LEADS SCORING CASE STUDY

Below steps were followed to proceed with assignment for creating a model for Leads Scoring Case Study.

Step 1: Importing libraries and Data

1. Imported all the necessary libraries.
2. Imported Leads Dataset using Pandas library.

Step 2: Data Understanding

1. Checked number of rows and columns in Dataset using ".shape".
2. Checked Data type of each column using ".info()".
3. Computed statistical aspects of dataset using ".describe()".

Step 3: Data Cleaning and Transformation

1. We have Dropped the columns which are no longer useful in dataset.
2. After removing the columns which are not required and seeing the top 5 rows of dataset we see that there are some columns having label as 'Select' which means customer has chosen not to answer this question. We will be replacing those lables 'Select' to null values.
3. Calculated missing values in dataset and removed columns having more than 30% of missing values.
4. Then in next step for missing values we have imputed missing values with values maximum number of a occurrences in a particular column.
5. One of the columns was having two identical names in different format. We changed the labels name into one format.
6. Changed binary variables into '0' and '1'.
7. Created dummy variables for multicategory variables.

Step 4: Checking for Outliers

1. Checked outliers and plotted box plot for variables having outliers.
2. Created bins for the variables having outliers.
3. Removed redundant and repeated variables in the dataset.

Step 5: Data Preparation

1. Split Dataset into train and test dataset.
2. Scaled Dataset using StandardScaler and checked for conversion rate.
3. After this plotted a heat map to check correlations among variables.
4. Found some variables which are highly correlated and dropped those variables.

Step 6: Model Building

1. Created a Logistic regression model.
2. Used RFE feature selection with 15 variables and again created a logistic regression model.
3. Assessing model with statsmodel and also calculated VIF.
4. We will be dropping columns which have high p values and recreating the model until we have a final model.
5. Plotted ROC Curve.
6. For our final model we checked the optimal probability cutoff by finding points and checking the accuracy, sensitivity and specificity.
7. We found one convergent point and we chose that point for cutoff and predicted our final outcomes.
8. We checked Precision and Recall for out final model and also plotted graph for Precision vs Recall tradeoffs.
9. Made predictions on test data set and predicted value was recorded.
10. We also calculated accuracy, sensitivity and specificity for test data set.
11. We found that score from out test data set is in acceptable range.
12. We also gave Lead Score to test data set to show that high lead score are hot leads where as low lead scores are not hot leads.

Step 7: Conclusion:

1. Both train and test set have accuracy, recall and sensitivity in acceptable range.
2. Top feature for good conversion rate are as below:
    - Last Notable Activity_Had a Phone Conversation.
    - Lead Origin_Lead Add Form.
    - What is your current occupation_Working Professional.