

# Learning to Predict Email Open Rates Using Subject and Sender

Daniel Vitor de Oliveira Santos<sup>a</sup> and Wladimir Cardoso Brandão<sup>b</sup>

*Institute of Exact Sciences and Informatics, Pontifical Catholic University of Minas Gerais,  
Dom José Gaspar Street, 500, Belo Horizonte, Brazil*

**Keywords:** Predictive Analytics, Email Marketing, Campaign Optimization, Text Mining, Machine Learning.

**Abstract:** The burgeoning daily volume of emails has metamorphosed user inboxes into a battleground where marketers vie for attention. This paper investigates the pivotal role of email subject lines in influencing open rates, a critical metric in email marketing effectiveness. We employ text mining and advanced machine learning methodologies to predict email open rates, utilizing subject lines and sender information. Our comparative analysis spans eight regression models, leveraging diverse strategies such as morphological text attributes, operational business factors, and semantic embeddings derived from TF-IDF, Word2Vec, and OpenAI's language models. The dataset comprises historical email campaign data, enabling the development and validation of our predictive models. Notably, the CatBoost model, augmented with operational features and dimensionally reduced embeddings, demonstrates superior performance, achieving a Root Mean Squared Error (RMSE) of 5.16, Mean Absolute Error (MAE) of 3.60, a Coefficient of Determination ( $R^2$ ) of 77.53%, and Mean Absolute Percentage Error (MAPE) of 14.73%. These results provide actionable insights for improving subject lines and email marketing strategies, offering practical tools for practitioners and researchers.

## 1 INTRODUCTION

The digital revolution has significantly transformed personal and professional communication. Among the various digital communication tools, email stands out as one of the most powerful and effective, particularly in marketing. The number of email users worldwide is approximately 4.3 billion in 2023 and is expected to exceed 4.8 billion by 2027 (The Radicati Group, 2023). Additionally, the total number of business and consumer emails sent and received daily surpassed 347 billion in 2023 and is projected to grow to over 408 billion by the end of 2027.


Email marketing plays a crucial role in acquiring new customers and retaining existing ones, standing out in a digital environment saturated with information. The effectiveness of email campaigns largely depends on the subject lines, which are among the first elements seen by recipients. Well-crafted subject lines, including personalization, emotional triggers and conciseness, can significantly increase open rates compared to generic subject lines (Stupar-Rutenfrans et al., 2019). However, many companies still rely on intuition rather than a systematic, data-driven ap-


proach to crafting these crucial elements.

The high return on investment (ROI) of \$40 for every dollar spent on email marketing reported by sources, such as Omnisend's 2022 data report, underscores the importance of optimizing every aspect of email campaigns, particularly the subject lines. This high ROI is attributable to the direct and personalized nature of email communication, which strengthens customer engagement and loyalty. Nevertheless, achieving such outcomes is heavily dependent on the effectiveness of the subject lines, as they significantly impact open rates and engagement (Advisor, 2024).

Despite its effectiveness, creating effective subject lines is fraught with challenges, primarily due to the heavy reliance on A/B testing. This method involves sending different versions of an email to a sample audience to determine which one yields higher open rates and engagement metrics. This approach is time-consuming, costly and fails to keep pace with the dynamic nature of consumer behavior, often requiring constant iteration and testing. Such methods can delay campaign launches and escalate operational costs, highlighting the need for more efficient strategies.

This paper addresses these challenges by proposing an analytical approach to deepen understanding the relationship between email subject lines and open rates. By leveraging comprehensive datasets from

<sup>a</sup>  <https://orcid.org/0009-0008-4116-6070>

<sup>b</sup>  <https://orcid.org/0000-0002-1523-1616>

previous email campaigns and employing advanced text mining and machine learning techniques, this research aims to develop predictive models to evaluate and maximize subject lines' effectiveness. These models are expected to provide actionable insights that could significantly reduce the reliance on inefficient and expensive A/B testing methods.

Eight different regression models including tree-based methods, ensemble techniques, kernel-based methods, and artificial neural networks are explored to identify the most effective approach for analyzing email marketing effectiveness. Each model type provides unique advantages: tree-based methods offer interpretability, ensemble techniques provide robustness, kernel-based methods capture complex relationships, and neural networks can learn non-linear patterns and handle large datasets. The research begins with extracting textual features from email subject lines and sender fields, then integrates operational features, including contextual and temporal data, along with the morphological attributes of the subject lines. In the final analysis phase, semantic features represented by embeddings are combined with operational features to assess their cumulative impact on model performance, supporting the central hypothesis that the semantic content of the subject line critically influences open rates, as evidenced by the development and testing of these models.

Historical email campaign data, integrating open rate metrics, were used to develop and validate the models and approaches. Among these, the CatBoost model was identified as particularly effective, incorporating operational features with advanced data processing techniques such as PCA and OpenAI embeddings. This model demonstrated substantial effectiveness in improving our understanding and optimization of email subject lines, which are critical for enhancing the performance of email marketing campaigns.

## 2 RELATED WORK

Research on email subject lines emphasizes their critical role in marketing efficacy, particularly how well-crafted lines can significantly boost user engagement. (Wainer et al., 2011) highlighted that subject lines that incite curiosity or promise utility tend to achieve higher open rates, demonstrating that recipients are more likely to engage with content that piques their interest or offers immediate value. Further studies like (Sahni et al., 2016) have shown that personalization, such as including the recipient's name, can increase open rates from 9.05% to 10.80%, leading to a 31% increase in sales leads and a 17% decrease in unsub-

scribes, illustrating a strong consumer preference for tailored email content.

Expanding on the importance of aesthetics, (Feld et al., 2013) noted that visual design elements are crucial, drawing a parallel between the impactful design in direct mail and engaging email subject lines. Meanwhile, advancements in machine learning have provided methods to enhance this effect; for example, (Balakrishnan and Parekh, 2014) employed a Random Forest model to predict open rates, achieving significant accuracy improvements with a RMSE of  $1.60 \times 10^{-2}$  and a correlation of 0.352, underscoring the potential to refine email marketing strategies through data-driven insights.

In the financial sector, (Conceição, 2019) applied a Random Forest algorithm to predict campaign success based on open rates, achieving an accuracy of about 82% by identifying effective subject line keywords. This complements psychological studies like those by (Miller and Charles, 2016), which explored how emotional triggers in subject lines, such as excitement or urgency, significantly influence user interactions by enhancing open rates.

Recent contributions by (Joshi and Banerjee, 2023) and (Paulo et al., 2022) further demonstrate the utility of integrating deep learning with traditional machine learning approaches. Joshi's Ngram-LSTM model excelled in sparse data conditions, achieving an accuracy of 84% within a 10% error tolerance. At the same time, Paulo's analysis across 140,000 email subject lines found that Random Forest models were most effective, achieving an accuracy of 62.4% and an F1-score of 62.2%. These studies highlight the importance of combining structural and semantic features of subject lines to refine the prediction of open rates, which this research builds upon by employing a multifaceted approach that blends semantic and morphological features with advanced machine learning techniques for enhanced predictive accuracy and deeper insights.

## 3 BACKGROUND

### 3.1 Email Marketing Concepts

Optimizing email marketing campaigns necessitates a deep understanding of key performance metrics. Metrics such as **Delivered** reflect the number of successfully received emails, highlighting the efficiency of email delivery systems. The **Open** metric measures recipient engagement by recording opened emails. Due to their reputation, the sender's identity (**Sender**) significantly influences open rates. The **Subject** line,

which captures immediate attention, along with the **Open Rate**—calculated as follows:

$$\text{Open Rate} = \left( \frac{\text{Number of Opens}}{\text{Number of Delivered}} \right) \times 100 \quad (1)$$

—both dictate the campaign’s success by reflecting the effectiveness of subject line strategies.

### 3.2 Dimensionality Reduction with PCA

Principal Component Analysis (PCA) is a statistical technique that reduces the dimensionality of datasets, aiming to preserve as much variability as possible. This reduction is accomplished by transforming the original variables into a new set of variables, linear combinations of the original variables and orthogonal to each other. These new variables, known as principal components, are ordered so that the first few retain most of the variation in all of the original variables, setting a foundational stage for handling high-dimensional data effectively (Hastie et al., 2009).

### 3.3 Curse of Dimensionality

The concept of the “curse of dimensionality,” first introduced by (Bellman, 1961), refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces. As the dimensionality increases, the volume of the space increases so exponentially that the available data becomes sparse. This sparsity is problematic for any method that requires statistical significance, as the data points are too dispersed to be representative. PCA effectively addresses this issue by reducing the dimensionality, concentrating the data into its most informative aspects and significantly diminishing the space in which the data points are spread. This reduction not only simplifies the computational demands but also enhances the effectiveness of subsequent analyses by mitigating the impact of the curse of dimensionality.

### 3.4 Text Embeddings

Following the theme of dimensionality reduction, text embeddings represent another crucial technique for managing complex data structures in natural language processing. Just as PCA helps to reduce the complexity of numerical data, text embeddings simplify the vast and intricate relationships within text data into manageable vectors. These vectors allow computational models to process and interpret large volumes of text efficiently, paving the way for deeper insights into language patterns and behaviors. Text embeddings are advanced techniques used to convert text

into a numerical format, enabling machines to understand and process language. These embeddings represent words or phrases as vectors in a high-dimensional space. Each dimension captures some aspect of the word’s meaning, and similar words are positioned closer together in this space, reflecting their semantic similarities. The process of creating embeddings involves analyzing a large corpus of text and learning to map words that have similar meanings close to each other. This mapping preserves the contextual nuances of words, allowing algorithms to recognize that certain words share similar contexts and meanings even if they never appear together in the text.

## 4 METHODOLOGY

The methodology employed in this paper is systematically illustrated in Figure 1 and comprises several critical stages: (1) Key Concepts and Data Overview, which provides a foundational understanding of the metrics and the dataset used; (2) Dataset Preparation and Feature Engineering, which involves cleaning, preprocessing, and engineering features from the data; (3) Embedding Generation and Text Preprocessing, which captures semantic information from the text data and prepares it for analysis; (4) Approaches, which detail the different strategies used to combine features and embeddings for model input; (5) Models, which describe the various machine learning models employed; and (6) Evaluation Metrics and Transformed Target Regressor, which presents all metrics employed and techniques to improve model performance.

### 4.1 Data Overview

The dataset, which includes over 4 billion emails delivered from 4,285 campaigns, was provided by Etus Media Holding (Etus Media Holding, 2024), a company specializing in email marketing management. The campaigns are between May 2022 and May 2024. This dataset, in Portuguese, pertains to Brazilian email campaigns and includes key attributes such as sender details, subject lines, delivery statistics, and user engagement metrics. Due to privacy and data security concerns for the clients involved, specific information about the data collection methods and the exact nature of the data cannot be fully disclosed. However, the general methodology involves extracting data through secure queries from a comprehensive email marketing database. This dataset is critical for subsequent preprocessing and analytical steps. The data is segregated into training (72%),

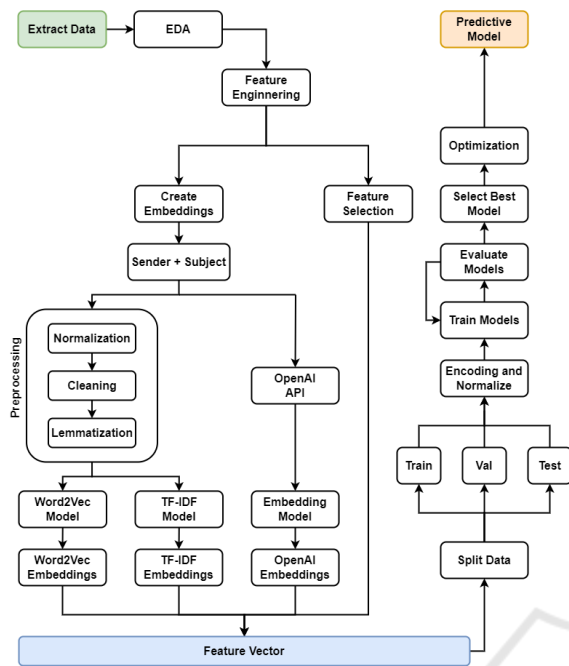


Figure 1: Flowchart of the methodology.

validation (8%), and test (20%) sets, ensuring comprehensive model training and validation. Data normalization and dimensionality reduction via PCA are critical steps in preparing this dataset for further analysis and model building.

## 4.2 Dataset Preparation and Feature Engineering

Following a thorough data overview, the primary goal in this phase is to prepare the dataset for robust analysis by ensuring data quality and extracting relevant features. Initially, entries with missing critical data, such as Subject, Sender, and Delivered, are removed. To ensure the reliability of our analysis, the focus is on campaigns that sent at least 100,000 emails. This threshold is selected based on the observation that campaigns with fewer emails often include pilot tests and may contain inconsistencies or experimental errors. In contrast, campaigns exceeding this email volume are generally considered validated and provide a more stable basis for statistical analysis, reducing the likelihood of anomalies that could skew the results.

Feature engineering enhances the model’s predictive capabilities. Created features include *sender open rate* (the average rate at which emails from the sender are opened), *sender count* (total campaigns from the sender), and *sender sum delivered* (total emails delivered by the sender). Temporal aspects such as *hour of the day* and *day of the week* are used alongside

email-specific metrics like *email open rate* and *delivered scaled* (normalized delivery volume), enriching the dataset with contextual and behavioral insights.

Feature selection, leveraging domain expertise and statistical methods, narrow down to the most impactful features, including detailed email metrics and temporal data points that significantly affect open rates. This careful preparation sets the stage for the next crucial step of embedding generation and text preprocessing, further refining the data for predictive modeling.

## 4.3 Embedding Generation

After preparing the dataset, this phase generates text embeddings to analyze the semantics of email subject lines, which are crucial for predicting open rates. We use three types of embeddings: TF-IDF, which highlights keyword importance; Word2Vec, capturing contextual word relationships; and OpenAI’s pre-trained models, providing advanced semantic understanding.

The Term Frequency-Inverse Document Frequency (TF-IDF) method transforms text data into numerical features by evaluating the importance of words within the subject lines relative to their frequency across all subject lines. This technique is beneficial for identifying keywords expected to influence email open rates, highlighting significant words while downplaying common terms that may not carry specific meaning in the context of email subjects (Salton and Buckley, 1988).

The Word2Vec model, introduced by Mikolov et al. (Mikolov et al., 2013), generates vector representations of words by considering the context in which they appear. This method captures semantic similarities between words, producing meaningful numerical embeddings. The continuous bag-of-words (CBOW) and skip-gram models are particularly effective at capturing nuanced relationships between words in a large corpus.

Employing a pre-trained model from OpenAI (OpenAI, 2024), embeddings are generated to encapsulate deeper semantic relationships within the text. These embeddings leverage large-scale language models trained on vast amounts of diverse text data, providing rich, contextualized representations of email subjects and senders. This approach is advantageous for capturing complex linguistic patterns and contextual nuances that simpler models might miss.

The text from the subject lines and the sender’s information are combined into a single textual input to generate these embeddings. This combined feature (Subject + Sender) ensures that the embeddings cap-

ture both the message content and the sender's influence, critical factors determining email open rates. By employing these three embedding techniques, the research aims to cover a broad spectrum of textual representations, from keyword importance and contextual semantics to deep contextual embeddings, leveraging the strengths of each method to enhance the predictive power of our models.

## 4.4 Approaches

Based on the extensive embeddings and preprocessing processes, this article employs several approaches to enhance the predictive power of the models by integrating various types of features and embeddings. Each approach is designed to optimize different aspects of the predictive modeling process, enabling a thorough examination of how various feature combinations can impact the accuracy of open rate predictions.

### 4.4.1 Morphologic

The first approach, *morphologic*, focuses on basic text features from email subjects, such as text length, word count, presence of special characters, numbers, and emojis, as well as the uppercase ratio, named entity count, letter ratio, vowel count, consonant count, and punctuation count. These features provide a foundational understanding of the structural elements of subject lines, which can be critical for assessing their effectiveness.

### 4.4.2 Operational

Building upon the Morphologic approach, the second approach, *operational*, extends the feature set by adding contextual and temporal information. This includes the number of emails delivered, scaled number of emails delivered, hour of the day, day of the week, time of day, hourly open rate, weekday open rate, and time of day open rate. By incorporating these features, the model gains insights into the broader operational context within which emails are sent and received, enhancing its ability to predict open rates based on real-world email campaign dynamics.

### 4.4.3 Text Preprocessing and Embedding

Before generating embeddings, the textual data from email subjects and sender information undergoes several preprocessing steps: normalization, cleaning, and lemmatization. Normalization involves converting all characters to lowercase and removing special characters and punctuation marks. Cleaning entails removing unnecessary characters such as numbers and

currency symbols (e.g., 'R\$'). Lemmatization, performed using the SpaCy library, reduces words to their base or root form, treating different word forms as a single entity. These preprocessing steps create a unified and cleaned textual input, vital for the generation of embeddings.

Building upon this foundation, the next set of approaches focuses exclusively on employing different types of embeddings: *TF-IDF*, *Word2Vec*, and *OpenAI* embeddings. These embeddings capture deeper textual semantics that may not be evident through structural analysis alone. TF-IDF and Word2Vec embeddings were trained on the dataset to highlight key terms and contextual meanings. Additionally, OpenAI embeddings, generated using the **text-embedding-3-small** model from OpenAI's API, provide a more sophisticated analysis of the text, leveraging large-scale language models.

This combined approach to text preprocessing and embedding generation sets the stage for developing advanced predictive models, moving forward to the strategies that will employ these embeddings to effectively forecast email open rates.

### 4.4.4 Combined

Finally, the most innovative approaches combine operational features with each type of embedding. For these strategies, PCA is utilized to reduce the dimensionality of the embeddings to six dimensions, effectively addressing the curse of dimensionality and enhancing our models' performance. By applying dimensionality reduction, we focus on the most informative features, thereby improving the efficiency and accuracy of our predictive algorithms.

By integrating these diverse strategies, the research aims to test various combinations of features and modeling techniques to identify the most effective approach for predicting email open rates. This comprehensive approach allows for evaluating each strategy's contribution to overall performance, paving the way for selecting the best-performing model in the subsequent phase of model development and assessment.

After the strategic development of predictive capabilities, the next crucial step involves selecting the most effective model, which will be covered in the subsequent subsection on models and evaluation.

## 4.5 Data Preprocessing

The dataset includes several categorical features, such as the sender and email address, which must be converted into a numerical format for the machine learning models to process. This conversion is achieved

using the One-Hot Encoding technique, transforming each categorical feature into a series of binary features representing unique categories. This step is crucial for preserving categorical information in a format that the models can effectively utilize.

Normalization is another essential step in data pre-processing, ensuring that numerical features have a consistent scale. Min-Max Scaling was applied to rescale features to a range between 0 and 1. This normalization improves the convergence of gradient-based algorithms and ensures that all features contribute equally during model training.

As previously discussed in the background section, PCA is applied to reduce the dimensionality of the embedding vectors and capture the most important features. This step simplifies the model and improves computational efficiency without significant loss of information. The number of dimensions was determined through an iterative method, testing ranges from 1 to 100 dimensions, with six being identified as the optimal choice. The embeddings are thus transformed into a smaller set of six principal components and then used as input features for the predictive models.

To improve the model performance and stabilize the variance of the target variable, each model was trained using the TransformedTargetRegressor technique. This technique applies a transformation to the target variable during training and inversely transforms the predictions back to the original scale. In this article, a logarithmic transformation was used due to its effectiveness in stabilizing variance and managing skewed distributions, as shown in Equation 2 and Equation 3.

The logarithmic transformation is applied as follows:

$$y_{\text{transformed}} = \log(y + 1) \quad (2)$$

The inverse transformation to retrieve the original target values is:

$$y_{\text{original}} = \exp(y_{\text{transformed}}) - 1 \quad (3)$$

The advantage of using TransformedTargetRegressor is that it allows models to better handle non-linear relationships and heteroscedasticity in the data. By stabilizing the variance, the models can achieve more accurate and reliable predictions, particularly when dealing with target variables that have a wide range of values.

## 4.6 Models

Advancing from the strategic integration of features and embeddings, this phase concentrates on the rigorous comparison and identification of the best-performing predictive model. Given the complexity

of email behavior and the nuances captured by various feature sets, selecting the most effective model is crucial for achieving high predictive accuracy in open-rate forecasting.

Various regression techniques have been selected based on their popularity in the literature and effectiveness in handling different types of data and complexities. These include tree-based methods, ensemble techniques, kernel-based methods, and neural networks, each offering unique advantages and suited for specific aspects of the analyzed data.

Tree-based methods, such as Decision Trees and Random Forest models, are appreciated for their interpretability and ability to handle numerical and categorical data. Decision Tree Regressor predicts the value of a target variable by learning simple decision rules inferred from the data features. At the same time, Random Forest improves predictive accuracy and controls over-fitting by constructing multiple decision trees and aggregating their predictions.

Ensemble techniques, including Gradient Boosting, LightGBM, XGBoost, and CatBoost, are employed for their robustness and high accuracy. These methods build models sequentially or in parallel, correcting errors made by previous models or combining multiple models to enhance overall performance. For instance, LightGBM and XGBoost are particularly noted for their efficiency with large datasets and their capability to optimize a differentiable loss function using gradient descent.

Kernel-based methods, represented by Support Vector Regressor (SVR), offer versatility in both linear and non-linear regression tasks through kernel tricks, making them effective in high-dimensional spaces.

Neural networks, such as the Multilayer Perceptron Regressor (MLP), are capable of modeling complex non-linear relationships due to their deep learning capabilities and multiple layers of neurons. The MLP model used in this article is configured with a specific architecture and learning parameters to optimize performance. Specifically, the MLP employs a single hidden layer consisting of 1024 units, utilizing the rectified linear unit (ReLU) activation function. Training is conducted using the Adam optimizer with a constant learning rate of 0.001. These settings are selected for their effectiveness in handling complex patterns and interactions within the data, enhancing the MLP's ability to learn and make accurate predictions.

Each model undergoes a thorough training and evaluation process using the prepared datasets, split into training, validation, and test sets as previously outlined. This rigorous evaluation includes cross-

validation and performance assessment using multiple metrics to ensure that the models perform well on training data and generalize effectively to unseen data.

#### 4.7 Evaluation Metrics

The evaluation of predictive models employed in this article utilizes a comprehensive set of metrics, each providing unique insights into the model's performance and the accuracy of the predictions. These metrics are essential for a thorough understanding of how well the models predict email open rates, and for selecting the best model based on empirical evidence.

Mean Absolute Error (MAE), as shown in Equation 4, is used to measure the average magnitude of the errors in a set of predictions, providing a straightforward indicator of prediction error magnitude without considering the direction of the errors.

Root Mean Squared Error (RMSE), given by Equation 5, takes the square root of the average squared differences between predicted and actual values. It is favored in scenarios where larger errors are more detrimental, as it gives greater weight to them, thus providing a more sensitive measure of prediction accuracy.

The Coefficient of Determination ( $R^2$ ), defined in Equation 6, is crucial for assessing the proportion of variance in the dependent variable that can be predicted from the independent variables. It helps determine the models' overall fit to the data, with values closer to 1 indicating a more accurate model.

Mean Absolute Percentage Error (MAPE), expressed in Equation 7, presents prediction accuracy as a percentage, offering a clear perspective on the relative size of the prediction errors concerning the actual values.

To further enhance model performance and address issues of non-linear relationships and heteroscedasticity, the TransformedTargetRegressor technique with a logarithmic transformation is applied. This transformation stabilizes the variance of the target variable, facilitating more reliable predictions. The logarithmic transformation is particularly effective in managing the effects of skewness in the target data, enhancing the model's ability to make accurate predictions across a range of values.

The equations for the evaluation metrics and the transformation process are as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6)$$

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (7)$$

This rigorous approach to model evaluation ensures that the predictive models developed in this article are effective in theoretical settings and robust and applicable in practical scenarios, providing reliable forecasts that can significantly enhance the efficacy of email marketing campaigns.

## 5 RESULTS

This section presents a comprehensive analysis of the performance and optimization of various machine learning models for predicting email open rates. The evaluation includes a range of performance metrics, model selection processes, optimization techniques, feature importance assessments, diagnostic evaluations, and performance comparisons across different data segments.

### 5.1 Model Performance Overview

This section provides an in-depth analysis of the performance of various machine learning models, employing different strategies with the primary objective of identifying the most effective approach for predicting email open rates. The evaluation focuses on comparing a diverse range of machine learning models and feature engineering strategies to determine which combination maximizes predictive performance. The models were evaluated based on four key metrics: RMSE, MAE,  $R^2$ , and MAPE as illustrated in Figure 2.

Among all evaluated models, CatBoost and LightGBM emerged as the most effective, particularly when integrating operational features, OpenAI embeddings, and PCA for dimensionality reduction. This combination not only mitigates the curse of dimensionality but also enhances the models' ability to discern nuanced patterns within the data, significantly improving predictive accuracy.

Detailed performance metrics analysis revealed that tree-based models such as Random Forest, Gradient Boosting, LightGBM, and CatBoost are particularly well-suited for this task, due to their superior predictive capabilities. These models benefit from ensemble learning techniques that significantly enhance accuracy and robustness. In contrast, strategies relying solely on basic text features, such as the morphologic strategy, proved less effective. This underscores

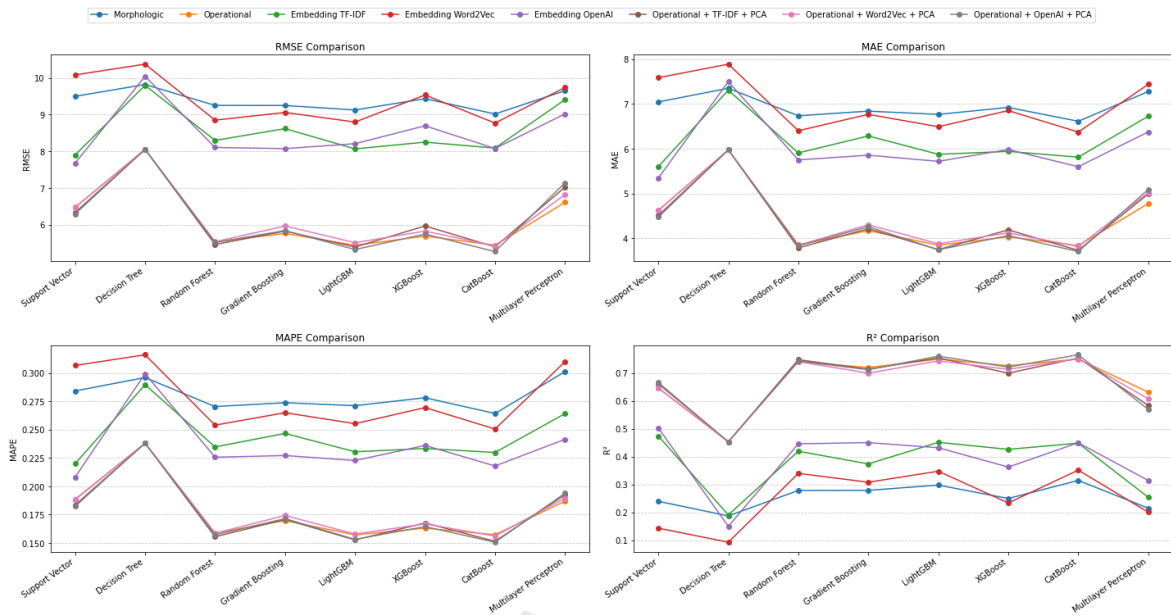


Figure 2: Comparison of RMSE, MAE,  $R^2$ , and MAPE for Different Models and Strategies.

the complexity of the data and the limitations of simple text attributes in fully capturing its nuances.

Models incorporating embeddings consistently outperformed those relying solely on basic text features, highlighting the critical role of semantic information in the prediction process. The integration of operational features with embeddings, especially when paired with PCA for dimensionality reduction, achieved the best results, providing a comprehensive solution that leverages both textual and operational data insights.

### 5.2 Best Model Selection

Following the detailed evaluation of various models, the CatBoost and LightGBM models consistently demonstrated the highest performance, particularly when incorporating operational features alongside OpenAI embeddings with PCA for dimensionality reduction. This section details the methodology used to identify and optimize the best-performing model, which holds substantial practical implications for enhancing email marketing strategies.

A comprehensive ranking system based on a heuristic aggregation of key performance metrics—RMSE, MAE, MAPE, and  $R^2$ —was implemented to determine the most effective model. For RMSE, MAE, and MAPE, where lower values signify better performance, models were ranked from best to worst. For  $R^2$ , higher values are better, so rankings were inverted to maintain consistency across all metrics, with lower ranks indicating superior performance.

mance.

The ranks for each metric were summed to derive an overall performance score for each model, facilitating a direct comparison of their effectiveness across diverse conditions. This ranking method, detailed in Table 1, is beneficial as it accounts for the error magnitude and the proportion of variance the model explains, providing a comprehensive view of model robustness and accuracy.

Table 1: Top 5 Best Models Based on Overall Performance.

Type	RMSE	MAE	$R^2$	MAPE
CB+E1	5.268	3.705	76.6%	15.0%
LGBM+E1	5.323	3.745	76.1%	15.3%
CB+E2	5.398	3.732	75.4%	15.1%
LGBM+E2	5.392	3.749	75.4%	15.2%
CB+E3	5.422	3.818	75.2%	15.6%

**Legend:** CB = CatBoost; LGBM = LightGBM; E1 = Operational + Embedding OpenAI + PCA; E2 = Operational + Embedding TF-IDF + PCA; E3 = Operational + Embedding Word2Vec + PCA.

The CatBoost model with the Operational + OpenAI + PCA strategy emerged as the best performer, achieving the lowest RMSE and one of the highest  $R^2$  values. This model’s ability to effectively integrate advanced feature engineering with powerful text embeddings and dimensionality reduction highlights its suitability for the task.



### 5.3 Model Optimization

Hyperparameter optimization for the CatBoost model was conducted using a Grid Search approach, which explored a range of hyperparameter combinations to maximize predictive performance and ensure robustness against overfitting. This process focused on key parameters that significantly influence the model's performance, such as iterations, maximum depth, and the evaluation metric, as shown in Table 2.

The specific hyperparameters tuned during this process included:

Table 2: Optimized Hyperparameters for CatBoost.

Hyperparameter	Value
Random Seed	42
Use Best Model	True
Evaluation Metric	RMSE
Iterations	6000
Max Depth	6

The optimization process prioritized the minimization of the root mean square error (RMSE) as the primary evaluation metric. This careful tuning aimed at balancing the model's bias and variance, crucial for achieving reliable predictions.

The outcomes of the optimization are highlighted in the Table 3, demonstrating the improved performance of the CatBoost model with these refined settings:

Table 3: Performance of Optimized CatBoost Model.

Metric	Value
RMSE	5.164
MAE	3.6097
$R^2$	77.53%
MAPE	14.73%

These results confirm the effectiveness of the Grid Search method in refining the model's hyperparameters, significantly enhancing its ability to generate highly accurate predictions. The optimized settings not only improved key performance metrics but also confirmed the model's enhanced capacity to capture complex patterns in the data, making it highly suitable for practical applications in predicting email open rates.

### 5.4 Feature Importance Analysis

The optimized CatBoost model provided crucial insights into the features that significantly influence email open rates. This analysis is instrumental in understanding the key drivers behind recipient engage-

ment and optimizing email marketing strategies accordingly.

Feature importance analysis identified critical predictors for email open rates. The most influential feature, *Sender Open Rate*, underscores the substantial impact of the sender's reputation on recipient behavior, emphasizing the importance of maintaining a high sender reputation to enhance open rates. Other notable features include the number of emails delivered (*Delivered* and *Delivered Scaled*), various open rate metrics (*Hour Open Rate* and *Email Open Rate*), and the principal components of the OpenAI embeddings (*OpenAI PCA 0*, *OpenAI PCA 1*, etc.). Table 4 lists these top features, offering actionable insights for marketers.

Table 4: Top 10 Most Important Features in the Optimized CatBoost Model.

Feature	Importance
Sender Open Rate	20.54%
Delivered	9.13%
Delivered Scaled	7.79%
Hour Open Rate	5.94%
Email Open Rate	5.42%
OpenAI PCA 0	4.49%
OpenAI PCA 1	3.63%
OpenAI PCA 2	3.49%
OpenAI PCA 5	3.45%
OpenAI PCA 4	3.29%

The inclusion of principal components from the OpenAI embeddings as important features demonstrates their effectiveness in capturing relevant semantic information from the text. These components help represent textual nuances that contribute to accurate open rate predictions. Additionally, the analysis reveals that contextual and temporal characteristics, such as hourly and email open rates, are crucial for model performance. These features provide valuable insights into when and how recipients interact with emails, enabling more precise predictions.

This comprehensive analysis underscores the need to combine advanced feature engineering with powerful textual representations to capture both contextual and semantic information, significantly enhancing the model's predictive power.

### 5.5 Model Diagnostics

To evaluate the performance of our optimized CatBoost model further, we conducted a detailed diagnostic analysis using two primary visualizations: the histogram of residual errors and the actual vs. predicted values plot. These visualizations provide in-

sights into the model’s error distribution and predictive accuracy.

### 5.5.1 Histogram of Residual Errors

The histogram of residual errors (Figure 3) displays the distribution of prediction errors, calculated as the difference between the predicted and actual values. Ideally, for a well-performing model, the residuals should be symmetrically distributed around zero, indicating no significant bias in the predictions.

In our analysis, the residual errors form a roughly normal distribution centered around zero, suggesting that the CatBoost model’s errors are unbiased. The presence of a few outliers, indicated by residuals at the far ends of the distribution, suggests that while the model performs well on average, it may struggle with certain extreme cases. Nonetheless, the majority of errors are close to zero, which is indicative of high model accuracy.

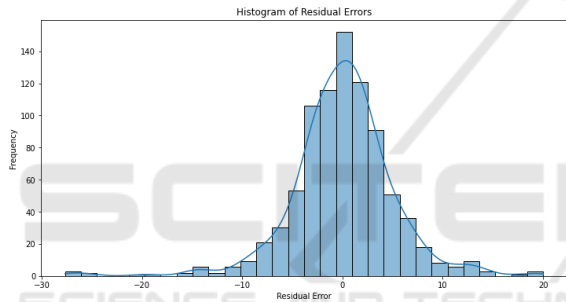


Figure 3: Histogram of Residual Errors.

### 5.5.2 Actual vs. Predicted Values

The actual vs. predicted values plot (Figure 4) provides a scatter plot of actual values against the model’s predictions. The ideal outcome is for all points to lie on the diagonal line where the predicted values equal the actual values. Deviations from this line represent prediction errors.

In our results, most points are clustered around the diagonal line, indicating that the model’s predictions closely match the actual values. This alignment demonstrates the model’s strong predictive capabilities. The spread of points along the diagonal also highlights that the model maintains consistent performance across the range of open rates. However, a few deviations from the line confirm the presence of some prediction errors, consistent with the residual error histogram analysis.

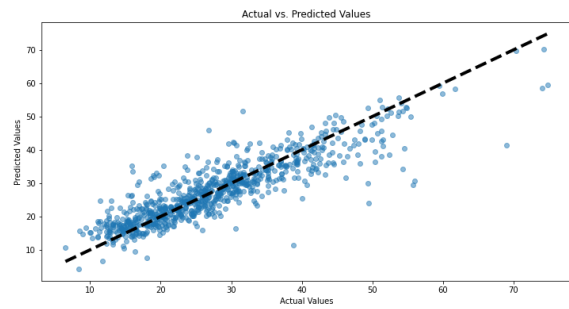


Figure 4: Actual vs. Predicted Values.

### 5.5.3 Discussion

The combination of these diagnostic plots supports the robustness and effectiveness of the CatBoost model. The symmetric distribution of residuals around zero and the close alignment of actual and predicted values demonstrate that the model has a high degree of accuracy and generalizes well to new data.

The analysis highlights that while the model performs exceptionally well overall, attention to outliers and extreme cases may provide opportunities for further refinement. These diagnostics confirm that the model’s strong performance metrics (RMSE, MAE,  $R^2$ , and MAPE) are indicative of its reliability and predictive power in real-world applications.

### 5.6 Comparison Across Quartiles

To gain deeper insights into the predictive model’s performance across different data segments, we analyzed the model by dividing the data into quartiles based on open rates. We used the *KBinsDiscretizer* to discretize the open rates into five quartiles labeled from A to E, with quartile A having the lowest and quartile E having the highest open rates. This method helps identify specific areas where the model performs better or worse, as detailed in Table 5.

Table 5: Performance Metrics Across Quartiles.

Cut	Min	Max	RMSE	MAE	MAPE
A	6.50	18.07	4.94	3.75	27%
B	18.17	23.73	4.21	2.90	14%
C	23.76	28.84	4.03	3.01	12%
D	28.85	35.84	4.31	3.10	10%
E	35.87	74.85	7.50	5.28	12%

The analysis presented in Table 5 demonstrates that the model achieves optimal performance within mid-range quartiles (B, C, and D), characterized by open rates typical of standard email campaigns. These quartiles exhibit significantly lower RMSE, MAE, and MAPE values, confirming the model’s ro-

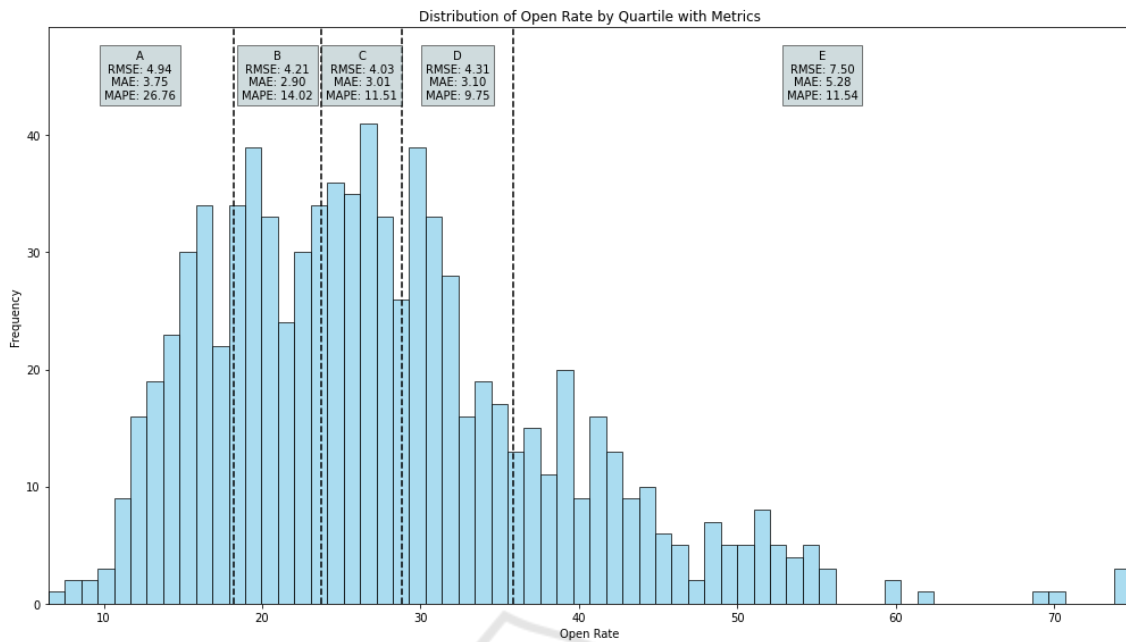


Figure 5: Distribution of Open Rate by Quartile with Metrics.

bust predictive accuracy for common campaign outcomes. In contrast, the highest quartile (E), which includes campaigns with exceptionally high open rates, presents substantial challenges, showing elevated RMSE and MAE values. This pattern suggests that extreme open rates introduce complexities that can compromise model accuracy, likely due to outlier effects or unique characteristics that are under-represented in the training data. Additionally, while the performance in quartile A, representing the lowest open rates, is relatively better than in quartile E, it is not as good as in the mid-range quartiles. This indicates that while the model handles low open rates reasonably well, there is still room for improvement in these segments.

Further illustrating the variability in model performance, Figure 5 depicts the distribution of open rates by quartile. This visual representation reinforces the findings from Table 5 and highlights the areas where the model demonstrates strong performance, also underscoring the necessity for further improvements, particularly in handling extreme values and outliers. Additionally, the distribution of campaigns is skewed towards the mid-range quartiles, B, C, and D. This is where the majority of the campaigns fall, as indicated by the higher frequency in the histogram. The model's better performance in these quartiles suggests that it is well-tuned for the most common scenarios, making it robust for practical applications.

Overall, this quartile-based analysis highlights the strengths and limitations of the model. It performs

well for the majority of campaigns but faces challenges with extreme open rates. These insights are crucial for understanding the model's applicability and for guiding future improvements, particularly in enhancing its capability to handle outliers and extreme cases more effectively.

## 6 CONCLUSION AND FUTURE WORKS

This paper applied text mining and machine learning techniques to predict email open rates from subject lines and sender, demonstrating notable predictive accuracy with the optimized CatBoost model. Enhanced by operational features and advanced text embeddings, particularly through PCA with OpenAI embeddings, the model achieved an RMSE of 5.164, MAE of 3.609,  $R^2$  of 77.53%, and MAPE of 14.73%. These results support the hypothesis that the semantic content of subject lines significantly impact open rates and highlight the importance of the sender's reputation, quantified by the *Sender Open Rate*. Although the model performs well on average open rates, it exhibits limitations with extreme values, suggesting the need of refinements to capture diverse behaviors.

Previous research emphasized the roles of curiosity and utility in subject lines (Wainer et al., 2011). This article builds on these insights by demonstrating that incorporating advanced text embeddings can significantly enhance predictive capabilities beyond

simple text features. Additionally, personalization improves open rates (Sahni et al., 2016), a finding corroborated by this research, which identifies the *Sender Open Rate* as a pivotal predictor. Moreover, this work extends previous work (Balakrishnan and Parekh, 2014) by integrating sophisticated machine learning techniques like CatBoost with operational features and PCA, thereby achieving improved predictive accuracy. Also, this paper confirms the utility of deep learning and embeddings (Joshi and Banerjee, 2023), and expands upon by illustrating that combining these embeddings with operational features effectively boosts model performance.

Future research should focus on enhancing the model's capability to handle extreme open rates, possibly through the use of more detailed datasets and advanced nonlinear modeling techniques. Exploring Large Language Models (LLMs) to refine text embeddings and their direct integration into regression tasks could markedly improve prediction accuracy and adaptability.

Optimizing this model could facilitate real-time adjustments based on campaign feedback, reducing reliance on traditional A/B testing, which is often time-intensive and costly. Such an approach enables highly personalized strategies that respond dynamically to individual recipient behaviors, optimizing ROI by enhancing the precision and effectiveness of email marketing campaigns. By merging cutting-edge machine learning technologies with email analytics, this methodology adapts to the evolving dynamics of consumer interactions in the digital era, offering significant enhancements over traditional methods.

Ultimately, this article bridges the gap between theoretical machine learning techniques and their practical application in marketing, providing a comprehensive methodology for enhancing email campaign strategies through data-driven insights, thus contributing to a more strategic and economically efficient approach to digital marketing.

## ACKNOWLEDGEMENTS

The authors thank the Pontifícia Universidade Católica de Minas Gerais – PUC-Minas and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — CAPES (CAPES – Grant PROAP 88887.842889/2023-00 – PUC/MG, Grant PDPG 88887.708960/2022-00 – PUC/MG - Informática, and Finance Code 001).

## REFERENCES

- Advisor, F. (2024). Best email marketing software 2024.
- Balakrishnan, R. and Parekh, R. (2014). Learning to predict subject-line opens for large-scale email marketing. pages 579–584.
- Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, NJ.
- Conceição, A. G. d. (2019). Main factors driving the open rate of email marketing campaigns. Dissertation for master in modelling, data analysis and decision support systems, University of Porto, Porto, Portugal.
- Etus Media Holding (2024). About Etus Media Holding.
- Feld, S., Frenzen, H., Krafft, M., Peters, K., and Verhoef, P. C. (2013). The effects of mailing design characteristics on direct mail campaign performance. *International Journal of Research in Marketing*, 30(2):143–159.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2nd edition.
- Joshi, S. and Banerjee, I. (2023). Ngram-lstm open rate prediction model (nlormp): Simple, effective, and easy to implement approach to predict open rates for marketing email. *International Journal of Digital Marketing*.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Miller, R. and Charles, E. (2016). A psychological based analysis of marketing email subject lines. In *2016 International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 58–65. IEEE.
- OpenAI (2024). Openai api.
- Paulo, M., Miguéis, V. L., and Pereira, I. (2022). Leveraging email marketing: Using the subject line to anticipate the open rate. *Expert Systems With Applications*, 207:117974.
- Sahni, N. S., Wheeler, S. C., and Chintagunta, P. K. (2016). Personalization in email marketing: The role of non-informative advertising content. *Marketing Science*, 35(2):216–233.
- Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.
- Stupar-Rutenfrans, S., de Koff, D., and van den Elst, J. (2019). The effect of subject lines on open rates of email marketing messages. *Advances in Social Sciences Research Journal*, 6(7):181–188.
- The Radicati Group, I. (2023). Email statistics report, 2023–2027.
- Wainer, J., Dabbish, L., and Kraut, R. (2011). Should i open this email? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 343–352. ACM.