22. Vasko V.T. Theoretical foundations of crop production / V.T. Vasko – St. Petersburg: Prifi-Inform, 2004. – 200 p.

23. Pigorev I. Ya. On the role of scientific concepts in agriculture / I. Ya. Pigorev, A. V. Naumkin, V. N. Naumkin [and others] // Bulletin of the Kursk State Agricultural Academy. – 2018. – No. 1. – P. 4-10.

UDC

ПРЕДСКАЗАНИЕ И ВЫЯВЛЕНИЕ КИБЕРПРЕ СТУПЛЕНИЙ С ПОМОЩЬЮ МАШИННОГО ОБУЧЕНИЯ

Лхагва Одончимег, *доцент кафедры информационных сетей и безопасности, доктор философии, MUST, Школа информационных и коммуникационных технологий, Улан-Батор, Монголия, e-mail: odno@must.edu.mn*

Реферат

Фишинговые веб-сайты являются распространенным методом социальной инженерии, который имитирует внешний вид надежных (URL)-страниц. Например, злоумышленники часто используют фишинговые методы, направляющие пользователей на мошеннические сайты или прокси-серверы, через подделку или отравление Системы доменных имен (DNS). В данном исследовании был составлен обзор текущего состояния киберпреступности в мире и в Монголии, а также проведено исследование для определения уровня образования, возраста и пола киберпреступников. Для выявления фишинговых атак были проведены оценки характеристик данных и сравнительный анализ особенностей фишинговых веб-сайтов. Также была установлена взаимосвязь между характеристиками с использованием методов машинного обучения, основанных на сходстве. Затем был обучен модельный алгоритм на основе метода логистической регрессии. Для обучения модели использовались 80 % данных, а 20 % были использованы для тестирования, что позволило подтвердить возможность выявления фишинговых веб-сайтов по показателям Precision, Recall и F1. Эксперимент показал, что наилучшей характеристикой стала 29-я по счету, которая позволила модели выявлять фишинговые сайты с точностью 93 %. Эта модель теперь способна предсказывать фишинговые и нефишинговые сайты с высокой точностью. Затем с помощью матрицы ошибок было проверено, действительно ли логистическая регрессия предсказала 93 % правильных результатов. Результаты показали, что из 2000 проверенных данных $ 950 + 930 = 1880 $ были предсказаны верно, что подтверждает точность модели в 93 %.

Ключевые слова: фишинговая атака, логистическая регрессия, матрица ошибок.

CyBER CRIME

Cyber crime and phishing attack

In most countries of the world, illegal access to the system, illegal interception of data, illegal intervention, distribution of malicious means by illegal use of computers, online fraud, and data breach are considered cybercrime. It is very commonplace throughout the world [6].

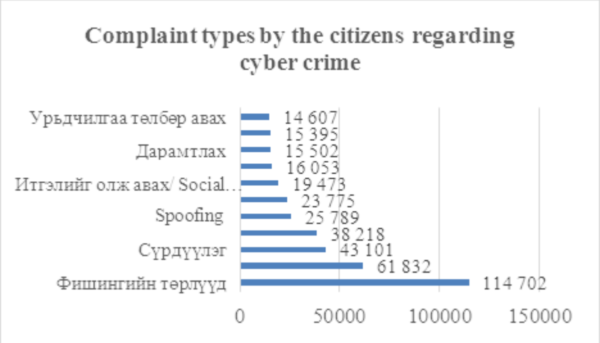


Figure 1 – Number of complaints and number of victims regarding cybercrime

Phishing is the attempt by individuals and by groups of people to obtain personal information of unsuspicious users by using social engineering techniques. In this, phishing emails are created that appear to be sent by the official organization or by familiar or known individuals. these emails offer to click the link of fraud website which appear to be legitimate. Later, they ask users to provide personal information such as login into account name and password which in turn leads to further risk. Additionally, these fraud websites contain harmful codes.

Phishing websites are a popular method of social engineering which imitates web pages with authentic url proving its reliability. For instance, attackers mostly use phishing methods and means by stealing and harming domain name system and aim towards sites and proxy servers which cheat users.

Social engineering

Social engineering attack

In social engineering attacks, interaction of people (social skills) can be used to access or destroy the information about the organization or its computer system. Maybe, the attacker seems to be humble and respectable, or he acquaints himself as the new employee, repairman, or researcher, and even he offers credentials to support this identity. By asking questions, he can gather enough information to gain access to the organization's network. If the attacker does not gather enough information from one source, he will connect with another source of the organization and will gain trust based on the information gathered from the first source.

Phishing attack

Phishing is a form of social engineering. A phishing attack uses email or malicious websites to obtain personal information by posing as a trusted organization. For instance, the attacker can send an email to a reputable credit card company or financial organization requesting an account statement. such an act mostly shows there is a problem. When users give the desired reply, attackers can use it to log into the account.

Vishing attack

Vishing attack is a social engineering method which supports voice communication. Confidential information can be disclosed by calling the user number through this method. It is possible to make advanced vishing attacks through voice over internet protocol resolutions (voip) and broadcasting services.

Smishing attack

Smishing is a form of social engineering which uses smile images when sending sms or text messages. Text message can contain links regarding web pages, email addresses, and phone numbers. Therefore, it is possible to call at the number, or to open email messages and browser windows automatically when click all the mentioned. Integration of emails, voice messages, and web browser functions which were created for the users increases the probability of becoming victims of malicious acts.

CURRENT SITUATION OF CYBERCRIME IN MONGOLIA

In 2020, Mongolia had 3,2 million active users of the internet.

That the citizens actively use systems such as Facebook, and Twitter whose servers are in foreign countries and which can't be directly regulated creates the risk for the citizens to become victims of crime and builds up the opportunity for those committing crime to conceal their illegal activities, to erase their tracks, and to change their images. [9].

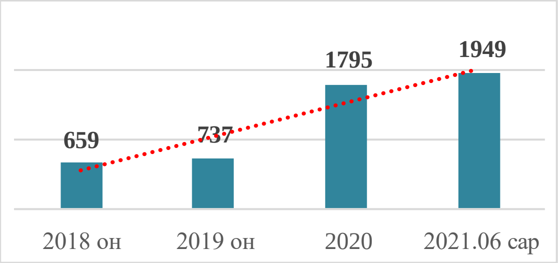


Figure 2 – Crimes committed online/ by dates/

Crimes committed online were compared by dates. They are as follows: 659 in 2018, 737 in 2019, 1795 in 2020, 1949 within 6 months of 2021. The average increase in 2018–2020 was 34,7 %.

A. Literature review in Mongolia:

There has been done many works regarding cybercrime by foreign volitional and professional LLM and JD researchers. Works concerned are quite rare in Mongolia. However, it is worth mentioning that recently there has been done research works and brochures by a few national researchers and respective organizations. For instance, the book "The feature of investigating the crime (cyber) regarding electronic information security" by T. Khaltar, PhD and expert in information security and cybercrime [2], research work "A detailed study on cybercrimes against electronic information security" [3]. In these works, the international definition of cybercrime and the current situation of cybercrime in Mongolia are compared. Also, in the book "Electronic law 2010" [4] by the lawyer L. Galbaatar, legal regulation in an electronic environment was compared with international best practices. In her research work "The study on the prediction of cybercrime through machine learning" L. Oyunchimeg, PhD, analyzed data attributes and did prediction by comparing machine learning correlation methods. In this sense, her work is the first among this type of research in Mongolia.

RESEARCH METHODOLOGY. MACHINE LEARNING

Machine learning is the technique that allows a computer to learn from its experience. In other words, machine learning is the automation of computer operation and the improvement of machine learning process based on its experience without any detailed programming and without any support by the humans.

Machine learning types:

They can be classified as below:

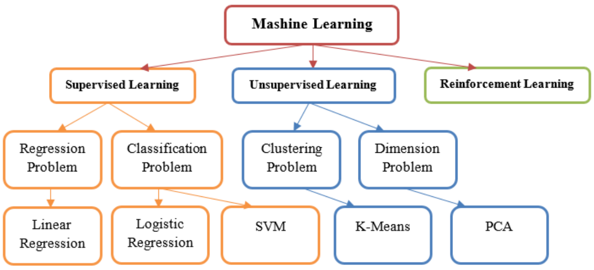


Figure 3 – Machine learning types

In this research work, we used methods of unsupervised machine learning.

B. Methods of supervised machine learning:

Pearson correlation

Correlation is the technique to analyze in detail the interdependence between numerical and irregular variables such as age and blood pressure.

Spearman correlation

It is nonparametric statistics, and its distribution doesn't depend on parameters. In most cases, nonparametric statistics evaluate data rather than real values, and it is related to Spearman correlation coefficient compared to Pearson correlation.

Kendall Tau correlation

It is very similar to the Spearman correlation coefficient. These 2 methods are nonparametric measures of correlation. Spearman and Kendall coefficients are calculated based on the ordered data rather than actual data. Like Pearson and Spearman correlations, Kendall Tau is always between –1 and +1. In this, –1 expresses a negative value between 2 variables and 1 indicates a positive correlation between 2 variables.

IV. DETECTION OF PHISHING ATTACK

A. Literature review regarding phishing detection

Since 2007, there have been published much research works which compared machine learning technique for predictive detection of phishing. In these works, several machine learning methods such as Logistic regression (LR), Classification and Regression trees (CART), Bayesian Additive Regression Trees (BART), Support Vector Machines (SVM), Random Forests (RF), Neural Networks (Net) were studied and compared their use regarding whether prediction was true. Since 2018, there has been done much research works using artificial intelligence techniques and it has become a very interesting research field.

B. Attributes to be phishing websites:

This research is concerned with defining the attributes of phishing websites and 10000 data gathered was classified. In this, there are 4 groups such as *Address Bar based Features, Abnormal Based Features, HTML and JavaScript based Features, Domain based Features. Every group contains websites with 10–16 attributes.*

V. EXPERIMENTATION PART

On Colab online platform, data was provided for machine to learn.

Data was first placed in Google drive => My drive folder, then imported the following to Colab.

Here, data will be read, some attributes of phishing websites will be evaluated, and specific attributes of phishing websites will be brought out. Next, csv placed in Google drive will be imported.

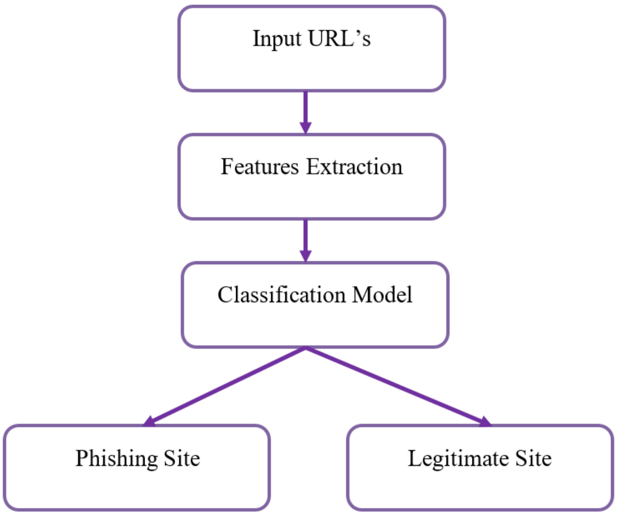


Figure 4 – Job sequence inserted into the machine

After the data was provided for machine to learn, the phishing site was labeled =1, and legitimate site was labeled =0.

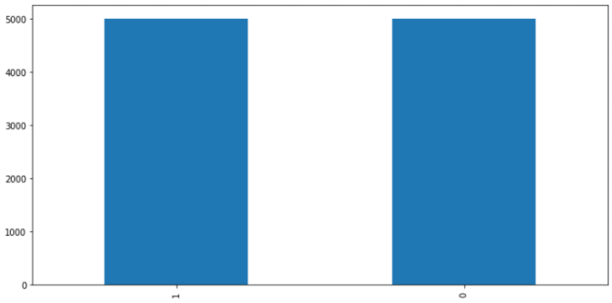


Figure 5. Comparison of phishing and non-phishing data

To identify whether it is phishing site from the attributes of dataset using 5 methods given below and to evaluate the prediction made through comparison:

- ➢ Spearman correlation
- ➢ Pearson correlation
- ➢ Cosine
- ➢ Kendall
- ➢ Mutial info

Then, every prediction made will be learned by machine using Logistic regression.

Evaluation by Spearman correlation:

In this, 4 correlation methods are used for comparison. This shows what functions have linear correlation when prediction was made by using only Spearman correlation.

If looking at the first 10 columns, it can be concluded that none of the features are strongly associated with labels. Although NumDash has significant negative effect on the labels which suggests that higher the number of dashes, the more likely it is a phishing site.

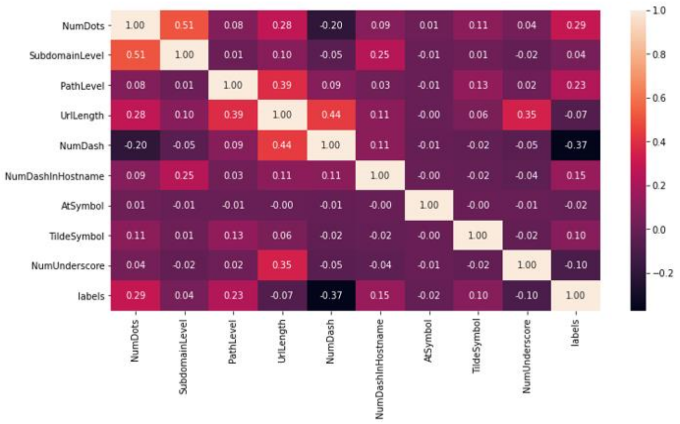


Figure 6 – Correlation heatmap of the first 0–10 columns

Regarding the next 20–30 columns, they still don't have any feature of strong correlation.

Columns 30–40

If looking at the heatmap below, there are several attributes that are linearly correlated to the dep variables.

- It shows that the higher the values of Insecure Forms, the more the probability for the site to be a phishing site.
- PctNull Self Redirect Hyperlinks shows positive correlation like Insecure Forms.
- Fequent Domain Name Mismatch shows there is mean linear correlation in positive direction.
- Submit Info to Email shows high probability for the sites which ask users to send their personal information by email to be phishing sites.

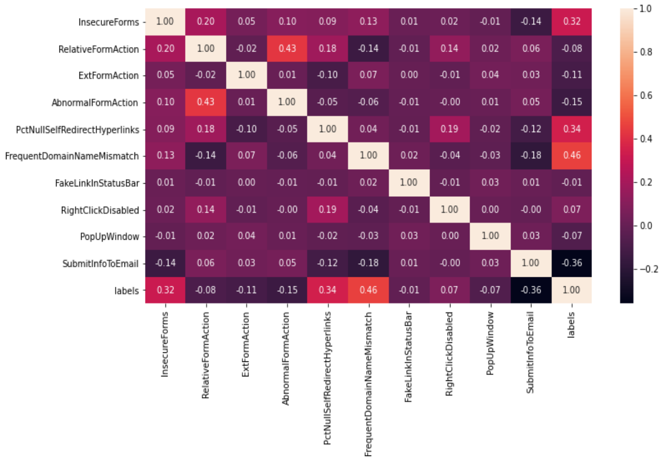


Figure 7 – Correlation heatmap of columns up to 30–40

Columns 40–50

In this group, the only column which can have some correlation with labels is PctExtNullSelfRedirectHyperlinksRT. It has a negative effect on labels and when the percentage of redirecting links is 0, the probability for the sites to be phishing increases.

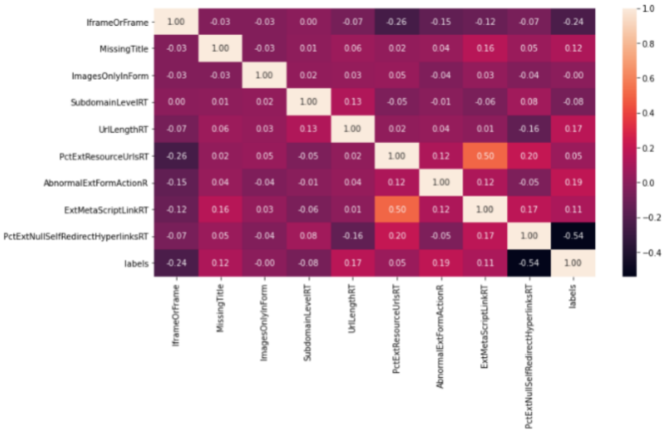
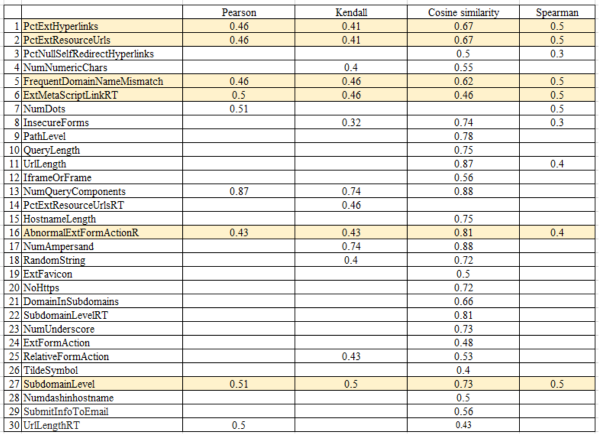


Figure 8 – correlation heatmap of columns 40–50

The table below shows detection made through comparison of similar correlation methods.

TABLE 1 – COMPARISON RESULT OF CORRELATION METHODS



		Pearson	Kendall	Cosine similarity	Spearman
1	PctExtHyperlinks	0.46	0.41	0.67	0.5
2	PctExtResourceUrls	0.46	0.41	0.67	0.5
3	PctNullSelfRedirectHyperlinks			0.5	0.3
4	NumNumericChars		0.4	0.55	
5	FrequentDomainNameMismatch	0.46	0.46	0.62	0.5
6	ExtMetaScriptLinkRT	0.5	0.46	0.46	0.5
7	NumDots	0.51			0.5
8	InsecureForms		0.32	0.74	0.3
9	PathLevel			0.78	
10	QueryLength			0.75	
11	UrlLength			0.87	0.4
12	IframeOrFrame			0.56	
13	NumQueryComponents	0.87	0.74	0.88	
14	PctExtResourceUrlsRT		0.46		
15	HostnameLength			0.75	
16	AbnormalExtFormActionR	0.43	0.43	0.81	0.4
17	NumAmpersand		0.74	0.88	
18	RandomString		0.4	0.72	
19	ExtFavicon			0.5	
20	NoHttps			0.72	
21	DomainInSubdomains			0.66	
22	SubdomainLevelRT			0.81	
23	NumUnderscore			0.73	
24	ExtFormAction			0.48	
25	RelativeFormAction		0.43	0.53	
26	TildeSymbol			0.4	
27	SubdomainLevel	0.51	0.5	0.73	0.5
28	Numdashinhostname			0.5	
29	SubmitInfoToEmail			0.56	
30	UrlLengthRT	0.5		0.43	

Also, it displays that 30 attributes with high correlation were detected. It is clear from the data detection that methods as Cosine and Kendall are the best. Also, it shows that the following are the most important attributes:

- ✓ *PctExtHyperlinks*
- ✓ *PctExtResourceUrls*
- ✓ *FrequentDomainNameMismatch*
- ✓ *ExtMetaScriptLinkRT*
- ✓ *AbnormalExtFormActionR*
- ✓ *SubdomainLevel*.

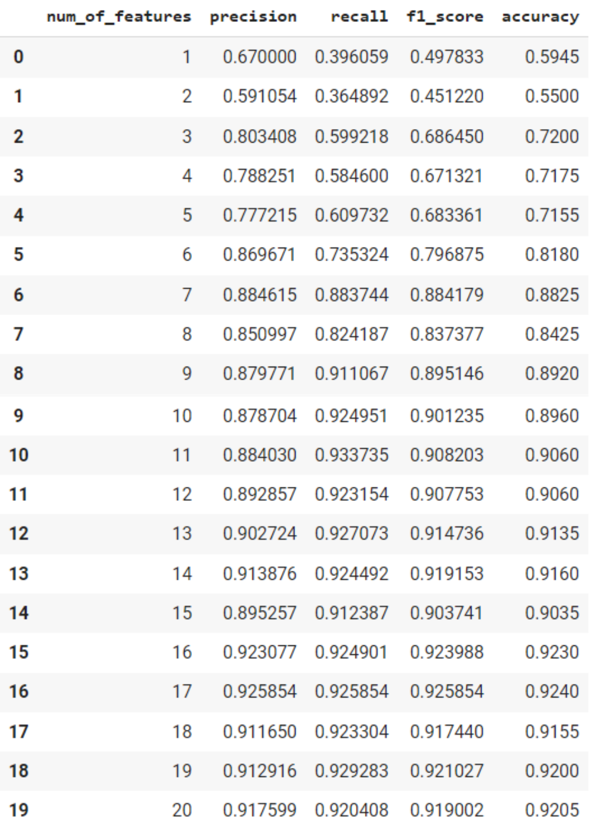
Predicting phishing sites

First using Logistic regression, the machine will learn to predict whether it is a phishing site. Evaluation measures will be accuracy_score, precision_score, recall_score, f1_score.

A. For the machine to learn Logistic models

It aims to perform learning process by repetition using Logistic regression model. For the machine, it requires several top N attributes. In this, evaluation measures such as ACCURACY, PRECISION, RECALL, F1 SCORE will be used.

Repetition starts from 1. All 50 attributes will be learned by the machine to obtain the most appropriate numerical attribute.



	num_of_features	precision	recall	f1_score	accuracy
0	1	0.670000	0.396059	0.497833	0.5945
1	2	0.591054	0.364892	0.451220	0.5500
2	3	0.803408	0.599218	0.686450	0.7200
3	4	0.788251	0.584600	0.671321	0.7175
4	5	0.777215	0.609732	0.683361	0.7155
5	6	0.869671	0.735324	0.796875	0.8180
6	7	0.884615	0.883744	0.884179	0.8825
7	8	0.850997	0.824187	0.837377	0.8425
8	9	0.879771	0.911067	0.895146	0.8920
9	10	0.878704	0.924951	0.901235	0.8960
10	11	0.884030	0.933735	0.908203	0.9060
11	12	0.892857	0.923154	0.907753	0.9060
12	13	0.902724	0.927073	0.914736	0.9135
13	14	0.913876	0.924492	0.919153	0.9160
14	15	0.895257	0.912387	0.903741	0.9035
15	16	0.923077	0.924901	0.923988	0.9230
16	17	0.925854	0.925854	0.925854	0.9240
17	18	0.911650	0.923304	0.917440	0.9155
18	19	0.912916	0.929283	0.921027	0.9200
19	20	0.917599	0.920408	0.919002	0.9205

Figure 9 – Logistic regression result when data up to 1–20 was given to be learned

B. Logistic regression result

Logistic regression is the binary classification algorithm to make the probability for everything to be true or false. Logistic regression calculates logit function. This logit function is only the probability record of events.

This method is used to predict whether the website is phishing or legitimate based on the given model.

As shown, the model has increased and decreased during the learning process. In this, the most important was to obtain the best feature among all measures. Another thing to notice is the performance of machine for Recall value was inconstant. To select the best N for which our model has the least variation in Precision and Recall values, the 7th attribute was selected from all well-performing measures.

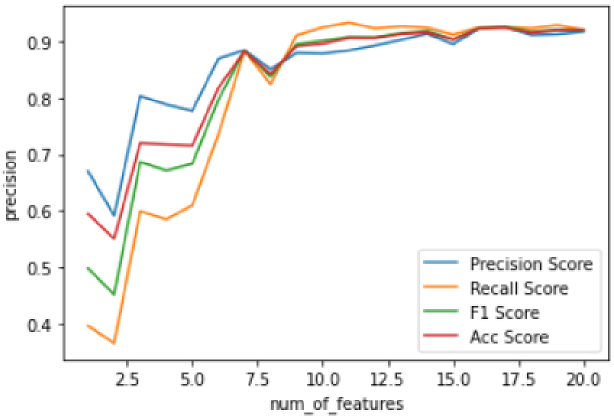


Figure 10 – Graph of Logistic regression result regarding the data 1–20

Learning will be continued through repetition starting from the 20th attribute to the 50th.

Learning result by Logistic regression regarding the data 20–50:

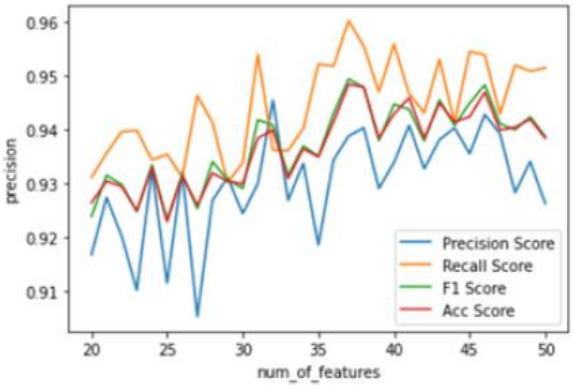


Figure 11 – Graph of Logistic regression result when data up to 20–50 was given to be learned

From the graph above, the 27th and 29th are N, the most appropriate attribute. In other words, it is better for the graph fluctuation regarding Precision and Recall measures to be less. Also, regarding whether it is a phishing site, the prediction at the 27th and 29th attributes was made correctly. It means that the machine detects phishing sites up to 93 %.

TABLE 2 – LOGISTIC REGRESSION RESULT

	Precision	**Accuracy**	**Recall**	**F1 score**
0	0,98	–	0,98	0,98
1	0,98	–	0,98	0,98
Logistic regression	0,93	0,93	0,93	0,93

That this Logistic regression model can now predict up to 93 % with 93 % accuracy, and 93 % recall proves high capability of this model to predict phishing and non-phishing sites. Now this prediction can be checked by Confusion matrix.

Model evaluation and error calculation

It is important to use independent verification and performance measures when using certain methods and techniques in machine learning. It cannot be said that the model will work best regarding unprecedented data after machine learns to process the dataset. Validation is the process of deciding whether quantitative results measuring the predicted correlation between variables is acceptable as a description of data.

Confusion matrix

There was an attempt to prove the results in machine learning by Confusion matrix.

Confusion matrix is N x N matrix which is used to evaluate classification model performance and N is the number of target classifications. Matrix compares actual target values with predicted values by machine learning model. It is possible to check in detail how well the classification model works, what types of errors it makes.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Precision is concerned with how many cases among those predicted correctly are positive.

$$\text{Precision} = \frac{TP}{TP+FP}$$

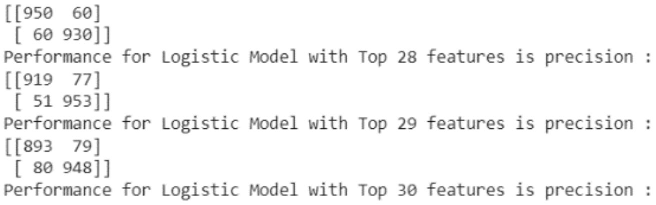
Recall shows how many of the actual positive cases the model correctly predicted.

$$\text{Recall} = \frac{TP}{TP+FN}$$

In practice, when F1 Score tries to increase Model precision, Recall can decrease or increase. It is because Precision and Recall have harmonic mean values, there is common understanding regarding these two measures.

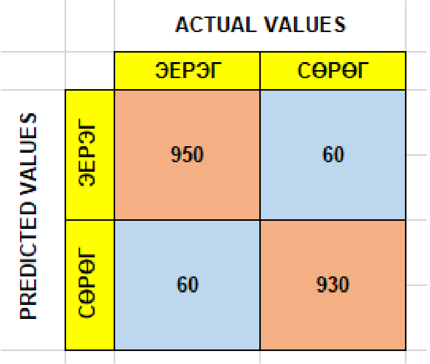
$$\text{F1 Score} = \frac{2}{\frac{1}{Recall}+\frac{1}{Precision}}$$

Now prediction up to 93 % at the 29th attribute will be checked by Confusion matrix.



```
[[950  60]
 [ 60 930]]
Performance for Logistic Model with Top 28 features is precision :
[[919  77]
 [ 51 953]]
Performance for Logistic Model with Top 29 features is precision :
[[893  79]
 [ 80 948]]
Performance for Logistic Model with Top 30 features is precision :
```

If looking at Confusion matrix results, 2000 data was tested and it proves that 950/60, 60/930 and sum of these are 2000. Because the learning at the 29th attribute started from 0, it can be seen as the 28th attribute. At the 29th attribute, prediction was done up to 93 % and it is proved by Confusion matrix.



		ACTUAL VALUES	
		ЭЕРЭГ	СӨРӨГ
PREDICTED VALUES	ЭЕРЭГ	950	60
	СӨРӨГ	60	930

True positive (TP)- from actual 2000 attributes, 950 were positive and this model predicted 950 positive attributes.

True negative (TN)- from actual 2000 attributes, 930 were negative and this model predicted 930 negative attributes.

False positive (FP)- from actual 2000 attributes, 60 were negative, but the model falsely predicted 60 attributes as positive.

False negative (FN)- predicted values were predicted falsely. Of the actual 2000 attributes, 60 were positive, but the model falsely predicted 60 attributes as negative.

From the above, it is clear when checking 2000 data by Confusion matrix, prediction was correct as given 950 + 930 = 1880. That logistic regression model predicted up to 93 % without any error was proved as true.

CONCLUSION

This research is concerned with the following:

✓ There was included the literature review regarding how cybercrime is at an international level by studying research works concerned.

✓ Quantitative indicators have been brought out through comparison by analyzing the current situation of cybercrimes in Mongolia.

✓ Literature review was done regarding cybercrime in Mongolia.

✓ A study was done to determine the education level, age, and gender of those committing cybercrimes.

✓ Machine learning methods and techniques were studied and compared.

✓ The review was written through the study on phishing attack as the main representative of social engineering.

✓ Data attributes were evaluated to detect phishing attack. In this, specific features of phishing sites were bought out using Pearson, Spearman, Kendall, and Cosine methods.

Based on Logistic regression method, model was learned by the machine. 80 % of data was provided for the machine to learn, and 20 % was experimented. It has been proved that the use of indicators such as Precision, Recall, F1 to detect phishing websites are highly possible.

From the experiment, it is obvious that attribute 29 is N- a very appropriate attribute. It means learning machine detects phishing sites by 93 % when processing the 29th attribute. The less the graph fluctuation regarding Precision and Recall measures is, the better it would be, and prediction of phishing sites was true. On the other hand, machine detects phishing sites when processing the 29th attribute.

This model can now predict up to 93 % and it proves that it has high capability to predict phishing and non-phishing sites.

References

1. Khaltar, T. Feature of invstigating cybercrimes against electronic information security / T. Khaltar. – Ulaanbaatar, 2019.

2. Khaltar, T. A deep analysis on cybercrimes against electronic information security / T. Khaltar. – Ulaanbaatar, 2019.

3. Galbaatar, L. Electronic law / L. Galbaatar. – Ulaanbaatar, 2010.

4. Zolzaya, D. Artificial intelligence and machine learning : student book / D. Zolzaya. – CRC Press. – 2021. – 154 c.

5. Odonchimeg, L. Determining whether the weather fluctuations affect cybercrime in Mongolia by using deep learning methods / L. Odonchimeg : the review of research work : International conference, 2021. – ICT 100.

6. United Nations / Comprehensive Study on Cybercrime. – 2013. – URL: https://www-unodc-org.translate.goog/unodc/en/organized-crime/ comprehensive-study-on-cybercrime.html (date of access: 21.10.2024).

7. United Nations / Cybercrime study. – 2013. – URL: https://www.unodc.org/documents/-organized-crime/unodc_ccpcj_eg.4_2013/cybercrime_study_ 210213.pdf (date of access: 21.10.2024).

8. International Telecommunication Union (ITU) / Understanding Cybercrime. – 2014

9. Statistical news. General police department. – URL: http://police.gov.mn/as/static (date of access: 21.10.2024).

10. National Cyber Security Index (NCSI). – URL: https://ncsi.ega.ee/ (date of access: 21.10.2024).

11. Council Of Europe. Convention on Cybercrime. [Budapest. November, 2001]

12. http://techcarpenter.blogspot.com/2017/07/machine-learning-explained.html

13. https://medium.com/datadriveninvestor/an-introduction-to-clustering-61f6930e3e0b

14. https://www.congress.gov/bill/113th-congress/senate-bill/2521/text

15. http://www.arcyber.army.mil/

http://www.dpr.go.id/dokjdih/document/uu/UU_2006_23.pdf