

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра теоретических основ
компьютерной безопасности и
криптографии

Прогнозирование торгового трафика нейросетевыми методами

АВТОРЕФЕРАТ

дипломной работы

студентки 6 курса 631 группы
специальности 10.05.01 Компьютерная безопасность
факультета компьютерных наук и информационных технологий
Князевой Марины Аркадьевны

Научный руководитель

доцент

И. И. Слеповичев

23.01.2021 г.

Заведующий кафедрой

д. ф.-м. н., доцент

М. Б. Абросимов

23.01.2021 г.

Саратов 2021

ВВЕДЕНИЕ

Актуальность. Задача прогнозирования является действительно актуальной, поскольку помогает при решении сразу нескольких задач:

- 1) позволяет оценить тенденции изменения того или иного фактора;
- 2) позволяет провести анализ рынков;
- 3) уменьшать риск принятия неверных, необоснованных, субъективных решений.

Сейчас в этой области активно ведутся разработки. Например, исследование нигерийского фондового рынка, статья для анализа которого указана в списке использованных источников.

Помимо областей экономики можно определить задачу прогнозирования и в других не менее важных сферах жизни человека. Решение этой задачи может способствовать наилучшему принятию решения в определённых ситуациях.

Цели и задачи. Целью работы является проверка возможности прогнозирования суммарного торгового трафика импорта для некоторых крупных стран нейросетевыми методами. Разработка новой модели и соответствующего ей метода прогнозирования.

Для этого необходимо решить следующие задачи:

1. проанализировать предметную область;
2. осуществить обзор моделей и методов прогнозирования временных рядов, выявить достоинства и недостатки каждого класса моделей;
3. выявить наиболее эффективные классы моделей прогнозирования, определить перспективные подходы, которые могут позволить устранить недостатки авторегрессионного класса моделей;
4. разработать собственную модель для прогнозирования на основе нейронных сетей; выполнить программную реализацию алгоритмов;
5. оценить эффективность предложенной модели прогнозирования при решении задачи прогнозирования различных временных рядов.

Объект и предмет исследования. Объектом исследования работы являются временные ряды. Предметом исследования является модель, которая построена нейросетевыми методами для прогнозирования временного ряда.

Метод исследования. Для решения поставленных задач в работе использованы методы математического моделирования, анализ временных рядов, регрессионный анализ, методы объектно-ориентированного программирования.

Результаты исследования. В ходе исследования были получены результаты:

1. выявлены недостатки моделей прогнозирования;
2. разработана модель и метод прогнозирования временных рядов для торгового трафика;
3. произведена оценка построенной модели с помощью метрик качества, выявлена область её применимости.

Структура работы. Дипломная работа состоит из введения, трёх разделов, заключения, списка использованных источников и восьми приложений.

В первой главе описаны основные понятия и определения: временные ряды, сезонность, цикличность, тренды.

Во второй главе рассмотрена классификация моделей прогнозирования, описаны статические и структурные модели прогнозирования, рассмотрены их недостатки и достоинства.

В третьей главе описана программная реализация нейросетевой модели, которая была сделана с использованием библиотек tensorflow/keras, языка программирования Python. Также в этой главе проанализирована область применения реализованной нейросетевой модели.

Общий объём работы – 111 страниц, из них 51 страница – основное содержание, включая 80 рисунков и 2 таблицы, список использованных источников из 36 наименований.

КРАТКОЕ СОДЕРЖАНИЕ

1 Основные понятия и определения

В этом разделе описываются основные термины, касающиеся временных рядов и которые потребуются для решения поставленной задачи прогнозирования.

Пусть дан временной ряд $y_0, y_1, \dots, y_t, \dots$, где $y_i \in \mathbb{R}^n$ – элемент временного ряда, то модель временного ряда будет определяться формулой:

$$\hat{y}_{t+d}(w) = \mathcal{F}(y_1, \dots, y_t; w),$$

где $d = 1, \dots, D$, D – горизонт прогнозирования, w – вектор параметров модели.

Особое внимание стоит уделить стационарным временным рядам, у которых распределение значений является инвариантным относительно момента времени, для которого оно построено.

Такие ряды характеризуются трендами, сезонной и циклической компонентой.

Необходимо разграничить также два понятия модель и метод прогнозирования. Метод прогнозирования – последовательность действий, которые нужно совершить для получения модели прогнозирования временного ряда. Модель прогнозирования – функциональное представление, которое адекватно описывает временной ряд и является основой для получения будущих значений процесса.

2 Классификация моделей временных рядов

В статистических моделях зависимость будущего значения от прошлого задается в виде некоторого уравнения. К ним относятся:

1. регрессионные модели;
2. авторегрессионные модели (ARIMAX, GARCH, ARDLN);
3. модель экспоненциального сглаживания.

В структурных моделях зависимость будущего значения от прошлого задается в виде некоторой структуры и правил перехода по ней. К ним относятся:

1. нейросетевые модели;

2. модели на базе цепей Маркова;
3. модели на базе классификационно-регрессионных деревьев.

Далее более подробно в подразделах 2 раздела рассматриваются соответствующие структурные и статические модели.

Регрессионный анализ используется для прогноза, анализа временных рядов, тестирования гипотез и выявления скрытых взаимосвязей в данных.

Линейная регрессия – одна из важнейших и широко используемых техник регрессии. Одним из его достоинств является лёгкость интерпретации результатов. Под многомерной регрессией понимают линейную регрессию в n -мерном пространстве, где объекты и признаки являются n -мерными векторами.

Для точности оценки регрессии как линейной, так и многомерной модели можно воспользоваться методом наименьших квадратов.

Этот метод заключается в минимизации квадрата евклидова расстояния между двумя векторами – вектором восстановленных значений зависимой переменной и вектором фактических значений зависимой переменной.

Нелинейная регрессия – это способ нахождения нелинейной модели взаимосвязи между зависимой переменной и набором независимых переменных.

Адаптивные методы прогнозирования временных рядов представляют из себя методы, цель которых заключается в построении самокорректирующихся (самонастраивающихся) экономико-математических моделей, которые способны отражать изменяющиеся во времени условия, учитывать информационную ценность различных членов временной последовательности и давать достаточно точные оценки будущих членов данного ряда. Такие модели предназначаются, прежде всего, для краткосрочного прогнозирования.

Простейшие адаптивные модели:

1. экспоненциальное сглаживание, модель Брауна;
2. модели линейного роста, которые включают в себя Модель Хольта, модель линейного роста Брауна, модель прогнозирования Дж. Бокса и Г. Дженкинса;
3. сезонные модели: модель Хольта-Уинтерса, модель Тейла-Вейджа.

Главное достоинство прогнозной модели Брауна состоит в том, что она способна последовательно адаптироваться к новому уровню процесса без значительного реагирования на случайные отклонения.

Недостаток: экспоненциальная средняя дает систематическую ошибку, когда временной ряд имеет тенденцию линейного роста

Модель Хольта-Уинтерса учитывает мультипликативный тренд и сезонность, и модель Тейла-Вейджа, в свою очередь, берёт во внимание аддитивный тренд и сезонность.

На основе модели Хольта Уинтерс создал свою прогностическую модель, которая способна учитывать экспоненциальный тренд и аддитивную сезонность.

В модели Хольта также необходимо решить задачу прогнозирования временного ряда. При существовании линейного тренда на данных временного ряда использование модели Брауна не имеет смысла. Для учёта влияния линейного тренда используют модель Хольта. Однако эта модель не учитывает сезонность.

Далее рассматриваются авторегрессионные модели:

- ARIMAX (autoregression integrated moving average extended);
- GARCH (generalized autoregressive conditional heteroskedasticity);
- ARDLN (autoregression distributed lag model).

Авторегрессионная (AR) модель – модель временных рядов, в которой значения временного ряда в данный момент линейно зависят от предыдущих значений этого же ряда. Авторегрессионный процесс порядка p (AR(p)-процесс) определяется следующим образом:

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

где $a_1, a_2, \dots, a_p$ – параметры модели (коэффициенты авторегрессии), c – постоянная (часто для упрощения предполагается равной нулю), а ε_t – белый шум.

ARIMAX-модель (autoregression integrated moving average extended) – математическую модель для анализа временных рядов, которая объединяет в

себе интегрированную авторегрессию, скользящее среднее и возможность учёта дополнительных внешних факторов.

Также используется модель скользящего среднего порядка q :

$$Z(t) = \frac{1}{q} (Z(t-1) + Z(t-2) + \dots + Z(t-q)) + \varepsilon_t,$$

где q – порядок скользящего среднего, ε_t – ошибка прогнозирования.

ARMA(p,q) – объединение в одной модели авторегрессии и скользящего среднего. Если в качестве входных данных используются не сами значения временного ряда, а их разность q -того порядка, то модель носит название авторегрессии проинтегрированного скользящего среднего. Такую модель называют ARIMA(p,d,q). Развитием модели ARIMA(p,d,q) является модель ARIMAX(p,d,q), которая записывается уравнением:

$$Z(t) = AR(p) + \alpha_1 X_1(t) + \dots + \alpha_s X_s(t),$$

где $\alpha_1, \dots, \alpha_s$ – коэффициенты внешних факторов $X_1(t), \dots, X_s(t)$. [10]

GARCH-модель (generalized autoregressive conditional heteroskedasticity) – модель, которая чаще всего используется для прогнозирования ситуации на финансовых рынках в условиях нестабильности. Для таких ситуаций характерно такое понятие, как дисперсия на различных интервалах наблюдения, которую также называют гетероскедастичностью. Расчёт дисперсии для GARCH-модели:

$$\sigma^2(t) = a + \sum_{i=1}^a b_i r_{t-1}^2 + \sum_{i=1}^p c_i \sigma_{t-1}^2,$$

где p – количество предшествующих оценок, которые влияют на текущее значение, c_i – весовые коэффициенты, отражающие степень влияния предыдущих оценок на текущие значения.

ALRDM-модель (autoregression distributed lag model) – авторегрессионная модель с распределённым лагом. Часто при моделировании процессов на изучаемую переменную влияют и текущие значения процесса, и его лаги, то есть значения временного ряда, предшествующие изучаемому моменту времени. ALRDM-модель описывается уравнением:

$$Z(t) = \phi_0 + \phi_1 Z(t-l-1) + \dots + \phi_p Z(t-l-p) + \varepsilon_t,$$

где $\phi_0, \dots, \phi_p$ – коэффициенты, l – величина лага. Модель обозначают $ALRDM(p, l)$, чаще всего применяют при экономических процессах.

В подразделе 2.4 можно видеть определения, касающиеся нейронных сетей, которые потребуются для решения поставленных задач.

Нейронная сеть – это математическая модель, а также её программное или аппаратное воплощение, которая построена по принципу организации и функционирования биологических нейронных сетей, которые, в свою очередь определяются как сети нервных клеток живого организма.

В общем случае выделяют три фундаментальных класса нейросетевых архитектур:

- 1) однослойные сети прямого распространения;
- 2) многослойные сети прямого распространения;
- 3) рекуррентные сети.

Сеть прямого распространения – сеть, в которой существует входной слой узлов источника, информация от которого передаётся на выходной слой нейронов.

Сети с обратными связями – искусственные нейронные сети, в которых выход нейрона может вновь подаваться на его вход.

Однослойные сети прямого распространения – класс нейронных сетей, у которых связи между нейронами не образуют замкнутые циклы.

Многослойные сети прямого распространения – класс нейронных сетей, характеризующийся существованием одного или нескольких скрытых слоёв.

Рекуррентная нейронная сеть – класс нейронных сетей, где связи между элементами образуют направленную последовательность и существует, по крайней мере, одна обратная связь.

По литературным данным в исследовании нейронных сетей рекуррентные нейронные сети с архитектурой LSTM являются наилучшим вариантом для предсказания временных рядов по нескольким причинам. В отличие от традиционных RNN, LSTM специально разработаны для устранения проблемы долгосрочной зависимости. Основная особенность LSTM – запоминание

информации в течение длительных периодов времени, отсюда следует, что они практически не требуют обучения. Это достигается за счет того, что повторяющийся модуль модели имеет комбинацию четырех слоев, взаимодействующих друг с другом.

В подразделе 2.5 и 2.6 определены модели на базе цепей Маркова и дерева CART.

Последовательность случайных переменных $X_1, \dots, X_n, X_{n+1}$ принимает форму цепи Маркова, если вероятность нахождения системы в состоянии x_{n+1} в некоторый момент времени $(n + 1)$ зависит исключительно от вероятности нахождения системы в момент времени n в состоянии x_n .

Можно сделать вывод о том, что цепи Маркова имеет смысл рассматривать как порождающие модели, которые состоят из множества попарно связанных друг с другом состояний. Каждый раз, когда система переходит в конкретное состояние, именно с ним ассоциируется выход системы.

Модели прогнозирования на основе цепей Маркова предполагают, что будущее состояние процесса зависит только от его текущего состояния. Процессы, которые моделируются цепями Маркова, относятся к процессам с короткой памятью.

CART (Classification and Regression Tree) представляет из себя алгоритм бинарного дерева решений, который предназначен для решения задач классификации и регрессии.

По модели CART прогнозируемое значение временного ряда зависит от предыдущих значений и некоторых независимых переменных.

3 Практическая часть

Из источника были предварительно скачаны данные, таблица MBSComtrade.csv. В файле находятся данные, характеризующие суммарный торговый трафик по месяцам, указанный в долларах, для различных стран с 1957 года по 2020 год. Таблица состоит из 25 столбцов (параметров).

Программа была написана на высокоуровневом языке программирования Python (версия 3.8.0). Для реализации программы на языке Python была

использована отдельная, виртуальная среда Anaconda, выступающая в качестве платформы для обработки данных. Jupyter Notebook выступает в качестве инструмента для интерактивной разработки и представления проектов в области Data Science

В программе используются следующие библиотеки и модули: matplotlib (3.2.1), tensorflow (2.2.0), pandas (1.0.3), scikit-learn (0.22.1), numpy (1.18.1).

В результате работы ноутбука LSTM_test_all_c.ipynb создавались папки для сохранения:

- 1) графиков, построенных для начальных данных из таблицы;
- 2) для сохранения графиков прогнозов;
- 3) для фиксации изменения метрик качества (MSE и MAE).

Также в результате работы ноутбука LSTM_test_all_c.ipynb были созданы два текстовых файла, которые включают в себя три оценки модели: значение функции потерь (MSE) и средней абсолютной ошибки (MAE).

Для выполнения нормализации полученных данных использовали StandartScaler. Далее происходит формирование тренировочной train выборки и тестовой test выборки.

Далее будет описана архитектура модели. Основной слой – блок из 64 ячеек с архитектурой LSTM. Слой Dropout был добавлен для регуляризации (с долей обнуляемых весов 0,3). Dense слой из 64 нейронов был использован для более аккуратного предсказания сложной зависимости.

В качестве функции активации был использован гиперболический тангенс (tanh). Рекуррентная функция активации в LSTM блоке – сигмоида.

Для обучения сети был выбран метод обратного распространения ошибки. Этот метод используется при обновлении весов в нейронной сети. В результате из каждой временной точки тренировочной выборки построен прогноз на три временные точки вперед для 16 стран с «качественными» данными. Область применимости моделей ограничена временными рядами, которые имеют годовую сезонность. Наличие единого тренда в данных ожидаемо повышает точность, аномальные изломы в тренде ухудшают качество прогноза.

ЗАКЛЮЧЕНИЕ

На сегодняшний день разработано множество моделей для решения задачи прогнозирования временного ряда, наибольшую применимость среди которых имеют авторегрессионные и нейросетевые модели.

В результате выполнения работы была спроектирована и реализована нейросетевая модель для прогнозирования торгового трафика. Также были рассмотрены понятия и определения из области нейронных сетей, проанализированы достоинства и недостатки моделей и методов прогнозирования временных рядов.

Для достижения цели были пройдены основные этапы проектирования от изучения предметной области до апробации модели.

В ходе работы был сделан вывод о том, что область применимости созданной модели ограничена временными рядами, которые имеют годовую сезонность. Также была получена оценка модели с помощью двух метрик качества, полученные результаты представлены в двух таблицах в приложениях.

При существовании более объёмной выборки, имеющую сезонность и тренд, можно было бы сравнить результаты с полученными результатами в ходе выполнения практической части работы, построенная модель является рабочей.