Brief Summary

Lead Scoring Case Study

Steps followed:

- Data Reading and Understanding
- Data Cleaning
- Data Transformation
- Data Analysis
- Data Preparation
- Building Logistic Regression model and calculating Lead score
- Model Evaluation

1. Data Reading and Understanding: First step was to load the given dataset to the jupyter file and analyse the data like shape of the dataset, datatype of the columns, and some statistical info about the data like mean, mode, media, outliers.

2. Data Cleaning - It was observed that there were some redundant columns in the dataset that we decided to remove.

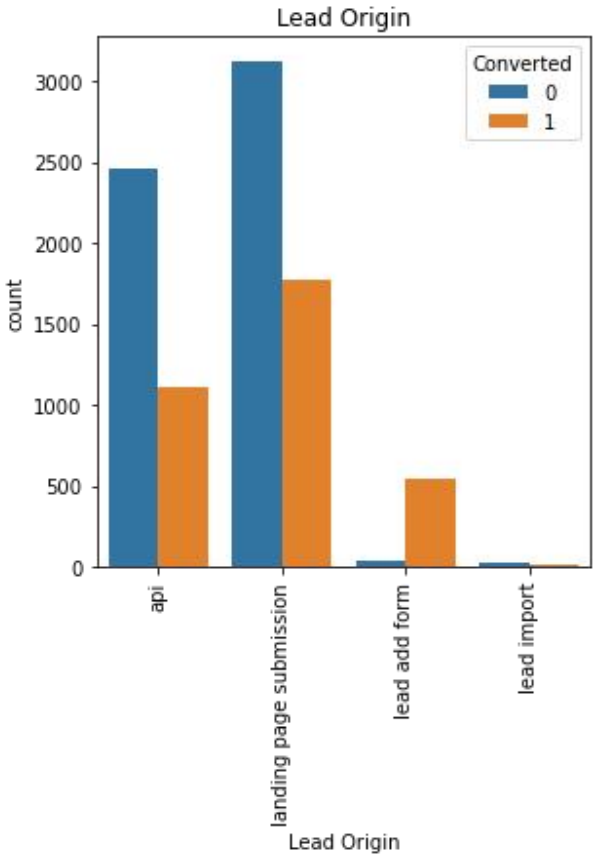
- There were some columns that were having a 'Select' label which showed that the customer didn't select any option. It was better to put it as null value because there were no suitable options present to select for the customer searching for.
- Outliers were observed in two columns which were handled by upper capping them due to the nature of the data.
- We removed the columns having missing values more than 35%.
- For the remaining categorical columns having missing values we replaced them using the mode value.

- Two columns had identical names which were taken care of by changing the column name into one format.

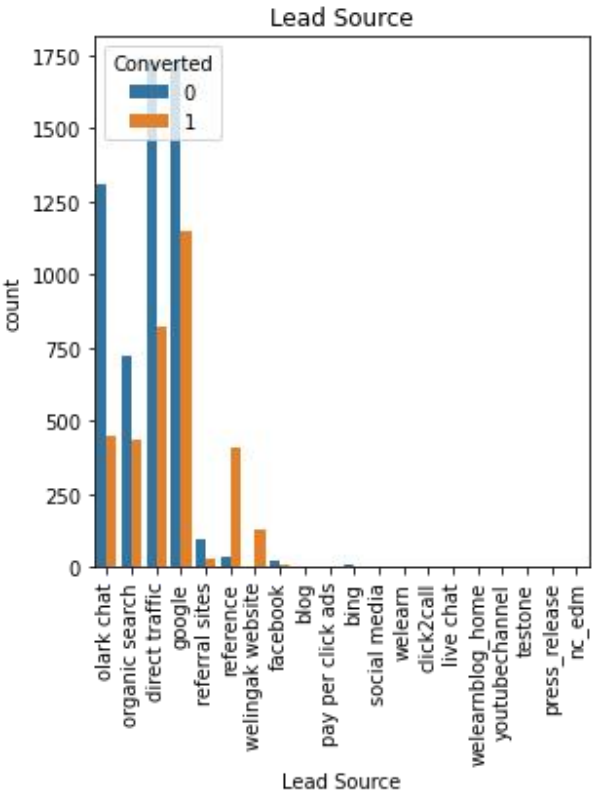
- After outlier treatment and further analysis, we decided to impute missing values in the numerical columns by their respective modes due to the nature of data.

3)Data Analysis

A: Lead origin



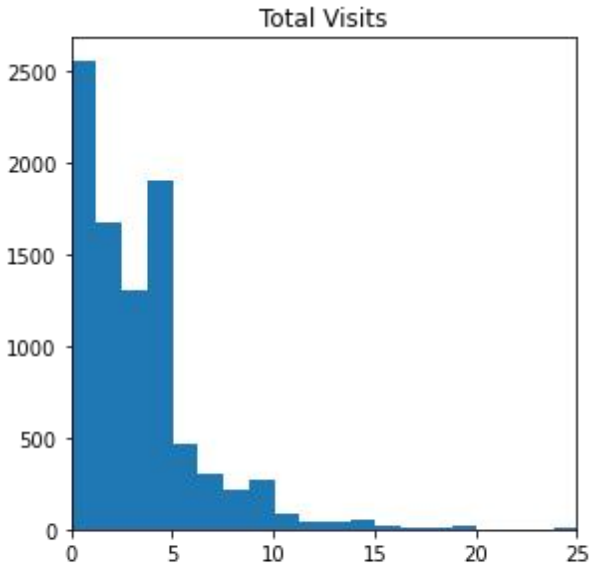
B: Lead source

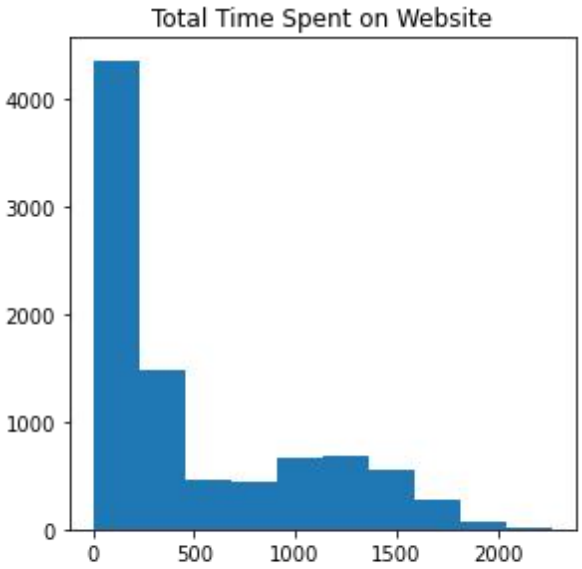


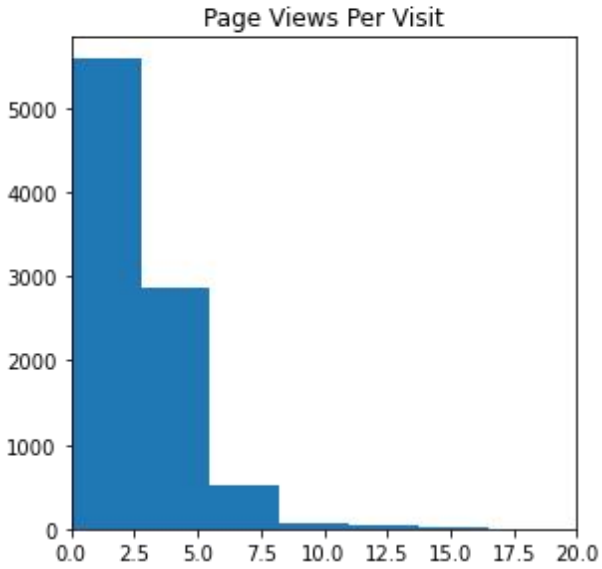
From the graphs : a)Major conversion in the lead origin is from Google.

b)Maximum conversion happened from Landing Page Submission.

NUMERICAL VARIALBLES







Data Preparation:

Dataset was split into train and test data.

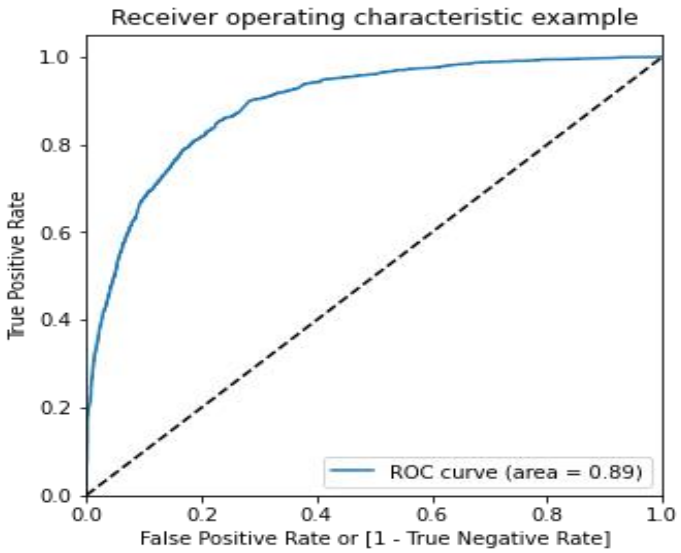
MinMax scaling of the data was done for further modelling

Building Logistic Regression Model and calculation of Lead score:

In the model obtained by logistic regression it was observed that many variables have high p-values. For the feature selection we used RFE as the number of variables are quite high and individually checking them was not efficient.

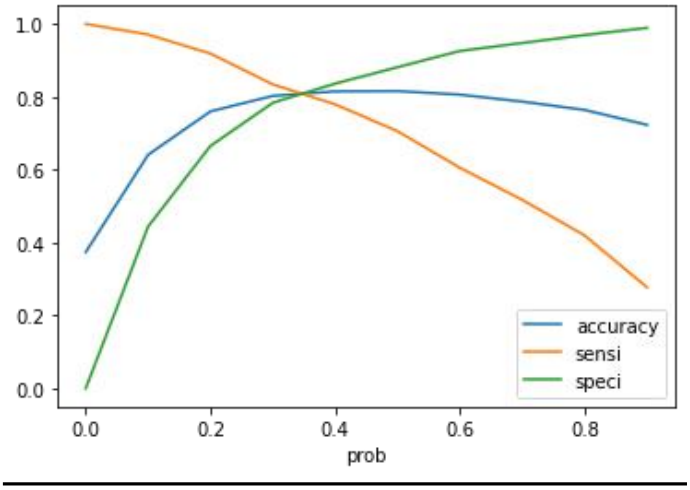
After RFE was done all the columns based on their ranking were selected and again modelling was done. All the features with p-value greater than 0.05 were dropped one by one and the model was built repeatedly.

ROC Curve



The curve is closer to the left side of the border than to the right side hence our model is having great accuracy. The area under the curve is 89% of the total area.

Probability Cutoff Point



Probability cutoff point was at around 0.35 as this is where the sensitivity, accuracy and specificity converged.

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

1. The total time spend on the Website.
2. Total number of visits.
3. When the lead source was:
    (a) Google
    (b) Direct traffic
    (c) Organic search
    (d) Welingak website
4. When the last activity was:
    (a) SMS
    (b) Olark chat conversation
5. When the lead origin is Lead add format.
6. When their current occupation is as a working professional.
7. Keeping these in mind the X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their courses.