Lead Scoring Case Study

Lead scoring is the process of identifying and prioritizing potential customers based on their likelihood to convert into paying customers. In this lead scoring case study, we built a logistic regression model to predict the conversion probability of leads.

The data cleaning phase involved handling unknown values 'Select' present in many categorical columns, dropping columns with more than 40% missing values, dropping categorical columns with more than 80% class imbalance, dropping rows with less than 2% missing values, performing outlier analysis on numerical columns, and limiting the data to the 99th percentile. Univariate analysis was performed to check the distribution of data in numerical and categorical columns.

In the data preparation phase, we created dummy variables for categorical columns, performed an 80:20 train: test split, and performed Minmax feature scaling on numerical columns.

For model building and evaluation, we started with the Generalized Linear Model (GLM) from the stats model and used Recursive Feature Elimination (RFE) to perform feature selection. We ended up with 15 estimators that had a p-value of less than 0.05 and a VIF around 5. We started building a logistic regression model with 60+ features (after categorical variable conversion) and ended up with 15 features with a recall rate around 80%.

To identify the threshold for categorizing leads as converted and not converted, we used the ROC curve and brute force calculation of accuracy, sensitivity, and specificity. We identified the threshold to be 0.35, which would classify leads with a probability greater than 0.35 as converted and those with a probability less than 0.35 as not converted.

In conclusion, the logistic regression model we built for lead scoring had a recall rate of around 80% and could accurately predict the conversion probability of leads. The identified threshold of 0.35 could be used to categorize leads as converted or not converted, enabling the sales team to focus their efforts on leads with a higher probability of conversion.