Lead Scoring for X Education

1. Loaded the given data. Identified the numerical values & categorical levels, number of missing values, data types for all the variables in the dataset. Obtained its shape.

2. Dropped those variables not having any quantitative or qualitative significance with respect to our problem statement.

3. Looked at the data imbalance. Detected if any outliers and plotted probability distribution for numerical variables, before & after imputing missing values with median. Dropped the variables having >30% missing values, including those numerical variables whose missing values were imputed. Retained those numerical variables which had <30% missing values before imputation.

4. Dropped those variables where >30% of the users have not selected any option, as the value 'Select' is treated as a missing value. Retained only the variable "Specialization" since <30% 'Select' values.

5. Dropped those variables with large class imbalance.

6. Selectively dropped the rows having missing values for each column.

7. Imputation of 'Select' values in the variable "Specialization" was not possible, as it is the most occurring value. Obtained using mode() of the variable "Specialization".

8. Performed univariate & bivariate analysis for both numerical & categorical variables. Recorded the observations.

9. Created dummy variables for all categorical levels, dropping the first level (for model training). Dropped the level 'Specialization_Select', since treated as a missing value. Dropped original categorical variables after creating its corresponding dummies. We now have no more missing values.

10. Performed a 70% - 30% split of our dataset into train data & test data.

11. Divided both train data & test data into input & output/target variable(s).

12. Detected if any outliers and plotted probability distribution for numerical variables to decide which scaling operation to undergo. Based on the observations (stated in the Jupyter/Python notebook / .ipynb file), decided to go with *standard()* scaler method to acquiesce the outliers and the wide range.

13. Scaled input & output numerical variables for both train data & test data, by standardization.

14. Plotted a heatmap to find the correlations between all categorical levels.

15. Used a combined approach of both automated (i) & manual (ii) methods to select the best features.
    i. Recursive Feature Elimination (RFE) – selected only true RFE Support features
    ii. Generalized Logistic Regression – selected features with p-value < 0.05 & VIF < 5

    * VIF = Variance Inflation Factor

16. Predicted the probability of lead conversion on train data, using arbitrary probability cut-off = 0.5.

17. Evaluated the predictions made on the train data using Sensitivity, Specificity, Accuracy, Precision, Recall, True Positive rate, False Positive rate, True Negative rate, False Negative rate, Positive Predictive value, Negative Predictive value performance metrics.

18. Plotted Receiver Operating Characteristic (ROC) curve for train data to obtain Area Under the Curve (AUC). Performance metrics and AUC good for train data, with arbitrary probability cut-off.

19. Plotted {Sensitivity, Specificity, Accuracy} vs. Probability cut-off to find the optimal cut-off point. Also, plotted {Precision, Recall} vs. Probability cut-off and, computed F1 score to address the correctness of prediction with respect to our problem statement, considering the precision - recall trade-off.

20. Predicted the probability of lead conversion on train data, using optimal probability cut-off = 0.44. Re-evaluated these predictions using ROC curve & performance metrics.

21. Selecting the same set of features in test data, as that done in train data.

22. Predicted the probability of lead conversion on test data, using optimal probability cut-off = 0.44, same as in train data. Evaluated these predictions using ROC curve & performance metrics.

23. Evaluation results yielded good & similar results in both train & test data.

24. Generated model equation by computing the predictive power of each of the selected features. Visualized the predictive power of the features using a bar plot.

25. Assigned a lead score for each ID, after combining the prediction values of train & test data.