Федеральное государственное бюджетное образовательное учреждение
высшего образования «Московский государственный университет имени
М.В.Ломоносова»
МЕХАНИКО-МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ
Кафедра Математической теории интеллектуальных систем

Курсовая работа

TODO

Выполнил:
студент 431 группы
Зенин В. О.

Научный руководитель:
к.ф.-м.н., н.с Половников В. С.

Москва - 2023

Оглавление

1. Введение

В настоящее время прогнозирование временных рядов является одной из наиболее актуальных задач в области анализа данных. Это связано с тем, что временные ряды могут отражать различные экономические, социальные и политические процессы, которые необходимо учитывать при принятии решений в различных сферах жизни. Существует множество методов прогнозирования временных рядов, каждый из которых имеет свои преимущества и недостатки.

Одним из методов прогнозирования временных рядов является использование AR (авторегрессионных) моделей [1]. AR модели позволяют учитывать прошлые значения временного ряда для предсказания его будущих значений. Авторегрессионный процесс задается следующим образом:

$$X_t = \beta_0 + \sum_{i=1}^{p} \beta_i X_{t-i} + \varepsilon_t,$$

где X_t – будущее значение, которое необходимо предсказать, β_i – параметры модели, p – порядок модели, независимые одинаково распределенные случайные величины $\varepsilon_t \sim N(0, 1)$.

Также широко распространён метод скользящего среднего (Moving Average, MA). Этот метод основан на усреднении значений временного ряда за определенный период времени. Модель скользящего среднего q-го порядка определяется как

$$X_t = \sum_{i=0}^{q} \beta_i \varepsilon_{t-i}$$

Для усреднения временного ряда кроме скользящего среднего можно использовать экспоненциальное сглаживание. Таким образом получается следующая модель [2]:

$$S_t = \begin{cases} C_t, t = 1 \\ C_{t-1} + \alpha \cdot (C_t - S_{t-1}) t > 1 \end{cases},$$

где S_t – сглаженный ряд, C_t – исходный ряд, α – коэффициент сглаживания, $\alpha > 0$ и обычно не превосходит 1 или 2.

Для рядов с выраженной сезонностью существует метод сезонной декомпозиции (Seasonal Decomposition) – он позволяет выделить сезонные компоненты из общего ряда, что даёт возможность прогнозировать поведение временного ряда в зависимости от сезонных факторов. В данном методе временной ряд делится на две составляющие: сезонную и трендовую. Сезонная составляющая представляет собой повторяющиеся колебания, связанные с сезонными факторами (например, сезонность продаж в розничной торговле). Трендовая составляющая отражает общую тенденцию развития ряда [3] [4].

Описанные ранее классические методы могут применятся как по-отдельности, так и вместе. Из последнего вытекают комбинированные модели ARMA, ARIMA, SARIMA. В качестве параметров необходимо задать факторы, которые будет использовать модель, например: значения временного ряда из прошлого, размер окна скользящего среднего, сдвиг для сезонности

(необходимо чтобы заданные факторы находились в одном и том же сезонном промежутке). Коэффициенты при заданных факторах вычисляются по методу наименьших квадратов [5].

Хорошо зарекомендовали себя методы на основе деревьев решений. Одним из преимуществ деревьев решений является их способность обрабатывать большие объемы данных и находить сложные зависимости между признаками и целевым значением. Они также могут быть легко интерпретированы и объяснены, что делает их полезными для задач прогнозирования временных рядов. Например, если имеются данные о продажах товаров в магазине за последние несколько лет, то можно использовать деревья решений для прогнозирования будущих продаж. Мы можем определить признаки, такие как цена товара, сезонность, количество конкурентов в районе и т.д., и использовать их для создания дерева решений. Каждый узел дерева будет принимать решение на основе значения признака, и на выходе получится прогноз будущих продаж.

Для алгоритмов на основе деревьев решений часто используется градиентный бустинг. Данных подход может быть описан следующим образом: построенное дерево имеет некоторую ошибку в своих предсказаниях, при наличии дифференцируемой функции ошибки можно определить поправочные значения, уменьшающие ошибку (градиент) и затем построить новое дерево, цель которого предсказать поправочные значения. Повторяя данную процедуру множество раз строится последовательность деревьев, в которой каждое новое дерево уточняет результат всех предыдущих [6].

Деревья решений не имеют представления о временной зависимости между наблюдениями. Чтобы сообщить им эту информацию необходимо закодировать время в признаках. Например, год, месяц, день недели, информация был ли день выходным или праздником – всё это может быть частью признаков, по которым будет строиться прогноз. Однако целевой признак, например такой как цена актива или величина продаж не должны включаться. В этом основное отличие от классических моделей, которые предсказывают целевую переменную основываясь на её же значениях в исторических данных.

Также существуют различные подходы на основе нейронных сетей, которые могут использоваться для работы с изменяющимися во времени данными. В процессе своего обучения они способны извлекать сложные нелинейные зависимости из данных и генерировать своё предсказание, основываясь на этом. В качестве таких данных можно использовать связанные с финансами временные ряды, например, цены различного рода активов.

Можно найти описание паттернов движения цены, полученные путём анализирования исторических данных биржевых котировок. Многие трейдеры используют их как основание для своих стратегий. Правила, образующиеся в результате найденных закономерностей, достаточно примитивны, как и сами паттерны. Если предположить, что кем-то найдена выгодная стратегия, то подобную способны найти и многие другие участники рынка, сводя на нет любую потенциальную выгоду. Вызывает интерес: способны ли нейронные сети находить паттерны и, тем самым, определять приносящие доход стратегии торговли, скрытые от большинства трейдеров.

Информация о классических финансовых инструментах во многом скрыта и хранится на биржах. Финансовые транзакции также скрыты за межбанковским обменом и не поддаются анализу. Однако существуют набирающие популярность криптофинансовые активы, инфор-

мация о которых, по своей природе, намного более открыта и может быть использована для анализа движения цены.

Цель данной работы – Исследование доступной публично информации о криптовалютах, построение нескольких архитектур нейронных сетей для анализа исторических данных и построение прогноза изменения будущей цены актива.

Основными задачами курсовой являются:

- Изучение существующих данных в блокчейне Bitcoin.
- Практическая реализация моделей на базе рекуррентных нейронных сетей и архитектуре трансформера.
- Постановка задач предсказания движения цены как задачи регрессии и классификации.
- Сравнение полученных результатов между собой и определение перспектив подобных исследований.

2. Данные и извлечение признаков

2.1. Информация из сети

В данной работе использованы дневные наблюдения о состоянии блокчейн сети Bitcoin с 10 мая 2020 года по 8 мая 2023 года, полученные с blockchain.com. Некоторые базовые признаки также вычислены заранее поставщиком данных. Их описания собраны в таблице 1.[1]

2.2. Подготовка данных

При работе с ценой актива часто используется логарифм цены,

$$\ln(\frac{x_t}{x_{t-1}})$$

позволяющий перейти от абсолютных значений к относительным. Смысл данного преобразования заключается в том, что успешная стратегия приносит доход в результате изменения цен, умноженных на вложенный капитал и именно доход имеет ключевое значение.

Входные данные для нейронных сетей следует скалировать. Однако некоторые признаки в наших данных имеют количественную природу, что выражается в почти линейном росте. Например, абсолютное значение добытых на момент времени t монет BTC. Больший смысл имеет изменение в добыче, так как оно потенциально способно дать сигнал о будущих движениях цены. Поэтому в нашем случае подобное преобразование уместно применить ко всем признакам.

2.3. Формирование обучающего множества

До логарифмирования имелось 1094 векторов, размерности 27 каждый. В результате преобразования наблюдение за первый день вырождается и остается 1093 вектора значений.

Для обучения использовались значения до 15 июня 2022 года. Для валидации – с 15 июня 2022 года по 20 января 2023 года. Для теста – с 20 января 2023 года по 8 мая 2023 года. Данные временные диапазоны выбраны чтобы обеспечить соотношение 70:20:10.

Целевой признак – market-price.

Сформируем из данных следующие пары:

$$(X_{[m;t]}, Y_t),$$

где $X_{[m;t]} = (x_{t-m}, x_{t-m-1}, ..., x_{t-2}, x_{t-1})$ – подпоследовательность длины m, x_t – вектор признаков для момента времени t, Y_t – значение целевого признака. Для задачи регрессии $Y_t = y_t$, где y_t логарифм цены. Для задачи классификации $Y_t = \begin{cases} 1, y_t > 0 \\ 0, y_t \leq 0 \end{cases}$

[1]Признаки, отмеченные (*) имеют не более 3 пропущенных значений, которые восстановлены линейной интерполяцией.

Список литературы

[1] James D. Hamilton. Time Series Analysis and Forecasting. Princeton University Press, 1994.

[2] Everette S. Gardner. Exponential smoothing: The state of the art. 01 Jan 1985-Journal of Forecasting (John Wiley & Sons, Ltd.)-Vol. 4, Iss: 1, pp 1-28

[3] Lovell, Michael C. Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis. Journal of the American Statistical Association, vol. 58, no. 304, 1963, pp. 993-1010.

[4] Robert Alan Yaffee, Monnie McGee. Introduction to Time Series Analysis and Forecasting: With Applications of SAS and SPSS. Academic press, 2000

[5] Hyndman, R.J., & Athanasopoulos, G. (2021) Forecasting: principles and practice, 3rd edition, OTexts: Melbourne, Australia.

[6] Panarese, A.; Settanni, G.; Vitti, V.; Galiano, A. Developing and Preliminary Testing of a Machine Learning-Based Platform for Sales Forecasting Using a Gradient Boosting Approach. Appl. Sci. 2022, 12, 11054.

[7] U Thissen, R van Brakel, A.P de Weijer, W.J Melssen, L.M.C Buydens. Using support vector machines for time series prediction. Chemometrics and Intelligent Laboratory Systems, Vol. 69, Iss: 1-2, 2003, pp 35-49.

Таблица 1: Признаки из блокчейн сети Bitcoin

Признак	**Описание**
total-bitcoins (*)	Количество добытых монет
market-price	Средняя цена в USD на крупнейших обменниках
trade-volume	Объем обменянных BTC (USD)
blocks-size	Размер сети блокчейна (MB)
avg-block-size	Средний размер блока (MB)
n-transactions-total	Количество транзакций
n-transactions-per-block	Среднее число транзакций на блок
n-payments-per-block	Среднее число наград за валидированный блок
median-confirmation-time	Медианное время, за которое обработанная транзакция добавляется к сети
avg-confirmation-time	Среднее время, за которое обработанная транзакция добавляется к сети
hash-rate	Мощность сети
difficulty	Относительная мера сложности сети – насколько трудно валидировать очередной блок
transaction-fees	Выплаченные BTC за валидацию блоков
transaction-fees-usd	Выплаченные USD за валидацию блоков
fees-usd-per-transaction	Среднея выплата в USD за валидированную транзакцию
cost-per-transaction	Общий доход майнеров, разделённый на количество транзакций
n-unique-addresses (*)	Количество уникальных адресов, используемых в сети
n-transactions	Количество подтвержённых транзакций за день
n-payments	Количество подтвержённых выплат за день
mempool-count	Количество неподтверждённых транзакций
mempool-growth	Рост хранилищая неподтверждённых транзакций
mempool-size	Размер хранилища неподтверждённых транзакций
n-transactions-excluding-popular	Количество транзакций, за исключением 100 самых популярных адресов
estimated-transaction-volume (*)	Оценочная стоимость транзакций (BTC)
estimated-transaction-volume-usd (*)	Оценочная стоимость транзакций (USD)