Insurance Premium Default

Predictive Analytics Report

Okonkwo, Onyeka
Onyeka.okonkwo22@gmail.com

Executive Summary

Problem Statement

Premium is the major source of income for insurance companies, without which their liquidity could be threatened. Therefore, companies need to mitigate the risk of payment defaults. To do this, it becomes imperative that insurance companies are able to identify what kinds of customers are likely to default and significant factors that affect them.

Hence, the goal of this report is to study historical data of the company, garner useful insights through various visualization techniques, develop a model that predicts the likelihood of a customer defaulting, and proffer strategic recommendations the company can use to minimize default.

Methods and modelling techniques

To achieve the above, the following methods were applied:

1. Data importation and assessment.
2. Descriptive statistics and Univariate/Bivariate analysis
3. Random Forest: It's a general procedure that uses multiple models (trees, in this case) to obtain a better predictive performance.
4. Logistic Regression modelling: The idea is to find a relationship between features and probability of particular outcome.
5. Naïve Bayes modelling: It's a classification technique based on the Bayes theorem. It assumes that predictors are independent of one another i.e. that the presence of a particular feature in a class is unrelated to the presence of any other features.
6. K Nearest Neighbours (KNN): It's an algorithm that stores all available cases and classifies new cases by a majority vote of its nearest neighbours. It can be used for both regression and classification problems.
7. Comparison of model performance (for 3 – 6)
8. Application of Ensemble methods i.e. Bagging and Boosting.
9. Comparison of 8 with top performer from 7

Findings and Recommendations

In the course of this analysis and report, there were some observations of which are highlighted, and recommendations which we believe will prove useful to the company.

1. The overall default rate is 6.3%.

2. Significant variables that affect default are late payments between 3-6 months, 6-9 months and above 12 months, income and percentage of payment made in cash.

3. The Random Forest model performed best in predicting customers default.

4. Low income customers, customers sourced from channel D and those within 20-30 years of age were more likely to default.

5. Management could adjust marketing effort and focus on customers within channel A, high income earners and cash-paying customers.

6. They could also consider setting up different insurance packages more tailored to low income earners and their younger customers, this has the potential of placing premium burden within affordable limits, hence reducing the probability of defaulting.

7. The company should set-up a customer support centre that checks-in on customers and reminds them when premium is due. This system could be automated.

8. Introduction of penalty on default may also be an effective deterrent to defaulting.

9. On automation, customers could also be assigned dashboards and various payment options ranging from direct debits to bank accounts or credit cards, prepayments, or standing orders on specific dates with their respective financial institutions.

Introduction

Insurance companies generate major revenue from the customers' Premium. When these customers default, it could lead to revenue losses for the company, this is because income to run operations and pay-out Claims become insufficient. Negative cash flow is a red flag for any company, so therefore, to meet its short term obligations, Insurance companies need to generate sufficient revenue.

To mitigate against the risk of defaults which may threaten the going concern of the entity, it would be beneficial for Insurance companies to know upfront which type of customers would default in premium payments. This knowledge would help the management make operational and strategic decisions, streamlining marketing efforts towards more promising customers. It would also identify the factors that cause high default rates, equipping management to propose strategies for reducing same.

Therefore, the objective of this report is to:

1. Identify factors that cause higher default rate.
2. Build a model that can predict the likelihood of customers defaulting on premium payment.
3. Propose strategies for reducing default rate.

To achieve this, we will perform preliminary exploratory analysis on the data provided using visualization, and build a predictive model to draw more insight.

Data Preparation

The data for this report was generated from historical data of customer base of the company. A sample of 79,853 customers were randomly chosen, with personal information relating to –

- ID of customer
- Percentage premium paid by cash
- Age of customers
- Income
- Marital status
- Vehicle owned – between 1 and 3
- Late payment of premium – 3-6 months; 6-12 months; more than 12 months
- Risk Score – Higher scores are better
- Number of dependants – Between 1 and 4
- Accommodation – Owned or Rented
- Number of premiums paid till date
- Sourcing Channel – How customers were acquired
- Residence Area Type – Urban or Rural
- Total premium paid
- Default – Yes or No

There are 17 variables presented with Default being the predicted variable. That is, 17 columns and 79,853 rows in total. Each row represents a single customer.

The following actions were performed on the data:

- Search for missing variables – None was found. Dataset was complete.
- Renaming of variables for easy reference and coding:

Previous State:	**Changed To:**
Marital Status	Marital_Status
Residence_Area_Type	Residence
perc_premium_paid_by_cash_credit	Perc_paid_in_cash
Age_in_days	Age
Count_3-6_months_late	Count_3_6_months_late
Count_6-12_months_late	Count_6_12_months_late

- Changing Age in days to Age by dividing by 365 days.
- Adjusting Perc_premium_by_cash to actual percentages by multiplying by 100.
- Removing ID column as it adds no fresh insight.
- Changing character string of Residence to Binary – 1 (urban), 0 (rural)
- Changing character string of Souring Channel to Binary – A – E represented by 1 – 5 respectively.
- Converting other misclassified categorical variables from Numeric to Factor.

The end results:

- 9 Numerical variables and 7 Categorical variables.
- 17 columns shrunk to 16.
- Rows remain the same.

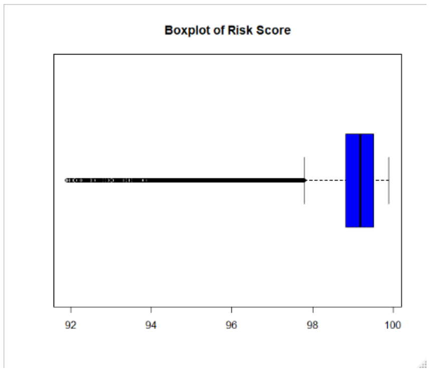
Exploratory Data Analysis

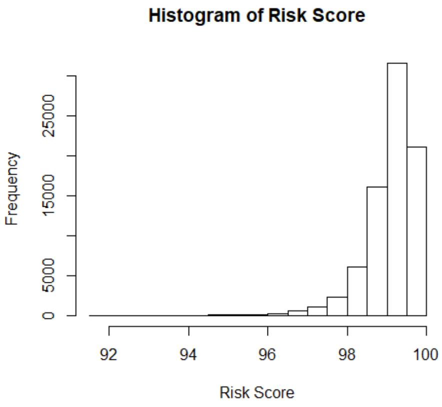
First step was to summarize the data. The following insights were garnered:

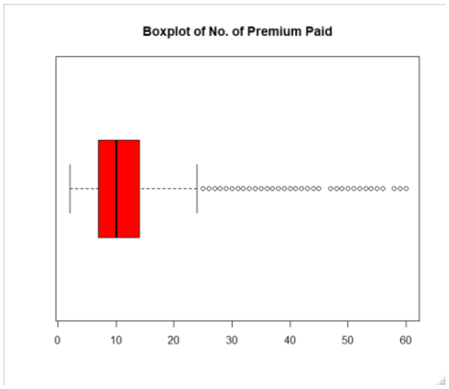
Variable	Insight
Income	Minimum is $24,030 with the max being $90,262,600. The median income is $166,560
Marital Status	50.1% of customers are unmarried, as against 49.8% married counterpart. There is no significant difference in the spread.
Number of Dependents	There’s a 25% spread among each customer with dependents ranging from 1 to 4.
Risk Score	The minimum score is 91.9, with a maximum of 99.89. The median score is also 99. If this is on a scale of 100, then the customers are credit worthy.
Number of premiums paid	Minimum number is 2, median is 10, while the maximum is 60 times. Sort out customers who pay more premium as they are outliers.
Sourcing channel	More sourced from channel A/1 i.e. 54% more than any other channel. Channel E/5 records only 0.76% of customers.
Residence	About 39% of customers are in the Rural area, as against 61% from urban area. Does this spread have any impact on the Default and premium paid?
Default	There’s a 6.3% default rate. Data is imbalanced and will be treated before model building.
Age	Minimum age is 21, maximum age of customers is 103, and the median age is 51 years. It appears there are more upper middle-aged customers.
Accommodation	About 49% live in rented apartments and 51% own their homes.
Vehicle owned	33% of customers own between 1-3 cars across board. No significant difference in spread.

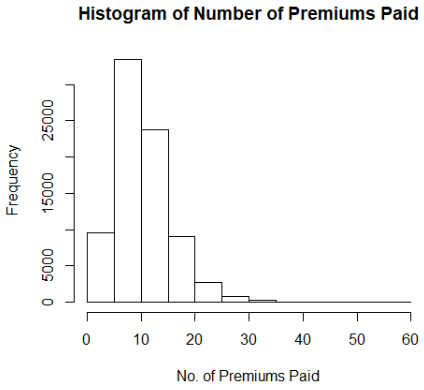
Data Visualization – Univariate and Bivariate Analysis

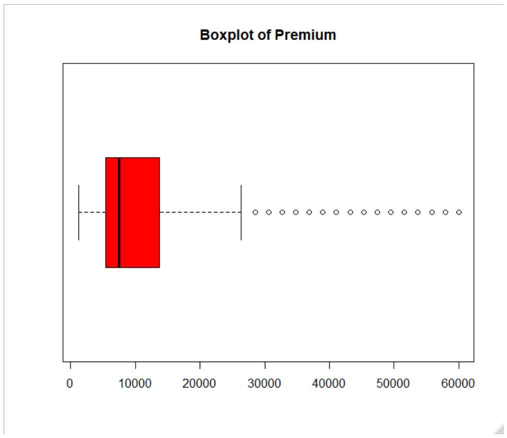
Furthermore the data was checked to ascertain its distribution. For this purpose the use of histograms and boxplots were employed to see the distribution of the data and check for the existence of outliers, while tables were used to check for counts of categorical variables.

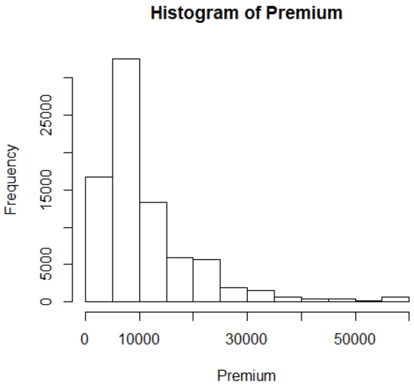


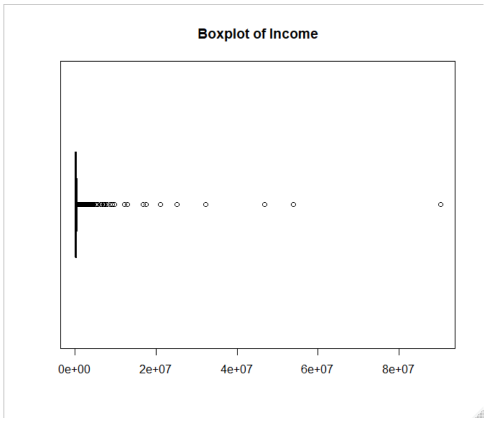


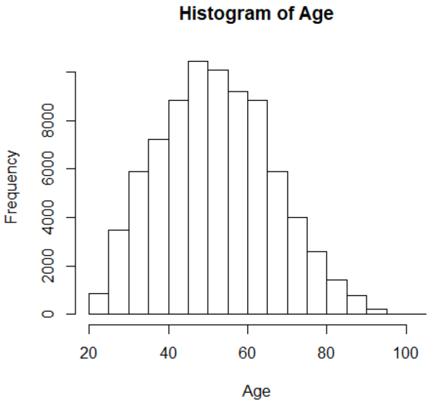






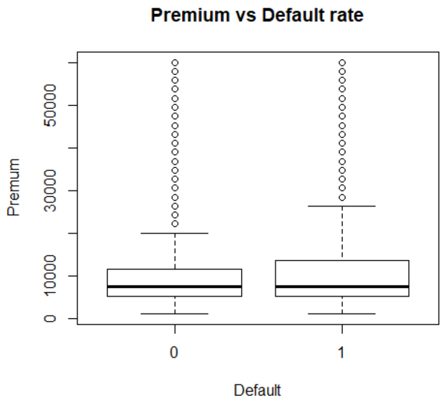


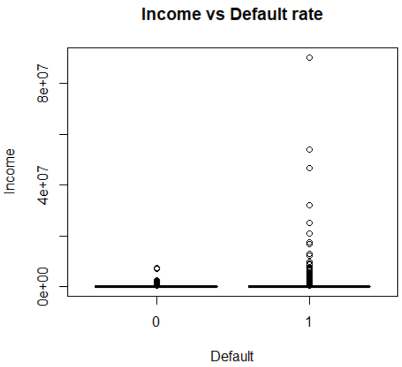


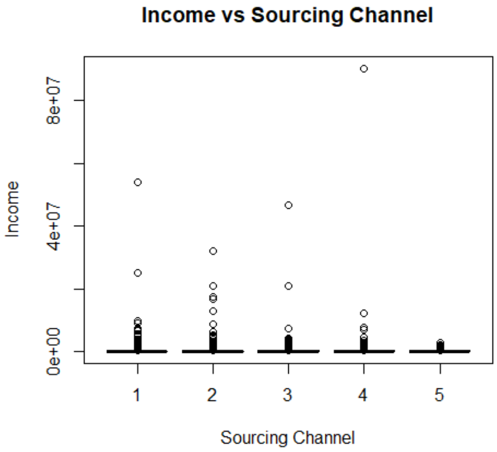


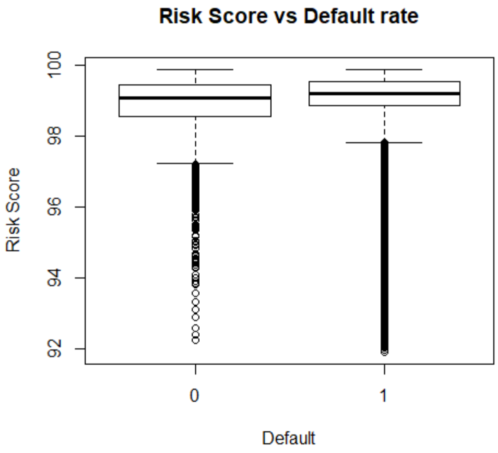
Variables	Insight
Income	• Outlier present • Majority of data fall on the left • Median income is around $100,000
Risk Score	• Outliers present • Majority of data is on the right with risk scores between 91 and 97 • The median risk score is around 99 and a maximum of 100.
Premium paid	• Outliers exist in data • Most information fall to the left • Highest count is around $10,000 • Outliers between $30,000 and $60,000
No. of premium paid	• Outliers exist in data • Histogram is skewed to the right • Outliers falls between 25 and 60
Age	• Normally distributed • Maximum age is 103 and minimum is 21. • Median age is 50. • Majority of customers are aged between 40 and 60 years.

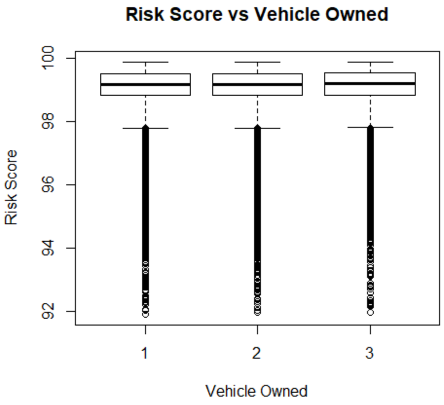
Bivariate Analysis







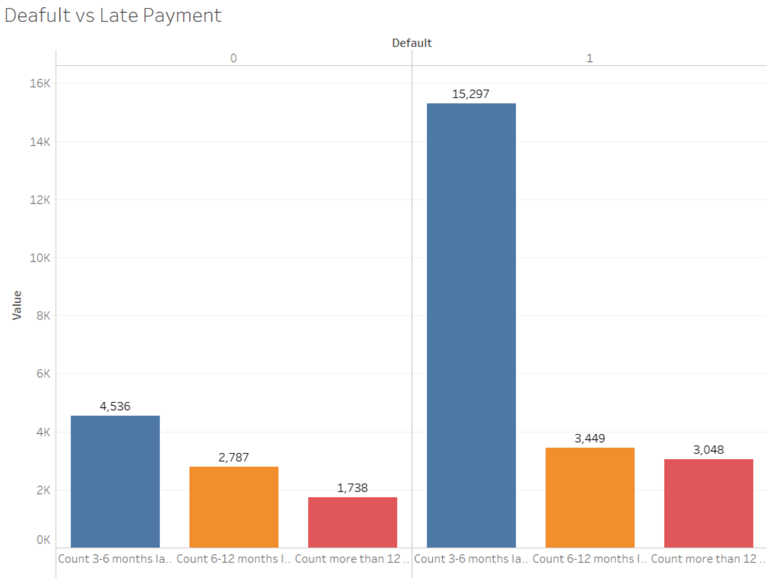




Variables	Insight
Premium vs Default	No significant difference. Median hovers just below $1000 for both defaulters and non-defaulters.
Income vs Default	No significant difference, although there are more outliers within non-defaulters.
Risk Score vs Vehicle owned	Median for all 3 is around 98/99. Data spread is similar.
Risk score vs Default	Median for non-defaulters is 99, and 98 for Defaulters.
Income vs Sourcing Channel	Groups 1-4 have significant outliers, especially group 4 with over $90m. Median falls within the same range of around $100,000. No significant difference beyond these.

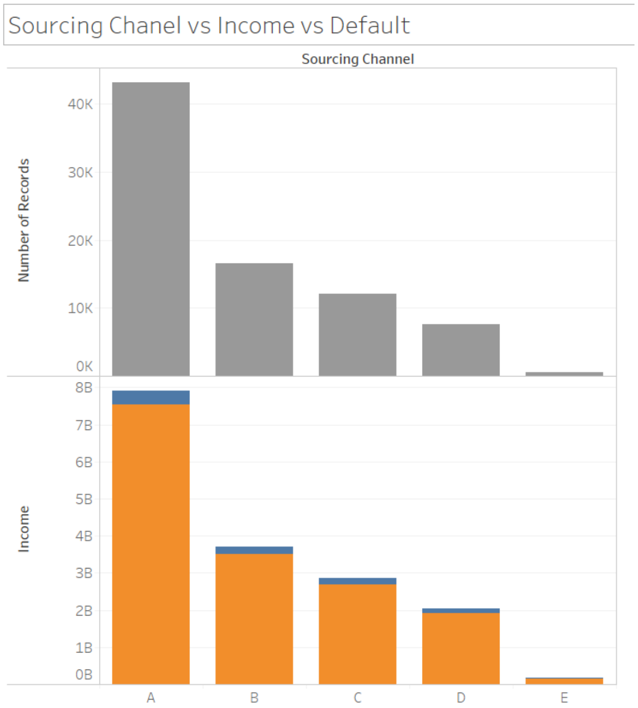
Further Graphical Visualizations

Default vs Late Payment Duration



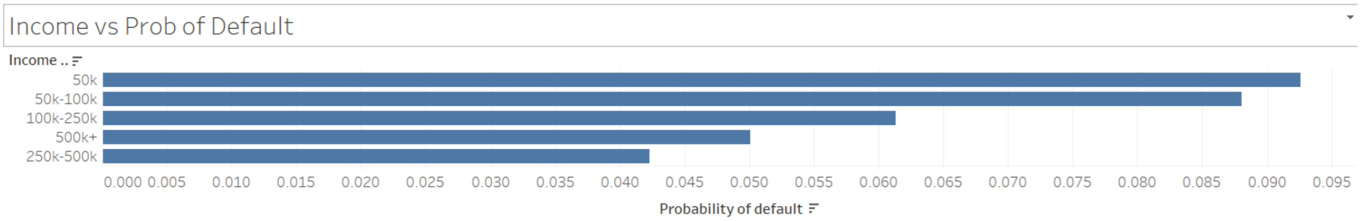
The graph above shows customers who have defaulted in payment between 3-6 months rank higher than others. They also have the highest default and non-default rate, although majority are non-defaulters.

Sourcing Channel vs Income vs Default



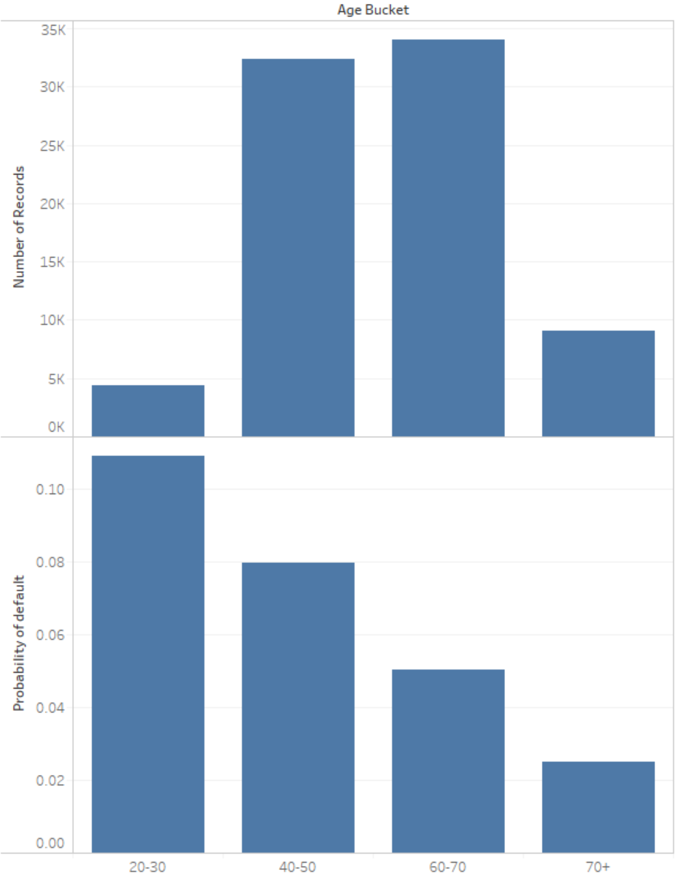
Majority of customers, as seen in the graph are sourced from channel A and this group also have the highest income earners when compared to channels B to E.

Income vs Probability of Default



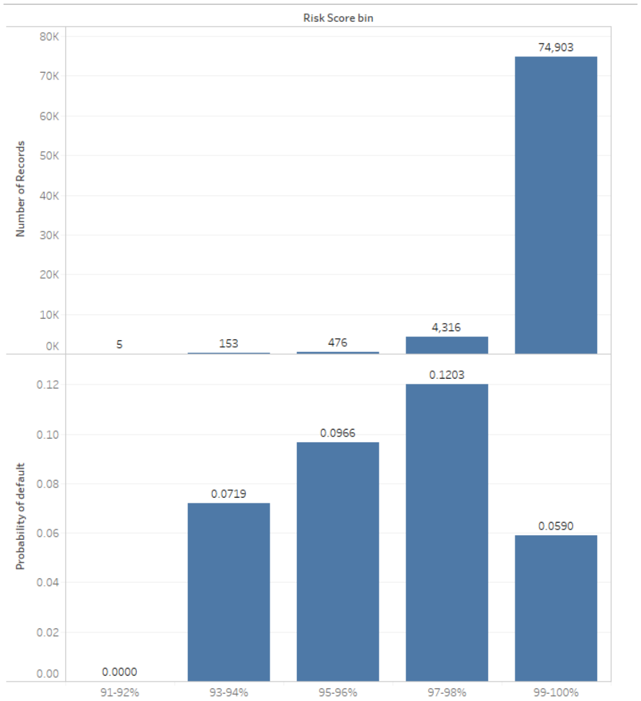
In the income and probability of default chart above, it is shown that customers with income within 50,000 have the highest likelihood to default, while customers between 250k-500k income bins have the lowest likelihood of default. It's interesting to see that customers who earn above 500,000 (i.e. 500k+) have a higher chance of default than the latter.

Age of Customers vs Probability of Default



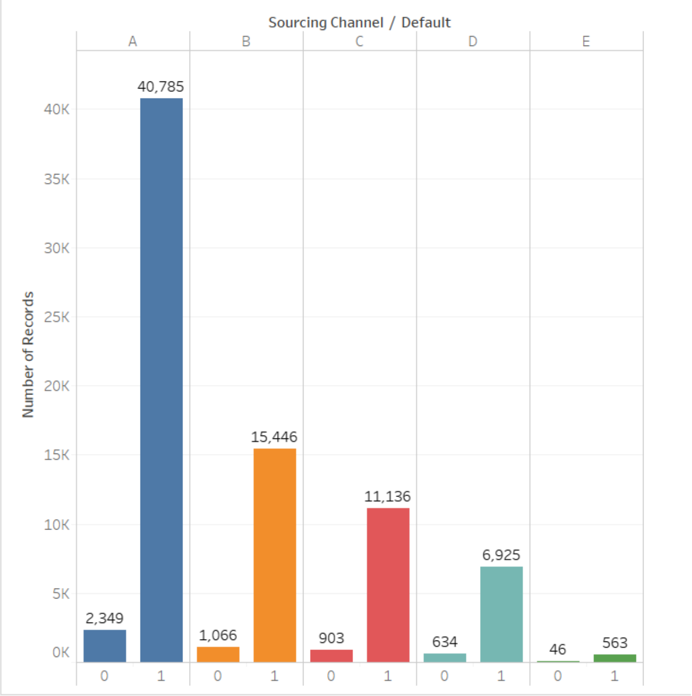
Customers between ages 20-30 years have the highest probability of default (10.8%) even though they have the lowest number of records. While customers aged 60-70 years have the highest number of records with 5% likelihood of default. It would be safe for the company to look into customers aged 40-50 years, since about 32,000 customers within the sample come from this group and the have the second highest likelihood to default in premium payment (7.9%).

Risk Score vs Probability of Default



The chart shows that customers with a risk score between 97-98% have a higher probability of default at 12%, compared to the other bins. This is irrespective of the fact that there are more customers within 99-100% risk score and yet the probability of default is the second lowest at 5%.

Sourcing Channel vs Default



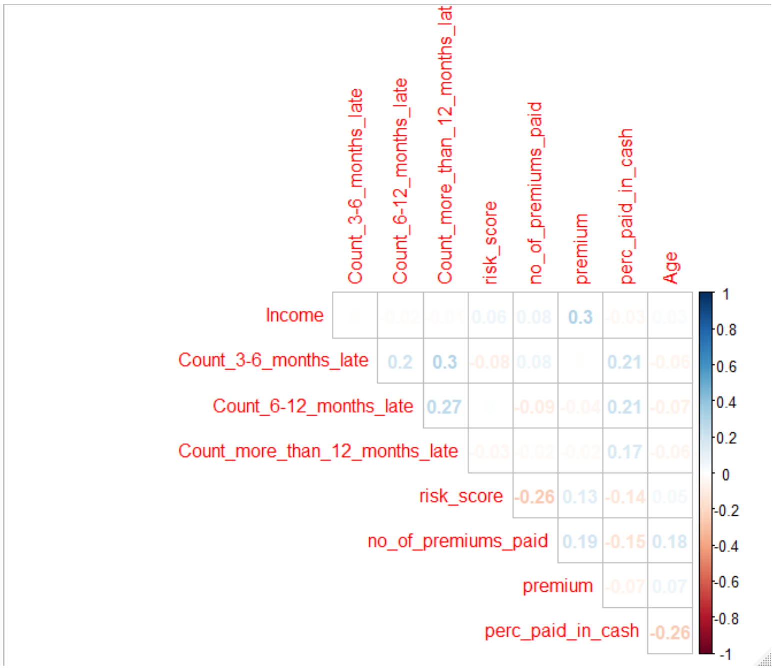
Default rate per Channel:

- A – 5.4% default rate
- B – 6.46% default rate
- C – 7.5% default rate
- D – 8.39% default rate
- E – 7.55% default rate

Channel D has the highest default rate even though customers within this bracket have the second lowest number. Next is channel E and then C. Channel A has the lowest default rate.

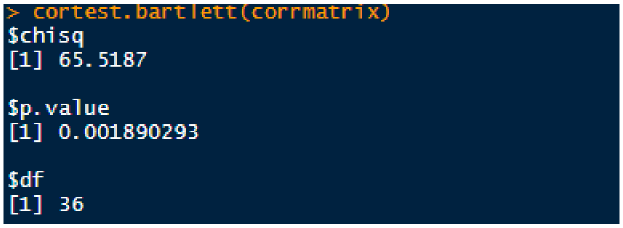
Test for Collinearity

We tested for collinearity between the independent variables and found none.



Among the variables, there are no related factors. Correlation coefficients fall under 0.2.

Using Bartlett Test



```
> cortest.bartlett(corrmatrix)
$chisq
[1] 65.5187

$p.value
[1] 0.001890293

$df
[1] 36
```

In a further test for correlation, we put forth the hypothesis:
H0: Predictors are not correlated
H1: Predictors are correlated

Since our P-value is less than 0.05, we are 95% certain to reject the null hypothesis and conclude there is no correlation among predictors.

Summary of Findings

Variables	Insight
Late Payment vs Default	Customers who have defaulted in payment between 3-6 months rank higher than others.
No. of dependent vs Sourcing channel	Customers from channel A/1 have more dependents.
Default vs Sourcing Channel	• A/1 – 5.4% default rate • B/2 – 6.46% default rate • C/3 – 7.5% default rate • D/4 – 8.39% default rate • E/5 – 7.56% default rate Channel D has the highest default rate even though customers within this bracket have the second lowest number. Next in line is channel E and then C. • Channel A has the highest customer representation (54%) and lowest default rate.
Default vs Residence	Default rate between urban and rural dwellers both hover around 6%. Residence does not affect possibility of default.
Risk Score vs Default	Customers with risk score at 97%-98% have the highest chance of default at 12%.
Age vs Default	• Customers between ages 20-30 years have the highest probability of default (10.8%) even though they have the lowest number of records. • The majority of customers fall between 40 and 60 years of age.
Income vs Sourcing channel vs Default	• Customers from channel A earn the highest income. • Customers with income around 50,000 are more likely to default. • Those with income above 250,000 are least likely to default.

- There is no correlation between the various variables although this could change after transformation.
- Outliers are present in most variables and will be treated.
- The total default rate within our sample is 6.3%.

Data Pre-processing

Treatment of Outliers

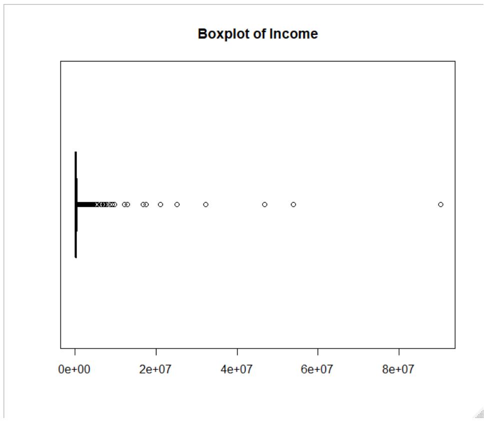
There were a number of variables with Outliers in the dataset. To treat for these, the variables were capped at various percentile largely based on judgement.

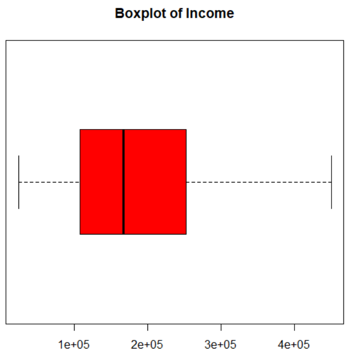
Variables treated are: Income, Age and Premium.

Boxplots are also included showing comparative spread of data among the variable before and after outlier treatment, and from the output it is evident that outliers were eliminated.

Income – Upper limit capped at the 95th percentile (450,050).

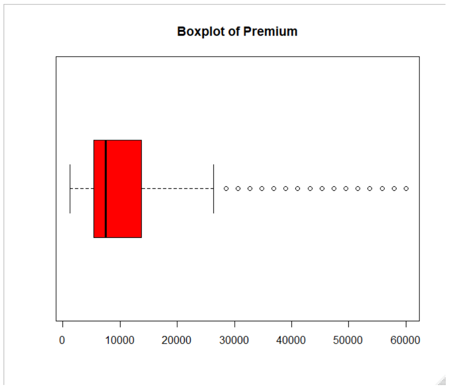
Before *After*

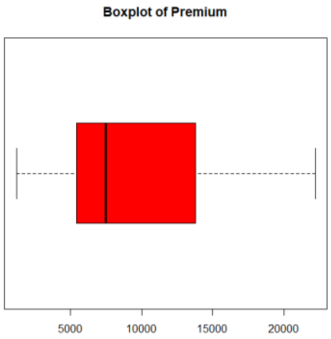




Premium – Upper limit capped at the 90th percentile (22,200)

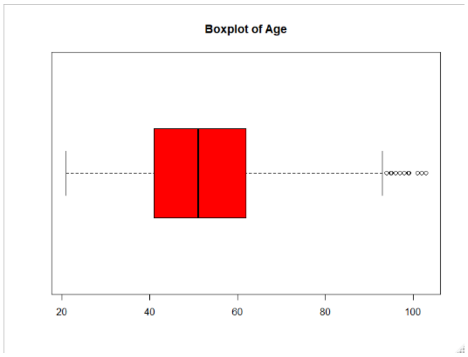
Before *After*

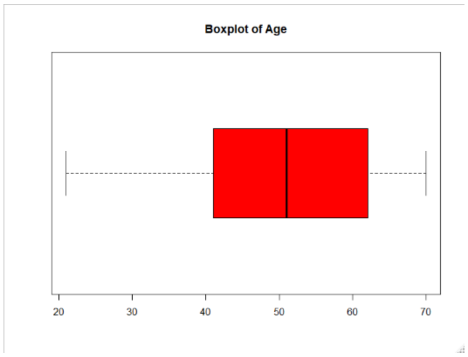




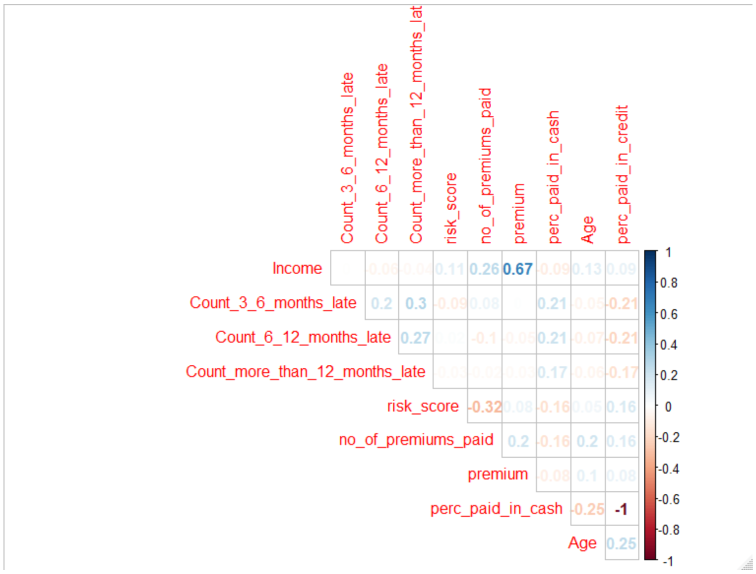
Age – Upper limit capped at 90th percentile (70)

Before *After*





<u>Using Correlation to check for relationship among variables after outlier treatments</u>



From our correlation matrix, there is now a positive correlation between Income and total premium paid at 0.67. And a perfect negative correlation between Percentage paid in cash and percentage paid via credit (the new variable created).

In light of this, Premium and Percentage paid in credit columns were removed from the data set, alongside ID column. This is to prevent any distortion in accuracy of models.

Analytical Approach

Modelling

The dataset is imbalanced – 6.3% default rate. To deal with this we can apply ensemble method like SMOTE, and try to use this to minimize variance in the dataset and bias in our predictive output.

Models to be applied are:

- Random Forest
- Logistic regression
- Naïve Bayes
- KNN

And ensemble techniques to tune the model will be:

- Bootstrap Aggregating
- Boosting

Evaluating the model

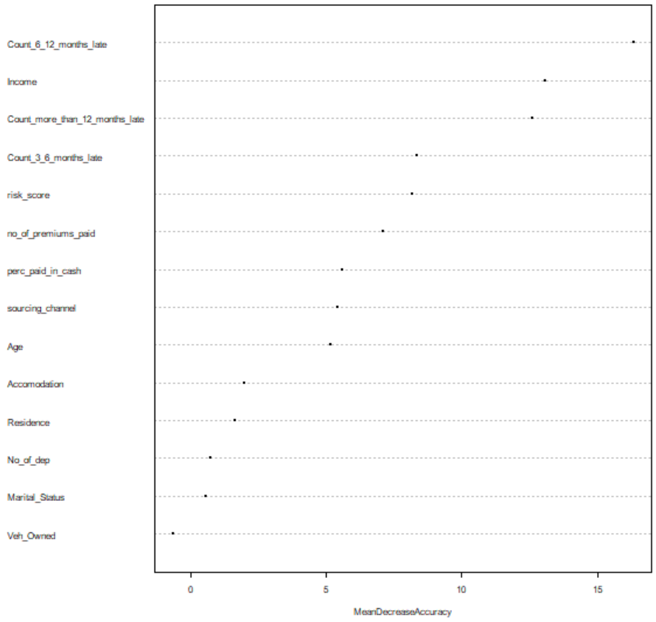
To evaluate the performance of our model, based on the business case, we will check for accuracy of prediction, specificity and sensitivity of our model in correctly identifying positives and labelling them.

This evaluation criteria will be used across all our various predictive models and then compared to see which performs best.

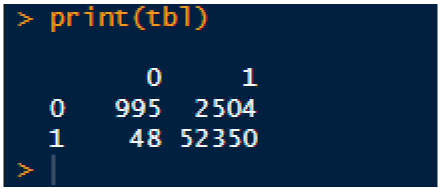
Random Forest

Random Forest is a versatile model known to manage unbalanced data well. It can also handle both categorical and classification problems. RF was initially run and then tuned to minimize the OOB error rate.

Important variables from model - Count 6-12 months, Income, Count more than 12 months, Count 3-6 months.



<u>Evaluation on Train Data</u>



```
> print(tbl)

        0     1
  0   995  2504
  1    48 52350
>
```

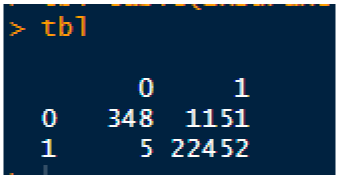
Accuracy – 95.4%.

Specificity – 95.4%

Sensitivity – 95.4%

All applied metric give similar output. Model is stable and performs well in correctly classifying the data and accuracy of prediction.

Model evaluation (on test data)



```
> tbl

        0     1
  0   348  1151
  1     5 22452
```

Accuracy – 95.2%. This means there's a 4.8% error rate in prediction, and our model succeeds significantly in predicting default rate.

Specificity (TNR) – 0.99. At 98.58% the model performs well in classifying customers who are likely to default in payment.

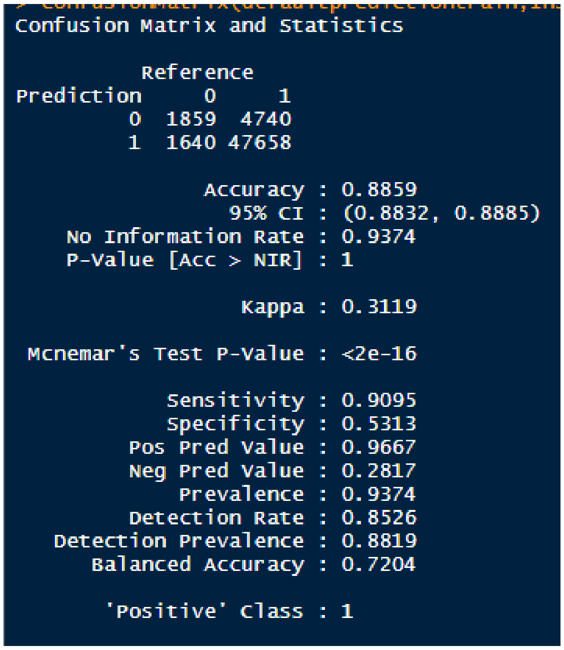
Sensitivity – (TPR) – 0.95. The model also does well in predicting the percentage of customers who will not default in payment of premium.

It can also be seen that based on the results of both the Train and Test set, our data is stable and nether overfit or underfit.

Logistic Regression

The idea of logistic regression is to find a relationship between features and probability of particular outcome, in this case, predicting customers default rate. Before building our logistic regression model, imbalance of data was corrected using Smote.
Important variable per model – Percentage paid in cash, Count 6-12 months, Count 3-6 months, Count more than 12 months.

Evaluation on Train Data



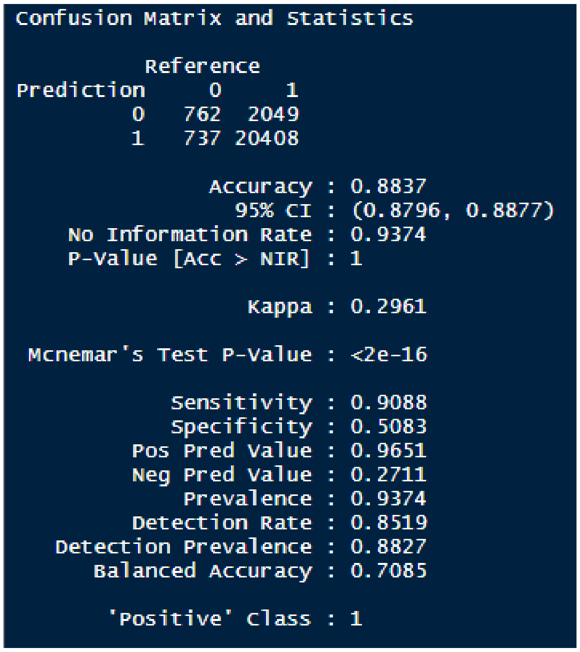
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0  1859  4740
         1  1640 47658
                                          
               Accuracy : 0.8859          
                 95% CI : (0.8832, 0.8885)
    No Information Rate : 0.9374          
    P-Value [Acc > NIR] : 1               
                                          
                  Kappa : 0.3119          
                                          
 Mcnemar's Test P-Value : <2e-16          
                                          
            Sensitivity : 0.9095          
            Specificity : 0.5313          
         Pos Pred Value : 0.9667          
         Neg Pred Value : 0.2817          
             Prevalence : 0.9374          
         Detection Rate : 0.8526          
   Detection Prevalence : 0.8819          
      Balanced Accuracy : 0.7204          
                                          
       'Positive' Class : 1               
```

Accuracy – 88.6%

Specificity – 0.53 i.e. ~ 53%

Sensitivity – 0.91 i.e. ~ 91%

Evaluation on Test Data



```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0   762  2049
         1   737 20408

               Accuracy : 0.8837
                 95% CI : (0.8796, 0.8877)
    No Information Rate : 0.9374
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2961

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

            Sensitivity : 0.9088
            Specificity : 0.5083
         Pos Pred Value : 0.9651
         Neg Pred Value : 0.2711
             Prevalence : 0.9374
         Detection Rate : 0.8519
   Detection Prevalence : 0.8827
      Balanced Accuracy : 0.7085

       'Positive' Class : 1
```

Accuracy – 88.3%. This means there's an 11.7% error rate in prediction, and our model succeeds significantly in predicting default rate.

Specificity (TNR) – 0.51 i.e. ~ 51%. The model is an average performer in predicting the likelihood of default premium payment.

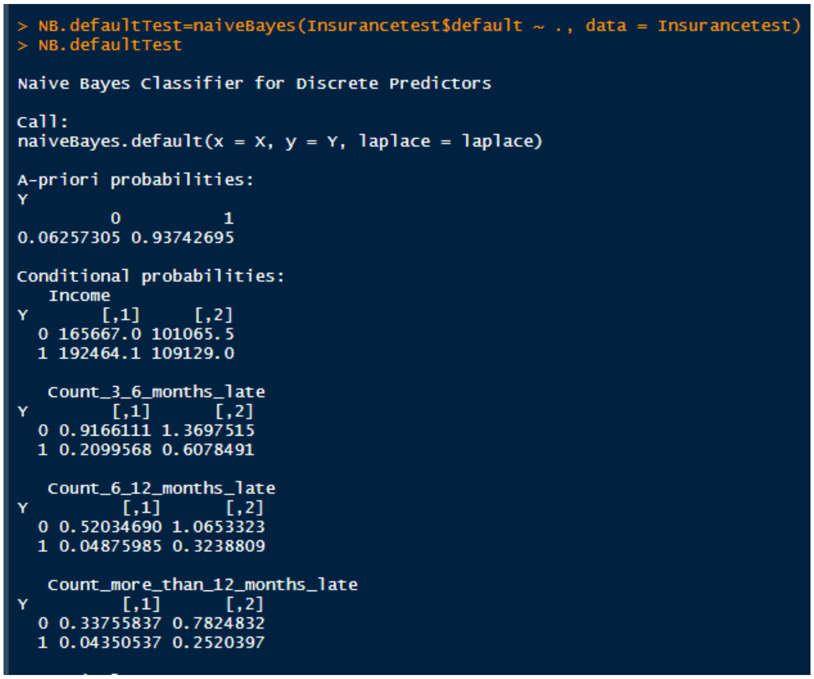
Sensitivity – (TPR) – 0.91 i.e. ~ 91%. As for predicting the probability that customers will not default, it does better with only a 9% error rate.

Note: Without adjusting for imbalance in data, performance of the model was greatly increased. However, because imbalance in data could skew the results and lead to overfitting the model, Smote was applied to adjust data.

From the results of both the Train and Test set, the minor difference in output indicates our model is stable.

Naïve Bayes

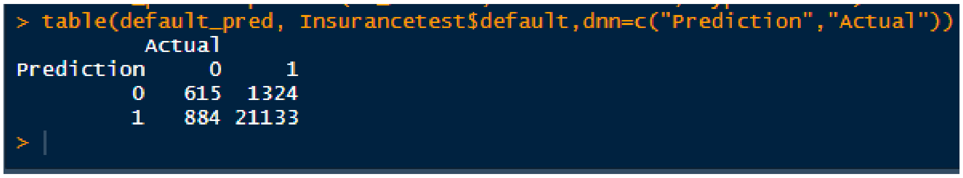
Naïve Bayes is a classification technique based on the Bayes theorem. It assumes that predictors are independent of one another i.e. that the presence of a particular feature in a class is unrelated to the presence of any other features.



```
> NB.defaultTest=naiveBayes(Insurancetest$default ~ ., data = Insurancetest)
> NB.defaultTest

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
         0          1 
0.06257305 0.93742695 

Conditional probabilities:
   Income
Y       [,1]     [,2]
  0 165667.0 101065.5
  1 192464.1 109129.0

   Count_3_6_months_late
Y        [,1]      [,2]
  0 0.9166111 1.3697515
  1 0.2099568 0.6078491

   Count_6_12_months_late
Y         [,1]      [,2]
  0 0.52034690 1.0653323
  1 0.04875985 0.3238809

   Count_more_than_12_months_late
Y         [,1]      [,2]
  0 0.33755837 0.7824832
  1 0.04350537 0.2520397
```

The above result shows that customers who have late payment between 3 – 6 months have a 92% likelihood of defaulting, while those who have been 6 – 12 months late have a 52% of defaulting.

Model evaluation (on test data)



```
> table(default_pred, Insurancetest$default,dnn=c("Prediction","Actual"))
          Actual
Prediction     0     1
         0   615  1324
         1   884 21133
> 
```

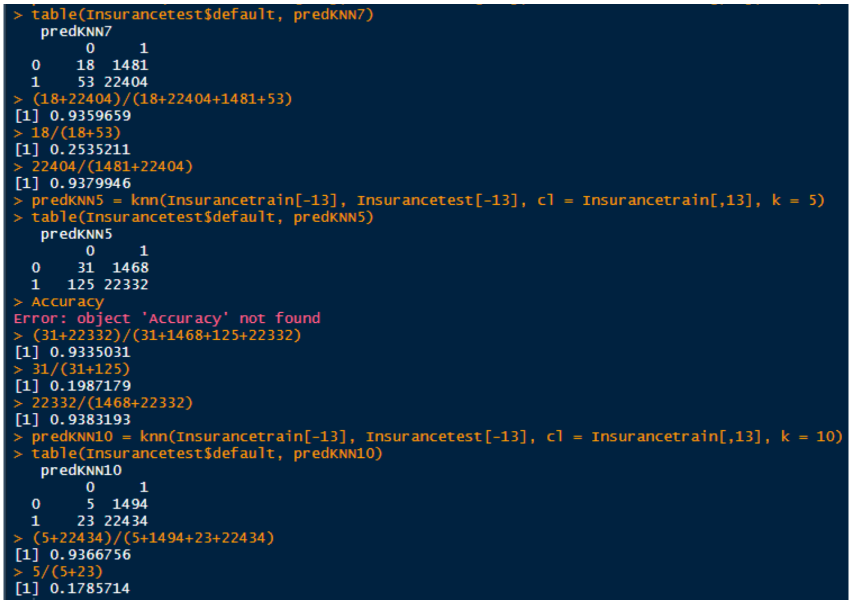
Accuracy – 90.8%. That is a 9.2% error rate. The model is a good performer.

Specificity (TNR) – 0.41. It doesn’t do so well in correctly classifying the defaulters. There is a 59% error.

Sensitivity – (TPR) – 0.94. The model does well in classifying non-defaulters with a5.9% error rate.

K Nearest Neighbors (KNN)

KNN is an algorithm that stores all available cases and classifies new cases by a majority vote of its nearest neighbors. It can be used for both regression and classification problems.



```
> table(Insurancetest$default, predKNN7)
   predKNN7
        0     1
  0    18  1481
  1    53 22404
> (18+22404)/(18+22404+1481+53)
[1] 0.9359659
> 18/(18+53)
[1] 0.2535211
> 22404/(1481+22404)
[1] 0.9379946
> predKNN5 = knn(Insurancetrain[-13], Insurancetest[-13], cl = Insurancetrain[,13], k = 5)
> table(Insurancetest$default, predKNN5)
   predKNN5
        0     1
  0    31  1468
  1   125 22332
> Accuracy
Error: object 'Accuracy' not found
> (31+22332)/(31+1468+125+22332)
[1] 0.9335031
> 31/(31+125)
[1] 0.1987179
> 22332/(1468+22332)
[1] 0.9383193
> predKNN10 = knn(Insurancetrain[-13], Insurancetest[-13], cl = Insurancetrain[,13], k = 10)
> table(Insurancetest$default, predKNN10)
   predKNN10
        0     1
  0     5  1494
  1    23 22434
> (5+22434)/(5+1494+23+22434)
[1] 0.9366756
> 5/(5+23)
[1] 0.1785714
```

After trying different levels of K, the optimal was picked at 7 because it gives the best value for specificity without compromising on overall accuracy and model fit.

Accuracy – 93.6%. Records a 6.4% error in correctly classifying defaulters and non-defaulters.

Specificity (TNR) – 0.25. Model performs poorly in picking the defaulters with a 75% error rate.

Sensitivity (TPR) – 0.94. Models does well in picking the non-defaulters with a 6.2% error.

Model Evaluation Metrics (as used in data) and Performance Judgement of Random Forest, Logistic Regression, Naïve Bayes and KNN

Confusion Matrix: We use this to test for the accuracy of the predictive powers of our model. The higher the accuracy percentage, the better the performance of the model.
Specificity: This is the proportion of negative results out of the number of samples which were actually positive.

Metrics	Random Forest	Logistic Regression	Naives Bayes	KNN
Accuracy	95%	88%	91%	94%
Specificity	0.99	0.51	0.41	0.25
Sensitivity	0.95	0.91	0.94	0.94

We see that Random Forest surpasses the other models in both accuracy test and TNR, with Logistic regression being the worst performer in accuracy test and KNN in specificity test.

Tuning the Model – Ensemble Methods

For this the use ensemble methods is applied which deals with the imbalance by training multiple models using the same algorithm.

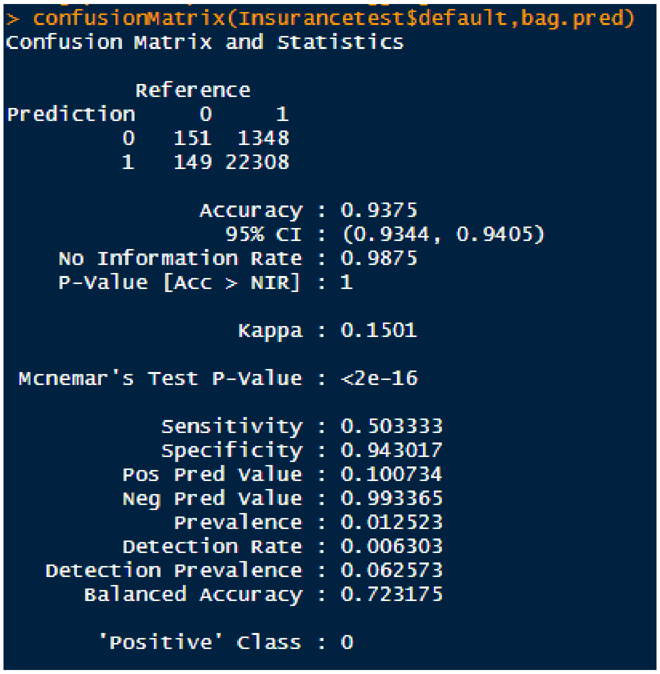
Hence:

- Bagging (Bootstrap aggregation): reduces the variance but retains some of the bias because we still sample the same learners.
- Boosting: helps reduce both bias and variance.

After building the models their performance is compared to the Random Forest model which was the best performer among the four used earlier.

Bagging or Bootstrap Aggregating

Bagging helps improve the performance of simple models and reducing overfitting of more complex models.



```
> confusionMatrix(Insurancetest$default,bag.pred)
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0   151  1348
         1   149 22308

               Accuracy : 0.9375
                 95% CI : (0.9344, 0.9405)
    No Information Rate : 0.9875
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1501

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

            Sensitivity : 0.503333
            Specificity : 0.943017
         Pos Pred Value : 0.100734
         Neg Pred Value : 0.993365
             Prevalence : 0.012523
         Detection Rate : 0.006303
   Detection Prevalence : 0.062573
      Balanced Accuracy : 0.723175

       'Positive' Class : 0
```

Accuracy – 93.8%. That's a 6.2% error rate in rightly predicting the defaulter and non-defaulters.

Specificity – 0.50. Model's performance is average. 50% of the time, it correctly classifies defaulters.

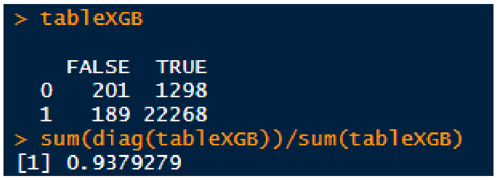
Sensitivity – 0.94. 94% of the time the model classifies the non-defaulters appropriately.

Boosting – XGBoost

The aim is to minimize the errors of the previous models. Boosting reduces both bias and variance.

Important variables - Percentage paid in cash, Count 6-12 months, Count more than 12 months, Count 3-6 months.

Model Evaluation – on test data



```
> tablexGB

     FALSE  TRUE
  0    201  1298
  1    189 22268
> sum(diag(tablexGB))/sum(tablexGB)
[1] 0.9379279
```

Accuracy – 93.8%. That’s a 6.2% error rate in rightly predicting the defaulter and non-defaulters.

Specificity – 0.52. Model’s performance is average. 52% of the time, it correctly classifies defaulters.

Sensitivity – 0.94. 94% of the time the model classifies the non-defaulters appropriately.

Precision – 94.5%

Comparison of Best Performer

Metrics	Random Forest	Bagging	Boosting
Accuracy	95%	94%	94%
Specificity	0.99	0.50	0.52
Sensitivity	0.95	0.94	0.94

Random Forest gives better predictive output than both Bagging and Boosting techniques, which emphasizes the versatility of the model to handle many kinds of data well.

Business Insights and Recommendation

The company sought to understand the probability of customers defaulting in premium payment, predict which customers will default, while also identifying significant variables that affect this. A sample of 79,853 customers were provided of which approximately 6.3% default. The dataset also consisted of information relating to customers personal details e.g. income, risk score, age, etc.

Data was explored and transformed before further analysis were carried out. From findings we see a relationship between income of customers and the premium paid. It was also discovered through modelling that significant variables that affect the probability of customers defaulting were to a great extent dependent on whether they had any late payments between 3-6 months, 6-9 months and above 12 months. Income and percentage of payment made in cash were also significant factors that affect default.

Furthermore, among customers who had a late payment record, those who fell within the 3-6 months bracket have a higher probability of defaulting.

Relating to income, customers who earned between $50,000 and $100,000 have a 93% chance of defaulting in payment. Therefore, more high income earners ($250,000+) are less likely to default in premium payment.

While the channels from where customers were sourced did not rank as an important predictor, it was found that those from channel D had the highest default rate, while customers from channel A had the lowest. Likewise with Age. Customers between 20 – 30 years were 11% more likely to default than older customers. In fact, the higher the age, the lower the likelihood of default. Perhaps there’s a relationship between the Sourcing Channels and Age of customers. Customers aged 40 – 50 years make up the second largest group (after 60 – 70years) with an 8% likelihood of default. The company should therefore pay closer attention to this group of customers, put measures in place like check-in phone calls and reminders to ensure they don’t default in payment.

Risk Score of customers should also be given some attention. Although customers with high Risk Score show a proclivity to repaying debts, we find that those who pay in cash are less likely to default. The management should consider sourcing customers who are high income earners and are more likely to pay the premium in cash. This will keep the company more liquid.

Other measures the company may take to reduce the default rate includes setting up an option for customers to place standing orders on their bank accounts and credit cards to automatically deduct premium payments when due. A customer service desk should also be set up to regularly check-in with customers, setting up reminders when payment is due. Introduction of fees on late payment could also be introduced, this will deter customers from defaulting in payment. The company should focus marketing efforts on high income, cash paying customers, and could consider introducing a different (perhaps bundled) insurance package for low income and younger customers. This could potentially place their premium payment within an affordable limit and hence reduce the possibility of late payments and subsequent default.