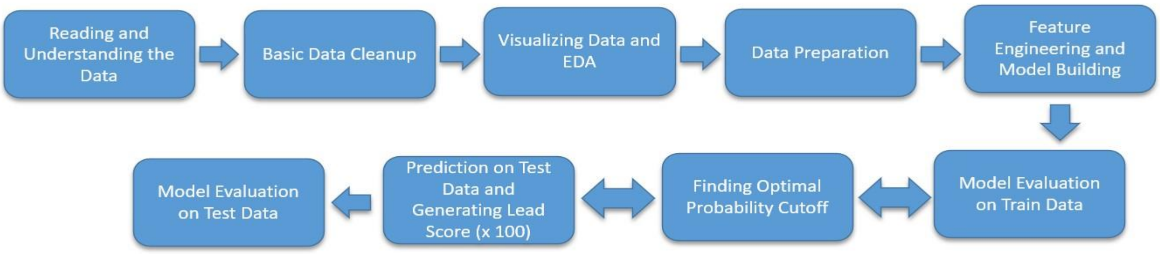
Summary Report


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 dropdowns. 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 35% missing value.
- Created new buckets/bins for the categorical variables having very high numbers of classes with few datapoints.
- Performed missing value treatment using **Business Understanding.** For **Specialization** and **Occupation** NaN values are replaced with a new category **Not Disclosed.**
- Renamed some column names to simpler names for convenience during EDA and Model building.

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:

- **Outlier Treatment:** By observing box plot and calculating different percentile values, identified

  2.8% of total data (< 5%) as outliers and removed those rows.
- **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:**

  ○ 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.

- **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 16 features and built 1st Logistic Regression model.
- Then manually eliminated the features one by one. 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 .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 cut-off 0.5)

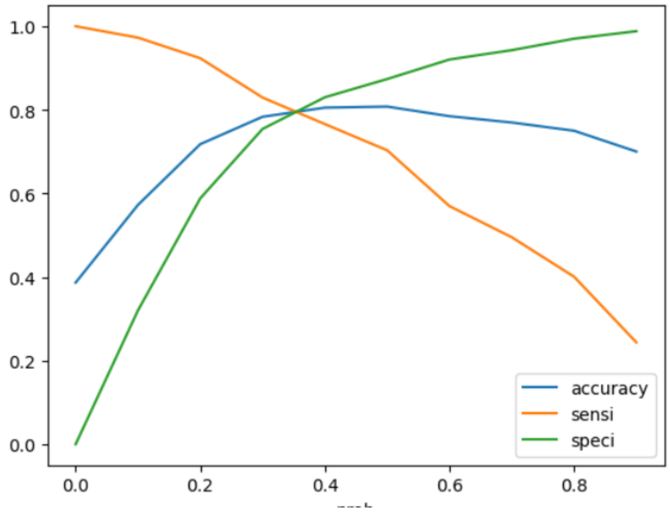
- Model has been used to predict the probability on training dataset and then used 0.5 as probability cut off to calculate our target (0 or 1).
- Calculated different evaluation metrics as below:

```
Overall model accuracy: 0.8077468115257744
Sensitivity / Recall: 0.7031758957654723
Specificity: 0.8736842105263158

Confusion matrix:
      True negative: 3403        False positive: 492
      False negative:  729       True positive: 1727
```

7. Finding Optimal Probability cut-off & 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 cut-off = 0.35.



```
Overall model accuracy: 0.7967249252086286
Sensitivity / Recall: 0.7992671009771987
```

```
Specificity: 0.7951219512195122
```

8. Prediction on Test Data & Generating Lead Score

- Performed MinMax Scaling on Test Data (only Transform) and kept only hose column which are present as predictor variables for final model.
- Using Model, we calculated the probability on Test dataset and used cut-off =0.35 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.

9. Model Evaluation on Test data & Interpretation

Calculated evaluation metrics on test data.

```
Confusion matrix:
      True negative: 1465       False positive: 279
      False negative: 242       True positive: 737

Overall model accuracy: 0.8077468811525744

Sensitivity / Recall: 0.7528089887640449

Precision: 0.7253937007874016
```