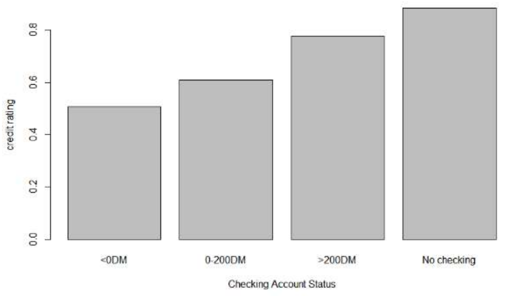
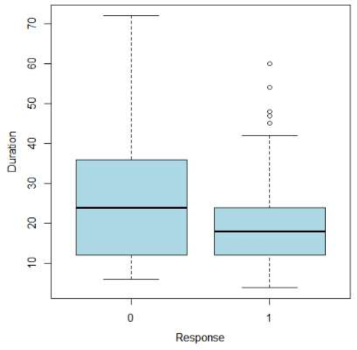
RA – Take Home Assignment (Quiz -2)
Submitted by Kanika Gupta

Part A
The Response variable is the probability of getting a good Credit Rating ("1")
It is evident from the summary statistic that there are no missing values in the dataset.

The explanatory variables

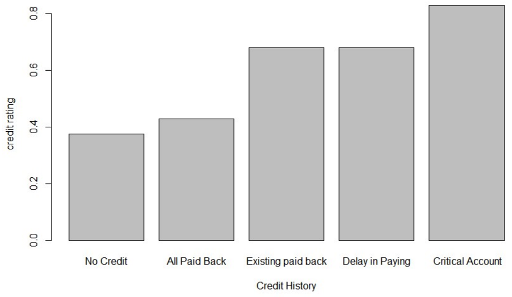
CHK_ACCT (Checking Account Status)
As indicated by the below graph, as the Amount in Account increases, the probability of being creditworthy increases. But the exception is that a person having no checking account is more creditworthy than with account balance which seems like exception.





DURATION (Duration of Credit in Months)
From this box plot used for continuous variable, we can see that as the Duration of Credit increases, the probability of getting credit rating decreases. This means that short duration credits are preferred by the lending agencies.

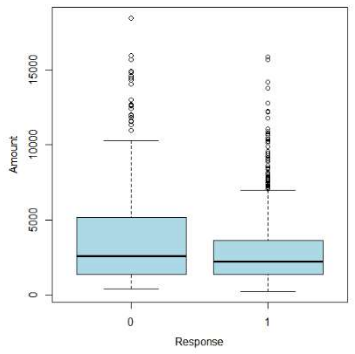
History (Credit History): Consumers with "critical" credit history show large increase in expected odds of creditworthiness. The odds for these customers relative to customers with fully paid credits, fully paid credits at this bank and those with existing credits paid back are greater. This result is counterintuitive as it suggests that consumers with worse credit history are less likely to default. This might be due to a form of "bias" associated with the way loans are issued - the bank may be more stringent when it comes to loaning a consumer with bad credit history, whereas consumers with good credit history do not face the same kind of scrutiny and may end up being issued a loan they eventually cannot repay. An alternative explanation is that there may be a data issue in which the categories were incorrectly labelled.

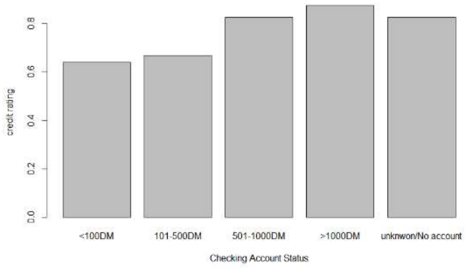


Purpose of Credit:

Purpose of Credit	Probability of getting Good Credit Rating
New_car	lower
Old_car	higher
Furniture	not significant
Radio/TV	not significant
Education	lower chances of getting approval
Retraining	not significant

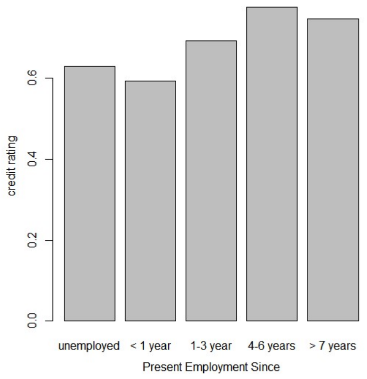
Credit Amount
The probability of getting good credit rating decreases as the credit amount increases, which means that the lower credit amounts are preferred.

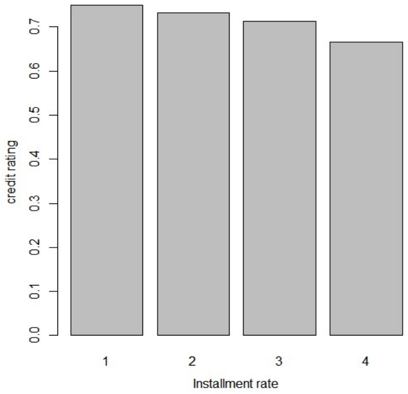




SAV_ACC (Average Balance is Savings Account):
From the bar graph we can see that as the Balance amount increases, the probability of getting Credit increases. The exception here is the unknown or no account.

Employment: The chances of getting credit approval increase as the employment years increase which is also intuitive as the stability increase with increase in employment years.





INSTALL_RATE (Instalment rate as % of income)- When a bank extends a loan, it ensures that the total outgo on EMIs does not exceed certain percentage limit of your take home salary/income. This is done to prevent you from defaulting by taking larger loans that you can reasonably service. This bar graph shows that as this percent increases the chances of getting load or good credit rating reduces.

Gender
We can also see from the graphs that if a person is male and single the chances of getting credit is higher than if the person is a single Female. While the credibility does not change when the person is male or female and married or widowed. The exception here is that when the person is male and divorced the chances are less of getting credit.

Guarantor and co applicant
If there is a guarantor then the chances of getting a credit increase.
As per the data applying with a co-applicant is reducing the chances of getting credit which is against the intuition.

Real Estate, Prop_unknown or none, Age

If a person has real estate the chance of getting credit increases while it reduces when a person has no property. The chances increase with the increase in Age as with age the financial stability increases.

Other_Install

If the person applying for a loan has other instalments running has lower chances of getting approval.

Rent, Own_Residence and Present Residence

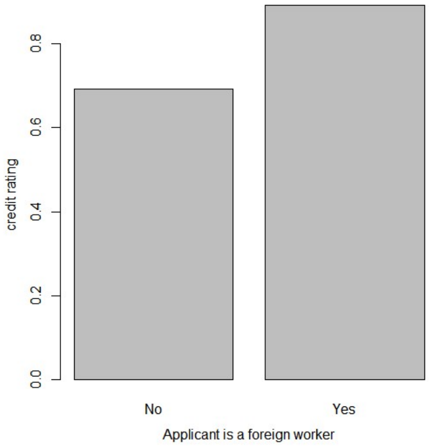
A person living on rent has lower chances while having own residence increase the chances. The Present_resident is not affecting the Credit Rating significantly.

Num_Credits, Job Nature, Num_Dependants and

These variables are not significant as per the initial data study and the Credit rating is not changing significantly for these variables.

Telephone and Foreign worker

If a person has a Telephone in his/her name or is a foreign worker, the chances of getting good credit rating are higher.

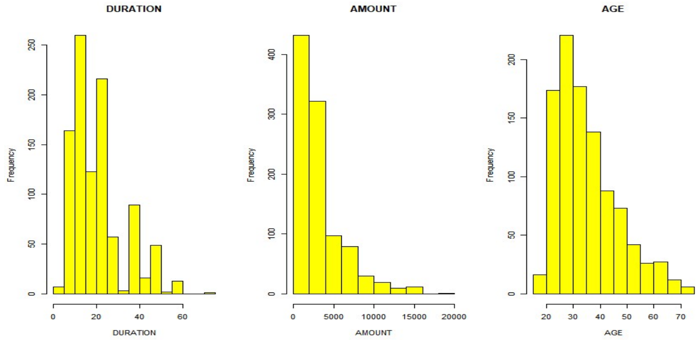


For the variables Age, Amount and Duration are converted to bins as the data is highly skewed

For age four brackets are created “1-20", "21-30","31-40","50+"

For Amount the brackets created are "1-4000", "4001-8000", "8001-12000","12000+"

For Duration, the brackets created are "1-20","21-40","41-60","60+"



Factors are created in R for the following variables with their corresponding base cases to see the effect of each level

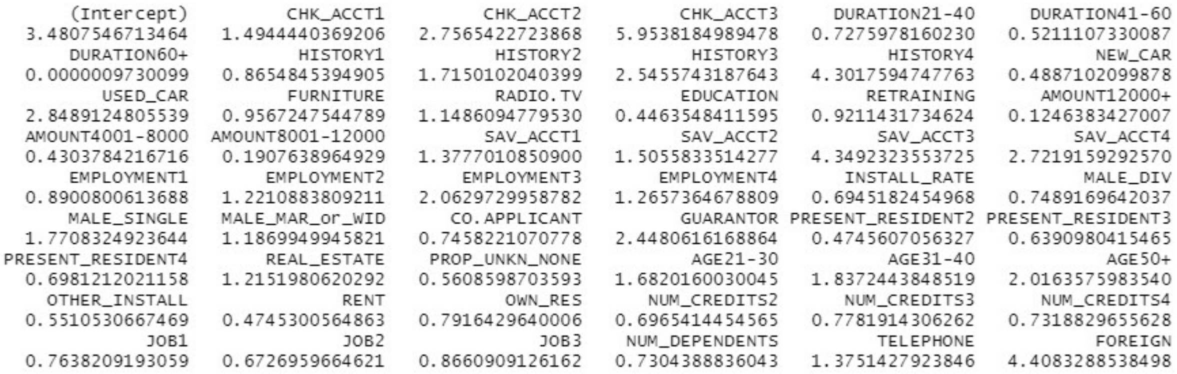
CHK_ACCT	0: < 100DM
History	0: No Credits Taken
SAV_ACCT	0: <100DM
Employment	0: Unemployed
Present_Resident	0: < Less than a year
Job	0: Unemployed/Unskilled-non-resident
NUM_CREDITS	1: Number of existing Credit is 1
Age	1-20 years
Amount	1-4000 DM
Duration	1-20 months

For unit level data logit transformation does not work. We must consider **maximum likelihood method** of estimation.

Part b.

Running the regression model using all the variables and result interpretation (Model can be seen from the R Output)

The odds ratios for all the variables are as below



```
      (Intercept)         CHK_ACCT1         CHK_ACCT2         CHK_ACCT3     DURATION21-40     DURATION41-60 
  3.4807546713464   1.4944440369206   2.7565422723868   5.9538184989478   0.7275978160230   0.5211107330087 
      DURATION60+          HISTORY1          HISTORY2          HISTORY3          HISTORY4           NEW_CAR 
  0.0000009730099   0.8654845394905   1.7150102040399   2.5455743187643   4.3017594747763   0.4887102099878 
         USED_CAR         FURNITURE          RADIO.TV         EDUCATION        RETRAINING      AMOUNT12000+ 
  2.8489124805539   0.9567247544789   1.1486094779530   0.4463548411595   0.9211431734624   0.1246383427007 
  AMOUNT4001-8000  AMOUNT8001-12000         SAV_ACCT1         SAV_ACCT2         SAV_ACCT3         SAV_ACCT4 
  0.4303784216716   0.1907638964929   1.3777010850900   1.5055833514277   4.3492323553725   2.7219159292570 
      EMPLOYMENT1       EMPLOYMENT2       EMPLOYMENT3       EMPLOYMENT4      INSTALL_RATE          MALE_DIV 
  0.8900800613688   1.2210883809211   2.0629729958782   1.2657364678809   0.6945182454968   0.7489169642037 
      MALE_SINGLE   MALE_MAR_or_WID      CO.APPLICANT         GUARANTOR PRESENT_RESIDENT2 PRESENT_RESIDENT3 
  1.7708324923644   1.1869949945821   0.7458221070778   2.4480616168864   0.4745607056327   0.6390980415465 
PRESENT_RESIDENT4       REAL_ESTATE    PROP_UNKN_NONE          AGE21-30          AGE31-40            AGE50+ 
  0.6981212021158   1.2151980620292   0.5608598703593   1.6820160030045   1.8372443848519   2.0163575983540 
    OTHER_INSTALL              RENT           OWN_RES      NUM_CREDITS2      NUM_CREDITS3      NUM_CREDITS4 
  0.5510530667469   0.4745300564863   0.7916429640006   0.6965414454565   0.7781914306262   0.7318829655628 
             JOB1              JOB2              JOB3    NUM_DEPENDENTS         TELEPHONE           FOREIGN 
  0.7638209193059   0.6726959664621   0.8660909126162   0.7304388836043   1.3751427923846   4.4083288538498 
```

AIC: 993.09 BIC: 1258.109 -2logL: 'log Lik.' 885.0899 (df=54)

The Regression Coefficient (as per the code output) which are greater than 0(less than 0) means that the Odds ratio of xi = 1 to xi = 0 keeping the other x's fixed is more than 1 (less than 1). The base odds will increase by the corresponding factor when the odds ratio is more than 1 and will reduce when the odds ratio is less than 1.From the model Summary(Code Output) we can say that the estimates of the coefficients for Check_account, Duration, History, New car, Used Car , Amount, Savings Account, Install _rate, Male Single, Guarantor, Present resident, Other Install and Foreign are significant as they have low p-values. The performance of each category is evaluated w.r.t. their base category.

For the variable **CHK_ACCT1(0-200DM)**, the odds ratio is 1.491, **CHK_ACCT2(>200DM)** the odds ratio is 2.756, **CHK_ACCT3(no Checking account)** the odds ratio is 5.953 which means that a change in Checking account status w.r.t the base case, keeping other variables constant leads to an increase in the probability of credit approval by their corresponding factors. It is surprising to see that the applicant with no Checking account are more creditworthy.

For the variable **New_car** the odds ratio is 0.488 which means that as this variable changes from 0 to 1, keeping other variables constant, the probability of credit approval reduces by a factor of 0.452. And when the purpose of loan is **Old_car** the chances of getting loan approval increases by a factor 2.84. The probability of getting Good credit rating decrease when the purpose of the loan is **Furniture**, **Education** and **Retraining**, and increase when the purpose is **Radio/TV**. For every, one unit increase in **Install_rate**, the chances of getting approval reduces.
As the **Duration and Amount** brackets are increasing the expected odds of creditworthiness are reducing. If the person has **Savings_ACCT** 3 (>1000DM), the chances of loan approval are highest compared to other levels w.r.t to the base category. It is surprising to see that the **Savings_Acct 4** (With no account) is more creditworthy than a **Savings Account 1**(101-500DM) and **Savings Account 2** (500-1000DM).
Bad Credit **History** (**HISTORY4**) has the highest odds ratio as compared to other levels. The chances of getting Credit approval are less if the person has **other_installments** running, is living on **Rent** or has an **Own Residence. Male_Single or Male_Mar or Wid** is more Creditworthy than the females with same status, while **Male_Div** is less creditworthy than the females who are divorced.
Creditworthiness is increasing as the **Age** is increading.
Loan approval chances reduce if there is a **Co-applicant** and increase if there is **Guarantor**.
With a unit increase in the number of **Num_Dependents**, the odds reduce by a factor of 0.69.
The chances increase it the Applicant has **telephone** in his/her name or is a **Foreign Worker**.
Similary the effect of the remaining variables can be interpreted.

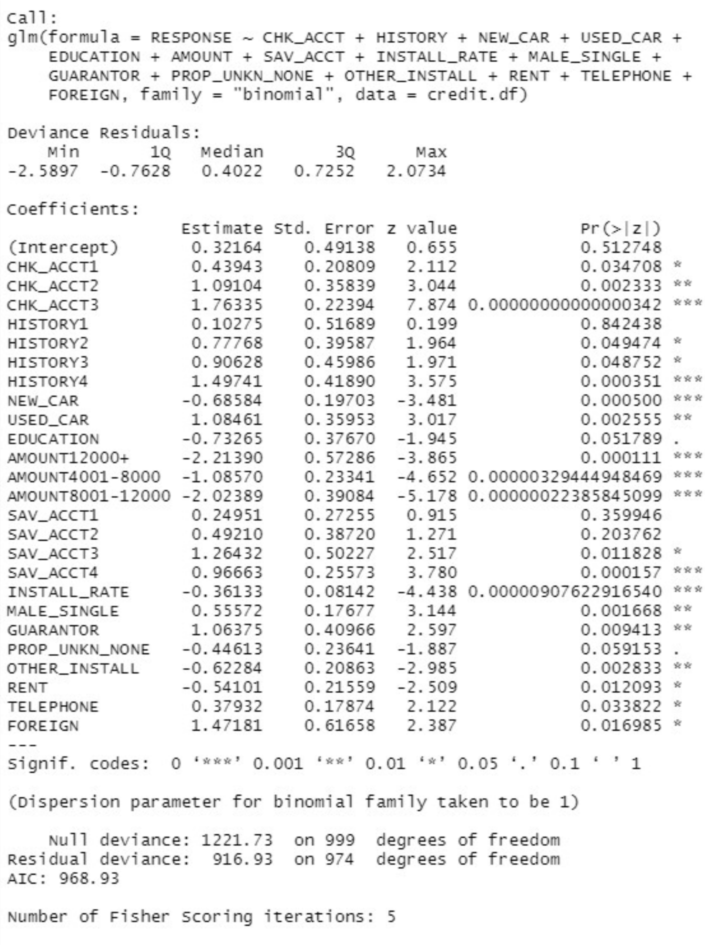
Part c

In evaluating whether a credit approval can be given to an applicant both explanation and prediction models are important. The Explanatory model helps to explain how the applicant got the rating, as in which factors where significant in deciding the credit rating 0(bad) or 1(good). While the prediction model is important to correctly predict that an applicant would be a defaulter or not. And in the banking scenario, it is more important to correctly predict a false positive than a false negative as a false positive will cause losses to the crediting agency.

Explanatory model

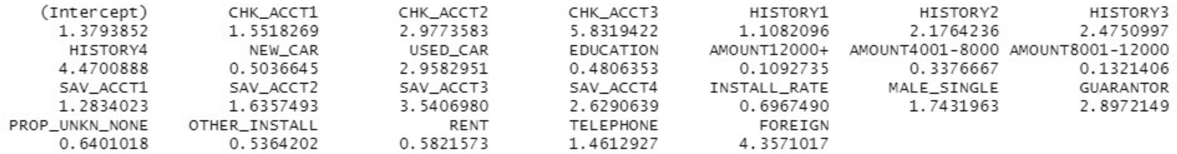
The full model is reduced using backward selection method for variable selection and the variable Present_resident is removed after that as it was not significant also seen from the initial data study and the AIC and BIC reduced further.

Model Summary



```
Call:
glm(formula = RESPONSE ~ CHK_ACCT + HISTORY + NEW_CAR + USED_CAR +
    EDUCATION + AMOUNT + SAV_ACCT + INSTALL_RATE + MALE_SINGLE +
    GUARANTOR + PROP_UNKN_NONE + OTHER_INSTALL + RENT + TELEPHONE +
    FOREIGN, family = "binomial", data = credit.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5897  -0.7628   0.4022   0.7252   2.0734

Coefficients:
                 Estimate Std. Error z value           Pr(>|z|)
(Intercept)       0.32164    0.49138   0.655           0.512748
CHK_ACCT1         0.43943    0.20809   2.112           0.034708 *
CHK_ACCT2         1.09104    0.35839   3.044           0.002333 **
CHK_ACCT3         1.76335    0.22394   7.874 0.0000000000000342 ***
HISTORY1          0.10275    0.51689   0.199           0.842438
HISTORY2          0.77768    0.39587   1.964           0.049474 *
HISTORY3          0.90628    0.45986   1.971           0.048752 *
HISTORY4          1.49741    0.41890   3.575           0.000351 ***
NEW_CAR          -0.68584    0.19703  -3.481           0.000500 ***
USED_CAR          1.08461    0.35953   3.017           0.002555 **
EDUCATION        -0.73265    0.37670  -1.945           0.051789 .
AMOUNT12000+     -2.21390    0.57286  -3.865           0.000111 ***
AMOUNT4001-8000  -1.08570    0.23341  -4.652 0.0000032944948469 ***
AMOUNT8001-12000 -2.02389    0.39084  -5.178 0.0000002238584509 ***
SAV_ACCT1         0.24951    0.27255   0.915           0.359946
SAV_ACCT2         0.49210    0.38720   1.271           0.203762
SAV_ACCT3         1.26432    0.50227   2.517           0.011828 *
SAV_ACCT4         0.96663    0.25573   3.780           0.000157 ***
INSTALL_RATE     -0.36133    0.08142  -4.438 0.0000090762291654 ***
MALE_SINGLE       0.55572    0.17677   3.144           0.001668 **
GUARANTOR         1.06375    0.40966   2.597           0.009413 **
PROP_UNKN_NONE   -0.44613    0.23641  -1.887           0.059153 .
OTHER_INSTALL    -0.62284    0.20863  -2.985           0.002833 **
RENT             -0.54101    0.21559  -2.509           0.012093 *
TELEPHONE         0.37932    0.17874   2.122           0.033822 *
FOREIGN           1.47181    0.61658   2.387           0.016985 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1221.73  on 999  degrees of freedom
Residual deviance:  916.93  on 974  degrees of freedom
AIC: 968.93

Number of Fisher Scoring iterations: 5
```

Odds Ratio



```
     (Intercept)        CHK_ACCT1        CHK_ACCT2        CHK_ACCT3         HISTORY1         HISTORY2         HISTORY3
       1.3793852        1.5518269        2.9773583        5.8319422        1.1082096        2.1764236        2.4750997
        HISTORY4          NEW_CAR         USED_CAR        EDUCATION     AMOUNT12000+  AMOUNT4001-8000 AMOUNT8001-12000
       4.4700888        0.5036645        2.9582951        0.4806353        0.1092735        0.3376667        0.1321406
       SAV_ACCT1        SAV_ACCT2        SAV_ACCT3        SAV_ACCT4     INSTALL_RATE      MALE_SINGLE        GUARANTOR
       1.2834023        1.6357493        3.5406980        2.6290639        0.6967490        1.7431963        2.8972149
  PROP_UNKN_NONE    OTHER_INSTALL             RENT        TELEPHONE          FOREIGN
       0.6401018        0.5364202        0.5821573        1.4612927        4.3571017
```

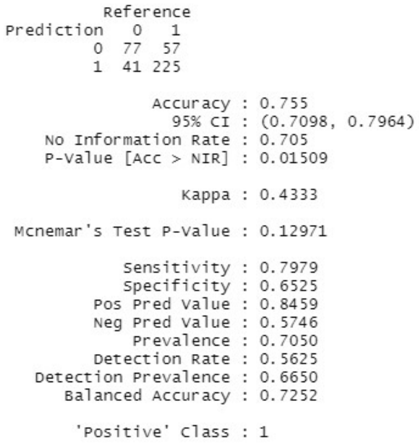
AIC: 968.93 BIC: 1096.536 -2logL: 'log Lik.' 916.9346 (df=26)

We can see that the AIC and BIC have both reduced from the full model. There is no further scope of removing any variable as it leads to higher AIC, BIC and Log lik values.

This model gives the insight that the credits for **Amount** in the bracket of “1-4000" has highest chance of loan approval as the odds ratio is less than 1 for other sub levels.
Among the **Purpose of loan** category only **New_car**, **Old_car** and education are **significant**.
The chances of getting Credit approval are less if the person has **other_installments**, is living on **Rent**, is applying for **New_car** or **Education** purpose. The chances increase it the Applicant has **telephone** in his/her name or is a **Foreign Worker**.
If the person has **Savings_ACCT** 3 (>1000DM), the chances of loan approval are highest compared to other levels w.r.t to the base category.
As the **Instalment** rate increases by a unit, the chances of getting credit approval reduces the base odds by a factor of 0.696.
If a person has no **Property** or the property is unknown the chances are less than if a person has a Property.
The gender of a person is significant in explaining the Credit approval Probability only when the person is single i.e. If a person is **Male and single**, the chances are higher to get a credit approval.
For the **CHK_ACCT** and **History** the odds ratio can be interpreted similarly for all the levels.

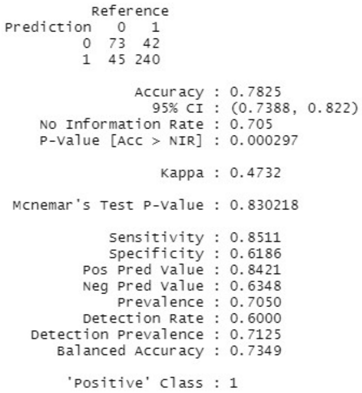
part d) Prediction model
The data is bifurcated into 60% Training and 40% Test data.
The customers in the validation set are classified into two categories, "Credit Rating- Good" and “Credit Rating- Not Good" using the model built from the training data.
Using the model, we got in part c) and applying that model to our Training data and then obtaining the prediction values for the validation data set, the Correlation between Predicted values and Actual values on the test data is **39.5%**
We have generated the confusion matrix to get the required Sensitivity and corresponding cut off values.The columns are the Actual values and rows are the Predicted values in the confusion matrix.

i) To get the Sensitivity as 80%, the Cut off is **0.639**



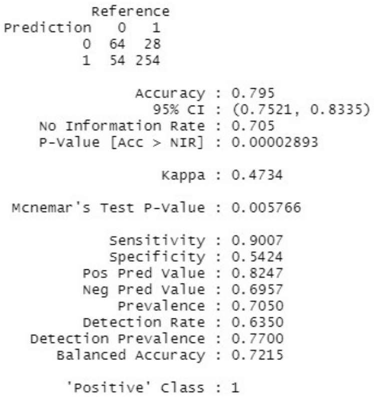
```
          Reference
Prediction   0   1
         0  77  57
         1  41 225

               Accuracy : 0.755
                 95% CI : (0.7098, 0.7964)
    No Information Rate : 0.705
    P-Value [Acc > NIR] : 0.01509

                  Kappa : 0.4333

 Mcnemar's Test P-Value : 0.12971

            Sensitivity : 0.7979
            Specificity : 0.6525
         Pos Pred Value : 0.8459
         Neg Pred Value : 0.5746
             Prevalence : 0.7050
         Detection Rate : 0.5625
   Detection Prevalence : 0.6650
      Balanced Accuracy : 0.7252

       'Positive' Class : 1
```

ii) To get the Sensitivity as 85%, the Cut off is **0.595**



```
          Reference
Prediction   0   1
         0  73  42
         1  45 240

               Accuracy : 0.7825
                 95% CI : (0.7388, 0.822)
    No Information Rate : 0.705
    P-Value [Acc > NIR] : 0.000297

                  Kappa : 0.4732

 Mcnemar's Test P-Value : 0.830218

            Sensitivity : 0.8511
            Specificity : 0.6186
         Pos Pred Value : 0.8421
         Neg Pred Value : 0.6348
             Prevalence : 0.7050
         Detection Rate : 0.6000
   Detection Prevalence : 0.7125
      Balanced Accuracy : 0.7349

       'Positive' Class : 1
```

iii) To get the Sensitivity as 90%, the Cut off is **0.52**



```
          Reference
Prediction   0   1
         0  64  28
         1  54 254

               Accuracy : 0.795
                 95% CI : (0.7521, 0.8335)
    No Information Rate : 0.705
    P-Value [Acc > NIR] : 0.00002893

                  Kappa : 0.4734

 Mcnemar's Test P-Value : 0.005766

            Sensitivity : 0.9007
            Specificity : 0.5424
         Pos Pred Value : 0.8247
         Neg Pred Value : 0.6957
             Prevalence : 0.7050
         Detection Rate : 0.6350
   Detection Prevalence : 0.7700
      Balanced Accuracy : 0.7215

       'Positive' Class : 1
```

Sensitivity = TP/(TP+FN)

Specificity = TN/ (TN + FP)

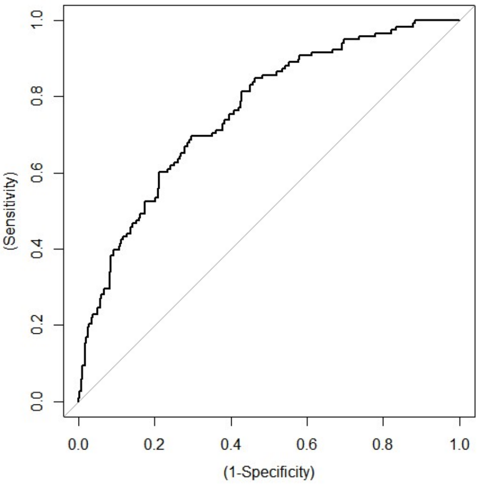
Sensitivity is the proportion of Predicted good Credit Rating to the actual good credit Ratings and specificity is the proportion of Predicted Bad Credit Rating to the Actual Bad Credit Ratings.

As the cut off value increases, the sensitivity reduces and the specificity increases. In the Credit rating problem specificity is more important to decide a cut off value. As it is more important to correctly predict the defaulters, i.e. those people who are more likely to be unable to repay the credit. Hence, the value of False positive which is predicting the Bad credit rating as a good Credit rating should be very low. This will help in reducing losses incurred due to bad loans.

Since the costs of a false positive (incorrectly saying that an applicant has a good credit rating) outweigh the benefits of a true positive (correctly saying that an applicant is a good credit risk) by a factor of 5, the PPR(positive prediction value) will be lower. The impact of a false positive has a higher weightage than the true positive, we must use the PPR to see the fitness of our model.

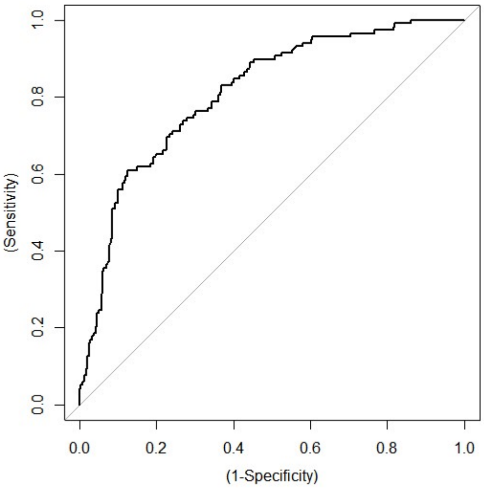
part e) ROC CURVE

The ROC curve for Full model i.e. part b)



Area under the curve(AUC): 0.7586

The ROC curve for Final model i.e. part d)



Area under the curve (AUC): 0.8089

The above graphs are the ROC curve which includes in one graph the performance of the model for all possible cut-off values. ROC stands for “Receiver Operating Characteristic”. The ROC curve as shown considers systematically all cut-off values from 0 to 100%. For each cut-off value it then measures the number of Goods below the cut-off and the number of Bads below the cut-off. It then plots these two numbers as x- and y-coordinates. A perfect model would show a ROC curve that consists of two straight lines: From (0,0) to (0,1) and from (0,1) to (1,1), i.e. very steep. A model with no predictive power would have a ROC curve that follows the diagonal, since that would imply that for every cut-off value we find an equal number of goods and bads, i.e. there is a perfect overlap in the two histograms.

From the above graphs we can see the Predictive power is higher for the part d) i.e. the reduced model and the same is evident from the higher AUC (Area under the curve value).