ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ СНИЖЕНИЯ ЭФФЕКТА ИНФОРМАЦИОННОЙ ФРАГМЕНТАЦИИ ПРИ ПРОГНОЗИРОВАНИИ ИНВЕСТИЦИОННОЙ ЭФФЕКТИВНОСТИ

Бекматов Акмал Курбонмахматович

Асистент Каршинского государственного технического университета

ORCID: 0009-0002-7212-0733

АННОТАЦИЯ

Рассматривается применение алгоритмов машинного обучения для прогнозирования инвестиционной эффективности в условиях информационной фрагментации. Проведён сравнительный анализ устойчивости методов к пропущенным значениям, структурным разрывам и разнородности источников. Установлено, что алгоритмы бустинга с встроенной обработкой NaN (CatBoost, LightGBM) превосходят традиционные методы в условиях несбалансированных панелей. Предложена схема аналитического конвейера, включающая обязательную диагностику механизма пропуска перед выбором алгоритма.

Ключевые слова: *машинное обучение; информационная фрагментация; XGBoost; LightGBM; CatBoost; пропущенные данные; инвестиционная эффективность; ансамблевые методы.*

APPLICATION OF MACHINE LEARNING METHODS TO REDUCE THE EFFECT OF INFORMATION FRAGMENTATION IN FORECASTING INVESTMENT EFFICIENCY

Akmal Kurbonmaxmatovich Bekmatov

Assistant, Karshi State Technical University

ORCID: 0009-0002-7212-0733

ABSTRACT

This paper examines the application of machine learning algorithms for forecasting investment efficiency under conditions of information fragmentation. A comparative analysis of the robustness of these methods to missing values, structural breaks, and heterogeneous data sources is conducted. The results indicate that boosting algorithms with built-in handling of missing values (CatBoost and LightGBM) outperform traditional methods in the context of unbalanced panel data. An analytical pipeline is proposed that includes mandatory diagnosis of the missing data mechanism prior to algorithm selection.

Keywords: machine learning; information fragmentation; XGBoost; LightGBM; CatBoost; missing data; investment efficiency; ensemble methods.

ВВЕДЕНИЕ

Согласно докладу UNCTAD (2024), глобальные ПИИ достигли 1,5 трлн долл. США, однако развивающиеся страны второй год подряд фиксируют их снижение — во многом вследствие геополитической фрагментации и неоднородности статистической базы. Аналитик, работающий с международными инвестиционными данными, неизбежно сталкивается с пропусками, структурными разрывами и методологической несопоставимостью источников. Традиционные регрессионные методы требуют полных сбалансированных панелей — условие, редко выполнимое на практике. Методы машинного обучения предлагают альтернативу, однако их эффективность зависит от корректного обращения с неполными данными. Цель статьи — сравнить алгоритмы МО по критерию устойчивости к информационной фрагментации и предложить методологическую рамку их применения в инвестиционном анализе.

Обзор литературы

Gu, Kelly и Xiu провели масштабное сравнение МО-алгоритмов на задаче оценки премий за риск и показали устойчивое превосходство ансамблевых методов над линейными моделями, особенно при высокой размерности признаков. Fujimoto et al. применили CatBoost для интерполяции нерандомных пропусков в базе ORBIS и установили, что механизм пропуска, а не его доля, определяет выбор метода восстановления. Систематический обзор Shehab et al. по протоколу PRISMA охватил 22 работы 2024–2026 гг. и подтвердил: LightGBM оптимален для высокочастотных рядов, CatBoost — при наличии категориальных переменных, ансамблевый стекинг стабильно превосходит одиночные алгоритмы. Обзор по управлению финансовыми рисками констатирует доминирование ансамблей в задачах кредитного скоринга и обнаружения мошенничества, отмечая «методологическую изолированность» большинства исследований, затрудняющую перенос результатов на развивающиеся рынки с заведомо более высокой фрагментацией данных.

Материалы и методы.

В качестве источников данных используются: UNCTAD FDI Statistics (потоки ПИИ, 1990–2023); World Development Indicators Всемирного банка (ВВП,

GFCF, 1960–2023); индикаторы ЕБРР по структурным реформам (1989–2023); база ORBIS (финансовая отчётность фирм, 2005–2022). Все источники характеризуются систематической неполнотой: ORBIS недопредставляет МСП развивающихся стран, UNCTAD фиксирует кондуитные потоки, статистика ЕБРР содержит разрывы 1990-х годов.

Перед обучением моделей необходима диагностика механизма пропуска. Выделяют три типа: MCAR (пропуск случаен), MAR (пропуск зависит от наблюдаемых переменных) и MNAR (пропуск зависит от самого пропущенного значения). В инвестиционных данных преобладает MNAR: компании с худшими показателями реже публикуют отчётность. Для диагностики применяется тест Литтла. При MCAR и MAR используются MICE и KNN-вменение; при MNAR — CatBoost с параметром nan_mode, обрабатывающим пропуски внутри алгоритма бустинга.

Базовая регрессионная постановка задачи: $\hat{y}_{it} = f(X_{it}) + \varepsilon_{it}$, где $\hat{y}_{it}$ — прогнозируемая инвестиционная отдача объекта i в период t; X_{it} — вектор признаков (возможно неполный); f — МО-алгоритм.

Для XGBoost регуляризованная функция потерь имеет вид:

$$L = \Sigma_i\, l(y_i, \hat{y}_i) + \Sigma_k\, \Omega(f_k), \quad \Omega(f) = \gamma T + \tfrac{1}{2}\lambda\|w\|^2,$$

где T — число листьев дерева; λ — коэффициент L2-регуляризации; w — веса листьев; γ — штраф за сложность.

Random Forest агрегирует прогнозы B деревьев:

$$\hat{y} = (1/B) \cdot \Sigma\_\{b=1\}\^\{B\}\ f\_b(x).$$

Ансамблевый стекинг объединяет прогнозы базовых моделей через мета-алгоритм второго уровня. В предлагаемой схеме базовые модели — RF, XGBoost, LightGBM; мета-алгоритм — гребневая регрессия. Интерпретация результатов осуществляется методом SHAP.

Схема аналитического конвейера

1. Сбор и аудит данных	→	2. Диагностика пропусков (MCAR/MAR/MNAR)	→	3. Вменение (MICE / KNN / CatBoost)	→	4. МО-модели + SHAP-интерпретация
UNCTAD, WB, EBRD, ORBIS	→	Little's test, паттерн-анализ	→	При MNAR — CatBoost; при MAR — MICE	→	RF, XGBoost, LightGBM, Stacking

Рисунок 1. Источник: составлено автором.

РЕЗУЛЬТАТЫ

Сравнительный анализ алгоритмов по методологическим критериям представлен в таблице 1. Прямые эмпирические метрики не приводятся ввиду отсутствия единой сопоставимой выборки; оценки основаны на опубликованных методологических характеристиках алгоритмов.

Таблица 1 — Сравнительный анализ алгоритмов МО по критериям устойчивости к фрагментации данных

Метод	Точность	Устойчивость к пропускам	Обработка NaN	Ограничения
Random Forest	Высокая	Средняя	Усреднение деревьев	Требует явного вменения при MNAR
XGBoost	Высокая	Высокая	Sparse-aware алгоритм	Чувствителен к гиперпараметрам
LightGBM	Очень высокая	Высокая	Гистограммный метод	Менее устойчив на малых выборках
CatBoost	Высокая	Очень высокая	nan_mode + Ordered Boosting	Больший объём памяти
LSTM	Высокая (ряды)	Низкая	Требует предобработки	Высокая вычислительная стоимость
Stacking	Наивысшая	Средняя	Зависит от базовых моделей	Риск переобучения

Источник: составлено на основе Gu et al.; Fujimoto et al.; Shehab et al..

CatBoost и LightGBM обладают наибольшей встроенной устойчивостью к пропускам за счёт явной обработки NaN внутри алгоритма бустинга. Random Forest требует предварительного вменения, однако агрегирование по деревьям снижает влияние единичных пропусков. LSTM демонстрирует высокий потенциал для временны́х рядов, но требует длинных и плотных обучающих выборок — условие, редко выполнимое при MNAR-данных. Стекинг стабильно превосходит одиночные алгоритмы, однако при малых выборках существенно возрастает риск переобучения мета-алгоритма.

ОБСУЖДЕНИЕ

Главное системное ограничение существующих работ — методологическая изолированность: алгоритмы тестируются на проприетарных данных развитых рынков, что затрудняет перенос результатов на развивающиеся экономики с более высоким уровнем фрагментации. Второе ограничение — игнорирование механизма пропуска: применение стандартного удаления строк (listwise deletion) при MNAR систематически завышает оценки

инвестиционной эффективности, поскольку наиболее прозрачные компании с наилучшими показателями публикуют полные данные.

Направлениями дальнейших исследований являются: разработка стандартизированных бенчмарк-наборов с контролируемыми паттернами фрагментации; интеграция SHAP-объяснений с пространственно-эконометрическими моделями для регионального инвестиционного анализа; применение генеративных моделей для аугментации неполных данных.

ЗАКЛЮЧЕНИЕ

Информационная фрагментация в инвестиционных данных носит системный характер и не устраняется выбором одного алгоритма. Алгоритмы бустинга, прежде всего CatBoost и LightGBM, структурно лучше приспособлены к таким данным, чем традиционные методы. Научная новизна состоит в предложении аналитического конвейера, включающего обязательный этап диагностики механизма пропуска как определяющего условия выбора алгоритма — в отличие от стандартных МО-пайплайнов, пропускающих этот шаг. Практическая применимость конвейера подтверждается его ориентацией на открытые данные UNCTAD, World Bank, EBRD и ORBIS, доступные большинству исследователей без специального доступа к проприетарным базам.

СПИСОК ЛИТЕРАТУРЫ

1. Gu S., Kelly B., Xiu D. Empirical Asset Pricing via Machine Learning // The Review of Financial Studies. — 2020. — Vol. 33, No. 5. — P. 2223–2273. — DOI: 10.1093/rfs/hhaa009.

2. Fujimoto S., Ishikawa A., Mizuno T., Watanabe T. Interpolation of Non-Random Missing Values in Financial Statements' Big Data Using CatBoost // Journal of Computational Social Science. — 2022. — Vol. 5, No. 2. — P. 911–938. — DOI: 10.1007/s42001-022-00165-9.

3. UNCTAD. World Investment Report 2024. — Geneva: United Nations, 2024. — URL: https://unctad.org/publication/world-investment-report-2024.

4. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // Proc. 22nd ACM SIGKDD. — New York: ACM, 2016. — P. 785–794. — DOI: 10.1145/2939672.2939785.

5. Shehab M. et al. Machine Learning and Deep Learning in Computational Finance. — arXiv: 2511.21588. — 2024.

6. Kurbonmaxmatovich, B. A. (2026). HUDUDLAR RIVOJLANISHIDA INVESTITSIYA SAMARADORLIGINI BAHOLASH MEXANIZMLARI. *O'ZBEKISTONDA FANLARARO INNOVATSIYALAR VA ILMIY TADQIQOTLAR JURNALI*, *4*(48), 117-121.