Project Report

Topic: Term Deposit Subscription Prediction

Chintan Thakkar - cdt303
Kushal Shah - ks5574

1) Problem Motivated with appropriate background source :

The term deposits are of prime interest for banks, since they can invest those funds in higher gain investments to make profits. The fact that a client investing in one plan at a bank is likely to invest in other attractive plans offered by the same bank, motivates the banks to persuade these clients to purchase the term deposit first and then offer them other promotional plans, and thus invest their marketing resources wisely.

We learn about the background by considering the Banking Term Deposit Dataset available in the UCI Dataset repository and the explanation given for all of it's features. Especially, we would like to explore details about economy metrics like employment variation rate, consumer price index, consumer confidence index and see how they impact the person's decision to invest in term plans.

2) Target Variable:

The target variable is 'y', which indicates whether the customer has subscribed for the Term Deposit Plan or not in the previous campaign.

Predictor variables:

VARIABLE	**DESCRIPTION**	**EXAMPLES**
age	age of client	numeric
job	type of job	categorical
marital	marital status	categorical
education	level of education	categorical
default	has credit in default?	categorical
housing	has housing loan?	categorical
loan	has personal loan?	categorical

Data related to last contact in the current campaign:

VARIABLE	DESCRIPTION	EXAMPLES
contact	contact communication type	categorical
month	last contact month of year	categorical'
day_of_week	last contact day of the week	categorical
duration	last contact duration, in seconds	numeric

Other Attributes:

VARIABLE	DESCRIPTION	EXAMPLES
campaign	number of contacts performed during this campaign and for this client	numeric
pdays	number of days that passed by after the client was last contacted from a previous campaign	numeric
previous:	number of contacts performed before this campaign and for this client	numeric
poutcome	outcome of the previous marketing campaign	categorical

Socio-Economic Attributes:

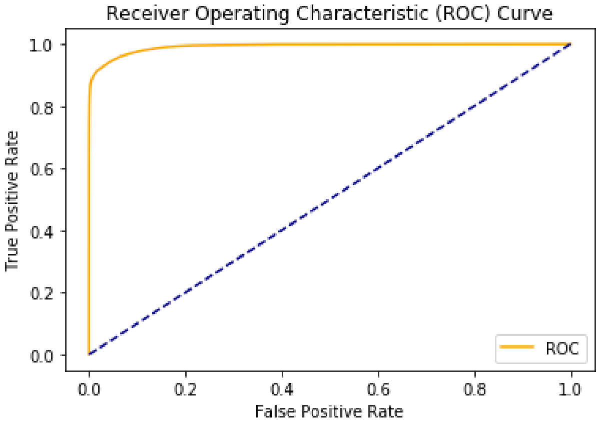
VARIABLE	DESCRIPTION	EXAMPLES
emp.var.rate	employment variation rate - quarterly indicator	numeric
cons.price.idx	consumer price index - monthly indicator	numeric
cons.conf.idx	consumer confidence index - monthly indicator	numeric
euribor3m	euribor 3 month rate - daily indicator	numeric
nr.employed	number of employees - quarterly indicator	numeric

3) Problem Statement:

The dataset represents the data of clients of a bank from their marketing campaign, and based on client's background, the bank would like to know whether the client would subscribe to a term deposit plan. Our goal is to predict if the client will subscribe to the Term Deposit plan or not, this can be achieved by classification.

4)Type of Model motivated:

Since the problem outcome required a binary classification, and Random Forest combines results of multiple smaller sub decision trees for optimal performance, we decided to test Random Forest algorithm and decision trees separately as well. Also, since data points distributed close to each other may tend to belong to the same target class, we decided to test KNN model. And since SVM can partition the 2 classes well at times by maximizing the margin around separation line, we wanted to evaluate it as well. So, to summarize we have used classification algorithms such as Random Forest Classification, K-nearest neighbors, Logistic Regression, Naives bayes' and SVM. We have tried the modelling using multiple kernel types in SVM, and observed that Linear performed than RBF kernel. For Random Forest Classification we used different values for n_estimators and observed that values above 25 performed quite well on the prediction accuracy.. We have also used K-fold Cross Validation to assess the performance of the above classification models and found that Random Forest Classification has the highest accuracy amongst all.



The above figure is the ROC Curve for Random Forest Classification with the ROC-AUC score as 0.992. This Model has the performance Accuracy of 95%, f1 Score: 0.949, Precision: 0.966, Recall: 0.933

5) Evaluation Approach:

We evaluated the performance of the models by assessing their prediction accuracy, precision, recall, F1-score and confusion matrix. Using these metrics, we plotted the ROC-AUC curve to get more clarity on the performance. Also, we ran k-fold cross validation on all the models under test and compared the average performance score over k-folds for each of the model for a comparative analysis.

6) Assumptions and Limitations:

We assume that there are no additional attributes affecting the outcome apart from the predictor attributes present in the dataset.

7) Problem in Scope of class:

Since the target variable is binary, there was no such problem faced in this dataset regarding the scope of the class.

8) Changes from original proposal:

We had decided to scale down numeric features such as cons.price.idx, cons.conf.idx, euribor and nr.employed. But we instead had to drop 3 of these 4 features because of high correlation. Also cons.conf.idx had values ranging from -50.8 to -26.9, so it did not require scaling down to a huge extent. We have just scaled the feature “duration” as it ranges from 0 to 4918 seconds. Rest of the things went pretty well as per our plan.

9) Figures:

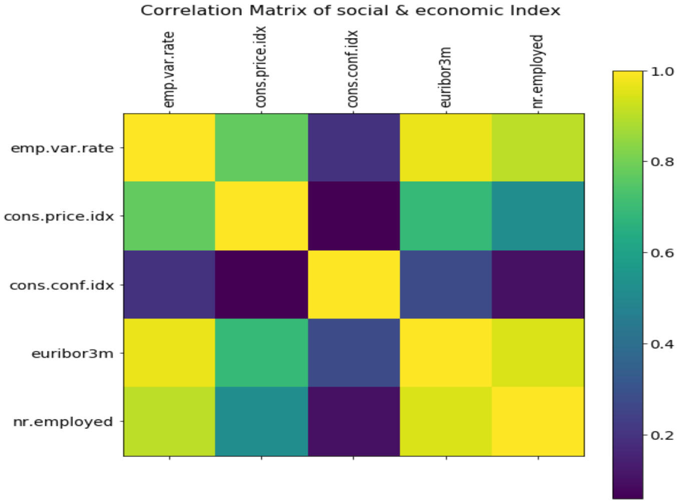
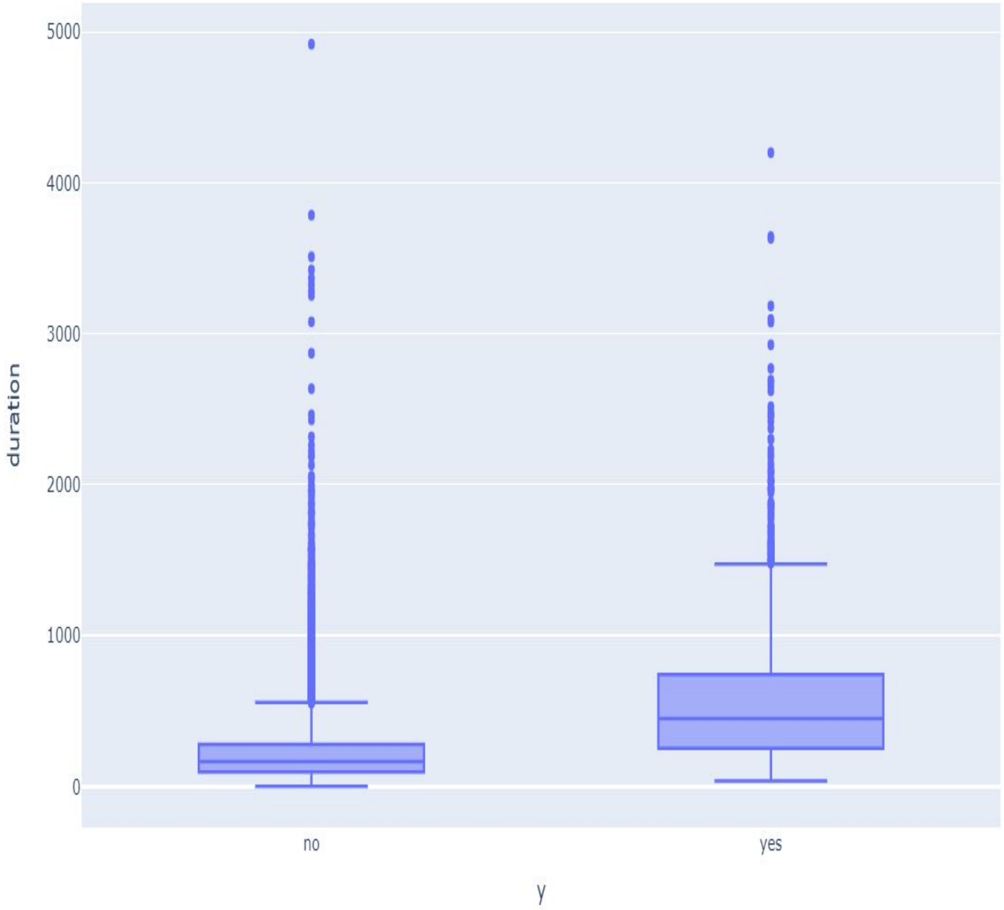


Fig 9.1

Impact of Call Duration on Outcome



As seen from the above graph, the average duration of call is approximately 7.5 minutes for the customers who subscribed to the Term Deposit, whereas it is 2.5 minutes for the customers who didn’t subscribe, and this correlates to real world scenario that the customers tend to talk longer when they are interested in a service, else they hang up the phone shortly.

10) Team Evaluation:

Net ID	Score
cdt303	4
ks5574	4