Hook Document: Spam or Ham Detection

DS4002 Case Study

GitHub Repository: https://github.com/natalieassaad/DS4002_CS3

Context & Motivation

Spam emails are unsolicited messages sent in bulk, often for commercial purposes. While some are simply irrelevant, many contain inappropriate or even malicious content designed to deceive recipients. In 2023, spam emails accounted for nearly half of global email traffic [1]. Although this seems like a trivial topic, it can cause large scale damage to both businesses and consumers. In "The Economics of Spam" by Microsoft researcher Justin Rao and Google employee David Reily, the researchers predict that American firms and consumers experience a loss of $20 billion annually due to spam [2]. Cyber attackers use spam emails as a vector for malware or phishing attempts, particularly with large-scale email campaigns. An estimated 91% of cyber attacks begin with a phishing email [3].

Given the depth of this issue, identifying and filtering malicious content before it reaches users is key in preventing harm caused by spam emails. Currently, email services like Gmail use a combination of techniques to filter spam emails such as identifying known spam signals, gathering historical user feedback, and content analysis [4]. While these methods are effective, malicious emails that do reach user inboxes can result in significant consequences such as identity theft, financial loss, cyber attacks, or data breaches.

Deliverable

In this case study, you investigate which of the three machine learning models (K-Nearest Neighbors, Logistic Regression, and Decision Trees) is the most effective at identifying spam emails based on precision and accuracy. In order to complete this case study, you will write the code for preprocessing the textual data as well as developing the KNN and Decision Tree models. You will submit a GitHub repository that includes materials such as your cleaned dataset, completed scripts, output files, and a one page write-up evaluating your model results/conclusions.

References

[1] Petrosyan, A. (2024, October 4). *Monthly share of spam in the total e-mail traffic worldwide from January 2014 to December 2023*. Statista. https://www.statista.com/statistics/420391/spam-email-traffic-share/

[2] Rao, Justin M., and David H. Reiley. 2012. "The Economics of Spam." *Journal of Economic Perspectives*, 26 (3): 87–110. DOI: 10.1257/jep.26.3.87

[3] G. Smith, "Top phishing statistics for 2025: Latest figures and Trends," StationX, https://www.stationx.net/phishing-statistics/.

[4] Mailtrap. (n.d.). *Gmail spam filter: How it works & how to avoid it*. Retrieved from https://mailtrap.io