|  |
| --- |
| Data Science Competency Training  Core Level Topic AssignmenT  Regression and model evaluation |

Assignment Submission

Name: Tan Min Sheng Vincent

To minimize loss from the bank’s perspective, the bank needs a decision rule regarding who to give approval of the loan and who not to. An applicant’s demographic and socio-economic profiles are considered by experienced credit analysts before a decision is taken regarding his/her loan application.

We would like to build a regression model which will be able to provide bank managers guidance for deciding whether to approve a loan or not to a prospective applicant based on his/her profiles.

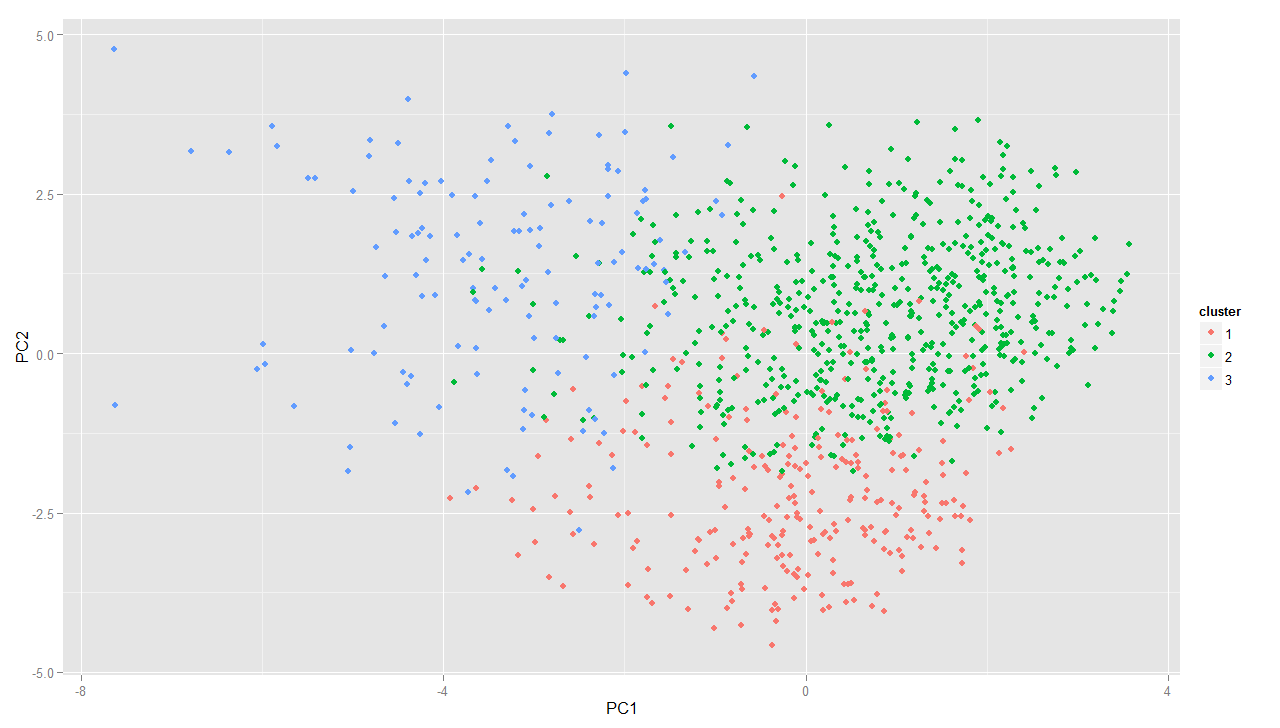
1) Assess the data and ensure that there are no defects in it.

* There are no missing values
* There are defects in the data:
  + Attribute 5: Credit asking for $’0’
  + Attribute 7: Extra value without category, ‘9’.
  + Attribute 9: Extra value without category, ‘7’ & No values from category 5, female, single
  + Attribute 13: Abnormally large outlier for age, ‘111’, 36 years from the closest value
  + Attribute 14: Extra value without category, ‘4’.
* Records with deflective entries are removed from the dataset as it’s not a lot of rows (4 in 1005 rows). So, removing them will not affect the dataset.

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

|  |  |  |
| --- | --- | --- |
| Attribute Name | Transformation | Result |
| Status.of.account | Break all categories in attribute into true false pairs |  |
| Credit.history | Break all categories in attribute into true false pairs |  |
| purpose.of.loan | Break all categories in attribute into true false pairs |  |
| Savings.account | Break all categories in attribute into true false pairs |  |
| Years.with.present.employer | Break all categories in attribute into true false pairs |  |
| Personal | * Combine Married, Separated, Divorced, Widowed into “Not Single” * Split status and sex into different columns | Status   * 1: Single * 0: Not Single   Sex   * 1: Male * 0: Female |
| Assets | Break all categories in attribute into true false pairs |  |
| Job.Classification | Break all categories in attribute into true false pairs |  |

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



The data can be segregated into 3 clusters using the k-means clustering technique which I believe might be caused by Age, Credit asked for and Loan duration.

4) Formulate and construct a suitable regression model.

Refer to R script

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

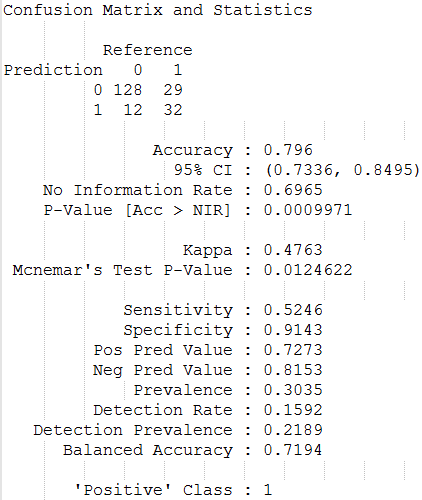
Refer to R script

6) Explain the best regression model that you have constructed.

As the aim of the regression model is to “To minimize loss from the bank’s perspective, the bank needs a decision rule regarding who to give approval of the loan and who not to”, the best regression model would be one that has a higher specificity.  
The model that have the highest specificity is on that only used Loan.Duration, Status.of.account and Savings.account as variables to predict whether to approve the loan.

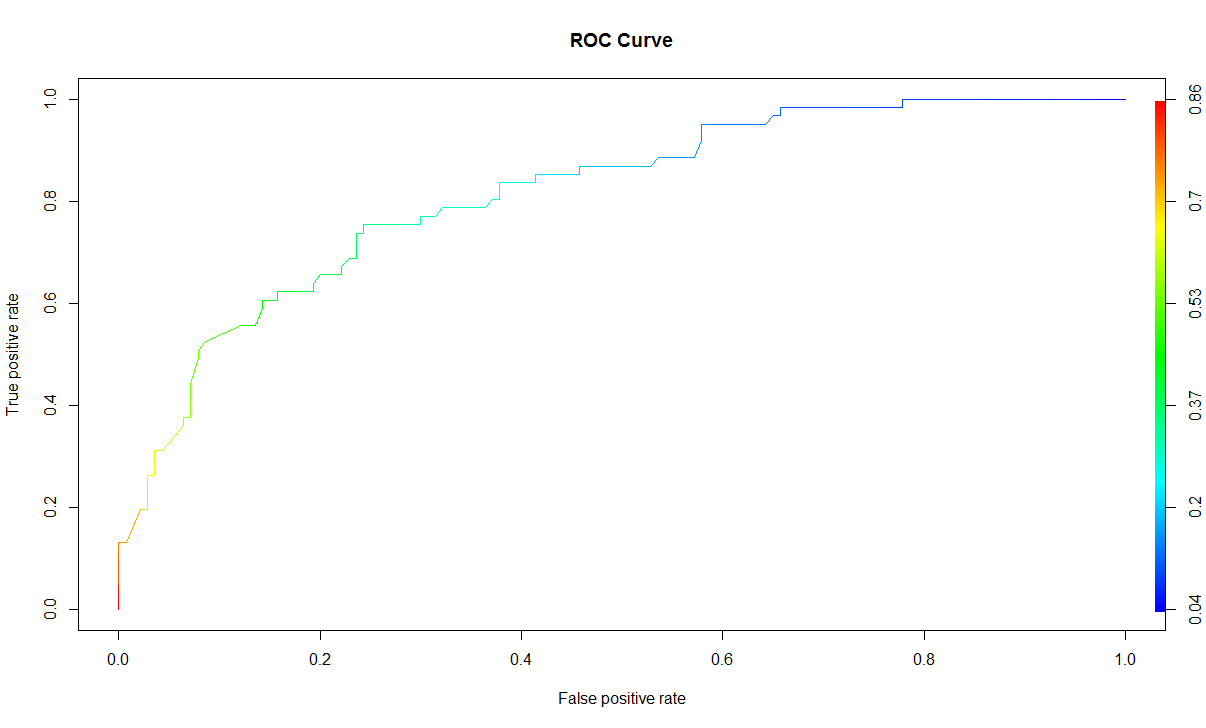
7) Assess the predictive ability of the model and comment on your assessment.

The above described model yields the confusion matrix as shown below:



With an accuracy of 0.796, I would say that the model is the most accurate of the three. Topped with the business goal stated as “minimizing loss”, the specificity of the model of 0.9143 would show that the model has indeed achieved its goal.

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?



AUC = 0.8150468

With an auc score of 0.8150468, the model can be said to be good at predicting whether to grant loans to individuals.