Lead Scoring Case Study

Team Members: Antara, Ashirwad, Anand and Shubhendu

Table of Contents

Background of X Education Company

- X Education is an education company that sells online courses
- Interested candidates land on the company website and browse for courses. The company also markets its courses on several websites and search engines like Google.
- Interested candidates might browse the courses or fill up a form for the course or watch some videos.
- When people fill up a form providing their email address or phone number, they are classified as a lead.
- Once 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.
- Typical lead conversion rate at X education is around 30%.

Problem Statement & Objective of the Study

Problem Statement:

- X Education's lead conversion rate is very poor at ~30%

Objective of the Study:

- The CEO has given a ballpark of the target lead conversion rate to be around 80%
- Make lead conversion process more efficient by identifying the most promising leads i.e., Hot Leads
- Help the sales team focus on communicating with the Hot Leads rather than making calls to everyone.

Analysis Approach

Data Cleaning:
Loading Data Set, understanding & cleaning data

EDA:
Check imbalance, Univariate & Bivariate analysis

Data Preparation
Dummy variables, test-train split, feature scaling

Model Building:
RFE for top 20 feature, Manual Feature Reduction & finalizing model

Model Evaluation:
Confusion matrix, Cutoff Selection, assigning Lead Score

Predictions on Test Data:
Compare train vs test metrics, Assign Lead Score and get top features

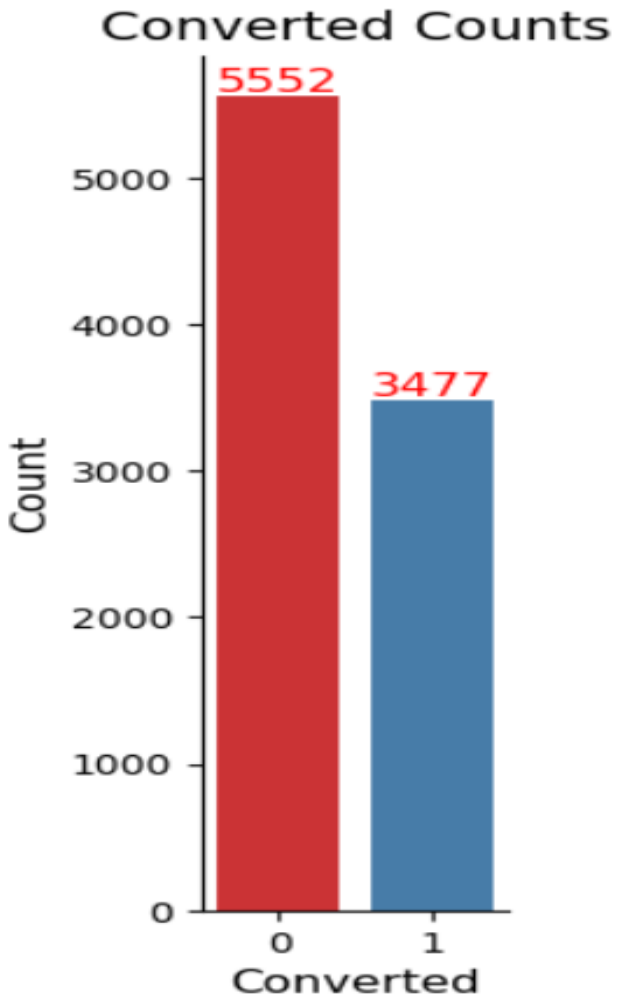
Recommendation:
Suggest top 3 features to focus for higher conversion

Data Cleaning

- "Select" level for 4 categorical variables has been replaced with Null value, as the value of “Select” does not represent any business value.
- Columns with over 30% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Dropped columns that don't add any insight or value to the study objective e.g., columns with only one value ‘No’.
- Imputation was performed with relevant statistical measure e.g., Median
- Additional categories were created for some variables with Null values e.g., Last Activity, Occupation
- Outliers in TotalVisits and Page Views Per Visit were capped at 99th percentile
- After cleaning, 98% of data has been retained

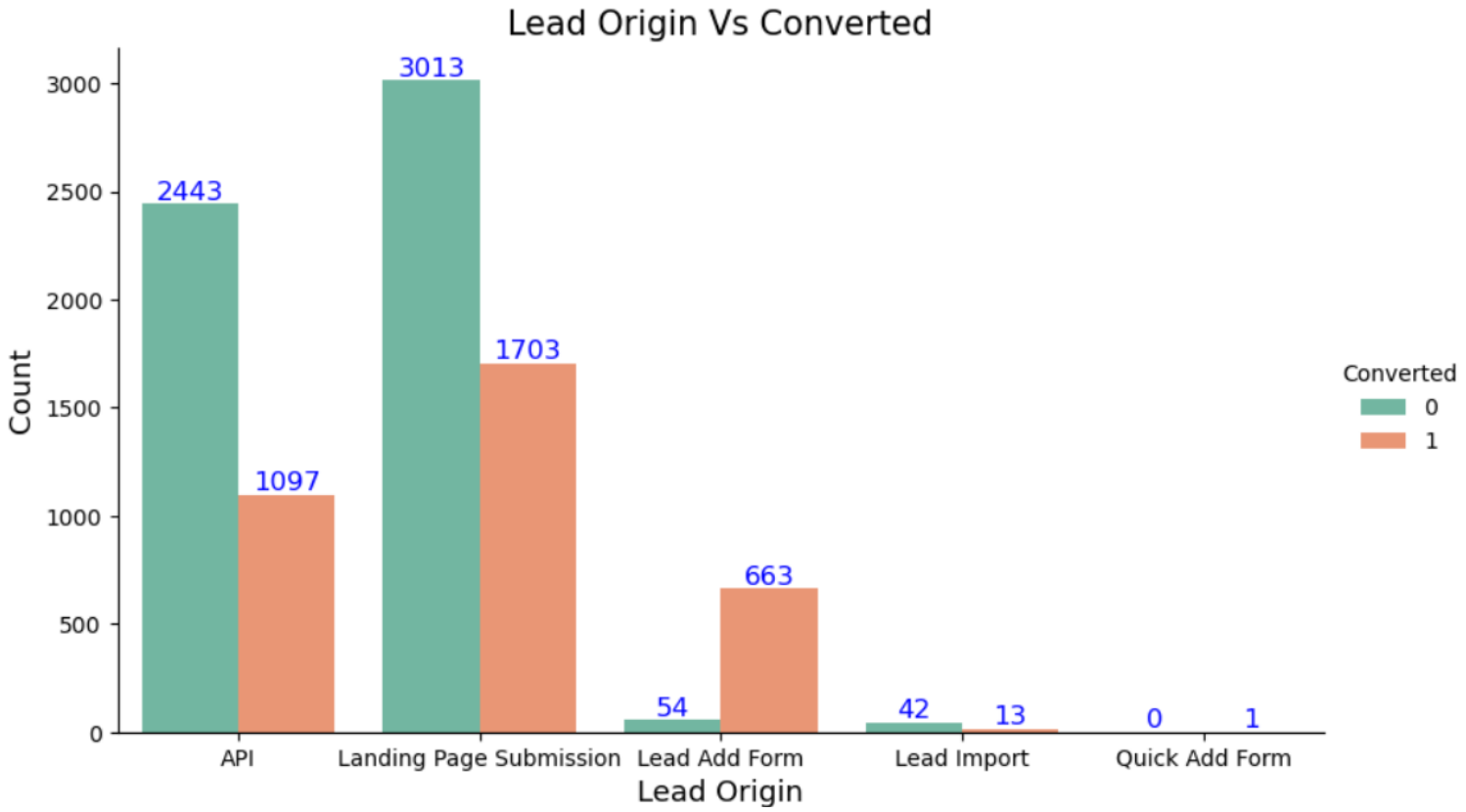
EDA - Data Imbalance

- Data is imbalanced while analyzing target variable since the cconversion rate is ~39%



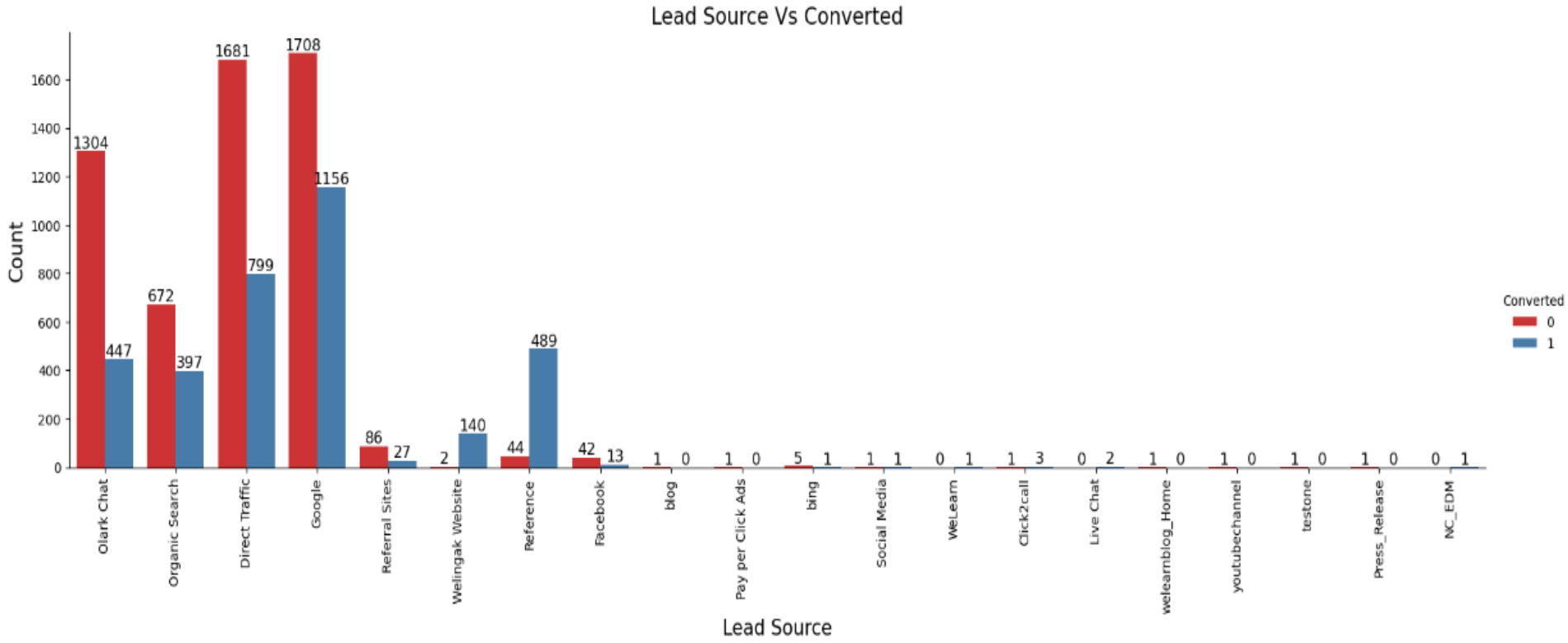
EDA - Univariate Analysis – Categorical Variables

- Maximum conversion happened from Landing Page Submissions followed by APIs



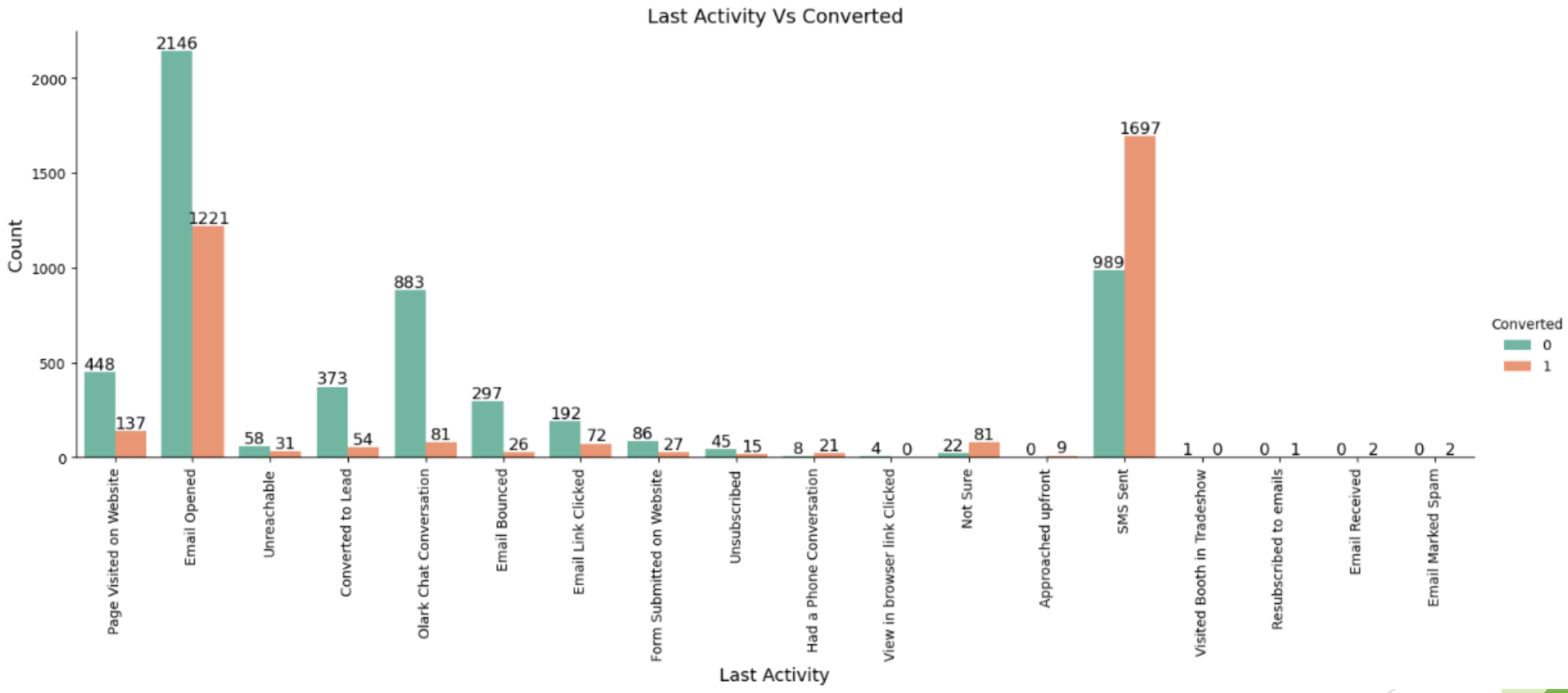
EDA - Univariate Analysis – Categorical Variables

- Maximum conversion happened from Goggle, Direct Traffic and References



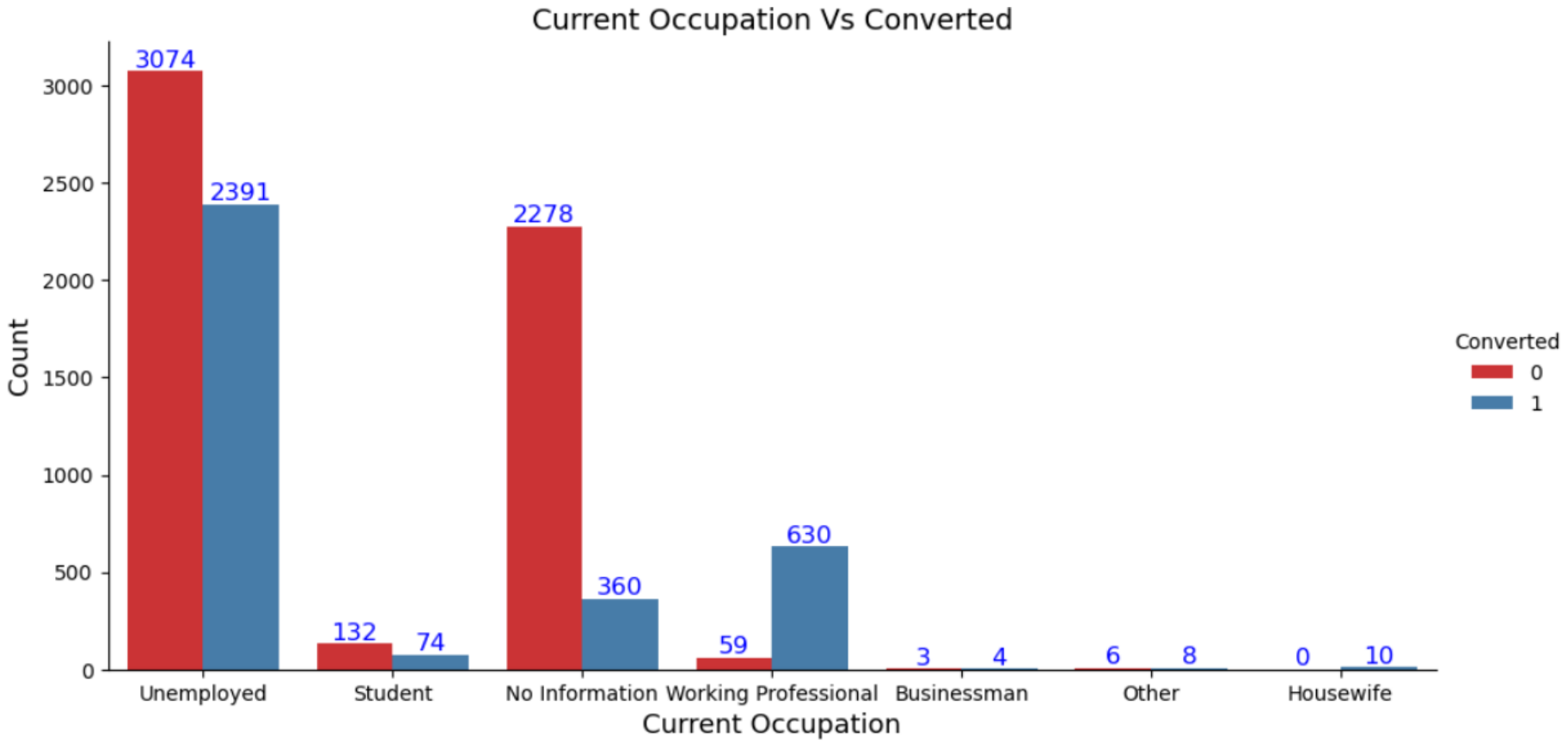
EDA - Univariate Analysis – Categorical Variables

- SMS Sent and Email Opened have the most conversions



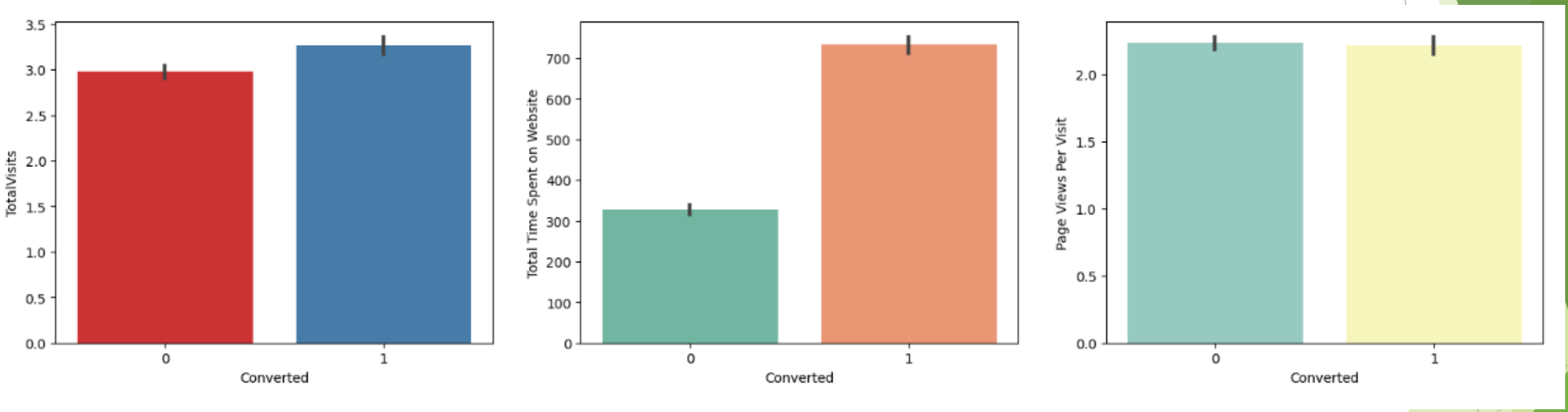
EDA - Univariate Analysis – Categorical Variables

- Unemployed and Working Professionals are the best target customers

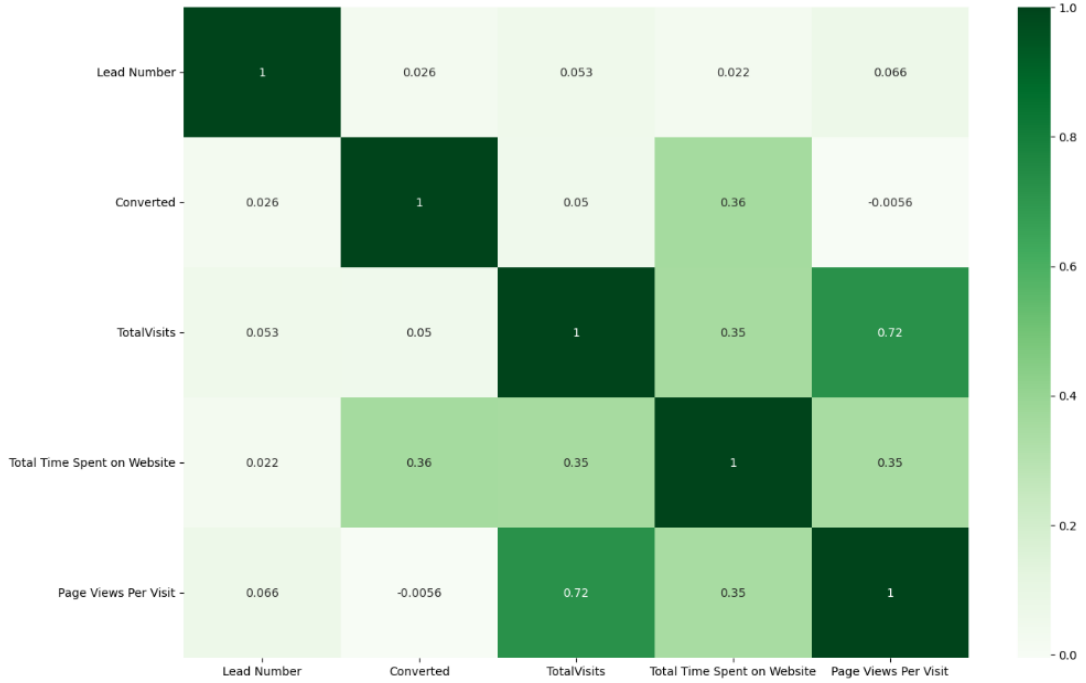


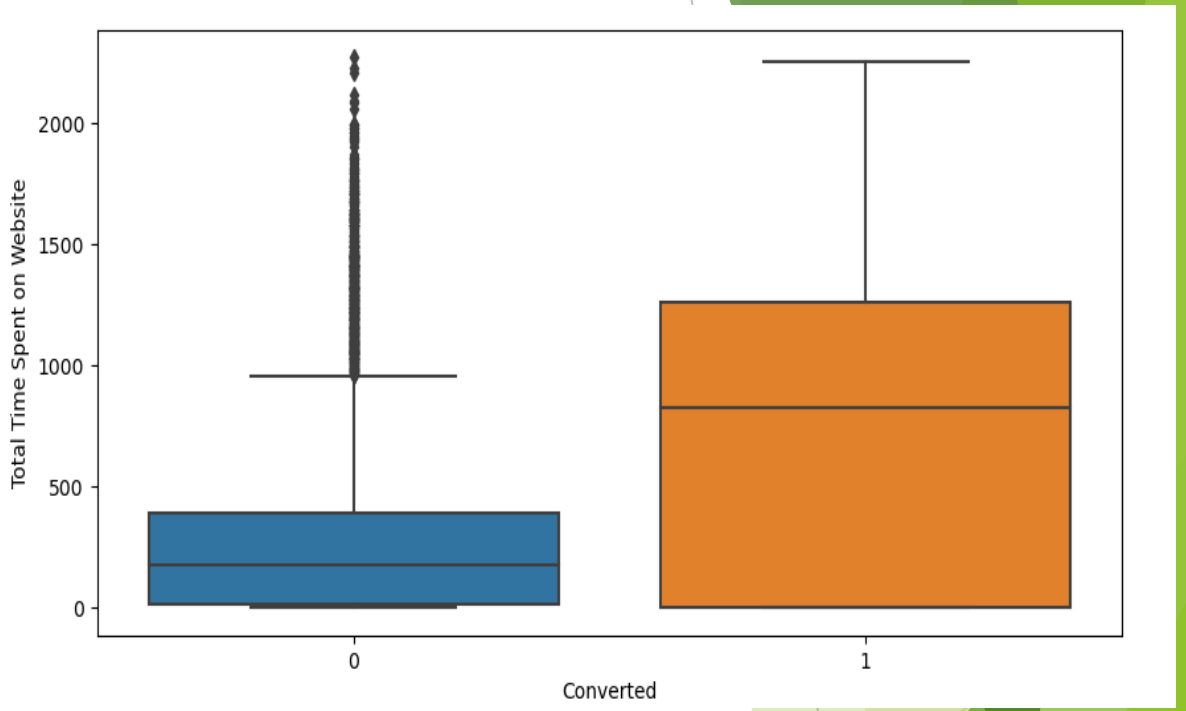
EDA - Bivariate Analysis – Numerical Variables

- More time spent on the website in contributing to higher conversions



EDA – Bivariate Analysis for Numerical Variables





- Past Leads who **spends more time on the Website** have a higher chance of getting successfully converted than those who spends less time as seen in the **box-plot**

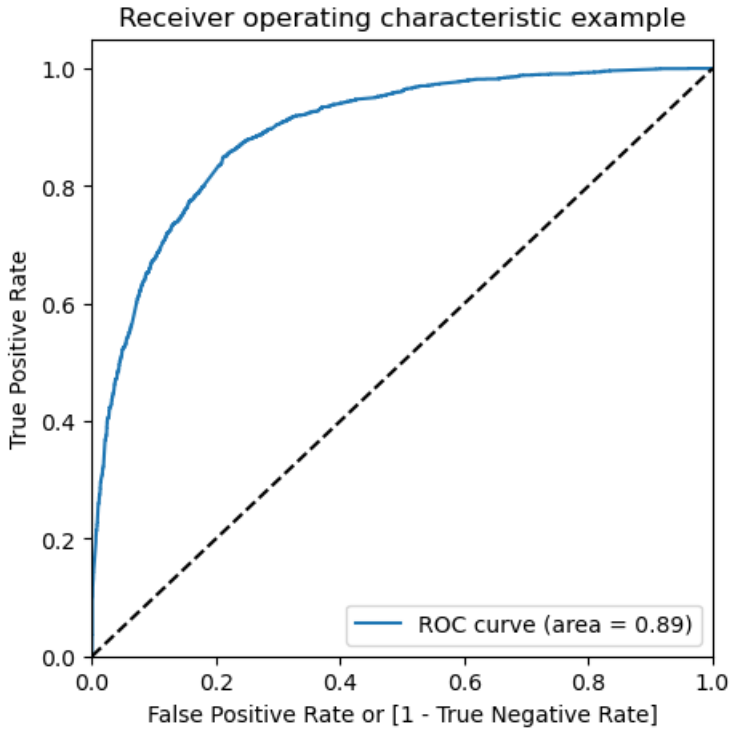
Data Preparation before Model building

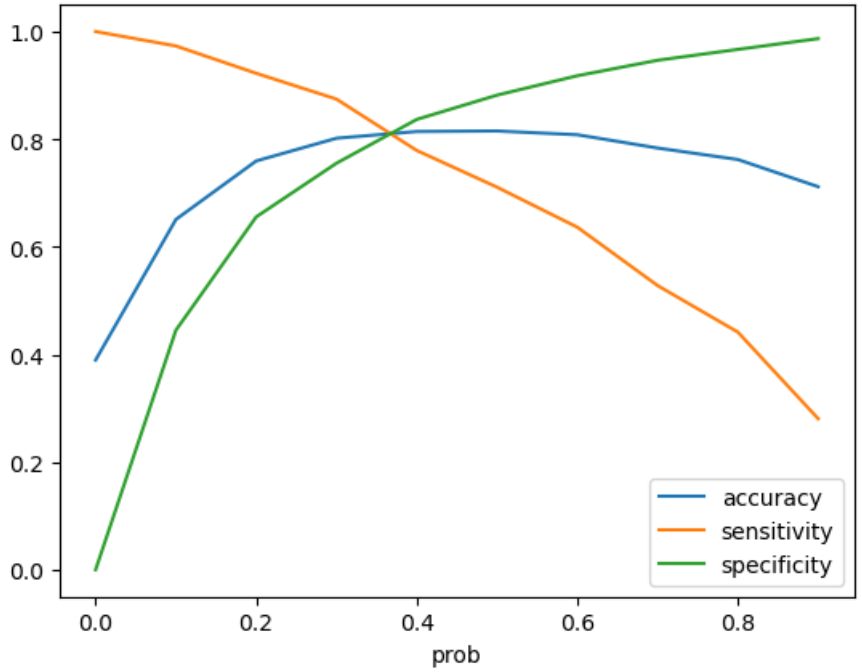
- Binary level ‘Yes/No’ categorical columns have been mapped to '1’ and ‘0'
- Created Dummy Variables for all the categorical columns where the no. of categories > 2
- Splitting Train & Test Sets
  - 70:30 % ratio was chosen for the split
- Feature scaling for continuous variables
  - MinMax scaler was used to scale the features

Model Building

- Feature selection performed using Recursive Feature Elimination (RFE)
- Manual Feature Reduction process was used to build models by dropping variables with insignificant p values
- Pre RFE – 70 columns & Post RFE – 11 columns
- Model 11 looks stable with,
  - significant p-values
  - Acceptable multicollinearity with VIFs less than 5

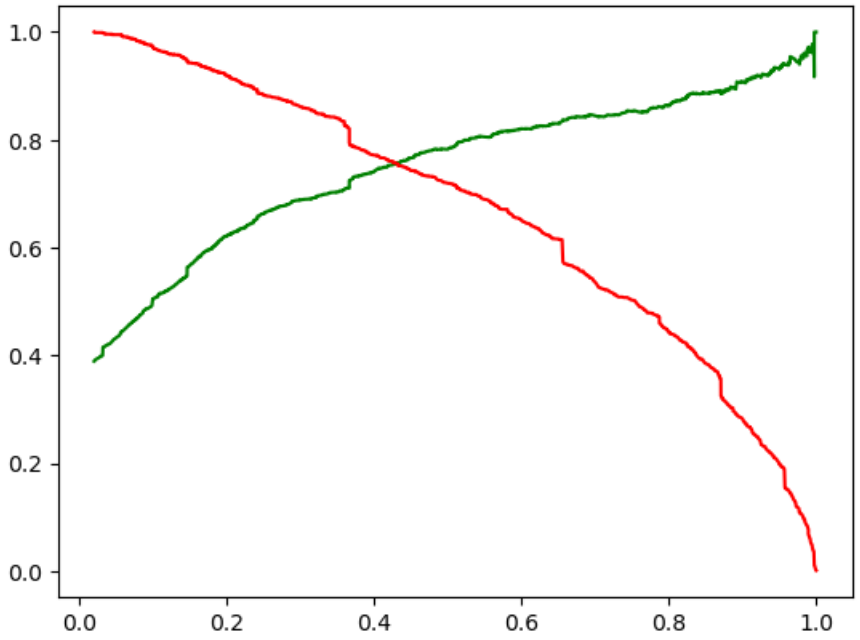
Model Evaluation - Train Data Set





- 0.39 has been taken as the cutoff after checking evaluation metrics coming from both plots
- Final prediction of conversions have a target of ~78%. Hence this is a good model.

Model Evaluation - Test Data Set



- Accuracy, Sensitivity and Specificity values of test set are around 81%, 77% and 83% which are approximately closer to the respective values calculated using trained set.
- Lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around ~78%.
- Therefore, overall this model seems to be good.

Recommendation based on Final Model

Focus Strategies:

1.Targeting the Top 30-40% of High Lead Scores:

Concentrating efforts on leads with high scores maximizes the chances of conversion, ensuring efficient resource allocation. Targeting leads with higher scores optimizes outreach, focusing on those more likely to convert, leading to better conversion rates.

2.Personalized Phone Call Script:

A personalized script enhances the quality of interactions, addressing specific needs and concerns, and fostering a positive impression.

3.Encouraging Customer Engagement on Portal:

Increased portal engagement signifies stronger interest and commitment, improving the likelihood of successful conversions.

Expansion Strategies:

1.Leveraging Existing Lead Feedback:

Insights from current customers can provide valuable information to refine strategies, overcome objections, and tailor communication approaches.

2.Targeting Working Professionals:

Diversifying the target audience taps into a potentially lucrative market, as working professionals may have higher conversion potential due to their specific needs and resources.

Thank You!