Summary Report of Lead Score Case Study

The analysis is done for X Education Company which sells online courses to industry professionals and to find ways to get more industry professionals to join their courses the basic data provided gave us a lot of information about how the potential customers visit the site, the time they spend their how they reached the site and conversion rate

The following are the steps used for analysis

Step 1: Importing Data and Inspecting the Data frame

The CSV file is read and statistical information is described by using describe and info function. As mentioned the file contains values as " Select" that values are present in many of the categorical variables. It may be because the customer did not select any option from the list, hence it shows 'Select'. 'Select' values are as good as NULL. So we converted these values to null values.

Step 2: Exploratory Data Analytics

Null values are imputed by using mean and median of the respective columns. The columns having null values greater than 40 % are dropped.
After knowing the value counts of each category of the column some of the columns has only one category so dropped that columns as well as we dropped the variables which are not significant for analysis and will not give any information to the model.

Categories that have less occurrences on the Last Notable Activity may be replaced as other notable activity and this column is similar to last activity so drop this col. **The conversion rate is maximum of last activity as "Email Opened" The conversion rate of SMS sent as last activity is maximum so we have to make a call to the lead that has opened their email and to whom SMS sent to increase the conversion rate.**

To improve the overall lead conversion rate, we need to focus on increasing the conversion rate of 'API' and 'Landing Page Submission' and also increasing the number of leads from 'Lead Add Form'

After plotting the boxplot for continuous columns it shows the point (values) beyond the upper limit (third quartile) these are the outliers. Outliers are found in the following columns

"Total Visits" Max value is 251.0

"Page Views Per Visit"-- Max value is 55.0

Outliers are handled by using Capping and flooring method

After plotting boxplot of continuous variable with target variable it is observed that Median for converted and not converted leads are the same for 'Page Views Per Visit' and for total visits 'Leads spending more time on the website' are more likely to be converted

Step 3: Data Preparation

The binary variables are converted into numeric as (Yes/No) to 0/1

For categorical variables with multiple levels, create dummy features (one-hot encoded). The categorical variable with more number of labels are clubbed into on label whenever it necessary these variables are “Tags”, “Specialization”, “Lead source”, and “Last Notable Activity.

Step 4: Model building

The dataset is split into train and test with 70:30 ratio by using “sklearn model_selection import train_test_split”.

The features with high magnitude will be more significant than the other features in the final model so all the features are scaled in one magnitude by using Standard Scalar method. By using “statsmodels.api” library constant is added to X_train then by applying logistic regression n fit the model and got the summary, but this select the so many features for analysis to avoid this we use RFE feature selection starting with 15 features .We will then optimize the model further by inspecting VIF and p-value of the features with 15 features

Now final Logistic Regression Model is built with 13 features. Features used in final model are

1. 'Do Not Email',
2. 'Total Time Spent on Website',
3. 'Lead Origin_Lead Add Form',
4. 'Last Notable Activity_Olark Chat Conversation',
5. 'Last Notable Activity_Other_Notable_activity',
6. 'Last Notable Activity_SMS Sent',
7. 'What is your current occupation_Working Professional',
8. 'Tags_Could_be_Potential',
9. 'Tags_Interested in full time MBA',
10. 'Tags_Interested in other courses',
11. 'Tags_Others_or_not_eligible',
12. 'Tags_Will revert after reading the email',
13. 'Tags_in touch with EINS'

"Tags_Interested in other courses","Tags_Interested in full time MBA"and "Tags_Could_be_Potential" are the top three categorical/dummy variables in the model which should be focused the most on in order to increase the probability of lead conversion with respect to the absolute value of their coefficient factors.

Step 4: Model Evaluation

Evaluation matrix as a confusion matrix is used to calculate Accuracy, Sensitivity and Specificity as well as Precision and Recall. Different cut-offs can be used depending on the use-cases (for E.g. when high sensitivity is required, when model have optimum precision score etc.)

Model evaluation is done as follows:

Result on Train data	**Result on Test data**
Accuracy is 84.21%	Accuracy : 84.55%
Sensitivity is 83.89%	Sensitivity : 85.88%
Specificity is 84.41%	Specificity : 83.71%
Precision : 85.78%	Precision : 77%
Recall 78.50%	Recall 85.88%
	Roc : 0.91

Conclusion:

The logistic regression model predicts the probability of the target variable having a certain value, instead of predicting the value of the target variable directly. Then a cut-off of the probability is used to obtain the predicted value of the target variable.

Here, the logistic regression model is used to predict the probability of conversion of a customer (lead).

Optimum cut off is chosen to be 0.3 i.e. Any lead with greater than 0.3 probability of converting is predicted as Hot Lead (customer will convert) and any lead with 0.3 or less probability of converting is predicted as Cold Lead (customer will not convert)

The final model has Sensitivity of 85.88% this means the model is able to predict 85.88% customers out of all the converted customers (Positive conversion) correctly.

The final model has Precision of 85.78% this means 85.78% of predicted hot leads are True Hot Leads.