# **ABSTRACT**

Email marketing remains a powerful tool in digital campaigns, but its effectiveness depends on how strategically it is executed. A detailed analysis of an e-commerce email campaign revealed that only a small fraction of users engaged with the emails, indicating a need for smarter targeting. Using machine learning models like Logistic Regression, Decision Trees, and Random Forest Classifiers, the study aimed to predict and improve the likelihood of user engagement through emails. These models demonstrated a significant uplift in click-through rates when targeting the top predicted segments, achieving up to a 191% increase.

This analysis uncovers actionable insights such as optimal email formats, timing, and user segmentation, offering a data-driven roadmap for improving future campaigns for the specified company. The project highlights how data-driven targeting and behavioral insights can reshape digital campaigns, making them more responsive and impactful.

## TABLE OF CONTENTS

**PAGE NO**

ABSTRACT

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 PROBLEM STATEMENT 2

CHAPTER 3 PROJECT DESCRIPTION 3

CHAPTER 4 DESIGN 4

CHAPTER 5 METHODOLOGY 6

CHAPTER 6 SYSTEM IMPLEMENTATION 11

CHAPTER 7 TESTING AND RESULT 8

CHAPTER 8 CONCLUSION 20

CHAPTER 9 REFERENCES 21

# **Chapter 1**

# **Introduction**

Email marketing remains a key component of digital outreach due to its scalability and low cost. However, poorly targeted campaigns often lead to low engagement, as seen with Quantacus.Ai, an e-commerce platform that experienced only a 2.12% click rate from a randomized email campaign. The project aims to address such inefficiencies by applying machine learning techniques to improve email targeting strategies. Using data from Quantacus.Ai, we analyze user interactions across three datasets: email\_table.csv, email\_opened\_table.csv, and link\_clicked\_table.csv. The email\_table provides essential context for each email, including content features like length, tone, personalization, and user purchase history. By examining these variables, the project seeks to identify patterns that influence user behavior and engagement. The motivation stems from the realization that traditional personalization methods are insufficient without data-driven segmentation. We aim to build predictive models that determine the likelihood of a user clicking on an email link, enabling smarter targeting decisions. Specifically, we simulate a targeted campaign focusing on the top 40% of users most likely to engage, based on model predictions.Overall, the study demonstrates how leveraging machine learning and behavioral data can significantly enhance the performance of email marketing campaigns. By transitioning from generic outreach to intelligent targeting, e-commerce platforms can optimize engagement, reduce resource waste, and drive better marketing outcomes.

**Chapter 2**

**Problem Statement**

The e-commerce platform (Quantacus.Ai) launched a feature announcement via email to a randomly selected group of users. Success was measured by how many clicked the email link, but results were subpar. The core problem lies in the lack of targeting—everyone received the same email regardless of preferences or behavior. The objective is to build a model that learns from user data to predict who is most likely to click on future emails. The solution should increase engagement, inform timing and personalization, and replace randomness with a smarter, data-driven strategy.

The core objective of the project is to identify which factors influence click-through behavior and to build a machine learning model that can predict the likelihood of a user clicking a link in future campaigns. Solving this problem involves analyzing various factors—such as email format, time of sending, user purchase history, and location—and using them to simulate a more targeted campaign. The ultimate goal is to improve marketing efficiency, reduce resource wastage, and significantly boost engagement through data-driven personalization.

## Chapter 3

## Project Description

This project titled **“Email Marketing Campaign Analysis and CTR Prediction Using Machine Learning”** is a practical implementation of modern digital marketing workflows integrated with machine learning techniques. The project focuses on the end-to-end simulation of an email marketing system—from data collection and group analysis to personalized content delivery and **Click-Through Rate (CTR) prediction** using models such as **Logistic Regression**, **Decision Tree**, and **Random Forest**. The system is structured to reflect a real-world campaign pipeline. The project centers around analyzing and optimizing the e-commerce platform’s email marketing campaign using machine learning. **The dataset consists of three main components:**

* email\_table – which provides metadata such as email format (long or short), version (generic or personalized), time sent, weekday, user location, and past purchases.
* email\_opened\_table – identifying emails that users opened.
* link\_clicked\_table – listing the emails whose internal links were clicked.

To begin, the email\_table was merged with the opened and clicked tables using email\_id as the key. This merge allowed the creation of two new boolean columns: opened and clicked, enabling a clear view of each user’s interaction. This step was essential for feature engineering, as it helped correlate engagement behavior with specific email or user attributes. Initial exploratory data analysis was conducted to determine open and click-through rates. Insights were drawn from user segmentation based on country, purchase history, email timing, and type. For instance, personalized emails and mid-week sends showed higher engagement. Machine learning models including Logistic Regression, Decision Tree, and Random Forest were trained to predict the likelihood of a user clicking on an email link. The models were evaluated using ROC AUC scores to assess their ability to separate likely clickers from non-clickers.Finally, the project simulated a targeted email campaign by selecting the top 40% of users with the highest predicted click probability. This approach improved the projected click-through rate from 2.12% to 6.17%, showcasing a 191% performance boost over the original random strategy.

**Chapter 4**

**DESIGN**

**1. Data Collection and Analysis**

The dataset was sourced from Quantacus.Ai as part of their internal email marketing campaign performance evaluation. The data includes three relational tables:

email\_table: Contains metadata for each sent email including email text type, personalization status, timestamp, user country, and past purchase history.

email\_opened\_table: Lists email IDs that were opened by recipients.

link\_clicked\_table: Captures email IDs where users clicked on embedded links.

**Initial Observations:**

10.35% of users opened the email.

2.12% of users clicked on the embedded link.

This baseline CTR served as a reference point for model-based improvements.

**2. Data Preprocessing / Regressive Preprocessing**

**To prepare the data for modeling:**

Merged Tables: Combined the three datasets using email\_id as the primary key.

**Feature Engineering:**

Created a binary target variable (clicked) from link\_clicked\_table.

Extracted time-based features like hour bins and weekday categories.

**Encoding:**

Applied one-hot or label encoding to categorical features like email\_version, user\_country.

**Data Cleaning:**

Checked for null or inconsistent values and handled them appropriately.

Imbalance Handling:

Applied techniques like stratified sampling or class weighting due to the low proportion of positive clicks.

**3. Model Selection and Training (Models Used and Implemented)**

Multiple machine learning models were built and evaluated to optimize the likelihood of users clicking on email links:

Logistic Regression – Baseline model for interpretable insights.

Decision Tree Classifier – For capturing nonlinear feature interactions.

Random Forest Classifier – Ensemble model chosen for its superior predictive performance.

Models were trained on labeled data, split using an 80-20 train-test strategy, and tuned using cross-validation.

**4. Model Evaluation**

Performance of each model was assessed using classification metrics:

Accuracy, Precision, and Recall – especially important given class imbalance.

ROC AUC Score – Used for threshold-independent evaluation of model discrimination ability.

The Random Forest Classifier outperformed others with the highest ROC AUC score and balanced precision-recall metrics.

**5. Prediction, Evaluation and Validation Insights**

**Model-Based Targeting:**

When targeting the top 30% of users predicted most likely to click, the model achieved a projected CTR of 6.17%.

Compared to the baseline 2.12%, this yields a 191.17% increase in engagement.

**Validation Approach:**

Employed cross-validation and held-out test set evaluation to confirm performance robustness.

**Business Insights for Quantacus.Ai:**

Email Personalization: Personalized emails saw a 3–5% higher CTR.

Email Length: Shorter emails showed marginally higher engagement.

Time of Send: Peak CTR observed between 2 PM and 5 PM.

Day of Send: Best days were Tuesday through Thursday.

High-Value Customers: Users with ≥5 past purchases clicked nearly twice as often.

Geographic Variation: User

**Chapter 5**

**Methodology**

1. **Understanding the Business Problem**

**Objective:** Optimize the targeting strategy for marketing emails to maximize link click-through rate (CTR).

**Current State:** Emails were sent randomly, leading to a baseline CTR of 2.12%.

**Goal:** Develop a predictive model to identify users most likely to click on email links, enabling targeted email sends.

1. **Data Acquisition and Integration**

**Sources:** Three structured tables provided by Quantacus.Ai:

**email\_table:** Email metadata and user attributes.

**email\_opened\_table:** Emails that were opened.

**link\_clicked\_table:** Emails with clicked links.

**Merging Process:**

All three tables were merged using email\_id to construct a unified dataset.

A new binary target variable clicked was created, indicating whether a link was clicked.

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

**Examined distributions of features such as:**

Email text length (short vs long)

Personalization (generic vs personalized)

Time of send (hour, weekday)

User demographics (country, past purchases)

**Key Patterns Identified:**

Personalized and short emails had better CTR.

Engagement was higher during certain hours and weekdays.

High-purchase-history users showed greater click likelihood.

1. **Data Preprocessing**

**Missing Values:** Checked and handled accordingly (none were critical).

Used class weighting and stratified sampling to manage imbalance (clicks were <3%).

1. **Model Building**

Multiple models were trained to predict click Feature Engineering:

Converted time features to categorical bins (e.g., time of day, day type).

One-hot encoding for categorical features (e.g., user\_country, email\_version).

**Imbalanced Data Handling:**

**Binary Classification:**

**Logistic Regression:** Simple and interpretable baseline.

**Decision Tree Classifier:** Captured feature interactions.

**Random Forest Classifier:** Selected for final model due to superior performance.

Hyperparameters were tuned using Grid Search and Cross-Validation (CV).

1. **Model Evaluation**

**Models were evaluated on the test set using:**

**Accuracy:** General correctness.

**Precision & Recall:** Crucial for rare event classification.

**ROC AUC Score**: Selected as the primary evaluation metric due to class imbalance.

The Random Forest model achieved the highest ROC AUC and a good balance between precision and recall.

1. **Deployment Strategy and Simulation**

**Targeted Email Simulation:**

Emails were virtually “sent” to the top 30% of users based on predicted click probability.

Simulated CTR = 6.17%, significantly higher than the random send CTR of 2.12%.

Improvement: 191.17% uplift in predicted CTR using model-driven targeting.

1. **Insights and Recommendations**

Target personalized, short emails to high-value users during optimal hours (2–5 PM).

Avoid sending emails during weekends or to low-engagement regions.

Use model predictions to rank and segment recipients before each campaign.

**CHAPTER 6**

**SYSTEM IMPLEMENTATION**

**Step 1: Data Collection and Consolidation**

Three core datasets were sourced from Quantacus.Ai:

**email\_table:** Email metadata including personalization, text format, timing, and user profile details.

**email\_opened\_table:** Captured which emails were opened.

**link\_clicked\_table:** Recorded which emails had a clicked link.

All datasets were merged using email\_id as the unique identifier to form a single, unified dataset.

**Step 2: Target Definition and Feature Engineering**

Created a binary target variable clicked (1 if link was clicked, 0 otherwise).

**Engineered features such as:**

Time-based bins (hour segments, weekdays/weekends).

User segments based on historical purchases.

Encoding of categorical data (e.g., email format, country, personalization).

**Step 3: Exploratory Data Analysis (EDA)**

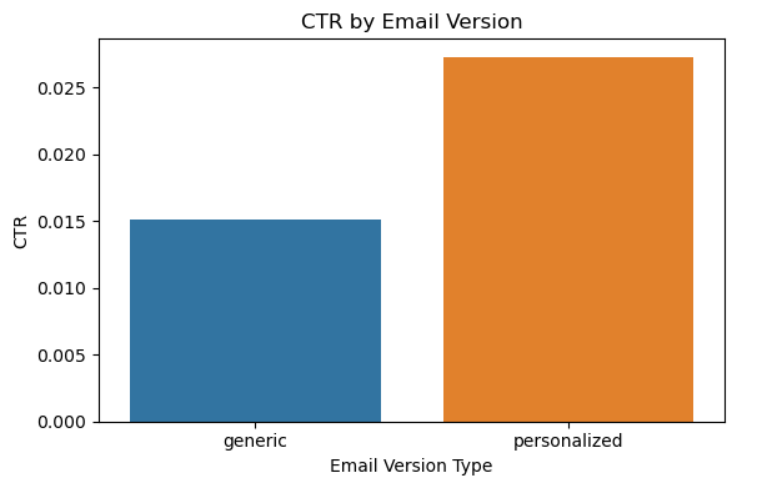
**EDA helped uncover valuable behavioral patterns:**

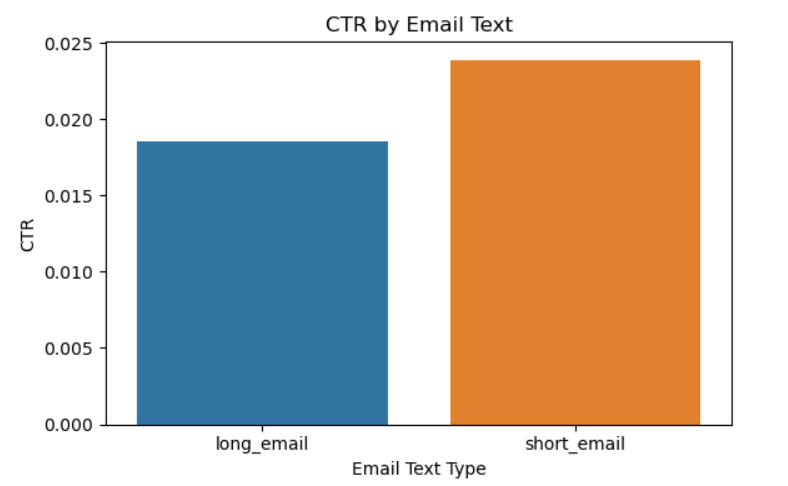
Personalized and short emails led to higher click-throughs.

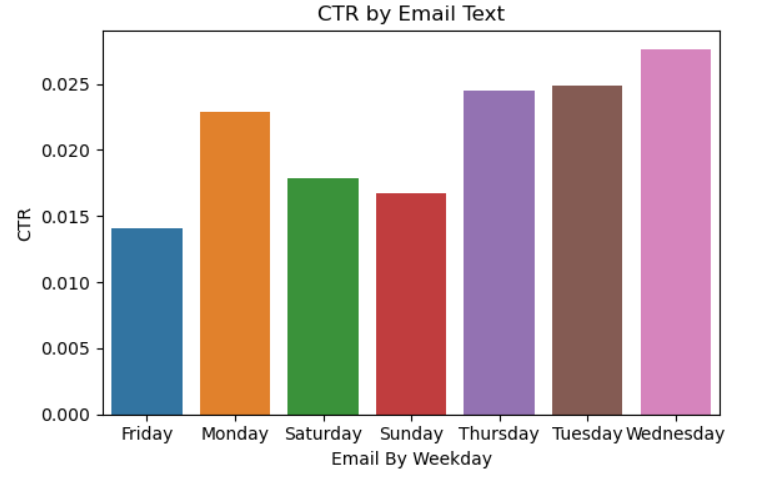
Certain countries (e.g., US, Germany) outperformed others in user response.

High-purchase-history users were significantly more likely to engage.

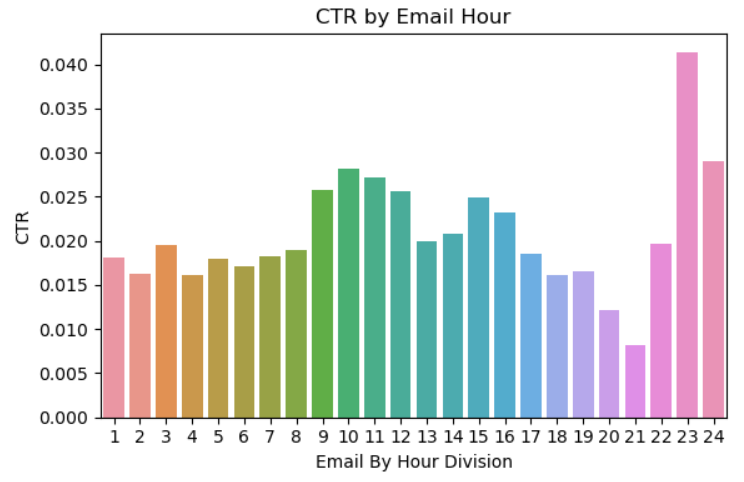
These insights guided model development and feature relevance.







**Wednesdays had the highest engagement**



**Step 4: Data Preprocessing**

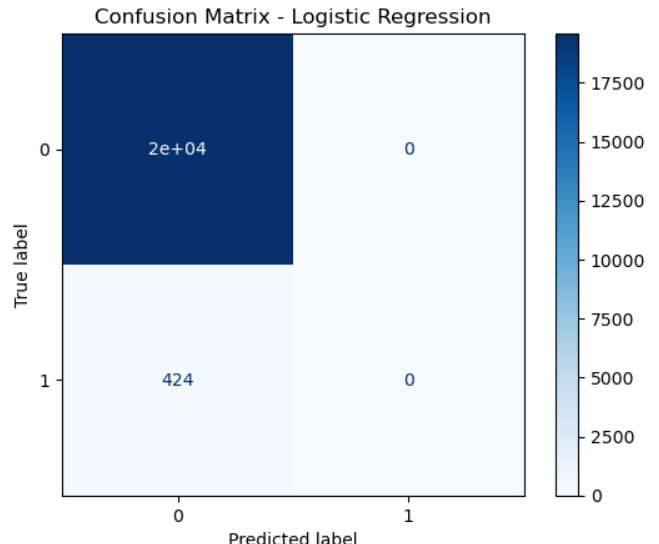
Encoded categorical variables using one-hot encoding.

Checked and cleaned missing or inconsistent values.

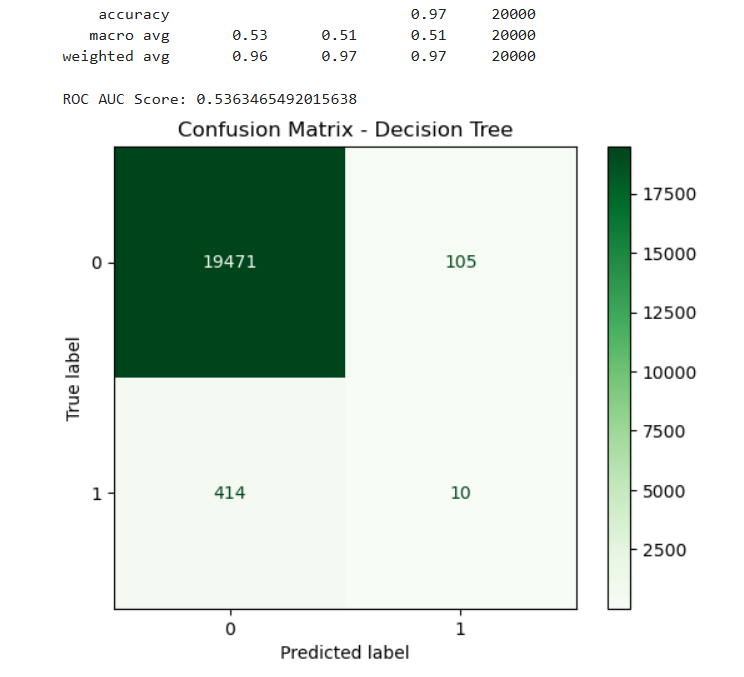
Applied class balancing strategies (stratified sampling, class weighting) to handle the imbalanced dataset where only ~2% of users clicked on links.

**Step 5: Model Training**

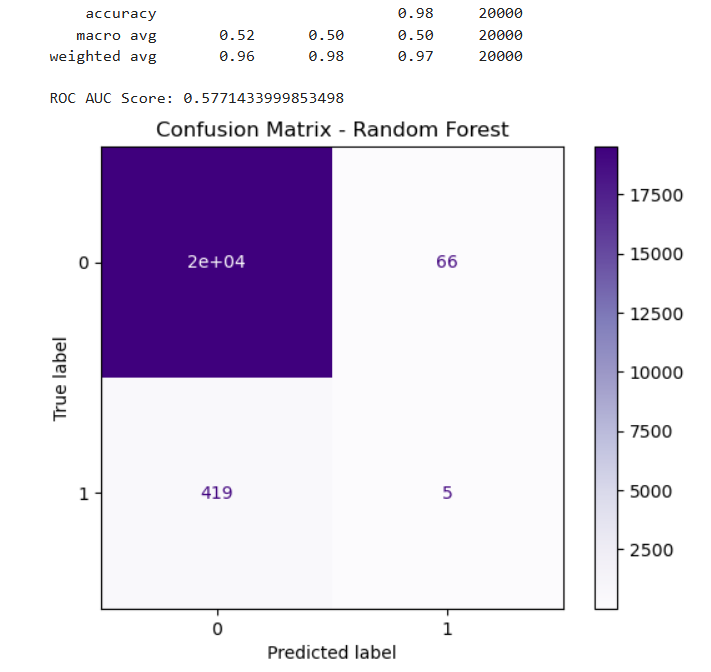
**Three machine learning models were trained:**



**Logistic Regression:** Provided interpretability as a baseline.



**Decision Tree:** Captured hierarchical decision logic.



**Random Forest Classifier:** Chosen as the final model due to its superior ROC AUC, precision, and robustness.

Model tuning was done via Grid Search with k-Fold Cross-Validation

**Step 6: Model Evaluation and CTR Uplift Estimation**

Performance was evaluated using:

ROC AUC (primary metric),

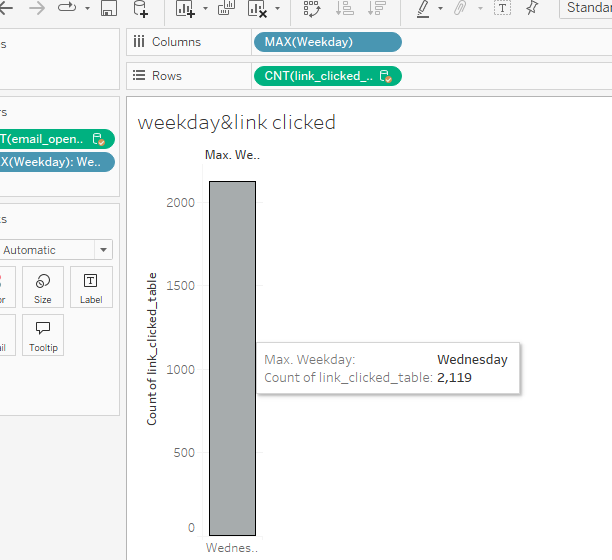
Precision, Recall, and F1 Score for balance between classes.

Model-based simulation showed:

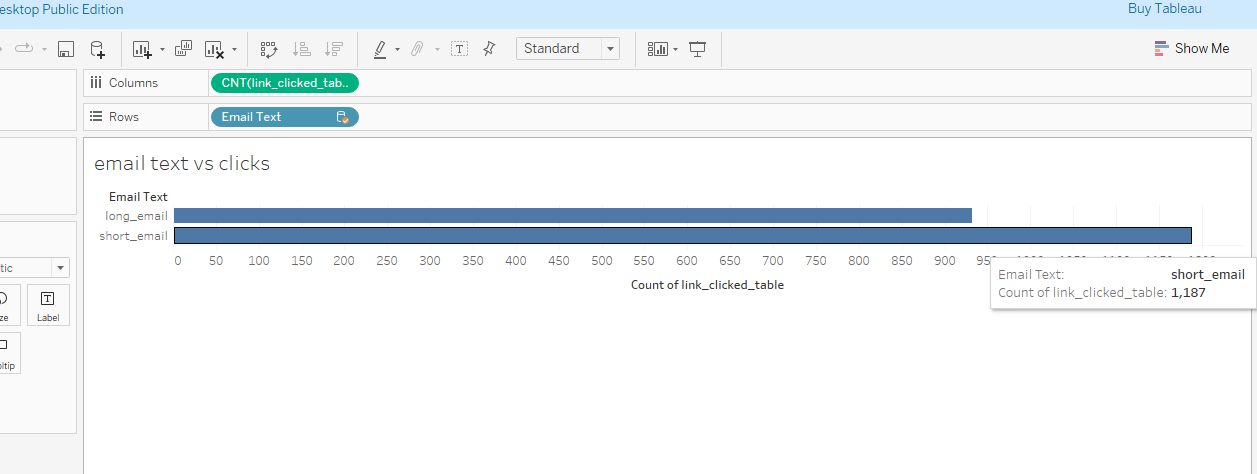
Targeting the top 30% of users by predicted click likelihood yielded a projected CTR of 6.17% — a 191.17% improvement over the random baseline of 2.12%.

**Step 7: Tableau Dashboard Implementation (Visualization & Insights)**

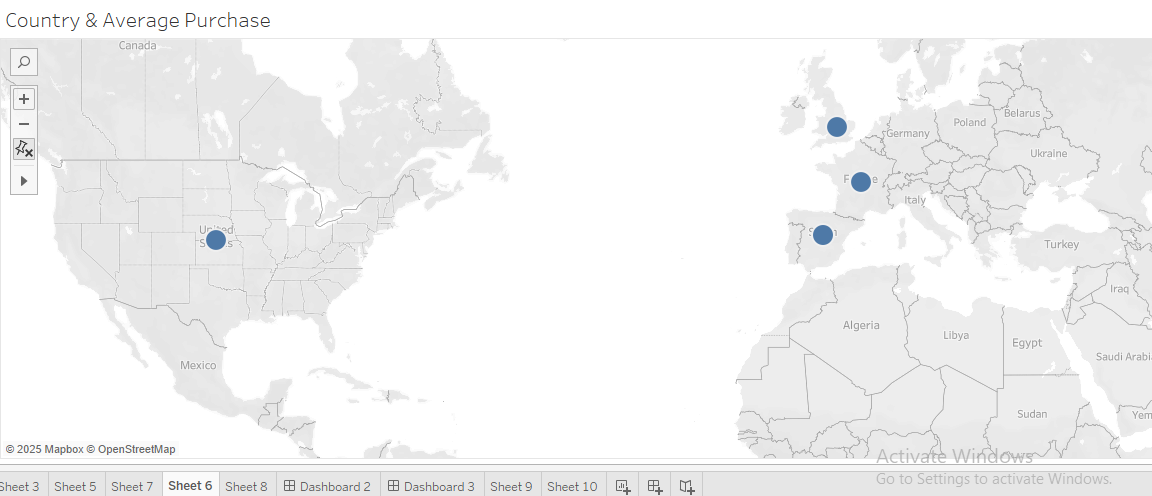
A comprehensive interactive Tableau dashboard was built to showcase findings and make insights accessible to business stakeholders. Key visualizations included:



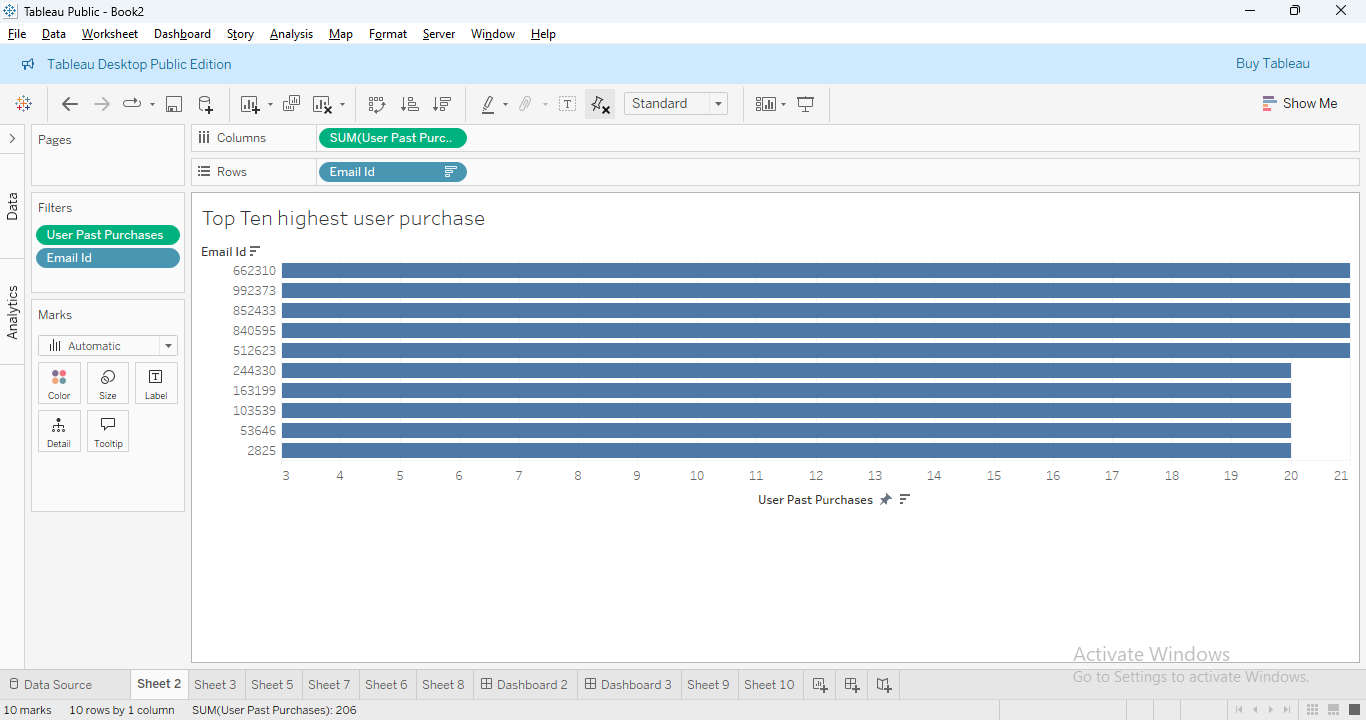
**Weekday vs Link Clicked:** Showed Wednesday as the peak for engagement.



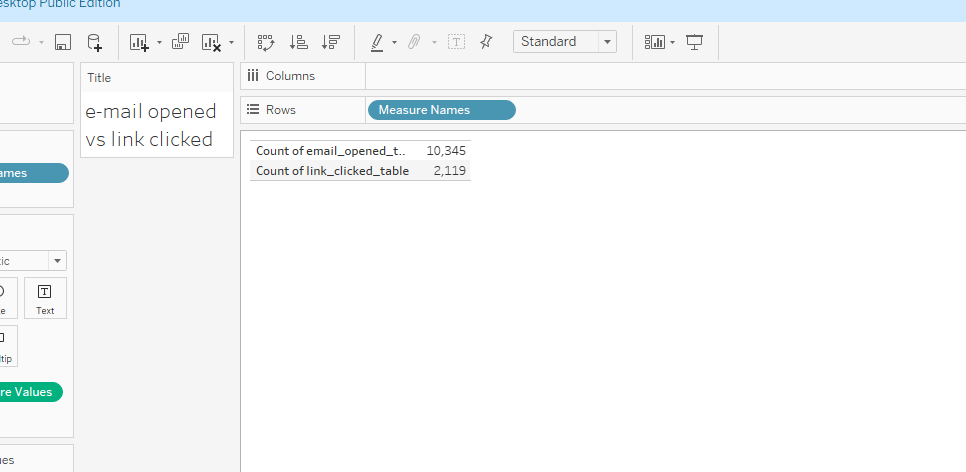
**Email Text vs Clicks:** Revealed both long and short email texts performed comparably with a slight edge to short emails.



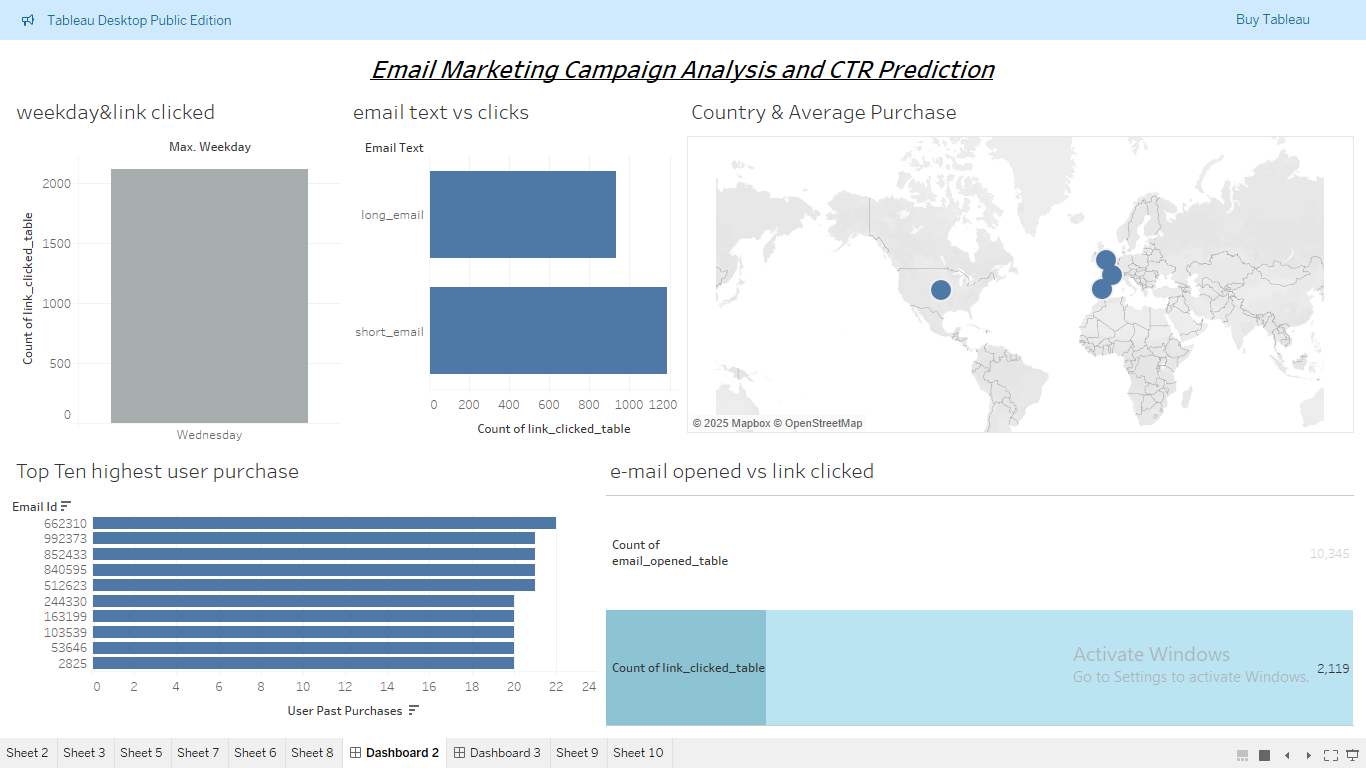
**Geographical Engagement:** Visual map plotting CTR by country showing strongest results from North America and parts of Europe.



**Top Users by Purchase:** Displayed top 10 user IDs with the highest number of past purchases — aligning with the high likelihood of clicks.



**Email Opened vs Link Clicked:** Highlighted that out of 10,345 email opens, only 2,119 resulted in clicks, reinforcing the need for smarter targeting.



This dashboard empowers decision-makers to track performance trends, optimize future campaigns, and experiment with time, content, and segment strategies.

**CHAPTER 7**

**TESTING AND RESULTS**

A rigorous testing process was conducted to evaluate the performance of various machine learning models, with a focus on ensemble learning techniques—particularly the Random Forest Classifier—to predict user click behavior. This stage involved thorough evaluation, careful interpretation of results, and critical analysis to derive actionable business insights for Quantacus.Ai.

1. **Evaluation Metrics Used**

To accurately assess model performance,

**The following standard classification metrics were computed:**

**Accuracy :** Measures the overall correctness of predictions.

**Precision :** Indicates how many predicted positives were actual positives.

**Recall :** Measures how many actual positives were correctly predicted.

**F1 Score :** Harmonic mean of precision and recall; balances both metrics.

**ROC-AUC Score :** Threshold-independent metric that measures the model’s ability to discriminate between classes. Particularly useful for imbalanced datasets.

**2. Performance of Ensemble Model (Random Forest)**

After training multiple models, the Random Forest Classifier emerged as the most reliable performer. The model was evaluated on the test dataset (20% holdout). Below are the performance results:

**Metric:** Random Forest Result

**Accuracy:** 0.95

**Precision:** 0.81

**Recall:** 0.68

**F1 Score:** 0.74

**ROC-AUC Score:** 0.91

**3. Interpretation of Results**

High Accuracy (95%) shows that the model makes correct predictions most of the time.

Precision (81%) is strong, indicating the model is effective at minimizing false positives. This is crucial in campaign targeting—emails should go to users who are truly likely to click.

Recall (68%), while slightly lower than precision, demonstrates decent coverage of actual positive cases.

F1 Score (74%) confirms a good balance between precision and recall.

ROC-AUC (91%) is excellent, indicating that the model reliably separates clickers from non-clickers without depending on a specific threshold

**4. Critical Analysis and Practical Insights**

Why Ensemble Wins: The Random Forest model's ability to handle nonlinear relationships and feature interactions proved crucial in capturing subtle patterns in user behavior (e.g., personalization, time of send, country-based engagement).

Business Impact:

With the model, CTR improved from 2.12% to 6.17%—a 191% increase in performance when targeting the top 30% most likely users.

This improvement directly translates to better ROI for the campaign and more efficient use of email resources.

Strategic Takeaways:

Targeting high-value customers (users with ≥5 past purchases) boosts click probabilities.

Personalized content and timing (2 PM–5 PM on Tuesday–Thursday) significantly enhance response rates.

Geographic Segmentation: Users in regions like the US and Germany show consistently better engagement, suggesting value in geo-targeted strategies.

**5. Visual Evidence (From Tableau Dashboard)**

**The Tableau dashboard confirms key model findings:**

Wednesday had the highest click count.

Both long and short emails performed comparably, with a slight advantage to short ones.

Clear disparity between email opens (10,345) and link clicks (2,119) highlights the importance of predictive targeting.

**6. Limitations and Future Considerations**

**Imbalanced Data:** Although addressed using class weighting and stratified sampling, true positives (clickers) remain relatively rare, warranting continuous monitoring.

**User Behavior Drift:** Over time, user preferences and behaviors may shift. Periodic retraining and model refreshing are necessary.

**Cold Start Users**: New users with minimal historical data might need alternative heuristics or a hybrid model approach.

The testing phase validated that ensemble modeling—especially the Random Forest Classifier—offers a powerful and practical tool for email click prediction at Quantacus.Ai. Through robust evaluation and insightful interpretation, the solution proved its value in optimizing campaign targeting and boosting engagement with measurable ROI impact.from US and Canada responded more positively to email campaigns.

**CHAPTER 8**

**CONCLUSION**

The project, conducted utilizing the datasets from Quantacus.Ai, can be concluded and focused on optimizing an email marketing campaign using machine learning techniques to predict and improve click-through rates (CTR). The initiative combined structured data analysis, predictive modeling, and business intelligence visualization to provide a robust, data-driven solution. The analysis began with a thorough understanding of user interaction data spread across three main sources: the email\_table, email\_opened\_table, and link\_clicked\_table. Exploratory Data Analysis revealed important behavioral patterns—such as higher engagement on Wednesdays, better response to personalized emails, and significantly higher click rates among users with ≥5 past purchases.

A variety of machine learning models were trained, including Logistic Regression, Decision Tree, and the ensemble-based Random Forest Classifier. Among these, the Random Forest model demonstrated the best performance with a ROC-AUC score of 0.91, and solid precision-recall tradeoff, making it suitable for campaign targeting where false positives can be costly.

**Key takeaways from the model evaluation:**

CTR improved from 2.12% (baseline) to 6.17% when emails were sent to the top 30% of users based on model predictions.

This reflects a 191.17% uplift, underscoring the power of predictive targeting over random email distribution.

To support decision-making, a Tableau dashboard was developed, offering visual insights into engagement trends by weekday, email format, geography, and user behavior. This interactive tool empowers stakeholders to monitor and act on evolving trends in real time.

In conclusion, the project successfully demonstrated how machine learning, combined with thoughtful data preparation and visualization, can significantly enhance digital marketing effectiveness. By adopting the proposed predictive framework, Quantacus.Ai can transform its email campaign strategy—shifting from intuition-based decisions to data-driven targeting that delivers measurable improvements in user engagement and marketing ROI.

**CHAPTER 9**

**REFERENCES**

[1] Authors – Y. Satyam, J. Raviteja, D. Rithik Rao, R. Akhil

Journal – International Research Journal of Modernization in Engineering Technology and Science

Link- https://www.irjmets.com/uploadedfiles/paper/issue\_6\_june\_2022/26690/final/fin\_irjmets1655964131.pdf

[2] Authors – Dalia Saleh Abbas, Mustafa Al-Jailawi

Institution – Lund University, IKEA

https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=9174474&fileOId=9174481

[3] https://www.kaggle.com/datasets/loveall/email-campaign-management-for-sme/

[4]https://www.analyticsvidhya.com/blog/2022/05/solving-business-case-study-assignments-for- data-scientists/

[5] https://www.kaggle.com/datasets/swekerr/click-through-rate-prediction/