**Leads Scoring Case Study**

A brief summary report in 500 words explaining how you proceeded with the assignment and the learnings that you gathered.

**Answer:**
Below are the steps how we have proceeded with our assignments:

**1. Data Cleaning:**

a. First step to clean the dataset we choose was to remove the redundant variables/features.

b. After removing the redundant columns, it was found that some columns are having label as 'Select' which means the customer has chosen not to answer this question. The ideal value to replace this label would be null value as the customer has not opted any option. Hence, we changed those labels from 'Select' to null values.

c. Removed columns having more than 45% null values

d. For remaining missing values, we have imputed values with maximum number of occurrences for a column most of the times

e. For Outlier treatment we used Median for capping

**2. Data Transformation:**

a. Changed the multicategory labels into dummy variables and binary variables into '0' and '1'.

b. Checked the outliers and created bins for them.

c. Removed all the redundant and repeated columns.

**3. Data Preparation:**

a. Split the dataset into train and test dataset and scaled the dataset.

b. After this, we plot a heatmap to check the correlations among the variables.

c. We worked on finding the some correlations and the highly correlated variables dropped.

d. Exploratory data analysis is done performing univariate and bivariate analysis to understand every variable

e. A few variables are dropped based on imbalance and perceived least impact on model

**4. Model Building:**

- We used RFE to reduce the feature count from 70 to 30
- We used VIF methods to eliminate unnecessary features
- We looked at P values and VIF (various combinations of High/Low P and High VIF lead to dropping the variables)
- We ran 12 iterations of Model Building dropping variables one by one based on the P values and coefficients

- Checked the precision and recall with accuracy, sensitivity and specificity for our final model and the tradeoffs.
- Prediction is made now in test set and predicted value was recoded.
- Model Evaluation was done by calculating accuracy sensitivity and specificity for various probability cutoffs
- After plotting the three, an optimal cut off point was found at 0.365
- We found the score of accuracy and sensitivity from our final test model is in acceptable range.
- We have given lead score to the test dataset for indication that high lead score are hot leads and low lead score are not hot leads
- Learnings gathered below:
    - *Test set is having accuracy, recall/sensitivity in an acceptable range.*
    - *In business terms, our model is having stability an accuracy with adaptive environment skills. Means it will adjust with the company's requirement changes made in coming future.*
    - *Top features for good conversion rate in the order relative importance is shown below (Lead Origin~ Lead Add Form, Lead Source~ Welingak Website, Last Activity ~ Had a Phone Conversation)*