Lead Scoring Case Study
Summary Report

Step 1: Reading and Understanding the Data

- Loading the data - Leads-data
- Make copy of leads data for further process if require
- Basic data inspection

Step 2: Data Cleaning and Preparation

- Data information
- Checking NULL value with percentage
- Drop High percentage value columns
- Duplicate check
- Unique value checking each column
- Drop all the columns which don't make any inference
- NULL value treatment for each column
- Impute with mode/median as per the require
- Clubbing some parameters which are not making sense individually

Step 3: EDA (Exploratory Data Analysis)

- Univariate analysis for each column
- Inference from each column for this case
- Note down all the columns which are not making any inference
- We also treating outliers here
- Capping the value where outliers are present
- At least dropping all the columns which are not making any use for deciding the leads

Step 4: Data Preparation

- Converting some binary variables (Yes/No) to 0/1
- For categorical variables with multiple levels, creating dummy Creation
- Dropping the repeated variables

Step 5: Model Building

- Train and Test Split
- Features Scaling of the data
- Checking the Lead Conversion Rate
- Model Building - 1 :- Running Your First Training Model
- Feature Selection Using RFE
- Model Building - 2 with RFE
- Model Building - 3 with RFE
- Model Building - 4 with RFE
- Checking the p-value for each model and once it will be less than 0.05 will stop model building
- Create confusion matrix
- Calculate accuracy
- Checking VIFs and removing all high VIFs
- We will stop RFE once VIF's less than 3
- Heatmap of the all feature variable to check the Correlation

Step 6: Model Evaluation

- Calculating Metrics beyond Accuracy
- Sensitivity & Specificity
- Plotting the ROC Curve
- Finding Optimal Cutoff Point
- Precision and Recall
- Precision and recall tradeoff
- Calculating the F1 score

Step 7: Making predictions on the test set

- Classification Report
- Plotting the ROC Curve for Test data
- Calculating the Area Under the Curve(GINI)

Step 8: Calculating Lead score for the entire dataset

- Lead Score = 100 * Converted Probability
- This needs to be calculated for all the leads from the original dataset (train + test)

Step 9: Determining Feature Importance

- Making the list of all feature
- Sorting these feature
- Find out TOP 3 features variable

Step 10: Conclusion

- Positive Coefficients features
- Negative Coefficients features
- TOP 3 features
- Conclusion and Recommendation

Team-

- Rohit Keshari
- Rahul Choudhary

---------End of Summary---------