**Module 4: Data Mining Project Intermediate Analytics**

**Author’s Name:** Group A( Shivani Sharma and Rupali Somani)

**Course Number:** ALY6015

**Course Title:** Intermediate Analytics

**Academic Term:** Fall 2019 CPS Analytics

**Instructor’s Name:** Joe Manseau

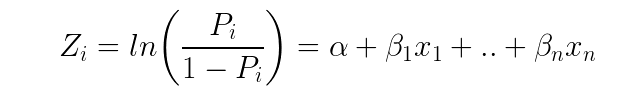
**Assignment Completion Date:** 12-03-2019



**Introduction**

This project is based on one of the application area of data mining i.e Logistic regression. If linear regression is used for predicting continuous variables. Logistic regression is used for binary classification and it is always based on assumption that logit transformation of the resulting variable is having a linear relationship wrt to predictor variables. It is a classic prediction modelling technique for binary categorical variables. There are various practical real time application areas of this model like for detecting spams, frauds in credit card, frauds in banking etc.

The concept behind this model is we need to take the odds of the events using following formula:



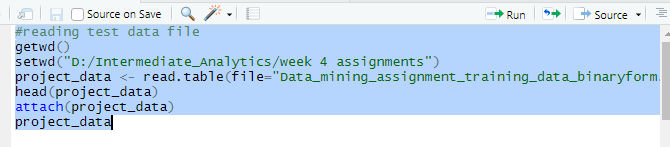
And P always lies between 0 and 1.

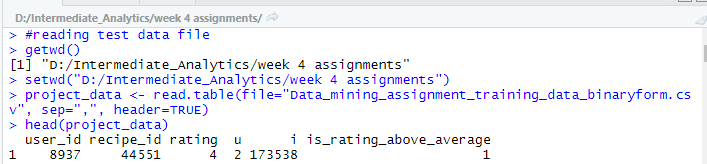
I will explain the step by step approach for this model in the analysis section using R programming language and at the end in conclusion will provide the summarized metrics of the model.

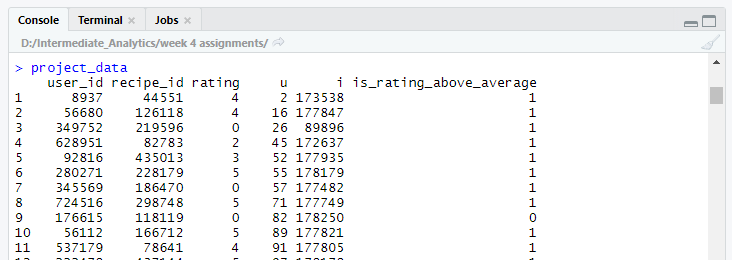
**Analysis**

I have picked up a data set for a real time scenario i.e a website called food.com. And it is showing up the records for various users and their reviews and ratings about the recipes. I have bifurcated the data and added up a column for is the rating above average? The problem statement that I wanted to solve here is that if the rating for a recipe is above average or below average. As based on this feedback the company will figure out which recipes needs improvement. In the new data after adding the column is rating above average or below average. The rating which is more than 3 is considered as above average and a value of 1 is assigned to that and the ratings which are below 3 are considered as below 3 and assigned a value 0. I have divided the data set into three groups i.e training ,test and validation. The division is 50 %,30% and 20 % respectively. Now I am going to start applying the model on training data set. Here is the step by step approach for this model:

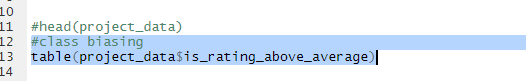
1. First of all we will import the csv file for processing. Here is the logic and output for that:

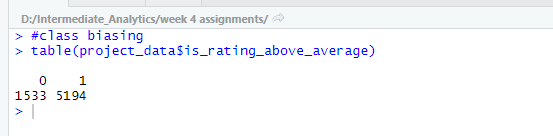




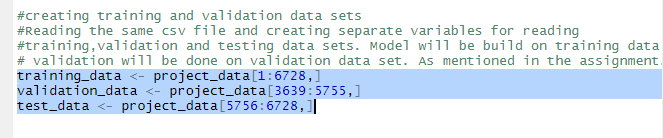


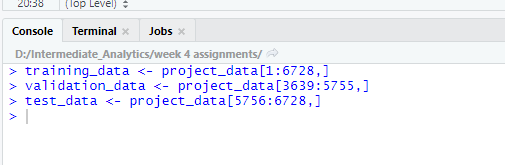
1. Now we have to identify the class biasing which can be achieved by verifying if a condition is observed with proportion of events is much smaller than the non-events and ideally it should be same. Following is the logic for verifying this:

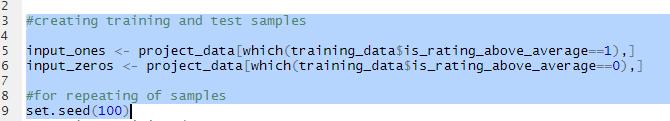


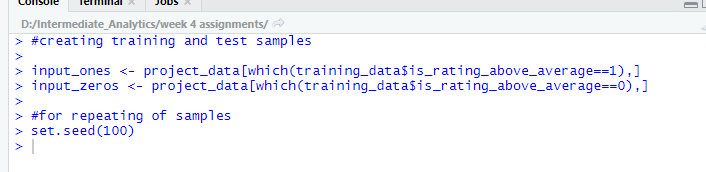


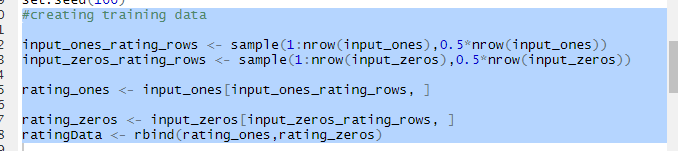
1. In this step we will divide the binary value column i.e is the rating above average or below average and then diving it into further classes i.e training, validation and test in the csv file of the data set. In further addition to this storing the data sets in different variables. Now we will use training data set for further calculations and create training and test samples. After that we will validate the accuracy of model using validation data set. Here is the logic:

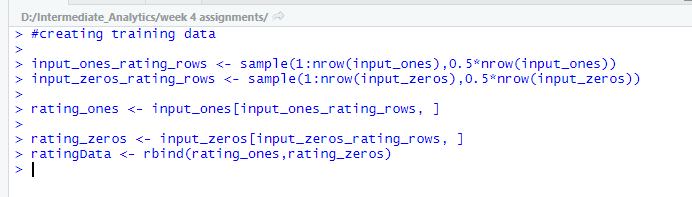


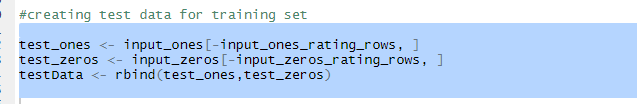


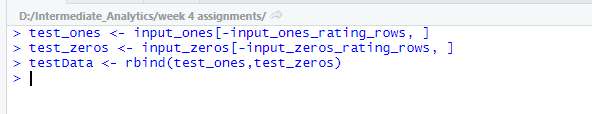




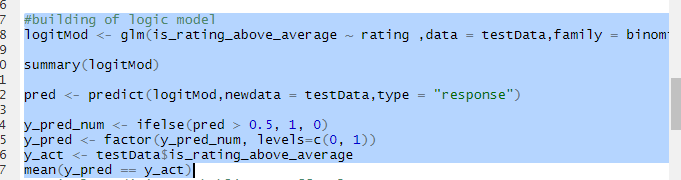


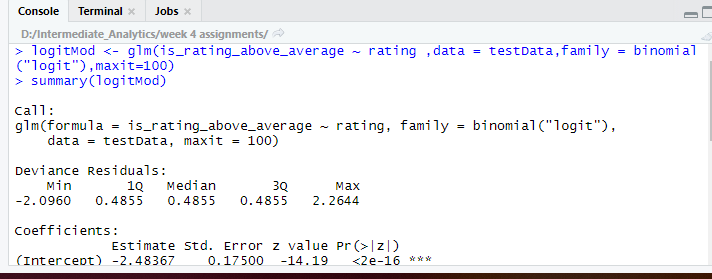


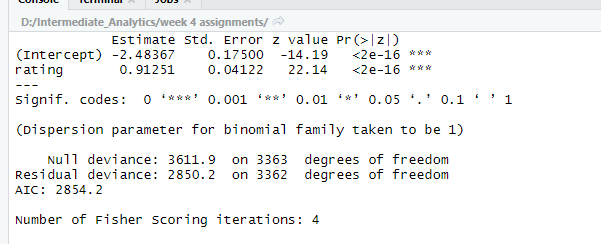


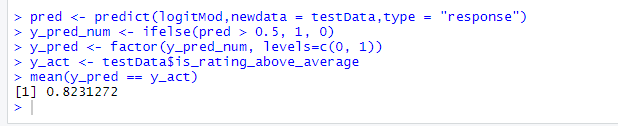


1. Now we will be applying the logit model and do the prediction for the training data set that we created.





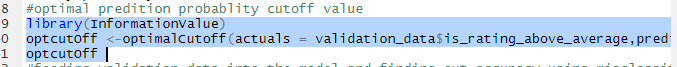


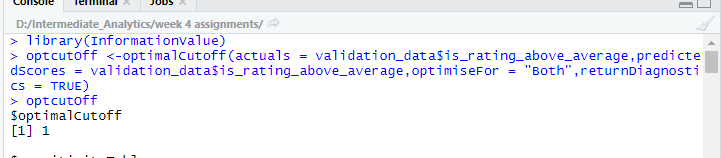


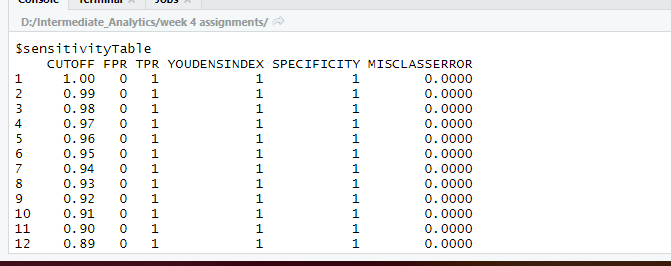
From the above data we can observe that the we have applied the formula on ratings based on the user id values which i.e which user provided what rating. The estimated values and standard deviation is very less and also the p value is coming out to be very small. And these observations are done for 4 iterations. Also the value of AIC is coming as 2854.2

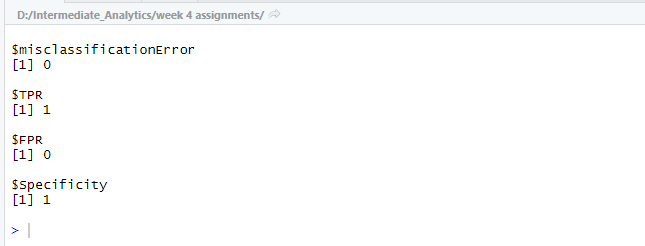
Also from the above results we can observe that the accuracy of the model is coming as 82.31% i.e proportion of y\_pred which is matching with y\_act i.e mean, which is good. The AIC value is coming as 2854.2.

Now I am going to do further analysis on the data and figure out the optimal prediction probability cut off value for the model applied. The default value for this cut off prediction is 1 or the ratio between 0’s and 1’s in our training data set. And in R language we have a function called optimal cut off in information value library to calculate the probability cut off value . Here is the logic for this:

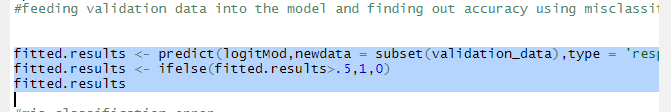


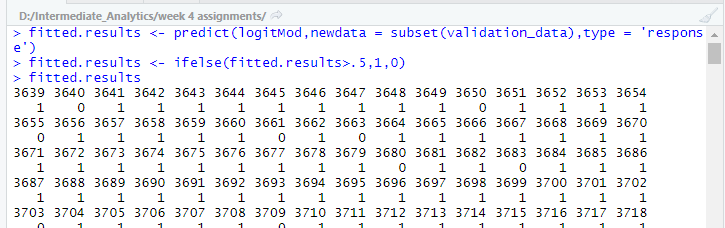




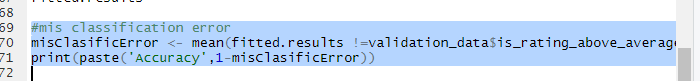


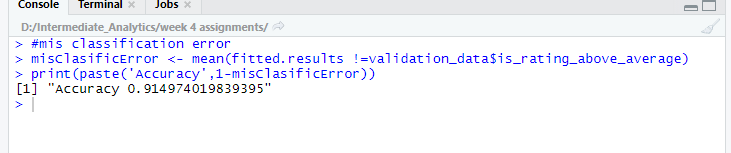
The optcutoff value is coming as 1.





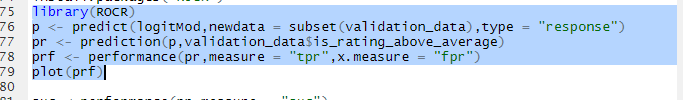
Now in order to check the accuracy of the model we will find out mis classification error by feeding validation data into the function.

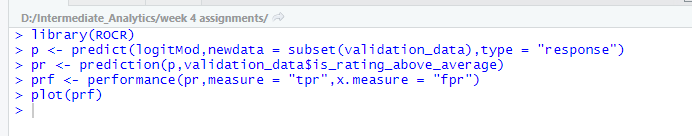


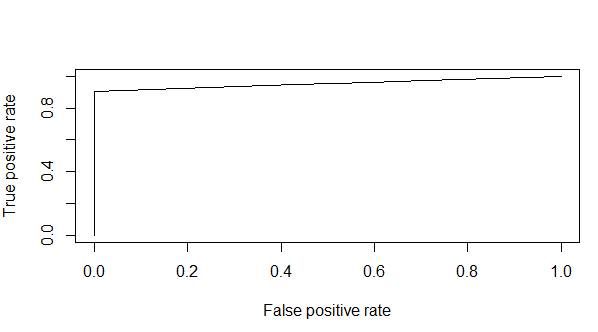


The accuracy is coming as 91% which is very good result.

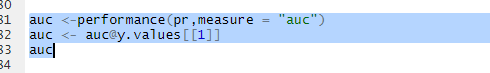
Now I want to plot ROC for this model. Receiver operating characteristics curve traces the percentage of true positives which are accurately predicted by our model. For a good model as the cut off value is lowered it should it should mark more of actual 1’s as positives and lessor of actual 0’s as 1’s. So ideally a good model curve of ROC should rise steeply and denotes that the TPR increases fast as the cut off decreases. More the area of the curve better is the prediction ability of the model. Here is the logic and output of the function:

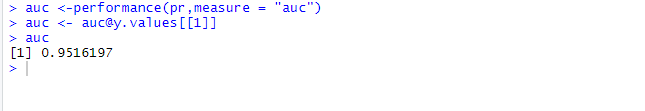




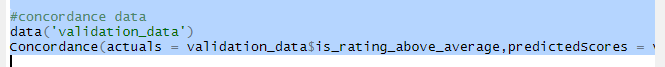


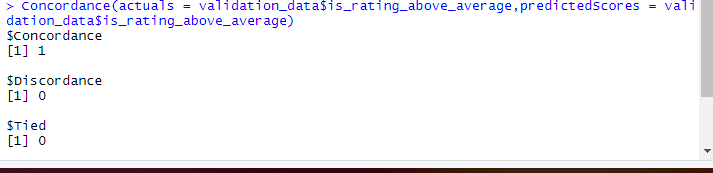
The model that I applied is having the ROC curve is more than 80% which is a pretty good result and the shape of the curve is as per the expected behavior and the curve is rising steeply as the cut off score is decreasing. Also the area of the curve is approximately .275 which is also a good coverage and hence better is the prediction. Now I am going to find the AUC value which measures the performance and the value is coming as 95% which is also a very good result for the fitting of our model.

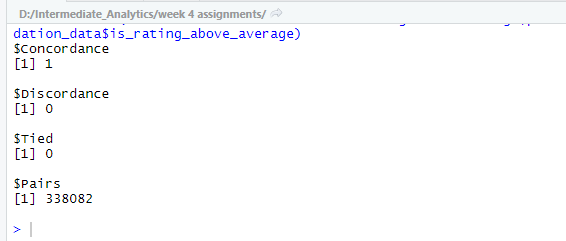




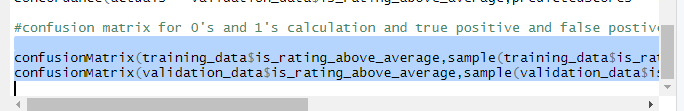
Now I am going to find out the concordance of the model which is the percentage of pairs whose score of actual positive values is greater than that of actual negatives. If a model is perfect then it is 100 %.

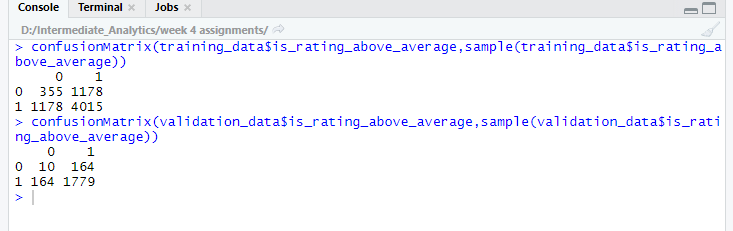






Now , I am going to calculate the confusion matrix for training data and validation data.





The rows in confusion matrix are the counting of predicted 0’s and 1’s values however columns are actuals. So, we have 355 out of 1533 instances predicted as below average which is 23 %. And 4015 are predicted as above average out of 5015 which is 77 %. In the training data set However there are cases of false positives and false negative as well i.e 1178 were predicted as above average and in actual they were below and 1178 were predicted as above average and in actual they were below average.

**Conclusion**

# At the end I would like to conclude that the model that I applied fits more than average as the ROC value is coming as 80% which is a good value and also the curve is steep which is having a larger area but it is becoming constant after a certain point. The concordance value is coming out as 1. The accuracy of the model is coming as 91 % which is stating that the model we applied is good. Optimal cut off value is more than default value which is also good i.e 1. And the results of confusion matrix are also significant which is fulfilling our objective of the problem statement i.e finding the fitness of model for all the ratings which is above average and validating the training data set results.

# References

1. Prabhakaran, Selva. “Eval(ez\_write\_tag([[728,90],'r\_statistics\_co-Box-3','ezslot\_4',109,'0','0']));Logistic Regression.” *Logistic Regression With R*, <http://r-statistics.co/Logistic-Regression-With-R.html>.
2. “HOME.” *IDRE Stats*, <https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/>.
3. Li, Shuyang. “Food.com Recipes and Interactions.” *Kaggle*, 8 Nov. 2019, <https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions>.
4. “Logistic Regression - A Complete Tutorial with Examples in R.” *Machine Learning Plus*, 15 Aug. 2019, <https://www.machinelearningplus.com/machine-learning/logistic-regression-tutorial-examples-r/>.
5. “Top 15 Evaluation Metrics for Machine Learning with Examples.” *Machine Learning Plus*, 18 Oct. 2018, <https://www.machinelearningplus.com/machine-learning/evaluation-metrics-classification-models-r/>.
6. “How to Perform a Logistic Regression in R.” *R*, 13 Sept. 2015, https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/.