Lead Scoring Assignment: Summary Report

Analysis is done for X Education company to assign the customers with a lead score such that the customers with higher lead score have a higher conversion possibility and the customer with lowest score have a lower possibility.

In order to achieve it, we have built ML model and assigned a lead score between 0 and 100 to each lead. Based on the logistics regression model and the requirement of the company, a cut-off lead score can be selected to classify if the lead is hot or not.

Following are the steps followed:

1. Data Cleaning:
   - Data was treated for null values. Either the columns with high null percentage was removed or were filled with mode values for categorical columns and median for continuous variables
2. Exploratory Data Analysis:
   - Univariate and Bivariate analysis was done against the target variable "Converted" column.
   - Outliers were replaced with IQR values of 5% and 95%
   - Columns which were not adding any information to the model were removed
3. Data Preparation:
   - The columns with yes or no variable are replaced with numeric values 1 or 0
   - Dummy variables are created for the categorical columns
4. Splitting the train and test datasets:
   - The data is divided into train and test data set with 70% and 30% of the data respectively
5. Scaling:
   - Standard Scaler is used to the numeric values to be in the range of -1 to 1
6. Model Building:
   - Feature selection using recursive feature elimination [RFE] for 15 columns from the dataset
   - Based on p values and variance inflation factor [VIF] the columns are removed manually such that all the columns p values are less than 0.05 and VIF is less than 5
7. Model Evaluation:
   - Confusion matrix is created using the predicted values and the target variable Converted column
   - Based on the confusion matrix Recall score, Precision score, Accuracy , Specificity and Sensitivity is calculated
   - Built the ROC curve using accuracy , specificity and Sensitivity to find the optimal cut off value which happens to be between 30-40% .
   - Optimal cut off value is derived to be 35%
   - With the above cut off the conversion rate has increased to 42%
8. Predictions:
   - Predictions were made on the test data set using the above columns arrived from the evaluation step.
   - Predictions for lead score was arrived with cut off score of 35% with Accuracy of 80%, Sensitivity of 79% and Specificity of 82%

9. Precision and Recall Curve:
    - These scores are calculated to reverify the test data predictions
    - Obtained Precision Score of 71% and recall score of 79%
10. Key Insights from the model:
    - There is around 1% difference on train and test data's performance metrics. Hence, we can say that there is not overfitting of the training data.
    - Sensitivity values are high which will predict the customers correctly who might get converted.
    - Specificity values are high which will ensure that customers on the verge of conversion are not selected
    - Depending on the business requirement, we can increase or decrease the probability threshold value which in turn will decrease or increase the Sensitivity and increase or decrease the Specificity of the model.
    - For the CEO's ball park number of 80% conversion rate can be achieved by changing the cut off to 10%.