Spam Email Classification

By Gautam S

Data Analysis, Model Training, and Evaluation

Project Goal & Workflow

01 Goal

Using the UCI Spambase dataset, we'll build an effective machine learning model to classify emails as Spam or Ham.

02 Exploratory Data Analysis (EDA)

03 Data Preprocessing & Feature Engineering

04 Model Training and Selection

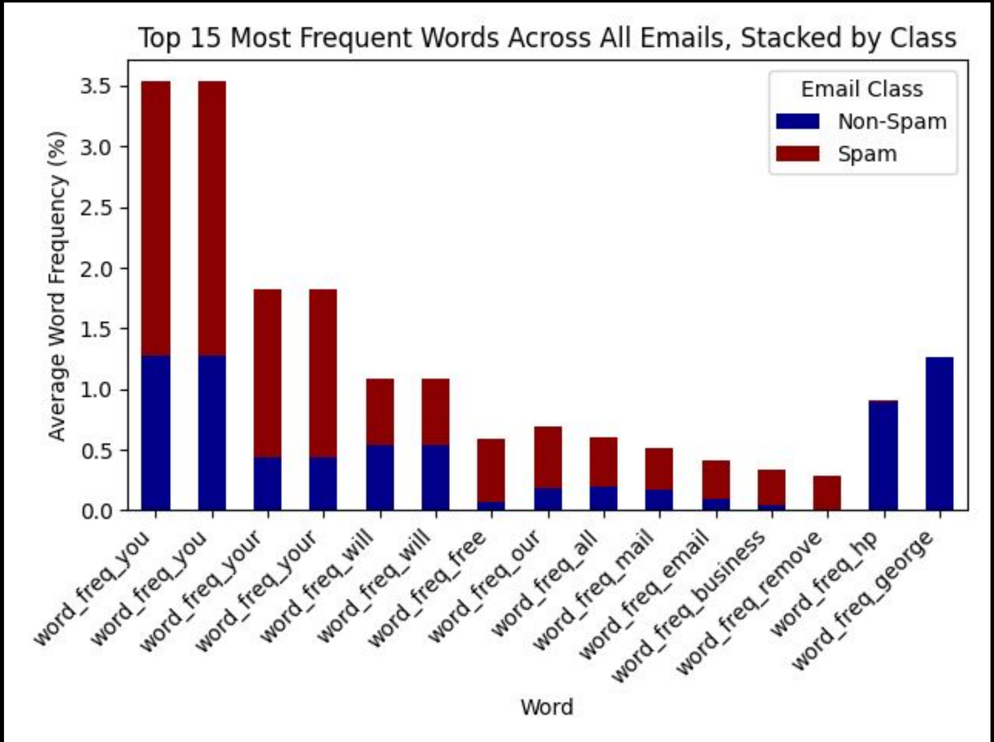
05 Final Evaluation on Unseen Data

01

Exploratory Data Analysis

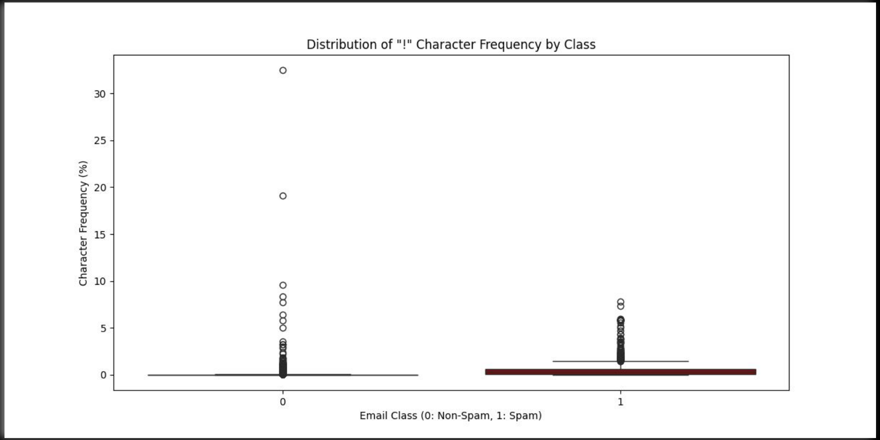
EDA: Spam Emails Use Specific Keywords

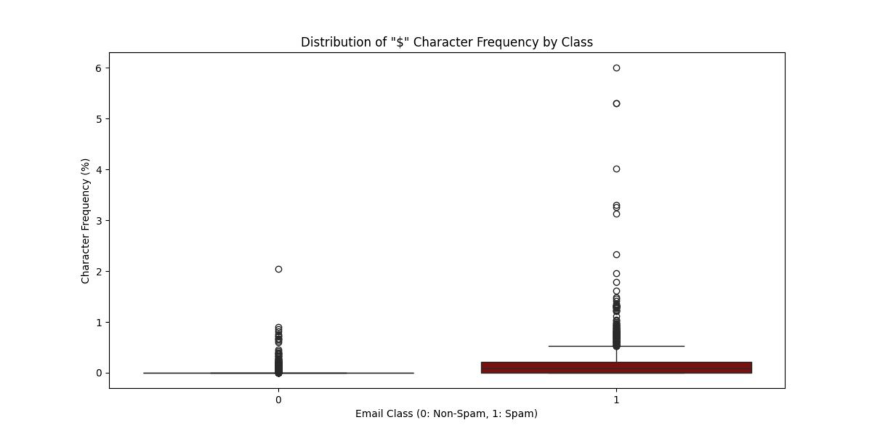
Exploratory Data Analysis reveals that spam emails frequently contain specific keywords, highlighting patterns in their content.



EDA: Spam Emails Emphasize Urgency & Money

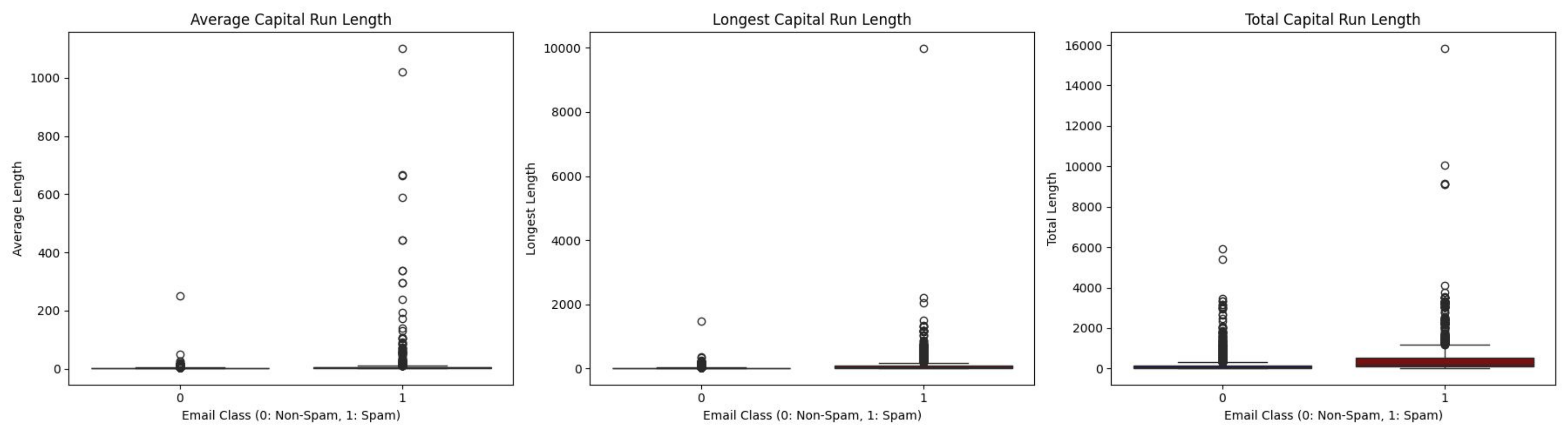
Spam emails often leverage psychological triggers, using words that convey urgency and financial gain to entice recipients.





EDA: Spam Relies on Excessive Capitalization

A common characteristic of spam emails is the overuse of capitalization, a tactic to grab attention and emphasize certain words.

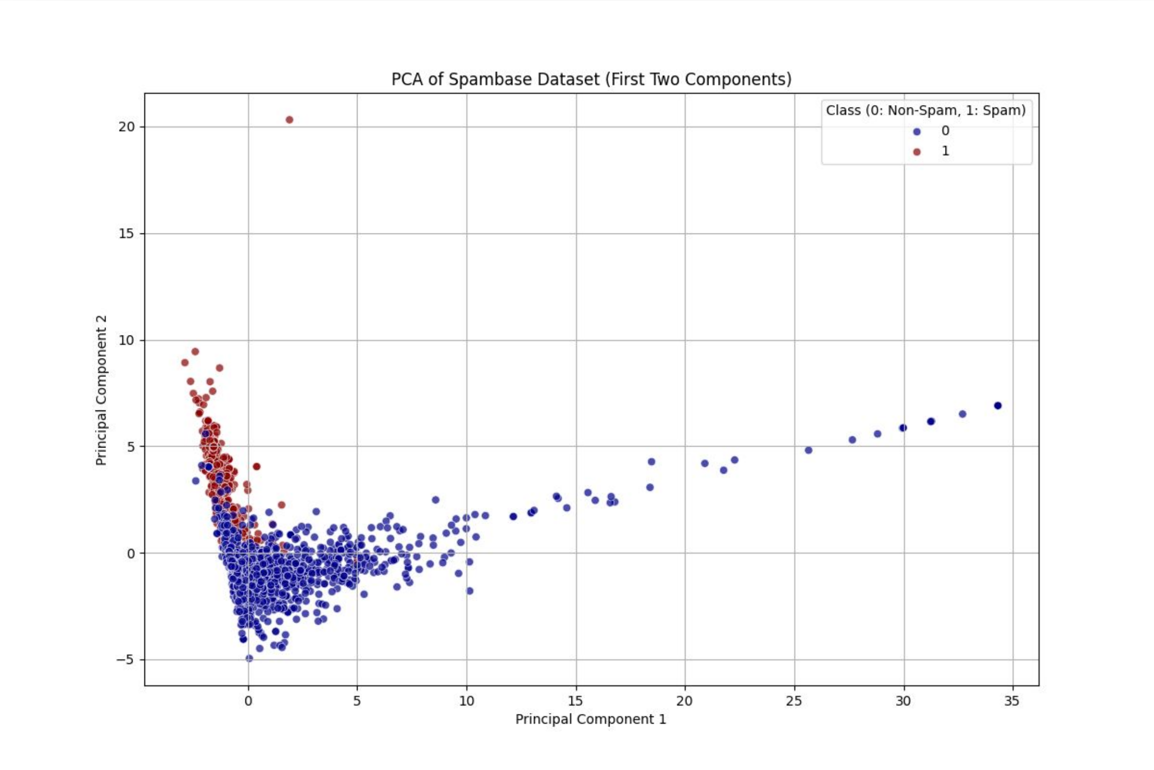


02

Preprocessing

Preprocessing: Can We Simplify the Features?

Principal Component Analysis (PCA) showed some class separation, but also significant overlap, indicating that reducing to 2 components loses too much critical information.



Significant Overlapping

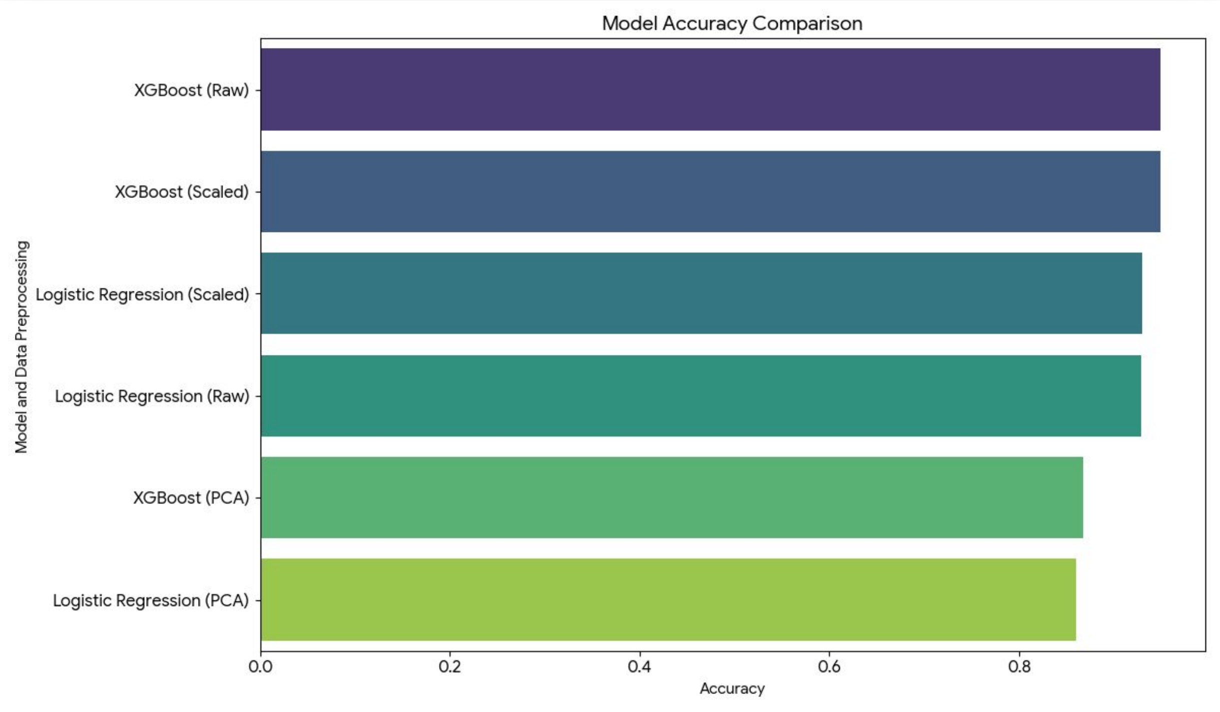
03

Model Training

Model Training & Results

Two models, Logistic Regression and XGBoost, were tested across three data variations: Raw, Scaled, and PCA

Data	Model	Accuracy	Precision
Scaled	XGBoost	0.9490	0.9965
Scaled	Logistic Regression	0.9294	0.9209
Raw	Logistic Regression	0.9283	0.9160
PCA	XGBoost	0.8283	0.8160
PCA	Logistic Regression	0.8599	0.8534



Performance of Different Models

Conclusion & Next Step

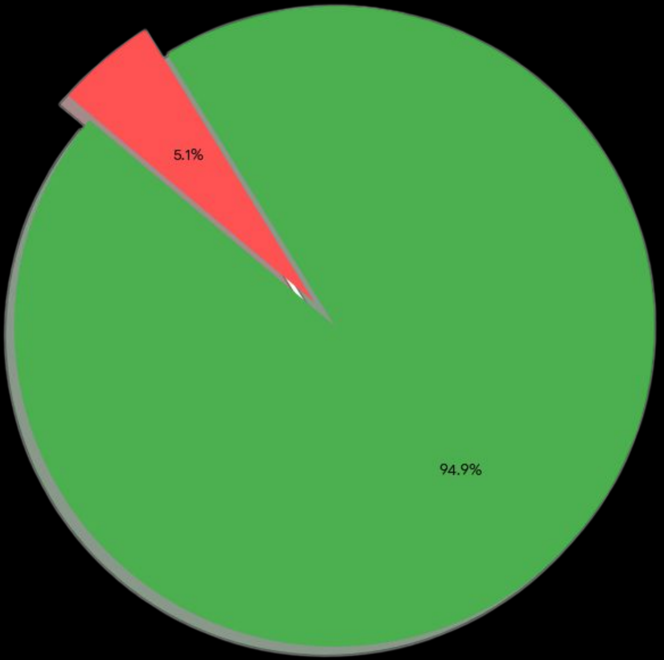
Best Model

XGBoost Classifier on raw data achieved the highest accuracy (94.9%) and precision (93.7%).

Key Takeaway

Feature scaling and PCA did not improve XGBoost performance, indicating raw features hold the most signal.

Best Model : XGboost with Raw DATA



Q&A

Thank You

Gautam S