ВЫЯВЛЕНИЕ ФИШИНГОВЫХ ИНТЕРНЕТ-ДОМЕНОВ С ПОМОЩЬЮ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ В РЕЖИМЕ ПОТОКОВОЙ ОБРАБОТКИ ДАННЫХ

Павлычев А. В.[1], Кузьминец К. В.[2]

DOI: 10.21681/2311-3456-2025-2-141-153

Цель работы: *целью исследования является разработка эффективного способа выявления фишинговых Интернет-доменов с помощью алгоритмов машинного обучения в режиме потоковой обработки данных.*

Метод исследования: *в рамках работы проведен анализ признаков, характеризующих произвольные Интернет-домены и разработан программный комплекс, позволивший собрать собственный датасет, содержащий набор признаков для более чем 250 000 доменов. На полученном наборе данных проведено обучение ряда моделей машинного обучения, которые сравнивались по точности и скорости работы. Выбранный классификатор использован при разработке программного прототипа, который прошел апробацию на выборке, состоящей из 1000 произвольных Интернет-доменов.*

Полученный результат: *разработан классификатор и программный прототип, позволяющий в рамках заданных параметров точности и скорости работы относить произвольный Интернет-домен к категории фишингового либо легитимного. Достоверность и обоснованность предлагаемых научных положений, результатов и выводов работы подкрепляется разносторонним изучением современного состояния предметной области, системным обоснованием предложенных моделей, не противоречащих известным положениям других авторов, проведением серии экспериментов, подтверждающих результаты теоретических изысканий. Собранный в рамках работы набор данных опубликован в открытом доступе на платформе Kaggle для использования исследователями для разработки различных интеллектуальных способов выявления фишинговых доменов.*

Научная новизна *заключается в разработке способа выявления фишинговых доменов в режиме потоковой обработки данных, что может быть использовано в системах обнаружения вторжений и при разработке веб-приложений, защищающих пользователей от нежелательного контента. На основании полученной модели реализован и апробирован прототип программного обеспечения, продемонстрировавший точность работы 98,5 % при средней скорости обработки одного ресурса 1,2 секунды.*

Ключевые слова: *фишинговые домены, интернет-безопасность, машинное обучение, потоковая обработка данных, алгоритмы классификации, выявление фишинга.*

Введение

Выявление фишинговых Интернет-ресурсов не теряет своей актуальности последнее десятилетие. Кража личных данных и финансовой информации с использованием поддельных веб-сайтов остается одним из самых эффективных и технически несложных видов Интернет-мошенничества.

Согласно как российским, так и зарубежным исследованиям, количество фишинговых атак продолжает ежегодно расти. Злоумышленники становятся более изобретательными, изучают поведение пользователей, все чаще используют психологические методы, чтобы убедить пользователей перейти на фишинговые сайты. Сами поддельные ресурсы становятся все более схожими с настоящими, что делает их обнаружение для обычных пользователей и даже антивирусных программ более сложным.

Актуальность темы подтверждает большое количество исследований, направленных на разработку методов и инструментов для обнаружения и предотвращения фишинговых атак.

Целью исследования является разработка способа выявления фишинговых Интернет-ресурсов с помощью алгоритмов машинного обучения в режиме потоковой обработки данных.

Объектом исследования является набор признаков и алгоритм выявления фишинговых Интернет-доменов, в совокупности дающие необходимое сочетание точности и скорости работы. В рамках работы проведен сравнительный анализ по современным публикациям на тему обнаружения фишинговых доменов и предложены варианты, обеспечивающие оптимальное сочетание точности и скорости

1 Павлычев Алексей Викторович, директор Центра информационной безопасности, доцент департамента информационной безопасности Института математики и компьютерных технологий ФГАОУ ВО «Дальневосточный федеральный университет» (ДВФУ), г. Владивосток, Россия. E-mail: pavlychev.av@dvfu.ru

2 Кузьминец Кирилл Вячеславович, главный специалист по защите информации ФГАОУ ВО «Дальневосточный федеральный университет» (ДВФУ), г. Владивосток, Россия. E-mail: kuzminetc.kv@dvfu.ru

работы. Новизной предлагаемого подхода является выявление фишинговых сайтов в режиме потоковой обработки данных, что может быть использовано в системах обнаружения вторжений и при разработке веб-приложений, защищающих пользователей от нежелательного контента.

В рамках исследования осуществлен поиск оптимального способа выявления фишинговых Интернет-ресурсов на основании заданного набора признаков. В качестве критериев оптимальности рассматриваются эффективность работы предлагаемого алгоритма по установленной метрике (пороговое значение – не менее 90%) и время, необходимое для обработки данных для одного домена (пороговое значение – не более 2 секунд).

Авторами разработан программный комплекс и собран собственный набор данных, содержащий признаки более чем для 200 000 Интернет-доменов, из которых легитимные ресурсы составляют не менее 100 000, фишинговые – также не менее 100 000.

Основные задачи, поставленные в рамках научно-исследовательской работы:

1. Обзор источников литературы по теме исследования с целью определения подхода к выявлению фишинговых Интернет-доменов с точки зрения заданных целевых характеристик;
2. Формирование выборки, содержащей легитимные и фишинговые Интернет-домены;
3. Обучение моделей машинного обучения и сравнение их эффективности по заданным критериям;
4. Выбор набора признаков, отвечающих условиям эффективности классификации и скорости работы;
5. Оптимизация модели на выбранном наборе признаков;
6. Проверка работы разработанного прототипа на произвольном наборе Интернет-доменов и оценка качества полученных результатов.

Способы выявления фишинговых Интернет-ресурсов

Согласно как российским, так и зарубежным исследованиям, количество фишинговых атак продолжает ежегодно расти, что стало одной из главных угроз 2023 года[3]. Злоумышленники становятся более изобретательными, изучают поведение пользователей, все чаще используют психологические методы, чтобы убедить пользователей перейти на фишинговые сайты. Сами поддельные ресурсы становятся все более схожими с настоящими, что делает их обнаружение для обычных пользователей и даже антивирусных программ более сложным. Помимо уже распространенных тактик и техник, на вооружении киберпреступников все чаще стали использоваться нейросети, такие как ChatGPT [1].

Проведенный анализ источников литературы по теме выявления фишинговых Интернет-ресурсов [2–8] позволяет выделить следующие основные подходы при исследовании данной предметной области:

1. Анализ URL-адресов. Исследователи анализируют URL-адреса веб-сайтов с целью выявления характеристик, которые могут указывать на фишинг, такие как поддельные домены, опечатки в названиях и т.д.
2. Сбор и анализ данных о доменах. Методы сбора данных о доменах, включая WHOIS-запросы и анализ DNS-записей, могут помочь выявить аномалии и признаки, указывающие на возможное использование домена для фишинга;
3. Анализ содержания веб-страниц: Исследователи анализируют HTML-код и содержание веб-страниц, чтобы выявить подозрительные элементы, такие как фальшивые логотипы, запросы на ввод личных данных и др.;
4. Машинное обучение. Многие исследования используют методы машинного обучения для создания моделей, способных автоматически классифицировать веб-сайты как фишинговые или легитимные на основе различных признаков, таких как содержание страниц и поведение пользователей;
5. Анализ поведения пользователей. Исследователи могут изучать поведение пользователей, их реакции на подозрительные веб-сайты и попытки перехвата данных;
6. Социальная инженерия. Исследования в области социальной инженерии позволяют понять, как злоумышленники манипулируют психологией пользователей для получения доступа к личным данным;
7. Анализ данных о фишинге. Исследователи собирают данные о фишинговых атаках и анализируют их, чтобы выявить новые тренды, тактики и уязвимости.

Приведенные методы и подходы часто комбинируются для создания более эффективных систем обнаружения и предотвращения фишинговых атак.

Наибольшую популярность в задаче выявления фишинговых Интернет-ресурсов получили алгоритмы машинного обучения ввиду их эффективности и несложной технической реализации. Проведено большое количество исследований [9–13], сравнивающих различные алгоритмы машинного обучения с точки зрения точности выявления фишинга.

При этом вопросы быстродействия обучения моделей и работы самих алгоритмов либо

3 В соответствии с исследованиями, число мошеннических сайтов в 2023 г. выросло на 86 %. URL: https://bi.zone/news/chislo-moshennicheskikh-saytov-v-2023-godu-vyroslo-na-86/

не рассматриваются, либо рассматриваются в качестве дополнительной информации, которой не уделяется должного внимания. Однако скорость обработки данных при применении методов машинного обучения в прикладных задачах защиты информации может иметь решающее значение, поскольку пользователи современных Интернет-ресурсов привыкли к быстрой работе веб-приложений и любые задержки в их работе могут привести к негативной реакции с их стороны, вплоть до отказа от дальнейшего использования ресурса.

Использование алгоритмов машинного обучения для выявления фишинговых сайтов является распространенной темой различных исследований. Точность моделей может достигать 90–98 % [14]. Однако критически важным моментом являются временные задержки, необходимые для работы классификатора, которые могут достигать 15–30 секунд [15] в зависимости от исследуемого набора признаков. Данные ограничения делают невозможным применение таких классификаторов в режиме реального времени.

Очевидно, что чем больше признаков исследуется, тем выше точность алгоритма, но и тем больше времени требуется классификатору на вынесение вердикта. Наибольшие временные затраты необходимы при изучении контента веб-приложений на предмет поиска различных элементов, демаскирующих фишинговые ресурсы: наличием межсайтовых скриптов, форм для сбора данных о банковских картах, картинки с изображением платежных систем и другие. Самые быстрые в вычислении те признаки, которые связаны с исследованием доменного имени или URL-адреса, такие как длина доменного имени, количество поддоменов, наличие спецсимволов и другие. Третью группу признаков составляют данные со сторонних ресурсов, дающие информацию о регистрационных данных (WHOIS-сервисы), а также информацию о DNS-записях.

Исследований, объектом которых являются именно временные задержки, необходимые на обработку различных признаков фишинговых сайтов, в открытом доступе найдено не было.

Для обучения классификаторов в рамках задачи поиска фишинговых сайтов с помощью алгоритмов машинного обучения исследователи используют два основных подхода:

1. Используют открытые датасеты. Наиболее известным является набор данных из UCI Machine Learning Repository[4], содержащий данные о 11,000 URL-адресах с 30 различными характеристиками. Ряд датасетов можно найти в открытом доступе в репозиториях Kaggle[5]. Плюсом использования открытых датасетов является простота доступа и возможность оценки предлагаемого алгоритма с работами других исследователей. К отрицательным сторонам данного подхода можно отнести их давность (большинство из них сформированы более 10 лет назад), а также относительно небольшие объемы, что может отрицательно сказаться на применении алгоритмов на новых Интернет-ресурсах;
2. С помощью различных инструментов исследователи собирают собственные датасеты из фишинговых и легитимных ресурсов. В большинстве случаев используются Интернет-ресурсы https://openphish.com/ и https://phishtank.org/. Данный подход более трудоемок, но дает возможность исследователям экспериментировать и собирать признаки, отсутствующие в открытых датасетах.

Указанные группы признаков, характеризующие Интернет-ресурсы, могут включать в себя следующие: наименование субъекта базы WHOIS, наименование организации из базы WHOIS, наименование центра сертификации, издавшего SSL-сертификат, корневой домен, домен второго уровня, количество поддоменов, длина домена, возраст домена (количество дней с момента регистрации домена), количество DNS-записей типа MX, количество DNS-записей типа TXT, количество DNS-записей типа A, количество DNS-записей типа NS, количество DNS-записей типа CNAME, наличие favicon у домена, количество цифр в доменном имени, количество дефисов в доменном имени, количество гласных в доменном имени, количество согласных в доменном имени.

Формирование набора данных

Для обучения модели поставлена задача сформировать датасет, который будет включать в себя информацию о признаках не менее чем 200 000 доменов, из которых:

- 100000 легитимных доменов;
- 100000 фишинговых доменов.

Легитимные домены получены из статистики Cloudflare Radar (https://radar.cloudflare.com/domains). Сервис занимается оценкой популярности Интернет-доменов, собирает и анализирует активность пользователей сети Интернет.

Фишинговые домены собраны с помощью двух web-ресурсов:

- https://openphish.com/
- https://phishtank.org/

Оба ресурса занимаются сбором и анализом фишинговых ресурсов. С ресурса Phishtank скачана и проанализирована база подтвержденных фишинговых ресурсов. С ресурса Openphish на постоянной основе производилась выгрузка бесплатного

4 UCI Machine Learning Repository: Phishing Websites Data Set. URL: https://archive.ics.uci.edu/ml/datasets/phishing+websites

5 Phishing Dataset for Machine Learning. URL: https://www.kaggle.com/datasets/shashwatwork/phishing-dataset-for-machine-learning

	ID	DOMAIN_NAME	TS_RATE_CALC	IP	WHOIS_NAME	WHOIS_ORG	WHOIS_REGISTRAR
0	23	000000000000000000000000.findyourjacket.com	2024-01-17 20:33:49.000	75.2.37.224	domain admin	whois privacy corp.	internet domain service bs corp.
1	24	00000000000000000gg.000webhostapp.com	2024-01-17 20:33:53.000	145.14.145.152	gdpr masked	gdpr masked	hostinger operations, uab
2	25	000000008838383839929292922222.ratingandreviews.in	2024-01-17 20:33:54.000	135.181.142.217	NaN	NaN	hosting concepts b.v. d/b/a openprovider
3	26	0000000095.godaddysites.com	2024-01-17 20:33:55.000	76.223.105.230	registration private	domains by proxy, llc	godaddy.com, llc
4	27	00000002.c1.biz	2024-01-17 20:33:58.000	185.176.43.106	NaN	NaN	domaincontext, inc.
...	...	...	...	...	...	...	...
258145	404095	bafybeidw6oz5afnunqnjucw3gxwu7umkszq354c5mbxq4...	2024-01-27 12:20:47.000	54.205.31.215	NaN	NaN	gandi sas
258146	404096	uspsuxw.top	2024-01-27 12:21:01.000	162.62.126.246	redacted for privacy	bang1	alibaba.com singapore e-commerce private limited
258147	404097	login.dosensosiologi.com	2024-01-27 12:21:03.000	103.153.182.185	domain admin	privacy protect, llc (privacyprotect.org)	hostinger operations, uab
258148	404098	currently-email-att-activation-update-6287483....	2024-01-27 12:21:12.000	199.34.228.40	NaN	NaN	NaN
258149	404099	royal-uk-ramhandtohand.codeanyapp.com	2024-01-27 12:21:15.000	45.55.112.74	registration private	domains by proxy, llc	godaddy.com, llc

258150 rows × 24 columns

Рис. 1. Набор данных с информацией о признаках Интернет-доменов

варианта данных. Данные в такой подписке обновляются каждые 12 часов.

Для автоматического сбора информации о фишинговых доменах разработан программный комплекс, который ежедневно загружал данные о новых ресурсах, вычислял признаки и записывал полученные результаты в базу данных. Итоговый датасет имеет вид, представленный на рисунке 1.

Схема таблицы представлена следующим образом:

- ID – порядковый номер строки в таблице;
- DOMAIN_NAME – доменное имя;
- TS_RATE_CALC – дата\время обработки записи;
- IP – IP-адрес, в который резолвится домен;
- WHOIS_NAME – имя владельца домена;
- WHOIS_ORG – данные организации, на которую зарегистрирован домен;
- WHOIS_REGISTRAR – данные организации-регистратора домена;
- CERT_ISSUER – организация, выпустившая сертификат;
- ROOT_DOMAIN – корневой домен;
- SECOND_DOMAIN – домен второго уровня;
- SUBDOMAINS_COUNT – количество поддоменов в домене;
- DOMAIN_LENGTH – длина домена;
- DOMAIN_AGE – возраст домена;
- DNS_MX_COUNT – количество DNS-записей типа MX;
- DNS_TXT_COUNT - количество DNS-записей типа TXT;
- DNS_A_COUNT - количество DNS-записей типа A;
- DNS_NS_COUNT - количество DNS-записей типа NS;
- DNS_CNAME_COUNT - количество DNS-записей типа CNAME;
- FAVICON – наличие у домена Favicon;
- SYMS_DIGS – количество цифр в домене;
- SYMS_DASH – количество дефисов в домене;
- SYMS_VOWELS – количество гласных в домене;
- SYMS_CONSONANTS – количество согласных в домене;
- IS_PHISHING – целевой признак принадлежности домена к фишингу (бинарный признак: 0 – легитимный домен, 1 – фишинговый домен).

В течение 6 месяцев таким способом удалось собрать данные о признаках 284 244 доменов. Собранный датасет опубликован в открытом доступе на платформе Kaggle и доступен по ссылке: https://www.kaggle.com/datasets/golner/phishing-domains.

Измерение скорости вычисления признаков

Для потоковой обработки данных критичное значение имеет скорость вычисления. В рамках исследования поставлена задача выбрать набор признаков, суммарное время вычисления которых не превышает 2 секунды.

Собранные в датасете признаки имеют различную скорость вычисления: ряд из них считается за доли секунды, на вычисление других может потребоваться время от 1 до 5 секунд и более, что связано с необходимостью получения данных со стороннего сервиса.

Для получения точных данных проведен эксперимент. Взяты произвольные 50 легитимных доменов и 50 фишинговых доменов и вычислена для каждого из них скорость выделения признаков.

WHOIS-признаки (онлайн-признаки, за один запрос к внешнему сервису получается информация о возрасте домена, наименовании владельца

домена и организации, на которую зарегистрирован домен, наименовании организации – регистратора домена).

- Максимальное время – 2,97 сек.
- Минимальное время – 0,15 сек.
- Среднее значение – 0,96 сек.

Наименование центра сертификации, издавшего SSL-сертификат (онлайн-признак, определяется путем направления запроса на сам домен).

- Максимальное время – 5,19 сек.
- Минимальное время – 0,01 сек.
- Среднее значение – 1,01 сек.

Количество DNS-записей типа MX (онлайн-признак, определяется путем направления запроса к внешнему сервису).

- Максимальное время – 1,0 сек.
- Минимальное время – 0,12 сек.
- Среднее значение – 0,20 сек.

Количество DNS-записей типа TXT (онлайн-признак, определяется путем направления запроса к внешнему сервису).

- Максимальное время – 1,0 сек.
- Минимальное время – 0,13 сек.
- Среднее значение – 0,22 сек.

Количество DNS-записей типа A (онлайн-признак, определяется путем направления запроса к внешнему сервису).

- Максимальное время – 1,0 сек.
- Минимальное время – 0,12 сек.
- Среднее значение – 0,20 сек.

Количество DNS-записей типа NS (онлайн-признак, определяется путем направления запроса к внешнему сервису).

- Максимальное время – 1,0 сек.
- Минимальное время – 0,12 сек.
- Среднее значение – 0,18 сек.

Количество DNS-записей типа CNAME (онлайн-признак, определяется путем направления запроса к внешнему сервису).

- Максимальное время – 1,0 сек.
- Минимальное время – 0,12 сек.
- Среднее значение – 0,17 сек.

Наличие favicon у домена (онлайн-признак, определяется путем направления запроса на сам домен).

- Максимальное время – 7,33 сек.
- Минимальное время – 0,01 сек.
- Среднее значение – 1,66 сек.

Для следующих признаков экспериментальное время вычисления время является идентичным:

- Корневой домен (офлайн-признак, вычисляется из строки адреса домена);
- Домен второго уровня (офлайн-признак, вычисляется из строки адреса домена);
- Количество поддоменов (офлайн-признак, вычисляется из строки адреса домена);
- Длина домена (офлайн-признак, вычисляется из строки адреса домена);
- Количество цифр в доменном имени (офлайн-признак, вычисляется из строки адреса домена);
- Количество дефисов в доменном имени (офлайн-признак, вычисляется из строки адреса домена);
- Количество гласных в доменном имени (офлайн-признак, вычисляется из строки адреса домена);
- Количество согласных в доменном имени (офлайн-признак, вычисляется из строки адреса домена);
- Максимальное время – 0,0 сек.
- Минимальное время – 0,0 сек.
- Среднее значение – 0,0 сек.

Для удобства анализа полученные данные приведены в табличной форме (Рис.2).

На основании полученного набора данных построены графики, наглядно отражающие скорость вычисления признаков (Рис. 3).

	WHOIS_SPEED	CERT_ISSUER_SPEED	ROOT_DOMAIN_SPEED	SECOND_DOMAIN_SPEED	SUBDOMAINS_COUNT_SPEED	DOMAIN_LENGTH_SPEED
0	1.30	0.46	0.0	0.0	0.0	0.0
1	2.97	0.69	0.0	0.0	0.0	0.0
2	0.81	0.49	0.0	0.0	0.0	0.0
3	1.27	0.01	0.0	0.0	0.0	0.0
4	0.27	0.29	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...
95	0.84	5.16	0.0	0.0	0.0	0.0
96	0.24	0.28	0.0	0.0	0.0	0.0
97	1.13	5.06	0.0	0.0	0.0	0.0
98	2.37	0.02	0.0	0.0	0.0	0.0
99	1.06	0.37	0.0	0.0	0.0	0.0

100 rows × 16 columns

Рис. 2. Набор данных с информацией о скорости вычисления признаков

DOI: 10.21681/2311-3456-2025-2-141-153

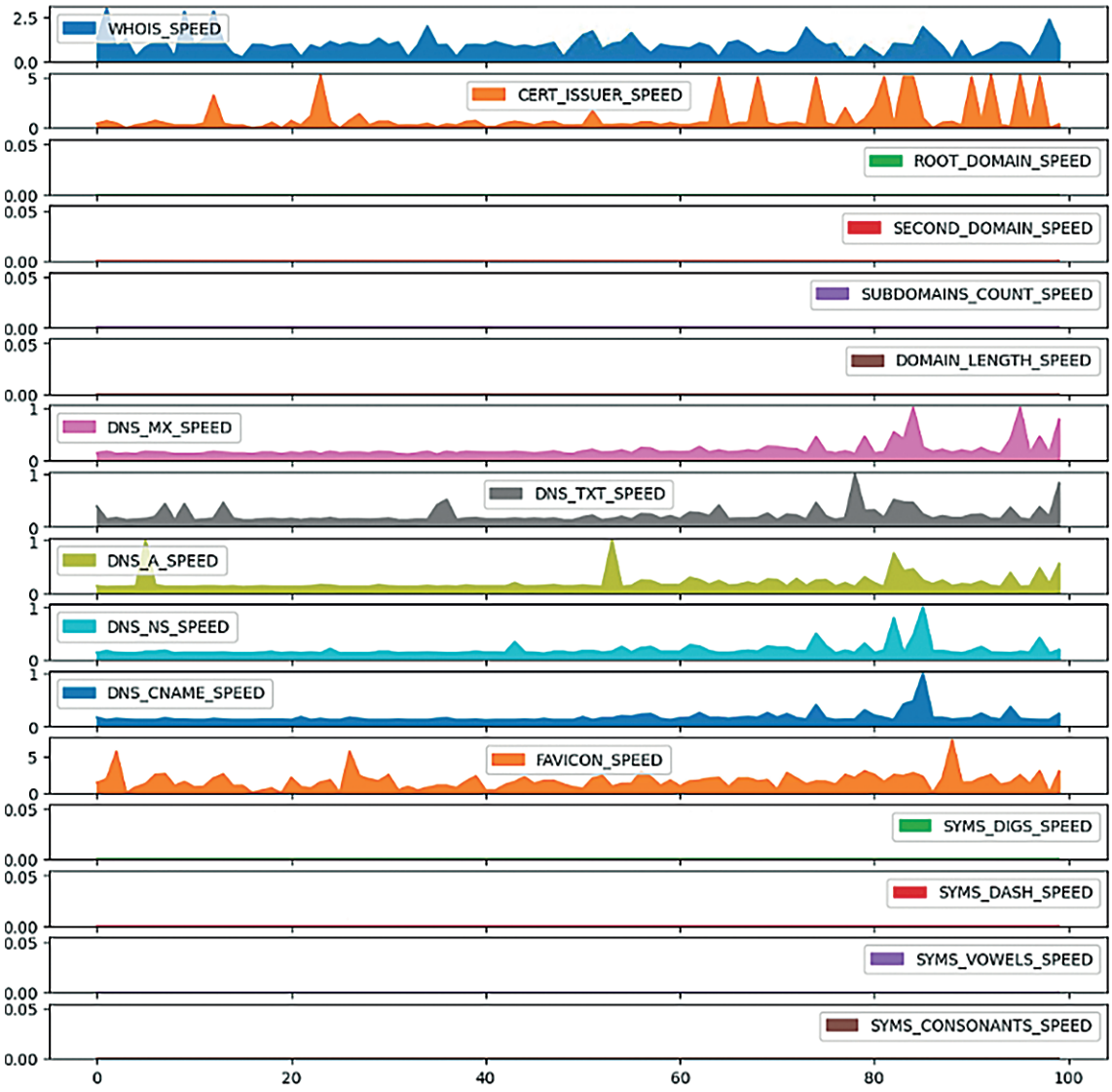

Рис. 3. Визуализация скорости получения признаков

Исходя из полученных результатов, можно сделать вывод, что наибольшие временные издержки необходимы для вычисления следующих признаков доменов:

- WHOIS-признаки (информация о возрасте домена, наименовании субъекта базы, наименовании организации);
- Наименование центра сертификации, издавшего SSL-сертификат;
- Наличие favicon у домена.

Полученные результаты использованы при формировании набора признаков, участвующих при обучении классификатора с целью поиска оптимального сочетания точности и скорости работы.

Обучение и сравнение классификаторов

На следующем этапе исследования в рамках задачи реализации прототипа модели выявления фишинговых Интернет-доменов необходимо на собранном датасете обучить классификаторы и провести их сравнительное исследование на основе эффективности их работы.

Для обучения моделей будем использовать следующие популярные алгоритмы машинного обучения:

- Логистическая регрессия (Logistic Regression);
- Наивный Байес (Naive Bayes);
- Метод опорных векторов (Support Vector Machine, SVM);
- Деревья решений (Decision Trees);
- Нейронные сети (Neural networks);
- Случайный лес (Random Forest);
- K-ближайших соседей (K-neighbors);
- Экстремальный градиентный бустинг (XGBoost);
- Категориальный бустинг (CatBoost).

Для решения задач классификации необходимо ввести метрики, по которым мы будем оценивать полученные модели.

В задачах классификации общепринятыми являются следующие определения:

TP – true positive (истинно-положительное решение): классификатор верно отнёс объект к рассматриваемому классу;

TN – true negative (истинно-отрицательное решение): классификатор верно утверждает, что объект не принадлежит к рассматриваемому классу;

FP – false positive (ложноположительное решение): классификатор неверно отнёс объект к рассматриваемому классу;

FN – false negative (ложноотрицательное решение): классификатор неверно утверждает, что объект не принадлежит к рассматриваемому классу.

Рассмотрим основные метрики.

Accuracy – доля правильных классификаций. Несмотря на очевидность и простоту, является одной из самых малоинформативных оценок классификаторов. Считается по следующей формуле:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

Recall – полнота, TPR (true positive rate), показывает отношение верно классифицированных объектов класса к общему числу элементов этого класса.

$$Recall = TP / (TP + FN) \quad (2)$$

Precision – точность, показывает долю верно классифицированных объектов среди всех объектов, которые к этому классу отнес классификатор.

$$Precision = TP / (TP + FP) \quad (3)$$

Fall-out – FPR (false positive rate), показывает долю неверных срабатываний классификатора к общему числу объектов за пределами класса. Характеризует, насколько часто классификатор ошибается при отнесении того или иного объекта к классу.

$$FPR = FP / (FP + TN) \quad (4)$$

Для получения более взвешенных оценок качества моделей зачастую используют различные виды агрегации точности и полноты.

Арифметическое среднее:

$$A = 1 / 2 * (Precision + Recall) \quad (5)$$

Минимум:

$$M = \min (precision, recall) \quad (6)$$

Гармоническое среднее, F-мера:

$$F = 2 (Precision * Recall) / (Precision + Recall) \quad (7)$$

К полученному датасету в соответствии с алгоритмом последовательно применены указанные выше алгоритмы машинного обучения.

В качестве инструмента использовался язык программирования Python 3.0 и библиотека для анализа данных Scikit-learn.

В качестве показателя эффективности в рамках текущего исследования предлагается использование гармонического среднего (F-меры), которое учитывает как ложноположительные, так и ложноотрицательные ответы классификатора, что имеет важное значение в задаче выявления фишинговых Интернет-доменов.

Обучение производилось на тренировочной выборке, проверка на тестовой с использованием модуля train_test_split. Соотношение между ними составило 75/25.

В Таблице 1 приведены значения эффективности разработанных классификаторов, полученные на тестовой выборках.

Полученные результаты демонстрируют, что наибольшую эффективность показывает алгоритм «случайный лес» (Random Forest) – более 96,1 %. Близкие по значению результаты работы также продемонстрировали алгоритмы категориального бустинга (CatBoost) – 95,88 % и экстремального градиентного бустинга (XGBoost) – 95,58 %.

Таблица 1.
Эффективность по F-мере рассмотренных классификаторов

Классификатор	Эффективность (F-мера)
RandomForestClassifier	0,961046205336391
CatBoostClassifier	0,9588614459698163
XGBClassifier	0,9557779912609625
DecisionTreeClassifier	0,9330471970002168
KNeighborsClassifier	0,8990827109609842
SVM	0,7996839071554742
MLPClassifier	0,7824692429266479
GaussianNBClassifier	0,7799281043726177
LogisticRegression	0,7784715981282345

В рамках исследования в дальнейшем будем использовать для классификации Интернет-доменов алгоритм машинного обучения «случайный лес», показавший наилучшие результаты.

Выбор набора признаков

С целью определения возможности использования полученной модели в режиме потоковой обработки данных на следующем этапе исследования необходимо выбрать оптимальный набор признаков для обучения классификатора, который должен отвечать следующим условиям:

- Алгоритм не должен быть зависим от каких-либо сторонних онлайн-ресурсов, которые могут стать недоступными или быть ограниченными по производительности;
- Точность обнаружения ≥ 90 %;

DOI: 10.21681/2311-3456-2025-2-141-153

	RandomForestClassifier	CatBoostClassifier	XGBClassifier	feature_mean_speed
SUBDOMAINS_COUNT	0.029352	0.035111	0.038132	0.0000
DOMAIN_LENGTH	0.036430	0.030094	0.024138	0.0000
DOMAIN_AGE	0.223115	0.142844	0.182882	NaN
DNS_MX_COUNT	0.022481	0.021672	0.079217	0.2038
DNS_TXT_COUNT	0.044477	0.032100	0.082146	0.2236
DNS_A_COUNT	0.050645	0.060399	0.058850	0.2018
DNS_NS_COUNT	0.073059	0.085002	0.089941	0.1874
DNS_CNAME_COUNT	0.022529	0.032080	0.070528	0.1744
FAVICON	0.008167	0.007507	0.011375	1.6676
SYMS_DIGS	0.034540	0.059328	0.091387	0.0000
SYMS_DASH	0.011532	0.021776	0.027319	0.0000
SYMS_VOWELS	0.021859	0.018412	0.018576	0.0000
SYMS_CONSONANTS	0.024608	0.016053	0.022747	0.0000
WHOIS_NAME_CODE	0.033321	0.018481	0.020215	NaN
WHOIS_ORG_CODE	0.090610	0.054479	0.043092	NaN
WHOIS_REGISTRAR_CODE	0.111215	0.105281	0.036728	NaN
CERT_ISSUER_CODE	0.057846	0.096228	0.051659	1.0111
ROOT_DOMAIN_CODE	0.034323	0.049707	0.024812	0.0000
SECOND_DOMAIN_CODE	0.069890	0.113445	0.026256	0.0000

Рис. 4. Значимость признаков для первой модели и скорость их вычисления

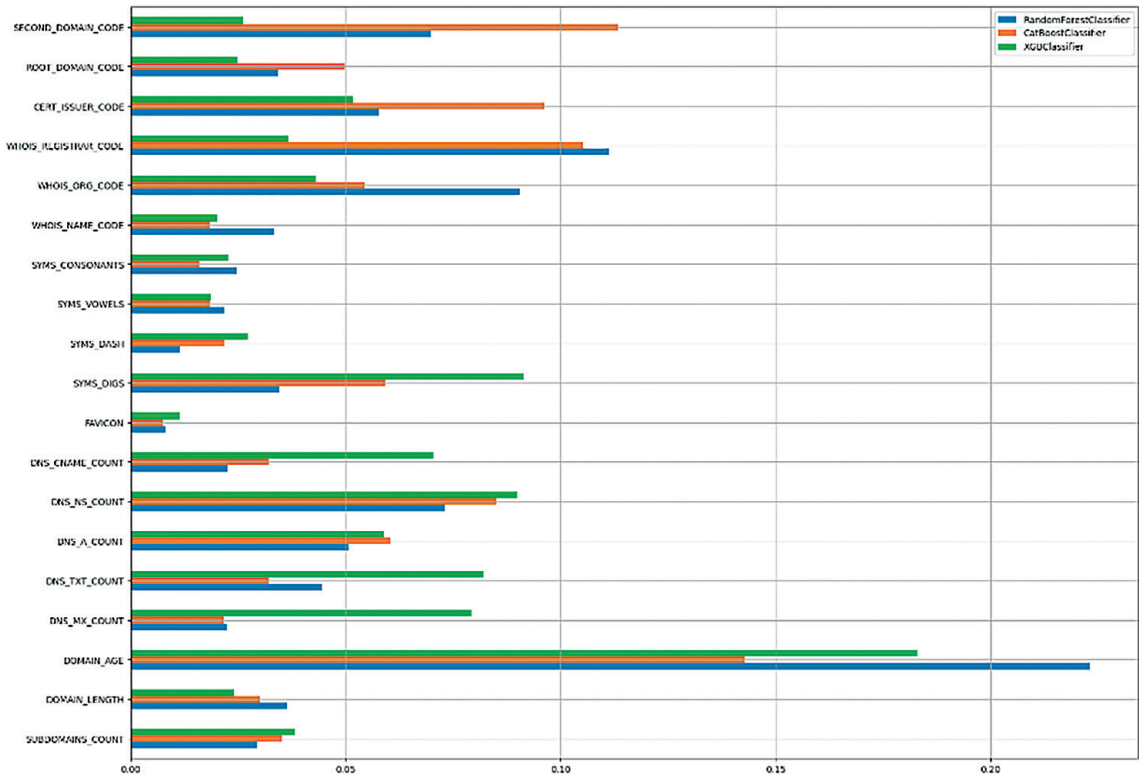

Рис.5. Гистограмма значимости признаков первой модели для различных классификаторов

- Достижимое время на обработку одного Интернет-домена (включая формирование признаков) ≤ 2 сек.

В рамках исследования предстоит выяснить «важность» признаков – какое значение каждый из них имеет в задаче классификации. Для этих целей используем стандартные атрибуты рассмотренных классификаторов – feature_importances_. Для наглядности также добавим среднюю скорость вычисления признаков, которую мы рассчитали ранее.

На первом этапе для обучения модели используем все 19 признаков. Результаты представлены на рисунках 4, 5.

Гистограмма наглядно демонстрирует, что для всех трех классификаторов, наибольшее значение имеет признак DOMAIN_AGE (входит в группу whois-признаков), наименьшее значение имеет признак FAVICON.

Как уже было рассчитано ранее, эффективность классификатора для всех 19 признаков составляет 96,10 %, что удовлетворяет поставленным условиям.

Для расчета скорости работы алгоритма просуммируем скорости вычисления признаков, рассчитанные ранее (признаки DOMAIN_AGE, WHOIS_NAME_CODE, WHOIS_ORG_CODE, WHOIS_REGISTRAR_CODE входят в группу признаков WHOIS). Результат составил 4,63 сек, что превышает пороговое значение.

Также в первой модели используется ряд онлайн-признаков, зависимые он стороннего WHOIS-сервиса, который может быть недоступен и негативно повлиять на работу классификатора.

Во второй модели удалим признаки, связанные с WHOIS-сервисом (DOMAIN_AGE, WHOIS_NAME_CODE, WHOIS_ORG_CODE, WHOIS_REGISTRAR_CODE). В обучении модели теперь участвует 15 признаков. Эффективность второй модели составляет 94,48 % (удовлетворяет поставленным условиям), скорость вычисления признаков – 3,67 сек. (не удовлетворяет условиям).

На следующем этапе удалим из модели самые «медленные» признаки – FAVICON (1,67 сек) и CERT_ISSUER_CODE (1,01 сек) для выполнения условия по скорости вычисления. В обучении модели теперь участвует 13 признаков. Эффективность третьей модели составляет 92,99 % (удовлетворяет поставленным условиям), скорость вычисления признаков – 0,99 сек. (удовлетворяет условиям).

В качестве контрольного эксперимента обучим модель, содержащую только офлайн-признаки, не требующие обращения к сторонним сервисам. В обучении модели теперь участвует только 8 признаков. Эффективность четвертой модели составляет 86,59 % (не удовлетворяет поставленным условиям), скорость вычисления признаков – 0,0 сек. (удовлетворяет условиям).

Сводные результаты подбора оптимального набора признаков указаны в таблице 2.

Таким образом, под условия, заданные в рамках исследования, подходит модель, содержащая 13 признаков: SUBDOMAINS_COUNT, DOMAIN_LENGTH, DNS_MX_COUNT, DNS_TXT_COUNT, DNS_A_COUNT, DNS_NS_COUNT, DNS_CNAME_COUNT, SYMS_DIGS, SYMS_DASH, SYMS_VOWELS, SYMS_CONSONANTS, ROOT_DOMAIN_CODE, SECOND_DOMAIN_CODE.

Таблица 2.

Результаты замеров эффективности и скорости моделей

Модель	Эффективность, %	Скорость, сек	Выполнение условия по независимости от внешних сервисов	Выполнение условия по эффективности	Выполнение условия по скорости
Первая модель, 19 признаков (полный набор признаков)	96,10	4,63	нет	да	нет
Вторая модель, 15 признаков (без whois-признаков)	94,48	3,67	да	да	нет
Третья модель, 13 признаков (без «медленных» признаков)	92,99	0,99	да	да	да
Четвертая модель, 8 признаков (только офлайн-признаки)	86,59	0,0	да	нет	да

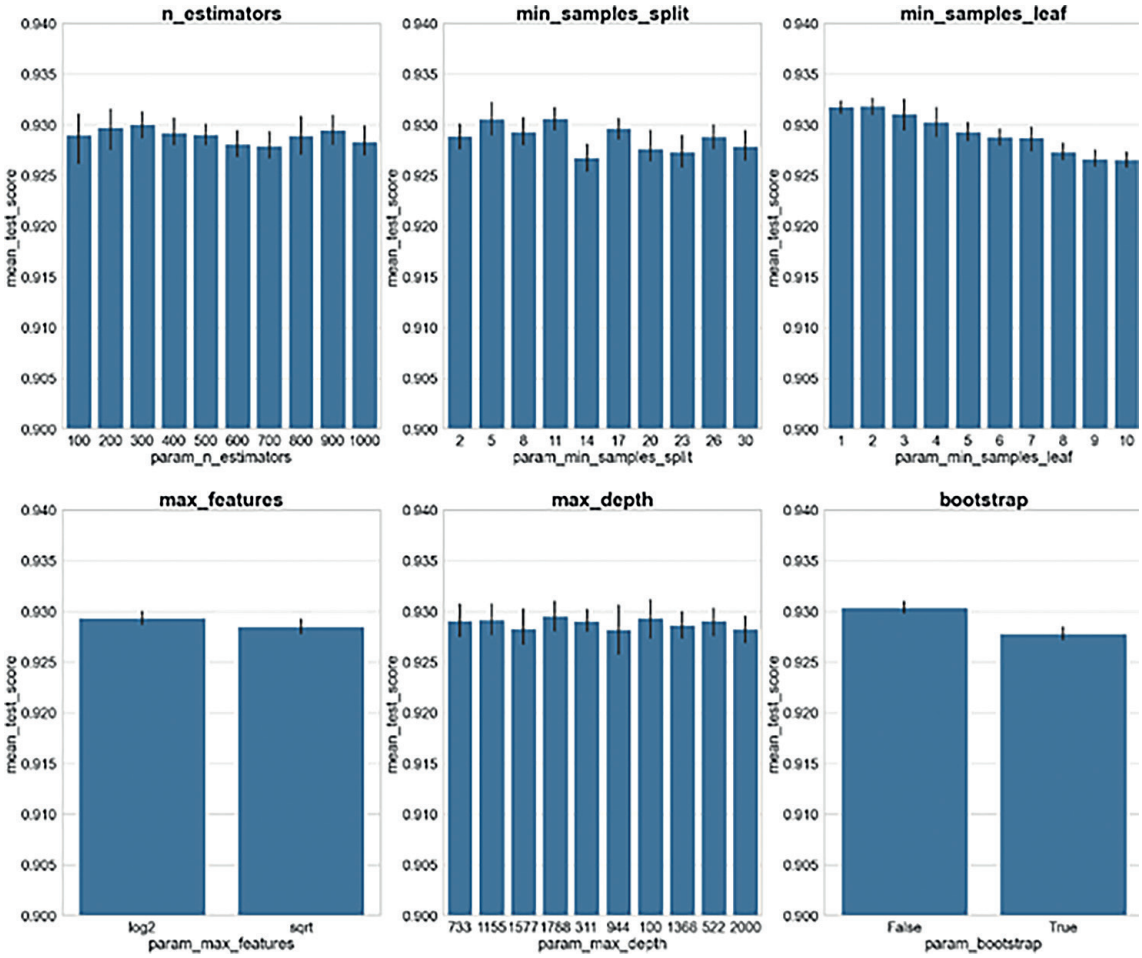

Рис. 6. Гистограммы использования параметров в ходе работы RandomizedSearchCV

Оптимизация классификатора

После определения набора признаков предстоит оптимизировать нашу модель. Для подбора гиперпараметров для RandomForest в библиотеке Scikit-learn содержится класс RandomizedSearchCV, который с помощью заданного словаря позволяет подобрать случайные гиперпараметры и выбрать из них лучший набор.

В рамках нашего исследования в результате использования RandomizedSearchCV получен следующий набор параметров:

```
{'n_estimators': 900,
 'min_samples_split': 11,
 'min_samples_leaf': 2,
 'max_features': 'log2',
 'max_depth': 733,
 'bootstrap': False}
```

Полученные значения гиперпараметров используются для обучения новой модели «случайного леса» и вычисления ее эффективности.

Итоговая эффективность модели составила 93,39%. Таким образом, путем подбора гиперпараметров удалось улучшить модель на 0,4 %.

Проверка работы полученной модели

В рамках исследования разработан программный прототип на языке Python, использующий разработанную модель машинного обучения для определения принадлежности произвольного доменного имени к фишинговым или легитимным. Интерфейс программы и пример работы приведены на рисунках 7 и 8.

```
D:\phishing_ml>python main.py
Usage:
  python script.py -f            # Process domains from domain_list.txt file
  python script.py <domain>      # Process a single domain
  python script.py               # Show this help message
```

Рис. 7. Интерфейс разработанного прототипа

```
############################################################
Total clear domains: 11
Total phishing domains: 355
Total non-existent domains: 134
### COMMON SPEED ###
Min speed: 0.12
Max speed: 7.0
Average speed: 1.22
```

Рис. 8. Пример работы разработанного прототипа

В рамках исследования проведена проверка 1000 Интернет-доменов, взятых из набора данных случайным образом.

Из 500 легитимных доменов:

- 490 доменов были правильно классифицированы как легитимные;
- 2 домена ошибочно классифицированы как фишинговые;
- 8 доменов на момент проверки не существовали;
- максимальная скорость проверки: 0,11 сек;
- минимальная скорость проверки: 5,27 сек;
- средняя скорость проверки: 1,16 сек.

Из 500 фишинговых доменов:

- 362 домена были правильно классифицированы как фишинговые;
- 9 доменов ошибочно классифицированы как легитимные;
- 129 доменов на момент проверки не существовали;
- максимальная скорость проверки: 0,11 сек;
- минимальная скорость проверки: 6,53 сек;
- средняя скорость проверки: 1,23 сек;

Таким образом, оценить качество работы модели можно на основе 863 доменов.

Проведем оценку работы прототипа по следующим показателям:

True-Positive – прототип классифицировал домен как фишинговый, и после проверки он действительно оказался фишинговым – 362;

True-Negative – прототип классифицировал домен как безопасный, и после проверки он действительно оказался безопасным – 490;

False-Positive – прототип классифицировал домен как фишинговый, но после проверки оказалось, что он безопасный – 2;

False-Negative – прототип классифицировал домен как безопасный, но после проверки выяснилось, что он фишинговый – 9.

Фактически, прототип, функционирующий на основе разработанной модели машинного обучения, ошибочно классифицировал 11 Интернет-доменов из 863, активных на момент проверки.

Значительное количество доменов (8 чистых и 129 фишинговых) на момент проверки не существовали. Это может быть связано с тем, что фишинговые домены часто имеют короткий срок жизни и могут быть отключены до проведения проверки.

На основе полученных метрик можно сделать следующие выводы об эффективности работы прототипа:

Accuracy = (362+490) / (362+490+2+9) = 0,987

Precision = 362 / (362+2) = 0,995

Recall = 362 / (362+9) = 0,976

F1_Score = 2 * (0,995 * 0,976) / (0,995+0,976) = = 0,985

Среднее время проверки (сек.) = (1,16 + 1,23) /2 = = 1,195

Таким образом, программный прототип в среднем обрабатывает один Интернет-домен за 1,2 сек., демонстрируя эффективность классификации на уровне 98,5 %.

Заключение

В рамках исследования проведен анализ различных способов и подходов к выявлению фишинговых Интернет-доменов, как с точки зрения точности обнаружения, так и скорости работы. Определен набор признаков, максимально полно характеризующих Интернет-домены и экспериментальным способом получена средняя скорость вычисления каждого из них.

На основании выделенного набора признаков сформирован набор данных, содержащий признаки более чем для 200 000 Интернет-доменов, из которых легитимные ресурсы составляют не менее 100 000, фишинговые – также не менее 100 000.

Полученный датасет использован для обучения различных классификаторов, наибольшую эффективность по F-мере среди которых показал алгоритм «случайного леса». Проведен ряд экспериментов, в результате которых выбран оптимальный набор признаков, обеспечивающий точность детектирования не менее 90% при времени обработки одного ресурса не более 2 сек. После подбора гиперпараметров финальные характеристики классификатора составили:

- точность детектирования – 93,39 %;
- скорость обработки доменного имени – 0,99 сек.

На основании проведенных теоретических результатов разработан программный прототип, который в ходе апробации показал высокую эффективность и надежность при классификации как легитимных, так и фишинговых доменов. Низкие уровни ложноположительных и ложноотрицательных срабатываний подтверждают эффективность модели.

Результаты проверки прототипа на 1000 произвольных Интернет-доменах показали эффективность на уровне 98,5 % при среднем времени обработки одного ресурса 1,2 сек. Указанная точность и скорость работы позволяет применять обученную модель в режиме потоковой обработки данных.

Таким образом, все поставленные задачи в рамках исследования решены, а цель работы достигнута.

Данная работа выполнена при поддержке «ИнфоТеКС Академия», программы поддержки научных исследований, проводимой АО «ИнфоТеКС».

Литература

1. Al-Hawawreh M., Aljuhani A., Jararweh Y. Chatgpt for cybersecurity: practical applications, challenges, and future directions //Cluster Computing. 2023. Т. 26. №. 6. С. 3421–3436. DOI: 10.1007/s10586-023-04124-5.
2. Vijayalakshmi M. et al. Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions // Iet Networks. – 2020. – Т. 9. – №. 5. – С. 235–246. DOI: 10.1049/iet-net.2020.0078.
3. Kalaharsha P., Mehtre B. M. Detecting Phishing Sites--An Overview // arXiv preprint arXiv:2103.12739. – 2021.
4. Basit A. et al. A comprehensive survey of AI-enabled phishing attacks detection techniques // Telecommunication Systems. – 2021. – Т. 76. – С. 139–154. DOI: 10.48550/arXiv.2103.12739.
5. Paliath S., Qbeitah M. A., Aldwairi M. Phishout: Effective phishing detection using selected features // 2020 27th International Conference on Telecommunications (ICT). IEEE, 2020. с. 1–5. DOI: 10.1109/ICT49546.2020.9239589.
6. Rashid J. et al. Phishing detection using machine learning technique //2020 first international conference of smart systems and emerging technologies (SMARTTECH). IEEE, 2020. С. 43–46. DOI:10.1109/SMART-TECH49988.2020.00026.
7. Селищев А. Д. Разработка программно-математического обеспечения системы информационноаналитического мониторинга фишинговых атак // Инновационные научные исследования. – 2020. – №. 12–2. – С. 33–39.
8. Афанасьева Н. С., Елизаров Д. А., Мызникова Т. А. Классификация фишинговых атак и меры противодействия им // Инженерный вестник Дона. – 2022. – №. 5 (89). – С. 3263–3277. DOI: 10.1007/s10586-023-04042-6.
9. Kumar M. et al. Machine learning models for phishing detection from TLS traffic // Cluster Computing. – 2023. – С. 1–15.
10. Tanimu J., Shiaeles S. Phishing Detection Using Machine Learning Algorithm //2022 IEEE International Conference on Cyber Security and Resilience (CSR). – IEEE, 2022. – С. 317–322. DOI: 10.1109/CSR54599.2022.9850316.
11. Shoaib M., Umar M. S. URL based phishing detection using machine learning // 2023 6th International Conference on Information Systems and Computer Networks (ISCON). IEEE, 2023. С. 1–7. DOI: 10.1109/ISCON57294.2023.10112184.
12. Uddin M. M. et al. A Comparative Analysis of Machine Learning-Based Website Phishing Detection Using URL Information // 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI). – IEEE, 2022. – С. 220–224. DOI: 10.1109/PRAI55851.2022.9904055.
13. Pujara P., Chaudhari M. B. Phishing website detection using machine learning: a review // International Journal of Scientific Research in Computer Science, Engineering and Information Technology. – 2018. – Т. 3. – №. 7. – С. 395–399.
14. Mathankar S. et al. Phishing Website Detection using Machine Learning Techniques // 2023 11th International Conference on Emerging Trends in Engineering & Technology-Signal and Information Processing (ICETET-SIP). – IEEE, 2023. – С. 1–6. DOI: 10.1109/PRAI55851.2022.9904055.
15. Omari K. Comparative study of machine learning algorithms for phishing website detection //International Journal of Advanced Computer Science and Applications. 2023. Т. 14. №. 9. DOI:10.14569/IJACSA.2023.0140945.

DETECTION OF PHISHING INTERNET DOMAINS USING MACHINE LEARNING ALGORITHMS IN REAL-TIMEDATA STREAMING

Pavlychev A. V.[6], Kuzminetc K. V.[7]

Keywords: *phishing domains, internet security, machine learning, real-time data streaming, classification algorithms, phishing detection.*

Objective: *the aim of this research is to develop an effective method for detecting phishing Internet domains using machine learning algorithms in real-time data streaming.*

Methodology: *the work involves an analysis of features characterizing arbitrary Internet domains, and the development of a software complex that allowed collecting a custom dataset containing a set of features for over 250,000 domains. Several machine learning models were)trained on the obtained dataset and compared in terms of accuracy and speed. The selected classifier was used to develop a software prototype that was tested on a sample of 1,000 arbitrary Internet domains.*

Results: *a classifier and software prototype were developed, enabling the categorization of arbitrary Internet domains as either phishing or legitimate within given accuracy and speed parameters. The validity and justification of the proposed scientific findings, results, and conclusions are supported by a comprehensive review of the current state of the field, systematic justification of the proposed models, which do not contradict known positions of other authors, and a series of experiments confirming the results of theoretical studies. The dataset collected during the work was published on the Kaggle platform for open access and use by researchers for developing various intelligent methods for detecting phishing domains.*

Scientific novelty: *the scientific novelty lies in the development of a method for detecting phishing domains in real-time data streaming, which can be used in intrusion detection systems and in the development of web applications that protect users from unwanted content. A software prototype was implemented and tested based on the obtained model, demonstrating an accuracy of 98.5% with an average processing speed of 1.2 seconds per resource.*

6 Aleksey V. Pavlychev, Director, Cybersecurity Center, Far Eastern Federal University (FEFU), Vladivostok, Russia. E-mail: pavlychev.av@dvfu.ru
7 Kirill V. Kuzminetc, Chief cybersecurity specialist, Far Eastern Federal University (FEFU), Vladivostok, Russia. E mail: kuzminetc.kv@dvfu.ru

References

1. Al-Hawawreh M., Aljuhani A., Jararweh Y. Chatgpt for cybersecurity: practical applications, challenges, and future directions //Cluster Computing. 2023. Т. 26. №. 6. S. 3421–3436. DOI: 10.1007/s10586-023-04124-5.
2. Vijayalakshmi M. et al. Web phishing detection techniques: a survey on the state-of-the-art, taxonomy and future directions // Iet Networks. – 2020. – Т. 9. – №. 5. – S. 235–246. DOI: 10.1049/iet-net.2020.0078.
3. Kalaharsha P., Mehtre B. M. Detecting Phishing Sites--An Overview // arXiv preprint arXiv:2103.12739. – 2021.
4. Basit A. et al. A comprehensive survey of AI-enabled phishing attacks detection techniques // Telecommunication Systems. – 2021. – Т. 76. – S. 139–154. DOI: 10.48550/arXiv.2103.12739.
5. Paliath S., Qbeitah M. A., Aldwairi M. Phishout: Effective phishing detection using selected features // 2020 27th International Conference on Telecommunications (ICT). IEEE, 2020. S. 1–5. DOI: 10.1109/ICT49546.2020.9239589.
6. Rashid J. et al. Phishing detection using machine learning technique // 2020 first international conference of smart systems and emerging technologies (SMARTTECH). IEEE, 2020. S. 43–46. DOI:10.1109/SMART-TECH49988.2020.00026.
7. Selishhev A. D. Razrabotka programmno-matematicheskogo obespechenija sistemy informacionnoanaliticheskogo monitoringa fishingovyh atak // Innovacionnye nauchnye issledovanija. – 2020. – №. 12–2. – S. 33–39.
8. Afanas'eva N. S., Elizarov D. A., Myznikova T. A. Klassifikacija fishingovyh atak i mery protivodejstvija im // Inzhenernyj vestnik Dona. – 2022. – №. 5 (89). – S. 3263–3277. DOI: 10.1007/s10586-023-04042-6.
9. Kumar M. et al. Machine learning models for phishing detection from TLS traffic // Cluster Computing. – 2023. – S. 1–15.
10. Tanimu J., Shiaeles S. Phishing Detection Using Machine Learning Algorithm // 2022 IEEE International Conference on Cyber Security and Resilience (CSR). – IEEE, 2022. – S. 317–322. DOI: 10.1109/CSR54599.2022.9850316.
11. Shoaib M., Umar M. S. URL based phishing detection using machine learning // 2023 6th International Conference on Information Systems and Computer Networks (ISCON). IEEE, 2023. S. 1–7. DOI: 10.1109/ISCON57294.2023.10112184.
12. Uddin M. M. et al. A Comparative Analysis of Machine Learning-Based Website Phishing Detection Using URL Information // 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI). – IEEE, 2022. – S. 220–224. DOI: 10.1109/PRAI55851.2022.9904055.
13. Pujara P., Chaudhari M. B. Phishing website detection using machine learning: a review // International Journal of Scientific Research in Computer Science, Engineering and Information Technology. – 2018. – Т. 3. – №. 7. – S. 395–399.
14. Mathankar S. et al. Phishing Website Detection using Machine Learning Techniques // 2023 11th International Conference on Emerging Trends in Engineering & Technology-Signal and Information Processing (ICETET-SIP). – IEEE, 2023. – S. 1–6. DOI: 10.1109/PRAI55851.2022.9904055.
15. Omari K. Comparative study of machine learning algorithms for phishing website detection //International Journal of Advanced Computer Science and Applications. 2023. Т. 14. №. 9. DOI:10.14569/IJACSA.2023.0140945.