BT2103 Optimisation Methods in Business Analytics
Academic Year 2022/23 Semester 2

Project Group 1

No.	Student Name	Matriculation No.
1.	Foo Yan Rong	A0233628A
2.	Han Seungju	A0243359Y
3.	Khor Kiat Siong	A0239007J
4.	Shin Jisu	A0244844Y

NUS
National University
of Singapore

Contents

1. Brief introduction of data set and data modelling problem

This dataset (‘cards.csv’) contains information on credit card clients in Taiwan between April 2005 and September 2005. The target feature, which is a default payment value, is predicted with binary value (0 = not default, 1 = default). It contains the data of 30,000 credit card holders and is described by 23 feature attributes. It includes demographic factors such as age, gender, education, and marital status, and continuous variables regarding credit such as credit and payment history.

2. Exploratory data analysis

The structure of the data is a (30000 x 25) sized data frame including ID column, without missing or null values.

2.1 Target Variable/Y/default payment next month

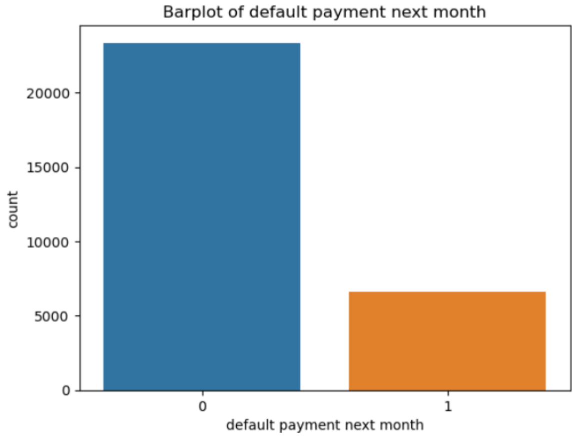
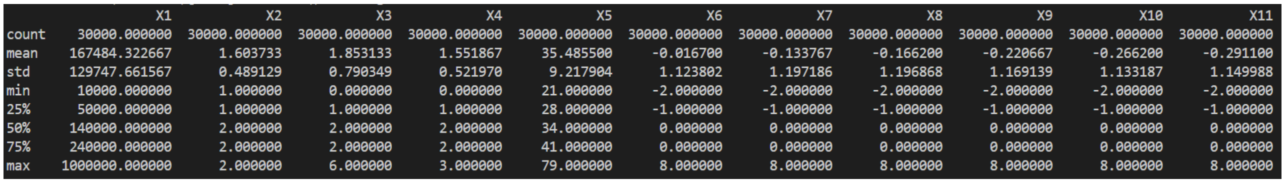


Figure 2.1: Number of clients who non-default (blue) and default (blue)

There are significantly more non-defaulters (77.9%) than defaulters (22.1%). Hence, the dataset is imbalanced which may cause the model to be biased towards predicting non-defaulters. This issue can be tackled by assigning a higher weight to the default data in the model or by assigning a lower weight to non-default data in the model. For the current project, this issue will be dealt with by tuning the weights of the classes in the model. The details are explained under 5.3 Model Training and Testing Method.

2.2 Basic Overview of Feature Variable Distribution



```
                   X1            X2            X3            X4            X5            X6            X7            X8            X9           X10           X11
count    30000.000000  30000.000000  30000.000000  30000.000000  30000.000000  30000.000000  30000.000000  30000.000000  30000.000000  30000.000000  30000.000000
mean    167484.322667      1.603733      1.853133      1.551867     35.485500     -0.016700     -0.133767     -0.166200     -0.220667     -0.266200     -0.291100
std     129747.661567      0.489129      0.790349      0.521970      9.217904      1.123802      1.197186      1.196868      1.169139      1.133187      1.149988
min      10000.000000      1.000000      0.000000      0.000000     21.000000     -2.000000     -2.000000     -2.000000     -2.000000     -2.000000     -2.000000
25%      50000.000000      1.000000      1.000000      1.000000     28.000000     -1.000000     -1.000000     -1.000000     -1.000000     -1.000000     -1.000000
50%     140000.000000      2.000000      2.000000      2.000000     34.000000      0.000000      0.000000      0.000000      0.000000      0.000000      0.000000
75%     240000.000000      2.000000      2.000000      2.000000     41.000000      0.000000      0.000000      0.000000      0.000000      0.000000      0.000000
max    1000000.000000      2.000000      6.000000      3.000000     79.000000      8.000000      8.000000      8.000000      8.000000      8.000000      8.000000
```

Figure 2.2.2: Descriptive statistics on variables X1-X11

X1/LIMIT_BAL

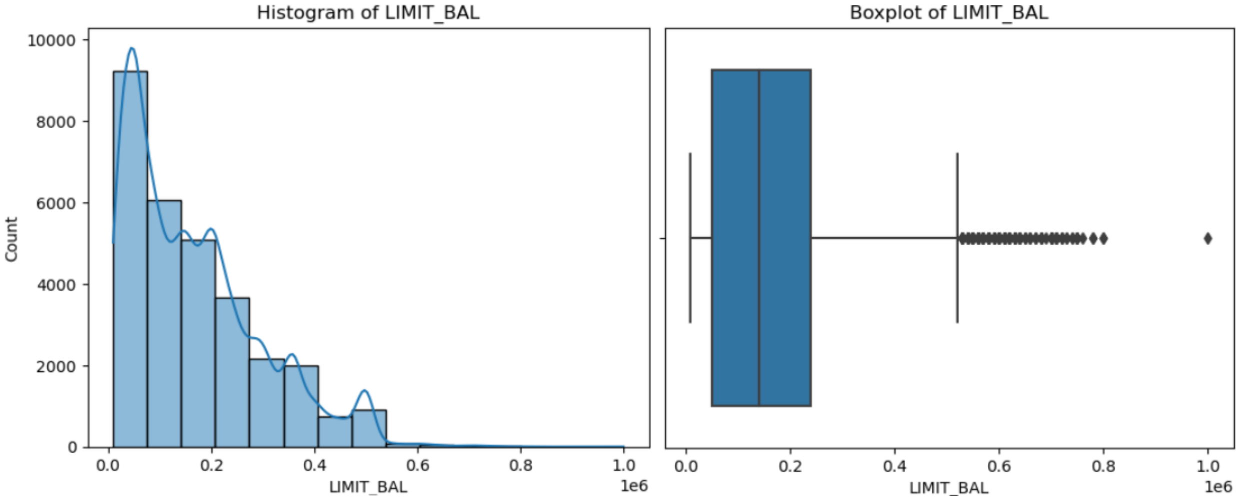


Figure 2.2.3: Left: Histogram of limit balance values for clients.
Right: Boxplot of limit balance values for clients

There are several potential outliers towards the right tail of the distribution according to the boxplot above. They represent clients with a high limit balance. They are not considered as outliers since it is possible for some high net-worth clients to have a high limit balance.

X2/SEX

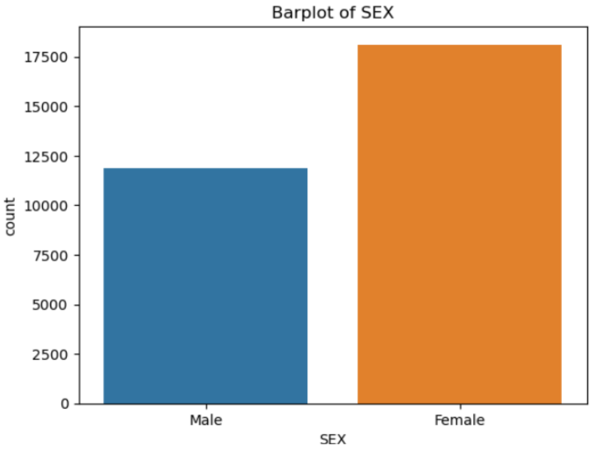


Figure 2.2.4: Barplot of client's sex

X3/EDUCATION

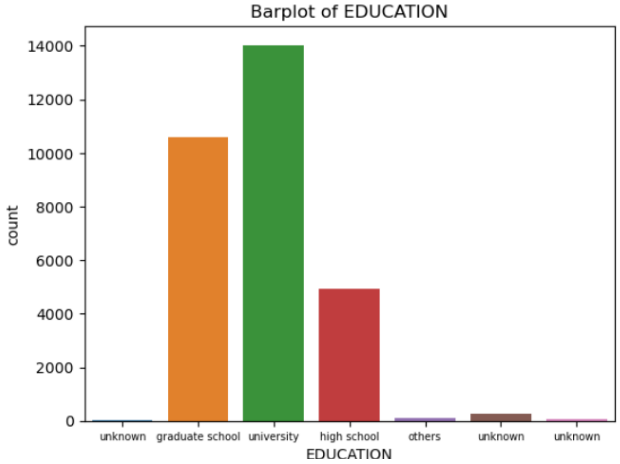


Figure 2.2.5: Barplot of client education levels

There are unknown labels present with values {0, 5, 6}.

X4/MARRIAGE

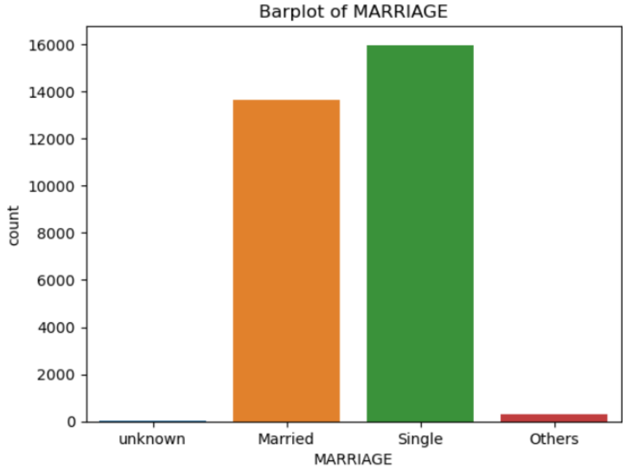


Figure 2.2.6: Barplot of client marriage status

There is an unknown label present with value {0}.

X5/AGE

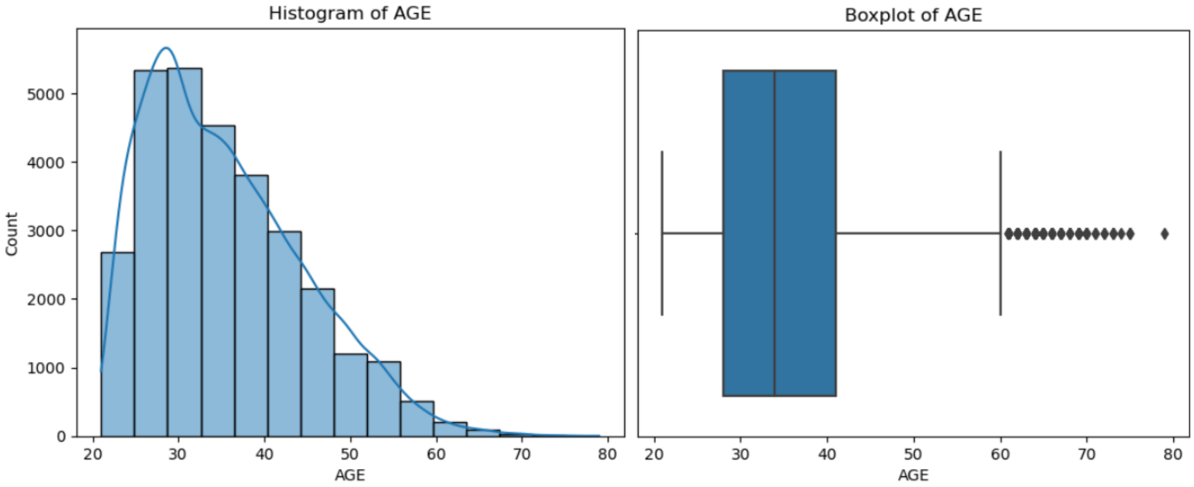


Figure 2.2.7: Left: Histogram of client age. Right: Barplot of client age

X6-X11/PAY_X

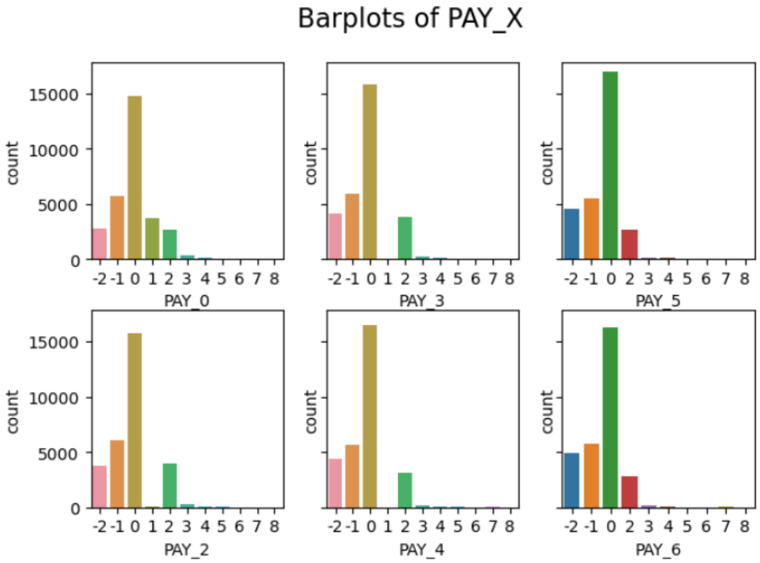


Figure 2.2.8: Barplots of Pay_X (X6-X11) values

There are unknown labels present with values {-2, 0}. PAY_2, PAY_3, PAY_4, PAY_5 and PAY_6 have very similar distributions. PAY_0 has a significantly greater proportion of data with value = 1 as compared to the other variables.

X12-X17/BILL_AMTX

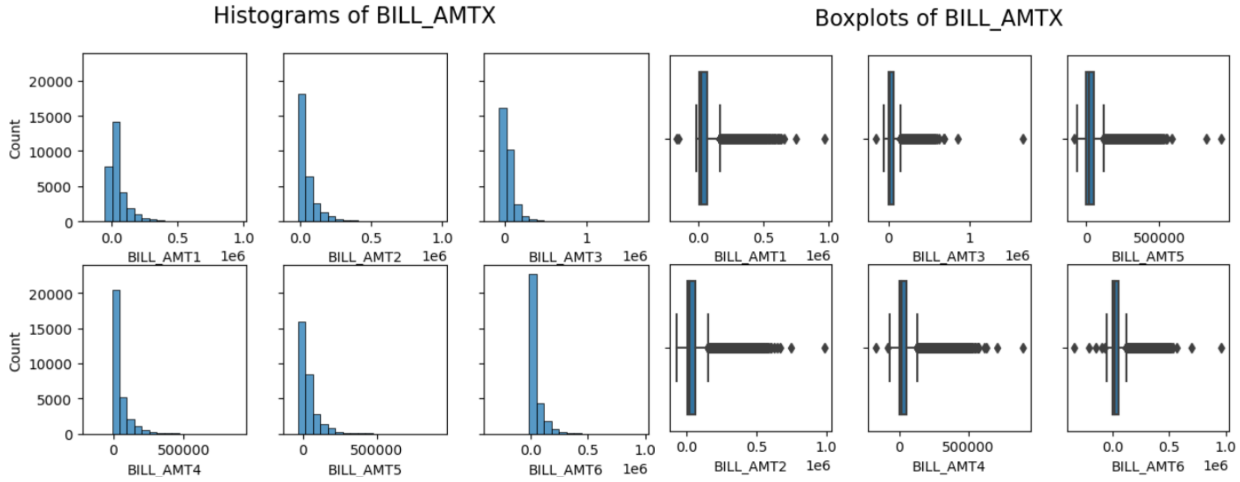
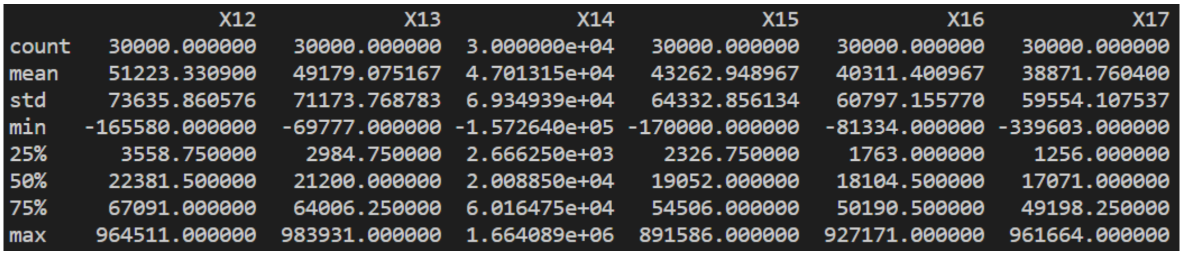


Figure 2.2.9: Left: Histogram of Bill_AmtX (X12-X17). Right: Boxplots of Bill_AmtX (X12-X17)



```
                X12            X13            X14            X15            X16            X17
count  30000.000000   30000.000000  3.000000e+04   30000.000000   30000.000000   30000.000000
mean   51223.330900   49179.075167  4.701315e+04   43262.948967   40311.400967   38871.760400
std    73635.860576   71173.768783  6.934939e+04   64332.856134   60797.155770   59554.107537
min  -165580.000000  -69777.000000 -1.572640e+05 -170000.000000  -81334.000000 -339603.000000
25%     3558.750000    2984.750000  2.666250e+03    2326.750000    1763.000000    1256.000000
50%    22381.500000   21200.000000  2.008850e+04   19052.000000   18104.500000   17071.000000
75%    67091.000000   64006.250000  6.016475e+04   54506.000000   50190.500000   49198.250000
max   964511.000000  983931.000000  1.664089e+06  891586.000000  927171.000000  961664.000000
```

Figure 2.2.10: Statistical description of X12-X17

There are several outliers in the data for variables "BILL_AMT1", "BILL_AMT2", "BILL_AMT3", "BILL_AMT4", "BILL_AMT5" and "BILL_AMT6". These outliers tend to be right-skewed. There are also negative bill amounts, which could be interpreted as the client having credit in his balance.

X18-X23/PAY_AMTX

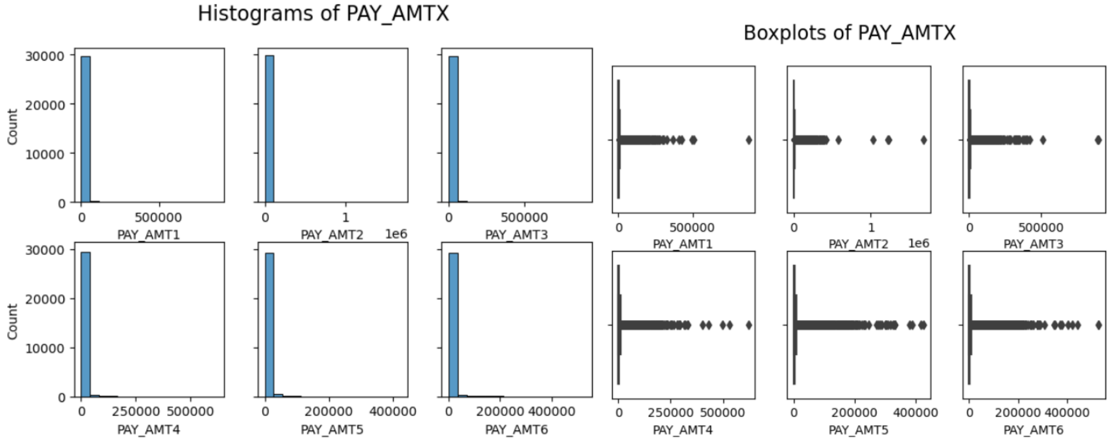
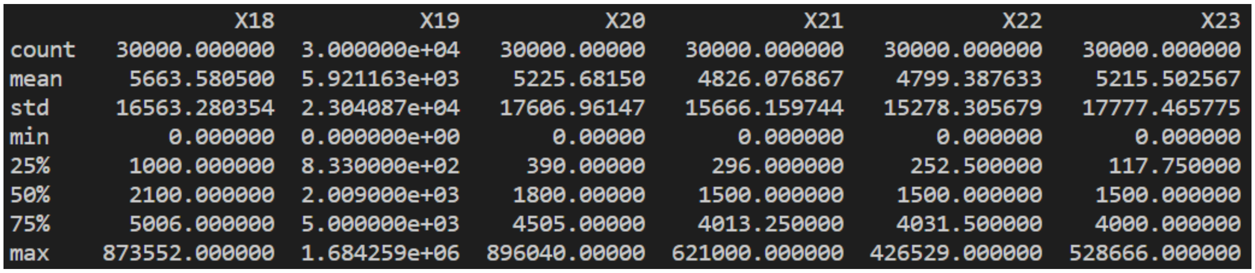


Figure 2.2.11: Left: Histogram of Pay_AmtX (X18-X23).

Right: Boxplots of Pay_AmtX (X18-X23)



```
                X18           X19          X20           X21           X22           X23
count  30000.000000  3.000000e+04  30000.00000  30000.000000  30000.000000  30000.000000
mean    5663.580500  5.921163e+03   5225.68150   4826.076867   4799.387633   5215.502567
std    16563.280354  2.304087e+04  17606.96147  15666.159744  15278.305679  17777.465775
min        0.000000  0.000000e+00      0.00000      0.000000      0.000000      0.000000
25%     1000.000000  8.330000e+02    390.00000    296.000000    252.500000    117.750000
50%     2100.000000  2.009000e+03   1800.00000   1500.000000   1500.000000   1500.000000
75%     5006.000000  5.000000e+03   4505.00000   4013.250000   4031.500000   4000.000000
max   873552.000000  1.684259e+06 896040.00000 621000.000000 426529.000000 528666.000000
```

Figure 2.2.12: Statistical description of X18-X23

There are several outliers in the data for variables "PAY_AMT1", "PAY_AMT2", "PAY_AMT3", "PAY_AMT4", "PAY_AMT5" and "PAY_AMT6". These outliers tend to be right-skewed. Unlike Bill_AmtX, pay amount has no negative values because clients cannot possibly pay negative amounts.

Outlier Analysis for Continuous Variables

There are several outliers for variables such as LIMIT_BAL, BILL_AMT and PAY_AMT. However, these extreme values are likely to be legitimate observations. Hence, they cannot be simply removed as removing them may skew the true distribution of these variables. Furthermore, these data may be critical in determining the target variable. Therefore, no outliers will be filtered out.

3. Data Pre-Processing

3.1 Unknown Labels

There are several categorical variables such as EDUCATION, MARRIAGE and PAY_X which have unknown labels. For variable EDUCATION, values {0, 5, 6} do not have an assigned label. Since the proportion of data for these values are rather small, they are combined into the label under “others”. Similarly, for MARRIAGE, value {0} does not have an assigned label. Hence, data with value 0 is combined into the label for “others” like it is shown below.

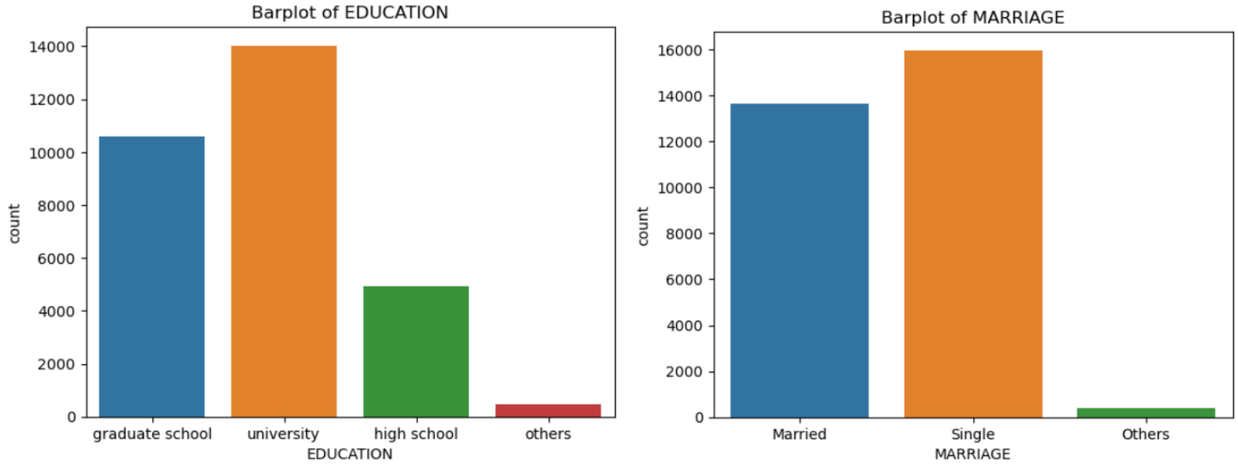


Figure 3.1.1: Left: Barplot of client education levels after replacing unknown values.
Right: Barplot of client marriage status after replacing unknown values.

For PAY_X, values {-2, 0} do not have an assigned label. However, the proportion of data with these values is large, hence they cannot be simply combined under a single value without understanding their representations. Even if the data is uninterpretable, the data can be used to extract potentially useful insights in our feature selection. Also, it is observed that data samples with values greater than 5 have a very low proportion. Hence, they can be combined together as having delayed payment 5 months or later. This helps to increase the simplicity of the feature variable without losing too much information.

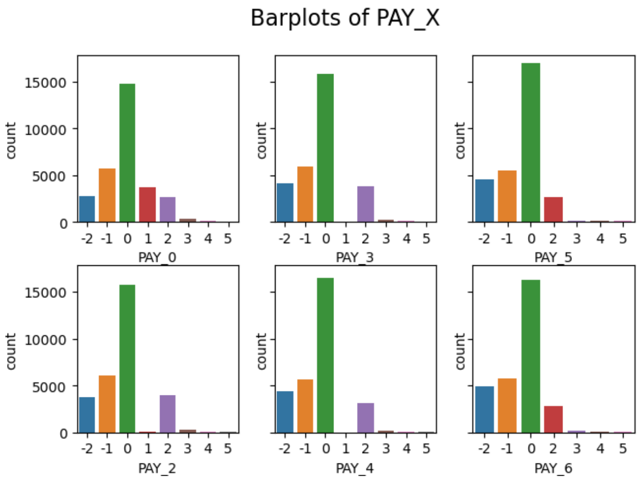


Figure 3.1.2: Barplots of Pay_X (X6-X11) after replacing values greater than 5

3.2 Z-Index Standardisation for Continuous Variables

Several variables in the dataset have different scales. For example, "AGE" has a range of [21,79] while variables such as "LIMIT_BAL", "BILL_AMT" and "PAY_AMT" have values in the magnitudes of 100,000 and 1,000,000. Depending on the model applied, feature scaling may be important to ensure that features with a large range and higher magnitude do not necessarily have a larger influence over the model. This is especially important for models such as logistic regression and neural networks that use gradient descent. Support vector machine is a distance-based algorithm, hence it is also affected by the range of features.

Due to the significant presence of outliers in several feature variables such as BILL_AMTX and PAY_AMTX, z-index standardisation is preferred over min-max scaling. The presence of outliers skews the ratio between data points, hence limiting the effectiveness of min-max-scaling. Therefore, standardisation (Z-score normalisation) is performed for each feature; X1 (LIMIT_BAL), X5 (AGE), and X12-X23 (BILL_AMT AND PAY_AMT).

3.3 Dummy Encoding for Categorical Variables

For multi-categorical variables such as EDUCATION, MARRIAGE and PAY_X, it needs to be ensured that the different categories are accurately represented in the model. For example, MARRIAGE is represented by {1 = married; 2 = single; 3 = others}. However, 1, 2 and 3 are simply numbers which represent a certain category and do not have a numerical relationship with one another.

Dummy encoding allows the transformation of these categorical variables into a set of binary variables. This ensures the nominal variables can be effectively used in our models. In this case, MARRIAGE can be represented by assigning 2 dummy variables for a total of 3 possible categories in MARRIAGE.

	MARRIAGE_2	MARRIAGE_3
MARRIED	0	0
SINGLE	1	0
OTHERS	0	1

Table 3.3.1: Encoding of marriage status

The other categorical features are also represented in a similar way.

4. Feature selection

4.1 Correlation between feature variables

High correlation between feature variables may cause multicollinearity problems which may be issues for models that assume independence of feature variables such as the Naive Bayes model.

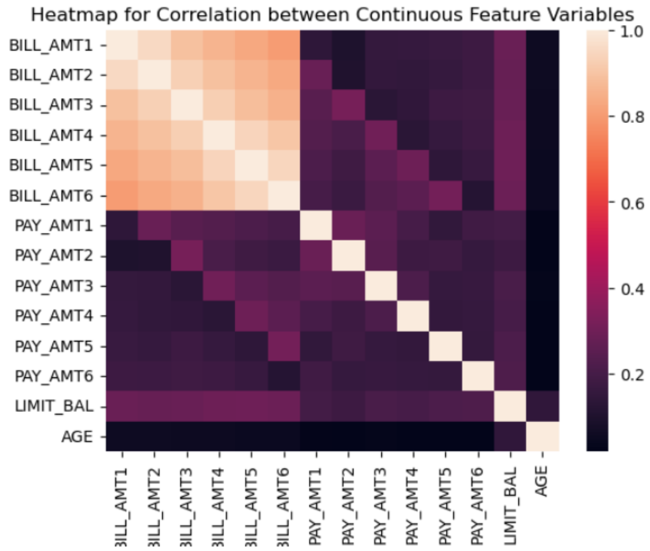


Figure 4.1.1: Correlation matrix between continuous variables

The observation shows that the 6 "BILL_AMT" feature variables have very high correlation with one another. Hence, the entire set of "BILL_AMT" attributes cannot be included into the model.

One consideration is to combine all the values of “BILL_AMT” into one variable “BILL_AMT_AVG” by taking the average across all “BILL_AMT” variables.

4.2 Correlation between feature and target variable

The higher the correlation between the feature and target variable, the more likely the feature variable is useful for prediction.

4.2.1 Categorical Variables

Chi2 test is to check for independence of the categorical feature variable with the target variable.

Feature Variable	**Chi2 value (3dp)**	**p-value**
SEX	47.709	4.945e-12
EDUCATION	163.217	1.233e-32
MARRIAGE	35.662	8.826e-08
PAY_0	5365.965	≈ 0.0
PAY_2	3474.467	≈ 0.0
PAY_3	2622.462	≈ 0.0
PAY_4	2341.470	≈ 0.0

PAY_5	2197.695	≈ 0.0
PAY_6	1886.835	≈ 0.0

Table 4.2.1.1: Chi2 statistic for categorical variables

Chi2 values for all feature variables suggest that there is some level of dependence between these features and the target variable (all p-values > 0.05). Hence, they are likely to be useful features to include in our model.

4.2.2 Continuous Variables (Data Discretization)

After performing data discretization on continuous feature variables to convert them into pseudo-categorical variables, it is possible to check for correlation with the target variable by performing the chi2 test once again.

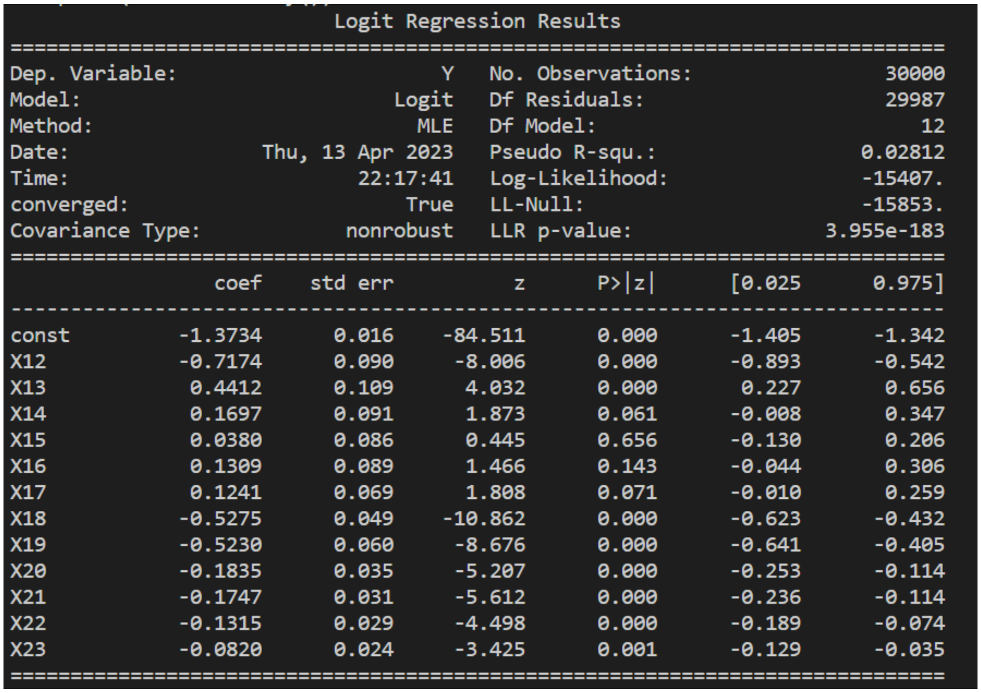
Feature Variable	**Chi2 value (3dp)**	**p-value**
LIMIT_BAL	745.022	1.443e-155
AGE	79.210	2.320e-13
PAY_AMT1	28.268	8.366e-0
PAY_AMT2	10.902	0.0914
PAY_AMT3	23.392	0.000675
PAY_AMT4	22.931	0.00345
PAY_AMT5	45.848	6.426e-07
PAY_AMT6	33.213	0.000123
BILL_AMT_AVG	22.852	0.00181

Table 4.2.2.1: Chi2 statistic for continuous variables

Chi2 values for all these feature variables except PAY_AMT2 seem to suggest that there is some level of dependence between these features and the target variable. Hence, they may be useful features to include in our model. PAY_AMT2 has p-value > 0.05, hence PAY_AMT2 may be a less useful feature. However, the Chi2 values might vary depending on how the values are binned. Hence, this method may not be the most reliable in finding correlation of the continuous feature variables with the target variable.

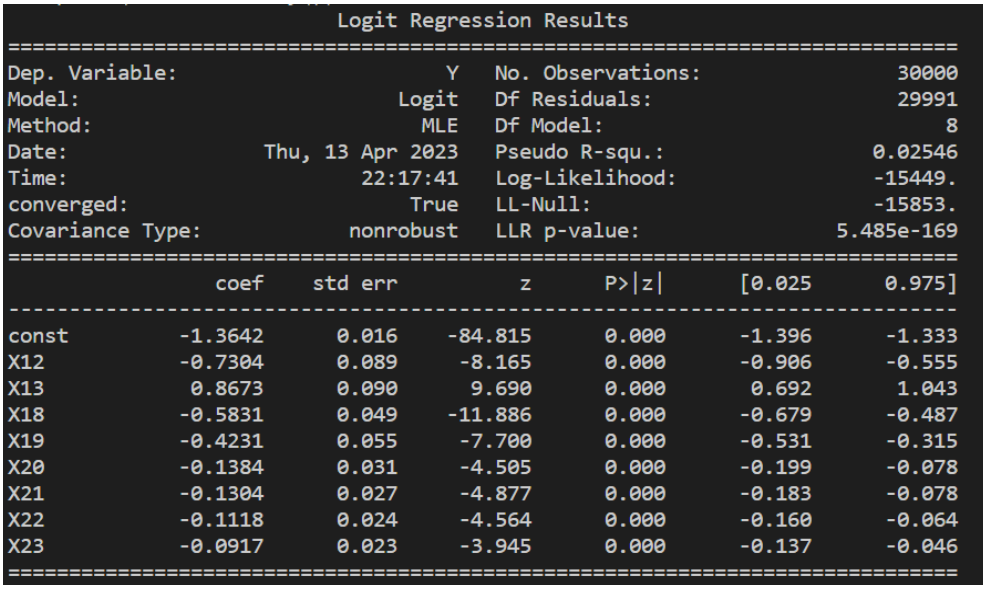
4.3 Logistic Regression

The purpose of logistic regression is to check for significant correlation between X12-X23 and credit card default (Y).



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      Y   No. Observations:                30000
Model:                          Logit   Df Residuals:                    29987
Method:                           MLE   Df Model:                           12
Date:                Thu, 13 Apr 2023   Pseudo R-squ.:                 0.02812
Time:                        22:17:41   Log-Likelihood:                -15407.
converged:                       True   LL-Null:                       -15853.
Covariance Type:            nonrobust   LLR p-value:                3.955e-183
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.3734      0.016    -84.511      0.000      -1.405      -1.342
X12           -0.7174      0.090     -8.006      0.000      -0.893      -0.542
X13            0.4412      0.109      4.032      0.000       0.227       0.656
X14            0.1697      0.091      1.873      0.061      -0.008       0.347
X15            0.0380      0.086      0.445      0.656      -0.130       0.206
X16            0.1309      0.089      1.466      0.143      -0.044       0.306
X17            0.1241      0.069      1.808      0.071      -0.010       0.259
X18           -0.5275      0.049    -10.862      0.000      -0.623      -0.432
X19           -0.5230      0.060     -8.676      0.000      -0.641      -0.405
X20           -0.1835      0.035     -5.207      0.000      -0.253      -0.114
X21           -0.1747      0.031     -5.612      0.000      -0.236      -0.114
X22           -0.1315      0.029     -4.498      0.000      -0.189      -0.074
X23           -0.0820      0.024     -3.425      0.001      -0.129      -0.035
==============================================================================
```

Figure 4.3.1: Logistic regression of Y against X12-X23



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      Y   No. Observations:                30000
Model:                          Logit   Df Residuals:                    29991
Method:                           MLE   Df Model:                            8
Date:                Thu, 13 Apr 2023   Pseudo R-squ.:                 0.02546
Time:                        22:17:41   Log-Likelihood:                -15449.
converged:                       True   LL-Null:                       -15853.
Covariance Type:            nonrobust   LLR p-value:                5.485e-169
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.3642      0.016    -84.815      0.000      -1.396      -1.333
X12           -0.7304      0.089     -8.165      0.000      -0.906      -0.555
X13            0.8673      0.090      9.690      0.000       0.692       1.043
X18           -0.5831      0.049    -11.886      0.000      -0.679      -0.487
X19           -0.4231      0.055     -7.700      0.000      -0.531      -0.315
X20           -0.1384      0.031     -4.505      0.000      -0.199      -0.078
X21           -0.1304      0.027     -4.877      0.000      -0.183      -0.078
X22           -0.1118      0.024     -4.564      0.000      -0.160      -0.064
X23           -0.0917      0.023     -3.945      0.000      -0.137      -0.046
==============================================================================
```

Figure 4.3.2: Logistic regression of Y against X12-X23 excluding X14-X17

As the result shows that the coefficient of X12, X13, X18-X23 for logistic regression is statistically significant at a 5% significance level, these variables are significantly correlated with Y. However, the coefficients of X14-X17 are not statistically significant at the 5% significance level. From the correlation matrix (Figure 4.3.1), it could be deduced that this would be reasonable since variables X12-X17 are significantly correlated with each other and some variables would be redundant in predicting Y. In addition, it could be suggested that default credit card clients are likely to not pay their bills punctually from an intuitive perspective.

4.4 Feature Importance Coefficient

Feature importance assigns a score to each feature variable based on how significant the feature is to the model when they are predicting a target variable. Given that the target variable is a binary variable, logistic regression model can be fit to retrieve the coefficient of importance. Coefficient values can be both positive and negative. A positive coefficient indicates a feature which predicts a defaulter and a negative coefficient indicates a feature which predicts a non-defaulter. In this case, features with the highest absolute coefficient value are considered the most important.

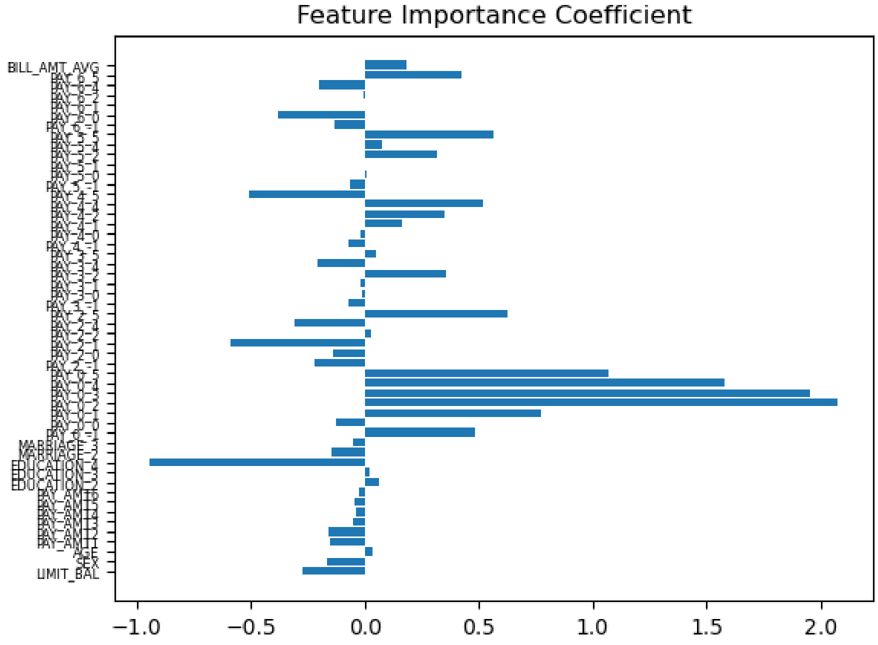


Figure 4.4.1: Feature importance coefficient for each feature

It could be noticed that EDUCATION_4 (others) has a significantly higher negative coefficient, indicating that a customer who is not in graduate school, university or high school is likely to be a non-default customer. This makes sense as these customers likely do not suffer a student debt as compared to customers who have been to graduate school, university or high school. Hence, EDUCATION is likely an important feature to be included in the model.

Also, PAY_0_2, PAY_0_3, PAY_0_4 and PAY_0_5 have significantly higher positive coefficient, indicating that a customer who has delayed payment for 2 or more months in the first month is likely to default, which makes sense due to their inability to pay on time. This also means that PAY_0 is likely a very important feature to be included in the model.

4.5 Forward and Backward Feature Selection

The forward and backward feature selection helps to filter features with greater importance to our model sequentially. In the forward feature selection, the most significant features are added sequentially. On the other hand, in the backward feature selection, the least significant features are removed sequentially.

Given n=5 features,

Forward feature selection, in order of greatest importance:
['PAY_0_1', 'PAY_0_2', 'PAY_0_3', 'PAY_2_2', 'PAY_5_2']

Backward feature selection, in order of greatest importance:
['PAY_0_1', 'PAY_0_2', 'PAY_0_3', 'PAY_2_2', 'PAY_5_2']

These features are likely to be more important for our model.

Whilst potential useful features are identified with several methods, all available features should be considered for model training to prevent from accidentally omitting a useful feature that might not have been captured using the methods stated above. Furthermore, different models may utilise features differently, hence the importance of a feature may differ across models. As a result, all available features in the models are kept.

5. Model selection

5.1 Supervised vs Unsupervised vs Reinforcement Learning

The training dataset provides labels of whether a client defaults or not. Hence, supervised learning models are used to predict future client defaults. These models include: Logistic Regression, Naive Bayes, Support Vector Machine, Decision Trees, and Neural Network.

5.2 Classification vs Regression

As the target variable is a binary variable of default (value of 1) and non-default (value of 0), the model selected must be a classification model.

5.3 Model Training and Testing Method

The performance of stratified k-fold cross-validation is conducted with k=10.The k-fold cross-validation is to ensure that the performance of the model does not depend on the way the training and test subsets were chosen. Since the class distribution is unbalanced, stratified k-fold is used to ensure that there are samples of the minority class (defaulters) present in the test set. In the worst case, there may be no samples of defaulters in the testing set.

To tackle the problem of class imbalance, each class can be weighed differently. Setting a higher weight to the default class (minority class) increases the penalty incurred by the model if it misclassifies a default sample. To calculate the weight to assign to each class, formula 5.3.1 is used:

$$weight_i = \frac{number\ of\ observations}{total\ number\ of\ classes * number\ of\ observations\ of\ i} \quad for\ i\ =\ 0\ and\ 1$$

$$weight_0 = \frac{30\ 000}{2\ *\ 23\ 364} = 0.642 \qquad weight_1 = \frac{30\ 000}{2\ *\ 6636} = 2.260$$

Formula 5.3.1

5.4 Models selected

5.4.1 Logistic Regression

It estimates the coefficients of the input variables, which represent the strength and direction of the relationship between each input variable and the output variable. Hence, it could be used to predict the probability of an output variable to be in one of the classes given a new set of input variables.

5.4.2 Gaussian Naive Bayes

A more general Naive Bayes model which focuses more on continuous variables, assuming that all the continuous variables are associated with each feature distributed in Gaussian Distribution.

5.4.3 Support Vector Machine

The support vector machine model can classify data with high-dimensional spaces, just like the dataset given. It is less susceptible to overfitting, which may allow the model to perform well with unseen test data.

5.4.4 Decision Tree

The decision tree model is easy to interpret and can handle both continuous and categorical variables. It does not rely on summary statistics, making it robust to outliers.

5.4.5 Neural Network (Multilayer Perceptron (MLP) classifier)

The model was run based on the stochastic gradient descent solver. Input layer of the model would be equal to the number of variables, which is 82, in our dataset after preprocessing. This is due to the large number of variables created after encoding the various categorical variables. Output layer of the model would have 1 node predicting whether the client defaults or not. Selection of hidden layers and number of nodes in each hidden layer: arbitrary selection of 15 nodes per hidden layer, and 2 hidden layers. Models with different hidden layers and nodes per hidden layers were iteratively run through and fitted with training data. However, it was found that the accuracy of each model did not differ by much.

6. Model evaluation

6.1 Accuracy Rate

Simple evaluation metric such as accuracy rate is used to give a general overview of the model performance. The average accuracy rate over the 10 iterations of the 10-fold cross validation is shown below.

$$\text{Accuracy rate} = \frac{Number\ of\ Correct\ Predictions}{Number\ of\ Test\ Data}$$

Model	**Accuracy Rate (1dp)**
Logistic Regression	77.6%
Gaussian Naive Bayes	80.5%
Support Vector Machine	77.9%
Decision Tree	73.4%
Neural Network (MLP)	81.9%

However, due to the class imbalance problem, accuracy may not be the best evaluation method to check on the model's performance. If the model has a heavy bias towards predicting non-default samples, then a high accuracy may be reflected even though the model may perform very poorly in predicting default samples.

6.2 Confusion Matrix

The confusion matrix helps us calculate the misclassification rate on the test set. It shows the true positive, false positive, true negative and false negative instances of our model on the test set. We set 1 (default) to be positive, and 0 (non-default) to be negative. Given that 10-fold cross-validation was performed, the average for each iteration is calculated (rounded off to nearest integer).

Model	**True Negative**	**False Positive**	**False Negative**	**True Positive**
Logistic Regression	1943	393	279	383
Gaussian Naive Bayes	2169	167	418	245
Support Vector	1996	340	317	346

Machine				
Decision Tree	1930	406	399	264
Neural Network	2214	121	425	238

6.3 Precision, Recall, Specificity, False Alarm Rate and F1-Score

Using the data provided from the confusion matrix, precision, recall, specificity and false alarm rate can be calculated. The average value for each evaluation method over the 10 iterations is shown below.

$$Recall\ =\ TPR\ \ =\frac{TP}{TP+FN}$$

$$Precision\ \ =\ \frac{TP}{TP{+}FP}$$

$$FNR\ =\frac{FN}{FN+TP}$$

$$Specificity\ \ =\ TNR\ \ =\frac{TN}{TN+FP}$$

$$False\ Alarm\ Rate\ \ =\ FPR\ =\frac{FP}{TN+FP}$$

$$\text{F1-Score} = \frac{2{*}precision{*}recall}{precision+recall}$$

Model	**Recall**	**Specificity**	**False Alarm Rate (1-specificity)**	**False Negative Rate (1-recall)**	**Precision**	**F1-Score**
Logistic Regression	57.8%	83.2%	16.8%	42.2%	49.5%	53.3%
Gaussian Naive Bayes	37.0%	92.8%	7.2%	63.0%	61.6%	45.3%
Support Vector Machine	52.7%	85.6%	14.4%	47.3%	50.8%	51.3%
Decision Tree	40.1%	82.7%	17.3%	59.9%	39.1%	39.6%
Neural Network	35.9%	95.0%	5.0%	64.1%	65.9%	46.5%

F1-score takes into account how the data is distributed. Given a class imbalance problem such as this, F1-score may be a good evaluation method to assess the performance of our model. In the context of credit default, it is of greater importance to be able to correctly identify the default cases. Hence, more emphasis should be placed on the recall score.

6.4 Average Class Accuracy

Similarly, average class accuracy ensures that poor performance on predicting the minority class (default) is captured. Thus, it may also be a good evaluation metric.

Model	**Average Class Accuracy**
Logistic Regression	70.5%
Gaussian Naive Bayes	64.9%
Support Vector Machine	68.7%
Decision Tree	61.3%
Neural Network	65.6%

6.5 ROC Index

ROC Index tells us how good the model is at differentiating between the two classes, which may be a good evaluation metric to use.

Model	**ROC Index**
Logistic Regression	70.5%
Gaussian Naive Bayes	64.9%
Support Vector Machine	68.8%
Decision Tree	61.2%
Neural Network	65.4%

7. Conclusion

The model should ideally predict client defaults with a high accuracy rate, as the goal is to predict accurately whether a client is likely to default or not. This means that maximising the recall value should have a high priority. With the imbalance of client default and non-default in the dataset, the model should ideally also perform well for average class accuracy and F1-Score.

Based on the evaluation methods, the selected models performed differently. While some models like the linear support vector machine had a high accuracy rate of 82.0%, it performed poorly in the average class accuracy rate of 64.3% compared to the other models, since it predicted a lot of results to be negative. In the context of credit default, having a low recall value is likely to be more detrimental.

Comparing the selected models, the logistic regression model is the best performing model, scoring the highest in both average class accuracy and F1-score. It also has the highest ROC index. While it suffers from lower precision and specificity compared to other models, it outperforms the other models in the recall score which is of greater importance in the context of credit default. The logistic regression would be the most effective in ensuring the defaulters are correctly classified out of the selected models.

8. Discussion on Point of Improvement

There are various areas that can be improved. First, the more data can be collected for the default samples. By having more data on default samples, the problem of class imbalance would be less extreme.

Second, to provide a greater interpretability, more labels on certain variables can be used. For example in the PAY_X variable, the label for data with values equal to 0 and -2 were not provided even though these data account for a significant proportion of the sample data.

Third, more new features could be created by combining existing features. Based on the interaction of features, more meaningful features could be created. For example, PAY_AMT / BILL_AMT may potentially be a useful feature to be included in the model as it shows the proportion of the bill amount paid by the client for a particular month.

Fourth, more experiments can be done by tuning the threshold for classification. By default, the threshold is 0.5 for the models. By experimenting with different thresholds for classification, a more reliable threshold to classify the samples could be found.

Lastly, for the Multi Perceptron neural network model, more evaluation on the speed of training could be conducted in order to determine the optimal amount of nodes and hidden layers to be used for training of the model. Excess nodes and hidden layers add more complexity to the model without providing much more predictive accuracy. Hence, eliminating these excess nodes and hidden layers can allow a faster training time whilst providing similar levels of predictive accuracy. The model could also be optimised by testing different activation methods. Currently the model used in the project utilises ReLu as the squashing function, which is a more efficient way compared to the sigmoid or hyperbolic tangent squashing functions. However, other squashing functions or optimisations of the ReLu could be tried and tested to see if it can provide a better level of accuracy, such as the leaky ReLu, parameterised ReLu, exponential ReLu.