**1. Objective**

The goal of this project was to analyze and optimize an email marketing campaign for an e-commerce company. The key objectives were:

* Analyze how users responded to a recent email campaign.
* Build a machine learning model to predict user engagement (clicks).
* Simulate improvements to click-through rate (CTR) using the model.
* Explore performance across different user segments.

**2. Data Overview**

Three datasets were provided:

* **email\_table.csv**: Metadata about emails sent.
* **email\_opened\_table.csv**: Records of emails that were opened.
* **link\_clicked\_table.csv**: Records of emails where the internal link was clicked.

*Note:*

From a total of 100000 emails sent to people, 10345 people opened it and from that 2119 people clicked on the link in that email.

Total emails sent = 10,0000

Emails opened = 10345

Links clicked = 2119

**Key features used:**

* email\_text (short/long)
* email\_version (generic/personalized)
* hour
* weekday
* user\_country
* user\_past\_purchases

Additional features were engineered:

* time\_of\_day (based on hour)
* purchase\_bucket (based on number of purchases)

**3. Campaign Performance Metrics**

* **Open Rate**: 10.35%
* **Click-Through Rate (CTR)**: 2.12%
* **Click-to-Open Rate**: 20.48%

These metrics were calculated by merging the datasets on email\_id.

**4. Predictive Modeling**

Two models were trained to predict the likelihood of a user clicking on an email:

* **Random Forest**
  + ROC AUC Score: 0.5892 (moderate performance)
* **XGBoost (Optimized)**
  + ROC AUC Score: **0.7255** (strong performance)
  + Outperformed Random Forest in distinguishing clickers from non-clickers.

*Note:*

Although, the ROC AUC score of **0.5892** of Random Forest model is just slightly better than random guessing (which is **0.5**). We optimised the score by using a different model, XG Boost which has a ROC AUC score of **0.7255**

**5. Simulated Optimization of Campaign**

| **% of users targeted** | **Model** | **Original CTR** | **Simulated CTR** | **Improvement** |
| --- | --- | --- | --- | --- |
| 30% | Random Forest | 2.12% | 2.12% | 1.00× |
|  | **XGBoost** | 2.12% | **4.35%** | **2.05×** |
| 20% | Random Forest | 2.12% | 3.77% | 1.78× |
|  | **XGBoost** | 2.12% | **4.97%** | **2.34×** |
| 10% | Random Forest | 2.12% | 4.03% | 1.90× |
|  | **XGBoost** | 2.12% | **5.96%** | **2.81×** |

*Note:*

XGBoost **consistently outperforms** Random Forest in predicting users likely to click on emails. At every targeting level (10%, 20%, 30%), XGBoost delivers **higher simulated CTRs** and greater improvements over the original 2.12% CTR. Notably, in the **top 30%** segment, Random Forest shows no improvement, **maintaining the original CTR**, while XGBoost still achieves a **4.35% CTR**—a 2.05× boost. Therefore, XGBoost should be the preferred model for future campaigns aiming to maximize user engagement through data-driven targeting.

**6. Segment Analysis & Insights**

**A. By Email Type and Version**

|  |  |  |
| --- | --- | --- |
| **Email Text** | **Version** | **CTR** |
| long\_email | generic | 1.37% |
| long\_email | personalized | 2.34% |
| short\_email | generic | 1.66% |
| short\_email | personalized | **3.12%** |

**Insights**:

* **Personalization emails** boosts CTR.
* Short emails **outperforms** long ones.
* **Short + personalized** emails perform best.

*Note:*

**Personalized emails** outperform generic ones in both short and long formats.

Example:

long\_email CTR improves from 1.37% → 2.34% with personalization.

short\_email CTR improves from 1.66% → 3.12% with personalization.

**Insights**: Including the **user’s name** ("Hi John") in the email significantly increases engagement.

**B. By Hour**

* CTR peaks during certain hours (especially late mornings ).
* Hour 23 **(11:00 PM)** is the best time to send emails to **maximize** user engagement.
* Evening hours between **8 PM and 9 PM** show the **lowest engagement**, suggesting those times should be avoided for email campaigns.

**7. Conclusion**

The project demonstrated how machine learning can:

* Enhance targeting in email campaigns.
* Nearly triple the CTR through intelligent user selection.
* Identify optimal email formats and timing strategies.

By using data-driven insights and modeling, marketing teams can create highly efficient and personalized campaigns.

**Questions**

**Q.1**: What percentage of users opened the email and what percentage clicked on the link within the email?

*Ans:*

* Percentage of users who opened the email:
  + Emails Sent = 100,000
  + Emails Opened = 10,345
  + Open Rate = (Emails Opened / Emails Sent) × 100 = (10,345 / 100,000) × 100 = 10.35%
* Percentage of users who clicked the link in the email:
  + Links Clicked = 2,119
  + Click-Through Rate (CTR) = (Users Who Clicked / Users Who Received the Email) × 100 = (2,119 / 100,000) × 100 = 2.12%

**Q.2**: Can you build a model to optimize how to send emails to maximize the probability of users clicking on the link inside the email?

*Ans:*

**Yes**, based on the data and insights from the campaign, a machine learning model can be built to optimize email sending for better engagement:

**Features to include in the model**:

* + - **Email Type and Version** (personalized vs. generic)
    - **Email Length** (short vs. long)
    - **Time of Day** (based on hour)
    - **Day of the Week**
    - **User History** (e.g., past purchases, engagement)
    - **User Location** (user\_country)

**Q.3**: By how much do you think your model would improve click-through rate (CTR)? How would you test that?

*Ans:*

Based on the **Simulated Optimization** section:

* + **XGBoost** model outperforms the **Random Forest** model consistently across different segments.
  + **Simulated CTR for XGBoost**:
    - At 30% targeting, the CTR improved from 2.12% to 4.35% (**2.05×** improvement).
    - At 20% targeting, the CTR improved from 2.12% to 4.97% (**2.34×** improvement).
    - At 10% targeting, the CTR improved from 2.12% to 5.96% (**2.81×** improvement).

**Q.4:** Did you find any interesting patterns on how the email campaign performed for different segments of users? Explain.

*Ans:*

* **Email Type and Version**:
  + **Personalized emails perform better**: Personalized emails outperformed generic ones in both short and long formats. For example:
    - **Long emails**: Generic (1.37%) vs. Personalized (2.34%).
    - **Short emails**: Generic (1.66%) vs. Personalized (3.12%).
  + **Short emails outperform long emails**: Short emails had a higher CTR than long ones across both versions (generic and personalized).
* **Best Time to Send**:
  + **Late-night (Hour 23)**: Emails sent around 11:00 PM had the highest engagement, suggesting that this is an optimal time to send emails for this specific user base.
  + **Avoid evening hours (8 PM - 9 PM)**: CTR was lower in the evening, especially between 8 PM and 9 PM, indicating that users may be less engaged during those times.
* **User Segmentation**:
  + Email performance was likely influenced by factors like user country and past purchases, though specific insights on those segments weren't fully provided. Future analysis could explore which user segments (e.g., based on purchasing behavior or geography) are more likely to engage with certain types of emails (personalized vs. generic).