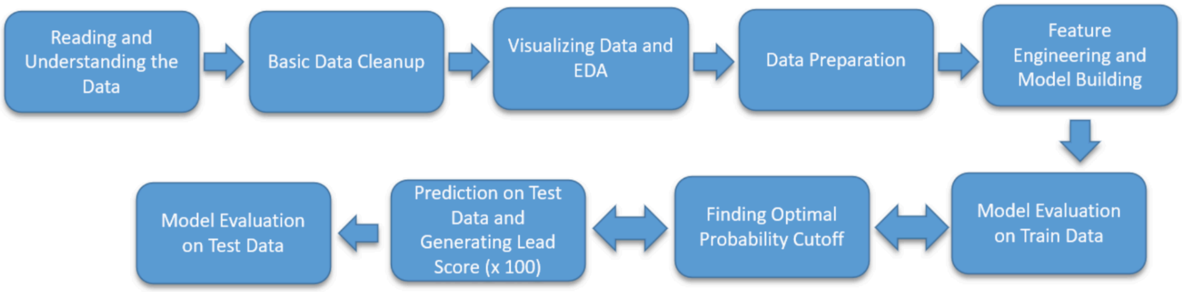
Summary Report

Overall Approach:


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. Data Overview

- The dataset consists of **9,240 records** in the `Leads.csv` file, containing **37 columns** (30 categorical and 7 numerical).

2. Data Cleaning

- Replaced 'Select' values with **NaN**, as it was identified as a default selection.
- Dropped columns with **only one unique value** due to lack of variance.
- Removed columns with **more than 40% missing values**.
- Grouped categorical variables with high cardinality into **bins**.
- Missing values in **Specialization** and **Occupation** were replaced with a new category, **Not Disclosed**.
- Simplified column names for better readability during **EDA and modeling**.

3. Exploratory Data Analysis (EDA)

- **Box Plots** were created for `TotalVisits`, `Total Time Spent on Website`, and `Page Views Per Visit`.
- **Pair Plot** was generated for all numerical variables.
- **Count Plots** were used to analyze categorical variables with respect to the conversion rates.
- Insights from these visualizations were documented in the **PPT and Jupyter Notebook**.

4. Data Preparation

- **Outlier Removal**: Identified and removed **2.8%** of total records as outliers.

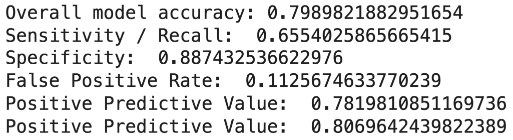
- **Train-Test Split**: Data split into **70% training** and **30% testing**.
- **Missing Value Imputation**:
    - **Median imputation** for numerical variables.
    - **Mode imputation** for categorical variables.
- **Categorical Encoding**:
    - Binary columns encoded as **0/1**.
    - Dummy variables created for multi-class categorical columns (`drop_first=True`).
- **Feature Scaling**: MinMax scaling applied to train data.
- **Feature Selection**:
    - **Variance Thresholding**: Removed features with variance **< 0.001**.
    - **Correlation Analysis**: Dropped highly correlated features.

5. Feature Engineering & Model Building

- **Recursive Feature Elimination (RFE)** identified the top **16 features**.
- **Logistic Regression models** were iteratively refined:
    - A total of **7 models** were built.
    - Features were manually eliminated based on **p-values (< 0.05)** and **VIF (< 5)**.
    - Accuracy and confusion matrix were reviewed after each model iteration.

6. Model Prediction & Evaluation (Training Data)

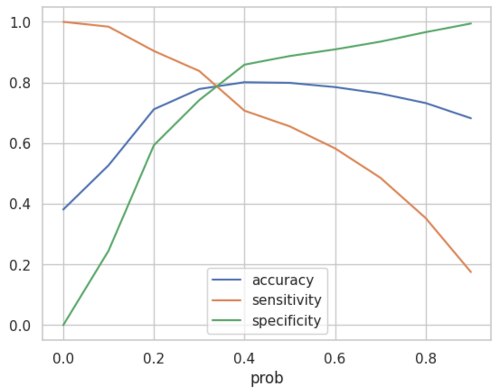
- **Model 7** was selected for predictions on training data using a **cutoff probability of 0.5**.
- Evaluation metrics were calculated.



```
Overall model accuracy: 0.7989821882951654
Sensitivity / Recall:  0.6554025865665415
Specificity:  0.887432536622976
False Positive Rate:  0.1125674633770239
Positive Predictive Value:  0.7819810851169736
Positive Predictive Value:  0.8069642439822389
```

7. Determining Optimal Probability Cutoff

- Sensitivity, specificity, and accuracy were plotted across different probability thresholds.



- The optimal probability cutoff was determined as **0.32**.

8. Prediction on Test Data & Lead Scoring

- MinMax scaling was applied to test data using the transformation from train data.
- Predictions were made using **Model 7** with a **cutoff of 0.32**.
- A **Lead Score (0-100)** was assigned based on probability (`probability * 100`), where higher scores indicate **hot leads** and lower scores indicate **cold leads**.

9. Model Evaluation on Test Data

- Performance metrics were computed on the test dataset to assess model effectiveness.

```
Model Evaluation Metrics on Test dataset
##################################################
Confusion Matrix:
True Negative: 1258      False Positive: 402
False Negative: 203      True Positive: 832

Overall model accuracy: 0.7755102040816326
Sensitivity / Recall:  0.8038647342995169
Specificity:  0.7578313253012048
False Positive Rate:  0.2421686746987952
Positive Predictive Value:  0.6742301458670988
Positive Predictive Value:  0.8610540725530459
```

Key Findings

The top 3 variables that contribute most towards the probability of a lead getting converted are:

- **Total Time Spent on Website**: Leads who spend more time on the website are more likely to convert, indicating strong engagement with the website content.
- **What is your current occupation (Working Professional)**: Leads who identify as working professionals have a higher conversion rate, suggesting they are more likely to seek professional development opportunities.
- **Lead origin (Other)**: Leads from sources categorized as 'Other' have a very high conversion rate, which may include effective lead acquisition methods like referrals.