ОРИГИНАЛЬНАЯ СТАТЬЯ

CC BY 4.0

DOI: 10.26794/2587-5671-2022-26-5-132-148
УДК 336.051/336.64(045)
JEL G32, C14, C63

Оценка стоимости компании с использованием методов машинного обучения

П.С. Коклев
Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

АННОТАЦИЯ

Целью работы является тестирование гипотезы о возможности применения методов машинного обучения для разработки моделей, позволяющих точно спрогнозировать рыночную капитализацию предприятия на основании данных трех основных форм финансовой отчетности: отчета о финансовых результатах, бухгалтерского баланса, отчета о движении денежных средств. **Научная новизна** исследования заключается в предложении альтернативного подхода к актуальной проблеме финансов — оценке стоимости компании. Проверить рассматриваемую гипотезу позволяет проведенное эмпирическое исследование, в рамках которого осуществлялось создание моделей с использованием наиболее популярных **методов** машинного обучения (Lasso, Elastic Net, KNN, Random Forest, SVM и др.). С целью определения лучшего подхода для оценки бизнеса сопоставлена эффективность различных методов на основании показателя R^2 (86,7% для метода *GBDT*). Информационной базой выступают данные отчетности компаний, котирующихся на биржах *NYSE* и *NASDAQ*. В ходе исследования автор также решает проблему интерпретируемости полученных моделей. Выявлены наиболее важные признаки — формы финансовой отчетности и конкретные их статьи, оказывающие наибольшее влияние на капитализацию предприятия. Три независимых способа оценки важности признаков указывают на особую ценность информации, содержащейся в данных отчета о финансовых результатах. В частности, наиболее значимыми для высокоточных прогнозов оказались данные о совокупном доходе (Comprehensive income). Также выделены наиболее предпочтительные методы трансформации и вменения отсутствующих данных для финансовой отчетности. Рекомендованы различные способы усовершенствования разработанных моделей с целью достижения еще более высокой точности оценок. Сделан **вывод**, что машинное обучение можно применять как более точный, непредвзятый и менее затратный способ оценки стоимости компании. Анализ важности признаков может быть использован для понимания и дальнейшего изучения процесса формирования стоимости компании.

Ключевые слова: оценка стоимости компании; относительная оценка; DCF; машинное обучение; искусственный интеллект; big data; регрессионный анализ; градиентный бустинг; деревья решений

Для цитирования: **Коклев П.С. Оценка стоимости компании с использованием методов машинного обучения. *Финансы: теория и практика.* 2022;26(5):132-148. DOI: 10.26794/2587-5671-2022-26-5-132-148**

ORIGINAL PAPER

Business Valuation with Machine Learning

P.S. Koklev
Saint Petersburg State University, Saint Petersburg, Russia

ABSTRACT

The aim of the article is to test the hypothesis about the applicability of machine learning **methods** to train models that allow to accurately predict the market capitalization of an enterprise based on data contained in three main forms of financial statements: *Income statement, Balance sheet, and Cash flow statement*. **The scientific novelty** of the study lies in the proposal of an alternative approach to the actual finance problem — business valuation. The conducted empirical study allows us to test the hypothesis under consideration. We train various models using the most popular machine learning **methods** (*LASSO*, *Elastic Net*, *KNN*, *Random Forest*, SVM, and others). To determine the best approach for assessing the value of a company, the effectiveness of different methods is compared based on the R^2 performance metric (86,7% for the *GBDT*). Financial statements data of *NYSE* and *NASDAQ* companies are used. The study also addresses the problem of the interpretability of the trained models. The most important features are identified — the forms of financial statements and their specific items that have the greatest impact on market capitalization. Three independent ways to determine feature importance indicate the significance of the information contained in the *Income statement*. In particular, *Comprehensive income* was the most important item for accurate predictions. Robust methods of variable normalization and missing data imputation are also highlighted. Finally, various ways of improving the developed

© Коклев П.С., 2022

models are recommended to achieve even higher accuracy of forecasts. The study **concludes** that machine learning can be applied as a more accurate, unbiased, and less costly approach to value a company. Feature importance analysis can also be used to understand and further explore the value creation process.
Keywords: business valuation; relative valuation; DCF; machine learning; artificial intelligence; big data; regression analysis; gradient boosting; decision trees

For citation: **Koklev P.S. Business valuation with machine learning. *Finance: Theory and Practice*. 2022;26(5):132-148. DOI: 10.26794/2587-5671-2022-26-5-132-148**

ВВЕДЕНИЕ

Отмечая способность искусственного интеллекта оказывать сильнейшее воздействие и полностью менять облик отраслей экономики и областей научного знания, его по праву называют электричеством XXI в[1]. Финансовый сектор и теория финансов также ощутили на себе влияние искусственного интеллекта. Машинное обучение, воплощая в себе самую продуктивную форму реализации идеи искусственного интеллекта на практике, уже довольно давно применяется финансовыми институтами для решения разнообразных задач. Первые исследования использования нейронных сетей для прогнозирования динамики движения цен фондового рынка опубликованы еще в 90-х гг. прошлого столетия (Kryzanowski, Galler, & Wright, 1993). Именно финансовый сектор стал основным «полигоном» для испытания новых методов. Наиболее благоприятным фактором для этого стало изобилие и разнообразие данных, агрегируемых финансовыми институтами в рамках своей обыкновенной деятельности. Кроме этого, именно финансовые рынки видятся наиболее привлекательным объектом приложения интеллектуальных усилий ведущих специалистов информационных и математических наук, ведь вознаграждение за успешное применение статистических методов на этом поприще будет максимальным.

Ежегодно публикуются сотни работ, фиксирующих случаи успешного применения методов искусственного интеллекта в финансовом секторе (Cao, 2020). Однако некоторые темы незаслуженно остаются без внимания. Так, аспекты применения машинного обучения для оценки стоимости компании недостаточно рассмотрены научным сообществом. Это может быть связано с несколькими причинами, одна из которых — сложность моделирования субъективного процесса оценки аналитиком стоимости предприятия. Некоторые исследователи полагают, что процесс оценки не может быть формализован (Damodaran, Investment Valuation: Tools and Techniques for Determining the Value of Any Asset, 2012). Тем не менее именно способность инкорпорировать нелинейные взаимоотношения между переменными, являющимися свойством сложных процессов (например, процесса формирования стоимости предприятия) — одна из главных сильных сторон машинного обучения. Доказано, что некоторые методы статистического обучения позволяют аппроксимировать непрерывные функции любого типа и сложности (Cybenko, 1989). Именно поэтому сложность и субъективность процесса формирования стоимости компании являются не преградой, а поводом для применения методов машинного обучения (*ML, Machine Learning*) для прогнозирования рыночной капитализации компании[2].

Данная работа посвящена проверке гипотезы о возможности применения методов машинного обучения с целью прогнозирования капитализации предприятия. Для этого проведено эмпирическое исследование и последующий анализ, в рамках которого сопоставлена эффективность разнообразных подходов контролируемого обучения для прогнозирования/присвоения рыночной капитализации компании на основании данных ее финансовой отчетности за восемь последних кварталов. То есть данных бухгалтерского баланса, отчета о финансовых результатах и отчета о движении денежных средств.

Первая часть основного раздела посвящена краткому обзору литературы и предпосылок для применения машинного обучения в контексте исследуемой проблемы. Далее рассмотрены методологические аспекты эмпирического исследования. Наконец, само исследование на основании данных 3945 компаний *NASDAQ* и *NYSE* и его итоги обсуждены в третьей части. В результате сравнительного анализа определены наиболее эффективные методы

[1] Сайт Stanford Business. URL: https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity (дата обращения: 29.11.2021).

[2] В рамках исследования под стоимостью компании понимается сумма рыночной капитализации компании и балансовой стоимости обязательств. Таким образом, подразумевая наличие данных балансовой стоимости обязательств, проблема прогнозирования стоимости компании идентична проблеме прогнозирования рыночной капитализации.

обучения, позволяющие прогнозировать стоимость компании с высокой точностью. Эффективность методов сопоставлена с помощью показателя R^2, рассчитанного для тестового множества, т.е. данных, не используемых при обучении моделей.

Высокая точность получаемых прогнозов может быть использована для решения разнообразных задач. Оценка стоимости частных предприятий, определение цены размещения акции перед *IPO*, создание инвестиционных стратегий[3], формирование новых подходов к отражению инвестиций в капитал других предприятий в бухгалтерском учете — далеко не полный список возможных точек приложения полученных в ходе исследования результатов. Помимо самих моделей, ценностью обладает и анализ важности признаков, позволяющий выделить независимые переменные — формы финансовой отчетности и конкретные их статьи, которые наиболее сильно влияют на итоговый прогноз стоимости. Это позволяет выявить самую ценную информацию, содержащуюся в финансовой отчетности, что, в свою очередь, может быть использовано как компаниями для более полного раскрытия наиболее релевантной для инвесторов информации, так и регулятором при разработке новых стандартов. Кроме того, важность признаков дает подсказки относительно самого процесса формирования стоимости компании. Наконец, в ходе работы автору удалось определить лучшие методы для предварительной обработки данных финансовой отчетности: нормализации независимых переменных и вменения отсутствующих данных.

1. МАШИННОЕ ОБУЧЕНИЕ ДЛЯ ОЦЕНКИ СТОИМОСТИ КОМПАНИИ

1.1. Предпосылки использования ML методов для оценки стоимости компании

Рассмотрим основные факторы, позволяющие предположить, что статистическое обучение может быть успешно использовано для определения капитализации компании. Используя термин «машинное обучение», как правило, имеют в виду следующее:

1. Применение разнообразного набора непараметрических статистических методов прогнозирования, способных инкорпорировать нелинейные отношения между независимыми переменными.

2. Использование регуляризации — способа штрафования сложных моделей с целью предотвращения проблемы переобучения[4].

3. Подбор оптимального набора гиперпараметров среди множества возможных вариантов спецификации модели. *Grid Search* и *Random Search* — наиболее популярные алгоритмы поиска оптимального набора гиперпараметров. Подбор гиперпараметров, как правило, осуществляется с использованием валидационного множества, получаемого в ходе перекрестной проверки (*k-fold cross-validation*).

Помимо фокуса на применении передовых алгоритмов, машинное обучение акцентирует внимание на верификации модели с помощью перекрестной проверки и использовании тестовой выборки для получения непредвзятой оценки качества модели. Таким образом смещается фокус с оценки параметров модели $\hat{\beta}$ на результат прогноза $\hat{y}$. Для оценки стоимости компании первичным является определение функции $h(x)$, позволяющей осуществлять высокоточный прогноз. Именно прогноз, а не оценка коэффициентов $\hat{\beta}$ имеет большую прикладную ценность. Специализируясь именно на задачах прогнозирования, машинное обучение идеально подходит для рассматриваемой проблемы. Многомерная природа *ML* методов позволяет заметно превосходить в гибкости традиционные эконометрические подходы. Эта гибкость помогает лучше аппроксимировать неизвестную функцию, отражающую процесс, формирующий стоимость компании.

Некоторые свойства методов машинного обучения делают их идеальными кандидатами для моделирования процессов с неизвестной или неопределенной формой. В первую очередь это разнообразие. Даже в рамках одного из множеств непохожих друг на друга методов исследователь имеет возможность выбирать бесконечное множество различных спецификаций модели. Например, гиперпараметр коэффициент скорости обучения (*learning rate*) может принимать бесконечное множество значений из области положительных действительных чисел. Помимо этого, отбор лучших моделей с помощью перекрестной проверки позволяет контролировать проблему переобучения и избегать ложных открытий, являющихся следствием неконтролируемого тестирования множества различных вариантов спецификации моделей (Gu, Kelly, & Xiu, 2020).

[3] В контексте разработки инвестиционных стратегий положительная разность между прогнозной и фактической стоимостью будет трактоваться как недооценка, а отрицательная — как переоценка компании.

[4] Переобучение — явление, когда построенная модель хорошо объясняет стоимость компании из обучающей выборки, но не способна делать качественные прогнозы для компаний, не участвующих в обучении модели.

Вопреки своей первоначальной специализации, аналитические методы, разработанные для работы с большими данными, особенно эффективны при работе с небольшими наборами данных (3945 наблюдений считается небольшим набором данных). Они заметно превосходят традиционные методы для социальных наук: *наименьших квадратов* (*МНК*) для проблем регрессии и *логистической регрессии* для задач классификации, которые не способны в своем первоначальном виде учитывать нелинейное воздействие независимых переменных на зависимую. Способность инкорпорировать многомерную природу данных является основной причиной превосходства непараметрических методов искусственного интеллекта.

Другим недостатком традиционных методов является склонность к переобучению. Полученная с помощью *МНК* модель делает, как правило, крайне неточные предсказания за пределами выборки. Кроме этого, интерпретируемость результатов перестала считаться преимуществом *МНК*. Использование метода регуляризации Тихонова (*Ridge*) (Тихонов, 1963), регрессии *Lasso* (Tibshirani, Regression shrinkage and selection via the lasso, 1996), а также их комбинации — *Elastic Net* позволяет увеличить качество прогнозирования, сохраняя возможность простой интерпретации полученных коэффициентов $\hat{\beta}$. С учетом всех недостатков некоторые исследователи полагают, что использование простой линейной регрессии в социальных науках должно быть сведено к минимуму (Hindman, 2015).

Используемые в исследовании объясняющие переменные — статьи финансовой отчетности — зачастую похожи друг на друга и являются сильными коррелятами[5]. Последствием нестрогой мультиколлинеарности является высокое значение стандартной ошибки для оцениваемых коэффициентов, что лишает *МНК* главного преимущества — интерпретируемости. Кроме того, *МНК* перестает работать, когда количество переменных стремится к числу наблюдений. Использование *ML* позволяет успешно работать с паталогическими для традиционных методов данными и отбирать жизнеспособные модели.

Обладая огромным потенциалом для прогнозирования стоимости компании, машинное обучение все же имеет некоторые недостатки и ограничения. Полученные в ходе применения моделей прогнозы являются измерениями. Сами по себе измерения не указывают на фундаментальный механизм, формирующий рыночную капитализацию компании. Процесс присвоения прогноза зачастую непрозрачен. Даже специалистам довольно сложно доступным языком описать логику формирования того или иного прогноза. Сложность коммуникации является одним из барьеров к применению машинного обучения к некоторым проблемам финансов. К счастью, проблеме интерпретируемости результатов уделяется особенное внимание в литературе. Современные методы оценки важности признаков (*feature importance*), рассмотренные в эмпирическом исследовании, позволяют нивелировать проблему «черного ящика» (Carvalho, Pereira, & Cardoso, 2019).

1.2. Обзор литературы

Важным вопросом является определение независимых переменных, с помощью которых должно осуществляться прогнозирование стоимости. Теория оценки стоимости активов дает вполне определенные указания. Академики и практики используют множество разнообразных подходов к оценке стоимости компании. Ведущий исследователь в этой области А. Дамодаран выделяет четыре подхода (Damodaran, Valuation approaches and metrics: a survey of the theory and evidence, 2007).

В первую очередь это *дисконтирование денежных потоков*, DCF. Данный подход определяет стоимость путем дисконтирования ожидаемых денежных потоков, генерируемых активами компаний.

Другой способ — метод бухгалтерской оценки — использует данные балансовой стоимости активов.

Третий и наиболее часто применяемый на практике метод — относительной оценки, подразумевающий использование рыночных данных стоимости похожих компаний (Pinto, Robinson, & Stowe, 2019). Для присвоения стоимости компании используют фондовые мультипликаторы.

Наконец, в рамках метода непредвиденного требования применяют модели оценки стоимости финансовых опционов.

Благодаря прочным теоретическим основам, именно дисконтированию денежных потоков уделено особое внимание в научном сообществе. Основы подхода предложены А. Маршаллом и О. Бём-Баверком, которые рассматривали концепцию приведенной стоимости в первой половине XX в. Бём-Баверк первым продемонстрировал расчет стоимости аннуитета в явной форме (Böhm-Bawerk,

[5] Примером коррелированных переменных могут служить показатели выручки и валовой прибыли. Более экстремальный пример — валюта баланса на дату последнего ежеквартального отчета и валюта баланса на дату предпоследнего ежеквартального отчета. Вследствие использования данных за восемь последних кварталов практически каждая из переменных является сильным коррелятом с семью другими переменными из матрицы плана X.

1903). В основе метода лежит предположение о том, что активы с крупными и прогнозируемыми денежными потоками должны обладать большей стоимостью, нежели активы с низкими и волатильными денежными потоками. Другими словами, стоимость актива является возрастающей функцией от размера ожидаемых денежных потоков и убывающей от ставки дисконтирования, отражающей риск и неопределенность денежных потоков.

Необходимо отметить, что на практике, применяя метод DCF для определения ожидаемых денежных потоков, прогнозирования темпов их роста, рентабельности капитала и др., используют данные именно финансовой отчетности. Следовательно, и обучение статистических моделей, прогнозирующих стоимость компании, должно осуществляться с использованием этих данных.

Настоящая работа имеет связь с целым рядом исследований 2000-х гг., посвященных приложению машинного обучения к финансовым рынкам. В 2009 г. Д. Атсалакис и К. Валаванис рассмотрели существующую литературу, посвященную использованию нейронных сетей для прогнозирования динамики фондового рынка (Atsalakis & Valavanisb, 2009). Позднее, в 2018 г., Ф. Чинг, Э. Камбриа и Р. Уэлш опубликовали обзор применения методов обработки естественного языка, *NLP* для финансового прогнозирования (Xing, Cambria, & Welsch, 2018). Среди работ, непосредственно связанных с оценкой стоимости активов, выделяется работа 2015 г. Б. Парка и Д. Бае, посвященная использованию машинного обучения, в частности алгоритма *AdaBoost* для прогнозирования стоимости жилой недвижимости в штате Вирджиния (Park & Bae, 2015).

Наиболее близкими по смыслу к прогнозированию капитализации предприятия с помощью статистических методов являются работы, посвященные прогнозированию фондовых мультипликаторов. В одной из первых публикаций по данной тематике М. Кисор и В. Уайбек использовали данные темпа роста прибыли, коэффициента дивидендных выплат и стандартного отклонения изменения прибыли на акцию (*EPS*) для определения отношения капитализации к чистой прибыли, мультипликатора P/E (Whitbeck & Kisor, 1963). Выборка состояла из 135 компаний. Полученное в 1963 г. уравнение регрессии имеет следующий вид:

$$P/E = 8,2 + 1,5g + 6,7\left(Payout\ ratio\right) - 0,2\sigma_{eps}.$$

Прогнозирование мультипликаторов для компаний российского фондового рынка рассматривалось в работе (Коклев, 2020). Автор отдает предпочтение прогнозированию мультипликатора $EV/Sales$, аргументируя это тем, что положительное значение выручки позволяет использовать максимально возможную выборку компаний.

Х. Йоши и Р. Чауха использовали детерминанты мультипликаторов для прогнозирования значения коэффициентов с помощью регрессии *Ridge* и *Lasso*. Скорректированный р-квадрат составил 70%. Отдельная тестовая выборка для верификации результатов не использовалась (Joshi & Chauha, 2020).

М. Ле Клер, А. Алфорд, С. Пенман, Д. Ниссим, А. Дамадоран и многие другие авторы разрабатывали модели прогнозирования мультипликаторов (Liu, Nissim, & Thomas, 2002). В рамках этих исследований независимыми переменными, как правило, являются различные коэффициенты, интегральные показатели, характеризующие деятельность, финансовую устойчивость и эффективность предприятия.

2. МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

С целью проверки рассматриваемой гипотезы необходимо решить следующую задачу: обучить модель оценки рыночной стоимости акционерного капитала предприятия. Не имея априорной информации относительно эффективности того или иного подхода, видится целесообразным рассмотреть максимально возможное число существующих на сегодняшний день методов машинного обучения. Определение лучшего метода можно осуществить путем сопоставления точности прогнозов для данных тестового множества. В наиболее общем виде мы описываем прогнозную стоимость компании как:

$$\hat{y} = h\left(x\right),$$

где h — функция стоимости компании, полученная методом машинного обучения; $\hat{y}$ — прогноз рыночной капитализации компании; x — вектор признаков компании, состоящий из данных финансовой отчетности.

2.1. Используемые методы машинного обучения

Существует довольно широкий спектр качественной литературы, подробно описывающий каждый подход как с точки зрения описания метода, так и самих вычислительных алгоритмов, позволяющих получить модель для заданной обучающей выборки (Hastie, Tibshirani, & Friedman, 2016). Поэтому не станем подробно останавливаться на каждом из рассмотренных в работе методов, а лишь кратко перечислим их. В *табл. 1* сведены двенадцать разнообразных подходов, призванных максимально репрезентативно представить

Таблица 1 / Table 1

Рассмотренные методы обучения моделей / Employed machine learning methods

Метод / Method	Имплементация алгоритма / Algorithm implementation	Функция потерь / Loss function	Количество исследованных спецификаций / Number of specifications investigated
Метод наименьших квадратов, МНК	scikit-learn*	MSE	10 000
Гребневая регрессия, Ridge	scikit-learn	$MSE + \alpha \|w\|_2^2$	10 000
Лассо регрессия, Lasso	scikit-learn	$MSE + \alpha \|w\|_1$	2000
Elastic Net	scikit-learn	$MSE + \alpha\rho \|w\|_1 + \frac{\alpha(1-\rho)}{2} \|w\|_2^2$	2000
Стохастический градиентный спуск, SGD	scikit-learn	MSE, Huber loss, epsilon insensitive, squared epsilon insensitive	2000
Регрессия с функцией потерь Хьюбера, Huber	scikit-learn	Huber loss	2000
Метод опорных векторов, SVM	scikit-learn	epsilon-insensitive	3000
Метод k-ближайших соседей, KNN	scikit-learn	–	2000
Дерево решений, Decision Tree, DT	scikit-learn	MSE, MAE, Poisson loss	10000
Случайный лес, Random Forest, RF	scikit-learn	MSE, MAE, Poisson loss	500
Сверхслучайные деревья, Extremely Randomized Trees, ERT	scikit-learn	MSE, MAE	500
Градиентный бустинг на основе деревьев решений, GBDT	CatBoost**	MSE	100

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

Примечание / Note: * Scikit-learn — библиотека для машинного обучения в Python. URL: https://scikit-learn.org/ (дата обращения: 15.11.2021); ** CatBoost — библиотека для обучения моделей деревьев решений на основе градиентного бустинга, разработанная ПАО Яндекс. URL: https://catboost.ai/ (дата обращения: 17.11.2021) / * Scikit-learn is a library for machine learning in Python. URL: https://scikit-learn.org/ (accessed on 15.11.2021); ** CatBoost is a gradient boosting on decision trees library, developed by Yandex. URL: https://catboost.ai/ (accessed on 17.11.2021).

идею использования *ML* для оценки стоимости компании.

Помимо метода, указаны и программные библиотеки, используемые для построения моделей в рамках каждого метода: *scikit-learn* для традиционных алгоритмов и *CatBoost* для градиентного бустинга. Помимо *CatBoost,* существует несколько других популярных библиотек имплементации алгоритма градиентного бустинга на основе деревьев решений: *XGBoost, H2O* и *LightGBM* от *Microsoft*. Выбор в пользу проекта *Яндекса* объясняется более высокой точностью получаемых моделей как при значении гиперпараметров по умолчанию, так и при их оптимизированном значении. Помимо

этого, процесс обучения с помощью *CatBoost* проходит, как правило, быстрее (Prokhorenkova, Gusev, Vorobev, Dorogush, & Gulin, 2017). В последнем столбце указано количество протестированных моделей в рамках каждого метода. Тестирование большого числа возможных спецификаций необходимо для определения наилучшего набора гиперпараметров. Отметим, что в *табл. 1* отсутствуют *нейронные сети*. Из-за неоднородной природы данных[6] финансовой отчетности этот метод оказался неэффективным ($R^2 = -8,0\%$).

2.2. Оценка качества моделей

Существуют серьезные основания полагать, что результаты регрессии зачастую могут являться следствием неконтролируемого тестирования множества вариантов спецификации модели (Ioannidis & Doucouliagos, 2013). Для получения желаемого результата исследователь задним числом выбирает «лучшую» модель. Формирование отдельной тестовой выборки лишает недобросовестного ученого опции подгона модели под данные с целью получения сенсационных результатов. В рамках настоящего исследования генеральная совокупность разделяется на два несовместных множества: обучающее и тестовое. Тестовое множество состоит из случайной выборки 20% данных — 788 компаний. Именно р-квадрат, рассчитанный для компаний i, принадлежащих тестовой выборке τ, будет являться справедливым показателем, характеризующим качество прогнозов. Сравнение эффективности различных *ML* методов будет осуществляться с помощью данного показателя.

$$R^2_{oos} = 1 - \frac{\sum_{i \in \tau} \left(y_i - \hat{y}_i \right)^2}{\sum_{i \in \tau} \left(y_i - \bar{y}_i \right)^2}.$$

Данные остальных 3155 компаний будут использоваться для обучения модели и перекрестной проверки в процессе оптимизации гиперпараметров с применением алгоритма случайного поиска (Random Search). Суть случайного поиска состоит в сопоставлении множества различных спецификаций в рамках метода путем случайного подбора гиперпараметров из заданного распределения (Bergstra & Bengio, 2012). Выбор лучшей модели из множества различных вариантов спецификации осуществляется на основе показателя р-квадрат R^2_{cv}, рассчитанного для валидационной выборки, представляющей собой динамически меняющуюся часть обучающей. Для валидационной выборки будет последовательно изыматься десятая часть обучающей.

2.3. Важность признаков

Различные способы оценки важности признаков призваны нивелировать проблему интерпретируемости, позволяя оценить вклад каждой независимой переменной — статьи финансовой отчетности как в общее качество модели, так и в конкретный прогноз. Рассмотрим три различных подхода:

1. Изменение значения прогноза (Prediction Values Change).
2. Определение важности с помощью перестановок (Permutation Importance).
3. SHAP (SHapley Additive exPlanations).

В рамках первого метода важность переменной определяется с помощью расчета среднего изменения прогноза при изменении признака. Чем больше среднее изменение прогноза, тем большая важность присваивается признаку. Этот процесс выполняется для каждой переменной.

Способ перестановки признаков (*Permutation Importance*) работает следующим образом: произвольным образом перемешиваются наблюдения для оцениваемой переменной. То есть случайным образом перетасовывается один столбец матрицы плана X, строки которой состоят из векторов признаков компаний. Остальные столбцы остаются неизменными. После перестановки рассчитываются прогнозные значения капитализации уже на основе измененной матрицы. Переменная признается важной, если точность прогнозов значительно снижается по сравнению с первоначальной, неизмененной матрицей. С другой стороны, признак считается неважным, если его перестановка не привела к значительному снижению R^2. Так, процесс случайной перестановки повторяется триста раз для каждой переменной (Breiman, 2001). Высокое число итераций перестановок для каждого признака необходимо для получения узких доверительных интервалов оценки среднего снижения качества модели.

Метод *SHAP* представляет собой подход, мотивированный идеей *Вектора Шепли*, принципа из теории игр, позволяющему вычислить оптимальное распределение выигрыша между игроками, учитывая их вклад в итоговый результат (Shapley,

[6] Неоднородность данных заключается в том, что различные статьи финансовой отчетности несут в себе разный экономический смысл. Нейронные сети, напротив, наиболее эффективны при работе с однородными данными. Например, изображениями, где каждый признак представляет собой пиксель.

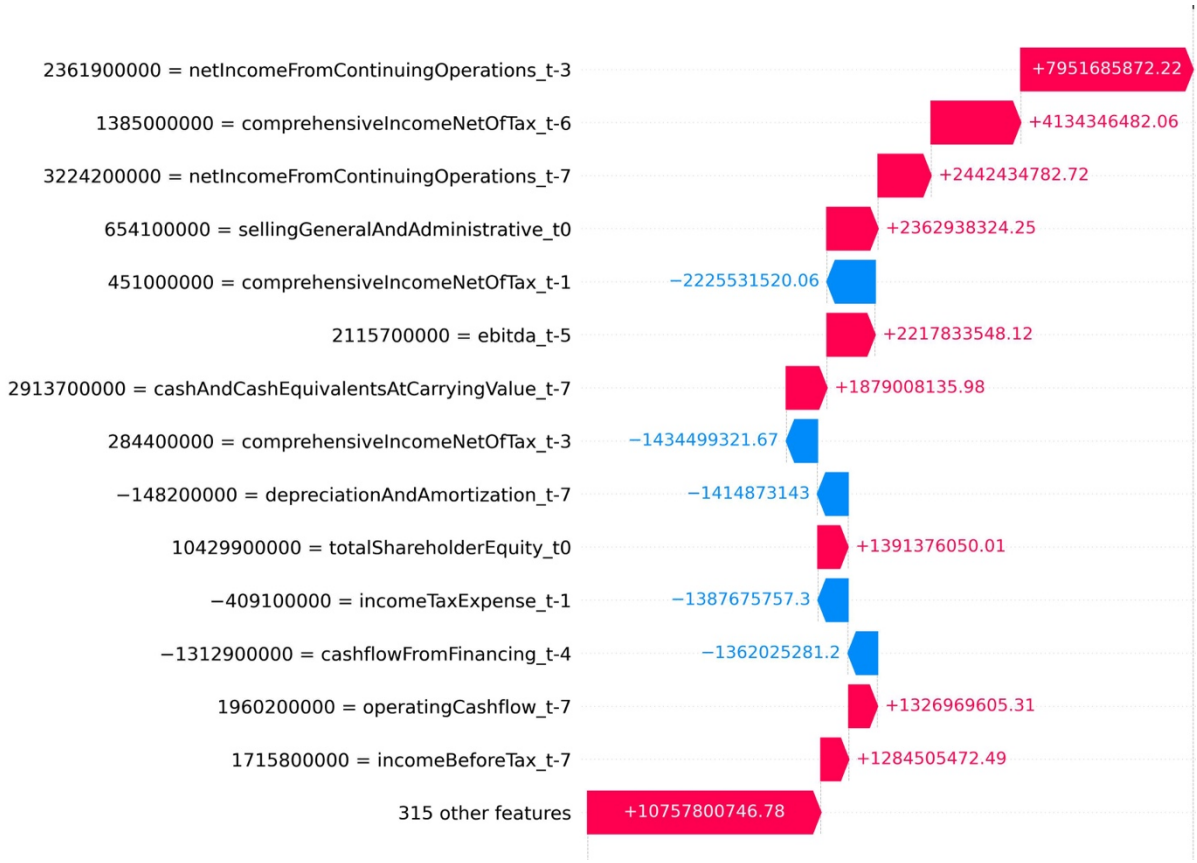

Рис. 1 / Fig. 1. **Оценка важности признаков методом SHAP для компании *Biogen Inc.* Локальные значения и вклад признаков указаны в долл. США. Модель GBDT / SHAP values based on feature importance for *Biogen Inc.* Numbers are given in US dollars. GBDT model**

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

2016). В контексте интерпретации моделей статистического обучения *SHAP* позволяет получить локальный (для конкретного наблюдения x) аддитивный для каждой статьи финансовой отчетности вклад в итоговый прогноз капитализации по сравнению с базовым уровнем (средним прогнозом модели). Иллюстрация принципа на примере биофармацевтической компании *Biogen Inc.* показана на *рис. 1*.

По вертикальной оси перечислены статьи финансовой отчетности и их значения, отсортированные по убыванию абсолютному значению *SHAP*. То есть представлены переменные, оказавшие наибольшее влияние на прогноз капитализации компании *Biogen*. Так, согласно методу, значение в 2,36 млрд долл. *Чистой операционной прибыли* три квартала назад (*netIncomeFromContinuingOperations_t-3*) увеличило прогнозную капитализацию компании на 7,95 млрд долл. по сравнению с базовым уровнем.

Наибольшее негативное влияние оказало значение *Совокупного дохода* в прошлом квартале (*comprehensiveIncomeNetOfTax_t-1*), понизившее прогнозную стоимость компании на 2,23 млрд долл. Суммарный вклад локальных значений для остальных 315 признаков, не представленных на графике индивидуально, составляет 10,76 млрд долл. с положительным знаком (последняя строка на *рис. 1*). Итоговый прогноз капитализации в данном примере составил 42,4 млрд долл.

Помимо анализа прогноза для отдельно взятой компании, возможно оценить и общую важность каждого из признаков путем расчета среднего абсолютного значения *SHAP* на основе всех наблюдений. Для оценки важности признаков в рамках каждого из трех методов сделан выбор в пользу использования всей генеральной совокупности, состоящей из объединения обучающего и тестового множеств.

3. ЭМПИРИЧЕСКОЕ ИССЛЕДОВАНИЕ НА ОСНОВЕ ДАННЫХ NYSE И NASDAQ

3.1. Данные

Независимыми переменными, с помощью которых прогнозируется капитализация, являются данные ежеквартальной финансовой отчетности (форма 10-Q) за последние восемь кварталов, полученные в ноябре 2021 г. Генеральная совокупность состоит из 3945 компаний, котирующихся на биржах *NYSE* и *NASDAQ*. Финансовые отчеты

предоставлены сервисом *Alpha Vantage*[7]. Всего матрица плана X содержит 329 столбцов, 328 из которых являются статьями бухгалтерского баланса, отчета о финансовых результатах и о движении денежных средств. Также набор предикторов включает в себя информацию о секторальной принадлежности компании. Так, вектор признаков для отдельной компании имеет следующий вид:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ \dots \\ x_{329} \end{pmatrix} = \begin{pmatrix} \text{Всего активов}_{t=0} \\ \text{Всего текущих активов}_{t=0} \\ \dots \\ \dots \\ \text{Сумма денежных средств на конец года}_{t=-7} \end{pmatrix}.$$

3.1.1. Отсутствующие данные

Для большинства компаний существуют пропущенные значения для некоторых признаков. Причин отсутствующих данных может быть несколько. Так, степень детализации раскрытия информации может варьироваться от компании к компании. С другой стороны, специфика хозяйственной деятельности фирмы может не подразумевать наличия конкретной статьи. К примеру, компания, не использующая финансовый лизинг, будет иметь пропущенные данные для признака capital Lease Obligations в каждом из восьми кварталов. Практически каждый из рассмотренных *ML* методов не может использоваться при наличии отсутствующих данных: обучение модели и последующее прогнозирование невозможно. Очевидным решением является исключение корпораций с отсутствующими данными из выборки. Альтернативный вариант — удаление переменной с отсутствующими данными. Однако недостатки таких радикальных решений заметно превышают их главное преимущество — простоту.

Наличие отсутствующих данных для заданного признака может являться объединяющей характеристикой для некоторой страты компаний. Скажем, для компаний одной отрасли. Следовательно, удаление компаний с пропущенными данными делает выборку нерепрезентативной для фондового рынка в целом. Полученные таким образом результаты и выводы распространялись бы только на специфическое подмножество компаний — предприятий с низким числом отсутствующих данных. Очевидно, что ценность результатов такого исследования была бы значительно ниже. Помимо этого, уменьшение числа наблюдений ведет к заметному снижению качества получаемых моделей.

В работе 2001 г. М. Банко и Э. Брилл на примере проблемы обработки естественного языка показали, что именно объем данных для обучения, а не выбор метода является главной детерминантной эффективности машинного обучения (Banko & Brill, 2001). Работая с небольшим набором данных, есть все основания полагать, что исключение компаний из выборки значительно снизит прогностическую силу обучаемых моделей. Принимая во внимание перечисленные факторы, удаление компании из выборки будет применяться только в самых патологических случаях — при отсутствии данных для девяноста и более статей отчетности.

Избежать удаления большего числа компаний позволяет замена отсутствующих данных вмененными значениями. Наиболее часто используемыми способами является вменение мерами центральной тенденции: средним значением или медианой. Такой подход также видится недостаточным для проблемы оценки стоимости компании, где даже небольшое увеличение прогностической силы модели дается с большим трудом.

Воспользуемся более сложным способом вменения — алгоритмом итеративного вменения (*Iterative Imputation*). Суть метода заключается в обучении множества вспомогательных моделей, где каждый признак прогнозируется с помощью остальных. Так, для каждого признака и для каждой компании пропущенные данные заменяются прогнозными значениями (Buck, 1960). Важным компонентом является определение метода, с помощью которого происходит обучение вспомогательных моделей. Не имея оснований предпочитать конкретный метод, выбор в пользу градиентного бустинга на основе деревьев решений видится наиболее здравым. На сегодняшний день именно *GBDT* считается самым передовым методом и выбором по умолчанию для работы с разнородными данными (Munkhdalai, Munkhdalai, Namsrai, Lee, & Ryu, 2019). Итеративное вменение применялось перед обучением финальной модели прогнозирования капитализации в рамках каждого из двенадцати рассмотренных методов за исключением самого *GBDT*. Дело в том,

[7] Alpha Vantage — API сервис финансовых данных. URL: https://www.alphavantage.co// (дата обращения: 28.11.2021).

что имплементация данного алгоритма библиотекой *CatBoost* имеет нативную поддержку отсутствующих данных. Таким образом, вменение для данного метода не является обязательным. Кроме этого, *GBDT* также не требует и масштабирования/стандартизации признаков.

3.1.2. Масштабирование признаков

Масштабирование или стандартизация признаков — необходимый этап предварительной обработки данных. Порядок величин для статей финансовой отчетности довольно высок и сильно варьируется. Типичные значения многих статей могут измеряться миллиардами. За редкими исключениями, эффективность *ML* методов заметно снижается, если размерность признаков сильно варьируется. Многие оптимизационные алгоритмы и функции затрат требуют нормализации независимых переменных. Существует довольно большое количество опций стандартизации данных, лучшие из которых заранее определить очень сложно. Всего было рассмотрено семь вариантов трансформации, представленных в библиотеке *scikit-learn*. Эксперименты с тренировочной выборкой показали, что самым стабильным и предпочтительным методом преобразования данных является квантильная трансформация (*Quantile Transformation*). Этот подход к стандартизации данных использует непараметрическое преобразование признака к равномерному распределению со значениями от нуля до единицы. Так, например, после преобразования максимальное значение для переменной *выручка* будет равно единице, а минимальное — нулю. Выручка для медианной компании примет значение 0,5. Для альтернативных методов стандартизации (*Standard Scaler*, *Max Absolute Scaler* и др.) р-квадрат R^2_{is} для итоговых моделей, рассчитанный на основе тренировочных данных, зачастую находился в отрицательной зоне и заметно уступал в качестве квантильной трансформации.

3.2. Результаты и обсуждение

Эффективность каждого из рассмотренных методов сопоставлена в *табл. 2* и на *рис. 2*. Как и предполагалось, из-за неспособности учитывать нелинейные взаимоотношения традиционные методы оказались сравнительно неэффективными (R^2_{oos} = 20,8% для *МНК*). По этой же причине применение различных способов регуляризации к линейной модели не привело к значительному улучшению результатов. Более того, эффективность *Lasso* и *МНК* оказалась практически идентичной. Только с помощью *Ridge* регрессии удалось незначительно улучшить результат — до 22,5%. Изменение функции потерь *MSE* на *Huber* только ухудшило качество прогнозов (17,0%).

Функция потерь Хьюбера (Huber, 1992) применяется при большом числе выбросов, что актуально для финансовой отчетности в необработанном формате. После стандартизации переменных набор данных перестал включать в себя экстремальные наблюдения. Следовательно, применение данной функции потерь не является необходимым. Использование стохастического градиентного спуска *SGD* с различными функциями потерь не привело к желаемому результату (17,6%). Метод опорных векторов *SVM,* используемый по большей части для задач классификации, предсказуемо оказался неэффективным для проблемы регрессии. Р-квадрат составил только 21,7%. Значительного улучшения качества моделей удалось достичь с использованием непараметрических методов: *k-ближайших соседей* и *дерева решений*.

Доля объясненной дисперсии для *KNN* составила 44,3%, а для дерева решений — 41,4%. Применение ансамблей деревьев решений позволило поднять точность прогнозов на качество иной уровень. Для классического алгоритма *Случайного леса* (*Random Forest)* коэффициент детерминации составил уже 73,2%. Для еще более рандомизированной его модификации, метода *сверхслучайных деревьев*, р-квадрат оказался меньше: 64,9%. Наконец, метод градиентного бустинга с использованием деревьев решений (*GBDT*) позволил получить сверхточные предсказания капитализации компаний из тестовой выборки ($R^2_{oos} = 86,7\%$).

Благодаря тому, что последняя модель способна давать наиболее качественную оценку стоимости компании, она представляет собой особый интерес с точки зрения анализа важности признаков. Столбчатая диаграмма (*рис. 3*) показывает наиболее важные статьи отчетности, при изменении значений которых наиболее сильно изменяется прогнозное значение стоимости компании (метод *Prediction Values Change*). Информация представлена в разрезе каждого финансового отчета: *о финансовых результатах*, *о движении денежных средств* и *бухгалтерского баланса*. По горизонтальной оси измеряется относительная важность признака.

Отметим, что наименьшую ценность для модели представляют данные *бухгалтерского баланса*. Интересно, что балансовая стоимость нематериальных активов с большим отрывом является наиболее важной переменной этой формы отчетности. Это

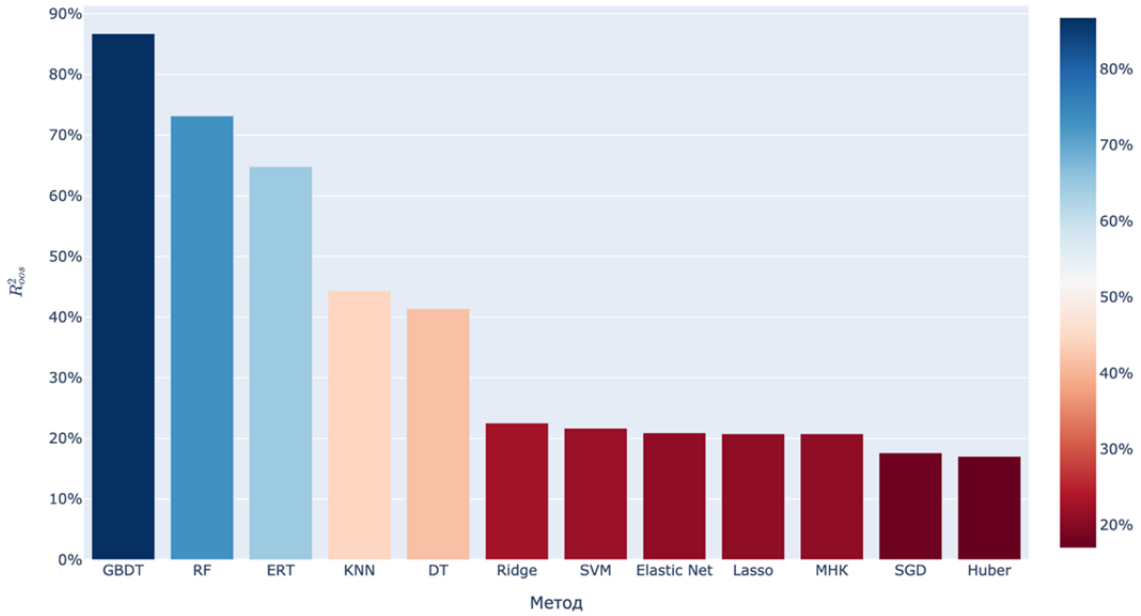

Рис. 2 / Fig. 2. **Эффективность различных методов машинного обучения на основе показателя р-квадрат, рассчитанного для тестового множества / Machine learning methods performance comparison based on the out of sample R-squared**

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

Таблица 2 / Table 2

Эффективность различных методов машинного обучения на основе показателя р-квадрат, рассчитанного для тестового множества / Machine learning methods performance comparison based on the out of sample R-squared

Метод	MHK	Ridge	Lasso	Elastic Net	Huber	SGD	SVM	KNN	DT	RF	ERT	GBDT
R^2_{oos}, %	20,8	22,5	20,8	20,9	17,0	17,6	21,7	44,3	41,4	73,2	64,9	86,7

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

противоречит мнению некоторых исследователей, которые считают балансовую стоимость нематериальных активов, в частности статью *гудвилл* наименее информативной при построении *DCF* моделей. В рамках отчета о движении денежных средств наиболее информативными признаками являются изменения в операционных активах за разные кварталы.

Из *рис. 3* также можно заметить, что совокупная значимость *отчета о финансовых результатах* значительно превосходит суммарную важность двух остальных отчетов. Несмотря на подверженность манипуляциям, зависимости от учетной политики, именно данные нижней части отчета оказываются наиболее важными при оценке стоимости компании. В соответствии с целями и задачами финансового учета *совокупный доход, чистая прибыль, EBITDA* действительно могут считаться ключевыми показателями деятельности предприятия, необходимыми при принятии инвестиционных решений. *Выручка от реализации продукции*, напротив, в число важных признаков не входит. Данные наблюдения также могут косвенно указывать на преимущества использования фондовых мультипликаторов, рассчитанных на основе прибыли[8].

Оценим влияние признаков способом перестановок *Permutation Importance*. Напомним, что важными признаками являются те, случайная перестановка которых привела к значительному снижению качества модели. Переменные, перестановка которых привела к наибольшему среднему падению р-квадрат, показаны на *рис. 4. Среднее снижение* качества

[8] Вопреки распространенному мнению о предпочтительности мультипликаторов, рассчитанных на основе выручки.

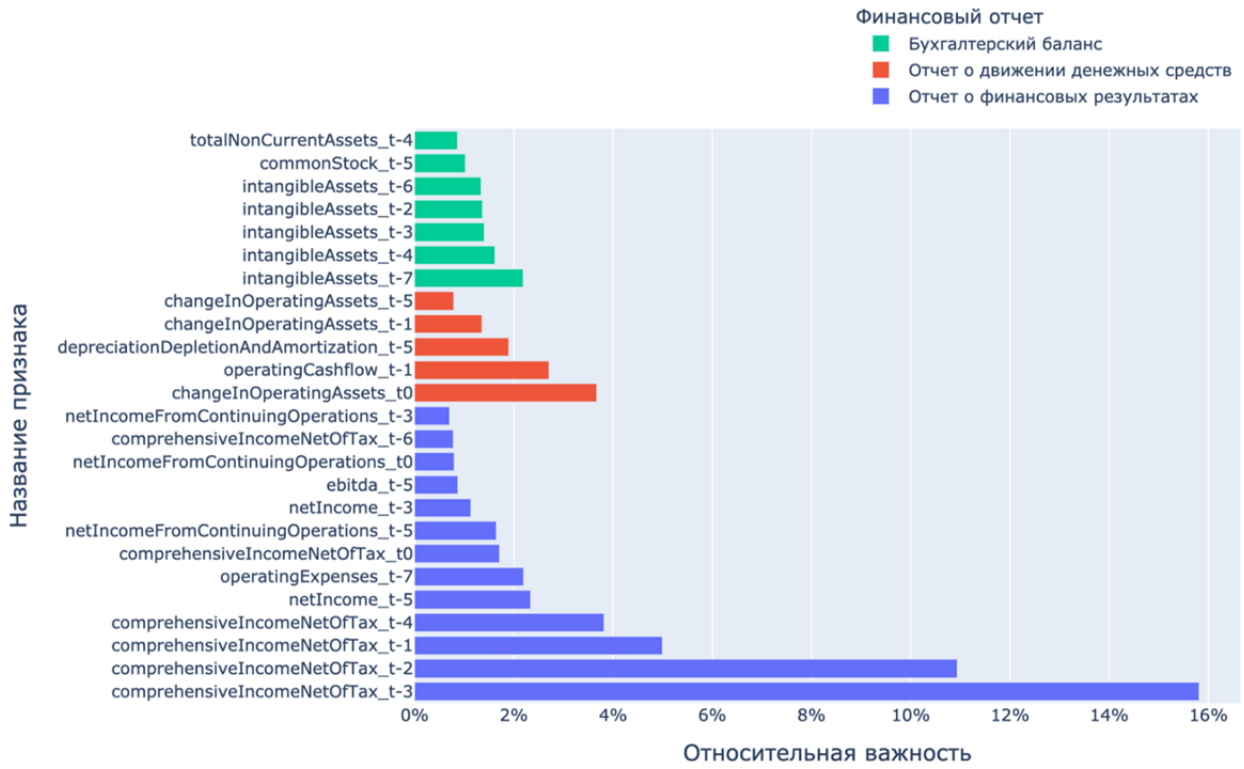

Рис. 3 / Fig. 3. **Относительная важность статей отчетности, оцененная методом Prediction values change / Relative feature importance. Prediction values change method**

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

модели рассчитано на основе трехсот случайных перестановок для каждого признака. Доверительные интервалы построены на основе двух стандартных отклонений.

Основные выводы, полученные разными способами оценки важности признаков, во многом пересекаются, увеличивая их достоверность. Очевидна низкая значимость данных *бухгалтерского баланса*: кумулятивная важность всех семи статей отчета не превышает значение пятой по значимости статьи *отчета о финансовых результатах*. Данные *совокупного дохода* (*Comprehensive Income*) за разные кварталы остаются ключевыми переменными в рамках каждого из методов оценки важности признаков. Для *отчета о движении денежных средств* главными полями являются *изменения в операционных активах* и *операционный поток денежных средств*.

Наконец, рассчитаем важность переменных методом *SHAP*. *SHAP* значения каждого отдельного наблюдения для наиболее важных признаков показаны на (*рис. 5*), где указаны влияния локальных значений признаков на прогноз. По вертикальной оси перечислены признаки. Красным цветом выделены наблюдения с высоким значением заданного признака, а синим — с низким. Позиция точек по горизонтальной оси показывает величину аддитивного вклада значения переменной в итоговый прогноз по сравнению с базовым уровнем.

Величина и знак влияния для большинства статей закономерен: высокое положительное значение статьи, как правило, увеличивает прогноз стоимости. Статья *совокупный доход* в очередной раз оказалась самой информативной при формировании прогноза. Отметим, что среднее абсолютное *SHAP* значение оказалось наибольшим для данных *совокупного дохода* именно за последний период (ближайший к настоящему дню). Вклад низких значений для большинства переменных, напротив, ассоциируется с меньшей капитализацией. Исключение составляют *выручка*, *краткосрочный долг* и *всего обязательств*.

*Таким образом, модель, а значит, и инвесторы поощряют компании, имеющие низкую долю заемного капитала. Это может указывать на то, что на практике для большинства компаний приведенная стоимость налогового щита ниже, чем потеря стоимости от риска финансового дистресса (*financial distress*), являющегося следствием использования заемных средств.*

Негативное влияние выручки объяснить сложнее. Это может являться артефактом выбранного набора данных. В топ двадцать важных факторов попали только данные выручки для седьмого квартала.

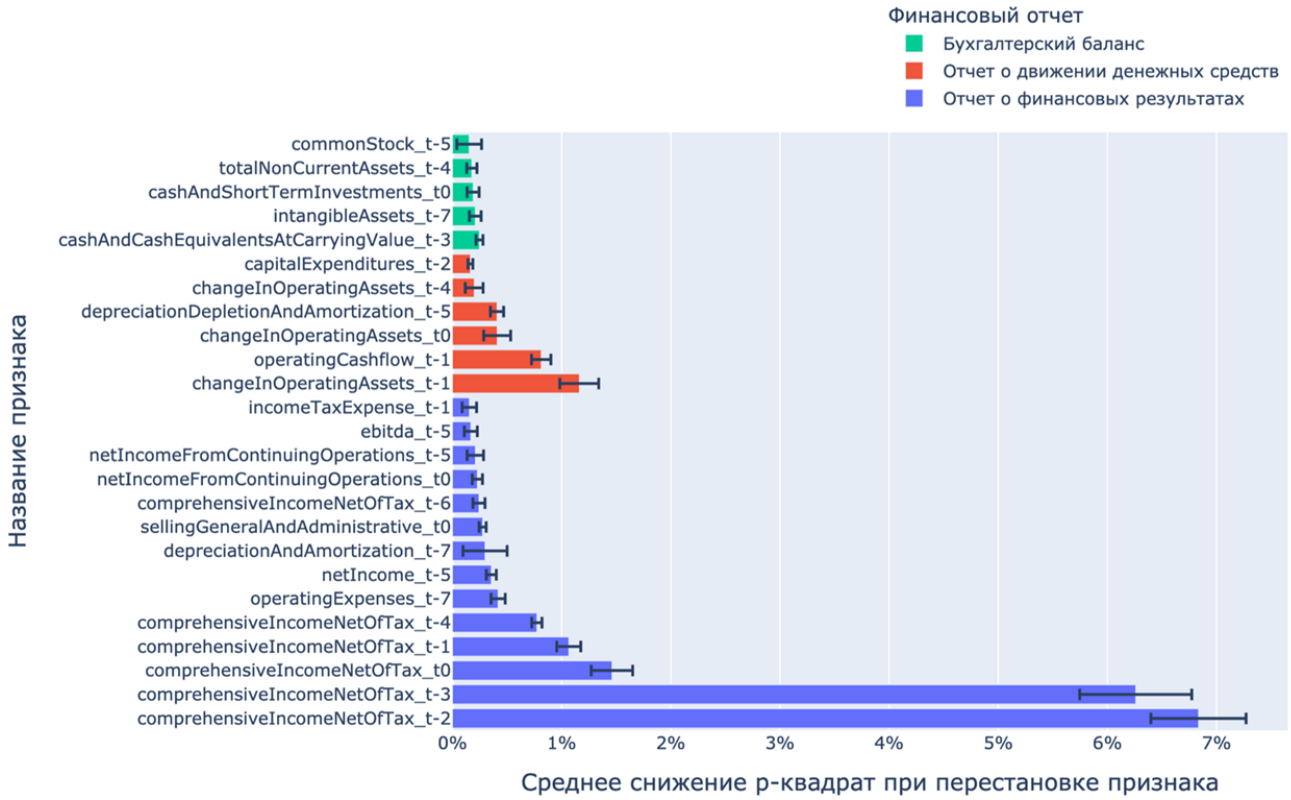

Рис. 4 / Fig. 4. **Важность признаков на основе метода перестановок / Feature importance. Permutation importance method**

Источник / Source: составлено автором / compiled by the author

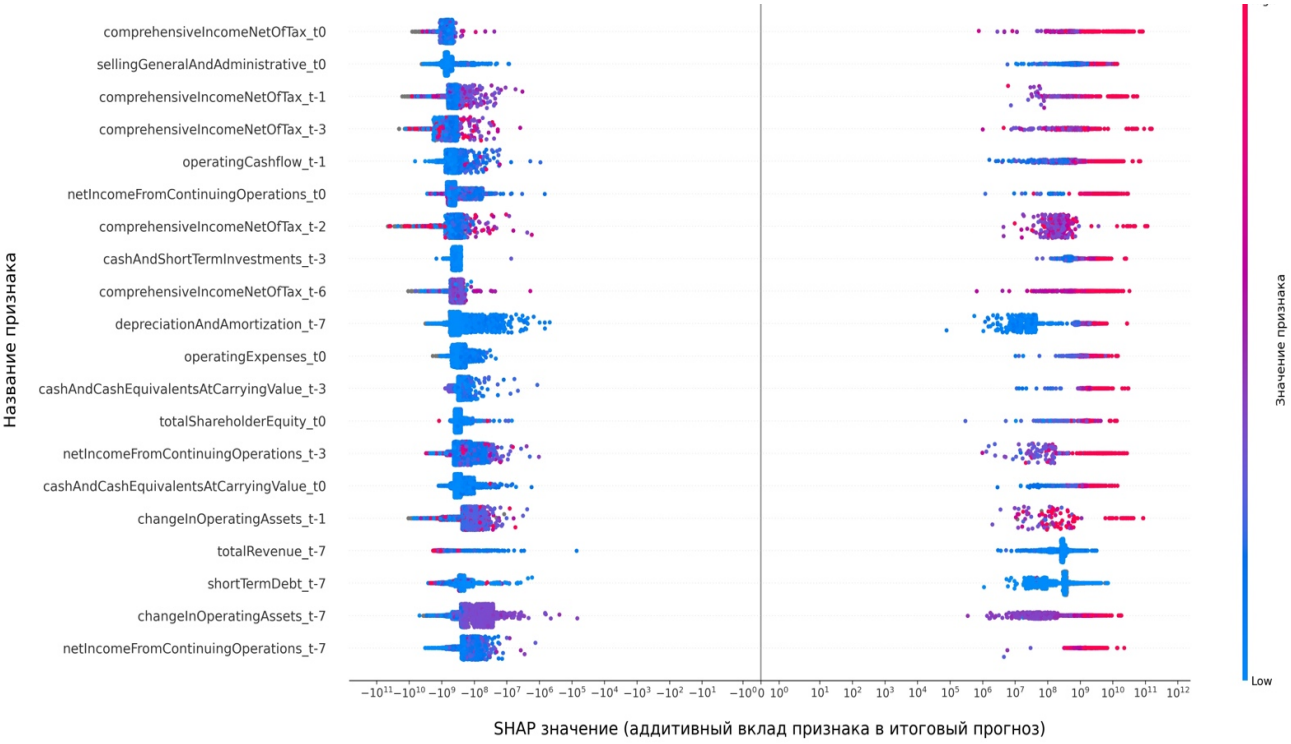

Рис. 5 / Fig. 5. **SHAP значения для наиболее важных признаков / SHAP values for the most important features**

Источник / Source: составлено автором / compiled by the author

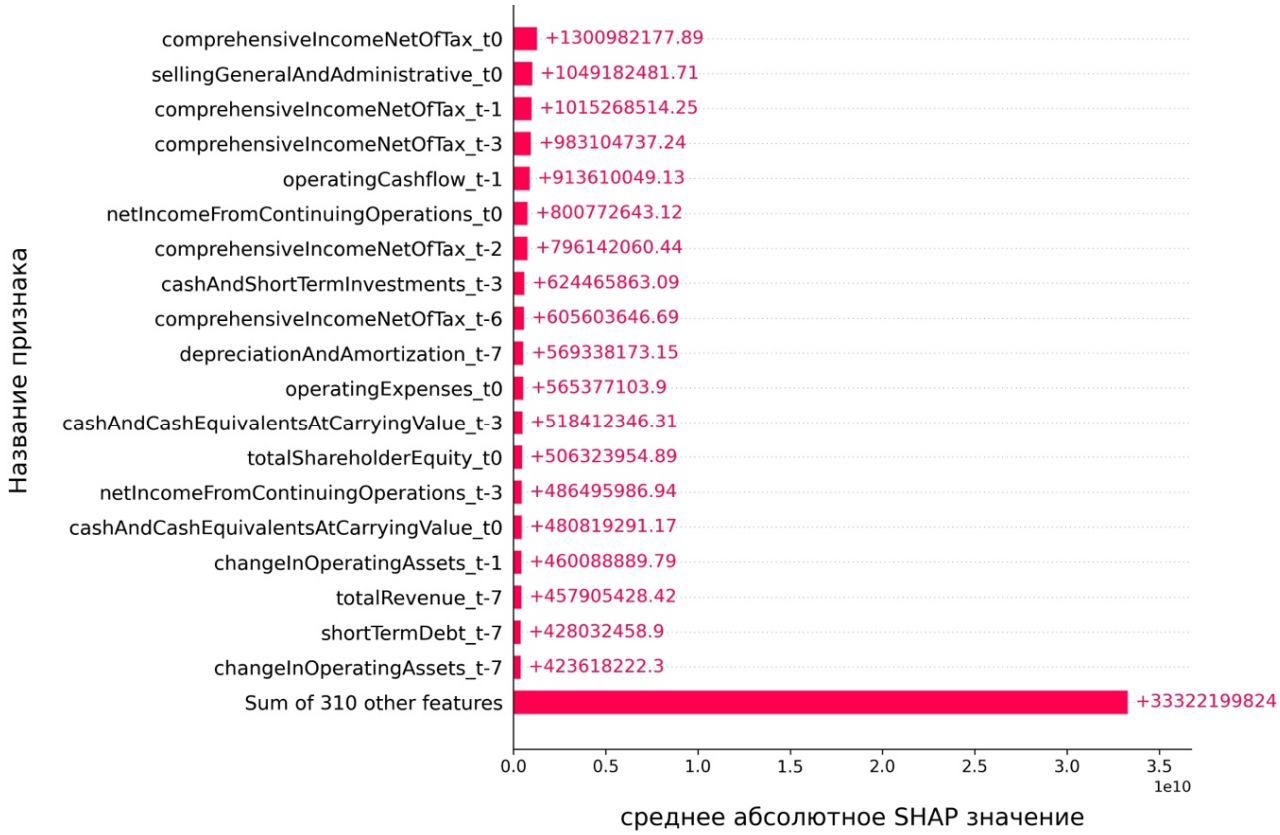

Рис. 6 / Fig. 6. **Среднее абсолютное SHAP значение для наиболее важных признаков / Mean absolute SHAP values for the most important features**

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

Показатели выручки за другие кварталы оказались сравнительно не важными. Вклад переменной, характеризующей отрасль компании, также оказался незначительным.

Признаки, отсортированные по среднему абсолютному *SHAP* значению, сведены в столбчатой диаграмме (*рис. 6*). Видно, что абсолютная важность некоторых признаков довольно велика, однако их относительный вклад в итоговый прогноз, как правило, невелик. Суммарный средний вклад девятнадцати наиболее важных признаков несравнимо ниже вклада остальных 310 признаков (последняя строка *рис. 6*). Таким образом, для формирования высокоточного прогноза используются не отдельные статьи, а данные всей финансовой отчетности за каждый из восьми кварталов.

3.3. Идеи дальнейшего рассмотрения проблемы

Усовершенствование моделей, полученных с помощью методов машинного обучения, может проводиться сразу по нескольким направлениям.

1. Инжиниринг новых признаков.
2. Поиск новых признаков.
3. Создание динамической версии модели.

В теории методы машинного обучения способны самостоятельно, без расчета вспомогательных признаков извлечь необходимую информацию из данных финансовой отчетности. Однако это применимо только для больших наборов данных (миллионы наблюдений). Работая в режиме нескольких тысяч наблюдений, расчет показателей ликвидности, оборачиваемости и рентабельности активов может позволить извлечь больше информации из имеющегося набора данных.

Помимо финансовых отчетов, ценными для оценки могут быть и другие данные, в том числе неструктурированные. Извлечение текстовой информации из годовых отчетов (Sehrawat, 2019), использование заменяющих переменных, описывающих корпоративное управление, дивидендную политику предприятия, также могут улучшить качество прогнозов (Ковалев & Драчевский, 2020). Включение данных финансовой отчетности для компаний как развитых, так и развивающихся рынков позволит увеличить выборку и получить более обобщенную модель (способную делать прогнозы для компаний разных размеров, секторов и регионов). В результате разработанные модели

могут быть применены для сложной проблемы оценки компаний российского фондового рынка (Abramishvili, Lvova, & Voronova, 2019).

Флуктуации фондового рынка указывают на сильное влияние макроэкономических факторов на рыночную капитализацию компаний. Создание динамической версии модели позволит учесть макроэкономические переменные и заметно увеличить выборку. Идея состоит в том, что компания в два разных момента времени представляет собой два разных наблюдения. К примеру, набор данных может включать в себя десять наблюдений по компании *Apple Inc.* в разные моменты времени. Каждому из десяти наблюдений будет соответствовать свой вектор признаков, состоящий из данных финансовой отчетности компании за восемь последних кварталов[9]. Кроме того, такой подход позволит расширить вектор признаков макроэкономическими данными, соответствующих моменту наблюдения. В результате применения данной процедуры для каждой компании размер выборки увеличится примерно в десять раз. Вкупе с учетом макроэкономических факторов, увеличение выборки позволит заметно увеличить точность прогнозов капитализации.

ВЫВОДЫ

Перечислим основные результаты и итоги исследования.

1. Благодаря способности учитывать нелинейные взаимоотношения, методы машинного обучения заметно превосходят традиционные эконометрические подходы и способны давать точные оценки стоимости компаний за пределами обучающей выборки.

2. Экономические выгоды от использования методов машинного обучения огромны. *ML* позволяет осмыслить сложный набор данных финансовой отчетности и экономить множество трудовых затрат высококвалифицированного персонала. Вместо траты десятков часов на создание сложных, многостраничных документов *MS Excel* с расчетом стоимости отдельно взятой компании методом *DCF* аналитик может получить точную и непредвзятую оценку сотен и даже тысяч компаний за несколько секунд.

3. Полученные модели могут быть использованы для решения прикладных задач финансового менеджмента, корпоративных финансов, бухгалтерского учета и в инвестиционном анализе при создании торговых стратегий. Положительная разница между прогнозной капитализацией и фактической может являться критерием для включения акций компании в портфель.

4. Различные способы оценки важности признаков единогласно указывают на особую ценность данных *отчета о финансовых результатах* и, в частности, статьи *совокупный доход*. Дальнейшие исследования, посвященные анализу важности признаков, позволят лучше понять процесс создания стоимости компании.

5. Совершенствование разработанных моделей может осуществляться по следующим направлениям: использование большей выборки, создание, добавление и инжиниринг новых признаков. Предложенная динамическая версия модели позволит поднять и без того высокую точность оценки стоимости компании на качественно иной уровень, предположительно оставив экспертный уровень далеко позади.

Результаты работы подтверждают рассматриваемую гипотезу. Высокое значение показателя R^2 для компаний из тестовой выборки недвусмысленно указывает на большой потенциал использования методов машинного обучения для оценки стоимости предприятия путем прогнозирования его рыночной капитализации на основании данных финансовой отчетности.

[9] Десять наблюдений компании Apple inc. не должны иметь общих ежеквартальных отчетов. Так, каждому наблюдению соответствует непересекающийся с другими наблюдениями двухлетний интервал (при использовании данных за восемь кварталов).

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Kryzanowski L., Galler M., Wright D. W. Using artificial neural networks to pick stocks. *Financial Analysts Journal*. 1993;49(4):21–27. DOI: 10.2469/faj.v49.n4.21
2. Cao L. AI in finance: A review. *SSRN Electronic Journal*. 2020. DOI: 10.2139/ssrn.3647625
3. Damodaran A. Investment valuation: Tools and techniques for determining the value of any asset. Hoboken, NJ: John Wiley & Sons, Inc.; 2012. 992 p.
4. Cybenko G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*. 1989;2(4):303–314. DOI: 10.1007/BF02551274
5. Gu S., Kelly B., Xiu D. Empirical asset pricing via machine learning. *The Review of Financial Studies*. 2020;33(5):2223–2273. DOI: 10.1093/rfs/hhaa009
6. Тихонов А.Н. О решении некорректно поставленных задач и методе регуляризации. *Доклады Академии наук*. 1963;151(3):501–504. URL: http://www.mathnet.ru/links/76d17d1b225aa6609693b033d8ad3c25/dan28329.pdf

Tikhonov A.N. On the solution of ill-posed problems and the regularization method. *Doklady Akademii nauk*. 1963;151(3):501–504. URL: http://www.mathnet.ru/links/76d17d1b225aa6609693b033d8ad3c25/dan28329.pdf (In Russ.).
7. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1996;58(1):267–288. DOI: 10.1111/J.2517–6161.1996.tb02080.x
8. Hindman M. Building better models: Prediction, replication, and machine learning in the social sciences. *The Annals of the American Academy of Political and Social Science*. 2015;659(1):48–62. DOI: 10.1177/0002716215570279
9. Carvalho D.V., Pereira E.M., Cardoso J.S. Machine learning interpretability: A survey on methods and metrics. *Electronics*. 2019;8(8):832. DOI: 10.3390/electronics8080832
10. Damodaran A. Valuation approaches and metrics: A survey of the theory and evidence. Hanover, MA: Now Publishers Inc.; 2007. 104 p.
11. Pinto J.E., Robinson T.R., Stowe J.D. Equity valuation: A survey of professional practice. *Review of Financial Economics*. 2019;37(2):219–233. DOI: 10.1002/rfe.1040
12. Böhm-Bawerk E. Recent literature on interest (1884–1899): A supplement to "Capital and interest". New York: The MacMillan Co.; 1903. 151 p.
13. Atsalakis G.S., Valavanisb K.P. Surveying stock market forecasting techniques — Part II: Soft computing methods. *Expert Systems with Applications*. 2009;36(3):5932–5941. DOI: 10.1016/j.eswa.2008.07.006
14. Xing F.Z., Cambria E., Welsch R.E. Natural language based financial forecasting: A survey. *Artificial Intelligence Review*. 2018;50(1):49–73. DOI: 10.1007/s10462–017–9588–9
15. Park B., Bae J.K. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*. 2015;42(6):2928–2934. DOI: 10.1016/j.eswa.2014.11.040
16. Whitbeck V.S., Kisor M., Jr. A new tool in investment decision-making. *Financial Analysts Journal*. 1963;19(3):55–62. DOI: 10.2469/faj.v19.n3.55
17. Коклев П.С. Влияние участия государства в акционерном капитале на стоимость компании. *Тенденции развития науки и образования*. 2020;(60–8):14–18. DOI: 10.18411/lj-04–2020–154
Koklev P.S. Impact of the state ownership in equity on company value. *Tendentsii razvitiya nauki i obrazovaniya*. 2020;(60–8):14–18. (In Russ.). DOI: 10.18411/lj-04–2020–154
18. Joshi H., Chauha R. Determinants and prediction accuracy of price multiples for South East Asia: Conventional and machine learning analysis. *Indonesian Capital Market Review*. 2020;12(1):42–54. DOI: 10.21002/icmr.v12i1.12051
19. Liu J., Nissim D., Thomas J. Equity valuation using multiples. *Journal of Accounting Research*. 2002;40(1):135–172. DOI: 10.1111/1475–679X.00042
20. Hastie T., Tibshirani R., Friedman J. The elements of statistical learning: Data mining, inference, and prediction. 2nd ed. New York: Springer-Verlag; 2016. 767 p. (Springer Series in Statistics). DOI: 10.1007/978–0–387–84858–7
21. Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. CatBoost: Unbiased boosting with categorical features. In: Proc. 32nd Int. conf. on neural information processing systems (NIPS'18). (Montréal, December 3–8, 2018). New York: Curran Associates Inc.; 2018:6639–6649. URL: https://proceedings.neurips.cc/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf
22. Ioannidis J., Doucouliagos C. What's to know about the credibility of empirical economics? *Journal of Economic Surveys*. 2013;27(5):997–1004. DOI: 10.1111/joes.12032
23. Bergstra J., Bengio Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*. 2012;13(2):281–305. URL: https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf
24. Breiman L. Random forests. *Machine Learning*. 2001;45(1):5–32. DOI: 10.1023/A:1010933404324
25. Shapley L.S. A value for n-person games. In: Kuhn H.W., Tucker A.W., eds. Contributions to the theory of games. Vol. II. Princeton, NJ: Princeton University Press; 2016:307–318. DOI: 10.1515/9781400881970–018
26. Banko M., Brill E. Scaling to very very large corpora for natural language disambiguation. In: Proc. 39th Annu. meet. of the Association for Computational Linguistics (ACL'01). (Toulouse, July 06–11, 2001). Stroudsburg, PA: Association for Computational Linguistics; 2001:26–33. DOI: 10.3115/1073012.1073017

27. Buck S.F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1960;22(2):302–306. DOI: 10.1111/j.2517–6161.1960.tb00375.x
28. Munkhdalai L., Munkhdalai T, Namsrai O.-E., Lee J.Y., Ryu K.H. An empirical comparison of machine-learning methods on bank client credit assessments. *Sustainability*. 2019;11(3):699. DOI: 10.3390/su11030699
29. Huber P.J. Robust estimation of a location parameter. In: Kotz S., Johnson N.L., eds. Breakthroughs in statistics: Methodology and distribution. New York: Springer-Verlag; 1992:492–518. (Springer Series in Statistics). DOI: 10.1007/978–1–4612–4380–9_35
30. Sehrawat S. Learning word embeddings from 10-K filings for financial NLP tasks. *SSRN Electronic Journal*. 2019. DOI: 10.2139/ssrn.3480902
31. Ковалев В.В., Драчевский И.С. Дивидендная политика как фактор управления стоимостью компании: сравнение тенденций на формирующихся рынках. *Вестник Санкт-Петербургского университета. Экономика*. 2020;36(1):95–116. DOI: 10.21638/spbu05.2020.105
Kovalev V.V., Drachevsky I.S. Dividend policy as a factor for managing company value: Comparing trends in emerging markets. *Vestnik Sankt-Peterburgskogo universiteta. Ekonomika = St Petersburg University Journal of Economic Studies (SUJES)*. 2020;36(1):95–116. DOI: 10.21638/spbu05.2020.105
32. Abramishvili N.R., Lvova N.A., Voronova N.S. Is it possible to assess the corporate market value in the emerging market? In: New challenges of economic and business development — 2019: Incentives for sustainable economic growth. Proc. 11th Int. sci. conf. (Riga, May 16–18, 2019). Riga: University of Latvia; 2019:12–21. URL: https://dspace.lu.lv/dspace/handle/7/48896 (дата обращения: 18.12.2021).

ИНФОРМАЦИЯ ОБ АВТОРЕ / ABOUT THE AUTHOR

Пётр Сергеевич Коклев — аспирант кафедры теории кредита и финансового менеджмента, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия
Petr S. Koklev — postgraduate student of the Department of Credit Theory and Financial Management, Saint Petersburg State University, Saint Petersburg, Russia
https://orcid.org/0000-0003-2594-7973
koklevp@gmail.com

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.
The author read and approved the final version of the manuscript.

Статья поступила в редакцию 24.01.2022; после рецензирования 11.02.2022; принята к публикации 27.04.2022.
Автор прочитал и одобрил окончательный вариант рукописи.
Conflicts of Interest Statement: The author has no conflicts of interest to declare.
The article was submitted on 24.01.2022; revised on 11.02.2022 and accepted for publication on 27.04.2022.