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%

Solution Summary:

Step1: Reading and Understanding Data
Read and analyze the data.

Step2: Data Cleaning:

- We imputed the 'Select' with NULL values.
- We dropped the variables that had high percentage of NULL values in them.
- This step also included imputing the missing values with median/mode values.
- Created new classification variables in case of categorical variables.
- The outliers were identified and capped.

Step3: Data Analysis

- Then we started with the Exploratory Data Analysis of the data set to get a feel of how the data is oriented.
- And inferences were drawn on those variables. Some inferences are:
    - Count of unemployed leads is high with more than 70% conversion rate.
    - Working professionals have great conversion rate.
    - None of the specialization has a major impact
    - Time spent on website gives high conversion rate

Step4: Creating Dummy Variables
We created the dummy data for the categorical variables.

Step5: Test Train Split:
The next step was to divide the data set into test and train sections with a proportion of 70-30% values.

Step6: Feature Rescaling

- We used the Standard Scaling to scale the original numerical variables.
- Then using the stats model we created our initial model, which would give us a complete statistical view of all the parameters of our model.

Step7: Feature selection using RFE:

- Using the Recursive Feature Elimination, we went ahead and selected the 20 top important features.
- Using the statistics model generated, we recursively tried looking at the P-values and VIFs in order to select the most significant features and dropped the insignificant ones.
- Finally, we arrived at the 15 most significant variables. The VIF's for these variables were also found to be good (all less than 5).
- We then created the dataframe with the actual Converted values and the predicted probabilities taking cutoff as 0.5. i.e. if the probability is greater than 0.5, the predicted value will be 1 else 0.
- Based on the above assumption, we derived the Confusion Metrics and calculated the overall Accuracy of the model.
- We also calculated the 'Sensitivity' and the 'Specificity' matrices to understand how reliable the model is.

Step8: Plotting the ROC Curve
We then tried plotting the ROC curve for the features and the curve came out be pretty decent with an area coverage of 87% which indicates a good predictive model.

Step9: Finding the Optimal Cutoff Point

- Then we plotted the probability graph for the 'Accuracy', 'Sensitivity', and 'Specificity' for different probability values.
- The intersecting point of the graphs was considered as the optimal probability cutoff point. The cutoff point was found out to be 0.38

- We could also observe the new values of the Accuracy: 78.85% Sensitivity: 78.94% Specificity: 78.80%
- Then we calculated the lead score and figured that the final predicted variables approximately gave a target lead prediction of 79%

Step10: Computing the Precision and Recall metrics

We also found out the Precision and Recall metrics values came out to be 69% and 77.10% respectively on the train data set.

Step11: Making Predictions on Test Set

Implementing the learnings on the test model and calculated the conversion probability based on the Sensitivity and Specificity metrics and found out the accuracy value to be 78.41%; Sensitivity= 77.10%; Specificity= 79.18%.

Step12: Final conversion rate and Recommendations

- From the model, we have optimized the conversion rate of approx**. 38%** to a final conversion rate of **79%.**
- The conversion went up by eliminating the features which were having a lot of noise in data and were not helping for lead conversion.
- Based on this model, the sales team will know the key indicators and the ideal lead score which they should analyze before reaching out to the customer.
- Along with boosting the conversion rate, this model will also help in increasing efficiency of the team and provide them with a more focused approach of lead conversion.