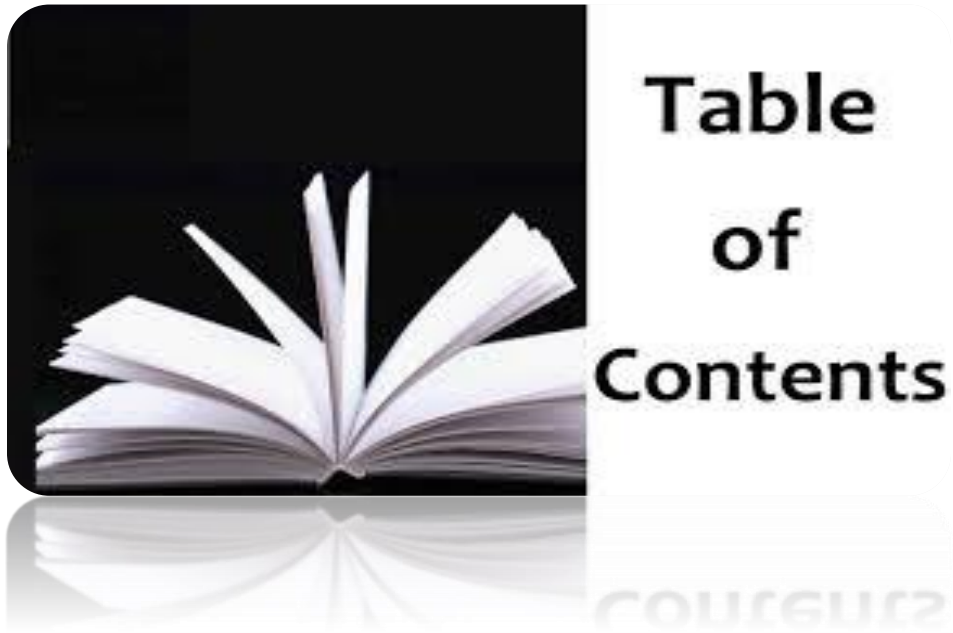
LEAD SCORING CASE STUDY

TABLE ON CONTEXT

- Problem Statement
- Goals of the case the case study
- Lead Conversion Process
- Steps Followed
- Data Wrangling
- EDA
- Data Preparation
- Model Building
- Model Evaluation
- Conclusion



PROBLEM STATEMENT

- X Education has a high number of leads, but their lead conversion rate is only about 30%.
- X Education is looking to increase the efficiency of their lead conversion process by identifying the most promising leads, also known as Hot Leads.
- The sales team wants to prioritize communication with these potential leads instead of reaching out to everyone.
- Our objective is to help X Education select the most promising leads - those with the highest likelihood of converting into paying customers.
- We will build a model to assign a lead score to each lead, with higher scores indicating a higher chance of conversion and lower scores indicating a lower chance of conversion.
- The CEO has set a target lead conversion rate of approximately 80%.

Goals of the Case Study

Few goals for this case study:

- Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

- There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Lead Conversion Process

The funnel shown here can be used to illustrate this:

- Since most leads are created in the top stage, when they are most likely to become paying customers, we need to appropriately nurture the potential leads in order to enhance lead conversion.
- By teaching them about the product and keeping in constant communication.

Steps Followed

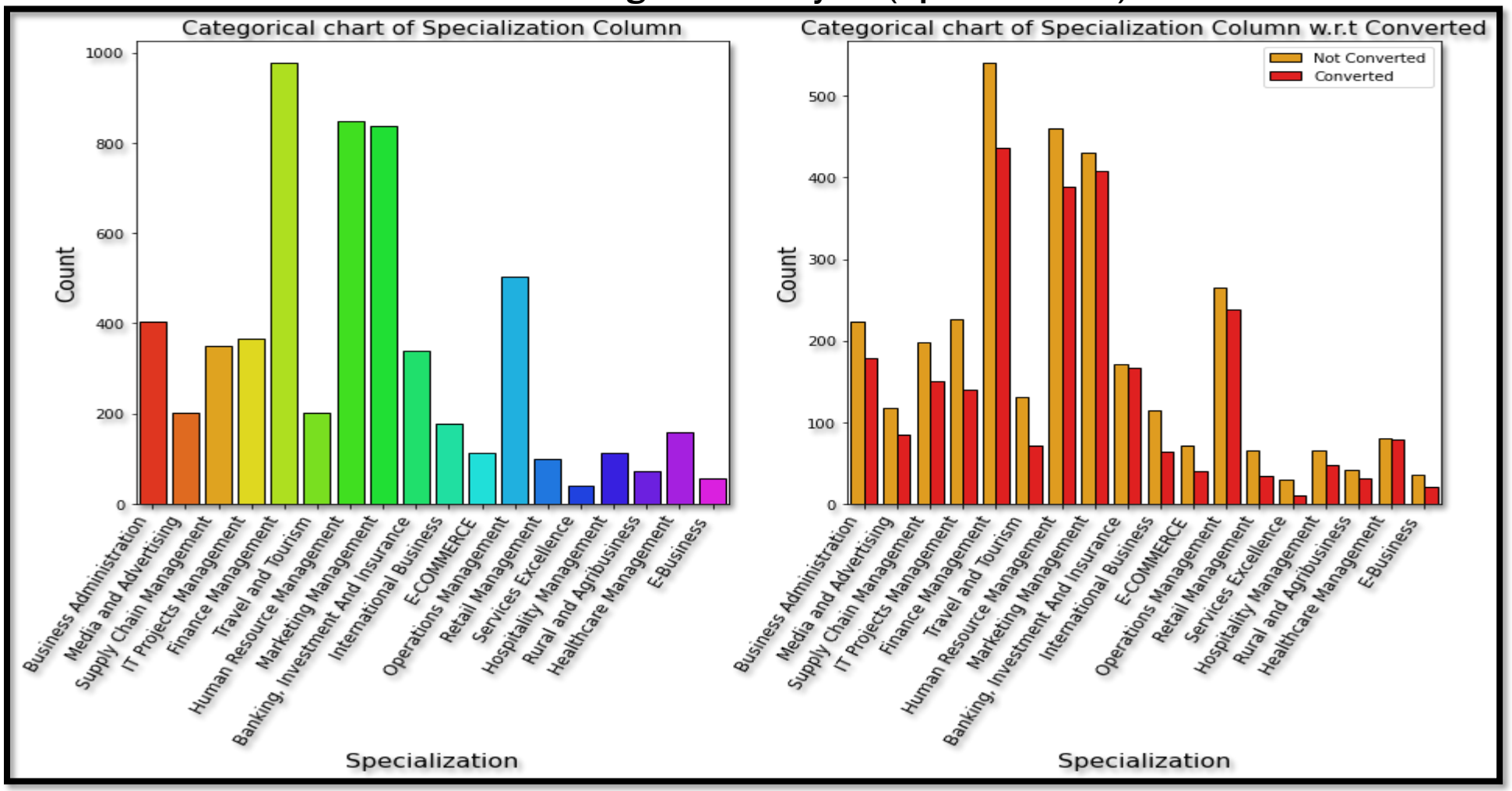
- Importing necessary libraries
- Importing the provided dataset
- Data Understanding & Cleaning
- Exploratory Data Analysis (Variables Inspection)
- Data Preparation
- Model Building (Logistic Regression)
- Model Evaluation (Logistic Regression Metrics)
- Model Testing
- Model Inference
- Conclusion based on our results

Data Wrangling :

- Importing dataset
- Check overall dataset and make key observations.
- Check overall dimensions of the dataset.
- Verify column formats and correct any inconsistencies found in dataset.
- Examine for any NULL values present in the dataset.
- Impute those rows or substitute mean or median values to deal with null values in data set.

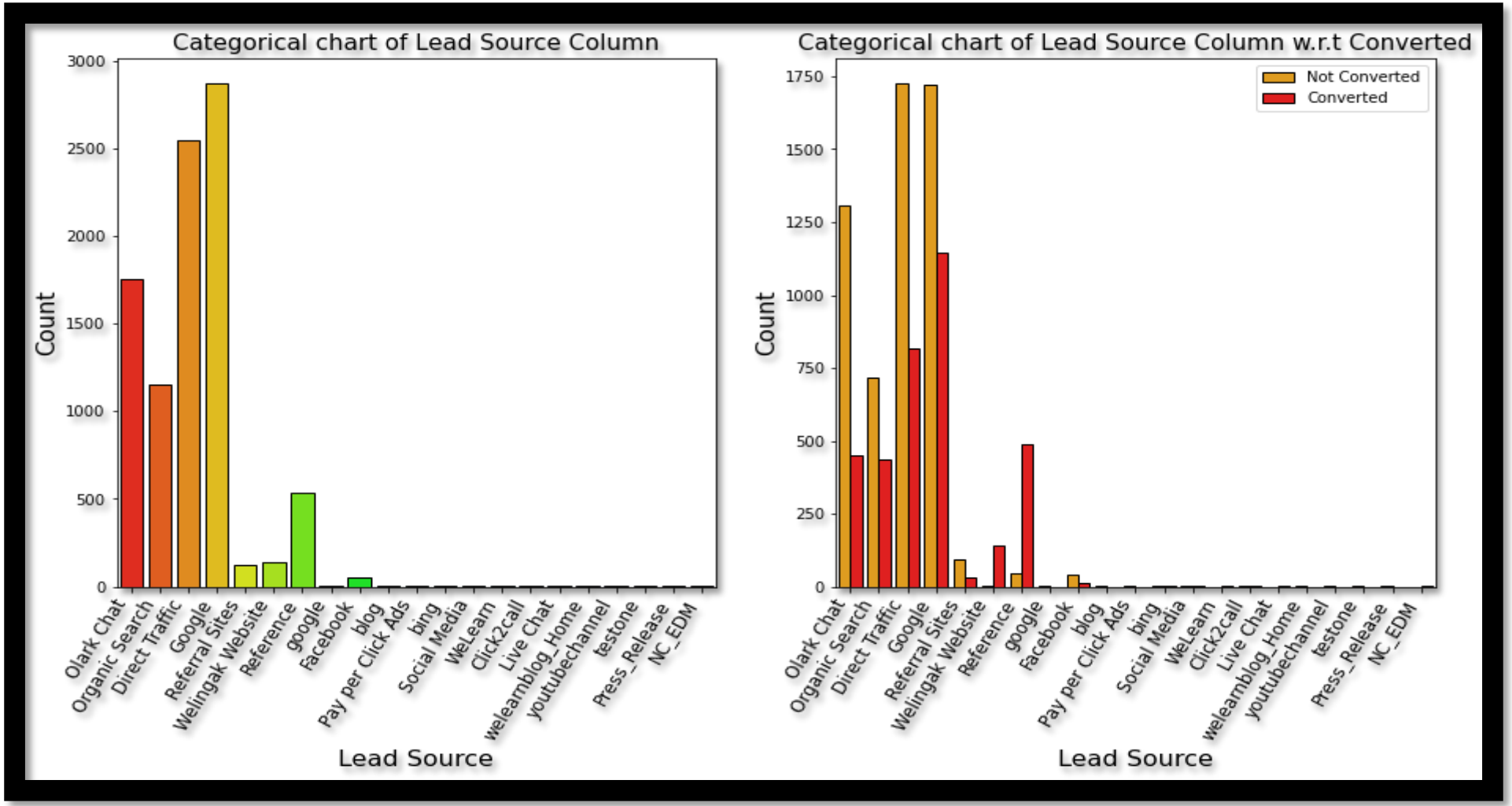
Exploratory Data Analysis

Bivariate Categorical Analysis (Specialization)



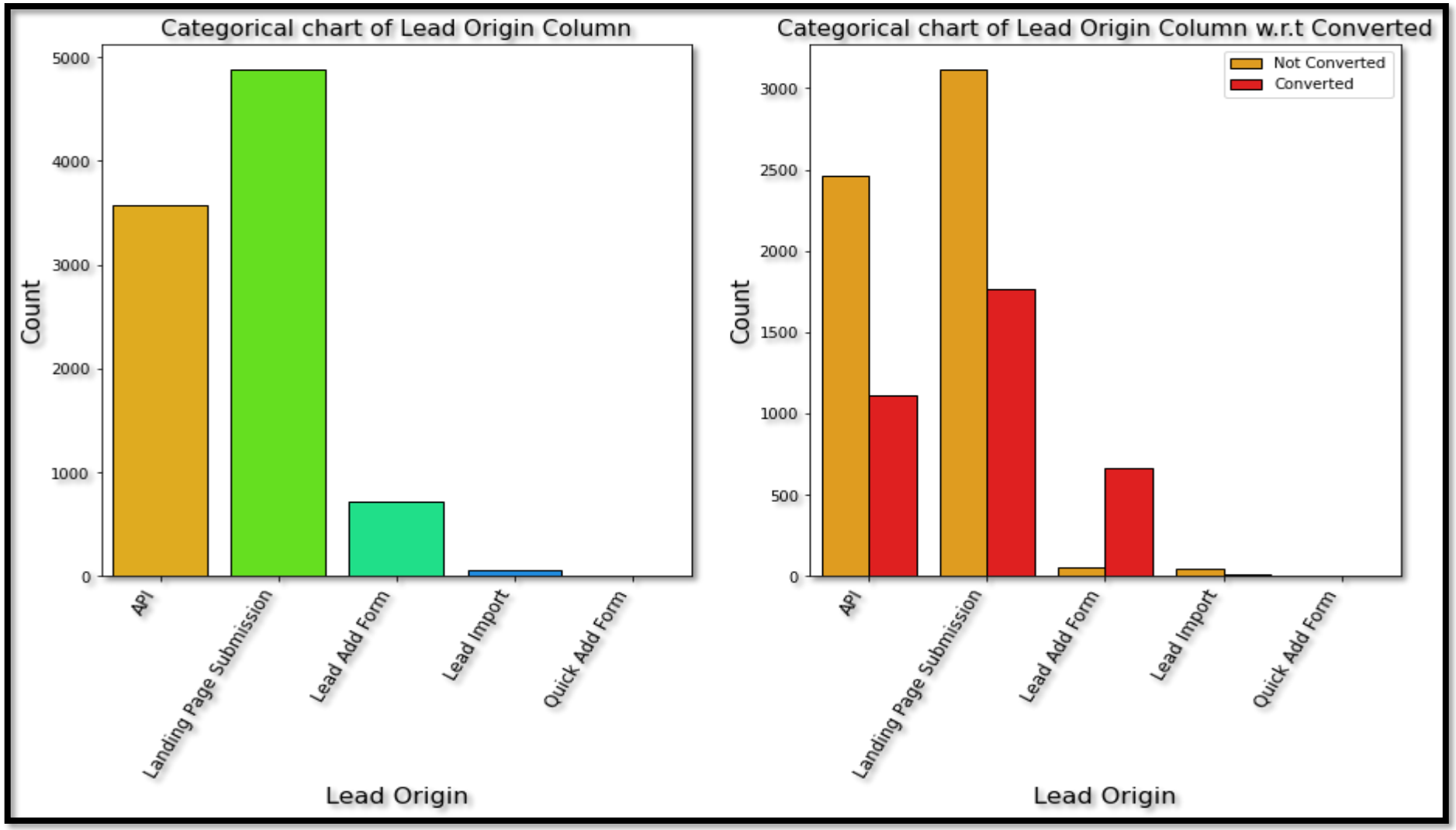
- Other (i.e select) category is having most followed by Finance ,Marketing , HR departments respectively where mostly customers are working.

Bivariate Categorical Analysis (Lead source)



– Most of the leads are from Google, but leads converted by referrals have the same conversion rate as Google leads.

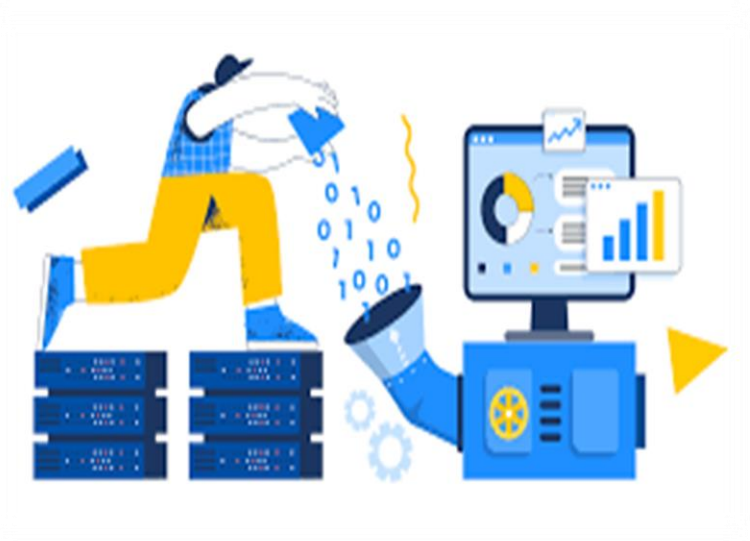
Bivariate Categorical Analysis (LEAD ORIGIN)



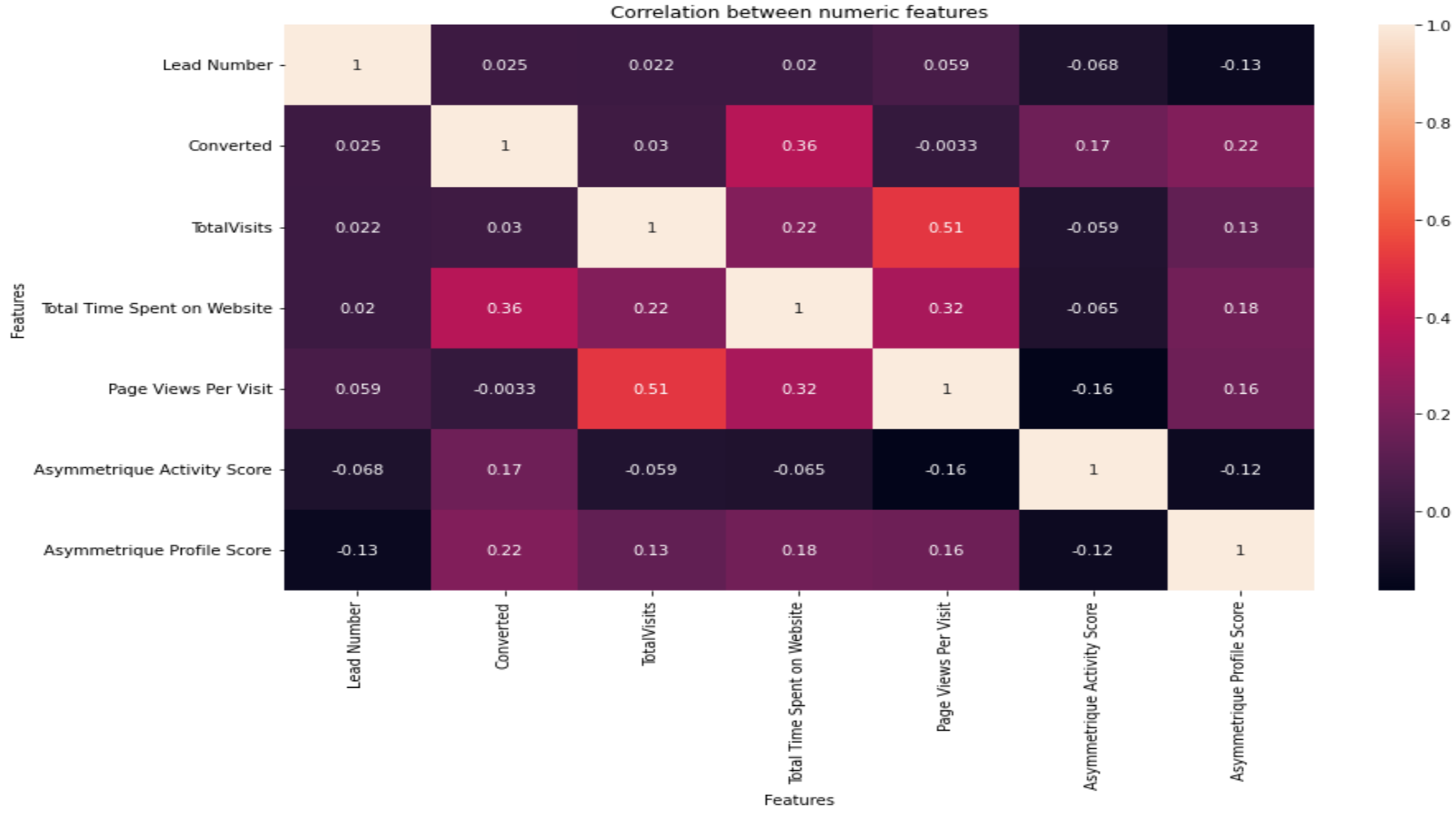
- Most Conversions are from Landng Page Submissions and Lead Add Form.

DATA PREPARATION

- Binary-level categorical columns were previously mapped to 1/0.
- Dummy features (one-hot encoded) were created for the following categorical variables: Lead Origin, Lead Source, Do Not Email, Last Activity, Specialization, Current Occupation, Tags, City, A Free Copy of Mastering the Interview, and Last Notable Activity.
- The dataset was split into training and test sets using a 70:30 ratio.
- Feature scaling was performed using the standardization method.



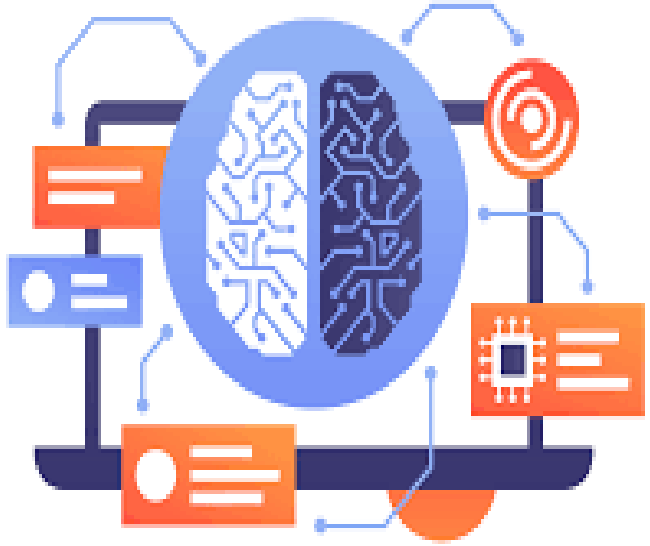
Correlation Plot



- Total Time spent on the website have highest correlation with the converted.

MODEL BUILDING

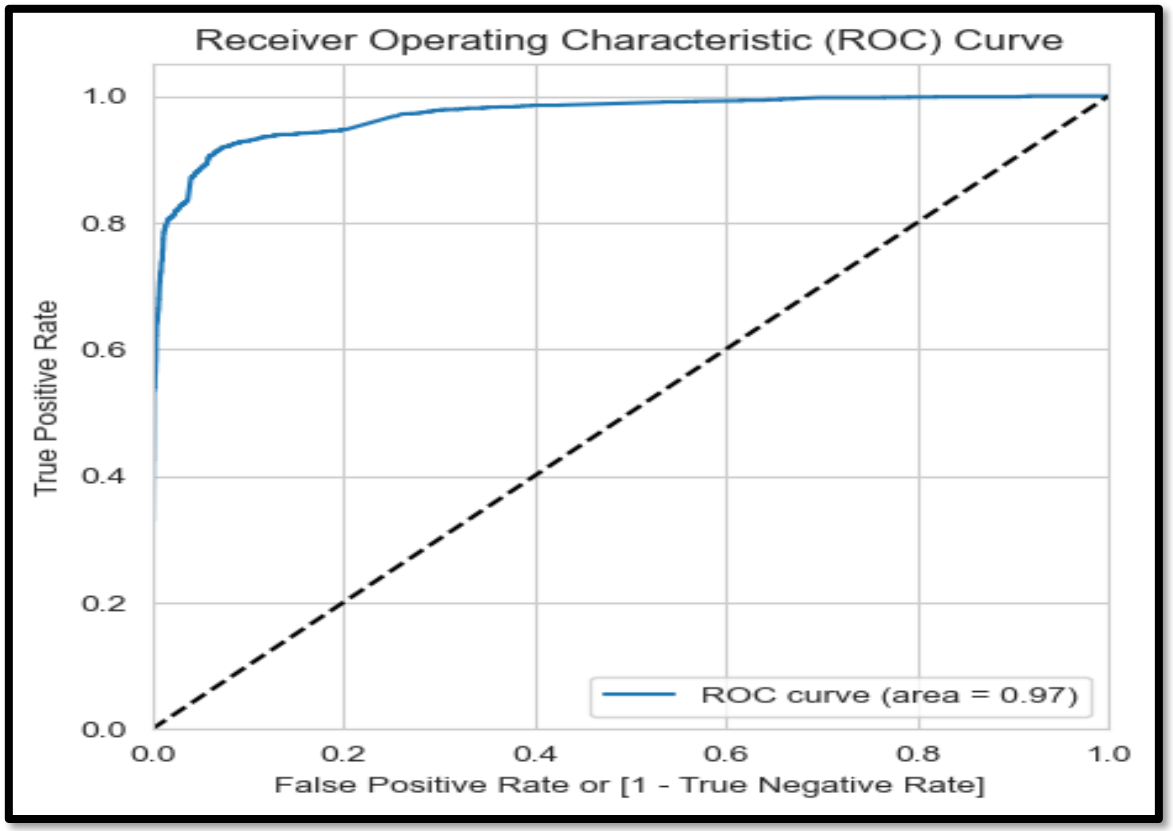
- The dataset has many dimensions and features, which can reduce model performance and increase computation time.
- Recursive Feature Elimination (RFE) was used to select important columns.
- Manual feature reduction was performed by dropping variables with p-values greater than 0.05.
- Model 3 was stable after three iterations, with all p-values < 0.05.
- No multicollinearity was detected, as all VIFs were < 5.
- Model 3 is the final model for evaluation and predictions.

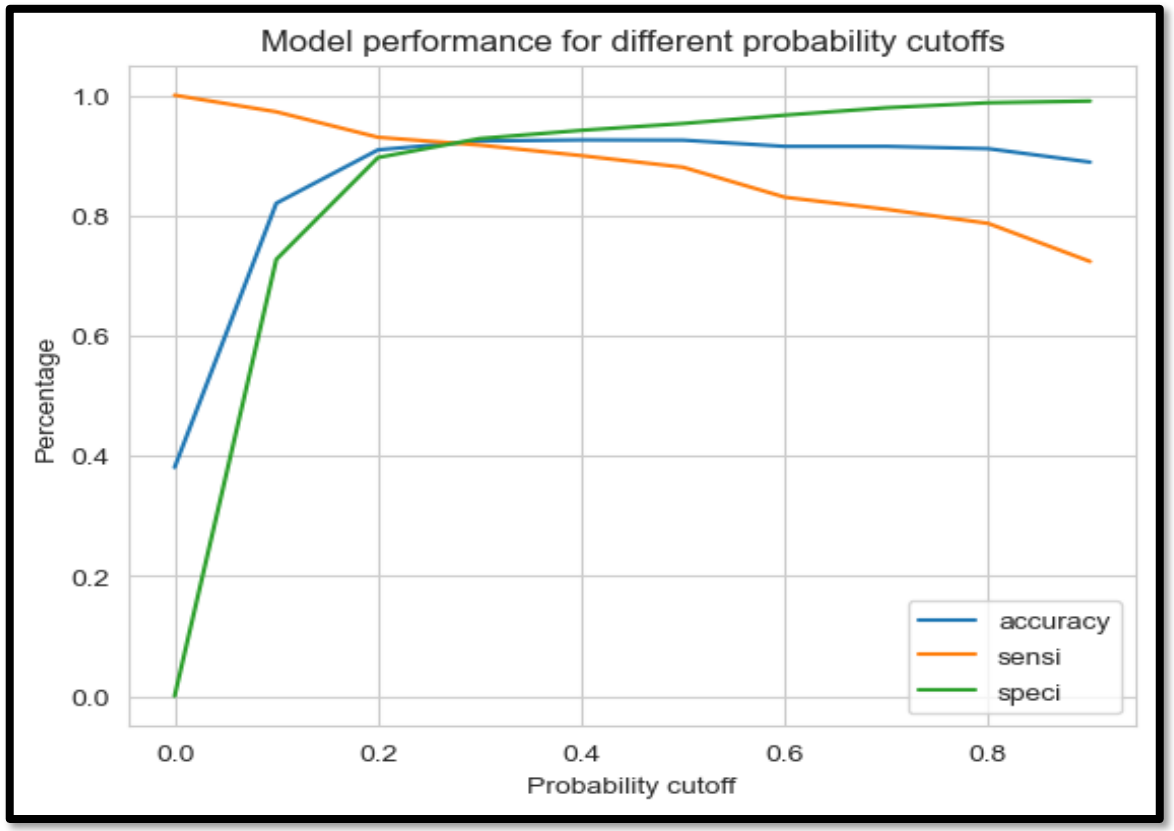


MODEL EVALUATION

Metrics	Scores
Accuracy score	0.91
F1- score	0.93
Precision score	0.89
Recall score	0.91

ROC CURVE





- Since the ROC curve value is 0.97, which is close to 1, this indicates that our predictive model is performing well.

- The ROC curve has a value of 0.97, indicating high performance.
- Training data accuracy is 92.34%.
- Sensitivity is 91.69%.
- Specificity is 92.73%.

Conclusion:

- Conversions are higher for leads from Google, Organic Search, Direct Traffic, and Referrals.
- SMS and Email marketing leads have higher conversion rates.
- The Lead Add Form generates qualifying leads and should be used across key areas.
- Leads with a Lead Score >0.35 tend to convert more, and the model accuracy score is 91%.
- The sales team should focus on working professionals for higher conversions.
- Leads spending more time on the website tend to convert more.
- Reducing the website bounce rate can increase customer engagement time and conversions.
- Landing Page Submissions and the Lead Add Form lead to more conversions.

THANK YOU