Summary Report for Lead_Score

Problem Statement: We need to find out the customers who have a high conversion percentage to become a part of the X education.

We started the process through Data cleaning, The Data cleaning process for Lead_score is huge. Firstly, we need to check for the percentage of the null value for the entire column. We found some columns even with the high percentage values, they are essential in the process for checking which customers have high chances of conversion.

The first challenge we encountered in the process of Data Cleaning is the null percentage for Lead_quality, we got ~51% null value for Lead_quality. We then started to analyze the column through the EDA(Exploratory Data Analysis) approach and filled the null_values through the percentage of the sub-categories in the Lead_quality. For the rest of the columns with the least percentage, we followed the same approach.

Then comes the main thing i.e., which columns should be dropped and which columns are more beneficial for our analysis. For every analysis we should be very precise and careful while removing the unwanted columns, because there are more chances, one mistake can completely give us wrong analysis at the end. Here, we removed the columns by checking how they are useful in calculating the conversion rate and also which columns will help us in getting the probability for the chance of the customer to join the course. We have two columns in the data set: “What matters more to you in choosing the course”, “Newspaper”. In both, the columns we have only one sub-category which has a high percentage of reputation, from these kinds of columns the probability and also conversion rate cannot be defined. We have removed these kinds of columns from the data set which doesn’t give much insight into the analysis further.

Once we have cleaned and removed the unwanted columns from the data-set. Here comes the data preparation part, we have created the dummy variables for the categorical variables which are present in the data-set after cleaning. This process is the first step towards data-modeling.Once the dummy variables are created, we scale the data using the Standardization method. We generally do scaling to make sure that every point in the data is in one place. This is one of the biggest challenges while preparing the data if the data is not scaled the modeling cannot give us the accurate outputs. After concatenating the dummy variables, we have dropped the actual columns to reduce the redundancy in the data set. This step should be done so that we don’t find any repetitions in the data while modeling.

The data is prepared, the final step towards the analysis is training-testing the data. We have divided the data into 70:30 percent for train and test respectively. Once the data gets split. We did the RFE(Recursive Feature Elimination) to check the columns which have high P-values. We remove the columns eventually. But what the challenging part we got to know is that we

should never remove all the columns at a time, as if we did that there might be a chance we will be losing the columns which are more important for final analysis. We need to remove each column which has the highest p-value and then again we need to check for the p-values for the rest of the columns. Now we need to use the final trained data with the test data to check the percentages in both the data sets.

This is the approach that we followed for the analysis of lead_score data.