Cleaning and Variable Analysis

1 Introduction

A crucial part of issuing credit cards is risk management and default detection within a bank. Though the law has stringent measures against credit card defaulting, it is still prevalent in many cases. It is often difficult to determine whether a given individual is at risk of defaulting given a data set, due to the highly complicated nature of the problem. Statistical models may assist in predicting if a person is likely to default given their history and other factors, to assist banks in minimising potential losses.

Predictive modelling is the process of applying statistical techniques, such as regression, to derive a model between a set of predictor variables, such as the financial history of an individual, and a response variable, such as whether or not they are at risk of defaulting. This is usually done by analysing a data set where the response is known for the predictor variables, to build a model that can then be applied to data where the response is unknown. In this project, the data set for training the model consists of 20,000 individuals, with their financial history, demographics and their defaulting status.

To produce a model, first the data was split into two sets in Section 2, with 15,000 observations being used as a training set, and the remaining 5,000 retained as a validation set. Sections **??** and **??** desribe the distributins of each variable in the data set, as well as any interactions present between each of the variables. This allows for a more accurate set of potential models to be produced in Section **??**, using the training data. The validation data is then used to evaluate the error of each model in Section **??**, with the best model and its error presented in Section **??**. The assumptions of the chosen model will be shown and discussed in Section **?? is there actually a prediction that needs to be made?**.

2 Cleaning and Preparing the Data

The following R code was used to read in the data set from `train.csv`.

```r
dat = read.csv("train.csv", skip = 1)
```

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:

```r
str(dat)
```

```
## 'data.frame':    20000 obs. of  25 variables:
##  $ ID                        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ LIMIT_BAL                 : int  20000 120000 20000 120000 70000 450000 60000 50000 210000 20000
##  $ SEX                       : int  2 2 1 2 2 2 1 1 1 2 ...
##  $ EDUCATION                 : int  2 2 1 2 2 1 1 2 1 1 ...
##  $ MARRIAGE                  : int  1 2 2 1 2 1 2 2 2 2 ...
##  $ AGE                       : int  24 26 24 39 26 40 27 33 29 22 ...
##  $ PAY_0                     : int  2 -1 0 -1 2 -2 1 2 -2 0 ...
##  $ PAY_2                     : int  2 2 0 -1 0 -2 -2 0 -2 0 ...
##  $ PAY_3                     : int  -1 0 2 -1 0 -2 -1 0 -2 2 ...
##  $ PAY_4                     : int  -1 0 2 -1 2 -2 -1 0 -2 -1 ...
##  $ PAY_5                     : int  -2 0 2 -1 2 -2 -1 0 -2 0 ...
##  $ PAY_6                     : int  -2 2 2 -1 2 -2 -1 0 -2 0 ...
##  $ BILL_AMT1                 : int  3913 2682 15376 316 41087 5512 -109 30518 0 14028 ...
##  $ BILL_AMT2                 : int  3102 1725 18010 316 42445 19420 -425 29618 0 16484 ...
##  $ BILL_AMT3                 : int  689 2682 17428 316 45020 1473 259 22102 0 15800 ...
##  $ BILL_AMT4                 : int  0 3272 18338 0 44006 560 -57 22734 0 16341 ...
##  $ BILL_AMT5                 : int  0 3455 17905 632 46905 0 127 23217 0 16675 ...
##  $ BILL_AMT6                 : int  0 3261 19104 316 46012 0 -189 23680 0 0 ...
##  $ PAY_AMT1                  : int  0 0 3200 316 2007 19428 0 1718 0 3000 ...
##  $ PAY_AMT2                  : int  689 1000 0 316 3582 1473 1000 1500 0 0 ...
##  $ PAY_AMT3                  : int  0 1000 1500 0 0 560 0 1000 0 16741 ...
##  $ PAY_AMT4                  : int  0 1000 0 632 3601 0 500 1000 0 334 ...
##  $ PAY_AMT5                  : int  0 0 1650 316 0 0 0 1000 0 0 ...
##  $ PAY_AMT6                  : int  0 2000 0 0 1820 1128 1000 716 0 0 ...
##  $ default.payment.next.month: int  1 1 1 1 1 1 1 1 1 1 ...
```

As shown, all variables are considered integers. This is not correct, as the only numeric variables are LIMIT_BAL, AGE, BILL_AMT and PAY_AMT. The other variables were converted to factors using the following R code:

```r
dat2 = dat
dat2$SEX = as.factor(dat$SEX)
dat2$EDUCATION = as.factor(dat$EDUCATION)
dat2$MARRIAGE = as.factor(dat$MARRIAGE)
dat2$PAY_0 = as.factor(dat$PAY_0)
dat2$PAY_2 = as.factor(dat$PAY_2)
dat2$PAY_3 = as.factor(dat$PAY_3)
dat2$PAY_4 = as.factor(dat$PAY_4)
dat2$PAY_5 = as.factor(dat$PAY_5)
dat2$PAY_6 = as.factor(dat$PAY_6)
dat2$default.payment.next.month = as.factor(dat$default.payment.next.month)
```

The data set now comprises of 14 numeric variables and 10 factors, as expected:

```r
str(dat2)
```

```
## 'data.frame':    20000 obs. of  25 variables:
##  $ ID                        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ LIMIT_BAL                 : int  20000 120000 20000 120000 70000 450000 60000 50000 210000 20000
##  $ SEX                       : Factor w/ 2 levels "1","2": 2 2 1 2 2 2 1 1 1 2 ...
##  $ EDUCATION                 : Factor w/ 7 levels "0","1","2","3",..: 3 3 2 3 3 2 2 3 2 2 ...
##  $ MARRIAGE                  : Factor w/ 4 levels "0","1","2","3": 2 3 3 2 3 2 3 3 3 3 ...
##  $ AGE                       : int  24 26 24 39 26 40 27 33 29 22 ...
##  $ PAY_0                     : Factor w/ 11 levels "-2","-1","0",..: 5 2 3 2 5 1 4 5 1 3 ...
##  $ PAY_2                     : Factor w/ 10 levels "-2","-1","0",..: 5 5 3 2 3 1 1 3 1 3 ...
##  $ PAY_3                     : Factor w/ 11 levels "-2","-1","0",..: 2 3 5 2 3 1 2 3 1 5 ...
##  $ PAY_4                     : Factor w/ 11 levels "-2","-1","0",..: 2 3 5 2 5 1 2 3 1 2 ...
##  $ PAY_5                     : Factor w/ 10 levels "-2","-1","0",..: 1 3 4 2 4 1 2 3 1 3 ...
##  $ PAY_6                     : Factor w/ 10 levels "-2","-1","0",..: 1 4 4 2 4 1 2 3 1 3 ...
##  $ BILL_AMT1                 : int  3913 2682 15376 316 41087 5512 -109 30518 0 14028 ...
##  $ BILL_AMT2                 : int  3102 1725 18010 316 42445 19420 -425 29618 0 16484 ...
##  $ BILL_AMT3                 : int  689 2682 17428 316 45020 1473 259 22102 0 15800 ...
##  $ BILL_AMT4                 : int  0 3272 18338 0 44006 560 -57 22734 0 16341 ...
##  $ BILL_AMT5                 : int  0 3455 17905 632 46905 0 127 23217 0 16675 ...
##  $ BILL_AMT6                 : int  0 3261 19104 316 46012 0 -189 23680 0 0 ...
##  $ PAY_AMT1                  : int  0 0 3200 316 2007 19428 0 1718 0 3000 ...
##  $ PAY_AMT2                  : int  689 1000 0 316 3582 1473 1000 1500 0 0 ...
##  $ PAY_AMT3                  : int  0 1000 1500 0 0 560 0 1000 0 16741 ...
##  $ PAY_AMT4                  : int  0 1000 0 632 3601 0 500 1000 0 334 ...
##  $ PAY_AMT5                  : int  0 0 1650 316 0 0 0 1000 0 0 ...
##  $ PAY_AMT6                  : int  0 2000 0 0 1820 1128 1000 716 0 0 ...
##  $ default.payment.next.month: Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

The data was then split into the predictor variables and response variable, as well as the training set and validation set.

```r
validInd = sample(1:nrow(dat), nrow(dat)/4)
train = dat[-validInd, ]
valid = dat[validInd, ]
trainX = train[ ,1:(ncol(dat)-1)]
trainY = train[ ,ncol(dat)]
validX = valid[ ,1:(ncol(dat)-1)]
validY = valid[ ,ncol(dat)]
```

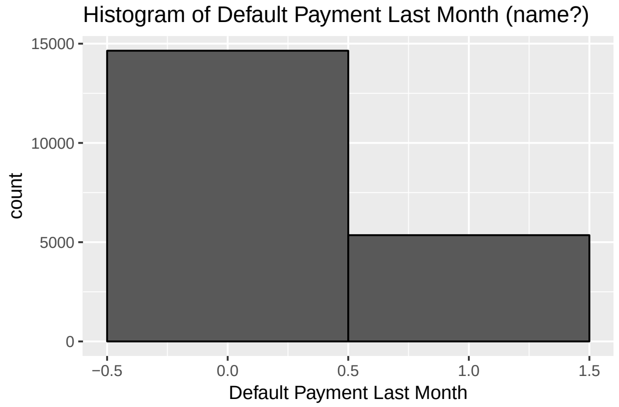
3 Univariate Analysis

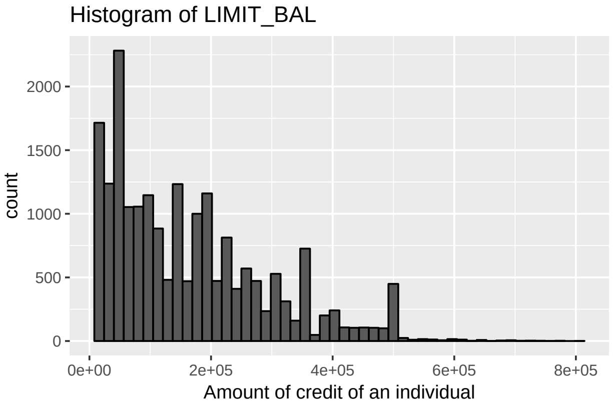
3.1 Variable Descriptions

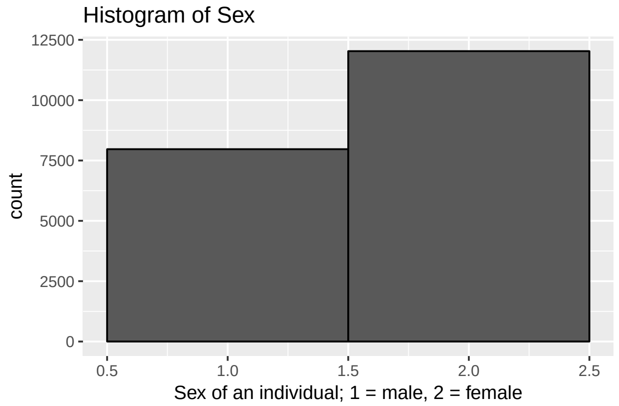
Table 1: Description of variables in the data set.

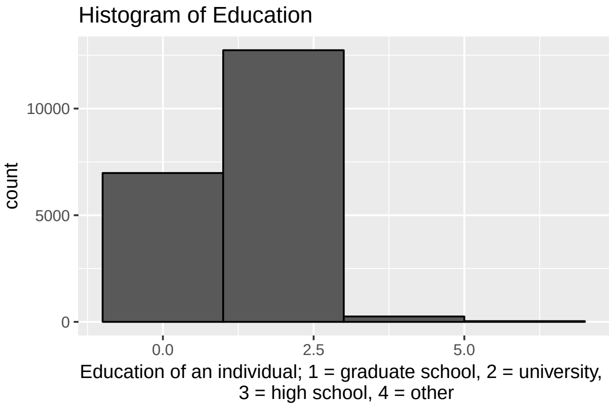
Variable Name	Data Type	Role in model	Description
default payment next month	Factor	Response	1 = a default payment, 0 = no default
LIMIT_BAL	Numeric	Predictor	Amount of credit of an individual, in NT dollars
SEX	Factor	Predictor	Sex of an individual; 1 = male, 2 = female
EDUCATION	Factor	Predictor	Education 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

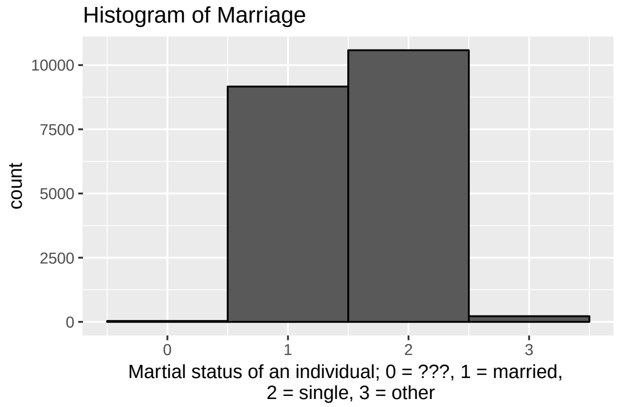
3.2 Univariate Plots

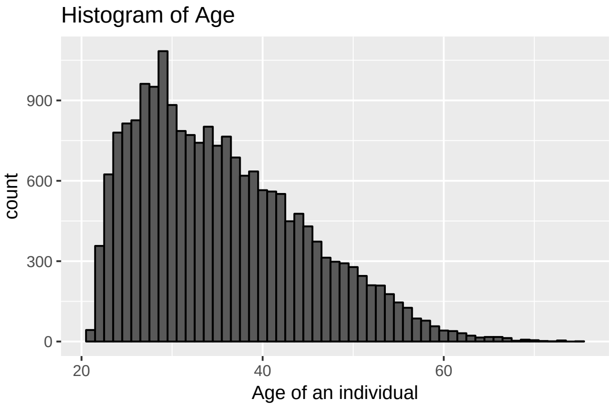

Histogram of Default Payment Last Month (name?)
count
Default Payment Last Month


Histogram of LIMIT_BAL
count
0e+00
2e+05
4e+05
6e+05
8e+05
Amount of credit of an individual


Histogram of Sex
count
Sex of an individual; 1 = male, 2 = female


Histogram of Education
count
Education of an individual; 1 = graduate school, 2 = university, 3 = high school, 4 = other


Histogram of Marriage
count
Martial status of an individual; 0 = ???, 1 = married, 2 = single, 3 = other


Histogram of Age
count
Age of an individual