Analysis on Portuguese Bank Marketing Data

- **Harshadkumar Choudhri**
- **Completed**

Summary:

Introduction:

The data is about Portuguese banking institution's direct marketing campaigns. This banking institute's campaigns were mainly based on the phone calls. Same clients being contacted more than one once to know the status of their client's bank term deposit whether they are subscribing (yes) or no. The aim is to increase campaign efficiency by identifying the main factors that affect the success of a campaign and predicting whether the campaign will be successful to a certain client. Developing a more efficient and precise campaign strategy which helps to reduce the costs and improve the profit. The variables in the data set are numeric and categorical. The dataset contains 45211 observations with 20 attributes.

Model and Results:

Since data were collected from phone call interview, there were many clients who refused to provide their personal information due to privacy factor. Hence, the data had few missing values in the variables like 'education', 'marital', 'job' and 'housing' which needs to be imputed and deleted if they are not strongly correlated with the target variable. I used contingency table and chi-square to identify the importance of available unknown values and imputation is done based on contribution to target variable. Moreover, before building the model and checking the accuracy-based validation dataset, data sampling is done to avoid the effect of oversampling. Since the data had 88.70% of responders as they accept the term deposit and 11.3% of non-responders as they reject the term deposit, hence dataset is sample replacement from the group with less data until the number equals to the larger group. Three different algorithms logistic regression, neural networks, and Decision tree(CART) are implemented to test the model of subscription of term deposit and measured the efficiency of based on test. Moreover, Ensemble method was used to compare the accuracy of different algorithms performance. Ensemble method chooses the maximum accuracy performance from all the three algorithms and neural networks gives the best model performance with 92.56% accuracy.

Recommendations:

In the light of overall test accuracy and based on ensemble model performance, neural networks perform better with 92.56% accuracy. It delivers the most powerful prediction ability. It is important to find the factors which are most important and influencing the customers decision to term subscription. According to performed analysis and model performance, we can suggest most influencing variables such as duration, number of employees in the bank (nr. employed), Employment rate (emp.var. rate), age and months in which the client is being contacted. Based on the coefficient signs of variables in logistic regression, "duration" has positive effect on people saying

“yes” because the longer is the call duration, the higher interest the customer will show to the term deposit. “nr. employed”, which is the number of employees in the bank, has positive effect for turning people to subscribe the term deposit as they can engage more clients on the telephonic conversion and quality calls might turn them to subscribe term deposit. Therefore, if the banks want to increase their lead generation, they should hire more people to work for them, improve the quality of conversation on the phone, and target the months such as may, the group of age between 30 to 35.

Introduction:

The data is about Portuguese banking institution’s direct marketing campaigns. This banking institute’s campaigns were mainly based on the phone calls. Same clients being contacted more than one once to know the status of their client’s bank term deposit whether they are subscribing (yes) or no. It is essential to understand the factors which involved in client’s account sign up for the short-term deposit and developing strategy based on that to get most promising leads to win them over. Hence, developing a more efficient and precise campaign strategy which helps to reduce the costs and improve the profit. The variables in the data set are numeric and categorical. The dataset contains 45211 observations with 20 attributes.It contains variables such as age, type of customer, month, marital status, housing loan, personal loan, education, and nr. employed.

Objective:

Our goal is to build a classifier to predict whether a client will subscribe to a term deposit or not. The aim is to increase campaign efficiency by identifying the main factors that affect the success of a campaign and predicting whether the campaign will be successful to a certain client. Moreover, developing a more efficient and precise campaign strategy which helps to reduce the costs and improve the profits.

Meet the data:

Data: Bank Marketing Data
Source: https://archive.ics.uci.edu/ml/datasets/bank+marketing

The data were collected from a Portuguese marketing campaign related with bank deposit subscription for 45211 clients with 12 categorical variables and 7 numerical variables and 20 features

Age(Numeric): Age of the client.
Job:('admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician')
Marital (Marital Status): Categorical: 'divorced','married','single','unknown';
Housing loan (Yes/ No): Has credit in default?
Personal loan (Yes / No): Has housing loan?
Contact(Cellular/Telephone): contact communication type (categorical: 'cellular’, ‘telephone')
Month: last contact month of year.

Duration (last contact duration in seconds)
campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
emp.var. rate: employment variation rate - quarterly indicator (numeric)
cons.price.idx: consumer price index - monthly indicator (numeric)
cons.conf.idx: consumer confidence index - monthly indicator (numeric)
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)
euribor3m: euribor 3-month rate - daily indicator (numeric)
nr. employed: number of employees - quarterly indicator (numeric)
previous: number of contacts performed before this campaign and for this client (numeric)

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

Data Preprocessing:

Since data were collected from phone call interview, there were many clients who refused to provide their personal information due to privacy factor. Hence, the data had few missing values in the variables like 'education', 'marital', 'job' and 'housing' which needs to be imputed and deleted if they are not strongly correlated with the target variable. I used contingency table and chi-square to identify the importance of available unknown values and imputation is done based on contribution to target variable. Moreover, before building the model and checking the accuracy-based validation dataset, data sampling is done to avoid the effect of oversampling. Since the data had 88.70% of responders as they accept the term deposit and 11.3% of non-responders as they reject the term deposit, hence dataset is sample replacement from the group with less data until the number equals to the larger group. To produce a proper model of this data, all variables must be able to be represented by the same attributes. Once all observations can be evaluated by the same variables, some selection criteria can be used to determine what variables are important to produce an unbiased model. Of course, this chosen model is compared to one that includes all variables, to ensure that only insignificant variables have been removed. Once variables have been selected, checks must be complete to confirm that the model does not violate any basic assumptions.

Methods used for analysis:

1)Logistic Regression

Logistic regression is known as probabilistic classification model used to predict the binary response from binary classifier. Logistic regression is independent of various class imbalance. Often it predicts better than several different classification algorithms. There are many categorical predictors have imbalance distribution in the data, so building a model using full set of predictors often degrade the performance of the system then the predictors of near zero variance.

2)Neural Network

Neural networks perform better in most of the occasions as it uses the weights(error) of the output to improve the accuracy of the system by reducing the variation in the system. 'Back propagation' algorithm is used in multilayer feed forward neural network. Each layer uses the output of previous layer to improve the performance. It is time consuming but gives the best accuracy among the different data mining algorithms. Moreover, neural networks are known as 'black box' prediction model because it lacks in finding the insights of relations among predictors and outcome variable.

3)Decision Tree(CART)

Decision tree is easy to use and understand. It produces the rules that are easy to interpret and implement. Moreover, variable selection & reduction is automatic, which helps in dimension reduction and achieve the better performance in the model. Decision tree do not require assumptions of statistical model, which is desirable. Moreover, it can work without extensive handling of missing data. The only reason, we should to opt different algorithm is accuracy, as it gives instable and poor predictive performance.

4) Ensemble method:

Generally, ensemble methods are used for combining two or more similar algorithms to produce the best model performance. Ensemble method always performs better than other algorithms because it takes the maximum accuracy among all the different implemented models. It basically cancels out the errors produced in the various algorithms as errors can be positive or negative, hence it gives the error free output in response of different inputs in terms of accuracy.

Graphical representation of numerical variables:

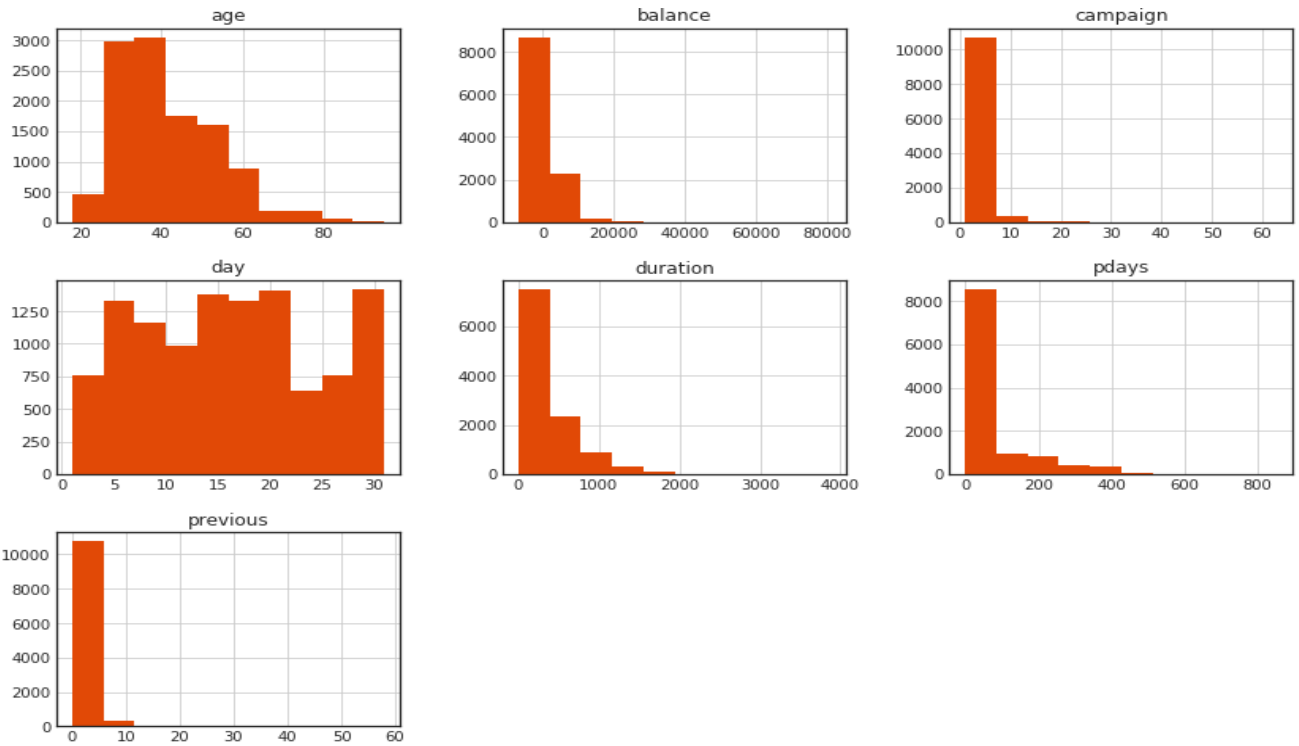

Figure1

Above plots shows the overall analysis of different predictors involved in predicting the term deposit subscription. **Figure 1** shows that, average age is 41 years old, minimum age is 18, and maximums age is 95. Moreover, mean balance is around 1500 and standard deviation is somewhat higher, which means balance is highly distributed all over the dataset. Also, duration is towards the lower side as mean is around 4 minutes after converting it into minute form. There are no much insights we can get from general graphical representation.

Analysis:

What percentage (%) of potential clients accepted to subscribe term deposits vs refused to subscribe term deposits?

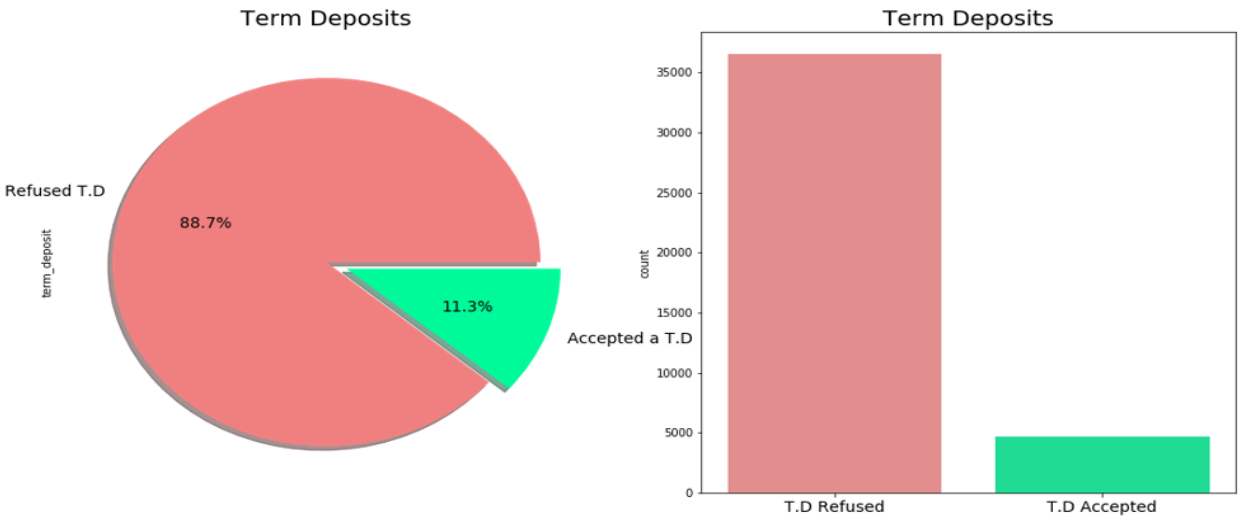

Figure 2

Figure 2 shows the percentage of potential clients accepted to subscribe the term deposit and refused to subscribe to term deposit. As we can see that, 88.7% of clients refused to take the term subscription whereas only 11.3% of potential clients accepts the term deposit. We can also infer that, count of term rejection is exceeds 35000 where as acceptance is somewhat around 5000.

Which Months has highest Subscription of term deposit?

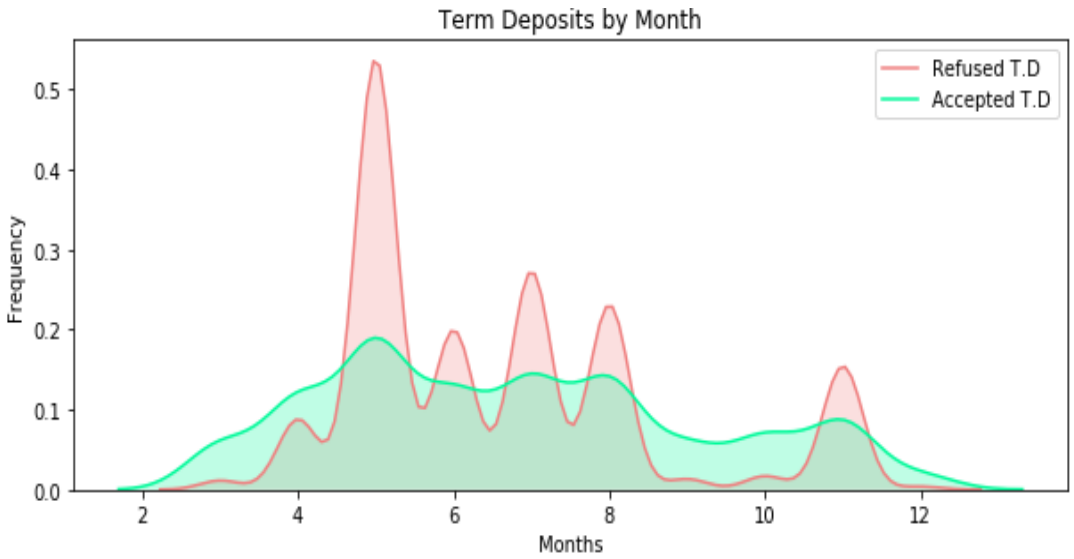

Figure 3

From figure 3, we can tell that may month was the highest marketing activity month with the higher term rejection ratio but also may is the month where highest term subscription is achieved. There is a wide gap between acceptance and rejection ratio of term deposit in may month. Moreover, the months like June, august, September, and December had the highest ratio as request were lower in these months,

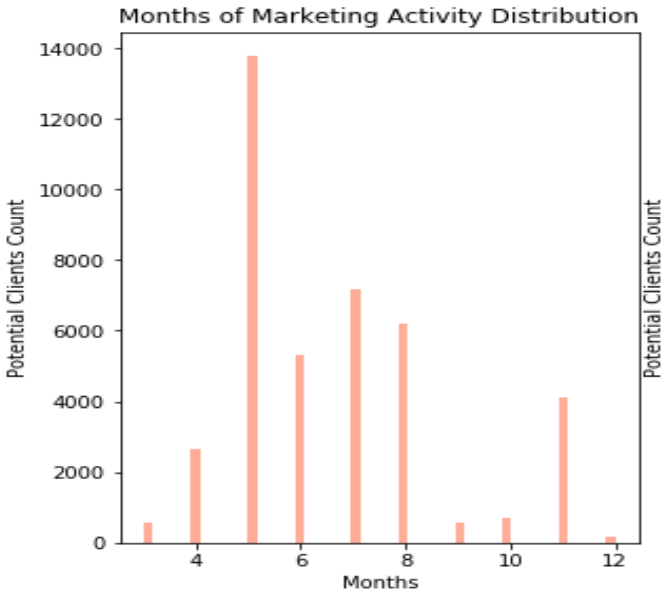

Figure 4

Above figure also describe the term acceptance and rejection ratio in different months and campaign's marketing activity. It shows that may month had 14000 potential clients count towards term subscription.

Which Seasons has the highest subscription of term deposits?

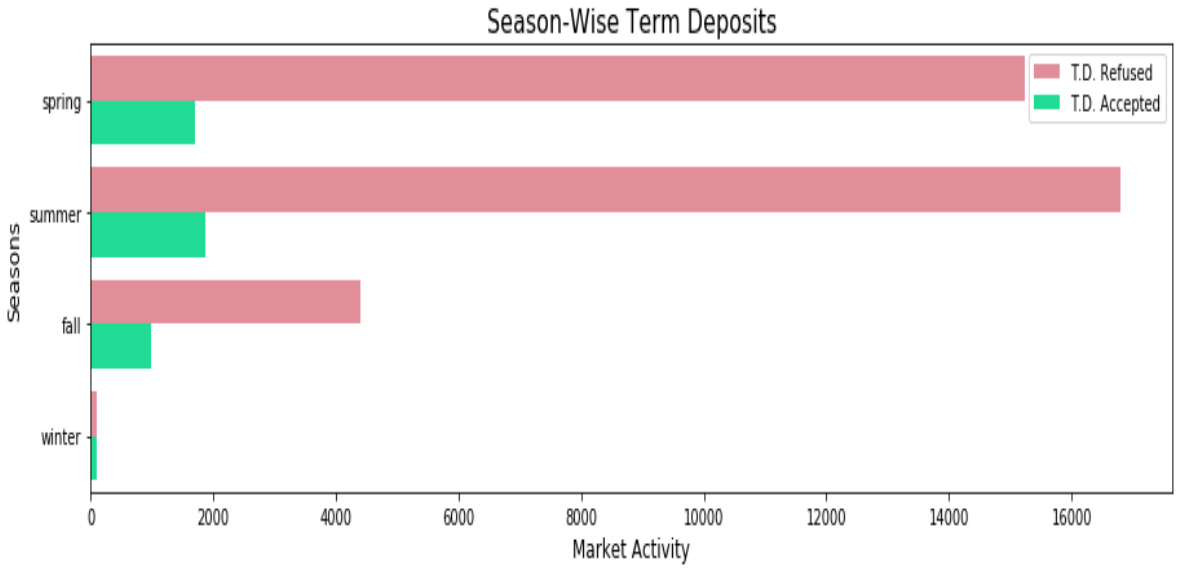

Figure 5

Figure 5 shows the season wise term deposit subscription, surprisingly clients prefer to subscribe more during spring and summer season. On the other hand. Winter season records the lowest subscription of term deposit. Hence, marketing campaign should be increased in winter and fall seasons to earn more subscriptions.

Is there any effect of Call Duration to Term Deposit Subscription?

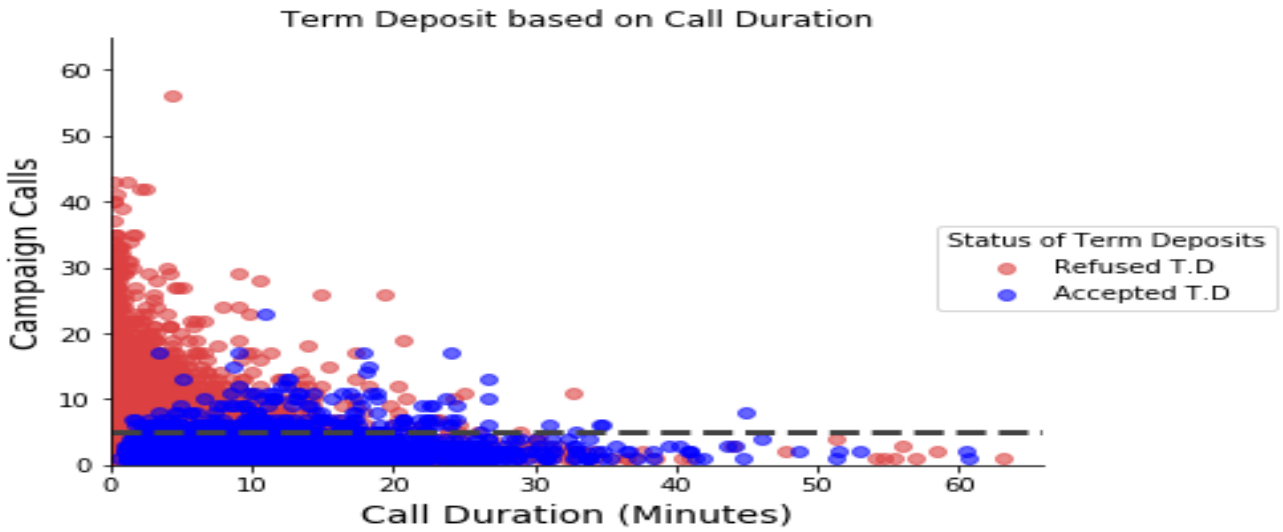

Figure 6

Above figure 6 shows the relation of call duration to term subscription of client. We can see that, if the client is contacted more on same marketing campaign than potential client will more likely to reject the term subscription. Moreover, if the customer is contacted more than 5 times then, it can result in lead generation towards term subscription. After the third call probability od acceptance decreases drastically.

How Age affects the subscription of term deposit?

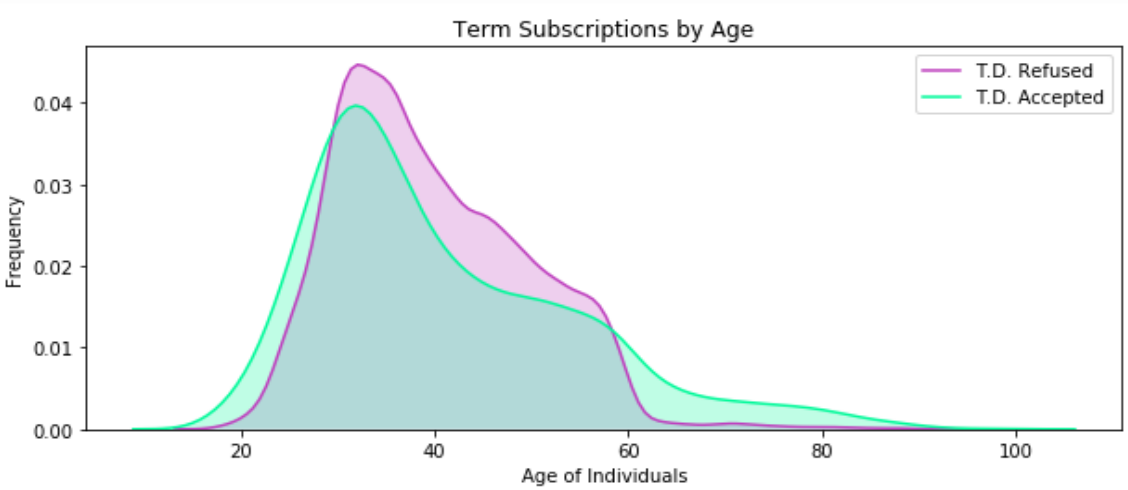

Figure 7

Client's age plays can play an important role in analyzing and suggesting marketing campaign to increase their lead generation. From the figure 7 we can see that, bank has targeted 30 to 35 years of individual and they succeeded in that, as most client posses job in this group of age. Interestingly, younger and older groups accept more term subscriptions, so next marking campaign should focus more on these group of people more to increases their subscription. Age 60 and higher accepts almost 75% of term deposits.

Correlation between variables:

Figure 8

Figure 8 shows the correlation among all the predictors and their contribution towards term deposit subscription. We can see that more green boxes represent more positive correlation between term subscription and predictors and more red shows the negative relation among them. We can see that, call duration adds more contribution to subscription of term deposit whereas previous campaign shows the opposite of call duration. Interestingly, nr.employed shows the negative relationships but after modeling that variable, I found it is very important variable to increase the term subscription. Moreover, relation of client's housing and credit loans makes significant difference in term deposit subscription. Hence, marketing campaign should focus on average and high balance profile clients to increase the likelihood of term subscription

Results:

Decision tree:

To check the model performance, I implemented decision tree and compared with the different leaves which produces the best accuracy performance. The maximum validation score measured as 0.9158(91.58%) for the 35 leaf's as it reduces after thereafter.

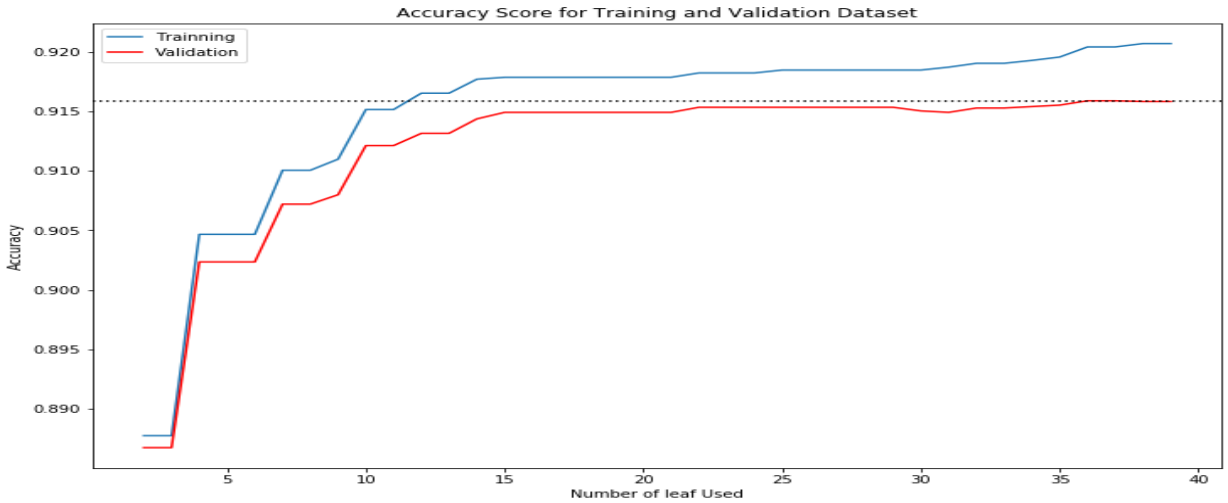

Figure 9

33	35.0	0.919553	0.915513
34	36.0	0.920403	0.915878
35	37.0	0.920403	0.915878

After implementing decision tree to check the influence predictors in term subscription, we found that duration of the client call, contact made in that month, month in the client contacted, client 's status in previous marketing campaign (previous)and the client's housing loan are the influencing factors in client's subscription towards the term deposit.

Neural Networks:

In neural networks, for training procedure, back propagation algorithm is used to produce the best accuracy model as it chosen as the best numerical optimal method. 5 cross validation is used to select the best parameters which uses the average AUC value in validation score for selection measurement. I used different activation methods such as 'relu' and 'logistic, but 'relu' produces the best accuracy as 93.59 in the validation set.

```
print(accuracy_score (yVald, nn.predict(xVald)))
print(accuracy_score (yTrain, nn.predict(xTrain)))
nn.score(xTrain, yTrain)
```

```
0.906470016994
0.935901586274

0.93590158627387499
```

The difficulties which I faced during neural network implementation is it consumes lot of time to test the model performance, and there is no selection measure for producing the best accuracy, as I check 50 different combinations of various layers and number of nodes with activation functions.

Logistic regression:

In implementing logistic regression method, I practiced 5=fold to 10-fold cross validation to tune the parameters for classifiers. After finalizing the tune model, I tested its performance on the testing set (validation set) for measuring the accuracy of the model and the best accuracy achieved as 91.37% on the validation set. I changed

the prediction threshold and choosing different sampled data as data is imbalanced. I also checked raw version of logistic model to judge the accuracy of the model, but I got around 0.87 to 0.89 accuracy score which it was less than the 0.9137 score achieved on 10-fold cross validation.

Ensemble method:

Generally, ensemble methods compare all the model accuracy to give the best model performance. Here, ensemble methods were used to compare the results from logistic regression, neural networks, ad decision tree classifier. By comparing the model accuracy and selecting the maximum accuracy, neural networks performed better that other algorithms by producing the best accuracy performance of 93.59 %. Ensemble method makes model more robust and stable thus ensuring the decent performance on the test cases. Ensemble method always performs better than other algorithms because it takes the maximum accuracy among all the different implemented models. It basically cancels out the errors produced in the various algorithms as errors can be positive or negative, hence it gives the error free output in response of different inputs in terms of accuracy.

Recommendation for next marketing campaign:

Ensemble method ensures that neural networks perform better with 92.56% accuracy. It delivers the most powerful prediction ability. It is important to find the factors which are most important and influencing the customers decision to term subscription. According to performed analysis and model performance, we can suggest most influencing variables such as duration of the call, number of employees working in the organization, month in which the contact is being made, also the employment rate. Hence, if the banks want to increase their lead generation, they should hire more people to work for them, improve the quality of conversation on the phone, and target the months such as may, the group of age between 30 to 35.

Actions should bank consider improving the number of term subscriptions are as follows:

1)**Months of marketing activity:** For the next marketing campaign, it will be wise for the bank to focus more in targeted months such as may, September, and October. However, may is the month were highest rejection is achieved in terms of subscription. Moreover, December should be taken into consideration because it was the month where lowest marketing activity is done.

2)**Campaign call:** Suggestion for the next marketing campaign would be, bank should implement a policy stating no more contact above 3 to the same client, else there are higher chance of turning same client into rejection of deposit. Bank should focus on other customers instead of calling same client to get the new customers.

3)**Age of the client:** The next marketing campaign should target potential clients in between age of 30 to 35 as they are more likely to subscribe to term deposit and bank should keep focus in mind about the rejection as well for the same age group. Moreover, youngest and the oldest aged customers are more likely to accept the term subscription.

4)**Seasonality:** Bank should target the seasons such as spring and summer for the marketing campaign as they give the higher subscription ration then the winter and fall. However, bank can change their marketing strategy in fall and winter seasons to increase the lead generations.

5)**Housing and credit loans:** By changing the relation of client's housing and credit loans makes significant difference in term deposit subscription. Hence, marketing campaign should focus on average and high balance profile clients to increase the likelihood of term subscription.

Developing a questionnaire during the calls plays an important role as it correlates mostly positive adding more contribution to increase the lead generation and achieving more term subscription. It does not guarantee that the long calls results in subscription of term deposit but adding quality into conversation with more interesting questions to the client makes him engage on to the phone call.

References:

- **https://archive.ics.uci.edu/ml/datasets/bank+marketing**
- **https://www.shutterstock.com/search/banking**
- **https://firefighters.co.nz/interest-rates/**
- **http://www.columbia.edu/~jc4133/ADA-Project.pdf**
- **https://www.kaggle.com/janiobachmann/bank-classifying-term-deposit-subscriptions/notebook**