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:

```
##  [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"
```

Table 2.1 List of variables

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

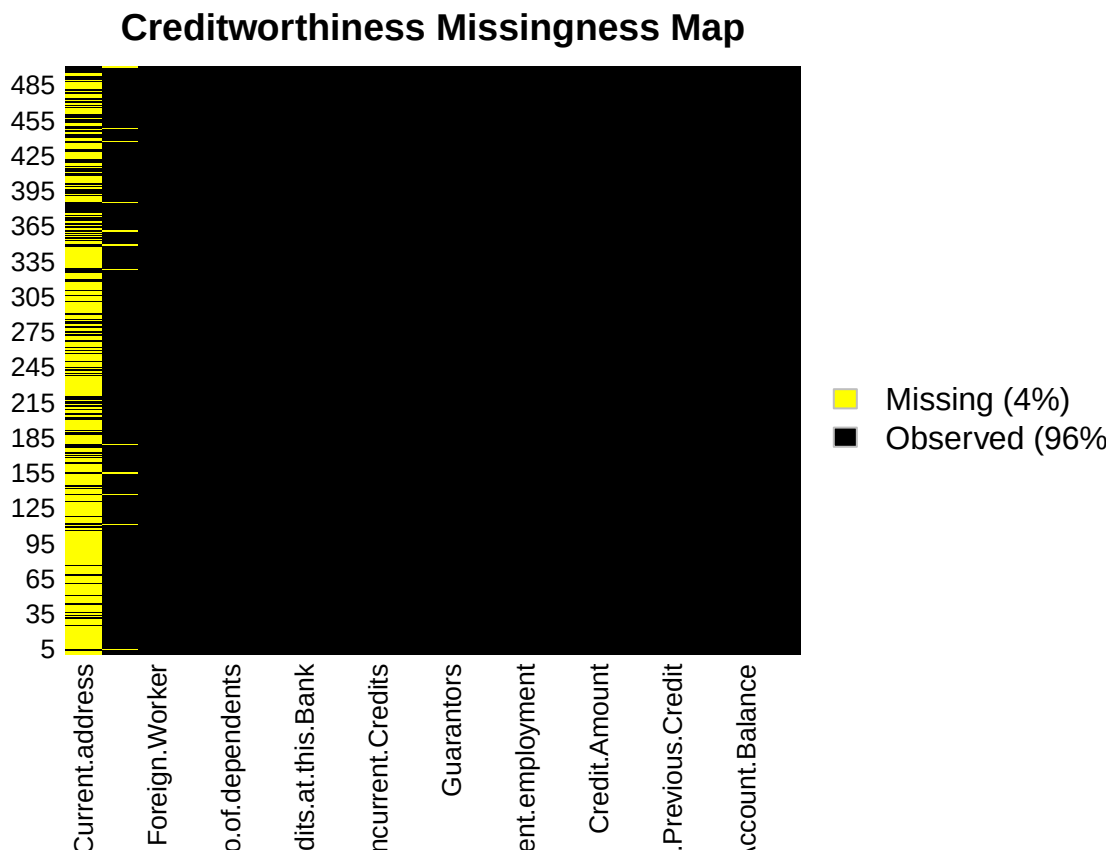

Fig 2.1 Missing Values

	NAs
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

Table 2.2 NA's per variable

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 of categorical variables to get a sense of variables that have zero variance or near zro variance that can affect our modeling.

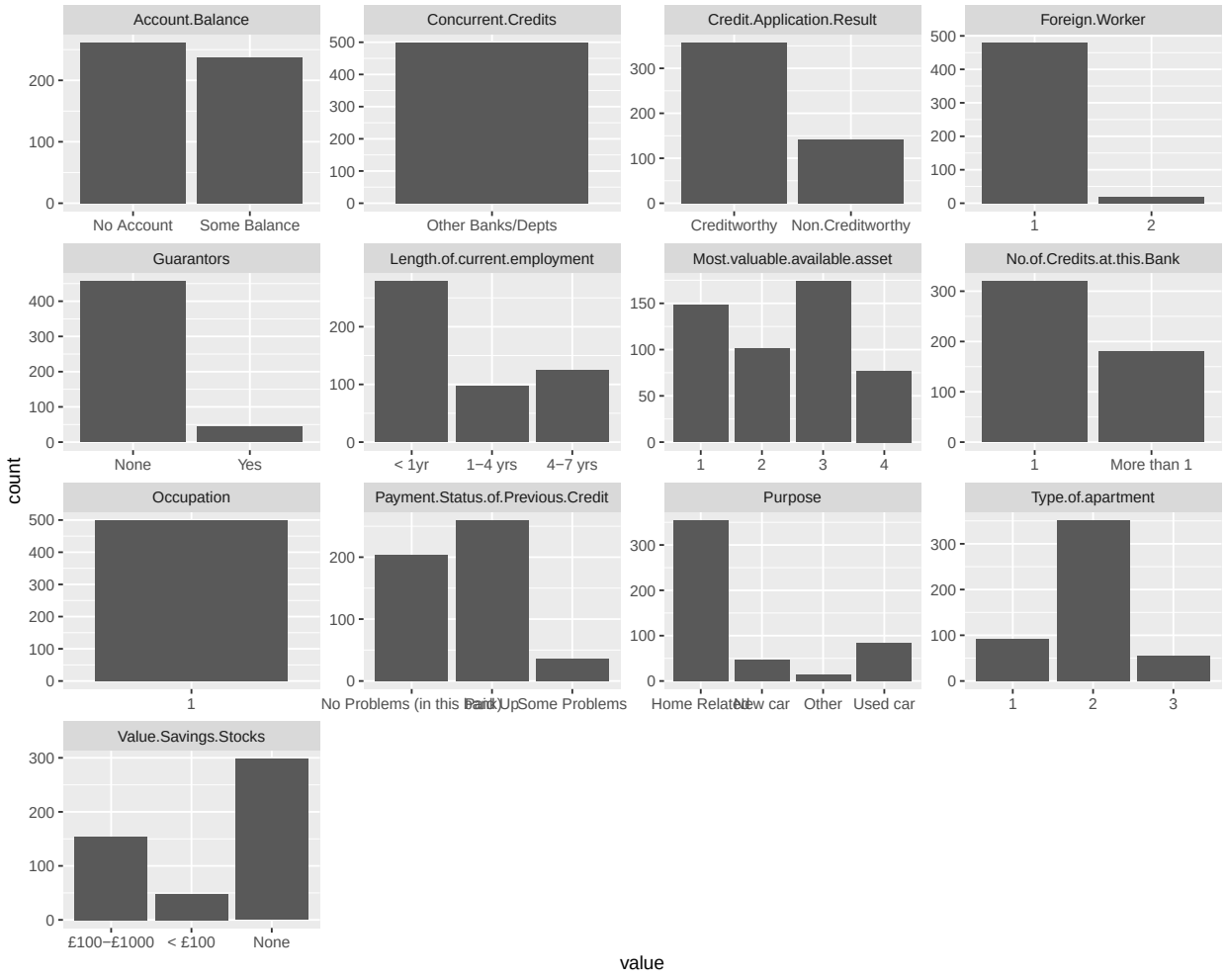

Fig_3.1 Bar charts of categorical variables

We can identify immediately 3 variables, 'Concurrent.Credits','Foreign.Worker', 'Guarantors' and 'Occupation'. To be sure of this we claculate the frequency ratio of the most occurrence over the second most occurrence value within the variables and also the percent unique to check their validity.

```
##                    freqRatio percentUnique zeroVar  nzv
## Concurrent.Credits   0.00000           0.2    TRUE TRUE
## Occupation           0.00000           0.2    TRUE TRUE
## Foreign.Worker      25.31579           0.4   FALSE TRUE
```

Table 3.1 Near Zero Variance predictor variables

We discovered that only three variables meet this criteria as shown in Table 3.1 above and these variables were dropped. We continue with our analysis by plotting the distribution of the continuous variables.

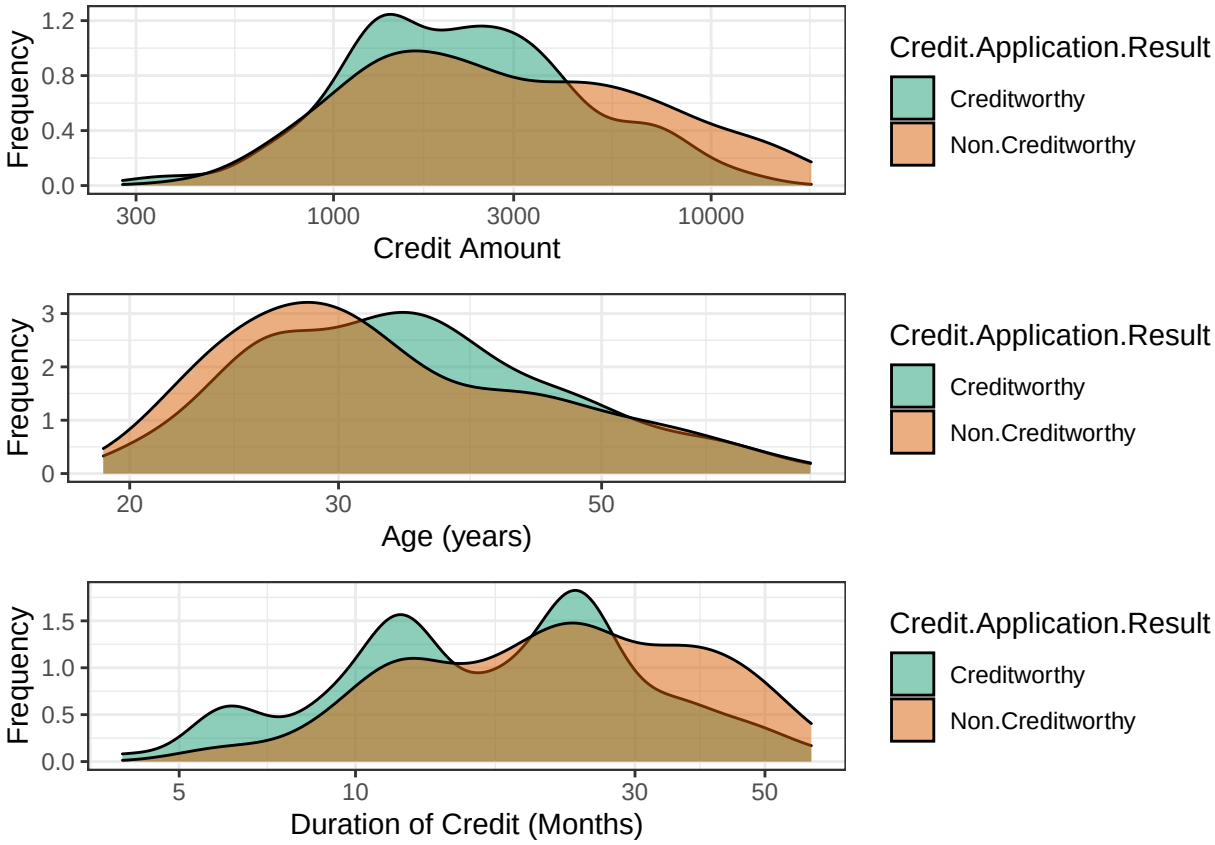

Fig 3.2 Density Plots

We can see that they are not normally distributed and it seems as if the central tendencies for each category in the three plots are different. We also see that the categories follow the same skew direction. The defined humps in Credit Amount and Duration of Credit were there are 3 suggests that there are about 2-3 defined groups in the loan applicants which will be verified with categorical plots subsequently. We also sense that these be strong predictors and this will be further verified with boxplots to visualize the difference in their means.

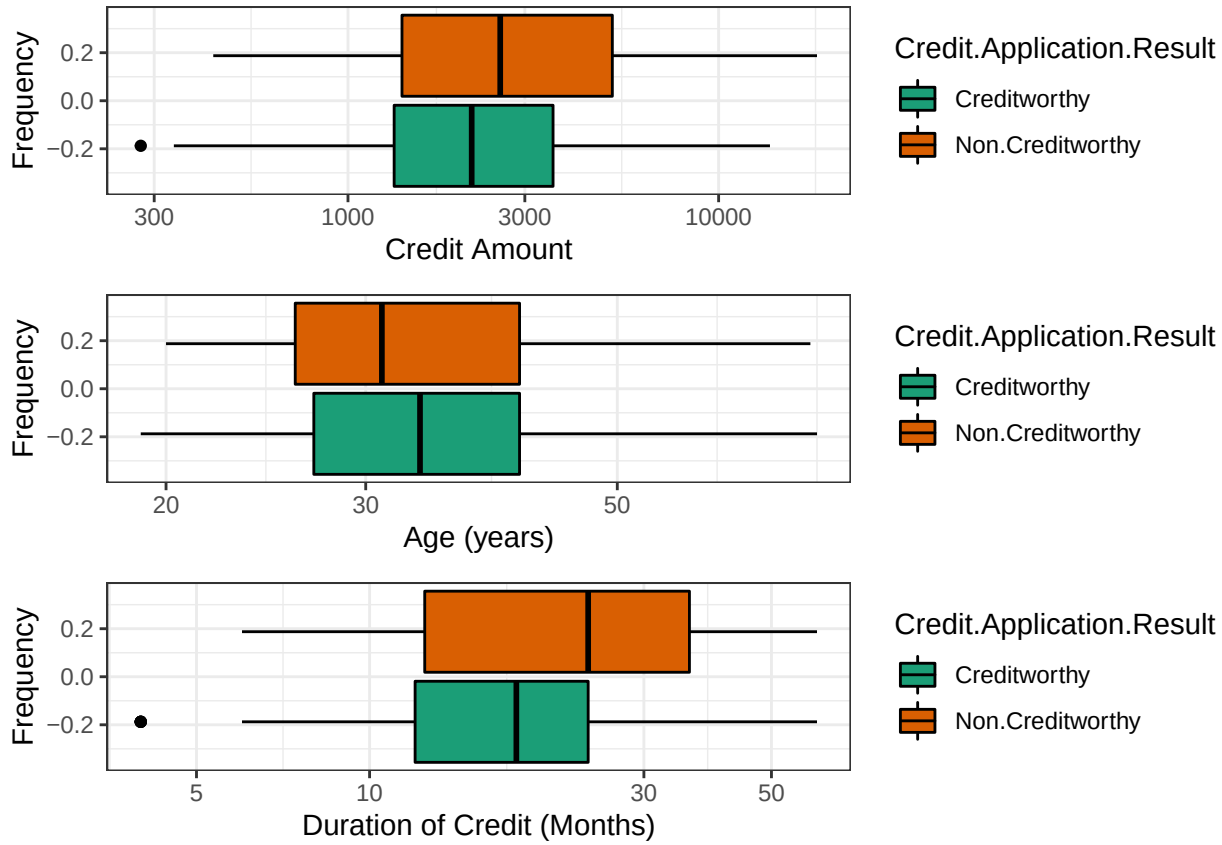

Fig 3.3 Box Plots

We discover that the difference is means for the two categories are evident for the 3 continuous predictor variables and Duration of Credit is the strongest of the three. This was tested using the t-test and their strength measured by the p-values is shown in the table below.

```
## # A tibble: 3 x 2
## # Groups:   var [3]
##   var                          p.value
##   <chr>                          <dbl>
## 1 Age.years                   0.247
## 2 Credit.Amount               0.000260
## 3 Duration.of.Credit.Month 0.0000207
```

Table 3.2 Strength of continuous predictor variables

Here we look for insights on some of the categorical variables by plotting some bar graphs.

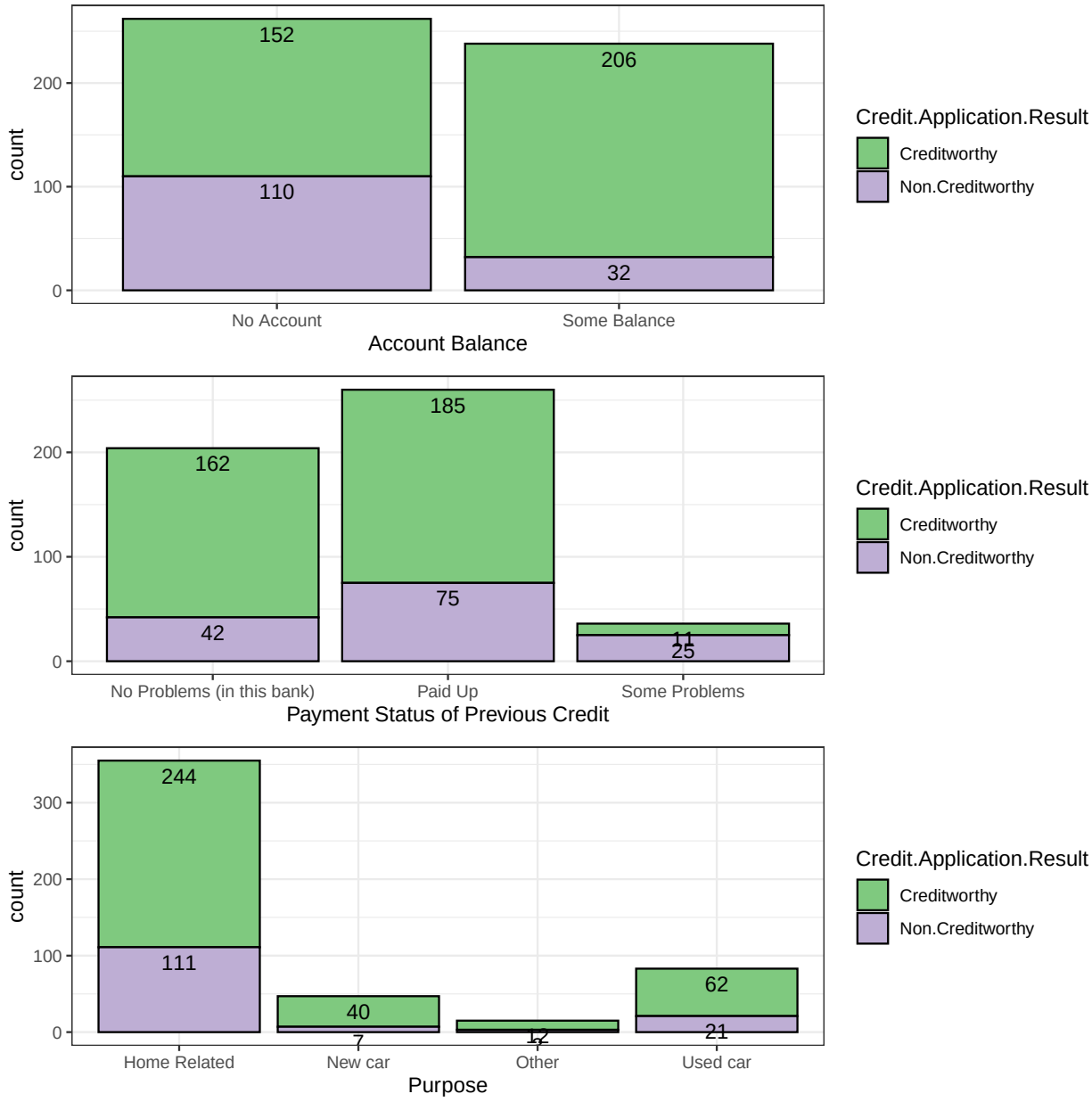

Fig 3.4 Barplots group 1

Some observations from Fig 3.4

Account Balance: there are more Creditworthy people with *Some Balance* than *No Account* and consequently less people Non.Creditworthy people with *Some Balance* than *No Account* . It suggests that having some amount in your account may determine creditworthiness.

Payment Status of previous Credit:This chart suggests that while there may be 25 accounts with *Some Problems* that are classified as Non.Creditworthy, having probably defaulted, there were 75 accounts that are now *Paid Up* but classified as Non.Creditworthy. The predictive model should be able to reduce this risk with an acceptable mis-classification rate.

Purpose: The bank has a bigger appetite for *Home Related* loans.

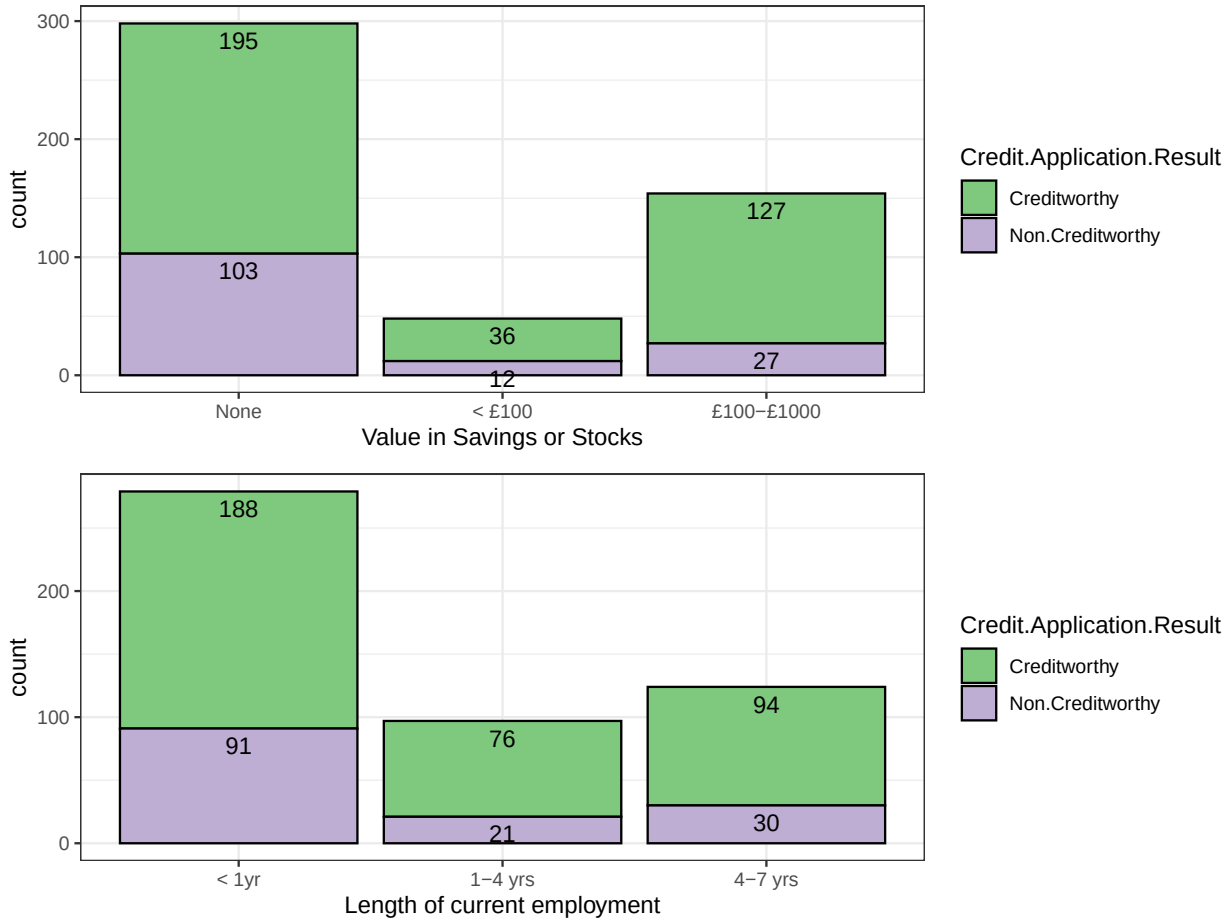

Fig 3.5 Bar plots group2

This group of charts seem to follow the same trend. It shows that people with less than a year in current may not have attained an acceptable credit score and without any savings further supports the suggestion that these may be new employees or those just starting out in their careers.

Section 4: Training the Model

We used 4 different algorithms to train the binary classification model and compared their ROC and Accuracy performance of metrics to determine the best performance for scoring the 500 new loan applications.The algorithms we used are:

1.Logistic Regression
2.Decision Tree
3.Random Forest
4.Boosted Tree

We did a train/test split of 70/30 for our external validation and set the hyper parameters to tune the different algorithms using repeated cross validation of 10 for to determine the accuracy of the models and class probability with two class summary to measure the ROC of the models.

1. Logistic Regression The model gave an accuracy of 0.73 with a 10-fold repeated cross validation as can be see in the model output below.

```
## Generalized Linear Model
##
```

```
## 351 samples
##  15 predictor
##   2 classes: 'Creditworthy', 'Non.Creditworthy'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 10 times)
## Summary of sample sizes: 316, 316, 316, 316, 316, 316, ...
## Resampling results:
##
##   Accuracy   Kappa
##   0.7329683  0.2616023
```

Table 4.1 Logistic Regression model Accuracy measure output

We measured the performance of the Logistic Regression using the ROC metric with the output of 0.745 as seen below.

```
## Generalized Linear Model
##
## 500 samples
##  15 predictor
##   2 classes: 'Creditworthy', 'Non.Creditworthy'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 10 times)
## Summary of sample sizes: 450, 451, 450, 450, 450, 450, ...
## Resampling results:
##
##   ROC        Sens       Spec
##   0.7487683  0.8964127  0.4016667
```

Table 4.1b Logistic Regression with ROC measure output

We applied the model to the hold-out sample and got an overall of 79.2% as seen in the Confusion Matrix and its statistics output below. This was an improvement on the accuracy of the model. However, the calculated Positive Predictive Value (PPV) of 81.5% and Negative Predictive Value (NPV) of 70.37% showed that the model is biased towards Creditworthy.

```
## Confusion Matrix and Statistics
##
##
## glm.pred           Creditworthy Non.Creditworthy
##   Creditworthy               99               23
##   Non.Creditworthy            8               19
##
##                Accuracy : 0.7919
##                  95% CI : (0.7179, 0.854)
##     No Information Rate : 0.7181
##     P-Value [Acc > NIR] : 0.02538
##
##                   Kappa : 0.4236
##
##  Mcnemar's Test P-Value : 0.01192
##
```

```
##             Sensitivity : 0.9252
##             Specificity : 0.4524
##          Pos Pred Value : 0.8115
##          Neg Pred Value : 0.7037
##              Prevalence : 0.7181
##          Detection Rate : 0.6644
##    Detection Prevalence : 0.8188
##       Balanced Accuracy : 0.6888
##
##        'Positive' Class : Creditworthy
##
```

Table 4.2 Confusion Matrix and Statistics output

2. Decision Tree

We used the Decision Tree algorithm to train the model and got an accuracy of 73.27% at a complexity parameter of 0.0375. This can be seen in the line plot of the tuning parameter in Fig 4.2 below.

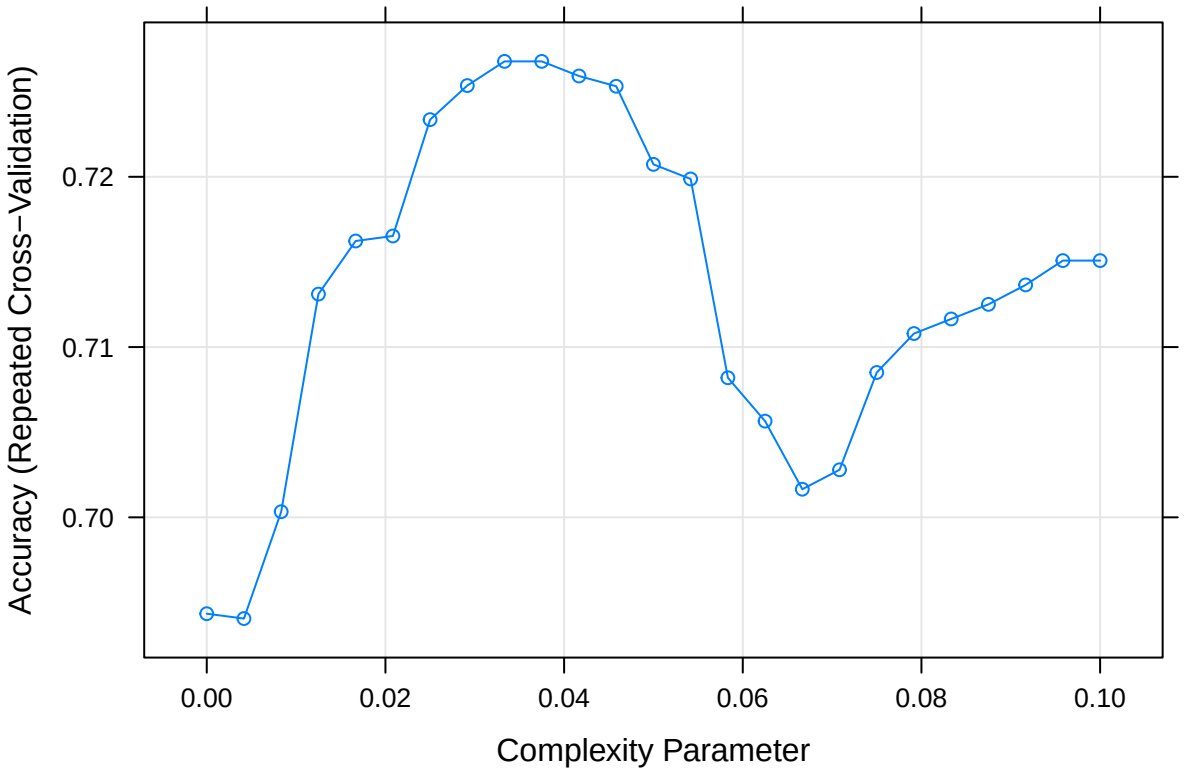

Fig 4.2 Decision Tree complexity parameter tuning plot

We can also see from the variable importance plot that *Payment.Status.of Previous.Credit* is the most important followed by *Duration.Credit.Month* and *Account.Balance* compared to the Logistic Regression model that showed *Account.Balance* as the most significant predictor.

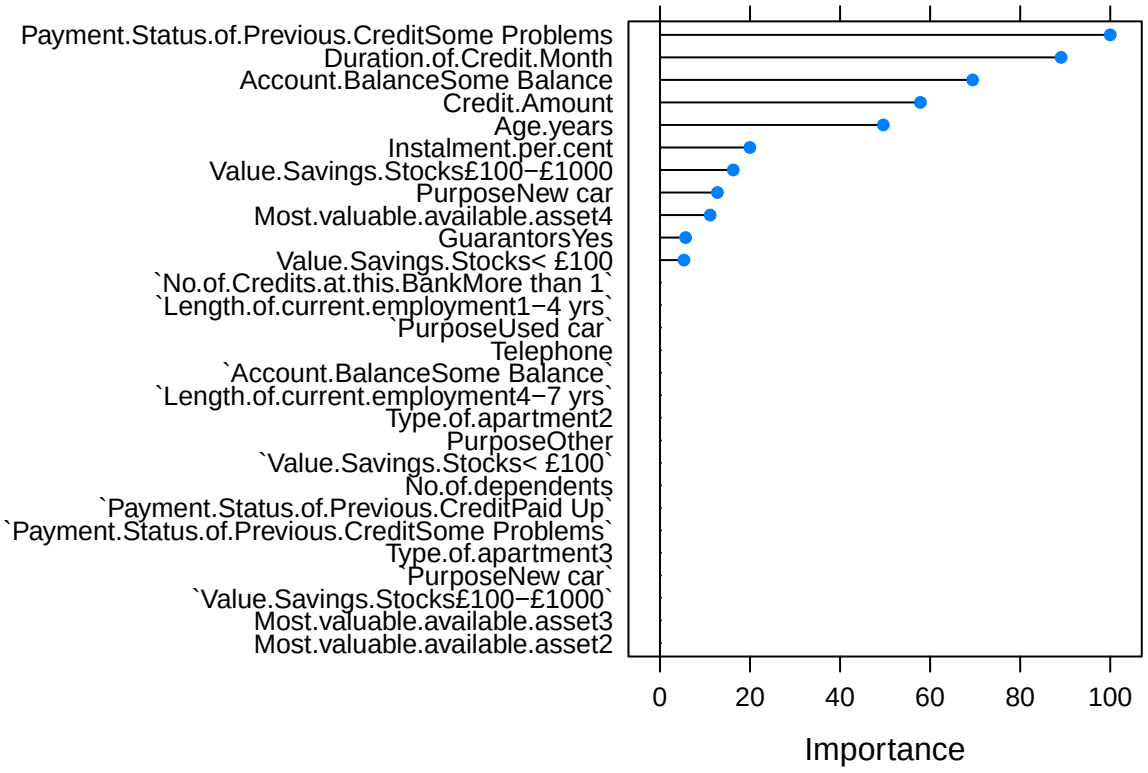

Fig 4.3 Decision Tree Variable importance plot

Next we tuned the Decision Tree model using the ROC metric. This gave us the optimal ROC of 69.39% at a complexity parameter value of 0.03333333. See the ROC Vs CP plot in Fig 4.4 below.

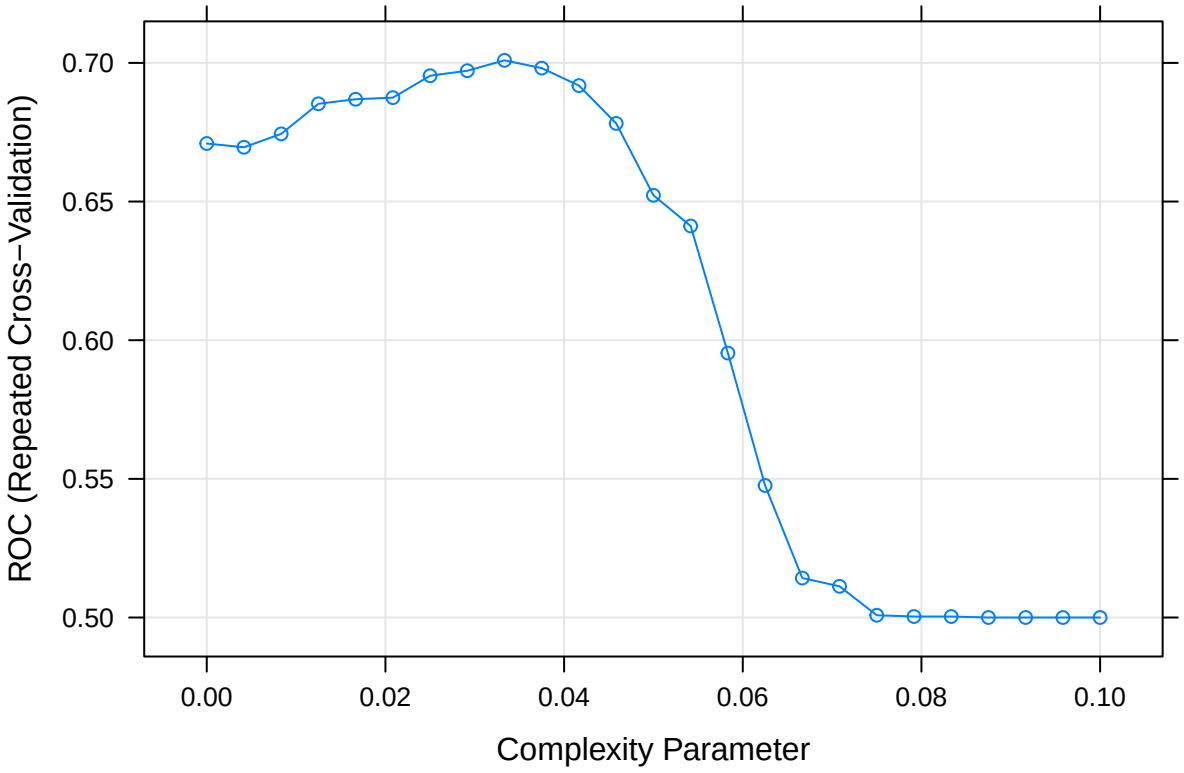

Fig 4.4 Decision Tree ROC vs CP plot

Applying the model with accuracy as metric of choice to the hold-out sample gave an overall accuracy of 78.52%. This mode did not perform better than the Logistic Regression.Similar to the Logistic Regression, it is biased towards the Creditworthy category from the values of the PPV and NPV.

```
## Confusion Matrix and Statistics
##
##
## tree.pred          Creditworthy Non.Creditworthy
##   Creditworthy              100               25
##   Non.Creditworthy            7               17
##
##                Accuracy : 0.7852
##                  95% CI : (0.7106, 0.8482)
##     No Information Rate : 0.7181
##     P-Value [Acc > NIR] : 0.039154
##
##                   Kappa : 0.3901
##
##  Mcnemar's Test P-Value : 0.002654
##
##             Sensitivity : 0.9346
##             Specificity : 0.4048
##          Pos Pred Value : 0.8000
##          Neg Pred Value : 0.7083
##              Prevalence : 0.7181
```

```
##          Detection Rate : 0.6711
##    Detection Prevalence : 0.8389
##       Balanced Accuracy : 0.6697
##
##        'Positive' Class : Creditworthy
##
```

3. Random Forest