Lead Scoring Case Study Using Logistic Regression

Submitted By:

Anuj Pansare

Onkar Ramdin

Nidhi Shah

Problem Statement

- X Education, an online course provider for industry professionals, receives numerous daily visits from interested individuals via various marketing channels like websites and search engines. These visitors can browse courses, fill out forms with contact details, or watch videos. When contact information is provided, the visitors become classified as leads. Leads also come from past referrals.
- The sales team at X Education then engages these leads through calls and emails, aiming to convert them into customers. However, the current lead conversion rate is low, at about 30%.

Objective

- The company aims to improve efficiency by identifying "Hot Leads"—those with the highest potential to convert. By focusing on these high-potential leads, X Education hopes to increase its lead conversion rate and improve the effectiveness of its sales efforts.

Solutioning Approach

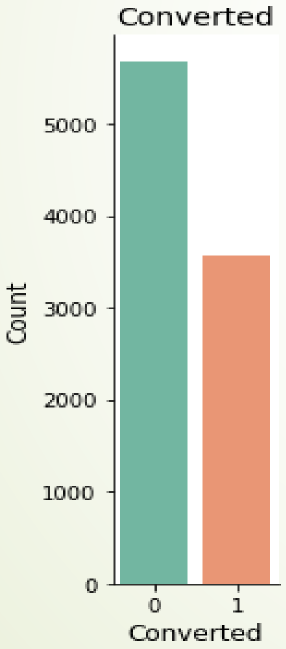
- **Data Collection and Preparation**
  - **Ingest Data**: Ingest the lead data from source file.
  - **Data Cleaning**: Handle missing values, remove duplicates, and correct any inconsistencies in the dataset.
  - **Feature Engineering**: Convert categorical variables into binary or dummy variables. For example, convert "Yes/No" variables to 1/0.
- **Exploratory Data Analysis (EDA)**
  - **Understand Data Distribution**: Visualize the distribution of different features using histograms, bar plots, and box plots.
  - **Correlation Analysis**: Use correlation matrices to understand the relationships between different variables.
  - **Identify Key Features**: Analyze which features are most relevant to lead conversion by examining their distributions and correlations with the target variable.
- **Model Building**
  - **Split Data**: Divide the data into training and testing sets, typically using an 70-30 split.
  - **Build Logistic Regression Model**: Use the training set to build the logistic regression model.
- **Feature Selection**
  - **Select Features**: Use Recursive Feature Elimination (RFE) method to select the most relevant features, and ensure selected features are not highly correlated with each other to avoid multicollinearity issues.
- **Model Evaluation**
  - **Evaluate Performance**: Use metrics such as accuracy, precision, recall and the ROC curve to evaluate model performance.

Data Collection and Preparation

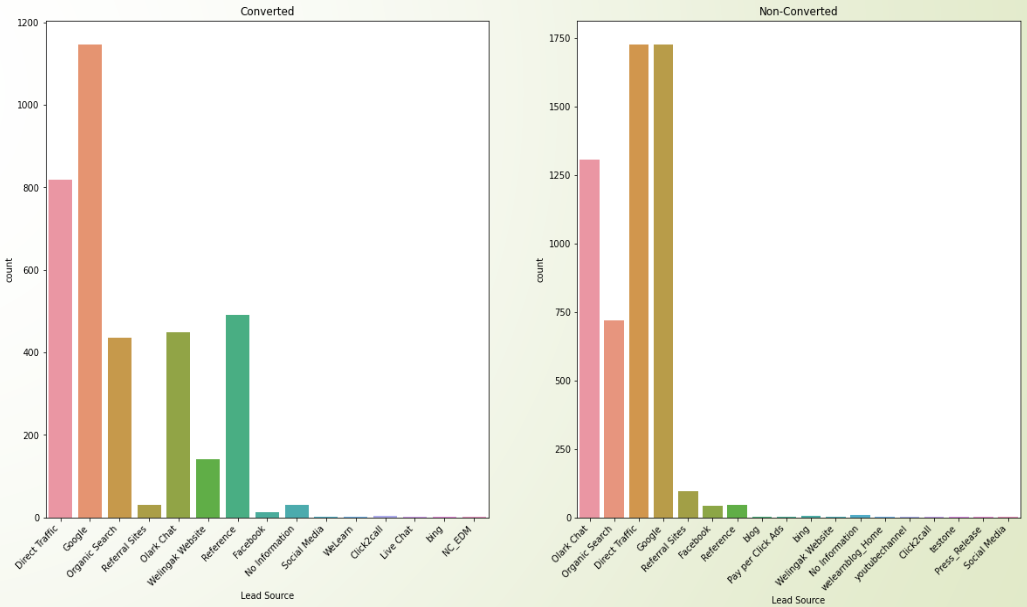
- Import the data from csv file and observe the data.
- Convert the categorical values to binary values using label/binary encoding method.
- Drop the categorical columns for which sum is either 0 or 1, as they are not going to influence the lead scoring.
- Handling null values,
  - Drop the columns having 30% or more null values in it.
  - For non-numeric columns having less percentage of null values, replace the null values with default value.
  - For numeric columns having less percentage of null values, replace the null values with median value.

Exploratory Data Analysis

Based on the current data, **lead conversion rate is approx. 38%**

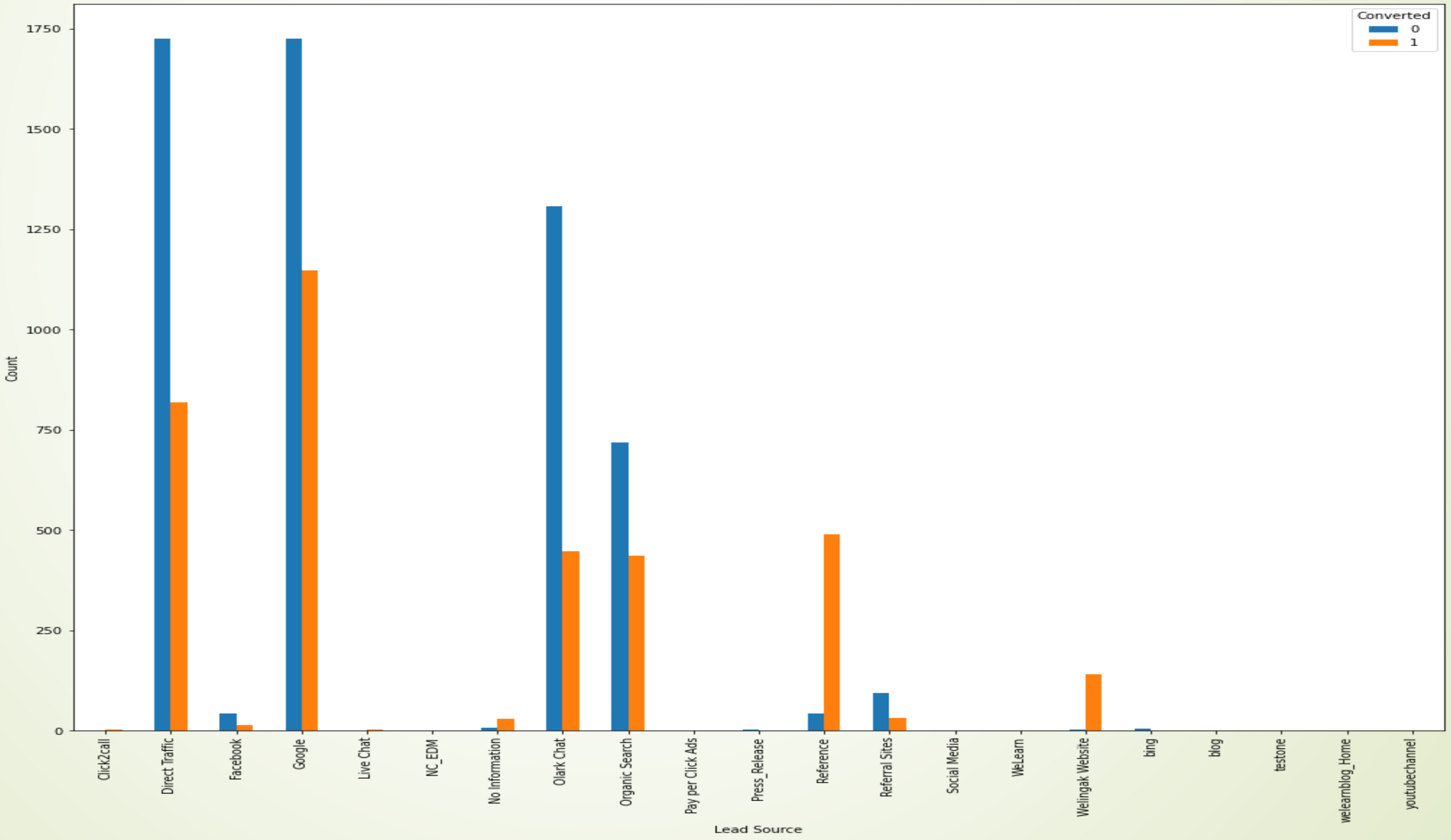


Majority of leads are coming from **Direct Traffic** and **Google,** followed by **Olark Chat**



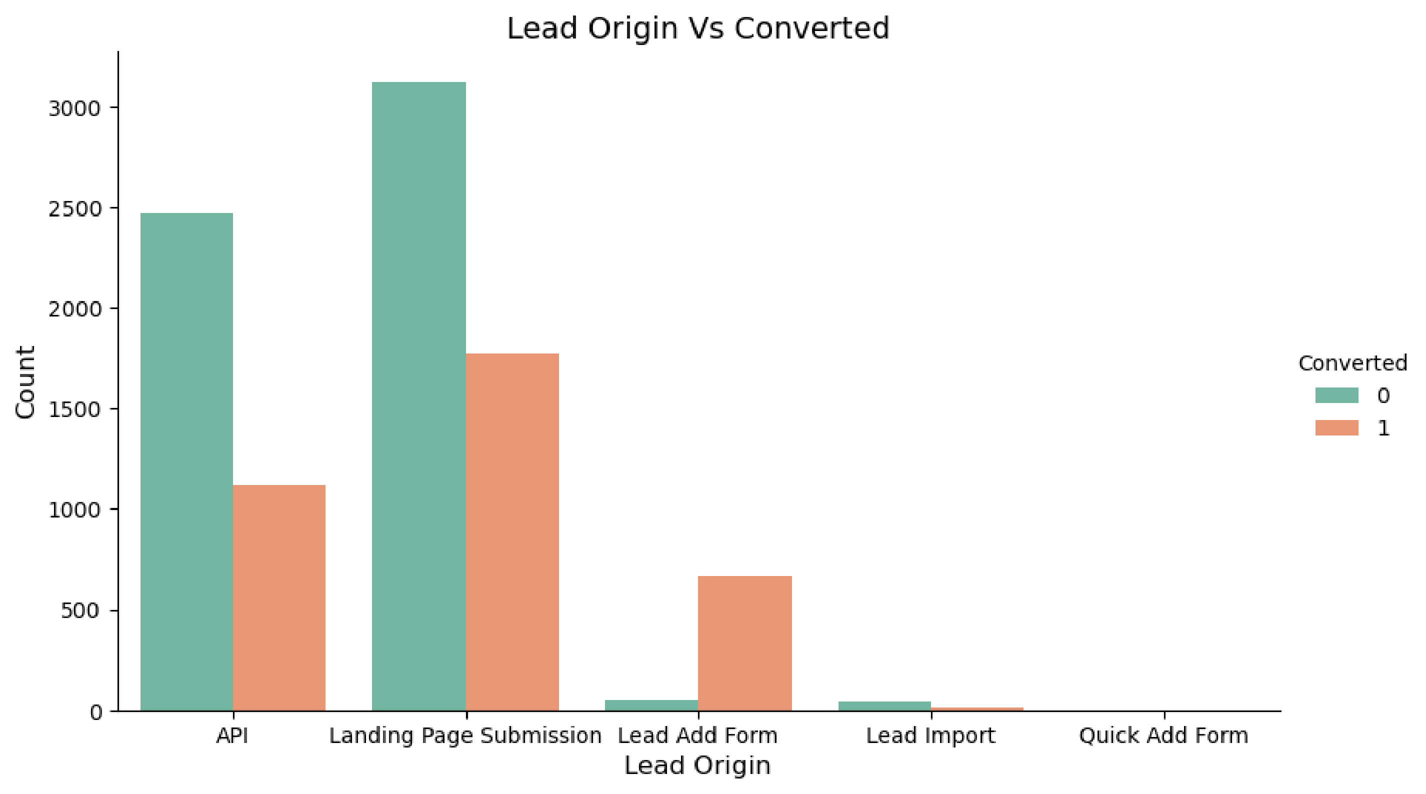
Exploratory Data Analysis

- Leads coming from Google are having highest conversion rate, followed by Direct Traffic and Olark Chat



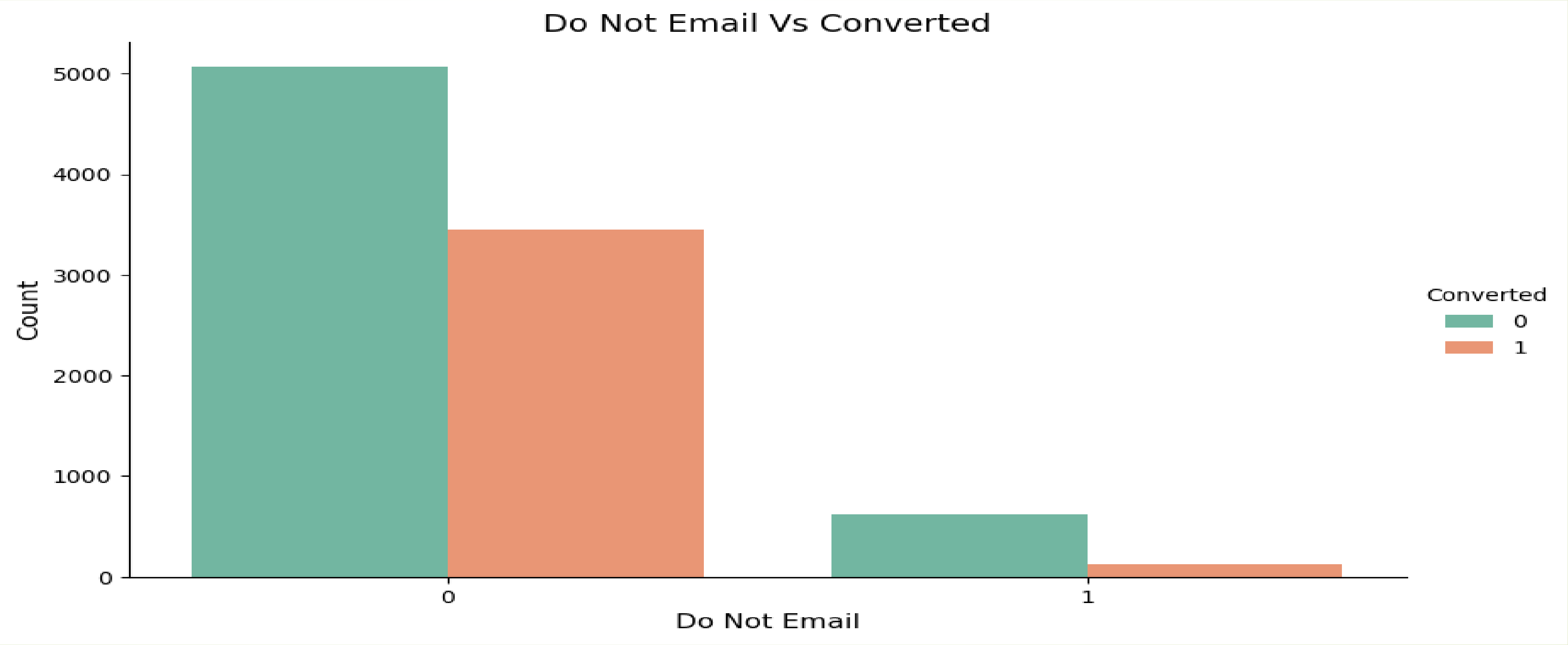
Exploratory Data Analysis

- Leads **originated** from **Landing Page Submission** are having **highest conversion rate**.



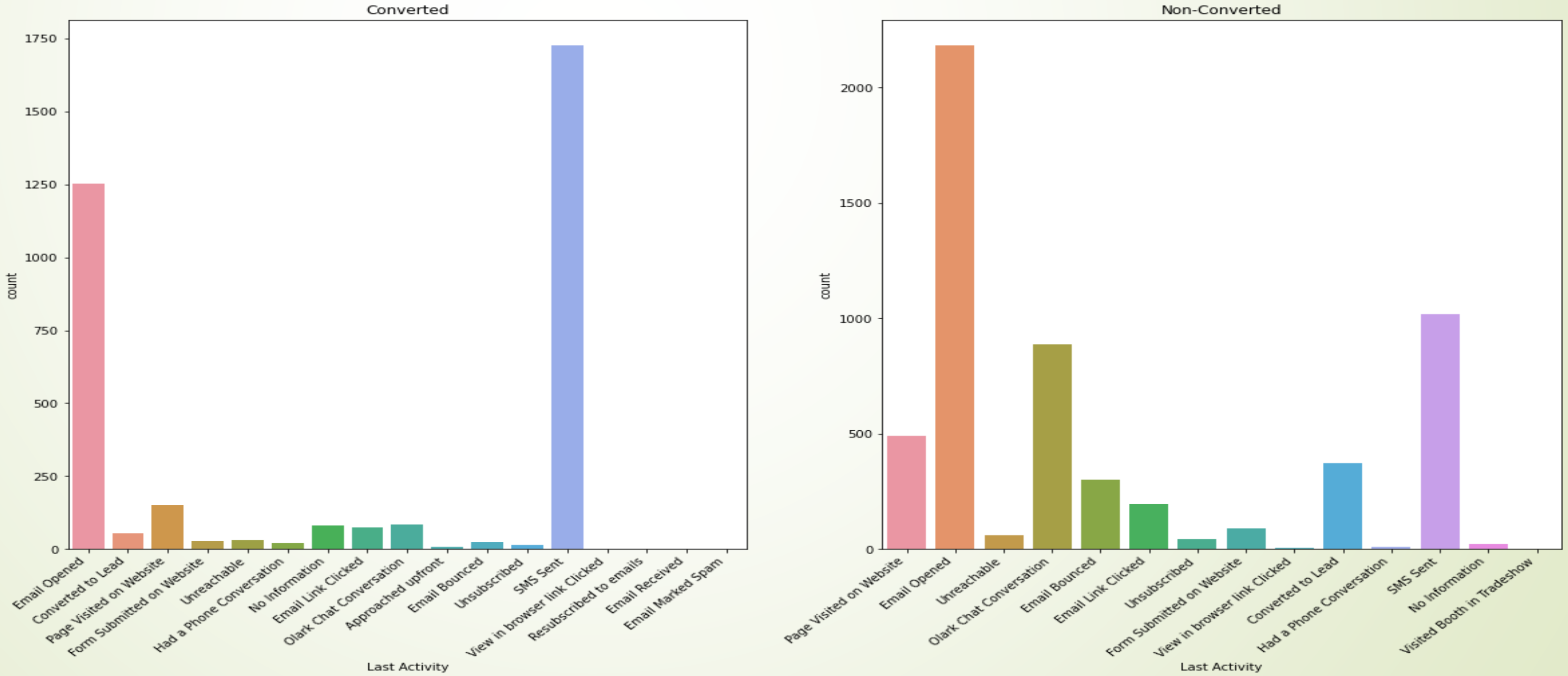
Exploratory Data Analysis

- The **majority of conversions occur among those leads who are allowed to be emailed** (where "Do Not Email" is 0). This suggests that **emailing potential leads significantly contributes to their conversion**.



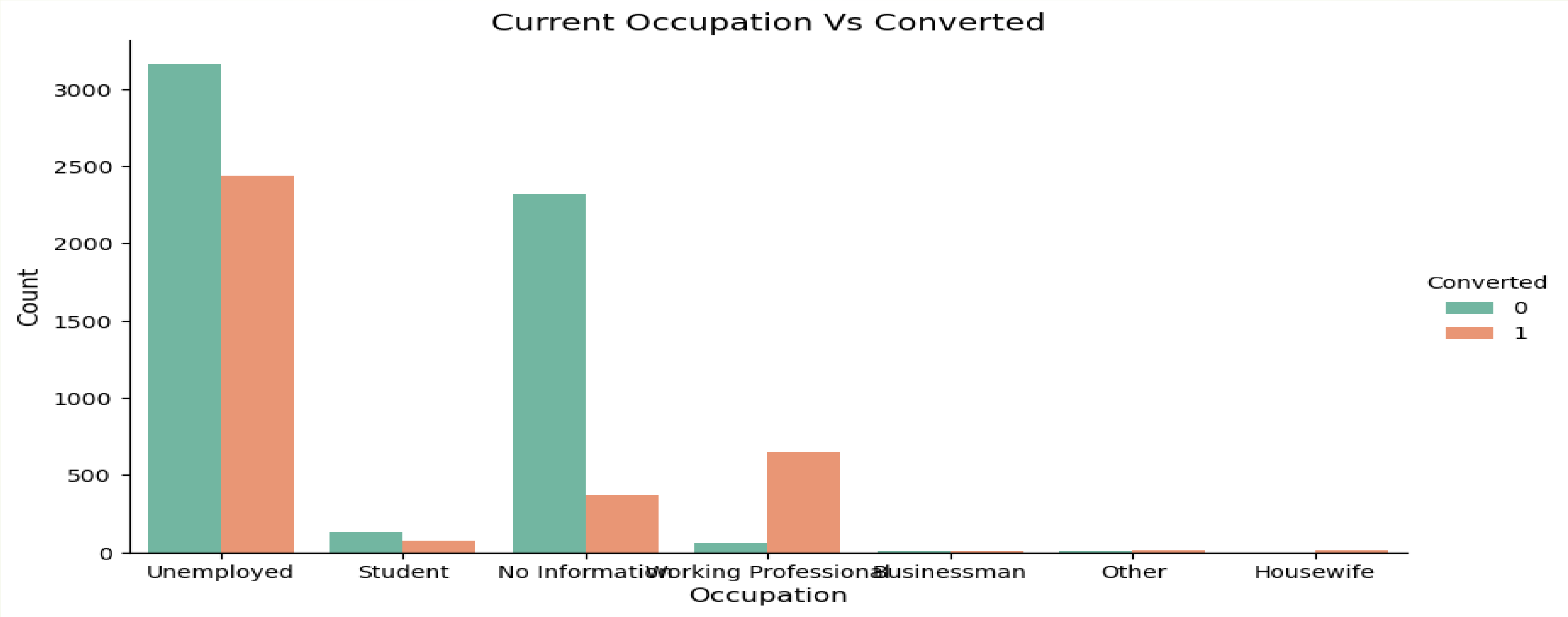
Exploratory Data Analysis

- The most common last activity for **converted leads is "SMS Sent" followed by “Email Opened”**. These activities are associated with a high number of conversions, indicating their effectiveness in the conversion process.



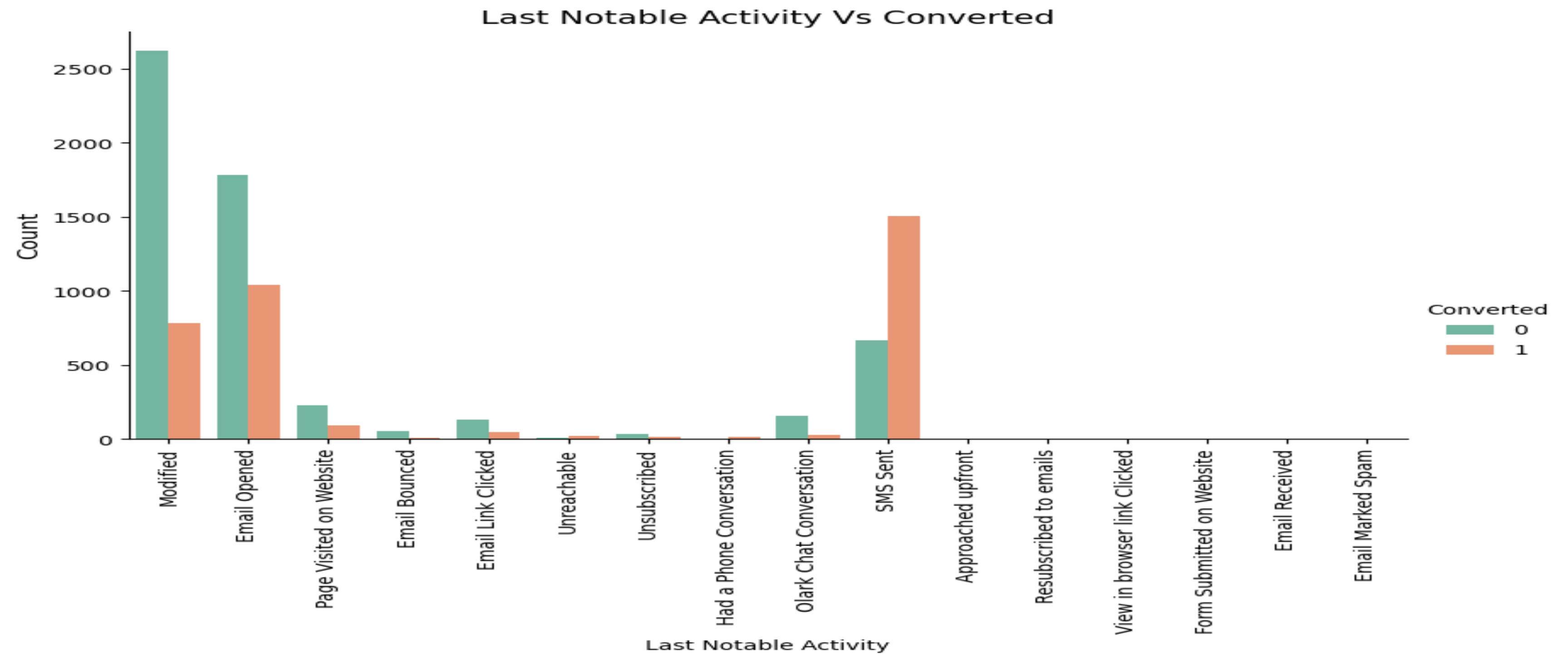
Exploratory Data Analysis

- **Working professionals have a substantial number of conversions** compared to non-conversions. **This group is an important segment** for the company's marketing efforts as they are more likely to convert.
- **A significant portion of the leads are unemployed**, with a notable number converting. This suggests that **unemployed individuals show a relatively high interest** in the courses offered.



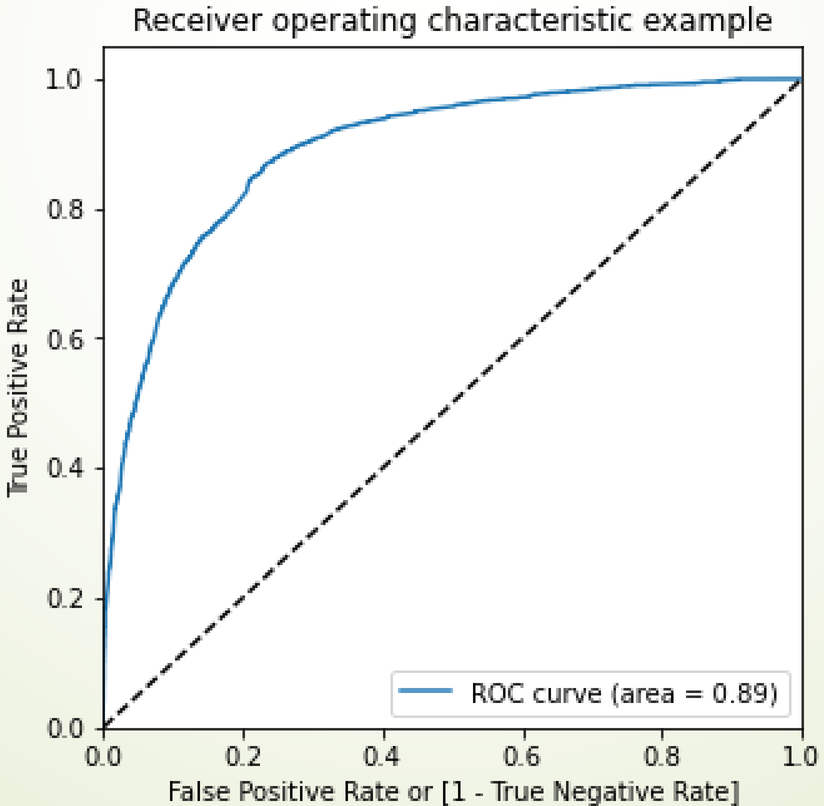
Exploratory Data Analysis

- Leads having last notable activity as “SMS Sent” **have a substantial number of conversions** compared to non-conversions, which suggest potentially effective channels for driving conversions.



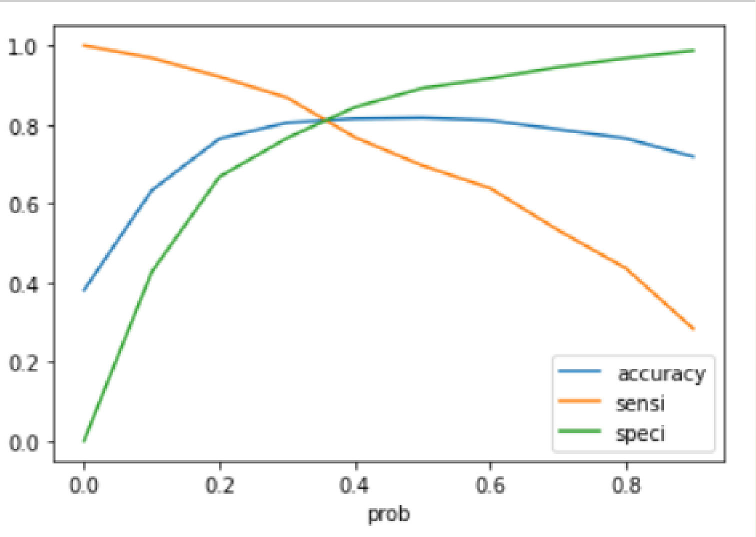
Model Evaluation – Train Data Set

- The ROC curve leans towards the top-left corner, which is ideal. This indicates that the model can correctly classify positive cases (high TPR) while keeping the number of incorrectly classified negative cases (FPR) low. The AUC of 0.89 is significantly higher than 0.5, suggesting the model performs much better than random guessing.



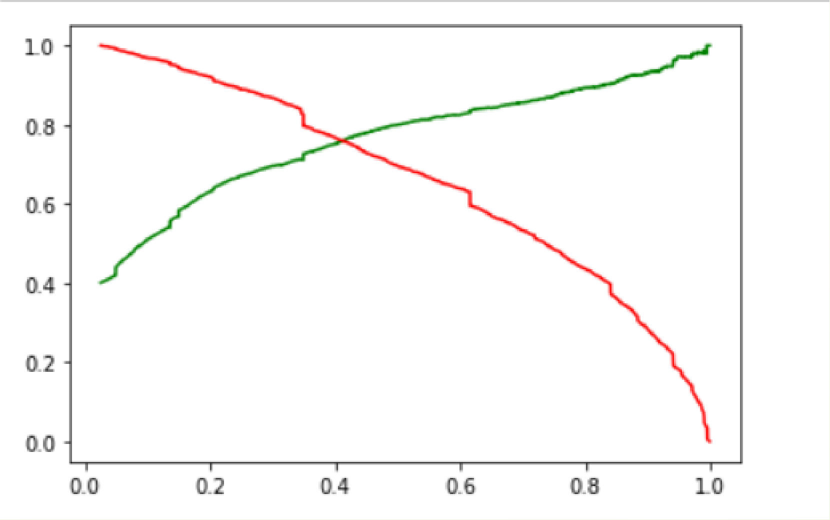
Model Evaluation – Train Data Set

- At very low probability thresholds (close to 0), the model tends to classify most instances as positive, resulting in high sensitivity but low specificity.
- As the threshold increases, the model becomes more conservative in classifying instances as positive, leading to higher specificity but lower sensitivity.
- The point where the accuracy peaks (around 0.35 to 0.40) suggests an optimal balance between sensitivity and specificity for this model, achieving the highest overall proportion of correct classifications.



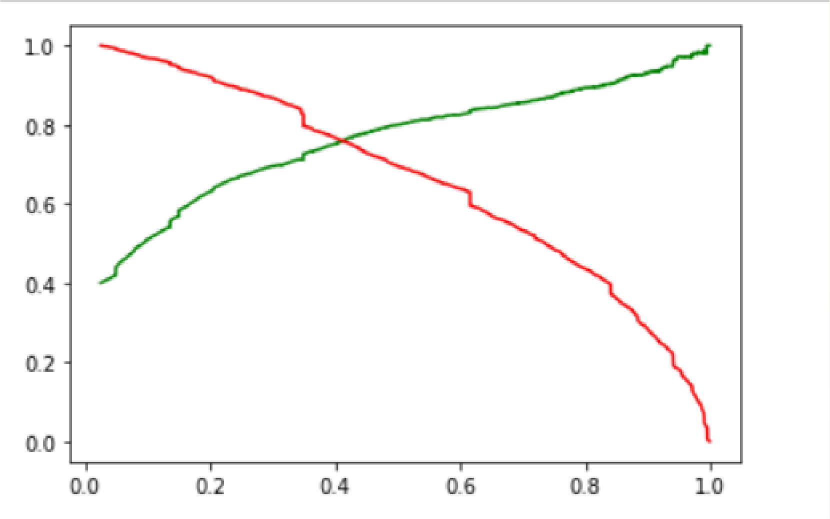
Model Evaluation – Train Data Set

- At very low probability thresholds (close to 0), the model classifies almost everything as positive, resulting in high recall but low precision.
- As the threshold increases, the model becomes more selective in classifying instances as positive. This increases precision but decreases recall.
- The crossing point i.e. 0.41 of the curves typically indicates a balance between precision and recall



Model Evaluation – Test Data Set

- At very low probability thresholds (close to 0), the model classifies almost everything as positive, resulting in high recall but low precision.
- As the threshold increases, the model becomes more selective in classifying instances as positive. This increases precision but decreases recall.
- The crossing point i.e. 0.41 of the curves typically indicates a balance between precision and recall



Conclusion

- The overall accuracy of the model is 80 % with very similar specificity and sensitivity between Train and Test Dataset
- The lead score calculation is provided in the model with a threshold of 37 for prediction.
- The model provides for a conversion rate on the test dataset at around 80%
- So the overall model satisfies the user requirements stated in the problem statement.