Научная статья
УДК 336.2
https://doi.org/10.36511/2588-0071-2024-4-60-68.

Методика выявления уклоняющихся от уплаты налогов организаций на основе метода «Случайный лес»

Литвиненко Александр Николаевич[1], Гармышева Александра Александровна[2]

[1, 2]Санкт-Петербургский университет МВД России, Санкт-Петербург, Россия
[1]Lanfk@mail.ru
[2]garmysheva.a@yandex.ru, https://orcid.org/0009-0005-1867-1042

Аннотация

В статье изложена авторская методика выявления уклоняющихся от уплаты налогов организаций. Методика основывается на анализе финансовых и нефинансовых факторов деятельности компаний с помощью применения метода машинного обучения «Случайный лес». Результаты исследования могут быть использованы правоохранительными и налоговыми органами как инструмент противодействия уклонению от уплаты налогов.

Ключевые слова: налоги, методика выявления, уклонение от уплаты налогов, «случайный лес», машинное обучение

Для цитирования

Литвиненко А. Н., Гармышева А. А. Методика выявления уклоняющихся от уплаты налогов организаций на основе метода «Случайный лес» // На страже экономики. 2024. № 4 (31). С. 60–68. https://doi.org/10.36511/2588-0071-2024-4-60-68.

Original article

Methodology for identifying tax evading organizations based on the method “Random Forest”

Alexander N. Litvinenko[1], Alexandra A. Garmysheva[2]

[1, 2]Saint Petersburg University of the Ministry of Internal Affairs of Russia, Saint Petersburg, Russian Federation
[1]Lanfk@mail.ru
[2]garmysheva.a@yandex.ru, https://orcid.org/0009-0005-1867-1042

Abstract

The article describes the author’s methodology for identifying tax evading organizations. The methodology is based on the analysis of financial and non-financial factors of companies’ activities using the “Random Forest” machine learning method. The results of the study can be used by law enforcement and tax authorities as a tool to counteract tax evasion.

Keywords: taxes, detection methods, tax evasion, Random Forest, machine learning

For citation

Litvinenko A. N., Garmysheva A. A. Methodology for identifying tax evading organizations based on the method “Random Forest”. *The Economy under Guard,* 2024, no. 4 (31), pp. 60–68. (In Russ.). https://doi.org/10.36511/2588-0071-2024-4-60-68.

Введение

Несмотря на предпринимаемые в Российской Федерации меры, проблема уклонения от уплаты налогов остается актуальной и требует постоянного внимания со стороны налоговых и правоохранительных органов. Выявление данных противоправных деяний имеет первостепенную важность для экономической безопасности страны, так как незаконное уменьшение налоговых обязательств налогоплательщиками влечет сокращение поступлений в бюджеты всех уровней, тем самым ставит под угрозу реализацию задач и функций государства.

Проблема выявления участников, уклоняющихся от уплаты налогов, имеет множество аспектов, главными из которых являются:

1) латентность;
2) большой объем данных и их обработка;
3) технологические вызовы и анонимизация.

Эффективное выявление и пресечение таких незаконных действий требуют комплексного подхода, включающего внедрение передовых технологий, совершенствование законодательства и улучшение координации между государственными органами. Однако ключевым аспектом современной борьбы с налоговыми преступлениями является применение новых технологий для выявления уклонений от уплаты налогов.

Анализ больших данных (*Big Data*), машинное обучение, искусственный интеллект (далее — ИИ), блокчейн могут быть применены для автоматизации процессов анализа информации, выявления аномалий, построения моделей поведения налогоплательщиков и прогнозирования рисков.

Целью научного исследования является создание методики, позволяющей выявлять организации, уклоняющиеся от уплаты налогов.

Выбор метода исследования

Традиционные математико-статистические методы многомерного анализа, которые проводятся в области экономических исследований, остаются фундаментом научных исследований и позволяют оценить степень взаимосвязь их систем показателей, дать представление о стохастических связях, прогнозировать [1, с. 50]. Однако для эффективного решения актуальных проблем, возникающих в эпоху стремительно развивающихся цифровых технологий, необходима разработка предложений, идущих в ногу со временем. Одним из таких является машинное обучение, которое основывается на обучении от данных и при неизменном наборе данных позволяет избегать однонаправленности результатов, что неминуемо в математико-статистических методах, основанных на строгих математических моделях и предположениях о распределении данных. Принимая во внимание преимущества и недостатки рассмотренных методов (см. табл. 1), мы сделали выбор в пользу метода машинного обучения.

Таблица 1

Матрица преимуществ и недостатков математико-статистических методов и методов машинного обучения

Table 1

The matrix of advantages and disadvantages of mathematical and statistical methods and machine learning methods

Метод	+	-
Математико-статистические методы	Интерпретируемость; теоретическая обоснованность; эффективность в случае ограниченного количества данных; контроль за ошибками	Неэффективность для больших объемов данных; трудность работы с неструктурированными данными; неприменимость к сложным явлениям; однонаправленность результатов
Методы машинного обучения	Автоматизация; обработка больших объемов данных; адаптивность; способность извлекать сложные закономерности; возможность улучшения с опытом	Неинтерпретируемость; опасность переобучения или недообучения; зависимость от качества данных; необходимость большого объема данных

Одним из набирающих популярность методов машинного обучения является метод «случайного леса» (*Random Forest*). Он относится к классу ансамблевых методов, которые объединяют несколько моделей для улучшения качества прогнозов. «Случайный лес» — это комбинация деревьев решений, где каждое дерево строится на основе подмножества обучающих данных и случайного подмножества признаков.

Таким образом, метод «Случайный лес» объединяет прогнозы множества деревьев решений, что часто приводит к более точным и стабильным результатам. Путем обработки больших объемов данных выявляются закономерности, паттерны и тенденции, которые используются для предсказания значений переменной ответа. Данный метод применяется в различных задачах по выявлению отклонений на основе анализа финансовых и нефинансовых показателей [2–4].

Содержание методики

Методика выявления уклоняющихся от уплаты налогов организаций состоит из следующих этапов:

1. Отбор финансовых и нефинансовых факторов деятельности организаций, косвенно указывающих на уклонение от уплаты налогов.
2. Анализ первичной информации о финансовых и нефинансовых факторах деятельности организаций.
3. Построение ансамбля решающих деревьев с использованием случайного подмножества переменных-предикторов для выявления уклоняющихся от уплаты налогов организаций.
4. Проверка достоверности построенной модели.

Отбор факторов деятельности организаций

На первом этапе на основе анализа официальных документов Федеральной налоговой службы Российской Федерации [5–8] были отобраны восемь факторов и показателей деятельности организаций, способных косвенно указывать на потенциальную возможность уклонения от уплаты налогов (табл. 2).

Таблица 2

Факторы, указывающие на направленность действий налогоплательщика на неправомерное уменьшение налоговой обязанности

Table 2

Factors indicating the direction of the taxpayer's actions towards an unlawful reduction of tax liability

Фактор	Показатель
Налоговая нагрузка [5]	Относительная налоговая нагрузка (ОНН): ОНН = Н/В*100, где Н — сумма начисленных налогов за год; В — выручка компании за год. Высокая/средняя/низкая — сравнивается по среднеотраслевому показателю (виду экономической деятельности)
Связи организации [6]	Количество связей организации
Финансовая устойчивость [5]	Коэффициент финансовой устойчивости $К_{фу} = \frac{СК+ДО}{ВБ}$, где СК — собственный капитал; ДО — долгосрочные обязательства; ВБ — валюта баланса
Признаки фирмы-однодневки [7]	Присутствуют/отсутствуют
Налоговые риски [5]	Присутствуют/отсутствуют
Выручка [5]	Выручка: $В = \frac{Ц}{К}$, где Ц — цена товара или услуги; К — количество проданных товаров или оказанных услуг
Проверки [5]	Количество проведенных проверок в отношении компании
Применение специального налогового режима [8]	Применяет / не применяет

Анализ первичной информации о финансовых и нефинансовых факторах

На втором этапе проведен анализ первичной информации о данных показателях деятельности 240 организаций, зарегистрированных в Санкт-Петербурге. В процессе выбора факторы были проверены на энтропию Шенона для разделения данных в узлах деревьев решений, входящих в состав случайного леса. С помощью кластерного анализа, проведенного на платформе *Loginom*, каждый показатель

На страже экономики ◆ 2024 ◆ № 4 (31) ◆ The Economy under Guard

разбит по нескольким диапазонам, заданным машинным алгоритмом. Далее собранные данные случайным образом разделены на тренировочную и тестовую выборки: 80 % компаний предназначены для построения обучающей модели «Случайный лес» и 20 % — для проверки ее достоверности.

Построение модели «Случайный лес»

Третий этап заключается в построении модели «Случайный лес», позволяющей определять организации, уклоняющиеся от уплаты налогов. В качестве технического способа реализации методики выбрана *low-code* платформа *Loginom*. Данный инструмент проводит анализ данных любого уровня сложности без программирования, имеет интуитивно понятный интерфейс и активную техническую поддержку. Главным преимуществом использования данной платформы является доступ к специальной библиотеке *Loginom Python Kits*, на которой изложены готовые компоненты для обучения и прогона алгоритмов и моделей, доступных в библиотеке *Scikit-learn*. Таким образом, платформа не обязывает пользователя иметь глубокие знания в программировании и позволяет использовать преднастроенные компоненты. Отметим, что выбор программного обеспечения и языка программирования пользователь может сделать самостоятельно, исходя из своих запросов. Это не повлияет на реализацию методики.

Код для построения случайного леса на *Python* с использованием *Scikit-learn:*

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.datasets import load_iris
# Загрузка набора данных
iris = load_iris()
X = iris.data
y = iris.target
# Разделение набора данных на обучающий и тестовый
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
# Создание модели случайного леса
model = RandomForestClassifier(n_estimators=5, random_state=0)
# Обучение модели
model.fit(X_train, y_train)
# Оценка точности модели на тестовом наборе данных
accuracy = model.score(X_test, y_test)
print(f'Точность модели: {accuracy}')
```

Оптимальное количество деревьев в модели «Случайного лес» зависит от конкретной задачи, размера набора данных и других параметров. Приведенный код создает модель случайного леса с пятью деревьями и оценивает ее точность на тестовом наборе данных. Необходимое достаточное количество деревьев установлено с помощью кросс-валидации, которая позволяет оценить производительность модели: насколько хорошо модель обобщает данные, не переобучаясь на конкретном наборе данных. Основная идея кросс-валидации заключается в том, что данные разбиваются на несколько подмножеств, называемых фолдами, и модель обучается и оценивается несколько раз, каждый раз используя разные фолды в качестве тестового набора данных. Благодаря кривой обучаемости

мы можем проследить зависимость метрики качества модели от количества деревьев (рис. 1). Это позволяет определить момент, когда увеличение количества деревьев (более пяти) перестает значительно улучшать качество модели.

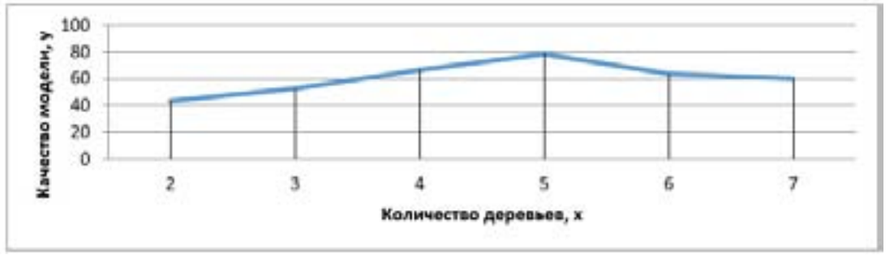

Рис. 1. Кривая обучения зависимости метрики качества модели от количества деревьев

Fig. 1. The learning curve of the dependence of the model quality metric on the number of trees

Визуализируем единичное дерево случайного леса (рис. 2), изучив которое, мы можем понять, как работает модель.

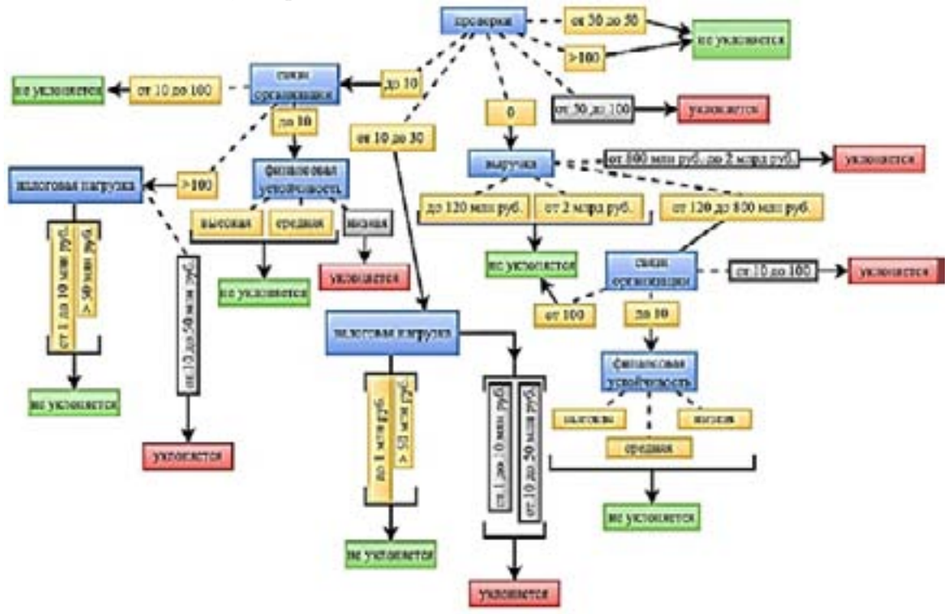

Рис. 2. Модель единичного дерева из ансамбля деревьев «Случайного леса» по выявлению уклоняющихся от уплаты налогов организаций

Fig. 2. The model of a single tree from an ensemble of trees of a "Random Forest" to identify tax evading organizations

Проверка достоверности построенной модели

На последнем этапе проверена достоверность построенной модели. В этих целях по предложенной модели проверено 20 % компаний, которые ранее были неизвестны машинному алгоритму. Достоверность оказалась достаточно высокой: 41 из 48 организаций метод «Случайный лес» охарактеризованы верно.

Применение методики

Для применения методики выявления уклоняющихся от уплаты налогов организаций пользователю необходимо сформировать необходимый пакет данных и загрузить их на вход модели «Случайный лес». Каждое дерево сделает собственный прогноз, а затем результаты прогнозов будут усреднены, и среднее значение станет окончательным предсказанием. Достоинствами разработанной методики являются автоматизация построения модели, адаптивность к любым изменениям и понятность, что облегчает использование и интерпретацию результатов для пользователей без специальных знаний в области программирования. В дальнейшем повышение эффективности метода «Случайный лес» возможно за счет расширения показателей и обучения на большем количестве уже известных данных о компаниях.

Заключение

Таким образом, разработанная методика, основанная на применении метода машинного обучения «Случайный лес», представляет собой эффективный инструмент противодействия уклонению от уплаты налогов, позволяющий анализировать финансовые и нефинансовые факторы, свидетельствующие о налоговых правонарушениях.

Результаты исследования могут быть использованы правоохранительными и налоговыми органами в целях оперативного выявления организаций, уклоняющихся от уплаты налогов. Это способствует снижению уровня теневой экономики и защите экономических интересов страны.

Список источников

1. Большакова Л. В., Яковлева Л. В. Современные математико-статистические методы обработки информации в научной и практической работе // Проблемы современной науки и образования. 2016. № 7 (49). С. 49–52.

2. Митяков Е. С. Машинное обучение в задачах обеспечения экономической безопасности // Развитие и безопасность. 2020. № 4 (8). С. 92–105.

3. Мячин Н. В., Ахмедов Т. Ч. Сценарии применения метода «дерево решений» на рынке микрофинансовых услуг // Современные проблемы обеспечения экономической безопасности хозяйствующего субъекта: сборник научных трудов Всероссийской научно-практической конференции (Москва, 17 февраля 2022 года). Москва: Московский университет Министерства внутренних дел Российской Федерации имени В. Я. Кикотя, 2022. С. 166–169.

4. Татевосян А. С. Методика выявления потенциальных участников нарушений внешнеэкономической деятельности на основе метода дерева решений // Социальные и экономические системы. 2023. № 5-1 (47). С. 128–135.

5. Об утверждении Концепции системы планирования выездных налоговых проверок: приказ Федеральной налоговой службы России от 30 мая 2007 года № ММ-3-06/333@ (в ред. от 10 мая 2012 года) // Финансовая газета. 21 июня 2007 год, 31 мая 2012 год.

6. О практике применения статьи 54.1 Налогового кодекса Российской Федерации: письмо Федеральной налоговой службы Российской Федерации от 10 марта 2021 года № БВ-4-7/3060@. URL: https://www.nalog.gov.ru/rn77/about_fts/about_nalog/10687108/ (дата обращения: 26.01.2024).

7. О рассмотрении обращения: письмо Федеральной налоговой службы Российской Федерации от 11 февраля 2010 года № 3-7-07/84 // СПС «КонсультантПлюс» URL: https://www.consultant.ru/document/cons_doc_LAW_98034/ (дата обращения: 26.01.2024).

8. О направлении обзора судебной практики, связанной с обжалованием налогоплательщиками ненормативных актов налоговых органов, вынесенных по результатам мероприятий налогового контроля, в ходе которых установлены факты получения необоснованной налоговой выгоды путем формального разделения (дробления) бизнеса и искусственного распределения выручки от осуществляемой деятельности на подконтрольных взаимозависимых лиц: Письмо Федеральной налоговой службы Российской Федерации от 11 августа 2017 года № СА-4-7/15895@ // Доступ из СПС «Гарант». URL: https://www.garant.ru/products/ipo/prime/doc/71654502/ (дата обращения: 26.01.2024).

References

1. Bolshakova L. V. Yakovleva L. V. Modern mathematical and statistical methods of information processing in scientific and practical work. *Problems of modern science and education*, 2016, no 7 (49), pp. 49–52. (In Russ)

2. Mityakov E. S. Machine learning in the tasks of ensuring economic security. *Development and security*, 2020, no 4 (8), pp. 92–105. (In Russ)

3. Myachin N. V., Akhmedov T. Ch. Scenarios for the application of the decision tree method in the microfinance services market. Modern problems of ensuring economic security of an economic entity: collection of scientific papers of the All-Russian Scientific and practical Conference (Moscow, February 17, 2022). Moscow: Moscow University of the Ministry of Internal Affairs of the Russian Federation named after V. Ya. Kikot, 2022. Pp. 166–169. (In Russ)

4. Tatevosyan A. S. Methodology for identifying potential participants in violations of foreign economic activitypatterns based on the decision tree method. *Social and economic systems*, 2023, no 5-1 (47), pp. 128–135. (In Russ)

5. On approval of the Concept of the on-site tax audit planning system: order of the Federal Tax Service of Russia no. MM-3-06/333@ of 30 May, 2007 (as amended. Of dated 10 May, 2012). *Financial Gazette*, 21 May, 2007, no 25; 31 May, 2012, no 22. (In Russ)

6. On the practice of applying Article 54.1 of the Tax Code of the Russian Federation: letter of the Federal Tax Service of Russia no. BV-4-7/3060@ of 3 Octomber, 2021. URL: https://www.nalog.gov.ru/rn77/about_fts/about_nalog/10687108 / (accessed 26.01.2024). (In Russ)

7. On consideration of the appeal: letter of the Federal Tax Service of Russia no. 3-7-07/84 of 11 February, 2010. Access from the reference legal system “ConsultantPlus” URL: https://www.consultant.ru/document/cons_doc_LAW_98034 / (accessed 26.01.2024). (In Russ)

8. On the direction of the Review of Judicial Practice Related to the Appeal by Taxpayers of Non-Normative Acts of Tax Authorities Issued as a Result of Tax Control Measures, during which the facts of obtaining unjustified tax benefits through formal division (fragmentation) of business and artificial distribution of proceeds from Activities carried out on con-

trolled interdependent persons: letter of the Federal Tax Service of Russia no. CA-4-7/15895@ of 8 November, 2017. URL: https://www.garant.ru/products/ipo/prime/doc/71654502. Access from the reference legal system "Garant". (accessed 26.01.2024). (In Russ)

Информация об авторах | Information about the authors

А. Н. Литвиненко — доктор экономических наук, профессор, профессор кафедры экономической безопасности Санкт-Петербургского университета МВД России

A. N. Litvinenko — Doctor of Sciences (Economy), Professor, Professor of the Department of Economic Security of the St. Petersburg University of the Ministry of Internal Affairs of Russia

А. А. Гармышева — адъюнкт кафедры экономической безопасности Санкт-Петербургского университета МВД России

A. A. Garmysheva — Postgraduate of the Department of Economic Security of the St. Petersburg University of the Ministry of Internal Affairs of Russia

Статья поступила в редакцию 23.08.2024; одобрена после рецензирования 05.10.2024; принята к публикации 12.12.2024.

The article was submitted 23.08.2024; approved after reviewing 05.10.2024; accepted for publication 12.12.2024.