PREDICTIVE MODELING FOR EMAIL MARKETING SUCCESS: OPTIMIZING CAMPAIGN DELIVERABILITY AND GOOGLE POSTMASTERS METRICS VIA TREE-BASED REGRESSION

By

Diana Olivia Arva, Bachelors of Science in Bioinformatics

A thesis submitted to the Graduate Committee of
Ramapo College of New Jersey in partial fulfillment
of the requirements for the degree of
Master of Science in Data Science
Spring, 2025

Committee Members:

Dr. Debbie Yuster, Advisor

Dr. Osei Tweneboah, Reader

Jessica Kulenguski, Reader

Alex Marsek, Reader

McKenna Fuller, Reader

Dedication

To my parents, thank you for your unwavering support and encouragement throughout these past five years. I am deeply grateful for everything you all have done to help me reach this milestone, this achievement would not be possible without you guys!

I would also like to thank my boyfriend for being there and cheering me on every step of the way, and for the countless hours spent listening to me talk about this research. Hopefully you picked up a thing or two about email marketing along the way!

Acknowledgments

To begin, I would like to sincerely thank my advisor, Dr. Debbie Yuster, for your invaluable guidance and consistent feedback throughout this entire process. Your support provided a strong foundation that kept me motivated and focused during the development of this thesis!

I would also like to thank Dr. Osei Tweneboah for sharing your expertise in machine learning and for helping to clarify complex topics along the way!

Lastly, I am grateful and thank Jessica Kulenguski, Alex Marsek, Mckenna Fuller, and everyone on the DSBI team for serving as my readers and for their continued support and encouragement at every stage of this journey!

Table of Contents

List of Tables

List Of Figures

Abstract

Email remains the most effective and widely used channel to engage with consumers, typically utilized through coordinated email campaigns. These campaigns consist of a series of emails sent to a target audience with the intent of leading users to interact with a call to action (CTA). The CTA is clearly defined and placed within the body of the email, and interaction can take various different forms. Some include signing up on a website, clicking on a provided link or starting a subscription service.

However, email campaigns face two core issues, the first is that a significant portion of emails within these campaigns fail to reach the recipient's inbox. The second is that companies are also striving to seek ways to optimize how engaging their emails are for the user. The primary goals of the research are to improve email delivery rates, maximize inbox placement within the target audience and enhance user engagement metrics.

To address those goals, the research investigated two key objectives: first, identifying which email factors most directly correlate with delivery success, and second, analyzing how modifications within an email campaign rollout can improve both delivery success and engagement metrics. The research followed a structured workflow of data collection and dataset construction, predictive modeling, and segmented tracking A/B testing.

The first phase of the research involved collecting and constructing a dataset to analyze and predict the factors correlated with email deliverability. The dataset consisted of 19 distinct email campaign features across approximately 300 campaigns. All features were extracted and compiled from historical campaign data stored in a digital marketing company's internal database. Key features include metrics such as click-through rate (CTR), which measures the rate

of users who clicked on the CTA out of those who opened the email, as well as other engagement metrics, delivery event data and third-party metric tools like Google Postmasters.

Following dataset construction, two tree-based machine learning models, Decision Tree and Random Forest regressors, were trained to predict email delivery, with the primary goal of obtaining insights into the key influencing factors. Model optimization and evaluation were additionally performed using hyperparameter tuning and 5-fold and 10-fold cross validation. The models identified top critical factors, with high email send frequency contributing 50.51% and 53.80% of the predictive power influencing email delivery success, respectively.

Insights from the predictive modeling then informed the design of an A/B test, which evaluated whether historically low-engaging users and those with past delivery issues were associated with lower engagement metrics. The results confirmed the hypothesis that users with historical poor performance lowered engagement metrics, and that suppressing emails to these users may help preserve Google Postmasters metrics, increasing the likelihood that emails reach the user’s inbox rather than being filtered as spam or blocked entirely.

This research contributes to the evolving field of marketing optimization by demonstrating how predictive modeling and experimental testing can identify and address campaign inefficiencies in large-scale email campaigns. It also highlights future research frameworks, including sentiment analysis of email content and segmentation strategies to target users within respective demographics, with the ultimate target of enhancing both engagement and deliverability in digital email marketing strategies.

Chapter 1: Introduction

In our data-driven society, email has dominated how companies have been able to advertise and communicate effectively through digital marketing, and it is regarded as one of the most reliable advertising methods (Kanellopoulos, T, 2025). It provides a simple yet effective way to engage its audience through special offers, advertisements, or in building relationships with the targeted audience. Email marketing has become one of the most popular and effective ways to drive customer engagement, with businesses collectively investing around $3 billion each year (Thomas, J. S., Chen, C., & Iacobucci, D, 2022). Given the immense value of email marketing, businesses advertise through email campaigns, which deploy a coordinated set of emails specifically curated for a target audience to interact with a call to action (CTA). The CTA is displayed to grab the user's attention and to act on the email, whether it may be through signing up on a website or starting a subscription service. The campaign's goal is to ensure that emails not only reach the inbox, but engage the user as well.

However, successfully delivering emails has become increasingly challenging, as many consumers regard email advertisements as an unwanted distraction that offers little value. Specifically, email campaigns face two core issues: a significant portion of emails fail to reach the recipient's inbox, limiting the amount of exposure each campaign can achieve. Additionally, for those emails that are successfully delivered, companies strive to optimize how engaging the content is for users. These core issues are not unique to the campaigns examined by the marketing company in this research, but are shared throughout the email marketing industry as a whole. In order to enhance and improve user engagement through bulk email sending, companies implement practical strategies using insights gleaned from mining historical data to tailor emails to include audience segmentation, targeted content creation, and optimal sending patterns.

To address these core issues, the research investigates two main research objectives: first, identifying which email campaign features most directly correlate with delivery success, and second, analyzing how modifications in campaign design can improve both delivery success and engagement metrics. These objectives aim to support three key primary goals: specifically in improving email delivery rates, maximizing inbox placement within the target audience, and enhancing engagement metrics. The features used to construct the dataset were extracted and compiled from historical campaign data stored in a digital marketing company's internal database. The email campaigns analyzed in this research comprised a multitude of different industries, including newsletters, health, financial, and promotional emails. Additionally, the campaigns varied significantly in volume, ranging from approximately five hundred email sends to over one million for the top performing campaigns. Decision Tree and Random Forest regressors were implemented using Python, along with model assessment methods such as k-fold cross-validation and feature importance analysis, to identify the key factors influencing the successful delivery of emails. Lastly, a segmented A/B test was conducted using insights derived from the regression models to evaluate whether users with low historical engagement and past delivery issues negatively affect engagement metrics, and that potential suppression of emails to these users may help preserve Google Postmasters metrics, thereby increasing the likelihood of successful inbox placement.

Throughout this research paper, I plan to investigate the growing area of bulk email sending through email campaigns, and provide an analysis of areas that negatively impact user engagement and experience. This groundwork will provide businesses with an enhanced understanding of bulk email optimization, and how to maximize user inbox placement within their advertising campaigns. Additionally, it will support the development of future marketing strategies aimed within any domain-related advertising space, allowing improved audience retention within their call to actions (CTA), as well as increasing brand awareness and ensuring a consistent return on investment.

Chapter 2: Background

2.1 Evolution of Email & its Role in Marketing

The first electronic message through email was sent in the 1970's by the American programmer Ray Tomlinson, widely regarded as the father of email. His invention was the introduction of the "@" symbol, which was used to separate the user from the domain name, allowing communication to be sent and received from two different computers (Taylor, J., 2024). This early form of digital communication was sent over ARPANET (Advanced Research Projects Agency Network), which was a precursor to the internet (Burtle, L., Head, S., & Lankford , S., 2013). Gary Thuerk was one of the first to experiment with email as a marketing tool by sending a mass email advertisement to around four hundred ARPANET users. It generated a shocking 13 million dollars in sales in 1978, proving early on that mass email sending is profitable (Church, C., 2023).

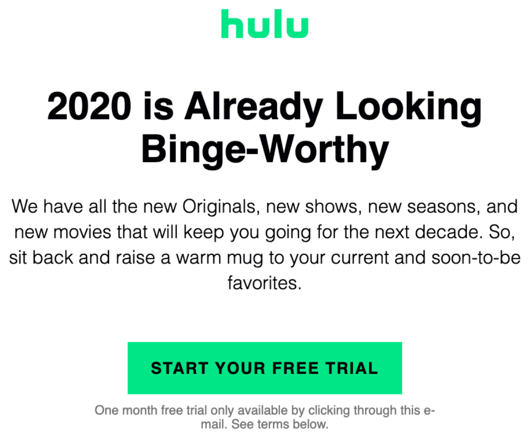
The early 1980s saw the introduction of the Domain Name System (DNS), which allowed the current state of email to become what it is today (A Digicert Company, 2024). In the context of emails, DNS provides security measures, such as DomainKeys Identified Mail (DKIM) and Sender Policy Frameworks (SPF), which third-party bulk email tools such as Google Postmasters provides to ensure the security and reliability of email transfers (A Digicert Company, 2024). The DNS system allows email sending to be accurate and secure for all users, enhancing its credibility as a reliable communication method and further supporting its growth.

2.2 Current State of Email Marketing

The 1990s marked the beginning of modern email marketing, partly due to the development and widespread use of several technological advancements including HTML, the popularization of internet service providers (ISPs), and the emergence of various email platforms (Taylor, J., 2024). The inventions and advancements in email technology from Tomlinson, Thuerk, and others have shaped email strategies by introducing strategically chosen phrases, images, and incorporating call to action (CTAs) within the email content.

Today, businesses primarily advertise through email campaigns, where an email campaign refers to a coordinated set of emails sent to a target audience with a clear objective: to engage users through call to actions, or CTAs. A well placed and crafted CTA can lead the recipient to act on the email through purchasing, signing up for a newsletter, or clicking on a provided link (Duarte, N., 2024). For example, Figure 1 shows an advertisement from the streaming platform Hulu. The flow of the email begins with an attention-grabbing sentence that is designed to intrigue the user. It then highlights a catchy proposition of Hull offering "new shows, new seasons, and new movies that will keep you going…". Finally, a clearly displayed green box serves as the call to action, prompting the user to start a free trial.

Figure 1. *Example Of A CTA Within an Email Advertisement*



hulu

2020 is Already Looking Binge-Worthy

We have all the new Originals, new shows, new seasons, and new movies that will keep you going for the next decade. So, sit back and raise a warm mug to your current and soon-to-be favorites.

START YOUR FREE TRIAL

One month free trial only available by clicking through this e-mail. See terms below.

For businesses to have the maximum return on investment, they employ email campaigns that also personalize their email lists by features such as demographics, purchasing behavior, or engagement levels (Kaddipudi, M., 2021). Customers who receive emails containing advertisements relevant to them are more likely to engage with the content, thus improving core metrics. Metrics include key performance indicators (KPIs) such as open rates, click-through rates (CTR), bounce rates, and unsubscribe rates, among many others. These are widely used among businesses that bulk send emails through campaigns to advertise.

2.3 Google Postmasters to Measure Email Campaign Success

Organizations establish teams which are tasked to analyze historical data to identify email messaging strategies that help drive engagement. Other than KPI's, companies use third-party metrics tools such as Google Postmasters, which provides insights into email performance. Google Postmasters provides metrics on email campaigns including domain and IP reputation, user-reported spam rate, and security metrics (Ferguson, A., & Hartmann, M., 2024). An email domain refers to the email address section after the @ symbol. They specifically examine the domain of the sender email, or the “from” email address that sends the advertisement into the recipient's inbox. These email domains must comply with authentication protocols such as DKIM, SPF, and DMARC (domain-based message authentication, reporting, and conformance) to verify the legitimacy of each domain (Dmarc, 2025). The metrics reported by Google Postmasters allow teams to further analyze their email campaigns and marketing strategies, specifically in an effort to reduce spam, as a lower reputation results in reduced odds of successful email delivery. A stronger reputation indicates that the emails are trustworthy, increasing the chances of landing in the inbox rather than the spam folder. The analysis conducted in the following chapters provides valuable insights into why it is crucial for email campaigns to target users that can provide high engagement, in order to avoid the emails being flagged and to maintain recognition as a reputable domain.

2.4. Cycle of a Marketing Email within an Email Campaign

A marketing email campaign typically launches a large volume of emails to a targeted group of users, or recipients. The volume can vary depending on each campaign, ranging from a few hundred emails to several millions. Before an email can reach its intended recipient, it must first pass through various authentication and security protocols. These protocols are managed by both ESP (email service provider) systems as well as external tools such as Google Postmasters. While Postmasters provides insights into email performance and sending reputation, one of its key roles is in ensuring that any emails that travel and land in a user's inbox are secure and safe from any potential spam or malicious content. Based on insights from Postmasters and other indicators such as previous spam rate or reputation, the recipient's email service provider will determine whether to either accept or reject the email. Rejection can occur in two ways: by completely blocking the email, or having the email get sent into the user's spam or junk folder.

If the email is successfully delivered to the inbox, the recipient then has the opportunity to engage with it, which includes opening the email, clicking on any placed call to actions (CTA), or choosing to unsubscribe. Each of these interactions is tracked and recorded in the digital marketing company's database. Performing analysis on the data can help evaluate the effectiveness of each email campaign, and determine user engagement and whether the emails are reaching the inbox as intended. The data additionally allows investigation in how to further refine campaigns to improve deliverability and overall performance.

Chapter 3: Methodology

This chapter outlines the methods used to collect, clean, and transform the email campaign data to ensure it is suitable for use in the following tree-based regression models. The results from these machine learning models will be then utilized in segmented A/B testing to examine whether changes to an email campaign improves performance. The data collection involved extracting campaign-related data tables using SQL queries in Snowflake, a data warehousing platform provided by the digital marketing company (Snowflake Documentation, 2025). To ensure data reliability, SQL Looker was used to verify the integrity of the collected data. Following data collection, this chapter provides a detailed description of the data schemas used, which helps to further contextualize the features involved in the research. Data cleaning and preprocessing methods, as well as model optimization methods for tree-based regression models were additionally discussed. The remaining sections of the chapter discuss any limitations encountered during the data collection phase and address any ethical considerations relevant to using the dataset.

3.1 Study Design and Data Collection

This study utilized a correlational research design, which was chosen to explore which characteristics of an email campaign are associated with success. Success is measured by the rate at which emails are successfully delivered to a user’s inbox. The following sections begin by introducing key industry standard metrics and variables that were used throughout the research. It will then explain the importance of Google Postmasters Tools, specifically domain reputation, and its role in improving the likelihood of email deliverability within campaigns. Additionally, it provides an overview into the collection of features included in the construction of both the send-level and campaign-level datasets.

3.2 Key Metrics and Variables

While the following datasets contain numerous variables, this research primarily focuses on a subset of key metrics and features most closely aligned with the research objectives. Table 1 highlights the key variables analyzed throughout the research, with click-through rate (CTR) identified as the most important. CTR represents the percentage of users who click on an advertisement, specifically the call to action (CTA), after opening the email. A higher CTR rate indicates that the advertisement is effective, relevant, and engaging to the target audience. The dataset also includes engagement metrics, which measure the percentage of users who take specific actions with the email. It also contains data on advertising reach for each email campaign, measuring the campaign's reach and sending patterns per user.

Table 1. *Table of Key Metrics & Variables*

Metric and Variables	Description
Click-through rate (CTR)	Represents the percentage of recipients who clicked on a CTA out of those who opened the email.
Open rate	Represents the percentage of recipients who opened the email out of the total delivered.
Unsubscribe rate	Represents the percentage of recipients who opted out after opening the email.
Bounce Rate	Represents the percentage of users who click on a CTA and visit a third-party website, but navigate away after viewing a single page.

CTR is typically analyzed alongside other specific engagement metrics such as open rate, unsubscribe rate and bounce rate. Open rate refers to the percentage of recipients who opened an email out of the total delivered. Similarly, the unsubscribe rate represents the proportion of users who opted out after opening the email. Lastly, bounce rate refers to the percentage of users who click on the CTA and visit a third-party website but navigate away after viewing only a single page.

Advertising reach, on the other hand, measures how extensively a campaign is distributed to its audience. This includes the number of distinct users who interacted within each campaign and the number of emails sent per each unique user. These metrics are important in identifying repetitive patterns, since there are instances where users receive or view the same advertisement several times. These features can inform decisions about email send frequency and targeting strategies. The key metrics and features were particularly selected as they are widely recognized industry standards for marketing companies as a whole. They allow further performance evaluation and engagement analysis into how the email campaigns are performing, as well as supporting comparison across different campaigns in terms of engagement growth or decline.

3.3 Google Postmasters

Due to one of the key objectives in investigating delivery success within the email campaigns, third-party metric tools such as Google Postmasters Tools, allows companies to monitor and improve the deliverability of emails to users. The platform provides an abundance of information (Table 2), with domain reputation being the most important when evaluating delivery outcomes in this research. The company’s internal database contained Google

Postmasters information for only a subset of emails, and the research utilized only those emails with valid Postmaster information.

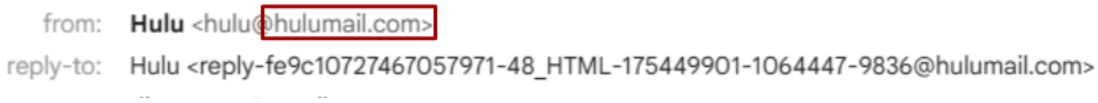
Table 2. *Key Metrics Provided by Google Postmasters*

Metric	Definition
Domain Reputation	Indicates how trustworthy the sender's domain is. A higher domain reputation reduces the likelihood of emails being filtered as spam or rejected.
IP Reputation	Assesses the trustworthiness of the sending IP address. A higher IP reputation increases the chances of emails reaching the recipient's inbox.
User Reported Spam Rate	The percentage of recipients who manually report emails from the sender as spam after receiving them in their inbox.
Security Authentication Protocols	Measures whether the sender's domain passes authentication checks like SPF, DKIM, and DMARC, which help verify the legitimacy of the senders.

Note: Definitions of Google Postmasters metrics were sourced from Ferguson, A., & Hartmann, M. (2024, January 20). Google Postmaster Tools—What It's and How It Can Help You. Microsoft.

Domain reputation is based on email domains, which refer to the portion of an email address that comes after the "@" symbol. Google Postmasters specifically examines the sender domain, or the "From" address used to send the email to the recipient. This is the section highlighted in red in Figure 2, and it is what Postmasters evaluates to assign a reputation score. Continuing with the example of the Hulu advertisement, *hulumail.com* is the domain that Postmasters will use to assess its reputation value.

Figure 2. *Example Of A Sender Domain*



from: **Hulu** <hulu@hulumail.com>
reply-to: Hulu <reply-fe9c10727467057971-48_HTML-175449901-1064447-9836@hulumail.com>

The domain reputation metric is categorical, and consists of four distinct values: Bad, Low, Medium and High. A High reputation value signals to Google Postmasters that the sender's emails are trustworthy, increasing the likelihood that the email reaches the recipient's inbox rather than being flagged as spam and filtered into the spam folder (Table 3). On the other hand, both Bad and Low reputations values indicate that emails sent from these sender domains have inconsistent sending practices, low engagement, or a history of being reported as spam by past recipients. Domains with a Medium reputation value, while it is more favorable than scoring either Low or Bad, still have a moderate chance of being flagged and filtered into the spam inbox rather than in the recipients inbox, reducing deliverability and limiting chances of further user engagement.

Table 3. *Definitions of Domain Reputation Values Scores*

Value	Definition
Bad	A history of sending extremely high volumes of emails or spam. Emails likely to be rejected or marked as spam.
Low	Known to send a considerable amount of emails. Emails are typically sent into the spam folder, with few rejected.
Medium	Generally sends reputable emails, though some are occasionally flagged as spam. These senders exhibit moderate deliverability and engagement rates.
High	Maintains a strong track record of low spam rates and high engagement. Compiles with Google Postmasters sender guidelines.

Note: Definitions of Google Postmasters metrics were sourced from Ferguson, A., & Hartmann, M. (2024, January 20). Google Postmaster Tools—What It's and How It Can Help You. Microsoft.

Scoring the sender's domain is done in an attempt to reduce spam and fraudulent emails, as lower reputation results in lower odds of email delivery. Google Postmasters examines several factors relating to each sender domain other than domain reputation, such as whether the email matches security protocols (SPF, DKIM, DMARC), user engagement with the emails, and user reported spam rate. Due to this, it is crucial that email campaigns send targeted and trustworthy emails that can comply with all of these factors, in order to prevent being flagged and demarcated as a reputable domain.

Figure 3. *Distribution of Domain Reputation Categories*

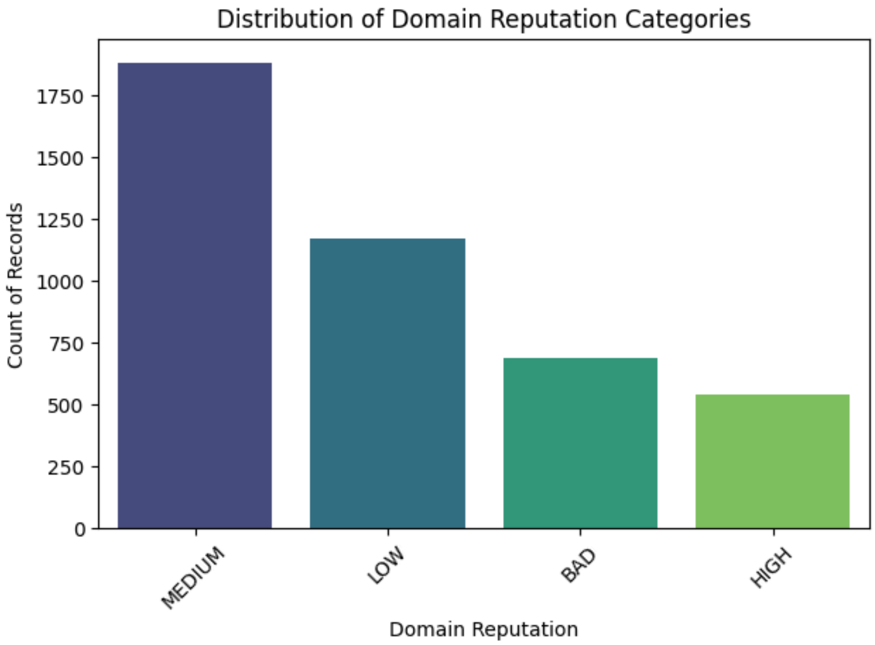


Figure 3 shows the percentage distribution of the domain reputation values across the dataset utilized in the subsequent sections of this chapter. To start, approximately 43% of the sender domains fall into the Medium category, suggesting that the domains have reasonable deliverability and user engagement. However, they are still occasionally flagged as potential spam and are either landing in the user's inbox or in the spam folder. Because of this, domains with a Medium reputation score should be viewed as a warning sign, since they are at risk of falling into the Low category. Around 27% of domains fall into the Low category, which represents a significant portion of senders whose emails are frequently flagged as potential spam and are unlikely to reach the recipient's inbox. Additionally, 16% of the sender domains are rated Bad, suggesting that these emails have an even greater chance of risk of being rejected altogether or sent to the spam folders. All together, the Bad, Low and Medium categories account for nearly 83% of all sender domain reputation scores within the dataset, with only 13% achieving a High reputation category.

The low distribution of domains rated High highlights the importance of improving domain reputation, which allows a significant increase in the chance of emails being successfully delivered to the recipient’s inbox rather than being filtered into the spam folder or rejected altogether from sending to the recipient. While the Medium value is the most common domain reputation scores, it reinforces the idea that most domains in the dataset are underperforming and are at risk of reduced email visibility by lowering the amount of successfully delivered emails in the inbox.

3.4 Dataset Collection & Construction

This section provides an overview of the two datasets constructed during the research. The campaign-level dataset was utilized to answer the first research objective, which was in identifying key factors that correlate with successful email delivery within a campaign. Both datasets were constructed using a one-month timeframe between 1/12/25 to 2/11/25, and were extracted from five internal tables of historical email data from the digital company's internal database. This section also includes a detailed analysis into the five source tables used within the data collection and construction process.

3.4.1 Send-level Dataset

The send-level dataset was constructed using a purposive (judgemental) sampling technique, which involved extracting email-level data, where each observation represents a single occurrence of an email sent to the inbox of a targeted user (Table 4). The same five tables used in the send-level dataset construction were subsequently used in the construction of the campaign-level dataset.

Table 4. *Table of Email-level Variables & Metrics*

Table of Email Variables & Metrics			
Key Performance Indicators (KPI's)			
● Send Date* ● Deploy ID* ● ISP (Internet Service Provider)* ● Total Delivered ● Total Failed	● Total Sends ● Total Clicks ● Total Opens ● Cached Opens* ● Total Distinct Users	● Open Rate ● Delivered Open Rate* ● Cached Open Rate ● Total Unsubscribes ● Unsubscribe Rate	● Delivered Rate ● Failed Rate ● Delivered Click Rate* ● Delivered Unsubscribe Rate* ● Click-through Rate (CTR)
Delivery Events ● UUID* ● Contact* ● ESP (Email Service Provider)* ● Event* ● Error Code* ● Error Category	**Email Info** ● Sender Email* ● Signal Type ● Offer Category ● Sends per Unique Clicker	**Campaign Info** ● Campaign Name	**Google Postmasters** ● Security Protocols (Binary)* ● Domain Reputation ● IP Reputation ● User Reported Spam Rate*

Note: Variables with an * were removed from the campaign-level dataset due to their granularity at the individual email level or variable was not needed anymore.

The send-level dataset contained a total of 35 distinct features, the majority of which were quantitative metrics as shown in the KPI table. In addition to these metrics, the KPI table included fields such as the send date in which the email was sent to the recipient, and the Deploy ID, a unique identifier tag that represents each email sent. It is used to link information from multiple data tables for the same email, with each email in the database having a unique Deploy

ID. ISP (Internet Service Provider) provides additional context into the specific provider utilized for each sent email. Both of the Delivered and Failed metrics are binary variables, populated if the email was successfully sent or not. To note, if future research involves extensive use of the send-level dataset, one of these variables could be removed to improve efficiency without data loss. Metrics such as opens, clicks and unsubscribes were populated with a value of 1 when an email was successfully delivered and the recipient took an action, such as opening the email, clicking a call to action (CTA), or unsubscribing.

In addition, both Open Rate and Cached Open Rate were included in the dataset. The cached open rate refers to instances where Gmail “pre-loads” or “pre-opens” the recipient’s email before the user actually opens it. Services like Gmail often do this to offer a more enhanced user experience by providing faster loading times or previews. However, this behavior may introduce false positives when companies try to measure how often users are genuinely interacting by opening their delivered email. In short, cached open rate refers to pre-processed opens, whereas open rate captures actual user engagement.

The delivery events table extracted information specific to email deliverability, which was one of the key objectives within the research to investigate. While the Deploy ID links to a specific email, the UUID (Universally Unique Identifier) links to the recipient’s email address from which the email was sent to. This allows the email address to be linked with other data tables to gather additional information on the recipient. The Contact field refers to the recipient's actual email address but was removed for privacy reasons, as discussed in Section 3.13. Since each email is still linked by UUID, the Contact name’s removal did not affect the data’s integrity.

ESP provides additional context as to what email service provider was used to send the email. Similar to the Delivered and Failed variables, the Event field indicates whether the email reached the inbox or was diverted to the spam folder. For example, an email sent to a spam folder is considered a failure. Error Code and Error Category serve similar purposes with the Error Category indicating the reason for delivery failure. Table 5 lists potential error categories that may occur when an email fails to be delivered, covering a variety of issues, including a full recipient inbox, excessively high email send frequency or rejection as spam.

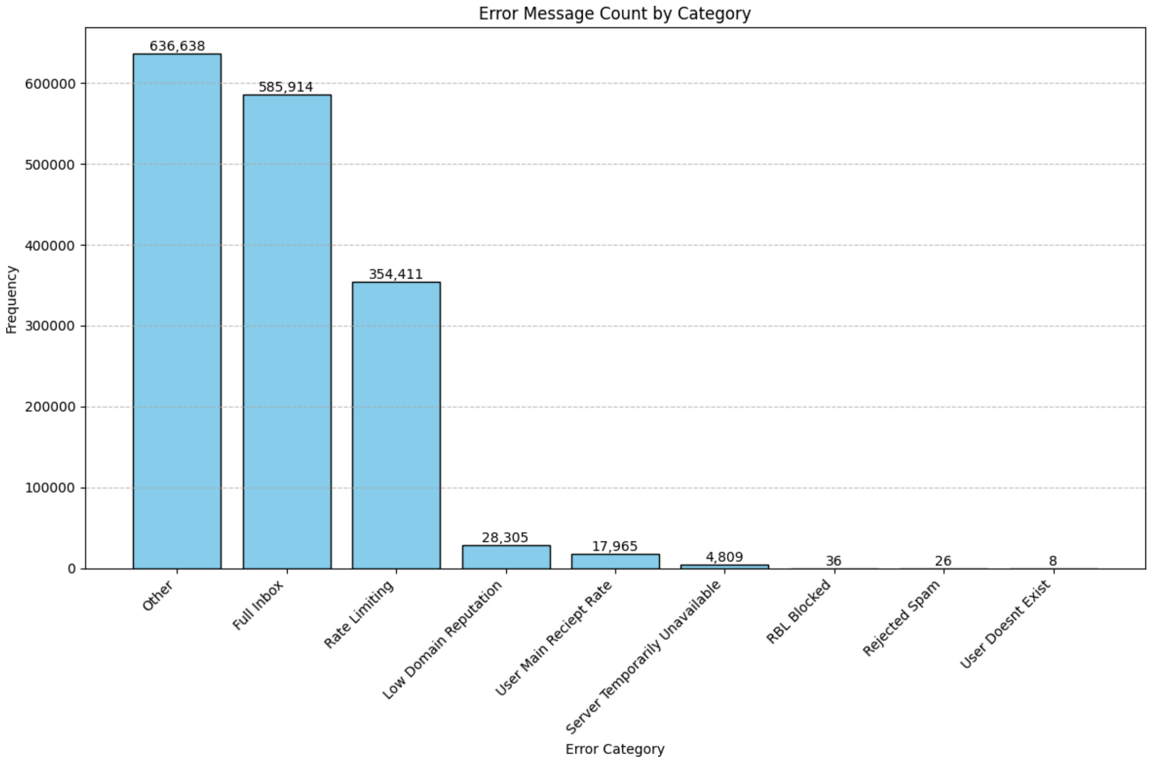
Table 5. *Common Error Category from Failed Email Sends*

Error Name	Description
User Mail Receipt Rate	Targets the mail server overall, often as a result of high send frequency, users are receiving emails at a rate that exceeds the recipient's server limit.
Rate Limiting	Targets individual recipients, often as a result of high send frequency, users are receiving emails at a rate that exceeds server limits.
Full Inbox	Indicates the recipient’s inbox is full and cannot accept new emails.
Rejected Spam	Rejected by the recipient server due to emails identified as spam.
Low Domain Reputation	Sender domain contains a poor reputation, and emails are rejected.
RBL Blocked	Sender IP address listed on an RBL list (Real-time Blackhole List), indicating possible spam activity.
Failed DKIM Authentication	Security protocol, email could not be verified as secure by the receiving email server.
Server Temporarily Unavailable	Receiving server is unavailable and cannot receive delivered emails.
Other	Any other error category not included within the categories mentioned above.
No Error	Indicates the email was successfully delivered to the recipient’s inbox.

The “Other” error category includes all failed emails that did not fall into the defined error categories listed above. These were grouped into a single category because they were not directly relevant to the purpose of the research objectives. This category encompassed a variety of less common or less relevant email failures, including internal system errors, potential third-party add-ons blocking the delivery of emails, and issues related to formatting or attachment policies such as image handling. These errors were all grouped into one category to maintain focus on the error categories that most directly aligned with the research objectives.

Besides the “other” error category, recipients with full inboxes were the source for a significant portion of rejected email attempts, totaling around 580,000 (Figure 4). Other common error categories included high email send frequency and a low domain reputation score, with approximately 350,000 and 28,000 instances, respectively. These high counts indicate key areas that could cause issues in email send reliability.

Figure 4. *Error Category Count by Category*



Note: the “no error” category was excluded from Figure 4 to focus exclusively on error categories associated with failed email sends.

Additional email-related features were included in the send-level dataset, such as the Sender Email. As discussed in Section 3.3, this metric is relevant because Google Postmasters evaluates the sender domain’s reputation. However, the feature was removed from the campaign-level dataset, as Postmasters metrics alone were deemed sufficient. Features such as Signal Type, Offer Category, and audience reach metric like Sends per Unique Clicker were explained further in detail in the following section on campaign-level data collection. The campaign-level dataset provided the Campaign Brand Name that links each email to its respective campaign and was used to aggregate the dataset at the campaign-level level for the following analysis. Finally, data from Google Postmasters information was additionally incorporated, with Security Protocols stored as binary variables, with a value of 1 indicating the email passed all necessary security checks.

Table 6. *Key Features of the Email Send-Level Dataset*

Deploy ID	ISP	Signal Type	Offer Category	Domain Reputation	Error Category	Event
—	Gmail	Other	Legal	LOW	other	Delivered
—	Gmail	Signups	Jobs	BAD	rate_limiting	Delivered
—	Yahoo	Signups	Finance	LOW	other	Failed
—	Gmail	Clicks	Loans	LOW	full_inbox	Delivered

Note: Values in the Deploy ID column have been hidden to maintain confidentiality; however the column is retained to represent the unique identifier for each observation in the dataset.

Table 6 provides a subset of the send level dataset, which ultimately consisted of a total of 35 distinct features, the majority of which were quantitative KPIs.

3.4.2 Campaign-level Dataset

The campaign-level data was then aggregated so that each observation within the dataset represented a unique email campaign rather than individual email sends. SQL was used to compile all necessary metrics and variables, with all the features drawn from the same five distinct tables provided by the internal company's database (Table 7).

Table 7. *Table of Campaign-level Variables & Metrics*

<table>
<tr><th colspan="4">Table of Campaign Variables & Metrics</th></tr>
<tr><td colspan="4"><b>Key Performance Indicators (KPI’s)</b></td></tr>
<tr><td>● Total Delivered*
● Total Failed*
● Total Sends
● Total Clicks</td><td>● Total Opens
● Total Unsubscribes
● Click-through rate (CTR)</td><td>● Total Distinct Users
● Delivered Rate*
● Failed Rate*</td><td>● Cached Open Rate
● Open Rate
● Unsubscribe Rate</td></tr>
<tr><td><b>Delivery Events</b>
● Error Category
(for failed email sends)</td><td><b>Email Info</b>
● Signal Type
● Offer Category
● Sends per Unique Clicker</td><td><b>Campaign Info</b>
● Campaign Name</td><td><b>Google Postmasters</b>
● Domain Reputation
● IP Reputation</td></tr>
</table>

Note: Variables with an * were removed from the predictive modeling dataset to prevent overfitting but were included in other areas of the research.

A total of 20 features were extracted when constructing the dataset, which was used to address the first research objective, which is identifying factors that correlate with the successful delivery of emails within a campaign. The key performance indicators (KPI) table contained the majority of the relevant metrics, including notable ones mentioned previously such as CTR, open rate and unsubscribe rate. Total counts of user engagement and performance were also included, such as the number of total email sends, delivered and failed emails, opens, clicks and unsubscribes.

Other tables included email delivery events, which contained the error category for failed email sends within a campaign. Additional email information was included in the campaign-level dataset, such as the signal type of the campaign, whether the intended action was to sign up or click through the email. The offer category of the advertisement was also included, such as identifying whether the campaign was health-related, financial or promotional. Lastly, advertising reach was included as well, such as the number of emails sent to each unique user. The campaign table was included to aggregate the send-level information into campaign-level, and lastly, the Postmasters table was used to add domain reputation and IP reputation.

The campaign-level data now included 309 unique observations across 20 distinct features, forming the basis for the subsequent analysis in this research. Table 8 displays a subset of the data, showcasing key features such as CTR and open rate used for the model development. Google Looker, a SQL-based business intelligence and analytics tool, was used to ensure that the extracted metrics aligned with the company's internal reporting standards (Google Cloud, 2025). It was employed throughout the development of the query to facilitate validation and consistency of the metrics.

Table 8. *Key Features of the Email Campaign-Level Dataset*

Campaign Name	Total Sends	Total Distinct Users	CTR	Open Rate	Error Category
Campaign 1	2,306,172	39,319	1.77	0.16	No Error
Campaign 2	682	105	11.12	0.11	User Mail Receipt Rate
Campaign 3	271	44	2.21	0.06	Rate Limiting
Campaign 4	5,870	506	0.95	0.17	No Error
Campaign 5	19,640	174	0.87	0.16	No Error

Note: Campaign names have been anonymized for confidentiality purposes.

The aggregation of the campaign-level dataset now allowed for the use of key metrics relative to each specific campaign. The send-level dataset could not support these metrics due to its high granularity. Additionally, for all categorical variables such as domain reputation and error categories, the most frequent values within each campaign were used to represent that respective category. For example, Table 8 shows that rate limiting is the most frequent error in Campaign 3, which is why it is represented as that corresponding value for that campaign.

3.5 Comprehensive Analysis of Data Tables

The table below provides an in-depth summary of all relevant data tables mentioned above that were utilized in the collection and construction of the dataset, including their respective uses for the research objectives:

Table 9. *Comprehensive Overview Of Data Tables Utilized*

Table Name	Description	Use for Research Objective
Google Postmasters Table	Stores domain-level email deliverability and reputation metrics sourced from Postmasters Tools.	Provides key information such as domain and IP reputation, and user reported spam rate, which directly impact deliverability rates.
Key Performance Indicators (KPI) Table	Contains engagement metrics per campaign, such as CTR, open rate and unsubscribe rate, aggregated from send-level data.	Offers key performance and engagement variables used to identify factors correlated with successful email delivery.
Email Delivery Events Table	Records final delivery outcomes for all emails and error categories for email delivery failures.	Identifies the type and frequencies of delivery failures, helping to diagnose the cause of delivery issues.
Campaign Level Table	Includes information such as campaign brand name, aggregates data to the campaign level.	Allows the analysis to be performed at the campaign-level rather than the individual email level.
Email Related Variables Table	Contains additional campaign details, such as the signal type and offer category of the email.	Included to further contextualize the type of each campaign for added information.

3.6 Data Cleaning

To start, out of the 20 distinct features in the campaign-level dataset, the campaign names and sends per unique clicker contained null values, accounting for 1.5% and 60.9% of the entire column, respectively (Table 10).

Table 10. *Percentage of Null/Missing Values within Dataset's Variables*

Campaign Name	Sends per Unique Clicker
1.50%	60.90%

For the campaign name variable, only observations with missing values were removed from the dataset due to the relatively low number of missing values. However, the sends per unique clicker column contained null values in more than half of the observations. As a result, the entire column was removed from the dataset.

Since the primary goal of the tree-based regression models was to identify features correlated with success, the target variable was calculated as the percentage of total emails within each campaign that were successfully delivered to a user's inbox. The calculation involved dividing the number of delivered emails by the total amount of sent emails, and multiplying it by 100 to convert it to a percentage (Figure 5).

Figure 5. *Calculation of Successfully Delivered Emails in Campaigns (Target Variable)*

$$\begin{array}{c}\text{Target}\\\text{Variable}\end{array} = \frac{\text{\# of successfully delivered emails}}{\text{\# of total emails sent}} \ \text{x} \ 100\%$$

Since the target variable was a continuous numerical value, regression-based machine learning models were used in the research. The dataset consisted of a wide range of campaign volume sizes, ranging from several hundred to over a million sends. In order to ensure that the campaigns were actively utilized within the company, a minimum requirement of at least 250 sends per email campaign was applied, and any campaigns that did not meet the threshold were removed from the dataset.

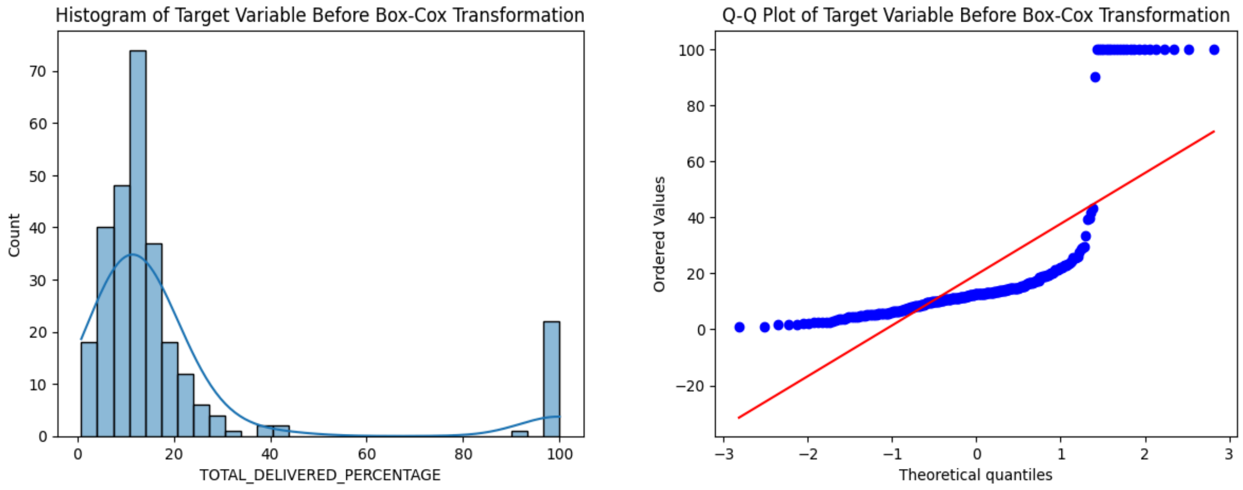
Other data preprocessing steps involved analyzing and correcting the datatypes of each variable to reflect the actual attribute types. The data frame contained two categorical features: error category and campaign name. In order to make them suitable for the following regression models, the error category was one-hot encoded to convert into a binary format. However in regards to campaign name, due to having 285 distinct campaigns, utilizing one hot encoding would significantly increase the dataset's dimensionality and computational complexity of the models. To prevent this, label encoding was instead applied, which converted the brand names into numerical values, where each number uniquely identified a campaign name. Lastly, to ensure that leakage did not occur within the regression models, any features related to the target variable, including the total counts of delivered and failed emails, along with email delivered and failed rates, were removed from the dataset but were still included in exploratory data analysis (EDA).

3.7 Data Transformation

As mentioned previously, due to the extensive range of email volume within each campaign, Figure 6a depicts the target variable within the research, which reveals a heavily right-skewed distribution. Most of the campaigns had a total delivery rate clustered between 9%

to 18%, with a small number of outliers displaying unusually high rates between 90% and 100%. Utilizing this skewed target variable could potentially cause the regression models to become biased or generalize poorly to the dataset. A Q-Q plot (Quantile-Quantile plot), shown in Figure 6b, was also graphed to visualize whether the data values within the target variable closely follow the line of best fit, indicated by the red diagonal line.

Figure 6a & 6b. *Histogram and Q-Q Plot of Target Variable Before Box-Cox Transformation*



a) **b)**

Before the transformation, the data points exhibited characteristic patterns of an exponential distribution, further confirming that the target variable is right-skewed. A Box-Cox Transformation was performed to stabilize variance and improve normality for the following models. In order to implement the transformation, the values must be positive and the data should be continuous, both of which were met by the target variable (Rossiter, D. G., 2019). Figure 7 presents the equation for Box-Cox Transformation, where *x* represents the original variable, and lambda (λ) is a parameter that is used to normalize the data based on its value.

When λ is either 0 or approximately 0, the Box-Cox equation performs a natural logarithmic transformation. For values other than 0, the equation applies a power transformation.

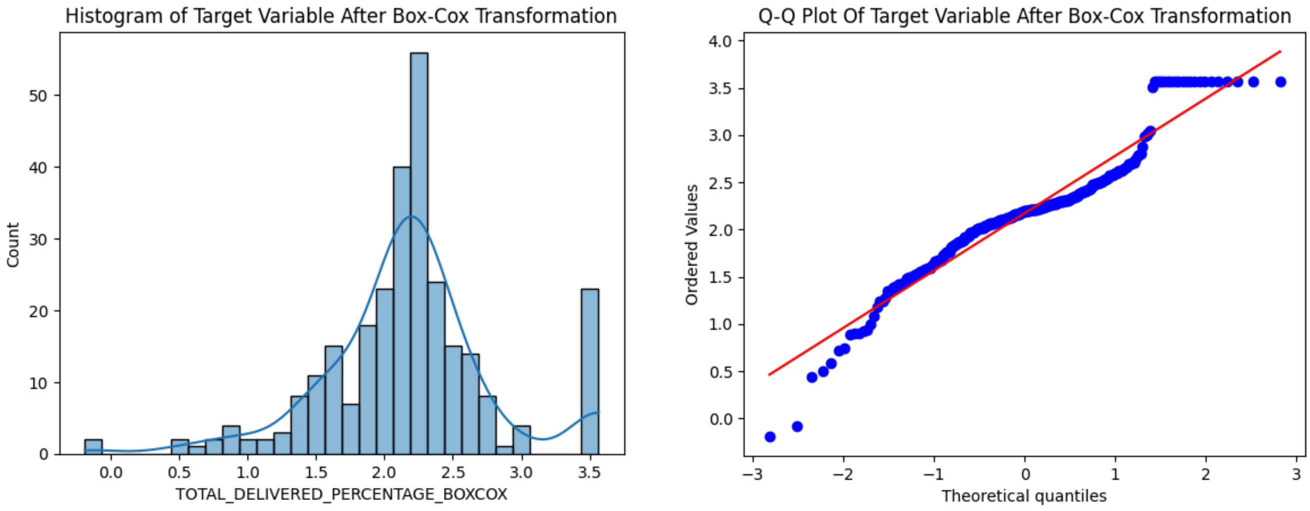
Figure 7. *Box-Cox Transformation Equation*

$$x(\lambda) = \frac{(x^{\lambda} - 1)}{\lambda} \, for \; \lambda \neq 0$$

$$x(\lambda) = \ln(x) \, for \; \lambda = 0$$

The Scipy Statistics library was utilized to calculate the optimal lambda value, which resulted in a lambda (λ) value of -0.116. Given that the resulting value is λ ≈ 0, the Box-Cox Transformation approximates a natural logarithm, as shown in the bottom equation in Figure 7. After the transformation, the x-axis scaling was adjusted and now depicts the data points to appear more normally distributed, resembling a Gaussian distribution (Figure 8a). Furthermore, the Q-Q plot (Figure 8b) indicates that the data points now align more closely along the line of best fit. While some irregularities remain within the target variable, the transformation significantly improved the skewness of the data values. However, it is important to note that the subsequent predictive modeling in this research does not require normalization, as tree-based machine learning models are not sensitive to factors such as skewed distributions.

Figure 8a & 8b. *Histogram and Q-Q Plot of Target Variable After Box-Cox Transformation*



a) **b)**

After data cleaning, filtering and transformation, the fully processed campaign-level dataset now contained a total of 19 features and 285 distinct campaigns which will be used in the following sections of the research.

3.8 Hyperparameter Tuning

Due to the heavily skewed nature of the dataset resulting from varying campaign volumes, hyperparameter tuning was implemented to optimize both machine learning regression models. More specifically, GridSearchCV (Grid Search with Cross-Validation) from the Scikit-learn library was utilized, as it tests and finds the best possible combinations of hyperparameters to yield the best model performance (Hyperparameter Tuning Using Gridsearchcv, 2020). The values used within the grid search were combined and calculated using k-fold cross validation, with the final parameters presented in Section 5.1 and 5.2 of the research.

3.9 Tree-based Regression Model

The research utilized two machine learning models: Decision Tree and Random Forest Regressor. Given the research’s continuous target variable, regression-based tree models were utilized and evaluated using common statistical metrics such as mean-squared error (MSE), root mean squared error (RMSE) and R^2 (coefficient of determination) to measure the proportion of variance explained by the models. The regression trees constructed predictions based on the MSE, using the value to select optimal splitting criteria during the process of the tree construction. The splitting criteria iterates over all features and a specific threshold to find the most accurate tree prediction. Figure 9 provides the splitting criteria utilized within the research’s regression based tree models, where *n* is the number of data points within each node, Y_i is the true target value for observation *i*, and $\hat{Y}_i$ is the predicted value for observation *i* (Farshad, K., 2024).

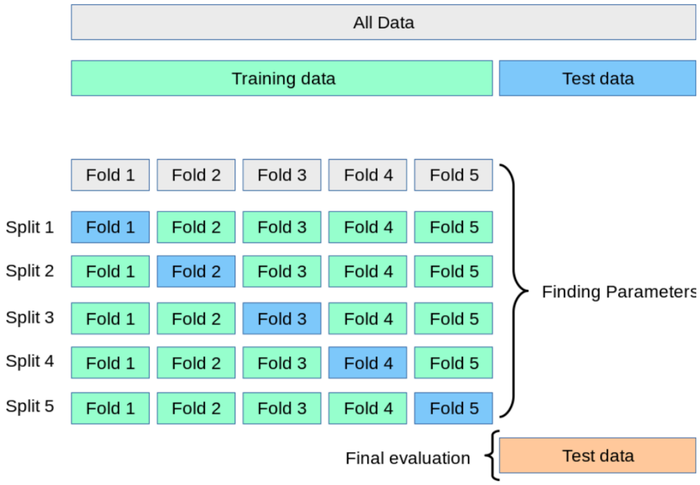
Figure 9. *Splitting Criteria for Regression-Tree Algorithms*

$$\mathrm{MSE} = \overset{\text{Mean}}{\boxed{\frac{1}{n}\sum_{i=1}^{n}}}(\overset{\text{Error}}{\boxed{Y_i - \hat{Y}_i}})^{\overset{\text{Squared}}{\boxed{2}}}$$

3.10 K-Fold Cross Validation

To assess the model's performance and ensure no risk of overfitting, k-fold cross validation (Figure 10) was employed within the modelling portion of the research. This technique was utilized to assess the model's ability to generalize to new, unseen data. The dataset is split into *k* subsets or folds, with one fold serving as the testing set, while the remaining *k-1* folds are used for training the model. This process is repeated *k* times, with each of the folds being used as the test set once (Kumar, A., 2024). Both 5-fold and 10-fold cross validation were utilized in the research. It was additionally applied to tune the hyperparameters of both models while also evaluating the performance across various hyperparameter configurations.

Figure 10. *Detailed Diagram of K-fold Cross Validation*



3.11 A/B Testing

Lastly, a segmented A/B test was conducted to assess whether users with historical low engagement and prior delivery errors would negatively impact key engagement metrics. Also known as split testing, A/B testing compares the performance of two versions of a campaign (A vs. B) to determine which version yields better results in terms of user engagement with the emails (Gallo, A., 2017). In this case, the normal group (A) consisted of users with typical engagement patterns, while the test group (B) included users with historically low performance. The A/B test did not involve a control and a treatment. Rather, the hypothesis was that Group A users would exhibit higher CTR and open rates rather than the Group B users. The eventual goal is to improve the marketing sender's domain reputation by reducing email sends that are unwanted or do not get delivered. Additionally, SQL Looker was used to visualize and track the KPI data over the course of the A/B test.

3.12 Data Limitations

The initial data collection process for this research involved extracting data covering three months, or one business quarter. This period was chosen to investigate performance trends within each email campaign over time. The modeling and analysis portions of the research were conducted using Databricks, a cloud-based data processing and analysis platform (Skaya, I., & Salet, M., 2025). However, due to memory limitations within the Databricks notebook environment, the dataset was reduced to approximately one month (4 weeks) of data. Despite reducing the dataset to one third of its original size, the timeframe was still sufficient for analyzing email campaign performance over time.

3.13 Ethical Considerations

Prior to the dataset being aggregated from the send-level to the campaign-level, the send-level data included information for each individual email sent, with each row representing a single delivery to a user's inbox. The data schema containing stored information on email delivery sends included personal information, such as the recipient's email address. Some of these email addresses included sensitive information, such as their first and last name. In order to protect the privacy of users, all email addresses were removed from the dataset before any aggregation or analysis was performed.

Chapter 4: Exploratory Data Analysis

This chapter presents a comprehensive overview of the exploratory data analysis (EDA) conducted on the aggregated and cleaned campaign-level dataset. The following sections identify key insights and patterns within the email campaigns and guide feature selection for the following predictive modeling. A statistical analysis was performed to examine the distribution of each variable and detect any potential outliers. Visualizations, such as heatmaps, boxplots and scatterplots, were employed to explore relationships between key variables and their correlation with respect to one another. The chapter additionally discusses the significance of these findings, and how they provide crucial insights for further exploration, specifically through the modeling of tree-based regression techniques in the following chapter.

4.1 Statistics Summary

Initial steps involved identifying the dimensions of the dataset and their respective data types, which were properly converted in the data preprocessing section of the methodology chapter. The statistical information examined within the dataset, including the count, mean and standard deviation, provided initial insights into some potential outliers. The statistical summary revealed key insights into the campaign-level data:

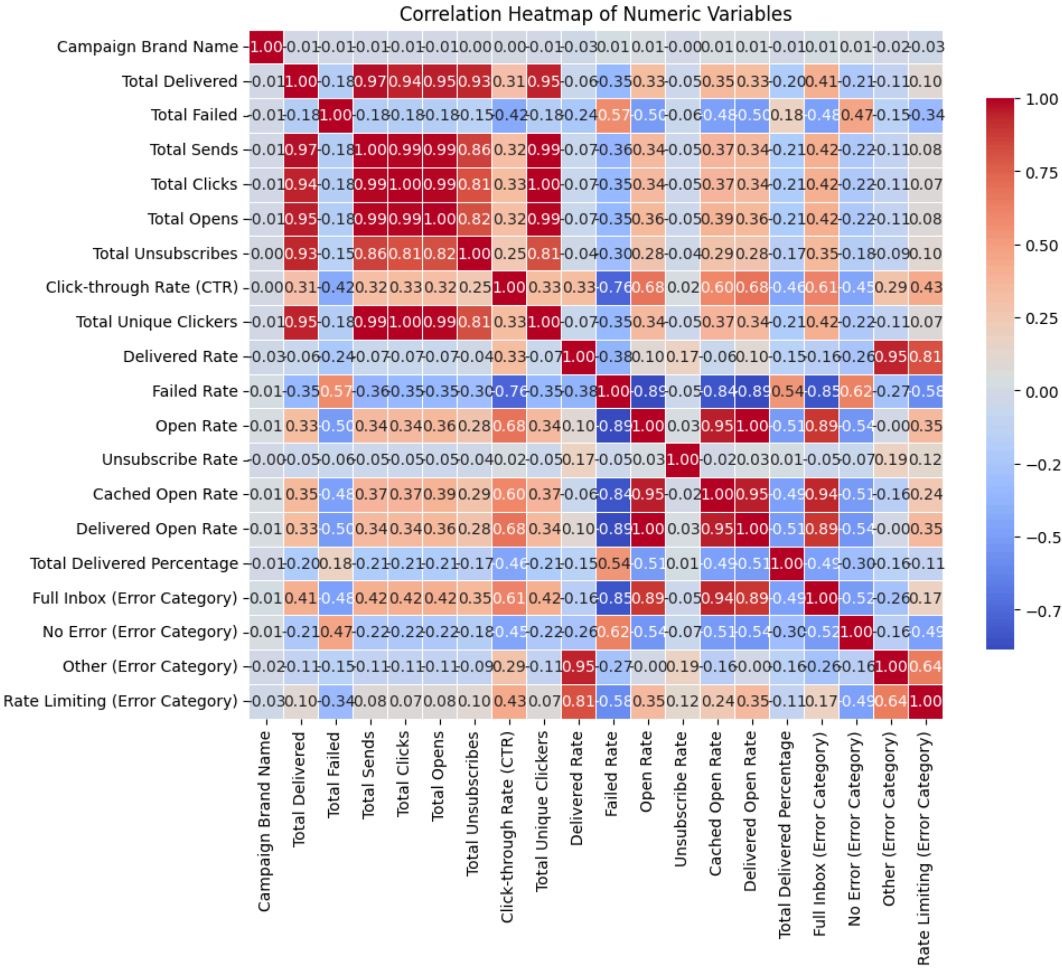
- Campaigns on average sent around 4,000 emails, but the dataset contained large outliers, with some campaigns reaching nearly 3 million recipients, significantly inflating the average values.

- Failure rates exhibited high variability, with 18% of campaigns experiencing no failures at all, suggesting the presence of potential outliers or inconsistencies in the data.
- Mean values for the total count of email sends, total delivered, and total failures were higher than the median values, indicating that most campaign deployments were small or medium-sized, while a small subset of large campaigns drove the majority of the volume (Appendix A).

4.2 Heatmap Correlation Matrix

A heatmap correlation matrix was used to identify strong intercorrelations among the independent variables. In Figure 11, scores closer to 1 (shown in red) indicate a strong positive correlation, whereas scores closer to -1 (shown in blue) indicate a strong negative correlation. The dark red cluster of values in the top left of the heatmap indicate that several variables, such as total email sends, clicks, opens, and unsubscribes, are not independent of each other. One possible explanation for the high collinearity among these features is that they all stem from total user email engagement. Since many of these metrics are calculated using total delivered as their denominator, they are inherently related to one another, and it is difficult to isolate any individually occurring effect within each variable.

Figure 11. *Correlation Heatmap Matrix of Numeric Variables within the Dataset*



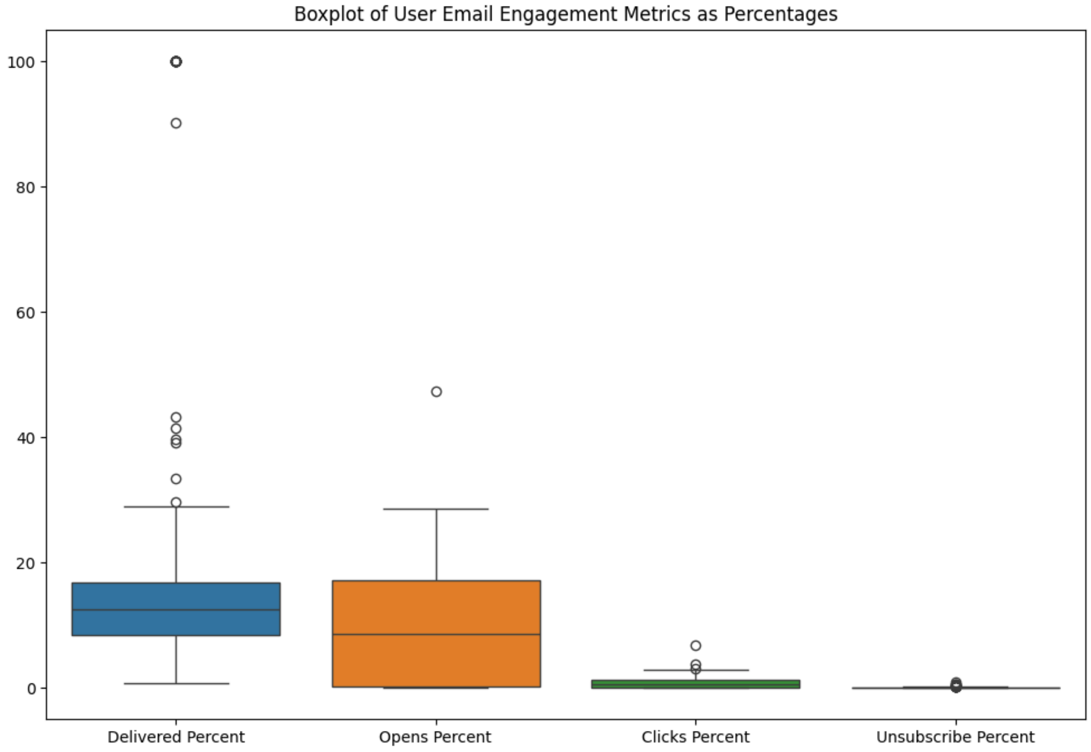
The heatmap additionally revealed insightful relationships into the correlation between key variables, one of which being the strong positive correlation (0.86) between total unsubscribes and total send volume (top-left of the heatmap). This correlation may point to several key contributing factors such as overexposure. When emails within a campaign are sent too frequently, users may feel overwhelmed, especially if the content is repetitive, which might lead users to disengage from the content and reduce inbox clutter by unsubscribing entirely from the email campaign. Another potential factor is that larger email campaigns have higher visibility, simply due to the fact that greater email volume creates more opportunities for users to opt-out from the email list, regardless of the quality of the campaign. This strong correlation between high send volume and high unsubscribe volume suggests that the increase in user opt-outs is largely driven by increased exposure, as a campaign that sends more emails inherently offers more opportunities for users to unsubscribe due to higher frequency of contact.

Additionally, a strong correlation of 0.68 between click-through rate (CTR) and open rate (central region of the heatmap) was found, suggesting that if users are intrigued enough to open an email, they are also more likely to engage with the CTA within the body of the email. This highlights the importance of emails containing an intriguing subject line to capture the audience's attention. After the user has opened the email, it should also contain a strong and visible CTA to guide users towards the next intended action. This strong correlation suggests that email campaigns with high open rates but low click-through rates may indicate that, while the subject line effectively captures user attention, the CTA is either weak or unclear, ultimately failing to convert the users' interest into meaningful engagement.

4.3 Boxplot of User Engagement Metrics

To provide a visual summary of user engagement within the email campaigns, a box plot was constructed to visualize the distribution of the four key metrics within the dataset, listed as delivered, opens, clicks, and unsubscribes, with each of them expressed as a percentage of total email sends.

Figure 12. *Boxplot of the Four Key User Email Engagement Metrics as Percentages*



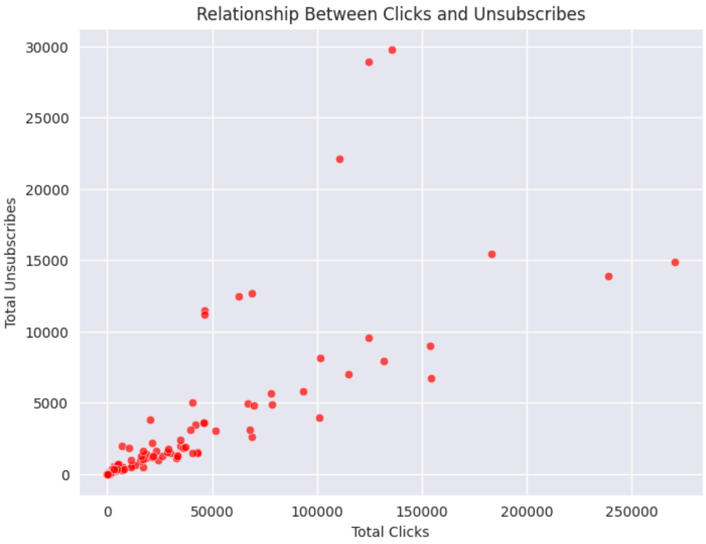
The percentage of emails delivered shows a central tendency of approximately 9% to 18%, indicating that only a very limited portion of emails are successfully delivered to users' inboxes (Figure 12). The presence of extreme values near the 90% to 100% mark suggests that a small subset of email campaigns were possibly either test sends or sent to a very small, highly curated list of users, thus achieving abnormally high delivery success. The 'opens' percentage box, while sharing an identical range with the delivered percentage, has a wider interquartile range (IQR) with a median of around 10%. An extreme outlier reaching approximately 50% shows that a certain campaign managed to capture half of the total user population to open the delivered emails, further indicating high performance irregularities.

Both clicks and unsubscribes metrics display a much smaller distribution with median values at approximately 1%. Nevertheless, the 'clicks' percentage box plot shows outliers between 5% and 7%, with low unsubscribe rates possibly due to the overall low number of emails that were successfully delivered, as indicated in the first box plot. Overall, the variance observed in the percentage of emails delivered and opened suggests further investigation to identify potential issues in regard to outliers within the dataset.

4.4 Relationship between Clicks and Unsubscribes

Scatterplots were used to visualize the relationship between key features within email campaigns, illustrating a positive correlation between the total number of clicks and total number of unsubscribes (Figure 13). The positive relationship implies that as users engage more with the campaign by clicking through the content, there is also an increase in the number of users opting out. This may suggest that after exploring the email content, some users determined it was either not relevant to their interests or that the frequency of emails was too high, leading them to unsubscribe from the campaign entirely. This positive correlation highlights that a high number of clicks does not necessarily indicate user satisfaction.

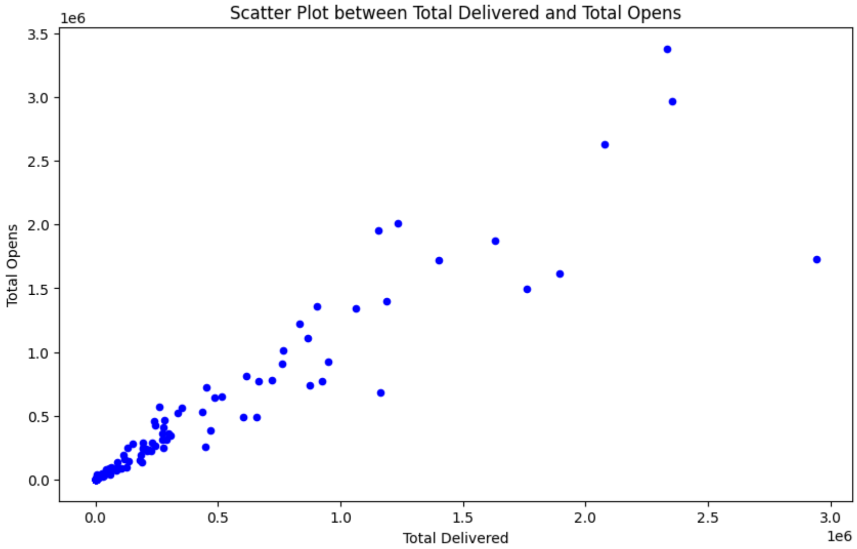
Figure 13. *Scatterplot visualizing the Relationship Between Clicks and Unsubscribes*



4.5 Relationship Between Delivered Emails and Total Opens

An additional scatter plot illustrates the positive relationship between the total number of delivered emails and the corresponding number of opens (Figure 14). A higher send volume simply provides more opportunities for recipients to engage with the emails, which can ultimately lead to increased conversions, sales or the campaign's overall objective achievement.

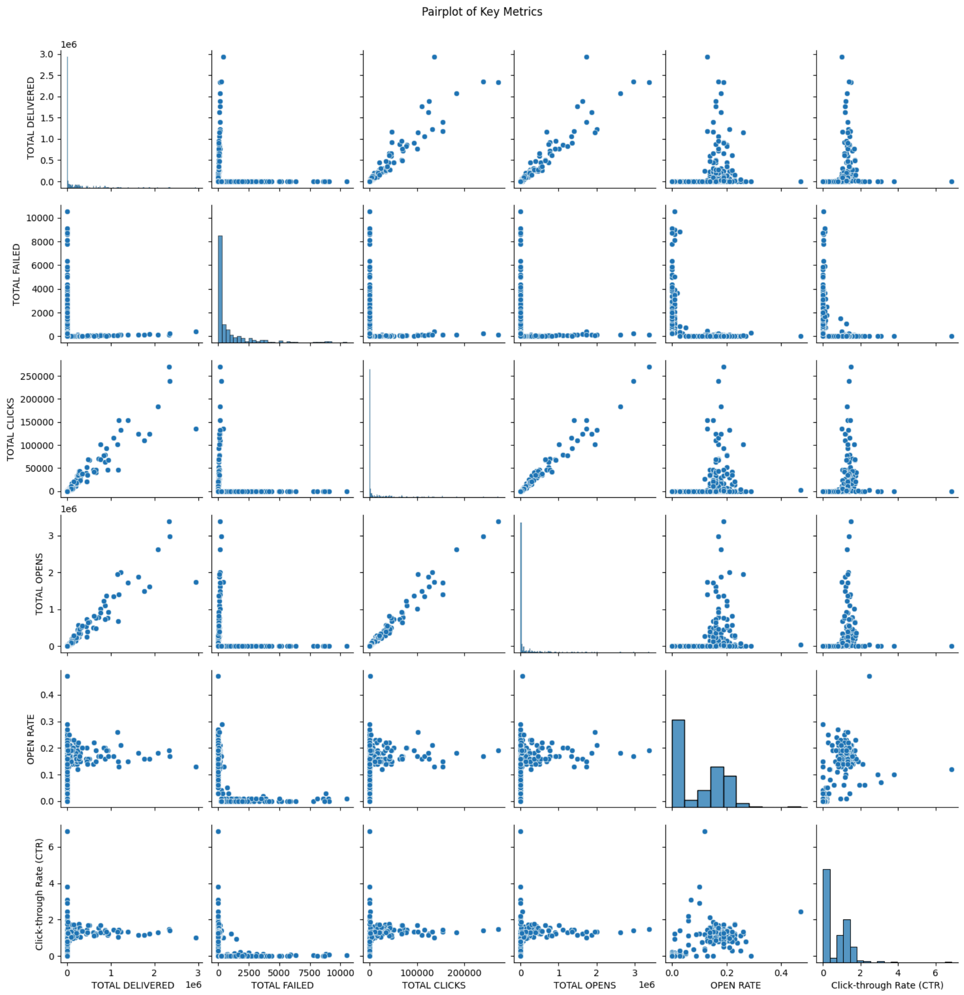
Figure 14. *Relationship Between Delivered Emails and Total Opens*



4.6 Pairplot of Key Metrics

A pairplot (Figure 15) was constructed to visually map out user interaction with the campaigns. One notable pattern observed across several scatter plots is the presence of L-shaped relationships, particularly between the failed email sends and key engagement metrics. This L-shaped distribution may stem from several underlying factors, the most prominent being that campaigns with high failure rates tend to exhibit very low user interactions. Conversely, campaigns with a high volume of successful deliveries typically demonstrate higher user engagement and lower failure rates. In other words, when a substantial portion of emails within the campaigns fail to send, it leaves little to no opportunity for user interaction. On the other hand, once a campaign reaches a critical number of successful deliveries, user interaction starts to increase sharply, resulting in graphs depicting an "L" shaped characteristic. The pairplot additionally shows that deliverability and openness drive user volume, whereas content relevance and clarity drive engagement through clicks, which highlights the need for both strong audience segmentation and well-timed email scheduling to maximize campaign performance.

Figure 15. *Pairplot of Key Email Performance Metrics*



The exploratory data analysis revealed several key insights, such as the presence of outliers within the campaign, which helped in understanding the structure of the research's extracted dataset. Visualization tools provided additional information on the strong correlation between variables, such as unsubscribes and total email sends, with boxplots and scatterplots further showing the datasets' range within user engagement. These findings provide a crucial understanding of the variability within the data, and provide a clear baseline to be used in the following regression-based modeling chapter.

Chapter 5: Tree-based Regression ML Models

This chapter provides an overview of the machine learning modelling techniques applied, and highlights the key results derived from the model performance metrics. Two regression-based machine learning techniques, Decision Tree and Random Forest Regressor, were implemented to predict the delivery rate for a given email campaign, measured by the number of emails delivered divided by the total number of emails sent. The models were then used to identify the top predictors of campaign delivery success. Hyperparameter Tuning was additionally performed to optimize both models using GridSearchCV. The chapter concludes with an analysis on the results from model assessment tools, including k-fold cross validation.

5.1 Decision Tree Regressor

A Decision Tree Regressor was trained to predict an email campaign's delivery rate, estimating the likelihood of emails successfully reaching recipients' inbox. The dataset was partitioned into training and testing sets using an 80:20 split ratio, applied uniformly across both the feature set and the target variable. Due to the substantial variability and skewness identified during Chapter 4, hyperparameter tuning was conducted using the GridSearchCV function from the Scikit-learn library. The function included searching across a predefined grid of hyperparameters to identify the optimal parameters for the model's performance. Five-fold and ten-fold cross-validation was employed to assess the model for potential overfitting using mean squared error. The hyperparameter combination that yielded the most optimal performance, shown in Table 11, was used to then train the decision tree regression model.

Table 11. *Optimal Hyperparameters for Decision Tree Regressor Identified via GridSearchCV*

Hyperparameter	**Value**
Criterion	Friedman MSE
Maximum Depth of Tree	5
Minimum Leaf Samples	10
Minimum Split Samples	2

The training R^2 value of 0.8118, as seen in Table 12, indicates that the decision tree regressor explains approximately 81.18% of the variance in successfully delivered emails within the training dataset. Similarly, a testing R^2 score of 0.7282 demonstrates that the model generalized well to new, unseen data, accounting for 72.82% of the variance in successful deliveries. The mean squared error (MSE) of 0.2532 and root mean squared error (RMSE) of 0.5031 further support the model's reliability, with high values as a result of the Box-cox transformation which applied a log transformation to the target variable. These error metrics offer no evidence of heavy overfitting and issues in the model accurately predicting email delivery success within the campaigns.

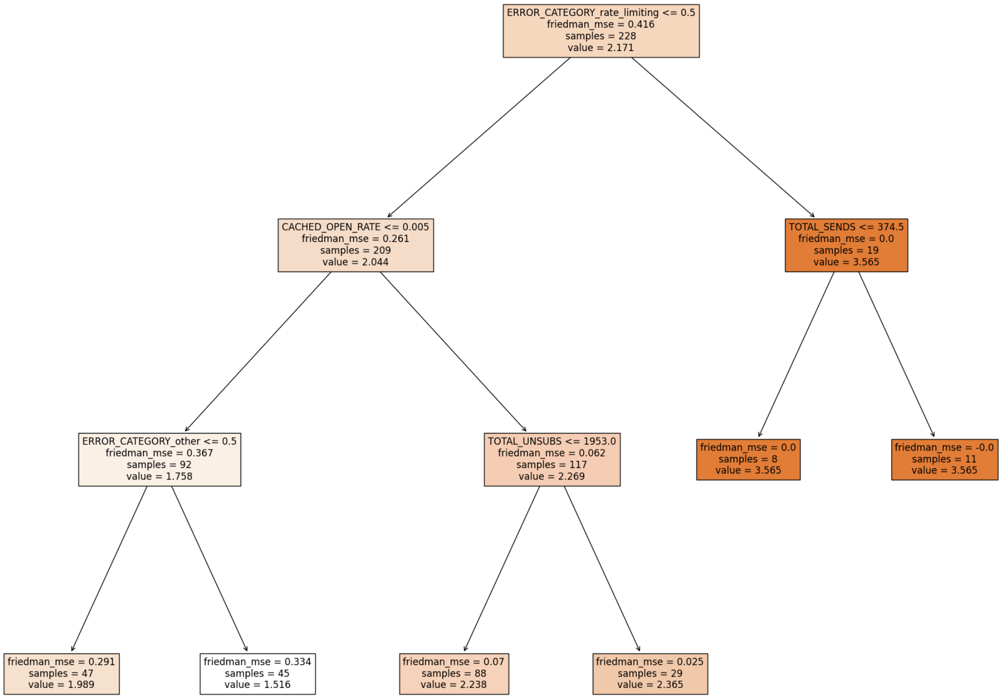
Table 12. *Decision Tree Regressor Model Result Outputs*

Metric	**Value**
Training R^2	0.8118
Testing R^2	0.7282
Mean Squared Error (MSE)	0.2532
Root Mean Squared Error (RMSE)	0.5031
R-squared (Testing)	0.7282

5.1.1 Visualization of Decision Tree Model

After fitting the Decision Tree Regressor, the splits were visualized using the Matplotlib library, and the maximum depth of the tree was reduced to three to enhance visibility of the tree. The final tree (Figure 16) indicates that rate limiting errors were the primary determining factors for delivery performance, suggesting that frequent email rates are associated with poor user engagement outcomes. Campaigns with a rate limiting value of <= 0.5 were then further split into additional subsections based on cached open rate, error types and subscribe metric values. These findings suggest that high email rates, high unsubscribe counts and low open rates contribute to user disengagement from the campaigns.

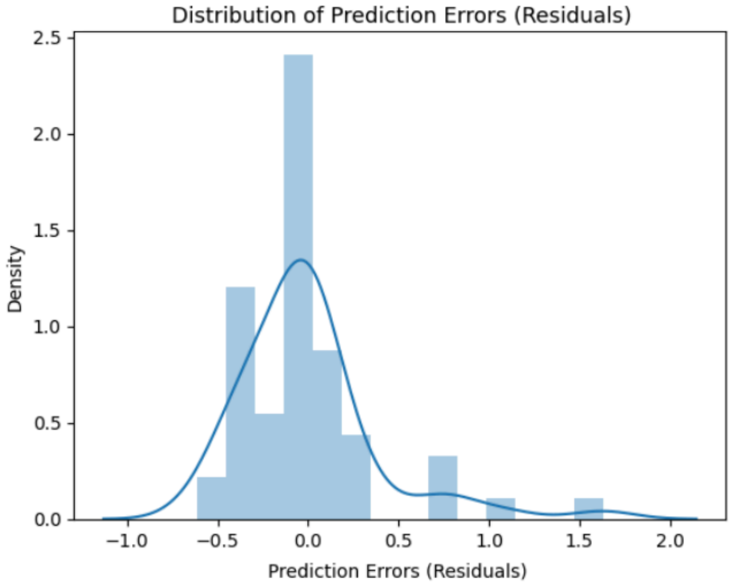
Figure 16. *Decision Tree Regressor Tree*



5.1.2 Decision Tree Regressor Performance Plots

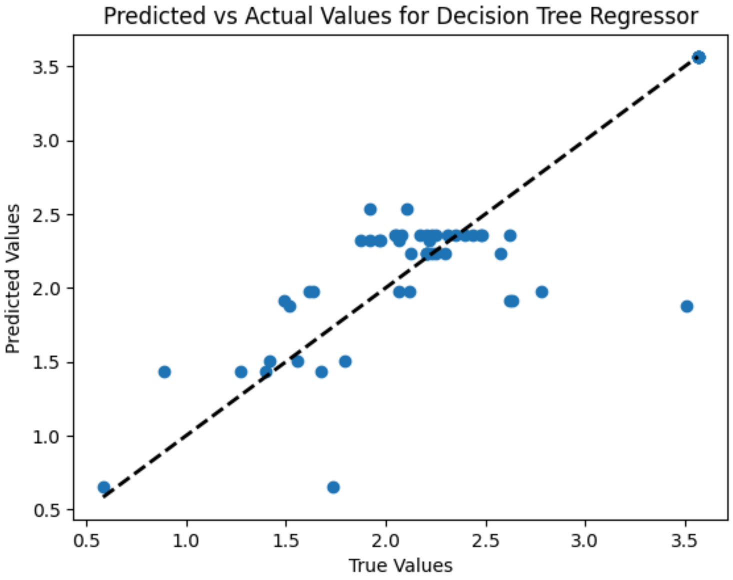
Additionally, to evaluate the decision tree regressor's performance beyond numerical metrics, the distribution of residual errors and a scatter plot comparing predicted vs actual values were visualized, depicted in Figure 17 and Figure 18, respectively. Figure 17 displays a histogram with a kernel density estimate (KDE) of the residuals, calculated as the difference between the actual and predicted values, where the actual values represent the true delivery percentages after the Box-Cox transformation, and the predicted values are the outputs generated by the model. The resulting distribution shows the model to be approximately normally distributed with some slight skewness resembling patterns of a right-skewed distribution. The peak of the bell-shaped curve reached y = 2.4 and centered at 0, with a tail indicating few outliers between 1.0 and 1.5 on the x-axis.

Figure 17. *Distribution of Residual Errors*



The second plot (Figure 18) presents a scatter plot comparing the relationship between the true versus predicted values from the model, with each dot representing a single observation. A diagonal reference line is also included to illustrate the ideal prediction pattern. The scatter plot shows a positive trend with a cluster of values aligned along the reference line at around 2.0 to 2.5 for both true (x) and predicted (y) values. However, several observations fall below the reference line, indicating slight underprediction by the model.

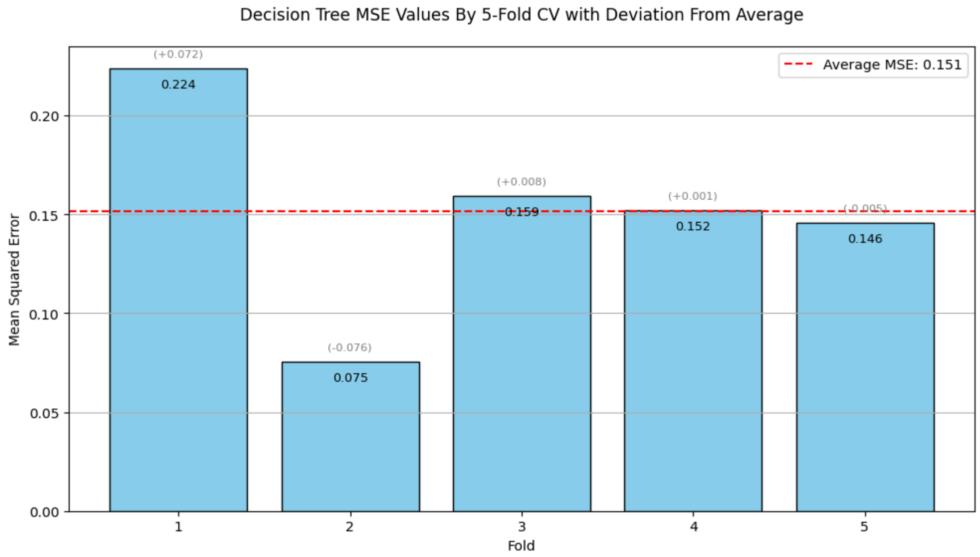
Figure 18. *Scatterplot comparing actual target variables with the decision tree model's prediction*



5.1.3 Cross Validation for Decision Tree Regressor Model

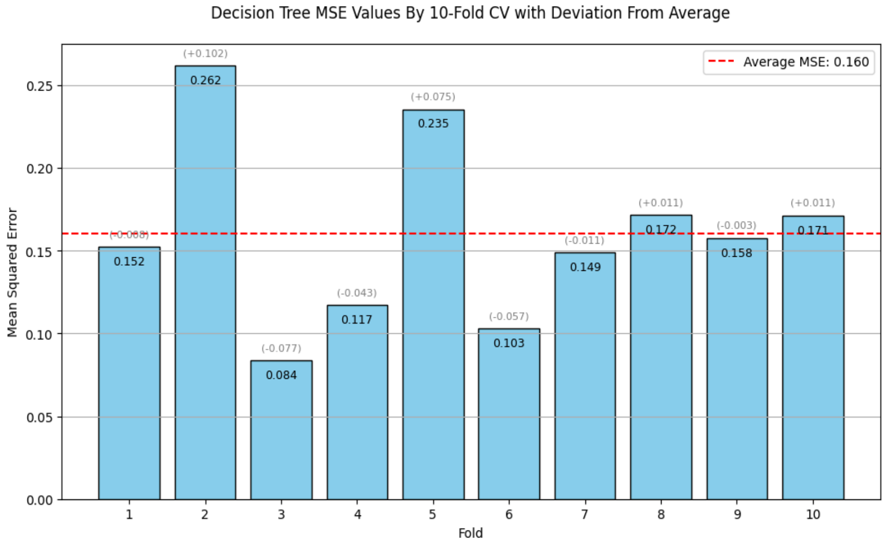
K-fold cross-validation was performed using both 5-fold and 10-fold splits to evaluate the model's generalizability on unseen data. The performance metric used was the mean squared error (MSE), where lower MSE values indicate stronger predictive performance. Randomly selected fold scores were examined to gain a deeper understanding of the model's behavior across different data partitions. The second fold (Figure 19) yielded an MSE value of 0.075, reflecting relatively low prediction error for that subset. This MSE value is notably lower than the average 5-fold cross-validation score of 0.151, suggesting that model performance varies substantially depending on the specific data split, as shown by the other folds within the figure.

Figure 19. *Decision Tree MSE Values by 5-Fold Cross Validation with Deviation From Averag*e



Additionally, a randomly selected fold score in the 10-fold cross-validation resulted in a MSE value of 0.235, exceeding the overall average of 0.160, indicating weaker performance for that particular split as shown in Figure 20. The significant difference between the individual fold scores from the average suggests that the decision tree model is sensitive and is not consistently reliable across different subsets of the data.

Figure 20. *Decision Tree MSE Values by 10-Fold Cross Validation with Deviation From Average*



5.1.4 Decision Tree Feature Importance

Feature importance analysis was used to identify which variables contributed most significantly to predicting the total percentage of successfully delivered emails within each campaign. Table 13 ranks the top five most influential predictors, with *Rate Limiting (Error Category)* as the most influential predictor, accounting for 50.5% of the model's decision-making process. Being of the highest importance, this feature indicates that nearly half of the explained variance in the model's prediction can be explained by rate limiting, or sending frequency in which emails are sent to a user's inbox within each campaign, as shown by the error category. Performance metrics such as total sends and total opens were additionally ranked among the top five most important features, possibly due to the frequency of sent emails having a negative impact on performance since failed email sends has a direct impact on user engagement with the emails.

Table 13. *Feature Importance Scores for Decision Tree Regressor*

Feature	Importance (%)
Rate Limiting (Error Category)	50.51%
Other (Error Category)	18.37%
Total Sends	8.76%
Delivered Open Rate	7.53%
Total Opens	4.39%

5.2 Random Forest Regressor

The same methods previously applied to the decision tree regressor were also implemented to train and evaluate a random forest regressor. It was employed to determine whether it could yield improved results when compared to the previous model, and to further assess the successful delivery of emails within a campaign. Both models utilized the same target variable and were trained with an 80:20 ratio train-test split and an identical random state of 42 for consistency. The hyperparameters used to optimize the random forest model via GridSearchCV are shown in Table 14.

Table 14. *Optimal Hyperparameters for Random Forest Regressor Identified via GridSearchCV*

Hyperparameter	Value
Max Depth	5
Minimum Leaf Samples	1
Minimum Split Samples	2
Number of Estimators	100

The random forest regressor outperformed the decision tree across all model performance metrics (Table 15). The model achieved a training R^2 score of 0.8652 and a testing R^2 score of 0.7856, performing stronger than the decision tree regressor by approximately +5.74%. This improvement was most likely due to the random forest ensemble technique, considering that it combines multiple decision tree result values together. Furthermore, the random forest MSE value was 0.2531, a minor decrease to the decision tree 0.2532 by a miniscule -0.04%. In contrast, the RMSE dropped more significantly from decision trees value of 0.5031 to 0.4812.

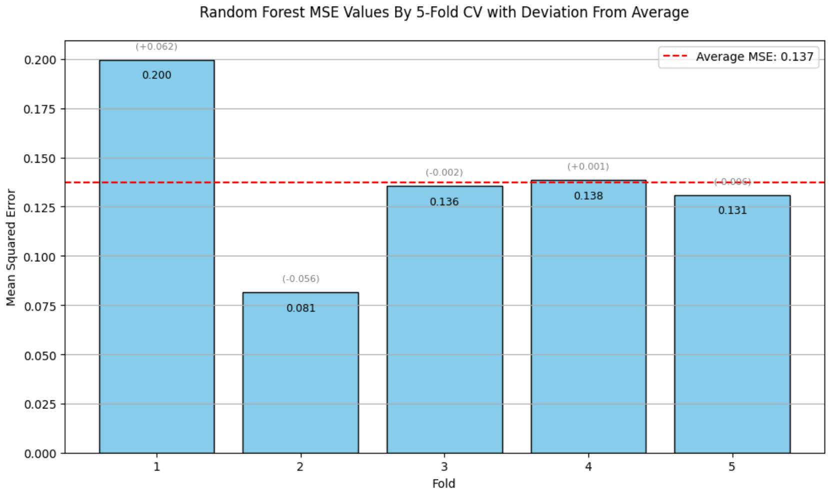
Table 15. *Random Forest Regressor Model Result Outputs*

Metric	**Average Value**
Training R^2	0.86527
Testing R^2	0.7856
Mean Squared Error (MSE)	0.2515
Root Mean Squared Error (RMSE)	0.4812
R-squared (Testing)	0.7856

5.2.1 Cross Validation

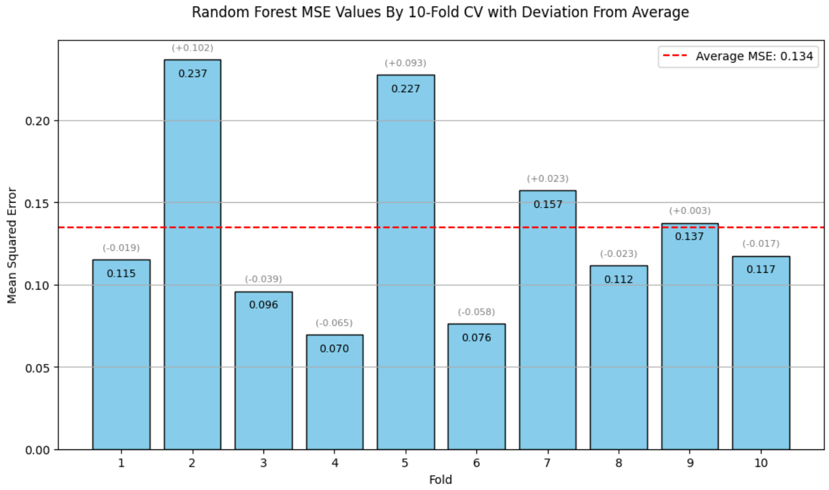
Similar to the decision tree, k-fold cross validation scores were used to assess the random forest model's consistency across various data splits. One randomly selected fold, the fifth fold (Figure 21), had a MSE value of 0.13, which was slightly lower compared to the overall average of 0.137, however showing similar amounts of variability within the folds when compared to the decision tree.

Figure 21. *Random Forest MSE Values by 5-Fold CV with Deviation From Average*



In contrast, a randomly selected fold, specifically the second fold in the 10-fold cross-validation, yielded a significantly higher MSE of 0.237 and performed worse than the 10-fold average score of 0.134, further indicating variability in the model's performance (Figure 22). Both the decision tree and random forest evaluation methods indicate that both models had substantial variability in performance between their respective folds.

Figure 22. *Random Forest MSE Values By 10-Fold CV with Deviation From Average*



5.2.2 Random Forest Feature Importance

Table 16 presents feature importance results that are largely similar to those of the decision tree model, with some slight differences. The *Rate Limiting (Error Category)* was again identified as the most influential predictor, contributing 53.8% to the model's decision-making process. Similar to what was seen in the decision tree model, this top feature reinforces the idea that sending frequent emails to users negatively impacts delivery performance. The total number of email sends followed as the second most impactful feature at 11.5%, with other high-influencing features including miscellaneous errors similar to decision tree scores at 11.1%, the cached open rate at 5.37%, and the total count of unsubscribes at 4.66%.

Table 16. *Feature Importance Scores for Random Forest Regressor*

Feature	Importance (%)
Rate Limiting (Error Category)	53.80%
Total Sends	11.50%
Other (Error Category)	11.10%
Cached Open Rate	5.37%
Total Unsubs	4.66%

Chapter 6: A/B Testing

This section provides a detailed overview of the construction, deployment, and analysis of the segmented A/B test using model insights performed in the research. The test aimed to evaluate the engagement behavior of historically low-performing users and assess how future email suppression might improve overall engagement and delivery success. The chapter begins by discussing the methodology used to split users into control and target groups, the timeframe of the test, and randomization criteria for data splitting. An 80:20 split in total email volume was implemented, with 80% of emails sent to users in the normal engagement pool, while the remaining 20% were grouped in a target group consisting of users flagged for historically low engagement.

This section also outlines the pipeline used to carry out the test and includes a visualization of the testing process. Additionally, it provides context into the company's internal practices for selecting email content for both the normal and target user pools using vector similarity searches. While the content selection approach was outside the scope of the segmented A/B test, it was used to determine whether historical user engagement patterns alone explained low performance, rather than differences in the quality of email content. The results of the segmented A/B test are then interpreted, focusing on differences in engagement metrics across normally engaging groups and historically low engaging ones. While no intervention was made to reduce the frequency of emails sent to either group during the test period, the chapter concludes with a comparative analysis of result outcomes, intended to inform future suppression strategies that may influence both engagement metrics and delivery outcomes using Google Postmasters.

6.1 Data Splitting

Due to several areas of research linking a high email send frequency to lower engagement metrics and decreased Postmaster scores, a segmented A/B test was conducted to investigate whether reducing the frequency of email sends to a particular group of low performing users has the potential to positively impact user performance metrics. Rather than a traditional A/B test, two groups were defined based on an internal criteria: a normal user pool (users deemed safe to email) and a flagged user pool. The test began with the creation of a “Do Not Contact” (DNC) list, which flagged and grouped a select number of users with historical data indicating low engagement metrics with past emails. In addition, users who have also returned delivery errors, such as ‘full inboxes’ and ‘user email does not exist’ were added into the DNC list. As a result, the DNC pool included users with past low engagement metrics and those with known deliverability issues with their emails.

Another list, containing users with consistent historical engagement and no delivery errors, were grouped into a normal user pool. During the test, a cap of 2 million total email sends per day was limited, with an 80:20 split: 80% of emails sent to users in the normal group, and 20% to users in the DNC user pool (Table 17). Given that both the normal user pool and DNC user pool contained varying group volume sizes, this 80:20 ratio was applied to email volume, not user count, ensuring the daily email sending limit followed the intended 80:20 split. Additionally, both user pools contained user information not related to a specific campaign, and included emails that either contained Google Postmasters information or did not. Because of this, changes in Google Postmasters following the evaluation of the segmented A/B test could not be directly measured. The following analysis was performed on the deployed segmented tracking A/B test from April 11 through April 22, 2025.

Table 17. *Description of Each User Group Pool for A/B Testing*

Groups	Description
Control (Normal)	Users in the normal send pool with consistent historical engagement patterns.
Target (Should Suppress)	Users from the DNC list flagged due to low engagement patterns or delivery errors.

The control group consisted of users from the original email send list with no modifications made to their email send patterns, comprising about 80% of the total email volume. The target group consisted of a randomly selected subset of users from the DNC list, making up the remaining 20% of the testing volume. These users had historical patterns of low engagement or deliverability errors with previously sent emails and were therefore added to the DNC user pool.

6.2 Randomization

Randomization was applied during the selection process of the *should suppress* user group. Each user within the DNC list was assigned a random number between 0 and 100. Since the segmented A/B test was based on total email volume rather than user count, users with a random number less than or equal to 20 (approximately 20% of the DNC population) were flagged as eligible but not yet selected to receive emails during the test period. The DNC user pool consisted of approximately 360,000 users, which was larger than the normal user pool of 289,000 (Table 18). The objective was to maintain an 80:20 split in daily email volume between the control and target groups.

Table 18. *User Pool Sizes*

Group	User Volume	Eligibility Criteria	Eligible User Pool	Email Volume
(Control) Normal	289,000	All users are included	289,000	1,600,000
(Target) Should Suppress	360,000	Random threshold ≤ 20	72,000	400,000

While 72,000 DNC users (20%) were flagged as eligible each hour, not all were sent emails. A second level of random sampling was applied to this eligible group to select only as many users as needed to fulfill the 20% daily email volume target of approximately 400,000 total daily emails. This two-step randomization process was performed to ensure that the A/B test split was based on total email volume, not user count. Every hour during the test, a new batch from the DNC user pool with a random number ≤ 20 was evaluated. From this batch, a further random sample was selected to meet the 20% email volume target. This process was repeated hourly to avoid any chance of bias from using a static subset of the DNC group throughout the test. Rotating the users also prevented any potential skewness that may arise from repeatedly using the same email recipients. Table 19 outlines the criteria used within the data split into separate testing (80%) and control (20%) groups. The overall goal was to improve key engagement metrics by examining the potential reduction in the number of emails sent to low engagement or users with past delivery issues.

Table 19. *Criteria Threshold for A/B Testing Split*

Criteria	Action
Target Group (20%)	Random selection from the DNC list (users with a number below the 20% threshold). Further random selection from eligible users to meet the 20% email volume criteria.
Control Group (80%)	Users from the normal pool, continued to monitor their engagement metrics at the normal sending pattern.

6.3 A/B Testing Split & Pipeline

Due to the 80:20 split in daily email volume, a skewed distribution between the two groups was observed (Table 20). The control group consisted of approximately 289,000 distinct users who collectively received 1.6M emails each day of the duration of the test. Additionally, the target group included about 72,000 distinct users pulled hourly from the DNC list, receiving 400,000 emails per day, with a total send volume of 2 million sends. On average, users from both groups received a total of 5.54 emails per day, rounded to approximately 6 for each user, which represents a relatively high frequency of email sends per recipient.

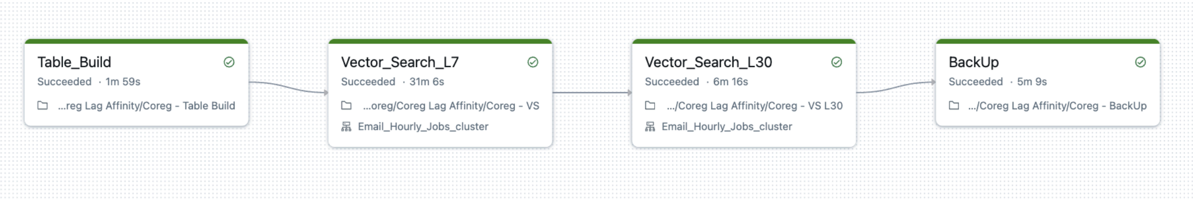
Table 20. *Volume Distribution of A/B Testing Groups*

Group	User Volume	Email Volume
(Control) Normal	289,000	1,600,000
(Target) Should Suppress	72,000	400,000

- **Average Number of Emails Sent per User:** 5.54 (approximately 6 emails)
- **Total Email Sample Size:** 2,000,000 emails per day

This test was monitored and analyzed over an 11-day period, focusing on key engagement metrics such as open rates and click through rates (CTR), to evaluate how the design variation impacted user behavior. To establish a reliable baseline for comparison, pre-test metrics were examined by analyzing the specific analytic pipeline or Directed Acyclic Graph (DAG) used during the A/B test. The DAG, referred to as coreg lag, computed lag-based features from past historical engagement metrics like open rate and CTR (Potters, C., & Rathburn, D, 2023). This preprocessing step ensured that both the control and target groups were accurately segmented and comparable based on their behavior baselines (Figure 23).

Figure 23. *Directed Acyclic Graph (DAG) for A/B Testing*


Table_Build
Succeeded · 1m 59s
...reg Lag Affinity/Coreg - Table Build
Vector_Search_L7
Succeeded · 31m 6s
...oreg/Coreg Lag Affinity/Coreg - VS
Email_Hourly_Jobs_cluster
Vector_Search_L30
Succeeded · 6m 16s
.../Coreg Lag Affinity/Coreg - VS L30
Email_Hourly_Jobs_cluster
BackUp
Succeeded · 5m 9s
.../Coreg Lag Affinity/Coreg - BackUp

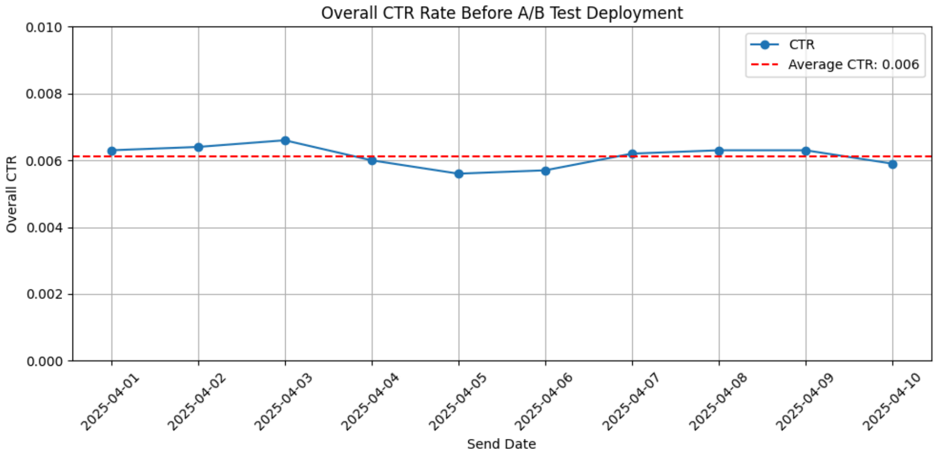
The first stage of the pipeline, labeled ‘Table Build’, involved constructing both a control table of users with normal historical engagement and a DNC table to identify recipients who had a history of either low engagement metrics or past delivery errors. This table was generated through a modified SQL query developed during the initial research section. Following the table creation, two pipelines, labeled ‘Vector Search L7’ and ‘Vector Search L30’, performed nearest-neighbor vector similarity searches using 7-day and 30-day historical windows. These two steps were part of the company’s internal practice for selecting email content to ensure consistency in content delivery across user groups. These steps were used to identify email offers for both the normal group and DNC-listed users by comparing their behavioral patterns to those of other users with similar historical engagement data.

This step in the DAG pipeline involved representing each DNC user as a feature vector, which incorporated key engagement metrics, such as open rates and CTR, across both a short term (7-day) and long term (30-day) timeframe. These feature vectors were processed through nearest-neighbor similarity searches on the Databricks DAG platform. Once these similar users were identified, the DAG pipeline chose email offers or content that those users had previously engaged with, and recommended these offers to the DNC users. This pipeline was also performed for the normal group to choose the best content, ensuring that the test focused solely on historical engagement patterns alone.

The L7 and L30 steps were part of the standard email delivery practices used for all users in A/B tests conducted by the company. The purpose of these steps was to standardize how emails were selected across groups while ensuring that the content had the potential to generate interaction based on historical patterns. Although DNC users had a history of low engagement, the DAG pipeline's purpose was to test whether this poor performance was due to their behavioral patterns and not the content they received. By removing content relevance as a factor in the user's historical poor performance, it ensured that the segmented A/B test focuses purely on how the DNC user's engagement metrics compare over time.

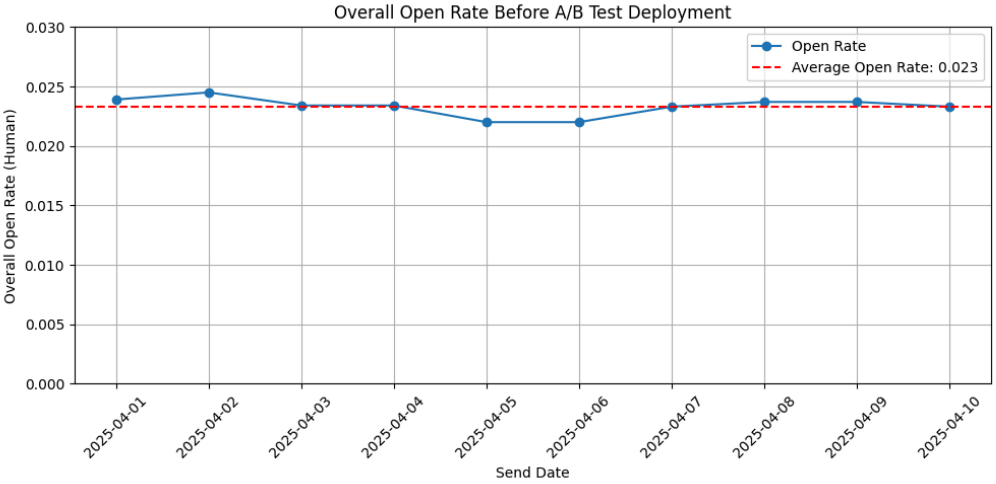
6.4 Before A/B Test Deployment

Figure 24. *Overall CTR Rate Before A/B Test Deployment*



Figures 24 and 25 display engagement metrics prior to the deployment of the A/B test. This was analyzed using the coreg lag DAG in order to compare changes and establish a baseline for comparison. The CTR rate from the period April 1-10 showed a relatively stable trend, averaging around 0.61%, with peaks reaching up to 0.65%. Similarly, open rate remained consistent throughout the same timeframe, averaging approximately 2.3%, and ranging between 2.2% and 2.6%.

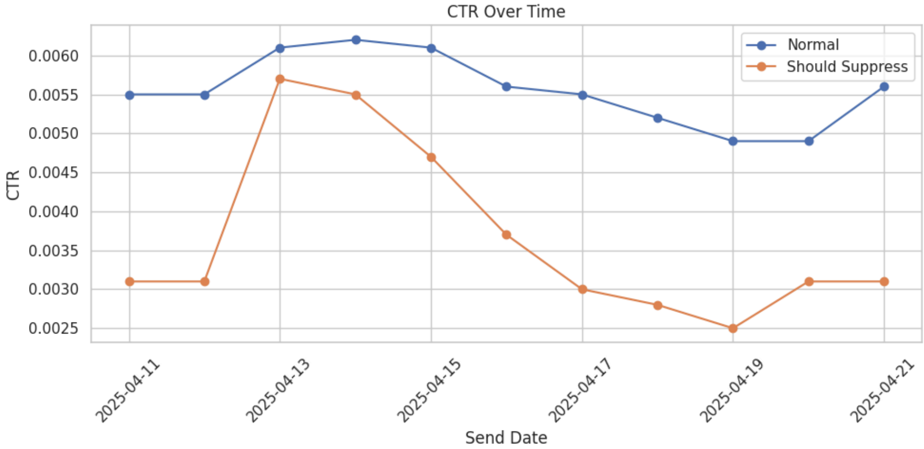
Figure 25. *Overall Open Rate Before A/B Test Deployment*



6.5 During A/B Test Deployment

The trends observed throughout the deployment of the A/B test supported the expectations of the segmenting approach (Figure 26). The normal group maintained a higher CTR throughout the test period, averaging at approximately 0.54%. Users who did not experience high frequency of sends and thus were not placed on the DNC list were more responsive to the delivered emails. Alternatively, the target group experienced a brief peak in CTR on 04/13, reaching around 0.57% before declining steadily to below 0.30% for the remainder of the test. This sudden peak in CTR appears to be an anomaly within the test and could be attributed to factors such as sample noise. These emails linked with lower engagement cause broader deliverability implications, as they can negatively affect external metrics tools like Google Postmaster scores. When engagement metrics such as CTR and open rate remain low, Google Postmaster might flag these email sender domains as poor quality. This can decrease the likelihood that the emails will land in the recipient's inbox, and have a higher chance of landing in the spam folder, or complete blocking of future emails by Google Postmasters.

Figure 26. *CTR Rate During A/B Test Deployment*



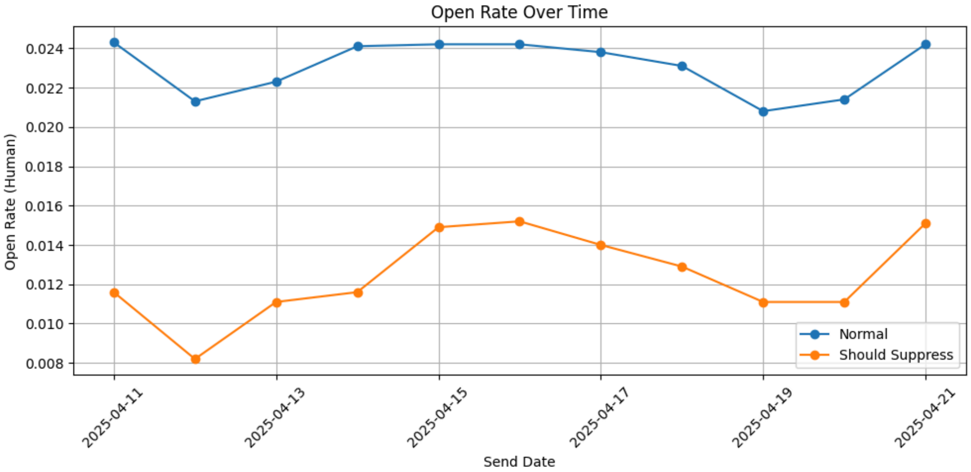
Analyzing open rates (Figure 27), the normal group consistently outperformed the should suppress group as well, maintaining strong performance between 2.1% and 2.5%, with an average of 2.2%. In contrast, the target group, who received fewer emails, experienced significantly lower open rates, averaging at around 1.5%. When paired with lower CTR trends observed in the group, it highlights a negative pattern of weak engagement. Poor engagement may trigger email providers to prioritize or flag certain emails as spam, reducing their visibility to future possible recipients. This effectively harms future campaigns from having high engagement rates, reinforcing the focus of maintaining and managing appropriate sending frequency patterns to potential users.

Additional day-of-week patterns were observed during the segmented A/B test. For both the normal and should suppress groups, open rate trends declined during the end of the workweek and over the weekends, particularly during the periods of April 11-13 and April 18-20, which encompassed Friday through Sunday. Trends from both these weekend periods show a decline in open rate ranging from 0.8% to 1.2%. In contrast, the weekday period range of

April 14-18 (Monday through Friday), showed an increase in open rates, ranging from 1.2% to 1.6%.

These trends suggest that users from both groups are generally more likely to engage with emails by opening them during weekdays, with a noticeable decline in open rate trends starting from Friday and continuing through the weekend, followed by an increase at the start of the new week. This pattern of user engagement can be further validated by analyzing a longer time period to determine whether the trend is significant. If consistent, it can support a more targeted and data-driven approach to scheduling email sends. This can help optimize email sending patterns and timing to further increase user engagement and overall campaign performance.

Figure 27. *Open Rate During A/B Test Deployment*



Overall, the findings from the segmented tracking A/B test support the hypothesis that reducing email frequency to flagged low engagement users has a measurable impact on the key engagement metrics. When comparing pre-test performance to the metrics observed during the deployment period, a noticeable decline between both metrics were observed. Prior to the test, CTR averaged around 0.61%, significantly higher than 0.54% exhibited during the test period.

Additionally, open rate saw a pre test rate of 2.3% compared to a slight decrease of 2.2%. This -11.48% and -4.35% drop of CTR and open rate respectively may be influenced by external or environmental factors not accounted for during testing.

It is important to note that the segmented A/B test focused on user pools rather than any specific email campaign, and as a result, changes in Google Postmasters domain reputation could not be directly measured. This is due to only a subset of emails within the company's dataset containing domain reputation scores, The emails used in the construction of datasets, exploratory data analysis (EDA), and predictive modeling were only included if there contained information stored on them from Google Postmasters . However, given the relationship between high engagement metrics and email deliverability, improving overall engagement performance may positively contribute to sender reputation and long-term delivery success.

Future work can involve investigating the decline in metrics during the A/B test. Future analysis could additionally implement a true A/B test framework, in which one subset of DNC users continues to receive emails, while the other have their emails suppressed. This will allow a more enhanced understanding into whether email suppression for historically low engaging users leads to an increase in overall campaign performance and the delivery success with emails.

Chapter 7: Discussion

This chapter provides a comprehensive summary of all findings from the research investigation, which includes exploratory data analysis (EDA), machine learning regression models, and A/B testing. The analysis identifies several key factors that influence the performance of email campaigns. Among the most impactful are high-frequency emails, high open rates paired with low click-through engagement, and their correlation with low sender reputation scores as measured by Google Postmasters. The chapter also reviews A/B testing results and outlines actionable further investigation.

7.1 High-Frequency Sends

The research identified that email deliverability is a critical issue impacting overall campaign performance (Table 20). To start, EDA showed a strong correlation between send volume and unsubscribe volume, identifying the first cause for decreased engagement and poor Postmasters scores. While this suggests that frequent email sends may overwhelm recipients and damage the sender reputation, the correlation is likely influenced by the sheer volume of emails, as campaigns with higher send volume will naturally yield a higher number of unsubscribes simply due to scale.

Table 21. *Summary of Analyses Identifying the Issue of High-frequency Sends*

Section	Method	Key Findings
Methodology	Boxplot of Error Category Count by Error Category *Send-level data	'Rate limiting' the third top most occurring error category within failed email sends.
Exploratory Data Analysis (EDA)	Heatmap Correlation Matrix	High correlation between total sends and total unsubscribes (0.86).
Tree-based Regression	Decision Tree Regressor	Feature selection scores 'rate limiting' highest with 50.51%.
Tree-based Regression	Random Forest Regressor	Feature selection scores 'rate limiting' highest with 53.80%.

Additionally, the regression models results reinforced these findings, with both decision tree and random forest regressors identifying *Rate Limiting (Error Category)* as the most significant predictor in identifying the target variable, or the percentage of successful email delivery within campaigns. This predictive model result directly affects the sender domain reliability with Google Postmasters. High unsubscribes, spam reports, and blocked email domains signal to Google that the domain is not well-received by recipients.

7.2 Low Further Engagement

Another major issue revealed throughout the analysis is low downstream engagement, in that the email content does not compel the recipient to take any action, likely focusing more on generating opens than on encouraging meaningful engagement through the call to actions (CTA). Future work within the area could include deeper content analysis to isolate factors within emails that influence click behavior. This can include placement of the CTA, content body wording or the relevancy of the CTA offer to the one advertised in the subject line. Further investigation may include building upon the A/B test to provide targeted content based on demographic qualities, such as users' age group, location or income level.

7.3 A/B Analysis Findings

The following summarizes the key findings discovered during the A/B test component of the research:

- High-frequency sending contributes to engagement fatigue and diminished performance.
- Suppressing emails to poor-performing recipients can positively impact domain-level metrics from Google Postmasters scores by improving sender reputation, thereby enhancing long term email deliverability.
- Comparison of pre-test and post-test engagement metrics suggests other external factors may also impact engagement, leading to another area of potential investigation.

The segmented A/B test confirmed that low-engaging recipients exhibited significantly lower open rates and CTR, supporting the hypothesis that historically disengaged users negatively impact overall key performance metrics. The overall investigation found the importance in curating audience sender lists and managing send frequency, which is key in maintaining strong engagement with users and having a successful email campaign to advertise.

Conclusion

The research set out to investigate the two key research objectives: identifying which email campaign features most directly correlate with delivery success, and how changes in campaign design can improve both delivery success and engagement metrics. The study utilized both statistical and visual exploration techniques to further reveal correlations between specific characteristics of bulk email sends and key performance metrics. Additionally, it employed regression-based machine learning techniques to identify core issues that can better inform future email marketing strategies. Although the research could not directly evaluate Google Postmasters domain reputation following the segmented A/B test, it suggests that improving engagement metrics by potentially suppressing historically low engaging users can enhance overall email deliverability.

The findings revealed that high-frequency email sends within campaigns negatively impact engagement and thus reduce the likelihood of successfully delivered emails to recipients, which may be a cause due to low Google Postmasters domain reputation scores. High send volumes led to email fatigue and increased unsubscribe rates, which collectively reduced overall deliverability. The research additionally revealed that open rates alone are not sufficient indicators of success within a campaign. While many campaigns achieve high open rates, few successfully converted those opens to clicks, highlighting a content performance gap that future research should address through additional investigation. Lastly, the research highlighted the importance of maintaining a strong sender reputation to maximize the likelihood of the email landing in a users' inbox, rather than in the spam folder. This was achieved by means of suppressing users with previous historical trends of low engagement. The results suggest how it is crucial for future marketing efforts to prioritize investigating historical user engagement trends before implementing send strategies.

The issues identified within this research offer multiple opportunities for further exploration. One key area includes determining why CTAs are underperforming by using techniques such as sentiment or textual analysis on the content or subject lines. This can uncover different aspects of the emails that may contribute to low engagement trends. Additionally, further investigation using demographics is another potential area to analyze and test, tailoring the content based demographic qualities such as age, location or income. This could additionally be integrated with A/B testing to determine which segments respond best to specific types of messaging. Overall, this research highlights the need for improved engagement-based user segmentation strategies, and how to develop a sustainable, optimized, and profitable digital marketing strategy over the long term.

References

Burtle, L., Head, S., & Lankford , S. (2013). *A Brief History of the Internet*. Board of Regents of the University System of Georgia. https://www.usg.edu/galileo/skills/unit07/internet07_02.

Church, C. (2023, May 26). *The History of Email Marketing (Infographic)*. Brafton. https://www.brafton.com/blog/email-marketing/the-history-of-email-marketing/

Dmarc. Org – Domain Message Authentication Reporting & Conformance. (n.d.). Retrieved April 29, 2025, from https://dmarc.org/

Duarte, N. (2024, August 8). *The secret to writing a call to action in a persuasive speech*. Duarte. https://www.duarte.com/blog/how-to-write-a-call-to-action-in-a-persuasive-speech/

Farshad, K. (2024, October 30). Understanding Decision Tree Regressor: An In-Depth Intuition. *Medium*. https://farshadabdulazeez.medium.com/understanding-decision-tree-regressor-an-in-depth-intuition-a1d3af182efd

Ferguson, A., & Hartmann, M. (2024, January 20). *Google Postmaster Tools—What It's and How It Can Help You*. Microsoft. https://learn.microsoft.com/en-us/dynamics365/customer-insights/journeys/google-postmaster

Gallo, A. (2017, June 28). A Refresher On A/B Testing. *Harvard Business Review*. https://hbr.org/2017/06/a-refresher-on-ab-testing

Hyperparameter Tuning Using Gridsearchcv. (2020, September). CodeSignal Learn. https://codesignal.com/learn/courses/introduction-to-machine-learning-with-gradient-boosting-models/lessons/hyperparameter-tuning-using-gridsearchcv

Kaddipudi, M. (2021). Journal of Emerging Technologies and Innovative Research. *Jetir*, *5*(10). https://www.researchgate.net/publication/351591033_Journal_of_Emerging_Technologies_and_Innovative_Research_ISSN_2349-5162

Kanellopoulos, T. (2025, January 13). Email Marketing Statistics 2024: ROI Insights & Trends. Competitors App. https://competitors.app/email-marketing-stats/

Key Concepts & Architecture | Snowflake Documentation. (n.d.). Retrieved May 7, 2025, from https://docs.snowflake.com/en/user-guide/intro-key-concepts

Kumar, A. (2024, August 16). K-fold Cross Validation In Machine Learning—Python Example. *Analytics Yogi*. https://vitalflux.com/k-fold-cross-validation-python-example/

Looker Business Intelligence Platform Embedded Analytics. (n.d.). Google Cloud. Retrieved May 7, 2025, from https://cloud.google.com/looker

Potters, C., & Rathburn, D. (2023, December 30). *Lagging indicator: Economic, business, and technical*. Investopedia. https://www.investopedia.com/terms/l/laggingindicator.asp

Rossiter, D. G. (2019, October 30). *Box-Cox Transformation*. Cornell University. https://www.css.cornell.edu/faculty/dgr2/_static/files/R_html/Transformations.html

Skaya, I., & Salet, M. (2025, March 31). *What Is Azure Databricks?* Microsoft. https://learn.microsoft.com/en-us/azure/databricks/introduction/

Taylor, J. (2024, February 21). *The History of Email Marketing*. Knak. https://knak.com/blog/history-of-email-marketing/

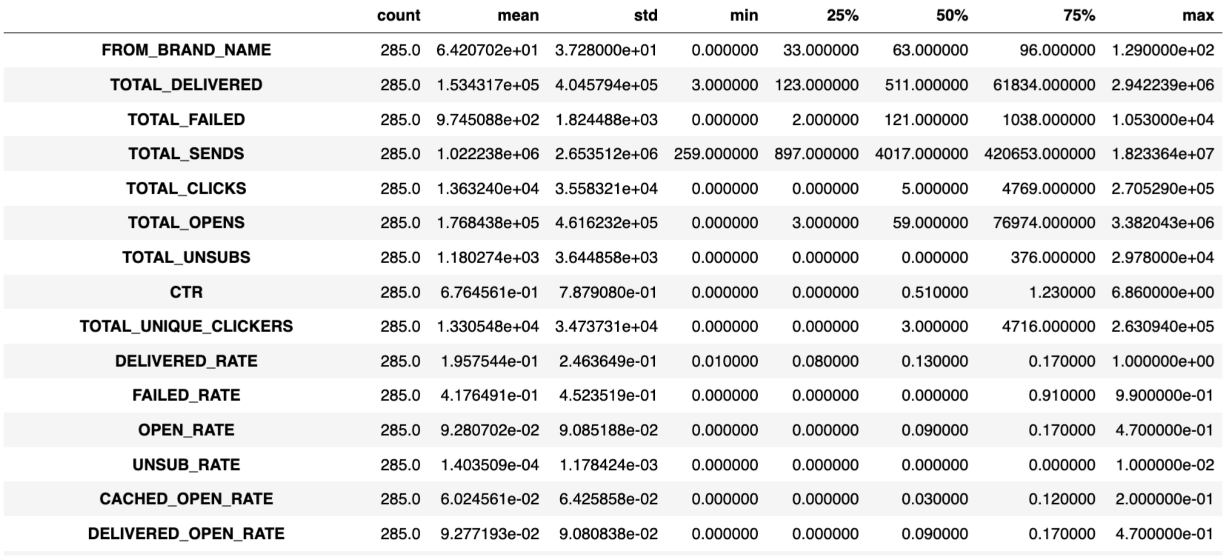
The Interplay Between DNS and Email: An Essential Guide for DNS Professionals. (2024, May 18). DNS Made Easy: A Digicert Company. https://dnsmadeeasy.com/resources/the-interplay-between-dns-and-email-an-essential-guide-for-dns-professionals

Thomas, J. S., Chen, C., & Iacobucci, D. (2022). Email Marketing As a Tool For Strategic Persuasion. *Journal of Interactive Marketing*, *57*(3), 377–392. https://doi.org/10.1177/10949968221095552

Ward, A. (2018, August 28). *How to Check Your Domain Reputation*. Postmark. https://postmarkapp.com/blog/how-to-check-your-domain-reputation

Appendices

Appendix A. *Statistical Summary of Numerical Variables within the Campaign-level Dataset*



	count	mean	std	min	25%	50%	75%	max
FROM_BRAND_NAME	285.0	6.420702e+01	3.728000e+01	0.000000	33.000000	63.000000	96.000000	1.290000e+02
TOTAL_DELIVERED	285.0	1.534317e+05	4.045794e+05	3.000000	123.000000	511.000000	61834.000000	2.942239e+06
TOTAL_FAILED	285.0	9.745088e+02	1.824488e+03	0.000000	2.000000	121.000000	1038.000000	1.053000e+04
TOTAL_SENDS	285.0	1.022238e+06	2.653512e+06	259.000000	897.000000	4017.000000	420653.000000	1.823364e+07
TOTAL_CLICKS	285.0	1.363240e+04	3.558321e+04	0.000000	0.000000	5.000000	4769.000000	2.705290e+05
TOTAL_OPENS	285.0	1.768438e+05	4.616232e+05	0.000000	3.000000	59.000000	76974.000000	3.382043e+06
TOTAL_UNSUBS	285.0	1.180274e+03	3.644858e+03	0.000000	0.000000	0.000000	376.000000	2.978000e+04
CTR	285.0	6.764561e-01	7.879080e-01	0.000000	0.000000	0.510000	1.230000	6.860000e+00
TOTAL_UNIQUE_CLICKERS	285.0	1.330548e+04	3.473731e+04	0.000000	0.000000	3.000000	4716.000000	2.630940e+05
DELIVERED_RATE	285.0	1.957544e-01	2.463649e-01	0.010000	0.080000	0.130000	0.170000	1.000000e+00
FAILED_RATE	285.0	4.176491e-01	4.523519e-01	0.000000	0.000000	0.000000	0.910000	9.900000e-01
OPEN_RATE	285.0	9.280702e-02	9.085188e-02	0.000000	0.000000	0.090000	0.170000	4.700000e-01
UNSUB_RATE	285.0	1.403509e-04	1.178424e-03	0.000000	0.000000	0.000000	0.000000	1.000000e-02
CACHED_OPEN_RATE	285.0	6.024561e-02	6.425858e-02	0.000000	0.000000	0.030000	0.120000	2.000000e-01
DELIVERED_OPEN_RATE	285.0	9.277193e-02	9.080838e-02	0.000000	0.000000	0.090000	0.170000	4.700000e-01

Appendix B. *Metric Summary of Categorical Variables within the Campaign-level Dataset*

Feature	Missing	Average Length
From Brand Name	1.55%	15.75
Error Category	0%	10.78

Appendix C. *Hyperparameter Grid For Decision Tree Regressor*

Parameter	Candidate Parameter Values				
Max Depth	5	10	15	29	None
Min Samples Split	2	5	10	20	—-
Min Leaf Samples	1	2	5	10	—-
Max Features	Sqrt	Log2	None	—-	—-
Criterion	MSE	Friedman MSE	MAE	—-	—-

Appendix D. *Hyperparameter Grid For Random Forest Regressor*

Parameter	**Candidate Parameter Values**				
Number of Estimators	50	100	200	300	—-
Max Depth	5	10	15	20	None
Minimum Split Samples	2	5	10	—-	—-
Minimum Leaf Samples	1	2	4	—-	—-
Maximum Features	Auto	Sqrt	Log2	None	—-
Bootstrap	True	False	—-	—-	—-