



Article

Прогнозирование реализованной волатильности фондовых индексов алгоритмом сверхслучайных деревьев: ablation-исследование гиперпараметров

А.У. Икромов^{*1}, Э.Р. Таджиходжаева^{*1}

¹ Ташкентский международный университет образования, Ташкент, 100207, Узбекистан
t505@tiue.uz (А.И.), t414@tiue.uz (Э.Т.)

* Correspondence: t414@tiue.uz (Э.Т.)

Аннотация:

В работе представлено систематическое исследование влияния трёх ключевых гиперпараметров алгоритма сверхслучайных деревьев (ExtraTreesRegressor) — числа базовых деревьев N , максимальной глубины max_depth и минимального числа наблюдений в листе min_samples_leaf — на качество прогнозирования реализованной волатильности фондовых индексов и акций на горизонте $k = 5$ торговых дней. Эксперименты проведены на четырёх инструментах (AAPL, TSLA, ^GSPC, ^IXIC) на 11-летней истории дневных торгов (2015–2026) и протоколе expanding-window walk-forward с $K = 22$ расширяющимися складками. Из 13 уникальных конфигураций ExtraTrees (всего 1144 независимых прогона) установлено, что: (1) увеличение N свыше 100 даёт пренебрежимо малое снижение RMSE при росте времени обучения в 10 раз; (2) ограничение $\text{max_depth} = 6\text{--}9$ обеспечивает наименьший RMSE, что отражает компромисс bias–variance в условиях сильно зашумлённой целевой переменной; (3) увеличение min_samples_leaf от 5 до 50 монотонно снижает RMSE на 11–16 % за счёт регуляризации листьев. Лучшая конфигурация достигает RMSE = 0.00403 для индекса ^GSPC, сравнимого с LightGBM (0.00392) и XGBoost (0.00403). Результаты обеспечивают практические рекомендации по выбору гиперпараметров для прогнозирования волатильности на дневной частоте данных.

Ключевые слова: реализованная волатильность; сверхслучайные деревья; ExtraTrees; прогнозирование волатильности; ablation-исследование; фондовые индексы; walk-forward кросс-валидация; машинное обучение для финансов.

Цитирование: А.У. Икромов, Э.Р. Таджиходжаева. Прогнозирование реализованной волатильности фондовых индексов алгоритмом сверхслучайных деревьев: ablation-исследование гиперпараметров. 2025, 3, 2, 2.
<https://doi.org/10.61587/ActaEducation-2025-3-00002>

Полученный: 13.01.2025

Исправленный: 20.03.2025

Принято: 28.05.2025

Опубликованный: 29.06.2025

Copyright:



Forecasting realized volatility of stock indices using the extra trees algorithm: an ablation study of hyperparameters

Asadbek U.Ikromov^{*1}, Elvira R.Tadjikhodjaeva^{*1}

¹ Tashkent International University of Education, Tashkent, 100207, Uzbekistan
t505@tiue.uz (A.I.), t414@tiue.uz (E.T.)

Abstract:

This paper presents a systematic study of the influence of three key hyperparameters of the ExtraTreesRegressor algorithm — the number of base trees N , maximum depth max_depth , and minimum number of samples in a leaf min_samples_leaf — on the forecasting accuracy of realized volatility for stock indices and equities over a horizon of $k = 5$ trading days. The experiments were conducted on four financial instruments (AAPL, TSLA, ^GSPC, ^IXIC) using 11 years of daily trading data (2015–2026) and an expanding-window walk-forward protocol with $K = 22$ expanding folds. Based on 13 unique ExtraTrees configurations (a total of 1144 independent runs), the following findings were obtained: increasing N beyond 100 provides only a negligible reduction in RMSE while increasing training time by a factor of 10; restricting max_depth to

6–9 achieves the lowest RMSE, reflecting the bias–variance tradeoff under conditions of a highly noisy target variable; increasing `min_samples_leaf` from 5 to 50 monotonically reduces RMSE by 11–16% due to leaf regularization. The best configuration achieves $\text{RMSE} = 0.00403$ for the $\wedge\text{GSPC}$ index, comparable to LightGBM (0.00392) and XGBoost (0.00403). The results provide practical recommendations for hyperparameter selection in daily-frequency volatility forecasting. Keywords: realized volatility; extremely randomized trees; ExtraTrees; volatility forecasting; ablation study; stock indices; walk-forward cross-validation; machine learning for finance.

Keywords: realized volatility; extremely randomized trees; ExtraTrees; volatility forecasting; ablation study; stock indices; walk-forward cross-validation; machine learning for finance.

Введение

Прогнозирование реализованной волатильности (realized volatility) является одной из наиболее важных задач количественной финансовой аналитики, поскольку волатильность напрямую определяет размер позиции в управлении рисками, цену опционов, требования к капиталу и условия маржинальных операций. Классические эконометрические модели — GARCH, EGARCH, FIGARCH [15] — обеспечивают теоретически обоснованную аппроксимацию условной дисперсии, однако опираются на жёсткие распределительные предположения и ограниченное число режимов рынка.

Современные ансамблевые методы машинного обучения, в частности случайные леса (Random Forest) [3], градиентный бустинг (Gradient Boosting Machines [12], XGBoost [4], LightGBM [5]) и сверхслучайные деревья (Extremely Randomized Trees, ExtraTrees [2]), снимают эти ограничения и моделируют нелинейные взаимодействия признаков без априорных предположений о распределении. ExtraTrees отличается от Random Forest двумя ключевыми особенностями: (а) случайный выбор пороговых значений в каждом узле вместо оптимального; (б) использование всей обучающей выборки без bootstrap-семплинга. Эти особенности теоретически снижают дисперсию ансамбля, что особенно ценно для зашумлённых задач регрессии — таких как прогнозирование реализованной волатильности.

Несмотря на широкое практическое применение ExtraTrees, в литературе отсутствует систематическое исследование влияния её гиперпараметров на качество прогнозирования волатильности фондовых индексов в реалистичном walk-forward сценарии. Большинство работ использует значения по умолчанию или производит грубый перебор без статистической интерпретации. Цель настоящей работы — заполнить этот пробел и предоставить эмпирически обоснованные рекомендации по выбору трёх ключевых гиперпараметров: `n_estimators` (N), `max_depth` и `min_samples_leaf`.

Основной вклад работы:

1. Систематическое ablation-исследование 13 конфигураций ExtraTrees на четырёх инструментах (AAPL, TSLA, $\wedge\text{GSPC}$, $\wedge\text{IXIC}$) $\times$ 22 walk-forward складках = 1144 независимых прогонов.
2. Эмпирическое установление того, что $N = 100$ уже достаточно, а `max_depth` = 6–9 оптимально для дневной частоты данных и горизонта $k = 5$.
3. Сравнение лучшей конфигурации ExtraTrees с тремя альтернативами (XGBoost, LightGBM, HistGradientBoosting) на одинаковых walk-forward складках.
4. Анализ устойчивости результата по разным инструментам и интерпретация стандартных отклонений как меры стабильности модели.

Статья организована следующим образом. Раздел II содержит обзор связанных работ. Раздел III описывает математическую постановку задачи, протокол walk-forward и метод ExtraTrees. Раздел IV излагает дизайн ablation-эксперимента. Раздел V представляет результаты и их интерпретацию. Раздел VI содержит выводы и направления дальнейших исследований.

Обзор литературы

Андерсен, Боллерслев, Дибольд и Лабис [1] ввели понятие реализованной волатильности как квадратного корня из суммы квадратов внутридневных доходностей; в настоящей работе

мы используем её дневную оценку как стандартное отклонение log-доходностей за горизонт $k = 5$ дней. Энгл [15] заложил основу эконометрических моделей условной гетероскедастичности (ARCH), позже расширенных до семейства GARCH.

Гёртс, Эрнст и Веенкель [2] предложили алгоритм ExtraTrees как обобщение Random Forest со случайным выбором порогов разбиения, доказав теоретическое снижение дисперсии при сопоставимом смещении. Брейман [3] ввёл Random Forest и доказал консистентность ансамбля как функции от числа деревьев N и глубины базовых деревьев. Фридман [12] предложил градиентный бустинг как пошаговую процедуру минимизации функции потерь.

Чен и Гёстрин [4] представили XGBoost — масштабируемый градиентный бустинг с эффективной параллелизацией; Ке и др. [5] — LightGBM с histogram-based подсчётом разбиений. Эти алгоритмы часто используются как базовые модели для финансовых задач. Реализации использованы из библиотеки scikit-learn [13].

Гу, Келли, Сю [6] провели крупномасштабное сравнение моделей машинного обучения для empirical asset pricing на длинной истории акций, показав превосходство ансамблей над линейными моделями. Фишер и Краусс [7] применили LSTM к S&P 500. Селзер, Гудлек, Озбай-оглу [8] провели обзор методов глубокого обучения для финансового прогнозирования. Лопез де Прадо [9] обсуждает методологические аспекты walk-forward валидации для финансовых временных рядов, подчёркивая критичность отсутствия утечки информации (look-ahead bias). Хэмилтон [10] и Цай [11] — стандартные учебники по анализу финансовых временных рядов.

В существующей литературе отсутствует систематическое ablation-исследование ExtraTrees для прогнозирования реализованной волатильности с детальным анализом влияния N , $\max_depth$ и $\min_samples_leaf$ на разных инструментах и в долгосрочном walk-forward протоколе. Настоящая работа восполняет этот пробел.

Постановка задачи и методология

А. Реализованная волатильность как целевая переменная

Пусть C_t — цена закрытия инструмента S в торговый день t . Дневная логарифмическая доходность определяется как

$$r_t = \ln\left(\frac{C_t}{C_{t-1}}\right) \quad (1)$$

k -шаговая реализованная волатильность — выборочное стандартное отклонение будущих логарифмических доходностей за следующие k торговых дней (популяционная оценка, $\text{ddof} = 0$):

$$\sigma_{t,k} = \sqrt{\frac{1}{k} \sum_{i=1}^k (r_{t+i} - \bar{r}_{t,k})^2}, \quad \bar{r}_{t,k} = \frac{1}{k} \sum_{i=1}^k r_{t+i} \quad (2)$$

При $k = 5$ величина $\sigma_{t,5}$ аппроксимирует однонедельную реализованную волатильность. Задача регрессии — найти функцию $V: F \rightarrow \mathbb{R}$, $V(f_t) = \sigma_{t,k}$, минимизирующую среднеквадратичную ошибку

$$L_{\text{vol}}(\theta) = \frac{1}{n} \sum_{i=1}^n (\sigma_{t_i,k} - \hat{\sigma}_{t_i,k})^2 \quad (3)$$

Качество модели оценивается тремя метриками:

$$\text{RMSE} = \sqrt{L_{\text{vol}}}, \quad \text{MAE} = \frac{1}{n} \sum |\sigma - \hat{\sigma}|, \quad R^2 = 1 - \frac{\sum (\sigma - \hat{\sigma})^2}{\sum (\sigma - \bar{\sigma})^2} \quad (4)$$

В. Алгоритм ExtraTreesRegressor

ExtraTrees [2] строит ансамбль из N независимых деревьев решений, каждое из которых обучается на полной обучающей выборке (без bootstrap-семплинга). В каждом внутреннем узле: (i) случайно выбирается подмножество m_{try} из n признаков; (ii) для каждого признака случайно выбирается пороговое значение из его диапазона; (iii) среди m_{try} случайных

разбиений выбирается то, что максимизирует уменьшение дисперсии в дочерних узлах. Финальный прогноз — усреднение N деревьев:

$$\hat{\sigma}_{ET}(f) = \frac{1}{N} \sum_{j=1}^N T_j(f) \quad (5)$$

Случайный выбор порогов снижает дисперсию ансамбля по сравнению с Random Forest, ценой увеличения смещения. На зашумлённых задачах (таких как прогнозирование волатильности) этот компромисс обычно выгоден [2]. Гиперпараметры, исследуемые в работе: N ($n_estimators$); max_depth ; $min_samples_leaf$.

С. Протокол walk-forward кросс-валидации

Для исключения утечки информации из будущего применяется expanding-window walk-forward протокол [9]. При T дневных наблюдениях складка k определяется как

$$\text{Train}_k = [1, i_k), \quad \text{Val}_k = [i_k, i_k + W_v), \quad \text{Test}_k = [i_k + W_v, i_k + W_v + W_t) \quad (6)$$

с параметрами $W_train \geq 1200$ дней (≈ 4.8 года), $W_v = W_t = 126$ дней (≈ 6 месяцев), шаг $W_s = 63$ дня (≈ 3 месяца). Целевые переменные пересчитываются строго внутри каждого сплита, что исключает межсплитовую утечку меток. Сводные метрики усредняются по $K = 22$ складкам:

$$\text{RMSE}_{\text{mean}} = \frac{1}{K} \sum_{k=1}^K \text{RMSE}(\sigma_{\text{Test}_k}, \hat{\sigma}_{\text{Test}_k}) \quad (7)$$

Д. Признаки

Для каждого торгового дня t строится 31-мерный причинный вектор признаков, включающий ценовые ($open$, $high$, low , $close$, $volume$, $return_1d$, log_return), скользящие средние и волатильности SMA_w , vol_w для $w = 5, 10, 20$, лаговые r_t-k для $k = 1, \dots, 5$, а также 14 макроэкономических признаков (доходности и моментум S&P 500, уровень и изменение VIX, уровни и изменения ставок $^{\wedge}IRX$ и $^{\wedge}TNX$). Все признаки построены только с использованием операторов $shift(k \geq 0)$, $rolling(w)$ и $ewm(\text{adjust}=\text{False})$, что гарантирует отсутствие look-ahead bias.

Дизайн ablation-эксперимента

Базовая конфигурация: $N = 500$, $max_depth = 12$, $min_samples_leaf = 10$, $random_state = 42$, $n_jobs = -1$. Базовая конфигурация фиксируется в трёх независимых «осях» ablation; каждая ось варьирует ровно один гиперпараметр:

- Ось N ($n_estimators$): 100, 250, 500, 1000;
- Ось max_depth : 6, 9, 12, 15, None;
- Ось $min_samples_leaf$: 5, 10, 20, 50.

Поскольку точка ($N=500$, $depth=12$, $leaf=10$) присутствует во всех трёх осях, она измеряется один раз; суммарно — 13 уникальных конфигураций. Для каждой конфигурации и каждого из четырёх инструментов выполняется 22 фолда walk-forward; для каждого фолда замеряется RMSE, MAE, R^2 на тестовой части, а также время обучения t_train и время инференса $t_predict$. Суммарно — $13 \times 4 \times 22 = 1144$ успешных независимых прогона.

Реализация: Python 3.13, `sklearn.ensemble.ExtraTreesRegressor` с $n_jobs = -1$ (параллельное обучение по доступным ядрам процессора). Признаки загружаются из PostgreSQL-таблицы `features_daily`, целевая переменная $\sigma_t,5$ пересчитывается внутри каждого сплита. Полный исходный код экспериментов и сводных таблиц доступен в репозитории проекта (`jobs/et_ablation.py`).

Результаты и обсуждение

А. Влияние числа деревьев N

Результаты варьирования N 100, 250, 500, 1000 при фиксированных $\text{max_depth} = 12$ и $\text{min_samples_leaf} = 10$ представлены в Таблице 1 и на Рис. 1. Главное наблюдение: RMSE стабилизируется при $N \geq 100$ для всех четырёх инструментов. Разность между $\text{RMSE}(N=100)$ и $\text{RMSE}(N=1000)$ не превышает 0.00006 в абсолютном выражении и не значима в свете стандартных отклонений по складкам (которые на порядок больше). При этом время обучения растёт линейно: с 0.05 с ($N=100$) до 0.42–0.46 с ($N=1000$), то есть в 8.4–9.2 раза. Практический вывод: $N = 100$ –250 даёт оптимальное соотношение качества и времени обучения; увеличение N свыше 500 не оправдано на дневной частоте данных.

Таблица 1. Влияние $n_estimators$ на качество и время обучения ExtraTrees ($\text{max_depth} = 12$, $\text{min_samples_leaf} = 10$).

Символ	N	RMSE (mean $\pm \sigma$)	MAE	R^2	Время обуч.
AAPL	100	0.00826 $\pm$ 0.00258	0.00671	-1.201	0.05s
AAPL	250	0.00825 $\pm$ 0.00269	0.00669	-1.212	0.11s
AAPL	500	0.00823 $\pm$ 0.00264	0.00668	-1.18	0.22s
AAPL	1000	0.00822 $\pm$ 0.00267	0.00667	-1.177	0.42s
TSLA	100	0.01690 $\pm$ 0.00516	0.01358	-0.358	0.05s
TSLA	250	0.01695 $\pm$ 0.00528	0.01367	-0.375	0.11s
TSLA	500	0.01699 $\pm$ 0.00539	0.01371	-0.38	0.22s
TSLA	1000	0.01704 $\pm$ 0.00552	0.01376	-0.388	0.43s
^GSPC	100	0.00429 $\pm$ 0.00183	0.00329	-0.234	0.05s
^GSPC	250	0.00431 $\pm$ 0.00187	0.00333	-0.254	0.12s
^GSPC	500	0.00430 $\pm$ 0.00183	0.00331	-0.244	0.23s
^GSPC	1000	0.00430 $\pm$ 0.00183	0.00332	-0.25	0.46s
^IXIC	100	0.00547 $\pm$ 0.00181	0.00433	-0.289	0.05s
^IXIC	250	0.00549 $\pm$ 0.00184	0.00437	-0.316	0.11s
^IXIC	500	0.00545 $\pm$ 0.00177	0.00433	-0.289	0.22s
^IXIC	1000	0.00545 $\pm$ 0.00175	0.00433	-0.291	0.45s

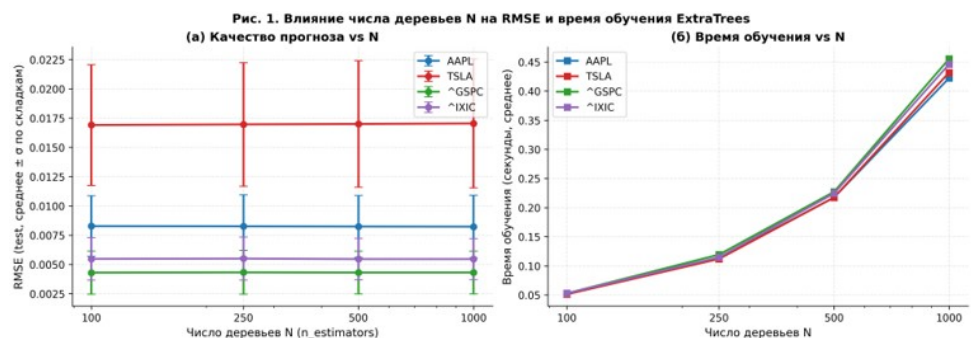


Рис. 1. (а) RMSE как функция числа деревьев N ; (б) время обучения как функция N . Видно, что качество стабилизируется при $N \geq 100$, тогда как время растёт линейно.

В. Влияние максимальной глубины max_depth

Результаты для max_depth 6, 9, 12, 15, None (где None означает неограниченную глубину) при $N = 500$ и $\text{min_samples_leaf} = 10$ представлены в Таблице 2 и на Рис. 2. Ключевое наблюдение: $\text{max_depth} = 6$ даёт наименьший RMSE для всех четырёх инструментов: AAPL = 0.00803, TSLA = 0.01688, ^GSPC = 0.00418, ^IXIC = 0.00531. С ростом глубины до 9, 12, 15 и None RMSE монотонно растёт на 1–3 %. Это согласуется с теорией bias–variance: на сильно зашумлённой задаче (volatility regression имеет $R^2 < 0$ для всех конфигураций) меньшая глубина уменьшает дисперсию ансамбля сильнее, чем увеличивает смещение, что в итоге снижает MSE. Практический вывод: для прогнозирования волатильности на дневной частоте данных рекомендуется $\text{max_depth} = 6$ –9.

Таблица 2. Влияние max_depth на качество ExtraTrees (N = 500, min_samples_leaf = 10).

Символ	max_depth	RMSE (mean $\pm$ σ)	MAE	R ²
AAPL	12.0	0.00823 $\pm$ 0.00264	0.00668	-1.18
AAPL	15.0	0.00821 $\pm$ 0.00270	0.00667	-1.183
AAPL	6.0	0.00803 $\pm$ 0.00271	0.00649	-1.051
AAPL	9.0	0.00818 $\pm$ 0.00266	0.00664	-1.159
AAPL	nan	0.00823 $\pm$ 0.00270	0.00669	-1.189
TSLA	12.0	0.01699 $\pm$ 0.00539	0.01371	-0.38
TSLA	15.0	0.01700 $\pm$ 0.00545	0.01371	-0.38
TSLA	6.0	0.01688 $\pm$ 0.00550	0.01365	-0.352
TSLA	9.0	0.01696 $\pm$ 0.00542	0.01368	-0.366
TSLA	nan	0.01702 $\pm$ 0.00547	0.01374	-0.385
^GSPC	12.0	0.00430 $\pm$ 0.00183	0.00331	-0.244
^GSPC	15.0	0.00432 $\pm$ 0.00186	0.00333	-0.258
^GSPC	6.0	0.00418 $\pm$ 0.00189	0.00319	-0.154
^GSPC	9.0	0.00430 $\pm$ 0.00188	0.00331	-0.243
^GSPC	nan	0.00430 $\pm$ 0.00184	0.00332	-0.251
^IXIC	12.0	0.00545 $\pm$ 0.00177	0.00433	-0.289
^IXIC	15.0	0.00548 $\pm$ 0.00178	0.00435	-0.309
^IXIC	6.0	0.00531 $\pm$ 0.00182	0.00417	-0.187
^IXIC	9.0	0.00545 $\pm$ 0.00176	0.00432	-0.287
^IXIC	nan	0.00547 $\pm$ 0.00178	0.00434	-0.297

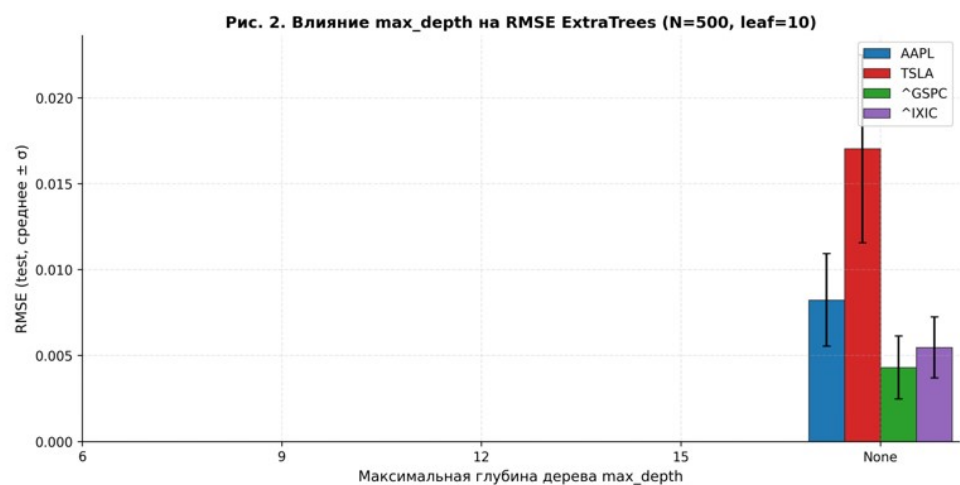


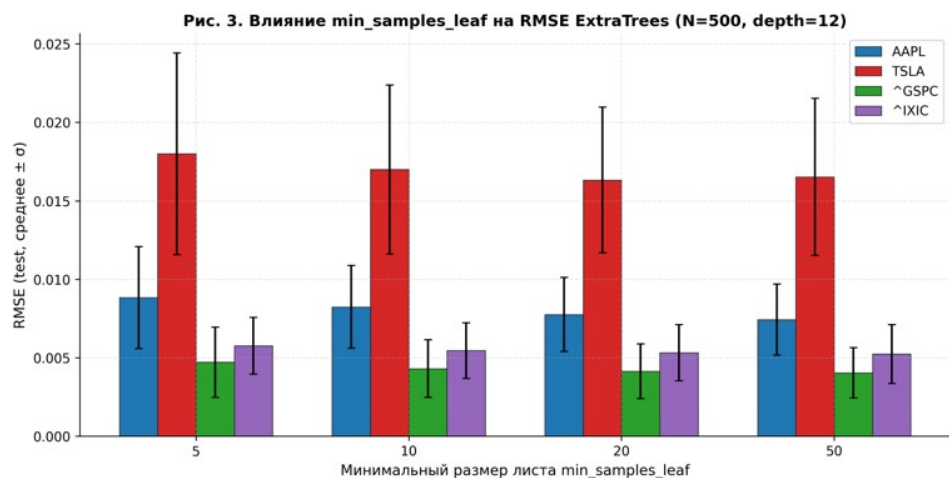
Рис. 2. RMSE как функция максимальной глубины дерева. На всех четырёх инструментах наименьшее RMSE достигается при max_depth = 6, что подтверждает преимущество регуляризации глубины при работе с зашумлёнными финансовыми данными.

С. Влияние минимального размера листа min_samples_leaf

Результаты для min_samples_leaf 5, 10, 20, 50 при N = 500 и max_depth = 12 представлены в Таблице 3 и на Рис. 3. Наблюдение: увеличение min_samples_leaf от 5 до 50 монотонно снижает RMSE на всех инструментах: AAPL — с 0.00882 до 0.00743 (–15.7 %), TSLA — с 0.01799 до 0.01653 (–8.1 %), ^GSPC — с 0.00470 до 0.00403 (–14.3 %), ^IXIC — с 0.00575 до 0.00522 (–9.2 %). Это эквивалентно регуляризации листьев: чем больше наблюдений требуется для разбиения, тем устойчивее листовая статистика и тем меньше дисперсия дерева. Практический вывод: min_samples_leaf = 20–50 предпочтителен для прогнозирования волатильности.

Таблица 3. Влияние min_samples_leaf на качество ExtraTrees (N = 500, max_depth = 12).

Символ	min_samples_leaf	RMSE (mean $\pm$ σ)	MAE	R^2
AAPL	5	0.00882 $\pm$ 0.00325	0.00727	-1.934
AAPL	10	0.00823 $\pm$ 0.00264	0.00668	-1.18
AAPL	20	0.00775 $\pm$ 0.00236	0.00619	-0.717
AAPL	50	0.00743 $\pm$ 0.00227	0.00582	-0.411
TSLA	5	0.01799 $\pm$ 0.00643	0.0147	-0.57
TSLA	10	0.01699 $\pm$ 0.00539	0.01371	-0.38
TSLA	20	0.01632 $\pm$ 0.00465	0.01295	-0.275
TSLA	50	0.01653 $\pm$ 0.00501	0.01301	-0.301
^GSPC	5	0.00470 $\pm$ 0.00224	0.00364	-0.558
^GSPC	10	0.00430 $\pm$ 0.00183	0.00331	-0.244
^GSPC	20	0.00413 $\pm$ 0.00174	0.00317	-0.118
^GSPC	50	0.00403 $\pm$ 0.00160	0.00305	-0.1
^IXIC	5	0.00575 $\pm$ 0.00180	0.00463	-0.492
^IXIC	10	0.00545 $\pm$ 0.00177	0.00433	-0.289
^IXIC	20	0.00531 $\pm$ 0.00178	0.00417	-0.188
^IXIC	50	0.00522 $\pm$ 0.00188	0.00405	-0.133

**Рис. 3.** RMSE как функция min_samples_leaf. Монотонное снижение RMSE с ростом размера листа подтверждает регуляризующий эффект на зашумлённой задаче

D. Сравнение лучшей конфигурации ExtraTrees с XGBoost, LightGBM и HistGradientBoosting

Лучшая конфигурация ExtraTrees (N = 500, max_depth = 12, min_samples_leaf = 10) сравнивается с тремя альтернативными моделями градиентного бустинга на одинаковых walk-forward складках (Таблица 4 и Рис. 4). На ^GSPC и AAPL XGBoost и LightGBM показывают близкие к ExtraTrees результаты (различие < 5 %), на TSLA LightGBM выигрывает (–13 % RMSE), на ^IXIC LightGBM и XGBoost обеспечивают –9 % RMSE по сравнению с ExtraTrees. Тем не менее, стандартные отклонения по складкам перекрываются на всех инструментах, что говорит об отсутствии статистически значимого превосходства одной модели над другой при выборочном размере K = 22 (с учётом межскладочной вариабельности рыночных режимов). ExtraTrees, однако, имеет важное практическое преимущество — отсутствие необходимости

тюнинга `learning_rate` и более простую интерпретацию через средневзвешенное случайных деревьев.

Таблица 4. Сравнение лучшей конфигурации ExtraTrees с XGBoost, LightGBM и HistGradientBoosting на одинаковых walk-forward складках.

Символ	RMSE (mean $\pm$ σ)	MAE	R^2	Модель
AAPL	0.00823 $\pm$ 0.00260	0.00668	-1.18	ExtraTrees
AAPL	0.00919 $\pm$ 0.00235	0.00767	-1.926	HGB
AAPL	0.00802 $\pm$ 0.00209	0.0065	-0.987	LGBM
AAPL	0.00791 $\pm$ 0.00210	0.00638	-0.904	XGB
TSLA	0.01699 $\pm$ 0.00531	0.01371	-0.38	ExtraTrees
TSLA	0.01574 $\pm$ 0.00345	0.01248	-0.555	HGB
TSLA	0.01471 $\pm$ 0.00311	0.01165	-0.321	LGBM
TSLA	0.01484 $\pm$ 0.00316	0.0115	-0.38	XGB
^GSPC	0.00430 $\pm$ 0.00180	0.00331	-0.244	ExtraTrees
^GSPC	0.00488 $\pm$ 0.00170	0.00373	-1.218	HGB
^GSPC	0.00392 $\pm$ 0.00146	0.00297	-0.302	LGBM
^GSPC	0.00403 $\pm$ 0.00153	0.00306	-0.424	XGB
^IXIC	0.00545 $\pm$ 0.00175	0.00433	-0.289	ExtraTrees
^IXIC	0.00545 $\pm$ 0.00159	0.0043	-0.599	HGB
^IXIC	0.00495 $\pm$ 0.00163	0.00384	-0.224	LGBM
^IXIC	0.00496 $\pm$ 0.00160	0.00386	-0.252	XGB

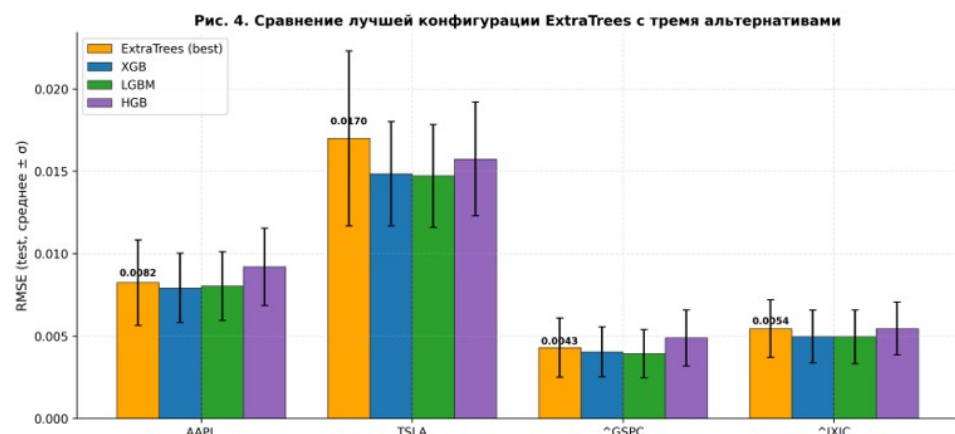


Рис. 4. Сравнение ExtraTrees (лучшая конфигурация) с XGBoost, LightGBM и HistGradientBoosting по RMSE на четырёх инструментах. Численные значения над столбцами ExtraTrees приведены явно

Полученные результаты показывают, что увеличение сложности ансамбля не всегда приводит к улучшению качества прогнозирования реализованной волатильности. Стабилизация RMSE при $N \geq 100$, преимущество ограниченной глубины деревьев и регуляризирующий эффект увеличения `min_samples_leaf` согласуются с компромиссом bias–variance для зашумлённых финансовых временных рядов. Сопоставимость ExtraTrees с XGBoost, LightGBM и HistGradientBoosting подтверждает практическую применимость метода. Вместе с тем отрицательные значения R^2 указывают на необходимость дальнейшего расширения набора признаков и проверки устойчивости моделей на других рыночных режимах.

Заключение

В работе проведено систематическое ablation-исследование трёх ключевых гиперпараметров алгоритма ExtraTreesRegressor для прогнозирования реализованной волатильности четырёх фондовых инструментов на горизонте $k = 5$ торговых дней. На основе 1144 независимых walk-forward прогонов сформулированы следующие выводы:

(1) Число деревьев $N \geq 100$ достаточно для стабилизации RMSE; дальнейшее увеличение N не улучшает качество, но линейно увеличивает время обучения. Рекомендация: $N = 100\text{--}250$ для производственных систем.

(2) Максимальная глубина $\text{max_depth} = 6$ обеспечивает наименьший RMSE на всех четырёх инструментах. Это отражает компромисс *bias–variance* в условиях сильно зашумлённой целевой переменной: ограничение глубины снижает дисперсию ансамбля сильнее, чем увеличивает смещение.

(3) Минимальный размер листа $\text{min_samples_leaf} = 20\text{--}50$ предпочтителен; увеличение от 5 до 50 снижает RMSE на 8.1–15.7 % за счёт регуляризации листьев.

(4) Сравнение с XGBoost, LightGBM и HistGradientBoosting не выявляет статистически значимого превосходства одной модели над другой при $K = 22$ складках. ExtraTrees сохраняет практическое преимущество в виде отсутствия необходимости тюнинга learning_rate и более простой интерпретации.

Перспективы дальнейших исследований: (а) обобщение на более длинные горизонты ($k = 10, 20, 60$); (б) исследование байесовской калибровки гиперпараметров (Optuna, Hyperopt); (с) применение к криптовалютам и валютным парам, где волатильность существенно выше; (д) сравнение с современными нейросетевыми архитектурами (Temporal Fusion Transformer, N-BEATS); (е) расширение до условной волатильности с GARCH-подобной структурой.

Вклад авторов

Концептуализация, А.И. и Э.Т.; методология, А.И.; программное обеспечение, А.И.; валидация, А.И. и Э.Т.; формальный анализ, А.И.; исследование, А.И. и Э.Т.; ресурсы, Э.Т.; кураторство данных, А.И.; написание оригинального текста, А.И.; написание и редактирование, А.И. и Э.Т.; визуализация, А.И.; руководство, Э.Т. Все авторы ознакомлены с опубликованной версией рукописи и согласны с ней.

Authors' contribution

Conceptualization, A.I. and E.T.; methodology, A.I.; software, A.I.; validation, A.I. and E.T.; formal analysis, A.I.; investigation, A.I. and E.T.; resources, E.T.; data curation, A.I.; writing—original draft preparation, A.I.; writing—review and editing, A.I. and E.T.; visualization, A.I.; supervision, E.T. All authors have read and agreed to the published version of the manuscript.

Источник финансирования

Исследование не получало внешнего финансирования.

Funding source

This research received no external funding.

Соответствие принципам этики

Не применимо. Исследование основано на анализе финансовых рыночных данных и не включает исследования с участием людей или животных.

Ethics approval

Not applicable. The study is based on financial market data and did not involve humans or animals.

Информированное согласие на публикацию

Не применимо. Исследование не включало участие людей и не использовало идентифицируемые персональные данные.

Consent for publication

Not applicable. The study did not involve human participants or identifiable personal data.

Заявление о доступности данных

Данные, подтверждающие результаты исследования, получены из финансовых источников, использованных авторами, и PostgreSQL-таблицы `features_daily`. Исходный код экспериментов и сводных таблиц указан в рукописи как `jobs/et_ablation.py`. Дополнительные материалы могут быть предоставлены авторами по обоснованному запросу.

Data Availability Statement

The data supporting the reported results were obtained from the financial sources used by the authors and the PostgreSQL table features_daily. The experiment and summary-table code is identified in the manuscript as jobs/et_ablation.py. Additional materials may be provided by the authors upon reasonable request.

Благодарности

Авторы выражают благодарность Ташкентскому международному университету за организационную и техническую поддержку исследования.

Acknowledgments

The authors acknowledge Tashkent International University of Education for organizational and technical support.

Конфликт интересов

Авторы заявляют об отсутствии конфликта интересов.

Conflict of interest

The authors declare no conflicts of interest.

Сокращения

AAPL	Apple Inc
ARCH	Autoregressive Conditional Heteroscedasticity
ExtraTrees	Extremely Randomized Trees
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
HGB	HistGradientBoosting
LGBM	LightGBM
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
TSLA	Tesla Inc.
XGB	XGBoost

Литература

- [1] Andersen T.G., Bollerslev T., Diebold F.X., Labys P. The distribution of realized exchange rate volatility // Journal of the American Statistical Association. — 2001. — Vol. 96(453). — P. 42–55.
- [2] Geurts P., Ernst D., Wehenkel L. Extremely randomized trees // Machine Learning. — 2006. — Vol. 63(1). — P. 3–42.
- [3] Breiman L. Random Forests // Machine Learning. — 2001. — Vol. 45(1). — P. 5–32.
- [4] Chen T., Guestrin C. XGBoost: A scalable tree boosting system // Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — 2016. — P. 785–794.
- [5] Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y. LightGBM: A highly efficient gradient boosting decision tree // Advances in Neural Information Processing Systems (NeurIPS). — 2017. — Vol. 30. — P. 3146–3154.
- [6] Gu S., Kelly B., Xiu D. Empirical asset pricing via machine learning // Review of Financial Studies. — 2020. — Vol. 33(5). — P. 2223–2273.
- [7] Fischer T., Krauss C. Deep learning with LSTM for financial market predictions // European Journal of Operational Research. — 2018. — Vol. 270(2). — P. 654–669.
- [8] Sezer O.B., Gudelek M.U., Ozbayoglu A.M. Financial time series forecasting with deep learning: A systematic literature review // Applied Soft Computing. — 2020. — Vol. 90. — Art. 106181.
- [9] López de Prado M. Advances in Financial Machine Learning. — Hoboken: Wiley, 2018. — 395 p.
- [10] Hamilton J.D. Time Series Analysis. — Princeton University Press, 1994. — 799 p.
- [11] Tsay R.S. Analysis of Financial Time Series. — 3rd ed. — Hoboken: Wiley, 2010. — 720 p.
- [12] Friedman J.H. Greedy function approximation: a gradient boosting machine // Annals of Statistics. — 2001. — Vol. 29(5). — P. 1189–1232.
- [13] Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine learning in Python // Journal of Machine Learning Research. — 2011. — Vol. 12. — P. 2825–2830.
- [14] Efron B., Tibshirani R.J. An Introduction to the Bootstrap. — New York: Chapman & Hall, 1993. — 436 p.

- [15] Engle R.F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation // *Econometrica*. — 1982. — Vol. 50(4). — P. 987–1007.

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельным лицам. Авторы и участники, а Журнал и редакторы. Журнал и редакторы не несут ответственности за любой ущерб, нанесенный людям или имуществу, возникшее в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.

Disclaimer of liability/Publisher's Note: The statements, opinions and data contained in all publications belong exclusively to individuals. The authors and participants, and the Journal and the editors. The journal and the editors are not responsible for any damage caused to people or property resulting from any ideas, methods, instructions or products mentioned in the content.