УДК 336.76

Е. В. Королева ORCID ID 0000-0003-2640-5906

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия, e-mail: koroleva_ev@spbstu.ru

А. С. Полубатонова

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия

А. Р. Сереженкова

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия

ДЕТЕРМИНАНТЫ ДОХОДНОСТИ БИРЖЕВЫХ ИНВЕСТИЦИОННЫХ ФОНДОВ: МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

Ключевые слова: доходность ETF, модели машинного обучения, ключевые признаки, прогнозирование.

Статья посвящена выявлению детерминант доходности биржевых инвестиционных фондов на основе моделей машинного обучения. Цель исследования – определить ключевые факторы доходности ETF и сопоставить эффективность современных нелинейных и ансамблевых алгоритмов. Эмпирическая база включает данные за 2024 год по 34 странам (105 наблюдений). В работе построены и сравнены семь моделей: SVR, регрессия на решающих деревьях, Random Forest, Gradient Boosting, XGBoost, Bagging и CatBoost. Для интерпретации результатов и оценки вклада признаков применялся SHAP-анализ. Наилучшее качество продемонстрировали модели Random Forest и регрессия на решающих деревьях, объяснившие около 65% вариации целевой переменной. SHAP-анализ показал, что спред bid-ask является наиболее значимым фактором и оказывает устойчивое положительное влияние на доходность. Процентная ставка характеризуется нелинейным воздействием: при экстремальных значениях выявляется отрицательный эффект, тогда как умеренные уровни связаны с ростом доходности. Коэффициент расходов демонстрирует стабильную отрицательную зависимость с доходностью. Гипотезы о значимом влиянии оборачиваемости портфеля и типа управления (активное/пассивное) эмпирического подтверждения не получили. Полученные результаты подтверждают целесообразность применения моделей машинного обучения, прежде всего ансамблевых моделей, для анализа и прогнозирования доходности ETF, обеспечивающих сочетание точности и интерпретируемости, и могут быть использованы при разработке инвестиционных стратегий с акцентом на спред bid-ask и контроль издержек.

E. V. Koroleva ORCID ID 0000-0003-2640-5906

Peter the Great St. Petersburg Polytechnic University, Saint Petersburg, Russia, e-mail: koroleva_ev@spbstu.ru

A. S. Polybatonova

Peter the Great St. Petersburg Polytechnic University, Saint Petersburg, Russia

A. R. Serezhenkova

Peter the Great St. Petersburg Polytechnic University, Saint Petersburg, Russia

DETERMINANTS OF EXCHANGE-TRADED FUNDS' RETURNS: MACHINE LEARNING MODELS

Keywords: ETF returns, machine learning models, feature importance, financial forecasting.

This paper explores the determinants of exchange-traded fund returns using machine learning models. The objective of the study is to identify key drivers of ETF returns and compare the performance of modern nonlinear and ensemble algorithms. The empirical base includes data for 2024 for 34 countries (105 observations). Seven models are constructed and compared: SVR, decision tree regression, Random Forest, Gradient Boosting, XGBoost, Bagging, and CatBoost. SHAP analysis was used to interpret the results and assess the contribution of features. The Random Forest and decision tree regression models demonstrated the best performance, explaining approximately 65% of the target variable variation. SHAP analysis revealed that spread bid-ask is the most significant factor and has a consistent positive impact on returns. The interest rate is characterized by a nonlinear impact: at extreme values, a negative effect is observed, while moderate levels are associated with increased returns. The expense ratio demonstrates a consistent negative relationship with returns. The hypotheses regarding the significant impact of portfolio turnover and management type (active/passive) were not empirically supported. The obtained results confirm the feasibility of using machine learning models, particularly ensemble methods, to analyze and forecast ETF returns, providing a combination of accuracy and interpretability, and can be used in developing investment strategies with a focus on spread bid-ask and cost control.

Введение

Развитие рынка биржевых инвестиционных фондов (ETF) является одним из ключевых трендов современной финансовой системы [1], предоставляя инвесторам доступ к диверсифицированным портфелям с относительно низкими издержками. В условиях растущей волатильности и усложнения рыночных взаимосвязей повышение точности прогнозирования доходности ETF становится актуальной задачей для управляющих компаний, аналитиков и частных инвесторов. Традиционные линейные модели, хотя и широко применяются для выявления основных зависимостей, зачастую оказываются недостаточно гибкими для учета сложных нелинейных взаимодействий между факторами, определяющими доходность [2].

В научной литературе представлен широкий спектр факторов, потенциально влияющих на эффективность ETF, однако результаты эмпирических исследований часто носят противоречивый характер, а влияние многих факторов остается недостаточно изученным в контексте современных нелинейных моделей машинного обучения. Это создает потребность в комплексном анализе, позволяющем не только оценить значимость отдельных предикторов, но и выявить скрытые паттерны в данных.

Целью данного исследования является моделирование факторов, определяющих доходность инвестиционных фондов, с применением набора современных нелинейных и ансамблевых моделей машинного обучения.

Исследование базируется на данных, полученных из предварительно построенной модели линейной регрессии, где были отобраны статистически значимые предикторы. Результаты работы позволят не только идентифицировать ключевые драйверы доходности ETF, но и оценить сравнительную эффективность различных моделей машинного обучения в задачах финансового прогнозирования.

Материалы и методы исследования

Для анализа были выбраны ключевые модели нелинейного моделирования: SVR, Decision Tree Regression и ансамблевые модели – Random Forest, Gradient Boosting, XGBoost, Bagging и CatBoost. Моделирование проводилось с помощью языка программирования Python.

Для выявления статистически значимых факторов, определяющих доходность, в качестве объекта исследования были выбраны биржевые инвестиционные фонды (ETF). Эмпирическая база исследования (за 2024 год) сформирована на основе данных финансовых ресурсов: Investing.com [3], StockAnalysis.com [4], Finance.yahoo.com [5], Investfunds.ru [6], TradingEconomics.com [7]. В выборку были включены ETF, отслеживающие динамику соответствующих инвестиционных индексов. Такой подход обеспечил широкое географическое и рыночное представительство. В итоговую выборку вошли фонды из 34 стран, в основном с развитой финансовой системой, что делает выборку репрезентативной для анализа глобальных тенденций и сравнительной оценки фондов из разных регионов. Общее количество наблюдений составило 105.

Доходность ETF формируется под воздействием комплекса взаимосвязанных факторов, среди которых исследования выделяют несколько ключевых детерминант. Во-первых, существенную роль играют характеристики самого фонда: высокая оборачиваемость портфеля приводит к росту транзакционных издержек [8], а повышенный коэффициент расходов (expense ratio) напрямую уменьшает итоговую доходность инвестора [10]. Во-вторых, существенную роль играет ликвидность, одним из ключевых прокси-показателей которой выступает bid–ask спред: расширение спреда отражает более высокие транзакционные издержки и информационную асимметрию, вследствие чего инвесторы, как правило, требуют дополнительную премию за владение низколиквидными активами [8, 9].

Таблица 1

Перечень исследуемых факторов, гипотез и метрик

Гипотеза	Ожидаемое влияние	Обозначение переменной	Характеристика переменной	Ед. измерения
-		Y_Return	Доходность ETF (разница между ценой акции фонда на конец года и ценой акции фонда на начало года, деленная на цену акции на начала года)	проц. пункты
H1	-	Costs	Коэффициент расходов	проц. пункты
H2	-	Turnover	Коэффициент оборачиваемости (отношение среднего объема торгов к числу чистых активов в годовом исчислении)	проц. пункты
H3	-	Interest rate	Ключевая ставка государственных банков различных стран (среднегодовое значение)	проц. пункты
H4	+	Bid–Ask spread	Спред bid-ask (разница между минимальной ценой продажи (Ask) и максимальной ценой покупки (Bid))	проц. пункты
H5	Active +	Active/passive	Активное (1) или пассивное (0) управление	1/0

Источник: составлено авторами.

Это, в свою очередь, формирует классическую дилемму между уровнем ликвидности и потенциальной доходностью. В-третьих, важна стратегия управления: активное управление, несмотря на возможность опережать рынок, сопряжено с дополнительными рисками и издержками [11]. Наконец, критическим внешним детерминантом являются процентные ставки. Их повышение не только ухудшает макроэкономические условия для инвестиций [12], но и снижает ожидаемую доходность отдельных классов активов (например, облигаций), что немедленно отражается на результатах соответствующих ETF [13]. Устойчивость данного эффекта подтверждается в различных рыночных контекстах [14]. Основные гипотезы, обозначения переменных, их характеристики и ожидаемое направление влияния представлены в таблице 1.

В процессе сбора возникла необходимость унификации валюты, так как данные о размере фондов и ВВП для некоторых стран были представлены в национальной валюте. Для решения этой проблемы все значения были переведены в доллары США, используя среднегодовой обменный курс, соответствующий периоду сбора данных.

Для оценки обобщающей способности использовано фиксированное разбиение выборки на обучающую и тестовую части в пропорции 80/20 [15]. Разбиение выполнялось случайно при фиксированном значении параметра начального зерна генератора случайных чисел (seed), что обеспечивает воспроизводимость результатов. Для модели SVR применялась стандартизация признаков (StandardScaler).

Результаты исследования и их обсуждение

В ходе исследования были построены и проанализированы семь моделей машинного обучения для прогнозирования доходности ETF. В результате, был произведен их сравнительный анализ и построена сводная таблица 2.

Наивысшее качество демонстрируют модели Random Forest и Decision Tree Regression. Данные модели объясняют около 65% дисперсии целевой переменной при минимальных значениях средней квадратичной ошибки (MSE = 0,011), что свидетельствует об их довольно высокой точности и устойчивости. Умеренное качество показывают XGBoost ($R^2 = 0,56$) и Gradient Boosting ($R^2 = 0,54$). Их метрики указывают на способность моделировать более половины изменчивости данных, однако результаты ниже, чем у Random Forest. Модель SVR практически не объясняет дисперсию данных, в то время как Bagging и CatBoost демонстрируют ограниченную эффективность.

Для интерпретации вклада признаков во всех моделях применялся SHAP-анализ. Анализ выявил высокую согласованность ключевых факторов, влияющих на прогноз.

Таблица 2

Сравнительная таблица результатов моделирования

Модели	**Нелинейные**		**Ансамблевые**				
	SVR	**Decision Tree Regression**	**Random Forest**	**Gradient Boosting**	**XGBoost**	**Bagging**	**Cat-Boost**
Факторы	Направление/степень влияния						
Спред bid-ask	+	+	+	+	+	+	+
Процентная ставка	-/+	-/+	-/+	-/+	-/+	-/+	-/+
Коэффициент расходов	-	-	-			-	-
Коэффициент оборачиваемости			-				
Тип управления							
Метрики качества							
R^2	0,012	0,653	0,654	0,543	0,558	0,419	0,419
MSE	0,03	0,011	0,011	0,014	0,014	0,018	0,014

Источник: составлено авторами.

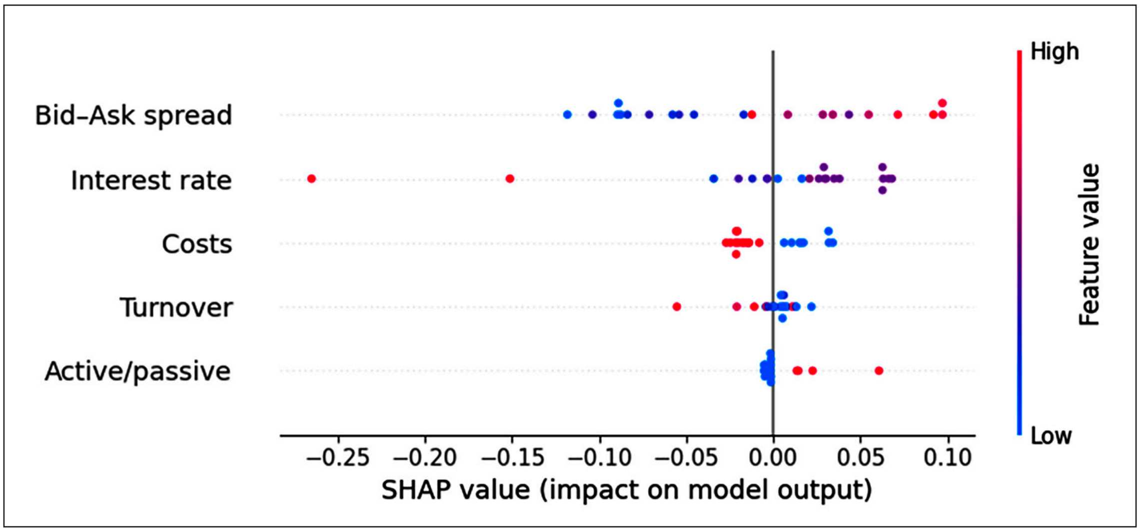

Рис. 1. Вклад факторов в построение модели Random Forest
Источник: составлено авторами

Спред bid-ask является наиболее значимым фактором во всех эффективных моделях. Более широкий спред устойчиво ассоциируется с положительным вкладом в прогнозируемую доходность, что согласуется с теоретическими ожиданиями и подтверждает гипотезу H4.

Процентная ставка оказывает сложное нелинейное влияние. Наибольший негативный эффект наблюдается при экстремально высоких и низких значениях ставки. Умеренный уровень ставки, напротив, способствует росту прогнозируемой доходности. Таким образом, гипотеза H3 о негативном влиянии подтвердилась лишь для крайних значений, в целом же характер влияния является амбивалентным (-/+).

Коэффициент расходов проявляет устойчивую отрицательную связь с доходностью ETF в большинстве моделей (Random Forest, Decision Tree, Bagging, CatBoost, SVR), что подтверждает гипотезу H1.

Коэффициент оборачиваемости и тип управления не продемонстрировали статистически значимого влияния на доходность ETF в рамках построенных нелинейных моделей. Согласно анализу SHAP, их вклад был минимальным или близком к нулю.

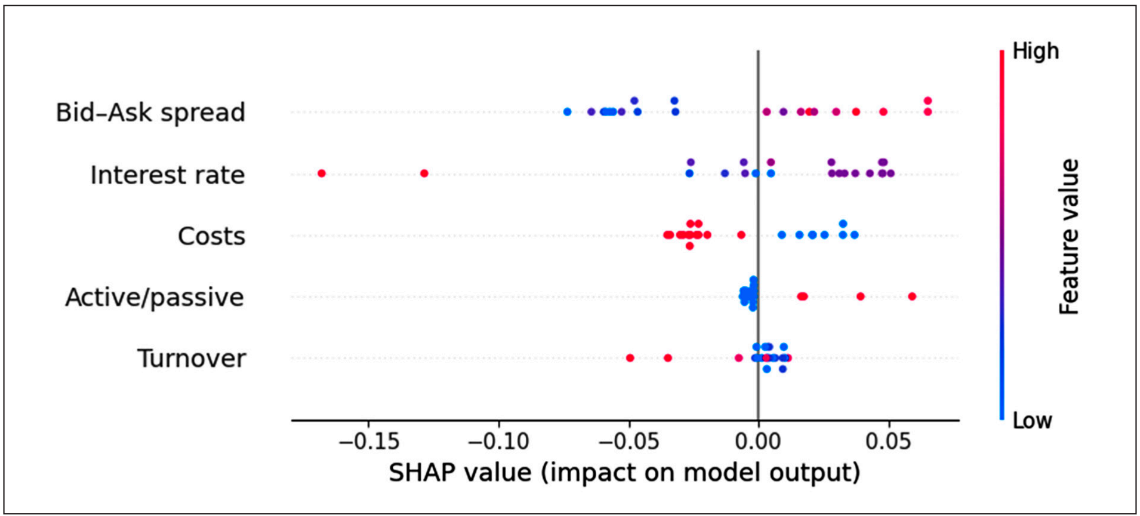

Рис. 2. Вклад факторов в построение модели Decision Tree Regression
Источник: составлено авторами

На рис. 1 в качестве примера представлена визуализация SHAP-значений для модели Random Forest, демонстрирующая вышеописанные закономерности.

Таким образом, гипотеза H1 (отрицательное влияние расходов) нашла полное подтверждение. Гипотеза H4 (положительное влияние спред bid-ask) подтвердилась в рамках исследования. Гипотеза H3 (отрицательное влияние процентной ставки) подтвердилась частично, выявив нелинейный характер зависимости. Гипотезы H2 (отрицательное влияние коэффициента оборачиваемости) и H5 (активное управление положительно влияет на доходность ETF) статистически не подтвердились в рамках данного исследования.

Для углубления анализа и проверки согласованности результатов был рассмотрен вклад факторов в модель Decision Tree, которая показала сопоставимое с Random Forest качество. Визуализация SHAP-значений для этой модели представлена на рис. 2.

Анализ графика выявляет как сходства, так и важные нюансы по сравнению с моделью Random Forest. Доминирование спреда bid-ask сохраняется. Данный фактор оказывает наиболее сильное и последовательно положительное влияние на прогноз.

Процентная ставка вновь демонстрирует ярко выраженный нелинейный эффект. Однако в модели Decision Tree паттерн проявляется еще более четко. Экстремальные значения (как высокие, так и низкие – красные точки в обеих частях распределения) имеют выраженный негативный вклад. Умеренные значения ставки (синие и голубые точки в центре) связаны с положительным влиянием на прогноз.

Коэффициент расходов показывает неоднозначные результаты. Высокие значения расходов не оказывают значимого влияния (точки сконцентрированы возле нуля), тогда как низкие значения (синие точки) в ряде случаев ассоциируются со снижением прогнозируемой доходности. Данный результат может свидетельствовать о том, что в рамках простой древовидной структуры модель уловила нелинейную зависимость, которая требует отдельного изучения.

Тип управления и коэффициент оборачиваемости, как и в других моделях, не демонстрируют значимого вклада, что подтверждается плотной концентрацией их SHAP-значений возле нуля.

Сравнительный анализ SHAP-визуализаций для двух лучших по метрикам моделей подтверждает высокую устойчивость ключевых выводов. Спред bid-ask является универсальным и наиболее значимым драйвером доходности ETF. Процентная ставка оказывает сложное нелинейное воздействие, где опасность для доходности представляют именно экстремальные значения, а не сам факт роста или падения. Статус второстепенных факторов для оборачиваемости и типа управления также не вызывает сомнений.

При этом обнаруженные вариации во влиянии коэффициента расходов (от стабильно негативного в Random Forest до неоднозначного в Decision Tree Regression) подчеркивают важность использования ансамблевых моделей. Они агрегируют множество решений, что позволяет нивелировать случайные паттерны, уловленные отдельным деревом, и дать более устойчивую и надежную оценку важности признаков.

Таким образом, Random Forest можно рекомендовать не только как модель с лучшими метриками, но и как инструмент, который обеспечивает наиболее сбалансированную и интерпретируемую картину влияния факторов.

Модель опорных векторов для регрессии (SVR) продемонстрировала крайне низкую прогнозную точность ($R^2 = 0{,}012$) при моделировании доходности ETF. Данный результат согласуется с выводами современных исследований, посвященных применению машинного обучения в финансовой аналитике, и может быть объяснен рядом причин.

Во-первых, эффективность SVR критически зависит от корректного выбора гиперпараметров, таких как параметр регуляризации C и параметры ядерной функции. Как отмечают López et al. [16], без тщательной оптимизации этих параметров модель склонна либо к переобучению, либо к недообучению, особенно при работе с зашумленными данными. В рамках данного исследования использовалась стандартная конфигурация SVR, что, вероятно, привело к ее неадекватной адаптации к специфике финансовых показателей.

Во-вторых, сравнительные исследования моделей машинного обучения для прогнозирования доходности активов указывают на превосходство ансамблевых алгоритмов, таких как Random Forest и градиентный бустинг, над SVR в условиях нелинейных зависимостей. Feng et al. [17] подчеркивают, что модели на основе деревьев способны автоматически выявлять сложные взаимодействия признаков, не требуя явной спецификации модели, что делает их более гибкими и точными при работе со структурированными финансовыми данными.

В-третьих, масштабный анализ, проведенный Gu et al. [2], показывает, что SVR часто уступает современным ансамблевым подходам в задачах эмпирического ценообразования активов, особенно при ограниченном объеме выборки. В данном исследовании объем данных составил 105 наблюдений, что является относительно малой выборкой для стабильной работы SVR, чувствительного к шуму и размеру обучающего набора.

Таким образом, низкие метрики SVR в данном исследовании не свидетельствуют об отсутствии взаимосвязей между факторами и доходностью ETF. Скорее, они отражают ограниченную применимость стандартной конфигурации SVR для моделирования сложных нелинейных зависимостей в условиях малой выборки. Полученный результат подтверждает целесообразность использования более гибких ансамблевых моделей, таких как Random Forest, для задач финансового прогнозирования.

Заключение

Настоящее исследование было направлено на моделирование факторов доходности биржевых инвестиционных фондов (ETF) с применением набора нелинейных и ансамблевых алгоритмов машинного обучения. Эмпирический анализ, проведенный на данных за 2024 год по 34 странам, позволил получить следующие основные результаты.

Наиболее высокую точность и устойчивость при прогнозировании доходности ETF в рамках проведенного исследования продемонстрировали ансамблевые методы машинного обучения – прежде всего Random Forest и Decision Tree Regression. Указанные модели обеспечили наилучшее качество аппроксимации целевой переменной, объяснив порядка 65% ее дисперсии при минимальных значениях ошибки прогнозирования. Напротив, модель SVR в базовой конфигурации показала низкую результативность, что указывает на ограниченность ее применения в подобных задачах без целенаправленной настройки гиперпараметров и подбора ядра.

Интерпретация результатов на основе анализа важности признаков с применением SHAP-методологии выявила высокую согласованность ключевых факторов, определяющих динамику доходности ETF. Спред bid-ask подтвердил статус наиболее значимой и устойчиво положительной детерминанты доходности. Процентная ставка характеризуется выраженным нелинейным эффектом: при экстремально высоких и низких значениях ее влияние на доходность носит отрицательный характер, тогда как умеренный уровень ставки ассоциирован с повышением доходности. Коэффициент расходов демонстрирует стабильную отрицательную связь с результативностью ETF, что соответствует теоретическим представлениям о снижении чистой доходности инвестора вследствие издержек фонда. При этом гипотезы о статистически значимом негативном влиянии оборачиваемости портфеля и о положительном влиянии активного управления на доходность ETF в рамках рассматриваемой выборки подтверждения не получили: вклад данных факторов в предиктивную модель оказался пренебрежимо малым.

Полученные результаты имеют практическую значимость для управляющих компаний и инвесторов. Random Forest может быть рекомендован как наиболее сбалансированный инструмент прогнозирования, сочетающий высокую точность с возможностью интерпретации и оценки вкладов факторов. При формировании инвестиционных стратегий и выборе ETF целесообразно уделять приоритетное внимание характеристикам ликвидности базовых активов и контролю уровня совокупных расходов фонда. Нелинейный характер влияния процентных ставок, выявленный в исследовании, предполагает необходимость более детального учета макроэкономической среды и режимов денежно-кредитной политики при оценке перспектив доходности ETF.

В целом результаты подтверждают потенциал современных ансамблевых моделей машинного обучения для выявления сложных нелинейных зависимостей на финансовых рынках и углубляют понимание факторов, определяющих доходность ETF. В качестве направлений дальнейших исследований представляется целесообразным расширение набора объясняющих переменных, увеличение объема выборки, а также проведение системной оптимизации гиперпараметров для более сложных алгоритмов (в частности, XGBoost и CatBoost) с последующим сравнением их прогностических и интерпретационных характеристик.

Библиографический список

1. Agarwal V. et al. Do ETFs increase the commonality in liquidity of underlying stocks? CFR Working Paper, 2021. № 21-04. URL: https://www.econstor.eu/handle/10419/232548 (дата обращения: 22.01.2026).

2. Gu S., Kelly B., Xiu D. Empirical asset pricing via machine learning // The Review of Financial Studies. 2020. № 33(5). P. 2223-2273. DOI: 10.1093/rfs/hhaa009.

3. Investing.com. котировки и финансовые новости [Электронный ресурс]. URL: https://ru.investing.com/ (дата обращения: 22.01.2026).

4. Stock Analysis. Free Online Stock Information for Investors [Электронный ресурс]. URL: https://stockanalysis.com/ (дата обращения: 22.01.2026).

5. Yahoo Finance. Stock Market Live, Quotes, Business & Finance News [Электронный ресурс]. URL: https://finance.yahoo.com/ (дата обращения: 22.01.2026).

6. InvestFunds. сайт про инвестиции и фондовые рынки [Электронный ресурс]. URL: https://investfunds.ru/ (дата обращения: 22.01.2026).

7. Tradingeconomics. Мировая платформа с лучшими графиками для трейдеров и инвесторов [Электронный ресурс]. URL: https://tradingeconomics.com/ (дата обращения: 22.01.2026).

8. Muñoz F., Vargas M., Marco I. Environmental mutual funds: Financial performance and managerial abilities // Journal of Business Ethics. 2014. Т. 124. №. 4. С. 551-569. URL: https://link.springer.com/article/10.1007/s10551-013-1893-x (дата обращения: 22.01.2026). DOI: 10.1007/s10551-013-1893-x.

9. Pástor Ľ., Stambaugh R. F. Liquidity risk and expected stock returns // Journal of Political economy. 2003. Т. 111. № 3. С. 642-685. URL: https://www.journals.uchicago.edu/doi/abs/10.1086/374184 (дата обращения: 22.01.2026). DOI: 10.1086/374184.

10. Калижников Ю. А., Примаков Е. В., Рязанцев А. А. Особенности отечественного рынка ETF // Modern Science. 2021. № 12-3. С. 70-85. URL: https://elibrary.ru/item.asp?id=47455959 (дата обращения: 22.01.2026).

11. Yuzvovich L., Frais V., Kodasheva G. Stock Exchange Mutual Funds as the Drivers of the Stock Market Development in Terms of Profitability and Alternatives // SHS Web of Conferences. EDP Sciences, 2021. Т. 93. С. 02027. URL: https://www.shs-conferences.org/articles/shsconf/abs/2021/04/shsconf_nid2020_02027/shsconf_nid2020_02027.html (дата обращения: 22.01.2026). DOI: 10.1051/shsconf/20219302027.

12. Audita A. Y., Iskandar R., Azis M. The effect of expense ratio, fund size and fund age on performance of ETF mutual funds with interest rate as a moderating variable // Fair Value: Jurnal Ilmiah Akuntansi Dan Keuangan. 2023. Т. 6. № 1. С. 55-66. URL: http://journal.ikopin.ac.id/index.php/fairvalue/article/view/3523 (дата обращения: 22.01.2026).

13. Ali U., Hirshleifer D. Opportunism as a firm and managerial trait: Predicting insider trading profits and misconduct // Journal of financial economics. 2017. Т. 126. № 3. С. 490-515. DOI: 10.1016/j.jfineco.2017.09.002.

14. Kemeç U., Kula V., Baykut E. Decoding ETF Market Movements: The Impact of Internal and External Factors // Ekonomi Politika ve Finans Araştırmaları Dergisi. 2025. Т. 10. № 3. С. 1122-1142. URL: https://dergipark.org.tr/en/pub/epfad/article/1718492 (дата обращения: 22.01.2026). DOI: 10.30784/epfad.1718492.

15. Рудзейт О. Ю., Зайнетдинов А. Р., Недяк А. В., Рагулин П. Г. Прогнозирование цены акции с помощью метода регрессионного анализа // Отходы и ресурсы. 2020. Т. 7. № 4. С. 14-14. URL: https://www.elibrary.ru/item.asp?id=44582635 (дата обращения: 22.01.2026).

16. López J., Maldonado S., Carrasco M. Double regularization methods for robust feature selection and SVM classification via DC programming // Information Sciences. 2018. Т. 429. С. 377-389. URL: https://www.sciencedirect.com/science/article/abs/pii/S0020025517310976 (дата обращения: 22.01.2026). DOI: 10.1016/j.ins.2017.11.035.

17. Feng G., He J., Polson N. G. Deep learning for predicting asset returns // arXiv preprint arXiv:1804.09314. 2018. DOI: 10.48550/arXiv.1804.09314.