LEAD SCORING CASE STUDY

PROBLEM STATEMENT

- An education company named X Education sells online courses to industry professionals.
- The company markets its courses on several websites and search engines like Google.
- When the people fill up a form providing their email address or phone number, they are classified to be a lead.
- Once these leads are acquired, employees from the sales team start making calls, writing emails, etc.
- The typical lead conversion rate at X education is around 30%
- Even though they get a high number of leads not all leads are converted

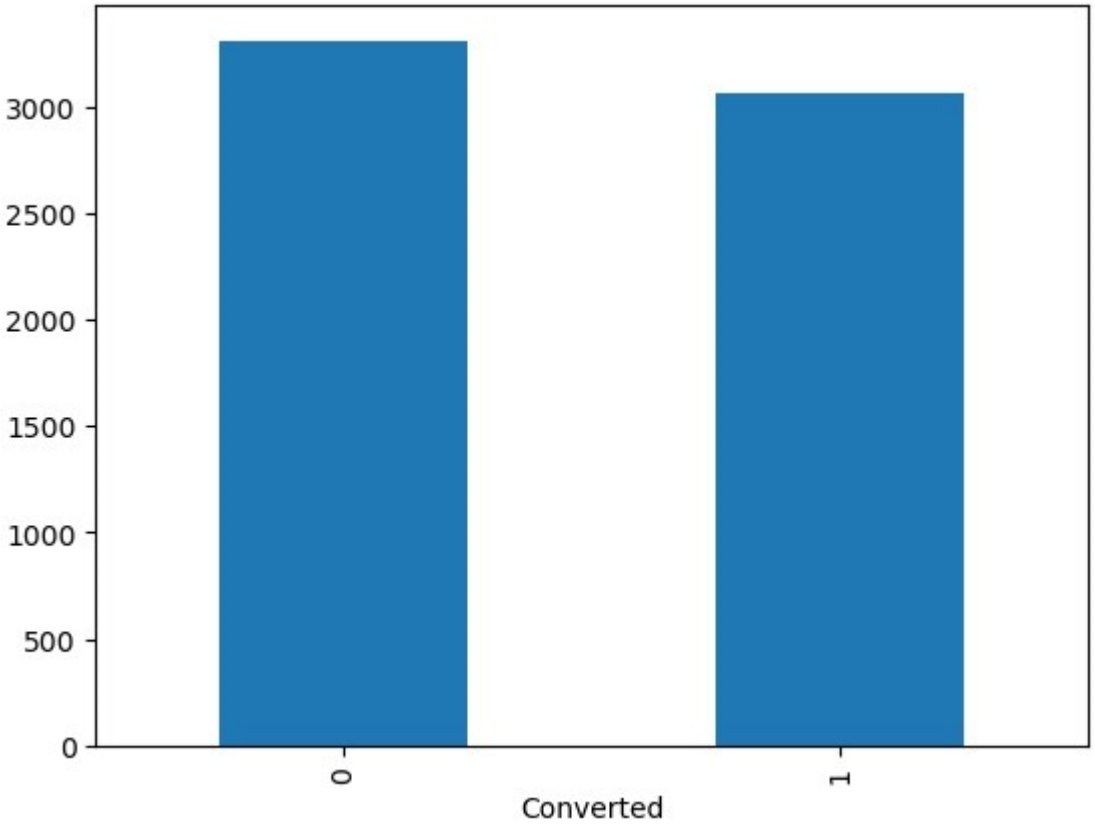
BUSINESS GOALS

- Lead X wants us to build a logistic regression 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

PROBLEM APPROACH

- Importing the data set and inspecting the data fields
- Data Pre-processing
- Exploratory Data Analysis
- Dummy variable creation for categorical features
- Test-Train split of the data set (70%-30%)
- Scaling the features
- Model building using Recursive Feature Elimination(RFE) and Variance Inflation Factor(VIF)
- Model Evaluation using Accuracy, Sensitivity, Specificity, Precision and Recall
- Making Predictions

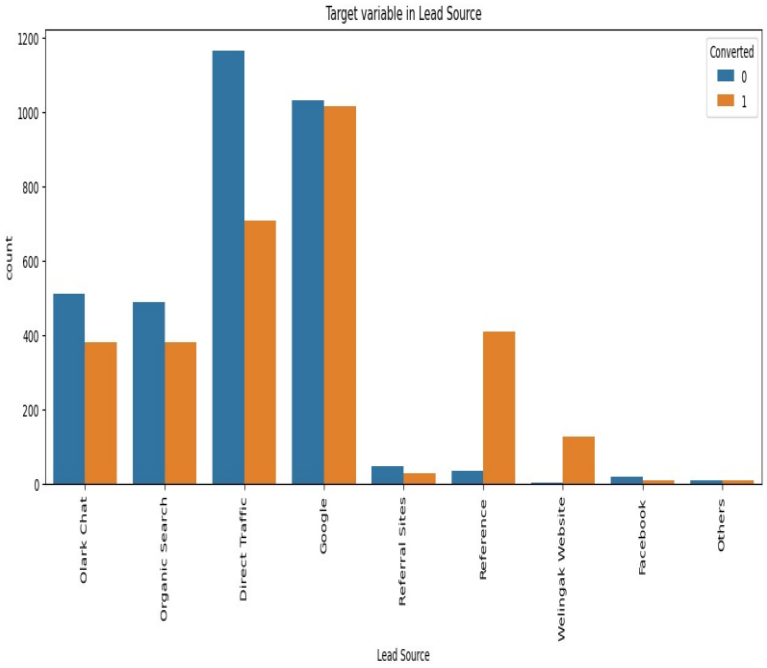
EDA- TARGET VARIABLE



THERE IS NO IMBALANCE IN THE TARGET FEATURE

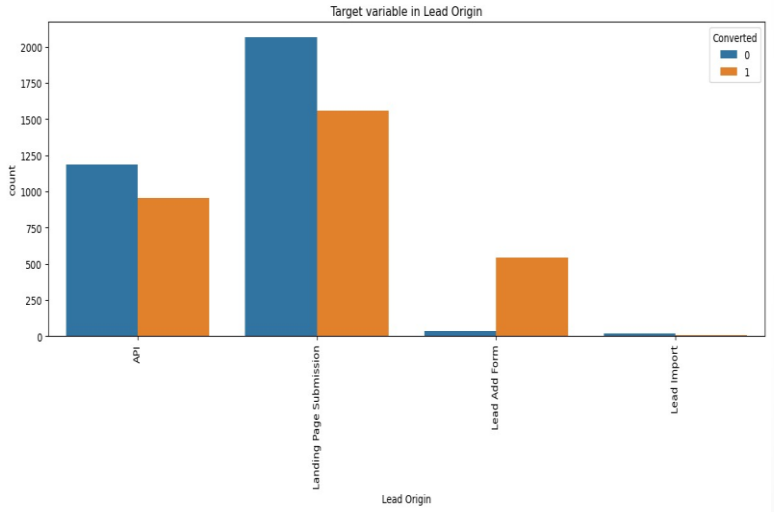
EDA- LEAD SOURCE and Lead ORIGIN

LEAD SOURCE



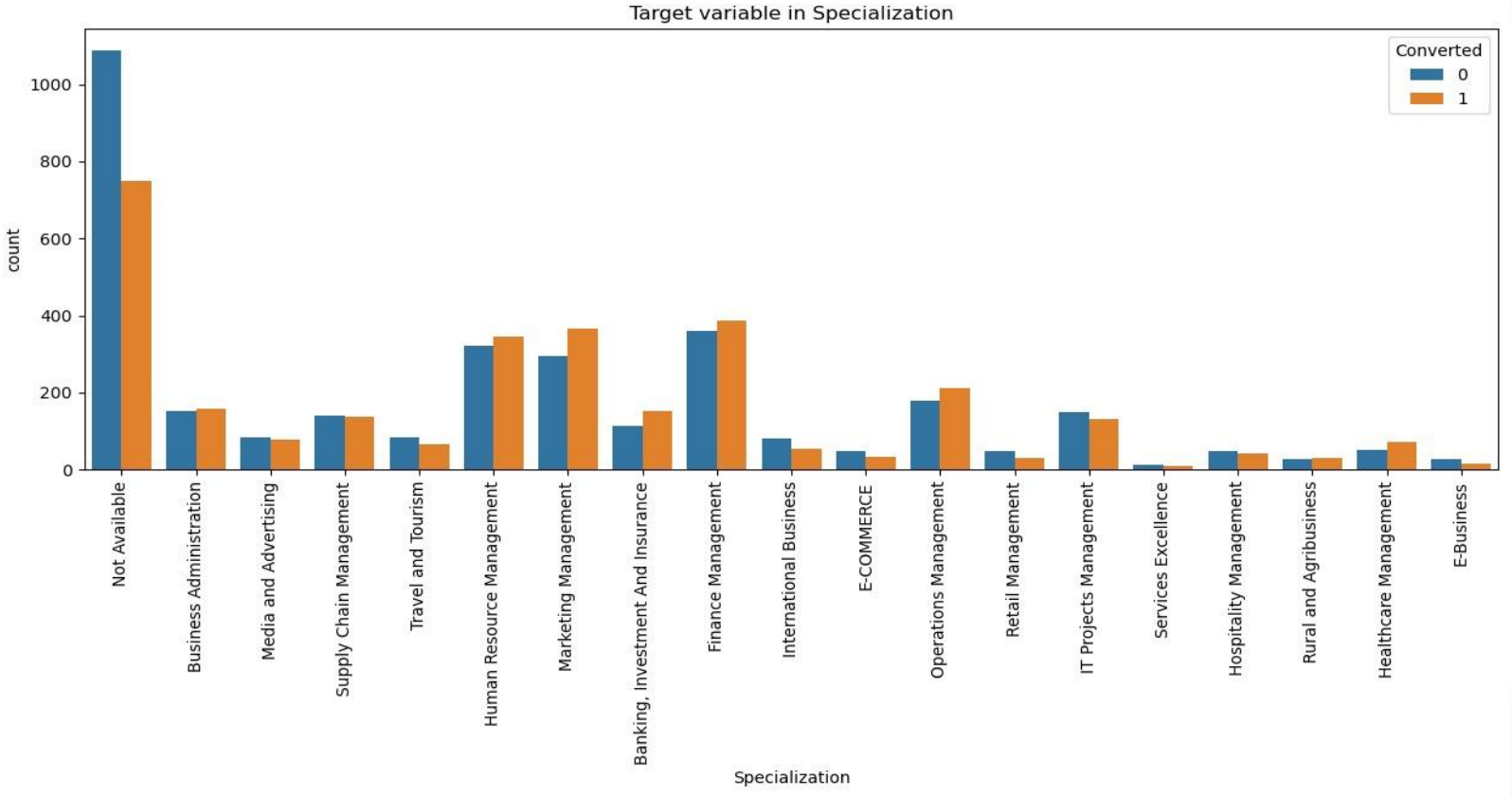
In Lead Source Google and other websites seem to attract more people to convert

LEAD ORIGIN



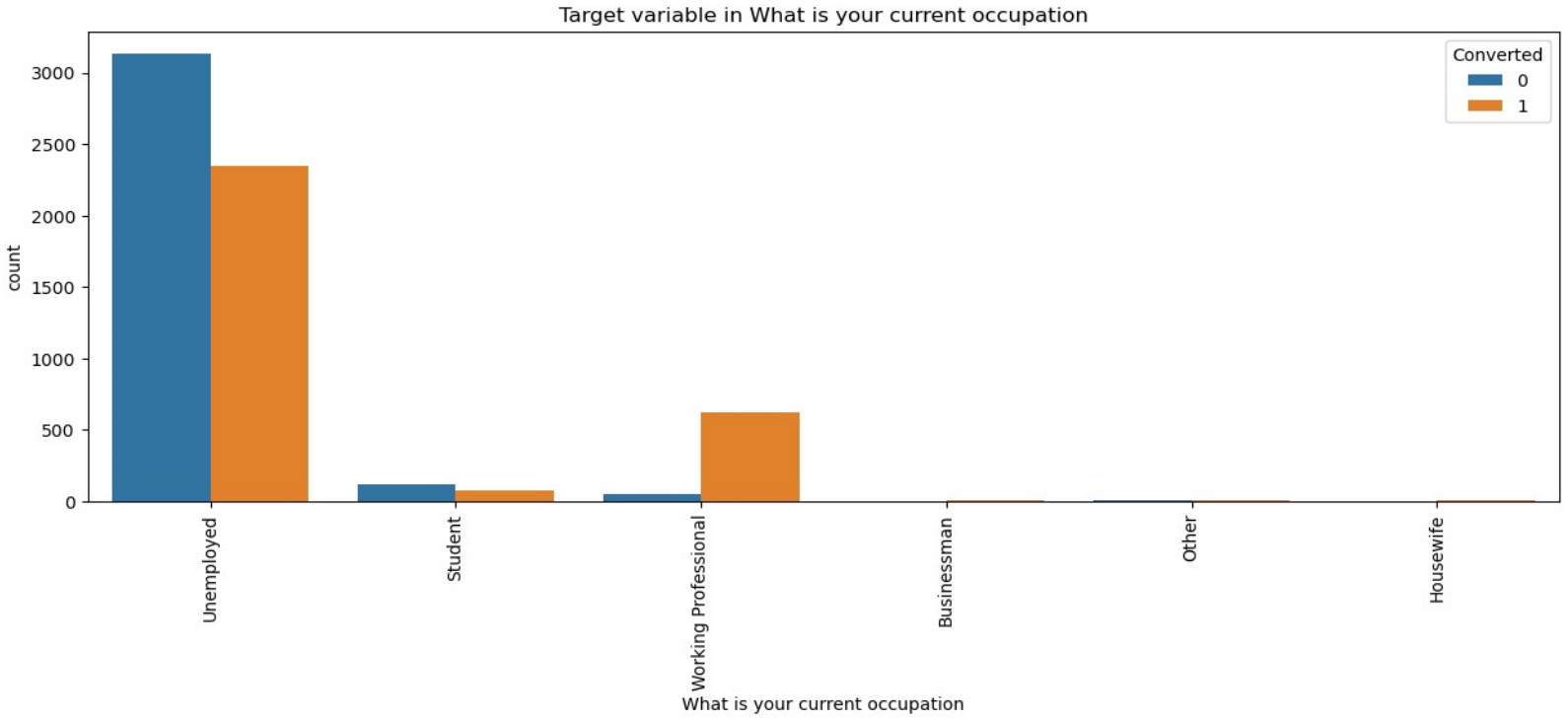
In Lead Origin most number of leads are Landing on Submission

TARGET VARIABLE IN SPECIALIZATION



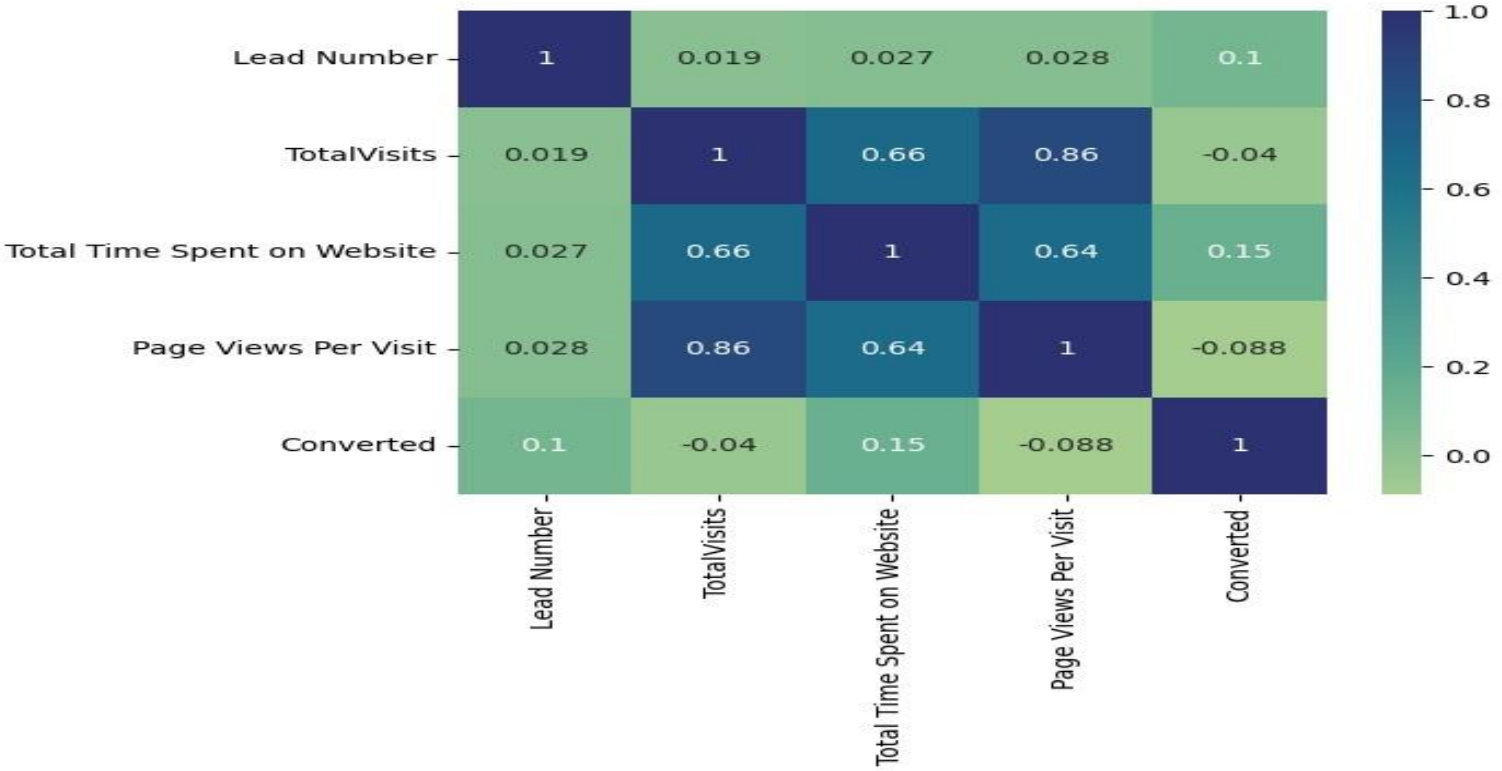
Leads from HR, Finance & Marketing management specializations are high probability to convert

TARGET VARIABLE IN OCCUPATION



Leads which are Unemployed are getting converted more

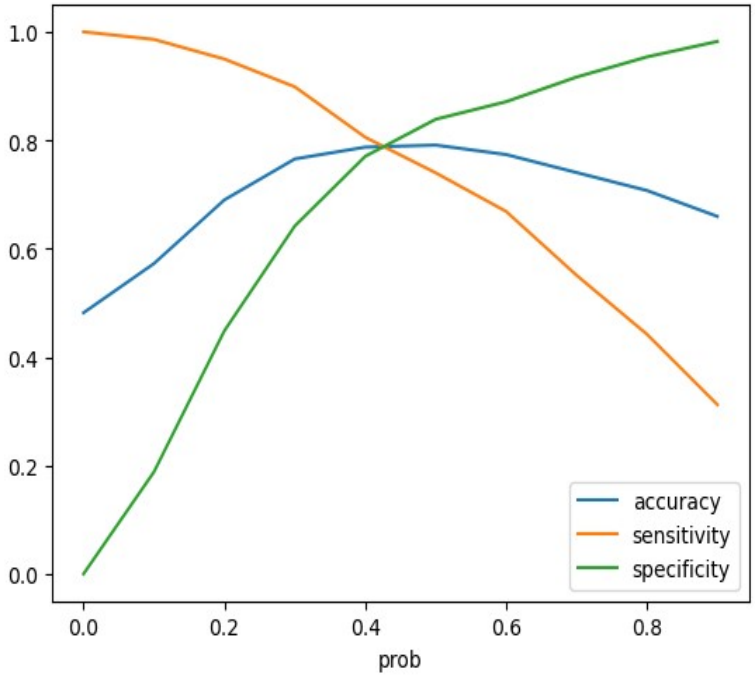
CORRELATION BETWEEN VARIABLES



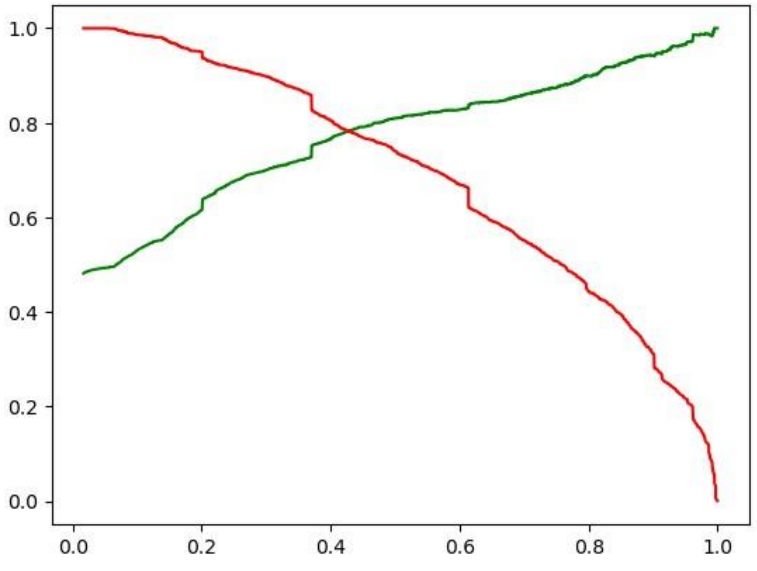
Correlation between numerical features and target variable is very low

MODEL EVALUATION

ACCURACY-SENSITIVITY-SPECIFICITY



PRECISION-RECALL TRADEOFF



For both the above metrics the optimal cutoff probability for conversion comes around 0.45

RESULTS

- Upon testing on the test data set following were observed
  - The accuracy of the model was 78.9%
  - The model was able to predict 76.8% of the true positives in each category
  - The model was able to predict 81.1% of the true negatives in each category
  - The precision of the model turned out be 79.1%

CONCLUSION

- People who spent more time on the website are likely to convert
- We see max number of leads are generated by google / direct traffic while the max conversion ratio is by reference and welingak website.
- Marketing management the Finance management have higher conversion rates so the company can focus on them more
- E-mails, SMS messages and alert messages can have high impact on lead conversion
- Most number of leads are generated for Unemployed people but they may be less likely to convert