Leading Score Summary

Problem Statement

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses.

The company markets its courses on several websites and search engines like Google. Once these people land on the website, they might browse the courses or fill up a form for the course or watch some videos. When these people fill up a form providing their email address or phone number, they are classified to be a lead. Moreover, the company also gets leads through past referrals. Once these leads are acquired, employees from the sales team start making calls, writing emails, etc. Through this process, some of the leads get converted while most do not. The typical lead conversion rate at X education is around 30%.

The CEO, In particular, has given a ballpark of the target lead conversion rate to be around 80%

Summary

- **Reading and Understanding the data**
  - Number of rows and columns.
  - Data types of each columns.
  - Checking first few rows how data looks.
  - Checking how the data is spread.
  - Checking for duplicates.

- **Cleaning data**
  - dropping variables that have all unique values and are of no relevance.
  - Dropping all the columns with more than 35% missing values.
  - Checking for missing values.
  - Handling missing value.

- **Data Visualization and Outliers Treatment**
  - We performed univariate analysis on categorical column to see which columns makes more sense and removed those columns whose variance is nearly zero.
  - Univariate analysis of numeric features.

- **Data preparation for modelling**
  - Creation of dummy variables.

- **Train Test split**
  - The split was done at 70% and 30% for train and test data respectively.

- **Feature Scaling**
  - We use Standard Scaler to scale numerical variables.
  - Plot heatmap to see the correlation matrix.

- **Model Building**
  - Running RFE with 12 Variables as Output.
  - Dropping column with high p-value.
  - Model 5 seems to be stable with significant p-values, we shall go ahead with this model for further analysis.
  - The ROC Curve should be a value close to 1. We are getting a
  - good value of 0.85 indicating a good predictive model
  - From the curve above, 0.3 is the optimum point to take it as a cut-off probability.
  - Accuracy => 76.93%, Sensitivity => 82.63%, Specificity => 73.43%
  - Precision score => 65.5%
  - Recall score => 82.6%

- **Conclusion**
  - While we have checked both Sensitivity Specificity as well as Precision and Recall Metrics, we have considered the optimal cut off based on Sensitivity and Specificity for calculating the final prediction.
  - Accuracy, Sensitivity and Specificity values of the test set are around 78%, 83% and 75% which are approximately closer to the respective values calculated using the trained set.
  - Also, the lead score calculated in the trained set of data shows the conversion rate on the final predicted model is around 80%
  - Hence overall this model seems to be good.