Summary

The data Provided by X Education for the analysis and to find ways to convert potential buyers into customers by different information provided by the company such as, buyers visiting the site, how much time they spend , how did they reach the site and what is their conversion rate .

AIM: To Improve the lead conversion by predicting the probability of lead conversion

Process Followed:

1) Exploratory Analysis
    a. Check for null/missing values. Imputing/removing null values.
    b. Drop columns if they consist of large missing values
    c. Handling outliers
    d. Checking for Correlation coefficients among variables
    e. Removing anomalies
2) Creating Dummy Variable & Feature scaling
3) Checking Multicollinearity
4) Feature Selection (Using RFE)
5) Model Building (Logistic Regression)
6) Model evaluation
7) Prediction

Process Detail:

- We started with exploratory analysis, find and resolve few of the anomalies.

  Few of them are:

- Column “Specialization” & “How did you hear about X Education”, “select” as a value.
- Columns “Search”, “Magazine” etc. has only one category. Simialry few of the column are highly imbalanced.

- Many Null values were present in the data.

- Univariate and bivariate analysis is done. Merged various category into few to reduce the dummy variable dimension in the final data.
- Correlation checked and it was not too much to bother in the data.
- Outliers also checked and capped. Done this in “Page Views Per Visit” & “TotalVisits” with value 20 & 50 respectively. As the observation above these range was too less.
- Created dummy variables of “Lead Origin", "Lead Source", "Do Not Email", "Last Activity", "Specialization", "What is your current occupation", "Tags", "Lead Quality", "City", "A free copy of Mastering The Interview", "Last Notable Activity”.
- Scaled feature with Standardization to bring them on same scale.
- Chose Logistic regression as Model and RFE as the feature selection technique (with top 15 important variable)
- Checked for Multicollinearity (using VIF) and found that the data do not have high multicollinearity.
- Evaluation of model was done on test dataset with F1 score with value 0.93. Also checked AUC – ROC which was 0.98

Learning and Conclusion

- The Variable that contributes the most into identifying potential buyer can be:

1. Total Time Spent on Website
2. The lead source when generated from: -
    - Google
    - Direct traffic
    - Organic Search
3. The lead origin was lead add form

- We infer through the analysis is that most of the lead originated through 'API', 'Landing page' did not get converted. However, they were the variable which were contributing to highest number of lead conversion
- We also notice that the users who went through Google searches and direct

searches converted the most are potential buyers. O-lark chat and organic also saw a significant conversion regarding potential buyer.

- The unemployed section of the buyers was more interested and likely to convert as a potential buyer. Also, working professionals can be considered as a good choice for lead conversion (potential customer)