Ruidong Chen
APAN 5205
Final Report

1. Introduction

Electronic communication has become an indispensable aspect of our daily routines. However, alongside the numerous advantages of email communication, there exists a persistent problem, spam emails, that continues to disturb users worldwide. These unsolicited, irrelevant, and often malicious messages inundate inboxes, causing considerable inconvenience, wasting valuable time, and posing potential security threats to both individuals and organizations. As email usage continues to increase, the battle against spam becomes increasingly crucial. Therefore, finding effective ways to predict and filter spam emails is of paramount importance in enhancing the overall email user experience and safeguarding the integrity of digital communication.

The primary focus of this project report lies in addressing the critical issue of spam emails through the application of machine learning methodologies. We intend to harness the power of various advanced machine learning techniques, including logistic regression and cluster analysis, to achieve accurate prediction and classification of spam emails. The key contributions of this project are outlined as follows:

1. Develop a robust statistical model capable of distinguishing between ham (non-spam) and spam emails with exceptional precision.
2. Identify and extract relevant features from email messages. This project seeks to explore and identify the most significant features that substantially contribute to the prediction of spam emails.
3. Conduct a comprehensive evaluation of multiple machine learning methods and identify the optimal approach that yields the highest accuracy in spam email classification. Through rigorous experimentation and analysis, we aim to determine the most effective and efficient technique to combat spam effectively.

2. Data Exploration

2.1. Data Description

The dataset under analysis can be accessed through the following link: https://www.kaggle.com/datasets/mfaisalqureshi/spam-email. It consists of 5572 observations, each comprising two essential columns. The first column denotes the category of the email, distinguishing between "ham" (non-spam) and "spam". The second column contains the textual content of the emails.

Among the 5572 emails, 747 of them are classified as spam, accounting for approximately 13.41% of the total dataset. To visually represent the distribution of words in the dataset, a word cloud is given below. The word cloud provides an intuitive glimpse into the prevalent terms and phrases within the emails, aiding in understanding the underlying patterns and characteristics associated with spam and non-spam messages.

2.2. Discussion of Potential Covariates

For conducting a more in-depth analysis, we have formulated the following independent variables:

1. The number of words in an email. This variable captures the word count of each email, ranging from 1 to 171, with an average of 15.58 words per email. The inclusion of this covariate enables us to assess whether the length of an email has any correlation with its classification as spam or non-spam.
2. Percentage of uppercase letters. We compute the proportion of uppercase letters in each email, which varies from 0% to 100%, with an average of 3.7%. This variable offers insights into whether the usage of uppercase letters serves as a potential indicator for identifying spam emails.
3. Percentage of positive words. To evaluate the sentiment of the emails, we calculate the percentage of positive words, derived by dividing the number of positive words by

the total count of positive and negative words. Words are categorized using the "bing" lexicon. This percentage ranges from 0% to 100%, with an average of 34.8%. By considering the emotional tone conveyed by the emails, we can assess whether the prevalence of positive language correlates with the occurrence of spam messages.

4. Emotions of words. Employing the "nrc" lexicon, we categorize words into ten emotions: anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise, and trust. We then calculate the number of words in each email associated with each emotion, resulting in ten separate variables. This data allows us to explore the emotional characteristics of both spam and non-spam emails and figure out whether specific emotional patterns are existing in either category.

Certainly, a comparison of the top 20 common words for all emails, spam emails, and ham emails would provide valuable insights into the textual characteristics of each category. The frequency of each word, denoted within parentheses, aids in understanding the characteristic of these terms within the dataset.

Rank	Top 20 common words for all emails	Top 20 common words for spam emails	Top 20 common words for ham emails
1	call (585)	call (352)	2 (324)
2	2 (513)	free (223)	gt (318)
3	ur (391)	2 (189)	It (316)
4	gt (318)	txt (160)	ur (247)
5	4 (316)	ur (144)	call (233)
6	It (316)	4 (129)	day (201)
7	free (283)	mobile (127)	time (199)
8	day (221)	text (125)	love (197)
9	time (218)	stop (121)	4 (187)
10	love (207)	claim (111)	u (170)
11	send (196)	reply (104)	lor (162)
12	text (193)	price (90)	home (161)
13	txt (174)	cash (76)	da (149)
14	u (170)	150p (71)	dont (134)

15	home (163)	send (71)	send (125)
16	lor (162)	won (70)	pls (115)
17	stop (158)	nokia (67)	night (111)
18	da (149)	urgent (62)	happy (107)
19	reply (147)	win (60)	hey (107)
20	dont (146)	tone (59)	hope (107)

Indeed, categorizing the words from the table into three distinct categories based on their frequency in spam and ham emails is a judicious approach to identify potentially features. The three categories are as follows:

1. The first category comprises words that exhibit a high frequency in spam emails but have a low frequency in ham emails. Examples of such words include "free", "txt", "mobile", "claim", "price", "cash", "150p", "won", "nokia", "urgent", "win", and "tone". These words may serve as strong indicators of spam content due to their prominence in spam emails while being relatively infrequent in ham emails.
2. The second category represents words that have a high frequency in ham emails but a low frequency in spam emails. Some of the words in this category are "gt", "day", "time", "love", "u", "home", "lor", "da", "dont", "pls", "night", "happy", "hey", and "hope". These words might be indicative of non-spam emails as they are commonly used in legitimate correspondence but less prevalent in spam messages.
3. The third category consists of words that have a moderate frequency in both spam and ham emails. Examples include "call", "2", "ur", "4", "It", "send", "text", "stop", and "reply". As these words are not distinctly associated with either spam or non-spam content, they might not offer significant power for predicting spam emails.

Given the observation that the third category words may not be very effective in predicting spam emails, we propose counting the frequency of each word from the first and second categories in every email. The results will be stored in separate variables, generating a total of 26 variables, with each variable corresponding to a specific word. This step allows for the creation of more focused and informative features that can contribute to building an effective predictive model to accurately identify spam and non-spam emails. By leveraging these variables, we can enhance the accuracy and efficiency of our machine learning methods in detecting and classifying spam emails with greater precision.

3. Results

This project employs three distinct models for predicting spam emails:

1. Logistic regression model (without frequencies of 26 common words). The first model involves a logistic regression analysis, utilizing all the independent variables discussed in Section 2 (excluding the frequencies of the 26 common words). This model aims to assess the predictive capability of the selected covariates in accurately classifying spam and non-spam emails.
2. Logistic regression model (with frequencies of 26 common words). The second model also employs a logistic regression analysis, but this time, it incorporates the frequency of the 26 common words as additional independent variables. By considering these specific words, the model can determine if they significantly contribute to enhancing the accuracy of spam email prediction.
3. K-means clustering with logistic regression. The third model involves a two-step process. Firstly, the data is segmented into different clusters using k-means clustering. Subsequently, logistic regression is performed independently within each cluster to classify spam and non-spam emails. Finally, the predictions from all clusters are combined to produce the overall classification outcome. This approach is particularly efficient for large datasets, as k-means clustering reduces computational complexity compared to hierarchical clustering.

To evaluate the performance of each model, prediction accuracy, which measures the proportion of correctly classified observations out of the total observations, serves as the primary criterion. By conducting these three models and analyzing their prediction accuracies, we aim to identify the most effective approach for spam email classification. The results obtained will contribute valuable insights towards the development of a robust and efficient spam emails detection system, ultimately enhancing the overall email user experience and information security.

Model 1: logistic regression model (without frequencies of 26 common words)

The confusion matrix is a crucial tool for evaluating the performance of a classification model. For model 1, the confusion matrix is presented as follows:

	Predict ham	Predict spam

True ham	4698	127
True spam	549	198

The overall accuracy of Model 1 is calculated as (4698 + 198) / (4698 + 127 + 549 + 198) = 87.87%. However, upon closer examination of the confusion matrix, it becomes apparent that model 1 exhibits significant weaknesses in predicting spam emails correctly. Specifically, out of the total 747 true spam emails, only 198 were accurately predicted as spam, while 549 spam emails were incorrectly classified as ham. This disparity in predicting spam emails correctly (true positives) and misclassifying them as ham (false negatives) indicates poor performance in identifying spam messages.

Model 2: logistic regression model (with frequencies of 26 common words)

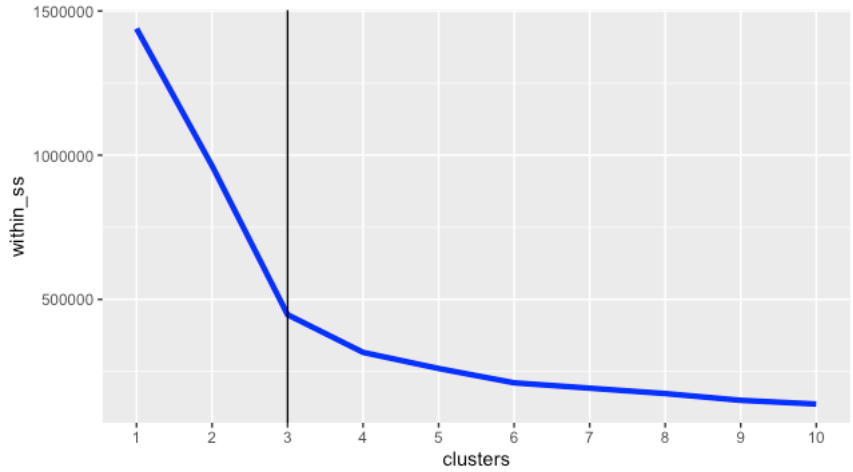
By incorporating the counts of the 26 common words (appearing in either spam or ham emails but not in both) into the logistic regression model, model 2 exhibits notable improvements in prediction accuracy. The confusion matrix is as follows:

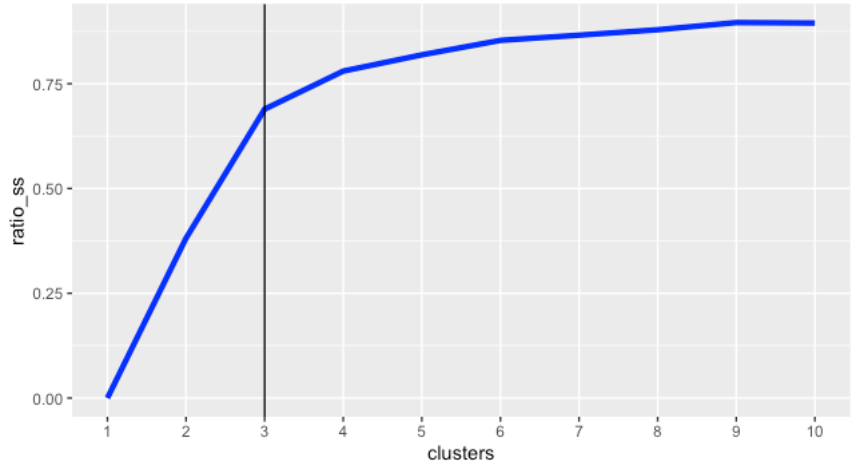
	Predict ham	Predict spam
True ham	4787	41
True spam	258	489

The overall accuracy of Model 2 is (4787 + 489) / (4787 + 41 + 258 + 489) = 94.63%. The accuracy has significantly increased compared to model 1, indicating that the addition of the 26 common words as features has enhanced the model's ability to classify emails correctly. The marked improvement in accuracy is evident in both spam and ham email predictions. Specifically, the number of correctly predicted spam emails (true positives) has risen significantly from 198 in model 1 to 489 in model 2. This substantial increase in true positive rate for spam classification demonstrates the effectiveness of incorporating the 26 common words' frequencies in improving the model's ability to distinguish spam from non-spam emails.

Model 3: k-means clustering with logistic regression

Model 2 achieves a commendable level of accuracy in predicting spam emails. However, we believe that the dataset can be better analyzed by partitioning it into different clusters, where emails within each cluster exhibit similar features. To achieve this, we proceed with model 3. In this approach, we first employ k-means clustering. The within SS plot and ratio plot both suggest a three-cluster solution.





Subsequently, we run a logistic regression model, incorporating all the features used in model 2, within each cluster. By combining the prediction results from each cluster, we arrive at the confusion matrix for model 3:

	Predict ham	Predict spam
True ham	4774	51
True spam	181	566

The overall accuracy of model 3 is (4774 + 566) / (4774 + 51 + 181 + 566) = 95.84%. Model 3 outperforms both model 1 and model 2, achieving the highest accuracy in classifying emails. The key advantage of model 3 lies in its ability to predict spam emails more accurately. Of the 747 true spam emails, model 3 correctly identifies 566 as spam, a notable improvement compared to model 2's prediction of 489. This enhanced spam prediction capability further validates the efficacy of k-means clustering in grouping similar emails

together and leveraging logistic regression within each cluster to obtain more accurate spam classifications.

However, model 3 does show a slightly lower accuracy in predicting ham emails, with 4774 out of 4825 true ham emails accurately classified. This trade-off between correctly predicting ham and spam emails highlights the model's focus on accurately identifying spam as its primary objective. Considering the project's purpose of effectively predicting spam emails, model 3 emerges as the most suitable choice. Its superior performance in spam classification, along with a highly competitive overall accuracy, makes it a promising and robust approach for addressing the critical issue of spam email detection.

4. Conclusion

In this project, we addressed the critical issue of predicting spam emails using machine learning methods. Our objective was to develop effective predictive models to distinguish between ham (non-spam) and spam emails, thereby enhancing the overall email user experience and mitigating potential security threats.

Three distinct predictive models were implemented and evaluated for their performance. In model 1, we applied logistic regression using selected independent variables, achieving an accuracy of 87.87%. However, this model exhibited limitations in correctly predicting spam emails, with a considerable number of false negatives. To address these shortcomings, model 2 introduced the frequencies of 26 common words as additional features in logistic regression. This enhancement led to a notable accuracy improvement, reaching 94.63%. The true positive rate for spam classification substantially increased, indicating that these common words played a crucial role in effectively identifying spam emails. Recognizing the potential benefits of clustering, we pursued model 3, combining k-means clustering with logistic regression. By partitioning the data into three clusters, we observed further accuracy enhancements. Model 3 achieved an impressive accuracy of 95.84%, particularly excelling in accurately predicting spam emails. With 566 out of 747 spam emails correctly classified, model 3 demonstrated superior spam prediction capabilities compared to the previous models.