Summary Report for X Education Lead Conversion Case Study

Authors: Purnima NS, Priya Raajendran, Rajesh Iyer

---

1. **Problem Statement**

   X Education, an online education provider, faces a challenge in improving its lead conversion rate, which currently stands at around 30%. The goal is to identify "Hot Leads" that are most likely to convert into paying customers. The target is to enhance the lead conversion rate to approximately 80% by building a model that assigns a lead score to each lead. This model will help the sales team focus their efforts on leads with higher potential.

---

2. **Data Cleaning Summary**
   - **Initial Data Examination:**
     - The dataset consists of various attributes including 'Lead Source', 'Total Time Spent on Website', 'Total Visits', and 'Last Activity', among others.
     - The target variable is 'Converted', indicating whether a lead was converted (1) or not (0).
   - **Removing Unnecessary Columns:**
     - Dropped index columns 'Prospect ID' and 'Lead Number' as they were not useful for analysis.
   - **Handling Null Values:**
     - Replaced 'Select' with NaN and then handled NaN values:
       - Dropped columns with 40% or more null values.
       - Imputed missing values for categorical variables based on their distribution and relevance:
         - Replaced 'City' nulls with 'Unknown'.
         - Replaced 'Specialization' nulls with 'None'.
         - Replaced 'Tags' nulls with 'None' and combined other tags into 'Other_Tags'.
         - Replaced 'What is your current occupation' nulls with 'Unemployed'.
         - Replaced 'What matters most to you in choosing a course' nulls with 'Better Career Prospects'.
       - Dropped columns with minimal information after imputation, such as 'Country'.
   - **Dropping Irrelevant Columns:**
     - Columns with minimal or redundant information were dropped, including 'Do Not Call' and other minor columns.

---

3. **Exploratory Data Analysis (EDA)**

Here's a summary of the key statistics for each feature, split by the conversion status (Converted = 0 or 1):

1. Total Time Spent on Website

- Converted = 0:
  - Median: ~550
  - 25th Percentile (Q1): ~200
  - 75th Percentile (Q3): ~850
- Converted = 1:
  - Median: ~1250
  - 25th Percentile (Q1): ~850
  - 75th Percentile (Q3): ~1550

2. Total Visits

- Converted = 0:
  - Median: ~3
  - 25th Percentile (Q1): ~2
  - 75th Percentile (Q3): ~5
- Converted = 1:
  - Median: ~4
  - 25th Percentile (Q1): ~3
  - 75th Percentile (Q3): ~6

3. Page Views Per Visit

- Converted = 0:
  - Median: ~3
  - 25th Percentile (Q1): ~2
  - 75th Percentile (Q3): ~4
- Converted = 1:
  - Median: ~4
  - 25th Percentile (Q1): ~3
  - 75th Percentile (Q3): ~5

This data suggests that users who convert tend to spend more time on the website, visit more often, and view more pages per visit compared to users who do not convert. These features could be useful in predicting conversion likelihood.

---

4. **Dummy Variables for Dependent Variables:**

   Created dummy variables for categorical features to be used in the logistic regression model. This includes converting categorical data such as 'Lead Source', 'Specialization', 'Tags', and others into numerical format to facilitate model training.

---

5. **Training and Test Data Split**
   - **Splitting the Data:**

    - The dataset was split into training and test sets using an 80-20 split.
    - Standard scaling was applied to normalize the features.

---

6. **Findings & Conclusion**
    - **Model Performance:**
        - A logistic regression model was trained to predict lead conversion.
        - The model's performance was evaluated using metrics such as accuracy, precision, recall, and ROC AUC.
    - **Significant Findings:**
        - Key features influencing lead conversion were identified through feature importance analysis. the top 5 features influencing lead conversion are:
            1. **Total Time Spent on Website**: This indicates how engaged a lead is with the content, with more time spent often correlating with higher conversion likelihood.
            2. **Lead Source**: The origin of the lead (e.g., Google, Social Media) affects the quality of the lead and their conversion probability.
            3. **Last Activity**: The most recent interaction or activity of the lead (e.g., Email Opened, Page Visited) helps gauge their interest and likelihood of conversion.
            4. **Specialisation**: The area of specialisation (e.g., Management Specializations) can impact the conversion rate based on how relevant the courses are to the lead's career goals.
            5. **Tags**: Tags associated with the lead (e.g., Interested in other courses, Will revert after reading the email) provide additional context on the lead's potential interest and engagement level.
        - Conversion rates varied significantly across different features, emphasising the need for targeted strategies.
    - **Conclusion:**
        - By focusing on leads with high predicted scores, X Education can improve its lead conversion rate.
        - The model provides actionable insights that can help prioritise high-potential leads, potentially increasing the conversion rate towards the target of 80%.
    - **Next Steps:**
        - Fine-tune the model based on additional data or feedback.
        - Implement the model in the sales process and monitor performance.
        - Continuously refine the model with new data to adapt to changing lead characteristics.