Case Study – Lead Scoring – Summary Report

Submitted by: Sangeeta Kalra & Vikram Mathur

<u>Process Followed</u> The process has been derived from the CRISP-DM Framework.

- Reading & Understanding Data
    a. Dataset shape, info, describe applied to study and understand it.
    b. Running Pandas Profiler – Studying the reports
- Data Preparation and EDA
    a. Handling Redundant Features
        i. Removing Features with Constant Values
        ii. Checking & Removing Features with 95% Constant Values
        iii. Removing ID Features having all Unique Values
        iv. Evaluating Features having High Missing Values
    b. Checking for Duplicates
    c. Checking & Treating Missing Values
    d. Checking & Fixing Data Types
    e. Outlier Analysis & Treatment
    f. Numerical Feature Analysis (Univariate, Bivariate and Segmented analysis)
    g. Categorical Feature Analysis (Univariate, Bivariate and Segmented analysis)

    Studying and understanding the data well so that business decisions can be taken while training the model where business sense needs to be applied.

- Feature Engineering – Data Preparation
    a. Derived Metrics
    b. Dummy Encoding for Unordered Categorical Variables
- Splitting Data into Train and Test Sets
- Training The Model
    a. Standardization – Min Max Scaling on the Numerical Variables
    b. RFE (Recursive Feature Elimination)
    c. Building model using statsmodel, for the detailed statistics
        i. Reiterating the steps to ensure appropriate features are included in the Model based on p-value and variance inflation factor.
- Model Evaluation
    a. Predictions – Predictions / Probabilities being aligned in a DataFrame with the Lead Number
    b. ROC (Receiver Operating Characteristics) Curve – to evaluate if the model is good or not.
    c. Finding Optimal Cut Off Point
        i. Reviewing Metrics like Accuracy, Sensitivity and Specificity.
        ii. Ballpark value of 80% on metrics
    d. Assign Lead Score to Train Set
- Predictions and Evaluation On the Test Set
    a. Evaluating the Test Set by running the model on the test dataset and evaluating all metrics
        i. Evaluating the ballpark value on 80% on metrics
    b. Assign Lead Score to Test Set

<u>Learnings</u>

- End to End understanding of how to approach a practical industry application. Understanding of the kind of business scenario's and projects get handed over and how to approach and deliver.

- Step by Step process to break the entire process of Machine Learning implemented using Logistic Regression
- Biggest learning was regarding data preparation. The importance of ensuring that maximum time is spent on this step to prepare the data as much as possible to avoid re-work in future.
- Further, visualizations are helping to train the eyes to perform better Exploratory Data Analysis.
- Clarity on varied elements of Logistic Regression. Right from clear understanding of metrics and using them as the benchmark to tweak model to be optimally trained.
- Ability to differentiate between different Machine Learning Technique and business and technical applicabilities of each.
- Ability to talk the language and express the work as a data scientist.
- Learning of the ipywidgets library to create a tool for quick univarate and bivariate analysis of Categorical variables.