**Email Marketing Campaign Optimization: Assignment Report**

**1. Objective**

The goal of this project is to optimize email targeting for an e-commerce platform by identifying which users are most likely to engage with email campaigns, specifically those who are most likely to click on links within the email.

**2. Data Overview**

The dataset consists of three primary tables originally provided as separate CSV files:

* email\_table.csv containing email content, user details, and send time
* email\_opened\_table.csv identifying which emails were opened
* link\_clicked\_table.csv identifying which emails had clicks on links

These were merged in **Excel** using the IF and MATCH functions to match email IDs and bring the open and click labels into a single consolidated file for analysis.

Final combined dataset columns:

* email\_id: unique identifier
* email\_text: short or long
* email\_version: personalized or generic
* hour, weekday, user\_country, user\_past\_purchases
* email opened (Y-1/N-0)
* link clicked (Y-1/N-0)

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

Key EDA steps included:

* Checking distribution of open and click responses
* Analysing click and open rates across:
  + Email content types
  + Send hours and weekdays
  + User past purchase history and geography
* Visualized segment-level CTR differences

EDA findings informed the feature engineering and modeling strategy.

**4. Methodology**

**Step 1: Data Preprocessing & Feature Engineering**

* Created working\_hours binary variable
* Created is\_weekend flag
* Added interaction term: personalized\_short
* One-hot encoded all categorical variables
* Checked for missing values and confirmed no major nulls

**Step 2: Model Building**

The initial modelling was done using **XGBoost** to predict P(click) directly. However, due to **severe class imbalance**, this model failed to predict any positive class (clicks).

To address this, we applied **SMOTE (Synthetic Minority Over-sampling Technique)** to balance the training data and then performed **threshold tuning** to optimize for F1-score. This helped drastically improve recall and the F1-score of the minority class.

After validating the improvement, we shifted to a **two-stage conditional model**:

* **Model 1**: Predicts probability of opening the email P(open) using LightGBM
* **Model 2**: Predicts probability of clicking on the link given the email was opened P(click | open)
* **Final Probability**: P(click) = P(open) \* P(click | open)

**Step 3: Threshold Tuning**

* Tuned the threshold for binary classification using F1-score
* Optimal threshold selected to maximize balance between precision and recall

**Step 4: Evaluation**

* Used classification report, ROC-AUC, and F1 score to evaluate prediction quality
* Evaluated CTR uplift by comparing model-based targeting to baseline

**5. Results**

* **Email Open Rate:** ~10.35%
* **Click-Through Rate (CTR):** ~2.12% (according to the data)

**Post-Modelling Evaluation**

* **Recall (Click class):** improved from 1% to 59%
* **F1-Score (Click class):** improved from 0.02 to 0.36 after threshold tuning

**Mean Predicted P(click):**

# Top N user targeting performance

* Top 5%: ~**12.73%** of the users are expected to click on the link
* Top 10%: ~**10.20%** of the users are expected to click on the link
* Top 20%: ~**7.95%** of the users are expected to click on the link
* Baseline: ~**2.12%** of the users clicked on the link as per the given data

This suggests that if we target the top 5% of the customers and study the patterns corresponding to their email id row in the given data, we can achieve up to **6x CTR uplift**.

The useful patterns are listed in the section below

**6. Key Insights from Segmentation**

* **Email Version**: Personalized + short performed best
* **Timing**: Weekday mornings showed higher open & click rates
* **User Behaviour**: Past purchasers more likely to engage
* **Geography**: US and UK users were most responsive

**7. Recommendations**

* Replace random campaign strategy with model-based targeting
* Focus on top 5% predicted users for highest ROI. Top 10% can also be taken as the expected CTR between the two differs by 2%.

**8. Conclusion**

A robust machine learning pipeline was built to optimize CTR using conditional modeling. The approach demonstrated strong potential for campaign uplift and forms the basis for a scalable, intelligent email marketing strategy.