Question 1

Base on the possible values, get the distinct value for each variable to check if there is any value that are erroneous then check the amount of wrong value to decide on whether to remove give them a valid value.

With identified variables with invalid value “Years.with.present.employer” , “Personal” & “Credit.asked.for..X.100.”.

With the amount of invalid value for these columns only >0.1% the invalid records will be remove as it will not have significant impact on the construction of the prediction model.

Question 2

Variable “Personal” was split into “Gender” and “MaritalStatus” column, for “MaritalStatus”; “Single” was grouped together and the rest was grouped there, it will be better if we can perform smaller grouping, but due the structure of the given data we are not able to do as such.

Variable “Purpose of load” was transform into smaller group by identifying their underlining purpose. such as grouping Car (New), Car (Old) &Vacation as one, as they can be categorizing under luxury. Thus this variable was binned into 5 categories.

Variable “Credit.offered” has reassigned the Rejected from ‘2’ into 0, so to fit the logic of “True and False” better.

Question 3



As the graph above has shown, there is a very clear grouping between the Credit Offered and Rejected group. As there space between these group is significant

Question 4

Through random sample, split the data into 80-20, 80% for training and 20% for testing.

As this problem is a classification problem with two possible output; “true” or “fasle”. Thus the model shall be logistic regression model. The initial model was trained with all the variables execept the “Credit.offered.” as it is the outcome that we want to predict.

And after trained the model, we look at each variable at their P-value, and found out that “Status.of.account” & “saving.account” have the lowest p-value which indicate that they have high association with the the credit will be offered or not.

The model is reviewe with ANOVA, through ANOVA we check on the contribution of each variable to the model, and found that “Status.of.account”, “Loan.Duration”,” Credit.history” & “Savings.accoun” contribute the most to the model.

Question 5

Using StepAIC stepwise algorithm to select the variables.

With the final predictor for the model, we build a new model with the final predictors.

And then evaluate if the model have multicollinearity by looking at the VIF value. All the variables greater than 1 for the VIF value, but the highest is around 1.2, which indicates there is a very very small coolinearity for these variable but it is too small the have any impact for the model.’

Question 6

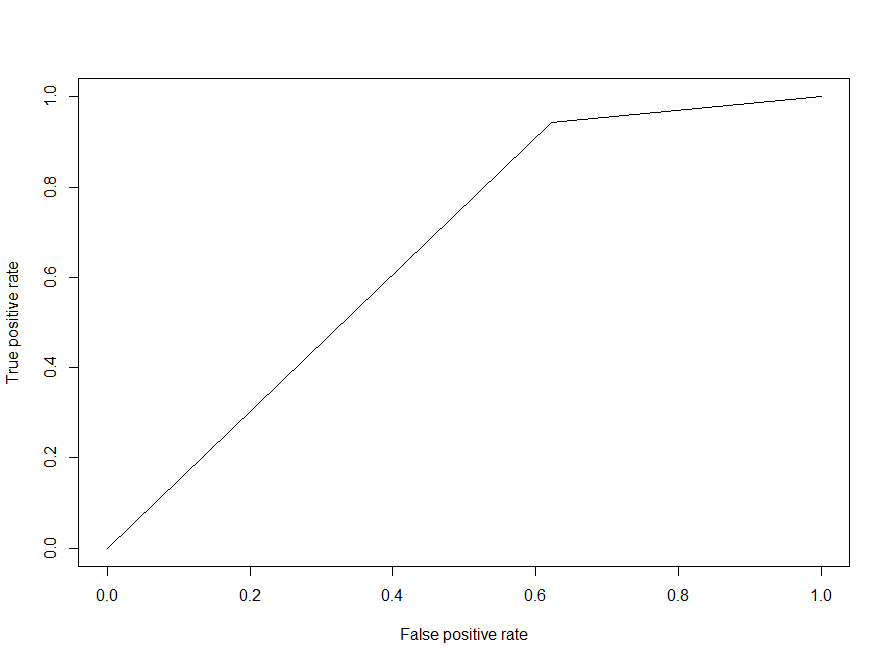
This model is built base on 11 predictors and all the predictor have certain level of association with the dependent variable. “Status.of.account” & “saving.account” predictor have the highest significant association with the dependent variable.

Question 7

The model has accuracy of ~77%, the accuracy may not be the actual accuracy as it is tested using the data from the same set of the original dataset, and the result can vary based on how the data is split. Thus the actual accuracy can only be aquire when the model went to production.

Question 8

The figure below is the ROC curve



The model has area under the curve (AUC) of ~0.66, it shows that prediction with the model is better than without one, the AUC value I not very ideal as having of value of 1 is the best, and there is a very big gap between the ideal of value 1 as compared to the model built.