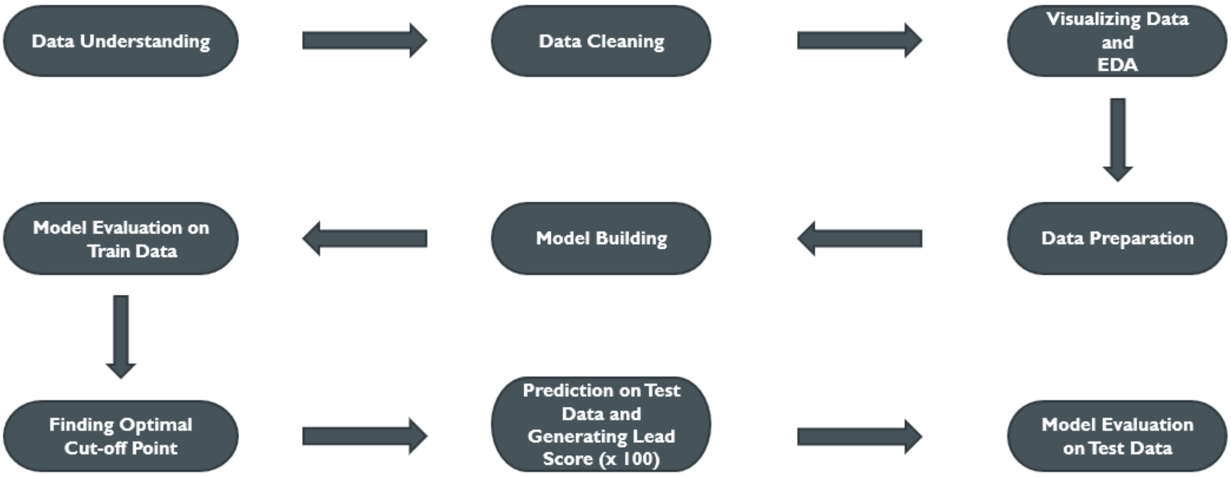
Summary Report

Arun – Prabanjan - Srushti


Data Understanding
Data Cleaning
Visualizing Data and EDA
Data Preparation
Model Building
Model Evaluation on Train Data
Finding Optimal Cut-off Point
Prediction on Test Data and Generating Lead Score (x 100)
Model Evaluation on Test Data

1. **Reading and Understanding the Data:**
   The initial data from the "leads.csv" file consists of 9240 records and comprises 37 columns, encompassing 30 categorical and 7 numerical variables.

2. **Basic Data Clean up :** Prior to the Exploratory Data Analysis phase, we conducted basic data cleanup tasks, which included the deletion of columns with blank values exceeding 40%, imputation of missing values with mode values, and renaming of values where necessary.

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:** The team address outliers by replacing values below or equal to the 5th percentile with the 5th percentile value itself, and values above or equal to the 95th percentile with the 95th percentile value.

- **Train-test Split:** The dataset has been divided into Train and Test sets in a 70:30 ratio. The Train dataset is utilized for training the model, while the Test dataset is used for evaluating the model's performance.
- **Missing Value Imputation:** Nominal categorical columns were imputed using the mode.
- **Categorical Variables Encoding:** Categorical columns with 'Yes' and 'No' values have been encoded as 1 and 0, respectively. For columns with more than two categories, dummy variables

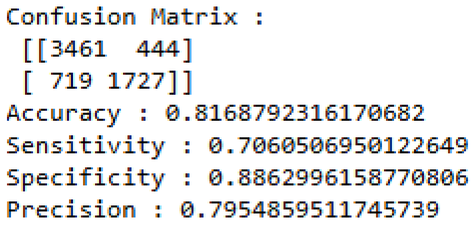
were created, and the original column along with the first dummy variable for each column were dropped from the data-frame.

- **Scaling Features:** Selected numerical variables were scaled using the Standard Scaler.

5. **Feature Engineering and Model Building** • Team initially built a Logistic Regression model using sklearn's linear model with these 20 features.
• Team then manually fine-tuned the model to ensure statistical significance by checking p-values (accepted if less than 0.05) and removed multicollinearity by examining Variance Inflation Factors (VIFs, accepted if less than 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.

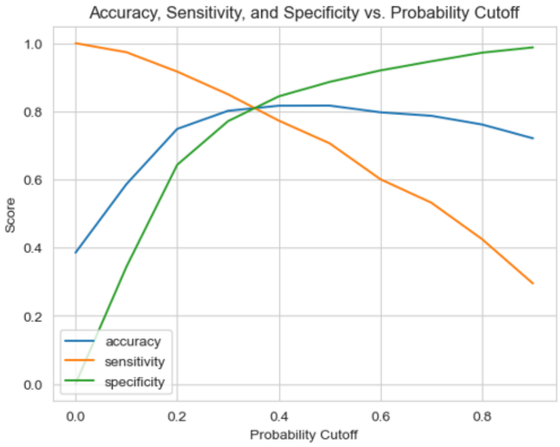
6. **Prediction & Model Evaluation: (on Training data with cutoff 0.5)**

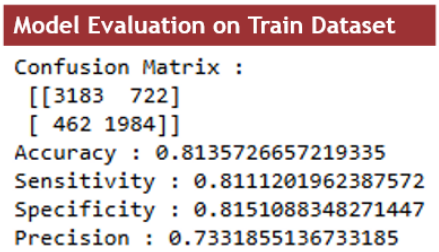
Model 9 has been used to predict the probability on training dataset and then used .5 as probability cut off to calculate our target (0 or 1).



```
Confusion Matrix :
 [[3461  444]
 [ 719 1727]]
Accuracy : 0.8168792316170682
Sensitivity : 0.7060506950122649
Specificity : 0.8862996158770806
Precision : 0.7954859511745739
```

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 = .35.



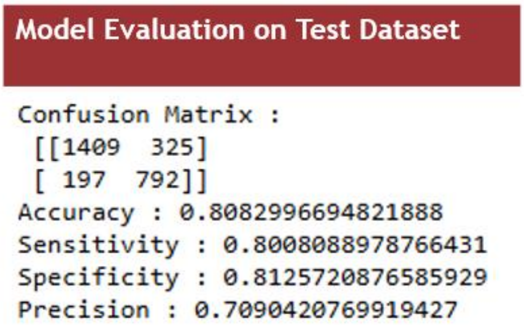


Model Evaluation on Train Dataset

```
Confusion Matrix :
 [[3183  722]
 [ 462 1984]]
Accuracy : 0.8135726657219335
Sensitivity : 0.8111201962387572
Specificity : 0.8151088348271447
Precision : 0.7331855136733185
```

8. **Prediction on Test Data & Generating Lead Score**

• Utilizing Model 9, we computed probabilities on the Test dataset and applied a cutoff of 0.35 to predict the target (0 or 1). Subsequently, we generated a Lead Score column ranging from 0 to 100 by multiplying the probabilities by 100. A higher score indicates a hotter lead, while a lower score suggests a colder lead.



Model Evaluation on Test Dataset

```
Confusion Matrix :
 [[1409  325]
 [ 197  792]]
Accuracy : 0.8082996694821888
Sensitivity : 0.8008088978766431
Specificity : 0.8125720876585929
Precision : 0.7090420769919427
```

9. **Top three variables** in the model which contribute most towards the probability of a lead getting converted:

- **Total Time Spent on Website:** Leads who spend more time browsing the website are more likely to be interested in the product or service and therefore have a higher conversion probability.
- **Lead Source:** Leads acquired through referrals tend to convert at a higher rate. Referrals often come with built-in trust and positive recommendations, indicating a stronger interest.
- **Current Occupation (in some cases):** Depending on your product or service, a lead's current occupation can be a significant factor. For instance, if you offer professional development courses, leads who are students might be less likely to convert compared to working professionals.