Lead Scoring Case Study

Harsha Vardhan Tamma
Manushree Jain
Arutla Uma Maheshwar Reddy

The 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.

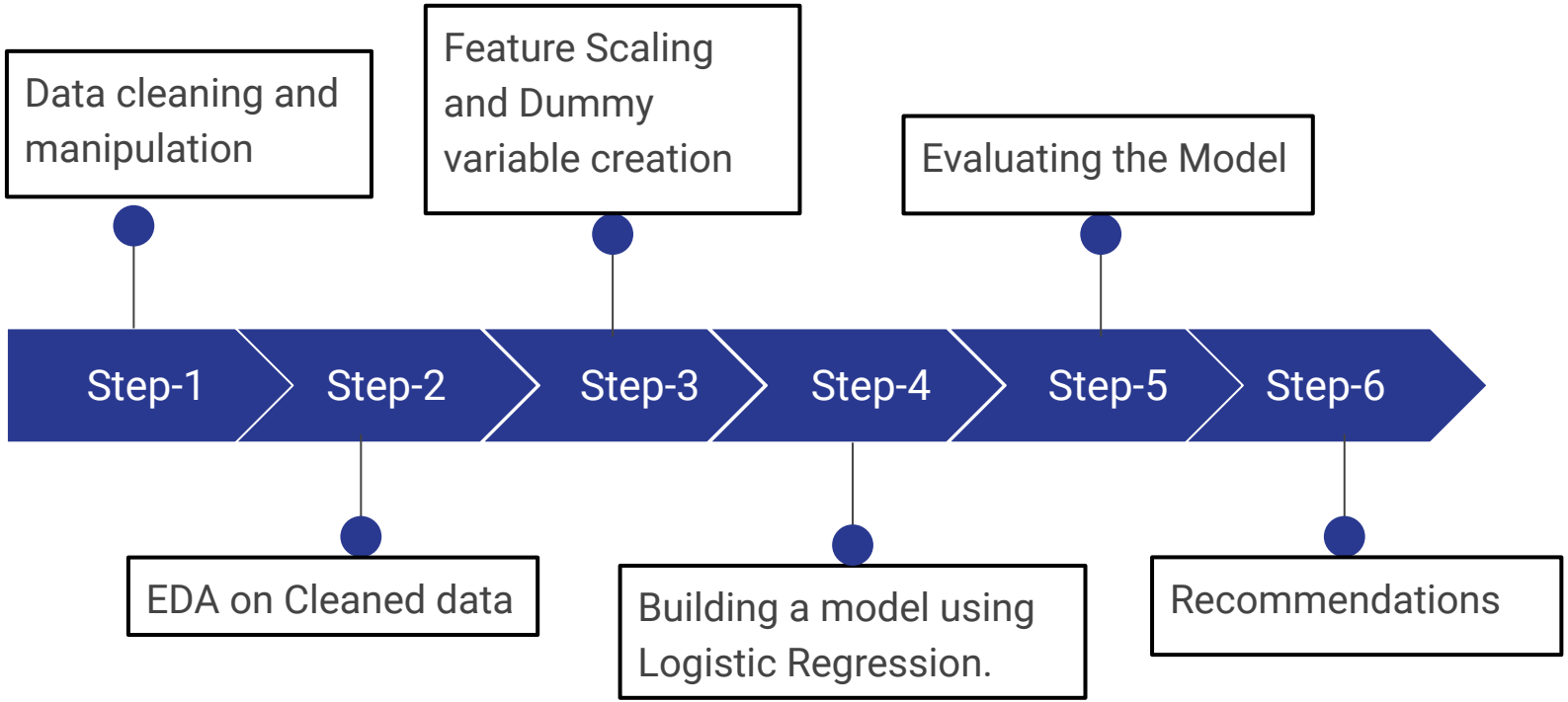
Business Objective

Objective 1	Objective 2	Objective 3
X education wants to know most promising leads.	X education wants us to build a Model which identifies the hot leads.	X education wants their model deployed for future use.

Goal

To achieve target lead conversion rate of ~80%

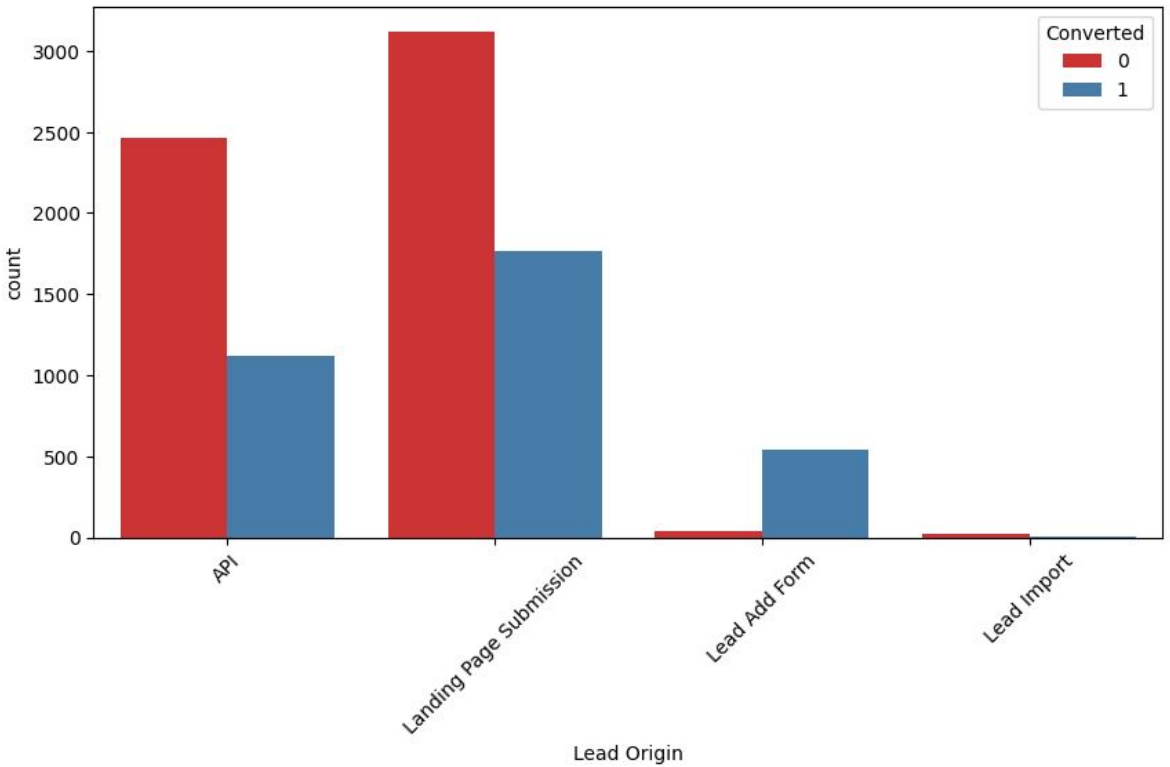
Implementation


Data cleaning and manipulation
Feature Scaling and Dummy variable creation
Evaluating the Model
Step-1
Step-2
Step-3
Step-4
Step-5
Step-6
EDA on Cleaned data
Building a model using Logistic Regression.
Recommendations

Data Cleaning and Manipulation.

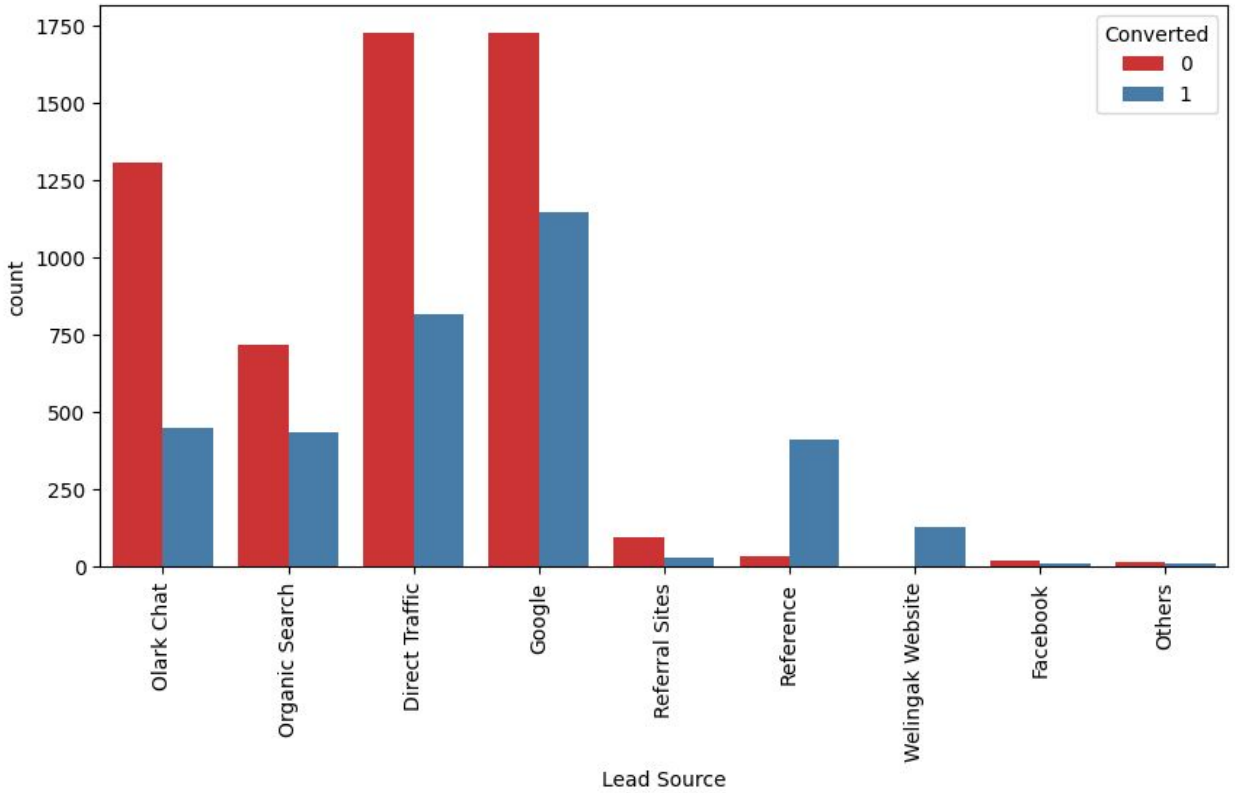
- ➔ Dataset has 9240 rows and 37 columns
- ➔ ‘Select’ level in 3 categorical variables are converted to NULL (NaN).
- ➔ Dropped columns with more than 40% of missing values.
- ➔ Missing values in columns “Specialization”, “Tags”, “What is your current occupation”, “Country”, “City” are imputed.
- ➔ Column “What matters most to you in choosing a course” has been dropped due to highly skewed data.
- ➔ For columns with less than 2% of missing values, only the rows containing the null values are dropped.
- ➔ After cleaning, we were able to retain ~98% of the data.

EDA - Univariate Analysis - Lead Origin



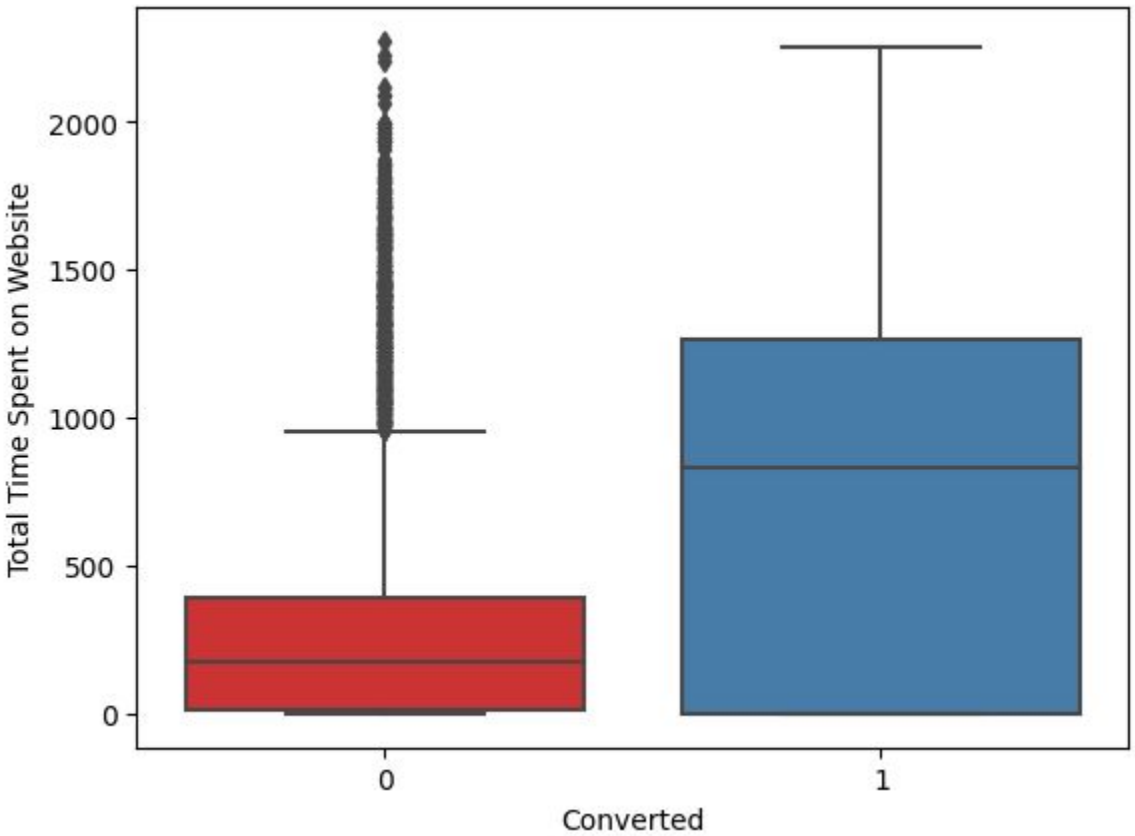
- ➔ API and Landing Page Submission have 30-35% conversion rate but count of lead originated from them are considerable.
- ➔ Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- ➔ Lead Import are very less in count.
- ➔ To improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin and generate more leads from Lead Add Form

EDA - Univariate Analysis - Lead Source



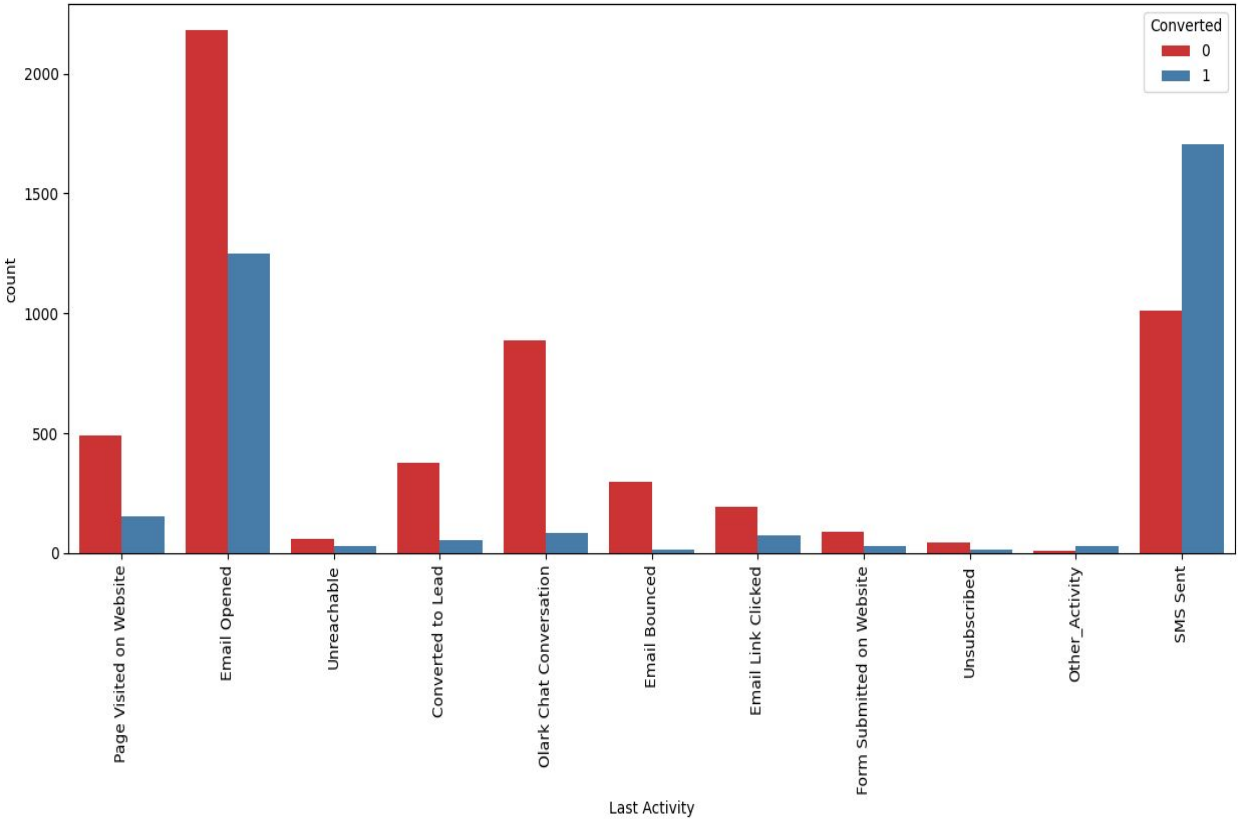
- ➔ Google and Direct traffic generates maximum number of leads.
- ➔ Conversion Rate of reference leads and leads through welingak website is high.
- ➔ To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website

EDA - Univariate Analysis - Total Time Spent on Website



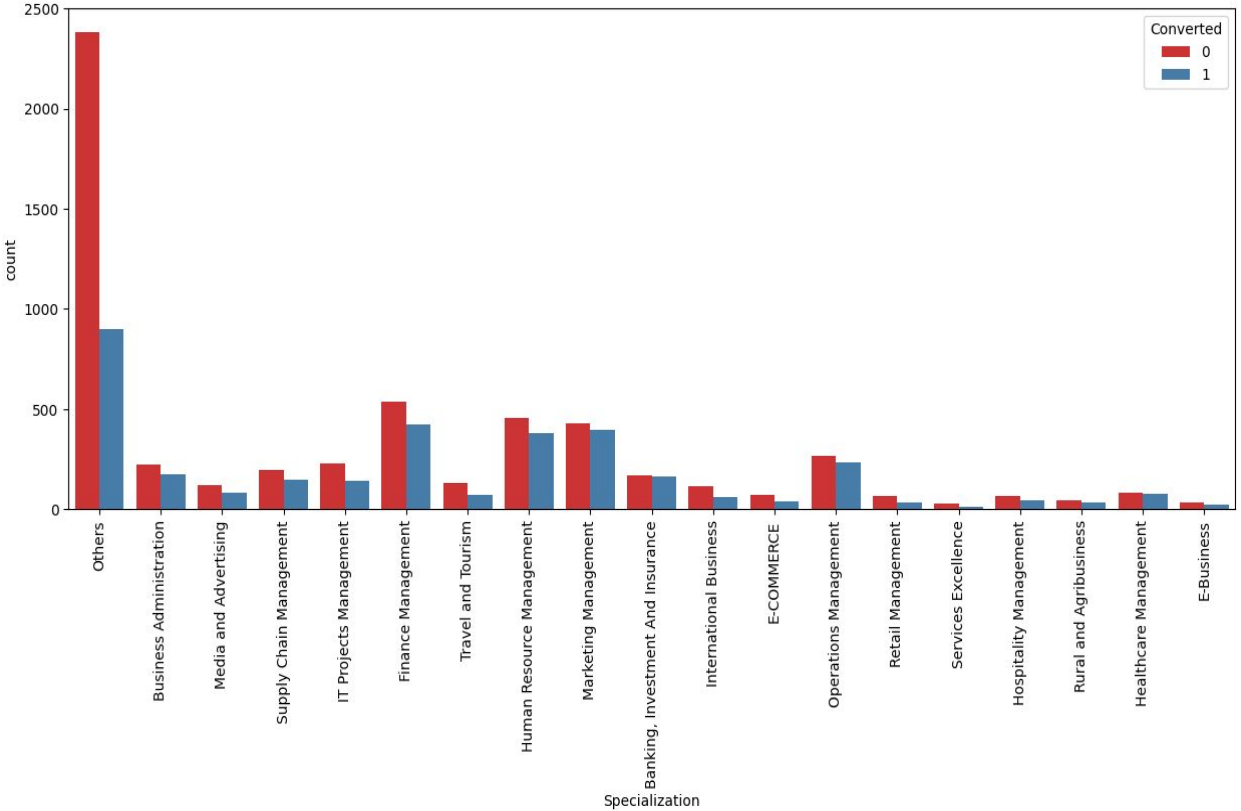
- ➔ Leads spending more time on the weblise are more likely to be converted.
- ➔ Website should be made more engaging to make leads spend more time.

EDA - Univariate Analysis - Last Activity



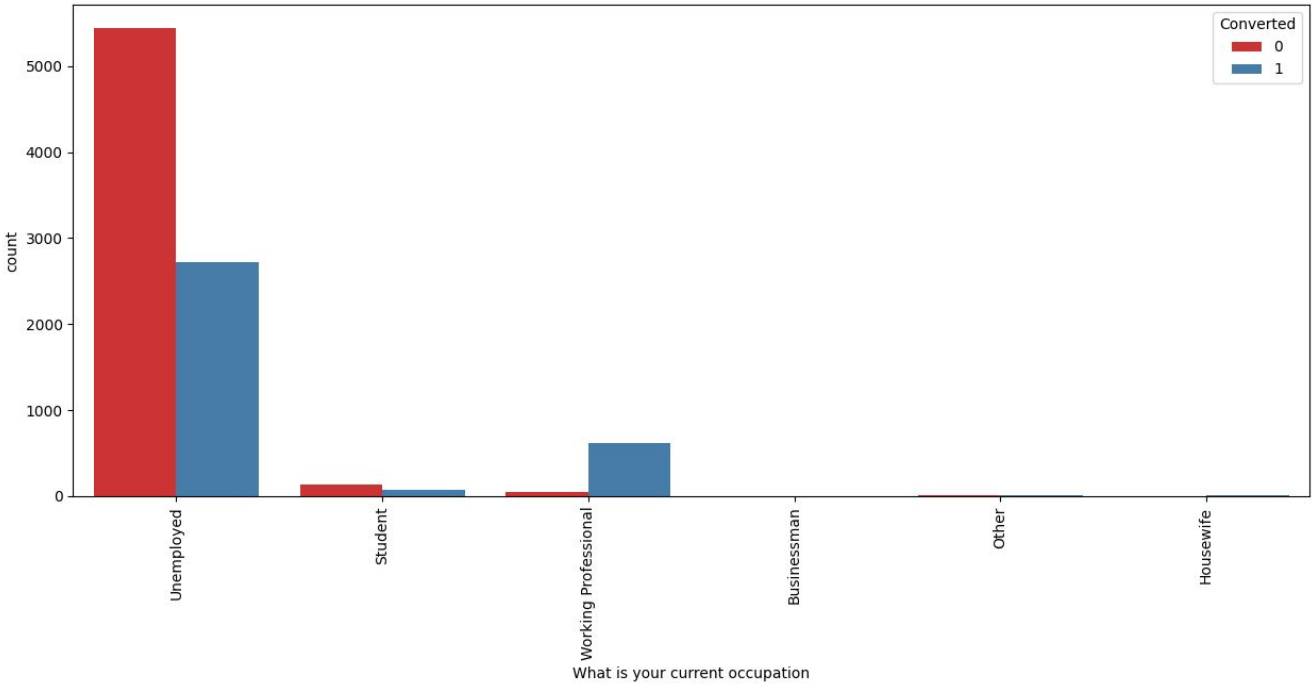
- ➔ Most of the lead have their Email opened as their last activity.
- ➔ Conversion rate for leads with last activity as SMS Sent is almost 60%.

EDA - Univariate Analysis - Specialization



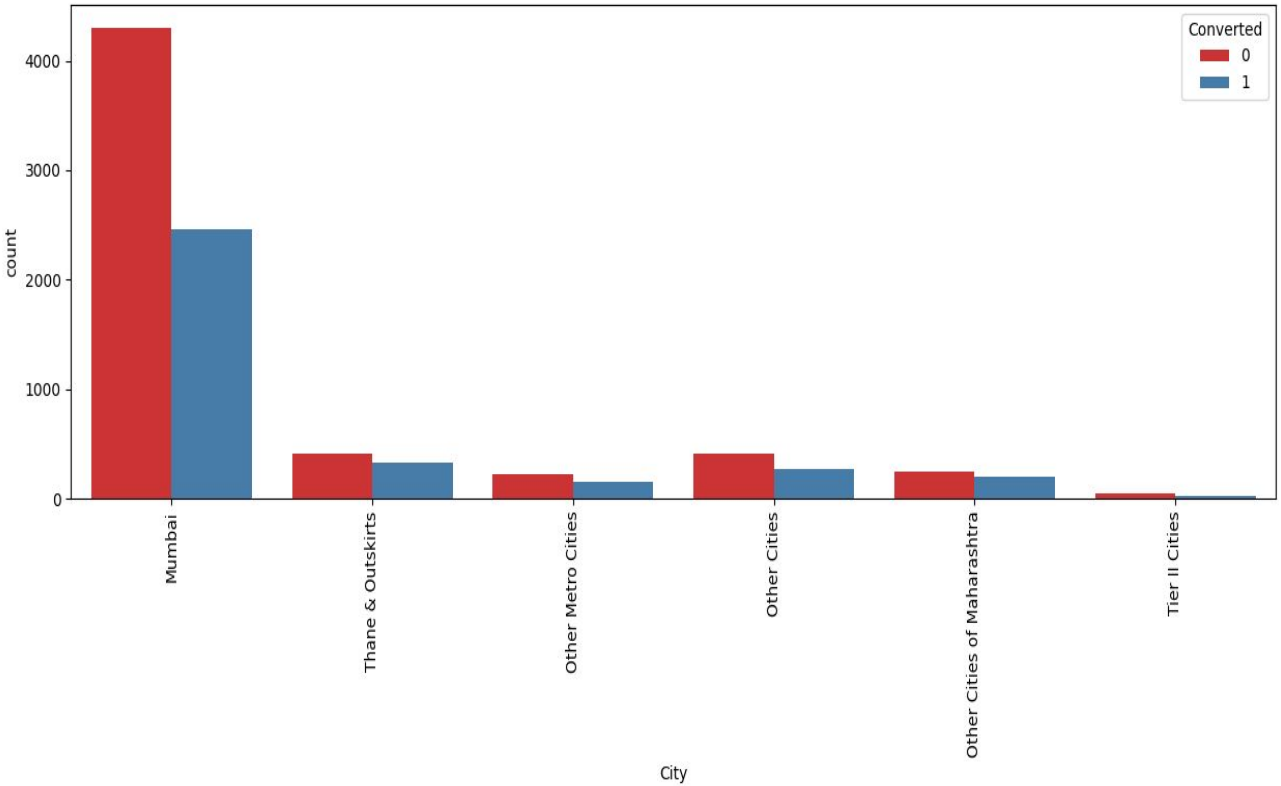
➔ Focus should be more on specialization with high conversion rate

EDA - Univariate Analysis - What is your current occupation



- ➔ Working Professionals going for the course have high chances of joining it.
- ➔ Unemployed leads are the most in numbers but has around 30-35% conversion rate.

EDA - Univariate Analysis - City



➔ More leads from Mumbai with around 50% conversion rate.

EDA - Conclusion

- ➔ Based on the univariate analysis we have seen that many columns are not adding any information to the model, hence we dropped them for further analysis.
  - ◆ 'Lead Number', 'Tags', 'Country', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', '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', 'A free copy of Mastering The Interview' are dropped.

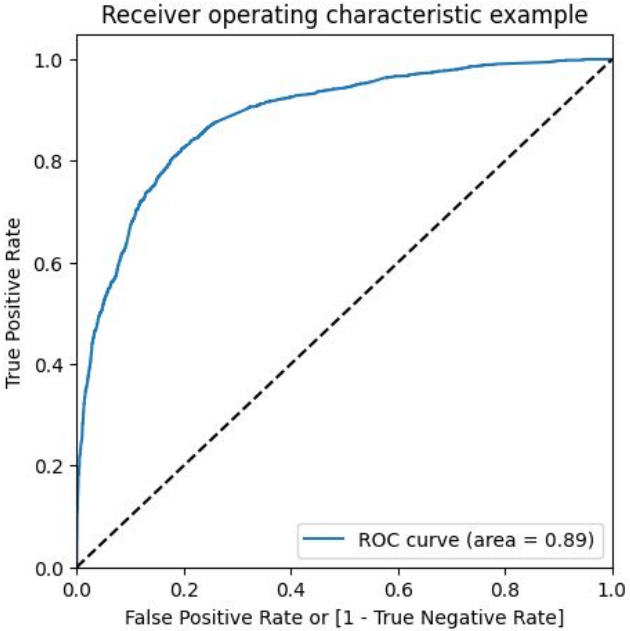
Feature Scaling and Dummy variable creation

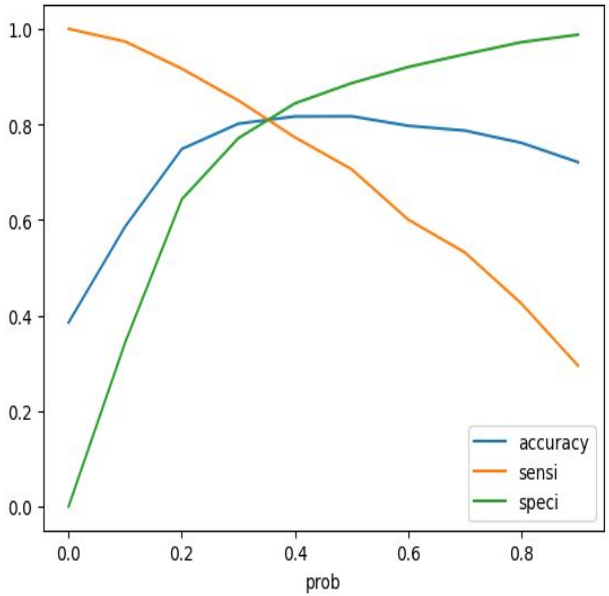
- ➔ Numerical variables are normalised.
- ➔ Dummy variables are created for Categorical variables.
  - ◆ 'Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'City', 'Last Notable Activity'
- ➔ Total rows and columns for analysis: 9074 rows x 71 columns

Model Building and Evaluation

- Data split into Train and Test sets
  - Split between Train and Test is 70:30
- Recursive Feature Elimination (RFE) is used with 15 variables as output.
- Model was built on Train data by removing the variables with P-Value greater than 0.05 and VIF greater than 5.
- Once the model is built, Predictions were done on Train and Test data
- Calculated Accuracy, Sensitivity, Specificity using an arbitrary value such as 0.5 as cut-off probability.
- Overall Accuracy was 81%.

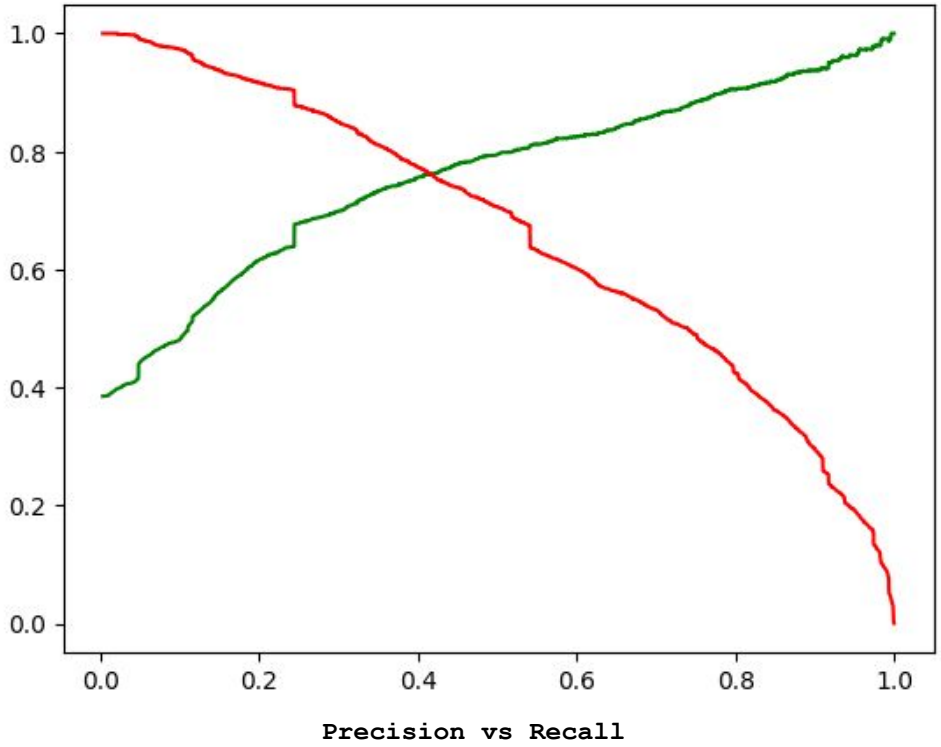
Model Evaluation





- ROC area is 0.89 which indicates our model is good.
- Optimal cutoff probability is a point where we get balanced sensitivity and specificity.
- From the second graph, Optimal Cut-off is at **0.42.**
- Final predictions were done using above Optimal cut-off.

Model Evaluation



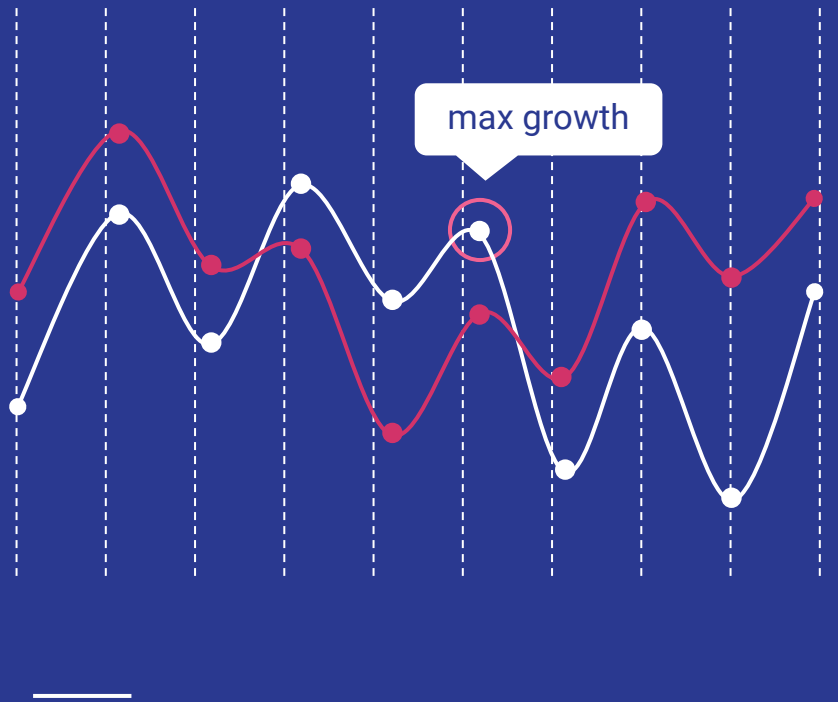
- ➔ Plotted trade-off curve between precision and recall.
- ➔ Assigned lead score to the test data.
- ➔ Created confusion matrix for Train and Test data.
- ➔ Comparing the values obtained for Train & Test:
  - ◆ Train Data:
    - Accuracy : 81.6 %
    - Sensitivity : 75.7%
    - Specificity : 85.2 %
  - ◆ Test Data:
    - Accuracy : 81.6 %
    - Sensitivity : 76.1 %
    - Specificity : 84.7 %

Observations

- ➔ 368 leads were hotleads whose lead score is greater than 85.
- ➔ We have achieved our goal of getting a ballpark of the target lead conversion rate to be around 80% .
- ➔ The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate of 80%

Impact

- There are 368 leads which can be contacted and have a high chance of getting converted, whose lead score is greater than 85.
- We have achieved our goal of getting a ballpark of the target lead conversion rate to be around 80% .
- The Model seems to predict the Conversion Rate very well and we should be able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate of 80%



Recommendations

- ➔ The company **should** make calls to leads from below sources as they are **more likely** to get converted:
  - Leads coming from the lead sources "Welingak Websites" and "Reference".
  - Leads who are the "working professionals".
  - Leads who spent "more time on the websites".
  - Leads coming from the lead sources "Olark Chat".
  - Leads whose “last activity” was “Others” and “SMS Sent”.
- ➔ The company **should not** make calls to the leads from below sources as they are **not likely** to get converted:
  - Leads whose “last activity” was "Olark Chat Conversation"
  - Leads whose “lead origin” is "Landing Page Submission"
  - Leads whose “Specialization” was "Others".
  - Leads who chose the option of "Do not Email" as "yes"

Thank You