LEAD SCORING CASE STUDY SUMMARY

Problem Statement

An education company named X Education sells online courses to industry professionals. A Company needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

To built a model wherein a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with 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%.

Solution Summary

Reading and Understanding the Data

Import and analyze the data

Data Cleaning

We drop the features with high percentage of null values; in our case we dropped the features which have more than 40% of null values present.

Imputation of missing values by creating the new variable, in case of categorical features. And removing the highly skewed features.

Exploratory Data Analysis

We started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented.

Dummy Variable Creation

We first converted the binary variable of categorical features to numerical binary feature.

Then created the dummy variable for all categorical features.

Splitting of Dataset to Train and Test

Now we divided the dataset to train and test split with proportion of 70-30% of size.

Standardising the features

The next step is to standardise the numerical features, in our case we used Standard Scalar from SKLEARN library.

Feature Elimination

Initially we removed the features with Recursive Feature Elimination method which drops down to top 18 features.

Then by using the statistics we recursively tried to looking at p-values in order to select most significant features and remove all insignificant features.

We also look for VIF value to eliminate the features.

Model Building

Finally we built the model with the final selected features and predicted the target variable probability.

Then we plotted the ROC curve and the curve came out be pretty decent with an area coverage of 89% which further solidifies the model.

Plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The intersecting point of the graphs was considered as the optimal probability cut-off value. The cut-off value was found out to be 0.358.

Evaluation of model

Based on the optimal cut-off value we predicted the target variable.

Now we need to evaluate the predicted value with actual value. Came out to be

- Accuracy → 80.9 %
- Sensitivity → 78.48 %
- Specificity → 82.35 %

Prediction and Evaluation on Test Data

Then we implemented the learning to the test set and calculated the conversion probability based on probability cut-off we found out the predicted target variable and the evaluation on test set is done.

- Accuracy → 80.13 %
- Sensitivity → 77.76 %
- Specificity → 81.65 %

Overall model performance

After combining the train and test set the overall performance of our model is,

- Accuracy → 80.67 %
- Sensitivity → 78.25 %
- Specificity → 82.14 %

HOT LEADS: The most potential leads which comes out to be 1601 out of 9074 leads.