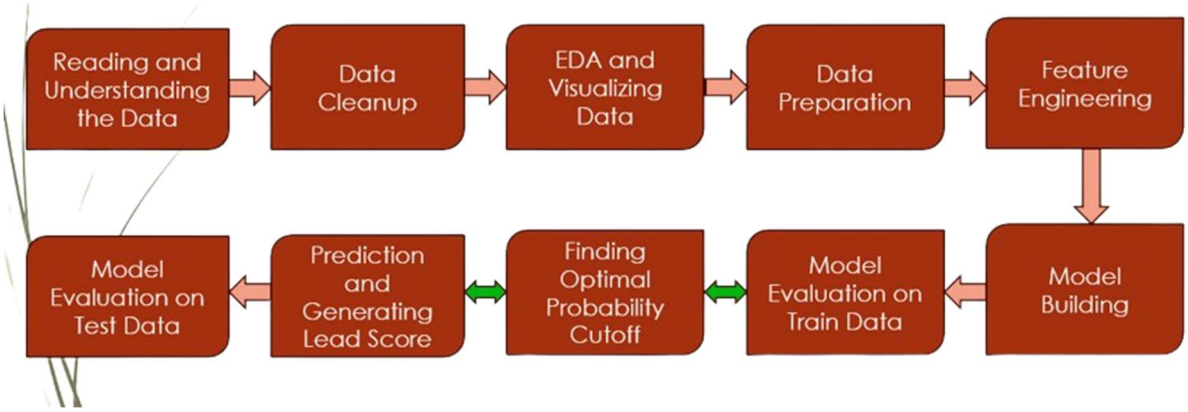
SUMMARY



1. Reading and Understanding the data:

The initial dataset in the 'leads.csv' file contains 9,240 records and includes 37 columns, consisting of 30 categorical columns and 7 numerical columns.

2. Basic Data Cleanup:

- Since 'Select' is not a valid category, it likely represents the default option in the form dropdowns. We replaced all instances of 'Select' with NaN values.
- We dropped the columns that had more than 40% of their values missing.
- Missing value treatment was done based on business understanding. For the columns Specialization, What is your current occupation, 'What matters most to you in choosing a course, Tags, City, NaN values were replaced with a new category called "Others."

3. Visualizing Data and EDA:

- Box Plot of TotalVisits, Total Time Spent on Website, Page Views Per Visit.
- Pair Plot of all Numeric variables.
- Count Plot of different categorical variables with Converted as label.
- Based on the plot we derived inferences and mentioned that in the PPT and the Jupyter Notebook.

4. Data Preparation:

- **Missing Value Imputation:** Missing values were imputed based on business understanding, using the mean or median for numerical columns, and either the mode or a new "Other" category for categorical columns.

- **Outlier Treatment:** Outliers in the numerical columns were identified by using pairplots and boxplots, and they were addressed based on business understanding.
- **Categorical Variables Encoding:**
    - Columns having binary classes replaced with 0,1.
    - Dummy variables (with drop_first=True) have been created for categorical columns having more than 2 classes.

- **Train-Test Split:** Dataset has been split into Train and Test in 70:30 ratio.
- **MinMax Scaling:** Performed MinMax Scaling (fit_transform) on train dataset and MinMax Scaling (transform) on test dataset for numerical variables only.
- **Model Building:** Creating first model with all the variables and then manual fine tune the model on the basis of high correlation, p-values and VIF.
- Created correlation heatmap and dropped variables having higher correlations.

5. Feature Engineering and Model Building:

- We used Recursive Feature Elimination (RFE) to select the top 20 features and built the initial Logistic Regression model based on those features.
- Then manually eliminated the features one by one. Total 3 models were built and after each model building p-values of all beta-coefficients and VIFs have been checked simultaneously, identified feature has been excluded in next model. Accepted p-value is lower than 0.05 and VIF < 5.
- After building each new model, we evaluated the overall accuracy and confusion matrix to compare its performance with the previous model.

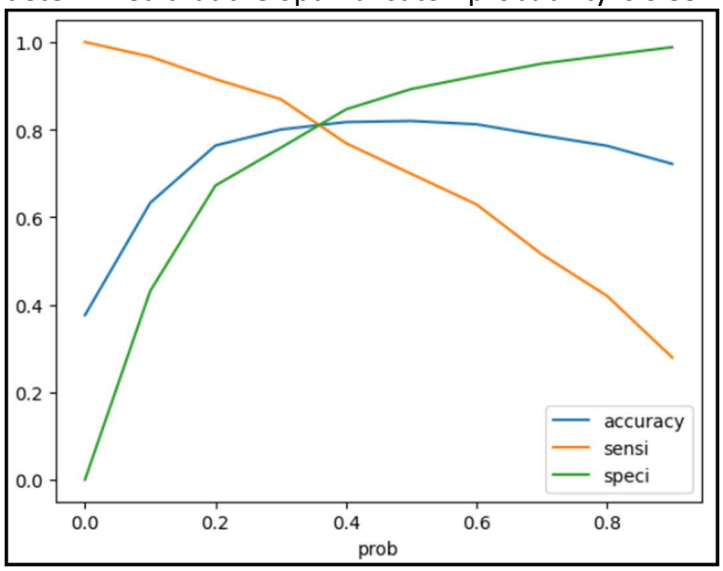
6. Prediction & Model Evaluation on Training Data (Cutoff- 0.05)

- Model 3 was used to predict probabilities on the training dataset, and a probability cutoff of 0.5 was applied to classify the target variable as 0 or 1.
- Calculated different evaluation metrics as below:

```
Model Evaluation Metrics
-----------------------------------------------
Confusion Metrics :
True Negative: 3504        False Positive: 422
False Negative: 713        True Postive: 1654
-----------------------------------------------
Model Accuracy: 0.82
Sensitivity: 0.7
Specificity: 0.89
False Positive Rate: 0.11
Precision: 0.8
Recall: 0.7
```

7. Finding Optimal Probability Cutoff & Evaluation on Train data

- We calculated specificity, sensitivity, and accuracy for the model at different probability cutoffs and plotted them in the graph below. From this graph, we determined that the optimal cutoff probability is 0.35.



8. Prediction on Test data and Generating Lead Score

- Performed MinMax Scaling on Test Data (only Transform) and kept only those columns which are present as predictor variables for final model.
- Using Model 3 we calculated the probability on Test dataset and used cutoff = 0.35 to predict the target (0,1). Created a column **Lead Score** (between 0 to 100) by doing **Conversion_prob*100.** A higher score means hot lead, lower score implies cold lead.

9. Model Evaluation on Test Data & Interpretation

- Calculated Evaluation metrics on Test data:

```
Model Evaluation Metrics
------------------------------------------------
Confusion Metrics :
True Negative: 1335        False Positive: 331
False Negative: 199        True Postive: 833
------------------------------------------------
Model Accuracy: 0.8
Sensitivity: 0.81
Specificity: 0.8
False Positive Rate: 0.2
Precision: 0.72
Recall: 0.81
```

- Top 3 variables which contribute most towards the probability of a lead getting converted:
    - Total time spent on website
    - Lead Origin_Lead Add Form
    - What is your current occupation_Working Professional