greatlearning
Learning for Life

PREDICTION OF PREMIUM DEFAULTERS:

The Renewal Of Premium Payment In An Insurance Company

NOVEMBER 8, 2020

By: Chinedu H Obetta

greatlearning
Learning for Life

Table of Contents:

greatlearning
Learning for Life

EXECUTIVE SUMMARY

General Scope:

The ability of any organization to generate and sustain revenue is one of the major determinant of its survival in the nearest future, and this is applicable to every organization. In the insurance industry, their revenue strengths are based on premiums collected from policyholders.

The objective of the project:

- To develop a model that will determine and predict, with high degree of confidence, policyholder that have a higher probability to default in premium payment. While the probability will help the insurance company in managing its cash flows over a given period, it will also help the company to engage the policyholder that have high propensity to default in premium payment.

- To explore the data for a general and preliminary overview of the policyholders and their propensity to renew the policy upon expiration. This will enable the insurance company to predict and track policyholder that will likely default in premium payments. The idea will also help the company in its cash flows management over a given period of time.

Brief Description Of Methods -The exploratory data analysis:

Brief Profile Of The Policyholders

1. There are 79, 853 insurance policyholders under review, out of which 50% are married while the balance of 50% are single.
2. The dataset also shows that only 6% of the policyholders defaulted in the renewal of their insurance policy over time.
3. We also observed that most of the policyholders are super rich. Our analysis revealed that closed to 40% of the policyholders earn more than $200,000 per annum while 3% earn less than $45,000 per annum.
4. We also noticed that the premium is meant for customers that are advanced in age. This is noting that over 70% of the policy holders are between 41 and 74 years old, 25% are Millennials while the balance of 6% are over 75 years old.
5. For every 10 policyholders, 2 of them failed to make a payment of their premium within 30 days. Thus, the company's late payment rate for the period under review is 20%.
6. Slight majority of the policyholders are urban dwellers while 40% resides in the rural communities.

7. A good number of policyholders reside in their own homes, while 60% owns their place of residence.
8. The customers of the company are sourced through 5 distinct channels. About 54% of the policyholders are sourced through only channels alone.

Chapter 1c. Final Insights & Recommendations

- It is revealed that the rate of default increases as the number of late payment increases. It is therefore important for the management to put measures or strategies that will encourage the policyholders to make an early or prompt renewals of their insurance premiums.

- It is also understood that the rate of default slightly vary according to the age range of the policy holders. The default rate decrease as the age range of the policy holders increases. The Silent generations (age between 75 & 95) hardly fail in the renewal of their insurance premium while the millennials are the highest defaulters. It is therefore recommended for the company to modify its prizing to reflect the identified trend of default rate across the age groups.

- Notwithstanding that the rate of default is a bit higher at the low income earners, it can be concluded that it is similar across all the income group.

- Prior to now, the company could only estimate 6% of its policyholders that default in the repayment of their insurance policy. With the model, we can comfortably predict 83% of the policy holders that default in the renewal of insurance policy. This is a significant improvement of 76%.

Chapter 2. Approach-Logical steps to final model selection.

To develop a model that will determine and predict, with high degree of confidence, policyholder that have a higher probability to default in premium payment, the following steps were followed:

a. **Set up working Directory & Environment Set up:** This is started at the onset of the R session. It makes the importing and exporting of the data and code files easier. It is the location of the folders in your PC. Ie, where the dataset and codes related to the project are stored.

b. **Install necessary Packages and Invoke Libraries:** Under this sections, all the necessary packages are installed and its associated libraries are invoked. Installing all the packages in same location increases the code readability.

c. **Import and Read the Dataset:** The given dataset is in excel format. Hence, the command "read.excel" is used for the importation of the file**.**

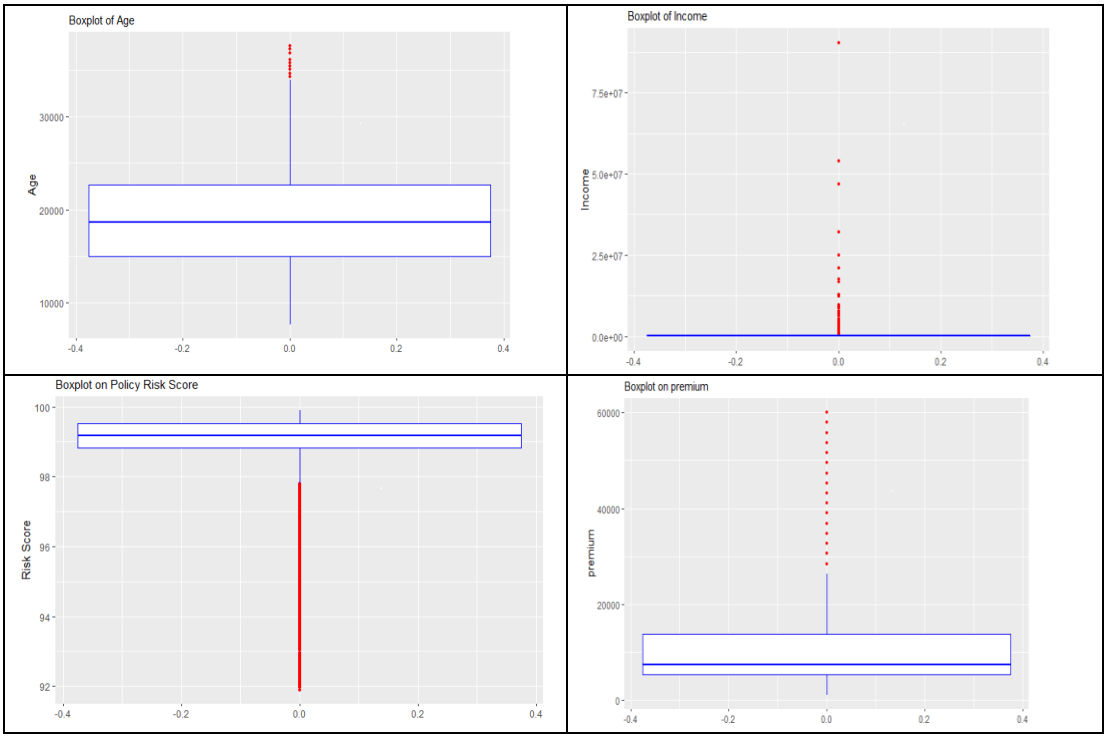
d. **Variable Identification:**

1. **ID:** Unique Policyholder's ID
2. **Cash. Credit:** The proportion of premium that was paid by cash payment
3. **Age:** The policyholder's Age in years.
4. **Income**: Income of the policyholders
5. **Marital Status:** The marital status of the policyholder (Married or Unmarried)
6. **Vehicle:** The number of vehicles owned by the policyholders.
7. **Dependents:** The number of dependent on the policyholders.
8. **late.pmt.3_6:** The number premium that was paid late in the last 3 to 6 months.
9. **late.pmt.6_12:** The number premium that was paid late in the last 6 to 12 months.
10. **late.pmt.More_12Mnth:** The number premium that was paid late 12 months ago.
11. **Accommodation:** Owns or rented place of accommodation
12. **Risk Score:** The risk score of the policy holders.
13. **No premium:** The total number of premium paid by the policyholder.
14. **Residence:** Whether the policyholder resides in urban or rural.
15. **Sources:** The channel through which the customer was sourced
16. **Premium:** The premium paid by the policyholders.
17. **Default:** This indicates policyholder that have renewed their premium.

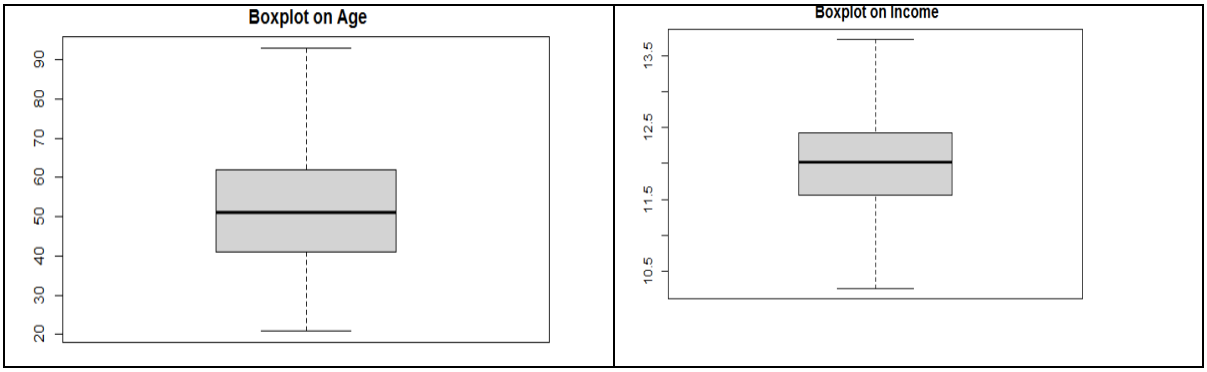
e. **Missing Value Identification**. This is no missing value in the observation.

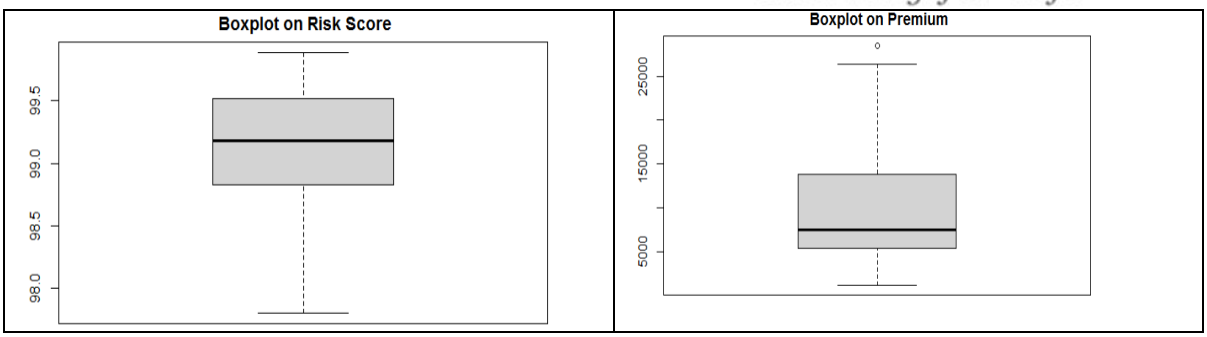
f. **Outlier Identification & Treatment : Checking For Outliers On Continuous Variables:**

- **The boxplot below shows the existence of outliers on the identified variables**



The boxplot below shows that the outliers have been treated. This is noting that outliers are no longer in the dataset. The details for the treatment of the outliers are attached as appendix on the project





f. Variable transformation.

Variable transformation otherwise known as the feature Engineering is required to generate business insight from the dataset. Here, we will modify existing features to get a better insights into the dependent variable “Default".

The Variable: AGE

- The age of the policy holders were recorded in days instead of years. Thus, the variable "Age" was transformed to be in years instead of days. This will give us more useful insight about the age of the policyholders and its relevant on the dependent variable.

Variable: late.pmt.3_6, late.pmt.6_12 and late.pmt.More_12Mnth

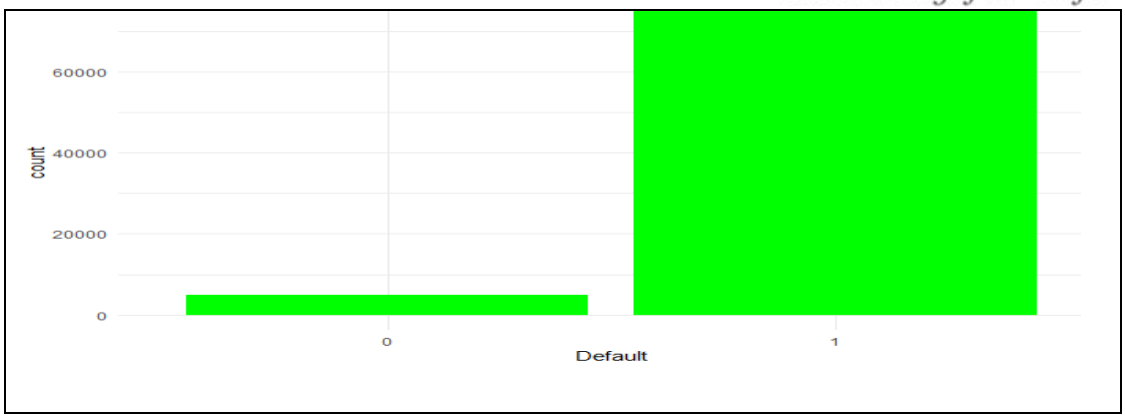
- These variables are similar as they measure the number of late payment over a given period of time.
- A new column which gives us the information about the total number of late payment by each of the policyholder is required and it would help us in the analysis to a more useful insight.

Variable: Income

- The variable "income' is tightly skewed to the left and no insight could be derived from the histogram. The box plot also shows is complex data structure. Thus, there was need for the data to be transformed before meaningful insight could be inferred from the analysis.

g. Exploratory Data Analysis- Univariate Analysis & Bi-Variate Analysis.

G1. Distribution of the dependent variable-Default

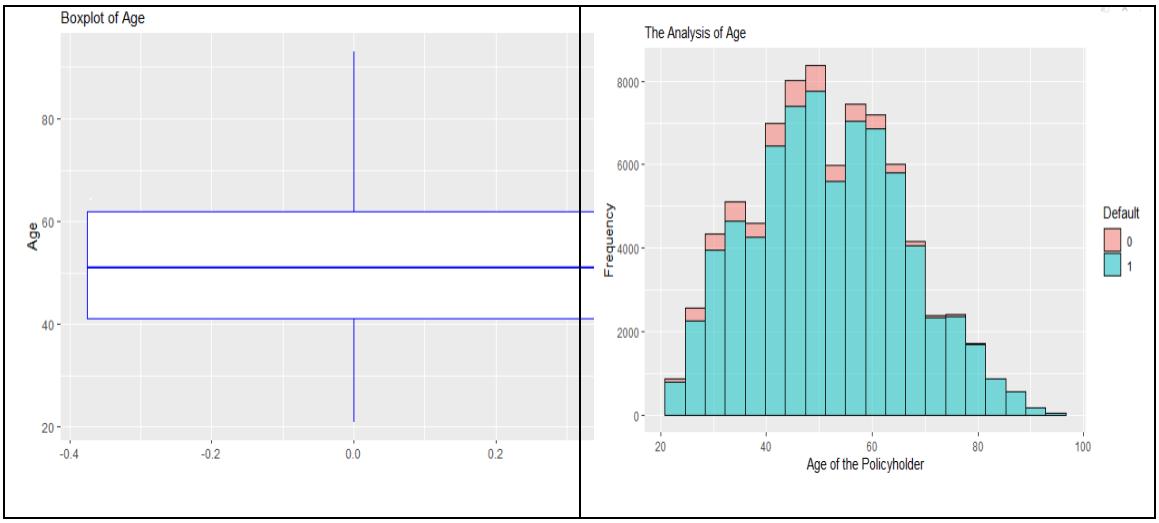


Observations:

- 6.3% of the policyholders under review defaulted in the renewed their insurance while the balance of 93.7% of the policyholders has renewed their premium.
- The distribution of the dependent variables appears imbalanced, thus, the tendency of the model to be skewed to the number of policyholders that renew their premium is very likely. While the output seems imbalanced, I am not sure if the imbalance will be significant.
- This calls for the dataset to be preprocessed further to ensure accurate development of predictive models. The model when developed will help the insurance company to predict policyholders that have very high probability to default in premium payment and appropriate action will then be taken on the identified customers or policyholders.

1.1.1 Univariate Analysis - Histogram and Boxplots

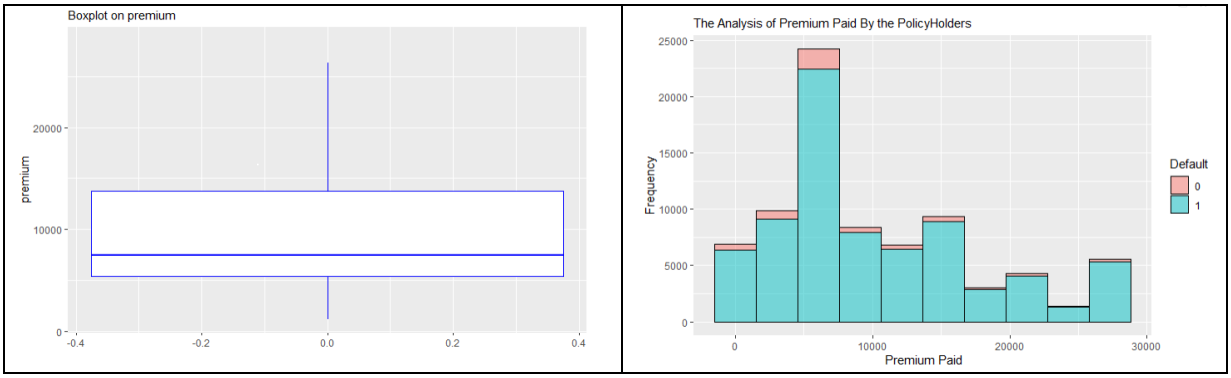
Insights On The Age Analysis:



- The average age of all the policyholders is 52 years old, the youngest and oldest policyholder is 21 and 103 years old respectively.

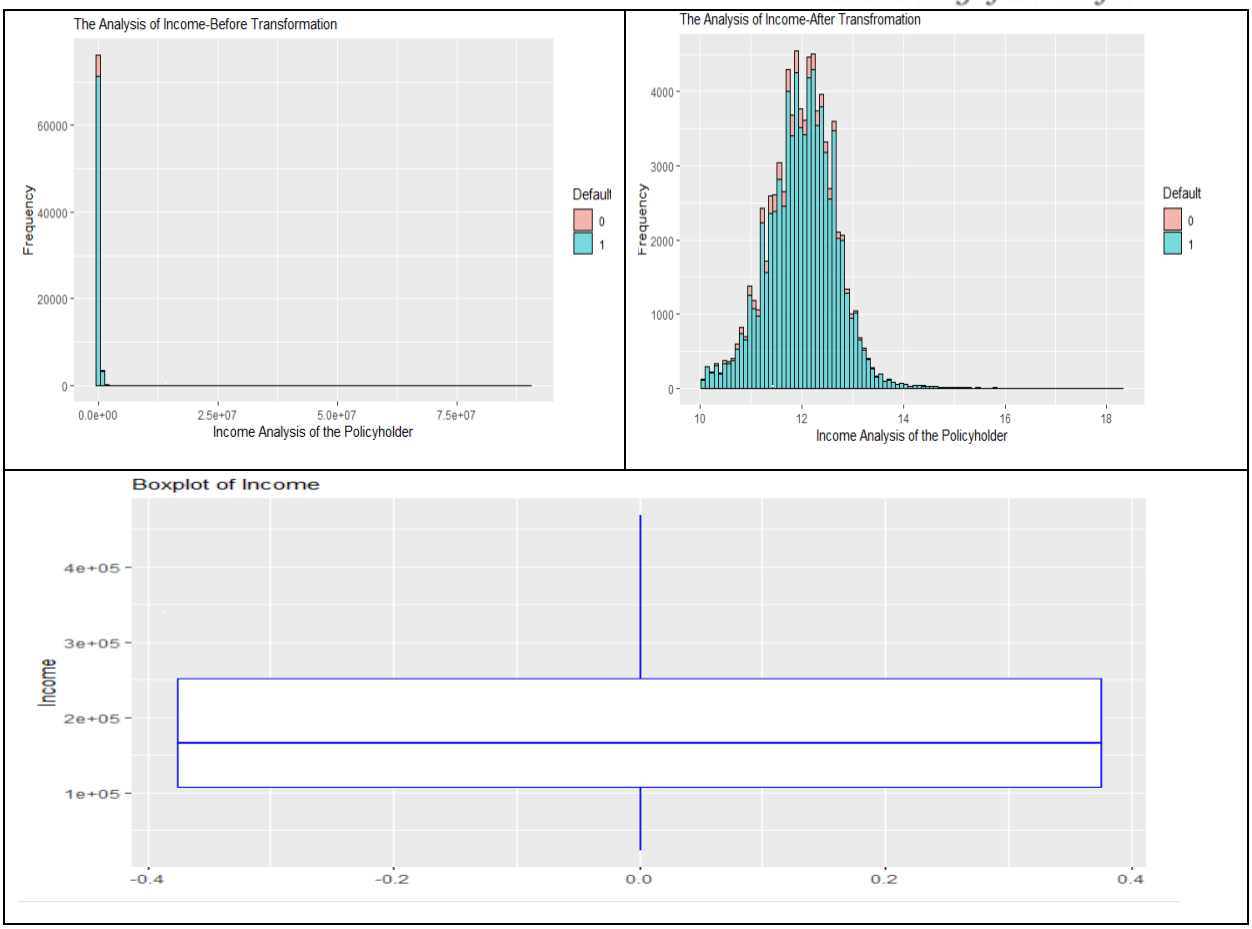
- There is no much difference in the age range of policyholders that renew their policy and those that do not renew theirs as shown
- Most of the policy holders are within the working age as 75% of all the policy holders are below 62 years old.

Insights On The Analysis Of Premium Paid By The Policyholder:



- There seems to be no difference in premium paid amongst the policyholders that renew their policy and those that do not
- The average premium paid by policyholders is USd10,988 and 75% of them pay less than USD13,800 .

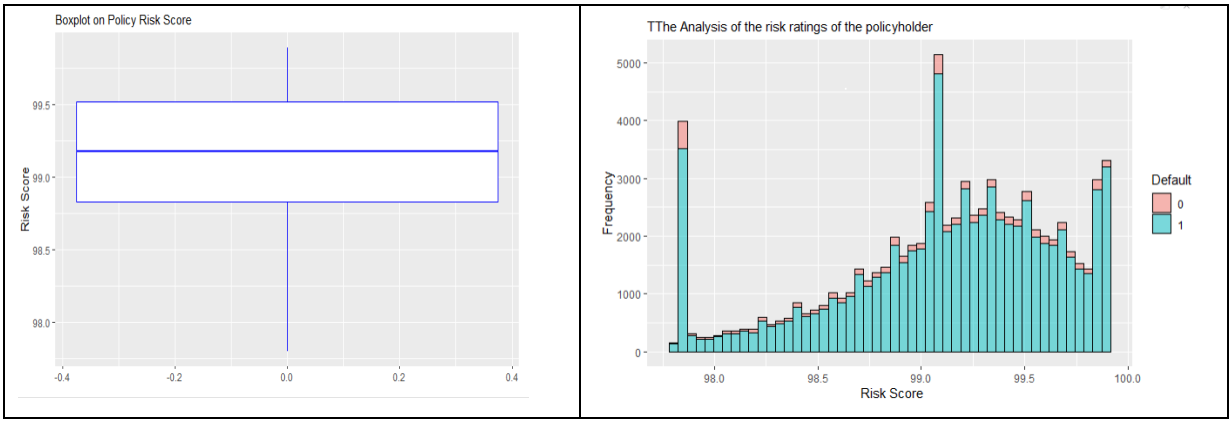
Insights On The Analysis Of The Income Of The Policyholders



Further transformation and the treatment of outliers revealed as follows;

- There is no much difference on the distribution of income between the policyholders that make premium payment and those that do not.
- The policyholders are high income earners as most of them earn over USD150,000. We also noticed that 75% of the policyholders earn up $252,000.00

Insights On The Analysis Of The Risk Score Of The Policyholders

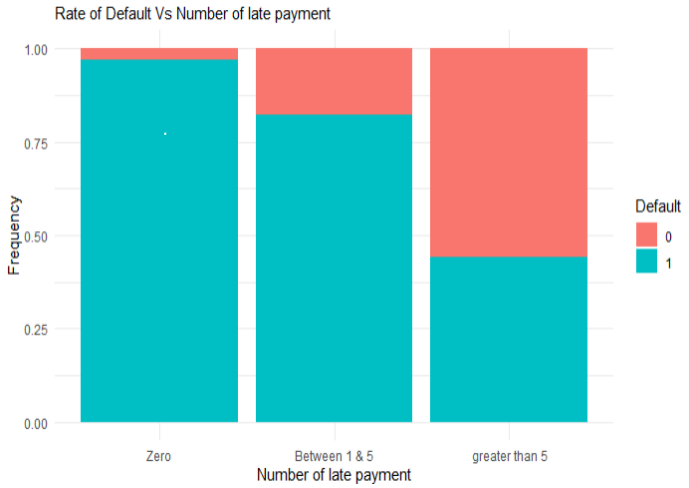


- The risk score is skewed to the right with an average risk rating of 99.08. The minimum and maximum risk score is 92.76 and 99.89 respectively.

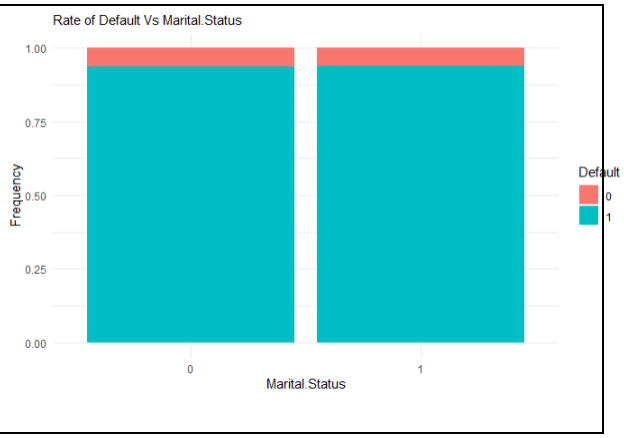
1.2 Bivariate Analysis

Let us plot percent stacked bar chart to see the effect of independent variables on the rate of default

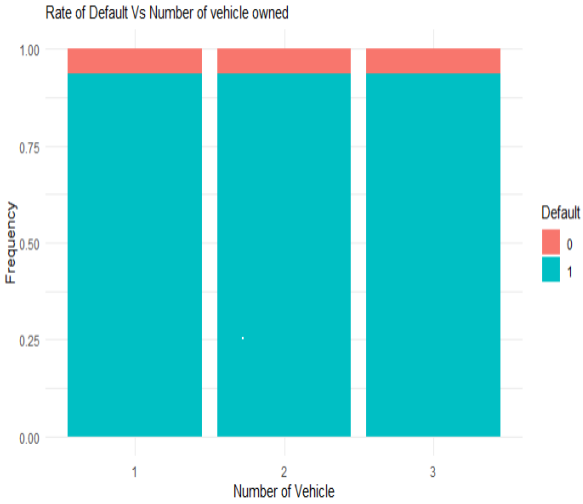
Rate of Default Vs Number of late payment

Insights	
• More than 50% of the policyholders that has a record of more than 5 late payment default in the premium payment. • It is also observed that most of the policyholders that do not have any record of late payment hardly miss their payment. • It is also inferred that the rate of default increases as the number of late payment increases.	

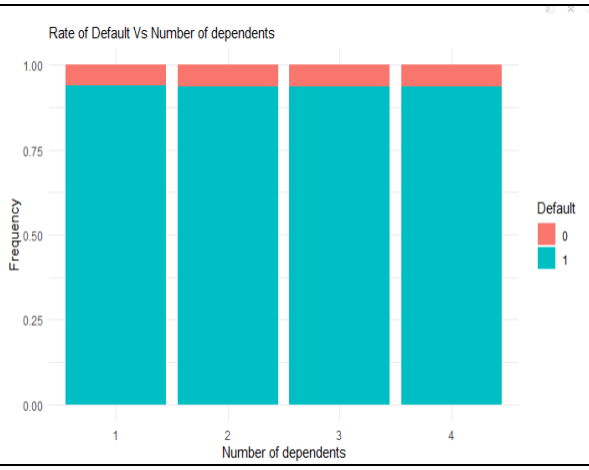
3.4.4: Rate of Default Vs Marital. Status

Insight:	
• The renewal of the insurance policy does not seem to be dependent on the marital status of the policyholder. This is noting that the distribution of those that failed to renew the premium payment is same across the marital status • The test statistic also confirms this as p-value is more than 0.05	

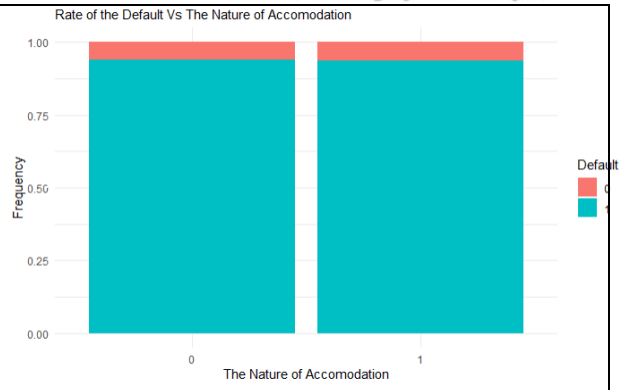
3.4.5: Rate of Default Vs Number of vehicle owned

Insight: • The distribution of those that defaulted is similar across the number of vehicles owned by the policy holders, thus, the renewal of the insurance policy does not seem to be dependent on the number of vehicles owned by the policyholder. • The test statistic also confirms this as p-value is more than 0.05.	

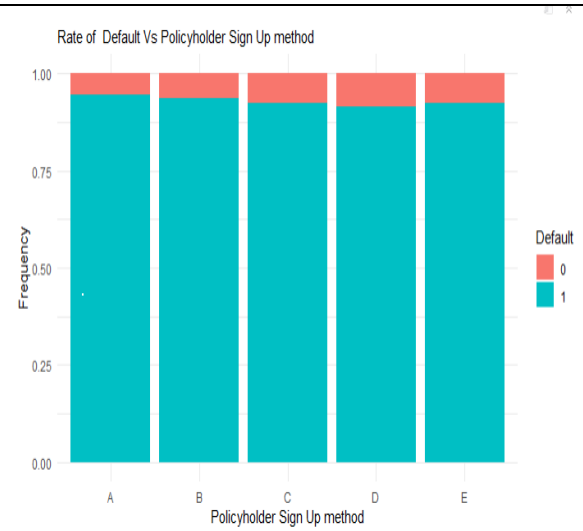
3.4.6: Rate Of Renewal Vs Number Of Dependents

Insight: • The payment of premium does not depend on whether the policyholder resides in a owned or rented apartment. This is noting that the distribution of the default is similar across the nature of the accommodation	

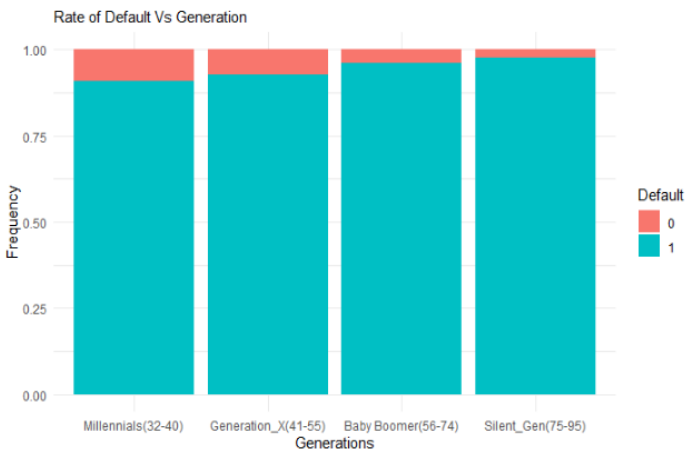
3.4.7: Rate of Default Vs The Nature of Accommodation

Insight: • The distribution of rate of default is similar across the policyholder's accommodation type. Thus, the nature of the accommodation does not seem to influence whether the policyholder will default or not	

3.4.8: Rate of Default Vs Policyholder Sign Up method

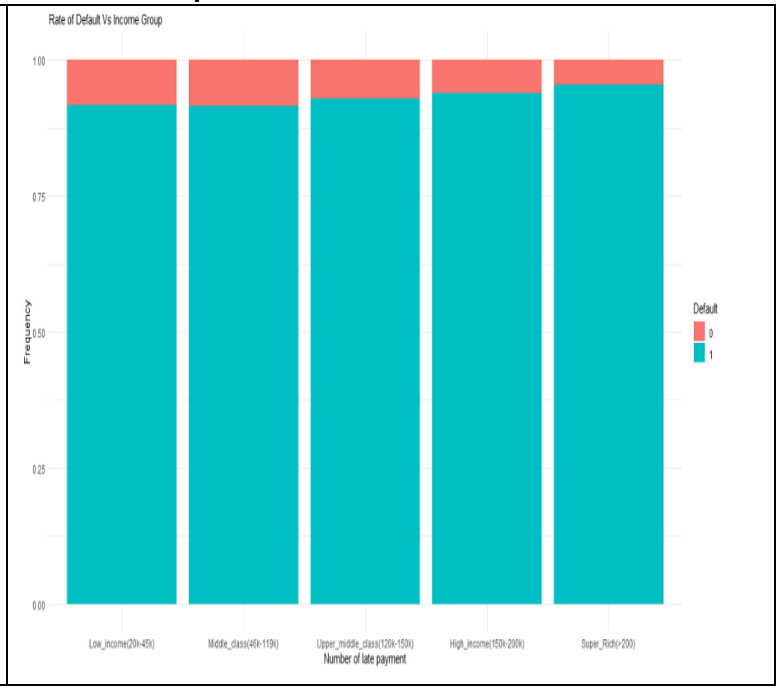
Insight: • The method through which policyholders are sourced does not have any significant effect on whether the policyholder will default or not.	

Rate of Default Vs Generations

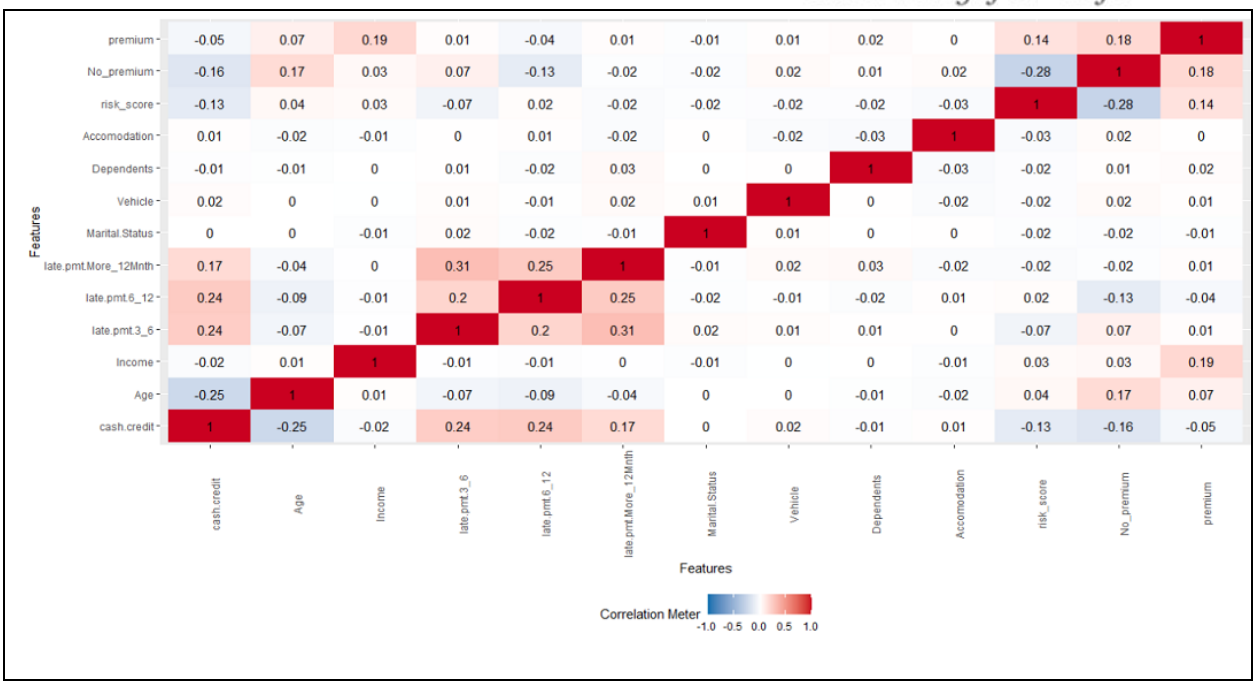
Insight: • It is also understood that the rate of default slightly vary according to the age range of the policy holders. The default rate decrease as the age range of the policy holders increases.	

Rate of Default Vs Rate of Default Vs Income Group

- Notwithstanding that the rate of default is a bit higher at the low income earners, it can be concluded that it is similar across all the income group.



Correlation Plot Between Numeric Variables in The Dataset



Insight

- It appears there is no strong correlation amongst the potential predictor variables as shown the correlation plot above. This means that the predictor variable are independent of each other, thus, the model will not be affected by the problem of multicollinearity.
- In other words, the independent variables do not have any effect on another. The increase or decrease in any of the variables do not have any effect on other variables

CHECKING FOR MULTICOLLINEARITY

- This is to confirm if there is any variables that are correlated with each other. Variables with high VIF (Variance Inflation Factor) will be excluded from the set.
- The test of multicollinearity shows that no variable from the 15 input variable has a collinearity problem, and this means none of the variables are related to each. As result, all the variables will be used in building or developing models for predictions.

- The test also reveals the linear correlation coefficients amongst the variables are between the ranges as shown below:

```
min correlation ( No_premium ~ Vehicle ):  -0.0001110382
max correlation ( late.pmt.More_12Mnth ~ late.pmt.3_6 ):  0.2755535
```

- The variables and their corresponding variance inflation factors are shown below;

Variables	VIF
cash.credit	1.208871
Age	1.143247
Income	1.070717
late.pmt.3_6	1.147808
late.pmt.6_12	1.117405
late.pmt.More_12Mnth	1.135962
Marital. Status	1.002923
Vehicle	1.002813
Dependents	1.003640
Accommodation	1.002268
risk_score	1.166147
No premium	1.246590
Sources	1.073409
Residence	1.004357
premium	1.166130

h. Model Development, Selection and Optimization:

Analytical approach: The prediction problem under review is classification as the data has a dependent variable with a distinct categories that the model needs to predict. In this case, the algorithms will learn a function to come up with a boundary that will help in classifying the data into class labels as defaulted or not defaulted.

Being a classification problem, there various algorithms that can be deployed to learn and subsequently classified the data into defaulted or not defaulted. Some of algorithms are as follows;

- **Logistic Regression:** This is most common learning algorithm for supervised learning and it is based on parametric statistics. The algorithm will classify whether the policyholders will default or not on the payment of premium. The classification will be based on the linear combination of the independent variables. The reliability of the algorithm will depend on whether the data complies with all its assumption. These include, the assumption that all the variables are independent etc.
- **Decision Trees:** This is another algorithm that we can use to predict whether the policyholder will default or not. The model uses a tree to represent a number of decision

paths and the outcome of the each path. The model is known for its interpretability and this means that it could enable the insurance the company to determine understand why some policyholder will default in the payment of premium. This insight could enable to develop a policy or guidance around it. The model has a challenge of overfitting of dataset.

- Naïve Bayes is a sophisticated Linear Classification models that can be used to predict the class variable. The model is also based on Bayesian theorems of conditional probabilities and it is subjected to some of the assumptions of the probability such as independencies of the variables.
- I will try other non-linear classification methods such as KNN and Random Forest for the prediction of the class variables.
- In addition, ensemble methods such as Boosting and Bagging will also be used to come up with robust results on whether the policyholder will default or not in premium payment

H2. Modelling Process validation

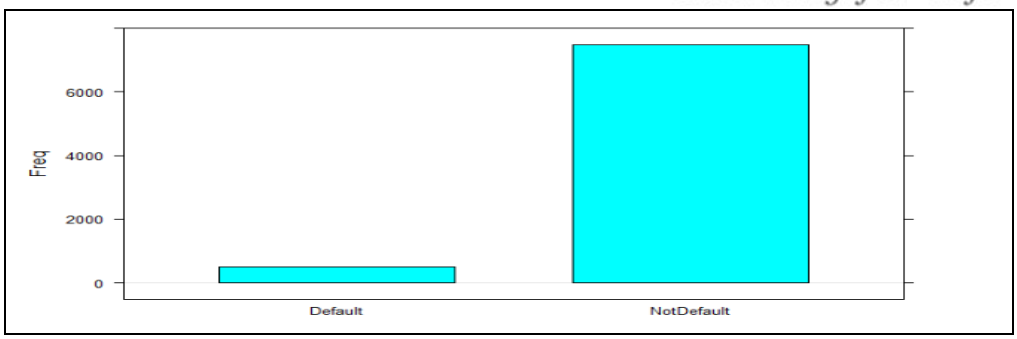
This is basically a process of evaluating and applying different models on training dataset and see which models that performs best for selection. In addition to finding the right model, we are also required to find the right parameters for the model. At the end, the selected model with the appropriate parameter will be evaluated on the testing dataset.

At this stage, let us get the dataset partitioned for model validation and noting the size of the dataset (79,853), there is need for us to take 5% of the whole data to speed-up the model building & iterations. Thereafter, the sample of the dataset is split into two categories in the ratio of 70:30 as explained below:

Training Dataset:	70% of the 10% data-subset. This represent 5,590 dataset
Testing Dataset:	30% of the 10% data-subset. This represent 2,395 dataset.

Model Selection:

- The selection of models and its corresponding parameters is also dependent on the problem definition and other features of the dataset. An exploratory review of the dependent variable 'Default' shows that the dataset is imbalanced as 6.3% of the policyholders under review defaulted in the renewal of their insurance while the balance of 93.7% of the policyholders has renewed their premium. See below.



- The problem under review is a classification problem as such, metrics like accuracy, precision, recall, specificity and AUC will be used for model comparisons. While an accuracy is an important metric for model comparison, it will not be useful for this dataset due to observed imbalance distribution of the dependent variable.

- **TREATMENT OF IMBALANCE OF THE DATASET:** To develop appropriate model that can generalize the dataset, it is important to ensure that the dataset is balanced. To balance the dataset, we are going to use two of the most widely used techniques such as k-fold cross-validation and resampling technique.
  -
  - The sampling technique is further divided into 3 such as Over Sampling, Under Sampling and SMOTE (Synthetic Minority Over Sampling Technique). The Over sampling technique which tries to randomly generate a sample of attributes from observation in the minority class in this case Defaulters will be used in this project. This is noting its predictive ability amongst the other two sampling techniques.
  - K-fold Cross-Validation is going to be used to address the problem of imbalance in the dataset. Given that our choice of sampling technique above will introduce randomness into the dataset, the cross validation technique will be deployed on the dataset before sampling technique.
  - It is important to note that these two methods of treating the imbalance in the dataset will be introduced in the control parameter for all the prospective models to aid model comparisons.

MODEL SELECTION: since the case under review is classification problem with two class labels, i will be evaluating the following algorithms;

- Logistics Regression
- Naïve Bayes
- Random Forest

- KNN and
- Xgboost.

Logistics Regression: This learning algorithm is based on parametric statistics and it will classify the class labels be based on the linear combination of the independent variables. The algorithm revealed that out of 16 predicator variables, only 8 were statistically significant in driving the default probabilities of the policy holders. These include;

- Cash. Credit,
- The Age of the policy holder.
- The Income,
- Number of Premium,
- The Policy Holders Means of subscribing (Sources)
- The number of late payment,
- The policy holder's Accommodation Status
- The number of vehicle owned by the policy holder.

Variable of Importance	**Predictors**	**Weight**
A review of the model's variables of importance shows that 6 variables contributed 90% predictability power of the algorithm to effectively derive the probability that a policy holder will default in the premium payment or not.		
	cash.credit	30%
	late.pmt	27%
	No_premium	10%
	Age	6%
	SourcesD	14%
	Income	4%
		90%

Model Performances:

		Sens	**Spec**	**Accuracy**
The overall performances of the model is marginally acceptable even though	Training Dataset	72%	78%	77%

the accuracy for overall prediction is in the 70s. The model was able to generalize the dataset to a certain extent as the sensitivity and specificity of the model in the both the training and testing dataset for the prediction of default are similar. Noting that the training error in the two dataset are more than 25%, thus, the model contain a lot of bias error and variance error	Testing Dataset	70%	77%	76%
	Difference	2%	1%	1%

Naïve Bayes: This is another machine learning algorithm that is based on statistical theorem and it operates based on Bayes Theorem.

		Sens	Spec	Accu
The overall accuracy of the model is a bit high when compared with the Logistic Regression. On the contrary, the model accuracy is not important metric for this case. Sensitivity is an important metric because we are dealing with financial data. It is expected that our model should be able to detect Defaulting customers. The sensitivity for both training and testing dataset is poor and this signifies the existence of high variance.	Training Dataset	64%	88%	86%
	Testing Dataset	56%	88%	86%
	Difference	8%	0%	0%

KNN-Models:

This is an algorithm that classify dependent variable based on the similarities of the out of the data. The similarity is determined by looking the distance that exist amongst the observations.

Even though, the KNN seems to be a better model (in terms of sensitivity) when compared with Naïve Bayes, the training and testing error are still high. Thus, there is a high bias and high variance.		Sens	Spec	Accu
	Training Dataset	75%	74%	74%
	Testing Dataset	70%	72%	72%
	Difference	5%	2%	2%

Random Forest is one of the ensemble models.

The model under review shows over fits the dataset. The model learnt a lot of noise at the training. This is evidenced as the metric of choice (sensitivity) drop to 9% at the unseen dataset from 98% at the training dataset. Thus, this is an overfit model and not useful for the dataset.		Sens	Spec	Accu
	Training Dataset	98%	100%	99%
	Testing Dataset	9%	99%	93%
	Difference	89%	1%	6%

Xtreme Gradient boosting Machines is also one of the ensemble models and it is selected due to its effectiveness in the management of bias and variance.

Just like the previous models, Xtreme Gradient boosting's error rate is also high. This is evidenced in all the metrics such as sensitivity and specificity. The model is however opened for comparisons.		Sens	Spec	Accu
	Training Dataset	74%	78%	78%
	Testing Dataset	66%	76%	75%
	Difference	8%	2%	3%

Summary of the models performances:

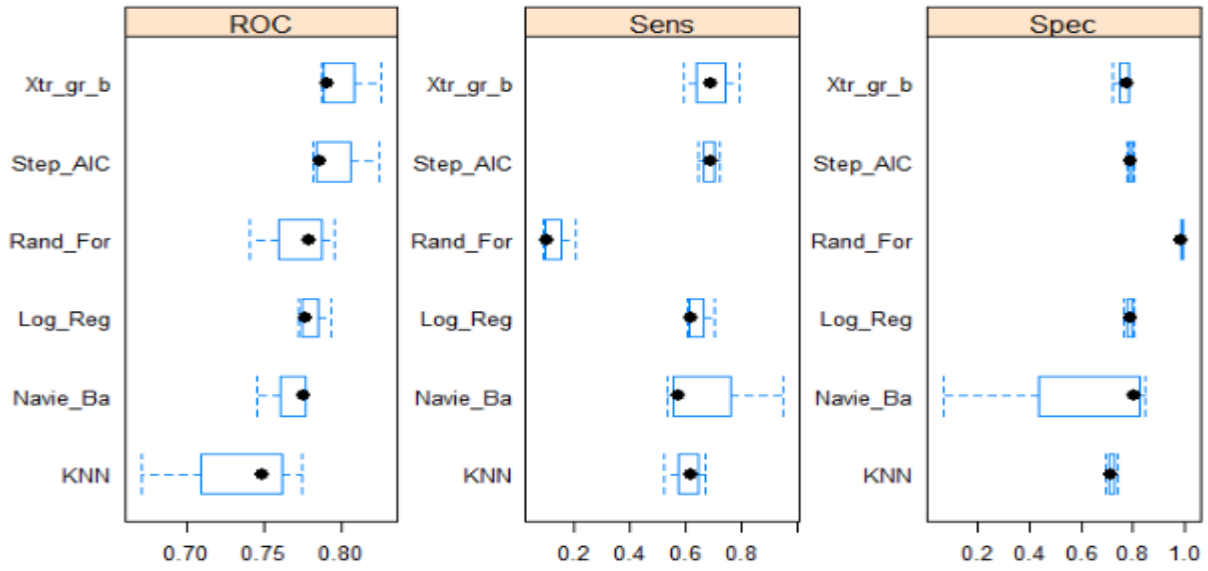
While the review of other metrics are important for model comparisons, emphasis will be placed on the sensitivity of each models. By sensitivity, we mean the ability of the model to correctly predict policyholder that defaulted in the renewal of the policy. Importance is placed on the sensitivity because it is better for the insurance company to correctly predict customers that will default than those that will not default.

The models comparisons are shown below:

```
ROC
               Min.    1st Qu.     Median       Mean   3rd Qu.       Max. NA's
Log_Reg   0.7718378 0.7745601 0.7772824 0.7805899 0.7849660 0.7926495    0
Step_AIC  0.7813094 0.7839379 0.7865663 0.7975273 0.8056362 0.8247061    0
Navie_Ba  0.7452779 0.7604615 0.7756451 0.7660427 0.7764251 0.7772050    0
KNN       0.6714599 0.7101177 0.7487755 0.7316544 0.7617516 0.7747278    0
Rand_For  0.7406691 0.7598258 0.7789825 0.7717414 0.7872776 0.7955727    0
Xtr_gr_b  0.7871292 0.7888105 0.7904918 0.8009730 0.8078948 0.8252979    0

Sens
               Min.    1st Qu.     Median       Mean   3rd Qu.       Max. NA's
Log_Reg   0.6101695 0.61542957 0.6206897 0.6459186 0.6637931 0.7068966    0
Step_AIC  0.6440678 0.66686148 0.6896552 0.6859536 0.7068966 0.7241379    0
Navie_Ba  0.5344828 0.55537697 0.5762712 0.6863433 0.7622735 0.9482759    0
KNN       0.5254237 0.57305669 0.6206897 0.6061757 0.6465517 0.6724138    0
Rand_For  0.0862069 0.09482759 0.1034483 0.1310150 0.1534191 0.2033898    0
Xtr_gr_b  0.5932203 0.64143776 0.6896552 0.6919930 0.7413793 0.7931034    0

Spec
               Min.    1st Qu.     Median       Mean   3rd Qu.       Max. NA's
Log_Reg   0.76201373 0.7750504 0.7880871 0.7844348 0.7956454 0.8032037    0
Step_AIC  0.77459954 0.7814645 0.7883295 0.7886358 0.7956539 0.8029782    0
Navie_Ba  0.06864989 0.4370709 0.8054920 0.5724039 0.8242809 0.8430699    0
KNN       0.69107551 0.7034988 0.7159221 0.7157574 0.7280984 0.7402746    0
Rand_For  0.98512586 0.9862700 0.9874142 0.9885557 0.9902707 0.9931271    0
Xtr_gr_b  0.72082380 0.7493008 0.7777778 0.7615476 0.7819095 0.7860412    0
```



A review of the model comparisons reveals the followings;

- KNN and Naïve Bayes are the worst performing models for the dataset.
- Random Forest and Logistic Regression are above average. Their AUC, Sensitivity and Specificity are relatively acceptable when compared with KNN and Random Forest.
- Extreme Gradient Boosting performed better than other models under review as its AUC, sensitivity and specificity are higher than others. Thus, Extreme Gradient Boosting is our best model for the dataset.

Extreme Gradient Boosting Model

- As shown below, the confusion matrix for the model shows that it misclassified 19 Defaulters out of the 75 Defaulters. While this is improvement from the base model of 6% to 74%, let's see if the sensitivity can be improved by adjusting the False Negative without compromising the accuracy of the model significantly.

Prediction	Default	NotDefault
Default	56	280
NotDefault	19	842

- Adjusting the probability of classification to 0.44, we were able to reduce the misclassification of Defaulters from 25 to 23. This increased the sensitivity of the model from 74% to 83% while the specificity reduced from 76% to an acceptable rate of 73%. The accuracy of the model dropped from 74% to 66% Kindly see the table below.

Prediction	Default	NotDefault
Default	62	388
NotDefault	13	734

Relevance and implementability of the conclusions and recommendations:

- Prior to now, the company could only estimate 6% of its policyholders that default in the repayment of their insurance policy. With the model, we can comfortably predict 71% of the policy holders that default in the renewal of insurance policy. This is a significant improvement of 66%.

- As pointed out earlier, the sources of onboarding customers contributes significantly in determining the probability that a customer will default or not. It is therefore important for the company to pay so much of attention to its source of onboarding customers.

- Some of the information or data collected did not impact so much in the projects. Thus, the company may stop obtaining the information from its customers. This will lead to reduction of the cost of collecting the data and the computing cost too.

- The customer's means of payment played a larger role in determining what customer will Default in the repayment of the premium or not.