

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ  
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ НАЦИОНАЛЬНЫЙ  
ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ  
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук

Триголос Алексей Павлович

**Применение методов машинного обучения и анализа текстовых данных для  
предсказания стоимости акций на российском рынке**

Выпускная квалификационная работа по направлению подготовки  
01.04.02 Прикладная математика и информатика образовательная программа  
магистратуры «Машинное обучение и высоконагруженные системы»

Научный руководитель:  
Каюмов Руслан Асхатович

Рецензент:  
Паточенко Евгений Анатольевич

Москва 2025

# Оглавление

<b>Оглавление</b>	2
<b>Аннотация</b>	4
<b>1 Введение</b>	5
<b>2 Основа работы</b>	5
2.1 Актуальность исследования	5
2.2 Объект и предмет исследования	6
2.3 Цель и задачи	6
2.4 Научная новизна	6
2.5 Методология и структура работы	7
<b>3 Обзор существующих решений</b>	7
3.1 Использование линейной регрессии	7
3.2 Современные методы машинного обучения	7
3.3 Новостной фон	8
3.4 Результаты исследований	9
<b>4 Сбор, обработка и анализ данных</b>	9
4.1 Тикеры эмитентов	9
4.2 Цены акций	9
4.3 Новости по акциям	18
<b>5 Модели</b>	22
5.1 Метрики	22
5.2 По ценам	24
5.3 С использованием новостей	25
5.4 SARIMAX	25
5.5 Ridge Regression	27
5.6 Random Forest Regression	29
5.7 XGBRegressor	30
5.8 LSTM	34
5.9 Результаты исследований	34

<b>6 Сервис . . . . .</b>	<b>37</b>
6.1 Архитектура . . . . .	37
6.2 Запросы API . . . . .	38
6.3 Общая структура сервиса . . . . .	39
6.4 Функционал . . . . .	40
<b>7 Как можно улучшить результат . . . . .</b>	<b>44</b>
<b>8 Заключение . . . . .</b>	<b>45</b>
<b>Список литературы . . . . .</b>	<b>47</b>

## **Аннотация**

Данная выпускная квалификационная работа посвящена разработке комплексного подхода к прогнозированию стоимости акций российских компаний с применением современных методов машинного обучения и обработки естественного языка. Основное внимание было уделено поиску наилучшей модели для получения наиболее качественных прогнозов и интеграции текстовых данных из новостных источников в архитектуру прогнозных моделей. В исследовании проанализированы акции 215 компаний, котирующихся на Московской бирже за период с 2010 по 2024 год с использованием более миллиона новостных статей «РИА Новости». Реализованы и сравнены модели Ridge Regression, Random Forest, XGBoost, SARIMAX и LSTM с гибридными подходами, демонстрирующие ошибку менее 3% по MAPE на однодневном горизонте прогнозирования. Особенностью работы стало создание автоматизированного веб-сервиса на базе FastAPI и Streamlit с интеграцией облачного хранилища данных для предоставления результатов работы в удобном виде.

# 1. Введение

По данным «Интерфакс» [3] к концу 2024 года число физических лиц, имеющих счета на Московской Бирже, превысило 35 миллионов человек, что составляет более 46% экономически активного населения страны. Еще в 2018 году данный показатель составлял менее 2%. Таким образом, можно говорить о том, что рост инвестирования в стране за 7 лет увеличился более чем в 23 раза. Такую тенденцию роста частных инвестиций граждан можно наблюдать и в остальном мире.

По своей сути человек - такое животное, которое хочет получить всё и сразу. Большинство людей не может и не хочет разумно оценивать свои шансы. Нам недостаточно получать хорошо, средне или как все. Мы хотим жить лучше, иметь больше перспектив, надеемся, что нам повезет. Но это может обернуться невероятным крахом.

Учитывая это и число людей пришедших на биржу, можно ожидать, что многие попробуют обыграть рынок или полезут в еще большие риски, что я не считаю правильным. Никто не может предсказать как поведет себя бумага, какие будут новости, как отреагирует толпа, поэтому и невозможно предугадать цену той или иной акции, как и предсказать погоду. Но мы же умеем прогнозировать погоду на ближайшие дни, так может, у нас есть шанс предугадать что произойдет с конкретными бумагами в ближайшее время?

В данной работе поставлена задача, на основе собранных данных по акциям компаний и их новостному фону, научиться прогнозировать цены акций различных эмитентов, выяснить, какие модели лучше подходят для получения наиболее приближенных к реальности результатов и определить, насколько целесообразно использовать новостные данные для повышения точности предсказания.

# 2. Основа работы

## 2.1. Актуальность исследования

Современные финансовые рынки характеризуются высокой волатильностью, где цены акций реагируют на макроэкономические события, корпоративные новости и рыночные спекуляции. Российский рынок ценных бумаг, несмотря на геополитические вызовы, сохраняет привлекательность для инвесторов, что требует разработки точных инструментов прогнозирования. Традиционные экономико-математические модели часто не учитывают качественные факторы, тогда как машинное обучение позволяет интегрировать текстовые данные

новостей для улучшения предсказательной способности.

## **2.2. Объект и предмет исследования**

Объектом исследования выступают акции компаний Московской биржи из секторов энергетики, телекоммуникаций, финансовых и других. Предмет исследования – методы машинного обучения для временных рядов с интеграцией семантического анализа новостных текстов.

## **2.3. Цель и задачи**

Цель работы – создание гибридной модели прогнозирования стоимости акций, комбинирующей технические индикаторы и семантические значения новостей. Для её достижения решались следующие задачи:

1. Сбор и предобработка исторических данных по 330 тикерам.
2. Сбор, обработка новостей и оценка их эмоциональной окраски с использованием большой языковой модели.
3. Реализация различных типов моделей: от линейных и скользящих средних до ансамблей и нейросетей.
4. Создание веб-интерфейса для визуализации прогнозов с архитектурой микросервисов на Docker.

## **2.4. Научная новизна**

1. Использование LLM-моделей для русского языка в контексте финансовой аналитики.
2. Разработка подхода для использования новостного контента и трендовой составляющей для обучения моделей прогнозирования.
3. Использование NLP-фич для российского рынка и выявление лучших результатов в предсказании цен акций множества эмитентов.

## **2.5. Методология и структура работы**

Исследование основано на принципах CRISP-DM с использованием Python-стека (Pandas, Scikit-learn, TensorFlow). Работа проходила по следующим этапам: изучение предыдущего опыта, сбор, обработка и анализ полученных данных, построение моделей и анализ результатов, реализация сервиса.

## **3. Обзор существующих решений**

### **3.1. Использование линейной регрессии**

Статья [12] посвящена применению классического метода линейной регрессии для прогнозирования цен акций. Основная цель - проверить гипотезу о том, что линейная регрессия способна с высокой точностью предсказать цену закрытия акций на основе исторических данных и финансовых показателей компаний. Теоретическая база статьи строится на использовании регрессионного анализа, который позволяет моделировать зависимость цены акции от различных факторов, учитывая многолетние данные, собранные с помощью библиотеки yfinance.

В исследовании подчеркивается актуальность точного прогнозирования в условиях растущей популярности биржевых торгов, что важно для принятия инвестиционных решений. Авторы достигли высокой точности модели - до 96% в прогнозировании, что подтверждает эффективность линейной регрессии в данной задаче. Однако отмечается, что для повышения точности необходимо учитывать дополнительные макроэкономические факторы и события, что является ограничением текущей модели.

Практическая значимость заключается в возможности использовать предложенный подход для поддержки решений инвесторов на российском рынке, адаптируя методику к специфике локальных данных.

### **3.2. Современные методы машинного обучения**

В работе [11] проведено сравнение классической статистической модели ARIMA и современных методов машинного обучения - моделей SVR (Support Vector Regression) и LSTM (Long Short-Term Memory) - для задачи краткосрочного прогнозирования стоимости акций российских компаний, таких как «Газпром», «Сбербанк», «Яндекс» и др.

Авторы выявили, что ARIMA не подходит для краткосрочного прогнозирования из-за низкой чувствительности к резким изменениям цен, тогда как модели машинного обучения, особенно LSTM, показывают высокую точность и адекватно реагируют на волатильность рынка. Модель LSTM, состоящая из двух рекуррентных слоев и полно связанных слоев с функцией активации ReLU, показала лучшие результаты по метрикам RMSE и R<sup>2</sup>, достигая точности предсказания в диапазоне от 70% до 96% для большинства компаний. При этом для «Сбербанка» точность была ниже из-за устойчивого восходящего тренда без значительной волатильности, что усложнило моделирование.

Для LSTM использовалась специальная подготовка данных с учетом временных шагов. Результаты свидетельствуют о перспективности применения глубоких рекуррентных нейронных сетей для краткосрочного прогнозирования на российском рынке акций. Кроме того, автор рекомендует дальнейшее улучшение моделей путем интеграции технических индикаторов и расширения выборки.

### 3.3. Новостной фон

Исследование [5] посвящено интеграции анализа новостных текстов и исторических данных о ценах акций для предсказания падений стоимости акций на бразильском фондовом рынке. Авторы рассматривают влияние новостных публикаций на поведение инвесторов и динамику цен, используя методы машинного обучения и текстового майнинга.

В работе анализируется широкий набор из 64 ценных бумаг, что значительно расширяет охват по секторам экономики, в отличие от многих исследований, ограниченных несколькими активами. Важным аспектом является применение 11 различных алгоритмов машинного обучения для классификации и прогнозирования, что позволяет выявить наиболее эффективные модели для предсказания финансовых потерь. В частности, исследование демонстрирует, что модели, основанные на анализе текстовых данных новостей (на португальском языке), превосходят традиционные стратегии "купить и держать" и скользящих средних в точности предсказания падений цен.

Это подтверждает гипотезу о том, что текстовый анализ новостей является ценным источником информации для прогнозирования краткосрочных изменений рынка, что актуально и для российского рынка акций, где новостной фон также влияет на волатильность. Авторы также подчеркивают важность временных окон публикаций и возврата, выявляя оптимальные горизонты для использования новостей в торговых решениях.

## 3.4. Результаты исследований

Учитывая анализ статей и работ, описанных выше, а также других исследований, что остались за кадром, я склоняюсь к построению нескольких видов моделей на большом объёме акций компаний. Расскажу вам про созданные мною более простые модели, такие как множественная линейная регрессия или скользящее среднее, так и про более сложные ансамбли и рекуррентные нейронные сети.

Кроме этого, я покажу какие результаты были мною получены в каждой из описанных моделей и продемонстрирую как они изменяются при добавлении новостных данных, которые предварительно были оценены большой языковой моделью.

# 4. Сбор, обработка и анализ данных

## 4.1. Тикеры эмитентов

Я стремлюсь понять, какая модель лучше приспособлена для получения более точного предсказания цен акций, поэтому посчитал необходимым получить как можно больше данных по различным бумагам. Таким образом, первостепенной задачей было получить наибольшее число тикеров компаний.

Для этого я использовал Python-библиотеку "beautifulsoup4". Изначально со страницы [1] я получал ссылки на каждую компанию. Это делалось по той причине, что на данной странице сайта описаны только названия в алфавитном порядке, но не указаны их тикеры. Обработав данные со страницы я получил 366 ссылок для получения более подробной информации.

Затем я уже обрабатывал каждую из страниц отдельно для извлечения тикера компании. К сожалению, не все данные оказались валидными, но данный этап сильно облегчил получение большого объема тикеров со значительным процентом правильности. А для хранения всех данных по данной работе я пользовался облачным хранилищем S3 от Яндекса.

## 4.2. Цены акций

Получив тикеры компаний мне было необходимо найти данные об их ценах за максимальный доступный период. Для этого я воспользовался ещё одной библиотекой "arimoex". Как раз с её помощью я и получал данные о стоимости той или иной бумаги на всём доступ-

ном промежутке времени.

Но не всё получилось так просто. Во-первых, не все полученные на прошлом этапе тикеры были валидны, из-за чего мне удалось получить данные только по 332 компаниям, тем не менее это большое число и составляет 90% от изначальных 366. А вторым неприятным моментом получилось то, что некоторые компании, которые торгуются долго на рынке, имеют короткую историю. Я предполагаю, что это связано с тем, что они базировались в других странах по каким-либо причинам, но из-за ситуаций произошедших в 2022 году, им пришлось редомицилироваться, то есть переехать из иной страны в Россию. Или возможен вариант смены названия или слияния с другой компанией, что также могло повлиять на историю.

Для каждой из компаний я получал множество строк как показано на «Рисунке 1». Здесь «TRADEDATE» хранит информацию о том, для какой даты актуальны значения в данной строке. Колонки «OPEN», «LOW», «HIGH», «CLOSE» показывают цены открытия, минимальную, максимальную и закрытия для соответствующего дня. В «VOLUME» указано, какое число бумаг было продано на бирже, а «VALUE» предоставляет информацию об общей стоимости всей дневной торговой сессии.

	TRADEDATE	OPEN	LOW	HIGH	CLOSE	VOLUME	VALUE
0	2013-03-25	96.00	96.00	101.14	98.79	593680	59340002.8
1	2013-03-26	98.58	97.08	99.31	97.20	1283550	126030358.8
2	2013-03-27	97.90	95.39	98.00	96.75	1261950	121835900.2
3	2013-03-28	96.38	95.72	98.66	98.59	1971410	192469794.9
4	2013-03-29	98.60	98.32	99.09	98.76	782000	77268860.1

Рисунок 1: Пример данных биржи об акции эмитента.

Получив данные, необходимо проверить целостность. В данном случае я проверял пропуски в данных, и картина получилась не такой хорошей. Как можно видеть на «Рисунке 2», большая часть акций имеет небольшой процент пропусков, менее 20%. В 90 бумагах вообще не было пропусков, или они составляли менее одного процента. В данных подсчётах учитывалось отсутствие данных в колонке «CLOSE» в любой из дней. Тем не менее процент пропусков более 10 уже будет плохо сказываться на достоверности данных при обучении и

валидации, особенно если учитывать бумаги, у которых пропуски под 100%.

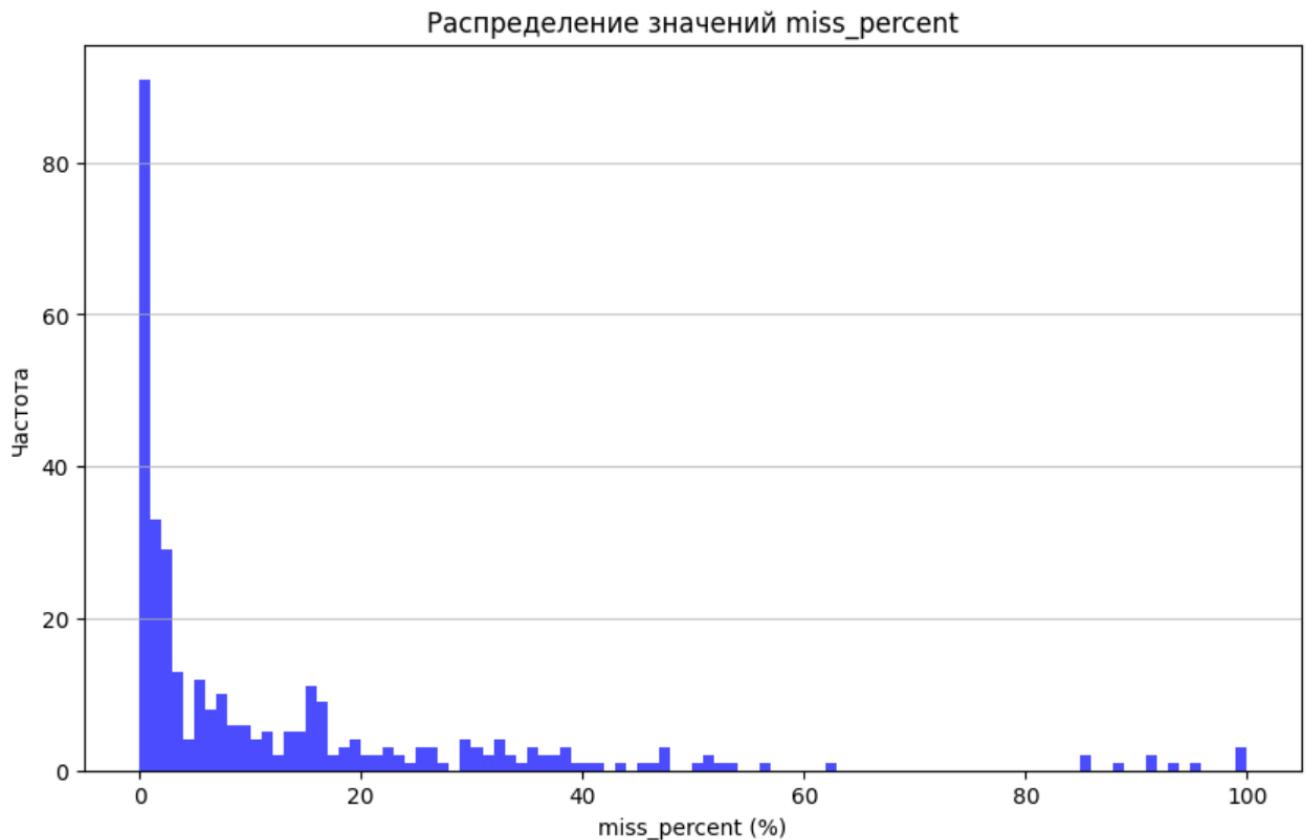


Рисунок 2: Процент пропусков в данных цен акций.

Кроме подсчёта пропусков я получил данные по количеству дней для каждой из бумаг. На «Рисунке 3» видно, что есть аномальный выброс данных на значении 2624. Это значит, что по какой-то причине у очень большого количества бумаг данные были получены не ранее определённого дня, примерно 10.5 лет назад. Не знаю, почему могло такое произойти, но приходится работать с этим. Предположительно, по многим бумагам данные просто не сохранились на Мосбирже, или же при какой-то активности с ними данные более ранних цен были утеряны.

Очевидно, что пропуски в начале данных тяжело заполнить какими-то правдоподобными значениями, это слишком трудно, ведь даже не на что опереться, поэтому я решил избавиться от них. После их устранения средний процент пропусков снизился с 12.87% до 12.23%, что не выглядит существенным улучшением. Если смотреть отдельно на бумаги, то минимальным улучшением было 0.01% или вообще без изменений, потому что не все акции имели пропуски, ну или они были не в начале последовательности. А вот лучший результат получился 84.4%

Простое удаление начальных пропусков позволило немного уменьшить их процент в

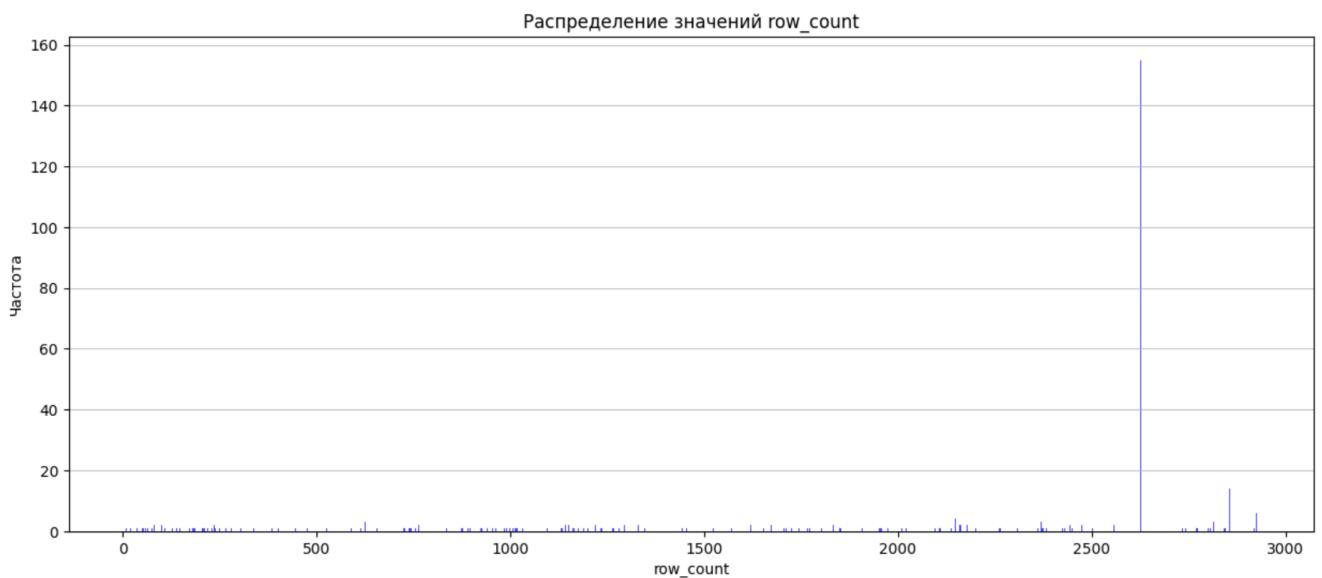


Рисунок 3: Число дней известных для каждой из бумаг.

акциях, что видно на «Рисунке 4» или на «Рисунке 5» с изменениями до 10%.



Рисунок 4: Изменение процента пропусков в данных цен акций.

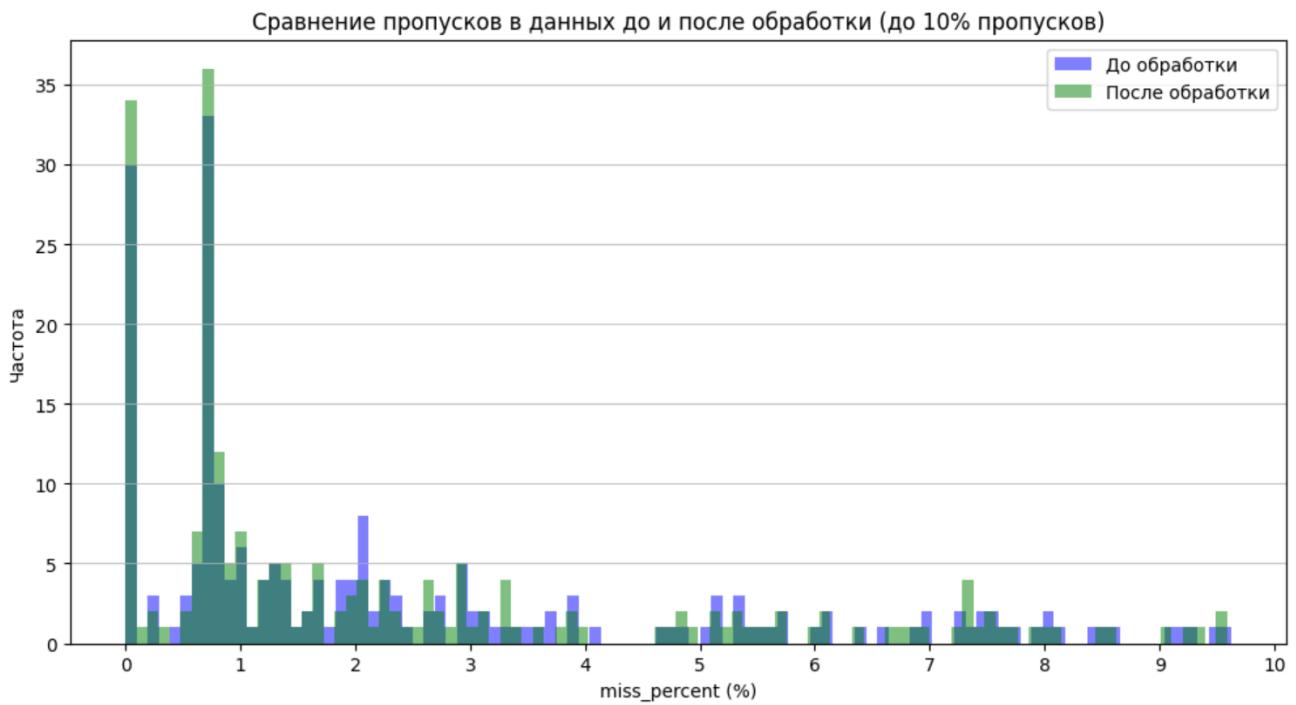


Рисунок 5: Изменение процента пропусков в данных цен акций для акций с пропуском менее 10%.

Кроме того, изменилось и число данных в акциях, как можно видеть на «Рисунке 6».

Даже в тот аномальный день не все акции имели валидные начальные значения, что делает появление такого выброса ещё более странным.



Рисунок 6: Изменение числа дней, известных для каждой из бумаг.

Если говорить об изменении цены акции, то по «Рисунку 7» можно видеть, что практически вся разница цен получается небольшой, это связано с тем, что сама стоимость бумаг не такая большая, и изменения не так сильны за день. Но некоторые бумаги имеют большую стоимость, поэтому видно, что изменения за день бывают и больше чем 750.

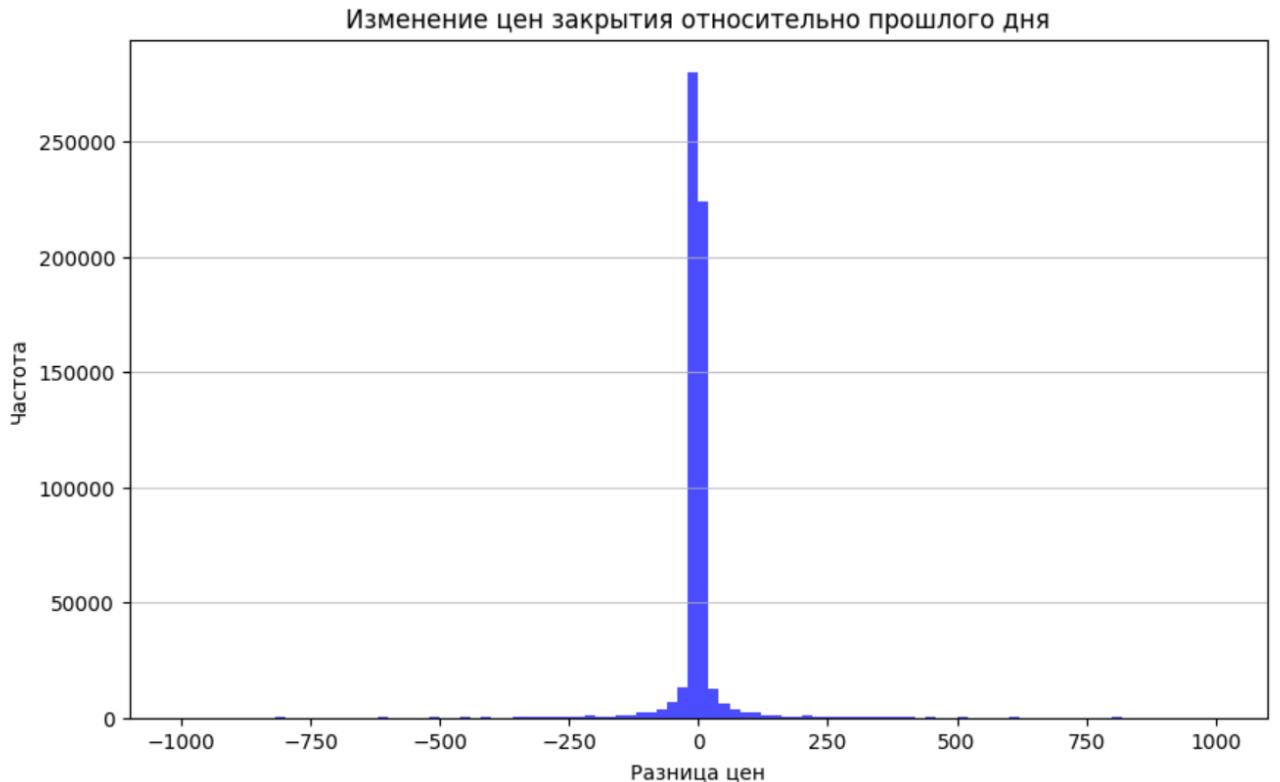


Рисунок 7: Изменение цен акций за один день.

Правда, удобнее смотреть изменение цен в процентах, что отображено в «Рисунке 8». Видно, что в процентном соотношении цена меняется незначительно, в преобладающем большинстве примерно на половину процента в день. Да, эти графики являются обрезанными по краям, потому что иначе их анализ был бы не таким удобным. Ведь бывают резкие изменения цен акций при новостях, отчётах или тех же дивидендных отсечках.

Кроме всего выше написанного, хочу представить графики по числу бумаг в обороте за один день «Рисунок 9» и по их стоимости «Рисунок 10», из чего следует, что примерный оборот в день составляет около 50 миллиардов.

Изменение цен закрытия относительно прошлого дня в процентах

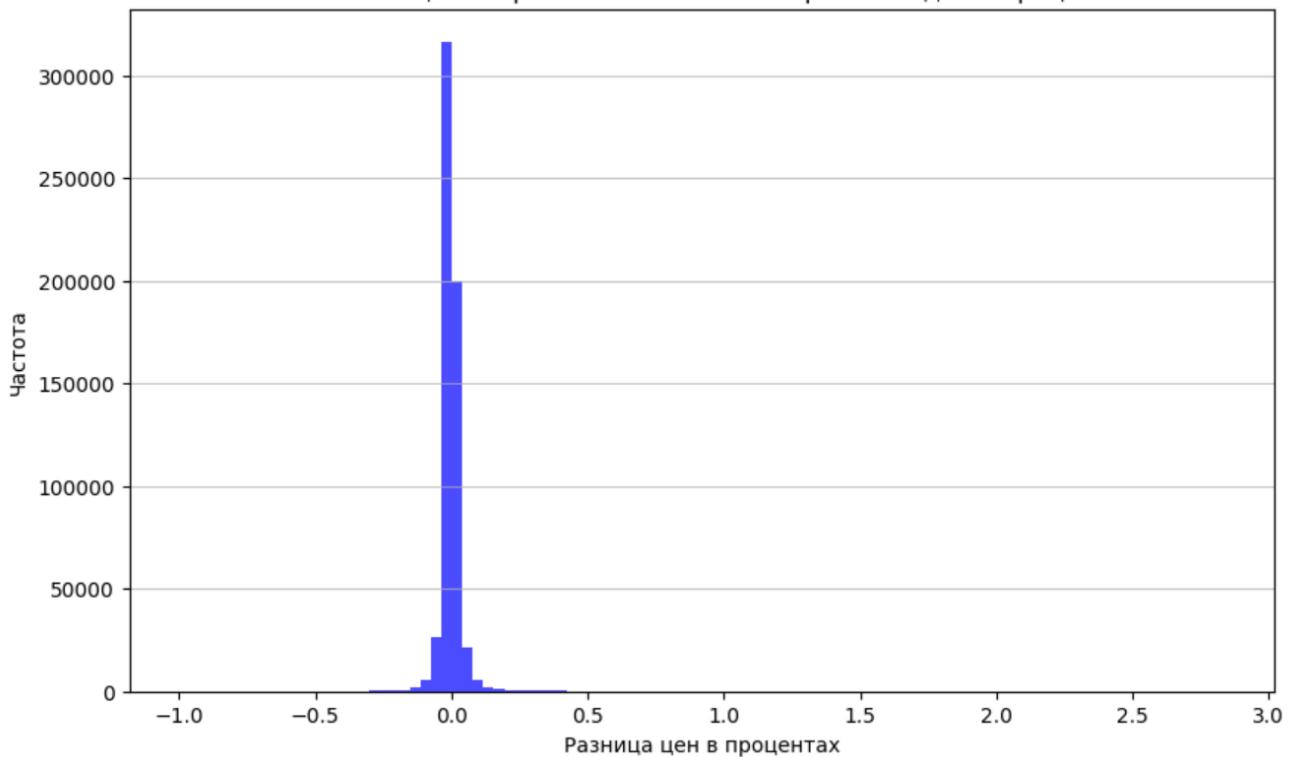


Рисунок 8: Изменение цен акций за один день в процентах.

Распределение значений VOLUME

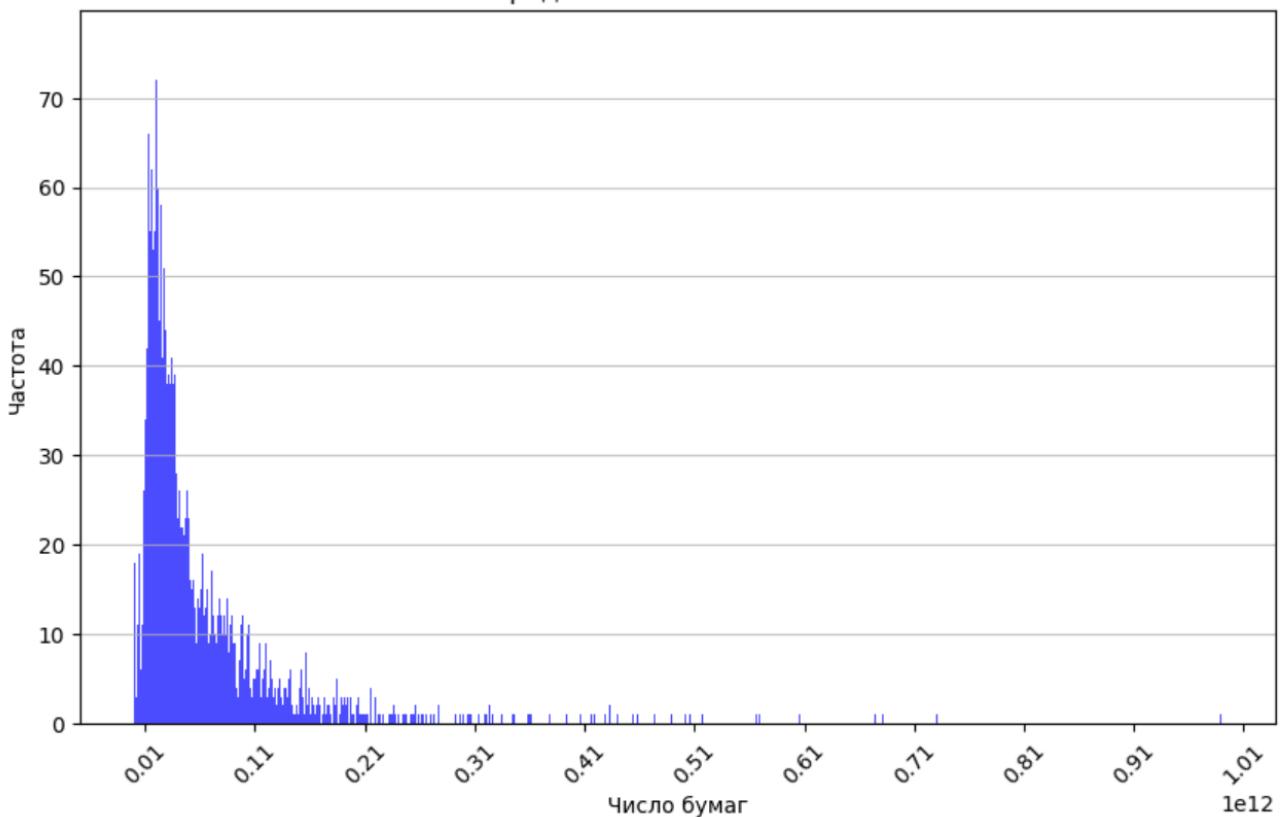


Рисунок 9: Оборот торгов за день.

### Распределение значений VALUE

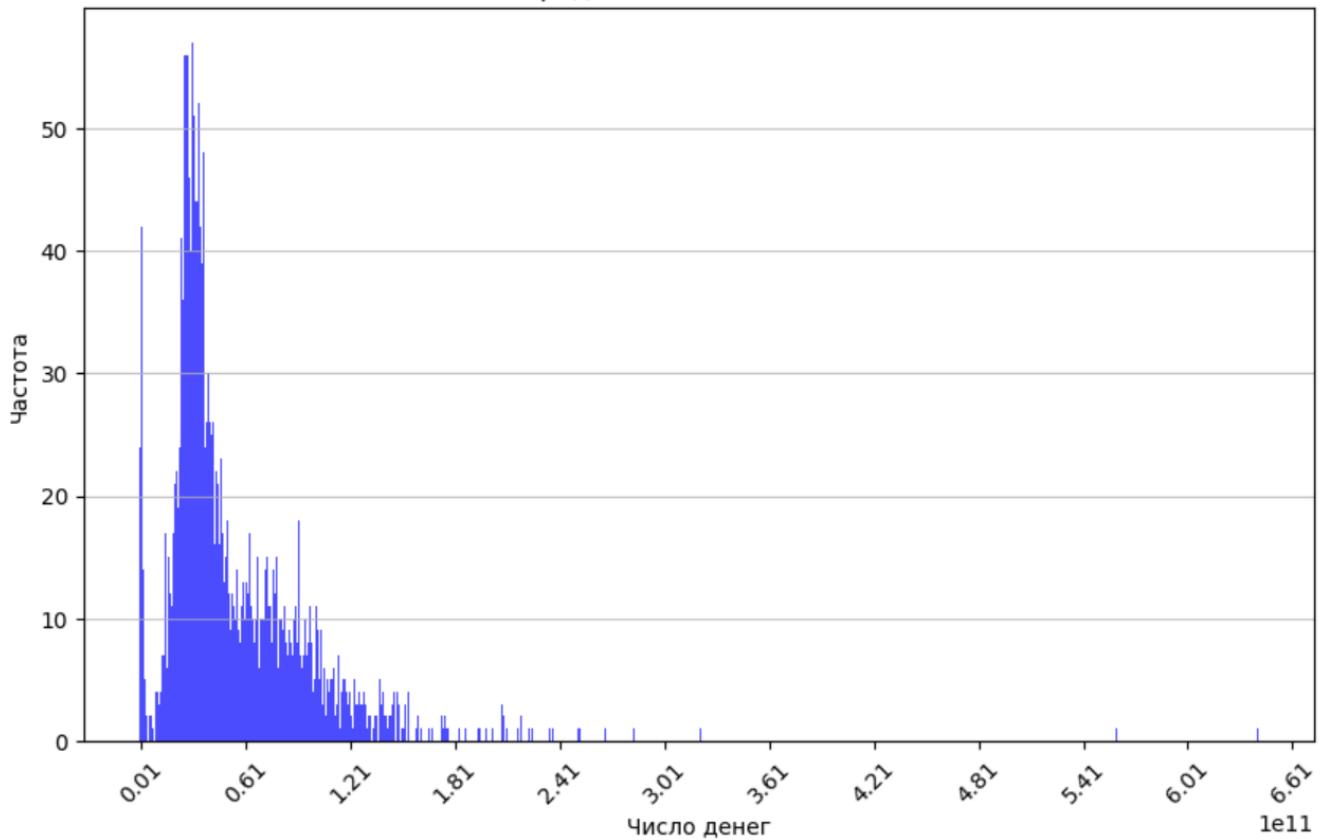


Рисунок 10: Объём торгов за день.

Для получения хоть сколько-то реальных предсказаний пришлось отказаться от всех бумаг, пропуски в которых составляли более 10%. Таким образом, их общее число сократилось до 215 единиц.

После чего оставался последний шаг в предобработке данных - это избавление от пропусков. По различным причинам большинство акций имеют пропуски, хотя хранят день, как будто полноценно торговались, что можно видеть на «Рисунке 11». Скорее всего, в данные дни намеренно приостанавливались торги, что и повлияло на такие разрывы.

Для избавления от пропусков я для всех бумаг создал скрипт по их заполнению, используя линейную интерполяцию. Так, например, акция с тикером RKKE имела более 8% пустых значений. Выполнив программу, я получил результат как на «Рисунке 12», на котором можно видеть прямые участки, которые были добавлены с помощью заполнения. Можно было бы попробовать придумать способ получше для заполнения пропусков, но это сильно усложняет, заставляет рассматривать каждый случай отдельно и сбивает нас с поставленной задачи.

В данной главе я рассказал, как собирал данные по ценам акций, обучение на которых будет позже. Предоставил анализ по распределению, пропускам и выявленной мною

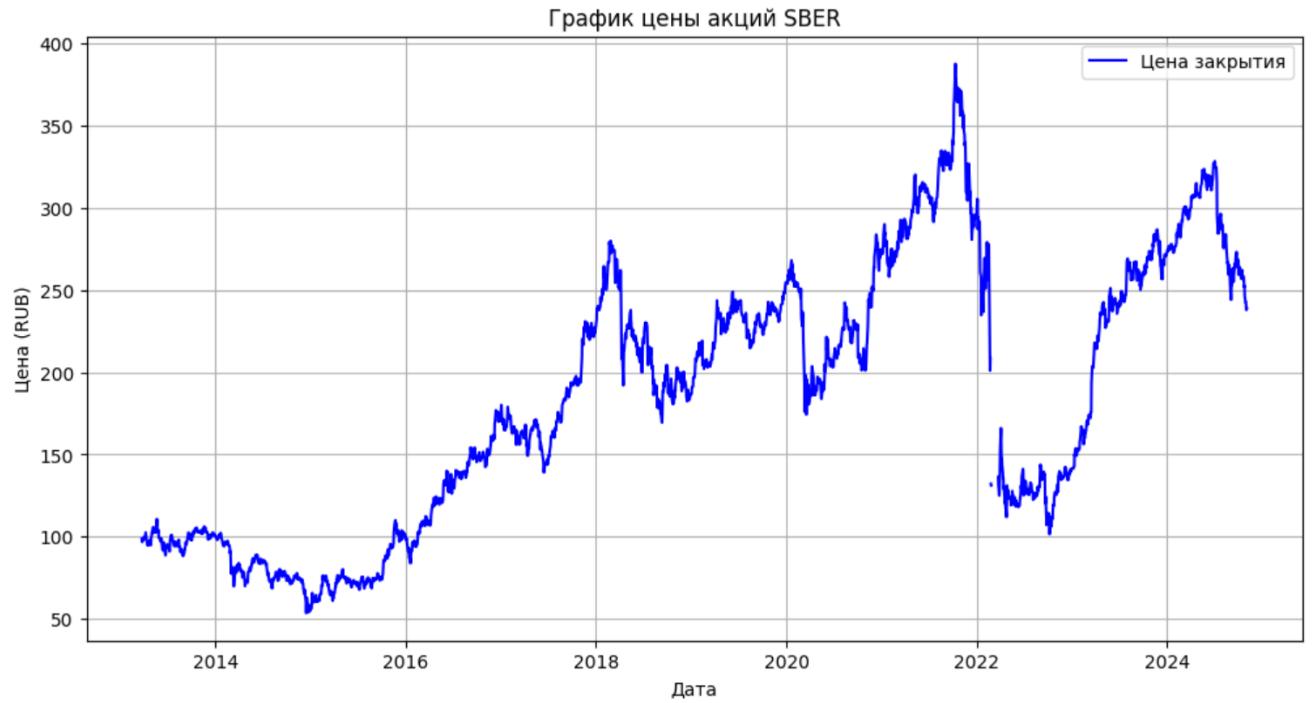


Рисунок 11: Цена закрытия акции Сбера.

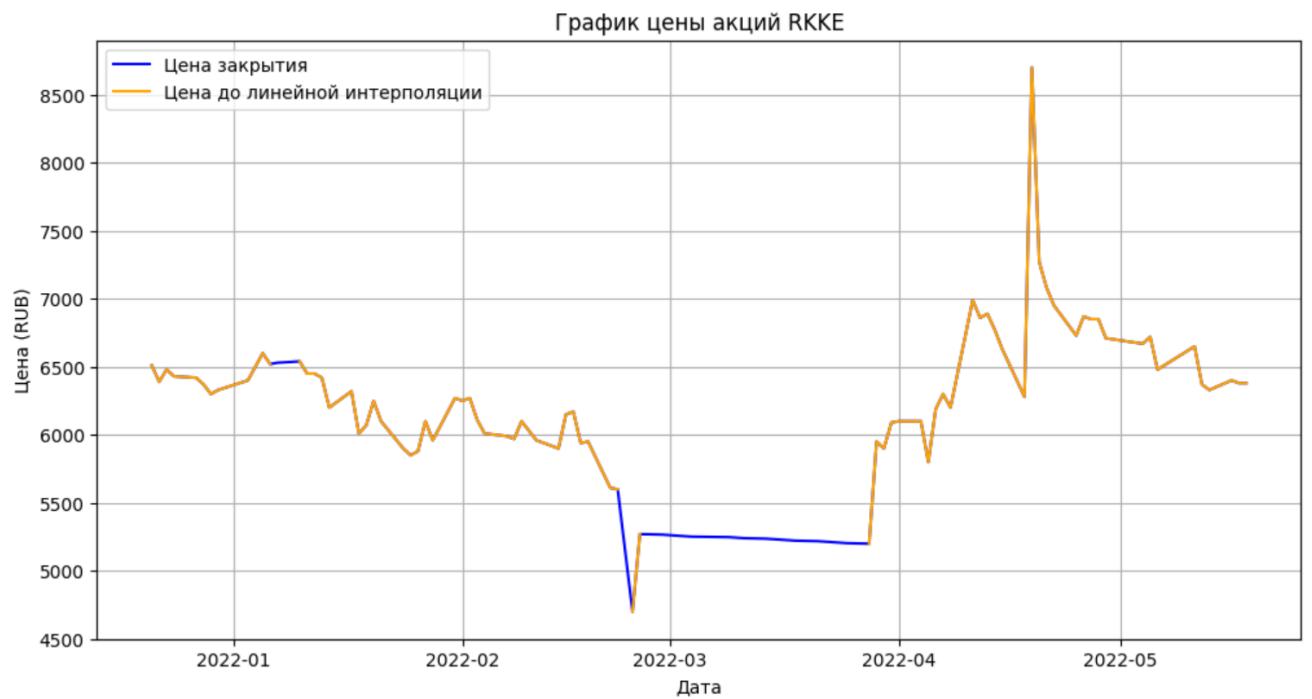


Рисунок 12: Заполнение линейной интерполяцией.

информации. И ещё предобработал полученную информацию, подготовив её к следующим действиям.

## 4.3. Новости по акциям

Для получения новостей я начал изучать доступные источники. Посмотрев различные сайты, статьи и массивы данных на Kaggle [6], я не нашёл того, что меня бы устроило. В основном данные были неактуальны для русского информационного поля, их было не так много или они были в малом количестве, поэтому я решил собрать свою базу новостей.

Для выбора источника я рассматривал несколько сайтов. В их число входили РИА [14], БКС [10], РБК [13], Smart-lab [9], МОЕХ [7], Finam [2] и Investing [4]. При выборе фаворита я смотрел на удобство получения данных, на полноту, то есть как много и часто появляются новости, и на количество исторических данных.

По итогу я выбрал использование РИА новостей [14]. Данные там хранятся начиная с 2010 года, каждый день было от 200 до 900 новостей, и способ получения данных был достаточно понятен. Да, на некоторых источниках можно было найти новостные данные конкретно по отдельным эмитентам, что могло бы упростить задачу, но я хотел получать больший набор данных, чтобы можно было выявить влияние информации на различные бумаги.

После того как я определился с выбором, я начал получать данные. Изначально мне было необходимо получить ссылки на сами новости, потому что они выдавались порционно по 20 штук на каждый из дней отдельно, например, как в источнике [15]. К сожалению, не все ссылки были читаемыми, и была трудность с количеством обращений на сайт, но мне удалось получить более двух миллионов ссылок.

Следующим этапом было получение уже самих новостей. Для этого я обращался по каждой из ссылок из предыдущего этапа и считывал содержимое. Я всё так же боролся с 429 ошибкой о превышении числа запросов за единицу времени. При получении страницы с новостью я вычленял из неё всю нужную информацию: заголовок, подзаголовок, аннотацию, цитату, текст. На самом деле видов информации там было больше, но я свёл к этим пяти типам. Кроме того, всё что не являлось заголовком или аннотацией, могло встречаться в новости более чем один раз, такие данные я объединял и все имеющиеся ссылки заменял на текст, чтобы не было какого-то странного значения для LLM.

К сожалению, не все новости были для меня полезны. РИА новости часто запускали опросы на разные темы, как на «Рисунке 13». Такие статьи не несут полезной информации, поэтому я просто сохранял ссылку и никак не обрабатывал информацию из неё.

Поделиться



© РИА Новости / Владимир Астапкович | Перейти в медиабанк



1 из 8

## Когда в нашей стране начали отмечать Новый год в ночь с 31 декабря?

- В 1492 году
- В 1700 году
- В 1767 году

Ответить

Рисунок 13: Опрос РИА новостей.

Кроме этого, некоторые новости были не такого формата, как преобладающее большинство, например, «Рисунок 14». Просмотрев большое количество подобных неподходящих данных, я понял, что они тоже редко обладают какой-то важной информацией. Поэтому я решил не усложнять код и не терять время на этом, такие новости я тоже не использовал в дальнейшем.

Так же встречались новости, которые включали в себя только аннотацию и заголовок. «Рисунок 15». Такие новости я на всякий случай записывал как неполные, но их содержание могло быть важным, поэтому, несмотря на такое малое количество информации, я их добавлял в данные для дальнейшей работы.

Все остальные новости были валидными. В общей сложности у меня получилось более 14 лет новостей, из которых чуть меньше 22 тысяч я принимал за неподходящие, что не должно сильно повлиять, ведь это менее одного процента.

Следующим этапом я приступил к оценке новостей. Для этого я начал выбирать модели LLM для использования и источники, где есть к ним бесплатный доступ, потому что у

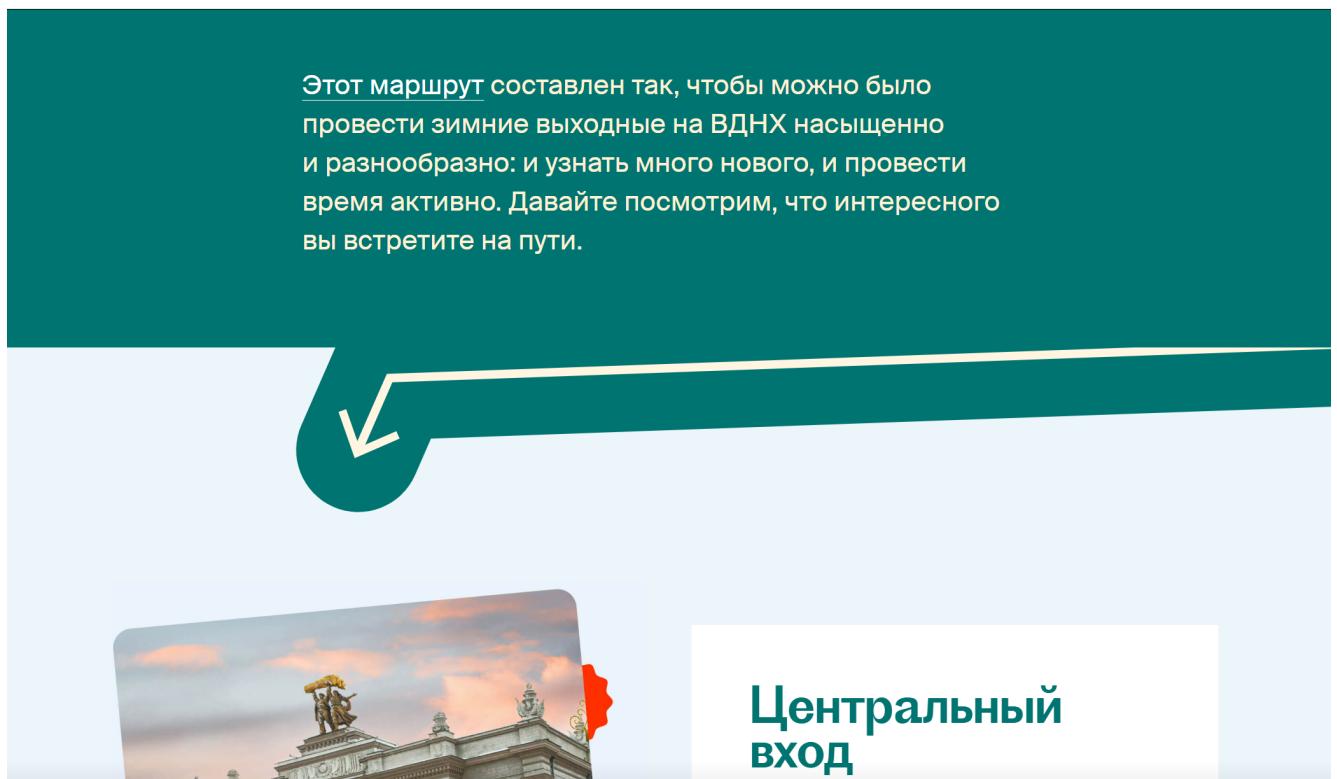


Рисунок 14: Нестандартный формат РИА новостей.

меня было более двух миллионов новостей, и промт запросы включали большое число токенов, обычно из этого и строится цена обращений. В качестве сервиса я нашёл, на то время бесплатный, SambaNova [8]. Доступ был без ограничений, не считая число запросов, с чем я справился. Но что более важно, там не было ограничений из-за моего использования из России.

Для использования их мощностей я получил API ключ и начал тестирование их моделей. Лучшим вариантом для меня стала модель «Meta-Llama-3.3-70B-Instruct». На тот момент она была второй по мощности, но обрабатывала в разы больше запросов за единицу времени. Кроме этого, я протестировал её на своих примерах с положительными, отрицательными и нейтральными новостями, и её ответы мне понравились.

Для получения наилучшего результата я создал промт, основными особенностями которого были следующие части:

1. Модель является опытным аналитиком с большим стажем на рынке и хорошо разбирается в новостях и их влиянии на российские компании.
2. Хорошо понимает специфику российской аудитории, реакцию людей на новость и как на какую компанию может это повлиять.



19:26 19.12.2024

Поделиться

## От "Орешника" до анекдотов. Ежегодная пресс-конференция Владимира Путина



Сегодня, 19 декабря, в Гостином Дворе в Москве состоялась традиционная пресс-конференция Путина, совмещенная с прямой линией. Президент отвечал на вопросы четыре с половиной часа. Самое интересное — в нашей подборке.

Рисунок 15: Новость аннотация.

3. Модель получает все тикеры, которые я собрал ранее, 215 штук, и может предложить свои, в случае дальнейшего обучения.
4. Обращает внимание на даты новостей и учитывает, есть ли влияние на российский рынок вообще.
5. Есть описание, как передана новость (заголовок, текст и другое), и предоставлена новость.
6. Ожидаемый ответ в виде тикера, важности новости для этой компании, сектора компании и причины, почему была выбрана именно такая важность этой новости для данного эмитента.
7. Важность должна быть от -1 до 1.

8. Учитывать возможное влияние не только на описанную компанию, но и на конкурентов или союзников из того же сектора или зависимых.

Как я описал ранее, меня устраивали её ответы, то есть и важность, и приведённая причина соответствовали новости. В качестве примера были такие результаты.

Важность 0.542, причина: «Новость о создании газового хаба в Турции может увеличить поставки газа в Европу, что положительно скажется на компании Газпром». Как можно видеть, в условиях давления на российский экспорт возможность создания хаба в Турции является большим плюсом для Газпрома, ведь эта страна является членом НАТО и входит в состав Евросоюза, то есть таким способом может помочь России продолжить продавать свой газ под другим флагом, не так сильно снижая его стоимость.

Важность -0.235, причина: «Снижение цен на газ в Европе может привести к снижению доходов Газпрома». Ну а вторая новость очевидно негативная для Газпрома. Поставки в Европу снижаются, продолжаются санкции и попытки отказа от всего российского.

Завершающим этапом было преобразование всех данных под каждый тикер компании, ведь было бы неудобно для использования обращаться к каждому дню и искать среди них нужные значения. Для решения этой задачи я создал скрипт, который выявляет все тикеры и для каждого из них сохраняет минимальную и максимальную важность, выявленную за день, время, в которое эти новости были опубликованы, общее количество новостей за день и их среднюю важность.

В данной главе я описал, как выбирал, получал и обрабатывал новости по желаемым эмитентам. Как их оценивал и предобрабатывал для дальнейшего обучения на них. В результате я получил файлы в облачном хранилище для каждого тикера с необходимой мне информацией на каждый из доступных дней.

## 5. Модели

### 5.1. Метрики

Для оценки моделей, основываясь на опыте предыдущих работ, мною было выбрано две метрики.

RMSE «Рисунок 16» является очень полезным и удобным способом отображения ошибки предсказаний от реальных значений в тех же единицах. Её суть в том, что мы складываем все квадраты разностей реального значения и спрогнозированного, после чего делим на ко-

личество предсказаний и извлекаем корень для получения ошибки в тех же значениях. В формуле  $n$  - число прогнозов,  $y_i$  - реальное значение,  $\hat{y}$  - прогноз модели.

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

Рисунок 16: Формула RMSE ошибки.

MAPE «Рисунок 17» является более наглядным способом для отображения средних ошибок, потому что не привязан к определенным числам, а показывает ошибку в процентах, что может быть полезно, когда для разных акций цена сильно отличается. Её формула немнога отличается: считается сумма модулей всех разностей реальных значений и предсказаний, деленных на реальное значение, происходит умножение на 100% и деление на количество предсказаний. В таком случае мы и получаем процент средней ошибки на всех предсказаниях. В формуле  $n$  - число прогнозов,  $y_t$  - реальное значение,  $\hat{y}$  - прогноз модели.

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \cdot 100\%$$

Рисунок 17: Формула MAPE ошибки.

Но, к сожалению, временные ряды, по моему мнению, считать чисто по этим формулам было бы неверно. Ведь очевидно, что любой человек, хоть сколько-то понимающий в фондовом рынке, сделает предсказание на завтра лучшее, чем любая самая умная модель, которая попытается предугадать, что будет с бумагой через год. Поэтому в своей работе я

считал эти ошибки не на всём промежутке предсказания, а для каждого дня высчитывалась своя ошибка отдельно. Это помогало мне вычислять точность предсказаний на каждый день и смотреть, какая модель лучше показывает результаты на каком промежутке. Если вернуться к моим формулам выше, то для подсчета ошибки просто убирается сумма и деление на число предсказаний, ведь каждый день считается отдельно.

## 5.2. По ценам

Для более точного предсказания я выделил группу бумаг, у которых известны данные о ценах более чем за 5 лет. Это было сделано для получения более точных предсказаний, ведь, как можно было видеть выше на графиках, для некоторых из акций были известны значения цен менее чем за месяц. Я не избавился от них, они тоже были обучены и проанализированы, просто в дальнейшем я буду показывать результаты только по тем 139 компаниям, данные по которым известны более пяти лет.

Биржевой рынок не может обеспечить стабильный поток данных, что сказывается на точности. В нашей стране есть праздники и многое другое, в каких-то неделях может быть 3 рабочих дня, где-то 6, добавляются торги выходного дня и вечерние, чего раньше не было. Такие изменения не позволяют удобно понять цикл, тем более его как-то запрограммировать. Я пробовал использовать более базовые модели, например: ForecastingHorizon, NaiveForecaster, ExponentialSmoothing и AutoETS для прогнозирования именно временных рядов, но из-за описанных проблем приходилось добавлять костыли, и результат мне не нравился.

Тем не менее, для всех моих итоговых моделей, кроме SARIMAX и LSTM, я использовал общий подход. Он заключался в том, что я брал даты и цены закрытия на каждый из дней. Дату я разбивал на год, месяц и день, для цены закрытия я делал лаги. В этом и заключалась основная сложность и непредсказуемость результатов. В году 365 дней, что уже не точно, а рабочих получается около 247. Следуя из этой, примерно верной, информации, я пробовал добавлять лаги, так у меня получилось 1, 2, 3, 4, 5 для первых пяти дней, 10 для лага в две недели, 21 - месяц, 62 - три месяца, 124 - полгода, 247 - год, но цикл может быть и более года, рынок может быть как медвежьим, так и бычьим. Поэтому я добавил ещё лаги на полтора, два и три года. Кроме этого, мне приходилось подчищать данные лагов, чтобы не было такого, что для предсказания на третий день использовалось реальное значение второго дня.

Алгоритм в основном сводился к следующему решению. Я определял параметры для поиска лучшей модели. Проходился по каждому из тикеров. Добавлял лаги, разбивал да-

ту, производил деление на тренировочную и тестовую части. Выполнял нормализацию, используя MinMaxScaler. И начинал GridSearchCV обучение, где таргетом было целевое значение, обучение было направлено на уменьшение RMSE ошибки, а за разбиение отвечал TimeSeriesSplit, который делил на 5 частей. После завершения обучения я итеративно на каждый день делал предсказания и считал ошибки. Для всех, кроме первого дня, я дополнял данные пустых лагов полученными ранее предсказаниями, чтобы третий день считал, что во второй день цена была равной той, что прогнозирует модель. После чего вся необходимая информация сохранялась в облачное хранилище.

### 5.3. С использованием новостей

Для предсказаний с новостями пришлось немного усложнить алгоритм. Во-первых, на каждый из дней добавлялись данные с информацией о новостях (максимальная и минимальная важности, их время, количество новостей и их средняя важность). Если для данной акции не было информации на этот день, то значения важностей устанавливались равными нулю, число новостей - единице, а время - на 14:00, что является примерно серединой торгового дня. После этого время разбивалось на часы и минуты, и добавлялись лаги по той же схеме, что описана в предыдущей главе.

Основным усложнением было то, что новости в будущем также не известны, получается, все добавляемые значения, то есть важности, число, время, должны также прогнозироваться. Таким образом, в предсказании получается уже 9 ожидаемых полей. И во время последнего этапа с оценкой и получением значений на будущие дни от модели лаги по новостям заполняются по тому же принципу, что и лаги по цене.

### 5.4. SARIMAX

SARIMAX (Seasonal Autoregressive Integrated Moving Average with eXogenous regressors) — это мощная модель временных рядов, которая используется для прогнозирования данных, имеющих сезонные компоненты и внешние регрессоры. SARIMAX является расширением модели ARIMA, которая включает в себя сезонные эффекты и возможность добавления внешних переменных.

S (Seasonal): Этот компонент добавляет сезонные эффекты в модель.

AR (Autoregressive): Этот компонент учитывает зависимость текущего значения временного ряда от его предыдущих значений.

I (Integrated): Этот компонент отвечает за разности временного ряда, чтобы сделать его стационарным. Стационарный ряд — это ряд, у которого статистические свойства, такие как среднее и дисперсия, не зависят от времени.

MA (Moving Average): Этот компонент учитывает зависимость текущего значения от предыдущих ошибок прогноза. Это позволяет модели корректировать свои предсказания на основе прошлых ошибок.

X (eXogenous regressors): Это возможность добавления внешних переменных, которые могут влиять на временной ряд.

Данная модель больше всего отличается от описанного выше шаблона: в ней не используются новости или какие-то ещё поля; необходимы только данные по ценам, но предсказания я так же оценивал ежедневно, чтобы иметь один формат результатов.

В данном случае я не вижу необходимости показывать результаты по отдельным бумагам. Предсказанные ею значения неплохи, но даже на малом промежутке ничего классного нет. Она редко показывают результаты лучше остальных моделей. Хотя у неё есть преимущество в малой требовательности к данным.

Если рассматривать средние результаты по 139 акциям старше пяти лет, то в «Таблице 1» видно, что ошибка в первый день в среднем составляет 3%, а к концу второй недели - уже более 12 с половиной. И дальше результат, очевидно, лучше не будет, что видно на «Рисунке 18», который показывает увеличение ошибки в первый месяц.

Таблица 1: Таблица средних ошибок SARIMAX на две недели.

День предсказания	RMSE	MAPE
1	82.211	3.018
2	154.788	5.100
3	254.078	6.384
4	314.963	6.573
5	268.571	6.475
6	374.613	8.622
7	486.077	11.278
8	641.263	12.515
9	692.472	12.436
10	612.356	12.491



Рисунок 18: MAPE ошибки SARIMAX на первый месяц.

## 5.5. Ridge Regression

Ridge - это техника, которая применяется в контексте линейной регрессии для борьбы с проблемой мультиколлинеарности и переобучения. Основная идея заключается в добавлении L2-регуляризации к функции потерь, что позволяет контролировать величину коэффициентов регрессии.

Обучение этой модели проходило, как было описано в начале этой большой главы, как без использования новостей, так и с ними. В отличие от SARIMAX количество используемых данных было увеличено, особенно с новостями, и это принесло свои плоды: точность предсказаний у данной модели в среднем лучше, и она чаще дает более точные предсказания. В некоторых случаях она лучше даже на горизонте в полгода и год. Для обучения параметр `alpha` подбирался в диапазоне от 0.0001 до 10000, а `fit_intercept` - либо `True`, либо `False`.

Правда, если сравнивать данную модель без использования новостей и с ними, то средние значения получались неожиданными. По «Таблице 2» видно, что почти во всех случаях точность предсказания лучше у модели без новостей. На самом деле, такое получается во многих моделях, что я объясню в завершение главы.

Таблица 2: Таблица средних ошибок Ridge на две недели.

День предсказания	Ridge RMSE	Ridge + news RMSE	Ridge MAPE	Ridge + news MAPE
1	<b>20.785</b>	43.508	<b>1.476</b>	1.715
2	<b>25.420</b>	52.559	<b>2.287</b>	2.917
3	53.785	<b>43.131</b>	<b>2.959</b>	3.619
4	<b>71.724</b>	94.299	<b>3.233</b>	3.670
5	<b>61.794</b>	137.609	<b>3.672</b>	4.413
6	<b>68.694</b>	136.872	<b>4.113</b>	5.458
7	<b>81.128</b>	141.974	<b>4.353</b>	5.631
8	<b>99.891</b>	164.282	<b>4.766</b>	5.861
9	<b>102.058</b>	198.797	<b>4.984</b>	6.296
10	<b>94.851</b>	198.242	<b>5.158</b>	6.225

Также можно посмотреть средние MAPE ошибки на ближайший месяц для обученной модели на «Рисунке 19» и для обученной с новостями модели на «Рисунке 20».



Рисунок 19: Средняя MAPE ошибка Ridge на первый месяц.

Более интересно то, что в процентах разница получается от половины до одного процента, но в некоторых случаях RMSE ошибка с новостями в два раза хуже. Предполагаю, что



Рисунок 20: Средняя MAPE ошибка Ridge с новостями на первый месяц.

причина в том, что новостная модель чуть хуже прогнозирует значения для акций с большей стоимостью, поэтому проценты почти не отличаются, а абсолютные значения различны в большей степени.

Хотел бы все же повторить, что это среднее значение, то есть для некоторых бумаг новостная модель показывает результаты лучше.

## 5.6. Random Forest Regression

Random Forest - представляет собой ансамблевый метод, который использует множество деревьев решений для принятия более точных решений. Основная идея заключается в том, что несколько деревьев, обученных на различных подмножествах данных, могут давать более надежные прогнозы, чем одно дерево.

Обучение происходило так же, как и для модели Ridge, за тем исключением, что параметрами для выбора были: n\_estimators - [50, 100, 200, 300, 400, 500], max\_features - [log2, sqrt], max\_depth - [None, 10, 20, 30], min\_samples\_split - [2, 5, 10], min\_samples\_leaf - [1, 2, 4], bootstrap - [True, False]

Результаты средних ошибок на превые две недели представлены в «Таблице 3». По

ней видно, что в первые дни случайный лес показывает результаты немного хуже, чем у модели Ridge, но после второго дня его средняя ошибка меньше, что ещё лучше заметно на «Рисунке 21». К сожалению, случайный лес с новостями показал не столь хороший результат «Рисунок 22», думаю, это связано с тем, что получается слишком много признаков относительно данных, что усложняет нахождение зависимостей.

Таблица 3: Таблица средних ошибок Random Forest на две недели.

День предсказания	Forest RMSE	Forest + news RMSE	Forest MAPE	Forest + news MAPE
1	<b>16.721</b>	81.013	<b>1.712</b>	4.006
2	<b>18.858</b>	83.729	<b>2.443</b>	4.654
3	<b>40.009</b>	92.666	<b>2.950</b>	5.300
4	<b>51.671</b>	85.500	<b>3.047</b>	5.339
5	<b>46.386</b>	95.172	<b>3.340</b>	5.794
6	<b>47.510</b>	109.671	<b>3.765</b>	6.312
7	<b>59.279</b>	107.751	<b>3.947</b>	6.543
8	<b>71.268</b>	89.686	<b>4.334</b>	6.820
9	<b>70.553</b>	100.799	<b>4.390</b>	7.241
10	<b>62.685</b>	120.310	<b>4.625</b>	7.295

Случайный лес показывает себя более стабильным в прогнозах на длинный горизонт, даже модель с новостями, имеющая плохие начальные прогнозы, через месяц имеет меньшую ошибку по сравнению с Ridge с новостями.

## 5.7. XGBRegressor

XGBRegressor (Extreme Gradient Boosting Regressor) является частью библиотеки XGBoost, предназначено для реализации алгоритмов градиентного бустинга. Этот метод позволяет создавать предсказательные модели, способные обрабатывать как линейные, так и нелинейные зависимости в данных.

XGBRegressor использует метод градиентного бустинга, который строит модели последовательно, каждая из которых исправляет ошибки предыдущей, что позволяет значительно улучшить точность предсказаний. В отличие от многих других алгоритмов, XGBRegressor включает механизмы регуляризации (L1 и L2), что помогает избежать переобучения модели.

Данная модель обучалась по тому же шаблону, что и две предыдущие, описанному в начале главы. Параметры подбора были следующими: n\_estimators - [100, 200, 300],

График средних ошибок в предсказаниях

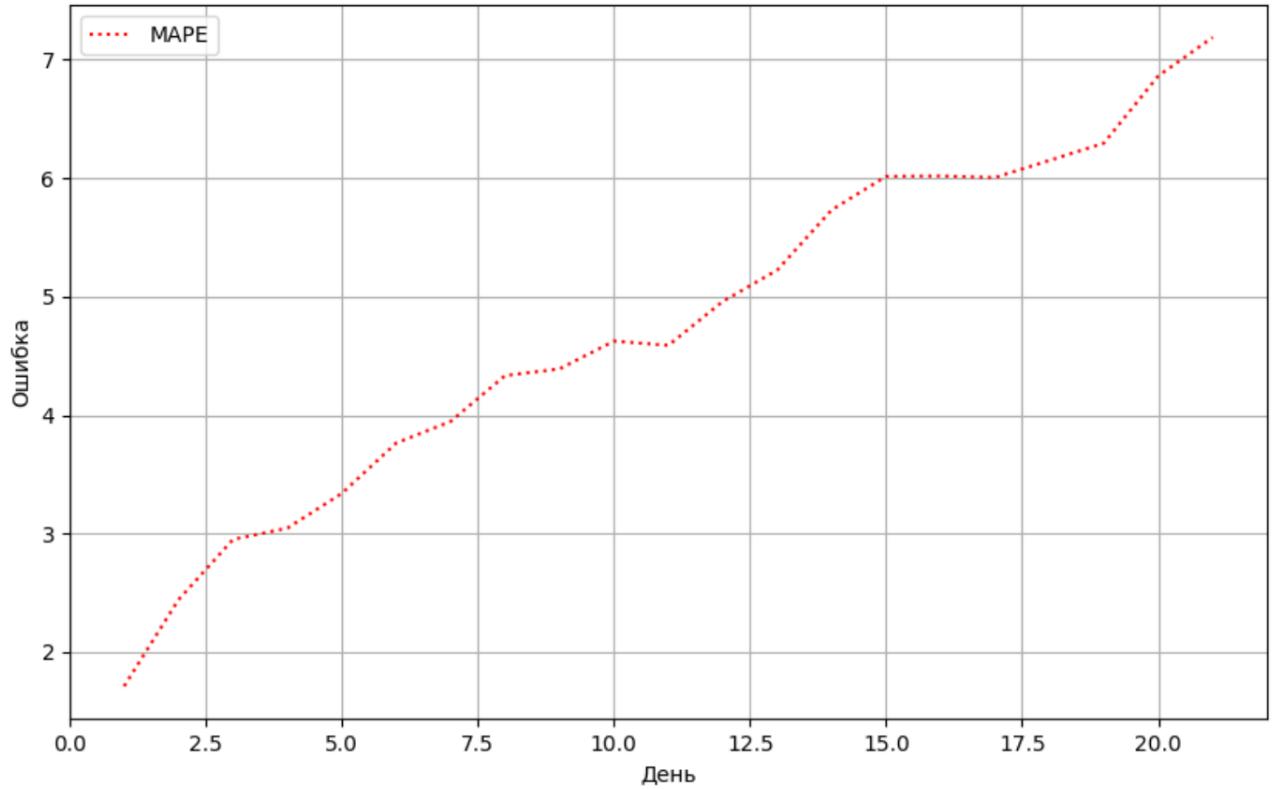


Рисунок 21: Средняя MAPE ошибка Random Forest на первый месяц.

График средних ошибок в предсказаниях

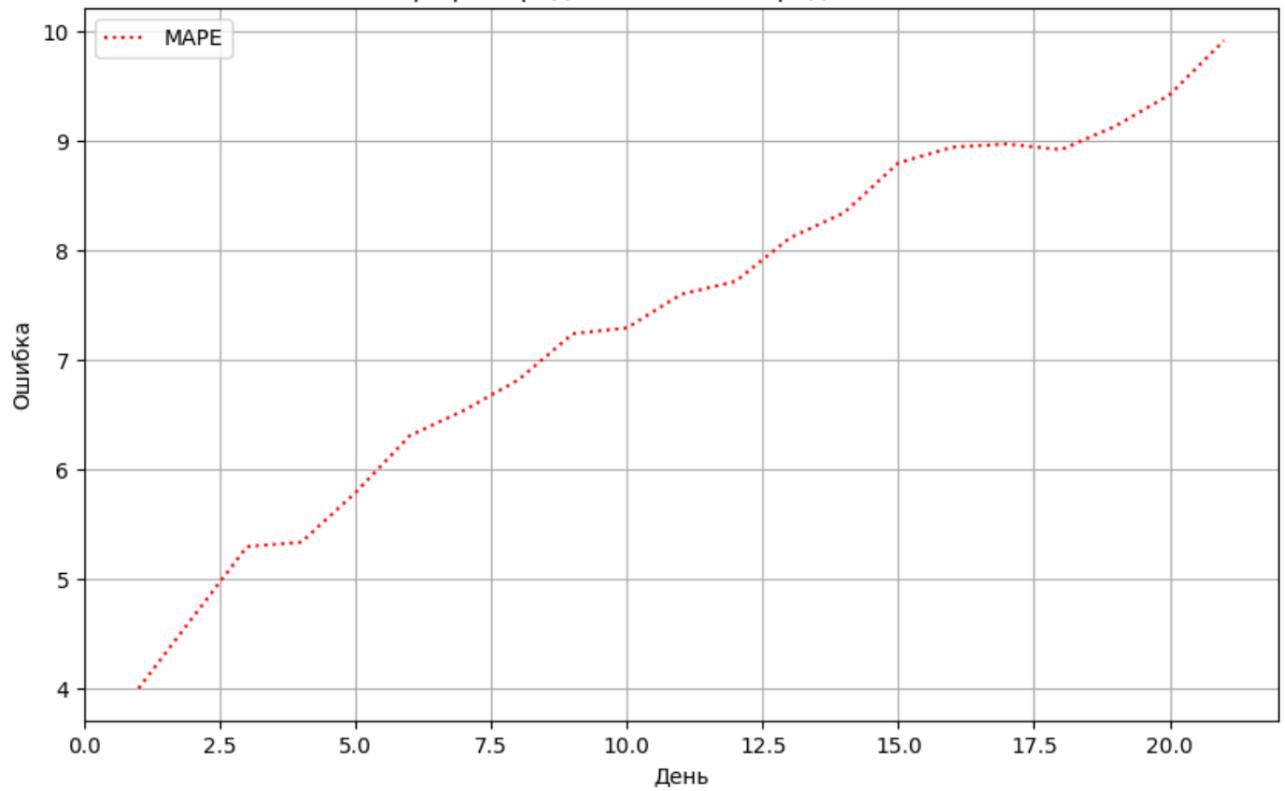


Рисунок 22: Средняя MAPE ошибка Random Forest с новостями на первый месяц.

learning\_rate - [0.01, 0.1, 0.2], max\_depth - [3, 5, 7], min\_child\_weight - [1, 3, 5], subsample - [0.5, 0.7, 1.0], colsample\_bytree - [0.5, 0.7, 1.0], gamma - [0, 1, 5].

Данная модель показала себя не хуже, чем случайный лес. Модель обученная с новостями, хоть и была менее точна в среднем, но оказалась примерно на процент точнее по MAPE, что можно увидеть в «Таблице 4».

Таблица 4: Таблица средних ошибок XGBRegressor на две недели.

День предсказания	XGB RMSE	XGB + news RMSE	XGB MAPE	XGB + news MAPE
1	34.296	<b>26.212</b>	<b>1.727</b>	2.819
2	48.190	<b>23.163</b>	<b>2.471</b>	3.780
3	78.342	<b>30.181</b>	<b>3.002</b>	4.229
4	97.238	<b>47.323</b>	<b>3.146</b>	4.325
5	91.990	<b>59.053</b>	<b>3.552</b>	4.743
6	92.665	<b>45.018</b>	<b>3.993</b>	5.040
7	111.372	<b>53.106</b>	<b>4.229</b>	5.444
8	124.604	<b>72.252</b>	<b>4.536</b>	5.826
9	130.286	<b>98.925</b>	<b>4.639</b>	6.294
10	132.155	<b>116.713</b>	<b>4.866</b>	6.424

В данном случае получилась обратная ситуация: в среднем модель с новостными данными лучше прогнозирует акции с большей ценой, поэтому и получается, что MAPE ошибка с новостями больше, чем без них, а вот в абсолютных значениях хуже себя показывает обычный градиентный бустинг.

На «Рисунке 23» и «Рисунке 24» можно увидеть результаты средних MAPE ошибок на месяц.

График средних ошибок в предсказаниях

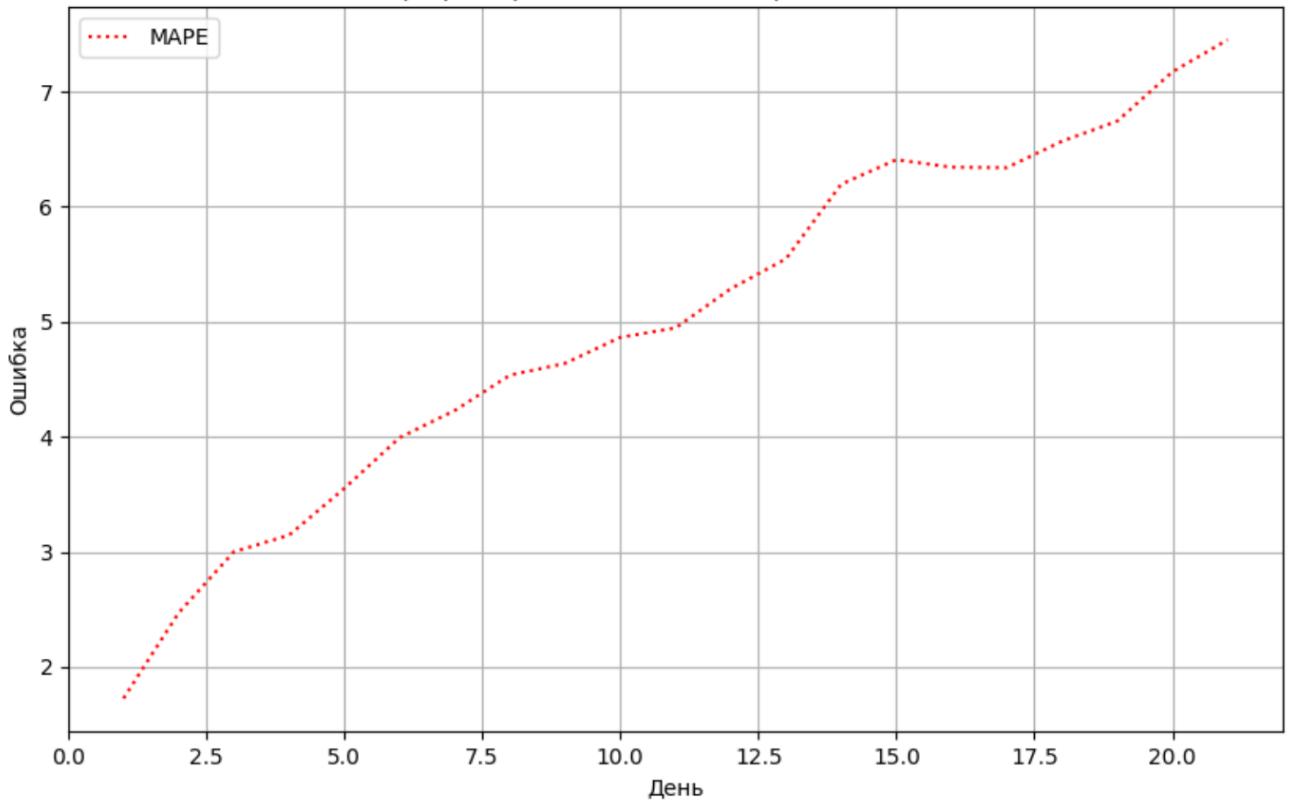


Рисунок 23: Средняя MAPE ошибка XGBRegressor на первый месяц.

График средних ошибок в предсказаниях

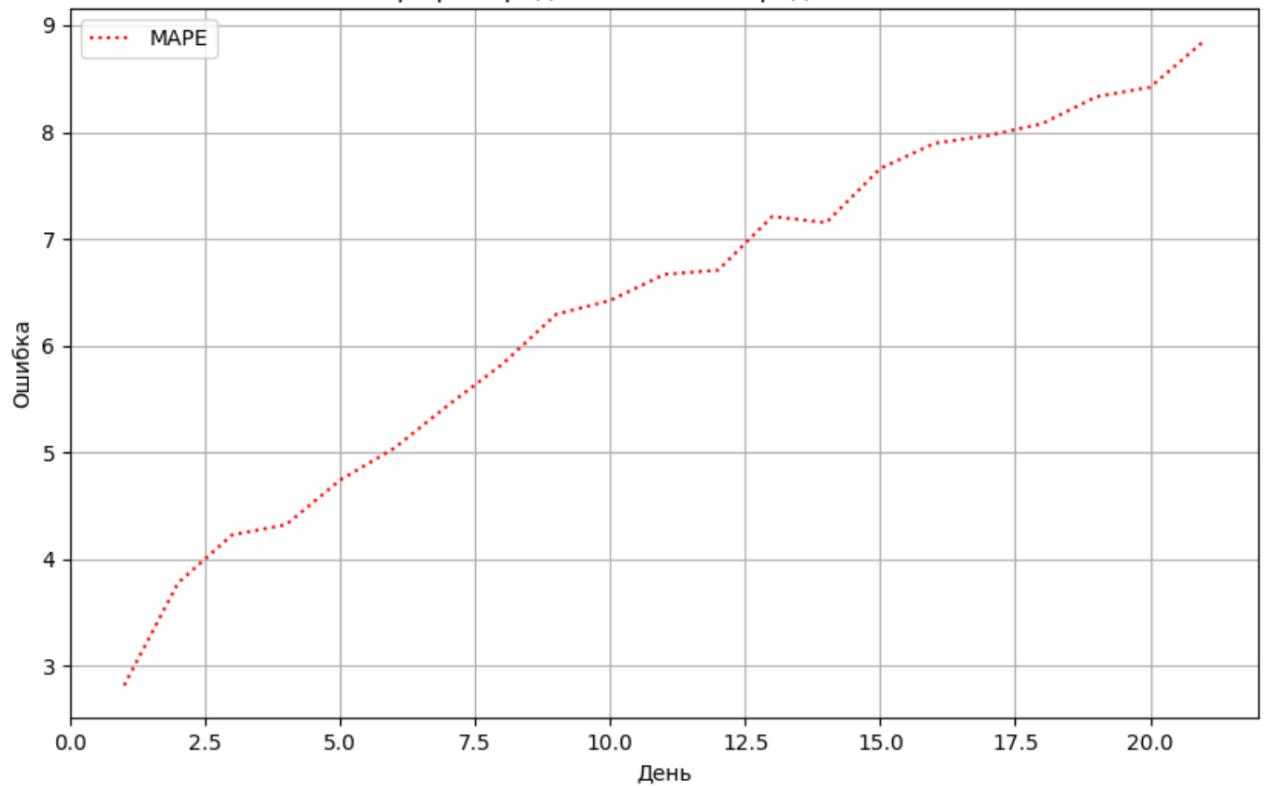


Рисунок 24: Средняя MAPE ошибка XGBRegressor с новостями на первый месяц.

## 5.8. LSTM

LSTM (Long Short-Term Memory) — это тип рекуррентной нейронной сети (RNN), предназначенный для обработки и предсказания последовательных данных. Вопрос касается принципов работы LSTM, его архитектуры и применения в задачах машинного обучения.

Основной подход был таким же, как и в предыдущих моделях, за исключением того, что вместо GridSearchCV я использовался Sequential.

Первый слой состоял из 48 нейронов с регуляризацией recurrent\_dropout=0.1, return\_sequences = True и инициализацией весов с помощью метода he\_normal. Затем следовал слой нормализации, и отключение 20% нейронов. После чего шел второй слой LSTM с 24 нейронами и return\_sequences = False. Далее еще один слой нормализации и два полносвязанных слоя: первый с функцией активации ReLU, а второй - с числом ожидаемых значений. Для обычной модели - единица (только стоимость), для модели с новостями - вывод девяти значений (стоимость и данные ожидаемых новостей).

Оптимизатор использовался Adam со скоростью обучения 0.001 и clipnorm=1.0. Функция потерь - Huber. Обратный вызов останавливается после 15 эпох, если валидационная ошибка не улучшается, восстанавливает лучшие веса. Также было использовано динамическое изменение скорости обучения: после 50 эпохи величина с каждым разом уменьшалась на 5%.

К сожалению, данная модель показала себя хуже остальных, хотя средняя ошибка за две недели не изменялась больше чем на процент, что видно по «Таблице 5». На «Рисунке 25» и «Рисунке 26» видно, что средняя MAPE ошибка за месяц изменилась менее чем на два процента.

Не сомневаюсь, что результаты по данной модели могут быть лучше, но это требует более тщательного подхода и большего количества времени. Но, не смотря на такие средние результаты, данная модель и её версия с новостями часто оказываются в числе лучших для некоторых акций.

## 5.9. Результаты исследований

Итак, основная причина того, что модели с новостями в среднем показывают результат хуже, чем без них. Я считаю, что это связано с тем, что сами новости в полном объёме для каждой из компаний сложно получить. Я пробовал для каждой новости получить все возможные влияния на каждую из бумаг, что не так просто. Но еще сложнее получить

Таблица 5: Таблица средних ошибок LSTM на две недели.

День предсказания	LSTM RMSE	LSTM + news RMSE	LSTM MAPE	LSTM + news MAPE
1	<b>179.778</b>	358.869	<b>12.284</b>	19.643
2	<b>176.325</b>	352.127	<b>12.432</b>	20.080
3	<b>194.466</b>	347.170	<b>12.219</b>	20.286
4	<b>193.421</b>	357.515	<b>12.000</b>	19.972
5	<b>190.941</b>	353.370	<b>12.091</b>	20.026
6	<b>193.543</b>	350.417	<b>12.283</b>	20.106
7	<b>204.219</b>	351.025	<b>12.167</b>	20.019
8	<b>216.102</b>	373.916	<b>12.117</b>	20.152
9	<b>220.227</b>	376.465	<b>12.148</b>	20.440
10	<b>214.584</b>	380.385	<b>12.237</b>	20.377



Рисунок 25: Средняя MAPE ошибка LSTM на первый месяц.

новости по менее популярным акциям. Это я к тому, что условный Сбер или Яндекс будут гораздо чаще фигурировать в новостных сводках, и влияние на их бизнес найти проще, чем на какую-то не столь большую компанию. Таким образом получается, что в модели число признаков становится слишком большим по отношению к числу данных, а их влияние



Рисунок 26: Средняя MAPE ошибки LSTM с новостями на первый месяц.

практически невидимо из-за сложного выявления новостной информации.

Не вижу смысла рассматривать конкретные бумаги, потому что сутью было выявить лучшие модели, но, чтобы не сложилось мнение, что lstm плох или что новости не вносят важные данные, я бы привел в пример данные по тикеру ALRS.

Для неё лучшими моделями по MAPE были следующие: на первый день Random Forest с новостями и ошибкой 0.964%, на пятый день SARIMAX с 0.1%, десятый (две недели), двадцать первый (месяц) и шестьдесят второй (квартал) дни в лидерах был LSTM с новостями с соответствующими процентами по MAPE 0.795, 0.403 и 2.026. На 124 (полгода) день в лидерах был Ridge с новостями с ошибкой 3.381, и через 247 дней (год) снова LSTM с новостями и ошибкой по MAPE 11.859.

На данном примере можно видеть, что в некоторых случаях, на некоторых бумагах, модели с новостями работают лучше, хотя в среднем результат у новостных моделей гораздо хуже. Также видно, что LSTM с новостями, показавшая худшие средние результаты, для данной бумаги на большем горизонте оказалась лучшей. Здесь была бумага эмитента Алроса, которая является достаточно крупной и известной в России, что помогло получить достаточно новостных данных для хорошего обучения и прогнозирования.

И в завершение этой главы я бы хотел по «Таблице 6» продемонстрировать, какие

места в среднем на разных горизонтах времени занимала каждая из моделей. Как видно, в среднем лучшей оказывалась модель Random Forest, от которой почти всегда слегка отставала XGBRegressor. Если смотреть среди новостных моделей, то фаворитом определенно была модель XGBRegressor.

Таблица 6: Таблица средних мест моделей на горизонты дней.

День	SARIMAX	Ridge	Ridge+	Forest	Forest+	XGB	XGB+	LSTM	LSTM+
1	4.78	<b>3.27</b>	3.85	3.73	5.33	4.06	4.47	7.37	8.14
5	5.28	4.01	4.72	<b>3.70</b>	5.14	4.06	4.24	6.35	7.50
10	5.59	4.12	4.53	3.92	5.21	<b>3.83</b>	4.42	6.02	7.36
21	7.01	4.15	4.38	<b>3.86</b>	4.95	3.94	4.45	5.34	6.93
62	8.15	3.99	4.88	<b>3.78</b>	5.01	3.81	4.39	4.55	6.44
124	8.73	4.14	4.60	<b>3.65</b>	4.95	3.68	4.04	4.24	6.01
247	8.08	4.45	4.62	<b>3.14</b>	4.26	4.06	3.95	3.76	5.35

Рассмотрев все вышеописанные модели и полученные результаты, я бы сказал, что в общем случае лучше выбирать Random Forest Regression модель, она достаточно хороша в начале и отлично работает на большем периоде. Лучше её может быть модель Ridge Regression, но только на коротком промежутке. Ну и если новостей по бумаге много, эмитент обсуждаемый или популярный, но не только ставший таким заметным, то можно попробовать XGBRegressor с новостями. Также имеет смысл использовать LSTM, но для работы с ним необходимо потратить много времени и сил. Ну а ARIMA-подобные модели я бы не рекомендовал, для биржевого рынка лучше иметь больше факторов, а данное семейство моделей не очень подходит.

## 6. Сервис

### 6.1. Архитектура

Не менее важен сервис. Для написания серверной части я выбрал Python-фреймворк FastAPI. Для большего удобства пользователям в качестве визуального представления клиентам я выбрал Streamlit. В качестве базы данных я решил не отходить от облачного хранилища Яндекс S3. Для придания более серьезного вида мною были добавлены Nginx и SSL-сертификат. А для большего удобства развертывания использовал Docker Compose.

FastAPI — это современный веб-фреймворк для создания API на Python, который позволяет разработчикам быстро и эффективно разрабатывать приложения. Он основан на стандартных Python-типаах и использует асинхронные функции, что обеспечивает высокую производительность.

Streamlit — это фреймворк с открытым исходным кодом, который позволяет разработчикам создавать веб-приложения для визуализации данных и взаимодействия с моделями машинного обучения. Он был разработан с целью упростить процесс создания приложений, позволяя сосредоточиться на логике и визуализации, а не на сложностях веб-разработки.

S3-хранилище Яндекс — это облачное решение для хранения и управления данными, которое предоставляет пользователям возможность хранить неструктурированные данные, такие как изображения, видео, резервные копии и другие файлы. Оно основано на принципах, схожих с Amazon S3, и предлагает высокую доступность, масштабируемость и безопасность.

Nginx — это мощный веб-сервер, который также может выполнять функции балансировщика нагрузки. Он позволяет распределять входящий трафик между несколькими серверами, что обеспечивает высокую доступность и масштабируемость приложений.

SSL (Secure Sockets Layer) сертификат — это цифровой сертификат, который подтверждает личность веб-сайта и шифрует данные, передаваемые между пользователем и сервером. Он обеспечивает защиту от перехвата данных и атак типа "человек посередине".

Docker Compose — это инструмент, который позволяет определять и запускать многоконтейнерные приложения Docker. Он использует файл конфигурации в формате YAML для описания сервисов, сетей и томов, необходимых для работы приложения.

Таким образом, получается следующая схема. Поднимается сервер с использованием Docker Compose, после чего становятся доступны FastAPI и Streamlit, на которые клиент может делать запросы. Далее Nginx перехватывает запросы и решает, куда их направить. После получения запроса на Streamlit от клиента идёт обращение на бэкенд. Далее FastAPI ищет информацию в своей памяти или делает запрос к S3 Яндекс. Получив от облачного хранилища данные по запросу, FastAPI их обрабатывает и возвращает данные в Streamlit или клиенту. В случае, если клиент делал запрос не напрямую на бэкенд, то полученный ответ Streamlit обрабатывает и отдаёт в удобном для понимания виде, после чего снова ожидается запрос от клиента.

## 6.2. Запросы API

Доступные запросы описаны в «Таблице 7». И на «Рисунке 28»

## Цикл обработки запросов

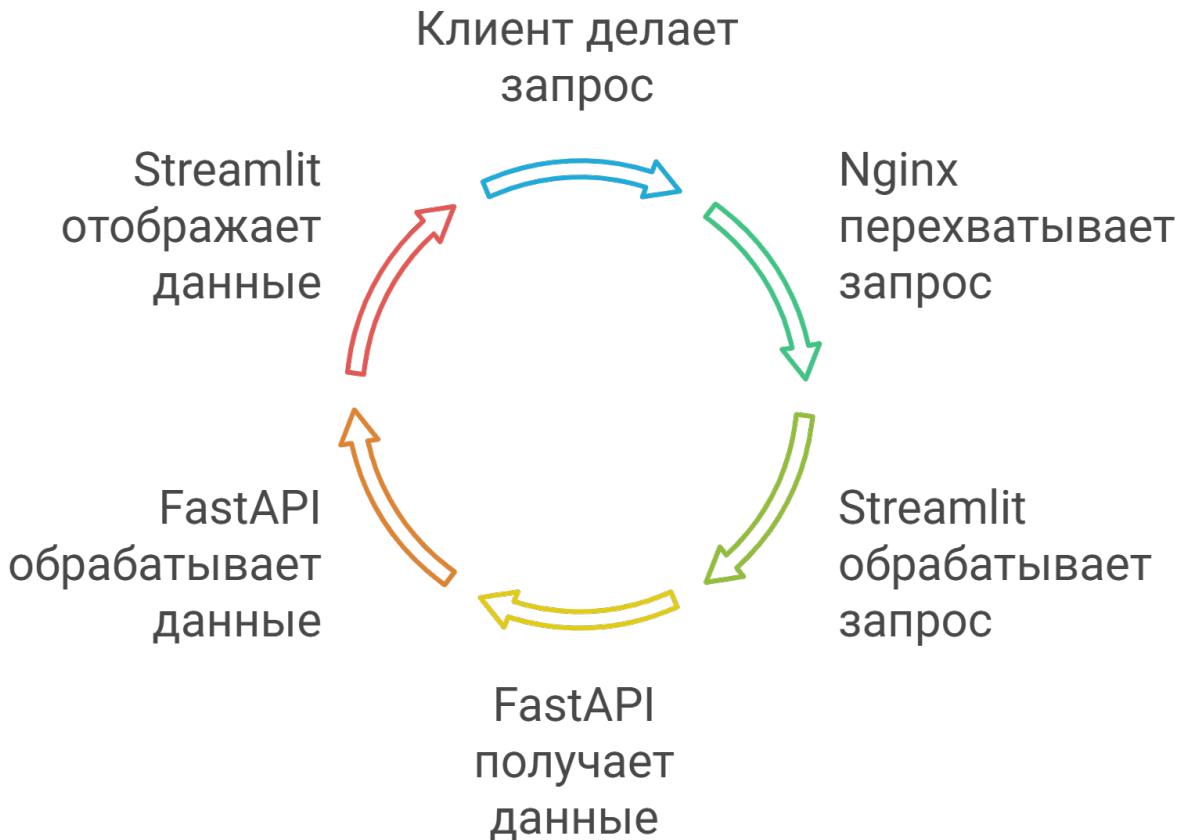


Рисунок 27: Цикл обработки запросов.

### 6.3. Общая структура сервиса

Сервис представлен тремя страницами:

- Акции - страница позволяет выбрать акцию и модель, а также посмотреть прогнозы и оценить их точность.
- Среднее - на этой странице можно увидеть, насколько хорошо та или иная модель справляется с предсказаниями и какие в среднем ошибки она имеет.
- Топ моделей - страница предоставляет информацию о местах, которые каждая из мо-

Таблица 7: Содержание запросов.

Запрос	Суть запроса	Принимает	Возвращает
/secids/	Получение тикеров компаний	-	Список доступных тикеров
/models/	Получение моделей	-	Список обученных моделей
/predict/{model}/{secid}/	Получение данных для конкретного тикера и модели	Тикер акции, модель	Прогноз цен акции, ошибки на каждый день, данные для соответствующих графиков
/predict_mean/{model}/{secid}/	Получение среднего графика для определенной модели на всех тикерах или старше пяти лет	Модель, горизонт	Прогноз средних ошибок на каждый день, данные для соответствующих графиков
/top_models/{duration}/	Среднее занимаемое место моделью в качестве прогнозов	Горизонт	Данные для графиков средних мест моделей на разные горизонты

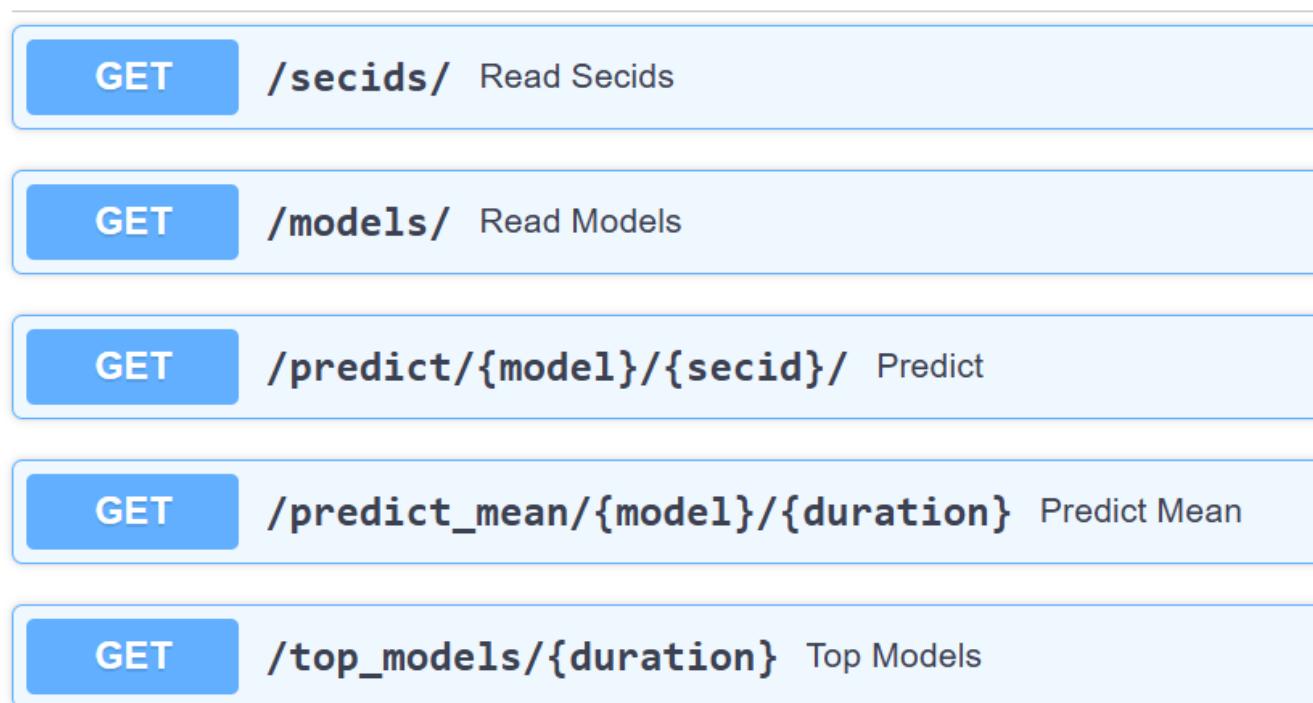


Рисунок 28: Доступные endpoint'ы.

делей занимает на разных горизонтах для множества доступных акций.

Каждая страница состоит из двух областей: левой боковой панели (сайдбара) с настройками и основной области, в которой происходит визуализация результатов.

## 6.4. Функционал

На странице с акциями можно выбрать тикер и модель, после чего, как на «Рисунке 29» будет отображена таблица с прогнозами и ошибками, график цен, а также графики

валидации с прогнозом и ошибки по периодам, как на «Рисунке 30».

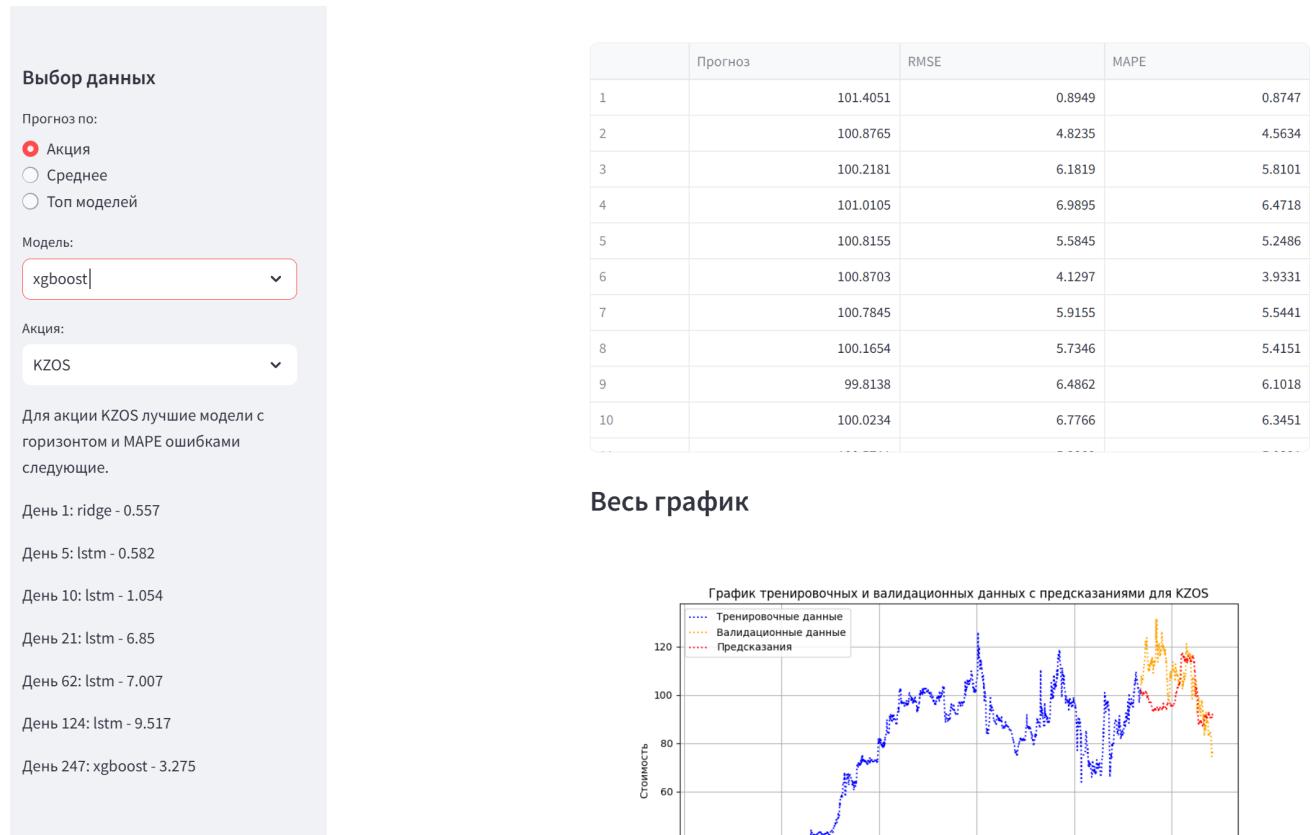


Рисунок 29: Страница акций: основные данные.

Для большего удобства слева продемонстрировано, на каких горизонтах какая модель для данной бумаги оказалась точнее.

Страница со средними результатами показывает полученные средний данные модели на множестве акций. На «Рисунке 31» видна таблица со средними ошибками и соответствующие графики, дополненные на «Рисунке 32».

И последняя страница рассказывает о том, какая модель на каком горизонте в среднем лучше показывала свои результаты, то есть как часто она занимала лучшие места. На «Рисунке 33» и «Рисунке 34», видны лидеры и то, как они меняются со временем.

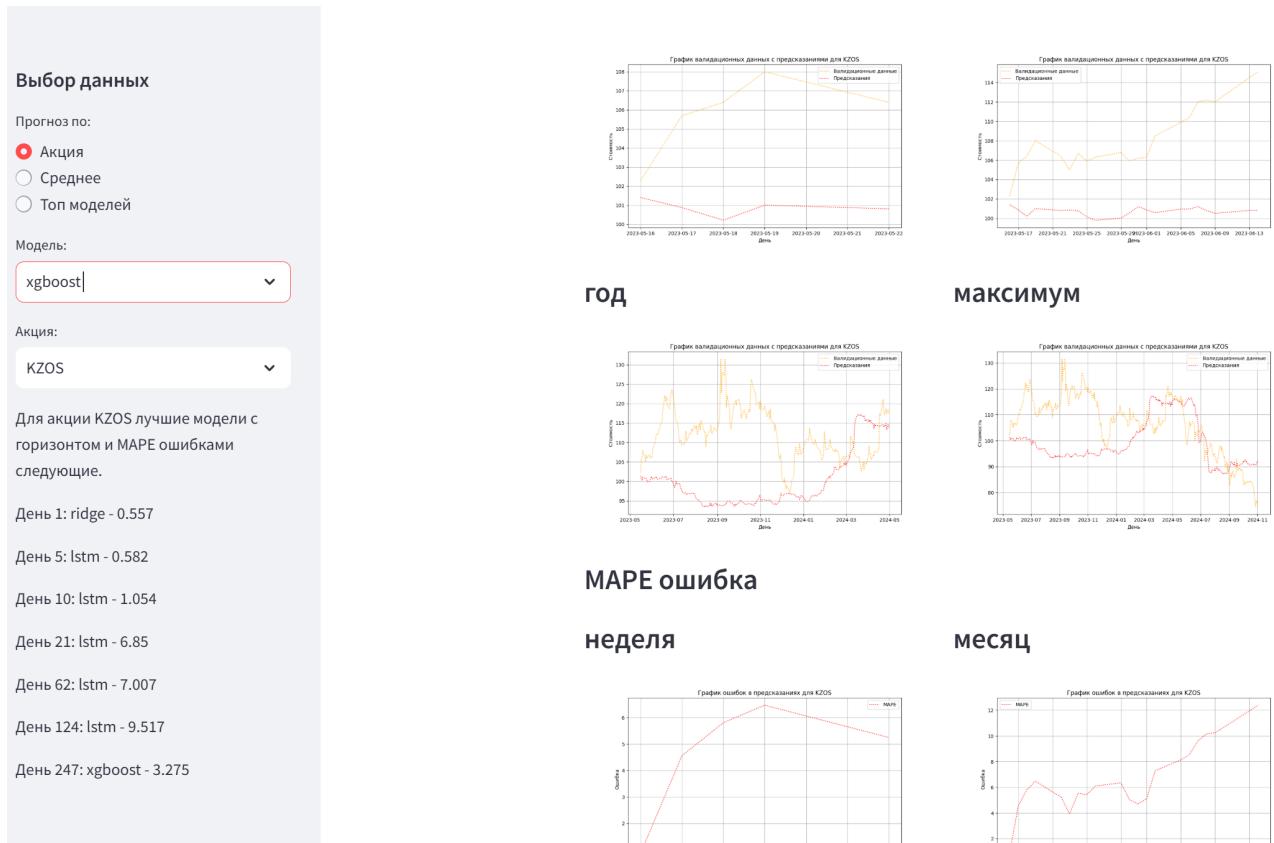


Рисунок 30: Страница акций: данные с удобными отрезками.

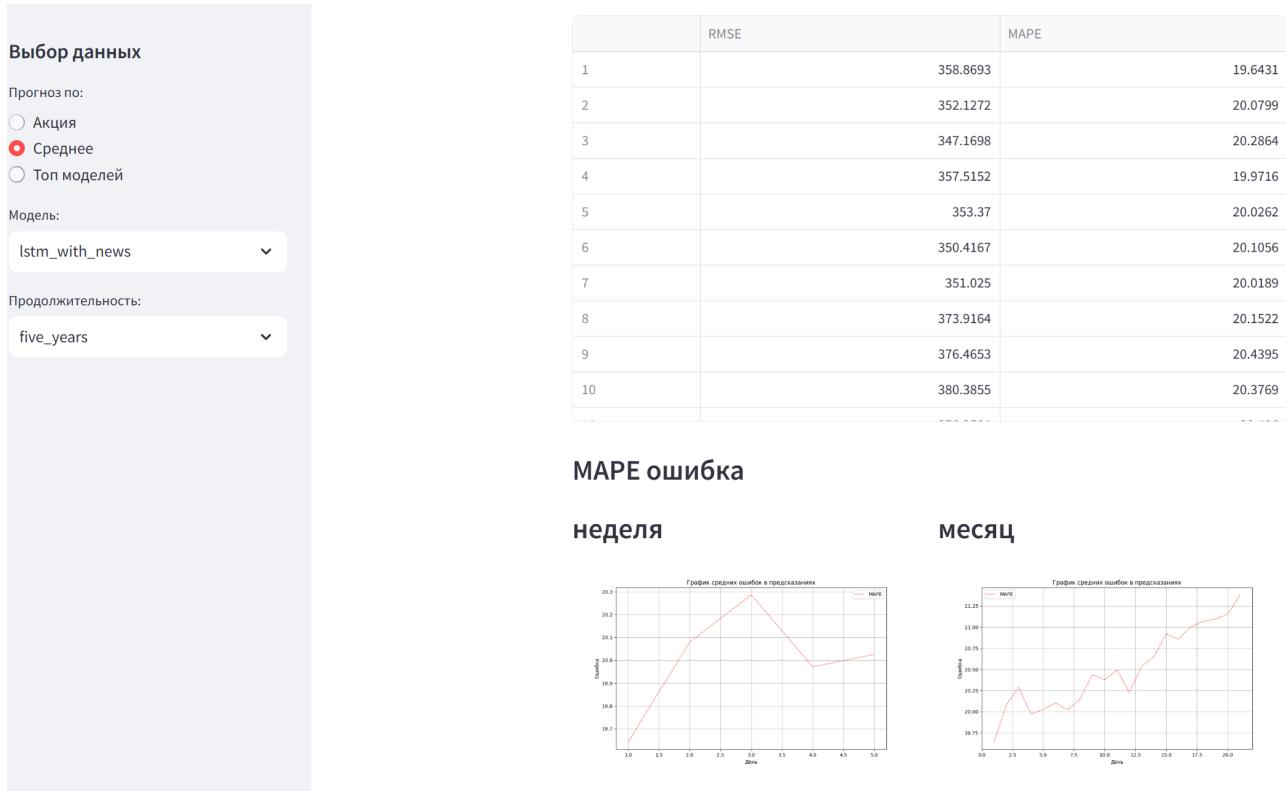


Рисунок 31: Страница средних данных по модели.

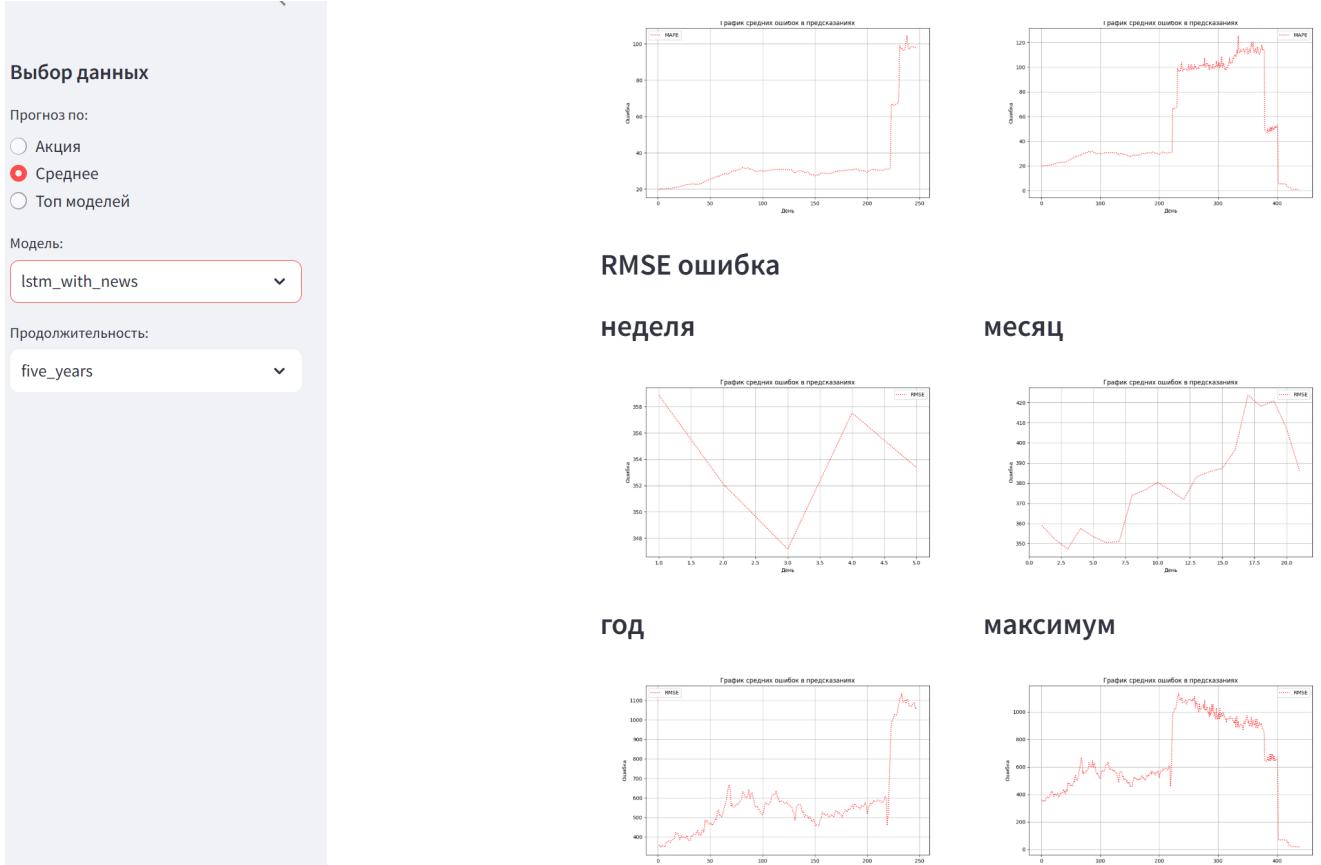


Рисунок 32: Страница графиков средних данных по модели.

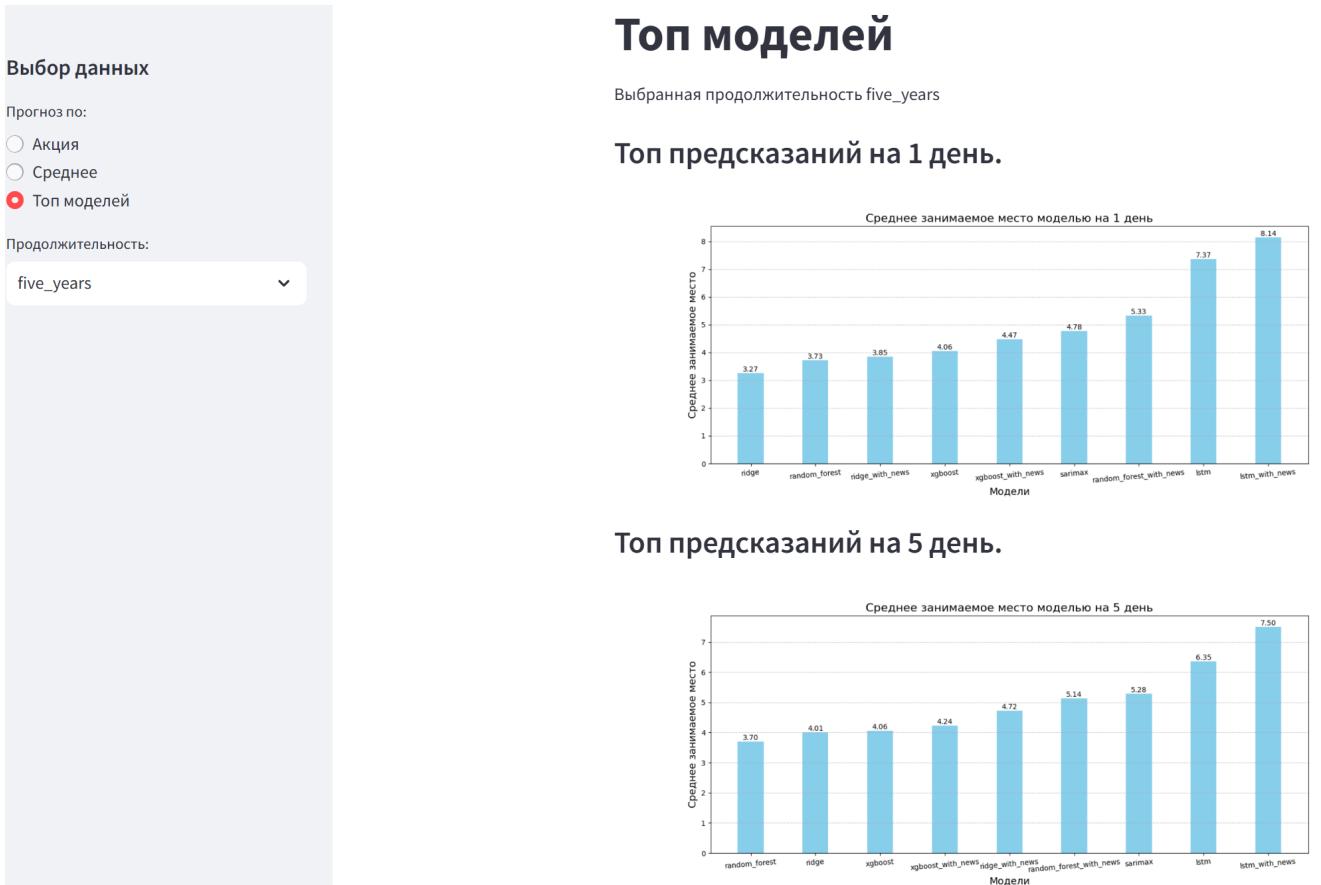
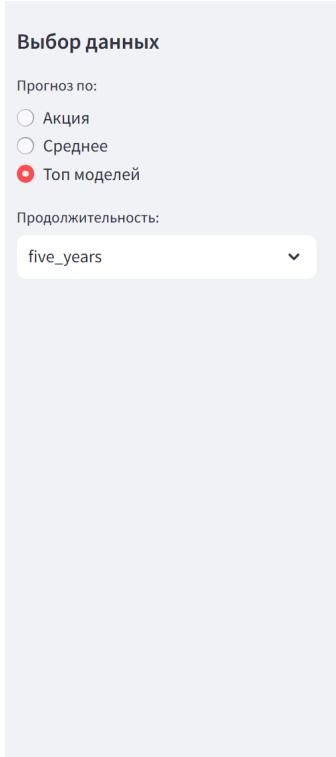
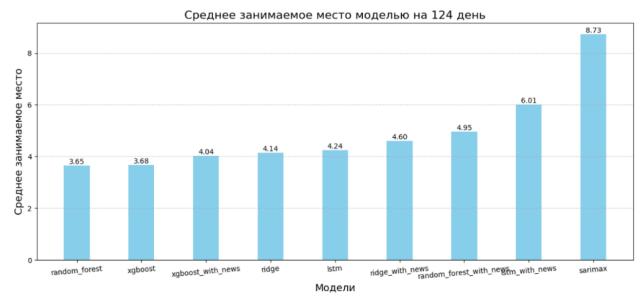


Рисунок 33: Страница топа моделей на коротком горизонте.



Топ предсказаний на 124 день.



Топ предсказаний на 247 день.

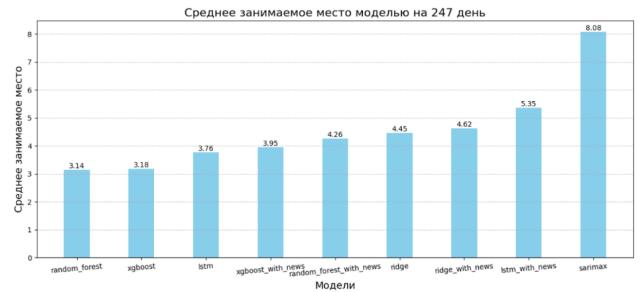


Рисунок 34: Страница лучших моделей на длинном горизонте.

## 7. Как можно улучшить результат

Для получения более точных прогнозов я вижу несколько вариантов.

Первое. Лучше выбирать бумаги, в которых вы лучше разбираетесь, то есть иметь экспертизу в той области, в которой работает компания. Это поможет вам самим понять, правдоподобен ли прогноз, учитывая текущее состояние компании, её окружение и менеджмент. Верите ли вы в такое и насколько соответствует вашему риск-профилю.

Второе. По возможности увеличивать количество данных по её цене, чтобы тренировочных данных было много. И постараться более здраво заполнять имеющиеся пропуски.

Третье. Использовать больше данных. Добавлять в модели параметры EBITDA, P/E и другие. Использовать известные дневные минимальные и максимальные цены, объём торгов и их стоимость.

Четвёртое. Лучше и точнее определять новостной фон. Учитывать, что по общим новостям тяжелее выявлять информацию о мелких или непопулярных компаниях. Возможно, лучше дополнять новостями для конкретных эмитентов.

Пятое. Попробовать подобрать параметры моделей лучше и построить более хорошую LSTM-модель.

## 8. Заключение

В заключение моего исследования хотелось бы отметить, что в рамках данной выпускной квалификационной работы мною были собраны, обработаны, проанализированы и подготовлены данные по 215 акциям и более чем 2 млн. новостей. Были обучены такие модели, как SARIMAX, Ridge, Random Forest, XGBoost и LSTM. Результаты были проанализированы и представлены в удобном формате в виде сервиса, как по отдельным бумагам, так и средние показатели для получения общей картины.

Результаты работы следующие. Лучшей моделью оказался случайный лес, он проигрывает Ridge только на начальном этапе, но уже через пару дней показывает более точные прогнозы. Кроме того, он является более устойчивым на большем горизонте. Также градиентный бустинг очень хорош, из полученных результатов видно, что данная модель немного хуже случайного леса, но все равно достаточно точна и тоже более стабильна в будущем. По поводу ARIMA-подобных моделей, не думаю, что они достаточно конкурентоспособны по сравнению с остальными. А LSTM в моей работе выглядит наиболее стабильной, но не столь точной, вполне возможно, что при построении хорошей нейронной сети результаты будут более точными.

Хотя в среднем модели с использованием новостей показали результат немного хуже, чем без них, я считаю, что новостные данные положительно влияют на результаты моделей, потому что они часто показывали хороший результат. Только для этого я бы советовал собирать как можно больше данных для вашей акции. Чем больше новостей, влияющих на ваши бумаги, вы сможете найти, тем больше информации модель сможет из этих данных получить. Если новостных данных по эмитенту не так много, то лучше рассматривать модели без использования новостей.

И напоследок, хотел бы сказать, что не стоит слепо верить прогнозам моделей, тем более на большом горизонте. Лучше стабильно инвестировать, а не пытаться предугадать, что произойдет на рынке завтра, ведь рано или поздно все пойдет не по плану. Рынок резко рухнет или поднимется, что предугадать невозможно. Вы определенно не успеете среагировать, что может привести к проблемам. Но, не пытаясь переиграть рынок и имея стратегию, вы будете готовы к любым взлётам и падениям.

## Список литературы

- [1] Fin-Plan. *Список акций Мосбиржи*. URL: <https://fin-plan.org/stocks/rus>. (дата обращения: 18.05.2025).
- [2] Finam. *Новости Finam*. URL: <https://www.finam.ru/>. (дата обращения: 18.05.2025).
- [3] Interfax. *Число физлиц с брокерскими счетами на МосБирже в 2024 году выросло на 5,4 млн*. URL: <https://www.interfax.ru/business/1003027?ysclid=mau9sga6le282523374>. (дата обращения: 18.05.2025).
- [4] Investiong. *Новости Investing.com*. URL: <https://ru.investing.com/>. (дата обращения: 18.05.2025).
- [5] Juvenal José Duarte<sup>1</sup> Sahudy Montenegro González<sup>1</sup> José César Cruz Jr. *Predicting Stock Price Falls Using News Data: Evidence from the Brazilian Market*. URL: <https://www.ncbi-hub.ru/10.1007/s10614-020-10060-y?ysclid=mav835u6bv826669042>. (дата обращения: 18.05.2025).
- [6] Kaggle. *Kaggle*. URL: <https://www.kaggle.com/>. (дата обращения: 18.05.2025).
- [7] МОЕХ. *Новости Московской биржи*. URL: <https://www.moex.com/ru/news/>. (дата обращения: 18.05.2025).
- [8] SambaNova. *LLM для оценки новостей*. URL: <https://cloud.sambanova.ai/dashboard>. (дата обращения: 18.05.2025).
- [9] SMART-LAB. *Новости SMART-LAB*. URL: <https://smart-lab.ru/news/>. (дата обращения: 18.05.2025).
- [10] БКС. *Новости БКС*. URL: <https://bcs-express.ru/category>. (дата обращения: 18.05.2025).
- [11] Шамраева В.В. *МАТЕМАТИЧЕСКИЕ МЕТОДЫ ПРОГНОЗИРОВАНИЯ ИЗМЕНЕНИЯ ЦЕНЫ АКЦИЙ И ИХ РЕАЛИЗАЦИЯ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ*. URL: <https://s.fundamental-research.ru/pdf/2024/11/43718.pdf>. (дата обращения: 18.05.2025).
- [12] Рудзейт О.Ю. Зайнетдинов А.Р. Недяк А.В. Рагулин П.Г. *Прогнозирование цены акции с помощью метода регрессионного анализа*. URL: <https://resources.today/PDF/14INOR420.pdf?ysclid=matxarpoky444560454>. (дата обращения: 18.05.2025).
- [13] РБК. *Новости РБК*. URL: <https://www.rbc.ru/>. (дата обращения: 18.05.2025).

- [14] РИА. *Новости РИА*. URL: <https://ria.ru/>. (дата обращения: 18.05.2025).
- [15] РИА. *Ссылки на новости*. URL: <https://ria.ru/20250518/>. (дата обращения: 18.05.2025).