**Автор: Владислав Стерхов**

**Сопроводительное письмо**

Квалификационная работа на тему:

«Анализ временных рядов на криптовалютной бирже»

В рамках данной квалификационной работы будут рассмотрены основные алгоритмы машинного обучения, такие как линейная регрессия, дерево решений. Также рассмотрены методы сингулярного спектрального анализа, проведены рассуждения на тему применения их с специфическими методами для временных рядов, таких как ARIMA, GARCH, SARIMAX. Построена одна из специфических моделей и приведены результаты предсказаний.

Цели проекта:

1. Реализовать модуль для подключения к криптовалютной бирже Binance и получения торговых данных, таких как: цена, стакан, объемы торгов;
2. Реализовать модуль агрегации торговых данных для получения математических величин, таких как: среднее, стандартное отклонение, дисперсия, максимум, минимум, сумма.
3. Провести анализ данных, для выявления параметров, которые косвенно влияют на цены торговой пары.
4. Разработать программный комплекс для анализа агрегированных данных, по принципам сингулярного спектрального анализа
5. Разработать модель машинного обучения для предсказания временных рядов

Данная работа имеет актуальность в сфере трейдинга криптовалют по следующим критериям:

1. Трейдинг – это сфера, где каждая секунда имеет огромное значение. Таким образом, постоянный мониторинг информации о торгах, с интервалом в 1 секунду, играет ключевую роль в следующих ситуациях:

* Своевременная реакция, на внезапные, крупные изменения на рынке, в следствии чего позиция своевременно может быть закрыта без крупных финансовых убытков.
* Возможность удачного входа в позицию – шорт или лонг, в следствии чего можно получить большую прибыль.

1. Возможность получать бесплатный и безлимитный доступ к торговой информации о всех торговых инструментах, которые прошли листинг на бирже Binance.
2. Возможность хранить информацию о торгах локально, без ограничения на объем и время.
3. В следствии доступности информации реализован эффективный расчет математических величин, перечисленных в работе ранее, скорость которого не сравнимо больше, чем скорость человеческого расчета. Также, становится возможным внедрение автоматизированной покупки и продажи активов, расчета рисков и прибыли, любых других формул лучших практик мирового трейдинга, эффективность которых будет максимально высокой, в сравнении в человеческой.
4. Возможность внедрения различных торговых стратегий, техническая часть которых может быть реализована на любых предобученных моделях машинного обучения.

В качестве основной метрики качества для модели ARIME выбран AIC. Метрика Информационный критерий Акаике (AIC) является одним из наиболее распространенных выборов для оценки качества модели ARIMA. Данная метрика выбрана в качестве основного инструмента оценки качества модели ARIMA по ряду причин:

1. Учитывает баланс между сложностью модели и ее точностью: AIC принимает во внимание сложность модели через количество параметров, необходимых для построения модели, и штрафует модели с большим числом параметров. Это позволяет найти оптимальный баланс между точностью предсказания и сложностью модели.

2. Обеспечивает сравнение различных моделей: AIC позволяет сравнить несколько моделей ARIMA, используя одну и ту же метрику. Это облегчает выбор наилучшей модели, исходя из минимизации AIC.

3. Устойчивость к переобучению: AIC имеет встроенную защиту от переобучения модели, так как штрафует модели с большим числом параметров. Это помогает избежать модели, которая "переучивается" на имеющихся данных и не обобщается хорошо на новые данные.

4. Объективный подход к выбору модели: AIC является статистическим критерием, который основывается на математических принципах. Это делает его более объективным и надежным выбором для оценки качества моделей ARIMA.

5. Широко используется в литературе и практике: AIC является широко используемым критерием оценки качества моделей ARIMA и встречается во многих источниках литературы. Это делает его удобным выбором для сравнения с другими исследованиями и использования в практических приложениях.

При сравнении двух моделей та, у которой меньше AIC, как правило, «лучше».

Формула AIC (Информационный критерий Акаике): 

где k — число параметров модели, L — максимизированное значение функции правдоподобия модели.

Кроме того, будут рассмотрены методы оптимизации и настройки параметров модели в качестве отбора наилучших AIC.

Для начала необходимо привести перечень данных, которые предоставляет криптовалютная биржа Binance для каждой валютной пары (инструмента):

* Цена
* Запросы на покупку в стакане (Asks)
* Запросы на продажу в стакане (Bids)
* Совершенные продажи за период времени
* Совершенные покупки за период времени

Подключение к криптовалютной бирже и получение исходных торговых данных располагается на github, по ссылке - <https://github.com/broom888/CryptoCurrencyTrading_public/tree/master/binance_kafka_producer>

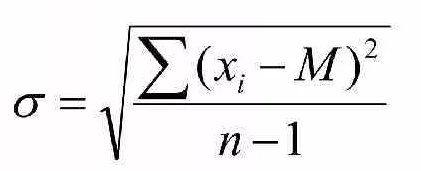
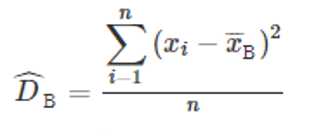
При анализе исходных торговых данных в первую очередь стоит обратить внимание на покупки и продажи, они могут содержать колоссальную разницу в объемах, за счет разницы в состоянии игрока. Эти параметры требуют усреднения и расчета математических величин.

Чем дольше набирается позиция, тем большим получится её объем и, соответственно, тем сильнее будет движение в ту или иную сторону. [3, с. 96]

Данная особенность набора данных может трактоваться как шумы, однако, ход торгов валютной пары не может быть удален, т.к. это влечет за собой потерю объективности. Эти данные должны быть учтены в модели и записаны как “Стандартное отклонение” и “Дисперсия”, а так-же более простые математические методы: Среднее, Максимум, Минимум, Сумма. Таким образом реализовано сглаживание шумов. Модуль агрегации данных располагается на github, по ссылке –

<https://github.com/broom888/CryptoCurrencyTrading_public/tree/master/aggregator_kafka_consumer>

Формулы для расчета математических величин:

* Стандартное отклонение  , где:
  + Xi – каждое наблюдаемое значение,
  + M – среднее арифметическое выборки X,,
  + n – размерность выборки.
* Дисперсия  , где:
  + Xi – каждое наблюдаемое значение,
  + Xb – среднее арифметическое выборки X,
  + n – размерность выборки.

В ходе агрегации исходных данных мы получаем два объекта: AggregatedTradeVolume, и более расширенный - Candle, с временных интервалом в 1 и 15 минут, соответственно. Модели данных для этих объектов представлены в приложении А и Б. После таких преобразований мы имеем усредненные данные, со всеми показателями отклонений от среднего.

Для задач фундаментального анализа используется модель Candle с 15 минутным интервалом данных. Из методов фундаментального анализа временных рядов на бирже можно выделить следующие методы:

1. Графический метод - данный метод используется для визуального изучения временных рядов, построения графиков и нахождения трендов, циклов и сезонностей.
2. Технический анализ - изучение графиков цен и объемов торгов, чтобы определить тренды и находить точки входа и выхода на рынок.

Реализация данных методов представлены по ссылкам:

* Для объекта Candle – <https://github.com/broom888/CryptoCurrencyTrading_public/blob/master/data_science/ShowCandlesSeries.ipynb>
* Для объекта AggregatedTradeVolume – <https://github.com/broom888/CryptoCurrencyTrading_public/blob/master/data_science/ShowAggEntity.ipynb>

Об особенностях ценообразования валютной пары можно получить больше информации после агрегации исходных торговых данных в объекты с математическими величинами, выведя корреляционную матрицу между ними.

Далее, для проведения анализа и предсказаний будет использоваться модель AggregatedTradeVolume с временных интервалов в 1 минуты. Следующим шагом мы проверим теоретические заблуждения, я том, что возможно предсказывать уменьшение или увеличение следующей цены, без каких либо преобразований, с помощью простых моделей, таких как:

* Линейная регрессия
* Дерево решений

Полученная точность этими методами близится к 0.5, и не соответствуют нашим требованиям об эффективности. Практика с простыми моделями и разделением на обучающее и тестовое множество представлено по ссылке – <https://github.com/broom888/CryptoCurrencyTrading_public/blob/master/data_science/LinearRegAndDecTree.ipynb>

По этому, хочется представить метод разложения временного ряда на главные компоненты – PCA (Principal Component Analysis). Данный метод содержит следующие преобразования ряда, разделенные на 4 основных шага:

1. Подготовка данных. Входные данные должны быть в формате временных рядов.

2. Формирование траекторной матрицы. Для этого следует выбрать длину окна, которая определяет размер траектории, и сформировать матрицу траекторий.

3. Выделение сингулярных чисел. Используя сингулярное разложение матрицы траекторий, можно выделить компоненты с наибольшими сингулярными числами.

4. Разделение компонент. Компоненты могут быть разделены на тренды, циклы и нерегулярные колебания путем анализа сингулярных векторов и выбора подходящих для каждой компоненты.

Данный метод, теоретически, можно использовать при обучении моделей ARIMA, GARCH, SARIMAX для предсказывания тренда, сезонности и шума, по отдельности, что может положительно влиять на динамику сделок. Реализация метода разложения временного ряда на главные компоненты представлена по ссылке на github – <https://github.com/broom888/CryptoCurrencyTrading_public/blob/master/data_science/PCA.ipynb>

Из аналогичных работ можно выделить работу с применением PCA для искусственного временного ряда, ссылка – <https://www.kaggle.com/code/leokaka/pca-for-time-series-analysis>. В данной работе проводится изучения метода PCA, представлены формулы расчетов. Успешно проведена реконструкция 12 элементов временного ряда из траекторной матрицы. Приведены размышления об подходах к декомпозиции временного ряда и производительности метода, на ряду c методом SSA. Также, успешно получены такие компоненты временного ряда как тренд, сезонность и шум, как временной ряд.

Далее, на рисунке 1, представлены результат декомпозиции искусственного временного ряда на главные компоненты – тренд, сезонность и шум. Компоненты выведены на график вместе с искусственным временным рядом.

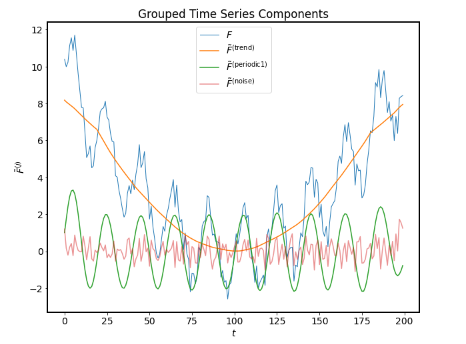


Рисунок 1 — Декомпозиция временного ряда на главные компоненты

На рисунке 2 изображены все три компонента – тренд, сезонность, шум, на отдельных графиках.

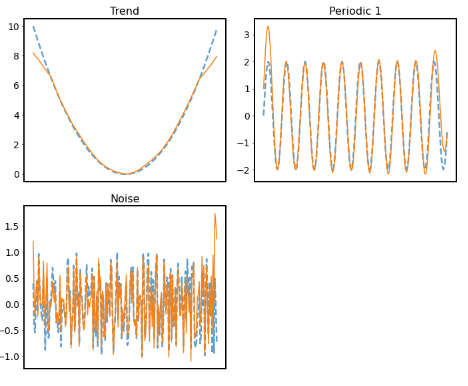


Рисунок 2 — Графики тренда, сезонности и шума

Построим модель ARIMA, для предсказания цены для представленного временного ряда BTCUSDT, содержащего данные за 400 минут. Реализация модели и предсказания представлена по ссылке на github – <https://github.com/broom888/CryptoCurrencyTrading_public/blob/master/data_science/ARIMA.ipynb>

Итоговая модель ARIMA имеет следующие показатели точности AIC: -5331.31815, что является наилучшими показателями подбираемых моделей из доступных значений параметров 5х5х5. При этом наилучшими параметрами оказались значения: 3, 0, 4.

Выбор наилучших параметров модели ARIMA выполнен с помощью циклического обучения моделей ARIMA на наборе данных и выборе наилучшего показателя Информационного критерия Акаике.

В ходе этой работы было проведено исследование возможностей применения машинного обучения в трейдинге. Были рассмотрены как классические методы машинного обучения, так и современные методы специализированные для временных редов.

Цели достигнуты, в результате удалось разработать комплексное решение для получения, хранения, анализа торговых данных и предсказания цены инструментов криптовалютной биржи Binance

Результатом внедрения данных методик в работу трейдера, несомненно, может являться большая прибыльность сделок. В целом, можно сделать выводы, что при запланированном развитии данного проекта гарантирована финансовая прибыль и окупаемость работ. Однако, необходимо учитывать сложность задачи и многообразие факторов, влияющих на рынок.

В заключение хочется отметить пути развития проекта, их можно перечислить по пунктам:

1. Создание системы для покупки\продажи активов на решении искусственного интеллекта.

Это позволяет улучшить скорость и точность торговли, а также уменьшить человеческий фактор.;

1. Расширение источников данных: Использование широкого спектра данных для обучения моделей, включая финансовые данные, новостные статьи, социальные медиа-посты и другие внешние факторы, которые могут влиять на рынок;
2. Интеграция с облачными ресурсами: Использование облачных вычислений для улучшения масштабируемости и производительности моделей. Это позволяет быстро обрабатывать большие объемы данных и выполнять сложные вычисления, необходимые для оптимальной торговой стратегии.
3. Улучшение управления рисками: Разработка моделей машинного обучения для определения и управления рисками в торговых операциях. Это может включать прогнозирование вероятности убыточных сделок, определение оптимального уровня стоп-лосса и т. д.

Используемая литература

1. Machine Learning for Algorithmic Trading – 2020г. Stephan J.
2. Практический анализ временных рядов -2021г. Aileen N.
3. Курс активного трейдера – 2019г. Герчик А.
4. Механизм трейдинга – 2016г. Мартынов А.

**Приложение А**

Модель данных AggregatedTradeVolume:

currency   
timestamp   
cost   
volume  
sellMaxQuantity  
sellAvgQuantity  
buyMaxQuantity  
buyAvgQuantity  
bidMaxQuantity  
bidMaxQCost  
bidAvgCost  
askMaxQuantity  
askMaxQCost  
askAvgCost

**Приложение Б**

Модель данных Candle:

currency   
openTimestamp   
closeTimestamp   
openCost   
closeCost   
minCost   
maxCost   
volatilityCost   
volatilityPercent   
askStandardDeviationCost   
askVarianceCost   
askAverageCost   
askStandardDeviationQuantity   
askVarianceQuantity   
askSumQuantity  
askAverageQuantity   
bidStandardDeviationCost   
bidVarianceCost   
bidAverageCost   
bidStandardDeviationQuantity   
bidVarianceQuantity   
bidSumQuantity   
bidAverageQuantity   
buyStandardDeviationCost   
buyVarianceCost   
buyAverageCost   
buyStandardDeviationQuantity   
buyVarianceQuantity   
buySumQuantity   
buyAverageQuantity   
sellStandardDeviationCost   
sellVarianceCost   
sellAverageCost   
sellStandardDeviationQuantity   
sellVarianceQuantity   
sellSumQuantity   
sellAverageQuantity