Lead Scoring Case Study Summary

Problem Description:

X Education sells online courses. Leads generated from various sources are captured. There is other metadata around lead that is captured for each lead. Team is assigned to nurture hot leads and convert such potential leads to confirmed opportunity (leads).

Approach:

We performed the Logistic Regression on the given dataset to solve the given problem of X Education Company.

Below are the steps followed to solve this problem:

Data Understanding and Analysis:

- ❖ There are columns with higher missing value in the data. Also, there are columns where the default value “Select” is populated. We will be initially considering these as missing values and apply the same missing value treatment for such values.
- ❖ Categorical columns where % missing value is less (<5) will be imputed with mode and where it is greater than 45% will be dropped.
- ❖ Quantitative variables where % missing value is less will be imputed with the median. Statistical analysis indicated that there isn't a significant difference between median and mean for these columns and hence imputing with median should not create issues.

Data Visualization and Preparation:

- ❖ We performed univariate analysis on categorical columns to see which columns make more sense and removed those columns whose variance is nearly zero. We performed bivariate analysis on categorical columns to see how they vary w.r.t Converted columns.
- ❖ Boxplot indicates that there are outliers in the dataset. We have used the IQR method to treat the outliers in the data set.
- ❖ Logistic regression uses numerical data, so we will convert the categorical data by using the following technique.

i. Dummy Variables – Categorical variables with low/moderate level will be treated using dummy variables.

ii. Label Encoding – We will use label encoding for variables with higher levels. This is to avoid drastic increase in dataframe size.

Model Building:

- ❖ We have used the Recursive Feature Elimination Technique to remove attributes and built a model on those attributes that remain.
- ❖ In this step we made the model stable by using stats library, where we checked the p-values to be less than 0.05 and vif values to be around 5.
- ❖ Variance inflation factor(vif) is used to treat the 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 0.5 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 the ROC curve to find the area under the curve.

Model Evaluation & Prediction on Train Set:

- ❖ In this step 5 we took 0.5 as the cut-off. 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.
- ❖ We can observe that 0.35 is the tradeoff between Precision and Recall. Thus we can safely choose to be considered a hot Lead.

Prediction on Test Set:

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

Train Data:

Accuracy: 80 %, Sensitivity: 80 %, Specificity: 81%, Precision :72 % Recall :80 %

Test Data:

Accuracy : 80 % Sensitivity : 80 % Specificity : 80% Precision : 72 % Recall : 80 %

Conclusion:

With the current cut off as 0.35 we have Precision around 72 % and Recall around 80%

From the above values We can conclude that the Model seems to predict the Conversion Rate very well and we should be able to assure the CEO in making good calls based on this model.

Number of Hot_leads are: 397