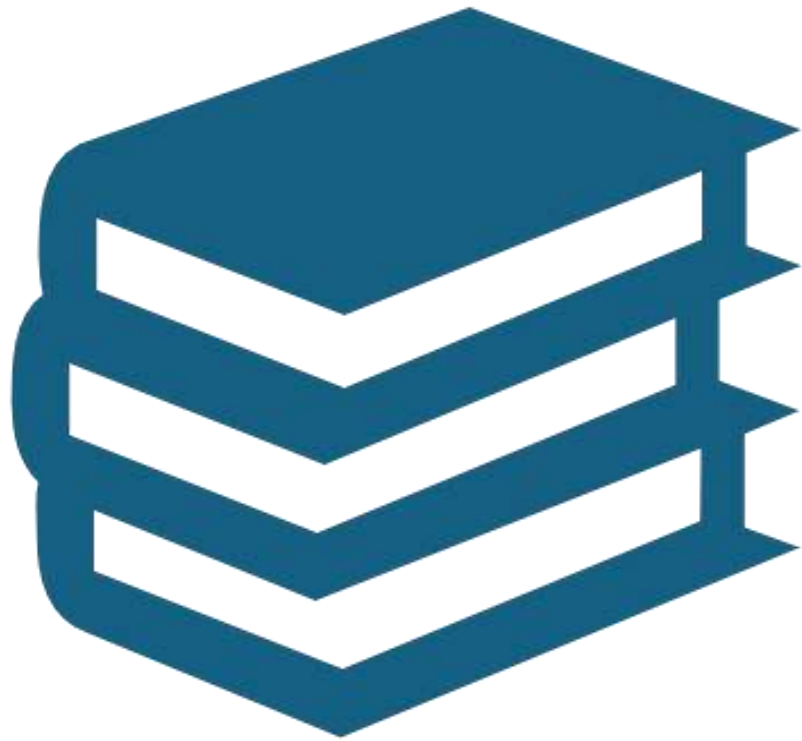


Problem Statement

- X Education, an online education company, faces a challenge in optimizing its lead conversion process. Despite generating a substantial number of leads through various channels such as website visits, form submissions, and past referrals, **the company struggles with a low conversion rate of approximately 30%**. This inefficiency results in a significant waste of resources as the sales team spends considerable time and effort reaching out to a large number of leads, many of whom do not convert into paying customers.
- To address this issue, X Education aims to identify the most promising leads—those with the highest likelihood of converting into paying customers, referred to as "Hot Leads." By accurately targeting these Hot Leads, the company expects to improve its overall lead conversion rate, **with a goal set by the CEO to achieve an 80% conversion rate.**
- The company seeks to implement a predictive model that assigns a lead score to each potential customer, indicating their likelihood of conversion. This model will enable the sales team to focus their efforts



Goals of the Case Study

1. 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.
2. There are some more problems presented by the company which our model should be able to adjust to if the company's requirement changes in the future so we will need to handle these as well. These problems are provided in a separate doc file.

Approach

Problem Understanding and Data Collection

- Understand the Problem and clarify the objective
- Collect Data: including demographic details, interaction history, lead source, and other features

Data Preprocessing

- Data Cleaning
- Encoding Categorical Variables into numerical format
- Create new features by combining existing ones if necessary
- Standardize or normalize the features

Exploratory Data Analysis (EDA)

- Calculate summary statistics for numerical and categorical features.
- Use plots like histograms, box plots, and correlation matrices to understand the distribution of features and relationships between them.
- Identify Key Features and analyze which features might have the most impact on lead conversion.

Splitting the Data

- Train-Test Split
- Further split the training set into training and validation sets to tune the model parameters

Model Building

- Fit a Logistic Regression model using the training data.
- Regularization: Apply regularization (L1 or L2) to prevent overfitting, especially if there are many features.
- Feature Selection: Use techniques like Recursive Feature Elimination (RFE) to identify the most important features and remove irrelevant ones.

Model Evaluation

- Performance Metrics
- Check Accuracy, Precision, Recall, and F1-Score and then create a Confusion Matrix

Model Interpretation

- Coefficients Analysis: Interpret the coefficients of the Logistic Regression model to understand the impact of each feature on lead conversion.

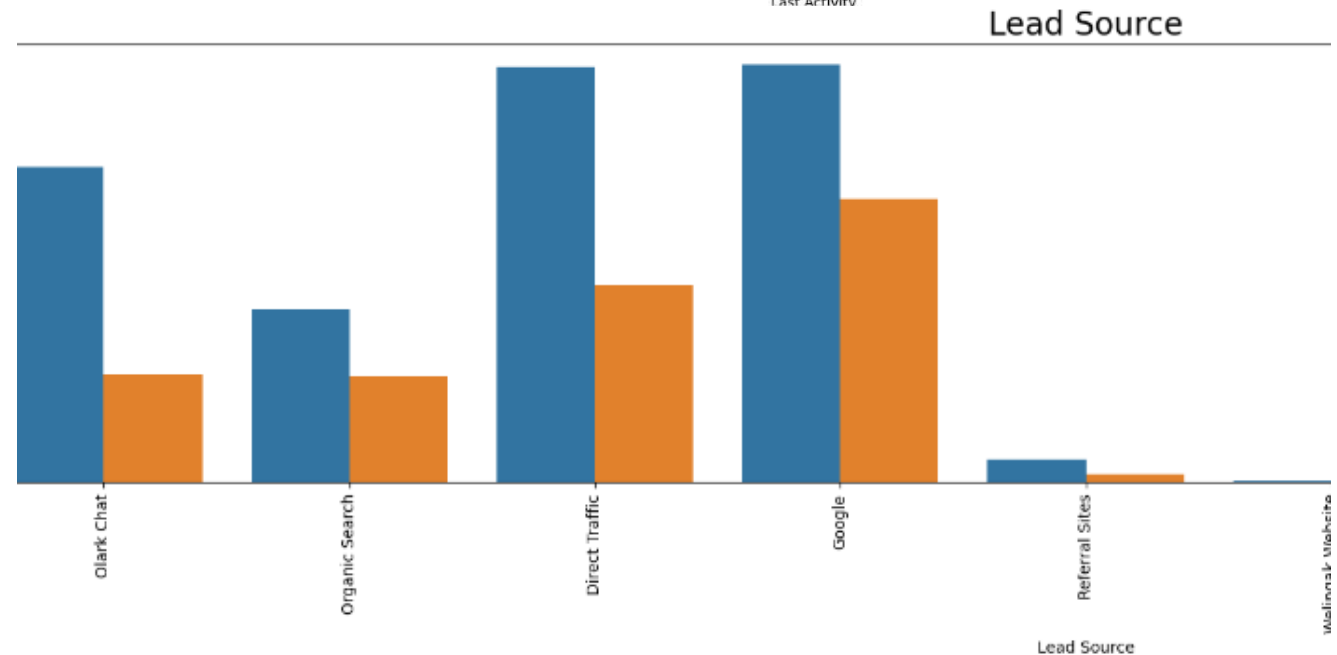
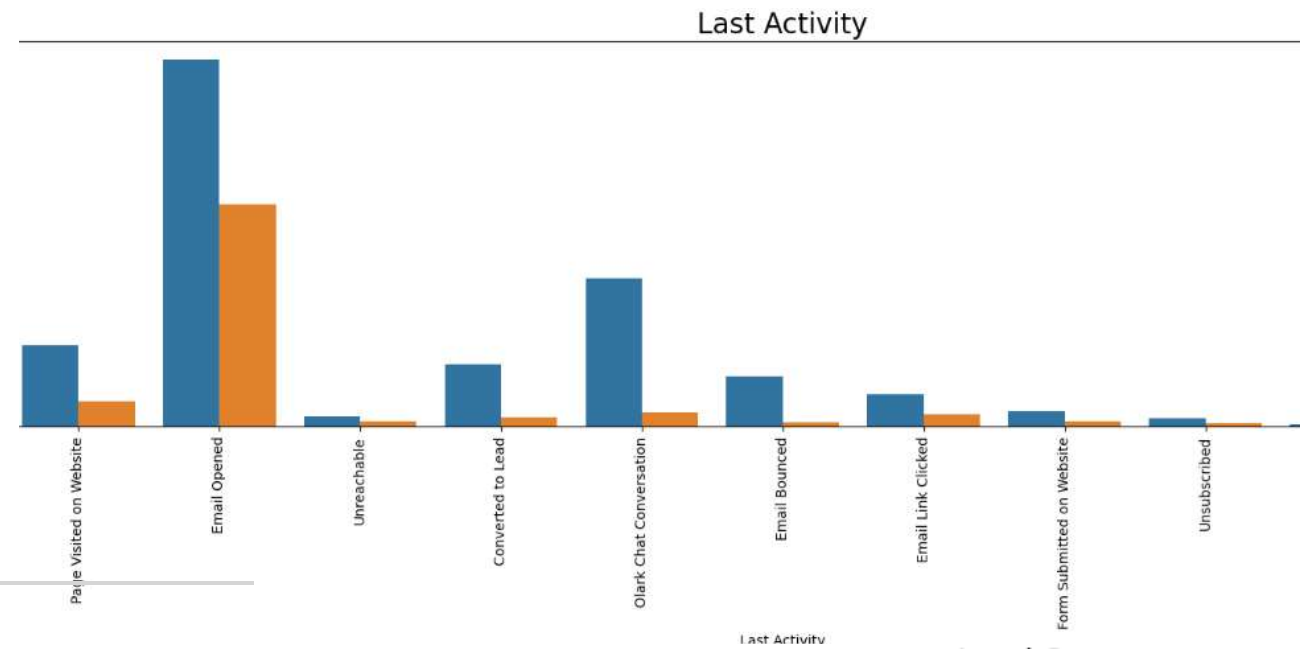
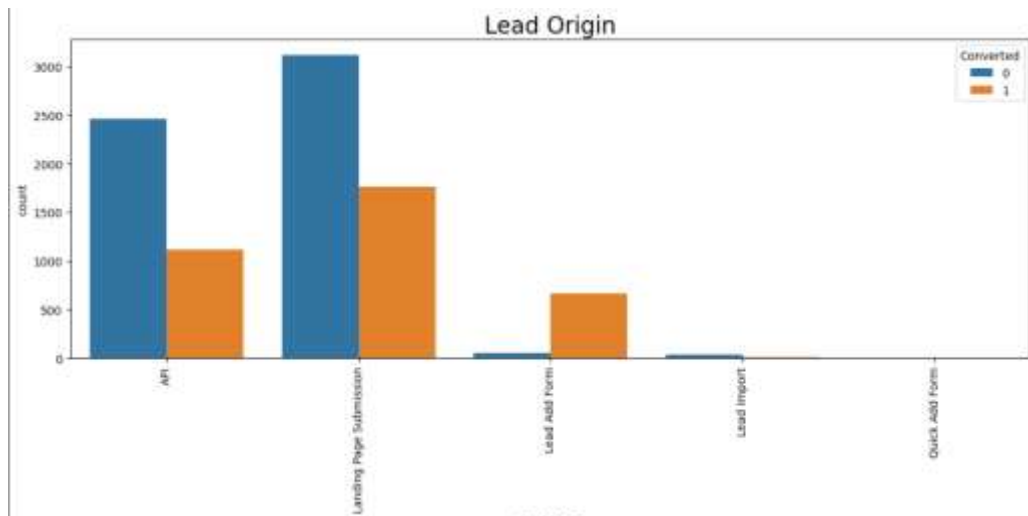
Model Deployment

- Final Model Selection
- Continuously monitor the model's performance and recalibrate it as necessary over time

Reporting and Documentation

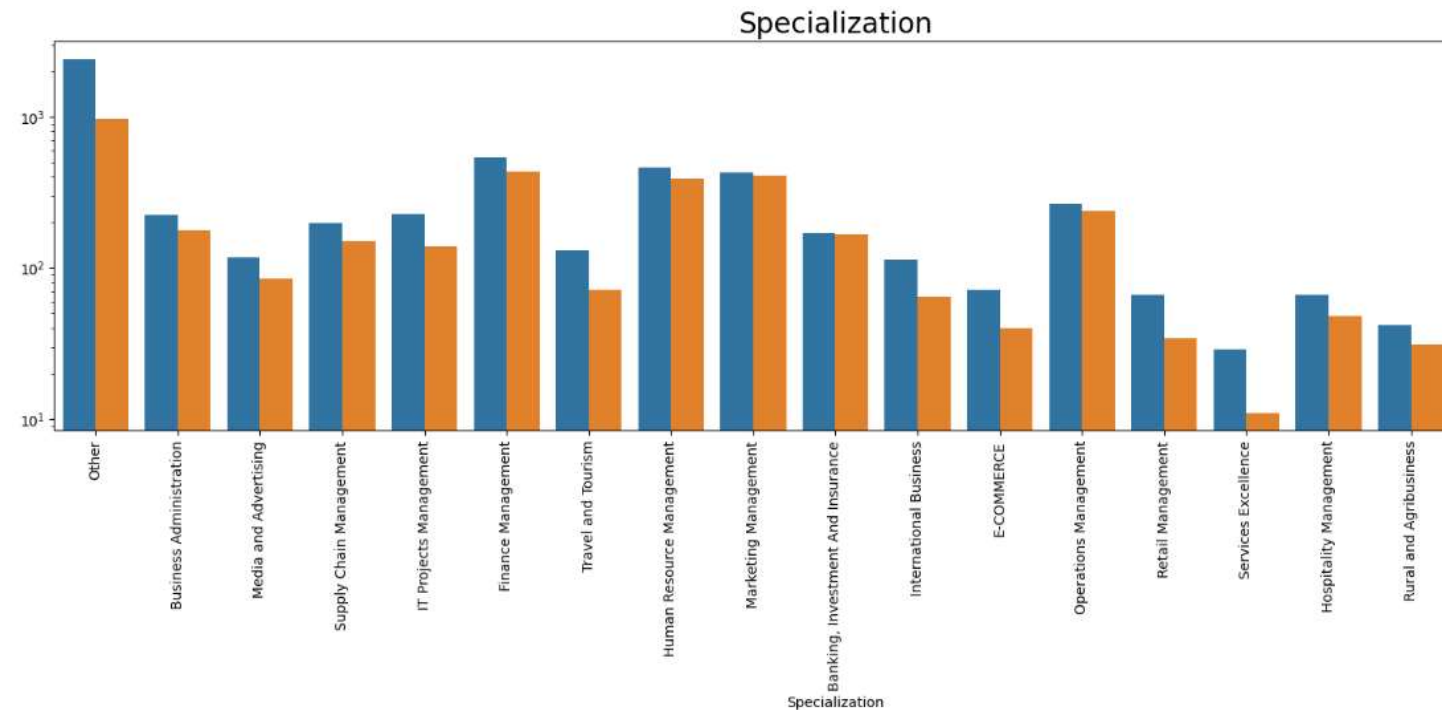
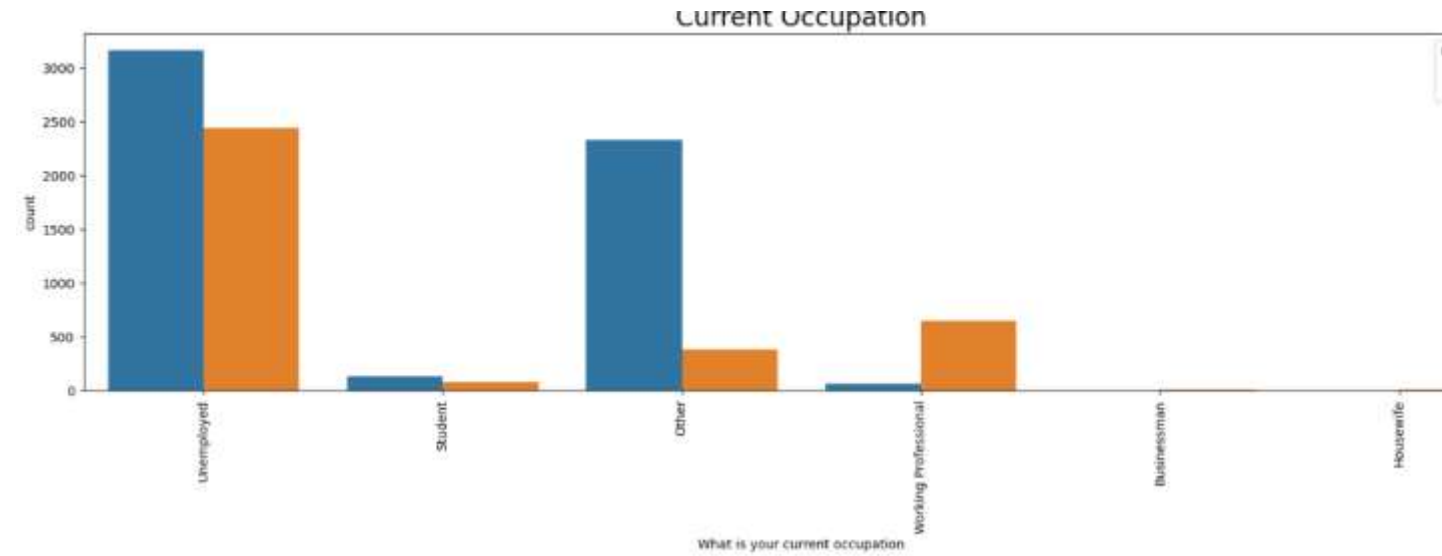
EDA of Categorical Variables

- ❖ Google and Direct traffic generates maximum number of leads
- ❖ Conversion rate of 'Reference' and 'Welingak Website' leads is high
- ❖ We should focus be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website
- ❖ Conversion rate for last activity of 'SMS Sent' is ~63%.
- ❖ Highest last activity of leads is 'Email Opened' .
- ❖ People who opted for mail option are becoming more leads.
- ❖ We have more unsuccessful conversion through 'API' 'Landing Page Submission' as compared to 'Lead Add Form'
- ❖ For 'Lead Add Form' number of conversion is more than unsuccessful conversion.
- ❖ Count of 'Lead Import' is lesser.
- ❖ So to improve overall lead conversion rate, we need to focus more on improving lead conversion of API and Landing Page Submission origin



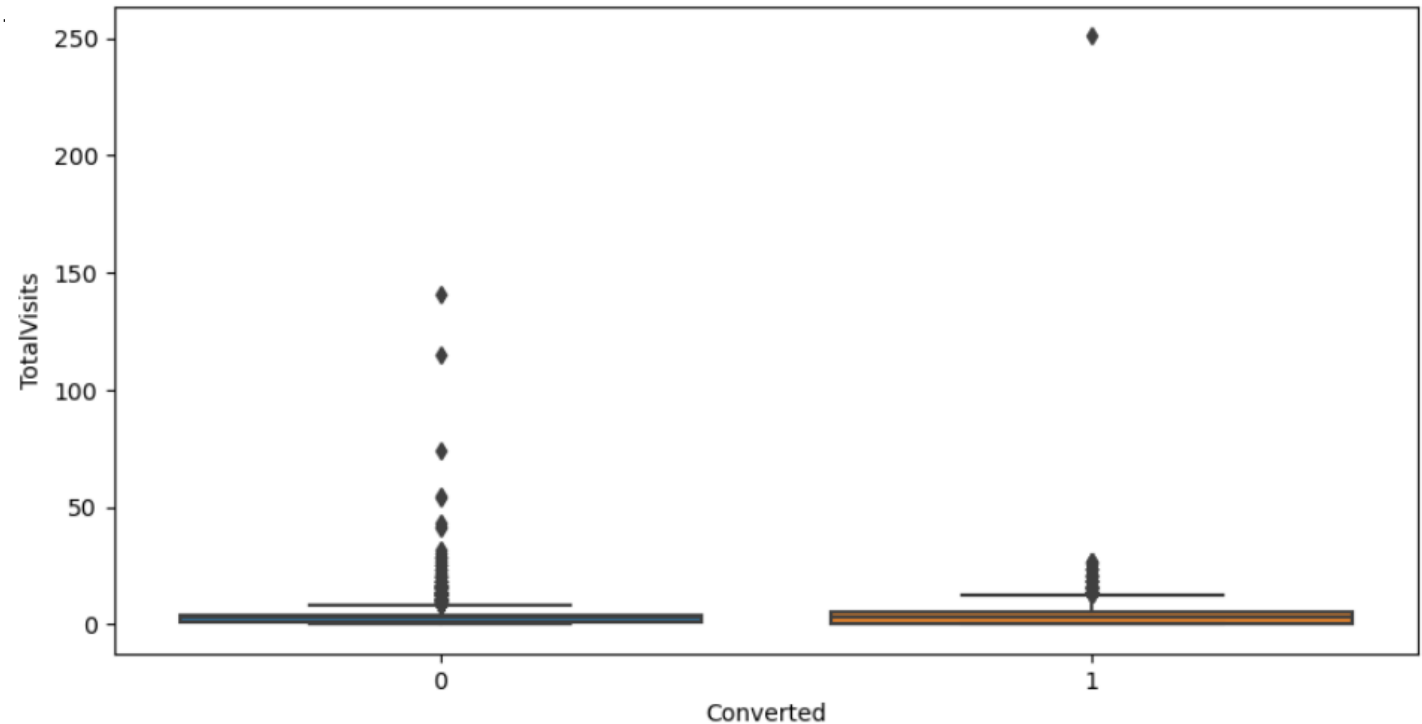
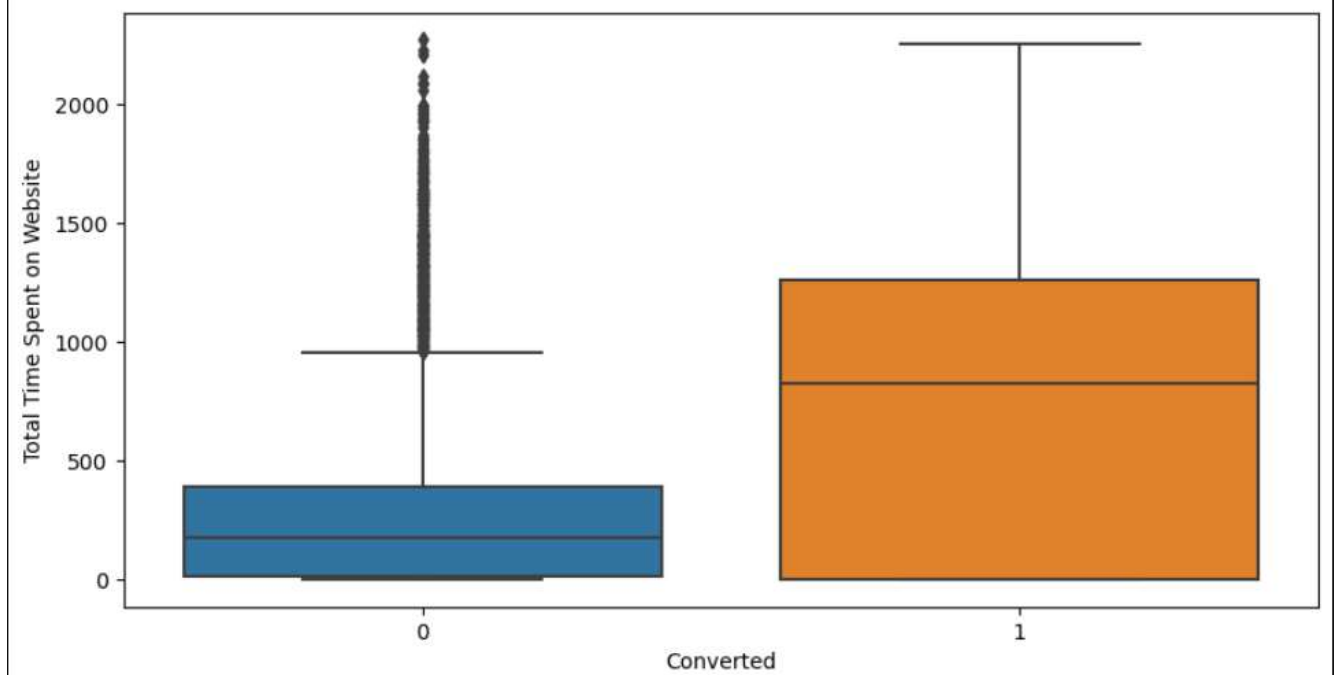
EDA of Categorical Variables

- ❖ 'Unemployed' leads are generating a greater number of leads and having ~45% conversion rate.
- ❖ Conversion rate is higher for 'Working Professionals'.
- ❖ 'Management' specialization are generating a greater number of leads along with Other category



EDA of Numerical Variables

- ❖ Leads spending more time on website are the ones which are getting converted the most.
- ❖ Median for converted and non-converted leads are close
- ❖ We can not say anything based on Total visit
- ❖ Similarly, we cannot say anything conclusive based on page viewers per visit



Correlation Check



'Lead Source_Facebook' and 'Lead Origin_Lead Import' have correlation of 0.98, this is the highest correlation value we have in the data



'Do Not Email' and 'Last Activity_Email Bounced' also have higher correlation.

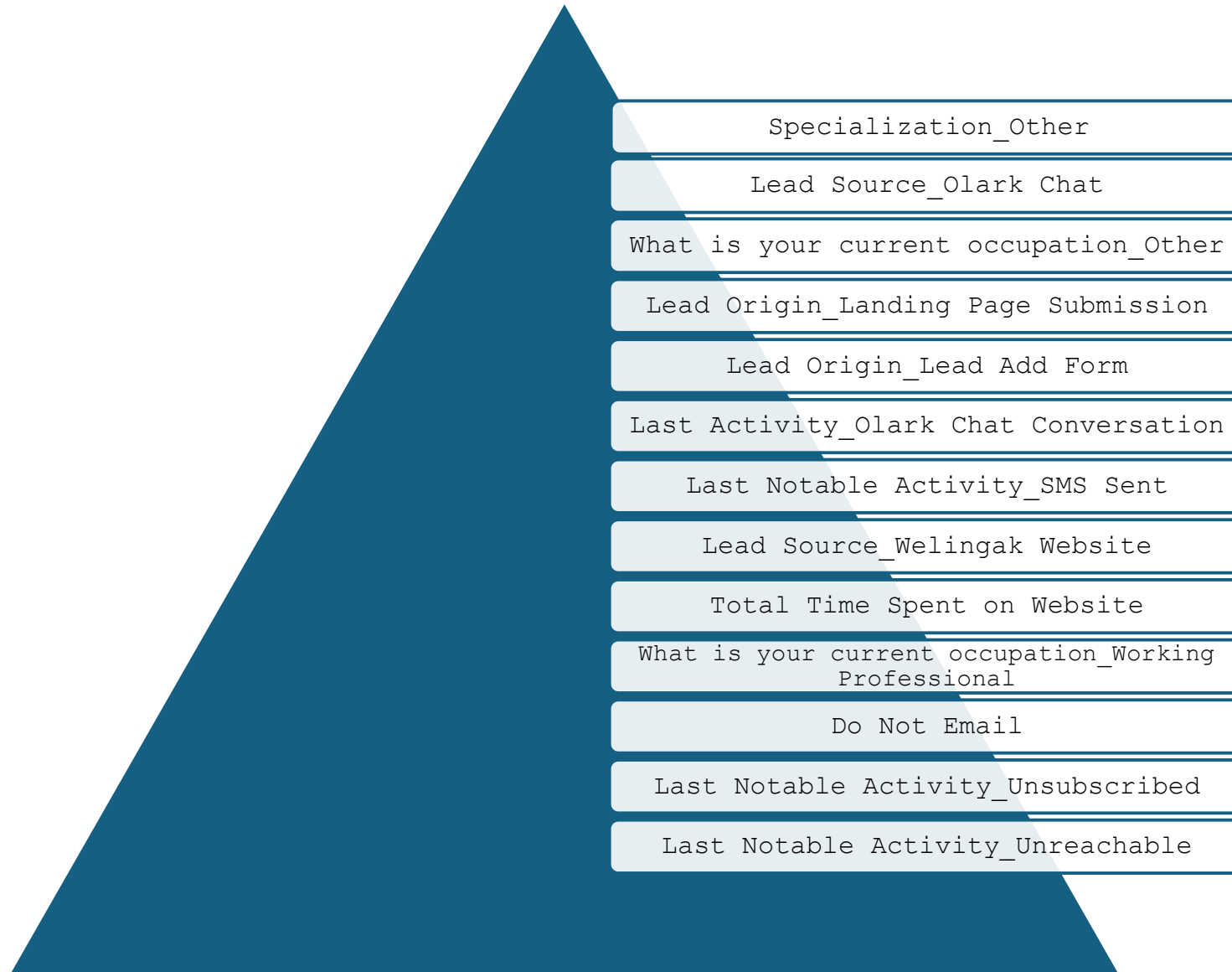


'Lead Origin_Lead Add Form' and 'Lead Source_Referance' have high correlation of 0.85.

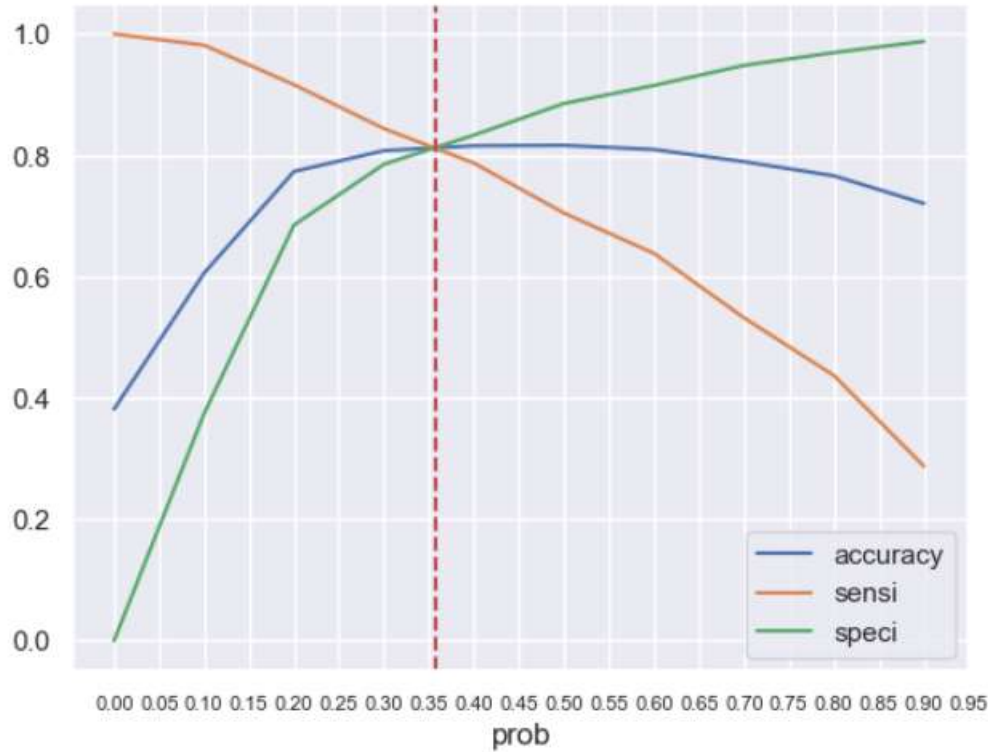


'TotalVisits' and 'Page Views Per Visit' have high correlation of 0.72.

Important variables for Conversion Rate after building the Logistic Regression Model

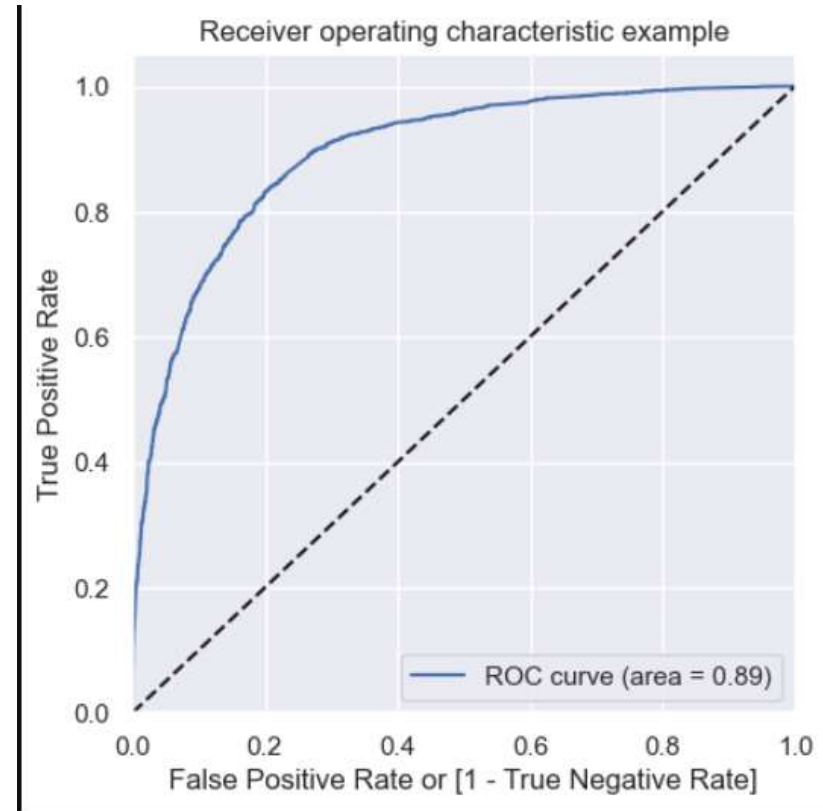


Model Evaluation - Train Data Set



From above graph we can see that cutoff point is 0.358 based on:

- Accuracy – 81%
- Sensitivity - 81%
- Specificity – 80%

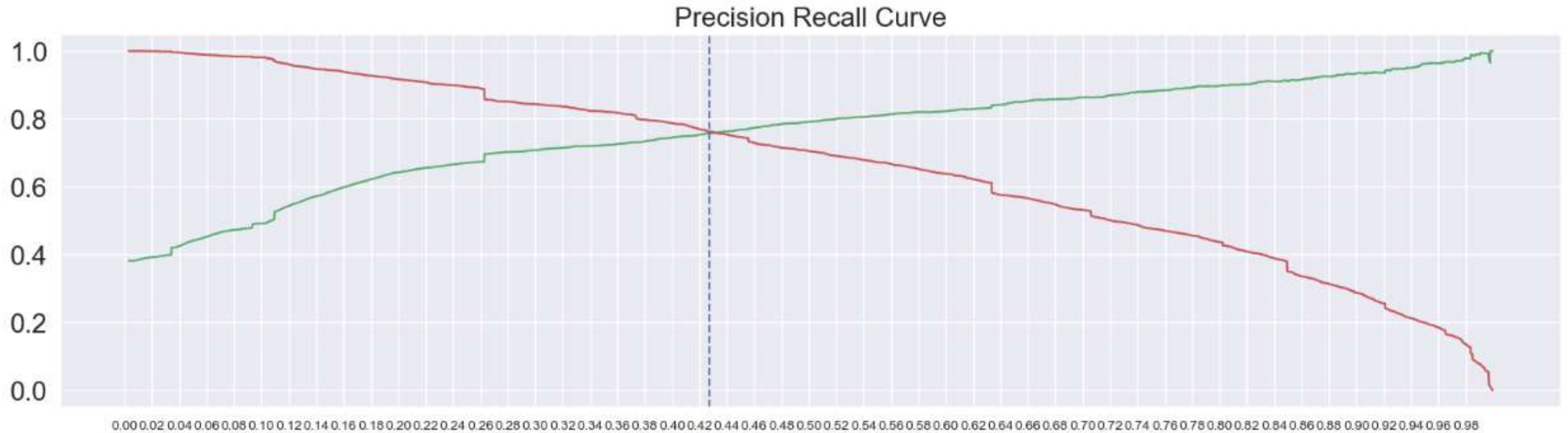


- We are getting a good value of 0.89 indicating a good predictive model.
- As ROC Curve should be a value close to 1.
- So overall this model seems to be performing well.

Confusion Matrix

317 9	756
439	198 9

Model Evaluation - Train Data Set



From above 'precision_recall_curve' we can see that cutoff point is 0.427.

Precision – 72%
Recall – 81%

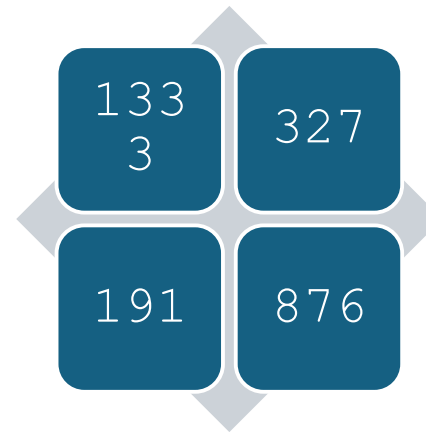
Confusion Matrix

334 3	592
577	185 1

Model Evaluation - Test Data Set

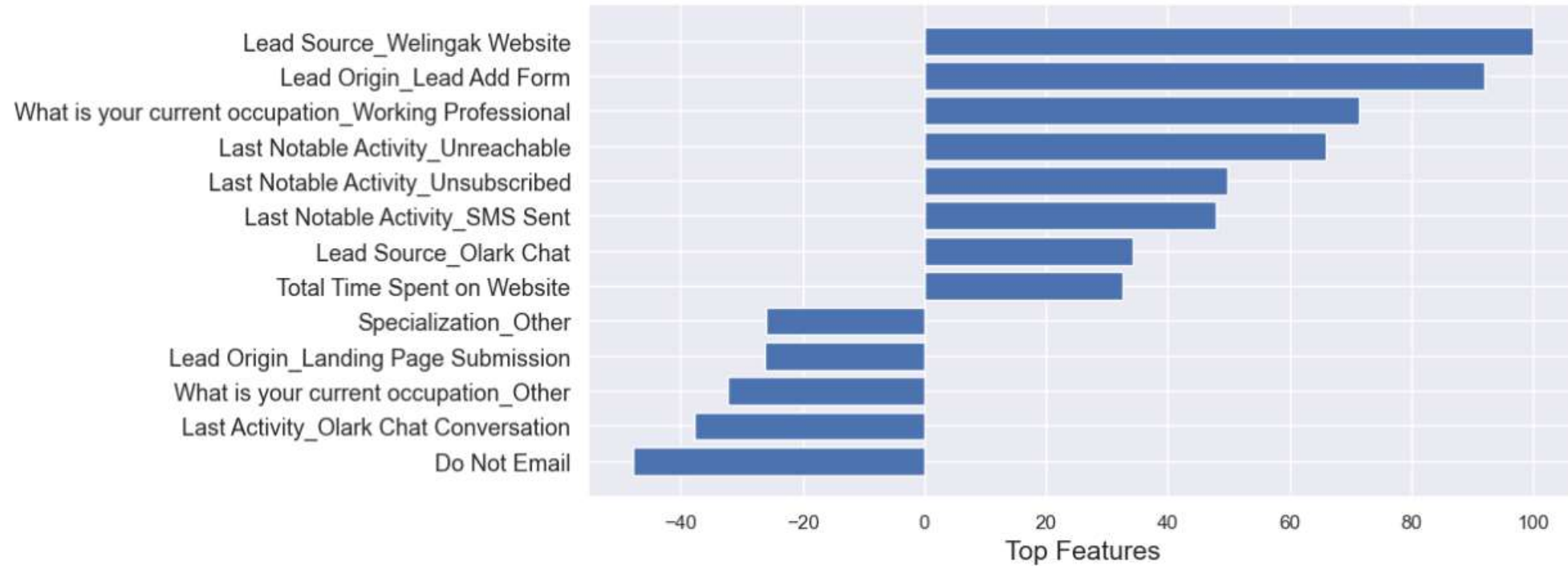
- After running the same configuration on Test data set, we get the final results as below
- Accuracy – 81%
- Sensitivity - 82%
- Specificity – 80%
- Precision – 0.72
- Recall – 0.82

Confusion Matrix



133 3	327
191	876

Final Model and Equation



Top 10 features for better conversion rate

Final Equation :

Converted = $0.261843 + (3.30 * \text{Lead Source_Welingak Website}) + (3.03 * \text{Lead Origin_Lead Add Form}) + (2.36 * \text{What is your current occupation_Working Professional}) + (1.64 * \text{Last Activity_Unsubscribed}) + (1.58 * \text{Last Activity_SMS Sent}) + (1.13 * \text{Lead Source_Olark Chat}) + (1.07 * \text{Total Time Spent on Website}) + (2.17 * \text{Last Notable Activity_Unreachable}) - (0.87 * \text{Lead Origin_Landing Page Submission}) - (0.85 * \text{Specialization_Other}) - (1.07 * \text{What is your current occupation_Other}) - (1.24 * \text{Last Activity_Olark Chat Conversation}) - (1.58 * \text{Do Not Email})$

Conclusion

To improve the potential lead conversion rate X-Education will have to mainly focus important features responsible for good conversion rate are :-

- ❖ Lead Source_Welingak Website : As conversion rate is higher for those leads who got to know about course from 'Welingak Website',so company can focus on this website to get more number of potential leads.
- ❖ Lead Origin_Lead Add Form: Leads who have engaged through 'Lead Add Form' having higher conversion rate so company can focus on it to get more number of leads cause have a higher chances of getting converted.
- ❖ What is your current occupation_Working Professional : The lead whose occupation is 'Working Professional' having higher lead conversion rate ,company should focus on working professionals nad try to get more number of leads.
- ❖ Last Activity_SMS Sent: Lead whose last activity is sms sent can be potential lead for company.
- ❖ Must target leads which are spending a lot of time on website, and visiting again and again and are working professionals. Students can be approached, but they will have a lower probability of converting because they will not be interested in a course specifically designed for working professional
- ❖ Do not target students, they might not be interested in a course specifically designed for working professionals.