DOI: 10.37943/12OYRS4391

Dinara Kaibassova
PhD, Acting Associate Professor of the Department of Information and Computing Systems
dindgin@mail.ru, orcid.org/0000-0002-8410-7758
Abylkas Saginov Karagandy Technical University, Kazakhstan

Margulan Nurtay
Computer Science master student
solano.lifan2@bk.ru, orcid.org/0000-0002-0786-6195
Abylkas Saginov Karagandy Technical University, Kazakhstan

Ardak Tau
Senior lecturer at the Information and Computing Systems Department
ardak.tau@mail.ru, orcid.org/0000-0003-4883-6328
Abylkas Saginov Karagandy Technical University, Kazakhstan

Mira Kissina
Senior lecturer at the Information and Computing Systems Department
motya.2002@mail.ru, orcid.org/0000-0003-2232-1203
Abylkas Saginov Karagandy Technical University, Kazakhstan

SOLVING THE PROBLEM OF DETECTING PHISHING WEBSITES USING ENSEMBLE LEARNING MODELS

Abstract. Due to the popularity of the easiest way to obtain personal information among attackers, phishing detection is becoming a popular area for research aimed at countering the implementation of such attacks. Malicious website detection is essential to prevent the spread of malware and protect end users from victims. Unfortunately, malicious URL detection still needs to be better understood due to a lack of features and inaccurate classification. Possible sources were examined in order to investigate the subject. Based on the collected information from previous studies, this study is devoted to solving the problem of detecting phishing websites using Ensemble Learning. The aim of the work is to choose the most optimal algorithm for classifying phishing websites using gradient boosting algorithms. AdaBoost, CatBoost, and Gradient Boosting Classifier were chosen as Ensemble Learning algorithms and were used to improve the efficiency of classifiers. Practical studies of the parameters of each algorithm for finding the optimal classification model are given. Research and experiments were carried out on a dataset containing information extracted from the contents of a URL: main URL, domain, directory, and file. A thorough Exploratory Data Analysis (EDA) was carried out, as a result of which the main dependencies and patterns of determining phishing resources were identified using correlation analysis. ROC AUC Score was chosen as an evaluation metric for the algorithms. The best result for predicting phishing websites was demonstrated by the AdaBoost Classifier algorithm, with an average ROC AUC score of 99%. The results of the experiments were illustrated in the form of graphs and tables.

Keywords: phishing detection, Ensemble Learning, imbalanced classification, gradient boosting.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 12, DECEMBER 2022

Introduction

Nowadays, digitalization has reached its highest level of development in many sectors and spheres of life. Almost every company, even small businesses, has its website, where they introduce its various online services. All data are stored in electronic form, starting with the constituent documents of the enterprise (charter, contract, etc.) and ending with the personal data of employees and electronic dossiers of customers. In this regard, ensuring the security of data stored in databases of web resources becomes more important every year.

According to the Google Safe Browsing Report [1], there were three million phishing websites in January 2022, up almost 3,100% from September 2010. Phishing is an attack primarily aimed at inexperienced users. In phishing, attackers lure end users into pre-configured links where they enter their personal information, such as banking and credit card information, and passwords in text forms. In this attack, attackers masquerade as trusted organizations, such as service providers, employees of an organization, or an organization's technical support team, so that end users never doubt their legitimacy. This is mainly done through emails asking for a system update, a message that the account has been blocked, a request for a prize, and so on. Thus, cybercriminals have been and continue to distribute false information and advertisements for years in order to attract users to visit malicious websites. Once the victim visits the malicious website, the attackers use various strategies to infect users, malicious payloads, or trick the victims into interacting with the attackers for financial fraud or other types of attacks. Moreover, attackers can use vulnerable websites to perform malicious activities. For example, a malware developer may inject cross-site scripting into a vulnerable website to steal sensitive information from a victim visitor or perform a phishing attack [2].

One of the main reasons why users fall prey to phishing attacks is a lack of awareness of security tokens (attackers can also spoof tokens), a lack of attention to the site's URL, or an inattention to the anomalous behavior of the site's toolbar [3]. The authors of [4] proposed to identify the difference between real and phishing websites by analyzing the URL of the website where the random forest classifier was used. In [5], a real-time phishing protection system was proposed that uses seven different classification algorithms and functions based on natural language processing (NLP). Implemented classification algorithms, such as the Random Forest algorithm with NLP-based features only, provide the best performance for detecting phishing URLs. The authors of [6] consider the effectiveness and suitability of using methods to detect phishing messages even before the end user reads the letter. Also, the vector representation of URLs was studied in [7]. In general, machine learning techniques have been used to classify web pages by systematically analyzing a set of features that reflect the characteristics of a malicious web page [8]. The main goal of phishing is to get end users to share their sensitive information. Due to the popularity of this method among attackers, phishing detection is becoming a popular area for research aimed at countering the implementation of such attacks. The main goal of phishing is to get end users to share their sensitive information. Due to the popularity of this method among attackers, phishing detection is becoming a popular area for research aimed at countering the implementation of such attacks.

Related works

There are various approaches to solving information security problems. Machine learning methods are used to classify web pages by systematically analyzing a set of features that reflect the characteristics of a malicious web page. The authors of [8] compared the performance of several classifiers, such as K-Nearest Neighbor (k-NN), Support Vector Machines (SVM), the C4.5 Tree, Classification and Regression Trees (CART), Logit boost Alternating Decision Tree (LADTree), and Naive Bayes Tree (NBTree).

DOI: 10.37943/12OYRS4391

Research [9] aims to improve the accuracy of malicious URL detection by designing and developing a malicious URL detection model based on cyber threat analytics using two-stage ensemble learning. Here, a two-stage ensemble learning model combines a random forest (RF) algorithm for pre-classification with a multi-level perceptron (MLP) for the final decision.

This study [10] implemented the analysis of the content of the relevant website using the TF-IDF matrix to develop an effective method for detecting phishing on websites based on URLs and achieved 90.68% efficiency when executing phisher fighter to implement the proposed method.

The research [11] proposed new ways to detect phishing websites using an ensemble learning approach. In particular, the authors applied Naïve Bayes Tree and Best First Tree algorithms to develop such models as Bootstrap Aggregation, Adaptive Boost Ensemble Learning, and Multi-boost Adaptive Boost Ensemble Learning. The performance results of developed models are illustrated among three different datasets, including various measures such as accuracy, F-Measure, AUC, and FPR. Therefore, tree ensemble approaches are viable methods that can be used to detect dynamic phishing websites.

Research methodology: The problem of detecting malicious websites has existed since the beginning of 2004. Authors propose various methods, but Machine Learning based detection method performs better than the methods [12,13]. The results of [8] show that Support Vector Machine (SVM) is the most accurate classification technique in MultiBoost and AdaBoost, while the K-Nearest Neighbor (kNN) technique is used in bagging and random subspace. In [14] authors analyze the impact of COVID-19 on various aspects related to cybersecurity and describe the chronology of COVID-19-themed cyberattacks launched around the world to determine the attacker's modus operandi and the consequences of attacks. The authors of this paper propose an intelligent system based on fuzzy logic and data mining to detect malicious URLs and phishing attacks on the topic of COVID-19. Many solutions have been proposed to locate these websites accurately. These decisions can be divided into three categories according to the source of the study:

1. Content-oriented approach. This method is based on the analysis of the text content of the page using Text Mining techniques. Using only the pure TF-IDF algorithm, 97% of phishing websites can be detected with 6% false positives.

2. URL-based approach. It uses page rank by analyzing the content of the URL, including domain elements, hosting, etc. This method can detect up to 97% of phishing sites.

3. A Machine Learning approach that is based on a statistical analysis of URL elements without checking in various web resource databases (various domain name spaces, SSL certificates). This approach can get up to 92% true positives and 0.4% false positives.

This section details the proposed approach. Combining the "opinions" of various machine learning algorithms on a given problem is believed to give better results than any individual approach. The structure of the proposed approach consists of several stages. First, the features extracted from the URL are grouped by several characteristics: URL body content, domain, directory, and file. After that, a wide feature vector is formed with numerical and categorical features according to the dataset. The URL is then checked for phishing using a gradient boosting algorithm. Figure 1 shows a model of the proposed approach to detecting phishing websites. The model includes dataset extraction, analysis and preprocessing blocks, and a results block for each algorithm.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 12, DECEMBER 2022

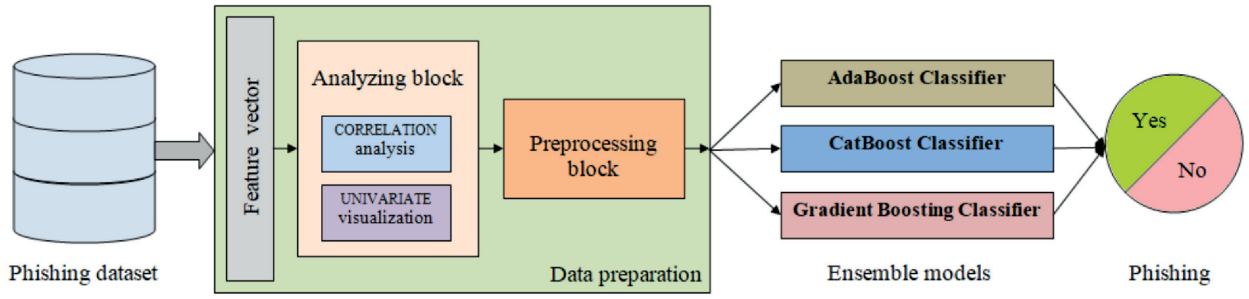

Figure 1. Model of the proposed approach

This work combines the use of a URL-based approach with machine learning. The URL, with its various features, is considered an object of study.

When considering machine learning algorithms for phishing website prediction, the focus was on ensemble models such as AdaBoost, CatBoost, and GradientBoosting Classifier. This choice is due to the good performance and reliability that are achieved using these models.

Ensemble methods are powerful tools that are used to build machine learning models. The ensemble is a machine learning technique where multiple models are trained to solve the same problem and combined to get better results. The basic premise is that the result of multiple models will be more accurate than the result of only one model.

AdaBoost (Adaptive Boosting) is an ensemble learning method that was originally created to improve the performance of binary classifiers. This algorithm was proposed by Yoav Freund and Robert Schapire in 1997 [15]. This algorithm enhances the relative weight of weak classifiers by merging them into a strong one and predicting a value again. Additionally, each next set of classifiers, called "ensemble", is based on objects that were incorrectly classified by previous sets.

The final classification equation looks like this:

$$F(x) = sign\left(\sum_{m=1}^{M} \theta_m f_m(x)\right)$$

Where f_m is the m-th weak classifier and m is its corresponding weight.

Another variety of gradient boosting algorithms is CatBoost, introduced by Yandex in 2017. The main advantage of CatBoost is that it works equally well with both numerical and categorical features [16]. The general idea of the CatBoost algorithm is that, firstly, ordered target encoding is performed, and secondly, when calculating predictions, examples with indices less than the one on which the prediction is to be obtained are used. At each boosting step, one decision table (tree) is trained. The depth of the tree is a hyperparameter, which means it is set in advance by the developer. If the depth is N, then the tree is a sequence of N features and separation thresholds and has 2N leaves.

Gradient Boosting Classifier is a gradient boosting algorithm that uses the principles of additivity and consistency when training models [17]. Its main difference from Adaptive Boosting is how it identifies weak models while training. While Adaptive Boosting uses weights assigned by the algorithm for this purpose, Gradient Boosting achieves this by calculating the loss function.

Dataset description: The dataset was taken from the open-source website github.com [18]. It contains more than 88 thousand websites marked as legal (legitimate) or phishing and allows developers to build models for solving the classification problem. The features in the dataset were the characteristics of the content of the HTTP request: URL address, domain,

DOI: 10.37943/12OYRS4391

request parameters, directories, and files. Initially, a detailed exploratory analysis of the data was carried out. The distribution of data by class is shown in Figure 2.

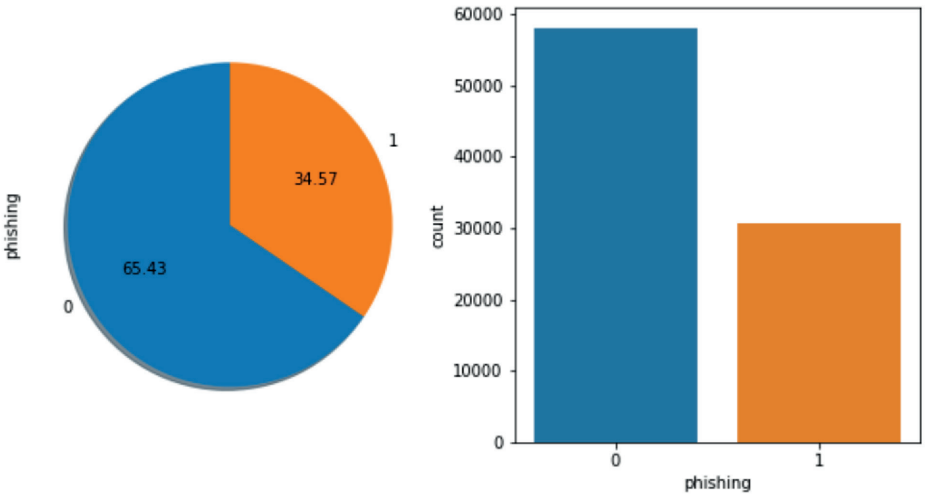

Figure 2. Classes distribution

In the chart in Figure 2, it can be clearly seen that the data is unbalanced, as the number of legitimate significantly exceeds the number of rows with phishing websites. It was decided to use the SMOTE method, which helps increase the number of instances for the minority class, to solve the problem of unbalanced classification.

A general description of the statistics for the features of the data set is presented in the table below.

Table 1. General statistics for the dataset

Parameter	qty_dot_ url	qty_ hyphen_ url	qty_ underline_ url	qty_ slash_url	...	time_ domain_ activation	time_ domain_ expiration	qty_ nameservers	qty_mx_ servers
Count	88647	88647	88647	88647	...	88647	88647	88647	88647
Mean	0.47601	0.27066	0.14905	0.86710	...	0.083928	0.20594	0.06832	0.14580
Std	0.89914	0.93488	0.95516	0.68378	...	1.010624	0.95576	0.96564	0.98150
Min	0.00000	0.00000	0.00000	0.00000	...	0.000000	0.00000	0.00000	0.00000
25%	0.00000	0.00000	0.00000	1.00000	...	0.000000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	1.00000	...	0.000000	0.00000	0.00000	0.00000
75%	1.00000	0.00000	0.00000	1.00000	...	0.000000	0.00000	0.00000	0.00000
max	18.0000	31.0000	32.0000	23.0000	...	80.000000	24.0000	154.000	28.0000

It should be noted that in this case, the most informative statistical indicator is the standard deviation, since it shows how stably the data is distributed in each of the features. The value of the standard deviation for all features does not exceed 1, which indicates the concentration of data around the mean value. Moreover, the first and second quartiles contain mostly zero values since more than 65% of the rows belong to legitimate website features.

In addition to the standard deviation and interquartile ranges shown in Table 1. Violin plot charts were used to estimate the scatter of the data. This method can be clearly seen in the example of the qty_dots_url (number of dots in the URL) feature in Figure 3.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 12, DECEMBER 2022

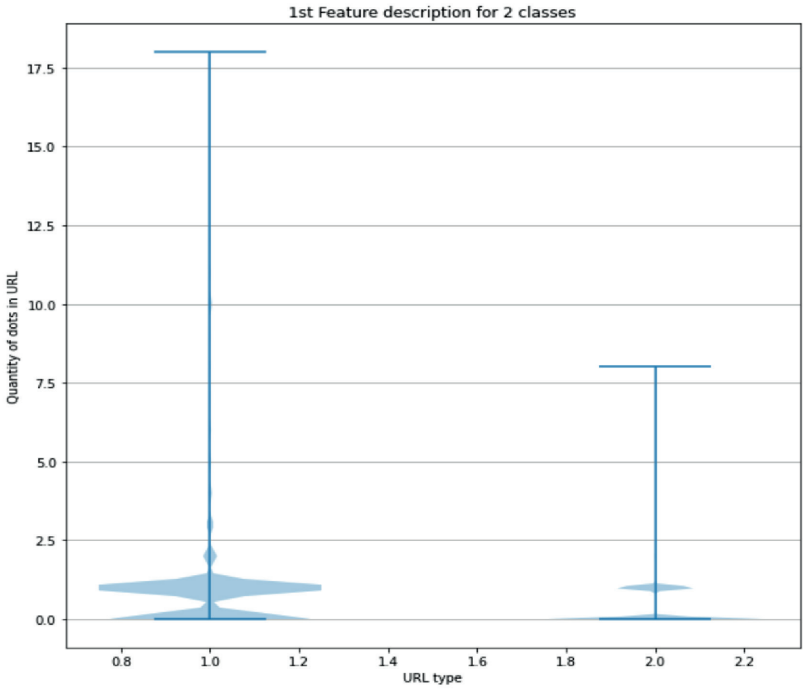

Figure 3. Violin plot for quantity of dots in URL

The violin plot clearly demonstrates that the maximum value of the number of dots in a URL is approximately 18 for phishing sites, while for non-phishing sites, this value is in the range of 8. This feature also shows that the majority of phishing URLs have a number of dots ranging from 0 to 3. In contrast, legitimate sites average about 1 dot per URL.

Exactly the same detailed statistical analysis was carried out for the remaining quantitative features of the dataset.

Further, the correlation matrix was built in the context of each of the characteristics of the URL (Figure 4, a, b, c, d).

Consider the correlation of features of the characteristics of the directories contained in the URL in detail. Signs of the number of hashtags and the number of equal signs in the address were highly correlated. Therefore, the values contained in the rows with these columns were checked, and as a result, the sign of the number of equal signs was removed, since it had a high standard deviation and a large number of zero values.

The corresponding preprocessing was carried out over the missing values, which in most cases with quantitative variables provided for the replacement of null values with the mean value (mean) for each feature column.

The dataset included categorical features, such as the presence of an SSL certificate, the indexability of the URL and Google domain, and the shortening of the URL.

For experiments, Google Colab with a Tesla T4 GPU was used.

DOI: 10.37943/12OYRS4391

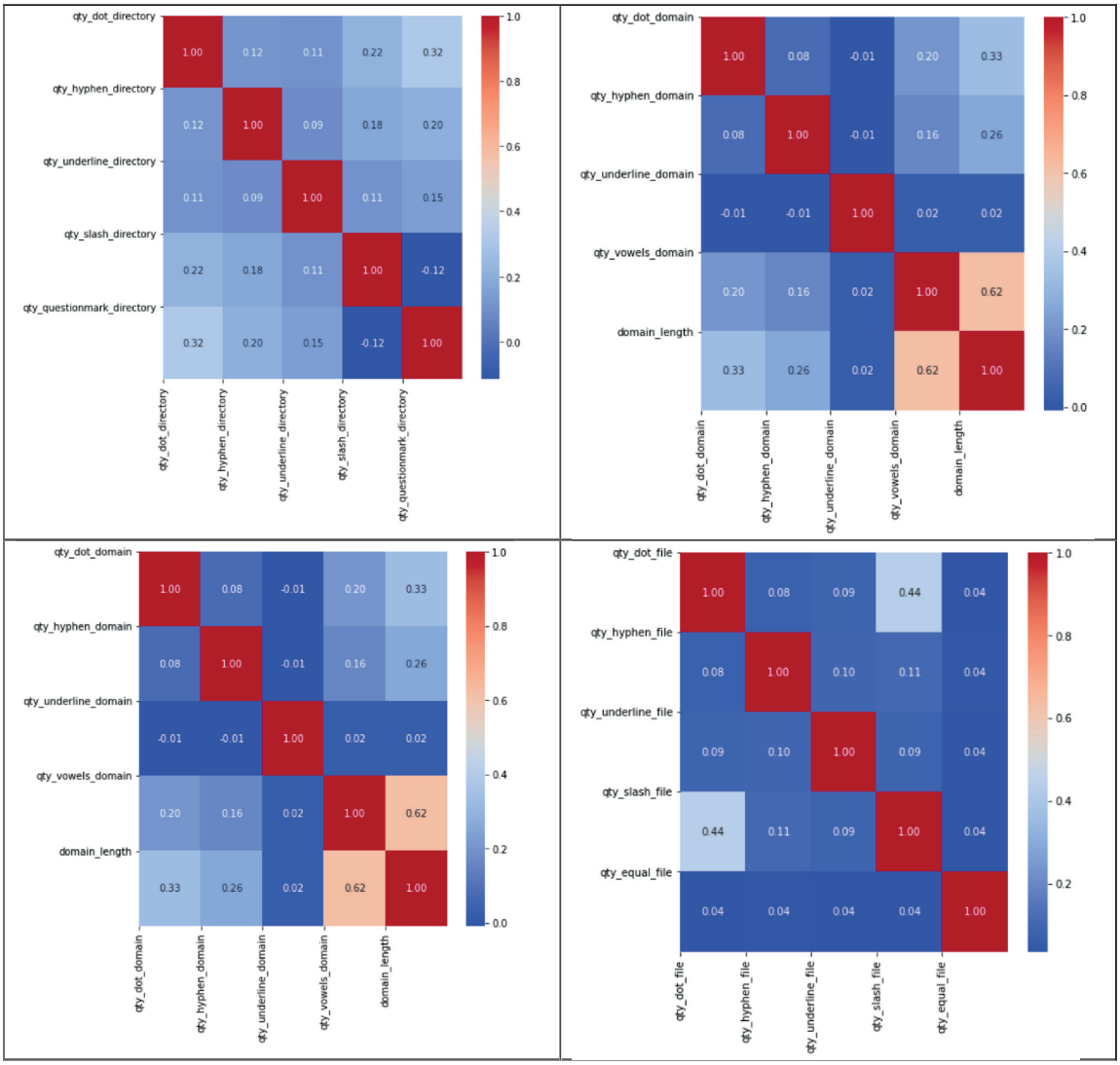

Figure 4. Correlation matrix of URL directory features

Results: For the preprocessed and analyzed dataset, it was decided to use the Stratified K-Fold cross-validation with 10 strata. When choosing the optimal training parameters, the GridSearchCV of the Python scikit-learn library was used. The parameters for each of the selected algorithms are revealed in the table:

Table 2. List of parameters of each algorithm for finding the optimal classification model

Algorithm	Parameters					
AdaBoost	Base estimator	Decision Tree Classifier				
	Max depth	2	4	6	8	10
	Min samples leaf	5			10	
	N estimators	10	50		250	1000
	Learning rate	0.01			0.1	
CatBoost	Depth	2	4	6	8	10
	Iterations	10	30	50	70	100
	Learning rate	0.01	0.02	0.03	0.04	
Gradient Boosting Classifier	Max depth	2	4	6	8	10
	Min samples leaf	5			10	
	N estimators	10	50		250	1000
	Learning rate	0.01	0.025	0.05	0.075	0.1

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 12, DECEMBER 2022

The results of GridSearchCV for finding the best parameters for the selected models are highlighted in the corresponding colors in the table above. The accuracies of the AdaBoost, CatBoost, and Gradient Boosting Classifier models were 99%, 98%, and 96%, respectively.

Each of the models was evaluated using the ROC AUC Score. The results of ROC AUC Score are presented in the figure below.

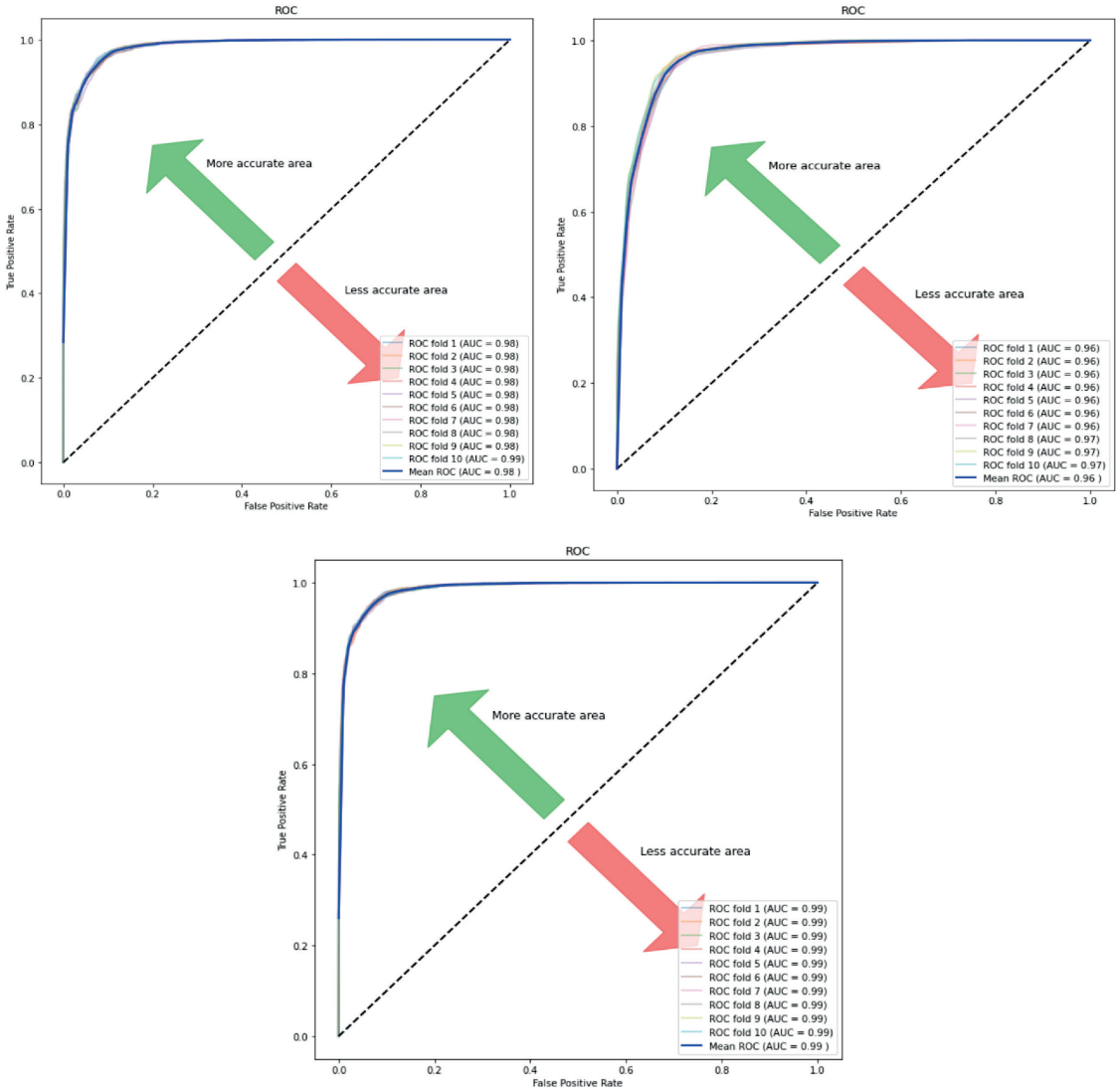

Figure 5. ROC AUC Score results

The results of the ROC AUC Score can be interpreted as follows:

- the average ROC AUC Score for Gradient Boosting Classifier turned out to be 0.98, while the range of difference between the obtained values was 0.001;
- for the model based on CatBoost, this figure was 0.96, which is less satisfactory compared to the previous model;
- the average ROC AUC Score for the model based on the AdaBoost was 0.99, which is the best indicator of the model's sensitivity and indicates the excellent predictive power of the model.

In general, it can be noted that all the selected algorithms illustrate approximately the same accuracy, however, in terms of performance and speed, AdaBoost Classifier displayed the best results.

Conclusion

This article regarded the application of ensemble learning models to the problem of detecting phishing websites. As part of the study, the goal was to select the most optimal algorithm for classifying phishing websites using gradient boosting algorithms. The authors propose an approach to using gradient boosting algorithms to solve a binary classification problem where numeric and textual data are present as analyzed data. The dataset was analyzed containing various features extracted from the content of a typical URL. The main regularities and dependencies of each attribute are revealed, and preprocessing and scaling of the data are carried out. In the course of the experiments performed, the model based on the AdaBoost Classifier based on the Decision Tree with an average ROC AUC Score of 99% showed the best accuracy. As a result of experimental and statistical analysis, it was found that the AdaBoost Classifier successfully detects phishing websites amongst the results of other algorithms. First of all, AdaBoost is not prone to overfitting and is most effective in binary classification tasks. Moreover, the main advantage of AdaBoost is its capability to generalize data because it is not always possible to build combinations that outperform fundamental algorithms. Also, it can identify objects for which the weights can take large values. AdaBoost can be enhanced to solve binary classification tasks using textual and numeric data simultaneously.

The proposed model can be used in the development of browser extensions or other advisory plug-in software to identify suspicious websites. While using this model in production, the database of phishing URLs will expand, and the accuracy characteristics of the model will constantly improve.

The next stage of the study is to analyze and select the optimal URL analysis model together with the content of the web page of this URL.

References

1. Google Transparency Report. (2022, March 2). Google Safe Browsing. Retrieved March 2, 2022, from https://transparencyreport.google.com/safe-browsing/overview?hl=en
2. Liu, M., Zhang, B., Chen, W., & Zhang, X. (2019). A survey of exploitation and detection methods of XSS vulnerabilities. *IEEE Access*, *7*, 182004–182016. https://doi.org/10.1109/ACCESS.2019.2960449
3. Rao, R.S., & Pais, A.R. (2019). Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Computing and Applications*, *31*(8), 3851–3873. https://doi.org/10.1007/s00521-017-3305-0
4. Laxmi Prasanna, K., Pradeepthi, K. V., & Saxena, A. (2022). Phishing URL Identification Using Machine Learning, Ensemble Learning and Deep Learning Techniques. In *Smart Intelligent Computing and Applications*, *Volume* 2 (pp. 573-582). Springer, Singapore. https://doi.org/10.1007/978-981-16-9705-0_56
5. Sahingoz, O., Buber, E., Demir, O., & Diri, B. (2019). Machine learning based phishing detection from URLs. *Expert Systems with Applications*, *117*, 345-357. https://doi.org/10.1016/j.eswa.2018.09.029
6. Abutaha, M., Ababneh, M., Mahmoud, K., & Baddar, S. (2021, May). URL Phishing Detection using Machine Learning Techniques based on URLs Lexical Analysis. In *2021 12th International Conference on Information and Communication Systems (ICICS)*, (pp. 147-152). IEEE. https://doi.org/10.1109/ICICS52457.2021.9464539
7. Huang, Z., Zhang, Y., Duan, R., & Wang, R. (2021, November). Research on Malicious URL Identification and Analysis for Network Security. In *2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC),* (pp. 418-422). IEEE.
8. Subasi, A., Balfaqih, M., Balfagih, Z., & Alfawwaz, K. (2021). A Comparative Evaluation of Ensemble Classifiers for Malicious Webpage Detection. *Procedia Computer Science, 194*, 272-279. https://doi.org/10.1016/j.procs.2021.10.082

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 12, DECEMBER 2022

9. Alsaedi, M., Ghaleb, F.A., Saeed, F., Ahmad, J., & Alasli, M. (2022). Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning. *Sensors*, *22*(9), 3373. https://doi.org/10.3390/s22093373
10. Vishva, E.S., & Aju, D. (2021). Phisher Fighter: Website Phishing Detection System Based on URL and Term Frequency-Inverse Document Frequency Values. *Journal of Cyber Security and Mobility*, *11*(1), 83–104. https://doi.org/10.13052/jcsm2245-1439.1114
11. Alsariera, Y., Balogun, A., Adeyemo, V., Tarawneh, O. & Mojeed, H. (2022). Intelligent tree-based ensemble approaches for phishing website detection. *Journal of Engineering Science and Technology*, *17*(1), 563–582.
12. Saleem, A., Vinodini, R., & Kavitha, A. (2021). Lexical features based malicious URL detection using machine learning techniques. *Materials Today: Proceedings, 47*, 163–166.
13. Zahra, S.R., Chishti, M., Baba, A. & Wu, F. (2021). Detecting Covid-19 chaos driven phishing/malicious URL attacks by a fuzzy logic and data mining based intelligence system. *Egyptian Informatics Journal*, *23*(2), 197–214. https://doi.org/10.1016/j.eij.2021.12.003
14. Tu, C., Liu, H., & Xu, B., (2017). AdaBoost typical Algorithm and its application research. In *MATEC Web of Conferences*, 139. 00222. EDP Sciences. https://doi.org/10.1051/matecconf/201713900222
15. Hancock, J., & Khoshgoftaar, T. (2020). CatBoost for big data: an interdisciplinary review. *Journal of Big Data*, *7*(1), 1-45. https://doi.org/10.1186/s40537-020-00369-8
16. Natekin, A. & Knoll, A., (2013). Gradient Boosting Machines, A Tutorial. *Frontiers in neurorobotics, 7*, 21. https://doi.org/10.3389/fnbot.2013.00021
17. Vrbančič, G., Fister, Jr. I., & Podgorelec, V. (2020). Datasets for Phishing Websites Detection. *Data in Brief, 33*, 106438. https://doi.org/10.1016/j.dib.2020.106438