Lead Scoring Case Study Summary

Problem Statement:

X Education sells online courses to industry professionals. X Education needs help in selecting the most promising leads, i.e., the leads that are most likely to convert into paying customers. The company needs a model wherein you 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, in particular, has given a ballpark of the target lead conversion rate to be around 80%

Summary

Step 1: Reading and Understanding the Data and Importing the dataset and looking at the first 6 records.

Step 2: **EDA**

- Looking at the distribution of the target variable we can conclude that It's not an imbalanced dataset
- Removing Columns with high null values except Lead Quality and Label the missing values with 'Not mentioned'
- Dropping Lead Profile column with 74.188312% null values and also unique values.
- Inference Maximum number of leads are generated by Google and Direct traffic. Conversion Rate of reference leads and leads through welingak website is high.
- To improve overall lead conversion rate, focus should be on improving lead conversion of olark chat, organic search, direct traffic, and google leads and generate more leads from reference and welingak website.
- Scatter plot to see the distribution between Total no of visits and Page views per visits. We have seen the outliers as well as trend.

Step 3: Dummy Variables Creation

- Encoding using get dummies for the categorical variables.
- Removed all the repeated and redundant variable.

Step 4: Test Train Split & Correlation Matrix

- Used Scaling to scale the original numerical variables.
- plot the heatmap to check the correlations among the variables.

Step 5: Model Building:

- We have achieved a pretty good Recall score and ROC-AUC score and creating a data frame with a probability score adjusting the optimal cut-off point as 0.7

- Experimentation with other classification models like Gradient Boosting Classifier, Decision Tree and Random Forest
- we have seen Logistic Regression performing well in test dataset. Random Forest giving us 94% accuracy on test data. Logistic Regression performed just 1% less compared to Random Forest
- So, we will choose Logistic Regression model this time because of the model interpretation. We can easily find the coefficient and p-value of individual feature by summarizing the model.
- Selecting top 15 features using a hybrid approach: after looking at p-value, VIF, RFE output and applying Random Forest's Feature Importance technique.
- Top 15 variables are well enough to explain 92.6% of model performance. So, we will ignore the other features which are having very less explanation power for this problem statement.
- We plot the ROC Curves and it summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds.
- Precision-Recall curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds.

Step 6: Conclusion:

- We have balanced dataset in this case study and here, the ROC-AUC score on test dataset is 98.1%, which is mind blowing.
- Features which contribute more towards the probability of a lead getting converted are:

    1. Tags_will revert after reading the email
    2. Total Time Spent on Website
    3. Tags_ringing