LEAD SCORING CASE STUDY

DONE BY: RAVNEET SINGH, RHENIUS RAJKUMAR, RUSHABH DAITHANKAR

INDEX

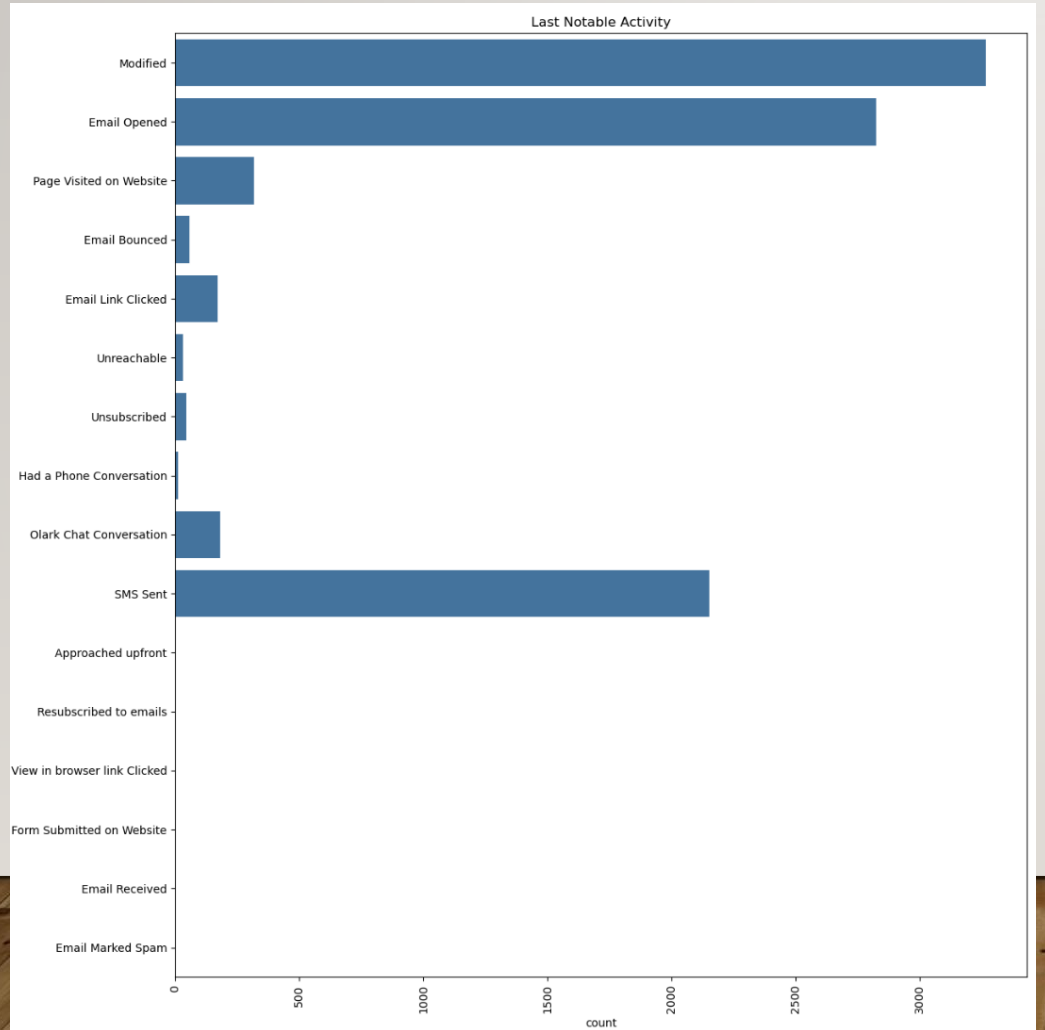
PROBLEM STATEMENT

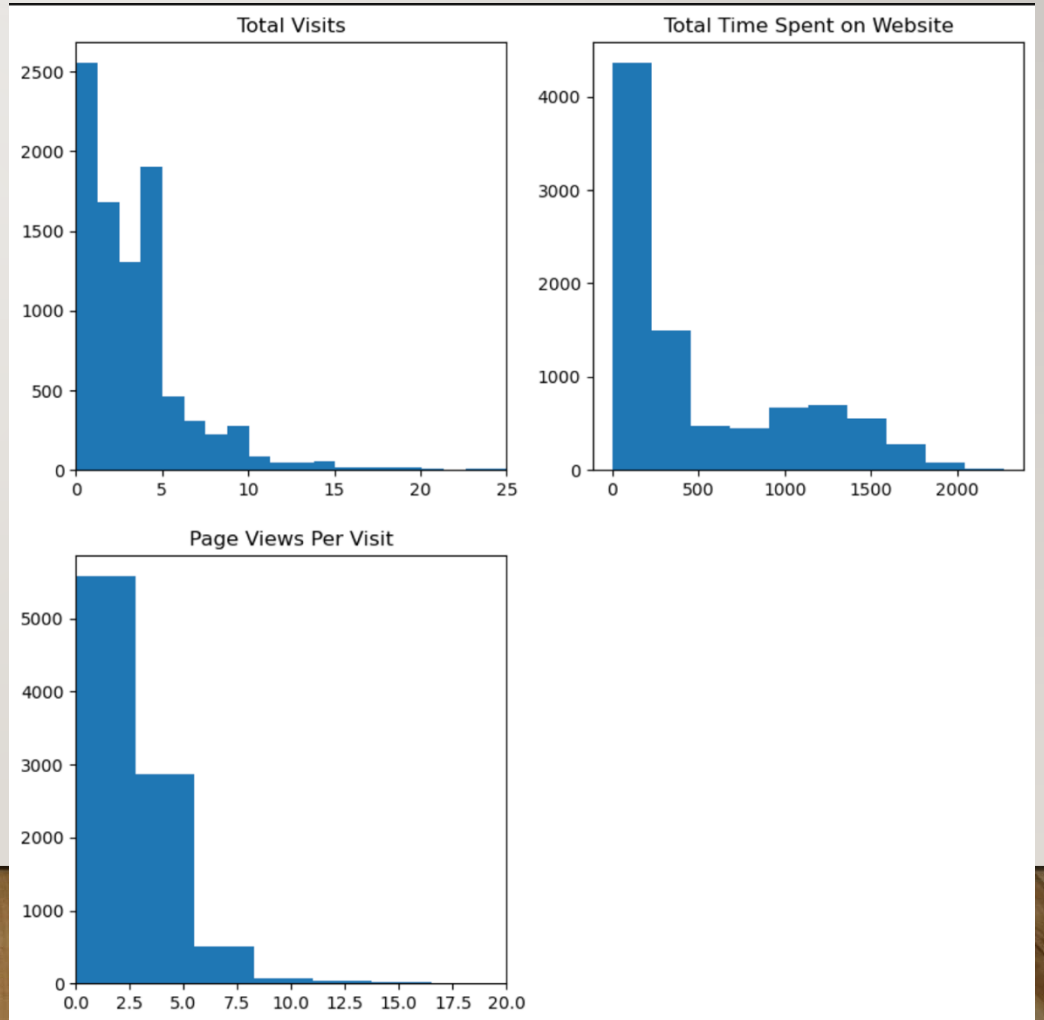
- Objective: Improve lead conversion rate from ~30% to 80%.
- Challenge: Identify "Hot Leads" likely to convert.
- Approach: Build a logistic regression model using historical lead data.
- Data: ~9,000 leads with attributes like Lead Source, Time on Website, and Last Activity.
- Outcome: Assign lead scores (0–100) to prioritize potential customers.
- Deliverables: Model, predictions, evaluation metrics, solutions document, and presentation.

DATE PREPARATION

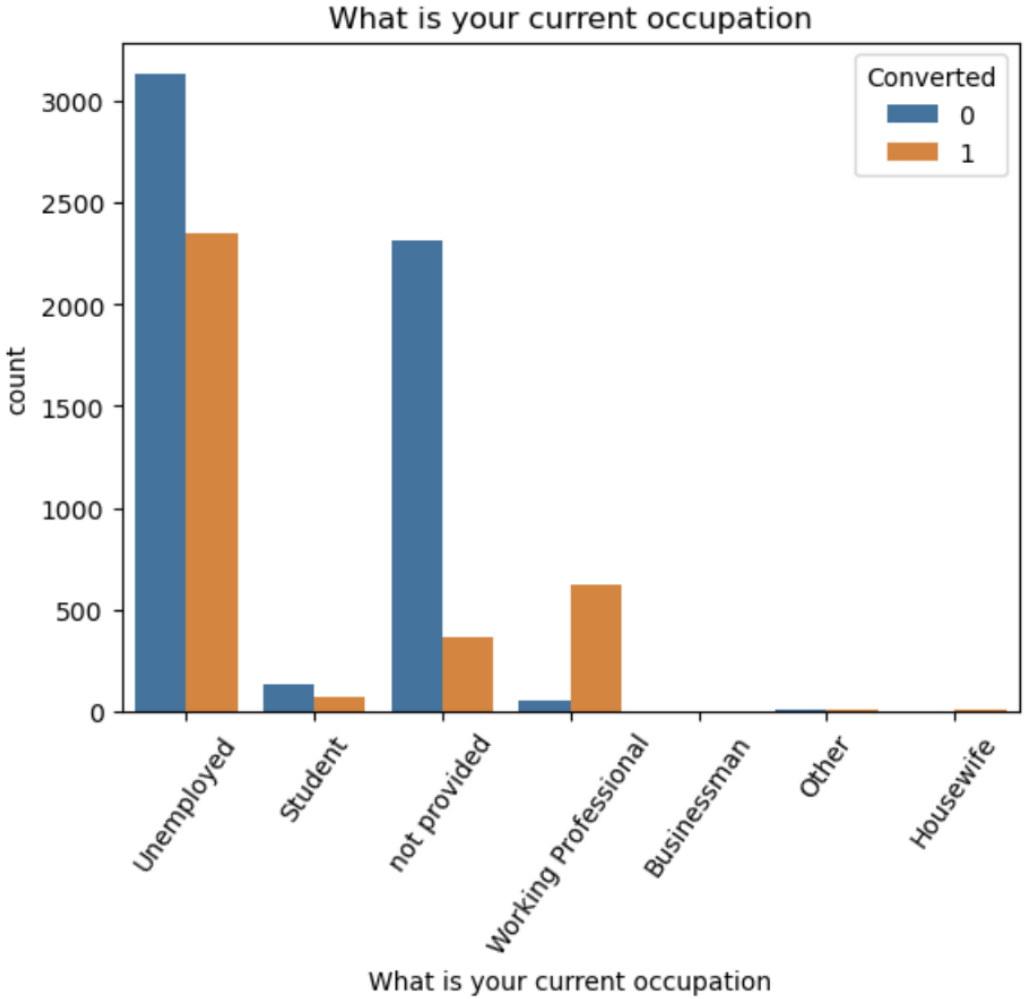
- Rows: 37, Columns: 9240
- Replacing the select with null
- Deleting columns with more than 35% of null values like: ‘Tags', 'Lead Quality’, and columns related to ‘Asymmetrique’
- Deleting the columns with only a single unique value: Like magazine, Newspaper, Digital Advertisement
- Deleting the unique columns: Prospect ID, Lead Number

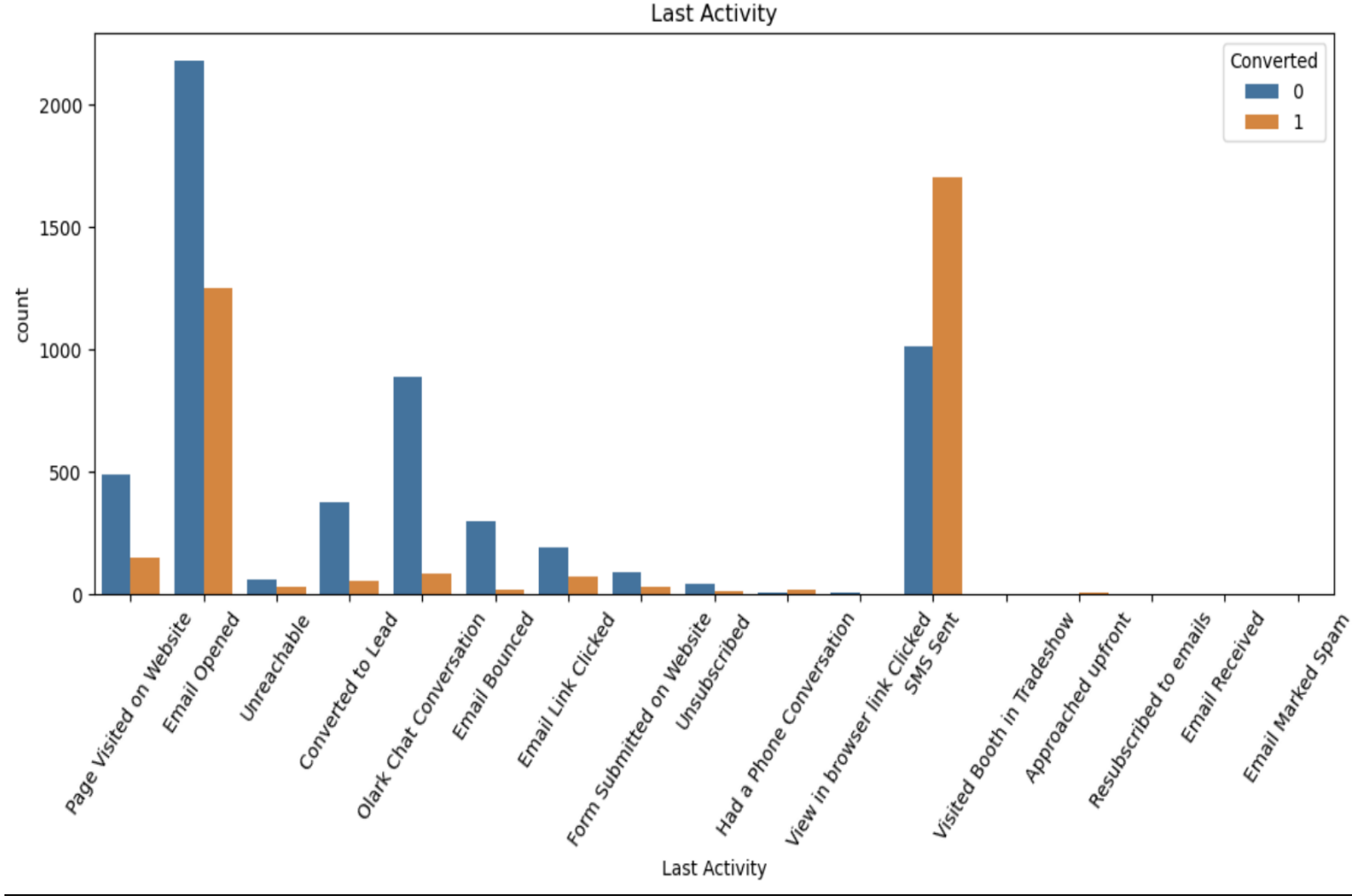
EXPLORATORY DATA ANALYSIS (1) - UNIVARIATE





EXPLORATORY DATA ANALYSIS (2) - BIVARIATE



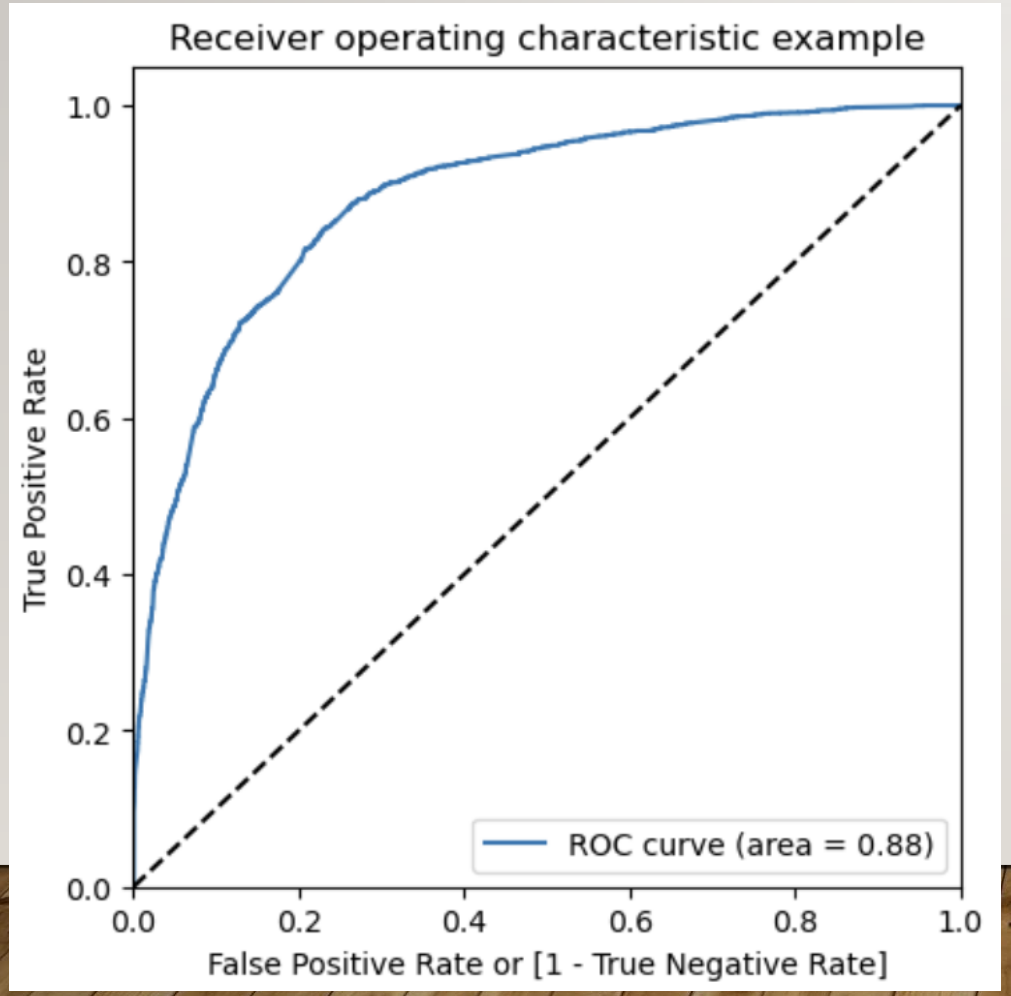


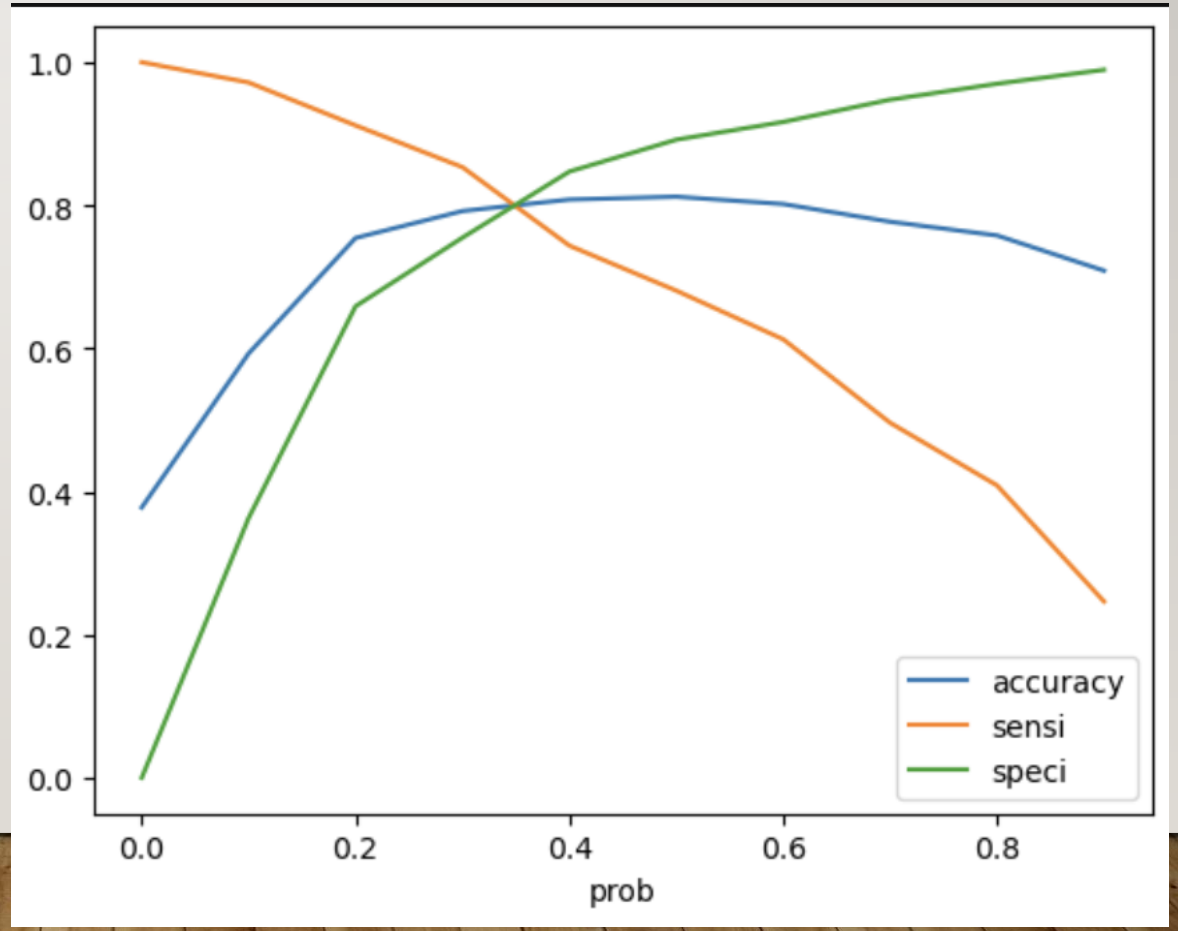
EXPLORATORY DATA ANALYSIS (3) - MULTIVARIATE

MODEL BUILDING

- Numerical columns scaled down using MinMaxScaler
- Dummy variables created using get_dummies
- Date was split into test-train at the ration 7:3 respectively
- Used RFE for feature selection, selected 15 variables
- Removed the columns where p-value > 0.05
- Overall accuracy ~80%

MODEL BUILDING – ROC CURVE





CONCLUSION

Columns the model highly depends on:

- Total Time Spent on Website
- Lead Source
  - Google
  - Organic Search
  - Direct Traffic
  - Referral Sites
  - Welingak Website
- Do Not Email
- Last Notable Activity
  - Email Opened
  - Modified
  - Email Link Clicked
- Lead Origin: Lead Add Form
- Last Activity: Olark Chat Conversation
- What is your current occupation: Working Professor