EMAIL PERFORMANCE PREDICTIONS WITHOUT CAMPAIGN HISTORY

Sourabh Khot [1], Venkata Duvvuri [1], Heejae Roh [2], and Anish Mangipudi [2]

[1] College of Professional Studies, Northeastern University
[2] Langley High School, Mclean, Virginia

ABSTRACT

Email will remain a vital marketing tool in 2024. Email marketing involves sending commercial emails to a targeted audience. It currently produces a significant ROI (return on investment) in the marketing sector [1]. This research paper presents a comprehensive study on predicting email open rates, focusing specifically on the influence of subject lines. The open-rate prediction algorithm SL^k relies on the semantic features of subject lines utilizing a seed dataset of 4500 anonymized subject lines from diverse business sectors. The algorithm integrates data preprocessing, tokenization, and a custom-built repository of power words and negative words to enhance prediction accuracy. In our experiments the actual open rate margin of error was tracking close to what's allowed as per input error giving confidence that SL^k can be directionally used for optimizing subject lines performance without prior history. The findings suggest that precise manipulation of subject line features can significantly improve the efficacy of email campaigns.

KEYWORDS

Email Marketing, Open Rate Prediction, Subject Line Analysis, Machine Learning, Natural Language Processing

1. INTRODUCTION

Email marketing is a form of direct marketing that uses email to communicate commercial or fund-raising messages to an audience [1]. The subject line of an email significantly influences whether a customer will open the email. Getting someone to open an email is challenging; the key to grabbing the recipient's attention is an excellent subject line [2]. Marketing involves anticipating or enlarging the demand structure for economic goods and services and satisfying this demand through the conception, promotion, exchange, and physical distribution of such goods and services [3].

Typical open rates for retail emails can vary based on several factors, including the industry, the quality of the email list, the relevance of the content, and the effectiveness of the subject lines. However, industry benchmarks can provide a useful reference point. Here are some general statistics on open rates for retail emails [4]:

David C. Wyld et al. (Eds): DSML, ARIA, NLP, CSEN, BIBC, EDTECH - 2024
pp. 01-20, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.142401

1.1. Industry Benchmarks for Retail Email Open Rates

1. General Retail:
    - Average Open Rate: Approximately 18–25%
    - Top Quartile: 25–30% or higher
    - Bottom Quartile: Below 15%
2. E-commerce:
    - Average Open Rate: Around 15–20%
    - Top Performers: 20–25% or higher
    - Lower Performers: Below 12%
3. Fashion and Apparel:
    - Average Open Rate: About 17–23%
    - Top Quartile: 23-28% or higher
    - Bottom Quartile: Below 14%
4. Health and Beauty:
    - Average Open Rate: Roughly 19-25%
    - Top Quartile: 25–30% or higher
    - Bottom Quartile: Below 16%
5. Home and Garden:
    - Average Open Rate: Around 20–26%
    - Top Quartile: 26–32% or higher
    - Bottom Quartile: Below 18%

These open rates show enough variation to make predicting open rates with greater accuracy imperative.

The main contributions of this research are to:

- Provide accurate predictions of email open rates to marketers without prior historical data
- Offer a clear and understandable approach for predicting open rates
- Assist marketers in optimizing their email subject lines for better engagement
- Evaluate the performance of the approach on proprietary benchmark datasets

1.2. Research Problem

Accurately predicting email open rates without historical data is essential for assisting marketers in optimizing their emails. Optimized subject lines lead to better engagement with marketing campaigns and increased revenue. Email open rates measure the proportion of recipients who open a given email relative to the total number of recipients. The study aims to learn and use individual preferences and aptitudes to create more personalized learning and leverage quantitative methods to simplify the complexity of each human brain, developing optimizing methods with software engineering solutions for the future.

2. RELATED WORK

Email marketing is a mature field but has generated renewed interest with the development of AI algorithms. The literature and ChatGPT research show that open rates are the key metric to get customers interested in email marketing. Open rates are the main focus of the study without prior campaign data. Additionally, the literature is filled with studies discussing the cost, features, predictors and process optimization of email marketing but requires larger datasets.

Email serves as a major communication channel, making its optimization worthwhile for firms. In 2020, individuals sent and received an average of 306 business and consumer emails per day [5]. Batra and Keller [6] discuss the fundamentals of email marketing, emphasizing the need for new technology to optimize this important communication channel. The relative importance of subject lines is highlighted, as in item 21 in Figure 1 by [7].

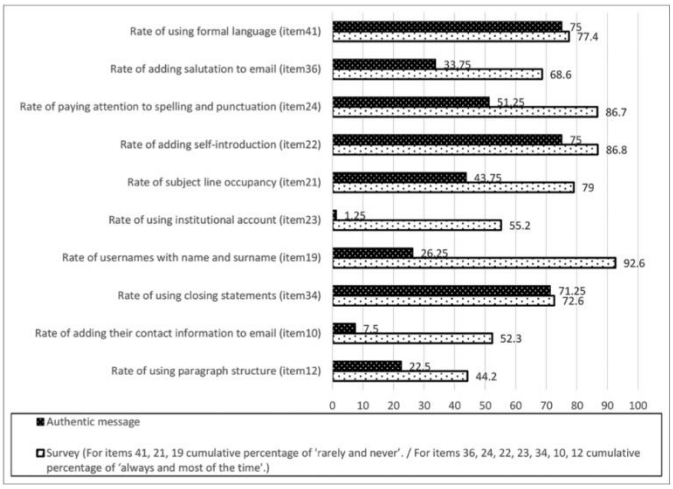


Figure 1: Importance of various attributes in emails.

Yasmin et al. [8] defined email marketing as direct digital marketing used to send ads, build brand and customer loyalty, build customer trust, and create brand awareness. The relatively low cost of email marketing compared to other forms of media exposure is a significant advantage. Companies can capture customers' attention by creating an attractive mix of graphics, text, and links for their products and services. Companies can increase sales, as illustrated by a strong correlation factor of 0.873 in Table 1.

Table 1. Correlation of Sales for email marketing.

		Increased Sales	Online Advertising	Email Marketing	Social Media	Text Messaging	Affiliate Marketing	Search Engine Optimization (SEO)	Pay Per Click (PPC)
Increased Sales	Pearson Correla-tion*(r)*	1	.869	.873	.840	.667	.560	.840	.560
	Sig. (2-tailed)		.056	.053	.075	.219	.326	.075	.326
	N	150	150	150	150	150	150	150	150

De Bruyn et al. [9] identified the pitfalls of not using Artificial Intelligence (AI) in marketing. AI offers opportunities to optimize and leverage new techniques for better marketing. Automating several manual tasks in marketing is crucial for improving efficiency. Trainor et al. [10] highlight the importance of technology-driven optimization for email marketing, which is vital for achieving firm objectives such as customer retention and satisfaction. Trainor et al.'s study integrated Information Technology (IT) with marketing capabilities and examined firm success metrics, leading to the development of AI-driven technology for email marketing optimization. Additionally, Trainor et al. pointed out that email marketing is important in a highly competitive environment, as shown in Figure 2.

Bala and Verma [11] conducted a survey on current and future trends in digital marketing, including email marketing. The research analyzed the effectiveness of email marketing across various domains. They noted that technology-driven segmentation of customer preferences can improve efficacy. However, the study did not address the subject line performance optimization using IT.

Rettie [12] reviewed email marketing literature to identify key advantages. They listed several measures that increase response rates in direct marketing and direct mail. Rettie conducted qualitative research among industry experts and analyzed various email marketing campaigns to identify factors associated with higher response rates. Key factors include subject line, email length, incentives, and the number of images, along with demographic and lifestyle data. These findings were used to create an email marketing process model based on [13].

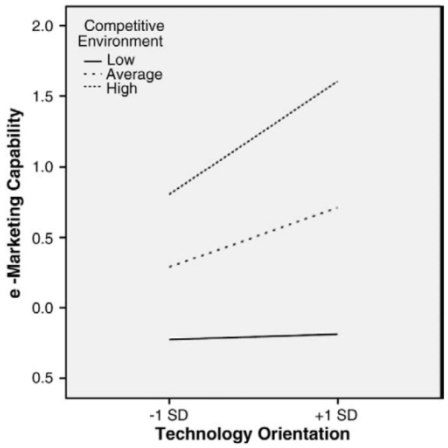


Figure 2. Importance of competition in email marketing prevalence

Balakrishnan and Parekh [14] introduced an approach for predicting email subject line open rates in large-scale email marketing campaigns. They leveraged syntactical, historical, and derived features from over a hundred thousand subject lines. Using a random forest model, they significantly enhanced prediction accuracy beyond traditional baselines. Their work highlights the unique challenges of email marketing and proposes a sophisticated model to address these complexities. That is, it emphasizes the potential of machine learning in optimizing email marketing strategies.

Paulo et al. [15] examined how subject lines influence open rates, employing machine learning models on a dataset of 140,000 emails. Structural and content features of subject lines were found to significantly affect open rates. The study utilized various machine learning techniques, including Random Forests, Decision Trees, and Neural Networks, to predict email open rates. These techniques contributed valuable insights for optimizing email campaigns for higher engagement.

Inspired by previous research, a repository of keywords was created. Each keyword ranged from 0.1 to 1.0 according to campaign types such as blast awareness, engagement, and highly targeted campaigns. Based on a threshold of 0.5, weights above 0.5 were classified as power words, and weights below 0.5 as negative words. This algorithm calculated the open rate for subject lines.

Direct marketing, the parent category of email marketing, involves reaching customers directly through emails or coupons and soliciting responses.

Predictive models have measured the response of such campaigns as shown in [16, 17]. LSTM models were used on multidimensional datasets with features like channel sources, websites, demographics, and more to predict response rates [16]. However, these models require large amounts of response data, which can be a limitation for businesses without such data.

Feature engineering is a crucial responsibility of AI models operating on large datasets, such as in marketing domains [18]. Feature engineering methods help build suitable predictors for predictions, although they can be cumbersome and elaborate. This research study avoids this complexity.

Artificial Intelligence models are often black boxes as they pose significant limitations when customers require explanations. There are significant challenges of black box models in high-stakes decision-making [19]. Although some argue that creating explanations for black box models is feasible, it is not easily achievable. Therefore, this study developed a white-box, rule-based algorithm based on keywords for the purpose, which is simple to understand and explain.

Fariborzi el al. [20] emphasized the importance of focusing on subject lines to grab recipients' attention. Marketing involves anticipating or enlarging the demand structure for economic goods and services and satisfying this demand through the conception, promotion, exchange, and physical distribution of such goods and services [3]. To that end, effective subject lines facilitate the exchange of goods.

Content marketing for emails follows guidelines similar to those for social media content marketing [21]. While content optimization is crucial for email performance, subject lines serve as the first gateway to emails. This first-door importance makes subject line optimization the first high-priority in the email marketing process. The complete literature survey analysis on dimension is outlined in Table 2.

Table 2: Literature Survey Analysis on Dimensions.

Dimension	**Papers**	**Pros**	**Cons**
Volume	Konuk (2021), Batra and Keller (2016)	Importance of emails	No methods to optimize
Cost	Yasmin et al. (2015)	Low cost of email marketing	No methods to optimize
Technology	De Bruyn et al. (2020), Trainor et al. (2011)	Optimizing email marketing with technology	No specific models or insights
Attributes	Bala and Verma (2018)	Predictors for improvement in responses	Geo-specific
Factors	Rettie (2002), Vriens et al. (1998)	Several email attributes contributing to improvement in response rates	Process dissection and no recommendations

Subject lines	Balakrishnan and Parekh (2014), Paulo et al. (2022), Fariborzi and Zahedifard (2021)	ML models for performance	Data-driven and feature engineering needed
AI	Sarkar and De Bruyn (2021), Campbell et al. (2020)	AI models for direct response modeling	Needs large data
Data	Kuhn and Johnson (2016), Rudin (2019)	Data engineering and limitations	No alternative approaches suggested
Content	Chaffey et al. (2013)	Content optimization like social media	No subject line optimization

3. METHODOLOGY

This section describes the data sources, tools, techniques, and steps used to develop our solution **SL^k**. The methodology consists of four main steps: preprocessing, tokenization, repository building, and open-rate prediction.

3.1. Methodology

We use various techniques, including data analysis, natural language processing, and machine learning, to develop our solution. The techniques we use include:

- Data curation: Several experts's commentaries and blogs were curated to identify high performing subject line seed keywords
- Data exploration: A technique that involves examining and summarizing the data structure, statistics, and distribution using descriptive methods and visualizations.
- Data preprocessing: A technique that involves transforming the data into a standard and clean form by applying various techniques, such as lowercasing, removing punctuation marks, special characters, emojis, personalization codes, numbers, and stopwords, and lemmatizing words.
- Data analysis: A technique that involves extracting features and insights from the data using various techniques, such as word weight assignment and campaign type classification.
- Data visualization: A technique that involves presenting the data analysis results using various techniques, such as word clouds, scatter plots, box plots, histograms, and bar charts.

3.1.1. Steps

After data curation, we follow four main steps to develop our solution:

- Preprocessing: We preprocess the subject lines using various techniques to transform them into a standard and clean form.
- Tokenization: We tokenize the subject lines using nltk.word_tokenize function to break them down into smaller units called tokens.
- Repository building: We build a repository of power words and assign word weight and campaign type based on our subject matter expertise. The word weight (0.0–0.1) represents the impact of the power word in the subject line to improve the email open rate based on the campaign type.
- Open rate prediction: We predict the open rate for each subject line using feature extraction and model selection techniques to classify them into five quality levels based on their potential open rate.

3.1.2. Tools

We use Python as the programming language to develop our solution. Python is a high-level programming language that supports multiple paradigms and has a rich set of libraries and modules for data analysis, machine learning, and natural language processing. We use the following Python libraries and modules in our solution:

- pandas: A library that provides data structures and tools for data manipulation and analysis.
- numpy: A library that provides numerical computing and linear algebra functions.
- matplotlib: A library that provides plotting and visualization functions.
- seaborn: A library that provides statistical data visualization functions.
- nltk: A module that provides natural language processing tools and resources.
- sklearn: A module that provides machine learning tools and algorithms.
- wordcloud: A module that provides word cloud generation functions.

4. SEED KEYWORDS CURATION

We use a large dataset of 4,500 anonymized subject lines *seed* datasets from various sources as listed below, including past email campaigns from various business sectors. The s*eed* dataset is divided into three subsets:

- Optinmonster: A subset of 100 subject lines from Optinmonster, a website that provides tools and resources for email marketing optimization.
- Wordstream.com: A subset of 200 subject lines from Wordstream.com, a website that provides online advertising and marketing solutions.
- Anonymized database: A random subset of 500 subject lines from sent email lists from various domains and sectors, such as finance, consumer goods, and health.

The seed dataset covers a wide range of topics, styles, and formats of subject lines, as well as different levels of open rates, sender information, and campaign types. Table 1 shows some descriptive statistics of the seed dataset.

Table 3. Descriptive statistics of the seed dataset.

Data Source	Number of Subject Lines	Average Number of Words	Average Number of Characters
Optinmonster	100	6.22	38.54
Wordstream.com	200	8.09	49.63
Anonymized Database	500	6.84	41.89
Total	800	6.85	41.92

5. OPEN RATE MODEL SL^K

5.1. Keyword-Based Solution (SL^k)

The SL^k subject line prediction algorithm consists of three main steps: preprocessing, repository building, and open-rate prediction algorithm.

5.1.1. Preprocessing

The tokenization step broke down the subject lines into smaller units called tokens, which are usually words or punctuation marks. Tokenization was done using nltk.word_tokenize function.

The preprocessing steps involved transforming the tokens into a standard and clean form. Techniques incorporated include lowercasing; removing punctuation marks, special characters, emojis, personalization codes, numbers, and stopwords; and lemmatizing words. The preprocessing step also involved identifying the language of each subject line using the langdetect library.

Preprocessing reduced the number of tokens in the data and removed irrelevant or noisy tokens that might affect the analysis. Table 2 shows the number of tokens obtained after each preprocessing step for each data source.

Table 4. Summary of tokens during Natural Language Processing Steps.

Subject Lines	All Tokens	Normalized Tokens	Tokens After Preprocessing	Tokens After Stopwords Removal	Tokens After Lemmatization
800	7016	7016	4642	2913	2913

The tokenization and processing steps allowed us to analyze the frequency and distribution of words in the seed data and identify the most common and influential words for subject line performance. The word cloud library allowed us to generate word clouds from the data, which are graphical representations of word frequency that give greater prominence to frequently appearing words. Figure 1 shows the word clouds generated.

Figure 3. Seed dataset of Subject lines keywords word cloud

5.1.2. Repository Building

The repository-building step involved creating a collection of 200 power words that can enhance subject line effectiveness based on word weight and campaign type. Power words are terms that can trigger emotional or psychological responses in the recipient, such as curiosity, urgency, or excitement.

The repository-building step consisted of two steps: word weight assignment and campaign type classification.

5.1.2.1. Word Weight Assignment

We assigned a numerical value to each word based on its influence or impact on subject line performance, defined by the open rate. Word weight was assigned using the average open rate of the subject lines that contain the word, the maximum open rate in the dataset, and our domain knowledge. This step allowed us to rank the words according to their importance and effectiveness for subject line performance.

5.1.2.2. Campaign Type Classification

For campaign type classification, we assigned a categorical value to each word based on its relevance or suitability for different types of email campaigns, such as blast awareness, engagement, or highly targeted campaigns. Campaign type was determined using domain knowledge of power words for each category. This step allowed us to categorize the words according to their purpose and effect for different email marketing goals and strategies.

5.1.3. Open Rate Prediction Algorithm SL^k

The open rate of an email marketing campaign depends primarily on the sender and their relationship with their recipients. To predict the open rate of a given subject line, we used the mean (μ) and s allowed margins (SD) of open rates from the sender's past campaigns for their target audience. Our algorithm **SL^k** takes 'μ' as the baseline open rate and then predicts the expected open rate based on the given subject line and the SD.

The algorithm predicts based on the number of power words and the weights of the power words from the previously built repository. We calculated the mean and s allowed margins SD based on similar campaigns run by the client. The open rate prediction step consisted of the following algorithm:

a. One power word → Assign predict rate as μ + 1SD
b. Two power words → Assign predict rate as μ + 2SD
c. One negative word → Assign predict rate as μ - SD
d. Two negative words → Assign predict rate as μ - 2SD

For a given campaign, power words are identified as words belonging to that campaign having an SM weight of more than 0.5. Similarly, negative power words are campaign words having an SM weight of less than 0.5.

Thus, this algorithm takes the subject line 'μ' and 'SD' as inputs, refers to the repository, and computes the predicted open rate.

6. EXPERIMENTS

This section covers the testing and experimentation of the algorithm with actual subject lines.
We built the algorithm in Python and hosted it on the cloud along with the repository. APIs were created so that the algorithm can be accessed using the internet. POST requests could be made to provide the

inputs of the subject line and mean and allowed margins in JSON format and get predicted open rate as output.

6.1. Dataset 1

6.1.1. Overview

A total of 2,142 past subject lines from retail sector of a proprietary dataset were used to test the algorithm. The mean open rate for each subject line was available, whereas three iterations with s allowed margins from 0.3, 0.5, and 1 were considered for this experimentation.

Each subject line, along with its mean open rate, was fed into the algorithm API via POST requests with the three allowed margins. The outputs of the predicted open rate and duration to compute were recorded. This data will be analyzed in the next section.

There are a total of 6,234 rows and 5 attributes. Three allowed margins (from 0.3, 0.5, and 1) for each subject line were requested through the API algorithm and the output was recorded. The attributes are as follows:s

1. predicted_open_rate: calculated based on the algorithm
2. subject_line: Given 2,142 subject_lines * 3
3. allowed_margin: 0.3, 0.5, and 1.0 for each subject_line
4. given_open_rate: mean open rate for each subject line
5. duration: time taken for the server to respond

6.1.2. Data Processing

A total of 192 (2.99%) data had errors while getting a response from the server. We dropped these 192 rows instead of imputing them. Because subject_line is the most important key attribute in this analysis, three columns were added so that subject_line can be analyzed in the form of numerical values. This allowed us to analyze how predicted_open_rate is related to the numerical traits of the subject line.

Table 5. Head(5) of Cleaned Data.

	predicted_open_rate	subject_line	allowed_margin	given_open_rate	duration	sl_length	punctuation_count	upper_case_proportion
1	12.64	WHOA! Up to 65% off.	0.3	12.79	6.33	20	2	27.8
2	12.54	WHOA! Up to 65% off.	0.5	12.79	22.40	20	2	27.8
3	12.29	WHOA! Up to 65% off.	1.0	12.79	22.40	20	2	27.8

4	13.28	WHOA. $29.95 and under sandals are here!	0.3	13.43	6.32	42	2	10.0
5	13.18	WHOA. $29.95 and under sandals are here!	0.5	13.43	19.28	42	2	10.0

The parameter 'sl_length' can determine how the length of the subject line affects the predicted_open_rate. Similarly, 'punctuation_count' allows you to see how the number of punctuations included in subject_line is related to predicted_open_rate. In this process, we excluded the periods (.) and dashes (–), which are commonly used in sentences. The parameter 'upper_case_proportion' captures the proportion of uppercase letters, as we have cases where the same sentence is expressed in uppercase and lowercase, such as 'FINAL CALL for $5 off!' and 'Final call for $5 off!' Likewise, 'upper_case_proportion' was added to understand how predicted_open_rate appears in uppercase and lowercase.

6.1.3. Results

We performed experiments on a total of 6,234 rows and 8 columns. The results are as shown in the table below.

Table 6. Average Experimental results on dataset1 with allowed margins=0.3,0.5,1.0

	predicted _open_rate	**alllowed_ margin**	**input actual _open_rate**	**duration**	**sl_length**	**punctuation _count**	**upper_case _proportion**
count	6234.00	6234.00	6234.00	6234.00	6234.00	6234.00	6234.00
mean	**17.88**	**0.60**	**17.28**	20.42	32.55	1.67	12.61
std	8.47	0.29	8.45	22.59	10.06	1.52	13.65
min	0.00	0.30	0.00	4.92	11.00	0.00	0.00
25%	13.96	0.30	13.48	5.99	26.00	1.00	3.45
50%	16.12	0.50	15.54	18.32	32.00	1.00	6.45
75%	18.96	1.00	18.30	23.08	39.00	2.00	17.65
max	91.37	1.00	89.37	373.96	111.00	16.00	90.91

The mean of 'given_open_rate' is 17.88 and 'predicted_open_rate' is 17.28. The average error margins calculated difference of predicted against given was 0.6 (rounded) and average allowed margin input was 0.6. Thus the actual margin of error was tracking close to what's allowed as per input error giving confidence that SL^k can be directionally used for optimizing subject lines performance without prior history.

Additionally, the response time from the server measured by 'duration' averaged 20.42 seconds, with a maximum response time of 373 seconds. 'sl_length' recorded a mean of 32.55 including spaces, with the shortest sentence being 11 characters (min) and the longest 111 characters (max). punctuation_count recorded a mean of 1.67, with a minimum of 0 and a maximum of 16. Likewise, 'upper_case_proportion' had a mean of 12.61, indicating that normal cases involved sentences with only the first letter capitalized.

The max of ‘upper_case_proportion’ was 90.91, which means most letters except blank spaces were written in uppercase.

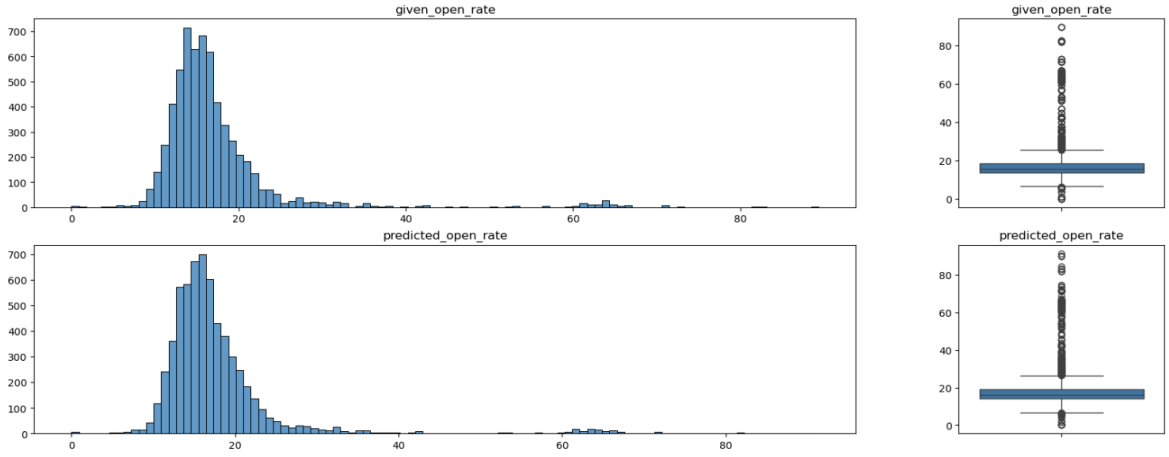


Figure 4. Dataset1 prediction results graphs

The parameters ‘given_open_rate’ and ‘predicted_open_rate’ show almost similar overall distribution. Both data are gathered around 17 (median of ‘given_open_rate’ is 15.54 and ‘predicted_open_rate’ is 16.12). The detailed distribution of data around 17 shows a slight transformation depending on the application of allowed_margin in the algorithm. Both data sets are right-skewed because ‘predicted_open_rate’ is mainly determined by ‘given_open_rate’ in the algorithm.

‘sl_length’ has a median of 32 and is relatively symmetrical. Many observations higher than 39 (Q3) appear right_skewed. Similarly, ‘punctuation_count’ has a median of 1 and is relatively concentrated between 0 and 2, with Q3 also at 2. The median of ‘upper_case_proportion’ is 6.45, with Q1 at 3.45 and Q3 at 17.65. Most graphs show right_skewed data because the median is small; however, the maximum number is comparatively large.

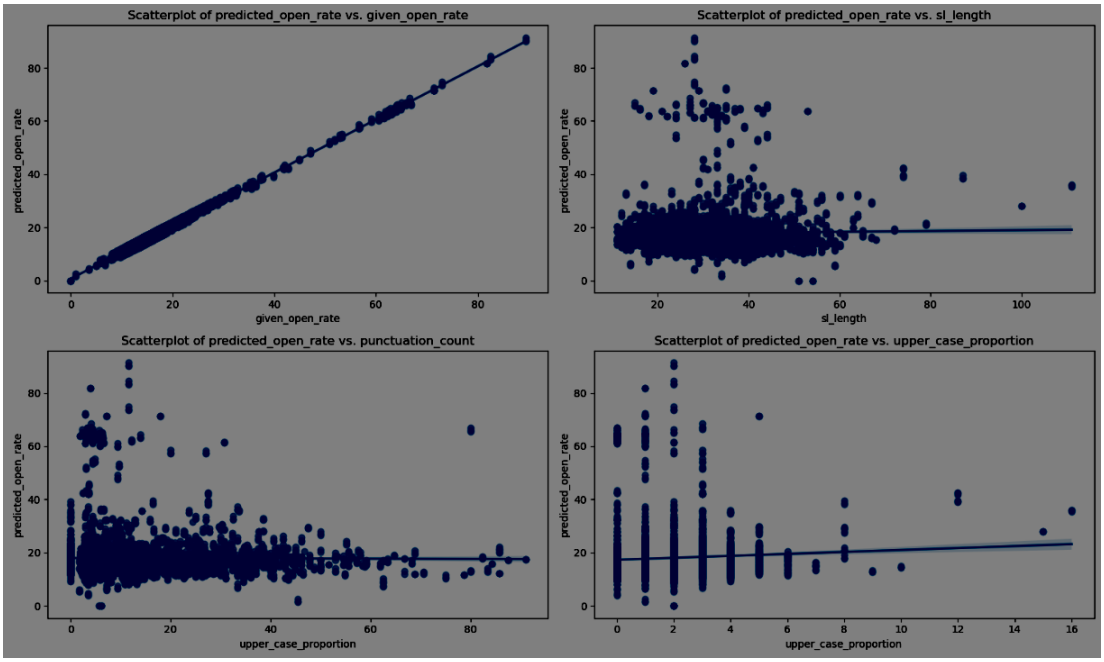
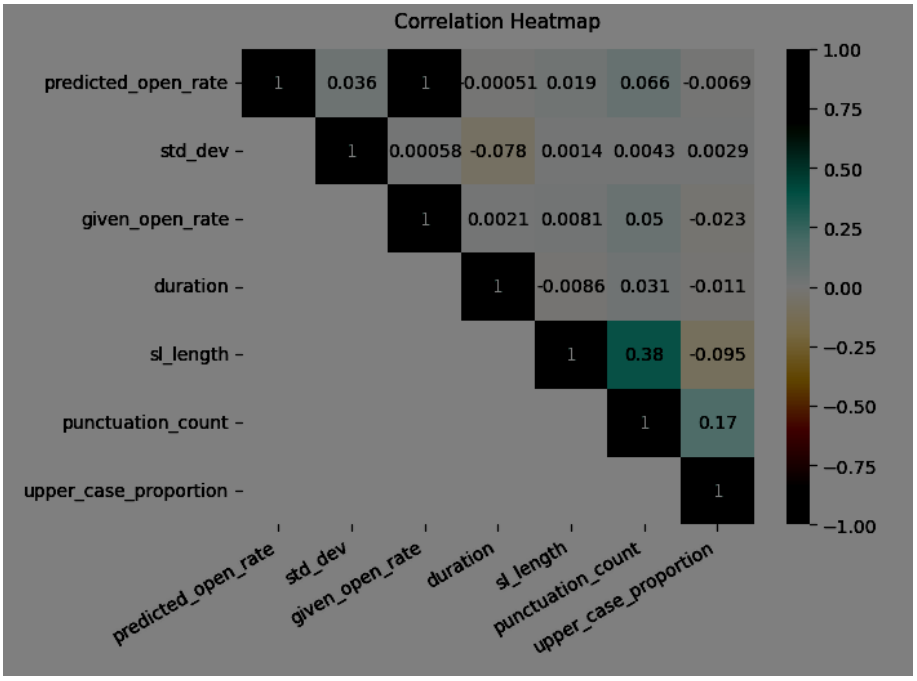


Figure 5. Dataset1 subject line length prediction relationship

The four graphs above represent the linear relationship between ‘predicted_open_rate’ and other numerical variables. In the upper left, ‘given_open_rate’ shows the scatter plot almost on the same line as ‘predicted_open_rate’, indicating high linearity and a strong standard correlation. Notably, ‘sl_length’, ‘upper_case_proportion’, and ‘punctuation_count’ have lower slopes than ‘given_open_rate’ but show a linear relationship to some extent. ‘sl_length’ and ‘punctuation_count’ show similar scatter plots.



The parameters ‘predicted_open_rate’ and ‘given_open_rate’ show a very high correlation of 1. This aligns with the almost perfect linearity seen in the scatter plot. Other variables that show a high correlation with ‘predicted_open_rate’ are punctuation_count (0.066), allowed_margin (0.036), and upper_case_proportion (-0.0069) in order.

6.2. Dataset 2

The second test case was conducted over a non profit organization email campaign subject lines, each comprising three iterations with varying allowed margins. The first set (test case 2.1) featured allowed margins of 1, 3, and 5, applied to three rows iteratively. The second set (test case 2.2) used allowed margins of 5, 7, and 10, applied to 10 subject lines, to assess their impact on the open rate prediction algorithm. The algorithm was tested using a total of 13 past subject lines, for which the mean open rates were available.

As with Experiment 1, each subject line, with its given open rate, was inputted into the algorithm API using POST requests, structuring the data to include the three allowed margins.

A total of 39 rows were generated with 5 attributes, with each subject line producing three results for the different allowed margins. The attributes remained consistent with those in Test Case 1, and we applied the same data preprocessing procedures.

6.2.1. Analysis

We performed descriptive analysis on a total of 39 rows and 8 columns. The results are shown in the table below:

1. predicted_open_rate: calculated based on the algorithm
2. subject_line: Given 2,142 subject_lines * 3
3. allowed_margin: 1, 3, and 5 for each subject_line
4. given_open_rate: mean open rate for each subject line
5. duration: time taken for the server to respond
6. sl_length: records the total length of the subject line
7. punctuation_count: number of punctuations, such as '!', '?', and '%', excluding dashes (–) and periods (.)
8. upper_case_proportion: number of uppercase letters/(sl_length - punctuation_count) * 100

Table 7. Whole rows of Cleaned Data with allowed margin = 1,3,5

	predicted_open_rate	subject_line	allowed margin	given_open_rate	duration	sl_length	punctuation_count	upper_case_proportion
1	53.4	Milaap '23 Campaign	1	53.4	7.62	19	1	11.1
2	53.4	Milaap '23 Campaign	3	53.4	7.40	19	1	11.1
3	53.4	Milaap '23 Campaign	5	53.4	8.22	19	1	11.1
4	44.3	Thanks Repeat donors!	1	43.3	7.75	21	1	10.0
5	46.3	Thanks Repeat donors!	3	43.3	7.52	21	1	10.0
6	48.3	Thanks Repeat donors!	5	43.3	7.48	21	1	10.0
7	48.3	C4PH's grant to Christian Bulan	1	48.3	7.83	31	1	16.7
8	48.3	C4PH's grant to Christian Bulan	3	48.3	7.70	31	1	16.7

9	48.3	C4PH's grant to Christian Bulan	5	48.3	7.91	31	1	16.7

Given only 3 rows for Set 1, we directly checked the results of the individual datasets, bypassing the descriptive analysis. In two subject lines, the allowed margins did not influence predicted_open_rate. This lack of effect can be attributed to the absence of either power words or negative words in the subject lines.

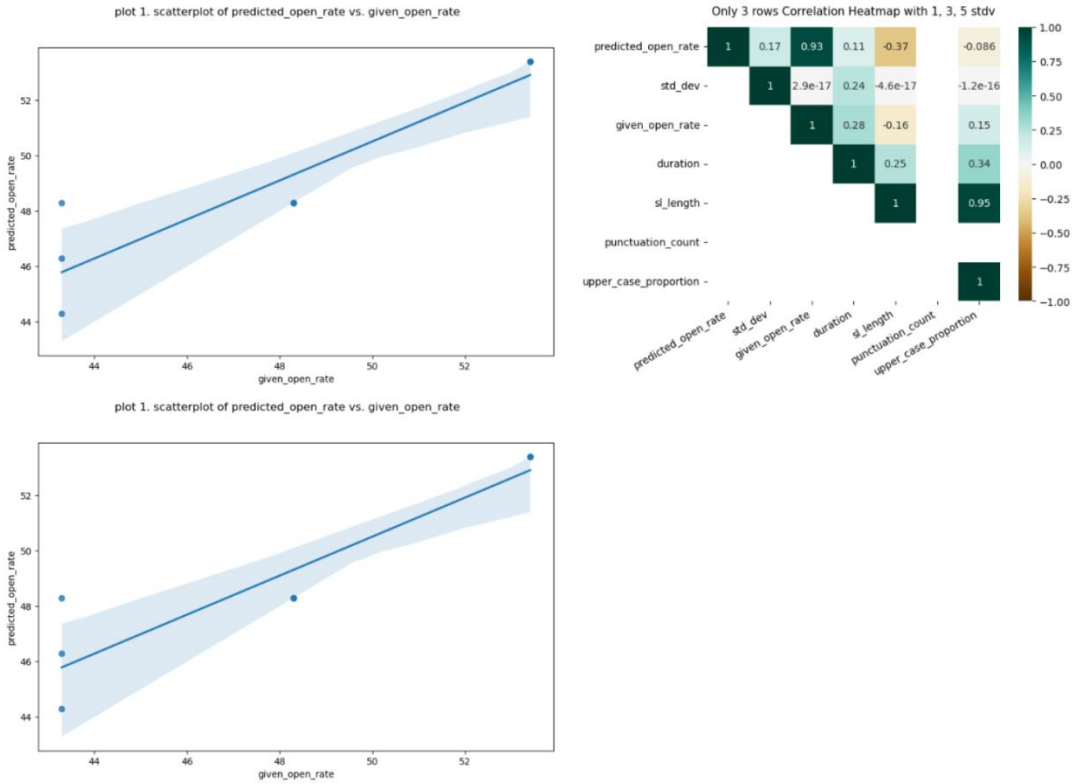


Figure 7. Dataset2 iteration 1 predictions correlation with actuals

Despite the small dataset, the graph on the left showed that the parameters predicted_open_rate and given_open_rate exhibit a linear relationship. Structurally, this occurred because the allowed margins is either added to or subtracted from given_open_rate, depending on whether a power word or negative word is present. It can be inferred that averaging predicted_open_rate for each subject line will establish a linear correlation with given_open_rate.

6.2.2. Results

The results of the dataset2 iteration 1 is tabulated as hereunder:

Table 8. Head (9) of Cleaned Data with with allowed margins=5,7,10

	predicted_open_rate	subject_line	allowed margin	given_open_rate	duration	sl_length	punctuation_count	upper_case_proportion
1	52.5	2022 year end newsletter (spring)	5	52.5	7.60	33	2	0.00
2	52.5	2022 year end newsletter (spring)	7	52.5	8.12	33	2	0.00
3	52.5	2022 year end newsletter (spring)	10	52.5	7.49	33	2	0.00
4	54.3	2021 year end newsletter	5	59.3	7.86	24	0	0.00
5	52.3	2021 year end newsletter	7	59.3	7.63	24	0	0.00
6	49.3	2021 year end newsletter	10	59.3	7.50	24	0	0.00
7	65.2	ICC Joola Summer Tournament awareness (last)	5	55.2	7.74	44	2	14.29
8	69.2	ICC Joola Summer Tournament awareness (last)	7	55.2	7.74	44	2	14.29
9	75.3	ICC Joola Summer Tournament awareness (last)	10	55.2	7.56	44	2	14.29

Observing the first nine rows, which represent the results for three subject lines, it is evident that as the size of the allowed margins increases, predicted_open_rate also fluctuates significantly. The interval of three between the allowed margins of 7 and 10 in the sequence of 5, 7, and 10 is wider than the increments of 1, 3, and 5 used in Test Case 2.1. The predicted_open_rate value exhibits greater fluctuations. This variation reflects the influence of power words or negative words more markedly due to the larger intervals.

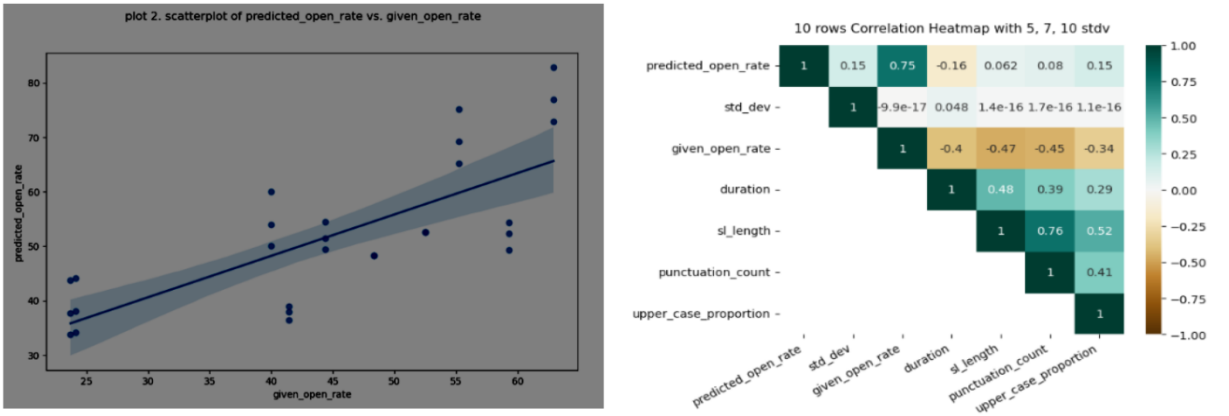


Figure 8. Dataset2 iteration 2 predictions correlation with actuals

Despite the limited size of the dataset, with only 10 rows, the linearity between predicted_open_rate and given_open_rate is again evident in the scatter plot. The correlation heatmap on the right displayed values of 0.75 and 0.15, respectively, in predicted_open_rate vs. given_open_rate and allowed margins. These results contrast with those from Test Case 1, where s allowed margins were set at 0.3, 0.5, and 1, yielding correlation values of 1 for 'given_open_rate' and 'predicted_open_rate' and 0.036 for 'allowed_margin' and 'predicted_open_rate'. The correlation for given_open_rate has significantly decreased in Test Case 2.2 compared to the 0.93 and 0.17 observed in Test Case 2.1. Although the gaps between allowed margins were only slightly wider, the correlation levels remained nearly constant. Further research is needed to verify the impact of different allowed margins on larger datasets in the future. We did not calculate mean error margins with actuals due to small dataset size and lack of statistical significance thereafter.

The research on predicting email open rates based on subject lines yielded several key findings:

- Subject Line Importance: The subject line significantly influences whether recipients open an email. Crafting an excellent subject line is crucial for improving open rates.
- Algorithmic SL^k Approach: We proposed an algorithm leveraging statistical techniques and domain knowledge with features like word weight, campaign type, and subject line characteristics to predict open rates accurately.
- High Correlation: The predicted open rate closely aligns with the historical given open rate, indicating a strong correlation between the two.
- Subject Line Attributes: Factors such as subject line length, punctuation count, and uppercase proportion impact the predicted open rate.

7. CONCLUSION

This research provides a comprehensive analysis of the factors influencing email open rates, with a particular focus on the impact of subject lines. Through the utilization of a seed keyword dataset and the application of natural language processing and machine learning techniques, the study offers predictive model for open rate prediction without any prior campaign data. The SL^k approach of creation of a keyword repository and the development of an algorithm for predicting open rates based on subject line features represent significant contributions to the field of email marketing analytics. In our experiments the actual open rate margin of error was tracking close to what's allowed as per input error giving confidence that SL^k can be directionally used for optimizing subject lines performance without prior history. These findings have practical implications for marketers seeking to optimize their email

campaigns for better engagement and can serve as a foundation for further studies in this domain. Ultimately, this paper lays out the importance of keyword-data-driven strategies in enhancing the performance of email marketing initiatives. In the experiments the actual margin was tracking close to what's allowed giving confidence that SL^k can be directionally used for optimizing subject lines performance without prior history.

REFERENCES

[1] Dave, Chaffey & P. R., Smith, (2013) *eMarketing eXcellence: Planning and optimizing your digital marketing*. Routledge. DOI: https://doi.org/10.4324/9780203082812.

[2] Jacquelyn, Smith, (2016, June 30) "The perfect way to write an email subject line, and 9 mistakes to avoid". *Business Insider*. Retrieved from https://www.businessinsider.com/perfect-way-to-write-email-subject-line-2016-6. Accessed: Jan. 6, 2024.

[3] Elham, Fariborzi & Eng Morvarid, Zahedifard, (2012) "E-mail marketing: Advantages, disadvantages and improving techniques". *International Journal of e-Education, e-Business, e-Management and e-Learning*, Vol. 2, No. 3, p232. Retrieved June 1, 2024 from https://www.researchgate.net/publication/269838654_E-mail_Marketing_Advantages_Disadvantages_andImproving_Techniques.

[4] Gary, J. Brunswick, (2014) "A chronology of the definition of marketing". *Journal of Business & Economics Research (Online)*, Vol. 12, No. 2, p105. DOI: 10.19030/jber.v12i2.8523.

[5] OpenAI, (2024) ChatGPT (May 21 version) [Large language model]. Available: https://chat.openai.com/chat. Accessed: Jan. 6, 2024. Sumeyye, Konuk, (2021) "E-Mail literacy in higher education academic settings", *International Journal of Education and Literacy Studies*, Vol. 9, No. 3, pp29-42. DOI: 10.7575/aiac.ijels.v.9n.3p.29.

[6] Rajeev, Batra & Kevin, Lane Keller, (2016) "Integrating marketing communications: New findings, new lessons, and new ideas". *Journal of Marketing*, Vol. 80, No. 6, pp122-145. DOI: https://doi.org/10.1509/jm.15.041.

[7] Konuk, Sümeyye. "E-Mail Literacy in Higher Education Academic Settings." *International Journal of Education and Literacy Studies* 9.3 (2021): 29-42.

[8] Afrina, Yasmin, Sadia, Tasneem & Kaniz, Fatema, (2015) "Effectiveness of digital marketing in the challenging age: An empirical study". *International Journal of Management Science and Business Administration*, Vol. 1, No. 5, pp69-80. DOI: 10.18775/ijmsba.1849-5664-5419.2014.15.1006.

[9] Arnaud, De Bruyn, Vijay, Viswanathan, Yeah Shan, Beh, Juergen, Brock & Florian, Von Wangenheim, (2020) "Artificial intelligence and marketing: Pitfalls and opportunities". *Journal of Interactive Marketing*, Vol. 51, No. 1, pp91-105. DOI: https://doi.org/10.1016/j.intmar.2020.04.007.

[10] Kevin, J. Trainor, Adam, Rapp, Lauren Skinner, Beitelspacher & Niels, Schillewaert, (2011) "Integrating information technology and marketing: An examination of the drivers and outcomes of e-marketing capability". *Industrial Marketing Management*, Vol. 40, No. 1, pp162-174. DOI: https://doi.org/10.1016/j.indmarman.2010.05.001.

[11] Madhu, Bala & Deepak, Verma, (2018). "A critical review of digital marketing". *International Journal of Management, IT & Engineering*, Vol. 8, No. 10, pp321-339. Retrieved from https://www.scirp.org/reference/referencespapers?referenceid=2929876 "Retrieved on 06/01/2024".

[12] Ruth, Rettie, (2002) "Email marketing: success factors". Retrieved from https://eprints.kingston.ac.uk/id/eprint/2108/1/paper.html "Retrieved on 06/01/2024".

[13] Marco, Vriens, Hiek R., Van der Scheer, Janny C., Hoekstra & Jan Roelf, Bult, (1998) "Conjoint experiments for direct mail response optimization". *European Journal of Marketing*, Vol. 32, No. (3/4), pp323-339. DOI: https://doi.org/10.1108/03090569810204625.Raju, Balakrishnan & Rajesh, Parekh, (2014). "Learning to predict subject-line opens for large-scale email marketing". *2014 IEEE International Conference on Big Data (Big Data)*, pp. 579-584. DOI: 10.1109/BigData.2014.7004277.

[14] Balakrishnan, Raju, and Rajesh Parekh. "Learning to predict subject-line opens for large-scale email marketing." *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, 2014.

[15] M., Paulo, V. L., Miguéis & Ivo, Pereira, (2022) "Leveraging email marketing: Using the subject line to anticipate the open rate". *Expert systems with applications*, Vol. 207, No. 117974. DOI: https://doi.org/10.1016/j.eswa.2022.117974.

[16] Mainak, Sarkar & Arnaud, De Bruyn, (2021) "LSTM response models for direct marketing analytics: Replacing feature engineering with deep learning". *Journal of Interactive Marketing*, Vol. 53, No. 1, pp80-95. DOI: https://doi.org/10.1016/j.intmar.2020.07.002.

[17] Colin, Campbell, Sean, Sands, Carla, Ferraro, Hsiu-Yuan (Jody), Tsao & Alexis, Mavrommatis, (2020) "From data to action: How marketers can leverage AI". *Business horizons*, Vol. 63, No. 2, pp227-243. DOI: https://doi.org/10.1016/j.bushor.2019.12.002.

[18] Max, Kuhn & Kjell, Johnson, (2019) *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC. DOI: https://doi.org/10.1201/9781315108230. Marcel, Sattler, (2021) "Why email marketing is crucial for businesses". Retrieved from https://www.forbes.com/sites/forbesagencycouncil/2021/05/03/why-email-marketing-is-crucial-for-businesses/?sh=f04b6b875a2d. Accessed: Jan. 6, 2024.

[19] Cynthia, Rudin, (2019) "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". *Nature Machine Intelligence*, Vol. 1, No. 5, pp206-215. DOI: https://doi.org/10.1038/s42256-019-0048-x.

[20] Fariborzi, Elham, and Morvarid Zahedifard. "E-mail marketing: Advantages, disadvantages andimproving techniques." *International Journal of e-Education, e-Business, e-Management and e-Learning* 2.3 (2012): 232.

[21] Andreia, Conceição & João, Gama, (2019) "Main factors driving the open rate of email marketing campaigns". *Discovery Science: 22nd International Conference, DS 2019, Split, Croatia, October 28–30, 2019, Proceedings 22*, Springer International Publishing, pp. 145-154. DOI: https://doi.org/10.1007/978-3-030-33778-0_12

AUTHORS

Sourabh Khot is a budding data scientist with a passion for data-driven decision-making. Based in Mountain View, CA, USA, Sourabh has a strong background in analytics, BI, and strategy. In his most recent role, Sourabh drove sales analytics for a $2 billion cloud business at Wipro. With a track record of delivering insights and driving business impact, Sourabh is well-equipped to tackle complex data challenges. Sourabh holds an MBA from IIM Shillong, an MPS Analytics from Northeastern University, and a degree in Electronics Engineering from MNIT Jaipur.

Venkata C. Duvvuri is a doctoral student in the Department of Technology Leadership and Innovation at Purdue University. Additionally, he is a Director of Data Science at Oracle Corporation in Redwood City, CA, USA. He loves teaching and is an adjunct faculty member at Northeastern University. He has held several leadership positions in data science at various companies. He holds a Master's degree in Computer Science from the University of Massachusetts Amherst and an MBA from the University of California, Davis.

Jay Roh obtained his Master's Degree in Data Analytics from Northeastern University. As part of the curriculum, he worked on various data science projects and machine learning models. He holds a Bachelor of Social Science in Media and Communication from Sungkyunkwan University, Seoul. Additionally, he served as a data researcher at the National Assembly of the Republic of Korea. Jay is highly passionate about the data field, with a particular interest in developing machine learning models and analyzing data with SQL and visualization tools.

Anish Mangipudi is a rising senior at Langley High School in northern Virginia, with plans to major in Electrical Engineering. He is currently working at the National Institute of Standards and Technology (NIST) in the Information Technology Laboratory, where he leverages his passion for machine learning to develop models in computational biology. Previously, Anish's enthusiasm for applying math and science to problem-solving led him to Marketbridge, a marketing science and consulting company, where he created R packages to curate scientific datasets for analytics. In his free time, Anish enjoys coding, playing basketball and football, and leading initiatives to build libraries at Boys and Girls Clubs in the DC area.