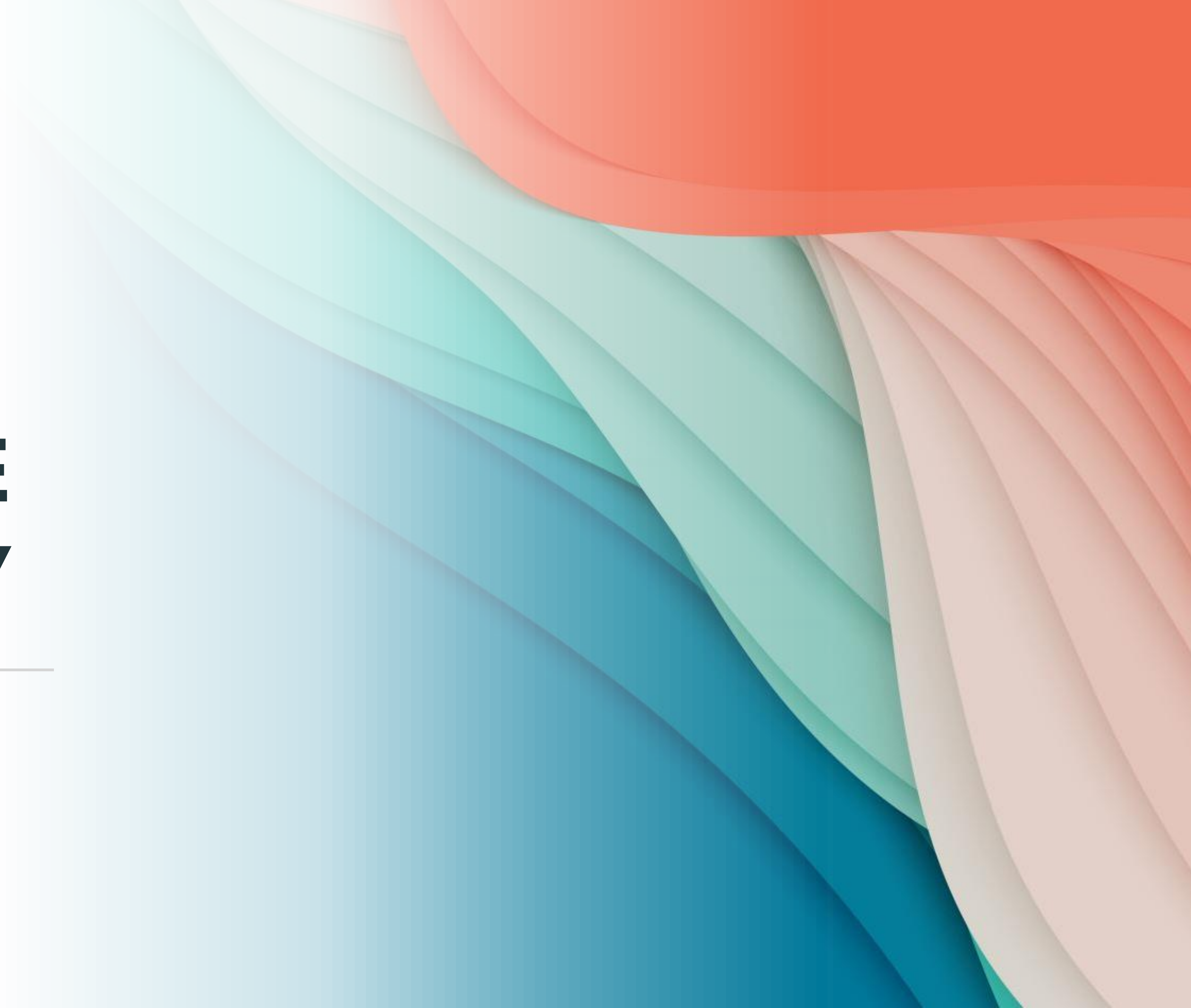
LEAD SCORE CASE STUDY



PRESENTED BY:

- K SAI KRANTH
- OMKAR PRALHAD PAWAR
- SANGAYYA HIREMATH

TABLE OF CONTENTS

- **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

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.

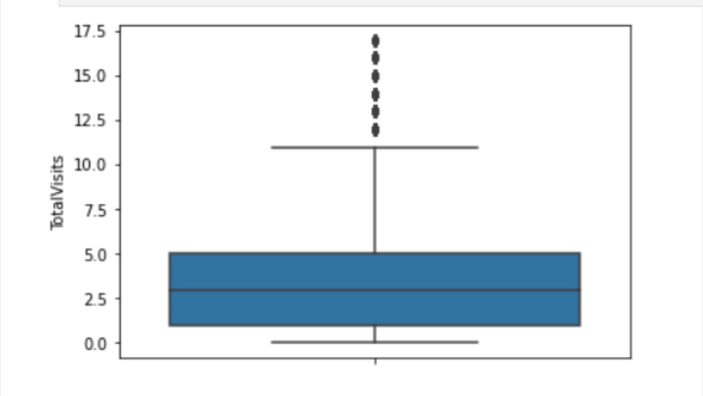
STEPS USED

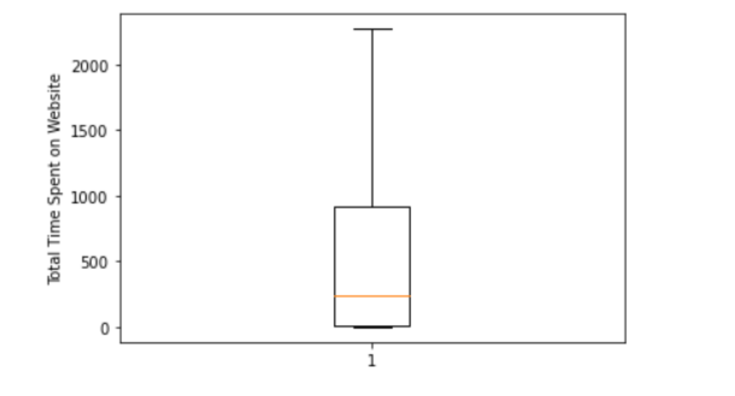
- Importing necessary libraries Importing the provided dataset
- Data Understanding & Cleaning
- Exploratory Data Analysis (Variables Inspection) (EDA)
- Data Preparation
- Model Building (Logistic Regression)
- Model Evaluation (Logistic Regression Metrics)
- Model Testing
- Model Inference
- Conclusion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

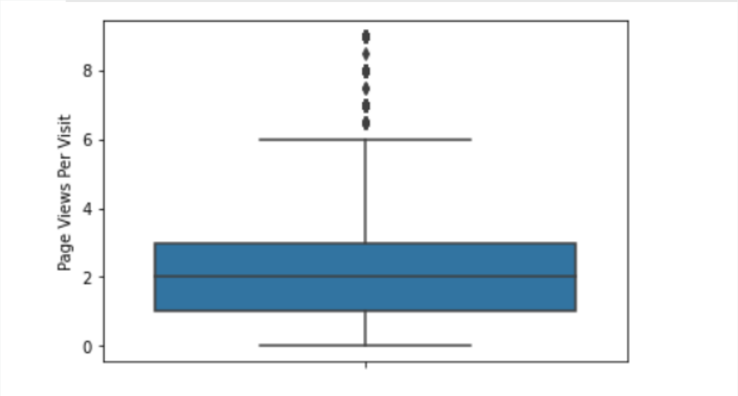
DATA WRANGLING

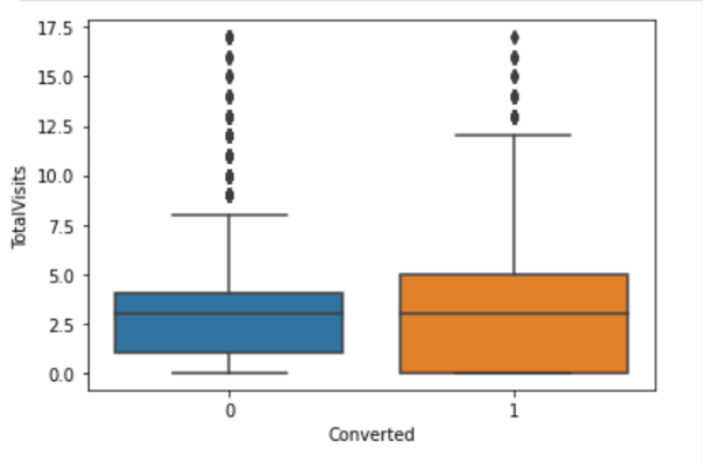
- Import dataset
- Go through the entire dataset and make key observations.
- Check overall dimensions of the dataset.
- Check column formats and correct any irregularities found in dataset.
- Check for any NULL values present in the dataset.
- Deal with NULL values by imputing those rows or replacing with mean or median values.

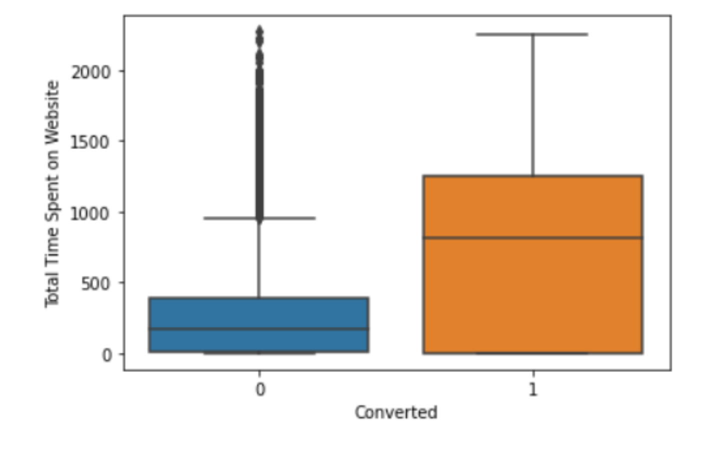
Outlier Check and Missing Value treatment in both Categorical and Numerical Columns:

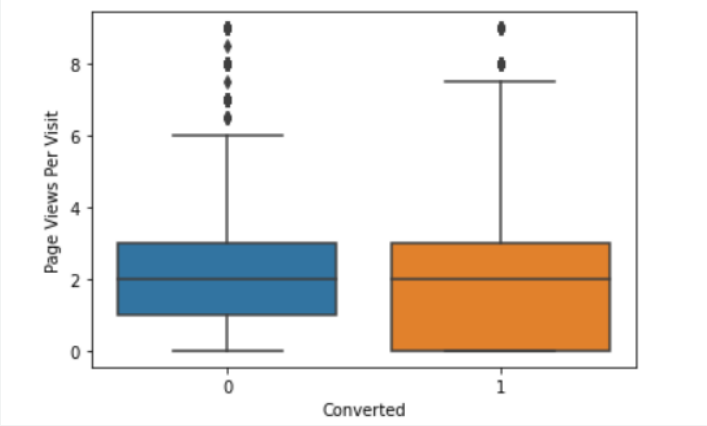






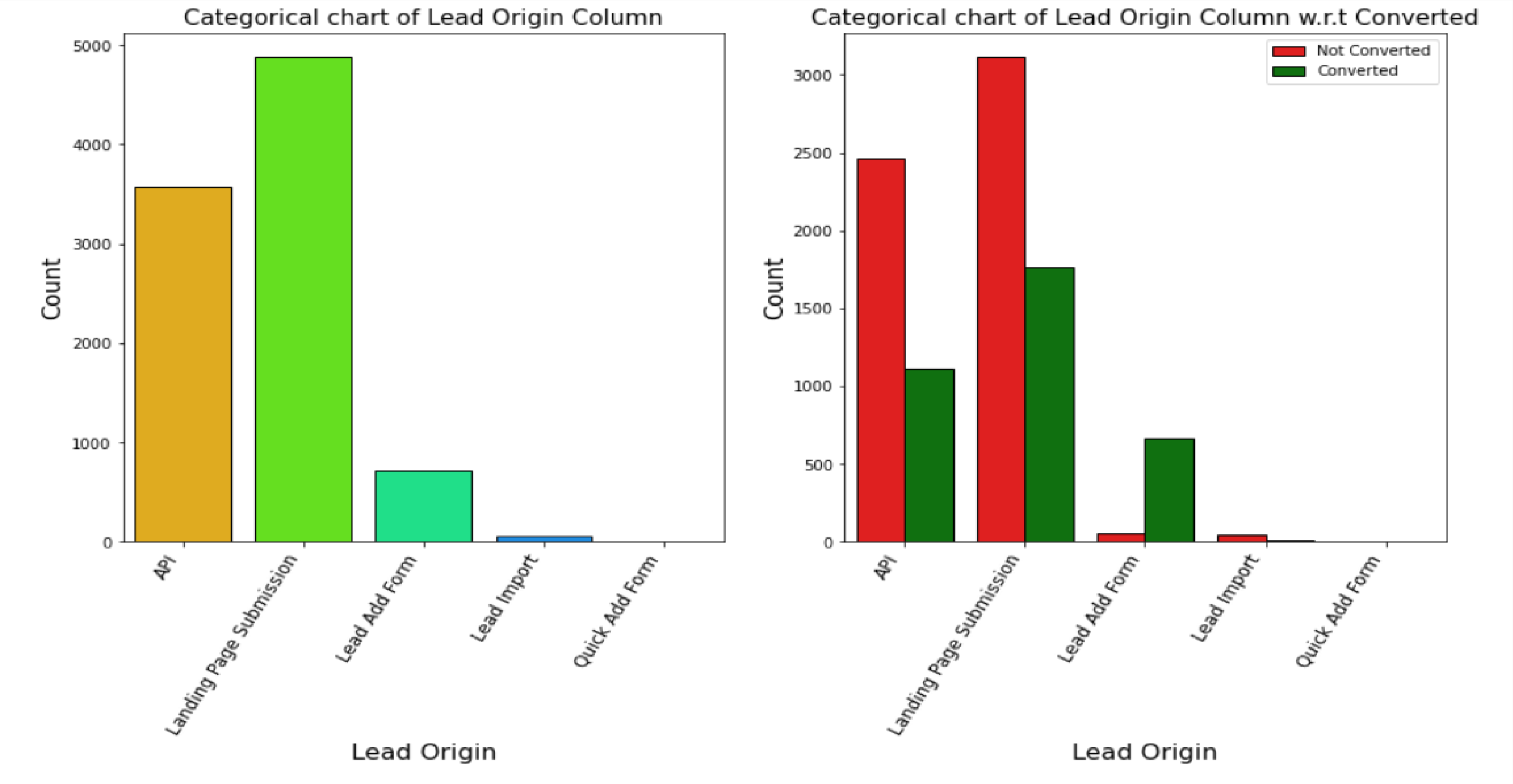






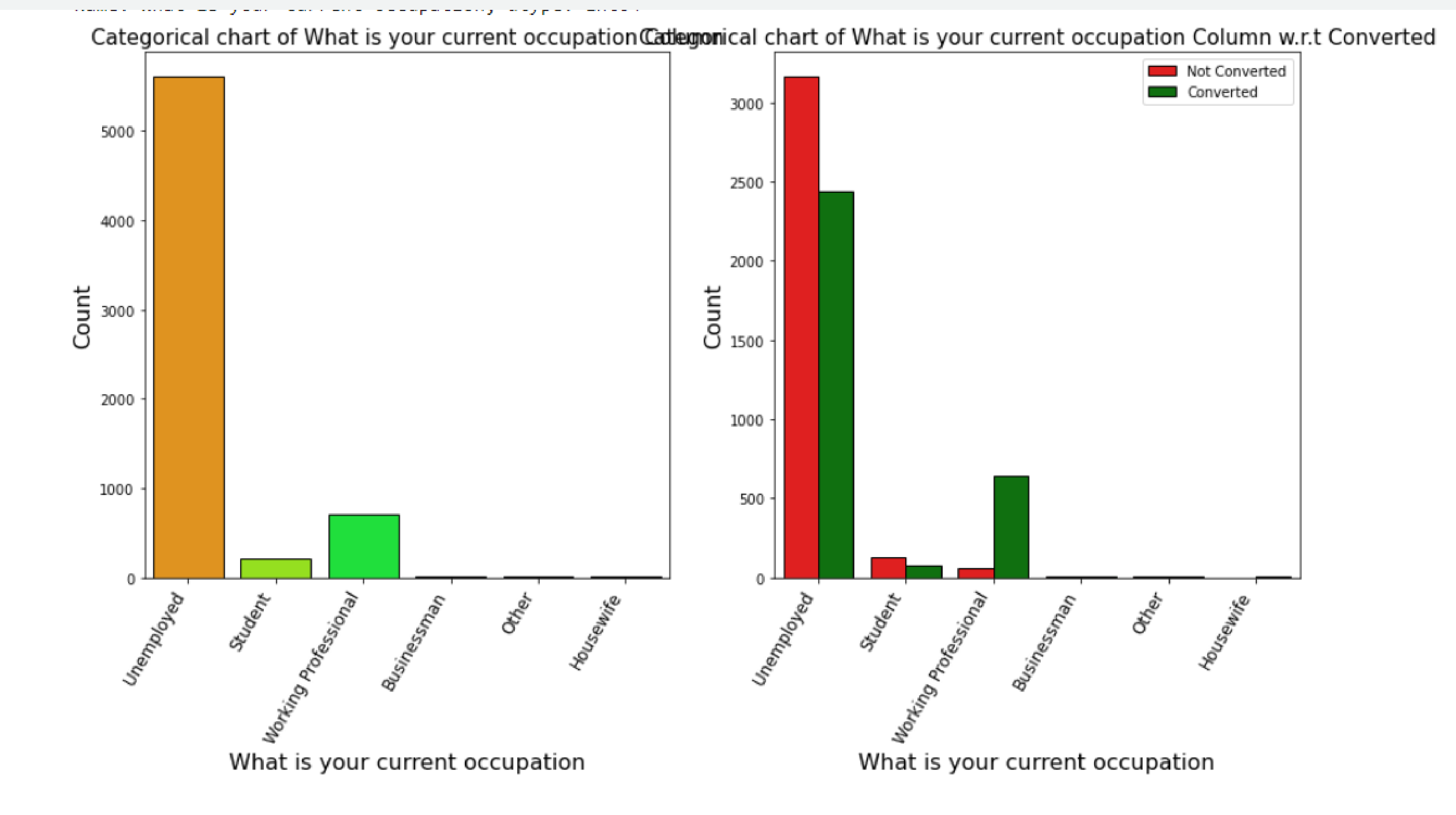
Exploratory Data Analysis (EDA)

Bivariate Categorical Analysis (LEAD ORIGIN)



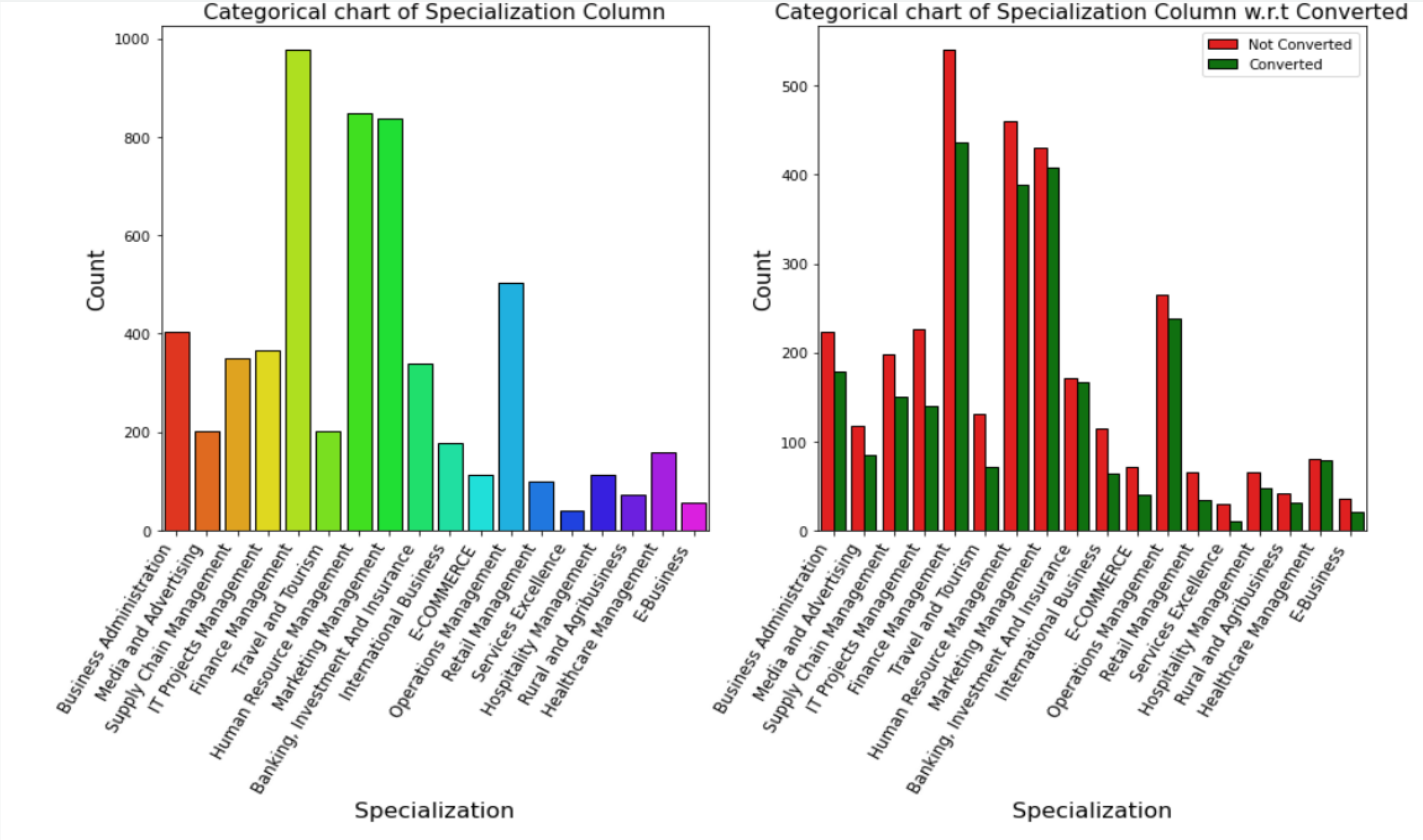
- Most Conversions are from Landing Page Submissions and Lead Add Form.

Bivariate Categorical Analysis (Occupation)



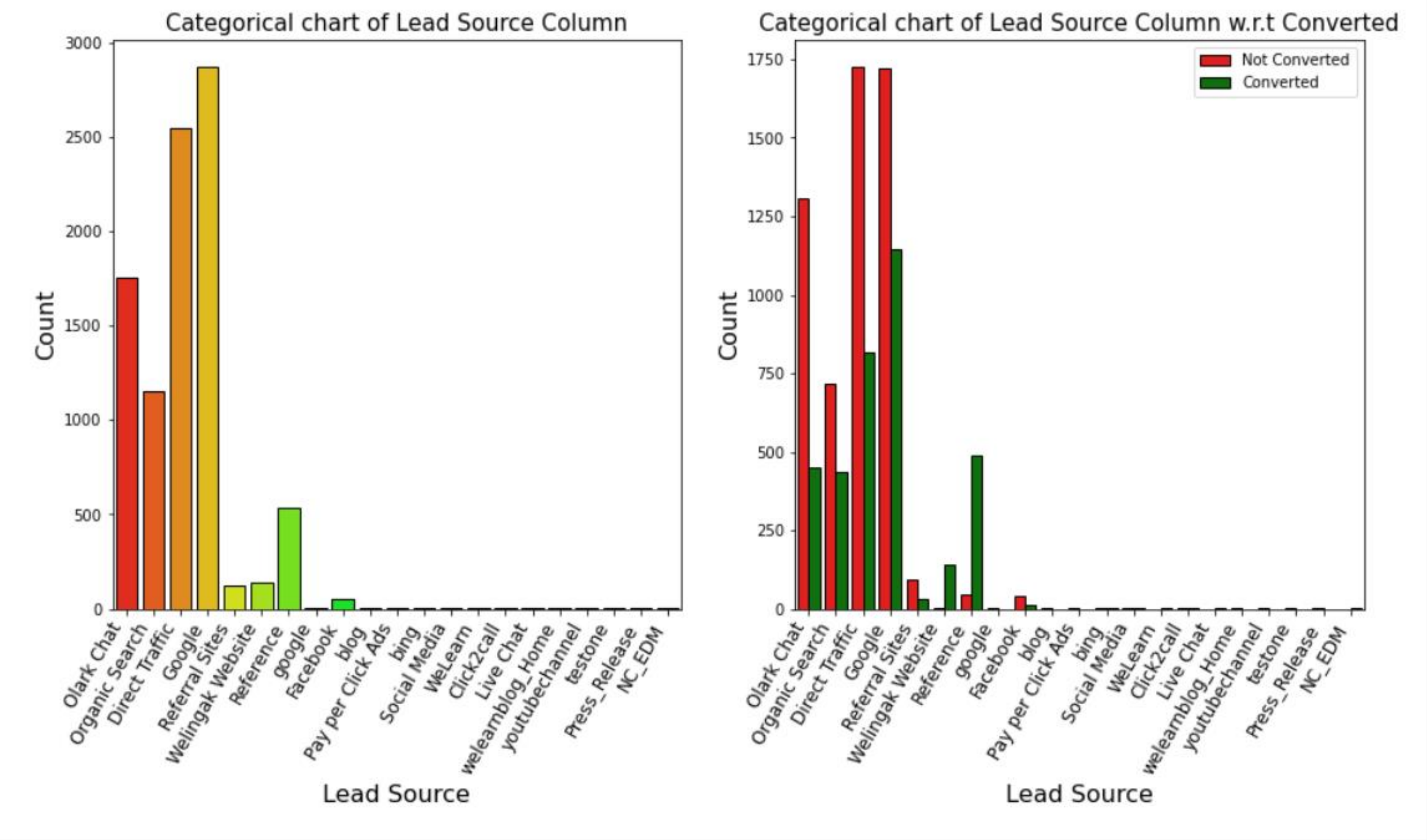
- Most of the customers are unemployed and Working Professionals are the most converted customers.

Bivariate Categorical Analysis (Specialization)



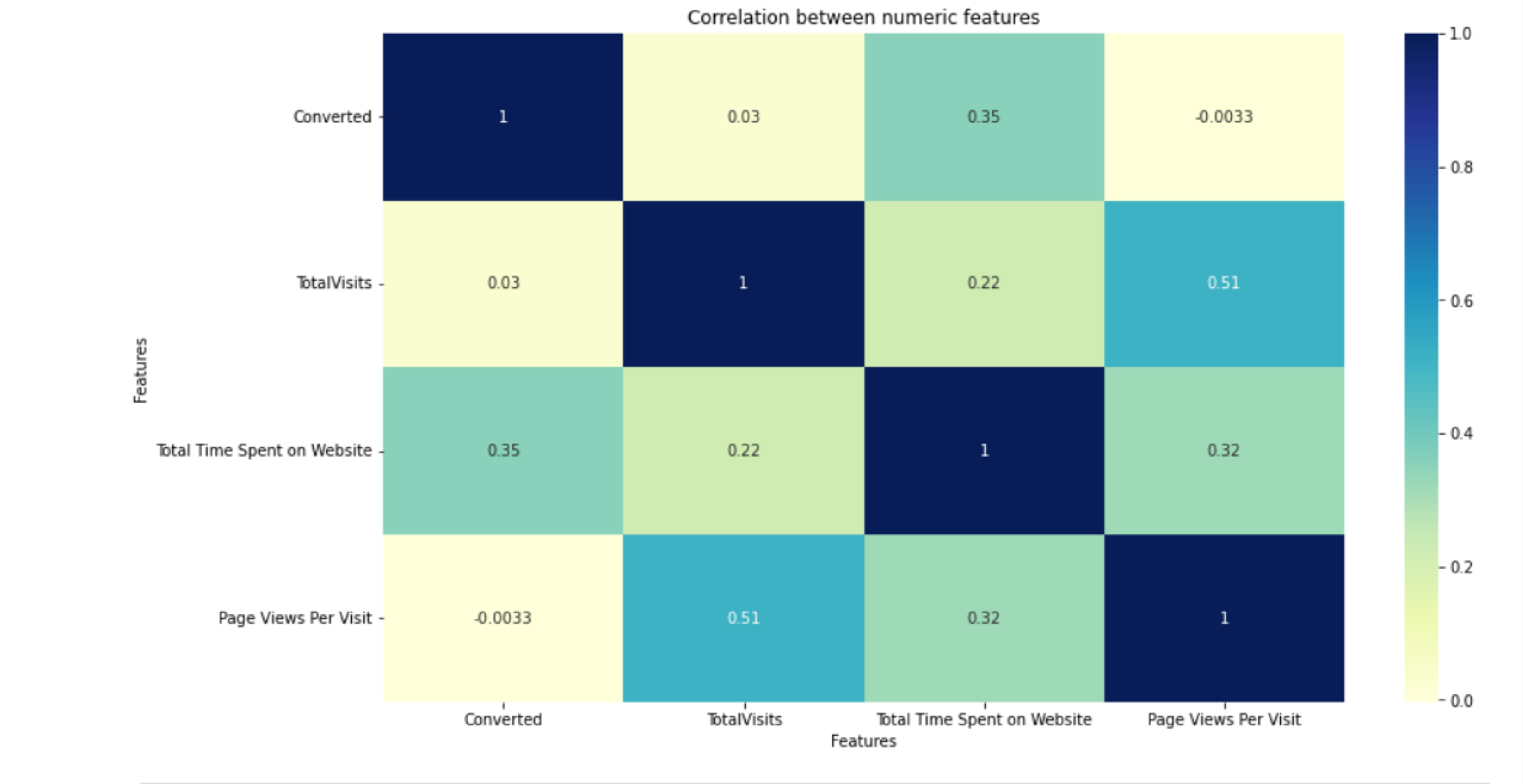
- Other (i.e. select) category is having most followed by Finance ,Marketing , HR departments respectively where mostly customers are working.

Bivariate Categorical Analysis (Lead Source)



- Top Lead sources are from sites like Google, Organic Search, Direct Traffic and Referrals.

Correlation Check



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

DATA PREPARATION

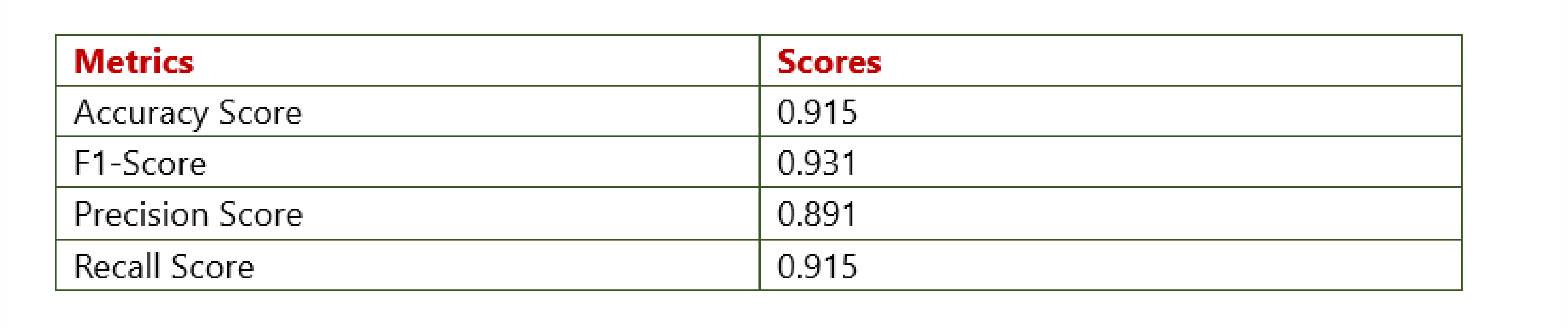
- Binary level categorical columns were already mapped to 1 / 0 in previous steps
- Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source , Do not email, Last Activity, Specialization, Current occupation, Tags ,City , A free copy of Mastering the Interview, Last Notable activity.
- Splitting Train & Test Sets - 70:30 % ratio was chosen form the split
- Feature scaling - Standardization method was used to scale the features
- Checking the correlations- Predictor variables which were highly correlated with each other were dropped.

MODEL BUILDING

Feature Selection

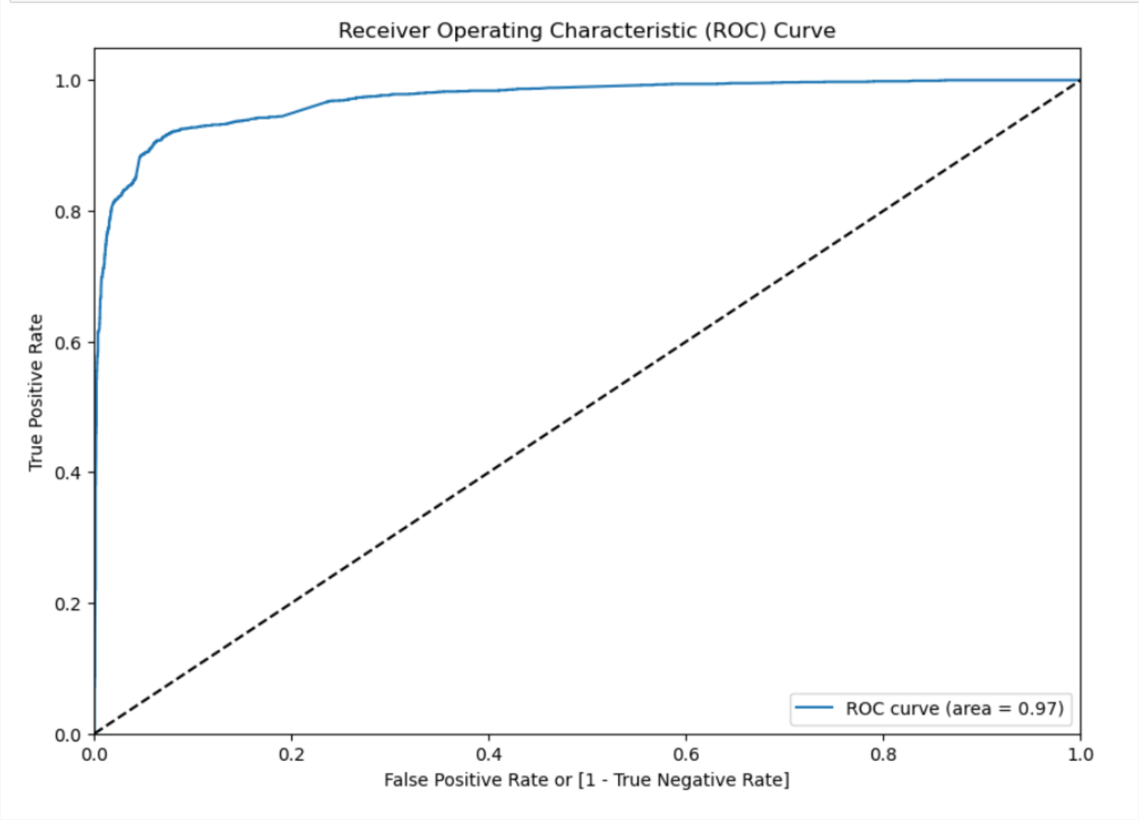
- The dataset has many dimensions and features.
- This can reduce model performance and increase computation time.
- Recursive Feature Elimination (RFE) is important to select only important columns.
- Manual Feature Reduction was used by dropping variables with p-value greater than 0.05.
- Model 3 looks stable after 3 iterations with significant p-values within the threshold (p-values < 0.05)
- There is no sign of multicollinearity with VIFs less than 5.
- Model 3 will be the final model used for Model Evaluation and predictions.

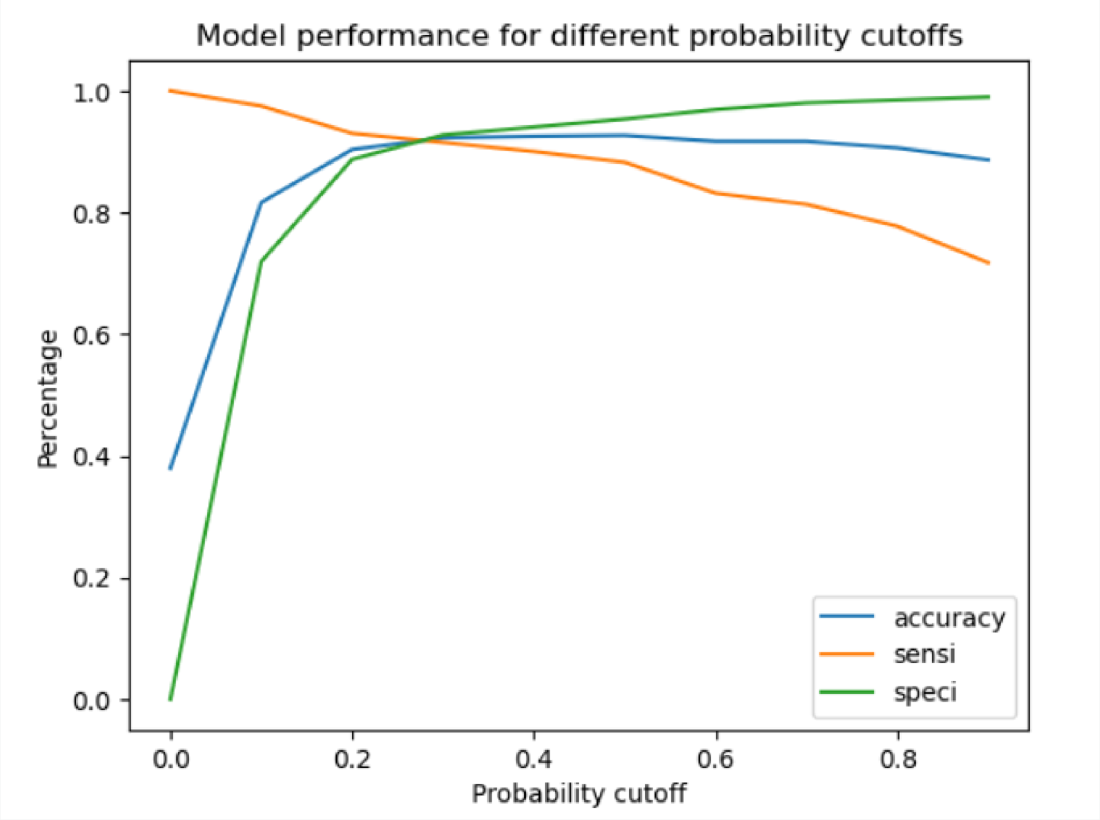
MODEL EVALUATION



Metrics	Scores
Accuracy Score	0.915
F1-Score	0.931
Precision Score	0.891
Recall Score	0.915

ROC CURVE





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

- Model appears to be performing well based on observations
- ROC curve has a value of 0.97, indicating high performance
- Training data accuracy achieved is 92.34%
- Sensitivity achieved is 91.69%
- Specificity achieved is 92.73%

Conclusion:

- Landing Page Submissions and Lead Add Form lead to more conversions.
- Conversions are higher for leads from Google, Organic Search, Direct Traffic, and Referrals.
- SMS and Email marketing leads have higher conversions.
- Finance, HR, Marketing, Operations, and Banking sector leads tend to convert more.
- "Better Career Prospects" option for career outcome leads to higher conversions.
- Leads spending more time on the website tend to convert more.
- Reducing website bounce rate can increase customer engagement time and conversions.
- Lead Add Form generates qualifying leads and should be used across key areas.
- Sales team should focus on working professionals for higher conversions.
- Leads with a Lead Score >0.35 tend to convert more and model accuracy score is 91%.

THANK YOU