Summary

1. Data Understanding
    - It is most important to understand the data and the context before we start to analyze it.
    - Before starting anything, we understood each column and checked its properties.
    - Here we find that many columns' data are based on the knowledge and intuitions of the sales team and will not help in model building.
    - We also felt that columns like Total Time Spent on Website, Lead Source, Lead Origin, Specialization, Occupation would be the probable variables in the final model.

2. Data Cleaning

- We removed the inconsistencies in the data to prepare it for analysis.
- We dropped columns added by the sales team.
- Columns having 'Select' value were replaced with NullValue.
- Many columns had very high Null values (up to 40%). We imputed the columns with low percentage of null with its mean or mode. Dropped columns with very high Null values.
- Lead Source Column had lots of values, most of which total to 9.5%. We combined and replaced these values with 'Others'.
- Few columns with Outliers like TotalVisits were soft-capped so that these values don't get extra weightage in the analysis
- Columns which have a single value with a very high percentage (like Country) were dropped as they are of little use in the model.

3. Data Analysis (EDA)
    - We visualized data in EDA using plots and graphs to see the pattern of each variable
    - We observed the behavior of numerical variables against the values of target variable using boxplot
    - We observed the behavior of categorical variables against the values of target variable using barplots
    - Here we found that Lead Origin has a very high conversion
    - We could find variables that have only (almost) single values and were dropped

4. Data Preparation
    - Categorical columns were converted to numerical columns
    - Columns with Yes/No values to 1/0 and Dummy variables were used for others

5. Train-Test Split
    - The data was split into train and test data in 70:30 ratio

6. Scaling
    - Scaling of train data was done using Min Max scaling so that all numeric variables have same unit.
7. Model Building
    - We did feature selection using RFE to get 20 best variables and a model was developed only using these variables.
    - VIF score was calculated to find the variables with high score
    - Variables with high p-value or VIF scores were rejected in a number of iterations.

8. Model Evaluation
    - The model was verified on test data.
    - Confusion matrix was created, and we calculated the accuracy.
    - Sensitivity (0.83) and Specificity (0.74) of the model was created.
    - We found the Optimum cut-off point using ROC curve
    - Accuracy, sensitivity and specificity for various probability cutoffs were calculated.
    - Plotting this, we found 0.3 as the cutoff probability.

9. Prediction
    - Predictions were made using 0.3 as cutoff probability with accuracy (77%), sensitivity (82%), specificity (74.8%)
    - Precision and recall were 66.9% and 82.6% and f-1 score is 74%

10. Conclusion
    - The model built was able to predict more than 80% of leads which were converted. Also, with the help of this model we can see that the attributes like ‘Total Time Spent on Website’, ‘Lead Origin’ and ‘Last Activity’ were among the top 3 to contribute towards the probability of lead conversion.