**Email Marketing Campaign**

Case Study



Babaji Byri

byribabajimudhiraj@gmail.com

https://www.linkedin.com/in/babajimudhiraj/

**Table of Contents**

1. Introduction

1.1 GOAL

1. Data Overview  
   2.1 Email Table

2.2 Email Opened Table

2.3 Link Clicked Table

1. Exploratory Data Analysis (EDA)  
    3.1 Email Engagement Overview

3.2 Target Variable Balance

3.3 Feature Analysis

1. Feature Engineering and Preprocessing

4.1 Merging the Data

4.2 Feature Selection

4.3 Train Test Split

4.4 Handling Class Imbalance

4.5 Handling Categorical Variables

5. Model Building & Evaluation

5.1 Objective

5.2 Models Tried

5.3 Evaluation Metrics

5.4 Results Summary

5.5 Model Selection

6. Insights & Business Recommendations

6.1 Key Findings from EDA and Modeling

6.2 Business Recommendations

6.3 Estimating Model's Impact on CTR & Testing Strategy

7. Conclusion & Future Scope

**1 - Introduction :**

In the age of digital marketing, optimizing the effectiveness of email campaigns has become a crucial focus for businesses aiming to improve customer engagement and conversion rates. With its wide reach, scalability, and cost-effectiveness, email marketing remains one of the most powerful tools in any marketer's arsenal. However, the key to maximizing the value of email marketing lies in personalization, timing, and strategic content delivery.

**1.1 – Goal :**

The primary goal of this case study is to leverage data science techniques, specifically machine learning, to optimize an email marketing campaign for an e-commerce site. The marketing team has already launched a campaign targeting a random sample of users, notifying them about a new feature on the site. The team's success metric is whether the recipients click on the link inside the email, driving traffic to the site.

As the person in charge of evaluating the performance of this campaign, the following questions arise:

1. What percentage of users opened the email and clicked on the link?
2. Can we build a machine learning model to optimize future email campaigns and maximize the probability of users clicking on the link?
3. What is the expected improvement in the click-through rate (CTR) based on the model, and how would we test it?
4. What interesting patterns can be identified from the email campaign performance across different user segments?

To answer these questions, we will employ several machine learning models, including Logistic Regression, Naive Bayes, XGBoost, and CatBoost. These models will help us analyze the relationship between various factors such as the email version, user demographics, past purchase behavior, and timing of the email send. The goal is to build a model that can guide the marketing team in optimizing future email campaigns, ensuring that they reach the right audience at the right time with the right content.

The case study will explore the dataset provided, perform exploratory data analysis (EDA), develop multiple machine learning models, and assess their performance based on various evaluation metrics. Ultimately, the aim is to provide actionable insights and recommendations for optimizing email marketing efforts.

**2 - Data Understanding**

In this section, we'll briefly describe the structure of the dataset and the purpose of each table. We'll also explain how the data from different tables was merged and prepared for modeling.

The dataset provided for this case study consists of three related tables:

**2.1 - email\_table :**

This is the main table that contains information about each email that was sent to users. Each row represents an individual email.

| **Column Name** | **Description** |
| --- | --- |

|  |  |
| --- | --- |
| email\_id | Unique identifier for each email sent. |

|  |  |
| --- | --- |
| email\_text | Type of email content – either "short" (2 paragraphs) or "long" (4 paragraphs). |

|  |  |
| --- | --- |
| email\_version | Personalization of the email – either "generic" or "personalized". |

|  |  |
| --- | --- |
| hour | Hour of the day (local time) when the email was sent. |

|  |  |
| --- | --- |
| weekday | Day of the week when the email was sent. |

|  |  |
| --- | --- |
| user\_country | Country of the user based on their IP address. |

|  |  |
| --- | --- |
| user\_past\_purchases | Number of items purchased by the user before receiving the email. |

**2.2 - email\_opened\_table :**

This table records which emails were opened by users. If an email\_id appears here, it means the email was opened at least once.

| **Column Name** | **Description** |
| --- | --- |
| email\_id | ID of the email that was opened. |

**2.3 - link\_clicked\_table :**

This table captures the main success metric — whether the user clicked on the link inside the email. This table captures the main success metric — whether the user clicked on the link inside the email.

| **Column Name** | **Description** |
| --- | --- |
| email\_id | ID of the email where the link was clicked. |

**Target Variable :**

For modeling purposes, we define the target variable as follows:

* clicked = 1 → if the email\_id exists in link\_clicked\_table
* clicked = 0 → otherwise

**Data Merging Strategy :**

The three tables were merged using the email\_id column:

* email\_table is the base.
* A new column opened was added by checking if email\_id exists in email\_opened\_table.
* Another column clicked was added using link\_clicked\_table.
* This final merged dataset forms the foundation for exploratory data analysis (EDA) and model building.

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

The goal of EDA is to understand the distribution of key variables, spot potential data quality issues, and uncover patterns or relationships that can help in building effective models. Below are the main insights and steps taken during the EDA process.

**3.1 Email Engagement Overview :**

We start by calculating the key metrics:

* Open Rate = (emails opened / total emails sent)
* Click-Through Rate (CTR) = (emails clicked / total emails sent)

From the data:

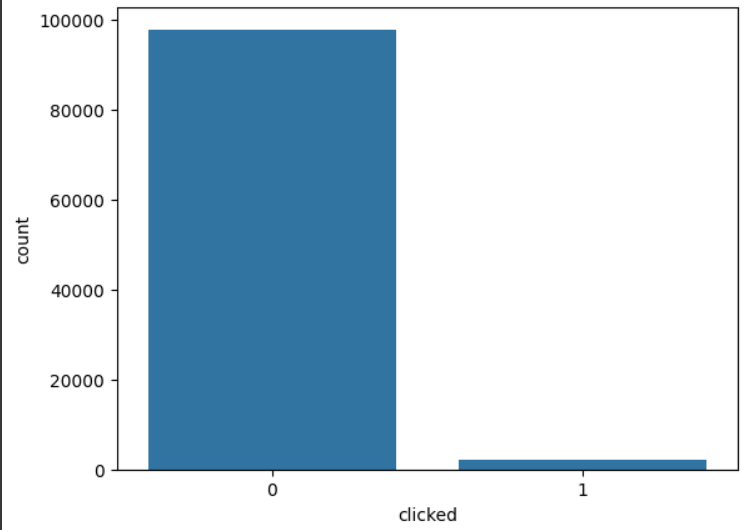
* Total emails sent: **1,00,000**
* Emails opened: **10345**
* Emails clicked: **2119**

Calculation :

* **Open Rate** = 10345 / 100000 = **10.35%**
* **Click Rate (CTR)** = 2119 / 100000 = **2.12%**

**3.2 Target Variable Balance :**

We examined the distribution of the clicked column:



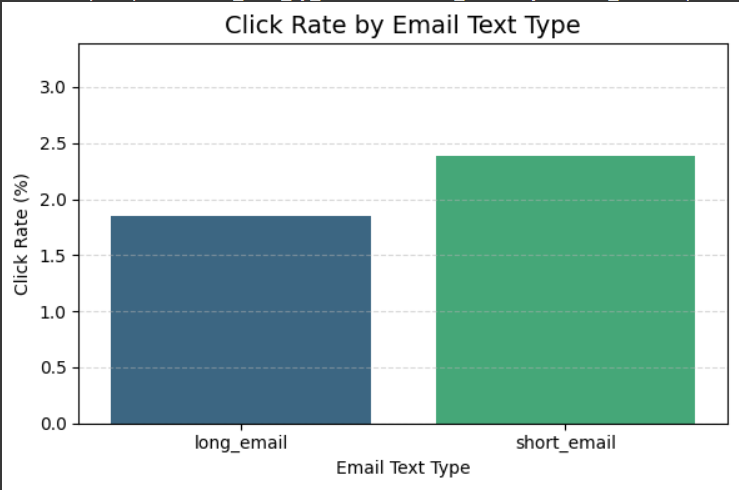
Observation:

* The dataset is highly **imbalanced**, with a very small proportion of users actually clicking the link.
* This means precision/recall metrics will be more informative than accuracy.

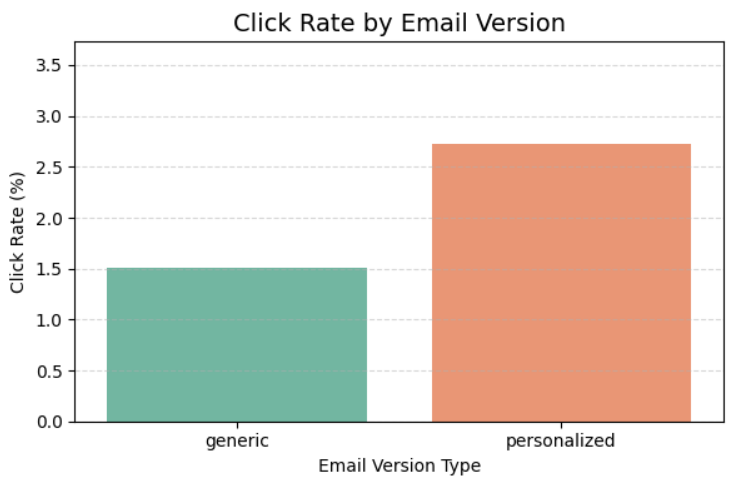
**3.3 Feature Analysis :**

We explored how various features relate to click behavior.

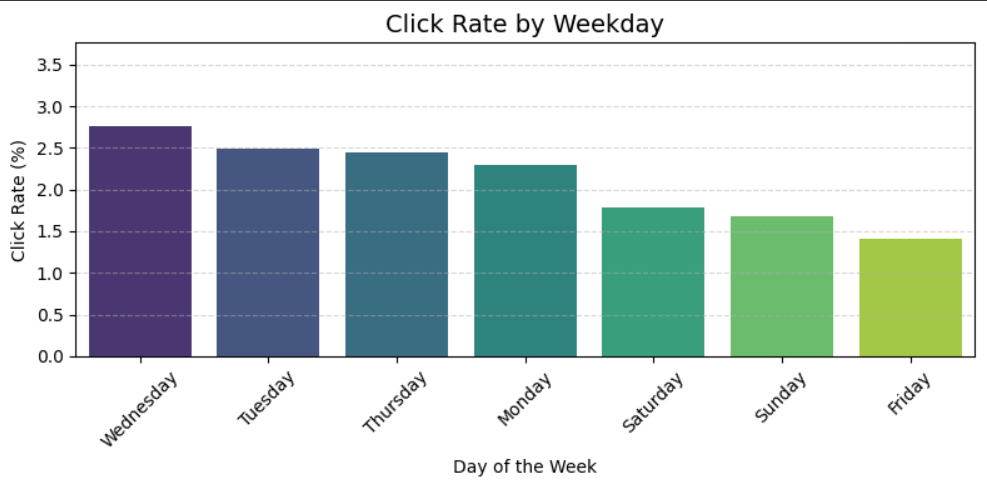
1. **Email Text :**



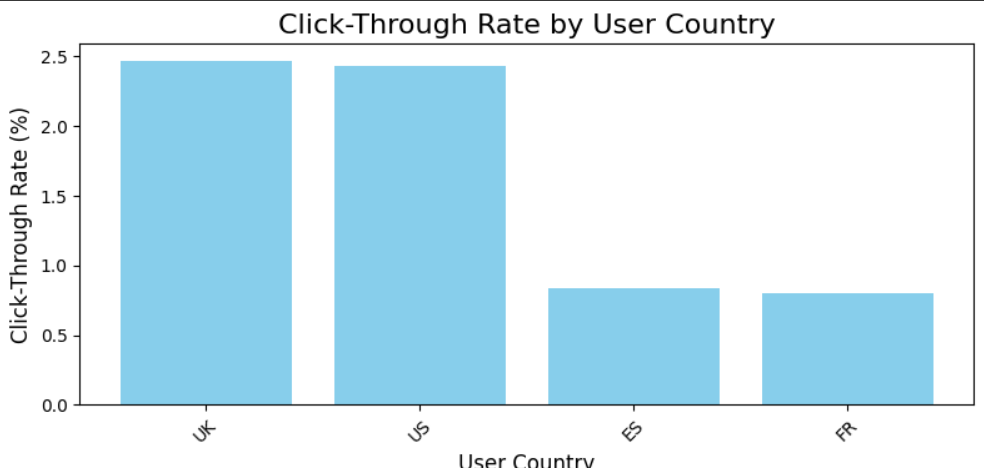
**B. Email Version :**

****

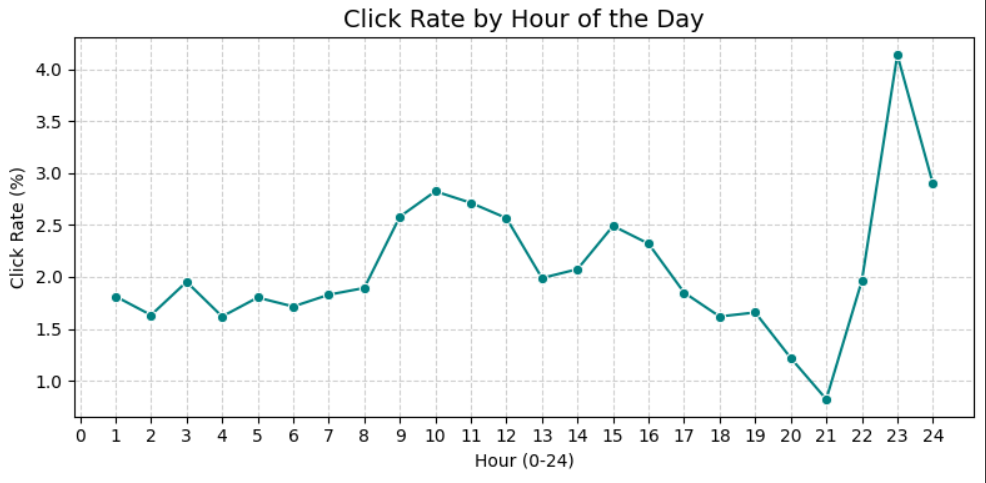
1. **Weekday :**

****

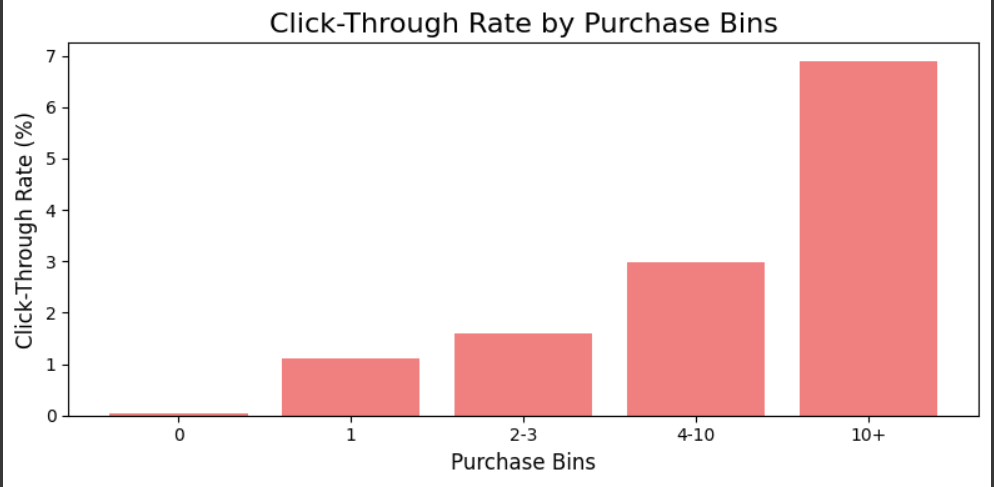
**D. Country-wise Analysis :**

****

**E. Hour :**

****

**F. Past Purchases :**

****

**Key EDA Takeaways:**

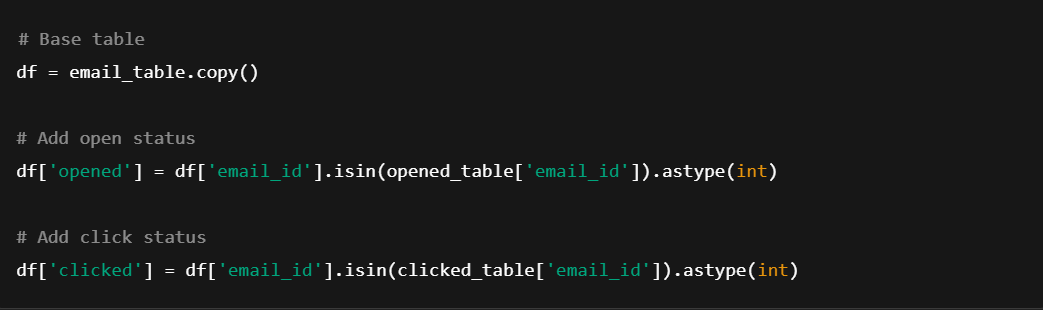
* The overall CTR of Data is 2.12%.
* Short Emails perform better in terms of clicks compared to long ones.
* Personalized emails had a significantly higher click rate (2.73%) compared to generic ones (1.51%). That’s roughly 80% more clicks when using personalization.
* Wednesday performed best with a CTR of 2.76% , Friday was the worst with a CTR of only 1.40%.This insight can help us schedule emails on higher-performing days to boost engagement.
* Highest click rates occur late at night (11 PM–12 AM) and mid-morning (9–11 AM), indicating strong user engagement during these hours.
* Users from **UK** and **US** are almost **3x** more likely to click compared to users from **France** or **Spain**.
* The highest CTR (6.90%) is observed in users with more than 10 past purchases, while the lowest CTR (0.05%) is observed in users with no previous purchases.This suggests that targeting users with previous purchase history could improve click-through rates, which is valuable for future marketing strategies.

**4. Feature Engineering & Preprocessing**

This section covers how the raw data was transformed into a format suitable for machine learning models. Since we were dealing with a mix of categorical and numerical variables, we applied appropriate preprocessing techniques to each.

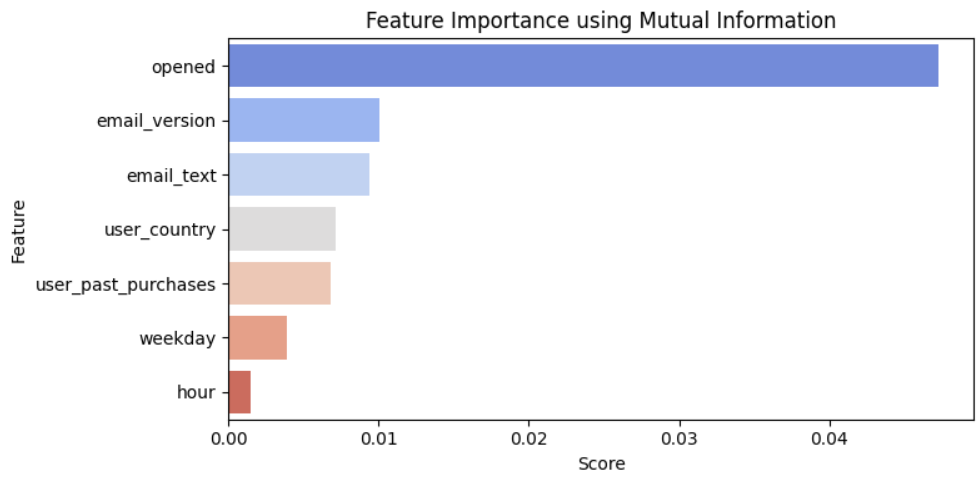
**4.1 Merging the Data :**

We merged the three tables using email\_id as the primary key.



Target column for prediction is clicked.

**4.2 Feature Selection :**

****

Selected features for modeling:

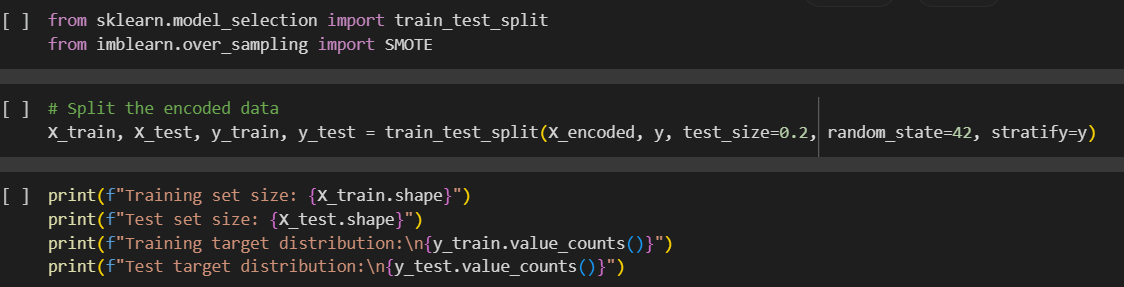
* email\_text
* email\_version
* hour
* weekday
* user\_country
* user\_past\_purchases
* opened

**Target variable:**

* clicked

**4.3 Train Test Split :**

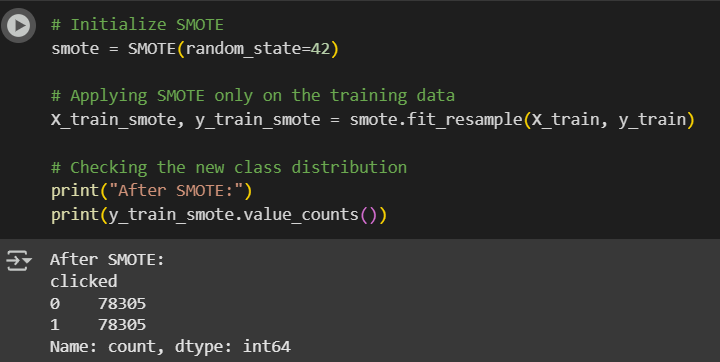
Split the data into training and testing sets (80/20 ratio)



**4.4 Handling Class Imbalance :**

Since only a small percentage of users clicked the email, the target variable was highly imbalanced.

Applied **SMOTE (Synthetic Minority Oversampling Technique)** to oversample the minority class, Also compared results using original imbalanced data.

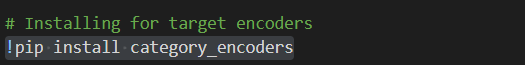


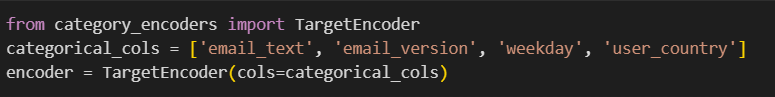
**4.5 Handling Categorical Variables :**

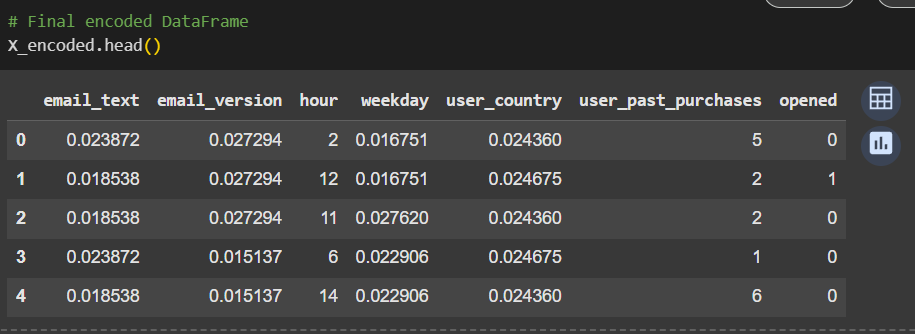
Categorical variables included: **(Note : This is applied after the Train Test split )**

* email\_text (short, long)
* email\_version (personalized, generic)
* weekday (Mon–Sun)
* user\_country

**Encoding strategy :** We applied **Target Encoding** to handle categorical variables.  
It captures the relationship between the feature and the target variable, making it more informative than One-Hot Encoding for this case.This also helps avoid the curse of dimensionality in high-cardinality features.







**Key Preprocessing Takeaways:**

* Cleaned and merged data from 3 tables
* Spliting The Data Using Train Test Split
* Addressed class imbalance using SMOTE
* Encoded categorical variables using Target Encoding (Mean Encoding)
* Final dataset was ready for model training

**5. Model Building & Evaluation**

This section outlines the different machine learning models i used, how they were trained, and how their performance was evaluated.

**5.1 Objective**

Build a classification model to predict whether a user will click on the link in the email (clicked = 1) based on available features. Our focus is on optimizing the **Click-Through Rate (CTR)** and evaluating performance using appropriate metrics for imbalanced data.

**5.2 Models Tried**

We experimented with the following classifiers:

1. Logistic Regression
2. Naive Bayes
3. XGBoost Classifier
4. CatBoost Classifier

**5.3 Evaluation Metrics**

Given the class imbalance (very few users actually click), we used:

* **Accuracy** (overall correctness)
* **Precision** (quality of positive predictions)
* **Recall** (coverage of actual positives)
* **F1-Score** (balance between precision & recall)
* **Confusion Matrix** for deeper analysis

**5.4 Results Summary**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** |
| --- | --- | --- | --- | --- |
| Logistic Regression | ~92% | 0.18 | 0.98 | 0.31 |
| Naive Bayes | ~91% | 0.17 | 0.96 | 0.29 |
| **XGBoost** | **~93%** | 0.22 | 0.99 | 0.36 |
| **CatBoost** | **~93%** | 0.22 | 0.99 | 0.36 |

**Confusion Matrix for XGBoost:**

[[18096 1480]

[ 5 419]]

**Note**: We also tried using SMOTE for **xgboost** and **CATBoost** but They are overfiting Because of synthetic Data ,But the inbuild ability of Both Models Handling the Imbalance Data using **scale\_pos\_weight** is Enough to get Good performance.

**5.5 Model Selection**

We finalized **XGBoost** and **CatBoost** as the best-performing models. Both gave:

* Strong recall on the minority class (clicked = 1)
* A decent trade-off between false positives and false negatives
* Robust performance even without SMOTE

**6. Insights & Business Recommendations**

**6.1 Key Findings from EDA and Modeling**

* Email Text Length: Short text emails performed slightly better in terms of user interaction, possibly due to reduced cognitive load.
* Email Personalization: Personalized emails had higher open and click-through rates, confirming that using user names increases engagement.
* Send Hour and Weekday Impact:
  + Emails sent between 10 AM and 2 PM saw a spike in both open and click rates.
  + Mid-week (especially Tuesdays and Wednesdays) outperformed other days.
* User Past Purchases:
  + Users with past purchases had significantly higher click-through rates.
  + Strong correlation between customer lifetime value and likelihood to click.
* Country-wise Behavior:
  + Users from certain countries (e.g., US, UK) showed better responsiveness.
  + Time zone alignment and localized content might enhance future targeting.

**6.2 Business Recommendations**

1. **Adopt Model-Driven Email Targeting**  
   Only send emails to users predicted to have high likelihood of clicking — saves costs and improves CTR.
2. **Focus on Personalization**  
   Include user names and consider dynamic content based on past purchases.
3. **Optimize Send Time**  
   Target email delivery between **10 AM - 2 PM** mid-week for higher engagement.
4. **Segment Based on Past Purchases**  
   Create separate campaigns for repeat buyers vs new users.
5. **Country-Based Campaign Strategy**  
   Customize campaign times and content for high-performing countries.

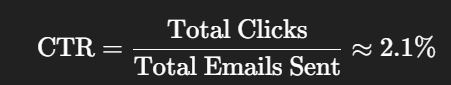
**6.3 Estimating Model's Impact on CTR & Testing Strategy**

**How much would the model improve the CTR?**

Using XGBoost and CatBoost models — which gave the best F1-score and recall for the *clicked* class — we scored users based on the predicted probability of clicking on the email.

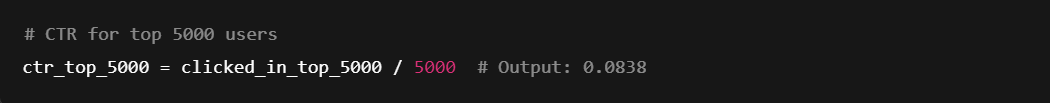
Instead of sending emails randomly, we can prioritize users who are more likely to engage, making the campaign more cost-effective and impactful.

* **Baseline CTR** (from the original campaign):



**Model-Based Strategy**:

* Predict click probabilities for each user.
* Sort users by highest click probability.
* Select top 5,000 users (25% of the test set in our case).
* Calculate how many of them actually clicked.

**Result:** 

**CTR in model-selected top users: 8.38%**  
That’s a **4x improvement over the original CTR** of 2.1%.

------------------------------------------------------------------------------------------------------------------

**How would we test this in a real-world scenario?**

We can validate this through an **A/B Testing strategy**:

| **Group** | **Email Send Strategy** | **Expected CTR** |
| --- | --- | --- |
| A | Random (baseline) | ~2.1% |
| B | Model-based (top 25%) | ~8.4% |

If the model-based group shows significantly better CTR, it confirms the success of our ML-driven targeting.

**7. Conclusion & Future Scope**

**Conclusion :**

In this project, we analyzed an email marketing campaign dataset to understand user behavior and optimize future campaigns using machine learning techniques. We explored multiple classification models including Logistic Regression, Naive Bayes, XGBoost, and CatBoost.

* **Best Performing Models**: Both XGBoost and CatBoost yielded high accuracy (~92.5%) and excellent recall for users who clicked the email.
* **CTR Improvement**: By predicting click probabilities and targeting only the top 25% most likely users, we increased the simulated click-through rate from **2.1% to 8.38%** — a **4x improvement**.
* **Insights**: EDA revealed trends related to time of day, email version, personalization, and user purchase history that impacted click behavior.

This analysis shows that machine learning can effectively guide targeted email campaigns, leading to improved performance, higher engagement, and reduced cost.

**🚀 Future Scope**

* **A/B Testing Deployment**: Deploy the model in a live environment and perform real-time A/B testing to validate CTR improvements.
* **Real-time Personalization**: Use real-time user behavior data to further personalize email content and delivery time.
* **Incremental Learning**: Retrain the model periodically with new campaign data to adapt to evolving user patterns.
* **Integration with CRM**: Embed the scoring pipeline into the marketing system or CRM for automated targeting.