Lead Scoring Case Study

SUMMARY

The Problem

X Education wants to build a model where they assign a lead score to each lead 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%.

The Solution

1. **Data Cleaning and EDA**
    - Attributes with very high missing values are dropped and those which are having less than 40% data missing are imputed by the median for categorical columns and mean for numerical columns.
    - Then those attributes are also dropped which don't give any insights. For dealing with outliers we take the help of IQR to cap some high outlier points.
    - Doing bivariate analysis to see the relationship between different attributes with the target column. After proper data cleaning, we left with 15 attributes including the target column for data modeling.
2. **Data Preprocessing**
    - We make the dummy variable for categorical columns and for numerical columns we standardize them using StandardScaler. After making dummy variables we left with 84 attributes.
    - For dealing with multicollinearity issues we drop some attributes which have more than 80% collinearity. After dropping high collinearity dummy variables we left with 79 attributes. Then we start our model building.
3. **Model Building**
    - We split the data into a 70-30 ratio for train and test data respectively.
    - After that, we take the help of RFE to select the top 15 features for making the first model, then eliminate those features which have insignificant p-value and high VIF.
    - Also, we add some important attributes which we find insightful during the EDA and run the model again and again. At last, we found 12 important features having low VIF and significant p-value.
4. **Model Evaluation**
    - After making the model we take different evaluation parameters like accuracy, sensitivity, specificity, precision, false positive rate, and negative predicted value to check the model performance. Also, we take the help of the ROC curve to find the best cutoff value for classifying the leads.
    - Our model accuracy, sensitivity, and specificity are above 90% and precision is around 86% so overall our model is doing great.
5. **Prediction Of The Test Set**
    - Finally, we take the cutoff as 0.27 and check different evaluation parameters in our testing data set
    - The difference between train and test evaluation parameters is not so elevated so we conclude that there are no overfitting issues and our model predicts so well.