Task: Given a dataset of phishing & normal emails, your task is to detect if a given email is a phishing email or not using a ML-led solution.

Dataset: Phishing and Normal mails in .eml format provided by Canary mail

Solution implemented by: Dhruvin Kakadia (dhruvinkakadia@gmail.com)

Approach

- Extract the mails from the eml files provided.

- Pre-process the mails by removing extra spaces, line breaks, etc and extracting the main content of the mail using html parsers like beautiful soup

- Extract elements like the links in the mail, links embedded in the images, and keywords like dear, account, eBay, PayPal, etc. which are especially important in detecting phishing mails.

- Generate feature vector for a particular mail with the help of the elements extracted in the previous step.

- Generate training and testing sets of the given mails which contains both, phishing and normal mails.

- Try out different Machine learning models like Support Vector Classifier, Logistic Regression, Decision Trees and ensemble models like Random Forest, Gradient boosting, XGBoost, AdaBoost, Extra Trees and Voting Classifier.

- Tune all the models to find the best hyperparameter space for the respective models using Grid search cross validation.

- Evaluate models by comparing their accuracies on the holdout set and select the best model after the evaluation

Attributes used for feature representation

- **Links:** Numerical attribute representing the total number of links present in a given mail.

- **Dears:** Categorical attribute representing if the term ‘dear’ is present in a given mail.

- **HTML:** Numerical attribute representing the number of HTML tags/elements in a given email.

- **Please:** Categorical attribute representing if the term ‘please’ is present in the given mail.

- **Account:** Categorical attribute representing if the term ‘account’ or ‘accounts’ is present in the given mail.

- **Images:** Numerical attribute denoting the number of images present in the given mail.

- **Bank:** Categorical attribute representing if the term ‘bank’ is present in the given mail.

- **Protect:** Categorical attribute representing if the term ‘protect’ is present in the given mail.

- **Click:** Categorical attribute representing if the term ‘click’ or ‘select’ or ‘visit’ is present in the given mail.

- **Ebay and Paypal:** Categorical attribute representing if either the term ‘ebay’ or the term ‘paypal’ is present in the given mail.

- **Thank You:** Categorical attribute representing if the term ‘thank you’ is present in the given mail.

Importance of attributes

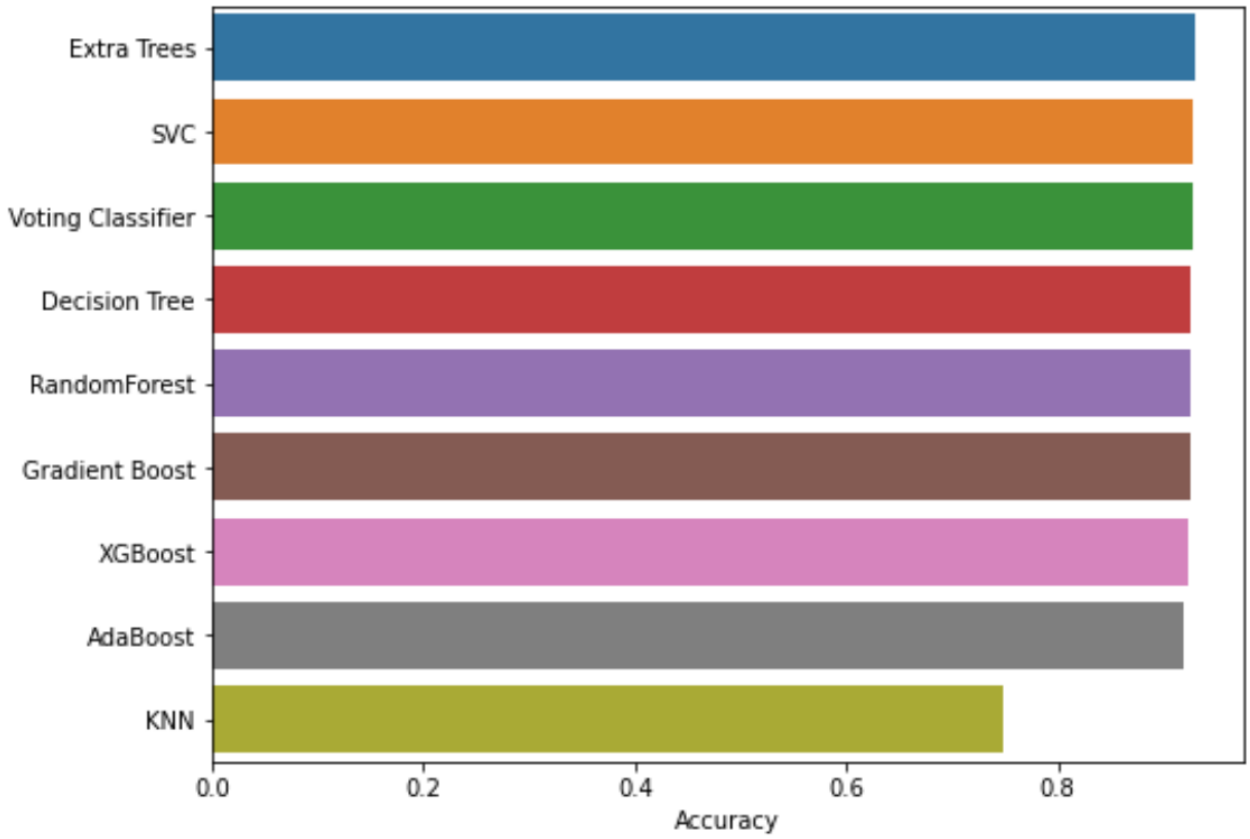
Importance of a particular attribute is given by the feature_importances_ attribute of the extra trees classifier. It is shown below

feature_importances_ returns a list of numbers, equal to the number of attributes in the dataset, and their sum is equal to one. This means that the number returned for a particular attribute is relative to other attributes.

The most important feature is the ‘Please’ feature with over 26% weightage, second most important feature being ‘Dears’ with over 16% weightage. ‘Dears’, ‘Thank You’ and ‘Please’ are the 3 most important features with their combined weightage in predicting if a mail is phishing or not being over 56%.

Evaluation of models

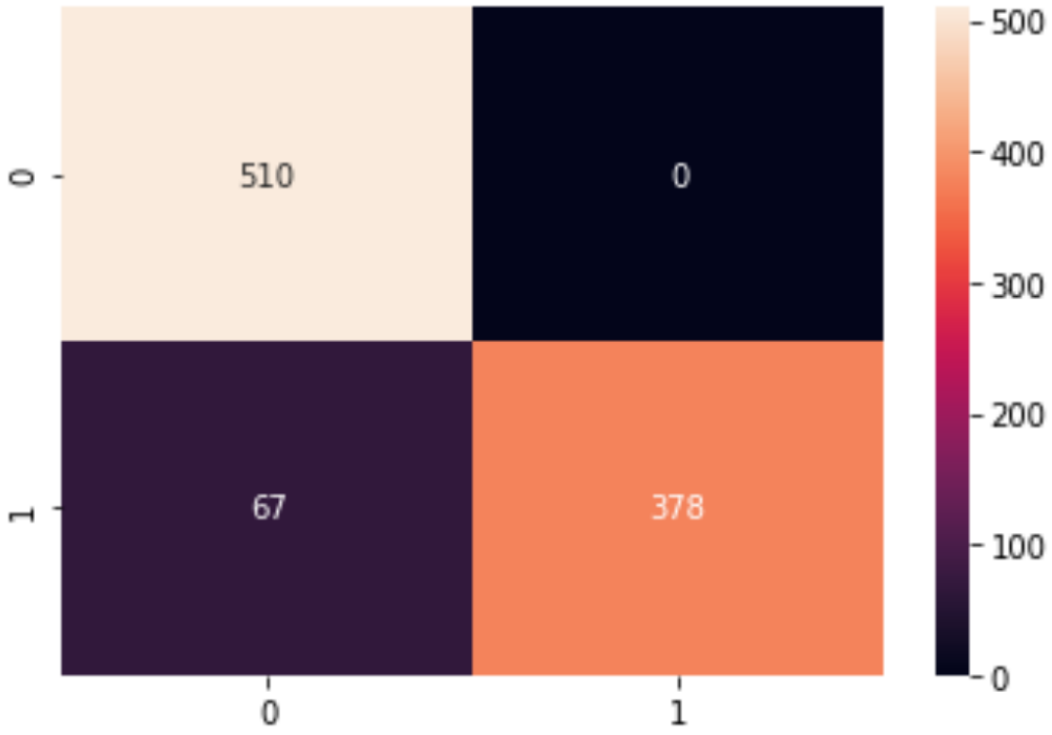
Below is the bar plot of the accuracies of all the models trained.



As you can see, the best model is Extra Trees Classifier with an accuracy of around 93%. Following are the accuracies of all the trained models:

```
Accuracies:
SVC - 0.9277486910994764
RandomForest - 0.9246073298429319
XGBoost - 0.9235602094240838
Gradient Boost - 0.9246073298429319
AdaBoost - 0.9193717277486911
Decision Tree - 0.9256544502617801
Extra Trees - 0.9298429319371728
KNN - 0.7476439790575916
Voting Classifier - 0.9277486910994764
```

Confusion matrix and Classification report for predictions on Extra Trees Classifier



The given model has high precision but a fairly lower recall, which could be further improved by extracting and adding more relevant features. The following is the classification report for the Extra Trees Classifier model.

	precision	recall	f1-score	support
0	0.88	1.00	0.94	510
1	1.00	0.85	0.92	445
accuracy			0.93	955
macro avg	0.94	0.92	0.93	955
weighted avg	0.94	0.93	0.93	955