Lead Scoring Case Study Summary

Problem Statement:

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

The company needs a model wherein you 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:

Step1: Reading and Understanding Data.

Read and analyze the information.

Step2: Data Cleaning:

We remove variables with a percentage of NULL values. This step also includes assigning missing values with medians, if necessary, in numerical variables and creating new categorical variables in categorical variables. Outliers are ide ntified and removed.

Step3: Data Analysis

We will begin exploring data analysis of the dataset to get an idea of how the data is driven. This step defines approximately 3 variables with only one value in each column. These changes are removed.

Step4: Creating Dummy Variables

We continue by creating dummy profiles for categorical variables.

Step5: Test Train Split

The next step is to divide the dataset into testing and training parts with values of 70-30%.

Step6: Feature Rescaling

We scale the number of primary variables using minmax scaling. The first model was created using the statistical model, which w ill allow us to see all the measurements of the model.

Step7: Feature selection using RFE

We select the 20 most important features using recursive feature removal. Using statistical results, we iterate to look at the P value to select the most significant results that should occur and discard non-significant results.

Finally, we have the 15 most important changes. VIFs of these variables were also found to be positive.
We will create a data frame with the variable value and we have the first hypothesis: If the value is greater than 0.5 it means 1, otherwise it will be 0.

Based on the above assumptions, it measures Uncertainty and calculates the overall accuracy of the model.

We also calculated the "sensitivity" and "specificity" matrices to understand the reliability of the model.

Step8: Plotting the ROC Curve

Then we tried to draw the ROC curve of the face, and the efficiency of the curve was very good with an area of 89%, which strengthened the performance of the model.

Step9: Finding the Optimal Cutoff Point

We then graph the "accuracy", "sensitivity" and "specificity" results for different values. The intersection of the graphs is considered the best possible cut. The cut-off point was found to be 0.37

According to the new results, we can observe that the model predicts approximately 80% correct values.

We can also evaluate new results such as "Accuracy = 81%, "Sensitivity = 79.8%" and "Specificity = 81.9%".

Score is also calculated and the final result Prediction is also calculated to deliver approximately 80% of the project estimate

Step10: Computing the Precision and Recall metrics.

We also found that the precision and recall metric values of the training data were 79% and 70.5%, respectively.
Based on real and repeated processing results, we obtained a cut-off value of approximately 0.42.

Step11: Making Predictions on Test Set

We then used the learning from the test model and calculated the probability of change based on expected and specific measurements and found a true value of 80.8%; Sensitivity = 78.5%; specific = 82.2%.