Name: Yehezkiel Efraim Darmadi
Unit: FIT5145
Student Number: 34078215
Have you selected a topic for Assignment 3 that is different from the one that you used for Assignment 1 (i.e., have you rewrote the first two sections of the report)?
I use the project from the assignment 1

Dataset link:
https://drive.google.com/file/d/10DYnurP9xAdohj81yH6TFOEQqLZxFl9H/view?usp=share_link

Kaggle dataset:
https://www.kaggle.com/datasets/jackksoncsie/spam-email-dataset/code?datasetId=3690036

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

1. Project Description

1.1. Introduction

Spam emails, often bulk messages associated with scams or advertising, represent significant threats to user privacy and security (Aslan et al., 2023). These emails are not only pervasive but also compromise the integrity of digital communication systems.

A common problem with real-world email datasets is the unequal balance between spam and non-spam messages. Such disparities can skew the performance of detection models, making it difficult to accurately identify spam (Qi et al., 2023).

This project aims to improve spam detection by addressing the issue of dataset imbalance. It employs a novel approach that integrates resampling with ensemble learning, as suggested by Qi et al. (2023). However, our methodology distinguishes itself by utilizing alternative techniques and models to enhance detection accuracy.

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 XGBoost via majority voting. Model success will be gauged by Sensitivity, F1, ROC ,and AUC scores in identifying spam emails (Tharwat, A., 2018).

1.3. Data Science Roles and Responsibilities

- **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**: Analyse 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. 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 organisations 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 anonymising 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.

3. Characterising and Analysing Data

3.1. Dataset Sources

Below are the potential datasets to be used for the project:

- Email Service Providers (ESPs): Collaborations with ESPs (e.g. Gmail, Yahoo, or Outlook) can provide access to a rich source of real-time email data, adhering to privacy and data protection laws.
- Cybersecurity Firms: Partnering with firms that specialize in internet security might provide datasets enriched with the latest spam threats.

3.2. Data Characteristics

- Volume: A combination of large datasets from Email Service Providers and Cybersecurity Firms.
- Variety: A combination of organized email data and labels, enhanced with varied metadata, content, user engagements, and detailed threat intelligence from ESPs and cybersecurity companies.
- Velocity: A blend of historical data and real-time streams from ESPs and Cybersecurity Firms, facilitating timely analysis of spam trends.
- Veracity: Generally high data quality with real-world accuracy from ESPs and cybersecurity firms.

3.3. Required Platforms, Software, and Tools

The solution utilizes Amazon Web Services (AWS) to offer robust, scalable, and efficient data storage and processing capabilities.

Data Storage:

- Amazon S3: A cost-effective, scalable storage solution with high durability and availability, integrating seamlessly with other AWS services.
- Amazon RDS (Relational Database Service): For structured data storage, offering automated backups, patching, and scaling.

Data Processing:

- Apache Spark: A robust open-source processing engine, well-suited for big data analytics, capable of handling both batch and real-time data processing and integrating with various data sources.
- AWS Glue: A fully managed ETL service that simplifies data preparation and transformation for analytics, connecting seamlessly with Amazon S3 and RDS while supporting data cleaning and integration.
- Jupyter Notebooks (with R): An interactive environment for data analysis and machine learning. It can be used with Apache Spark for exploratory data analysis and model development.
- AWS Lambda: A serverless compute service that automatically runs code in response to events and manages underlying resources, ideal for real-time data processing.

3.4. Data Analysis and Statistical Method

Our spam detection system tackles data imbalance, where non-spam emails often exceed spam emails, causing biased predictions. We employ resampling techniques such as SMOTE, Bootstrapping, and random under-sampling to balance the dataset and enhance data quality. Additionally, we use ensemble learning to integrate diverse models, boosting performance, accuracy, and robustness in machine learning applications (Zhang et al., 2018).

The project leverages ensemble learning by integrating Random Forest, SVM, and XGBoost. Random Forest prevents overfitting in high-dimensional data, SVM excels in managing complex classification boundaries, and XGBoost offers speed and accuracy in large-scale applications. This combination enhances predictive accuracy and model robustness, adapting to evolving spam tactics.

We use balanced evaluation metrics (Sensitivity, F1 score, ROC, and AUC) to ensure the model's reliability and accuracy (Tharwat, A. 2018). This careful calibration maintains trust in the system's performance, even as spam threats become more dynamic.

The high-level output of this approach is a spam detection model with high accuracy, effectively distinguishing between spam and non-spam emails. Using ensemble learning, the model adapts to new spamming techniques. It achieves high sensitivity, precision, recall, F1 scores, and AUC, resulting in minimal false negatives. This creates a robust, efficient, and secure email environment, significantly reducing risks and inefficiencies associated with spam emails.

Overall, our strategic design enhances email security and efficiency by effectively managing spam in a dynamic digital environment, ensuring emerging threats are countered with precision.

3.5. Demonstration

The dataset used for the demonstration is sourced from Kaggle because it provides a large, well-labelled collection of emails, making it ideal for building and evaluating spam detection models.

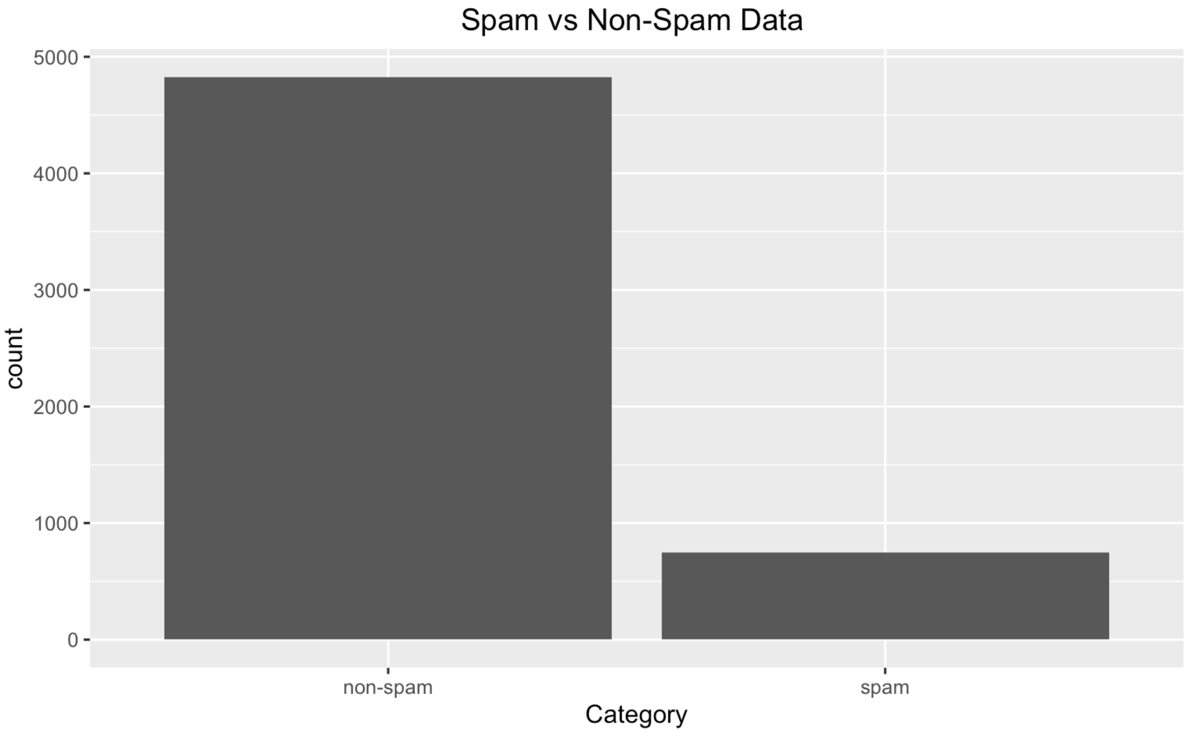

Figure 1. Biased dataset towards the "non-spam".

Figure 1 shows that the Kaggle dataset contains 5,158 entries with email content and spam/non-spam labels, showing an imbalance with only 13% spam. Resampling is needed to balance the dataset and avoid bias. Bootstrapping, which resamples the minority dataset by repeatedly drawing samples with replacement until it matches the majority, was used. Future comparisons will be made with SMOTE and Random Under-sampling to determine the best method.

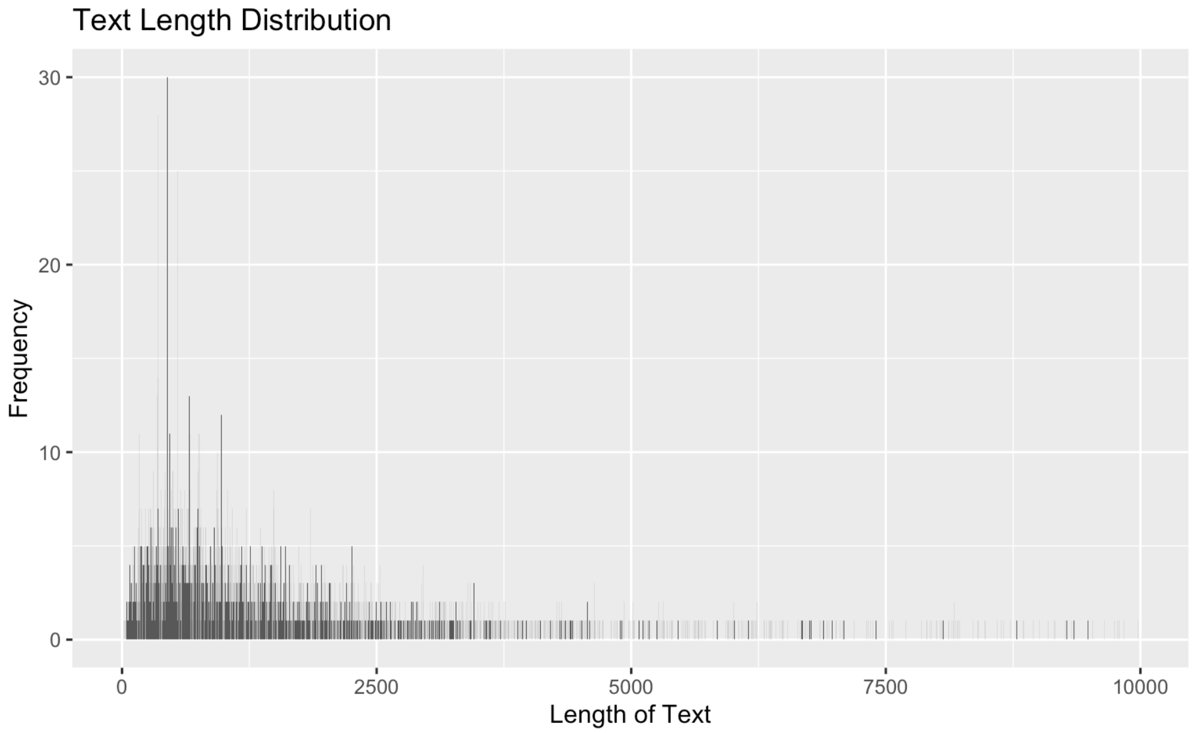

Figure 2. Text Length Distribution

Figure 2 shows a wide range of text lengths with a standard deviation of 2042.64. It can be seen that the distribution is skewed to the right, indicating mostly shorter emails. Thus, it can be said that many texts are brief, with a high frequency under 2500 characters.

There are a few emails with significantly longer text lengths, extending up to 10,000 characters, but these occur much less frequently. There are even texts over 40,000 characters that were excluded to prevent biasing the model.

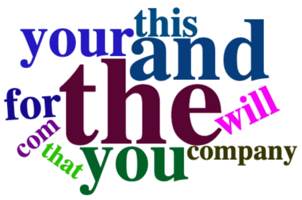

Figure 3. Top 10 Most Common Words in Spam Email

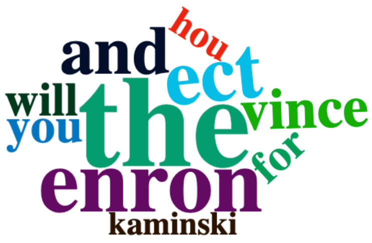

Figure 4. Top 10 Most Common Words in Not Spam Email

The email was pre-processed using TF-IDF to assess the relevance of words in a text or corpus. Figure 3 displays a word cloud for spam emails, it featuring words like "you", "the", "this", "and", "for", and "will". These words are typically associated with persuasive or marketing-related content, aimed at engaging or directing the recipient to take action.

Figure 4 showcases a word cloud for non-spam emails, highlighting words such as "vince", "kaminski", "enron", "hou", etc. The presence of specific proper nouns like "enron" and "kaminski" points to the dataset possibly including corporate communications, particularly from entities like enron. This contrast in vocabulary between the two clouds underlines the linguistic differences between spam and not spam email, offering insights for improving spam detection algorithms.

The analysis results and TF-IDF data, along with text length, will be used to build three models: Random Forest, SVM, and XGBoost. The predictions from these models will be combined to create an Ensemble Learning model.

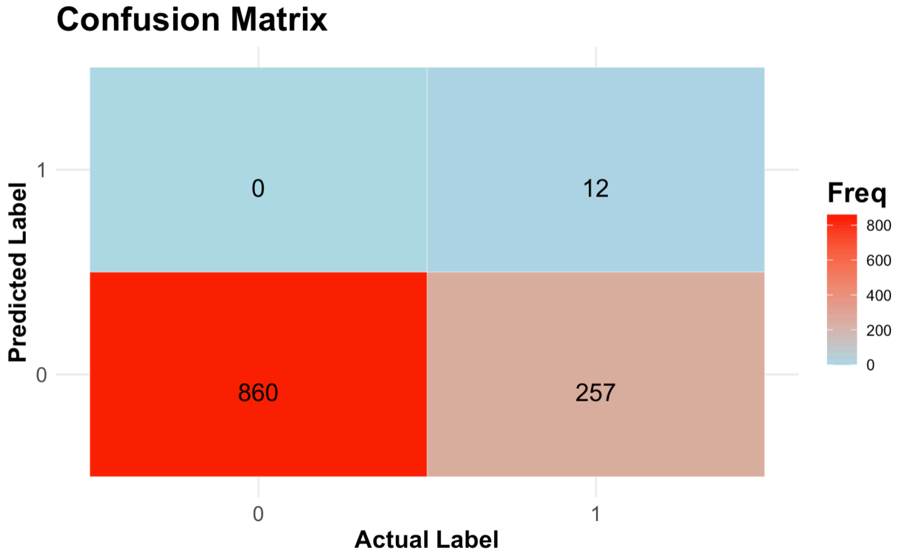

Figure 5. Confusion Matrix from Ensemble Learning

Metric	Value
Sensitivity	0.04
F1_score	0.08

Table 1. Evaluation Metrics

Figure 5 and Table 1 evaluate the ensemble learning model, showing its failure to detect any spam emails. It correctly classified many non-spam emails but mislabeled spam as non-spam due to imbalanced test data reflecting real-world conditions. This is evidenced by a low sensitivity score of 0.04 and an F1 score of 0.08, highlighting the model's ineffectiveness in predicting true spam.

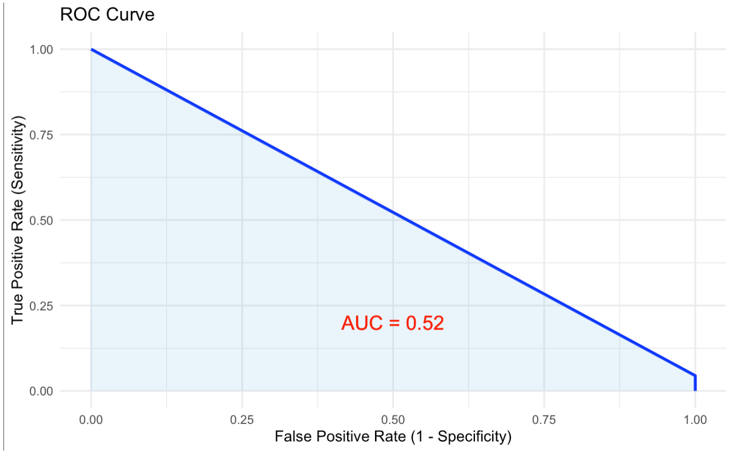

Figure 6. ROC and AUC for Ensemble Learning

The ROC curve and an AUC of 0.52 further confirm the model's poor performance, barely outperforming random guessing in distinguishing between spam and non-spam. This diagonal ROC curve suggests that the model lacks discriminative power, underscoring the urgent need for improvements in feature engineering, algorithm optimization, and addressing data imbalance to enhance spam detection capabilities.

The model's performance fell short of project expectations due to the limited scope of the Kaggle dataset, which was primarily used for demonstration. Plans are in place to enhance the model by incorporating real-world data from potential datasets, aiming to significantly improve its effectiveness.

4. Data Governance and Management

4.1. Data Governance

The project follows the Cross-Industry Standard Process for Data Mining (CRISP-DM methodology (Saltz, J. S., 2021), a structured approach to data science projects encompassing Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment. This ensures a thorough and consistent process, promoting best practices and successful outcomes.

Practices for Data Governance:

- Security and Confidentiality: Employ encryption and strict access controls to safeguard data.
- Compliance and Ethics: Ensure compliance with laws like GDPR and HIPAA, and uphold ethical standards through regular audits.
- Data Quality and Risk Management: Monitor and maintain data quality, and manage risks with a comprehensive breach response plan.

Below are the potential ethical concerns with the use of the data:

1. Privacy and Anonymity: Ensuring that anonymization techniques are robust enough to prevent the re-identification of individuals, safeguarding user privacy despite stringent security measures.

2. Informed Consent: Guaranteeing that data is collected with informed consent, where participants understand how their data will be used and can withdraw their consent at any time.
3. Data Breach and Incident Response: Managing data breaches effectively by notifying affected individuals promptly and taking corrective actions to mitigate impacts.

4.2. Data Management

Following the CRISP-DM guidelines, the management of data throughout its lifecycle is structured and documented. This standardization helps in maintaining the quality and usability of data across different stages of the project.

Practices for Data Management:

- Collection & Storage: Data is stored in secure Amazon S3 buckets (Amazon Web Services, 2024), with integrity checks to ensure accuracy.
- Processing & Quality: Data processing is performed using Apache Spark, with version-controlled scripts and regular quality assessments to maintain high data standards.
- Archiving & Deletion: Post-project, data is archived or securely deleted following organizational policies to prevent misuse.

Reference

Amazon Web Services. (n.d.). Amazon S3 security. Retrieved May 17, 2024, from https://aws.amazon.com/s3/security/

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

Jackksoncsie. (n.d.). Spam email Dataset [Dataset]. https://www.kaggle.com/datasets/jackksoncsie/spam-email-dataset

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

Saltz, J. S. (2021). CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 2337-2344). doi: 10.1109/BigData52589.2021.9671634

Tharwat, A. (2021). Classification assessment methods. Applied Computing & Informatics, 17(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Zhang, D., Jiao, L., Bai, X., Wang, S., & Hou, B. (2018). A robust semi-supervised SVM via ensemble learning. Applied Soft Computing, 65, 632–643. https://doi.org/10.1016/j.asoc.2018.01.038