Lead Scoring Case Study Summary

<u>Problem Statement:</u>

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e. the leads that are most likely to convert into paying customers.

The company needs a model wherein you a lead score is assigned to each of the leads such that the customers with higher lead score have a higher conversion chance and the customers with lower lead score have a lower conversion chance.

The CEO has given a ballpark of the target lead conversion rate to be around 80%

<u>Solution Summary:</u>

Step1: **Reading and Understanding Data**.

Read the data into Dataframe and analyse the Dataframe Data.

Step2: **Data Cleaning**:

We replaced default values like 'Select' in all columns with Null values. We dropped the variables that had more than 30% percentage of NULL values in them. This step also included imputing the null values less than 30% with Mode values of the columns in case of case of categorical variables.

Step3: **Outliers**:

The outliers in continuous numeric variables were identified and removed. While removing we also Ensured not large percentage of rows are dropped.

Step4: **Data Normalising**

In the categorical Variables columns, some of the values in that column occur less than 1%. Those values can be grouped together thereby reducing the number of dummy variables during dummy variables creation

In some of the categorical Variables where one values occur more than 99%. These columns can be removed.

Step5: **Exploratory Data Analysis**

Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented. In this step, the following are observed for:

Numeric Continuous variables:

We plotted Pairplot & Boxplot to see the relationship between continuous variable columns, Following are the observations

1. From the PairPlot we can see the variable "Page views per visit" and "TotalVisits" are propotional.

2. From the PairPlot histogram we can see the 'TotalVisits”,"Page Views per visit" and "Total Time Spent on Website" are concentrated more when value is Zero.
3. From the BoxPlot we can see 'TotalVisits" and "Total Time Spent on Website" has outliers for NON- Converted leads compared to Compared to Converted leads.
4. From the BoxPlot we can see 50 percentiles i.e. median of 'TotalVisits' of Converted Leads lie between 3 & 4. While for Non converted leads median of "Total Time Spent on Website" lie approximately at 2.
5. From the BoxPlot we can see 50 percentiles i.e. median of 'Total Time Spent on Website' of Converted Leads lie between 500 & 1000. While for Non converted leads median of "Total Time Spent on Website" lie between 0 and 500.
6. From the BoxPlot we can see 50 percentiles i.e. median of 'Page Views per visit' of Converted Leads lie approximately at 2. Also, for Non converted leads median of "Page Views per visit" lie approximately at 2 (same).

Categorical Variables:

We plotted Bar graph to see the relationship between categorical variable columns with hue over target variable Converted, Following are the observations

1. Lead Origin from "Landing Page submission" has high leads with 36% conversion rate, however "Lead Add Form" tops the chart with higher conversion rate of 94% of its total leads get converted.
2. Through 'Google' company gets higher of number of leads in that 39% gets converted.
3. Leads which company gets from other means have higher conversion rate of 79%.
4. People who are ok to receive mail are the 92% of the total leads and on which 39% of leads gets converted.
5. Last activity like "1. Email Opened (38%)", "2. SMS Sent (30%)" and "3. Olark Chat (11%)" are the top 3 activities through which most of leads are generated. However, the lead conversion rate follows the order "1. SMS Sent (63%)”, "2. Email Opened (36%)" and "3. Olark Chat (8%)"
6. Indian resident has more probability to be converted to lead with 38%.
7. 90% leads are generated by Unemployed however they have less conversion rate of 33%. So, education fees might be the issue for the unemployed. Feasible loan options and placement after completion of course will encourage more unemployed leads to get converted.
8. Working Professionals are the low lead generator however most of their leads (91%) get converted. This is obvious that working professional looks for course only when they are in need and they are also capable of affording.
9. Receiving a free mastering interview copy do not have any significant observation, since who don’t receive that copy have higher conversion rate 39%.
10. Last Notable activity like "1. Modified (37%)", "2. Email Opened (31%)" and "3. SMS Sent (24%)" are the top 3 activities through which most of leads are generated. However, the lead conversion rate follows the order "1. SMS Sent (69%)”, "2. Email Opened (36%)" and "3. Modified (20%)"

Step6: **Data Preparation:**

1. Converting categorical column which has binary variables (Yes/No) to 0/1.
2. Creating dummy variables for categorical columns which are not binary in nature.
3. Removing Sequences/Id columns which are not helpful for analysis.

Step7: **Correlation between independent variables:**

The following are the observation on finding the correlation between feature variables

1. Last Activity_Email Bounced & Do Not Email --Also Do Not Email is highly co-related to converted
2. Page Views Per Visit & Total Visits are highly corelated --Also Total Visits highly co-related to converted
3. Lead Source_Others & Lead Origin_Lead Add Form are highly corelated --Also Lead Origin_Lead Add Form is highly correlated to Converted
4. Last Activity_Email Opened & Last Notable Activity_Email Opened are highly corelated --Also Last Activity_Email Opened is highly correlated to Converted
5. Last Activity_Page Visited on Website & Last Notable Activity_Page Visited on Website are highly corelated --Also Last Activity_Page Visited on Website is highly correlated to Converted
6. Last Notable Activity_SMS Sent & Last Activity_SMS Sent are highly corelated-- Also Last Notable Activity_SMS Sent is highly correlated to Converted
7. What is your current occupation_Unemployed & What is your current occupation_Working Professional are highly corelated--Also What is your current occupation_Working Professional is highly correlated to converted
8. Last Notable Activity_Email Opened & Last Activity_Email Opened are highly corelated-- Also Last Activity_Email Opened corelated to Converted

Step8: **Test Train Split**:

The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step9: Feature Rescaling

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

Step10: **Feature selection using RFE**:

Using the Recursive Feature Elimination, we went ahead and selected the 20 top important features. Using the statistics generated, we recursively tried looking at the P-values to select the most significant values that should be present and dropped the insignificant values.

Finally, we arrived at the 13 most significant variables. The VIF's for these variables were also found to be good.

We then created the data frame having the converted probability values and we had an initial assumption that a probability value of more than 0.5 means 1 else 0.

Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.

We also calculated the '**Sensitivity**' and the '**Specificity**' matrices to understand reliability of the model.

Step11: Plotting the ROC Curve

We then tried plotting the ROC curve for the features and the curve came out be decent ROC Curve with an area coverage of 88% which further solidified the of the model.

Step12: Finding the Optimal Cut-off Point

Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values. The intersecting point of the graphs was considered as the optimal probability cut-off point. The cut-off point was found out to be 0.35

Based on the new value we could observe that close to 80% lead conversion.

We could also observe the new Metrics of, 'accuracy=81%, 'sensitivity=79.6%', 'specificity=81.2%'.

Step13: **Making Predictions on Test Set**

Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 81.6%; Sensitivity=81.8%; Specificity= 81.4%.

On the Test Dataset we could observe that close to 81.8% lead conversion