Lead Scoring

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- When these people fill up a form providing their email address or phone number, they are classified to be a lead
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.
- The company requires you to build a model wherein you need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

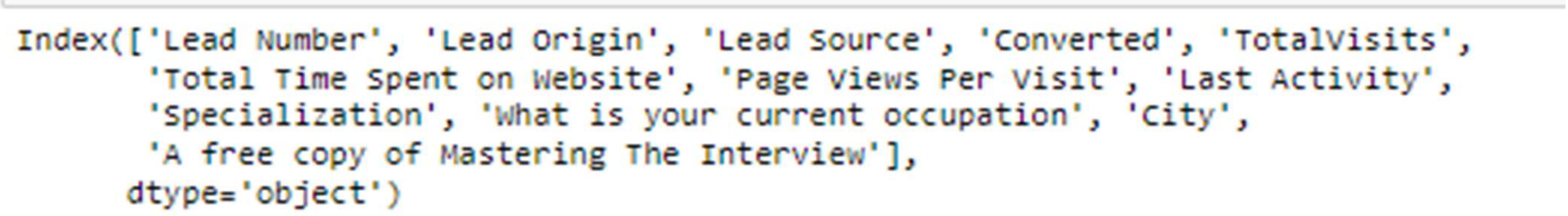
Objective :

Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads

Data Preparation

- Dropped the features which are part of Lead profiling activities like Tags,Lead Profile, Lead Quality, all Asymmetrique scores.
- Dropped all highly skewed variables which are having more than 90% of the entries related to one category
- Page Views Per Visit, Total Visits & Total Time Spent On Website are the only continues variables retained after dropping unnecessary columns.
- In all the below features, there are certain categories which are very less in count. All those are grouped into single category ‘Others.
    - Last Activity
    - Lead Source
    - Lead
    - Specialization
    - City
- In some of the features there were ‘Select’ category which are converted into Null values

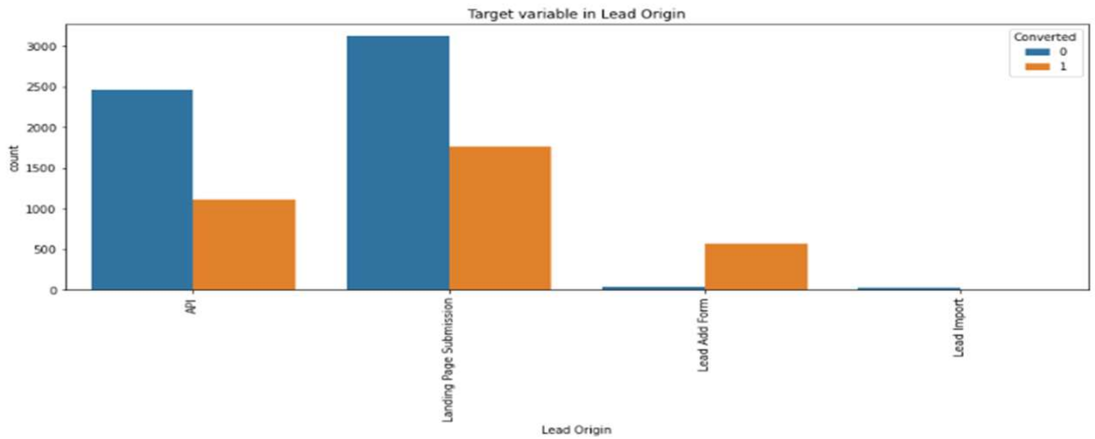
- Below are the Features that will be used for Modelling. All the categorical variables are converted to dummy variables

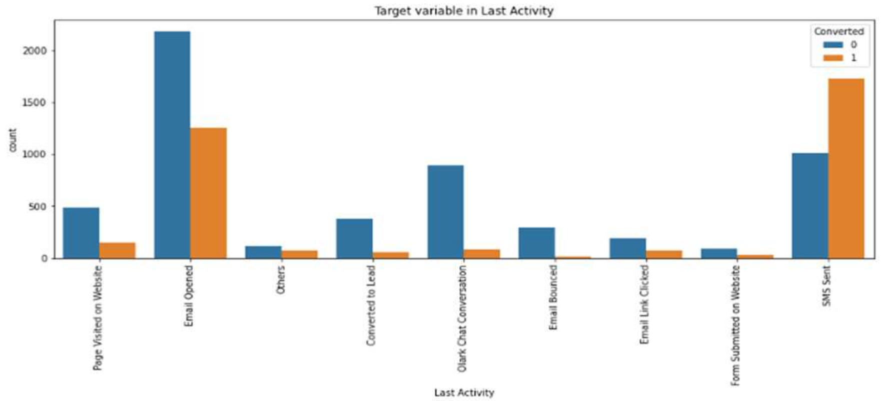


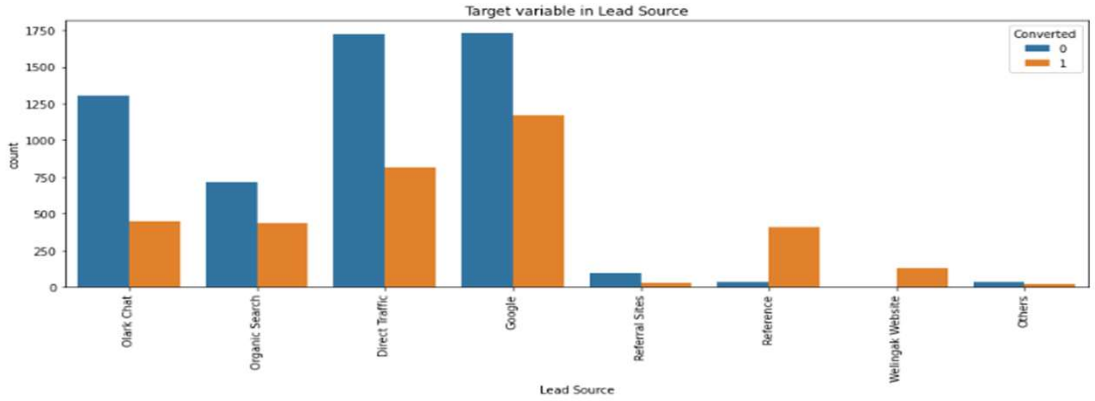
```
Index(['Lead Number', 'Lead Origin', 'Lead Source', 'Converted', 'TotalVisits',
       'Total Time Spent on Website', 'Page Views Per Visit', 'Last Activity',
       'Specialization', 'What is your current occupation', 'City',
       'A free copy of Mastering The Interview'],
      dtype='object')
```

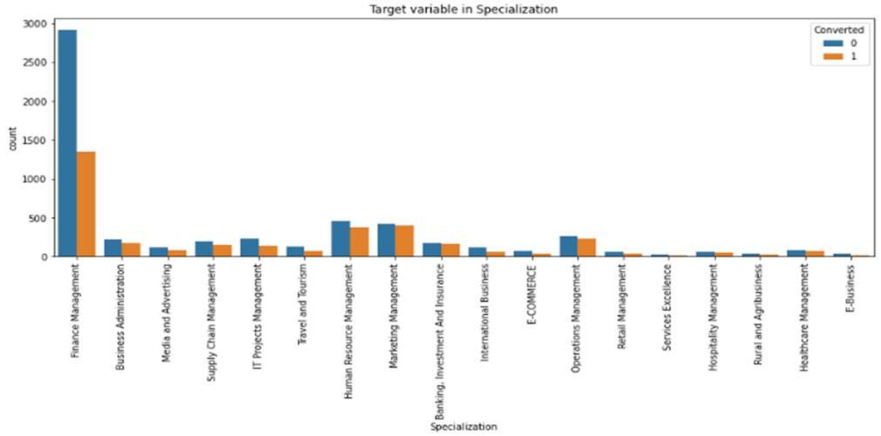
EDA

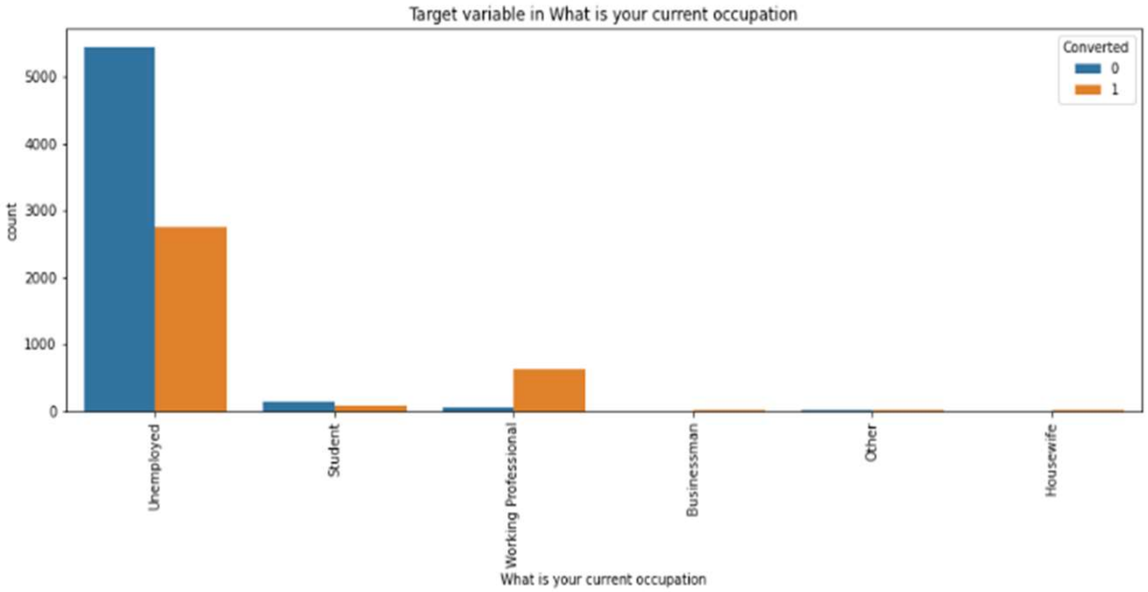
- With these Bivariate analysis of categorial variables, the following observations are evident.

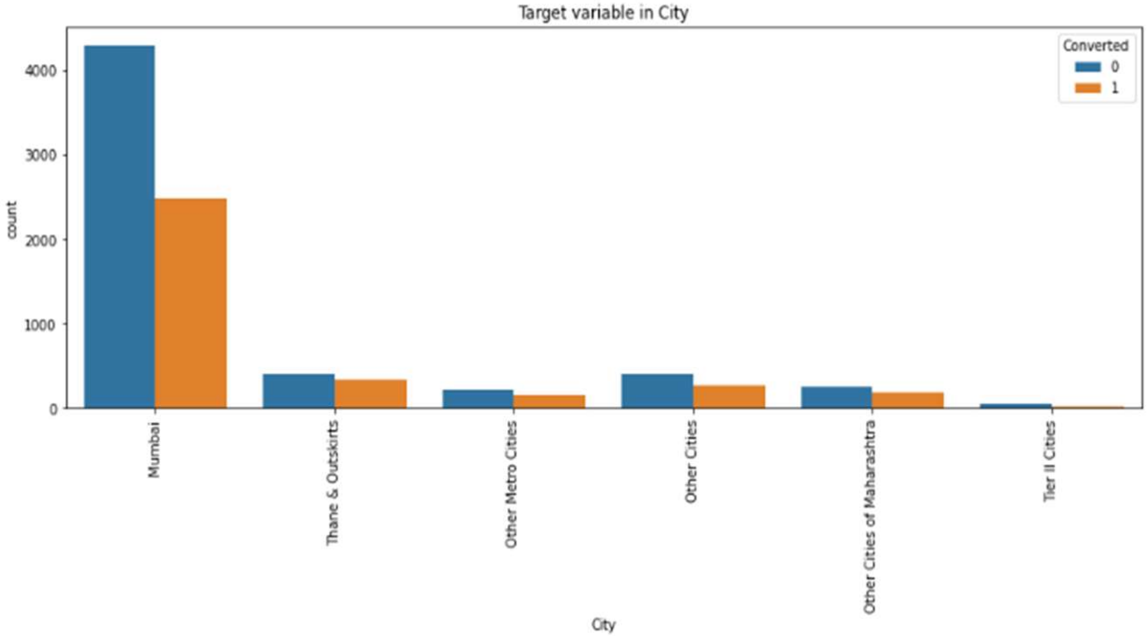

Target variable in Lead Origin


Target variable in Last Activity


Target variable in Lead Source


Target variable in Specialization



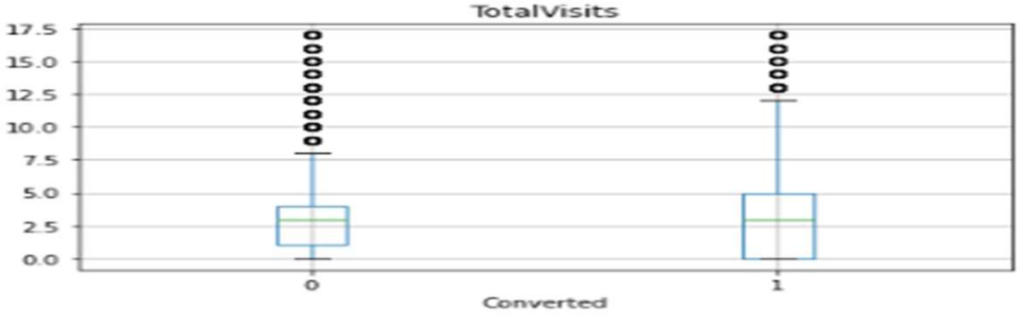


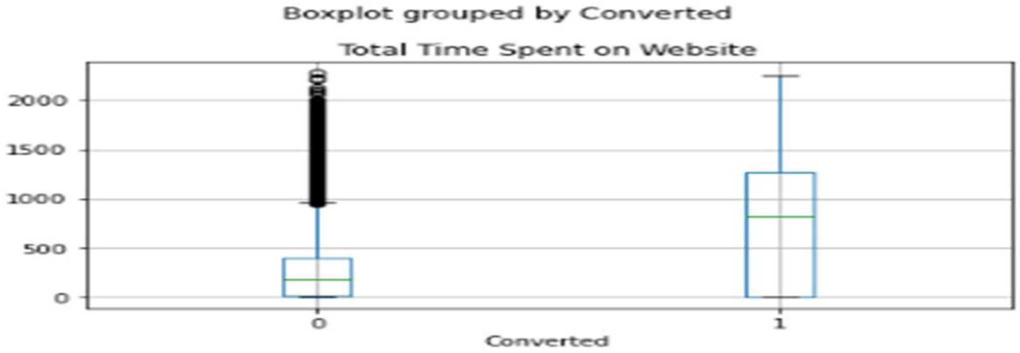
From these analysis, below are observations :

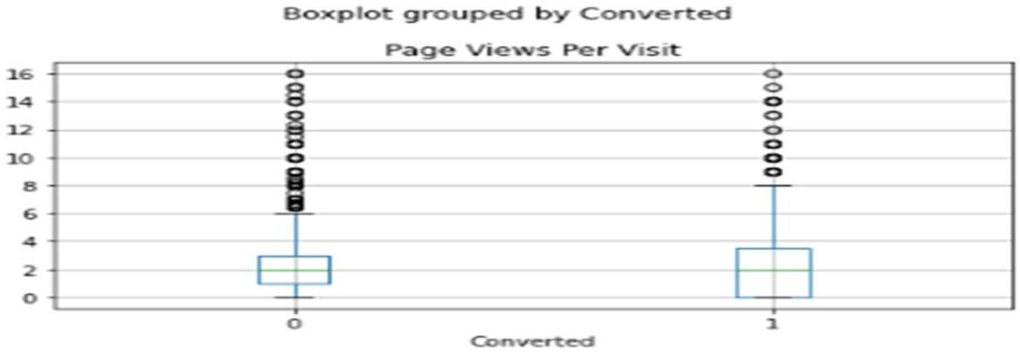
- •Lead origin of 'Lead Add Form' influences more for conversion of the lead
- •'Welingak Website' and 'References' lead sources are reliable for lead conversion
- •'Reference is reliable for a lead conversion
- •'SMS Sent' activity is an higher reliable for a lead to be converted
- •'Finance Management' specialization is not leading for a good conversion rate. Working professional are reliable for lead conversion
- •'Unemployed' candidates as expected not ideal candidates for a lead conversion
- •City is not playing any influence in lead conversion

Final Set of categorical are converted to dummy variables and the highly corealted bariables are dropped

EDA







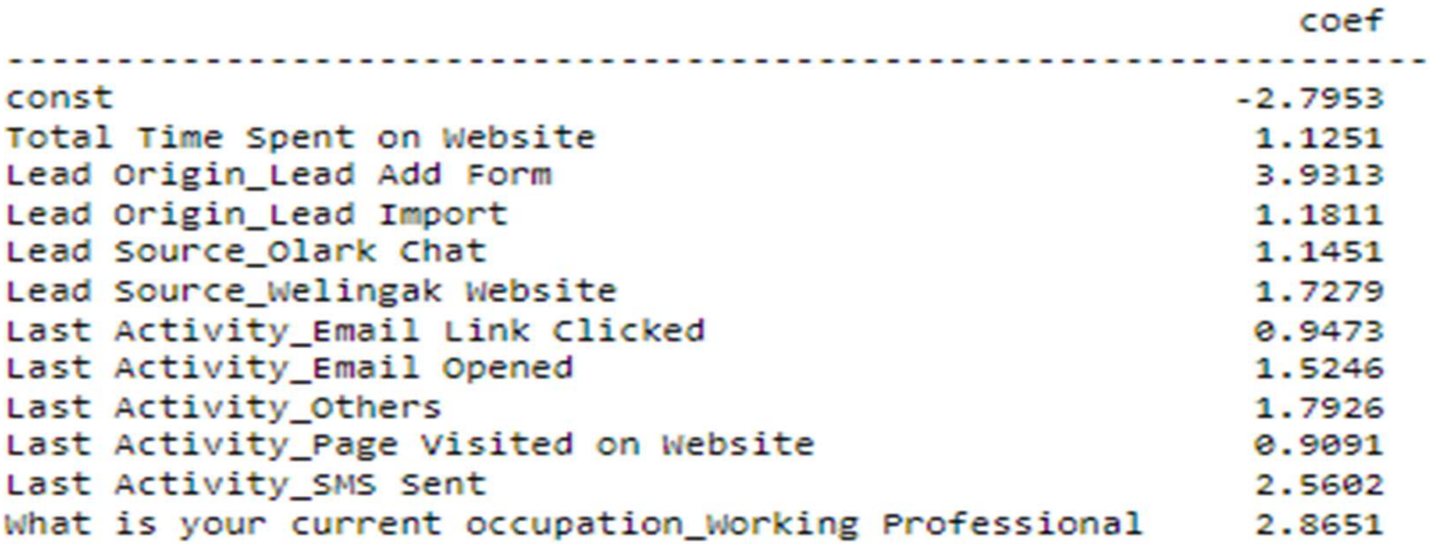
When Bivariate analysis is performed on continuous variables, below are the observations :

•Total Visits are quite higher for the converted deals
•Time Spent on the website is clearly huge for the converted deals
•Pages per visit also quite spread on higher side for the converted deals

Final data frame combining continuous features and dummy variables out of categorical formed with 51 columns, which is directly used for mode building

Model Building

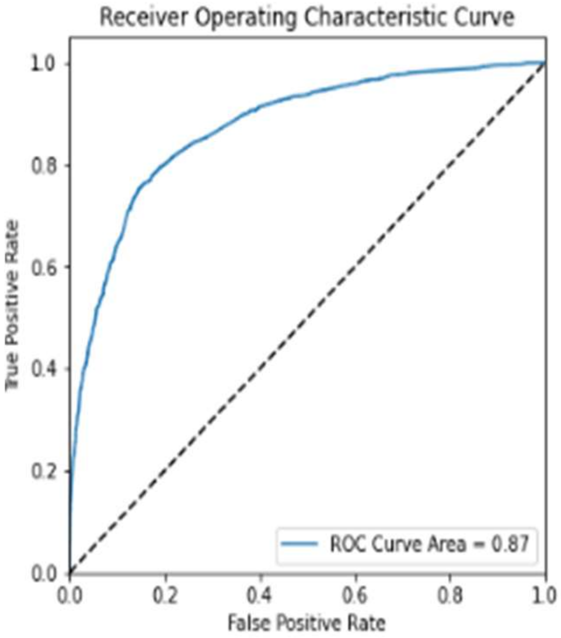
- To bigin with, identified 15 features with the help of RFE method.
- After 5 iterations of eliminating with high p-value and VIF, a final model is achieved with 11 variables and the variables are listed with coefficients :

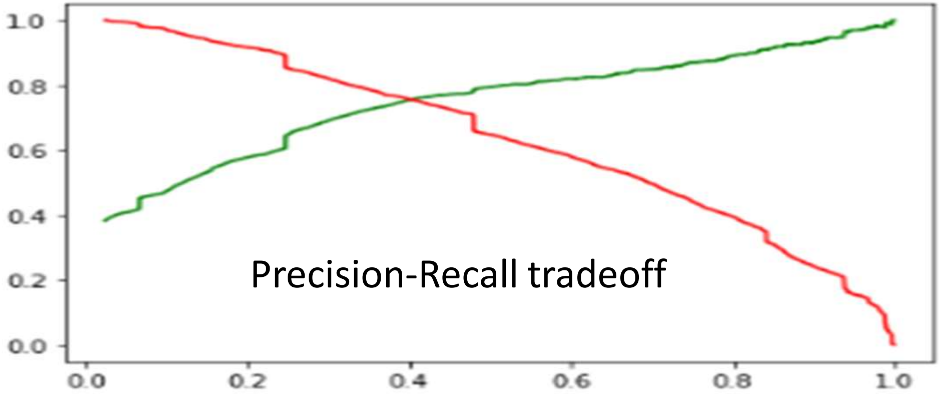


```
                                                      coef
------------------------------------------------------------
const                                              -2.7953
Total Time Spent on Website                         1.1251
Lead Origin_Lead Add Form                           3.9313
Lead Origin_Lead Import                             1.1811
Lead Source_Olark Chat                              1.1451
Lead Source_Welingak Website                        1.7279
Last Activity_Email Link Clicked                    0.9473
Last Activity_Email Opened                          1.5246
Last Activity_Others                                1.7926
Last Activity_Page Visited on Website               0.9091
Last Activity_SMS Sent                              2.5602
What is your current occupation_Working Professional     2.8651
```

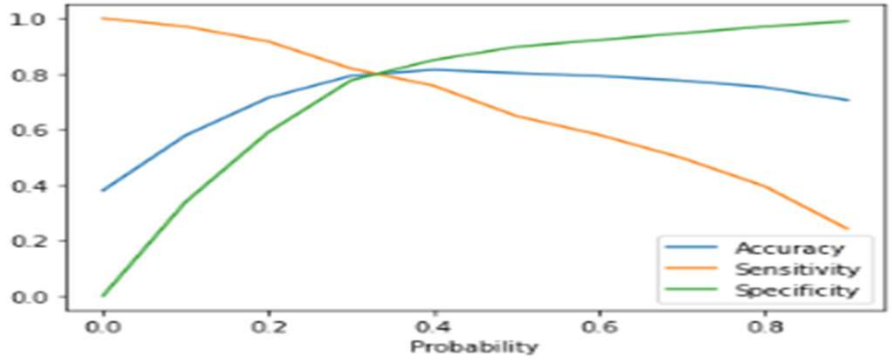
Identify the probability cut-off

- With ROC area of 0.87, the model seems to be good one.
- The Precision_Recall_Curve and Sensitivity_Specificity tradeoff are plotted to identify the optimal cut-off probability so as to keep the metric score good.



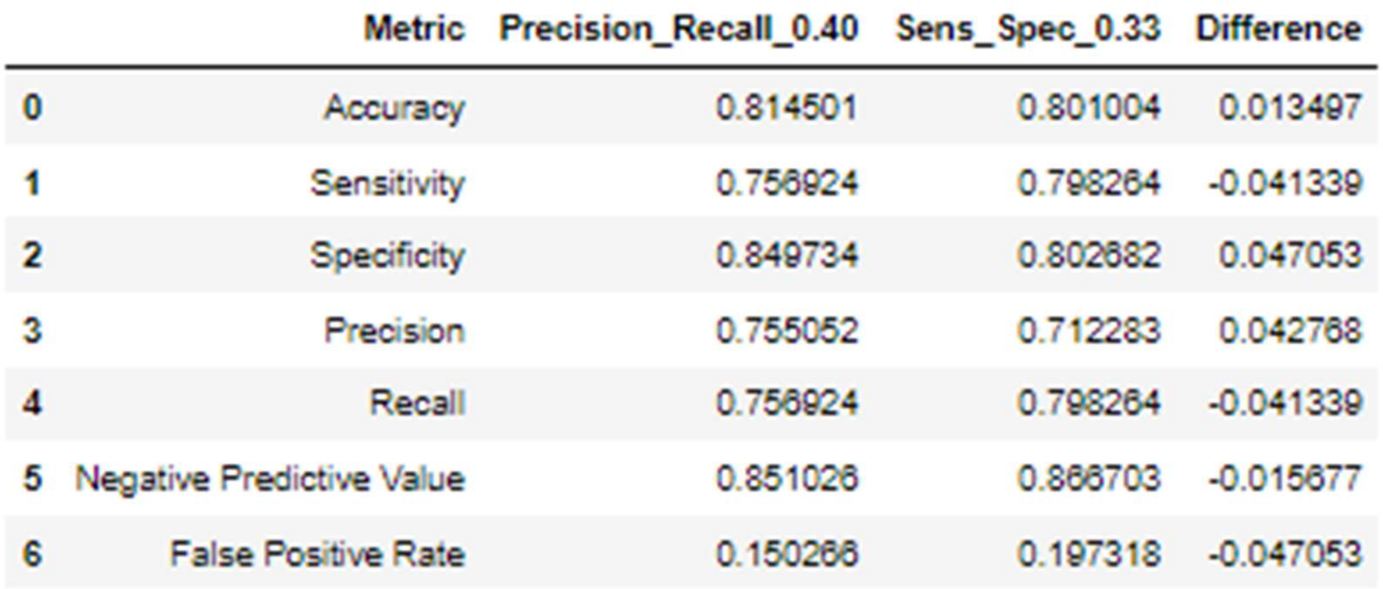


Precision-Recall tradeoff



Identify the probability cut-off

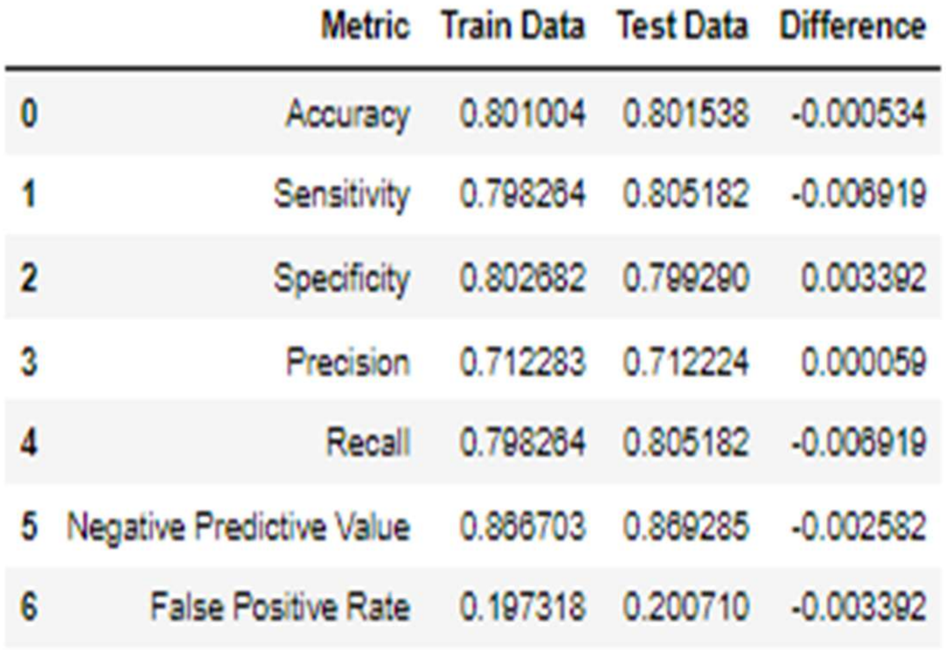
- 0.40 is found to be optimal cut-off probability from Precision_Recall_Curve and 0.33 from Sensitivity_Specificity tradeoff
- Objective of the case study is to identify about 80% potential lead candidates. Hence we should target high Recall and should be around 80%. So we stick to probability cut-off of 0.33



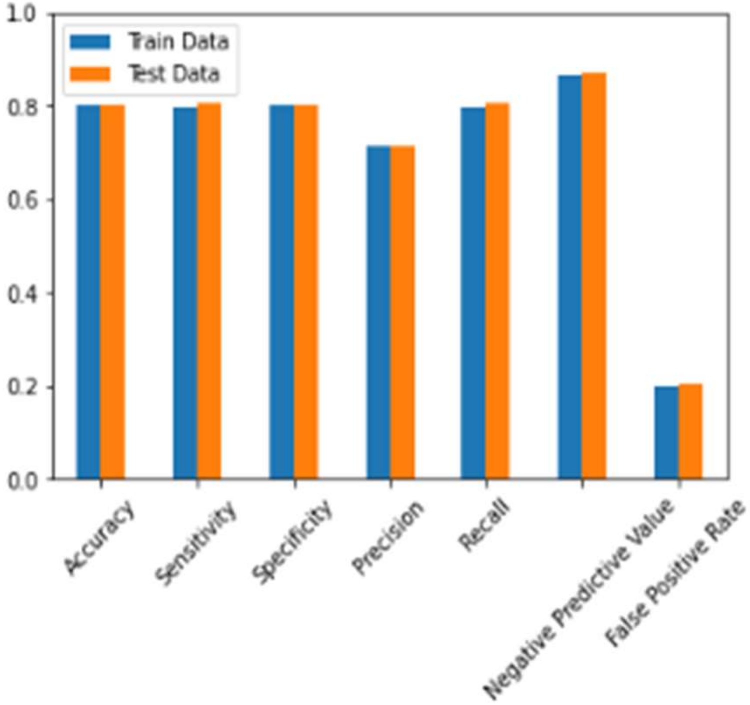
	Metric	Precision_Recall_0.40	Sens_Spec_0.33	Difference
0	Accuracy	0.814501	0.801004	0.013497
1	Sensitivity	0.756924	0.798264	-0.041339
2	Specificity	0.849734	0.802682	0.047053
3	Precision	0.755052	0.712283	0.042768
4	Recall	0.756924	0.798264	-0.041339
5	Negative Predictive Value	0.851026	0.866703	-0.015677
6	False Positive Rate	0.150266	0.197318	-0.047053

Prediction Metric on Test data

- The final model is applied on the test data with the probability cut-off 0.33 and the metric seems to good with not much drop compared to train data.
- Probability also converted to lead score ranging from, 0 - 100 for all the leads of train and test data



	Metric	Train Data	Test Data	Difference
0	Accuracy	0.801004	0.801538	-0.000534
1	Sensitivity	0.798264	0.805182	-0.006919
2	Specificity	0.802682	0.799290	0.003392
3	Precision	0.712283	0.712224	0.000059
4	Recall	0.798264	0.805182	-0.006919
5	Negative Predictive Value	0.866703	0.869285	-0.002582
6	False Positive Rate	0.197318	0.200710	-0.003392



Key Input to the business

- The below features have high influence over model as 9 out of 11 features in he final model are from these three
- features
  - Lead Origin
  - Las activity
  - Lead source
- Especially the below category of each above features seems to be very important in identifying the potential leads:
  - Lead Origin – Add form
  - Las activity – SMS sent
  - Lead source – Welingak website
- In a vital situation where business needs to target highly potential candidates for lead conversion, it is recommended to target with lead score over 70. Lead score over 90 is very highly potential candidate for lead conversion.

Thank you