Business Objective:

- Help X Education to select the most promising leads(Hot Leads), i.e. the leads that are most likely to convert into paying customers.
- To build a logistic regression model to assign a lead score value between 0 and 100 to each of the leads.

Approach:

The approach has been to divide the entire case study into various checkpoints to meet each of the sub-goals. The checkpoints are represented as below:

- Understanding the Data Set & Data Cleaning
- Exploratory Data Analysis
- Data Preparation for Modelling
- Feature Selection using RFE
- Model Building
- Predicting the conversion Probability on Train dataset
- Finding optimum probability threshold
- Plotting the ROC curve and calculating AUC
- Evaluating the model on train dataset
- Making Prediction on test dataset
- Evaluating the model on Test dataset
- Lead Score calculation on complete data
- Determining Important Features

Logistic Model metrics on train dataset:
We evaluated the train model on different metrics as below:

- Accuracy: 0.9208
- Sensitivity: 0.9207
- Specificity: 0.9209
- Precision: 0.8765
- Recall: 0.9207
- F1 score: 0.8980

Logistic Model metrics on test dataset:
We evaluated the train model on different metrics as below:

- Accuracy: 0.9052
- Sensitivity: 0.8988
- Specificity: 0.9092
- Precision: 0.8636
- Recall: 0.8988
- F1 score: 0.8809

Important Features:
We used model features coefficients to determine the important features. Higher the value, higher the importance. We came up with 3 important features which will affect the lead conversion.

- Tags_Closed by horrizon
- Tags_Lost to EINS
- Tags_Will revert after reading the email

Challenges:

1. Data was skewed and had null values so cleaning data was a major challenge. We started with 37 columns but ended up with only 15 features for model building.
2. Making final model with significant p-value (less than 0.05) and no multicollinearity (vif less than 3) and high sensitivity
3. Accurate prediction on test dataset with high accuracy and high sensitivity

Learnings:
1. Data cleaning is most important part for analysis and model building
2. Exploratory Data Analysis helps to understand and visualise data better
3. Scaling data helps to bring all the variables at same scale
4. Start with enough features for model building
5. Drop variables with high significance and with multicollinearity
6. Metrics are important to evaluate the model performance
7. Business understanding is equally important to evaluate any model