Name: Yehezkiel Efraim Darmadi
Unit: FIT5145
Student Number: 34078215

Optimizing Spam Email Detection: A Dual Strategy of Resampling and Ensemble Learning

1. Project Description

1.1. Introduction

Spam emails, typically bulk messages linked to scams or advertising, pose serious threats to user privacy and security (Aslan et al., 2023). This project aims to enhance spam detection by tackling dataset imbalance with a unique combination of resampling and ensemble learning as suggested by Qi et al.'s (2023) with the difference in employing alternative techniques and models.

The Kaggle dataset, comprising 5,158 entries categorized into email content and spam/non-spam labels, demonstrates a label imbalance, with spam accounting for only 13% of the data, as shown in the figure below.

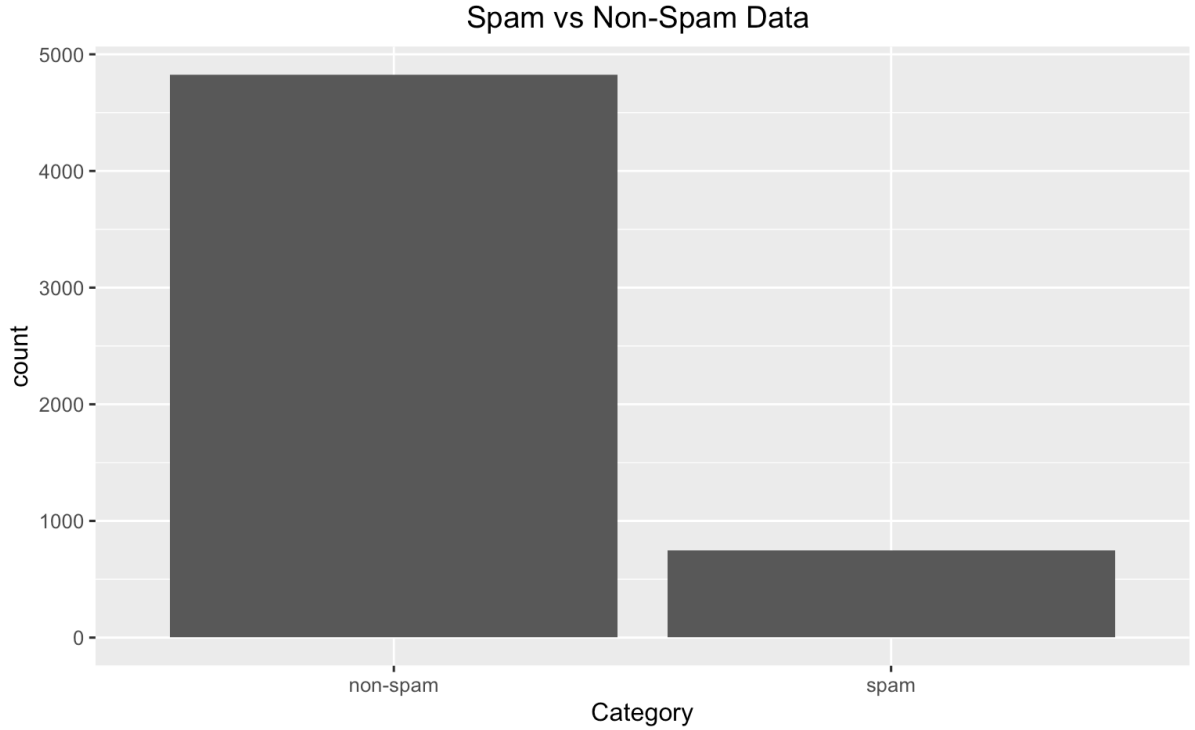

Figure 1. Biased dataset towards the "non-spam".

1.2. Objective

The goal is to boost the accuracy of spam detection models by mitigating dataset imbalance and using ensemble learning. The project will assess different resampling methods (random under-sampling, SMOTE, bootstrapping) to even out the dataset and combine Random Forest, SVM, and Logistic Regression via majority voting. Model success will be gauged by accuracy, precision, recall, and F1 scores in identifying spam emails.

1.3. Data Science Roles and Responsibilities

- **Project Manager**: Oversees the project's overall execution, ensuring timelines and deliverables are met.
- **Data Scientists**: Develop and tune the machine learning models as described in the methodology section, conduct experiments with resampling techniques, and implement ensemble learning strategies.
- **Data Analysts**: Analyze the spam email dataset for insights and patterns, perform exploratory data analysis, and assist in evaluating model performance.
- **Data Engineers**: Ensure the data pipeline is robust, automate data collection and pre-processing steps, and manage database integration.
- **Data Protection Officer**: Ensures the project adheres to data privacy and security standards, evaluating the ethical implications of data usage and model deployment.

1.4. Methodology

- **Data Collection**: Compile a comprehensive dataset of emails, categorizing them into spam and non-spam.
- **Data Cleaning**: Remove duplicates, correct inconsistencies, and handle missing values.
- **Data Analysis**: Explore the dataset to identify patterns and imbalances.
- **Model Building**: Develop machine learning models using Random Forest, SVM, and Logistic Regression.
- **Resampling**: SMOTE, random under-sampling, and bootstrapping to balance the dataset.
- **Ensemble Learning**: Combine the predictions from Random Forest, SVM, and Logistic Regression via majority voting.
- **Evaluation**: Measure the models' performance using accuracy, precision, recall, and F1 scores.

1.5. Expected Outcome

The project aims to deliver a highly accurate spam detection model, capable of adapting to the evolving nature of spam emails. Additionally, it seeks to provide insights into the most effective strategies for dealing with imbalanced datasets in the context of spam email detection.

2. Business Model

2.1. Introduction

Email is vital for communication, yet spam emails threaten efficiency and security, leading to significant challenges including productivity losses and security risks. In 2021, the FBI noted losses of around USD2.4 billion from email scams and business compromises, with the IC3 reporting 3,729 ransomware incidents costing over USD49 million. The total financial impact of spam emails in 2021 was nearly USD2.45 billion, highlighting the importance of this project's improvements in spam detection for safer and more efficient email use (Lanctot, A., & Duxbury, L., 2022; Omotehinwa, T. O., & Oyewola, D. O., 2023).

2.2. Beneficiaries

The project offers advantages to cybersecurity companies as a product and email service providers as a service, along with businesses and individuals seeking spam-free inboxes. Additionally, it provides the data science community, particularly those focused on NLP and ensemble learning, with valuable insights on managing imbalanced datasets, thereby promoting more secure digital communications for all involved parties.

2.3. Value Proposition

The project minimizes manual spam filtering efforts. If assuming 13% of a daily 100-minute email session is spam, and the model reduces this by 90%, it saves about 12 minutes each day. Moreover, reducing phishing risks could lead to significant financial benefits for businesses. Enhanced spam detection also improves data security and assists organizations in meeting data protection regulations, providing additional advantages (Sharabov et al., 2024).

2.4. Challenges

This project's challenges include keeping up with spammers who change their methods often, protecting email privacy, and fine-tuning the detection system to avoid wrongly flagged emails and missed spam. These issues are key to creating a spam detector that works well and respects user privacy.

2.5. Mitigation Strategies

To overcome the challenges, strategies include anonymizing data to protect privacy, adhering to data laws, and integrating online learning for updates. Additionally, continuous model monitoring, regular updates, and forming partnerships with email service providers for better data access are essential for a robust and privacy-aware spam detection system.

2.6. Conclusion

"Optimizing Spam Email Detection" holds the promise of significantly advancing email security through innovative data science strategies. By addressing key challenges and leveraging the strengths of ensemble learning and resampling techniques, it offers a powerful tool against the ever-present threat of spam emails.

Reference

Aslan, Ö., Aktuğ, S. S., Ozkan-Okay, M., Yilmaz, A. A., & Akin, E. (2023). A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions. Electronics, 12(6), 1333. https://doi.org/10.3390/electronics12061333

Lanctot, A., & Duxbury, L. (2022). Measurement of Perceived Importance and Urgency of Email: An Employees' Perspective. Journal of Computer-Mediated Communication, 27(2). https://doi.org/10.1093/jcmc/zmac001

Omotehinwa, T. O., & Oyewola, D. O. (2023). Hyperparameter Optimization of Ensemble Models for Spam Email Detection. Applied Sciences, 13(3), 1971-. https://doi.org/10.3390/app13031971

Sharabov, M., Tsochev, G., Gancheva, V., & Tasheva, A. (2024). Filtering and Detection of Real-Time Spam Mail Based on a Bayesian Approach in University Networks. Electronics (Basel), 13(2), 374-. https://doi.org/10.3390/electronics13020374

Qi, Q., Wang, Z., Xu, Y., Fang, Y., & Wang, C. (2023). Enhancing Phishing Email Detection through Ensemble Learning and Undersampling. Applied Sciences, 13(15), 8756. https://doi.org/10.3390/app13158756