Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Основная образовательная программа
«Прикладная математика и информатика»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Исследовательский проект на тему

Алгоритмический трейдинг: высокочастотные, низкочастотные и глобал макро стратегии

Выполнил студент группы 191, 4 курса,
Мельников Артём Максимович

Руководитель ВКР:
Директор института количественных финансов НИУ ВШЭ,
Злотник Андрей Александрович

Москва 2023

Содержание

1 Аннотация

1.1 Аннотация

В данном исследовании рассматривается применение машинного обучения в высокочастотной торговле (HFT), в частности, прогнозирование цен для валютных пар ETH/USD и USD/ETH. Цель исследования - оценить эффективность различных предикторов в рамках различных моделей для улучшения стратегий HFT. Исследование включает в себя построение и сравнение моделей машинного обучения для прогнозирования рыночных тенденций и движения цен, начиная с реализации существующих стратегий, чтобы установить эталоны и понять текущую практику. Под руководством опытных количественных трейдеров эти модели итеративно дорабатываются на основе исторических данных для повышения эффективности. Эффективность моделей тщательно оценивается с помощью бэктестинга и симуляции, что позволяет разработать сложные модели прогнозирования для HFT и получить представление о сильных и слабых сторонах различных предикторов в быстро меняющейся торговой среде.

1.2 Abstract

This research investigates the application of machine learning in High-Frequency Trading (HFT), specifically targeting price prediction for the ETH/USD and USD/ETH currency pairs. It aims to evaluate the effectiveness of different predictors within various models for enhancing HFT strategies. The study includes building and comparing machine learning models to predict market trends and price movements, starting with the implementation of existing strategies to establish benchmarks and understand current practices. Under the guidance of experienced

quantitative traders, these models are iteratively refined using historical data for better performance. The effectiveness of the models is rigorously assessed through backtesting and simulation, aiming to develop sophisticated prediction models for HFT and gain insights into the strengths and weaknesses of different predictors in fast-paced trading environments.

1.3 Ключевые слова

Algorithmic trading, High frequency trading (HFT), Machine learning, Limit order book, Market Making, Cryptocurrency, Ethereum

Алгоритмическая торговля, Высокочастотная торговля (HFT), Машинное обучение, Ордербук, Маркет-мейкинг, криптовалюта

2 Введение

В постоянно меняющемся ландшафте финансовых рынков роль технологий и алгоритмических стратегий становится все более заметной. Высокочастотная торговля (HFT), характеризующаяся быстрым исполнением сделок и передовыми алгоритмическими стратегиями, стоит в авангарде этой технологической революции. В частности, в сфере торговли криптовалютами такие пары, как ETH/USD и USD/ETH, представляют собой уникальные проблемы и возможности из-за присущей им волатильности и динамики рынка.

Несмотря на значительные достижения в области HFT, существующие исследования часто не позволяют учесть все нюансы криптовалютных рынков. Традиционные модели, эффективные для обычных классов активов, с трудом адаптируются к хаотичным движениям цен и быстрым изменениям, характерным для цифровых валют. Этот пробел подчеркивает необходимость инновационных подходов, специально разработанных для криптовалютного трейдинга с высокими ставками.

Данное исследование направлено на устранение этого пробела путем разработки и сравнения спектра моделей машинного обучения для прогнозирования цен в HFT, с особым акцентом на пары ETH/USD и USD/ETH. Наш подход охватывает широкий спектр моделей, начиная от линейных регрессий и заканчивая сложными LSTM-сетями и методами градиентного усиления, используя при этом обширный набор признаков. К ним относятся дисбаланс orderbook, доходность за прошлые периоды, а также набор технических индикаторов, тщательно отобранных с учетом их актуальности для криптовалютных рынков.

Значение этого исследования заключается не только в повышении точности прогнозирования стратегий HFT, но и в более широком понимании применения машинного обучения на волатильных рынках. Новым аспектом нашей работы является разработка собственной метрики оценки. Эта метрика, предназначенная для балансировки различных результатов прогнозирования, тесно связана с конкретными целями управления рисками и доходностью в HFT.

Кроме того, наша методология удовлетворяет критической потребности в адаптации к реальному времени. В быстро меняющейся сфере криптовалютной торговли способность динамически настраивать модели с учетом новых данных имеет первостепенное значение. Данное исследование изучает этот аспект, предлагая идеи по управлению чрезмерной подгонкой и обеспечению вычислительной эффективности в реальных сценариях.

Таким образом, данная работа не только вносит вклад в область машинного обучения на финансовых рынках, но и предлагает новый взгляд на применение этих методов в уникальном и сложном мире криптовалютной торговли.

3 Обзор литературы

3.1 Высокочастотная торговля (HFT)

3.1.1 Optimal high-frequency trading with limit and market orders[1]

В данном исследовании изучается оптимальная стратегия для мелкого трейдера для получения прибыли на рынке, где цены меняются случайным образом и быстро. Трейдер должен найти компромисс между получением

лучшей цены с помощью лимитных заявок и быстрее исполнением рыночных заявок. Предполагается, что рынок следует за цепью Маркова, где цены зависят только от их текущего состояния и не зависят от их прошлой истории. Поступление рыночных заявок моделируется как процесс Пуассона, и трейдер может размещать лимитные ордера на разных уровнях книги лимитных заявок (LOB). В исследовании используется подход динамического программирования, чтобы найти оптимальную стратегию для различных функций полезности и параметров рынка. Анализируются два типа функций полезности: экспоненциальная и энергетическая. Оптимальная стратегия зависит от соотношения двух параметров: lambda и mu. В исследовании сравниваются свои результаты с другими существующими моделями высокочастотной торговли и показывает, что она способна уловить некоторые особенности, которые не присутствуют в других моделях. В работе также обсуждаются некоторые расширения и ограничения модели, такие как включение риска товарных запасов, влияние на рынок и асимметричная информация.

3.1.2 Price Dynamics in a Markovian Limit Order Market[2]

Авторами представлена математическая модель для изучения изменения цен на рынке с лимитными и рыночными ордерами. Модель предполагает, что поступление ордеров происходит по пуассоновскому процессу и что состояние книги лимитных ордеров может быть представлено цепью Маркова. В статье определены четыре типа событий, которые могут изменить состояние книги лимитных заявок, а именно: рыночные заявки на покупку, рыночные заявки на продажу, лимитные заявки на покупку и лимитные заявки на продажу.

В статье выделяются два сценария размещения лимитных ордеров, в

лучшем случае и на расстоянии одного тика, которые по-разному влияют на поведение цен и очередей. В статье выводятся формулы для различных величин, связанных с ценами и очередями, с использованием методов теории очередей и теории вероятностей. Некоторые из этих величин включают распределение продолжительности между изменениями цен, распределение и автокорреляцию изменений цен, а также вероятность восходящего движения цен.

Статья сравнивает свою модель с некоторыми реальными данными с различных рынков и показывает, что она может отражать некоторые важные особенности динамики цен. Однако в статье также обсуждаются некоторые ограничения модели, такие как предположение о стационарности рынка и пренебрежение эффектом влияния на цены. В конце статьи обсуждаются некоторые возможные расширения и применения модели.

В целом, статья предоставляет полезную основу для изучения динамики цен на рынке лимитных заказов и проливает свет на сложные взаимодействия между различными типами ордеров и их влияние на цены.

3.1.3 High-frequency trading in a limit order book[3]

В статье предлагается математическая модель, помогающая дилерам котировать цены и управлять запасами на рынке ценных бумаг. В работе используется стохастическая система управления для выведения оптимальных котировок спроса и предложения для дилера, торгующего одним активом в книге лимитных заявок.

Модель предполагает, что средняя цена актива следует геометрическому броуновскому движению с постоянной волатильностью, и моделирует процесс поступления заявок как процесс Пуассона с различной интенсивностью

для заявок на покупку и продажу. Прибыль дилера определяется как разница между стоимостью его запасов и остатком наличности минус штраф за риск.

В работе показано, что оптимальные котировки спроса и предложения являются линейными функциями запасов дилера с наклонами, зависящими от гаммы, волатильности, скорости поступления заявок и коэффициентов влияния на рынок. Гамма вводится как фактор риска, который измеряет чувствительность прибыли дилера к колебаниям цен. В статье обсуждается, как гамма влияет на компромисс между риском запасов и риском сделки и как ее можно откалибровать с помощью исторических данных или моделей микроструктуры рынка.

Приводятся численные примеры для различных значений гаммы и рыночных параметров и сравниваются с эмпирическими данными по акциям Нью-Йоркской фондовой биржи. Авторы также проводят моделирование методом Монте-Карло для проверки устойчивости стратегии при различных сценариях скачков цен, отмены заказов и ограничения запасов.

В целом, в статье представлена математическая модель, которая помогает дилерам котировать цены и управлять запасами на рынке ценных бумаг. Модель учитывает волатильность рынка, скорость поступления заказов и другие параметры и вводит гамму в качестве фактора риска для измерения чувствительности прибыли дилера к колебаниям цен. В статье приводятся численные примеры и симуляции, иллюстрирующие эффективность предложенной стратегии.

3.2 Regression модели

3.2.1 Forecasting crude oil prices with a large set of predictors: Can LASSO select powerful predictors?[4]

Авторы статьи решают проблему прогнозирования цен на нефть на основе исторической цены и нескольких выбранных параметров. Стоимость Texas Intermediate часто используется в качестве прокси для цен на нефть. В качестве параметров модели используются многие экономические показатели (ключевая ставка, общее количество денег в экономике, инфляция и т.д.), несколько ключевых значений на нефтяном рынке (объем импорта/-экспорта), а также ряд параметров, полученных с помощью технического анализа (они показывают наилучшие результаты).

Затем они обучают Lasso и Elastic net (модификация Lasso). Затем они сравнивают свои обученные модели с несколькими другими статистическими методами (Ridge, PCрегрессии, комбинированные методы предсказания моделей и динамическое взвешивание моделей). Если смотреть на выбранные метрики (в основном, вневыборочный R^2), то модели, обученные авторами, работают лучше сравниваемых методов. Подводя итоги своей работы, авторы рассматривают обученные модели и показывают, как модели Lasso и Elastic net выбирают наиболее надежные параметры из представленных.

3.2.2 Forecasting China's stock market volatility with shrinkage method: Can Adaptive Lasso select stronger predictors from numerous predictors?[5]

В данной статье используется метод Adaptive Lasso для прогнозирования волатильности китайского рынка ценных бумаг. Для измерения волатиль-

ности авторы использовали реализованную волатильность. Это было сделано с использованием ряда параметров, описывающих состояние фондовой биржи и экономики Китая. Модель, обученная исследователями, показала более точные прогнозы, чем другие модели. Сравнение проводилось на вневыборочных данных. Согласно эмпирическим наблюдениям, Adaptive Lasso работает даже лучше в периоды низкой волатильности. Далее авторы проверили устойчивость своей модели, используя несколько методов, признанных в данной области: различные горизонты прогнозирования, разделение волатильности на высокую и низкую, различные способы измерения волатильности (кроме RV) и т.д.

3.2.3 Applying Lasso Linear Regression Model in Forecasting Ho Chi Minh City's Public Investment[6]

Авторы данной статьи пытаются спрогнозировать расходы государственного бюджета в зависимости от прагматических параметров. Государство играет важную роль в экономике, и возможность прогнозировать бюджет позволяет более активно подходить к вопросам планирования. Для этого исследователи используют ряд популярных методов машинного анализа: OLS, Ridge и Lasso регрессионные модели. В качестве параметров они используют метрики, показывающие общее состояние экономики: цены на нефть, биржевые индексы, ключевую ставку ЦБ и т.д. Затем выбирается лучший метод на основе метрики размера ошибки: среднеквадратичной ошибки (RMSE) или средней абсолютной процентной ошибки (MAPE). В частности, Лассо показал наилучший результат при прогнозировании бюджета государственных расходов города Ho Chi Minh.

4 Методология

4.1 Метрика

4.1.1 Интуиция

Предлагаемая метрика направлена на оценку эффективности модели высокочастотного трейдинга (HFT) путем учета нюансов влияния ее прогнозов на торговые действия. Основная идея, лежащая в основе этой метрики - ее способность дифференцированно оценивать правильные и неправильные прогнозы, предлагая индивидуальную оценку которая соответствует особой важности различных результатов в сфере высокочастотной торговли. частотного трейдинга. Такая оптимизация позволяет привести метрику в соответствие с общей целью повышения прибыльности модели HFT.

4.1.2 Формула

Формально метрика определяется как

$$\text{metric}(y_{\text{true}}, y_{\text{pred}}) = \frac{\sum (\text{confusion matrix} \odot \text{weight matrix})}{\sqrt{\text{total action count}}}$$

Где:

- $\odot$ обозначает умножение по элементам.
- Матрица несоответствий для $(y_{\text{true}}, y_{\text{pred}})$ отражает положительные, отрицательные, ложно положительные и ложно отрицательные предсказания модели. (TP, TN, FP, FN)
- Матрица весов присваивает определенные веса различным результатам предсказания.

- Матрица попаданий получается путем поэлементного умножения матрицы несоответствий и весов.

- Метрический результат рассчитывается путем нормализации суммы матрицы попаданий на квадратный корень из общего количества действий, обеспечивая сбалансированную оценку, которая учитывает влияние предсказаний модели.

4.2 Обзор данных

4.2.1 Сбор данных и методология

Мы сознательно выбрали торговую пару USD/ETH, поскольку Ethereum занимает видное место в криптовалютном ландшафте, что делает его ключевым ориентиром для анализа. Выбор авторитетного поставщика криптовалютных данных облегчил доступ к историческим данным с особым акцентом на динамику высокочастотной торговли парой USD/ETH с мая по август 2022 года.

Выбор источника данных и доступ к нему Мы выбрали авторитетного поставщика криптовалютных данных, ориентируясь на торговую пару USD/ETH. Этот выбор обусловлен значительным влиянием Ethereum на рынок и его широким распространением, что делает его оптимальным эталоном для анализа высокочастотной торговли. Доступ к историческим данным был упрощен благодаря документации по API провайдера, а для аутентификации были получены необходимые API-ключи.

Временное и гранулярное определение Чтобы уловить нюансы динамики рынка, мы точно определили временной диапазон (с мая по август

2022 года) и гранулярность, сосредоточившись на ценах OHLC и объемах торгов. Такой подход учитывает высокочастотную природу криптовалютной торговли, обеспечивая детальную основу для последующего анализа.

Запросы к API, пагинация и хранение Для получения исторических данных выполнялись систематические запросы к API, при этом учитывались потенциальные ограничения и применялась пагинация для обеспечения полного охвата. Собранный набор данных, считающийся важным для представления пары USD/ETH в указанный период, был эффективно сохранен в формате CSV. Этот формат не только обеспечивает доступность, но и облегчает хранение в локальных базах данных и электронных таблицах, поддерживая различные аналитические подходы.

Обеспечение качества Признавая важность целостности данных, был применен надежный процесс проверки качества для устранения пробелов, несоответствий и аномалий в наборе данных. Полученный набор данных, отличающийся надежностью, был тщательно задокументирован, что способствовало повышению прозрачности и заложило прочную основу для воспроизводимости.

Эта оптимизированная методология с особым акцентом на торговую пару USD/ETH обеспечивает целостность собранных данных и позволяет нашему исследованию внести значимый вклад в более широкий дискурс криптовалютных исследований.

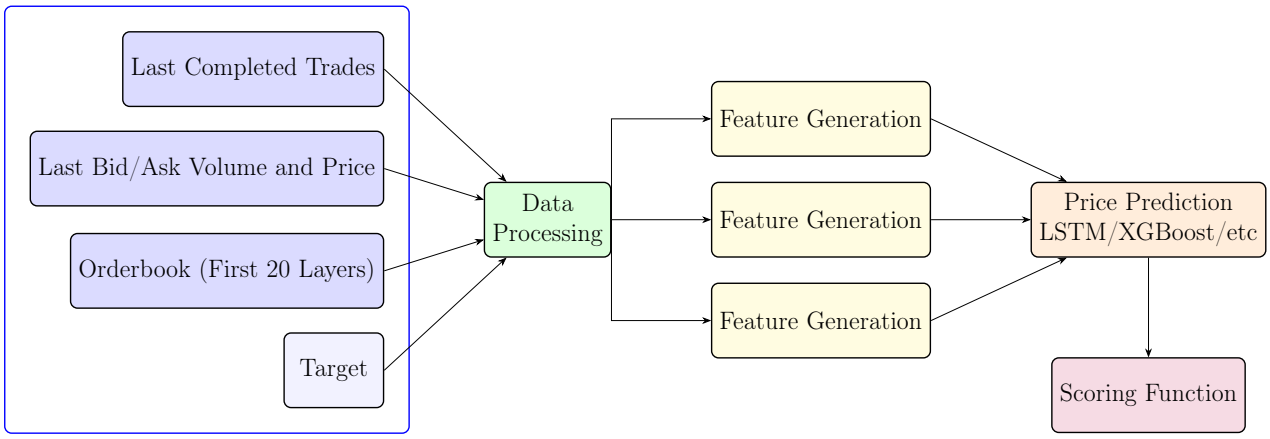

Рис. 5.1: Архитектура высокочастотного ML бота для прогнозирования цен

5 Разработка модели

5.1 Специфика признаков

5.1.1 Выбор размера временного окна

В нашей модели мы используем различные временные окна для выделения разнообразных аспектов динамики рынка. Отражение этих аспектов происходит благодаря разнообразию функций, включенных в наш анализ. Этот подход допускает выявление закономерностей и тенденций на различных временных горизонтах, что является ключевым фактором в адаптации модели к разнообразным условиям рынка.

Особый акцент ставится на гибкость модели в адаптации к различным временным характеристикам рыночной ситуации. Короткие временные окна, такие как 100ms, 250ms и 500ms, быстро реагируют на недавние изменения, в то время как более длинные окна, такие как 1000ms и 2000ms, фиксируют более продолжительные тенденции. Присутствие этих разнообразных временных окон в нашей модели обеспечивает ее адаптивность и эффективность на различных временных масштабах, что существенно повышает ее универсальность в анализе и прогнозировании динамики финансовых рын-

ков.

5.1.2 Выбор между спотовым рынком и рынком фьючерсов

В нашем подходе к финансовому моделированию мы выступаем за включение характеристик как спотового, так и фьючерсного рынков, чтобы достичь более тонкого и всестороннего понимания динамики рынка. Эта стратегия согласуется с нашим убеждением, что хорошо продуманная модель должна использовать сильные стороны обоих рынков, объединяя информацию, поступающую в режиме реального времени со спотовых рынков, и перспективную информацию, поступающую с фьючерсных рынков. Синтез этих различных точек зрения повышает аналитические возможности и эффективность модели. Вот три убедительные причины, подтверждающие нашу позицию:

Всестороннее понимание динамики рынка Сочетание характеристик спотового и фьючерсного рынков обеспечивает комплексное понимание динамики рынка. Спотовые рынки предлагают данные в режиме реального времени, отражающие текущие условия, в то время как фьючерсные рынки дают представление об ожиданиях и настроениях рынка в течение определенного времени. Такая двойная перспектива повышает способность модели улавливать как немедленные реакции, так и ожидаемые будущие движения, что позволяет получить более целостное представление о финансовых рынках.

Улучшенное управление рисками и прогнозирование Использование особенностей спотовых и фьючерсных рынков способствует совершенствованию стратегий управления рисками. Фьючерсные рынки, часто ис-

пользуемые для хеджирования, позволяют модели оценивать позиции участников рынка в отношении потенциальных будущих движений. Такая интеграция повышает предсказательную способность модели за счет использования информации с обоих рынков, что позволяет более тонко оценивать меняющиеся рыночные условия и поддерживать более эффективные стратегии снижения рисков.

Временной анализ и анализ настроений Синергия между спотовыми и фьючерсными рынками позволяет проводить временной анализ и оценку настроений. Спот-рынки предоставляют информацию в режиме реального времени, в то время как фьючерсные рынки предлагают перспективу. Такое сочетание позволяет модели адаптироваться к меняющимся рыночным условиям, предвидеть изменения в настроениях инвесторов и получить более глубокое понимание поведения рынка на различных временных горизонтах. По сути, включение функций спотового и фьючерсного рынков повышает сложность и адаптивность финансовых моделей, позволяя им отражать все сложности современных финансовых рынков.

5.2 Описание всех признаков

Начнем обзор признаков с изучения простых функций, постепенно переходя к более сложным. В ходе дальнейшего развития этой дискуссии необходимо отметить значительный вклад исследований, представленных в [7] и [8]. В этих работах представлены фундаментальные идеи и методологии, которые являются неотъемлемой частью углубленного понимания темы нашего исследования

Order Book Imbalance Признак OBI в нашей модели высокочастотной торговли определяет нормализованную разницу в объемах торгов между лучшим предложением (B) и лучшим спросом (A). Математически это выглядит следующим образом:

$$OBI = \frac{B - A}{B + A}$$

Эта метрика служит индикатором динамики рынка в реальном времени: положительные значения свидетельствуют о преобладании интереса к покупке, а отрицательные - о преобладании интереса к продаже. Функция OBI повышает адаптивность нашей модели к быстро меняющимся условиям портфеля заказов, способствуя своевременному принятию решений в динамичных торговых средах.

Log Order Book Imbalance LOBI определяет логарифмическое соотношение совокупных объемов спроса (B) и предложения (A):

$$\text{LOBI} = \log\left(\frac{\text{B}}{\text{A}}\right)$$

По сравнению с традиционным дисбалансом книги заказов (OBI), функция LOBI вводит логарифмическое преобразование, повышая чувствительность к изменениям в динамике книги заказов. Такая повышенная чувствительность позволяет LOBI улавливать тонкие изменения в рыночных дисбалансах. Обе характеристики дают ценные сведения, при этом LOBI предлагает более тонкую перспективу, которая может быть особенно полезна в быстро меняющейся и волатильной торговой среде. Выбор между LOBI и OBI зависит от конкретных требований торговой стратегии и желаемого

уровня чувствительности к колебаниям портфеля заказов.

Trade Period Cumulative Imbalance Используя функции get_l_r и trades_perio вычисляются совокупные объемы покупки (trades_buy и продажи (trades_sell) в заданном окне оглядки. Математически это выражается как:

$$\text{IMB}(\text{trades}, \text{lookback\_window}) = \text{trades\_buy} - \text{trades\_sell}$$

Эта функция дает представление о временном торговом дисбалансе, помогая нашей модели оценить преобладающее давление на покупку или продажу в течение указанного периода обратного просмотра. Совокупные объемы торгов способствуют всестороннему пониманию динамики рынка, способствуя принятию обоснованных решений в сценариях высокочастотной торговли.

Линия Накопления/Распределения ADL измеряет кумулятивный поток капитала в актив или из него. В варианте ADL_{timed} мы захватываем эту аккумуляцию или распределение за указанный временной интервал, в то время как ADL_{last} предоставляет снимок последнего наблюдаемого значения. Математически это представлено как:

$$\text{ADL}_{timed}(t) = \text{ADL}_{timed}(t-1) + \frac{(\text{close} - \text{low}) - (\text{high} - \text{close})}{\text{high} - \text{low}} \times \text{volume}$$

$$\text{ADL}_{last} = \text{ADL}_{timed}(T)$$

Индекс Среднего Направления ADX количественно определяет силу тренда на рынке. ADX_{timed} и ADX_{last} представляют значения ADX за определенный временной интервал и последнее наблюдаемое значение соответственно. Расчет включает сглаживание направленного движения с течением времени и выражается как:

$$\mathrm{ADX}_{timed}(t) = (1 - \alpha) \times \mathrm{ADX}_{timed}(t - 1) + \alpha \times \left| \frac{\mathrm{DI\_plus} - \mathrm{DI\_minus}}{\mathrm{DI\_plus} + \mathrm{DI\_minus}} \right|$$

$$\mathrm{ADX}_{last} = \mathrm{ADX}_{timed}(T)$$

Осциллятор Шанда Моментума CMO измеряет момент актива. CMO_{timed} и CMO_{last} охватывают этот момент за указанный временной интервал и как последнее наблюдаемое значение соответственно. Расчет представлен как:

$$\mathrm{CMO}_{timed}(t) = \frac{\mathrm{sum\_up} - \mathrm{sum\_down}}{\mathrm{sum\_up} + \mathrm{sum\_down}}$$

$$\mathrm{CMO}_{last} = \mathrm{CMO}_{timed}(T)$$

Скорость Изменения ROC количественно определяет процентное изменение цены за указанный период времени. ROC_{timed} и ROC_{last} представляют значения ROC за временной интервал и последнее наблюдаемое значение соответственно. Математически это выражается как:

$$\mathrm{ROC}_{timed}(t) = \frac{\mathrm{close}(t) - \mathrm{close}(t - \mathrm{period})}{\mathrm{close}(t - \mathrm{period})} \times 100$$

$$\mathrm{ROC}_{last} = \mathrm{ROC}_{timed}(T)$$

Моментум MOM измеряет скорость изменения цены актива. MOM_{timed} и MOM_{last} представляют моментум за временной интервал и как последнее наблюдаемое значение соответственно. Расчет дан следующим образом:

$$\mathrm{MOM}_{timed}(t) = \mathrm{close}(t) - \mathrm{close}(t - \mathrm{period})$$

$$\mathrm{MOM}_{last} = \mathrm{MOM}_{timed}(T)$$

Индекс Относительной Силы Индекс относительной силы (RSI) оценивает величину последних изменений цены. RSI_{timed} и RSI_{last} представляют значения RSI за временное окно и как последнее наблюдаемое значение, соответственно. Формула для RSI выглядит следующим образом:

$$\mathrm{RSI}_{timed}(t) = 100 - \frac{100}{1 + \frac{\mathrm{avg\_gain}}{\mathrm{avg\_loss}}}$$

$$\mathrm{RSI}_{last} = \mathrm{RSI}_{timed}(T)$$

Stochastic RSI (Timed and Last) Стохастический индекс относительной силы (RSI) - это вариант традиционного RSI, используемый для определения условий перекупленности или перепроданности. ‘Stochastic RSI (Timed)‘ отражает значение за определенный временной интервал, а ‘Stochastic RSI (Last)‘ представляет самое последнее значение. Расчет производится следующим образом:

$$\text{Stochastic RSI} = \frac{\text{RSI} - \min(\text{RSI})}{\max(\text{RSI}) - \min(\text{RSI})}$$

Линейные регрессионные признаки Эти характеристики получены из линейного регрессионного анализа ценовых движений. ‘Интерцепт (Timed)‘ и ‘Интерцепт (Last)‘ относятся к y-интерцептам, ‘Склон (Timed)‘ и ‘Склон (Last)‘ - к наклонам, а ‘Коэффициент корреляции (Timed)‘ и ‘Коэффициент корреляции (Last)‘ - к коэффициентам корреляции линий регрессии. Общая форма линейной регрессии такова:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Линия накопления/распределения (AC) . Индикатор AC отслеживает динамику спроса и предложения, суммируя объем с поправкой на относительное положение закрытия в диапазоне high low.

$$\text{AC} = \text{Previous AC} + \text{Объем} \times \frac{(\text{Close} - \text{Low}) - (\text{High} - \text{Close})}{(\text{High} - \text{Low})}$$

Изменение цен (PC) Эта характеристика показывает скорость изменения цены за определенный период.

$$\text{PC} = \frac{\text{Текущая цена} - \text{Предыдущая цена}}{\text{Предыдущая цена}}$$

Custom Oscillator Indicator (COIN) COIN - это собственный осциллятор, который объединяет различные рыночные индикаторы для создания композитного рыночного сигнала.

Индикаторы волатильности (VI) Индикаторы волатильности, обозначаемые как VI_0, VI_1, VI_2 и VI_3, количественно определяют волатильность рынка на различных временных интервалах.

Адаптивная логистическая регрессия (ALogReg) Это модель логистической регрессии, которая адаптирует свои параметры на основе недавнего поведения рынка.

Относительный объем (RVol) RVol сравнивает текущий объем торгов со средним объемом за то же время дня или сессии.

$$\text{RVol} = \frac{\text{Текущий объем}}{\text{Средний объем}}$$

Range Kurtosis (RK) RK оценивает тенденцию ценовых диапазонов к образованию выбросов, отражающих экстремальные ценовые движения.

$$K = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Jump Variance (JumpVar) JumpVar выделяет и количественно оценивает дисперсию, связанную со скачками цен, отдельно от общей волатильности рынка.

5.3 Обзор предиктивной модели

Выбор правильных моделей прогнозирования для высокочастотной торговли (HFT), особенно для прогнозирования движения цен в валютных парах ETH/USD и USD/ETH, требует тщательного рассмотрения сильных сторон моделей и их соответствия характеристикам данных.

1 Линейная регрессия

Почему Линейная регрессия используется в качестве базовой модели для сравнения. Она представляет собой простой, но мощный метод, который позволяет установить базовый уровень эффективности предсказания перед применением более сложных моделей.

Актуальность Хотя линейная регрессия может быть неспособна улавливать сложные нелинейные зависимости, присущие рынку криптовалют, она обеспечивает важный эталон для оценки производительности более сложных моделей.

2 Сети длинной кратковременной памяти (LSTM)

Почему LSTM - это тип рекуррентной нейронной сети (RNN), которая особенно хорошо справляется с данными временных рядов, что очень важно для HFT. Они могут улавливать долгосрочные зависимости и закономерности в движении цен, что делает их хорошо подходящими для волатильной и последовательной природы криптовалютных рынков.

Актуальность Идеально подходит для моделирования временных последовательностей, присущих данным финансового рынка, что крайне важно для точного прогнозирования цен.

3 Градиентные бустинговые машины (например, XGBoost, LightGBM)

Почему Эти алгоритмы известны своей высокой производительностью и эффективностью при работе с различными типами данных и распределениями. Можели устойчивы к переобучению и могут

обрабатывать большие наборы данных с множеством признаков, что типично для HFT.

Актуальность Их способность моделировать сложные нелинейные зависимости очень важна для непредсказуемого и быстро меняющегося криптовалютного рынка.

4 Случайный лес

Почему Метод ансамбля, который строит несколько деревьев решений и объединяет их для получения более точного и стабильного прогноза. Random Forest может обрабатывать большое количество признаков и относительно невосприимчив к перестройке, что делает его надежным выбором.

Актуальность Случайность в выборе признаков позволяет охватить различные аспекты рынка, обеспечивая всестороннее видение для прогнозирования цен.

Каждая из этих моделей обладает уникальными преимуществами, такими как обработка последовательных данных, улавливание сложных взаимосвязей или эффективная обработка высокоразмерных данных, что делает их хорошо подходящими для динамичной и сложной природы HFT на криптовалютных рынках. Выбор модели в конечном итоге зависит от конкретных характеристик набора данных, доступных вычислительных ресурсов и конкретных целей торговой стратегии.

6 Эксперименты и оценка качества

6.1 Подбор гиперпараметров

В нашем исследовании HFT, сфокусированном на валютных парах ETH/USD и USD/ETH, после первоначального этапа оценки моделей мы определили подмножество моделей машинного обучения, которые демонстрируют наибольшие перспективы с точки зрения точности прогнозирования и вычислительной эффективности. Учитывая их потенциал, мы решили углубиться в оптимизацию этих моделей с помощью строгой настройки гиперпараметров.

6.1.1 LSTM

После тщательных экспериментов и оценок мы определили наиболее эффективные гиперпараметры для нашей LSTM-модели, используемой в высокочастотной торговле. Окончательные параметры, которые значительно повысили производительность нашей модели, приведены в таблице 6.1.

Полный обзор процесса настройки гиперпараметров и выбранных значений приведен в таблице 6.1.

Гиперпараметр	**Предлагаемый диапазон**
Количество слоев	1-4
Количество нейронов в слое	50-50
Скорость обучения	0.1, 0.01, 0.00
Размер партии	16, 32, 64, 128, 25
Коэффициент отсева	0 - 0.5
Функция активации	tanh, relu, sigmoid
Эпохи	Основано на ранней остановке
Оптимизатор	SGD, Adam, RMSprop

Таблица 6.1: Диапазоны для гиперпараметров LSTM в модели HFT

Гиперпараметр	Предлагаемый диапазон
Скорость обучения (learning rate)	0.01, 0.05, 0.1, 0.2
Количество деревьев ($n_estimators$)	100, 200, 300, 500
Глубина дерева (max_depth)	3, 6, 9, 12
Min вес в узле (min_child_weight)	1, 3, 5, 7
Коэффициент подвыборки (subsample)	0.6, 0.7, 0.8, 0.9
Регуляризация L1 (alpha)	0, 0.001, 0.005, 0.01
Регуляризация L2 (lambda)	1, 1.5, 2, 3

Таблица 6.2: Диапазоны для гиперпараметров XGBoost в модели HFT

6.1.2 XGB

Мы выбрали оптимальные гиперпараметры для нашей модели XGBoost в высокочастотной торговле после тщательной настройки и анализа. Выбранные параметры, подробно описанные в таблице 6.2, представляют собой сбалансированную комбинацию для достижения наилучшей производительности в нашем конкретном торговом контексте.

6.2 Оценка признаков

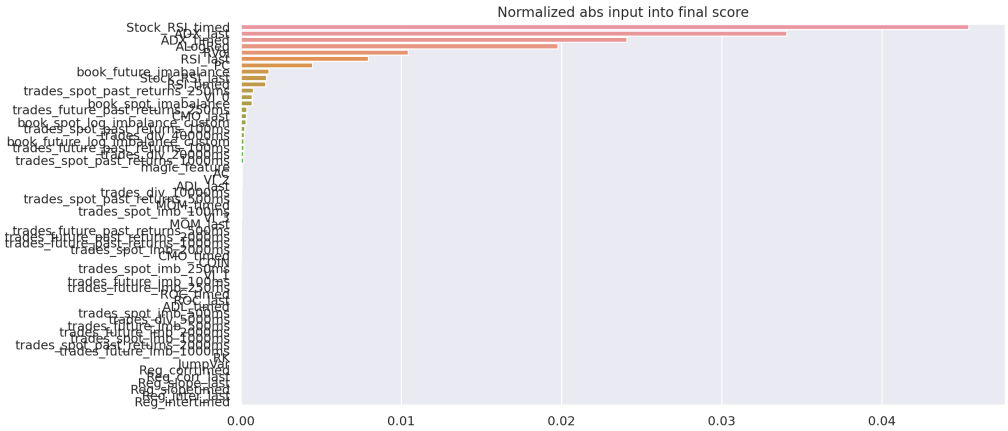

Рис. 6.1: Визуальное представление важности признаков в модели линейной регрессии.

Мы провели углубленный анализ значимости каждой характеристики в нашей модели высокочастотной торговли, изучив их влияние на результаты, предсказанные нашей эталонной линейной регрессионной моделью. Этот

метод оценки позволил получить ценные сведения о том, какие характеристики оказывают наибольшее влияние на формирование прогнозов модели, что позволило нам оптимизировать набор характеристик для повышения эффективности модели. Для наглядного представления этого анализа признаков, пожалуйста, обратитесь к иллюстрации, представленной на рисунке 6.1.

6.3 Ключевая метрика

Model	**Size**	**Metric**	**Inference Time (ms)**
all zeros	0.070KB	0.0	14.01
ideal model	0.070KB	2825.38	24.89
model_random	0.071KB	-75.91	63.25
linear	2.590KB	9.41	608.06
LSTM	5.000MB	320.36	33969.46
XGB	1.370MB	158.04	5445.39

Таблица 6.3: Final Metrics, Model Sizes, and Inference Times

В нашем исследовании высокочастотной торговли (HFT) мы используем линейную модель в качестве эталона для оценки других сложных моделей, таких как LSTM и XGB. Такой бенчмаркинг очень важен для понимания компромисса между простотой и вычислительной эффективностью и точностью прогнозирования в более сложных моделях.

В то время как линейная модель с ее относительно скромным временем вывода 608,06 мс и метрикой производительности 9,41 служит в качестве базовой, модель LSTM демонстрирует превосходную точность прогнозирования с метрикой 320,36. Однако это достигается за счет значительно большего времени вывода, составляющего 33 969,46 мс (около 34 секунд). Такая длительная обработка данных свойственна LSTM, работающей по принципу последовательной обработки данных, что является существенным ограни-

чением в высокоскоростной сфере HFT, где решения должны приниматься практически мгновенно.

Напротив, модель XGB обеспечивает более благоприятный баланс между точностью и скоростью. Несмотря на более низкую производительность (158,04) по сравнению с LSTM, ее время вывода значительно меньше - 5 445,39 мс (около 5,4 секунды). Древовидная структура XGB способствует более быстрому принятию решений, что в большей степени соответствует требованиям быстрого темпа работы в среде HFT.

Таким образом, несмотря на то, что LSTM может превосходить по точности, гибкость модели XGB делает ее более практичным и эффективным выбором для реальных торговых сценариев, где важна быстрая реакция на колебания рынка. Этот анализ подчеркивает, что при выборе моделей для HFT-приложений важно учитывать как точность, так и операционную целесообразность.

6.4 P&L

В нашем исследовании весовые коэффициенты матрицы несоответствий играют ключевую роль в интерпретации эффективности наших торговых алгоритмов, выходя за рамки простых показателей точности. Связывая эти веса с потенциальной прибылью и убытками (P&L) сделок, мы эффективно преодолеваем разрыв между теоретической эффективностью модели и практическими финансовыми последствиями. Такое согласование позволяет нам построить в реальном времени график 6.2 прибылей и убытков для наших алгоритмов, предлагая в реальном времени ощутимое представление их эффективности в динамичной среде высокочастотной торговли (HFT).

График 6.2 P&L, основанный на взвешенных результатах матрицы несо-

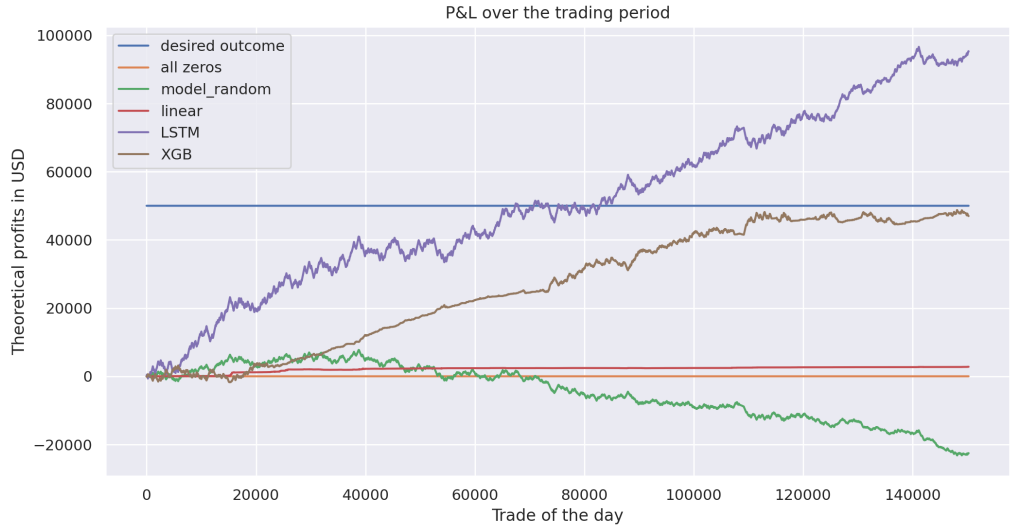

Рис. 6.2: График теоретической прибыли и убытков торговых алгоритмов

ответствий, служит мощным инструментом для визуализации финансового воздействия наших торговых стратегий. Он не только иллюстрирует потенциальные прибыли и убытки с течением времени, но и дает бесценное представление о соотношении риска и вознаграждения каждой модели. Такое практическое применение результатов наших исследований свидетельствует об актуальности и применимости нашей работы в реальных торговых сценариях.

7 Заключение

7.1 Результаты

Наше исследование в области высокочастотной торговли (HFT) с использованием машинного обучения показало значительные результаты. Мы успешно разработали и протестировали несколько моделей, среди которых LSTM и XGB выделялись своей эффективностью. Особенно заметной была высокая точность модели LSTM, однако ее применение в реальной торговле ограничено из-за значительного времени инференса. С другой стороны,

модель XGB, несмотря на немного меньшую точность, демонстрировала гораздо более высокую скорость обработки данных, что делает ее более подходящей для применения в HFT. Также был проведен анализ веса признаков и их влияния на торговые стратегии, что позволило создать реальный график P&L (прибыль и убытки), демонстрирующий практическую применимость наших исследований.

7.2 Возможные направления дальнейшего развития

В качестве дальнейшего развития исследования мы видим несколько перспективных направлений. Во-первых, оптимизация и улучшение скорости обработки данных моделью LSTM, возможно, за счет использования методов уменьшения размерности или параллельных вычислений. Во-вторых, более глубокий анализ влияния экономических и новостных факторов на торговые стратегии может привести к разработке более сложных и адаптивных моделей. Также интерес представляет применение гибридных моделей, сочетающих преимущества различных подходов. Наконец, расширение области исследования на другие финансовые инструменты и рынки может открыть новые возможности для торговых стратегий и их оптимизации.

8 Список литературы

1. F. Guilbaud and H. Pham, "Optimal high-frequency trading with limit and market orders," *Quantitative Finance*, vol. 13, no. 1, pp. 79–94, 2013.

2. R. Cont and A. de Larrard, "Price dynamics in a markovian limit order market," *SIAM Journal on Financial Mathematics*, vol. 4, no. 1, pp. 1–25, 2013.

3. M. Avellaneda and S. Stoikov, "High-frequency trading in a limit order book," *Quantitative Finance*, vol. 8, no. 3, pp. 217–224, 2008.

4. Y. Zhang, F. Ma, and Y. Wang, "Forecasting crude oil prices with a large set of predictors: Can lasso select powerful predictors?," *Journal of Empirical Finance*, vol. 54, pp. 97–117, 2019.

5. C. Liang, Y. Xu, Z. Chen, and X. Li, "Forecasting china's stock market volatility with shrinkage method: Can adaptive lasso select stronger predictors from numerous predictors?," *International Journal of Finance & Economics*.

6. N. N. Thach, L. H. Anh, and H. N. Khai, *Applying Lasso Linear Regression Model in Forecasting Ho Chi Minh City's Public Investment*, pp. 245–253. Cham: Springer International Publishing, 2021.

7. A. Ntakaris, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Mid-price prediction based on machine learning methods with technical and quantitative indicators," *PLOS ONE*, vol. 15, p. e0234107, June 2020.

8. J. Albers, M. Cucuringu, S. Howison, and A. Y. Shestopaloff, "Fragmentation, price formation, and cross-impact in bitcoin markets," 2021.

9 Приложение

9.1 Исходный код

Чтобы поддержать прозрачность и облегчить сотрудничество в этой области, мы разместили исходный код наших исследований в области Высокочастотной Торговли (HFT) в открытом доступе на GitHub. GitHub - это

широко признанная веб-платформа для контроля версий и совместной разработки программного обеспечения.

Наш репозиторий содержит те части наших моделей и алгоритмов, которые не опираются на проприетарную информацию или библиотеки и доступны для публичного просмотра и использования. Он представляет собой ресурс для исследователей и практиков, желающих ознакомиться с общими методологиями, применяемыми в нашей работе. Это позволяет коллегам из научного сообщества исследовать, адаптировать и развивать наши подходы, соблюдая при этом условия соглашения о неразглашении (NDA), регулирующего доступ к более чувствительным аспектам наших исследований.

https://github.com/Melnikovartem/non-nda-hft-hse

9.2 График финальной метрики для различных моделей

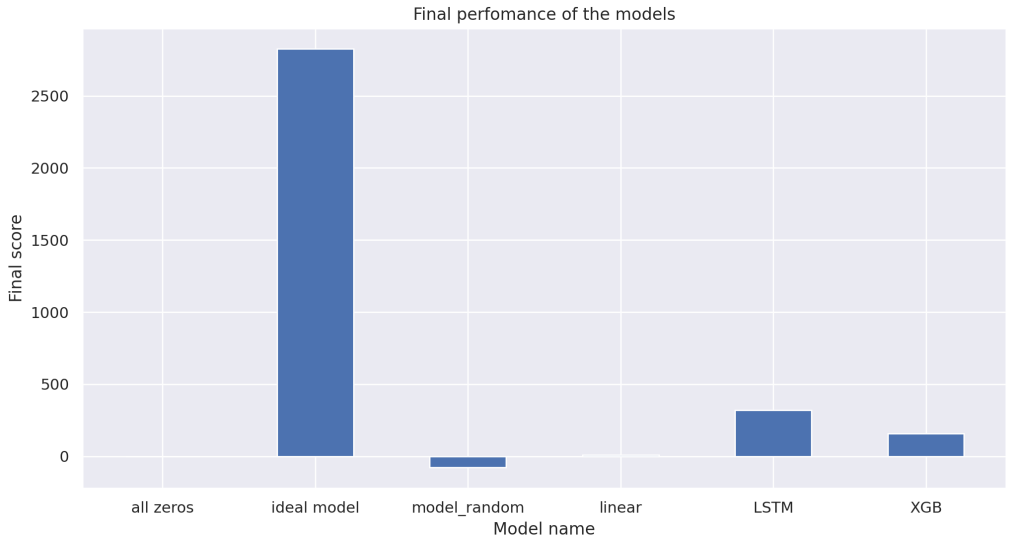

Рис. 9.1: Гистограмма итогового score для каждой модели

9.3 Финальная таблица со всеми признаками

book_future_imabalance book_future_log_imbalance_custom book_spot_imabalance book_spot_log_imbalance_custom	trades_future_imb_100ms trades_future_imb_250ms trades_future_imb_500ms trades_future_imb_1000ms
trades_future_imb_2000ms trades_future_past_returns_100ms trades_future_past_returns_250ms trades_future_past_returns_500ms	trades_spot_imb_100ms trades_spot_imb_250ms trades_spot_imb_500ms trades_spot_imb_1000ms
trades_future_past_returns_1000ms trades_future_past_returns_2000ms trades_spot_past_returns_250ms trades_spot_past_returns_1000ms	trades_spot_imb_2000ms trades_spot_past_returns_100ms trades_spot_past_returns_500ms trades_spot_past_returns_2000ms
trades_div_5000ms trades_div_10000ms trades_div_20000ms trades_div_40000ms	ADL_timed ADL_last ADX_timed ADX_last
CMO_timed CMO_last MOM_timed MOM_last	ROC_timed ROC_last RSI_timed RSI_last
Stock_RSI_timed Stock_RSI_last Reg_corrtimed Reg_slope_last	Reg_intertimed Reg_slopetimed Reg_inter_last Reg_corr_last
AC PC COIN ALogReg	VI_0 VI_1 VI_2 VI_3
RVol JumpVar	RK

Таблица 9.1: Полный список признаков в модели HFT, сгруппированных для наглядности