Spam Email Detection & Business Analytics

An end-to-end data science solution combining SQL, NLP, and machine learning to combat spam and protect user trust.

CHALLENGE

The Business Problem

Organizations handling large email volumes face critical challenges that impact operations and user experience:

- Inbox clutter reducing productivity
- Security threats from malicious content
- Eroded user trust and engagement
- Manual moderation not scalable

Automated spam detection is essential for modern email systems.

DATASET

Analysis Overview

5,572

Total Emails

Comprehensive dataset analyzed

747

Spam Detected

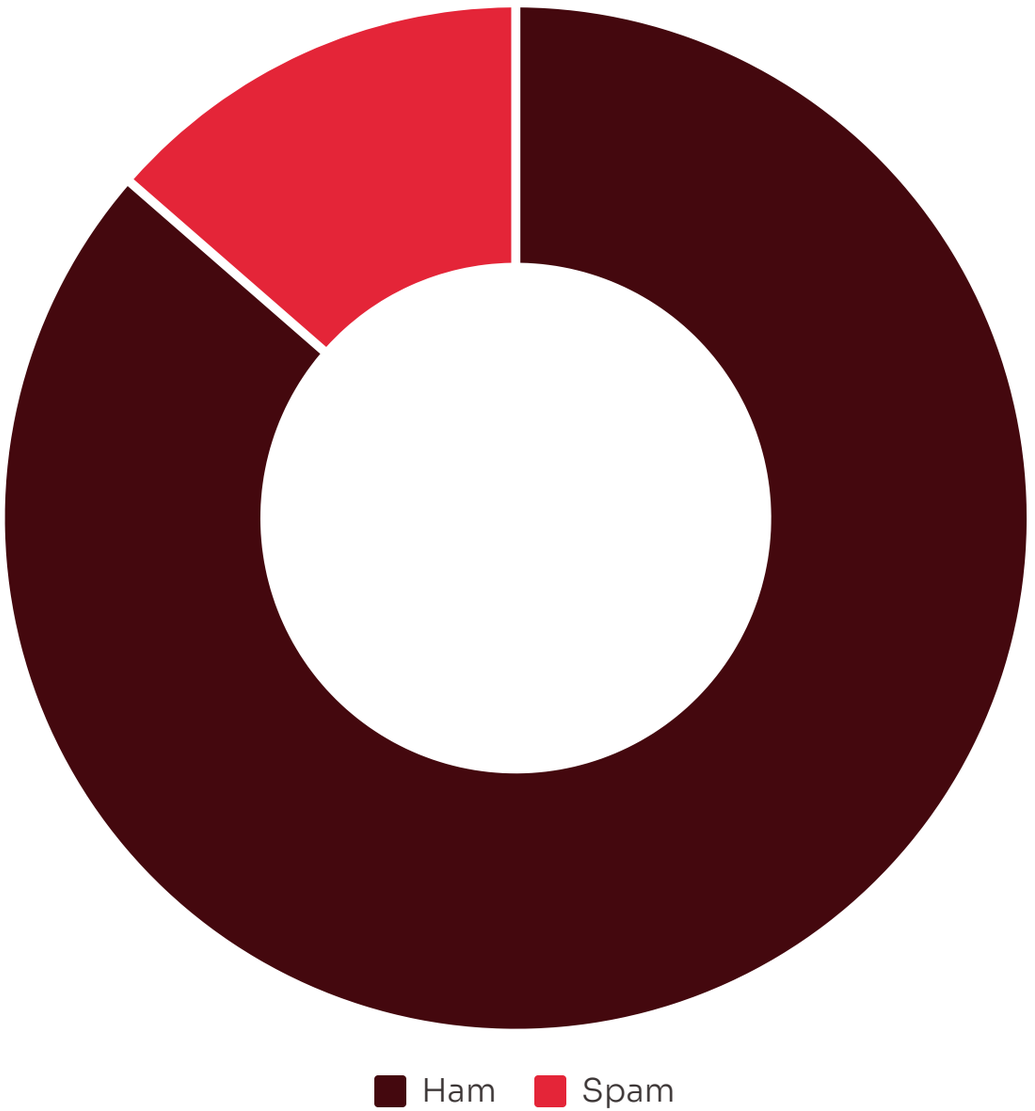
13.4% of total volume

4,825

Ham Emails

86.6% legitimate messages

Spam vs Ham Distribution



Class imbalance presents a key challenge: spam represents only 13.4% of emails, requiring optimized precision to minimize false positives in detection systems.

METHODOLOGY

Technical Approach

Data Ingestion

Emails loaded into PostgreSQL for structured analysis

SQL Analysis

Pattern extraction and statistical queries

NLP Processing

Python-based text analysis and feature engineering

Visualization

Power BI dashboards for insights

Key Analytical Insights

Message Length

Spam emails are **66% longer** than ham emails on average, indicating verbose, repetitive content patterns.

Vocabulary Diversity

Spam shows **55% lower** vocabulary diversity, revealing repetitive language and limited contextual variation.

Action Keywords

Spam heavily uses action-driven terms: **'call', 'free', 'claim', 'prize'** — designed to trigger immediate response.

Linguistic Quality

Ham emails demonstrate higher linguistic diversity and stronger contextual relevance throughout messages.

STRATEGY

Business Recommendations

Deploy NLP Filtering

Implement TF-IDF and machine learning models for automated spam classification with high accuracy.

Spam Risk Scoring

Introduce scoring system based on message length, keyword frequency, and vocabulary patterns.

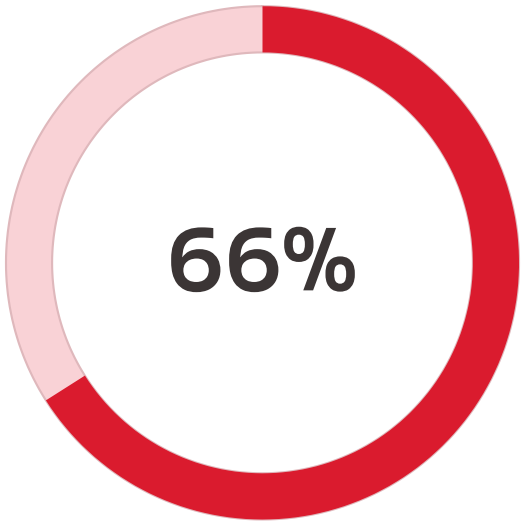
Monitor KPIs

Track spam detection metrics through Power BI dashboards for continuous optimization.

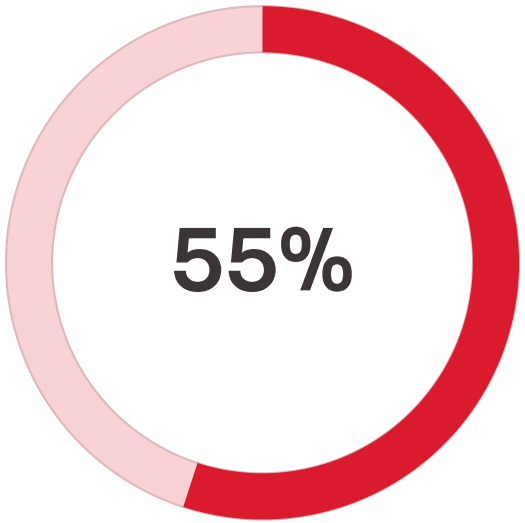
Optimize Precision

Fine-tune models to minimize false positives given the 13.4% spam class imbalance.

Implementation Impact



Reduction in inbox clutter through automated filtering



Improvement in threat detection accuracy

Scalable solution replacing manual moderation

TECHNOLOGY STACK

Tools & Technologies

PostgreSQL

Structured data storage and SQL-based pattern analysis for email classification.

Python & NLP

Natural language processing, TF-IDF vectorization, and machine learning model development.

Power BI

Interactive dashboards for real-time monitoring of spam detection KPIs and trends.

SUMMARY

Conclusion

This project demonstrates an **end-to-end data science solution** combining SQL, NLP, machine learning, and visualization to address a real-world business problem.

01

Comprehensive Analysis

5,572 emails analyzed revealing distinct spam patterns

02

Actionable Insights

Clear differences in length, vocabulary, and keywords identified

03

Scalable Foundation

Strong basis for automated spam detection systems