DSCT Assignment 3 Report

By Zheng Yi Tao

Question 1

Assess the data and ensure that there are no defects in it.

Referring to code, I have checked if there were any missing records, if there are any non-numeric values (because this dataset has all input as numbers). Also, I have changed the column names to something more readable.

Question 2

Determine if any categorical variables will need to be transformed, combined or split up.

Things that I have done:

1) Convert all qualitative variables from int to factor.
2) Changed the target variable Credit_Offered to 0 and 1. Where 0 is Reject, and 1 is Accept.
3) Changed the levels of all the categorical variables to more meaningful labels.
4) Decreased the number of categories for "Purpose of Loan" to just 4 instead of 10:
    a. Car
    b. Household
    c. Education
    d. Business
5) Personal variable has been split into Personal AND Gender
    a. Personal contains the marital status, which are then been converted into dichotomous variables using cSplit_e function.
6) Assets Variable has been split into dichotomous variable too.
    a. Reason is to reduce each category to just 1 input instead of multiple.
7) Noticed that there are some outliers in the records with data that has no meaning. They are all converted into NAs and the rows are removed, since there are only 3 rows affected.

Question 3

Determine if there are any natural grouping of the credit applications using clustering techniques (k-means / MDS).

Choice of clustering algorithm

First of all, there is a problem with using k-means with this dataset. K-means clusters data based on the nearest mean between observations using Euclidean distance to calculate distances between continuous variables. However, with a dataset that has mixed data types (continuous and categorical), Euclidean distance will not be able to produce sensible result. This is because categorical variables do not provide meaningful distance between itself.

For example, we have a categorical variable that have a binary option of either Female or Male. But to decide the distance between being a Male or a Female has no meaning as compared to the distance between a low and high temperature where we can decide how much cooler or hotter between data.

R programming has provided Cluster library that contains a distance metric called **Gower distance** used through the **daisy** function which works for both continuous and categorical variables. Gower distance works for differently for each variable type by scaling the data to fall between 0 and 1. It then use a linear combination of weights to calculate the final distance matrix.

The metrics used for each datatype are different:

- quantitative (interval): range-normalized Manhattan distance
- ordinal: variable is first ranked, then Manhattan distance is used with a special adjustment for ties
- nominal: variables of *k* categories are first converted into *k* binary columns and then the Dice coefficient is used

(Source: https://www.r-bloggers.com/clustering-mixed-data-types-in-r/)

Figure 1 show the result of clustering using Gower Distance where the number of clusters:

Figure 1: Clusters where k = 5

Any number of clusters that are more than 5 (See Figure 2) will remove any distinct clusters between the data making the cluster meaningless. Thus, there are 5 natural grouping seen from this dataset even though there are some points overlapping with other clusters.

Figure 2: Clusters where k = 7

Question 4

Formulate and construct a suitable regression model.

1) A training and testing set were created with a 80:20 ratio where 80% of the data goes into the training set and 20% of the data goes into the test set.
2) A model was created using R's glm function where Credit_Offered is the target variable while all the other variables are the predictors.
3) The model has decided that the 3 most significant variables where the p-value is less than 0.01 are "Account_Status", "Loan_Duration" and "Credit_History".
4) A Chi-square test was done also using anova function to determine the relative importance of each predictor. Result was consistent that the above 3 variables are the most significant.

Question 5

Step through your model and determine if there is any multicollinearity in the predictors and adjust your model accordingly.

StepAIC function is used and a new model with the adjusted variables are created. (Shown in the code)

```
> vif(modelRefined)
                    GVIF Df GVIF^(1/(2*Df))
Loan_Duration   1.032799  1        1.016267
Account_Status  1.048066  3        1.007855
Credit_History  1.057611  4        1.007026
```

Figure 3: Check for Multicollinearity using vif function

Figure 3 shows the result of vif function, where generally if the vif value is < 4.0, it has no multicollinearity relation.

Question 6

Explain the best regression model that you have constructed.

	Odds Ratio	2.5 %	97.5 %
Loan_Duration	0.9677043	0.9544734	0.9808909
Account_Status-1	0.7927431	0.3284800	1.8603954
Account_Status100	1.1347462	0.4734475	2.6485821
Account_Status200	2.2145950	0.7832617	6.3676658
Account_StatusNone	5.0719525	2.0509730	12.5222402
Credit_HistoryPaid	1.1998162	0.4261271	3.4182886
Credit_HistoryPaying_Duly	2.7789392	1.2747234	6.2886832
Credit_HistoryDelay_before	2.5906196	1.0369295	6.7052072
Credit_HistoryCritical	4.4133588	1.9302148	10.4466409

Figure 4: Odds ratio and 95% Confidence Interval

Using the coef and confint function, we can see the Odd Ratio of the 3 most significant variables. Beginning with Account_Status, it has shown that people with no existing checking account has up to 5.07 more odds of getting accepted with the loan. While people with their account status less than 0, which indicating that the account owns the bank money (because negative) will have the odds of only 0.7 of being accepted.

Looking at Credit_History, if it is a critical account, or that the credits exist in other bank, the odds of being accepted is 4.4, which is higher than if the account has paid back credits or have not taken credits at all which have the odds of 1.1. Lastly, while Loan_Duration is not as high odd ratio as the other two, it is more significant than other variables in the dataset. Loan duration has a 0.96 odds ratio of been accepted for the loan.

Thus, the 2 biggest factor that contributes to a person being accepted for the loan are those who have no checking account or people who have critical account / other credits existing in other bank.

Question 7

Assess the predictive ability of the model and comment on your assessment.

1) Using the predict function, the new model is being used to test against the test set where the target variable is removed from the test set.
2) After the predict function is completed, all the values that are above 0.5 are considered as 1 referring to “Accept” while the rest are 0 referring to “Reject”.
3) Within the caret package, the confusionMatrix function is used to determine the accuracy of the model. A confusion matrix plot is being generated (Seen in Figure 4):

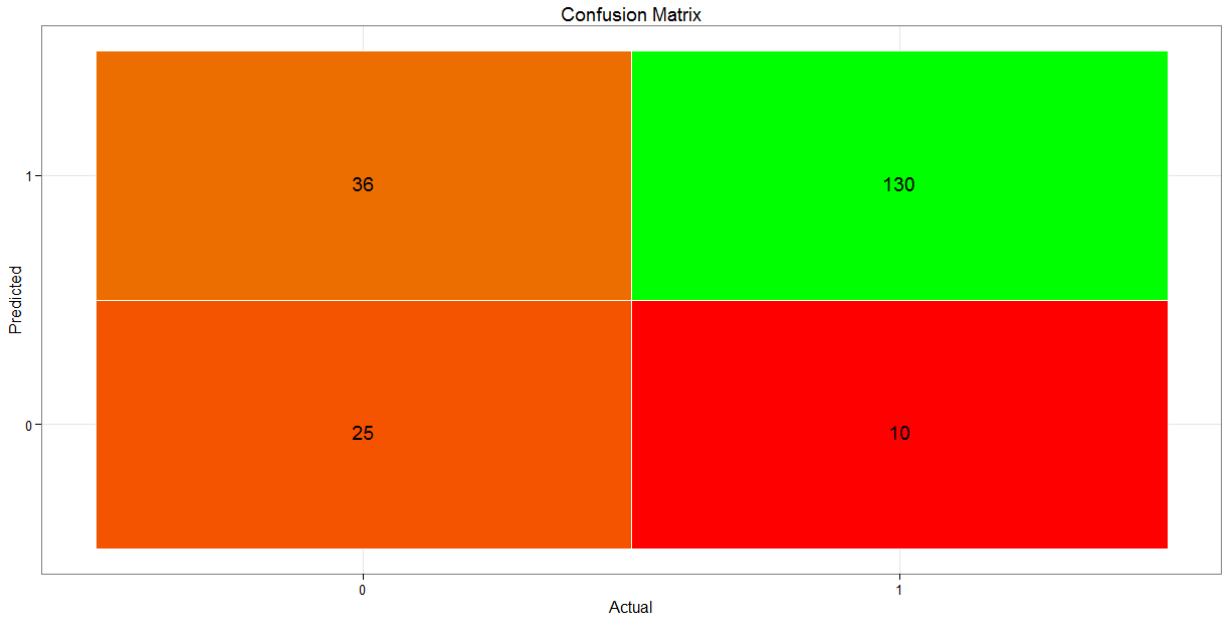

Figure 5: Confusion Matrix

```
               Accuracy : 0.7711
                 95% CI : (0.7068, 0.8273)
    No Information Rate : 0.6965
    P-Value [Acc > NIR] : 0.0115654

                  Kappa : 0.3847
 Mcnemar's Test P-Value : 0.0002278

            Sensitivity : 0.9286
            Specificity : 0.4098
         Pos Pred Value : 0.7831
         Neg Pred Value : 0.7143
             Prevalence : 0.6965
         Detection Rate : 0.6468
   Detection Prevalence : 0.8259
      Balanced Accuracy : 0.6692

       'Positive' Class : 1
```

Figure 6: Confusion Matrix calculations

Looking at the output of confusion matrix's calculation, we can see that the Accuracy is 77.1% with the confidence interval of 95% between 0.7 to 0.82 which is acceptable.

However, as pointed out by the codebook that it is worse to class a customer as good when they are bad, than it is to class a customer as bad when they are good – I have decided to look at the sensitivity instead. The reason is that sensitivity calculates how much True Positive was predicted among all the existing positives. Having sensitivity at 92.8%, it shows that most of the time the positive results will be indeed true positive and not false which is good for the bank.

Question 8

Construct a ROC curve and calculate the AUC to evaluate the regression model that you have built. Is the model you have constructed effective?

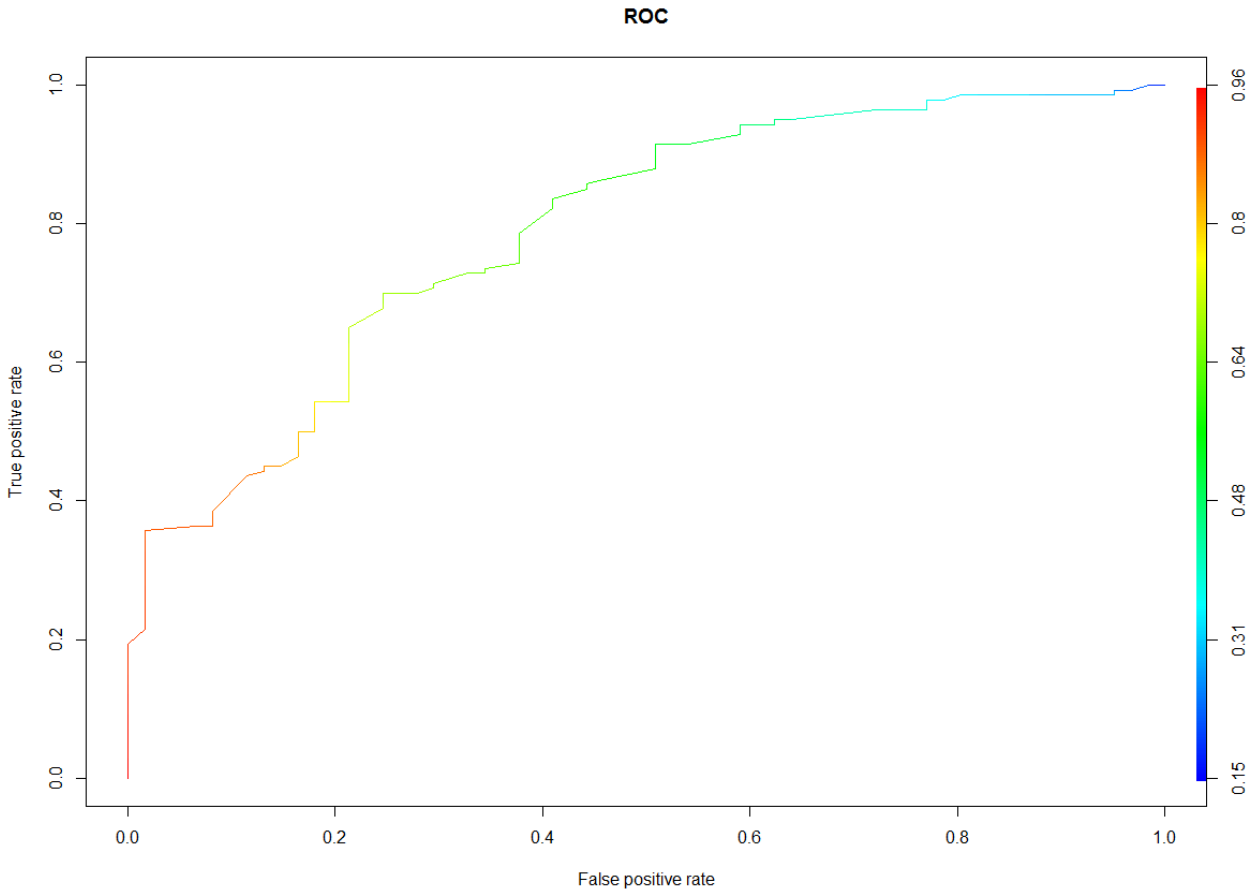

Figure 7: ROC Curve

The ROC Curve has been generated with the ROCR library as seen in Figure 6. The AUC values were calculated to be 78.98% where 100% is perfect, and it is nearer to 100% than 50%. Which shows that the model that was constructed is effective.