Spambase

Group 3: Collin Ladina, Judd Douglas, Matthew Karnatz, Ryan Cullen, Shaan Patel

Context and Question

- **Context**: According to EmailToolTester.com for 2023, **160 billion** spam emails were sent every day, with that number continuing to rise year after year.

- **Dataset Overview**: “Spambase” in the UCI Machine Learning Repository; 4,601 instances and 57 features include word frequencies, character frequencies, and capital letter percentages. The target variable is binary: 1=spam, 0=non-spam.

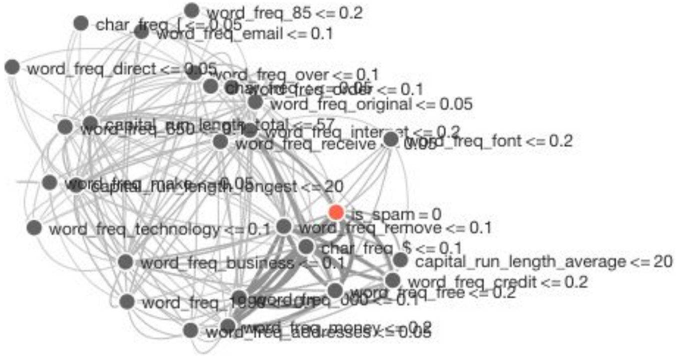
- **Problem Statement**: Can we accurately classify an email as spam or non-spam using the features provided in the dataset? This is crucial for improving email security and reducing the clutter in inboxes.

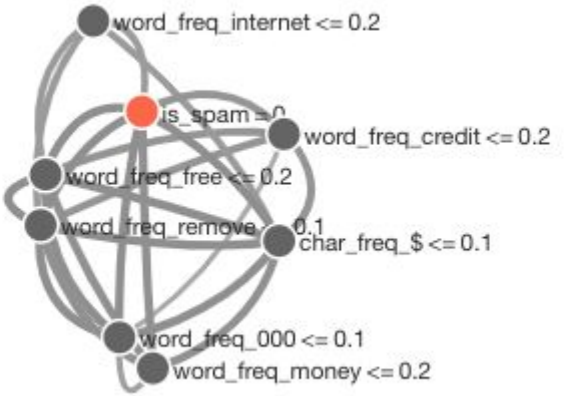
Data and Variables

- **Key Variables**: Frequencies of specific words (e.g., "make," "address"), characters (e.g., '!'), and percentage of capital letters.
- **Outliers**: Outliers are present, particularly in capital letter percentages, but they will be kept for analysis to maximize prediction coverage.
- **Missing Values**: Missing values will be the median of the feature involved to ensure all data is handled and considered but in the best way possible

Descriptive Visualizations

Purpose: Use visualizations to explore relationships between features and the target variable.





Model Selection

- **Chosen Model: Boosted**
- **Justification:** It limited the amount of Type I errors
  - **Logistic Regression** for simplicity and interpretability (binary classification).
  - **Random Forest** for handling feature complexity and potential interactions between word frequencies.
  - **Boosted** for maximizing accuracy and reducing false positives

Next Steps

- Ready to proceed with model building and further evaluation. The dataset and exploratory analysis provide strong foundations for effective spam classification.

Questions?