Lead Scoring Case Study

SUBMITTED BY

MINI SAXENA

RACHIN SALIM

Table of Contents

Problem Statement

- X Education is an online Educational Company that sells online courses to industry professionals. The company markets its courses on several websites and search engines like Google.

- Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.

- Although 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.

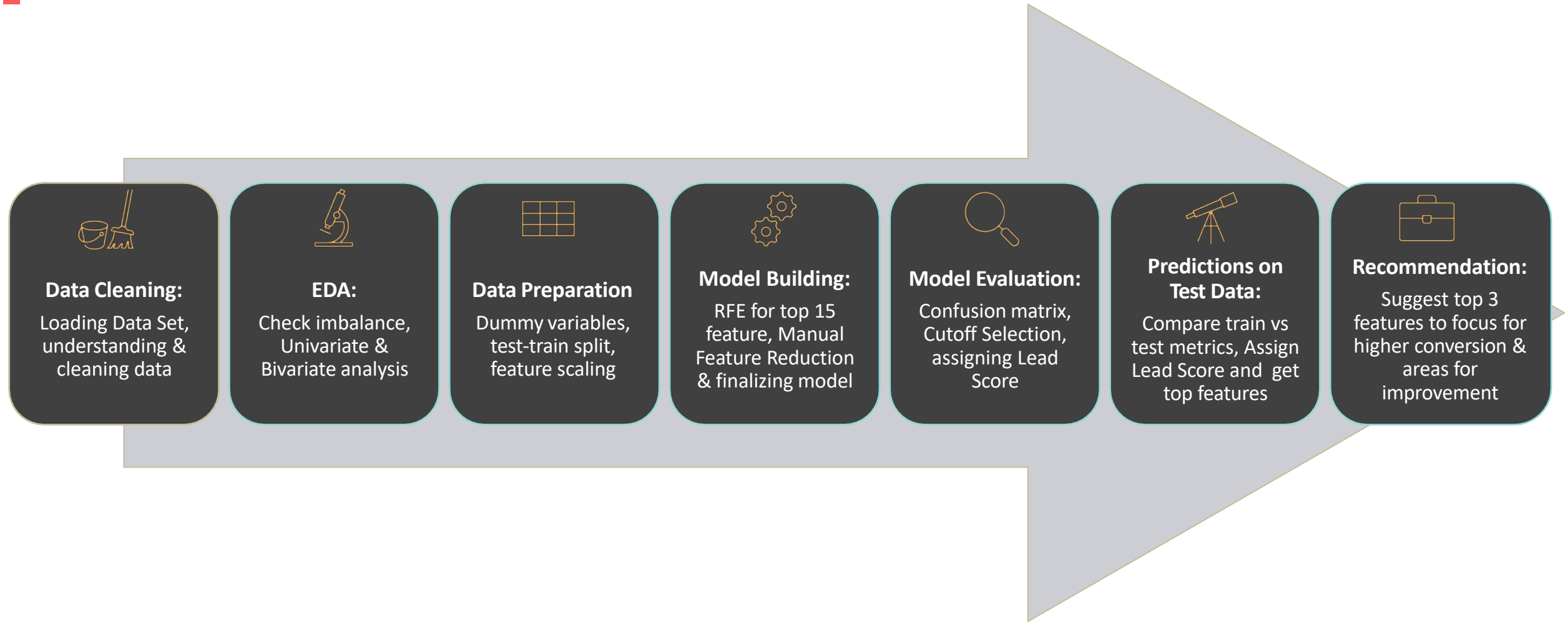
Business Objective of the Study

- X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers.

- 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 a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

- The CEO has given a ballpark of the target lead conversion rate to be around 80%.

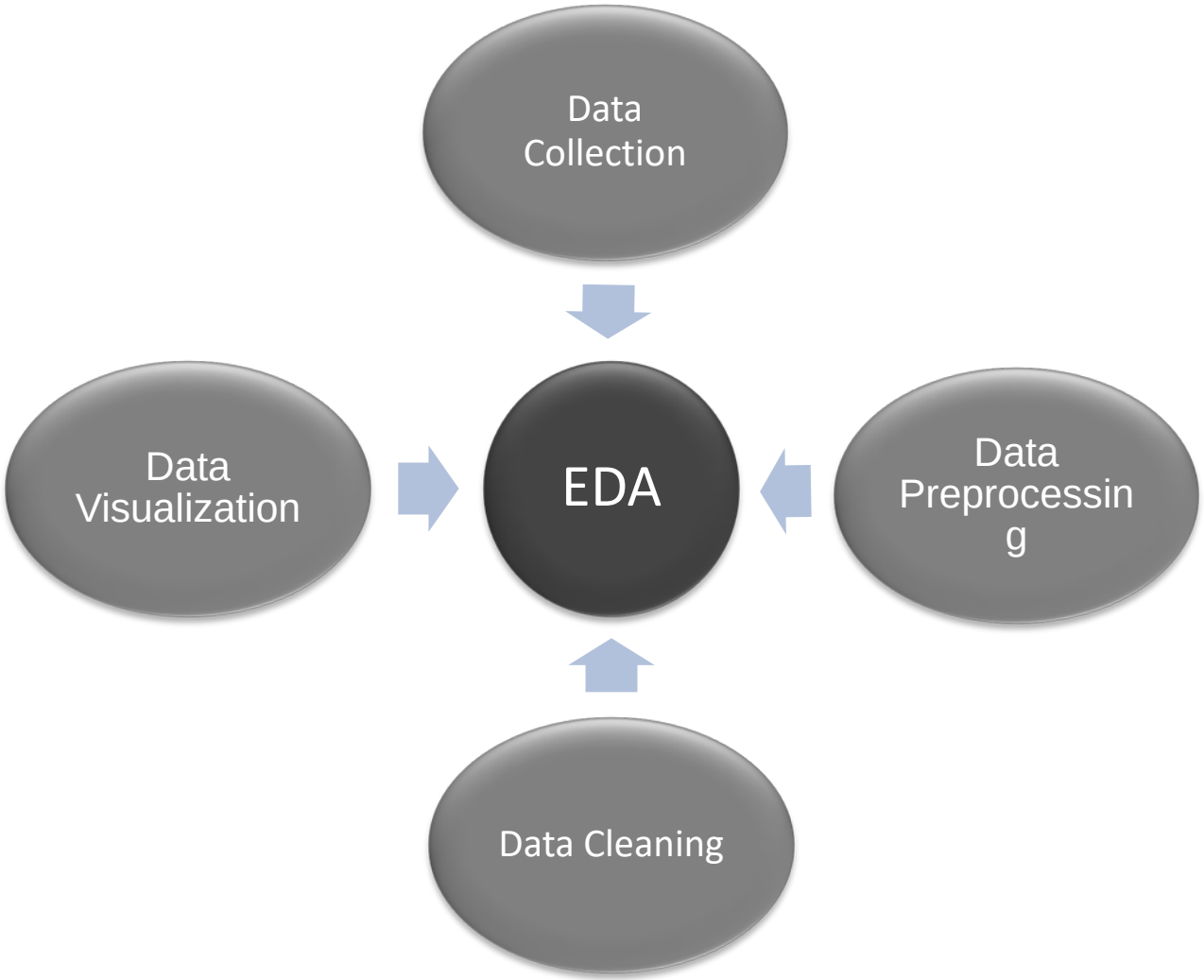
Analysis Approach



Data Cleaning and Data Preparation

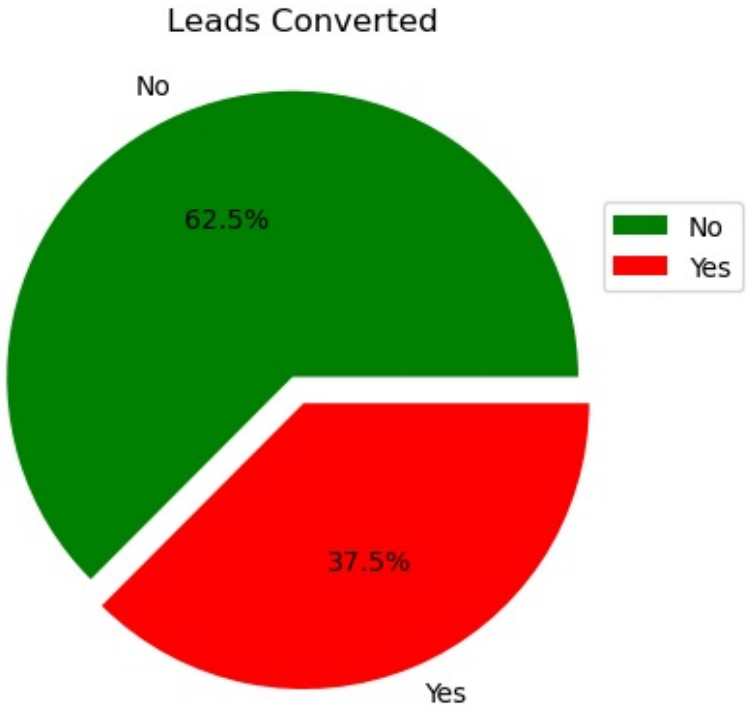
- Import Data from the Source
- Data Understanding
- Check and handle duplicate data.
- Check and handle NAN values and missing values.
- Drop columns, if it contains more than 40% of missing values.
- Impute the values, if necessary.
- Check and handle outliers in data.
- Drop columns, that are not useful for the analysis.

Exploratory Data Analysis(EDA)



EDA

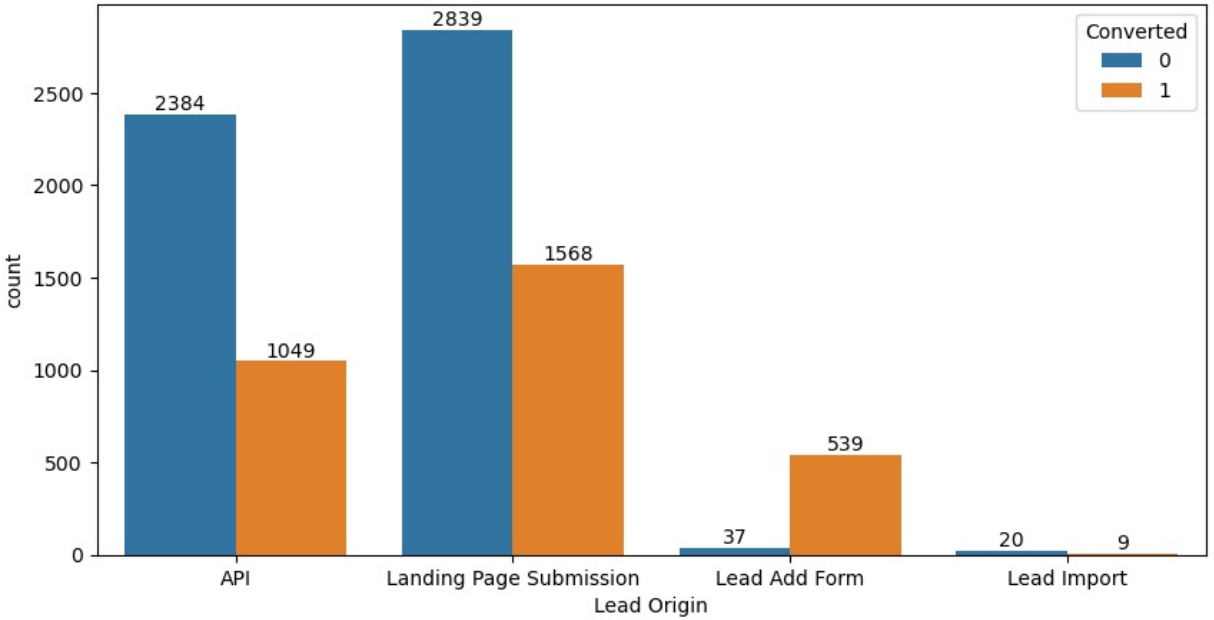
- Univariate data analysis: value count, distribution of variables, etc.
- Bivariate data analysis: correlation coefficients and pattern between the variables etc.
- Heat map for finding Correlation.

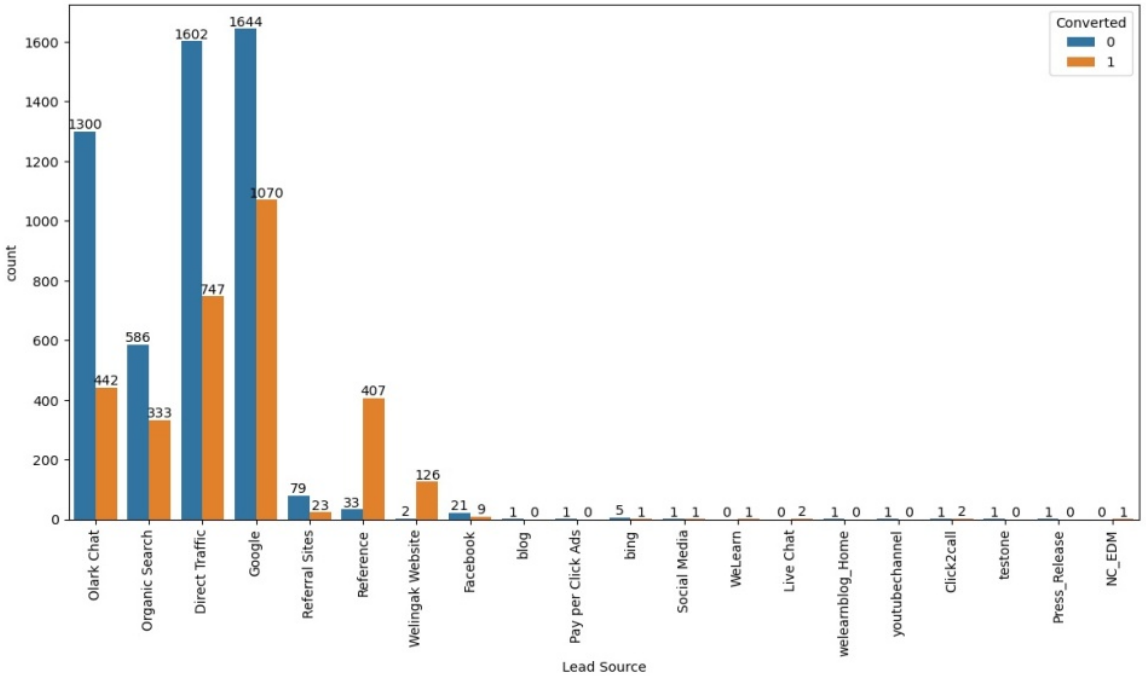


- Conversion rate is of 37.5%, meaning only 38.5% of the people have converted to leads.
- While 62.5% of the people didn't convert to leads.

EDA

Univariate Analysis – Categorical Variables





Lead Origin:

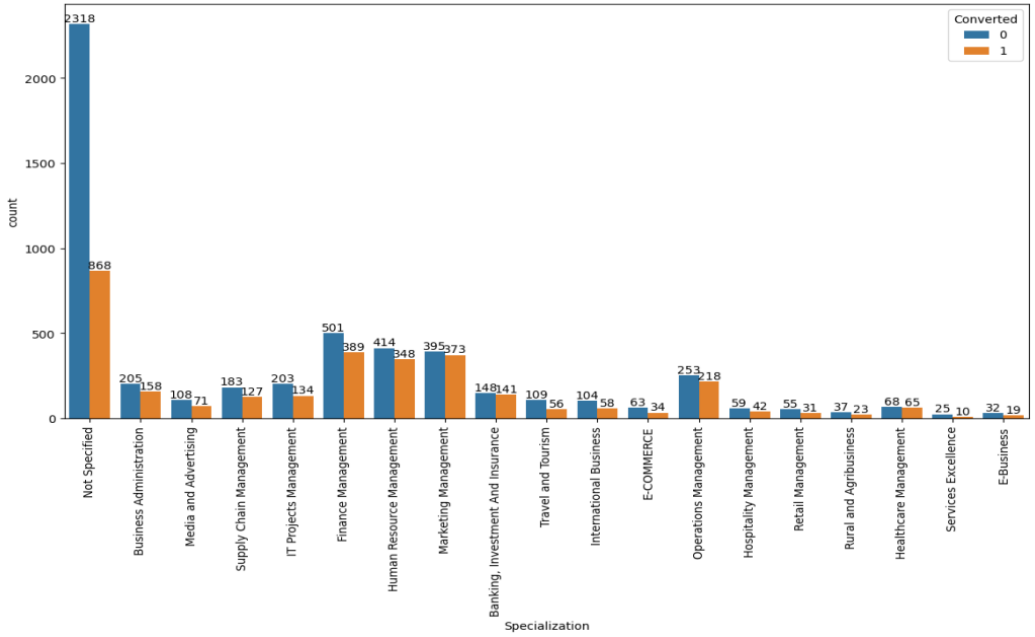
- Landing Page Submission and API has the highest lead count, but their conversion rate is within a range of 30-40%.
- Lead Add Form has the highest conversion rate of more than 93% but overall lead count is less.
- Lead import has the least lead count.

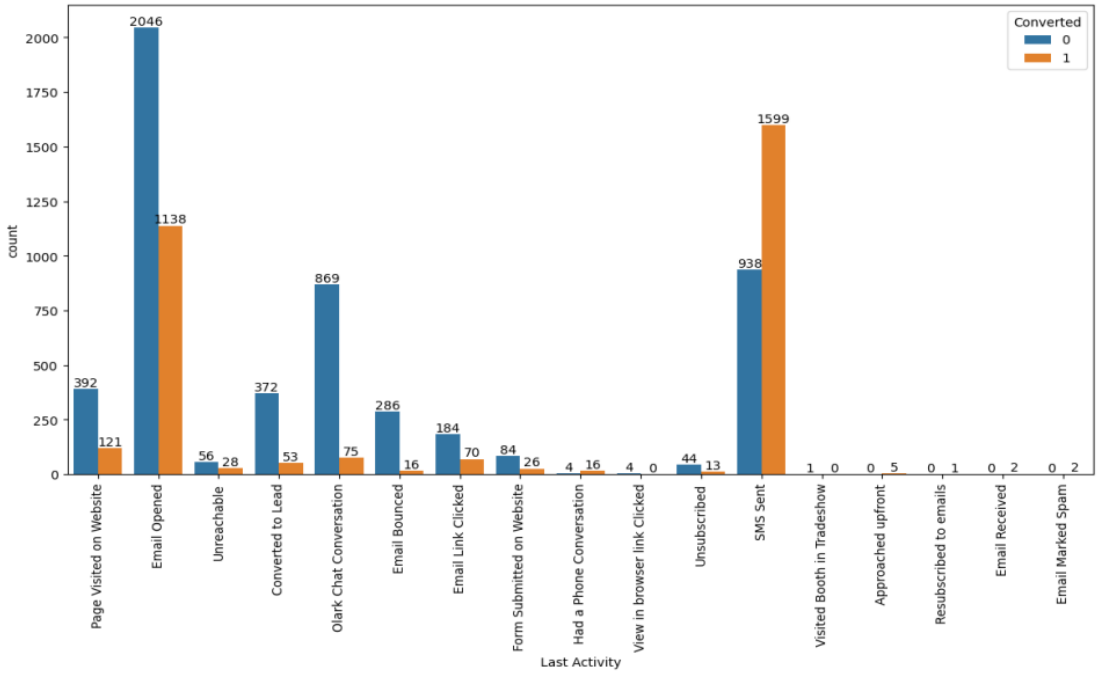
Lead Source:

- Google and Direct Traffic has the highest lead count but the conversion rate is within a range of 30-40%.
- Reference has the highest conversion rate of more than 92% but overall lead count is less.

EDA

- Univariate Analysis - Categorical Variables





Specialization:

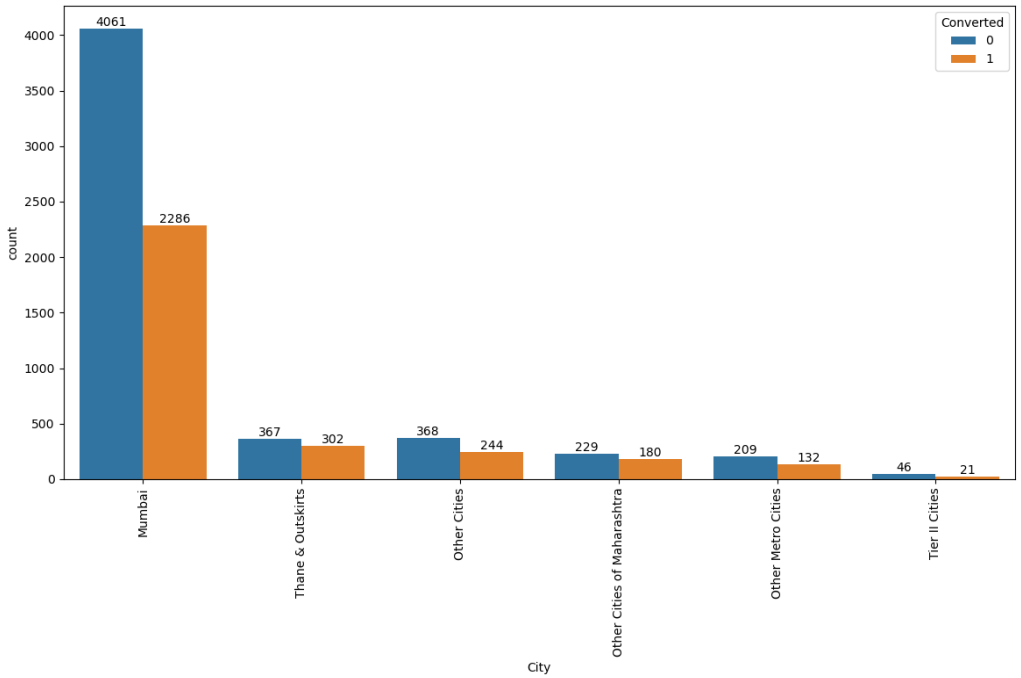
Specializations having management roles has higher leads and their conversion rates are also better.

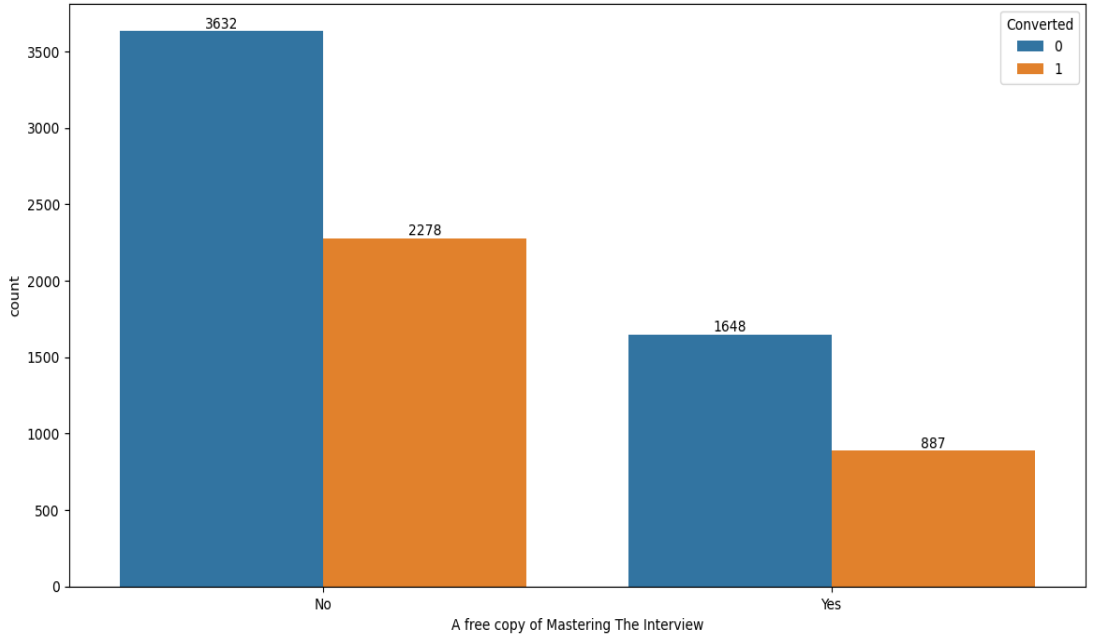
Last Activity:

- Email opened has the highest lead count but the conversion rate is within a range of 30-40%.
- SMS sent has the highest conversion rate of more than 62% with second highest lead count

EDA

- Univariate Analysis – Categorical Variables





City:

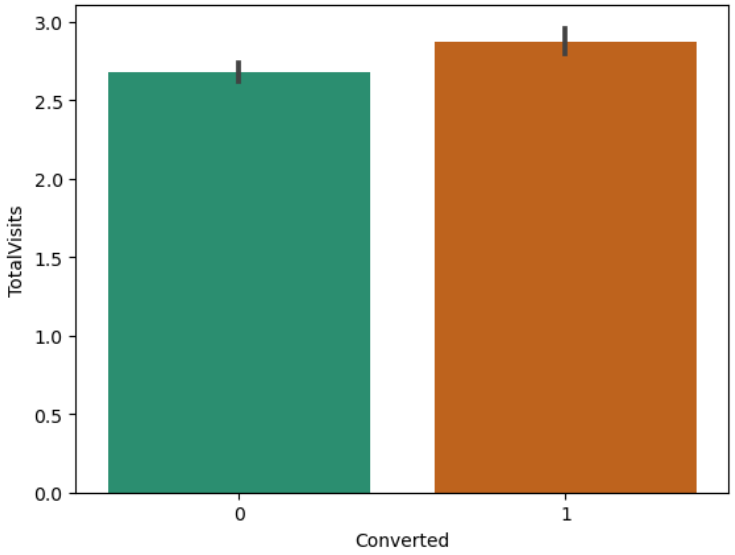
- Mumbai has the highest lead count, but the conversion rate is just above 36%.
- For all others conversion rate is better but lead count is very less compared to Mumbai

A free copy of Mastering The Interview:

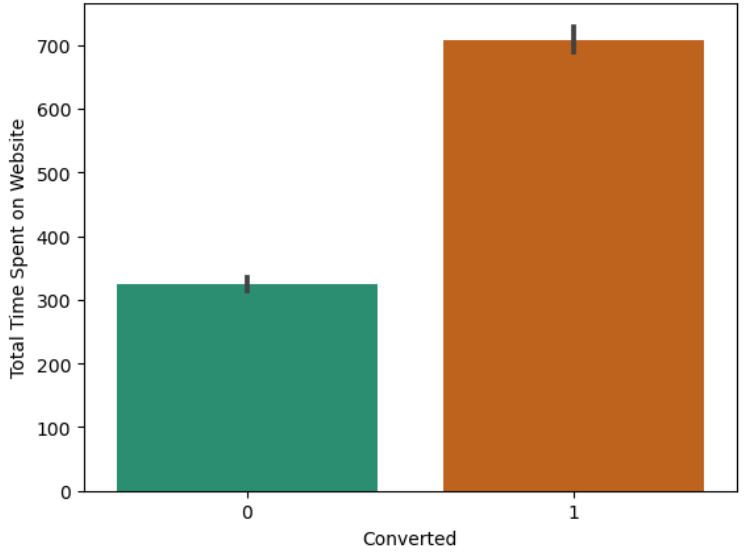
Although No has more lead count, the conversion rate is similar for both customers who wants a free copy of Mastering The Interview or not..

EDA

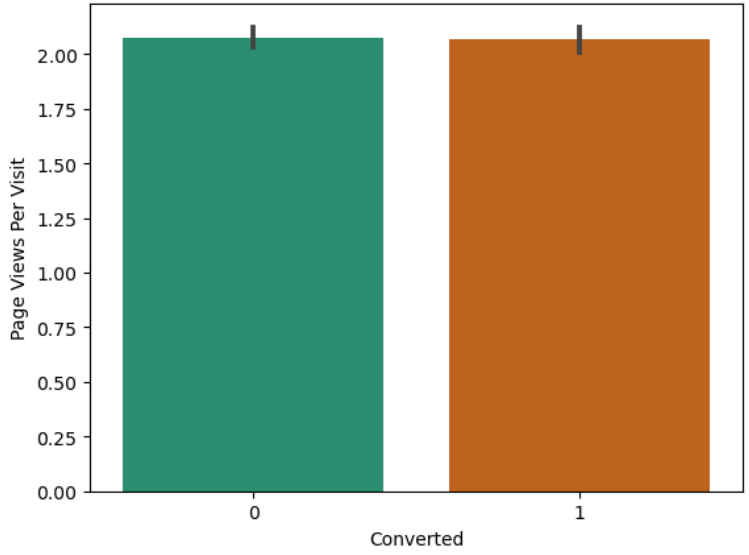
- Bivariate Analysis – Categorical Variables



Total Visits



Total Time Spend on Website



Pages per Visit

- Customers/Leads with more Time Spent on Website are more likely to be converted.
- Customers/Leads with more visits on the website are more likely to be converted.
- There is no noticeable difference in customer/leads conversion rate W.R.T Page Views Per Visit

Data Preparation for Model Building, Test-Train Split and Feature Scaling

- Converting Binary Variables into 1 and 0.
- Creating Dummy variables for categorical variables.
- Concat the Dummy variables to the Cleaned Dataset.
- Dropping the first columns and columns for which dummies were created
- Splitting the dataset into Test and Train Data.
- Do feature Scaling on the scaling required variables.

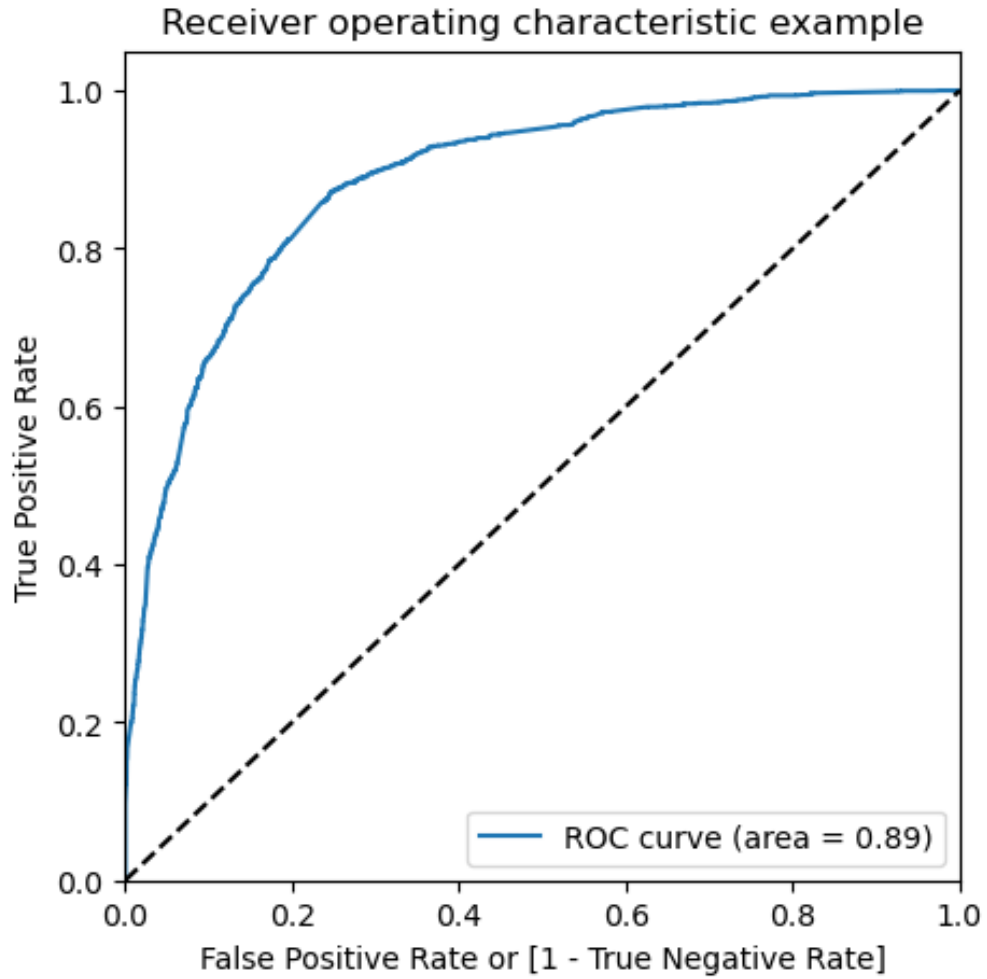
Model Building

- Create a Logistic regression model using the prepared data.
- Use RFE to select the best 15 variables.
- Assessing the model with StatsModels
- Creating a data frame that contain the names of the feature variables and their respective VIFs
- Drop variable having VIF> 5 and create a new model.
- Repeat the above steps and prepare the final model with variables having P-Value < 0.05 and VIF < 5.
- Create a confusion matrix and find the overall accuracy with a cut-off value of 0.5.

Model Evaluation

Evaluating the model by using different metrics - Specificity and Sensitivity

- Finding the Sensitivity and Specificity.
- Plotting the ROC curve

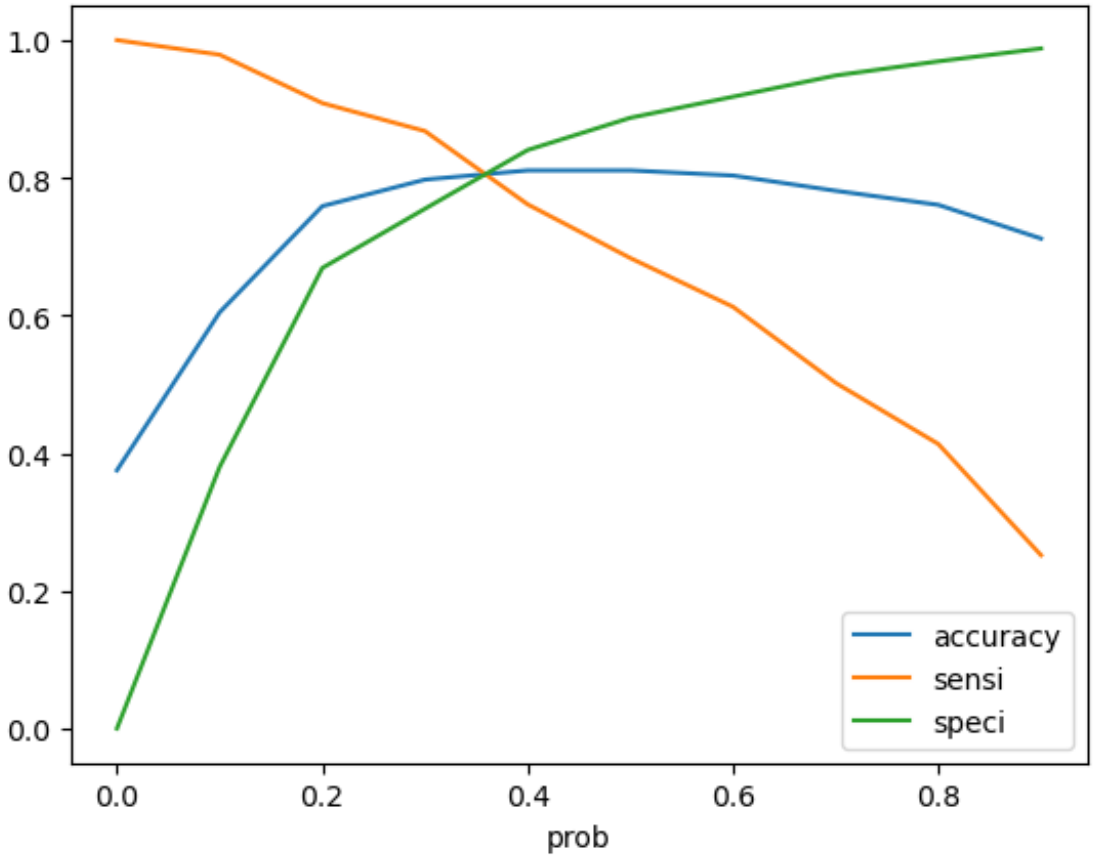


Model Evaluation

Evaluating the model by using different metrics - Specificity and Sensitivity

Finding Optimal Cut off value by plotting accuracy sensitivity and specificity for various probabilities.

The plot is showing an optimal cut off 0.35 based on Accuracy, Sensitivity and Specificity.

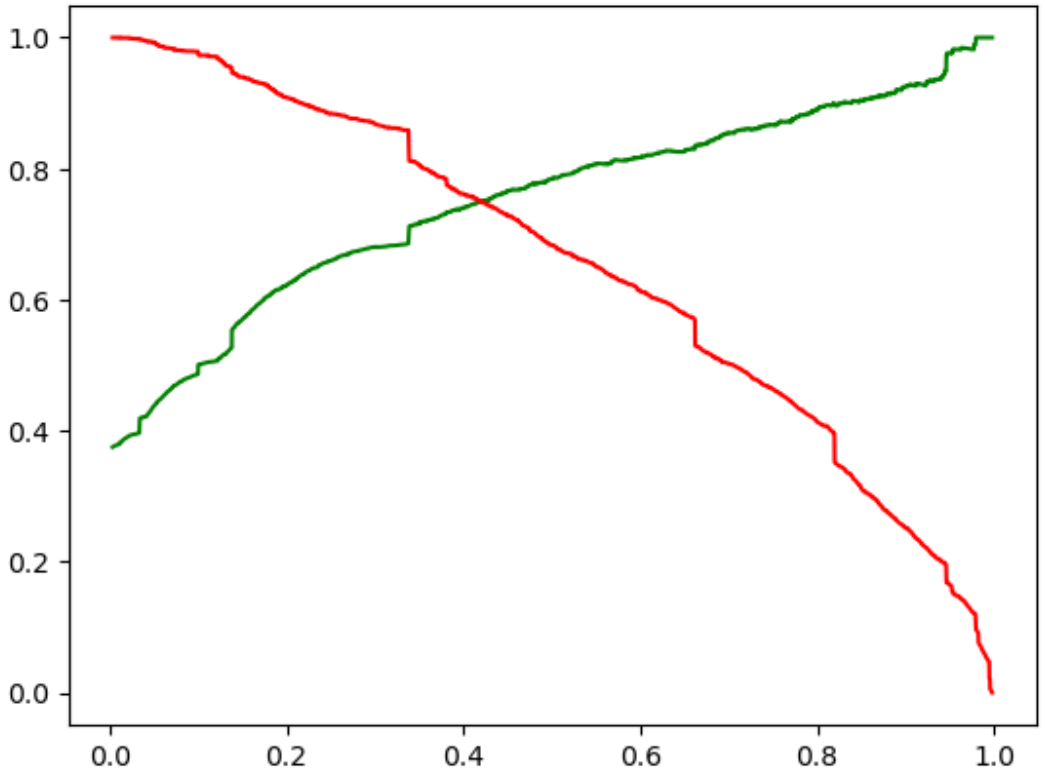


Model Evaluation

Evaluating the model by using different metrics - Precision and Recall

Finding the Precision and Recall.

The plot is showing an optimal cut off value of 0.4 based on Precision and Recall



Model Evaluation

Evaluating the model by using different metrics - Specificity and Sensitivity or Precision and Recall

Final Confusion Matrix.

3277	417
702	1515

Accuracy: 80.76%

Sensitivity: 80.55%

Specificity: 80.88%

Precision: 78.41%

Recall: 68.33%

Model Evaluation

Sensitivity and Specificity on Test Dataset

Confusion Matrix.

Accuracy: 81.93%

Sensitivity: 80.55%

Specificity: 80.88%

1333	253
205	743

Conclusion

- While we have checked both Sensitivity-Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
- Accuracy, Sensitivity and Specificity values of test set are around 82%, 81% and 82% which are approximately closer to the respective values calculated using trained set.
- Also the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 78%
- There are 511 Hot Leads. They should be targeted as they have a high chance of getting converted
- Hence overall this model seems to be good.

Recommendation based on Final Model

- As per the problem statement, increasing lead conversion is crucial for the growth and success of X Education. To achieve this, we have developed a regression model that can help us identify the most significant factors that impact lead conversion.
- We have determined the following features that have the highest positive coefficients, and these features should be given priority in our marketing and sales efforts to increase lead conversion.
  - Lead Origin_Lead Add Form: 3.49
  - Lead Source_Welingak Website: 2.49
  - Current_occupation_Working Professional: 2.31

- We have also identified features with negative coefficients that may indicate potential areas for improvement. These include:
  - Last Activity_Olark Chat Conversation : -1.52
  - Last Activity_Email Bounced : -1.42

Recommendation based on Final Model

To increase our Lead Conversion Rates

- Focus on features with positive coefficients for targeted marketing strategies.
- Develop strategies to attract high-quality leads from top-performing lead sources.
- Optimize communication channels based on lead engagement impact.
- Engage **working professionals** with tailored messaging.

- More budget/spend can be done on **Welingak Website** in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage providing more references.

To identify areas of improvement

- Analyze negative coefficients in specialization offerings.

Thank You!