Predicting Spam Emails

DS 4002 - 02/17/2025

Group 2: Ella Thomasson (Leader), Maggie Crowner, Emily McMahon

Motivation and Goals

Motivation : Spam emails have the intent to harm by encouraging users to download malicious software

Goal : Investigate the categorization of an email as spam or not spam. Be able to predict, with 80% accuracy, whether or not an unseen email is spam based on the contents of the email.

- **Hypothesis** : We can detect spam emails with 80% accuracy

Research Question : How accurately can we predict whether or not an email is spam based on the contents of the email? How do the accuracy results vary across different classifiers?

Data Acquisition/Explanation

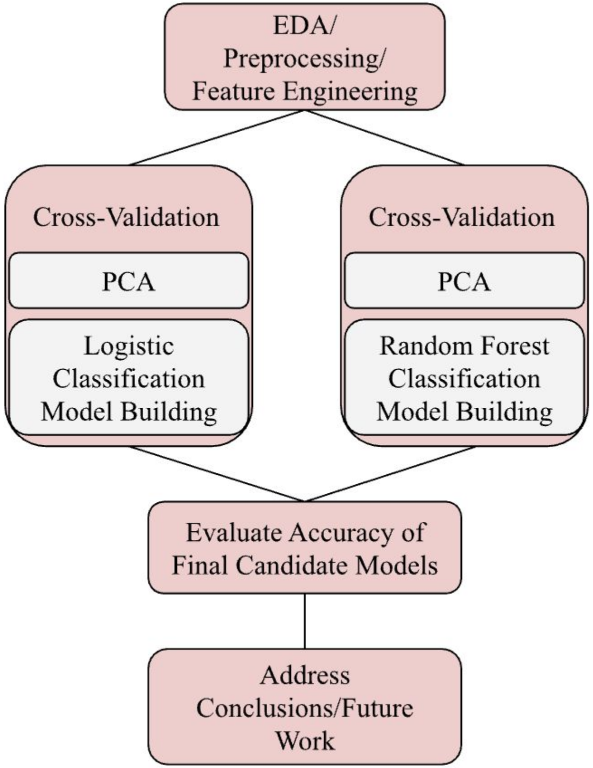
84 Rows(58 non-spam and 26 spam) and 3 Columns

Title	str	A string containing the subject line of the email.
Text	str	A string containing the body text of the email.
Type	str	Whether or not the email is spam - "spam" or "not spam" (response variable)

84 Rows(58 non-spam and 26 spam) and 104 columns

Text Sentences	str	All instances of punctuation marks that indicate the end of a sentence
Text Words	str	The number of words in each email's body
Company	dummy	Represents whether or not the email body has "company" in it
...	...	...
Type	str	Whether or not the email is spam

Analysis Plan



Justification

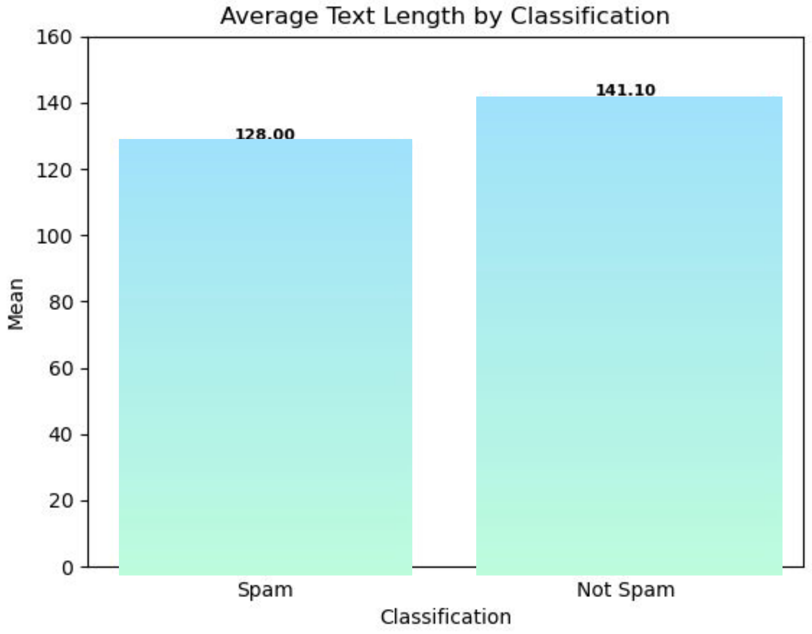
- Used 5-fold CV to obtain more accurate estimates on how the model will perform on unseen data
- Used PCA to decrease dimensionality before jumping into model building(bc #col>#rows)

Tricky Analysis Decisions

- Deciding which variables to include in the model
- Cut off values for dummy variables(10)- arbitrary
- Number of Components for PCA(0.8)- arbitrary

Bias and Uncertainty

- Small sample size
- Unequal number of spam and non-spam emails in the data



Results

Metric	LC Value	RF Value	Meaning
Accuracy	0.7619	0.7143	The probability of correctly predicting spam/not-spam emails
Precision	0.6500	0.6667	The probability of correctly predicting the positive class (a spam email)
Recall	0.5000	0.1538	Proportion of actual spam cases that the model correctly classifies as spam
F1-Score	0.5652	0.2500	The harmonic mean of precision and recall
ROC-AUC	0.8345	0.6963	The test's ability to distinguish between spam and non-spam individuals

LC Conf Matrix	Predicted Not Spam	Predicted Spam
Actual Not Spam	51	7
Actual Spam	13	13

RF Conf Matrix	Predicted Not Spam	Predicted Spam
Actual Not Spam	56	2
Actual Spam	22	4

Conclusions

- Would choose the Logistic Regression Model based on the accuracy metrics
- We did not quite meet our 80% accuracy goal, but we got fairly close with a 76.2% accuracy
- The RF model seems to not detect "spam" emails well, based on recall/F1-score

Next Steps

- Tune parameters in the RF
- Add more variables(or use a larger threshold for words to include within the models)
- Obtain a dataset with a larger number of rows
- Try different classification models(Neural Networks, SVM, Naive Bayes, etc.)
- Would Lasso have been a better method than PCA?

References & Resources

Github : https://github.com/maggiecrowner/DS4002-Project1/tree/main

References :

[1] "What Is Spam Filtering? How Do Spam Filters Work?," *Fortinet*. https://www.fortinet.com/resources/cyberglossary/spam-filters.

[2] Ghazala Nasreen, Muhammad Murad Khan, M. Younus, B. Zafar, and Muhammad Kashif Hanif, "Email spam detection by deep learning models using novel feature selection technique and BERT," *Egyptian Informatics Journal/Egyptian Informatics Journal* , vol. 26, pp. 100473–100473, Jun. 2024, doi: https://doi.org/10.1016/j.eij.2024.100473.

[3] A. Ajay, "What are the Advantages and Disadvantages of Random Forest?," *Pickl.AI*, Sep. 30, 2024. https://www.pickl.ai/blog/advantages-and-disadvantages-random-forest/.

[4] A. Khan, "Email Spam Detection with Machine Learning: A Comprehensive Guide," Medium, Mar. 22, 2024. https://medium.com/@azimkhan8018/email-spam-detection-with-machine-learning-a-comprehensive-guide-b65c6936678b

[5] GeeksforGeeks, "CrossValidation vs. Bootstrapping," GeeksforGeeks, Jun. 27, 2024. https://www.geeksforgeeks.org/cross-validation-vs-bootstrapping/

Thank You for Listening! Questions?