Summary Report

1. In order to help X-Eduction select the most promising leads logistic regression was performed on the given datset.

2.The first process performed was EDA(Exploratory Data Analysis) -

- The variables which had null values more than 70% were ropped
- For the remaining variables with null values < 70% either mode or a new category 'Others' was used to impute the missing values
- Few columns had 'Select' level. It was replaced by Nan
- The outliers in the continuous numeric columns was handled using 1.5 IQR method
- Dummy vaiables were used to represent categorical variables with two or more categories

3.The next step was model building

- Created the train test split
- Performed feature scaling for the numeric continuous variables
- Checked for correlations. A few variables were highly correlated
- Built the first model and checked for the p-values. Some of the variables were not significant
- Used RFE to identify the 15 most relevant featurs
- Using the above , identified features and built the next iteration of the model
- Iterated this process until the p-values and VIF were within acceptable limits

4. Makings Predictions

- The conversion probabilites were computed on the train set
- The various metrics – Accuracy , Sensitivity , Specificity were computed by using a conversion threshold of 0.5
- Next ROC curve and trade-off bewtween Precision and Recall was used to find the optimal probability and the metrics were recomputed
- Using this optimal probability, predictions were made on the test set and metrics compared to validate that the model is a good fit

5. Calculating Lead Score

- For each of the leads in the dataset a lead score was assigned using the formula 100 * Conversion Probability

6. Top Features

- From the final model the relative co-efficients were calculated for each of the selected features
- The Top 3 features which contribute most towards the lead getting converted were identified
- Similarly the Bottom 3 features i.e. The features that need improvement to convert a lead were identified

7. Learnings

- The importance of EDA prior to model building.
  - ✔ Key insights from inspecting the data help to treat the data correctly
  - ✔ If a feature has the same value across all the rows in a dataset then it does not contribute to predictions and can be dropped
  - ✔ Data Cleaning – imputing missing values , outlier treatment , feature scaling helps to build a efficient model
- RFE is a good technique to identify the key features for model building
- Understanding the significance of the various metrics is important to determine a optimal

cut-off for the converted probability

■ Plotting ROC curve , Pecision-Recall curve is a good way to determine optimal cut-off

■ Technique to identify the Top features that contribute positively towards the problem in hand. Similarly technique to identify the Top features that need most improvement .