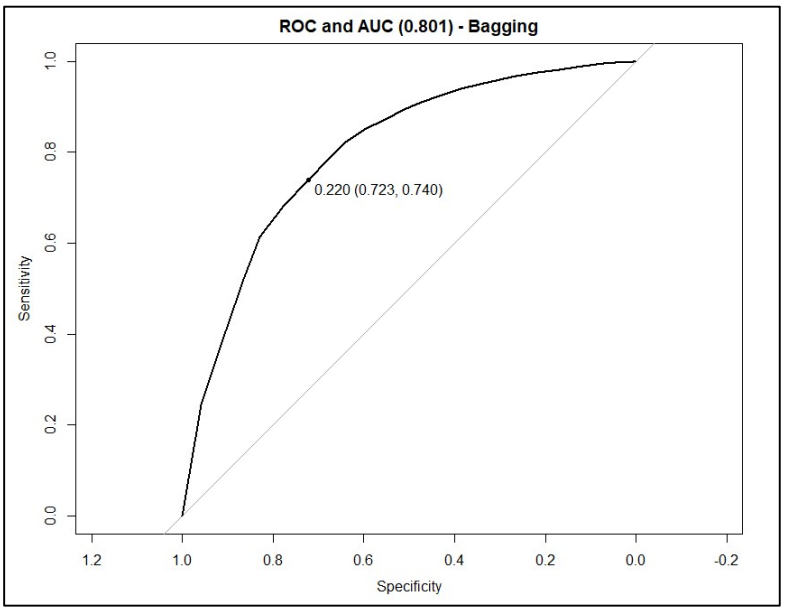
```
Confusion Matrix and Statistics - Bagging

          Reference
Prediction     0     1
         0   534  1393
         1   715 17320

               Accuracy : 0.894
                 95% CI : (0.89, 0.899)
    No Information Rate : 0.937
    P-Value [Acc > NIR] : 1

                  Kappa : 0.282

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.4275
            Specificity : 0.9256
         Pos Pred Value : 0.2771
         Neg Pred Value : 0.9604
             Prevalence : 0.0626
         Detection Rate : 0.0268
   Detection Prevalence : 0.0965
      Balanced Accuracy : 0.6766

       'Positive' Class : 0
```

Percentage of premiums paid as cash is still the most important variable. However, number of premiums paid is now the second most important variable with age being the third most important. From age to late_3_6_months.1, the variables have a similar level of importance

(below). The sourcing channels are now the least important variables with channel E the least important.



The point closest to the top left corner has a specificity of 0.723 and a sensitivity of 0.740 with and AUC of 0.801.



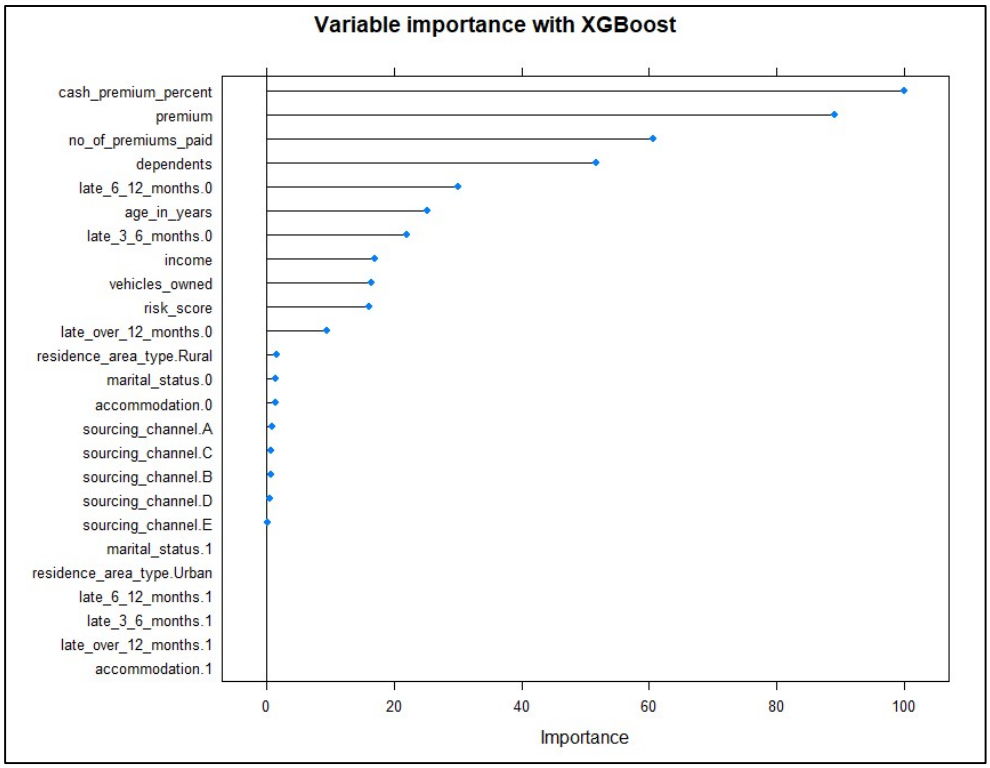
v. Boosting

Boosting refers to a group of algorithms using average weights to make weak learners into stronger learners.

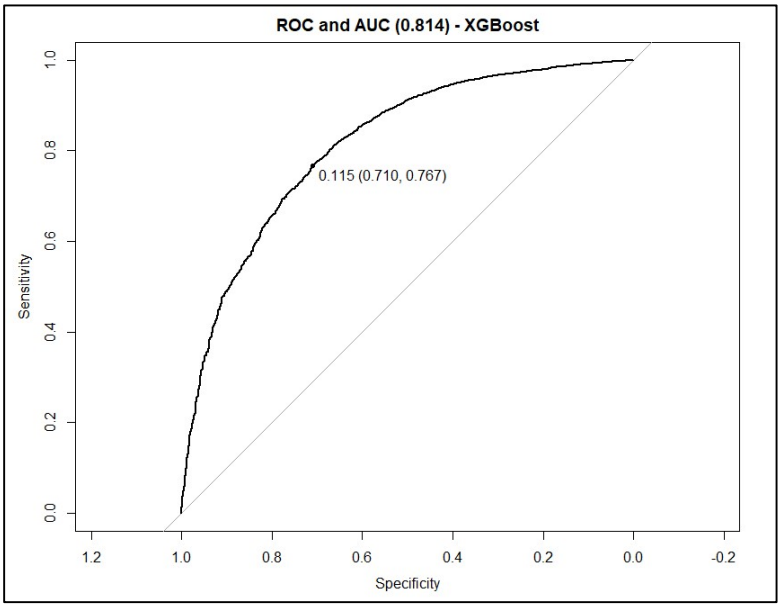
The accuracy of this model is high relative to the other models at 92.4% with a low sensitivity of 0.314. There are 392 customers correctly predicted as defaulters and 857 incorrectly identified as non-defaulters when they are.

```
Confusion Matrix and Statistics - Boosting

          Reference
Prediction     0     1
         0   392   663
         1   857 18050

               Accuracy : 0.924
                 95% CI : (0.92, 0.927)
    No Information Rate : 0.937
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3

 Mcnemar's Test P-Value : 7.41e-07

            Sensitivity : 0.3139
            Specificity : 0.9646
         Pos Pred Value : 0.3716
         Neg Pred Value : 0.9547
             Prevalence : 0.0626
         Detection Rate : 0.0196
   Detection Prevalence : 0.0529
      Balanced Accuracy : 0.6392

       'Positive' Class : 0
```

For boosting, percentage of premiums paid as cash is still the most important variable. Premium and number of premiums paid are second and third respectively. Some of the months when late payments were made are now closer to the bottom of variable importance. The least important variable is owned accommodation.



The area under the curve is .0814 with the point closest to 1 with a specificity of .710 and a sensitivity of 0.767.



9) Model Comparison

The five models below are compared based on accuracy, sensitivity, specificity and AUC (below).

Model	Accuracy	Sensitivity	Specificity	AUC
KNN	0.938	0.063	0.996	0.760
Naïve Bayes	0.937	0.000	1.000	0.832
Logistic Regression	0.854	0.597	0.872	0.823
Bagging	0.894	0.428	0.926	0.801
Boosting	0.924	0.314	0.965	0.814

Based on accuracy alone, KNN is the best performing model with the Logistic Regression model the lowest performing. When focusing on sensitivity, then Logistic Regression is the best performing model, with Naïve Bayes having the lowest sensitivity of zero. The logistic model also has the lowest specificity. Naïve Bayes has the highest AUC with KNN having the lowest KNN.

Focusing on our problem, we need a model that performs well in identifying customers that default. This means the model with the highest sensitivity i.e. Logistic Regression becomes the candidate model for our case. Based on this, the Naïve Bayes model should not be considered at all because it fails to correctly identify defaulting customers with a sensitivity of 0.

Compared to the other models, the Logistic Regression model has the second highest AUC which means it is good at predicting defaults.

10) Interpretation from the best model

The summary of the logistic regression model used is shown below. The results show that there is a positive relationship between default and the number of late payments in the different months. The strongest positive relationship is with late payments made between 6 and 12 months as shown by the coefficient. Percentage of premiums paid as cash has the largest negative relationship with defaults at -0.82046. This means customers with larger proportions of premiums paid as cash are more likely to default. The largest negative and positive relationship variables above are also the two most important variables.

There are variables that are not significant at a level of significance of 5%. Accommodation for example is one such variable as well as residential area type and dependents. These variables are also the some of the least important variables in the model as we saw before. These variables can be dropped to further refine the model as they do not contribute to the prediction of performance of the model.

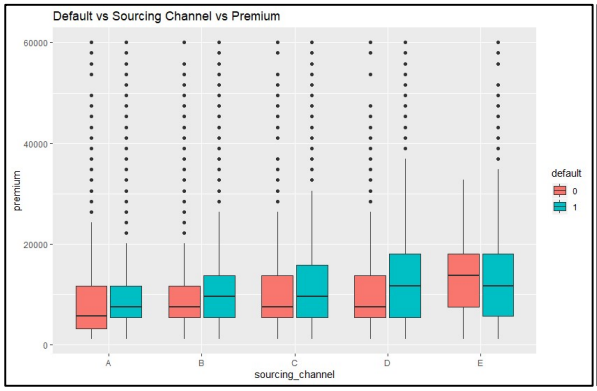
```
Call:
NULL

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.782  -0.785   0.409   0.755   3.026

Coefficients: (7 not defined because of singularities)
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -2.52516    0.15260  -16.55  < 2e-16 ***
late_3_6_months.0          0.96870    0.02613   37.08  < 2e-16 ***
late_3_6_months.1               NA         NA      NA       NA
late_6_12_months.0         1.71455    0.04156   41.26  < 2e-16 ***
late_6_12_months.1              NA         NA      NA       NA
late_over_12_months.0      1.16745    0.04231   27.59  < 2e-16 ***
late_over_12_months.1           NA         NA      NA       NA
marital_status.0          -0.09983    0.02277   -4.39  1.2e-05 ***
marital_status.1                NA         NA      NA       NA
accommodation.0            0.03595    0.02275    1.58  0.11403
accommodation.1                 NA         NA      NA       NA
sourcing_channel.A        -0.31153    0.13856   -2.25  0.02456 *
sourcing_channel.B        -0.27629    0.13966   -1.98  0.04789 *
sourcing_channel.C        -0.40964    0.14007   -2.92  0.00345 **
sourcing_channel.D        -0.48812    0.14143   -3.45  0.00056 ***
sourcing_channel.E              NA         NA      NA       NA
residence_area_type.Rural -0.00953    0.02340   -0.41  0.68378
residence_area_type.Urban       NA         NA      NA       NA
age_in_years               1.23732    0.07065   17.51  < 2e-16 ***
vehicles_owned             0.02134    0.02847    0.75  0.45354
dependents                 0.00768    0.03152    0.24  0.80757
income                     0.17912    0.01800    9.95  < 2e-16 ***
risk_score                 0.19997    0.01426   14.02  < 2e-16 ***
no_of_premiums_paid       -0.15592    0.01444  -10.80  < 2e-16 ***
premium                   -0.04084    0.01763   -2.32  0.02057 *
cash_premium_percent      -0.82046    0.01477  -55.54  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 64945  on 48736  degrees of freedom
Residual deviance: 47151  on 48718  degrees of freedom
AIC: 47189

Number of Fisher Scoring iterations: 5
```

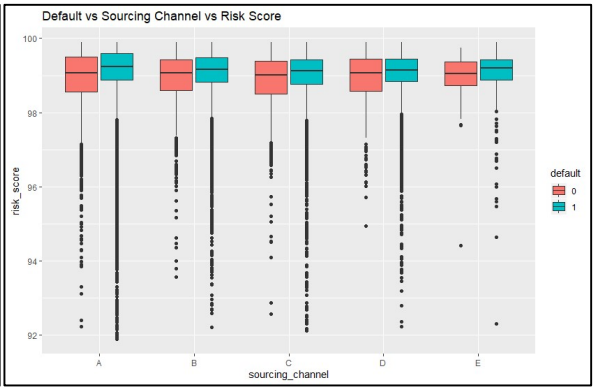
11) Business Insights and Recommendations

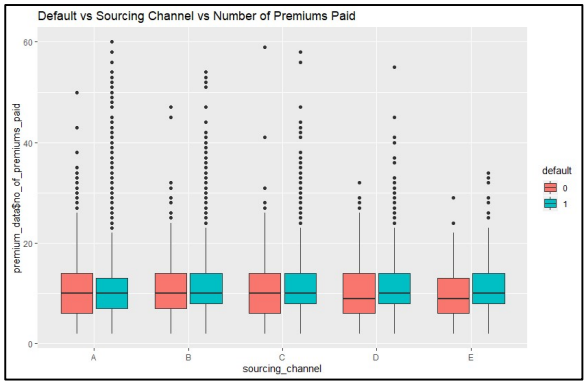
- The most important variable that was identified by all model is percentage of premiums paid as cash. The business can embark on marketing campaigns that encourage customers to move away from using cash for their premiums. This can be done by freezing premium increases for customers that use methods like debit orders which do not involve cash.
- Premiums paid as cash can be used by the business as a qualification criterion for customers that are eligible to get insurance. If a customer fails to meet as certain level of premium not paid as cash, then they are automatically disqualified from purchasing insurance. This will ensure that fewer customers default allowing for more predictable revenue. It can also be used to identify customers that qualify and make it easy for them to sign up for the business' products. For example, a customer can be assessed when submitting an online application and get preapproval making the process customer centric.
- The business can shorten the administration process of acquiring customers by excluding information that is not beneficial in the application process. For example, excluding information of whether a customer stays in rural or urban areas is not beneficial because it's been identified as one of the least important variables.
- The business can modify its distribution channels by focusing more on channels that bring more customers and decreasing or closing sourcing channels that are associated with defaults. For example, the company can do away with sourcing channel E which has consistently been one of the least important variables in all models. This refocus will give the business an additional opportunity to increase revenue.
- The business can create a communications system that notifies customers when payments are due and when they are late. This will raise awareness amongst customers as late payments are an important factor in most of the models. The business can also implement debit orders to deal with late payments as they are automatic and do not rely on customer memory. This is also another way to increase revenue.
- The model selected is not the most accurate model but the model that performs well on identifying customers that are likely to default. The focus is to reduce risk associated with defaulting customers so a model that performs well in this regard is more desirable than one that has the highest overall accuracy. This means the business should be comfortable in reduced accuracy which may mean missed opportunities but have confidence the model is providing the solution to identifying defaulting customers.

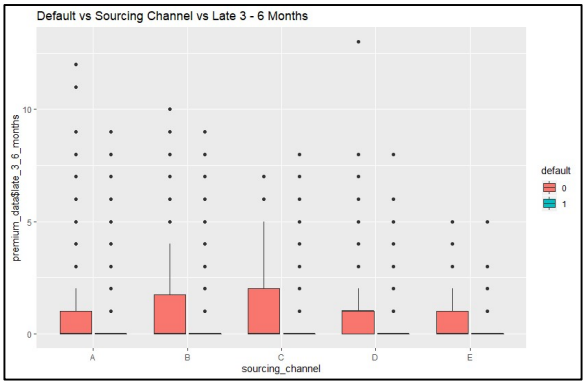
In conclusion, our resulting model focused on being able to identify defaulting customers than overall accuracy. This means that there could more customers that may be identified as defaulters when they are not. This means there are missed opportunities with these potential customers. The model though reduces the risk associated with lower revenues from defaulting customers which makes it relevant for our situation. Other variables like type of employment i.e. part-time or full-time employment can be further incorporated in future studies to see how they have an impact on defaults.

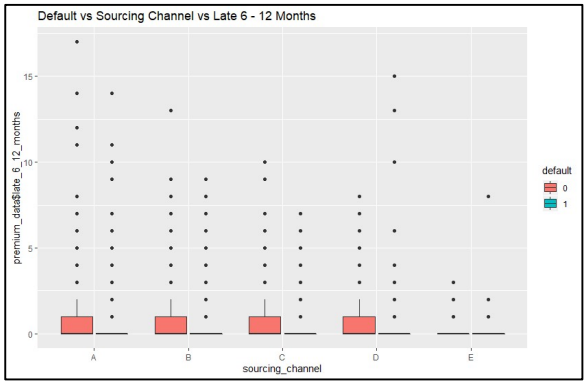
12) Annexures – Additional Graphs

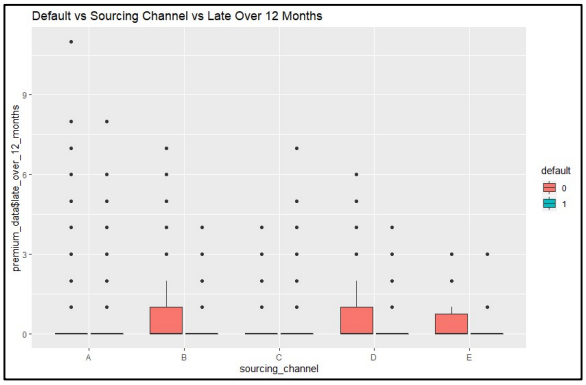


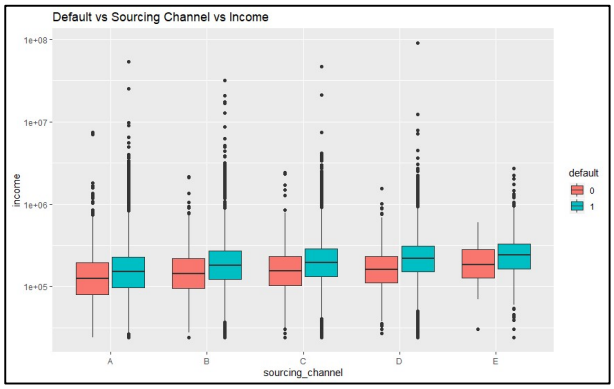


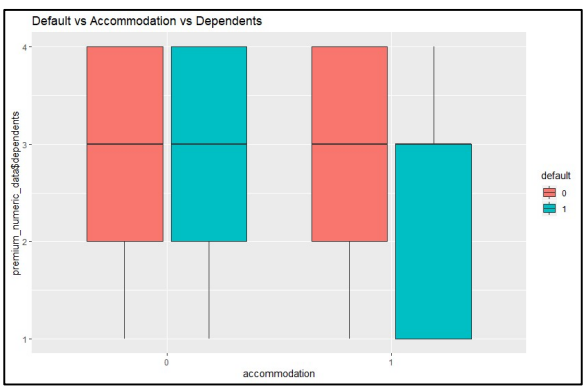


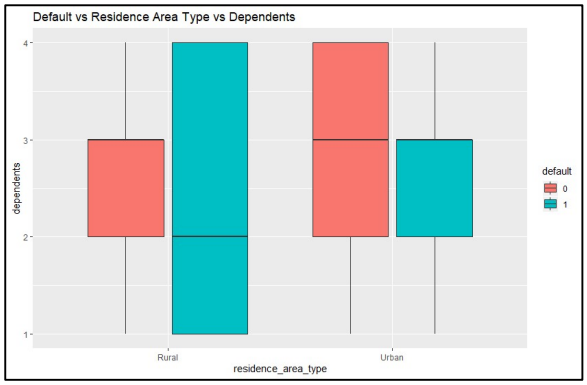


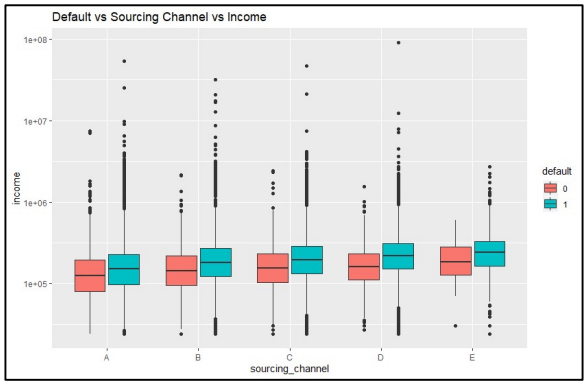






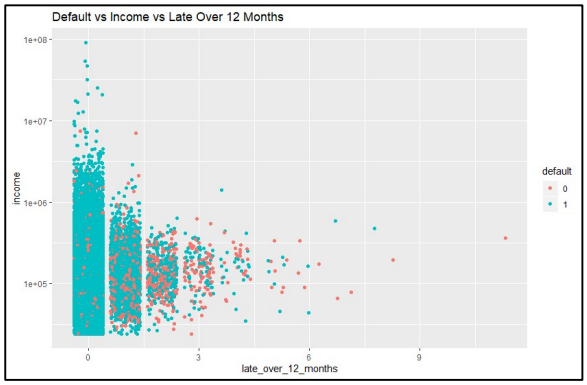


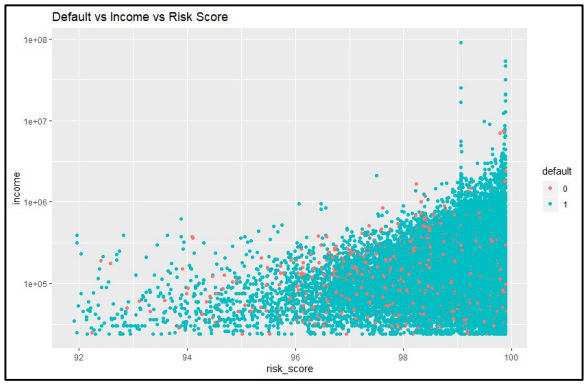

Default vs Residence Area Type vs Dependents
dependents
residence_area_type
Rural
Urban
default
0
1

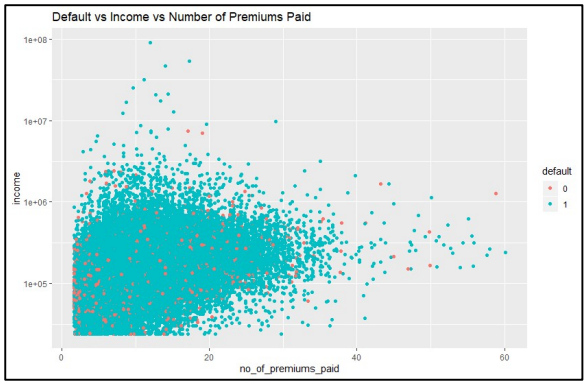

Default vs Sourcing Channel vs Income
income
sourcing_channel
A
B
C
D
E
default
0
1

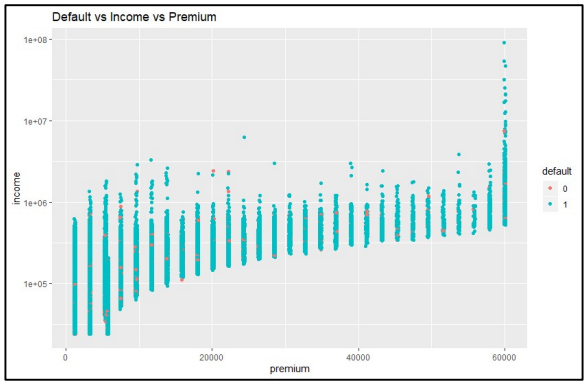

Default vs Income vs Late 3 - 6 Months
income
late_3_6_months
default
0
1

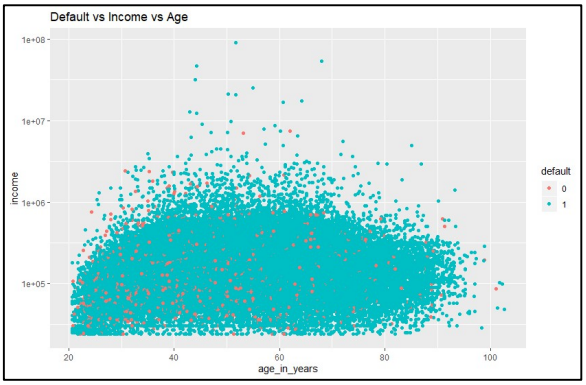

Default vs Income vs Late 6 - 12 Months
income
late_6_12_months
default
0
1

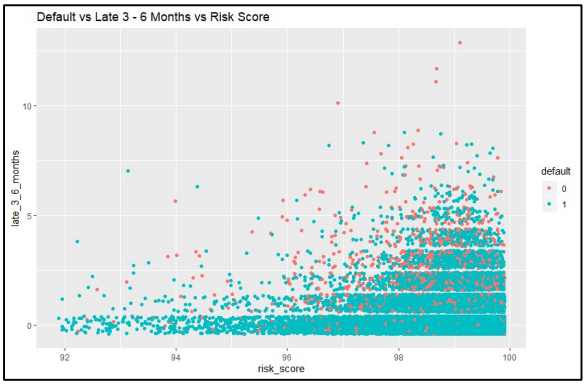

Default vs Income vs Late Over 12 Months
income
late_over_12_months
default
0
1


Default vs Income vs Risk Score
income
risk_score
default
0
1

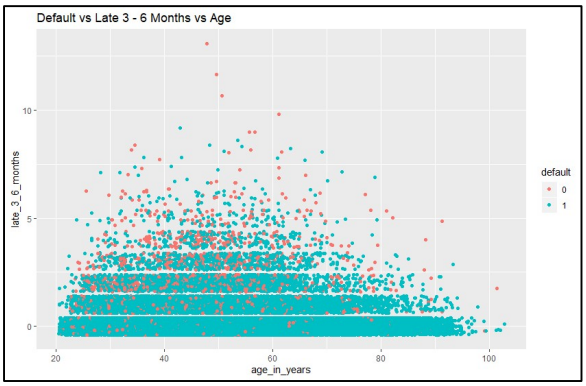

Default vs Income vs Number of Premiums Paid
income
no_of_premiums_paid
default
0
1

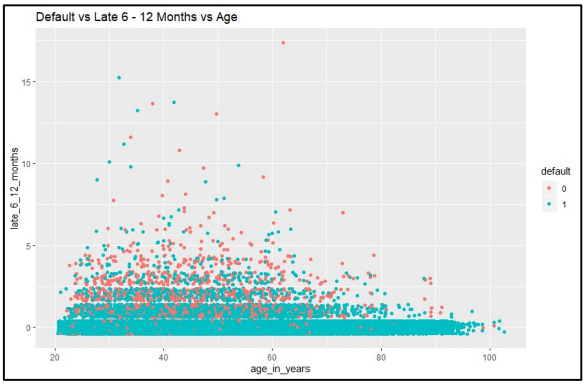

Default vs Income vs Premium
income
premium
default
0
1

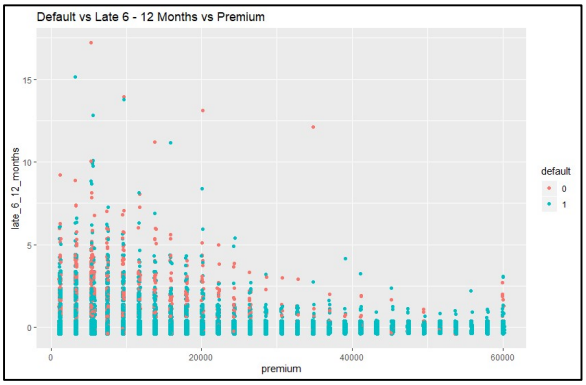

Default vs Income vs Age
income
1e+08
1e+07
1e+06
1e+05
20
40
60
80
100
age_in_years
default
0
1

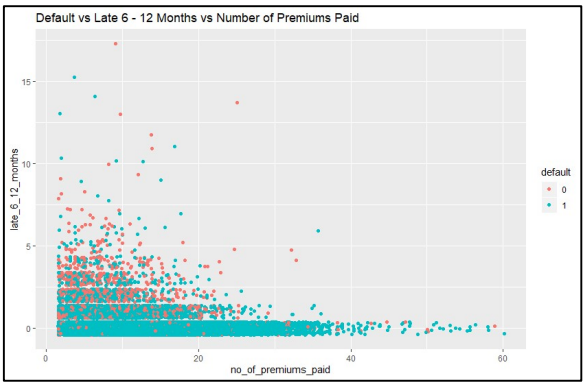

Default vs Late 3 - 6 Months vs Risk Score
late_3_6_months
10
5
0
92
94
96
98
100
risk_score
default
0
1

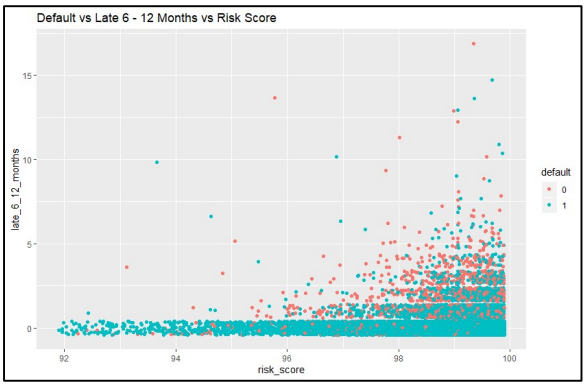

Default vs Late 3 - 6 Months vs Premium
late_3_6_months
10
5
0
0
20000
40000
60000
premium
default
0
1

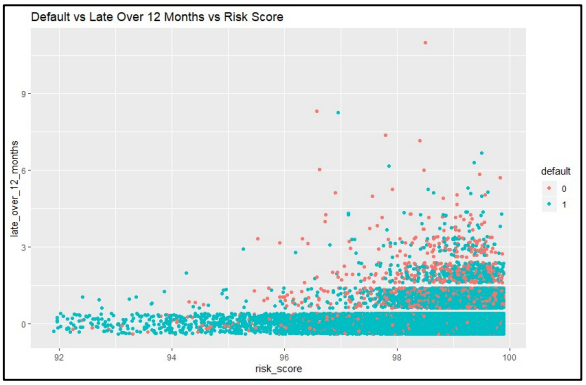

Default vs Late 3 - 6 Months vs Age
late_3_6_months
10
5
0
20
40
60
80
100
age_in_years
default
0
1

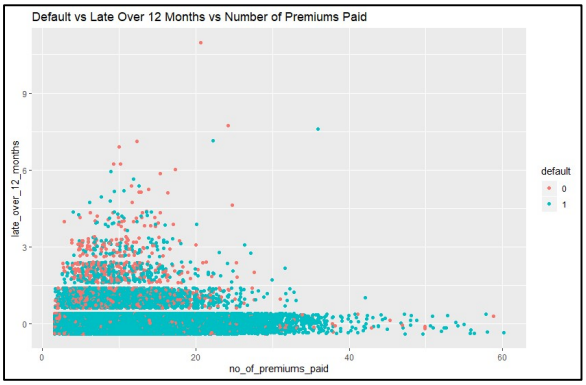

Default vs Late 6 - 12 Months vs Age
late_6_12_months
15
10
5
0
20
40
60
80
100
age_in_years
default
0
1

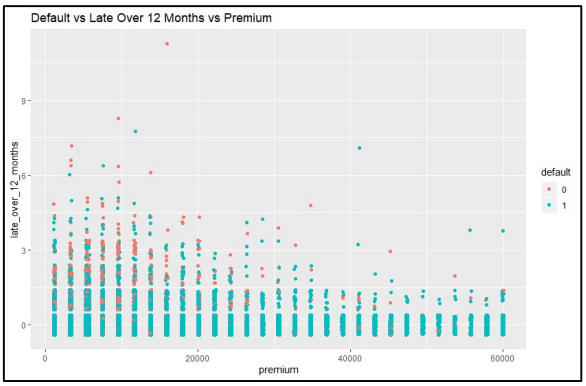

Default vs Late 6 - 12 Months vs Premium
late_6_12_months
15
10
5
0
0
20000
40000
60000
premium
default
0
1

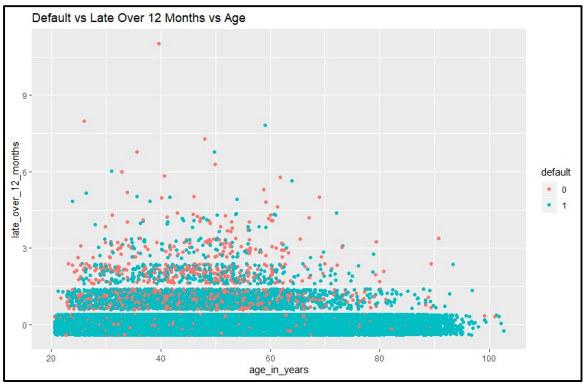

Default vs Late 6 - 12 Months vs Number of Premiums Paid
late_6_12_months
15
10
5
0
0
20
40
60
no_of_premiums_paid
default
0
1

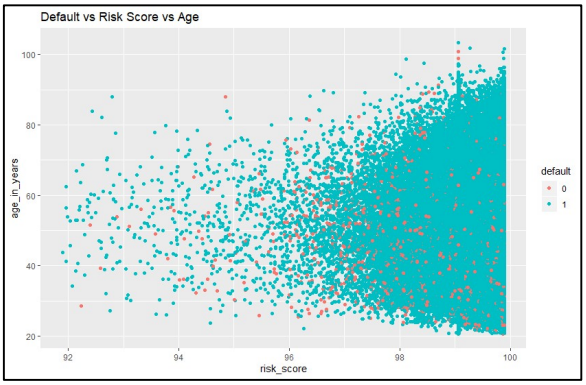

Default vs Late 6 - 12 Months vs Risk Score
late_6_12_months
15
10
5
0
92
94
96
98
100
risk_score
default
0
1

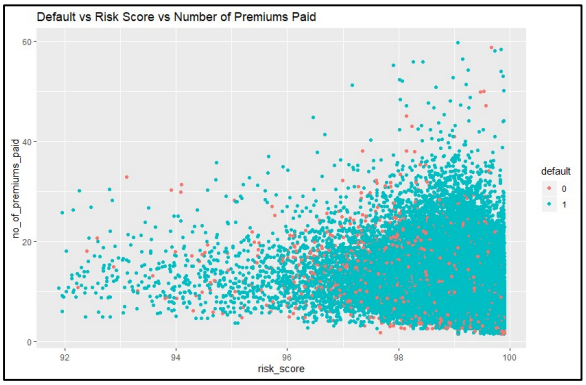

Default vs Late Over 12 Months vs Risk Score
late_over_12_months
risk_score
default
0
1

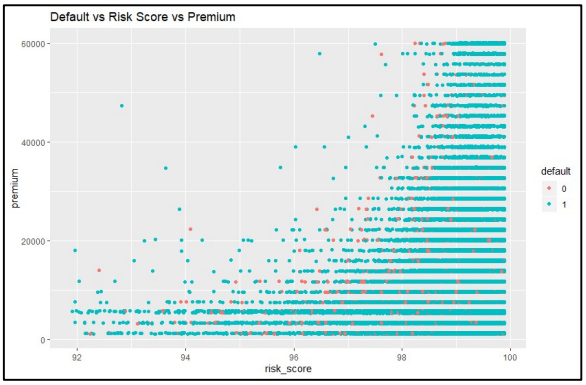

Default vs Late Over 12 Months vs Number of Premiums Paid
late_over_12_months
no_of_premiums_paid
default
0
1


Default vs Late Over 12 Months vs Premium
late_over_12_months
premium
default
0
1

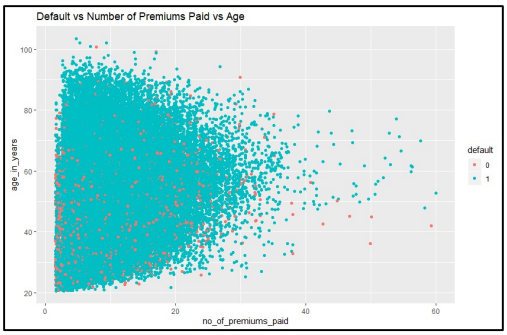

Default vs Late Over 12 Months vs Age
late_over_12_months
age_in_years
default
0
1


Default vs Risk Score vs Age
age_in_years
risk_score
default
0
1


Default vs Risk Score vs Number of Premiums Paid
no_of_premiums_paid
risk_score
default
0
1


Default vs Risk Score vs Premium
premium
risk_score
default
0
1


Default vs Risk Score vs Age
age_in_years
risk_score
default
0
1


Default vs Number of Premiums Paid vs Age
age_in_years
no_of_premiums_paid
default
0
1