Summary

1. Introduction:

This study was conducted to enhance efficiency of a lead qualification process, in order to attract more industry professionals in online courses provided by X Education. The historical lead conversion data is used to develop a Python-based Model solution that could predict high precision of a lead converting into a customer.

- Problem Statement:

The existing lead qualification process was labor-intensive, time-consuming, and hence hindered the ability of the sales team to prioritize high potential leads. The objective here was to implement a Machine Learning model to automatically score and rank leads based on their likelihood of conversion.

2. Approach :

1. ***Data Understanding***
    1.1. Extracted the existing lead details and their conversion outcomes.
2. ***Data Cleaning***
    2.1. Replaced Null values, found in the data with, Others.
    2.2. Columns with a relatively high percentage of null values were dropped.
    2.3. Missing values imputed with Median/Mode to obtain new classification of categorical variables.
3. ***Exploratory Data Analysis***
    3.1. Derived percentage of retaining rows while dropping all the irrelevant categorical variables.

3.2. Found a few numerical variables to have outliers prompting the data to be considered between 5 and 95 percentiles.

4. ***Data Preparation***
   4.1. New Dummy variables created for Categorical value
5. ***Training and Validation***
   5.1. Data was split into train and test data sets with a proportion of 70-30% respectively.
6. ***Feature Scaling***
   6.1. Original numerical values were scaled using MinMaxScaler.
7. ***Model Building***
   7.1. Used Recursive Feature Elimination to get the top 20 most relevant variables.
   7.2. Model was created on those variables which have seemingly low VIF values and low probability.
   7.3. Landed on a model with 12 variables that were in direct or inverse relation with the probability of a lead getting converted into a buyer.
8. ***Model Evaluation***
   8.1. Initial assumption for assigning '0' or '1' as a lead score, was based upon the probability being less than or more than 0.5 respectively.
   8.2. Created the data frame having the converted probability values.
   8.3. Confusion metrics were derived based on these initial assumptions.
   8.4. Accuracy, sensitivity and specificity of the model were calculated to evaluate the model's reliability.
9. ***ROC curve***
   9.1. Model achieved is highly accurate since it has a larger area under the ROC curve.

3. Results :

- The lead scoring model achieved an impressive **80.4** % accuracy in predicting lead conversion.
- The sales team should now be able to efficiently prioritize high-scoring leads, resulting in **118 %** improvement in the conversion rates.

4. Learnings :

- Emphasized the importance of clean and relevant data for accurate predictions.
- Highlighted the significance of thoughtful feature engineering in refining model performance.
- Underlined the necessity of rigorous training and validation procedures for robust model outcomes.

5. Conclusion

Sales teams to selectively focus on leads having a higher conversion potential, providing substantial boost in lead conversion rates.