CS 286

PROJECT REPORT

By

Nishita Narvekar & Adil Khan

Table of Contents

Aim

The aim of this project is to identify whether a website is a phishing website or a legitimate website.

Dataset

The dataset which we are gonna use for this task is Phishing Websites Data Set. [Link: https://archive.ics.uci.edu/ml/datasets/phishing+websites#] It has 30 features & 11056 observations. Here is a short description for each feature in the data set:

Features

The features can be categorically split up into four categories based on their relevance.

Address Bar based Features

- Using the IP Address: If an IP address is used as an alternative of the domain name in the URL then it can be a phishing URL.
- Long URL to Hide the Suspicious Part: Phishers can use long URL to hide the doubtful part in the address bar
- Using URL Shortening Services : Phishing is accomplished by means of an “HTTP Redirect” on a domain name that is short, which links to the webpage that has a long URL.
- URL’s having “@” Symbol: Using “@” symbol in the URL leads the browser to ignore everything preceding the “@” symbol and the real address often follows the “@” symbol.
- Redirecting using “//”: The existence of “//” within the URL path means that the user will be redirected to another website.
- Adding Prefix or Suffix Separated by (-) to the Domain: The dash symbol is rarely used in legitimate URLs. Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage.
- Sub Domain and Multi Sub Domains: The URL is classified as “Suspicious” since it has one sub domain, it is classified as “Phishing” since it will have multiple sub domains. Otherwise, if the URL has no sub domains, we will assign “Legitimate” to the feature.
- HTTPS: Uses https and Issuer Is Trusted and Age of Certificate is greater than 1 Year then the website is legitimate
- Domain Registration Length: Based on the fact that a phishing website lives for a short period of time, we believe that trustworthy domains are regularly paid for several years in advance. In our dataset, we find that the longest fraudulent domains have been used for one year only.
- Favicon: If the favicon is loaded from a domain other than that shown in the address bar, then the webpage is likely to be considered a Phishing attempt.
- Using Non-Standard Port: If all ports are open, phishers can run almost any service they want and as a result, user information is threatened.

- The Existence of “HTTPS” Token in the Domain Part of the URL: The phishers may add the “HTTPS” token to the domain part of a URL in order to trick users.

Abnormal Based Features

- Request URL: Request URL examines whether the external objects contained within a webpage such as images, videos and sounds are loaded from another domain.
- URL of Anchor: If the <a> tags and the website have different domain names. This is similar to request URL feature.
- Links in <Meta>, <Script> and <Link> tags: Given that our investigation covers all angles likely to be used in the webpage source code, we find that it is common for legitimate websites to use <Meta> tags to offer metadata about the HTML
- Server Form Handler (SFH): SFHs that contain an empty string or “about:blank” are considered doubtful because an action should be taken upon the submitted information.
- Submitting Information to Email: Web form allows a user to submit his personal information that is directed to a server for processing. A phisher might redirect the user’s information to his personal email.
- Abnormal URL: For a legitimate website, identity is typically part of its URL.

HTML and JavaScript based Features

- Website Forwarding: The fine line that distinguishes phishing websites from legitimate ones is how many times a website has been redirected.
- Status Bar Customization: Phishers may use JavaScript to show a fake URL in the status bar to users.
- Disabling Right Click : Phishers use JavaScript to disable the right-click function, so that users cannot view and save the webpage source code.
- Using Pop-up Window: It is unusual to find a legitimate website asking users to submit their personal information through a pop-up window.
- IFrame Redirection: IFrame is an HTML tag used to display an additional webpage into one that is currently shown.

Domain based Features

- Age of Domain: the minimum age of the legitimate domain is 6 months.
- DNS Record: If the DNS record is empty or not found then the website is classified as “Phishing”, otherwise it is classified as “Legitimate”.
- Website Traffic : legitimate websites ranked among the top 100,000.
- PageRank: 95% of phishing webpages have no PageRank
- Google Index: Usually, phishing webpages are merely accessible for a short period and as a result, many phishing webpages may not be found on the Google index.

- Number of Links Pointing to Page : 98% of phishing dataset items have no links pointing to them
- Statistical-Reports Based Feature: Host Belongs to Top Phishing IPs or Top Phishing Domains then it's a Phishing website otherwise it is a Legitimate

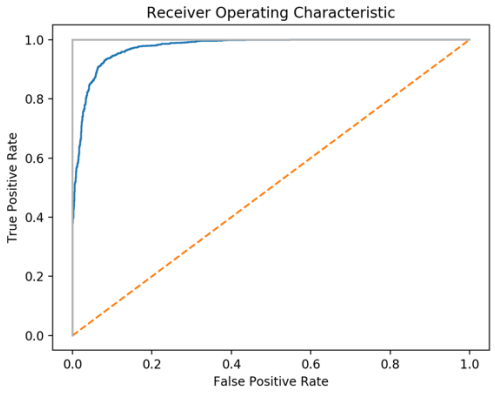
Experiments

I. SVM for all 30 features

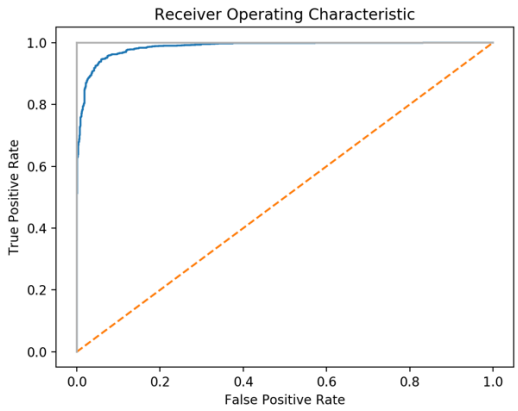
Our approach was to initially just run the model for all the 30 features and see the results. So, we included all the 30 features in the SVM model and observed the following results for linear, rbf & polynomial kernel:

Kernel	Accuracy (%)	AUC
linear	92.32	0.9208
rbf	94.03	0.9379
poly (degree = 3)	94.28	0.9409
poly (degree = 6)	95.15	0.95

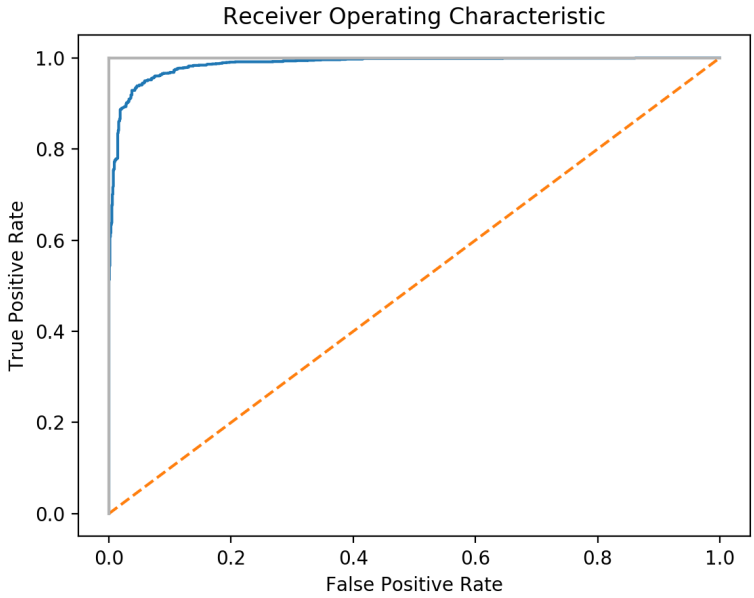
We got the best Accuracy & AUC scores for the polynomial kernel with degree = 6.
Figure below shows the ROC graphs for linear, rbf & polynomial kernel.



linear



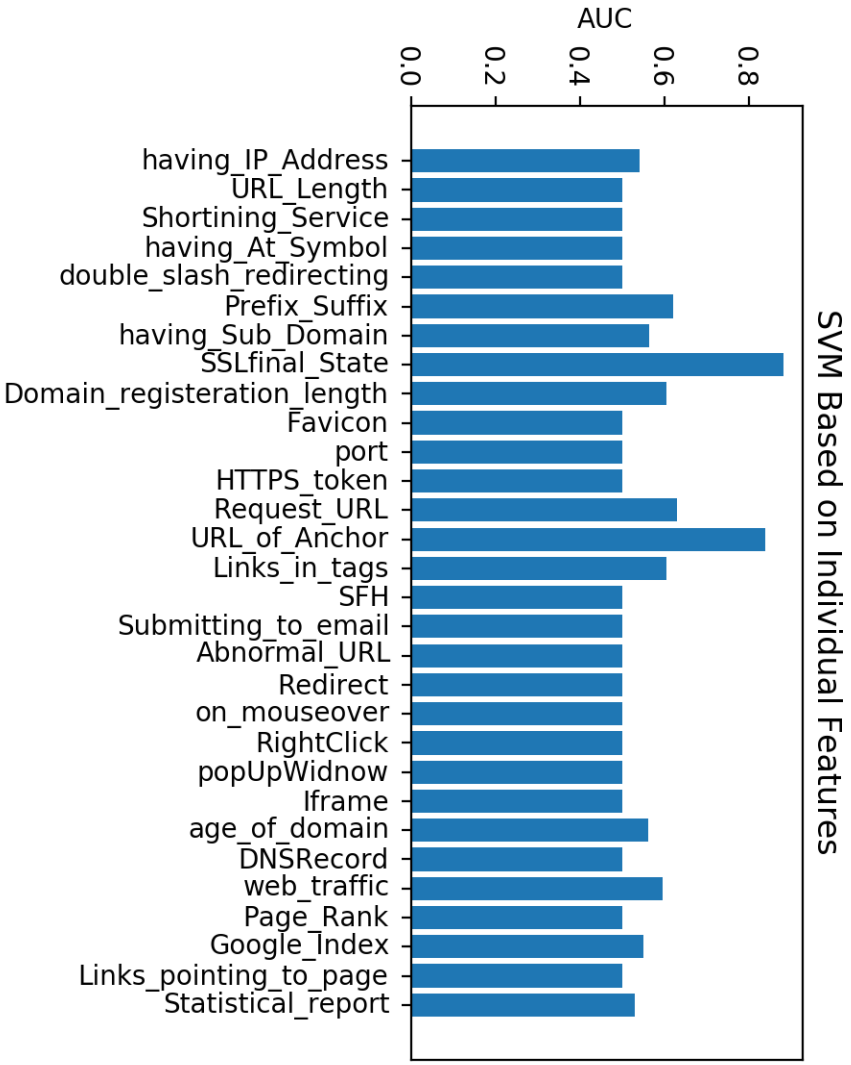
rbf



polynomial

II. SVM for each individual feature

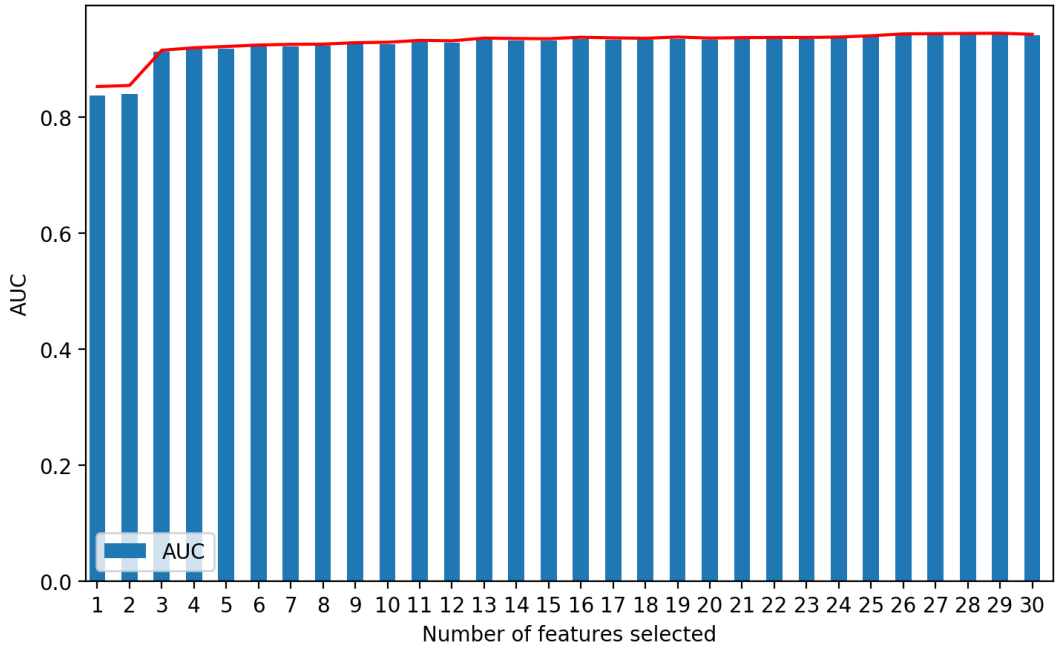
In the next experiment, we computed AUC scores for multiple SVM models of individual features to gain a better understanding of which feature weighs more in the machine learning model.



We found out that the feature SSLfinal_State has the highest AUC score (0.88), URL_of_Anchor second highest with an AUC score of 0.82.

III. RFE Results for SVM

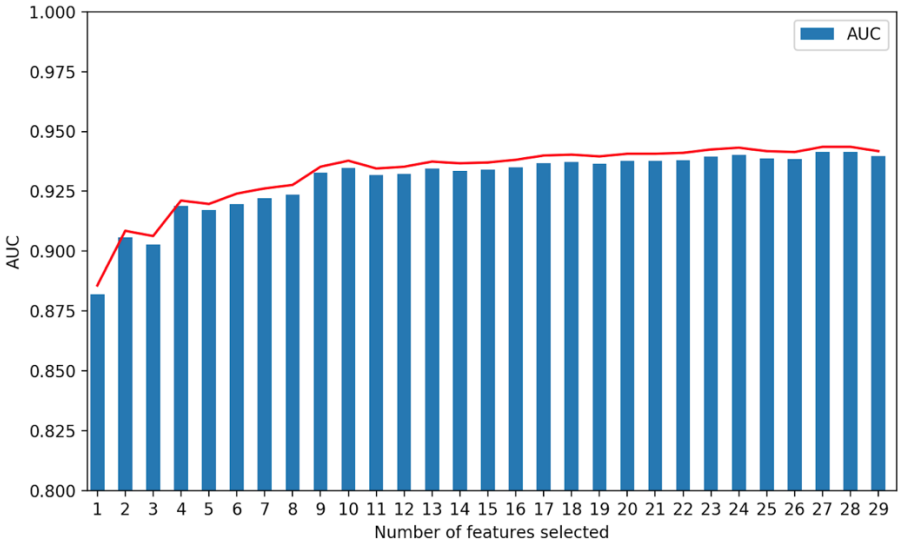
We are achieving an accuracy of 94% with all the 30 features. Can we do better if not worse with less number of features. We will use RFE to reduce the number of features and compute models with the reduced datasets.



We observed that the accuracy starts from 85% with just one feature selected, increases steadily till 5 features and then improves by just a little for every feature addition. This suggests that the model is best suited for the range of features from 5 – 10. However, the best AUC & Accuracy score is observed at all 30 features.

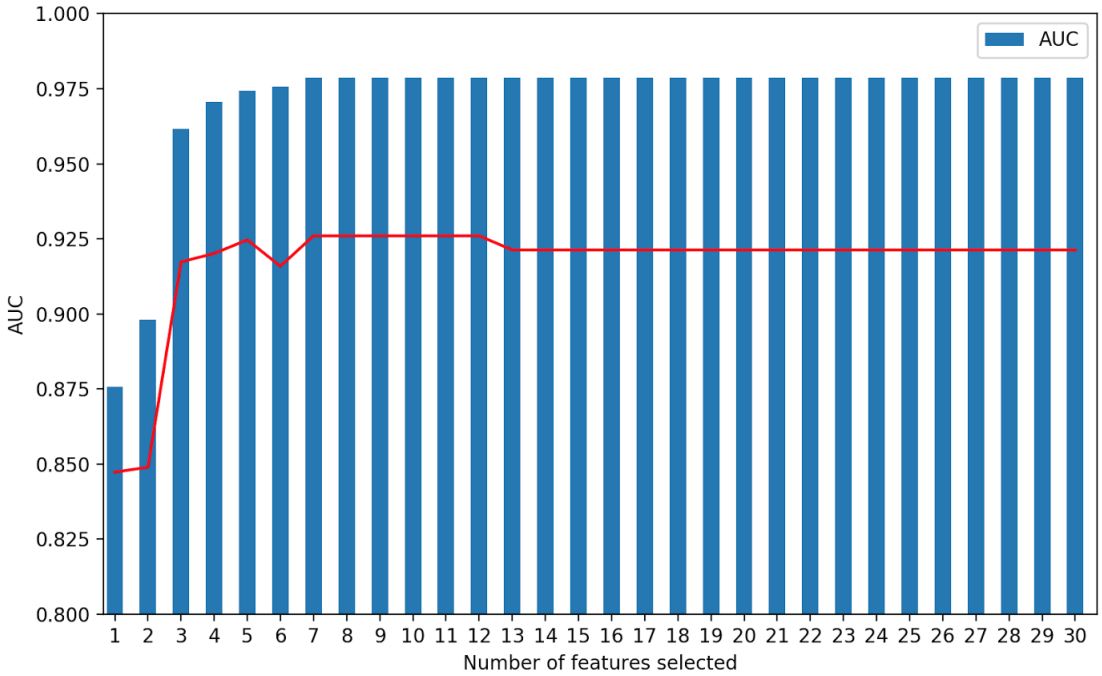
III. UFS Results for SVM

We observe similar results with features selected by the UFS algorithm



IV. AdaBoost

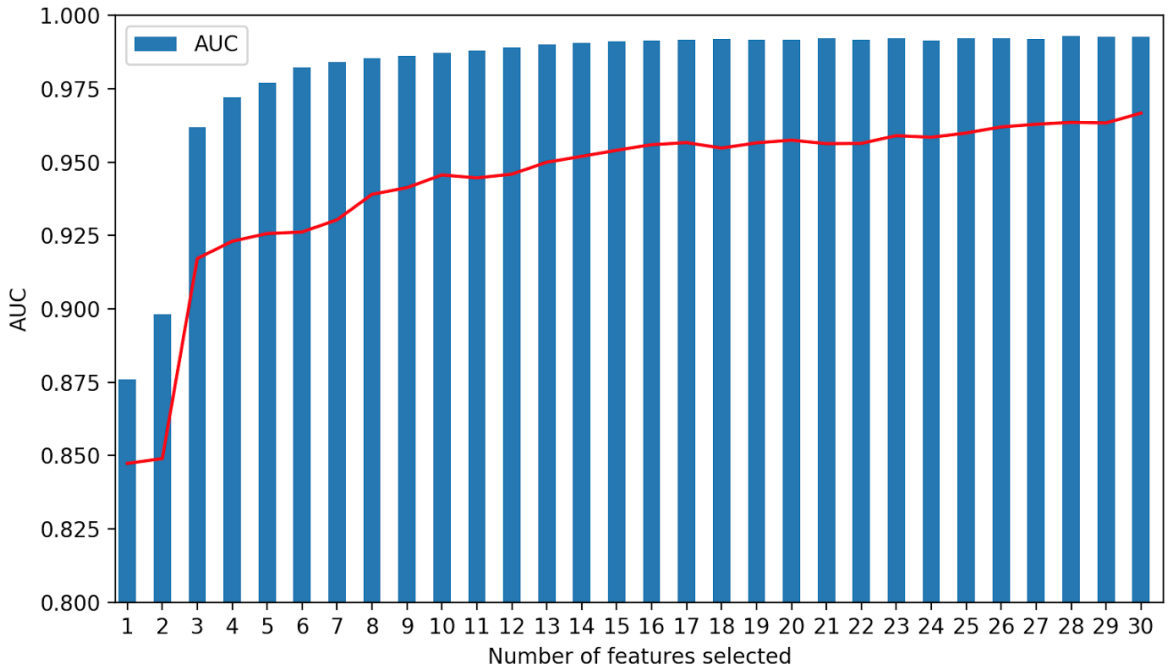
We wanted to compare the result of combining many weak classifiers versus a strong classifier. We chose Decision Stump which is a weak learner. A weak learner performs relatively poorly, its accuracy is just barely above chance. Weak learners are computationally simple. The core principle of AdaBoost is to fit a sequence of weak learners. An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. The predictions from all the weak learners are then combined through a weighted majority vote to produce the final prediction. At a given step, those training examples that were incorrectly predicted by the boosted model induced at the previous step have their weights increased, whereas the weights are decreased for those that were predicted correctly. As iterations proceed, examples that are difficult to predict receive ever-increasing influence. Each subsequent weak learner is thereby forced to concentrate on the examples that are missed by the previous ones in the sequence



From the above graph, we can see that and AUC of 0.98 can be achieved with 6 features. Thus a weak classifier like decision stump can give great results when a sequence of weak classifiers are combined together. The number of estimators that we used for the AdaBoost classifier was 8. The AdaBoost classifier provided much better results as compared to SVM.

IV. Voting Classifier

After combining a sequence of weak classifiers, we wanted to compare its results by combing a series of strong classifiers. Voting Classifier is used to combine conceptually different machine learning classifiers and use a majority vote or the average predicted probabilities (soft vote) to predict the class labels. Such a classifier can be useful for a set of equally well performing model in order to balance out their individual weaknesses. Soft voting returns the class label as argmax of the sum of predicted probabilities. When weights are provided, the predicted class probabilities for each classifier are collected, multiplied by the classifier weight, and averaged. The final class label is then derived from the class label with the highest average probability. We used the following strong classifiers and weights -
Gaussian Naive Bayes (w=1)
Decision Tree Classifier (w=1)
Random Forest Classifier(w=2)



From the above graph, we see can see that we get an AUC score of 0.98 with just 5 features and an AUC of 0.99 with 8 features. The combination of strong classifier worked slightly better than the series of weak classifier. The decision stump is a very simple algorithm with less computational complexity on the other hand its computationally more complex to train 3 strong classifiers.

Summary

We were able to predict whether a website is a phishing website or a legitimate website with a accuracy of 99%. The ensemble methods performed much better than SVM. The voting classifier performed the best in terms of AUC score. The computational complexity of three strong classifiers (Voting Classifier) is much higher than a simple classifier like decision stump.