Lead Scoring Case Study

Using Logistic Regression

Team: Suresh Kumar Kumar Singh, Shoyeb Patwekar & Sourabh Bhatt

Agenda

- Problem Statement
- Approach
- EDA
- Model Evaluation
- Observations
- Conclusion

Problem Statement

1. 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.
2. The typical lead conversion rate at X education is around 30%. Company wishes to identify the hot leads to maximize the conversion rate
3. CEO has ballpark target of 80%.
4. Hot leads will help sales team to make calls to targeted audience instead of all public

Problem Solving Approach

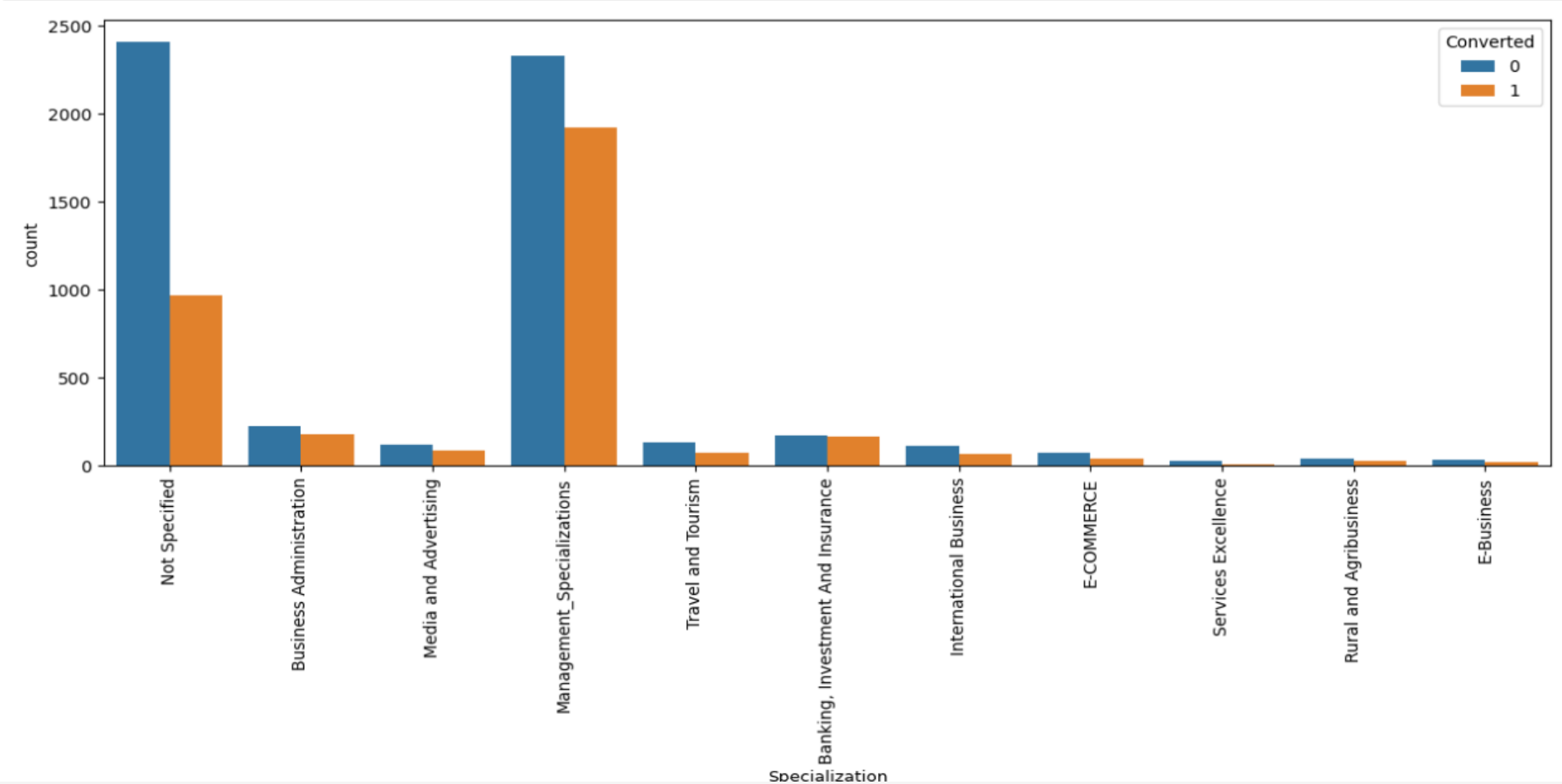
- Read and Inspect the Leads data
- Data clean-up and preparation
- EDA
- Dummy variables
- Test Train Split
- Feature Scaling
- Feature Selection using RFE
- Model Evaluation
- Precision and Recall

Data Clean up and Preparation

- Columns having “Select” means null, was replaced with NaN
- Columns more than 45% Null Values dropped
- Rows having missing data dropped, less than 2%
- New Category introduced, e.g. Not Specified for NULL value in Specialization
- Aggregated few columns like Management for Specialization
- Missing value for occupation is replaced with Unemployed

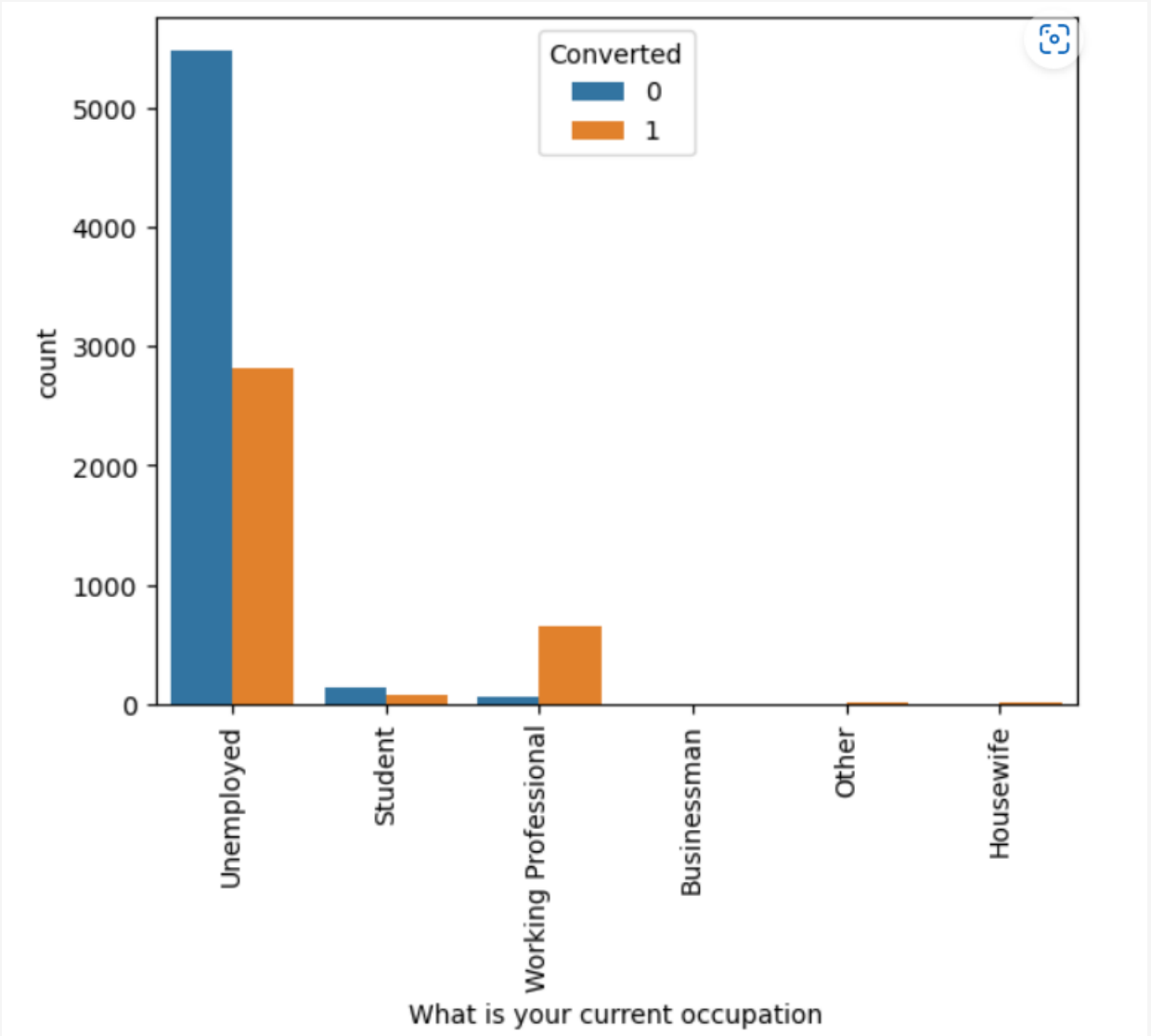
EDA

Specialization



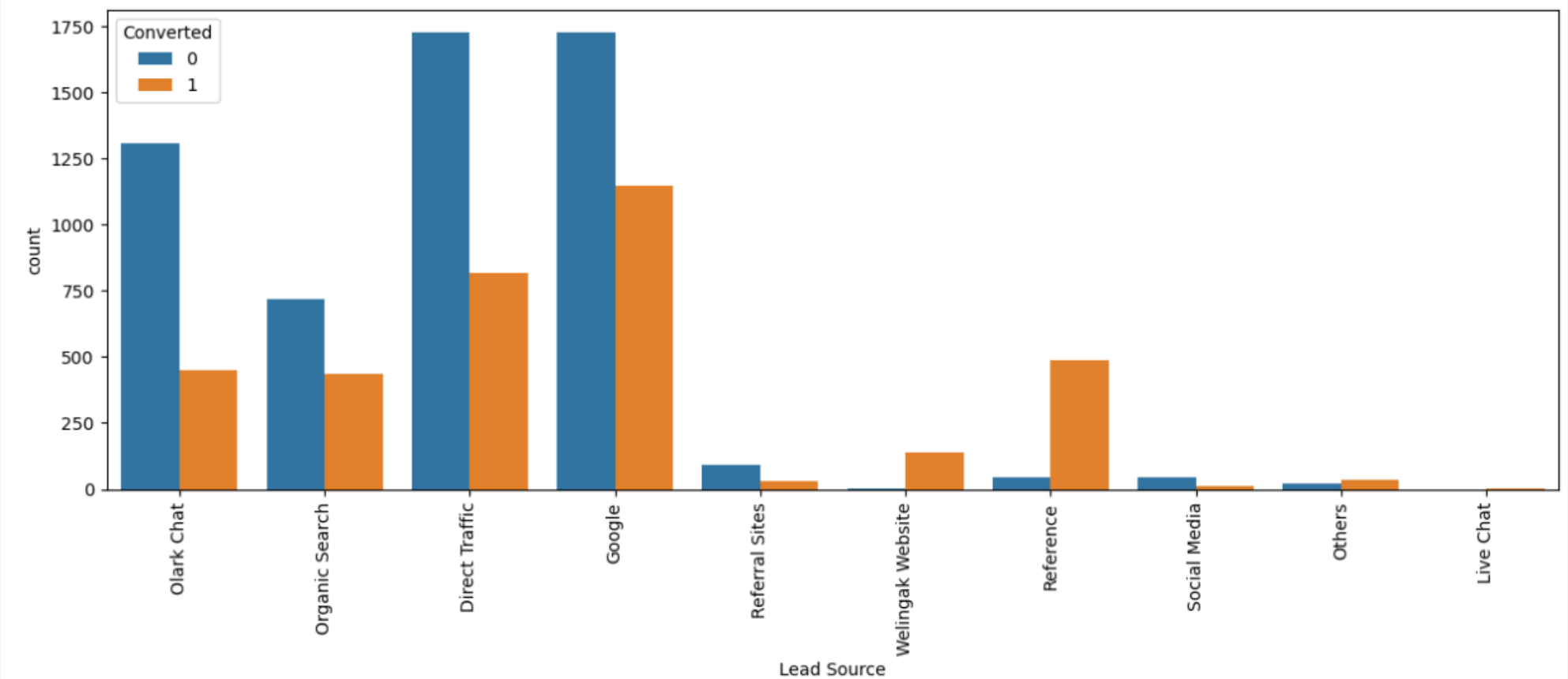
Occupation

- Unemployed People are quantitatively more converted
- Working Professionals have high conversion rate



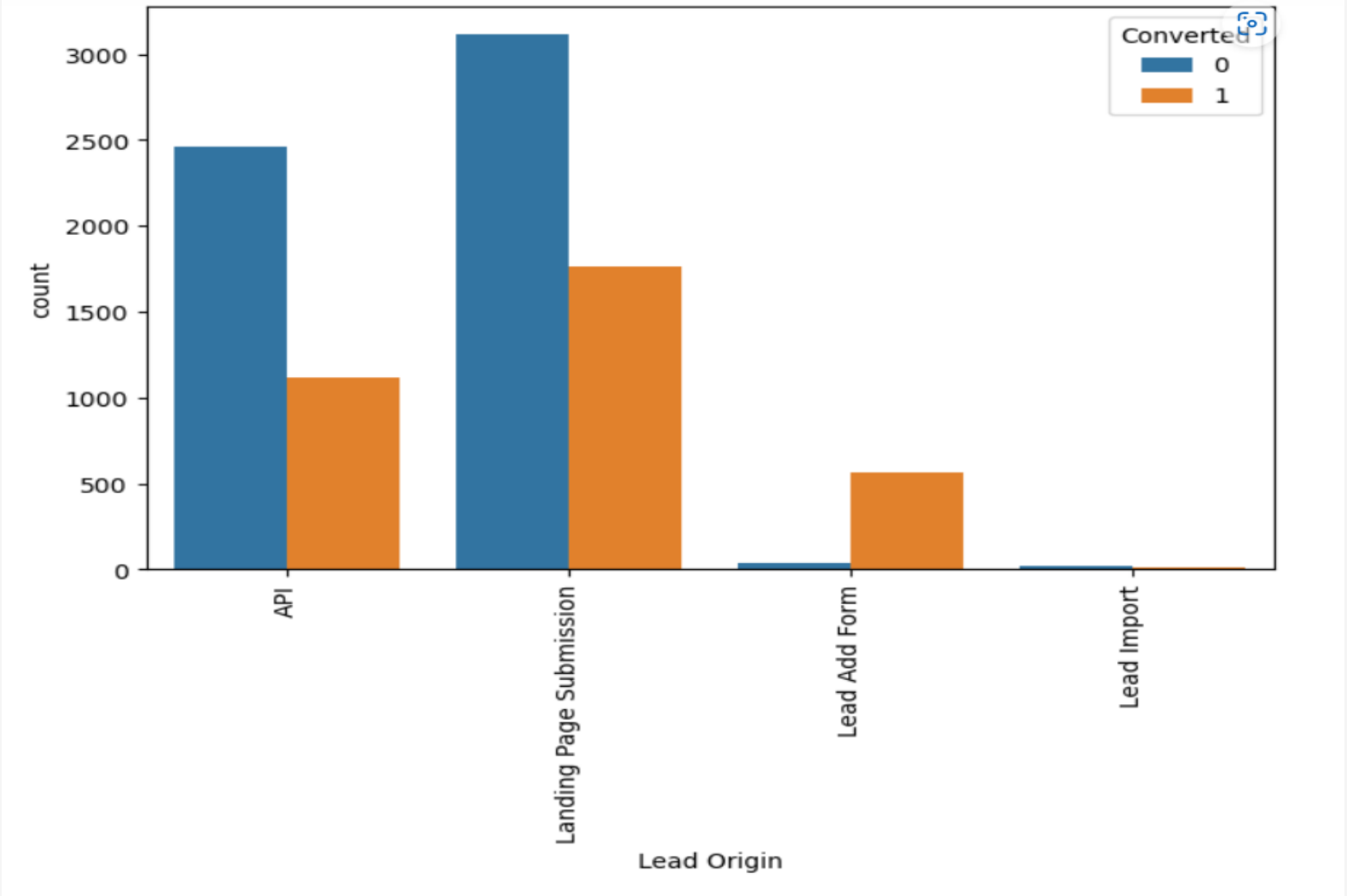
Lead Source

- Maximum number of leads are generated by Google and Direct traffic and significant conversion, less lead from welingak & reference website but high % of conversion rate



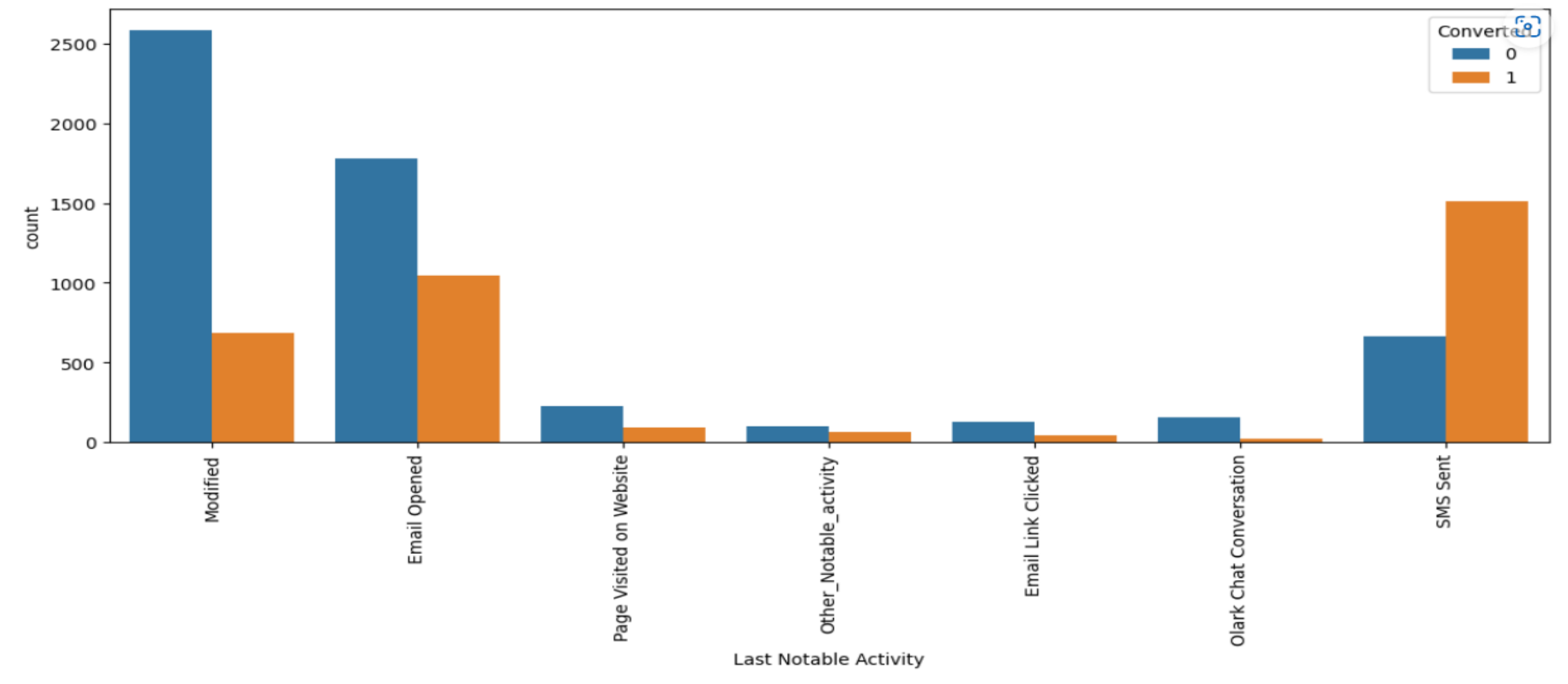
Lead Origin

- API and Landing Page Submission bring higher number of leads as well as conversion.



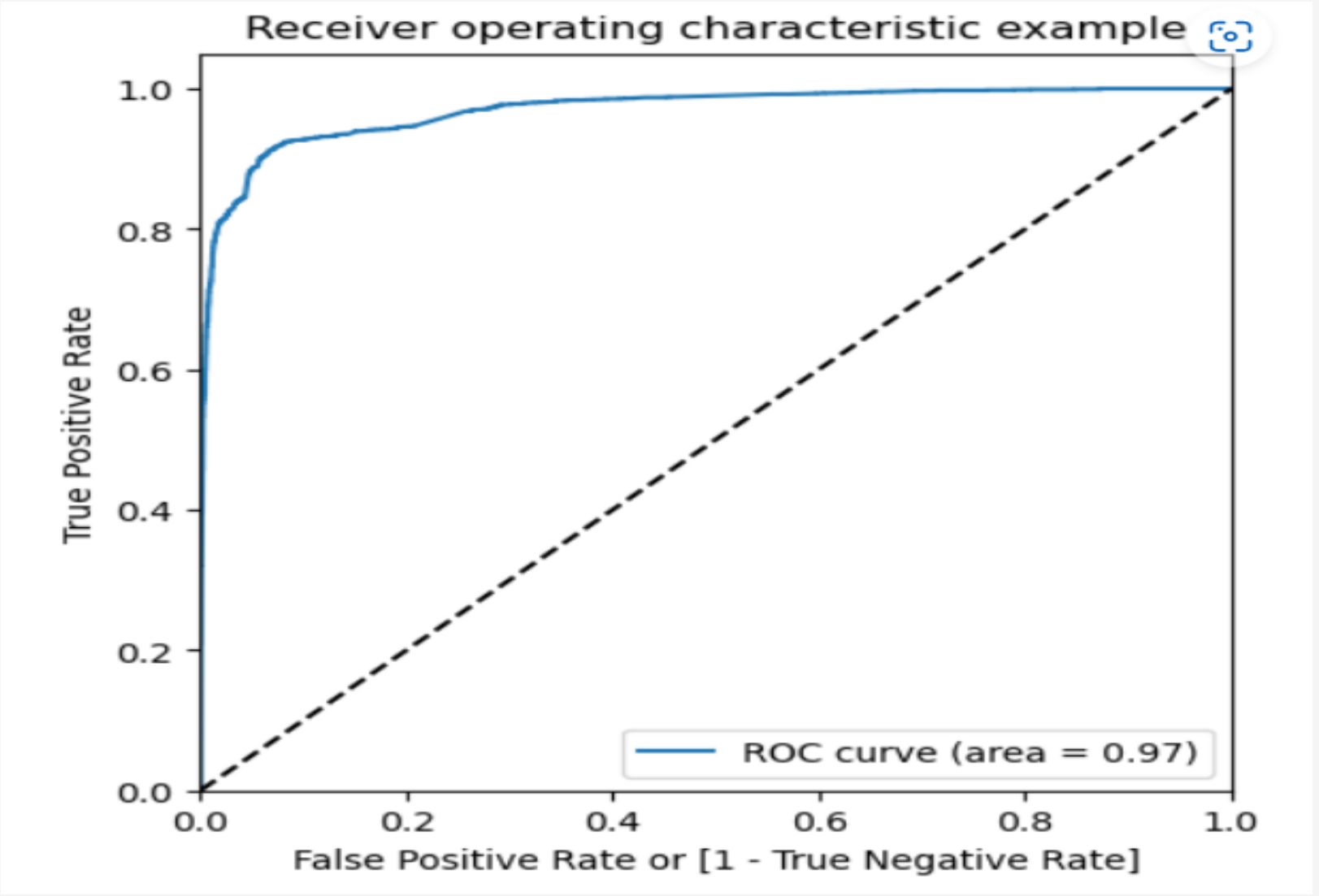
Last Notable Activity

- SMS Sent , Email Opened are top notable activities



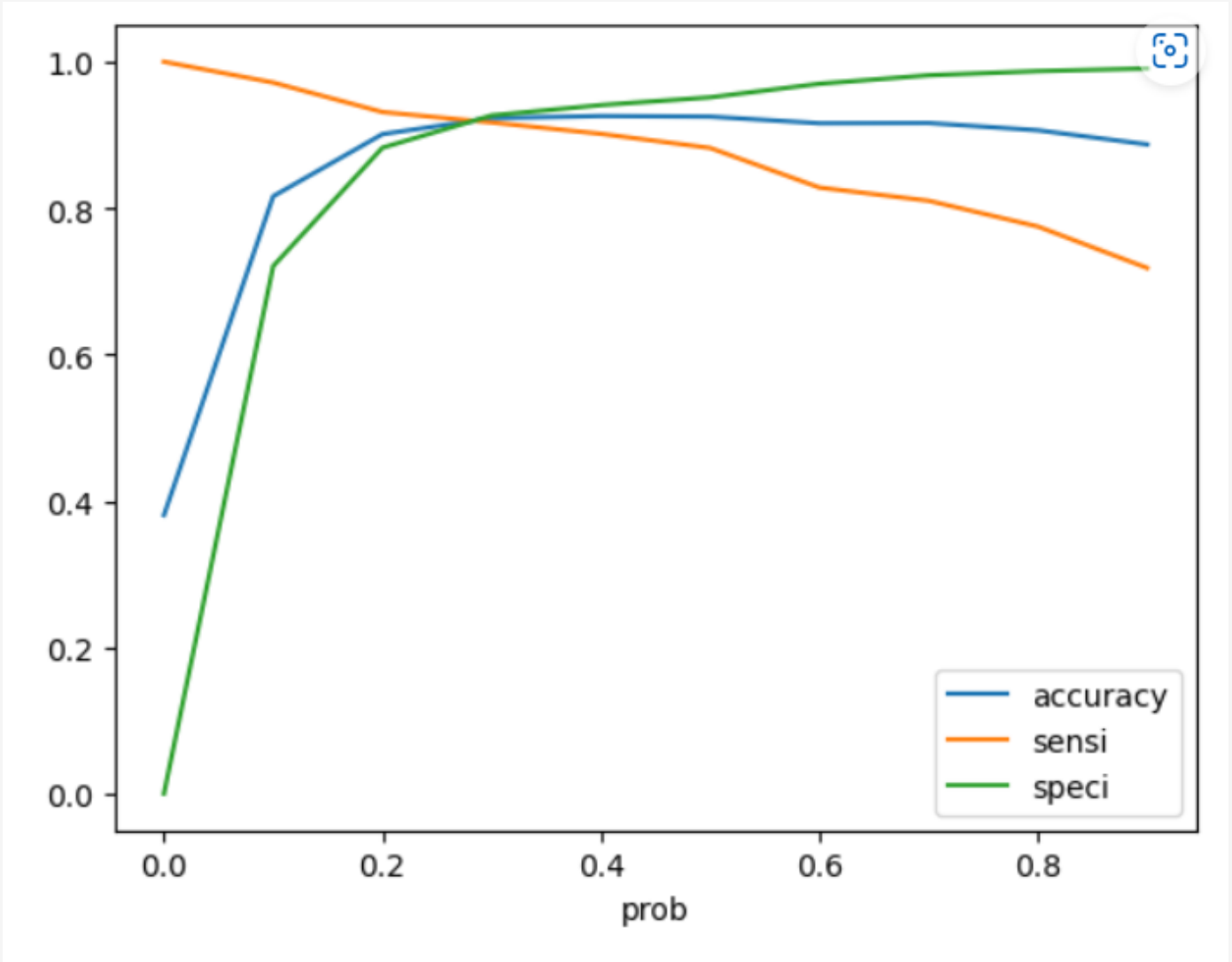
Model Evaluation

ROC Curve



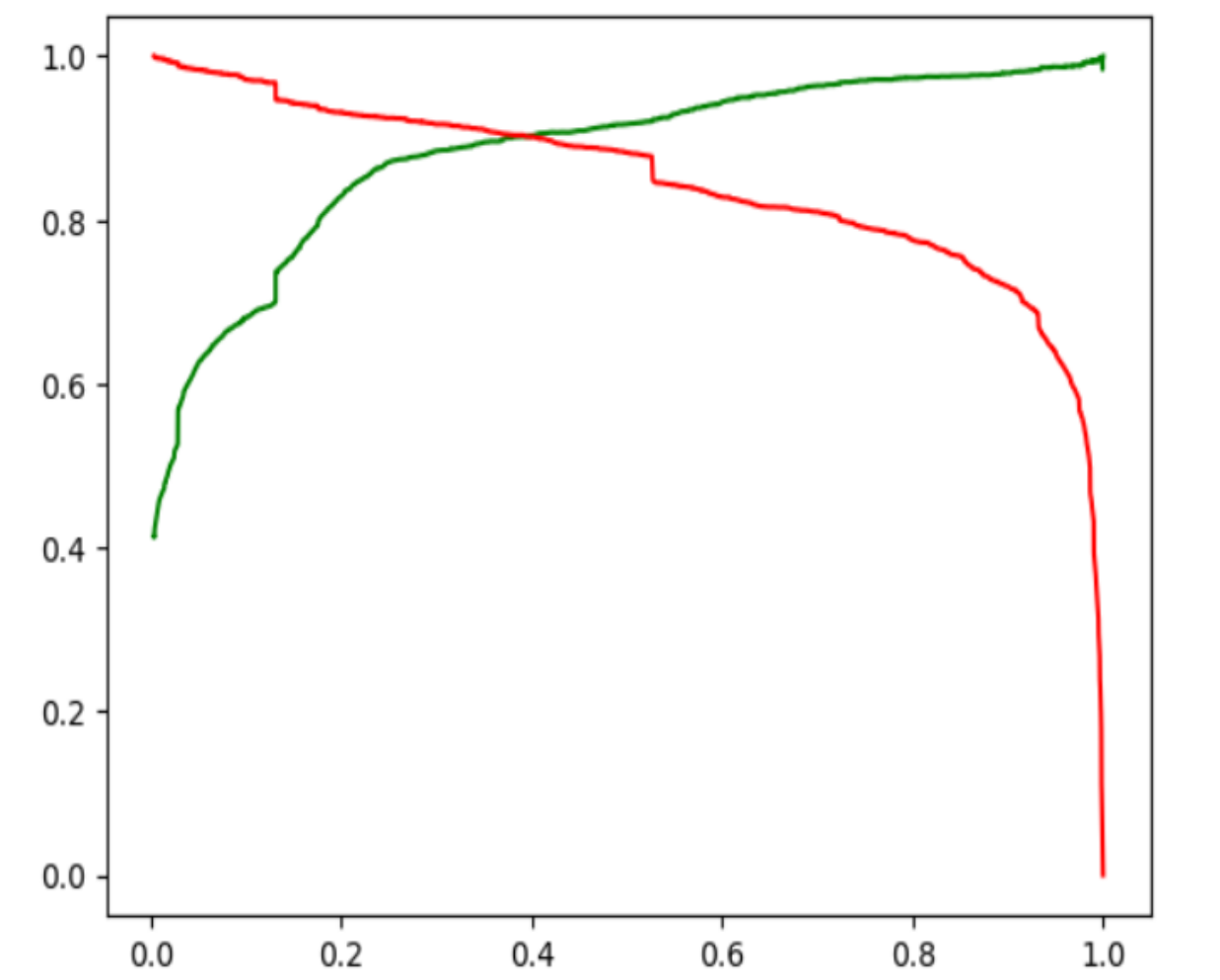
Selection of Optimal Cut off value

- Cut off value 0.3 Selected based on Accuracy, Sensitivity and Specificity
- Accuracy 92.3%
- Sensitivity 91.6%
- Specificity 92.6%



Precision And Recall

- Precision and Recall trade off should be between 0.3 and 0.5



Observations

After running the model on the Train Data:

- Accuracy : 92.29%
- Sensitivity : 91.70%
- Specificity : 92.66%

After running the model on the Test Data:

- Accuracy : 92.78%
- Sensitivity : 91.98%
- Specificity : 93.26%