Lead Scoring Case Study

Submitted by : Rishupriya Srivastava

Contents

Problem Statement

Challenge:	• X Education faces a substantial gap in lead conversion despite a high volume of generated leads.
Current Conversion Rate:	• Only 30% of acquired leads successfully convert into paying customers.
Inefficiency Concern:	• Resource-intensive efforts on all leads without targeted focus on potential conversions.
Objective:	• Increase lead conversion efficiency by identifying and prioritizing 'Hot Leads.'
Lead Scoring Model Requirement:	• Develop a predictive model assigning lead scores based on conversion likelihood.
CEO's Target:	• Aim for a target lead conversion rate of around 80%.
Expected Impact:	• Improve lead conversion rates.

Objective

- Develop a Lead Scoring Model:
- Clearly define the criteria that contribute to a lead being considered promising.
- Understand the CEO's target lead conversion rate of approximately 80%.
- Collect and preprocess relevant data on leads, ensuring that it is clean, accurate, and comprehensive.
- Split the dataset into training and testing sets.
- Iteratively fine-tune the model to enhance its predictive accuracy.
- Ensure that the lead scoring model is interpretable and can provide explanations for the assigned scores.
- Implement a system to monitor the model's performance over time.
- Regularly communicate the progress and results of the lead scoring model to the CEO and other relevant stakeholders.
- Document the entire process, from data collection to model development and deployment.

Problem Approach

- **Data Acquisition and Inspection:** Importing data and thoroughly inspecting the data frame for initial insights.

Data Preparation: Preparing the data for analysis, handling missing values, and ensuring data integrity.

Exploratory Data Analysis (EDA): Conducting EDA to gain a deeper understanding of data patterns and characteristics.

Dummy Variable Creation: Creating dummy variables to effectively represent categorical data.

Test-Train Data Split: Dividing the dataset into training and testing sets for model development and evaluation.

Feature Scaling: Standardizing or normalizing features to ensure consistent scale across variables.

Correlation Analysis: Examining correlations between variables to identify relationships and potential insights.

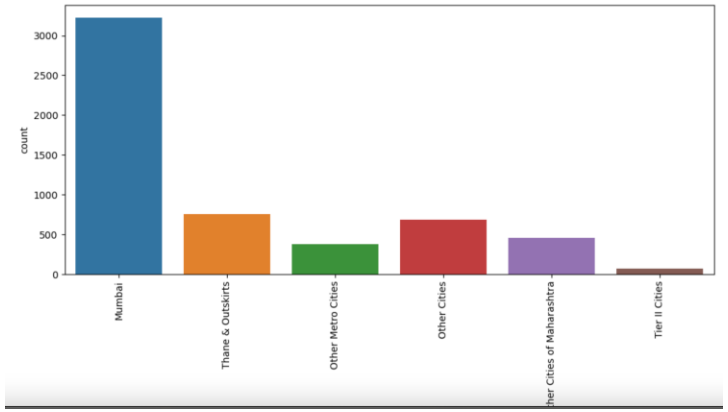
Model Building: Using Recursive Feature Elimination (RFE), Rsquared, Variance Inflation Factor (VIF), and p-values for optimal feature selection.

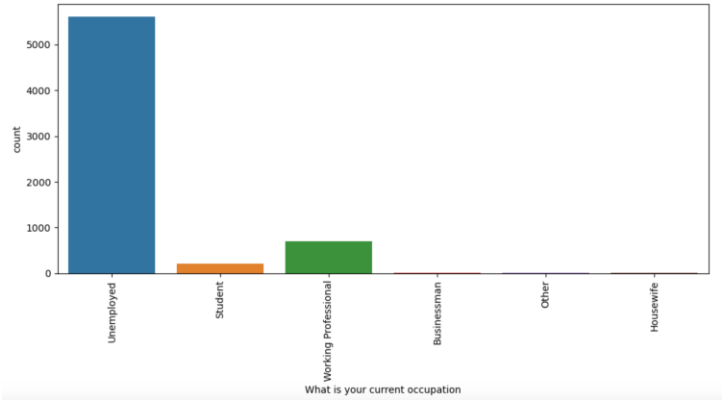
Model Evaluation: Assessing the model's performance using various metrics and validation techniques.

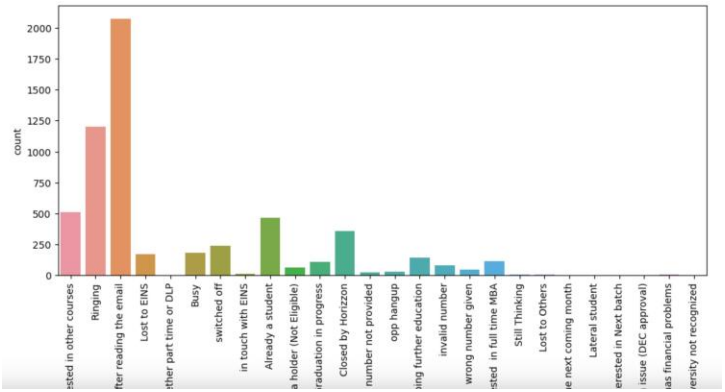
Predictions on Test Set: Applying the trained model to make predictions on the test set for performance validation.

EDA – Data Cleaning

- **Data Cleaning:** Replacing "Select" values in the feature column with null, as it indicates that the lead did not select any option.
- **Dropping Columns:** Removing columns with more than 40% missing values.
- **Removing Unwanted Columns:** Further elimination of unnecessary columns.
- **Checking for Duplicates:** Ensuring there are no duplicate entries in the dataset.
- **Checking for Categorical Data:** Identifying and handling categorical data appropriately.
- **Checking for Highly Skewed Data:** Assessing the skewness of data distribution in columns.
- **Grouping Low-Frequency Values:** Grouping values with low frequencies to prevent sparsity.
- **Checking for Outliers:** Identifying and managing outliers in the dataset.



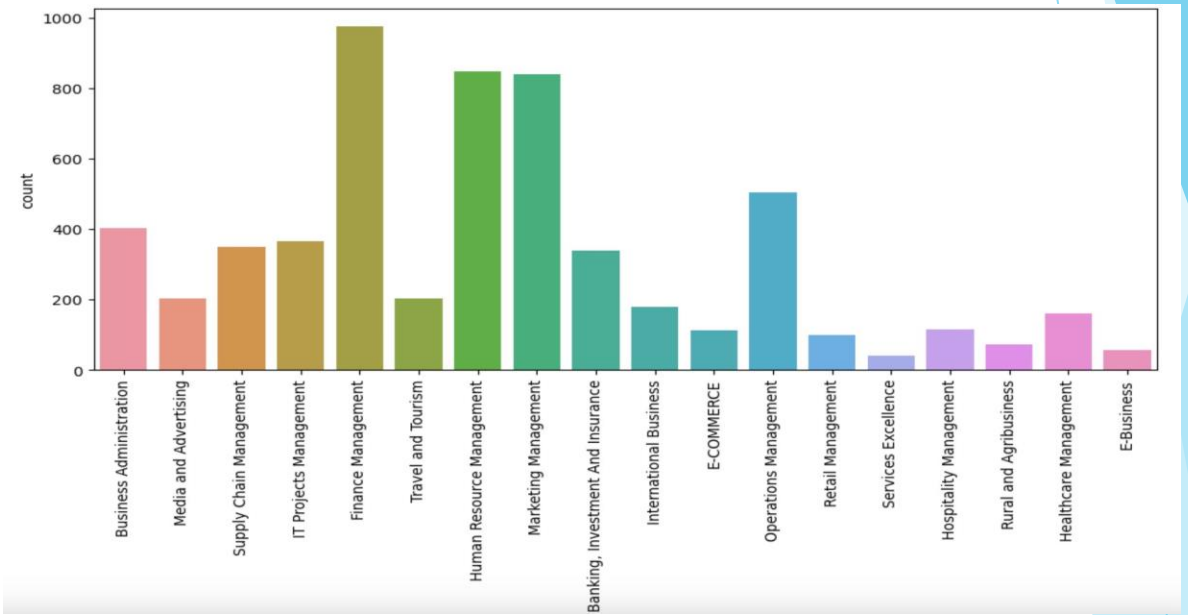




EDA – Reference Graphs for Data Cleaning

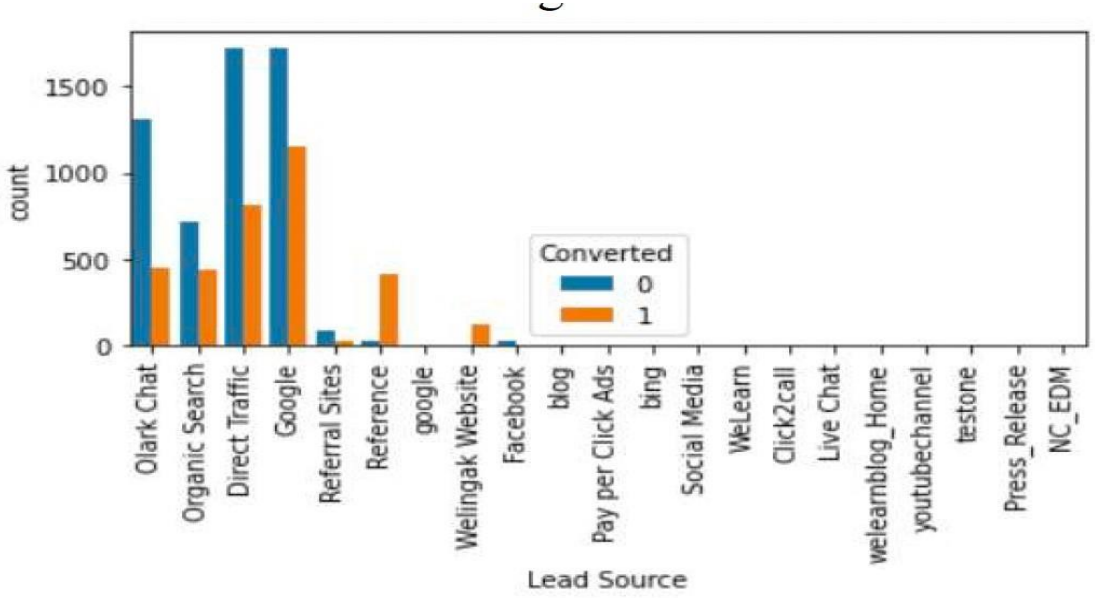
Specialization

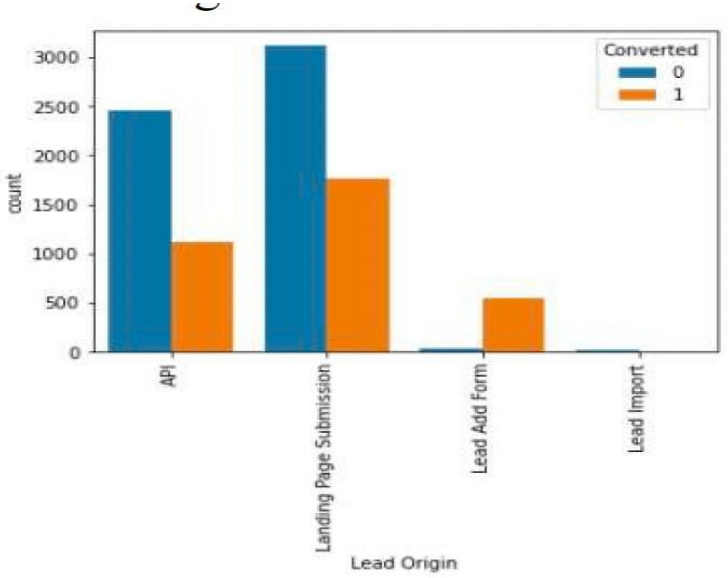
- Leads specializing in HR, Finance, and Marketing have a higher probability of conversion.



Lead Source and Lead Origin

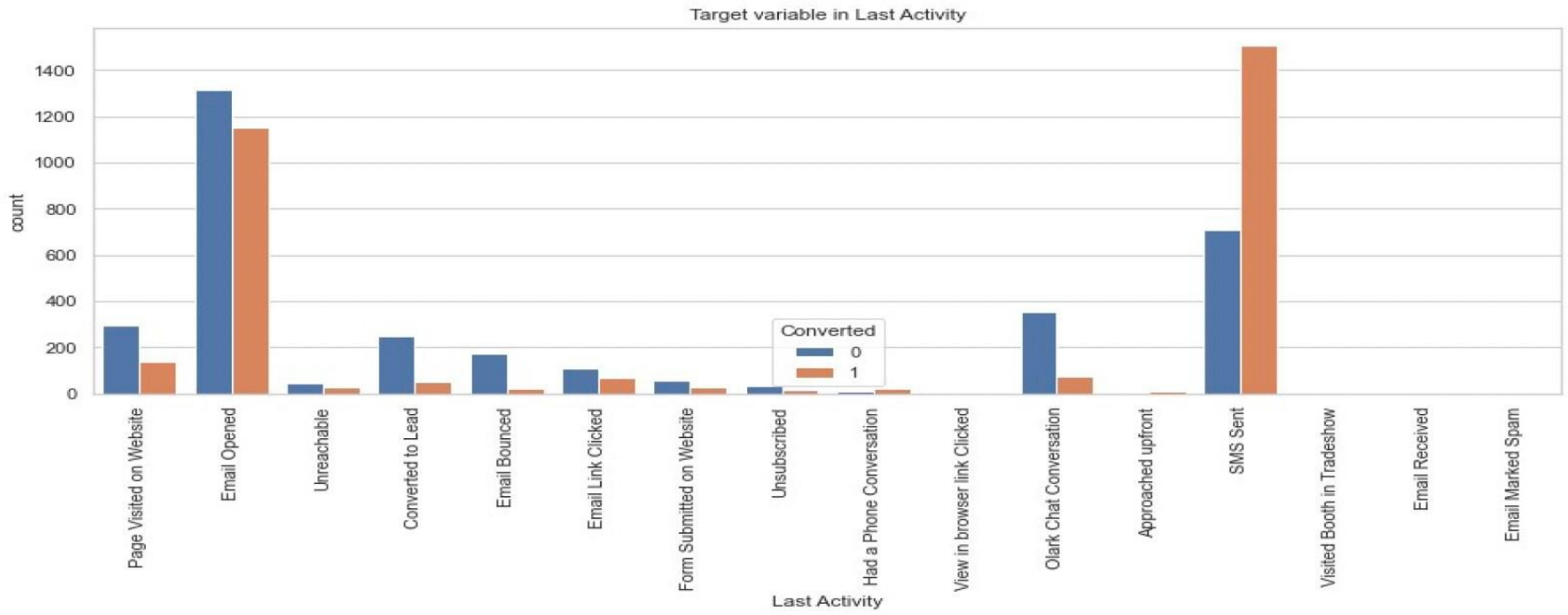
- Lead source clearly shows Google and Direct Traffic has high probability for conversion
- Whereas in Lead Origin most number of leads are landing on Submission.





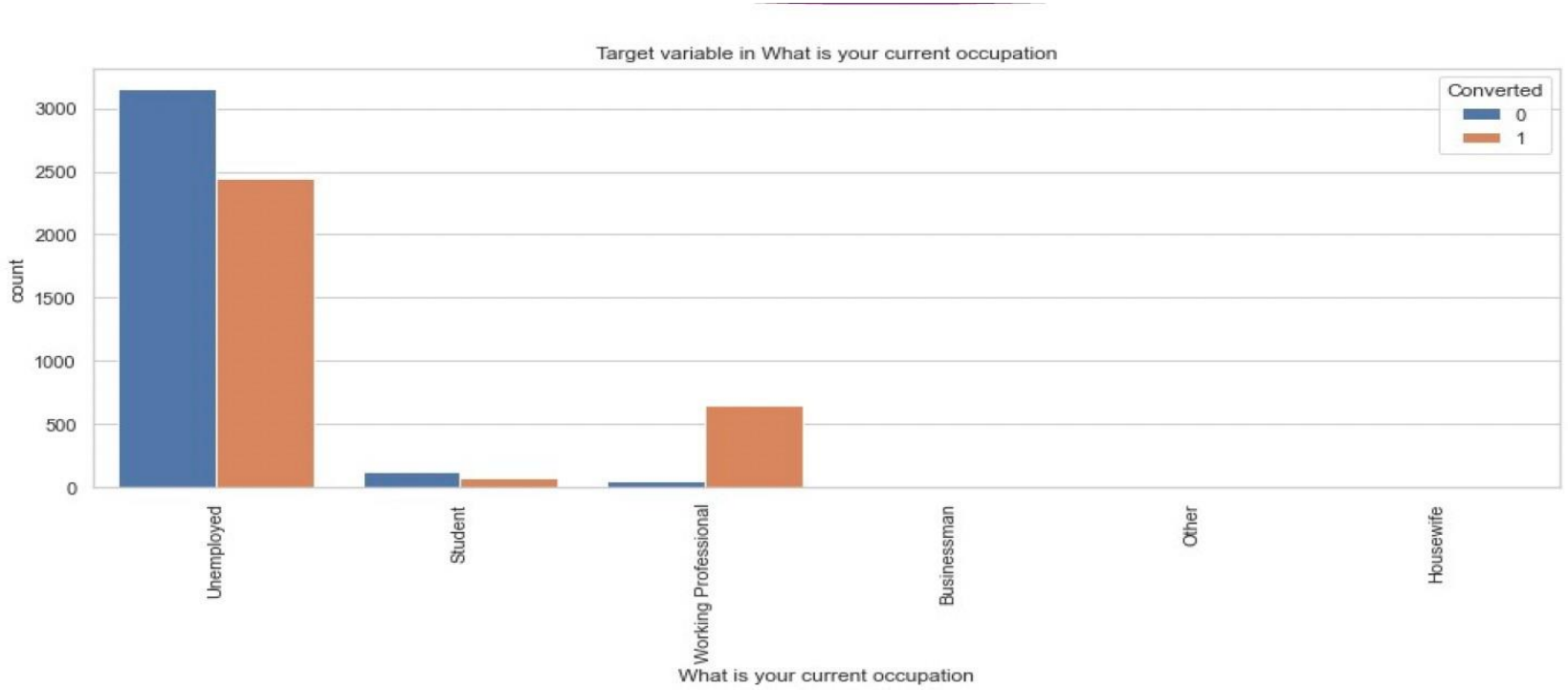
Last Lead Activity

- Leads which are opening email have high probability to convert, same as sending SMS would also benefit.



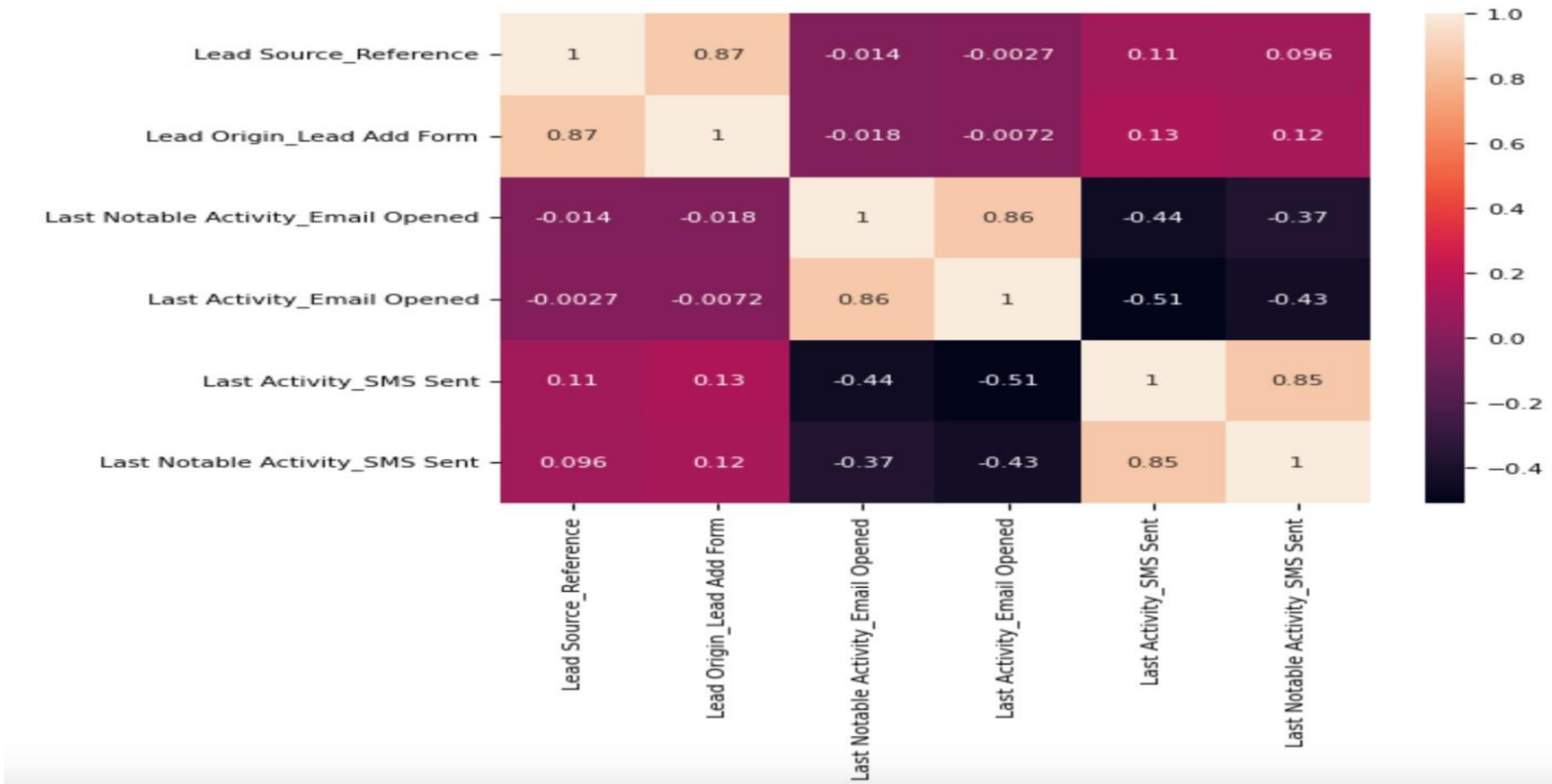
Occupation

- Unemployed individuals show a higher level of interest in joining the course, making them more likely to become leads.

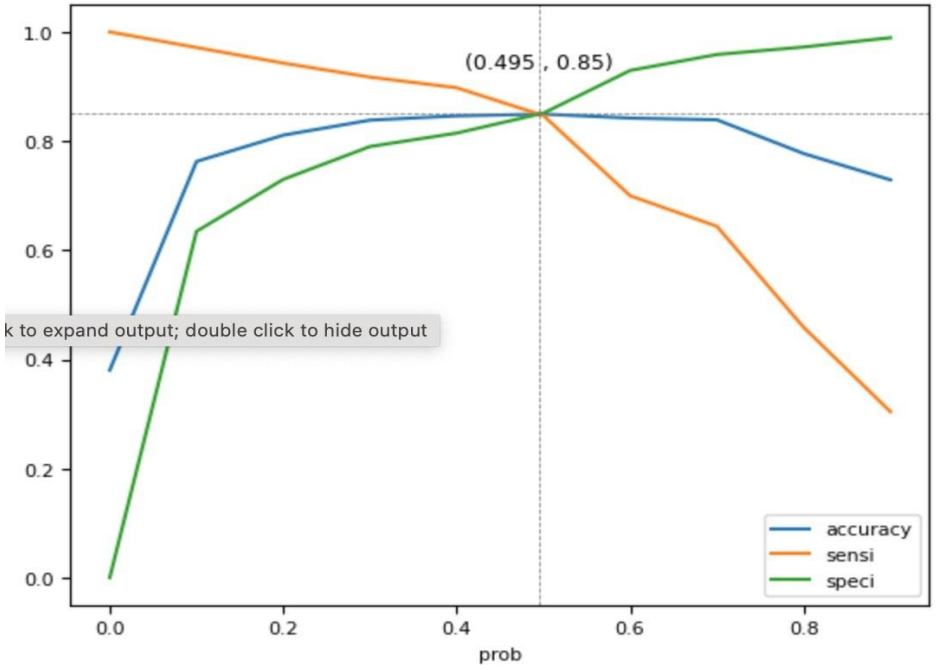


Correlation

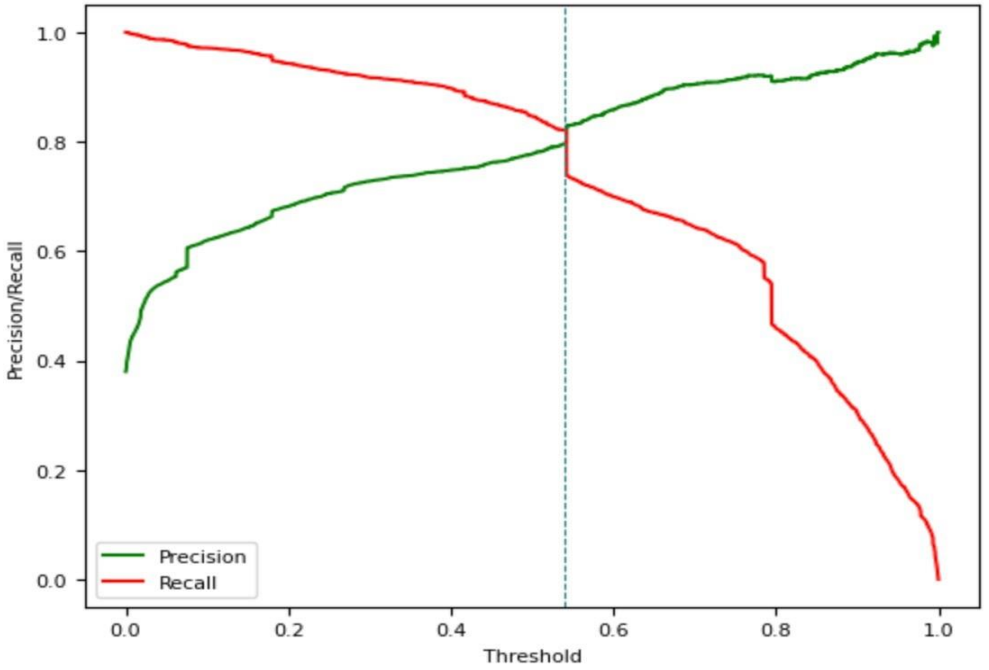
- There does not seem to be correlation



Model Evaluation



- The point 0.495 is approx point where all the 3 curve meet.
- So the 0.495 seems to be `Optimal cutoff point` for probability threshold.



The above graph shows the trade-off between Precision and Recall.

Observation

- Train Data :
  - Accuracy – 85.01
  - Sensitivity – 85.46
  - Specificity – 84.73
- Test Data :
  - Accuracy – 85.01
  - Sensitivity – 85.46
  - Specificity – 84.73

Conclusion

- **Consistent Performance:**
  - Evaluation metrics indicate consistent model performance in both training and testing datasets.
- **High Sensitivity:**
  - The model achieves a commendable sensitivity of 85.46% with a cut-off value of 0.49.
  - This indicates the model's ability to accurately identify potential leads that convert.
- **Exceeding CEO's Target:**
  - The model surpasses the CEO's target sensitivity of around 80%.
  - This showcases the effectiveness of the model in meeting key performance goals.
- **High Accuracy:**
  - The model demonstrates an accuracy rate of 85.01%.
  - This aligns closely with the study's objectives, affirming the model's reliability in predicting lead conversions.
- **Robust Performance Summary:**
  - In summary, the model exhibits robust performance, surpassing sensitivity targets and achieving high accuracy.
  - This instills confidence in the model's utility for lead identification at X Education.