Lead Score Case Study Assignment

Summary Report

N T Shrish Surya

Megha Sugunan

Annepu Supraja

The motive behind the assignment is to build a Logistic Regression model that is able to predict a ‘lead score’ for each of the customers such that the customers with a higher lead score have a higher conversion chance and the customers with a lower lead score have a lower conversion chance.

We have followed the steps given below for completing the assignment:

1. Exploratory Data Analysis: After importing the necessary libraries, we first load the dataset into a data-frame and view the dataset. By looking at the dataset dictionary, we understood what the features represent and what values they contain. We then started checking the null value percentage of the dataset. We noticed that there were columns with a missing value percentage of over 45%. So we dropped these columns.

2. Categorical Feature Analysis: We then went through the rest of the columns and found out that some of the features were very imbalanced (‘Country’, ‘What matters most to you in choosing a course’). So, we dropped these columns. In some of the columns such as (“Specialization”), there were some values having similar tags (xxxx_managenent, yyyy_management). So we grouped these under a single feature for better analysis. There were also some duplicate values which were either renamed or replaced using a

lambda function. After this, we were able to see that there were only 2 columns with less than 2% of null values remaining. So we just dropped these values.

3. Numerical Feature Analysis: After the categorical feature analysis, we were left with 14 columns, out of which 4 were numeric (including the target variable). So, we first create a correlation matrix. We then check these features for outliers using a boxplot. 2 out of the 3 numeric columns had outliers. We then proceeded with the outlier treatment on these columns and removed the 1% and 99% quartile to remove the outliers. We finally plotted these features as a boxplot against the target variable and derived insights which is explained in the ppt.

4. Dummy Variable and Data Scaling: Now that the analysis is done, we then move on with the dummy variable creation for the categorical columns and data scaling for the numerical columns (excluding target variable). Once the dummy variables were created, we will drop the original column. We then split the dataset into train_set and test_set with a ratio of 7 : 3.

5. Model Building: Once the data split is done, we start building the model using ‘statsmodel’ and ‘ref’. We first get the ref ranking of our new dataset and select the first 15 ranked columns. We then create a new dataframe containing the selected features and fit the dataframe in our model and get back the p-value. We then drop the columns which are having a very high p-value and move on to check the vif scores of the selected columns and drop them if there are unusual scores assigned to these columns. Once these checks are done to ensure that there is low multicollinearity between the features, we can check the accuracy of the model. If we are satisfied with the accuracy, we can move with the test set predictions and evaluate the model.

We were able to get an accuracy of over 90% with our test set predictions, indicating that our model performs well.