А.Г. Дьяконов[1], А.М. Головина[2]

ПОИСК ЗАКОНОМЕРНОСТЕЙ И ВАЖНОСТИ ПРИЗНАКОВ В ДАННЫХ ВИКТИМИЗАЦИОННОГО ОПРОСА

Введение

Сегодня методы машинного обучения и анализа данных проникли практически во все сферы деятельности человека и позволяют строить прогнозы, описывать закономерности в данных, проводить оптимизацию бизнес-процессов. При обработке результатов социологических опросов они используются не так часто. Видимо, это связано с тем, что подобные опросы «заточены» под конкретную цель, например, подтвердить или опровергнуть определённую гипотезу, поэтому традиционно используется аппарат математической статистики (для оценки достоверности гипотезы, представительности выборки и т.п.).

В данной работе сделан обзор подходов для нахождения нетривиальных зависимостей в данных, а также методов анализа важности признаков. Подходы проверены на реальных данных соцопроса, что позволило не только найти закономерности в этих данных, но и выявить недостатки опроса, а также проиллюстрировать ограниченность некоторых методов (особенно при некорректном использовании).

В машинном обучении с учителем (Supervised Learning) [1] исходная информация задаётся, как правило, матрицей $X = ||x_{ij}|| \in \mathbb{R}^{m\times n}$ и вектором $y = (y_1, \dots, y_m) \in \mathbb{R}^m$. В матрице представлены признаковые описания объектов $x_i = (x_{i1}, \dots, x_{in}) \in \mathbb{R}^n$ (есть также большой класс задач, в котором объекты изначально не заданы признаками), $i \in \{1,2, \dots, m\}$, m – число объектов, n – число признаков, а в векторе y – т.н. «целевые значения» или значения целевого признака объектов. Множество $\{(x_i, y_i)\}_{i=1}^m$ называется обучающей выборкой. Задача обучения с учителем заключается в построении алгоритма, который осуществляет отображение $a: \mathbb{R}^n \to \mathbb{R}$ и в идеале обладает условием корректности [2]

$$a(x_i) = y_i,\ i \in \{1,2, \dots, m\}, \qquad (1)$$

хотя часто возможность точного выполнения (1) связана с избыточной сложностью модели алгоритмов – параметрического множества $\{a_w\}$, в

[1] Москва, МГУ имени М.В. Ломоносова, профессор, e-mail: djakonov@mail.ru
[2] Москва, МГТУ имени Н.Э.Баумана, доцент, e-mail: nastya_gm@mail.ru

котором ищется алгоритм a. На практике при обучении (т.е. настройке параметров w модели) решают такую задачу оптимизации:

$\frac{1}{m}\sum_{i=1}^{m} L(a_w(x_i), y_i) + R(a_w) \underset{w}{\rightarrow} \min,$

где функция ошибки L формализует схожесть ответа алгоритма и целевого значения, а R – штрафует алгоритм за сложность.

Мы рассмотрели постановку задачи в простейшем случае, на практике не все признаки могут быть вещественные, почти всегда бывают категориальные признаки, значения которых определяют принадлежность объекта к какой-нибудь категории, например «образование», «профессия» и т.п. [3] В рассматриваемой в данной работе прикладной задаче большинство признаков как раз категориальные, но описание техники взаимодействия с категориальными признаками для изложения материала не принципиально. Отметим, что в описанной постановке задача машинного обучения с учителем называется регрессией, а если целевой признак категориальный, то классификацией.

В разделе 1 представим методы обнаружения закономерностей в данных с помощью сведения этой задачи к задаче обучения с учителем, а также методы обнаружения ошибок с помощью анализа погрешностей решения описанных задач. Отдельно опишем методы обнаружения артефактов в данных: пропусков, аномальных значений и т.п. В разделе 2 кратко изложим методы селекции признаков и некоторые методы оценки важности признаков. В разделе 3 опишем данные национального виктимизационного опроса и результаты оценки важностей признаков, а в разделе 4 – найденные важные признаки и закономерности, неточности и ошибки в данных.

1. Анализ данных средствами машинного обучения

На практике часто нет целевого признака, а есть данные, собранные в признаковой матрице X, и интерес представляет нахождение закономерностей (нетривиальных и практически полезных или интересных зависимостей в данных). Эту задачу можно свести к машинному обучению с учителем: зависимость можно формализовать в виде выражения

$$x_{ij} = f(x_{it_1}, \dots, x_{it_k}) + \varepsilon_i,\ i \in \{1,2,\dots,m\}, \qquad (2)$$

для некоторого подмножества признаков $J = \{t_1, \dots, t_k\}$:

$\emptyset \neq J \subseteq \{1,2,\dots,n\}\backslash\{j\}$, фиксированного индекса $j \in \{1,2,\dots,n\}$, ε_i – шум (погрешность представления), f – некоторая функция. Формально можно взять $J = \{1,2,\dots,n\}\backslash\{j\}$ (если равенство (2) выполнено для некоторого множества J, то оно выполнено и для любого его надмножества), но интересны закономерности, содержащие небольшое число признаков, т.е. когда мощность $|J|$ небольшая.

Далее через $X[:, J]$ будем обозначать подматрицу матрицы X, образованную столбцами из множества J. Естественный подход нахождения закономерностей состоит в следующем:

1) решаем задачу машинного обучения с j-м целевым признаком и признаковой матрицей $X[:, J]$, $J = \{1,2, \dots, n\} \backslash \{j\}$;

2) если качество решения достаточно высокое (т.е. погрешность в (2) небольшая), то определяем множество признаков J минимальной мощности, для которого решение всё ещё достаточно качественное (как вариант – находим все такие множества признаков).

Второй пункт можно выполнять итерационно, постепенно понижая мощность признаков. Описанный метод имеет ряд особенностей. Во-первых, он позволяет искать зависимости определённого вида, который определяется выбором модели $\{a_w\}$, в которой мы ищем подходящий алгоритм. Например, с помощью линейной регрессии ищутся линейные зависимости в данных

$$x_{ij} = \sum_{t \in J} c_t x_{it} + \varepsilon_i,\ i \in \{1,2, \dots, m\},$$

а с помощью обобщённой линейной регрессии с признаками – мономами ограниченной степени [1] – полиномиальные зависимости. Наконец, можно использовать сложные нелинейные модели, здесь популярны модели, основанные на ансамблировании (т.е. совместном использовании) решающих деревьев: случайные леса [4], бустинг над деревьями [5]. В каждом методе есть параметры отвечающие за сложность, например глубина деревьев, что позволяет не переобучаться при поиске закономерностей.

Для оценки, насколько хорошо найдена зависимость (2) используют стандартные схемы перекрестной проверки [1], например, алгоритм обучают на множестве объектов $I\colon\ \emptyset \neq I \subset \{1,2, \dots, m\}$ (здесь для простоты объект x_i отождествляем с его номером i), а ошибку оценивают на множестве объектов $I' = \{1,2, \dots, m\} \backslash I$. Процедуру проводят несколько раз, например, так, чтобы объединение всех использованных непересекающихся множеств I равнялось $\{1,2, \dots, m\}$. В случае, если процедура проводится k раз и мощности всех I примерно равны, т.е. множество $\{1,2, \dots, m\}$ делится на k примерно равных частей, её называют k-fold-проверкой [1]. В результате такой k-fold-проверки мы получаем вектор ответов $(a_1, \dots, a_m)$ алгоритма на всех элементах обучающей выборки: $a_i = a((x_{it})_{t \in J})$, $i \in \{1,2, \dots, m\}$ (строго говоря, здесь несколько алгоритмов, обученных на разных I). Это позволяет вычислить ошибку

$$\frac{1}{m} \sum_{i=1}^{m} L(a_i, x_{ij}),$$

На практике интерес представляют объекты, для которых значения $L(a_i, x_{ij})$ превосходят значения ошибок на остальных объектах, поскольку

это может быть связано с некорректными значениями признаков (например, аномально большими, ошибочными и т.п.) [6]. При вещественном целевом признаке обычно используют функцию ошибки MSE (mean squared error): $L(a_i, y_i) = (a_i - y_i)^2$, при категориальном – любую метрику качества, применяемую в задачах классификации (площадь по ROC-кривой AUC, логистическую ошибку log_loss) [1]. Для обнаружения аномалий в последнем случае учитывают вероятности принадлежности объектов к классам (практически любой современный алгоритм классификации оценивает также и их).

Сделаем несколько замечаний:

1. Изначально все признаки в признаковой матрице даны в разных масштабах, поэтому метрики типа MSE нельзя использовать для сравнения признаков (например, насколько один лучше другого восстанавливается по значениям остальных признаков). Однако, можно привести все признаки в один масштаб с помощью нормализации, следует также делать винсоризацию [7].

2. Описанный подход позволяет также находить схожие признаки (дубликаты или сильно скореллированные). Ясно, что для поиска нетривиальных зависимостей, необходимо устранить дубликаты из признаковой матрицы.

Опишем также метод анализа артефактов, под артефактами мы понимаем особые значения признаков, например, неизвестные (в различных библиотеках анализа данных они кодируются специальными значениями типа NaN – «Not a Number») или аномальные (несвойственные для данного признака, например, температура тела 3700 C или доход –7 руб.). Пусть в -м признаке есть артефактные значения $\{x_{i_1 j}, \dots, x_{i_k j}\}$ в позициях из $I = \{i_1, \dots, i_k\}$, тогда поставим задачу машинного обучения с признаковой матрицей $X[:, J]$ при $J = \{1,2, \dots, n\} \setminus \{j\}$ и целевым вектором $q = (q_1, \dots, q_m) \in \{0,1\}^m$: $q_i = 1 \Leftrightarrow i \in I$. Если указанную задачу удастся решить с достаточно большим качеством, то наличие артефактного значения определяется значением других признаков. По описанной выше схеме можно определить, от каких именно признаков и как зависят артефакты.

Также сведение к задаче машинного обучения подходит для обнаружения некоторых видов «утечек данных» (data leakages) [8]. Например, можно в качестве целевого признака выбрать служебный признак «номер строки» $(1,2, \dots, m)$, если его значения достаточно хорошо восстанавливаются (можно сравнить с восстановлением случайной перестановки его значений) по признаковой матрице X, то значит, что порядок объектов неслучаен (часто объекты упорядочены по времени возникновения) или есть признаки, отвечающие за порядок. На практике это может быть нежелательным эффектом, поэтому объекты

перемешивают, и подобные признаки удаляют или заменяют содержащими меньше информации (например, вместо точной даты – день недели и месяц).

2. Селекция признаков и важности признаков

В описанных в разделе 1 подходах важную роль играло нахождение оптимального подмножества признаков для решения конкретной задачи машинного обучения, такое нахождение называется селекция / отбор признаков (Feature Selection) [1, 9]. Основная цель селекции – удаление избыточных и нерелевантных (шумовых) признаков. На практике (безотносительно к рассматриваемым в данной работе задачам) отбор признаков необходим по следующим причинам:

- интерпретация (ответ на вопрос, от чего и как зависит ответ алгоритма),
- скорость работы алгоритмов (чем меньше признаков, тем быстрее работают алгоритмы на признаковых данных),
- борьба с переобучением (многие методы, например линейные, не рекомендуется использовать при корреляции между признаками),
- повышение качества (шумовые признаки могут снижать качество решения задачи),
- удешевление решения (если вычисление значений признаков дорогостоящая процедура, это случается, например, в медицине и геологии),
- моделирование (когда хотим, чтобы решение зависело от определённых признаков).

Отметим, что не следует путать отбор признаков с извлечением признаков (feature extraction) и сокращением размерности (dimensionality reduction), в перечисленных процедурах, как правило, генерируются новые признаки. При отборе – удаляются уже имеющиеся. Есть три основные группы методов селекции [9]:

- Фильтры (filter methods) – методы, которые по отдельному признаку и, возможно, целевому признаку получают оценку его важности. Как правило, они не ориентированы на конкретные модели алгоритмов машинного обучения.
- Обёртки (wrapper methods) – методы, которые получают оценки качества признаков с помощью анализа работы алгоритмов машинного обучения на подмножествах признакового пространства. Они ориентированы на конкретные модели алгоритмов машинного обучения (хотя и работают с моделью, как с чёрным ящиком).
- Встроенные (embedded methods) – являются частью методов машинного обучения. Качество признаков получается автоматически параллельно настройке модели.

В данной работе мы рассмотрим способы оценки важностей признаков (feature importance), в первую очередь, для ансамблей решающих деревьев, поскольку нас интересует поиск нелинейных закономерностей. Селекция с помощью важностей заключается в удалении (возможно, итерационном) неважных признаков и относится ко второй или третьей группе методов селекции, в зависимости от особенностей реализации.

Решающие деревья [10] строят, как правило, рекурсивно, проводя расщепления в листьях, т.е. разбивая лист R на подлистья

$$R_{\text{left}} = \{x_i \in R | x_{ij} < \theta\},\ R_{\text{right}} = \{x_i \in R | x_{ij} \geq \theta\}$$

(значения -го признака сравниваются с порогом θ), минимизируя функцию

$$Q(R, j, \theta) = |R| \cdot H(R) - |R_{\text{left}}| \cdot H(R_{\text{left}}) - |R_{\text{right}}| \cdot H(R_{\text{right}}), \qquad (3)$$

где H – мера неоднородности (impurity). В задачах классификации используются, как правило, энтропия:

$$H(R) = -\sum_k p_k \log_2 p_k,$$

или функция Джини

$$H(R) = \sum_k p_k (1 - p_k) = 1 - \sum_k p_k^2,$$

здесь p_k – доля объектов -го класса в рассматриваемой области, суммирование идёт по всем классам. Мера неоднородности минимальна (и обращается в ноль) только если все объекты принадлежат одному классу. В задаче регрессии в качестве меры неоднородности используют, как правило, дисперсию целевых значений объектов из области

$$H(R) = \text{var}(\{y_i | x_i \in R\}).$$

При построении дерева перебираются признаки j из некоторого множества (определяется конкретным алгоритмом), для каждого признака перебираются пороги θ из некоторого множества (также определяется алгоритмом). Для каждой пары признак-порог (j, θ) вычисляется (3) и выбирается пара, соответствующая максимальному значению (3), которое называется уменьшением неоднородности по этому признаку на этом расщеплении. Один из стандартных способов определения важности признака j – сумма уменьшений неоднородностей по всем расщеплениям, в которых он участвовал (иногда она нормируется на сумму важностей всех признаков). Такая важность называется важностью по неоднородности (impurity-based importance) и была предложена в [4] для оценки важностей признаков в случайном лесе. Ещё одно распростанённое название – Gini importance (когда в качестве неоднородности используется Джини).

Если алгоритм машинного обучения является композицией деревьев, например, случайный лес (Random Forest) или бустинг (Gradient Boosting), то важность признака равна среднему важностей по всем деревьям, входящим в обучение (MDI – mean decrease impurity). Описанный выше подход к оцениванию важностей реализован в библиотеке scikit-learn [11]. Из преимуществ важности по неоднородности стоит отметить, что она автоматически вычисляется при синтезе алгоритма (правда, подходит только для алгоритмов, основанных на решающих деревьях), из недостатков – её смещённость в сторону признаков с большим числом значений [12].

Другой естественный способ оценки важности признаков – перестановочная важность PFI (Permutation Feature Importance) – базируется на простой идее: признак важный, если перестановка его значений снижает качество алгоритма машинного обучения [4]. Перестановка предпочтительнее других преобразований данных (например, зашумления признака), поскольку не меняет распределение значений признака. Кроме того, можно использовать произвольный алгоритм машинного обучения (не обязательно основанный на решающих деревьях), его не надо переобучать при оценке важности. Реализация соответствующей функции есть в библиотеке scikit-learn [11]. Изменение качества можно наблюдать на обучающей или отложенной выборках (можно также использовать любые валидационные схемы). Кроме того, часто при реализации алгоритмов, в которые встроена процедура бутстрепа [4], например случайного леса, снижение качества вычисляют на OOB-подвыборке [4]. Таким образом, есть возможность строить алгоритм и оценивать качество признаков, имея фиксированную обучающую выборку (и не используя дополнительные данные).

Сделаем несколько замечаний:

1. Если есть несколько сильно коррелированных признаков, то перестановка значений одного из них может слабо влиять на качество решения. Один из вариантов решения этой проблемы – кластеризация признаков (по функции схожести «корреляции») и формирование признакового пространства из представителей кластеров [11]. Эта проблема исследована также в [13].

2. Ясно, что оценка важности зависит от перестановки и на практике лучше сделать несколько перестановок, вычислить выборочное среднее важностей и дисперсию. Оценка «заточена» под конкретную модель алгоритмов и функцию ошибки.

Иногда вместо перестановочной важности используют важность удалением (Drop-column importance): перестановку заменяют на удаление соответствующего столбца. Естественно, на редуцированной признаковой матрице алгоритм нужно учить заново, поэтому метод очень трудоёмкий,

тем не менее, некоторые исследователи предпочитают использовать именно его [14].

3. Данные национального виктимизационного опроса

Описанные в работе методы поиска закономерностей и оценки важностей признаков применялись для анализа национального виктимизационного опроса. Это первый подобный масштабный опрос, проведённый в РФ, и до настоящего времени методы машинного обучения и анализа данных в подобных опросах не применялись. Данные для анализа собраны Институтом проблем правоприменения (ИПП) при Европейском Университете в Санкт-Петербурге (ЕУ СПб) и представляют результаты телефонного опроса респондентов не моложе 18 лет по технологии CATI на основании простой случайной выборки телефонных номеров [15]. Респонденты опрашивались независимо от гражданства, но, по понятным причинам, большинство (98.1%) были гражданами РФ. Вопросы делились на анкетные (пол, возраст, социо-демографический профиль), основной («были ли жертвой преступления» – точная формулировка ниже) и уточняющие вопросы о преступлениях (где, когда и т.п.). Анкета составлена специальным образом профессионалами, учитывая многие особенности, например специфику русского языка [15], хотя мы покажем, что в данных есть неточности вызванные, в том числе, не совсем корректной методикой опроса. Ниже для краткости некоторые вопросы немного перефразированы.

Основной вопрос интервью звучал так: Q75 «Вспомните, пожалуйста, было ли такое, что вас обокрали, вас побили, вам угрожали, вы стали жертвой насилия, мошенничества или других преступлений в России за последние 5 лет?» (здесь и далее нумерация вида «Q75» соответствует нумерации вопросов в данных, которые выложены в свободный доступ [16]). В табл. 1 показана статистика ответов на него – 18% респондентов не ответили на этот вопрос «Нет». Отметим, что «преступление» трактуется довольно широко, в частности, преступлением является СМС-мошенничество (например, когда приходит СМС с просьбой перевести деньги).

Да	3001 (17.8%)
Нет	13776 (81.9%)
Не помню / затрудняюсь ответить	41 (0.2%)

Табл. 1 Статистика ответов на основной вопрос.

Для поиска закономерностей в данных использовалась признаковая матрица: строки соответствовали респондентам, а столбцы – вопросам.

При применении описанной выше методики анализа данных к результатам опроса возникает одна проблема: опрос не проводился по схеме «все вопросы всем респондентам», в том числе, и по причине, что на некоторые вопросы типа «Поймали ли преступника?» не у всех есть ответы (например, у тех, кто не был жертвой). Поэтому формально выявлялись зависимости от целых групп нерелевантных признаков. Например, при восстановлении значений признака Q66 «Можно ли сказать, что в Вашем случае преступление произошло через телефон / интернет (например, злоумышленники Вам звонили или писали с просьбой перевести им деньги) или к Вашему случаю это неприменимо?», высокая важность была у признаков вопросов

Q30 – про риск погибнуть в результате нападения,

Q26_1-3 – про физический ущерб, телесные повреждения,

Q25_1, Q25_2, Q25_3 – про оружие,

которые не задавались, тем, кто стал жертвой «электронного» преступления (и имели для соответствующих строк значения «NaN»). Эта проблема решается удалением соответствующих признаков.

Стоит также сказать про интерпретацию применяемых методов. Возникают естественные возражения против осмысленности модели, которая, например, по описанию человека определяет, сталкивался ли он с преступлением:

- данных не так много (если респондент был жертвой преступления, то проходил полный опрос – таких было 3001 человек, из остальных только у 3719 спрашивали анкетные данные),
- постановка задачи не совсем корректна, т.к. если человек становился жертвой преступления, то это был совершившийся факт (в котором, впрочем, есть доля случайности), а если нет, то это могло случиться с ним вскоре после интервью, т.е. его анкетные описания – это не описания человека, с которым точно ничего не случилось.

Однако, модель потребуется нам для нахождения закономерностей в данных, а не для предиктивной аналитики (поэтому мы даже не будем описывать качество моделей). Она может и должна верифицироваться другими методами, но мы оставим за рамками этой статьи обоснования и вычисления статистической значимости выводов. Как отмечено в [17], применяемые в машинном обучении методы оценки важности признаков не предоставляют возможности оценки статистической значимости. Кроме того, в подобной постановке традиционно решаются задачи в банковской отрасли, например, в скоринге (по описанию клиента определить, вернёт ли он кредит). Практика показывает, что модели машинного обучения в подобных постановках вполне разумны.

Заметим, что в реальности, данных для построения модели было даже меньше, чем описаний 3001 + 3719 респондентов. Во-первых,

использовались только описания респондентов, которые чётко ответили на вопрос «были ли они жертвой преступления» («да» или «нет»). Во-вторых, некоторые вопросы задавались не всем, а в ответах некоторых, кроме понятного «затрудняюсь ответить» есть незаполненные поля. В итоге, для построения моделей отобрано 5796 анкетных данных.

На рис. 1 показана оценка важностей признаков по неоднородности методом случайного леса [4] при разных значениях его гиперпараметра «число признаков, рассматриваемых при расщеплении» (max_features) для целевого признака Q75. Случайный признак «rnd» (его значения равномерно распределены на отрезке [0, 1]) был искусственно добавлен к данным – и видно, что он имеет самую большую важность (что вызвано большим числом его уникальных значений). Подобный пример неадекватности оценки MDI приводился в [14], но даже там случайный признак не был лучшим с подобным отрывом. Интересно, что применение для оценки метода PFI не решает проблему, если использовать одни и те же данные для настройки алгоритма и для оценки важностей, см. рис. 2 (такую оценку мы назвали «простой»). Эта проблема решается, при многократном разбиении матрицы данных на обучающую и матрицу для оценки важностей, см. рис. 3 (Hold-out-оценка). Применялось 10 разбиений с 20%-й валидационной выборкой и 2 перестановками значений для каждого разбиения.

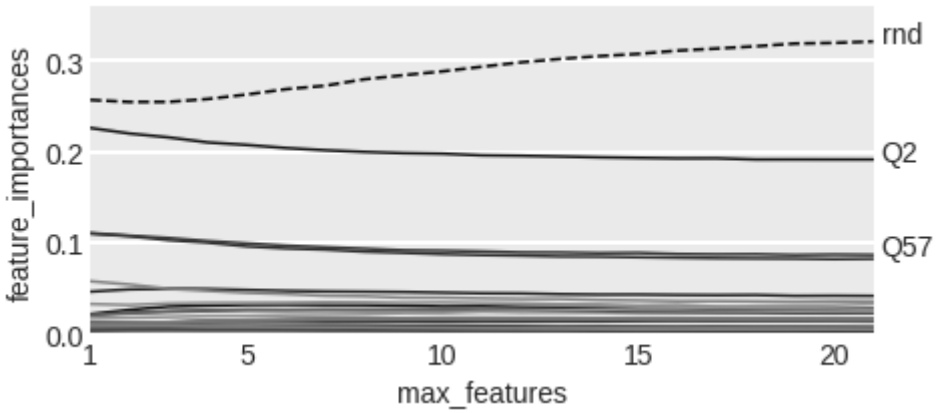

Рис. 1 Оценка важностей признаков методом MDI

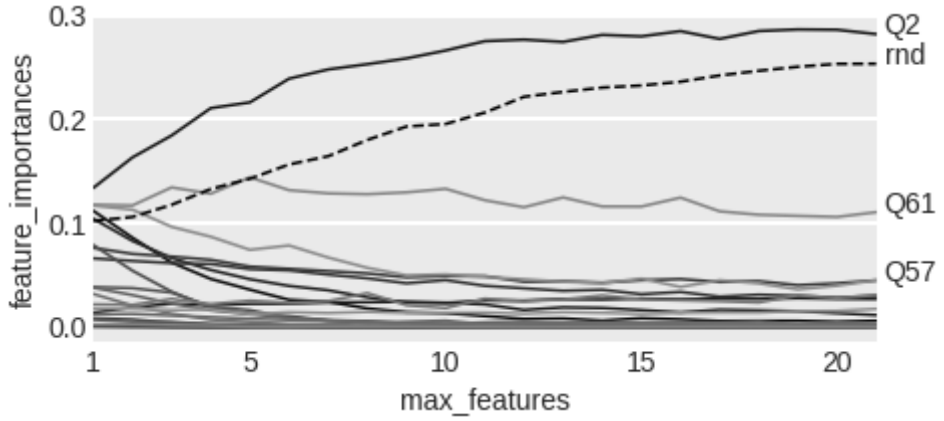

Рис. 2 Простая оценка важностей признаков методом PFI

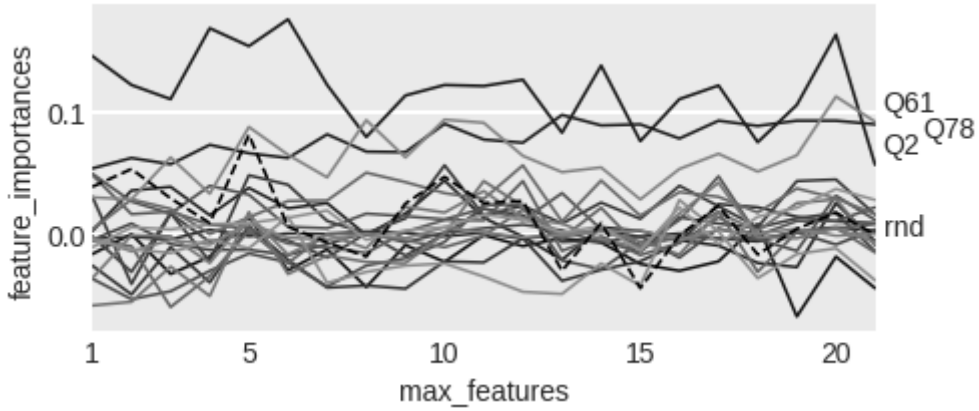

Рис. 3 Hold-out-оценка важностей признаков методом PFI

Приводимые графики практически не зависят от критерия расщепления (приведены для Джини). На рис. 4 показаны средние значения важностей (усреднялись по всем значениям параметра max_features) для разных мер неоднородностей, в том числе, когда целевой признак считался регрессионным (MSE). Интересно, что в задаче классификации выбор меры неоднородности не сильно влияет на оценку важности, а вот переход к задаче регрессии делает несколько признаков очень важными. Для градиентного бустинга над деревьями выводы аналогичные. На рис. 5 показаны оценки важности для случайного леса (RF), градиентного бустинга из библиотеки LightGBM (lgb) – важности усреднены для разных значений параметра num_leaves (число листьев в деревьях), градиентного бустинга из библиотеки XGBoost (xgb) – важности усреднены для разных значений параметра max_depth (глубина деревьев). Тройка самых важных признаков для всех алгоритмов одинакова.

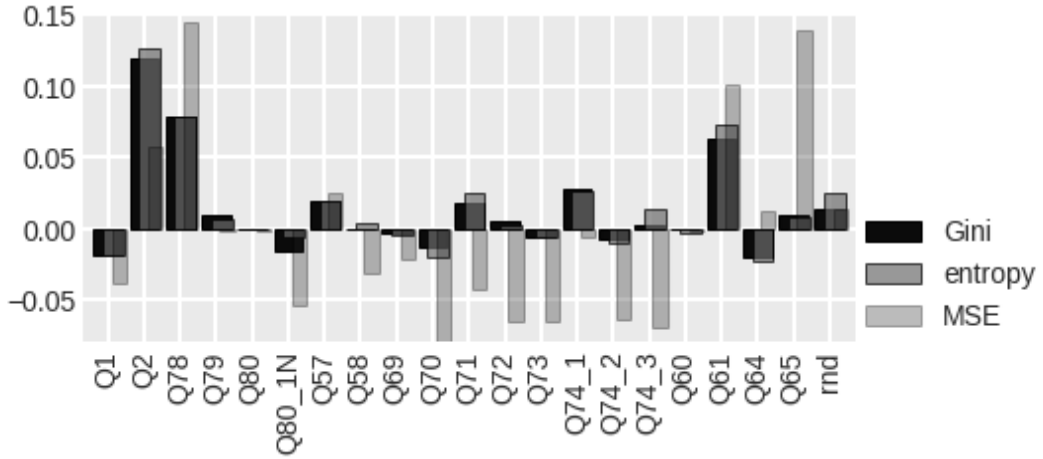

Рис. 4 Hold-out-оценка важностей при разных критериях расщепления

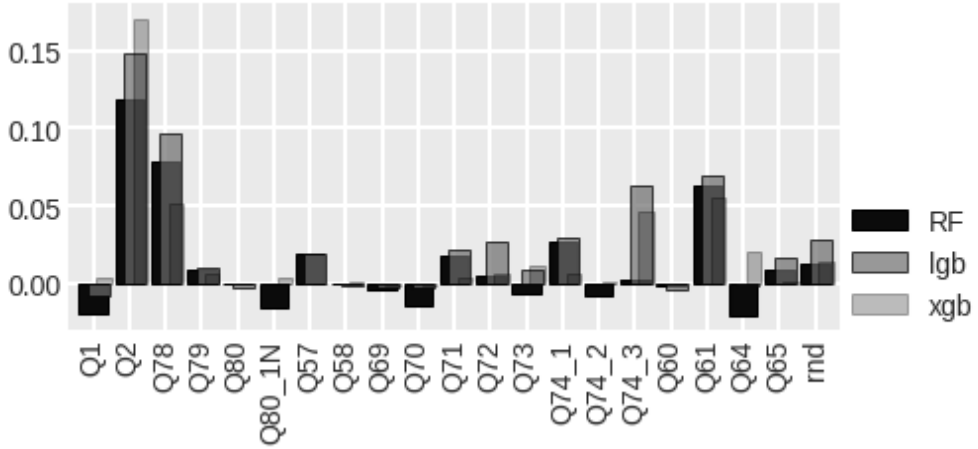

Рис. 5 Hold-out-оценка важностей для разных моделей алгоритмов

4. Выявленные закономерности и несоответствия в данных

Рассмотрим найденные важные признаки для основного вопроса – Q75. Один из самых важных признаков соответствует вопросу о возрасте Q2 «Скажите, пожалуйста, сколько Вам полных лет?» На рис. 6 показано распределение возрастов респондентов. Отдельно изображено число мужчин и женщин каждого дискретного ответа «число полных лет» (признак «пол» не считался моделью важным). До 40 лет (включительно) среди респондентов больше мужчин (4230 против 4108), а после 40 существенно больше женщин (5186 против 3294), в [15] отмечено, что это согласуется со статистикой по РФ. Как часто бывает в опросах, «круглые числа» (20, 30 и т.п.) называют чаще (многие округляют свой возраст), см. рис. 6. На рис. 7 показано распределение по возрастам жертв и не-жертв преступлений – плотности получены методом Парзена (с поправкой на ограничение 18+ при опросе, у молодых больше шансов стать жертвой преступления).

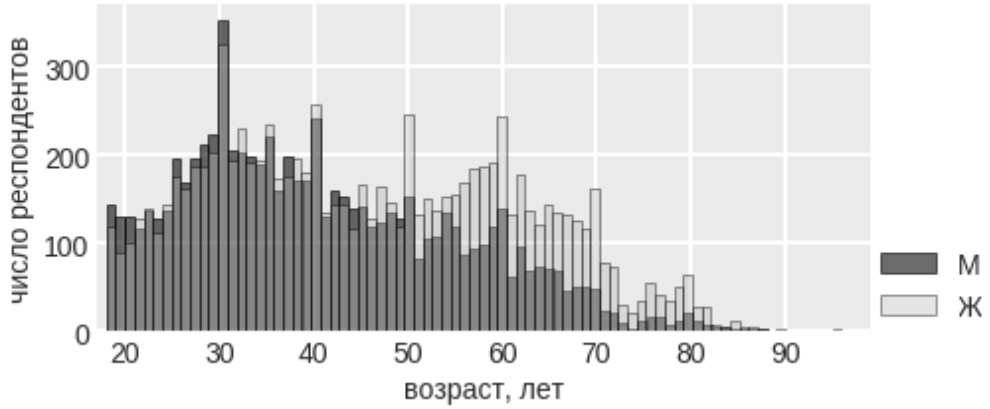

Рис. 6 Распределение респондентов по возрасту и полу

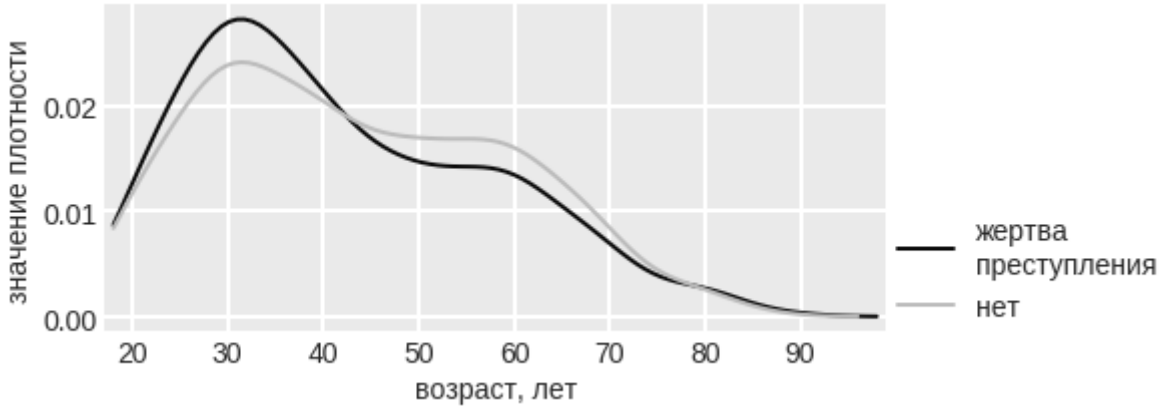

Рис. 7 Распределение возрастов жертв преступлений и остальных респондентов

Ещё один важный признак, как ни странно, связан с вопросом Q61 «Какое у Вас образование?». Интуитивно, чем выше уровень образования, тем меньше шансов стать жертвой преступления, но статистика говорит об обратном, причём проценты существенно отличаются. Среди опрошенных (напомним, что не всем не-жертвам задавались вопросы), которые чётко ответили на основной вопрос Q75, среди людей с (м.б. незаконченным) высшим образование – более 51% жертв преступлений, среди людей со средним специальным/техническим или начальным профессиональным образованием – 44% жертв, а среди людей с полным средним или ниже – 40%. Причём данная закономерность не зависит от возраста респондента.

Также важный признак соответствует вопросу Q78 «Скажите, Вы проживаете один или с кем-то?», см. табл. 2 – статистика подтверждает, что одинокие сталкиваются с преступлениями чаще (указан процент жертв среди тех, кто чётко ответил на основной вопрос и кому задавался вопрос Q78 – это не соответствует проценту среди населения). Интересно, что те, кто не говорит о личной жизни (это можно заметить и в ответах на другие вопросы), существенно реже сталкивались с преступлениями

(15.5% против 44.6% и 47.8%), но эта гипотеза нуждается в проверке на большей по объёму выборке.

	жертва	не-жертва	неизвестно
Один	509 (47.8 %)	573	6
С кем-то	2481 (44.6 %)	3076	31
Затрудняюсь / не скажу	11 (15.5 %)	60	0

Табл. 2 Статистика ответов на вопрос о сожительстве Q78 и основной Q75

Перечислим некоторые найденные с помощью описанной методологии закономерности. Женщины чаще становятся жертвами «электронных» преступлений, см. табл. 3. При ответе на вопрос «Q10 Кем злоумышленник Вам приходился?» 11 женщин назвали преступниками своих мужей, ни один мужчина не назвал преступницей жену (отметим, что на этот вопрос ответило всего 373 человек). Аналогичная ситуация с сожителями: только у двух мужчин был ответ на вопрос о личности преступника «сожитель» (против 13 у женщин), причём, судя по описанию преступления, это мог быть сожитель-мужчина.

	электронное преступление	нет	неизвестно
женщины	561 (34.9%)	1045	12
мужчины	342 (24.9%)	1030	11

Табл. 3 Статистика ответов на вопрос о «электронных» преступлениях

Предложенная техника также позволила указать на возможные проблемы с формулировками ответов, например, у вопроса Q57 «Как бы вы определили свой уровень дохода?» есть заранее заданные варианты ответа «едва сводим концы с концами», «денег не хватает на продукты» и т.п., которые естественным образом упорядочены организаторами соцопроса, но, видимо, этот порядок не соответствует тому, как воспринимают эти описания респонденты. Косвенное подтверждение этому – данный признак очень важный и среди опрошенных «бедных» около 50% жертв, а среди опрошенных богатых – меньше 45%, но в остальном зависимость немонотонная (что может быть связана также и с малым объёмом выборки).

При анализе данных опроса было найдено довольно много несоответствий. Некорректна категория «электронные преступления»: у респондентов спрашивали про преступления совершённые «через телефон или интернет», многие путали это с преступлениями, в которых

фигурировали телефон или интернет. Например, в 15 таких случаях в качестве материального ущерба называлась «техника», их описания: «кража телефона у моего ребенка», «украли телефон», «отдали планшет в ремонт частному мастеру, который планшет не вернул» и т.п. Подобное несоответствие выявлено с помощью анализа редких категорий (см. раздел 1). Аналогично, анализируя вопрос о том, кем был преступник, удалось найти преступления, помеченные как электронные, но вряд ли таковыми являющиеся: «не оплата коммунальных платежей квартирантами», «поджог дома дочки», «человек на судебном заседании под камеру угрожал расправой» и т.п. Из простых ошибок можно отметить несоответствие описания данным, например, для вопроса Q10 код «7» соответствует ответу «Сосед или соседка», а код 8 – ответу «Сожитель», а не наоборот (что подтверждается описанием преступлений) [16].

Самая интересная закономерность, вскрывающая методический просчёт в проведении опроса – странности в распределении времён преступлений. На вопрос Q16 «В каком примерно месяце это было?» 20.7% опрошенных ответили, что не помнят точно месяц преступления. У остальных, распределение числа преступлений по месяцам показано на рис. 8, на котором видны «сезонные колебания»: ослабевание преступной деятельности с мая по июнь. Для «электронных» преступлений наблюдается такой же эффект.

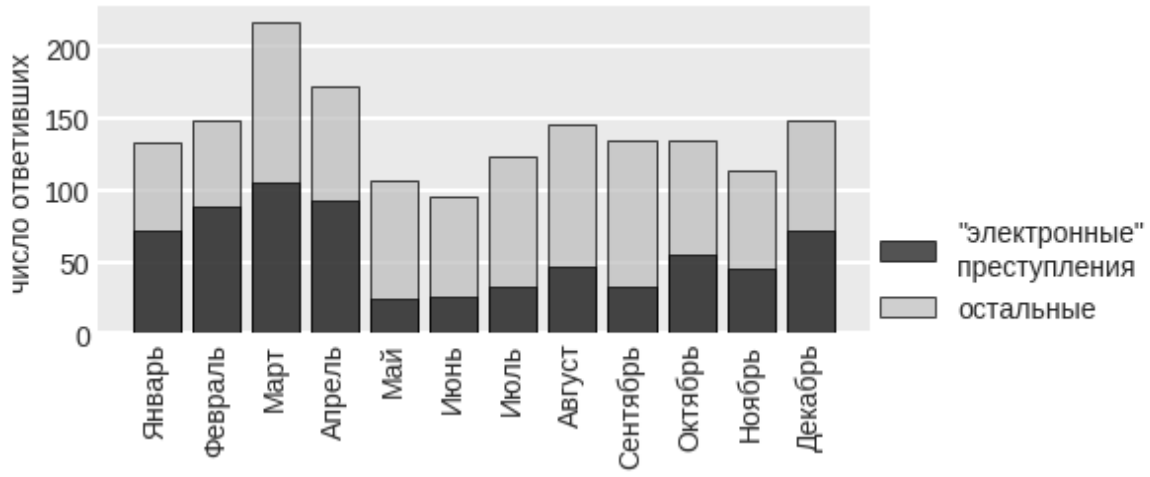

Рис. 8 Распределение числа преступлений в зависимости от типа по месяцам

Одна из возможных гипотез наблюдаемого эффекта – опрос проводился с марта по май (что подтвердилось), поэтому респонденты точно помнили месяцы недавних преступлений (февраль-апрель) и не помнили месяцы давнишних (май-июль). Косвенно эта гипотеза подтверждается также анализом ответов тех, кто назвал время года

преступления: 13.5% не помнили и время года, по ответам помнивших время года, но не месяц, в каждом сезоне было примерно поровну преступлений, что не соответствует ответам помнивших точно месяц. Для последних, весна – самое насыщенное преступлениями время года, а лето – наоборот (разница довольно существенна), см. рис. 9. Здесь перевод ответа «месяц» во «время года» осуществлялся по принятой календарной схеме (например, лето – это июнь, июль и август).

Вывод, который напрашивается по результатам анализа: подобные опросы следует проводить не в фиксированный временной отрезок, а на протяжении всего года, при этом в данных точно указывать дату опроса.

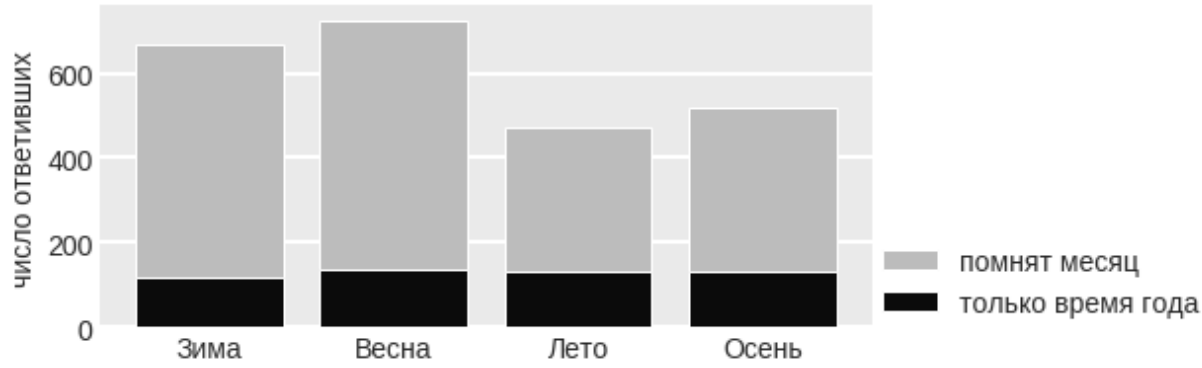

Рис. 9 Распределение ответов про время года преступления тех, кто точно помнил месяц и тех, кто точно помнил лишь время года

Заключение

В работе описана методика анализа данных с помощью решения задач машинного обучения с учителем. В качестве иллюстрации они применена для исследования данных виктимизационного опроса, анализ средствами машинного обучения подобных данных сделан впервые. Также для удобства написана библиотека визуализации ответов на вопросы, выложенная в общий доступ [18].

Найдены интересные закономерности в данных опроса, например, зависимость возможности стать жертвой от возраста, одиночного проживания и т.п. Получены интересные гипотезы, которые требуют дополнительной проверки на большем массиве информации, например, что люди, которые уклоняются от определённых ответов, с большей вероятностью не были жертвами преступлений.

Выявлены некоторые недостатки опроса, например, его проведение в течение небольшого промежутка времени в конце весны. В результате, респонденты лучше помнили весенние и зимние преступления (возможно, что у летних и осенних преступлений также больше неточностей в описаниях). Также обнаружены многочисленные неточности в данных. Часть из них связана с тем, что респонденты ошибочно понимали вопросы, например путали «преступление через телефон» и

«преступление с телефоном», другая часть – с ошибками внесения информации и неправильными кодовыми таблицами.

В работе получен пример, когда важность случайного признака существенно выше важностей остальных признаков. Ранее подобное наблюдалось на модельных данных: случайные признаки могли иметь неожиданно высокую важность, но здесь она максимальна и данные реальные. Кроме того, подтверждено известное на практике правило: важность следует оценивать на отложенной выборке и метод PFI предпочтительнее метода MDI.

В данной работе не приведены результаты анализа действий респондентов во время / после преступлений, что может быть потенциально хорошей областью для дальнейших исследований. Не показана оценка важности с помощью популярного метода SHAP [19] (результаты аналогичны приведённым, но вычисления существенно более трудоёмкие).

Литература

1. *Hastie T., Tibshirani R., Friedman J.* The Elements of Statistical Learning: Prediction, Inference and Data Mining. Springer Verlag. 2009. Second Edition.
2. *Журавлёв Ю.И.* Об алгебраическом подходе к решению задач распознавания или классификации //Проблемы кибернетики, 1978, Вып. 33, с.5-68.
3. *Дьяконов А.Г.* Методы решения задач классификации с категориальными признаками //Прикладная математика и информатика, 2014, № 46, с.103-127.
4. *Breiman L.* Random Forests //Machine Learning, 2001, 45(1), p.5-32.
5. *Friedman J.H.* Greedy function approximation: A gradient boosting machine //Annals of Statistics, 2000, № 29, p.1189-1232.
6. *Дьяконов А.Г., Головина А.М.* Выявление аномалий в работе механизмов методами машинного обучения //Аналитика и управление данными в областях с интенсивным использованием данных: труды XIX Международной конференции DAMDID/RCDL'2017 (10-13 октября 2017 года, г.Москва), ФИЦ ИУ РАН Москва, 2017, с.469-476.
7. *Alexandropoulos S.-A.N., Kotsiantis S.B., Vrahatis M.N.* Data preprocessing in predictive data mining //The Knowledge Engineering Review, 2019, v.34, №1, p.1–33.
8. *Kaufman S., Rosset S., Perlich C., Stitelman O.* Leakage in data mining: Formulation, detection, and avoidance //ACM Transactions on Knowledge Discovery from Data (TKDD), 2012, 6(4) , p.15:1-15:21.

9. *Guyon I., Elisseff A.* An introduction to variable and feature selection //Journal of Machine Learning Research, 2003, №3, p.1157-1182.
10. *Quinlan J.R.* Learning Efficient Classification Procedures and Their Application to Chess End Games //Machine Learning, 1983, p.463–482.
11. https://scikit-learn.org/stable/ (библиотека для машинного обучения)
12. *Strobl C., Boulesteix A.-L., Zeileis A., Hothorn T.* Bias in random forest variable importance measures //BMC Bioinformatics, 2007, 8(1), 25.
13. *Strobl C., Boulesteix A.-L., Kneib T., Augustin T., Zeileis A.* Conditional Variable Importance for Random Forests // BMC Bioinformatics, 2008, 9(1), p.307.
14. *Parr. T., Turgutlu K., Csiszar C., Howard J.* Beware Default Random Forest Importances // https://explained.ai/rf-importance/
15. *Веркеев А.М., Волков В.В., Дмитриева А.В., Кнорре А.В., Кудрявцев В.Е., Кузнецова Д.А., Кучаков Р.К., Титаев К.Д., Ходжаева Е.А.* Как изучать жертв преступлений? // Мониторинг общественного мнения: Экономические и социальные перемены, 2019, №2, с.4-31.
16. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/C2OTH9 (данные национального виктимизационного опроса)
17. *Van der Laan M.J.* Statistical Inference for Variable Importance. The International Journal of Biostatistics, 2006, 2(1). p.1-31.
18. https://github.com/Dyakonov/visualization (библиотека для визуализации данных опроса)
19. *Lundberg Scott M., Lee Su-In A.* Unified Approach to Interpreting Model Predictions // Advances in Neural Information Processing Systems, 2017, 30, p.4765-4774.