Lead Scoring Case Study

Problem Statement

X Education sells online courses to industry professionals. There are many leads generated through multiple Lead Generation System however, they need help in filtering the most promising leads, i.e., the leads that are most likely to convert into paying customers.

X Education wants to build a model wherein 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%

Solution Approach

- **Reading and Understanding Data**. Reading the data from the csv file and analyse the data using primary steps like describe, info, checking null values, type of data etc.

- **Data Cleaning**: We took the base percentage of 35% and dropped the variables that had above 35% of NULL values.

  Additionally, wherein there were certain important columns and had null values, we replaced the null values with 'not provided' since they were having data type as 'object'.

- Checking for columns with unique values and dropping them as they will not impact our analysis.
- **Creating Dummy Variables Once the outliers/ null values were removed and we had proper sanitized data,** we started data-modelling with the first step of creating dummy data for the categorical variables.
- **Test Train Split**: Once the dummies were created, we distributed the data set into test (70%) and train sections (30%).
- **Feature Rescaling,** we used the Min Max Scaling to scale the original numerical variables.
- we created our initial model, using stats model which would give us a complete statistical view of all the parameters of our model.
- **Feature selection using RFE**: Using the RFE we selected the top 15 important features.
- Using the statistics generated, we recursively tried looking at the P-values and dropped the insignificant variables.
- We filtered the variables and **calculated the 12 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.

- We derived the Confusion Metrics and calculated the overall Accuracy of the model and calculated the '**Sensitivity**' and the '**Specificity**' matrices to understand how reliable the model is.
- **Plotting the ROC Curve,** We then tried plotting the ROC curve for the features and the checked that the curve area coverage is 88% thus suggesting a good model
- We plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values.
- **With the current cut off as 0.35 we have accuracy, sensitivity, and specificity of around ~81%**
- **With the current cut off as 0.35 we have Precision around ~79% and Recall around ~70%**

Conclusion:

Below three variables are the top variables which contributed most towards the probability of a lead getting converted.

- TotalVisits
- Total Time Spent on Website
- Lead Origin_lead add form

We should focus more on below three dummy variables.

- Lead Origin_lead add form
- Lead Source_olark chat
- Lead Source_welingak website

Also, we can apply below strategy for no-working professionals

- Other than working professionals we can call an offer discount to customers who wanted the course but unable to join due to fund issues.
- We can work on students and offer management/communication skill courses along with main course to convert more leads.