BUSINESS REPORT- PREDICTING DEFAULT RISK

Section 1: Business Understanding

Business Situation

Our bank receives 200 loan applications per week, but due to a financial scandal that hit a competitor the credit risk unit of the bank will be processing 500 applications this week.The influx of new credit applications is a great opportunity the bank wants to immediately pursue.

The Complication

The bank will want to maintain there processing turnaround time while ensuring that the credit risk unit is able to effectively determine creditworthy applications, while reducing the risk of default by effectively determining non-creditworthy applications.

Key Decision that needs to be made

The Head of the credit risk department needs to decide if a loan should be approved for each of the 500 loan applications received this week.

Approach

This project is data rich; it has readily available information that can be used to predict creditworthiness of the 500 loan applications. The data will be acquired internally from already processed loan applications,'customers-to-score' and the data from the 500 loan applications yet to be reviewed,'customers-to-score'. The two sets of data include personal details about the customer, such as their age and how long they have been at their current job. It will also include details on the individual's banking and credit history, such as their account balance, number of credits at this bank, and their payment status of previous credit.

We will use the data set with already processed loan application to build a binary classification predictive model to determine if a customer is creditworthy or non-creditworthy.

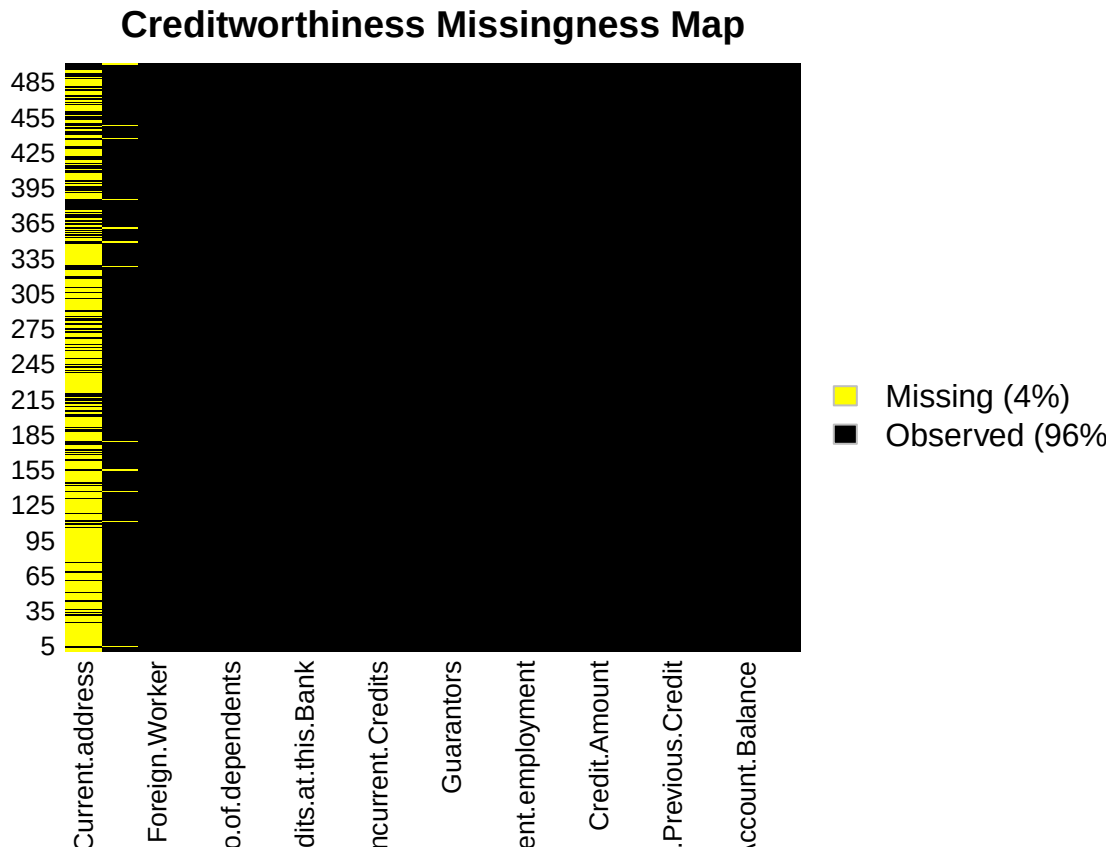
Section 2: Data Structure & Quality

The data we used to train the model was an equivalent sum of 500 loan applications with 19 variables that includes the outcome variable. The 19 variables included in the data set are listed as follows:

```
names(train)
```

```
##  [1] "Credit.Application.Result"          "Account.Balance"
##  [3] "Duration.of.Credit.Month"           "Payment.Status.of.Previous.Credit"
##  [5] "Purpose"                            "Credit.Amount"
##  [7] "Value.Savings.Stocks"               "Length.of.current.employment"
##  [9] "Instalment.per.cent"                "Guarantors"
## [11] "Duration.in.Current.address"        "Most.valuable.available.asset"
## [13] "Age.years"                          "Concurrent.Credits"
## [15] "Type.of.apartment"                  "No.of.Credits.at.this.Bank"
## [17] "Occupation"                         "No.of.dependents"
## [19] "Telephone"                          "Foreign.Worker"
```

We checked the data structure and quality to check for missing values.




```
##                                    colSums.is.na.train..
## Credit.Application.Result                              0
## Account.Balance                                        0
## Duration.of.Credit.Month                               0
## Payment.Status.of.Previous.Credit                      0
## Purpose                                                0
## Credit.Amount                                          0
## Value.Savings.Stocks                                   0
## Length.of.current.employment                           0
## Instalment.per.cent                                    0
## Guarantors                                             0
## Duration.in.Current.address                          344
## Most.valuable.available.asset                          0
## Age.years                                             12
## Concurrent.Credits                                     0
## Type.of.apartment                                      0
## No.of.Credits.at.this.Bank                             0
## Occupation                                             0
## No.of.dependents                                       0
## Telephone                                              0
## Foreign.Worker                                         0
```

The missing values plot showed that the variables *Duration.in.Current.address* had more than 50% of its values missing and *Age.years* had less than 5% missing values. We dropped *Duration.in.Current.address* and imputed the missing *Age.years* with the median value. Character variables were also encoded as factors.

Section 3: Exploratory Data Analysis

Next we performed some exploratory data analysis to generate some insights from our internal data. We started with some distributions to get a sense of the numeric data.