Проект по построению рекомендаций на фондовом рынке

# Аннотация проекта (Executive summary)

Задачей проекта является формирование рекомендации на фондовом рынке с применением средств машинного обучения.

В рамках проекта проверялись следующие стратегии:

* Принцип “купи и держи” подразумевает идею купить акции и ожидать длительный период в расчете получить доход от роста стоимости ценной бумаги и от выплачиваемых дивидендов.
* “Колебания стоимости акций в периоды дивидендных выплат” - идея строиться на предположении, что за несколько дней до выплаты дивидендов происходит рост стоимости ценной бумаги. Задача приобрести ценную бумагу до роста и продавать максимально близко к экс дивидендной дате (продажа по повышенной цене).
* Стратегия “Корреляция различных финансовых инструментов в общей отрасли” - идея строиться на предположении, что компании одной отрасли могут быть связаны общими событиями и менять тренды движения котировок в общую сторону.
* Стратегия “Прогнозирование тренда на основе предыдущих данных” - прогнозирование направления тренда на основе предыдущих цен. Идея заключается в том что при построении модели машинного обучения, моделью будут выявлены закономерности по которым будут строиться рекомендации.

В качестве входных данных выбраны ценные бумаги нефтегазовой отрасли с 2017 по 2021 года.

Основным параметром измерения экономическая оценки стратегии является среднедневная доходность стратегии. Дополнительными параметрами являются общее кол-во сделок за рассматриваемый период, кол-во прибыльных сделок, максимальные потери от худшей сделки, средняя длительность сделки, максимальная просадка и другие.

| Стратегия | Реализация | Среднедневная доходность |
| --- | --- | --- |
| Принцип “купи и держи” | Измерение разницы цен между началом и концом года | 0,042% |
| “Колебания стоимости акций в периоды дивидендных выплат” | Измерение разницы цен в период за 20 дней и за 2 дня до экс-дивидендной даты | 0,142% |
| “Корреляция различных финансовых инструментов в общей отрасли” | Реализация нейронной сети (на tensorflow) по трем архитектурам: Dense, LSTM, Conv1D | * Dense: 0,49% * LSTM: 0,43% * Conv1D: 0,12% |
| “Прогнозирование тренда на основе предыдущих данных” | Реализация линейной модели (Sklearn.linear\_model) | 0,1% |

Метрикой измерения точности моделей машинного обучения выбрана метрика accuracy.

Построенные модели по стратегиям “Корреляция различных финансовых инструментов в общей отрасли” и “Прогнозирование тренда на основе предыдущих данных” выведены в демонстрационный телеграм-канал: “<https://t.me/MTSProject_StackMarket>”.

Публикация осуществляется с помощью разработанного телеграм-бота, который использует построенные модели и данные о котировках ценных бумаг за текущие сутки.

Пример рекомендаций телеграм-канала:

“Рекомендация на 31.08.2021:

По стратегии "Корреляция в отрасли":

- Покупка LKOH (уверенность модели Conv1D: 55%)

- Покупка SIBN (уверенность модели LSTM: 60%)”

Основным минусом стратегий (включая дивидендную стратегию) является “подгонка” результатов под “идеальные” показатели. Так например выбранные для дивидендной стратегии периоды открытия/закрытия сделки подобраны для прошлых лет по нефтегазовой отрасли, попытки применить эти же параметры для других ценных бумаг будут демонстрировать большие убытки. Стратегии построенные с применением машинного обучения аналогично “подстраиваются” под значения. Для выработки стабильно неубыточной стратегии необходимо проведение дополнительных изысканий.

Проектные файлы:

* Каталог проекта GDrive:

<https://drive.google.com/drive/folders/1ByM63BCupxb182WIY_HY14-o_77_vdGU?usp=sharing>

* Ноутбук 2-го ДЗ:

Оценка модели выбор baseline:

<https://colab.research.google.com/drive/1mqEq6vQ4UkRBfo4lQ1vCNt3PqK3iGIDy>

* Ноутбуки 3-го ДЗ:
* Стратегия “Корреляция различных финансовых инструментов в общей отрасли”:

<https://colab.research.google.com/drive/1kb-Z6fqNWNnj-QI55aoqHJmmv8Krtz1X>

* Стратегия “Прогнозирование тренда на основе предыдущих данных”:

<https://colab.research.google.com/drive/181bki-MltxcgnP4WF2jY8ImGnJL_HayO>

* Демостенд, телеграм-бот:

<https://colab.research.google.com/drive/1B6yMBvgncmd9D9pqa4K2UgNVtfzhVYMk>

* Каталог стратегий:

<https://drive.google.com/drive/folders/1HoaAHngrGChdC4BE41BkgOY2wek6IYWu>

* Каталог Демостенда:

<https://drive.google.com/drive/folders/1IOQC2d-4iIDx1UcQzsVFVummPw5DlFkq>

* Канбан доска: <https://gitlab.com/mts.project2/stock.market/-/issues>

# Оглавление:

[**Аннотация проекта (Executive summary)**](#_heading=h.376cm4rgrc3s) **1**

[**Оглавление:**](#_heading=h.3znysh7) **2**

[**Постановка задачи**](#_heading=h.tyjcwt) **4**

[Проблематика](#_heading=h.3dy6vkm) 4

[Бизнес-задача](#_heading=h.1t3h5sf) 4

[Разработать неубыточную стратегию инвестирования денежных средств с использованием общедоступных финансовых инструментов.](#_heading=h.4d34og8) 4

[Математическая задача](#_heading=h.2s8eyo1) 4

[Обоснование выбора проекта](#_heading=h.17dp8vu) 4

[Этапы реализации проекта](#_heading=h.3rdcrjn) 5

[Инструменты проекта](#_heading=h.26in1rg) 5

[Система Управление версиями](#_heading=h.lnxbz9) 5

[**Изучение предметной области**](#_heading=h.35nkun2) **6**

[Подбор инструментов для решения задачи](#_heading=h.1ksv4uv) 6

[Вид анализа (технический / фундаментальный анализ)](#_heading=h.44sinio) 6

[Финансовый инструмент](#_heading=h.2jxsxqh) 7

[**Базовая стратегия.**](#_heading=h.z337ya) **7**

[В качестве базовой стратегии будет использована стратегия: “Колебания стоимости акций в периоды дивидендных выплат”.](#_heading=h.1y810tw) 7

[**Общее описание:**](#_heading=h.4i7ojhp) **7**

[Построение модели Базовой (первичной) стратегии](#_heading=h.2xcytpi) 8

[**Исходные данных**](#_heading=h.1ci93xb) **8**

[Описание используемых данных](#_heading=h.3whwml4) 8

[**Валидация данных**](#_heading=h.2bn6wsx) **9**

[Анализ входного датасета](#_heading=h.qsh70q) 9

[Определение волатильности цен по каждой компании](#_heading=h.3as4poj) 10

[По итогам расчета волатильности для используемых инструментов можно сделать вывод, о том: что акции LKOH имеют значительно сильную волатильность по сравнению с другими инструментами. При этом волатильность остальных компаний находится на приемлемом уровне.](#_heading=h.1pxezwc) 11

[Определение количества и силы гэпов (резкий скачок цены, например из-за дивидендов)](#_heading=h.2p2csry) 11

[Определить какие из гэпов дивидендные какие по другим причинам.](#_heading=h.147n2zr) 11

[**Экономический эффект**](#_heading=h.3o7alnk) **12**

[Описание экономической оценки модели](#_heading=h.23ckvvd) 12

[Расчет экономической оценки базовой модели](#_heading=h.ihv636) 13

[Метрики и оценка качества моделей. Фиксация baseline.](#_heading=h.32hioqz) 15

[Расчет и сравнение просадок базовой модели и стратегии “купи и держи”](#_heading=h.1hmsyys) 16

[**Стратегии и гипотезы**](#_heading=h.41mghml) **17**

[Стратегия “Корреляция различных финансовых инструментов в общей отрасли”.](#_heading=h.2grqrue) 20

[Общее описание:](#_heading=h.vx1227) 20

[Анализ рынка на применимость стратегии:](#_heading=h.3fwokq0) 21

[Подбор архитектуры модели:](#_heading=h.1v1yuxt) 24

[Анализ детализации входных данных](#_heading=h.4f1mdlm) 26

[Анализ сегментации данных.](#_heading=h.2u6wntf) 27

[Промежуточная оценка модели LSTM с учетом сегментирования данных](#_heading=h.19c6y18) 28

[Сравнительный анализ точности модели и экономической оценки. Анализ воспроизводимости прогноза](#_heading=h.3tbugp1) 30

[Воспроизводимость модели Dense:](#_heading=h.28h4qwu) 31

[Воспроизводимость модели LSTM:](#_heading=h.nmf14n) 31

[Воспроизводимость модели Conv1D:](#_heading=h.37m2jsg) 32

[Анализ входных таймфреймов](#_heading=h.1mrcu09) 33

[Подбор гиперпараметров модели](#_heading=h.46r0co2) 34

[Подготовка выходных данных/классификация ответов](#_heading=h.2lwamvv) 35

[Развитие стратегии и модели](#_heading=h.111kx3o) 35

[**Вывод по итогам моделирования стратегии “Корреляция в отрасли”**](#_heading=h.ulr8yvkep520) **36**

[Стратегия “Прогнозирование конкретной цены на основе предыдущих данных”](#_heading=h.206ipza) 36

[Линейные модели](#_heading=h.4k668n3) 36

[Стратегия “Прогнозирование тренда на основе предыдущих данных”](#_heading=h.2zbgiuw) 36

[Использование разницы цены n(10) предыдущих часов](#_heading=h.1egqt2p) 37

[**Создание демонстрационного стенда**](#_heading=h.3ygebqi) **37**

[**Выбор способа демонстрации решения**](#_heading=h.2dlolyb) **37**

[**Описание демонстрационного стенда**](#_heading=h.sqyw64) **38**

[**Описание структуры каталогов проекта:**](#_heading=h.3cqmetx) **40**

[**Термины и сокращения**](#_heading=h.1rvwp1q) **41**

# 

# Постановка задачи

## Проблематика

В настоящее время наблюдается повышенный интерес к инвестированию среди населения. Прибыльность инвестиций напрямую зависит от риска. Торговля акциями (противоположность подходу “купи и держи”) является высокорисковым инструментом. В целях минимизации рисков разрабатываются торговые стратегии, которые на основе различных данных формируют сигнал на вход в позицию (открытие ордера на покупку либо на продажу актива). Применение машинного обучение в торговых стратегиях является актуальным направленим.

## Бизнес-задача

### Разработать неубыточную стратегию инвестирования денежных средств с использованием общедоступных финансовых инструментов.

* Разработанная стратегия должна быть доходнее принципа “купи и держи”
* Предполагается максимизировать средневную доходность выбранной стратегии

## Математическая задача

* Разработать модель анализа котировок для прогнозирования сигнала на вход в позицию (открытие ордера на покупку либо на продажу).
* Ранжирование сигналов по “уверенности” модели в них, установка порога “уверенности” для целевого класса
* Определение условий закрытия позиции (ордера).
* Метрика качества – среднедневная доходность за период использования финансовых активов при работе с выбранной стратегией

## Обоснование выбора проекта

Причиной выбора данного проекта являются:

* Доступность исходных данных. Много источников исходных данных с открытым доступом.
* Наличие большого количества материалов по данной тематике.
* Относительная простота проверки результата работы модели в боевых условиях
* Общий интерес команды к данной теме

## Этапы реализации проекта

Определены следующие этапы реализации проекта:

1. Подготовительный (31.07.2021 - 13.08.2021)

* Сбор и первичная валидация данных
* Построение базовой модели для формирования baseline
* Определение системы оценок экономической эффективности моделей

1. Определение эффективных стратегий (итерационный этап) (14.08.2021 - 26.08.2021)

* Выдвижение гипотез и формирование стратегий
* Подготовка датасетов, конструирование признаков (feature engineering)
* Создание моделей, подбор гиперпараметров
* Проверка эффективности модели

1. Формирование сервиса/ Проверка выбранных стратегий в бою (с 26.08.2021 по 29.08.2021)

## Инструменты проекта

Проектная работа:

* Каталог проекта GDrive:

<https://drive.google.com/drive/folders/1I8mtxrngsSO7jI7NhjGva8d51z0nlTdp?usp=sharing>

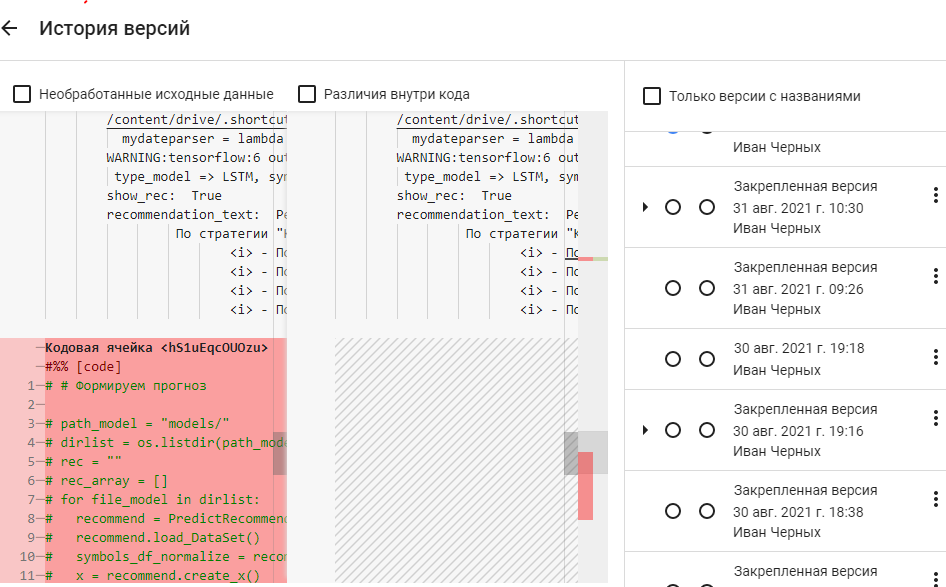
* Ноутбук: <https://colab.research.google.com/drive/1dBTQKRW7grEa2Maf5MEA5Wp_G9vJzNwl#scrollTo=gY_WggoOrcUa>
* Канбан доска: <https://gitlab.com/mts.project2/stock.market/-/issues>

### Система Управление версиями

В качестве системы управления версиями выбрано решение Google по следующим причинам:

* основное хранилище проекта является Google Drive
* разработка ведется в Google Colab
* в качестве дальнейшего развития проекта планируется использовать планировщик задач [Google Scheduler](https://cloud.google.com/scheduler) (из пакета Google Cloud Platform).
* вся проектная документация ведется в Google Docs, где также используется системам версионности Google.

Инструменты Google фиксируют все изменения и предоставляют удобные средства управления версиями (сравнение, восстановление и прочее).



# Изучение предметной области

## Подбор инструментов для решения задачи

### Вид анализа (технический / фундаментальный анализ)

Существует две методики прогнозирования цен акций:

* Технический анализ (ТА) – это способ прогнозирования будущей стоимости с помощью графиков и индикаторов, здесь рассматривается поведение цены акций и выявляются его разнообразные паттерны (используется анализ временных рядов).
* Фундаментальное исследование (ФА) – это метод прогнозирования цен с помощью финансовых показателей деятельности компании и вычисления внутренней стоимости фирмы. В этом случае аналитики оценивают информацию, которая больше относится к компании, чьи акции торгуются на бирже, нежели к самим акциям.

Оба вида анализа имеют преимущества и недостатки.

Среди основных минусов ТА можно выделить:

* Противоречивость выводов при использовании различных методов.
* Запаздывание сигналов (тренд замечается слишком поздно).
* Высокий уровень неопределенности при прогнозах.

Среди основных минусов ФА можно выделить:

* Использование методов ФА требует достаточно глубоких знаний по финансам и экономике, бухгалтерской отчётности компаний и т.п.
* Фундаментальный анализ не учитывает множество факторов, которые могут оказывать реальное влияние на ценовые движения акций. Поэтому часто ценовые движения могут идти вопреки прогнозам фундаментальных аналитиков.
* Для каждого случая нужно индивидуальное решение.
* Занимает достаточно много времени.

С учетом отсутствия у команды глубоких знаний по финансам и трудностей при сборе всех факторов для проведения ФА, основным видом применяемого анализа выбирается Технический анализ.

Основные методы технического анализа:

* Классический анализ — определяет варианты ценовых изменений с помощью начертания различных линий и фигур на ценовом графике.
* Индикаторный анализ — определяет варианты изменения цены с помощью математических функций, называемых индикаторами.
* Анализ объемов — определяет варианты дальнейшего изменения цены с точки зрения объема сделок, прошедших с активом и приведшим цену к определенным изменениям.
* Свечной анализ — определяет варианты развития ценового движения, исходя из комбинаций японских свечей.
* Анализ сезонных закономерностей (например зависимость от дивидендных выплат).

### Финансовый инструмент

В качестве основного финансового инструмента выбран Фондовый рынок: Акции крупных российских и зарубежных компаний (с невысокой волатильностью). Выбор конкретных компаний производится на основе схожести рода деятельности компаний (например: нефтяная промышленность, банковская сфера и т.д.).

Выбор таймфрейма (интервал времени ценового графика) зависит от конкретно выбранной стратегии инвестирования.

# Базовая стратегия.

### 

### В качестве базовой стратегии будет использована стратегия: “Колебания стоимости акций в периоды дивидендных выплат”.

# Общее описание:

Существуют стабильные компании, которые выплачивают дивиденды акционерам в фиксированные даты, определяемые на акционерных собраниях за несколько месяцев до выплат. Информация о точных датах дивидендных выплат доступна в открытых источниках.

Дивгэп - стратегия инвестирования, опирающаяся на взаимодействие со следующими фактами:

1. За несколько дней до экс-дивидендной даты цена акции как правило возрастает, за счет увеличения количества акционеров, желающих приобрести ценную бумагу для получения дивидендных выплат.
2. После экс-дивидендной даты цена акции падает, так как акционеры получили выплаты и могу избавиться от акций.

Возможные сценарии поведения инвестора:

1. Приобретать ценную бумагу до роста и продавать максимально близко к экс дивидендной дате (продажа по повышенной цене)
2. Приобретать ценную бумагу после экс дивидендной даты, когда цена акции упала после выплаты дивидендов, и удержание для ожидания дальнейшего роста (восстановление цены).

### Построение модели Базовой (первичной) стратегии

Перед выполнением основных работ по моделированию необходимо построить базовую модель, которая определит baseline (“исходный уровень”) оценки эффективности. В дальнейшем все построенные модели будут оцениваться относительно этого baseline.

Требование к построению baseline:

1. Простота формирования (чем проще тем лучше)
2. Экономический эффект baseline должен быть выше, чем стратегия по принципу “купи и держи” (см. раздел “Оценка экономического эффекта”)
3. Решение должно соответствовать рассматриваемой стратегии

Выбранная базовая стратегия “Колебания стоимости акций в периоды дивидендных выплат” (см. [п. 3.1.1.](#_heading=h.z337ya)) позволяет быстро построить базовую ("прикидочную") модель для определения baseline, без необходимости применения машинного обучения.

Шаги построения базовой модели:

1. Загрузка исходных данных: даты закрытия дивидендных реестров + дневные значения Close по акциям компаний из нефтегазовой отрасли.
2. На основе дат закрытия дивидендных реестров (экс-дивиденды) определить дату потенциальной покупки как дата экс-дивиденды - 20 будних дней. Т.е. следуем принципу открываем сделку на покупку акций за 20 будних дней до дат закрытия дивидендных реестров.
3. Дату закрытия сделки на покупку определяем как равную за 2 дня до даты закрытия дивидендных реестров.
4. Получившиеся пары дат открытия-закрытия сделки на покупку проверяем на значениях цен акций.
5. Сравниваем доходность стратегии со средней доходностью по акциям (см. раздел “Оценка экономического эффекта”). В случае если доходность базовой модели выше, чем средняя доходность акций считаем полученные результаты как baseline (“исходный уровень”) оценки эффективности для последующего моделирования.

# Исходные данных

## Описание используемых данных

Для построения модели базовой стратегии (см.2.2.1.) используются следующие данные:

* Дневные котировки акций по выбранным компаниям
* Перечень дат закрытия дивидендных реестров по каждой компании

Для расчета экономической оценки базовой модели дополнительно используются следующие данные:

* Дивидендная доходность акций в год

Все данные собираются с 2017 года по 2021 включительно по 5 компаниям ("ROSN", "LKOH", "GAZP", "SNGS", "TATN").

Описание источников данных:

| Наименование данных | Данные по каким акциям | Период | Где используются | Как используются | Источник данных | Способ получения |
| --- | --- | --- | --- | --- | --- | --- |
| Дневные котировки акций по выбранным компаниям | ROSN  LKOH  GAZP  SNGS  SNGSP  SIBN  TATN  NVTK | 2017 -2021 | * Базовая модель * Модель “купи и держи” * Все исследуемые стратегии | * Расчет доходности базовой стратегии * Расчет доходности стратегии “купи и держи” * Обучение всех созданых моделей | Брокер Московской биржи | Автоматизированная выгрузка (частично ручная) |
| Перечень дат закрытия дивидендных реестров по каждой компании | * Базовая модель | * Расчет дат открытия и закрытия сделки для базовой модели * Определение кол-ва дивидендных выплат | <https://www.dohod.ru/ik/analytics/dividend> | Вручную (данных немного) |
| Дивидендная доходность акций в год | * Модель “купи и держи” | * Расчет доходности стратегии “купи и держи” (необходимо для экономической оценки) | <https://smart-lab.ru/q/shares_fundamental4/?field=div_yield> | Вручную (данных немного) |
| Дневные котировки акций по выбранным компаниям | ROSN  LKOH  GAZP  SNGSP  SIBN  NVTK | Ежедневный | * Стратегия “Корреляция в отрасли | - Ежедневное прогнозирование направление тренда | Брокер Московской биржи | Автоматизированная выгрузка (частично ручная) |

# Валидация данных

## Анализ входного датасета

Первичный анализ данных проведен с использованием библиотеки “Pandas Profiling” (версия 3.0.0).

Результаты анализа доступны по ссылке:

<https://drive.google.com/file/d/12jAyoCBx_TqJhX8etUvVoj4jyEaWAXXR/view?usp=sharing>

По итогам анализа используемого датасета (с использованием ProfileReport) можно сделать следующие выводы:

- пропущенных данных нет;

- данные дат уникальны (так и должно быть, они выступают в роли индекса в датасете);

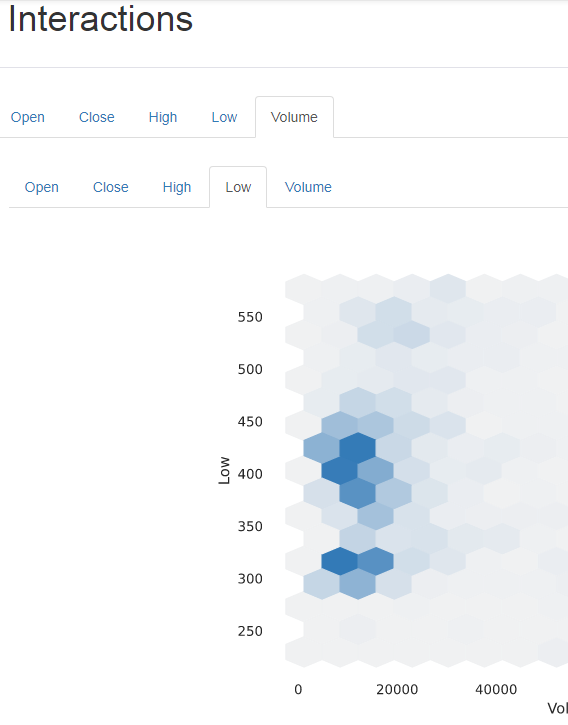
- очевидная корреляция между ценами High,Low,Open,Close

- наличие более сильного "Взаимодействия" данных между Volume и ценами в определенных значениях цены 300 и 400 обуславливается тем, что просто количеством данных с этими диапазонами цен больше, чем с другими ценами.

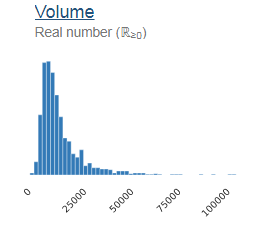
- Объемы (Volume) больше значения 500000 можно считать выбросами. При использовании в стратегиях информации по объемам это необходимо учитывать

Таким образом при первичном анализе можно сказать, что данные корректны, однако зависимостей не выявлено.

**Взаимодействие данных между Volume и ценами в определенных значениях цены 300 и 400.**



**Объемы (Volume) больше 500000 можно считать выбросами.**



## Определение волатильности цен по каждой компании

Волатильность – это изменчивость цены. Резкое падение или рост цены обычно приводит к росту волатильности. Когда цена долго колеблется вокруг определенной отметки, волатильность снижается.

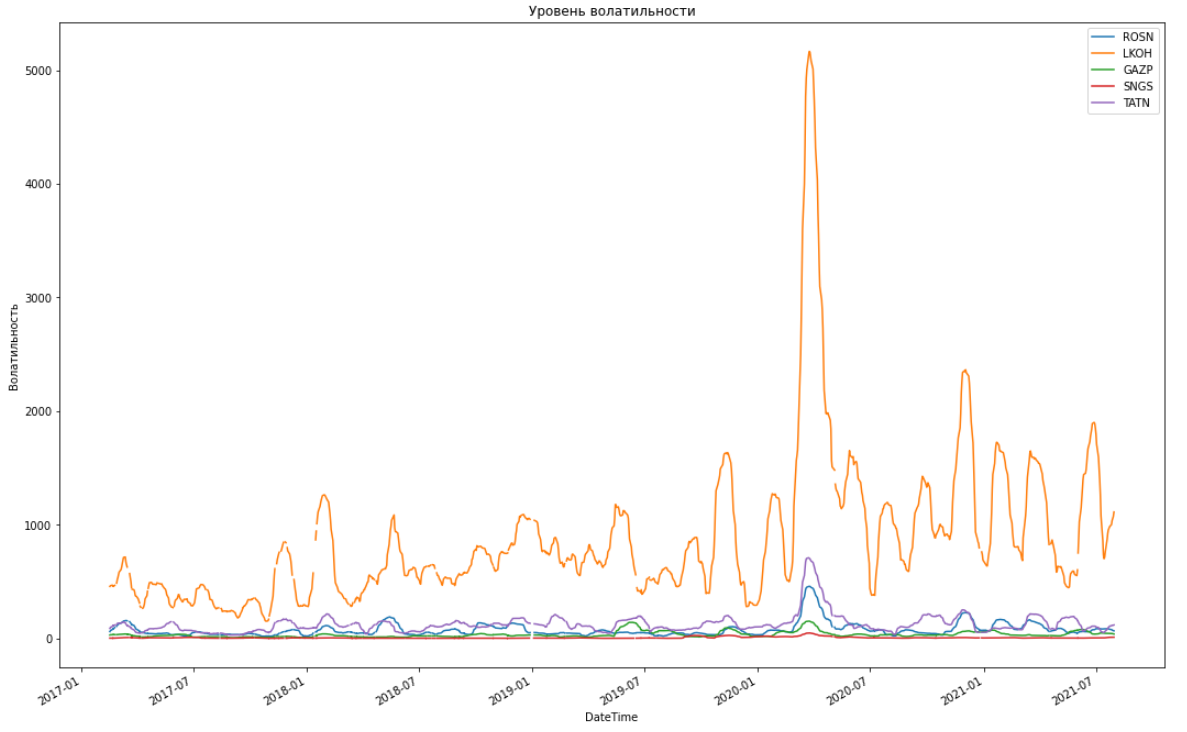
Чем выше волатильность, тем выше риски скачков цены, таким образом волатильность определяет надежность инструмента.

При создании торговых систем с использованием машинного обучение оптимальнее использовать менее волатильные инстурменты как менее рискованные и с более “спокойным” движением цены.

Расчёт исторической волатильности

Волатильность пропорциональна стандартному отклонению доходности финансового инструмента и обратно пропорциональна квадратному корню временного периода :





## По итогам расчета волатильности для используемых инструментов можно сделать вывод, о том: что акции LKOH имеют значительно сильную волатильность по сравнению с другими инструментами. При этом волатильность остальных компаний находится на приемлемом уровне.

Расчет волатильности необходимо делать для всех используемых в проекте инструментов.

## 

## Определение количества и силы гэпов (резкий скачок цены, например из-за дивидендов)

## Определить какие из гэпов дивидендные какие по другим причинам.

# Экономический эффект

## Описание экономической оценки модели

Эффективность торговой стратегии будет оценивается относительно среднедневной доходности финансовых инструментов за один и тот же промежуток времени. Т.е. выполняется сравнение реализованной торговой стратегии с принципом “купи и держи”.

Принцип “купи и держи” подразумевает идею купить акции и ожидать длительный период в расчете получить доход от роста стоимости акций и от выплачиваемых дивидендов. Т.е. доходность определяется как сумма роста цены акций и выплаченных дивидендов.

Расчет доходности осуществляется в процентах. При рассмотрении доходности стратегий и экономической оценки допускается, что при торговле несколькими инструментами инвестиционные денежные средства равномерно распределяются между различными акциями, таким образом проценты доходности одних акций могут сравниваться с доходностью других. При расчете оценки эффективности будет также использоваться доходность в процентах.

При оценке эффективности необходимо учитывать временной период работы стратегии. Так, например, принцип “купи и держи“ использует финансовые активы весь год, т.е. использовать их в других инвестициях не представляется возможным, и. напротив, базовая дивидендная стратегия использует финансовые активы фиксированное кол-во календарных дней умноженное на количество дивидендных выплат в течении года. Таким образом расчет экономической оценки должен учитывать среднюю доходность за день.

Расчет экономической оценки:

* Экономическая оценка (E)
* Доходность (P, Profit)
* Доходность рассматриваемой модели (PM)
* Кол-во календарных дней в году использования фин.активов рассматриваемой моделью (DM)
* Доходность от повышения рыночной цены(PMV, market value)
* Доходность от выплаты дивидендов (DY).
* Годовая доходность принципа “купи и держи” = (PMV + DY)
* Кол-во календарных дней в году использования фин.активов принципом “купи и держи”(всегда 365)

E(за год) = PM/DM - (PMV + DY) / 365 (ед.измерения % доходность в день)

Таким образом экономическая оценка это разность между среднедневной доходностью рассматриваемой модели и среднедневной доходностью принципа “купи и держи”. Если E положительное, тогда рассматриваемая модель экономически эффективнее принципа “купи и держи”. Величина E показывает эффективность модели, чем больше значение, тем больше экономический эффект.

Помимо экономической эффективности дополнительными критериями успешности при построении модели являются:

* количество прибыльных сделок, отражающее стабильность доходности (в большинстве сделок прослеживается четко выраженная доходность).
* количество совершенных сделок.
* минимальная просадка в убыток, как по времени так и по величине (в период работы любой торговой стратегии допустимо краткосрочный уход в убыток с последующим возвращением в доходность)

Описание данных оценок представлено ниже.

## Расчет экономической оценки базовой модели

Для расчета экономической оценки базовой модели выбран 5-летний период с 2017 по 2021. В качестве рассчитываемых параметров будут использоваться средние значения за 5 лет. Это позволит проводить единую оценку моделей на различных временных периодах.

E(базовой модели) = PM/DM - (PMV + DY) / 365

Доходность в процентах по принципу "купи и держи"

|  | **2017** | **2018** | **2019** | **2020** | **2021** | Сумма | Среднее |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ROSN** | -25,73% | 45,16% | 5,605 | -4,77% | 24,89% | **45,15%** | **9,03%** |
| **LKOH** | -0,91% | 46,05% | 23,215 | -17,87% | 20,84% | **71,32%** | **14,26%** |
| **GAZP** | -15,31% | 16,11% | 64,065 | -17,77% | 35,09% | **82,18%** | **16,43%** |
| **SNGS** | -11,99% | -3,69% | 88,395 | -29,95% | -10,08% | **32,68%** | **6,53%** |
| **TATN** | 13,06% | 51,58% | 3,815 | -33,86% | -5,31% | **29,28%** | **5,85%** |
| **Среднегодовая рыночная доходность по всем акциям за весь период (PMV):** | | | | | | | **10,42%** |

На основе указанных данных о доходности можно сделать следующие вывод:

* В 2017 и 2020 годах доходность по принципу “купи и держи” была убыточна
* Не для всех акций данный принцип работает успешно, например, у “SNGS” прослеживается резкий перепад годовых доходностей, что мешает делать долгосрочные прогнозы

Доходность от выплаты дивидендов в процентах по принципу "купи и держи":

|  | **2017** | **2018** | **2019** | **2020** | **2021\*** | Сумма | Среднее |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ROSN | 3,6% | 6,0% | 7,4% | 1,6% | 1,6% | **20,20%** | **4,04%** |
| LKOH | 6,4% | 5,0% | 8,8% | 5,0% | 5,0% | **30,20%** | **6,04%** |
| GAZP | 6,2% | 10,8% | 5,9% | 5,9% | 5,9% | **34,70%** | **6,94%** |
| SNGS | 2,3% | 2,4% | 1,3% | 1,9% | 1,9% | **9,80%** | **1,96%** |
| TATN | 8,3% | 11,5% | 8,5% | 4,3% | 4,3% | **36,90%** | **7,38%** |
| **Среднегодовая доходность от выплаты дивидендов (DY):** | | | | | | | **5,27%** |

\*Доходность по дивидендам за текущий год (2021) не может быть определена до окончания года, поэтому для 2021 года используются данные 2020 года

Доходность в процентах по базовой дивидендной стратегии

|  | **2017** | **2018** | **2019** | **2020** | **2021** | Сумма | Среднее |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ROSN** | 9,64% | 17,52% | 3,23% | 15,80% | 0,48% | **46,67%** | **9,33%** |
| **LKOH** | 8,03% | 13,52% | 5,87% | 11,01% | 3,25% | **41,68%** | **8,34%** |
| **GAZP** | 6,44% | 8,39% | 5,76% | 2,21% | 8,25% | **31,05%** | **6,21%** |
| **SNGS** | 8,85% | 0,66% | 5,97% | -2,50% | -6,89% | **6,09%** | **1,22%** |
| **TATN** | 11,58% | 0,43% | 9,69% | -10,51% | 3,55% | **14,74%** | **2,95%** |
| **Среднегодовая доходность базовой модели по всем акциям за весь период (PM):** | | | | | | | **5,6%** |

На основе указанных данных о доходности можно сделать следующие вывод:

* Доходность базовой дивидендной стратегии является более стабильной чем принцип “купи и держи”
* Дивидендная стратегии не всегда является выгодной так например для “SNGS” два года подряд данная стратегия являлась убыточной.

Количество дней в году использования фин.активов рассматриваемой моделью (DM) зависит от количества дивидендных выплат за рассматриваемый период.

Количество дивидендных выплат с 2017 по 2021 года.

|  | **2017** | **2018** | **2019** | **2020** | **2021** | Сумма | Среднее |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ROSN** | 2 | 2 | 2 | 1 | 1 | 8 | 1,6 |
| **LKOH** | 2 | 2 | 2 | 2 | 1 | 9 | 1,8 |
| **GAZP** | 1 | 1 | 1 | 1 | 1 | 5 | 1 |
| **SNGS** | 1 | 1 | 1 | 1 | 1 | 5 | 1 |
| **TATN** | 2 | 3 | 3 | 1 | 1 | 10 | 2 |
| **Среднее за год кол-во дивидендных выплат одной компанией (DM):** | | | | | | | **1,48** |

Среднее кол-во дивидендных выплат по всем акциям за 5 лет составляет 1,48 в год . Базовая модель подразумевает покупку акций за 20 рабочих дней и продажу за два рабочих дня (включительно) до даты экс-дивидендов, что составляет 19 дней. Формула перевода рабочих дней в календарные: КД = 7/5 \* РД

Таким образом:

**DM** = Среднее кол-во див.выплат \* на кол-во календарных дней = **1,48 \* 7/5\* 19** = **1,48 \* 26,6** = **39,37**

Дивидендная модель задействует денежные средства в среднем на 39,37 кал.дней в году, тогда как принцип “купи и держи” 365 кал.дней.

Итого:

* Доходность базовой модели (PM): **5,6%**
* Кол-во дней в году использования фин.активов базовой моделью (DM): **39,37**
* Среднегодовая доходность от повышения рыночной цены по принципу “купи и держи” (PMV): **10,42%**
* Доходность от выплаты дивидендов (DY): **5,27%**

**E(базовой модели)** = PM/DM - (PMV + DY) / 365 = **5,6/39,37** - (**10,42 + 5,27)/365 =**

**= 0,142**- **0,042 = 0,1% в день**,

где

- PM/DM - среднедневная доходность базовой модели составляет **0,142% в день**

- (PMV + DY)/365 - среднедневная доходность “купи и держи” составляет **0,042% в день (15,33% годовых)**

**Экономическая эффективность базовой модели относительно принципа “купи и держи” составляет 0,1% в день .**

Эффективность построенных в дальнейшем модели будет сравниваться с экономической эффективность базовой модели.

Необходимо отметить, что для большинства стратегий можно определить годовую доходность, умножив дневную доходность на 365, однако особенность дивидендной стратегии заключается в том, что она работает только в периоды выплаты дивидендов, а не в течении всего года и определение годовой доходности осуществляется по усложненной схеме и учитывает периоды див.выплат всех акций.

Важно отметить, что все расчеты применимы только к выбранному пулу финансовых инструментов (акции 5-ти нефтегазовых компаний).

## Метрики и оценка качества моделей. Фиксация baseline.

Основная оценка последующих экспериментальных моделей будет производиться по среднедневной доходности.

Среднедневная доходность Базовой модели составляет **0,142%**, что является довольно высоким показателем (можно выразить как 51,83% годовых, однако к базовой модели не применим перевод в годовую доходность).

Среднедневная доходность принципа “купи и держи” составляет **0,042%** (15,33% годовых).

Таким образом:

* если у анализируемой модели среднедневная доходность выше чем у Базовой модели – значит модель построена успешно и полученный результат учитывается в дальнейшей работе.
* если у анализируемой модели среднедневная доходность ниже чем у Базовой модели, но при этом выше, чем у принципа “купи и держи” – модель построена удовлетворительно, надо искать способы улучшения.
* если у анализируемой модели среднедневная доходность ниже, чем у принципа “купи и держи” – модель считается неприемлемой, требуется либо отказаться от модели, либо внести критические изменения.

Для “успешно” и “удовлетворительно” построенных моделей также производится оценка дополнительных параметров качества модели:

* количество прибыльных сделок, отражающее стабильность доходности. В большинстве совершенных сделок прослеживается четко выраженная доходность. Нестабильная доходность выражена наличием большого количества убыточных сделок, даже если при этом итоговая доходность положительная.
* количество совершенных сделок. Слишком малое количество сделок является отрицательным показателем стратегии (например если за год всего две сделки, то это неудовлетворительно).
* минимальная просадка в убыток, как по времени так и по величине (в период работы любой торговой стратегии допустимо краткосрочный уход в убыток с последующим возвращением в доходность).

По итогам анализа каждой модели формируется аналитический отчет с включением произведенных оценок экономической эффективности и качества моделей.

## Расчет и сравнение просадок базовой модели и стратегии “купи и держи”

Просадка акций – это падение цены акции относительно её предшествующего максимального значения.

По сути, просадка акций это есть потенциальное уменьшение капитала инвестора.

При этом бывает два типа просадки:

* Максимальная историческая просадка
* Текущая просадка

Максимальная историческая просадка – это максимальное снижение цены акций, выраженное в процентах, относительно своего предшествующего исторического максимума.

Текущая просадка акций – это величина текущего снижения акций относительно их предшествующего исторического максимума. Величина текущей просадки определяется как разница между предшествующей абсолютной максимальной ценой и текущей ценой акции.

Нас будет интересовать максимальная историческая просадка и максимальная просадка после покупки в период работы стратегии. В качестве критерия хорошей стратегии введём среднедневную просадку, минимизация которой приведёт к качественному результату.

Таким образом, расчёты будут проводиться по формулам:

* для исторической просадки – **DDh** = (max - min\_after) / max / period
* для просадки после покупки – **DDp** = (first - min) / first / period,

где

* max – максимальная цена за указанный период
* min\_after – минимальная цена, которая достигалась после max
* first – цена в начале периода (цена покупки)
* min – минимальная цена за указанный период
* period – время работы стратегии (365 для "купи и держи", ~40 для дивидендной)

Для базовой стратегии:

* **DDh = 0.000720**
* **DDp = 0.000501**

Для дивидендной стратегии:

* **DDh = 0.001064**
* **DDp = 0.000772**

Как мы видим, обе метрики меньше (то есть лучше) у базовой стратегии, которая, в свою очередь, является менее доходной. Эти результаты согласуются с утверждением о наличии более высокой доходности у более рисковых стратегий

# Стратегии и гипотезы

Моделирование производится на основе выбранной торговой стратегии. В рамках реализации проекта будет осуществлен анализ различных стратегий, при моделировании которых будут использоваться подходы машинного обучения.

Моделирование каждой стратегии будет включать анализ набора признаков, формирование и различное разбиение входных данных, подбор архитектуры модели и выбор гиперпараметров.

При построении модели выполняются следующие виды оценок эффективности:

1. Наименее значимая оценка. Оценка на тестовых данных через встроенные функции tensorflow - метод evaluate, в который подается входные данные X и реальные ответы (Y/таргеты). evaluate производит на X производит расчет прогнозируемых ответов (прогнозируемые Y/таргеты) и осуществляет сверку с реальными ответами. Результат сверки выдается в соответствии с построенной модели в виде accuracy.
2. Средне значимая оценка. Формирование матрицы ошибок (Confusion matrix). В рамках проекта прогнозирования движения цены точность того что предсказанное значение Y соответствует действительности (precision) важнее, чем факт выявления наибольшего числа верных Y (recall). Построение матрицы ошибок будет осуществляется собственной функцией, т.к. необходимо учитывать порог при котором предсказанное значение Y будет отнесено либо к классу 1 (сигнал на рост цены) либо к классу 0 (сигнал на падение цены). Порог используется в качестве более показателя уверенности модели в предсказании (например к классу 1 относятся только те предсказания которые имеют значение выше 0.6 ,а к классу 0 все значения которые ниже 0.4). Все значения не вошедшие в пороговые значения (у которых значения между 0.4 и 0.6) помечаются как неопределенные и не участвуют в построении матрицы ошибок Confusion matrix.
3. Наиболее значимая оценка. Оценка экономической эффективности модели согласно правилам определенным в соответствующем разделе.

Каждая стратегия для оценки экономической эффективности должна определять дату открытия сделки (на покупку/long или на продажу/short) и дату закрытия сделки.

В качестве определения даты закрытия сделки возможно использовать следующие подходы:

| Наименование | Описание | Плюсы | Минусы |
| --- | --- | --- | --- |
| Константное время закрытия | Использовать константное время закрытия сделки (например если для прогнозирования используется дневной таймфрейм, то константное время закрытия сделки можно устанавливать 1-2 дня, либо например открытие сделки утром в прогнозный день, а закрытие вечером того же дня). | * Простота реализации | * Непредсказуемость движения цены. Возможно закрытие в невыгодной позиции |
| Пороги Stop Loss и Take Profit | Использовать пороги Stop Loss (цена снизилась ниже порога - фиксируем потери и закрываем сделку) и Take Profit (цена повысилась и достигла заданного порога - фиксируем прибыль и закрываем сделку) | * Простота реализации * Применимость при относительно высокой волатильности (т.е. если порог SL задан больше чем средняя волатильность, тогда средние падения цены не влияют на срабатывание порога) | * Тяжело выбрать оптимальные значения порогов, либо могут быть * При резких падениях цены ниже порога идет закрытие в убыток (StopLoss) даже при росте общего среднего тренда * При сильном росте тренда порог Take Profit не ограничит возможную прибыль (проблема упущенной выгоды) |
| Плавающие пороги закрытия Trailing Stop Loss | Аналог подхода “Пороги Stop Loss и Take Profit” но порог Stop Loss является динамическим, т.е. при повышении цены порог Stop Loss также повышается, это позволяет извлечь больше потенциальной прибыли. Возможно улучшение подхода за счет усложнения алгоритма динамического изменения порогов. | * Защита от проблем упущенной выгоды как при обычных порогах рост | * При сильных резких падениях цены идет закрытие в убыток (StopLoss) даже при росте общего среднего тренда * Сложность реализации |
| Использование сигнала индикаторов для выбора времени закрытия | Существует множество описанных стратегий по закрытию сделки по сигналу от индикаторов (При этом могут использоваться различные индикаторы) | * Большое количество описанных стратегий по закрытию сделок с использованием индикатора | * Сложность реализации * Необходимость наличия дополнительных источников данных в виде сигналов индикаторов * Необходимость выбора подходящих стратегий из множества доступных |
| Использование той же модели машинного обучения, что и для определения времени открытия | Определение даты закрытия с использованием той же модели, с помощью которой определяется сигнал на открытие сделки. Т.е. когда модель прогнозирует, что будет рост - открываем сделку на покупку, пока модель предсказывает рост сделка не закрывается, как только идёт прогноз на окончание роста, тогда сделка закрывается. | * Простота реализации | * Возможны частые ложные срабатывания, т.е. закрытие сделки сразу после открытия, даже при общем росте тренда. |
| Создание отдельной модели машинного обучения для определения сигнала на закрытие | Построение полноценной модели машинного обучения, с проведением отдельного обучения и подбора архитектуры и гиперпараметров | * Эффективность выбора сигнала закрытия измерима и полностью зависит от эффективности построенной модели | * Сложность реализации |

Оценка исследуемых моделей осуществляется по следующим показателям:

* accuracy - точность полученная на валидационных данных при обучении модели. Accuracy на обучаемых данных не используется, т.к. при обучении модели склонны к переобучению, что приводит к неверному определению точности.
* evaluate accuracy - точность полученная встроенными средствами tensorflow при проверке на тестовых данных.
* Оценка точности Precision, также известное как PPV (positive predictive value) по матрице ошибок (Confusion Matrix). Расчет матрицы ошибок разработан с учетом анализа пороговых значений. Порог используется в качестве более показателя уверенности модели в предсказании (например к классу 1 относятся только те предсказания которые имеют значение выше 0.6 ,а к классу 0 все значения которые ниже 0.4). Все значения не вошедшие в пороговые значения (у которых значения между 0.4 и 0.6) помечаются как неопределенные и не участвуют в построении матрицы ошибок Confusion matrix. Если порог указан 0.5, тогда матрица ошибок работает по классической схеме.
* Оценка NPV (negative predictive value) по матрице ошибок (Confusion Matrix).
* Экономическая оценка на примере среднедневной доходности. Рассчитывается согласно правилам раздела оценка экономической эффективности. При этом правилом открытия сделки является наличие двух подряд уверенных сигнала на открытие, аналогично для закрытия сделки необходимо от модели получить два уверенных сигнала на закрытие.
* Количество совершенных сделок.
* Средняя длительность сделки. Слишком длительные сделки демонстрируют низкое качество модели (например если модель строится на дневном таймфреме, а средняя продолжительность сделок 30 дней, либо вероятнее всего модель неверно выявляет направления трендов). Аналогично если среднее время сделки слишком маленькое (например 1 день), тогда также высокая вероятность неверного определения трендов.
* Количество прибыльных сделок, отражающее стабильность доходности. Нестабильная доходность выражена наличием большого количества убыточных сделок, даже если при этом итоговая доходность положительная. Измеряется процентное соотношение прибыльных сделок к общему числу сделок (прибыльные и убыточные).
* Самая убыточная сделка. Сделка с самым большим убытком.
* Максимальная просадка (Drawdown) разница между ценой покупки и самым нижним значением цены в период совершения сделки.

## Стратегия “Корреляция различных финансовых инструментов в общей отрасли”.

### Общее описание:

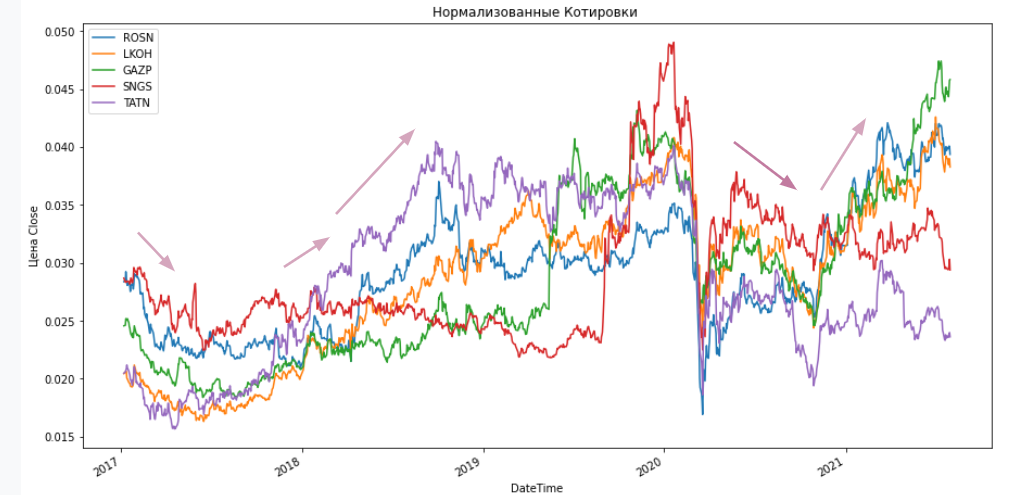
Стоимость акций сильно зависят от внешних факторов (политические, экономические, новостные и прочие факторы). При этом компании одной отрасли могут быть связаны общими событиями и менять тренды движения котировок в общую сторону. Например, при выходе положительной новости про нефтяной рынок стоимость акций нефтяных компаний вероятно будет подниматься на возрастающем тренде. Однако разные компании реагируют на изменения с различной скоростью. Выявление компаний реагирующих на изменения рынка с различной скоростью позволит предсказывать изменение стоимости акций у компаний, которые медленно реагируют на изменение цены на фоне компаний которые реагируют быстро.

Другими словами стратегия “Корреляции акций компаний общей отрасли” подразумевает выявление таких компаний у которых еще не изменилась стоимость акций на фоне ярко выраженных изменений других компаний этой же отрасли. Например, при выходе положительной новости про нефтяной рынок стоимость акций 5 из 6 нефтяных компаний выросла, таким образом можно сделать предположение, что у последней 6-ой компании также в скором времени изменится стоимость акций.

Ноутбук стратегии: <https://colab.research.google.com/drive/17LwOcrnJuMRzPBIdEzpYbzRvns9X69k3#scrollTo=A-0TQPLzhBuG>

### Анализ рынка на применимость стратегии:

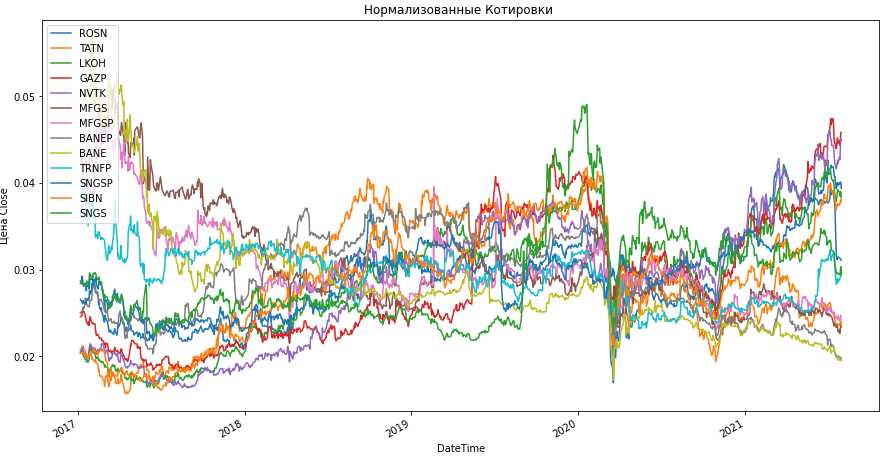
Первичный анализ применимости стратегии “Корреляции акций компаний общей отрасли” на дневных таймфреймах с 2017 по 2021 года для нефтяных компаний: 'ROSN', 'LKOH', 'GAZP', 'SNGS', 'TATN'. Визуализация нормализованных цен акций (нормализация нужны для возможности сравнения различных акций).



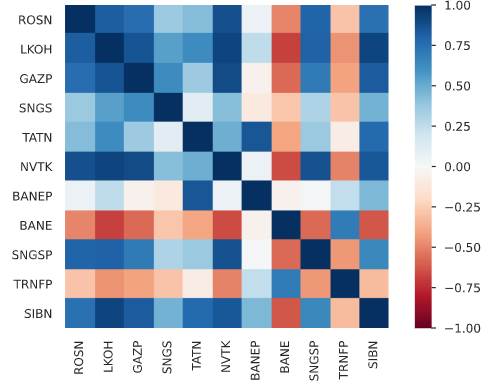
По итогам визуализации котировок по различным компаниям можно сделать вывод, что имеет место наличие общих трендов. Следующий шаг анализа требует:

* добавить в анализ дополнительные российские компании нефтяной отрасли.
* более детально проанализировать общие трендовые движения на коротком промежутке времени.
* изучить конкретные примеры роста/ падения тренда на предмет, особенности именно выбранной отрасли, а не факт движения всего фондового рынка (например из-за кризисных ситуаций).
* выявить примеры компании, которые реагируют на изменения отраслевого рынка с “задержкой”.

Добавление дополнительных нефтяных компаний усложняет анализ, но позволяет выявить больше возможных корреляций

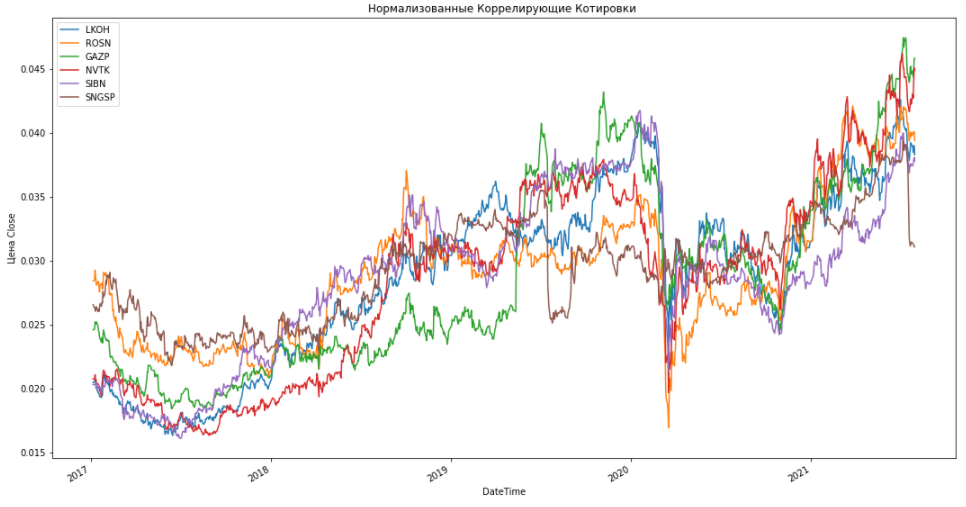


Над выбранными символами (компаниями) проведена проверка корреляции нормализованных значений котировок:



Выявлена сильная корреляция между: LKOH, ROSN, GAZP, NVTK, SIBN, SNGSP. С учетом чего можно сделать вывод, что совпадения между трендами связано не только с общим рынком, но также и с отраслевыми событиями (Если бы зависело только от общего рынка, тогда бы корреляция была бы у всех инструментов).

Дальнейший анализ будет производиться с выбранными символами.



Для выбранных коррелирующих инструментов визуально также прослеживается совпадение направлений трендов.

Далее необходимо выявить примеры компании, движение котировок которых возможно определить на примере движения котировок других компаний этой же отрасли (например, акции которые реагируют на изменения отраслевого рынка с “задержкой” относительно реагирования других компаний).

Визуальный анализ довольно трудоемкий, в связи с чем поиск зависимых компаний будет производиться через моделирование машинного обучения. Выбор зависимых компаний будет осуществляться через поиск наибольшего процента предсказания на валидационных данных. Будут применены три вида модели: 1. только слои Dense 2. Слои LSTM и Dense. 3. Слои Conv1D и Dense. Для LSTM и Conv1D входные данные поступают в виде векторов значений. Модели будут строиться как бинарные классификаторы, где необходимо определить движение цен пойдет вверх или вниз. Искомой метрикой для валидационных данных выбрана метрика accuracy (так как число выборки для двух классов примерно одинаковое, если бы порядок кол-ва элементов сильно различался бы, тогда необходимо было бы применять метрику precision). Модели для поиска зависимых компаний будут максимально упрощенными, т.к. для задачи определения наиболее зависимых компаний необходимо определить относительную эффективность, а не абсолютную.

Для поиска зависимых компаний, каждая компания из рассматриваемого набора будет по очереди выступать в роли компании для которой осуществляется предсказание направления движения цены:

Рассматривается весь период с 2017 по 2021 года. Каждая модель запускается как минимум дважды с 200 эпохами.

| Наименование | Dense | LSTM+Dense | Conv1D+Dense |
| --- | --- | --- | --- |
| **LKOH** | **0.6174** | **0.6261** | **0.6000** |
| ROSN | 0.5348 | 0.5652 | 0.5739 |
| GAZP | 0.5391 | 0.5478 | 0.5826 |
| NVTK | 0.5522 | 0.5478 | 0.5391 |
| **SIBN** | **0.6043** | **0.6261** | 0.5913 |
| SNGSP | 0.5696 | 0.5130 | 0.5565 |

\* при экспериментах accuracy достигло значения 0.6261 (LKOH и SIBN на модели LSTM)

По итогам построения первичных моделей, можно сделать вывод, что наиболее зависящими от других символов в своей отрасли являются: LKOH, SIBN. Эти символы будут использованы в моделировании.

### Подбор архитектуры модели:

При подборе архитектуры будут рассматриваться архитектуры с использованием слоев Dense (полносвязанный слои), LSTM (разновидность рекуррентных сетей для анализа временных рядов), Conv1D (сверточные слои).

Подбор архитектуры будут осуществляться для символа **LKOH**. Рассматривается весь период с 2017 по 2021 года. Каждая модель запускается как минимум дважды с 200 эпохами.

**Dense (Упрощенная модель):**

| **№** | **Модель** | **Лучший результат accuracy:** |
| --- | --- | --- |
| 1 | Dense(10, activation='relu', kernel\_initializer='he\_normal')  Dense(8, activation='relu', kernel\_initializer='he\_normal')  Dense(1, activation='sigmoid') | **0.6174** |
| 2 | Dense(100, activation='relu', kernel\_initializer='he\_normal')  Dense(1, activation='sigmoid') | **0.6521 (стабильный результат)** |
| 3 | Dense(400, activation='relu', kernel\_initializer='he\_normal')  Dense(20, activation='relu', kernel\_initializer='he\_normal')  Dense(1, activation='sigmoid') | **0.6608 (нестабильный результат)** |
| 4 | Dense(800, activation='relu', kernel\_initializer='he\_normal')  Dense(1, activation='sigmoid') | **0.6521 (стабильно)**  **0.6695/0.6782 (нестабильно)** |
| 5 | Dense(800, activation='relu', kernel\_initializer='he\_normal')  BatchNormalization()  Dense(1, activation='sigmoid') | **0.6521/0.6695(стабильно)**  **0.6782/0.6869 (нестабильно)** |
| 6 | Dense(800, activation='relu', kernel\_initializer='he\_normal')  Dropout(0.1)  BatchNormalization()  Dense(1, activation='sigmoid') | **0.6521/0.6695(стабильно)**  **0.6869/0.6956 (нестабильно)**  **Данная архитектура модели для Dense принимается как предварительно итоговая.** |

**LSTM (Упрощенная модель):**

Выбор длины входного вектора. Длина входного вектора соответствует временным отсчетам, т.е. если таймфрейм выбран как дневной, то длина входного вектора 5 будет соответствовать 5 дням.

| **№** | **Модель** | **Лучший результат accuracy:** |
| --- | --- | --- |
| Выбор длины входного вектора | | |
| 1 | #Длина входного вектора 5  LSTM(5, activation='tanh')  Dense(1, activation='sigmoid') | **0.60/0.6173 (стабильно)**  **0.6260 (нестабильно)** |
| 2 | #Длина входного вектора 7  LSTM(5, activation='tanh')  Dense(1, activation='sigmoid') | **0.6228 (стабильно)**  **0.6666/0.6754 (нестабильно)** |
| 3 | #Длина входного вектора 10  LSTM(5, activation='tanh')  Dense(1, activation='sigmoid') | **0.5877/0.6140 (стабильно)**  **0.6228 (нестабильно)** |
| 4 | #Длина входного вектора 8  LSTM(5, activation='tanh')  Dense(1, activation='sigmoid') | **0.6140/** **0.6315 (стабильно)**  **0.6666/0.6754 (нестабильно)**  **По итогам длина входного вектора для LSTM по умолчанию будет использоваться 8 элементов (соответствует 8 дням)** |
| Подбор архитектуры LSTM | | |
| 5 | LSTM(7, activation='tanh')  Dense(1, activation='sigmoid') | **0.6228/0.6403 (стабильно)**  **0.6578/0.6666 (нестабильно)** |
| 6 | LSTM(7, activation='tanh')  LSTM(5, activation='tanh')  Dense(1, activation='sigmoid') | **0.6228/0.6403 (стабильно)**  **0.6578 (нестабильно)** |
| 7 | LSTM(20, activation='tanh')  LSTM(5, activation='tanh')  Dense(1, activation='sigmoid') | **0.6228 (стабильно)** |
| 8 | LSTM(20, activation='tanh')  LSTM(10, activation='tanh')  Dense(1, activation='sigmoid') | **0.6315/0.6403 (стабильно)**  **0.6666 (нестабильно)** |
| 9 | LSTM(20, activation='tanh')  BatchNormalization()  LSTM(10, activation='tanh')  BatchNormalization()  Dense(1, activation='sigmoid') | **0.6140/0.6403(стабильно)**  **0.7017 (нестабильно)** |
| 10 | LSTM(20, activation='tanh')  Dropout(0.1)  LSTM(10, activation='tanh')  BatchNormalization()  Dense(1, activation='sigmoid') | **0.6315/0.6403/0.6578 (стабильно)**  **0.6754/0.6842/0.7037 (нестабильно)**  **Данная архитектура модели для** LSTM **принимается как предварительно итоговая.** |
|  |  |  |

**Conv1D (Упрощенная модель):**

| **№** | **Модель** | **Лучший результат accuracy:** |
| --- | --- | --- |
| 1 | Conv1D(128,3, activation='relu' )  MaxPool1D(pool\_size=2)  Conv1D(64,3, activation='relu' )  MaxPool1D(pool\_size=2)  Conv1D(32,3, activation='relu' )  BatchNormalization()  Dense(1, activation='sigmoid') | **0.6621 (нестабильно)** |
| 2 | Conv1D(20,3, activation='relu')  MaxPool1D(pool\_size=2)  Conv1D(10,3, activation='relu' )  BatchNormalization()  Dense(1, activation='sigmoid') | **0.6018/0.6296 (стабильно)**  **0.6574/0.6759 (нестабильно)**  **Данная архитектура модели для** Conv1D **принимается как предварительно итоговая.** |

### Анализ детализации входных данных

В качестве входных данных были выбраны дневные отсечки (значения цены раз в день), что достаточно для обучения и валидации данных, однако для качественного тестирования модели этих данных недостаточно. Данные используются с 2017 по 2021 год.

При использовании дневных данных используется следующее распределение входных данных:

* обучение: 958 значений,
* валидационные данные: 108 значений,
* тестовые данные: 55 значений.

Для проведения более качественного анализа работы модели, необходимо увеличить объем датасета. Увеличение датасета возможно произвести за счет перехода на меньший таймфрейм: например вместо дневных отсчетов будут использоваться четырехчасовые отсчеты.

При переходе вместо дневного таймфрейма на четырехчасовой объем входных данных значительно увеличился до следующих показаний:

* обучение: 2776 значений,
* валидационные данные: 310 значений,
* тестовые данные: 224 значений.

Соответственно корреляция между различными компаниями сохраняется.

Однако точность моделей сильно упала:

* для модели Dense стабильных показаний accuracy 0.65/0.66 до 0.56.
* для модели LSTM стабильных показаний accuracy 0.64/0.65 также до 0.56.

Данное поведение возможно объяснить по следующим причинам:

1. Ценовые котировки изменяются внутри дня, что дает более частые колебания между повышением и снижением цены и как следствие увеличивается частота изменения целевой функции (Y/таргета). Т.е. например при уверенном возрастающем тренде внутри дня цена может уходить ниже предыдущей, что для фиксации целевого класса будет означать нисходящий тренд, что в свою очередь будет противоречить общему восходящему тренду. Данная проблема может быть решена за счет реализации более сложного определения целевого класса (например на основе не только текущей и предыдущей цены а используя большее кол-во цен и используя средние значения цены).
2. Выбранные модели плохо работают с таймфреймами ниже дневных. Либо из-за особенности самих данных, либо из-за архитектуры модели. Решается за счет изменения архитектуры моделей.
3. Также допустимо, что проблема является более критической и высокие проценты точности на дневном таймфреме являются проблемой малого количества данных. Решается только изменением стратегии либо полного пересмотра архитектуры моделей.

По итогам проведенного анализа принято решение остановиться на дневном таймфрейме.

### Анализ сегментации данных.

В рамках анализа сегментации данных проверяется предположение, что исходные данные расположенные ближе (по временной шкале) к прогнозируемым периодам больше влияют на прогнозирование, чем данные расположенные дальше (по временной шкале) от прогнозируемых периодов.

Проверка заключается в обучение модели на входных данных за различные периоды времени и проверка лучших значений accuracy:

| Модель | 2017-2021 | 2018-2021 | 2019-2021 | 2020-2021 |
| --- | --- | --- | --- | --- |
| Dense | **0.7064** | 0.7023 | 0.6779 | 0.6764 |
| LSTM | 0.7130 | 0.7229 | 0.7414 | **0.8182** |
| Conv | Для модели Conv не проводился данный анализ | | | |

\* обучающая выборка до 30.04.2021 (из нее 10% это валидационные данные) с 01.05.2021 по 01.08.2021 тестовая выборка.

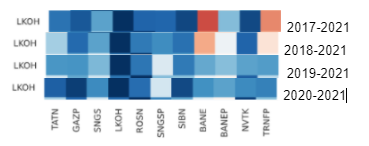
По итогам анализа сегментации входных данных у модели Dense выявлена стабильное ухудшение точности прогноза. Основной причиной снижения точности является уменьшение объема обучающей выборки, т.к. изначально объем обучающей выборки равен примерно 4,5 года, таким образом исключение даже одного года сильно влияет на объем входных данных. При этом модель Dense для предсказания использует только данные предыдущей временной отсечки (цены предыдущего дня).

Напротив модель LSTM при отбрасывании более старых данных демонстрирует повышение оценки точности. Вероятной причиной может служить особенность рынка ценных бумаг, что в последние года у рассматриваемого символа LKOH корреляция с другими компаниями выше чем в предыдущие года.

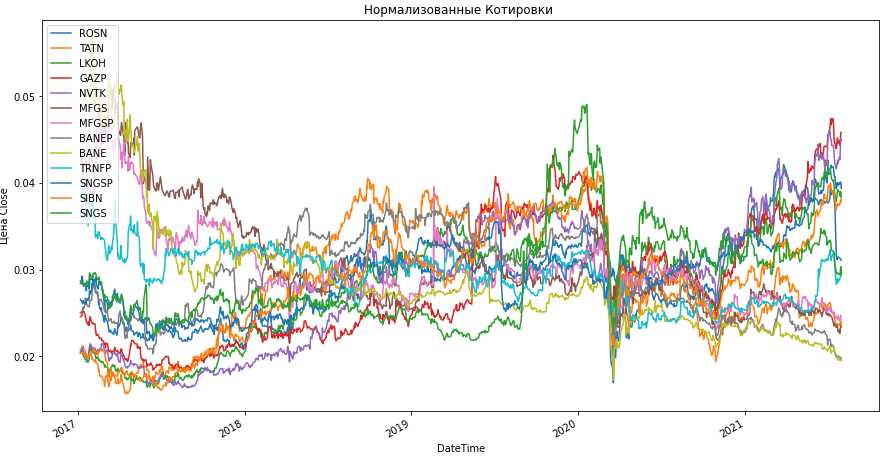
Для более точного определения причины такого поведения проанализировано поведение модели LSTM на временных срезах 2017-2018 (0.55 accuracy) и 2017-2019 (0.6533), таким образом просто уменьшение объема входных данных не влияет на рассматриваемое поведение LSTM.

Причиной такого поведения могут являться сами входные данные.

Анализ корреляции LKOH с другими компаниями в различных срезах



Наблюдается изменение корреляции LKOH с другими отраслевыми компаниями при уменьшении рассматриваемого периода. Данное явление может быть связано с общемировой ситуацией в период пандемии с 2020 по 2021 года, что заставляло все рынки резко падать в 2020 году и в дальнейшем расти в период 2021 года (что также наблюдается при ручной визуализации графика цен).



Таким образом модель LSTM выявила сильную зависимость рынка, не выявив слабую отраслевую зависимость между различными символами. Для подтверждения необходимо произвести оценку моделей по выбранной сегментации данных.

.

### Промежуточная оценка модели LSTM с учетом сегментирования данных

Промежуточная оценка модели LSTM с различным сегментированием данных. Проверка ситуации резкого повышения показателя accuracy при ограничении входных данных. Оценка моделей осуществляется на тестовых данных (тестовая выборка: с 01.05.2021 по 01.08.2021), которые модель не могла видеть ранее.

Оценка осуществляется по следующим показателям:

* accuracy - точность полученная на валидационных данных при обучении модели.
* evaluate accuracy - точность полученная встроенными средствами tensorflow при проверке на тестовых данных.
* Оценка точности Precision, также известное как PPV (positive predictive value) по матрице ошибок (Confusion Matrix). Расчет матрицы ошибок разработан с учетом анализа пороговых значений. Порог используется в качестве более показателя уверенности модели в предсказании (например к классу 1 относятся только те предсказания которые имеют значение выше 0.6 ,а к классу 0 все значения которые ниже 0.4).
* Оценка NPV (negative predictive value) по матрице ошибок (Confusion Matrix).
* Экономическая оценка на примере среднедневной доходности. Рассчитывается согласно правилам раздела оценка экономической эффективности.
* Количество совершенных сделок.
* Средняя длительность сделки.
* Количество прибыльных сделок, отражающее стабильность доходности. Измеряется процентное соотношение прибыльных сделок к общему числу сделок (прибыльные и убыточные).
* Самая убыточная сделка. Сделка с самым большим убытком.
* Максимальная просадка (Drawdown) разница между ценой покупки и самым нижним значением цены в период совершения сделки.

| Модель LSTM по различным срезам | 2017-2021 | 2018-2021 | 2019-2021 | 2020-2021 | 2017-2019 |
| --- | --- | --- | --- | --- | --- |
| accuracy | 0.7130 | 0.7229 | 0.7414 | 0.8182 | 0.6533 |
| evaluate accuracy | 0.5818 | 0.5273 | 0.5636 | 0.3818 | 0.4808 |
| Precision / PPV (positive predictive value) (порог 0.5) | 0.5833 | 0.5143 | 0.5789 | 0.3158 | 0.4808 |
| NPV (negative predictive value) (порог 0.5) | 0.5806 | 0.55 | 0.5556 | 0.4167 | 0.3333 |
| Precision / PPV (positive predictive value) (порог 0.6) | 0.4444 | 0.4828 | 0.6364 | 0.3333 | 0 (нет с таким порогом) |
| NPV (negative predictive value) (порог 0.4) | 0.6538 | 0.5714 | 0.5652 | 0.4815 | 0 (нет с таким порогом) |
| Среднедневная доходность | 0.13% | -0.46% | -0.75% | -0.63% | 0.34% |
| Количество сделок | 4 | 4 | 5 | 5 | 1 |
| Средняя длительность сделки | 6 дней | 10 дней | 3 дня | 3 дня | 29 дней |
| Количество прибыльных сделок | 75% | 25% | 60% | 60% | 100% |
| Максимальная просадка (Drawdown) | 2.83% | 5.5% | 5.22% | 3.1% | 5.22% |

\* на данном этапе оценки, среднедневная доходность (включая просадку и кол-во сделок) может быть не лучшим показателем, т.к. не определены четкие правила открытия и закрытия сделки, используются базовые условия, которые необходимо в дальнейшем пересмотреть.

Оценка LSTM модели при сегментации данных демонстрирует низкую экономическую оценку при высоких показаниях точности (accuracy), что дополнительно подтверждает сильное влияние именно рынка в периоды с 2020 по 2021 года, а не зависимость между отраслями. При дальнейшем моделировании необходимо учитывать данную особенность.

### Сравнительный анализ точности модели и экономической оценки. Анализ воспроизводимости прогноза

Критически важно при реализации модели обеспечить воспроизводимость результатов, т.е. стабильную согласованность обучения модели с итоговой экономической оценкой. Неприемлема ситуация, когда при одинаковых входных данных и одинаковых параметрах модели, но при повторных обучениях модели и запусках прогнозирования итоговая экономическая оценка значительно отличается.

Если такая ситуация встречается необходимо выявить причину такого поведения и при необходимости пересмотреть процесс создания модели от формирования целевых показателей до экономической оценки.

Одним из способов проверки воспроизводимости модели в текущем проекте, является: отражение на графике цен прогнозируемых времен открытия и закрытия сделок при различных обучениях и запусках одной модели. Анализ предлагаемых моделью областей открытия и закрытия осуществляется вручную.

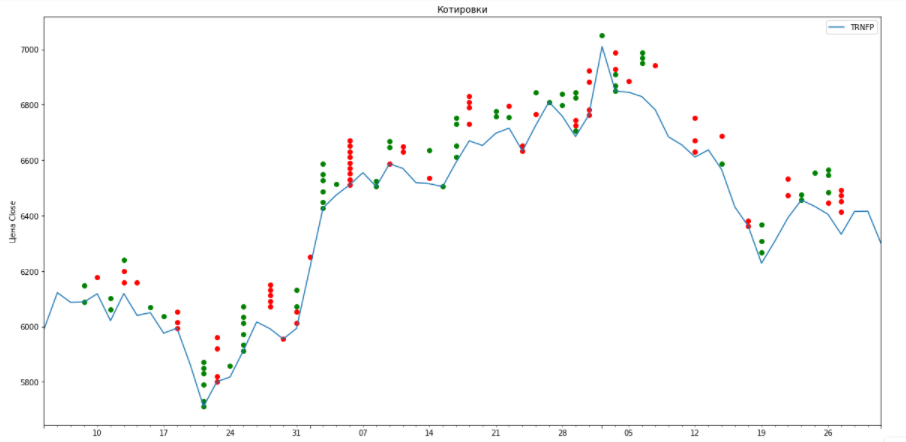
Одновременно с проведением анализа воспроизводимости производится проверка корректности программного кода на предмет скрытых ошибок, а также проверка правил формирования входных и целевых данных.

Данный анализ дополнительно позволяет определить влияние повышения точности модели при обучения на итоговую экономическую оценку (другими словами изменение прибыли при изменении точности модели)

Механизм проверки: по одной архитектуре и на одинаковых данных производиться несколько обучений моделей (вероятнее всего точность обученных моделей будет различаться незначительно). По каждой обученной модели определяются даты открытия и даты закрытия, которые выводятся на единый график. Если большая часть дат открытия/закрытия совпадает, значит модель уверенно воспроизводит предсказания. Анализ проводится на символе LKOH.

Анализ каждой модели производится на тестовых данных (тестовая выборка: с 01.05.2021 по 01.08.2021), которые модель не могла видеть ранее на этапах обучения и валидации.

#### **Воспроизводимость модели Dense:**



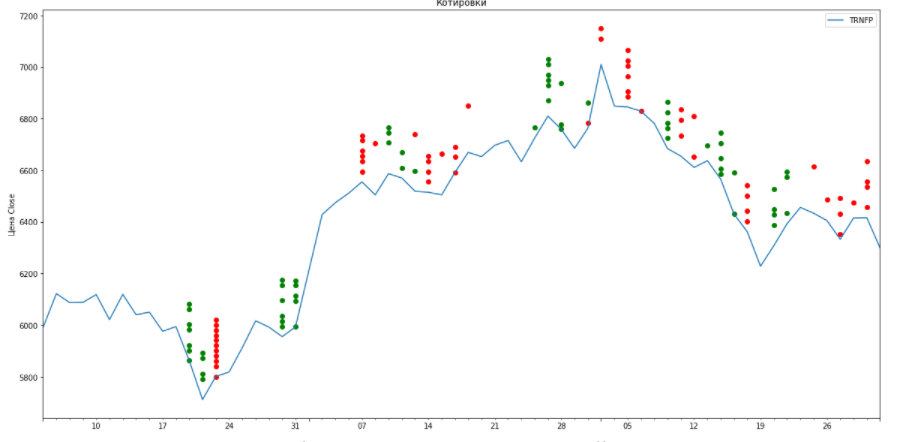
Наблюдается положительная воспроизводимость модели Dense, т.е. модели с разным уровнем точности предсказывают по схожим правилам с некоторой погрешностью.

Сравнительная таблица точности модели и экономической оценки:

| **accuracy** | 0,5505 | 0,5872 | 0,5963 | 0,6055 | 0,6147 | 0,6239 | 0,6606 | 0,6697 | 0,6789 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Среднедневная доходность:** | 0,40% | 0,18% | 0,26% | 0,24% | 0,63% | 0,66% | 0,20% | 0,65% | 0,49% |
| **Кол-во сделок:** | 8 | 9 | 7 | 8 | 8 | 2 | 8 | 7 | 7 |
| **Кол-во прибыльных сделок:** | 62% | 67% | 71% | 62% | 75% | 100% | 62% | 71% | 71% |
| **Худшая сделка:** | -7,11% | -3,47% | -2,35% | -3,47% | -1,13% | 0% | -3,85% | -3,19% | -1,13% |
| **Средняя длительность сделки:** | 6 | 4 | 3 | 4 | 3 | 2 | 3 | 3 | 1 |
| **Максимальная просадка за все сделки:** | 7,61% | 4,03% | 3,54% | 4,03% | 2,10% | 3,31% | 3,91% | 3,75% | 3,31% |

Наблюдается наличие связи точности и экономической оценки, чем выше точность модели, тем выше экономические характеристики модели. При этом встречаются случаи (выбросы) когда при высокой точности доходность снижается.

#### **Воспроизводимость модели LSTM:**



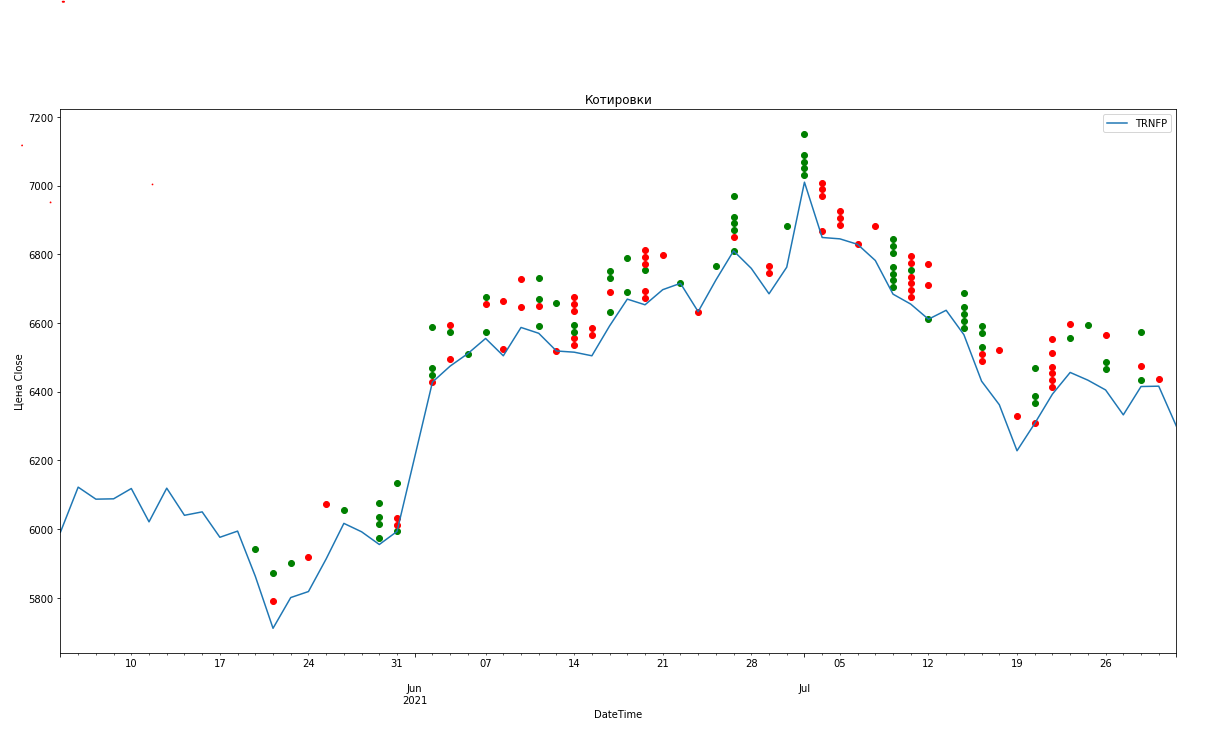
Наблюдается высокая воспроизводимость модели LSTM, т.е. модели с разным уровнем точности предсказывают по схожим правилам с некоторой погрешностью.

Сравнительная таблица точности модели и экономической оценки:

| **accuracy** | 0,5648 | 0,5833 | 0,5926 | 0,6019 | 0,6204 | 0,6296 | 0,6389 | 0,6481 | 0,6759 | 0,6944 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **evaluate accuracy** | 0,5455 | 0,5636 | 0,4909 | 0,6 | 0,5273 | 0,6 | 0,4909 | 0,4545 | 0,5091 | 0,5091 |
| **Precision/PPV (порог 0.5)** | 0,5208 | 0,5484 | 0,4889 | 0,5714 | 0,5238 | 0,6 | 0,48 | 0,4444 | 0,5 | 0,5 |
| **NPV (порог 0.5)** | 0,7143 | 0,5833 | 0,5 | 0,65 | 0,5294 | 0,6 | 0,5 | 0,4643 | 0,5152 | 0,5217 |
| **Precision/PPV (порог 0.6)** | 0 | 0,6 | 0 | 0,875 | 0 | 0 | 0,4444 | 0,4348 | 0,8571 | 0,5 |
| **NPV (порог 0.4)** | 0 | 0,4167 | 0 | 0,5 | 0,5 | 0 | 0,5217 | 0,4118 | 0,4615 | 0,5 |
| **Среднедневная доходность** | 0,26% | -0,59% | -0,17% | 0,08% | 0,12% | -0,07% | 0,40% | 0,24% | -0,14% | 0,43% |
| **Кол-во сделок** | 3 | 5 | 4 | 3 | 5 | 4 | 5 | 6 | 6 | 4 |
| **Кол-во прибыльных сделок** | 67% | 40% | 25% | 67% | 40% | 50% | 60% | 67% | 17% | 75% |
| **Худшая сделка** | -1,70% | -2,58% | -1,07% | -1,06% | -1,06% | -1,06% | -3,33% | -4,40% | -3,15% | -4,82% |
| **Средняя длительность** | 17 | 3 | 10 | 9 | 5 | 7 | 6 | 5 | 4 | 8 |
| **Максимальная просадка** | 3,10% | 3,10% | 3,10% | 3,10% | 3,10% | 3,31% | 6,76% | 4,92% | 3,37% | 5,33% |

Наблюдается наличие связи точности и экономической оценки, чем выше точность модели, тем выше экономические характеристики модели. При этом встречаются случаи (выбросы) когда при высокой точности доходность снижается.

#### **Воспроизводимость модели Conv1D:**



Наблюдается неоднозначная воспроизводимость модели Conv1D, т.е. модели с разным уровнем точности в некоторых случаях предсказания схожи, а в некоторых случаях предсказания противоречат.

Сравнительная таблица точности модели и экономической оценки:

| **accuracy** | 0,5648 | 0,5833 | 0,5926 | 0,6019 | 0,6111 | 0,6296 | 0,6389 | 0,6481 | 0,6667 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **evaluate accuracy** | 0,4909 | 0,4727 | 0,3818 | 0,4 | 0,4909 | 0,4727 | 0,5455 | 0,4727 | 0,5091 |
| **Precision/PPV (порог 0.5)** | 0,4848 | 0,4706 | 0,3793 | 0,3929 | 0,4857 | 0,4643 | 0,5357 | 0,4706 | 0,5 |
| **NPV (порог 0.5)** | 0,5 | 0,4762 | 0,3846 | 0,4074 | 0,5 | 0,4815 | 0,5556 | 0,4762 | 0,5217 |
| **Precision/PPV (порог 0.6)** | 0,5 | 0,5 | 0,3846 | 0,375 | 0,4333 | 0,4815 | 0,5 | 0,4 | 0,5 |
| **NPV (порог 0.4)** | 0 | 0,5 | 0,4444 | 0,4545 | 0,5333 | 0,4583 | 0,5 | 0,2 | 0,6667 |
| **Среднедневная доходность** | 0,37% | -0,05% | 0,16% | -0,38% | 0,05% | -0,19% | -0,17% | 0,12% | -0,52% |
| **Кол-во сделок** | 5 | 7 | 6 | 9 | 7 | 6 | 8 | 9 | 8 |
| **Кол-во прибыльных сделок** | 60% | 57% | 33% | 33% | 43% | 50% | 38% | 56% | 25% |
| **Худшая сделка** | -4,59% | -3,15% | -2,63% | -2,63% | -2,63% | -2,30% | -2,35% | -1,09% | -2,58% |
| **Средняя длительность** | 6 | 5 | 4 | 2 | 5 | 4 | 3 | 4 | 3 |
| **Максимальная просадка** | 6,14% | 3,37% | 5,50% | 5,50% | 5,50% | 3,54% | 3,54% | 3,31% | 3,54% |

Наблюдается наличие общей связи точности и экономической оценки. Однако отсутствует ярко выраженная зависимость. Часто встречаются противоречивые соотношения.

### Анализ входных таймфреймов

В качестве входных данных могут использоваться данные за различный промежуток времени (таймфреймы) от 1 минуты до дней. Использование низких таймфреймов используется при высокочастотной торговле, что не подходит для данного проекта. В текущей стратегии будут рассмотрены следующие таймфреймы: Дневной (D1), четырехчасовой (H4), часовой (H1).

Оценка использования различных таймфреймов. Период данных с 2017 по 2021:

| Модель | **Dense** | | | **LSTM** | | | **Conv1D** | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Таймфреймы | D1 | H4 | H1 | D1 | H4 | H1 | D1 | H4 | H1 |
| **accuracy** | 0,6606 |  |  | 0,6944 |  |  | 0,6481 |  |  |
| **evaluate accuracy** | 0,5625 |  |  | 0,5091 |  |  | 0,4727 |  |  |
| **Precision/PPV (порог 0.5)** | 0,6053 |  |  | 0,5 |  |  | 0,4706 |  |  |
| **NPV (порог 0.5)** | 0,5 |  |  | 0,5217 |  |  | 0,4762 |  |  |
| **Precision/PPV (порог 0.6)** | 0,4167 |  |  | 0,5 |  |  | 0,4 |  |  |
| **NPV (порог 0.4)** | 0,4118 |  |  | 0,5 |  |  | 0,2 |  |  |
| **Среднедневная доходность** | 0,35% |  |  | 0,43% |  |  | 0,12% |  |  |
| **Кол-во сделок** | 7 |  |  | 4 |  |  | 9 |  |  |
| **Кол-во прибыльных сделок** | 86% |  |  | 75% |  |  | 56% |  |  |
| **Худшая сделка** | -1,23% |  |  | -4,82% |  |  | -1,09% |  |  |
| **Средняя длительность** | 4 |  |  | 8 |  |  | 4 |  |  |
| **Максимальная просадка** | 2,28% |  |  | 5,33% |  |  | 3,31% |  |  |

### Подбор гиперпараметров модели

Поиск гиперпараметров и построение архитектуры будут производиться с символами LKOH (основной) и SIBN (дополнительный).

Подбор гиперпараметров

Модель LSTM

Параметр Batch - “Размер пакета”. return\_sequences = False

| Batch | 1 запуск | 2 запуск | 3 запуск |
| --- | --- | --- | --- |
| 10 | 0.6347 (70 эпоха) | 0.57 | 0.53 |
| 12 | 0.6434 (80 эпоха) | 0.60 | 0.54 |
| 14 | 0.55 | 0.60 | 0.6173 (130 эпоха) |
| 17 | 0.6173 (80 эпоха) | 0.59 | 0.56 |
| 20 | 0.60 | 0.60 | 0.55 |
| 22 | 0.54 | 0.58 | 0.6173 (10 эпоха) |
| 24 | 0.59 | 0.53 | 0.58 |
| 32 | 0.58 | 0.54 | 0.53 |
| 52 | 0.57 | 0.58 | 0.56 |

Оптимальное значение Batch в диапазоне между 10 до 20. По умолчанию значение будет 12.

Параметр для LSTM - return\_sequences. результат работы LSTM либо вектор значений либо одно значение на фильтр.

Значение параметра return\_sequences = True

| Batch | 1 запуск | 2 запуск | 3 запуск |
| --- | --- | --- | --- |
| 10 | 0.60 | 0.6260 | 0.59 |
| 12 | 0.55 | 0.56 | 0.59 |
| 14 | 0.60 | 0.5826 | 0.59 |
| 16 | 0.58 | 0.5826 | 0.59 |

Явных признаков улучшения не выявлено, значение по умолчанию будет return\_sequences = False.

### Подготовка выходных данных/классификация ответов

Модель проекта представляет собой бинарную классификацию, где 1 является сигналом на восходящий тренд, а 0 является сигналом на нисходящий тренд. При этом правило классификации зависит от направления изменения цены, по умолчанию изменение текущего значения от предыдущего. В рамках оптимизации модели, необходимо произвести анализ изменения точности модели при изменении способов классификации.

### Развитие стратегии и модели

Возможные направления улучшения и развития стратегии “Корреляция в отрасли” и построенной модели:

* Оптимизировать программный код (Переписать формирование данных прогнозирования и рекомендаций в виде структуры класса, а не в виде готового текста. Все проверки моделей в один класс обернуть. Схожие функции по разным моделям перенести в функции. Добавить тестов и прочее)
* Увеличить кол-во компаний на основе которых производится обучение моделей.
* Произвести детальный подбор гиперпараметров.
* Выбрать оптимальный таймфрейм (сейчас используется дневной таймфрейм)
* Произвести ряд экспериментов с входными данными (изменить размер входного вектора для временных рядов и сверточной сети, обогатить данные).
* Произвести ряд экспериментов с более качественным определением целевой функции (таргетные данные).
* Настроить планировщик задач [Google Scheduler](https://cloud.google.com/scheduler) (из пакета Google Cloud Platform) для полностью автоматической загрузки данных и публикации рекомендаций на канале.

### Вывод по итогам моделирования стратегии “Корреляция в отрасли”

По итогам моделирования стратегии “Корреляция в отрасли” возможно сделать следующие выводы:

* Общие значения метрики accuracy колеблются от 0.60 до 0.70 для их повышения необходимо внести изменения в архитектуру, сделать подбор гиперпараметров и провести ряд экспериментов с входными данными и целевой функцией.
* Модель Conv1D даже при хороших значениях accuracy имеет нестабильный уровень воспроизводимости.
* Наиболее стабильной и воспроизводимой по результатам является модель LSTM.

## Стратегия “Прогнозирование конкретной цены на основе предыдущих данных”

Мы стремимся прогнозировать ежедневные скорректированные цены закрытия, используя данные за предыдущие N дней.Такие прогнозы позволят улучшить любую стратегию инвестирования, нивелируя потери от просадок.

При покупке финансовых инструментов, Инвестор “резервирует” свои средства на определённый срок. Пользуясь прогнозами во время этого периода, инвестор сможет повысить доходность за счёт своевременной покупки и продажи инструментов, ориентируясь при этом на основную стратегию

### Линейные модели

В ходе экспериментов доказывается очевидное суждение о высокой корреляции цен в соседнии дни. С учётом данных за предыдущие N дней, подбираются коэффициенты линейной регрессии. Главным минусом линейных моделей в предсказании цен по такому принципу - это неулавливание тренда, что очень сильно снижает доходность от данной стратегии. При наличие тренда модель предсказывает его же продолжение, замыкая таким образом круг и не давая прогнозов неожиданным скачкам, преобладающим в ценах на фондовой бирже

## Стратегия “Прогнозирование тренда на основе предыдущих данных”

Данный принцип позволит улучшить любую стратегию, повышая доходность во время срока работы последней путём продажи и покупки инструментов в зависимости от прогнозов.

### Использование разницы цены n(10) предыдущих часов

Для анализа был составлен датасет, в строках которого располагаются n разниц цен между i и i+1, i+1 и i+2 ... часами; предсказывается тренд в час после этих десяти

Для отбора моделей используются метрики accuracy (основная), precision, recall и f-measure. Для исследования были взяты модели из библиотеки scikit-learn. Наилучшие результаты показали следующие модели: **KNeighborsClassifier** (number = 51, distance\_metric = euclidean) – 0,593 accuracy; **MLPClassifier** (hidden\_layers\_sizes = (200,)) – 0,554 accuracy.

После простых моделей были использованы ансамбли. Наилучший результат показал **Voting Classifier.** Модели, входящие в состав: **KNeighborsClassifier, MLPClassifier, AdaBoostClassifier** (n\_estimators = 88)**, GradientBoostingClassifier** (n\_estimators=88, learning\_rate=1.0, max\_depth=1) – 0,633 accuracy

# Создание демонстрационного стенда

## Выбор способа демонстрации решения

Особенностью проекта является отсутствие необходимости поддержки интерактивности с пользователями. Результатом работы моделей является предсказание на открытие и закрытие сделки на покупку акций компаний. В связи с чем возможны следующие варианты реализации демонстрационного стенда:

| Варианты демостенда | Описание применения | Плюсы | Минусы |
| --- | --- | --- | --- |
| Рекомендательный веб-сайт | Веб сайт, где пользователь может посмотреть предлагаемые даты для покупки/продажи определенного финансового инструмента. Возможно размещать графическое представление о статистике работы каждой модели, сведения о доходностях, просадки, результаты прошедших рекомендаций. | * Расширяемый функционал. Возможность поэтапного добавления нового функционала | * Сложность реализации * Необходимость хостинга |
| **Телеграм-канал с автопостингом рекомендаций** | Автопостинг сообщений с рекомендациями о покупке/продаже акций по различным стратегиям | * **Простота реализации** * Возможность реализации на Google Colab | * Ограниченный функционал. Только публикация сообщений. |
| Телеграм-бот | Поддержка интерактивности с пользователями в части вывода дополнительной информации: статистика по стратегиям и моделям, сведения о доходностях, просадки, результаты прошедших рекомендаций | * Простота реализации * Возможность реализации на Google Colab | * Неудобство использования и ограниченный функционал. |
| Торговые сигналы | Формирование на основе рекомендаций модели торговых сигналов либо внешних торговых индикаторов | * Возможность проверить работу моделей на бою | * Необходимость размещения сигнала (специальные площадки) * Требование специальных знаний в области трейдерства |

В рамках проекта важнейшим критерием выбора способа публикации решения является затраченное время на реализацию. В связи с чем реализация демонстрационного стенда будет осуществляться через **телеграм-канал с автопостингом рекомендаций**, как самый малозатратный способ по времени.

## Описание демонстрационного стенда

Создание телеграм-канала осуществляется встроенными средствами Телеграмм. Для обеспечения автоматической публикации рекомендаций необходимо реализовать телеграм-бота, который будет осуществлять публикацию сообщений в телеграмм канале.

Для формирования рекомендаций необходимо реализовать Модуль предсказания на основе построенных моделей. При чем Модуль предсказания должен обеспечивать возможность масштабирования за счет добавления новых моделей предсказания. Т.е. при добавлении новой модели предсказания, модуль должен автоматически формировать предсказание на основе добавленной модели предсказания.

Общая архитектура решения:

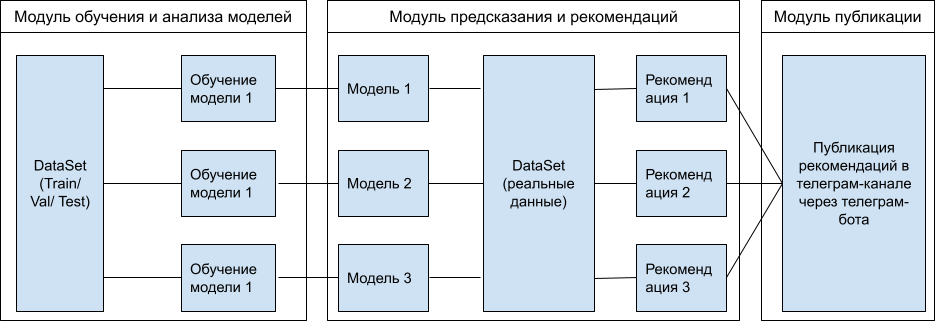
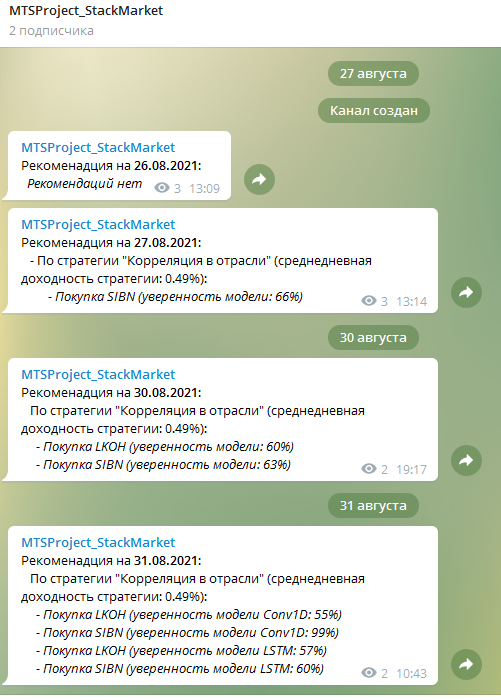


Схема работы решения:

1. Ежедневно формируются и выгружаются данные о котировках (используются дневные данные по компаниям из газонефтяной отрасли). Выгрузка осуществляется в проектную папку, в качестве хранилища данных выступает Google Drive.
2. Ежедневно (после загрузки данных) телеграм-бот запускает модуль предсказаний и рекомендаций, который используя загруженные данные строит прогнозы по всем доступным моделям (например по стратегии “Корреляция различных инструментов в общей отрасли” доступно 6 моделей. Для двух символов LKOH и SIBN, построенных по различным архитектурам нейронных сетей: Dense, Conv1D, LSTM).
3. Полученный прогноз проверяется на достижение необходимого порога для формирования прогноза (по умолчанию: либо два дня подряд должен быть сигнал на покупку либо два дня подряд сигнал на продажу).
4. Сформированные рекомендации телеграм-бот публикует на телеграм-канале: “<https://t.me/MTSProject_StackMarket>”.

Пример работы телеграмм-канала <https://t.me/MTSProject_StackMarket>:



,где:

* “среднедневная доходность стратегии” - является средним статическим значением, определенным на этапе обучения и тестирования стратегии.
* “Уверенность модели” - это значение прогноза модели, не является гарантированной величиной движения рынка.

В стратегии “Корреляция в отрасли” рекомендации производятся по двум символам LKOH и SIBN по трем различным моделям: Dense, LSTM, Conv1D.

Итоговое решение об использовании рекомендаций принимается пользователем самостоятельно.

# Описание структуры каталогов проекта:

* MTS.Project - общий каталог (<https://drive.google.com/drive/folders/1Q-nRvqwdaRBV1lJAk4aq3DiCZBJISdsV?usp=sharing>)
  + read.me MTS.Project - текущее описание проекта
  + MTS\_Project\_1\_3.ipynb - общий код проекта (<https://colab.research.google.com/drive/1wX5Cx0qnUF89fbjs24Qn8UDxdHO3mzbK>

)

* + DataSet/ - каталог с исходными данными
    - /GazNeft\_H1/ - каталог с данными крупных компаний акций по нефтегазовой отрасли с **часовыми** отсчетами
    - /GazNeft\_D1/ - каталог с данными крупных компаний акций по нефтегазовой отрасли с **дневными** отсчетами
    - Validation/ - каталог с результатами работы PandasProfile
  + Strategy/ - каталог рассматриваемых стратегий
    - Linear\_regression/ - каталог стратегии “Прогнозирование конкретной цены на основе предыдущих данных”
    - Correlations\_in\_Industry/ - каталог стратегии “Корреляция различных финансовых инструментов в общей отрасли”
    - Trend\_classification/ - каталог стратегии “Прогнозирование тренда на основе предыдущих данных”
  + Prod/ - каталог публикации и демонстрации результатов в телеграм-канале
    - models - используемые модели для прогнозирования и рекомендаций
    - Normalizer - каталог с правилами нормализации данных по различным финансовым инструментам
    - DataSet/ - каталог с исходными данными
      * /GazNeft\_H1/ - каталог с данными крупных компаний акций по нефтегазовой отрасли с **дневными** отсчетами
        + DD\_mm\_YYYY/ - Ежедневные каталоги с данными за текущие сутки в формате DD\_mm\_YYYY (например 26\_08\_2021)
    - PredictRecommendation.py - модуль рекомендаций по стратегии “Корреляция в отрасли”
    - MTSProject\_StackMarket\_bot.ipynb - телеграм-бот, публикующий данные в телеграм-канал (<https://colab.research.google.com/drive/1JZu-2kSus9VO46ZoF7MVvLQC0chdzlm5?usp=sharing>)
    - config.py - конфигурационные данные телеграм-канала и бота (в том числе токены доступа)
  + ConfCall/ - каталог с итогами встреч

# Термины и сокращения

| Наименование | Описание |
| --- | --- |
| **“Отсечка”, “Дата закрытия реестра акционеров” (Record Date)** | Дивидендная отсечка — это дата, в которую эмитент объявляет список акционеров, которым причитаются выплаты. Право на получение дивидендов за определённый период имеют не все акционеры, а только те, кто владеет ценными бумагами на дату закрытия реестра. Собственник, который прошел регистрацию позже этого срока, выплат не получит. Поэтому срок закрытия реестра акционеров, имеющих право на дивиденды, называют «датой отсечки» — Record Date. Она может совпадать с экс-дивидендной датой при модели биржевых торгов «Т + 0». |
| **Экс-дивидендная дата (Ex-Dividend Date)** | Акции можно приобрести по договору купли-продажи между физическими лицами или на рынке ценных бумаг — через брокера. Второй способ наиболее распространён. Его особенность в том, что купленные ценные бумаги зачисляются на счёт инвестора не сразу, а только через 2 рабочих дня. Модель торговли «Т + 2» применяется не только в России, но и на международных биржах. |
| **Таймфрейм** | Интервал времени ценового графика. Например дневной таймфрейм - это значение цены на конец дня, часовой - на конец часа и т.к. далее |
| **Символы** | Под символами подразумевается кодовое-символьное обозначение компании, акции по которой рассматриваются, например: LKOH - ПАО "Нефтяная компания "ЛУКОЙЛ", GAZP - ПАО "Газпром" |
|  |  |