1. Brief Introduction

1.1 Data Set

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. It contains payment information of 30,000 credit card holders obtained from a bank in Taiwan. The target feature (column Y) to be predicted is binary valued 0 (= not default) or 1 (= default).

1.2 Data Modeling Problem

The data modeling problem is to determine the accuracy of 4 different models (refer to Chapter 5) used for credit risk assessment in predicting the probability of default for a credit card holder.

2. Data Exploration Analysis

2.1 Overview of Data

There are a total of 30,000 observations and 25 variables in the original data set, before any modifications. There is 1 dependent factor: *'default_payment_next_month'*, and 24 other independent variables.

```
str(data)
```

```
'data.frame':   30000 obs. of  25 variables:
 $ ID                        : int  1 2 3 4 5 6 7 8 9 10 ...
 $ LIMIT_BAL                 : int  20000 120000 90000 50000 50000 50000 500000 100000 140000 20000 ...
 $ SEX                       : int  2 2 2 2 1 1 1 2 2 1 ...
 $ EDUCATION                 : int  2 2 2 2 2 1 1 2 3 3 ...
 $ MARRIAGE                  : int  1 2 2 1 1 2 2 2 1 2 ...
 $ AGE                       : int  24 26 34 37 57 37 29 23 28 35 ...
 $ PAY_0                     : int  2 -1 0 0 -1 0 0 0 0 -2 ...
 $ PAY_2                     : int  2 2 0 0 0 0 0 -1 0 -2 ...
 $ PAY_3                     : int  -1 0 0 0 -1 0 0 -1 2 -2 ...
 $ PAY_4                     : int  -1 0 0 0 0 0 0 0 0 -2 ...
 $ PAY_5                     : int  -2 0 0 0 0 0 0 0 0 -1 ...
 $ PAY_6                     : int  -2 2 0 0 0 0 0 -1 0 -1 ...
 $ BILL_AMT1                 : int  3913 2682 29239 46990 8617 64400 367965 11876 11285 0 ...
 $ BILL_AMT2                 : int  3102 1725 14027 48233 5670 57069 412023 380 14096 0 ...
 $ BILL_AMT3                 : int  689 2682 13559 49291 35835 57608 445007 601 12108 0 ...
 $ BILL_AMT4                 : int  0 3272 14331 28314 20940 19394 542653 221 12211 0 ...
 $ BILL_AMT5                 : int  0 3455 14948 28959 19146 19619 483003 -159 11793 13007 ...
 $ BILL_AMT6                 : int  0 3261 15549 29547 19131 20024 473944 567 3719 13912 ...
 $ PAY_AMT1                  : int  0 0 1518 2000 2000 2500 55000 380 3329 0 ...
 $ PAY_AMT2                  : int  689 1000 1500 2019 36681 1815 40000 601 0 0 ...
 $ PAY_AMT3                  : int  0 1000 1000 1200 10000 657 38000 0 432 0 ...
 $ PAY_AMT4                  : int  0 1000 1000 1100 9000 1000 20239 581 1000 13007 ...
 $ PAY_AMT5                  : int  0 0 1000 1069 689 1000 13750 1687 1000 1122 ...
 $ PAY_AMT6                  : int  0 2000 5000 1000 679 800 13770 1542 1000 0 ...
 $ default.payment.next.month: int  1 1 0 0 0 0 0 0 0 0 ...
```

Out of the 30,000 observations, the data has no missing values.

```
In [8]: sum(is.na(data))
```

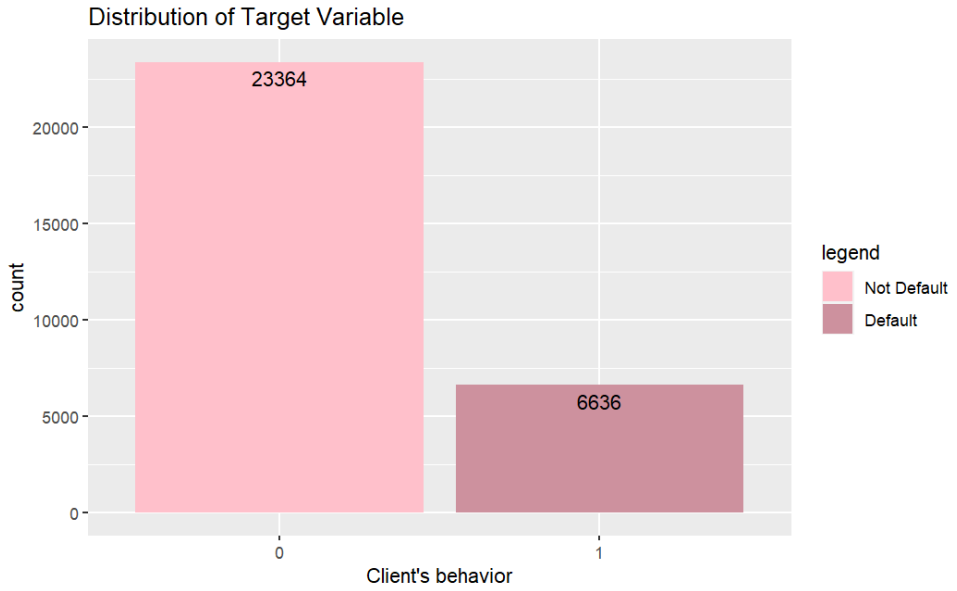
```
0
```

The 25 variables in the data set can be further split into continuous and categorical variables:

Continuous variables	Categorical variables
ID, LIMIT_BAL, AGE, BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6, PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6	*SEX, EDUCATION, MARRIAGE, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, default_payment_next_month*

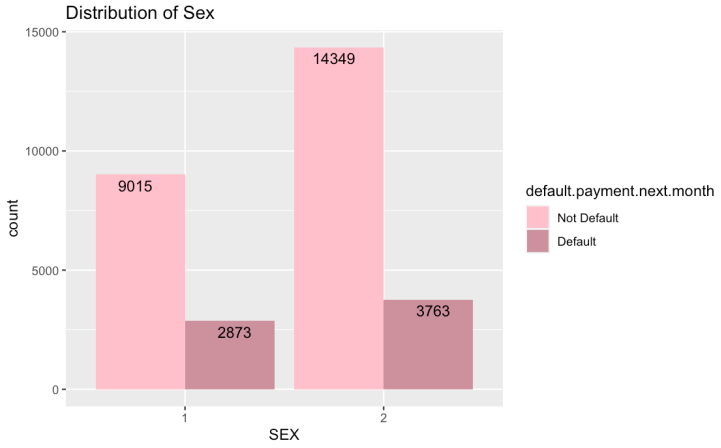
2.2 Distribution of Target Variable – Client's Behaviour (*'default_payment_next_month'*)

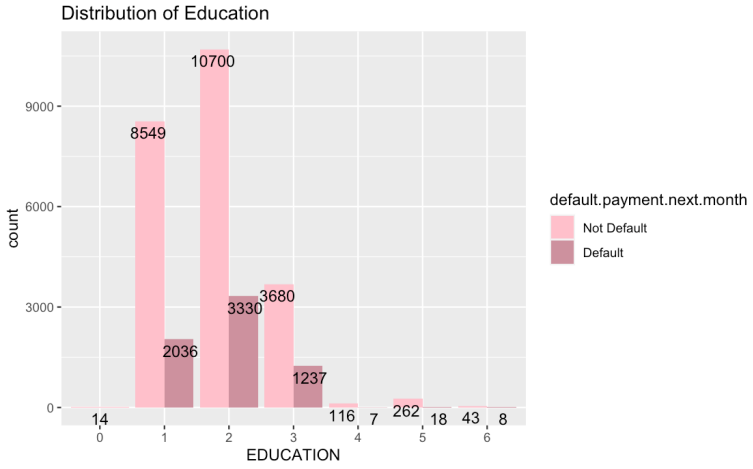
In order to visualise the distribution of the target variable, we plotted the following bar chart.

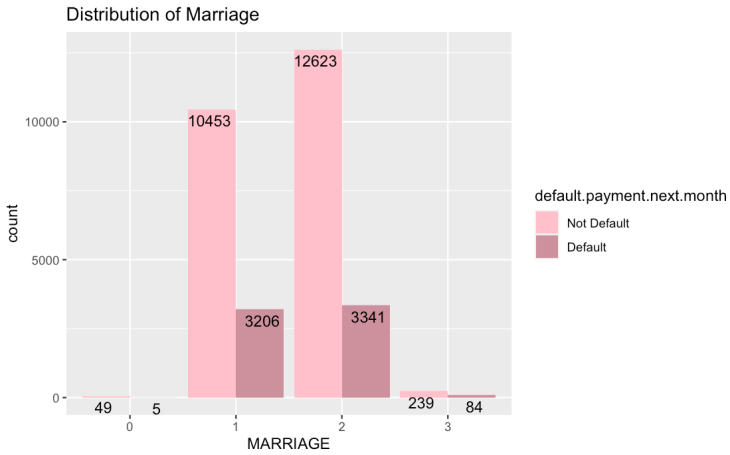


We can see from the above plot that there are 23,364 clients classified under *'Not Default'* and 6,636 clients classified under *'Default'*, which accounts for 22.1% of the total number of clients.

2.3 Distribution of Categorical Variables



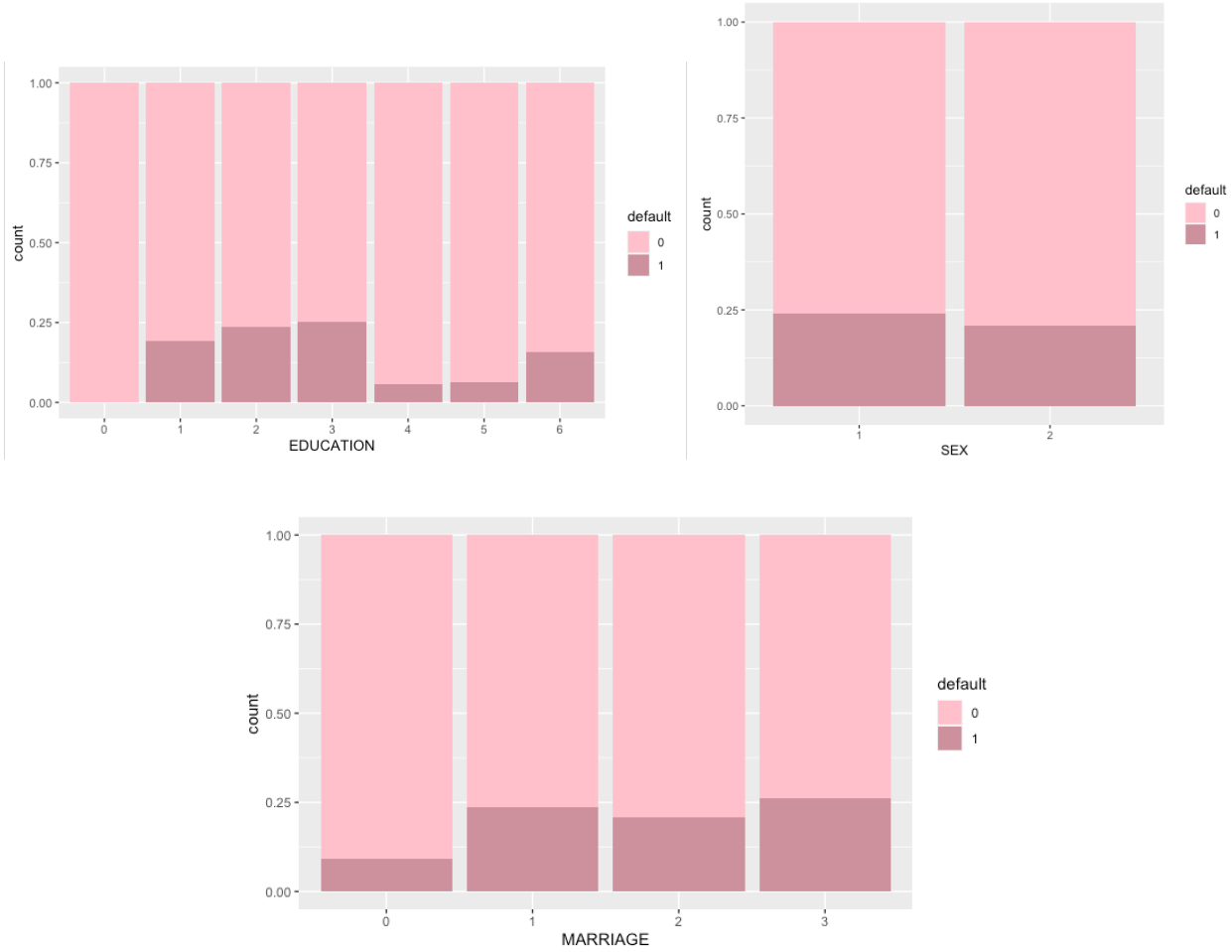




From the plots, we made the following observations:

1. All categorical variables have the same distribution between *'Default'* and *'Not Default'*.
2. There are more female (18,112) than male (11,888) clients in this dataset.
3. Most clients have a university (14,030) or graduate school (10,585) education level.
4. There are more single clients (15,964) than married clients (13,659) in this dataset.

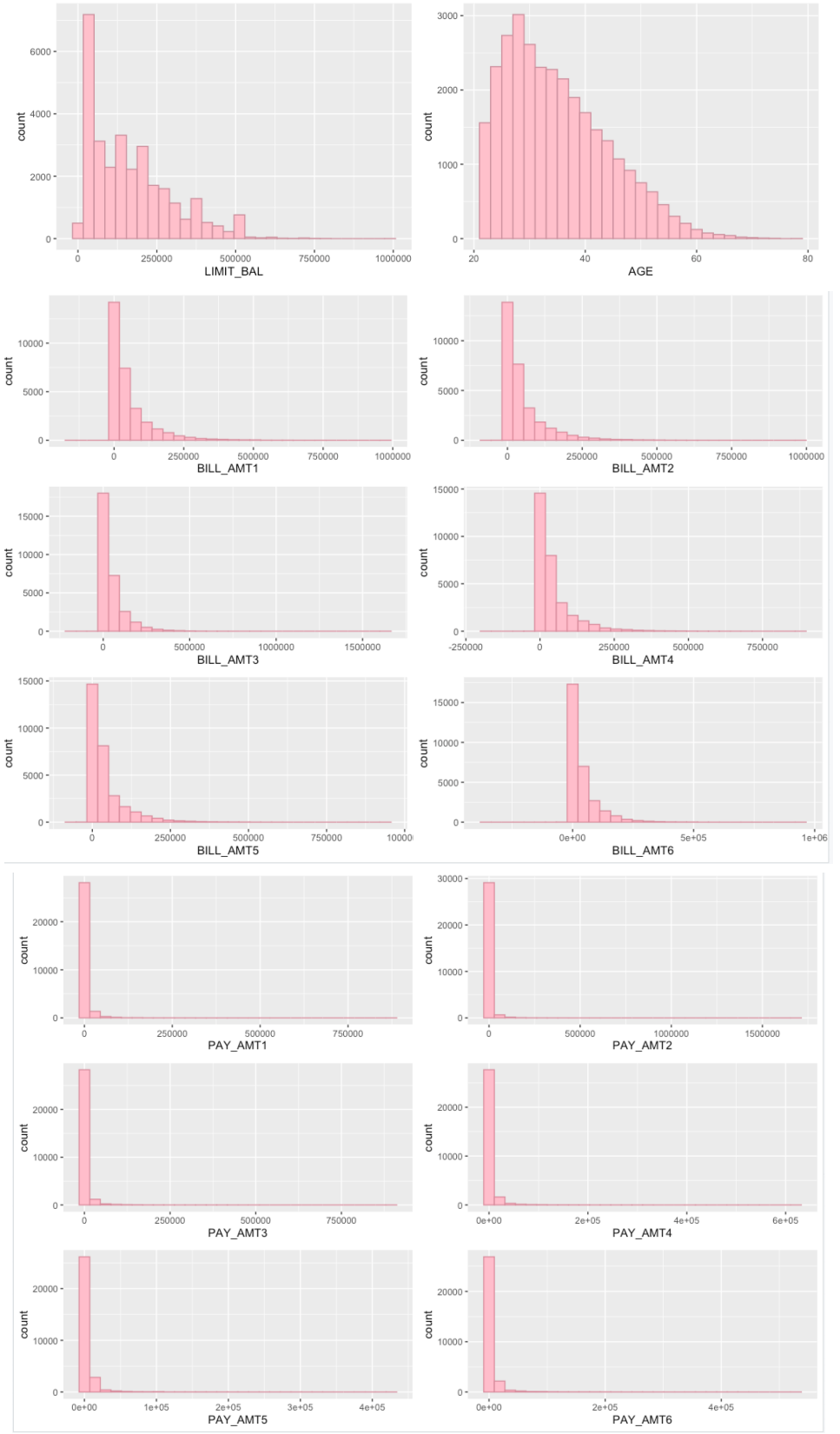
We have identified several inconsistencies in the dataset which will be addressed in Chapter 2.5.



The plots above show the ratio of the dependent variable, *'default_payment_next_month'*, for each categorical variable. From this, we can infer that:

1. Credit card holders with lower education levels are more likely to default their payments.
2. Males have a slightly higher tendency to default their payments compared to Females.
3. Married credit card holders are slightly more likely to default their payments compared to single card holders.

2.4 Distribution of Continuous Variables



From the histograms above, we can observe that the majority of the credit card holders:

- have a limit balance less than $50,000,
- are aged 25 to 45 years old,
- have a bill statement of less than $50,000 and
- made a previous statement of less than $5,000.

We can also take note that some of the bill statements recorded in the data have negative values.

2.5 Inconsistencies in Variables

2.5.1 V3: *'EDUCATION'*

According to the legend provided, *'EDUCATION'* is supposed to have the following values: 1 = Graduate; 2 = University; 3 = Highschool; 4 = Others. When unique values are returned, there are 0, 5 and 6 values that are unexplained.

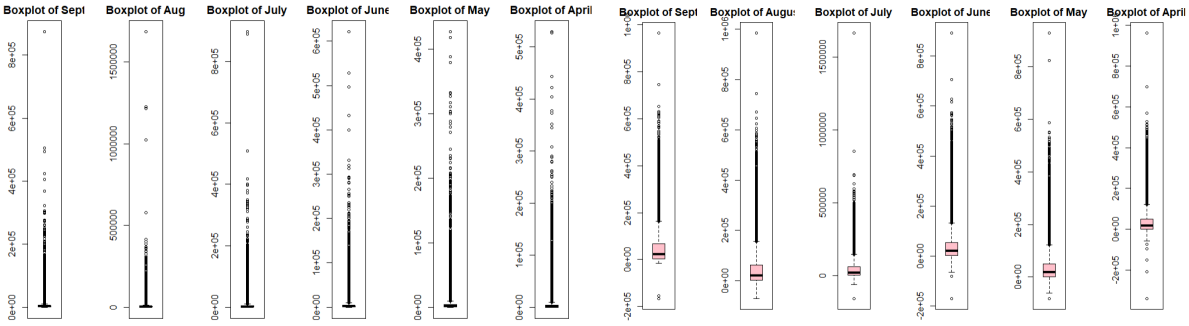
2.5.2 V4: *'MARRIAGE'*

According to the legend provided, *'MARRIAGE'* is supposed to have the following values: 1 = Married; 2 = Single; 3 = Others. When unique values are returned, there are 0 values that are unexplained.

2.5.3 V6-V11: *'PAY_0'* to *'PAY_6'*

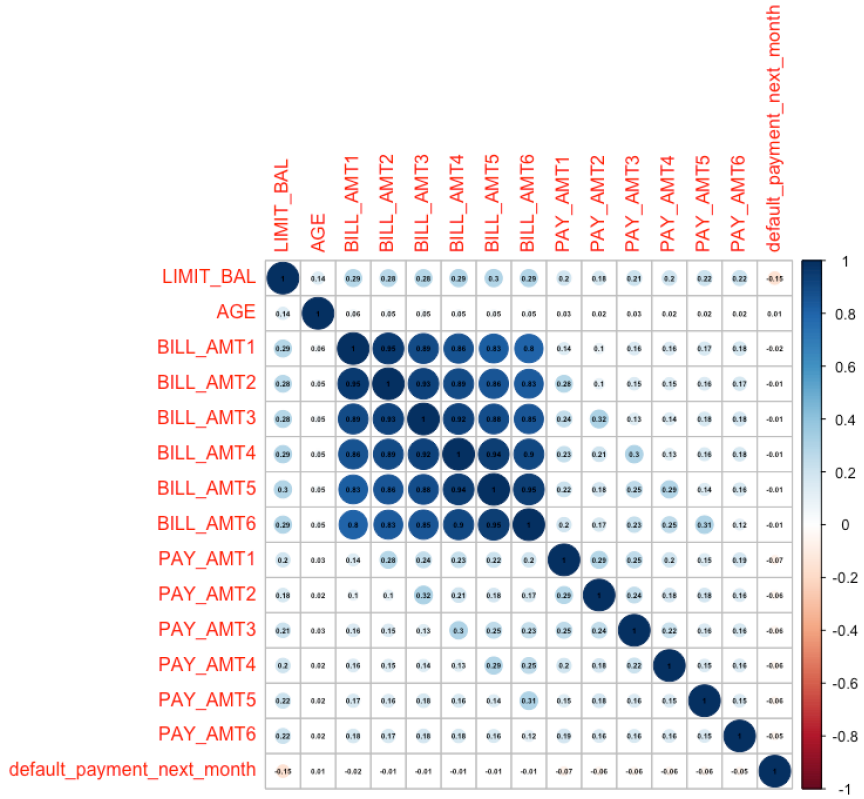
According to the legend provided, the range for *'PAY_0'* to *'PAY_6'* lies between [-1, 9], excluding 0. Upon returning unique values, values of 0 and -2 were observed.

2.6 Outliers in continuous variables



From the boxplots, we can see that there is a significant number of outliers. These outliers in '*PAY_AMT1'* to '*PAY_AMT6'* and '*BILL_AMT1'* to *'BILL_AMT6'* may not represent erroneous data, as there may be clients that are more affluent and are able to make such high payments.

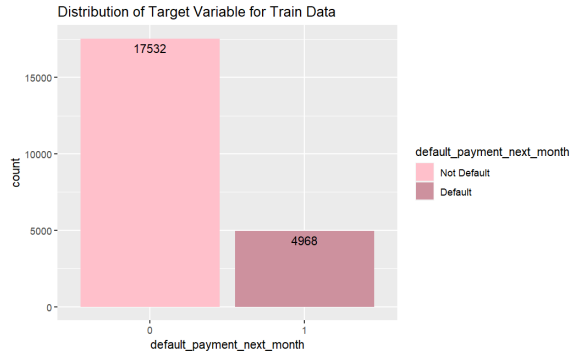
2.7 Correlation and Multicollinearity



From the corrplot, we can see that

- *'BILL_AMT1'* to *'BILL_AMT6'* are highly correlated with one another. This may result in multicollinearity which reduces the precision of the estimated coefficients and weakens the statistical power of our regression model. Hence, we can consider removing some of these features to prevent multicollinearity.
- All the continuous independent variables have little correlation to the target variable '*default_payment_next_month'.*

2.8 Class Imbalance



In the full dataset, there are 23,364 clients classified under *'Not Default'* (77.88%) and 6,636 clients classified under *'Default'* (22.12%). In the training dataset, there are 17,532 clients

classified under *'Not Default'* (77.92%) and 4,968 clients classified under *'Default'* (22.08%). Hence, our dataset is imbalanced in an approximate 8:2 ratio.

The large imbalance between *'Default'* and *'Not Default'* would lead to difficulties in predicting clients with *'Default'* as there is much less training data available for that. Furthermore, the large number of *'Not Default'* classification would cause inaccuracy in prediction as the models would achieve 77.88% accuracy in the event that all clients are predicted to be *'Not Default'* by the model. This is known as the Maximum Chance Criterion. This means there is a risk that the model is considered to be accurate even if it predicts the default status wrongly.

The Maximum Chance Criterion measures the proportion of the class with the largest size. Clearly, if we classified all individuals into the largest group, *'Not Default',* we could get a hit ratio of 77.88% without doing any work. Hence, one should have a hit rate of at least as much as the Maximum Chance Criterion rate.

As such, we will be evaluating our models based on average class accuracy as opposed to accuracy to account for these imbalances.

3. Data Pre-Processing

3.1 Making Categorical Variables factors

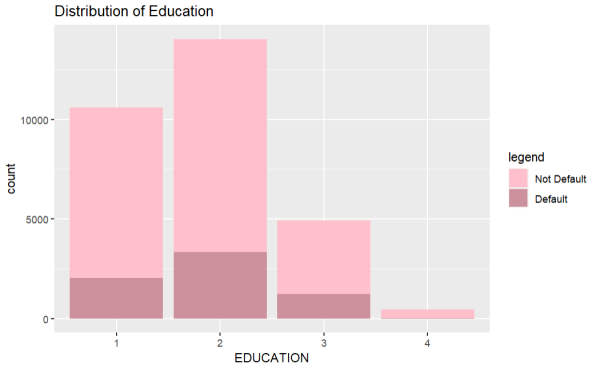
In order for categorical data to be represented as factors, we have changed the identified categorical values: *'SEX', 'EDUCATION', 'MARRIAGE', 'PAY_0' to 'PAY_6'*, originally represented by numbers to factors. As the outcome variable *'default_payment_next_month'* is also categorical, we have also converted it into a factor.

3.2 Re-categorising erroneous data

As mentioned previously in Chapter 2.5, we have decided to re-categorise the variables that have inconsistencies as per below.

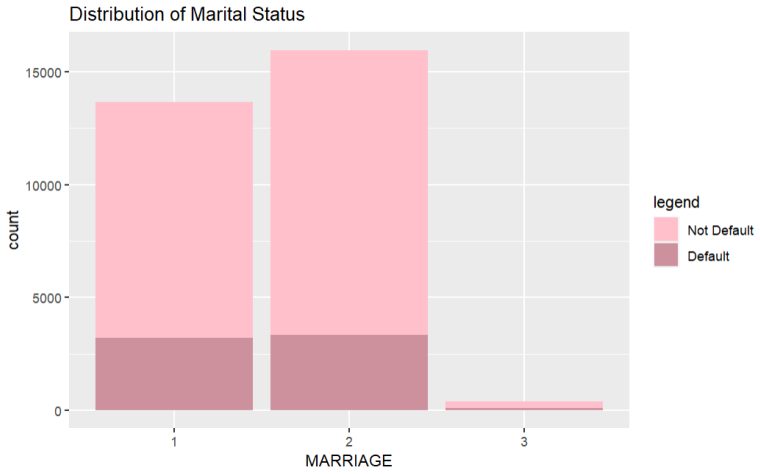
3.2.1 V3: *'EDUCATION'*

We have decided to classify the unexplained data values of 0, 5 and 6 values to 4 = Others.



3.2.2 V3: *'MARRIAGE'*

We have decided to classify the unexplained data values of 0 values to 3 = Others.



3.2.3 V3: *'PAY_0'* TO *'PAY_6'*

We have categorised values of -2, -1 and 0 to 0, meaning that payment is made on time.
For instance, the distribution for 'PAY_0' after recategorising is:

-2	-1	0	1	2	3	4	5	6	7	8
0	0	23182	3688	2667	322	76	26	11	9	19

3.3 Min-Max Scaling

This is the formula for min-max scaling which scales continuous variables for both train and test data between the range of 0 to 1:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

This allows for a fairer contribution from each variable equally as opposed to in the different models we are testing such as support vector models and logistic regressions. In addition, the data will end up with smaller standard deviations, which can suppress the effects of outliers.

4. Feature Selection

4.1 Filter Method

4.1.1 Selection using Statistical Tests (t-test and chi-square test)

For this filter method, the selection of features is independent of any machine learning algorithms. Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable. For the model that we are trying to predict, our target response variable is categorical while our independent variables contain both categorical as well as continuous variables.

For the independent categorical variables, which include education, sex, and marriage, we will use a Chi-Square test to conclude if these categorical variables are independent with whether or not customers default their payment the next month. If they are not independent, they should be included in the model as features.

For the independent continuous variables, we will use a t-test to conclude they are statistically significant in predicting whether or not customers default their payment the next month.

Included Variables	Excluded Variables
SEX, EDUCATION, MARRIAGE, LIMIT_BAL, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, BILL_AMT1, BILL_AMT2, BILL_AMT3, PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6	AGE, BILL_AMT4, BILL_AMT5, BILL_AMT6

4.2 Wrapper Method

4.2.1 Selection using Boruta Method

Boruta creates a set of shadow variables that shuffles each column randomly, creating a new column for each variable. This results in double the number of variables. After which, all these variables are used to create a random forest model that ranks each variable's importance. In comparison to the shadow variable's importance, a variable will be removed from the dataset if it is deemed to be less important. This process will repeat and end only when all variables have been deemed to be confirmed or rejected.

Included Variables	Excluded Variables
LIMIT_BAL, AGE, EDUCATION, MARRIAGE, PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6	SEX

4.2.2 Selection using Stepwise AIC Backward Regression

We build a regression model from a set of candidate predictor variables by removing predictors based on Akaike Information Criterion (AIC), in a stepwise manner until there is no variable left to remove any more.

Included Variables	Excluded Variables
LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_0, PAY_2, PAY_3, PAY_5, BILL_AMT2, BILL_AMT5, PAY_AMT1, PAY_AMT2	PAY_4 , PAY_6, BILL_AMT1, BILL_AMT3, BILL_AMT4, BILL_AMT6, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6,

4.2.3 Selection using Lasso Regression

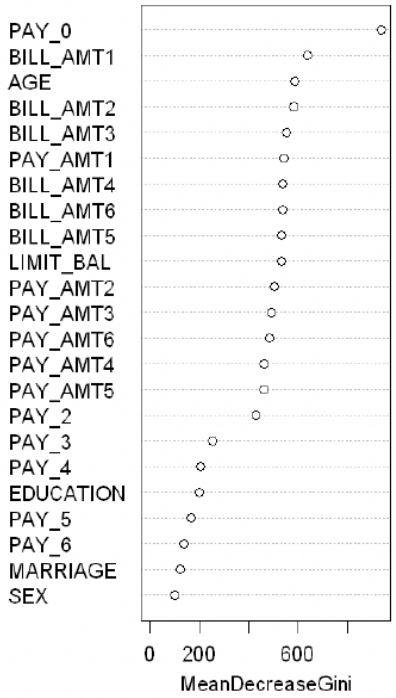
Lasso regression is a classification algorithm that uses shrinkage in simple and sparse models (i.e models with fewer parameters). In Shrinkage, data values are shrunk towards a central point such as the mean. If the predictor coefficient is shrunk all the way to zero, it means that the predictor is completely removed from the model because it does not impact the model enough.

Included Variables	Excluded Variables
LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, BILL_AMT1, PAY_AMT1, PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6	BILL_AMT2, BILL_AMT3, BILL_AMT4, BILL_AMT5, BILL_AMT6

4.3 Embedded Method

4.3.1 Selection using Random Forest

We used Random Forest to rank the variables according to their Mean Decrease Gini (IncNodePurity), a measure of variable importance based on the Gini impurity index used for calculating the splits in trees. From the plot, we include the top 10 ranking variables.



Included Variables	Excluded Variables
PAY_0, BILL_AMT1, AGE, BILL_AMT2, LIMIT_BAL, BILL_AMT3, PAY_AMT1, BILL_AMT4, BILL_AMT5, BILL_AMT6	PAY_AMT2, PAY_AMT3, PAY_AMT4, PAY_AMT5, PAY_AMT6, MARRIAGE, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, SEX, EDUCATION

5. Model Selection

5.1 Neural Network Model

Neural network is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain. Neural networks consist of node layers consisting of an input layer, one or more hidden layers, and an output layer. Each node connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated and will send data to the next layer of the network. Else, no data will be passed to the next layer of the network.

	Accuracy		Average Class Accuracy		Area Under ROC Curve	
Selection Method	Train Data	Test Data	Train Data	Test Data	Train Data	Test Data
All features	82.2%	81.41%	75.88%	73.47%	0.6518	0.6585
Filter Method	82.23%	81.37%	75.76%	73.36%	0.6547	0.659
Boruta Method	82.23%	81.33%	76.16%	73.37%	0.6499	0.6555
Stepwise AIC Backward Regression	82.14%	81.43%	75.33%	73.81%	0.6563	0.651
Lasso Regression	82.22%	81.41%	75.72%	73.47%	0.6546	0.6586
Random Forest	82.06%	81.4%	76.54%	75.18%	0.6412	0.63

Using features chosen by Random Forest gives the most accurate Neural network model based on average class accuracy.

5.2 Naive Bayes Model

Naïve Bayes is a machine learning algorithm based on the Bayes Theorem and probability, used for classification. It assumes that every feature is independent of one another.
Using the e1071 package's naive_bayes() function, we are able to create a model using the Naive Bayes algorithm and use that to calculate the relevant information below.

	Accuracy		Average Class Accuracy		Area Under ROC Curve	
Selection	Train Data	Test Data	Train Data	Test Data	Train	Test Data

Method					Data	
All features	76.83%	76.84%	67.41%	67.13%	0.7425	0.7374
Filter Method	76.84%	76.41%	67.43%	67.05%	0.7421	0.7361
Boruta Method	76.79%	76.41%	67.37%	67.06%	0.7402	0.7365
Stepwise AIC Backward Regression	78.25%	78.07%	68.75%	68.61%	0.747	0.7397
Lasso Regression	76.87%	76.51%	67.48%	67.17%	0.742	0.7348
Random Forest	80.95%	80.12%	72.26%	70.90%	0.7381	0.7289

Using features chosen by Random Forest gives the most accurate Naive Bayes model based on average class accuracy.

5.3 Logistic Regression Model

Logistic regression is an example of supervised learning. It is a statistical method used to calculate or predict the probability of a binary event occurring based on previous observations. It's a type of regression analysis and is a commonly used algorithm for solving binary classification problems. The logistic regression model computes a sum of the input features and calculates the logistic of the result. The output of logistic regression is always between 0 and 1, which is suitable for a binary classification task. An optimal probability threshold is then used to classify the logistic regression output. If the output is greater than or equal to the threshold, the output is classified as “1”, else it is classified as “0”.

	Accuracy		Average Class Accuracy		Area Under ROC Curve	
Selection Method	Train Data	Test Data	Train	Test Data	Train Data	Test Data
All features	82.26%	81.23%	75.76%	73.51%	0.6561	0.6453
Filter Method	82.28%	81.2%	75.66%	73.36%	0.6584	0.647

Boruta Method	82.26%	81.2%	76.13%	73.72%	0.6514	0.6397
Stepwise AIC Backward Regression	82.16%	81.39%	75.25%	73.63%	0.6589	0.6525
Lasso Regression	82.27%	81.2%	75.78%	73.43%	0.6563	0.6454
Random Forest	82.15%	81.37%	76.60%	74.82%	0.642	0.6327

Using features chosen by Random Forest gives the most accurate Logistic Regression model based on average class accuracy.

5.4 Support Vector Machine Model

The Support Vector Machine (SVM) is able to perform classification for both linearly and nonlinearly separable cases, depending on the set kernel in the function. Kernel function is used to transform the training set of data from a non-linear decision surface to a linear equation with a higher number of dimensions.

Considering that we are unable to determine whether our data set is linearly or nonlinearly separable, we will perform SVMs with different kernels first to account for both cases on all the variables in the data set without removing any features to decide on the most suited kernel. Following which, we will select the best kernel to be used with the dataset before applying it on the selected features.

	Accuracy		Average Class Accuracy	
Different Kernels	Train Data	Test Data	Train Data	Test Data
Linear	82.32%	81.53%	66.40%	65.95%
Polynomial	82.45%	81.08%	65.16%	63.65%
Radial Basis	82.71%	81.41%	66.61%	65.06%
Sigmoid	73.68%	72.64%	60.02%	59.04%

From the above results, we are able to determine that the Radial Basis kernel is the most suited for this data set through comparing the average class accuracy.

Using Radial Basis kernel:

	Accuracy	Average Class Accuracy

Selection Method	Train Data	Test Data	Train Data	Test Data
All features	82.71%	81.41%	66.61%	65.06%
Filter Method	82.69%	81.44%	66.61%	65.08%
Boruta Method	82.73%	81.36%	66.61%	65.07%
Stepwise AIC Backward Regression	82.45%	81.49%	65.87%	65.01%
Lasso Regression	82.69%	81.35%	66.67%	65.06%
Random Forest	82.20%	81.37%	64.35%	63.14%

Using features chosen by Lasso Regression gives the most accurate SVM model based on average class accuracy.

6. Model Evaluation

From Chapter 5, we have the Accuracy, Average Class Accuracy and area under ROC curve of the models. To improve our evaluation, we included the use of F1-Score and Kappa Statistic.

6.1 F1-Score

F1-Score is the weighted average of Precision and Recall. Thus, this score takes both false positives and false negatives into account. It performs well when dealing with imbalanced data. A higher score would indicate high Precision and Recall because it gives a balanced weight to both Precision and Recall.

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

6.2 Kappa Statistic

We have also included the Kappa Statistic, which takes into account how closely the predicted values are to the actual values, controlling for any true positives or true negatives that occur by chance. A score of 1 means that there is perfect agreement between the predicted and actual values, and a score of 0 means that there is no agreement between the predicted and actual values.

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

6.3 Final summary of evaluations

Models	Feature Selection Method	Average Class Accuracy		F1-Score		Kappa Statistic	
		Train Data	Test Data	Train Data	Test Data	Train Data	Test Data
Neural Network	Random Forest	76.54%	75.18%	0.456	0.432	0.363	0.338
Naive Bayes	Random Forest	72.26%	70.90%	0.500	0.488	0.386	0.368
Logistic Regression	Random Forest	76.60%	74.82%	0.457	0.460	0.365	0.343
SVM	Lasso Regression	66.67%	65.06%	0.492	0.460	0.399	0.358

From the table above, the Neural Network model (using the features selected by Random Forest) produces the highest average class accuracy for the training and test datasets while the Naive Bayes model (using the features selected by Random Forest) produces the highest F1-Score and Kappa Statistics for the training and test datasets.

As previously mentioned in Chapter 2.8, the dataset is highly imbalanced in terms of the number of default and not default clients As such, as opposed to considering just the accuracy of the models, we consider Average Class Accuracy, F1-Score and Kappa Statistics as a suitable evaluation metric in this case.

From the table above, we can see that Naive Bayes model with the Random Forest method for feature selection yields the highest results comparatively for F1-score and Kappa Statistic. Although the Average Class Accuracy for the Naive Bayes model falls slightly short in terms of accuracy, in this context of helping the Bank to predict its clients' default payments, a greater emphasis should be placed on making less false negative predictions (predicted: *'Non-Default'*, Actual: *'Default'*).

Thus, we conclude that the Naive Bayes model is the best model for predicting the probability of default for a credit card holder.

7. Limitations and Improvements

7.1 Imbalance in PAY_2, PAY_4 to PAY_6 factor levels in train and test set

The factor levels in PAY_0 and PAY_6 are significantly different as shown below:

```
 [1] -2  0 -1  2  3  5  4  7  1  6
 [1]  2  0 -1 -2  3  5  7  1  4  6  8
 [1] -2  2  0 -1  3  5  7  1  4  6
 [1] -1  0 -2  2  3  4  5  7  6  1  8
[1] -2  2  0 -1  4  3  7  5  6
 [1] -2  0 -1  2  3  5  4  7  8  6
[1] -2  2  0 -1  3  6  4  5  7
 [1] -2  2  0 -1  3  6  4  7  8  5
```

When we try to make the aforementioned variables factors, it results in factors with different levels that causes models to not be able to run, for example SVM. In order to account for that, SVM does not take into account the variables above as factors, but as values.

To improve this, we could have a more random split for train and test datasets to ensure that provides an equal level of factors for the categorical variables.

7.2 Multicollinearity

From Chapter 2.7, we can see that '*BILL_AMT1'* to *'BILL_AMT6'* are highly correlated with one another which may result in multicollinearity. Hence, we may need to remove some of these features to prevent multicollinearity. However, feature selection using Filter Method will include '*BILL_AMT1', 'BILL_AMT2'* and *'BILL_AMT3'* as important variables and feature selection using Boruta Method will include all '*BILL_AMT1'* to *'BILL_AMT6'* as important variables. As such, the statistical significance of the independent variable may be undermined due to the presence of multicollinearity.

To improve this, we could remove more of the highly correlated variables. We can do so by performing analysis designed for highly correlated variables, such as Principal Components Analysis.

7.3 Limitations of Models

Model	Limitations
Neural Network Model	Neural network composed of many interconnected processing nodes. Hence, it is difficult to determine what is the proper network structure and understand how the node weights result in the predicted output.
Naive Bayes Model	Naive Bayes assumes all variables are independent. However, the features in this dataset are not independent of each other, as seen from the correlation matrix as there is still some correlation between each variable.
Logistic Regression Model	Logistic Regression assumes linearity between the dependent variable and the independent variables, which we are unable to confirm. It also requires observations to be independent of one another. Furthermore, as the predictions from this model are based on the independent variables, if they are not properly identified, the model will have little predictive value.

SVM	SVM does not work very well when the dataset is noisy. As the support vector classifier works by placing data points above and below the hyperplane, there is no probabilistic clarification for the classification.

In addition, in order to compare model performances more fairly and avoid overfitting, k-fold cross-validation could be implemented to ensure that no lucky splits occur.

7.4 Limitation in Dataset

As seen from the correlation matrix in chapter 2.7, most of the independent variables have a small correlation with the target variable. Hence, variables with a possibly higher correlation with the target variable could have been added to the dataset such as client's income, types of loan and length of credit history. These variables will be able to help predict whether the client defaults the payment more effectively.

8. References

Advantages and Disadvantages of ANN in Data Mining. (2021, July 7). GeeksforGeeks. Retrieved November 15, 2022, from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-ann-in-data-mining/

Advantages and Disadvantages of Logistic Regression. (2022, August 23). GeeksforGeeks. Retrieved November 15, 2022, from https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/

Amazon AWS. (n.d.). *What is a Neural Network? AI and ML Guide - AWS*. Amazon AWS. Retrieved November 15, 2022, from https://aws.amazon.com/what-is/neural-network/

Chauhan, N. S. (2022, April 8). *Naïve Bayes Algorithm: Everything You Need to Know*. KDnuggets. Retrieved November 15, 2022, from https://www.kdnuggets.com/2020/06/naive-bayes-algorithm-everything.html

default of credit card clients Data Set. (2016, January 26). UCI Machine Learning Repository. Retrieved November 15, 2022, from https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Kernel SVM - machine learning in R. (2019, May 19). RPubs. Retrieved November 15, 2022, from https://rpubs.com/markloessi/497544

Major Kernel Functions in Support Vector Machine (SVM). (2022, February 7). GeeksforGeeks. Retrieved November 15, 2022, from https://www.geeksforgeeks.org/major-kernel-functions-in-support-vector-machine-svm/

Meyer, D., Chang, C., & Lin, C. (n.d.). *svm function*. RDocumentation. Retrieved November 15, 2022, from https://www.rdocumentation.org/packages/e1071/versions/1.7-12/topics/svm

Stewart, Z. (2022, April 12). *Building a Student Intervention System | Machine Learning, Deep Learning, and Computer Vision*. Ritchie Ng. Retrieved November 15, 2022, from https://www.ritchieng.com/machine-learning-project-student-intervention/

Support vector machine in Machine Learning. (2020, December 22). GeeksforGeeks. Retrieved November 15, 2022, from https://www.geeksforgeeks.org/support-vector-machine-in-machine-learning/

3.4 Using Colors in a Bar Graph. (n.d.). R Graphics Cookbook. Retrieved November 15, 2022, from https://r-graphics.org/recipe-bar-graph-colors

What are Neural Networks? - Singapore. (2020, August 17). IBM. Retrieved November 15, 2022, from https://www.ibm.com/sg-en/cloud/learn/neural-networks

Yang, Z. (2022, May 1). *Disadvantages of Artificial Neural Networks And Workarounds*. The Analytics Club. Retrieved November 15, 2022, from https://www.the-analytics.club/disadvantages-of-artificial-neural-networks

The F1 score | towards data science. (n.d.). Retrieved November 18, 2022, from https://towardsdatascience.com/the-f1-score-bec2bbc38aa6

Cohen Kappa score python example: Machine learning. Data Analytics. (2022, June 13). Retrieved November 18, 2022, from https://vitalflux.com/cohen-kappa-score-python-example-machine-learning/

Variable selection methods. (n.d.). Retrieved November 18, 2022, from https://cran.r-project.org/web/packages/olsrr/vignettes/variable_selection.html

INSEAD course: Data Analytics for Business. Insead_Analytics. (n.d.). Retrieved November 18, 2022, from http://inseaddataanalytics.github.io/INSEADAnalytics/

Team, G. L. (2022, October 31). *A complete understanding of lasso regression*. Great Learning Blog: Free Resources what Matters to shape your Career! Retrieved November 18, 2022, from https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/

Frost, J. (2022, July 22). *Multicollinearity in regression analysis: Problems, detection, and solutions*. Statistics By Jim. Retrieved November 18, 2022, from https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/