Lead Scoring Case Study

DS C68 Batch

Submitted by:
Gowtham M
Harish N
Richa G

Upgrad Assignment

- Problem Statement and Objective
- Solution Methodology
- EDA (Visualizations)
- Correlation
- Model Building
- Model Evaluation
- Observation & Conclusion

Problem Statement & Objective

Problem Statement:

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.
- To make this process more efficient, the company wishes to identify the most potential leads, also known as ‘Hot Leads’.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

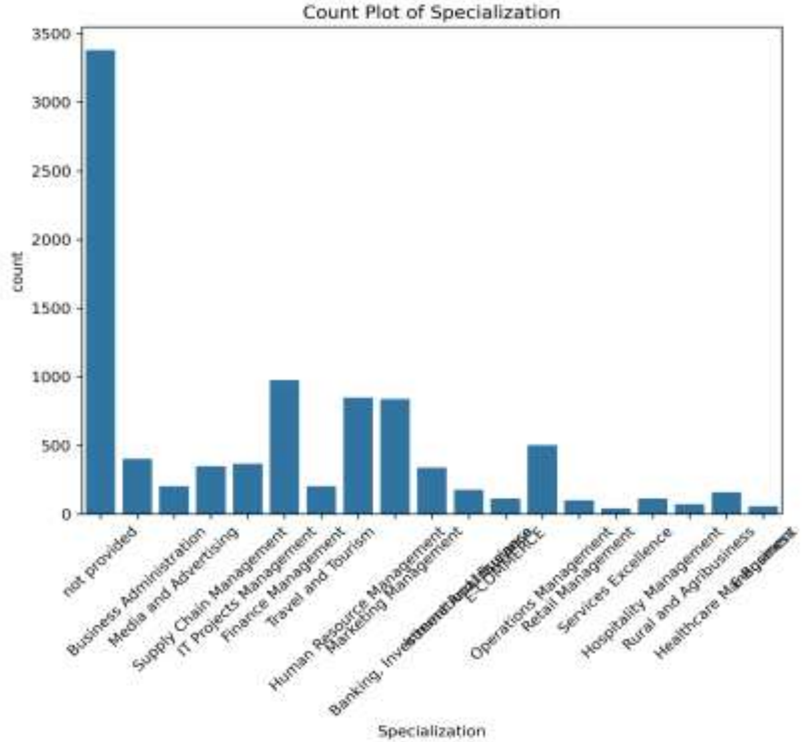
Objective:

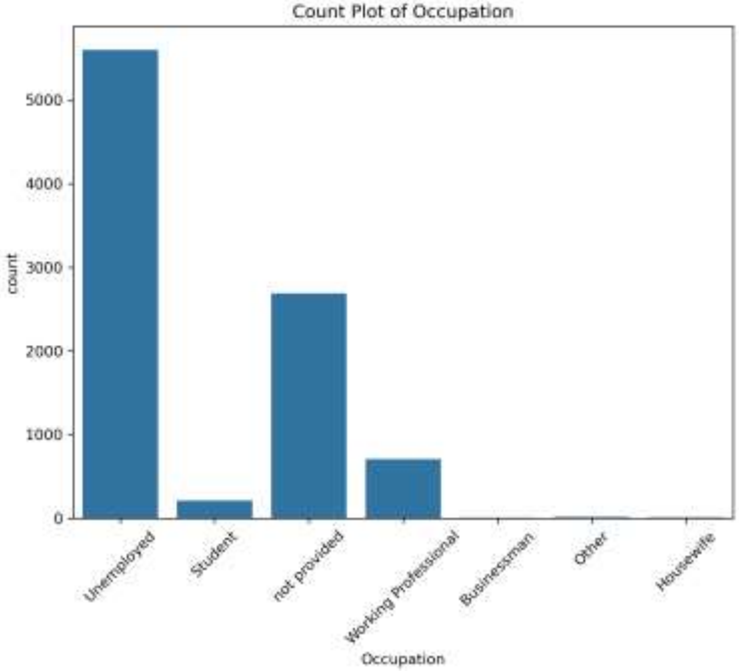
- X education wants to know most promising leads. For that they want to build a Model which identifies the hot leads. CEO of X education want to achieve a lead conversion rate of 80%.
- Once completed, Deployment of the model for the future use.

Solution Methodology

- ❑ Importing data and inspecting the data frame
- ❑ Data preparation(Data cleaning and data manipulation)
- ❑ EDA(Univariate data analysis, Bivariate data analysis)
- ❑ Dummy variable creation
- ❑ Test-Train split
- ❑ Feature scaling
- ❑ Correlations
- ❑ Model Building (Logistic Regression used ,with statistical evaluation - RFE,VIF and p- values)
- ❑ Model Evaluation
- ❑ Making predictions on test set

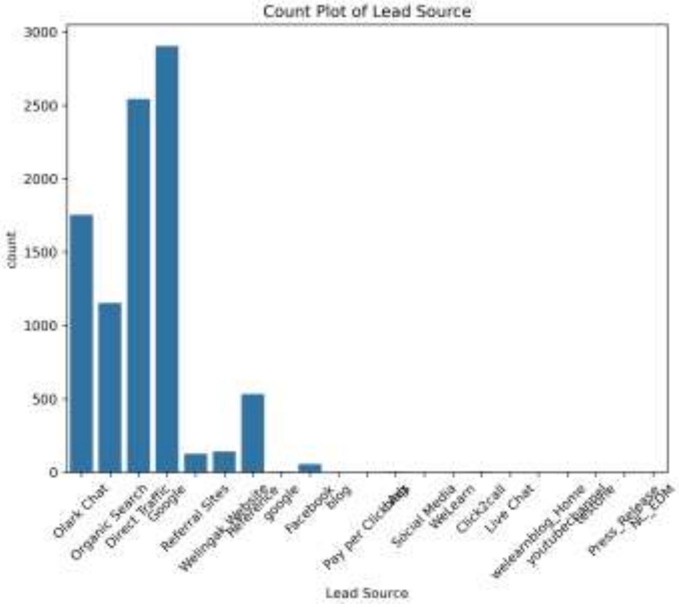
EDA – Variables with “Select” datapoints

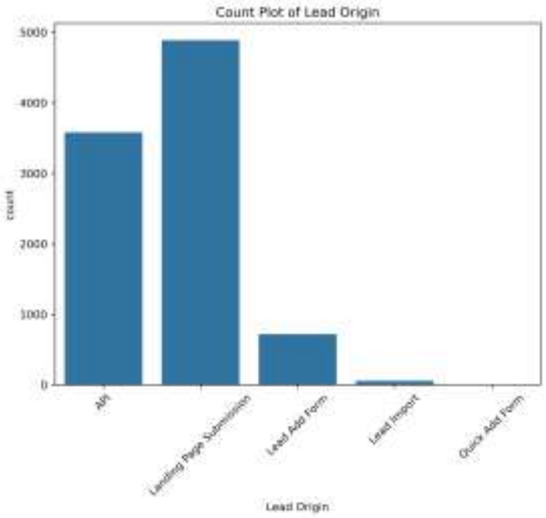


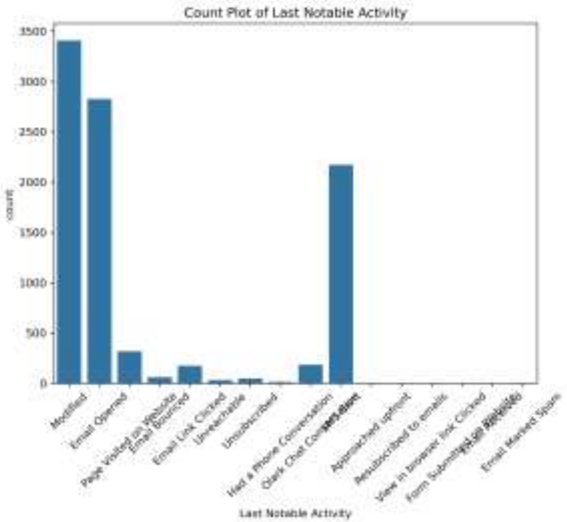


- Variables like Specialization, Occupation and How did you ear about X education has “Select” as one of the option which is similar to NAN values , Hence they are imputed as “not provided”

EDA - Univariate Analysis - Categorical

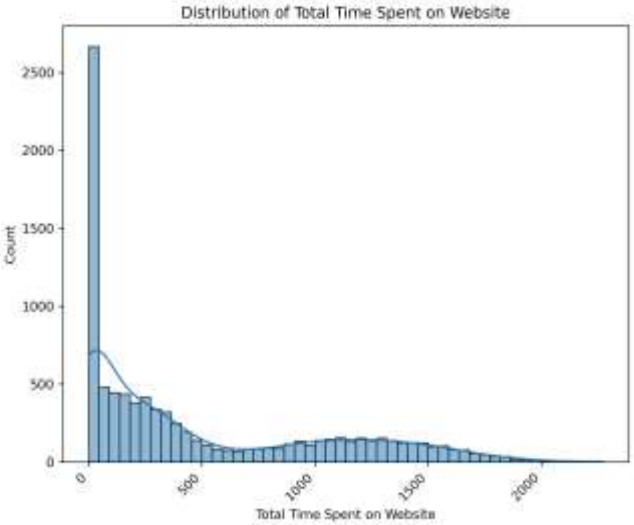


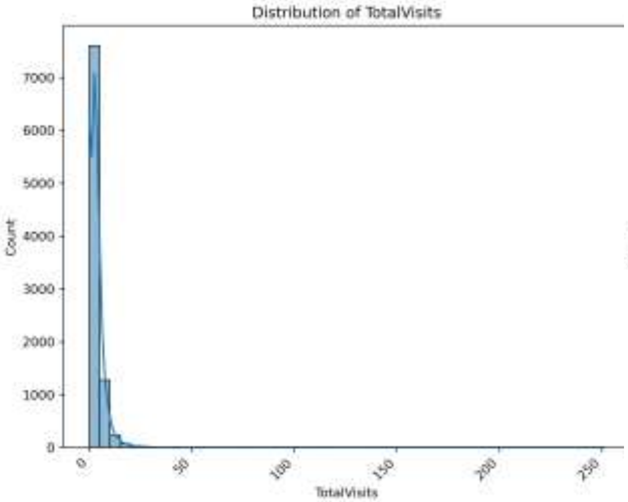


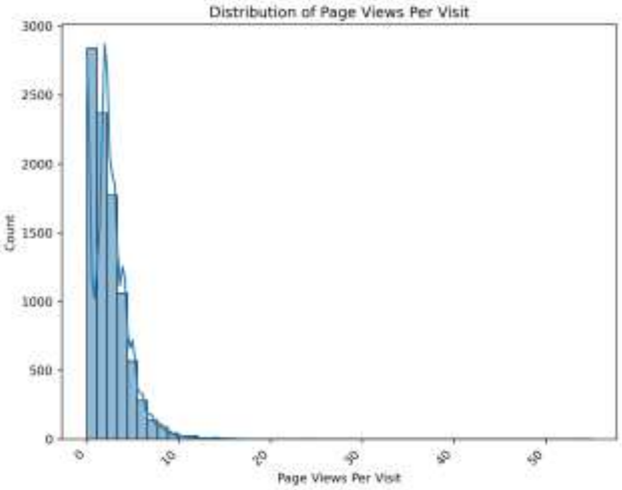


- In lead source the leads through google & direct traffic high probability to convert
- Whereas in Lead origin the greatest number of leads are landing on submission
- X Education gets a lot of leads, its lead conversion rate is very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

EDA - Univariate Analysis - Numerical

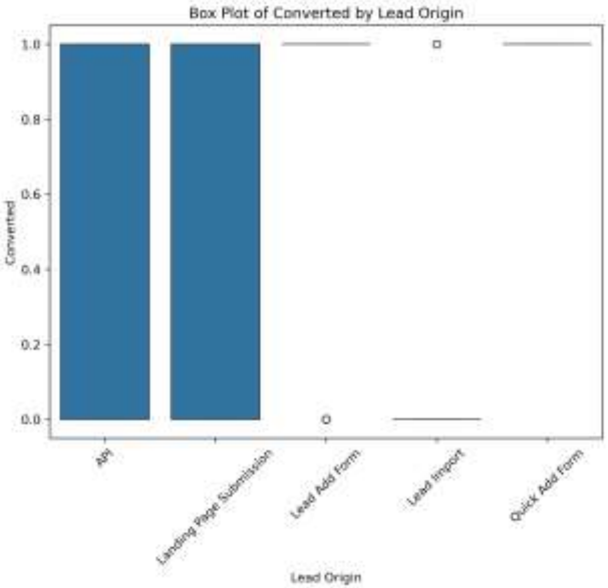


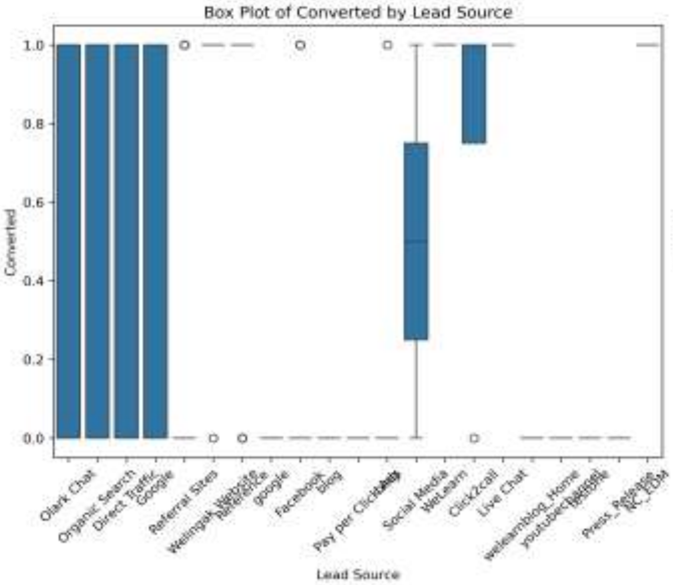


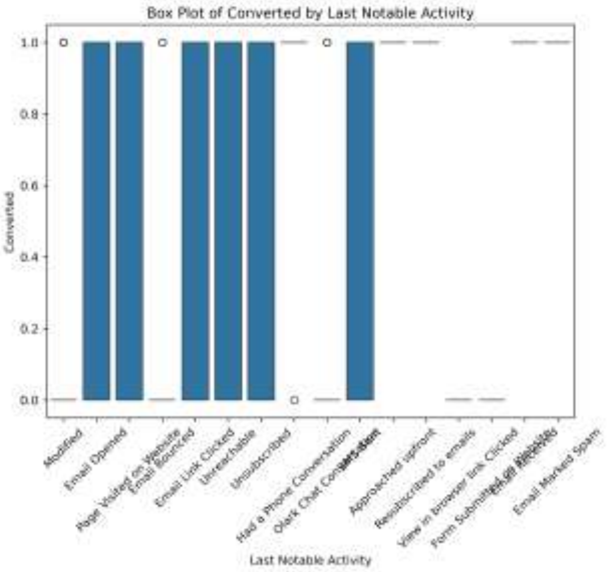


- 'Total Time Spent on Website','TotalVisits','Page Views Per Visit' displays a downward trend.
- All the three variables looks insightful as the trend forms a pattern over count.
- Numeric variables are highly useful to understand the distribution of the count.

Bivariate Analysis – Target(‘Converted’)

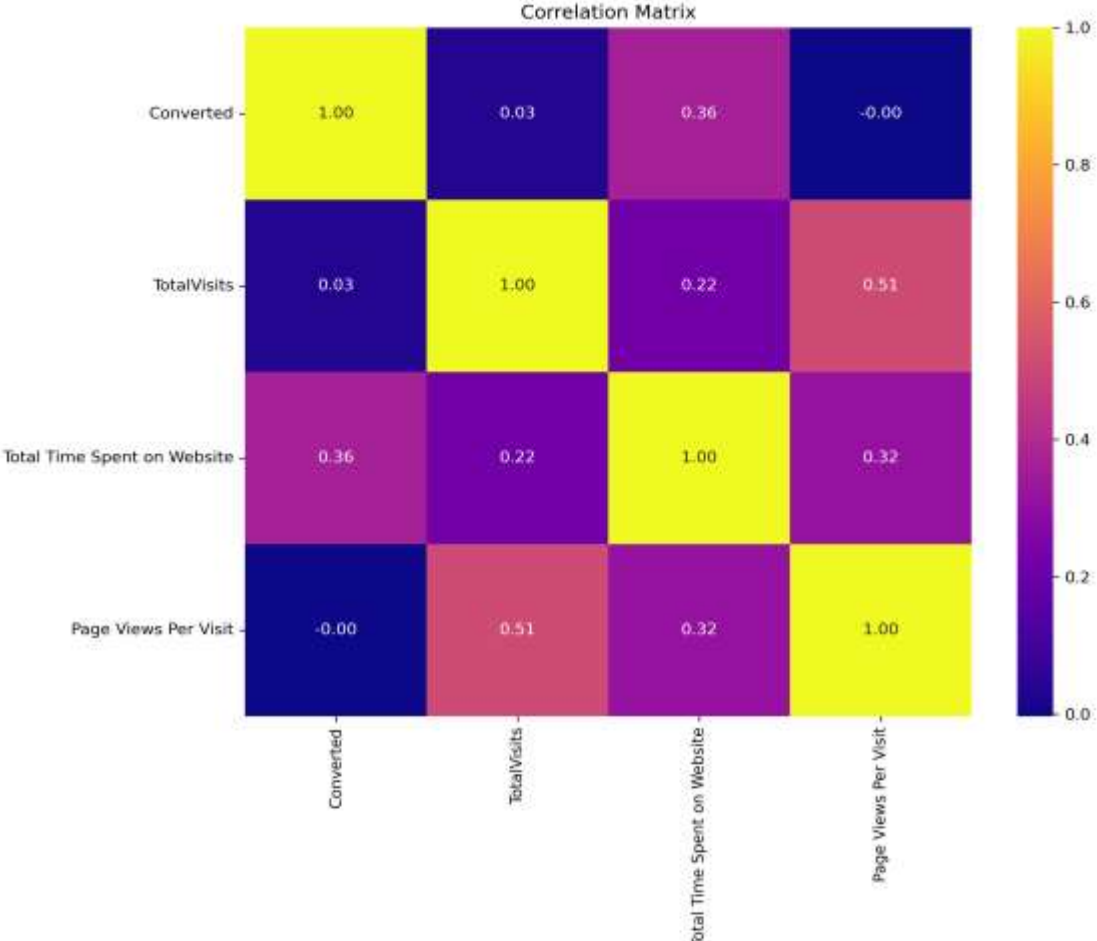






- 'Lead Origin’ has insights on distribution based on API and 'Landing page submission’
- 'Lead Source' shows promising insights about conversion and major sources like Olark, google, we chat etc. which needs to be explored.
- 'Last Notable activity' shows definitive information about lead conversion.

Correlations



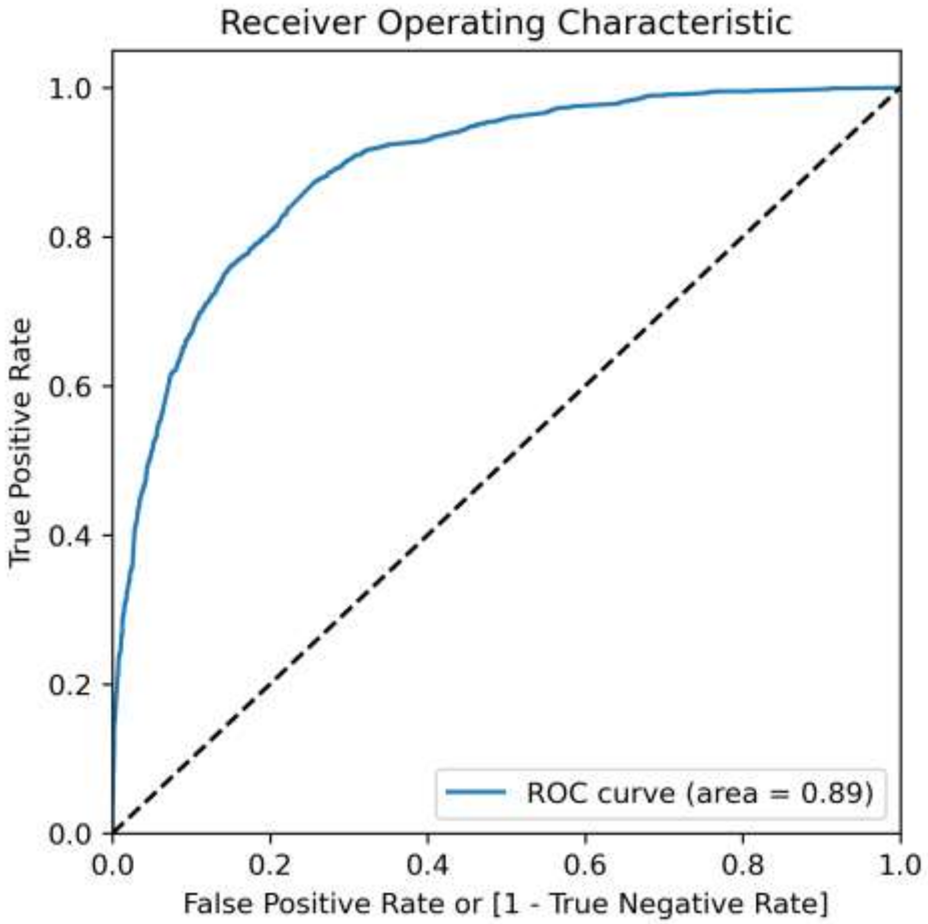
- Correlation between Total Visits and Page view visit is observed at 0.51 which is comparatively lower.

- No other major correlation observed.

Model Building

- ❑ Splitting the Data into Training and Testing Sets
- ❑ The first basic step for regression is performing a train-test split, we have chosen 70:30 ratio.
- ❑ Use RFE for Feature Selection
- ❑ Running RFE with 15 variables as output
- ❑ Building Model by removing the variable whose p- value is greater than 0.05 and vif value is greater than 5
- ❑ Predictions on test data set
- ❑ Overall accuracy 81%

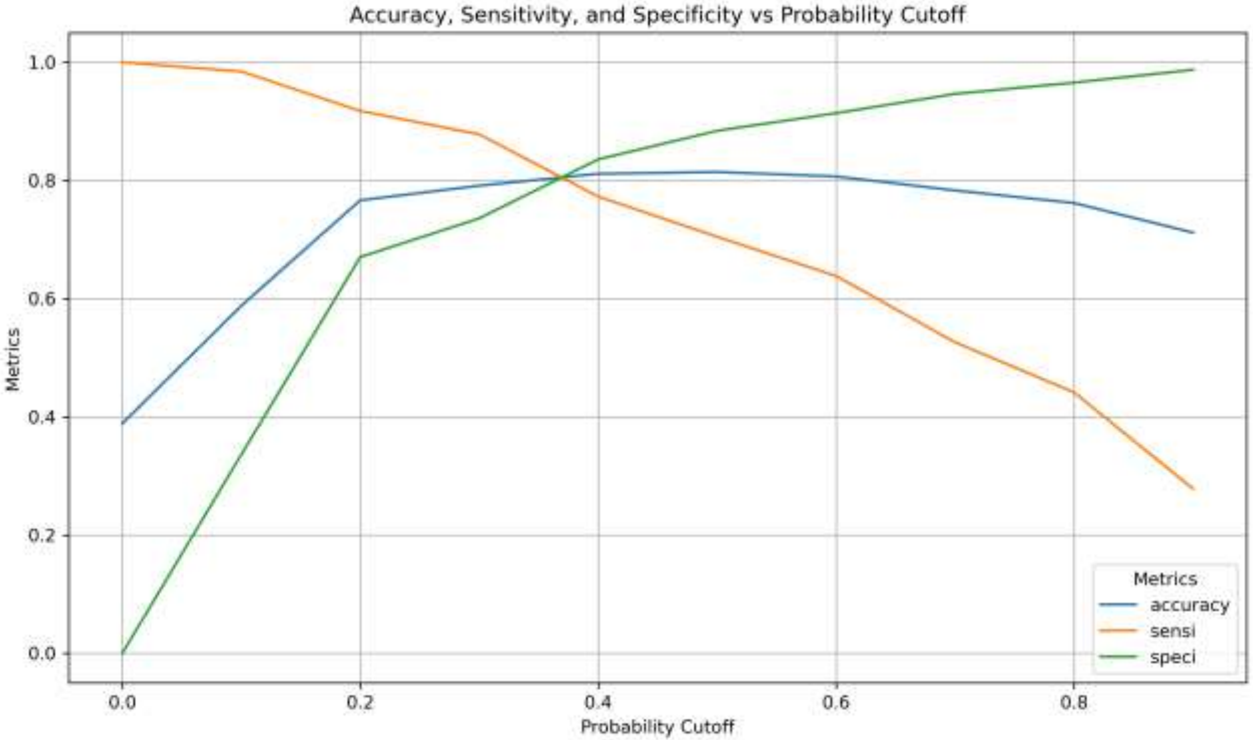
Model Evaluation –ROC curve



- A Receiver Operating Characteristic (ROC) curve is a graph that shows the performance of a model across all thresholds by plotting the true positive rate (TPR) against the false positive rate (FPR). The closer the curve is to the upper left corner, the better the model's overall accuracy

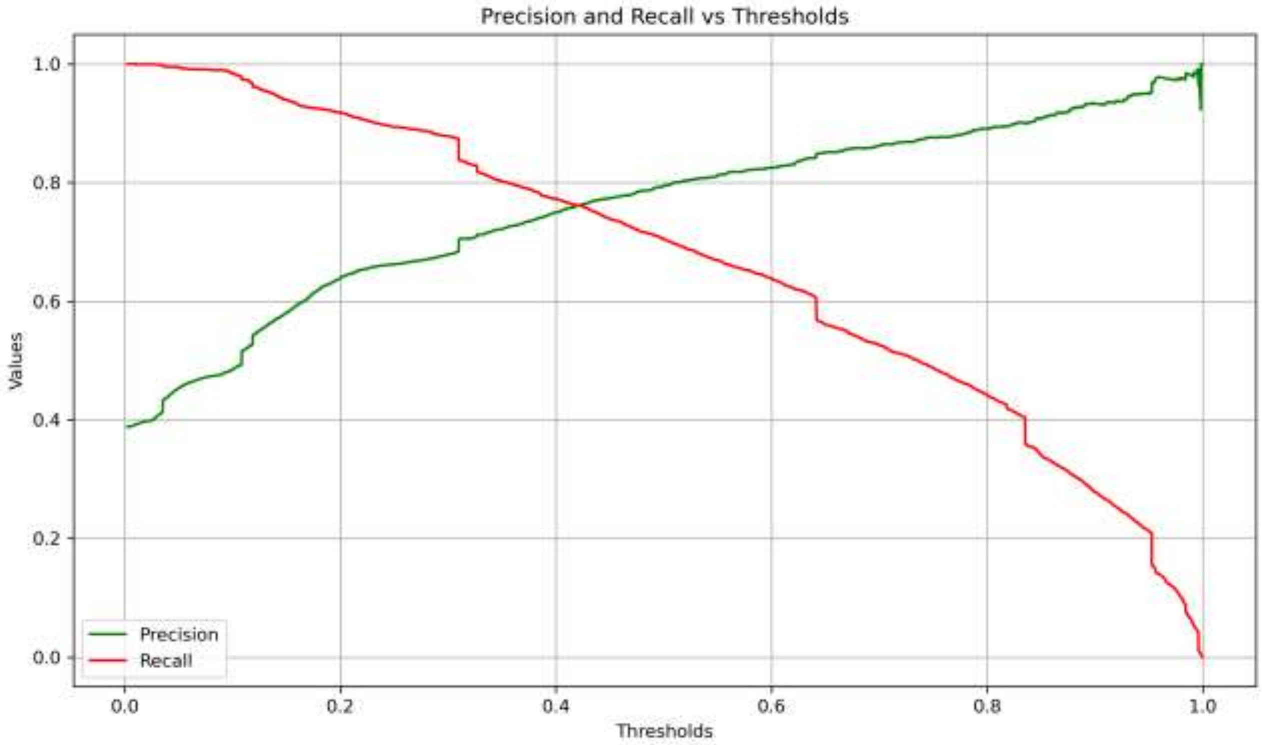
- ROC area under the curve values are in acceptable range(0.89)

Model Evaluation –ROC curve



- To find optimum cutoff for the Probability conversion rate , we have used a function to determine the best possible value as '0.38'

Model Evaluation-Precision & Recall tradeoff



- To find optimum cutoff for the Probability conversion rate , we have used a function to determine the best possible value as ‘0.42'

Observation

Train Data:

Accuracy : 80%

Sensitivity :78%

Specificity : 82%

Test Data:

Accuracy : 81%

Sensitivity : 78%

Specificity : 82%

FINAL SELECTED FEATURES

- ❑ Total visits.
- ❑ The total time spend on Website.
- ❑ When the lead origin was:
- Lead Add form
- Landing Page submission
- ❑ When the last notable activity was:
- Had a phone conversation
- Unreachable
- ❑ When the last activity was:
- SMS sent
- Olark chat conversation
- ❑ When the Lead source is "olark chat".
- ❑ When their current occupation is as a working professional

Conclusion

Below mentioned category of People should be reached for Lead conversion

- ❑ Customer spending lots of time on the website. (Also to improve the metric website User interface and functionality can be made more user friendly and appealing.)
- ❑ Customer Visiting the site regularly. (This metric will improve automatically if the above metric is improved).
- ❑ People identified as Lead(Lead origin) from "Lead Add Form" source have more potential of conversion.
- ❑ People receiving a "Phone call" as "Last Notable Activity" are higher chance of conversion. So, reaching out people with above conditions are ideal.
- ❑ Customer whose occupation is "Working professional" have higher chance of conversion.

If the above factor are taken into account Lead scoring rate can be vastly improved.

Thank You!