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

Name: Yehezkiel Efraim Darmadi

Unit: FIT5145

Student Number: 34078215

Email: ydar0001@student.monash.edu

Introduction to Spam Email Issues

- What are spam emails?

Spam emails are unwanted bulk messages often linked to scams or advertising, posing serious threats to user privacy and security (Aslan et al., 2023) .

- Challenges posed by the dataset:

A common problem with real-world email datasets is the unequal balance between spam and non-spam messages

- The impact of spam emails:

In 2021, the FBI reported losses of approximately USD 2.4 billion due to email scams, highlighting the severe financial and security implications of spam emails (Lanctot, A., & Duxbury, L., 2022; Omotehinwa, T. O., & Oyewola, D. O., 2023) .

Project Objectives

- Goal:

To boost the accuracy of spam detection models by mitigating dataset imbalance and utilizing ensemble learning.

- Outcome:

Develop a highly accurate spam detection system that can adapt to the evolving nature of spam emails.

Data Science Teams

- Data Scientists:

Develop, tune, and implement machine learning models; conduct experiments with resampling techniques.

- Data Analysts:

Analyze spam email datasets for insights, perform exploratory data analysis.

- Data Engineers:

Manage data collection and preprocessing, ensure robust data pipeline integration.

- Data Protection Officer:

Ensure compliance with data privacy and security standards.

Required Platforms, Software, and Tools

- The solution focuses on leveraging AWS-Based Data Storage & Processing Capabilities.
- Data Storage Solutions:
  - Amazon S3: Scalable storage for vast data quantities, offering high durability and seamless integration with AWS services.
  - Amazon RDS: Manages structured data with automated backups, patching, and scalability.
- Data Processing Tools:
  - Apache Spark: Robust open-source engine for both batch and real-time data processing, excellent for big data analytics.
  - AWS Glue: Fully managed ETL service, simplifying data preparation and integration with Amazon S3 and RDS.
  - Jupyter Notebooks (with R): Interactive platform for data analysis and model development, compatible with Apache Spark.
  - AWS Lambda: Serverless computing service, ideal for event-driven and real-time data processing.

Business Model: Beneficiaries, and Challenges

- Beneficiaries:

Cybersecurity companies, email service providers, businesses, and individuals.

- Value:

Reduces manual spam filtering efforts, minimizes phishing risks, and improves email communication efficiency.

- Challenges:

Rapid evolution of spammer tactics, maintaining email privacy, and avoiding misclassification.

- Mitigation:

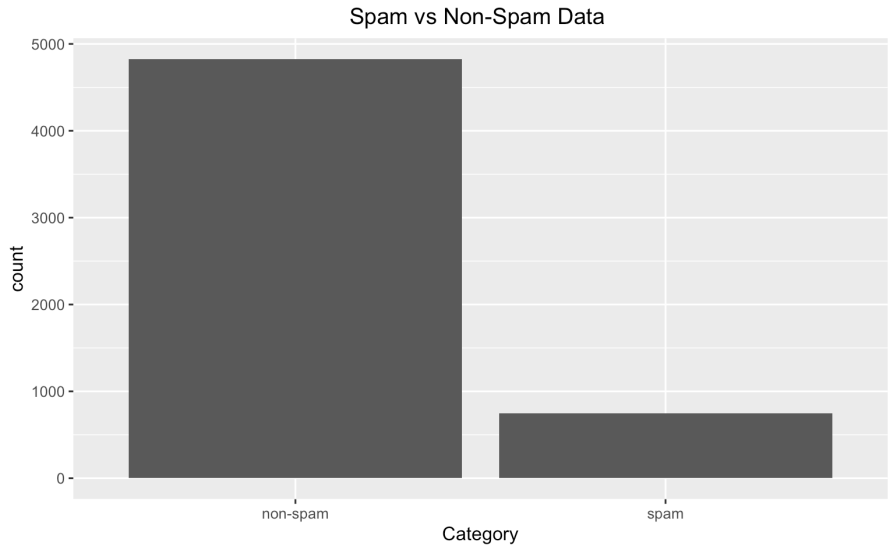
Anonymizing data, complying with data protection laws, and implementing continuous model learning and updates.

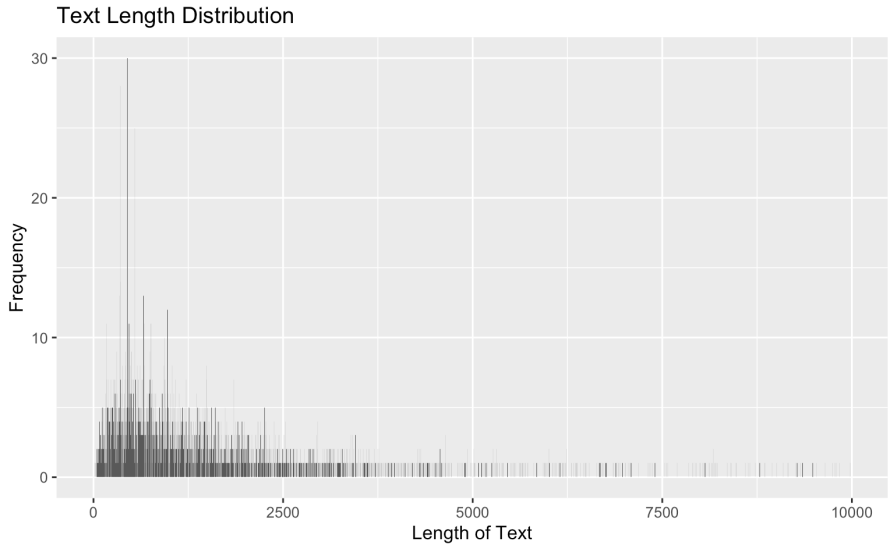
Data Characteristic

- Dataset sources:

Email Service Providers (ESPs) and Cybersecurity Firms.

- Data Characteristics:
  - Volume: A combination of large datasets from Email Service Providers and Cybersecurity Firms.
  - Variety: A combination of organized email data and labels, 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.





Data Analysis

- The Kaggle dataset consists of 5,158 email entries, with a significant imbalance: only 13% are classified as spam.
- Resampling Techniques:
  - SMOTE
  - Random Under-sampling
  - **Bootstrapping** resampling the minority class until it equates the majority.
- Distribution skewed to the left: most emails are under 2,500 characters, indicating shorter text lengths predominantly and have variance.

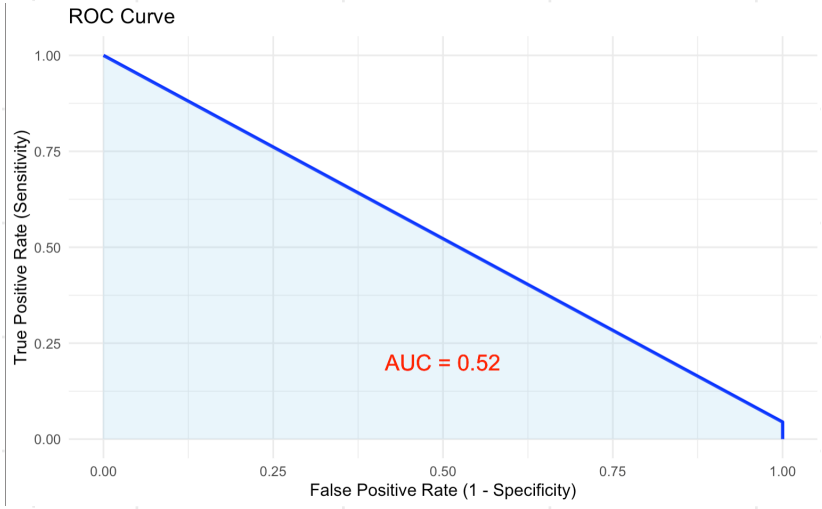
Data Analysis

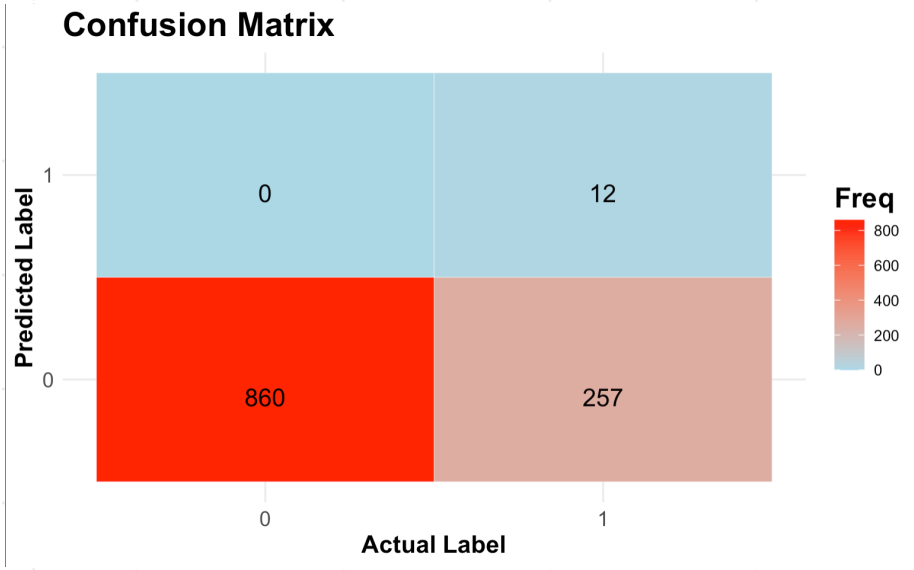
- TF-IDF Preprocessing:
  - Utilized to evaluate the relevance of words within the emails, enhancing the understanding of content significance.
- Word Cloud Analysis:
  - Spam Emails: Features persuasive or marketing-related words like "you", "the", "this", "and", "for", "will", indicating intent to engage or prompt action.
  - Non-Spam Emails: Dominated by proper nouns such as "vince", "kaminski", "enron", "hou", suggesting inclusion of corporate communications, especially from entities like Enron.
- Insights:
  - Clear linguistic distinctions between spam and non-spam emails provide crucial insights for refining spam detection algorithms.

Top 10 Most Common Words in Spam Email

Top 10 Most Common Words in Not Spam Email

Machine Learning Model





- Machine Learning Models:
  - Random Forest
  - SVM
  - XGBoost
- Evaluation Metics
  - F1 Score
  - ROC and AUC
  - Specificity

Metric	Value
Sensitivity	0.04
F1_score	0.08

- Sensitivity of 0.04 and F1 score of 0.08 indicate the inability to predict spam email.
- A ROC curve with an AUC of 0.53 indicates performance barely above random guessing, showing significant limitations in model discrimination.
- Key Issues:
  - Influenced by the imbalanced nature of the test data which reflects real-world conditions.
- Future Directions:
  - Plans to enhance the model using real-world data for better performance (more data).
  - Focus on refining feature engineering, optimizing algorithms, and addressing data imbalance to improve detection capabilities.

Data Governance

- Cross-Industry Standard Process for Data Mining (CRISP-DM) Framework:
  - Follows a structured approach to data science:
    - Business Understanding
    - Data Understanding
    - Data Preparation
    - Modelling
    - Evaluation
    - Deployment
- Data Governance Components:
  - Security and Confidentiality:
    - Utilizes encryption and strict access controls to protect data integrity.
  - Compliance and Ethics:
    - Complies with regulations such as GDPR and HIPAA, reinforced by regular audits to maintain ethical standards.
  - Data Quality and Risk Management:
    - Continuously monitors and maintains data quality and manages potential risks with a well-prepared breach response plan.

Data Management

- Aligned with CRISP-DM
- Structured Data Lifecycle Management:
  - Structured and documented, ensuring quality and usability at all project stages.
  - Standardization helps maintain the quality and usability of data across different stages of the project.
- Key Data Management Practices:
  - Collection & Storage:
    - Secure storage in Amazon S3 buckets with integrity checks to guarantee data accuracy.
  - Processing & Quality:
    - Data processing through Apache Spark, utilizing version-controlled scripts and regular quality assessments to uphold high data standards.
  - Archiving & Deletion:
    - Post-project, data is either archived or securely deleted in accordance with organizational policies to prevent misuse.