Lead Score Case Study

Group Members
1.Roslin
2.Roshan
3.Pavithra

Problem Statement

- X Education, a provider of online courses to industry professionals, is facing a challenge with its lead conversion rate.
- The company's objective is to improve efficiency by identifying 'Hot Leads,' the most promising prospects.
- X Education plans to build a predictive model to identify these 'Hot Leads' effectively.
- By concentrating the sales team's efforts on these predicted 'Hot Leads,' the company aims to enhance its lead conversion rate and overall sales performance.

Solution Approach

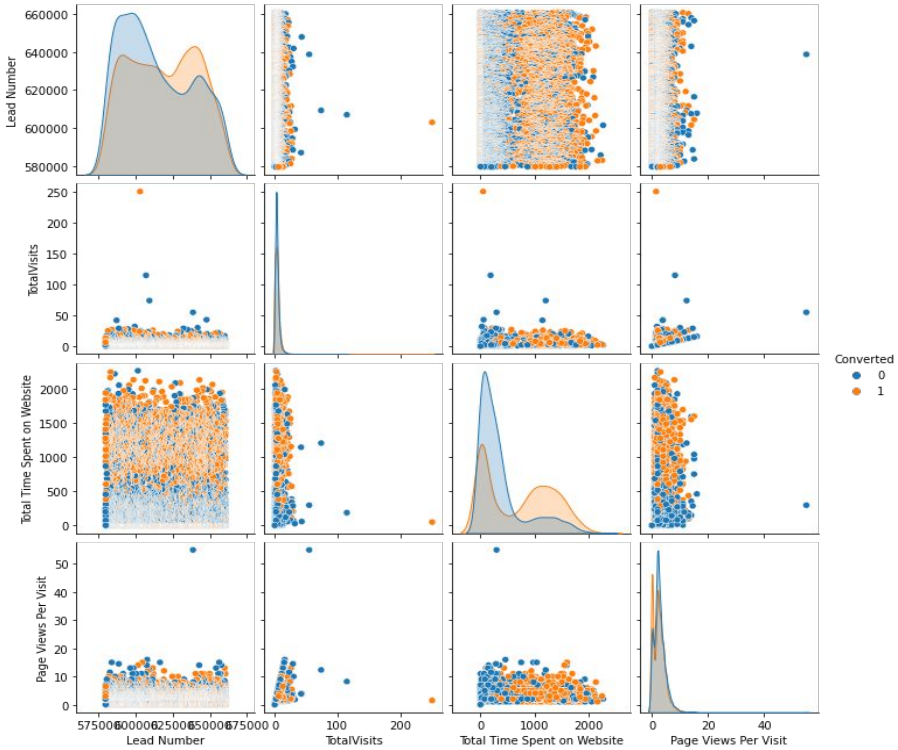
1. Data Cleaning
2. EDA
3. Data Preparation
4. Building Model
5. Conclusions

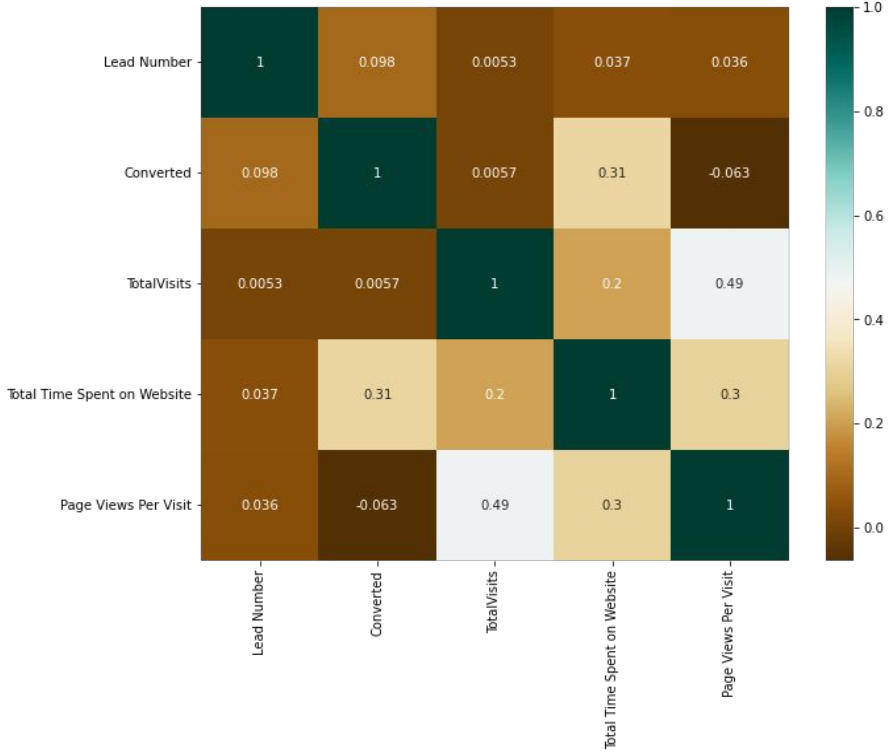
Data Cleaning

1. By looking at the null percentages, columns with high nulls (>30%) have been removed. For other columns, rows with null values have been removed.
2. Categorical columns with high proportion of irrelevant data (eg. Select, NaN) have been removed
3. Columns with a majority of a single value have been dropped

EDA (1/3)

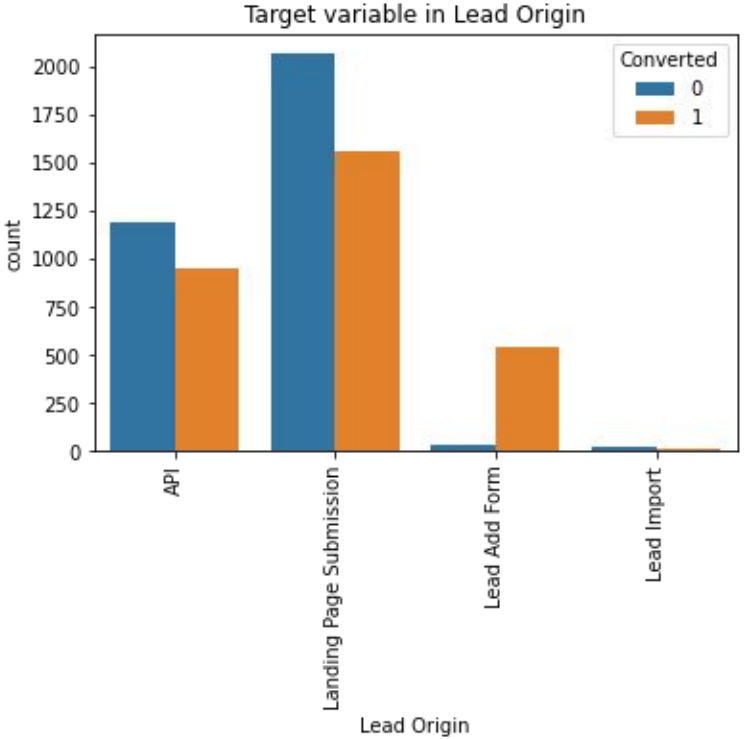
1. Pair plot and Heat Map are plotted to find the correlation between variables.
2. ‘Total time spent on the website’ has high correlation with ‘conversion’.

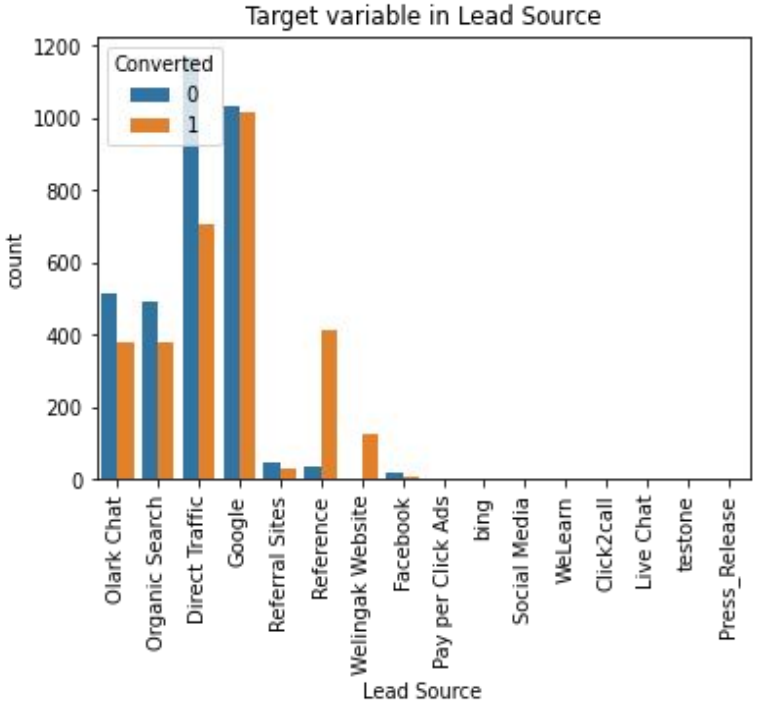




EDA (2/3)

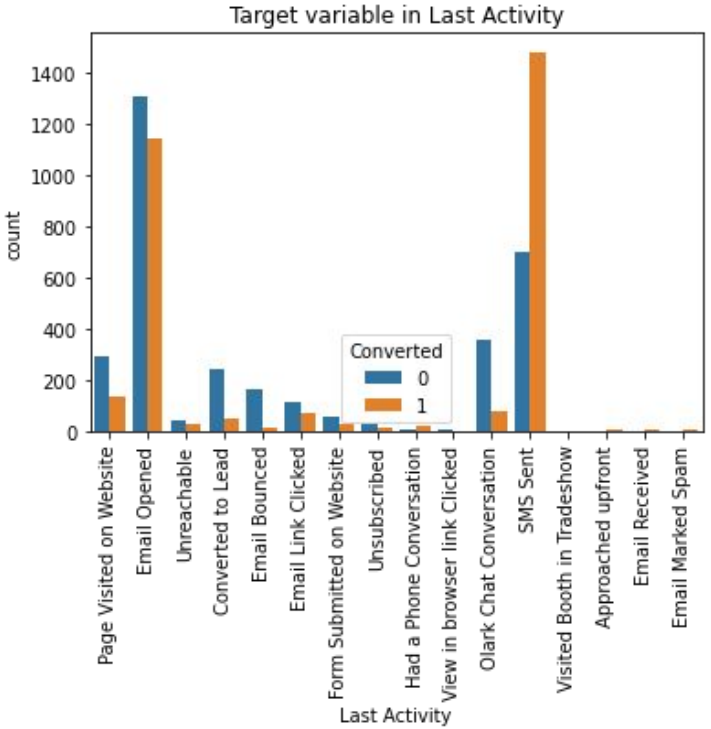
3. The conversion rate is high from the ‘Lead Add Form’ within Lead origin and from ‘References’ within lead source

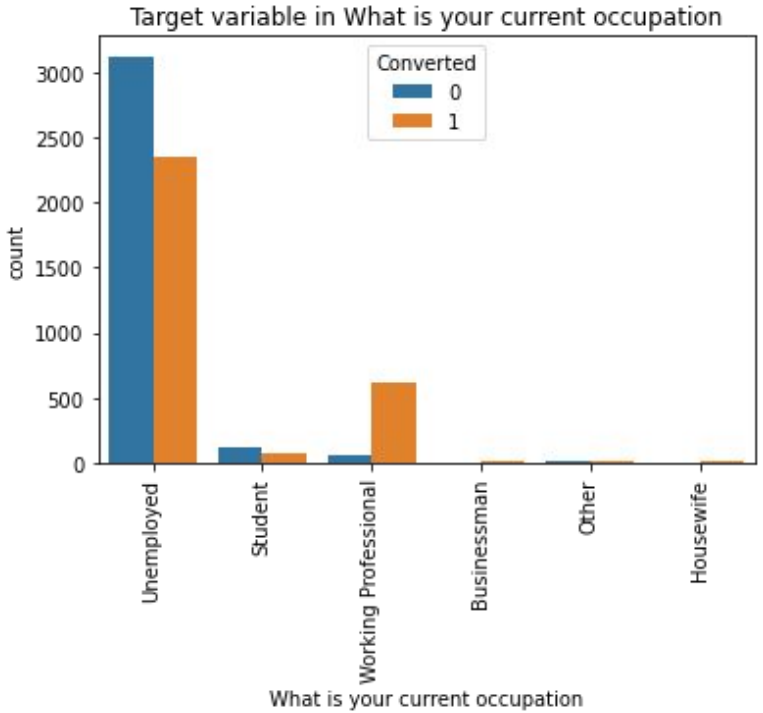




EDA (3/3)

4. Conversion rate is high when the candidate's last activity is 'SMS sent'. It is also very high when the applicant is a 'Working professional'





Data Preparation

1. Dummy variables are created for columns whose data type is Object such as:
    a. Lead Origin
    b. Lead Source
    c. Do Not Email
    d. Last Activity
    e. Specialization
    f. What is your current occupation
    g. A free copy of Mastering The Interview
    h. Last Notable Activity

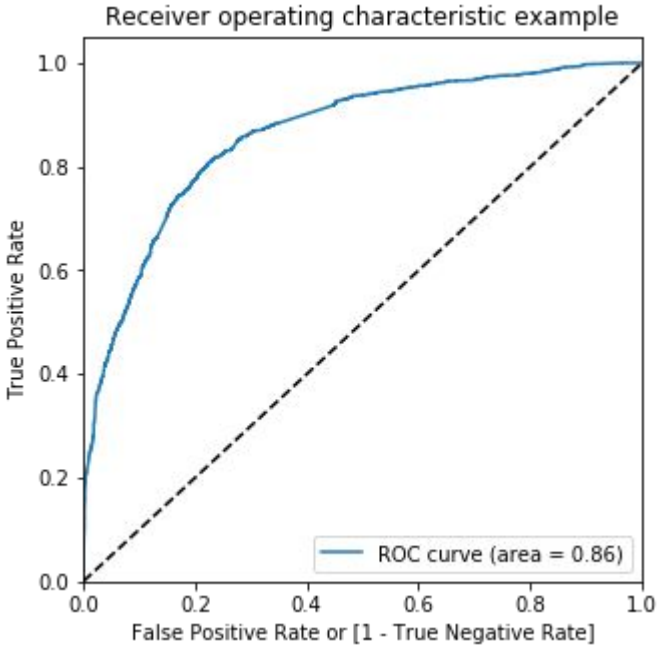
2. After creating the Dummy Variables the number of columns went upto 75.

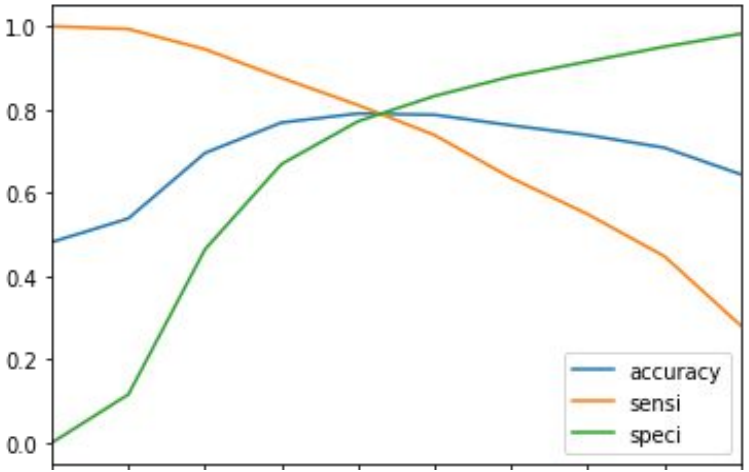
Building Model(1/2)

1. The first step for regression is splitting of data into Training and Test Sets. We have chosen 70% of data as Training set and remaining 30% of data as Test set.
2. Numeric variables present in the dataset are scaled using MinMaxScaler.
3. RFE is used for feature selection. The RFE is runned by taking 15 variables as output. This gives the top 15 features for model building.
4. Building Model by removing the variable whose p-value is greater than 0.05 and VIF value greater than 5.
5. The model is evaluated by creating a confusion matrix and calculating other metrics such as overall accuracy, sensitivity, specificity etc.

Building Model(2/2)

6. ROC curve is plotted as below
7. Predictions made on the test data set resulted in overall accuracy of 78% for the model





Conclusions

1. For both Optimal and Precision Evaluation Method, Accuracy is 78%
2. The consistency in Evaluation values gives confidence on the Model
3. Variables - 'TotalVisits', 'Total Time Spent on Website','Lead Origin_Lead Add Form', 'Lead Source_Olark Chat','Lead Source_Welingak Website' are directly proportional to the model performance