Lead Scoring Case Study

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 mostpotential 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.

Business Objective

- X education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Solution Methodology

Data cleaning and data manipulation.

- 1. Check and handle duplicate data.
- 2. Check and handle NA values and missing values.
- 3. Drop columns, if it contains large amount of missing values and not useful for the analysis.
- 4. Imputation of the values, if necessary.
- 5. Check and handle outliers in data.

EDA

- 1.Univariate data analysis: value count, distribution of variable etc.
- 2.Bivariate data analysis: correlation coefficients and pattern between the variables etc.

- Feature Scaling & Dummy Variables and encoding of the data.
- Classification technique:
- logistic regression used for the model making and prediction.
- Validation of the model.
- Model presentation.
- Conclusions and recommendations.

Data Manipulation

- Single value features like Magazine, Receive More Updates About Our Courses, Update me on Supply

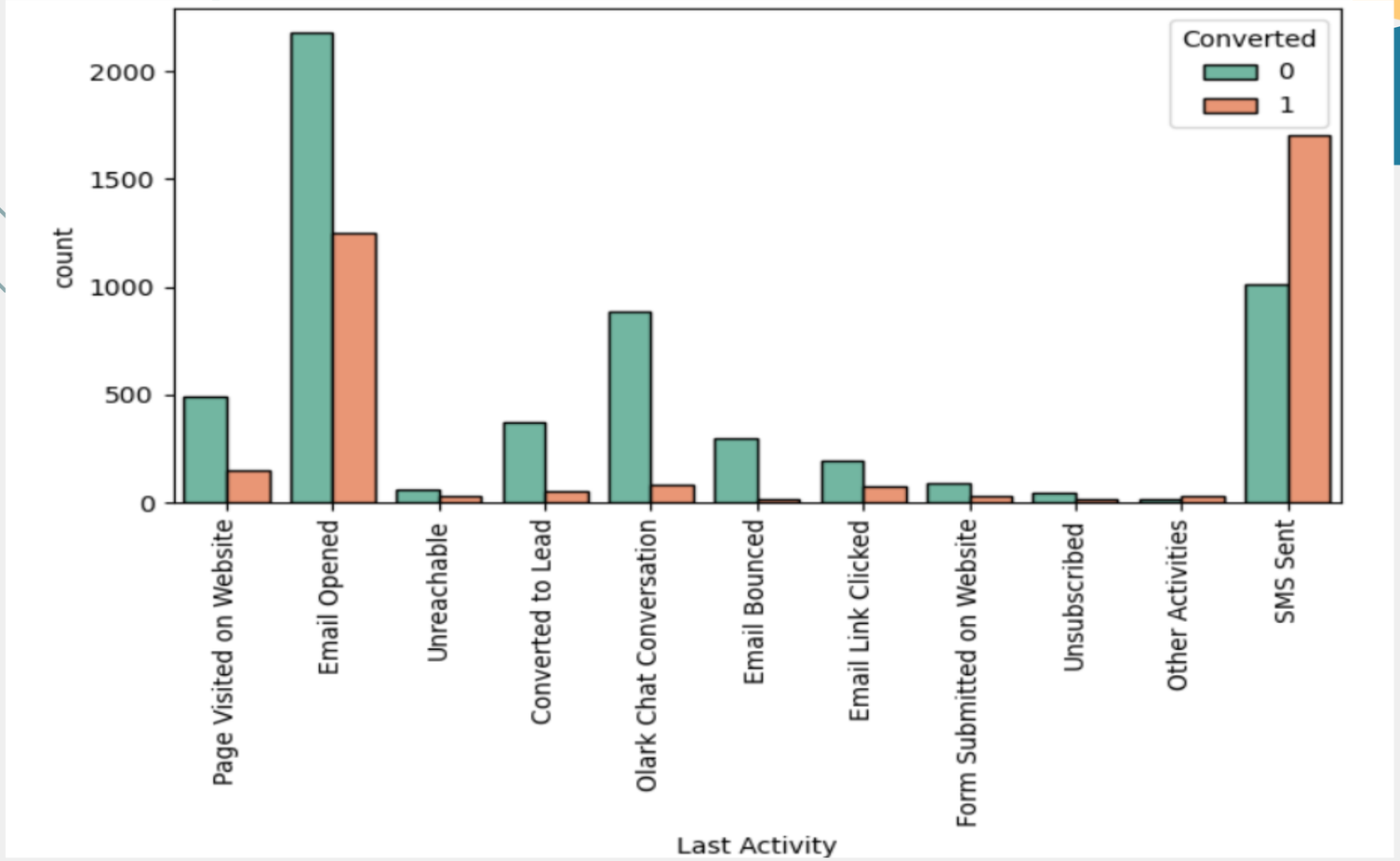
- Chain Content Get updates on DM Content, I agree to pay the amount through cheque etc. have been dropped.

- Removing the Prospect ID and Lead Number which is not necessary for the analysis.

- After checking for the value counts for some of the object type variables, we find some of the features which has no enough variance, which we have dropped,the features are: Do Not Call, What matters most to you in choosing course, Search, Newspaper Article, X Education Forums, Newspaper,Digital Advertisement etc.

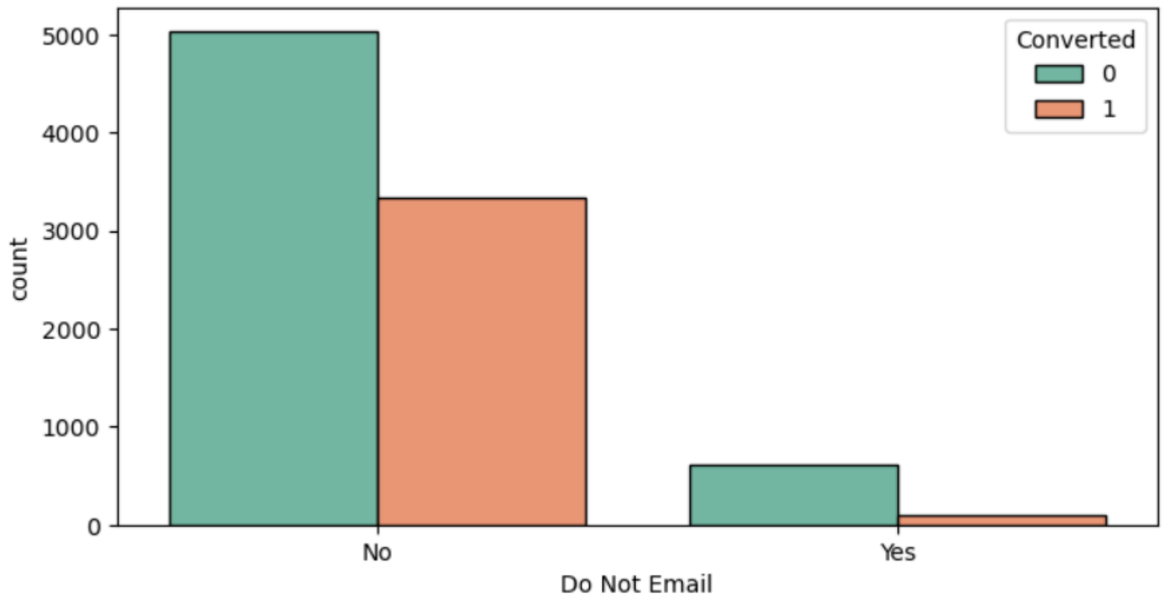
- Dropping the columns having more than 35% as missing value such as How did you hear about X Education and Lead Profile.

EDA



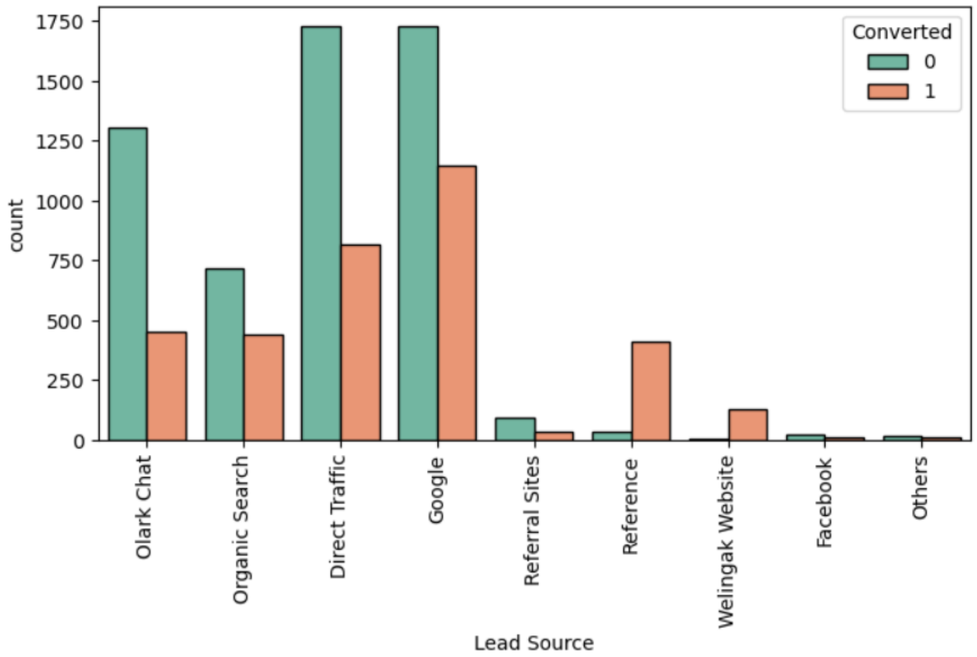
Inferences

- Conversion rate for Email Opened is ~60%
- SMS Sent has high Conversion rate



Inferences

- most of them are 0 so we cannot make any inferences

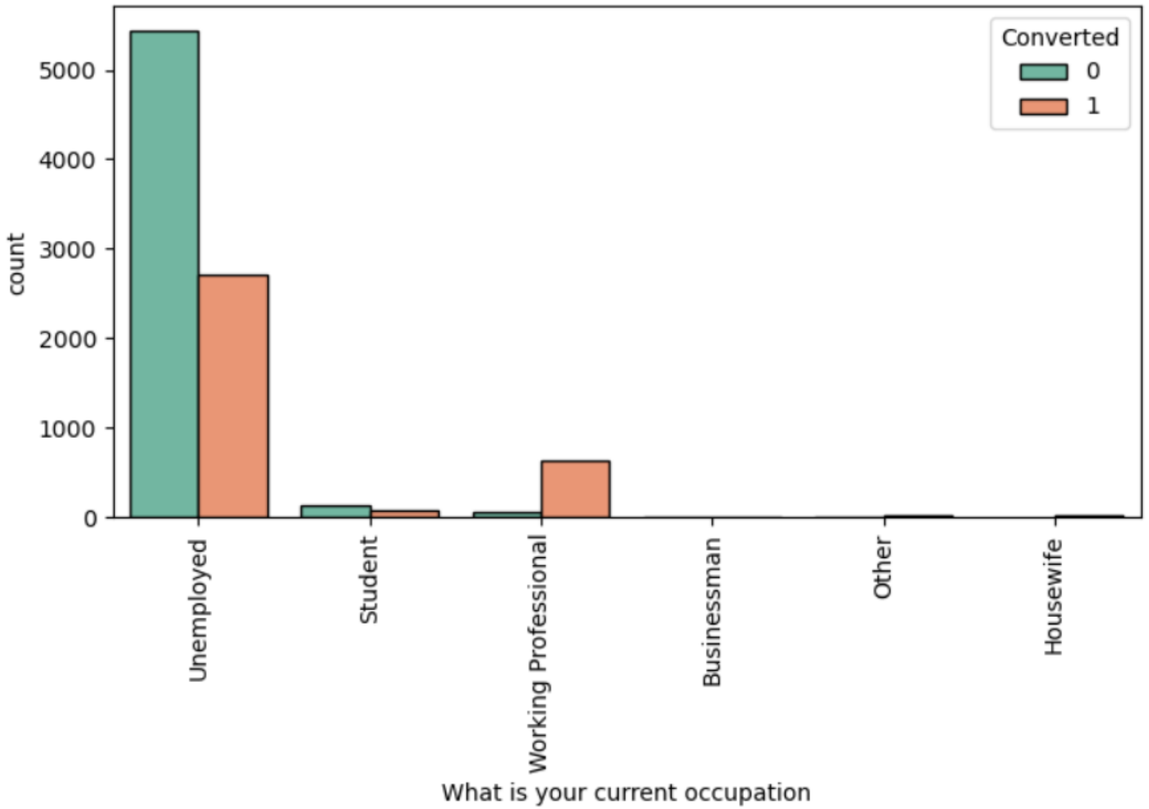


Inferences

•Google and Direct Traffic has the highest conversion rate

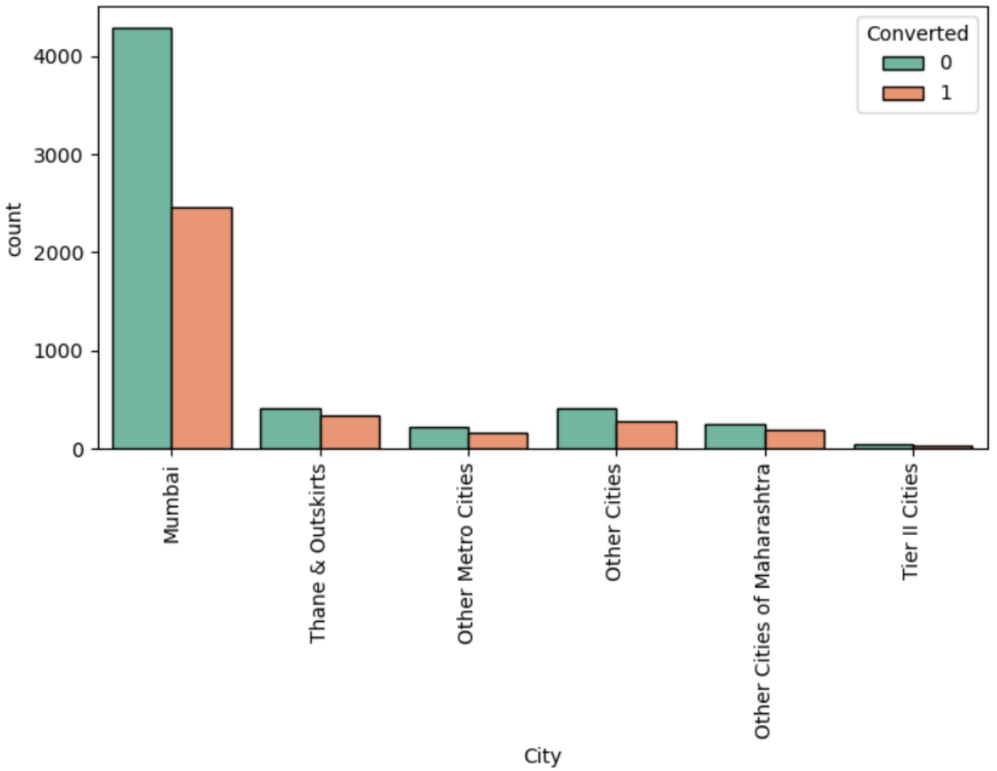
•References and Welingak website has high conversion rate but low counts

Categorical Features



Inferences

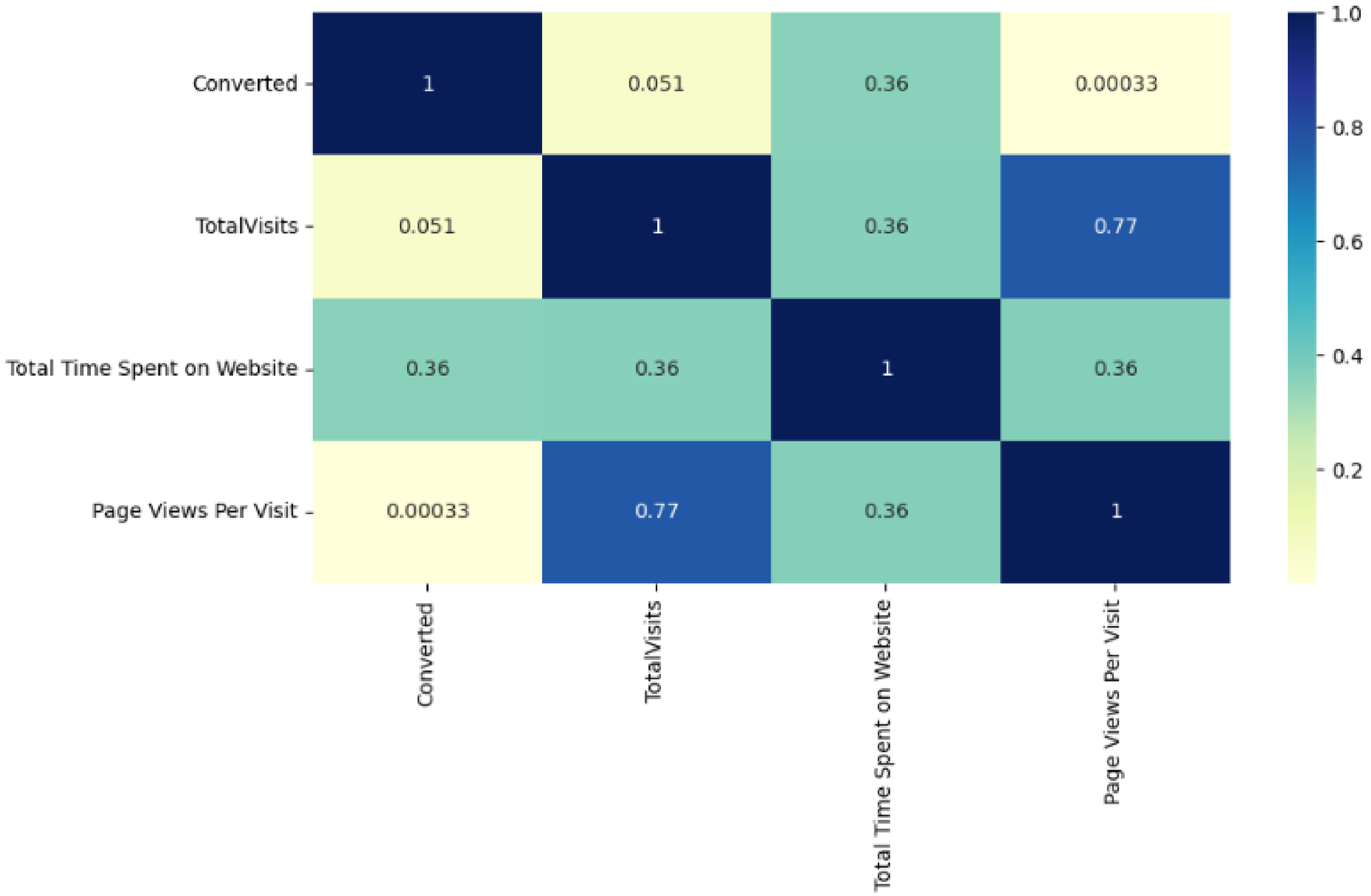
- from the above countplot we can say that Unemployed has high count and 30-35% conversion rate and Working Professional has high conversion rate



Inferences

- Mumbai has the conversion rate of ~50%, Remaining values has Considerable conversion rates

Numerical Features



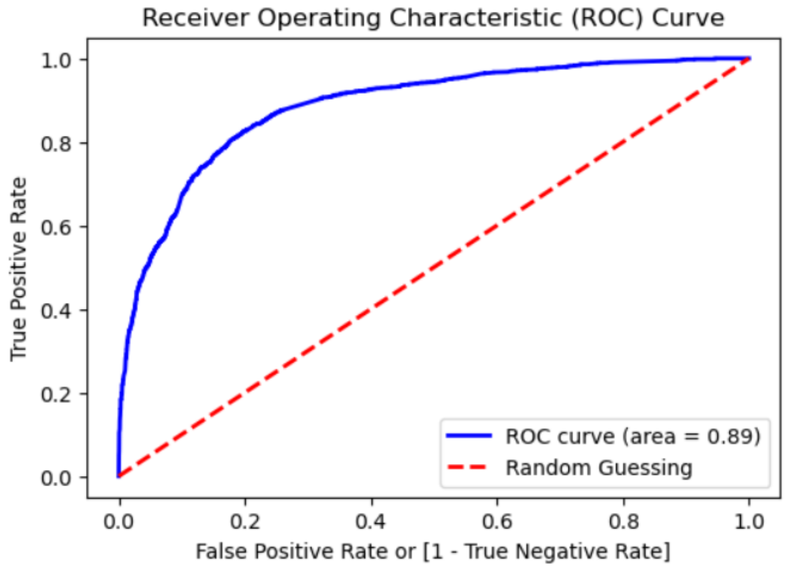
Data Conversion

- Numerical Variables are Normalized
- Dummy Variables are created for object type variables

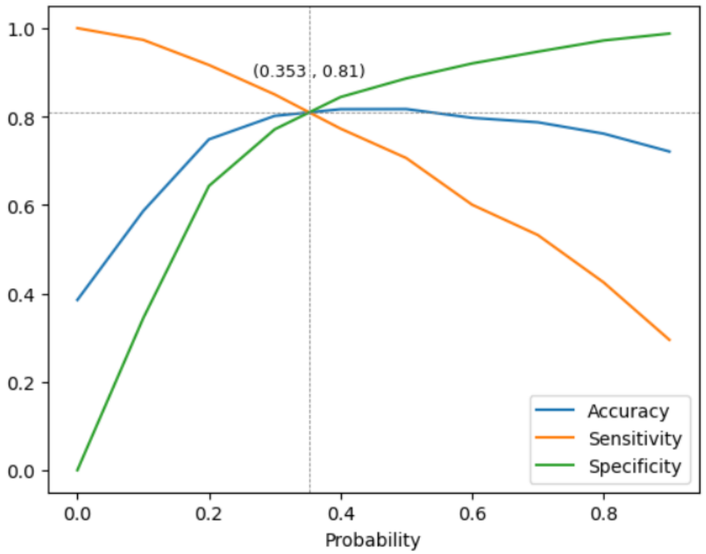
Model Building

- 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 79.43% on test data, and 79.17% on train data set.

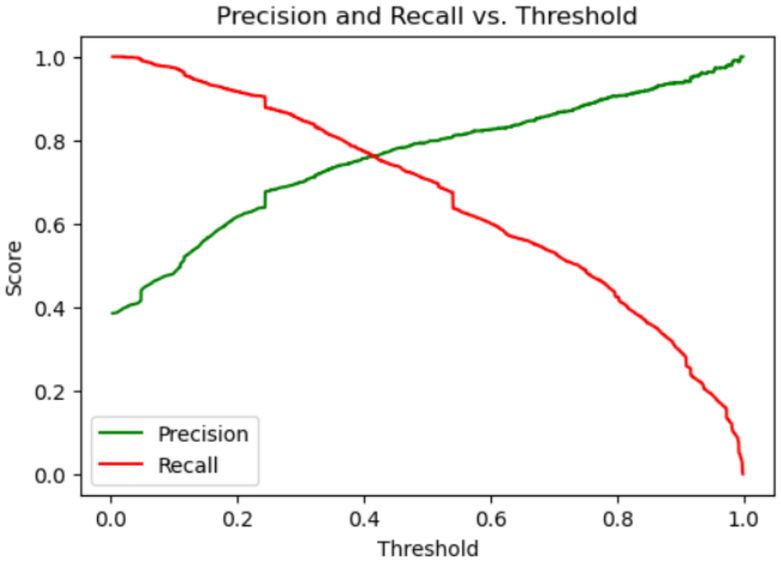
ROC Curve



The area under ROC curve is 0.89 which indicates our model is a good one



From the above graph we can take 0.354 as optimal cutoff value



from the above graph we can say there is a trade off between Precision and Recall

Finding Optimal Cut off Point

- Optimal cut off probability is that
- Probability where we get balanced sensitivity and specificity.
- From the second graph it is visible that the optimal cut off is at 0.35.
- From Precision and recall chart suitable threshold is near to 0.39
- Although the values are same at both cutoff 0.35 and 0.38, so kept 0.35.

Conclusion

It was found that the variables that mattered the most in the lead converted are
Lead Origin
    Lead Add Form
- Do Not Email
  - Yes
- Last Activity
  - Converted to Lead
  - Olark Chat Conversation
- What is your current occupation_
  - Unemployed
  - Working Professional
- Tags
  - Busy
  - Closed by Horizzon
  - Lost to EINS
  - Will revert after reading the email
  - in touch with EINS
- Last Notable Activity
  - SMS Sent

THANK YOU