Lead Score Case Study

Submitted by :

Anjali Singh

RS Catherin Jenefer Sen

Aloke KR Das

Olive Praisy

LEAD SCORE CASE STUDY FOR X EDUCATION

CONTENTS

PROBLEM STATEMENT

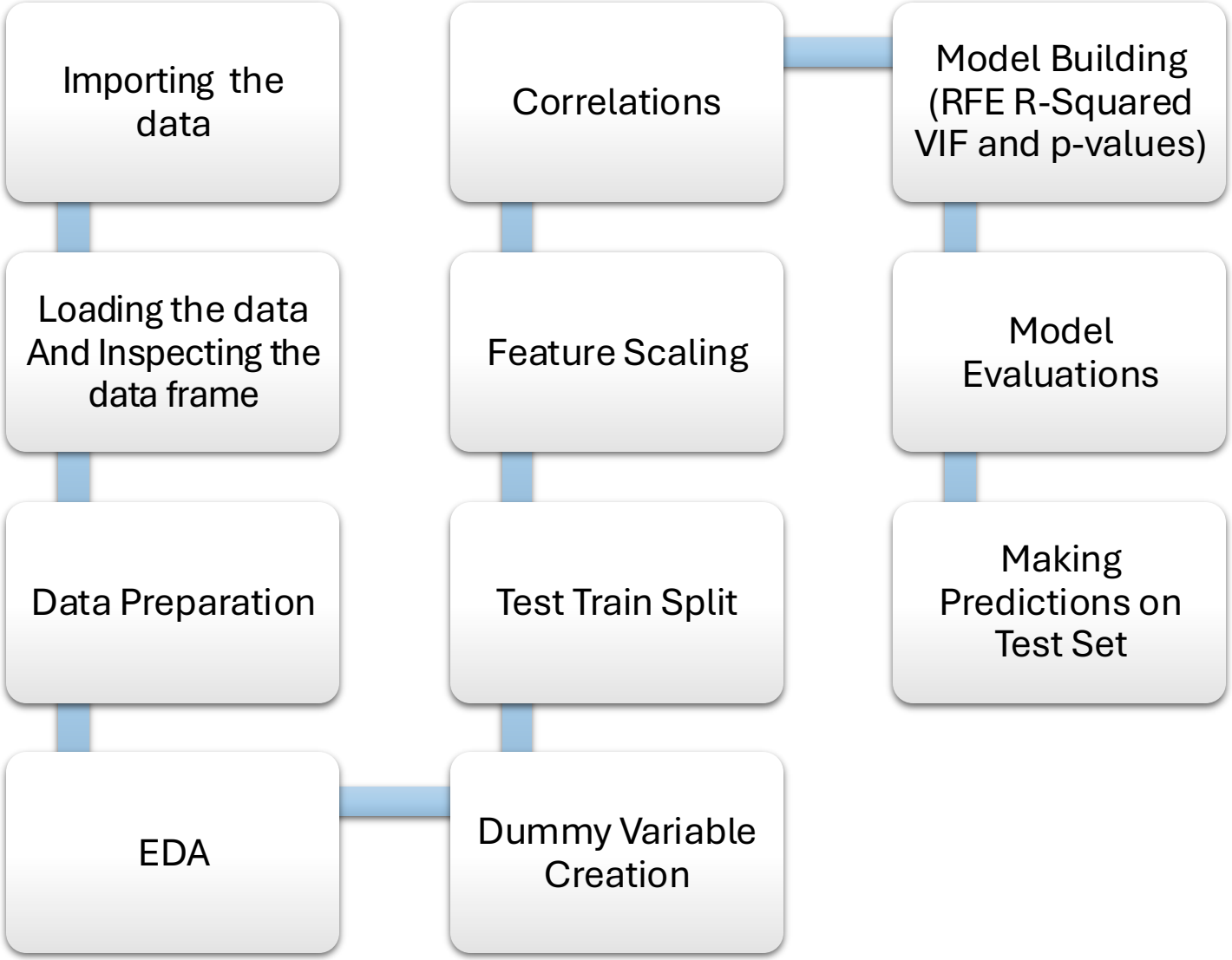
X Education, an online course provider for industry professionals, attracts numerous visitors to its website daily through marketing efforts on various platforms, including search engines like Google. These visitors may browse courses, fill out inquiry forms, or watch informational videos. When visitors provide their email addresses or phone numbers via these forms, they are classified as leads. Additionally, the company receives leads through past referral. Once leads are acquired, the sales team engages with them through calls and emails. However, despite acquiring a significant number of leads, X Education's lead conversion rate is relatively low, averaging around 30%. For instance, out of 100 leads acquired in a day, only about 30 are converted into customers. To improve efficiency, X Education aims to identify the most promising leads, referred to as 'Hot Leads.' By focusing on these high-potential leads, the company hopes to increase its lead conversion rate, allowing the sales team to prioritize their efforts on the most likely prospects instead of reaching out to all leads indiscriminately.

BUSINESS OBJECTIVE

- Lead X wants us to create a model that assigns a lead score between 0 and 100 to each potential customer. This will help them identify the most promising leads and improve their conversion rate.
- The CEO aims for an 80% lead conversion rate.

Additionally, they want the model to handle future needs, such as managing peak times, fully utilizing staff, and planning next steps after reaching the target

PROBLEM APPROACH



EDA - DATA CLEANING

Handling Duplicates

Upon checking for duplicates, we found that the dataset is clean and does not contain any duplicate entries.

Handling Missing Values

Deleting Columns: Removed columns with a high percentage of missing values

Median Imputation: Used median for filling missing values in skewed distributions.

Other Techniques: Applied additional methods as needed.

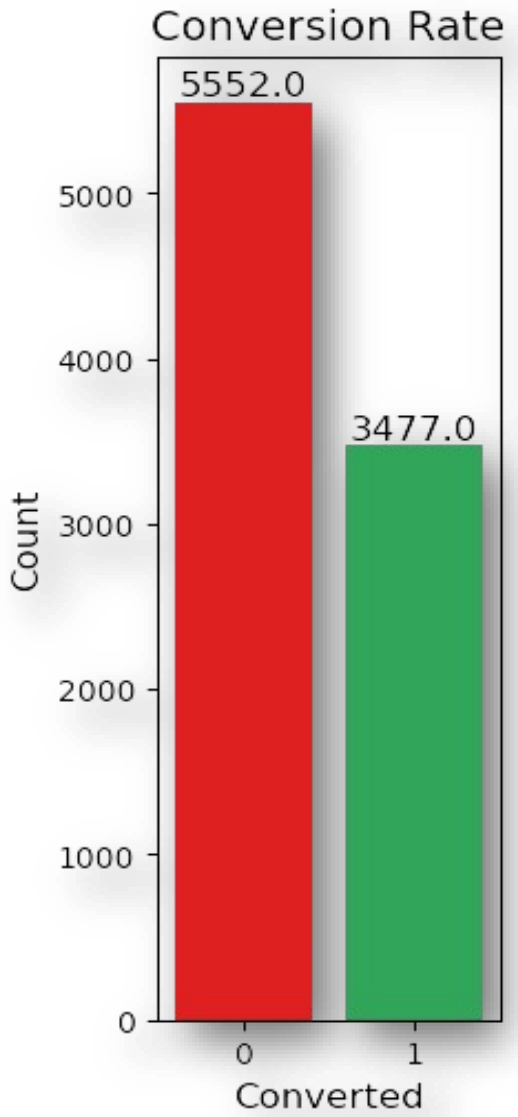
Handling Outliers:

Box Plots: Visualized box plots to detect the presence of outliers

Quantile Analysis: Used quantiles (min, 25%, 50%, 75%, 90%, 99%, max) to identify and manage outliers effectively.

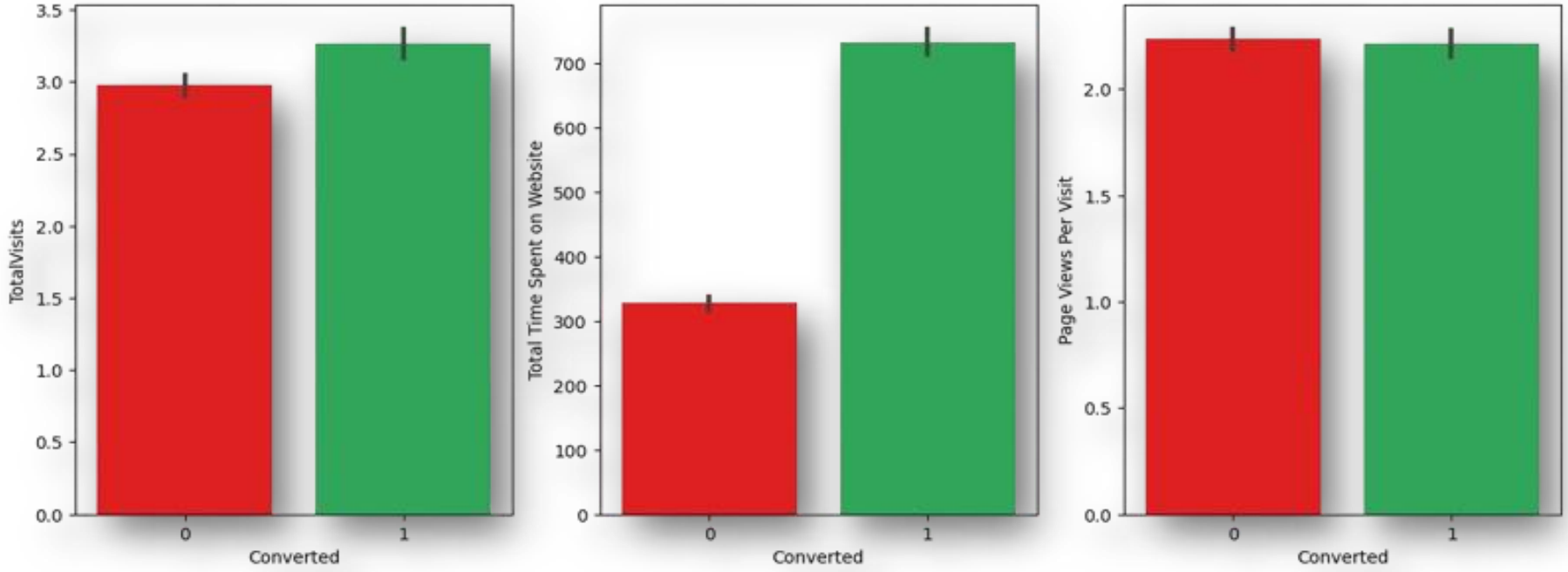
✓ Successfully addressed all missing values efficiently, achieving an impressive data retention rate of **97.72%** post-cleaning

DATA ANALYSIS INSIGHTS

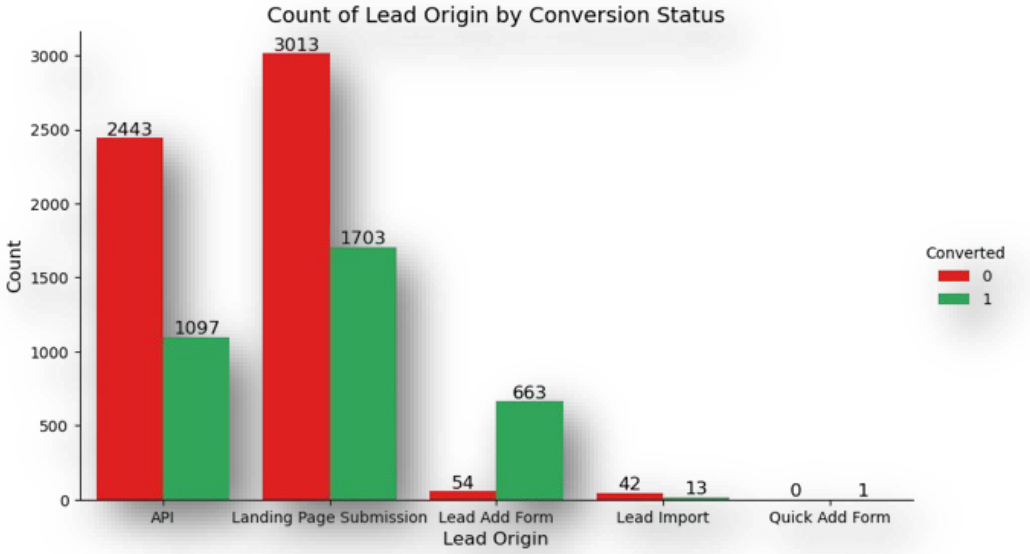


Prior to model development, we observed a baseline conversion rate of approximately **39%.**

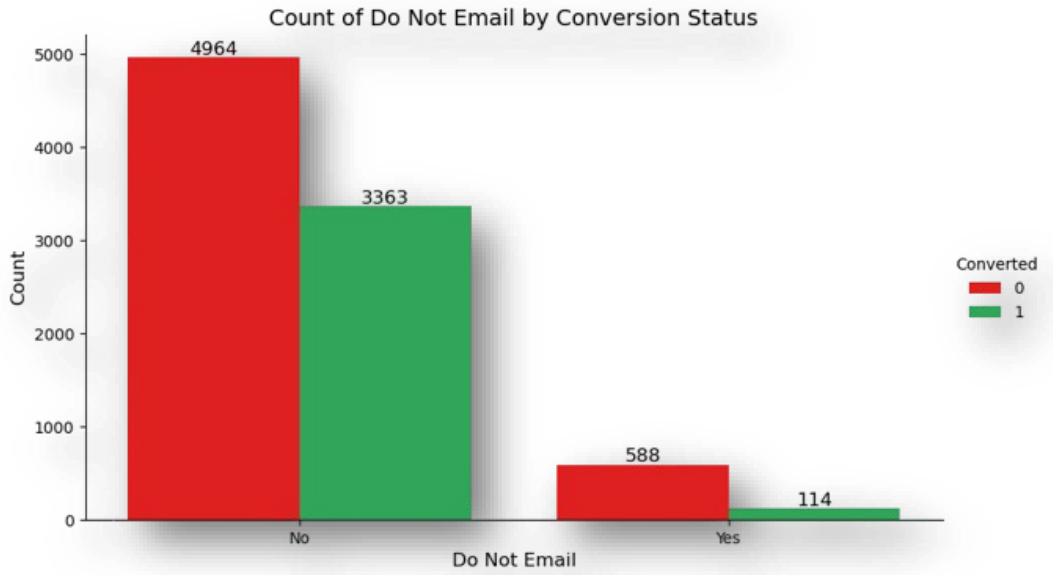
The features Total Visits, Total Time Spent on Website, and Page Views Per Visit demonstrated higher conversion rate

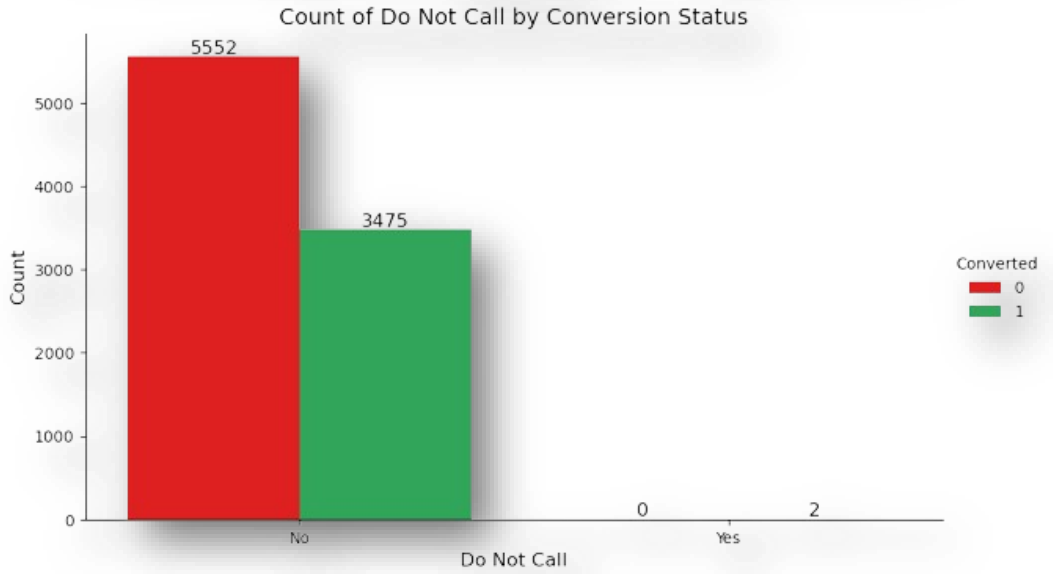


The highest conversion rate in the Lead Origin category was achieved through Landing Page Submissions

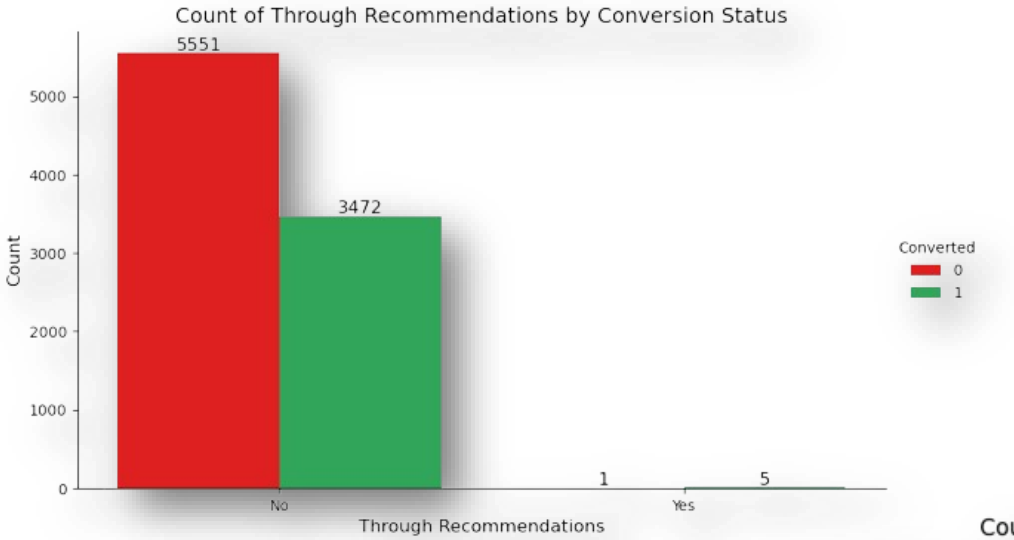


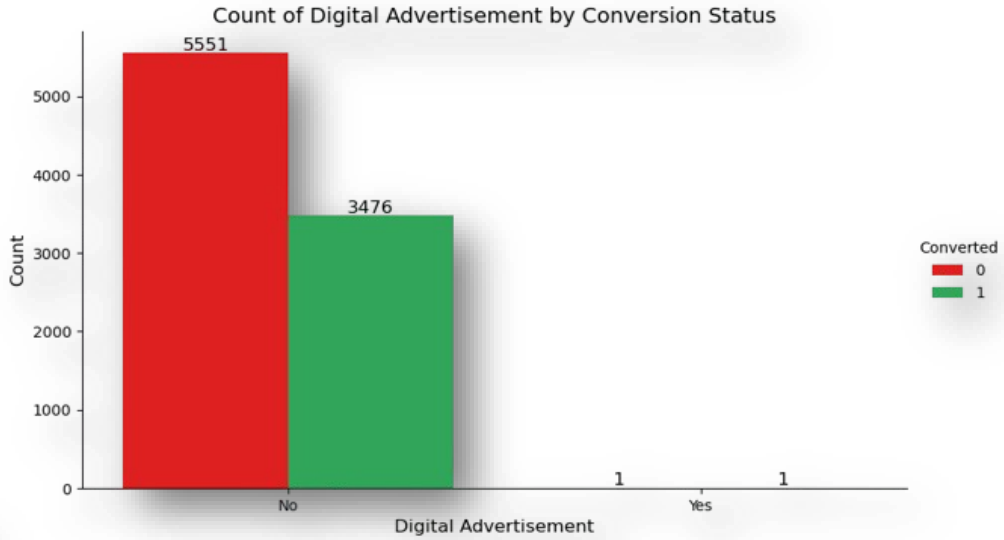
Majority of conversions were driven by calls and emails

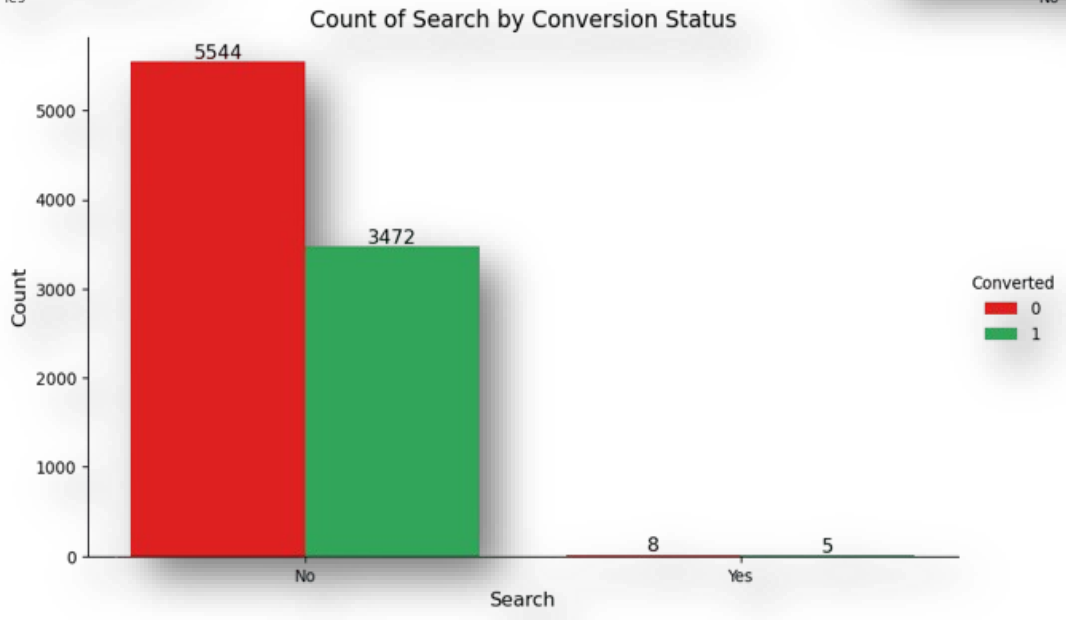


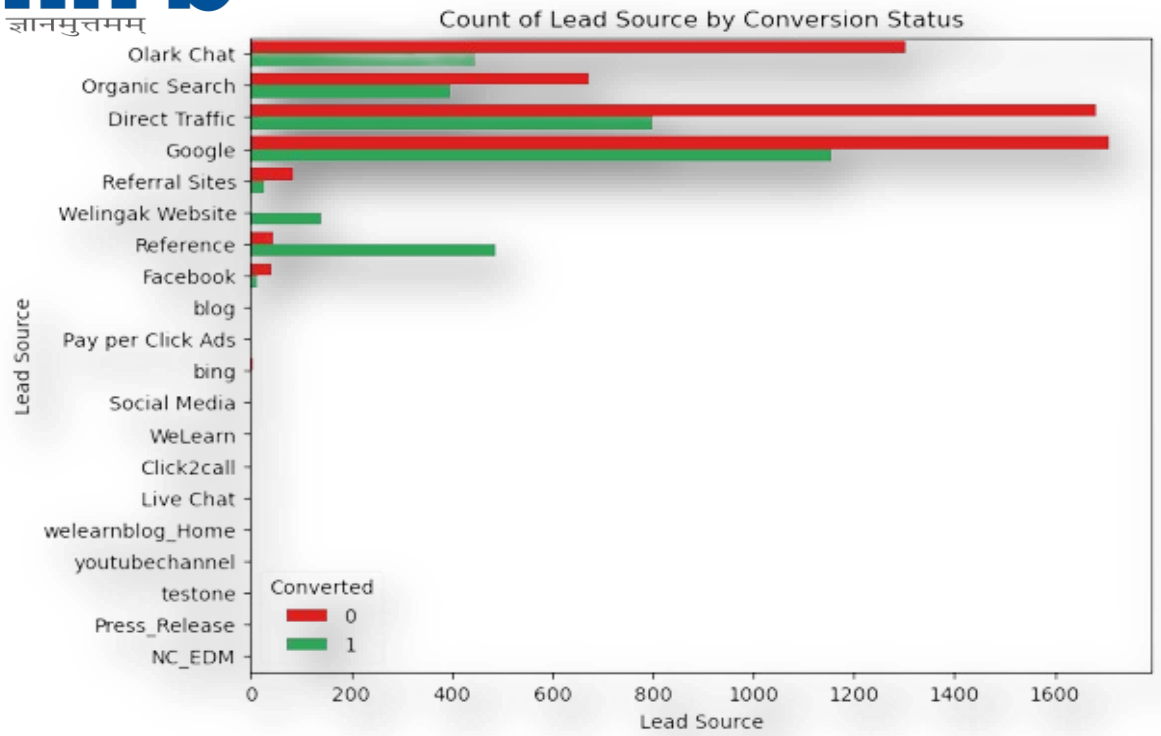


Conversions were minimally impacted by search, digital advertisements, and recommendations.

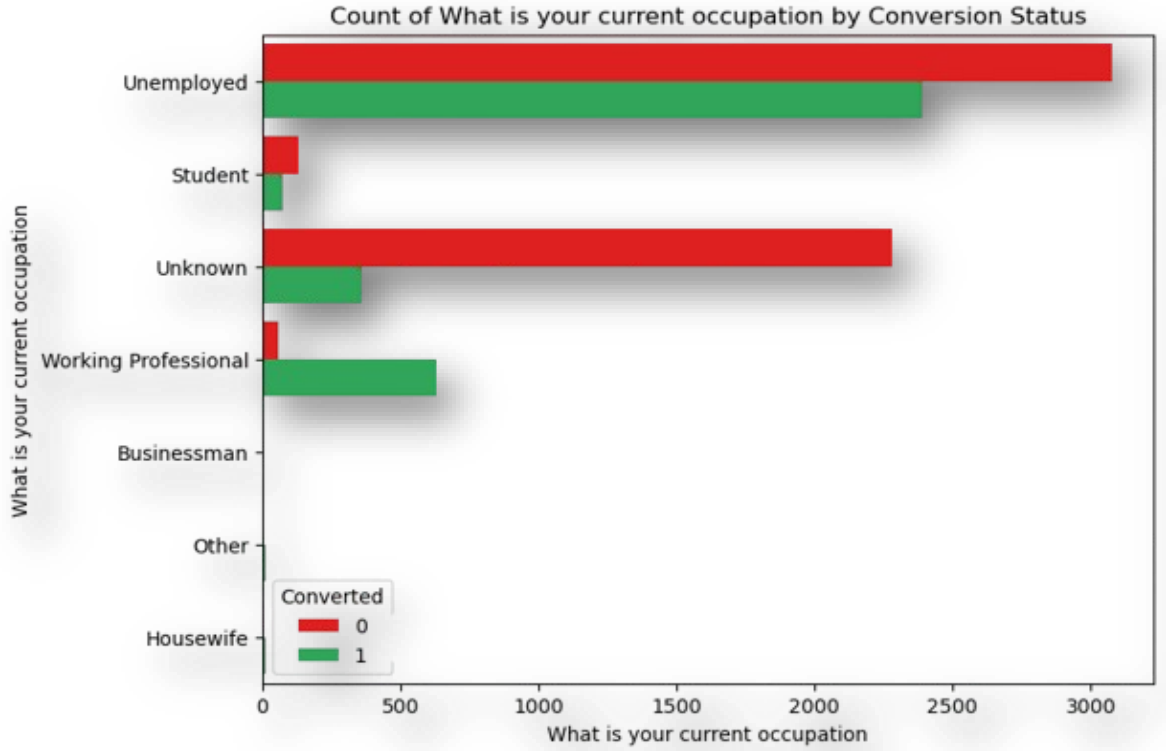








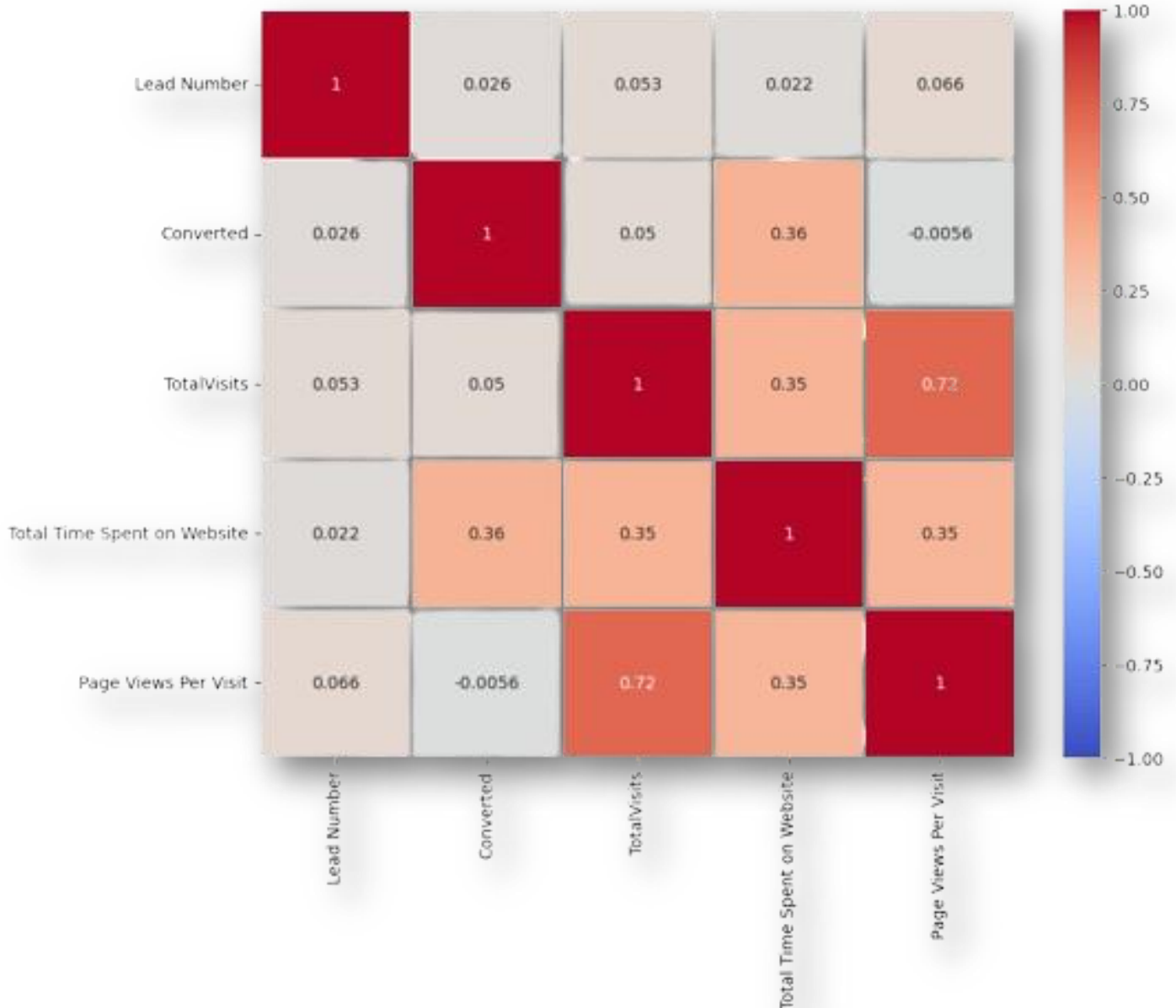
The major conversions in the lead source are primarily driven by Google



More conversions occurred among individuals who are unemployed.

CORRELATION

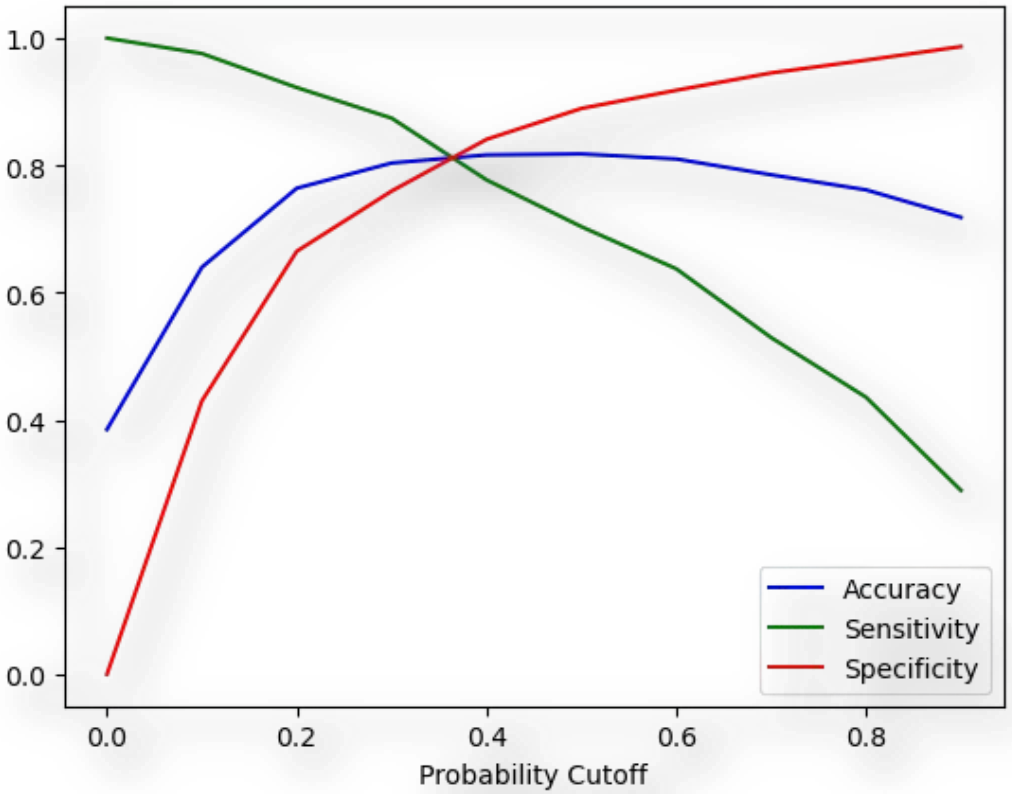
- Converted lead has highest correlation with time spent on website
- Total visits seems be highly correlated to page views per visit followed by total time spent on website
- Time spent on website alao seems to have correlation with page views per visit



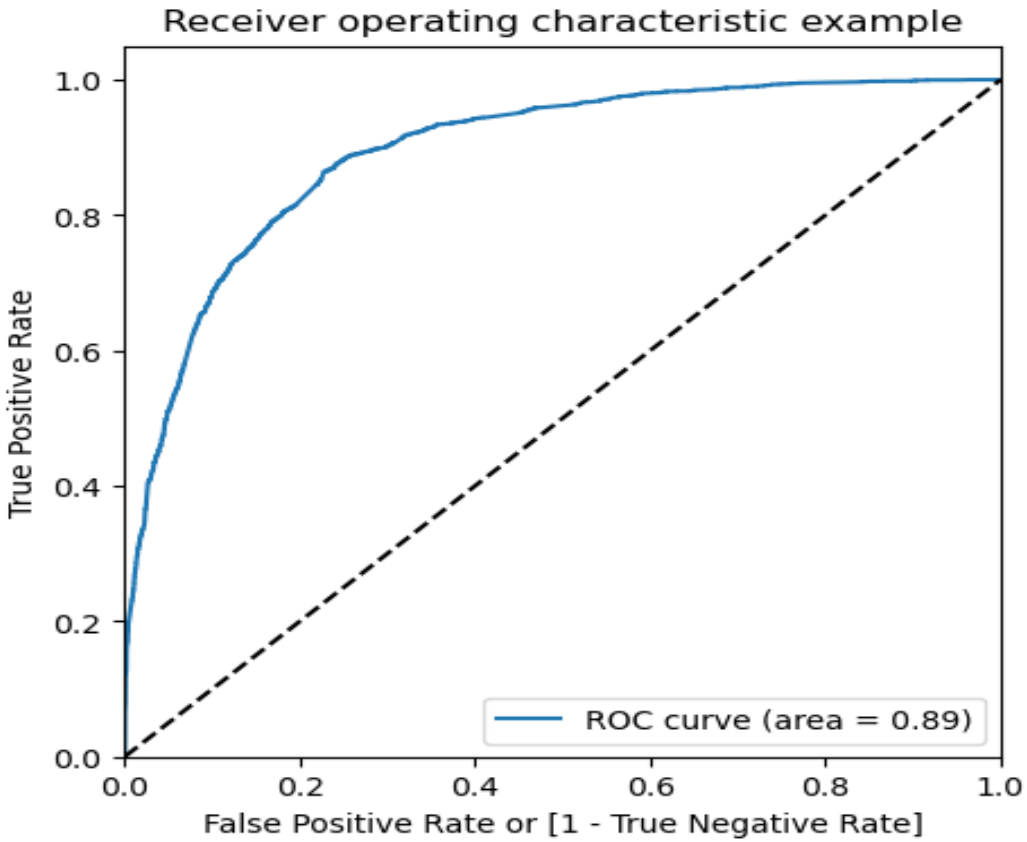
Features Impacting Conversion Rate

- Total Visits
- Total Time Spent on Website
- Do Not Email - Yes
- Lead Origin - Lead Add Form
- Lead Source-Olark Chat
- Lead Source - Welingak Website
- Last Activity - Email Opened
- Last Activity - Had a Phone Conversation
- Last Activity - SMS Sent
- What is your current occupation - Unknown
- What is your current occupation - Working Professional
- Last Notable Activity - Modified
- Last Notable Activity - Olark Chat Conversation
- Last Notable Activity - Unreachable

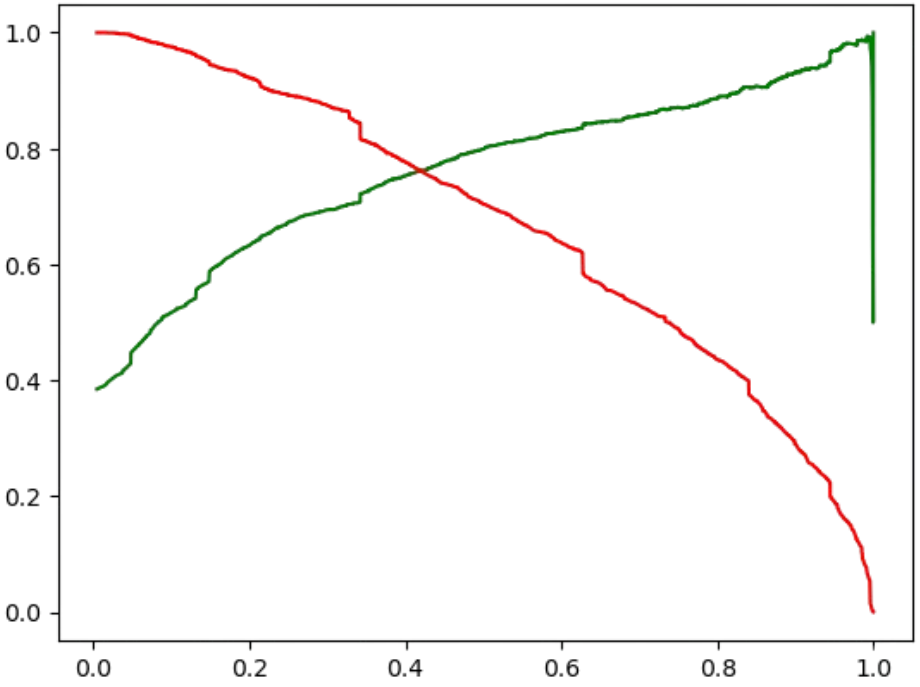
ROC Curve for Model Performance Evaluation



With an AUC of 0.86, the model is performing well



MODEL EVALUATION : PRECISION AND RECALL ON TRAINING DATA



The graph illustrates an optimal cutoff point of 0.353, derived by balancing precision and recall.

Confusion Matrix:

3146	742
460	1972

- Accuracy ; 80.98%
- Sensitivity : 81.09%
- Specificity : 80.92%
- Precision : 72.92%
- Recall : 70.35%
- Conversion Rate : 81.09%

MODEL EVALUATION ON TEST DATA SET

Confusion Matrix:

1365	299
203	842

- Accuracy ; 81.47%
- Sensitivity : 80.57%
- Specificity : 82.03%
- Precision : 73.8%
- Recall : 80.57%
- Conversion Rate : 80.57%

RECOMMENDATIONS

- Google is identified as a major source of conversions. Increase investment in Google Ads and enhance search engine optimization (SEO) efforts to drive more leads from this channel
- The highest conversions are observed from landing page submissions. Prioritize optimizing and promoting landing pages to capture and convert leads more effectively.
- People who prefer phone conversations and SMS tend to convert better. Tailor communication strategies to include more phone and SMS follow-ups for higher engagement and conversion rates.
- Given the positive impact of phone calls and SMS, implement strategies to increase these types of interactions. Train sales and support teams to follow up effectively through these channels.
- Higher engagement metrics like total visits and time spent on the website correlate with better conversion rates. Enhance website content and user experience to increase engagement and conversions.
- Utilize the insights from feature impacts and conversion rates to inform marketing and sales strategies. Regularly update and analyze data to adapt strategies and maintain high conversion rates.

CONCLUSION

We evaluated the model using both Sensitivity-Specificity and Precision-Recall metrics, selecting the optimal cutoff based on Sensitivity and Specificity for final predictions. The test set shows Accuracy, Sensitivity, and Specificity values of approximately 81.5%, 80.5%, and 82%, Precision of 74%, and Recall of 80.54%, which are close to those from the training set. The lead scoring on the training data indicates a conversion rate of around 80.57% for the final model. Overall, this model performs well and meets the expected criteria.