PREDICTING BANK MARKETING OUTCOME FOR TERM DEPOSIT SUBSCRIPTION

Table of Contents

Abstract

In this study, we examine a bank telemarketing dataset obtained from UCI repository. The data was from a previous marketing campaign run by a bank to promote term deposits to its clients. In this study, we try to predict whether or not a client would subscribe to a term deposit. Prediction was done with the help of various classification models. We explored a couple of these models and found out that a decision tree classifier with a test size of 0.50 provided the highest accuracy (87 percent) out of the other models employed in this study.

Introduction

Marketing is an essential part of today's business philosophy. Any product or service if not marketed to the right audience will result in its quick demise, regardless of how good the product or service is. For this reason, it is important to make sure that marketing campaigns are optimized and strategized in such a way that it appeals to the target audience. This is especially important in the banking sector, where there are hundreds of banks offering similar services competing with each other to grab the attention of the populace. Banks are constantly trying to undercut each other's prices and trying to reach people first. This competition has grown fiercer with each passing year. Every year millions of dollars are being pumped by banks into their marketing campaigns to try and get hold of a significant share in the market. This figure is expected to rise in the future as new marketing platforms and innovations keep rising up. If these marketing campaigns fail to garner any attention or fail to acquire new clients, then this could spell serious trouble for the banks. So, it's important to make sure that their marketing efforts squandered by bad marketing strategies. In this study, we explore a bank telemarketing campaign dataset containing various client centric features and the advertised feature, term deposit (y). The aim is to use the data obtained from clients of a previous marketing campaign, to try and predict whether or not they would end up taking a term deposit. By being able to accurately predict the outcome, the banks can then use this information to specifically target the clients with attributes that yield successful outcomes. This in turn is going to bring the marketing cost down leading to more profits for the bank. During the study, we also explore some correlations between pairs of features, this will enable us to gain some useful insights that can be put into action for future marketing campaigns.

Methodology

Data Acquisition

The bank marketing dataset used for this study was taken from the UCI Machine Learning Repository. It contains 45,211 observations across 17 features. The features consist of both categorical and numerical features. The categorical features include education, marital status, job etc. The numerical features include age, balance, duration etc. The feature, term deposit is the target variable in this study. Term deposit consists of Yes (subscribed to term deposit) and No (didn't subscribe to the term deposit) values.

No.	Feature Name	Feature Type	Feature Description
1	Age	Numeric	Age of the client
2	Job	Categorical	Job of the client
3	Marital	Categorical	Marital Status of the client
4	Education	Categorical	Education background of the client

5	Default	Categorical	Client has credit in default? (Yes/No)
6	Balance	Numeric	Client bank balance
7	Housing	Categorical	Client has housing loan? (Yes/No)
8	Loan	Categorical	Client has personal loan? (Yes/No)
9	Contact	Categorical	Mode of contact
10	Day	Numeric	Last contact day of the week
11	Month	Categorical	Last contact month of year
12	Duration	Numeric	Last contact duration, in seconds
13	Campaign	Numeric	Number of contacts with the client performed during this campaign and also outside it.
14	Pdays	Numeric	Number of days prior to which the client was last contacted from a previous campaign
15	Previous	Numeric	Number of contacts performed before this campaign
16	Poutcome	Categorical	Outcome of the previous marketing campaign
17	Y(Term Deposit)	Categorical	Client subscribed a term deposit? (Yes/No) - Target Variable

Table 1. Descriptive statistics of each feature

Data Preparation

The first step was to import all the required libraries like pandas, matplotlib and seaborn. The pandas library offers data structures and operations for manipulating the data. While matplotlib and seaborn are both python data visualization libraries used for plotting graphs later on during data exploration. The data was then loaded onto the notebook and put into a pandas dataframe. Following this, the dataframe was then compared with the csv file to check if the imported data was equivalent to the one present in the file.

The next step was to clean the numerical features in the dataset. The numerical features were first checked for any missing values by using the isnull() function. The isnull() returns a Boolean value based on whether or not a particular cell contains a missing value. The total count of all the Boolean values was taken from each feature using value_counts() and it resulted in all the features showing false meaning that there are no null values present in any of the numerical features. Since no ranges for numerical features are mentioned, we cannot check the features for outliers. Next, we remove features which contain majority unknown values as they contribute next to nothing to our study. Features like pdays contain more than 36000 unknown values, which is more than three fourths of the data, the main reason for this is that most of the people in this campaign are being contacted for the very first time, as seen from the previous feature. Because of all these reasons, the pdays feature is dropped from the dataset. There are features like contact and poutcome which also harbor a lot of unknown values but removing these features will result in the loss of valuable data which can affect the result.

Finally, we clean the categorical features. Since categorical data can contain extra whitespaces, they were first stripped off any possible extra whitespaces using the str.strip() function. Before this function was called, the objects were first grouped based on the object data type and passed into a lambdas expression one by one, where the function was called on them. Next, the categorical features were also checked for any missing values by using isnull() function. The result shows that there are no null values in any of the categorical features in the dataset.

Data Exploration

Under data exploration, we take an in-depth look at some of the features in the dataset. We have used the seaborn library for visualizing the graphs. Count plots were employed for visualizing categorical features while distribution plots were used to visualize numerical features.

First, we explore some of the relevant categorical features. Job is the first categorical feature in the dataset, as you can see there are a wide range of jobs undertaken by the clients involved in this campaign. People who do blue-collar and management jobs make up the majority of the dataset. While entrepreneurs, students and housemaids make up a small percentage of the clients. It's also important to note that there is a small percentage of clients whose jobs are unknown, but they only make up a small portion of the data, that excluding them won't make much of a difference.
The next categorical feature is marital status, where from the graphs we can see that there is a large percentage of married clients compared to single and divorced clients. Divorced clients make up the smallest demographic among the clients.

The next categorical feature is education which showcases the educational background of all the clients. From the graph, we can see that clients with secondary and tertiary educational backgrounds make up most of the clients in the dataset. While people with primary educational background are comparatively less in number.
Following this, we have the housing loan feature which shows the number of clients that have taken a housing loan and those clients who have not. Looking at the graph, you can see that the number of clients who have taken a housing loan is more than the number of clients who haven't taken one. Up next, we have the loan feature which shows whether the clients have taken a loan or not. The number of clients that haven't taken a loan is far greater than those who have not.

The categorical feature contact, which represents the mode of contact through which the client was contacted during the campaign. Judging by the graph, we can say that majority of the clients were contacted through the cellphone, followed by a small percent contacted on the telephone. This infographic does make a lot of sense as nowadays cellphones are lot more widespread than the telephones. It's also important to note that there 13,000 clients whose contact mode is unknown.

Month is the next categorical feature in the dataset. It shows the month during which the client was last contacted. From the graphs, we can see that almost 14,000 people were last contacted during the month of May, which is the highest among all the other months. This could be because the campaign began in May, as a result of which a large amount of calls were made right at the beginning to get the campaign off to a solid start. December recorded the lowest amount of calls in the year, this could possibly be because of the holiday season taking place at that time.

The final categorical feature is Poutcome. It contains the outcome of the previous campaign the client was a part of. The poutcome for almost 36,000 clients is unknown, this maybe because the majority of the clients are involved in a marketing campaign for the very first time. Apart from the unknown outcomes, there are around 5000 outcomes that were a failure in the previous campaign, followed by a few success and other outcomes.

Finally, we explore the relevant numerical features. Age is the first numerical feature in the dataset and judging by the graph we can see that there are a lot of clients between the age of 20 to 40. We can also see that there is a massive dip in the histogram after the age of 60 suggesting that there is comparatively less number of clients above the age of 60. By observing the distribution curve, we can see that age is normally distributed among the clients. The next numerical feature is duration. Duration shows the last contact duration with the client in seconds. From the distribution plot, we can observe that almost 90 percent of the calls lasted between 0 to 500 seconds. While the remaining 10 percent lies beyond 500 seconds. Duration however is not normally distributed as the distribution curve is skewed towards the right (positive skew).

Feature Relations

In this section, we will be exploring some interesting relations between pairs of features in the dataset.

- Married clients are more likely to take a term deposit than single & divorced clients.

  To visualize the relation between marital status and term deposit, a count plot was first plotted, and the values were divided based on the y (term deposit feature). Looking at the graph we can see that clients, who have not taken up a term deposit are more in number compared to those who have, in all three of the categories. This is especially noticeable in the married section, where the ratio between the clients who have said yes or no to the term deposit is the highest out of the three. This could be because married clients have a family to support and thus don't have surplus amount of money to put into a term deposit. The ratios for both single and divorced clients are almost similar to each other. Thus, we can conclude by stating that there is a higher chance that single and divorced clients are more likely to take up a term deposit. Thereby proving the above hypothesis wrong.

- Clients with a tertiary educational background are more likely to take a term deposit compared to clients with primary or secondary educational backgrounds.

  The count plot shows the relation between a client's educational background and term deposit. We can see that clients that have said no to the term deposit are greater in number compared to the ones that said yes. Secondary in particular has around 20,000 people, who has not taken the term deposit. Comparing the ratios between yes and no of all the categories, secondary has the highest ratio followed by tertiary and primary, who roughly similar to each other. Thereby, we can come to the conclusion that clients of tertiary educational backgrounds are more likely to take a term deposit than clients of secondary educational backgrounds but at the same time, they are equally as likely as clients of primary educational backgrounds to take the deposit. Thereby proving the above hypothesis wrong.

- Clients doing blue-collar jobs are more likely to not take a term deposit compared to clients doing white-collar jobs.

  We can clearly see from the graph that a large percentage of blue-collar clients have not taken a term deposit. This pattern does make sense as blue-collar jobs provide less income compared to other jobs. As a result of which majority of the blue-collar clients tend to shy away from investing a sum of money as term deposit. Out of the clients that have taken a term deposit, clients with management jobs have the highest percentage. But when we take the ratio of yes to no for the term deposit, student clients have lowest ratio. Also, across the board, people who have said no to the term deposit are far greater than the ones who said yes. So, its better to target clients with white-collar jobs in the future marketing campaigns.

- Clients who have taken up a term deposit have a higher average bank balance compared to those who have not.

  Judging by the bar plot, we can see that the mean bank balance for clients who had taken up a term deposit is way higher that the ones who have not. The average bank balance for clients who had taken a term deposit is around 1800 as opposed to 1300 for those who had said no to a term deposit. From this correlation, we can draw a conclusion that in the future it's better to target clients with a bank balance of 1800 and above.

- Retired clients have a higher average call duration compared to others.

  From the bar plot, it's clearly evident that retired along with unemployed clients have a higher average call duration than the rest. This maybe because retired and unemployed clients may not be as busy as other clients with jobs. As a result of this, retired clients can take some time to listen to the caller instead of quickly ending the call. This longer call duration would allow banks more time to persuade the client into take a term deposit. So, by targeting the retired population in the next campaign, there is a high chance that the bank can gain some new customers.

- Clients whose previous campaign outcome was successful are more likely to take a term deposit in the current campaign.

  So, this count plot shows the outcome of the previous campaign that the client was a part of and also the outcome (term deposit) for the marketing campaign used in this study. From the graph, we can clearly see that there are almost 40,000 clients, whose previous campaign outcome was unknown. This could be because a majority of the clients are being contacted for the very first time in this campaign. Looking at the failure attribute, we can see that clients who said no, both times are much higher than the clients who said no for the previous campaign and yes for the current one. So, in the future campaigns, it might be better to decrease the amount of calls made to clients who declined both times. Coming to the success attribute, we can see that the number of clients who were a success in both the campaigns are slightly more than the clients who were a success only in the previous campaign. Thereby, we can conclude that by reaching out to clients who were a success in the previous campaigns, there's a higher chance that the client may be success in the current campaign. Thus, the above hypothesis is proven right.

- Increasing the number of calls to a client is not going to increase the probability of the client taking up a term deposit.

  From the bar plot, we can see that the clients, who said yes to a term deposit required only an average of about 2 calls to be sold on the idea, while an average of about 2.8 calls were made for the ones that said no. More number of calls were being made to clients who said no, in an effort to persuade them further into taking up the deposit. But now, since we know that only 2 calls are being made to clients who said yes, there's no point in calling a client for the third time if he/she has said no the last 2 times. With this vital information, the bank can then enforce that only 2 marketing calls be made per client per campaign. This greatly helps the marketing team to shift their focus sooner to other potential clients out there. Thus we can conclude that the above hypothesis is correct.

- Clients who haven't taken a personal loan are more likely to take the term deposit than clients who have taken a personal loan.

  Judging by the count plot, we can see that clients who haven't taken a personal loan and have taken a term deposit are more than clients who have taken both. This correlation does make sense as people usually take loans so they can put that money to use rather than give it back to the bank to keep as a term deposit. So, for future campaigns the marketing team must target clients who haven't taken up a personal loan from the bank, as there is a higher probability that they will take the term deposit compared to clients who have taken a loan.

- Clients above the age of 40 are more likely to take a term deposit than clients below the age of 40.

  The bar plot shows the relation between age and term deposit. From the graph, we can see that there is no significant difference between the mean ages of clients who said yes and no to the term deposit. The mean age for both are around 40 meaning there is correlation between age and term deposit. We can conclude by saying that age does not factor in when it comes to a client's decision on whether or not to take a term deposit. Thereby, proving the above hypothesis wrong.

- Clients who don't have credit in default are more likely take a term deposit than those clients who have.

  The count plot shows the relation between the features, default and term deposit. From the graph, we can see that clients who haven't defaulted their credit and have taken the term deposit are more than the clients who have defaulted their credit and have taken the term deposit. The latter category, in fact has zero clients. This could have been because no clients who have defaulted was called during the campaign, but that assumption is incorrect as the bank has reached out to clients who defaulted and every single one of them have said no. So, we can conclude by saying that it is pointless to approach clients who have defaulted, with the term deposit as we already know that they are going to say no.

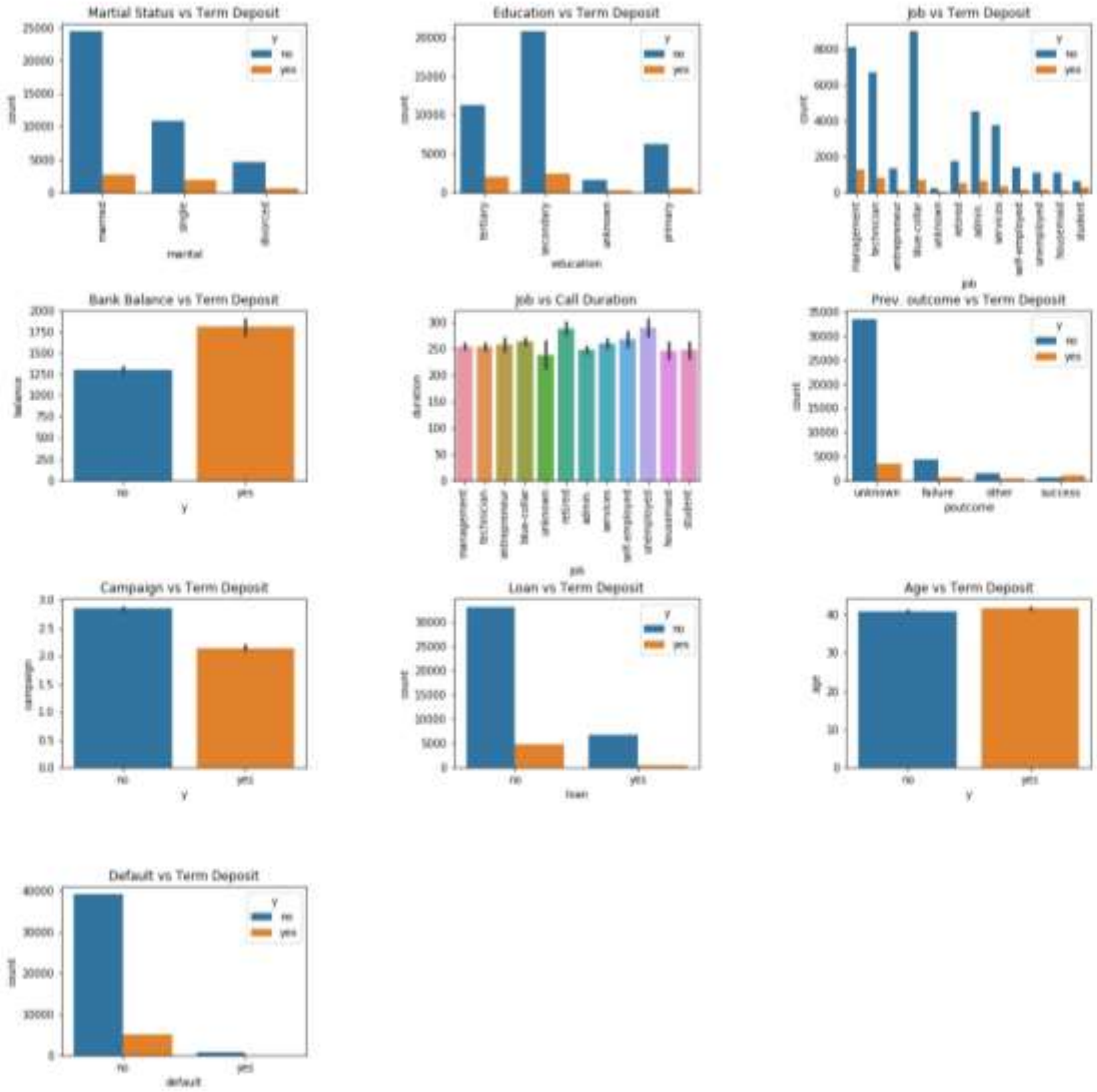


Figure 1. Relationship between the features(attributes)

Data Modeling

In Data modelling, machine learning models are used to infer the predictions based on the dataset provided to it. This model adopts classification or clustering algorithms for predicting or classifying class labels. And for this project, we have considered two classification models, K-nearest neighbour classifier and Decision tree classifier. We used samples from UCI with bank marketing dataset that contains marketing campaign information for term deposit and the label that indicates yes or no based on the outcome. The dataset contains about 45211 records, while each record represents an instance with 17 attributes and the class attribute with two values: Yes and No. The classes are distributed as 88% of the total records for "No" and 12% for "Yes" class.

Experiment

The data set was transformed to cater to the needs of the classification algorithm. This task was done by splitting the categorical values into separate binary features of its own class type. For example, the month feature was converted from one column to 12 columns, with each of the 12 columns containing binary values. This process was carried was out for job, marital status, education, contact, month and poutcome. The numerical columns were left as is since there are already in a suitable format required by the models. After the process of converting categorical features to boolean features, the number of attributes increased from 17 columns to 39 columns however, there were no missing observations. Among the 39 columns was the class label, which we had taken care by storing the feature in another data frame for carrying out the train test split.

K-nearest Neighbour Classification

A k-nearest classification was carried out on the dataset. It was tested with various test sizes such as 80:20, 50:50, 60:40. Using trial and error approach we found out that a k value of 7 gave the best precision, recall and f1 scores. Classification was carried out using a p value of 2 (Euclidean Distance), the algorithm was set auto, the weights were set to uniform and the default weights (Uniform) was used. From carrying out the classification, we found out that a test size 80:20 provided the best results when it comes to recall, precision and f1 scores.

Decision Tree Classifier

A decision tree is a classifier expressed as a recursive partition of the input space based on the values of the attributes. Each internal node splits the instance space into two or more sub-spaces according to certain function of the input attribute values. Each leaf is assigned to one class that represents the most appropriate or frequent target value.

Instances are classified by traversing the tree from the root node down to a leaf according to the outcome of the test nodes along this path. Each path can be transformed then into a rule by joining the tests along this path. For example, one of the paths in Fig. 1 can be transformed into the rule: “If Outlook is Sunny and Humidity is Normal then we can play tennis”. The resulting rules are used to explain or understand the system well.

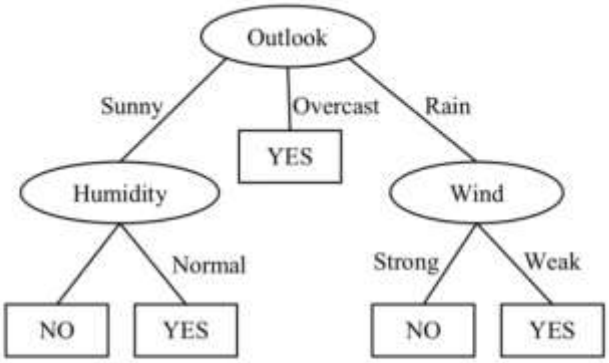


Figure 2. Decision tree example

Fig. 2 shows an example for a simple decision tree for “Play Tennis” classification. It simply decides whether to play tennis or not (i.e. classes are Yes or No) based on three weather attributes which are outlook, wind and humidity [1].

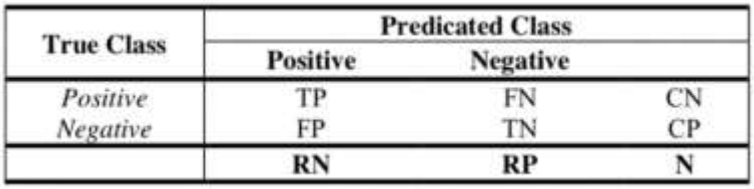
Based on the understanding of the algorithm, we conduct the experiment with our dataset. The initial run for the model was performed with no parameters provided. we could see that the accuracy score was matching the k-nearest neighbour accuracy score however, there was an over-fitting issue in the model. This problem was approached by using the combination of parameters to set constraints on the tree so that size shrinks. The final combination which proved the best among the set of attributes was to choose criterion as entropy and restrict the maximum number of leaf nodes to 10. The reason for choosing criterion as entropy is because the class labels are categorical, and it provides better efficiency with categorical variables when compared to the Gini index, which is intended for numerical variables.

Results

The outcomes from the models were recorded with various parameters set. This data was compared with each other to find the best model. The scores were evaluated on the basis of the confusion matrix, precision, recall, and f1-score. And in order to test the model, k-fold cross-validation technique was adopted. The testing was executed with 20%,40% and 50% as the test sizes. Among the three, there was a better outcome with test size being 50%.

Confusion matrix

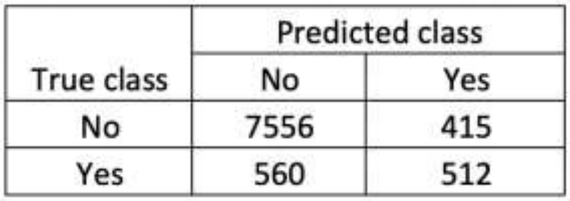
The output data of a classification model are the counts of correct and incorrect instances with respect to their previously known class. These counts are plotted in the confusion matrix as shown in table 1.



True Class	**Predicated Class**		
	Positive	**Negative**	
Positive	TP	FN	CN
Negative	FP	TN	CP
	RN	**RP**	**N**

Table 2 concept of confusion matrix

In table 2, TP (True Positives) is the number of instances that correctly predicted as positive class. FP (False Positives) represents instances predicted as positive while their true class is negative. The same applies for TN (True Negatives) and FN (False Negatives). The row totals, CN and CP, represent the number of true negative and positive instances and the column totals, RN and RP, are the number of predicted negative and positive instances respectively. Finally, N is the total number of instances in the dataset. From the Table 23 we can infer that there were 512 true positives, 7556 true negatives, 415 false positives and 560 in the testing set of data.



	Predicted class	
True class	No	Yes
No	7556	415
Yes	560	512

Table 3 confusion matrix for the bank marketing data set

Precision

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. In this experiment we get the precision as 53% for class label yes. It is calculated as shown below.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall

Recall is the ratio of correctly predicted positive observations to all the observations in actual class. In our experiment, we got the reading to be 49% for class label yes and 94% for class label no. It is calculated as shown below.

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Classification error rate

Error rate is the total number of wrong predictions out of all possible predictions. In this experiment, we observed an error rate of 12 % which is nominal for the current data set.

f1-score

F1-score is the harmonic mean of precision and recall which in our dataset was 50% for class label “yes” and 94% for class label “no”.

Discussion

In the figure below, we have the decision tree diagram for the bank marketing data set with test size. The algorithm initially splits the data based on the call duration, which is the root node. And since we have the maximum leaf node parameter set, there are 10 leaf nodes. The criterion parameter was set as entropy, reason being the target variable is categorical. We could observe that the leaf nodes have low entropy, meaning the classes are pure and homogenous. However, there are leaf nodes with high entropy due to the skewness in the target data. This issue can be corrected by either by making changes to the data or changes to the algorithm or incorporating the combination of both [3] which is further enhancement task of this experiment. And also, we observed that the accuracy score, precision, recall, and f1-scores was minutely better by 1% for test size 20% from that test size 50% scores. But we choose to keep 50% as the test size for the model by considering more number of leaf nodes having low entropy values(50% in this case).

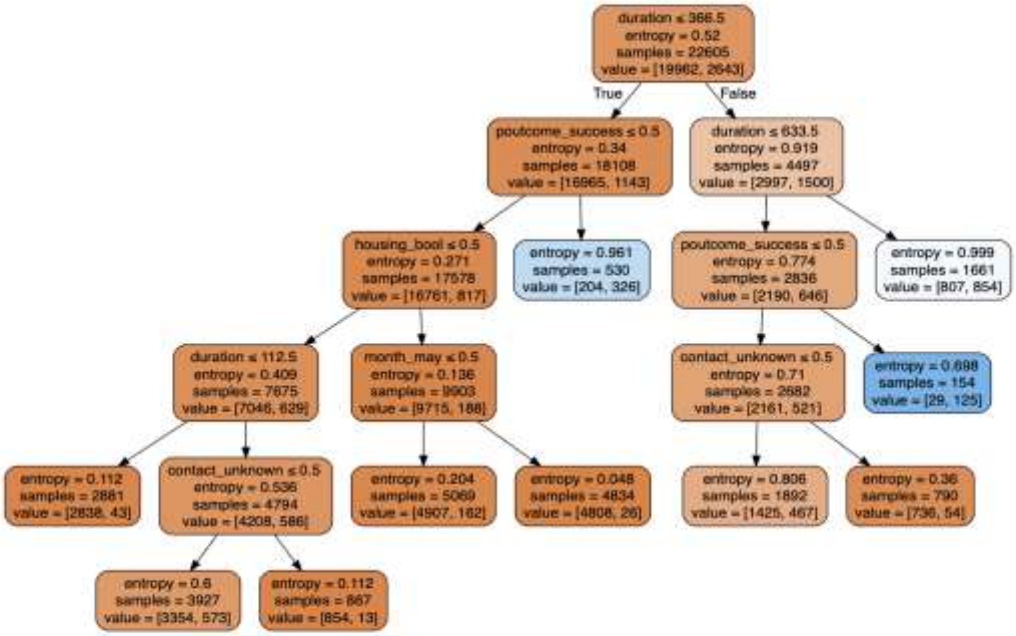


Figure 3 Decision tree for bank marketing data set.

Conclusion

The experiment is carried out to find out whether the customer would subscribe to term deposit or not. We start with exploring each of the features and the relationships associated with them. Following this, we run the models by incorporating classification algorithms like k-nearest neighbour and decision tree algorithms. We test the models using the k-folds validation method and also varying the parameters to the algorithms and also the thresholds of the same. In this process, we make note all the results and decide to choose 50% of the dataset to be the test size for the decision tree algorithm with maximum leaf nodes as 10 and criterion as entropy. This combination provided the best fit among the models. The model predicts that customers whose talking duration is more than 5 min are most likely to be interested regardless of their previous outcomes. However, the model can be fine-tuned further to provide higher recall rates and much more lower entropy values in the leaf nodes, which is the future scope of this experiment.

References

[1] T. Mitchel, (1997), Machine Learning, USA, McGraw Hill.

[2] https://www-users.cs.umn.edu/~kumar001/dmbook/ch4.pdf

[3] https://link.springer.com/article/10.1007/s13748-016-0094-0