Lead Scoring Case Study

Submitted by:

Pranav Sakhala

Vasu Patrot

Rishikesh 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

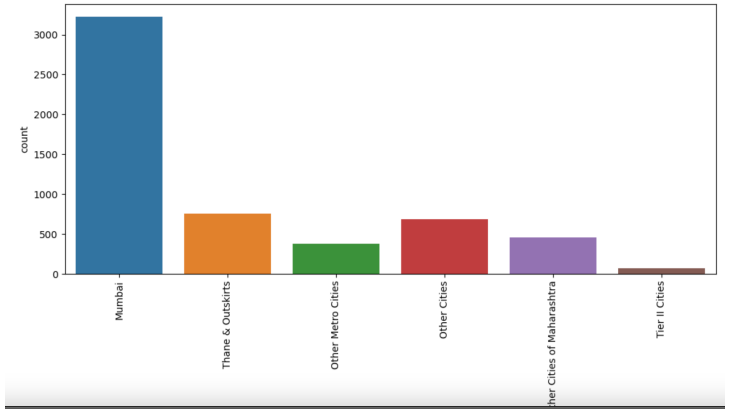
- Develop a Lead Scoring Model:
- Clearly define the criteria that contribute to a lead being considered promising.
- Understand the CEO's target lead conversion rate of approximately 80%.
- Collect and preprocess relevant data on leads, ensuring that it is clean, accurate, and comprehensive.
- Split 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.

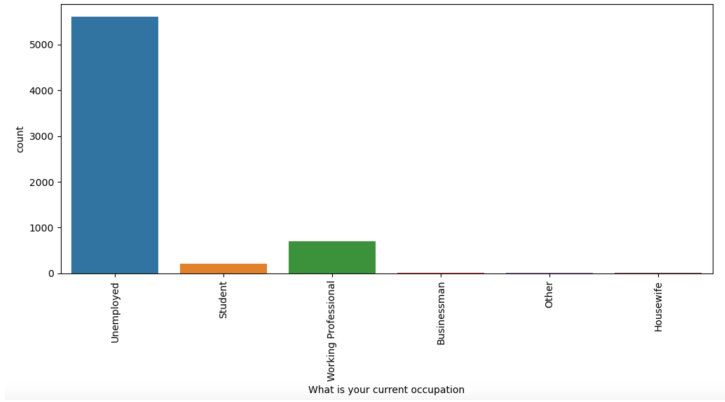
Problem Approach

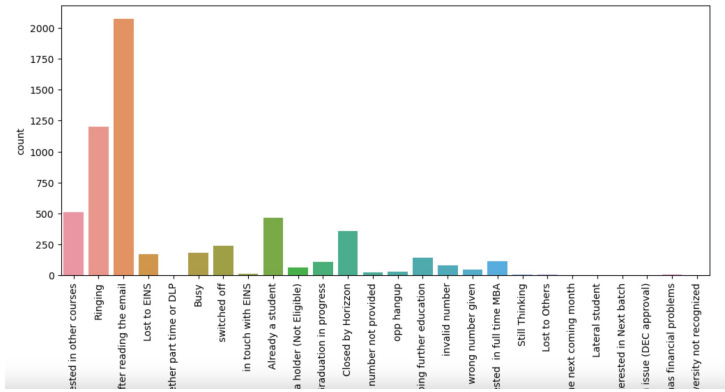
- **Data Acquisition and Inspection:** Importing data and thoroughly examining the data frame for insights.
- **Data Preparation:** Preparing the data for analysis, addressing missing values and ensuring data integrity.
- **Exploratory Data Analysis (EDA):**Conducting EDA to gain a deeper understanding of the data's patterns and characteristics.
- **Dummy Variable Creation:** Creating dummy variables to represent categorical data effectively.
- **Test-Train Data Split:** Splitting 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:** Utilizing Recursive Feature Elimination (RFE), Rsquared, Variance Inflation Factor (VIF), and p-values for optimal feature selection.
- **Model Evaluation:** Assessing the model's performance through 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 :** As there are various values in feature column as Select which is same as null because lead had not selected any option over there so we can replace this with null value.
- **Dropping the Columns** : Columns with values missing > 40% are dropped
- Removing more unwanted Columns
- Checking for Duplicates
- Checking for Categorical data
- Checking for Columns with Highly Skewed Data
- Grouping Low Frequency Values
- Checking for Outliers



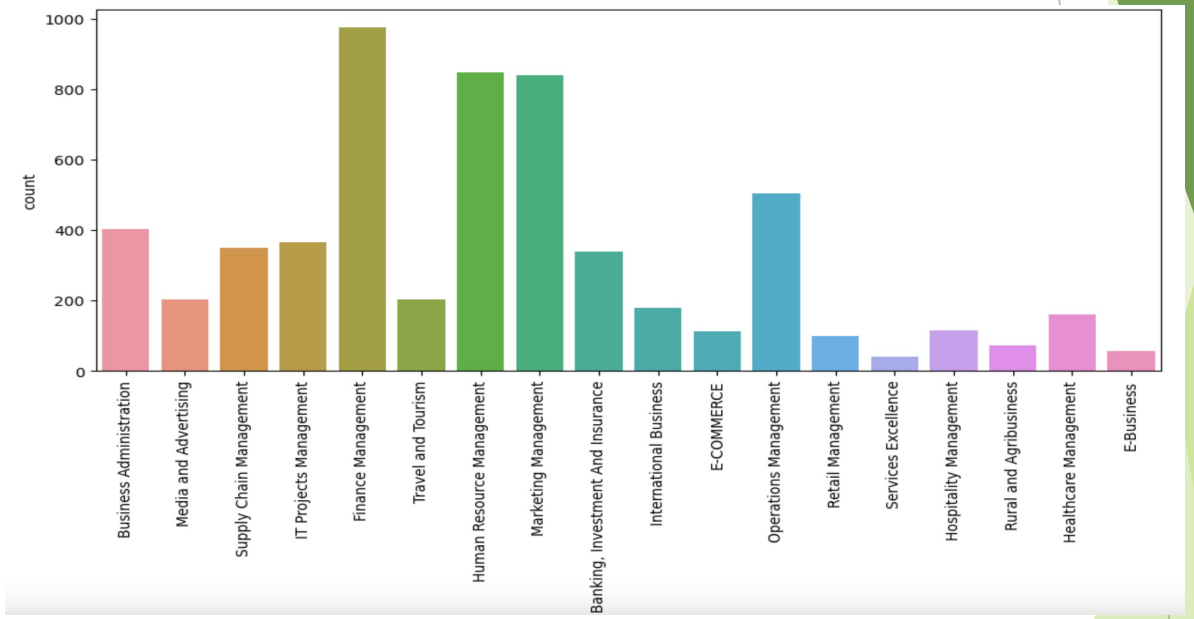




EDA – Reference Graphs for Data Cleanin

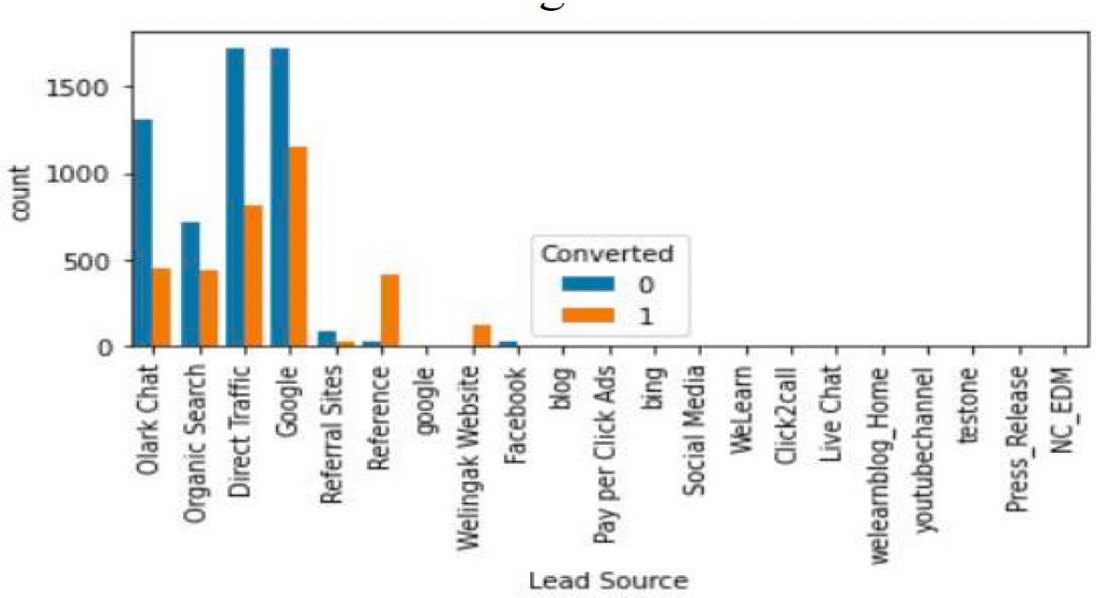
Specialization

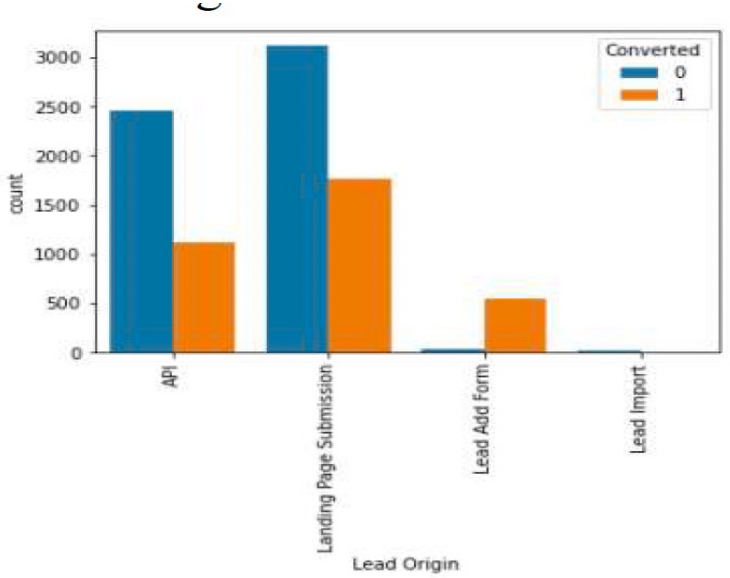
- Leads from HR Finance and Marketing spcilizations have higher probability for conversion



Lead Source and Lead Origin

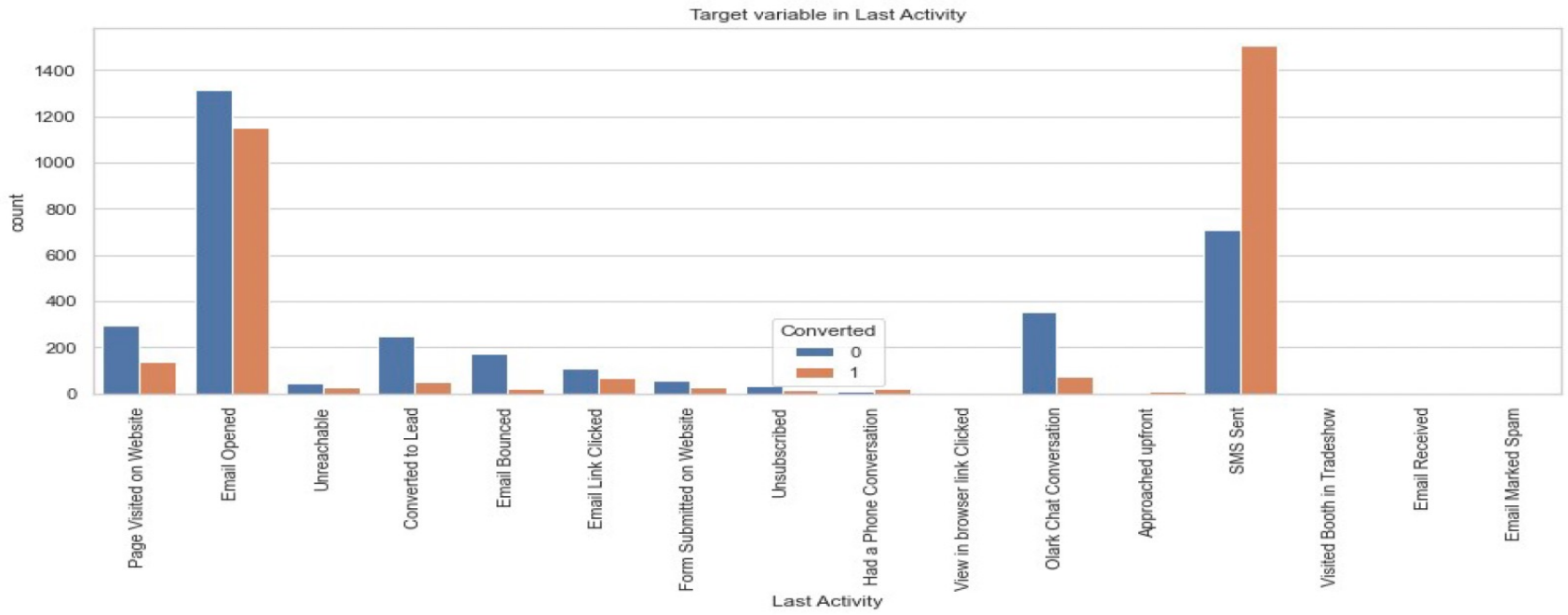
- Lead source clearly shows Google and Direct Traffic has high probability for conversion
- Whereas in Lead Origin most number of leads are landing on Submission.





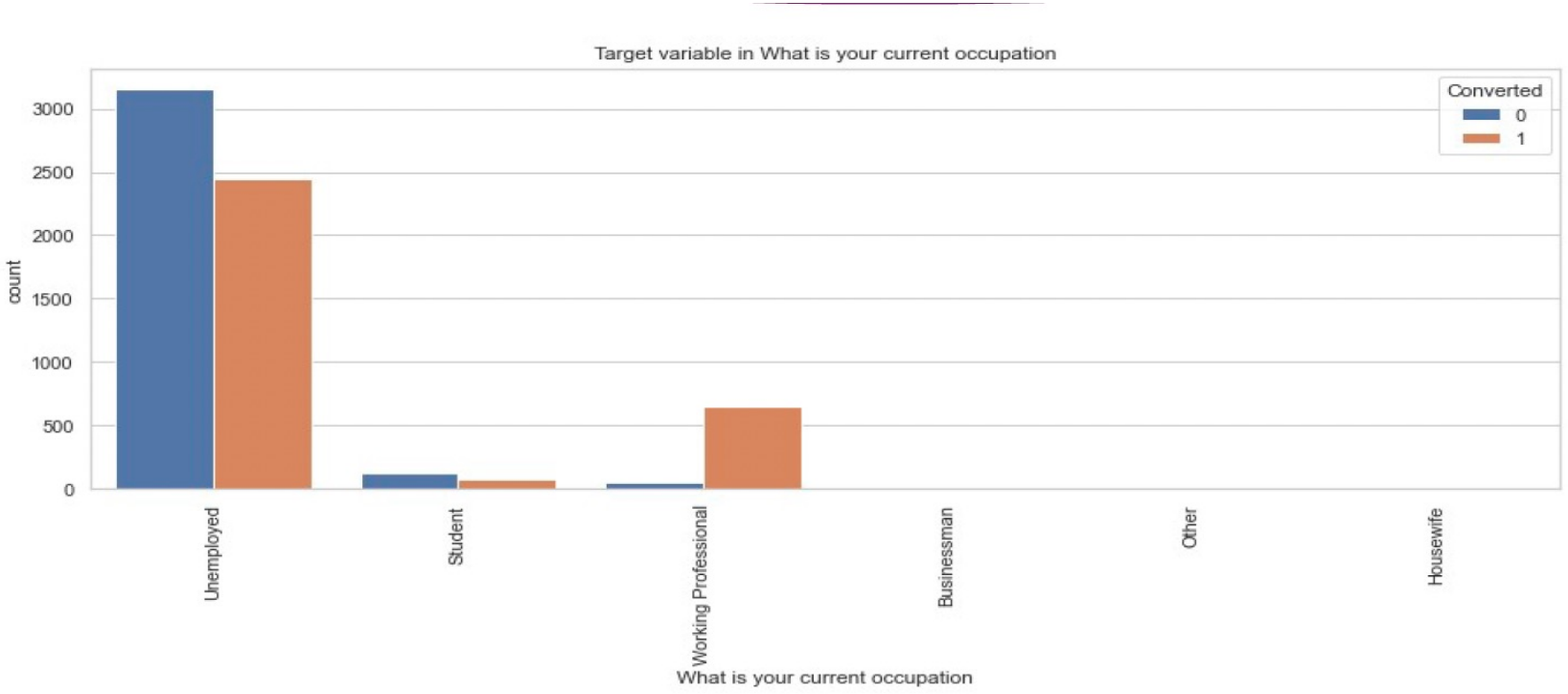
Last Lead Activity

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



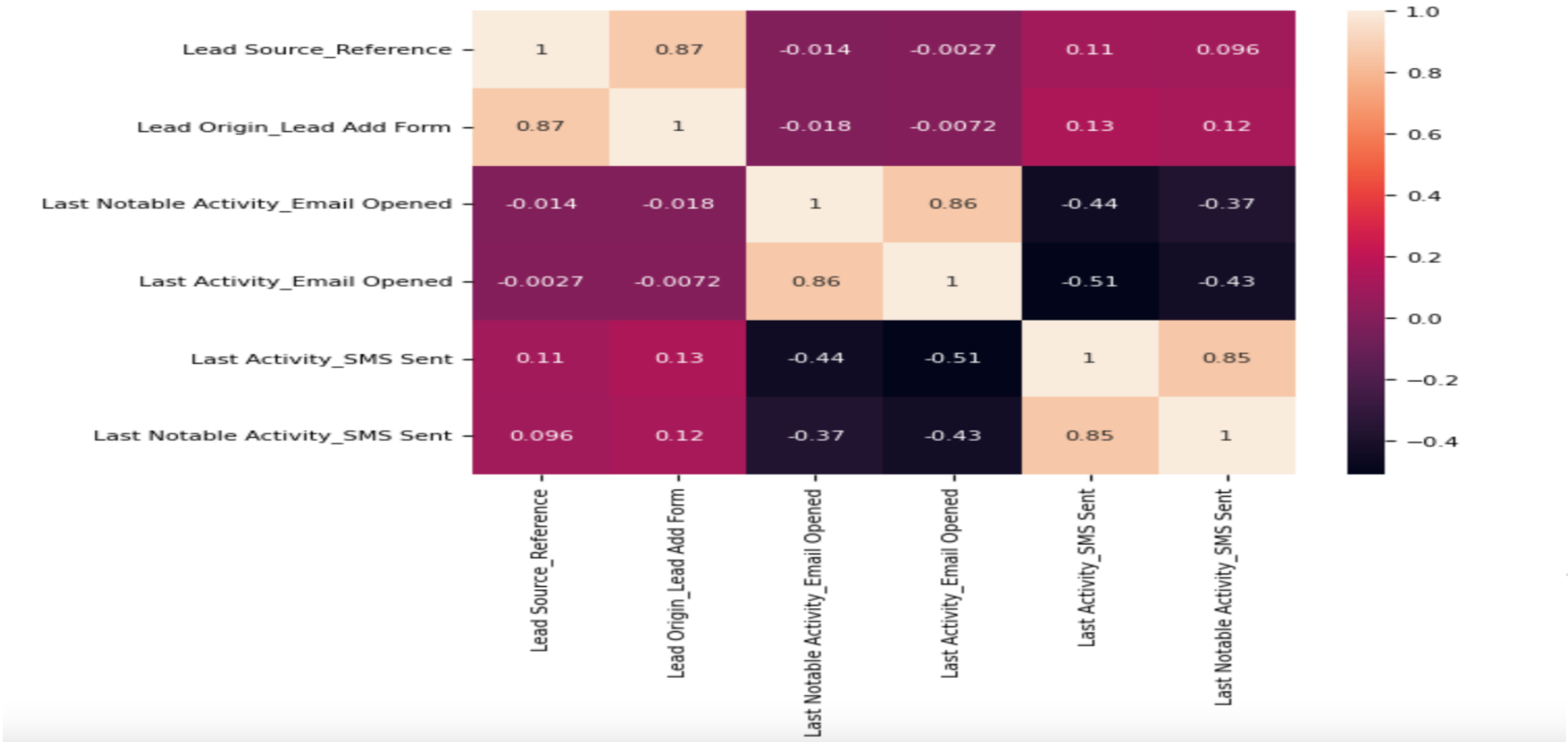
Occupation

- Leads which are unemployed are more interested to join the course.

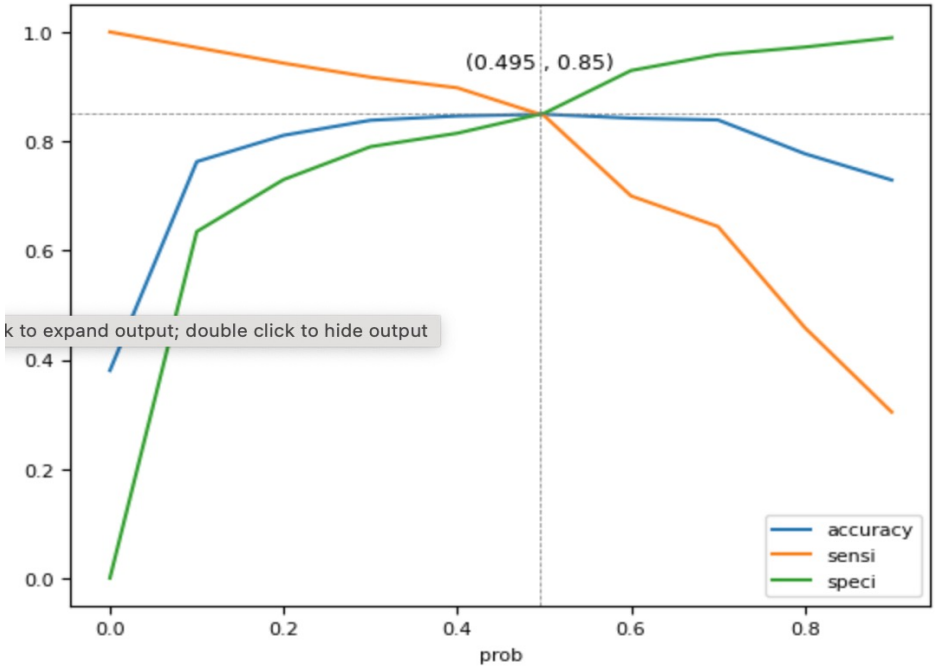


Correlation

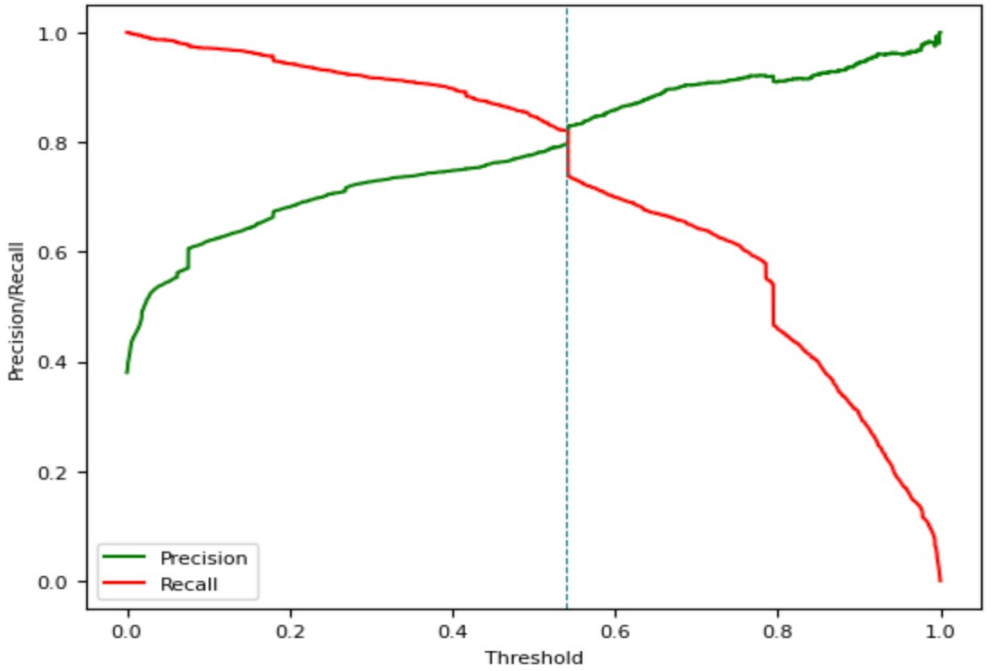
- There does not seem to be correlation



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

- The evaluation metrics reveal a noteworthy consistency in the model's performance across various measures in both the training and testing datasets.
- The model demonstrated a commendable sensitivity of 85.46%, using a cut-off value of 0.49, indicating its capability to accurately identify potential leads that convert.
- The CEO's target sensitivity of around 80% has been surpassed, showcasing the model's effectiveness in meeting key performance goals.
- With an accuracy rate of 85.01%, the model aligns closely with the objectives set for the study, affirming its reliability in predicting lead conversions.
- In summary, the model exhibits robust performance, exceeding specified sensitivity targets and achieving high accuracy, instilling confidence in its utility for lead identification at X Education.