NUS

National University of Singapore

BT2103 AY 22/23 SEM 2 GROUP 43

Steffi Lim En Qi A0238209A

Yong Duan Kang A0240563L

Adele Ng A0239037A

Introduction

Background

Credit risk analysis is a crucial process in the financial industry that involves assessing the likelihood of a borrower defaulting on their loan. Inaccurate credit risk assessments can result in losses, and bankruptcy as evident during the global financial crisis of 2008. Machine learning and big data analysis has emerged as powerful tools for analysing credit risk, enabling lenders to make more informed lending decisions.

Problem statement

The objective is to identify an accurate predictive model that can effectively evaluate the likelihood of loan defaults by new customers. This will enable lenders worldwide to make well-informed decisions in customer selection and loaning, thereby minimizing potential losses.

Dataset Details

The dataset is a collection of payment information of 30,000 credit card holders from a bank in Taiwan. The data consists of information such as default payments, the demographic of the credit card holders, payment history and bill statements from April 2005 to September 2005. The breakdown of the 25 feature attributes is as follows:

<u>Information of the customers:</u>

ID: ID of each customer

X1 LIMIT_BAL: Amount of given credit in NT dollar for both the individual customer and their family

X2 SEX: Gender (1 = male; 2 = female)

X3 EDUCATION: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others, 5 = unknown, 6 = unknown)

X4 MARRIAGE: Marital status (1 = married; 2 = single; 3 = others)

X5 AGE: Age of the customer

Information on the delay of past payment (referred to a specific month):

X6 PAY_0: Repayment status in September 2005 (-2 = no consumption, -1 = pay duly, 0 = use of revolving credit, 1 = payment delay for one month, 2 = payment delay for two months, ..., 8 = payment delay for eight months, 9 = payment delay for nine months and above)

X7 PAY_2: Repayment status in August 2005 (same scale as above)

X8 PAY_3: Repayment status in July 2005 (same scale as above)

X9 PAY_4: Repayment status in June 2005 (same scale as above)

X10 PAY_5: Repayment status in May 2005 (same scale as above)

X11 PAY_6: Repayment status in April 2005 (same scale as above)

Information on the amount of bill statement (a monthly report that credit card companies issue to credit card holders each month):

X12 BILL_AMT1: Amount of bill statement in September 2005 in NT dollar

X13 BILL_AMT2: Amount of bill statement in August 2005 in NT dollar

X14 BILL_AMT3: Amount of bill statement in July 2005 in NT dollar

X15 BILL_AMT4: Amount of bill statement in June 2005 in NT dollar

X16 BILL_AMT5: Amount of bill statement in May 2005 in NT dollar

X17 BILL_AMT6: Amount of bill statement in April 2005 in NT dollar

Information on the amount of previous statement:

X18 PAY_AMT1: Amount of previous statement in September 2005 in NT dollar

X19 PAY_AMT2: Amount of previous statement in August 2005 in NT dollar

X20 PAY_AMT3: Amount of previous statement in July 2005 in NT dollar

X21 PAY_AMT4: Amount of previous statement in June 2005 in NT dollar

X22 PAY_AMT5: Amount of previous statement in May 2005 in NT dollar

X23 PAY_AMT6: Amount of previous statement in April 2005 in NT dollar

Variable to be predicted:

Y default.payment.next.month: Default payment (0 = no, 1 = yes)

Discrepancies in **X6 - X11** column values:

Looking at columns **X6 - X11**, values of 0 and -2 are in the dataset but are not included in the description. We found that out of 30,000 data points, 14,737 contains 0 while 2759 contains -2.

After researching online, we found that this discrepancy has been addressed by the professor who created the dataset. It states that **0** represents the use of revolving credit while **-2** represents no consumption.

Source:
https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset/discussion/34608

Data Pre-Processing

Checking column names

In the given dataset, we observed that there is inconsistent naming of the column names. Hence, we renamed **X6 PAY_0** to **X6 PAY_1** and **default.payment.next.month** to **DEFAULT** for a more systematic naming convention.

Checking for column type

We call the function **sapply(credit, class)** to check what kind of data is represented in each column. All the column has a data type of “integer”.

Checking for missing values

Using **is.na(credit)**, we found that there are no missing values from either of the columns.

Checking for duplicate records

First, we checked for duplicated data through the customer’s ID using the function **duplicated(credit.dup)** and we found that there are 35 duplicated records. After removing the duplicated records, there are now 29,965 data points.

Identifying outliers

We plotted a boxplot to identify any outliers. There is a point on the graph which shows that the **LIMIT_BALANCE** of a customer is 1,000,000. We choose to include this point in our dataset as we observed that this customer has been paying on time or using his revolving credit. He did not default either.

Checking for categorical data

We identify all columns that contain categorical data and check that they do not contain discrepancies.

X2 SEX is consistent.

For **X3 EDUCATION,** it was observed in the original dataset that values 0, 5, and 6 were not labeled. As a result, we made the decision to group these values under the category "Others" which is represented by value 4.

For **X4 MARRIAGE,** we observed entries of value 0. As a result, we group this value under the category “Others” which is represented by value 3.

For columns which represent the information on the delay of past payment (referred to a specific month), **X6 PAY_0, X7 PAY_2, X8 PAY_3, X9 PAY_4, X10 PAY_5** and **X11 PAY_6,** as mentioned above, there are some discrepancies with the data. It is observed in the original dataset that values of 0 and -2 are observed and not labelled. We also found that there are 14,737 entry points contains 0 while 2759 contains -2. As there are a significant number of data points, we did not remove them.

Checking numerical columns

X5 AGE is a column which contains numerical data, age of the customers. Hence we created separate bin with intervals of 5. Our bins are as follows: 0: [20, 25), 1: [25, 30), 2: [30, 35), 3: [35, 40), 4: [40, 45), 5:[45: 50), 6:[50: 55), 7:[55: 60), 8:[60: 65), 9:[65: 70), 10:[70: inf).

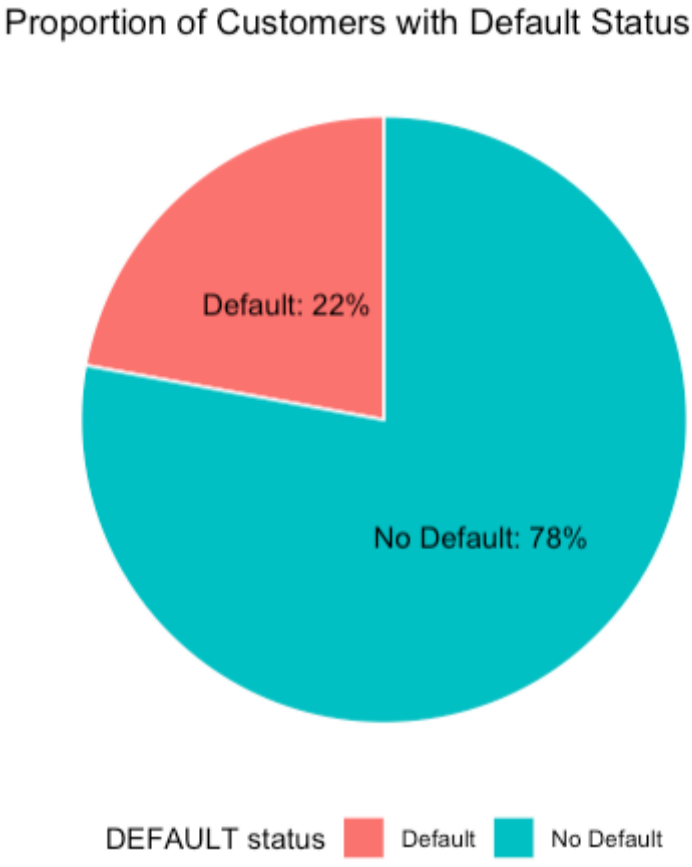
Exploratory Data Analysis

Upon conducting initial investigations on the dataset, we discovered some patterns with the customers' demographics and the default payments. The data was largely clean and there were few missing values.

Our findings are as such:

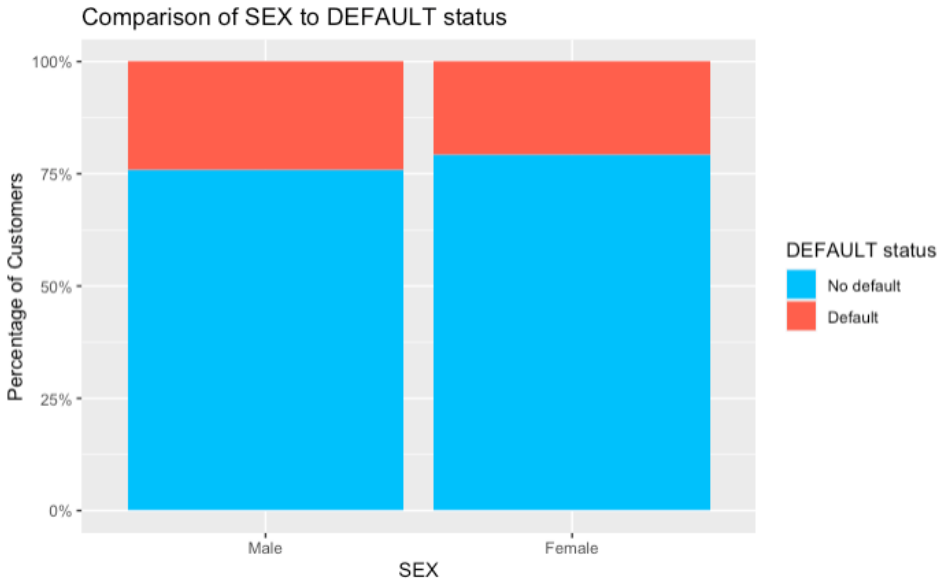
1. Overall Breakdown of Default Status

We observed that out of 29965 data points, there are 23373 customers (78%) with non-default payments and 6592 customers (22%) with default. However, this is not a cause for concern as we would expect fewer defaults as compared to non-default payments.



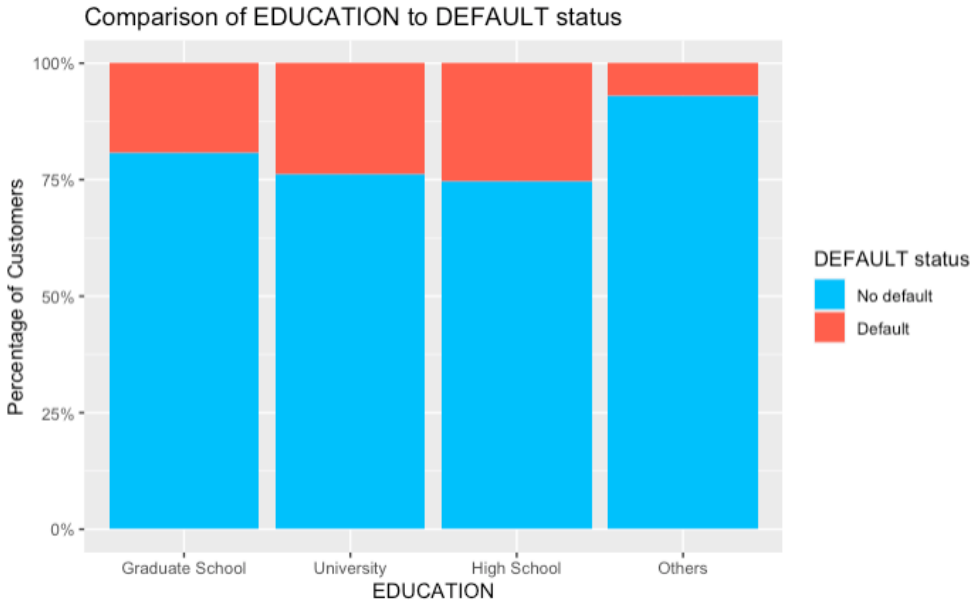
2. Gender and Default Status

We observed that the proportion of males with default payments is higher than the proportion of females with default payments.



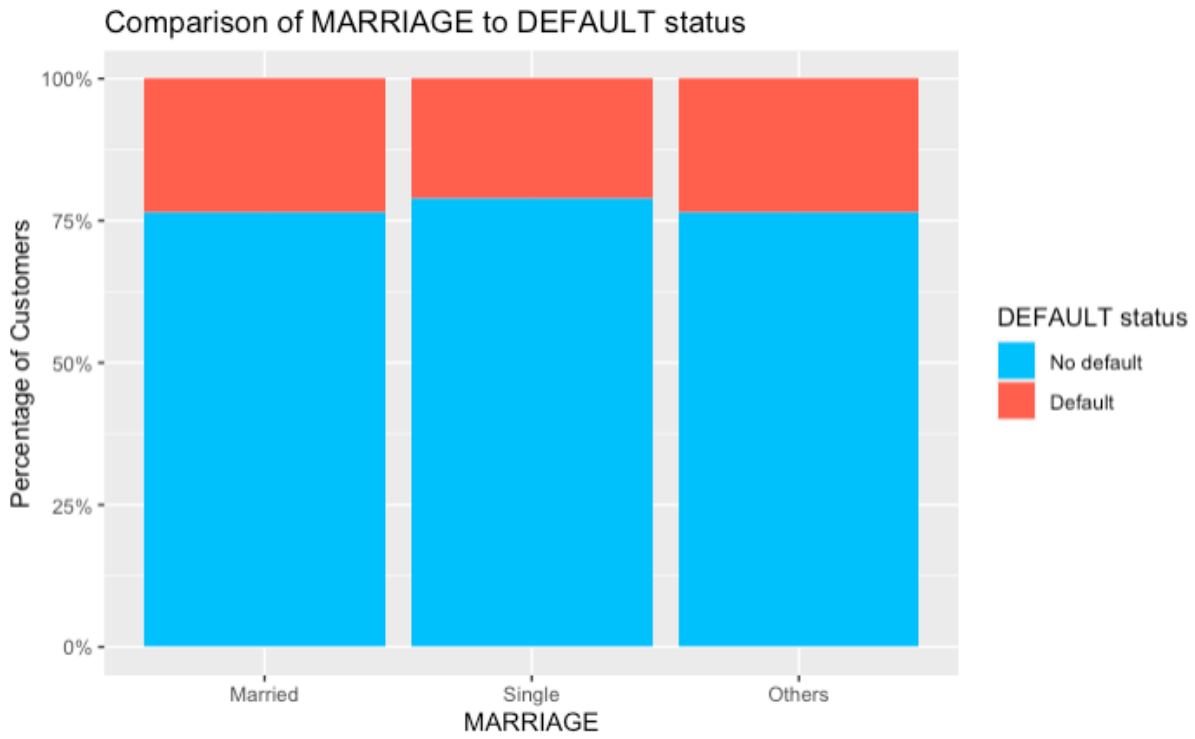
3. Education Levels and Default Status

We observed a negative correlation between the level of education of credit card holders and the percentage of default payments. This means that the higher the education level, the less likely an individual is to default on payments. This is likely due to the fact that higher education levels should result in better-paying jobs, enabling those with higher education to pay back the money and not default.



4. Marriage and Default Status

We observe that married couples have a higher proportion of default payments as compared to single people. One reason this might be the case is that married couples are more likely to take out loans to buy a house which they are unable to pay back on time resulting in default payments.



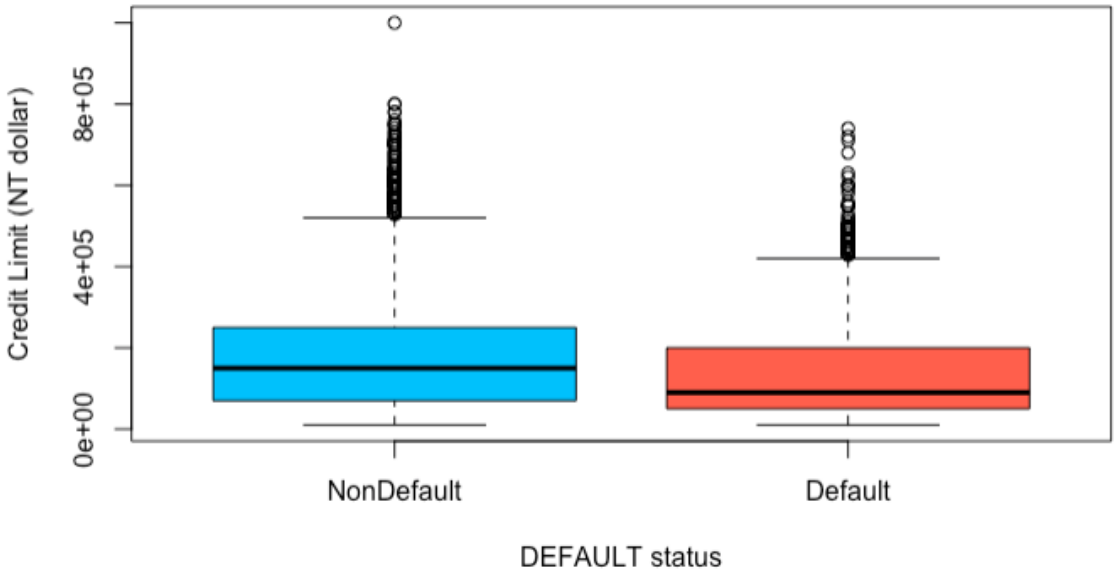
We also conducted a Chi-squared test to test whether there is a statistical difference between the proportion of default to non-default payments in customers who are married and single. It is observed that the p-value of the Chi-Squared test is 1.074*10^-6 which is lower than the significance level of 0.05. Hence there is sufficient evidence to reject the null hypothesis and conclude that the difference between the number of default and non-default payments differ for customers who are married and single is statistically significant.

```
        Pearson's Chi-squared test

data:  cont_table
X-squared = 27.489, df = 2, p-value = 1.074e-06
```

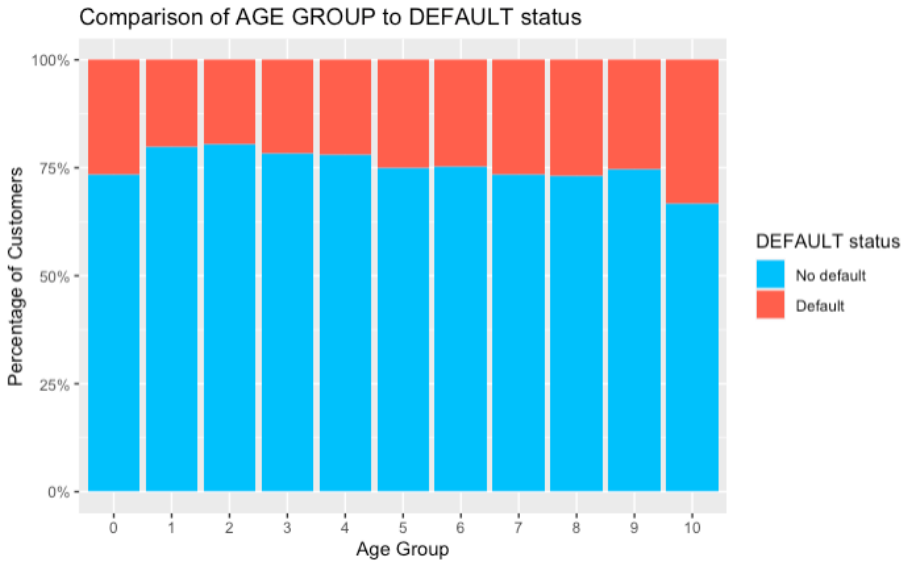
5. Credit Limit and Default Status

We observe that individuals with a higher credit limit are less likely to default on their payments.



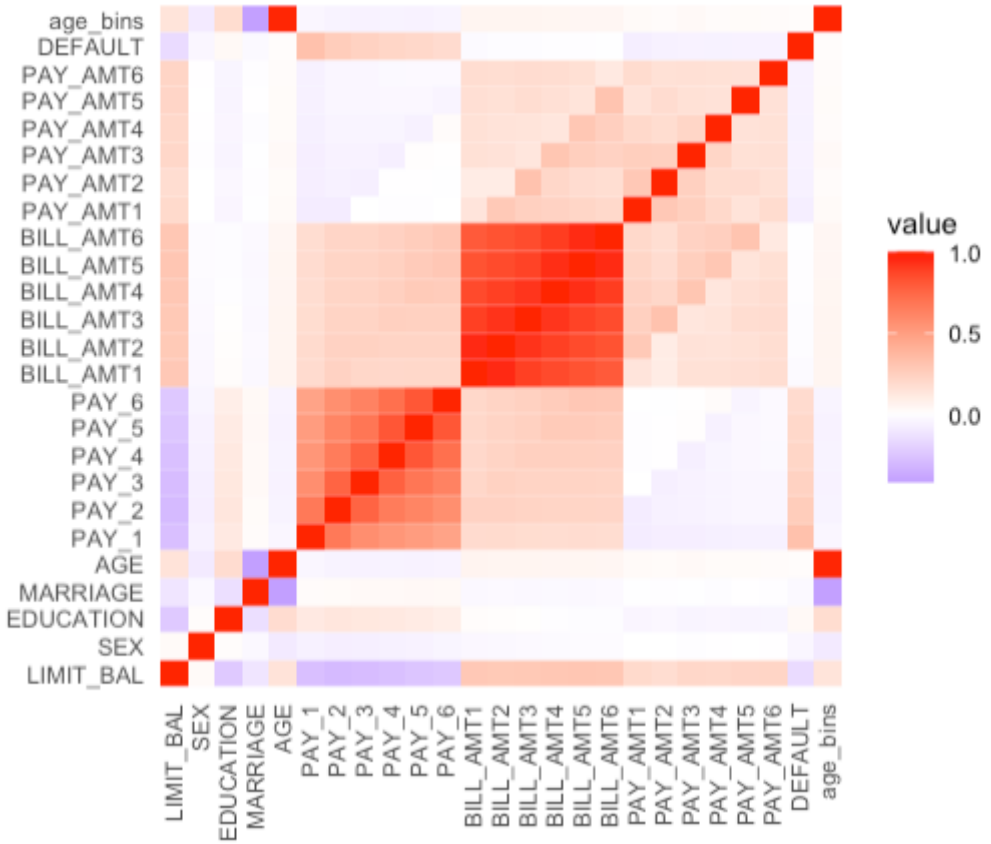
6. Age and Default Status

Our analysis indicates that the proportion of default payments tends to increase notably among customers aged 30 and above.



Feature Selection

We use feature selection to automatically or manually select features which contribute most to our prediction variable, **DEFAULT.** We drop the variable **ID** as it is ID of each customer and does not help us in predicting the variable to be predicted, **DEFAULT**. We also drop the variable **DEFAULT** as this is the variable that we will be predicting. Now, we plot out a correlation matrix between all the other variables to identify which ones will be used as feature variables.



After looking at the correlation matrix, we decided to include all the other variables to help us in the prediction.

Model Selection

Our plan is to run four models, Generalized Linear Model, Support Vector Machine, Neural Network, and Random Forest. We will provide a detailed explanation of why we have selected each of these models and describe their respective advantages and disadvantages. Additionally, we will generate a confusion matrix for each model that will show the number of outcomes predicted by the models. In this analysis, we will treat the default case with a value of 1 as the positive outcome.

- **True Positive (TP):** an instance that had a positive target value and was predicted to have a positive target value.
- **True Negative (TN):** an instance that had a negative target value and was predicted to have a negative target value.
- **False Positive (FP):** an instance that had a negative target value and was predicted to have a positive target value.
- **False Negative (FN):** an instance that had a positive target value and was predicted to have a negative target value.

1. Simple Logistic Regression - Generalised Linear Model (GLM)

GLM is a type of simple logistic regression that is versatile and flexible that can handle a wide range of variables which includes binary and count. The variable we want to predict, DEFAULT, is a binary variable which makes GLM an appropriate model to use. It is also a parametric model, which assumes a specific distribution response variable using maximum likelihood estimation. Below is the confusion matrix that shows the performance of the GLM model.

GLM Confusion Matrix		Predicted	
		0	1
Actual	0	TN = 6766	FP = 215
	0	FN = 1530	TP = 478

Advantages:

- Many ways to regularize the model to tolerate some errors and avoid over-fitting
- We do not have to worry about correlated features
- It can easily take in a new data using an online gradient descent method

Disadvantages:

- It requires observations to be independent of one another

2. Support Vector Machine (SVM)

SVM is an algorithm that is used for classification tasks that involves finding a hyperplane that separates data points into distinct classes in an N-dimensional space where N is the number of features. We chose SVM as it is able to identify a margin of separation between the default and non-default classes. Below is the confusion matrix that shows the performance of the SVM model.

GLM Confusion Matrix		Predicted	
		0	1
Actual	0	TN = 6681	FP = 300
	0	FN = 1344	TP = 664

Advantages:

- The SVM model has parameters to prevent the model from overfitting as well as tolerate errors
- SVM models are less affected by outliers and noise in the data.

Disadvantages:

- SVM model may not perform well when dealing with imbalanced datasets, where one class is more prevalent. In this case, there are more default cases as compared to non-default cases.

3. Neural Network (NN)

NN are a type of machine learning algorithm that can identify complex patterns and extract insights from large datasets. NN needs to rely on training data to learn and improve its accuracy over time. With methods like backpropagation, the NN model will allow for classifying and clustering data at high velocity, which is an extremely powerful tool. Below is the confusion matrix that shows the performance of the NN model.

GLM Confusion Matrix		Predicted	
		0	1
Actual	0	TN = 6981	FP = 0
	0	FN = 2008	TP = 0

Advantages:

- They are more flexible than linear models as they do not require restrictive assumptions about the form of the basic model.
- Neural networks can learn from the training data and adjust their parameters to new data patterns, allowing them to improve their performance over time.

Disadvantages:

- NN can be difficult to interpret as the relationship between the inputs and the outputs are not always clear.
- As NN learns from training data, they would usually need large amounts of data points in order to produce meaningful interpretations.
- They can be prone to overfitting the training data set.

4. Random Forest

Random forest is a machine learning algorithm that is widely used for classification and regression tasks. It uses an ensemble of decision trees to make predictions. It is a powerful tool for data analysis and prediction as it can handle large datasets with many features and can capture complex relationships between variables.

GLM Confusion Matrix		Predicted	
		0	1
Actual	0	TN = 6565	FP = 416
	0	FN = 1272	TP = 736

Advantages:

- Random forest is suitable for handling imbalanced datasets, as it can accurately classify data even with a significant class imbalance.
- It can handle datasets with a large number of features without overfitting, making it a good choice for high-dimensional data.

Disadvantages:

- It is computationally expensive and slow especially dealing with large datasets or with many features.
- It can lead to overfitting, resulting in poor generalisation performance on the data.

Selecting the Best Model

	GLM	SVM	NN	RF
Accuracy	0.8058738	0.8171098	0.7766159	0.8122149
Sensitivity	0.2380478	0.3306773	0	0.3665339
Specificity	0.9692021	0.9570262	1	0.9404097
Precision	0.6897547	0.6887967	NaN	0.6388889
Harmonic Mean	0.382212	0.4915213	0	0.527478
Average Class Accuracy	0.603625	0.6438518	0.5	0.6534718

In order to select the best model, we tabulated the Accuracy, Precision, Sensitivity, Specificity, Average Class Accuracy as well as Harmonic Mean of each model. Each row represents the results of the metrics used to compare the different models used. We first compare the accuracy rate between the different models and we found that the SVM model has the highest accuracy rate of 0.8171098.

However, as mentioned above, the dataset is significantly imbalanced with more customers with non-default payments as compared to customers with default payments. Hence, classification accuracy can mask poor performance as the performance of the non-default level overwhelms the performance of the default level.

Therefore, in this case, the Accuracy rate may not be a good metric to help us to identify the best model as it does not tell us the underlying distribution of the dataset. We use average class accuracy and the harmonic mean is computed as a secondary comparison. Harmonic mean helps to tackle biases by data imbalances by being sensitive to values that are lower than the average.

Overall we see that out of the 4 models, Random Forest performs the best out of the other 3 models with the highest average class accuracy and harmonic mean and thus, based on our analysis we can conclude that it is the best model to predict the variable **DEFAULT.**

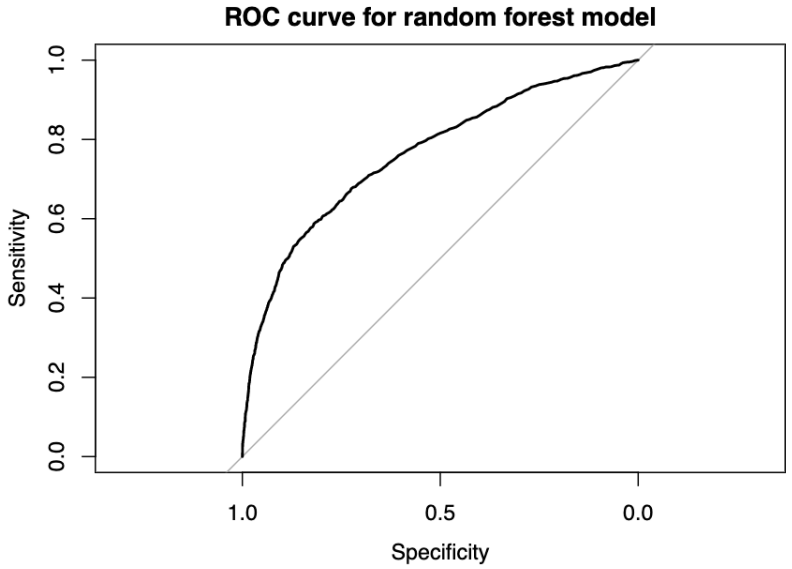
Model Evaluation

We further evaluated the performance of the selected model through the cross-validation approach and by plotting the ROC curve. The approach we will be using is K-fold cross-validation due to its ability to handle large datasets.

K-fold cross-validation involves dividing the dataset into k subsets and then training the model on k-1 folds and testing it on the remaining folds. This process is repeated k times, with each fold used once as the test set. The results of each iteration are then averaged to obtain a more robust estimate of the model's performance. We will also be able to obtain a range of accuracy scores across multiple training and test sets. This is useful as it can help prevent overfitting as we will be using different subsets of input data and evaluating the model performance after.

We assigned the value of k to be 10 and we observed an accuracy of 0.814883 or 81.5%, which is higher than the initial accuracy of 0.8122149.

Additionally, we also plotted the ROC curve and obtained the ROC index, which is the area under the graph.



The ROC index has a value of 0.7641886 which is greater than 0.7. Thus we can safely conclude that the selected model, Random Forest, is a strong model.

Areas for Improvement

1. Naive Bayes (NB)

NB is another model that we can consider using when selecting our models. NB is based on Baye's theorem which states that the probability of the hypothesis given and evidence is proportional to the product of the probability of the evidence given the hypothesis and the prior probability of the hypothesis. This model is suitable as the variables that are included are categorical variables. It is also easier to implement as it does not require such a large data sample to provide accurate results given that all samples are independent of each other. It is also robust to irrelevant features, allowing it to provide good performance even if some features are redundant. However, given the dataset, it is difficult to ensure that the features will be entirely independent of one another. Hence if we are guaranteed that the features are entirely independent of one another, we would have included this in one of the models that we chose.

2. Resampling (Tomek Links)

As mentioned above, the dataset that was given to us is imbalanced with a significant difference between the number of default and non-default payments. We resample to achieve a balanced data set and we can either oversample the minority class or undersample the majority class. However, if we oversample the minority class, it would lead to overfitting and duplicated values, which would pose a problem for machine learning models.

Hence we will choose to undersample the majority class, non-default payments. We can implement Tomek Links which involves identifying pairs of examples from different classes that are the nearest neighbours to each other. If a pair of examples is found to be a Tomek Link, the example from the majority class is removed. Random Under Sampling may remove informative examples and may not always be effective as compared to Tomek Links where it removes the data points that are closer to the boundary between classes.

3. Principal Component Analysis (PCA)

As the number of features increase, the data becomes more sparse and the volume of the feature space grows exponentially. In the dataset given, we have 23 variables and hence creating a large multidimensional space. With PCA, we can reduce the number of features in the dataset while retaining most of the important information. We can use it to identify patterns and relationships between variables in a lower-dimensional space. However, by implementing PCA, the number of variables will be greatly reduced and hence causing the lowering of the accuracy while improving the readability of the data.