DOI: 10.17323/2587-814X.2021.3.48.59

Подготовка данных для машинного анализа ключевых показателей эффективности территориальных менеджеров

А.Ю. Владова [a,b]
E-mail: ayvladova@fa.ru

Е.Д. Шек [c]
E-mail: 1399selena@gmail.com

[a] Институт проблем управления им. В.А. Трапезникова РАН
Адрес: 117997, г. Москва, ул. Профсоюзная, д. 65

[b] Финансовый университет при Правительстве Российской Федерации
Адрес: 125993, г. Москва, Ленинградский проспект, д. 49

[c] Российский экономический университет им. Г.В. Плеханова
Адрес: 117997, г. Москва, Стремянный переулок, д. 36

Аннотация

Существенная трансформация операционной деятельности компаний – дистрибьюторов продуктов и услуг обусловлена изменениями в технологии получения и обработки данных. На данный момент работа представителей этих компаний в значительной степени оцифрована: например, автоматически фиксируется время нахождения в дороге, количество и места встреч с клиентами. При этом эффективность работы территориальных менеджеров, не совершающих прямые продажи, по-прежнему вынужденно оценивают с помощью опросов, экспертов и затратных двойных визитов, хотя наличие объемной выборки данных позволяет с помощью статистического анализа выявить как недостаточные, так и завышенные значения показателей эффективности работы. Исходные данные: реляционная база данных, накапливающая информацию о 28 категориальных, количественных, геолокационных и временных параметрах активностей территориальных менеджеров за год. На основе имеющихся данных созданы синтетические признаки (широта и долгота – индекс, регион, улица, дом; по идентификаторам вычислены суммы активностей; по временным признакам определены сезон года, день недели и период суток). Методика проведения статистического анализа включала три стадии: сбор и обработку первичных данных, обобщение и группировку обработанной информации, формулирование статистических гипотез и интерпретацию результатов. Для моделирования уровня искажения информации об активности менеджеров использован вероятностный подход. В результате с помощью построенного облака тегов выделены: наиболее популярный сезон для проведения рекламных кампаний; наиболее продуктивные отделы и территориальные представители; дни недели, на которые приходятся наибольшее количество контактов с клиентами. Установлено наличие значительного числа записей о проведении встреч в выходные дни. В результате проведенной разведки данных сформулирована статистическая гипотеза о возможности выявления территориальных менеджеров, искажающих

количество и параметры встреч. Для выявления скрытых взаимосвязей создан набор синтетических целых, действительных и категориальных переменных. Выявлены сомнительные данные (например, работа в выходные дни или ночью). Полученный обобщенный набор данных сгруппирован по признаку идентификатора активности территориального представителя и построено распределение признака. По каждому территориальному менеджеру просуммированы целые и действительные признаки и выявлены выбросы, характеризующие неэффективную работу или искажение данных. Таким образом, наличие объемной выборки данных об истории перемещений и активностях позволяют по косвенным признакам оценить эффективность работы территориальных менеджеров дистрибьюторской компании.

Ключевые слова: машинное обучение; ключевой показатель эффективности; база данных; временной ряд; геолокация; обучение без учителя; торговый представитель; b2b.

Цитирование: Владова А.Ю., Шек Е.Д. Подготовка данных для машинного анализа ключевых показателей эффективности территориальных менеджеров // Бизнес-информатика. 2021. Т. 15. № 3. С. 48–59.
DOI: 10.17323/2587-814X.2021.3.48.59

Введение

Система количественно измеримых ключевых показателей эффективности (key performance indicators, KPI) наиболее эффективна в крупных компаниях, где по сравнению с мелкими предприятиями сложнее выделить вклад каждого работника. Главным положительным моментом внедрения KPI является возможность количественного анализа деятельности работников и последующее планирование, а одним из отрицательных моментов является то, что при отсутствии контроля работник в состоянии адаптировать свои показатели под требования KPI.

Ежедневная работа территориального менеджера (ТМ) в рамках рекламной кампании продукта или услуги состоит из нескольких встреч (активностей) с клиентами, в ходе которых он демонстрирует рекламную презентацию в специальной программе. Эта программа фиксирует дату, время начала и длительность демонстрации, а также определяет географические координаты ТМ с помощью GPS-наблюдения. После проведенной активности ТМ заносит в распределенную базу данных отчет о посещении, устанавливая ряд дополнительных параметров, к числу которых относятся:

- ✦ характер активности (например, индивидуальный, в группе, дистанционный);
- ✦ сегментация клиента по месту работы и специальности, отношению к рекламируемому продукту или услуге;
- ✦ рекомендуемое количество активностей с клиентом.

Таким образом, специфика работы ТМ такова, что факт проведения активности не может быть отслежен работодателем, а характер дополнительных параметров представляет собой частное мнение ТМ.

С другой стороны, естественные для менеджеров по продажам и хорошо контролируемые параметры обратной связи, такие как объем продаж, количество заключенных договоров, закрытие сделок, количество новых клиентов и выполнение показателей плана, нерелевантны для ТМ, участвующих в процессе ознакомления клиентов с продуктом или услугой. Таким образом, из-за невозможности контроля результатов работодатель вынужден переходить к контролю процесса – тех действий, которые совершаются для достижения цели. Это соблюдение работниками регламента и нормативов компании, качество обслуживания и общения с клиентами, знание продукта, возможность самостоятельно принимать решения. Однако вопрос оценки эффективности работы ТМ остается открытым.

1. Существующие исследования

Согласно данным платформы Dimensions [1], которая предоставляет доступ к информации со всего мира о результатах работ по грантам, публикациям, патентам и другим источникам, количество публикаций и патентов в области оценки эффективности территориальных менеджеров с помощью современных методов анализа данных растет от года к году (рисунок 1). Некоторый спад в 2020 году объясняется одно-двухлетним периодом (в среднем), который проходит с момента подачи статьи или заявки до публикации или выдачи патента.

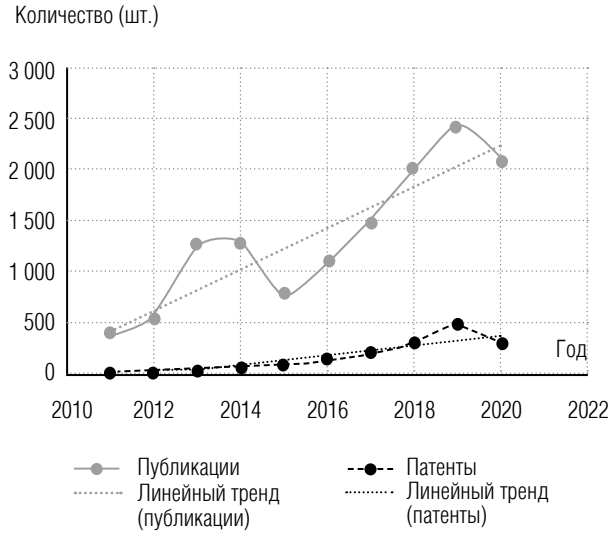

Рис. 1. Динамика числа публикаций и патентов

Существующие исследования в области оценки эффективности работы ТМ, а также выявления зависимостей между контролируемыми показателями и результативностью работы сотрудников базируются в основном на анализе стандартных методов контроля, вычисления числовых показателей результативности и использования аналитических и CRM-систем.

В статье [2] оценена эффективность подходов к управлению продажами в трех компаниях. Исследованы обоснованность распределения задач и продуктов между менеджерами, обоснованность фокусирования на определенных территориях, а также планирование презентаций продукта. Информация для исследования получена в ходе интервью с руководителями отделов продаж и маркетинга, наблюдения за работой медицинских представителей, а также изучения отчетов о деятельности организации и данных о рынке.

В исследовании [3] определены такие методы контроля результативности, как телефонные звонки, внешний и внутренний аудит, двойные визиты, в ходе которых оцениваются и отрабатываются профессиональные и коммуникативные навыки сотрудника и его умение провести презентацию продукта, ежедневные отчеты в системе CRM и их оценка путем проведения срезов, оценка финансовой отчетности в используемой информационной системе.

Статья [4] настаивает на интеллектуальном планировании деятельности ТМ, поскольку штат ТМ является одним из наиболее затратных по оплате труда и обучения. Выявлено, что работа ТМ в значительной степени оцифрована, так как в CRM-системах фиксируются многие параметры – от времени нахождения в пути до количества встреч в месяц. Существующие аналитические решения выстраивают маршруты до клиентов на основе времени в пути, возможной отмены встречи и других параметров, позволяя определить оптимальную последовательность обхода клиентов. При этом комплексный анализ собранных данных не всегда осуществляется должным образом. Кроме того, ключевой параметр эффективности деятельности ТМ, – человеческий фактор, – с трудом поддается анализу.

Авторы работы [5] рассмотрели цифровые технологии, применимые в сфере здравоохранения. Предложенные меры связаны с внедрением систем контроля KPI для оценки деятельности работников сферы здравоохранения (например, результата проведенной консультации или телефонного разговора администратора клиники). С использованием данных систем появляется возможность определить, какой отдел работает наиболее эффективно, какая услуга приносит больше прибыли.

В статье [6] описано применение гибридной процедуры, основанной на методе *k*-средних и дерева решений, для прогнозирования производительности сотрудников на следующий год. Использованы такие факторы, как личность, пунктуальность, красноречие и т.д. Алгоритм прогнозирует количество сотрудников, выбранных для повышения или увольнения, и помогает выявить неэффективных сотрудников.

Статья [7] посвящена описанию человеческого поведения как цепочке математических моделей – фильтров Калмана, упорядоченных цепью Маркова. Эти модели используют для распознавания поведения человека по сенсорным данным и краткосрочного прогнозирования его действий (например, последующих действий водителей на основе подготовительных движений). Авторы применили наивный байесовский классификатор к набору данных о продажах глобальной транспортно-экспедиторской компании за три года. Классификация проводилась по трем классам: «не справляется с обязанностями», «справляется с обязанностями» и «показывает выдающиеся результаты». Авторы предлагают использовать наивный байесовский классификатор для оценки работы специалистов по продажам с привлечением большего количества информации из CRM-систем.

В работе [8] KPI формируются на основе фрагментарных знаний о бизнес-процессах. Авторы статьи представили метод, позволяющий оценить свойства

KPI до внедрения и включающий сочетание целевого и концептуального моделирования.

Таким образом, анализ литературы показал, что наряду с высоким уровнем цифровизации бизнес-процессов, активность ТМ по-прежнему оценивают проведением опросов, мнениями экспертов и затратными двойными визитами. При этом наличие объемной выборки данных позволяет с помощью статистического анализа выявить как недостаточные, так и завышенные показатели эффективности работы ТМ.

2. Исходные данные

Исходные данные аккумулированы B2B-компанией, являющейся дистрибьютором лекарственных препаратов и БАДов. Компания имеет два бизнес-отдела (RX и OTC) и пять подразделений (T1, T2, T3, T4, T5). Маркетинговая деятельность ТМ компании ориентирована на ознакомление лиц, принимающих решения, с характеристиками продукции (в рамках брендовых кампаний). Параметры, автоматически формирующиеся в ходе встречи с клиентом, и параметры из отчета ТМ о результатах встречи заносятся в распределенную базу данных, управляемую CRM. В ней ведется учет ТМ, клиентов, продуктов и услуг. Имеющаяся выборка содержит более трехсот тысяч наблюдений по 28 различным признакам, к числу которых относятся:

- ✦ уникальные цифровые и/или символьные идентификаторы ТМ, их активностей, клиентов и компаний, в которых они работают, а также рекламных кампаний, в рамках которых проведена активность;
- ✦ категориальные признаки типа активности и компании, специальности, подразделения, категории и целевой группы клиента, рекламной кампании, флаги выполнения активности, информации о присутствии менеджера на активности;
- ✦ временные признаки начала и окончания активности, длительность показа презентации, договоренности по бренду, время определения координат;
- ✦ целые (количество встреч, результаты рекламной кампании, флаги), вещественные и геолокационные признаки: долгота и широта места проведения активности.

Таким образом, имеются умеренно разнородные цифровые данные с преобладанием служебной информации (*рисунок 2а*). Из-за пропусков в данных некоторые типы определились некорректно. После преобразования значений одного типа в значения другого типа (как явного приведения типов, так и определения по дате встречи, времени суток, дня недели, сезона года) соотношение целых, вещественных, временных и категориальных признаков усилилось в пользу целых и категориальных (*рисунок 2б*).

3. Методика статистического анализа

Традиционно выделяют три этапа статистического исследования [9, 10]:

- ✦ сбор и обработка первичных данных;
- ✦ статистическая сводка и группировка обработанной информации;
- ✦ постановка статистических гипотез и интерпретация результатов.

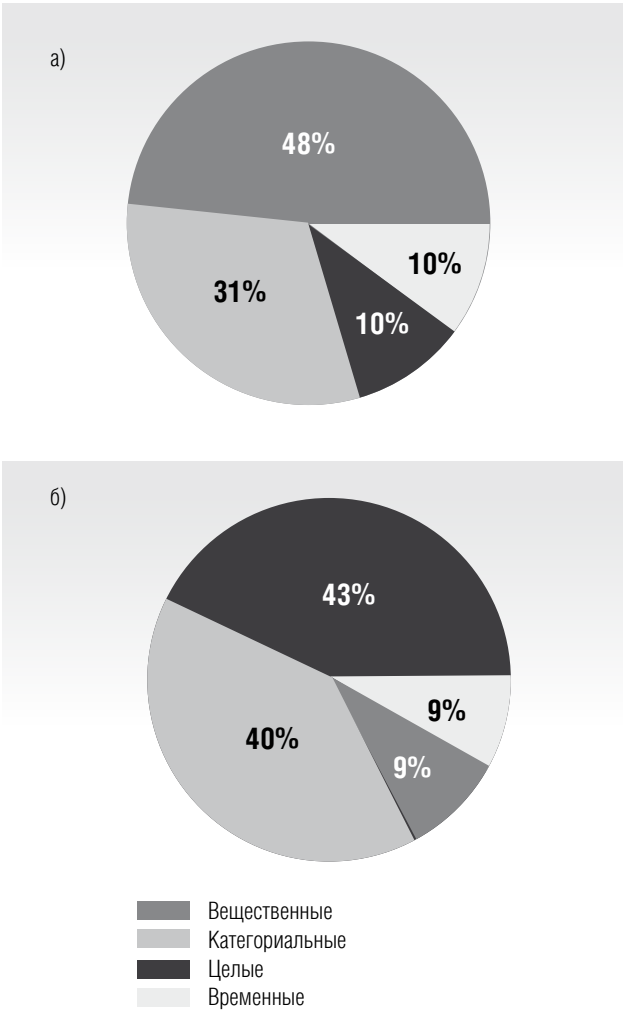

Рис. 2. Соотношение типов признаков:
а) до преобразования; б) после преобразования

♦ *Таблица 1* детализирует основные этапы исследования с помощью подходов, предлагаемых методами разведки данных [11, 12] и машинного обучения [13, 14].

Таблица 1.

Этапы статистического исследования

1.	**Сбор и обработка первичных данных**	• Аудит данных, обработка пропусков и выбросов
		• Типизация и кодирование признаков
		• Синтез временных и географических признаков
		• Построение облака тегов
2.	**Сводка и группировка обработанной информации**	• Профилирование признаков
		• Подбор вида и параметров распределений признаков
		• Отбор слабокоррелируемых признаков
		• Группировка данных
3.	**Гипотезы и интерпретация результатов**	• Постановка статистической гипотезы
		• Моделирование
		• Интерпретация результатов

Ниже представлены основные результаты применения детализированной методики статистического анализа для решения поставленной задачи.

4. Этап 1: Сбор и обработка первичных данных

4.1. Аудит данных

Моделирование временных рядов предполагает исследование связи значений признаков во времени. С помощью разновидности точечного графика – диаграммы запаздывания [15] изучены автокорреляции двух признаков идентификаторы ТМ и идентификаторы активностей (*рисунок 3*).

Большее число точек в области диагоналей предполагает более сильную автокорреляционную связь. Точки, сосредоточенные в середине или распространенные по всей площади рисунка, предполагают слабые связи. Неопределяемая хаотичная структура точек диаграммы указывает на то, что данные случайны.

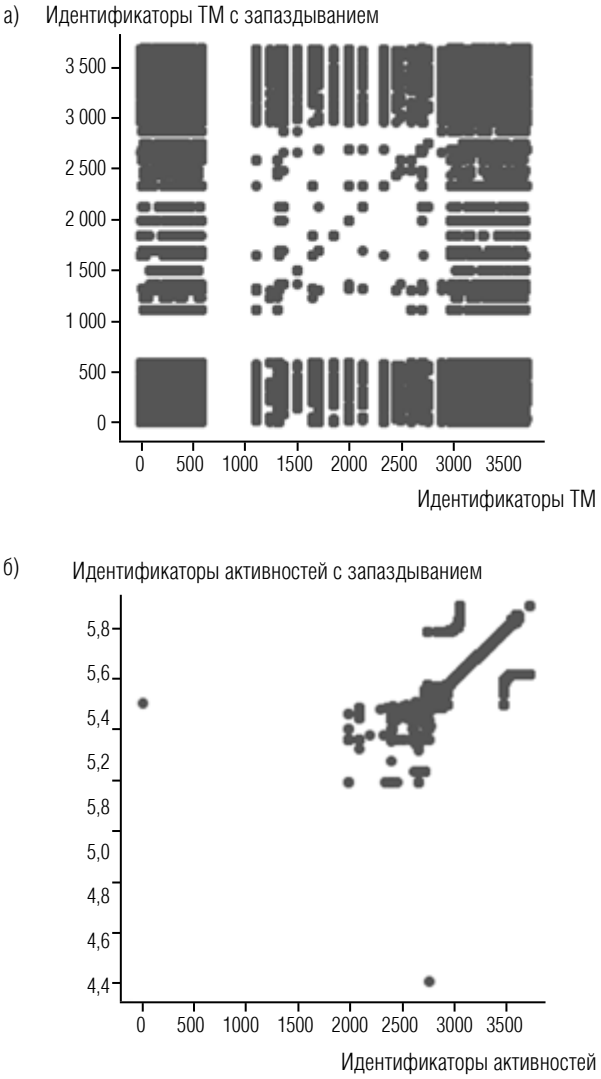

Рис. 3. Неслучайность данных признаков:
а) идентификаторы ТМ, запаздывание 30 наблюдений;
б) идентификаторы активностей, запаздывание 50 наблюдений

Диаграммы показывают наличие структуры у обоих категориальных признаков, что говорит о неслучайности данных. Существование автокорреляционных связей подчеркивается второй диагональю, проходящей через верхний правый и нижний левый углы. Установлено, что для больших данных такой тип графика строится намного быстрее, чем классический автокорреляционный вариант.

4.2. Синтез временных и географических признаков

По временным признакам, отражающим моменты начала и окончания презентации, синтезированы целый признак (порядковый день в году) и категориальные признаки (день недели, сезон года и период

Рис. 4. Карта активностей двух бизнес–отделов

суток, когда проведена активность). По географическим признакам широты и долготы в среде Tableau[1] построена карта активностей бизнес-отделов (*рисунок 4*), а также сформированы синтетические признаки (почтовый индекс, адрес и регион) [16, 17].

На основании картографических данных можно заключить, что, например, большую часть активностей проводит бизнес-отдел OTC. Он же эксклюзивно работает на территории Сахалина и Калининградской области.

4.3. Построение облака тегов

Для анализа значений 12 категориальных и временных признаков, преобразованных в категориальные, с помощью библиотеки WordCloud[2] языка Python построено облако тегов (*рисунок 5*).

Рис. 5. Облако тегов категориальных признаков

С помощью построенного облака тегов выделены:

- наиболее популярный сезон для проведения рекламных кампаний – зима (осенний и весенний сезоны идут со значительным отставанием);
- преимущественный интервал времени суток для встреч с клиентами – день (на втором месте – утро);
- наиболее продуктивный бизнес-отдел по количеству активностей – RX;
- преимущественный тип общения с клиентом – визит (на втором месте – консультация),
- дни недели, на которые приходятся наибольшее количество контактов с клиентами – вторник и четверг (на втором месте – среда и пятница). Установлено наличие значительного количества записей о проведении встреч в воскресенье;
- наиболее частыми клиентами являются поликлиники и аптеки.

5. Статистическая сводка и группировка обработанной информации

С помощью библиотеки pandas-profiling оценены взаимосвязи и построены гистограммы распределений признаков распределенной базы данных [18–20]. *На рисунке 6* представлены гистограммы распределений активностей двух бизнес-отделов дистрибьюторской компании по типам клиентов.

[1] http://elearning.tableau.com/
[2] https://python-graph-gallery.com/wordcloud/

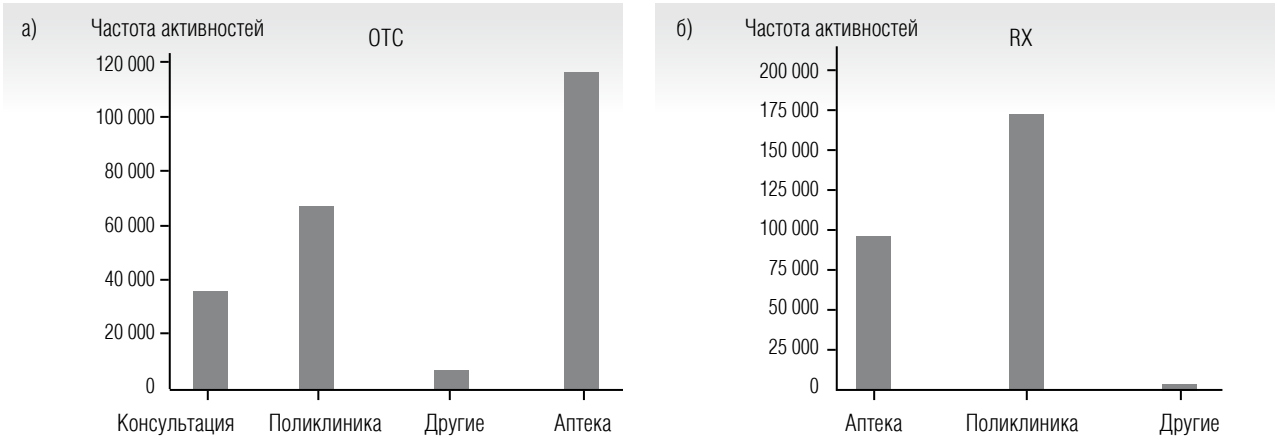

Рис. 6. Распределение частоты активностей
с разными типами клиентов по двум бизнес–отделам: а) OTC; б) RX

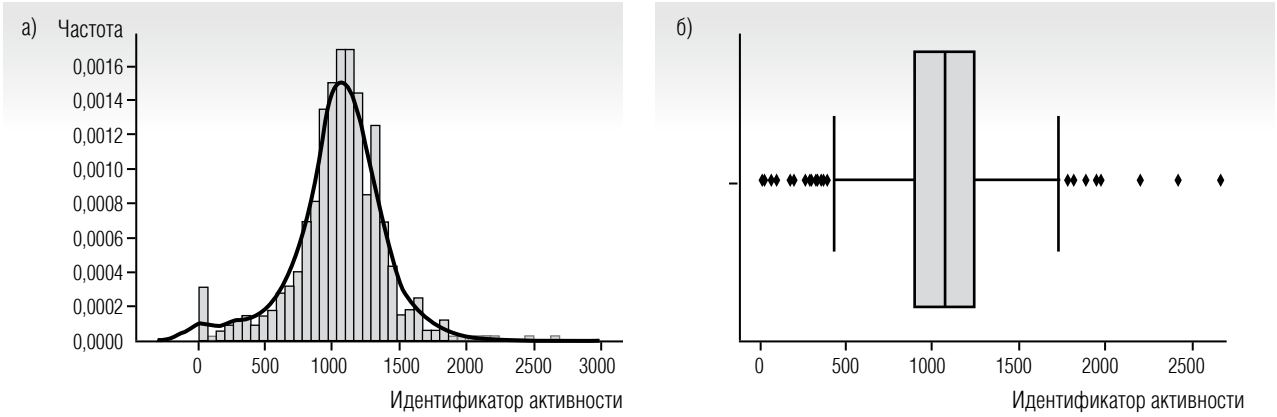

Рис. 7. Распределение признака «идентификатор активности ТМ»:
а) длинный правый хвост; б) выбросы

Наиболее интересным представляется близкое к нормальному распределение признака, характеризующего активность ТМ (*рисунок 7а*). Эта случайная величина проходит комбинированный тест на нормальность “normaltest” библиотеки SciPy, сочетающий тесты Д’Агостино и Пирсона на эксцесс и асимметрию [21], если уменьшить хвосты, убрав выбросы (*рисунок 7б*).

Преобразуем нормализованное распределение признака X «идентификатор активности ТМ» в стандартное нормальное распределение X_n по формуле:

$$X_n \sim Norm(0;1) = \frac{X - m_x}{\sigma_x}, \qquad (1)$$

где m_x – математическое ожидание признака идентификатор активности ТМ;

σ_x – стандартное отклонение признака идентификатор активности ТМ.

Распределение признака «идентификатор активности ТМ» позволило сформулировать статистическую гипотезу об оценке уровня искажения параметров активностей.

6. Гипотезы и интерпретация результатов

Для моделирования уровня искажения параметров активностей используем вероятностный подход [22–24]. Пусть гипотеза H_1 – активность проведена, а альтернативная гипотеза H_2 – активность не проведена. Пусть событие A – попадание параметров проведенной или не проведённой активности ТМ в базу данных. Тогда по формуле полной вероятности вероятность события A вычисляется следующим образом:

$$P(A) = P(H_1)P(A|H_1) + P(H_2)P(A|H_2) \qquad (2)$$

Пусть уровень нормированного идентификатора активности X_n у определенного ТМ принимает значение p. Обозначим вероятность того, что ТМ внес параметры несостоявшейся активности в базу данных, через P_m. Тогда, учитывая (1), перепишем (2) в следующем виде:

$$P(A) = p \cdot 1 + (1-p) P_m \quad (3)$$

Поскольку многие ТМ вводят информацию о встречах в распределенную базу данных, мы имеем дело с потоком информации. Для исходных настроек модели желательно знать портрет среднего ТМ, а именно – уровень его идентификатора активности p. Допустим, что наиболее распространенной является ситуация, когда нормированный идентификатор активности ТМ составляет 0,5. Пусть нормально распределенная величина $p \sim N(0{,}5;\ 0{,}1)$ моделирует входной поток. Ясно, что ТМ не может кардинально менять параметры активностей, иначе они станут выбросами. Пусть ТМ изменяет нормированный идентификатор активности в s раз, $s \in [0;\ 1]$ по отношению к среднему значению, повышая его до p':

$$p' = p + (1-p)s. \quad (4)$$

В результате с вероятностью $P_m = (1 - p)\ s$ ТМ искажает данные, и с вероятностью $1 - P_m = p$ вводит неискаженные данные. *Рисунок 8* показывает результаты моделирования в случае, если ТМ искажают информацию на 10 %.

Дальнейший анализ направлен на подтверждение гипотезы о том, что на основе исходных данных представляется возможным выявить ТМ, наблюдения по которым выбиваются из имеющейся выборки. Отметим, что по сформированному набору данных с учетом временного признака также возможно выполнить прогнозирование по известным моделям [25].

7. Дискуссия

Исходные признаки сгруппированы по идентификатору ТМ, таким образом, мы рассматриваем все наблюдения, которые относятся к конкретному менеджеру. При этом было выделено несколько паттернов для генерации признаков в соответствии с их смысловой сущностью [26, 27]: числовые признаки, обозначающие количество, были просуммированы, также по ним были выделены среднее и медиана. Среди таких признаков – длительность показа презентации, признак договоренности по бренду, время определения координат ТМ. По признакам, которые являются идентификаторами, сгенерирован признак количества уникальных значений, которые встречались среди наблюдений по ТМ [28].

Таким образом, получены новые данные, для которых можно выделить ТМ с суммами числовых признаков, являющихся выбросами (*рисунок 9*).

Необходимо отметить, что предложенный подход, основанный на группировке данных и выполнении статистического теста, наиболее эффективен для отсева сильно искаженных параметров и значительно менее эффективен для выявления слабо искаженных параметров. Поэтому на данном этапе исследования (до разработки системы

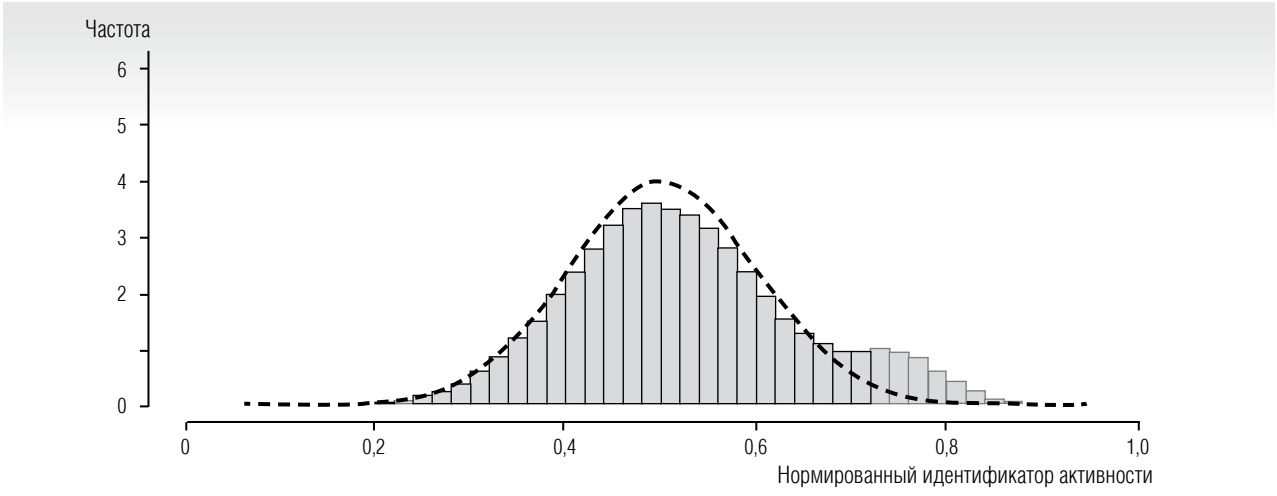

Рис 8. Результат искажения идентификатора активности на 10%

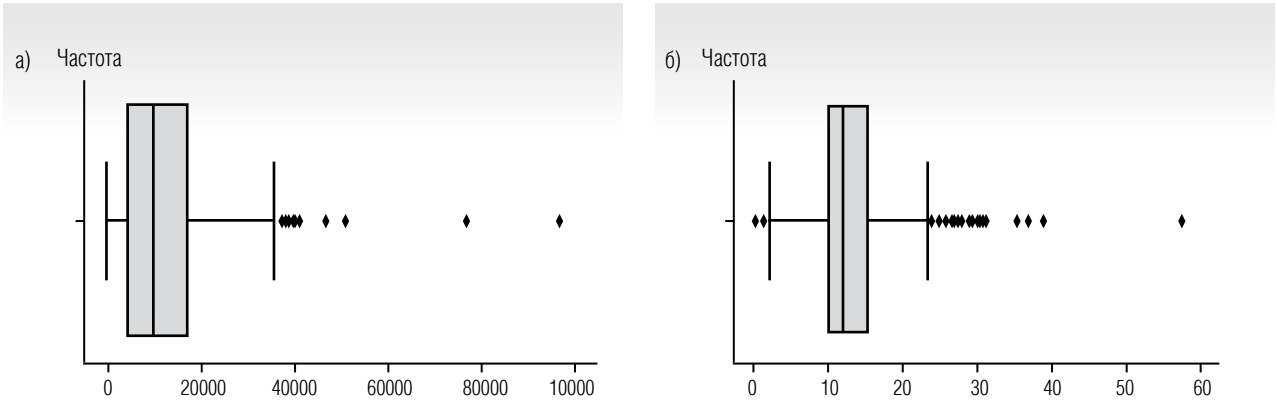

Рис. 9. Выбросы признака:
а) договоренность по бренду, усл. ед.; б) длительность активности, дни

статистических тестов или отдельной математической модели выявления слабо искаженных параметров) предлагается оценивать ключевые показатели эффективности ТМ по интегральной характеристике, учитывающей как результаты, полученные в соответствии с предложенным подходом, так и экспертные оценки. Использование комбинации статистических и экспертных (например, двойные визиты) оценок в рамках пилотного проекта позволит разметить имеющиеся данные с учетом слабо искаженных параметров и уточнить методику.

Заключение

В результате проведенной разведки данных поставлена статистическая гипотеза о возможности выявления ТМ, искажающих количество и параметры встреч. Для выявления скрытых взаимосвязей создан набор синтетических целых, действительных и категориальных переменных. Выявлены сомнительные данные (например, работа в выходные дни или ночью). Полученный обобщенный набор данных сгруппирован по признаку идентификатор активности территориального менеджера и построено распределение признака. Проведено вероятностное моделирование активности территориальных менеджеров для оценки уровня искажения информации. По каждому менеджеру просуммированы целые и действительные признаки и выявлены выбросы, характеризующие неэффективную работу или искажение данных. Для снижения ошибки второго рода предложено при разметке имеющихся данных использовать комбинацию статистических и экспертных оценок. ■

Литература

1. Bode C., Herzog C., Hook D., McGrath R. Dimensions Report. Cambridge, MA: Digital Science, 2018.
2. Димитриади Н.А. Эффективность предпринимательского проекта: аудит систем управления продажами // Учет и статистика. 2007. № 2 (10). С. 142–147.
3. Баева О.Н., Хомякова С.Г. Управление удаленными работниками: опыт фармацевтических компаний // Baikal Research Journal. 2015. Т. 6. № 5. С. 7. DOI: 10.17150/2411-6262.2015.6(5).18.
4. Баранов Р.А. Интеллектуальное планирование работы медицинских представителей // Цифровая экономика. 2018. [Электронный ресурс]: https://www.comnews.ru/digital-economy/content/116565/2018-12-10/intellektualnoe-planirovanie-raboty-medicinskih-predstaviteley (дата обращения: 22.11.2020).
5. Iljashenko O., Bagaeva I., Levina A. Strategy for establishment of personnel KPI at health care organization digital transformation // IOP Conference Series: Materials Science and Engineering. 2019. No 497. Article ID 012029. DOI: 10.1088/1757-899X/497/1/012029.
6. Sarker A., Shamim S.M., Zaman Md.S., Rahman Md.M. Employee's performance analysis and prediction using k-means clustering & decision tree algorithm // Global Journal of Computer Science and Technology: Software and Data Engineering. 2018. Vol. 18. No 1. P. 1–6.

7. Pentland A., Liu A. Modeling and prediction of human behavior // Neural Computation. 1999. Vol. 11. No 1. P. 229–242.

8. Roubtsova E., Michell V. A method for modeling of KPIs enabling validation of their properties // Proceedings of the 5th ACM SIGCHI Annual International Workshop on Behaviour Modelling – Foundations and Applications. Montpellier, France, 2 July 2013. Article ID 3. DOI: 10.1145/2492437.2492440.

9. Calixto N., Ferreira J. Salespeople performance evaluation with predictive analytics in B2B // Applied Sciences. 2020. Vol. 10. No 11. Article ID 4036. DOI: 10.3390/app10114036.

10. Владова А.Ю. Кластерный анализ изменений пространственного положения трубных секций магистрального нефтепровода по данным внутритрубных обследований // Безопасность труда в промышленности. 2018. № 1. С. 22–25.

11. Larose D.T., Larose C.D. Data mining and predictive analytics. Wiley, 2015.

12. Leskovec J., Rajaraman A., Ullman J.D. Mining of massive datasets. Stanford, CA: Stanford InfoLab, 2014.

13. VanderPlas J. Python data science handbook. O'Reilly Media, 2016.

14. Joshi P. Artificial intelligence with Python. Packt, 2017.

15. Brownlee J. Introduction to time series forecasting with Python: how to prepare data and develop models to predict the future. Machine Learning Mastery, 2017.

16. Ko I., Chang H. Interactive visualization of healthcare data using Tableau // Healthcare Informatics Research. 2017. Vol. 23. No 4. P. 349–354. DOI: 10.4258/hir.2017.23.4.349.

17. Audric S., De Bellefon M.-P., Durieux E. Descriptive spatial analysis / Handbook of spatial analysis. Theory and practical application with R // Insee Methodes. 2018. No 131. P. 3–30.

18. Brink H., Richards J.W., Fetherolf M. Real-world machine learning. Manning, 2016.

19. Ismail A. How to use Pandas-Profiling on Google Colab / Python in Plain English, 2020. [Электронный ресурс]: https://python.plainenglish.io/how-to-use-pandas-profiling-on-google-colab-e34f34ff1c9f (дата обращения 31.10.2020).

20. Skiena S.S. The data science design manual. Springer, 2017.

21. Вадзинский Р. Статистические вычисления в среде Excel. СПб: Питер, 2008.

22. Brownlee J. Statistical methods for machine learning. Machine Learning Mastery, 2020.

23. Денежкина И.Е., Зададаев С.А. Проверка статистических гипотез с использованием средств визуализации в среде RStudio // Системный анализ в экономике – 2018. Сборник трудов V Международной научно-практической конференции-биеннале. Москва, 21–23 ноября 2018 г. Под общ. ред. Г.Б. Клейнера, С.Е. Щепетовой. С. 181–184.

24. Соловьев В.И. Анализ данных в экономике: теория вероятностей, прикладная статистика, обработка и визуализация данных в Microsoft Excel. М.: КноРус, 2019.

25. Лукашин Ю.П. Адаптивные методы краткосрочного прогнозирования временных рядов. М.: Финансы и статистика, 2003.

26. Flach P. Machine learning: The art and science of algorithms that make sense of data. Cambridge: Cambridge University Press, 2012.

27. McKinney W. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. O'Reilly Media, 2012.

28. Владова А.Ю., Владов Ю.Р. Диджитализация маркетинговых кампаний // Материалы 2-й Международной научно-практической конференции «Цифровая трансформация промышленности: тенденции, управление, стратегии – 2020». Екатеринбург, 27 ноября 2020 г. С. 67–74.

Об авторах

Владова Алла Юрьевна

доктор технических наук;

ведущий научный сотрудник, Институт проблем управления им. В.А. Трапезникова РАН, 117997, г. Москва, ул. Профсоюзная, д. 65;

профессор департамента математики, Финансовый университет при Правительстве Российской Федерации, 125993, г. Москва, Ленинградский проспект, д. 49;

E-mail: ayvladova@fa.ru

ORCID: 0000-0002-8556-3798

Шек Елена Дмитриевна

студентка, Российский экономический университет им. Г.В. Плеханова, 117997, г. Москва, Стремянный переулок, д. 36;

E-mail: 1399selena@gmail.com

Data preprocessing for machine analysis of sales representatives' key performance indicators

Alla Yu. Vladova [a,b]
E-mail: ayvladova@fa.ru

Elena D. Shek [c]
E-mail: 1399selena@gmail.com

[a] V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences
Address: 65, Profsoyuznaya Street, Moscow 117997, Russia

[b] Financial University under the Government of the Russian Federation
Address: 49, Leningradsky Prospect, Moscow 125993, Russia

[c] Plekhanov Russian University of Economics
Address: 36, Stremyanny Lane, Moscow 117997, Russia

Abstract

Significant transformation of the operational activity of product and service distributors is driven by changes in data-receiving and processing technology. At present, the work of these companies' representatives is digitized to a large extent: for example, the road time, the number and places of meetings with customers are automatically recorded. At the same time, the productivity of managers who do not make direct sales is usually evaluated with the help of surveys, experts and costly double visits, although the existence of large data samples makes possible the use of statistical analysis to identify both insufficient and inflated values of performance indicators. Source data: a relational database that accumulates information about 28 categorical, quantitative, geolocation and temporal parameters of sale representatives' activities for the last year. Based on available data, we created synthetic features (the latitude and longitude features produced the index, region, street, and house features; based upon identifiers we calculated the sum of activities of sales representatives; according to temporary features we defined the season of the year, the day of the week and the period of day features). The methodology for statistical analysis consists of three main stages: collection and processing of primary data; summary and grouping processed information; setting statistical hypotheses and interpreting the results. A probabilistic approach was used to model the level of distortion of sale representatives' activities. As a result, with the built tag cloud we highlighted: the most popular season for advertising campaigns; the most productive departments and sale representatives; days of the week with the largest number of contacts to customers. We established a significant number of records about meetings with clients at the weekends. As a result of the data mining, we made a statistical hypothesis about the possibility of identifying the sale representatives who distort the number and parameters of meetings. A set of synthetic integer, real and categorical features was created to identify hidden relationships. Doubtful data (such as working at weekends or at night) were revealed. The resulting aggregated dataset is grouped by a sale representative's activity ID and the distribution of this feature is plotted. For each sale representative, integer and real features are summarized and outliers that characterize inefficient performance or distortion of data have been detected. Thus, the presence of a large sample of data on the history of movements and activities allowed us to evaluate the productivity of the distribution company's sales representatives based upon indirect features.

Key words: machine learning; key performance indicator; database; time series; geolocation; unsupervised learning; sales representative; b2b.

Citation: Vladova A.Yu., Shek E.D. (2021) Data preprocessing for machine analysis of sales representatives' key performance indicators. *Business Informatics*, vol. 15, no 3, pp. 48–59. DOI: 10.17323/2587-814X.2021.3.48.59

References

1. Bode C., Herzog, C., Hook D., McGrath R. (2018) *Dimensions Report.* Cambridge, MA: Digital Science.
2. Dimitriadi N.A. (2007) Efficiency of an entrepreneurial project: audit of sales management systems. *Accounting and Statistics*, no 2, pp. 142–147 (in Russian).

3. Baeva O.N., Khomyakova S.G. (2015) Managing remote workers: the experience of pharmaceutical companies. *Baikal Research Journal*, vol. 6, no 5, pp. 7. DOI: 10.17150/2411-6262.2015.6(5).18 (in Russian).

4. Baranov R.A. (2018) Intelligent planning of the work of medical representatives. *Digital Economy*. Available at: https://www.comnews.ru/digital-economy/content/116565/2018-12-10/intellektualnoe-planirovanie-raboty-medicinskih-predstaviteley (accessed 22 November 2020) (in Russian).

5. Iljashenko O., Bagaeva I., Levina A. (2019) Strategy for establishment of personnel KPI at health care organization digital transformation. *IOP Conference Series: Materials Science and Engineering*, no 497, article ID 012029. DOI: 10.1088/1757-899X/497/1/012029.

6. Sarker A., Shamim S.M., Zaman Md.S., Rahman Md.M. (2018) Employee's performance analysis and prediction using k-means clustering & decision tree algorithm. *Global Journal of Computer Science and Technology: Software and Data Engineering*, vol. 18, no 1, pp. 1–6.

7. Pentland A., Liu A. (1999) Modeling and prediction of human behavior. *Neural Computation*, vol. 11, no 1, pp. 229–242.

8. Roubtsova E., Michell V. (2013) A method for modeling of KPIs enabling validation of their properties. Proceedings of the *5th ACM SIGCHI Annual International Workshop on Behaviour Modelling – Foundations and Applications. Montpellier, France, 2 July 2013*, article ID 3. DOI: 10.1145/2492437.2492440.

9. Calixto N., Ferreira J. (2020) Salespeople performance evaluation with predictive analytics in B2B. *Applied Sciences*, vol. 10, no 11, article ID 4036. DOI: 10.3390/app10114036.

10. Vladova A.Yu. (2018) Clustering analysis of changes in the spatial position of the trunk oil pipeline sections based on the in-line inspection datasets. *Occupational Safety in Industry*, no 1, pp. 22–25 (in Russian).

11. Larose D.T., Larose C.D. (2015) *Data mining and predictive analytics*. Wiley.

12. Leskovec J., Rajaraman A., Ullman J.D. (2014) *Mining of massive datasets*. Stanford, CA: Stanford InfoLab.

13. VanderPlas J. (2016) *Python data science handbook*. O'Reilly Media.

14. Joshi P. (2017) *Artificial intelligence with Python*. Packt.

15. Brownlee J. (2017) *Introduction to time series forecasting with Python: how to prepare data and develop models to predict the future*. Machine Learning Mastery.

16. Ko I., Chang H. (2017) Interactive visualization of healthcare data using Tableau. *Healthcare Informatics Research*, vol. 23, no 4, pp. 349–354. DOI: 10.4258/hir.2017.23.4.349.

17. Audric S., De Bellefon M.-P., Durieux E. (2018) Descriptive spatial analysis. *Handbook of spatial analysis. Theory and practical application with R*. Insee Methodes, no 131, pp. 3–30.

18. Brink H., Richards J.W., Fetherolf M. (2016) *Real-world machine learning*. Manning.

19. Ismail A. (2020) How to use Pandas-Profiling on Google Colab. *Python in Plain English*. Available at: https://python.plainenglish.io/how-to-use-pandas-profiling-on-google-colab-e34f34ff1c9f (accessed 31 October 2020).

20. Skiena S.S. (2017) *The data science design manual*. Springer.

21. Vadzinsky R. (2018) *Statistical calculations in the Excel environment*. Saint Petersburg: Piter (in Russian).

22. Brownlee J. (2020) *Statistical methods for machine learning*. Machine Learning Mastery.

23. Denezhkina I.E., Zadadaev S.A. (2018) Testing statistical hypotheses using visualization tools in the R Studio environment. Proceedings of the *V International Scientific and Practical Conference "System Analysis in Economics". Moscow, Russia, 21–23 November* 2018, pp. 181–184 (in Russian).

24. Soloviev V.I. (2019) *Data analysis in economics: probability theory, applied statistics, data processing and visualization in Microsoft Excel*. Moscow: KnoRus (in Russian).

25. Lukashin Yu.P. (2003) *Adaptive methods of short-term time series forecasting*. Moscow: Finance and Statistics (in Russian).

26. Flach P. (2012) *Machine learning: The art and science of algorithms that make sense of data*. Cambridge: Cambridge University Press.

27. McKinney W. (2012) *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O'Reilly Media.

28. Vladova A.Yu., Vladov Yu.R. (2020) Digitalization of marketing campaigns. Proceedings of the *2-nd International Scientific and Practical Conference "Digital Transformation of Industry: Trends, Management, Strategies – 2020". Ekaterinburg, Russia, 27 November 2020*, pp. 67–74 (in Russian).

About the authors

Alla Yu. Vladova

Dr. Sci. (Tech.);

Leading Researcher, V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, 65, Profsoyuznaya Street, Moscow 117997, Russia;

Professor, Department of Mathematics, Financial University under the Government of the Russian Federation, 49, Leningradsky Prospect, Moscow 125993, Russia;

E-mail: ayvladova@fa.ru

ORCID: 0000-0002-8556-3798

Elena D. Shek

Student, Plekhanov Russian University of Economics, 36, Stremyanny Lane, Moscow 117997, Russia;

E-mail: 1399selena@gmail.com