# Lead Score Case Study

## Summary Report:

We have built a Logistic Regression model to predict the leads conversion where column **"Converted"** is our dependent feature while other columns are independent features.

**Step 1 - Data Import:**
As our data is in CSV format so we import it using pandas.read_csv() function.
**Step 2 - Inspecting the Dataframe:**
We check the data quality by using python inbuilt function like describe(), info(), isnull(), and attributes like shape, columns, dtype (for data types) etc. Outliers are checked by plotting a box-plot. For missing values and improper data like 'select' we have imputed it with mean() for numeric columns and mode() for categorical columns.
**Step 3 – Data Preparation:**
After cleaning data we execute data preparation steps where we map features with binary data with 1's and 0's. For categorical features with more than 2 values we create dummy features.
**Step 4 - Test-Train Split:**
After data preparation step we divide the data into X and y where X contains all the independent features and y contains the dependent feature. Once data has been loaded into X and y; we split X and y into X_train, X_test, y_train and y_test. We split the X and y data for train and test by 70% and 30% respectively.
**Step 5 - Feature Scaling:**
Once Train Test split is performed we need to perform feature scaling for numeric features. Here for feature scaling we have used Standard Scaler. ***Note:*** *We do not perform scaling on dummy features.*
**Step 6 - Model Building:**
Here we build our 1st Logistic Regression model using GLM model. By adding constant to X_train we pass X_train and y_train to the GLM model where we also pass the families attribute with value as "Binomial family", then we fit the model and look at the summary statistics like P Values and coefficients. We also check the VIF (Variance Inflation Factor) and based on VIF and P Value we drop the features with high P value and VIF.
**Step 7: Feature Selection Using RFE:**
RFE will be used for feature elimination. Providing the number of features to be kept the RFE is performed on all independent features and by RFE we identify the top n features as it was mentioned in RFE. Here we have provided 15 features. The top 15 features are LS-Welingak Website, LA-SMS Sent etc.
**Step 8: Rebuilding models using RFE columns:**
Once we are done with feature selection using RFE we select the model which gives the best accuracy, sensitivity, specificity, precision and recall.
**Step 9: Finding Optimal Cutoff Point:**
After model selection we select the optimal cut off point for the model i.e. where sensitivity, specificity and accuracy intersect each other. We also plot ROC curve to check the AUC we get from the final model.
**Step 10: Making predictions on the test set:**
We run the model on the test set and check its metrics for prediction on unseen data.