Project Report

On

Bank Marketing

By : Yatish Anvekar

Background & Introduction

- The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

Buisness Problem

- The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

- There are two datasets: 1) bank-full.csv with all examples, ordered by date (from May 2008 to November 2010). 2) bank.csv with 10% of the examples (4521), randomly selected from bank-full.csv. The smallest dataset is provided to test more computationally demanding machine learning algorithms (e.g. SVM).

- The classification goal is to predict if the client will subscribe a term deposit (variable y).
- Number of Instances: 45211 for bank-full.csv (4521 for bank.csv)
- Number of Attributes: 16 + output attribute.
- Attribute information:
- Output variable (desired target): 17 - y - has the client subscribed a term deposit? (binary: "yes","no

Approch to Data Cleaning

- First we check the information of given dataset as it tells us how many rows and columns are present and what aree the datatypeswhether they are object int or float
- then we check for nullvalues present
- After that we check the summary statistics .this part will tells aboutthe statistics of the dataset i.e mean median max value min values and also it will tell if there are any outlier present or not
- We also check the correlation of our dataset to check correlation ofcolumns with each other.
- There are numeric and categorical features
- I convert the categorical value to numeric using mapping and label encoding
- After converting in int I move to model building
- I check the outliers and then remove the outlier using zscore
- I checked for skewness and remove it
- Also check the distribution of output
- We delete the contact column

Visualization

- We plot correlation matrix via heatmap to see the correlationOf the columns with each other
- We also visualize the column viabar count plot
- We see the number of defaulters and non defaulters visa countplot
- We plot histogram to display the shape and spread of continuoussample data
- We also see the distribution of data and skewness
- We check the outliers using boxplot

Modelling part

- We know that this is classification problem so we use accuracy score ,classification report and confusion matrix as our evaluation matrix
- We also see the AUC score and also plot the AUC_ROC curve forour final model
- We see the precesion and recall value along with f1 score
- First we see the result
- I use logisticregression with cross validation and hyperparamter tuning
- Used cross validation for various model
- We also use random forest classifier as our evaluation model without usinghyperparametertuningas my computer got hanged

Conclusion

- Our dataset consist of categorical and numeric features
- The month of may have the highest number of clients
- The least number of clients are in the dec month which says the end of the year there are less clients
- as we can see there are more number of clients btw the age of 30-60
- there are less number of defaulter
- Blue collor and management job people clients are more in number
- Married clients are more
- different model have been trained and tested ,out of which i am going with randomforestclssifier