Lead Scoring Case Study

Group Members:-

Kiran Ghadigaonkar

Jayant Kumar Karan

Preeti Singh

Problem Statement

- X Education sells online courses to **Industry Professionals**.
- This 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 starts 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%.

Business Objective

- X Education needs help in identifying the most promising leads, i.e. the leads that are most likely to convert into paying customers.
- The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.
- The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

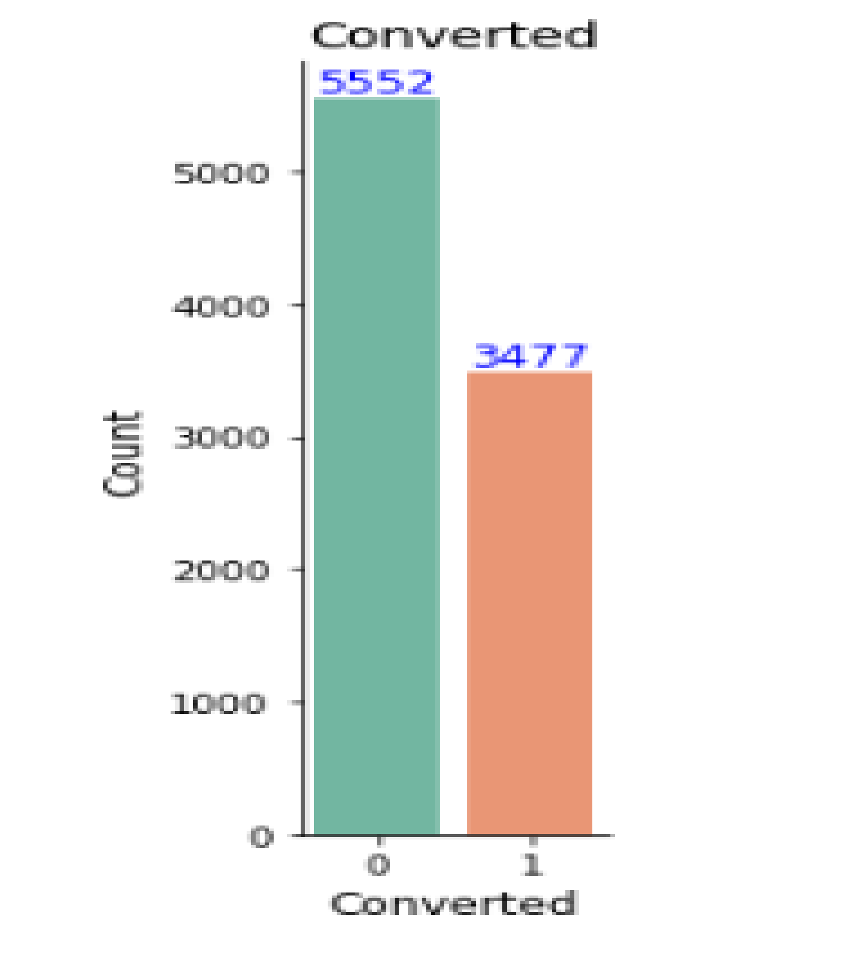
Solution Methodology

- Reading and Understanding the data
- Data cleaning and Data Manipulation
  - ❑ Check and handle Duplicate values and Missing values.
  - ❑ Drop columns, if it has a higher no. of missing values since, it won't help analysis.
  - ❑ Imputation of the missing values, if required.
  - ❑ Outliers Treatment
- EDA (Exploratory Data Analysis)
  - ❑ Univariate Data Analysis: Value count, distribution of variables, etc.
  - ❑ Bivariate Data Analysis: Correlation Coefficients and Patterns between the variables, etc.
- Data Preparation
  - ❑ Dummy Variables Creation, Test-Train Split, Feature Rescaling.
- Model Building
  - ❑ Classification technique: Model Building and Prediction using Logistic Regression.
- Model Evaluation
  - ❑ ROC Curve, Optimal Cut-off Point, Confusion Matrix, Precision, and Recall.
- Making Prediction on Test Set
- Conclusions and Interpretations for Higher Conversion.

Data Manipulation

- Total Number of Rows =37, Total Number of Columns =9240.
- Unique 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.
- After checking for the value counts for some of the object type variables, we find that, some of the features which have no enough variance, have been dropped like 'Newspaper Article', 'X Education Forums', 'Newspaper', etc.
- Dropping the columns having more than 30% as missing value such as 'Specialization', 'How did you hear about X Education', 'Tags', 'Lead Quality', 'Lead Profile','City', 'Asymmetrique Activity Index', 'Asymmetrique Profile Index','Asymmetrique Activity Score', 'Asymmetrique Profile Score', etc.

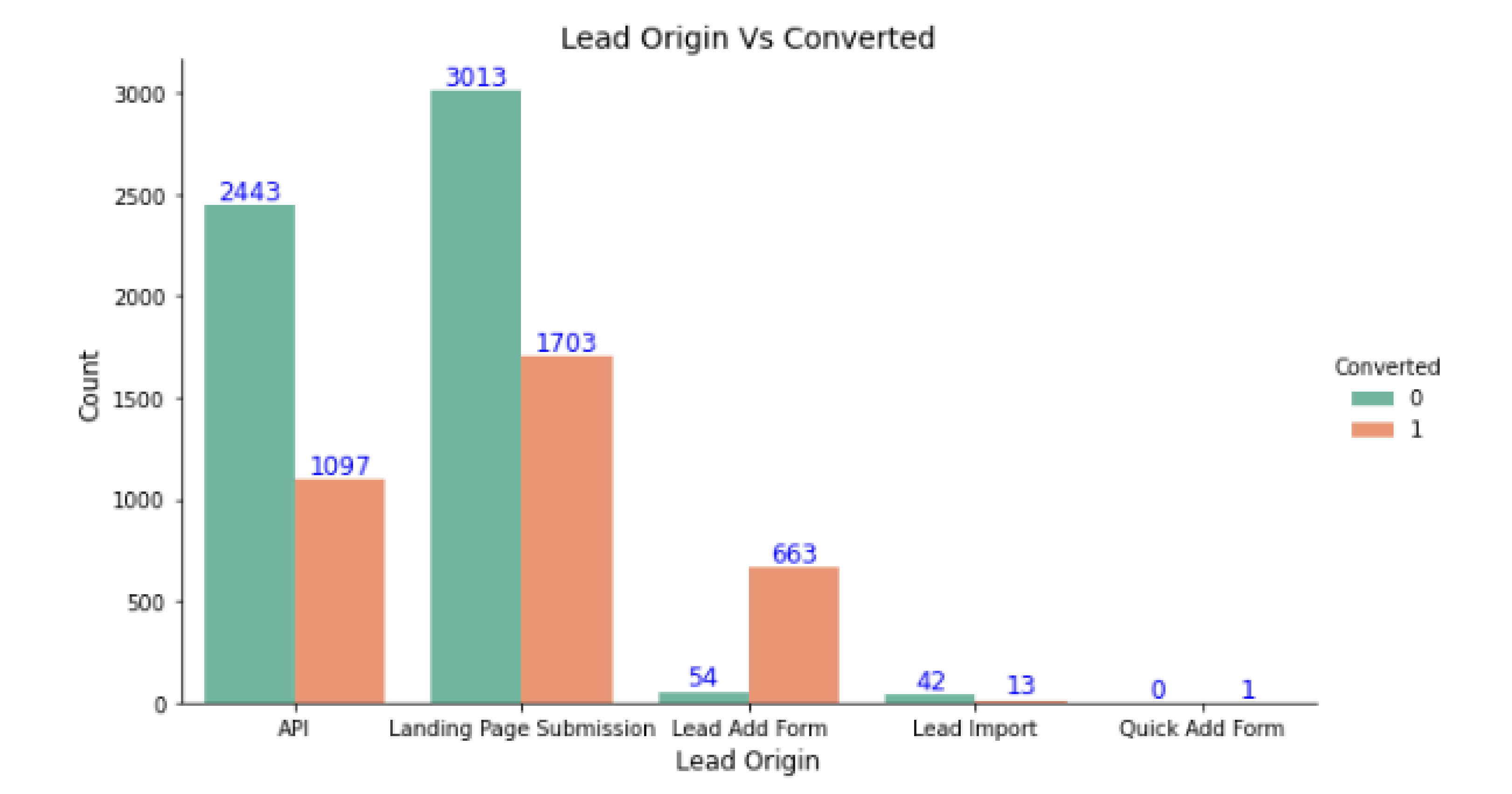
EDA



Overall Conversion

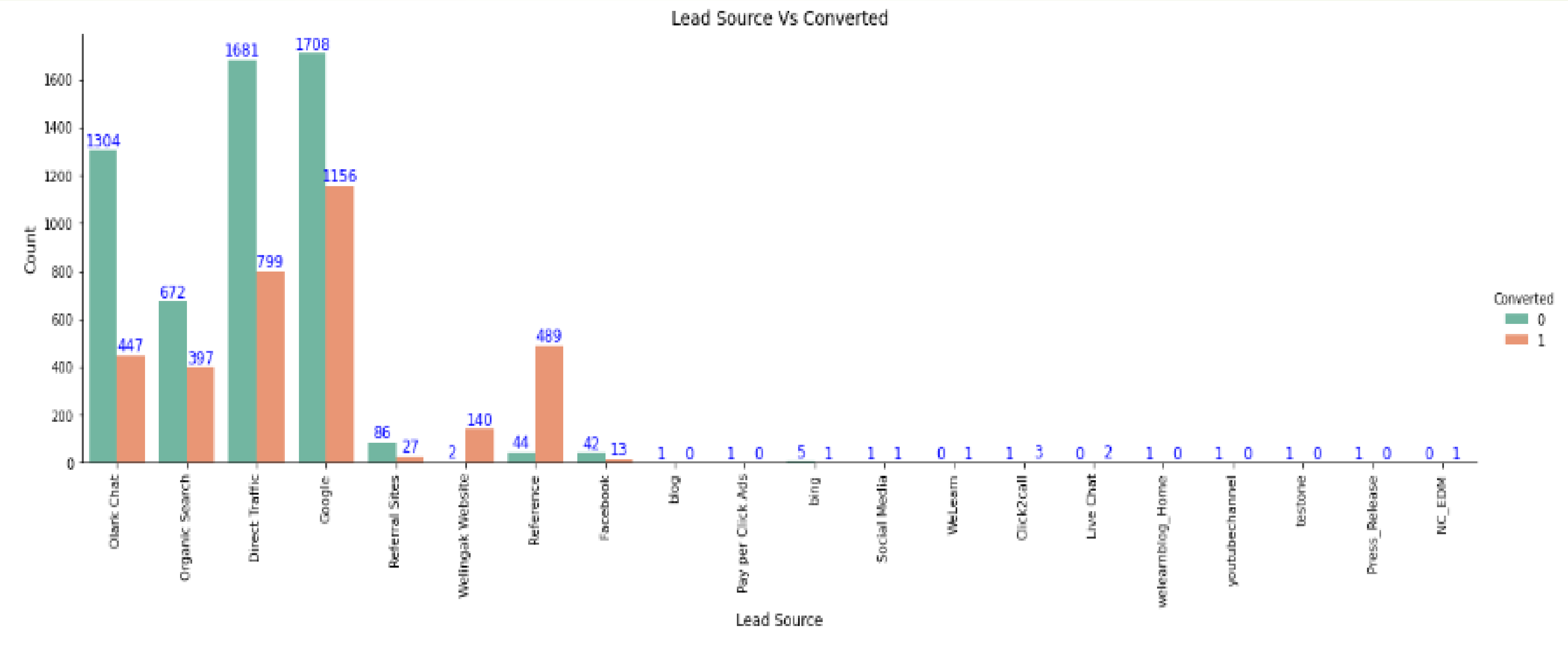
From the above graph, the overall conversion rate is around **39%.**

EDA – Categorical Variable Relation



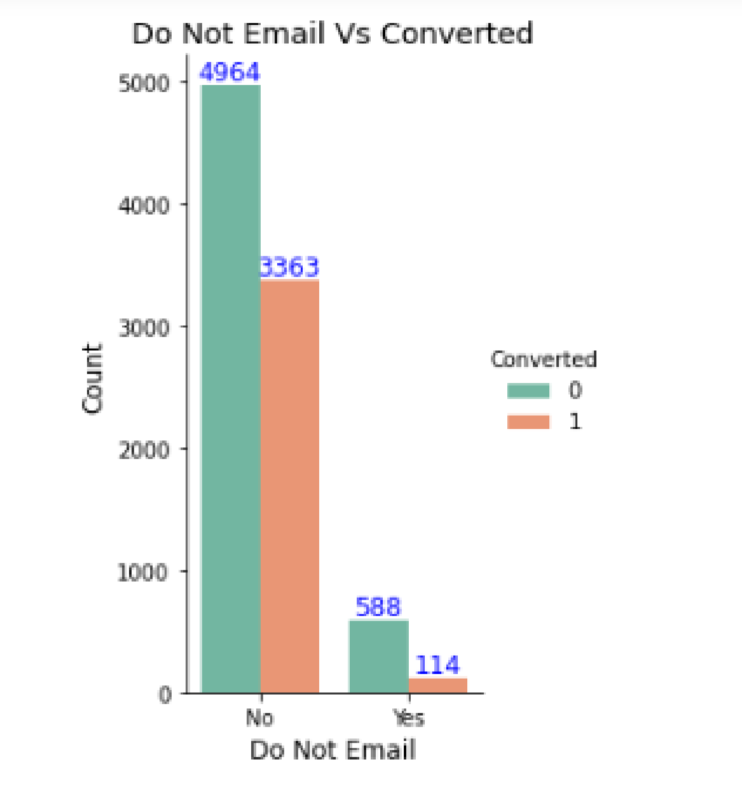
From the above graph, **Landing Page Submission** has maximum conversion Whereas, there is only one lead from **Quick Add Form** which got converted.

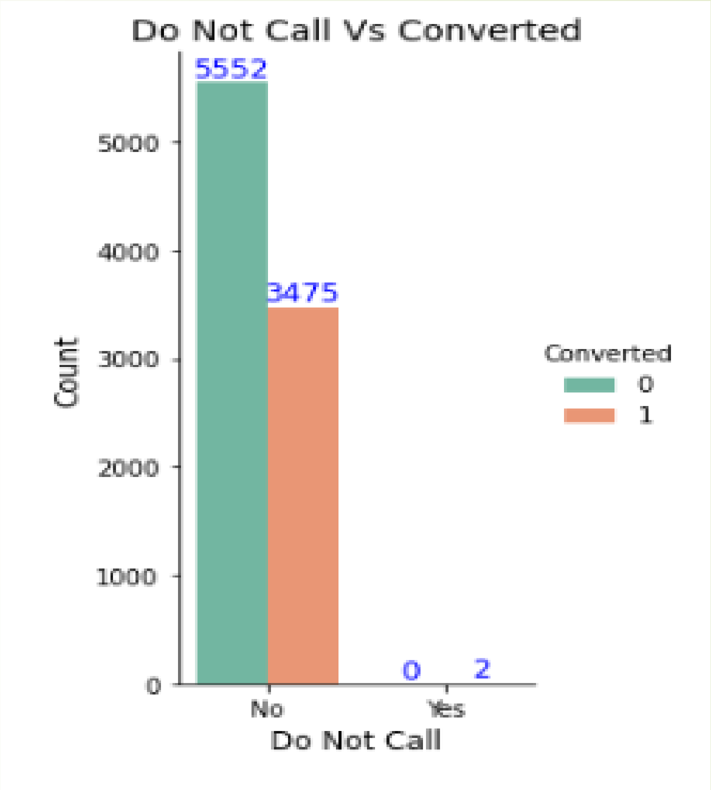
EDA – Categorical Variable Relation



From the above graph, we can see that the major conversion in the lead source is from Google.

EDA – Categorical Variable Relation

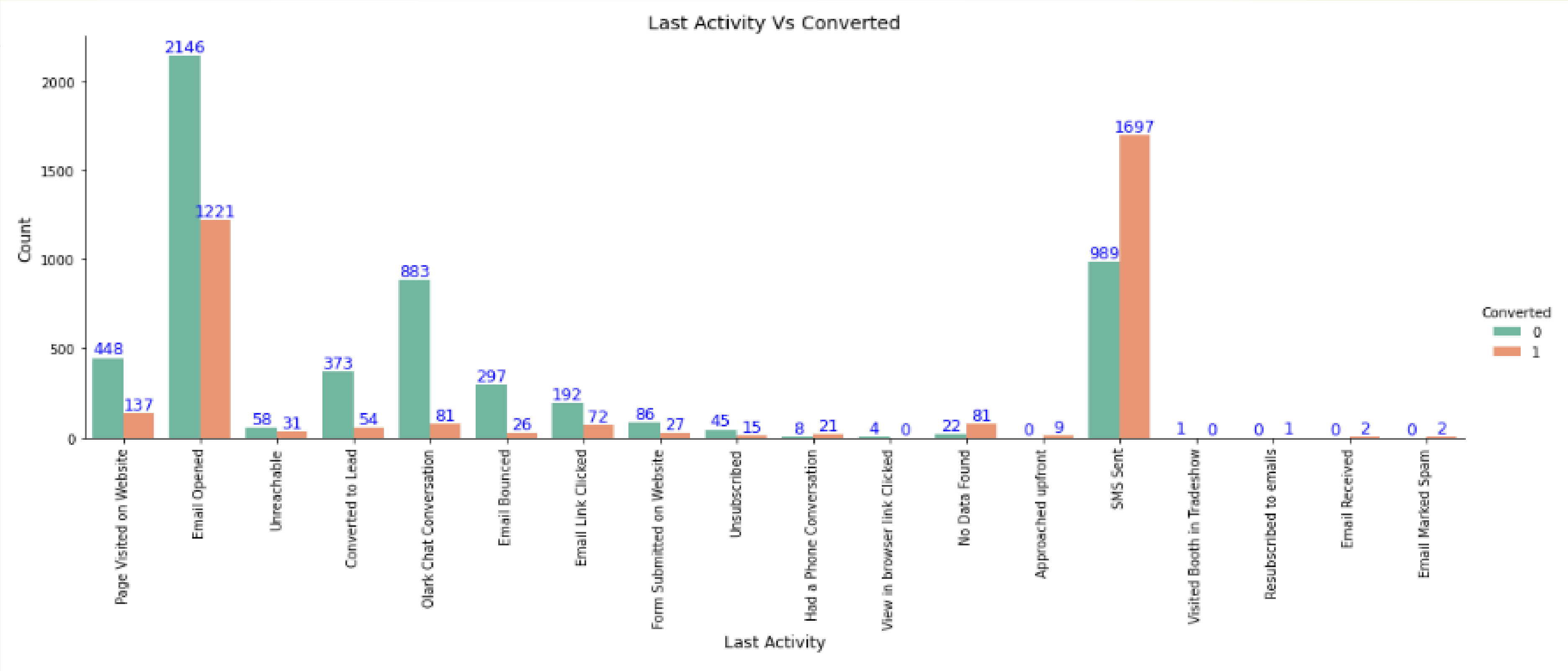




As we can see, the major conversions have been done through emails.

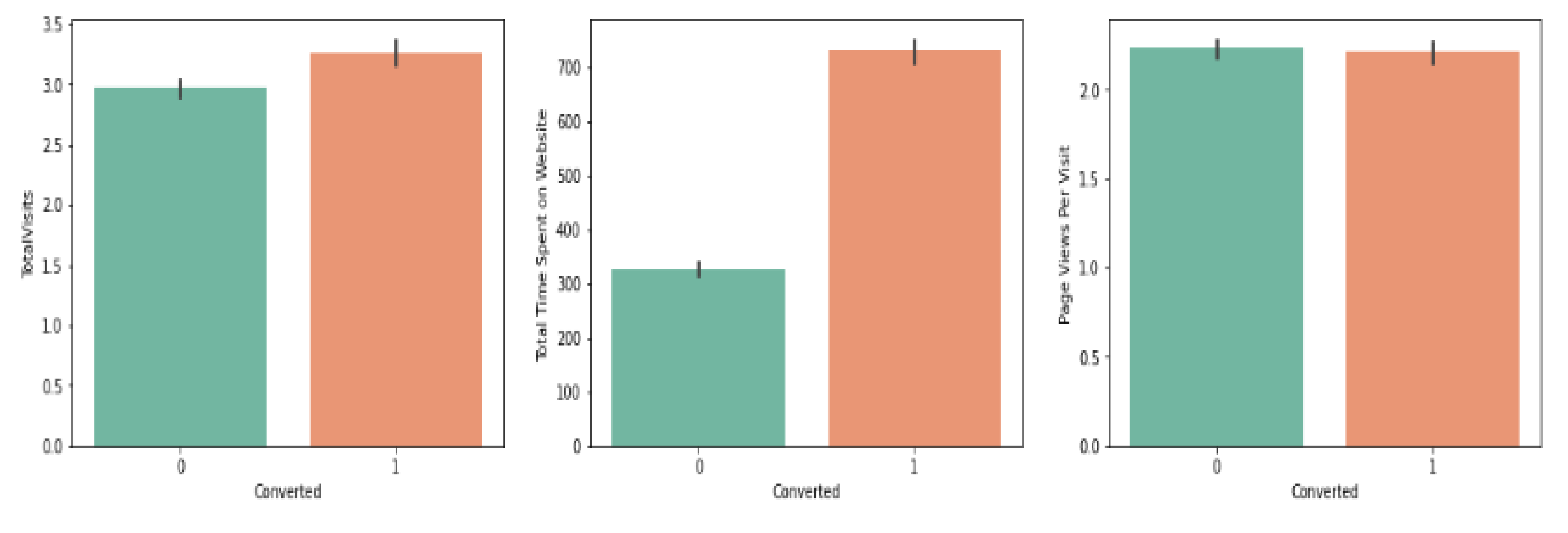
As we can see, the major conversions were done through Calls. Also, 2 leads who opted for ‘Do Not Call’, still got converted.

EDA – Categorical Variable Relation



As per the above graph, the Last Activity value of 'SMS Sent' had more conversion.

EDA – Numeric Variable Relation



The conversion rates seem high for Total Visits, Total Time Spent on Website, and Page Views Per Visit.

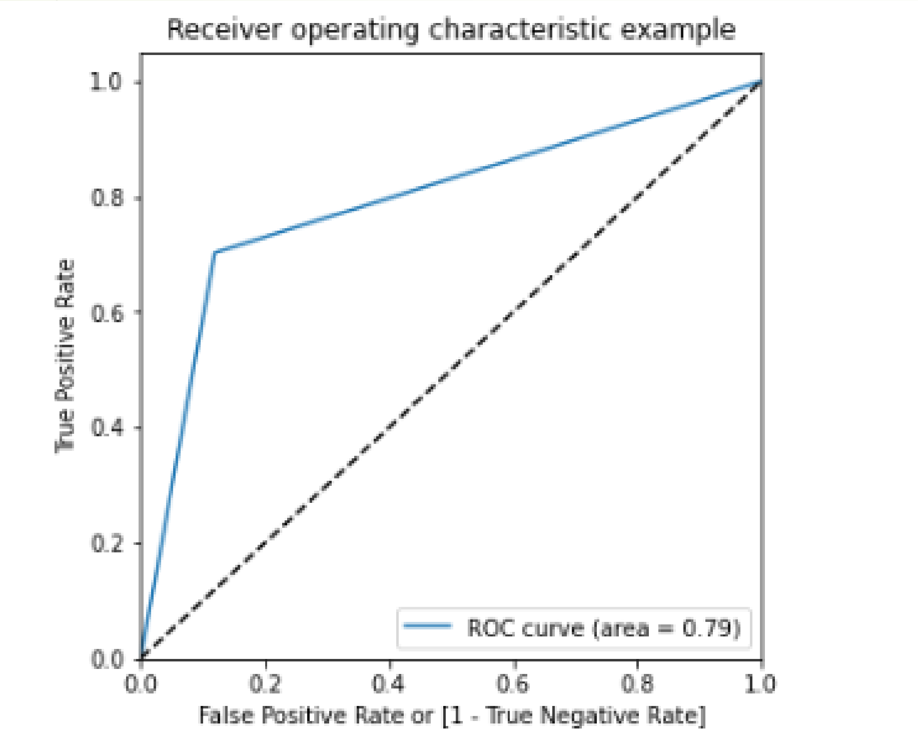
Data Preparation

- Numerical Variables are Normalised.
- Dummy Variables are created for variables with object data type.
- Total Rows for Analysis: 9029
- Total Columns for Analysis: 71

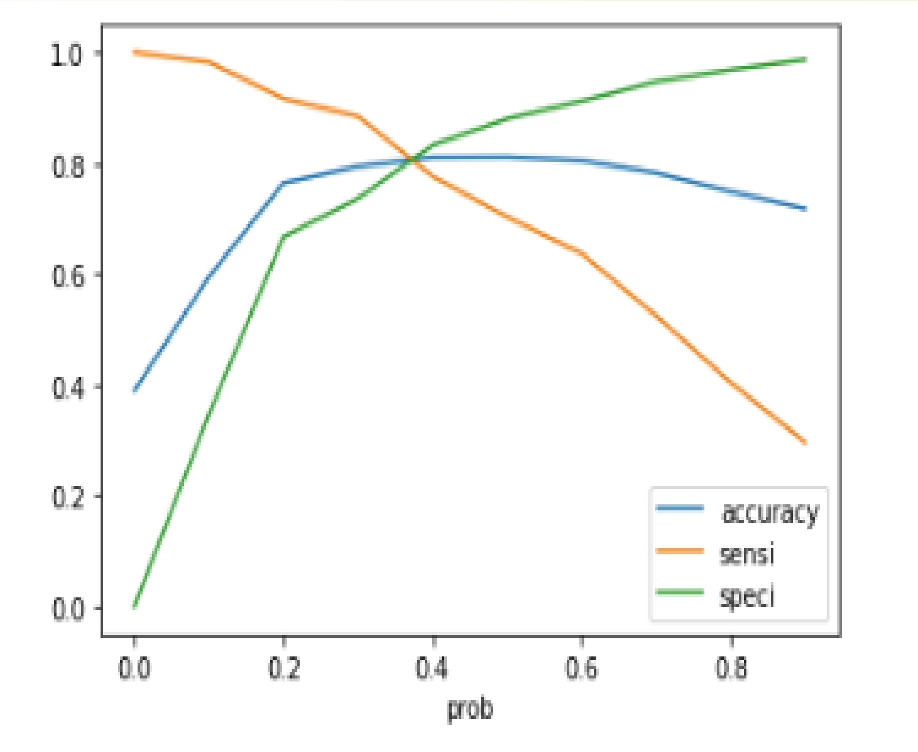
Model Building

- Splitting the Data into Training and Testing Sets.
- Train-Test split with 70:30 ratio.
- Feature Selection using RFE.
- Running RFE with 15 variables as output.
- Model Building by dropping the variable with p- value greater than 0.05 or VIF value greater than 5.
- Making Predictions on test data set.
- Overall accuracy 48%

ROC Curve



As we can see that, The area under ROC curve is 0.79 which seems good as it lies within (0,1).



As we can see from the curve above, 0.35 is the optimum point to take as a cutoff probability.

Conclusions

As we can see that, below are the Potential leads that can be preferred to increase chances of conversion:-

- Leads with maximum Total Visits
- Leads with maximum Total Time Spent on the website
- Leads with Lead Source as Google
- Leads with Last Activity as SMS
- Leads with Current Occupation as Working Professional