SUMMARY

Introduction:

X Education, an online education company, approached us to help improve their lead conversion rate. The company receives a significant number of leads daily, but the conversion rate is relatively low. Our goal was to build a logistic regression model that assigns a lead score to each potential lead, indicating the likelihood of conversion. The target conversion rate set by the CEO was around 80%.

Solution Approach:

Logistic regression model of ML is considered for solving the requirement of lead conversion. As the Target converted variable is a categorical variable.

Steps Performed:

For building a logistic regression model following steps were followed:

- Data Understanding
- Data Cleaning and missing value treatment
- EDA
- Data preparation required for Model building – Dummy variable creations and any other data conversions from categorical to binary.
- Train and Test data split with a ratio 70-30 respectively.
- Numerical data scaling using Min Max scaler.
- Model building using RFE.
- Re-iterate the model building by dropping the high p value and high VIF .
- Ensure final model is with p values less than 0.05 and VIF less than 5.
- Make Prediction, calculate the probability and Metrics calculations with 0.5 cutoff
- Plot ROC curve and finding optimal cut-off using accuracy, sensitivity, and specificity.
- Make Final prediction and metric calculation with optimal cutoff value.
- Build the test model with the 30% data. Scale the numerical data and make predictions.
- Calculate the metrics using the optimal cutoff.
- Metrics like accuracy, sensitivity, specificity, Precision and Recall are calculated
- Comparison of Train and Test calculations. Ensure they are matching

- **Model Observations:**
  - Accuracy percentage is 80% in both Train and Test data. This means the predicted data is correct.
  - Sensitivity percentage is 81% and 79% in train and test data respectively. This value implies that the model has predicted the positive instances correctly.
  - Specificity percentage is 79% and 80% in train and test data respectively. This value implies that the model has predicted the negative instances correctly.

  - Precision percentage is 71% and 70% in train and test data respectively. This value implies that the model has predicted positively.
  - Recall percentage is 81% and 79% in train and test data respectively. This value implies that model has predicted true positives and minimized false negatives.
  - Lead score is calculated using the conversion probability multiplied with 100

- **Data Observations:**
  - Potential leads that are getting converted are spending enough time in website.
  - Occupation is an important variable as the courses designed and offered by the company should suit the occupation. Working professionals have converted more in order to maximize their career prospects.
  - Lead sourcing and lead activity and lead origin are important variables.

- **Learnings:**
  - Learnings from this case study from business perspective
    - How logistic models can be built to solve the business problems like lead scoring.
    - Although not all data provided was relevant for the case study, it was a learning to understand how it can be used in model building.
    - Using the ROC curve, optimal cutoff and how metrics can be used to measure the model.