A brief summary report explaining how we proceeded with the assignment and the learnings that we gathered.

1 Reading and Understanding data

- Imported the dataset.
- There were some columns in the dataset which contains a value **Select**. This value is replaced with Nulls.

2 Data Preparation

- Converted all the binary variables to 0 and 1.
- Created dummy features for categorical variables with multiple levels
- Next checked the outliers of the dataset. The columns Total Visits,Page Views Per Visit have outliers. Removed the outliers using soft capping at higher end
- After fixing the outliers, we split the dataset into train and test set.
- In order to keep all the variables in the same scale we did standardisation
- Checked the correlation of the dataset. Some variables were highly correlated. Those variables were removed

3 Test Train Split

We split the dataset to train and test set. Train dataset was used to train the model and test dataset was used to evaluate the model.

4 Feature Scaling

Standardisation is done to bring all the variables on the same scale.

5 Variance Thresholding

The accuracy, specificity and sensitivity was calculated for various values of probability threshold and plotted. The optimum point for cut off was 3.5

6 Looking at Correlation

Checked the correlation of dataset using heat map. Highly correlated variables are removed

7 Model Building

Build the model with all variables. Many variables were insignificant. Dropping one by one is time consuming. So used an approach called Recursive Feature Elimination.

8. Feature Selection Using RFE

We have chosen RFE with 15 variables and continued dropping the variables one by one until we reached a model where all the variables were significant with low VIF

9. Plotting the ROC Curve

After building the final model, created an ROC curve

10 Finding Optimal cut off point

Here we choose the probability cut off which is the point where accuracy, sensitivity and specificity meet.

0.35 was the cut off point. Then we evaluated the model on train dataset using a confusion matrix

Overall accuracy of the model is 0.79

We got the precision and recall score as 0.70 and 0.80 respectively.

11 Making Predictions on test set

Here we calculate the lead score to find the hot leads. Higher the value, higher the chance of getting converted.

The accuracy of the test model was 0.80

We got the precision and recall score as 0.72 and 0.79 respectively.

The values got on the test data is approximately equal to train set. So overall the model seems to be good.

***Top three variable which increases the probability of lead getting converted are*:**

a. Lead Origin_Lead Add Form

b. Last Activity_SMS Sent

c. Last Notable Activity_Modified