UC Berkeley Practical Application #2

Data Interpretation

- A Portuguese Bank did a telephone marketing campaign to attract customers to set up a long-term deposit account with good interest rate
- A total of 17 campaigns occurred between May 2008 and Nov 2010

Data Analysis and Clean Up Procedure

- Check for Missing values and delete\replace the missing values depending on what seems to fit the situation
    - There are no missing values in this dataset
- Perform Univariate and Bivariate Analysis of the existing features
- Check for data imbalance
- Understand datatype of each feature. Convert Categorical data to numerical data
- Check the unique values in all categorical features. Value 'unknown' found in many categorical features. Exclude them from analysis. More than 30000 rows still available for analysis even after removing the rows with value 'unknown'
- Categorical Feature poutcome has more than 35000 rows with value 'non existent'. It does not add much value to the analysis. Remove it from analysis
- The following features default, housing, loan and y have binary data no and yes. Using mapping to convert them to 0 and 1 respectively.
- For feature 'contact' there are two values, telephone and cellular. Check how many are from telephone compared to cellular. One third of the data is contacted via telephone. Include both in analysis

Data Imbalance Analysis

- Data below shows that from the output data most of the customers did not choose to open the account
- The data is imbalanced with 88.74% of the clients saying No

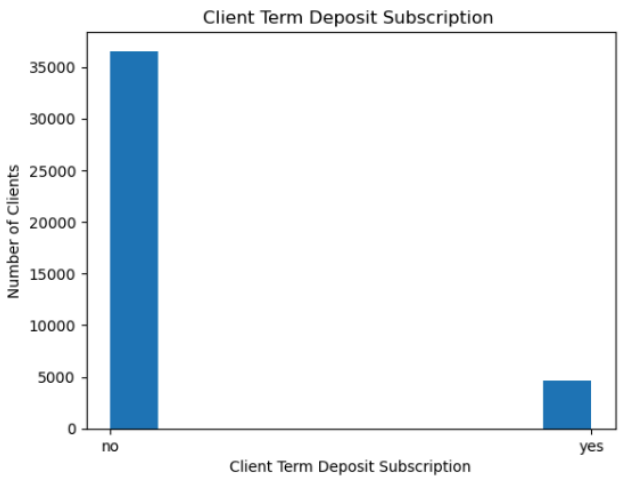


Figure 1: Distribution of the Client Term Deposit Subscription. Data shows imbalance with 88.74% not opening an account

Categorical Data Analysis

- Univariate Analysis shows

Feature Name	Critical Features
Job	Retired and students more likely to open a term deposit account
Marital Status	Single People are slightly more likely by 3% to open a term deposit account
Education	Illiterate data shows large variance and cannot be used to conclude if it is a critical feature. All other parameters showed similar likelihood of opening a term deposit account
Contact Method	People contacted by cellular phone are more likely by 10% to open a term deposit account
Month	September, October, December and March showed higher likelihood compared to other mother. March showed the maximum rate of opening a term deposit account
Day of Week	Day of week not very relevant. All days shows similar likelihood of customers opening a term deposit account

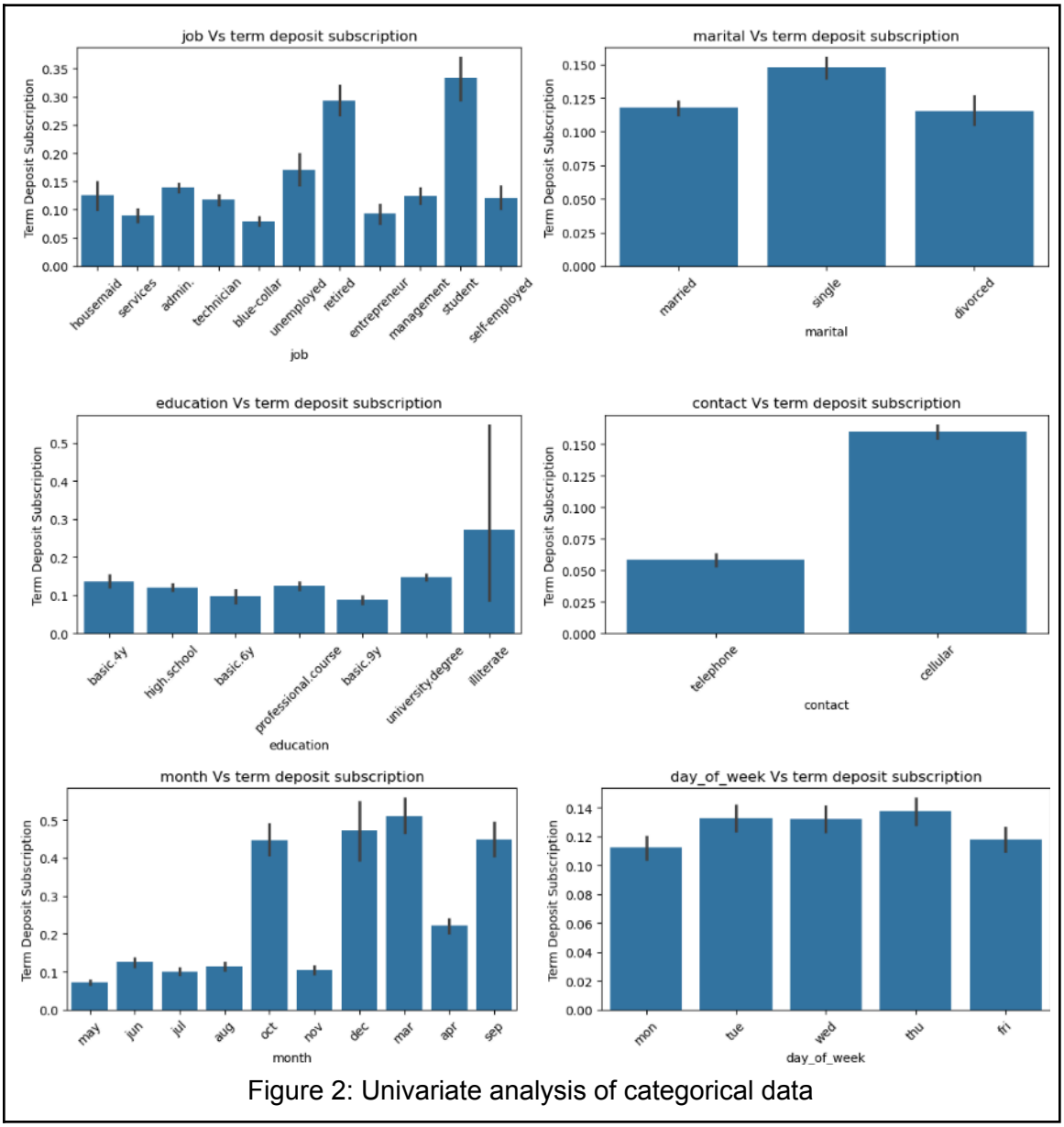


Figure 2: Univariate analysis of categorical data

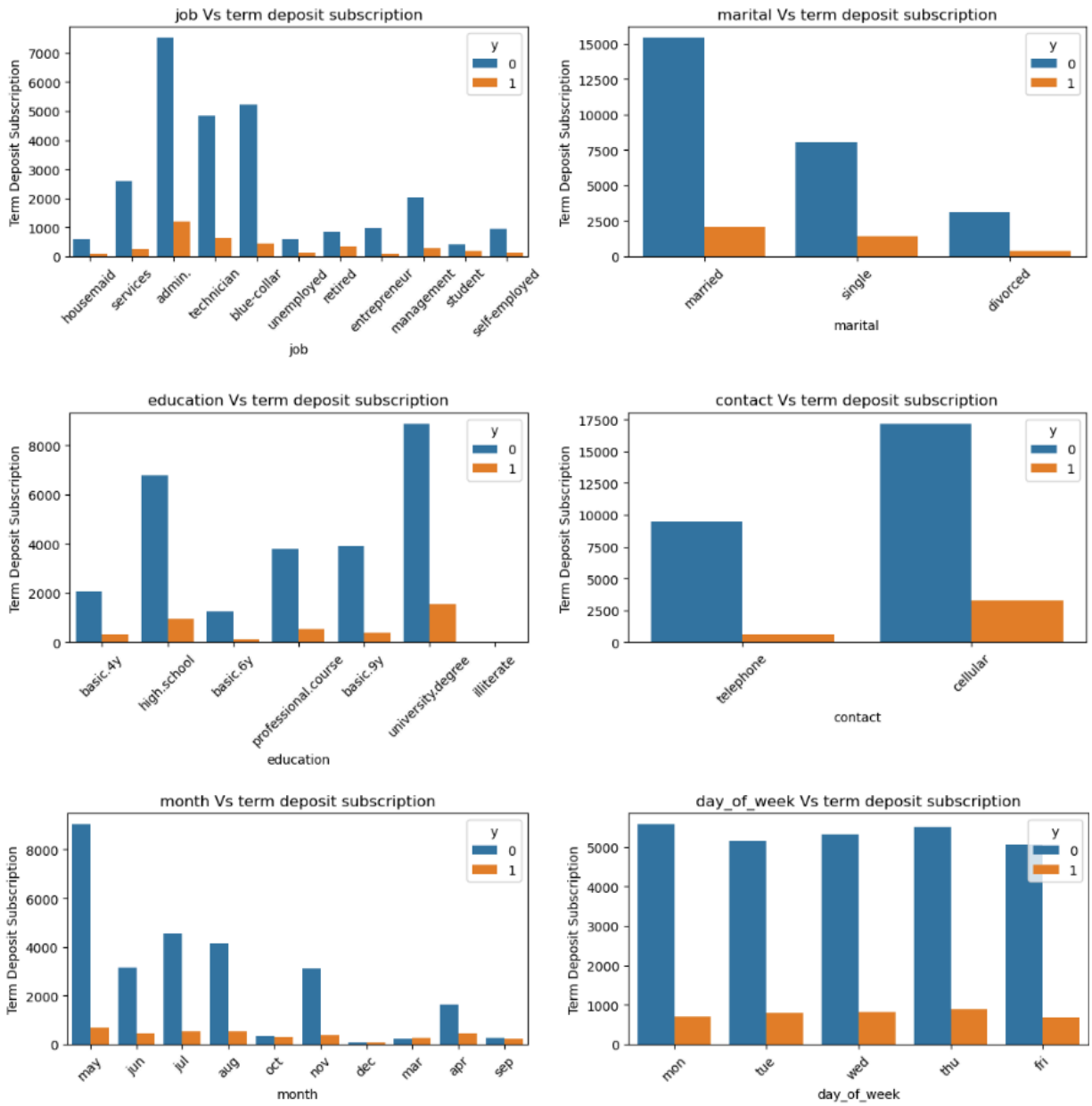


Figure 3: Bar plot of categorical data

Numerical Data Correlation Analysis

The features ‘previous’ shows positive impact on the term subscription with the optimum value at #3. The histogram shows the data imbalance and cannot be conclusive. The same goes for the other parameters too.

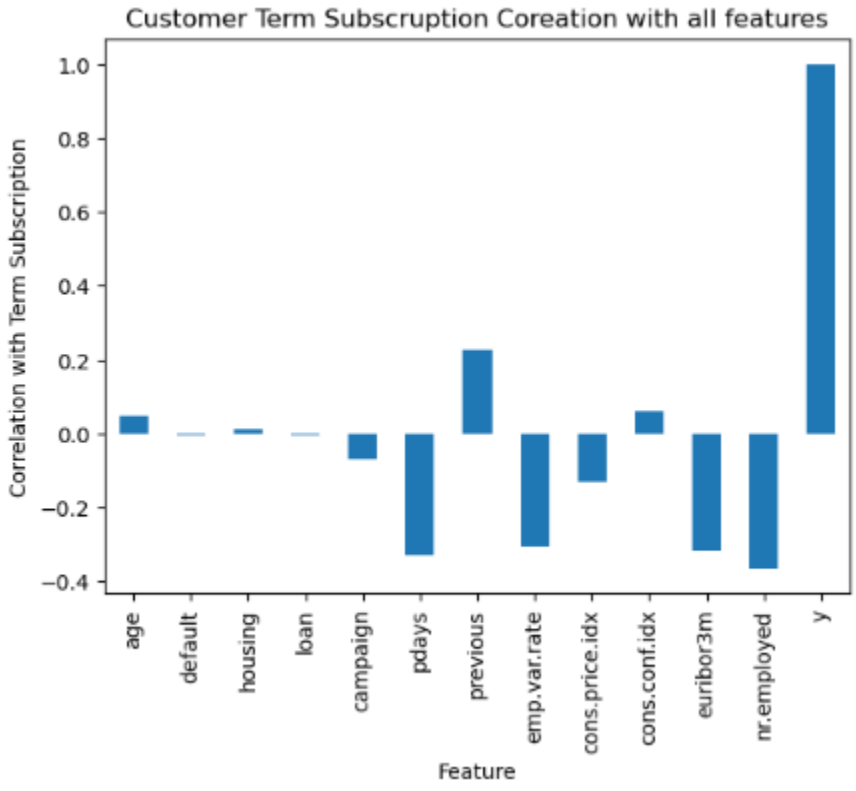


Figure 4: Correlation Matrix

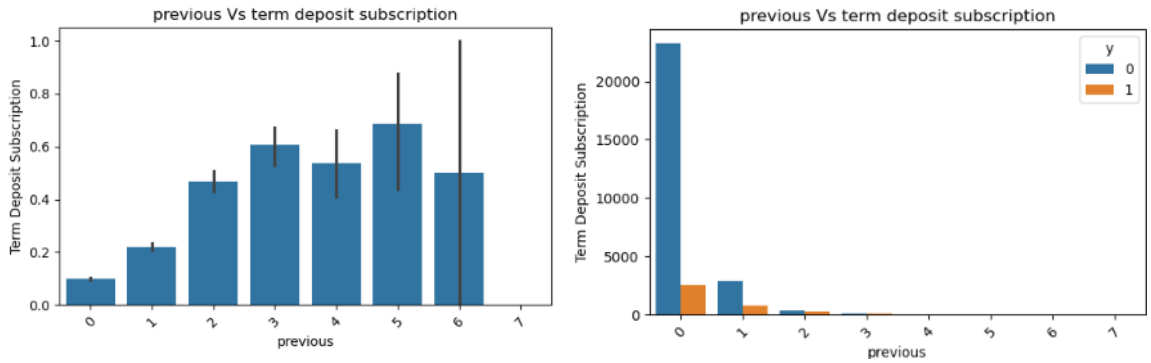


Figure 5: ‘previous’ Feature Vs Term Deposit Subscription

Business Objective

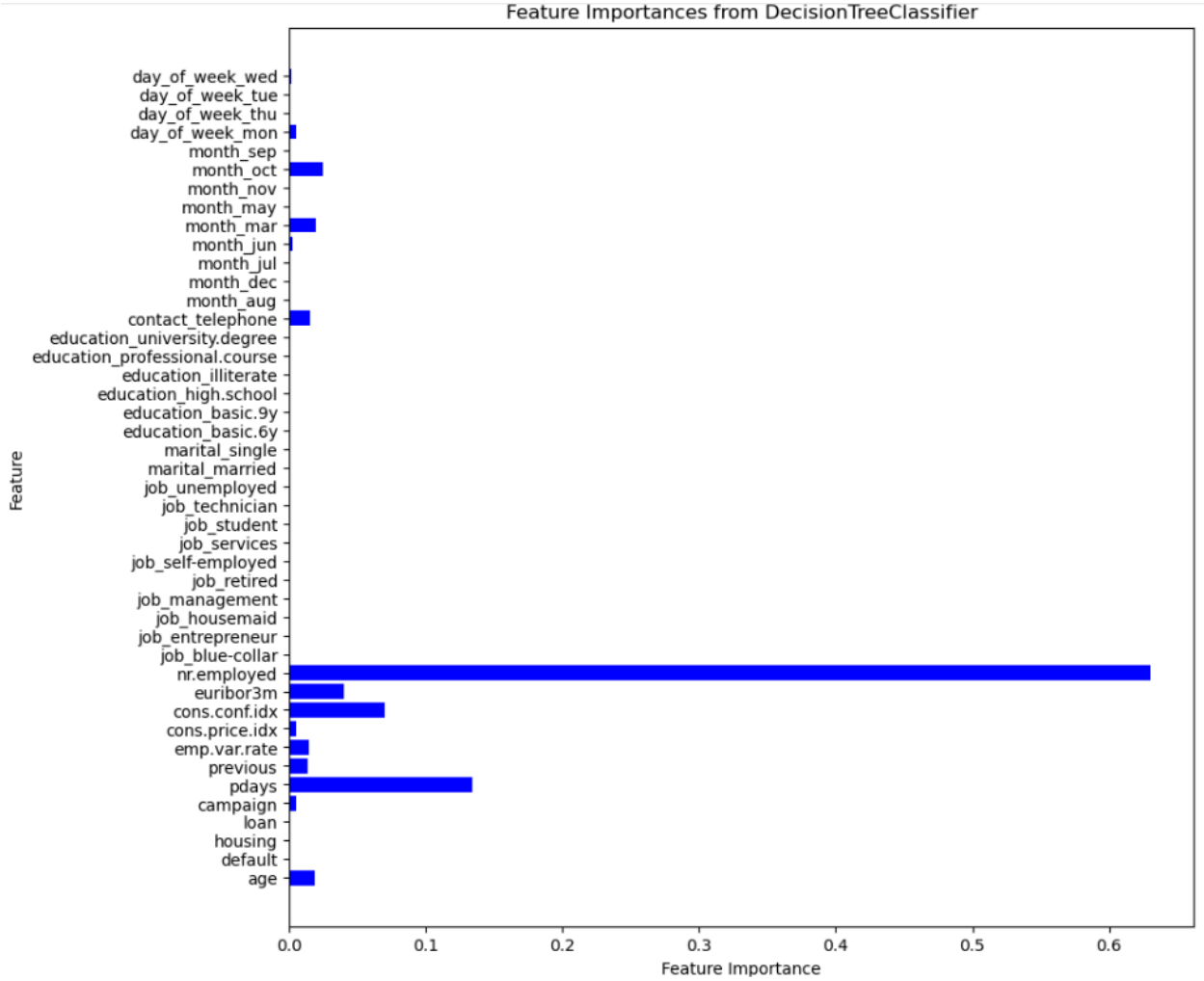
The business goal is to find a model that can explain success of a contact during the marketing campaign, that is if the client subscribes the deposit. Identify what factors used in the marketing can have an effective impact on getting the client to subscribe. The model can improve campaign efficiency by identifying the main characteristics that affect success, helping in better management of the available resources (e.g. human effort, phone calls, time etc.) and selection of buying customers.

Model #1

Baseline Model: Accuracy: 0.8697. This is expected since there is significant imbalance in the data.

Table below shows the Accuracy for the different models and the corresponding parameter. The best test accuracy was obtained for the Decision Tree Classifier with a max depth of 5. The test accuracy of 0.8871 is very close to the baseline model accuracy of 0.8697. The features impacting the decision tree shows the economic conditions and pdays shows strong dependency.

	Model	Parameter	Train time	Train Accuracy	Test Accuracy
0	knn	{'knn__n_neighbors': 7}	1.818759	0.900451	0.876025
1	logisticregression	{'logisticregression__C': 0.1}	0.674568	0.889340	0.884388
2	decisiontreeclassifier	{'decisiontreeclassifier__max_depth': 5}	0.195054	0.892825	0.887176
3	ridgeclassifier	{'ridgeclassifier__alpha': 10}	0.184065	0.889094	0.884880
4	lasso	{'lasso__alpha': 1}	0.160577	0.000000	-0.000182



Model #2

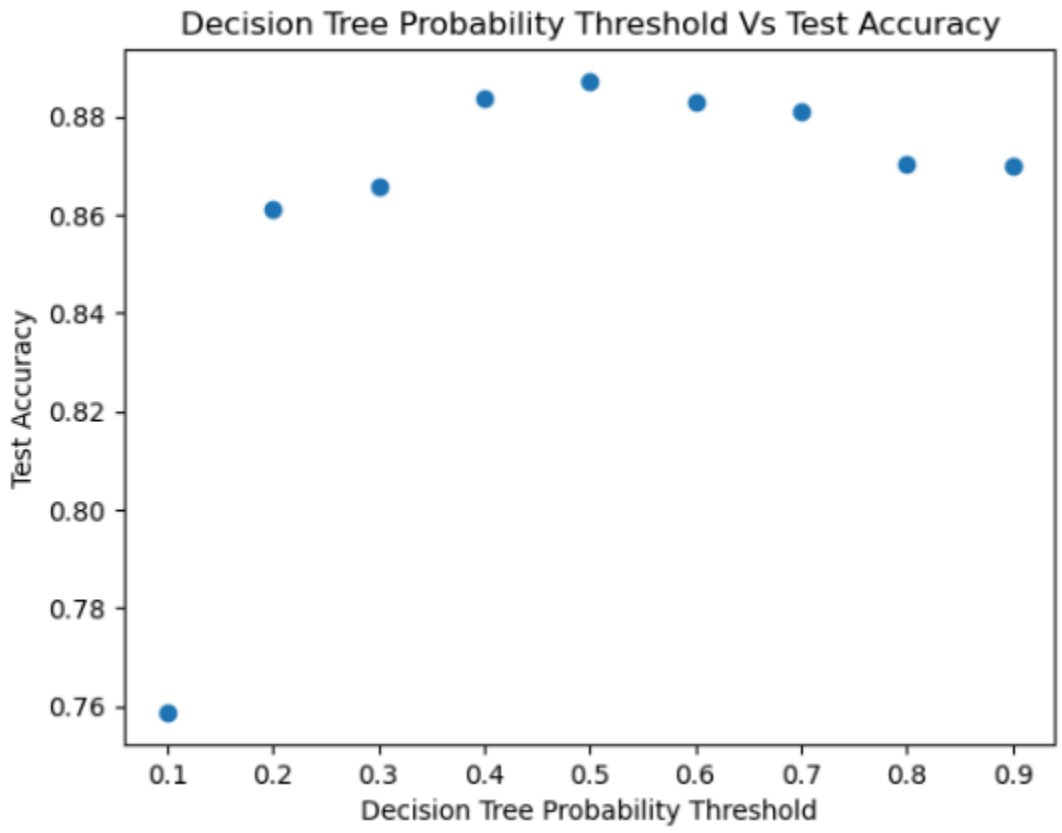
Remove all socio-economic features
- The hypothesis is the general population may not necessarily in sync with the national economy and may not influence personal decision.

Table below shows that there is no significant improvement with this model.

	Model	Parameter	Train time	Train Accuracy	Test Accuracy
0	knn	{'knn__n_neighbors': 7}	3.820498	0.893440	0.874877
1	logisticregression	{'logisticregression__C': 0.1}	0.183254	0.886880	0.880125
2	decisiontreeclassifier	{'decisiontreeclassifier__max_depth': 5}	0.240990	0.889709	0.877829
3	ridgeclassifier	{'ridgeclassifier__alpha': 1}	0.276784	0.886880	0.879469
4	lasso	{'lasso__alpha': 1}	0.210712	0.000000	-0.000182

Decision Tree Probability Threshold Tuning

Decision Tree Probability Threshold tuning shows that the Test Accuracy was optimum at decision threshold of 0.5



Conclusion

Due to the data imbalance the model prediction is close to the baseline prediction. The features impacting the decision tree shows the economic conditions and pdays shows strong dependency.