Siddhartha Shankar
https://github.com/sid0712
https://www.linkedin.com/in/sid-business/
469-509-4807
siddhartha.shankar@icloud.com

Problem Statement:

Implementing supervised learning models using Logistic Regression, Scalar Vector Classifier (SVC with different kernels), Decision Tree Classifier and AdaBoost Classifier on the dataset to predict whether a client will subscribe a term deposit or not and come up with the best model based on model evaluation parameters of classifiers. The Dataset can be downloaded from UCI_ML_Bank_Marketing.

Description of Dataset:

- Dataset consists of 45211observations and 21 attributes. Description can be found by visiting the link.

- Dependent variable (y) is binary categorical variable, whether a client will subscribe a term deposit or not. The main features are client demographic data, information of current campaign and social and economic context attributes.

- There are few attributes which have missing values labelled as 'unknown'. For example, in "default" attribute, the total amount of "yes" response is very small i.e. three clients only. However, the number 'unknown' status is quite large which is 8,598 in total. In this case, it is not advisable to make imputation because of the rare population of "yes" response. Therefore, missing value i.e. 'unknown' label is treated as a new category to come up with the best classifier model.

- 11.27% of the 'y' variable is 1 i.e. 11.27% of the clients have subscribed to a term deposit. Because of this, ROC (Receiver Operating Characteristic) curve and AUC (Area Under Curve) is preferred as the performance measurement over the overall power of prediction (accuracy). Secondly, we would like to minimize false positive rate (FPR) as we don't want to predict that a client subscribed to term deposit if it hasn't actually; because, in that case, we may lose him in other future campaigns.

Reason for selecting the dataset

In the modern world, the increasing number of marketing campaigns over time has reduced their effects on the public. First, due to competition, positive response rate to mass campaigns are typically very low, according to a recent study, less than 1% of the customers are influenced by the campaigns. Second, direct marketing has drawbacks, such as causing negative attitude towards banks due to intrusion of privacy. Targeting potential segment of clients who are more prone to subscribing a term deposit will save time and money with relatively high success rate.

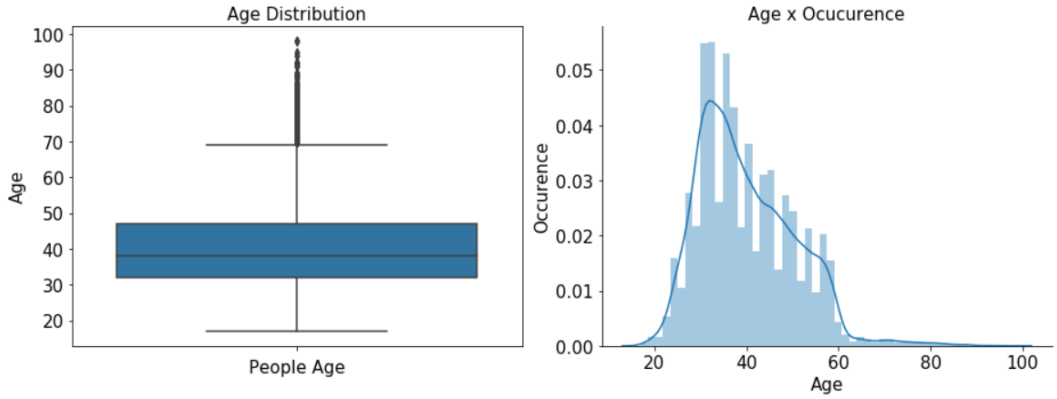
Our objective is to build a classifier to predict whether a client will subscribe a term deposit. If the classifier has high accuracy, it would help the banks in better management of resources by focusing on the potential customers "picked" by the classifier and improve their efficiency a lot. Additionally, classifier can provide significant attributes which are influential to customers' decision, so that a more efficient and precise strategy can be designed to improve the profits.

Siddhartha Shankar
https://github.com/sid0712
https://www.linkedin.com/in/sid-business/
469-509-4807
siddhartha.shankar@icloud.com

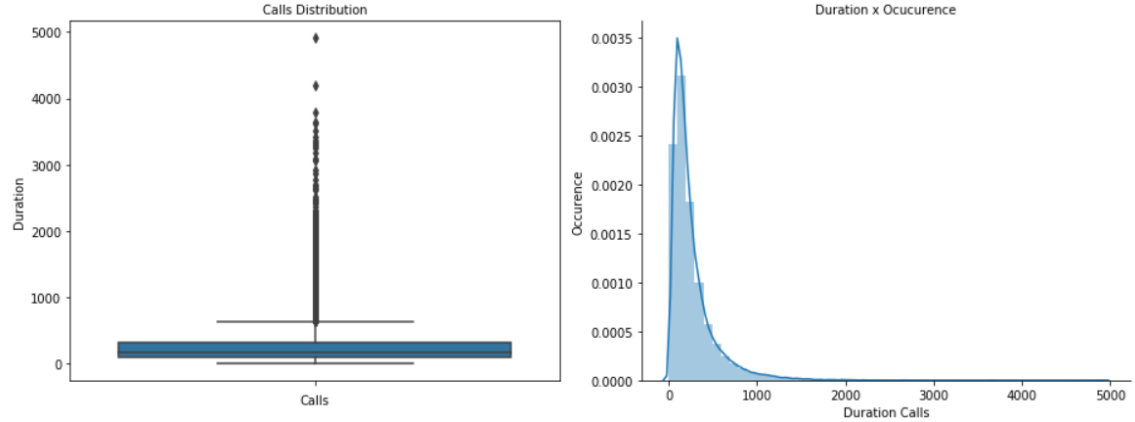
Preparation of Dataset:

Treatment of categorical and continuous variables

- Categorical variables like 'job', 'education', 'default', 'housing', 'loan', 'contact', 'campaign', 'previous', 'poutcome' is labelled as numbers in alphabetical order using label encoder package in sklearn.preprocessing. Missing values which is labelled as 'unknown' is also treated as on of the classes in each of the variables.
- 'Age' has 78 different values. Hence, it is converted into 4 classes based on the respective ages. There are 469 outliers viz. 1.14 % of the total dataset. Hence, it is ignored.



- 'duration' is also not normally distributed. Hence, it is converted into 4 classes based on median, IQR duration of calls.



- 'month', 'day of week' is also labelled as numbers using label encoder package in sklearn.preprocessing for the ease of classification.
- Rest of the continuous variables are scaled as per StandardScaler from Sklearn.preprocessing
- Dependent Variable 'y' is converted into binary variables which denotes whether a client subscribed to a term deposit or not.
- Dataset is divided into train and test sets with 80:20 ratio. No. of rows in training set = 32950 and test = 8238

K-fold cross-validation is performed with 10 splits and shuffle = False. This would help us to evaluate the models or algorithms to come up with the best classifier model.

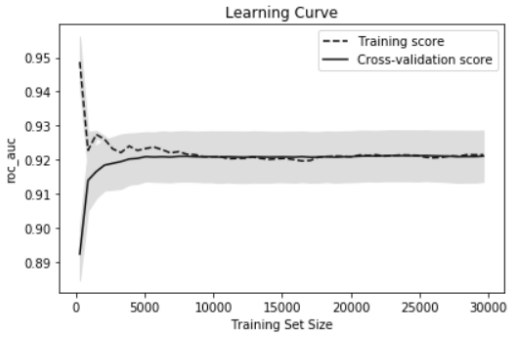
Siddhartha Shankar
https://github.com/sid0712
https://www.linkedin.com/in/sid-business/
469-509-4807
siddhartha.shankar@icloud.com

Model Statistics:

I. Logistic Regression

Classification Report & Confusion matrix *Learning curves to evaluate model performance*

```
[[7116  166]
 [ 603  353]]
              precision    recall  f1-score   support

           0       0.92      0.98      0.95      7282
           1       0.68      0.37      0.48       956

    accuracy                           0.91      8238
   macro avg       0.80      0.67      0.71      8238
weighted avg       0.89      0.91      0.89      8238

Accuracy Score: 90.7
Cross-Validation Score: 90.0
```

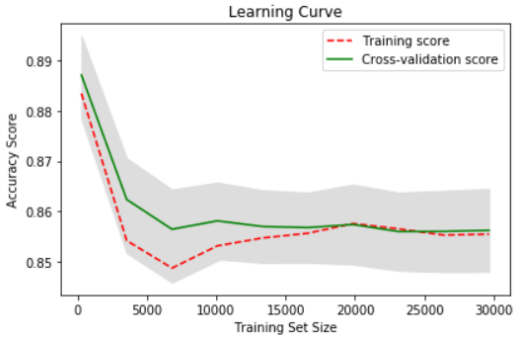


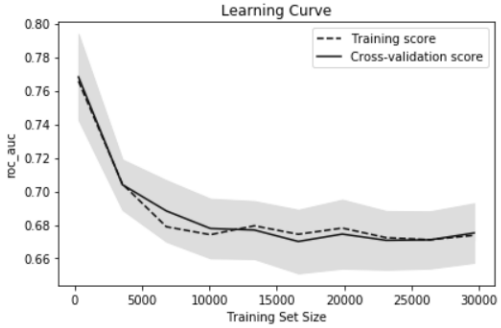
Model Accuracy on test dataset = 90.7%. Cross validation (CV) score = 90.0% and AUC = 0.92

II. Support vector Machines: (kernel = sigmoid)

Classification Report & Confusion matrix *Learning curves to evaluate model performance*

```
[[6676  606]
 [ 595  361]]
              precision    recall  f1-score   support

           0       0.92      0.92      0.92      7282
           1       0.37      0.38      0.38       956

    accuracy                           0.85      8238
   macro avg       0.65      0.65      0.65      8238
weighted avg       0.85      0.85      0.85      8238

Accuracy Score: 85.39999999999999
Cross-Validation Score: 88.7
```





Model Accuracy on test dataset = 85.40%, Cross validation (CV) score = 88.7%

III. Support vector Machines: (kernel = linear)

Classification Report & Confusion matrix *Learning curves to evaluate model performance*

```
[[7167  115]
 [ 758  198]]
              precision    recall  f1-score   support

           0       0.90      0.98      0.94      7282
           1       0.63      0.21      0.31       956

    accuracy                           0.89      8238
   macro avg       0.77      0.60      0.63      8238
weighted avg       0.87      0.89      0.87      8238

Accuracy Score: 89.4
Cross-Validation Score: 90.0
```

Model Accuracy on test dataset = 89.40%, Cross validation (CV) score = 90.0%

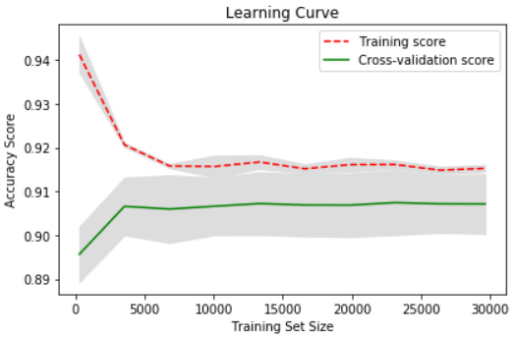
Siddhartha Shankar
https://github.com/sid0712
https://www.linkedin.com/in/sid-business/
469-509-4807
siddhartha.shankar@icloud.com

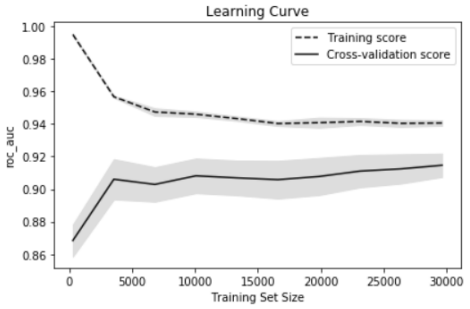
IV. Support vector Machines: (kernel = gaussian)

Classification Report & Confusion matrix

```
[[7163  119]
 [ 668  288]]
              precision    recall  f1-score   support

           0       0.91      0.98      0.95      7282
           1       0.71      0.30      0.42       956

    accuracy                           0.90      8238
   macro avg       0.81      0.64      0.69      8238
weighted avg       0.89      0.90      0.89      8238

Accuracy Score: 90.4
Cross-Validation Score: 90.7
```

Learning curves to evaluate model performance





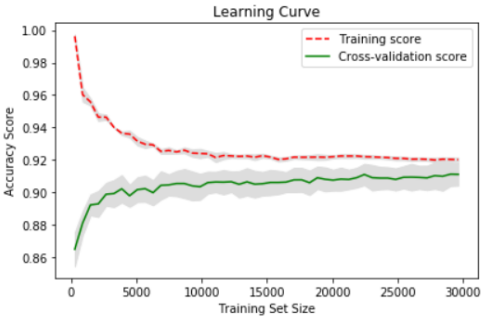
Model Accuracy on test dataset = 90.40%, Cross validation (CV) score = 90.7%

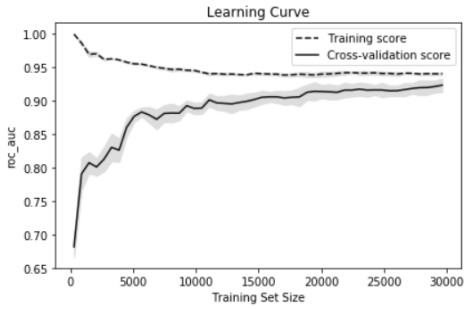
V. Decision Tree Classifier

Classification Report & Confusion matrix

```
[[7062  220]
 [ 512  444]]
              precision    recall  f1-score   support

           0       0.93      0.97      0.95      7282
           1       0.67      0.46      0.55       956

    accuracy                           0.91      8238
   macro avg       0.80      0.72      0.75      8238
weighted avg       0.90      0.91      0.90      8238

Accuracy Score: 91.10000000000001
Cross-Validation Score: 91.10000000000001
```

Learning curves to evaluate model performance

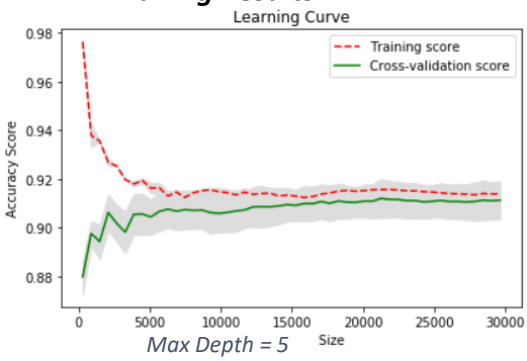




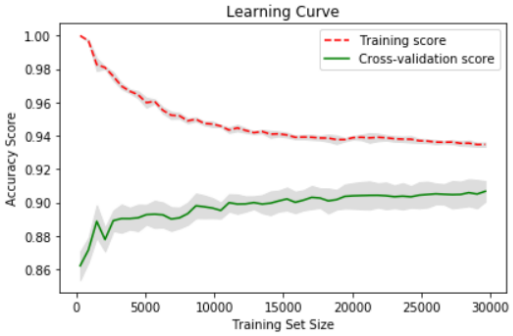
Model Accuracy on test dataset = 91.10%. Cross validation (CV) score = 91.10% and AUC = 0.92

Pruning Results

Max Depth = 1



Max Depth = 5



Max Depth = 10

With increase in tree depth, data tends to overfit; however, with less tree depth, data tends to underfit. Hence, as per the learning curve, max_depth = 7 is taken for tree as overfit-underfit trade-off.

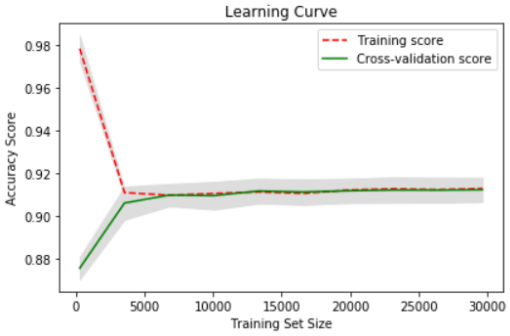
Siddhartha Shankar
https://github.com/sid0712
https://www.linkedin.com/in/sid-business/
469-509-4807
siddhartha.shankar@icloud.com

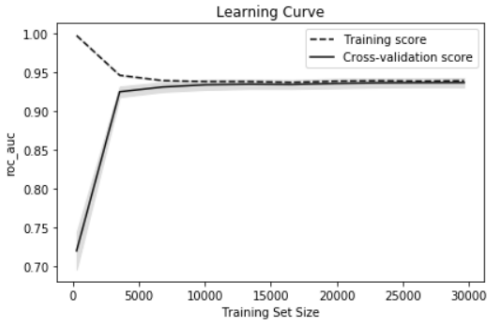
VI. AdaBoost Classifier

Classification Report & Confusion matrix

```
[[7110  172]
 [ 558  398]]
              precision    recall  f1-score   support

           0       0.93      0.98      0.95      7282
           1       0.70      0.42      0.52       956

    accuracy                           0.91      8238
   macro avg       0.81      0.70      0.74      8238
weighted avg       0.90      0.91      0.90      8238

Accuracy Score: 91.10000000000001
Cross-Validation Score: 91.2
```

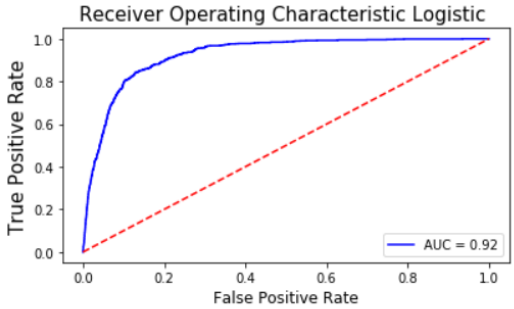
Learning curves to evaluate model performance

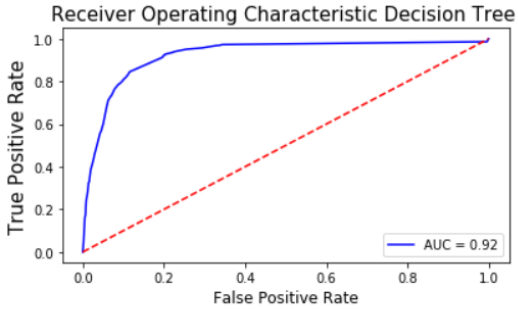


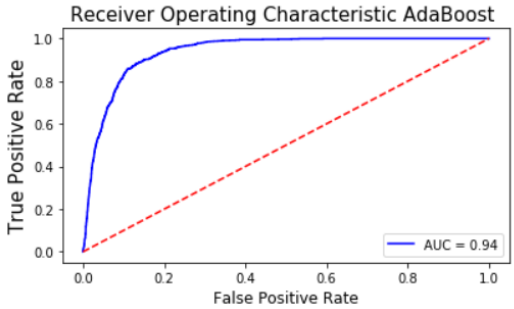


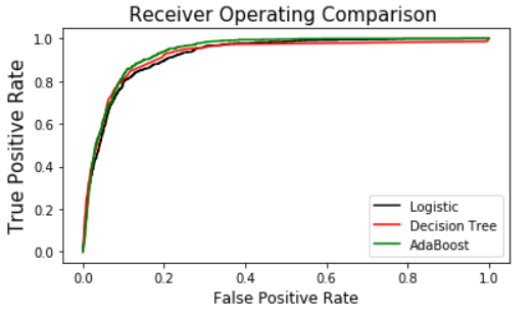
Model Accuracy on test dataset = 91.10%. Cross validation (CV) score = 91.20% and AUC = 0.94

ROC (Receiver Operating Characteristic) and AUC comparison between models









Cross-Validation Accuracy based on Accuracy

Models	Accuracy
AdaBoost Classifier	91.2
Decision Tree Classifier	91.1
SVM: kernel=Gaussian	90.7
Logistic Model	90.0
SVM: kernel=linear	90.0
SVM: kernel=sigmoid	88.7

Cross- Validation score helps us to validate the model with cross-validation score. Higher the Cross-Validation score, better the model is in performing classification.

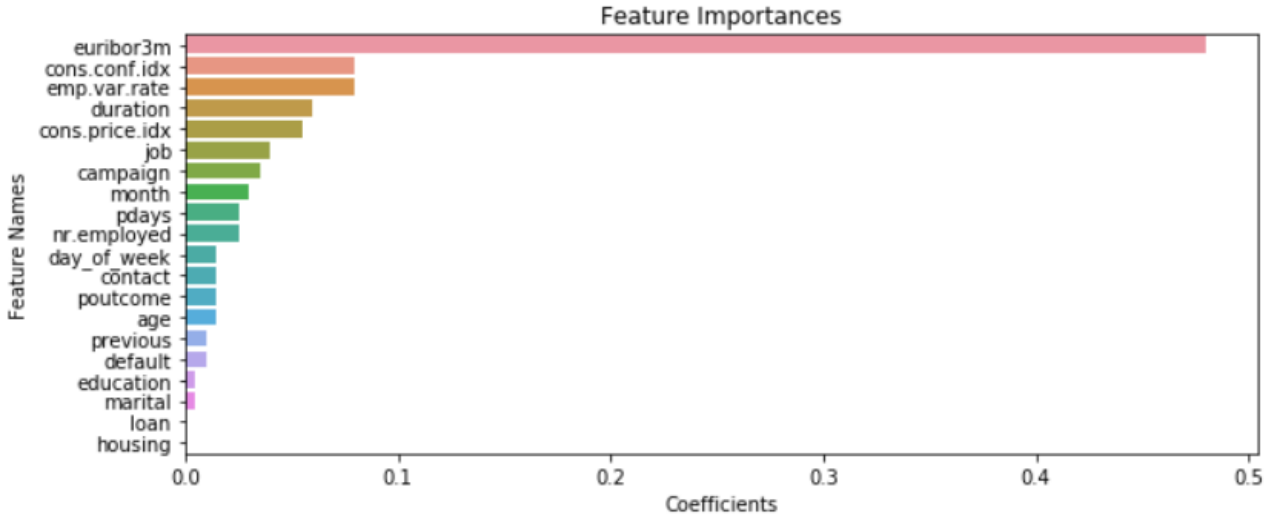
Based on ROC curve, AUC and Cross Validation (CV) score, *AdaBoost Classifier* performs best for the Banking Campaign : Term Deposit Dataset.

Siddhartha Shankar
https://github.com/sid0712
https://www.linkedin.com/in/sid-business/
469-509-4807
siddhartha.shankar@icloud.com

Since, Boosting is adaptive algorithm. Adaboost algorithm gives weight to the weak learners in algorithm and high weight to those learners who perform well. Hence, AdaBoost classifier tends to fit the dataset very well in comparison to other classifier model. This can be seen with high AUC and accuracy score as well and less fpr at tpr = 0.99.

Conclusion:

Recommendation based on Dataset II: Bank Campaign: Term deposit subscription



According to the important feature as per AdaBoost Classifier, the most significant attributes are 'duration', 'nr.employed', 'euribor3m', 'cons.conf.idx', 'cons.price.idx ,'emp.var.rate'. This is because the longer the conversations on the phone, the higher interest the customer is showing for the term deposit subscription. 'nr.employed', which is the number of employees in the bank is also significant. This can be because the more employees the bank have, the more influential and prestigious the bank is. 'euribor3m' is another important variable, which denotes the euribor 3-month rate. This indicator is based on the average interbank interest rates in Eurozone. It suggests that the higher the interest rate the more willingly customer will subscribe to the term deposit. Employment variation rate 'emp.var.rate' suggests that change of the employment rate will make customers less likely to subscribe a term deposit. This makes sense because the employment rate is an indicator of the macroeconomy. A stable employment rate denotes a stable economic environment in which people are more confident to make their investment.

Therefore, if banks want to increase their client base with regards to the term deposit, they should hire more people to work for them, improve the quality of conversation on the phone and run their campaigns when interest rates are high and macroeconomic environment is stable.