X Education - Lead Scoring Case Study

Detection of hot leads to concentrate more of marketing efforts on them, improving conversion rates for X education

Team Members:,Ritik Mehta, Mitalee & Mehul Mehtani

Table of Contents

- Background of X education company
- Problem statement & objective of the study
- Suggested ideas for lead conversion
- Analysis approach
- Data cleaning
- EDA
- Data preparation
- Model building (RFE & manual fine tuning)
- Model evaluation
- Recommendations

Background of X education company

- X Education is an online education company that offers courses to industry professionals.
- The company attracts potential learners through various marketing channels, including Google and other websites.
- Upon landing on the website, visitors can browse courses, watch videos, or fill out forms with their contact information.
- Providing contact details classifies visitors as leads.
- The sales team engages with these leads through calls and emails to convert them into customers.
- The lead conversion rate at X Education is approximately 38.5%.

Problem statement & objective of the study

- **Problem Statement:**
  - X Education is facing a challenge with its lead conversion rate, which currently stands at a low 38%.
  - The company aims to enhance the lead conversion process by identifying the most potential leads, also referred to as "Hot Leads."
  - The sales team seeks to focus their efforts on communicating with these high-potential leads instead of reaching out to every lead.
- **Objective of the Study:**
  - The primary objective is to assist X Education in selecting the most promising leads, those with the highest likelihood of becoming paying customers.
  - The task involves building a model that assigns a lead score to each lead, indicating their conversion probability.
  - Leads with higher scores are expected to have a greater chance of conversion, while those with lower scores are less likely to convert.
  - The CEO has set a target lead conversion rate of around 80%, serving as a benchmark for success.

Suggested ideas for lead conversion

Leads Grouping

- Leads are categorised according to their propensity or likelihood to convert into paying customers.
- This categorisation process leads to the creation of a targeted group known as "hot leads."
- The hot leads are those with the highest potential to convert, receiving special attention and focus from the

Better Communication

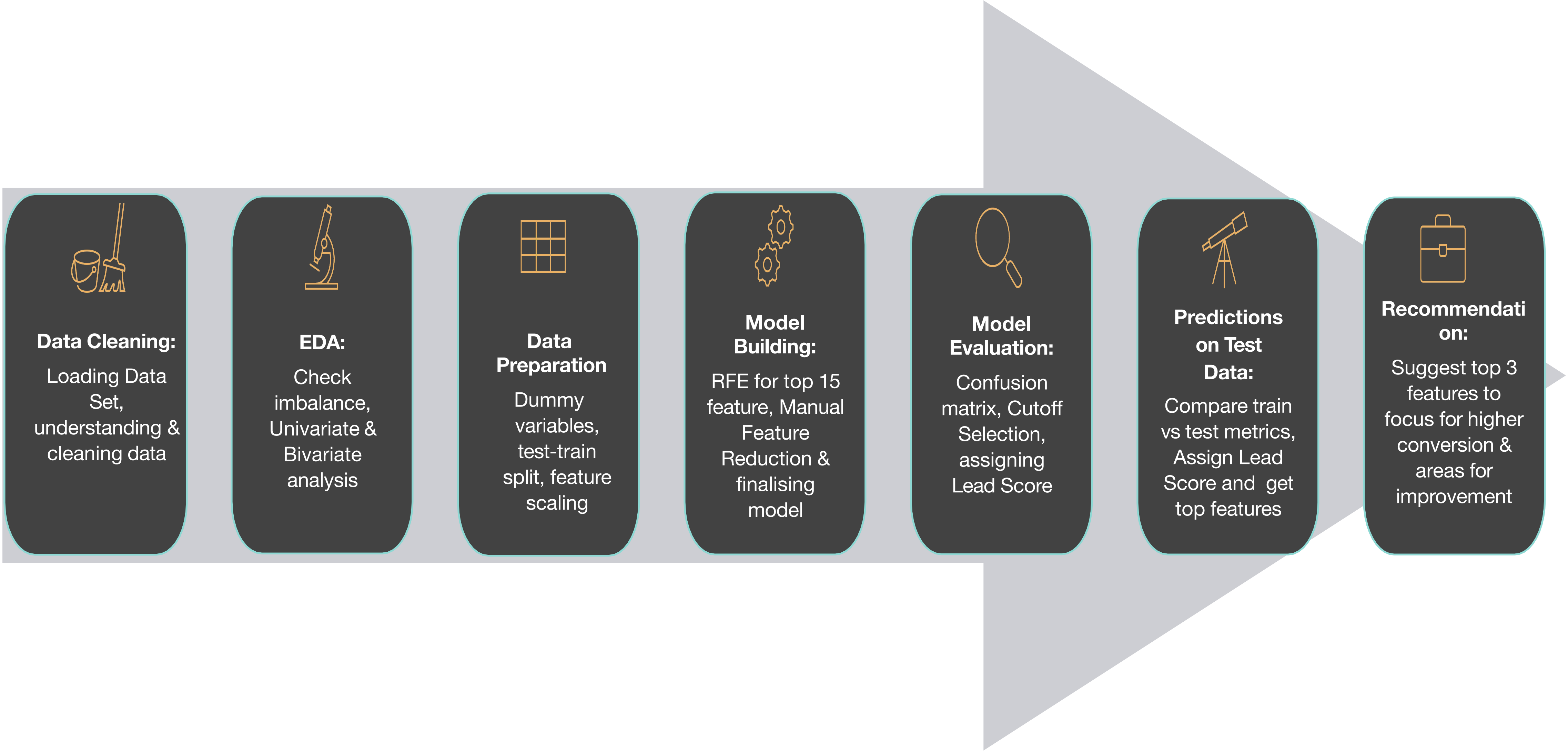
- By narrowing down the pool of leads, we can engage with a smaller but more promising group.
- This strategic approach enables us to have a more significant impact on lead conversion.
- Focusing our efforts on this smaller pool of leads increases the likelihood of successful communication and

Boost Conversion

- Concentrating on hot leads, those with a higher likelihood to convert, would result in a higher conversion rate.
- This targeted approach allows us to increase our chances of achieving the 80% conversion objective.
- By focusing on leads more likely to convert, we can improve our overall

We aim to achieve an 80% conversion rate, and to accomplish this, we need a high sensitivity in identifying hot leads.

Analysis approach



Data Cleaning:

Loading Data Set, understanding & cleaning data

EDA:

Check imbalance, Univariate & Bivariate analysis

Data Preparation

Dummy variables, test-train split, feature scaling

Model Building:

RFE for top 15 feature, Manual Feature Reduction & finalising model

Model Evaluation:

Confusion matrix, Cutoff Selection, assigning Lead Score

Predictions on Test Data:

Compare train vs test metrics, Assign Lead Score and get top features

Recommendation:

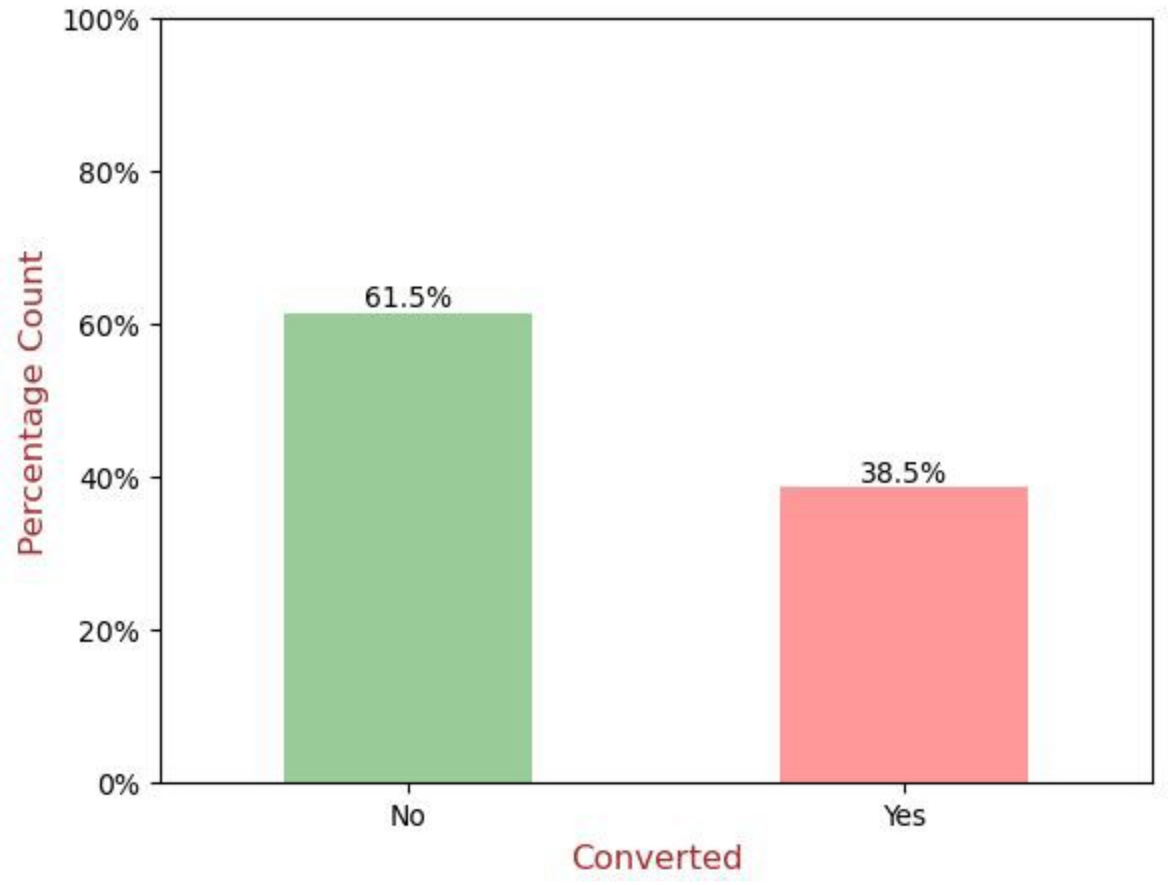
Suggest top 3 features to focus for higher conversion & areas for improvement

Data cleaning

- The **"Select"** level in some categorical variables indicates null values when customers didn't choose any option from the provided list.
- Columns with over 40% null values were removed from the dataset.
- Missing values in categorical columns were handled based on value counts and specific considerations.
- Columns that didn't contribute any valuable insights to the study objective, such as "tags" and "country," were dropped.
- Imputation techniques were used to handle missing values in certain categorical variables.
- Additional categories were created for certain variables to enhance the dataset.
- Columns like "Prospect ID" and "Lead Number," which had no relevance for modeling, were dropped, as well as those with only one category of response.
- Numerical data was imputed using the mode after checking the distribution.
- Skewed category columns were checked and dropped to avoid introducing bias in logistic regression models.
- Outliers in **"TotalVisits"** and **"Page Views Per Visit"** were treated and capped to mitigate their impact on the analysis.
- Invalid values were corrected, and data was standardised for specific columns like "lead source," ensuring consistency (e.g., "Google" and "google" were standardised).
- Low-frequency values were grouped together as "Others" to simplify the categorical variables.
- Binary categorical variables were mapped to numerical values for easier analysis.
- Various other data cleaning activities were performed to ensure data quality and accuracy.

EDA

Data is imbalanced while analysing target variable.

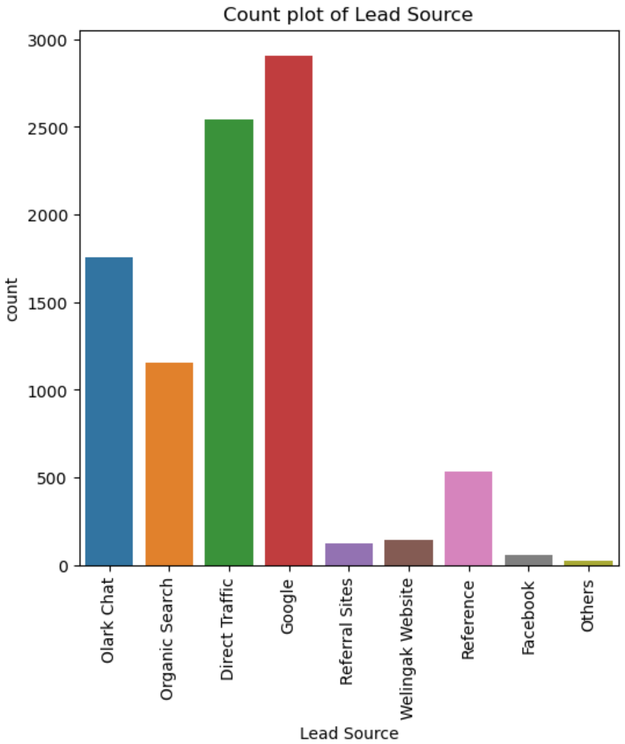


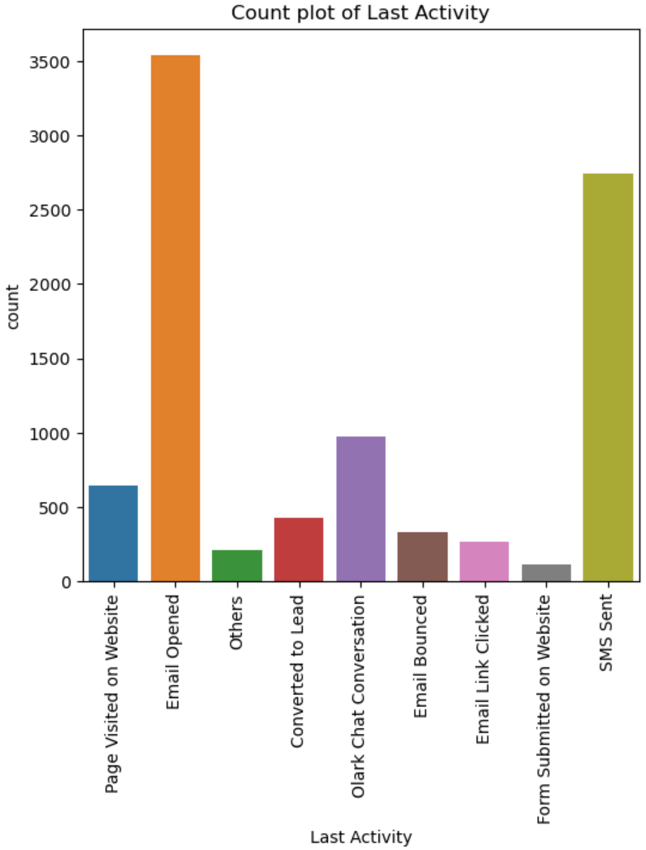
Insights

- The conversion rate stands at 38.5%, indicating that only a minority, or 38.5% of the people, have converted to leads.
- On the other hand, the majority, which is 61.5% of the people, did not convert to leads.

EDA

Univariate analysis – categorical variables





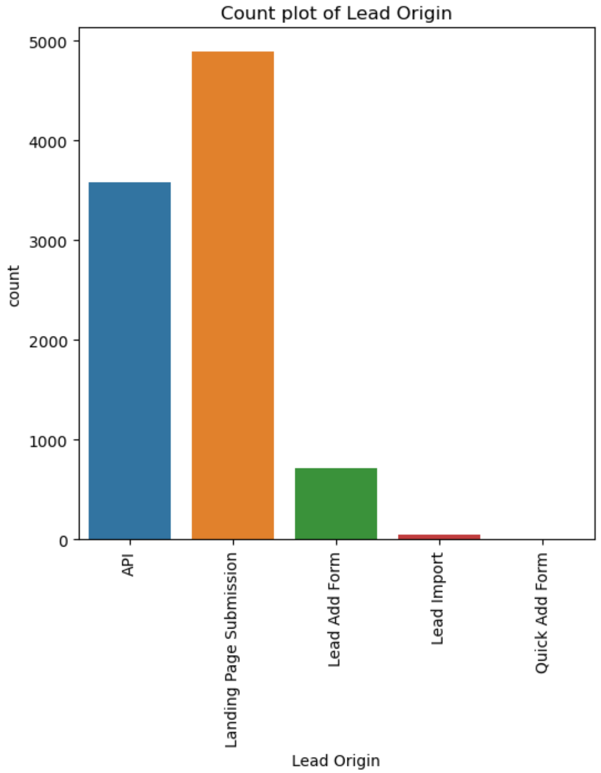
Insights

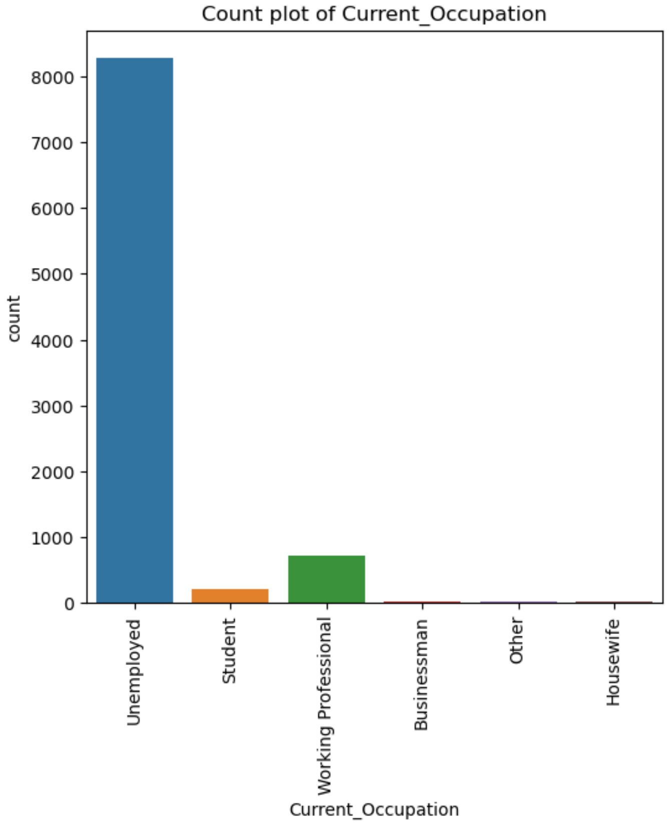
The most commonly used lead source used is Google followed by Direct Traffic

- **Most of the customers have actively engaged in the activities of SMS Sent and Email Opened.**

EDA

Univariate analysis – categorical variables



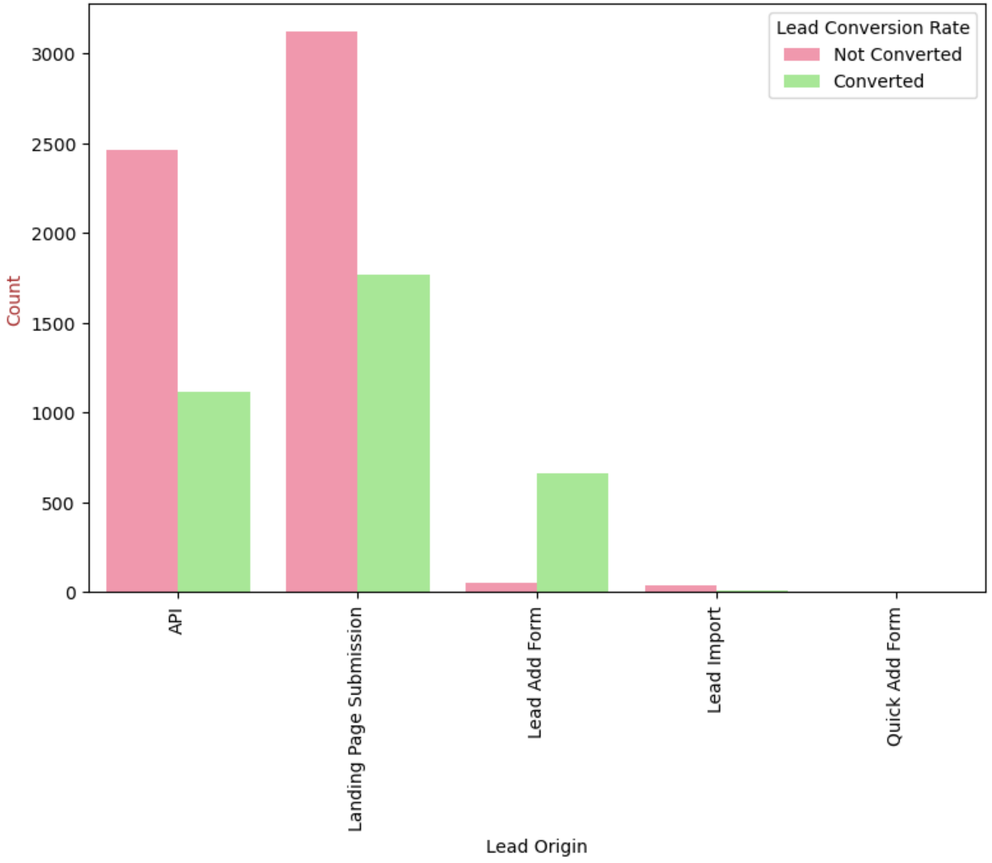


Insights

Landing Page Submission is the most popular lead origin amongst all followed by API.

It is observed Most people belong to the ‘Unemployed’ category

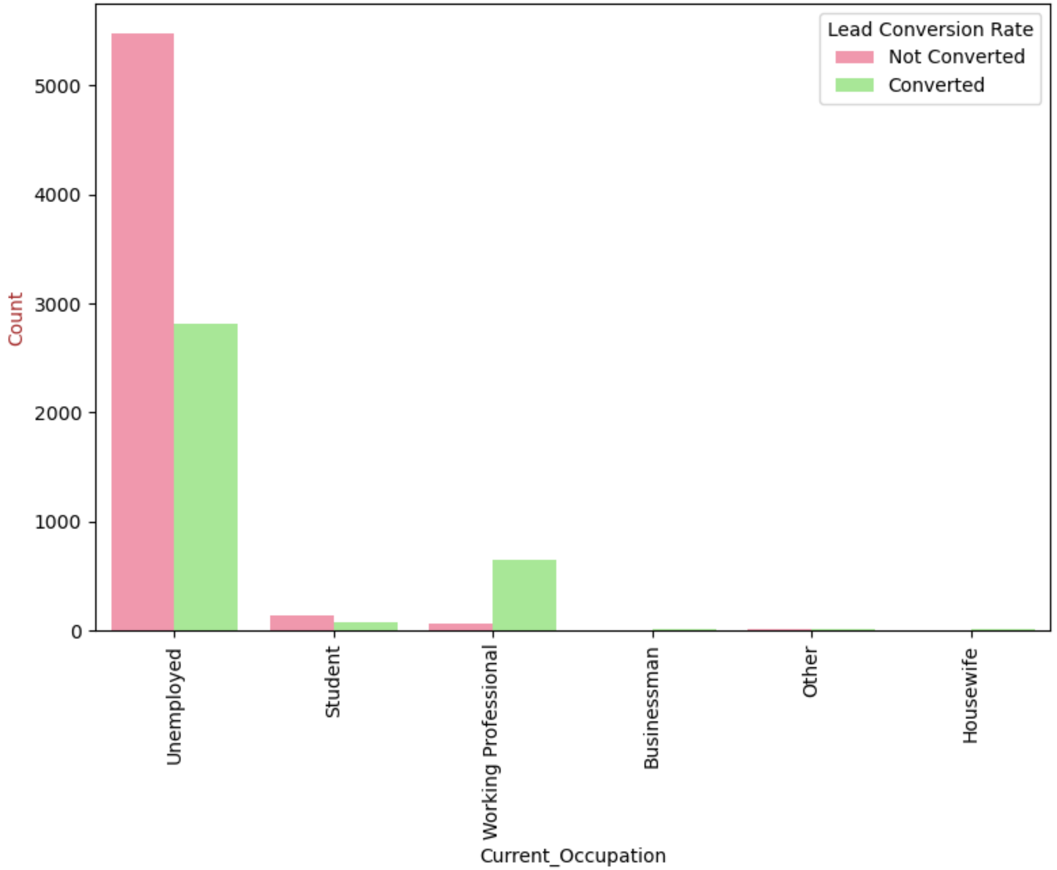
EDA – bivariate analysis for categorical variables



Insights

Lead Origin:Lead Conversion Rate is high in Landing Page Submission ,followed by API.Lead Import and Quick Add form do not add any value to our conversion

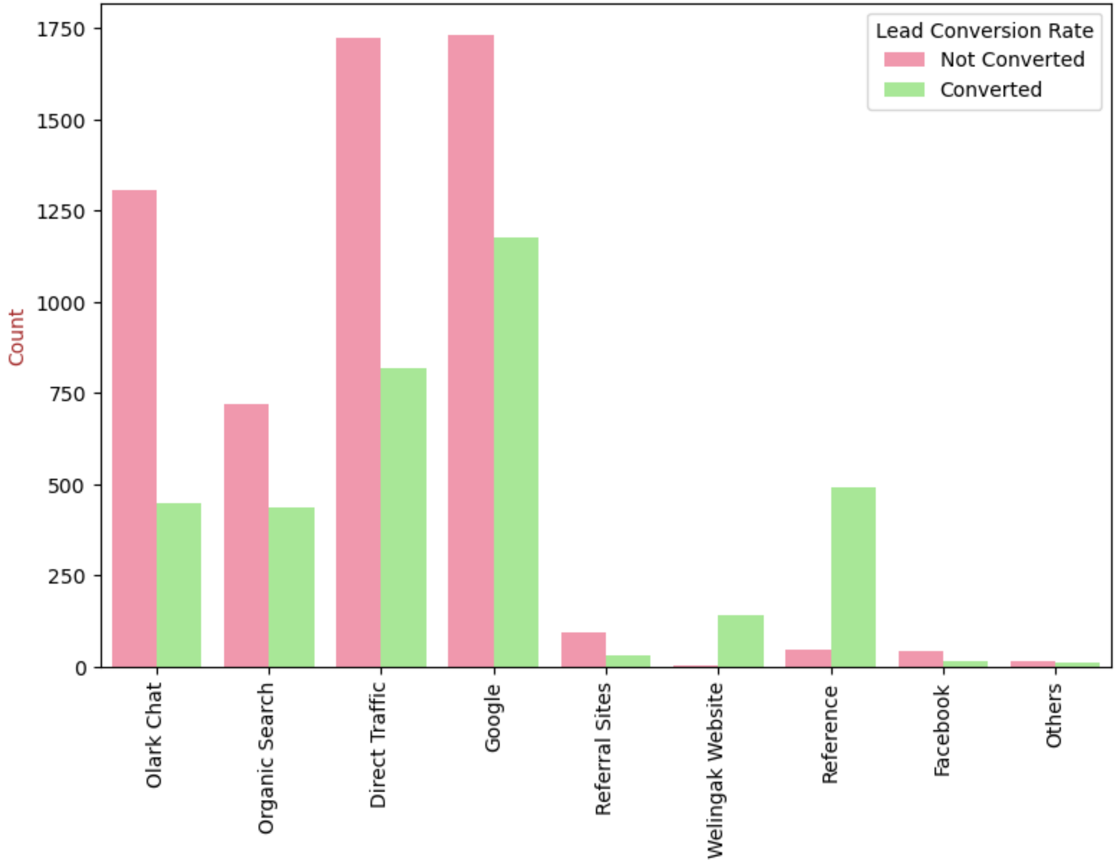
EDA – bivariate analysis for categorical variables



Insights

Most lead conversions are from users who are unemployed followed by working professionals .The rationale behind this is those who are 'unemployed' are looking to be employable and those who are 'working professionals' are looking for a chance to upskill themselves

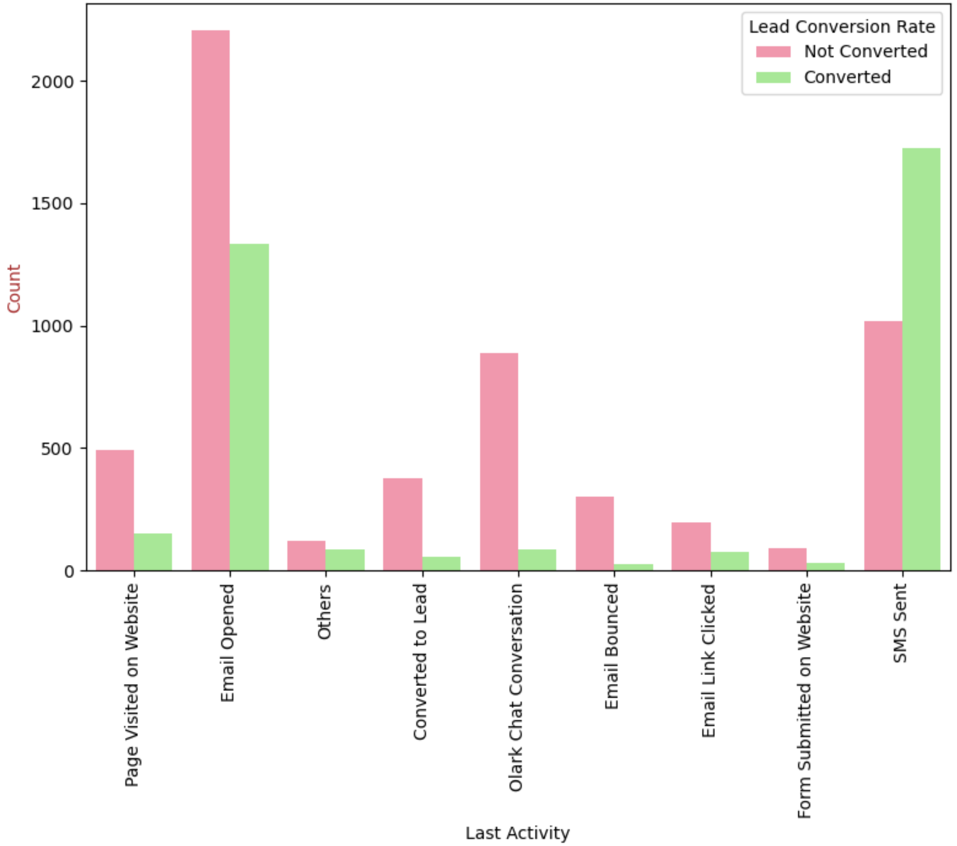
EDA – bivariate analysis for categorical variables



Insights

Lead Source: Most lead conversions are from sources namely:
1.Google
2.Direct Traffic
3.Olark Chat

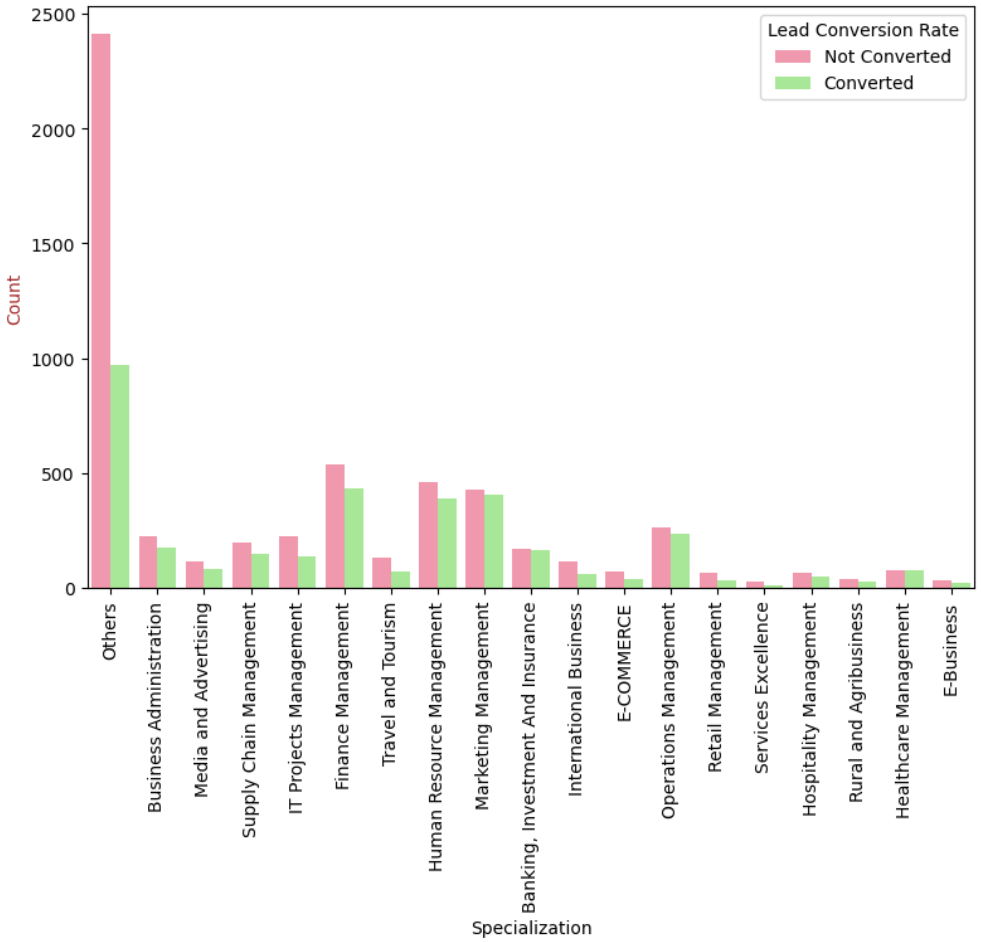
EDA – bivariate analysis for categorical variables



Insights

Those who have their Last Activity as SMS have a higher chance of lead conversion.
Email opened also has a positive response on lead conversion

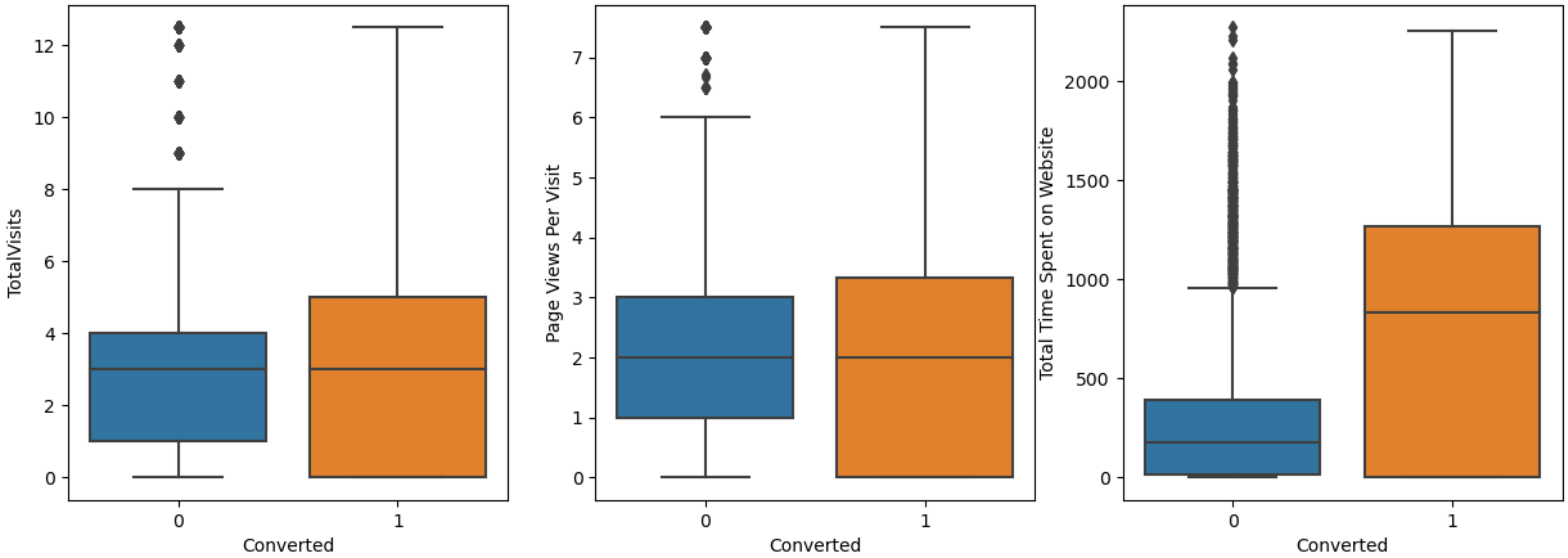
EDA – bivariate analysis for categorical variables



Insights

Specialisation: Marketing Management, HR Management,Finance Management show good contribution.

EDA – bivariate analysis for numerical variables



Insights

- The box-plot analysis indicates that past leads who spend more time on the website have a greater likelihood of successful conversion compared to those who spend less time.
- The data suggests that the amount of time spent by past leads on the website positively correlates with their chances of converting successfully.

Data preparation before model building

- Binary level categorical columns were previously encoded as 1 or 0 in the earlier steps.
- Dummy features were created for categorical variables, namely Lead Origin, Lead Source, Last Activity, Specialisation, and Current occupation, using one-hot encoding.
- The dataset was split into training and test sets using a 70:30 ratio for model evaluation.
- To ensure fair comparisons and model performance, feature scaling was applied using the standardisation method.
- Correlations among predictor variables were checked to identify any highly correlated features.

Model building

Feature selection

- Due to the dataset's high dimensionality and large number of features, it's essential to perform Recursive Feature Elimination (RFE) to select only the most important columns for modeling.
- The RFE process reduced the initial 49 columns to 23 columns, significantly reducing computational time and improving model performance.
- After RFE, a manual feature reduction process was employed, dropping variables with p-values greater than 0.05 to ensure statistical significance in the model.
- Among the different models tested, Model 4 demonstrated stability and met the criteria of having significant p-values below the threshold (p-values < 0.05) and no indication of multicollinearity, with Variance Inflation Factors (VIFs) below 5.
- Therefore, "lr7" is selected as the final model for Model Evaluation and will be used for making predictions in further analyses.

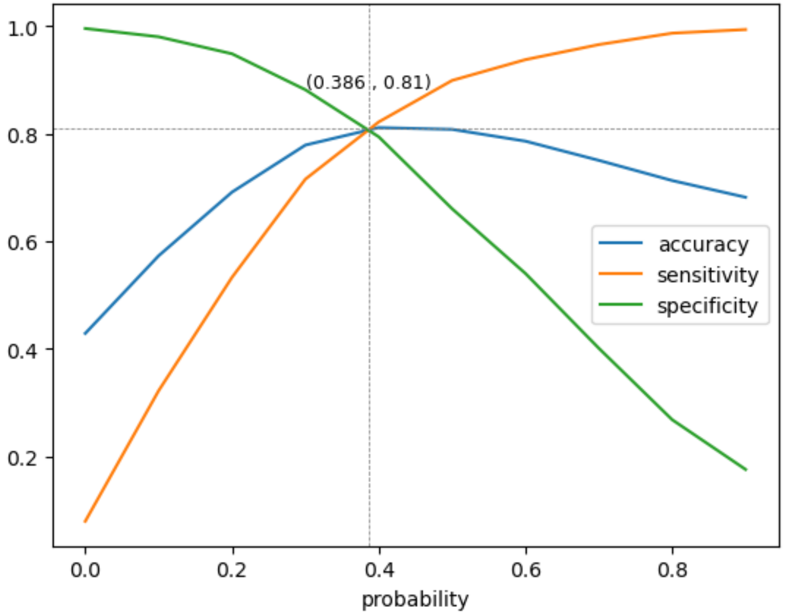
Model evaluation

Train data set

- **After analyzing the evaluation metrics from both plots, a cutoff value of 0.386 was determined as the appropriate threshold.**
- **This decision was based on a careful examination of the results from the evaluation metrics, which led to the selection of 0.386 as the optimal cutoff point.**

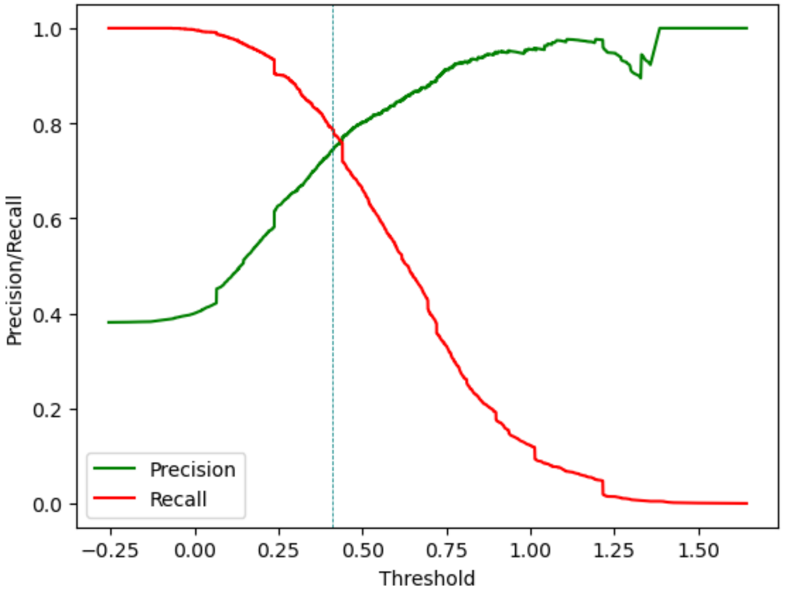
Confusion matrix & evaluation metrics with 0.386 as cutoff

Accuracy: 0.808
Sensitivity: 0.795
Specificity: 0.81



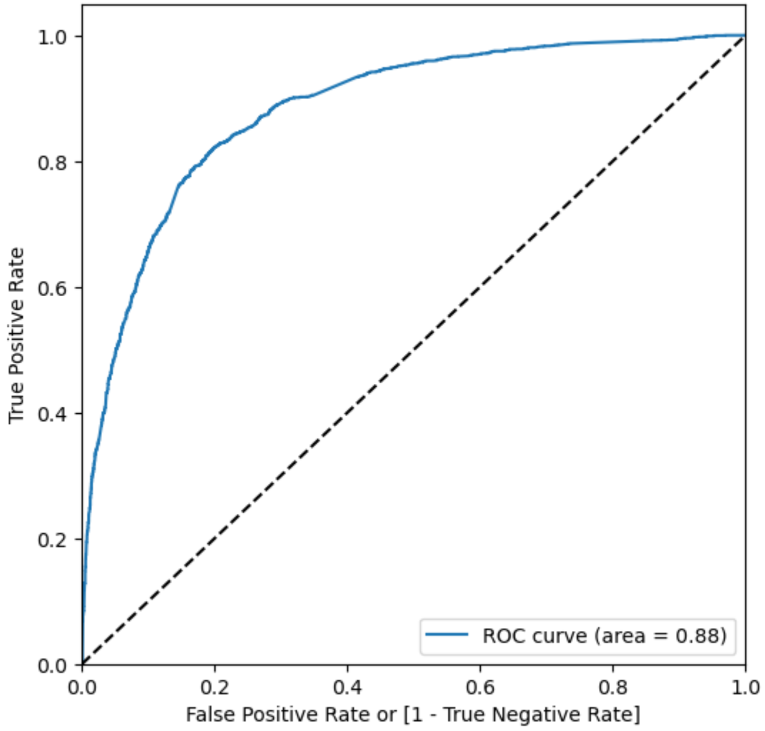
Confusion matrix & evaluation metrics with 0.40 as cutoff

Accuracy: 0.81
Recall: 0.795
Precision: 0.73

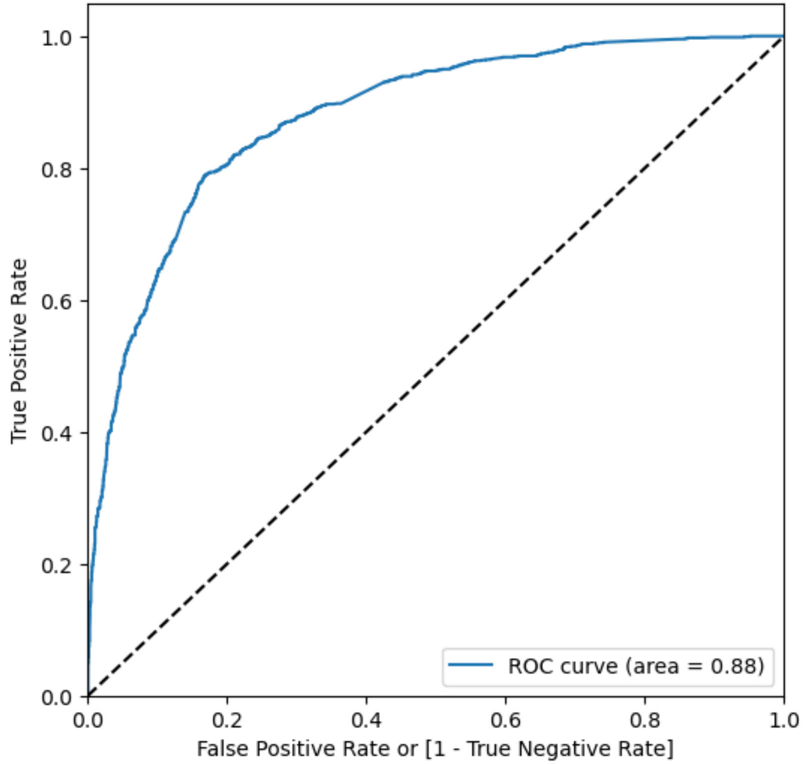


Model evaluation

ROC curve – train data set



ROC curve – test data set



Insights

- The model's AUC is 0.88, indicating a strong predictive ability. The ROC curve is near the top-left corner, reflecting a high true positive rate and low false positive rate at all thresholds.

- With an AUC of 0.88, the model performs well in predicting. The curve near the top-left corner signifies a model with high true positive rate and low false positive rate across all thresholds.

Model evaluation

Confusion matrix & metrics

FOR TRAIN DATA SET		vs	FOR TEST DATA SET
Accuracy ----	81%	:	80.8%
Sensitivity----	80.9%	:	79.5%
Specificity---	81%	:	81%

Insights

- **Using a cutoff value of 0.386, the model obtained a sensitivity of 80.9% on the train set and 79.5% on the test set.**
- **Sensitivity reflects how accurately the model identifies converting leads out of all potential leads.**
- **The CEO's target sensitivity of approximately 80% was met by the model.**
- **Furthermore, the model achieved an accuracy of 80.8%, aligning with the study's objectives.**

Recommendation based on final model

To increase lead conversion:

- Prioritise features with positive coefficients for targeted marketing strategies.
- Focus on top-performing lead sources to attract high-quality leads.
- Optimise communication channels based on lead engagement impact.
- Tailor messaging to engage working professionals effectively.
- Allocate more budget to Welingak Website advertising.
- Offer incentives for providing successful references.

To identify areas for improvement:

- Analyse specialisation offerings with negative coefficients.
- Review and improve the landing page submission process.