Approach & Analysis

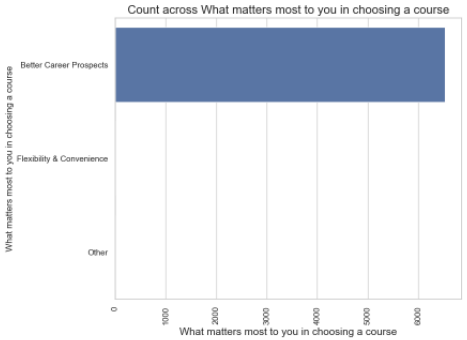
Problem Statement

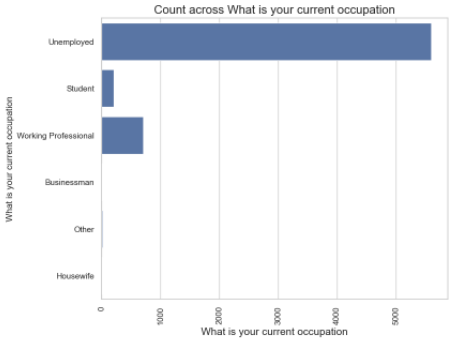
1. An education company named X wants to improve its lead conversion from the current 30% to around 80% by identifying the most potential leads.
2. The company wants to build a model wherein one need to assign a lead score to each of the leads such that the customers with a higher lead score have a higher conversion chance.

Analysis Approach

1. Import the data and inspect the data frame
2. EDA & Data Prep/Analysis
    - Prepare the Data (e.g. handling nulls etc.)
    - Dummy variable creation
    - Visualization
    - Correlation
    - Dummy Variables & Feature scaling
3. Model Building (RFE Rsquared VIF and pvalues)
4. Model Evaluation
5. Making predictions on test set

Visualization outcomes - 1

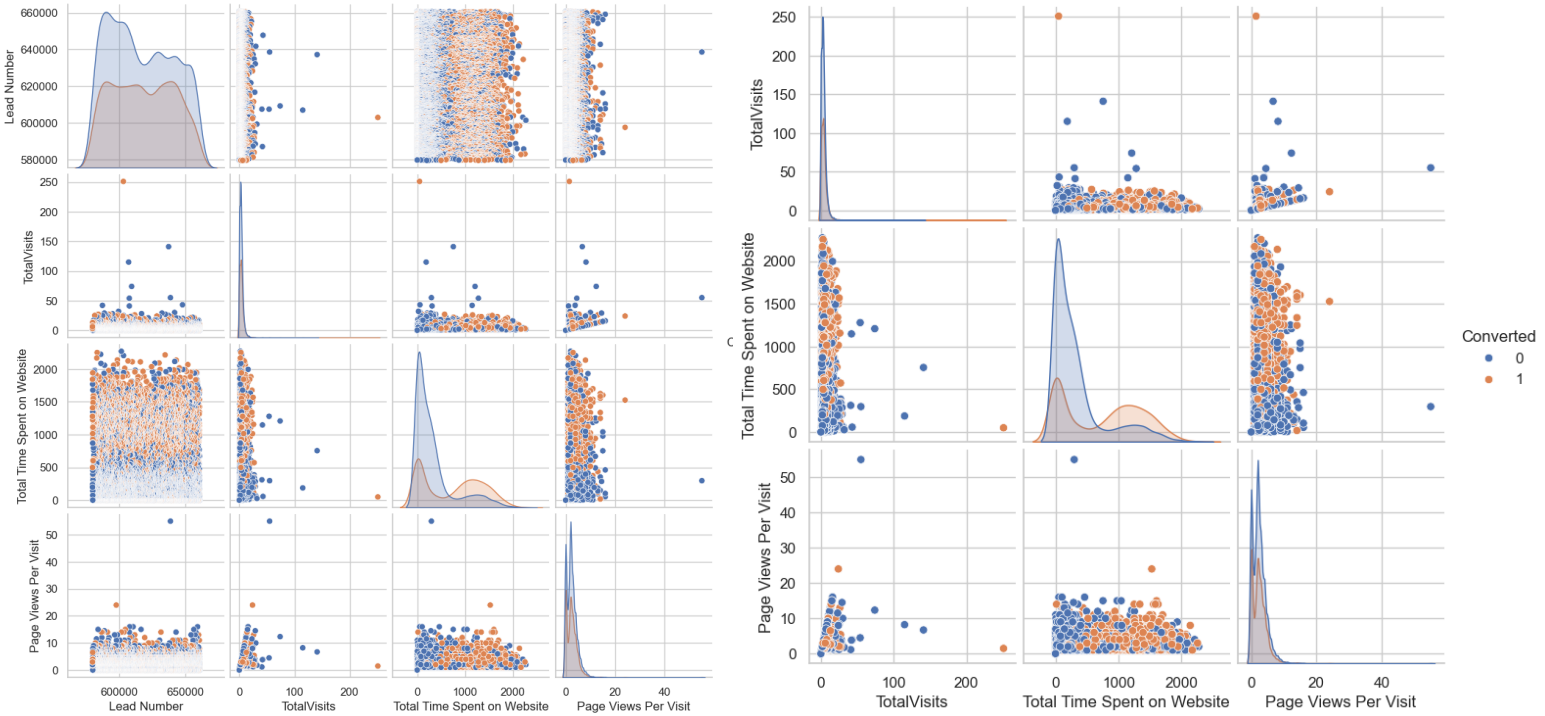




As it can be seen that the levels of "Lead Profile" and "How did you hear about X Education" have a lot of rows which have the value Select which is of no use to the analysis

So it's best that we drop them.

Visualization outcomes - 2

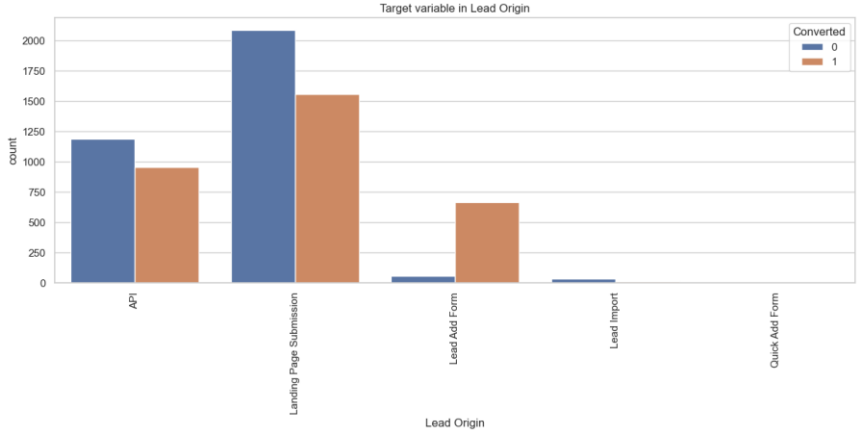


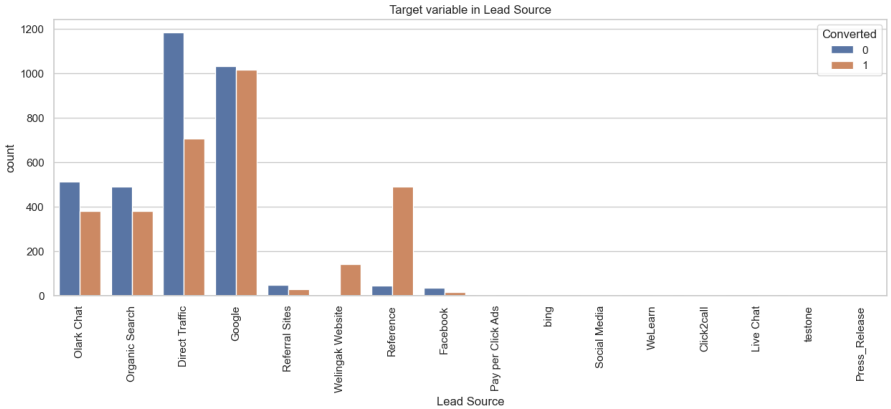
Columns that can be dropped: 'Do Not Call', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque'

Correlation (Quants)

Eventhough the Total Visits and Page Views per visit are showing a relationship, they are not that big. Besides, each tell a different story. Hence continue to include both in the modeling.

Categorical Features Visualization

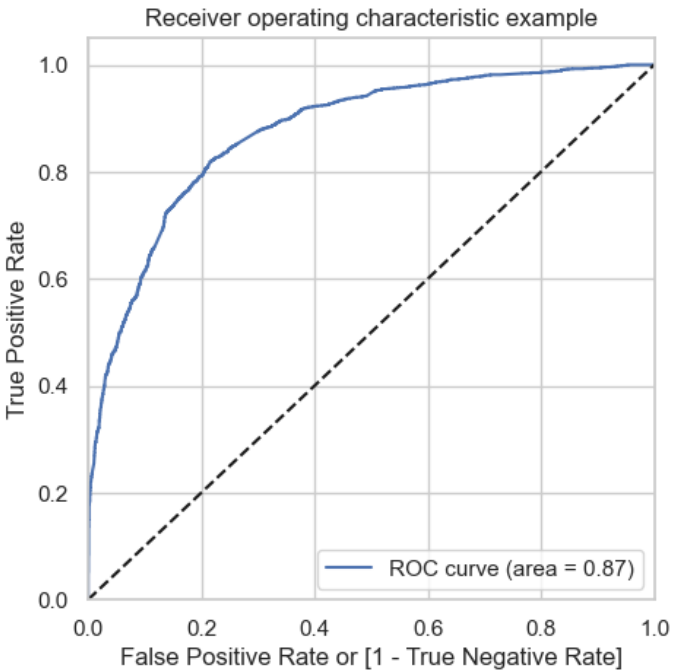




- High conversion leads are high on landing page submission.
- Leads through google and direct traffic has higher probability to convert.

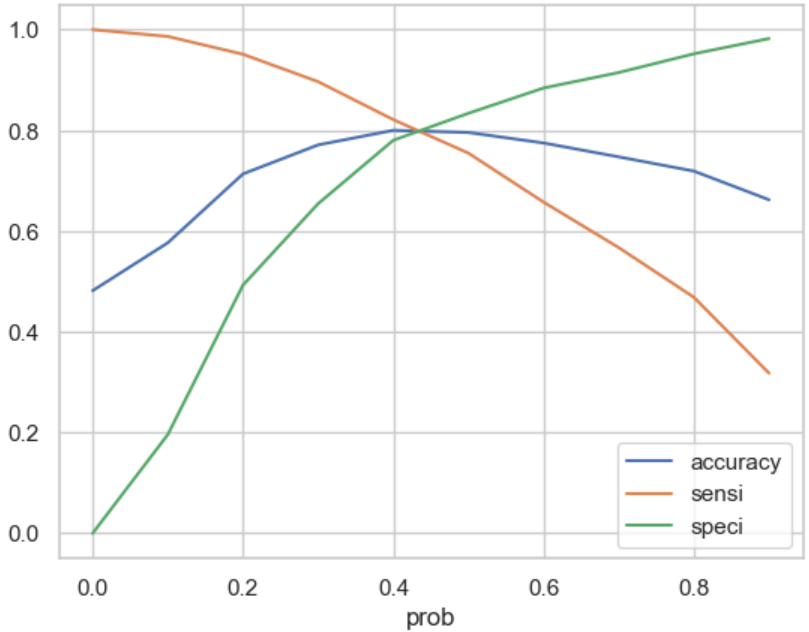
There are multiple other graphs, which can be seen in the jupyter notebook.

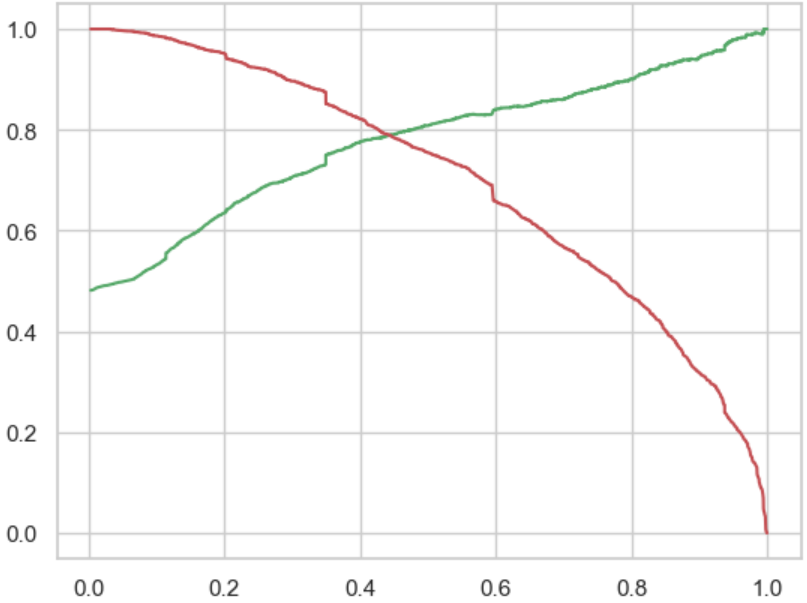
ROC Curve



- ROC Curve is decent
- Probability & Recall is also decent.

Cut Off





Around 0.42-0.44, you get the optimal values of the three metrics.

Results in business terms

ROC details are:

- ROC = 0.86
- Accuracy = 78%
- Precision = 77%
- Recall = 77%

Some of the top 3 variables are

- Total Visits
- Total Time Spent on Website
- Lead Source

Some of the top 3 categorical/dummy variables

- Lead Source
- Lead Origin
- Last Activity