Lead Score Case Study

Goals for the case study:

a. Build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted.

b. There are some more problems presented by the company which your model should be able to adjust to if the company's requirement changes in the future so you will need to handle these as well. These problems are provided in a separate doc file. Please fill it based on the logistic regression model you got in the first step. Also, make sure you include this in your final PPT where you'll make recommendations.

Provided data is Leads.csv

a. It contains 1940 rows and 37 columns in which 7 columns are numeric and 30 are categorical.
b. Conversion rate of the provided data is 39%.
c. Leads Data Dictionary.csv is also provided which is the metadata of the main dataset.

Data Cleaning and Preparation

1. We replace ‘select’ with nan since select is default value for some columns which is equivalent to null values.
2. Drop all the unnecessary columns
    a. Asymmetrique_Profile_Index
    b. Asymmetrique_Activity_Index
    c. Asymmetrique_Activity_Score
    d. Asymmetrique_Profile_Score
    e. Lead_Profile
    f. Tags
    g. Lead_Quality

h. How_did_you_hear_about_X_Education
i. City
j. Lead_Number

3. Now, we check the unique values in the columns and drop the columns containing just 1 unique value.

Conclusion

After cleaning the provided data and creating the logistic regression model we use RFE method to choose the final variables to work with. After training the model on the selected data we test the same on Test data which gives us approximately 81% of accuracy.