Lead Scoring Case Study Summary

Solution/ Approach:

- Following are the steps performed while achieving the business objective:
  - Reading and Understanding the Dataset
  - Cleaning the Dataset
  - Exploratory Data Analysis
  - Data Preparation
  - Model Building and Analysis
  - Final Conclusions and Recommendations
- Reading and Understanding the Dataset:
  - The dataset provided is large enough with 9240 rows and 37 columns.
  - The columns contain combination of categorical and continuous variables.
- Cleaning the Dataset:
  - There are various columns present in the data which are having null values as well as some values which having 'Select'.
  - We have deleted all those columns having more than 35% of missing values. We have deleted the rows where missing data is below 5%. We have handled and imputed the null values in the columns having less than 35% data. We have removed columns having skewness.
- Exploratory Data Analysis:
  - Univariate and Bivariate Analysis of the columns is performed against the Converted column for better visualization.
  - Outliers are handled using Univariate Analysis for Continuous Columns. The outlier values are capped at 95% quantile value.
- Data Preparation:
  - Categorical columns are converted into dummy variables before model building.
  - The data is split into train data and test data at 70% bias value.
  - The numerical columns are scaled using Standard Scaler.
  - There are some columns having higher correlation between each other. Some of these columns are dropped and some columns are left to analyse during RFE and Model Building.
- Model Building:
  - We have built the model using RFE for top 20 features first and then dropped the features manually by looking at significance and VIF values.
  - We have analysed the model based on VIF and significance and dropped all such factors one by one to reach to the stable model.
- Model Performance Analysis on Train and Test Data:
  - We have analysed the optimum threshold values using cut-off between Sensitivity and Specificity as well as Precision and Recall.
  - We have then predicted the values using optimum threshold probability as 0.38. We have got ~81% sensitivity from the model. This is also confirmed using Precision-Recall trade-off.
  - We have then made predictions on test data also, using the model.
  - We have got ~81% sensitivity score on test data which is as exact that was predicted by the CEO.

- Lead score is calculated based on the predicted probability.

- Final Conclusion and Recommendations:
  - Major indicators that a lead will get converted to a hot lead:
    - Lead_Source_Welingak website - A lead sourced from Welingak Website
    - Lead_Source_Reference – A referral from past customers.
    - Occupation_Working Professional - Working professionals.
    - Last_Activity_Had a phone conversation - A lead already had a phone call.
  - Major indicators that a lead will NOT get converted to a hot lead:
    - Last_Activity_Email bounced - Customer who bounced their email.
    - Last_Activity_Olark chat conversation – Customer having Olark Chat Conversation.
    - Last_Notable_Activity_Modified
  - Recommendations:
    - **The company should use a leads score threshold of 38 to identify "Hot Leads" as at this threshold, Sensitivity Score of the model is around ~81% which is as good as CEO's target of 80%.**