Summary Report

Analysis is done for an education company named x education which sells online courses for industry professionals. We have helped sales team of this company in finding out hot leads which have more chances of becoming paying customers. Initially the conversion rate was just as low as 38% and we have increased that to 86% with the help of logistic regression model. Initially leads were too high in number and we have filtered these leads into hot leads which saves lots of time and money. Hot leads are the people which are very likely to become the paying customers or we can term them as conversions.

The steps we've used in this analysis are:

1. **Data Sourcing, Cleaning and Preparation**:
   We already have the data so the first step that we have performed is to read and understand it. After that we have cleaned the data by checking for the null values, performing outlier treatment, checking for duplicate values, checking data types etc. Moving on we have standardize data into correct format. Now that data is cleaned and ready, we move onto further analysis.

2. **EDA- Univariate and Bivariate Analysis**:

   EDA is termed as the most crucial step for data analysis. In this step we have tried to get the insights from the data with the help of univariate and bivariate analysis. Firstly, we have performed univariate analysis in which we have analysed every single numerical column. Then we have analysed two or more columns together to get more insights about the data in bivariate analysis. Also, we have removed some columns with the help of EDA which we don't find useful. Further we perform feature scaling and test train split.

3. **Feature Scaling and Test Train Split:**

   In this step we have created dummy variables for the categorical columns to get a clear picture of useful and useless variables. Along with that we have also included drop first equals to true to make our data more concise and understandable. After that, we have divided our data set into train data and test data. On train data we will do our modelling and finally test that model on test data set. We have use sklearn library for this and for feature scaling we have used standard scaler from sklearn.

4. Model Building and Evaluation:

The first step in model building is feature selection so we have selected top 15 features which are necessary for our analysis with the help of RFE which is an automated process. Now with the help of stats model we have checked the p value and the coefficient value for our variables and also removed some variables which have high p value. Then we have built a logistic model with the help of GLM and calculated predicted value for the train data set from the model. Further we have calculated confusion metrices and VIF values for our final variables which turned out to be great. Then we have calculated different metrices such and accuracy, sensitivity, specificity, recall, precision, false positive rate, true positive rate etc. and then plotted roc curve and other two accuracy, sensitivity, specificity graph and precision and recall graph for train data set to find the optimal threshold value for the conversion probability. Finally, we have assigned a lead score to leads for the train data set.

5. **Prediction on Test Data Set:**

   At last, we have predicated values on the test data set using the accuracy, sensitivity and specificity

   Cut off threshold value and assigned lead score to leads which all turned out to be quite similar with the train data set. So overall the model was good and we have successfully managed to increase the conversion rate up to 86% by filtering out the important leads as hot leads which saves time for the sales team and money for the x education.