Name-Siddharth Singhal

Overall Approach

```
[Reading and Understanding the Data] → [Basic Data Cleanup] → [Visualizing Data and EDA] → [Data Preparation] → [Feature Engineering and Model Building]

[Model Evaluation on Test Data] ← [Prediction on Test Data and Generating Lead Score (x 100)] ↔ [Finding Optimal Probability Cutoff] ↔ [Model Evaluation on Train Data]
```
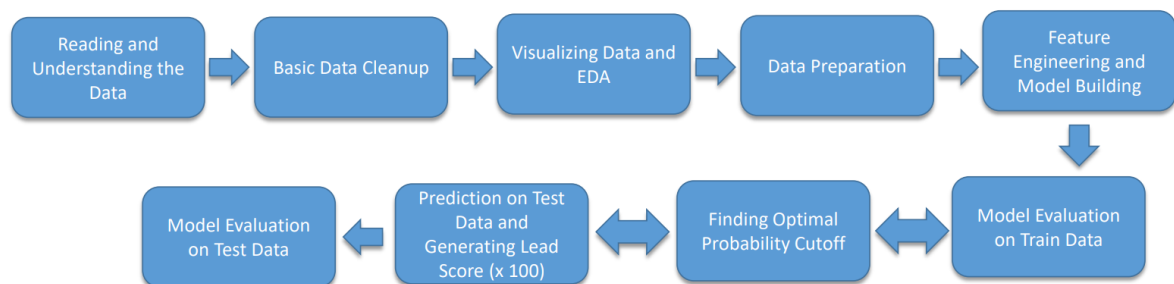
3

Summary

The objective is to develop a model for Company X Education to identify strategies that can effectively convert potential customers. The process involves comprehensive data analysis and model validation to accurately target the right audience and improve conversion rates. The following steps were undertaken:
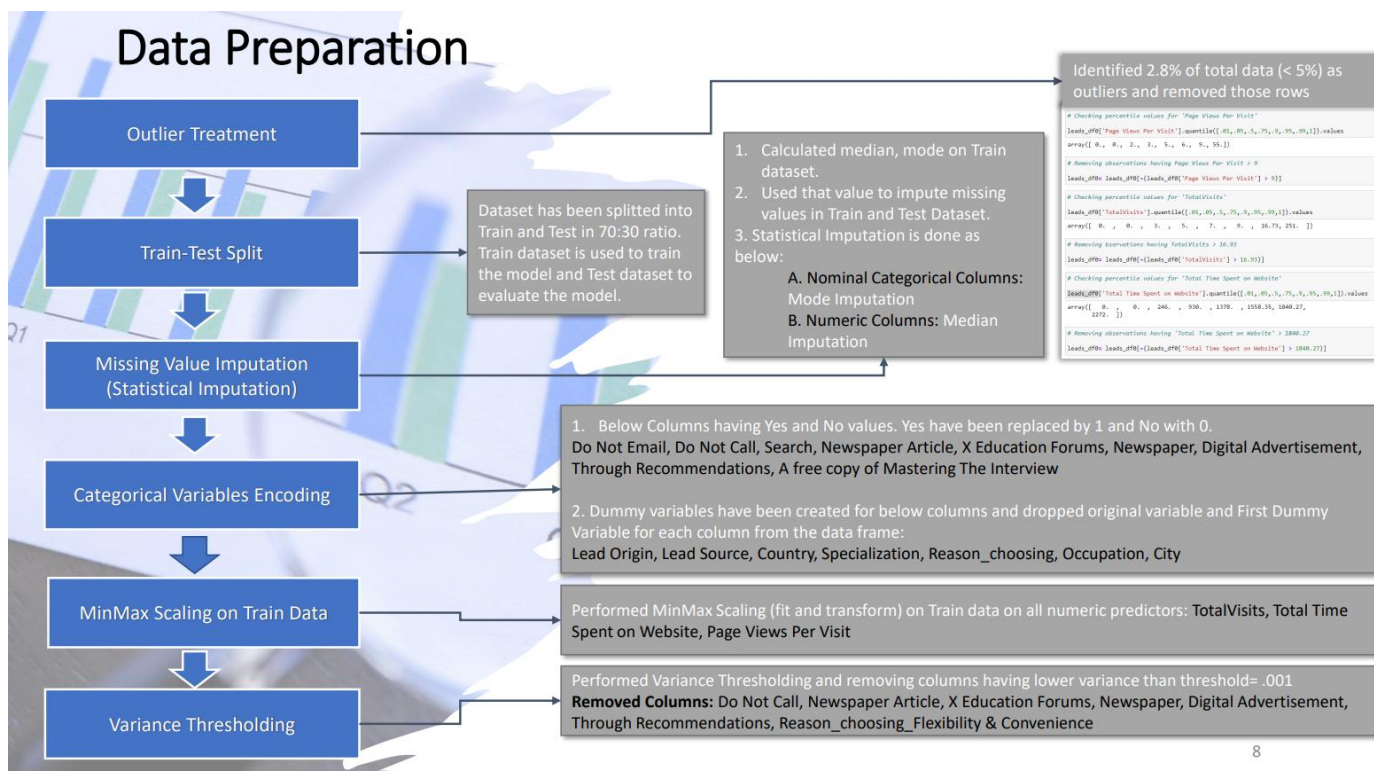
1. Exploratory Data Analysis (EDA):

- **Null Values**: We performed a quick assessment of missing data, removing columns where more than 45% of values were missing.
- **Handling Missing Data**: Instead of discarding rows with missing values—which would result in significant data loss—we substituted 'not provided' for missing entries.
- **Country Data**: Since 'India' was the most frequent value among non-missing entries, we replaced 'not provided' values with 'India'. However, given that nearly 97% of the data was from India, the column was ultimately dropped.
- **Data Preparation**: We addressed numerical variables, managed outliers, and created dummy variables where necessary.

2. Train-Test Split & Scaling:

- **Data Split**: The data was divided into training and testing sets, with a 70% to 30% split, respectively.
- **Scaling**: Min-max scaling was applied to the variables ['TotalVisits', 'Page Views Per Visit', 'Total Time Spent on Website'].

3. Model Development:

- **Feature Selection**: Recursive Feature Elimination (RFE) was employed to select the top 15 relevant features.
- **Variable Refinement**: After RFE, further manual elimination was performed based on Variance Inflation Factor (VIF) values and p-values.
- **Model Accuracy**: A confusion matrix was used to assess the model's accuracy, which was found to be 92.78%.

Data Preparation

Outlier Treatment → Identified 2.8% of total data (< 5%) as outliers and removed those rows

Train-Test Split → Dataset has been splitted into Train and Test in 70:30 ratio. Train dataset is used to train the model and Test dataset to evaluate the model.

Missing Value Imputation (Statistical Imputation) →
1. Calculated median, mode on Train dataset.
2. Used that value to impute missing values in Train and Test Dataset.
3. Statistical Imputation is done as below:
   - A. Nominal Categorical Columns: Mode Imputation
   - B. Numeric Columns: Median Imputation

Categorical Variables Encoding →
1. Below Columns having Yes and No values. Yes have been replaced by 1 and No with 0.
Do Not Email, Do Not Call, Search, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, A free copy of Mastering The Interview

2. Dummy variables have been created for below columns and dropped original variable and First Dummy Variable for each column from the data frame:
Lead Origin, Lead Source, Country, Specialization, Reason_choosing, Occupation, City

MinMax Scaling on Train Data → Performed MinMax Scaling (fit and transform) on Train data on all numeric predictors: TotalVisits, Total Time Spent on Website, Page Views Per Visit

Variance Thresholding → Performed Variance Thresholding and removing columns having lower variance than threshold= .001
Removed Columns: Do Not Call, Newspaper Article, X Education Forums, Newspaper, Digital Advertisement, Through Recommendations, Reason_choosing_Flexibility & Convenience

4. Model Evaluation:

- **Sensitivity – Specificity Evaluation**:
  - **Training Data**:
    - Using the ROC curve, an optimal **cutoff value of 0.3** was identified, resulting in:
      - Accuracy: 92.29%
      - Sensitivity: 91.70%
      - Specificity: 92.66%
      - Precision:88.43%
      - Recall:91.66%

o **Test Data**:
  ▪ The model yielded:
    ▪ Accuracy: 92.63%
    ▪ Sensitivity: 91.49%
    ▪ Specificity: 93.32%
    ▪ Precision:89.19%
    ▪ Recall:91.49%

Key Factors Influencing Conversion:

- **Lead Source**:
  o Total Visits
  o Total Time Spent on Website
- **Lead Origin**:
  o Lead Add Form
- **Lead Source**:
  o Direct Traffic
  o Google
  o Welingak Website
  o Organic Search
  o Referral Sites
- **Last Activity**:
  o Do Not Email (Yes)
  o Last Activity (Email Bounced)
  o Olark Chat Conversation

The model demonstrates strong predictive ability regarding conversion rates, providing Company X with a robust tool for making informed decisions.

Recommendations:

- Diversify Market Focus: Recognize that most responses are concentrated in Mumbai, India. To tap into potential customers, broaden your market focus beyond Mumbai.

- Leverage Mumbai Market: Given the strong inclination towards Mumbai, introduce special offers or discounts to boost sales in this region.

- Target Corporate Sectors: Noting that management professionals frequently search for courses, tailor your marketing strategies towards corporate clients to enhance conversion rates.

- Engage Working Professionals: Focus on working professionals to drive higher conversions. Also, consider the 50% conversion potential among unemployed individuals seeking to upskill, which could be a valuable opportunity.

- Optimize Lead Sources: Acknowledge that most leads are generated through Google and direct traffic, while the least come from live chat. Adjust strategies to enhance underperforming channels.

- Capitalize on High Conversion Websites: The Welingak website has the highest conversion rate. Increase efforts to drive traffic to this site.

- Boost Lead Conversion: Improve lead conversion by maximizing leads from referrals and the Welingak website.

- Focus on Key Traffic Channels: Concentrate on Olark chat, organic search, direct traffic, and Google leads to further increase lead conversion rates.

- Analyze Activity Correlations: Be aware of the high correlation between "Last Notable Activity_SMS Sent" and "Last Activity_SMS Sent," which could provide insights into customer behavior.

- Utilize Website Engagement Data: The median time spent on the website is 244 minutes. Use this information to refine engagement strategies.

- Maximize API and Landing Page Submissions: API and landing page submissions generate a significant number of leads and conversions. Focus on these channels to drive more leads.

- Enhance Lead Add Form Utilization: The Lead Add Form shows a very high conversion rate, although the volume of leads is not high. Increasing the number of leads through this form could substantially boost overall conversion rates and growth.