Lead Score Case Study

Santoshi Rupa Bagadi

Problem Statement

- X Education sells online courses to industry professionals.
- X Education gets a lot of leads, its lead conversion rate is very poor. For eg., if 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:

- X Education wants to know most promising leads. For that, they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.

Approach

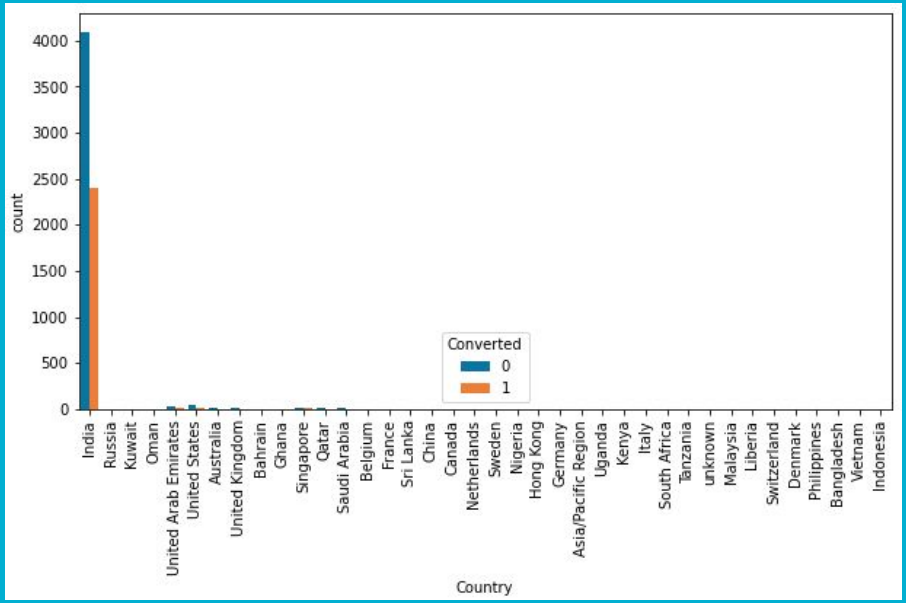
- Reading the dataset
- Cleaning the dataset
  - The null values in the data are first cleaned. Columns with more than 40% null values are removed.
  - Remaining null values in the columns are imputed with the mean and median for numerical columns and mode for categorical columns.
  - The columns which have "Select" are also replaced by null values.
- EDA:
  - A number of irrelevant categorical columns are found in the data on performing a basic EDA.
  - Numerical columns are good with almost no outliers.

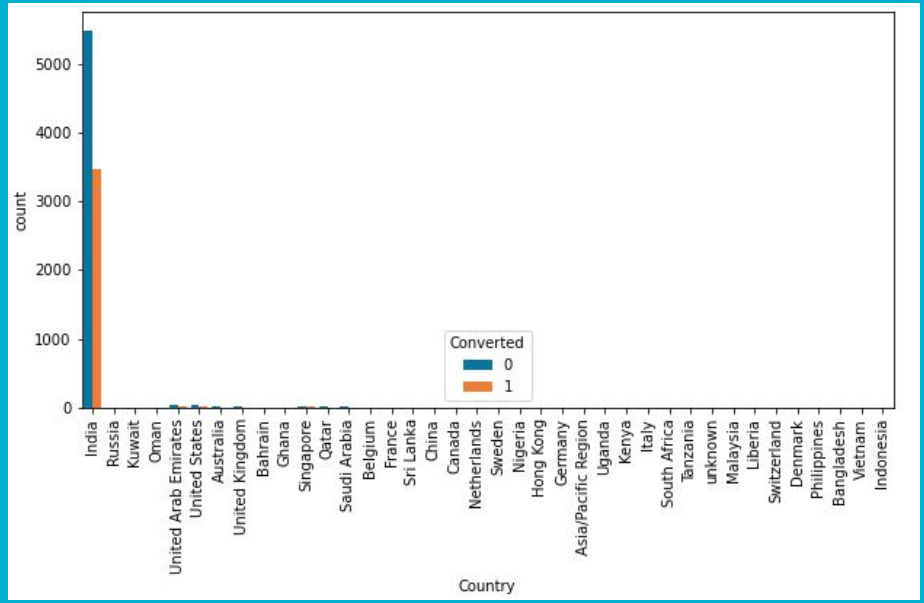
Approach

- Creating Dummy variables:
  - Dummy variables are created for categorical columns.
  - For numerical columns Standardscaler is used.
- Train - test split:
  - The data was split into train and test data in 70:30 ratio.
- Model building:
  - Using RFE we selected the most relevant 15 variables.
  - ELiminating features manually using the VIF and p-values. Features with vif <5 and p-value < 0.05 are kept.
- Prediction:
  - Prediction was done on the test data
- Precision and Recall.

EDA

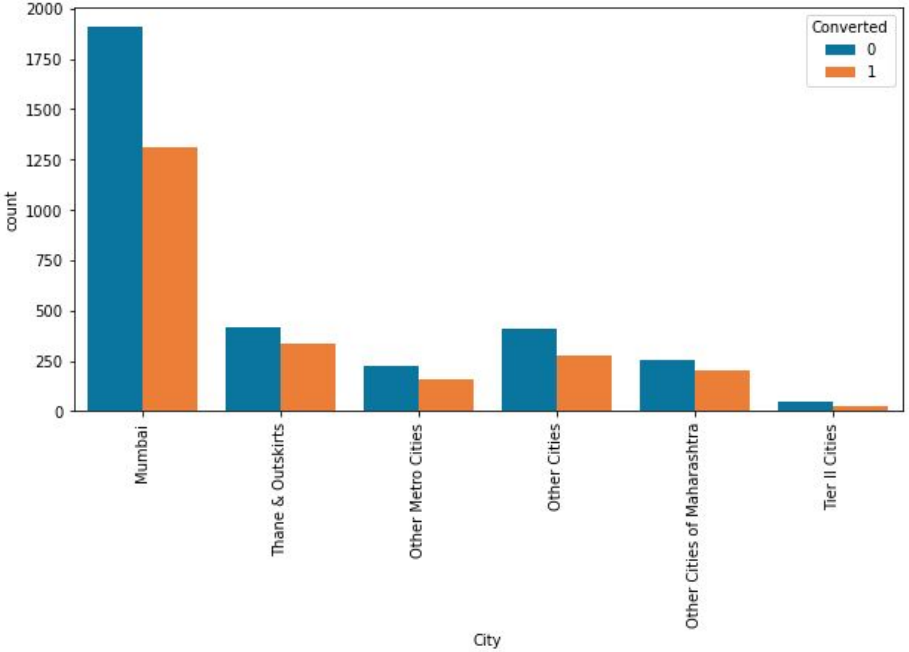
Categorical Columns - Country

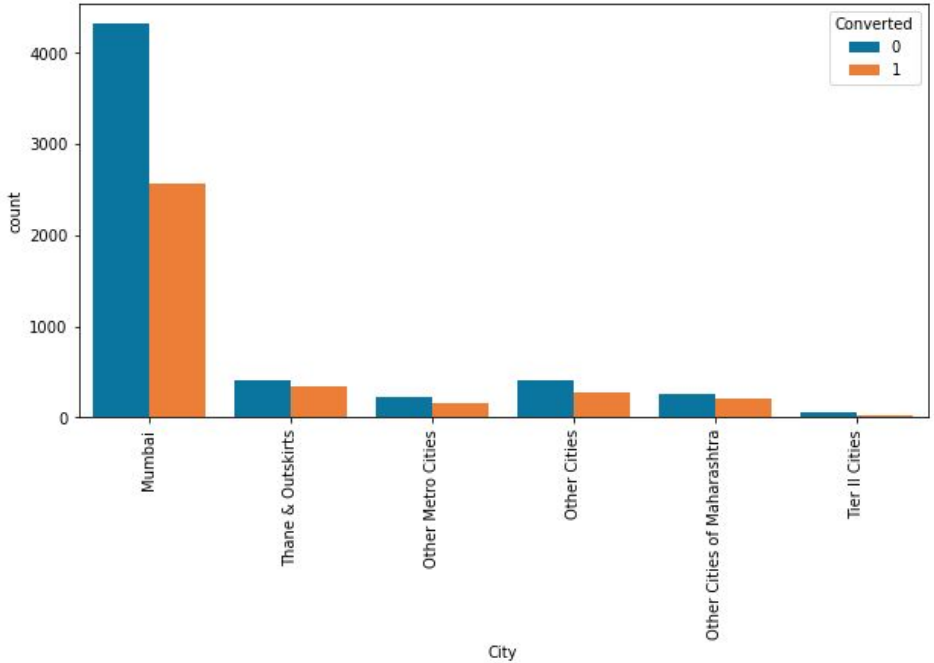




EDA

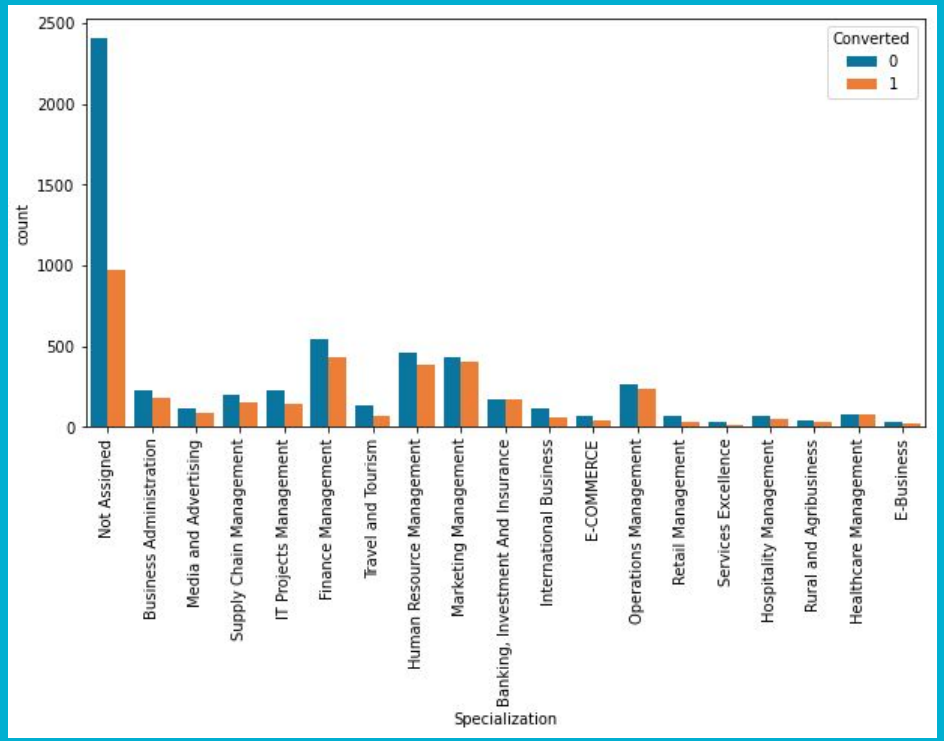
Categorical Columns - City

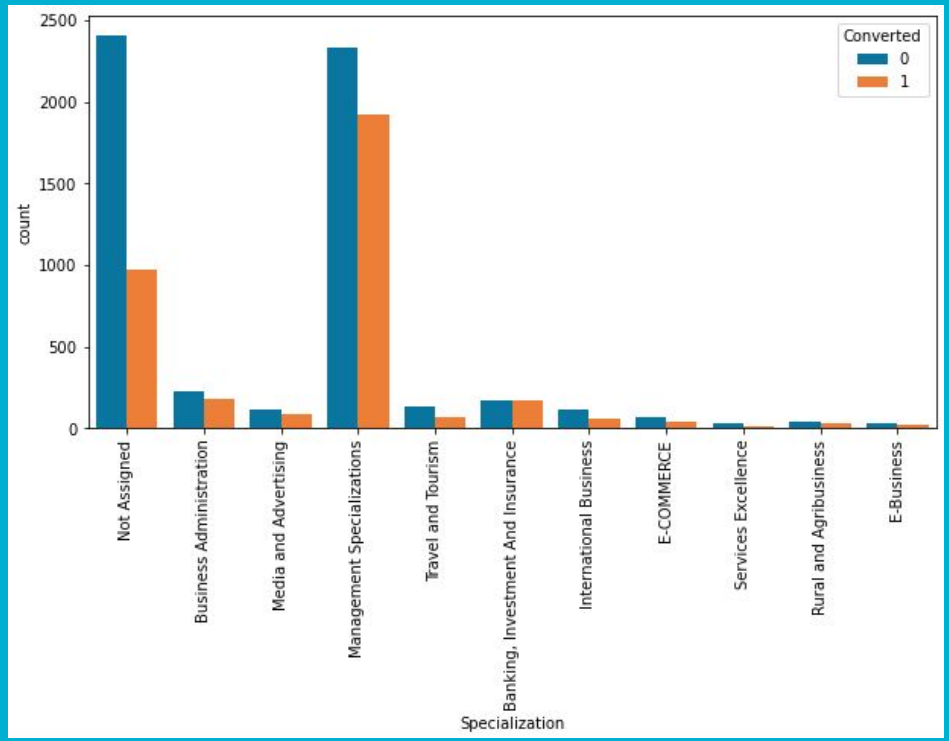




EDA

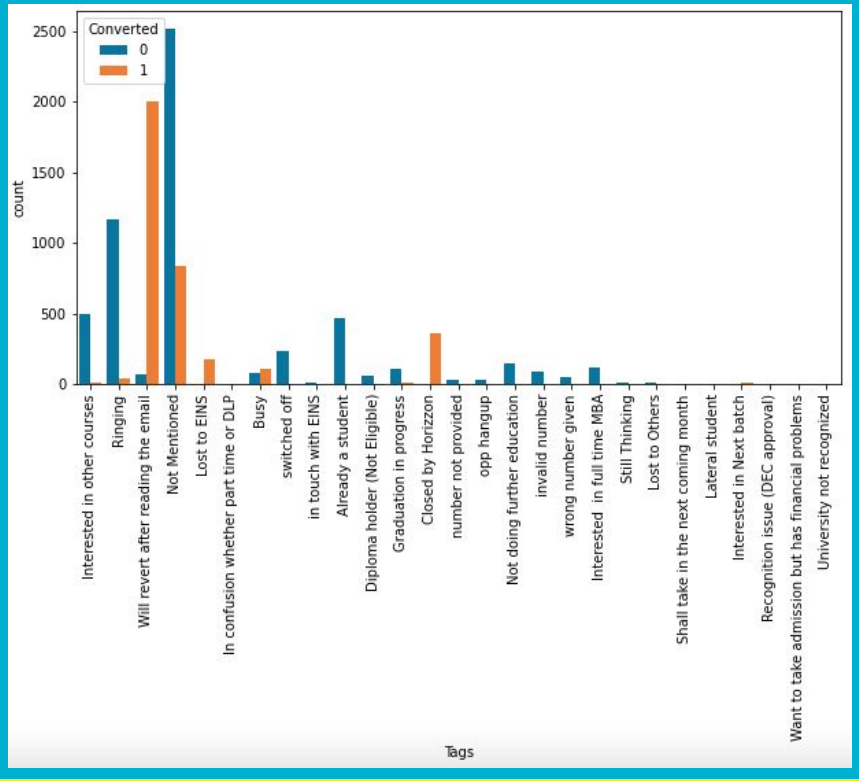
Categorical Columns – Specialization

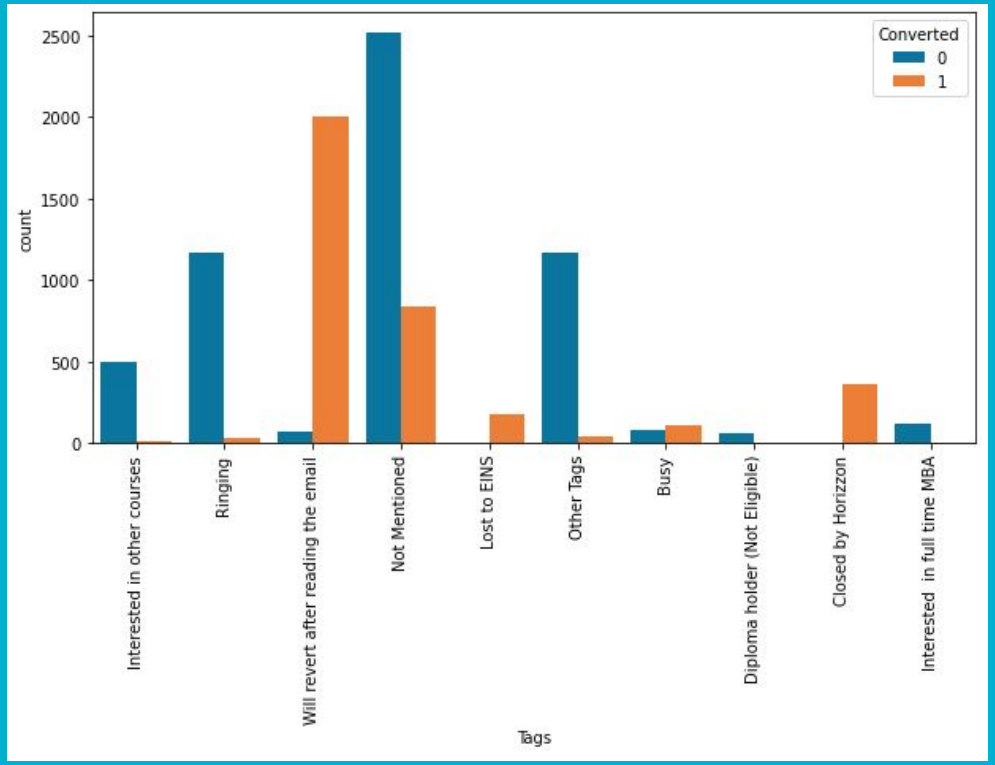




EDA

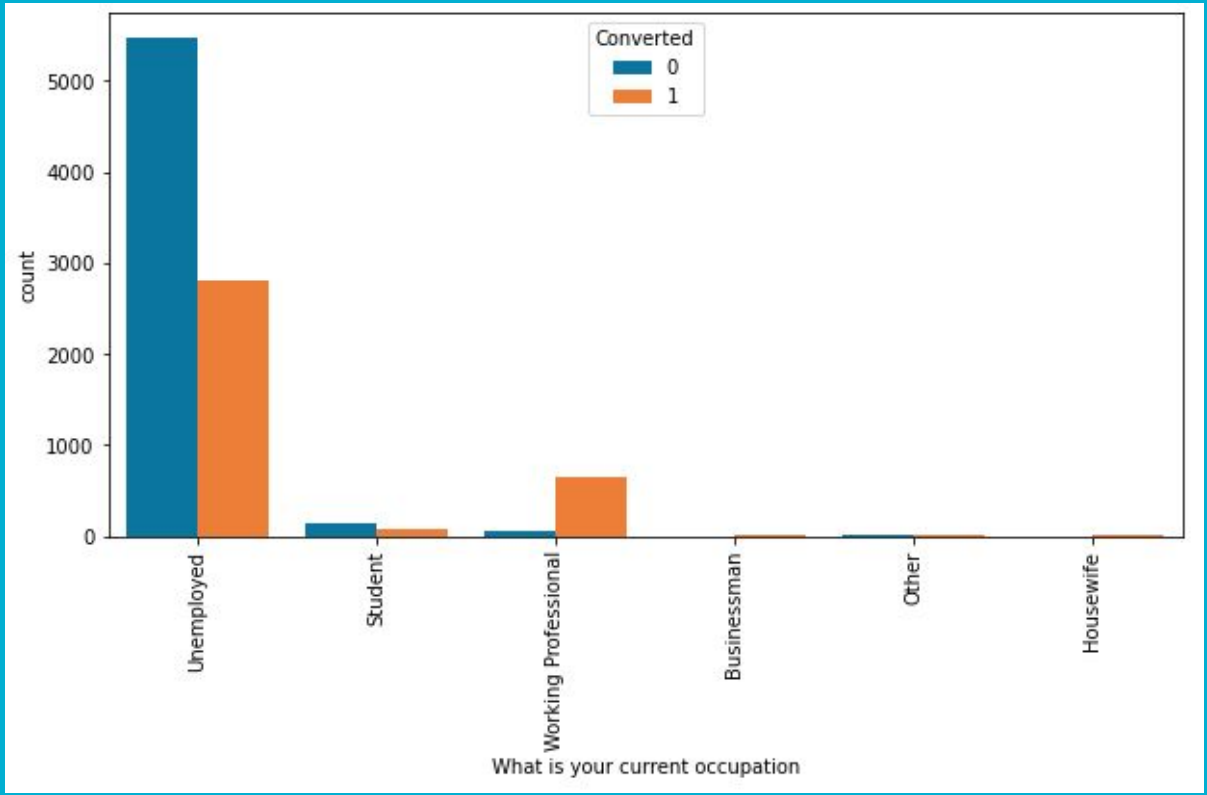
Categorical Columns - Tags





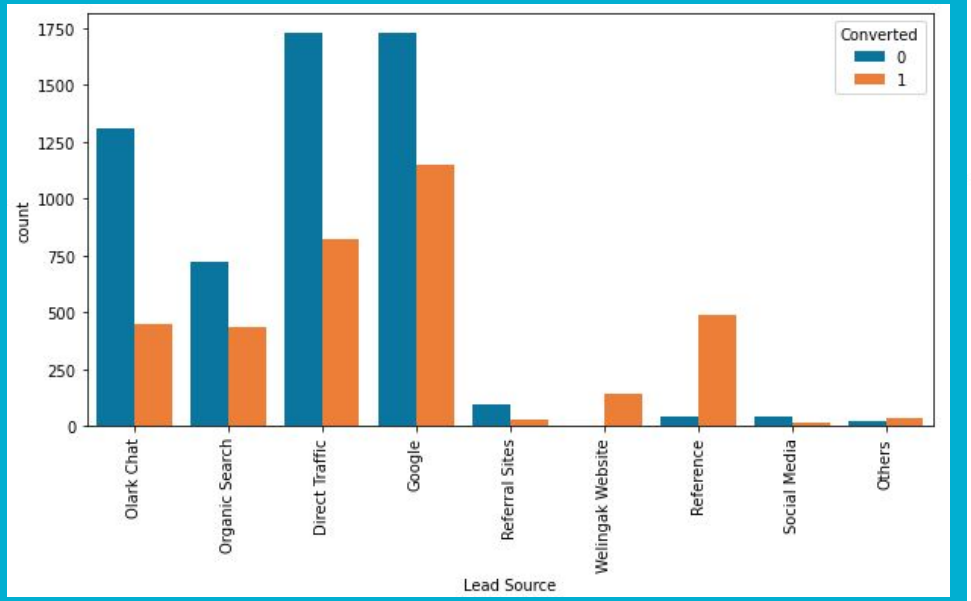
EDA

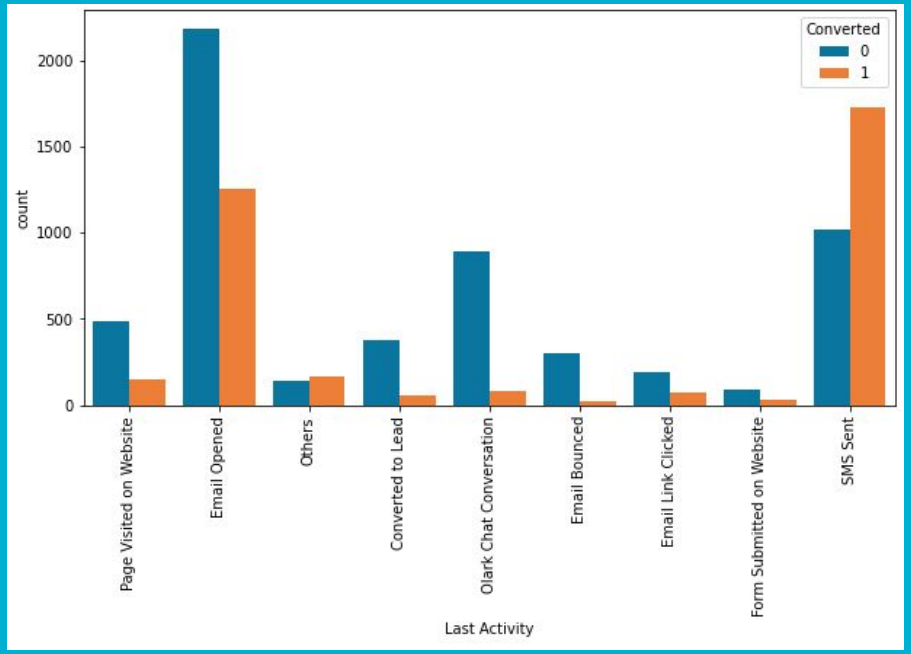
Categorical Columns – Current Occupation



EDA

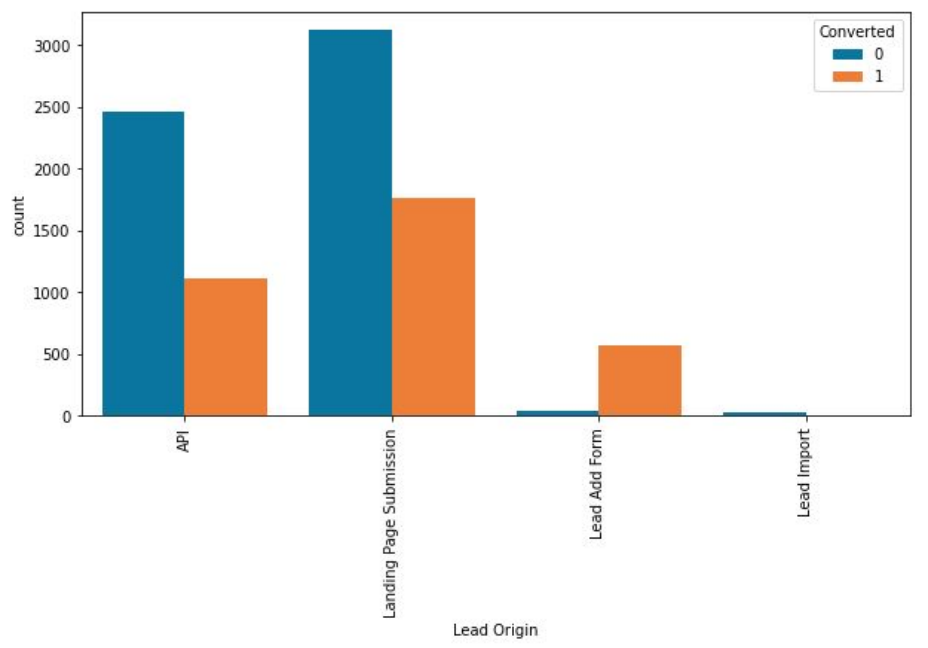
Categorical Columns - Lead Source & Last Activity

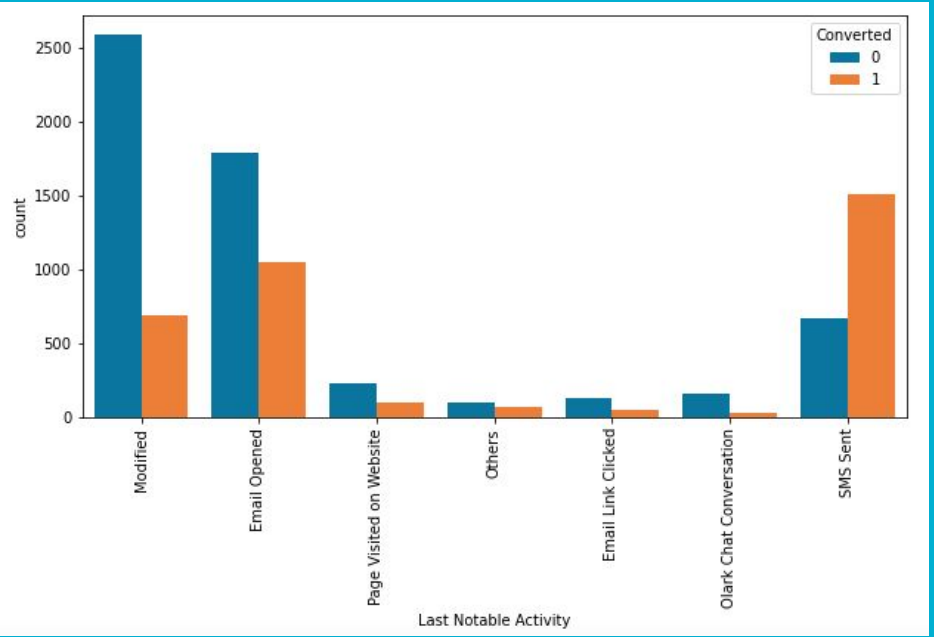




EDA

Categorical Columns – Lead Origin & Last Notable Activity





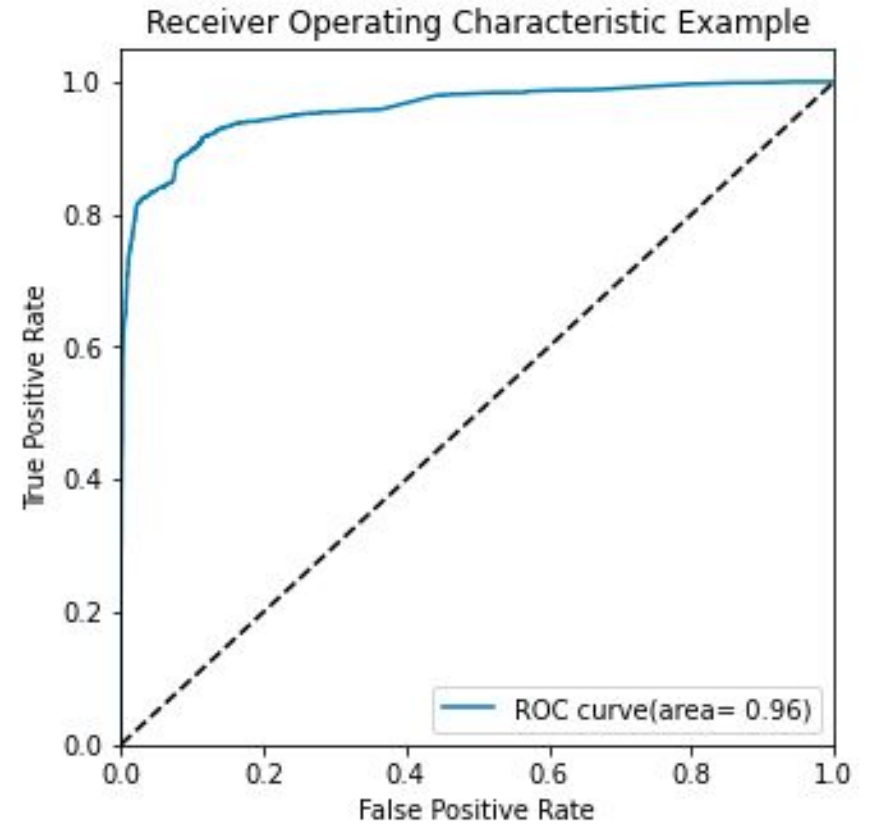
EDA

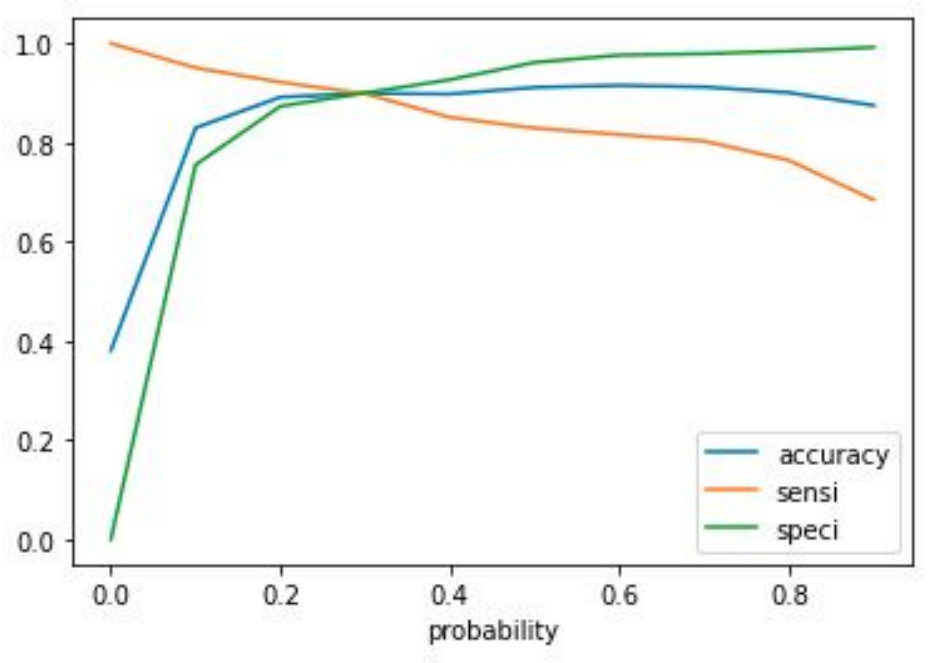
Numerical Columns: Correlation among numerical variables

Model Building

- Splitting the data into Train and Test sets
- The Train-Test split is done in 70:30 ratio.
- Using RFE, selected 15 relevant features.
- Building model by removing the variables whose p-value is greater than 0.05 and VIF value greater than 5.
- Performed predictions on the test data set.
- Overall accuracy: 89.91 %

ROC Curve


Receiver Operating Characteristic Example
True Positive Rate
False Positive Rate
ROC curve(area= 0.96)
0.0
0.2
0.4
0.6
0.8
1.0


accuracy
sensi
speci
probability
0.0
0.2
0.4
0.6
0.8
1.0

Conclusion

- The Variables that contribute to the probability of a lead to get converted are:
  - Lead Origin_Lead Add Form
  - Total Time Spent on Website
  - Last Notable Activity_SMS Sent
  - Tags_Will revert after reading the email
  - Last Notable Activity_Modified
  - Tags_Ringing
- X Education has to take these features into consideration and flourish as they have a very high chance to get almost all potential buyers to change their mind and buy their courses.