Lead Scoring Case Study Summary Report

Problem Statement: To make a logistic regression model to assign a lead score between 0 and 100 to each of the leads that can be used by company "X Education" (that sells online courses to professionals) to target potential leads. A higher score would mean that the lead is hot, i.e. is most likely to convert whereas a lower score would mean that the lead is cold and will mostly not get converted

Approach Steps:

1. **Step 1:** Reading and Understanding the Data:
    - The Data set included 37 attributes with max 9240 entries per attribute
    - The Data Imbalance percentage for Target variable "converted" is "1.59%", which is very good for analysis
    - Numeric variables were visualized through Pair-Plot and Correlation matrix.
    - Categorical variables were visualized through Boxplot: "*Lead Origin*", "*Lead Source*", "*Last Activity*", "what is your current Occupation", "Tags", "Lead Quality", "Lead Profile", are seen as variables contributing for conversion.
    - 5 of the variables had constant value "No". They were removed in Data Preparation step.
2. **Step 2:** Data Preparation:
    - No Empty Rows or Columns were present.
    - No Duplicate Rows were found.
    - During analysis, we didn't come across any need for data type change for any of the attributes.
    - "Select" values were converted to NaN values.
    - Checked for Missing Values.
        - 17 attributes were missing data. Out of which 2 attributes with missing percent > 70% were removed.
        - Missing value imputation:
            - Missing values for 'Asymmetrique_Activity_Index', 'Asymmetrique_Profile_Index', 'Tags', 'What_matters_most_to_you_in_choosing_a_course', 'What_is_your_current_occupation' are replaced by its respective max occurrence value.
            - Missing values for ''Specialization' is replaced by value "Others" assuming data is missing as the customer's specialization may not be in the list for selection.
            - Rest around 2% of missing values is removed by dropping the corresponding rows.
    - Dropping attributes
        - dropping columns "How did you hear about X-Education" and "Lead Profile" as their missing values % is high.
        - "Prospect ID" and "Lead Number" are unique identifiers of the leads and not of importance for analysis.

- "Receive More Updates About Our Courses", "Update me on Supply Chain Content", "Get updates on DM Content", "Magazine" and "I agree to pay the amount through cheque" have only single value and hence of no importance. As verified above
- "Country" and "City" of the leads are not of importance for analysis as its for online course.
- 'Asymmetrique Activity Score', 'Asymmetrique Profile Score', 'Last Notable Activity' are duplicate of 'Asymmetrique Activity Index', 'Asymmetrique Profile Index', 'Last Activity' respectively and hence dropped

- Handling Categorical Variables
  - Binary Mapping
    Attributes with “Yes /No” values were converted to “1/0” respectively.
  - Dummy Variable Creation
    Dummy variables created for remaining categorical variables with more levels.
- Outlier Analysis was performed.

Step 3: Model Building with Preprocessing steps:

- Splitting the data into train and test
- Feature scaling
- Calculated Conversion rate: 38%
- Dropped highly correlated variables
- Feature selection using RFE – ran RFE with 15 variables as output
- Dropped columns with high p value, in the following order:
  - Invalid number
  - Number not provided
  - Wrong number given
  - 03.Low
- Created a dataframe with the actual conversion flag and the predicted probabilities
- Created new column 'predicted' with 1 if Conversion_Prob > 0.5 else 0
- Checked VIFs value to confirm the selected features are correct and does not need any further fine tuning
- Calculated model metric parameters and plotted the ROC curve, as well as the Sensitivity vs Specificity vs Accuracy curve to calculate the optimal point/cut off probability (in our case, it came out to be 0.2)
- Then we ran the Train and Test set prediction, to validate our model, and to see the Precision vs Recall Tradeoff
- Finally we ran prediction on the test data set and calculated and assigned lead score to each leads as required by the problem statement