Lead Scoring Case Study

SUBMITTED BY :

1. **Alok Kumar**
2. **Varad Wavare**
3. **Abhishek Anand**

Contents

- Problem statement
- Problem approach
- EDA
- Correlations
- Model Evaluation
- Observations
- Conclusion

Problem Statement

- An education company named X Education sells online courses to industry professionals.
- On any given day, many professionals who are interested in the courses land on their website and browse for courses. They have process of form filling on their website after which the company that individual as a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.Through this process, some of the leads get converted while most do not.
- X Education's current lead conversion rate is approximately 30%.
- Now, this means if, say, they acquire 100 leads in a day, only about 30 of them are converted. To make this process more efficient, the company wishes to identify the most potential leads, also known as Hot Leads.
- If they successfully identify this set of leads, the lead conversion rate should go up as the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone

Business Objective

- Lead X wants us to build a model to give every lead a lead score between 0 -100 . So that they can identify the Hot leads and increase their conversion rate as well.
- The CEO want to achieve a lead conversion rate of 80%.
- They want the model to be able to handle future constraints as well like Peak time actions required, how to utilize full man power and after achieving target what should be the approaches.

Solution Approach

1. *Data Cleaning:*

   - Replaced "Select" values with NaN and removed rows with missing values.

   - Encoded categorical variables using one-hot encoding.

2. *Feature Engineering:*

   - Standardized numerical features.

   - Selected top 10 features using Recursive Feature Elimination (RFE).

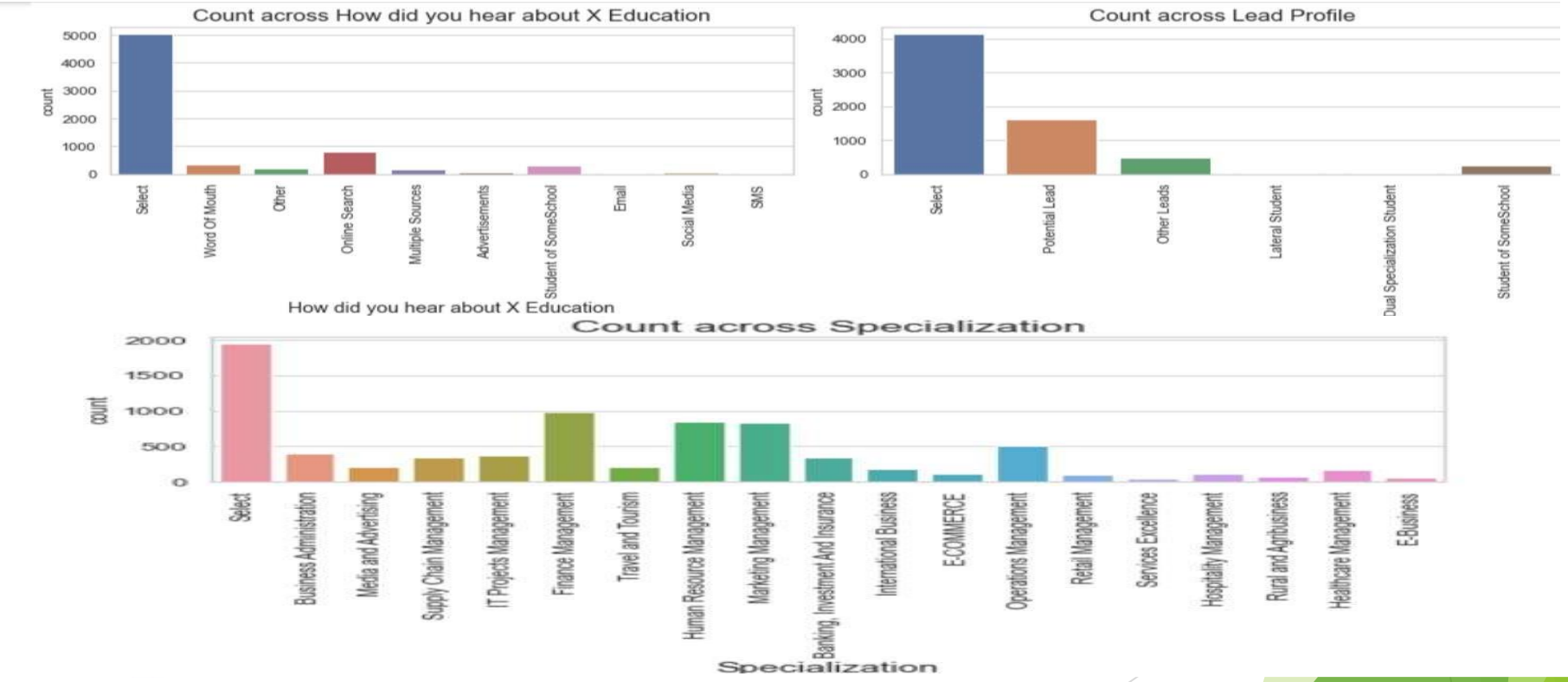
3. *Model Development:*

   - Logistic regression model trained and tested on 70%-30% split.

4. *Evaluation:*

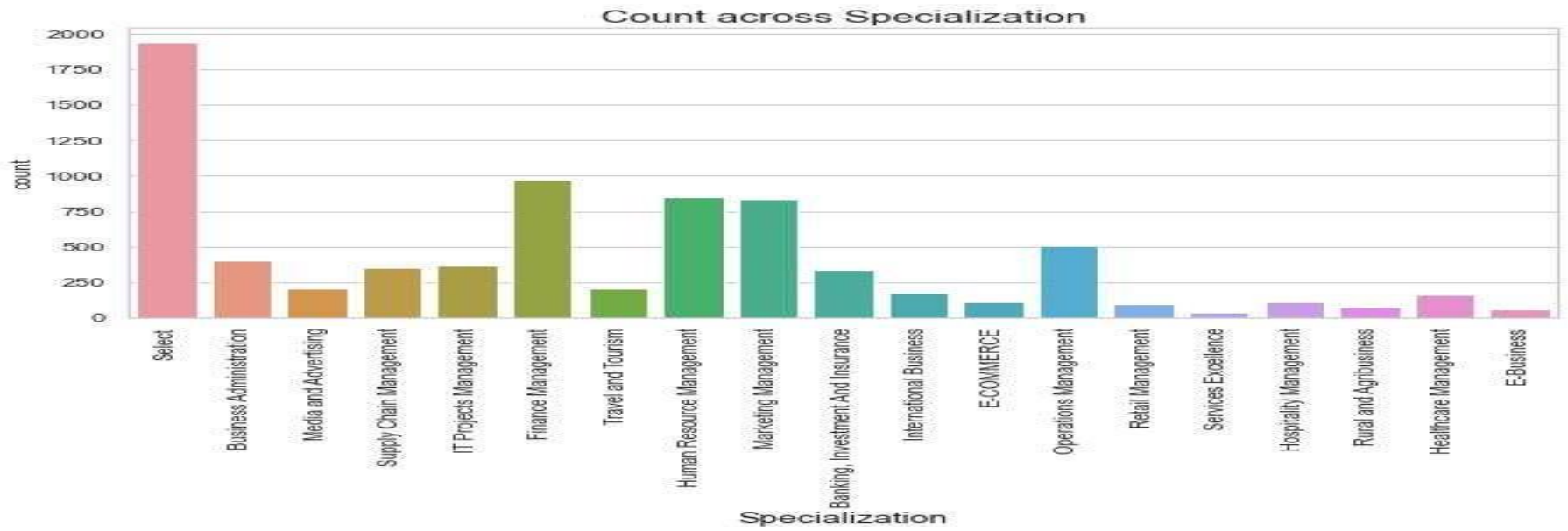
   - Assessed model using accuracy, precision, recall, F1-score, and ROC-AUC.

EDA – Data Cleaning



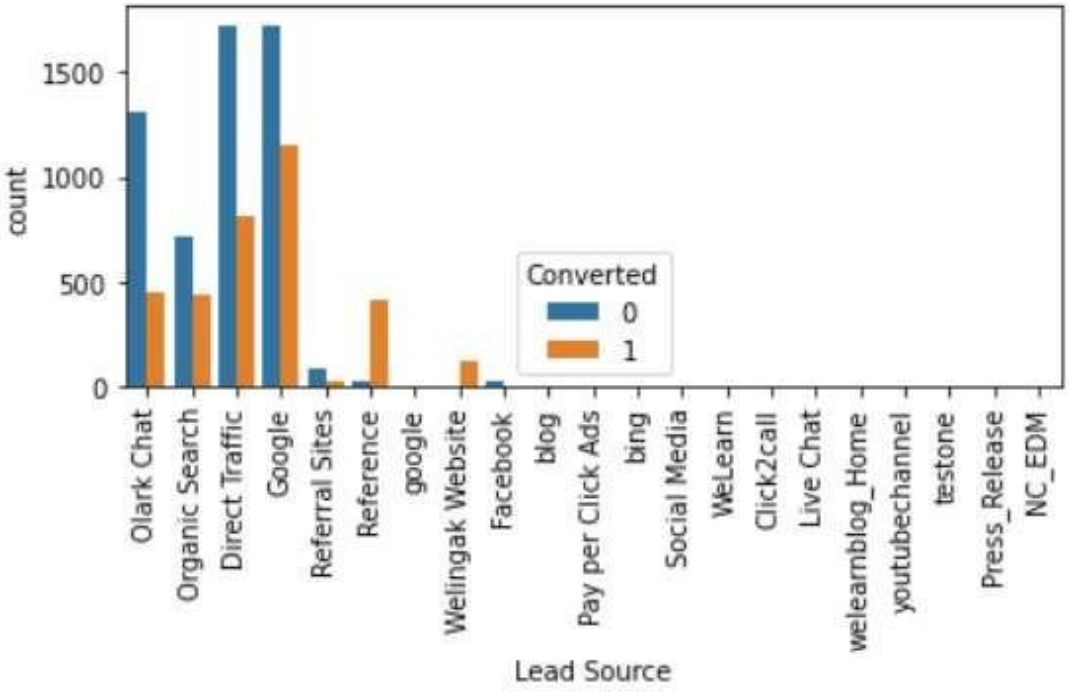
Specialization

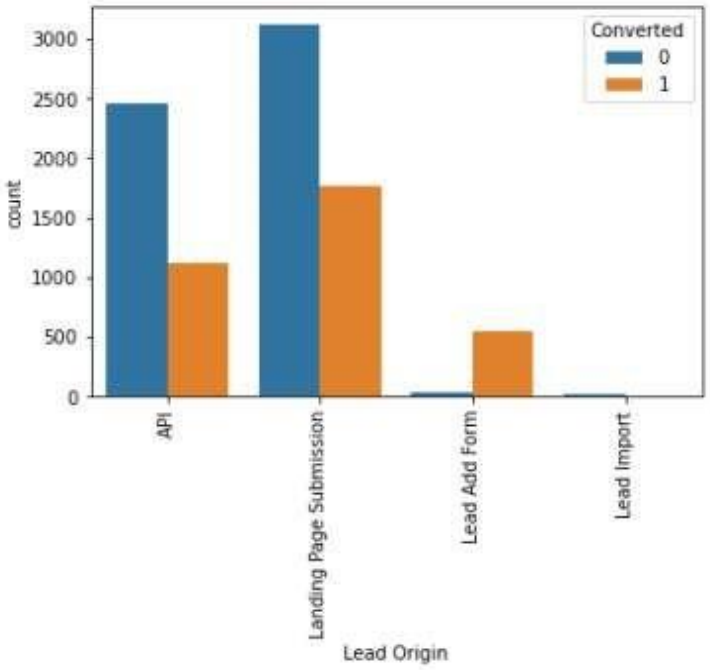
Leads from HR, Finance & Marketing management specializations are high probability to convert



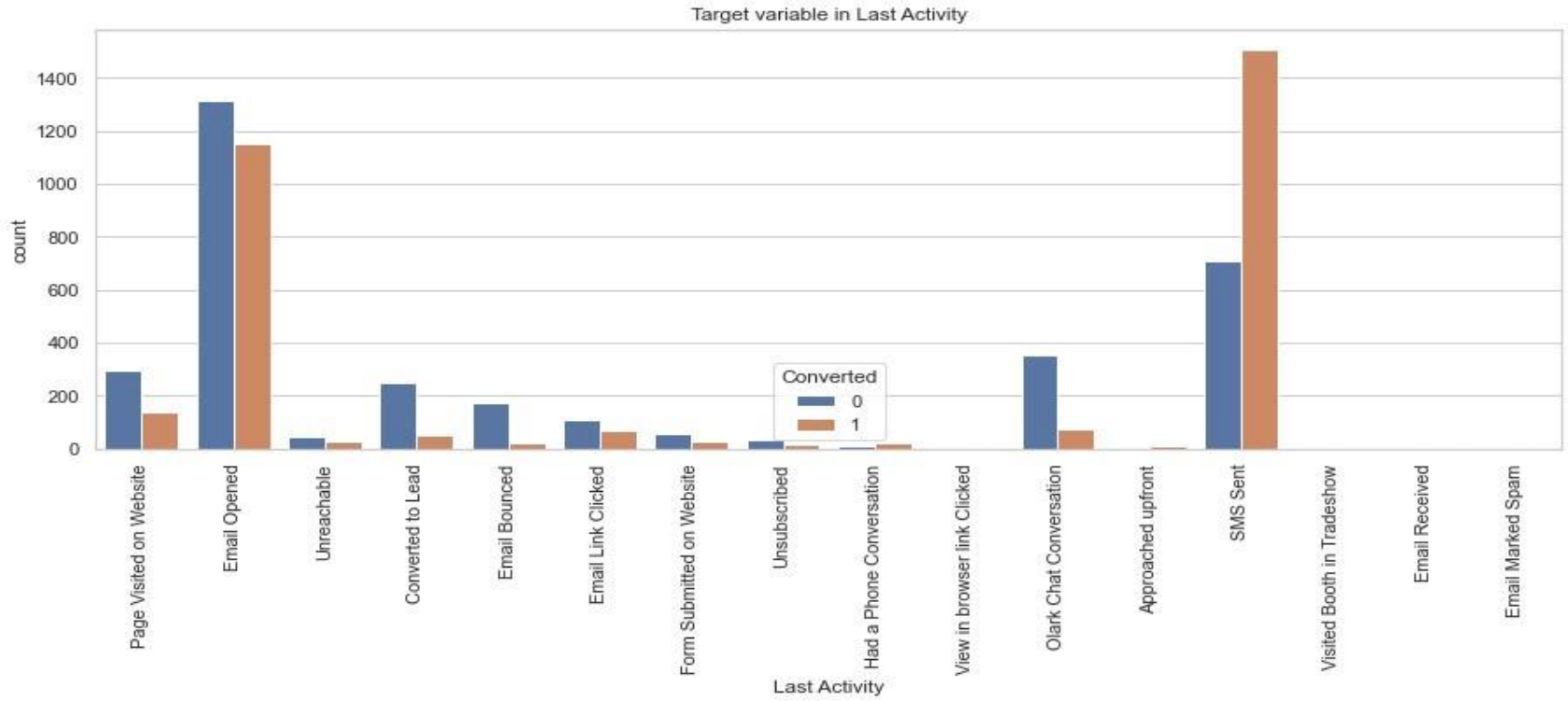
Lead Source & Lead origin

Whereas in Lead origin most number of leads are landing on submission



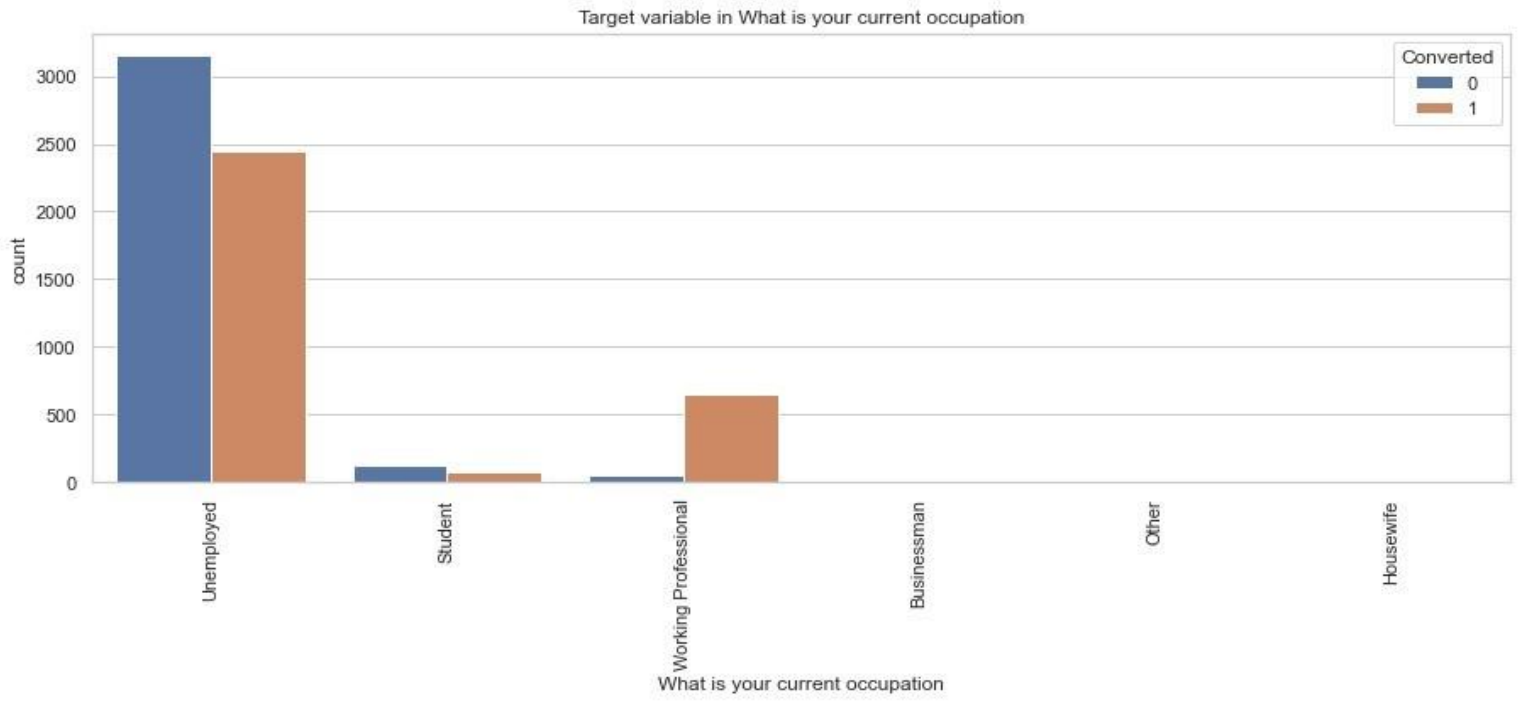


Last lead Activity

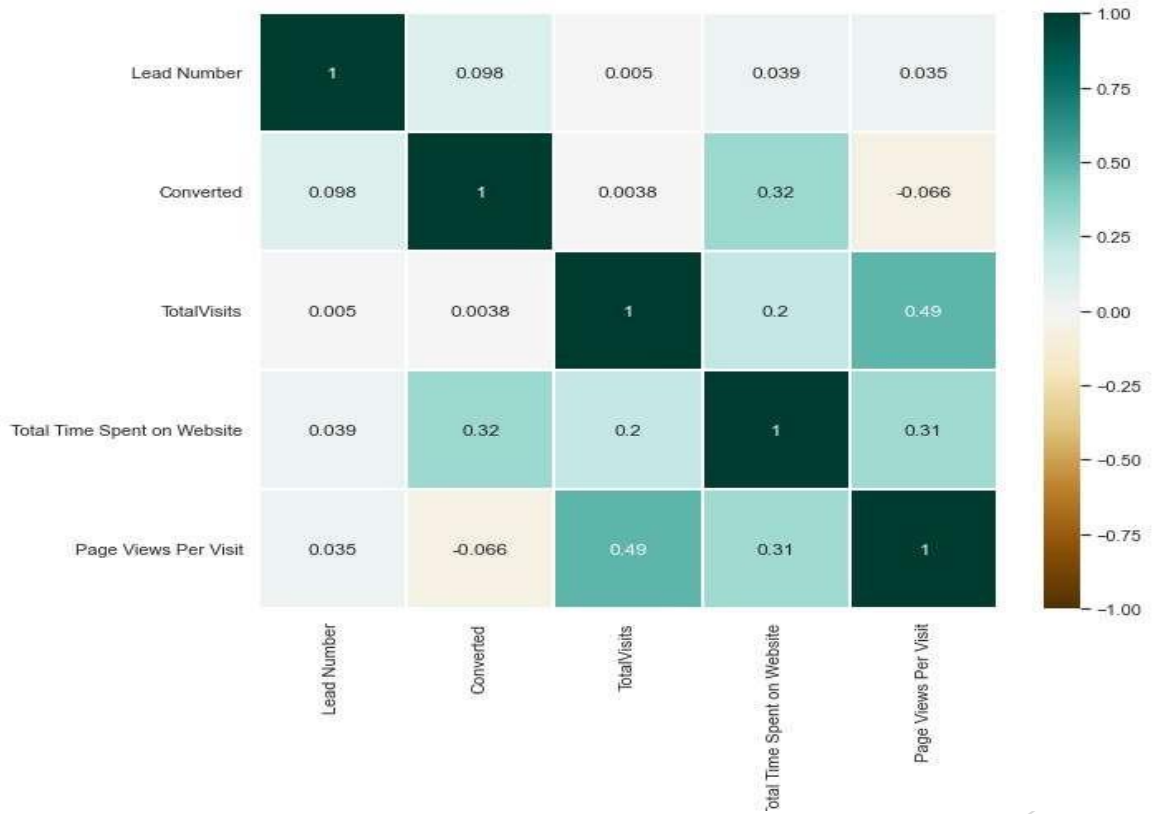


Last What is Your Occupation

hers.



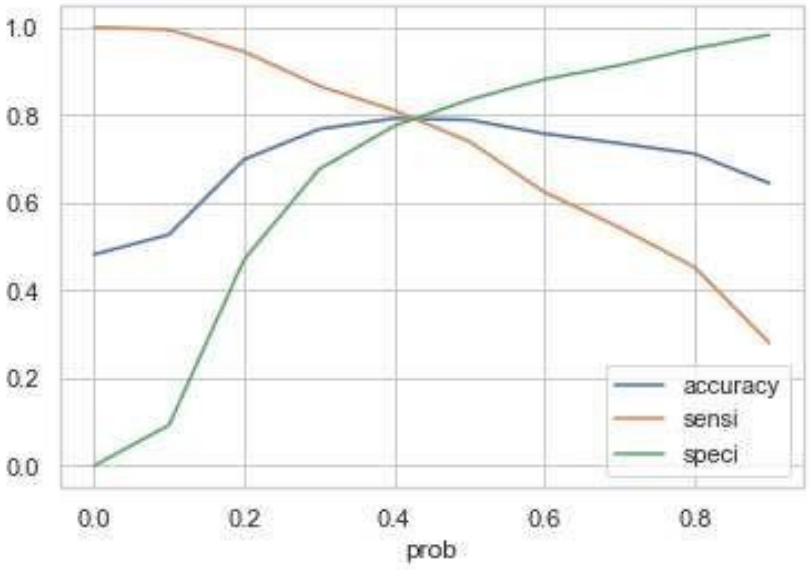
Correlation

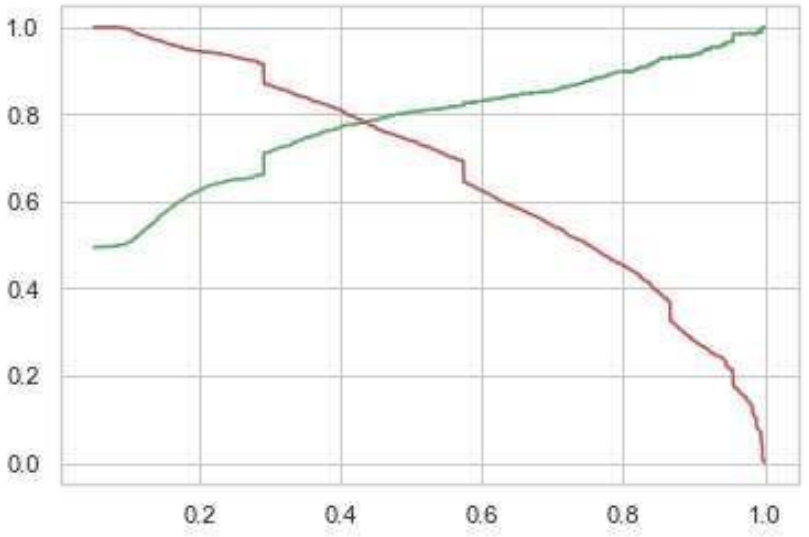


Model Evaluation

0.42 is the tradeoff between Precision and Recall -

Thus we can safely choose to consider any Prospect Lead with Conversion **Probability higher than 42 % to be a hot Lead**





Observations

Train Data:

Accuracy : 80%
Sensitivity : 77%
Specificity : 80%

Test Data:

Accuracy : 80%
Sensitivity : 77%
Specificity : 80%

Final Features list:

- Lead Source_Olark Chat
- Specialization_Others
- Lead Origin_Lead Add Form
- Lead Source_Welingak Website
- Total Time Spent on Website
- Lead Origin_Landing Page Submission
- What is your current occupation_Working Professionals
- Do Not Email

Key Learnings

- Importance of clean and complete data for model performance.
- Feature importance analysis offers actionable insights for marketing and sales optimization.
- Business-aligned threshold adjustments can maximize operational efficiency..

Conclusion

- The logistic regression model provides a reliable mechanism to score leads.
- Strategic adjustments based on business goals can enhance efficiency and conversion rates.
- We see max number of leads are generated by google / direct traffic. Max conversion ratio is by reference and welingak website.
- Leads who spent more time on website, more likely to convert.
- Most common last activity is email opened. highest rate = SMS Sent. Max are unemployed. Max conversion with working professional.