Lead Scoring Case Study

Problem Statement: Lead Scoring case study aims at identifying leads that have more probability to be converted into sales. The X-Education organization is facing issue with low lead conversion rates and the management wants to identify and put efforts behind the leads that have more potential to be converted. To solve this problem, we have been given a dataset of existing leads with multiple parameters and we need to build a ML model to identify hot leads.
Since, we would be classifying a lead as a potential hot or cold, this becomes a Logistic Regression example.

Approach

EDA

1. Data Cleaning: As per the best practices, the first step in order to build the model is to understand and clean the data. As part of Data Cleaning exercise, first of all Null values were removed, by wither deleting the columns with more than 40% null values (including "Select" entries) were dropped, remaining columns were imputed with modes and medians depending upon the nature of column and by deleting the rows where the percentage of null values was very low.
   Further data cleaning was done by doing Spell checks and also by dealing with outliers.

2. Univariate Analysis: After data cleaning, univariate analysis was performed to understand several trends with respect to all the variables. It was identified that few values did not have any variance and they were all or mostly similar and hence those columns were deleted.

3. Bivariate Analysis: A Bivariate analysis was performed to understand how the target variable "Converted" varies with other variables and what are the trends.

Model Building:

1. Data Clearning: As first step for model building, data preparation was done by converting all binary variables to 0 and 1 and also creating dummy variables for other categorical variables. Scaling was done for numerical variables.

2. Model Building: Model building was done with several iterations by checking p-values and VIF values and dropping variables that had higher p-values and VIF values. This resulted in a model with around 80% accuracy. However, the model had low sensitivity (64%), which means that True positive ratio was low as compared to overall Positive values detected. This would result in unnecessary effort for reaching out to customers who actually had less chances of converting. This was happening as a random cut-off of 0.5 was used to consider a Yes and No.

3. ROC Curve: In order to determine optimum cut-off, ROC curve method was used. This resulted in identifying a cut-off to be around 0.35.

4. Model was re-tested with a cut-off of 0.35 and this time all Accuracy, Specificity and Sensitivity were around 80%

5. Conversion Probabilities were multiplied by 100, to get a Conversion score.

Results

1. We had a model with around 80% accuracy.
2. Model was able to predict Conversion to an accuracy of around 80%
3. The test data also resulted in similar numbers, proving that model was not overfit and was general enough
4. We had 15 features in the model.