Lead Score Case Study

Tripurana saratchandra das

Problem Statement

1. Industry professionals can purchase online courses from X Education.
2. X Education has an extremely low lead conversion rate despite receiving a lot of leads.
   For instance, only roughly thirty of the 100 leads they obtain each day are converted.
3. The organization wants to find the most promising leads, or "Hot Leads," in order to streamline this procedure.
4. The lead conversion rate should increase if they are able to locate this group of leads since the sales staff will now be spending more time corresponding with the prospects rather than calling everyone.

The following technical steps are used:-

1.Data Cleaning:

- As a first step in cleaning the dataset, we decide to eliminate any redundant features or variables.
- The option "Select" has to be replaced with a null value because it did not provide us with much information, and the data set was mostly clean aside from a few null values.
- Removed the high null value proportion of above 40%.
- Determined how many distinct categories there were in each category column.

2. Exploratory Data Analysis

3.Dummy Variables

4.Scaling

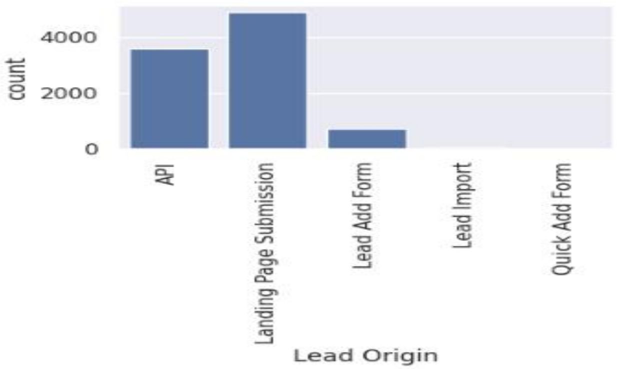
5.RFE-Based Feature Selection

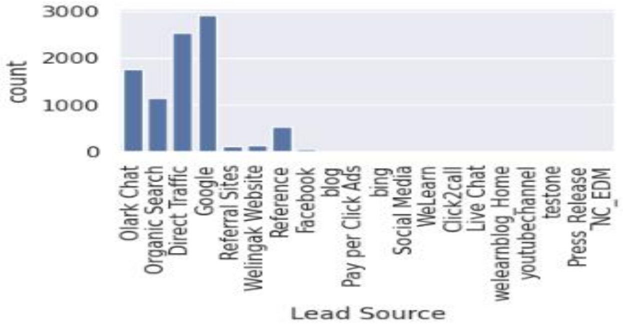
6.Model Building

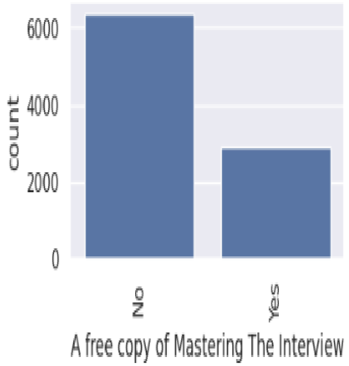
7.Test Set Prediction

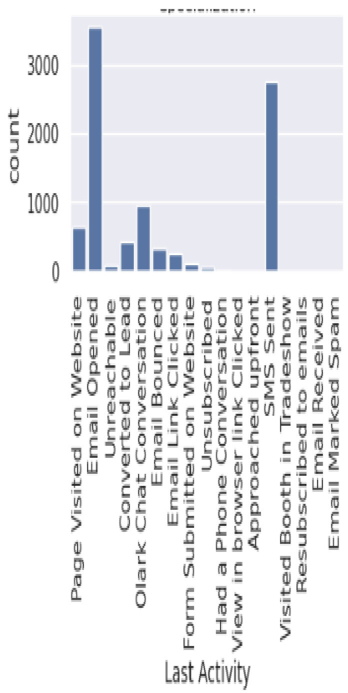
EDA

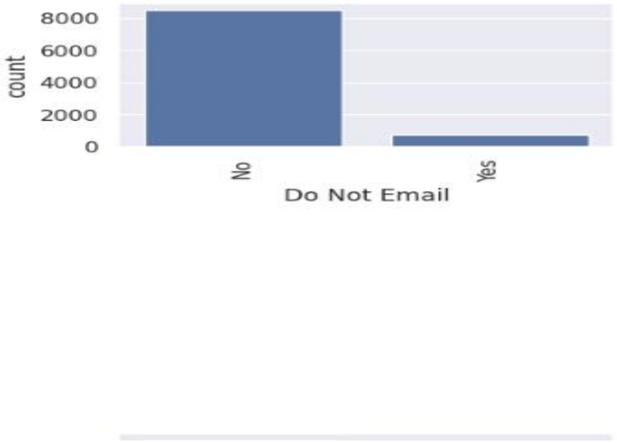
- Univariate Analysis (Categorical)

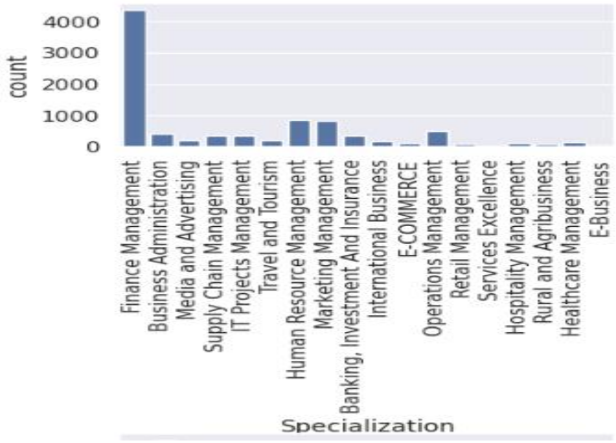




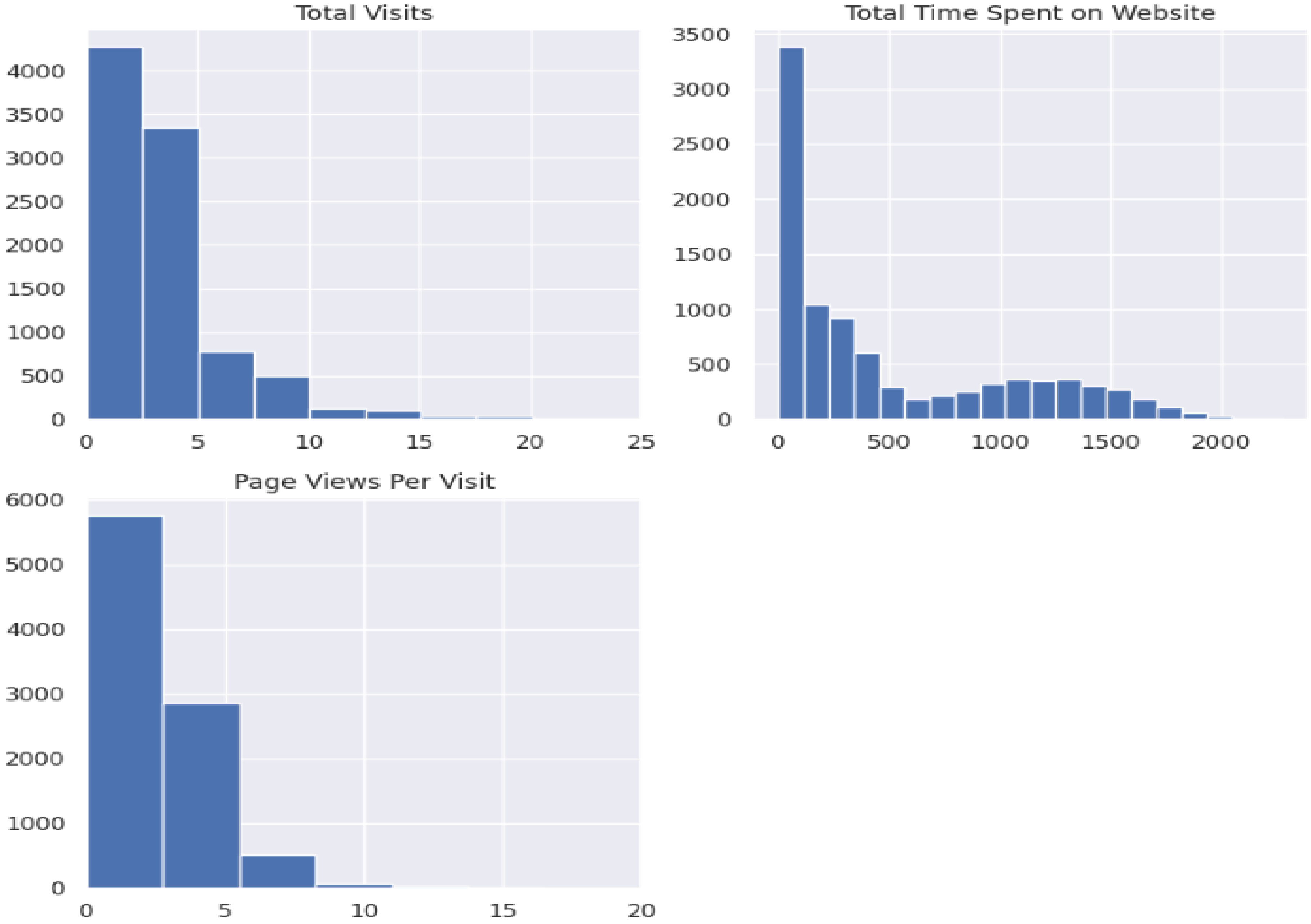




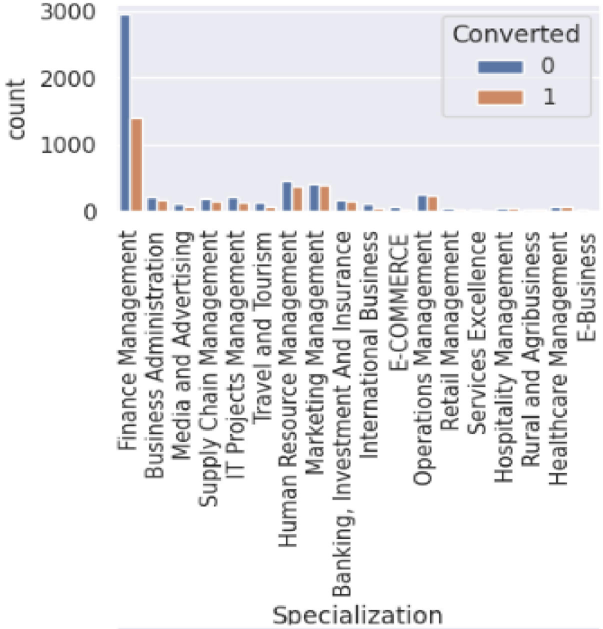


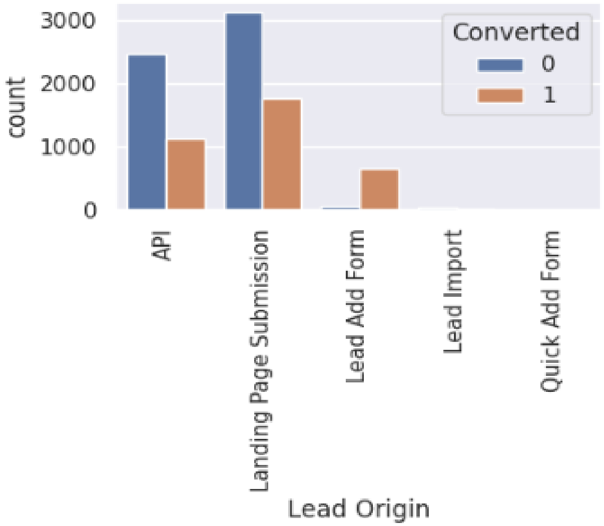


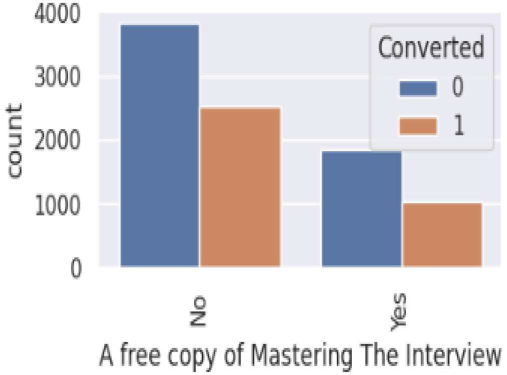
- Univariate Analysis(Contenious)

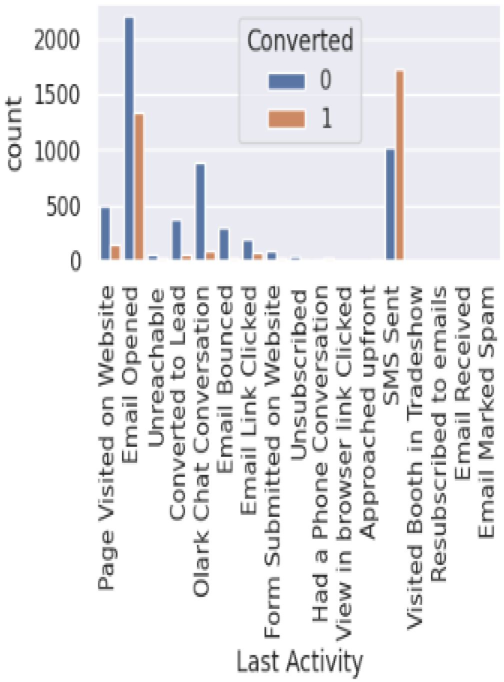


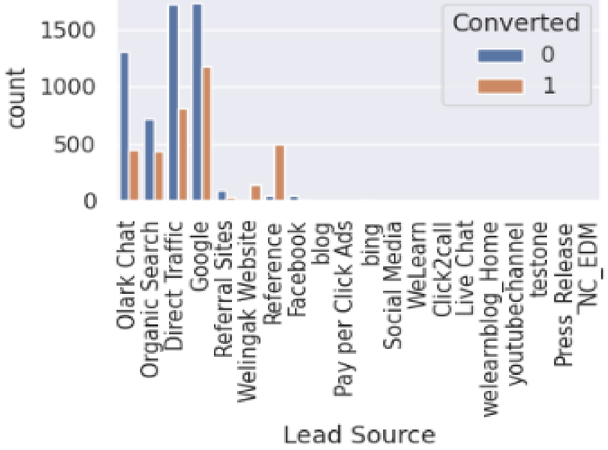
- Bivariate Analysis(with respect to 'converted')

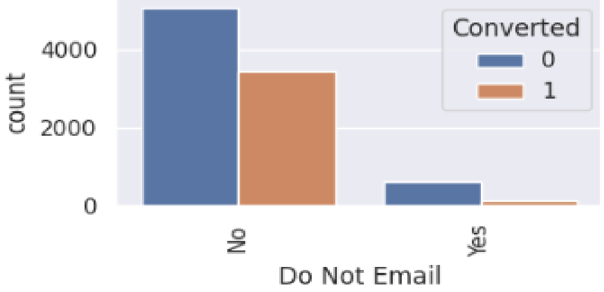












Data Preparation

1.Dummy Variables:

- For each of the categorical columns, dummy variables are produced.

2.Scaling:

- Scaled the data for continuous variables using standard scalar.

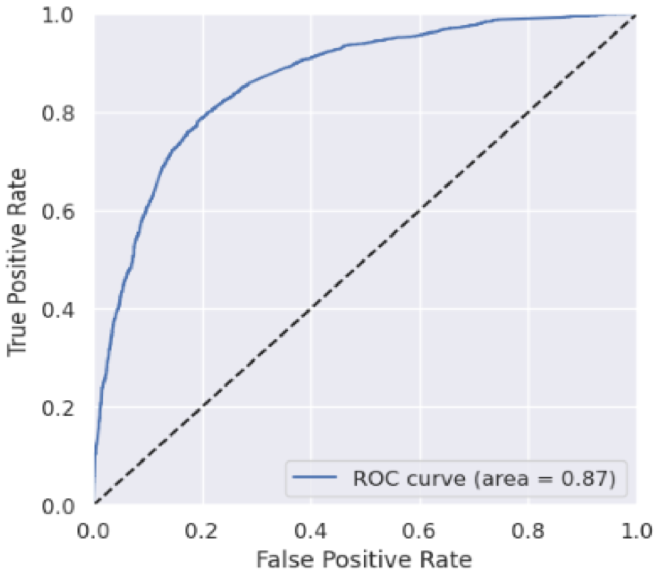
3.RFE-Based Feature Selection:

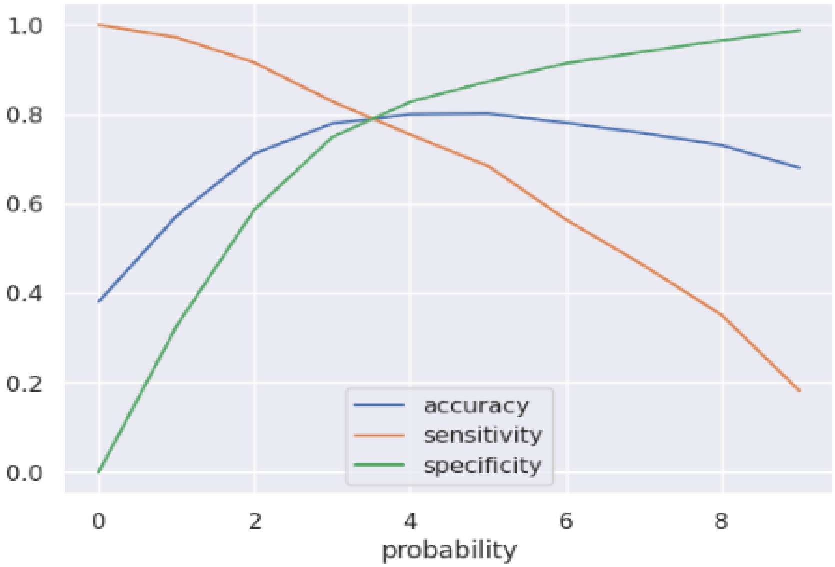
- By using RFE with provided 20 variables. It gives top 20 relevant variables.

Model Building:

- Subsequently, the variables with a VIF less than 5 and a p-value of 0.05 were retained, while the unnecessary characteristics were manually eliminated based on the VIF values and p-value.

Model Evaluation





Finding Optimal Cut off Point :

Probability where we get balanced sensitivity and specificity. From the second graph it is visible that the optimal cut off is at 3.8

Test Set Prediction:

- A confusion matrix was made. Subsequently, the accuracy, sensitivity, and specificity were determined by utilizing the ROC curve to determine the ideal cut-off value, which was approximately 80%.

Conclusion

It was found that the variables that mattered the most in the potential buyers are:

- The total time spend on the Website.
- Total number of visits.
- When the lead source was:

a. Google b. Direct traffic c. Organic search d. Olark chat

- When the last activity was:

a. SMS b. Olark chat conversation

- When the lead origin is Lead add format.