Прогнозирование доходности акций с использованием гибридной модели BiLSTM, ансамблирования TabM и анализа новостного фона

RMSE_INF

5 октября *2025* г.

Цель проекта и бизнес-ценность

Ключевая цель

Разработать эффективную модель для краткосрочного прогнозирования доходности акций российского рынка *(горизонт 20 дней)*, объединяющую ценовые данные и анализ новостного контекста.

Бизнес-ценность

Модель направлена на повышение точности прогнозирования, что обеспечит принятие более обоснованных инвестиционных решений, позволит автоматизировать торговые стратегии и улучшить управление портфелем.

Ключевые гипотезы

Опережающий сигнал новостей

Анализ тональности новостного фона *(NLP)* предоставит опережающие сигналы, значительно повышающие точность прогнозирования по сравнению с моделями, основанными исключительно на ценовых данных.

Эффективность гибридной модели

Гибридная архитектура *BiLSTM + TabM* превзойдет базовые подходы благодаря способности извлекать сложные временные зависимости и использовать эффективное ансамблирование для повышения точности.

Синергия данных

Комбинация ценовых данных, технических индикаторов и результатов сентимент-анализа обеспечит более робастную и точную модель прогнозирования.

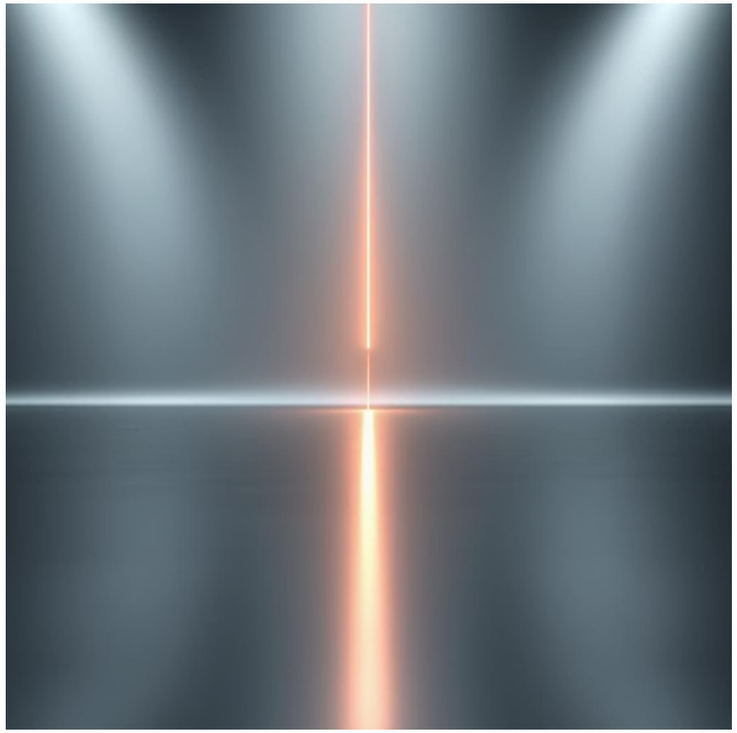
Источники данных и временные параметры

Источники данных

- Ценовые данные (свечи *OHLCV): open, high, low, close, volume.* Период*: 19.06.2020 – 08.09.2025 (19* тикеров*).*
- Новостные данные*:* дата публикации, заголовок, текст

Временные интервалы

Для формирования прогноза модель использует последовательности данных за *60* дней, предшествующих дате прогноза (период *[t-60, t-1]).*



Предотвращение утечки данных

Строгий временной протокол: Защита от утечек данных
Ключевое правило: Информация за день T (цены и новости) используется для прогноза ТОЛЬКО на следующий день T+1.
Симуляция реальности: Модель не "заглядывает в будущее". Она использует сегодняшнюю информацию для прогноза на завтра, как это делает реальный трейдер.

Протокол экспериментов

01

Разделение данных

Данные разделяются на обучающую *(15* тикеров, *~80%)* и валидационную *(4* тикера, *~20%)* выборки. Такое разделение на уровне тикеров позволяет объективно оценить обобщающую способность модели.

02

Оценка Модели

Производительность модели оценивается на независимой валидационной выборке *(4* тикера*).* В качестве основной метрики используется средняя абсолютная ошибка *(MAE).*

Сравнение результатов: Гибридная Модель и Базовый Прогноз

Базовая модель	Прогнозирование на основе средней исторической доходности	*0.014302*	-
Гибридная модель *(BiLSTM-TabM)*	Использование полного набора признаков *(*ценовые данные *+* новостные факторы*)*	*0.013642*	*+4.61%*

Гибридная модель показала значительное улучшение, снизив среднюю абсолютную ошибку *(MAE)* на *4.61%* по сравнению с базовой моделью. Этот результат подтверждает высокую предсказательную способность предложенного подхода.

Эффективность и практическое применение модели

Оптимизация ресурсов

- **Предобработка новостей:** занимает мене *20* минут с использованием *gpt-4o-mini* и около *5* минут *RuBERT*
- **Обучение модели:** требует примерно *15* минут на *GPU (GPU P 100).*
- **Требуемая видеопамять *(RAM):*** *16* ГБ при размере пакета *(batch_size) 128.*

- BERT для новостей

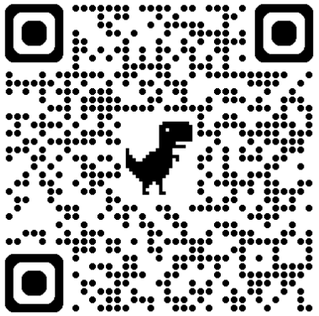
Ключевые сценарии использования

- **Ранжирование активов:** ежедневная оценка акций для выявления потенциальных лидеров и аутсайдеров рынка.
- **Основа для алго-трейдинга:** автоматизация принятия торговых решений и отправки приказов.

Ключевые преимущества и инновации

Оптимизированная гибридная архитектура

Архитектура сочетает *BiLSTM* для эффективного извлечения временных паттернов и *TabM,* имитирующую ансамбль *MLP*-моделей. Это обеспечивает улучшенную регуляризацию и значительно повышает общую производительность.



О TabM

Расширенный Feature Engineering

Мы используем комплексный набор из 85 признаков, включающий ценовые данные, результаты сентимент-анализа (RuBERT), технические индикаторы и свечные паттерны.

1. Базовые ценовые признаки (6 признаков)
   - open, high, low, close: Цены открытия, максимума, минимума и закрытия.
   - volume: Объем торгов.
   - return: Дневная доходность

2. Признаки анализа тональности (6 признаков)
Источник: Модель mxlcw/rubert-tiny2-russian-financial-sentiment (на базе RuBERT), примененная к классифицированным новостям.
   - sentiment_label: Итоговая метка (Негатив/Нейтраль/Позитив).
   - sentiment_score: Непрерывный скор от -1 до +1.
   - positive_prob, negative_prob, neutral_prob: Вероятности каждого класса.
   - confidence: Уверенность модели в своем прогнозе.

3. Технические индикаторы (11 признаков)
Признаки, отражающие динамику и моментум ценового движения.
   - Скользящие средние: SMA и EMA с разными периодами (10, 20, 50).
   - Осцилляторы: RSI (индекс относительной силы) и MACD (с сигнальной линией).
   - Волатильность: ATR (средний истинный диапазон).

4. Свечные паттерны (61 признак)
Источник: Библиотека TA-Lib.
Описание: Огромный набор бинарных признаков

5. Контекстные и временные признаки (2 признака)
   - day_of_week: День недели
   - ticker_code

Оценка эффективности: Снижение ошибки прогнозирования

Базовая модель *(Наивный подход)*	*0.014302*	-
Предложенный подход *(BiLSTM-TabM + Новости)*	*0.013642*	*↓ 4.61%*

Представленный ансамблевый подход с разделением весов демонстрирует значительное снижение средней абсолютной ошибки прогноза. Это улучшение напрямую ведет к формированию более точных и потенциально прибыльных торговых сигналов.

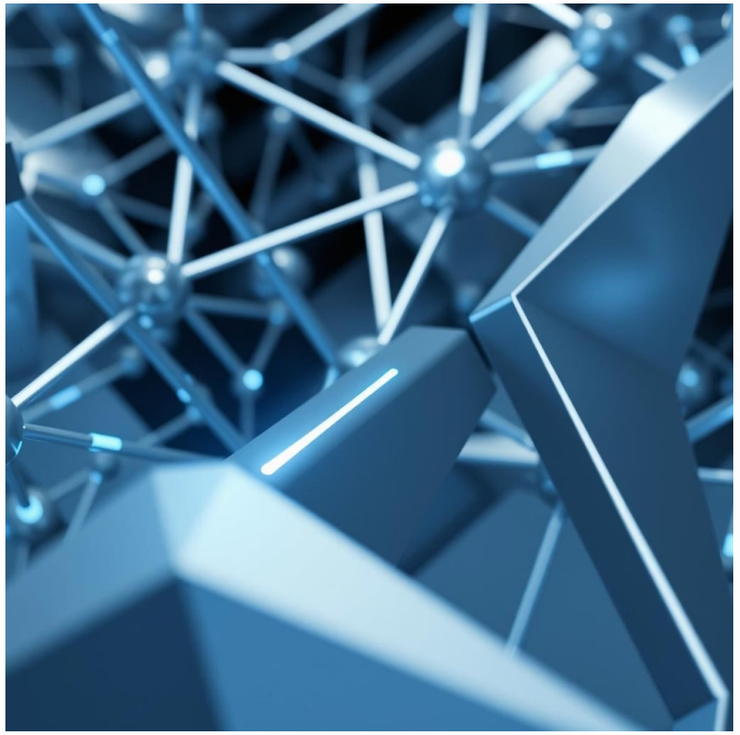
Выводы и ограничения

Основные выводы

- **Влияние новостей:** Интеграция сентимент-анализа новостей существенно повышает точность модели *(около 1.5%).*
- **Эффективное ансамблирование:** *TabM* улучшает качество моделей, работающих с табличными и временными данными.

Дальнейшие шаги

- Внедрение роллинг-валидации для оценки производительности.
- Оптимизация гиперпараметров *TabM* для повышения эффективности.
- Увеличиваем размер *train*
- Перход на *finBERT*



Ограничения текущего подхода

- **Макроэкономические факторы:** Модель не учитывает влияние макроэкономических показателей.
- **Горизонт прогноза:** Прогнозирование ограничено краткосрочным периодом *(до 20 дней).*
- **Качество *NLP*:** Результаты зависят от возможностей модели *mxlcw/rubert-tiny2-russian-financial-sentiment*

- Репозиторий с кодом