Lead Scoring Case Study

Problem Statement:

An education company named X Education sells online courses to industry professionals. On any given day, many professionals who are interested in the courses land on their website and browse for courses. The typical lead conversion rate at X education is around 30%.
The main Agenda is to help sales team and make the process more efficient by identifying "Hot Leads" based on the data provided with ball park target of lead conversion around 80%.

How to solve the problem:

The best way to solve this problem is by analysing the data and generating the insights based on the logistic regression model to understand which variable has more influence on the conversion rate. Depending on this information we can analyse the probability of the conversion of a particular lead which can be used as a lead score to analyse the possibility of that lead to be converted as customers. Depending on the lead score it will be more easy for sales team to prioritise a particular lead with better time efficiency.

Steps taken to solve the problem:

- **Initially we started with importing libraries, reading and understanding the data to analyse it based on the attributes provided in the data frame.**
- **After analysing the data we started with data cleaning and EDA process.**
    - Initially we replaced the 'Select' level in the attributes to null value as it does not provide any specific information. When a particular customer ignores to fill the information in a particular attribute, 'Select' level gets generated. Thus it should be replaced with Null.
    - We as well checked for the uniqueness for the variable 'Lead Number' and 'Prospect ID' to understand if there is any repetition of variables.
    - We started removing columns with more than 50% of null values as it had more null values and less information. (Columns dropped -> 'How did you hear about X Education', 'Lead Profile', 'Lead Quality')
    - Later we found find the percentage of unique categories in each column. The columns having more than 80% occurrence to a specific level were removed as it does not provide any significant information. (Columns dropped -> 'Last Activity', 'Country', 'What is your current occupation', 'What matters most to you in choosing a course', 'Tags', 'Do Not Call', 'Search', 'Magazine', 'Newspaper Article', 'X Education Forums', 'Newspaper', 'Digital Advertisement', 'Through Recommendations', 'Receive More Updates About Our Courses', 'Update me on Supply Chain Content', 'Get updates on DM Content', 'I agree to pay the amount through cheque', 'City')
    - We as well removed columns which are filled by sales team and are not system generated. (Columns dropped -> 'Prospect ID', 'Asymmetrique Activity

Index', 'Asymmetrique Profile Index', 'Asymmetrique Activity Score', 'Asymmetrique Profile Score')
    - Converting Yes, No values in column with 1 & 0 for better analysis (Columns -> 'Do Not Email', 'A free copy of Mastering The Interview')
    - For better analysis dropping rows having more than 3 null values.
    - Imputing missing values with imputation techniques (Lead Source – Mode, TotalVisits – Mode, Page Views Per Visit – Mean)
    - Finally there are more than 90% rows retained with no null values.
    - Graphical representation of data based on the conversion
- **Data Preparation:**
    - Creating dummy variables for categorical columns. (columns-> 'Lead Origin', 'Lead Source', 'Specialization', 'Last Notable Activity')
    - Removing categories in columns with less than 5% probability.
    - Performing test-train split.
    - Performing standard scaling of numerical continuous data using standard scaler.
    - Checking for conversation rate (38.5%)
    - Exploring the correlation between all the variables using heat map. Dropping column 'Lead Origin_API' as it is highly correlated.
- **Model Building:**
    - Implementing Logistic regression model
    - Using RFE to perform top 15 variable selection
    - Building a Logistic model with good sensitivity
    - Understanding the VIF and p-values for columns
    - Optimal probability cut-off with 0.3
    - Confusion Matrix:
        - Sensitivity -> 82.8
        - Specificity -> 77.5
        - Precision -> 76.6
        - Recall -> 67.7
    - Analysing model performance over test data
        - Accuracy -> 80.4
        - Sensitivity -> 74.6
        - Specificity -> 84.2
    - Implementing the score variable based on the lead conversion probability.