Statistical Modelling III - Group Assignment

1148120 - M. Vincent, 1721235 - S. Carnie-Bronca,

Introduction

Managing the risk of default has always been a crucial aspect business for banks and other institutions offering credit card loans. Although in most jurisdictions there are legal consequences discouraging credit card default, it is still a critical issue that lenders have to contend with. Predicting whether or not a given borrower is likely to default on their credit card debt is complex problem. Statistical models can play an important role in minimising the risk of default. Lenders can use information about previous customers, including whether or not they have defaulted on their loans, to assess the likelihood that a potential new customer will default, given the known characteristics of that that individual.

Predictive modelling is the process of applying statistical techniques to reveal the relationship between a response variable and one or more predictor variables, inclulding whether a relationship exists at all, in order to make predictions given new data. There are a great number statistical techniques that can be used for this purpose, however most of these techniques will fall into one of two categories; regression or classification. In regression models, the response is a continuous variable. In classification models, the response is a categorical variable. When the response can only take two possible values, the problem is one of *binary classification*. Predicting whether or not a customer will default on their credit card loans is an example of a binary classification problem. In this project, we use a dataset consisting of 20,000 individuals, with their financial history, demographics and whether or not they have defaulted, and use this data to train a model to predict whether a new customer is likely to default on a loan.

To generate the model, we used `caret`, a machine learning package for R. Functions from the `caret` package were used to help determine which predictors should be used in the model, and then to train and evaluate potential models. Before we began training models, we first performed exploratory data analysis. Sections 2 and 3 descibe the distributions of each variable in the data set, as well as relationships between predictor variables. This preliminary analysis allowed for a more efficient model selection process, as described in Section 4. The performance of several models that we considered is included in this section, with the best model presented in Section 5.

1. Cleaning and Preparing the Data

By inspecting the data, and using the `is.na` command on each variable, it was determined that there were no missing values in the data, so cleaning was not required. Next, each variable was inspected to ensure it is of the correct type. The data type for each variable is displayed in Table 1.

As shown, all variables are currently considered integers. This is not correct, as the only numeric variables are `LIMIT_BAL`, `AGE`, `BILL_AMT` and `PAY_AMT`. The other variables were therefore converted to factors. The converted variable types, and a brief description of the data, are displayed in Table 2.

2. Univariate Analysis

2.1. Variable Descriptions

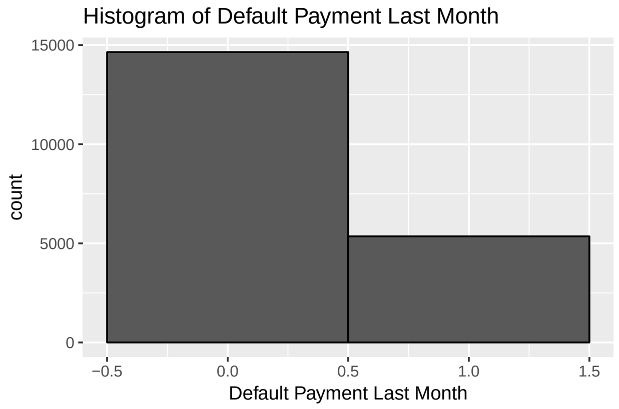
Table 1: Converted Variable Types

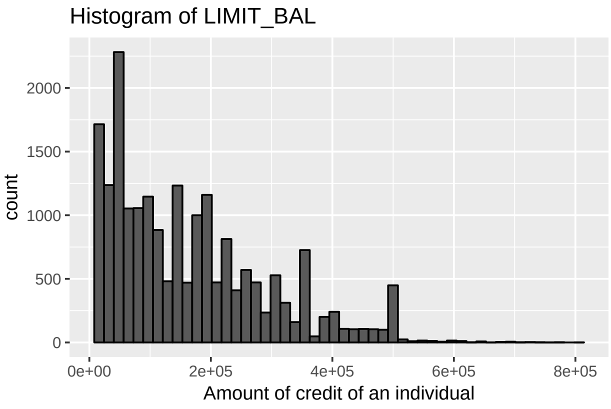
	Raw Data Type	Converted Data Type
LIMIT_BAL	integer	integer
SEX	integer	factor
EDUCATION	integer	factor
MARRIAGE	integer	factor
AGE	integer	integer
PAY_0	integer	factor
PAY_2	integer	factor
PAY_3	integer	factor
PAY_4	integer	factor
PAY_5	integer	factor
PAY_6	integer	factor
BILL_AMT1	integer	integer
BILL_AMT2	integer	integer
BILL_AMT3	integer	integer
BILL_AMT4	integer	integer
BILL_AMT5	integer	integer
BILL_AMT6	integer	integer
PAY_AMT1	integer	integer
PAY_AMT2	integer	integer
PAY_AMT3	integer	integer
PAY_AMT4	integer	integer
PAY_AMT5	integer	integer
PAY_AMT6	integer	integer
y	integer	factor

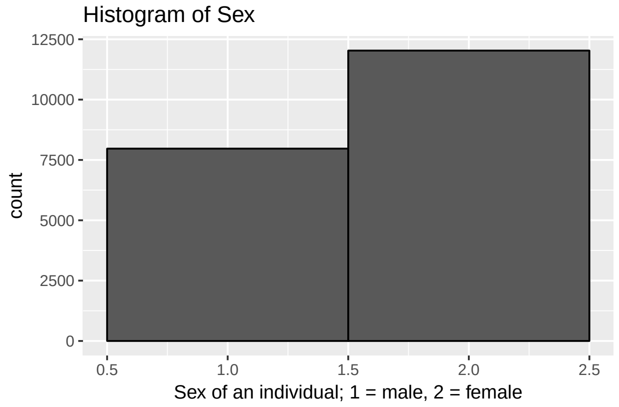
Table 2: Description of variables in the data set.

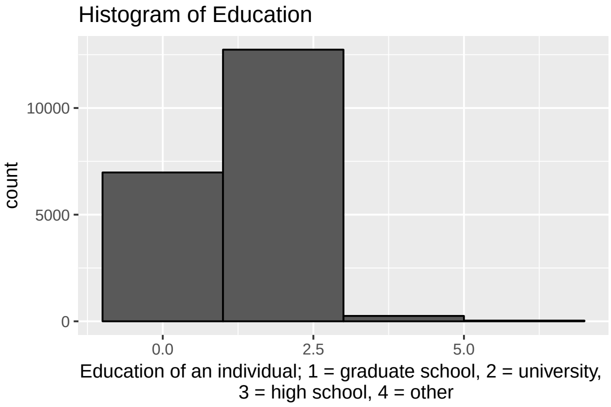
Variable Name	**Data Type**	**Role in model**	**Description**
default payment next month	Factor	Response	1 = a default payment, 0 = no default
LIMIT_BAL	Numeric	Predictor	Amount of credit of an individual, in NT dollars
SEX	Factor	Predictor	Sex of an individual; 1 = male, 2 = female
EDUCATION	Factor	Predictor	Education status of an individual; 1 = graduate school, 2 = university, 3 = high school, 4 = other education
MARRIAGE	Factor	Predictor	Martial status of an individual; 1 = married, 2 = single, 3 = other
AGE	Numeric	Predictor	Age of an individual
PAY_0 to PAY_6	Factor	Predictor	History of payment of an individual, from April (PAY_6) to September (PAY_0) 2015; -1 = on time, other values are months of delay in repayment
BILL_AMT1 to BILL_AMT6	Numeric	Predictor	Amount of bill statement, from April (BILL_AMT6) to September (BILL_AMT1) 2015, in NT dollars
PAY_AMT1 to PAY_AMT6	Numeric	Predictor	Amount of previous payment, from April (PAY_AMT6) to September (PAY_AMT1) 2015, in NT dollars

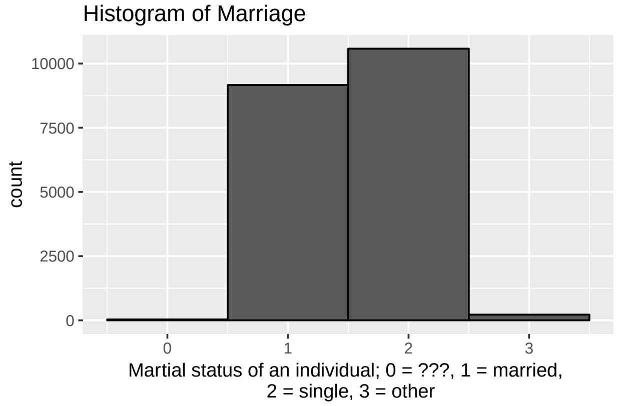
2.3. Univariate Plots

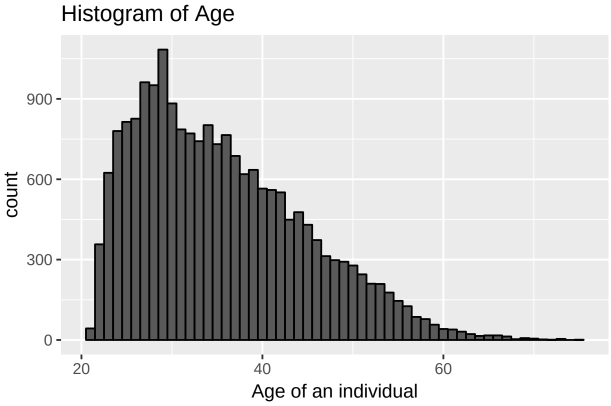












3. Bivariate Analysis

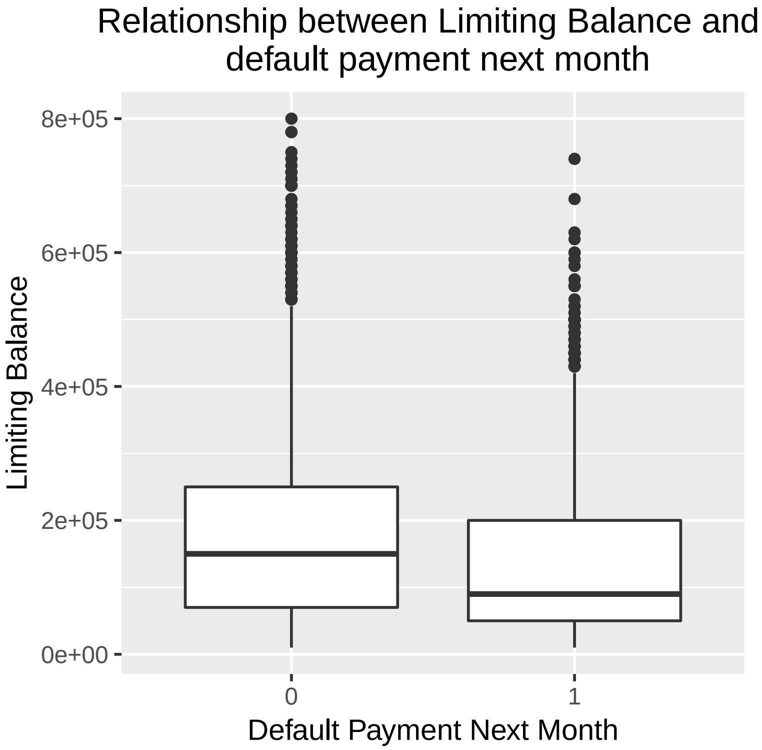
In order to determine the nature and strength of the relationship between each predictor variable and the response variable, default payment next month, plots of each of these relationships will be produced and investigated. Determing the relative associations of these predictor variables is important, as it allows us to consider the implication of their inclusion in the final model.

3.1. Continous Variables

We will begin by considering the relationship of the continuous predictor variables against default payment next month. These relationships will be analysed using side-by-side box plots.

3.1.1. Limiting Balance against default payment next month

Given limiting balance is the amount of credit available to an individual, in NT dollars, the main areas of interest will be whether having a higher or lower limiting balance will influence defaulting next month's payment. We will investigate this using side-by-side box plots, shown below.



From observing the box plots above, it is clear that both levels of default payment next month are positively skewed. We also see that the medians for both levels lie marginally closer to the lower quartile value than the upper. Although the upper and lower quartile values for those who do not default next month's payment lie above those who do respectively, the interquartile range for both levels overlap. Consider the median values for both levels shown below.

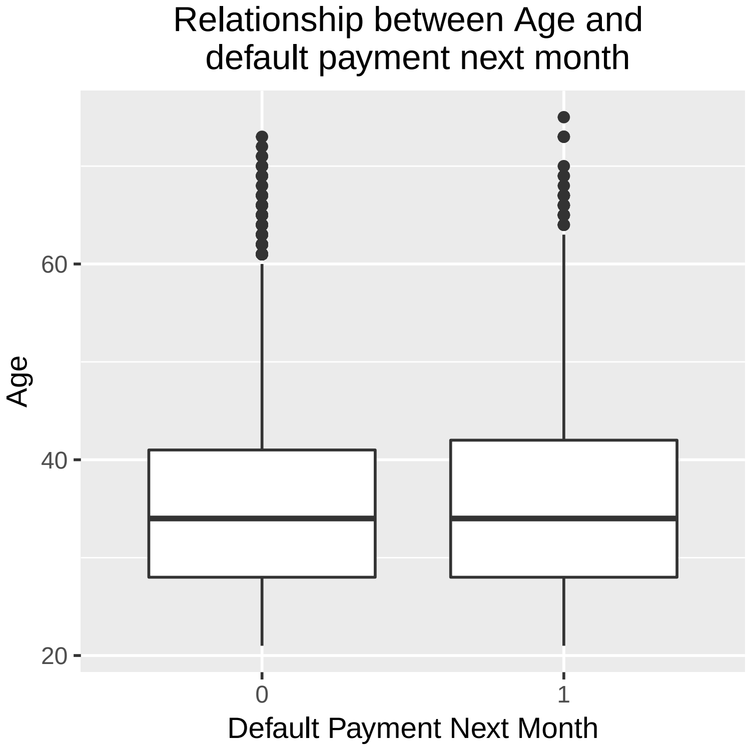
```
## [1] 150000
```

```
## [1] 90000
```

The median limiting balance for those who do not default next month's payment is 150000, and 90000 for those who do default next month's payment. Thus it could be suggested that having a larger limiting balance amount may decrease the likelihood of defaulting next month's payment, however there is not sufficiently strong evidence to definitively conlude this. It should also be noted that both levels have potential outlier candidates.

3.1.2. Age against default payment next month

The next variable to be considered is Age, another integer variable taking values between 21 and 75. The relationship between Age and default payment next month will be investigated through side-by-side box plots.



From observing the side-by-side box plots, we notice that both are positively skewed. Furthermore, the spread and median values for both levels of default payment next month are very similar, suggesting that there is no apparent trend for defaulting next month's payment based on Age.

```
##   0%  25%  50%  75% 100%
##   21   28   34   41   73
```

```
##   0%  25%  50%  75% 100%
##   21   28   34   42   75
```

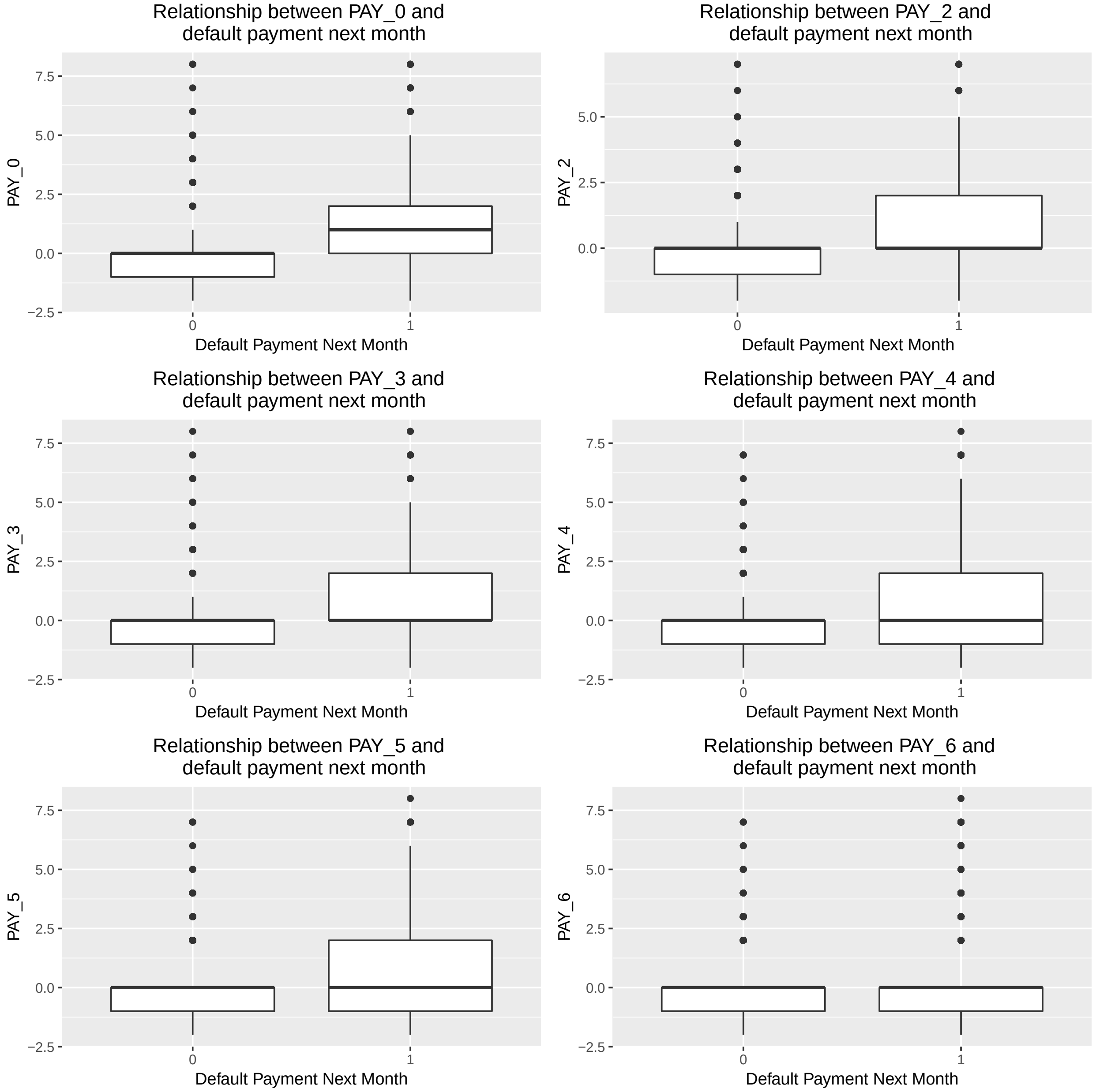
Upon further inspection, the median values and the lower quartile values are the same for both levels, begin 34 and 28 respectively. Additionally, the upper quartile values for those who do default next month's payment and those who do not are 42 and 41 respectively. This further supports the suggestion that age does not influence whether or not next month's payment is defaulted. There are also potential outlier candidates for both levels.

3.1.3. Pay_X against default payment next month

PAY_X is a measure of an individual's payment history, with:

1. PAY_6 corresponding to the individuals payment in April.
2. PAY_5 corresponding to the individuals payment in May.
3. PAY_4 corresponding to the individuals payment in June.
4. PAY_3 corresponding to the individuals payment in July.
5. PAY_2 corresponding to the individuals payment in August.
6. PAY_0 corresponding to the individuals payment in September.

Each PAY_X is an integer variable, taking values between -2 and 8 (corresponding to month's the payment of that month was overdue). Hence, the relationship between Pay_X and default payment next month will be investigated using side-by-side boxplots.



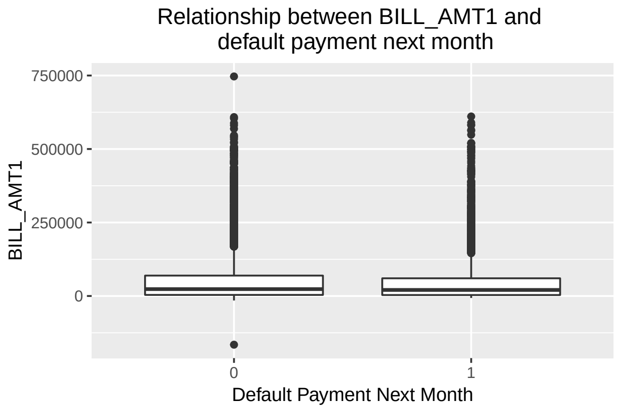
Observing these plots, we see that they are all positively skewed, with relatively low median values for both factor levels. Furthermore, other than for PAY_0, the median for both levels are equal. However, the upper quartile for all plots, other than PAY_6, is larger for those who do default next month's payment. Intuitively this makes sense, as we would expect defaulting next month's payment to be more likely if previous month's payments have also been late. There are also outlier candidates present in each of the plots.

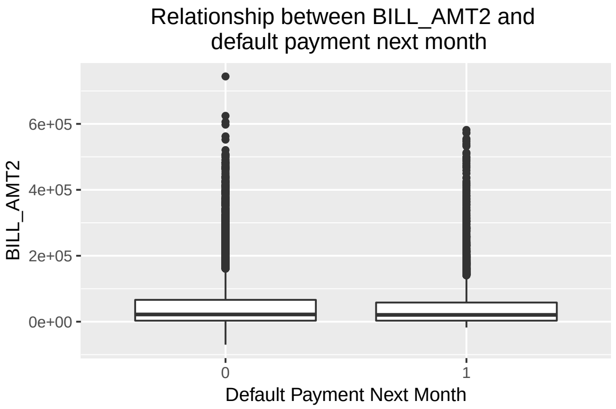
3.1.4. BILL_AMTX against default payment next month

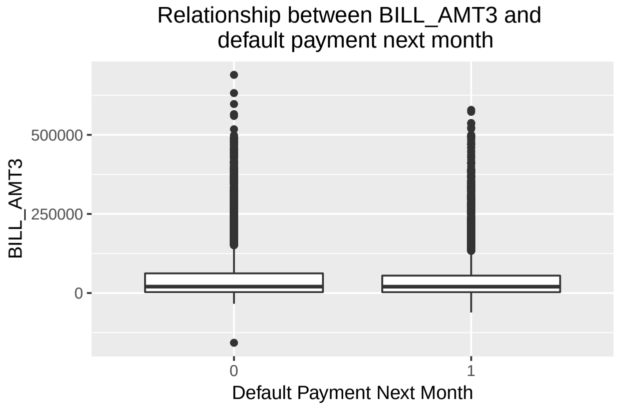
BILL_AMTX is the amount appearing on the bill statement as follows:

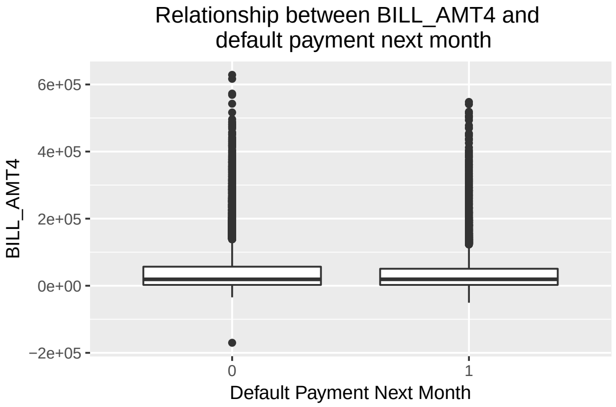
1. BILL_AMT6 corresponding to the individuals statement in April.
2. BILL_AMT5 corresponding to the individuals statement in May.
3. BILL_AMT4 corresponding to the individuals statement in June.
4. BILL_AMT3 corresponding to the individuals statement in July.
5. BILL_AMT2 corresponding to the individuals statement in August.
6. BILL_AMT1 corresponding to the individuals statement in September.

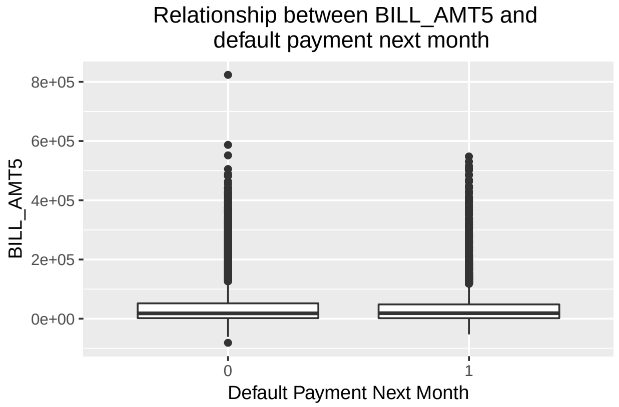
As BILL_AMTX is a numeric variable, we will investigate its relationship with default payment next month using side-by-side boxplots.

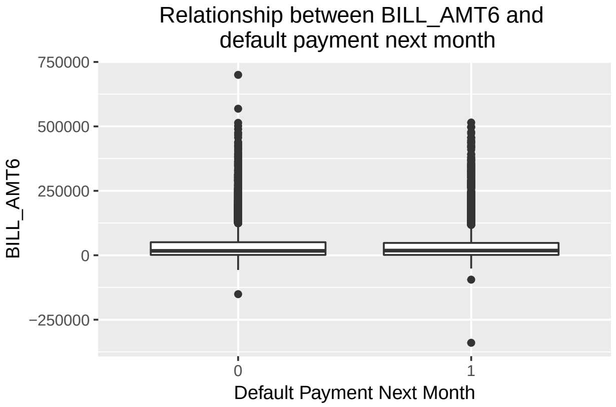












We see that all of these plots are positively skewed. Additionally, all side-by-side boxplots show that the median values for both levels of default payment next month lie withtin very close proximity of each other. Furthermore, there is also close similarities with the interquartile spread in each plot.

```
## [1] 23462.5 21907.5 20243.5 19102.5 18026.5 17006.5
```

```
## [1] 20588.5 20629.5 20053.5 19249.0 18574.0 18130.5
```

Inspecting the median values for the 6 plots shows that the median value for BILL_AMT1, BILL_AMT2 AND BILL_AMT3 are higher for those who do not default next month's payment, where as this is reversed for BILL_AMT4, BILL_AMT5 and BILL_AMT6. This suggests that having a higher bill amount in recent month's will, in general, reduce the likelihood of defaulting next month's payment. However, the overlap in interquartile spread suggests that this influence is minimal. There are once again outlier candidates for both levels of default payment next month.

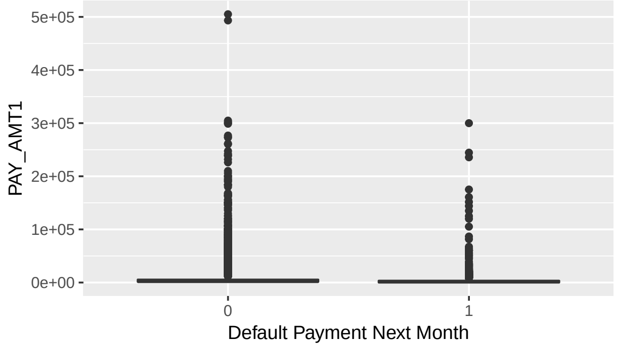
3.1.5. PAY_AMTX against default payment next month

PAY_AMTX is the amount of the previous payment, with:

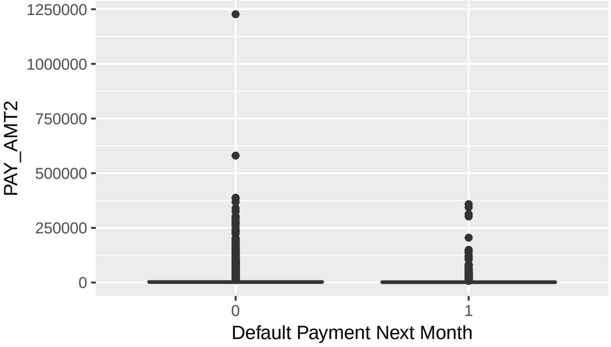
1. PAY_AMT6 corresponding to the payment made in April.
2. PAY_AMT5 corresponding to the payment made in May.
3. PAY_AMT4 corresponding to the payment made in June.
4. PAY_AMT3 corresponding to the payment made in July.
5. PAY_AMT2 corresponding to the payment made in August.
6. PAY_AMT1 corresponding to the payment made in September.

As PAY_AMTX is a numeric variable, we will investigate its relationship with default payment next month using side-by-side boxplots.

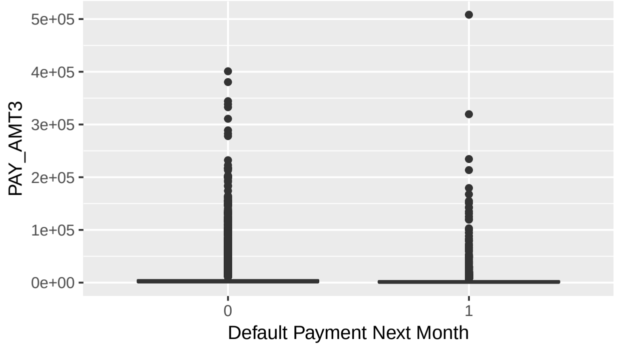
Relationship between PAY_AMT1 and default payment next month



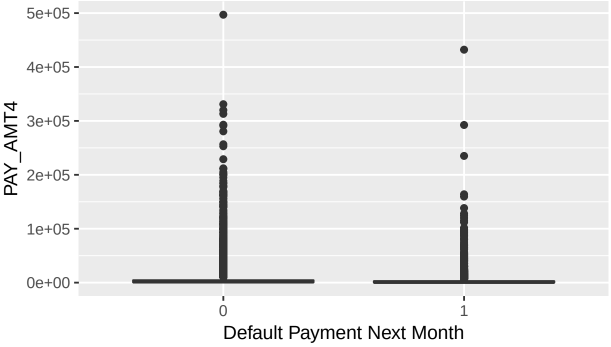
Relationship between PAY_AMT2 and default payment next month



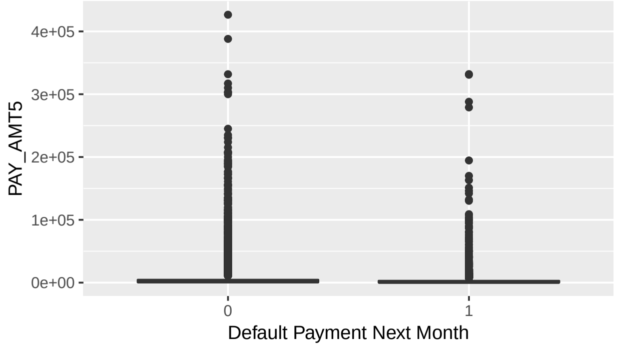
Relationship between PAY_AMT3 and default payment next month



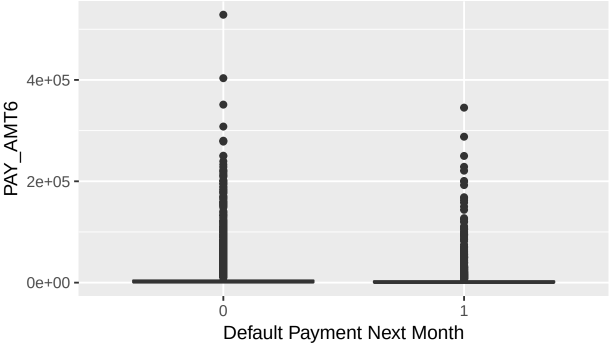
Relationship between PAY_AMT4 and default payment next month



Relationship between PAY_AMT5 and default payment next month



Relationship between PAY_AMT6 and default payment next month



We see that all of these plots are positively skewed. Additionally, all side-by-side boxplots show that the median, lower quartile and upper quartile values lie within very close proximity.

```
## [1] 2500.0 2244.0 2000.0 1780.0 1800.5 1766.5
```

```
## [1] 1639.5 1591.0 1250.0 1000.0 1000.0 1000.0
```

Upon further inspection, the median values are marginally larger for those who do not default next month's payment, so there may be a very small influence. There are outlier candidates for both levels of default payment next month.

3.2. Categorical Variables

We now consider the relationship between the categorical predictor variables and the response variable. These will be explored using side-by-side box plots where appropriate, otherwise analysing the mean values.

3.2.1. Sex against default payment next month

The next variable to be considered is Sex. As Sex and default payment next month are both integer variables with two levels (0 and 1 for default payment next month, 1 and 2 for sex) their relationship will be investigated by considering the default rates for both males and females.

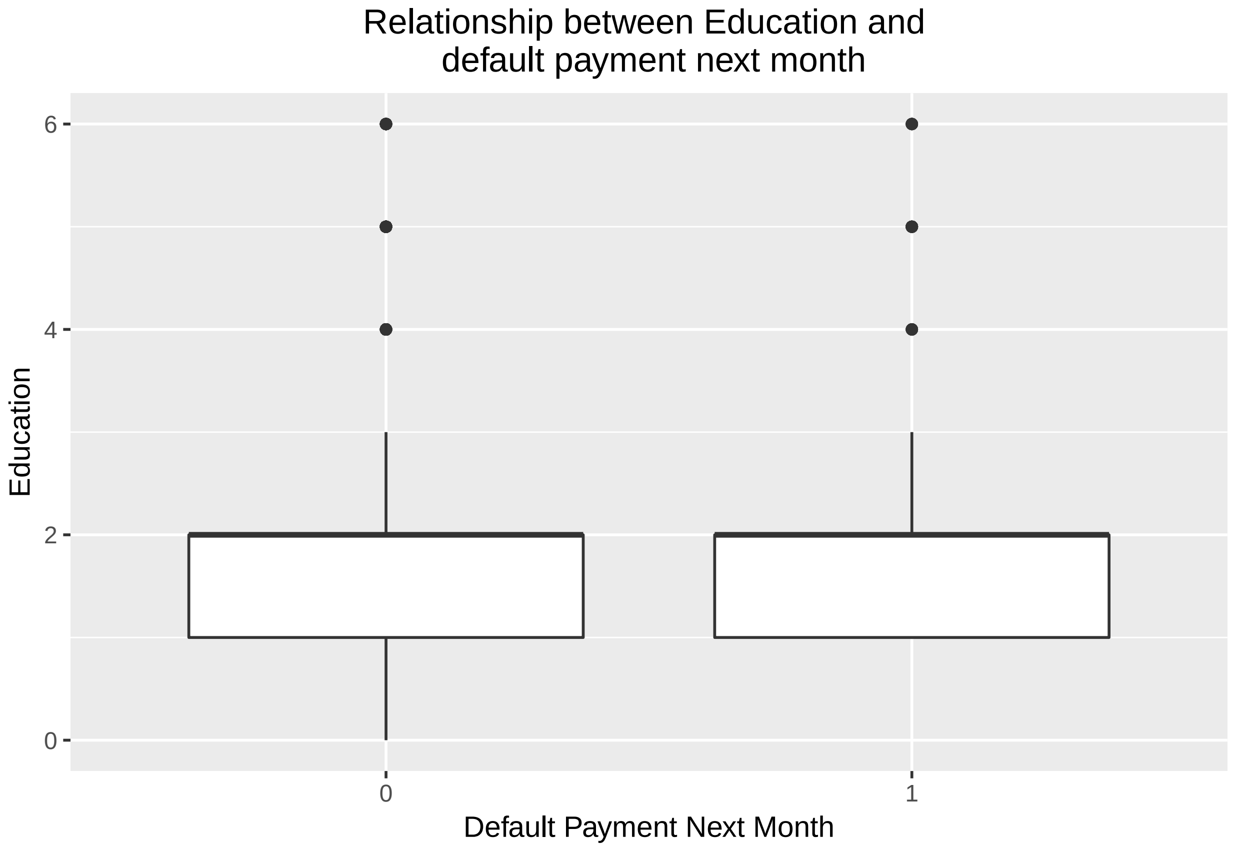
```
## [1] 0.2910026
```

```
## [1] 0.252265
```

Considering the above output, approximately 29.10% of males defaulted next month's payment, whereas only 25.23% did. This suggests that males may be more likely, on average, to default next month's payment.

3.2.2 Education against default payment next month

Education is another integer variable, taking 7 different levels, ranging from 0 to 6. We will once again consider side-by-side boxplots to investigate their relationship.



```
## [1] 1.84098
```

```
## [1] 1.89279
```

The mean level of education for those who do not default next month's payment is 1.84, whereas for those who do it's 1.89. Due to how close these are in value, it is unreasonable to suggest that the level of education of an individual will effect whether or not they default.

3.2.3. Marriage against default payment next month

We will now consider Marriage, another integer variable with four levels (0 up to 3). Because there are only four levels, no plot will be considered to investigate their relationship.

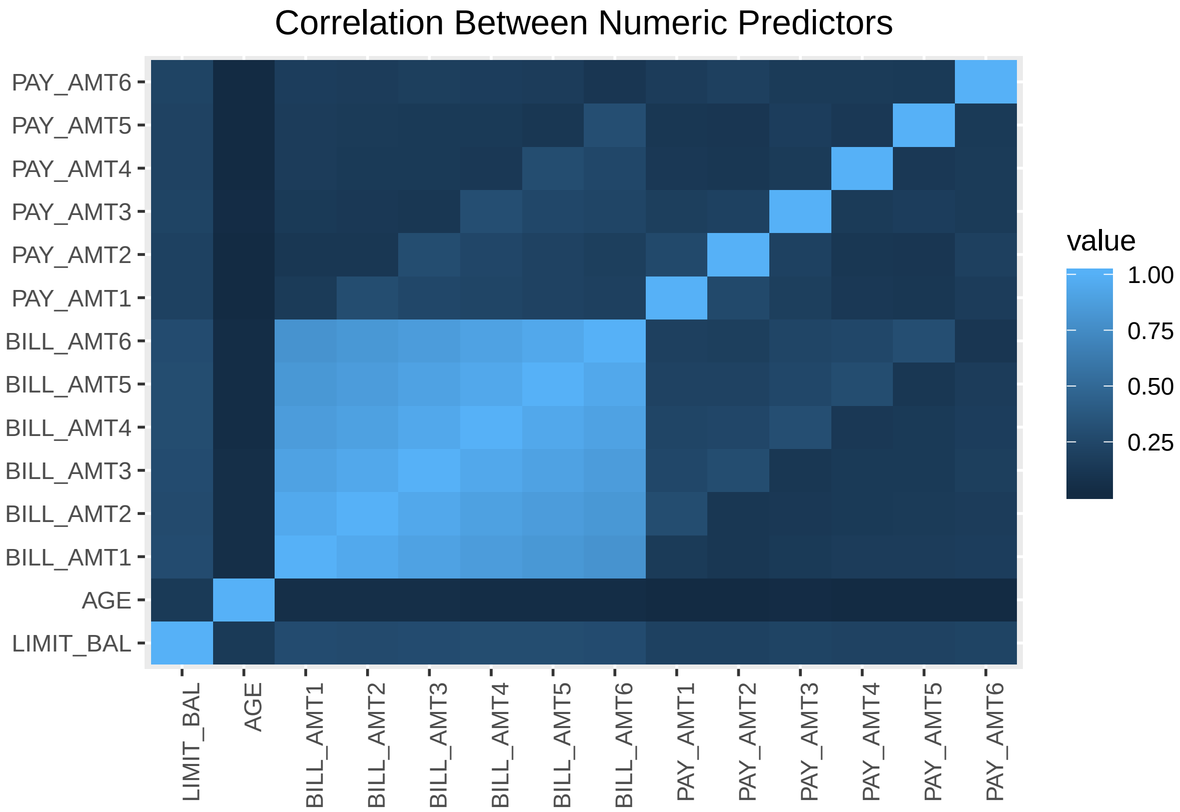
```
## [1] 1.557968
```

```
## [1] 1.525962
```

The mean marriage value for those who do not default next month's payment is 1.56, whereas this value is 1.53 for those who do. These values are once again so close in value that there is no suggestion of a relationship between these two variables.

3.3. Bivariate Analysis Between Predictors

In order to determine the strength of the relationships between predictors, we look at the correlation between the numeric predictor variables. This is shown in the graphic below.



	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6
BILL_AMT1	1.00	0.95	0.90	0.86	0.83	0.80
BILL_AMT2	0.95	1.00	0.94	0.89	0.86	0.83
BILL_AMT3	0.90	0.94	1.00	0.94	0.90	0.86
BILL_AMT4	0.86	0.89	0.94	1.00	0.94	0.90
BILL_AMT5	0.83	0.86	0.90	0.94	1.00	0.94
BILL_AMT6	0.80	0.83	0.86	0.90	0.94	1.00

Looking at the correlations, we see that there is high correlation between each of the BILL_AMT_x variables. The precise correlations for the BILL_AMT_x variables have been included in a table.

4. Model Fitting

4.1. Feature Engineering

Choosing an appropriate subset from the available predictor variables is an important step in selecting an accurate model. While the exclusion of important predictors will produce an incorrect model, adding too many redundant predictor terms can overcomplicate the model, making it harder to interpret without improving its accuracy. It can also reduce the statistical accuracy of parameter estimates. In machine

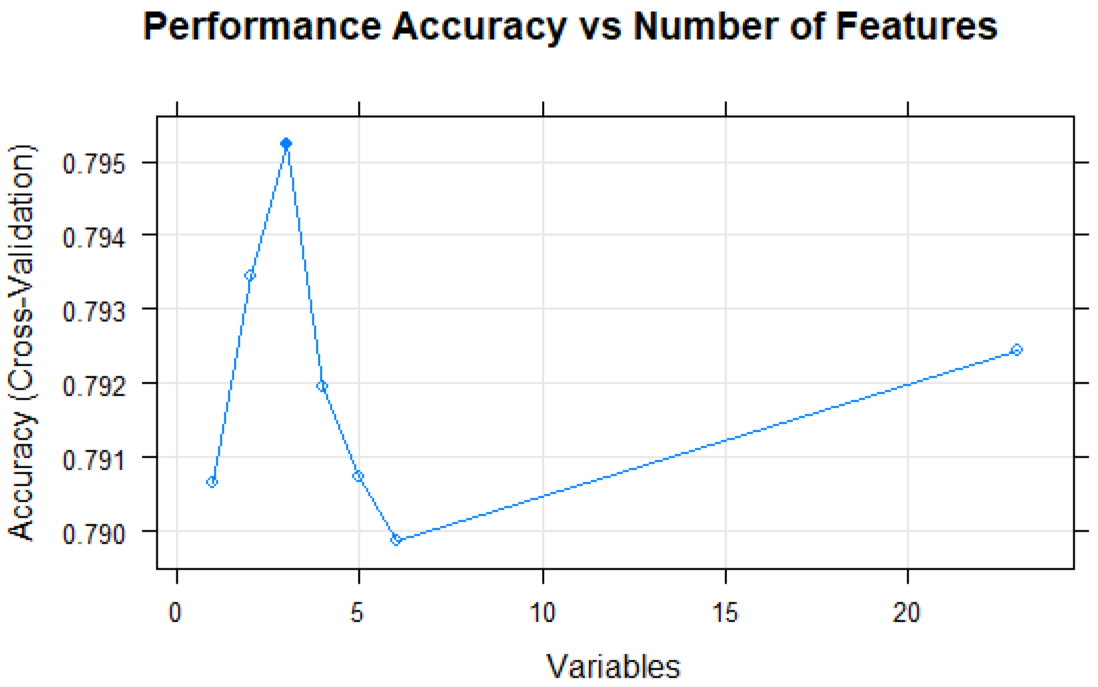


Figure 1: Accuracy against number of features for random forest model

learning, the process of selecting predictor variables for a model is know as *feature selection* ('feature' being a commonly used term for predictor variables).

The other activity that makes up feature engineering is *Feature extraction*. This is the process of building new features out of the available predictor variables in the data set. Additional features that we considered were interaction terms, which are often appropriate when there is an observed relationship between features. We note from the previous section that there is a strong correlation between the `BILL_AMT` variables.

We perform feature selection on the credit dataset using the Recursive Feature Elimination, which is a backwards selection technique that begins with all available features, and iteratively eliminates the least important ones. The model that was used to select features in this way was a random forest.

For the random forest model we examined using Recursive Feature Elimination, it was first found that the most accurate model consisted of just three features, as shown in Figure 1. `PAY_0`, `PAY_2` and `PAY_3`. It was decided that a random forest model including just these three predictors would be trained.

Unlike the random forest model, both the logistic regression model and the single decision tree model were able to be trained in a reasonable timeframe with all features included. The optimal feature set for each of these models was therefore examined at the model training and evaulation stage.

4.2. Model Selection

Given that this problem is one of binary classification, the following types of models were considered, all of which are commonly applied to binary classification problems

- Logistic regression
- Linear Support Vector Machine
- Kernelised Support Vector Machine
- Decision tree

Table 3: Test Accuracy of Trained Models

Model	Accuracy
Logistic Regression	79.95
Decision Tree	79.30

Table 4: Test Accuracy of All Trained Models

Model	Accuracy
Logistic Regression	79.95
Reduced Logistic Regression	79.93
Decision Tree	79.30
Reduced Decision Tree	79.46

- Random forest
- AdaBoost

Each of the models considered was trained using the `train()` function from the `caret` package. After attempting to train support vector machines, an AdaBoost model, and the reduced random forest with the selected features above, none were found to be able to train in a reasonable timeframe, and so had to be abandoned.

Consequently, the two models that we trained initially were a logisitic regression model and a decision tree model, both with all features included.

In the above R code, a logistic model was generated using `caret`, as well as using the standard R `glm()` function. As it proved difficult to use caret's functionality to automatically select optimal features using our preferred criterion, we instead used the stepwise AIC function. None of the features identified in this model exhibited a meaningful correlation, and so we did not include interactiion terms between any of the features.

We also trained a decision tree model using only those predictors identified through the recursive feature elminiation exercise for random forest, to check the accuracy of a more parsimonious feature set.

4.3. Model Testing and Evaluation

The trained models were tested against a portion of the original data set that had been randomly selected for testing purposes. Predicted values for the response variable are obtained, and then the accuracy of these predictions is stored using the `confusionMatrix()` function.

5. Final Model

Having attempted several different approaches, we found that only two could be trained in a practical timeframe. These were logistic regression and a single decision tree. Our preliminary feature selection work was not directly applicable to our final model selection decisions, given that it proved too resource-intensive to train even a reduced random forest model on the data set. However, given the random forests are an aggregation of decision trees, we chose to try training a decision tree with only the three features that were identified as important for the random forest.

To obtain a more parsimonious logistic regression model, we applied stepwise selection using the Akaike Information Criterion to determine feature significance. The features that were identified through this process were then extracted, and a new model was trained, using cross validation, on the reduced feature set.

In total, we trained four different models; two decision trees and two logistic regression models. We trained these models several times as they were each being developed, and typically found little variability in terms of their classification accuracy. However, the logisitic regression models performed slightly better than the decision tree models, and reduced logistic model showed very similar or slightly better perfomance than the full model. In the interests of model simplicity, we have opted to use the reduced logistic regression model to make our predictions.

```r
# Save the final model
saveRDS(mdl_logistic.2, "./final_model.rds")
```