Сетевой научный журнал «Ресурсы» https://resources.today
Russian Journal of Resources

2026, Том 13, № 3 / 2026, Vol. 13, Iss. 3 https://resources.today/issue-3-2026.html

URL статьи: https://resources.today/PDF/19ECOR326.pdf

DOI: 10.15862/19ECOR326 (https://doi.org/10.15862/19ECOR326)

2.3.4. Управление в организационных системах (технические науки)

Ссылка для цитирования этой статьи:
Гаджимагомедов, Ю. У. Методы и алгоритмы анализа поведения субъектов в организационных системах цифровой коммерции / Ю. У. Гаджимагомедов, А. А. Михеев, С. М. Володин // Ресурсы. — 2026. — Т. 13. — № 3. — URL: https://resources.today/PDF/19ECOR326.pdf. DOI: 10.15862/19ECOR326.

For citation:
Gadzhimagomedov Yu.U., Mikheev A.A., Volodin S.M. Methods and algorithms for analyzing the behavior of subjects in organizational systems of digital commerce. *Russian Journal of Resources*. 2026;13(3): 19ECOR326. Available at: https://resources.today/PDF/19ECOR326.pdf. DOI: 10.15862/19ECOR326. (In Russ., abstract in Eng.).

УДК 519.816

Гаджимагомедов Юнус Умрудинович
АНО ВО «Московский финансово-юридический университет МФЮА», Москва, Россия
Преподаватель кафедры информационные системы и технологии
E-mail: 29520749@s.mfua.ru

Михеев Александр Алексеевич
АНО ВО «Московский финансово-юридический университет МФЮА», Москва, Россия
Преподаватель кафедры информационные системы и технологии
E-mail: recuptw@gmail.com

Володин Сергей Михайлович
АНО ВО «Московский финансово-юридический университет МФЮА», Москва, Россия
Кандидат технических наук, доцент
E-mail: Volodin.S@mfua.ru

Методы и алгоритмы анализа поведения субъектов в организационных системах цифровой коммерции

Аннотация. Статья посвящена анализу поведения пользователей цифровой торговой системы и выделению сессий с повышенной вероятностью покупки для поддержки маркетинговых решений. Цель исследования — экспериментально оценить подход к классификации пользовательских сессий по факту покупки как основу аналитического модуля торговой системы. Исследовательские вопросы состояли в том, какой из сравниваемых алгоритмов обеспечивает лучший баланс точности и полноты выявления покупок, какова структура его ошибок и какие признаки вносят наибольший вклад в прогноз. Эмпирической базой послужил открытый набор Online Shoppers Purchasing Intention Dataset: после удаления 125 полностью совпадающих строк анализировались 12 205 сессий, описанных 68 входными признаками. Задача бинарной классификации решалась методами логистической регрессии (Logistic Regression), дерева решений (Decision Tree) и случайного леса (Random Forest) на едином стратифицированном разделении данных в соотношении 80 к 20 с учётом несбалансированности классов. Наилучший результат показал случайный лес: F1-мера составила 0,6618, площадь под кривой ошибок — 0,9281; на тестовой выборке модель выявила 316 из 382 покупок, сформировав 257 ложных положительных прогнозов. Наибольшую внутреннюю оценку важности получил показатель ценности страницы PageValues. Сделан вывод, что модель

пригодна для ранжирования и аналитической сегментации сессий, однако её практическое применение требует независимой валидации, проверки доступности показателя ценности страницы до момента покупки, обучения модели без него, настройки порога с учётом стоимости ошибок и экспериментальной оценки дополнительного экономического эффекта.

Ключевые слова: машинное обучение; цифровая торговля; пользовательское поведение; прогнозирование покупки; Random Forest; анализ данных; информационные системы

Введение

Современная цифровая торговая система накапливает значительный массив сведений о поведении пользователей: о просмотре страниц, длительности взаимодействия с каталогом, источниках переходов и совершённых покупках. В связи с этим перед аналитиками стоит практическая задача *выделения сессий с повышенной вероятностью покупки* для поддержки маркетинговых решений. Для её решения, на наш взгляд, необходимо оценить не только качество прогнозов, но и *характер ошибок модели*, учитывая ограниченность ресурсов для взаимодействия с пользователями.

Задача прогнозирования покупок по данным о посещениях интернет-магазина рассматривается в научной литературе на протяжении длительного времени. В частности, D. Van den Poel и W. Buckinx показали, что сведения о навигации по сайту и истории посещений позволяют прогнозировать совершение покупки при следующем визите [1]. W.W. Moe и P.S. Fader описали изменение вероятности конверсии в зависимости от накопленной истории посещений [2]. Отметим, что особенностью таких данных является малая доля сессий с покупкой, поэтому при обучении и оценке моделей необходимо учитывать *несбалансированность классов* [3].

Среди исследований последних лет выделим работу C.O. Sakar и соавторов, в которой предложена система, сочетающая прогноз покупательского намерения по агрегированным характеристикам посещения с оценкой вероятности ухода пользователя по последовательности действий [4].

M.S. Satu и S.F. Islam исследовали влияние преобразования и отбора признаков на качество классификации покупательского намерения [5]. Данные исследования показывают, что результат зависит как от алгоритма, так и от подготовки данных, а прямое сопоставление опубликованных метрик требует одинаковых процедур обучения и оценки.

В научных исследованиях отечественных авторов также растёт интерес к данной проблематике. В частности, Д. Г. Лагерев и соавторы построили модель склонности посетителя интернет-магазина к покупке на основе технических данных о визитах [6]; О. А. Мамиев, Н. А. Финогенов и Г. Б. Сологуб сравнили методы машинного обучения для прогнозирования вероятности и суммы покупки по данным электронной коммерции [7]. Р. С. Святов использовал событийные данные интернет-магазинов и учитывал несбалансированность классов с помощью стратифицированной кросс-валидации и настройки порога классификации [8]. Е. А. Шипилова и соавторы анализировали траектории поведения пользователей онлайн-сервисов на основе графов переходов и воронок конверсии [9].

Цель исследования — экспериментально оценить подход к классификации сессий по факту покупки как основу аналитического модуля торговой системы. Прикладной вклад работы, по нашему мнению, состоит в сравнении классификаторов на едином разделении данных, а также в рассмотрении структуры ошибок и значимости признаков для определения условий применения модели. Подчеркнём, что работа представляет собой эксперимент с существующими алгоритмами на открытых данных.

1. Материалы и методы

1.1 Исходные данные и предварительная обработка

Эмпирической базой исследования послужил набор Online Shoppers Purchasing Intention Dataset из UCI Machine Learning Repository (репозиторий машинного обучения Калифорнийского университета в Ирвайне).[1] Исходный CSV-файл (Comma-Separated Values — текстовый формат табличных данных с разделителями) содержит 12 330 строк и 18 столбцов: 17 входных признаков и целевую переменную Revenue. Каждая строка описывает отдельную сессию, при этом Revenue принимает значение 1 при наличии покупки и 0 при её отсутствии. Набор включает сведения о количестве и длительности просмотра страниц, показателях веб-аналитики, месяце посещения, типе посетителя и технических параметрах доступа.

Подготовка данных выполнена нами на Python с использованием библиотеки pandas [10]. Проверка исходной таблицы выявила отсутствие пропусков и наличие 125 полностью совпадающих строк, которые были удалены методом drop_duplicates до разделения выборки. В результате очищенный набор содержит 12 205 сессий, в том числе 1 908 — с покупкой и 10 297 — без покупки. Логические столбцы Weekend и Revenue преобразованы в значения 0 и 1.

Категориальные признаки Month, VisitorType, OperatingSystems, Browser, Region и TrafficType преобразованы методом one-hot encoding с параметром drop_first=True, исключающим один базовый уровень каждой категории. Отметим, что числовые коды технических параметров и источника трафика трактовались как категории. Итоговая таблица содержит 69 столбцов: 68 входных признаков и Revenue. Последний исключён из матрицы признаков X и использован только как целевая переменная y.

Особо отметим, что в эксперименте состав категорий определён по полному набору данных. При внедрении кодирование следует обучать на обучающей части и фиксировать обработку новых категорий. Масштабирование для линейной модели, в отличие от определения категорий, выполнено только по обучающей части, как описано далее.

1.2 Организация эксперимента

Модели реализованы средствами библиотеки scikit-learn [11]. Данные разделены функцией train_test_split с параметрами test_size=0.2, random_state=42 и stratify=y. Обучающая выборка включает 9764 сессии, тестовая — 2441; доли покупок составляют соответственно 15,63 и 15,65 %. Стратификация сохраняет близкое соотношение классов. Поскольку все модели обучались и оценивались на одинаковых частях набора, различия метрик относятся к выбранным алгоритмам и их настройкам в пределах данного разделения.

Для сравнения нами выбраны три классификатора:

- *Logistic Regression* — линейный ориентир; перед её обучением признаки стандартизированы с помощью StandardScaler, включённого в Pipeline, при этом параметры масштабирования рассчитаны на обучающей части и затем применены к тестовой; установлен max_iter=3000;
- *Decision Tree* — одиночное дерево решений, ограниченное глубиной max_depth=6;
- *Random Forest* — ансамбль из n_estimators=200 деревьев с max_depth=10; обучение выполнялось с n_jobs=-1.

[1] Sakar C., Kastro Y. Online Shoppers Purchasing Intention Dataset // UCI Machine Learning Repository. — 2018. — DOI 10.24432/C5F88Q. — URL: https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset (дата обращения: 07.10.2026). — Текст: электронный.

Ансамбль позволяет сопоставить одиночное дерево с совокупностью деревьев, обучаемых на случайных подвыборках [12; 13].

Для всех трёх классификаторов задан параметр class_weight="balanced", учитывающий частоты классов при обучении, и random_state=42; остальные настройки оставлены по умолчанию. В исходном эксперименте использована версия scikit-learn 1.9.0, что подтверждено метаданными сохранённой модели. Подчеркнём, что при воспроизведении необходимо фиксировать версии зависимостей, поскольку одинаковые параметры и начальное состояние генератора не гарантируют совпадения результатов между версиями библиотеки.

Гиперпараметры заданы в коде непосредственно: автоматический поиск, кросс-валидация и обучение без PageValues не выполнялись. Итоговый алгоритм выбирался по максимальному F1-score на тестовой части, поэтому данная часть использована и для сравнения вариантов, и для представления их качества. Дополнительная независимая выборка для подтверждения качества выбранной модели отсутствует; это ограничение учитывается нами при интерпретации результатов [14].[2]

Качество классификации оценивалось по метрикам accuracy, precision, recall, F1-score и ROC-AUC (Area Under the Receiver Operating Characteristic Curve — площадь под ROC-кривой, или кривой ошибок):

- *precision* показывает долю реальных покупок среди положительных прогнозов;
- *recall* — долю выявленных покупок среди всех фактических;
- *F1-score* обобщает баланс precision и recall;
- *ROC-AUC* характеризует ранжирование сессий по прогнозному баллу при различных порогах [15].

Метрики положительного класса рассчитаны для Revenue=1. Для дискретных прогнозов использован стандартный метод predict без отдельной настройки порога; ROC-AUC рассчитана по выходу predict_proba.

2. Результаты

2.1 Результаты исследовательского анализа

Общая конверсия в очищенном наборе составила 15,63 %. При сравнении типов посетителей максимальное значение наблюдалось у New_Visitor — 24,93 %, а среди представленных месяцев — у ноября (25,49 %). Отметим, что эти показатели описывают различия внутри исследуемого набора: их нельзя автоматически переносить на аудиторию другого магазина или трактовать как доказательство устойчивой сезонной закономерности за несколько лет.

Наиболее выраженная положительная корреляция числового признака с Revenue выявлена для PageValues: коэффициент составил 0,4919. Связь с покупкой также наблюдалась у показателей взаимодействия с товарными страницами, в то время как BounceRates и ExitRates имели отрицательную корреляцию с целевой переменной. Подчеркнём, что выявленные зависимости являются *статистическими ассоциациями* и сами по себе не устанавливают причинного влияния длительности просмотра или числа посещённых страниц на покупку.

[2] Cross-validation: evaluating estimator performance // scikit-learn documentation. — URL: https://scikit-learn.org/stable/modules/cross_validation.html (дата обращения: 20.09.2026). — Текст: электронный.

По результатам корреляционного анализа признаки не исключались: все классификаторы получили одинаковый набор из 68 входных столбцов.

2.2 Сравнительная оценка моделей

Показатели классификаторов на одной и той же тестовой выборке представлены в таблице 1. Значения округлены до четырёх знаков после запятой; исходные результаты сохранены в файле model_metrics.csv.

Таблица 1

Качество классификации пользовательских сессий

Модель	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	0,8677	0,5515	0,8272	0,6618	0,9281
Logistic Regression	0,8464	0,5058	0,7958	0,6185	0,9083
Decision Tree	0,8374	0,4884	0,8246	0,6134	0,9096

Составлено авторами по результатам вычислительного эксперимента

Random Forest получила максимальные значения F1-score (0,6618) и ROC-AUC (0,9281), а также наибольшие precision и recall среди сравниваемых моделей. Logistic Regression достигла F1-score 0,6185, Decision Tree — 0,6134. Таким образом, ансамбль показал лучший результат в выбранной конфигурации. Вместе с тем особо отметим, что превосходство не проверялось статистическими тестами и не означает, что Random Forest будет лучшим алгоритмом при любом разделении данных или на другой торговой платформе.

Показатель accuracy Random Forest составил 86,77 %. Для сравнения, правило, относящее все 2441 тестовые сессии к классу «без покупки», правильно классифицировало бы 2059 наблюдений, то есть получило бы accuracy 84,35 %, но не выявило бы ни одной покупки. Разница составляет 2,42 процентного пункта. Данный расчёт наглядно показывает, почему оценка только по accuracy скрывает особенности работы с редким положительным классом [16].

ROC-кривые сравниваемых моделей представлены на рисунке 1. Они построены по вероятностным выходам всех трёх моделей на тестовой части и показывают соотношение доли выявленных покупок и доли ложных положительных прогнозов при изменении порога.

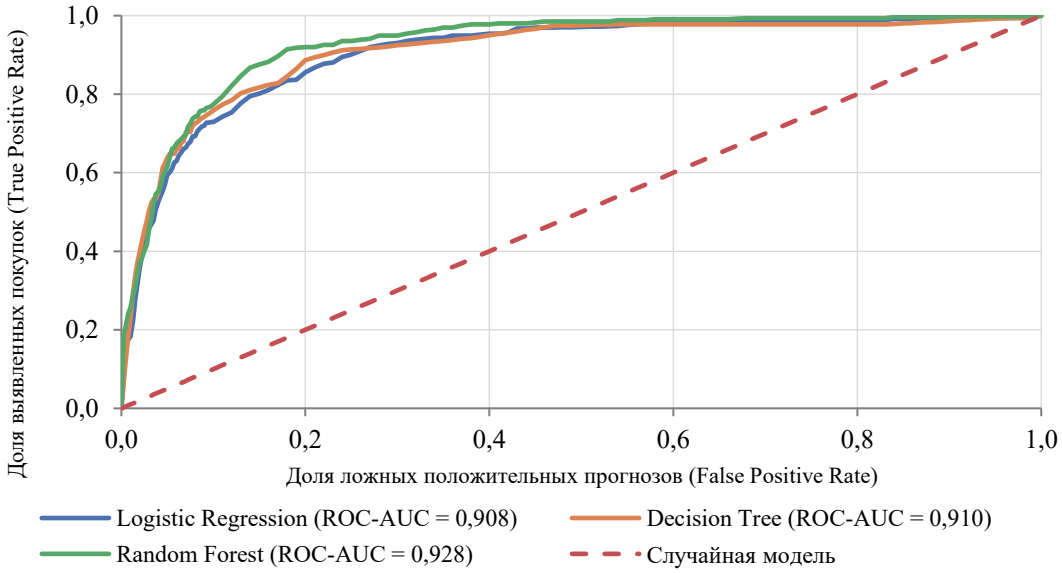


Рисунок 1. *ROC-кривые сравниваемых моделей*
(составлено авторами по результатам вычислительного эксперимента)

ROC-AUC ансамбля выше, чем у линейной модели и одиночного дерева. При этом ROC-AUC не определяет precision в конкретной рабочей точке и не подтверждает калибровку вероятностей. В связи с этим для выбора маркетингового действия необходимо дополнительно рассматривать ошибки при выбранном пороге, а не только площадь под ROC-кривой [17].

Матрица ошибок Random Forest приведена в таблице 2. Из 382 фактических покупок модель правильно выявила 316 и пропустила 66. Из 2059 сессий без покупки 1802 классифицированы верно, а 257 ошибочно отнесены к положительному классу.

Таблица 2

Матрица ошибок модели Random Forest

		Предсказанный класс	
		Без покупки	С покупкой
Фактический класс	Без покупки	**1802**	257
	С покупкой	66	316

Составлено авторами по результатам вычислительного эксперимента

Из 573 положительных прогнозов 316 соответствовали покупкам: precision равна 55,15 %, а 44,85 % выделенного сегмента составляют сессии без покупки. Recall 82,72 % показывает, что высокая полнота сопровождается значительным числом ложных положительных прогнозов. На наш взгляд, при платных коммуникациях или скидках допустимость этого компромисса зависит от стоимости действия.

2.3 Значимость признаков и интерпретация

Для Random Forest рассчитаны встроенные оценки feature_importances_, основанные на уменьшении неоднородности узлов деревьев. Значения нормированы так, что их сумма равна единице. На рисунке 2 показаны 15 признаков с максимальными оценками; анализ относится к сохранённой модели исходного эксперимента.

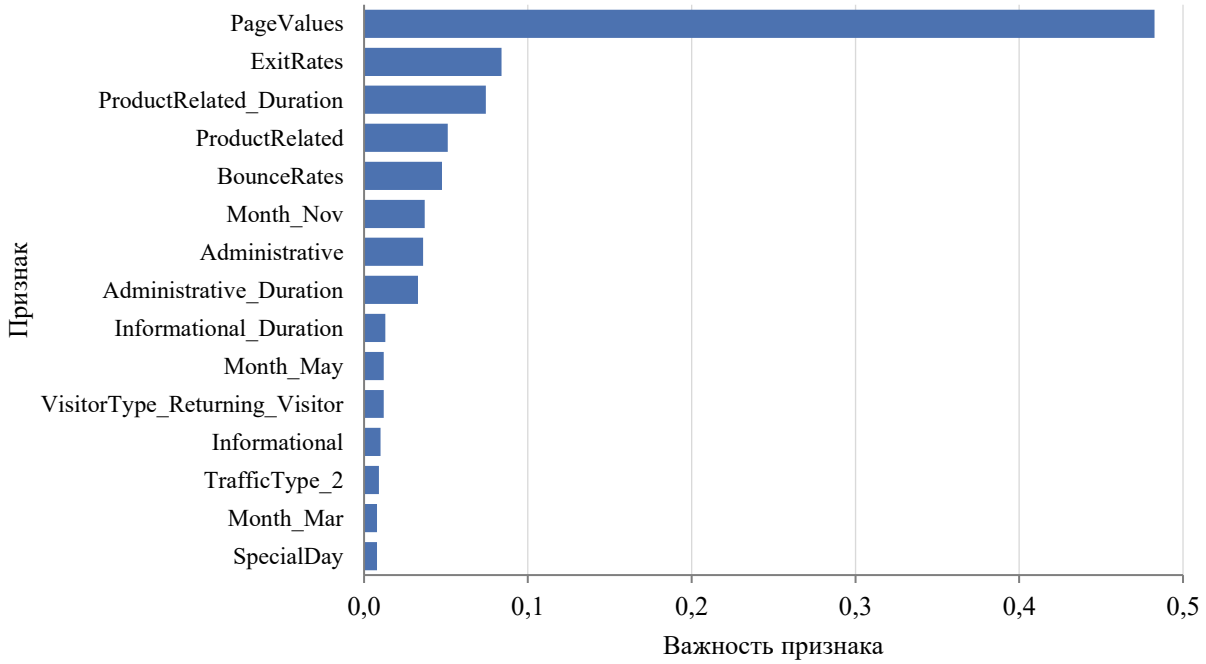


Рисунок 2. *Важность признаков в модели Random Forest (составлено авторами по результатам вычислительного эксперимента)*

19ECOR326
Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Оценка PageValues составила 0,4827. Следующие по величине значения получили ExitRates — 0,0840, ProductRelated_Duration — 0,0744, ProductRelated — 0,0510 и BounceRates — 0,0475. Особо отметим, что значение 0,4827 отражает долю в нормированной внутренней оценке важности модели и не означает, что PageValues объясняет 48,27 % покупок или изменения выручки.

Встроенная важность зависит от структуры обученных деревьев и особенностей признаков: она не показывает направление зависимости и может быть смещена в пользу переменных с большим числом возможных значений [18]. Поэтому вывод о наиболее значимых признаках требует дополнительной проверки, например перестановочной оценкой на данных, не использованных для обучения[3], или другими методами интерпретации моделей машинного обучения [19]. В настоящем эксперименте такая проверка не выполнялась.

Обсуждение

3.1 Использование результатов в торговой системе

Предлагаемый нами сценарий применения модели включает передачу характеристик сессии в аналитический модуль, преобразование признаков и получение прогнозного балла. На основе этого балла сотрудник может ранжировать сессии и выбирать действие с учётом бюджета. Отметим, что сервис прогнозирования и его интеграция с действующей платформой в проекте не реализованы; сценарий относится к дальнейшему внедрению.

Выбор порога классификации зависит от стоимости ошибок. Во-первых, ложный положительный прогноз может привести к ненужной коммуникации или скидке. Во-вторых, ложный отрицательный прогноз исключает сессию с покупкой из выделенного сегмента. При условных стоимостях c_{FP} и c_{FN} суммарная оценка для полученной матрицы ошибок имеет вид

$$L = 257 \times c_{FP} + 66 \times c_{FN}, \quad (1)$$

где:

L — суммарная стоимость ошибок;

c_{FP} — условная стоимость одного ложного положительного прогноза;

c_{FN} — условная стоимость одного ложного отрицательного прогноза.

Коэффициенты c_{FP} и c_{FN} в исследовании не измерялись; формула (1) описывает стоимость ошибок, а не прибыль или окупаемость.

Подчеркнём, что классификация факта покупки не показывает, изменит ли предложение поведение конкретного пользователя: часть выделенных покупателей могла бы совершить покупку и без воздействия [20; 21]. Для оценки дополнительного эффекта коммуникации нужен отдельный эксперимент с контрольной группой [22]. В проекте не проводились A/B-тесты (контролируемые эксперименты со случайным разделением пользователей на группы), не измерялись рост конверсии, выручки и экономия бюджета. Поэтому практическая значимость полученного результата заключается в возможности *аналитической сегментации*, а экономический эффект остаётся предметом последующей проверки.

3.2 Ограничения исследования

Проведённый эксперимент имеет ряд ограничений, которые необходимо учитывать при интерпретации его результатов.

[3] Permutation feature importance // scikit-learn documentation. — URL: https://scikit-learn.org/stable/modules/permutation_importance.html (дата обращения: 01.10.2026). — Текст: электронный.

Издательство «Мир науки» \ Publishing company «World of science» http://izd-mn.com

Во-первых, эксперимент основан на одном наборе и одном случайном разделении. Временная проверка, кросс-валидация, доверительные интервалы и независимая оценка после выбора алгоритма отсутствуют. Следующий этап исследования предполагает кросс-валидацию на обучающей части и оценку выбранной модели на отдельной тестовой выборке. Полученные метрики характеризуют данный эксперимент и не гарантируют результат эксплуатации.

Во-вторых, UCI описывает PageValues как производный показатель ценности страницы, связанный с транзакциями.[4] Это не доказывает утечку целевой переменной [23], однако для раннего прогноза необходимо проверить доступность признака до события покупки. Качество классификации агрегированных сессий пока не подтверждает столь же точный прогноз в начале посещения. Обучение без PageValues позволит оценить зависимость результата от этого показателя.

В-третьих, вероятностные выходы не калибровались: перед расчётом ожидаемой выручки нужно проверить соответствие прогнозов частотам покупок [24]. Внедрение модели также требует проверки на данных выбранной платформы, фиксации преобразований признаков и версий зависимостей, организации мониторинга качества и переобучения.

Заключение

Подводя итог вышеизложенному, сформулируем основные выводы исследования.

1. *Поставленная цель — экспериментально оценить подход к классификации сессий по факту покупки как основу аналитического модуля торговой системы — достигнута в пределах выбранной схемы эксперимента.* На 12 205 очищенных сессиях сравнены три классификатора с единым набором из 68 входных признаков.

2. *Наилучший результат в выбранной конфигурации показала модель Random Forest*, получившая максимальные F1-score 0,6618 и ROC-AUC 0,9281. На тестовой выборке из 2441 сессии она выявила 316 покупок, пропустила 66 и сформировала 257 ложных положительных прогнозов при 1802 верных отрицательных.

3. *Модель выделяет сегмент с повышенной долей покупок, однако высокая полнота сопровождается ложными прогнозами.* Для применения модели необходимы независимая валидация, проверка доступности PageValues и обучение без него, настройка порога под стоимость действия и экспериментальная оценка дополнительного экономического эффекта.

ЛИТЕРАТУРА

1. Van den Poel, D. Predicting online-purchasing behaviour / D. Van den Poel, W. Buckinx. — DOI 10.1016/j.ejor.2004.04.022 // European Journal of Operational Research. — 2005. — Т. 166, № 2. — С. 557–575.
2. Moe, W.W. Dynamic conversion behavior at e-commerce sites / W.W. Moe, P.S. Fader. — DOI 10.1287/mnsc.1040.0153 // Management Science. — 2004. — Т. 50, № 3. — С. 326–335.
3. He, H. Learning from imbalanced data / H. He, E.A. Garcia. — DOI 10.1109/TKDE.2008.239 // IEEE Transactions on Knowledge and Data Engineering. — 2009. — Т. 21, № 9. — С. 1263–1284.

[4] Online Shoppers Purchasing Intention Dataset // UCI Machine Learning Repository. — URL: https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset (дата обращения: 07.10.2026). — Текст: электронный.

4. Sakar, C.O. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks / C.O. Sakar, S.O. Polat, M. Katircioglu, Y. Kastro. — DOI 10.1007/s00521-018-3523-0 // Neural Computing and Applications. — 2019. — Т. 31, № 10. — С. 6893–6908.
5. Satu, M.S. Modeling online customer purchase intention behavior applying different feature engineering and classification techniques / M.S. Satu, S.F. Islam. — DOI 10.1007/s44163-023-00086-0 // Discover Artificial Intelligence. — 2023. — Т. 3, № 1. — С. 36.
6. Лагерев, Д. Г. 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. — DOI 10.25559/SITITO.14.201804.911-922 // Современные информационные технологии и ИТ-образование. — 2018. — Т. 14, № 4. — С. 911–922. — URL: http://sitito.cs.msu.ru/index.php/SITITO/article/view/458 (дата обращения: 19.10.2026).
7. Мамиев, О. А. Использование методов машинного обучения для решения задач прогнозирования суммы и вероятности покупки на основе данных электронной коммерции / О. А. Мамиев, Н. А. Финогенов, Г. Б. Сологуб. — DOI 10.17759/mda.2020 20100403 // Моделирование и анализ данных. — 2020. — Т. 10, № 4. — С. 31–40. — URL: https://psyjournals.ru/journals/mda/archive/2020_n4/Mamiev_Finogenov_Sologub (дата обращения: 19.10.2026).
8. Святов, Р. С. Прогнозирование покупательского поведения пользователей интернет-магазинов на основе событийных данных / Р. С. Святов. — DOI 10.26102/2310-6018/2025.51.4.064 // Моделирование, оптимизация и информационные технологии. — 2025. — Т. 13, № 4. — URL: https://moitvivt.ru/ru/journal/article?id=2129 (дата обращения: 19.10.2026).
9. Шипилова, Е. А. Анализ и моделирование траекторий поведения пользователей онлайн-сервисов с использованием платформы RETENTIONEERING / Е. А. Шипилова, Е. Е. Некрылов, Т. В. Курченкова. — DOI 10.12737/2219-0767-2022-15-4-82-93 // Моделирование систем и процессов. — 2022. — Т. 15, № 4. — С. 82–93. — URL: https://zh-szf.ru/ru/nauka/article/55589/view (дата обращения: 19.10.2026).
10. McKinney, W. Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter / W. McKinney. — 3rd ed. — Sebastopol: O'Reilly Media, 2022. — 582 с. — ISBN 978-1-098-10402-3. — URL: https://wesmckinney.com/book/ (дата обращения: 19.10.2026).
11. Scikit-learn: Machine Learning in Python / F. Pedregosa, G. Varoquaux, A. Gramfort [et al.] // Journal of Machine Learning Research. — 2011. — Т. 12. — С. 2825–2830. — URL: https://jmlr.org/papers/v12/pedregosa11a.html (дата обращения: 19.10.2026).
12. Breiman, L. Random forests / L. Breiman. — DOI 10.1023/A:1010933404324 // Machine Learning. — 2001. — Т. 45, № 1. — С. 5–32.
13. Hastie, T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction / T. Hastie, R. Tibshirani, J. Friedman. — 2nd ed. — New York: Springer, 2009. — XXII, 745 с. — DOI 10.1007/978-0-387-84858-7.
14. Varma, S. Bias in error estimation when using cross-validation for model selection / S. Varma, R. Simon. — DOI 10.1186/1471-2105-7-91 // BMC Bioinformatics. — 2006. — Т. 7. — С. 91.

15. Fawcett, T. An introduction to ROC analysis / T. Fawcett. — DOI 10.1016/j.patrec.2005.10.010 // Pattern Recognition Letters. — 2006. — Т. 27, № 8. — С. 861–874.
16. Широкова, Е. В. Сравнение метрик бинарной классификации для оценки эффективности алгоритмов машинного обучения / Е. В. Широкова. — DOI 10.37882/2223-2966.2025.02.40 // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. — 2025. — № 2. — С. 164–169. — URL: http://nauteh-journal.ru/files/46836277-1b33-4a58-abf6-32059dd9af91 (дата обращения: 19.10.2026).
17. Saito, T. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets / T. Saito, M. Rehmsmeier. — DOI 10.1371/journal.pone.0118432 // PLoS ONE. — 2015. — Т. 10, № 3. — С. e0118432.
18. Strobl, C. Bias in random forest variable importance measures: Illustrations, sources and a solution / C. Strobl, A.-L. Boulesteix, A. Zeileis, T. Hothorn. — DOI 10.1186/1471-2105-8-25 // BMC Bioinformatics. — 2007. — Т. 8. — С. 25.
19. Шевская, Н. В. Объяснимый искусственный интеллект и методы интерпретации результатов / Н. В. Шевская. — DOI 10.26102/2310-6018/2021.33.2.024 // Моделирование, оптимизация и информационные технологии. — 2021. — Т. 9, № 2. — URL: https://moitvivt.ru/ru/journal/article?id=1005 (дата обращения: 19.10.2026).
20. Devriendt, F. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics / F. Devriendt, D. Moldovan, W. Verbeke. — DOI 10.1089/big.2017.0104 // Big Data. — 2018. — Т. 6, № 1. — С. 13–41.
21. Азарнова, Т. В. Применение uplift-моделирования в задачах повышения эффективности маркетинговых коммуникаций / Т. В. Азарнова, Е. В. Ярош. — DOI 10.26102/2310-6018/2025.50.3.031 // Моделирование, оптимизация и информационные технологии. — 2025. — Т. 13, № 3. — URL: https://moitvivt.ru/ru/journal/article?id=2023 (дата обращения: 19.10.2026).
22. Kohavi, R. Controlled experiments on the web: survey and practical guide / R. Kohavi, R. Longbotham, D. Sommerfield, R.M. Henne. — DOI 10.1007/s10618-008-0114-1 // Data Mining and Knowledge Discovery. — 2009. — Т. 18, № 1. — С. 140–181.
23. Kaufman, S. Leakage in data mining: Formulation, detection, and avoidance / S. Kaufman, S. Rosset, C. Perlich, O. Stitelman. — DOI 10.1145/2382577.2382579 // ACM Transactions on Knowledge Discovery from Data. — 2012. — Т. 6, № 4. — С. 15.
24. Niculescu-Mizil, A. Predicting good probabilities with supervised learning / A. Niculescu-Mizil, R. Caruana. — DOI 10.1145/1102351.1102430 // Proceedings of the 22nd International Conference on Machine Learning (ICML'05). — New York: ACM, 2005. — С. 625–632.

Gadzhimagomedov Yunus Umrudinovich
Moscow University of Finance and Law (MFUA), Moscow, Russia
E-mail: 29520749@s.mfua.ru

Mikheev Aleksandr Alekseevich
Moscow University of Finance and Law (MFUA), Moscow, Russia
E-mail: recuptw@gmail.com

Volodin Sergey Mikhaylovich
Moscow University of Finance and Law (MFUA), Moscow, Russia
E-mail: Volodin.S@mfua.ru

Methods and algorithms for analyzing the behavior of subjects in organizational systems of digital commerce

Abstract. The article analyzes user behavior in a digital commerce system in order to identify sessions with an increased probability of purchase and support marketing decisions. The aim is to experimentally evaluate an approach to classifying user sessions by the fact of purchase as the basis of an analytical module of a trading system. The research questions were which of the compared algorithms best balances precision and completeness of purchase detection, what the structure of its errors is, and which features contribute most to the prediction. The empirical basis was the open Online Shoppers Purchasing Intention Dataset: after removing 125 duplicated rows, 12 205 sessions with 68 input features were analyzed. Logistic regression, a decision tree and a random forest were compared on a single stratified 80 to 20 split, taking class imbalance into account. The random forest showed the best result: the F1 score was 0,6618 and the area under the receiver operating characteristic curve was 0,9281; on the test sample it detected 316 of 382 purchases with 257 false positive predictions. The page value indicator PageValues received the highest importance score. The model is suitable for ranking and analytical segmentation of sessions, but its application requires independent validation, verification that the page value indicator is available before the purchase, training without it, cost-based threshold tuning and experimental evaluation of the economic effect.

Keywords: machine learning; digital commerce; user behavior; purchase prediction; Random Forest; data analysis; information systems

Дата публикации: 30.10.2026