Machine Learning Engineer Nanodegree

Predictive Analysis On Bank Marketing Campaign

M.Alfred Raju

February 25th 2019

1. Definition

Project Overview: Bank marketing is known for its nature of developing a unique brand image, which is treated as the capital reputation of the financial academy. It is very important for a bank to develop good relationship with valued customers accompanied by innovative ideas which can be used as measures to meet their requirements[1]. In banks, huge data records information about their customers .This data can be used to create and keep clear relationship and connection with the customers in order to target them individually for definite products or banking offers .Usually ,the selected customers are contacted directly through: personal contact ,telephone cellular, mail, and email or any other contacts to advertise the new product/service or give an offer, this kind of marketing is called direct marketing. The objective of direct marketing in retail banking is to attract new customers, to create a direct customer-bank communication to promote an offer or obtain customer information, and to strengthen a long-term relationship with the customer[2]. Although the bulk of banks' direct marketing spend is still devoted to acquiring new customers, banks are paying more attention to existing customers via direct mail and e-mail, according to a new report from Mintel Comperemedia[3]. During 2008, banks increased direct mail offers crossselling additional products and services to current clients by 57% over 2007. During the same period, acquisition direct mail rose 7%. From the literature, the direct marketing is becoming a very important application in data mining these days. The data mining has been used widely in direct marketing to identify prospective customers for new

products, by using purchasing data, a predictive model to measure that a customer is going to respond to the promotion or an offer [4]. Data mining has gained popularity for illustrative and predictive applications in banking processes. The data set used for this project is well known as bank marketing from the University of California at Irvine (UCI)[5].The source of the data is from: [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014 .

Problem Statement:

The customer's huge electronic data is the source for all bank marketing campaigns. The size of these data sources is impossible for a human analyst to come up with interesting information that will help in the decision-making process. Data mining models are completely helping in the performance of these campaigns.

- The main aim of my project is to find out which of the classifiers best suits for this dataset.

- The task is to find out the best classifier that would predict the best campaign for bank marketing.

- Considering f1score as the evaluation metric I would find the f1score of each model and select the best model with highest f1score.

- The experiment will be carried out using some classifiers such as:logistic regression,Decision Tree,adaptive boosting.

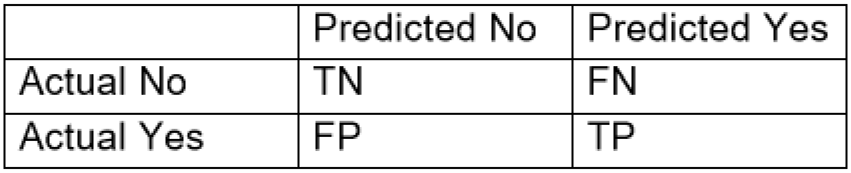
- This is to find out which of the classifiers would best suit the dataset in terms of classifying the pre-processed data,trained data,testing and make best prediction using the model obtained from the training process.

Metrics: The performance of each classification model is evaluated using two statistical measures; F1 score and accuracy score

The performance of a model cannot be assessed by considering only the accuracy, because there is a possibility for misleading. Therefore this experiment considers the F1 score along with the accuracy for evaluation.. This is because depending on the context , sometimes it is more important that an algorithm does not wrongly predict .

Thus, here we will use F-1 score as a performance metric, which is basically the weighted harmonic mean of precision and recall. Precision and Recall are defined as: It is using true positive (TP), true negative

(TN), false positive (FP) and false negative (FN). The percentage of Correct/Incorrect classification is the difference between the actual and predicted values of variables. True Positive (TP) is the number of correct predictions that an instance is true, or in other words; it is occurring when the positive prediction of the classifier coincided with a positive prediction of target attribute. True Negative (TN) is presenting a number of correct predictions that an instance is false, (i.e.) it occurs when both the classifier, and the target attribute suggests the absence of a positive prediction. The False Positive (FP) is the number of incorrect predictions that an instance is true. Finally, False Negative (FN) is the number of incorrect predictions that an instance is false. Table below shows the confusion matrix for a two-class classifier



	Predicted No	Predicted Yes
Actual No	TN	FN
Actual Yes	FP	TP

Precision=TP/ (TP+FP), Recall=TP/ (TP+FN), where

TP=True Positive

FP=False Positive

FN=False Negative

In the same vein, F-1 score is:

$$F1 = 2 \text{ x } \frac{Precision * Recall}{Precision + Recall}$$

We can use F-1 score as a metric that considers both precision and recall

Accuracy score:

Classification accuracy is defined as the ratio of the number of correctly classified cases and is equal to the sum of TP and TN divided by the total number of cases (TN + FN + TP + FP).

$$Accuracy = \frac{TP + TN}{TN + FN + TP + FP}$$

2. Analysis

Data Exploration: The data is associated with direct marketing campaigns of a Portuguese banking institution. phone calls were the main source of the marketing campaigns. Frequently, more than one contact to the same client was required, so as to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Input variables:

1.age (numeric)

2. job : type of job (categorical:'admin.','bluecollar','entrepreneur','housemaid','management','retired','selfemployed','services','student','technician','unemployed','unkn own')

3. marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

4. education (categorical:'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course ','university.degree','unknown')

5. default: has credit in default? (categorical: 'no','yes','unknown')

6. housing: has housing loan? (categorical: 'no','yes','unknown')

7. loan: has personal loan? (categorical: 'no','yes','unknown')

8. contact: contact communication type (categorical: 'cellular','telephone')

9. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10. day_of_week: last contact day of the week (categorical:'mon','tue','wed','thu','fri')

11. duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

Other attributes:

1. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

2. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

3. previous: number of contacts performed before this campaign and for this client (numeric)

4. poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent','success')

Social and economic context attributes

1. emp.var.rate: employment variation rate - quarterly indicator (numeric)
2. cons.price.idx: consumer price index - monthly indicator (numeric)

3. cons.conf.idx: consumer confidence index - monthly indicator (numeric)

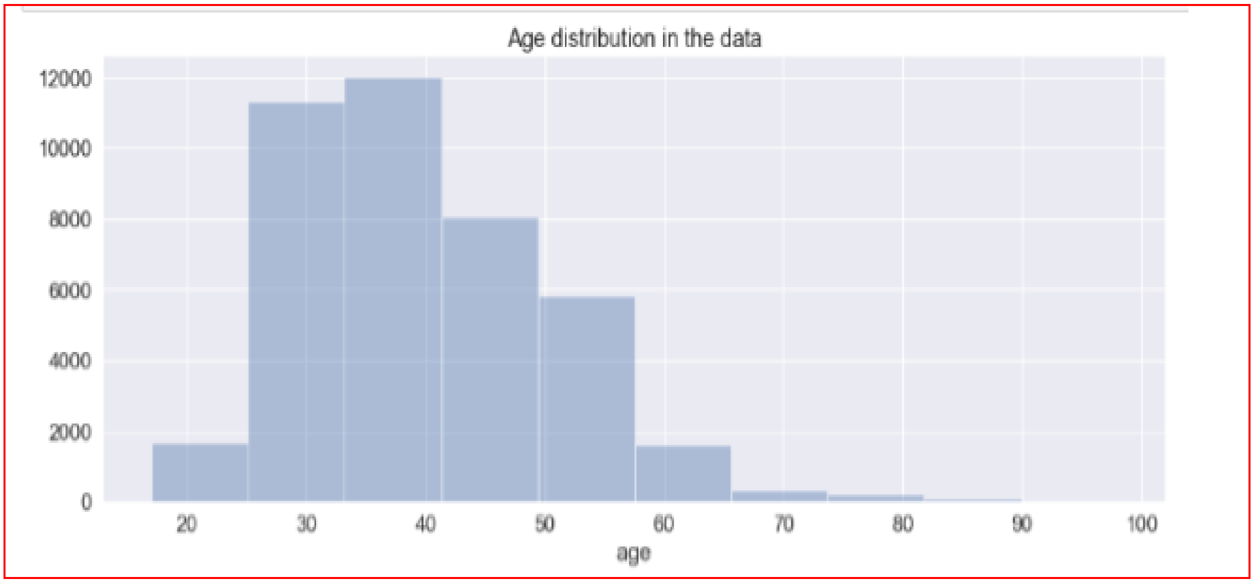
4. euribor3m: euribor 3 month rate - daily indicator (numeric)

5. nr.employed: number of employees - quarterly indicator (numeric)

Missing Attribute Values: There are several missing values in some categorical attributes, all coded with the "unknown" label. These missing values can be treated as a possible class label or using deletion or imputation techniques.

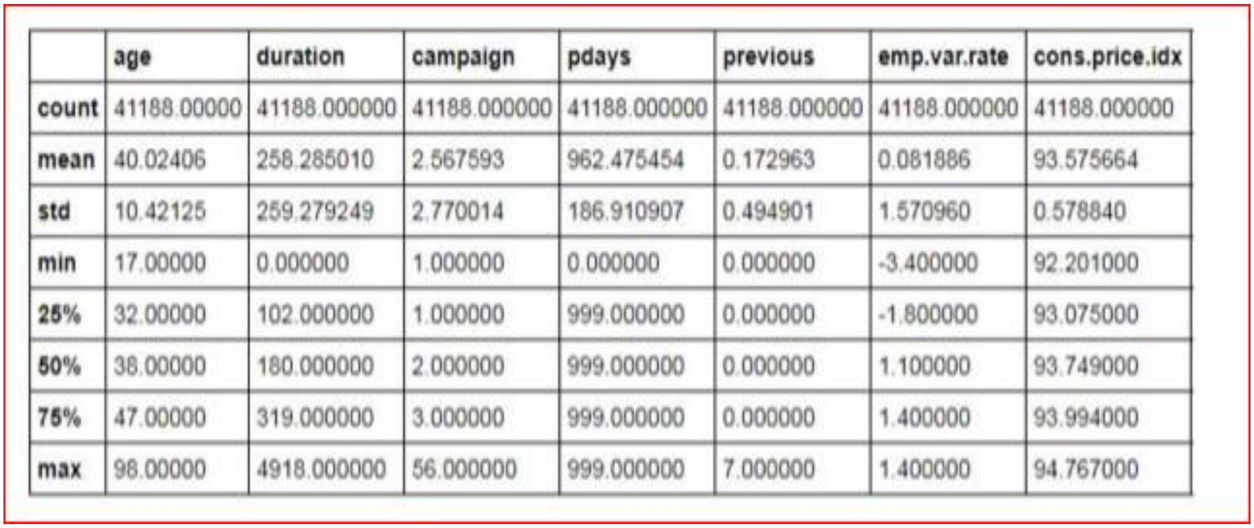
Output variable (desired target):

1. y - has the client subscribed a term deposit? (binary: 'yes' ,'no')



The above figure represents the age distribution of the data and from above we can understand that the data is lightly left skewed and this means that the most of the people are in the age group of 25-45. The dataset has 21 columns and 41188 rows, with 20 features, and one response variable. The number of subscribers is 4640 and the number of customers who did not subscribe is 36548. Based on this we can know that the response rate of customers is about 11.27%, this makes the

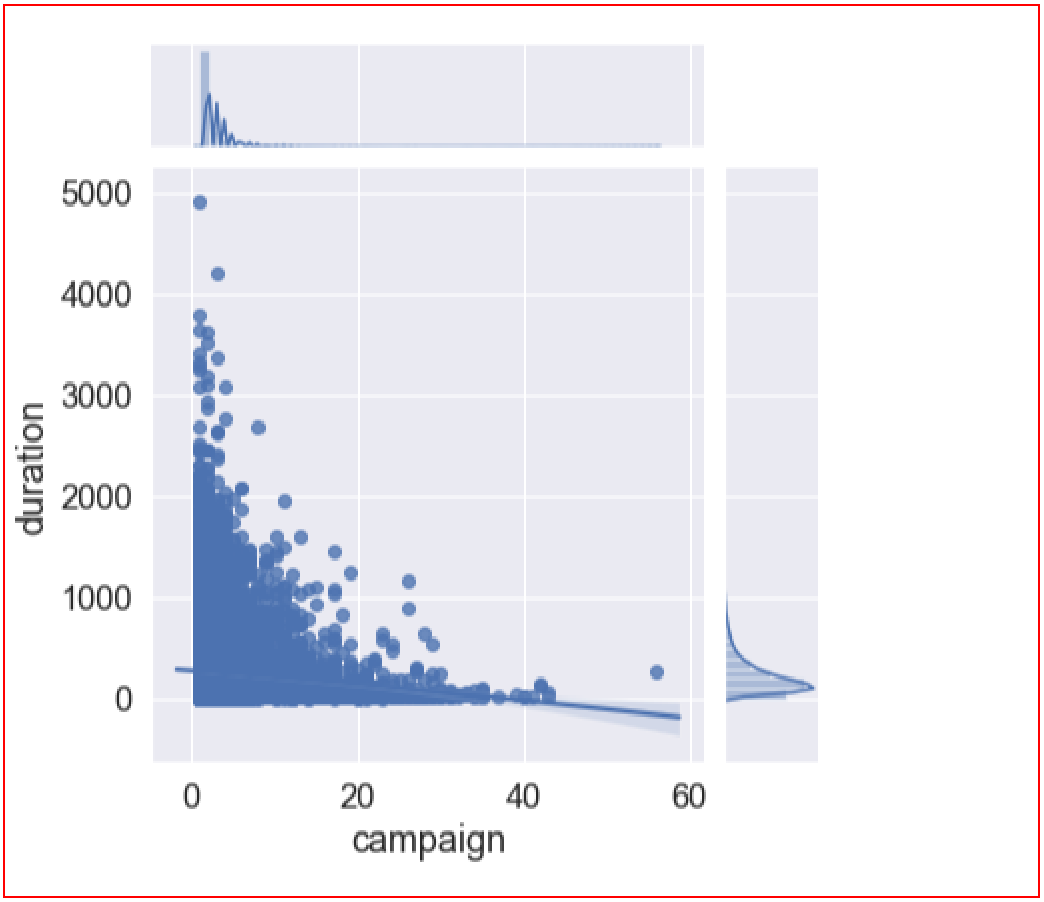
dataset very imbalanced.So,we are using f1 score as the evalution metrics



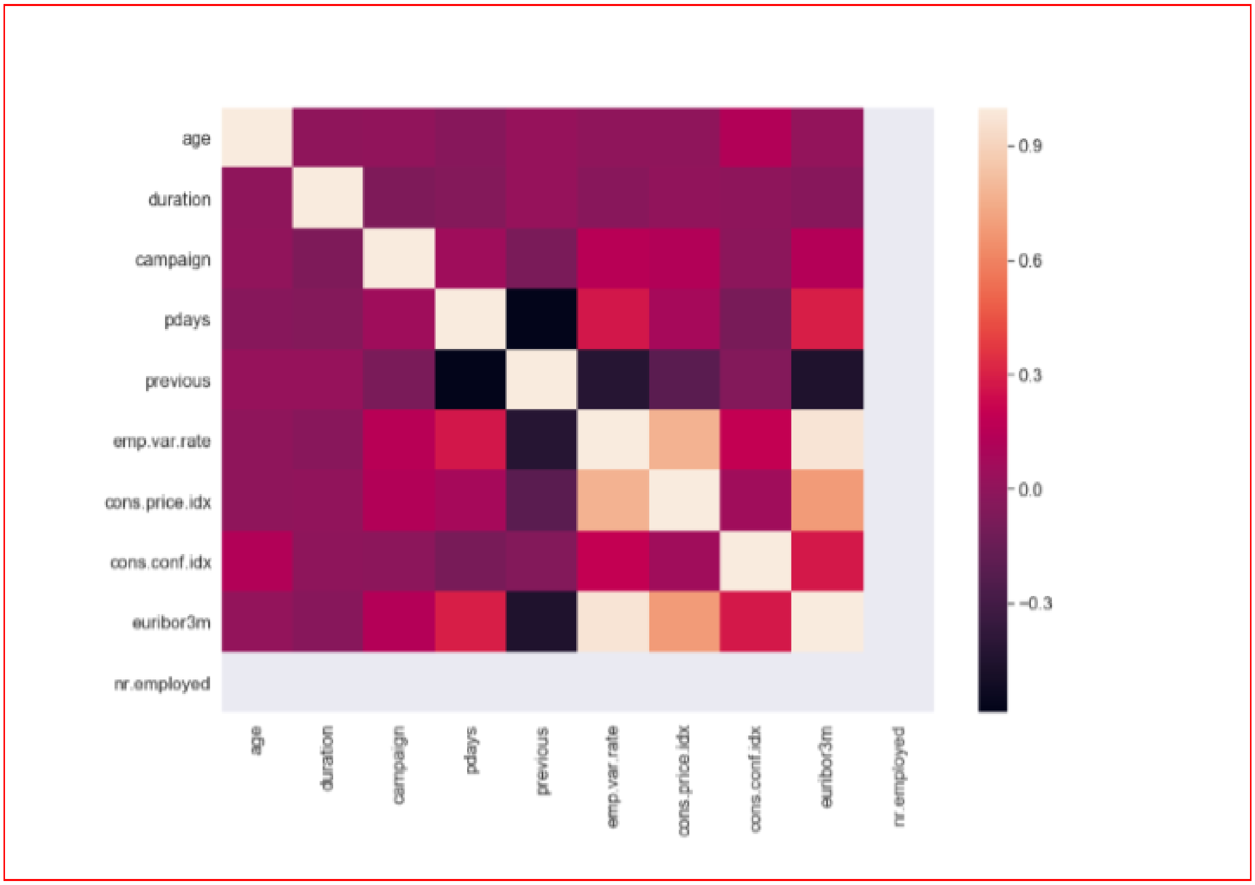
	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx
count	41188.00000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	40.02406	258.285010	2.567593	962.475454	0.172963	0.081886	93.575664
std	10.42125	259.279249	2.770014	186.910907	0.494901	1.570960	0.578840
min	17.00000	0.000000	1.000000	0.000000	0.000000	-3.400000	92.201000
25%	32.00000	102.000000	1.000000	999.000000	0.000000	-1.800000	93.075000
50%	38.00000	180.000000	2.000000	999.000000	0.000000	1.100000	93.749000
75%	47.00000	319.000000	3.000000	999.000000	0.000000	1.400000	93.994000
max	98.00000	4918.000000	56.000000	999.000000	7.000000	1.400000	94.767000

From the above table we can observe that all the attributes have same count and ‘duration’ has the highest variance and it can be considered as the best predictor when compared with others.

Exploratory Visualization:



The above graph is a joint grid for duration and campaign. The joint grid is a grid for drawing a bivariate plot with marginal univariate plots.



We use heatmap to visualize correlation of features of the data. There may be complex and unknown relationships between the variables in your dataset.It is important to discover and quantify the degree to which variables in your dataset are dependent upon each other. This knowledge can help you better prepare your data to meet the expectations of machine learning algorithms.So it is important to identify the correlation between features.But however from the above observation we can conclude that there is no much strong correlation.

Algorithms and Techniques: From my understanding,I observed that the given dataset is a typical supervised learning problem and the following algorithms can be applied to predict which is the best campaign to contact customers to subscribe deposits.The one among these which makes the best prediction can be selected as the best fit for the problem and this can be done with the help of the f1 scores of these classifiers.

The algorithms which I used here are logistic regression,Decision tree and AdaBoost classifiers.Lets understand the concepts of these algorithms in detail: First let’s know about Logistic Regression,

Logistic Regression: There are many classification tasks done routinely by people. For example, classifying whether an email is a spam or not, classifying whether a tumour is malignant or benign, classifying whether a website is fraudulent or not, etc. These are typical examples where machine learning algorithms can make our lives a lot easier. A really simple, rudimental and useful algorithm for classification is the logistic regression algorithm.

Sigmoid Function (Logistic Function) Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be class variable, i.e 0-no, 1-yes. Therefore, we are squashing the output of the linear equation into a range of [0,1]. To squash the predicted value between 0 and 1, we use the sigmoid function.

$$z = \theta_0 + \theta_1 \cdot x_1 + \theta \cdot x_2 + \cdots \qquad g(x) = \frac{1}{1+e^{-x}}$$

Linear Equation and Sigmoid Function

$$h = g(z) = \frac{1}{1+e^{-z}}$$

Squashed output-h

It is the go-to method for binary classification problems (problems with two class values).So it can be used for this problem as this is also a binary classification problem . Random Forest: Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it’s simplicity and the fact that it can be used for both classification and regression tasks.

DecisionTree:

Classification trees are the classification counterparts to regression trees. They are both commonly referred to as "decision trees" or by the umbrella term "classification and regression trees (CART)."

The real world application is predicting which stocks to buy based on past performace

Strengths: As with regression, classification tree ensembles also perform very well in practice. They are robust to outliers, scalable, and able to naturally model non-linear decision boundaries thanks to their hierarchical structure.

Weaknesses: Unconstrained, individual trees are prone to overfitting, but this can be alleviated by ensemble methods.

The decision tree can handle both numerical and categorial data.It is a good candiate for our dataset and handle in the data which hasn't been normalized. **Adaptive Boosting**: Ada-boost classifier combines weak classifier algorithm to form strong classifier. A single algorithm may classify the objects poorly. But if we combine multiple classifiers with selection of training set at every iteration and assigning right amount of weight in final voting, we can have good f1 score for overall classifier. The metric we use to evaluate these models is f1 score.The accuracy score is described as a ratio of the number of correctly predicted instances in divided by the total number of instances in the dataset multiplied by 100 to give a percentage (e.g. 95% accurate).Here we use 10-folds cross validation which means we split the entire data set into 10 parts i.e., perform training on 9 and testing on 1 and repeat this for all the combinations of train-test splits.

Benchmark Model: This is the most important step as you compare your final model with this benchmark model to see if your model is better, same or worse than this. Below table highlights performances of various models that were tried with their f1score and accuracy. The logistic regression model with default parameters yields 50% f1 score on training data. So I will consider it as benchmark and try to beat the benchmark with hyperparameter tuning of the model selected as the best model(i.e., tuning the one with highest f1score among the three mentioned).

Algorithm	F1 score	Accuracy
Logistic Regression	50	90
Decision Tree	52	88
AdaBoost	49	91

3. Methodology Data Preprocessing:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.Here I will split the data into features and target/label columns to prepare the data and perform data cleaning also check for quality of given data.To check the goodness of the model I created,I will split the data into training and validation sets and check the accuracy of the best model.I would further split the training data into two of which 70% will be used to train models and 30% will be used to train our models and 30% will be used as a validation set. There are several non-numeric columns that need to be converted. Many of them are simply yes/no, e.g.housing. These can be reasonably converted into 1/0 (binary) values. Other columns, like profession and marital, have more than two values, and are known as categorical variables. The recommended way to handle such a column is to create as many columns as possible values (e.g. profession_admin, profession_blue-collar, etc.), and assign a 1 to one of them and 0 to all others. These generated columns are sometimes called dummy variables, and we will use the pandas.get_dummies() function to perform this transformation)

Several Data preprocessing steps like preprocessing feature columns, identifying feature and target columns,data cleaning and creating training and validation data splits were followed and can be referenced for details in attached jupyter notebook **Implementation:** The step by step procedure I followed is:

1)Input data

2)Explore dataset

3)Data Cleaning

4)Model Evaluation

5)Train Classifier

6)Review Metrics

7)Validate Model

8)Conclusion

While exploring the dataset, we first load the libraries and data then we peek at the training data and know the dimensions of the data. We will then observe the overview of responses and overall response rate and see the statistical summary. Now we visualize the data of exploratory analysis. Now we perform data preprocessing, here we preprocess feature columns, identify feature and target columns, perform data cleaning and then do training and validation data split. In this step of evaluating models, we first build the models and use 10-fold cross validation to estimate accuracy. Now based on the accuracies we will select the best model here I used three models Logistic Regression,Decision Tree and Adaptive Boosting. Of which Decision Tree is the one with highestf1score and then we tune this model and use it to make predictions on the validation set.
Refinements:

1)We perform data cleaning to remove unwanted or noisy data which helps us to improve the quality of the data and improve decision making process.

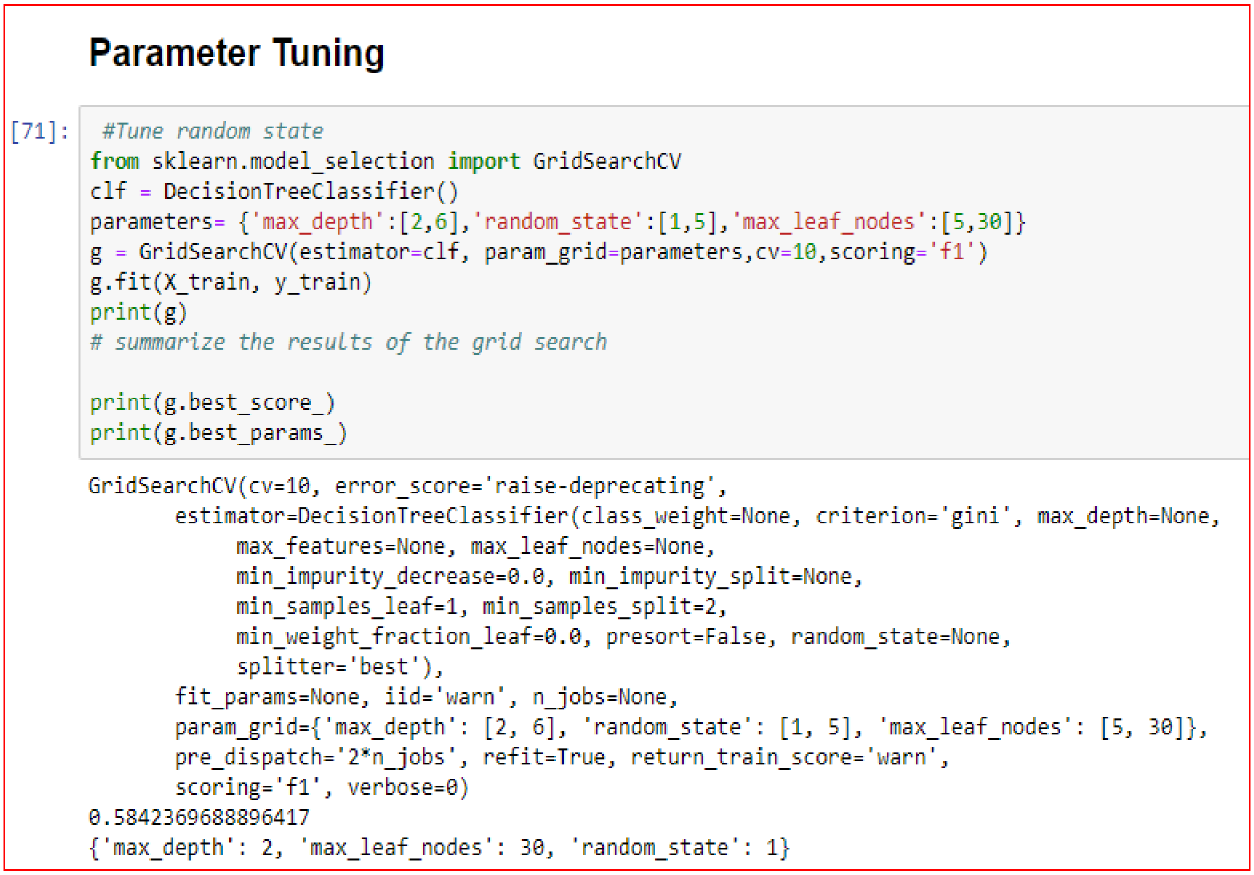
2) Here I split the data into features and target/label columns to prepare the data and perform data cleaning also check for quality of given data. To check the goodness of the model I created, I split the data into training and validation sets and check the accuracy of the best model.I would further split the training data into two of which 70% will be used to train models and 30% will be used to train our models and 30% will be used as a validation set.

3)We also used k-fold cross validation technique. This method attempts to maximize the use of available data for training and then testing a model. It is particularly useful for assessing model performance, as it provides a range of accuracy scores across different data sets. We performed hyper tuning of parameters of Decision Tree from scikitlearn library and the parameters tuned were shown in below table:

Parameter	Description	Values tested	Best value
Max_depth	The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.	[2,6]	2
Max_leaf_nodes	The minimum number of samples required to split an internal node: • If int, then consider *min_samples_split* as the minimum number. • If float, then *min_samples_split* is a fraction and *ceil(min_samples_split * n_samples)* are the minimum number of samples for each split.	[5,30]	30
Random_state	If int, random_state is the seed used by the random number generator; If RandomState instance, random_state is the random number generator; If None, the random number generator is the RandomState instance used by *np.random*.	[1,5]	1

4. Results

Model Evaluation and Validation: We applied and compared Logistic Regression vs the rest of the models. The metrics we used are calculated using sklearn wrapper so can be trusted for the model performance. Our end goal was to have a tuned model that could beat the benchmark model which it did successfully with a very good margin. So the solution described below is satisfactory to our initial expectations. We generated a final model with above list of tuned parameters. The output of this tuned model came just about 0.8% higher in f1score vs the untuned model. Code snippet of final model shown below:



Parameter Tuning

```
[71]:  #Tune random state
      from sklearn.model_selection import GridSearchCV
      clf = DecisionTreeClassifier()
      parameters= {'max_depth':[2,6],'random_state':[1,5],'max_leaf_nodes':[5,30]}
      g = GridSearchCV(estimator=clf, param_grid=parameters,cv=10,scoring='f1')
      g.fit(X_train, y_train)
      print(g)
      # summarize the results of the grid search

      print(g.best_score_)
      print(g.best_params_)
```

```
GridSearchCV(cv=10, error_score='raise-deprecating',
       estimator=DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best'),
       fit_params=None, iid='warn', n_jobs=None,
       param_grid={'max_depth': [2, 6], 'random_state': [1, 5], 'max_leaf_nodes': [5, 30]},
       pre_dispatch='2*n_jobs', refit=True, return_train_score='warn',
       scoring='f1', verbose=0)
0.5842369688896417
{'max_depth': 2, 'max_leaf_nodes': 30, 'random_state': 1}
```

The comparision between different models can be clearly depicted by the table below:

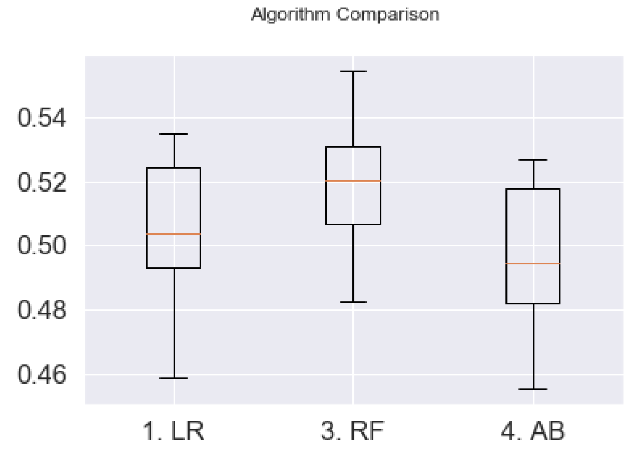
Algorithm	F1 score
Logistic Regression	50.5
Decision Tree	52.2
AdaBoost	49.6
Tuned Decision Tree	58.4

The recall rate is quite low but this can be improved by providing more weight to the positive labels `1`, this in turn decreases the overall accuracy but the model performance becomes robust to classify the outcome labels.

Justification:

I considered Logistic Regression as my benchmark model and attempted to cross its f1score .From the above results,it is clear that the best model I have chosen i.e., Decision tree was able to beat the f1score of the benchmark model(before and after tuning).So I am satisfied with the ability of the model I have chosen and tuned Decision Tree is the best one for the predictive analysis of bank marketing.

V. Conclusion Free-Form Visualization :



The comparision between three different models I used is visualized in the above graph.It explains that the top range is nearly 55% and the LR is the lowest performer with its range 53%.

Reflection:

In view of the data I analyzed and the domain I would like to give some suggestion to improve the duration:

1)The duration plays an important role among people saying yes i.e only interested people would have long duration and the banks should target the customers with noticeable duration and also on the customers who have positively reacted in the past campaigns. Preparing the data i.e., data cleaning and data processing is the time consuming part of this problem.And After the data is cleaned and prepared,the toughest part is to chose the best algorithm out of the three models I have chosen for this problem as all the three have the highest accuracy than my benchmark model.

After all the analysis I have chosen Decision Tree as the best model for this problem.

Improvements:

The improvements we could do are:

- To work with a subset of features that are high in variable importance pareto.
- To review metrics like log-loss to understand how quick the model is able to tune.
- To perform tuning to improve recall rate to improve overall prediction performance of the model.

References:

[1]https://www.tutorialspoint.com/bank_management/bank_management_marketing.htm [2]https://www.retailbanking-academy.org/media/uploads/NEWBRANDING-DOCS/modules/RBII/RB_II_New_201_Small.pdf

[3]http://www.mintel.com/comperemedia [4]Eniafe Festus Ayetiran, “A Data Mining-Based Response Model for Target Selection in Direct Marketing”, I.J.Information Technology and Computer Science, 2012, 1, 9-18. [5]https://archive.ics.uci.edu/ml/datasets/bank+marketing