Submission by: Pavana Lakshmi Venugopal
UBIT ID: 50464513

Part 1: Data Exploration and Preprocessing

Here Universal Banks wants to explore ways of converting its liability customers to personal loan customers. Also, the goal is to predict whether a customer accepted a personal loan that the bank offered in the last campaign or did not. This gives the bank a better idea on the customers who will take the personal loan for sure. The column Personal Loan represents the target variable. 480 customers accepted the personal loan and 637 did not accept the loan. NA values are present in this data set, and they are in the columns Experience and Income. In Education column Undergraduate category appears more frequently(389). Age and experience predictors are highly corelated. So, we must drop either one. In my case I have dropped Age column. And I did drop the column ID as I did not see any relevance to the prediction required. It is just random numbers which will not be useful. I flagged the categorical values. In this data set Education column needed to be flagged. I did scale the data. We need to normalize our predictors mainly because our model would not understand the difference between 200$ and 200Kgs, it just understands numbers, so we need get the values of the predictors on a similar scale. I used Z-score normalization technique here.

Part 2: k-NN

The training performance (F1 score) of the first model where k = 5 is 0.9427710843373495 and testing performance (F1 score) is 0.8960573476702508. I would say that there is an overfit issue. There is decrease in F1 score from Training to Testing hence I would say it is not performi ng well. Optimal value of k is 3 which is better than what we got for k=5. The f1 score is 0.90252 7.

Part 3: Logistic regression and model comparison

The training performance (F1 score) of the logistic regression model is 0.8828828828828829 and testing performance (F1 score) is 0.861111111111111. This model is not overfit. Percentage is slightly decreased from 0.88 to 0.86 which is not a huge difference. The same can be verified from classification summary as well. Between the logistic regression model and the k-NN model I would choose k-NN model because when we compare f1 score we already know the optimal value that is 90% with k=3 so I would choose k-NN model with the optimal value. When comparing with logistic regression, value is less that is around 86% which is lesser than 90% of k-NN value with k=3.