SUMMARY

This analysis is done for X Education to boost its business **by predicting its 'Hot Lead',** which increases the conversion rate which is at 30% in the current situation. We need to find the most important features which affect converting the leads to take up its courses. Also, find the loopholes in current business marketing to get high conversions.

To achieve the above-mentioned goals, we have used the **Logistic regression model** to predict the hot Leads and the most important features affecting the business.

The following are the steps used to build the model:

1. Cleaning data:

The data has features which have more than 41% null values. Some of the features also have select as their values which means the customer didn't select any option so, it can be treated as null values. Initially, the features having null values have been dropped later the 'select' was replaced with Nan and the features having more than 51% null values were dropped. The remaining null values were imputed with random values from the features so that we will be retained with data and the distribution of the data remains unchanged.

2. EDA:

A quick EDA was done to check the condition of our data. It was found that a lot of elements in the categorical variables were irrelevant so, they were dropped later. The numeric values also have outliers which were handled by taking the values in the 99th percentile. Based on our EDA we were able to get a glance at some of the important features that are affecting the conversion rate.

3. Train-Test split:

The split was done at 70% and 30% for train and test data respectively.

4. Column transformer:

Column transformer was created to create dummy variables for categorical features using OneHotEncoder and the Standard scalar is applied to scale the numerical features.

5. Model Building:

A total of four models were trained.

Model1: The first model was trained using all the features. We were able to achieve accuracy up to 81% but a very recall of 69% and precision of 76%.

Model 2: The feature selection was done Using the RFE. 15 features were selected and used for the next model in which we were able to achieve an accuracy of 80% and still a very low recall of 67% and precision of 76%.

Model 3: Feature importance is calculated using a Random forest algorithm and the top 20 features were used for the next model. The accuracy score got dropped so insignificant and high VIF features were dropped from the model.

Model4: Features with good significance and low VIF were used to achieve an accuracy of 79%.

6. Model Evaluation:

A confusion matrix was made. Later on the optimum cut-off value (using the ROC curve) was used to find the accuracy, sensitivity and specificity which came to be around 80% each.

7. Prediction:

The prediction was done on the test data frame with an optimum cut-off of 0.37 with accuracy, sensitivity and specificity of 80%.

8. Precision – Recall:

This method was also used to recheck and a cut-off of 0.41 was found with a Precision of around 73% and a recall of around 75% on the test data frame.