**MINI PROJECT 2**

**CLASSIFICATION MODELLING ON MARKETING CAMPAIGN DATASET FOR A BANK**

**Introduction**

For this mini project 2, we will be working on a bank marketing dataset containing information of several thousands of direct phone marketing, performed by a banking institution in Portugal to offer their new bank term deposit product. The original data has more than 45000 with 20 features. We will be working on a 20% sample of the dataset with 17 features.

We will conduct some following tasks:

* Data wrangling and Preparation
* Feature Scaling and Selection
* Implement classifier techniques
* Implement other pipelines required
* Make assumption and prediction based on analyzed data

1. Data Wrangling

We clean data as instructions:

* From the dataset, the “duration” columns do not imply the last duration contact hence dropping duration as suggested by the documentation

*df.drop('duration', axis=1, inplace=True)*

* Merge 2 columns “month” and “date” and convert it to numeric data type, we assume that the last contact date was nov 18th 2010, we will calculate number of dats since last contact performed until the last date assume above

*last\_date = "18 Nov, 2010"   
dt\_1 = dt.strptime(last\_date, "%d %b, %Y")   
df['last\_contact'] = df['date'].apply(lambda x: (dt\_1 - dt.strptime(x, "%d %b, %Y")).days)*

* Convert month first letter to upper case

*df['month'] = df['month'].apply(lambda x: x.capitalize())*

*df['date'] = df['day'].astype(str) + " " + df['month'] + ", 2010"*

* Wedrop unacessary columns

*df.drop(columns={'day', 'date', 'month'}, axis=1, inplace=True)*

* For categorical variables, converting all variable into dummy variables.
* Replace hyphen “-“ with underscore ”\_”

1. The best classifier model and features selected

We performed following features selection method:

* + LogisticsRegression
  + Linear SVM
  + SelectKBest with mutual\_info\_classif

Among those techniques, we would choose the method that give us the least number of features and have the highest accuracy value.

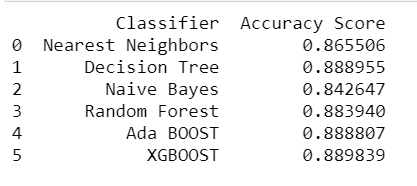
Based on the features selection method performed, all of the models result in a accuracy score of 0.89 when using the Logistic Regression Classifier. In this case, the best the model was the one that selected the least number of features, method 3 kbest selecting 10 features out of 30 with the high accuracy 0.88

After performing feature selection with 10 features selected, we use the function to get the columns selected and save it to another dataset name” feature1.csv”. Then, we do robust scaler for feature scaling to normalize range of variable.

The next step we train the new dataset with various classification models and select the best model. First of all we divided it into training and test data set with the ratio 75:25. Second, we use standard K-folds for cross validation, after that we use other machine learning pipeline to implement those classifier models below:

* + Naïve Bayes classifier
  + KNN
  + Decision tree classifier
  + Random forest classifier
  + Ada BOOST classifier
  + XGBOOST classifier

We have a result

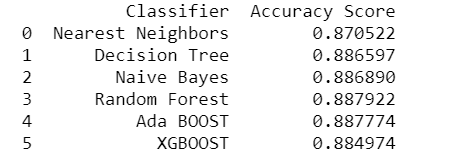


For the question 6, we do the same steps as the previous 5 one, we use less features number and implement new feature selection with random forest and get the accuracy score with 0.88899. Then I use Robust Scaler to scale the independent variable selected above and save into “feature2.csv” file.

Then we use feature2.csv file for new data frame to train again multiple classification model and select the best model which has high accuracy from train dataset. We do it again by splitting dataset into train/test set with ratio 75:25. We use k-fold to cross validation on training set only, after that use pipeline to implements below model:

* + Naïve Bayes classifier
  + KNN
  + Decision tree classifier
  + Random forest classifier
  + Ada BOOST classifier
  + XGBOOST classifier

The accuracy we gain from those models are:



XGBoost was the Best model in Step 5 with accuracy of 0.889839 and Decision Tree was the best model in step 6 0.888365 of accuracy

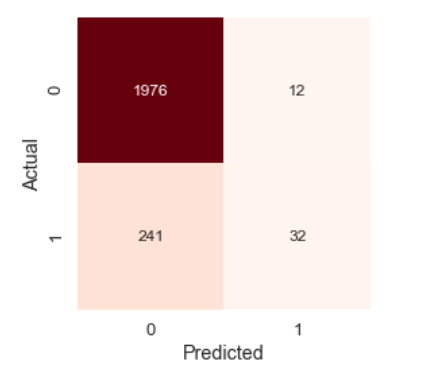
In step 5, 10 features were selected while in step 6, only 6 features were selected.

We can see that best models from each step are different classifiers, and that the accuracy result of best model in step 6 was slightly less than better model in step 6 even with less number of features.

1. Analysis of the chosen model performance

We would analyze the classifier model performance by producing the following:

* + Accuracy
  + Confusion matrix
  + Classification report
  + ROC curve



The result is telling us that we have 1976 + 32 correct predictions and 12 + 241 incorrect predictions.

Those figures means actual number of clients reject new term deposit after being called by the bank (label=0) is 2108 people (1976+32) and decision tree classifier predict 1976 clients do not agree to join.

The actual number of clients who agree to join new term deposit (label=1) is 273(241+32) and decision tree classifier predict 32 clients agree to join.

To Understand better a binary classification models other then accuracy, is interesting to rely on metrics like:

* **Recall Score**: The ratio of positives labeled by the model to all positives, TP / (TP + FN)
* **Precision Score**: The ratio of correctly labeled positives by the model to all positives, TP / (TP + FP)
* **F1 Score**: Harmonic mean o both Recall and precision, 2 \* (Recall \* Precision)/ (Recall + Precision)
* **Support**: The number of occurrences of each class in y\_test

This way let's plot the Classification report that tells us all of this metrics

As you can see the model is better to predict class 0 than class 1, in this problem is more likely that most people will reject than sign into the new banking product and consequently the dataset samples unbalanced related to dependent variable, you can see at the chart below the proportion of each class

That's why, even the model has high accuracy, predicting class 0 very accurately, is not so good predicting class 1, that for the bank is more important for the business to predict the class 1 than 0. Some techniques can help solve this problem, like oversampling and undersampling, and will be discussed in Step 9 (Additional Step)

1. Discuss your result and their managerial implications to the banking institution
2. Comment on the results on the expected outcome of the client mentioned at step 8 at the Jupyter notebook requirement and suggest some managerial strategies for the bank institution.