Summary – Lead Scoring Case Study

This analysis is done for X Education which sells online courses to industry professionals. The dataset provided to us helped us to know how potential customers visit the site, how much time they spend and conversion rate etc.

In order to analyze the data, we followed the below mentioned process:

1. Data understanding and inspection: We went through the Lead dataset and dictionary to derive an understanding of the data, no. of rows and columns, missing values etc.

2. Data cleaning: While dealing with null values, we dropped columns having greater than 40% null values and tried imputation for the rest based on the most frequent value in general. After this, the null value percentage in columns reached to as low as below 2% and we resorted to drop these rows to make a fair analysis.

3. Exploratory data analysis: After checking for any data duplicates, we noticed a 38% conversion rate. We performed univariate and bivariate analysis. In the process we came to know that many categorical values were irrelevant. We checked the numeric variables too for outliers etc. This way we were able to draw some inferences to set further progress note in the analysis.

4. Dummy variables: We converted some binary variables ("Yes/no") to 1,0, created dummy variables for categorical variables and dropped the columns for which dummy was created. For numeric variables we used Standardscaler.

5. Train-test data split: The split was done in the ratio of 70%-30% for train and test data respectively.

6. Model building: Firstly, we did feature selection using RFE with 20 variables as output. Then we started the rigorous assessment of models. We used Stats models and checked based on the p-value and VIF values. We dropped the columns with very high p-values and VIF values. Continuing in this fashion, we ended up selecting the 9^{th} model where we had 12 variables. The p-values of all variables were below 0 and the VIF values were also pretty low.

7. Model evaluation: Next, we made prediction on the train set and to find the predicted labels took conversion probability cut-off as 0.5 arbitrarily. Next, we made the confusion matrix and checked for accuracy, sensitivity and specificity of our model. In order to improve our sensitivity % we used ROC curve to find optimal cut-off point which came out to be 0.34. After model evaluation, assigning lead score to train and test data and making predictions on the test data we found the accuracy, sensitivity and specificity percentage for both train and test data to be around 80% each.

Thus, we achieved our goal of getting a ballpark of the target lead conversion rate to be around 80%. The Model seemed to predict the Conversion Rate very well which should be able to give the CEO confidence in making good calls based on this model to get a higher lead conversion rate of 80%.

8. Inference: Finally, we found important features like time spent on website, lead source, occupation, which should be contacted or not based on conversion probability.