Lead Scoring Assignment - Summary

Problem Statement:

Build a logistics regression model that predicts the probability of the lead conversion, which is Lead Score. Lead score represents the probability of lead conversion. High probability means hot lead and low probability represents a cold lead.

Solution Approach:

Broadly the solution approach contains the following steps

1. Data understanding and visualization
2. Data imputation and Data cleaning
3. Exploratory Data Analysis
4. Model Building and Prediction
5. Model Evaluation and Key observations

1. Data understanding and visualization:

As data understanding is key in model building, we started with that first. Identified the missing values percentage column wise from the leads data. This helped further in data cleaning process.
Like mentioned in the problem statement, we checked for the **Select** value presence in the categorical columns. User did not select any option, in such scenario value might be stored as **Select**. This value is equivalent to null. There are 4 columns that contain this value,

a) Specialization
b) How did you hear about X Education
c) Lead Profile
d) City

In the above columns, replaced all the select values with numpy NaN

2. Data imputation and Data cleaning

Next is the data imputation (null value treatment), dropped all the columns with null value % > 45%. In some of the columns (City, Specialization, Tags, What is your current occupation) where null values are high, identified the less occurred groups and using all those groups created a separate group called Others. In some columns the categories are very less and only one value is present more than 95%. This means it is highly skewed, we dropped these columns (Country,..)
In the numerical columns filled the missing values with median value.

3. Exploratory Data Analysis:

In this step we started with checking the correlation between the numerical variables. TotalVisits and Total Time Spent on Website are highly correlated. Next, we identified the outliers in the numerical data and removed them. After removing the outliers 90.5% data is retained. Did some analysis on the data and some key findings are

a) Lead Add Form has high conversion but less number of leads
b) Google and Direct traffic generate high number of leads
c) On specializations - The emphasis should be on diverse specialities with low lead generation but high conversion rates.

4. Model Building and Prediction:

Started with creating dummy variables for all the categorical columns in the cleaned data. After creating dummy variables there are 59 columns. Now data is ready for train test split. Split the data with the ratio 70:30 => training data – 70%, test data – 30%.

Data Scaling: Using standard scaler, data scaling is performed on the numerical columns ('TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website').
After this step using RFE (Recursive Feature Elimination) selected 12 features columns that are more important in predicting the target variable. Now using the 12 feature columns (X_train), target variable(y_train) model training is done. From the model summary identified a high p-value column, and is dropped from the train data. Next, calculated the VIF's (Variance Inflation Factor) of feature columns to identify the multicollinearity and all are under normal range.

5. Model evaluation and Key observations:

After training the model, next step is model evaluation. Using trained model, generated the Lead Score column which represents the probability of a lead conversion. Here is the model evaluation summary report

a) ROC – 0.96(Higher the value better the classification model performance)
b) Probability Cut-off – 0.34(Lead_Score above this value are classified as 1) c) Model Accuracy – 88.88
d) Confusion Matric –

Predicted	Not Converted	Converted
Actual		
Not Converted	1445	155
Converted	124	785

e) Precision – 83.51
f) Recall – 86.35

Key Observations:

1. There is above 80% chance that our predicted leads will be converted
2. When compared to the model we derived, our Logistic Regression Model is decent and accurate enough with approximately 88% Accuracy on Test Set, 85 % Sensitivity, and 90 % Specificity