SUMMURY REPORT

To solve the given problem of X Education Company, we performed Logistic Regression on the given dataset. We followed below steps to generate score against each lead.

1. Loaded the data into jupyter notebook.

2. Performed missing value treatment for all the columns and outlier treatment on numeric columns after dropping the columns which had more than 70% missing value.

3. After that we performed EDA on each categorical column to visualize data for data imbalance.

4. We dropped high data imbalanced columns.

5. Performed train, test split using sklearn library.

6. Scaled the train data using Standard Scaler.

7. Performed RFE on scaled data for coarse tuning and selected 15 features.

8. Used statsmodels GLM method to perform Logistic Regression.

9. We checked the coefficients, p-value and VIF of the selected features.

10. Performed 2 iterations after dropping high p-value features one by one.

11. In this final model, we perform prediction based on the selected features.

12. We assumed the probability cutoff of 0.5 to be converted a lead and based on that we created predicted conversion. If the predicted probability is > 0.5, lead will be converted else not.

13. Based on this we created the confusion matrix and checked for accuracy (92%), specificity (96%) and sensitivity (86%).

14. Then based on this model, we plotted the Receiver operating characteristic (ROC) curve to check the balance between True Positive Rate (TPR) and False Positive Rate (FPR). ROC curve is plotted by changing the probability cutoff value from 0.0 to 1.0. For each probability value we plot TPR against FPR to plot the curve. Higher the area under the curve means better model. Our model covered 96% area under the curve.

15. To find the balance between sensitivity and specificity, we created confusion matrix from 0.0 to 0.9 probability of every 0.1 increase. Then we plotted accuracy, sensitivity and specificity of each tenth probability. The intersection point of these 3 curves is the optimal point where sensitivity and specificity will be balanced. We took this cutoff as 0.2.

16. Then we reevaluated the model based on 0.2 probability cutoff and found accuracy (92%), specificity (94%) and sensitivity (87%).

17. After that we performed scaling by transform the test data.

18. Performed prediction of lead conversion on test data.

19. Evaluated the prediction on test data by accuracy, specificity and sensitivity matrix. And we found accuracy -92%, specificity - 94% and sensitivity - 87%.

20. We determined that our model is working equally well on unseen data.

21. Then we created lead conversion score = (conversion probability * 100) to give a score between 0 to 100 where higher the value means the lead is “hot” and there is high possibility that the lead can be converted.

There are many learning we gathered from this assignment. These are as below:

1. How to handle missing value and outliers in a data set.

2. How to create dummy variables/labels on categorical columns.

3. How to use python libraries to perform logistic regression on selected features.

4. How to choose best model based on balanced sensitivity and specificity.