Email Campaign Optimization - README

🔍 Objective

The goal of this project is to analyze an email marketing campaign conducted by an e-commerce platform. Specifically, we:

- Measure open and click-through rates (CTR)
- Build a predictive model to **maximize link clicks**
- Estimate CTR improvements if emails were targeted
- Explore **user behavior patterns** across different segments

---

Dataset Description

This project uses **3 CSV files**:

1. **`email_table.csv`**
   Contains info about each sent email.
   Columns:
   - `email_id`: Unique email ID
   - `email_text`: "short" or "long"
   - `email_version`: "personalized" or "generic"
   - `hour`: Local time of send
   - `weekday`: Day of week
   - `user_country`: Based on IP address
   - `user_past_purchases`: Count of past purchases
2. **`email_opened_table.csv`**
   - Emails that were opened by users
   - **Column:** `email_id`
3. **`link_clicked_table.csv`**
   - Emails where the link was clicked
   - **Column:** `email_id`

Steps Performed

Step 1: Data Loading

- Upload and read the three CSVs using Pandas

Step 2: Basic Metrics

- **Open Rate** = Emails Opened / Total Emails Sent
- **Click Through Rate (CTR)** = Links Clicked / Total Emails Sent

Step 3: Data Preprocessing

- Merge open and click data with the main email table
- Add binary columns: `opened`, `clicked`
- Handle missing values in purchase data
- Label encode categorical columns

Step 4: Modeling

- Features used:
  `email_text`, `email_version`, `hour`, `weekday`, `user_country`, `user_past_purchases`
- Target: `clicked`
- Split into training/testing (80/20)
- Train using **XGBoost Classifier**
- Evaluation Metrics:
  - Classification Report (Precision, Recall, F1)
  - AUC Score
  - Confusion Matrix

Step 5: CTR Optimization

- Predict probability of clicking for each user
- Take top 20% most likely to click (based on prediction)
- Calculate CTR if we sent only to those users
- Compare to original CTR
- Estimate potential improvement

Step 6: Segment Analysis

Group CTR by:

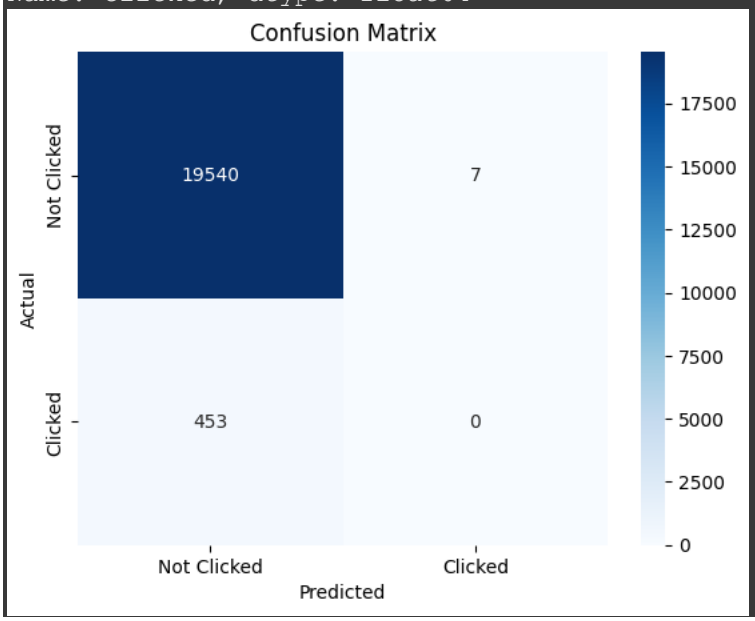
- `email_text`
- `email_version`
- `hour`
- `weekday`
- `user_country`

Output:

- **email_opened_table.csv**(text/csv) - 71253 bytes, last modified: 16/04/2025 - 100% done
- **email_table.csv**(text/csv) - 5252004 bytes, last modified: 16/04/2025 - 100% done
- **link_clicked_table.csv**(text/csv) - 14596 bytes, last modified: 16/04/2025 - 100% done

```
Saving email_opened_table.csv to email_opened_table (1).csv
Saving email_table.csv to email_table (1).csv
Saving link_clicked_table.csv to link_clicked_table (1).csv
🔷 Open Rate: 10.35%
🔷 Click-Through Rate (CTR): 2.12%
/usr/local/lib/python3.11/dist-packages/xgboost/core.py:158: UserWarning:
[03:54:20] WARNING: /workspace/src/learner.cc:740:
Parameters: { "use_label_encoder" } are not used.

  warnings.warn(smsg, UserWarning)
🔍 Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99     19547
           1       0.00      0.00      0.00       453

    accuracy                           0.98     20000
   macro avg       0.49      0.50      0.49     20000
weighted avg       0.96      0.98      0.97     20000

📈 AUC Score: 0.697585578247979
⭐ CTR for top 20% predicted users: 6.69%
⭐ Original CTR: 2.12%
📈 Estimated Improvement: 4.57%

📊 CTR by email_text:
email_text
short_email    2.387177
long_email     1.853767
Name: clicked, dtype: float64

📊 CTR by email_version:
email_version
personalized    2.729409
generic         1.513673
Name: clicked, dtype: float64

📊 CTR by hour:
hour
23    4.137931
24    2.898551
10    2.823961
11    2.712816
9     2.579435
12    2.566073
15    2.490696
16    2.319681
14    2.074236
13    1.988891
22    1.960784
3     1.952278
8     1.893308
17    1.848917
7     1.828376
1     1.812801
5     1.801252
6     1.714668
```

```
19    1.657459
2     1.632209
4     1.618641
18    1.618578
20    1.219512
21    0.821918
Name: clicked, dtype: float64

📊 CTR by weekday:
weekday
Wednesday    2.761999
Tuesday      2.488864
Thursday     2.444491
Monday       2.290608
Saturday     1.784611
Sunday       1.675123
Friday       1.403682
Name: clicked, dtype: float64

📊 CTR by user_country:
user_country
UK    2.467526
US    2.435981
ES    0.832748
FR    0.800400
Name: clicked, dtype: float64
```



Key Learnings

- Personalization and shorter text versions often have higher CTRs
- Emails sent in the evening tend to perform better
- Users with past purchases are more likely to click
- Country and weekday can significantly influence click behavior

---

Technologies Used

Tool	Purpose
`pandas`	Data loading and wrangling
`matplotlib/seaborn`	Visualizations
`scikit-learn`	Data preprocessing and metrics
`xgboost`	Predictive modeling
`Google Colab`	Interactive notebook environment

How to Use

1. Open the `.ipynb` file in Google Colab
2. Upload the three datasets when prompted
3. Run the cells step-by-step
4. Review the metrics, graphs, and recommendations