Phishing Website Detection by Using Gradient Boosting Algorithm

Introduction:

Gradient boosting is a method standing out for its prediction speed and accuracy, particularly with large and complex datasets. From Kaggle competitions to machine learning solutions for business, this algorithm has produced the best results. We already know that errors play a major role in any machine learning algorithm. There are mainly two types of error, bias error and variance error. Gradient boost algorithm helps us minimize bias error of the model .Before getting into the details of this algorithm. This algorithm starts by building a decision stump and then assigning equal weights to all the data points. Then it increases the weights for all the points which are misclassified and lowers the weight for those that are easy to classify or are correctly classified. A new decision stump is made for these weighted data points. The idea behind this is to improve the predictions made by the first stump. I have talked more about this algorithm here. Read this before starting this algorithm to get a better understanding .The main difference between these two algorithms is that Gradient boosting has a fixed base estimator .

1. **Loss function**: To reduce errors in prediction, we need to optimize the loss function. Unlike in AdaBoost, the incorrect result is not given a higher weightage in gradient boosting. It tries to reduce the loss function by averaging the outputs from weak learners.

2. 2. **Weak learner**: In gradient boosting, we require weak learners to make predictions. To get real values as output, we use regression trees. To get the most suitable split point, we create trees in a greedy manner, due to this the model over fits the dataset.

3. 3. **Additive model**: In gradient boosting, we try to reduce the loss by adding decision trees. Also, we can minimize the error rate by cutting down the parameters. So, in this case, we design the model in such a way that the addition of a tree does not change the existing tree.

   Finally, we update the weights to minimize the error that is being calculated.

Gradient Boosting classification Algorithm:

The rise in the quantity of unsolicited bulk emails (UBEs) has turned out to be a brutal risk to the global economy and security due to the invasion of scientific progressions and the raised ease in communication, mainly through emails. Now a days, purchasing domain names similar to google.com, youtube.com, amazon.com is very easy. Some attackers develop a domain name or follow the homograph technique by alternating the same characters in the domain, e.g.,"google.com". Representing domain in web address as tiny URL (short URL name), misspeland sub-domain in the domain name of URL link are the type of spoofing attack. Excitement offers are offered to register new pages. However, it redirects to certain new encrypted pages. The malware is automatically downloaded into users' systems in encrypted executable files or folders. Then the computer starts behaving abnormally. This type of attack is called a "ransom" attack. So basically, phishing types are based on social engineering and malware, with malicious code content technically called subterfuge. Always scammers search for the current issues, welfare benefit, surrounding situations, trade association, unemployment, financial relief packages, government agency, free software packages, growing technology, and health issues. In such a favorable environment, they target the users very easily through social networks. They act very differently according to the various current situations. They keep trying other options if the present action does not work. In the month of June 2020, the CERT-In (Indian Computer Emergency Response Team) warned the government agencies, departments, trade association's users, and laymen on the possible major phishing attack where millions of Indians could be targeted by fake emails and text messages asking for the free Covid-19 test.Furthermore, the online transactions payment apps like PhonePe, Google Pay, and Paytm are the other sources for email phishing during this Covid-19 pandemic. The best way to avoid phishing is to primarily make the network system very strong by mitigating the risk through the spam filter, firewall, anti-virus, or applying the blacklist method. On the other hand, it is challenging to assure complete security about the phishing email from the internet stigma. It is becoming challenging to predict the hacker's actual target and forged site, bad URL. Some of the topical research has evolved in recent years based on machine learning techniques to study the behavior of phishing URLs.

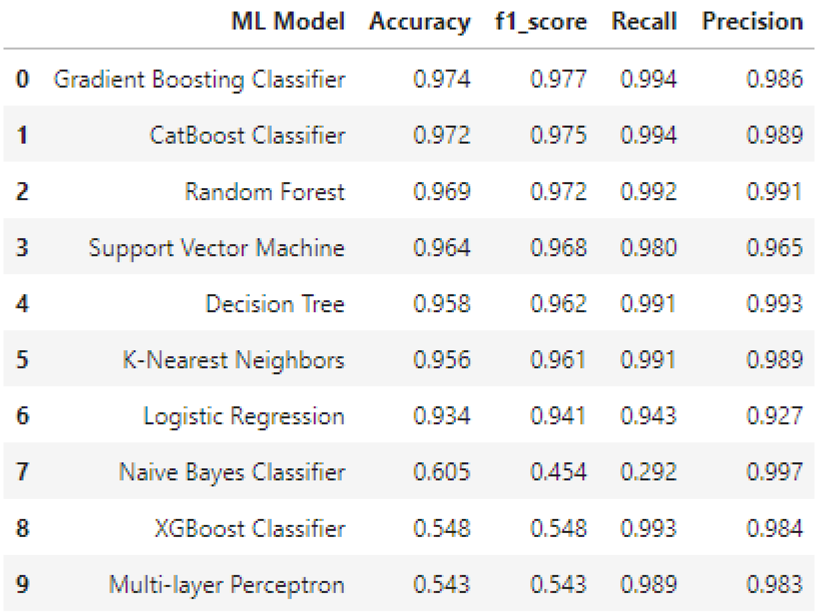
Accuracy:

The final take away form this project is to explore various machine learning models, perform Exploratory Data Analysis on phishing dataset and understanding their features.

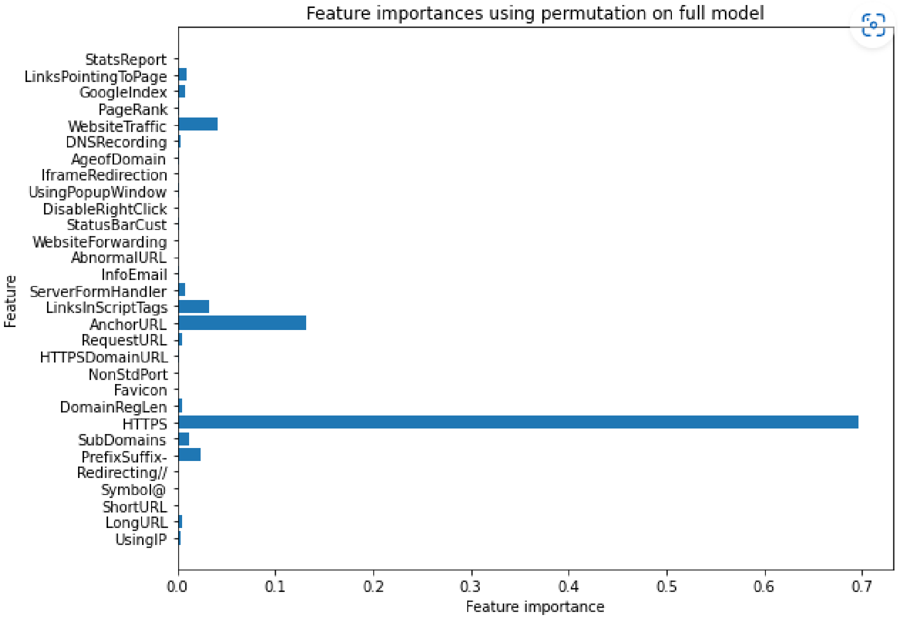
2. Creating this notebook helped me to learn about Cloud , about the features affecting the models to detect whether URL is safe or not, also I came to know how to tuned model and how they affect the model performance.

3. The final conclusion on the Phishing dataset is that the some feature like "HTTTPS", "Anchor URL", "Website Traffic" have more importance to classify URL is phishing URL or not.

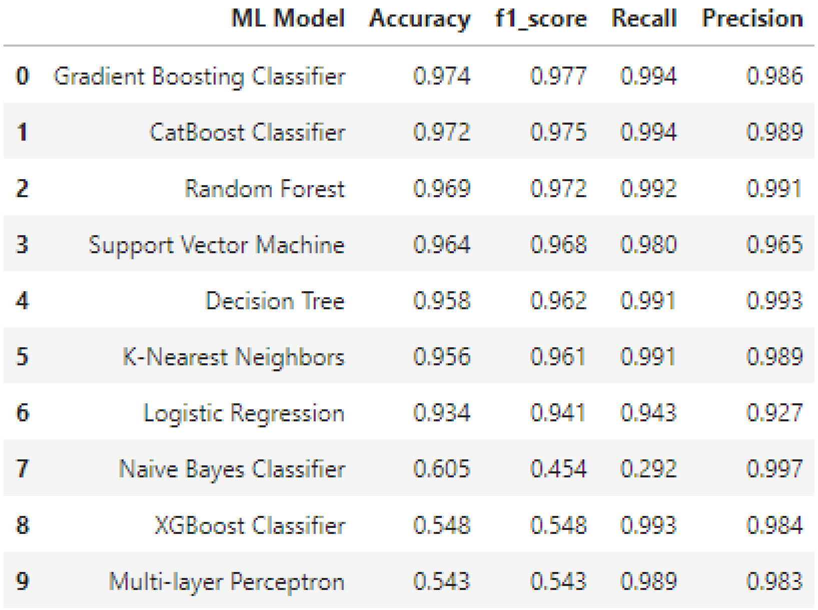
4. Gradient Boosting Classifier correctly classify URL up to **97.4%** respective classes and hence reduces the chance of malicious attachments.



	ML Model	Accuracy	f1_score	Recall	Precision
0	Gradient Boosting Classifier	0.974	0.977	0.994	0.986
1	CatBoost Classifier	0.972	0.975	0.994	0.989
2	Random Forest	0.969	0.972	0.992	0.991
3	Support Vector Machine	0.964	0.968	0.980	0.965
4	Decision Tree	0.958	0.962	0.991	0.993
5	K-Nearest Neighbors	0.956	0.961	0.991	0.989
6	Logistic Regression	0.934	0.941	0.943	0.927
7	Naive Bayes Classifier	0.605	0.454	0.292	0.997
8	XGBoost Classifier	0.548	0.548	0.993	0.984
9	Multi-layer Perceptron	0.543	0.543	0.989	0.983



Prediction level:



	ML Model	Accuracy	f1_score	Recall	Precision
0	Gradient Boosting Classifier	0.974	0.977	0.994	0.986
1	CatBoost Classifier	0.972	0.975	0.994	0.989
2	Random Forest	0.969	0.972	0.992	0.991
3	Support Vector Machine	0.964	0.968	0.980	0.965
4	Decision Tree	0.958	0.962	0.991	0.993
5	K-Nearest Neighbors	0.956	0.961	0.991	0.989
6	Logistic Regression	0.934	0.941	0.943	0.927
7	Naive Bayes Classifier	0.605	0.454	0.292	0.997
8	XGBoost Classifier	0.548	0.548	0.993	0.984
9	Multi-layer Perceptron	0.543	0.543	0.989	0.983

If we collect observations and calculate a 97.4% prediction interval based on that sample, there is a 95% probability that a future observation will be contained within the prediction interval. Conversely, there is also a 3% probability that the next observation will not be contained within the interval.

Conclusion:

Nowadays, phishing websites are increasing rapidly and causing more damage to the users and organizations. It is becoming a biggest threat to people's daily life and the networking environment. In these attacks, the intruder puts on an act as if it is trusted organization with an intention to purloin liable and essential information. Phishing website is a mock website that looks similar in appearance but different in destination. The unsuspected users post their data thinking that these websites come from trusted financial institutions. Hence, there is a need for efficient mechanism for the detection of phishing website. In our project, we developed a model that can be mainly used in determining the website's as either phishing or legitimate by using the features extraction techniques from the URL. These features are compared with the features present in the features extraction dataset and validated accordingly. Here, in our project we applied the algorithms like Gradient Boost, Cat Boost and Random Forest on the model that has been developed. During testing, it has been observed that the system has performed well and as expected. This paper aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 97.4% detection accuracy using Gradient boost classifier with lowest false positive rate.