Summary

X Education is an education company that sells online courses to industry professionals. They seek assistance in identifying the most promising leads most likely to become paying customers. They need a model that assigns a lead score to each potential customer, indicating their likelihood of conversion. The higher the lead score, the greater the chances of conversion. The CEO has set a target conversion rate of around 80%.

Data Reading and Understanding

- Imported a few libraries which are required in the initial part.
- Imported the data and examined its shape, data types, and descriptive statistics to get an overview.

Data Cleaning

- There were select values in a few columns. These values were in the data because the customer didn't select any option. It is as good as Null. Hence, we converted the values to Null.
- Checked the values in the data and dropped all the columns that had null values of more than 40%.
- After dropping a few columns, we imputed the null values in a few columns with the mode and also replaced some null values with "Others"
- All the columns which have null values less than 2% we dropped the rows.
- Checked the data imbalance and also checked the duplicate values.

EDA

- Did categorical analysis using counterplot and numerical with boxplot.
- Checked for the outliers and treated them using Capping.
- Checked the correlation between the columns.

Data Preparation

- Converted all the Yes and No values to binary values.
- Created dummy variables for categorical columns.
- Splitting the data into Train and Test sets with a 70:30 ratio.
- Scaling the variables with a Standard scaler.
- Did feature selection using RFE and selected 20 features.

Model Building

- Initiated model building and systematically removed columns with P-values greater than 0.05 one by one.
- Verified the VIF values and confirmed that all columns had VIF values less than 5.
- Model 4 was selected as the final model with 17 columns. All the columns had p-values less than 0.05 and VIF less than 5.

Model Evaluation

- Predicted the values on the train set.
- Choose an arbitrary cut-off probability point of 0.5 to find the predicted labels.
- Made the confusion matrix and calculated Accuracy, Sensitivity, Specificity, positive predictive value Negative predictive value.
- Plotted the ROC curve and got a value of 0.89.
- Created columns with different probability cutoffs and calculated accuracy sensitivity and specificity for various probability cutoffs.
- Plotted accuracy sensitivity and specificity for various probabilities and chose 0.35 as the cutoff probability.
- Created a confusion matrix using the new cut-off and calculated accuracy, sensitivity, specificity, positive predictive value, and negative predictive value.
- Also showed the Precision and recall tradeoff.
- Got the Accuracy: 81% Sensitivity: 81% Specificity: 80.4%

Prediction of Test Data

- Did scaling of the test data and predicted using the final model.
- Created a confusion matrix using the new cut-off and calculated accuracy, sensitivity, and specificity.
- Assigned lead score to Test Data.
- After running the model on the test data, we obtain the Accuracy: 80.5 % Sensitivity: 80.8 % Specificity: 80.3 %.

Recommendations

- Lead Add form has approx 90% conversion rate but the leads are less. X Education should focus on this.
- Reference leads and those from the Welingak website have high conversion rates. So, the company should focus on this more.
- The company should make calls to the leads who are the `working professionals` as they are more likely to get converted.