ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

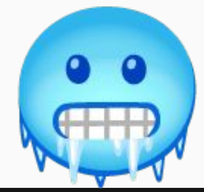
Предсказание цен акций на 30 дней по топ компаниям MOEX

17.12.2025

Команда

Ступников Дмитрий

Садоян Давид



Савин Владислав

Клюшова Ульяна

Описание и цель проекта

Данные: 2 датасета с ценами 19 компаний и новостям с finam.ru за 5 лет

Бизнес-ценность: обоснованные инвестиционные решения, автоматическая агрегированная обработка новостей и цен вместе для управления портфелем

Цель

Разработать эффективную модель для краткосрочного прогнозирования цен акций российского рынка (горизонт - 30 дней), объединяющую ценовые данные и анализ новостного контекста

Если кто захочет заработать - принимаем депозиты с комиссией 5% от прибыли
писать @Dmitriy_000

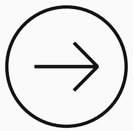
Ключевая гипотеза

Синергия ценовых данных, технических индикаторов и опережающего сигнала новостей

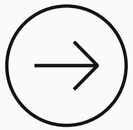
Более точные и **надёжные** предсказания цен по сравнению с расчётом лишь из цен и индикаторов

Данные

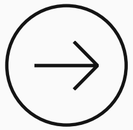
Рыночные данные



Источник: moex.com



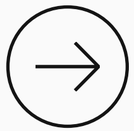
Тикеры: 19 публичных компаний (верхушка moex рейтинга)



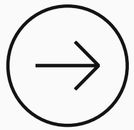
Признаки:

- `open`, `close`, `high`, `low` — цены открытия, закрытия, дневной максимум и минимум
- `volume` — объём торгов
- `date` — торговый день
- `ticker` — идентификатор компании

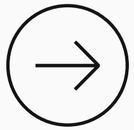
Новостные данные



Источник: finam.ru

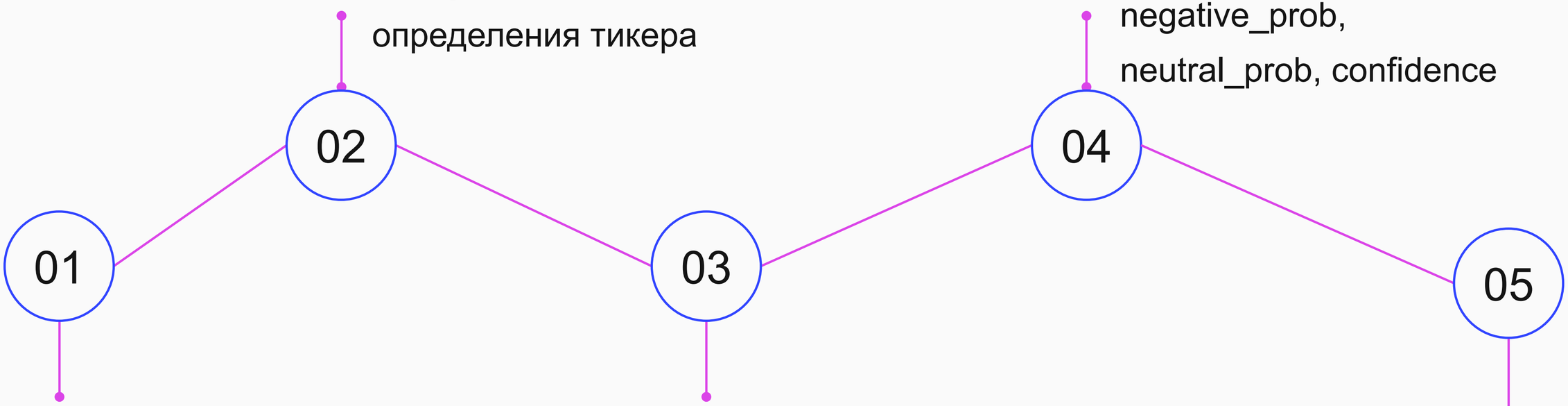


Признаки: `date,title, publication`



Задача: связать новости с конкретными компаниями и извлечь тональность

Обработка новостей



date, title, publication

Прогнали каждую новость через Chatgpt-4o-mini для определения тикера

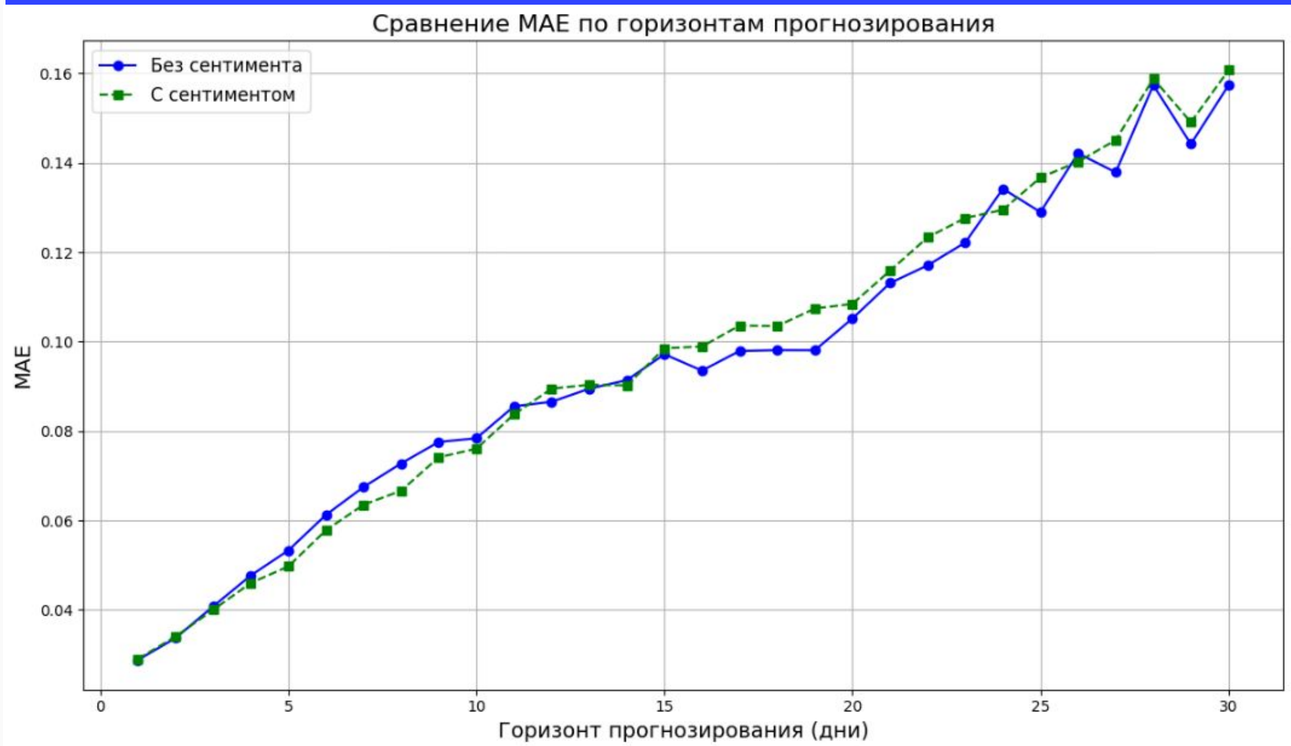
Прогнали новости через rubert-tiny2-russian-financial-sentiment для определения тональности каждой

date, title, publication, ticker, sentiment_score, positive_prob, negative_prob, neutral_prob, confidence

Объединили с ценами, получили цены за день + новость за день

Бейзлайн

- В качестве первой модели был выбран catboost
- Результаты 1 дня: MAE = 0.017
- Результаты 30 дня: MAE = 0.1574
- Пока прогноз на более долгий срок уж слишком плохой из-за сложных временных зависимостей
- На бейзлайне гипотеза себя не оправдала: новости не особо влияют



Модели

- Catboost
- Random Forest
- Ridge

Feature engineering

88 признаков: ценовые данные, результаты новостного анализа, технические индикаторы и свечные паттерны

1. Базовые ценовые признаки (5 штук):
   open, high, low, close, volume

2. Признаки анализа тональности (6 штук):
   sentiment score, sentiment label,
   positive/negative/neutral prob, confidence

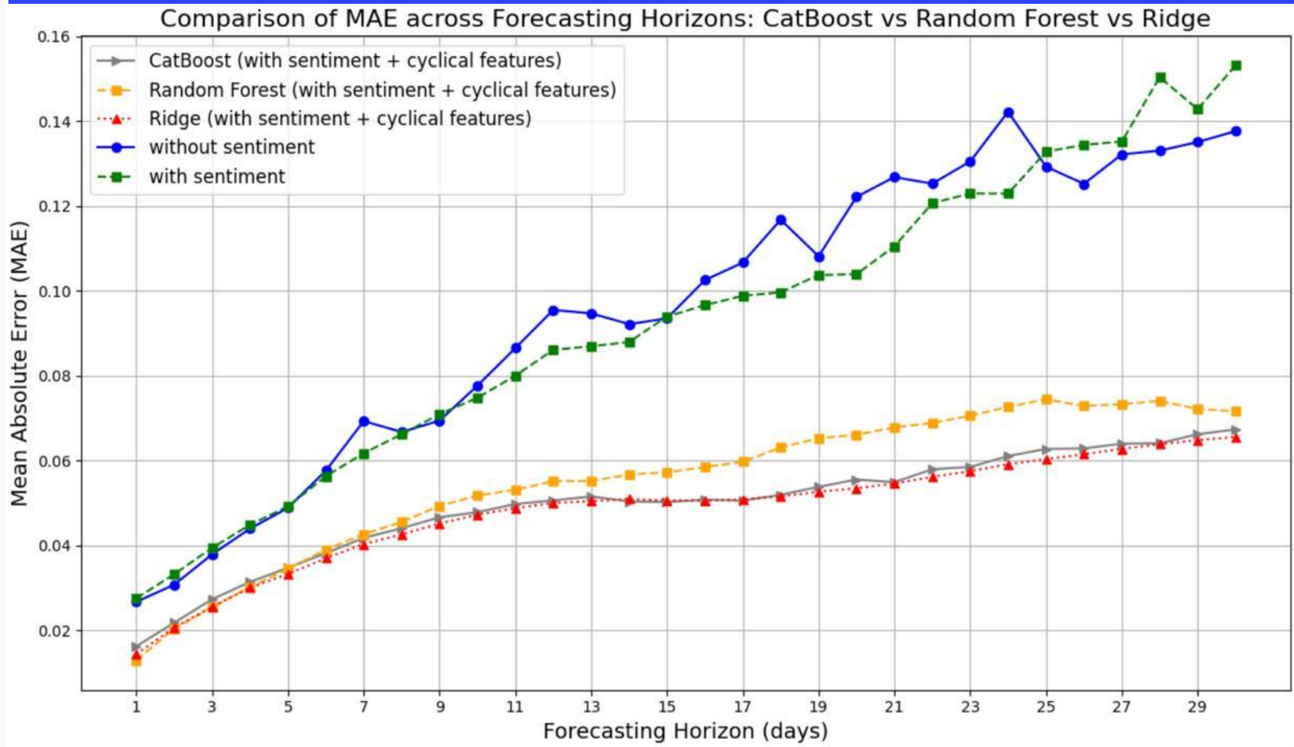
3. Технические индикаторы (11 штук):
   SMA, EMA, RSI, MACD, ATR

4. Свечные паттерны (60 штук):
   Технические паттерны - бинарные признаки

5. Временные признаки (6 штук):
   month, weekday, month_sin, month_cos,
   weekday_sin, weekday_cos

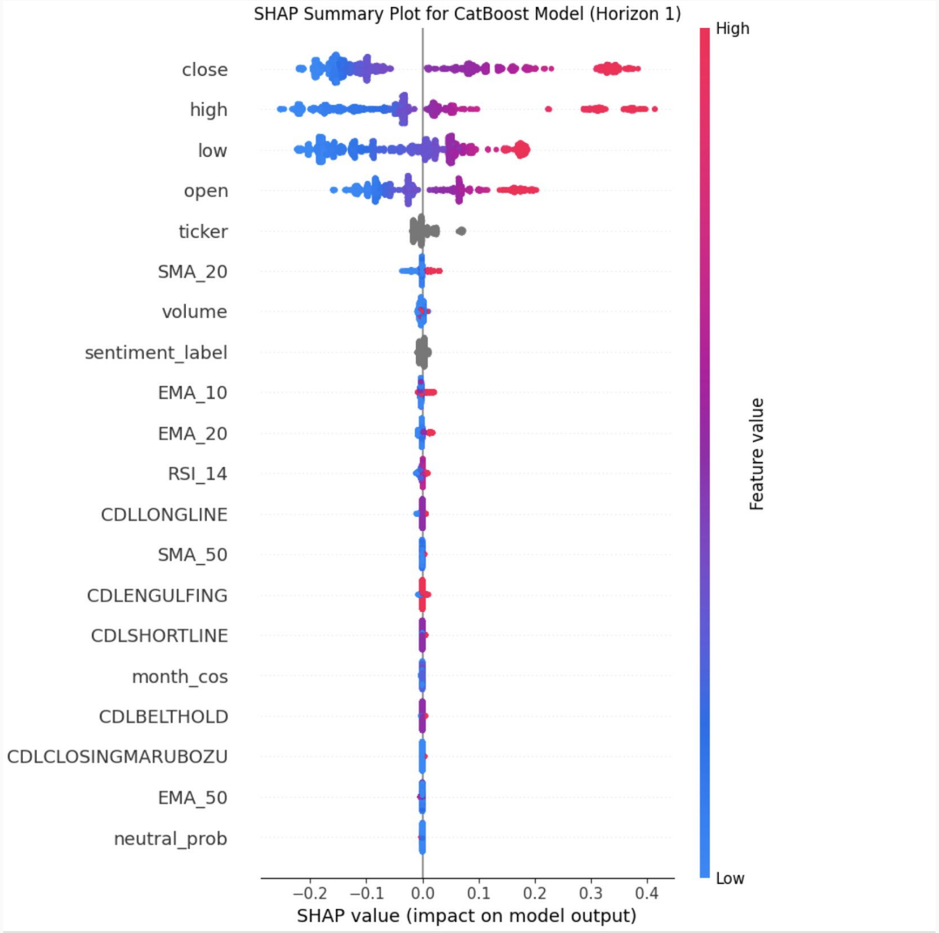
Новые успехи!

- Сравнивается catboost, random forest и ridge
- Результаты, особенно на долгосрок сильно улучшились
- Новые фичи себя отлично показали, разница в более чем 2 раза
- Random forest остаёт, а вот CatBoost и Ridge идут абсолютно вровень
- Формально - мы ошибаемся лишь на 6% в цене через 30 дней, что неплохой результат

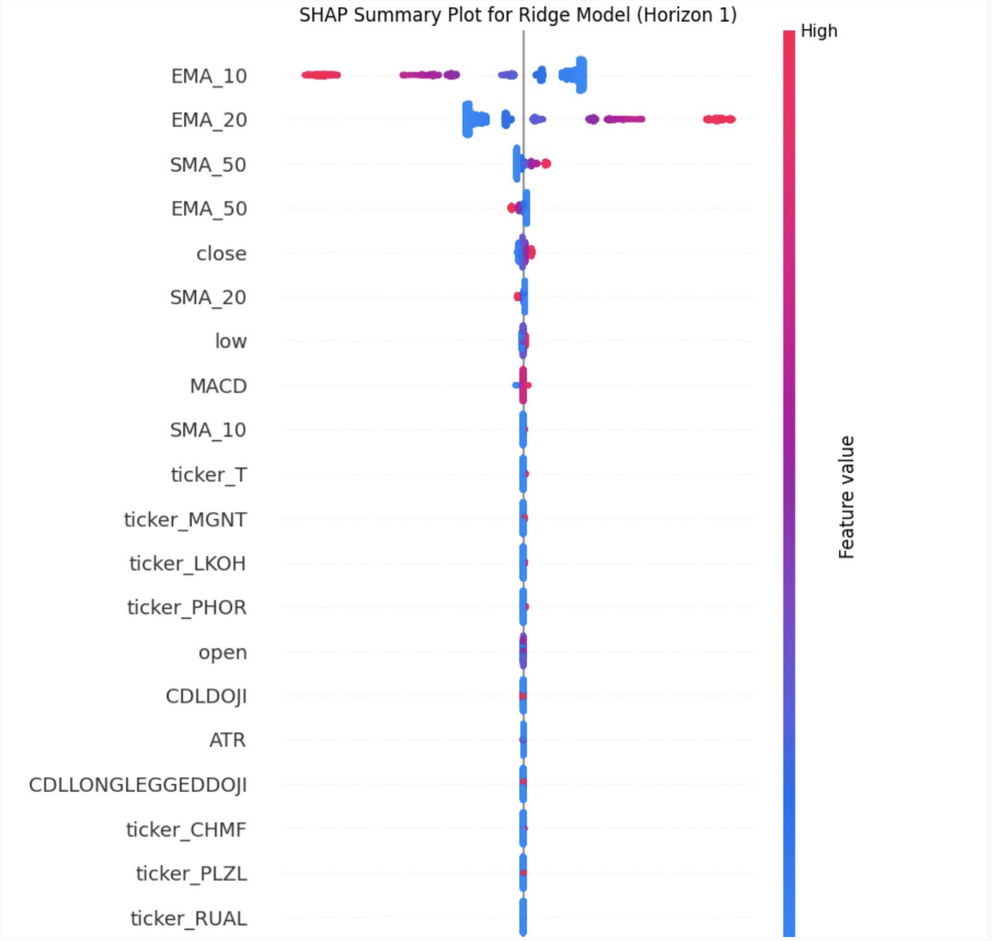


Интерпретация

Catboost

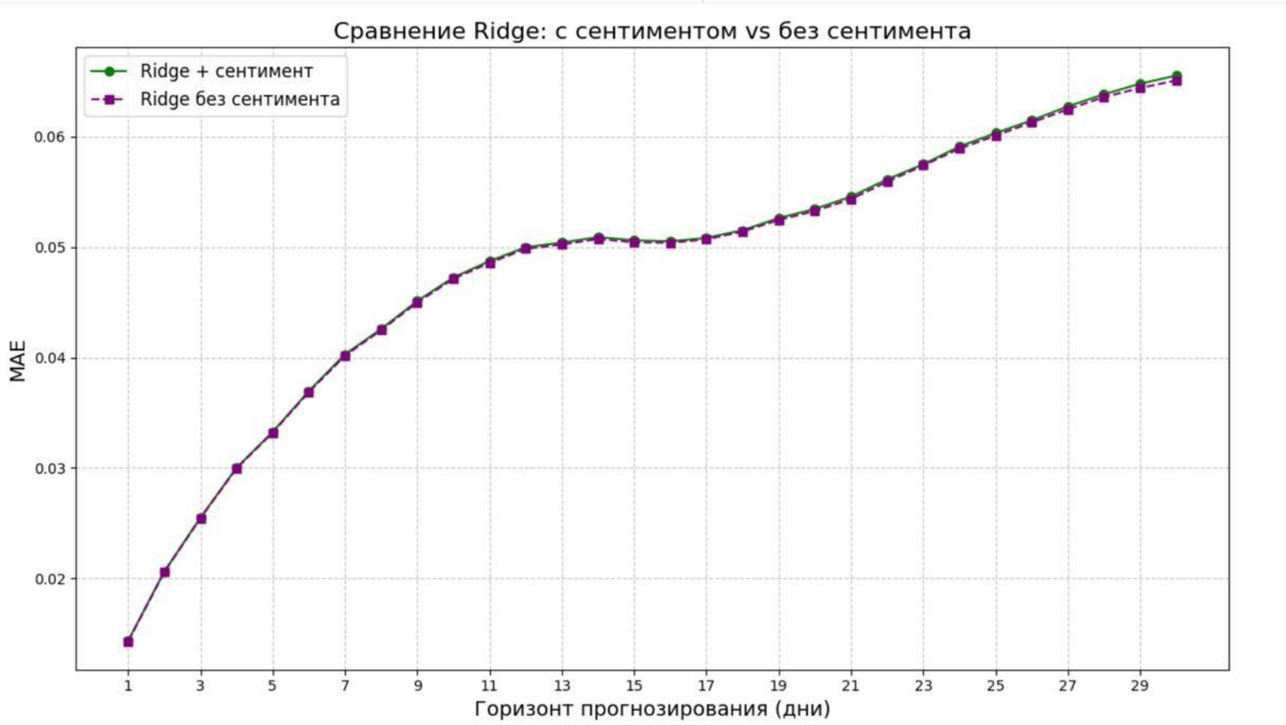


Ridge



Влияние новостей

- Итого: новости не влияют на прогноз цены с 1 до 30 дней
- Из-за ликвидных акций
- Из-за концентрации российского рынка на торговле “На новостях”
- Такая обработка новостей не учитывает их накопительный эффект и “сюжеты”
- Большая часть рынка торгует по индикаторам =)



Выводы и ограничения

Основные выводы

Влияние новостей
- Интеграция сентимент-анализа не оказала влияния

Технические индикаторы спасают catboost
- Из-за невозможности анализа глубоких временных паттернов, индикаторы существенно улучшают работу модели

Линейные модели тоже могут пободаться с ансамблями
- При этом ridge гораздо более интерпретируем и быстр

Гипотеза опровергнута

Такой вид анализа новостей не помогает предсказывать цены, что говорит о достаточно эффективном рынке на ликвидных акциях. Новости могут быть применимы только в очень краткосрочной торговле (10-15 минут)

Ограничения текущего и что можно сделать еще

Макроэкономические и корпоративные факторы
- Модели не учитывают общую экономическую ситуацию в стране и ситуацию в компании

Качество сентимент-анализа
- Также результаты зависят от возможностей NLP и ограниченности одного источника новостей

Возможные нововведения

Улучшенная обработка новостей со всех источников и сбор макроэкономических факторов + подсчёт мультипликаторов