Lead Scoring Case Study using Logistic Regression

By

ANAND M C

Table Of Contents

Introduction

- X Education is an online education provider targeting industry professionals.
- Their courses attract interest from numerous visitors who land on their website daily.
- Courses are promoted across various platforms, including popular websites and search engines like Google.
- Upon visiting the website, individuals may explore available courses, fill out course forms, or watch instructional videos.
- Visitors who provide contact information through form submissions are identified as leads.
- The sales team then engages with these leads through calls, emails, and other communication channels.
- While some leads convert into paying customers, the majority do not.
- X Education typically achieves a lead conversion rate of around 30%.

Problem Statement & Objective of the Study

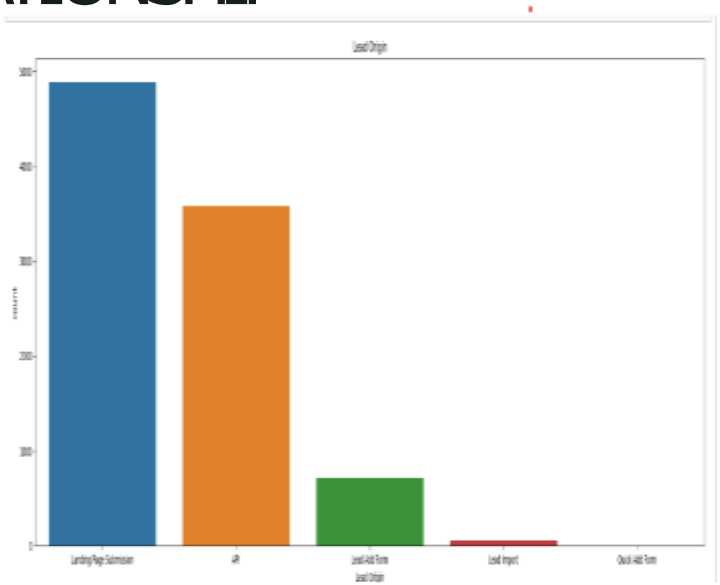
- X Education receives numerous leads but struggles with a low lead conversion rate, currently at approximately 30%.
- To enhance the lead conversion process, X Education aims to identify high-potential leads, often referred to as "Hot Leads."
- The sales team seeks to prioritize communication with these potential leads rather than making calls to every lead indiscriminately.
- X Education seeks to develop a lead scoring model to identify the most promising leads, those with the highest likelihood of converting into paying customers. The model aims to assign lead scores where higher scores indicate a greater chance of conversion, aligning with the company's goal of achieving an 80% lead conversion rate as suggested by the CEO

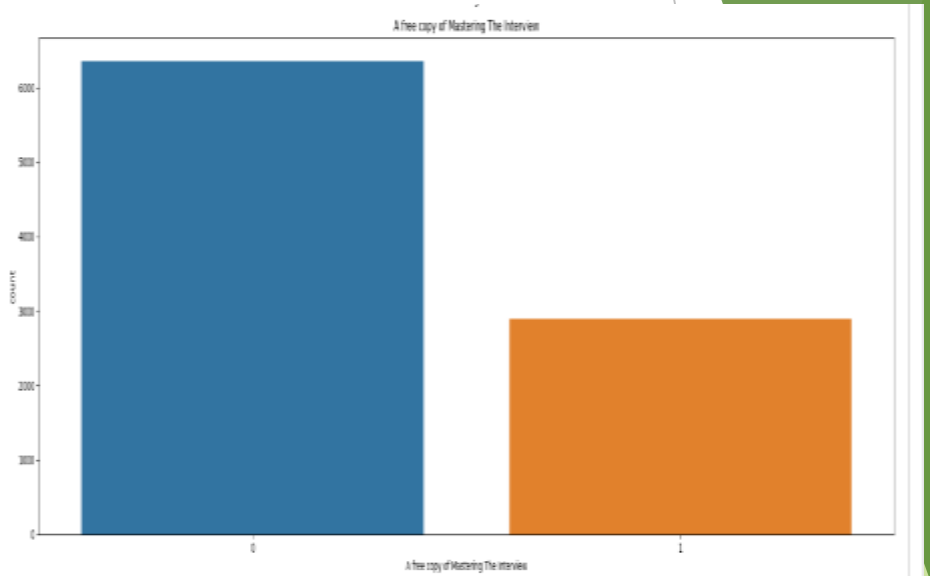
Approach

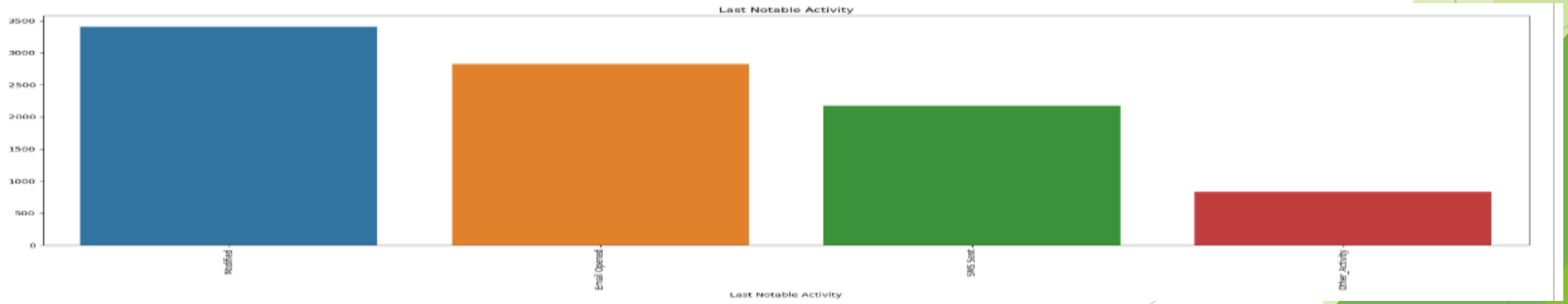
- Understanding the Business Problem and Objective
- Importing and Gaining Insights from the Data
- Data Cleaning and Manipulation
- Addressing Missing Values by Removing Columns with High Missing Value Rates and Dropping Features Not Relevant for Analysis
- Identifying Outliers
- Explore Different Classes of Features using value_counts() for categorical variables.
- Visualize Numerical Variables with a Pair Plot (sns.pairplot()) to examine relationships and distributions.
- Visualize Categorical variables using with a count plot to examine relationships.

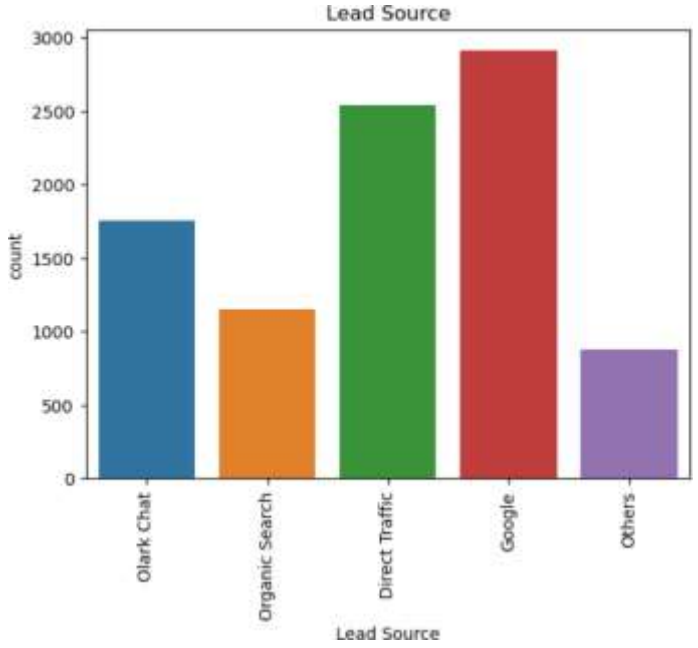
- Perform Univariate and Bivariate Analysis to uncover patterns and relationships between variables.
- Create Dummy Variables (pd.get_dummies()) for categorical features with more than two classes.
- Scale Numerical Features using StandardScaler from sklearn.preprocessing to ensure all features are on a similar scale.
- Apply Logistic Regression (sklearn.linear_model.LogisticRegression) as a classification technique for modeling.
- Use Recursive Feature Elimination (RFE) (sklearn.feature_selection.RFE) to automatically select the top features for modeling.
- Validate the Logistic Regression model and present the results effectively.

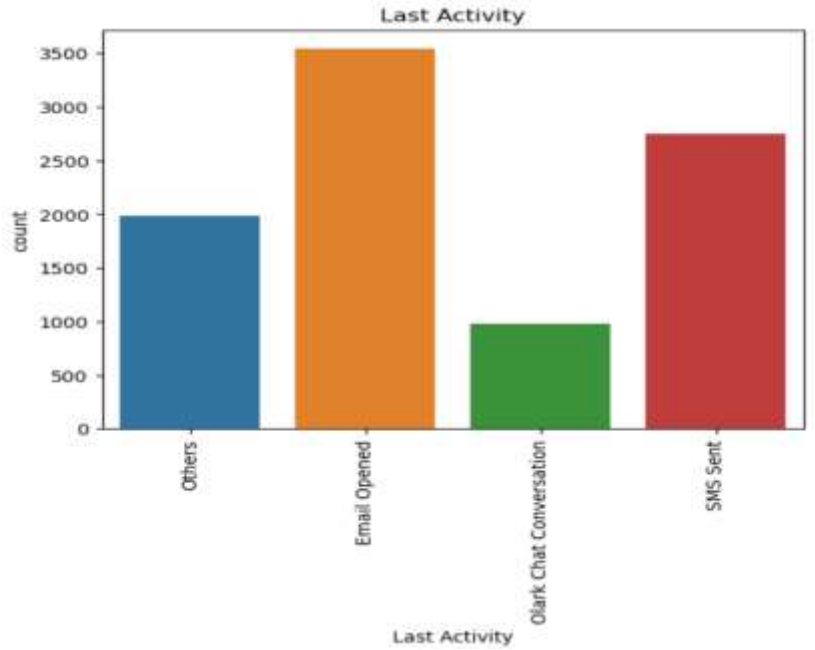
EDA AND CATEGORICAL/NUMERICAL VARIABLE RELATIONSHIP

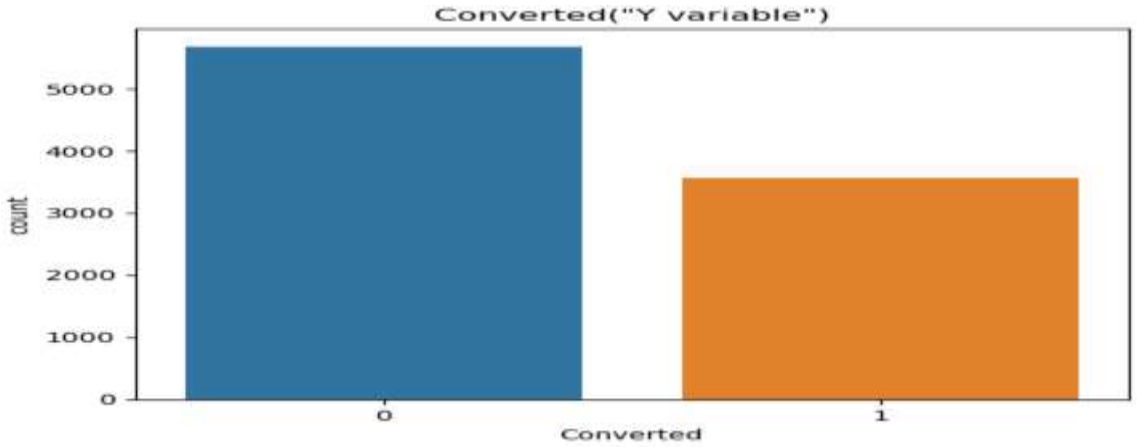






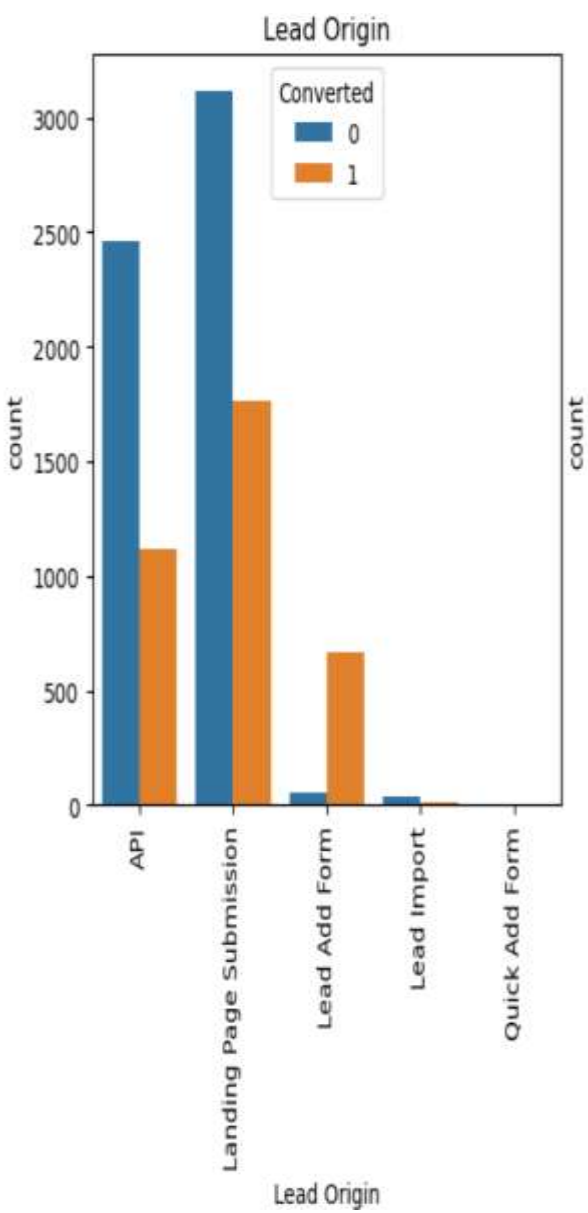

Lead Source
count
3000
2500
2000
1500
1000
500
0
Olark Chat
Organic Search
Direct Traffic
Google
Others
Lead Source

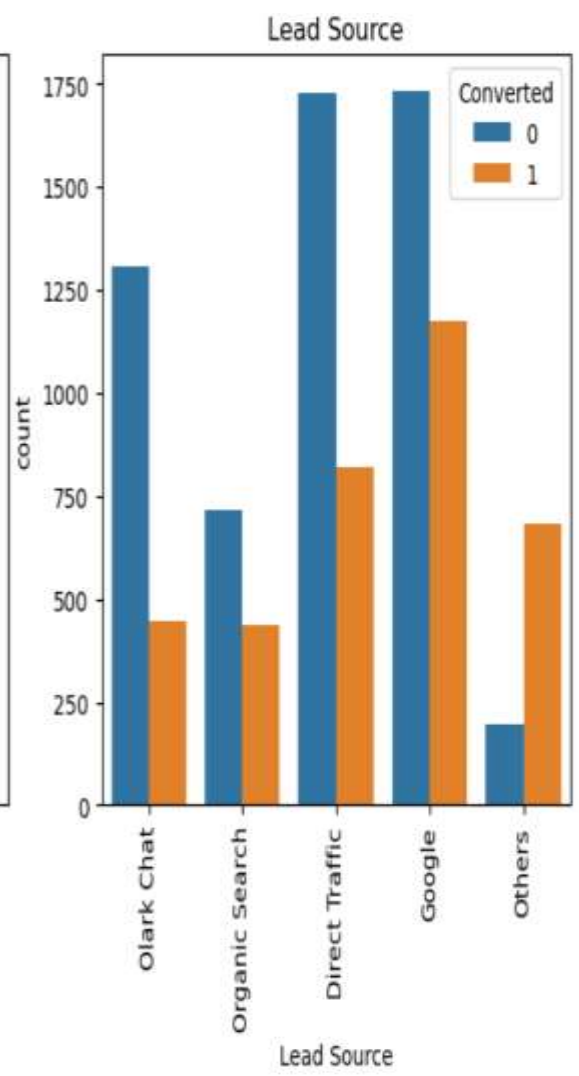

Last Activity
count
3500
3000
2500
2000
1500
1000
500
0
Others
Email Opened
Olark Chat Conversation
SMS Sent
Last Activity

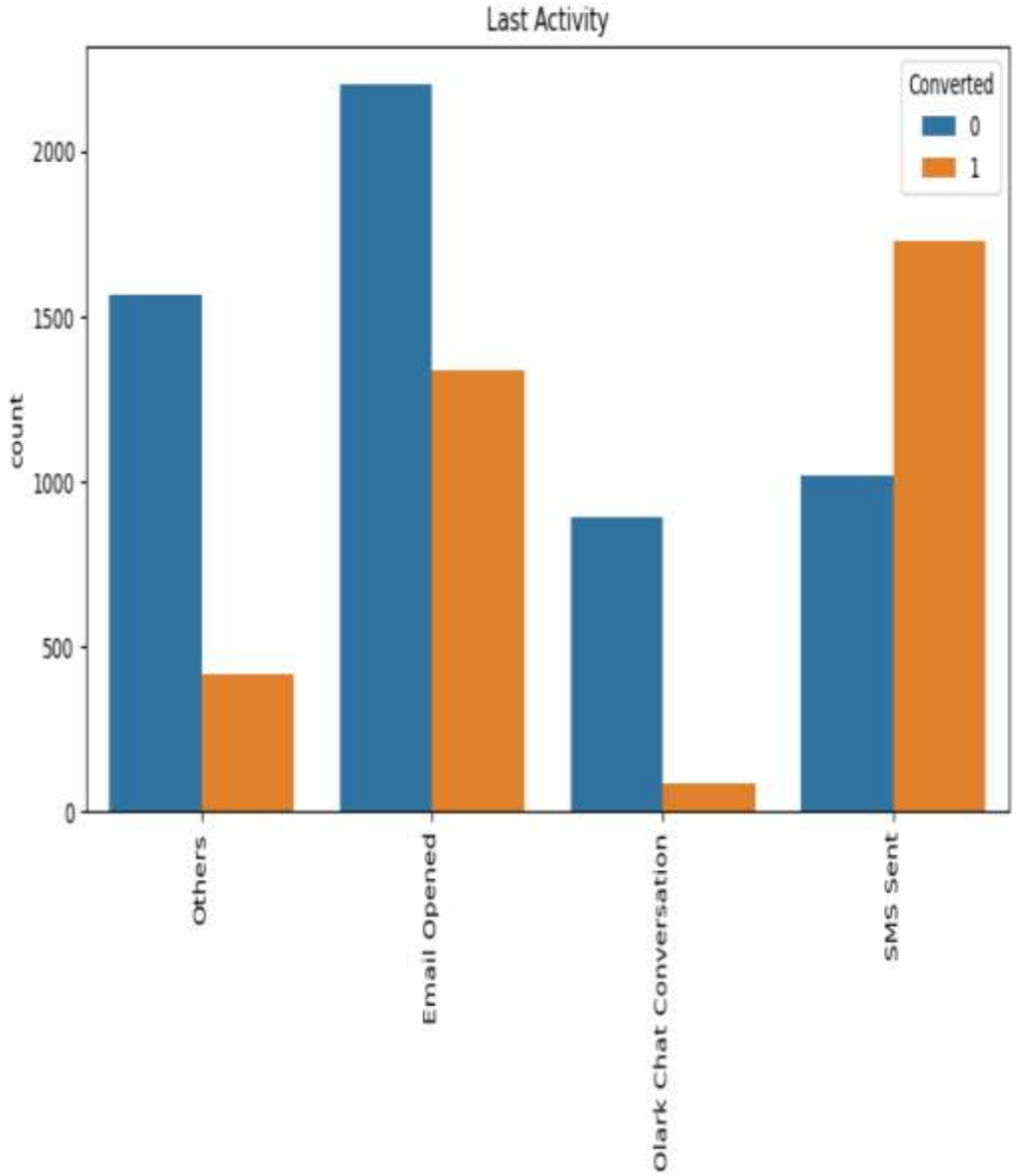

Converted("Y variable")
count
5000
4000
3000
2000
1000
0
0
1
Converted

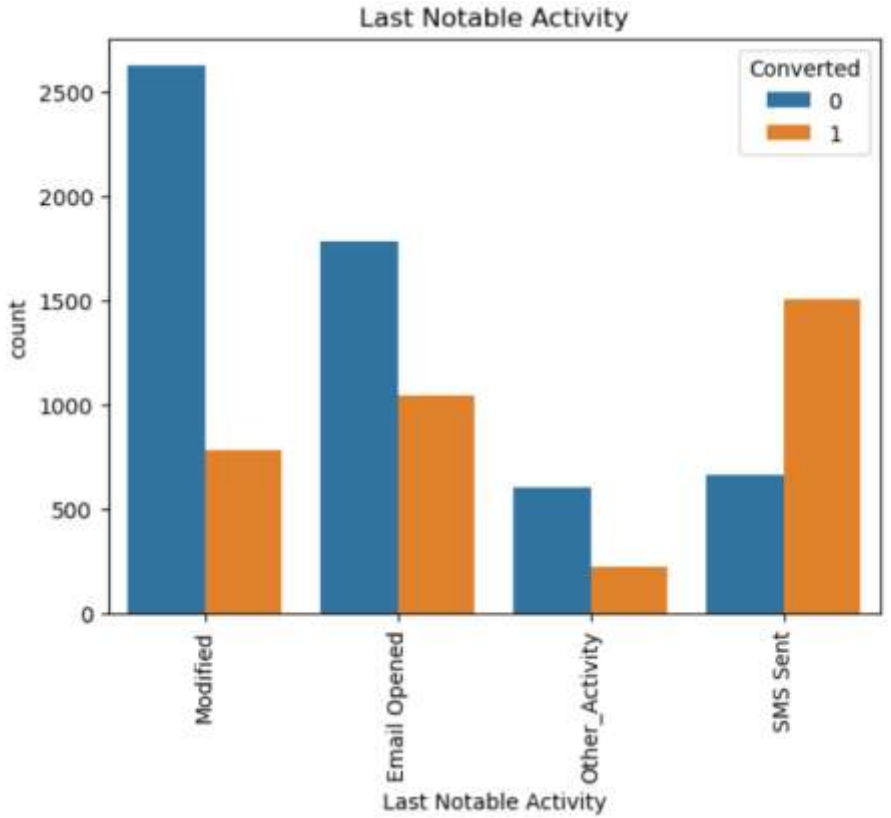
Observation

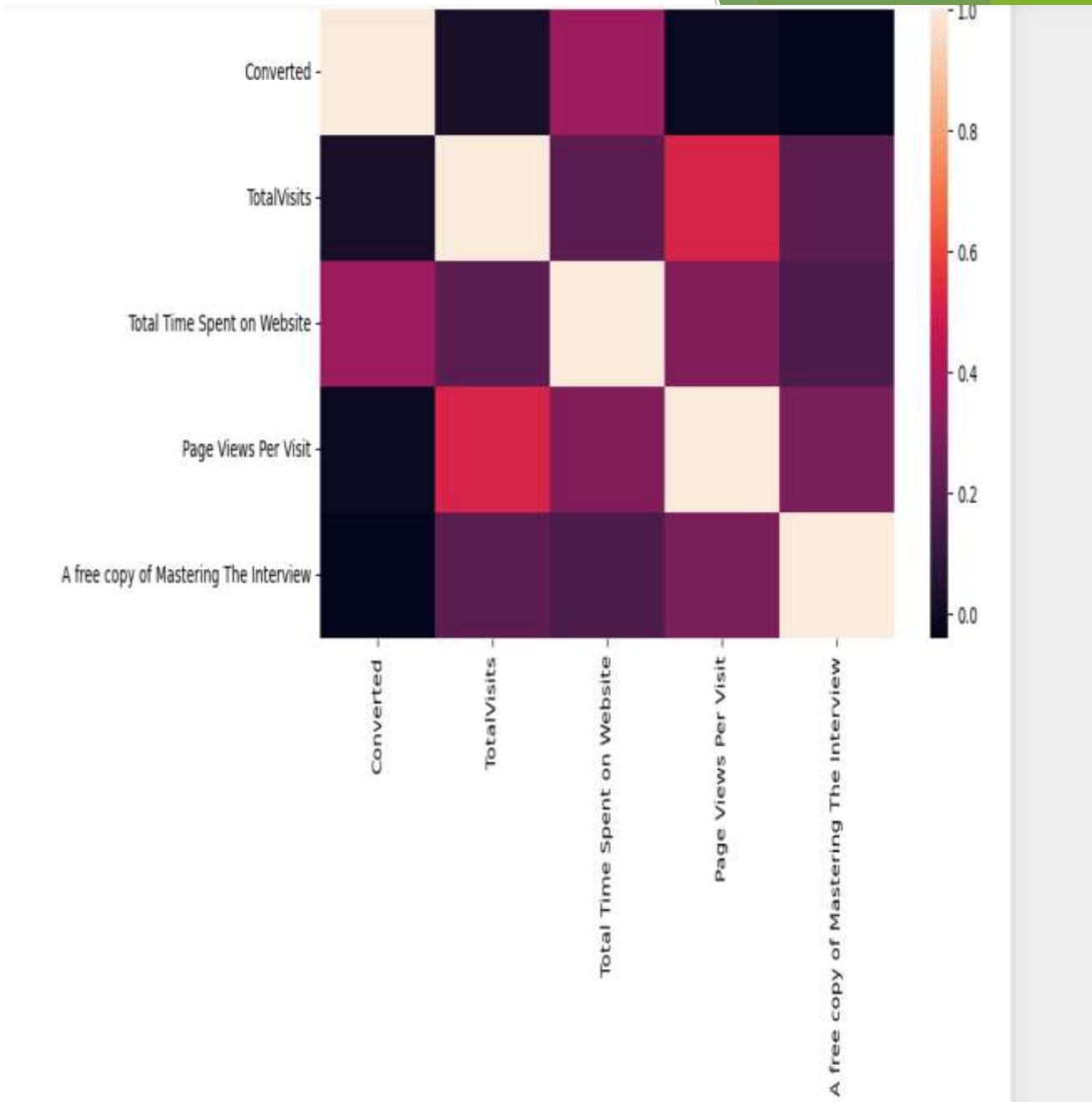
- The most common lead origin is through landing page submission, followed by leads generated from modified and email opened statuses.
- In terms of lead source, Google and direct traffic contribute significantly to the leads obtained.
- The most frequent last activity observed is email opened, followed by SMS sent.
- Typically, leads have 2 to 5 page views per visit, indicating moderate engagement.
- The average time spent on the website is relatively low, and most visitors have fewer than 5 total visits recorded.


Lead Origin
Converted
0
1
count
3000
2500
2000
1500
1000
500
0
API
Landing Page Submission
Lead Add Form
Lead Import
Quick Add Form
Lead Origin


Lead Source
Converted
0
1
count
1750
1500
1250
1000
750
500
250
0
Olark Chat
Organic Search
Direct Traffic
Google
Others
Lead Source


Last Activity
Converted
0
1
count
2000
1500
1000
500
0
Others
Email Opened
Olark Chat Conversation
SMS Sent


Last Notable Activity
Converted
0
1
count
2500
2000
1500
1000
500
0
Modified
Email Opened
Other_Activity
SMS Sent
Last Notable Activity


Converted
TotalVisits
Total Time Spent on Website
Page Views Per Visit
A free copy of Mastering The Interview
1.0
0.8
0.6
0.4
0.2
0.0

Observation

- Lead origin landing pages yield higher value but lower conversion rates.
- Lead ad forms exhibit higher conversion rates.
- Direct traffic as a lead source delivers significant value but lower conversion rates, as does Google.
- Focus is on increasing overall conversion rates.
- Email opened as the last activity shows lower conversion rates compared to SMS sent.
- Heatmap analysis reveals multiple correlated variables.

Data Preparation before Model building

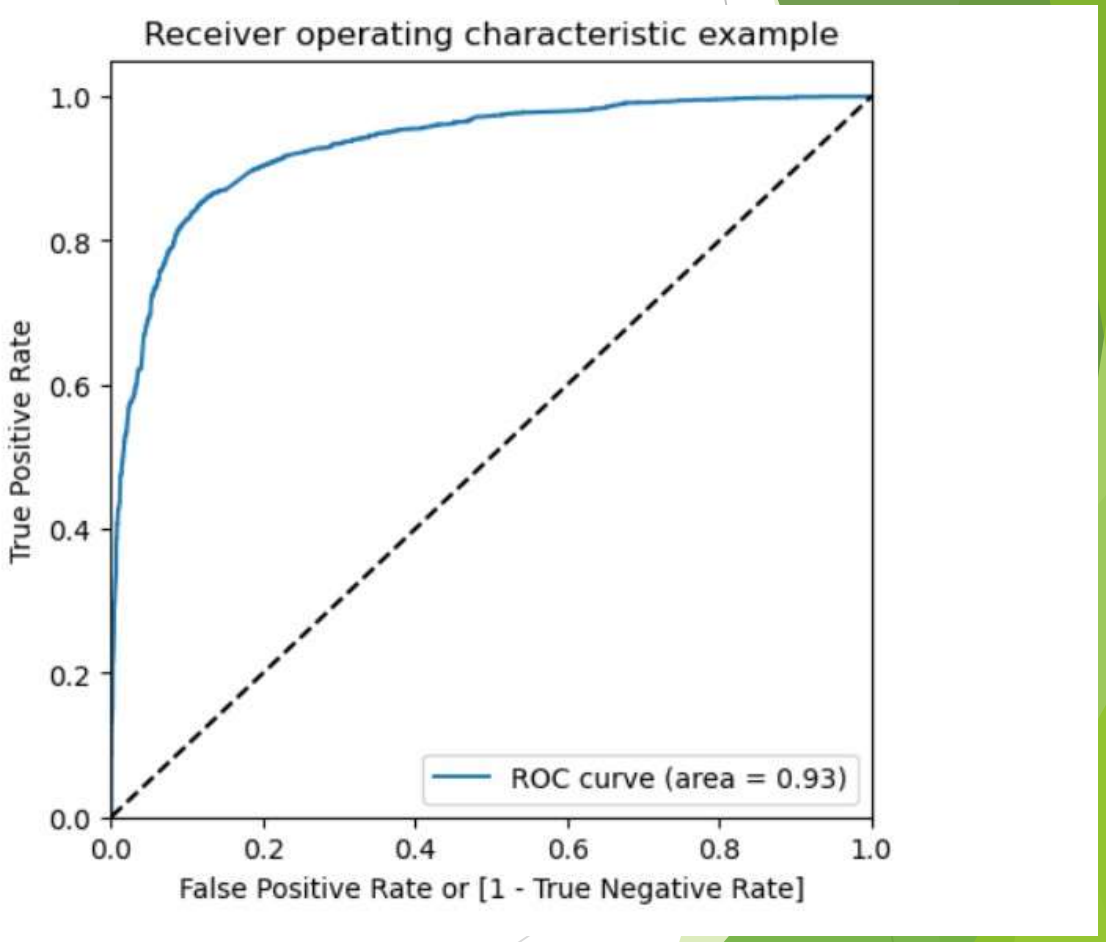
- Split the data into 'X' (independent variables) and 'Y' (dependent variable).
- Further divide 'X' and 'Y' into training and testing datasets using a 70-30 ratio, where 70% is allocated to training and 30% to testing.
- Utilize StandardScaler to scale down non-binary variables to ensure consistent scaling.
- Apply Recursive Feature Elimination (RFE) to select the top 15 features for modeling.
- Initially make predictions on the training data and subsequently on the test data.
- Iteratively refine models to improve accuracy.
- Use list comprehension to create a DataFrame named 'Actual', 'Predicted', and 'Converted Predicted' for analysis.
- Determine a cutoff value of 0.4 to classify prediction results.
- The final model achieves an accuracy score of 0.83 on the test data after multiple iterations, considering multicollinearity and high P-values in the model.

Model Evaluation

ROC Curve - Train Data Set

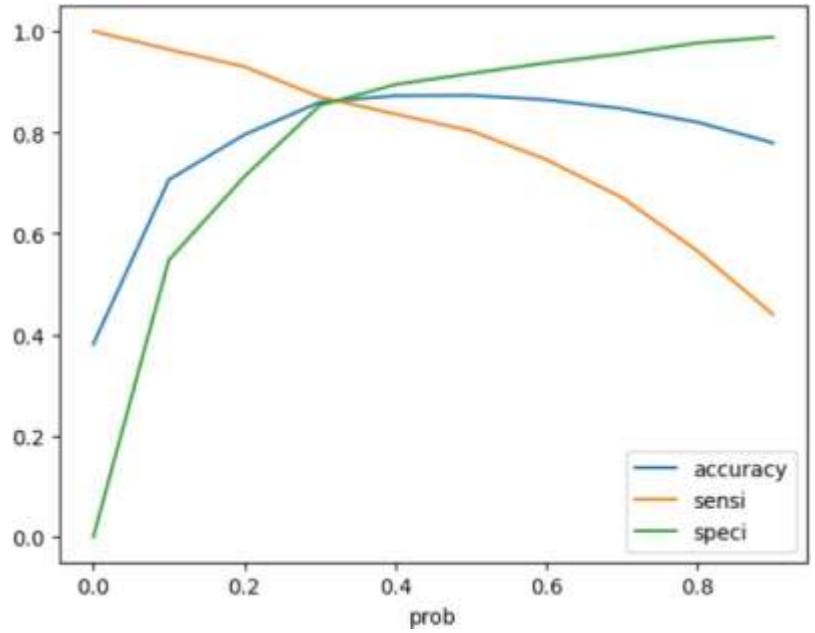
- Area under ROC curve is 0.88 out of 1 which indicates a good predictive model.

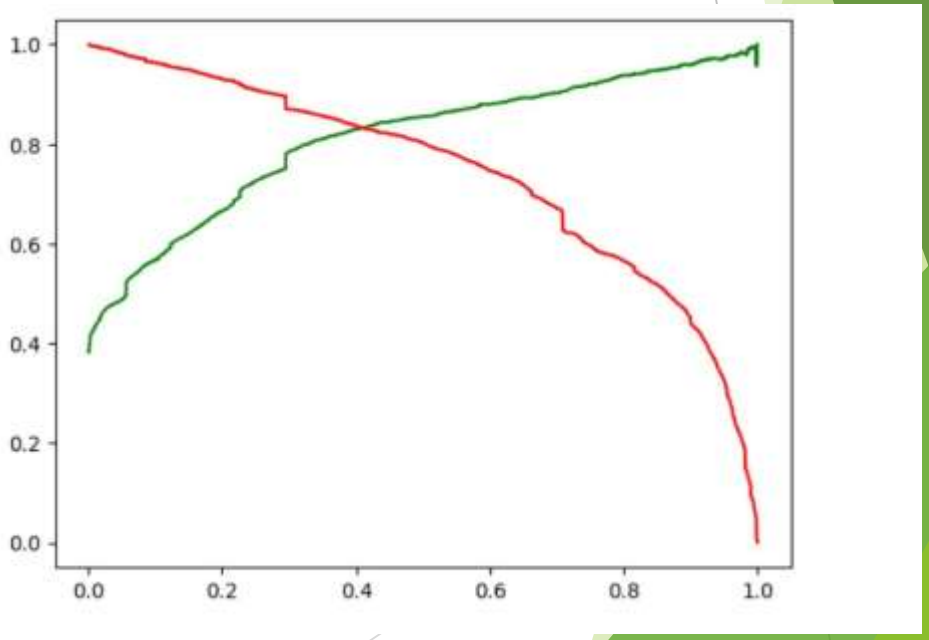
- The curve is as close to the top left corner of the plot, which represents a model that has a high true positive rate and a low false positive rate at all threshold values.



Inference:

1. With the help of below plot, we can select a threshold probability value by having a trade_off between accuracy, Sensitivity and Specificity.
2. It clearly says, we can take Threshold probability **x = 0.3**.





Conclusion

- **Train - Test**

Train Data Set:

- Accuracy: 85%
- Sensitivity: 86%
- Specificity: 85%

Test Data Set:

- Accuracy: 83%
- Sensitivity: 81%
- Specificity: 89%

The evaluation matrices are pretty close to each other so it indicates that the model is performing consistently across different evaluation metrics in both test and train dataset.

Conclusion

- The model attained a sensitivity of 86% on the training set and 81% on the test set.
- In this context, sensitivity represents the percentage of correctly identified converting leads out of all potential converting leads.
- The CEO of X Education established a target sensitivity of approximately 80%.
- Furthermore, the model achieved an accuracy rate of 83%, aligning closely with the study's objectives.

- In accordance with the outlined challenge, enhancing lead conversion stands as a pivotal factor for X Education's growth and prosperity.
- To facilitate this objective, we've constructed a regression model pinpointing the most influential factors affecting lead conversion.
- Our analysis has identified the following features with the highest positive coefficients.
- Accordingly, these features should be prioritized in our marketing and sales endeavors to bolster lead conversion rates.

- Total Time Spent on Website
- Lead Origin_Lead Add Form
- Lead Source_Olark Chat
- Tags_Other Reasons
- Tags_Will revert after reading the email
- Lead Profile_Potential Lead
- Lead Profile_diploma_dual_n_SomeSchool_lead
- Last Notable Activity_Other_Activity
- Last Notable Activity_SMS Sent

- **To enhance our Lead Conversion Rates**:
- Concentrate on features exhibiting positive coefficients to guide targeted marketing strategies effectively.
- Formulate tactics aimed at attracting high-quality leads from the most effective lead sources.
- Refine communication channels to maximize impact based on lead engagement metrics.
- Tailor messaging to resonate with working professionals effectively.
- Increase advertising investment specifically on the Welingak Website.
- Implement referral programs with incentives or discounts to encourage more referrals.
- Focus marketing efforts aggressively towards working professionals due to their higher conversion rates and potentially stronger financial capacities.

Thank you

THANKS