Lead Scoring Case Study Summary

Problem Statement

X Education has appointed you to help them select the most promising leads, i.e. the leads that are most likely to convert into paying customers. The company requires you to build a model wherein you need to assign a lead score to each of the leads 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. The CEO, in particular, has given a ballpark of the target lead conversion rate to be around 80%.

Process Followed

1. Understanding the domain/variables. – Going through the data dictionary and understanding all the variables

2. Import/Load the data. – Created a data frame using pandas to load the leads.csv file

3. Check the structure/metadata of the data. – Inspect the data frame using functionalities such as head, describe, Info, shape etc.

4. Perform Exploratory Data Analysis – Sweet Viz was used to perform EDA to understand zero variance variables and quantum of missing values

5. Data cleaning and Preparation – Zero variance columns were removed and rows with null values were removed. Columns not important for the model based on data understanding was also removed. Dummy variables were created for all categorical variables.

6. Train - Test split of data – The data was split into 70% for Train data and 30% for test data

7. Scaling of variables – Min Max scaler was used for scaling of all continuous variables was done

8. Building a model using Logistic Regression and calculate Lead Score – Logistic Regression was conducted using statsmodel and feature selection was done using Recursive Feature Selection (RFE) to bring the features from 75 to 15. Finally manual iterative approach was done to reduce the p value below 0.05 and VIF below 5 to arrive at a final list of 11 features

9. Model Evaluation on train data using metrics like Accuracy, Sensitivity, Specificity or

Precision and Recall – Model Evaluation was done using ROC first to arrive at optimal cutoff of 0.42 for accuracy, sensitivity and specificity of 79,79.3 and 79 respectively and 0.42 for Precision and Recall method with values 78 and 77.8 respectively

10. Using the final model on the test data – The model was finally run on the test data set

11. Model Evaluation on Test data using metrics like Accuracy, Sensitivity, Specificity – Values obtained were 78.5, 78 and 79 respectively