Solution Summary

To perform this assignment, I have used the following steps:

1. Understanding the data and the domain
2. Data Cleansing
3. EDA
4. Data Preparation
5. Model Building
6. Model Evaluation

Understanding the data and the domain

Before proceeding with the solution, the features in the data are checked in the data dictionary provided to get a detailed understanding of the data. The csv file is loaded into a pandas Dataframe to draw information like number of rows, columns, datatypes, statistical information of numerical columns etc.

Data Cleansing

Once the data has been loaded, the data has been cleaned by removing the columns which have more than 40%missing values and the remaining null values are imputed depending on the type of data. Few columns were dropped as they do not seem necessary for the analysis or model development.

EDA

After cleaning the data, various visualizations were used to perform univariate, bivariate and multivariate analysis to drew inferences from the data. With the help of EDA few more unnecessary columns were identified and dropped and the outliers in continuous columns.

Data Preparation

Once EDA is completed, data preparation steps like creating dummies for non-binary categorical columns, splitting the data as train and test with 80:20 ratio, scaling using min-max scaler were performed.

Model Building

With the help of RFE from sklearn , 15 features were selected for the first model and the model is generated. Based on the p-value and VIF the features are dropped one by one and a new model is generated with the left over features. The steps are iterated until all features in the model have optimum p-value and VIF.

Model Evaluation

Once the model is finalized, the Conversion probabilities are predicted. An arbitrary cut-off of 0.5 was selected to filter the leads initially and the confusion matrix is generated. The accuracy,

sensitivity and specificity were calculated. Then an optimum cut-off has been selected and the final predicted values are generated on train data. Once this is done the test data was used to predict the conversion probability. Later the Recall and Precision were calculated on based on precision_recall curve the optimum cut-off was selected again and the train and test data were again evaluated with the help of evaluation metrics.

Conclusion

With the help of the final model an accuracy of 81%, a sensitivity/recall of 77%, a precision of 77% and a specificity of 84% were achieved. The model has achieved a sensitivity of 77% which is close to the desired 80% conversion rate of the X Education Company.