SUMMARY REPORT

The purpose of this assignment was to help company X education to find out the most promising leads. i.e., leads that are most likely to convert into paying customers.

Data:

The data provided had 9240 data points. It consisted of various attributes like Lead Source, Total Time Spent on Website, Total Visits, Last Activity, etc. The target variable, in this case, is the column 'Converted' which tells whether a past lead was converted or not wherein 1 means it was converted and 0 means it wasn't converted

It was decided to perform logistic regression to arrive at the results. The following steps were undertaken:

Data sanitization:

There were few null values in the data. Columns with more than **40% null values** were dropped. Some of the null values instead of dropping were imputed as 'no info' to prevent data loss. Additionally, few unnecessary columns like *'prospect_ID'* which is a unique identifier were dropped.

EDA:

For numerical variables, the outlier check looked satisfactory. Univariate and bivariate analysis were done for the variables to look for any trends.

Dummy variables:

Dummy variables were created for the categorical variables.

Scaling:

Scaling was performed on the numerical variables using **minmaxscaler** method.

Test-Train Split:

Split was done with **33% as test data and 67% as the train data**.

Model Building:

Optimal number of features was computed as 28 which gives **0.94 accuracy**. Later RFE was done to get **28 features** that we will be using for model building. Later VIF and P values was compared to drop few more columns to finally arrive at features which had P values less than 0.05 and VIF < 5.

Model Evaluation:

Using ROC curve optimum cut off value was used to find the accuracy, sensitivity, specificity which came to be approximately 90%. Additionally Precision and F-score was also calculated and that came out around 90% as well.

Prediction:

Prediction was done on test data with accuracy, sensitivity, specificity, precision and F-score at approx. 90%. Precision-Recall cut-off was found to be ~ 0.45.

Conclusion:

Some of the features that matter the most which company X should leverage are:

1. Total time spent on the website.
2. Total number of visits.
3. Last activity was SMS.
4. Source is welingak website
5. Country is Germany
6. Tags closed by horizon