Assignment 01 : Prototype of Spam Classifier

Executive Summary

1. Hypothesis

- Given that the vocabulary structure and word frequencies of a spam message is quite different from a regular message, machine learning models should be able to distinguish between them well.
- The list of words with high frequencies in Spam messages have very little to no overlap with the regular messages.

2. Experiments

a) Custom Heuristics Model (Experimental)

- Interestingly, even a simple heuristics-based model is able to accurately predict messages 95% of the times. However precision score for spam messages is on the lower side at 79% (on validation data).
  - The heuristic function maps each message to spam/ ham scores which are based on the incidence of spam and ham vocabulary in the message.

b) Machine Learning Models

- While all three ML models (Decision Tree, Logistic Regression and SVM) are yielding good results and keep the 'False Positive Rate' well within the benchmark of 10%; LR and SVM do better (% FPR - 0.1%, 0.3% respectively on unseen test data).
- However, SVM does better on Recall compared to LR (80% vs 73% on test data)
- Given than the performance of base models is already quite good, hyper parameter tuning to find the best model (in case of Decision Tree), precision-recall curve to adjust classification thresholds (in case of Logistic Regression) aren't yielding a significant upside.

3. Model Choice

- Both Logistic Regression and SVM are good choices, with a slight edge for SVM for maintaining a balance between Precision and Recall.
- Other aspects such as computational complexity could be considered while making a choice of model for production.