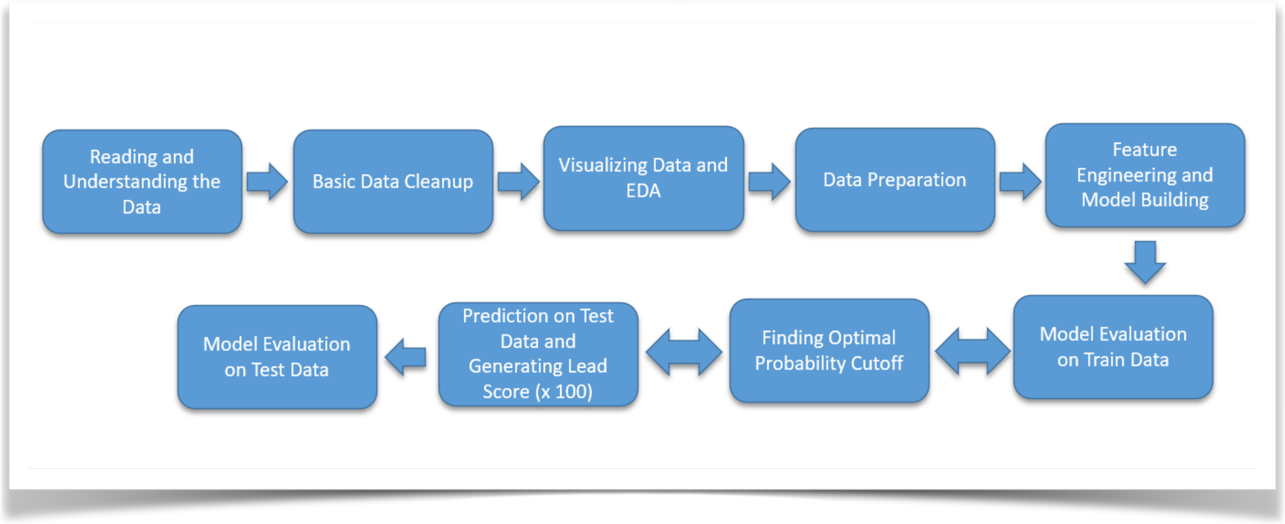
Summary Report


Reading and Understanding the Data
Basic Data Cleanup
Visualizing Data and EDA
Data Preparation
Feature Engineering and Model Building
Model Evaluation on Train Data
Finding Optimal Probability Cutoff
Prediction on Test Data and Generating Lead Score (x 100)
Model Evaluation on Test Data

1. Reading and Understanding the Data :-

Initial data with 9240 records in “leads.csv” file has 37 columns which include 30 categorical and 7 numerical columns are available.

2. Basic Data Clean up :-

- As ‘Select’ is not a valid class, we can conclude that the Select might be the default value set in the form drop downs. We replaced ‘Select’ with NaN.

- Columns having only one unique value does not have any variance, hence we dropped these columns.

- Dropped the columns having more than >3000 missing value.

- Performed missing value treatment using **Business Understanding.** For **Specialisation** and **Occupation** NaN values are replaced with a category **Not Disclosed.**

- Renamed some column names to simpler names for convenience during EDA and Model building.

3. Visualising Data and EDA :-

- Pair Plot of all Numeric variables.
- Sub Plot of all the categorical columns.
- 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 :-

- **Train-Test Split:** Dataset has been split into Train and Test in 70:30 ratio.
- **Missing Value Imputation (Statistical Imputation):** Calculated median, mode on Train dataset.
  Used that value to impute missing values in Train and Test Dataset.
  Performed Mode Imputation
  for Categorical columns and Median imputation for Numeric variables.
- **Categorical Variables Encoding:**

o Columns having binary classes replaced with 0,1
o Dummy variables (with drop_first=True) have been created for categorical columns having more than 2 classes.

- **Performed MinMax Scaling** on Train data(other than dummy).
- **Performed Variance Thresholding**, removed columns having lower variance than
  threshold=.001
- **Created correlation heatmap** and dropped variables having higher correlations.

5. Feature Engineering and Model Building :-

- RFE has been used to get top 15 features and built 1st LogisticRegression model.

- Then manually eliminated the features one by one. 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 .05 and VIF < 5.

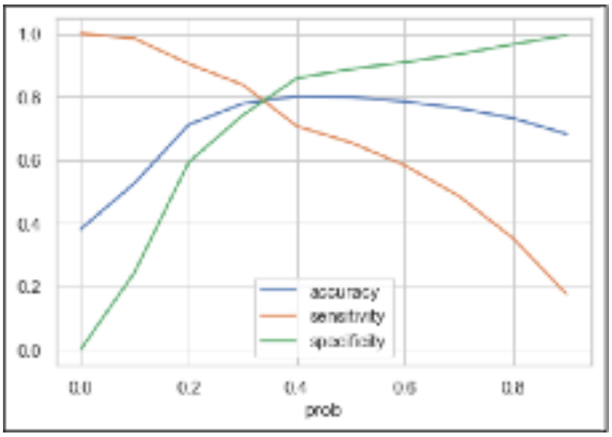
- Checked Overall model accuracy, Confusion Matrix after each new model, to understand how
the new model is performing in compared to the previous one.

6. Prediction & Model Evaluation: (on Training data with cutoff .40)

- Model has been used to predict the probability on training dataset and then used .40 as probability cut off to calculate our target (0 or 1).

7. Finding Optimal Probability cutoff & Evaluating on Train Data

• Calculated specificity, sensitivity, and accuracy for our model for different cut-off probabilities and then plotted that in below graph. From the graph we got optimal probability cutoff = .40



8. Prediction on Test Data & Generating Lead Score

- Performed MinMax Scaling on Test Data (only Transform) and kept only column which are present as predictor variables for final model.

- We calculated the probability on Test dataset and used cutoff =.42 to predict the target (0,1). Created a column **Lead Score** (between 0 to 100) by doing **prob*100.** A higher score means hot lead, lower score implies cold lead.