MIST7770 Final Project “Spambase”
Group 3: Ryan Cullen, Judd Douglas, Matthew Karnatz, Collin Ladina, Shaan Patel

Context and Question

Spam emails have become a consistent burden on the inboxes of most users. In 2023, some estimates reported that 160 billion up to 320 billion spam emails were sent each day and “94% of malware is delivered via this medium” (Forbes, EmailTester). This number continues to proliferate, and while many of these emails may not contain phishing attacks or malware, they are still a massive inconvenience for recipients who do not know how to deal with them. Because of this, an accurate classification method of spam emails would greatly improve the quality of user experience and security on the internet.

The primary objective of this project is to accurately classify incoming emails as ‘spam’ or ‘not spam’, while limiting the prevalence of Type I errors. Our problem statement is: *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

Our data source, Spambase, comes from the UC Irvine Machine Learning Repository. It consists of 4601 emails with 57 features that capture the word, character, symbol, and case frequencies. Our target variable is binary, where 0 is ‘not spam’ and 1 is ‘spam’. In order to handle outliers in our data, we ____________. Since the missing numeric data does not follow any noticeable pattern, the missing values are replaced with the median value of the feature involved. We do this because the median value is particularly robust to outliers and is more representative of the data. This also helps preserve the distribution of our data, whereas using the mean would skew the distribution. This will help ensure that all the data is handled and considered in the proper way.

Descriptive Visualizations

Show how your variables are related to the target using scatter plots, bar graphs, point graphs, etc. Should you keep all the variables?

The purpose of using descriptive visualizations is to help illustrate the distribution of the variables, as well as the associations they have with each other. Since certain words, phrases, symbols, or mannerisms (case) are often used together, we can visualize their relationships to help understand the model prediction power. For example, an all caps message designed to grab

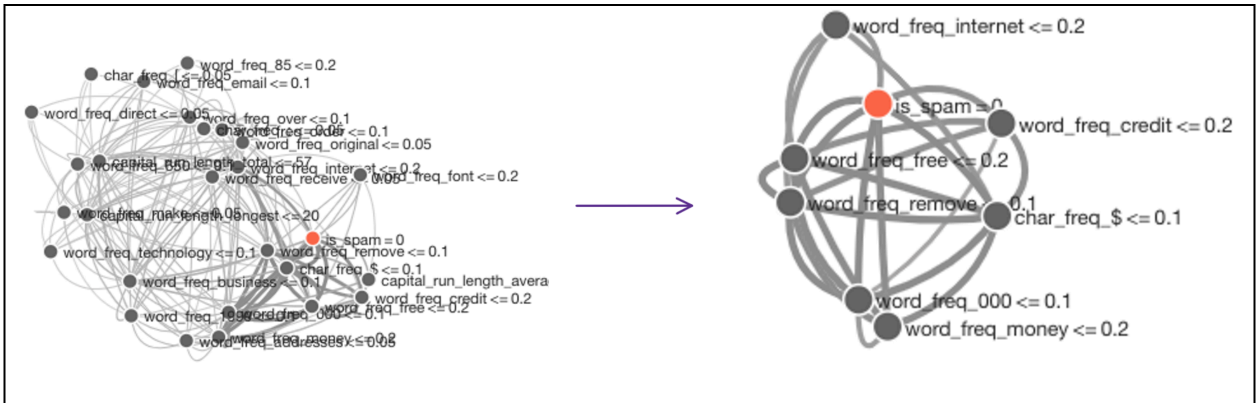


Figure 1. Association web with increasing leverage

the users attention may be convoluted with dollar signs, exclamation marks, or the word "free". This relationship is visualized by the association webs in figure 1, and is further clarified by removing some extraneous variables by increasing the leverage in the BigML software.

Method

• Describe why these techniques are useful for your project and adequate for the data that you are analyzing.

• Choose one or a set of metrics to evaluate the performance of your model and briefly discuss why you have chosen such metric(s). For example, if your target variable is a class, which performance measures will you use and why?

• Explain the process you adopted for building the model – i.e., did you split the data into training and testing sets? Do you have unbalanced data? What did you do to prevent leakage and address overfitting?

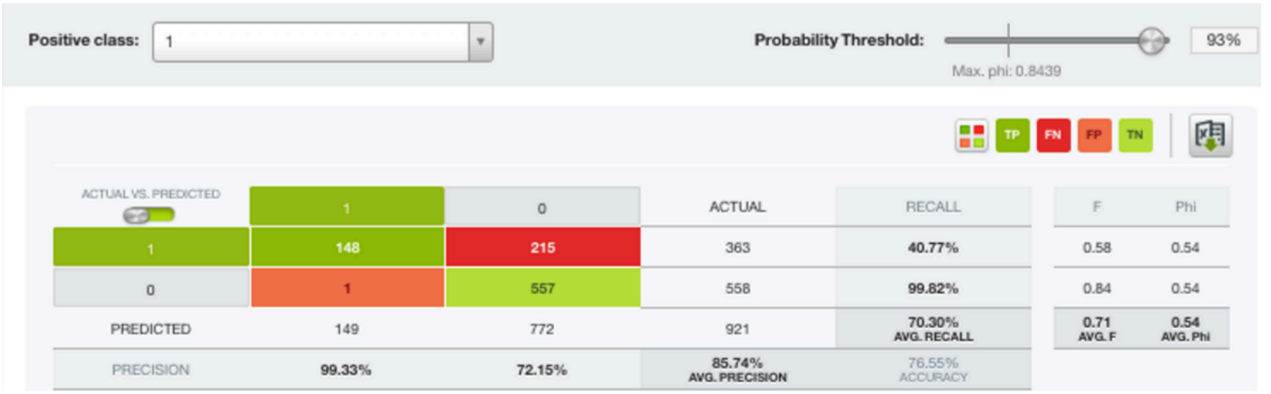
When exploring our data and deciding on what type of data analysis technique we should use to build our model, a few factors were taken into account. First, due to the binary classification goal of our project, we contemplated using logistic regression. It is a straightforward method that gives us an easily interpretable output where we can see the individual contribution of each feature. It also works well with text data. The second method we debated using was a random forest. This ensemble approach is a powerful predictor when analyzing non-linear relationships like frequencies. A random forest would help incorporate patterns found within these spam emails to correctly identify spam. Finally, we looked at using a gradient boosted tree model. Boosting is a viable approach to answering our problem statement since the iterative approach it takes translates to higher performance by implementing error correction. But most importantly, it also serves to limit the Type I errors (false positives), which we emphasize as an important function of the model. For these reasons, we ultimately decided that the boosted model was the best fit in this scenario. This choice was further reinforced as it outperformed its competitor models after running each technique in BigML.

To begin the Gradient boosting method, we first split the Spambase data into a training, testing, and valuation set with the respective proportions of _70___15_15_??? This is very useful to measure the performance of the model in an unbiased manner. We made further provisions to cull other problems, such as overfitting, by _____

Results

Report the results of your final model. If you use regressions, make sure to state clearly whether a relationship is statistically significant. Discuss the direction of the relationship along with the amount of variance explained.

• Write a paragraph (slide) or two with the answer(s) to your question(s) based on your model results. Explain how the model helped you to answer your question(s) (e.g., if your question is, what is the relationship between X and Y? then focus on the statistical significance and direction of the relationship and refer to the coefficient estimates and p-values or CI of the model to discuss these in the text).



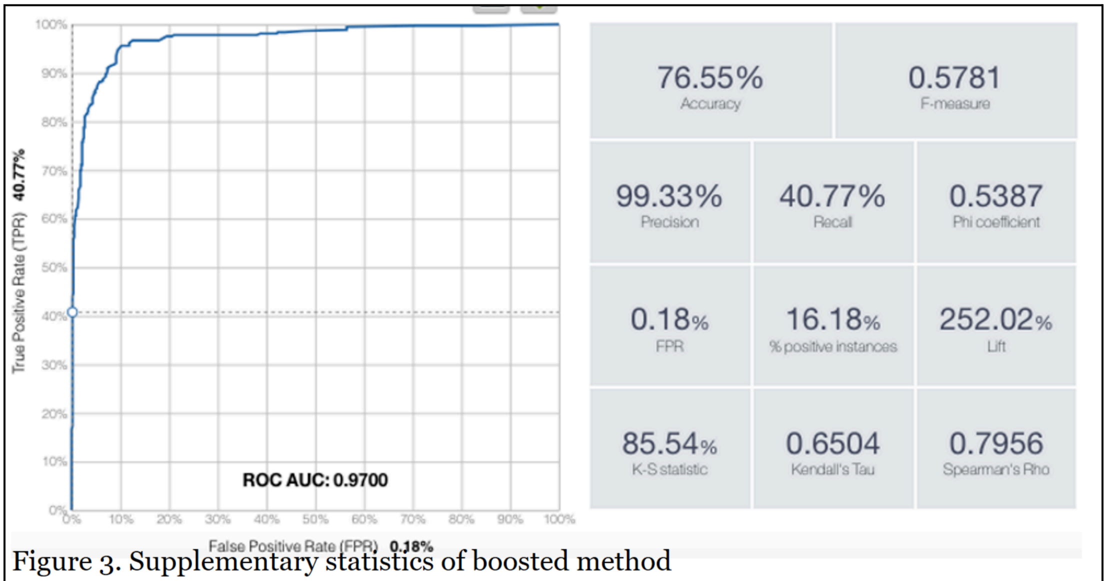
Positive class: 1

Probability Threshold: 93%

Max. phi: 0.8439

ACTUAL VS. PREDICTED	1	0	ACTUAL	RECALL	F	Phi
1	148	215	363	40.77%	0.58	0.54
0	1	557	558	99.82%	0.84	0.54
PREDICTED	149	772	921	70.30% AVG. RECALL	0.71 AVG. F	0.54 AVG. Phi
PRECISION	99.33%	72.15%	85.74% AVG. PRECISION	76.55% ACCURACY		

Figure 2. Correlation matrix of boosted method



76.55% Accuracy	0.5781 F-measure	
99.33% Precision	40.77% Recall	0.5387 Phi coefficient
0.18% FPR	16.18% % positive instances	252.02% Lift
85.54% K-S statistic	0.6504 Kendall's Tau	0.7956 Spearman's Rho

Figure 3. Supplementary statistics of boosted method

Limitations

• Identify one or two limitations in your work.

• Would you like to have collected additional data that you could have used in your model (be specific about what data you wish you had and why it would have been useful)?

• Is there something holding you back from making an inference about causality?

Although our model is robust and performs well, that is not without a few caveats. We are limited in our prediction power under certain conditions _____

- Languages other than english
- We treat each email as a unique sender with no name or reputation of malice. A database containing past offenders would help identify new ones if the IP address is the same.
    - It may also be nice to have a catered model for each business with the features tuned toward
- Does not directly transcribe image text, so pictures/attachments could bypass model
- Model could be outdated
- 2601 observations is relatively small. More data is likely out there, just not public.

Recommendations

- *Make a set of final recommendations that are based on your analysis of the data and question(s), as well as the inherent limitations described.*
- *What are the key takeaways and your call to action? Present a brief list of high-level points.*

Along with an employee awareness program to train employees to be cognizant of potential spam threats, this model should be continuously updated