Lead Scoring Case Study Summary

Problem Statement

An education company names X Education sells online courses to industrial professionals. The company wants to identify the potential customers, also known as “Hot Leads” who have a high conversion chance. The company wants us to build a model where we assign a lead score to each lead between 0 and 100 such that the customers with a higher lead score have a high conversion chance and the customer with a lower lead score have a lower conversion chance.

Approach:

Data Cleaning:

- The dataset had some values as ‘select’ the categorical columns. We replaced the ‘select’ values with Nan values because as per the dictionary they are equivalent to null values.
- Then we checked for the percentage of null values present in each column and dropped the columns having more than 45% of nulls.
- For the remaining columns nulls to impute the null values we have created new categories for the nulls.
- Then we checked for the outliers in numerical columns in and, removed the outliers using the capping method.

Exploratory Data Analysis:

- We have done Univariate and Bivariate analyses on the both numerical and categorical columns. The bivariate analysis was done concerning the target variable.
- We found that some of the categorical columns are irrelevant for the model, as some were sales generated and some were highly skewed. Hence, we dropped such columns.

Data Preparations:

- We have dropped all the identified irrelevant columns.

- The **dummy variables** were created for the remaining categorical columns. Then we dropped the highly correlated columns.
- We have **split the dataset** into train and test datasets with a ratio of 7:3 respectively.
- Then we have scaled the continuous variables of the train data set using Standard Scaler.

Model Building:

- RFE was used to select the top 15 variables for the build model. Then rest of the variables were eliminated manually on basis of their p-values and VIF. The variables with p-values less than 0.05 and VIF less than 5 were kept.
- At last we got the model with the top 7 significant features.

Model Evaluation:

- We have used specificity and sensitivity metrics.
- We have found the optimal cut-off to be 0.3 using the sensitivity, specificity, and accuracy curve. The sensitivity, specificity, and accuracy were around 80 %.
- The area under the ROC curve was around 83%.
- The final conversion rate on train data was around 85%.
- We have also checked the precision and recall with accuracy, sensitivity, and specificity for our final model on the train set.

Prediction on test dataset:

- Then we implemented the learnings to the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics.
- Found the accuracy value to be 70.4%; Sensitivity= 84.4%; Specificity= 62.39%.
- The final conversion rate on test data was around 84%.

Conclusion:

- The features that contribute the most towards the probability of leads getting converted are :

    - Lead Source_Welingak Website.
    - Current_Occupation_Working Professional
    - Current_Occupation_Other

- The lead score calculated for the test dataset and train dataset shows the conversion rate of 84% and 85%, which meets the expectation of the CEO has given a ballpark of the target lead conversion rate to be around 80%
- The sensitivity of the model is around 84%, which will help us to select the leads that have more chances of getting converted.