Санкт-Петербургский государственный университет

Бородкин Вячеслав Павлович
Выпускная квалификационная работа

Прогнозирование возраста пользователя в социальной сети: ансамбль классификаторов

Уровень образования: бакалавриат
Направление 02.03.03 «Математическое обеспечение и администрирование информационных систем»
Основная образовательная программа СВ.5006.2017 «Математическое обеспечение и администрирование информационных систем»

Научный руководитель:
доцент кафедры информатики, к.т.н. М. В. Абрамов

Рецензент:
Ассистент факультета информационных технологий
и программирования Университета ИТМО,
к.т.н. С. Б. Муравьев

Санкт-Петербург
2021

SAINT-PETERSBURG STATE UNIVERSITY

Borodkin Vyacheslav Pavlovich
Graduate qualification work

Predicting the age of a user in social media: an ensemble of classifiers

Education level: bachelor's degree
Direction 02.03.03 «Software and Administration of Information Systems Software Engineering»
Basic educational program CB.5006.2017 «Software and Administration of Information Systems Software Engineering»

Scientific supervisor:
Associate Professor,
PhD in Engineering Sciences, Maksim Abramov

Reviewer:
Assistant of the Faculty of Information Technology and Programming, ITMO University,
PhD in Engineering Sciences, Sergey Muravyov

Saint-Petersburg
2021

Оглавление

Введение

Актуальность темы. В настоящее время все больше людей пользуются социальными сетями: по данным на 2021 год, в мире насчитывается 4.2 миллиарда пользователей социальных сетей, что составляет более 50% всего населения Земли [15]. Из отчета Федерального Бюро Расследований за 2020 год видно, что с увеличением числа пользователей, растет и количество зарегистрированных киберпреступлений [3]. Как следствие, увеличиваются материальные потери: убытки жертв таких преступлений на 2020 год составили более 4 миллиардов долларов. Для сравнения, в 2019 эта сумма составила 3.5 миллиарда [3].

Вопрос безопасности в мире информационных технологий выходит на первый план. Для защиты своих клиентов, IT-компании увеличивают бюджет, предотвращение кибератак [20]. Кроме того, страны повышают цифровую грамотность своего населения. Например, в Российской Федерации на специальной платформе проводится акция "Цифровой Диктант" (`https://digitaldictation.ru/`). Таким образом, данное направление становится одним из самых актуальных. Хотя вопросу безопасности уделяется большое внимание, количество преступлений продолжает расти. Можно предположить, что это связано не столько с безуспешностью борьбы с кибератаками, сколько о развитии данной отрасли криминального мира.

По данным Verizon за 2020 год, 22% всех атак содержали элементы социальной инженерии [14]. Из отчета компании видно, что в большинстве всех зафиксированных преступлений использовалось два вида нападений: фишинг и претекстинг. Процентное соотношение данных методов составляет примерно 90% к 10% соответственно. Представленная статистика показывает, что при кибератаках на информационные системы, уязвимости, связанные с человеческим фактором, часто используются злоумышленниками.

Успешность подобных методов сильно зависит от информации о целевом пользователе, которой владеет злоумышленник. Например, данные о возрасте человека. Так, интересы, а значит и сферы давления

на слабые места детей и взрослых сильно отличаются [11]. Исследования компании Google показывают, что подростков больше интересует развлекательный контент, а людей более старшего возраста интересуют видео развивающего и прикладного характера. К примеру, люди возрастной категории 25-29 лет смотрят обучающие видео чаще, чем люди 13-17 лет [21]. Таким образом, можно предположить, что детей и подростков скорее заинтересует реклама какой-нибудь бесплатной игры, чем образовательные курсы. Опираясь на вышесказанное, можно допустить, что знание возраста потенциальной жертвы может помочь злоумышленникам.

Для предсказания возможных вариантов атак и их упреждения необходимо понимать, к какой возрастной категории относится пользователь. Эту информацию можно получить из аккаунтов людей в социальных сетях. Усложняется выявление возраста тем, что далеко не каждый заполняет поле «Возраст» или «Дата рождения» на своей странице в различных социальных сетях. Таким образом, вопрос определения возраста по данным, полученным со страницы пользователя, является актуальной областью исследования для предотвращения социоинженерных атак.

Целью данной работы является улучшение существующих (повышение точности и применимости) и разработка новых подходов восстановления возраста пользователей социальной сети для опосредованного повышения вероятности предотвращения социоинженерных атак. В том числе, подходов, использующих методы ансамблирования.

Для достижения этой цели были поставлены следующие **задачи:**

- Произвести идентификацию социальных сетей, которые будут наилучшим образом соответствовать критериям (популярность в русскоязычном сегменте интернет; отсутствуют значительные ограничения на работу с данными, такие как, например, невозможность их получения ранее реализованными способами)

- Провести сравнительный анализ существующих методов и подходов, для выявления применимых к выбранным социальным сетям

- Адаптировать существующие методы для возможности их применения на выбранных социальных сетях и провести сравнение их работы на собранных со страниц пользователей данных

- Синтезировать новые методы восстановления возраста для опосредованного повышения вероятности предотвращения социоинженерных атак

- Провести сопоставление разработанных методов с адаптированными для выявления преимуществ и недостатков

- Внедрить разработку в существующий комплекс для анализа пользователей социальных сетей

Объект исследования: аккаунты пользователей в социальных сетях.

Предмет исследования: методы восстановления возраста пользователей в социальных сетях.

Научная новизна заключается в том, что предложены новые способы агрегации данных и методы восстановления пропущенных возрастов пользователей.

Теоретическая и практическая значимость исследования. Теоретическая значимость заключается в создании новых подходов восстановления возрастов, которые позволят уточнить получаемые оценки для пользователей в социальной сети «ВКонтакте» (`https://vk.com`).

Практическая значимость заключается в интеграции полученных подходов в существующий комплекс программ для оценки защищенности пользователей информационных систем от социоинженерных атак. Это позволит повысить устойчивость к кибератакам, содержащим элементы социальной инженерии.

Структура и объем работы. Данная работа состоит из введения, - глав, заключения, списка используемой литературы и словаря терминов. Общий объем работы — 38 страницы.

В главе 1 проводится описание предметной области, обоснование актуальности целей и задач.

В главе 2 описываются существующие подходы к восстановлению пропущенных атрибутов пользователей.

В главе 3 описывается теоретическая составляющая проводимых расчетов.

В главе 4 представлено описание реализации рассмотренных методов.

1. Описание предметной области

В этой главе описывается проблематика увеличения количества социоинженерных атак. Проводится анализ социальных сетей. Приводятся описания поставленной цели и задач, которые необходимо выполнить для ее достижения. Идентифицируется социальная сеть для дальнейшей работы.

1.1. Восстановление атрибутов пользователей

Количество кибератак, содержащих элементы социальной инженерии, увеличивается с каждым годом [14]. Для их предотвращения необходимо знать как можно более подробную информацию о пользователях, по отношению к которым применяются эти атаки. Некоторые социальные сети позволяют не указывать или же скрывать полную информацию о пользователе. Так, например, социальная сеть «ВКонтакте» допускает скрытие возраста от других пользователей. Другим примером является сеть «Одноклассники» (`https://ok.ru`). Для регистрации в ней не требуется никакой личной информации, кроме телефона, имени и возраста. При этом, достоверность последних двух не проверяется. Для того, чтобы получить информацию, не предоставляемую самим пользователем, можно использовать агрегацию других доступных данных и с их помощью проводить дальнейшее восстановление пропущенных атрибутов пользователей [17, 7, 4, 23].

В настоящее время существует множество работ, связанных с восстановлением пропущенных атрибутов пользователей в социальных сетях [17, 7, 4, 23, 18, 9]. Однако, среди существующих обзоров довольно большая часть методов предназначена для использования в рамках социальных сетей, не столь популярных в России [22]. А именно «Twitter» (`https://twitter.com`) и «Facebook» (`https://facebook.com`). Часть методов [4, 23, 18] затрагивает сети, более распространенные в русскоязычном сегменте, но данные подходы не имеют большой применимости при в силу недостаточной точности результатов.

Улучшение существующих и реализация новых способов восстанов-

ления возраста, применимых к популярным в России социальным сетям и дальнейшее внедрение в существующий комплекс для анализа защищенности пользователей позволит уточнять возрастные оценки пользователей, что в свою очередь может помочь в выработке персональных рекомендаций по защите от социоинженерных атак.

1.2. Обоснование целей и задач работы

Основываясь на обосновании актуальности предотвращения социоинженерных атак, видится важным исследование методов восстановления возрастов пользователей. Зная предполагаемый возраст человека, можно будет предположить, какой именно социоинженерной атаке он подвержен. Предположительно, это поможет снизить количество преступлений, совершаемых с помощью социальных сетей.

Целью данной работы является улучшение существующих (повышение точности и применимости) и разработка новых подходов восстановления возраста пользователей социальной сети для опосредованного повышения вероятности предотвращения социоинженерных атак. В том числе, подходов, использующих методы ансамблирования.

Задачи для достижения поставленной цели:

- Произвести идентификацию социальных сетей, которые будут наилучшим образом соответствовать критериям (популярность в русскоязычном сегменте интернет; отсутствуют значительные ограничения на работу с данными, такие как, например, невозможность их получения ранее реализованными способами)
- Провести сравнительный анализ существующих методов и подходов, для выявления применимых к выбранным социальным сетям
- Адаптировать существующие методы для возможности их применения на выбранных социальных сетях и провести сравнение их работы на собранных со страниц пользователей данных
- Синтезировать новые методы восстановления возраста для опо-

средованного повышения вероятности предотвращения социоинженерных атак

- Провести сопоставление разработанных методов с адаптированными для выявления преимуществ и недостатков
- Внедрить разработку в существующий комплекс для анализа пользователей социальных сетей

1.3. Выбор социальной сети

Для выбора и адаптации алгоритмов восстановления возраста пользователей, необходимо было выбрать социальную сеть. Изначальный список был взят из опроса, проведенного компанией Mediascope в 2019 году [22]. В нем были выделены самые популярные сети России. Среди них «ВКонтакте», «Instgram» (`https://instagram.com`), «Одноклассники», «Facebook». Социальная сеть «ВКонтакте» предоставляет интерфейс VK Api (`https://vk.com/dev/methods`) для получения данных со страниц пользователей и сообществ. С помощью существующих методов, для отдельных пользователей можно получить информацию с их страниц, при условии, что сам пользователь указал ее, а его страница является публичной. Для сообществ можно получить списки их участников. Уже реализованный способ сбора данных сильно упрощает работу с социальной сетью. Исходя из того, что первое место в вышеуказанном списке занимает сеть «ВКонтакте», имея при этом удобный интерфейс для получения сведений о пользователях, было принято решение остановить выбор именно на этой социальной сети.

1.4. Вывод по главе

В данной главе приведено обоснование актуальности исследования методов восстановления атрибутов пользователей социальных сетей. Представлено обоснование цели и поставленных задач. Проведен анализ популярных в Российской Федерации социальных сетей и выбрана одна конкретная («ВКонтакте») для дальнейшей работы.

2. Используемые подходы

В этой главе приводится формальное описание задачи. Проводится анализ некоторых из существующих подходов к агрегации данных и методов для восстановления атрибутов пользователя.

2.1. Формализация задачи

Математически задачу восстановления возраста пользователя можно интерпретировать следующим образом: каждому пользователю социальной сети соответствует набор параметров X. Необходимо построить алгоритм $a : X_1 \to Y$, где $X_1 \subset X$ - подмножество параметров, $Y \in N$ - предполагаемый возраст. В качестве параметров могут выступать возрасты друзей пользователя, его интересы, подписки и т. д. Рассмотрим некоторые из возможных подходов восстановления пропущенных атрибутов пользователей в социальных сетях.

2.2. Вычисление моды атрибутов окружения пользователя

Первый метод основан на вычислении моды атрибута у окружения пользователя. Мода - значение на множестве, принимаемое наибольшее число раз. Если несколько значений встречаются во множестве одинаковое количество раз и при этом, никакое другое значение не принимается чаще - множество называется мультимодальным. Визуально определить моду можно, построив график зависимости частоты вхождения элемента в выборку от его значения. Значение элемента, встречающееся наибольшее количество раз и будет модой. Так, на рисунке 1 представлен график для множества значений A, где $A = \{2, 1, 4, 2, 5, 5, 3, 2, 3\}$.

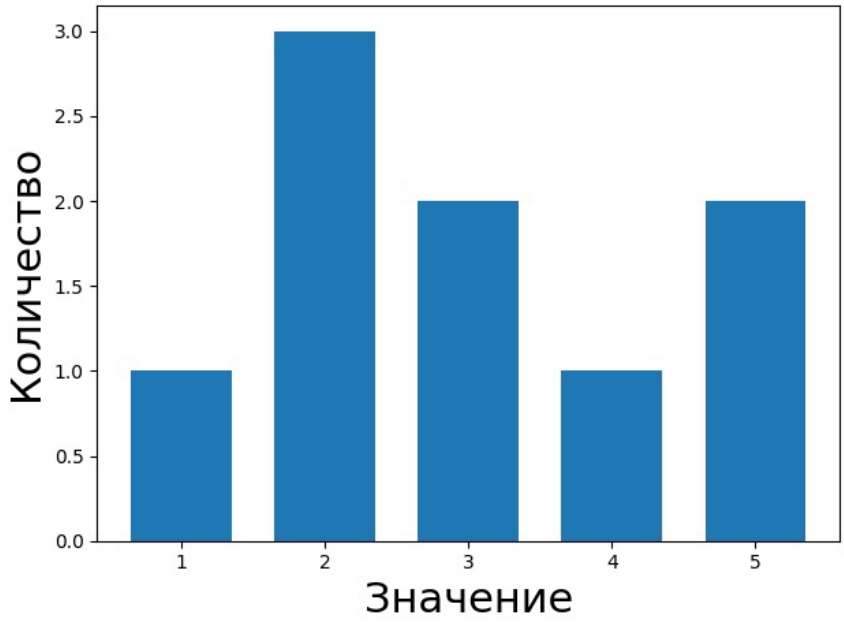

Рис. 1: График зависимости количества вхождений от значения

Одной из работ, в которой рассматривается данный метод, написана Трофимовичем Ю.С., Козловым И.С. и Турдаковым Д.Ю. [23]. В конкретном случае алгоритм применялся для восстановления города проживания пользователя. Однако, его можно использовать и для восстановления возраста. Идея алгоритма заключается в следующем:

1. Для интересующего нас пользователя строится социальный граф. Его глубина определяется заранее

2. В узлах этого графа указываются известные значения атрибута, который необходимо восстановить (рис. 2)

3. Если в узле неизвестен необходимый атрибут, то при наличии связей с другими узлами, значение пропущенного атрибута выставляется равным моде узлов, связанных с рассматриваемым (рис. 3)

4. Шаг 3 повторяется до тех пор, пока количество пропущенных атрибутов не станет равным нулю или не перестанет изменяться (рис. 4)

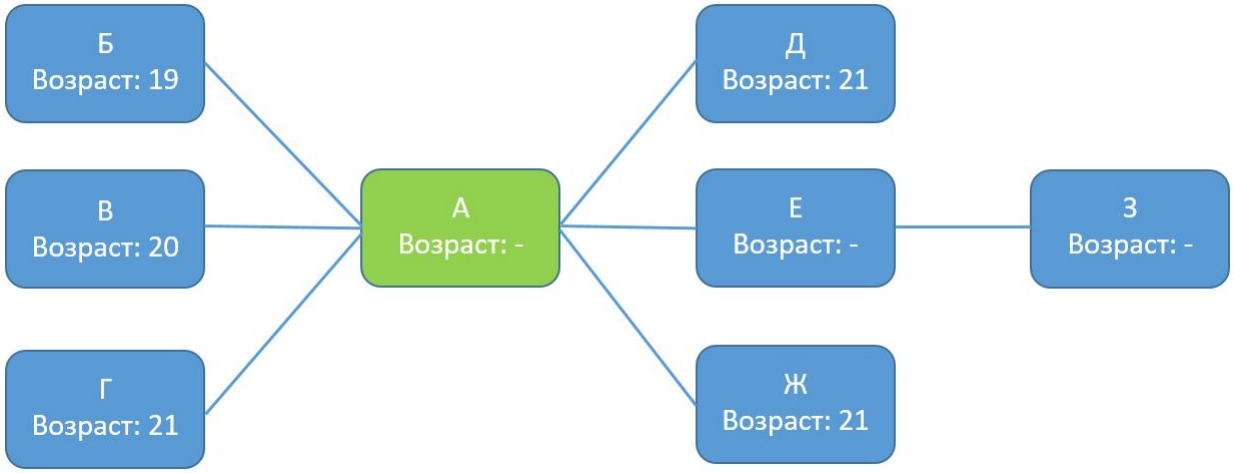

Рис. 2: Пример изначального соц. графа.

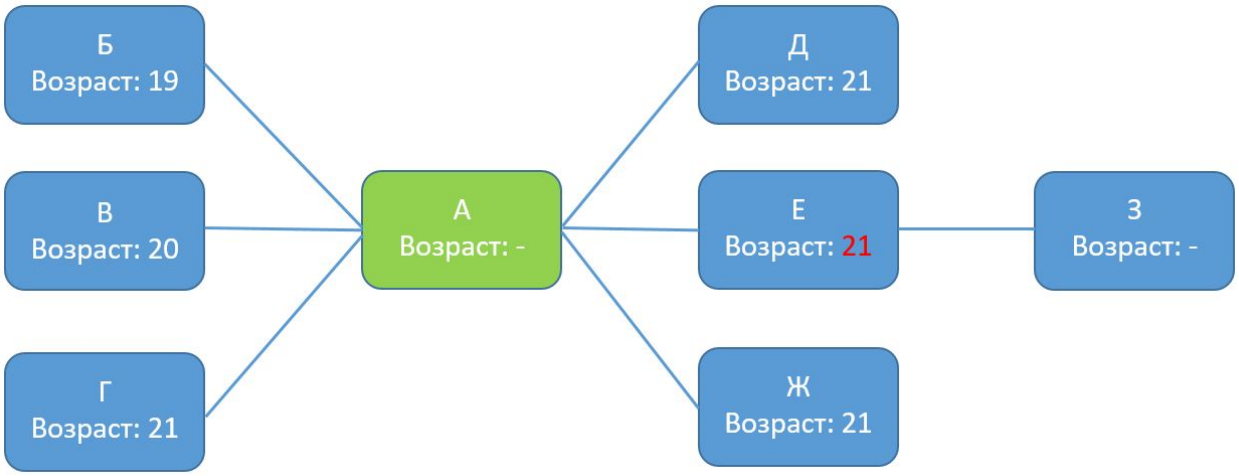

Рис. 3: Пример шага алгоритма.

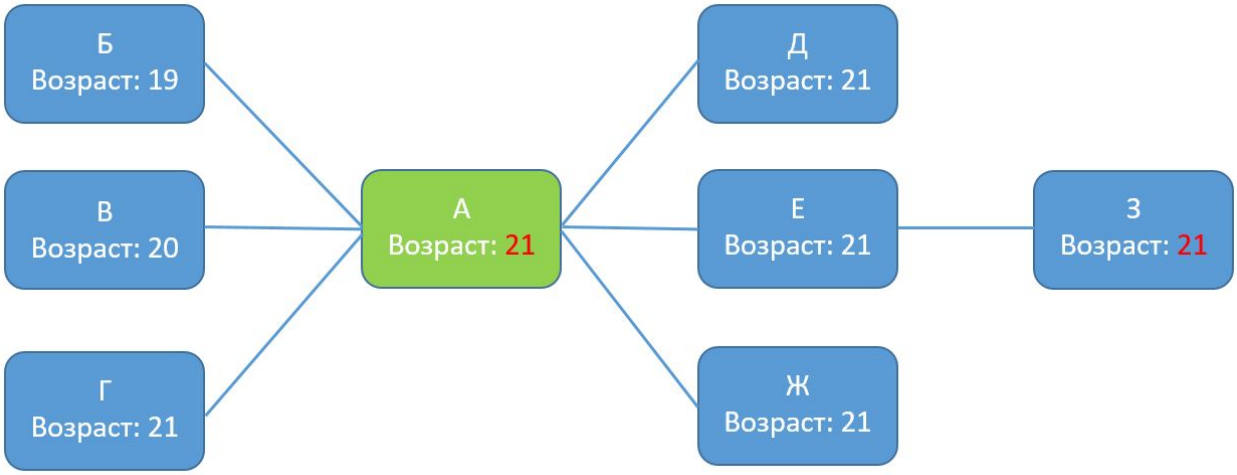

Рис. 4: Пример получения конечного результата.

Данный метод применим для социальных графов, где у каждого узла множество связей с известными значениями. Заметим, что при маленьком количестве связей увеличивается вероятность мультимодальности. Решением этой проблемы является взятие среднего значения мод.

2.3. Кластеризация

Идея следующего алгоритма состоит в том, что люди, объединенные по какому-либо признаку, будут иметь схожие атрибуты. К примеру, в приведенной статье [17] был создан классификатор, восстанавливавший значения пропущенных полей студентов двух университетов из заданного набора возможных значений, например, номер общежития, в котором проживал студент или же направление его обучения. Кластеризация проводилась, основываясь на годе поступления в университет, направлении обучения и номере общежития, в котором проживали рассматриваемые студенты. Эта же идея лежит в основе статьи [13]. Ее авторы, используя идею кластеризации, попытались предсказать объемы продаж в магазинах, имея в качестве начальных данных информацию о продажах прошлых месяцев. Алгоритм выставления значения интересующего атрибута, используя объединение в кластеры, можно представить следующим образом:

1. Выбор атрибутов, по которым будет происходить кластеризация. При этом они не должны совпадать с атрибутами, которые необходимо восстановить

2. Разделение множества на группы, опираясь на выбранные признаки

3. Выявление зависимостей между атрибутами и принадлежности конкретной группе

4. Восстановление атрибутов, учитывая общую тенденцию в группе

Опираясь на идею кластеризации, можно создать классификатор, основанный на делении пользователей на группы. Признаки, по которым будет происходить кластеризация, необходимо выбирать, основываясь на особенностях конкретной социальной сети.

2.4. Метод средних значений сообществ пользователя

Этот способ восстановления возраста был придуман в рамках данной работы, основываясь на вышеуказанных подходах. Метод средних значений сообществ пользователя заключается в восстановлении возраста, опираясь на средний возраст сообществ, в которых состоит пользователь. На первом шаге данного метода необходимо подготовить начальные данные. Под начальными данными подразумеваются:

- Список сообществ, в которых состоит пользователь
- Количество участников в каждом из представленных сообществ
- Список возрастов участников каждого сообщества

Первый шаг состоит из сбора необходимых для работы алгоритма данных. Для ускорения подсчетов было принято решение принимать во внимание только те сообщества, количество участников которых не превышает 500 человек. После получения сообществ рассматриваемого пользователя, определялась принадлежность каждого из них классу, зависящему от количества участников. Номер класса соответствовал целой части от деления аудитории сообщества на 100 (1).

$$Class = [\frac{PeopleInGroup}{100}] \tag{1}$$

Затем вычислялась мода возрастов каждой группы, в которой состоит рассматриваемый пользователь. Таким образом, после проделанных выше расчетов, человеку приписывалась таблица следующего вида: После этого рассчитывались средние значения возрастов по каждому из

Сообщество	Класс	Значение моды
Group_1	Class_1	Value_1
Group_2	Class_2	Value_2
Group_3	Class_3	Value_3
...	...	...

Таблица 1: Сообщества пользователя

классов групп. На выходе пользователю сопоставлялся кортеж из 5 значений. К получившимся значениям применялся классификатор, основанный на дереве решений. При обучении дерева использовалась кроссвалидация, а также был выполнен перебор таких параметров как:

1. Глубина дерева
2. Максимальное количество экземпляров в листе
3. Минимальное количество экземпляров в листе

Перебор и выбор лучших параметров производился посредством библиотеки Scikit-learn.

2.5. Ансамбль классификаторов

Существует несколько вариантов формирования ансамблей алгоритмов [12, 10, 2]. Один из них - голосование классификаторов. Ключевая идея такого способа заключается в том, что при наличии нескольких результатов, правильным считается тот, который получило большее количество классификаторов. Так как предсказанные значения возраста пользователя несколькими методами могут не пересекаться, а, следовательно, ансамбль, основанный на голосовании не будет применим. Вторым способом использования ансамбля является использование результатов нескольких классификаторов, основываясь на точности их по отдельности. Подбор весов было решено осуществить с помощью линейной регрессии, во избежание ручного подбора. Последний рассмотренный метод похож на предыдущий, но его идея заключается не в подборе весов, а обучении нового классификатора, основываясь

на результатах предыдущих. В рамках этой работы будет рассмотрен ансамбль, основывающийся на алгоритме случайного леса.

2.6. Вывод по главе

В данной главе приведено формальное описание поставленной задачи. Представлены методы, основанные на вычислении моды значений окружения пользователя, кластеризации. Описана собственная модель восстановления возраста, использующая возрасты людей, состоящих в сообществах, участником которых является рассматриваемый пользователь. Предложен метод ансамблирования отдельных методов восстановления возраста. Приведены статьи, в которых использовались уже применяющиеся методы. Описан общий вид входных данных для каждого алгоритма. Таким образом был проведен сравнительный анализ существующих подходов к восстановлению атрибутов пользователей и приведены собственные методы.

3. Теоретическая часть

В этой главе будет описана идея способа кросс-валидации для обучения моделей, рассмотрены метрики, опираясь на которые будет происходить сравнение качества реализуемых классификаторов.

3.1. Кросс-валидация

Одним из способов обучения моделей на маленьких наборах данных является кросс-валидация [1]. Основная идея обучения с использованием кросс-валидации заключается в следующем: пусть есть набор данных X. Его можно разделить на X_{train} и X_{test}, где X_{train} - выборка, на которой будет происходить обучение моделей и перебор параметров, а X_{train} - финальная выборка, на которой будет проверяться конечная модель (рис. 5). Затем, тестовая выборка делится еще раз, на заранее

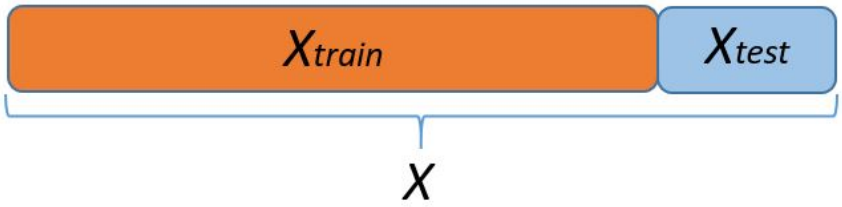

Рис. 5: Разбиение всех данных.

выбранное количество по возможности равных частей (рис. 6). Также на вход подаются параметры модели, которые необходимо перебрать. Далее, создаются модели со всеми возможными вариациями перебираемых параметров. Каждая модель теперь будет обучаться k раз, где k – количество частей, на которые было разбито тестовое множество. При этом, в качестве обучающего набора данных будут предоставлены только $k-1$ частей всего обучающего множества. На рисунке они отмечены оранжевым цветом. Последняя же, синяя часть, будет выступать в качество тестового набора для каждой конкретной модели с уникальным набором параметров. Затем, результаты каждой такой модели, полученные с k разных наборов данных, усредняются. Лучшей моделью считается та, которая даст наивысшие средние результаты.

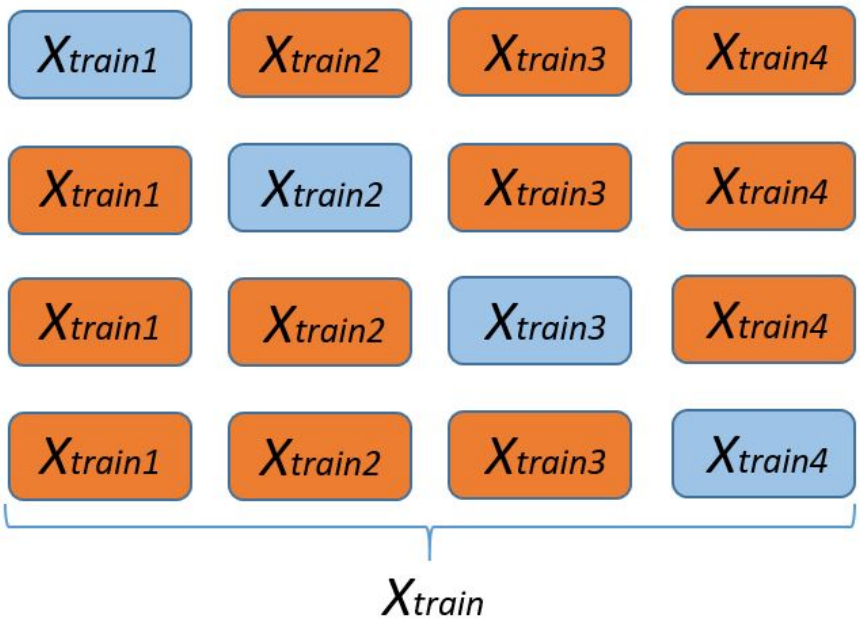

Рис. 6: Разбиение тестового множества.

Для небольших наборов данных кросс-валидация помогает избежать переобучения. Другими словами, это ситуация, когда классификатор вместо общей закономерности выявил частные случаи и использует их для предсказаний.

3.2. Метрики оценки качества методов

Для определения точности методов рассчитывались коэффициент детерминации, среднеквадратическая ошибка и средняя абсолютная ошибка [6]. Для первого формула выглядит следующим образом:

$$R^2 = 1 - \frac{D[y|x]}{D[y]} = 1 - \frac{\sigma^2}{{\sigma_y}^2},$$

где R^2 - коэффициент детерминации, ${\sigma_y}^2$ - дисперсия случайной величины, а σ^2 - условная дисперсия зависимой переменной. Среднеквадратическая ошибка высчитывается так:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2,$$

где MSE - среднеквадратическая ошибка, Y_i - правильные значе-

ния возрастов, а $\hat{Y_i}$ - предполагаемые значения. Средняя абсолютная ошибка:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i}),$$

где MAE - средняя абсолютная ошибка, Y_i - правильные значения возрастов, а $\hat{Y_i}$ - предполагаемые значения. Данные метрики служат для определения качества регрессионных моделей, поэтому они не использовались в ряде вышеперечисленных работ, а именно тех, где речь идет о восстановлении географических атрибутов, например, города проживания пользователя [4, 23]. Заметим, что коэффициент детерминации указывает на зависимость между двумя наборами данных. Коэффициент 1 означает функциональную зависимость и служит эталоном для классификаторов, 0 же означает отсутствие связи между наборами. Значения среднеквадратической и средний абсолютной ошибок наоборот, должны стремиться к 0, так как отображают разницу попарно взятых значений двух наборов данных. С помощью средней абсолютной ошибки можно понять, насколько в среднем ошибается классификатор, однако, минусом этой метрики является то, что отрицательные ошибки предсказаний компенсируются положительными. Среднеквадратическая ошибка же лишена этого минуса. Также, с помощью этой метрики можно идентифицировать большие ошибки, так как разница между предсказанным и реальным значением возводится в квадрат [19]. Данные метрики часто используются в работах с регрессионными моделями для определения их точности[16, 5, 8].

3.3. Вывод по главе

В данной главе описан принцип кросс-валидации, используемый в дальнейшей разработке, обосновано его применение. Приведены метрики, которые будут использоваться для сравнительного анализа реализуемых моделей ($MSE,\ MAE,\ R^2$), обоснован их выбор.

4. Реализация

В настоящей главе перечисляются инструменты, используемые для реализации моделей. Описывается процесс сбора данных пользователей. Приводятся реализации методов восстановления возрастов пользователей социальной сети «ВКонтакте». Проводится сравнительный анализ реализованных моделей.

4.1. Используемые программные инструменты

Для реализации рассмотренных методов был выбран язык программирования Python 3.8. Причиной такого выбора является специализация языка на работе с данными, большой выбор библиотек, направленных на их агрегацию и дальнейшую работу с классификаторами. Также, важным было наличие удобной библиотеки для работы с VK Api 11.9.1.

Список средств, используемых библиотек:

- Matplotlib (`https://matplotlib.org`) Для построения графиков
- Библиотеки Pandas(`https://pandas.pydata.org`) и NumPy (`https://numpy.org`) для обработки данных
- Scikit-learn (`https://scikit-learn.org`) для обучения классификаторов
- Библиотека VK Api — для получения данных из социальных сетей

4.2. Сбор данных

В данной главе рассмотрены реализации методов, основываясь на специфике выбранной социальной сети.

Метод мод реализовывался на уже имеющемся наборе данных. Он состоял из 247247 записей о пользователях сети «ВКонтакте». Каждая запись содержала следующие поля:

- Имя

- Фамилия
- Ссылка на профиль
- Дата рождения
- Город проживания
- Родной город
- Закрыт ли аккаунт (доступна ли информация о нем)
- Пол пользователя

Профили людей с закрытыми аккаунтами были исключены из дальнейшего рассмотрения не представляли интереса для исследований, по причине отсутствия информации. Поэтому база данных сократилась до 220135 записей. Далее, были убраны все пользователи, информация о дате рождения которых была неизвестна. Таким образом список людей стал включать данные о 79762 пользователях (рис. 7). Конечная таблица содержала следующие столбцы:

- Имя
- Фамилия
- Ссылка на профиль
- Идентификатор пользователя
- Дата рождения
- Возраст

На получившемся наборе уже и производились дальнейшие исследования.

Для метода мод необходимо было собрать данные о возрастах друзей пользователей. Таким образом, количество запросов, которые необходимо было сделать, чтобы получить значение моды возрастов друзей

	firstName	lastName	link	birthday	id	age
0	[illegible]	[illegible]	[illegible]	24.1.1998	[illegible]	22
1	[illegible]	[illegible]	[illegible]	20.8.1978	[illegible]	42
2	[illegible]	[illegible]	[illegible]	30.4.1998	[illegible]	22
3	[illegible]	[illegible]	[illegible]	30.12.1985	[illegible]	35
4	[illegible]	[illegible]	[illegible]	19.8.1996	[illegible]	24
...	...	...	...	...	...	...
79757	[illegible]	[illegible]	[illegible]	15.2.1976	[illegible]	44
79758	[illegible]	[illegible]	[illegible]	15.6.2002	[illegible]	18
79759	[illegible]	[illegible]	[illegible]	8.11.2003	[illegible]	17
79760	[illegible]	[illegible]	[illegible]	8.11.2005	[illegible]	15
79761	[illegible]	[illegible]	[illegible]	23.3.2004	[illegible]	16

79762 rows × 6 columns

Рис. 7: Данные после начальной обработки.

одного пользователя равно $1+N$, где N - количество друзей. Первый же запрос необходим для получения списка друзей. Для метода средних значений возрастов групп пользователя, чтобы получить необходимые данные для одного человека, необходимо было сделать $1 + N + \sum_{i=1}^{N}(K_i)$ запросов, где N − количество групп, в которых состоит пользователь, K_i − количество участников в $i-$ой группе. Первый запрос необходим для получения списка сообществ пользователя, следующие N, чтобы получить списки пользователей, состоящих в каждой из групп, последние $\sum_{i=1}^{N}(K_i)-$ для получения возрастов участников всех групп. Так как VK Api имеет ограничение на 3 запроса в секунду, отправленных от имени одного человека, весь сбор данных происходил в 7 потоков с аккаунтов разных людей.

4.3. Реализация метода «Мод»

Каждому человеку из итогового набора данных ставился в соответствие список его друзей, полученный посредствам VK Api. Из каждого такого списка исключались люди, информацию о дате рождения которых нельзя было получить из-за настроек приватности. Затем, в каждом получившемся списке высчитывалась мода возрастов. Получившееся значение присваивалось рассматриваемому пользователю (рис. 8). Для улучшения результатов, последний шаг присваивания выполнялся

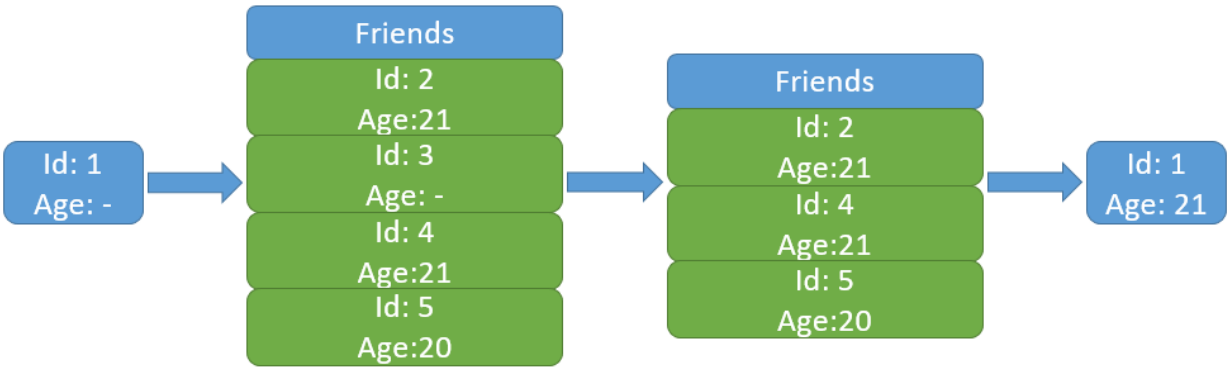

Рис. 8: Реализация метода мод.

только в том случае, если хотя бы у 5 друзей рассматриваемого пользователя возраст совпадал с рассчитанным значением моды. Если же множество возрастов друзей являлось мультимодальным - рассчитывалось среднее значение мод.

4.4. Реализация метода кластеризации

Основываясь на алгоритме, представленном выше, было принято решение производить деление на группы, основываясь на сообществах, находящихся в подписках у рассматриваемого пользователя. Сервис «ВКонтакте» предлагает классификацию сообществ на следующие тематики:

- Новости
- Спорт

- Музыка
- Блогеры
- Радио и телевидение
- Развлечения
- Игры и киберспорт
- Наука и технологии
- Мода и красота
- Культура и искусство
- Благотворительность
- Бренды

Идея изменения алгоритма, основываясь на особенностях выбранной социальной сети, заключается в выборе тематик, чаще всего появляющихся в списке сообществ пользователя. Затем, используя метод дерева решений сделать вывод о возрасте человека. Но в процессе реализации возникла проблема, не позволившая в дальнейшем развить данный алгоритм. В большинстве случаев администраторы сообществ выставляют настройки приватности таким образом, что тематику группы нельзя узнать с помощью VK Api. Классификация же сообществ вручную являлась бы субъективной, что могло повлиять на дальнейший анализ данных. Поэтому было принято решение отказаться от данного способа.

4.5. Реализация метода средних значений сообществ пользователя

Так как для данного метода, чтобы получить итоговое значение для одного человека - необходимо было сделать большое количество запросов:

- Получить список сообществ
- Для каждой из групп получить список пользователей, состоящих в ней
- Для каждого пользователя получить его возраст

было принято решение отобрать тех пользователей, которые состоят не более чем в 50 сообществах, чтобы ускорить сбор данных. Также, одной из проблем являлось то, что администраторы многих сообществ выставляют настройки приватности таким образом, что становится невозможно получить список пользователей через запросы средствами VK Api. Учитывая вышесказанное, был собран датасет, содержащий в себе записи о 4114 людях и средних значениях возрастов участников сообществ, в которых они состоят. Учитывая структуру записи данных, описанных в главе 2, датасет принял следующий вид:

	name	class0	class1	class2	class3	class4	real_age
0	[illegible]	34	0	0	24	28	23
1	[illegible]	30	27	0	0	0	25
2	[illegible]	30	34	0	35	34	25
3	[illegible]	31	30	0	0	0	31
4	[illegible]	0	0	0	0	35	37
...	...	...	...	...	...	...	...

Рис. 9: Пример записей для метода средних значений сообществ.

Затем, эти данные были разделены на две выборки: тестовую и обучающую соответственно. Процент записей, отправленных в тестовую выборку составил 30%. Таким образом, результаты данного метода проверялись на 1235 записях. На тестовой выборке был обучен классификатор, используя кросс-валидацию и подбор параметров дерева, дающих лучшие результаты. Таблица параметров представлена ниже:

Параметр	Диапазон	Лучшее значение
Максимальная глубина	1-10	5
Минимальное количество экземпляров в листе	1-10	1

Таблица 2: Параметры дерева решений

Для последующего вычисления метрик предсказания модели и реальный возраст пользователя были записаны в отдельную таблицу, представленную на рисунке 10, где $real\_age$ – возраст, указанный на странице, $groups\_method$ – возраст, предсказанный моделью.

	real_age	groups_method
764	27	27
2402	35	22
396	49	33
2325	37	33
2203	39	27
...	...	...
1042	31	36
2085	44	41
18	33	32
2041	42	32
966	28	33

Рис. 10: Предсказание возраста методом средних значений сообществ.

4.6. Реализация ансамбля

Для реализации ансамбля нужны данные, к которым применимы все методы, которые будут объединены. Таким образом, если при построении ансамбля будут использованы методы мод и средних значений групп, необходимо было отобрать людей, списки друзей и групп

которых возможно получить, используя VK Api. При этом, условия, накладываемые на количество сообществ и участников в них, описанные в реализации метода средних значений сообществ. Чтобы получить такое записи, было решено применить метод мод к людям из тестовой выборки, полученной ранее. На выходе был получен датасет, содержащий в себе идентификатор человека, его возраст и два значения возраста, полученные разными методами.

	real_age	friends_method	groups_method
1	25	25	27
4	37	36	33
9	31	25	27
10	24	27	32
11	34	33	36
...	...	...	...
2741	38	32	33
2742	33	31	32
2744	41	32	41
2749	32	36	33
2750	32	32	33

Рис. 11: Таблица предсказания возрастов пользователей двумя методами.

На рисунке 11 представлена таблица, значения колонок которой отображают следующее:

- $real\_age$ - настоящий возраст пользователя
- $friends\_method$ - значение возраста, полученное, методом мод возрастов пользователей
- $groups\_method$ - значение возраста, полученное, методом средних значений возрастов сообществ пользователя

Далее необходимо было подобрать релевантные веса, для значений, полученных каждым путем. Для этого датасет так же был разделен на

тестовую и обучающую выборку. Создано два классификатора, один из которых был основан на линейной регрессии, а второй на методе случайного леса. При обучении классификатора, основанного на методе случайного леса, использовался перебор параметров и выявление тех, которые дадут лучшие результаты. Перебираемыми параметрами являлись:

Параметр	Диапазон	Лучшее значение
Количество деревьев	10-50	20
Максимальная глубина	1-10	5
Минимальное количество экземпляров в листе	1-10	4
Минимальное количество элементов, достаточное для разделения листа	2-20	18

Таблица 3: Параметры случайного леса

Выбранные значения параметров указаны в таблице.

4.7. Сравнение результатов

На итоговое сравнение выставлены 4 следующих способа определение возраста пользователя в социальной сети:

- Метод средних возрастов сообществ
- Метод мод
- Ансамбль, основанный на линейной регрессии
- Ансамбль, основанный на случайном лесе

Так как среднеквадратическая и средняя абсолютная ошибка сильно зависят от исходных данных, необходимо проводить замеры на наборах одинакового размера. Поэтому, конечные расчеты и сравнение результатов были проведены для всех реализованных методов на одних и тех

же данных: над тестовой выборкой ансамбля, а затем над всем датасетом, к которому применимы вышеуказанные методы. Метрики рассчитывались с помощью встроенных средств библиотеки Scikit-learn. Для тестового набора данных результаты получились следующими:

	MSE	$R2$	MAE
Groups method	143.05	0.10	8.05
Friends method	227.09	0.09	10.61
Linear regression	**137.36**	**0.14**	8.10
Forest	152.66	0.04	**7.78**

Таблица 4: Метрики для тестового объема данных

Ниже приведена таблица результатов для всего объема данных:

	MSE	$R2$	MAE
Groups method	81.49	0.11	5.88
Friends method	81.59	0.11	**5.18**
Linear regression	73.86	0.19	6.12
Forest	**72.44**	**0.21**	**5.18**

Таблица 5: Метрики для всего объема данных

Основываясь на полученных данных, можно сделать вывод, что использование ансамблей улучшает качество предсказания возраста пользователя. Можно заметить, что минимальные значения среднеквадратической и средней абсолютной ошибки принадлежит модели, основанной на случайном лесе. Ей же принадлежит максимальное из представленных значений коэффициента детерминации. Это говорит о том, что использование данного ансамбля улучшает предсказания, отдельных моделей, взятых для обучения ансамбля. Общие неточности могут быть обусловлены заведомо ложными введенными пользователями данными

(социальная сеть «ВКонтакте» не проверяет пользователей на достоверность предоставляемой информации). Однако, уже существующие методы, реализованные и примененные к тем же данным, на которых проверялись новые алгоритмы показали результаты хуже. Это говорит о возможности применения предложенных алгоритмов восстановления возрастов пользователей в реальных системах. В данный момент ведётся тестирование работы синтезированных методов восстановления возраста в существующем комплексе для анализа пользователей социальных сетей (`sea.dscs.pro`).

4.8. Вывод по главе

В данной главе перечислены инструменты, используемые для реализации моделей. Описан процесс сбора данных пользователей. Приведены реализации методов восстановления возрастов пользователей социальной сети «ВКонтакте». Проведен сравинтельный анализ реализованных моделей. Лучшие результаты были показаны методами ансамблирования. Описан дальнейший этап разработки - внедрение в существующий комплекс для анализа пользователей социальных сетей.

Заключение

В рамках дипломной работы были реализованы существующие методы восстановления возраста пользователей. Предложено два новых способа восстановления возраста пользователей, позволившие повысить точность классификаторов.

Для достижения представленных результатов была поставлена следующая **цель**: улучшить существующие (повысить точность и применимость) и разработать новые подходов восстановления возраста пользователей социальной сети для опосредованного повышения вероятности предотвращения социоинженерных атак, в том числе, подходов, использующих методы ансамблирования. Решены следующие **задачи**:

- Произведена идентификация социальных сетей. Выбрана сеть, лучше всего удовлетворяющая рассматриваемым критериям. Данные, полученные со страниц её пользователей использовались для дальнейшей работы
- Проведен анализ существующих методов и подходов
- Рассмотренные методы адаптированы под конкретную социальную сеть и проведено сравнение их точности на тестовых данных
- Синтезированы новые методы восстановления возраста пользователей социальных сетей
- Проведено сопоставление разработанных методов с адаптированными
- Осуществленно внедрение синтезированных методов в существующий комплекс для анализа пользователей социальных сетей

Таким образом был выполнен ряд поставленных задач, способствующих достижению цели выпускной квалификационной работы. Реализованные методы помогут повысить точность восстановления возраста

пользователей социальной сети «ВКонтакте», что в свою очередь позволит улучшить определение типов кибератак, которым подвержены различные пользователи.

Список литературы

[1] Boehm U. Matzke D. Gretton-M. Castro S.-Cooper J. Skinner-M. Strayer D. Heathcote A. Real-time prediction of short-timescale fluctuations in cognitive workload // Cognitive Research: Principles and Implications Volume 6, Issue 1, December 2021, Article 30, doi:10.1186/s41235-021-00289-y.

[2] Derbentsev V. Babenko V. Khrustalev-K. Obruch H.-Khrustalova S. Comparative Performance of Machine Learning Ensemble Algorithms for Forecasting Cryptocurrency Prices// International Journal of Engineering Volume: 34, Issue: 1, Pages: 140-148, doi:10.5829/ije.2021.34.01a.164.

[3] Internet crime report 2020. — URL: `https://www.ic3.gov/Media/PDF/AnnualReport/2020_IC3Report.pdf` (дата обращения: 26.04.2021).

[4] Kaveeva A.D. Gurin K.E. Vkontakte' local friendship networks: Identifying the missed residence of users in profile data// Monitoring Obshchestvennogo Mneniya: Ekonomicheskie i Sotsial'nye Peremeny Volume 145, Issue 3, May-June 2018, Pages 78-90, doi:10.14515/monitoring.2018.3.05.

[5] Meng T. Huang R. Lu-Y. Liu H.-Ren J. Zhao-G. Hu W. Highly sensitive terahertz non-destructive testing technology for stone relics deterioration prediction using SVM-based machine learning models // Heritage Science Volume 9, Issue 1, December 2021, Artcicle 24, doi:10.1186/s40494-021-00502-7.

[6] Metrics and scoring: quantifying the quality of predictions. — URL: `https://scikit-learn.org/stable/modules/model_evaluation.html` (дата обращения: 28.11.2020).

[7] Mulders D. de Bodt C.-Bjelland J. Pentland-A. Verleysen M.-de Montjoye YA. Inference of node attributes from social network

assortativity// Neural Computing Applications Volume: 32, Issue: 24, Pages: 18023-18043, doi:10.1007/s00521-018-03967-z.

[8] Müller P.L. Odainic A. Treis-T. Herrmann P.-Tufail A. Holz-F.G. Pfau M. Inferred retinal sensitivity in recessive Stargardt disease using machine learning // Scientific Reports Volume 11, Issue 1, December 2021, Artcicle 1466, doi:10.1038/s41598-020-80766-4.

[9] Nie L. Zhang L. Wang M. Hong R.-Farseev A. Chua T. Learning user attributes via mobile social multimedia analytics // ACM Transactions on Intelligent Systems and Technology Volume 9, Issue 1, December 2021, Artcicle 24, doi:10.1145/2963105.

[10] Pirizadeh M. Alemohammad N. Manthouri M. Pirizadeh M. A new machine learning ensemble model for class imbalance problem of screening enhanced oil recovery methods// Journal of Petroleum Science and Engineering Volume 198, Article 108214, doi:10.1016/j.petrol.2020.108214.

[11] Samuel C. McQuade James P. Colt Nancy B. Meyer. Cyber Bullying: Protecting Kids and Adults from Online Bullies. — 2020.

[12] Seryasat OR. Kor I. Zadeh HG. Taleghani AS. Predicting the number of comments on Facebook posts using an ensemble regression model// International Journal of Nonlinear Analysis and Applications Volume: 12, Issue: 24, Pages: 49, doi:10.22075/IJNAA.2021.4796.

[13] Tirta H. Perdana N.J. Mulyawan B. Sparepart sales clusterization and prediction using automatic clustering algorithm // IOP Conference Series: Materials Science and Engineering, Volume 1007, Issue 1, Article 012191, doi:10.1088/1757-899X/1007/1/012191. — 2020.

[14] Verizon Data Breach Investigations Report. — 2020. — URL: `https://enterprise.verizon.com/resources/reports/2020-data-breach-investigations-report.pdf` (дата обращения: 28.11.2020).

[15] Web Canape. — 2021. — URL: https://www.web-canape.ru/business/vsya-statistika-interneta-i-socsetej-na-2021-god-cifry- (дата обращения: 26.04.2021).

[16] Wei C.-N. Wang L.-Y. Chang X.-Y. Zhou Q.-H. A prediction model using machine-learning algorithm for assessing intrathecal hyperbaric bupivacaine dose during cesarean section // BMC Anesthesiology Volume 21, Issue 1, December 2021, Artcicle 116, doi:10.1186/s12871-021-01331-8.

[17] You Are Who You Know: Inferring User Profiles in Online Social Networks// WSDM 2010 - Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, doi:10.1145/1718487.1718519 2010, Pages 251-260.

[18] Гомзин А.Г. Кузнецов С. Д. Метод автоматического определения возраста пользователей с помощью социальных связей //Труды института системного прогаммирования РАН Том: 28, номер: 6, страницы: 171-184 doi:10.15514/ISPRAS-2016-28(6)-12. — 2016.

[19] Как правильно выбрать метрику оценки для моделей машинного обучения: часть 1 Регрессионные метрики. — URL: https://www.machinelearningmastery.ru/how-to-select-the-right-evaluation-metric-for-machine-learning-m (дата обращения: 28.11.2020).

[20] Кибербезопасность 2019-2020. — URL: https://www.ptsecurity.com/ru-ru/research/analytics/cybersecurity-2019-2020/ (дата обращения: 28.11.2020).

[21] Новое поколение интернет-пользователей: исследование привычек и поведения российской молодежи онлайн. — 2017. — URL: https://www.thinkwithgoogle.com/intl/ru-ru/consumer-insights/consumer-trends/novoe-pokolenie-internet-polzovatelei-issledovanie-privychek-i-p (дата обращения: 26.04.2021).

[22] Социальные сети в цифрах. — 2019. — URL: https://mediascope.net/upload/iblock/f97/18.04.2019_Mediascope_Екатерина%20Курносова_РИФ+КИБ%202019.pdf (дата обращения: 28.11.2020).

[23] Трофимович Ю.С. Козлов И.С. Турдаков Д.Ю. Подходы к определению основного места проживания пользователей социальных сетей на основе социального графа. — URL: https://cyberleninka.ru/article/n/podhody-k-opredeleniyu-osnovnogo-mesta-prozhivaniya-polzovatelej (online; accessed: 28.11.2020).

Приложение А: список терминов

Социоинженерная атака - разновидность кибератаки, при которой воздействие происходит не на машину, а на человека.

Фишинг - способ интернет-мошенничества, при котором злоумышленник пытается получить логин и пароль пользователя, путем рассылки писем от имени популярных сервисов и компаний.

Претекстинг - способ мошенничества, при котором целью злоумышленника является получение информации о банковской карте жертвы с помощью звонков и смс-сообщений.

Социальный граф - граф, узлами которого являются аккаунты пользователей в социальных сетях, а ребрами — связи между этими пользователями (например, дружба).

Ансамблирование - техника в машинном обучении, основанная на объединении нескольких моделей для повышения точности.

Датасет - структурированный набор данных.