Lead Scoring Case Study

September 18, 2023

Overview

“X Education” specializes in selling online courses to industry professionals but faces a significant challenge with a low lead conversion rate, where out of 100 leads acquired in a day, only about 30 are converted into paying customers.

- To enhance efficiency, the company aims to distinguish high-potential leads, often referred to as 'Hot Leads,' from the larger pool.
- Identifying these 'Hot Leads' is crucial, as focusing resources and efforts on these prospects is expected to increase the lead conversion rate by directing the sales team's attention toward leads more likely to convert.
- X Education seeks to implement a lead scoring model that assigns scores to leads based on relevant attributes and behaviors to achieve this objective.
- The goal is to create a data-driven approach that helps the company prioritize leads with a higher likelihood of becoming paying customers, ultimately improving the overall conversion rate.
- Continuous monitoring and adjustment of the lead scoring model will be essential to maintain and further enhance lead conversion rates over time.

Business Objective

Lead Scoring Precision

The company's objective is to develop a lead scoring model that assigns scores to potential leads, reflecting their likelihood of converting into customers. A higher lead score signifies a greater potential for conversion, while a lower score indicates a lower likelihood of conversion.

Conversion Rate Target

The primary goal of this model is to achieve a lead conversion rate of 80% or higher, emphasizing the company's desire to focus on the most promising leads and optimize their conversion efforts.

Promising Lead Prioritization

The lead scoring model will play a pivotal role in helping the company identify and prioritize leads with the greatest potential, ultimately leading to more effective and efficient sales and marketing strategies.

Solution Approach:

Import Data.
Data Inspection.
Data Cleaning and Manipulation:

- Duplicate Data Handling.
- NA and Missing Value Management.
- Column Dropping for Missing Values.
- Value Imputation (if required).
- Outlier Detection and Treatment.

Exploratory Data Analysis (EDA):

- Univariate Data Analysis.
- Data imbalance Ratio.
- Correlations Matrix.

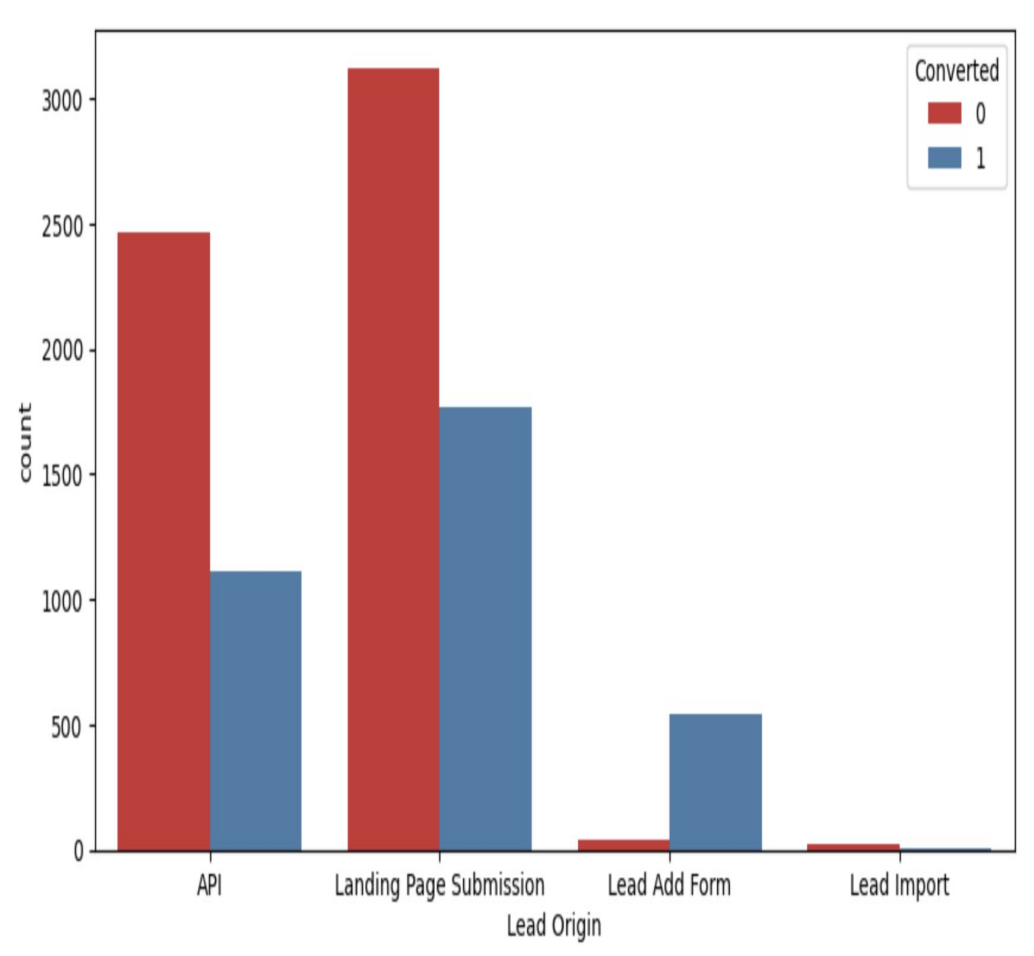
Feature Scaling, Dummy Variables, and Data Encoding.
Classification Technique:

- Logistic Regression for Modeling and Prediction.

Model Validation.
Model Evaluation.
Conclusions and Recommendations.

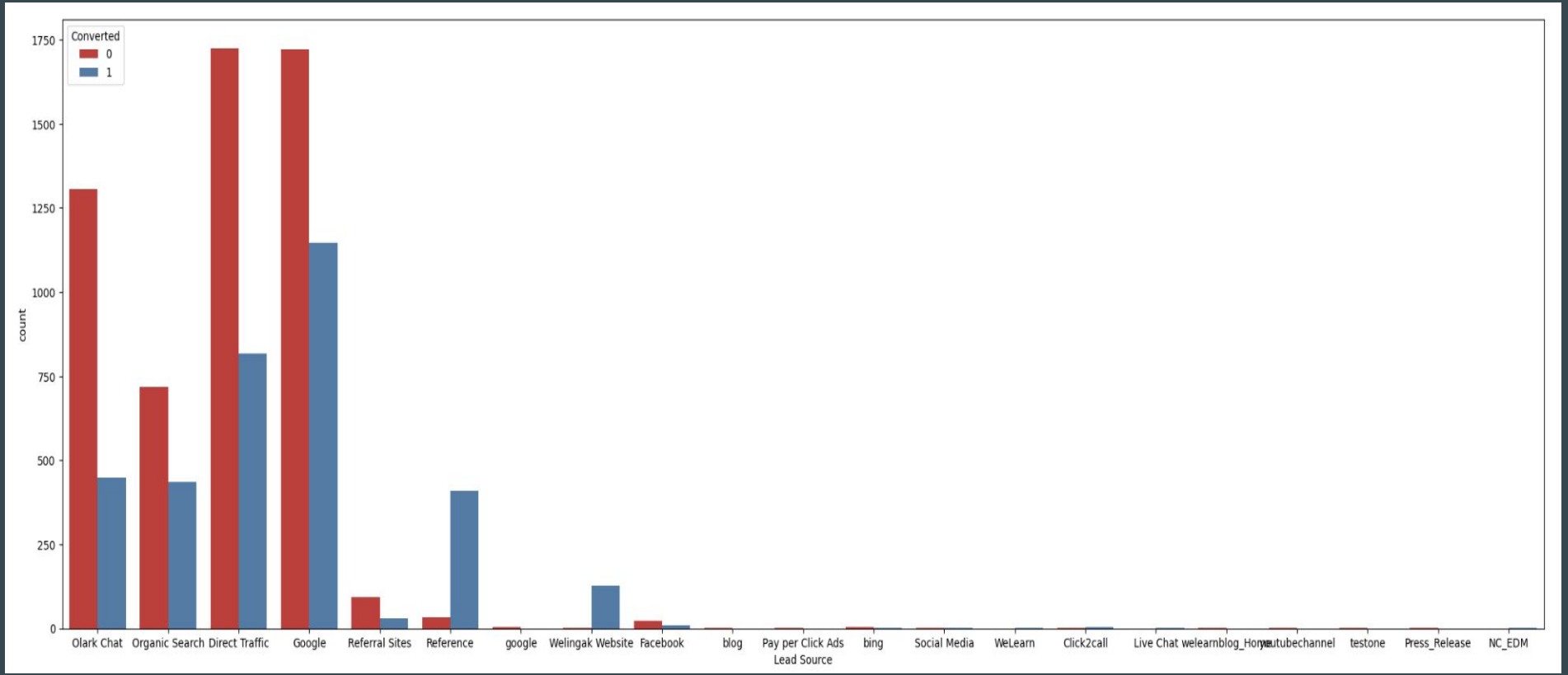
Exploratory Data Analysis

After import, inspect, and cleaning the dataset, we started exploring the data. As “Converted” column is our target variable, we did the comparison with all the independent variables and based on that we have extracted the best features or variables. Univariate analysis of those have been done, visualized and kept this on next slides.

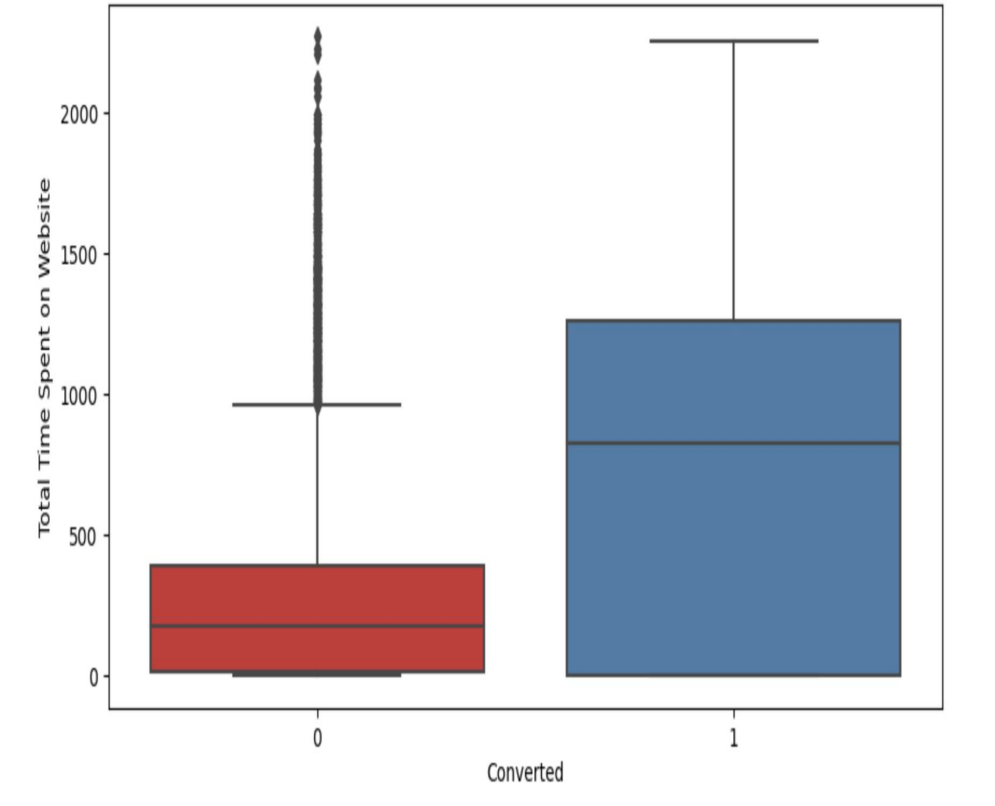


Univariate Analysis :

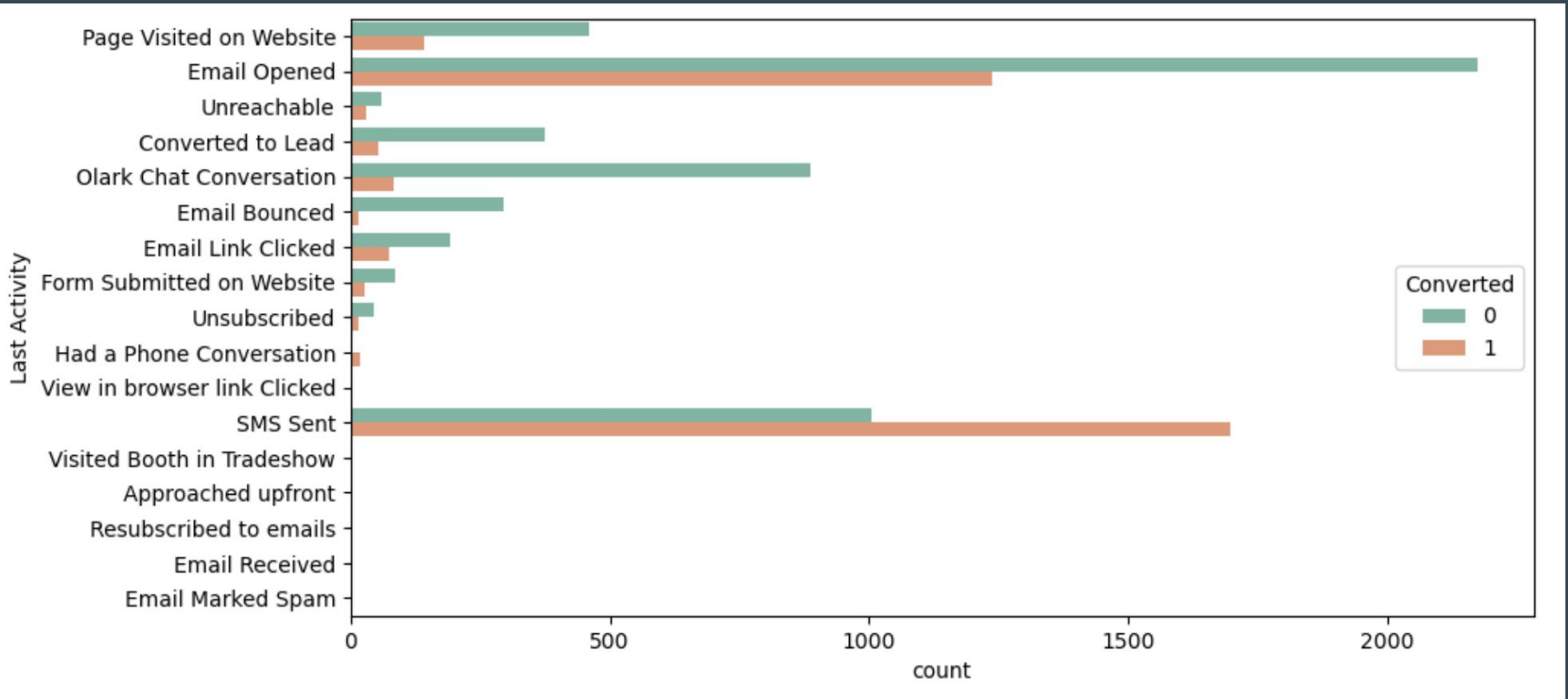
- API and landing page submissions can have a significant impact on increasing the lead conversion rate, as depicted in the plot.



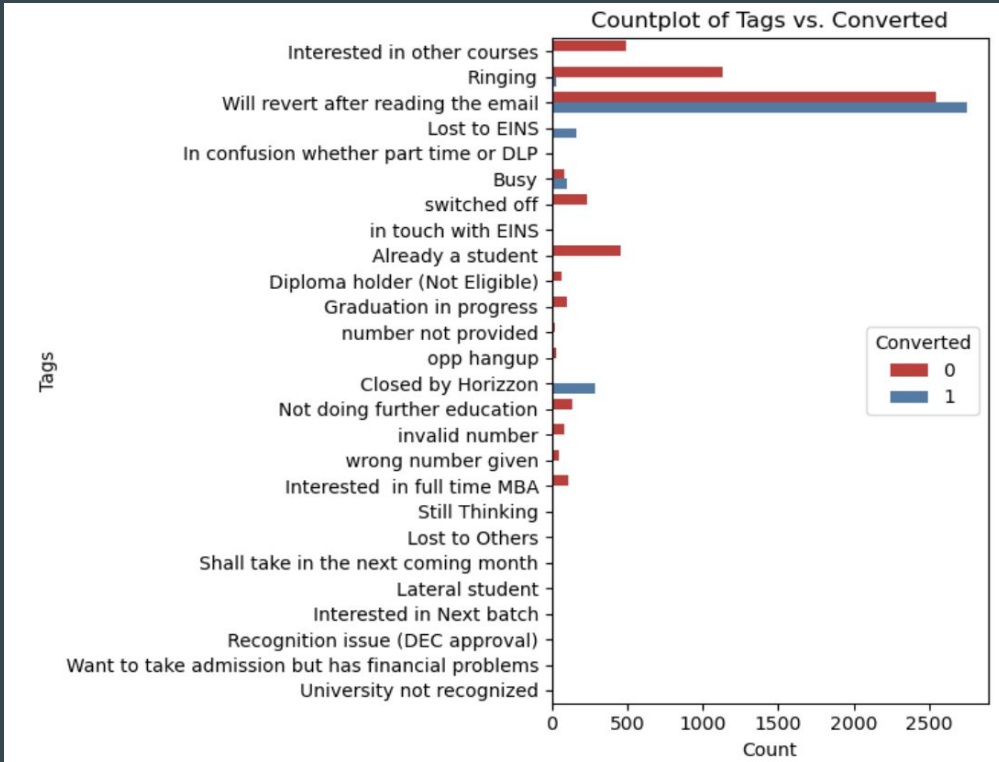
Inference : To improve the lead conversion rate, Team should focus on Google, Direct Traffic, Olark Chat, Organic Search, and References. Where Google, Direct Traffic, and Olark Chat are the top three platform to get more lead conversions.

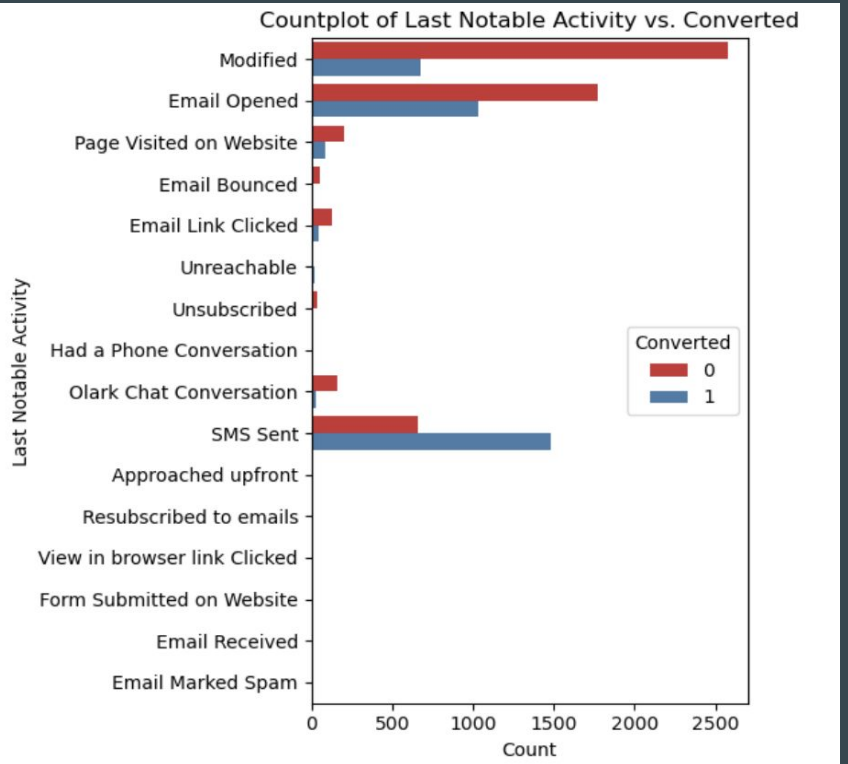


"Website visits" should be more focusable as its lead conversion rate in higher side.

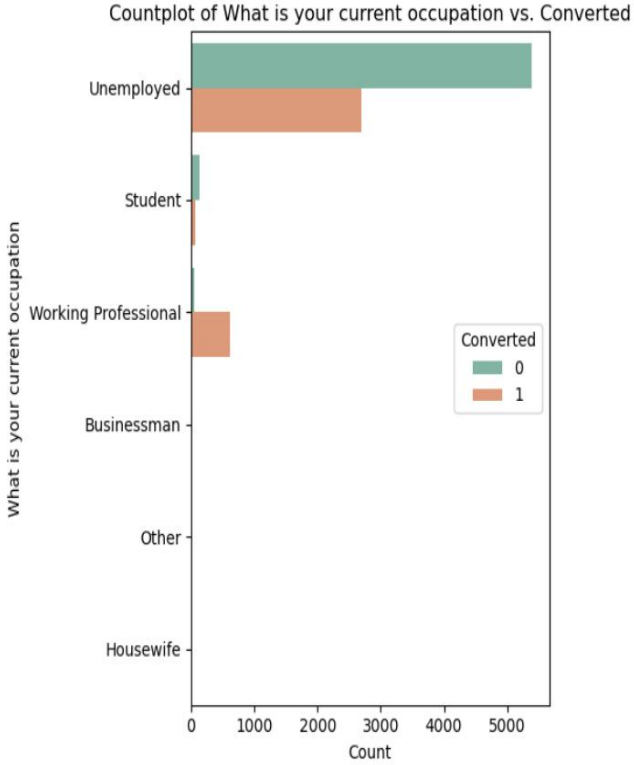


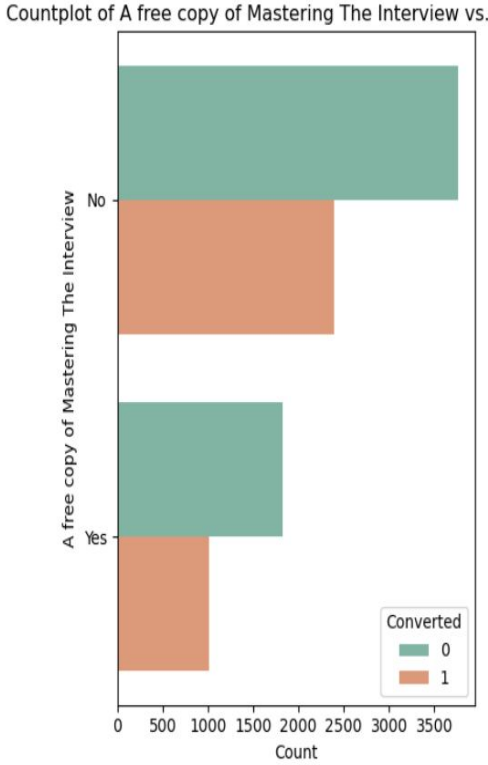
As per the value_count and visualization, we mark that "Email Opened" and "SMS Sent" are having higher lead conversions rate.

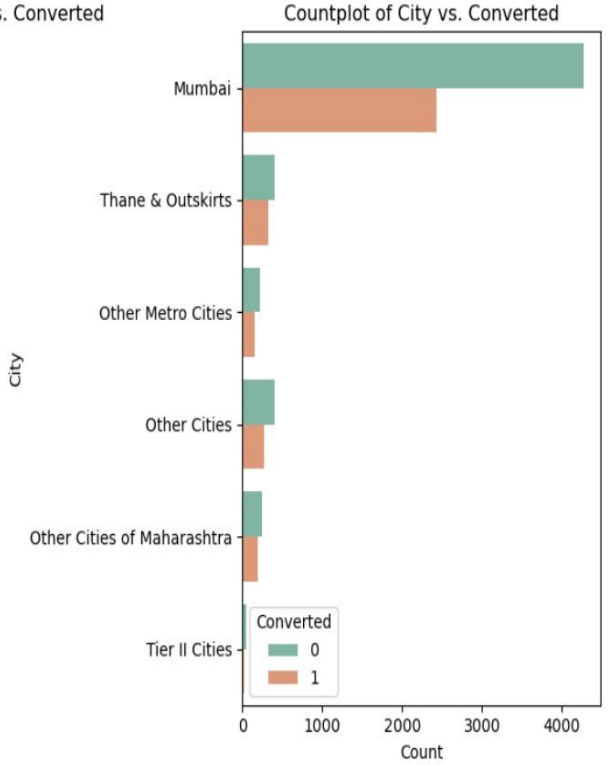




Inference 1 : "Tags" : Lead conversion rate is higher after leads read the email. Mail should be more focusable area.
Inference 2 : "Last Notable Activity" : Most of the student leads are converted after they opened the email content and sms. Where the “Modified” states that mostly non-targeted.





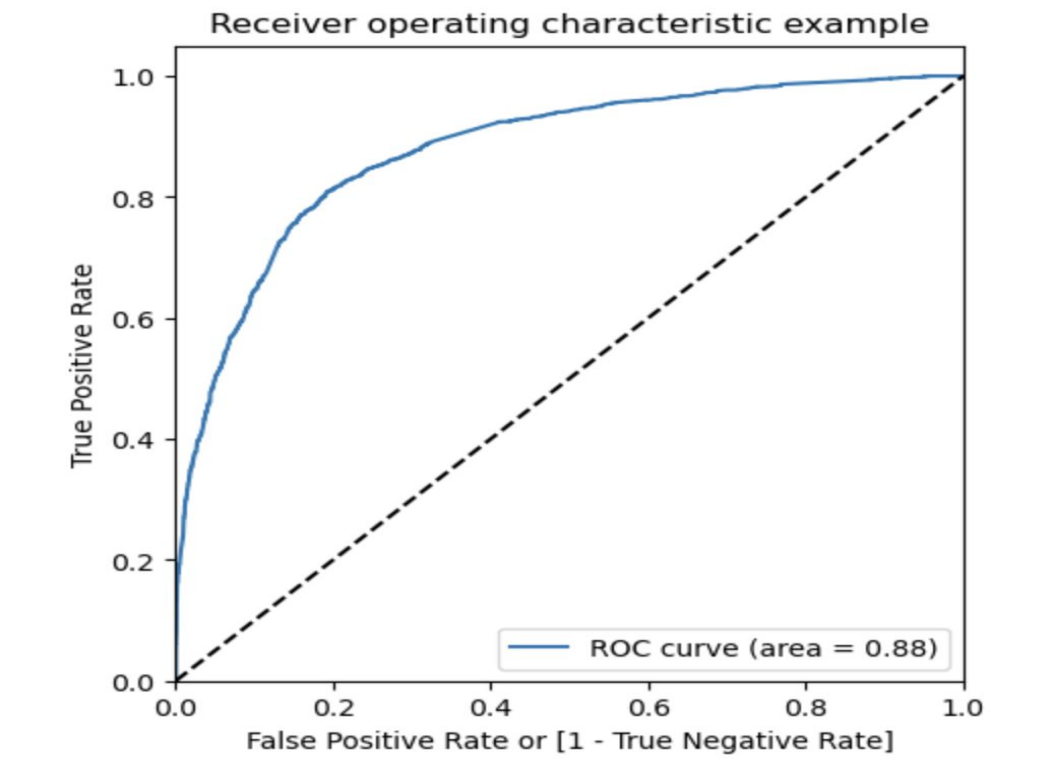


Inference 1 : "Unemployed" and "Working Professional" leads are more converted leads.¶
Inference 2 : "Free Mastering interview" copy is not much significant for lead conversion and non-conversion.¶
Inference 3 : "Mumbai" and its outskirts are having more lead conversion rate.

Model Building

Splitting the dataset into training and testing subsets.
Scaling variables within the training set.
Constructing the initial model.
Employing Recursive Feature Elimination (RFE) to remove less relevant variables.
Developing an updated model based on RFE results.
Eliminating variables with high p-values to refine the model.
Calculating Variance Inflation Factor (VIF) for all remaining columns.
Making predictions using the training dataset.
Evaluating model performance using metrics like Accuracy, Sensitivity, Specificity, and Precision-Recall.
Extending predictions to the test dataset.
Applying the same evaluation metrics to the test predictions.
Comparing the performance metrics between the training and test datasets to assess model generalization.

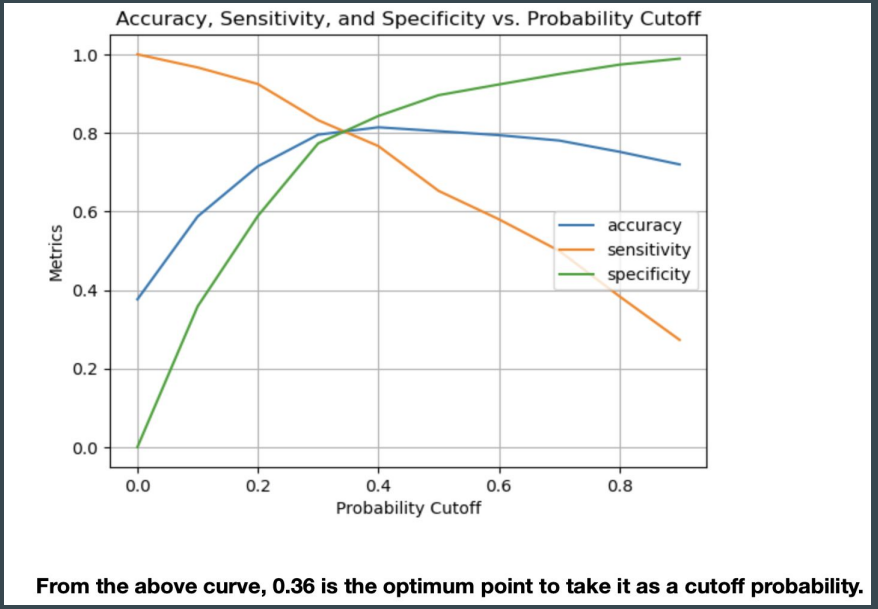
Model Evaluation



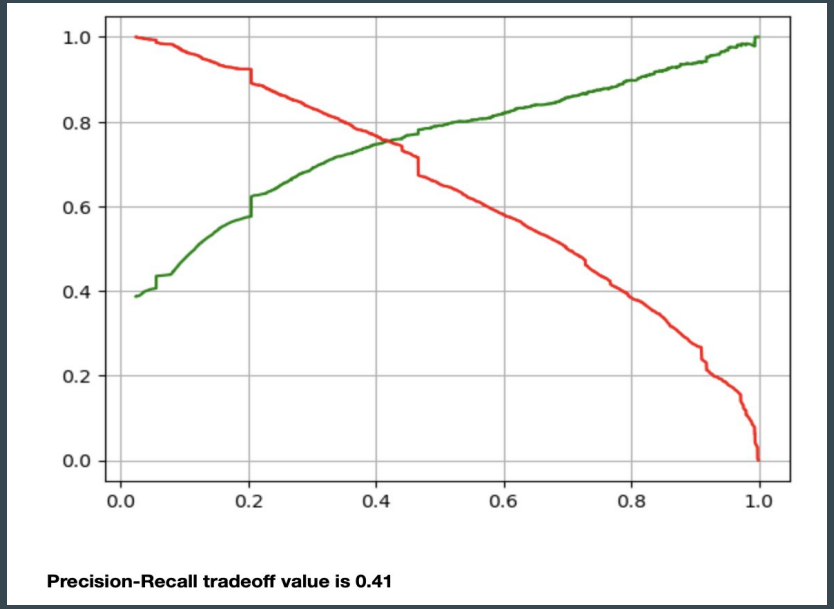
ROC Curve

- An ROC curve demonstrates the tradeoff between sensitivity and specificity.
- If ROC curve closer to the left hand-side and top border of ROC space, then our model would be more accurate.
- If ROC curve closer to the 45-degree diagonal of the ROC space, then our model would be less accurate.
- Next slide will demonstrate how this graph and the optimal threshold value is important for all the evaluation metrics.

Model Evaluation



From the above curve, 0.36 is the optimum point to take it as a cutoff probability.



Precision-Recall tradeoff value is 0.41

Based on above plots and their optimum threshold value, we can evaluate our below metrics.

Train Data

- Accuracy - 80.85
- Sensitivity - 79.21
- Specificity - 81.83

Test Data

- Accuracy - 80.68
- Sensitivity - 79.55
- Specificity - 81.39

Train Data

- Precision - 72.45
- Recall - 79.21

Test Data

- Precision - 71.59
- Recall - 79.55

Top features for building the best model

Lead Source_Welingak Website	6.367897
Lead Source_Reference	3.663536
What is your current occupation_Working Professional	2.707551
Last Activity_Had a Phone Conversation	2.234673
Last Activity_SMS Sent	1.222154
Lead Source_Olark Chat	1.113905
Total Time Spent on Website	1.094917
const	-0.385572
Last Activity_Converted to Lead	-1.063116
Lead Origin_Landing Page Submission	-1.093208
Specialization_Others	-1.120530
Last Activity_Olark Chat Conversation	-1.472820
Last Activity_Email Bounced	-1.968026

dtype: float64

Identifying the paramount features for constructing the optimal model.

Conclusion

Linear Regression Model	• We have created a good model which shows accuracy, Sensitivity, and Specificity comes under 80% of rate.
Target Counts	• There are 479 leads which are having "lead conversion rate" of greater than or equal to 80% after our model predicted correctly.
Top Lead Sources	• Welingak Website, Reference, Working Professional.
Connectivity Mediums	• Website, SMS, Olark Chat.

Recommendations

- Prospective leads demonstrating above-average website engagement durations exhibit a propitious inclination towards conversion, warranting strategic outreach for increased lead conversions.
- The Welingak website emerges as a prime locus for cultivating affirmative conversion prospects, underlining its instrumental role in lead generation.
- Reference leads stand as a commendable reservoir for bolstering lead conversion endeavors, substantiating their intrinsic value in the conversion ecosystem.
- The judicious deployment of SMS messaging and Olark chat functionalities can wield a substantial influence on the augmentation of lead conversions.
- The efficacy of the sales team's direct communication endeavors, particularly via telephone outreach, is underscored by the notably elevated phone conversion rates observed within the ambit of favorable leads.
- Focusing targeting efforts on the cadre of working professionals and unemployed leads, who display a proclivity towards pursuing educational courses, emerges as a promising avenue for fostering positive lead conversions.

The Team

Biplab Mondal

Bibudhendhu Mishra

Indrajit Bose