SUMMARY REPORT

Lead Score Case study:

- Read the Problem statement and get an overall idea about the firm to predict best lead conversion variables.

Reading and Understanding the Data

- Import all required Libraries.
- Import the data set and dictionary file for better understanding of features.
- Read the dataset info, shape, describe and duplicates for to get an idea towards data.

Data cleaning

- Then we started EDA and found maximum variables with 30% more null values In it so removed those features from dataset
- Started filling the missing data with mean, median and mode with respect type of data and column preference
- Unique data contained variables and skewed data contained variables has been removes since it will make our model biased towards some variables.
- The categorical dummy variables having below 100 occurrences also removed from the data set.
- Outliers also treated and kept the values below 99% mark.

Creating dummy variables

- Here we created dummy variables for categorical variables.

Test Train Split

- Dividing the data set into Train Dataset (70%) and Test Dataset (30%)

Feature Scaling

- We have used Standard Scaler to scale the original numerical variables. Then using states model we created our initial model, which would give us complete statistical view of all the parameters.

Feature Selection RFE and Model Building

- Using the recursive feature elimination we went ahead and selected top 15 important features from dataset.

- Using the statistics generated, we recursively tried looking at the p values in order to select the most significant values that should be present and dropped the insignificant variables.
- Finally, we arrived at the 11 most significant variables. The VIF's for these variables are below 5 and P also found to be good.
- We then created a data frame having converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0.
- Based on above assumption, we derived the confusion matrix and calculated overall accuracy for the model
- We also calculated sensitivity and specificity metrics to understand how reliable the model is

Plotting the ROC curve

- Here we plotted ROC curve to get the overall efficiency of our model covered area, the curve came out to be decent with an area coverage of 88% its good.

Finding the Optimal cutoff point

- After plotting the accuracy , sensitivity and specificity we got an intersection point at 0.365 that's the cutoff point
- Based on the value we observed 79% of values are rightly predicted by the model.
- We can also see for train data Accuracy :80%, Sensitivity :79% and Specificity : 81% And also calculated Precision: 72% , and Recall : 79%
- Here we got the final predicted conversion rate at 79.5% which is good model.

Making Perditions on test Data set

- Same modeling methods implemented on test dataset too with respect to that we got accuracy , sensitivity and specificity are respectively at 80% , 84% and 82%
- And also found that precision : 74% and recall : 84%
- Finally we got the percentage of final predicted conversions on test data as 84%