Summary

Analysis performed using provided lead dataset of X Education. Goals are build a logistic regression model to assign a lead score between 0 and 100 to each of the leads which can be used by the company to target potential leads and achieve conversion rate of 80 % hot to converted leads.

Steps of the Analysis:

1. Importing necessary libraries

2. Read and inspect the data:

Loaded and understood the head, size and numerical columns statistics of the data.

2.1 Data Cleaning:

There is no duplicated row in the data. Converted "Select" to Null because it is as good as a null value. Removed columns with 40% null values. Checked row null values greater than 70%.Imputed some categorical columns with appropriate data. Removed outliers in the numerical columns.

3. Exploratory Data Analysis (EDA):

Ensured the data balance after cleaning and it is coming around 38%. Performed Univariate analysis categorical and Numerical variables.

Variables with high conversion rate:

- Lead Add form – Lead Origin
- Reference and Welingak Website – Lead source
- SMS Sent – Last Activity

Variables that increase probability of lead conversion:

- API and Landing Page Submission – Lead Origin
- Olark Chat ,Organic Search, Direct Traffic, Google and Referral Sites – Lead Source
- Page Visited on Website, Olark Chat Conversation and Email Opened – Last Activity
- Finance Management, Human Resource Management and Marketing Management – Specialization

From Correlation matrix TotalVisits and Page Views Per Visit is highly correlated

4. Data Preparation

Created Dummy Variable for Categorical columns: Lead Origin, Lead Source, Last Activity, and Specialization, What is your current occupation, City, and Last Notable Activity. Split the data with 70% -train and 30% -test. Scaled the numerical variable used standard Scaler

5. Model building

RFE done to attain the top 20 relevant variables.9 time's model retrained in order to remove high VIF and p value variable.

6. Model Evaluation

Confusion matrix created and accuracy is 81% and specificity was good (~88%) but our sensitivity was only 69% this is because of 0.5 cut off

7. Prediction on Test data

With optimum cut off of 0.34 performed prediction on test data and Accuracy: 81.2 %, Sensitivity: 82.6 %, Specificity: 80.3 %

This help to achieve the goal of getting a ballpark of the target lead conversion rate to be around 80%

8. Conclusion

The company should make calls to the leads coming from below categories:

- Lead Origin - Lead Add Form
- What is your current occupation- Working Professional
- Lead Source - Welingak Website
- Last Notable Activity-SMS Sent
- Last Activity- Other_Activity
- Total Time Spent on Website
- Lead Source- Olark Chat
- Last Activity -Email Opened

The company should make calls to the leads coming from the lead Orgin "Lead Add Form" as they are more likely to get converted and "Landing Page Submission" are not likely to get converted.

The company should make calls to the leads who are the "working professionals" as they are more likely to get converted.

The company should make calls to the leads coming from the lead sources "Welingak Websites" and "Olark Chat" as they are more likely to get converted.

The company should make calls to the leads whose last activity was SMS Sent,"Email Opened" "and "Other_Activity" and as they are more likely to get converted.

The company should make calls to the leads who spent "more time on the websites" as these are more likely to get converted.

The company should not make calls to the leads whose lead origin is "Landing Page Submission" as they are not likely to get converted.

The company should not make calls to the leads whose Specialization was "Others" as they are not likely to get converted.

The company should not make calls to the leads whose last activity was "Olark Chat Conversation" as they are not likely to get converted.

The company should not make calls to the leads who chose the option of "Do not Email" as "yes" as they are not likely to get converted.