

DOI: 10.26794/2587-5671-2020-25-5-186-199

УДК 336.7(045)

JEL E58, G21, C53

# Сравнительный анализ методов машинного обучения при идентификации признаков вовлеченности кредитных организаций и их клиентов в сомнительные операции

Ю.М. Бекетнова

Финансовый университет, Москва, Россия

<https://orcid.org/0000-0002-1005-6265>

## АННОТАЦИЯ

В сфере финансового мониторинга для принятия эффективных управленческих решений необходимо оперативно получать объективные оценки хозяйствующих субъектов (в частности, кредитных организаций). Автоматизация процесса выявления недобросовестных кредитных организаций на основе методов машинного обучения позволит контролирующим органам оперативно выявлять и пресекать противоправную деятельность. **Цель** исследования состоит в обосновании возможностей применения методов и алгоритмов машинного обучения для автоматической идентификации недобросовестных кредитных организаций. Для этого необходимо подобрать математический инструментальный анализ данных о кредитных организациях, позволяющий проводить диагностику вовлеченности банка в процессы отмывания преступных доходов. Проведен сравнительный анализ результатов обработки данных о деятельности кредитных организаций **методами** классификации – логистической регрессии, деревьев решений. Использован метод опорных векторов, нейросетевые методы, Байесовские сети (двухклассовая сеть Байеса) и поиска аномалий – алгоритм одноклассовой машины опорных векторов и алгоритм обнаружения аномалий на основе метода главных компонент. Приведены результаты решения задачи классификации кредитных организаций с точки зрения возможной вовлеченности в процессы отмывания денежных средств, результаты анализа данных о деятельности кредитных организаций методами выявления аномалий. Проведен сравнительный анализ результатов, полученных при применении различных современных алгоритмов классификации и поиска аномалий. Сделан **вывод** о том, что алгоритм поиска аномалий на основе метода главных компонент показал более точные результаты по сравнению с алгоритмом одноклассовой машины опорных векторов. Из рассмотренных алгоритмов классификации наиболее точные результаты показал алгоритм двухклассового усиленного дерева решений (*Adaboost*). Приведенные результаты исследования могут быть использованы Банком России и Росфинмониторингом для автоматизации выявления недобросовестных кредитных организаций.

**Ключевые слова:** сомнительные операции; отмывание доходов; банк; кредитная организация; методы выявления аномалий; машинное обучение

**Для цитирования:** Бекетнова Ю.М. Сравнительный анализ методов машинного обучения при идентификации признаков вовлеченности кредитных организаций и их клиентов в сомнительные операции. *Финансы: теория и практика*. 2021;25(5):186-199. DOI: 10.26794/2587-5671-2021-25-5-186-199

## Comparative Analysis of Machine Learning Methods to Identify Signs of Suspicious Transactions of Credit Institutions and Their Clients

Yu. M. Beketnova

Financial University, Moscow, Russia

<https://orcid.org/0000-0002-1005-6265>

## ABSTRACT

In the field of financial monitoring, it is necessary to promptly obtain objective assessments of economic entities (in particular, credit institutions) for effective decision-making. Automation of the process of identifying unscrupulous credit institutions based on machine learning methods will allow regulatory authorities to quickly identify and suppress

illegal activities. The **aim** of the research is to substantiate the possibilities of using machine learning methods and algorithms for the automatic identification of unscrupulous credit institutions. It is required to select a mathematical toolkit for analyzing data on credit institutions, which allows tracking the involvement of a bank in money laundering processes. The paper provides a comparative analysis of the results of processing data on the activities of credit institutions using classification **methods** – logistic regression, decision trees. The author applies support vector machine and neural network methods, Bayesian networks (Two-Class Bayes Point Machine), and anomaly search – an algorithm of a One-Class Support Vector Machine and a PCA-Based Anomaly Detection algorithm. The study presents the results of solving the problem of classifying credit institutions in terms of possible involvement in money laundering processes, the results of analyzing data on the activities of credit institutions by methods of detecting anomalies. A comparative analysis of the results obtained using various modern algorithms for the classification and search for anomalies is carried out. The author **concluded** that the PCA-Based Anomaly Detection algorithm showed more accurate results compared to the One-Class Support Vector Machine algorithm. Of the considered classification algorithms, the most accurate results were shown by the Two-Class Boosted Decision Tree (AdaBoost) algorithm. The research results can be used by the Bank of Russia and Rosfinmonitoring to automate the identification of unscrupulous credit institutions.

**Keywords:** suspicious transactions; money laundering; bank; credit institution; anomaly detection methods; machine learning

**For citation:** Beketnova Yu.M. Comparative analysis of machine learning methods to identify signs of suspicious transactions of credit institutions and their clients. *Finance: Theory and Practice*. 2021;25(5):186-199. DOI: 10.26794/2587-5671-2021-25-5-186-199

## ВВЕДЕНИЕ

Кредитные организации [и другие субъекты Федерального закона от 07.08.2001 № 115-ФЗ «О противодействии легализации (отмыванию) доходов, полученных преступным путем, и финансированию терроризма»<sup>1</sup>] – первый рубеж борьбы с незаконной финансовой деятельностью. Они призваны собирать и анализировать информацию о своих клиентах и их финансовых операциях и сообщать в случае выявленной подозрительной активности в Росфинмониторинг. Однако кредитные организации также могут быть вовлечены в незаконную деятельность, осознанно покрывая теневые схемы, или по причине слабости системы внутреннего контроля.

Процессы отмывания денежных средств затрагивают различные отрасли экономики и отражаются, например, на объемах уплаченных налогов и сборов, выведенных за рубеж денежных средств, на динамике создания и ликвидации российских юридических лиц, внешней торговле, общем состоянии преступности, тенденциях миграции и т.п.

Росфинмониторинг является центральным звеном российской системы противодействия отмыванию доходов, осуществляющим взаимодействие по этому вопросу с другими подразделениями финансовой разведки, международными организациями, федеральными органами исполни-

тельной власти, правоохранительными органами и кредитными организациями.

В процессе финансового мониторинга можно выявить следующие негативные проявления со стороны и/или в отношении кредитных организаций:

- использование банковской инфраструктуры для организации схем по оказанию теневых финансовых услуг и отмыванию доходов;
- вывод денежных средств из банка – например, выдача заведомо невозвратных кредитов перед банкротством или преднамеренное банкротство;
- банкротство кредитной организации.

На практике, как правило, приходится иметь дело с комбинацией этих составляющих при различной степени их проявления. Поэтому здесь и далее будем использовать термин «девиантная составляющая деятельности кредитной организации», имея в виду одно или несколько вышеописанных негативных проявлений.

Специфика анализа данных финансового мониторинга и выявления лиц и организаций, вовлеченных в отмывание доходов, заключается в следующем:

1. Девиантные субъекты стремятся скрыть свою причастность к противоправной деятельности и активно маскируются под законопослушных участников, используя при этом современные информационные технологии и экспертные знания профессионалов в области финансов, юриспруденции и пр.

2. В связи с п. 1 для выявления латентных правонарушений аналитикам Росфинмонито-

<sup>1</sup> Федеральный закон от 07.08.2001 № 115-ФЗ «О противодействии легализации (отмыванию) доходов, полученных преступным путем, и финансированию терроризма». Справочно-правовая система «Консультант Плюс». URL: [http://www.consultant.ru/document/cons\\_doc\\_LAW\\_32834/](http://www.consultant.ru/document/cons_doc_LAW_32834/) (дата обращения: 04.10.2020).

ринга приходится анализировать большие объемы гетерогенных данных.

3. Проведение такого рода анализа по каждому проверяемому объекту требует от аналитика глубоких профессиональных знаний и опыта практической работы в предметной области, с одной стороны, и значительных временных ресурсов — с другой.

Аналитикам Росфинмониторинга при анализе кредитных организаций приходится оперировать большими объемами информации. Приведем некоторую статистику для иллюстрации этого утверждения.

По официальной информации Банка России по состоянию на 01.01.2020 г. на территории Российской Федерации осуществляют деятельность 442 кредитные организации и 618 филиалов<sup>2</sup>. Банковская отчетность содержит сотни параметров. Кроме того, в федеральной базе данных Росфинмониторинга по каждой кредитной организации ведется собственный учет информации — до 50 полей данных.

В дополнение к исходным данным в федеральной базе данных Росфинмониторинга формируется статистика по каждому виду объектов и дополнительная идентификационная информация — адреса, данные документов, удостоверяющих личность, и прочее. Ведомственные данные Росфинмониторинга обогащаются сведениями из различных государственных реестров, информацией о внешнеэкономической деятельности субъектов, налоговыми сведениями.

Эффективность системы противодействия отмыванию доходов во многом зависит от ее способности своевременно выявлять тенденции и закономерности в деятельности субъектов, что обуславливает необходимость оперативного получения объективных оценок работы объектов финансового мониторинга.

Традиционно в органах государственной власти применялся подход к проверкам объектов наблюдения, заключающийся в последовательной оценке экспертом одного объекта проверки за другим. Подобные оценки могут обладать экспертным субъективизмом [1, 2]. Кроме того, такой подход является ресурсоемким и требует существенных временных затрат.

Возрастающий объем поступающей информации (приблизительно на 20% ежегодно) приводит

к снижению оперативности ее обработки. Лица, ответственные за принятие решений, вынуждены работать с субъективными результатами анализа при растянутых сроках их получения [3].

Анализ задач Росфинмониторинга по противодействию отмыванию доходов показал, что фактическая потребность в количестве объектов, подлежащих анализу, многократно превышает возможности аналитиков. Данная проблемная ситуация требует определения приоритетов проверок.

Гетерогенный характер информационных ресурсов и значительный объем исключают возможность их ручной обработки.

Требуется переход от последовательных экспертных проверок единичных объектов к параллельным массовым автоматизированным проверкам с учетом современных методических и инструментальных возможностей в условиях цифровой трансформации государственного управления.

Необходимым инструментом для этого могут служить современные методы анализа данных и машинного обучения. Автоматизация процесса выявления недобросовестных кредитных организаций на основе методов машинного обучения позволит контролирующим органам оперативно выявлять и пресекать противоправную деятельность.

Проблематика легализации преступных доходов через банковский сектор на сегодняшний день довольно хорошо изучена. Риск-ориентированному банковскому надзору в целях противодействия отмыванию доходов и финансированию терроризма посвящено немало публикаций, например [4–8]. В основном исследования в данной области нацелены на усовершенствование правил внутреннего контроля в банках и реализации принципов «знай своего клиента». В то же время девиантная составляющая деятельности кредитной организации может быть обусловлена не только противоправной деятельностью ее клиентов, но и действиями менеджмента банка. Исследования, посвященные вопросам автоматизации выявления недобросовестных банков в целях противодействия отмыванию доходов и финансированию терроризма, встречаются довольно редко. Данная статья нацелена на заполнение этого пробела, так как в отличие от существующих исследований в ней проанализированы вопросы автоматизации выявления вовлеченности банков в легализацию доходов.

Гипотеза исследования: диагностика вовлеченности банка в процессы отмывания преступных

<sup>2</sup> Справочник по кредитным организациям. URL: [https://www.cbr.ru/banking\\_sector/credit](https://www.cbr.ru/banking_sector/credit) (дата обращения: 14.11.2020).

доходов может проводиться методами машинного обучения.

Цель исследования состоит в обосновании возможностей применения методов и алгоритмов машинного обучения для автоматической идентификации недобросовестных кредитных организаций. Для ее достижения необходимо подобрать математический инструментарий для анализа данных о кредитных организациях, позволяющий проводить диагностику вовлеченности банка в процессы отмывания преступных доходов.

## АНАЛИЗ РАБОТ ОТЕЧЕСТВЕННЫХ И ЗАРУБЕЖНЫХ УЧЕНЫХ ПО ПРИМЕНЕНИЮ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

К решению задачи выявления девиантных объектов финансового мониторинга — компаний-однодневок, банков, служащих площадками по отмыванию доходов, кредитных организаций на грани банкротства и выводящих финансовые активы, недобросовестных участников рынка ценных бумаг — можно подойти как к задаче классификации.

Обратимся к методам и алгоритмам, позволяющим решать подобные задачи в сфере финансового мониторинга.

### Логистическая регрессия

Логистическая регрессия является хорошо изученным и широко применяемым в статистике методом. В современных исследованиях, освещенных в публикациях [9–11], логистическая регрессия применяется в комбинации с другими методами или для сравнения с ними [12, 13].

### Деревья решений

Деревья решений обладают рядом преимуществ при решении задач классификации:

- эффективны в вычислениях и использовании компьютерной памяти, что делает их пригодными для работы с большими объемами данных;
- выбор функций интегрирован в процессы обучения и классификации;
- являются непараметрическими моделями, что позволяет обрабатывать данные с различным распределением.

Одной из существенных проблем при решении задач классификации является несбалансированность классов. Ее решению посвящены исследования [14–18].

*Двухклассовый лес решений (Two-class decision forest)* является конкретной реализацией леса решений, которая работает путем построения нескольких деревьев решений и последующего голосования за самый популярный выходной класс. Голосование является одним из наиболее известных методов получения результатов в ансамблевой модели. Деревья с высокой достоверностью прогноза будут иметь больший вес в окончательном решении ансамбля. Леса решений успешно применяются в социально-экономических исследованиях. Примером может служить работа [19], в которой рассмотрены вопросы обеспеченности населения энергоресурсами.

Алгоритм *AdaBoost* создает ансамбль слабых классификаторов, вызываемых каскадно, каждое следующее дерево исправляет ошибки предыдущего, а предсказания основаны на совокупности деревьев. Алгоритм *AdaBoost* успешно применяется при исследовании данных в области здравоохранения [20], кредитования [21], выявления мошеннических действий с пластиковыми картами [22, 23].

### Метод опорных векторов

Алгоритм двухклассовой машины опорных векторов (*Two-class Support Vector Machine*) создает модель двоичной классификации с использованием алгоритма машины опорных векторов. Двухклассовая машина опорных векторов является алгоритмом обучения с учителем на размеченных данных.

### Нейронные сети

Классификация с использованием нейронных сетей является методом обучения с учителем, поэтому для него требуется набор данных с тегами, который включает столбец меток. Алгоритм двухклассовой нейронной сети (*Two-class neural network*) используется для предсказания бинарных результатов, например есть ли у пациента определенное заболевание, может ли машина выйти из строя в течение определенного периода времени или является ли тот или иной объект финансового мониторинга девиантным.

### Байесовские сети

Байесовские сети применяются в основном для решения диагностических задач. Например, их часто используют в медицине, кредитном скоринге [24, 25] и в других задачах, требующих оценки рисков.

Алгоритм двухклассовой сети Байеса (*Two-class Bayes Point Machine*) использует байесовский подход к линейной классификации, он эффективно



аппроксимирует теоретически оптимальное байесовское среднее для линейных классификаторов (с точки зрения эффективности обобщения), выбирая один «средний» классификатор, точку Байеса.

#### Комбинированные методы

Комбинирование алгоритмов классификации при решении практических задач также может демонстрировать хорошие результаты, примеры представлены в исследованиях [26–28].

В статье [29] предложена концепция объединения алгоритмов бустинга и машины опорных векторов, которая показала высокую производительность. В работе [30] исследовано применение ансамблевого обучения на основе *XGBoost* для обработки данных с носимых датчиков о действиях пожилых людей.

В статье [31] предложен комбинированный классификатор с использованием метода опорных векторов, метода главных компонент и метода  $k$ -ближайших соседей для обработки данных в газоанализаторе. В работе [32] изучена эффективность ансамблей классификаторов в задачах прогнозирования оттока клиентов.

#### Поиск аномалий

Еще одним перспективным направлением решения задач финансового мониторинга является выявление аномалий. Данная группа методов получила широкое распространение в современных условиях.

В задачах поиска аномалий требуется, зная набор значений признаков  $x^{(1)}, x^{(2)}, \dots, x^{(m)}$ , находить объекты, сильно отличающиеся от остальных. Самые популярные задачи поиска аномалий — определение мошеннических транзакций [22, 33–35], выявление технических неполадок [36], обнаружение вторжений в сеть, проверка значений, введенных в систему, поиск нарушителей налогового законодательства и т. д.

Одноклассная машина опорных векторов (*One-Class Support Vector Machine*) создает модель, которая обучается на данных, которые имеют только один класс, являющийся «неаномальным». Он выводит свойства «неаномальных» случаев и по этим свойствам может предсказать, какие кейсы отличаются от обычных.

Алгоритм обнаружения аномалий на основе метода главных компонент (*PCA-Based Anomaly Detection*) анализирует доступные данные, чтобы определить, что представляет собой «нормальный» класс, и применяет метрики расстояния для выявления случаев, которые представляют аномалии

[37, 38]. Для обнаружения аномалий вычисляется проекция значений на собственные векторы вместе с нормализованной ошибкой восстановления. Нормализованная ошибка используется в качестве показателя аномалии. Чем выше ошибка, тем более аномальный случай.

Алгоритмы классификации широко применяются для анализа данных в медицине и здравоохранении [39–41].

Таким образом, можно заключить, что любая предметная область обладает спецификой, а решение практических задач требует подбора математического аппарата, в связи с чем данное исследование актуально.

## МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

В выборку данных для анализа вошли 334 кредитные организации, у 51 из них были отозваны лицензии. Исследуются показатели банковской отчетности формы № 101, в частности рассматриваются следующие группы показателей:

- вложения в ценные бумаги;
- вложения в капиталы других организаций;
- кредиты физическим лицам;
- кредиты предприятиям и организациям;
- просроченная задолженность;
- основные средства и нематериальные активы;
- прочие активы;
- вклады физических лиц;
- средства предприятий и организаций.

Анализируется срез данных за три месяца до отзыва лицензии — на основании многолетней практики противодействия отмыванию доходов данный период признан оптимальным, так как, с одной стороны, девиантная составляющая деятельности банка успевает достаточно ярко проявиться (экспериментально установлено, что она *начинает* проявляться примерно за 6 месяцев до отзыва лицензии), а с другой — еще остается достаточный запас времени для принятия мер.

Проведен ряд экспериментов по обработке данных при помощи алгоритмов бинарной классификации:

- двухклассового усиленного дерева решений;
- двухклассовой машины опорных векторов;
- двухклассовой логистической регрессии;
- двухклассового дерева решений;
- двухклассовой нейронной сети;
- двухклассовой сети Байеса.

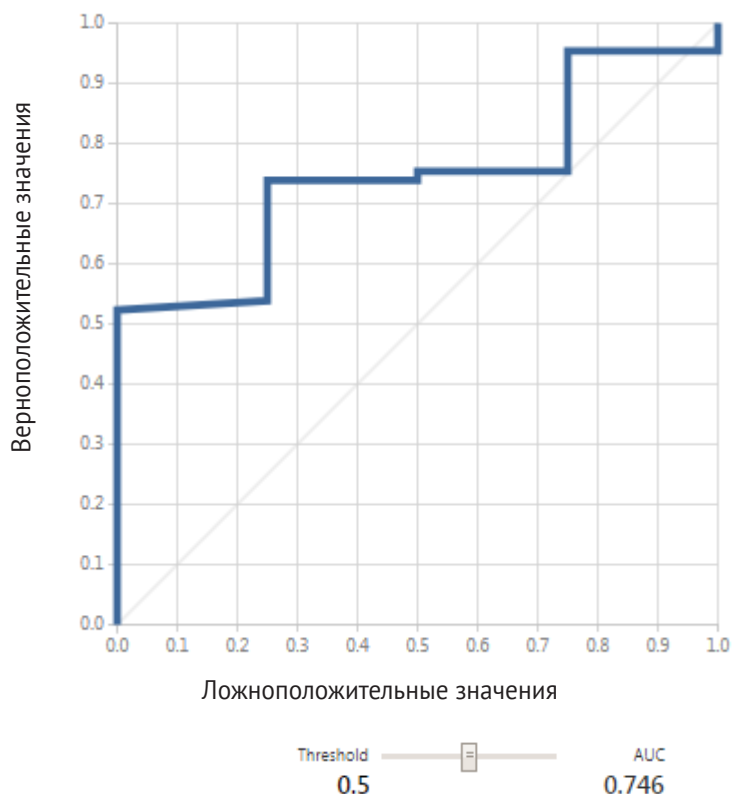


Рис. 1 / Fig. 1. ROC-кривая и показатели точности для модели двухклассовое усиленное дерево решений / ROC curve and accuracy metrics for a Two-Class Boosted Decision Tree model

Источник / Source: составлено автором / compiled by the author.

И для обнаружения аномалий — одноклассовой машины опорных векторов и поиска аномалий на основе метода главных компонент.

Для определения качества полученных моделей воспользуемся ROC-кривой (от англ. *receiver operating characteristic* — рабочая характеристика приемника) — это график, позволяющий оценить качество бинарной классификации. ROC-кривая отображает соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущие признак, и долей объектов от их общего количества, не несущих признака, ошибочно классифицированных как несущие признак.

Количественную интерпретацию ROC-кривой дает показатель *AUC* (от англ. *area under ROC curve* — площадь под ROC-кривой) — площадь, ограниченная ROC-кривой и осью доли ложно положительных классификаций [42–44]. Чем больше *AUC*, тем качественнее алгоритм, равенство показателя 0,5 равносильно случайному гаданию.

Эксперименты выполнялись на платформе для машинного обучения *Microsoft Azure Machine Learning Studio*. В *MS Azure ML* реализованы возможности анализа данных различными методами,

включая методы классификации, регрессии, кластерного анализа и поиска аномалий.

## РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Продемонстрируем классификацию банков по подготовленным данным при помощи алгоритма двухклассового усиленного дерева решений.

На рис. 1 проиллюстрирована ROC-кривая построенной модели, значение показателя точности *AUC* составляет 0,746.

Построим для сравнения двухклассовую машину опорных векторов (см. рис. 2).

Показатель *AUC* для построенной модели равен 0,619.

Построим теперь модели двухклассовой логистической регрессии и двухклассового дерева решений. ROC-кривые представлены на рис. 3.

Показатель точности моделей *AUC* равен 0,588 для модели двухклассовой логистической регрессии и 0,538 — для двухклассового дерева решений соответственно.

Построим модель двухклассовой нейронной сети (*Two class neural network*). ROC-кривая и показатель точности представлены на рис. 4. Показатель точности моделей *AUC* равен 0,654. Отметим,

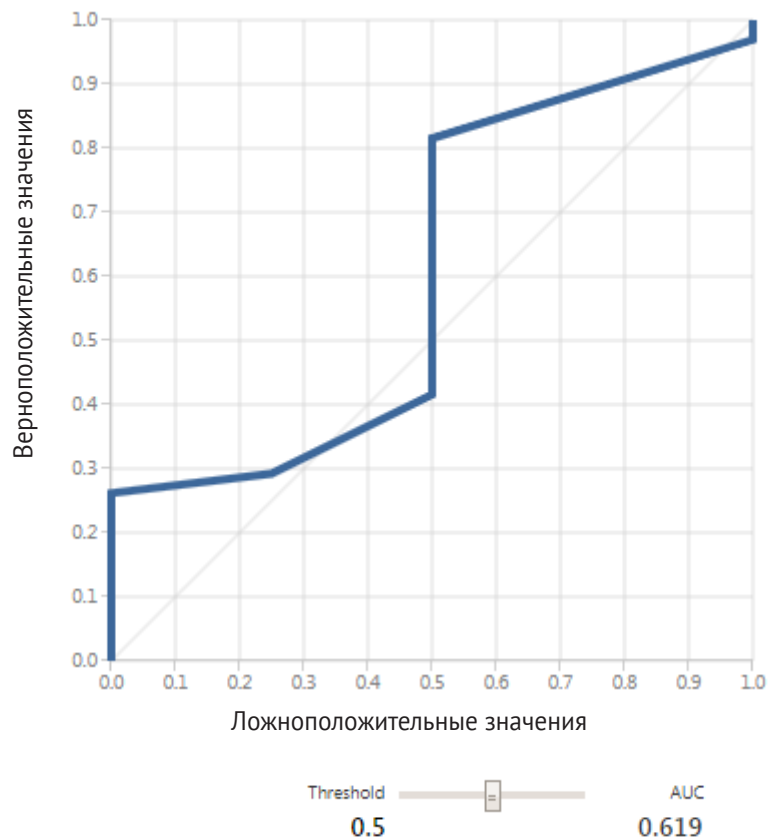


Рис. 2 / Fig. 2. ROC-кривая и показатели точности для модели двухклассовая машина опорных векторов / ROC curve and accuracy metrics for a Two-Class Support Vector Machine model

Источник / Source: составлено автором / compiled by the author.

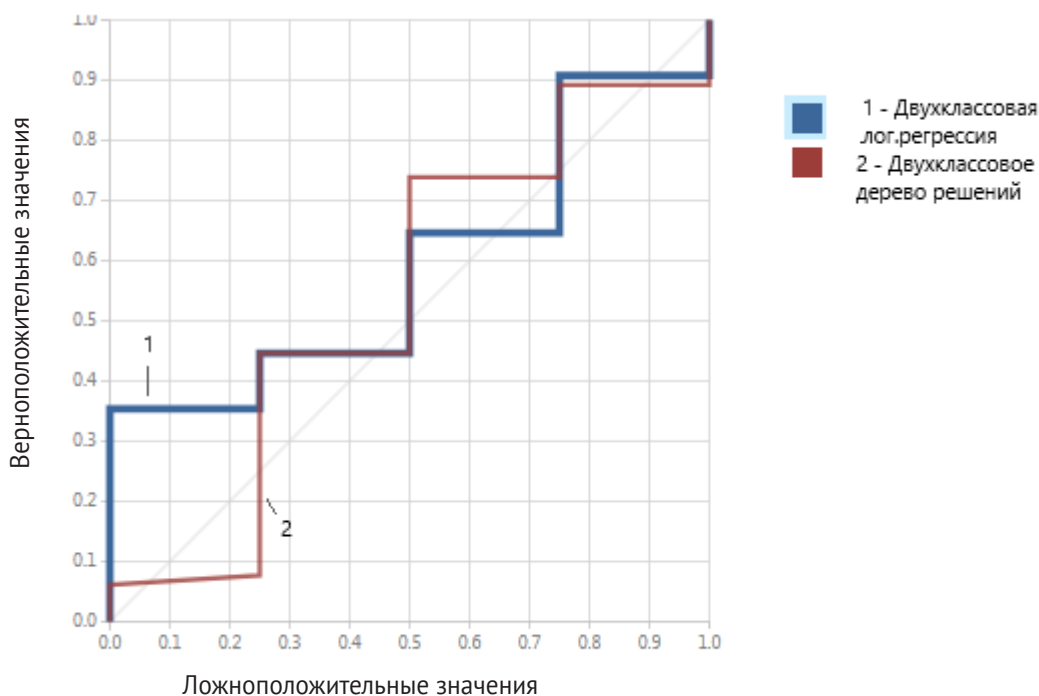


Рис. 3 / Fig. 3. ROC-кривые для моделей двухклассовой логистической регрессии и двухклассового дерева решений / ROC curves for Two-class Logistic Regression and Two-Class Decision Tree models

Источник / Source: составлено автором / compiled by the author.

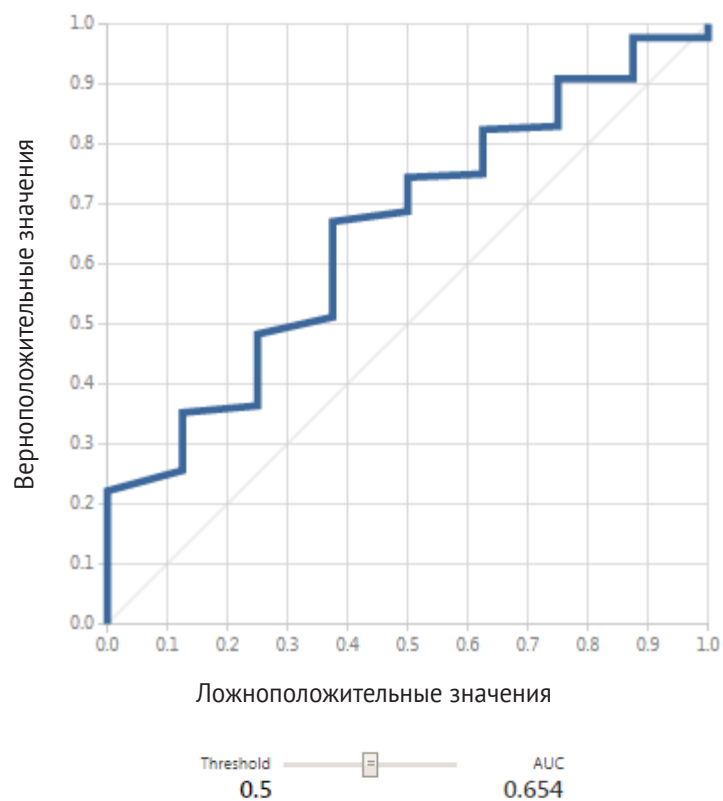


Рис. 4 / Fig. 4. ROC-кривая и показатель точности для двухклассовой нейросети / ROC curve and accuracy metrics for a Two-Class Neural Network

Источник / Source: составлено автором / compiled by the author.

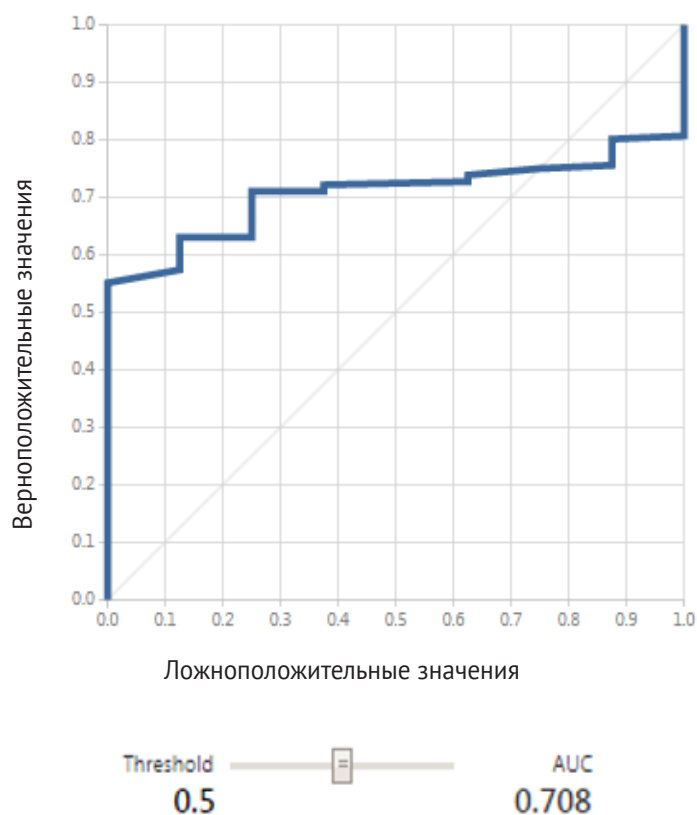


Рис. 5 / Fig. 5. ROC-кривая и показатель точности для двухклассовой сети Байеса / ROC curve and accuracy metrics for a Two-Class Bayes Point Machine

Источник / Source: составлено автором / compiled by the author.



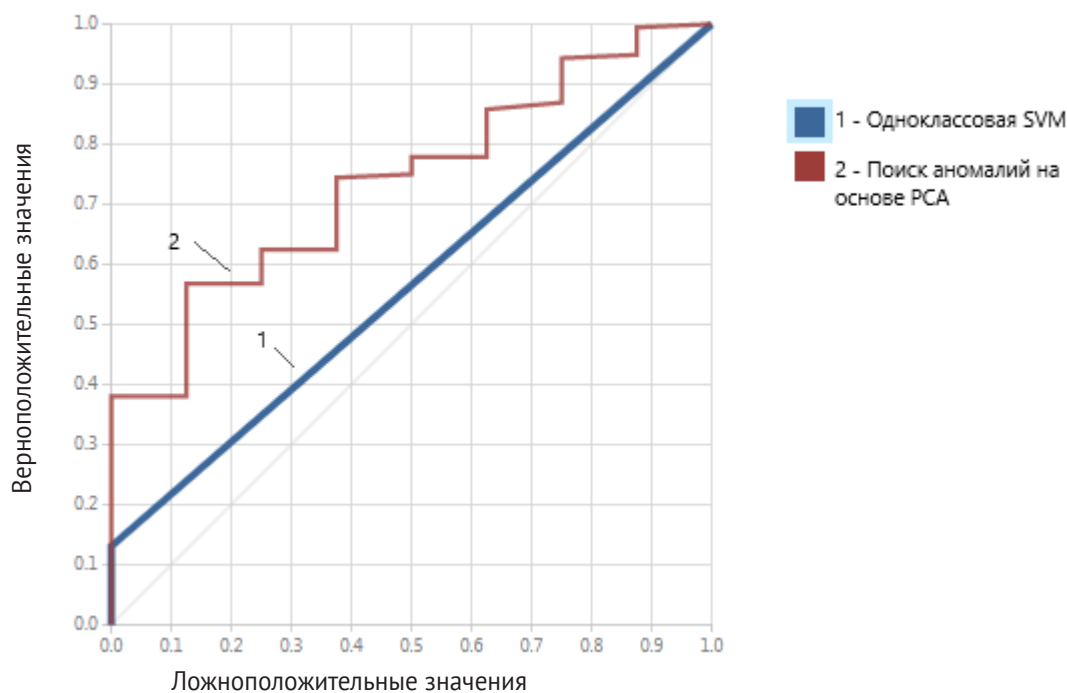


Рис. 6 / Fig. 6. ROC-кривые для моделей одноклассовой машины опорных векторов и поиска аномалий на основе метода главных компонент / ROC curves for One-Class Support Vector Machine models and PCA-Based Anomaly Detection method

Источник / Source: составлено автором / compiled by the author.

что нейросеть построена для трех скрытых слоев. Увеличение числа скрытых слоев привело к ухудшению показателей качества точности модели.

Построим модель двухклассовой сети Байеса (*Two-class Bayes Point Machine*). ROC-кривая и показатель точности представлены на рис. 5. Показатель точности моделей *AUC* равен 0,708.

Теперь рассмотрим алгоритмы поиска аномалий — одноклассовую машину опорных векторов (*One-Class Support Vector Machine*) и алгоритм поиска аномалий на основе метода главных компонент (*PCA-Based Anomaly Detection*). ROC-кривые для моделей одноклассовой машины опорных векторов и поиска аномалий на основе метода главных компонент представлены на рис. 6. Показатель точности моделей *AUC* равен 0,683 для одноклассовой машины опорных векторов и 0,739 — для алгоритма поиска аномалий на основе метода главных компонент соответственно.

В табл. 1 сведены показатели качества по всем реализованным моделям.

Можно сделать вывод, что из рассмотренных алгоритмов классификации наиболее точные результаты показал алгоритм двухклассового усиленного дерева решений, а алгоритм поиска аномалий на основе метода главных компонент дает более точные результаты по сравнению с алгоритмом одноклассовой машины опорных векторов.

Выгрузим предсказанные вероятности отзыва лицензии, полученные с помощью алгоритма *Adaboost*. Результат сравнения сведем в табл. 2.

В целях оценки качества результатов были экспертно выбраны 50 недевиантных кредитных организаций с хорошей репутацией и 100 кредитных организаций, о вовлеченности в сомнительные операции которых известно. Большое количество объектов с наименьшими значениями рейтинга оправдано тем, что в задачах финансового мониторинга, как правило, требуется находить именно девиантные объекты.

Из 100 кредитных организаций, определенных как «девиантные» (у 51 кредитной организации была отозвана лицензия в течение трех месяцев, еще у 46 банков лицензия была отозвана в течение 4–10 месяцев) *Adaboost* также классифицировал как девиантные 82 кредитные организации.

Из 50 кредитных организаций, определенных как «недевиантные», *Adaboost* классифицировал как недевиантные 50 банков.

Таким образом, ошибка первого рода для *Adaboost*:

$$\text{False Positive Rate} = FP / (TN + FP) = 0 / (50 + 0) = 0.$$

Ошибка второго рода:

$$\text{False Negative Rate} = FN / (TP + FN) = 18 / (82 + 18) = 0,18.$$

Таблица 1 / Table 1

Показатели AUC для алгоритмов оценки девиантной составляющей деятельности кредитных организаций / AUC metrics for algorithms for detecting deviations in the activities of credit institutions

	Алгоритмы обнаружения аномалий / Anomaly detection algorithms		Алгоритмы классификации / Classification algorithms					
	Одно-классовая машина опорных векторов	Обнаружение аномалий на основе метода главных компонент	Двух-классовая логистическая регрессия	Двух-классовое дерево решений	Двух-классовое усиленное дерево решений	Двух-классовая нейросеть	Двух-классовая машина опорных векторов	Двухклассовая сеть Байеса
Значение AUC	0,683	0,739	0,588	0,538	0,746	0,654	0,619	0,708

Источник / Source: составлено автором / compiled by the author.

Таблица 2 / Table 2

Результаты оценки кредитных организаций (фрагмент таблицы) / The results of the assessment of credit institutions (fragment of the table)

Наименование банка / Bank	Отозвана ли лицензия Банком России / Has the license been revoked by the Bank of Russia	Девиантный ли банк согласно результатам работы алгоритма Адабуст / Is the bank deviant according to the AdaBoost algorithm
ПАО Сбербанк	Не отозвана	Не девиантный
Банк ВТБ (ПАО)	Не отозвана	Не девиантный
АО «Райффайзенбанк»	Не отозвана	Недевиантный
Банк ГПБ (АО)	Не отозвана	Недевиантный
АО «АЛЬФА-БАНК»	Не отозвана	Недевиантный
ПАО РОСБАНК	Не отозвана	Недевиантный
АО «Россельхозбанк»	Не отозвана	Недевиантный
...		
АО КБ БРТ	Отозвана	Недевиантный
АО Банк РКБ	Отозвана	Девиантный
АО АКБ ГАЗСТРОЙБАНК	Отозвана	Девиантный
АО КБ ТЕТРАПОЛИС	Отозвана	Девиантный
ОАО КБ МВКБ	Отозвана	Девиантный
ООО КБ МЕЖТРАСТБАНК	Отозвана	Девиантный
АО КБ Приско Капитал Банк	Отозвана	Девиантный

Источник / Source: составлено автором / compiled by the author.

## ВЫВОДЫ

Итак, в статье определена специфика анализа объектов финансового мониторинга, в частности кредитных организаций, которая обусловлена высокой латентностью девиантных субъектов и их деятельности, большим объемом и гетерогенным характером требующей анализа и интерпретации информации, высокими требованиями к профессиональным знаниям и практическому опыту экспертов-аналитиков, а также существенными временными затратами на проведение анализа каждого отдельно взятого субъекта.

**Автоматизация процесса выявления недобросовестных кредитных организаций на основе методов машинного обучения позволит контролирующим органам оперативно выявлять и пресекать противоправную деятельность.**

В сфере финансового мониторинга для принятия эффективных управленческих решений необходимо оперативно получать объективные оценки хозяйствующих субъектов (в частности, кредитных организаций), что с учетом специфики предметной области исключает последовательную ручную проверку деятельности субъектов. Автоматизация процесса выявления недобросовест-

ных кредитных организаций на основе методов машинного обучения позволит контролирующим органам оперативно выявлять и пресекать противоправную деятельность.

В целях автоматизации анализа данных о кредитных организациях подобран математический инструментарий, позволяющий проводить диагностику вовлеченности банка в процессы отмывания преступных доходов. Для этого проведен сравнительный анализ результатов обработки данных о деятельности кредитных организаций методами классификации — логистическая регрессия, деревья решений (алгоритмы двухклассовый лес решений, *Adaboost*), метод опорных векторов (алгоритм двухклассовая машина опорных векторов), нейросетевыми методами (алгоритм двухклассовой нейронной сети), Байесовские сети (алгоритм двухклассовой сети Байеса) и поиска аномалий — алгоритм одноклассовой машины опорных векторов и алгоритм обнаружения аномалий на основе метода главных компонент.

Алгоритм поиска аномалий на основе метода главных компонент показал более точные результаты по сравнению с алгоритмом одноклассовой машины опорных векторов. Из рассмотренных алгоритмов классификации наиболее точные результаты показал алгоритм двухклассового усиленного дерева решений (*Adaboost*).

Тем самым подтверждена гипотеза исследования о возможности диагностики вовлеченности банка в процессы отмывания преступных доходов методами машинного обучения.

Приведенные результаты исследования могут быть использованы Банком России и Росфинмониторингом для автоматизации выявления недобросовестных кредитных организаций.

## СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Куркина Е. П., Шувалова Д. Г. Оценка риска: экспертный метод. *Проблемы науки*. 2017;(1):63–69. Kurkina E. P., Shuvalova D. G. Risk assessment: Expert method. *Problemy nauki*. 2017;(1):63–69. (In Russ.).
2. Захарьян А. Г. Экспертная оценка комплексной устойчивости коммерческого банка. *Финансовые исследования*. 2004;(9):14–19. Zakharyan A. G. Expert assessment of the complex sustainability of a commercial bank. *Finansovye issledovaniya*. 2004;(9):14–19. (In Russ.).
3. Бекетнова Ю. М., Крылов Г. О., Денисенко А. С. Проблемы управления и поддержки принятия решений в государственных органах власти на примере Росфинмониторинга. *Информатизация и связь*. 2018;(2):82–88. Beketnova Yu. M., Krylov G. O., Denisenko A. S. The Problems of management and decision support in the government authorities on the example of the Rosfinmonitoring. *Informatizatsiya i svyaz' = Informatization and Communication*. 2018;(2):82–88. (In Russ.).
4. Клочко А. Н., Логвиненко Н. И., Кобзева Т. А., Киселева Е. И. Легализация средств, полученных преступным путем, в сфере банковской деятельности. *Криминологический журнал Байкальского государственного университета экономики и права*. 2016;10(1):194–204. DOI: 10.17150/1996–7756.2016.10(1).194–204

- Klochko A.N., Logvinenko N.I., Kobzeva T.A., Kiselyova E.I. Legalizing proceeds from crime through the banking system. *Kriminologicheskii zhurnal Baikal'skogo gosudarstvennogo universiteta ekonomiki i prava = Criminology Journal of Baikal National University of Economics and Law*. 2016;10(1):194–204. (In Russ.). DOI: 10.17150/1996–7756.2016.10(1).194–204
5. Кононова Н.П., Патласов О.Ю., Кононов Э.Д. Риск-ориентированный подход в сфере противодействия отмыванию доходов и финансированию терроризма. *Наука о человеке: гуманитарные исследования*. 2016;(2):183–189. DOI: 10.17238/issn1998–5320.2016.24.183  
Kononova N.P., Patlasov O. Yu., Kononov E.D. The risk-focused approach in the sphere of counteraction to laundering of the income and to financing terrorism. *Nauka o cheloveke: gumanitarnye issledovaniya = The Science of Person: Humanitarian Researches*. 2016;(2):183–189. (In Russ.). DOI: 10.17238/issn1998–5320.2016.24.183
  6. Кузнецова Е.И., Бурькин Д.В., Мастерова С.А. Риск-ориентированный внутренний контроль кредитных организаций в сфере противодействия легализации доходов, полученных преступным путем. *Вестник экономической безопасности*. 2017;(2):299–302.  
Kuznetsova E.I., Burykin D.V., Masterova S.A. Risk-oriented internal control of credit institutions in the field of combating the legalization of proceeds from crime. *Vestnik ekonomicheskoi bezopasnosti = Vestnik of Economic Security*. 2017;(2):299–302. (In Russ.).
  7. Пряхин Г.Н., Амелешин К.А. Совершенствование методов противодействия легализации доходов, полученных преступным путем, и финансированию терроризма в банковской системе. *Вестник Челябинского государственного университета*. 2019;(3):28–34. DOI: 10.24411/1994–2796–2019–10304  
Pryakhin G.N., Ameleshin K.A. Improvement of methods of countering the legalization of criminal income and financing terrorism in the banking system. *Vestnik Chelyabinskogo gosudarstvennogo universiteta = CSU Bulletin*. 2019;(3):28–34. (In Russ.). DOI: 10.24411/1994–2796–2019–10304
  8. Филатова И.В. Применение риск-ориентированного подхода в целях противодействия легализации (отмыванию) доходов, полученных преступным путем. *Вестник Московского университета МВД России*. 2019;(1):233–236. DOI: 10.24411/2073–0454–2019–10055  
Filatova I.V. Application of a risk-based approach to counteract the legalization (laundering) of proceeds from crime. *Vestnik Moskovskogo universiteta MVD Rossii = Vestnik of Moscow University of the Ministry of Internal Affairs of Russia*. 2019;(1):233–236. (In Russ.). DOI: 10.24411/2073–0454–2019–10055
  9. Pavlidis N.G., Tasoulis D.K., Adams N.M., Hand D.J. Adaptive consumer credit classification. *Journal of the Operational Research Society*. 2012;63(12):1645–1654. DOI: 10.1057/jors.2012.15
  10. Yap B.W., Ong S.H., Husain N.H.M. Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications*. 2011;38(10):13274–13283. DOI: 10.1016/j.eswa.2011.04.147
  11. Khemais Z., Nesrine D., Mohamed M. Credit scoring and default risk prediction: A comparative study between discriminant analysis & logistic regression. *International Journal of Economics and Finance*. 2016;8(4):39. DOI: 10.5539/ijef.v8n4p39
  12. Li Z., Tian Y., Li K., Zhou F., Yang W. Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems with Applications*. 2017;74:105–114. DOI: 10.1016/j.eswa.2017.01.011
  13. Louzada F., Anacleto-Junior O., Candolo C., Mazucheli J. Poly-bagging predictors for classification modelling for credit scoring. *Expert Systems with Applications*. 2011;38(10):12717–12720. DOI: 10.1016/j.eswa.2011.04.059
  14. Siers M.J., Islam M.Z. Class imbalance and cost-sensitive decision trees: A unified survey based on a core similarity. *ACM Transactions on Knowledge Discovery from Data*. 2021;15(1):4. DOI: 10.1145/3415156
  15. Bunkhumpornpat C., Sinapiromsaran K. Density-based majority under-sampling technique. *Knowledge and Information Systems*. 2017;50(3):827–850. DOI: 10.1007/s10115–016–0957–5
  16. Devi D., Biswas S., Purkayastha B. A cost-sensitive weighted random forest technique for credit card fraud detection. In: The 10<sup>th</sup> Int. conf. on computing, communication and networking technologies (ICCCNT). (Kanpur, July 6–8, 2019). New York: IEEE; 2019. DOI: 10.1109/ICCCNT45670.2019.8944885
  17. Zhang S. Multiple-scale cost sensitive decision tree learning. *World Wide Web*. 2018;21(6):1787–1800. DOI: 10.1007/s11280–018–0619–5
  18. Zhu B., Baesens B., vanden Broucke S.K.L.M. An empirical comparison of techniques for the class imbalance problem in churn prediction. *Information Sciences*. 2017;408:84–99. DOI: 10.1016/j.ins.2017.04.015



19. Hurst W., Montanez C.A.C., Shone N. Towards an approach for fuel poverty detection from gas smart meter data using decision tree learning. In: Proc. 2020 3<sup>rd</sup> Int. conf. on information management and management science (IMMS). (London, Aug. 7–9, 2020). New York: Association for Computing Machinery; 2020:23–28. DOI: 10.1145/3416028.3416034
20. Al-Yarimi F.A.M., Munassar N.M.A., Al-Wesabi F.N. Electrocardiogram stream level correlated patterns as features to classify heartbeats for arrhythmia prediction. *Data Technologies and Applications*. 2020;54(5):685–701. DOI: 10.1108/DTA-03-2020-0076
21. Malhotra D.K., Malhotra K., Malhotra R. Evaluating consumer loans using machine learning techniques. In: Lawrence, K.D., Pai D.R., eds. *Applications of Management Science*. Bingley: Emerald Publishing Ltd; 2020;20:59–69. DOI: 10.1108/S 0276-897620200000020004
22. Yang Y., Liu C., Liu N. Credit card fraud detection based on CSat-related AdaBoost. In: Proc. 2019 8<sup>th</sup> Int. conf. on computing and pattern recognition (ICCP'19). (Beijing, Oct. 23–25, 2019). New York: Association for Computing Machinery; 2019:420–425. DOI: 10.1145/3373509.3373548
23. Tran P.H., Tran K.P., Huong T.T., Heuchenne C., Tran P.H., Le T.M.H. Real time data-driven approaches for credit card fraud detection. In: Proc. 2018 Int. conf. on e-business and applications (ICEBA 2018). (Da Nang, Feb. 23–25, 2018). New York: Association for Computing Machinery; 2018:6–9. DOI: 10.1145/3194188.3194196
24. Wu W.-W. Improving classification accuracy and causal knowledge for better credit decisions. *International Journal of Neural Systems*. 2011;21(4):297–309. DOI: 10.1142/S 0129065711002845
25. Zhu H., Beling P.A., Overstreet G.A. A Bayesian framework for the combination of classifier outputs. *Journal of the Operational Research Society*. 2002;53(7):719–727. DOI: 10.1057/palgrave.jors.2601262
26. Marqués A.I., García V., Sánchez J.S. Two-level classifier ensembles for credit risk assessment. *Expert Systems with Applications*. 2012;39(12):10916–10922. DOI: 10.1016/j.eswa.2012.03.033
27. Vukovic S., Delibasic B., Uzelac A., Suknovic M. A case-based reasoning model that uses preference theory functions for credit scoring. *Expert Systems with Applications*. 2012;39(9):8389–8395. DOI: 10.1016/j.eswa.2012.01.181
28. Akkoç S. An empirical comparison of conventional techniques, neural networks and the three-stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research*. 2012;222(1):168–178. DOI: 10.1016/j.ejor.2012.04.009
29. Turchetti Maia T., Pádua Braga A., de Carvalho A.F. Hybrid classification algorithms based on boosting and support vector machines. *Kybernetes*. 2008;37(9/10):1469–1491. DOI: 10.1108/03684920810907814
30. Wu Y., Qi S., Hu F., Ma S., Mao W., Li W. Recognizing activities of the elderly using wearable sensors: A comparison of ensemble algorithms based on boosting. *Sensor Review*. 2019;39(6):743–751. DOI: 10.1108/SR-11-2018-0309
31. Faleh R., Gomri S., Othman M., Aguir K., Kachouri A. Enhancing WO<sub>3</sub> gas sensor selectivity using a set of pollutant detection classifiers. *Sensor Review*. 2018;38(1):65–73. DOI:10.1108/SR-12-2016-0273
32. Lee S.-C., Faloutsos C., Chae D.-K., Kim S.-W. On detecting frauds in comparison-shopping services. In: Proc. 26<sup>th</sup> Int. conf. on world wide web companion (WWW'17 Companion). (Perth, Apr. 3–7, 2017). Geneva: IWWWC Steering Committee; 2017:811–812. DOI: 10.1145/3041021.3054219
33. Sohony I., Pratap R., Nambiar U. Ensemble learning for credit card fraud detection. In: Proc. ACM India joint int. conf. on data science and management of data (CoDS-COMAD'18). (Goa, Jan. 11–13, 2018). New York: Association for Computing Machinery; 2018:289–294. DOI: 10.1145/3152494.3156815
34. Lucas Y., Portier P.-E., Laporte L., Calabretto S., Caelen O., He-Guelton L., Granitzer M. Multiple perspectives HMM-based feature engineering for credit card fraud detection. In: Proc. 34<sup>th</sup> ACM/SIGAPP Symp. on applied computing (SAC'19). (Limassol, Apr. 8–12, 2019). New York: Association for Computing Machinery; 2019:1359–1361. DOI: 10.1145/3297280.3297586
35. Li Q., Xie Y. A behavior-cluster based imbalanced classification method for credit card fraud detection. In: Proc. 2019 2<sup>nd</sup> Int. conf. on data science and information technology (DSIT 2019). (Seoul, July 19–21, 2019). New York: Association for Computing Machinery; 2019:134–139. DOI: 10.1145/3352411.3352433
36. Ray S., Wright A. Detecting anomalies in alert firing within clinical decision support systems using Anomaly/Outlier Detection Techniques. In: Proc. 7<sup>th</sup> ACM Int. conf. on bioinformatics, computational biology, and health informatics (BCB'16). (Seattle, Oct. 2–5, 2016). New York: Association for Computing Machinery; 2016:185–190. DOI: 10.1145/2975167.2975186



37. Geiger B. C., Kubin G. Relative information loss in the PCA. In: Proc. IEEE information theory workshop (ITW). (Lausanne, Sept. 3–7, 2012). New York: IEEE; 2012:562–566. DOI: 10.1109/ITW.2012.6404738
38. Howard M. C. A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*. 2016;32(1):51–62. DOI: 10.1080/10447318.2015.1087664
39. Khan H., Srivastav A., Mishra A. K. Use of classification algorithms in health care. In: Tanwar P., Jain V., Liu C.-M., Goyal V., eds. Big data analytics and intelligence: A perspective for health care. Bingley: Emerald Publishing Ltd; 2020:31–54. DOI: 10.1108/978-1-83909-099-820201007
40. Deepa B. G., Senthil S. Constructive effect of ranking optimal features using Random Forest, Support Vector Machine and Naïve Bayes for breast cancer diagnosis. In: Tanwar P., Jain V., Liu C.-M., Goyal V., eds. Big data analytics and intelligence: A perspective for health care. Bingley: Emerald Publishing Ltd; 2020:189–202. DOI: 10.1108/978-1-83909-099-820201014
41. Ramaswami G., Susnjak T., Mathrani A., Lim J., Garcia P. Using educational data mining techniques to increase the prediction accuracy of student academic performance. *Information and Learning Sciences*. 2019;120(7/8):451–467. DOI: 10.1108/ILS-03-2019-0017
42. Эскиндаров М. А., Соловьев В. И., ред. Парадигмы цифровой экономики: Технологии искусственного интеллекта в финансах и финтехе. М.: Когито-Центр; 2019. 325 с.  
Eskindarov M. A., Soloviev V. I., eds. Paradigms of the digital economy: Artificial intelligence technologies in finance and FinTech. Moscow: Cogito-Center; 2019. 325 p. (In Russ.).
43. Lee J.-S. AUC 4.5: AUC-based c4.5 decision tree algorithm for imbalanced data classification. *IEEE Access*. 2019;7:106034–106042. DOI: 10.1109/ACCESS.2019.2931865
44. Zhang Q., Wang J., Lu A., Wang S., Ma J. An improved SMO algorithm for financial credit risk assessment — Evidence from China’s banking. *Neurocomputing*. 2018;272:314–325. DOI: 10.1016/j.neucom.2017.07.002

## ИНФОРМАЦИЯ ОБ АВТОРЕ / ABOUT THE AUTHOR



**Юлия Михайловна Бекетнова** — кандидат технических наук, доцент Департамента информационной безопасности, Финансовый университет, Москва, Россия  
**Yuliya M. Beketnova** — Cand. Sci. (Eng.), Assoc. Prof., Information Security Department, Financial University, Moscow, Russia  
 beketnova@mail.ru

*Статья поступила в редакцию 25.12.2020; после рецензирования 03.03.2021; принята к публикации 17.08.2021.*

*Автор прочитала и одобрила окончательный вариант рукописи.*

*The article was submitted on 25.12.2020; revised on 03.03.2021 and accepted for publication on 17.08.2021.*

*The author read and approved the final version of the manuscript.*