Метод оценивания признаков для моделей машинного обучения в задаче идентификации поддельных веб-сайтов

М.В. Макаров, Д.А. Андреев, И.В. Антонов, В.А. Лондиков, А.И. Юдов

Псковский государственный университет

Аннотация: В статье рассматривается проблема отбора признаков при обучении моделей машинного обучения в задаче идентификации поддельных (фишинговых) веб-сайтов. В качестве решения предлагается набор из ряда ключевых метрик: эффективность, надёжность, отказоустойчивость и скорость извлечения. Подобный подход к оцениванию позволяет ранжировать признаки по категориям с последующим отбором для обучения моделей машинного обучения, в зависимости от специфики предметной области и ограничений.

Ключевые слова: метод оценивания признаков, модель машинного обучения, идентификация фишинговых веб-сайтов, метрика, эффективность, надёжность, отказоустойчивость, скорость извлечения.

Введение

Фишинг (Phishing) является одним из самых распространённых видов современных кибератак [1]. Целью фишинга является получение конфиденциальных данных пользователей, например логина и пароля или данных банковских карт, путём социальной инженерии и создания мнимой легитимности под видом известного бренда или службы [2]. Одной из разновидностей фишинга является публикация в сети «Интернет» копии веб-сайта, страницы или документа, внешне неотличимого от оригинального [3]. Когда пользователи попадают на такие веб-сайты, например, путём получения рассылки подставных писем или сообщений в социальных сетях со ссылкой на поддельный (далее – фишинговый) веб-сайт, то они доверяют увиденному и вводят свои персональные данные, не подозревая, что отправляют их злоумышленникам.

Ранее традиционные методы борьбы с фишингом включали в себя составление чёрных списков, модерируемых вручную. С тех пор злоумышленники научились обходить такие методы, ограничивая время жизни фишинговых ресурсов до нескольких часов. Из-за этого составление

чёрных списков хоть и является относительно надёжным, но уже малоэффективным средством, поскольку время реакции на атаку слишком велико, чтобы защитить от неё многих пользователей [4–6]. Современные методы борьбы обычно предлагают защиту в реальном времени при помощи применения эвристик и методов машинного обучения (Machine Learning – ML) [7,8]. К таким средствам предъявляются высокие требования: они должны давать ответ гарантированно, в короткие сроки (желательно до нескольких секунд), при этом точность должна быть как можно выше, а ложные срабатывания на легитимных веб-сайтах – отсутствовать или быть минимальными. Сами признаки должны как можно дольше оставаться актуальными для обеспечения защиты в долгосрочной перспективе. Поэтому правильный отбор эвристик или признаков для обучения моделей машинного обучения является крайне важной задачей для обеспечения достойных показателей.

В данной статье рассматривается результат оценки 82 признаков по 4 метрикам: эффективность, надежность, отказоустойчивость и скорость извлечения, а также их описание и соответствующие результаты обучения моделей машинного обучения.

Описание признаков

Признаком для модели машинного обучения будем считать численное или категориальное значение, извлечённое тем или иным образом из данных, связанных с фишинговым ресурсом. Признаки, по источнику извлечения, можно условно разделить на 3 группы, которые основаны [9]: на гиперссылках (based on Uniform Resource Locator – URL-based), на гипертексте/содержании страницы (based on HyperText Markup Language – HTML-based) и на информации из сети (based on network – network-based).

Признаки первой группы (URL-based) извлекаются из ссылки на фишинговый ресурс. Все они, так или иначе, являются количественными или

качественными характеристиками строкового значения. Например, длина доменного имени (domain name), количество дефисов и точек, наличие адреса интернет-протокола (Internet Protocol address – IP address, далее IP-адрес) и т.п. С одной стороны, это самые быстро извлекаемые признаки, с другой – их значимость может поддаваться сомнению, поскольку качественно обфусцированная ссылка может сбить с толку даже профессионала.

Отдельной проблемой является использование злоумышленниками, во-первых, т.н. «сокращателей ссылок», а во-вторых, использование публичных сервисов вместо полного хостинга (hosting) веб-сайта. В первом случае, оригинальная ссылка подменяется ссылкой на сервис, который далее осуществляет перенаправление пользователя. Другая ситуация связана с использованием сторонних сервисов, основной домен (domain) которых легитимен, но поддомен и его содержание принадлежат злоумышленнику. Подобные примеры представлены в таблице №1.

Таблица №1

Примеры легитимных и фишинговых ссылок на ресурсы

	Легитимный ресурс	Фишинговый ресурс
Собственный домен	https://www.flowable.com/	http://15110077.com/
Похожий домен	https://web.whatsapp.com/	http://www.qkc-wswhatsapp.cc/
Сокращатель ссылок	https://bit.ly/4VASV9Z	https://bit.ly/3UUJClW
Сервис	https://java-online-compiler.github.io/	https://pragatiakhare.github.io/Amazon-Clone/

Использование сокращателей ссылок, по возможности, решалось следованием всем перенаправлениям и принятию последней ссылки в цепи. Использование легитимных сервисов обрабатывалось без изменений.

Вторая группа признаков (HTML-based) связана с обработкой гипертекста (Hypertext) или объектной модели документа (Document Object

Model – DOM) содержимого рассматриваемого ресурса. Это может быть, например, наличие или отсутствие определённых ключевых слов (tags, далее – теги), атрибутов или прослушиваний событий, в частности наличие тега <*form*> (описывает форму/анкету заполнения на странице), содержащего конфиденциальные данные. Также могут рассматриваться содержимое ссылок или скачиваемых скриптов и модулей, ссылка на иконку веб-сайта и т.п. Теоретически возможна обработка естественного языка и текста на странице, однако для обеспечения универсальности и независимости от языка, такие признаки не рассматривались.

Признаки третьей группы (network-based) являются самыми долго извлекаемыми и относительно ненадёжными, однако информативными. К данной группе относятся такие сведения о веб-сайте, как время регистрации домена и сертификата уровня защищенных сокетов (Secure Socket Layer certificate – SSL certificate, далее SSL-сертификат), срок их истечения, данные о компании, на которую зарегистрирован домен, количестве IP-адресов и др. Данные, в первую очередь, извлекаются при помощи протокола WHOIS (сетевой протокол, который используется для получения регистрационных данных о владельцах доменных имен и IP-адресов) [9].

Всего в рамках исследования рассматривалось 82 признака. Из них, из первой группы извлекалось 25 признаков, из второй – 36, из третьей – 21.

Метод оценивания признаков

Каждый из приведённых признаков оценивался по 4 метрикам. Ранжирование признаков по этим метрикам, в сочетании с возможностями фильтрации, позволяет проводить быстрый первичный отбор в соответствии с требованиями предметной области.

Первая метрика – эффективность, которая измеряет «информативность» признака, его влияние на качество определения фишинга, т.е. на точность модели. Для этого использовались

нормализованные результаты алгоритма информационного выигрыша (Information Gain – IG) и алгоритма нечетких и приближенных множеств (Fuzzy Rough Set – FRS) [10]. Все показатели приводились к шкале от 0 до 1 (где 1 – максимальный вклад среди приведённых). Метрика рассчитывалась, как среднее гармоническое между двумя показателями, т.е.:

$$M_{ef}^{i} = \frac{1}{\frac{1}{\frac{IG^{max}}{IG^{i}}} + \frac{1}{\frac{FRS^{max}}{FRS^{i}}}}, \quad (1)$$

где M_{ef}^{i} – показатель эффективности признака i; IG^{max}, FRS^{max} – максимальные значения среди всех признаков по каждому из алгоритмов; IG^{i}, FRS^{i} – значения признака i для соответствующего алгоритма.

Вторая метрика – надёжность, отражающая способность признака противодействовать ложным срабатываниям модели на легитимных веб-сайтах. Значение метрики определяется разными способами, в зависимости от типа используемых данных.

Для логических типов (вида «да»/«нет», 0 или 1) используется вычисление разности распространённости противоположных значений в соответствии с формулой:

$$M_{pr}^{i,bool} = \sqrt{|leg^2 - phi^2|}, \quad (2)$$

где $M_{pr}^{i,bool}$ – показатель надёжности признака i логического типа; leg – процентное содержание положительных значений среди легитимных веб-сайтов; phi – процентное содержание положительных значений среди фишинга.

Для категориальных признаков используется похожий подход. Каждая категория разбивается на отдельный столбец по принципу бинарного кодирования (binary encoding), далее значение по каждой категории рассчитывается по формуле логического типа, после чего итоговое значение показателя равняется среднему взвешенному от всех категорий:

$$M_{pr}^{i,cat} = \sum M^{i,j} \cdot weight_j, \quad (3)$$

где $M_{pr}^{i,cat}$ – показатель надёжности признака i категориального типа; $M^{i,j}$– показатель надёжности категории по формуле (2); $weight_j$ – вес, процентное содержание категории j.

Остальные признаки являются числовыми и рассчитываются, как процент пересечения интервалов, определённых первым и третьим квантилями, рассчитанных для легитимных и фишинговых веб-сайтов.

Третья метрика – отказоустойчивость, которая представляет собой показатель надёжности признака к пропускам. Предполагается, что проблем во время самого извлечения признака возникает статистически мало, а потому на метрику влияют преимущественно внешние запросы:

$$M_{ro}^{i} = fullness \cdot (1 - p_{content}) \cdot (1 - p_{dns}) \cdot (1 - p_{API}), \quad (4)$$

где M_{ro}^{i} – показатель отказоустойчивости признака i; $fullness$ – процент валидных значений в датасете (dataset) среди признака i; $p_{content}$ – величина риска ошибки запроса по протоколу передачи гипертекста (HyperText Transfer Protocol request – HTTP request, далее – HTTP-запрос) к рассматриваемому домену (0,1 за каждый запрос); p_{dns} – величина риска ошибки протокола WHOIS или иного запроса к системе доменных имен (Domain Name System – DNS) (0,2 за каждый запрос); p_{API} – величина риска ошибки запроса к стороннему сервису (0,3 за каждый запрос).

Четвёртая метрика – скорость извлечения. Некоторые признаки могут извлекаться за доли миллисекунд (подсчёт числа точек в ссылке), в то время как другие могут требовать нескольких секунд ожидания (например, запрос по протоколу WHOIS). Ввиду большой диспропорции, метрика изменяется не линейно, а на основе логарифмической шкалы (так, значению 1 равносильно время 1 мс, значению 2 – 10 мс, значению 3 – 100 мс и т.д.):

$$M_{ex}^{i} = \log_{10}(milliseconds) + 1\ , \quad (5)$$

где M_{ex}^{i} – показатель скорости извлечения признака i; $milliseconds$ – 75-ый процентиль времени на извлечение признака среди тестовых записей.

Использование именно 75-ого процентиля связано с полученным распределением времени в период экспериментов: до 10% тестов имеют аномально высокие показатели времени работы, при этом больше 60% – находятся примерно на одном, низком, уровне. На значении 75-ого процентиля находится незначительно более завышенное значение, чем самое распространённое, для создания резерва требуемого вычислительного ресурса.

Источник данных

Для проведения исследования был подготовлен датасет из 1500 фишинговых ресурсов и 3000 легитимных.

Фишинговые ресурсы были собраны при помощи веб-сайтов www.phishtank.com и www.openphish.com, которые регулярно предоставляют ссылки на действующие веб-сайты мошенников. Ввиду скоротечности ресурсов, информация о них собиралась в течение двух недель с сохранением HTML-содержимого веб-сайта, его зависимостей, а также информации из источников DNS в момент получения ссылки. Итоговый набор данных был отфильтрован от ссылок и данных, по которым были скачаны не страницы фишинговых ресурсов, а формы об ошибках, открытых к покупке доменов.

Ссылки на легитимные ресурсы были получены с помощью открытых баз Majestic и Trunco, предоставляющих топ самых популярных веб-сайтов в интернете. Были отобраны топ-100 самых известных веб-сайтов, поскольку они чаще всего подвергаются фальсификации злоумышленниками для фишинговых веб-сайтов, а также 2900 случайных ссылок.

В момент осуществления анализа и обучения моделей машинного обучения количество используемых фишинговых и легитимных ресурсов выравнивалось, тем самым соотношение было 1:1.

Результаты оценивания признаков

В таблице №2 приведён список всех рассмотренных 82 признаков, включая их описания, значения по всем метрикам, а также дополнительное значение корреляции признака с целевым для оценки распространённости и влияния признака в определении фишинговых или легитимных ресурсов.

В каждой колонке метрики жирным шрифтом выделены лучшие 10 значений признаков. При равенстве значений выделялись те, у чьих признаков выше показатель первой метрики («эффективность»), поскольку такой признак, как ожидается, будет демонстрировать более высокую точность.

Таблица №2

Признаки и их значения по метрикам

№	**Описание признака**	M_{ef}^{i}	M_{pr}^{i}	M_{ro}^{i}	M_{ex}^{i}	K^{1}
1	Длина ссылки (в символах)	**0,35**	0,86	**1**	**0**	0,42
2	Длина домена	0,22	0,67	**1**	**0**	0,41
3	Кол-во уровней доменов	0,01	0	1	0	0,12
4	Кол-во точек (.) в ссылке	0,03	0	1	0	0,19
5	Кол-во дефисов (-) в ссылке	0,13	**1**	**1**	**0**	0,36
6	Кол-во косых черт (/) в ссылке	0,11	**1**	**1**	**0**	0,35
7	Кол-во точек (.) в домене	0,01	0	1	0	0,12
8	Кол-во дефисов (-) в домене	0,11	**1**	1	0	0,36
9	Кол-во цифр в домене	0,12	**1**	**1**	**0**	0,34
10	Кол-во точек (.) в пути ссылки	0,05	**1**	1	0	0,23
11	Кол-во дефисов (-) в пути ссылки	0,07	**1**	1	0	0,27
12	Кол-во наборов случайных символов[2]	0,14	0	**1**	**0**	0,35
13	Кол-во наборов случайных символов в пути ссылки	0,14	0	**1**	**0**	0,41
14	Указание IP-адреса в ссылке	0	0	1	0	-
15	Указание порта в ссылке	0	0	1	0	-
16	Указание секции авторизации в ссылке	0	0	1	0	-

№	Описание признака	M_{ef}^{i}	M_{pr}^{i}	M_{ro}^{i}	M_{ex}^{i}	K^{1}
17	Использование защищённого протокола передачи гипертекста	0,26	0,83	1	0	-0,53
18	Указание расширения файла в пути	0	0	1	0	-
19	Указание фрагмента/якоря (#) в ссылке	0	0	1	0	-
20	Указание поддомена www.	0,14	0,50	**1**	**0**	-0,43
21	Домен первого уровня (.ru, .com, т.д.)	**0,35**	0,28	**1**	**0**	0,03
22	Домен первого уровня есть в белом списке[3]	0,09	0,61	1	0	-0,35
23	Домен первого уровня есть в чёрном списке[4]	0	0	1	0	0,41
24	Домен первого уровня в списке (-1, 0, 1)[5]	0,09	0,58	1	0	-0,35
25	Расширение файла (.html, .php, т.д.)	0	0	1	0	0,41
26	Кол-во символов страницы	**0,99**	**1**	0,90	2,67	-0,46
27	Отключена индексация (robots.txt)	0	0,16	0,80	3,63	0,07
28	Иконка находится на другом домене	0	0	0,90	2,66	-
29	Имеется тег <base>	0	0	0,90	0	-
30	Кол-во мета-тегов <meta>	0,24	0	0,90	2,67	-0,27
31	Кол-во ссылок <link>	0,22	0	0,90	2,67	-0,32
32	Кол-во скриптов <script>	0,20	0	0,90	2,67	-0,32
33	Процент внешних связей <link>	0,26	0	0,90	2,64	0,11
34	Процент внутренних связей <link>	0,27	0	0,90	2,67	-0,02
35	Процент встроенных[6] скриптов	0,27	0	0,90	2,67	-0,02
36	Процент связей на внешние скрипты	0,20	**1**	0,90	2,62	-0,06
37	Процент связей на внутренние скрипты	0,27	0	0,90	2,60	-0,05
38	Кол-во упомянутых доменов в связях	0,06	0	0,90	2,65	-0,23
39	Процент внутренних связей от всех	0,27	0	0,90	2,67	-0,02
40	Процент самой частой внешней связи	0,27	0	0,90	2,60	0,28
41	Имеется тег <noscript>	0,07	0,32	0,90	0,00	-0,30
42	В тег <noscript> вложено изображение	0	0	0,90	2,60	-
43	Имеется вложенный веб-сайт <iframe>	0,03	0	0,90	0,00	-0,22
44	Имеется форм <form>	0,02	0,29	0,90	2,60	-0,17
45	Форма не имеет защищённого протокола передачи гипертекста	0	0	0,90	2,67	-
46	«Тело» веб-сайта пустое	0	0	0,90	2,66	-
47	«Тело» веб-сайта имеет до 5 тегов	0,06	0,25	0,90	2,67	0,27
48	Имеется атрибут style у любого тега	0	0,14	0,90	2,67	-0,02
49	Имеется атрибут nonce у любого тега	0	0	0,90	2,61	-
50	Есть активное прослушивание событий	0	0,12	0,90	2,59	0,04

№	**Описание признака**	M_{ef}^{i}	M_{pr}^{i}	M_{ro}^{i}	M_{ex}^{i}	K^{1}
51	Есть отключение прослушивания событий	0	0	0,90	2,61	-
52	Отключено событие mouseover	0	0	0,90	2,59	-
53	Отключено событие contextmenu	0	0	0,90	2,57	-
54	Процент внешних ссылок <a>	**0,30**	**1**	0,90	2,67	-0,04
55	Процент внутренних ссылок <a>	**0,37**	0	0,90	2,67	-0,39
56	Процент пустых ссылок <a>	0,28	0	0,90	2,70	0,21
57	Кол-во упомянутых доменов в ссылках	0,15	**1**	0,90	2,63	-0,32
58	Процент ссылок на внутренний домен	**0,37**	0	0,90	2,67	-0,39
59	Процент ссылок на самый частный внешний домен	**0,34**	0	0,90	2,62	0,11
60	Кол-во прослушиваний событий	0	1	0,90	2,61	0,05
61	Кол-во отключенных событий	0	1	0,90	2,65	0,07
62	Кол-во IP-адресов	0,01	0	0,80	3,75	0,01
63	Кол-во IP-адресов версии 4	0	0	0,42	3,77	-
64	Кол-во IP-адресов версии 6	0	0	0,42	3,66	-
65	Кол-во используемых серверов имён	0,09	0	0,80	3,96	-0,14
66	Кол-во записей в серверах имён	0,07	0	0,80	3,90	0,06
67	Кол-во статусов, определяемых корпорацией по управлению доменными именами и IP-адресами	0,07	0	0,80	3,99	0,07
68	Кол-во перенаправлений по ссылке	0,10	**1**	0,46	3,63	0,34
69	Время регистрации веб-сайта	**1**	0,74	0,44	3,95	0,52
70	Время истечения срока действия веб-сайта	**1**	0	0,44	3,95	-0,10
71	Информация о веб-сайте имеет электронную почту	0,03	0,50	0,80	3,94	-0,20
72	Информация о веб-сайте имеет название организации	0,01	0,31	0,80	3,93	-0,10
73	Имеется хоть одно перенаправление	0,10	0,27	0,45	3,42	0,35
74	Имеется перенаправление со сменой домена (cross-origin)	0	0	0,45	3,42	-
75	Страна регистрации домена	0	0	0,56	4,73	-
76	Срок существования домена в днях	**0,83**	0,75	0,80	3,93	-0,55
77	Домену менее 30 дней	0,01	0,19	0,80	3,95	0,14
78	Домену менее 7 дней	0,01	0,14	0,80	3,93	0,10
79	Срок регистрации SSL-сертификата	0,13	0,22	0,80	3,04	-0,03
80	Версия сертификата SSL-сертификата	0	0	0,80	3,04	-
81	Используемый алгоритм SSL	0,02	0,26	0,80	3,04	0,12
82	В SSL применён алгоритм цифровой	0	0,24	0,80	3,04	-0,08

№	Описание признака	M_{ef}^{i}	M_{pr}^{i}	M_{ro}^{i}	M_{ex}^{i}	K^{1}
	подписи по эллиптической кривой					

[1]*значение коэффициента корреляции Пирсона для числовых признаков и категориальной кросс-энтропии для категориальных признаков, корреляция относительно целевого признака (чем ближе корреляция к 1, тем более прямая зависимость признака с фишинговыми ресурсами, и наоборот, чем ближе к -1, тем больше признак связан с легитимными веб-сайтами), прочерк означает, что во всех тестовых записях оказалось ровно одно одинаковое значение, и корреляцию определить нельзя.*

[2]*В данной работе, случайным набором символов называется последовательность из 5 или более букв и цифр, которая НЕ состоит из комбинаций: СГС, СГ, ГС, СС, ГГ, где С – согласная, Г – гласная, и все буквы строчные, кроме первой.*

[3]*en, ru, fr, uk, de, com, gov, edu – вручную отобранные популярные домены.*

[4]*me, br, cl – вручную отобранные домены, часто используемые фишингом.*

[5]*равен -1, если домен в чёрном списке, 1 – домен в белом списке, и 0 – в остальных случаях.*

[6]*имеется в виду тег <script>, код которого не импортируется через src, а приведён непосредственно в элементе внутри страницы веб-сайта.*

Как можно наблюдать из метрики эффективности, более востребованными являются числовые признаки, однако и некоторые категориальные признаки могут оказывать значительное влияние (домен первого уровня, использование защищённого протокола передачи гипертекста (HyperText Transfer Protocol Secure – HTTPS, далее – протокол HTTPS)). Самые высокие значения связаны со временем существования домена, что обусловлено созданием фишинга на короткий срок. Также довольно высокую значимость имеют признаки, связанные с анализом ссылок в HTML-коде страницы, что можно объяснить склонностью

фишинговых ресурсов к имитации только одной страницы и «стиранием» ссылки на все остальные.

Метрика надёжности демонстрирует в среднем более низкую востребованность признаков второй группы (HTML-based). Фишинг выделяет заметно более высокое, как видно из корреляции, содержание специальных символов (дефисы, точки), а также более длинные ссылки. Наличие поддомена www и протокола HTTPS, наоборот, чаще означает легитимность веб-сайта. Интересным наблюдением является высокое влияние количества символов в HTML-коде – фишинговые ресурсы объёмнее, чем легитимные.

Метрика отказоустойчивости однозначно закрепляет признаки первой группы (URL-based) как наиболее надёжные, что логично. Признаки второй группы (HTML-based) так же обладают высокой отказоустойчивостью, лишь незначительно уступая первой. Признаки третьей группы (network-based) извлекаются с наибольшим риском неполучения данных.

Метрика скорости извлечения во многом коррелирует с третьей метрикой. Признаки первой группы извлекаются почти мгновенно, признаки второй группы – значительно медленнее. Сложнее всего ситуация обстоит с признаками третьей группы, поскольку они подразумевают различные запросы к сторонним сервисам с долгим ожиданием.

При этом можно наблюдать, что 19 признаков вовсе не имеют корректных оценок по метрикам, что можно обнаружить по прочеркам в колонке корреляции. Это означает, что во всём датасете эти признаки имели одинаковое значение для всех записей. В некоторых случаях это являлось ожидаемым результатом (например, для количества IP-адресов, использования редкого тега <base> (позволяет изменить гиперссылку по умолчанию при скачивании дополнительных ресурсов веб-сайтом)), однако в других ситуациях подобное демонстрирует изменение тенденции.

Стоит отметить, что злоумышленники постоянно меняют свои алгоритмы и подходы к созданию фишинговых ресурсов для усложнения их идентификации автоматическими системами. Так, например, признак наличия IP-адреса ещё несколько лет назад присутствовал и служил хорошим показателем фишинга [7,8], однако на данный момент, среди 1500 актуальных фишинговых адресов не было найдено ни одного, использующего IP-адрес вместо домена. Аналогичным примером является признак скачивания логотипа веб-сайта с внешнего (часто легитимного) домена. Поэтому в данной статье рассматривается намного больше признаков, чем это минимально необходимо, поскольку в будущем они могут оказаться востребованными.

Результаты обучения моделей машинного обучения

После выполнения анализа было обучено суммарно 6 моделей полносвязной нейронной сети (Fully-Connected Neural Network – FCNN), для всех использовалась одна и та же конфигурация гиперпараметров.

Первая модель обучалась на всех 82 признаках без исключений, показатели точности которой служат в качестве «точка отсчёта» для определения результативности метрик.

Следующие четыре модели обучались на лучших признаках по каждой из 4 метрик соответственно (при равенстве результатов брались метрики с наивысшей метрикой эффективности), при этом количество признаков могло отличаться путём отбора лучшей модели среди всех, которые обучались с включением от 1 до 30 лучших признаков.

Последняя модель также была построена по принципу уменьшения признакового пространства посредством метода рекурсивного исключения признаков (Recursive Feature Elimination – RFE), однако признаки сортировались по алгоритму IG. Результаты данной модели приведены для

сравнения, поскольку RFE и IG являются частой конфигурацией при фильтрации признаков [8].

Для оценки точности моделей использовались следующий набор показателей:

- точность – процент верных ответов от общего количества тестов;
- кучность – отношение верно определённых фишинговых веб-сайтов, которые приняты моделью за фишинг, ко всем веб-сайтам (чем ближе к 1, тем меньше легитимных веб-сайтов было ошибочно идентифицировано как фишинг);
- полнота – отношение верно определённых фишинговых веб-сайтов ко всем реальным фишинговым веб-сайтам (чем ближе к 1, тем меньше фишинговых веб-сайтов было идентифицировано в качестве легитимных);
- кривая ошибок – метрика оценки качества бинарной классификации (binary classification) (принимает значения от 0 до 1, где 1 – модель справляется идеально, 0 – справедливо обратное).
- Среднее время ожидания ответа – время между началом запроса на проверку веб-сайта на фишинг и моментом получения ответа моделью.

Результаты обучения всех моделей приведены в таблице №3.

На основе полученных результатов стоит отметить, что использование метрики эффективности позволяет увеличить точность модели, в частности на 1,3% по сравнению с моделью, использующей все признаки для обучения, и на 3,5% – с моделью с частью признаков, отобранных по RFE и IG. Модель, которая обучена на признаках, отобранных по метрике надёжности, имеет более высокий показатель кучности, что означает меньшее количество ложных срабатываний на легитимных веб-сайтах, при этом прирост кучности составляет 0,029 и 0,028 по сравнению с первой и последней моделями соответственно.

Таблица №3

Результаты обучения моделей машинного обучения по показателям

Модель	Точность, %	Кучность	Полнота	Кривая ошибок	Среднее время ожидания ответа, мс
Все признаки	96,6	0,958	0,974	0,993	9011
Метрика эффективности	97,9	0,974	0,987	0,997	918
Метрика надёжности	97,2	0,987	0,961	0,994	2753
Метрика отказоустойчивости	95,1	1	0,909	0,991	56
Метрика скорости извлечения	91,6	0,945	0,896	0,979	47
Часть признаков (RFE, IG)	94,4	0,959	0,935	0,995	2587

По метрике отказоустойчивости была получена модель, обученная преимущественно на признаках первой и второй групп (URL-based и HTML-based), что позволило добиться максимальной кучности за счёт некоторого снижения точности. Наиболее быстрой моделью оказалась модель, которая обучена по признакам, отобранным по метрике скорости извлечения, однако это привело к значительному уменьшению точности данной модели.

Заключение

Данное исследование продемонстрировало перспективность предложенного метода оценивания признаков для моделей машинного обучения в задаче идентификации фишинговых веб-сайтов. Разработанные метрики (эффективность, надёжность, отказоустойчивость и скорость извлечения) позволяют не просто ранжировать признаки, но и целенаправленно формировать вектор признаков в зависимости от конкретных требований к системе защиты, в число которых могут входить максимальная точность, минимальное время отклика или устойчивость к

отказам внешних сервисов. Несмотря на то, что метрики не дают однозначных рекомендаций по наилучшему набору признаков, тем не менее, они предоставляют более качественный результат, чем отбор стандартными методами. Подобный системный подход открывает путь к созданию более адаптивных и эффективных моделей машинного обучения.

Также представляется прогрессивным направление по разработке самоактуализирующихся систем, способных динамически перестраивать свой набор признаков на основе актуальных метрик. В условиях, когда злоумышленники постоянно меняют тактику, такие системы могли бы автоматически отслеживать снижение информативности одних признаков и подключать другие, более релевантные, признаки для обеспечения непрерывной и своевременной защиты без постоянного вмешательства человека. Практическая реализация вышеуказанных механизмов позволит значительно повысить устойчивость антифишинговых решений к эволюции кибератак.

Проведённое исследование дополнительно актуализирует знания о востребованных признаках для борьбы с фишинговыми атаками. Представленные результаты могут послужить отправной точкой и справочными материалами для других исследователей и разработчиков, которые ведут работу по созданию или усовершенствованию собственных алгоритмов обнаружения фишинга.

Исследование выполнено в рамках реализации гранта ФГБОУ ВО «Псковский государственный университет» по мероприятию «Выполнение прикладных научных и научно-методических исследований (проектов) молодыми учеными по приоритетным направлениям Программы развития университета на 2025-2036 годы» (приказ № 0905-001 от 05.09.2025).

Литература

1. Крюкова И. В., Алимамедов Э. Н. Фишинг как вид интернет-мошенничества // Наукосфера. 2021. № 2 (2). С. 196–201.

2. Афанасьева Н. С., Елизаров Д. А., Мызникова Т. А. Классификация фишинговых атак и меры противодействия им // Инженерный вестник Дона, 2022, № 5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7641/.

3. Chiew K. L., Yong K. S. C., Tan C. L. A survey of phishing attacks: Their types, vectors and technical approaches // Expert Systems With Applications. 2018. Vol. 106. P. 1–20.

4. Селиверстов В. В., Корчагин С. А. Анализ актуальности и состояния современных фишинг-атак на объекты критической информационной инфраструктуры // Инженерный вестник Дона, 2024, № 6. URL: ivdon.ru/ru/magazine/archive/n6y2024/9277/.

5. Sheng S., Holbrook M., Kumaraguru P., Cranor L., Downs J. Who Falls for Phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Atlanta: ACM, 2010. P. 373–382.

6. Skula I., Kvet M. Domain Blacklist Efficacy for Phishing Web-page Detection Over an Extended Time Period // Proceedings of the 33rd Conference of Open Innovations Association FRUCT. Zilina: IEEE, 2023. P. 257–263.

7. Zhang Y., Hong J., Cranor L. CANTINA: A Content-Based Approach to Detecting Phishing Web Sites // International World Wide Web Conference Committee. Banff: ACM, 2007. P. 639–648.

8. Mourtaji Y., Pr. Bouhorma M., Pr. Alghazzawi D. Perception of a New Framework for Detecting Phishing Web Pages // Proceedings of the Mediterranean Symposium on Smart City Application. Tangier: Springer Cham, 2017. Article 11. P. 1–6.

9. Tian Y., Yu Y., Sun J., Wang Y. From Past to Present: A Survey of Malicious URL Detection Techniques, Datasets and Code Repositories // Computer Science Review. 2025. Vol. 58. Article 100810. P. 1–37.

10. Zabihimayvan M., Doran D. Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection // IEEE International Conference on Fuzzy Systems. New Orleans: IEEE, 2019. P. 1–6.

References

1. Krjukova I. V., Alimamedov Je. N. Naukosfera, 2021, No 2 (2), pp. 196–201.

2. Afanas'eva N. S., Elizarov D. A., Myznikova T. A. Inzhenernyj vestnik Dona, 2022, No. 5. URL: ivdon.ru/ru/magazine/archive/n5y2022/7641/.

3. Chiew K. L., Yong K. S. C., Tan C. L. Expert Systems With Applications, 2018, Vol. 106, pp. 1–20.

4. Seliverstov V. V., Korchagin S. A. Inzhenernyj vestnik Dona, 2024, No. 6. URL: ivdon.ru/ru/magazine/archive/n6y2024/9277/.

5. Sheng S., Holbrook M., Kumaraguru P., Cranor L., Downs J. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Atlanta: ACM, 2010, pp. 373–382.

6. Skula I., Kvet M. Proceedings of the 33rd Conference of Open Innovations Association FRUCT. Zilina: IEEE, 2023, pp. 257–263.

7. Zhang Y., Hong J., Cranor L. International World Wide Web Conference Committee. Banff: ACM, 2007, pp. 639–648.

8. Mourtaji Y., Pr. Bouhorma M., Pr. Alghazzawi D. Proceedings of the Mediterranean Symposium on Smart City Application. Tangier: Springer Cham, 2017, Article 11, pp. 1–6.

9. Tian Y., Yu Y., Sun J., Wang Y. Computer Science Review, 2025, Vol. 58, Article 100810, pp. 1–37.

10. Zabihimayvan M., Doran D. IEEE International Conference on Fuzzy Systems. New Orleans: IEEE, 2019, pp. 1–6.

Авторы согласны на обработку и хранение персональных данных.

Дата поступления: 8.12.2025

Дата публикации: 24.01.2026