X Education - Lead Scoring Case Study

IDENTIFICATION OF HOT LEADS TO FOCUS MORE ON THEM AND THUS LIFTING THE CONVERSION RATIO FOR X EDUCATION FROM 30 TO 80%

Group Members:
- Ranabir Khan
- Ketan Chinchalkar
- Kamal Sharma

Background

X Education Company

- X Education , An education company named sells online courses to industry professionals
- Many interested professionals land on their website
- The company markets its courses on several websites 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

Background

X Education Company

- 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%

Problem Statement

X Education Company

- X Education gets a lot of leads but its lead conversion rate is very poor
- 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

Problem Statement

X Education Company

- We will help them to select the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- We are required to build a model wherein we need to assign a lead score to each of the leads such that the customers with higher lead score have a higher conversion chance
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be 80%.

Lead – Conversion Process

Proposed Solution

Selection of Hot Leads → Communicating with Hot Leads → Conversion of Hot Leads

Leads Clustering

We cluster the leads into certain categories based on their tendency or probability to convert, thus, getting a smaller section of hot leads to focus more on.

Focus Communication

Since we would have a smaller set of leads to have communication with, we might make more impact with effective communication.

Increase conversion

Since we focused on hot leads, which were more probable to convert, we would have a better conversion rate, and hence we can achieve the 80% target.

Solution

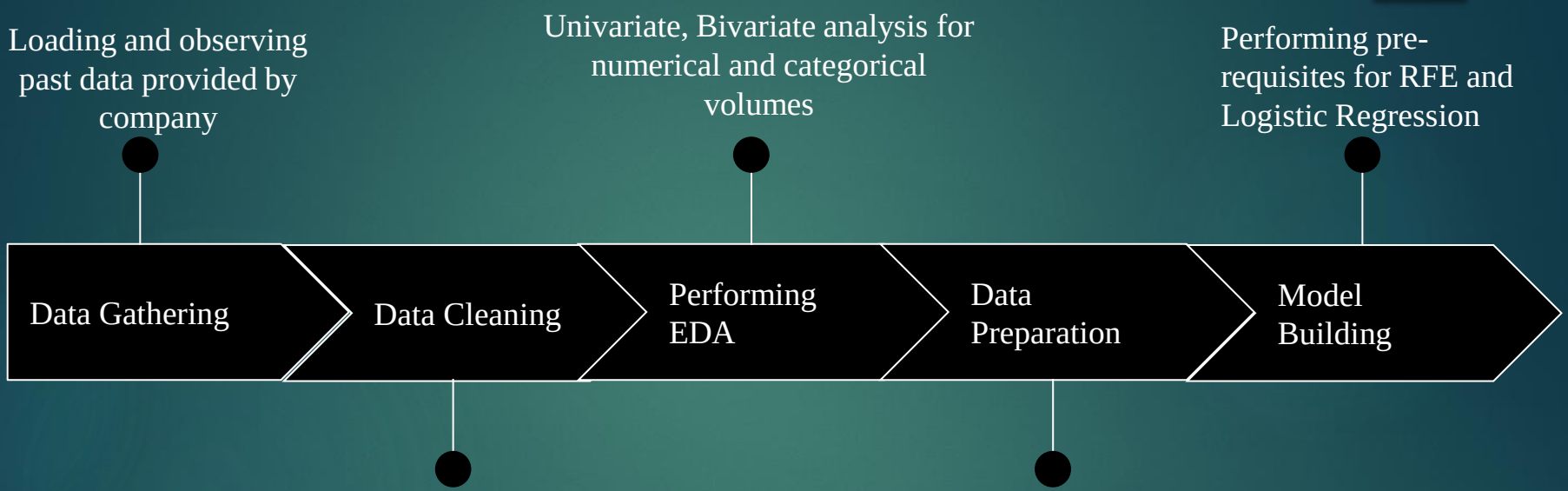
Selection of Hot Leads

For our Problem Solution, the crucial part is to accurately identify hot leads.

The more accurate we obtain the hot lead, the more chance we get of higher conversion ratio.

Since we have a target of 80% conversion rate, we would want to obtain a high recall and a good precision in obtaining hot leads.

Implementation


Loading and observing past data provided by company
Univariate, Bivariate analysis for numerical and categorical volumes
Performing pre-requisites for RFE and Logistic Regression
Data Gathering
Data Cleaning
Performing EDA
Data Preparation
Model Building
Duplicate removal, null value treatment, unnecessary column removal, etc.
Outlier Treatment, Feature-Normalization

Selection of top 20 features using RFE

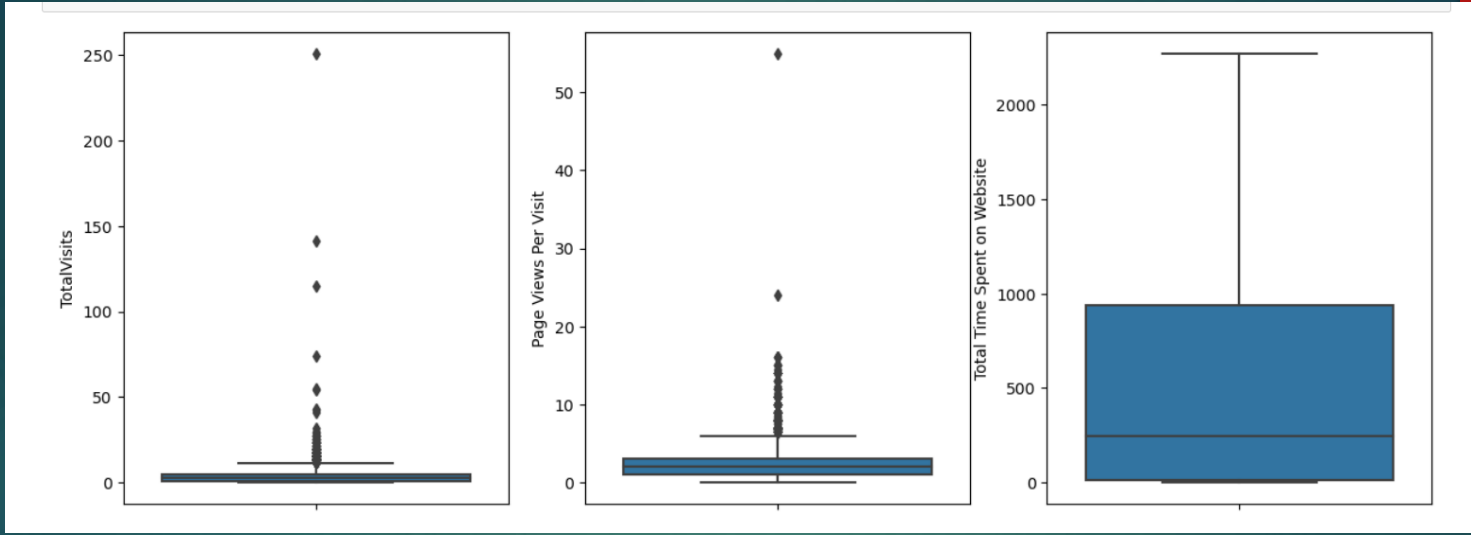
Reduction of features for model training iterations based on pvalue and VIF statistical variables

Verifying our Final Model Accuracy – Precision, Recall etc.

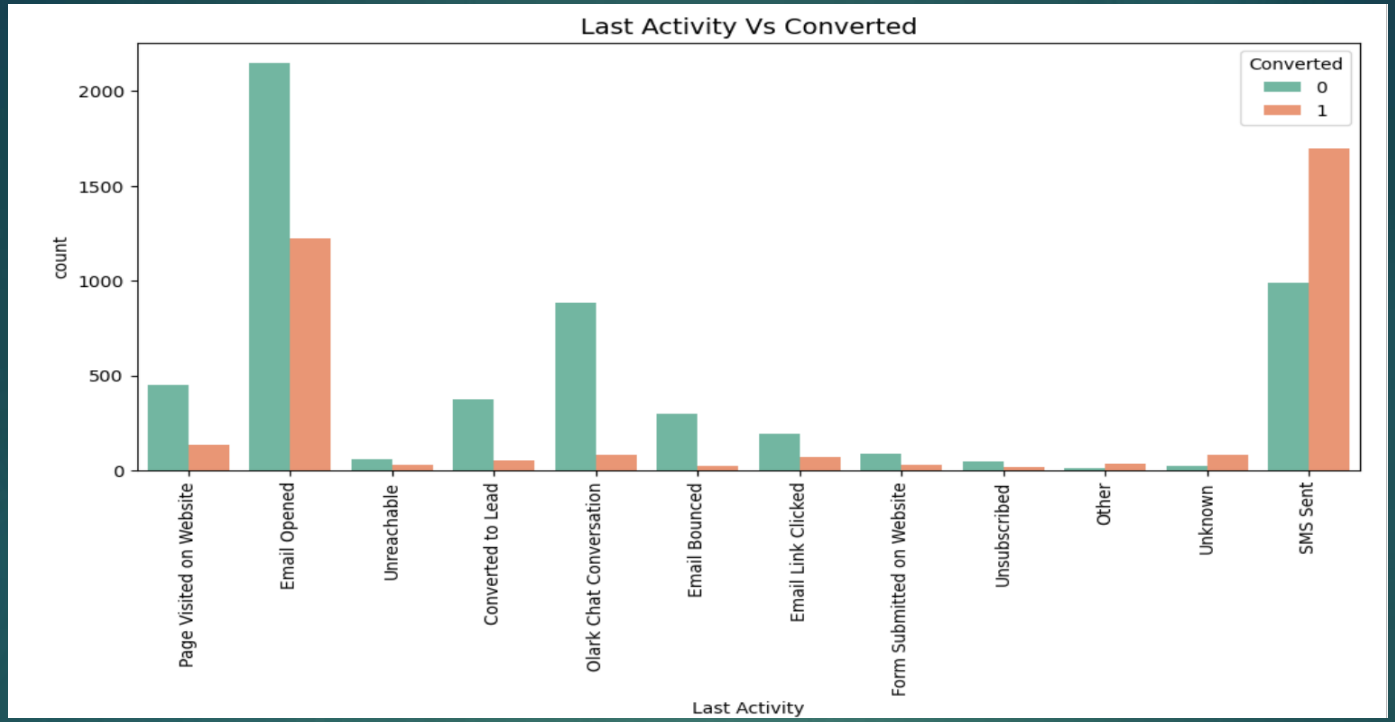
Model building using logistic regression for selected features

Final Model Analysis with training data and evaluate performance on Test Data. With ROC curve and plot of sensitivity and specificity for various cut off find the optimal cutoff

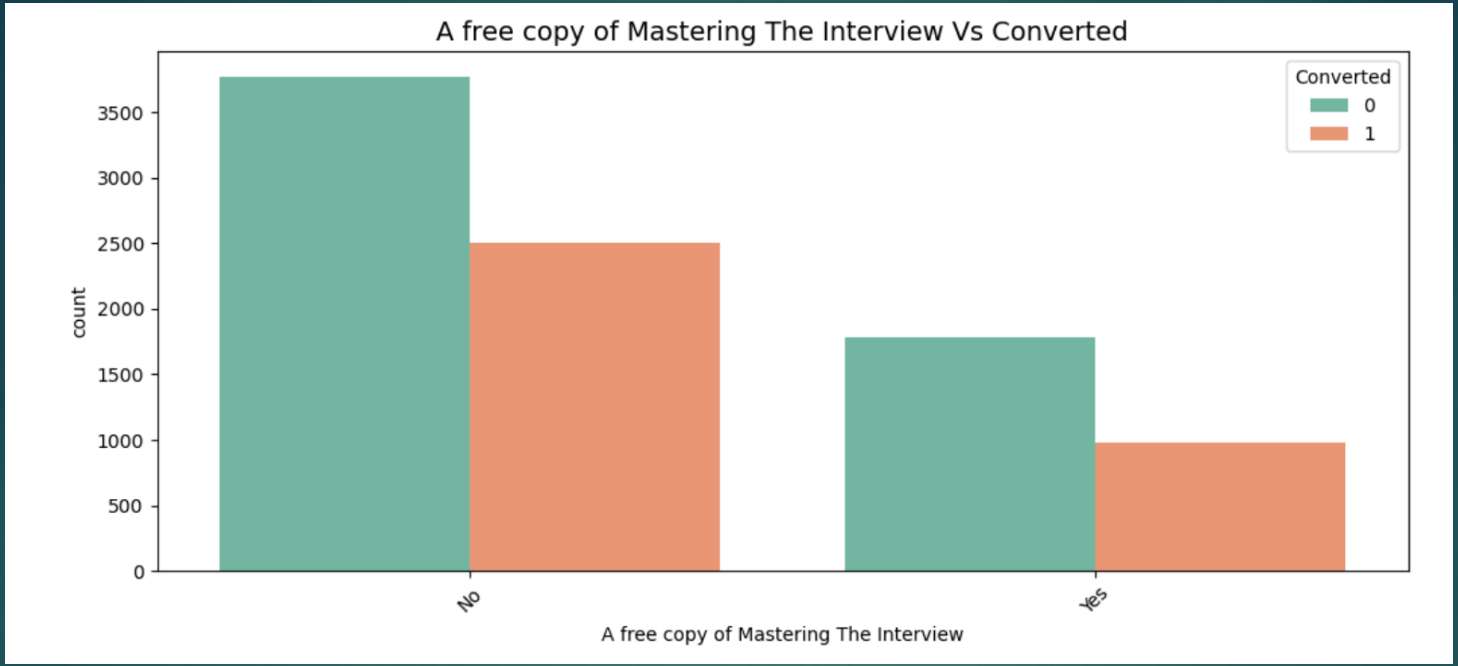
Data Visualization



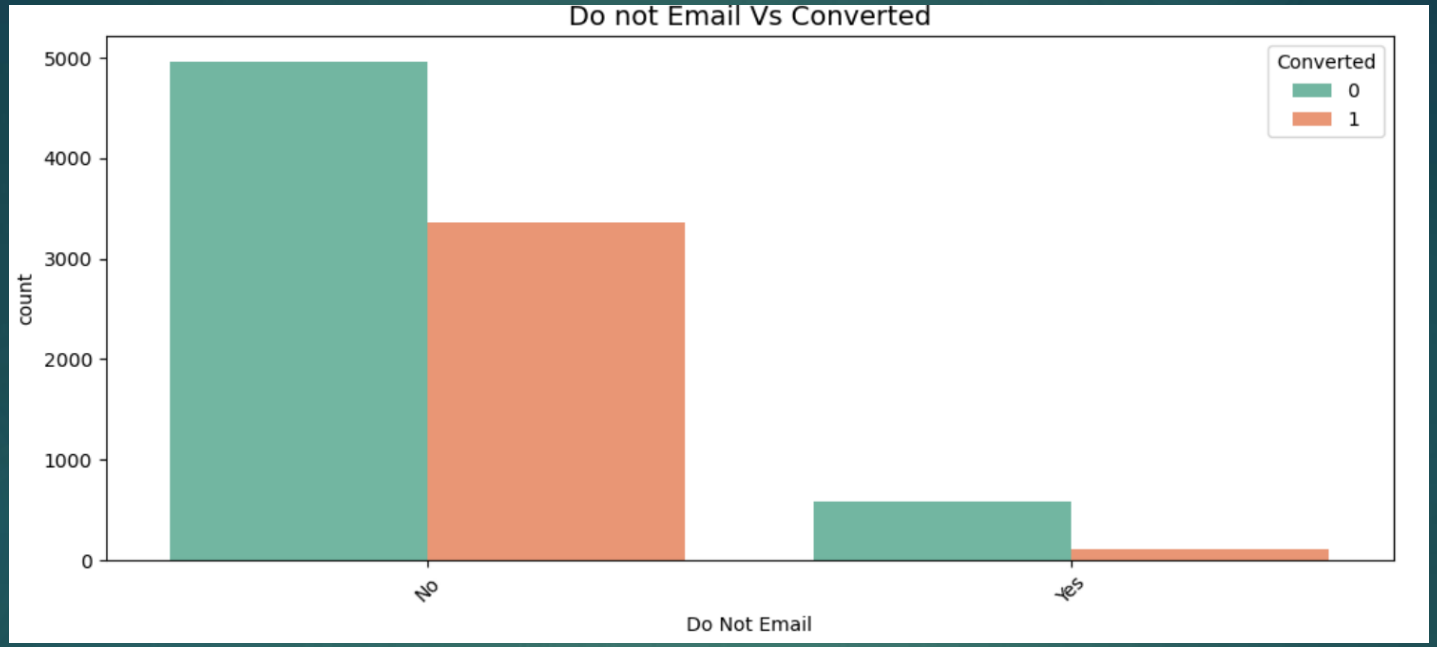
Outlier Analysis for Numerical Features



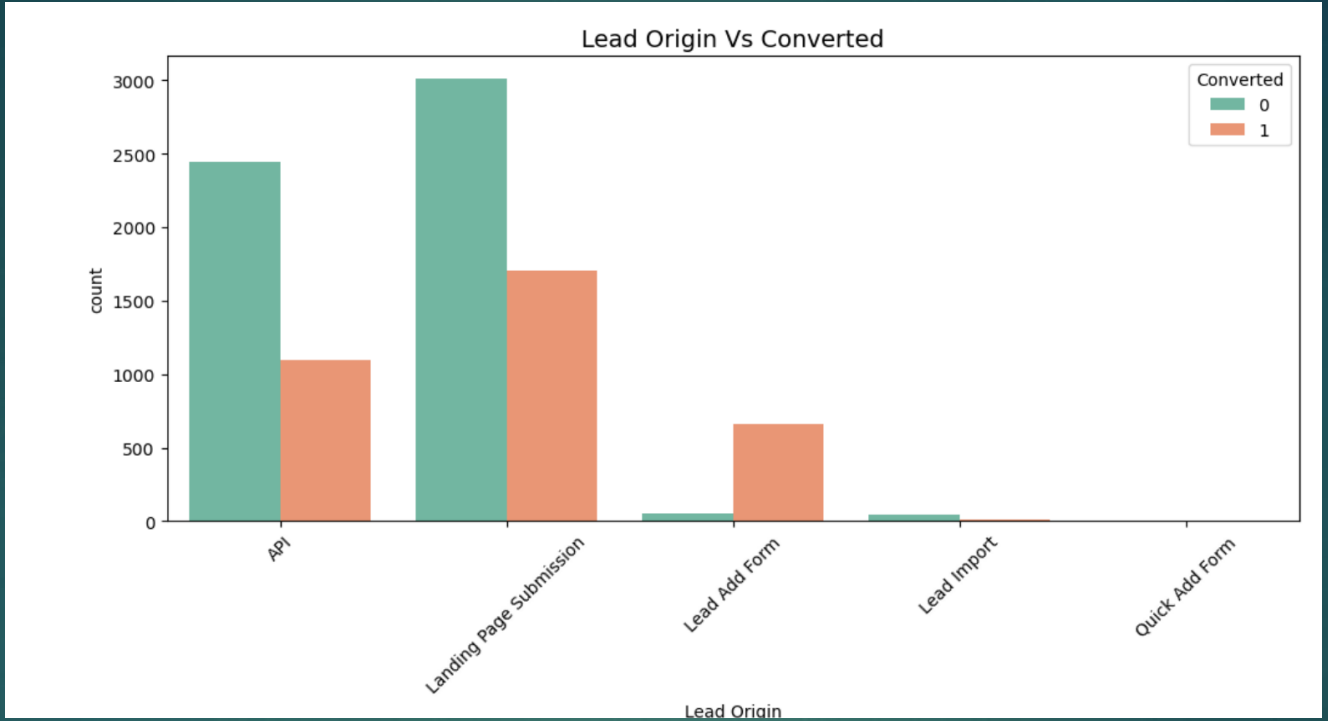
EDA plots depicting variation in categorical column (Last Activity) for those who Converted and those who didn't.



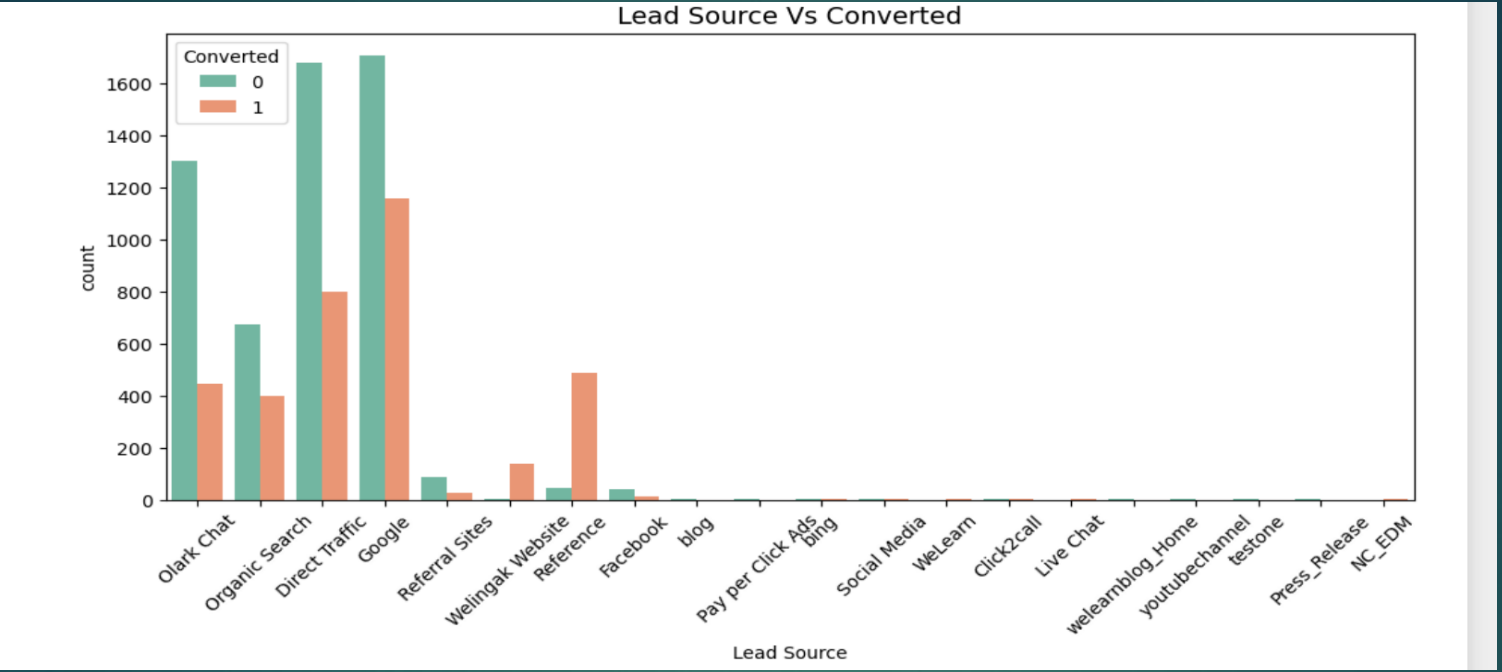
EDA plots depicting variation in categorical column (A free copy of Mastering The Interview) for those who Converted and those who didn't).



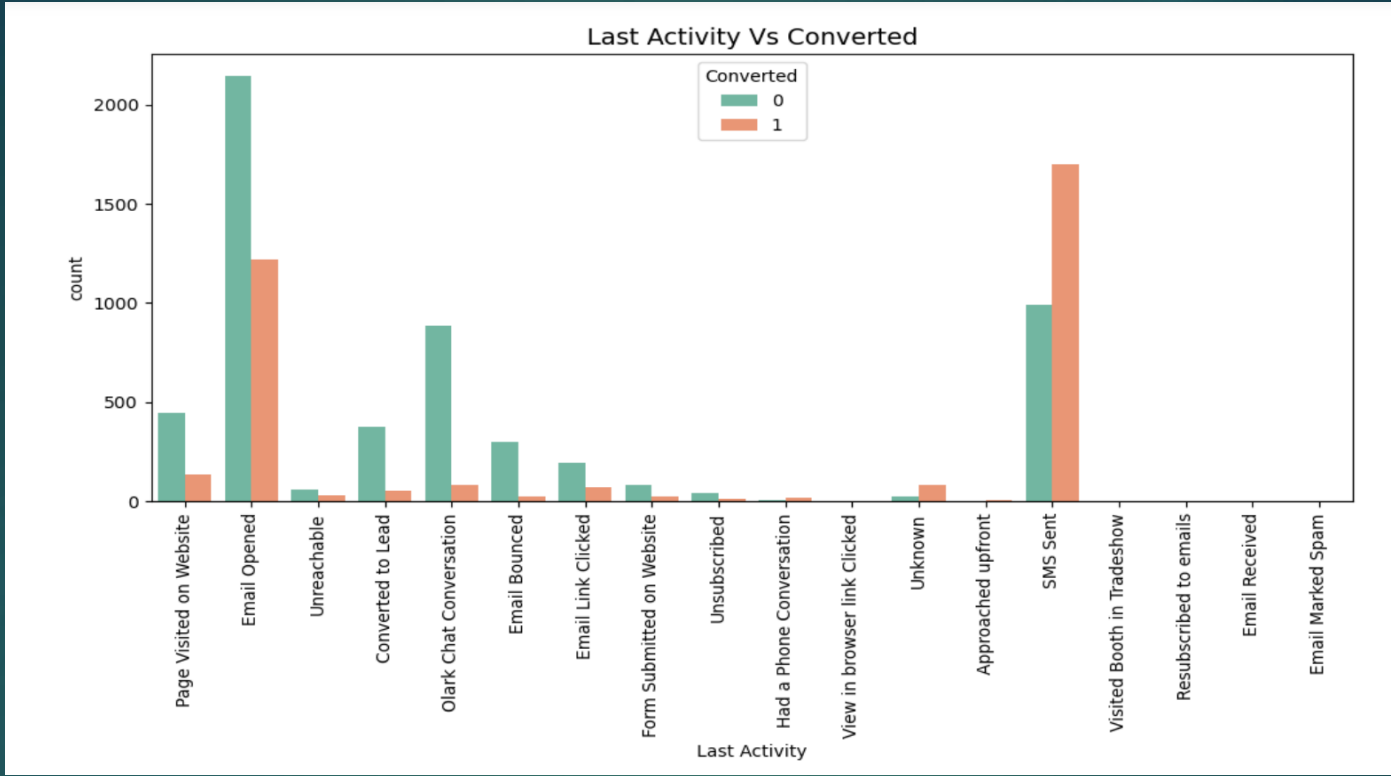
EDA plots depicting variation in categorical column (Do Not Email) for those who Converted and those who didn't.



EDA plots depicting variation in categorical column (Lead Origin) for those who Converted and those who didn't.

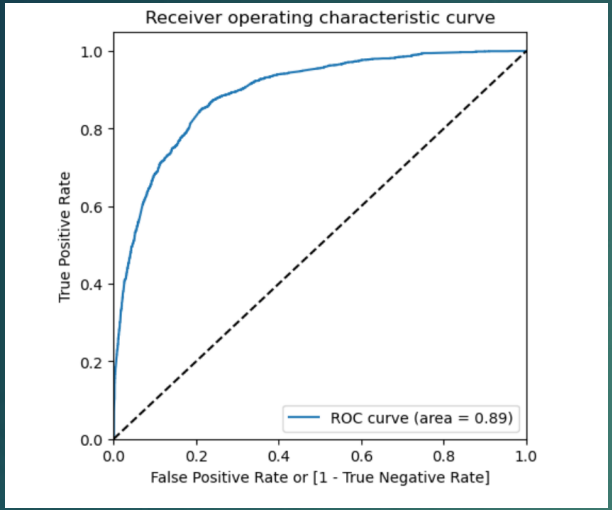


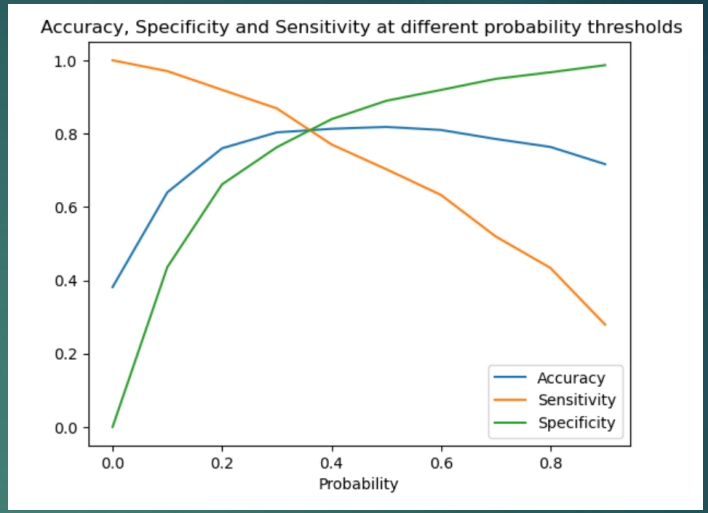
EDA plots depicting variation in categorical column (Lead Source) for those who Converted and those who didn't.



EDA plots depicting variation in categorical column (Last Notable Activity) for those who Converted and those who didn't.

Model ROC curve and Threshold





Logistic Regression Final Model Parameters

Area under ROC = 0.89

Probability cut-off = 0.36

Inference / Conclusion

Model Analysis

Performance of our Final Model

Overall accuracy on Test set: 0.816

Sensitivity/Recall of our logistic regression model: 0.819

Specificity of our logistic regression model: 0.813

Precision of our logistic regression model: 0.74 (which is OK as recall is important for the model)

Inferences from Model

Business Insights Derived from our Model

Top 4 variables in model, that contribute towards lead conversion are: (Sales team should focus more on leads with the below features for more conversions)

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Working Professional
- Lead Source_Welingak Website.

Conclusion

OUR LOGISTIC REGRESSION MODELBUILT IS DECENT AND ACCURATE ENOUGH WITH 81.6% ACCURACY ON TEST SET, 81.9% SENSITIVITY AND 81.3 % SPECIFICITY. THE THRESHOLD CUT OFF SUGGESTED IS 0.36.

THIS MODEL CAN BE DEPLOYED TO IDENTIFY THE HOT LEADS AND PUT THE SALES FOCUS ACCORDINGLY. X-EDUCATION CEO AND SALES TEAM SHOULD FOCUS ON THE LEADS WHICH SHOW AS TARGETED CONVERSIONS COMING OUT FROM THE MODEL AND OPTIMIZE THE SALES TEAM EFFORT AND BRING MORE CONVERSIONS