**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

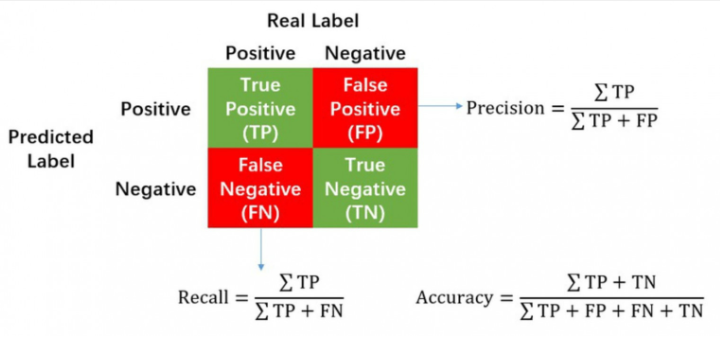
Данная работа имеет актуальность в сфере прогнозирования биржевых торгов, в соответствии с тем, что применение машинного обучения в связке с классическими методами трейдинга может привести к улучшению результатов торговли и повышению прибыли.

В результате будет предоставлены такие метрики качества, как: AIC, а для простых моделей будет приведен F1 Score.

* Формула AIC (Информационный критерий Акаике): 

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

* Формула меры точности теста F1 Score: 



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

Из аналогичных работ можно выделить работы по применению SSA для модели ARIMA, которая в будущем может использоваться для развития этого проекта. Работа содержит практические указания использования результатов SSA как параметры для обучающейся модели ARIME. Работа располагается по ссылке - <https://media.neliti.com/media/publications/323515-ssa-and-arima-for-forecasting-number-of-40972d4c.pdf>

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 реализовано 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]

Для агрегации торговых данных нам потребуются следующие математические методы: среднее, стандартное отклонение, дисперсия, максимум, минимум, сумма. Модуль агрегации данных располагается на github, по ссылке –

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

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

Полученный F1 Score 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 для применения, в качестве наилучших параметров, для таких моделей, как: ARIMA, GARCH, SARIMAX, и д.р. Реализация данного метода представлена по ссылке на github – <https://github.com/broom888/CryptoCurrencyTrading_public/blob/master/data_science/PCA.ipynb>

Таким образом, можно преступать к построению модели ARIMA, с предсказанием 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

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.

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, что полученные навыки, в ходе разработки и анализа, благотворно повлияют на дальнейшую разработку проекта, учитывая, что использование программных средств в инвестиционном процессе требует высокого уровня экспертизы разработчика.

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