LEAD SCORE CASE STUDY

Read the data first .

Understand the datatypes for each of the columns present in the data.

There are almost 9240 rows and 37 columns

Out of these 37 columns 7 are numerical and 30 are categorical

First challenge was to find and treat the Null and missing values present in the dataset

While cleaning the data it was found that we had 4 columns which had the category as "Select" we have replaced them with Unemployed , Others Specified and Others

Missing values that were greater than 45% of total values were dropped from the dataset

Also dropped certain columns like Magazine, newspaper

Replace the categories with few columns under one umbrella so that we do not create too many dummies (Lead Origin, Lead Source, Specialization, Tags)

Categorical Variables:

- Treated null values with Mode
- If there were multiple categories within a column have reduced them to 5-6
  Eg. Other_tags, Others etc
- It the categories within columns showed highly skewed toward a specific category then there were instantly dropped (Newspaper, Digital Advertisement, Through Recommendations etc)

Continuous Variables

- Treated Null values with Mean
- Prospect ID was dropped as we have Lead Number which also provided the same information

After making all these changes the total rows and columns left were (9074, 21)

Learnings were **"We need to understand the data and the columns and what details these columns provide and treat them accordingly that would help in Model Building"**

Outlier Treatment

Capped the outliers by way of quantiles

For TotalVisits and Page Views Per Visit

Converted 2 columns to Binary using def function to replace (Yes : 1 & No: 0)

When all the cleaning and treatment was done we have created dummies for the categorical variables we need to use in the Model

('Lead Origin', 'Lead Source', 'Last Activity', 'Specialization', 'What is your current occupation', 'Tags', 'City', 'Last Notable Activity')

To train the Model we have split the data into Train dataset and Test dataset in the ration 70:30 ratio.

Used Standard Scaler for scaling the continuous variables and took the heatmap to see the correlation in the data.

Used unstack to find the top 10 highly correlated variables and dropped them from Train and test data set.('City_Others' ,'Last Notable Activity_SMS Sent', 'Lead Source_Reference',

'City_Select', 'Last Activity_Email Bounced', 'Last Notable Activity_Others_Notable_activity')

By using RFE we got 15 variables on that we start building the Logistic Regression Model.

Removed variables whose p value is 0.05 and VIF is more than 3.

Note : Need to remove variable one by one and then run the Model again and also the VIF

Lead Score is calculated by multiplying converted_probability with 100

Then we calculate Accuracy, Sensitivity, Specificity, False Positive Rate, Positive predictive value and Negative Predictive Value.

Then we built the ROC curve to find out if the Model is good or not.

ROC curve 0.95 which indicate that the model is good. Then we should check if the Model is overfitting or not.

We found the optimal threshold as 0.3 to get balanced sensitivity and specificity.

We see the graph trade off of Precision and Recall.

Finally, we run our model on test data set and parent data set also (to get customer wise probability score) and see the metrics.