LEAD SCORE CASE STUDY

By Pooja Sharma

Table of Contents

- Problem Statement
- Analysis Approach
- Data Cleaning
- EDA
- Data Preparation
- Model Building
- Model Evaluation
- Recommendations

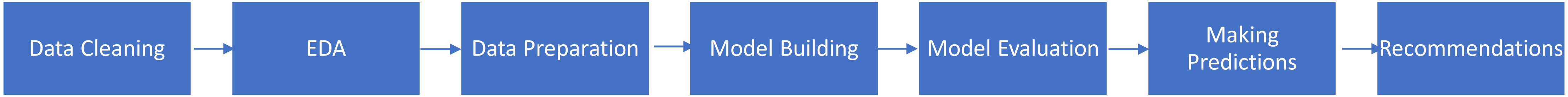
PROBLEM STATEMENT

X Education, an online course provider catering to industry professionals, is grappling with a low lead conversion rate of 30%. Despite attracting a substantial number of leads, the company aims to boost efficiency by identifying and prioritizing 'Hot Leads'—those with a higher likelihood of conversion. The objective is to achieve a target lead conversion rate of 80% by implementing a predictive lead scoring model. This model will assign scores to leads, allowing the sales team to focus efforts on prospects with the greatest potential for conversion, ultimately streamlining the lead nurturing process.

Key Objectives:

- Develop a predictive lead scoring model to assign scores to leads.
- Prioritize 'Hot Leads' with higher scores for targeted engagement.
- Achieve a target lead conversion rate of 80%, optimizing the overall conversion process and maximizing the impact of the sales team's efforts.

ANALYSIS APPROACH



Data Cleaning → EDA → Data Preparation → Model Building → Model Evaluation → Making Predictions → Recommendations

DATA CLEANING

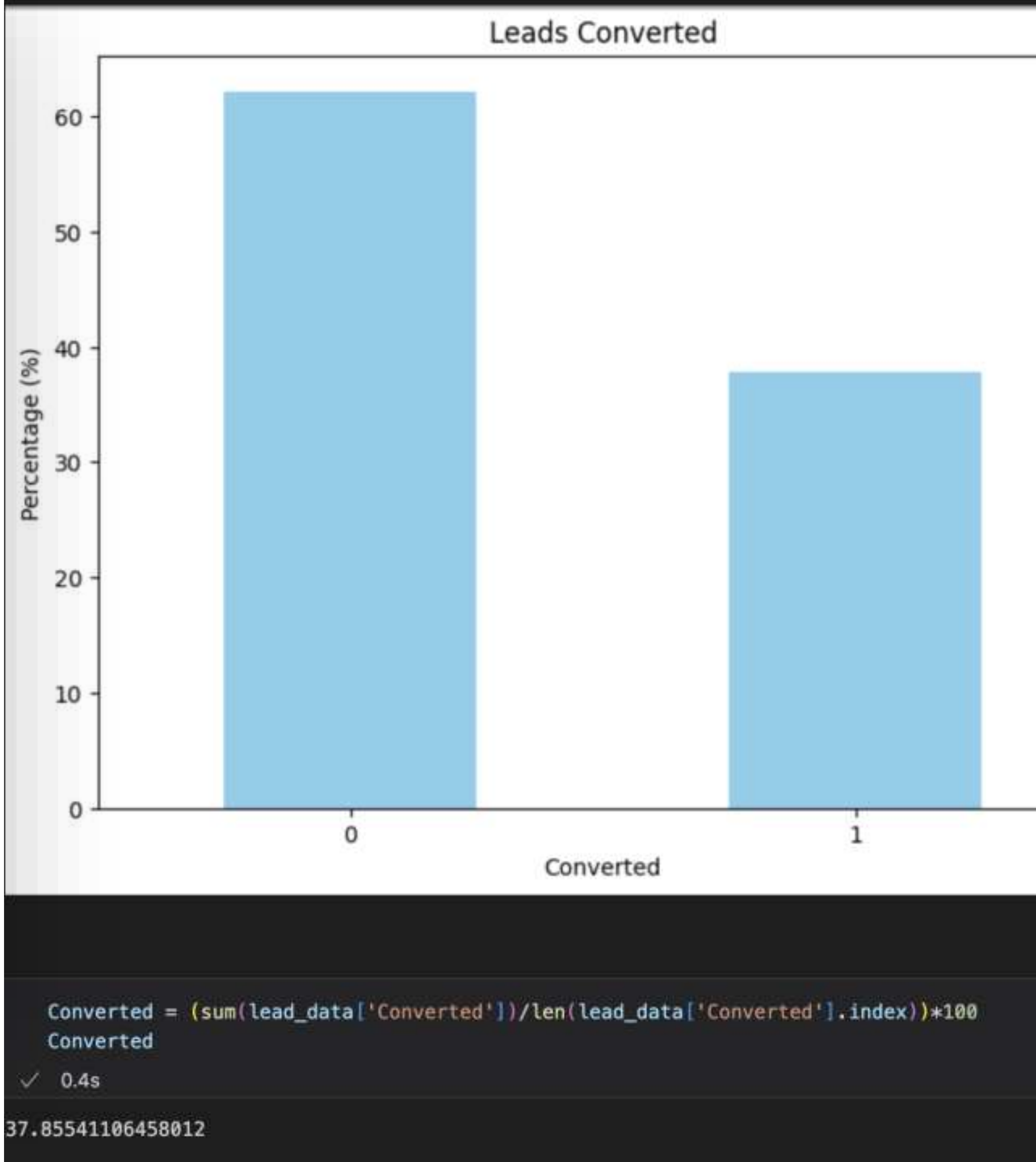
- "Select" level represents null values for some categorical variables, as customers did not choose any option from the list.
- Columns with over 40% null values were dropped.
- Missing values in categorical columns were handled based on value counts and certain considerations.
- Drop columns that don't add any insight or value to the study objective (tags, country)
- Imputation was used for some categorical variables.
- Additional categories were created for some variables.

DATA CLEANING

- Skewed category columns were checked and dropped to avoid bias in logistic regression models.
- Outliers in Total Visits and Page Views Per Visit were treated and capped.
- Low frequency values were grouped together to “Others”.
- Fixed Invalid values & Standardized data in columns by checking typo errors etc. (lead source has Google, google)

EDA

- There is imbalance in data as can be seen from the chart that only 37.8% of the leads have been converted.

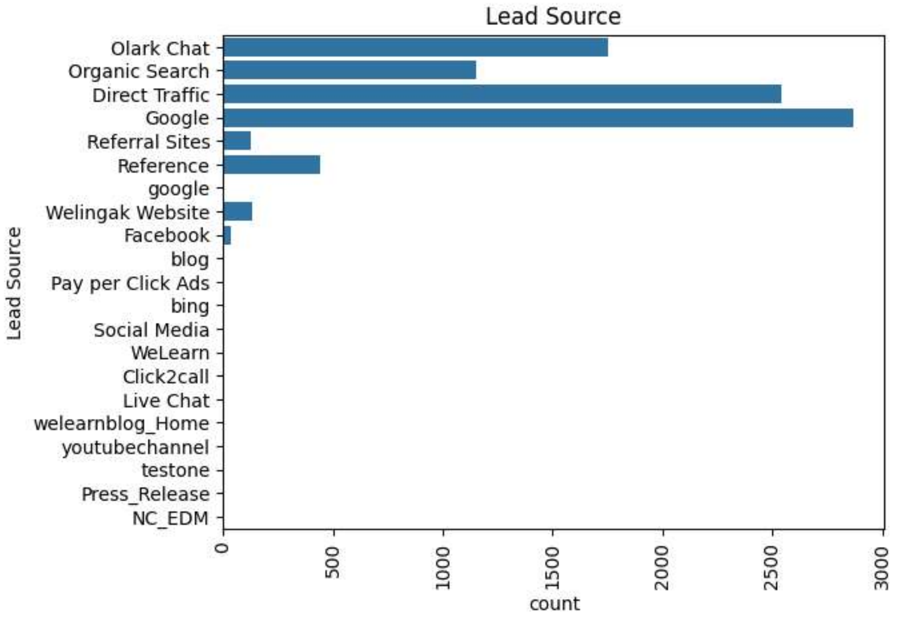


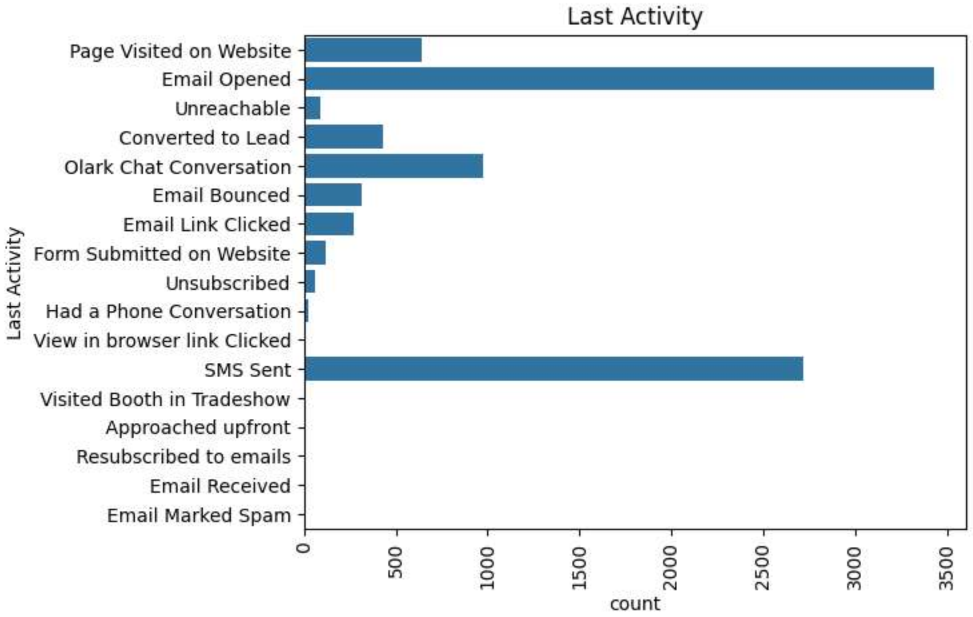
```
Converted = (sum(lead_data['Converted'])/len(lead_data['Converted'].index))*100
Converted
```

```
37.85541106458012
```

EDA

Univariate Analysis



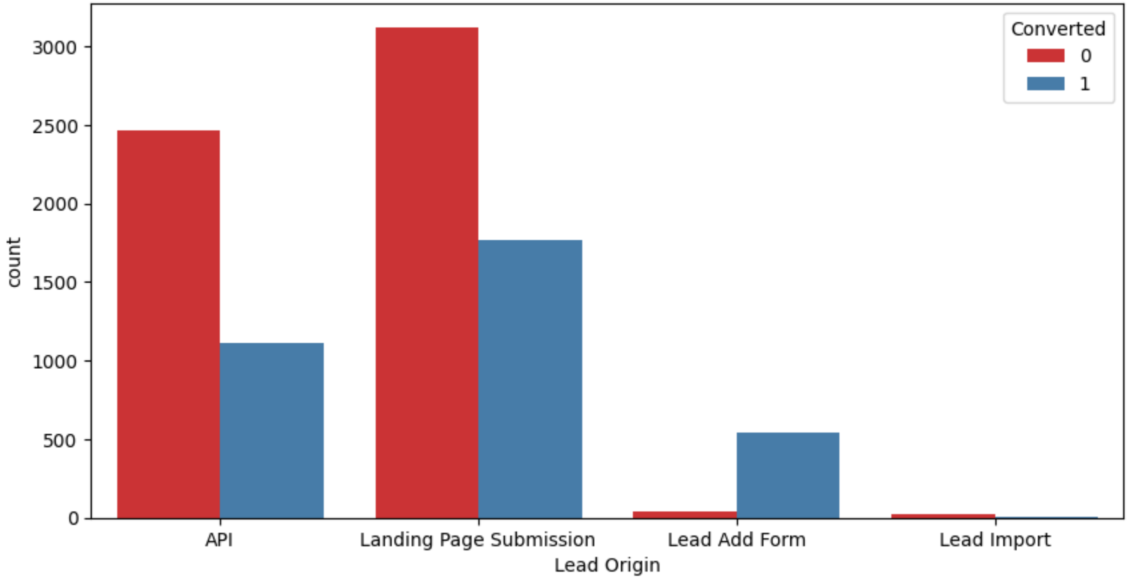


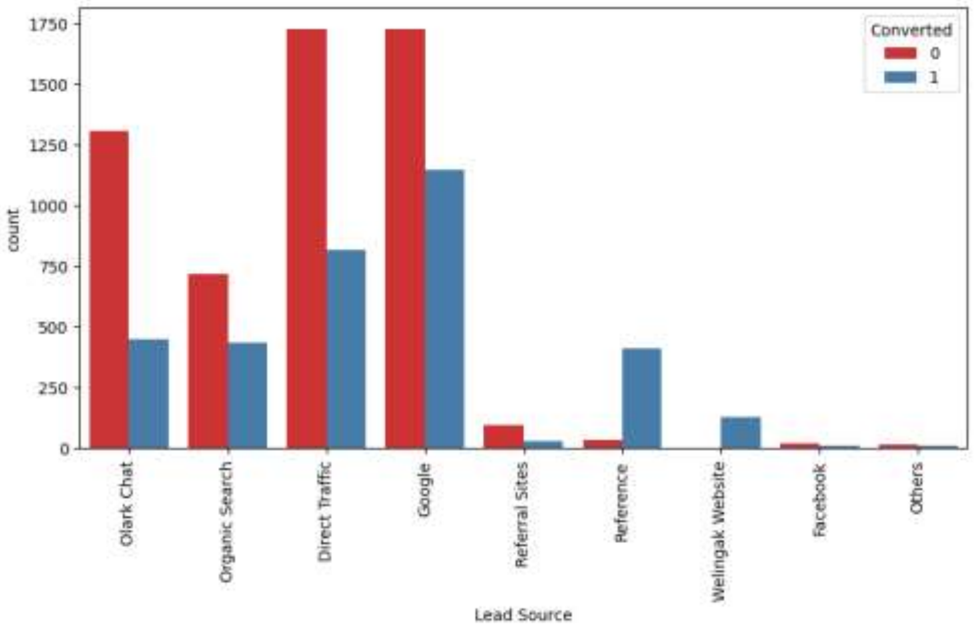
Highest Lead Source is Google followed by Direct traffic.

Last Activity for most is Email Opened followed by SMS sent.

EDA

Bivariate Analysis





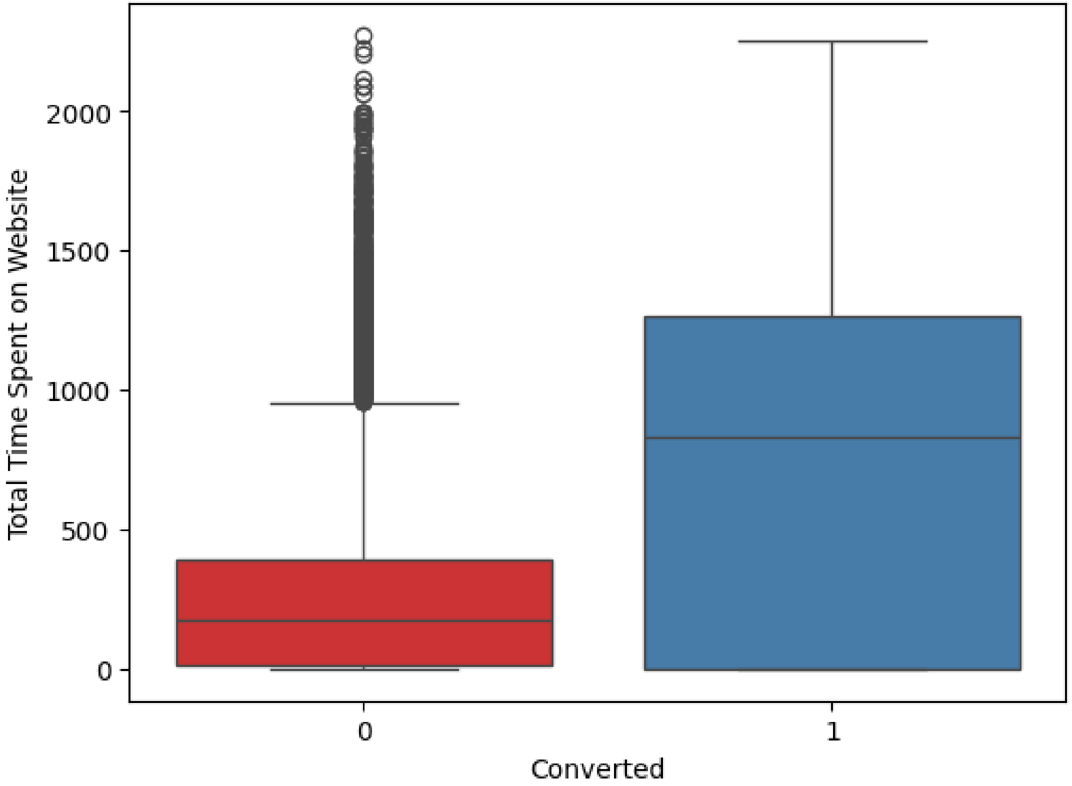
We can note that

- API and Landing Page Submission have 30-35% conversion rate but count of lead originated from these is considerable.
- Lead Add Form has more than 90% conversion rate but count of lead are not very high.
- Lead import has very less count.

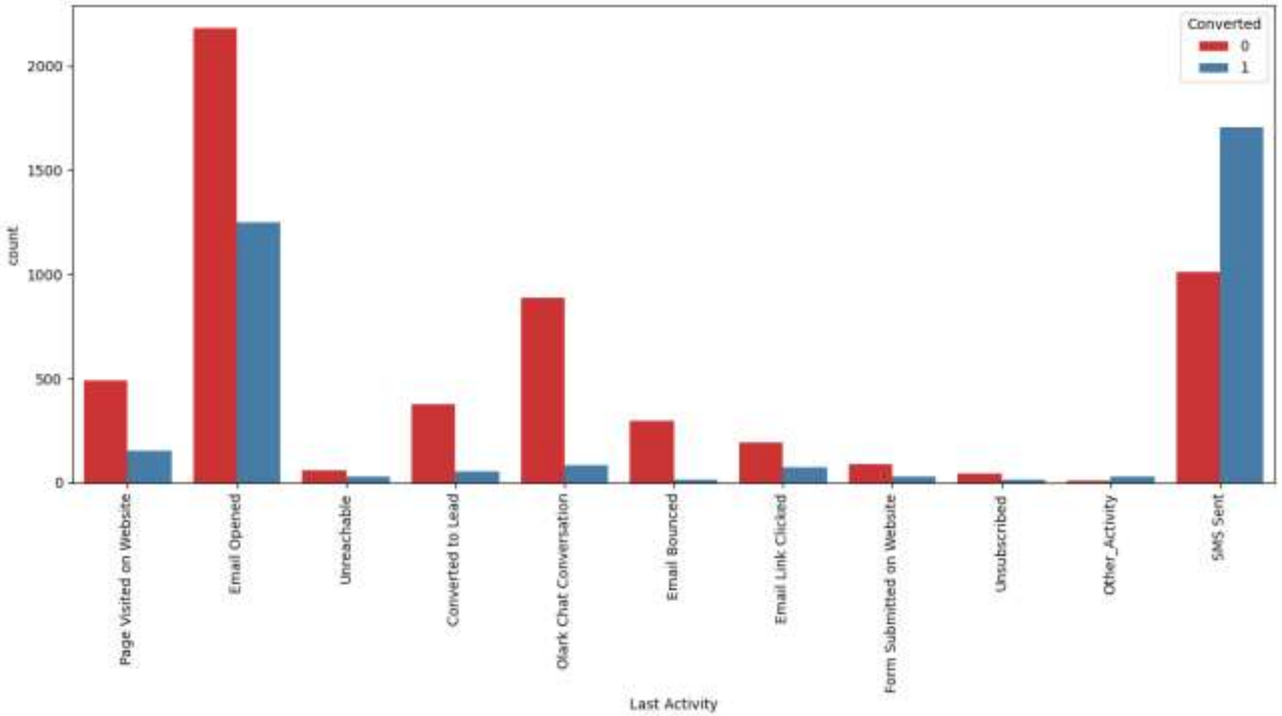
We can note that

- Google and Direct traffic generates maximum number of leads.
- Conversion Rate of reference leads and leads through welingak website is high.

Bivariate Analysis



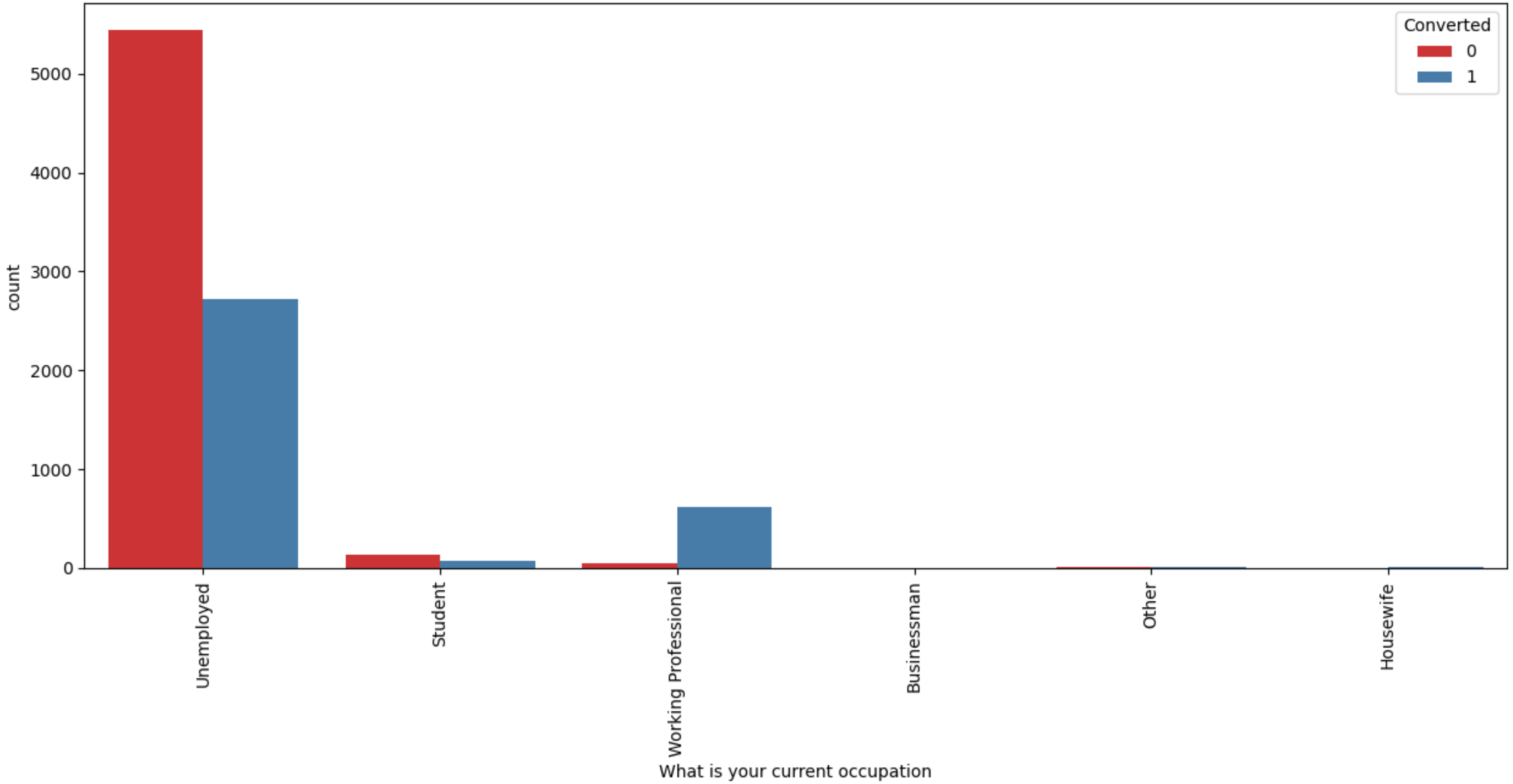
Leads spending more time on the website are more likely to be converted.



We can note that

- Most of the leads have Email Opened as their last activity.
- Conversion rate for leads with last activity as SMS sent is almost 60%.

Bivariate Analysis



We can note that

- Working professionals have a high chance of joining the course.

- Unemployed leads are the most in numbers but conversion rate is around 30-35%.

DATA PREPARATION

- Binary categorical variables were converted to 0 and 1.
- Dummy variables were created for categorical variables.
- Data was split into train and test sets (0.7:0.3)
- Feature scaling was done using Standard Scaler method.

FEATURE SELECTION

Feature Selection using RFE

- The dataset contains many features which will reduce overall model performance.
- We have used Recursive Feature Elimination to reduce the number of features to significant ones.
- We fine tune the model manually
- We were left with 20 columns after RFE.

MODEL BUILDING

- WE manually dropped less significant variables one by one based on their p values. (p>0.05 were dropped)
- Model 9 is our final model as p values are all closer to 0 and VIF values are less than 5 for all variables.
- We are left with 12 features after this process.

MODEL EVALUATION

Results :

Comparing the values obtained for Train & Test:

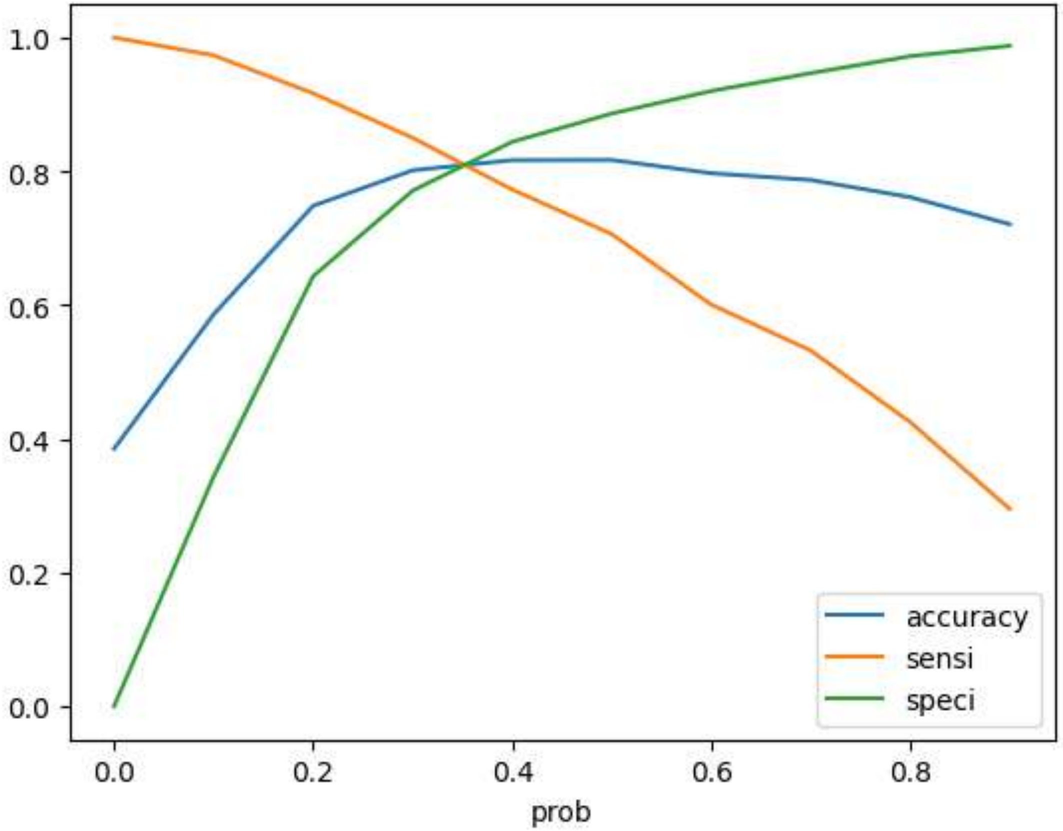
Train Data:

- Accuracy : 81.0 %
- Sensitivity : 81.7 %
- Specificity : 80.6 %

Test Data:

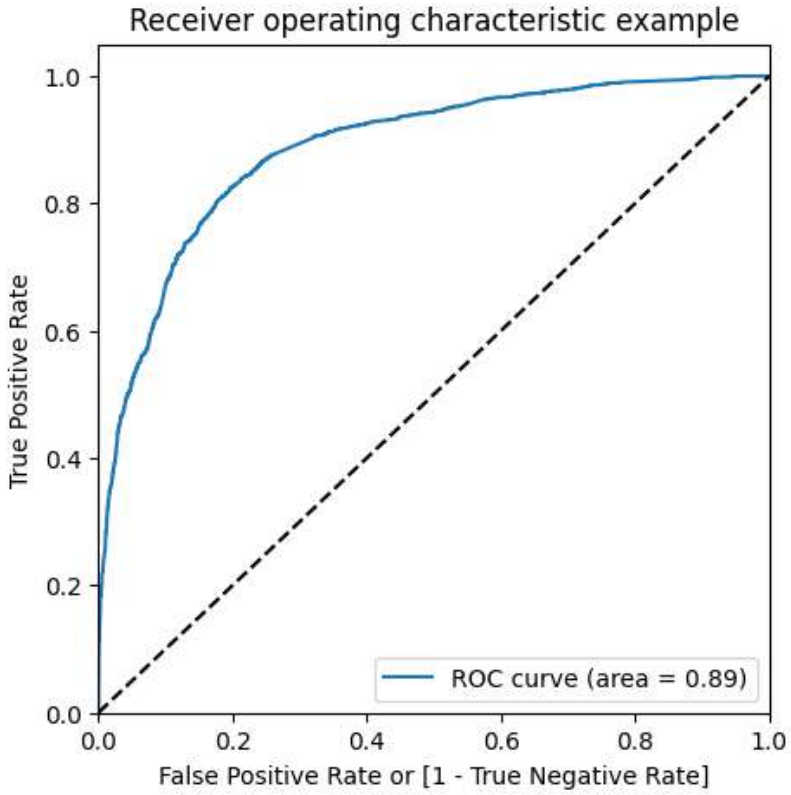
- Accuracy : 80.4 %
- Sensitivity : 80.4 %
- Specificity : 80.5 %

Cutoff probability is 0.34

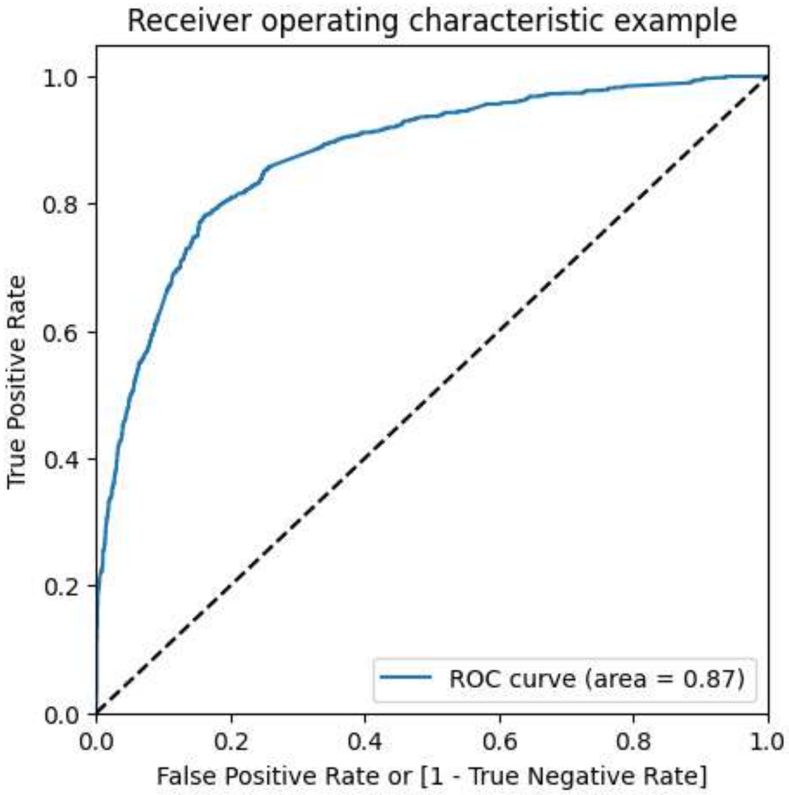


ROC Curves

Area under the ROC curve for training set is 0.89 which indicates a good predictive model.



Area under the ROC curve for test set is 0.87 which indicates a good predictive model.



RECOMMENDATIONS

Make Calls

- The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Reference" as these are more likely to get converted.
- The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.
- The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.
- The company should make calls to the leads coming from the lead sources "Olark Chat" as these are more likely to get converted.
- The company should make calls to the leads whose last activity was SMS Sent as they are more likely to get converted.

Do not Call

- The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.
- The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.
- The company should not make calls to the leads whose Specialization was "Others" as they are not likely to get converted.
- The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.

THANK YOU