

Солбонов Артем Олегович, магистрант, Забайкальский государственный университет, г. Чита

**СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЭФФЕКТИВНОСТИ АНСАМБЛЕВЫХ
И ЛИНЕЙНЫХ МОДЕЛЕЙ ПРИ ПРОГНОИЗИРОВАНИИ ЦЕНЫ НА
ВОЛАТИЛЬНЫХ ВАЛЮТНЫХ РЫНКАХ**

Аннотация

В статье рассматривается задача сравнительного анализа современных прогнозных архитектур при моделировании цены дневных свечей в условиях высокой волатильности валютного рынка. В качестве исследуемых алгоритмов выбраны классическая линейная модель ARIMA, а также ансамблевые методы машинного обучения: Random Forest, XGBoost, LightGBM и CatBoost. Вычислительный эксперимент проведен на фиксированном оптимальном временном интервале обучения глубиной в 1 месяц (22 торговых дня). На основе метрик MAE и RMSE выявлены ключевые закономерности изменения предиктивной способности моделей в зависимости от математической структуры алгоритма и характера исследуемого актива.

Annotation

The paper considers the problem of comparative analysis of modern forecasting architectures when modeling the price of daily candlesticks under high volatility conditions of the foreign exchange market. The classic linear ARIMA model and ensemble machine learning methods (Random Forest, XGBoost, LightGBM, and CatBoost) were selected as the studied algorithms. The computational experiment was carried out on a fixed optimal training time interval with a depth of 1 month (22 trading days). Based on the MAE and RMSE metrics, the key regularities of changes

in the predictive capability of models depending on the mathematical structure of the algorithm and the nature of the asset under study are revealed.

Ключевые слова: машинное обучение, цена, ансамблевые модели, градиентный бустинг, CatBoost, LightGBM, ARIMA, временные ряды.

Keywords: machine learning, close, ensemble models, gradient boosting, CatBoost, LightGBM, ARIMA, time-series.

Введение и постановка задачи. Моделирование и краткосрочное прогнозирование цены на международном валютном рынке представляет собой задачу анализа нестационарных временных рядов, осложненную присутствием высокого уровня шума. В финансовой инженерии долгое время доминировал подход, основанный на линейных статистических моделях авторегрессии и скользящего среднего [1]. Однако точность таких методов снижается в периоды высокой волатильности и геополитической нестабильности.

С развитием анализа данных возникла необходимость комплексной оценки альтернативных подходов [2]. Эта работа посвящена решению задачи сравнительной оценки пяти разнородных моделей (линейная ARIMA и ансамблевые алгоритмы Random Forest, XGBoost, LightGBM, CatBoost) применительно к прогнозированию дневной цены. В качестве фиксированного диапазона обучения используется месячное окно (22 торговых дня).

Методология и описание алгоритмов. Архитектура системы решает задачу регрессии для цены в момент времени $t + 1$. Шаг прогноза строго равен одному дню. Входные признаки включают в себя лаговые смещения за предыдущие два дня, а также котировки спотового контракта на золото и индекса доллара.

В рамках работы проводится оценка следующих моделей:

- ARIMA (Autoregressive Integrated Moving Average). Реализует линейную комбинацию прошлых значений ряда и ошибок прогноза. Выступает в качестве классической модели [3].

– Random forest (Случайный лес). Построен на принципе бэггинга и случайных подпространств признаков. Алгоритм снижает дисперсию прогноза за счет усреднения ответов множества несвязанных решающих деревьев [4].

– Градиентный бустинг (XGBoost, LightGBM, CatBoost). Алгоритмы последовательного построения деревьев, где каждое последующее дерево минимизирует остаточную ошибку предыдущих шагов [5].

Математические обоснования метрик. Для количественной оценки исследуемых ансамблевых моделей и установления степени их отклонения от базового бенчмарка в работе используются две классические метрики регрессии: средняя абсолютная ошибка (MAE) и среднеквадратичная ошибка (RMSE).

Средняя абсолютная ошибка (MAE) вычисляется по формуле:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

где y_i – фактическое значение цены валютного инструмента на i -м шаге тестирования; $\hat{y}_i$ – прогнозное значение цены, сгенерированное моделью; n – общий объем тестовой выборки. Данная метрика позволяет оценить средний модуль отклонения прогноза от реального рынка, являясь линейным индикатором, где все индивидуальные ошибки имеют одинаковый вес.

Среднеквадратичная ошибка (RMSE) рассчитывается следующим образом:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

В отличие от MAE, метрика RMSE более жестко штрафует прогнозную систему за появление крупных единичных ошибок (промахов), так как разность между фактом и прогнозом возводится в квадрат перед процедурой усреднения. Использование комбинации данных метрик необходимо для выявления склонности моделей к генерации критических экстраполяционных выбросов в периоды аномальной рыночной волатильности.

Экономическое обоснование признаков. Выбор межрыночных индикаторов в качестве входящих признаков моделей машинного обучения обусловлен спецификой макроэкономической природы валютной пары USD/ZAR. Южноафриканский рэнд (ZAR) традиционно относится к категории товарных (сырьевых) валют с высокой развивающейся экономикой. ЮАР является одним из крупнейших мировых экспортеров благородных металлов, в связи с чем динамика спотовых цен на золото оказывает прямое влияние на платежный баланс страны и на курс национальной валюты. Интеграция лаговых значений золота позволяет моделям машинного обучения улавливать долгосрочные межрыночные тренды.

С другой стороны, индекс доллара США (DXY), оценивающий отношение американской валюты к корзине из шести ведущих мировых валют, выступает в качестве глобального барометра рыночного риска. В периоды геополитической нестабильности или ужесточения монетарной политики ФРС США происходит массовый отток капитала из рискованных активов развивающихся рынков (включая ZAR) в защитные инструменты долларовой зоны. Таким образом, включение DXY в признаковое пространство предоставляет ансамблевым моделям контекст глобальной ликвидности, что критически важно для стабилизации нелинейных прогнозов в моменты резких структурных сдвигов.

Вычислительный эксперимент и анализ результатов. Для проведения вычислительного эксперимента были использованы исторические данные котировки валютной пары USD/ZAR, характеризующейся высоким уровнем волатильности. Обучение моделей проводилось на фиксированном временном интервале глубиной в 1 месяц (22 торговых дня). Разделение выборки на обучающую и тестовую реализовано в хронологическом порядке без случайного перемешивания, что исключает риски заглядывания в будущее.

Полученные метрики ошибок прогнозирования отображены в таблице 1.

Таблица 1 – метрики ошибок на месячном интервале для USD/ZAR.

Модель	Окно обучения	MAE	RMSE
ARIMA	месяц	0.006115	0.007200
Random forest	месяц	0.005419	0.007500
XGBoost	месяц	0.005991	0.007600
LightGBM	месяц	0.006394	0.008000
CatBoost	месяц	0.006018	0.007700

Анализ полученных данных позволяет выявить следующие ключевые закономерности:

- Превосходство бэггинга на зашумленных интервалах. Наилучший результат по метрике абсолютной ошибки зафиксирован у Случайного леса (MAE = 0.005419). Это можно объяснить тем, что механизмы построения глубоких независимых деревьев и усреднения их предсказаний эффективно выполняют роль низкочастотного фильтра, сглаживая агрессивный рыночный шум, свойственный южноафриканскому рэнду на месячной дистанции.

- Специфика линейного оператора и среднеквадратичной ошибки. Модель ARIMA показала наилучшее значение по метрике RMSE (0.007200). Поскольку формула среднеквадратичной ошибки возводит остатки в квадрат, она крайне чувствительна к крупным единичным промахам. Строгая математическая структура линейной авторегрессии минимизирует дисперсию таких крупных выбросов, в то время как нелинейные древовидные модели бустинга (XGBoost, LightGBM) склонны к локальному переобучению на спекулятивных ценовых всплесках.

- Сравнительный анализ алгоритмов бустинга. Внутри группы градиентного бустинга лидерство удерживает XGBoost (MAE = 0.005991), незначительно опережая CatBoost. Легковесный алгоритм LightGBM показал наибольшую ошибку среди ансамблей (MAE = 0.006394), что объясняется его стратегией полистового роста (leaf-wise), которая на малых объемах данных (22 бара) и высоком уровне шума может приводить к избыточной подгонке под локальные паттерны.

Углубленный анализ бустинга. Разница в эффективности бустинговых моделей (XGBoost, LightGBM, CatBoost) на изолированном месячном контуре обучения также тесно связана с внутренними алгоритмическими различиями их реализации:

- Алгоритм XGBoost использует классическую стратегию построения деревьев по уровням (leaf-wise). Это обеспечивает более жесткий контроль геометрии дерева, предотвращая глубокое разрастание отдельных ветвей на малых выборках. Именно эта особенность позволила ему показать стабильный результат ($MAE = 0.0059991$).

- В свою очередь библиотека CatBoost опирается на построение симметричных решающих деревьев (oblivious trees). В таких структурах на каждом уровне дерева используется один и тот же признак разделения, что накладывает сильное регуляризационное ограничение. Для зашумленного ряда USD/ZAR на коротком окне обучения это послужило эффективным барьером против переобучения, выведя модель на второе место среди бустингов.

- Архитектура LightGBM использует полистовой рост (leaf-wise) с ограничением по максимальной глубине. При наличии высокого стохастического шума и малой длины выборки (22 торговых дня) данный подход склонен находить ложные нерелевантные закономерности в единичных листьях, что и привело к фиксации наихудшей точности ($MAE = 0.006394$) в ансамблевой группе.

Заключение. В результате проведенного исследования была успешно подтверждена гипотеза о высокой эффективности ансамблевых методов машинного обучения при краткосрочном прогнозировании цены на волатильных валютных рынках. Экспериментально доказано, что в рамках локального месячного контекста алгоритм Случайного леса обеспечивает максимальную точность за счет устойчивости к шуму, в то время как линейная модель ARIMA остается наиболее стабильным инструментом с точки зрения минимизации крупных ошибок прогнозирования (RMSE).

Литература

1. Брукс, К. Информативный анализ временных рядов в финансовых приложениях / К. Брукс ; пер. с англ. – М. : Кембридж Пресс, 2019. – 720 с.
2. Джеймс, Г. Введение в статистическое обучение с примерами на языке R / Г. Джеймс, Д. Уиттон, Т. Хастис, Р. Тибширани ; пер. с англ. С. Э. Мастицкого. – М. : ДМК Пресс, 2017. – 440 с.
3. Хайндман, Р. Д. Прогнозирование: принципы и практика / Р. Д. Хайндман, Г. Атанасопулос ; пер. с англ. — Мельбурн, Австралия : OTexts, 2018. — 570 с.
4. Breiman, L. Random Forests / L. Breiman // Machine Learning. – 2001. – Vol. 45, No. 1. – P. 5–32.
5. Chen, T. XGBoost: A Scalable Tree Boosting System / T. Chen, C. Guestrin // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2016. – P. 785–794.

Literature

1. Brooks, C. (2019), Informativnyj analiz vremennyh ryadov v finansovyh prilozheniyah [Introductory Econometrics for Finance], Translated from English, Moscow, Cambridge Press, 720 p. (In Russ.)
2. James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017), Vvedenie v statisticheskoe obuchenie s primerami na yazyke R [An Introduction to Statistical Learning: with Applications in R], Translated by S. E. Mastickij, Moscow, DMK Press, 440 p. (In Russ.)
3. Hyndman, R. J. and Athanasopoulos, G. (2018), Prognozirovanie: principy i praktika [Forecasting: principles and practice], Translated from English, Melbourne, Australia, OTexts, 570 p. (In Russ.)
4. Breiman, L. (2001), "Random Forests", Machine Learning, Vol. 45, No. 1, pp. 5–32.

5. Chen, T. and Guestrin, C. (2016), "XGBoost: A Scalable Tree Boosting System", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794.