Lead Scoring Case Study

Summary-

1. Data Reading and Understanding:

Here we tried to get the look and feel of the data, we observed following things:

● Number of rows and columns

● Data types of each columns

● Checking first few rows how data looks

● Checking how the data is spread

● Checking for duplicates, if any

2. Data Cleaning

Here we checked for discrepancies in the dataset.

● Checking for null values and imputing them with appropriate methods

● We used mode imputation for categorical columns.

● We used median imputation for numerical columns, if there is skewness in

the columns

● We have dropped the columns which are highly imbalanced/skewed.

● We have grouped the values of columns with low frequency into “Others”

3. Data Visualization and Outlier Treatment

● We performed bivariate analysis on categorical columns to see how they vary w.r.t Converted column.

● We performed bivariate analysis on numerical columns with Converted column to see how the leads are related to these columns.

● We have used IQR method to treat the outliers in the data set.

● In this step we also plotted the correlation matrix/heatmap to identify the columns which are correlated.

● We have detected the outliers and treated them by removing the value above 95%ile and below 1%ile.

4. Feature Scaling

At this stage our data was very clean and no outliers. We know that logistic regression takes the input parameters as numerical values. Hence, we converted all the categorical columns to numerical.

● Columns which have only two levels “Yes” and “No” were converted to numerical using binary mapping.

● Columns which have more than two levels were converted to dummies using the pd.get_dummies function. Now, the data contained only numerical columns and dummy variables.

● Before proceeding for model building, we have rescaled all numerical columns by using the standard Scaler method.

5. Model Building

We have used Recursive Feature Elimination Technique to remove attributes and built a model on those attributes that remain. RFE uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute. In this step we made the model stable by using the stats library, where we checked the p-values to be less than 0.05 and VIF values to be under 5. Variance inflation factor(VIF) is used to treat multicollinearity.

Once the stable model was created, we predicted probabilities on the train set and created a new column predicted with 1 if probability is greater than .5 else 0.

We calculated the confusion matrix on this predicted column to the actual converted column. We also calculated the metrics sensitivity, specificity, precision, recall and accuracy. We also plotted ROC curve to find the area under the curve.

6. Model evaluation on Train Set

● In step 5 we took 0.5 as the cut-of. To confirm that it was the best cut, we calculated the probabilities with different cut-offs.

● With probabilities from 0.0 to 0.9, we calculated the 3 metrics -accuracy, sensitivity and specificity

● To make predictions on the train dataset, optimum cutoff of 0.35 was found from the intersection of sensitivity, specificity and accuracy.

Predictions on Test Dataset:

After finalizing the optimum cutoff and calculating the metrics on train set, we predicted the data on the test data set. Below are the observations:

Train Data:

● Accuracy: 80%

● Sensitivity: 81%

● Specificity : 81%

Test Data:

● Accuracy: 81%%

● Sensitivity: 70%

● Specificity: 87%

8. Final Observation:

The Model seems to predict the Conversion Rate very well. We should be able to help the education company select the most promising Leads or the Hot Leads