LEAD SCORING CASE STUDY

SUMMARY REPORT

Problem Statement –

X Education sells online courses to industry professionals. The company wants to identify potential leads to improve their conversion rate for which they need a Machine Learning model that can assign scores to each of the Leads based on their conversion probability. Which means Leads with higher score demonstrate high chance of converting into a customer. And Leads with low score demonstrate low chance of converting into a customer.

Summary Report –

The approach taken to build the Model is as follows:

Reading and Inspecting the Data:

a. Import libraries and analyze the data.
b. Check for duplicates in the data set.
c. Check the shape of the data set.
d. Check the datatypes of each column in the data set.
e. Check the statistical summary of numerical columns in the data set.
f. Check the count or percentage of missing values in each column of the data set.

1. Data Cleaning:

a. Replace select values in the data set with null values (i.e., np.nan)
b. Re-calculate the null value percentage of all the columns and drop the columns with missing values more than 40%.
c. Since we are trying to predict the model based on the information provided by the Lead, drop all the columns with information generated by sales team.
d. Impute columns with null values less than 40%.
    - Use median for numerical column
    - Use mode for categorical column.
    - If missing values in a categorical column are not suitable to impute with mode value, create a new category for the missing values.
e. Drop columns which are highly skewed or having only one value.

2. Data Analysis:

a. Perform analysis on categorical columns with respect to the target variable.
b. Perform analysis on numerical columns to detect outliers.
c. Remove outliers from the numerical columns.

3. Dummy Variable Creation:

a. Columns having “Yes” and “No” should be convert to binary.
b. Columns having more than two levels were converted to dummy variables with one variable dropped from the column as a redundant variable.

4. Train-Test Split:

a. Split the dataset into train data and test data in 70:30 ratio.

5. Feature Scaling:

a. Scale the numerical features using Standard scaler.
b. Check the percentage of converted customers.

6. Feature Selection:

a. Use RFE for feature selection. Start with 20 features.
b. Check accuracy at cut-off=0.5.
c. Remove features one by one based on p-value and VIF.
    - P-value should be less than 0.05
    - VIF should be less than 5.
d. Check Accuracy, Sensitivity and Specificity for this model.

7. ROC Curve:

a. Plot ROC curve to find the area under the curve.

8. Find Optimal Cut-off:

a. Calculate Accuracy, Sensitivity, and Specificity at different probabilities.
b. Plot a graph for Accuracy, Sensitivity, and Specificity to identify the optimal cut-off point. That is the intersection point of three values.
c. Re-evaluate the metrics using the optimal cut off point.

9. Make Prediction on Test Data:

a. Draw predictions on the test dataset using the model and optimal cut-off point.
b. Calculate the Accuracy, Sensitivity, and Specificity for test data.

10. Lead Score Calculation:

a. Calculate the lead score for the whole data set.
b. Create a data-frame with cut-off and conversion
c. Check percentage conversion.