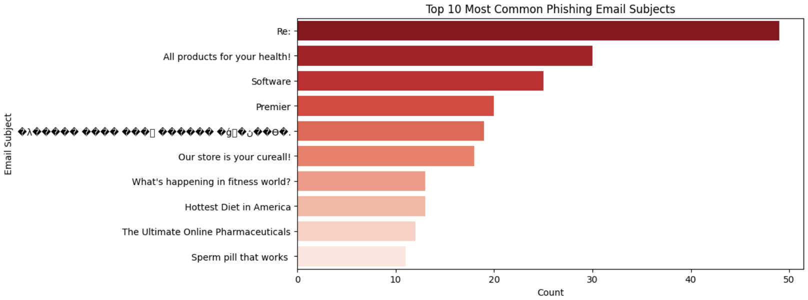
# Project Results

## Research Question

What characteristics make an email most likely to be a phishing attempt?

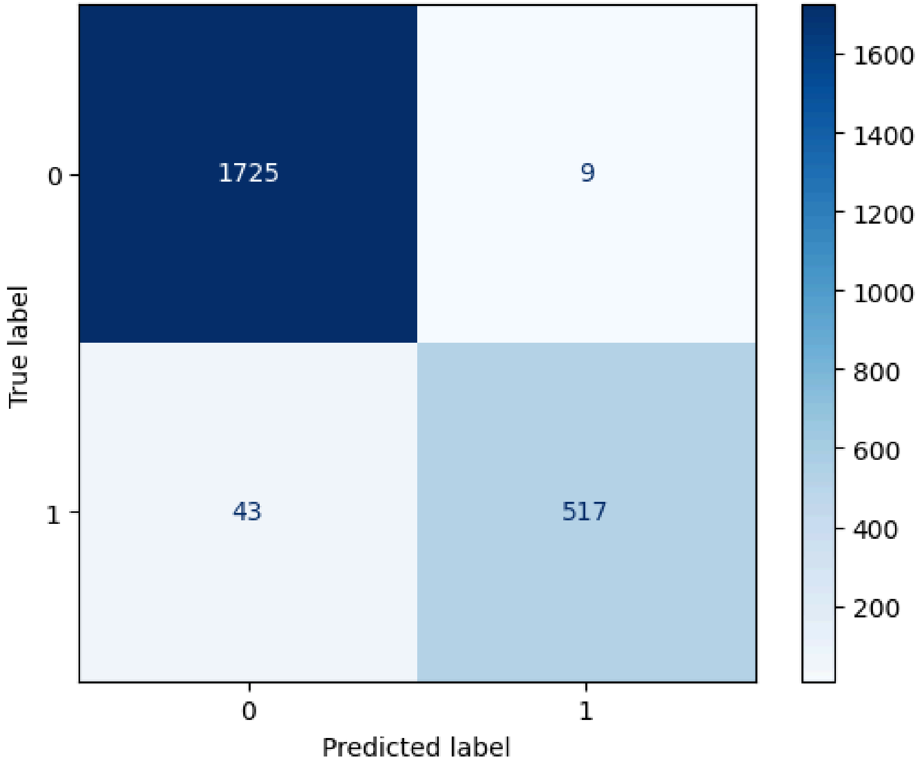
## Final Results



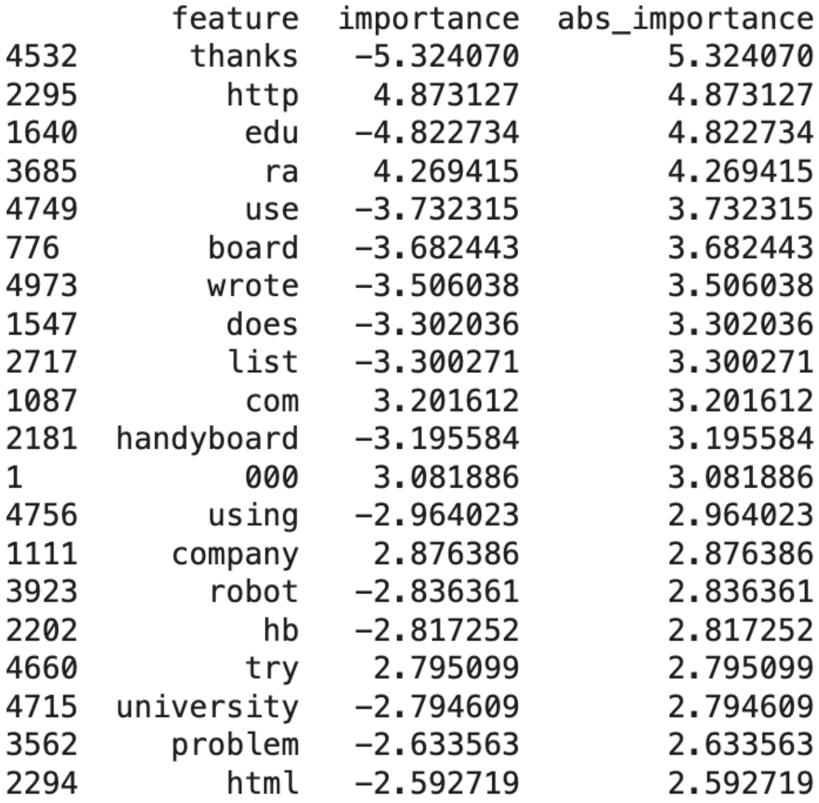
Before delving into full analyses, we wanted to determine which of the three models performed the best. These models included support vector machine (SVM) and TF-IDF vectorization, random forest, and linear regression. Of these, the linear regression had the highest accuracy of 97.69% and the highest precision of 98.29%. Additionally, this model had a recall percentage of 0.9232, meaning it was able to successfully identify positive phishing emails 92.32% of the time. Finally, linear regression had the second highest F1 score of 0.9521 and the highest ROC AUC score of 0.9982, meaning this model was incredibly successful in predicting what emails were and were not phishing.

| | **Linear Regression** | **Random Forest** | **SVM/TF-IDF Vectorization** |
|---:|:---:|:---:|:---:|
| **Accuracy** | 0.9773 | 0.9769 | 0.9647 |
| **Precision** | 0.9829 | 0.9828 | 0.9650 |
| **Recall** | 0.9232 | 0.9210 | 0.9647 |
| **F1 Score** | 0.9521 | 0.9509 | 0.9648 |
| **ROC AUC Score** | 0.9982 | 0.9982 | 0.9871 |

To better visualize the accuracy of this linear regression model we settled on, we decided to create a confusion matrix to analyze the relationship between predicted and true labels, either phishing or not. We discovered that out of 1,734 non-phishing emails, our model correctly labeled 1,725 of them, with only 9 being misidentified. Similarly, out of 560 phishing emails, our model correctly labeled 517 and only incorrectly labeled 43. Though the accuracy was higher in determining non-phishing emails, we can still say that this linear regression model is very effective overall.



Knowing this, we decided to use linear regression to determine which characteristics are indicative of a phishing email. This is the most important factor in being able to determine, and predict, whether an email is phishing or not. We decided to look at feature importances, in which a positive importance value means an increased chance the email is phishing, and a negative importance value means an increased chance the email is not phishing. Interestingly, we found that characters like “thanks” and “edu” are two of the most important features present in non-phishing emails. This is likely due to the fact that legitimate correspondence have human touches like saying thanks or including any educational links. Furthermore, “http”, “com”, “000”, and “company” were highly indicative of phishing emails. Many of these emails include links to external and insecure websites indicated by the “http” instead of “https” and likely less reputable domains such as “com” instead of “edu” or “gov”. The presence of triple zeros is likely due to the presence of large numbers with multiple zeros present at the end, used to grab the recipient's attention, either malicious saying they owe money or trying to pull at their greed by stating they could be endowed with money if they respond. Finally, “company” is a very general term, and if a specific company were to be contacting someone they would be much more likely to utilize their actual name instead.



```
         feature  importance  abs_importance
4532      thanks   -5.324070        5.324070
2295        http    4.873127        4.873127
1640         edu   -4.822734        4.822734
3685          ra    4.269415        4.269415
4749         use   -3.732315        3.732315
776        board   -3.682443        3.682443
4973       wrote   -3.506038        3.506038
1547        does   -3.302036        3.302036
2717        list   -3.300271        3.300271
1087         com    3.201612        3.201612
2181  handyboard   -3.195584        3.195584
1            000    3.081886        3.081886
4756       using   -2.964023        2.964023
1111     company    2.876386        2.876386
3923       robot   -2.836361        2.836361
2202          hb   -2.817252        2.817252
4660         try    2.795099        2.795099
4715  university   -2.794609        2.794609
3562     problem   -2.633563        2.633563
2294        html   -2.592719        2.592719
```

We also determined the prevalence of different categories of special characters present in phishing and non-phishing emails. Overall, we really only found a higher occurrence of uppercase letters in phishing emails, which is likely a tactic used to better grab the recipient's attention and persuade them to think the email contains pertinent and time-sensitive information.