Summary

X Education receives a lot of leads, but only about 30% of those leads actually become customers. The business wants us to create a model in which we score each lead individually so that leads with higher scores have a higher chance of converting. The CEO aims to convert leads at a rate of about 80%.

Data Cleaning:

- Columns containing more than 40% null values were eliminated. To determine the appropriate course of action, the values of categorical columns were reviewed. If imputing data resulted in an imbalance, then the column was either removed, a new category called others was created, or the most commonly occurring value was inputted. Any columns that did not provide valuable information were also removed.

- Categorical data represented by numbers were filled with the mode value, and columns that contained only one unique response from customers were removed.

- Different tasks such as addressing atypical data, correcting inaccurate information, categorizing infrequently occurring values, and converting binary data were completed.

EDA:

- The verification of data imbalance revealed that merely 38.5% of the leads were converted.

- I analyzed individual variables as well as the relationship between two variables, for both categorical and numerical data. Words such as 'Lead Origin', 'Current occupation' and 'Lead Source', among others. Share valuable observations regarding the impact on the objective variable.

- The more time spent on a website, the greater the likelihood of converting leads into customers.

Data Preparation:

- I generated artificial characteristics (in one-hot encoding format) to represent variables that are categorical.

- Dividing the data into two sets, one for training and one for testing, in a ratio of 70:30.

- Using standardization to scale features.

- I removed some columns as they had strong correlation with each other.

Model Building:

- The RFE technique was employed to decrease the number of variables from 48 to 15. By doing this, the dataframe will become easier to handle.

- Models were constructed using the method of Manual Feature Reduction, wherein variables with a p-value of over 0.05 were removed.

- Before achieving the stable Model 4, a total of 3 models were constructed, in which their p-values were greater than 0.05. There is no indication of multicollinearity as the VIF value is less than 5.

- After considering various models, we ultimately chose logm4, which included 12 variables. We employed logm4 to make predictions on both the train and test sets.

Model Evaluation:

- A matrix of confusion was created and a threshold of 0.345 was chosen after considering the accuracy, sensitivity, and specificity chart. This resulted in an approximately 80% level of accuracy, specificity, and precision. While relying on precision and recall metrics, the performance indicators were around 75%.

- The CEO requested to increase the conversion rate to 80% in order to address a business problem, but when we analyzed the precision-recall view, the metrics decreased. Therefore, we will opt for the sensitivity-specificity perspective as our ideal threshold for making conclusive forecasts.

- The train data was given a lead score using 0.345 as the threshold.

Making Predictions on Test Data:

- Creating forecasts for the examination by utilizing the last model and scaling.

- The evaluation measures for both the training and testing are almost equal and stand at approximately 80%.

- The lead score was given.

- Top 3 features are:
  - Lead Source_Welingak Website
  - Lead Source_Reference
  - Current_occupation_Working Professional

Recommendations:

- More budget/spend can be done on Welingak Website in terms of advertising, etc.
- Incentives/discounts for providing reference that convert to lead, encourage to provide more references.
- Working professionals to be aggressively targeted as they have high conversion rate and will have better financial situation to pay higher fees too.