SUMMARY OF LEAD SCORING CASE STUDY

By: Tushar Joshi & Apoorva Bhatla

This summary report includes the steps that were followed to solve the business problem and the conclusions drawn.

Step 1: Reading and Understanding the data.

- We loaded the data in the notebook, and went through all the columns, as per the data dictionary.
- Then we do the initial inspection in terms of checking the shape, data types, null values, duplicates and numerical and categorical columns

Step 2: Data Cleaning and Preprocessing

- We first inspect for the 'select' word in the data frame. And convert it into NaN
- Then column wise inspection of null Values, and drop the columns with more than 30% null values
- Drop Highly Skewed Values, and inspect the categorical columns and Numerical columns and impute them appropriately by replacing missing values with mode, or dropping in case of very less missing values
- Perform EDA on how each variable plays a part in converting a lead. And Outliers analysis
- Drop the columns created by the sales team.
- Convert the Yes/No columns with 1/0 and drop the original columns
- Create Dummy Variables and confirm if the Dataset has Null Values

Step 3: Train-Test Split.

- Perform the Train-Test Split and assign the variables to X, y sets
- Perform Feature Scaling of Numerical Variables and check the conversion Rate
- Analyze the correlations
- Build a Logistic Regression Model, using GLM function and assign the constant

Step 4: Model Building using RFE and Manual Feature Elimination

- Do feature scaling using RFE and select top 15 features
- Build our model iteratively until we get the optimum P-Value<0.05 and VIF value
- We found the Y-Value and Conversion Probabilities based on our Train Data Set
- We assigned a random value of 0.5 to predict if the conversion is 1 or 0

Step 5: Confusion Matrix and Parameters.

- Using these values, we derived the confusion Matrix. And Accuracy, sensitivity and specificity
- Based on this we plotted a graph to find the cut-off point, which in this case was
- Based on the cutoff point We created another confusion matrix, and analyzed the sensitivity, specificity value as well as Precision and Recall
- Compared these values to Business requirements to make decisions
- Then we plot a graph for precision and Recall Tradeoff

Step 6: Making Predictions on the Test Set.

- We Scale and Transform the test data
- Derive the test predicted values for Y set
- Create a dataset of predicted values and conversion probability
- Create a Confusion Matrix and derive the parameters, accuracy, sensitivity and specificity
- Compare on the train data and find that all the values are satisfactory
- Create a Lead score, to associate the potential lead, if he will convert or not

Conclusions.

Train set data

- Accuracy =74.65%
- Sensitivity=87.52%
- Specificity=66.25

Test Set Data

- Accuracy =73.96%
- Sensitivity=87.27%
- Specificity=66.53%

Inference

- The CEO expected a conversion rate of 80%, we see that our sensitivity, which is the conversion rate is 87%, so, its a very good model
- The Accuracy, Precision and Sensitiviy for both train and test set are within range.
- In business terms, this model has the ability to adapt to the company's needs of the future.
- The most Important features for positive lead conversion are:
    - **Lead Origin_Lead Add Form**
    - **What is your current occupation_Working Professional**
    - **Lead Source_Welingak Website**