

УДК 004.032

DOI: 10.25559/SITITO.14.201804.911-922

ИССЛЕДОВАНИЕ СКЛОННОСТИ ПОЛЬЗОВАТЕЛЯ ИНТЕРНЕТ-МАГАЗИНА К ПОКУПКЕ НА ОСНОВЕ ТЕХНИЧЕСКИХ ДАННЫХ О ВИЗИТАХ ПОСЕТИТЕЛЕЙ ИНТЕРНЕТ-МАГАЗИНА

Д.Г. Лагерев, И.А. Савостин, В.Ю. Герасимчук, М.С. Полякова
Брянский государственный технический университет, г. Брянск, Россия

THE STUDY OF THE PURCHASE PROPENSITY OF THE USER OF THE ONLINE STORE ON THE BASIS OF TECHNICAL DATA ON VISITS OF VISITORS TO THE ONLINE STORE

Dmitri G. Lagerev, Igor A. Savostin, Vyacheslav Yu. Gerasimchuk, Marina S. Polyakova
Bryansk State Technical University, Bryansk, Russia

© Лагерев Д.Г., Савостин И.А., Герасимчук В.Ю., Полякова М.С., 2018

Ключевые слова

ИАД; скоринг; ассоциативные
правила; логистическая
регрессия; временные ряды.

Аннотация

В статье описывается разработка сервиса автоматизации проведения маркетинговых кампаний на основе интеллектуального анализа данных о посещениях сайта. Исходные данные о визитах пользователей были получены из сервиса «Яндекс.Метрика». Для выполнения интеллектуального анализа данных было необходимо выполнить достаточно сложную предобработку. На основе разведочного анализа данных было принято решение разделить работу на следующие этапы: анализ, построение ансамбля моделей, оценка эффективности моделей и выбор лучшей, публикация веб-сервиса с целью дальнейшего использования в CMS интернет-магазина при проведении маркетинговых кампаний.

Интеллектуальный анализ данных выполнялся на основе аналитической платформы Loginom [5], которая позволяет как создавать аналитические модели, так и представлять их в виде веб-сервисов. Для интеллектуального анализа данных использовались три подхода: ассоциативные правила, временные ряды и скоринг. Наилучшие результаты показала скоринговая модель, она отлично себя показала на тестовом множестве, при этом экспертный анализ коэффициентов подтверждает её корректность и применимость. При необходимости разработанная аналитическая модель может быть обучена на данных о посещаемости любого интернет-магазина, предоставленного сервисом «Яндекс.Метрика».

Об авторах:

Лагерев Дмитрий Григорьевич, кандидат технических наук, доцент кафедры информатики и программного обеспечения, Брянский государственный технический университет (241035, Россия, г. Брянск, ул. Харьковская, д. 10-Б), ORCID: <http://orcid.org/0000-0002-2702-6492>, lagerevdlg@mail.ru

Савостин Игорь Анатольевич, магистрант, кафедра информатики и программного обеспечения, Брянский государственный технический университет (241035, Россия, г. Брянск, ул. Харьковская, д. 10-Б), ORCID: <http://orcid.org/0000-0002-4846-2285>, iansav7@gmail.com

Герасимчук Вячеслав Юрьевич, магистрант, кафедра информатики и программного обеспечения, Брянский государственный технический университет (241035, Россия, г. Брянск, ул. Харьковская, д. 10-Б), ORCID: <http://orcid.org/0000-0003-3165-1917>, slava-gerasimchuk@rambler.ru

Полякова Марина Сергеевна, магистрант, кафедра информатики и программного обеспечения, Брянский государственный технический университет (241035, Россия, г. Брянск, ул. Харьковская, д. 10-Б), ORCID: <http://orcid.org/0000-0003-1992-4839>, JeanGM@yandex.ru



Keywords

IAD; scoring; association rules; logistic regression; time series.

Abstract

The paper describes the development of an automatization service for marketing campaigns on the basis of web usage data mining. The source web usage data were obtained from Yandex.Metrica service. On the basis of exploratory data analysis it was decided to divide the work into the following stages: analysis, building an assembly of models, assessing the efficiency of the models and selecting the best one, and publishing the web service with the purpose of its further usage in the Internet shop CMS within marketing campaigns. The data mining was conducted on the basis of Loginom Analytics Platform which allows to create analytic models as well as to present them out as web services. For the data mining 3 approaches - association rules, time series and scoring - were used. The scoring model demonstrated the best results, it showed excellent results on test data, with that the coefficient expert analysis confirms its correctness and applicability. In case of necessity the developed analytic model can be learned on web usage data of any Internet shop provided by Yandex.Metrica service.

Введение

Современную жизнь трудно представить без интернет-магазинов. Пользователи интернета привыкли совершать покупки онлайн, сравнивая характеристики и читая отзывы, не выходя из дома.

Однако разработка сайта, аренда помещений, поисковая оптимизация и продвижение, поддержка его работы, синхронизация баз данных, работа склада и служб, а также доставка требуют вложений [23]. Очевидно, что существование магазина зависит как от вышеуказанных расходов, так и от доходов от продаж, которые любой магазин стремится максимизировать. Максимизация включает в себя рекламу, оптимизацию самого сайта и поисковую оптимизацию, систему скидок, которые мотивируют клиентов совершать покупки именно в этом магазине.

Цель исследования

В данной работе рассмотрена проблема определения склонности пользователя интернет-магазина к покупке, что может быть использовано для определения целесообразности предложения ему скидки. В дальнейшем, это может мотивировать его к покупкам в данном магазине.

Теоретический анализ

Было решено выполнить ИАД: построить ансамбль моделей в Deductog по выборке, накопленной за несколько месяцев работы магазина.

1. Основные этапы проведения исследования;
2. Получение данных.
3. Трансформация в пригодное для обработки состояние.
4. Загрузка в ПО для анализа.
5. Построение моделей.
6. Сравнение моделей и отбор лучших результатов.
7. Выводы.

После получения и предобработки данных будет производиться интеллектуальный анализ по трём направлениям: ассоциативные правила, временные ряды и скоринг [15, 17].

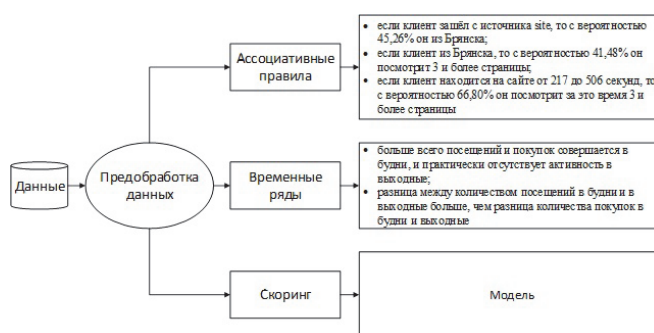


Рис. 1. Процесс получения результатов
Fig 1. The process of obtaining results

Получение данных

Исходные данные были предоставлены фирмой, деятельность которой связана с SEO-оптимизацией [22]. Вся информация была получена специалистами организации с помощью сервиса Яндекс.Директ и его инструмента – Яндекс.Метрика. «Яндекс.Директ» — система контекстной рекламы на страницах «Яндекса» и сайтах партнеров Рекламной системы Яндекса. «Яндекс.Метрика» — бесплатный интернет-сервис компании Яндекс, предназначенный для оценки посещаемости веб-сайтов и анализа поведения пользователей. На данный момент Яндекс.Метрика является второй по размеру системой веб-аналитики в Европе [1].

Подготовка данных для обработки

Для любых методов интеллектуального анализа данных огромную роль играет качество исходных данных. Наиболее частыми проблемами, с которыми приходится сталкиваться специалистам являются: пропуски, выбросы, экстремальные значения, а также неудобная или непригодная для обработки форма представления [24].

В имеющейся выборке по интернет-магазину присутствуют два файла — данные визитов и данные клиентов. Данные визитов имеют следующие столбцы:



Таблица 1.1. Столбцы данных о визитах

Table 1.1. Visitor Data Columns

ym:pv:watchID — идентификатор просмотра	ym:pv:title — заголовок страницы
ym:pv:URL — адрес страницы	ym:pv:referrer — источник трафика
ym:pv:browser — браузер	ym:pv:browserCountry — страна
ym:pv:deviceCategory — тип устройства	ym:pv:regionCity — город
ym:pv:lastTrafficSource — тип источника трафика	ym:pv:lastSearchEngineRoot — поисковая система
ym:pv:lastAdvEngine — рекламный движок	ym:pv:link — флаг перехода по ссылке
ym:pv:download — флаг загрузки	ym:pv:lastSocialNetwork — социальная сеть
ym:pv:clientID — идентификатор пользователя	ym:pv:lastSocialNetworkProfile — профиль в социальной сети
ym:pv:goalsID — достигнутые пользователем цели	

Данные клиентов, в свою очередь, состоят из:

Таблица 1.2. Столбцы данных о клиентах

Table 1.2. Customer data columns

ym:s:visitID — идентификатор визита	ym:s:dateTime — дата и время
ym:s:isNewUser — флаг: новый пользователь, или нет	ym:s:startURL — точка входа
ym:s:endURL — точка выхода	ym:s:pageViews — количество просмотренных страниц
ym:s:visitDuration — длительность визита	ym:s:bounce — флаг отказа сайта
ym:s:regionCity — город	ym:s:goalsID — идентификаторы целей
ym:s:goalsSerialNumber — номера целей в серии	ym:s:lastTrafficSource — последний тип источника трафика
ym:s:lastReferralSource — реферальный источник трафика	ym:s:lastSearchEngineRoot — поисковая система
ym:s:lastSocialNetwork — социальная сеть	ym:s:deviceCategory — тип устройства
ym:s:browser — браузер	

Было произведено удаление столбцов с информацией об адресе с названием страниц и с номером цели в серии, поскольку с точки зрения интеллектуального анализа они не содержат какой-либо полезной информации. После вышеописанных действий становится неясной необходимость в файле визитов для скоринговой модели, поэтому было принято решение для интеллектуального анализа далее работать только с файлом клиентов.

Полученные данные представляют собой два текстовых csv-файла [2], с разделителями - «;», записями, содержащими данный символ, что затрудняет импорт строк, имеющих пустые значения. Также были обнаружены проблемы при импорте поля «refereg», которое содержало длинные ссылки с символом разделителя, что воспринималось обработчиком как новый столбец. Так как в самой ссылке для скоринговой модели смысловую нагрузку не нес только домен, было принято решение избавляться от лишних частей адреса в данном поле.

В связи со всеми вышеперечисленными проблемами становится ясно, что прежде чем осуществить импорт данных из файлов в Deductor [3] (для дальнейшего анализа) было принято решение произвести подготовку данных. Исходя из знаний и

привычек автора, для ликвидации вышеописанных проблем были использованы стандартные инструменты командной строки Linux: bash, sed, awk, grep, nl, head, tail, cat и другие.

В итоге, для приведения в порядок файла визитов была сформирована команда из листинга 1.1:

```
cat visits.csv | awk -F ';' '{print $1»»$4»zzz»»$(NF-12)»»$(NF-11)»»»$(NF-10)»»»$(NF-9)»»»$(NF-8)»»»$(NF-7)»»»$(NF-6)»»»$(NF-5)»»»$(NF-4)»»»$(NF-2)»»»$(NF+1-1)»»}' | sed -r 's/\.+zzz//g' | sed 's/zzz//g' | sed -r 's/https?:\//g' | awk -F ';' '{ for(i = 1; i < NF; i++) { if( $i == "" ) { printf "null;" } else { printf "%s;" $i } }; printf "\n" }' | sed -r 's/ym:pv://g' > visits_csv
```

Листинг 1.1. Обработка файла визитов

Listing 1.1. Visits file processing

Аналогичным образом был обработан файл клиентов.

После выполнения вышеперечисленных обработок были получены файлы, которые были импортированы сначала в табличный процессор (листинг 1.1) для наглядной оценки результатов, а затем и в Deductor.

Таблица 1.3. Предобработанные входные данные

Table 1.3. Pre-processed input

Watch ID	referrer	browser	browser Country	Device Category	regionCity	Last Traffic Source	Last Search Engine Root	last Adv Engine	link	download	Client ID	Goals ID
1	google	chrome	ru	1	Chita	organic	google	null	0	0	1	null
2	site	chrome	ru	1	Chita	internal	null	null	0	0	2	null
3	site	safari_mobile	ru	2	Moscow	internal	null	null	0	0	3	null
4	null	safari_mobile	ru	2	Moscow	direct	null	null	0	0	4	null
5	site	safari_mobile	ru	2	Moscow	internal	null	null	0	0	5	[1, 5]
6	site	safari_mobile	ru	2	Moscow	internal	null	null	0	0	6	null
7	null	safari_mobile	ru	2	Moscow	direct	null	null	0	0	7	null
8	site	safari_mobile	ru	2	Moscow	internal	null	null	0	0	8	null
9	site	safari_mobile	ru	2	Moscow	external	null	null	1	1	9	null
10	author	yandex_browser	null	1	Saint Petersburg	referral	null	null	0	0	10	null



Если на данном этапе загрузить данные в Deductor, то можно увидеть, что столбец с целями становится малоинформативным из-за своего формата. Очевидно, что достигнутые пользователем цели нужно как-то конкретизировать, чтобы получить более эффективные результаты при использовании временных рядов и скоринга.

Для определения, необходимых для анализа целей, был применен достаточно простой подход. Исходя из таблицы 1.3, в качестве события принимается достижение целей из воронки заказов, т. е., начинающиеся с 260. Остальные будут считаться несобытием.

Таблица 1.4. Описание целей исходного ресурса
Table 1.4. Description of the source resource objectives

Описание цели	Идентификатор	Идентификатор (воронка продаж)
Переход в корзину (url содержит cart.php)	1592926	26021419
Просмотр 3+ страниц	16225015	
Просмотр контактов	16225020	
Кнопка купить		26021414
Оформление заказа		26021424

Ассоциативные правила

Перед тем, как применить скоринговую модель для данной задачи, было так же проведено исследование, суть которого заключалась в поиске ассоциативных правил, с целью выявления каких-либо закономерностей в исходных данных [13, 15].

Ассоциативные правила - правила, служащие для количественного описания взаимной связи между двумя или более событиями. Впервые задача поиска ассоциативных правил была предложена для нахождения типичных шаблонов покупок, совершаемых в супермаркетах, поэтому иногда ее еще называют анализом рыночной корзины (market basket analysis) [7].

Ассоциативное правило формируется в виде: «Если условие, то следствие». Условие и следствие часто называются соответственно левосторонним (left-hand side - LHS) и правосторонним (right-hand side - RHS) компонентами ассоциативного правила.

Поддержка – это число транзакций, которые содержат как условие, так и следствие:

$S(A \rightarrow B)$ = количество транзакций, содержащих A и B/ общее количество транзакций.

Достоверность – мера точности правила, определяющаяся как отношение количества транзакций, содержащих и условие, и следствие, к количеству транзакций, содержащих только условие:

$C(A \rightarrow B)$ = количество транзакций, содержащих A и B/ количество транзакций, содержащих только A.

Алгоритмы поиска ассоциативных правил предназначены для нахождения всех правил $X \rightarrow Y$, причем поддержка и достоверность этих правил должны быть выше некоторых наперед определенных порогов, называемых соответственно минимальной поддержкой (minsupport) и минимальной достоверностью (minconfidence) [6].

Алгоритм Apriori был предложен в 1994 г. (Agrawal и Srikant, исследовательская группа Almaden IBM) и в настоящее время является одним из наиболее популярных алгоритмов поиска ассоциативных правил [7].

Простой алгоритм Apriori состоит из трех шагов:

1. Объединение. Просмотр базы данных и определение частоты вхождения отдельных товаров;
2. Отсечение. Те наборы, которые удовлетворяют поддержке и достоверности, переходят на следующую итерацию с двухкомпонентными наборами;
3. Повторение. Предыдущие два шага повторяются для каждой величины набора, пока не будет повторно получен ранее определенный размер.

Для поиска ассоциативных правил было подготовлено три таблицы на основе исходных данных: таблица с визитами, таблица с информацией о посетителях и таблица, полученная слиянием двух предыдущих. По таблице с визитами были выявлены правила, которые не могут быть объяснены. Как правило, причиной тому могут служить либо аномальные значения, либо скрытые знания. Такие правила нельзя напрямую использовать для принятия каких-либо решений. Здесь необходим дополнительный анализ.

В двух других таблицах были найдены следующие правила:

- если клиент зашёл с источника site, то с вероятностью 45,26% он из Брянска;
- если город клиента, посетившего сайт, - Брянск, то с вероятностью 48,13% он зашёл на сайт с источника site;
- если клиент из Брянска, то с вероятностью 41,48% он посмотрит 3 и более страницы;
- если клиент находится на сайте от 217 до 506 секунд, то с вероятностью 66,80% он посмотрит за это время 3 и более страницы.

Очевидно, что полученные ассоциативные правила, к сожалению, не представляют какой-либо практической пользы. Возможно, результаты могли быть куда более информативными, если бы в исходных данных было бы больше полей с дополнительной информацией о пользователе.

Анализ временных рядов

Помимо поиска ассоциативных правил, была предпринята попытка поиска зависимостей в данных, имеющих временную привязку, т.е. попытаться найти связь между днями недели и количеством посещений [16]. Для этой задачи больше всего подходит анализ и декомпозиция временных рядов [20, 25].

Временной ряд [4] - последовательность наблюдений за изменениями во времени значений параметров (атрибутов, признаков, показателей) некоторого объекта или процесса.

Для анализа временных рядов посещаемости сайта не имеет значения тип устройства, браузер и откуда пользователь перешел на сайт. Значимыми являются только значения, определяющие количество посещений в тот или иной момент времени, так называемая временная шкала и количество достижений конечной цели, то есть количество покупок, совершенных в определенных промежутки времени. Поэтому для анализа временных рядов использовался только файл с данными о клиентах.

Поскольку дата и время в чистом виде не представляют интереса для анализа, они были преобразованы в дни недели с помощью обработчика «Преобразование дат» [9]. После преобразования даты в дни недели следует анализ количества посещений и покупок в тот или иной день для выявления сезонности и тренда, что поможет спрогнозировать спрос на продукцию с целью распределения различных акций по времени, бла-



годаря чему можно будет максимизировать прибыль компании. Для анализа полученных ранее данных используется обработчик «декомпозиция временного ряда».

В ходе декомпозиции была получена диаграмма, четко демонстрирующая сезонность в течение недели, и линии тренда демонстрируют посещаемость и покупки в течение анализируемого промежутка времени (рис. 2).

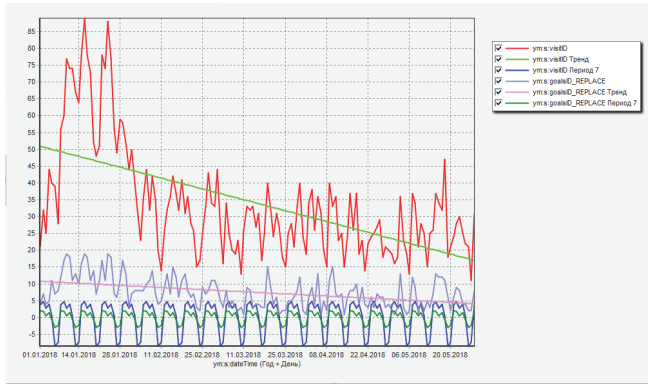


Рис. 2. Диаграмма декомпозиции временного ряда
Fig. 2. Time series decomposition diagram

Исходя из диаграммы декомпозиции [8] видно, что больше всего посещений и покупок на сайте совершается в будни: больше всего посещений во вторник и в четверг, меньше – в понедельник и в среду, достаточно мало – в пятницу, и практически отсутствует активность в выходные. В то же время разница между количеством посещений в будни и в выходные больше, чем разница количества покупок в будни и выходные. Наклон линий тренда показывает, что число посещений в рассмотренный промежуток постепенно уменьшается от начала к концу, однако при этом число покупок на сайте уменьшается куда меньше (лишь несущественно понижается) и остается к концу промежутка незначительно меньшим, чем в начале рассмотренного промежутка времени. Таким образом полученные данные позволят маркетологам лучше подобрать моменты для проведения акций и даже распределить по дням недели, какую рекламу размещать на сайте.

В то же время разница между количеством посещений в будни и в выходные больше, чем разница количества покупок в будни и выходные. Наклон линий тренда показывает, что число посещений в рассмотренный промежуток постепенно уменьшается от начала к концу, однако при этом число покупок на сайте уменьшается несущественно и остается к концу промежутка незначительно меньшим, чем в начале рассмотренного промежутка времени. Таким образом полученные данные позволят маркетологам лучше подобрать моменты для проведения акций и даже распределить по дням недели, какую рекламу размещать на сайте.

Задача классификации и скоринг

Перед тем, как была разработана вышеописанная модель, было так же создано ещё две, но в последствии они показали себя хуже, чем итоговая [26]. В одной из предшествующих моделей квантование времени суток было произведено вручную, удалены малозначимые, с точки зрения автора, поля, остальные - автоматически распределены в конечные классы. Полученная

таким образом модель даже на обучающей выборке показала себя хуже, чем итоговая. Другая модель была сформирована почти как итоговая, но без корректировок конечных классов. Следует отметить, что коэффициенты остались в допустимых пределах, несмотря на повышение информационного индекса. Также довольно удачным оказалось и автоматическое определение конечных классов. Но в таком случае при получении каких-то дополнительных знаний, необходимо более тщательно анализировать результат построения модели. Таким образом, благодаря двум предыдущим моделям автором была сформирована схема скоринга для итоговой модели (рис. 3).



Рис. 3. Схема скоринга
Fig. 3. Scoring scheme

В полученной выборке в столбцах `um:s:regionCity`, `um:s:lastReferralSource` и `um:s:lastSearchEngineRoot` были пропуски. С помощью обработчика «Замена значений» они были заменены на «unknown».

Данные в столбце, содержащим дату и время, в исходном виде не представляют большой пользы для анализа. Поэтому было принято решение разбить это поле на день недели и время суток с помощью обработчика «Преобразование даты».

Основной недостаток столбцов `um:s:visitDuration` и `um:s:pageViews` заключался в том, что они имели слишком большое множество уникальных значений. При анализе гистограммы стало понятно, что ряд значений можно разделить на интервалы. Это было сделано с помощью обработчика «Квантование» [10].

Согласно гистограммам, были определены следующие интервалы для данных полей:

- `pageViews`:
 - 1;
 - 2 — 7;
 - от 7.
- `visitDuration`:
 - 0;
 - 1 — 13;
 - 14 — 16;
 - от 16.

Когда требуется снизить количество уникальных значений поля, используется квантование — разбиение множества значений на отрезки. Одной из разновидностей квантования является автоматическое разбиение множества на конечные классы по средствам определения индексов `WoE` (Weight of Evidence). Но объединение в интервалы не подходит, и надо анализировать



вес каждого значения, поэтому был применён обработчик «Конечные классы». Он автоматически формирует классы значений, согласно их весам. Так же доступен интерактивный режим.

На вход обработчика «Конечные классы» были поданы поля `ym:s:dateTime`, `ym:s:regionCity`, `ym:s:browser`, `ym:s:lastReferralSource`, `ym:s:lastSearchEngineRoot`. Рассмотрим поле `dateTime` (часы). Данное поле содержит значения часа, в который пользователь попал в выборку. После работы обработчика и легкой ручной корректировки, получились результаты, показанные на рисунке 4. Автоматически значения распределились без учета специфики типичного распорядка дня. Ручная правка была необходима для более четкого выделения времени «обеда», времени пути на работу и с ней, времени после работы. Ночное время алгоритм определил вполне точно.

Распределение остальных вышеупомянутых полей было осуществлено автоматически и ручной корректировки не потребовало.

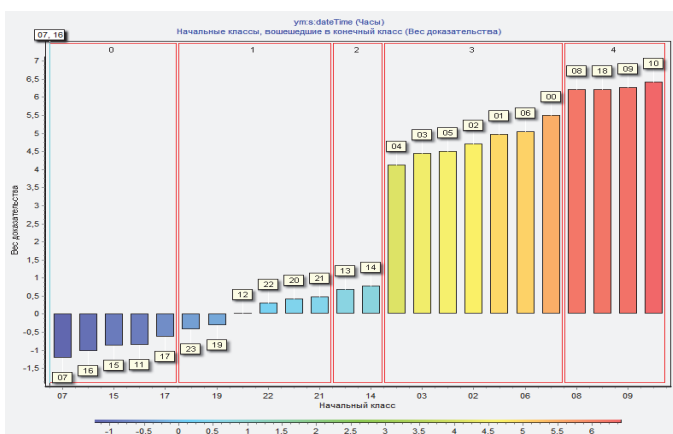


Рис. 4. Конечные классы поля `dateTime` (Часы)

Fig 4. End `dateTime` field classes (Hours)

Так как в выборке на 31 событие приходилось более пяти тысяч несобытий, было принято решение использовать отбор со смещением [18, 19]. При этом, события были отобраны в отношении 70% от оригинальной выборки. Несобытия — 10% от оригинальной выборки. Это позволило относительно снизить дисбаланс событий/несобытий. Все неотобраные строки, далее, будут добавлены в тестовую выборку. Данная операция была совершена с помощью обработчика «Сэмплинг» [11].

Для проверки корректности модели требуется тестовая выборка. Для ее формирования был применён обработчик «Слияние с узлом». Методом «Full outer join» выборки до сэмплинга и после. Объединенная выборка будет состоять из обучающей и тестовой.

Для построения скоринговой модели применялся обработчик «Логистическая регрессия» [12]. В качестве входных параметров использовалась объединенная выборка, был задан размер обучающей выборки и поправочные коэффициенты, чтобы компенсировать сэмплинг. Остальные параметры обработчика остались с настройками по умолчанию. После построения модели были внесены корректировки масштаба для учета смещения вероятностей событий/несобытий в результате сэмплинга. Для корректировки был использован стандарт FICO, который применяется в модели неявно.

Стандарт FICO был разработан американской компанией, занимающейся скорингом. Первоначальной целью этого стандарта была оценка кредитных рисков. В данном случае, так как был использован сэмплинг, то произошло смещение вероятностей. Следовательно, чтобы модель давала адекватные результаты на выходе, в последствии автоматически был применено масштабирование согласно данному стандарту.

На выходе была получена размеченная, согласно работе скоринговой модели, с двойным набором полей выборка. Дополнительно, обработчик предоставляет отчет по регрессии, ROC-кривую и прочие данные.

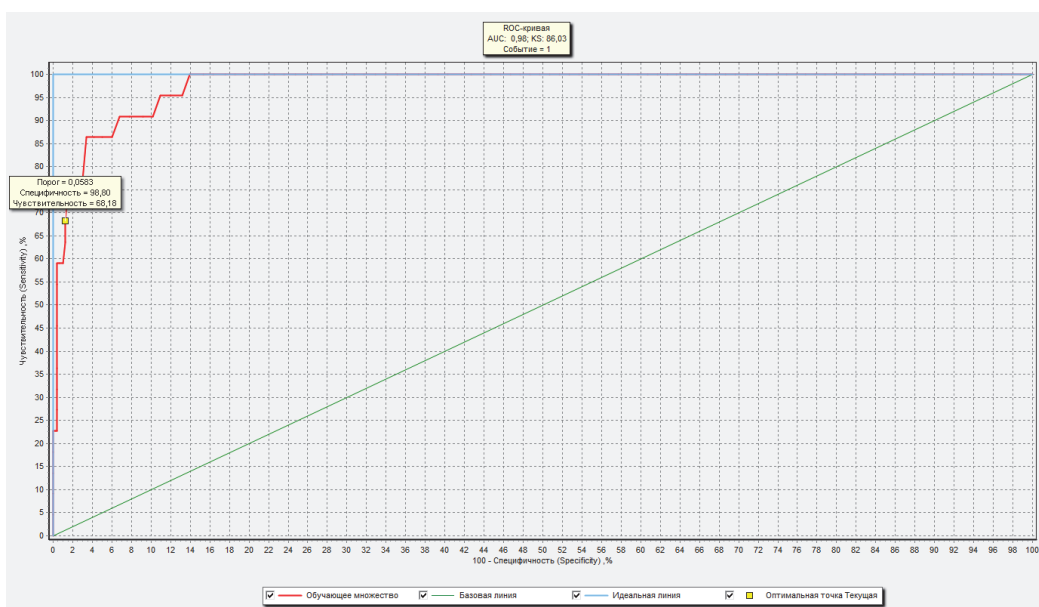


Рис. 5. ROC-кривая

Fig. 5. ROC curve



Далее, размеченная выборка была преобразована в формат, который пригоден для работы — сценария, предназначенного для оценки качества работы модели.

Рассмотрим наиболее интересные графики, предоставлен-

ные сценарием после обработки результатов скоринга: объединенные ROC-кривые (рис. 6), Lift-кривые (рис. 7), кривые распределения событий/несобытий (рис. 8).

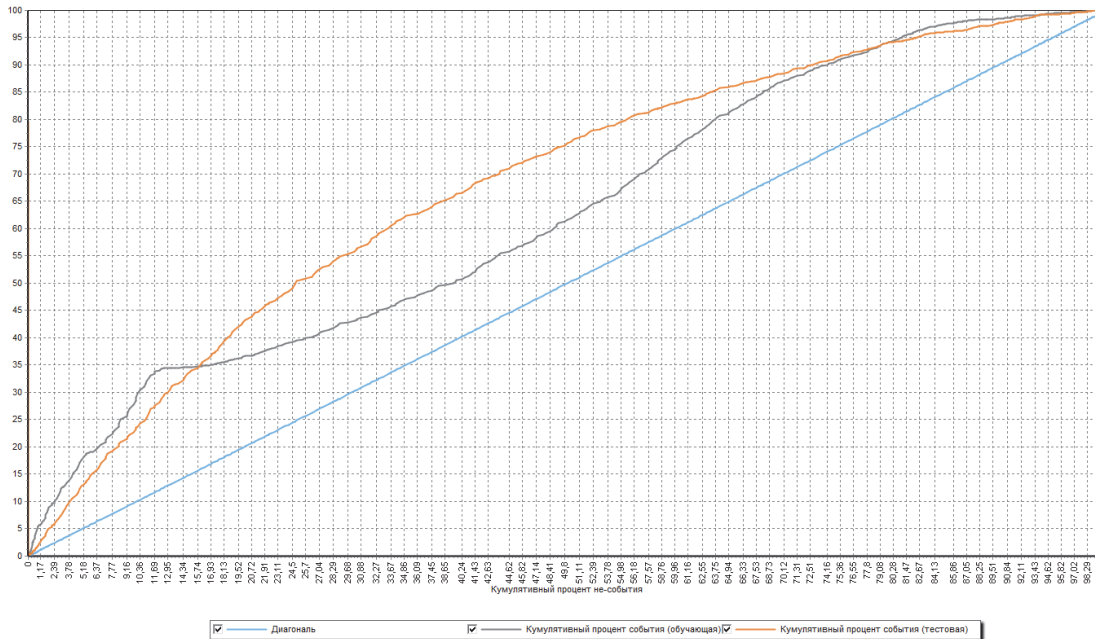


Рис. 6. Объединённые ROC-кривые

Fig. 6. Combined ROC curves

ROC-кривая на тестовой выборке в середине далеко уходит от кривой на обучающей выборке потому, что данных было очень мало, и модель просто не смогла обучиться в полной мере [14, 27]. Несмотря на этот факт, можно отметить, что начало и конец кривых совпадают. Это говорит о том, что в сложных ситу-

ациях модель может давать сбои, но в более простых она будет добросовестно работать. Следовательно, её можно использовать при автоматизации принятия решений о выдаче пользователю предложений различного рода скидок.

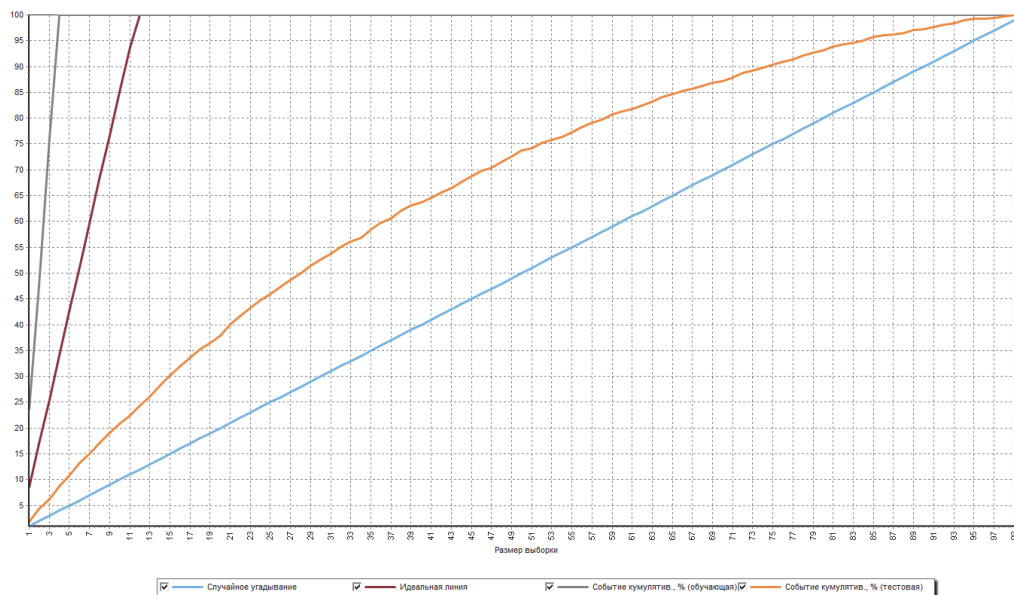


Рис. 7. Объединённые Lift-кривые

Fig 7. Combined Lift Curves



Lift-кривые показывают переобучение модели на обучающем множестве. То же показывает ROC-кривая в отчете о каче-

стве регрессии. В связи с недостаточным объемом данных модель получилась обученной под конкретные случаи.

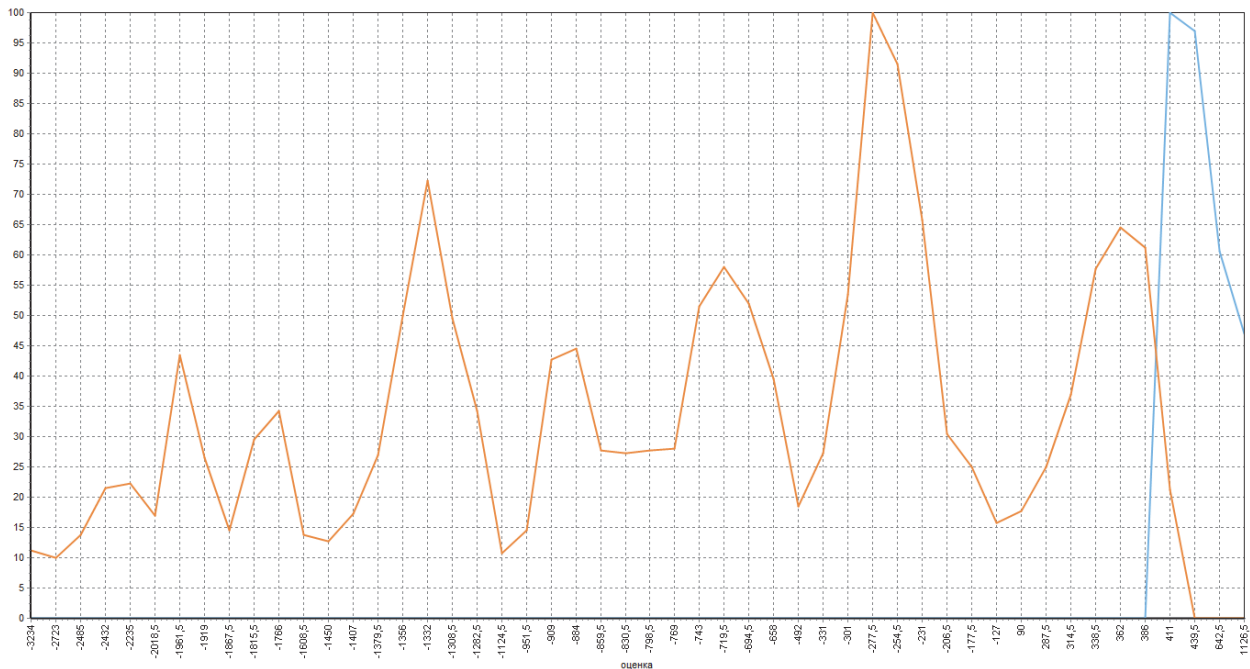


Рис. 8. Распределение событий/несобытий
Fig. 8. Event / non-event distribution

Если обратить внимание на распределение событий и несобытий, то здесь очень остро отобразилась нехватка данных, в которых на 5 тысяч записей приходилось всего 31 событие.

На основе результатов работы скоринговой модели была создана таблица, показывающая степень влияния атрибутов на итоговый скоринговый балл.

Таблица 2.1. Влияние атрибутов на скоринговый балл
Table 2.1. Impact of attributes on scoring score

Атрибут	Коэффициент	Влияние
ym:s:lastTrafficSource+	113.6022776	Сильное
ym:s:visitDuration+	81.15379678	
ym:s:dateTime (Часы) Метка+	62.35127532	
ym:s:dateTime (День недели)+	5.815291396	
ym:s:bounce+	4.291692094	Среднее
ym:s:pageViews+	2.86877764	Слабое
ym:s:isNewUser+	0.758787746	
ym:s:deviceCategory+	0.557128221	

Результаты

Согласно распределению коэффициентов логистической регрессии, можно сделать следующие выводы:

- вероятность попадания в воронку продаж больше, если пользователь зашел на сайт с компьютера, нежели чем с мобильного устройства, или, тем более, телевизора. Данное утверждение вполне логично: как правило, совершать покупки, сидя за компьютером, просто удобнее. Имеет смысл подтолкнуть пользователя к посещению сайта с компьютера;
- лучший день для покупок — такой, чтобы от заказа до получения и применения товара проходило 2-3 дня —

среднее время доставки из интернет-магазинов. Например, если товар может быть применен на выходных — наибольший спрос будет в среду;

- чем дольше пользователь на сайте — тем выше вероятность попадания в воронку продаж. Соответственно, продвижение сайта в поиске недостаточно, нужно ещё и сделать так, чтобы пользователь задержался на нём. Этому могут поспособствовать хороший дизайн, удобное расположение элементов, описание — ключевые факторы, влияющие на успех сайта;
- лучшая гарантия высокой вероятности покупки — рекомендация авторитетной для пользователя личности. Как показал анализ, вероятность попадания в воронку продаж выше, если пользователь пришел из социальной сети или тематического сообщества, в котором ему посоветовали именно этот сайт и именно этот товар. Мнениям других людей верят больше, чем просто рекламе на сайте, в том числе и таргетированной. Соответственно, стоит обратить внимание на способ рекламирования ресурса;
- вероятность покупки выше, если клиент зашел на сайт после нескольких часов работы — с 11 часов утра и до обеда, а также и после него. В обед тоже вероятность есть, но меньше. В ночные часы посетители склонны больше к изучению товара, нежели к покупке;
- размещение рекламы в тематических ресурсах — более ценно. Если продукт действительно качественный, то люди его обязательно упомянут в подобном месте (как и сам сайт);
- если товар относительно дорогой, то стоит предла-



гать его в тех регионах, где средний доход населения, или необходимость в данном товаре больше. Нет особого смысла тратиться на рекламу там, где всё равно товар не будут покупать. Дополнительно стоит отметить, что распорядок дня относительно более обеспеченных и относительно менее обеспеченных клиентов может сильно отличаться. И это стоит учитывать;

- стоит обратить внимание, что тестирование сайта и SEO-оптимизация отражаются в статистике. При анализе выборки, поле, обозначающее город или регион непременно будет отражать ещё и города, из которых тестировали и оптимизировали сайт.

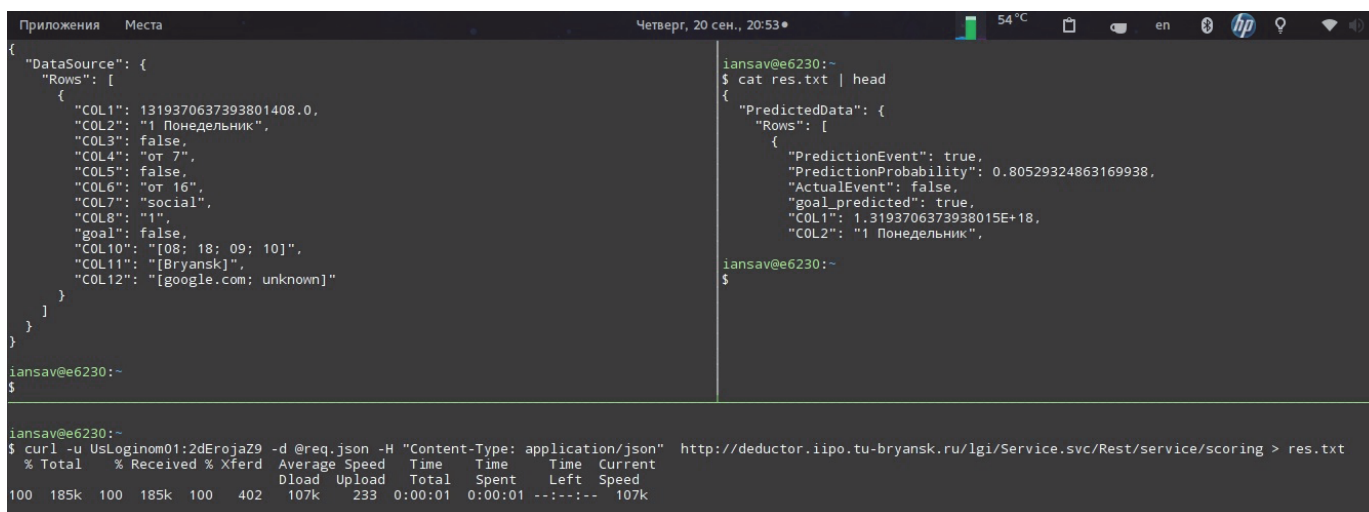
Сравнение моделей

Исследуя весь вышеописанный ансамбль моделей, можно сделать вывод, что анализ с использованием ассоциативных правил и временных рядов, к сожалению, оказался неинформативным. По сравнению с ними, скоринговая модель показала

довольно хороший результат, но не лучший в связи неоднородностью данных [26, 29]. В связи с этим, пригодным инструментом для дальнейшего применения остаётся только скоринговая модель.

Разработанный веб-сервис

На основе результатов данного исследования был создан веб-сервис, который может динамически на каждый rest-запрос давать ответ в формате json. В запросе, который отправляется на сервер, содержатся параметры пользователя, для которого необходимо решить вопрос о предложении скидки. Ответ сервера содержит в себе все данные, являющиеся результатом работы Loginom после логистической регрессии, в которых присутствуют коэффициенты, анализ «что, если..» и остальные параметры скоринговой модели. После разбора json-ответа от сервиса можно понять, стоит ли предлагать скидку текущему пользователю или нет. Сервис находится на сервере deductor.iipo.tu-bryansk.ru, что видно из URL-запроса на рисунке 9.



```
DataSource: {
  "Rows": [
    {
      "COL1": 1319370637393801408.0,
      "COL2": "1 Понедельник",
      "COL3": false,
      "COL4": "от 7",
      "COL5": false,
      "COL6": "от 16",
      "COL7": "social",
      "COL8": "1",
      "goal": false,
      "COL10": "[08: 18; 09: 10]",
      "COL11": "[Bryansk]",
      "COL12": "[google.com; unknown]"
    }
  ]
}

iansav@e6230:~$ cat res.txt | head
{"PredictedData": {
  "Rows": [
    {
      "PredictionEvent": true,
      "PredictionProbability": 0.80529324863169938,
      "ActualEvent": false,
      "goal_predicted": true,
      "COL1": 1.3193706373938015E+18,
      "COL2": "1 Понедельник"
    }
  ]
}

iansav@e6230:~$
iansav@e6230:~$ curl -u UsLoginom01:2dErojaZ9 -d @req.json -H "Content-Type: application/json" http://deductor.iipo.tu-bryansk.ru/lgi/Service.svc/Rest/service/scoring > res.txt
% Total % Received % Xferd Average Speed Time Time Time Current
Dload Upload Total Spent Left Speed
100 185k 100 185k 100 402 107k 233 0:00:01 0:00:01 --:--:-- 107k
```

Рис. 9. Результат работы веб-сервиса

Fig. 9. Web Service Result

Выводы

Благодаря скоринговой модели выяснилось, что лучший день для покупок — такой, чтобы от заказа до получения и применения товара проходило 2-3 дня — среднее время доставки из интернет-магазинов. Также был сделан вывод, что для повышения числа покупок необходимо размещать рекламу на источниках со схожей тематикой и желательно, чтобы там были рекомендации от реальных покупателей. Помимо этих выводов были сделано ещё много других, например, оказалось, что вероятность покупки выше, если клиент зашёл на сайт после нескольких часов работы и т.д.

Временные ряды и ассоциативные правила оказались менее результативными, но с их помощью были выявлены зависимости, связанные с днями недели – больше всего покупок и посещений происходит в будни.

Перспективы

В последствии разработанную модель можно интегрировать в работу сайта для предложения скидки его посетителю по средствам веб-сервиса на Loginom и других различных технологиях, так как доступны коэффициенты модели.

Как уже отмечалось выше, выборка была не особо пригодна для анализа: отношение событий/несобытий вынудило использовать сэмплинг, и, в последствии, корректировать модель. Но даже это не гарантирует абсолютно объективный результат. Как видно из графиков, предоставленных сценарием оценки качества моделей, из-за крайне малого количества событий, и, соответственно, выборки, модель переобучилась, что обусловило отличную работу в некоторых диапазонах значений, но недостаточно хорошую в других. Но даже при столь малой выборке результат получился вполне уместным - коэффициенты вполне объяснимы логически, а значит, модель не противоречит здра-



вому смыслу, что говорит о её корректности.

Так как распределение событий и несобытий было крайне неравнозначным, то в дальнейшем для анализа необходимо около 400 событий на 400 несобытий. Тогда можно быть уверенным в том, что модель нормально обучается и работает. В данном случае, такого баланса не было. Поэтому пришлось выделить около 500 несобытий и 22 события. Остальное - в тестовую выборку. Соответственно, учиться модели было практически не по чему.

Резюмируя все вышесказанные недостатки, можно прийти к одному основному — слишком малый объём данных. В исследовательских целях можно работать и с имеющийся в распоряжении выборкой, но в практических целях, где результат работы модели может повлиять на какие-либо бизнес-процессы предприятия, необходим сбор подобных данных как минимум в течение полугода [28].

Список использованных источников

- [1] Яндекс.Метрика // Яндекс. [Электронный ресурс]. URL: <https://yandex.ru/support/metrika/> (дата обращения: 03.11.2018).
- [2] *Shafraonovich Y.* Common Format and MIME Type for Comma-Separated Values (CSV) Files // RFC, 4180. IETF, (October 2005). 8 p. URL: <https://www.ietf.org/rfc/rfc4180.txt> (дата обращения: 03.11.2018).
- [3] Deductor. Продвинутая аналитика без программирования // BaseGroup Labs. [Электронный ресурс]. URL: <https://basegroup.ru/deductor/description> (дата обращения: 03.11.2018).
- [4] Временной ряд // Loginom. [Электронный ресурс]. URL: <https://wiki.loginom.ru/articles/time-series.html> (дата обращения: 03.11.2018).
- [5] Loginom. [Электронный ресурс]. URL: <https://loginom.ru> (дата обращения: 03.11.2018).
- [6] *Паклин Н.Б., Орешков В.И.* Бизнес-аналитика: от данных к знаниям. СПб.: Питер, 2013. 704 с.
- [7] *Шахиди А.* Введение в анализ ассоциативных правил // BaseGroup Labs. [Электронный ресурс]. URL: <https://basegroup.ru/community/articles/intro> (дата обращения: 03.11.2018).
- [8] *Елисеева И.И.* Эконометрика. 2-е изд. М.: Финансы и статистика, 2005. 576 с.
- [9] Дата и время // BaseGroup Labs. [Электронный ресурс]. URL: <https://basegroup.ru/deductor/function/algorithm/date> (дата обращения: 03.11.2018).
- [10] Квантование // BaseGroup Labs. [Электронный ресурс]. URL: <https://basegroup.ru/deductor/function/algorithm/quantification> (дата обращения: 03.11.2018).
- [11] Сэмплинг // BaseGroup Labs. [Электронный ресурс]. URL: <https://basegroup.ru/deductor/function/algorithm/sample> (дата обращения: 03.11.2018).
- [12] *Паклин Н.* Логистическая регрессия и ROC-анализ — математический аппарат // BaseGroup Labs. [Электронный ресурс]. URL: <https://basegroup.ru/community/articles/logistic> (дата обращения: 03.11.2018).
- [13] *Telikani A., Shahbahrami A.* Data sanitization in association rule mining: An analytical review // Expert Systems with Applications. 2018. Vol. 96. Pp. 406-426. DOI: 10.1016/j.eswa.2017.10.048
- [14] *Cook J.A.* ROC curves and nonrandom data // Pattern Recognition Letters. 2017. Vol. 85. Pp. 35-41. DOI: 10.1016/j.patrec.2016.11.015
- [15] *Fatta Di G.* Association Rules and Frequent Patterns // Encyclopedia of Bioinformatics and Computational Biology. Reference Module in Life Sciences. 2019. Vol. 1. Pp. 367-373. DOI: 10.1016/B978-0-12-809633-8.20333-6
- [16] *Salles R., Belloze K., Porto F., Gonzalez P.H., Ogasawara E.* Nonstationary time series transformation methods: An experimental review // Knowledge-Based Systems. 2019. Vol. 164. Pp. 274-291. DOI: 10.1016/j.knosys.2018.10.041
- [17] *de Vasconcelos J.B., Rocha A.* Business analytics and big data // International Journal of Information Management. 2018. DOI: 10.1016/j.ijinfomgt.2018.10.019
- [18] *Faraway J.J., Augustin N.H.* When small data beats big data // Statistics & Probability Letters. 2018. Vol. 136. Pp. 142-145. DOI: 10.1016/j.spl.2018.02.031
- [19] *Butler A.E., Copnell B., Hall H.* The development of theoretical sampling in practice // Collegian. 2018. Vol. 25, issue 5. Pp. 561-566. DOI: 10.1016/j.colegn.2018.01.002
- [20] *Hyndman R.J., Athanasopoulos G.* Forecasting: Principles & Practice. University of Western Australia, 2014. 138 p.
- [21] *Биллиг В.А., Корнеева Е.И., Сябро Н.А.* Ассоциативные правила. Сравнительный анализ инструментария // Программные продукты, системы и алгоритмы. 2016. № 2. С. 1-41. DOI: 10.15827/2311-6749.16.2.2
- [22] *Гроховский Л.О., Севостьянов И.О., Иванов Д.В., Фиронов Ф.А.* Продвижение порталов и интернет-магазинов. СПб.: Питер, 2015. 224 с.
- [23] *Романенкова О.Н.* SEO-продвижение как эффективный инструмент интернет-маркетинга // Стратегии бизнеса. 2014. № 1(3). С. 120-124. URL: <https://elibrary.ru/item.asp?id=21618536> (дата обращения: 03.11.2018).
- [24] *Любцын В.Н.* Повышение качества данных в контексте современных аналитических технологий // Вестник ЮУрГУ. Серия: Компьютерные технологии, управление, радиоэлектроника. 2012. № 23(282). С. 83-86. URL: <https://elibrary.ru/item.asp?id=17869672> (дата обращения: 03.11.2018).
- [25] *Губанов В.А.* Сравнение методов сезонной корректировки временных рядов // Научные труды: Институт народнохозяйственного прогнозирования РАН. 2010. № 8. С. 149-169. URL: <https://elibrary.ru/item.asp?id=15193311> (дата обращения: 03.11.2018).
- [26] *Попкова А.А.* Оценка эффективности и сравнение моделей бизнес-аналитических проектов с применением технологий Data Mining // Экономика, статистика и информатика. Вестник УМО. 2013. № 4. С. 184-190. URL: <https://elibrary.ru/item.asp?id=20211211> (дата обращения: 03.11.2018).
- [27] *Попов А.А., Холдонов А.А.* Структурная оптимизация нечетких регрессионных моделей с минимизацией ошибки прогноза на тестовой выборке // Вестник НГИЭИ. 2018. № 2(81). С. 7-19. URL: <https://elibrary.ru/item.asp?id=32509226> (дата обращения: 03.11.2018).
- [28] *Скиба С.А., Лойко В.И.* Социальный скоринг // Научный журнал КубГАУ. 2013. № 91. С. 1508-1517. URL: <https://elibrary.ru/item.asp?id=20678980> (дата обращения: 03.11.2018).
- [29] *Баннх А.А.* Методика оценки качества скоринговой мо-



дели с учетом доходности кредита // Вестник Удмуртского университета. Серия «Экономика и право». 2014. № 3. С. 21-24. URL: <https://elibrary.ru/item.asp?id=22933317> (дата обращения: 03.11.2018).

Поступила 03.09.2018; принята в печать 20.11.2018;
опубликована онлайн 10.12.2018.

References

- [1] Yandex.Metrika. Available at: <https://yandex.ru/support/metrika/> (accessed 03.11.2018). (In Russian)
- [2] Shafranovich Y. Common Format and MIME Type for Comma-Separated Values (CSV) Files. *RFC*, 4180. IETF, (October 2005). 8 p. Available at: <https://www.ietf.org/rfc/rfc4180.txt> (accessed 03.11.2018).
- [3] Deductor. BaseGroup Labs. Available at: <https://basegroup.ru/deductor/description> (accessed 03.11.2018). (In Russian)
- [4] Time series data. Loginom. Available at: <https://wiki.loginom.ru/articles/time-series.html> (accessed 03.11.2018). (In Russian)
- [5] Loginom. Available at: <https://loginom.ru> (accessed 03.11.2018). (In Russian)
- [6] Paklin N.B., Oreshkov V.I. Business intelligence: from data to knowledge. St. Petersburg: Peter, 2013. 704 p. (In Russian)
- [7] Shahidi A. Introduction to the analysis of associative rules. BaseGroup Labs. Available at: <https://basegroup.ru/community/articles/intro> (accessed 03.11.2018). (In Russian)
- [8] Eliseeva I.I. Econometrics. 2nd ed. M.: Finance and Statistics, 2005. 576 p. (In Russian)
- [9] Date and time. BaseGroup Labs. Available at: <https://basegroup.ru/deductor/function/algorithm/date> (accessed 03.11.2018). (In Russian)
- [10] Quantization. BaseGroup Labs. Available at: <https://basegroup.ru/deductor/function/algorithm/quantification> (accessed 03.11.2018). (In Russian)
- [11] Sampling. BaseGroup Labs. Available at: <https://basegroup.ru/deductor/function/algorithm/sample> (accessed 03.11.2018). (In Russian)
- [12] Paklin N. Logistic regression and ROC analysis - mathematical apparatus. BaseGroup Labs. Available at: <https://basegroup.ru/community/articles/logistic> (accessed 03.11.2018). (In Russian)
- [13] Telikani A., Shahbahrami A. Data sanitization in association rule mining: An analytical review. *Expert Systems with Applications*. 2018; 96:406-426. DOI: 10.1016/j.eswa.2017.10.048
- [14] Cook J.A. ROC curves and nonrandom data. *Pattern Recognition Letters*. 2017; 85:35-41. DOI: 10.1016/j.patrec.2016.11.015
- [15] Fatta Di G. Association Rules and Frequent Patterns. *Encyclopedia of Bioinformatics and Computational Biology*. Reference Module in Life Sciences. 2019; 1:367-373. DOI: 10.1016/B978-0-12-809633-8.20333-6
- [16] Salles R., Belloze K., Porto F., Gonzalez P.H., Ogasawara E. Non-stationary time series transformation methods: An experimental review. *Knowledge-Based Systems*. 2019; 164:274-291. DOI: 10.1016/j.knosys.2018.10.041
- [17] de Vasconcelos J.B., Rocha A. Business analytics and big data. *International Journal of Information Management*. 2018. DOI: 10.1016/j.ijinfomgt.2018.10.019
- [18] Faraway J.J., Augustin N.H. When small data beats big data. *Statistics & Probability Letters*. 2018; 136:142-145. DOI: 10.1016/j.spl.2018.02.031
- [19] Butler A.E., Copnell B., Hall H. The development of theoretical sampling in practice. *Collegian*. 2018; 25(5):561-566. DOI: 10.1016/j.colegn.2018.01.002
- [20] Hyndman R.J., Athanasopoulos G. Forecasting: Principles & Practice. University of Western Australia, 2014. 138 p.
- [21] Billig V.A., Korneeva E.I., Syabro N.A. Association Rules. Compared Analysis of the Tools. *Programmnye produkty, sistemy i algoritmy*. 2016; 2.:1-41. (In Russian) DOI: 10.15827/2311-6749.16.2.2
- [22] Grokhovsky L.O., Sevostyanov I.O., Ivanov D.V., Fironov F.A. Promotion of portals and online stores. SPb.: Peter, 2015. 224 p. (In Russian)
- [23] Romanenkova O.N. SEO-promotion as an effective tool for Internet Marketing. *Business strategies*. 2014; 1(3):120-124. Available at: <https://elibrary.ru/item.asp?id=21618536> (accessed 03.11.2018). (In Russian)
- [24] Lyubitsyn V.N. Improvement in Data Quality in the Context of Modern Analytical Technologies. *SUSU Bulletin. Series "Computer technologies, management, electronic engineering"*. 2012; 23(282):83-86. Available at: <https://elibrary.ru/item.asp?id=17869672> (accessed 03.11.2018). (In Russian)
- [25] Gubanov V.A. Comparison of Time Series Seasonal Adjustment Methods. *Scientific Articles – Institute of Economic Forecasting Russian Academy of Sciences*. 2010; 8:149-169. Available at: <https://elibrary.ru/item.asp?id=15193311> (accessed 03.11.2018). (In Russian)
- [26] Popkova A.A. Assessment of efficiency and comparison of models of business analytical projects with application of the datamining technologies. *Economics, Statistics and Informatics. Vestnik UMO*. 2013; 4:184-190. Available at: <https://elibrary.ru/item.asp?id=20211211> (accessed 03.11.2018). (In Russian)
- [27] Popov A.A., Holdonov A.A. Structural optimization of fuzzy regression models with minimizing the error on the test sample. *Herald NGIEI*. 2018; 2(81):7-19. Available at: <https://elibrary.ru/item.asp?id=32509226> (accessed 03.11.2018). (In Russian)
- [28] Skiba S.A., Loiko V.I. Social Scoring. *Scientific Journal of KubSAU*. 2013; 91:1508-1517. Available at: <https://elibrary.ru/item.asp?id=20678980> (accessed 03.11.2018). (In Russian)
- [29] Bannykh A.A. The method of quality estimate of the scoring model with consideration of the credit profitability. *Bulletin of Udmurt University. Series Economics and Law*. 2014; 3:21-24. Available at: <https://elibrary.ru/item.asp?id=22933317> (accessed 03.11.2018). (In Russian)

Submitted 03.09.2018; revised 20.11.2018;
published online 10.12.2018.



About the authors:

Dmitri G. Lagerev, Candidate of Engineering Sciences, Associate Professor of Informatics and Software Engineering Department, Bryansk State Technical University (10-B Harkovskaya Str., Bryansk 241035, Russia), ORCID: <http://orcid.org/0000-0002-2702-6492>, lagerev dg@mail.ru

Igor A. Savostin, Postgraduate Student, of Informatics and Software Engineering Department, Bryansk State Technical University (10-B Harkovskaya Str., Bryansk 241035, Russia), ORCID: <http://orcid.org/0000-0002-4846-2285>, iansav7@gmail.com

Vyacheslav Yu. Gerasimchuk, Postgraduate Student, of Informatics and Software Engineering Department, Bryansk State Technical University (10-B Harkovskaya Str., Bryansk 241035, Russia), ORCID: <http://orcid.org/0000-0003-3165-1917>, lava-gerasimchuk@rambler.ru

Marina S. Polyakova, Postgraduate Student, of Informatics and Software Engineering Department, Bryansk State Technical University (10-B Harkovskaya Str., Bryansk 241035, Russia), ORCID: <http://orcid.org/0000-0003-1992-4839>, jeangm@yandex.ru



BY

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted reuse, distribution, and reproduction in any medium provided the original work is properly cited.

