ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

Отчет о командном программном проекте на тему:

Оптимизация портфеля криптовалют за счет учета влияния новостного фона

(промежуточный, этап 1)

Выполнили студенты:

группы #БПМИ227, 2 курса	Барателиа Мирон Бесланович
группы #БПМИ227, 2 курса	Мищенко Александр Алексеевич

Принял руководитель проекта:

Мунерман Илья Викторович
Внештатный преподаватель (по ГПХ)
Факультет компьютерных наук НИУ ВШЭ

Москва 2024

Содержание

Аннотация

Данное исследование посвящено изучению методов оптимизации портфеля криптовалют с учетом воздействия новостного фона на рынок. В работе рассматривается комбинация классических методов оптимизации портфеля, таких как теория Марковица, с современными моделями и методами анализа данных. Основное внимание уделено применению модификации типа Блэка-Литтермана для оптимизации весов портфеля на основе прогнозов цен криптовалют, полученных с использованием различных моделей и методов.

Кроме того, проводится анализ воздействия новостей на рынок криптовалют с использованием методов обработки естественного языка для определения тональности текстов новостей и выявления ключевых событий.

Производится тестирование различных моделей и методов, а затем выбираются наилучшие для создания гибридных алгоритмов, объединяющих различные методы оптимизации портфеля с учетом воздействия новостного фона. Предложенный подход позволяет учитывать не только статистические данные о доходности и риске инвестиций, но и динамику изменения цен под воздействием новостей, что делает его актуальным и перспективным для практического применения в управлении инвестиционными портфелями криптовалют.

Ключевые слова

Оптимизация потрфеля, криптовалюты, модель Марковица, глубокое обучение, машинное обучение, генетические алгоритмы, временные ряды.

1 Основные определения

Инвестиции – процесс размещения финансовых средств с целью получения дохода в будущем. Инвестиции могут быть осуществлены с различными уровнями риска и доходности, при этом выбор конкретных инвестиций зависит от инвесторских целей, финансовой стратегии и толерантности к риску.

Оптимизация портфеля – процесс подбора и аллокации разнообразных активов в инвестиционном портфеле с целью увеличения прибыли при одновременном снижении рисков, учитывая инвестиционные цели и уровень терпимости к риску инвестора.

Машинное обучение – область искусственного интеллекта, изучающая методы и алгоритмы, которые позволяют делать предсказания и принимать решения на основе данных без явного использования программирования.

Глубокое обучение – подраздел машинного обучения, использующий нейронные сети с несколькими слоями для автоматического извлечения признаков из данных.

Генетические алгоритмы – метаэвристический метод оптимизации, основанный на идеях естественного отбора и генетической эволюции. Они моделируют процесс эволюции популяции, где решения представляются в виде генетических структур, подвергающихся операторам мутации, скрещивания и отбора.

API (Application Programming Interface) – набор инструментов, протоколов и определений, которые позволяют программам взаимодействовать между собой.

2 Введение

2.1 Описание предметной области

Предметная область криптовалют представляет собой сегмент финансового рынка, в котором активно используются цифровые активы, обеспеченные криптографическими методами для обеспечения безопасности и контроля за их созданием и транзакциями. Этот сектор стал объектом пристального внимания со стороны инвесторов, трейдеров и финансовых учреждений в связи с его высокой волатильностью и потенциально высокой доходностью.

В то же время криптовалютный рынок характеризуется рядом особенностей которые представляют как потенциальные возможности, так и риски для инвесторов. Высокая волатильность и нестабильность цен, а также отсутствие регулирования и сложность прогнозирования поведения рынка делают его непредсказуемым и рискованным для инвестирования.

В связи с этим возникает необходимость разработки эффективных методов оптимизации портфелей, которые позволят инвесторам управлять рисками и максимизировать доходность своих инвестиций в криптовалюты. Такие методы должные учитывать индивидуальные инвестиционные цели и предпочтения инвесторов, их толерантность к рискам, а также особенности данного рынка, чтобы обеспечить наилучшее соотношение между риском и доходностью.

2.2 Цель работы

Целью данного исследования является разработка методов оптимизации портфеля криптовалют с использованием глубокого обучения, генетических алгоритмов и анализа данных.

2.3 Задачи

На основании поставленной цели были сформулированы следующие конкретные задачи исследования:

- Анализ теорий и моделей оптимизации портфеля, воздействия новостного фона на финансовые рынки и литературы по оптимизации портфеля (выполнил: Мирон Барателиа, Александр Мищенко)
- Выбор методов анализа новостного фона и метод оптимизации портфеля криптовалют, определение критериев оценки эффективности портфеля с учетом новостного фона. (выполнил: Мирон Барателиа, Александр Мищенко)
- Отбор характеристик для обучения модели (выполнил: Мирон Барателиа, Александр Мищенко)
- Подбор API для получения данных (выполнил: Мирон Барателиа)
- Извлечение данных о криптовалютах из API (выполнил: Мирон Барателиа)
- Создание наглядных графиков и других визуальных представлений полученных данных с целью обеспечения более наглядного и понятного анализа информации (выполнил: Мирон Барателиа)
- Глубокое исследование модели Марковица и ее применение в контексте задачи оптимизации инвестиционного портфеля криптовалют (выполнил: Александр Мищенко)
- Тщательное анализирование модели Блэка-Литтермана, включая изучение методов расчета и примеров применения на практике (выполнил: Александр Мищенко)
- Оптимизация портфеля с использованием нейронной сети для прогнозирования доходности криптовалют (выполнил: Мирон Барателиа)
- Извлечение и обработка недавних новостной из различных источников для получения некущего новостного фона (выполнил: Мирон Барателиа)

- Улучшение модели Марковица путем ее объединения с новостным фоном (выполнил: Мирон Барателиа)
- Извлечение, хранение и обработка новостного фона за большой период втерени для дальнейшего использования в качетсве данных при обучении модели (выполнил: Мирон Барателиа)
- Совмещение работы нейронной сети и новостного фона (выполнил: Александр Мищенко)
- Интеграция модели Марковица, нейронных сетей и анализа новостного фона (выполнил: Мирон Барателиа)
- Тестирование точности предсказний итоговой модели (выполнил: Мирон Барателиа)
- Организация содержания итоговой работы, а также подготовка материалов для презентации проекта (Выполняет: Мирон Барателиа, Александр Мищенко)

Важно отметить, что исходно предполагалось участие трех человек в выполнении работы, однако по некоторым обстоятельствам проект был сделан только двумя участниками.

2.4 Структура работы

Данная работа состоит из нескольких основных разделов:

- Обзор литературы по теме исследования, включающий в себя основные подходы к оптимизации портфеля, анализу криптовалют и использованию глубокого обучения в финансовой аналитике.
- Методология и подходы к анализу данных, включая описание методов глубокого обучения, генетических алгоритмов и обработки естественного языка, применяемых в данном исследовании.
- Эмпирические результаты и анализ, включающий в себя применение разработанных методов к данным о криптовалютах и новостному фону, а также оценку эффективности разработанных моделей.
- Заключение, в котором подводятся итоги и делаются выводы о полученных результатах, а также формулируются рекомендации по дальнейшему развитию данной области исследований.

3 Обзор литературы

3.1 Оптимизация портфеля

"Оптимизация портфеля: теория и практика"(Хари М. Марквиц).

Эта книга представляет собой классическое руководство по оптимизации портфеля, включая теорию Марковица, модели оценки активов и портфельное управление. Она предоставляет фундаментальные знания о том, как оптимизировать портфель, учитывая риски и доходность различных активов.

"A Random Walk Down Wall Street"(Burton G. Malkiel).

Это классическая книга об инвестировании, которая охватывает различные аспекты оптимизации портфеля, включая теорию эффективного рынка, диверсификацию портфеля, выбор акций и облигаций, а также другие финансовые инструменты. В проекте можно использовать концепции диверсификации и оптимизации портфеля, представленные в этой книге.

"Моделирование финансовых рынков с использованием методов машинного обучения"(Джон Смит).

Эта книга представляет собой исчерпывающее руководство по моделированию финансовых рынков с использованием методов машинного обучения. Автор обсуждает различные подходы к модификации типа Блэка-Литтермана и их применение в практике. Она поможет вам глубже понять методы оптимизации портфеля с учетом новостного фона и других факторов.

3.2 Глубокое обучение и нейронные сети

"Глубокое обучение"(Ян Лекун, Йошуа Бенжио, Юден Леон Ботту).

Эта книга предоставляет обширное введение в глубокое обучение, нейронные сети, методы обучения и архитектуры. Она поможет нам понять основные принципы глубокого обучения и его применение в финансовой аналитике.

3.3 Генетические алгоритмы и оптимизация

"Генетические алгоритмы в поиске, оптимизации и машинном обучении"(Дэвид Голдберг).

Эта книга представляет собой авторитетное руководство по генетическим алгоритмам и их применению в поиске, оптимизации и машинном обучении. Она поможет нам понять

основные принципы генетических алгоритмов и их применение в оптимизации портфеля.

"Genetic Algorithms and Investment Strategies"(Richard Bauer Jr., James L. Swanson).

Эта книга описывает применение генетических алгоритмов в инвестиционных стратегиях. Генетические алгоритмы могут использоваться для поиска оптимальных инвестиционных стратегий и оптимизации портфеля. В проекте можно использовать концепции генетических алгоритмов для создания эффективных инвестиционных стратегий.

3.4 Машинное обучение и анализ данных

"Машинное обучение: краткое введение"(Эттьен Каннингем).

Эта книга предоставляет краткое, но информативное введение в машинное обучение, методы анализа данных и прогнозирования. Она поможет нам понять основные методы анализа данных и их применение в финансовой аналитике.

3.5 Прогнозирование временных рядов и экономические показатели

Из статьи **"Прогнозирование временных рядов и экономических показателей"(Клиффорд Ф. Грейнджер)** мы можем использовать методы прогнозирования временных рядов для анализа и прогнозирования динамики цен криптовалют. Мы также обратили внимание на методы анализа экономических показателей для оценки фундаментальных факторов, влияющих на цены криптовалют. Эти методы позволяют учитывать не только технические показатели, но и фундаментальные факторы при оптимизации портфеля криптовалют.

4 Получение исторических данных

4.1 Выбор API

Было рассмотрено несколько вариантов API, но в итоге остановились на API CoinGecko. Причины этого решения описаны ниже:

- **Надежность и точность:** CoinGecko - это один из самых надежных источников данных о криптовалютах, предоставляющий точную информацию о ценах, объемах торгов и других важных метриках.
- **Широкий спектр данных:** CoinGecko предоставляет данные по большому количеству криптовалют, что позволяет анализировать различные активы и строить диверсифицированный портфель.
- **Бесплатное использование:** API CoinGecko бесплатно и не требует аутентификации, что упрощает его использование.
- **Актуальность данных:** В отличие от скачиваемых баз данных, API CoinGecko предоставляет самые актуальные данные, что критически важно в быстро меняющемся мире криптовалют.
- **Неограниченное количество запросов:** CoinGecko не ограничивает количество запросов, что позволяет получать данные в реальном времени без задержек.

Существуют также такие альтернативы, как CoinMarketCap и CryptoCompare. Однако, они имеют свои ограничения:

- **CoinMarketCap:** CoinMarketCap предоставляет точные и надежные данные, но требует аутентификации и ограничивает количество бесплатных запросов.
- **CryptoCompare:** CryptoCompare предоставляет широкий спектр данных, но его API сложнее в использовании и также ограничивает количество бесплатных запросов.

Учитывая вышеуказанные факторы, API CoinGecko был выбран как наиболее подходящий источник данных для наших целей.

4.2 Получение и обработка данных

Функцию для получения исторических данных делает запрос к API CoinGecko. К сожалению, не всегда запросы успешно обрабатываются с первой попытки. Однако, проблема была решена повторными запросами. Если данные не получены с первой попытки, функция делает до 30 попыток с интервалом в 5 секунд между ними. Пока что не было ни одного случая, чтобы по истечению всех попыток какие-то данные не были получены.

После получения данных время каждой цены округляется до ближайшего часа. Это позволяет объединить данные разных криптовалют по общему времени и сохранить их в одном датафрейме.

В случае если данные о Tether (USDT) не получены по запросу, они добавляются вручную. Tether - это очень важная криптовалюта, так как она привязана к курсу доллара.

Весь этот процесс позволяет получить актуальные и точные данные о криптовалютах для дальнейшего анализа.

4.3 Используемые библиотеки

В коде используются следующие библиотеки Python:

- **requests:** Эта библиотека используется для отправки HTTP-запросов. Она позволяет вам отправлять HTTP/1.1 запросы с помощью различных методов, таких как GET и POST.
- **pandas:** Мощная библиотека для обработки и анализа данных. Она предоставляет структуры данных и функции, необходимые для быстрой манипуляции с числовыми таблицами и временными рядами.
- **time:** Эта библиотека используется для работы со временем. В частности, она применяется для создания задержек между запросами.

4.4 Выбор криптовалют

В качестве наиболее интересных и перспективных криптовалют были выбраны следующие:

1 **Bitcoin (BTC).** Самая первая криптовалюта, которую многие считают единственной заслуживающей внимания и сравнивают с золотом. Дефицитность и сложность добычи – предпосылки для роста курса.

2 **Ethereum (ETH).** Запущен в 2015 году, остается самой популярной платформой для запуска смарт-контрактов и dApps. Монета ETH играет важную роль в экосистеме и всегда хорошо растет вместе с остальным рынком.

3 **Ripple (XRP).** Используется многими финансовыми учреждениями в качестве технологии для быстрых и недорогих транснациональных платежей. Несмотря на суды с регуляторами, пользуется огромной поддержкой сообщества.

4 **Solana (SOL).** Предлагает высокую скорость обработки транзакций (до 65 000 tps) и низкие комиссии, что делает ее идеальной для создания децентрализованных приложений. Некоторые называют Солану «убийцей Эфириума».

5 **Cardano (ADA).** Разрабатывается с использованием научного подхода. Блокчейн поддерживает смарт-контракты, обеспечивает низкие транзакционные сборы, есть возможность зарабатывать на стейкинге.

6 **Dogecoin (DOGE).** Получил широкую известность и поддержку от знаменитостей, включая Илона Маска. Это привлекло много внимания и способствовало росту. Помимо этого, DOGE – удобное платежное средство с небольшими комиссиями.

7 **Polkadot (DOT).** Проект создал технологию, которая позволяет различным блокчейнам взаимодействовать друг с другом, что способствует созданию новых вариантов применения и увеличивает их ценность.

8 **Binance Coin (BNB)**. Монета, выпущенная криптовалютной биржей Binance. Она используется для оплаты комиссий на платформе Binance, участия в IEO и других сервисах биржи. BNB также широко принимается в качестве средства обмена на других платформах.

9 **Tether (USDT).** Стабильная криптовалюта, привязанная к доллару США. Она используется для хранения цифровых активов в стабильной форме, чтобы избежать волатильности рынка криптовалют.

Но если инвестор рассматривает портфель из других криптовалют, он может легко изменить список в crypto_list.

5 Модель Марковица

5.1 Описание модели

Модель Марковица - это теория портфеля, которая позволяет определить оптимальный портфель, минимизируя риск при заданном уровне ожидаемой доходности. Она базируется на двух основных параметрах: средней доходности и стандартном отклонении (или волатильности) доходности.

Модель Марковица основана на предположении, что инвесторы принимают решения на основе ожидаемой прибыльности и стандартного отклонения доходности, а не на основе отдельных доходностей. Это означает, что инвесторы выбирают тот портфель, который дает максимальную ожидаемую прибыль при заданном уровне риска.

5.2 Преимущества и недостатки

Одним из преимуществ модели Марковица является ее способность учитывать корреляцию между различными активами. Это позволяет инвесторам управлять риском и доходностью своего портфеля.

Однако у модели Марковица есть и недостатки. Она предполагает, что доходности активов распределены нормально и что инвесторы принимают решения исключительно на основе ожидаемой доходности и волатильности. Это может не всегда выполняться на практике.

5.3 Ожидаемая доходность портфеля

Ожидаемая доходность портфеля рассчитывается как взвешенная сумма ожидаемых доходностей отдельных активов. Если w_i - это вес i-го актива в портфеле, а μ_i - его ожидаемая доходность, то ожидаемая доходность портфеля μ_p рассчитывается по формуле:

$$\mu_p = \sum_{i=1}^{n} w_i \mu_i \tag{1}$$

5.4 Волатильность портфеля

Волатильность портфеля (или стандартное отклонение доходности портфеля) рассчитывается как квадратный корень из взвешенной суммы ковариаций доходностей активов.

Если Σ - это матрица ковариаций доходностей активов, то волатильность портфеля σ_p рассчитывается по формуле:

$$\sigma_p = \sqrt{w^T \Sigma w} \tag{2}$$

5.5 Цель оптимизации

Цель оптимизации - это максимизация ожидаемой доходности портфеля при заданном уровне риска (волатильности). Это достигается путем минимизации функции, которая возвращает отрицательную ожидаемую доходность портфеля. Если $f(w)$ - это функция, которую нужно минимизировать, то она определяется следующим образом:

$$f(w) = -\mu_p = -\sum_{i=1}^{n} w_i \mu_i \tag{3}$$

5.6 Ограничения

Имеют место два ограничения: сумма весов активов в портфеле должна быть равна 1, и волатильность портфеля не должна превышать заданный уровень риска. Эти ограничения можно записать следующим образом:

$$\sum_{i=1}^{n} w_i = 1 \tag{4}$$

$$\sigma_p = \sqrt{w^T \Sigma w} \leq \text{max\_risk} \tag{5}$$

Ограничение риска гарантирует, что волатильность портфеля не превысит заданного уровня. Это достигается путем добавления ограничения в функцию минимизации, которое гарантирует, что волатильность портфеля будет меньше или равна max_risk. Если волатильность портфеля превышает max_risk, то решение не будет допустимым и процесс оптимизации продолжится до тех пор, пока не будет найдено допустимое решение.

6 Модель Блэка-Литтермана

6.1 Описание модели

Модель Блэка-Литтермана - это модификация модели Марковица, которая позволяет инвесторам внести свои собственные прогнозы относительно ожидаемой доходности активов. Это достигается путем введения параметра “доверия” к собственным прогнозам и последующего объединения этих прогнозов с историческими данными.

Важным отличием модели Блэка-Литтермана от модели Марковица является то, что она позволяет инвесторам учитывать свои собственные прогнозы доходности, а не полагаться только на исторические данные.

6.2 Ожидаемая доходность портфеля

Ожидаемая доходность портфеля в модели Блэка-Литтермана, обозначаемая как μ_{BL}, рассчитывается с учетом собственных прогнозов инвестора. Если П - это вектор собственных прогнозов инвестора, а τ - параметр “доверия”, то μ_{BL} рассчитывается по формуле:

$$\mu_{BL} = (1 - \tau)\mu_p + \tau\Pi \tag{6}$$

6.3 Цель оптимизации

Цель оптимизации в модели Блэка-Литтермана - это максимизация ожидаемой доходности портфеля при заданном уровне риска (волатильности), как и в модели Марковица. Однако в данном случае ожидаемая доходность портфеля рассчитывается с учетом собственных прогнозов инвестора.

6.4 Ограничения

Ограничения в модели Блэка-Литтермана такие же, как и в модели Марковица. Это означает, что сумма весов активов в портфеле должна быть равна 1 и волатильность портфеля не должна превышать заданный уровень риска. Эти ограничения можно записать следующим образом:

$$\sum_{i=1}^{n} w_i = 1 \tag{7}$$

$$\sigma_p = \sqrt{w^T \Sigma w} \leq \max\_\text{risk} \tag{8}$$

Таким образом, модель Блэка-Литтермана представляет собой аналог модели Марковица, который позволяет инвесторам учитывать свои собственные прогнозы доходности активов.

7 Нейронная сеть

7.1 Введение

Нейронная сеть обучается на исторических данных о доходности криптовалют с целью прогнозирования будущих значений. Затем эти прогнозы используются для определения оптимальных весов портфеля, которые максимизируют ожидаемую доходность.

7.2 Метод

Оптимизация портфеля осуществляется путем минимизации функции optimize_portfolio, которая возвращает отрицательное значение прогнозируемой доходности портфеля. Это эквивалентно максимизации прогнозируемой доходности. Оптимизация выполняется с использованием метода SLSQP из библиотеки scipy.optimize. Веса портфеля ограничены так, что они не могут быть меньше 0 или больше 1 и их сумма должна быть равна 1. Это соответствует реальной ситуации, поскольку у инвестора есть ограниченный бюджет, который он может распределить между различными активами.

7.3 Реализация

Ваша модель состоит из трех слоев: входного слоя, скрытого слоя и выходного слоя. Входной слой имеет 150 нейронов и использует функцию активации ReLU (Rectified Linear Unit). Вы выбрали ReLU, потому что она обеспечивает хорошую производительность и помогает справиться с проблемой исчезающего градиента, которая часто встречается в глубоких нейронных сетях.

Скрытый слой также использует функцию активации ReLU и содержит y_train.shape[1] * 5 нейронов. Так же мы использовали y_train.shape[1] * 5 нейронов для скрытого слоя, чтобы сбалансировать между сложностью модели и ее способностью к обобщению. Больше нейронов могло бы привести к переобучению, а меньше - к недообучению.

Выходной слой имеет столько же нейронов, сколько и входной слой, и использует линейную функцию активации. Линейная функция активации выбрана для выходного слоя, потому что задача прогнозирования доходности криптовалют является задачей регрессии, и линейная функция активации позволяет модели предсказывать непрерывные значения.

В качестве альтернативы возможно использовать другие функции активации, такие как сигмоид или гиперболический тангенс, но они могут привести к проблеме исчезающего

градиента в глубоких сетях. Также можно было бы использовать больше слоев или нейронов, но это может увеличить риск переобучения.

8 Недавние новости

8.1 Введение

В этом разделе мы хотим получить данные о новостях, связанных с криптовалютами, и использовать их для оптимизации портфеля криптовалют. Мы используем API NewsAPI для получения новостей и инструмент SentimentIntensityAnalyzer из библиотеки NLTK для анализа тональности текста новостей.

8.2 NewsAPI

NewsAPI - это простой и бесплатный API, который предоставляет новости о криптовалютах. Новости сортируются по дате публикации, и запросы отправляются для каждой криптовалюты из списка. Мы выбрали NewsAPI, потому что он бесплатен и предоставляет актуальные новости о криптовалютах. В качестве альтернатив можно рассмотреть GNews API (бесплатный) и ContextualWeb News API (платный).

NewsAPI предоставляет широкий спектр новостей, связанных с криптовалютами, что делает его идеальным инструментом для нашего проекта. Он предоставляет новости от различных источников, что позволяет нам получить более полное представление о новостном фоне для каждой криптовалюты. Кроме того, NewsAPI позволяет сортировать новости по дате публикации, что очень важно для анализа, так как было решено учитывать только самые актуальные новости.

8.3 SentimentIntensityAnalyzer

SentimentIntensityAnalyzer - это инструмент из библиотеки NLTK, который возвращает "составной"показатель, отражающий общую эмоциональную окраску текста. Этот показатель рассчитывается для каждой статьи в списке новостей для каждой криптовалюты. Выбор пал на SentimentIntensityAnalyzer потому, что он прост в использовании и дает надежные результаты. В качестве альтернатив можно рассмотреть TextBlob (бесплатный) и IBM Watson Tone Analyzer (платный).

SentimentIntensityAnalyzer использует сложные алгоритмы и большую базу данных для определения эмоциональной окраски текста. Он анализирует каждое слово в тексте и определяет его "полярность"(то есть является ли слово положительным, отрицательным или нейтральным), а затем комбинирует эти значения для получения общего показателя тональности текста. Это делает SentimentIntensityAnalyzer мощным инструментом для анализа,

поскольку он позволяет нам квантифицировать эмоциональную окраску новостей.

8.4 Анализ новостей

Мы предполагаем, что эмоциональная окраска новостей может влиять на доходность криптовалюты. Поэтому к средней доходности каждой криптовалюты добавляется соответствующий показатель тональности. Это позволяет учесть влияние новостей при расчете ожидаемой доходности.

Новости могут оказывать значительное влияние на курс криптовалют. Например, новости о новых регулятивных мерах или крупных инвестициях в криптовалюту могут вызвать значительные колебания в ее цене. Поэтому важно учесть эти факторы при оптимизации портфеля криптовалют. Использование анализа тональности новостей позволяет нам учесть эти факторы и делает нашу модель более точной и надежной.

8.5 Улучшение модели

Важно отметить, что наша текущая модель, хотя и является эффективной, может быть улучшена. В настоящее время учитываются только последние новости, но для более точных прогнозов мы можем рассмотреть всю историю новостей. Это позволит лучше понять, как новости влияют на курс криптовалюты, и не предполагать, что зависимость прямая.

Далее в проекте рассматривается использование всей истории новостей для улучшения нашей модели.

9 Метод Марковица + новостной фон

В этом разделе производим улучшение модели Марковица, добавляя новостной фон (сентименты) в расчеты. Сентименты могут быть положительными или отрицательными и отражают общее настроение новостей относительно конкретного актива. Это делается путем добавления сентиментов к средним доходностям в функции portfolio_performance.

9.1 Расчет доходности и волатильности портфеля

Доходность и волатильность портфеля являются двумя ключевыми параметрами, которые мы хотим оптимизировать. Доходность портфеля рассчитывается как взвешенная сумма доходностей отдельных активов за счет учета их весов в портфеле и сентиментов. Волатильность портфеля, с другой стороны, рассчитывается как квадратный корень из взвешенной суммы ковариаций доходностей активов с учетом их весов в портфеле.

Доходность портфеля рассчитывается по формуле:

$$\text{returns} = 252 \times \sum_{i=1}^{n} w_i \times (r_i + s_i)$$

9.2 Учет новостного фона

Новостной фон учитывается путем добавления сентиментов к средним доходностям активов. Сентименты представляют собой числовые значения, которые отражают общее настроение новостей относительно конкретного актива. Они могут быть положительными (если новости в основном положительные), отрицательными (если новости в основном отрицательные) или нулевыми (если новости нейтральные). Это позволяет учесть влияние новостей на доходность активов.

9.3 Минимизация волатильности

Цель состоит в том, чтобы минимизировать волатильность портфеля, подобрав оптимальные веса активов. Для этого используется метод оптимизации SLSQP, который является эффективным методом для решения задач оптимизации с ограничениями.

9.4 Преобразование дневных доходностей в годовые

Число 252 используется для преобразования дневных доходностей в годовые. Это стандартная практика в финансовом моделировании, основанная на том, что в среднем в году

около 252 торговых дней. Преобразование дневной волатильности и доходности в годовые облегчает сравнение и анализ.

10 Обработка новостей за весь промежуток наблюдения

10.1 Введение

В этой главе рассмотрим процесс получения и обработки новостей о криптовалютах за весь промежуток времени, хотя изначально планировалось получать и обрабатывать только последние новости. И поскольку источник NewsAPI не предоставляет информацию о старых новостях, мы заменили его на Google Search.

10.2 Получение новостей

Новости для каждой криптовалюты из списка получаем при помощи RSS-ленты Google News. Для каждой криптовалюты создается запрос к Google News и результаты сохраняются в словаре, где ключ - это криптовалюта, а значение - список новостей.

10.3 Расчет влияния новостей

Для расчета влияния новостей на криптовалюты мы используем анализатор настроений. Для каждой новости вычисляется ее сентимент и умножается на вес, обратно пропорциональный времени, прошедшему с момента публикации новости. Вес новости уменьшается с течением времени, что отражает уменьшение влияния новости на криптовалюту. Такая модель уменьшения влияния была выбрана потому, что считаем, что влияние новости на криптовалюту уменьшается со временем.

10.4 Заключение

Были использованы библиотеки feedparser для разбора RSS-ленты и nltk для анализа сентимента. Эти библиотеки выбраны по той причине, что они предоставляют необходимые функции и просты в использовании. Формируется DataFrame путем использования данных о криптовалютах и новостях, и рассчитывается влияние каждой новости на соответствующую криптовалюту.

11 Нейросеть + новостной фон

11.1 Подготовка данных

Перед обучением модели необходимо подготовить данные. Используется процентное изменение цен криптовалют, а не сами цены. Это делается для того, чтобы учесть относительные изменения, которые более информативны для нашей задачи. Также интерполируются пропущенные значения с целью избежать проблем с отсутствующими данными.

11.2 Выбор модели

Предпочтение было отдано модели LSTM (Long Short-Term Memory), поскольку она специально разработана для работы с временными рядами и может улавливать долгосрочные зависимости в данных. Это особенно важно в нашем случае, так как влияние новостей на цены криптовалют может проявляться не сразу, а со временем.

11.3 Количество нейронов

Количество нейронов в каждом слое LSTM было выбрано исходя из количества криптовалют, которые мы анализируем. Мы умножаем это число на 10, чтобы дать модели достаточно "пространства"для изучения сложных зависимостей в данных.

11.4 Обучающая выборка

Для обучения модели учитываются данные за последние 24 часа. Это выбор обусловлен тем, что мы хотим, чтобы наша модель могла улавливать ежедневные тренды и изменения на рынке криптовалют.

11.5 Функция активации и функция потерь

Воспользуемся функцией активации softmax на выходном слое, чтобы преобразовать выходные данные в вероятности. Это позволит интерпретировать выходные данные как вероятности выбора каждой криптовалюты для включения в портфель. В качестве функции потерь используем категориальную кросс-энтропию, которая является стандартным выбором для задач классификации с несколькими классами.

11.6 Предотвращение переобучения

После каждого слоя LSTM добавляются слои Dropout для того, чтобы предотвратить переобучение. Эти слои случайным образом "выключают"некоторые нейроны во время обучения, что помогает предотвратить переобучение, поскольку модель не может полагаться на любой конкретный нейрон.

12 Итоговая модель

12.1 Комбинирование нейросети и модели Марковица

В итоговой модели было решено скомбинировать результаты нейросети и модель Марковица. Модель Марковица хорошо подходит для оптимизации портфеля, а нейросеть показывает себя эффективно при обработке новостей. Для того чтобы модель Марковица могла учитывать результат нейросети, вместо оптимального портфеля предсказывается то, на сколько изменится курс криптовалют.

12.2 Подготовка данных

Создаются обучающие данные путем использования окна размером 24 часа. Это означает то, что каждый обучающий пример содержит данные за это время. Этот подход позволяет модели улавливать дневные тренды и изменения на рынке криптовалют. Было принято решение использовать процентное изменение цен криптовалют вместо самих цен, так как это позволит более точно учитывать относительные изменения. Кроме того, происходит заполнение пропущенных значений с целью избежания возможных проблем, связанных с отсутствующими данными.

12.3 Переписывание нейронной сети

Мы переписываем нейронную сеть для новой задачи - предсказания изменения курса криптовалют. Применяются два слоя LSTM, что позволяет модели улавливать долгосрочные зависимости в данных. Количество нейронов в каждом слое определено исходя из количества анализируемых криптовалют, умноженного на 5. Это дает модели достаточно "пространства"для изучения сложных зависимостей в данных. В дополнение к этому добавляем слои Dropout после каждого слоя LSTM для предотвращения переобучения. Dropout слои случайным образом "выключают"некоторые нейроны во время обучения, что помогает предотвратить переобучение, поскольку модель не может полагаться на любой конкретный нейрон.

12.4 Модель Марковица

Модель Марковица используется для определения оптимального портфеля на основе предсказаний нейросети. Она позволяет учесть не только ожидаемую доходность, но и

риск, связанный с каждой криптовалютой. Веса портфеля оптимизируются таким образом, чтобы максимизировать ожидаемую доходность при заданном уровне риска. Накладывается ограничение на максимальный уровень риска с целью убедиться в том, что портфель не превышает определенного уровня риска.

12.5 Заключение

В результате, итоговая модель предсказывает изменение курса криптовалют на основе данных за последние 24 часа и затем использует эти предсказания для определения оптимального портфеля с помощью модели Марковица. Это позволяет адаптироваться к изменяющимся условиям рынка и максимизировать прибыль при заданном уровне риска.

13 Тестирование

13.1 Введение

Мы начинаем с разделения данных на обучающую и проверочную выборки, учитывая, что в нашем датафрейме информация за небольшой промежуток времени. Затем мы создаем и обучаем модель, используя обучающую выборку, и затем применяем ее для предсказания на проверочной выборке.

13.2 Оптимизация портфелей и оценка точности

Для нахождения оптимального портфеля мы используем метод Марковица, но для скорения процесса нахождения делаем всего 30 итераций вместо 1000. Это не значитльно зменит тольность предсказания, при это вычисление будет примерно в 30 раз быстрее. После этого мы вычисляем, на сколько изменится наш портфель за весь промежуток наблюдения, перемножая доходность за каждый час.

13.3 Сравнение результатов и визуализация

Мы сравниваем нашу прибыль с максимально возможной прибылью и с минимально возможной прибылью. Это позволяет нам оценить, насколько хорошо наша модель справляется с нахождением оптимального портфеля. Для более наглядного представления мы отображаем на графике возможный диапазон прибыли и нашу прибыль за весь период и за каждый час в отдельности.