Predicting the Credit Scores of Individuals

20129999

Introduction

In this analysis, we study the models for the associations between credit scores of individuals and other measurements. The data set has been divided into two groups, Train and Test. The relations between 20 measurements and the credit scores from Train have been checked to find a generalised model for predicting and classifying the risk of Test. From the modelling process, we find that Status, Duration, History, Purpose, Amount, Savings, Disposale, Personal, OtherParties, Age, Plans, Housing, Telephone and Foreign have close relations with credit scores of individuals. While the connection between Employment, Residence, Existing, Dependants, Property, Job and the credit scores may not be significant. Effectiveness of several linear models have been verified by mixtures of explanatory factors and covariates. The best-fitted model among them has been chosen during the process.

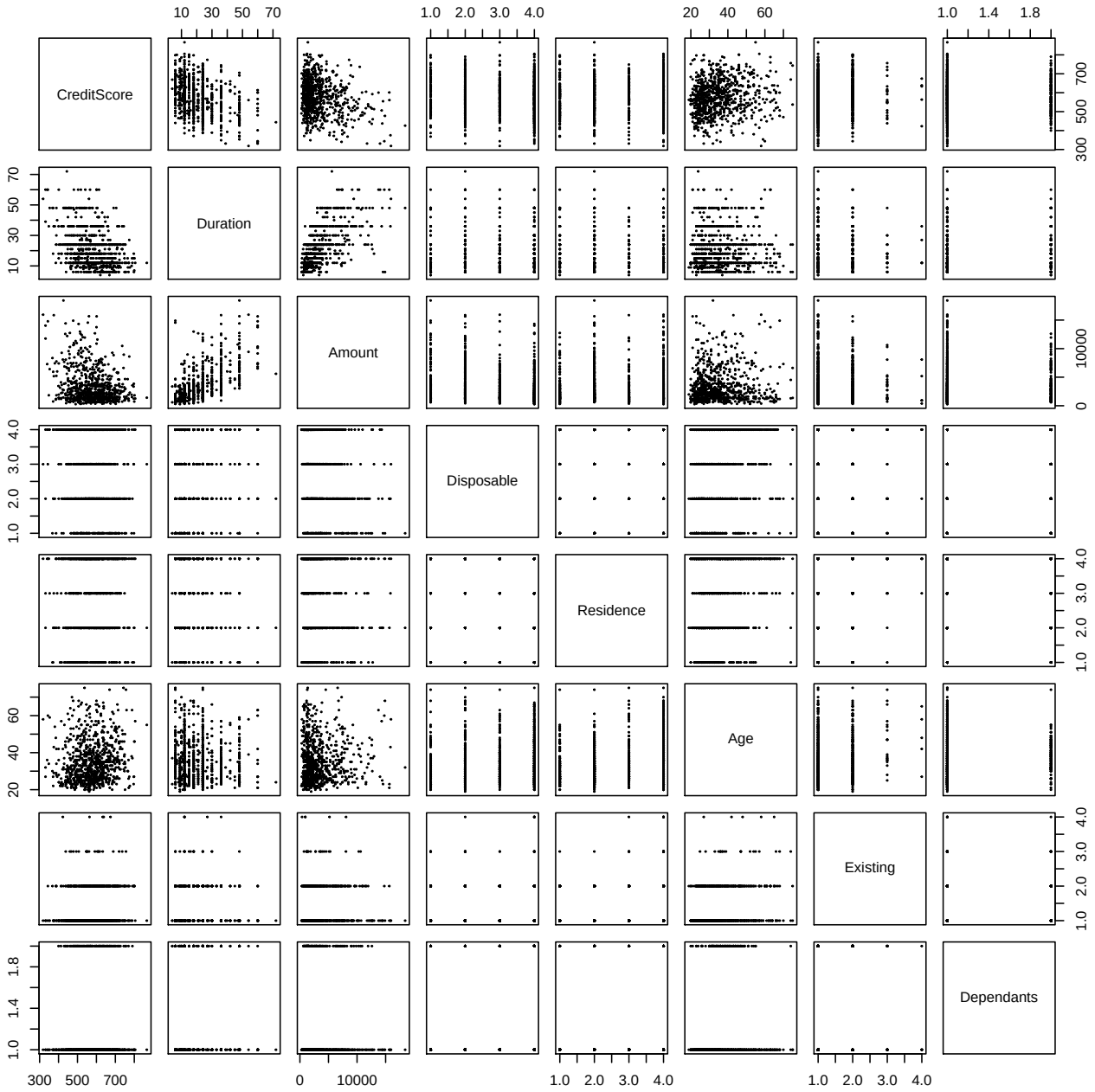
Exploratory Analysis

The data set consists of 20 measurements on 800 people. For each individual, the credit score has been measured, along the following 20 variables:

- Status: status of current account balance (“Negative”, “Small”, “Large”, “None”)
- Duration: duration of requested loan in months.
- History: status of previous loan history (“A”, “B”, “C”, “D”, “E”)
- Purpose: purpose of loan (“NewCar”, “UsedCar”, “Other”, “Furniture”, “Television”, “Domestic”, “Repairs”, “Education”, “Training”, “Business”)
- Amount: amount requested in Euros.
- Savings: balance of savings account (”Low”, ”Medium”, ”Large”, ”VeryLarge”, ”Unknown”)
- Employment: time in current employment (”Unemployed”, ”Short”, ”Medium”, ”Long”, ”VeryLong”)
- Disposable: the monthly repayment installments as a percentage of annual disposable income.
- Personal: personal status; (“M:DivSepMar”, ”F:DivSepMar”,”M:Single”, ”F:Single”)
- OtherParties: other parties with an interest (”None”, ”Coapp”, ”Guarantor”)
- Residence: full years in current residence.
- Property: most valuable significant asset (”House”, ”Savings”, ”Car”, ”None”)
- Age: age of applicant.
- Plans: other current loan plans (”Bank”, ”Stores”, ”None”)
- Housing: ownership status of accommodation (”Rent”, ”Own”, ”RentFree”)
- Existing: number of existing credits at this bank.
- Job: level of current job (”Unemployed”, ”Unskilled”, ”Skilled”, ”Management:Self”)
- Dependants: number of dependants.
- Telephone: whether the applicant has a registered phone in their name (”No”, ”Yes”)
- Foreign: whether the applicant is a foreign worker (”Yes”, ”No”)

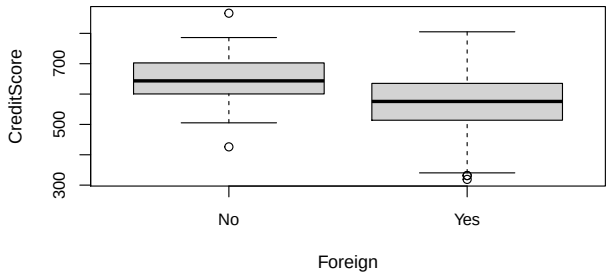
We should treat Duration, Amount, Disposable, Residence, Age, Existing and Dependants as continuous variables and should treat the other measurements as categorical variables (factors).

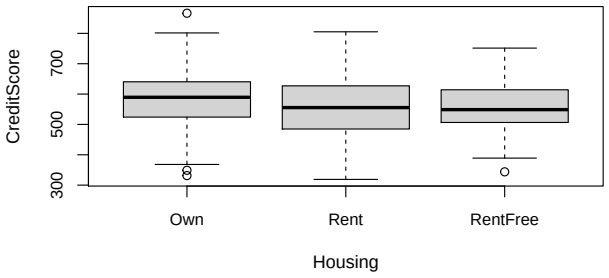
The following plot demonstrates the pairwise scatter plot for the continuous measurements.

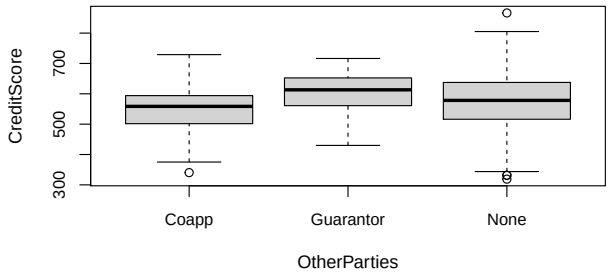


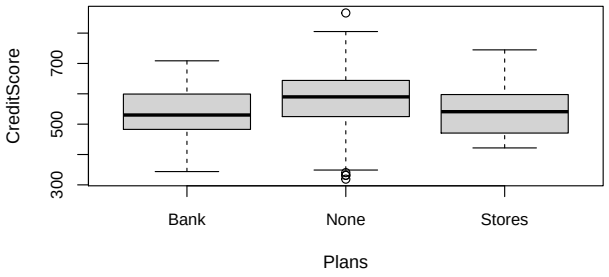
From the plot for the continuous measurements, we can see the association between several covariates and the credit scores of individuals including duration of requested loan, amount requested and age of applicant. Individuals with a shorter duration of requested loan and lower amount requested tend to have higher credit scores; the younger individuals are inclined to have a lower credit scores. However, the correlation between disposable income, years in current residence, existing credit scores in the bank, number of dependants and credit scores is not clear, and it is important to note that there is a obviously positive correlation between duration and amount. Hence, the relationship among these covariates and credit scores would be verified while modelling.

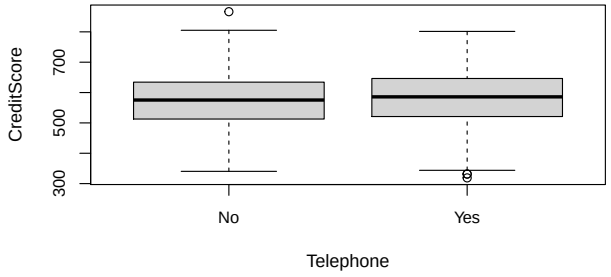
The factors (categorical variables) are considered.

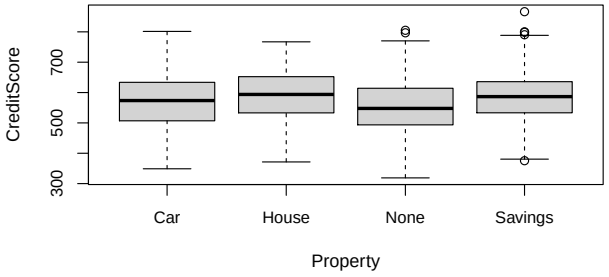

CreditScore
300
500
700
No
Yes
Foreign

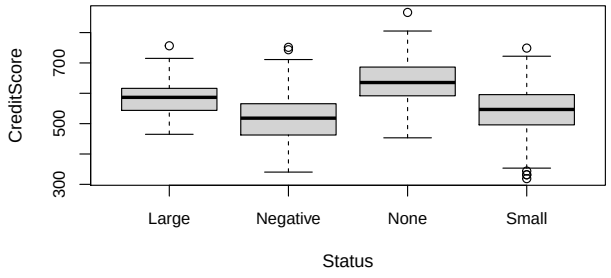

CreditScore
300
500
700
Own
Rent
RentFree
Housing

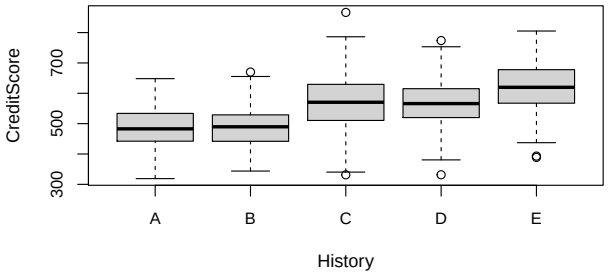

CreditScore
300
500
700
Coapp
Guarantor
None
OtherParties


CreditScore
300
500
700
Bank
None
Stores
Plans

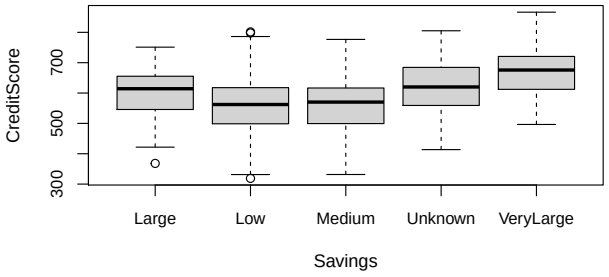

CreditScore
300
500
700
No
Yes
Telephone

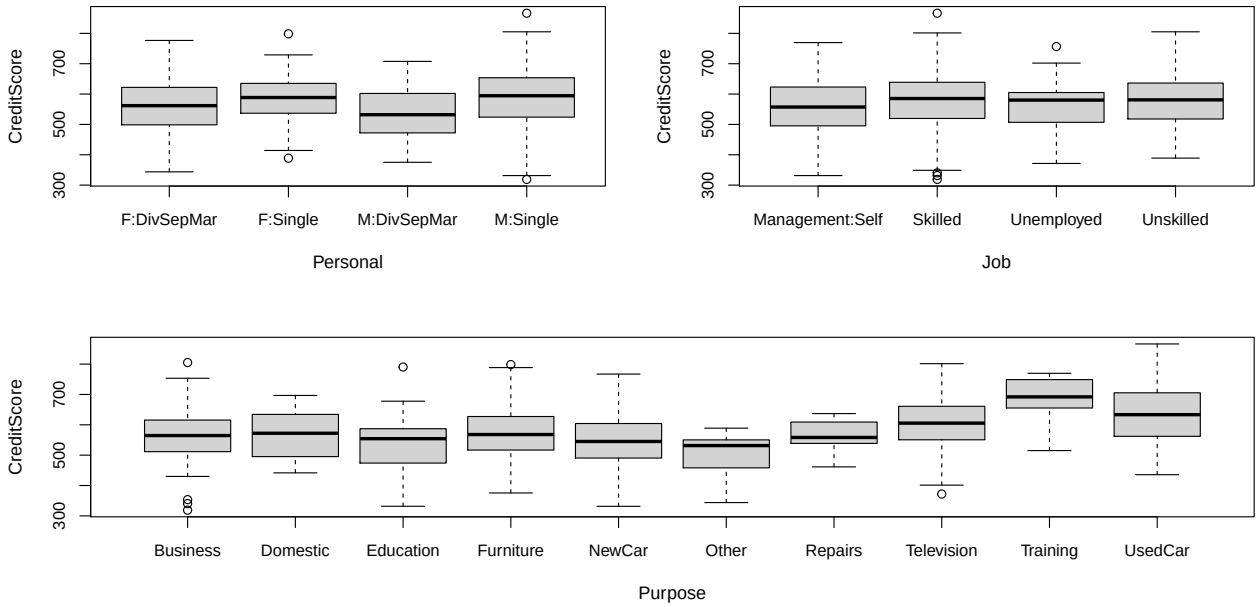

CreditScore
300
500
700
Car
House
None
Savings
Property


CreditScore
300
500
700
Large
Negative
None
Small
Status

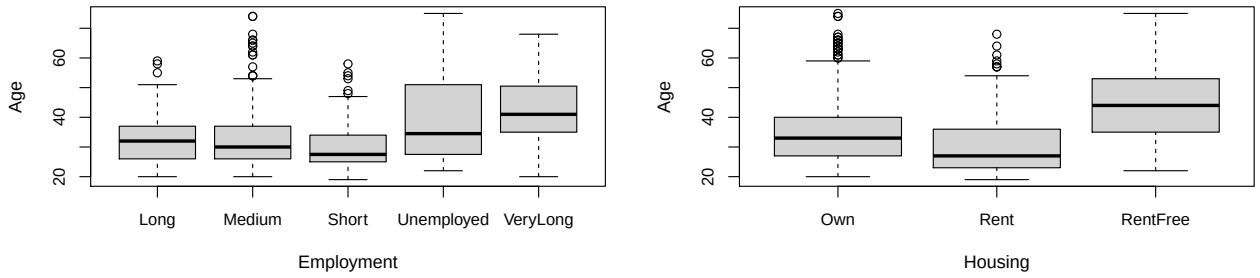

CreditScore
300
500
700
A
B
C
D
E
History


CreditScore
300
500
700
Long
Medium
Short
Unemployed
VeryLong
Employment


CreditScore
300
500
700
Large
Low
Medium
Unknown
VeryLarge
Savings



According to the box-plots above, for individuals, it appears there is an association between Status, History, Savings, Purpose, Employment, Personal, Otherparties, Property, Plans, Housing, Foreign and credit scores. Individuals who have large status of current account balance, critical delays of previous loan (level E in history), very large balance of savings, very long time in their current occupations (employment), training as purpose, guarantors as the other parties, their own houses, no plans, houses as the most valuable property, are single (personal), or not a foreigner tend to have better credit scores on average. Aside, the difference of two factors, Telephone and Job, are too exiguity; therefore, we cannot conclude whether this can be an indicator of the association with credit scores. Furthermore, it should again be noted that there is lots of correlation between factors and covariates. For example, elder people tend to have very long time in their current occupations and live in rent-free houses.



Similar relationships can be found between other variables — plots omitted, and these relationships would be examined by significance tests when attempting to draw conclusions or fitting models.

Modelling

We start with a full model (Model 1) with all 20 variables.

```
model_1 <- lm(CreditScore ~., data = Train)
```

```
summary(model_1)$coef
```

```
Coefficients (in part):
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.302e+02  2.437e+01  25.861  < 2e-16 ***
StatusNegative     -4.881e+01  8.175e+00  -5.971 3.64e-09 ***
```

```
Duration              -1.505e+00  2.031e-01  -7.410 3.41e-13 ***
HistoryE               7.737e+01  9.848e+00   7.857 1.36e-14 ***
PurposeNewCar         -4.467e+01  7.006e+00  -6.376 3.18e-10 ***
Amount                -6.289e-03  9.555e-04  -6.582 8.74e-11 ***
SavingsVeryLarge       6.422e+01  1.134e+01   5.663 2.12e-08 ***
EmploymentMedium      -1.491e-01  5.365e+00  -0.028 0.977831
EmploymentShort        8.074e-01  6.173e+00   0.131 0.895973
...
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the coefficient table, we find that Status, Duration, History, Purpose, Amount, Savings, Disposable, Personal, Plans, Housing, Telephone and Foreign have significance after controlling the other measurements. The p-value indicates a mild evidence of association between age and credit scores. However, the others measurements including Employment, Residence, Property, Existing, Job and Dependants have no obvious association between the credit scores of individuals. The p-values of these variates are quite big which show that there are high probabilities to accept the null hypothesis (the coefficient of the variate equals to zero). We will remove the variates, Residence, Existing, Dependants and Employment in the next model based on our exploratory analysis; we cannot recognize the relationship between these variates and credit scores from the plots. Further analysis will be made to check whether the simpler model will still have high accuracy and better generalisation.

```
model_2 <- lm(CreditScore ~ Status + Duration + History + Purpose
              + Amount + Savings + Disposable + Personal + OtherParties
              + Property + Age + Plans + Housing + Job + Telephone + Foreign,
              data = Train)
```

```
anova(model_1, model_2)
```

```
## Analysis of Variance Table
##
## Model 1: CreditScore ~ Status + Duration + History + Purpose + Amount +
##     Savings + Employment + Disposable + Personal + OtherParties +
##     Residence + Property + Age + Plans + Housing + Existing +
##     Job + Dependants + Telephone + Foreign
## Model 2: CreditScore ~ Status + Duration + History + Purpose + Amount +
##     Savings + Disposable + Personal + OtherParties + Property +
##     Age + Plans + Housing + Job + Telephone + Foreign
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1    751 1794621
## 2    758 1806447 -7    -11826 0.707 0.6662
```

```
summary(model_2)$coef
```

```
Coefficients (in part):
                       Estimate Std. Error t value Pr(>|t|)
PropertyHouse         -7.781e+00  4.889e+00  -1.592  0.11191
PropertyNone           5.700e-01  8.290e+00   0.069  0.94520
PropertySavings        2.536e+00  4.856e+00   0.522  0.60174
JobSkilled             1.749e+00  5.658e+00   0.309  0.75725
JobUnemployed         -1.321e+00  1.349e+01  -0.098  0.92197
JobUnskilled           6.862e+00  6.924e+00   0.991  0.32198
```

```
...
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We find the p-value is 0.6662 after using F-test between two nested models, so we do not reject the null hypothesis that all the coefficient of Residence, Existing, Dependants and Employment are all zero. In other words, there is no evidence that these variates have an association with credit scores. From the coefficient table of model 2, there is still no clear clue that Property and Job having significance by holding other variables fixed. We would eliminate these two factors, apply the F-test again and check the R^2_{adj} improves or not.

```
model_3 <- lm(CreditScore ~ Status + Duration + History + Purpose
              + Amount + Savings + Disposable + Personal + OtherParties
              + Age + Plans + Housing + Telephone + Foreign,
              data = Train)
```

```
anova(model_2, model_3)
```

```
## Analysis of Variance Table
##
## Model 1: CreditScore ~ Status + Duration + History + Purpose + Amount +
##     Savings + Disposable + Personal + OtherParties + Property +
##     Age + Plans + Housing + Job + Telephone + Foreign
## Model 2: CreditScore ~ Status + Duration + History + Purpose + Amount +
##     Savings + Disposable + Personal + OtherParties + Age + Plans +
##     Housing + Telephone + Foreign
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    758 1806447
## 2    764 1818970 -6    -12522 0.8758 0.5121
```

```
summary(model_1)$adj.r.squared
```

```
## [1] 0.7105007
```

```
summary(model_2)$adj.r.squared
```

```
## [1] 0.7112841
```

```
summary(model_3)$adj.r.squared
```

```
## [1] 0.7115658
```

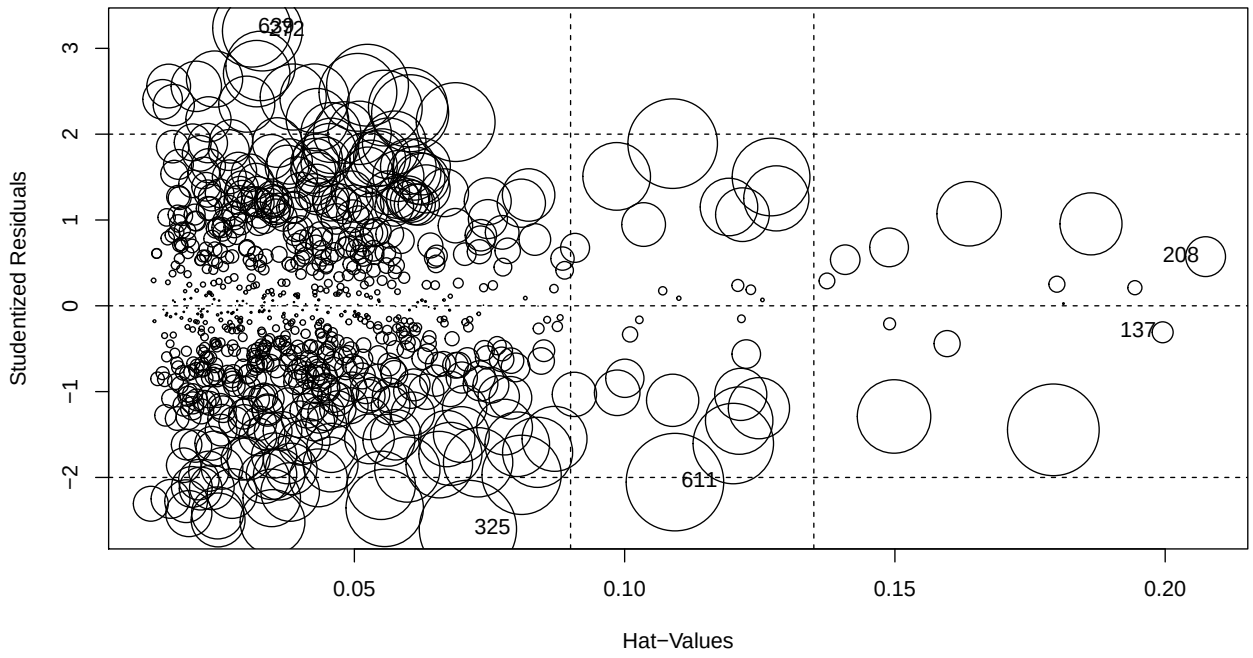
Again, we can find the p-value is 0.5121 ,so we do not reject the null hypothesis. As we can notice from the output of R^2_{adj}, Model 3 does improve by a very small amount comparing to Model 1 and Model 2. It reflects a fact that the remaining input variables are providing a useful information and are worth including in the model. In addition, we can use the AIC as the criterion to testify selection of models.

```
model_1_step_1 <- step(model_1)
```

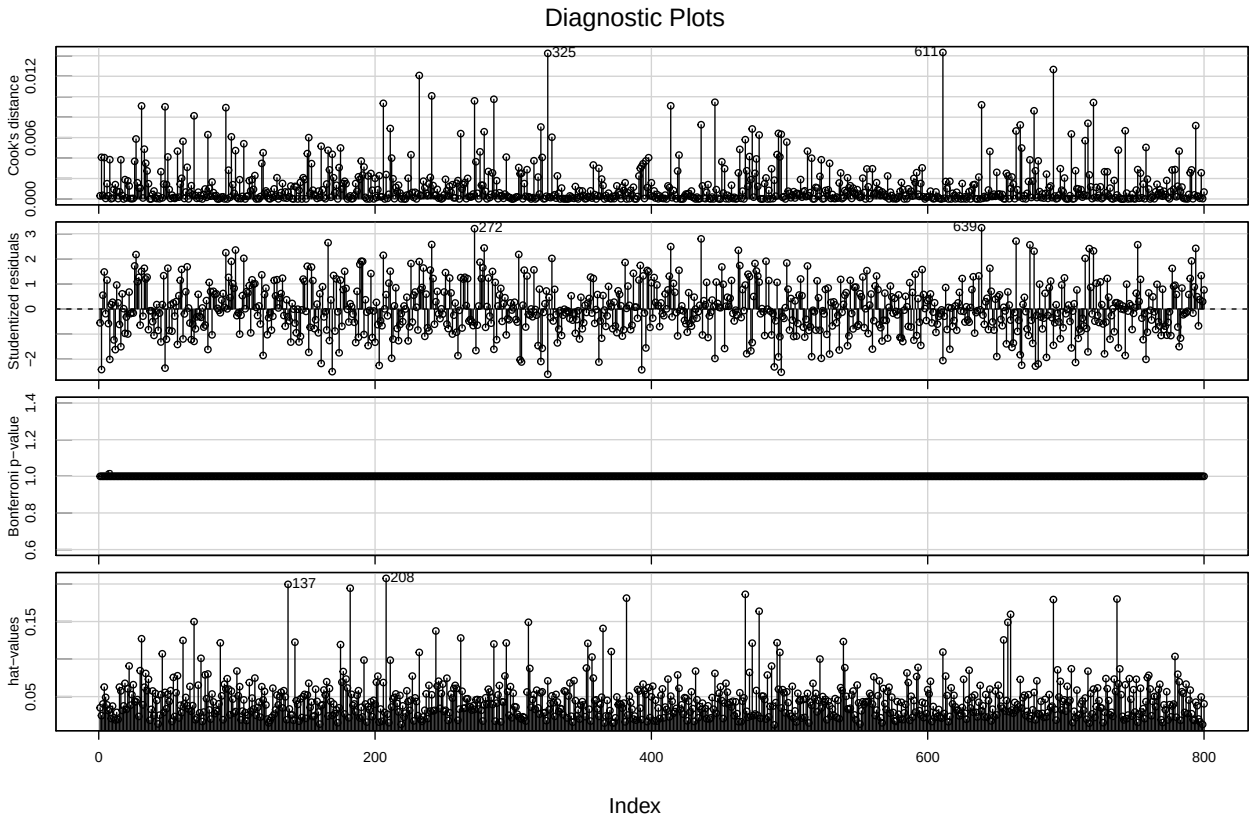
```
AIC=6258.31
CreditScore ~ Status + Duration + History + Purpose + Amount +
```

```
    Savings + Disposable + Personal + OtherParties + Residence +
    Age + Plans + Housing + Existing + Telephone + Foreign

Step:  AIC=6256.56
CreditScore ~ Status + Duration + History + Purpose + Amount +
    Savings + Disposable + Personal + OtherParties + Age + Plans +
    Housing + Existing + Telephone + Foreign

Step:  AIC=6255.34
CreditScore ~ Status + Duration + History + Purpose + Amount +
    Savings + Disposable + Personal + OtherParties + Age + Plans +
    Housing + Telephone + Foreign
```

In general, the smaller AIC presents a better model. The result consists with Model 3 which contains the same input variates so we choose Model 3 as the best model so far. Then, we can use the diagnostic tools to find out the influential data for revising our model.




```
##         StudRes        Hat        CookD
## 137 -0.3087751 0.19953768 0.0006609682
## 208  0.5728890 0.20747923 0.0023888261
## 272  3.2016675 0.03297630 0.0095937359
## 325 -2.5996094 0.07100426 0.0142404449
## 611 -2.0545249 0.10930796 0.0143290486
## 639  3.2373574 0.03102072 0.0092057726
```



It is immediately obvious that there are six very influential points. Whereas we are unable to correct them, the only thing we can do is to remove the offending data points and refit the model.

```
Train2 <- Train[-c(137,208,272,325,611,639), ]

model_4 <- lm(CreditScore ~ Status + Duration + History + Purpose
              + Amount + Savings + Disposable + Personal + OtherParties
              + Age + Plans + Housing + Telephone + Foreign,
              data = Train2)

summary(model_4)$adj.r.squared
```

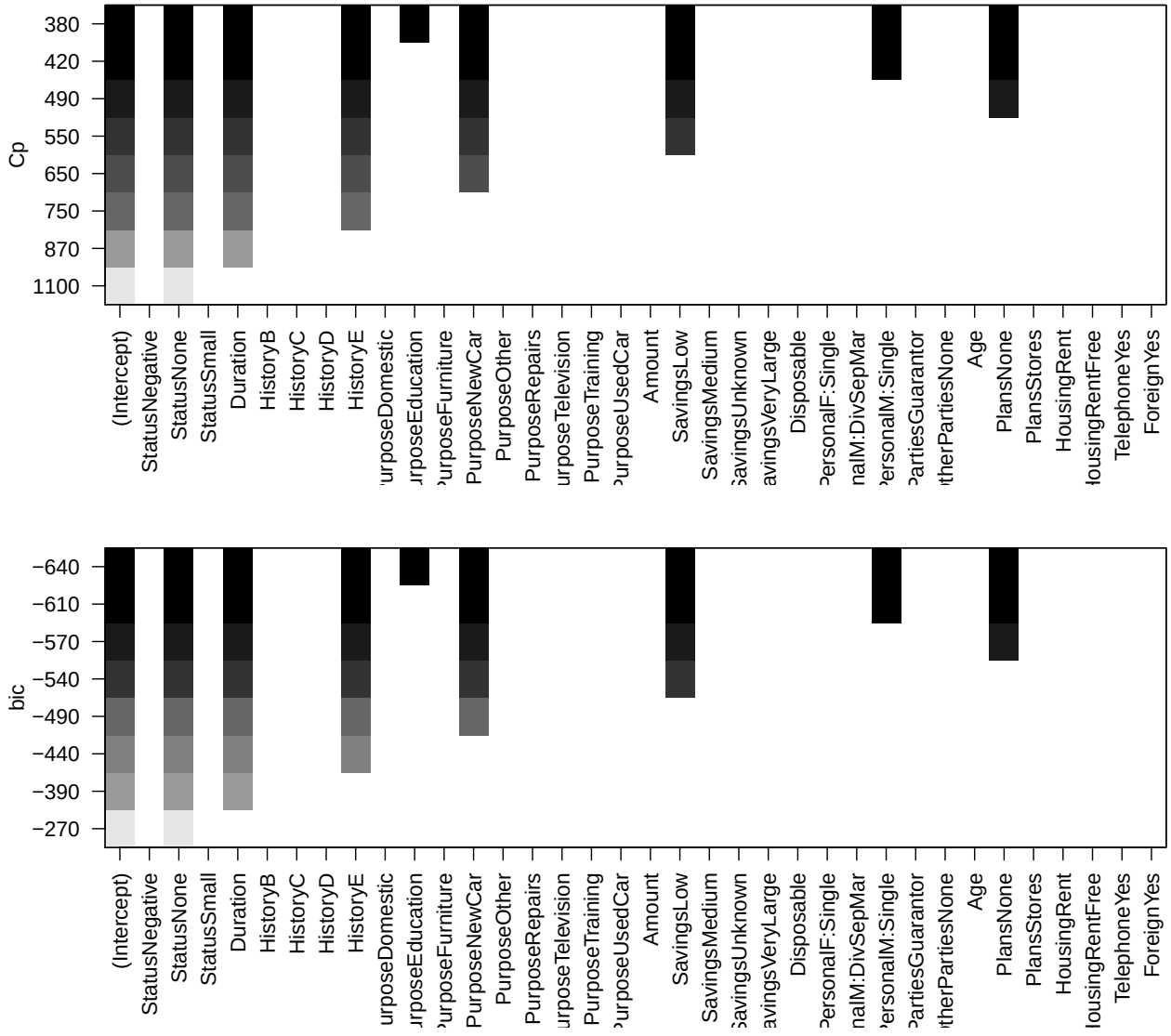
```
## [1] 0.7173649
```

```
compareCoefs(model_3, model_4)
```

```
(in part)              Model 1 (SE)       Model 2 (SE)
(Intercept)              618.9 (20.9)       616.2 (20.5)
StatusNegative           -49.80 (8.03)      -50.55 (7.90)
StatusNone                38.40 (7.82)       37.90 (7.69)
StatusSmall              -28.57 (8.14)      -28.50 (8.01)
Duration                 -1.465 (0.20)      -1.472 (0.20)
HistoryB                   -3.9 (12.2)        -3.9 (12.1)
HistoryC                 44.03 (9.37)       43.27 (9.21)
...
```

Having removing these data points, all the variates are still remain significant and the R^2_{adj} increases by a small value. It shows that although eliminating high leverage points does have an effect on the estimated coefficients, it is difficult to specify the main reason influencing the credit scores of these six individuals because there are 14 different variates involved in the model. As the results, we should leave them in the data set.

In order to further narrow down the model, the best subset regression is used to find better and simpler models. The higher Cp and the lower BIC values can indicate a valid and generalised model. In the following part, models are tested with two other criteria correspondingly: Cp, and BIC.



According to the Cp and BIC plots, we are suggested to test the following three simpler models:

```
model_5 <- lm(CreditScore ~ Status + Duration, data = Train)
model_6 <- lm(CreditScore ~ Status + Duration + History, data = Train)
model_7 <- lm(CreditScore ~ Status + Duration + History + Purpose, data = Train)
```

For the data set Train, mean squared prediction error of each model is:

Model 1 (full model)

```
## [1] 2243.276
```

Model 3

```
## [1] 2273.712
```

Model 5

```
## [1] 4703.737
```

Model 6

```
## [1] 4165.024
```

Model 7

```
## [1] 3487.166
```

As reported by above test set, Model 1 (full model) gives the best performance. The suggested models given by Cp and BIC can be helpful if we have limited data, but the consequence is that these models omit too many useful variates which leads to larger mean square errors comparing to full model and Model 3. Both full model and Model 3 are worth further analyses since the difference of mean squared prediction error of these models is quite small. Thus, both these two models are chosen for predictions.

Predictions

In the prediction part, Test data is used.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   380.3   526.0   596.1   592.6   656.6   794.7
```

Model 3: CreditScore ~ Status + Duration + History + Purpose + Amount + Savings + Disposable + Personal + OtherParties + Age + Plans + Housing + Telephone + Foreign

The mean squared prediction error and summary of Model 3 are:

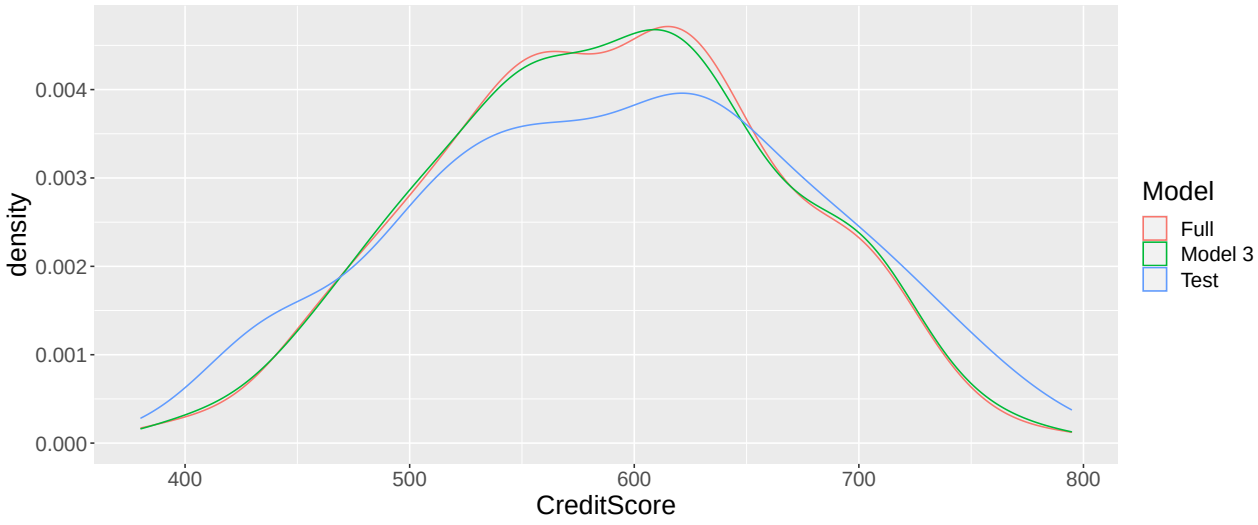
```
## [1] 2431.625
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   395.5   538.4   590.9   589.1   640.3   785.3
```

In the Model 1 (full model), mean squared prediction error and summary are:

```
## [1] 2460.575
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   385.6   538.6   592.1   588.9   640.2   791.3
```



The mean squared prediction error on Test in Model 3 (2431.625) is less than it is in the full model (2460.575) which indicates Model 3 fits a bit better. From the quartile table, there is no great difference between Test, full model and Model 3. However, it is easy to notice that predictions of full model and Model 3 are both tighter than Test data is. This means that both two models predict less variably comparing to the reality. The non-linear model would be able to over come this problem and make more precise forecasts, but it increases the complexity of the modelling process. After balancing the simplicity and effectiveness of modelling process, Model 3 is still the best model.

Classification

We now classify the individuals into two groups, bad risk and good risk, by the best model's predicted credit scores on Test. Individuals will be sorted into "1" if predicted credit score is less than 500 and "0" otherwise.

```
Test_risk <- ifelse(Test$CreditScore < 500,1,0)
Predictions_risk <- ifelse(predictions3 < 500,1,0)
sum(Test_risk == Predictions_risk) / length(Test_risk)
```

```
## [1] 0.86
```

The output shows the proportion of the individuals are correctly classified by the best model.

Summary

We found an evidence that Status, Duration, History, Purpose, Amount, Saving, Disposable, Personal, OtherParties, Age, Plans, Housing, Telephone and Foreign have association with credit scores of individuals; the other variables were not found to be associated with credit scores.

The data provided is initially dividing into Train and Test. If the ratio of the data set are changed, the predictive model may be changed as well. Our modelling process is based on exploratory analysis, F-test, AIC criterion and mean squared prediction error of Train. Although, Cp and BIC provide simpler models, their effectiveness are not quite good since they sacrifice too many useful information. By comparing all the potential models, we find that the best model, Model 3, is informative and with relatively better R^2_{adj}.

In the prediction section, the best model has a slight better performance than full model by considering the simplicity and effectiveness. It has 86% of accuracy while classifying the bad/good risks of individuals.