X Education - Lead Scoring Case Study

Team Members: Megha Bose, Rohithsaran V, Rishabh Katoch

Table of Contents

Background of X Education Company

- X Education, an online education company, specializes in selling courses to professionals in various industries. On a daily basis, the company attracts numerous professionals who visit their website and explore the available courses. To reach a wider audience, X Education markets its courses on various websites and search engines, including Google.
- When these professionals land on the website, they have the opportunity to browse through the courses, fill out a form expressing their interest, or watch informative videos. Once they provide their email address or phone number by filling out a form, they are categorized as leads. At this stage, the company has acquired potential customers who have shown interest in their courses.
- The sales team at X Education takes charge of these leads, initiating communication through phone calls, emails, and other means. The primary objective is to convert these leads into paying customers. However, it should be noted that the conversion rate for leads at X Education is typically around 30%, indicating that only a fraction of the acquired leads end up making a purchase.

- This information provides an overview of the lead acquisition and conversion process at X Education, highlighting the company's focus on engaging professionals, nurturing leads, and ultimately converting them into paying customers.

Problem Statement & Objective of the Study

Problem Statement:

- X Education gets a lot of leads, its lead conversion rate is very poor at around 30%
- X Education wants to make lead conversion process more efficient by identifying the most potential leads, also known as Hot Leads
- Their sales team want to know these potential set of leads, which they will be focusing more on communicating rather than making calls to everyone.

Objective of the Study:

- To help X Education select the most promising leads, i.e., the leads that are most likely to convert into paying customers.
- The company requires us to build a model wherein we 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%.

Suggested Ideas for Lead Conversion

Leads Grouping

- Leads are strategically categorized based on their propensity or likelihood to convert, resulting in a focused group of highly promising leads.

Better Communication

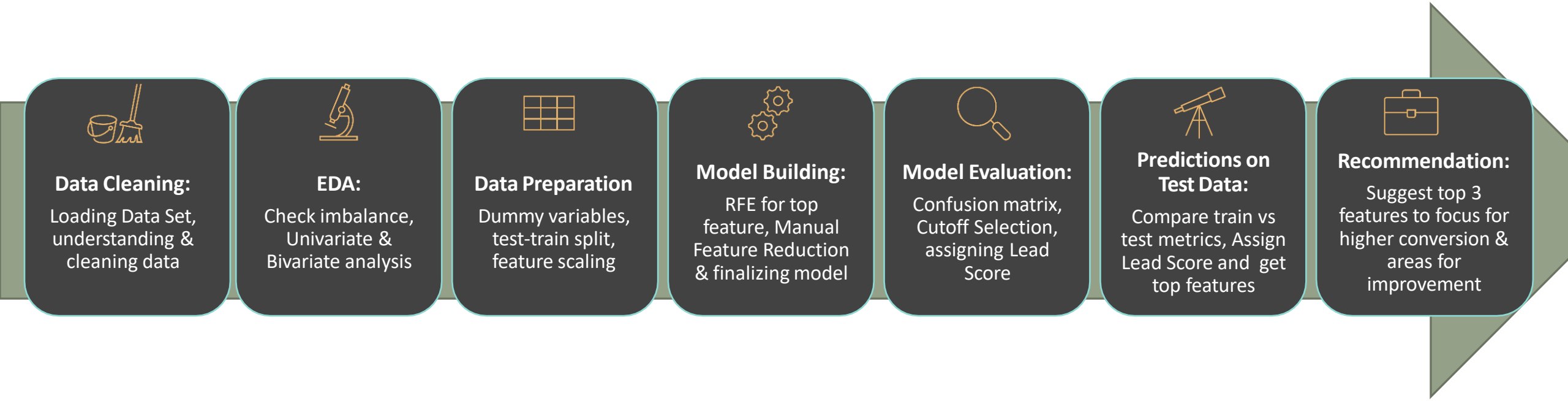
- By narrowing down our pool of leads, we can focus our communication efforts on a smaller, more targeted group. This approach enables us to make a greater impact and increase the effectiveness of our interactions with potential customers.

Boost Conversion

- By focusing on hot leads with a higher likelihood to convert, we can achieve a greater conversion rate and reach the objective of 80%.

To achieve our 80% conversion rate target, obtaining a high sensitivity in capturing hot leads is crucial.

Analysis Approach



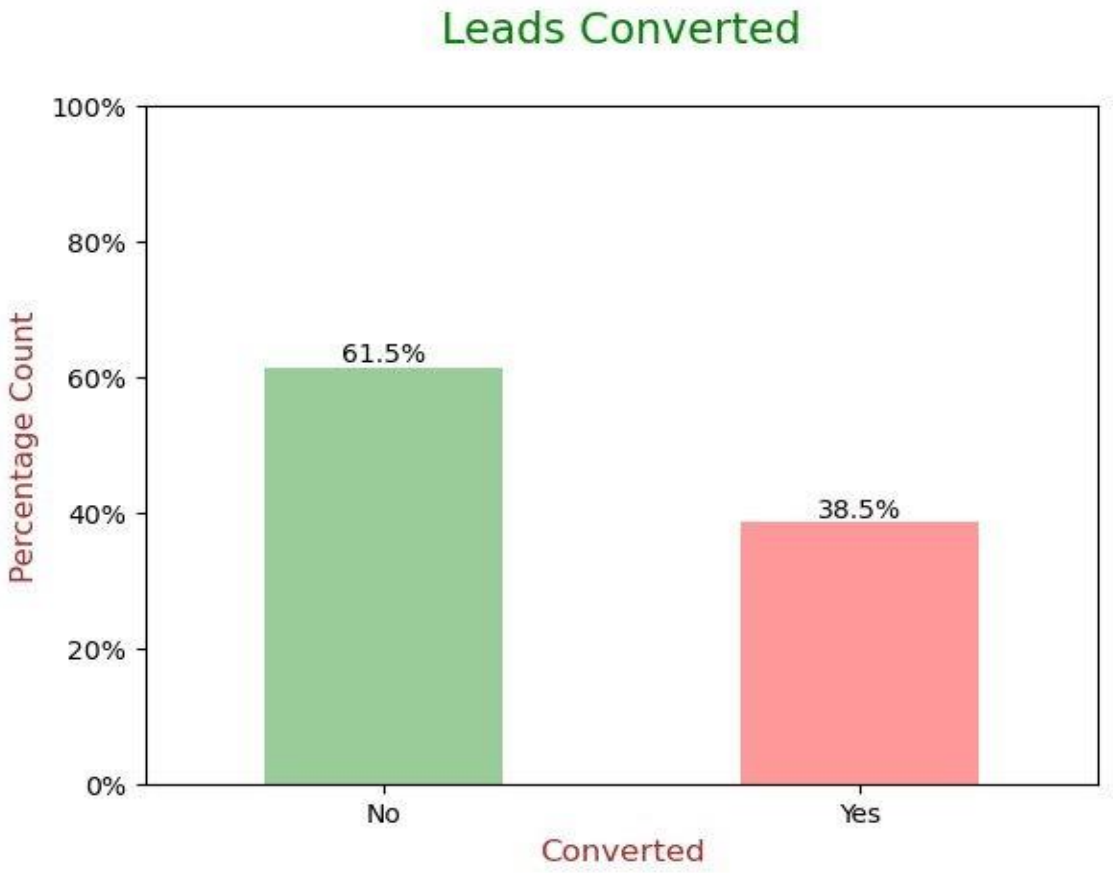
Data Cleaning

- The "Select" level represents null values for some categorical variables, indicating that customers did not choose any option from the provided list.
- Columns with over 30% null values were dropped from the dataset.
- Missing values in categorical columns were handled based on value counts and specific considerations.
- Columns that did not provide valuable insights or contribute to the study objective (such as tags and country) were dropped.
- Imputation techniques were applied to handle missing values in certain categorical variables.
- Additional categories were created for variables as needed.
- Columns that were not useful for modeling purposes (such as Prospect ID and Lead Number) or contained only one category of response were dropped.
- Numerical data was imputed with the mode after assessing the distribution.

Data Cleaning

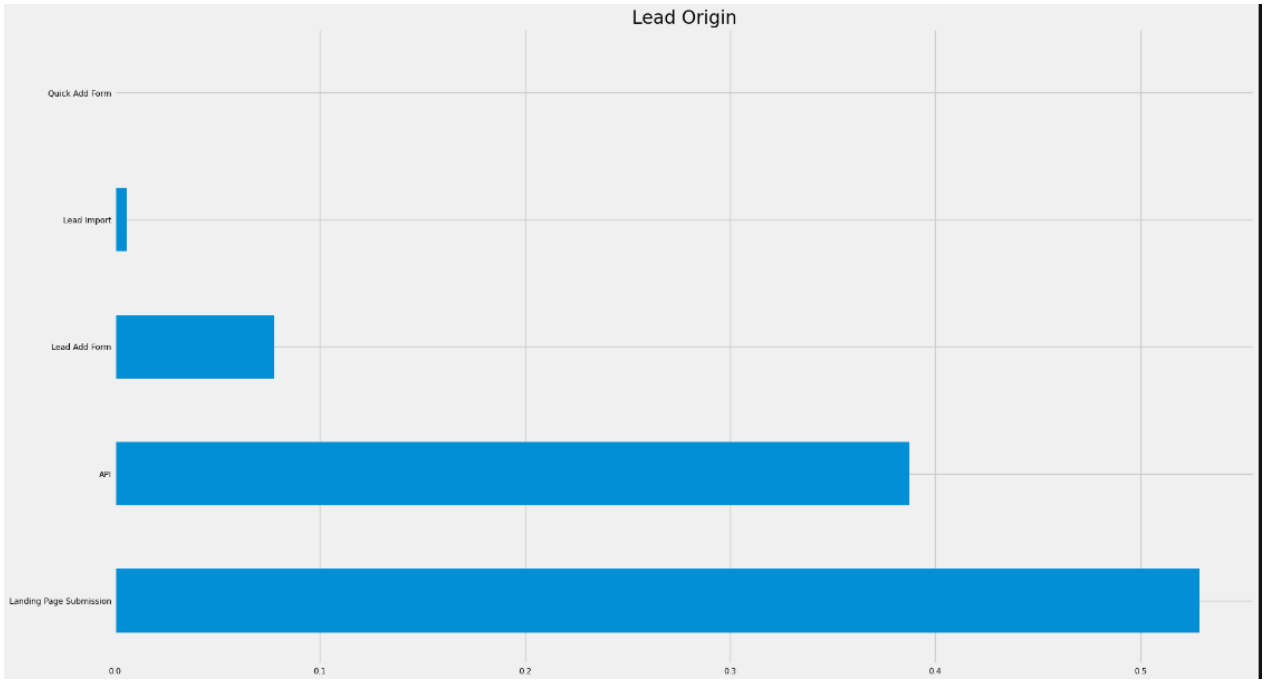
- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers in **TotalVisits** and **Page Views Per Visit** were treated and capped.
- Invalid values were fixed and data was standardized in some columns, such as lead source.
- Low frequency values were grouped together to “Others”.
- Binary categorical variables were mapped.
- Other cleaning activities were performed to ensure data quality and accuracy.
    - Fixed Invalid values & Standardizing Data in columns by checking casing styles, etc. (lead source has Google, google)
    - Dropping the columns with skewed columns as the combination of Not selected and one category > 60%

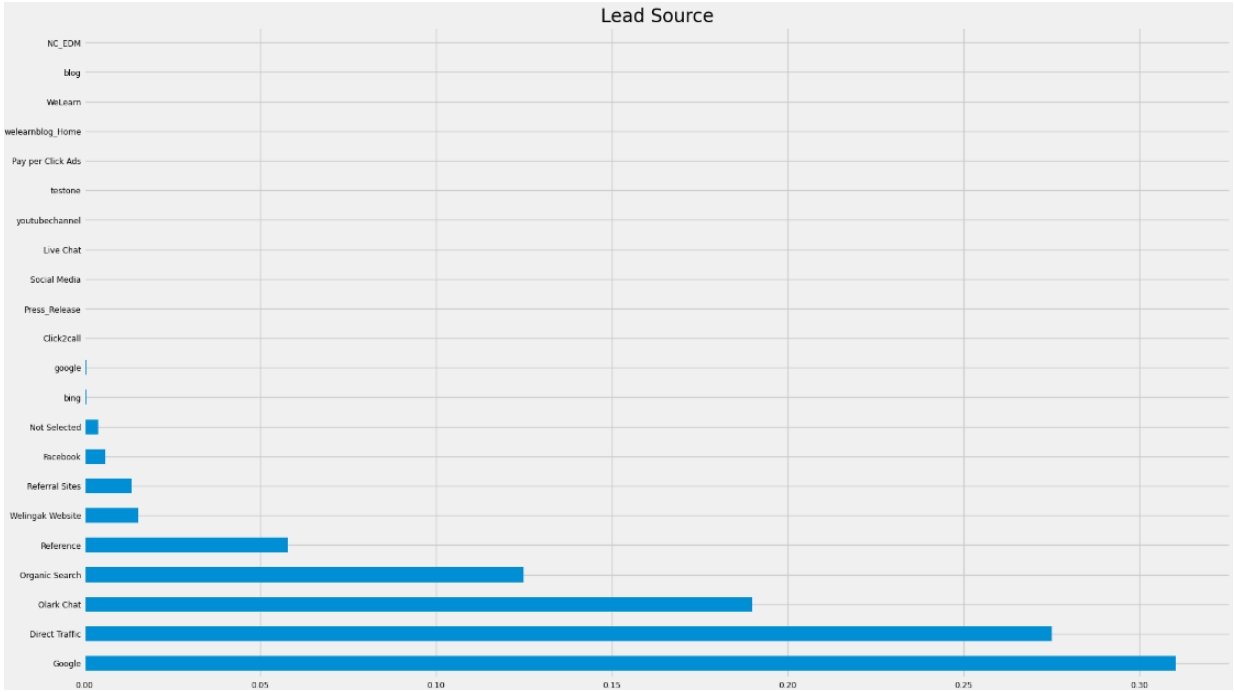
EDA



- Data is imbalanced while analysing target variable.
- Conversion rate is of 38.5%, meaning only 38.5% of the people have converted to leads.(Minority)
- While 61.5% of the people didn't convert to leads. (Majority)

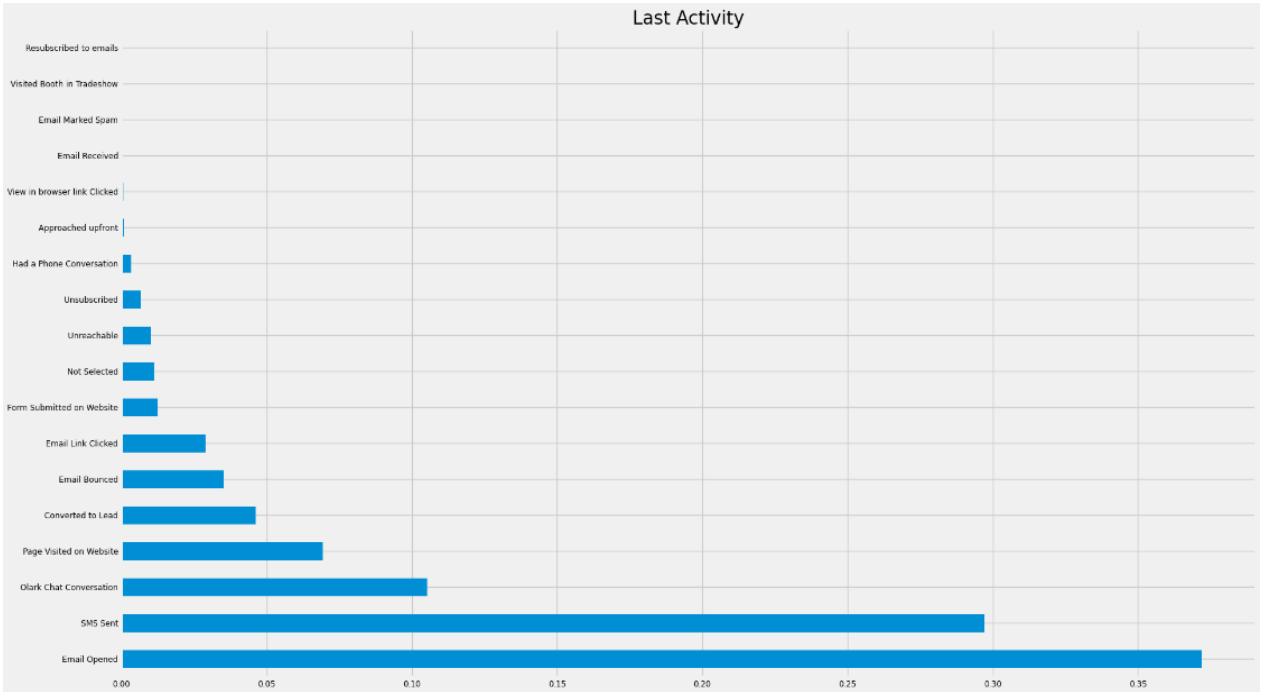
EDA Univariate Analysis – Categorical Variables

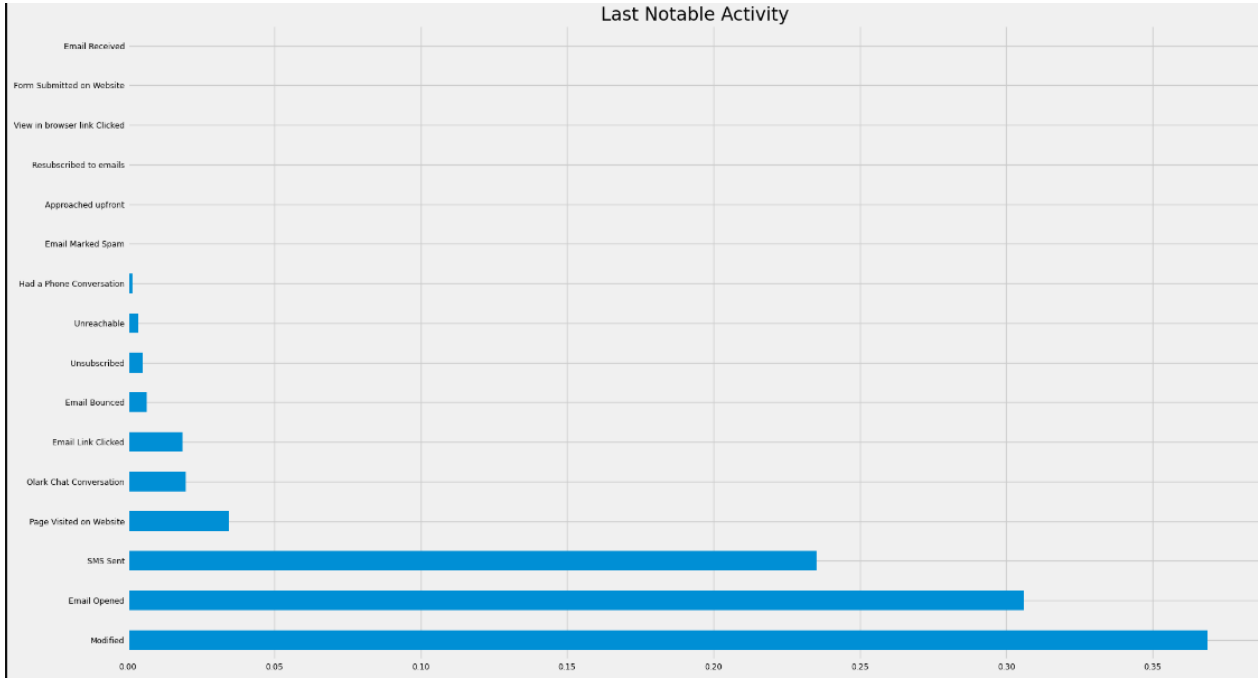




- **Lead Origin:** 55% of lead origin from landing page submission
- **Lead Source :** 33% of lead source is coming from google

EDA Univariate Analysis – Categorical Variables



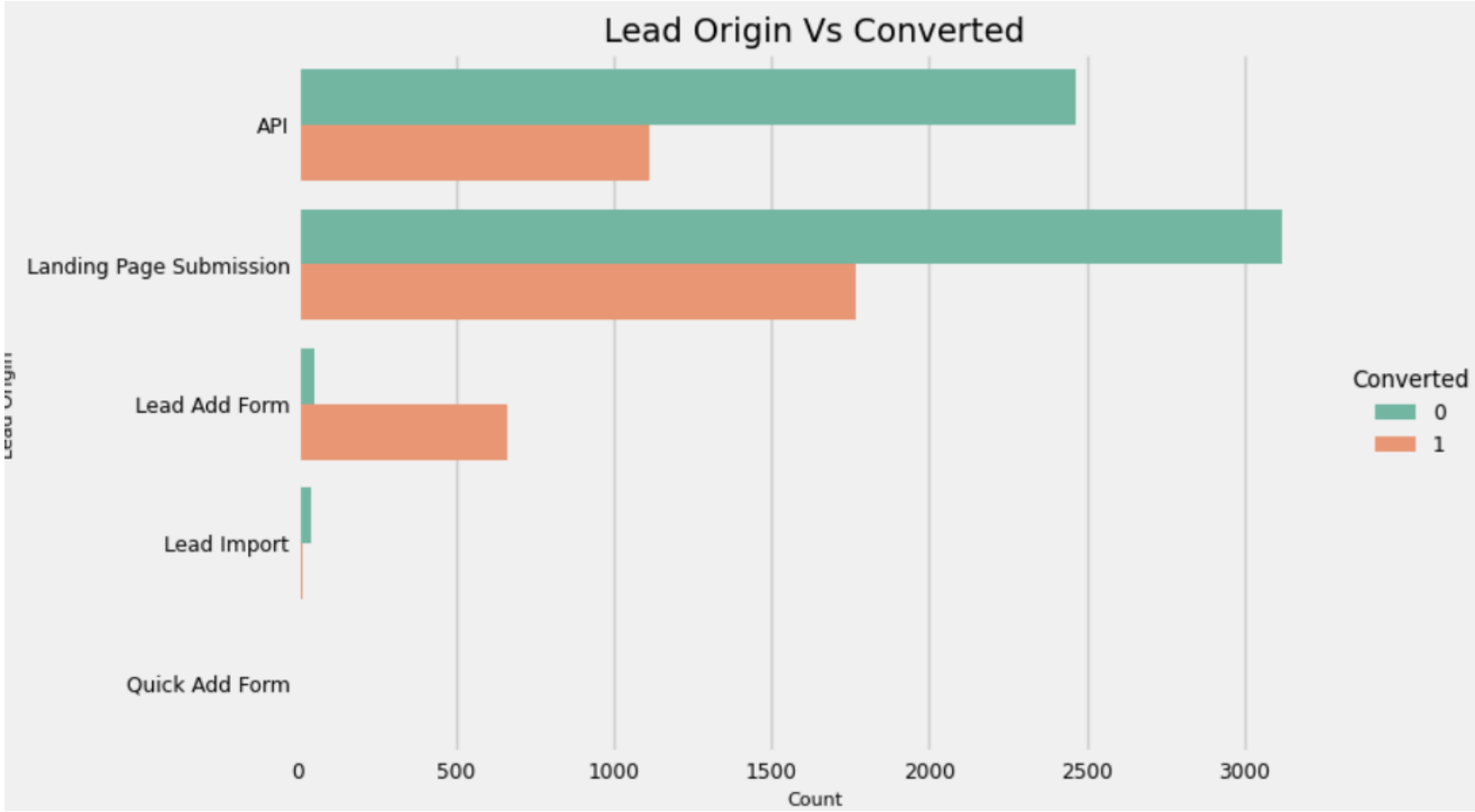


- **Lead Activity:** Email opened is the last activity of 38% of the users
- **Last Notable activity:** 38% of tge users have last notable activity as Modified

EDA – Bivariate Analysis for Categorical Variables

Lead Origin:

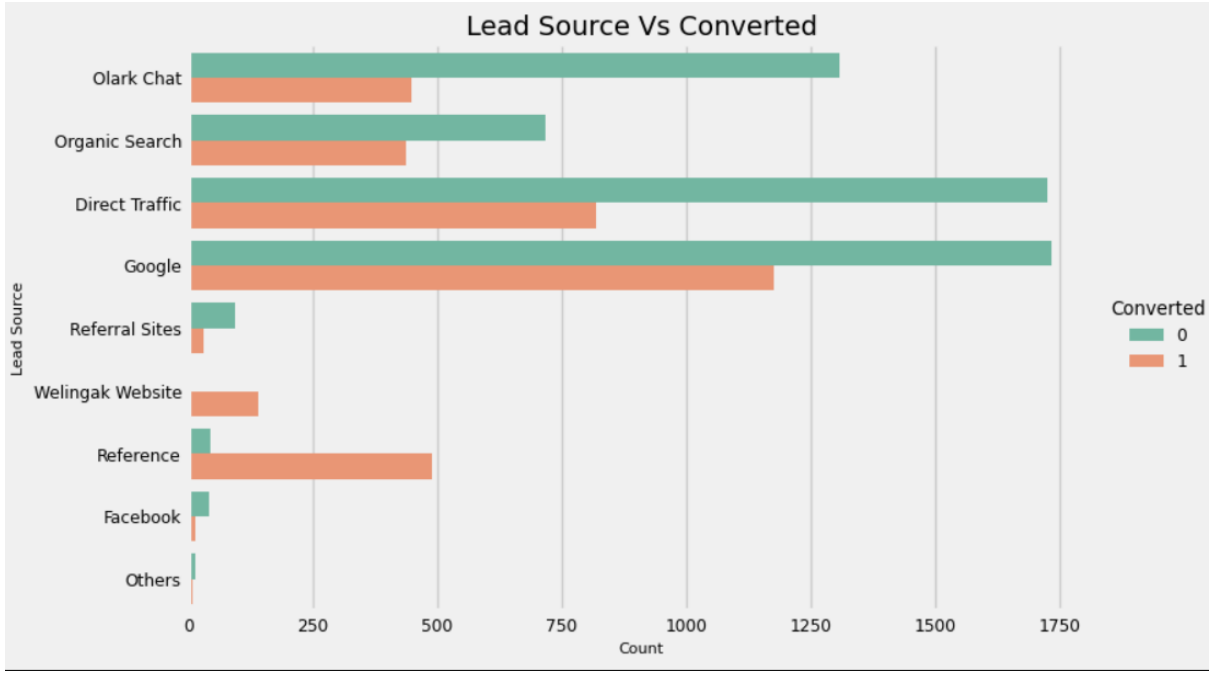
- Around 52% of all leads originated from *"Landing Page Submission"* with a **lead conversion rate (lead conversion rate) of 36%**.
- The *"API"* identified approximately 39% of customers with a **lead conversion rate (lead conversion rate) of 31%**.



EDA – Bivariate Analysis for Categorical Variables

Lead Source:

- *Google: Among the customers, 31% originate from Google. This source has a lead conversion rate of 40%, which is relatively high.*
- *Direct Traffic: Direct Traffic accounts for 27% of the customers, resulting in a lead conversion rate of 32%. Although this conversion rate is lower than Google, it still contributes significantly to lead conversions.*
- *Organic Search: Organic Search brings in approximately 12.5% of the customers, with a lead conversion rate of 37.8%. While the conversion rate is high, the customer contribution is relatively low compared to other sources.*
- *Reference: Reference has the highest lead conversion rate at 91%. However, only around 6% of the customers originate from this source. Despite the lower customer volume, it proves to be a highly effective lead source in terms of conversion rate.*

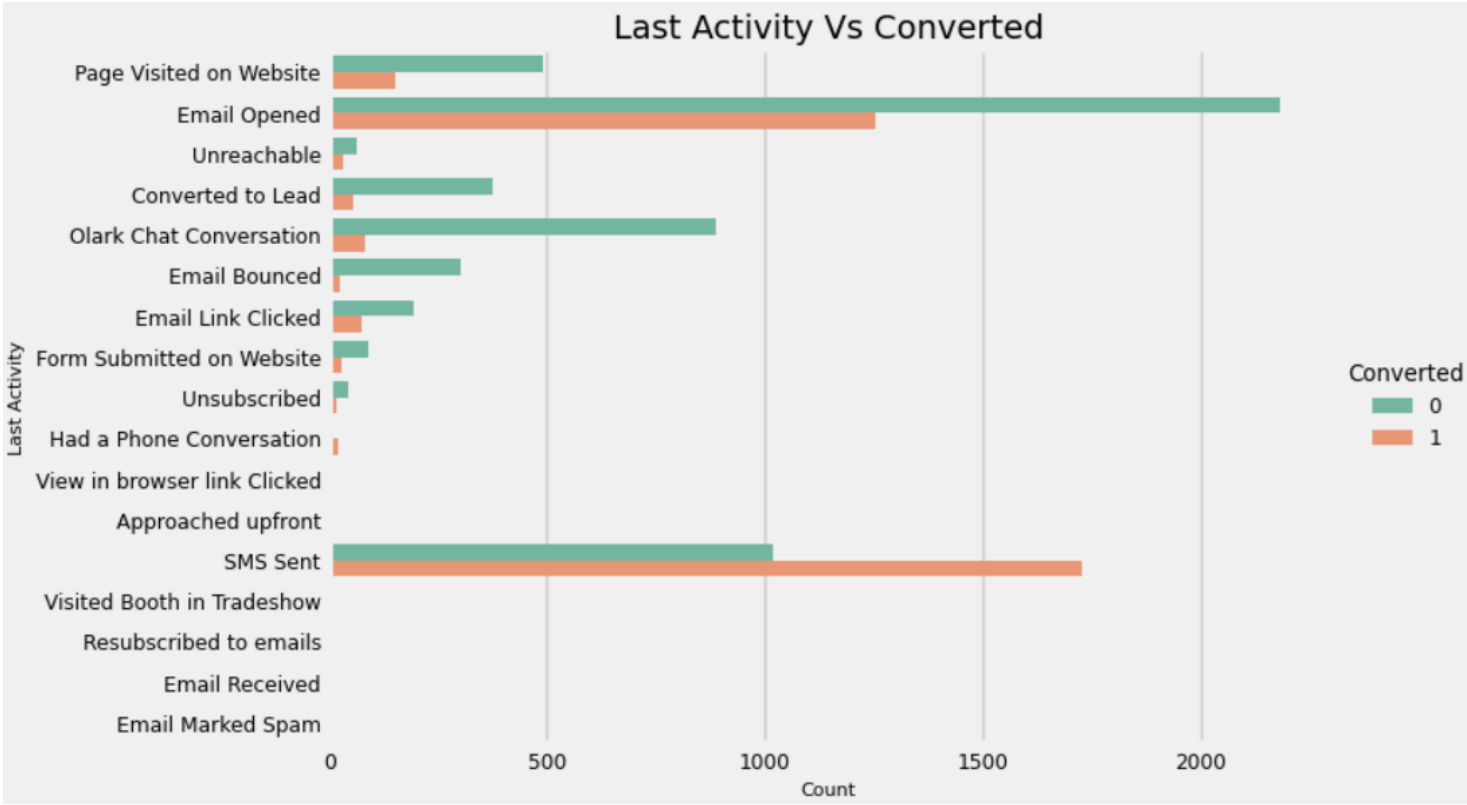


EDA – Bivariate Analysis for Categorical Variables

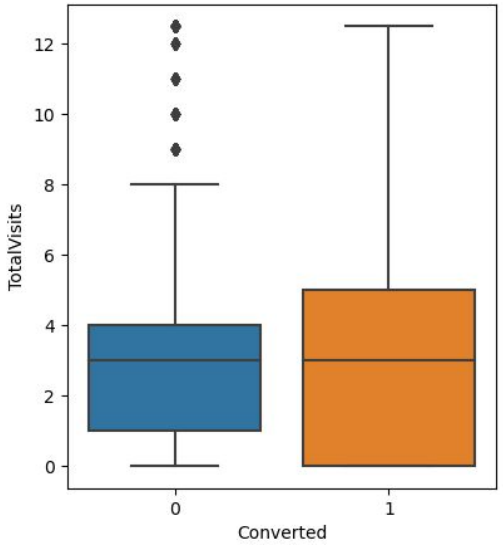
Last Activity:

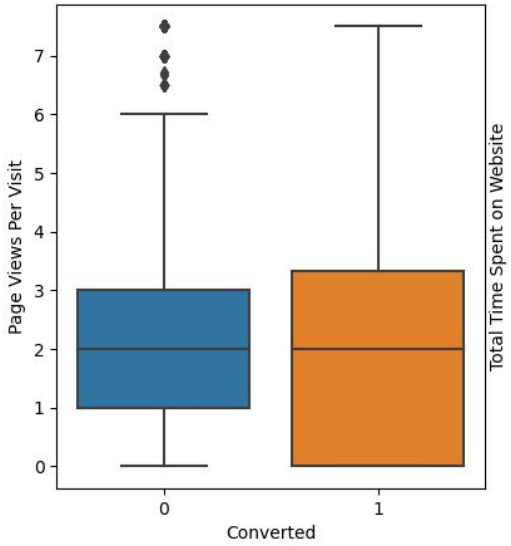
SMS Sent: The "SMS Sent" activity has a high lead conversion rate of 63%. This activity accounts for 30% of the last activities performed by customers. It proves to be an effective communication channel for driving conversions.

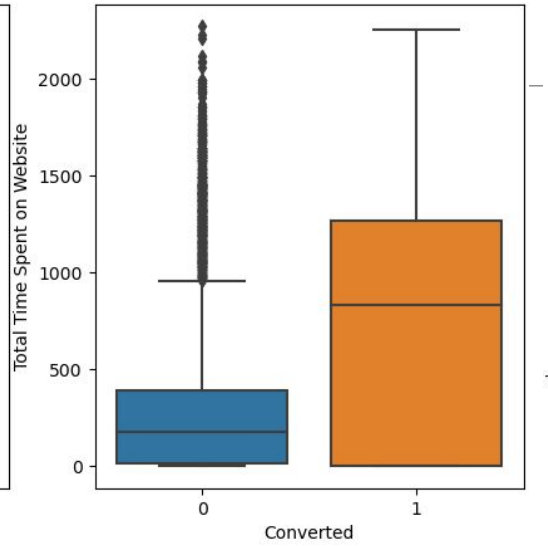
Email Opened: Among the last activities performed by customers, "Email Opened" contributes 38%. This activity has a lead conversion rate of 37%. It indicates that email communication is a significant factor in generating leads and achieving conversions.

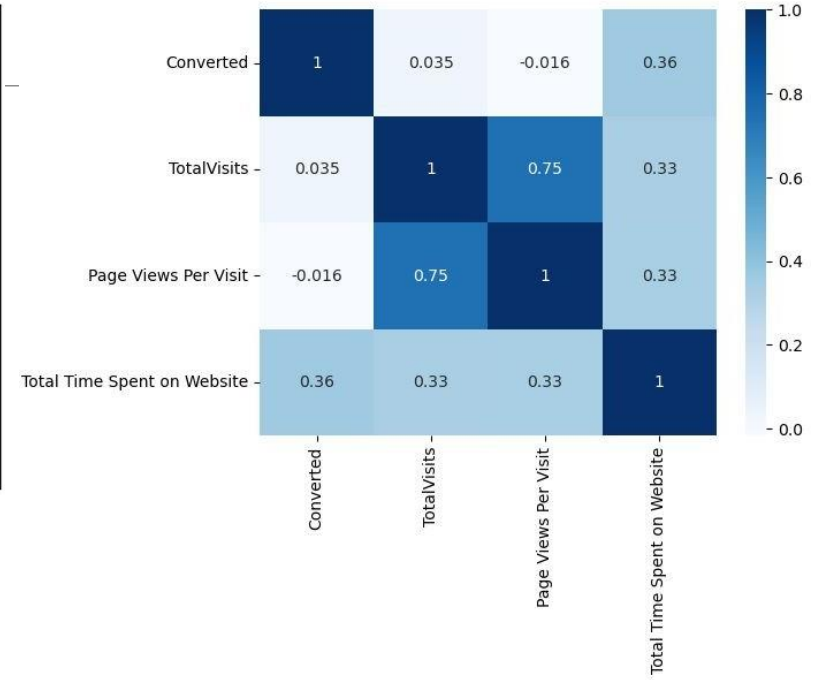


EDA – Bivariate Analysis for Numerical Variables









- Past leads who spend more time on the website have a higher chance of successful conversion compared to those who spend less time. This relationship is evident from the box plot analysis.
- The box plot shows that leads who spend more time on the website tend to have a higher median and a wider distribution of conversion rates. This indicates a positive correlation between website engagement and lead conversion.
- Conversely, leads who spend less time on the website have a lower median and a narrower distribution of conversion rates. This suggests that their conversion chances are relatively lower.

Data Preparation before Model building

- Binary level categorical columns were previously mapped to 1/0 in the earlier steps, ensuring consistency and compatibility for modeling.
- Dummy features were created for categorical variables such as Lead Origin, Lead Source, Last Activity, and last notable activity. This one-hot encoding technique allows for better representation of categorical data in the modeling process.
- The dataset was split into train and test sets, following a 70:30 ratio. This division ensures that a significant portion of the data is used for training the model, while the remaining portion is reserved for evaluation and testing.
- Feature scaling was applied to the dataset. Standardization, a common method, was used to scale the features, ensuring they have similar scales and preventing any undue influence from variables with larger values.
- Correlations between predictor variables were examined. Highly correlated variables, such as Lead Origin_Lead Import and Lead Origin_Lead Add Form, were identified and subsequently dropped from the dataset. This step helps to eliminate redundancy and minimize multicollinearity in the model.

Model Building

Feature Selection

- Given the large number of features and the high dimensionality of the dataset, it is crucial to perform feature selection to improve model performance and reduce computational time. Recursive Feature Elimination (RFE) is a popular technique used to select the most important features.
- By applying RFE, we were able to reduce the number of columns from 48 to 15, retaining only the most relevant and informative features for our model. This step helps to streamline the dataset and improve the efficiency of subsequent modeling processes.
- In addition to RFE, manual fine-tuning of the model was conducted. This involved further evaluation and elimination of variables based on their statistical significance and relevance to the target variable. By considering p-values and other statistical metrics, we ensured that only the most significant features were retained in the final model.
- The outcome of the feature selection process resulted in a refined dataset with 15 important columns, significantly reducing dimensionality and enhancing the model's performance and interpretability.

Model Building

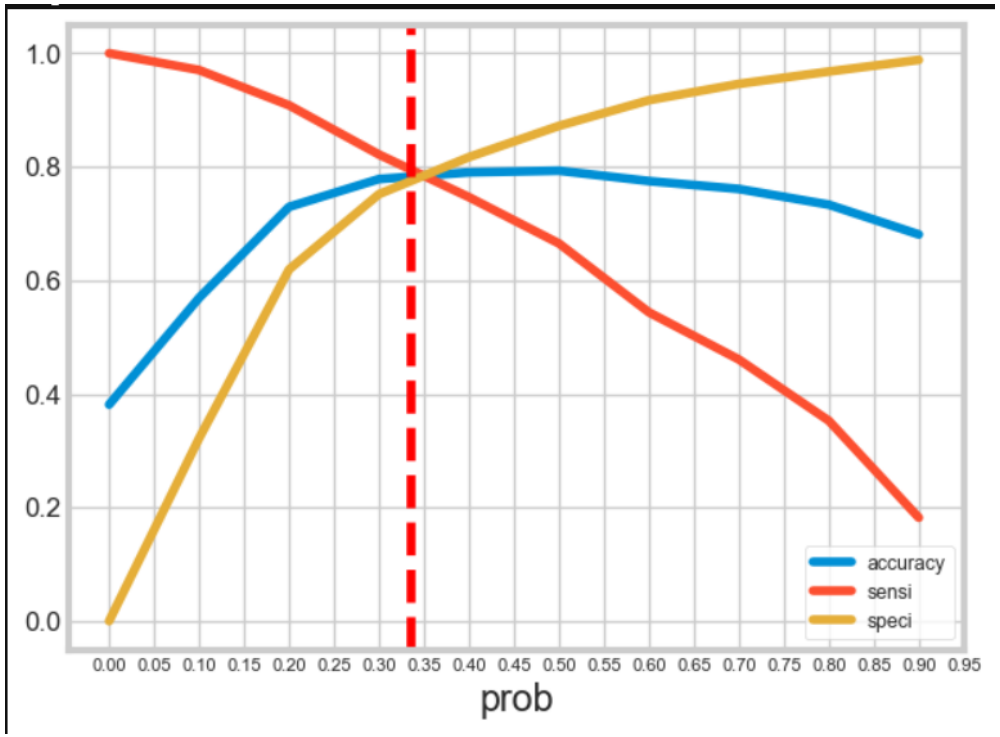
Manual Feature Reduction and Final Model Selection:

- To further refine the model, a manual feature reduction process was implemented based on the p-values of the variables. Variables with p-values greater than 0.05 were systematically dropped from the model. This step helped to eliminate less significant variables that did not contribute significantly to the prediction of the target variable.
- After four iterations, Model 10 emerged as a stable model. It exhibited significant p-values within the desired threshold (p-values < 0.05), indicating the statistical significance of the retained variables. Additionally, Model 10 demonstrated no signs of multicollinearity, as evidenced by Variance Inflation Factors (VIFs) below 5. This further validated the stability and reliability of the model.
- As a result, logm10 was selected as the final model for further evaluation and analysis. This model will be used to assess the performance of the predictive model and make accurate predictions on new data.
- By applying a combination of manual feature reduction, statistical analysis, and stability assessment, we have arrived at a robust and effective model, logm10, which will serve as the foundation for evaluation and prediction in our analysis.

Model Evaluation

Train Data Set

It was decided to go ahead with 0.34 as cutoff after checking evaluation metrics coming from both plots



Train Data Set Confusion Matrix

```
array([[3114,  888],
       [ 509, 1957]], dtype=int64)
```

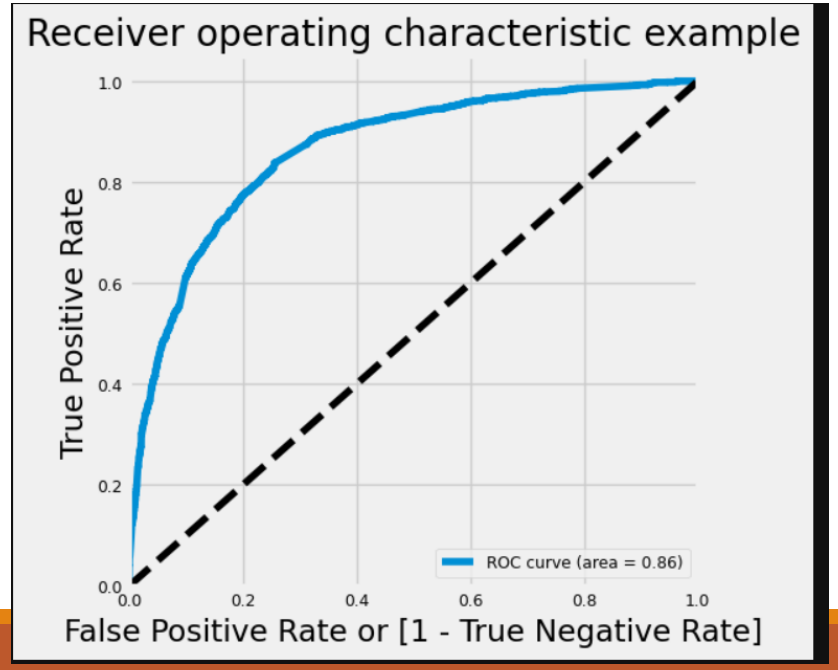
Test Data Set Confusion Matrix

```
array([[1310,  367],
       [ 226,  869]], dtype=int64)
```

Model Evaluation

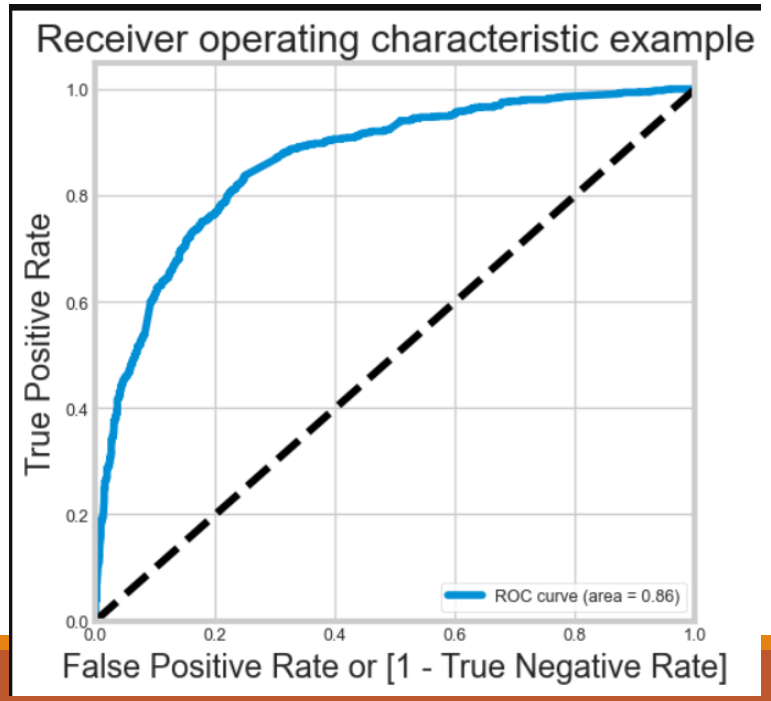
ROC Curve – Train Data Set

- Area under ROC curve is 0.86 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



ROC Curve – Test Data Set

- Area under ROC curve is 0.86 out of 1 which indicates a good predictive model.
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



Model Evaluation

For the training data, the model's performance metrics are as follows:

Accuracy: 0.784 (78.4%)

Precision: 0.688 (68.8%)

Recall: 0.794 (79.4%)

F1 Score: 0.737 (73.7%)

These metrics indicate how well the model performs on the training data. The accuracy represents the overall correctness of the model's predictions, while precision measures the proportion of correctly predicted positive instances out of the total predicted positives. Recall, or sensitivity, indicates the proportion of actual positive instances that were correctly predicted. The F1 score combines precision and recall into a single metric that balances both measures.

It's important to consider these metrics when evaluating the model's performance on the training data, as they provide insights into the model's ability to capture patterns and make accurate predictions. These metrics can help assess the effectiveness of the model in achieving the desired objective and guide any necessary adjustments or improvements.

The model's performance metrics for the given classification task for test data are as follows:

Accuracy: 0.786 (78.6%)

Precision: 0.703 (70.3%)

Recall: 0.794 (79.4%)

F1 Score: 0.746 (74.6%)

These metrics provide an evaluation of the model's effectiveness in predicting and classifying the target variable. The accuracy represents the overall correctness of the model's predictions, while precision measures the proportion of correctly predicted positive instances out of the total predicted positives. Recall, also known as sensitivity, indicates the proportion of actual positive instances that were correctly predicted. The F1 score combines precision and recall into a single metric that balances both measures.

These performance metrics can be used to assess the effectiveness of the model in achieving the desired outcome and can guide further improvements or decision-making processes related to the classification task.

Recommendation based on Final Model

- Based on our model, we have identified the features with their corresponding coefficients that have a significant impact on lead conversion. These coefficients indicate the relative importance of each feature in influencing the conversion rate. Here are the updated coefficients for the parameters you provided:
- **Positive Impact:**
  - TotalVisits: 1.27
  - Total Time Spent on Website: 6.63
  - Lead Source_Olark Chat: 2.26
  - Lead Source_Reference: 7.57
  - Lead Source_Welingak Website: 10.00
  - Last Activity_Had a Phone Conversation: 3.47
  - Last Activity_SMS Sent: 1.23
  - Last Notable Activity_Unreachable: 1.61
- **Negative Impact:**
  - Last Activity_Email Bounced: -2.89
  - Last Activity_Olark Chat Conversation: -1.62
  - Last Notable Activity_Email Link Clicked: -2.06
  - Last Notable Activity_Email Opened: -1.08
  - Last Notable Activity_Modified: -2.41
  - Last Notable Activity_Olark Chat Conversation: -2.07
  - Last Notable Activity_Page Visited on Website: -2.01
- These coefficients indicate the influence of each feature on lead conversion. Features with positive coefficients have a positive impact on conversion, while features with negative coefficients have a negative impact. By prioritizing the features with higher positive coefficients and addressing the areas related to negative coefficients, we can enhance our marketing and sales efforts to maximize lead conversion for X Education.

Recommendation based on Final Model

To increase our lead conversion rates and maximize our marketing efforts, we should consider the following strategies:

1.Lead Source Optimization: Identify the top-performing lead sources based on their positive coefficients and allocate resources to optimize lead acquisition from these sources. Implement tailored campaigns and tactics to attract high-quality leads.

2.Communication Channel Optimization: Analyze the impact of different communication channels on lead engagement, considering the coefficients. Focus on optimizing the channels that have shown a positive influence on lead conversion, ensuring effective and tailored communication with potential customers.

3.Increased Investment in Welingak Website: Allocate a higher budget for advertising and promotion of the Welingak Website. This platform has demonstrated a significant positive impact on lead conversion, making it a valuable investment for maximizing conversion rates.

4.Incentivize Referrals: Offer incentives or discounts to customers who provide references that convert into leads. Encourage them to provide more references by emphasizing the benefits of their referrals and the positive impact they can have on lead conversion.

To identify areas for improvement, focus on the following:

1.Specialization Offerings: Analyze the negative coefficients associated with specialization offerings. Review and enhance the specialization options, ensuring they align with the preferences and demands of potential customers.

2.Landing Page Submission Process: Review the landing page submission process to identify any areas that may hinder lead conversion. Streamline the process, remove any potential barriers, and optimize the user experience to improve conversion rates.

Thank You!