

**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ  
РОССИЙСКОЙ ФЕДЕРАЦИИ**  
Федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«Московский государственный технический университет имени Н.Э. Баумана  
(национальный исследовательский университет)»

**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА**  
**по курсу**  
**«Data Science»**

Слушатель

Новицкая Виктория Денисовна

Москва, 2023

## СОДЕРЖАНИЕ

|   |    |
|---|----|
| ВВЕДЕНИЕ .....                            | 1  |
| 1 Аналитическая часть .....               | 3  |
| 1.1 Постановка задачи .....               | 3  |
| 1.2 Описание используемых методов .....   | 5  |
| 1.2.1. Логистическая регрессия .....      | 5  |
| 1.2.2. Метод опорных векторов (SVM) ..... | 6  |
| 1.2.3. Случайный лес .....                | 6  |
| 1.2.4. Градиентный бустинг .....          | 7  |
| 1.2.5. Анализ тональности .....           | 8  |
| 1.3 Разведочный анализ данных .....       | 8  |
| 2 Практическая часть .....                | 13 |
| 2.1 Предобработка данных .....            | 13 |
| 2.2 Разработка и обучение модели .....    | 19 |
| 2.3 Нейронные сети .....                  | 26 |
| 2.4 Разработка приложения .....           | 30 |
| 2.5 Создание удаленного репозитория ..... | 31 |
| ЗАКЛЮЧЕНИЕ .....                          | 32 |
| Библиографический список .....            | 33 |

## ВВЕДЕНИЕ

Тема данной работы - Анализ новостных потоков для прогнозирования движения цены цифровых активов.

Цель данной работы состоит в изучении взаимосвязи между реакцией на движение цен криптовалют BTC и ETH после публикации новостей применяя дополнительные признаки на основе временного сдвига. Основной задачей является прогнозирование движения цен после выхода новости с использованием моделей машинного обучения, включая нейронные сети. Разработанная модель должна прогнозировать реакцию движения цены покупка/продажа после выхода новости через время.

Учитывая высокую волатильность цифровых активов, точное прогнозирование цен является сложной и актуальной проблемой. В последние годы эти активы привлекли внимание инвесторов и стали значимой частью их инвестиционных портфелей. Однако, высокая волатильность цен на эти активы затрудняет прогнозирование рисков и оценку потенциальной прибыли. В свете этого, актуальность разработки эффективных методов прогнозирования движения цен на цифровые активы возрастает.

В дальнейшем перспективы развития и использования цифровых валют, а также технологии, которые стоят за цифровыми валютами, которые имеют огромное количество кейсов в использовании, применяемые в области финансов, где также имеет развитие в сфере искусств, игр, частично логистики при легализации такой индустрии, которая имеет большое количество прорывных технологий настанет новой этап для экономики РФ и целого мира.

В ходе исследования будут рассмотрены различные подходы к анализу текстовых данных, а также проведен анализ имеющихся данных для определения корреляции между новостными событиями и изменением цен на цифровые активы.

Данная работа включает следующие этапы: обзор существующих исследований в области анализа новостных потоков и прогнозирования цен

цифровых активов, выявление основных методов и инструментов анализа текстовых данных, применяемых для данной цели, оценка их применимости и эффективности. В результате проведенного исследования будут определены ключевые факторы, влияющие на цены цифровых активов, а также предложены наиболее эффективные методы прогнозирования движения цены, которые могут быть использованы для минимизации рисков и оптимизации инвестиционных решений.

# 1 Аналитическая часть

## 1.1 Постановка задачи

В данной работе исследуются новостные потоки и исторические данные биткойна и эфириума. Данные по тематическим новостям были собраны с сайта Reuters, а исторические данные с сайта Binance.

В качестве исходных данных были использованы три датасета: данные по новостям (news), содержащие 2466 строк с заголовками, датами и текстами новостных статей за период с 14 мая 2021 года по 20 апреля 2023 года; данные по цифровым активам (btcusdt\_1m и ethusdt\_1m), содержащие информацию о цене открытия, максимальной и минимальной цене, цене закрытия и объеме торгов за каждую минуту за тот же период.

Для достижения поставленной цели были проведены следующие этапы анализа данных:

Предварительная обработка данных: были удалены дубликаты, пропуски и выбросы. Для текста: удаление стоп-слов, токенизация и лемматизация. Также проведен анализ текста новостей, чтобы извлечь ключевые слова, темы и сущности. Использовался метод обработки естественного языка (NLP).

Исследование свойств каждого признака: для каждого признака были проанализированы распределения, дисперсии, средние значения, минимальные и максимальные значения, а также выявлены аномалии и выбросы.

Исследование зависимостей между признаками: были проанализированы корреляционные матрицы и построены графики зависимости между признаками.

Визуализация данных: были построены графики изменения цены цифровых активов и количества новостных статей.

Датасеты по цифровым активам содержат по 6 признаков:

- Метку времени (timestamp),
- Цену открытия (open\_price),
- Максимальную цену (high\_price),
- Минимальную цену (low\_price),

- Цену закрытия (close\_price)
- Объем торгов (volume).

Общий объем выборки составляет 1017611 строк.

Датасет с новостями содержит 3 признака:

- Заголовок (Headline),
- Дату (Date)
- Текст новостных статей (Text).

Всего в датасете содержится 2466 строк.

После предварительной обработки данных

Данные по новостям и цифровым активам будут использованы для построения модели, которая позволит прогнозировать движение цен цифровых активов на основе новостных потоков.

Исследовалась зависимость между новостными потоками и ценами цифровых активов, используя статистические методы, корреляционный анализ и методы машинного обучения. Были определены ключевые факторы, влияющие на изменение цен, и выявлены возможные закономерности и тренды. Так же требуется разработать приложение, делающее удобным использование данных моделей специалистом предметной области.

Таблица 1 — Описание всех признаков датасетов

| Название   | Файл                      | Тип данных | Непустых значений | Уникальных значений |
|------------|---------------------------|------------|-------------------|---------------------|
| Headline   | news                      | object     | 2466              | 2466                |
| Date       | news                      | object     | 2466              | 2466                |
| Text       | news                      | object     | 2466              | 2466                |
| timestamp  | btcusdt_1m/<br>ethusdt_1m | object     | 1017611           | 1017611             |
| open_price | btcusdt_1m/<br>ethusdt_1m | float64    | 1017611           | 1017611             |

|             |                           |         |         |         |
|-------------|---------------------------|---------|---------|---------|
| high_price  | btcusdt_1m/<br>ethusdt_1m | float64 | 1017611 | 1017611 |
| low_price   | btcusdt_1m/<br>ethusdt_1m | float64 | 1017611 | 1017611 |
| close_price | btcusdt_1m/<br>ethusdt_1m | float64 | 1017611 | 1017611 |
| volume      | btcusdt_1m/<br>ethusdt_1m | float64 | 1017611 | 1017611 |

## 1.2 Описание используемых методов

### 1.2.1. Логистическая регрессия

Логистическая регрессия - это статистический метод для анализа набора данных, в котором есть одна или несколько независимых переменных, которые определяют результат. Результат измеряется с помощью дихотомической переменной (0 или 1). В контексте нашей задачи, логистическая регрессия может использоваться для определения вероятности того, что цена цифрового актива увеличится или уменьшится, основываясь на новостных данных.

Достоинства:

- Простота и легкость интерпретации;
- Хорошо работает с линейно разделимыми данными;
- Быстрая обучающая скорость и прогнозирование.

Недостатки:

- Предполагает линейную связь между признаками и логарифмом шансов;
- Может страдать от мультиколлинеарности и переобучения.

Области применения: Бинарная классификация, прогнозирование вероятностей.

### **1.2.2. Метод опорных векторов (SVM)**

Метод опорных векторов (SVM) - это мощный алгоритм машинного обучения, используемый для классификации и регрессии. SVM работает путем создания гиперплоскости, которая максимизирует расстояние между двумя классами. В контексте нашей задачи, SVM может использоваться для определения влияния новостных данных на изменение цены цифрового актива.

Достоинства:

- Хорошо работает с малыми наборами данных;
- Эффективно работает в пространствах высокой размерности;
- Может использовать различные ядра для построения нелинейных разделяющих гиперплоскостей.

Недостатки: Требуется настройка гиперпараметров, таких как параметр регуляризации и ядро.

Более медленный в обучении и прогнозировании по сравнению с другими методами машинного обучения, который используется для предсказания численных значений. Для анализа новостных потоков и прогнозирования цен цифровых активов, регрессия может быть использована для создания модели, которая будет прогнозировать цены цифровых активов на основе новостных данных. Регрессия может быть линейной, множественной, полиномиальной, решающего дерева и др. Основное преимущество регрессии состоит в том, что она позволяет предсказывать будущие значения цен на основе уже имеющихся данных.

### **1.2.3. Случайный лес**

Случайный лес - это ансамблевый метод машинного обучения, основанный на деревьях решений. Он строит несколько деревьев решений и объединяет их результаты для повышения точности и устойчивости к переобучению. В контексте нашей задачи, случайный лес может использоваться для определения влияния новостных данных на изменение цены цифрового актива.



Достоинства:

- Автоматический отбор признаков и определение их важности;
- Устойчив к переобучению, особенно при использовании большого количества деревьев;
- Хорошо работает с нелинейными данными и пропусками в данных.

Недостатки: Может быть медленным при работе с большими наборами данных и могут быть сложными и трудными для интерпретации.

Области применения: Классификация, регрессия и определение важности признаков. Метод может работать с нелинейными данными и пропусками в данных.

#### **1.2.4. Градиентный бустинг**

Градиентный бустинг - это ансамблевый метод машинного обучения, который строит несколько слабых моделей (обычно деревьев решений) последовательно, исправляя ошибки предыдущих моделей. В контексте нашей задачи, градиентный бустинг может использоваться для определения влияния новостных данных на изменение цены цифрового актива.

Достоинства:

- Обычно дает лучшие результаты по сравнению с другими ансамблевыми методами.
- Может автоматически определять важность признаков.
- Хорошо работает с нелинейными данными.

Недостатки:

- Более подвержен переобучению, особенно при использовании небольшого количества данных.

Может быть медленным в обучении и прогнозировании.

Требует настройки гиперпараметров, таких как глубина дерева, скорость обучения и количество деревьев.

Области применения:

Классификация, регрессия и определение важности признаков.

### **1.2.5. Анализ тональности**

Анализ тональности текста является важным методом для анализа новостных потоков. Он используется для определения эмоциональной окраски текста и выявления настроений в отношении конкретных событий, компаний или активов. Для анализа тональности могут использоваться различные методы, такие как правила на основе словарей, методы машинного обучения и глубокое обучение. Для обработки текстовых данных могут использоваться методы предобработки текста, такие как токенизация, удаление стоп-слов и лемматизация.

### **1.3 Разведочный анализ данных**

С учетом высокой частоты исторических данных по цифровым активам (каждая минута) предлагается применить ресемплирование данных по часу для более обобщенного анализа (`btcusdt_1m` и `ethusdt_1m`). Это позволило снизить уровень шума и сделать анализ более наглядным и информативным. Обнаружились 7 пропущенных значений в каждом из датасетов. Поскольку пропущенных значений немного, можно их удалить. В итоге датафреймы `eth_1h` и `btc_1h` количество значений в переменных равно 16962 строк после удаления пропусков. Все признаки имеют тип `float64`, то есть вещественный.

Для каждой переменной в датасетах (цена открытия, максимальная и минимальная цена, цена закрытия, объем торгов) построены гистограммы распределения. Это позволило оценить форму распределения, наличие асимметрии и выявить возможные выбросы.

Построила гистограммы распределения, которые представлены на рисунке 1-2 для каждой переменной на главной диагонали графика. Эти графики позволяют оценить степень взаимосвязи между параметрами и определить наличие выбросов в данных .

---

Histogram of BTCUSDT

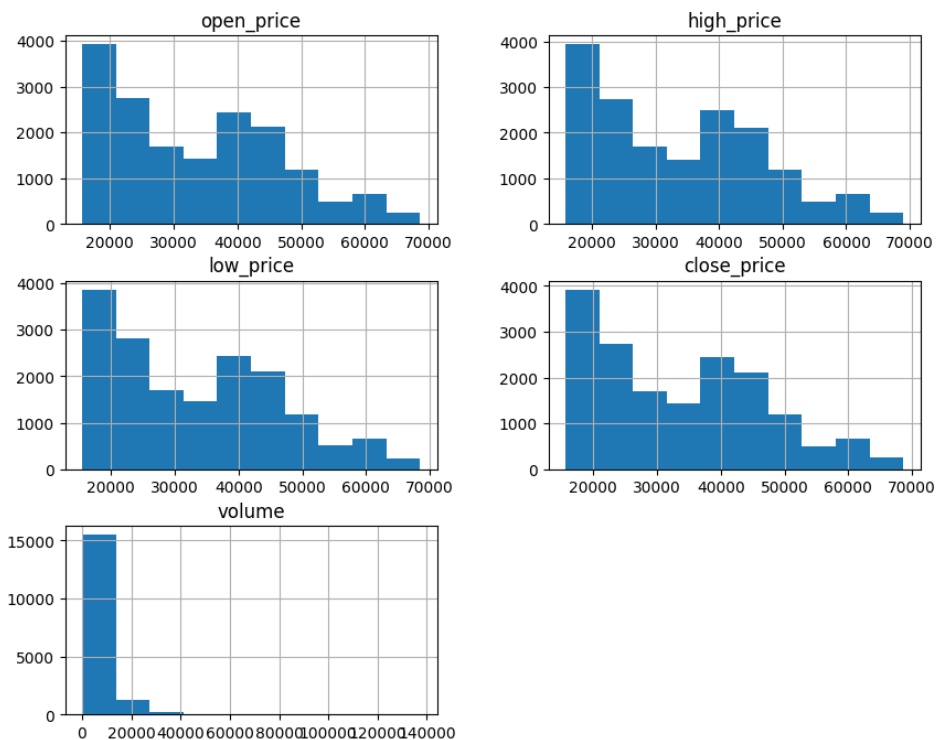


Рисунок 1 -Гистограммы распределения BTCUSDT

---

Histogram of ETHUSDT

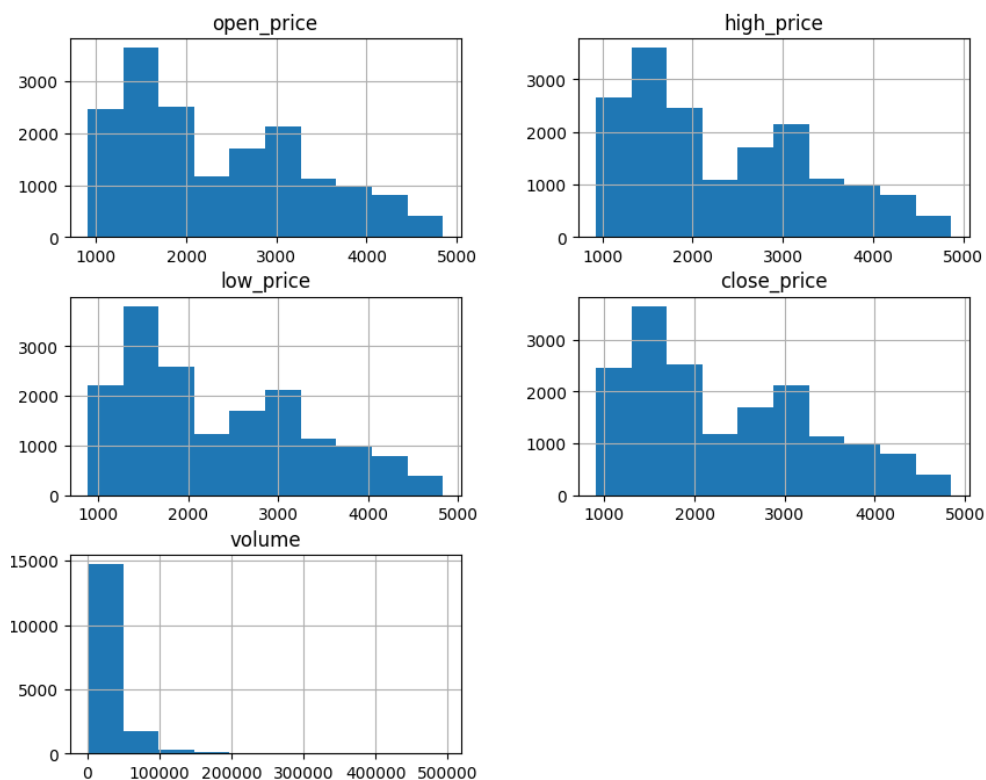


Рисунок 2- Гистограммы распределения ETHUSDT

На гистограммах распределения присутствует небольшая скошенность данных. Это означает, что распределение данных не является симметричным, и имеет тенденцию к более высоким значениям. Относительно других столбцов в таблице, можно заметить, что объем торгов (volume) имеет значительную положительную скошенность и острый пик распределения. Это говорит о том, что большинство значений объема торгов находятся на низком уровне, а на более высоких уровнях находится меньшее количество значений, и они сильно концентрируются вокруг среднего значения и имеют малую дисперсию.

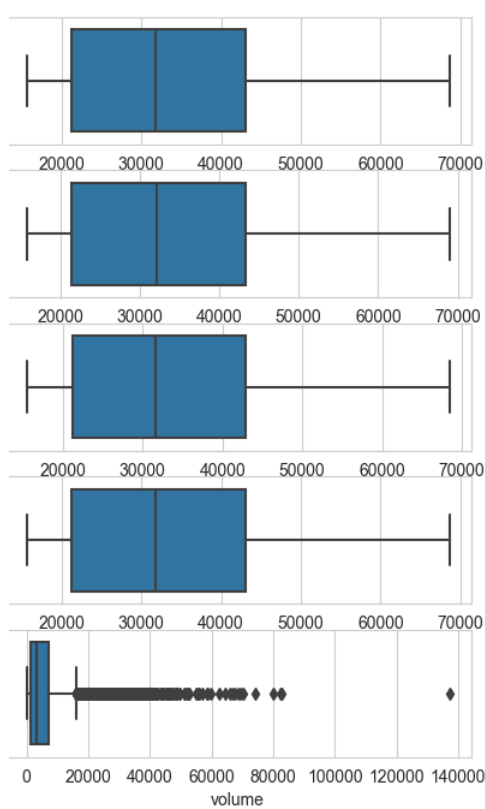


Рисунок 3- Диаграммы ящика с усами BTCUSDT

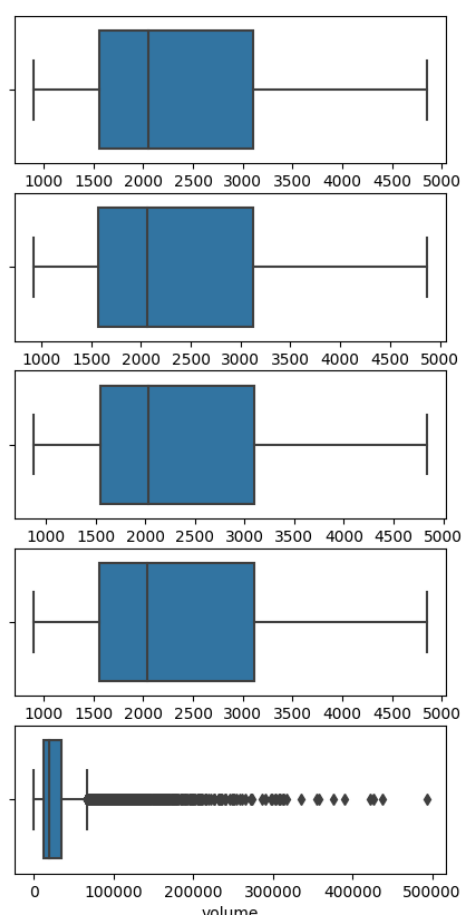


Рисунок 4 - Диаграммы ящика с усами ETHUSDT

На основании диаграмм ящика с усами, можно сделать следующие выводы:

- в колонке `open_price` и `close_price` нет выбросов, распределение цен достаточно равномерное;
- в колонках `high_price` и `low_price` нет выбросов, распределение цен достаточно равномерное;
- в колонке `volume` есть несколько очень больших выбросов, которые сильно смещают распределение. Это еще подтверждает, что средние значения больше медиан, что подтверждает наличие выбросов с большими значениями.

Для дальнейшего анализа таких данных необходимо их исключить.

Для исключения выбросов в колонке `volume` используется логарифмическая трансформация, так как данные имеют сильную скошенность, перед применением методов для удаления выбросов, методом межквартильного расстояния и 3-х сигм.

Применив эти методы на датасете было найдено:

Метод 3-х сигм: Ethereum выбросов - 5489, Bitcoin выбросов - 1174

Метод межквартильных расстояний: Ethereum выбросов - 94223, Bitcoin выбросов – 79103

Так же нас интересует описательная статистика датасета. Она представлена в таблице 2 и 3. Она в численном виде отражает то, что мы видим на гистограммах.

Таблица 2 - Описательная статистика признаков датасета btc\_1h

|       | open_price | high_price | low_price | close_price | volume     |
|-------|------------|------------|-----------|-------------|------------|
| count | 16962.000  | 16962.000  | 16962.000 | 16962.000   | 16969.000  |
| mean  | 33551.562  | 33720.968  | 33374.829 | 33550.293   | 5604.734   |
| std   | 13074.528  | 13146.971  | 12996.847 | 13074.007   | 6552.025   |
| min   | 15648.230  | 15769.990  | 15476.000 | 15649.520   | 0.000000   |
| 25%   | 21294.260  | 21382.587  | 21198.552 | 21295.115   | 1509.958   |
| 50%   | 31871.880  | 32105.935  | 31682.945 | 31868.020   | 3210.236   |
| 75%   | 43130.635  | 43373.947  | 42900.525 | 43129.385   | 7345.691   |
| max   | 68635.120  | 69000.000  | 68451.190 | 68633.690   | 137207.189 |

Таблица 3 - Описательная статистика признаков датасета eth\_1h

|       | open_price | high_price | low_price | close_price | volume     |
|-------|------------|------------|-----------|-------------|------------|
| count | 16962.000  | 16962.000  | 16962.000 | 16962.000   | 16969.000  |
| mean  | 2357.265   | 2372.000   | 2341.637  | 2357.159    | 28119.207  |
| std   | 983.614    | 989.443    | 977.175   | 983.563     | 28480.147  |
| min   | 904.260    | 928.360    | 881.560   | 904.250     | 0.000000   |
| 25%   | 1558.157   | 1566.535   | 1549.277  | 1558.155    | 12022.119  |
| 50%   | 2046.480   | 2061.220   | 2028.245  | 2045.455    | 19472.262  |
| 75%   | 3109.315   | 3126.642   | 3091.907  | 3109.325    | 33622.763  |
| max   | 4846.940   | 4868.000   | 4833.190  | 4846.710    | 493227.883 |

Для визуального анализа взаимосвязи между активами BTCUSDT и ETHUSDT построена матрица корреляции для объединенного датафрейма цифровых активов. На рисунке 5 видно корреляцию между переменными, которые близки к 1, то это говорит о сильной положительной линейной связи, и так как у объема отрицательная корреляция, то это может затруднить для прогнозирования движения цен, в дальнейшем не будем учитывать данное значение.

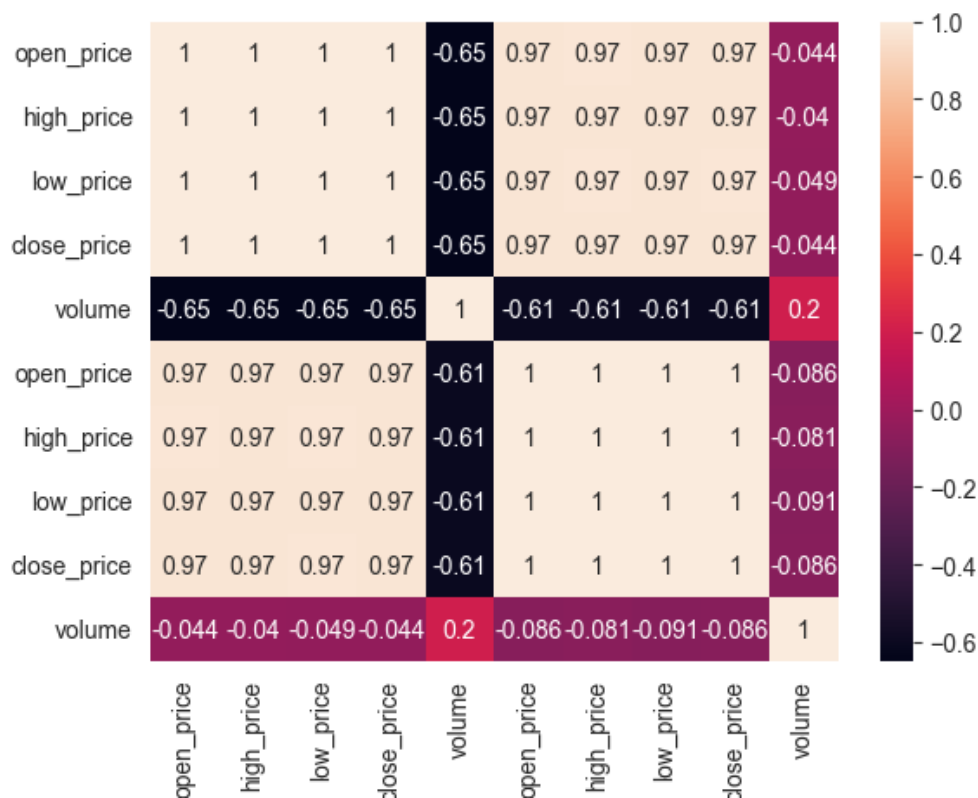


Рисунок 5 - Матрица корреляции

## 2 Практическая часть

### 2.1 Предобработка данных

Первым шагом была проведена предобработка данных, включающая в себя удаление стоп-слов, очистки текстовых данных от пунктуации, токенизацию и лемматизацию текста. Также проведен анализ текста новостей, чтобы извлечь ключевые слова, темы и сущности. Для этой цели была использована библиотека NLTK (Natural Language Toolkit) . Результат списка слов, отсортированный по частоте появления в тексте:

– Многие слова связаны с финансами и инвестициями, такие как "bank", "trading", "company", "fund", "business", "lender", "liquidity", "balance" и "portfolio".

– Некоторые слова указывают на временные рамки, такие как "thursday", "2021", "november" и "2023".

– Несколько слов связаны с географией и государствами, такие как "russia", "american" и "ambassador".

– Есть слова, связанные с технологиями и криптовалютами, такие как "token", "nonfungible" и "block".

– Также есть слова, которые указывают на события или действия, такие как "sanctions", "meeting", "filing", "covid19", "lawmakers" и "scrutiny".

Этот анализ дает представление о том, какие темы могут быть обсуждаемы в новостных статьях. Основываясь на этих данных, можно сделать вывод, что новости в основном связаны с финансовыми рынками, инвестициями, политикой и криптовалютами.

На графике в соответствии с рисунком 6 позволяет оценить, что чаще всего новостные события выходят в ноябре и меньше всего в апреле.



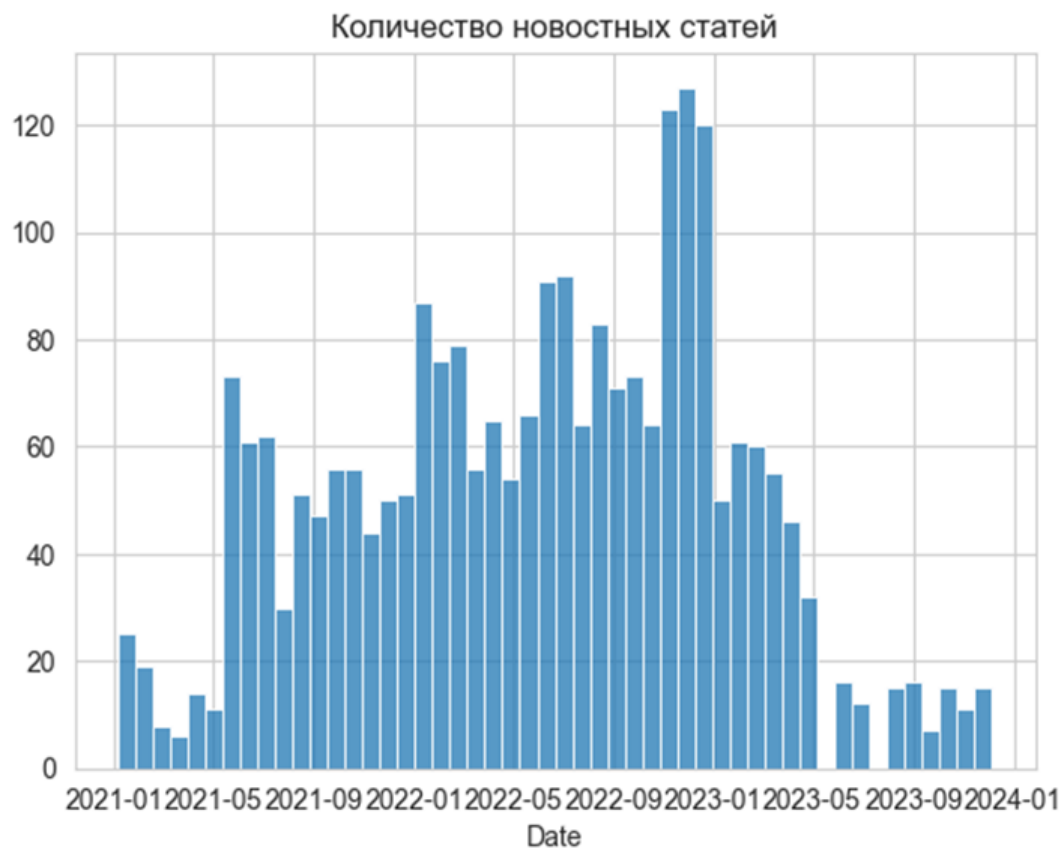


Рисунок 6 - Гистограмма новостей

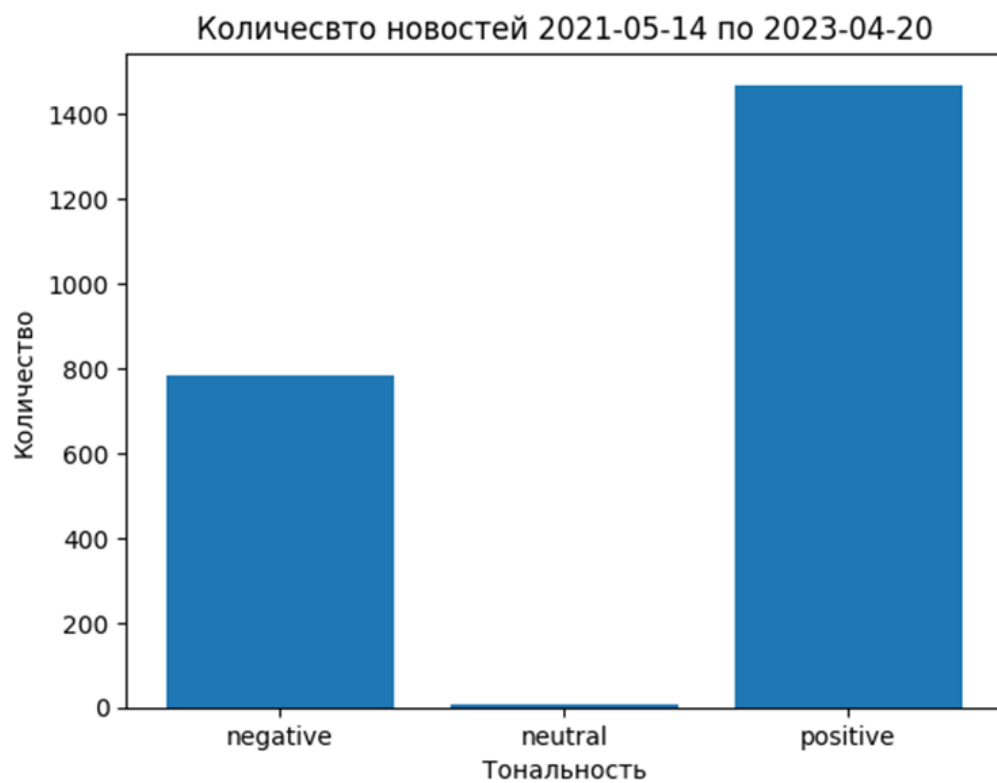


Рисунок 7 - Тональность новостей

Количество новостей принявшее окраску тональности:

- positive -1598
- negative -861
- neutral -7

Так как новостей с нейтральной тональностью всего 7, то для лучшей работы модели, данные новости будут удалены. Признак –‘Sentiment’, содержащий значение 1, если новость положительная, и 0, если новость отрицательная, превратили эти значения с помощью LabelEncoder().

В данном разделе приводятся графики распределения с рисунка 8 -11 для каждого признака до и после нормализации.

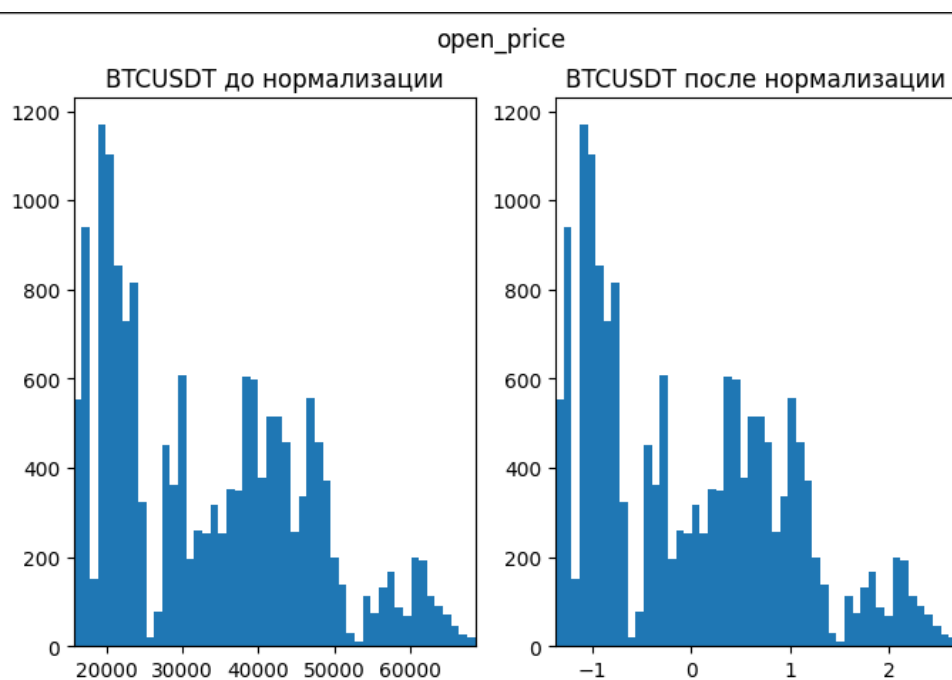


Рисунок 8 –график до и после нормализации признака ‘open\_price’

Минимальное значение open\_price до нормализации: 15648.23

Максимальное значение open\_price до нормализации: 68635.12

Минимальное значение open\_price после нормализации: -1.3693694621621764

Максимальное значение open\_price после нормализации: 2.6834308589373888

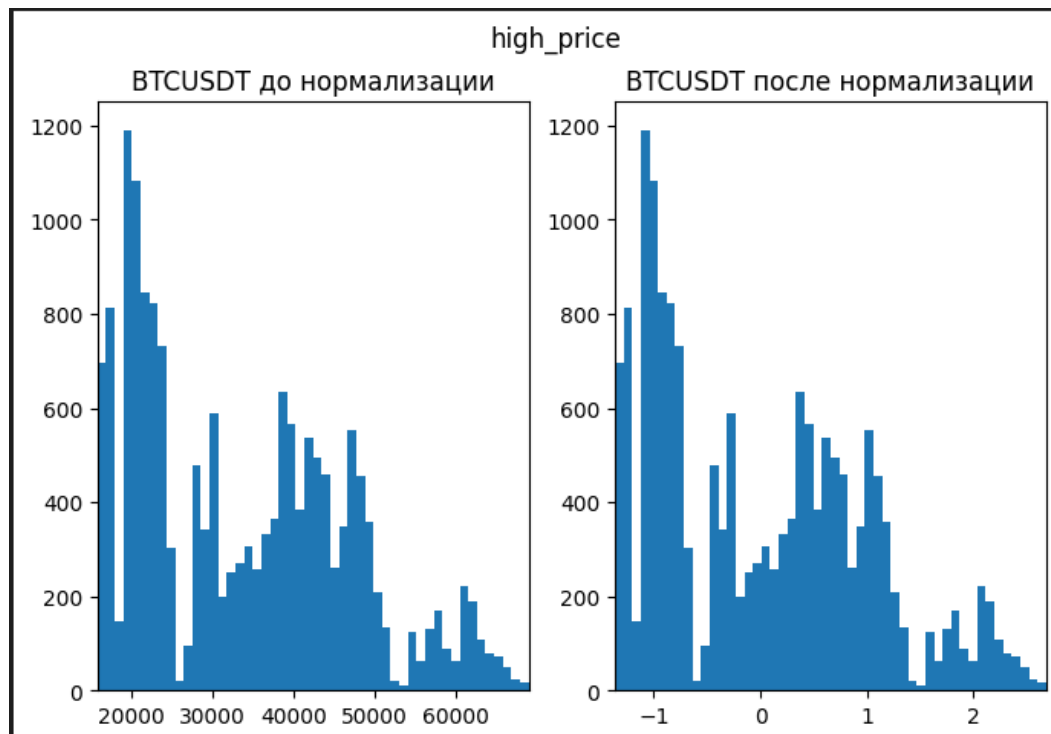


Рисунок 9 - график до и после нормализации признака 'high\_price'

Минимальное значение high\_price до нормализации: 15769.99

Максимальное значение high\_price до нормализации: 69000.0

Минимальное значение high\_price после нормализации: -1.365448134141976

Максимальное значение high\_price после нормализации: 2.6835132550032728

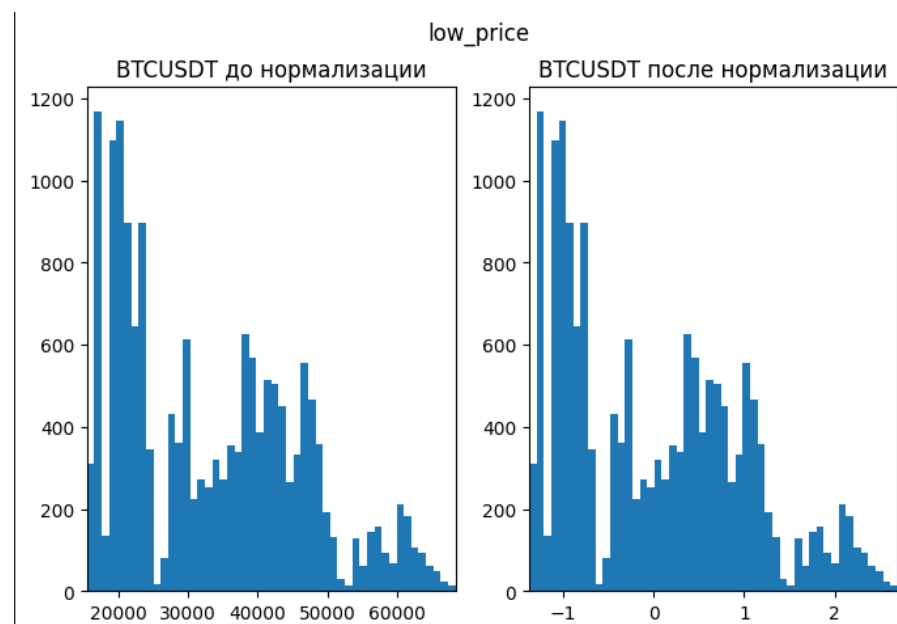


Рисунок 10 - график до и после нормализации признака 'low\_price'

Минимальное значение low\_price до нормализации: 15476.0

Максимальное значение low\_price до нормализации: 68451.19

Минимальное значение low\_price после нормализации: -  
1.3772076567050684

Максимальное значение low\_price после нормализации:  
2.6989157412259557

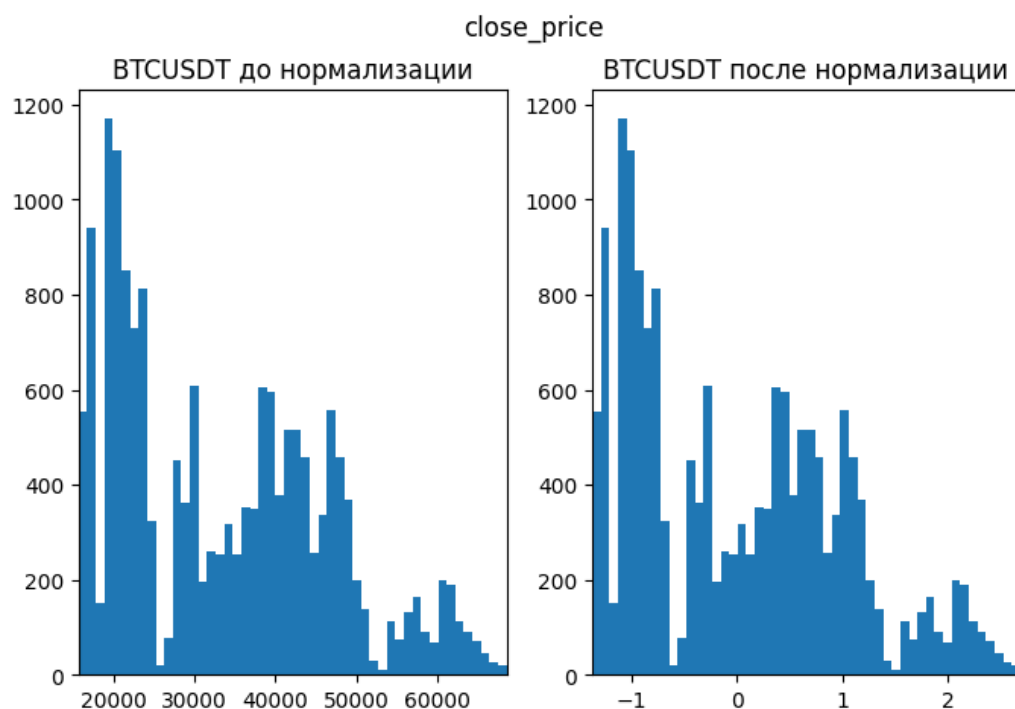


Рисунок 11- график до и после нормализации признака 'close\_price'

Минимальное значение close\_price до нормализации: 15649.52

Максимальное значение close\_price до нормализации: 68633.69

Минимальное значение close\_price после нормализации: -  
1.3692284113264401

Максимальное значение close\_price после нормализации:  
2.683525508767177

Процесс нормализации данных и btcusdt и ethusdt было выполнено успешно. После нормализации, данные были преобразованы так, что они теперь находятся в определенном диапазоне.

## 2.2 Разработка и обучение модели

Для прогнозирования движения цены после выхода новости, можем рассмотреть использование следующих моделей машинного обучения:

- Логистическая регрессия;
- Дерево решений;
- Случайный лес;
- Градиентный бустинг;
- Метод опорных векторов (SVM).

Поскольку данные состоят из цен с интервалом в 1 минуту, были созданы дополнительные признаки на основе временного сдвига. Добавила столбец с разницей цен закрытия за 15, 30 и 60 минут до и после выхода новости.

Входных признаков (X):

- Цена открытия (open\_price);
- Максимальная цена (high\_price);
- Минимальная цена (low\_price) ;
- Цена закрытия (close\_price);
- Разница цены закрытия за 15, 30 и 60 минут после выхода новости;
- Тональность новости (положительная или отрицательная), с бинарными значениями, где положительная новость будет иметь значение 1, а отрицательная - 0.

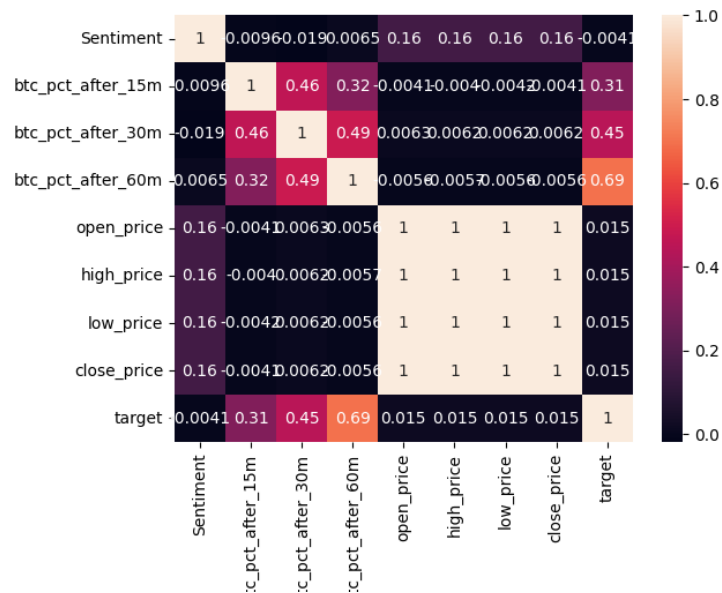


Рисунок 12 - корреляционная матрица

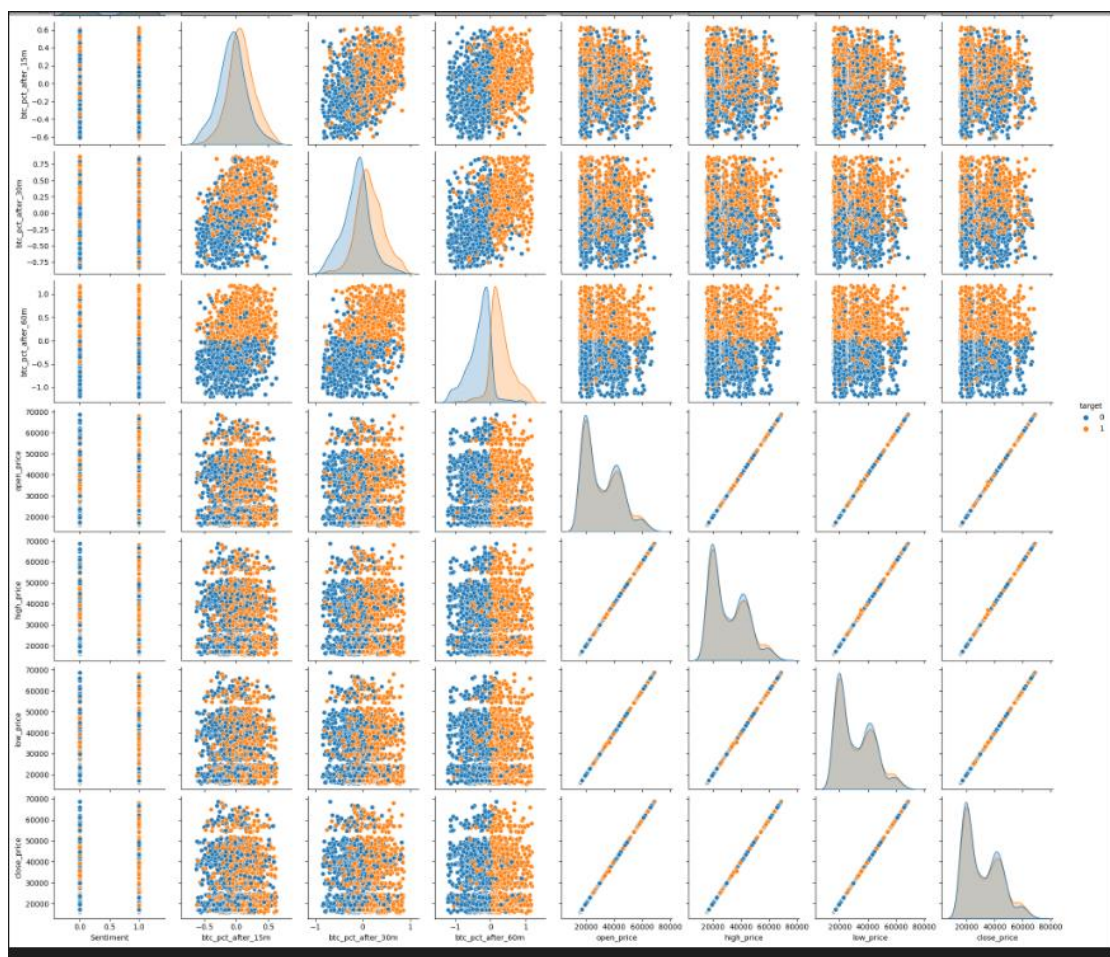


Рисунок 13 - Матрица рассеяния

Целевой переменной (y) будет столбец с изменением цены (например, за 15 минут) после выхода новости, преобразованный в бинарный формат (1 - если

цена увеличилась, 0 - если цена уменьшилась). Далее представлены результаты Логистической регрессии на рисунке 14 и 15.

```
[[254 92]
 [103 227]]
```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.71      | 0.73   | 0.72     | 346     |
| 1            | 0.71      | 0.69   | 0.70     | 330     |
| accuracy     |           |        | 0.71     | 676     |
| macro avg    | 0.71      | 0.71   | 0.71     | 676     |
| weighted avg | 0.71      | 0.71   | 0.71     | 676     |

Рисунок 14 - Результат при изменении цены 60 мин после выхода новости на активе btcusdt

```
✓ 0.1s
[[238 105]
 [100 233]]
```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.70      | 0.69   | 0.70     | 343     |
| 1            | 0.69      | 0.70   | 0.69     | 333     |
| accuracy     |           |        | 0.70     | 676     |
| macro avg    | 0.70      | 0.70   | 0.70     | 676     |
| weighted avg | 0.70      | 0.70   | 0.70     | 676     |

Рисунок 15- Результат при изменении цены 60 мин после выхода новости на активе ethusdt

Сравнение на рисунке 12 и 13:

В первом результате общая точность составляет 71%, что немного выше, чем 70% во втором результате. В первом результате полнота для класса 1 составляет 69%, что ниже, чем 70% во втором результате. В первом результате средняя точность для обоих классов составляет 71%, что немного выше, чем 70% во втором результате. Однако, точность для класса 1 в первом результате составляет 71%, что выше, чем 69% во втором результате. В первом результате средняя F1-мера составляет 72% для класса 0 и 70% для класса 1, что выше, чем 70% и 69% соответственно во втором результате.

Основываясь на этом сравнении, можно сказать, что первый результат является немного лучше сбалансированным между классами и имеет немного лучшую общую точность, полноту и F1-меру. В то время как точность и полнота для класса 1 во втором результате немного ниже, разница между двумя результатами не является существенной. Таким образом, оба результата показывают схожие показатели точности, полноты и F1-score.

На рисунке 14 результаты показывают, что модель «Дерево решений» смогла идеально предсказать движение цены на тестовой выборке. Матрица ошибок показывает, что не было ложных срабатываний или пропусков. Точность, полнота, и F1-мера для обоих классов равны 1, что свидетельствует о превосходном качестве предсказаний модели.

|              |   |           |        |          |         |
|--------------|---|-----------|--------|----------|---------|
| [[199 147]   |   |           |        |          |         |
| [119 211]]   |   |           |        |          |         |
|              |   | precision | recall | f1-score | support |
|              | 0 | 0.63      | 0.58   | 0.60     | 346     |
|              | 1 | 0.59      | 0.64   | 0.61     | 330     |
| accuracy     |   |           |        | 0.61     | 676     |
| macro avg    |   | 0.61      | 0.61   | 0.61     | 676     |
| weighted avg |   | 0.61      | 0.61   | 0.61     | 676     |

Рисунок 16 – результаты модели "Дерево решений" на BTCUSDT

|              |   |           |        |          |         |
|--------------|---|-----------|--------|----------|---------|
| [[207 136]   |   |           |        |          |         |
| [132 201]]   |   |           |        |          |         |
|              |   | precision | recall | f1-score | support |
|              | 0 | 0.61      | 0.60   | 0.61     | 343     |
|              | 1 | 0.60      | 0.60   | 0.60     | 333     |
| accuracy     |   |           |        | 0.60     | 676     |
| macro avg    |   | 0.60      | 0.60   | 0.60     | 676     |
| weighted avg |   | 0.60      | 0.60   | 0.60     | 676     |

Рисунок 17 - результаты модели "Дерево решений" на ETHUSDT



В первом результате общая точность составляет 61%, что немного выше, чем 60% во втором результате. В первом результате полнота для класса 1 составляет 64%, что немного выше, чем 60% во втором результате. В первом результате средняя точность для обоих классов составляет 61%, что немного выше, чем 60% во втором результате. Однако, точность для класса 1 в первом результате составляет 59%, что ниже, чем 60% во втором результате. В первом результате средняя F1-мера составляет 60% для класса 0 и 61% для класса 1, что выше, чем 61% и 60% соответственно во втором результате.

При использовании данной модели были получены аналогичные и так как разница между двумя результатами не является существенной, буду анализировать для данных bitcoin.

Результаты показывают на рисунке 15, что модель «Случайный лес» справилась с задачей классификации идеально на тестовых данных, достигнув 100% точности. Значения precision, recall и f1-score равны 1.00 для обоих классов, что указывает на идеальное предсказание.

```
..  [[228 118]
      [111 219]]
```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.67      | 0.66   | 0.67     | 346     |
| 1            | 0.65      | 0.66   | 0.66     | 330     |
| accuracy     |           |        | 0.66     | 676     |
| macro avg    | 0.66      | 0.66   | 0.66     | 676     |
| weighted avg | 0.66      | 0.66   | 0.66     | 676     |

Рисунок 18 - модель Случайный лес

Точность (precision) для класса 0 составляет 0.67, что означает, что 67% примеров, предсказанных как класс 0, действительно являются классом 0.

Точность для класса 1 составляет 0.65, что означает, что 65% примеров, предсказанных как класс 1, действительно являются классом 1.

Полнота (recall) для класса 0 составляет 0.66, что означает, что 66% примеров класса 0 были правильно классифицированы.

Полнота для класса 1 составляет 0.66, что означает, что 66% примеров класса 1 были правильно классифицированы.

F1-мера, которая является средним гармоническим между точностью и полнотой, составляет 0.67 для класса 0 и 0.66 для класса 1. Это показывает, что модель имеет сбалансированную производительность для обоих классов.

Общая точность модели (accuracy) составляет 66%, что означает, что 66% всех примеров были правильно классифицированы.

Средние показатели (macro avg) для точности, полноты и F1-меры составляют 0.66. Это свидетельствует о сбалансированной производительности модели для обоих классов.

В целом, модель показывает среднюю производительность с точностью 66%. Модель имеет схожую точность, полноту и F1-меру для обоих классов, что говорит о том, что она сбалансирована и не имеет существенного смещения в сторону одного из классов.

```
[[249  97]
 [113 217]]
```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.69      | 0.72   | 0.70     | 346     |
| 1            | 0.69      | 0.66   | 0.67     | 330     |
| accuracy     |           |        | 0.69     | 676     |
| macro avg    | 0.69      | 0.69   | 0.69     | 676     |
| weighted avg | 0.69      | 0.69   | 0.69     | 676     |

Рисунок 19 – модель «Градиентный бустинг»

Точность (precision) для класса 0 составляет 0.69, что означает, что 69% примеров, предсказанных как класс 0, действительно являются классом 0.

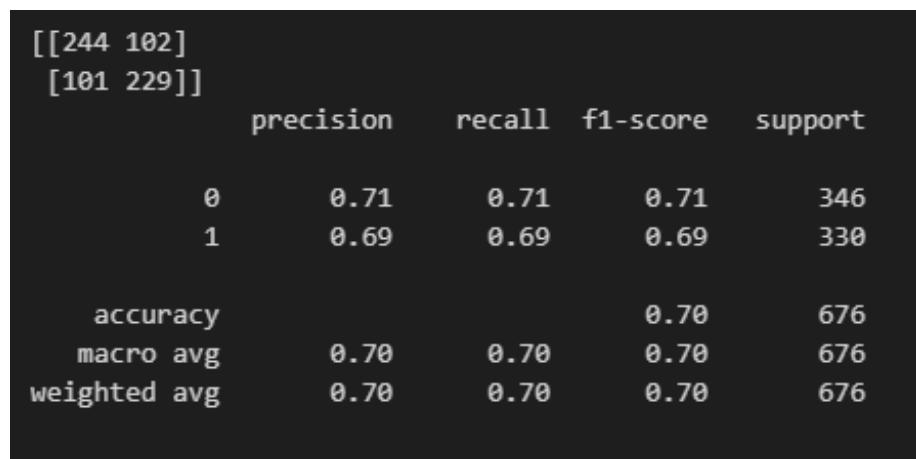
Точность для класса 1 составляет 0.69, что означает, что 69% примеров, предсказанных как класс 1, действительно являются классом 1.

Полнота (recall) для класса 0 составляет 0.72, что означает, что 72% примеров класса 0 были правильно классифицированы.

Полнота для класса 1 составляет 0.66, что означает, что 66% примеров класса 1 были правильно классифицированы.

F1-мера, которая является средним гармоническим между точностью и полнотой, составляет 0.70 для класса 0 и 0.67 для класса 1. Это показывает, что модель имеет сбалансированную производительность для обоих классов. Общая точность модели (accuracy) составляет 69%, что означает, что 69% всех примеров были правильно классифицированы. Средние показатели (macro avg) для точности, полноты и F1-меры составляют 0.69. Это свидетельствует о сбалансированной производительности модели для обоих классов.

В целом, модель градиентного бустинга показывает достаточно хорошую производительность с точностью 69%. Модель имеет схожую точность, полноту и F1-меру для обоих классов, что говорит о том, что она сбалансирована и не имеет существенного смещения в сторону одного из классов.



```

[[244 102]
 [101 229]]
precision    recall  f1-score   support

   0         0.71     0.71     0.71     346
   1         0.69     0.69     0.69     330

 accuracy          0.70     676
 macro avg         0.70     0.70     0.70     676
weighted avg         0.70     0.70     0.70     676

```

Рисунок 20 - Результаты модели SVM

Точность (precision) для класса 0 составляет 0.71, что означает, что 71% примеров, предсказанных как класс 0, действительно являются классом 0.

Точность для класса 1 составляет 0.69, что означает, что 69% примеров, предсказанных как класс 1, действительно являются классом 1.

Полнота (recall) для класса 0 составляет 0.71, что означает, что 71% примеров класса 0 были правильно классифицированы.

Полнота для класса 1 составляет 0.69, что означает, что 69% примеров класса 1 были правильно классифицированы.

F1-мера, которая является средним гармоническим между точностью и полнотой, составляет 0.71 для класса 0 и 0.69 для класса 1. Это показывает, что модель имеет сбалансированную производительность для обоих классов.

Общая точность модели (ассигасу) составляет 70%, что означает, что 70% всех примеров были правильно классифицированы.

Средние показатели (macro avg) для точности, полноты и F1-меры составляют 0.70. Это свидетельствует о сбалансированной производительности модели для обоих классов.

Таблица 3 - Отчет о классификации

|                               | precision | recall  | F1-score | accuracy |
|-------------------------------|-----------|---------|----------|----------|
| Логистическая регрессия       | 0 - 71%   | 0 - 73% | 0 - 72%  | 71%      |
|                               | 1 - 71%   | 1 - 69% | 1 - 70%  |          |
| Дерево решений                | 0 - 63%   | 0 - 58% | 0 - 60%  | 61%      |
|                               | 1 - 59%   | 1 - 64% | 1 - 61%  |          |
| Случайный лес                 | 0 - 67%   | 0 - 66% | 0 - 67%  | 66%      |
|                               | 1 - 65%   | 1 - 66% | 1 - 66%  |          |
| Градиентный бустинг           | 0 - 69%   | 0 - 72% | 0 - 70%  | 69%      |
|                               | 1 - 69%   | 1 - 66% | 1 - 67%  |          |
| Метод опорных векторов (SVM). | 0 - 71%   | 0 - 71% | 0 - 71%  | 70%      |
|                               | 1 - 69%   | 1 - 69% | 1 - 69%  |          |

## 2.3 Нейронные сети

### MLP

Модель MLP имеет сравнительно высокую точность на обучающей выборке (71.83%) и на валидационной выборке (71.01%). Это говорит о том, что модель достаточно хорошо обобщает данные и может быть использована для классификации с заданной точностью.

Потери на обучающей выборке (0.5636) и на валидационной выборке (0.5918) не сильно отличаются, что указывает на отсутствие переобучения. Если бы потери на обучающей выборке были намного меньше, чем на валидационной, это могло бы свидетельствовать о переобучении модели.

В целом, модель MLP с 50 эпохами обучения показывает хорошие результаты в данной задаче классификации с точностью около 71%.

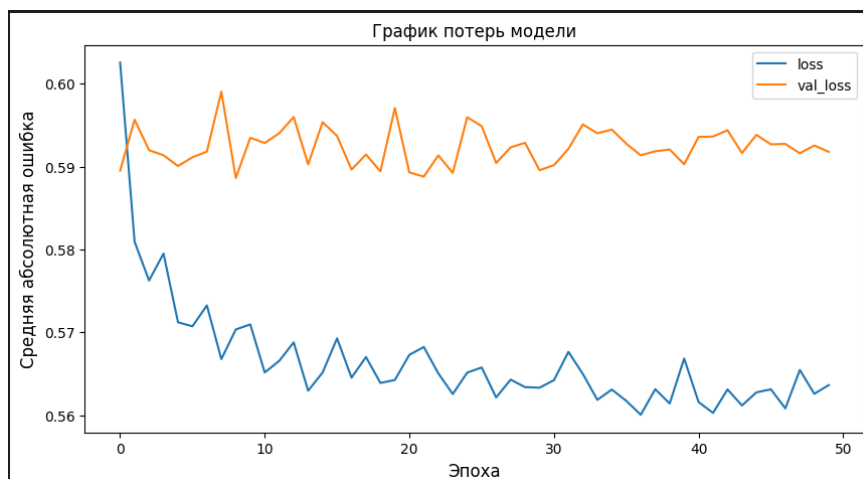


Рисунок 21 - График потерь модели "MLP"

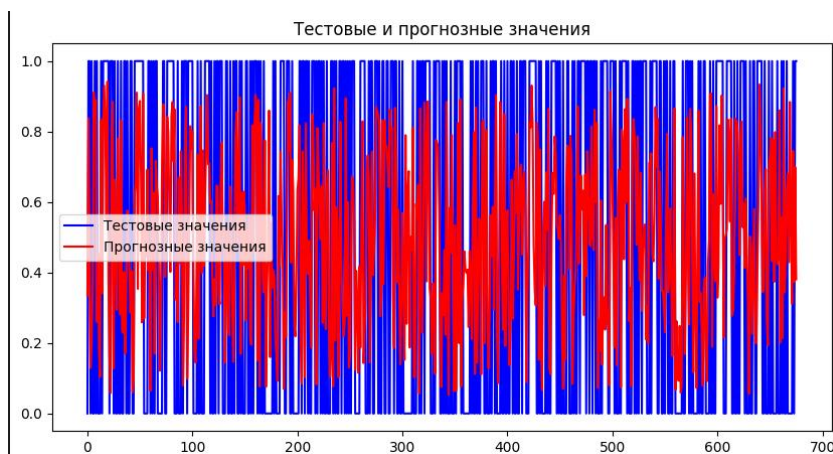


Рисунок 22 - Визуализация оригинальных и предсказанных значений модели "MLP"

## LSTM

Модель LSTM имеет сравнительно высокую точность на обучающей выборке (71.26%) и на валидационной выборке (70.56%). Это говорит о том, что модель достаточно хорошо обобщает данные и может быть использована для классификации с заданной точностью.

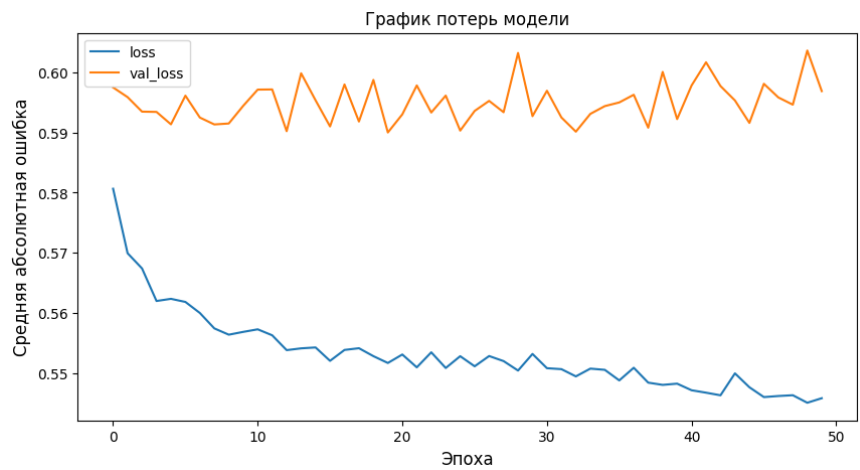


Рисунок 23 - График потерь модели "LSTM"

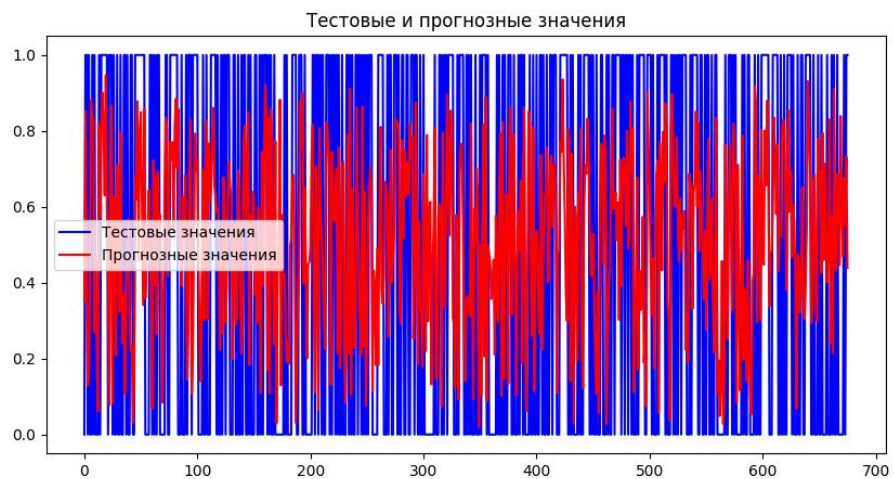


Рисунок 24 - Визуализация оригинальных и предсказанных значений модели "LSTM"

## CNN

Количество обучаемых параметров составляет 4,353(представлено на рисунке 23), что является относительно небольшим количеством, что может быть положительным фактором в плане скорости обучения и предсказания.

| Layer (type)                    | Output Shape  | Param # |
|---------------------------------|---------------|---------|
| conv1d_1 (Conv1D)               | (None, 5, 32) | 128     |
| max_pooling1d_1 (MaxPooling 1D) | (None, 2, 32) | 0       |
| flatten_1 (Flatten)             | (None, 64)    | 0       |
| dense_41 (Dense)                | (None, 64)    | 4160    |
| dense_42 (Dense)                | (None, 1)     | 65      |
| Total params: 4,353             |               |         |
| Trainable params: 4,353         |               |         |
| Non-trainable params: 0         |               |         |

Рисунок 25 - Структура модели CNN

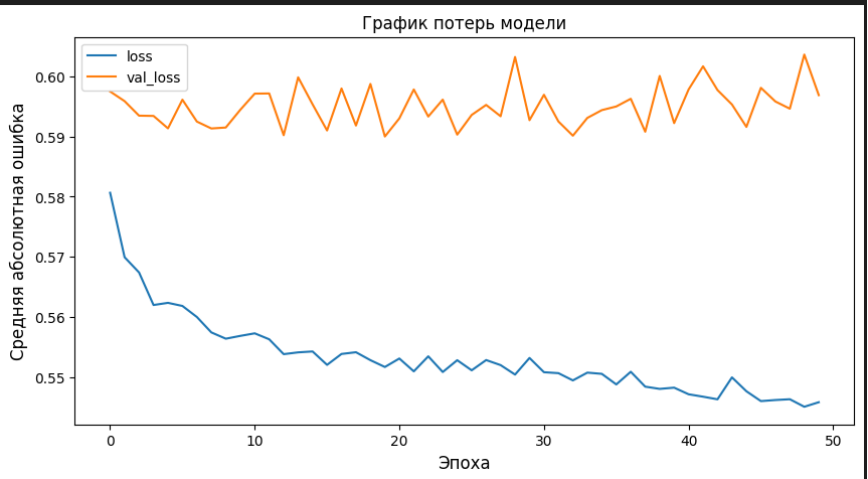


Рисунок 26-График потерь модели "CNN"

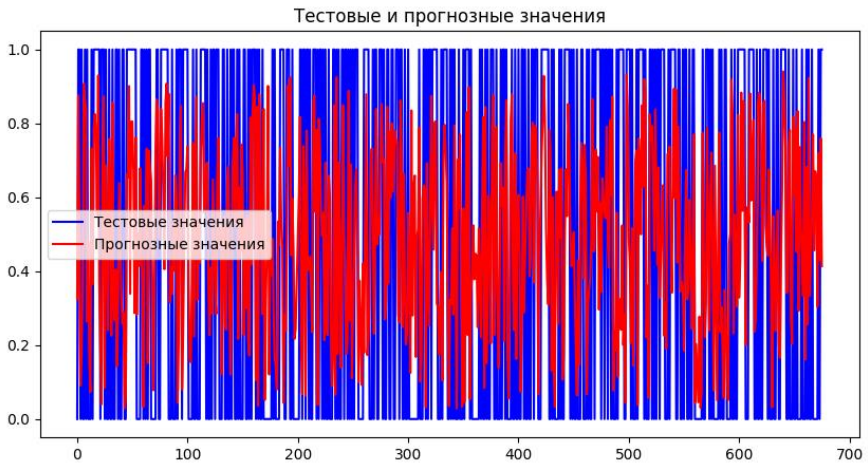


Рисунок 27 - Визуализация оригинальных и предсказанных значений модели "CNN"

Модель CNN имеет хорошую точность на обучающей выборке (72.46%) и на валидационной выборке (70.27%). Это говорит о том, что модель хорошо

обобщает данные и может быть использована для классификации с заданной точностью.

В целом, модель CNN показывает хорошие результаты в данной задаче классификации с точностью около 70.27%

Сравнивая результаты модели CNN с другими рассмотренными моделями, например, LSTM и MLP, можно заметить, что точность CNN немного ниже, чем у модели MLP, но выше, чем у модели LSTM.

## **2.4 Разработка приложения**

Функционал приложения:

Это веб-приложение, созданное на основе Flask, использует предварительно обученную модель нейросети (MLP) для прогнозирования движения цены (покупка/продажа) криптовалюты (BTCUSDT) после выхода новости на основе введенных пользователем данных.

Пользователь вводит следующие данные как показано на примере на рисунке 28 :

- 1) Sentiment (настроение новости): 0 (негативное) или 1 (позитивное);
- 2) BTCUSDT After 15m: изменение цены BTCUSDT через 15 минут после новости (например, 0.5, 0.6);
- 3) BTCUSDT After 30m: изменение цены BTCUSDT через 30 минут после новости (например, 0.5, 0.6);
- 4) Open Price: цена открытия BTCUSDT (например, 50000, 51000);
- 5) High Price: максимальная цена BTCUSDT (например, 51000, 52000);
- 6) Low Price: минимальная цена BTCUSDT (например, 49000, 50000);
- 7) Close Price: цена закрытия BTCUSDT (например, 51000, 52000).



Enter the data to get a price movement prediction

Sentiment:

BTCUSDT After 15m:

BTCUSDT After 30m:

Open Price:

High Price:

Low Price:

Close Price:

Рисунок 28 – Ввод данных

Краткая инструкция использования:

Запустите веб-приложение, выполнив команду `app.run(debug=True)` в терминале или командной строке.

Откройте веб-браузер и перейдите по адресу, указанному в терминале (обычно `http://127.0.0.1:5000/`).

Введите данные в поля ввода на основе подсказок (placeholder) и нажмите кнопку "Get Prediction".

После отправки формы, приложение обработает введенные данные и выполнит прогноз с использованием предварительно обученной модели нейросети.

На странице будет отображен результат прогноза: "Buy" (Покупать) или "Sell" (Продавать) криптовалюты.

```
{
  "Prediction": "Sell"
}
```

Raw Parsed

Рисунок 29 -Результат прогноза

## 2.5 Создание удаленного репозитория

Для данного исследования был создан удаленный репозиторий на GitHub, который находится по адресу <https://github.com/VicNovak/Crypto-News-Sentiment-Analysis-Price-Prediction>. На него были загружены результаты работы: исследовательский notebook, код приложения Flask и парсеров (Binance и Reuter news), датасет news.

## ЗАКЛЮЧЕНИЕ

В результате проведенного анализа было установлено, что прогнозирование цены цифрового актива на основе анализа новостных потоков является возможным, однако требует более тщательного исследования и улучшения модели.

В ходе исследования были рассмотрены различные модели машинного обучения для прогнозирования движения цены после выхода новости, такие как логистическая регрессия, дерево решений, случайный лес, градиентный бустинг, метод опорных векторов (SVM), нейронные сети типа MLP, LSTM и CNN. Общая точность моделей находится в диапазоне 66-71%.

Сравнивая результаты всех рассмотренных моделей, можно сделать вывод, что модель MLP показала наилучшую точность и в дальнейшем планируется провести дополнительные эксперименты с другими моделями машинного обучения и улучшить процесс предобработки данных, чтобы повысить точность и надежность прогнозирования цены цифровых активов на основе анализа новостных потоков.

В перспективе развития можно рассмотреть следующие направления:

1) Автоматизация анализа новостей: Интеграция модели с системами парсинга и анализа новостей в режиме реального времени для автоматического сбора и обработки информации.

2) Разработка стратегий торговли: Создание автоматической торговой системы на основе прогнозов модели для определения оптимальных точек входа и выхода из рынка.

3) Факторный анализ: Использование модели для анализа влияния различных факторов на цену, таких как экономические показатели, глобальные события или макроэкономические данные.

4) Расширение диапазона прогнозируемых активов: Адаптация модели для прогнозирования движения цен других криптовалют или традиционных финансовых инструментов, таких как акции, облигации, товары и валюты.

## Библиографический список

- 1 Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербург, 2021. - 416 с.: ил.
- 2 Плас Дж. Вандер Python для сложных задач: наука о данных и машинное обучение. — СПб.: Питер, 2018. — 576 с.: ил.
- 3 Герон, О. Прикладное машинное обучение с использованием алгоритмов Scikit-Learn и TensorFlow. М.: Символ-Плюс, 2019 - 420 с.: ил.
- 4 Чоллет, Ф. Глубокое обучение с использованием Python. Машинное обучение с использованием библиотеки Keras. М.: Символ-Плюс, 2018- 384 с.: ил.
- 5 Шабанов, П. Машинное обучение и анализ данных с Python. М.: Издательство Солон-Пресс, 2018 - 512 с.: ил.
- 6 Браун, М. Практический анализ данных на Python: изучаем Data Science и машинное обучение. СПб.: Питер, 2019 - 400 с.: ил.
- 7 Рашка, С. Python и машинное обучение: руководство для начинающих. М.: Издательство Эксмо, 2018 - 512 с.: ил.
- 8 Болонкин, А. Python для анализа данных: от простых вычислений до машинного обучения. М.: Издательство Вильямс, 2019 - 384 с.: ил.
- 9 Светльник, Р. Машинное обучение с использованием Python: руководство по созданию систем искусственного интеллекта. М.: Издательство Эксмо, 2020 - 480 с.: ил.
- 10 Мюллер, А., Гвидо, С. Введение в машинное обучение с использованием Python: руководство для специалистов по работе с данными. СПб.: Питер, 2020 - 416 с.: ил.
- 11 Маккинни, У. Python для анализа данных: анализ и визуализация данных в примерах. СПб.: Питер., 2020 - 608 с.: ил
- 12 Брюс, П. Практическая статистика для специалистов Data Science: Пер. с англ. / П. Брюс, Э. Брюс. — СПб.: БХВ-Петербург, 2018. — 304 с.: ил.

- 13     Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.
- 14     Документация по языку программирования python: – Режим доступа: <https://docs.python.org/3.8/index.html>.
- 15     Документация по библиотеке numpy: – Режим доступа: <https://numpy.org/doc/1.22/user/index.html#user>.
- 16     Документация по библиотеке pandas: – Режим доступа: [https://pandas.pydata.org/docs/user\\_guide/index.html#user-guide](https://pandas.pydata.org/docs/user_guide/index.html#user-guide).
- 17     Документация по библиотеке matplotlib: – Режим доступа: <https://matplotlib.org/stable/users/index.html>.
- 18     Документация по библиотеке seaborn: – Режим доступа: <https://seaborn.pydata.org/tutorial.html>.
- 19     Документация по библиотеке sklearn: – Режим доступа: [https://scikitlearn.org/stable/user\\_guide.html](https://scikitlearn.org/stable/user_guide.html).
- 20     Документация по библиотеке keras: – Режим доступа: <https://keras.io/api/>.
- 21     Руководство по быстрому старту в flask: – Режим доступа: <https://flaskrussian-docs.readthedocs.io/ru/latest/quickstart.html>.
- 22     Loginom Вики. Алгоритмы: – Режим доступа: <https://wiki.loginom.ru/algorithms.html>.
- 23     Петров, Д., Иванов, В. Применение методов машинного обучения для прогнозирования криптовалютных рынков с использованием Python // Журнал исследований в области компьютерных наук. – Т. 3, № 2. – 2021 -С. 125-137.
- 24     Калинин, А., Смирнов, М. Анализ текстовых данных в социальных сетях для прогнозирования финансовых рынков с использованием Python и машинного обучения // Журнал искусственного интеллекта и машинного обучения. – Т. 2, № 4. - 2020– С. 84-93.
- 25     Andre Ye. 5 алгоритмов регрессии в машинном обучении, о которых вам следует знать: – Режим доступа: <https://habr.com/ru/company/vk/blog/513842/>.

26 Li Y., Islambekov U., Akcora C., Smirnova E., Gel Y.R., Kantarcioglu M. Dissecting ethereum blockchain analytics: what we learn from topology and geometry of the ethereum graph?// society for industrial and applied mathematics publications. – 2020 .– 523-531 c.

27 Mammadov I., Akbulaev N., Hemdullayeva M. Correlation and regression analysis of the relation between Ethereum price and both its volume and Bitcoin price // Journal of structured finance. – 2020. – 46-56 c.

28 Anwar S., Anayat S., Butt S., Butt S., Saad M., Generation analysis of blockchain technology: Bitcoin and Ethereum // International journal of information engineering and electronic business.-2020.-№4.-30-39 c.

29 Faqir-Rhazoui Y., Arroyo J., Hassan S., A comparative analysis of the platforms for decentralized autonomous organizations in the Ethereum blockchain // Journal of internet services and applications. – 2021. - №1. – 9 c.

30 Meynkhhard A., Effect of Bitcoin volatility on altcoins pricing // Advances in intelligent systems and computing . – 2020.- 652-664 c.