Lead Score Assignment Summary

In order to boost the conversion rate from regular Leads to paying customers, this analysis is carried out for X Education. A substantial amount of data was provided by X Education in the form of the data set. There needs to be an increase in the company's 30% conversion rate to 80%. We must create a model in this case that will range from 0 to 100 in its Lead score. A higher score will have a greater possibility of conversion, whereas a lower score will have a much smaller chance of becoming a successful paying customer.

The following technical steps are used for the analysis:

1. Data cleaning.

A small number of columns with a significant percentage of null values were dropped.

Identified the outliers and dropped accordingly.

Following those procedures, data was retained in up to 68% of cases, and analysis was done on this cleaned data.

2. EDA.

EDA was performed on the cleansed data by plotting several plot styles and analysing the continuous and categorical variables.

For better comprehension, a univariate analysis was conducted on the target variable.

Some of the insights are follows:

- People spending more time are promising Leads.
- The Lead Origin- Landing Page Submission has the highest conversion rate among others.
- Google has the highest conversion rate.
- Leads whose Last Activity was SMS sent had the best conversion rate.
- Lead from Specialization who are unknown/Select columns has the highest rate of conversion.
- Person who are unemployed has the highest conversion rate comparatively to working professional.

3. Dummy variable creation, Train-Test split and Scaling.

Scaling and the creation of dummy variables were done for categorical columns. The scaling was carried out to place all the features within a similar range. The split was made with 30% for the test and 70% for the train.

4. Model Building and Predictions

The RFE technique was used to pick the features, and five modelling attempts were made until the VIF and p-values fell below acceptable limits.

Final Accuracy was 78%, Sensitivity was 79%, Specificity was 77% for the Test set. Precision was 77% and Recall was 78%.

Cut off was chosen to be 0.43 and prediction was made out of it.

5. Conclusion.

The variables which are important for potential Leads are Current occupation is “unemployed”, “Total time spend on the website”, “LeadOrigin_Lead Add Form”, “Last Activity as SMS sent” and “total visits”.