Assignment-2(Classification) Report

1. Yes,we need data pre-processing.The nominal columns needed to be scaled because there is huge variation in the scale of any two nominal features and categorical features needed to be encoded.The feature ‘duration’ was dropped due to data-leak and the feature ‘day’ and ‘month’ were dropped and combined into a single nominal feature ‘date’ which represents both their information.The data was scaled using MinMax scaler.The value -1 in ’ pdays’ was changed to a large value that is 9999 to make it more meaningful with respect to the model.

2. AUC was used as a metric to judge the performance of different model.The data was split into training and validation sets and different models with different max_depth were trained on the data.The model with highest AUC and least variance on both training and validation set was chosen to select the most optimal max_depth. Decision Tree with max_depth=3 gave the following result.

   training AUC=0.6693402947517585
   validation AUC=0.6560306201380679

   Which has the least variance among other values of max_depth.

3. Random Forest model is chosen best with respect to AUC score.No accuracy alone would not be a good metric because of the high imbalance of the output classes in the data.Predicting majority class itself would give a high accuracy,it give less information about how well the model classifies minority class.Random forest with max_depth=3 gave high AUC score of 0.7279 as compared to
   0.6674 of Decision Tree and 0.7079 of Naive Bayes.

4. It tells how well the model is capable of distinguishing between classes.A good model would have AUC in the range of 0.75 to 0.95 where as a poor model would AUC score in the range of 0.5.The AUC score is strongly affected by the misclassification of minority class which is our objective to reduce hence it is useful for the admissibility of a model in this problem.For example the Decision Tree model has an AUC score of 0.664 and it was poor in classifying minority class While Random Forest has an AUC score of 0.7279 and is better at classifying minority class.

5. The first model with max_depth=5 when trained on training data and tested on validation set gave the following results.

   validation AUC=0.7166980457404732
   training AUC=0.7643071301862234

   Where as the second model with max_depth=40 gave the following results:

validation AUC=0.6758030853656756
training AUC=0.9996587596739028
It is evident that there is a high variation in training and validation accuracy in the second model hence it is overfitting whereas in first model both are comparable.This is because increasing the depth of the trees would make the algorithm less general.

Results And Conclusions:

- **Decision Tree:**

  The training split was trained on DecsionTree was tested on testing data.The AUC score on both testing and training data was compared to determine an optimal value of max_depth as 5.The AUC of ROC curve is plotted in “AUC_DecisionTree.jpg”.

AUC score=0.6674

- **RandomForest:**

The training split was trained on Random Forest with 100 trees was tested on testing data.The AUC score on both testing and training data was compared to determine an optimal value of max_depth as 3.The AUC of ROC curve is plotted in “AUC_RandomForest.jpg”

AUC score=0.7279

- **Naive Bayes**

The training split was trained on Gaussian Naive Bayes was tested on testing data..The AUC of ROC curve is plotted in “AUC_NaiveBayes.jpg”.

AUC score=0.7079

- **Bagging with Decision Tree(max_depth=5)**

AUC score=0.7560 ,the AUC of ROC curve is plotted in “AUC_BaggingClassifier_3.jpg”.

- **Bagging with Decision Tree(max_depth=40)**

AUC score=0.991,the AUC of ROC curve is plotted in “AUC_BaggingClassifier_40.jpg”

Submitted by:
Rusafid Mirza Pottengad.
2015B4A70572G.