Факультет компьютерных технологий

18 августа 2025 г.

Прогнозирование цены и направления движения акций с использованием ML и DL.

Маматов Алмаз
mamatovalmaz@mail.ru

Описание задачи проекта

Проект посвящён созданию и оценке моделей ML и DL для прогнозирования цены и направления движения акций ПАО «Татнефть» на основе исторических котировок с техническими индикаторами, макроэкономических данных (валюты, нефть, индексы), а также новостного фона с эмбеддингами, рассчитанными моделью RuBERT.

Используются как классические алгоритмы ML (CatBoost, XGBoost), так и современные архитектуры DL (LSTM, GRU, CNN, Temporal Fusion Transformer) и их ансамбли.

Целевые признаки сформированы как для задач классификации (рост/падение), так и для регрессии (прогноз цены), что позволяет комплексно оценить потенциал моделей и выбрать лучшие решения или их комбинации.

Ценность задачи

- Финансовая выгода — прогнозирование динамики акций помогает принимать обоснованные решения, снижать риски и повышать доходность.
- Интеграция текстового анализа — учёт новостного фона позволяет получить дополнительную информацию, часто упускаемую алгоритмами, работающими только с числами.
- Развитие ML/DL в финансах — проект демонстрирует возможности современных моделей и ансамблей для работы с временными рядами и текстовыми данными.
- Практическая применимость — решение можно интегрировать в трейдинговые платформы как дополнительного помощника в принятии решений, дополняющего традиционный анализ и опыт трейдера.

Постановка задачи

В рамках проекта необходимо:

- Собрать и объединить данные: котировки, макроэкономические показатели, новости.
- Обработать их: расчёт индикаторов, эмбеддинги новостей (RuBERT), заполнение пропусков, масштабирование и т.д.
- Сформировать таргеты:
  - классификация: `target_5d_dir` (через 5 дней);
  - регрессия: `target_5d` (цена 5 дней).
- Обучить и сравнить модели: CatBoost, XGBoost, LSTM, GRU, CNN, TFT, ансамбли.
- Настроить гиперпараметры (Optuna) для повышения точности.
- Оценить по метрикам (Accuracy, Precision, Recall; RMSE, MAE, MAPE).
- Сформировать выводы и рекомендации для применения.

Часть 1: Сбор и обработка данных

Этап направлен на создание целостного и согласованного набора данных, который объединяет рыночную, макроэкономическую и новостную информацию. Главная цель — сформировать информативный и качественный датафрейм, который можно напрямую использовать для построения и обучения прогнозных моделей.

Блок 1: Подготовка котировок и макроданных

Собраны исторические котировки акций ПАО «Татнефть», а также цены нефти, валютных пар и ключевых фондовых индексов. Источником выступил сервис investing.com, что позволило получить данные за весь доступный период наблюдений. Все ряды приведены к единому временному формату с учётом торговых дней, очищены от дубликатов, пропусков и аномалий. Такая обработка обеспечивает корректность дальнейшего расчёта технических индикаторов и объединение с другими источниками информации.

Блок 2: Расчёт технических и производных признаков

На основе очищенных временных рядов рассчитаны индикаторы тренда (EMA, SMA), волатильности (Bollinger Bands, ATR), торговой активности (OBV, объёмные средние) и других параметров. Цель — отразить в данных не только абсолютные значения цен, но и динамику их изменений, направление движения, уровни перекупленности или перепроданности рынка. Это позволяет моделям улавливать более сложные закономерности, чем при работе только с ценой.

Блок 3: Объединение рыночных данных

Все собранные ряды — акции, нефть, валюты, фондовые индексы — синхронизированы по датам и объединены в единую таблицу. Такой подход даёт возможность моделям анализировать взаимосвязи между инструментами и учитывать, как изменения в одном сегменте (например, курса USD) отражаются на другом (акции компании).

Блок 4: Сбор новостных данных

Чтобы учесть влияние информационного фона на цену акций, собран массив новостей из Telegram-каналов и исторических публикаций с сайта rosinvest.com за период с 2012 по 2025 годы. Отбирались только те материалы, которые имеют отношение к компании, отрасли или макроэкономическим событиям. Это позволяет интегрировать в модель факторы, которые не видны напрямую в ценовых рядах, но могут определять краткосрочную динамику.

Блок 5: Объединение новостных источников

Новости из разных источников сведены в единый хронологический ряд. Такой формат обеспечивает непрерывность новостного потока и упрощает сопоставление его структуры с движением рыночных показателей, что важно для анализа причинно-следственных связей между событиями и изменениями цены.

Блок 6: Преобразование текстов в эмбеддинги

Тексты новостей обработаны моделью RuBERT, которая преобразует их в векторные числовые представления — эмбеддинги. Такой формат позволяет алгоритмам машинного обучения оперировать смысловым содержанием текстов и интегрировать новостной фон в расчёты наравне с числовыми рыночными данными.

Блок 7: Финальное объединение данных

Рыночные, макроэкономические и новостные признаки объединены в единый датафрейм. Все данные синхронизированы по датам, очищены и приведены к единому формату. Итоговый датафрейм содержит как числовые рыночные ряды, так и семантические признаки из новостей, что обеспечивает максимальную полноту входной информации для этапа построения прогнозных моделей.

Часть 2: Выбор и обоснование финальной модели

В этой части решается задача бинарной классификации:
предсказать, будет ли цена акции через 5 торговых дней выше текущей.

Рассмотренные подходы:

- ML: CatBoost, XGBoost, LightGBM, ансамбли (Stacking)
- DL: CNN, GRU, LSTM, Temporal Fusion Transformer

Оценка проводилась на отложенной части выборки при хронологическом сплите.

Блок 1: Подготовка данных и валидации

- Все признаки смещены на $+5$ шагов, чтобы исключить утечки.
- Деление по времени: $\sim$ 80% train, $\sim$ 20% test.
- Валидация: хвост train ($\sim$ 15%).
- Баланс классов учтён через `scale_pos_weight`.
- Классы распределены почти равномерно (около 50/50).

Блок 2: Подбор гиперпараметров (Optuna)

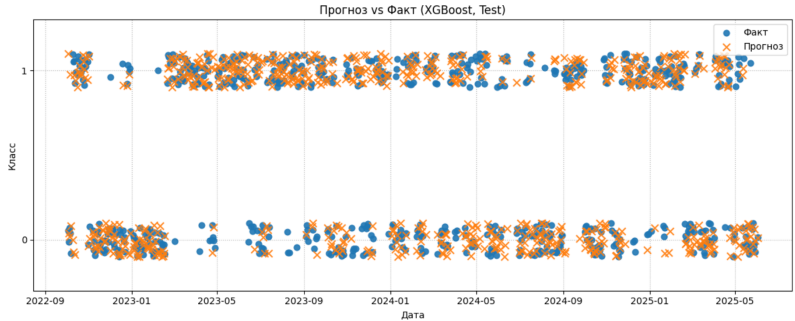
- Целевая функция: минимизация `valid logloss`.
- Использованы параметры: `eta`, `max_depth`, `min_child_weight`, `subsample`, `colsample_bytree`, регуляризации λ, α, γ.
- Ранняя остановка: 75 шагов без улучшения.
- Лучший результат достигнут на 659-й итерации (logloss минимален).

Блок 3: Сравнение моделей (тест)

Модель	**Accuracy**	**Precision**	**Recall**	**F1**
Stacking (XGB+CAT+LGB)	0.7679	0.7729	0.7904	0.7815
CatBoost	0.7872	0.7807	0.8272	0.8033
CNN	0.7720	0.7994	0.7754	0.7872
GRU	0.7406	0.7249	**0.8443**	0.7801
LightGBM	0.7708	0.7851	0.7762	0.7806
LSTM	0.7031	0.7923	0.6168	0.6936
TFT	0.7683	0.7500	0.6364	0.6885
XGBoost	**0.7961**	**0.8000**	0.8159	**0.8079**

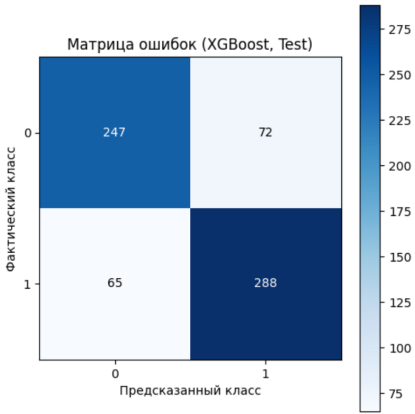
Итог: **XGBoost** показал лучший баланс Accuracy, Precision и F1 при высоком Recall.

Блок 4: Прогноз vs Факт по времени

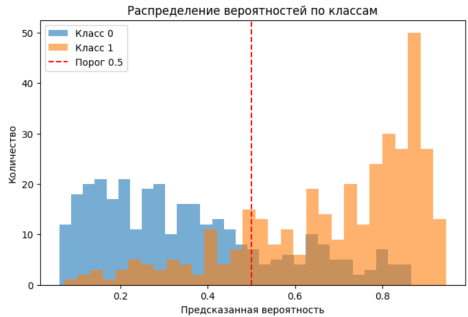


Большинство прогнозов совпадает с фактом; ошибок заметно меньше и они равномерно распределены по времени.

Блок 5: Матрица ошибок и вероятности



Ошибки FP и FN распределены примерно равномерно — подтверждается баланс Precision и Recall.



Класс 0 сосредоточен в зоне 0.1–0.4, класс 1 — в 0.6–0.9; пересечения около порога 0.5 — источник ошибок.

Блок 6: Точечный прогноз (пример)

Дата: **2025-06-02**

- Вероятность роста: **0.1164**
- Прогноз: **падение**
- Факт: **падение**

Прогноз совпал с фактом — модель адекватно реагирует на конкретные даты. Такие проверки удобны для ручной валидации.

Блок 7: Сравнение с исследованиями

В таблице приведены результаты близких исследований по задаче прогноза направления цены.

Работа	Рынок	Гор.	Модель	Результат
Моя (XGBoost)	MOEX (тех+макро+новости)	5д	XGBoost	Acc 0.796, F1 0.808
Sadorsky (2021)	ETF золото/серебро	5–20д	RF, Bagging	Acc 0.80–0.90
Basher & Sadorsky (2022)	Bitcoin, золото	5–20д	RF, Bagging	Acc 0.75–0.80 (5д), >0.90 (15д)
Yıldırım et al. (2021)	EUR/USD	5д	Hybrid LSTM	Acc 0.79–0.84 (часть дат)
Хубиев & Семёнов (2025)	MOEX	5д	LSTM, XGB	Acc 0.45–0.52
ВКР НИУ ВШЭ (2025)	MOEX (15 акц.)	t+1	LSTM+Boruta	Acc 0.58–0.66

Вывод: XGBoost сопоставим с лучшими ансамблями деревьев и превосходит большинство LSTM на горизонте t+5.

Блок 8: Итоговый вывод по части 2

- Финальная модель для бинарной классификации (t+5) — **XGBoost**.
- Достигает Accuracy ≈ 0.80 и F1 ≈ 0.81 на тесте.
- Ошибки FP и FN сбалансированы; распределение вероятностей чётко разделяет классы.
- Модель устойчива: не переобучается, стабильна при хронологическом сплите.
- Простая в эксплуатации: быстро обучается, легко интегрируется в пайплайн, интерпретируема по признакам.
- По сравнению с исследованиями XGBoost демонстрирует результат, сопоставимый с лучшими ансамблями деревьев и выше большинства решений на LSTM.
- Подходит как практическая модель-прототип для применения в реальной торговой системе.

Часть 3: Выбор и обоснование финальной модели

В этой части решается задача регрессии:
предсказать цену акции ПАО «Татнефть» через 5 торговых дней.

Рассмотренные подходы:

- ML: CatBoost, XGBoost
- DL: CNN, GRU, LSTM, Temporal Fusion Transformer (TFT)

Оценка проводилась на отложенной части выборки при хронологическом сплите.
По метрикам (MAE, RMSE, MAPE) финальной выбрана модель **TFT**.

Блок 1: Подготовка данных для регрессии

На данном этапе исходный датафрейм преобразован для решения задачи прогнозирования цены через 5 торговых дней с использованием TFT.

Основные шаги:

- формирование целевой переменной `target_5d`;
- очистка и переименование признаков;
- удаление неполных наблюдений;
- масштабирование входных данных (кроме `date` и `target_5d`);
- сохранение итогового датафрейма в CSV.

Итог: подготовлен нормализованный набор данных, соответствующий входным требованиям TFT.

Блок 2: Формирование календарных признаков

Для адаптации данных к требованиям TFT добавлены календарные и служебные переменные.

Основные шаги:

- генерация индекса времени `time_idx`;
- разложение даты: год, квартал, месяц, день, день недели;
- индикаторы: выходной, начало/конец месяца, день года;
- задание идентификатора временного ряда `group_id`;
- финальное упорядочивание и сохранение в CSV.

Итог: получен структурированный датафрейм, полностью совместимый с входным форматом TFT.

Блок 3: Финальная подготовка данных для TFT

На завершающем этапе данные были приведены к формату, требуемому библиотекой `pytorch-forecasting`. Выполнено разделение признаков на статические, временные известные заранее и временные неизвестные. Для обеспечения корректного обучения реализован хронологический сплит на обучающую и валидационную выборки, после чего данные преобразованы в объект `TimeSeriesDataSet`, автоматически управляющий нормализацией и структурой входных последовательностей.

Блок 4: Инициализация и обучение TFT

На данном этапе выполняется построение и обучение модели **Temporal Fusion Transformer** на подготовленных данных. Архитектура и параметры модели задаются автоматически из структуры датафрейма, с последующей настройкой ключевых гиперпараметров (скорость обучения, размер скрытых слоёв, число голов внимания, dropout и др.).
Для обеспечения устойчивости обучения применяются:

- **RMSE** как функция потерь и **MAE** как дополнительная метрика;
- колбэки ранней остановки и сохранения наилучшей версии модели;
- контроль градиентов и ограничение числа эпох.

В результате модель обучается на `train_loader` с контролем качества на `val_loader`, автоматически сохраняя оптимальный чекпойнт по метрике *val_loss*, который далее используется для тестирования и прогнозов.

Блок 5: Скользящее тестирование TFT

Для оценки прогноза использовано **скользящее тестирование** с окном длиной 90 дней: на каждом шаге модель обучается на последних наблюдениях и предсказывает цену акции через 5 торговых дней.
Особенности реализации:

- загрузка обученной TFT из сохранённого чекпойнта;
- формирование скользящих окон на основе полного ДФ;
- прогноз только на **будущем периоде** (2025 год и далее);
- сбор и анализ ошибок (MAE, RMSE, MAPE), формирование итогового датафрейма.

Качество прогноза (ручной подбор vs Optuna)

Подход	MAE	RMSE	MAPE
Ручной подбор	18.25	22.37	2.63%
Optuna	23.99	29.02	3.41%

Вывод: ручная настройка обеспечивает более точные прогнозы и меньший уровень ошибок, тогда как автоматическая оптимизация через Optuna приводит к излишнему сглаживанию динамики и росту ошибок что показано далее на графиках.

Прогноз vs Факт и распределение ошибок

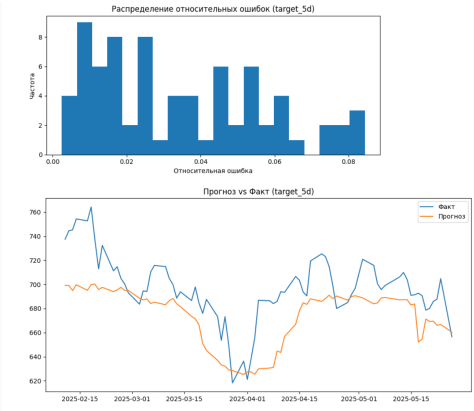


Рис. 3 – Прогноз vs Факт (Optuna)

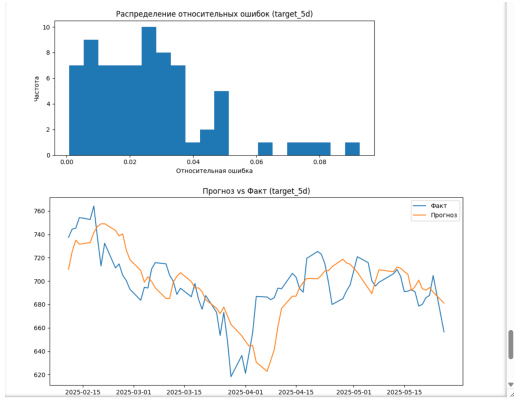


Рис. 4 – Прогноз vs Факт (ручной подбор)

Вывод: при ручной настройке модель лучше отслеживает динамику, ошибки локализованы и меньше по абсолютной величине.

Итоговый вывод по скользящему тестированию TFT

- **Ручная настройка гиперпараметров** обеспечивает более низкие ошибки (MAE ≈ 18, RMSE ≈ 22, MAPE ≈ 2.6%), что делает модель применимой на горизонте **t+5**.
- Автоматическая оптимизация через Optuna приводит к **сглаживанию динамики** и росту ошибок (MAPE > 3.4%).
- Графический и табличный анализ подтверждает: при ручной настройке модель ближе следует за фактической ценовой динамикой и допускает меньше крупных отклонений.

Заключение: модель TFT с ручным подбором гиперпараметров является оптимальным вариантом для прогноза цен акций ПАО «Татнефть» на горизонте 5 торговых дней.

Блок 6: Точечный прогноз TFT на заданную дату

На примере прогноза цены акции ПАО «Татнефть» на дату **06.06.2025** проверена работа обученной модели. Модель использует последние 90 наблюдений, календарные признаки формируются автоматически.

Фактическая цена: 639.10 ₽

- **Ручная настройка гиперпараметров:** прогноз 676.85 ₽, ошибка 37.75 ₽ (5.91%).
- **Optuna (автопоиск):** прогноз 592.70 ₽, ошибка 46.40 ₽ (7.26%).

Вывод: обе модели ошиблись больше, чем в среднем по скользящему тестированию, однако при ручной настройке ошибка ниже. Точечные проверки удобны для валидации и позволяют убедиться в устойчивости модели на конкретных датах.

Сравнение с исследованиями

Цель: определить оптимальную модель для прогноза цен акций на горизонте $t + 5$.

Работа	**Рынок**	**Гор.**	**Модель**	**Метрики**	**Результат**
Моя модель (TFT, ручн.)	MOEX (акции, индикаторы, макро, новости)	**t+5**	TFT	MAE / RMSE / MAPE	**18.25 / 22.37 / 2.63%**
Lim & Zohren (2021)	S&P 500	1–20	TFT, LSTM	RMSE / MAPE	TFT < LSTM, MAPE ∼3–5%
Sezer et al. (2020)	NYSE, MOEX, FOREX	1–15	CNN, LSTM, GRU	RMSE	CNN/GRU: 30–40; гибриды <30
Fischer & Krauss (2018)	S&P 500	t+1	LSTM vs RF	RMSE / R^2	LSTM RMSE ∼25; RF лучше
Díaz Berenguer et al. (2024)	NYSE	1–5	TFT + новости	RMSE / MAPE	TFT + FinBERT > LSTM, DeepAR

Краткий анализ исследований

Lim & Zohren (2021) Показано, что TFT устойчивее LSTM на горизонтах > 5 дней и обладает преимуществом интерпретируемости.

Sezer et al. (2020) Обзор 150+ работ и эксперименты на NYSE, MOEX, FOREX. Гибриды CNN+LSTM показали RMSE < 30, лучше отдельных моделей.

Fischer & Krauss (2018) Сравнение LSTM и RF на S&P 500 ($t + 1$). LSTM RMSE ~ 25, но ансамбли деревьев оказались стабильнее.

Díaz Berenguer et al. (2024) TFT в сочетании с новостями (FinBERT) превзошёл LSTM и DeepAR. Подтверждена ценность учёта новостного фона и причинности.

Итоговый вывод

Моя модель *Temporal Fusion Transformer* **(ручная настройка)** на горизонте $t + 5$ показала метрики: **MAE $\approx$18, RMSE $\approx$22, MAPE $\approx$2.6%**.

В сравнении с исследованиями:

- TFT подтверждает устойчивость и интерпретируемость (Lim & Zohren, Díaz Berenguer et al.);
- Ошибки моей модели (RMSE $\sim$22) меньше, чем у большинства классических LSTM/GRU (RMSE 25–40, MAPE 3–5%);
- Результаты сопоставимы или лучше зарубежных исследований, что подтверждает конкурентоспособность подхода для MOEX.

Заключение: финальная модель **TFT с ручной настройкой гиперпараметров** является оптимальным выбором для прогноза цен акций на горизонте 5 торговых дней.

Направления улучшения модели (1/4)

Работа с признаками (feature engineering)

- Более строгая фильтрация слабозначимых признаков позволит снизить уровень шума и уменьшить риск переобучения.
- Увеличение числа информативных признаков:
  - **Технические индикаторы**: Chaikin Money Flow (CMF), Money Flow Index (MFI), Williams %R, Momentum, Keltner Channels, Ichimoku Cloud.
  - **Макроэкономические ряды**: индекс потребительских цен (CPI), индекс деловой активности (PMI), торговый баланс, уровень безработицы, индекс доверия инвесторов.
  - **Кросс-рыночные факторы**: динамика фондовых индексов (S&P 500, RTS, MOEX), золото, газ, акции конкурентов (Лукойл, Роснефть).
- Важным шагом остаётся оценка важности признаков (например, SHAP, permutation importance), чтобы выявить реально значимые переменные.

Направления улучшения модели (2/4)

Оптимизация гиперпараметров и архитектуры

- Качество временных моделей сильно зависит от параметров: `encoder_length`, `hidden_size`, `batch_size`, число слоёв.
- Их точная настройка может заметно повысить результаты.
- Проблема: ограничение видеопамяти (ошибки OOM), что требует осторожного выбора размеров модели и параметров обучения.

Направления улучшения модели (3/4)

Использование агрегированных эмбеддингов новостей

- Вместо простого усреднения эмбеддингов можно применять агрегирование с учётом важности слов (attention) или по тональности.
- Такой подход позволяет отражать именно «финансовый сигнал» новости и усиливает вклад текстовых данных в прогноз.

Направления улучшения модели (4/4)

Изменение логики применения модели (бинарная классификация)

- Повышение порога классификации (например, до 0.8) позволит выдавать сигналы только при высокой уверенности.
- Зона неопределённости (0.4–0.6) может трактоваться как «нет сигнала».
- Такой подход ближе к реальной торговой логике, где надёжность прогноза важнее частоты сигналов.

План развития проекта

1. **Дальнейшее улучшение качества**
   - Доработка признаков, гиперпараметров и архитектуры моделей для повышения точности и устойчивости прогнозов.

2. **Расширение охвата прогнозов**
   - Прогноз не только на 5 дней, но также на **1 день** и **30 дней**.
   - Подключение большего числа акций Московской биржи (например, Лукойл, Роснефть) для проверки переносимости модели.

3. **Вывод модели в продакшн**
   - Реализация автоматического пайплайна: загрузка котировок и новостей, предобработка, генерация признаков.
   - Регулярное обновление модели или её перетренировка по мере накопления новых данных.
   - Автоматическая генерация торговых сигналов и интеграция в интерфейс (дашборд с визуализацией и метриками).
   - Возможность backtesting и сравнения прогнозов с фактическими результатами в реальном времени.