**ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ**

**НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**

**«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

Факультет компьютерных наук

Образовательная программа «Программная инженерия»

УДК 004.942

|  |  |  |
| --- | --- | --- |
| СОГЛАСОВАНО |  | УТВЕРЖДАЮ |
|  |  |  |
| Приглашенный преподаватель департамента больших данных и информационного поиска факультета компьютерных наук,  кандидат физико-математических наук |  | Академический руководитель образовательной программы «Программная инженерия»  профессор департамента программной инженерии, канд. техн. наук |
| (С.А. Буденный) |  | (В.В. Шилов) |
| «07» апреля 2023 г. |  | «07» апреля 2023 г. |

**Отчет**

**по исследовательскому курсовому проекту**

на тему Оценка нелинейной связности в динамике рыночной капитализации публичных компаний в задаче определения синтетического рыночного индекса

по направлению подготовки бакалавров 09.03.04 «Программная инженерия»

|  |
| --- |
| Выполнил  студент группы БПИ201  образовательной программы  09.03.04 «Программная инженерия» |
| Насыхова Анастасия Артемовна |
| «07» апреля 2023 г. |

Москва 2023

**РЕФЕРАТ**

Отчет 42 с., 47 рис., 7 табл., 7 источн., 1 прил.

Ключевые слова: Кластеризация временных рядов, метрика сходства временных рядов, прогнозирование цен на акции, создание синтетического индекса.

Изучение динамики рыночной капитализации публичных компаний имеет большое значение в финансах и экономике. Это позволяет анализировать и прогнозировать поведение финансовых рынков и инвестиций, а также выявлять тенденции и закономерности на фондовом рынке. В этой научной статье было проведено исследование возможности использования иерархической кластеризации на основе DTW (Dynamic Time Warping) для улучшения прогнозов модели TFT (Temporal Fusion Transformer) для финансовых временных рядов. Для оценки качества прогнозов использовалась метрика МАРЕ (средняя абсолютная относительная ошибка). Экспериментальные результаты показали, что использование временных рядов из одного кластера при обучении модели позволяет добиться более точных прогнозов по сравнению с моделью, обученной только на одном временном ряду. Исследование вносит ценный вклад в область анализа временных рядов и прогностического моделирования, представляя новый метод повышения точности прогностических моделей.

Объект исследования: нелинейные связности в динамике рыночной капитализации публичных компаний. Предмет исследования: кластеризация временных рядов на основе их сходства по структуре. Под сходством подразумевается задача поиска последовательностей с одинаковыми законами изменения.

Цель проекта: проверка гипотезы о том, что качество предсказания временного ряда одной компании можно улучшить за счет использования схожих временных рядов других компаний.

Метод или методология проведения работы: в процессе работы будет проведен ряд экспериментов, целью которых является анализ способов расчета расстояний между временными рядами и алгоритмов кластеризации. Усредненные результаты станут основой для создания синтетического индекса, который будет участвовать в исследовании способов его применения. Результаты проекта:

1. кластеры, полученные в результате проведения экспериментов;
2. новый метод улучшения качества прогнозирования временного ряда.

Апробация результатов: публикация научной статьи.

СОДЕРЖАНИЕ

[1. ОСНОВНЫЕ ТЕРМИНЫ, ОПРЕДЕЛЕНИЯ И СОКРАЩЕНИЯ 4](#_Toc132148679)

[2. ВВЕДЕНИЕ 5](#_Toc132148680)

[3. ОБЗОР И АНАЛИЗ ИСТОЧНИКОВ 6](#_Toc132148681)

[3.1. РАССТОЯНИЕ МЕЖДУ ВРЕМЕННЫМИ РЯДАМИ 6](#_Toc132148682)

[3.2. КЛАСТЕРИЗАЦИЯ 6](#_Toc132148683)

[3.3. ПРОГНОЗИРОВАНИЕ ЦЕН НА АКЦИИ 7](#_Toc132148684)

[4. МЕТОДОЛГОИЯ 9](#_Toc132148685)

[4.1. АЛГОРИТМ КЛАСТЕРИЗАЦИИ 9](#_Toc132148686)

[4.1.1. ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ 9](#_Toc132148687)

[4.1.2. K-MEANS 9](#_Toc132148688)

[4.2. МОДЕЛЬ ПРОГНОЗИРОВАНИЯ 10](#_Toc132148689)

[4.3. ПОСТАНОВКА ЗАДАЧИ 11](#_Toc132148690)

[4.4. ПРЕДЛАГАЕМЫЙ ПОДХОД 11](#_Toc132148691)

[5. ЭКСПЕРИМЕНТЫ 13](#_Toc132148692)

[5.1. ДАТАСЕТ 13](#_Toc132148693)

[5.2. МЕТРИКИ 14](#_Toc132148694)

[5.2.1. СХОДСТВО ВРЕМЕННЫХ РЯДОВ 14](#_Toc132148695)

[5.2.2. КАЧЕСТВО ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ 15](#_Toc132148696)

[5.3. РЕЗУЛЬТАТЫ 17](#_Toc132148697)

[5.3.1. КЛАСТЕРИЗАЦИЯ ВРЕМЕННЫХ РЯДОВ 17](#_Toc132148698)

[5.3.1.1. ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ 17](#_Toc132148699)

[5.3.1.2. K-MEANS 18](#_Toc132148700)

[5.3.1.3. FEATURE BASED 20](#_Toc132148701)

[5.3.1.4. ПОДВЕДЕНИЕ ИТОГОВ ЭТАПА 21](#_Toc132148702)

[5.3.2. СИНТЕТИЧЕСКИЙ ИНДЕКС 25](#_Toc132148703)

[5.3.3. ПРОГНОЗИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ 26](#_Toc132148704)

[6. ЗАКЛЮЧЕНИЕ 40](#_Toc132148705)

[7. СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 41](#_Toc132148706)

[ПРИЛОЖЕНИЕ А 42](#_Toc132148707)

# ОСНОВНЫЕ ТЕРМИНЫ, ОПРЕДЕЛЕНИЯ И СОКРАЩЕНИЯ

В настоящем отчете о НИР применяют термины с соответствующими определениями, указанными в таблице 1.

Таблица 1 – Основные термины и их определения

|  |  |
| --- | --- |
| Термин | Определение |
| Dynamic Time Warping (DTW) | Алгоритм, позволяющий найти оптимальное соответствие между временными последовательностями. |
| Временной ряд | Собранный в разные моменты времени статистический материал о значении каких-либо параметров исследуемого процесса. |
| Датасет | Обработанный и структурированный набор данных. |
| Кластеризация | Процесс группировки множества физических или абстрактных объектов в похожие классы. |
| Научно-исследовательская работа (НИР) | Работа научного характера, связанная с научным поиском, проведением исследований, экспериментами в целях расширения имеющихся и получения новых знаний, проверки научных гипотез, установления закономерностей, проявляющихся в природе и в обществе, научных обобщений, научного обоснования проектов. |
| Рыночная капитализация публичных компаний | Текущая оценочная стоимость всех акций компании, находящихся в публичном обращении. |
| Тикер | Краткое название компании в биржевой информации котируемых инструментов. |
| K-Means | Алгоритм кластеризации, подробное описание алгоритма можно найти в разделе Методология. |

# ВВЕДЕНИЕ

Выявление взаимосвязей между компаниями важно для понимания того, как изменения в одной компании могут повлиять на другие компании отрасли или на рынок в целом. Понимание этих взаимосвязей может помочь принимать более обоснованные инвестиционные решения, поскольку инвесторы могут лучше предсказывать, как поведут себя рынки в будущем. В связи с этим растет интерес к углубленному изучению взаимоотношений между компаниями, в частности, для выявления нелинейных взаимосвязей, которые характеризуются сложными закономерностями и требуют использования специальных методов анализа. В данной статье представлены результаты исследования нелинейных взаимосвязей между компаниями и обсуждаются их возможные практические применения.

Цель исследования – проверка гипотезы о том, что качество предсказания временного ряда одной компании можно улучшить за счет использования схожих временных рядов других компаний.

Задачи проекта:

1. Анализ способов расчета расстояний между временными рядами;
2. Создание датасета;
3. Определение способов кластеризации;
4. Создание синтетического индекса;
5. Применение синтетического индекса для улучшения модели предсказания цен на акции.

# ОБЗОР И АНАЛИЗ ИСТОЧНИКОВ

При работе с временными рядами возникает множество стандартных трудностей машинного обучения: высокая размерность входных данных, корреляции в данных, пропущенные значения. Однако в задаче кластеризации последовательностей существует пара дополнительных трудностей. Во-первых, в строках может быть разное количество отсчетов. Во-вторых, при работе с последовательными данными у нас появляется больше свободы в определении сходства одного объекта с другим, поскольку способ изменения данных с течением времени также вносит свой вклад.

## РАССТОЯНИЕ МЕЖДУ ВРЕМЕННЫМИ РЯДАМИ

**Time series similarity measures and time series indexing [1].** В этой статье рассматриваются способы вычисления расстояний между временными рядами. У большинства из них есть ключевой недостаток: метрика не учитывает сдвиг в данных. Это означает, что если в одном временном ряду есть последовательность из другого ряда, похожая по структуре, но расположенная в другом временном интервале, метрика будет считать эти ряды разными. И поскольку важно найти общую закономерность из-за поведения ряда для задачи определения синтетического индекса такие показатели использовать нельзя. К счастью, в статье также описывается метрика, которая решает эту проблему – DTW (Dynamic Time Warping).

**Recent Techniques of Clustering of Time Series Data: A Survey [2].** В более поздней статье авторы исследуют последние разработки в области кластеризации временных рядов. Самое главное, что исследователи уделяют особое внимание корреляции показателей расстояния с алгоритмами кластеризации. В статье представлены различные подходы к работа с временными рядами. Чаще всего для финансовых данных в качестве показателя выбирается простое евклидово расстояние и модифицируется. В качестве алгоритма используется K-Means значение. Под ним авторы подразумевают базовый алгоритм, адаптированный для работы с последовательными данными.

## КЛАСТЕРИЗАЦИЯ

**Clustering Time Series Data — A Survey [3].** В статье рассматриваются основные подходы к кластеризации временных рядов. В целом, кластеризацию временных рядов можно разделить на 2 типа:

* Feature based – подход, основанный на характеристиках, при котором объекты группируются на основе статистических данных, собранных из временных рядов;
* Raw data – подход, основанный на необработанных данных, при котором кластеризация применяется к векторам временных рядов без каких-либо преобразований.

Оба метода хорошо зарекомендовали себя при работе с финансовыми данными, но первый подход используется редко, поскольку он не учитывает скрытые связи между временными рядами.

**Time-series clustering - A decade review [4].** Авторы представляют всеобъемлющий обзор исследований по кластеризации временных рядов за последнее десятилетие. Они сосредоточены на методах кластеризации данных из разных областей. Они подтверждают результаты ранее проведенных исследований о том, что стандартные алгоритмы хорошо подходят для финансовых временных рядов. Интересной особенностью этой статьи является то, что помимо успешных решений, исследователи приводят ряд неудачных экспериментов, а также рассказывают о возможных причинах неудач. Наиболее распространенными причинами являются особенности последовательных данных: высокая размерность, высокий уровень шума в данных, сильные корреляции между значениями.

## ПРОГНОЗИРОВАНИЕ ЦЕН НА АКЦИИ

Многие проблемы прогнозирования связаны с анализом временых рядов. В нашем случае данные временного ряда могут быть определены как хронологическая последовательность наблюдений за ежедневной ценой закрытия акций. Большинство способов решения этой проблемы можно разделить на два класса: линейные и нелинейные модели.

**Stock price prediction using LSTM, RNN and CNN-sliding window model [5].** В 2017 году исследователи из Индии применили самый популярный линейный алгоритм и несколько различных архитектур нейронных сетей для прогнозирования цен на акции некоторых индийских компаний. Проведя серию экспериментов, они показали, что алгоритмы глубокого обучения способны выявлять скрытые закономерности и лежащую в их основе динамику в данных с помощью процесса самообучения, что приводит к лучшему прогнозированию временных рядов.

В последнее время в области временной обработки данных наблюдается растущий интерес к использованию рекуррентных нейронных сетей (RNN) и моделей-трансформеров в сочетании с методами слияния для моделирования временных зависимостей. В рамках этого направления была предложена новая модель, получившая название «Temporal Fusion Transformer Model (TFTM)», в которой информация из нескольких источников эффективно комбинируется для прогнозирования временных рядов.

**Temporal Fusion Transformers for interpretable multi-horizon time series forecasting [6].** Авторы статьи представляют модель TFTM для прогнозирования многомерных временных рядов. TFTM – это гибридная модель, которая сочетает в себе функциональность трансформеров и методы объединения, такие как стробированное слияние и декомпозиция временных рядов на тренд, сезонность и шум. Чтобы протестировать модель, были проведены эксперименты с несколькими наборами данных, включая данные потребления энергии, цены на акции и количество пассажиров на воздушном транспорте. Результаты показали, что TFTM превосходит стандартные модели прогнозирования временных рядов, такие как ARIMA, LSTM и GRU.

# МЕТОДОЛГОИЯ

В этом разделе описывается общая методология исследования, включая алгоритмы кластеризации и модель прогнозирования.

## АЛГОРИТМ КЛАСТЕРИЗАЦИИ

Кластеризация – это процесс группировки сходных временных рядов вместе для выявления лежащих в их основе закономерностей и структур, в данном исследовании используется алгоритм иерархической кластеризации.

## ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

Иерархическая кластеризация – один из наиболее распространенных алгоритмов кластеризации временных рядов. Этот алгоритм рассматривает все временные ряды как набор точек в n-мерном пространстве и строит иерархическую кластерную структуру, где каждый кластер состоит из более похожих рядов данных.

Алгоритм кластеризации можно разделить на два этапа:

* Cоздание матрицы расстояний. Первый этап заключается в создании матрицы расстояний между временными рядами. Расстояние между временными рядами может быть определено различными способами.
* Кластеризация. Каждый временной ряд образует свой собственный кластер. Затем алгоритм находит наименьшее расстояние между всеми возможными парами кластеров и объединяет их в новый кластер. Этот процесс продолжается до тех пор, пока все временные ряды не будут объединены в один кластер или пока не будет достигнуто указанное пользователем количество кластеров.

## K-MEANS

K-Means – это алгоритм кластеризации, используемый для разделения набора объектов на заданное количество кластеров. Он основан на минимизации суммы квадратов расстояний между объектами и центрами кластеров. Чтобы применить алгоритм к временным рядам, нужно выбрать метрику, по которой будут оцениваться расстояния между рядами. Обычно используются евклидово расстояние или DTW.

После инициализации начальных центроид k кластеров алгоритм выполняет следующие шаги:

1. Каждый временной ряд относиться к ближайшему по расстоянию до его центроиды кластеру;
2. Для каждого кластера вычисляется новый центр тяжести как среднее значение всех строк в этом кластере;
3. Шаги 1 и 2 повторяются до тех пор, пока центроиды не перестанут меняться или не будет достигнуто максимальное количество итераций.

Одним из основных параметров алгоритма является выбор начальных центроидов. Хорошей практикой является использование случайных точек из общего набора данных, а также проведение нескольких запусков с разными начальными центроидами, чтобы более точно определить оптимальное количество кластеров и избежать попадания в локальный оптимум.

Ограничения K-Means заключаются в необходимости явного указания количества кластеров, в чувствительности к выбросам и неоднородности кластеров по размеру. Однако при правильном выборе параметров и понимании особенностей данных K-Means может быть эффективным инструментом для кластеризации временных рядов.

* + 1. **FEATURE BASED ПОДХОД**

Feature based подход к кластеризации временных рядов заключается в использовании набора признаков, описывающих каждый временной ряд. В дальнейшем полученные статистики кластеризуются обычным K-means алгоритмом, как простые табличные данные.

## МОДЕЛЬ ПРОГНОЗИРОВАНИЯ

После кластеризации временных рядов для прогнозирования временных рядов использовалась модель Temporal Fusion Transformer (TFT) из пакета darts.

Temporal Fusion Transformer (TFT) – это attention-based архитектура нейронной сети, которая может включать множество источников информации, таких как временные характеристики и события календаря, в процесс прогнозирования временных рядов. Основная идея заключается в том, что TFT предсказывает значения временных рядов в разные моменты времени одновременно, принимая во внимание контекст, используя метод преобразования в качестве основы для архитектуры модели.

Особенностью TFT является то, что он использует гибридную архитектуру нейронной сети, которая сочетает в себе сверточные, рекуррентные и трансформирующие блоки. На вход модели поступает временной ряд, который затем преобразуется в многомерный контекстный вектор, содержащий информацию о последовательности сигналов временного ряда.

Далее осуществляется многомерное декодирование, которое осуществляется с использованием блоков-трансформеров, причем каждый блок принимает в качестве входных данных не только выходные данные предыдущего блока, но и вектор контекста. Это позволяет модели учитывать контекст при формировании прогнозов для каждого момента времени в будущем и использовать метрики для оценки прогнозов на всех временных этапах для оптимизации.

TFT использует дополнительные механизмы для улучшения качества прогнозов, включая авторегрессию и многозадачное обучение. Авторегрессионное обучение используется для улучшения генерации прогнозов для различных временных интервалов, а многозадачное обучение позволяет модели справляться с одновременным прогнозированием нескольких временных рядов, используя общую информацию о данных.

Таким образом, TFT-модель является эффективным, гибким и мощным инструментом для прогнозирования временных рядов. Она продемонстрировала высокое качество прогнозов на нескольких наборах реальных данных, а также может быть применена в различных областях, включая экономику, финансы, медицину и транспорт.

## ПОСТАНОВКА ЗАДАЧИ

Существует гипотеза о том, что поведение временных рядов цен на акции может быть одинаковым для разных компаний на разных временных интервалах. Если гипотеза верна, использование информации о другой компании может улучшить прогнозирование цен рассматриваемой компании. Ключевой целью работы является проверка этой гипотезы.

Цель будет достигнута с использованием кластерного подхода, который позволит эффективно группировать схожие данные финансовых временных рядов и использовать их для создания синтетических индексов, которые, как ожидается, будут отражать уникальные характеристики каждой группы.

Для оценки эффективности предлагаемого метода будут проведены эксперименты с реальными наборами данных финансовых временных рядов, сравнивающие точность прогнозирования с использованием полученного индекса и без него. Результаты прогнозирования будут оцениваться с использованием таких показателей, как средняя абсолютная процентная ошибка (MAPE), средняя абсолютная ошибка (MAE) и симметричная средняя абсолютная процентная ошибка (SMAPE).

Методы кластеризации показали многообещающие результаты в отражении базовой структуры финансовых временных рядов, но их эффективность в повышении точности прогнозирования досконально не исследована. В целом, данная работа демонстрирует потенциал использования методов кластеризации для повышения точности прогнозирования финансовых временных рядов и анализа динамики рыночной капитализации публичных компаний.

## ПРЕДЛАГАЕМЫЙ ПОДХОД

В этой работе можно выделить три глобальных этапа.

1. Кластеризация временных рядов. Будут исследоваться 3 вида кластеризации: Иеррахическая, K-Means, Feature based. Показателем сходства временных рядов являются евклидово расстояние и DTW.
2. Построение базовой модели для прогнозирования цен на акции. На этом этапе важно выбрать гиперпараметры для модели TFT из пакета darts.
3. Использование полученных кластеров для улучшения прогнозирования цен на акции. Обучение TFT-модели на основе списка временных рядов, составленный из данных компаний из одного индекса.

# ЭКСПЕРИМЕНТЫ

## ДАТАСЕТ

Набор данных, использованный в этой работе, был собран с сайта Yahoo Finance [7] за 7-летний период. Набор данных состоит из данных временных рядов цен закрытия компаний, включенных в индекс NASDAQ-100. В дополнение к временным рядам, набор данных также включает атрибуты, перечисленные в таблице 2.

Таблица 2 – Описание свойств набора данных

|  |  |  |
| --- | --- | --- |
| Атрибут | Описание | Тип |
| Тикер | Уникальный идентификатор, используемый для идентификации компании на бирже | str |
| Краткое название | Краткое название компании | str |
| Страна | Страна, в которой зарегистрирована компания | str |
| Рынок | Фондовая биржа, на которой зарегистрирована компания | str |
| Сектор | Сектор, к которому принадлежит компания | str |
| Отрасль | Предоставляет более конкретную классификацию в рамках сектора | str |
| Временные ряды | Цена закрытия акций за каждый день за последние 7 лет (2014-2023) | float64 |

На рисунках 1 и 2 изображены представления датасета.

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 1 – Табличное представление датасета

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 2 – График временных рядов компаний

Набор данных обеспечивает всестороннее представление о фондовом рынке, объединяя данные временных рядов с характеристиками конкретных компаний.

Перед непосредственным анализом все данные были предварительно обработаны, чтобы удалить любые недостающие значения, выбросы и шумы. Данные также были нормализованы (каждый ряд был поделен на свое максимальное значение), чтобы гарантировать, что каждый временной ряд имел одинаковый масштаб.

## МЕТРИКИ

### С**ХОДСТВО ВРЕМЕННЫХ РЯДОВ**

Чтобы сгруппировать данные временных рядов, необходимо использовать меру подобия, которая может эффективно фиксировать временные взаимосвязи между точками данных. В этом исследовании DTW и евклидово расстояние использовалось в качестве показателя сходства для сравнения различных временных рядов.

#### ЕВКЛИДОВО РАССТОЯНИЕ

Расстояние, которое определяется как квадратный корень из суммы квадратов разностей между соответствующими элементами ряда, в формуле (1).

(1)

где n – количество элементов в ряду, а x и y – два временных ряда.

Евклидово расстояние обладает рядом преимуществ, таких как простота и интуитивность, а также хорошая возможность сравнивать временные ряды разной длины. Однако у него также есть некоторые недостатки, такие как чувствительность к выбросам и смещению, которые могут снизить качество кластеризации.

#### DTW

Динамическое искажение времени (DTW) - это алгоритм выравнивания временных рядов, который позволяет измерить сходство между двумя последовательностями, учитывая возможное смещение (сдвиг по времени) между ними.

Суть метода заключается в том, что две последовательности сравниваются попарно, а затем выравниваются относительно друг друга. В процессе выравнивания каждая точка одной последовательности соотносится с точкой другой последовательности, которая расположена на самом близком расстоянии во времени, даже если она значительно отличается от их исходного положения. На рисунке 3 изображено непосредственно само отображение точек одного ряда точке другого.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 3: Сопоставление двух временных рядов после вычисления DTW

Метод DTW работает следующим образом:

1. Сначала вычисляется предварительное расстояние между двумя временными рядами.
2. Затем строится матрица расстояний, где каждый элемент соответствует расстоянию между парой точек во временном ряду.
3. Вычисляется оптимальный путь с минимально достаточным расстоянием, сравниваются точки в двух временных рядах.
4. Итоговое расстояние вычисляется как сумма расстояний между точками на оптимальном пути.

### КАЧЕСТВО ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ

Для оценки качества прогнозирования временных рядов использовались метрики MAE, MAE, SMAPE.

#### MAPE

Средняя абсолютная процентная ошибка (MAPE) - это показатель, который используется для оценки точности моделей прогнозирования временных рядов. Он рассчитывается как среднее абсолютное значение процентной ошибки во всех точках прогноза. Математически MAPЕ рассчитывается по формуле (2).

(2)

где actual - фактическое значение во временном ряду, predicted - значение, предсказанное моделью, а n - количество точек во временном ряду.

Основным преимуществом использования MAPE является тот факт, что MAPE позволяет оценить точность модели прогнозирования в процентах, что является понятным и простым для интерпретации.

#### MAE

Средняя абсолютная ошибка (MAE) измеряет среднюю абсолютную разницу между прогнозируемыми значениями и фактическими значениями. MAE рассчитывается по формуле (2).

(2)

где actual - фактическое значение во временном ряду, predicted - значение, предсказанное моделью, а n - количество точек во временном ряду.

Преимущества показателя MAE заключаются в том, что он легко интерпретируется и не чувствителен к выбросам в данных, поскольку при расчете используется абсолютное значение. Кроме того, он является аддитивным и масштабируемым (т.е. его можно усреднять по разным диапазонам, например, для каждой выборки или для каждого класса в многоклассовой классификации).

По сравнению с другими методами, такими как MSE (среднеквадратичная ошибка), MAE лучше подходит для оценки малых прогнозируемых значений, поскольку он не увеличивает ошибку в случае больших прогнозов.

#### SMAPE

Симметричная средняя абсолютная процентная ошибка (SMAPE) - это процентное отклонение прогноза от фактического значения. Рассчитать SMAPE можно по формуле (3).

(3)

где actual – фактическое значение во временном ряду, predicted – значение, предсказанное моделью, а n – количество точек во временном ряду.

SMAPE лучше всего использовать для оценки качества прогнозов, когда важно достичь баланса между малыми и большими значениями, а также уменьшить влияние выбросов. SMAPE также устойчив к выбросам, поскольку использует количество абсолютных изменений ошибок, а не их площадь. Более того, это более интерпретируемый показатель, чем RMSE (среднеквадратичная ошибка), который не обеспечивает визуальной интерпретации, такой как SMAPE.

## РЕЗУЛЬТАТЫ

### КЛАСТЕРИЗАЦИЯ ВРЕМЕННЫХ РЯДОВ

Для этого исследования мы использовали данные о ценах акций компаний, входящих в индекс NASDAQ-100, за период с 6 апреля 2020 года по 5 апреля 2023 года. Были отобраны только компании, акции которых торгуются на фондовой бирже в течение всего периода времени. Для каждой компании был сформирован временной ряд, представляющий дневное значение цены акций на момент закрытия торгов.

### ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ

Иерархическая кластеризация была выполнена на основе расстояния DTW между парами временных рядов. Количество кластеров не было определено заранее, но было рассчитано на основе минимального порогового значения расстояния между рядами для попадания в один кластер.

Наилучший результат кластеризации был получен на расстоянии DTW, равном 2.5, было получено 4 кластера. Рисунок 4 иллюстрирует результат распределения компаний по кластерам.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 4 – Визуализация кластеров в результате иерархической кластеризации с помощью tSNE

В результате этого эксперимента стало понятно, что оптимальное число кластеров равно 5, оно будет использовано в дальнейших экспериментах.

### K-MEANS

Кластеризация на основе K-Means была выполнена на основе евклидова расстояния и DTW. На рисунках 5 и 6 изображены результаты.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 5 – Визуализация кластеров в результате K-Means + евклидово расстояние с помощью tSNE

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 6 – Визуализация кластеров в результате K-Means + DTW с помощью tSNE

Заметно, что евклидово расстояние справляется хуже, потому что оно не учитывает сдвиг во временных рядах. С DTW результат лучше, но из-за чувствительности алгоритма к выбросам, некоторые компании попадают в разные кластеры, хотя между ними небольшое расстояние.

В целом, оба подхода (иерархическая кластепризация и K-Means) дают неплохие результаты, но будем считать, что результаты первого эксперимента лучше, потому что в нем не нужно было знать число кластеров. Если нужно будет применить разрабатываемый подход в другой задаче, иерархическая кластеризация будет более адаптивна.

### FEATURE BASED

Подход хорошо зарекомендовал себя в других задачах, но он не способен выявить нелинейные связности между компаниями, из-за чего в данной задаче его применять не стоит. На рисунке 7 четко видно отсутствие четкой границы между кластерами.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 7 – Визуализация кластеров в результате feature based кластеризации с помощью tSNE

### ПОДВЕДЕНИЕ ИТОГОВ ЭТАПА

Результаты эксперементов с кластеризацией показали, что компании из индекса NASDAQ-100 можно разделить на 4 – 5 кластера. Наиболее репрезентативные результаты показала иерахическая кластеризация. Ее результаты станут основой для создания синтетических индексов.

В таблицах 3 – 6 отображено финальное распределние компаний по кластерам, которое будет использоваться в дальнейших экспериментах.

Таблица 3 – Компании, попавшие в кластер №1

|  |  |  |  |
| --- | --- | --- | --- |
| Название компании | Сектор | Отрасль | Страна |
| Apple Inc. | Technology | Consumer Electronics | United States |
| Analog Devices, Inc. | Technology | Semiconductors | United States |
| Automatic Data Processing, Inc. | Industrials | Staffing & Employment Services | United States |
| American Electric Power Company | Utilities | Utilities – Regulated Electric | United States |
| Amgen Inc. | Healthcare | Drug Manufacturers – General | United States |
| ANSYS, Inc. | Technology | Software – Application | United States |
| Activision Blizzard, Inc | Communication Services | Electronic Gaming & Multimedia | United States |
| Broadcom Inc. | Technology | Semiconductors | United States |
| Astrazeneca PLC | Healthcare | Drug Manufacturers – General | United Kingdom |
| Biogen Inc. | Healthcare | Drug Manufacturers – General | United States |
| Booking Holdings Inc. Common St | Consumer Cyclical | Travel Services | United States |
| Cadence Design Systems, Inc. | Technology | Software – Application | United States |
| Costco Wholesale Corporation | Consumer Defensive | Discount Stores | United States |
| Copart, Inc. | Consumer Cyclical | Auto & Truck Dealerships | United States |
| Cisco Systems, Inc. | Technology | Communication Equipment | United States |
| CSX Corporation | Industrials | Railroads | United States |
| Cintas Corporation | Industrials | Specialty Business Services | United States |
| Dollar Tree, Inc. | Consumer Defensive | Discount Stores | United States |
| Electronic Arts Inc. | Communication Services | Electronic Gaming & Multimedia | United States |
| Exelon Corporation | Utilities | Utilities – Regulated Electric | United States |
| Fastenal Company | Industrials | Industrial Distribution | United States |
| Fiserv, Inc. | Technology | Information Technology Services | United States |
| Fortinet, Inc. | Technology | Software – Infrastructure | United States |
| Gilead Sciences, Inc. | Healthcare | Drug Manufacturers – General | United States |
| Honeywell International Inc. | Industrials | Conglomerates | United States |
| Keurig Dr Pepper Inc. | Consumer Defensive | Beverages – Non-Alcoholic | United States |
| The Kraft Heinz Company | Consumer Defensive | Packaged Foods | United States |
| KLA Corporation | Technology | Semiconductor Equipment & Materials | United States |
| Marriott International | Consumer Cyclical | Lodging | United States |
| Microchip Technology Incorporat | Technology | Semiconductors | United States |
| Mondelez International, Inc. | Consumer Defensive | Confectioners | United States |
| Monster Beverage Corporation | Consumer Defensive | Beverages – Non-Alcoholic | United States |
| Old Dominion Freight Line, Inc. | Industrials | Trucking | United States |
| O'Reilly Automotive, Inc. | Consumer Cyclical | Specialty Retail | United States |
| Palo Alto Networks, Inc. | Technology | Software – Infrastructure | United States |
| Paychex, Inc. | Industrials | Staffing & Employment Services | United States |
| PACCAR Inc. | Industrials | Farm & Heavy Construction Machinery | United States |
| Pepsico, Inc. | Consumer Defensive | Beverages – Non-Alcoholic | United States |
| Regeneron Pharmaceuticals, Inc. | Healthcare | Biotechnology | United States |
| Ross Stores, Inc. | Consumer Cyclical | Apparel Retail | United States |
| Starbucks Corporation | Consumer Cyclical | Restaurants | United States |
| Seagen Inc. | Healthcare | Biotechnology | United States |
| Sirius XM Holdings Inc. | Communication Services | Entertainment | United States |
| Synopsys, Inc. | Technology | Software – Infrastructure | United States |
| T-Mobile US, Inc. | Communication Services | Telecom Services | United States |
| Texas Instruments Incorporated | Technology | Semiconductors | United States |
| Verisk Analytics, Inc. | Industrials | Consulting Services | United States |
| VeriSign, Inc. | Technology | Software – Infrastructure | United States |
| Vertex Pharmaceuticals Incorpor | Healthcare | Biotechnology | United States |
| Walgreens Boots Alliance, Inc. | Healthcare | Pharmaceutical Retailers | United States |
| Xcel Energy Inc. | Utilities | Utilities – Regulated Electric | United States |

Таблица 4 – Компании, попавшие в кластер №2

|  |  |  |  |
| --- | --- | --- | --- |
| Название компании | Сектор | Отрасль | Страна |
| Charter Communications, Inc. | Communication Services | Telecom Services | United States |
| Illumina, Inc. | Healthcare | Diagnostics & Research | United States |
| Intel Corporation | Technology | Semiconductors | United States |
| Meta Platforms, Inc. | Communication Services | Internet Content & Information | United States |
| Netflix, Inc. | Communication Services | Entertainment | United States |
| Splunk Inc. | Technology | Software – Infrastructure | United States |

Таблица 5 – Компании, попавшие в кластер №3

|  |  |  |  |
| --- | --- | --- | --- |
| Название компании | Сектор | Отрасль | Страна |
| Adobe Inc. | Technology | Software – Infrastructure | United States |
| Autodesk, Inc. | Technology | Software – Application | United States |
| Applied Materials, Inc. | Technology | Semiconductor Equipment & Materials | United States |
| Amazon.com, Inc. | Consumer Cyclical | Internet Retail | United States |
| ASML Holding N.V. - New York Re | Technology | Semiconductor Equipment & Materials | Netherlands |
| Comcast Corporation | Communication Services | Telecom Services | United States |
| Cognizant Technology Solutions | Technology | Information Technology Services | United States |
| DexCom, Inc. | Healthcare | Medical Devices | United States |
| eBay Inc. | Consumer Cyclical | Internet Retail | United States |
| Alphabet Inc. | Communication Services | Internet Content & Information | United States |
| Alphabet Inc. | Communication Services | Internet Content & Information | United States |
| IDEXX Laboratories, Inc. | Healthcare | Diagnostics & Research | United States |
| Intuit Inc. | Technology | Software – Application | United States |
| Intuitive Surgical, Inc. | Healthcare | Medical Instruments & Supplies | United States |
| Lam Research Corporation | Technology | Semiconductor Equipment & Materials | United States |
| lululemon athletica inc. | Consumer Cyclical | Apparel Retail | Canada |
| MercadoLibre, Inc. | Consumer Cyclical | Internet Retail | Uruguay |
| Micron Technology, Inc. | Technology | Semiconductors | United States |
| NetEase, Inc. | Communication Services | Electronic Gaming & Multimedia | China |
| NVIDIA Corporation | Technology | Semiconductors | United States |
| NXP Semiconductors N.V. | Technology | Semiconductors | Netherlands |
| QUALCOMM Incorporated | Technology | Semiconductors | United States |
| Skyworks Solutions, Inc. | Technology | Semiconductors | United States |
| Workday, Inc. | Technology | Software – Application | United States |

Таблица 6 – Компании, попавшие в кластер №4

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Название компании | Сектор | | Отрасль | | Страна | |
|  | |  | |  | |  | |
| Align Technology, Inc. | | Healthcare | | Medical Devices | | United States | |
| Advanced Micro Devices, Inc. | | Technology | | Semiconductors | | United States | |
| Baidu, Inc. | | Communication Services | | Internet Content & Information | | China | |
| CrowdStrike Holdings, Inc. | | Technology | | Software – Infrastructure | | United States | |
| Datadog, Inc. | | Technology | | Software – Application | | United States | |
| DocuSign, Inc. | | Technology | | Software – Application | | United States | |
| JD.com, Inc. | | Consumer Cyclical | | Internet Retail | | China | |
| Moderna, Inc. | | Healthcare | | Biotechnology | | United States | |
| Marvell Technology, Inc. | | Technology | | Semiconductors | | United States | |
| Match Group, Inc. | | Communication Services | | Internet Content & Information | | United States | |
| Okta, Inc. | | Technology | | Software – Infrastructure | | United States | |
| Pinduoduo Inc. | | Consumer Cyclical | | Internet Retail | | China | |
| PayPal Holdings, Inc. | | Financial Services | | Credit Services | | United States | |
| Atlassian Corporation | | Technology | | Software – Application | | Australia | |
| Tesla, Inc. | | Consumer Cyclical | | Auto Manufacturers | | United States | |
| Zoom Video Communications, Inc. | | Technology | | Software – Application | | United States | |
| Zscaler, Inc. | | Technology | | Software – Infrastructure | | United States | |

### СИНТЕТИЧЕСКИЙ ИНДЕКС

Каждый кластер представляет собой отдельный синтетический индекс. На рисунках 8 – 11 синей линией обозначена линия индекса кластера. Индекс рассчитывается как среднее значение между временными рядами, которые попадают в один и тот же кластер.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 8 – Синтетический индекс 1

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 9 – Синтетический индекс 2

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 10 – Синтетический индекс 3

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 11 – Синтетический индекс 4

### ПРОГНОЗИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ

Первой задачей этого этапа стало обучение базовой модели TFT на одном временном ряду, было важно правильно подобрать гиперпараметры. Путем ручного перебора наилучшее значение MAPE было у модели с input\_chunck\_length = 7, output\_chunk\_length = 1, n\_samples = 100.

Дальнейшие исследования показали, что индекс может быть использован для улучшения прогнозов котировок компаний. Наиболее эффективные результаты показал способ с обучением не просто на синтетическом индексе, а с обучением на списке временных рядов компаний из одного индекса.

Чтобы убедиться в верности гипотезы эксперименты были повторены для 12 случайных компаний (для 3 компаний из каждого индекса). Итоговые результаты показали, в среднем более точные результаты показывает разработанный в этой научной статье подход. Для некоторых компаний небольшое улучшение также демонстрируют модели, обученные на случайном индексе, но стоит отметить, что для большинства из этих моделей прогноз либо незначительно улучшился, либо, наоборот, ухудшился.

В таблице 7 показано сравнение показателей для модели, обученной только на одном временном ряду (столбец «Базовый»), для модели, обученной на временных рядах компаний из одного кластера (столбец «Индекс»), и для модели, обученной на данных случайных компаний из разных индексов (столбец «Рандомный»).

Таблица 7 – Результаты предсказания модели

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Название компании | Базовый | | | Индекс | | | Рандомный | | |
| MAPE | MAE | SMAPE | MAPE | MAE | SMAPE | MAPE | MAE | SMAPE |
| Кластер 1 | | | | | | | | | |
| Fortinet, Inc. | 4,64 | 0,04 | 4,80 | **2.67** | **0.02** | **2.72** | 3,06 | 0,02 | 3,15 |
| Automatic Data Processing, Inc. | 2,73 | 0,02 | 2,78 | **1.95** | 0,02 | **1.96** | 2,58 | 0,02 | 2,59 |
| Apple Inc. | 3,69 | 0,03 | 3,80 | **1.94** | **0.02** | **1.98** | 3,69 | 0,03 | 3,75 |
| Кластер 2 | | | | | | | | | |
| Charter Communications, Inc. | 3,31 | 0,01 | 3,34 | **2.79** | 0,01 | 2.82 | 2,98 | 0,01 | 3,01 |
| Intel Corporation | 5,35 | 0,02 | 5,53 | **3.33** | **0.01** | **3.35** | 3,95 | 0,02 | 4,09 |
| Netflix, Inc. | 4,65 | 0,02 | 4,82 | **3.46** | 0,02 | 3,51 | 4,08 | 0,02 | 4,19 |
| Кластер 3 | | | | | | | | | |
| Intuit Inc. | 3,74 | 0,02 | 3,77 | 3,70 | 0,02 | 3,74 | *4.13* | 0,02 | *4.15* |
| Adobe Inc. | 3,72 | 0,02 | 3,80 | **3.43** | 0,02 | **3.48** | 3,62 | 0,02 | 3,66 |
| Skyworks Solutions, Inc. | 3,87 | 0,02 | 3,95 | **3.14** | 0,02 | **3.19** | *4.37* | 0,02 | *4.53* |
| Кластер 4 | | | | | | | | | |
| Marvell Technology, Inc. | 5,41 | 0,02 | 5,50 | **4.83** | 0,02 | **4.91** | *5.43* | *0.03* | *5.6* |
| Baidu, Inc. | 4,76 | 0,02 | 4,66 | 4,40 | 0,02 | 4,38 | 4,52 | 0,02 | 4,51 |
| Match Group, Inc. | 6,18 | 0,01 | 5,99 | 4,57 | 0,01 | 4,52 | 4,77 | 0,01 | 4,74 |

Рисунки 12 – 36 иллюстрируют фактические и прогнозируемые цены акций 12 компаний, перечисленных в таблице 7.

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 12 – Кластер 1: Fortinet, Inc. [Модель, обученная на одном временном ряде]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 13 – Кластер 1: Fortinet, Inc. [Модель, обученная на временных рядах из одного индекса]

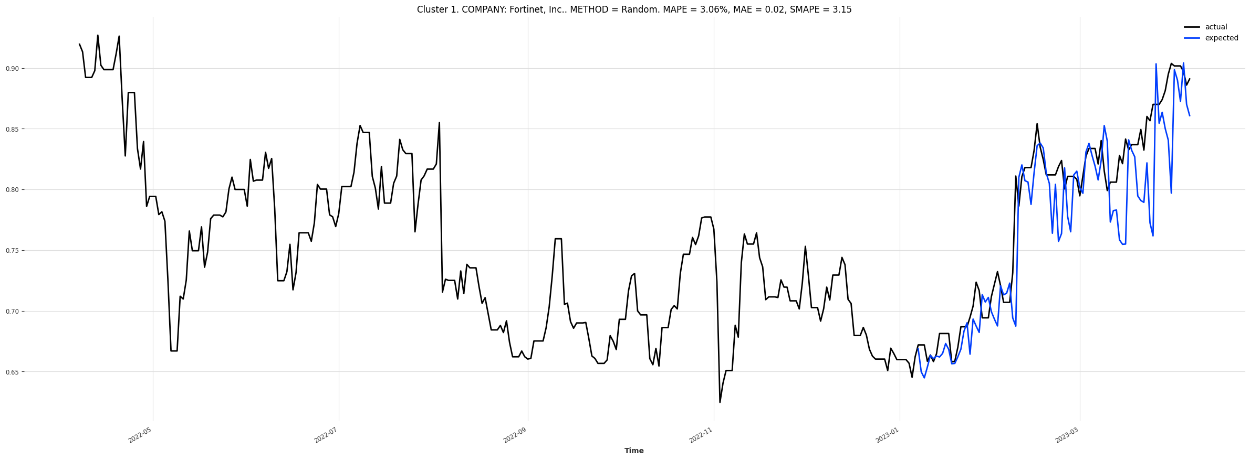


Рисунок 14 – Кластер 1: Fortinet, Inc. [Модель, обученная на временных рядах из рандомных индексов]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 15 – Кластер 1: Automatic Data Processing, Inc. [Модель, обученная на одном временном ряде]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 16 – Кластер 1: Automatic Data Processing, Inc. [Модель, обученная на временных рядах из одного индекса]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 17 – Кластер 1: Automatic Data Processing, Inc. [Модель, обученная на временных рядах из рандомных индексов]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 18 – Кластер 1: Apple Inc. [Модель, обученная на одном временном ряде]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 19 – Кластер 1: Apple Inc. [Модель, обученная на временных рядах из одного индекса]



Рисунок 20 – Кластер 1: Apple Inc. [Модель, обученная на временных рядах из рандомных индексов]

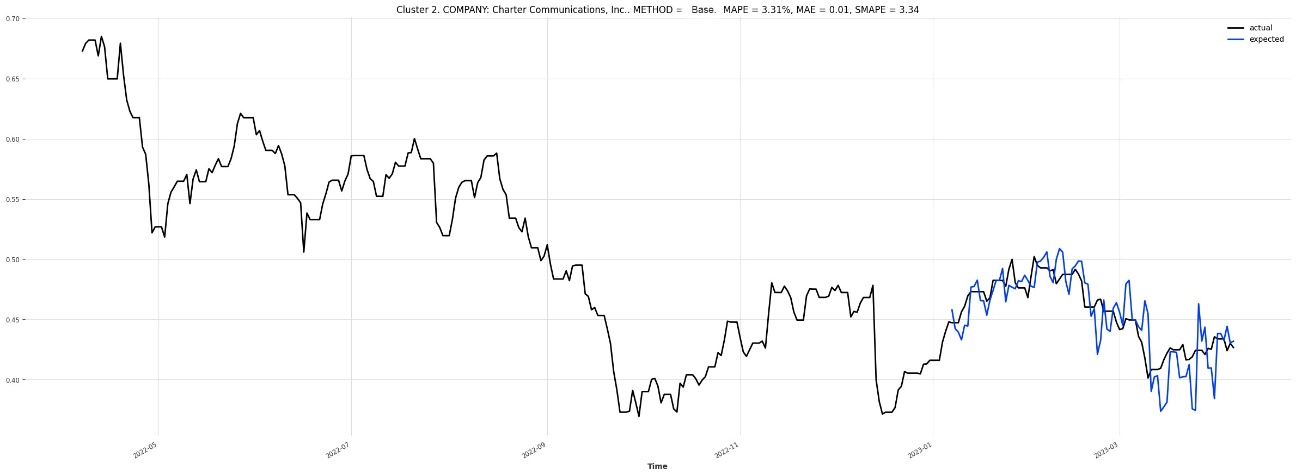


Рисунок 21 – Кластер 2: Charter Communications, Inc. [Модель, обученная на одном временном ряде]

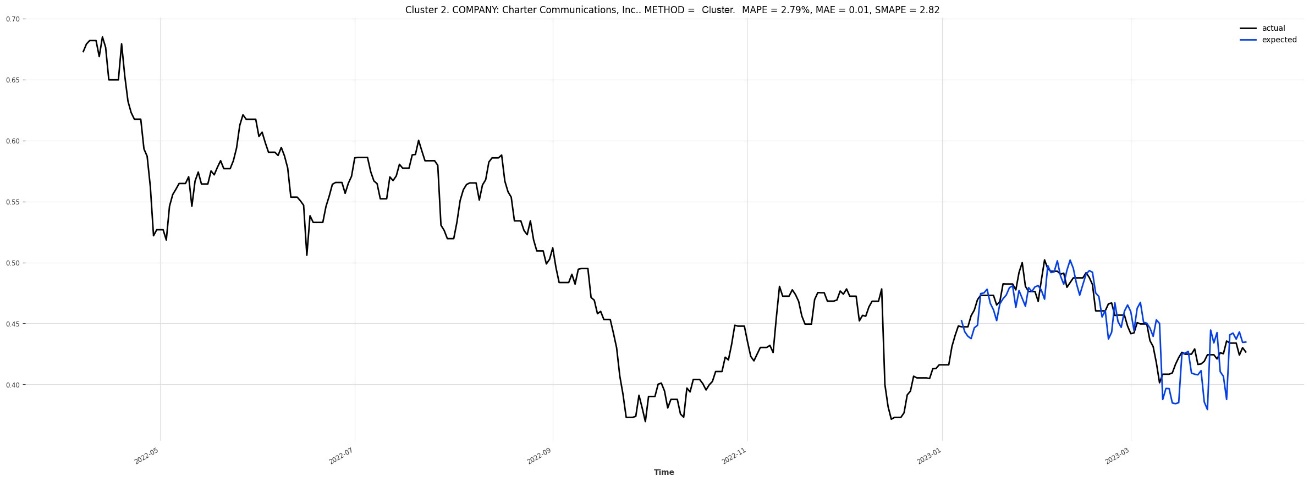


Рисунок 22 – Кластер 2: Charter Communications, Inc. [Модель, обученная на временных рядах из одного индекса]

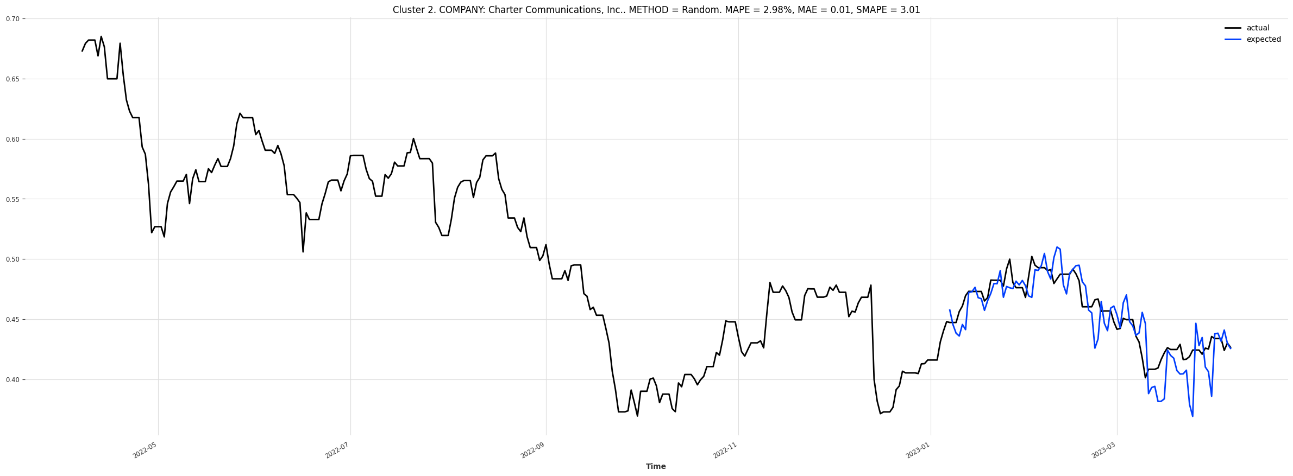


Рисунок 23 – Кластер 2: Charter Communications, Inc. [Модель, обученная на временных рядах из рандомных индексов]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 24 – Кластер 2: Intel Corporation [Модель, обученная на одном временном ряде]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 25 – Кластер 2: Intel Corporation [Модель, обученная на временных рядах из одного индекса]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 26 – Кластер 2: Intel Corporation [Модель, обученная на временных рядах из рандомных индексов]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 27 – Кластер 2: Netflix, Inc. [Модель, обученная на одном временном ряде]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 28 – Кластер 2: Netflix, Inc. [Модель, обученная на временных рядах из одного индекса]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 29 – Кластер 2: Netflix, Inc. [Модель, обученная на временных рядах из рандомных индексов]

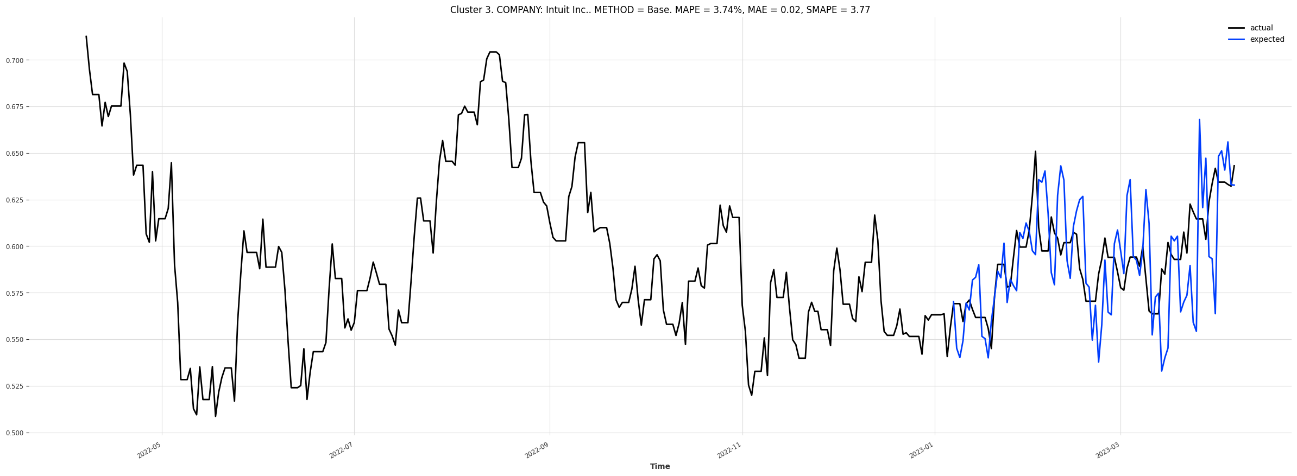


Рисунок 30 – Кластер 3: Intuit Inc. [Модель, обученная на одном временном ряде]

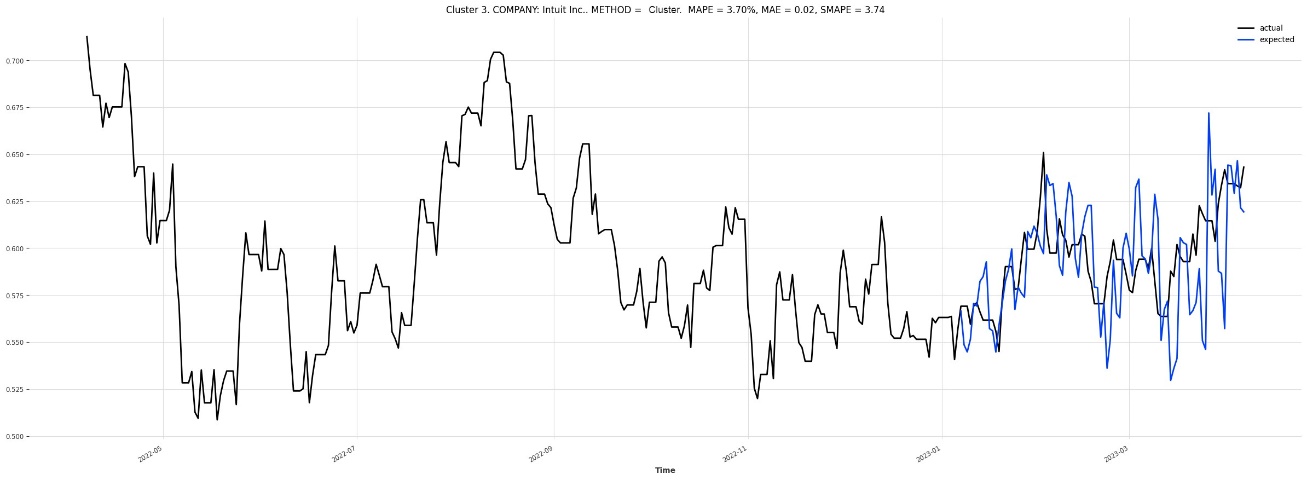


Рисунок 31 – Кластер 3: Intuit Inc. [Модель, обученная на временных рядах из одного индекса]

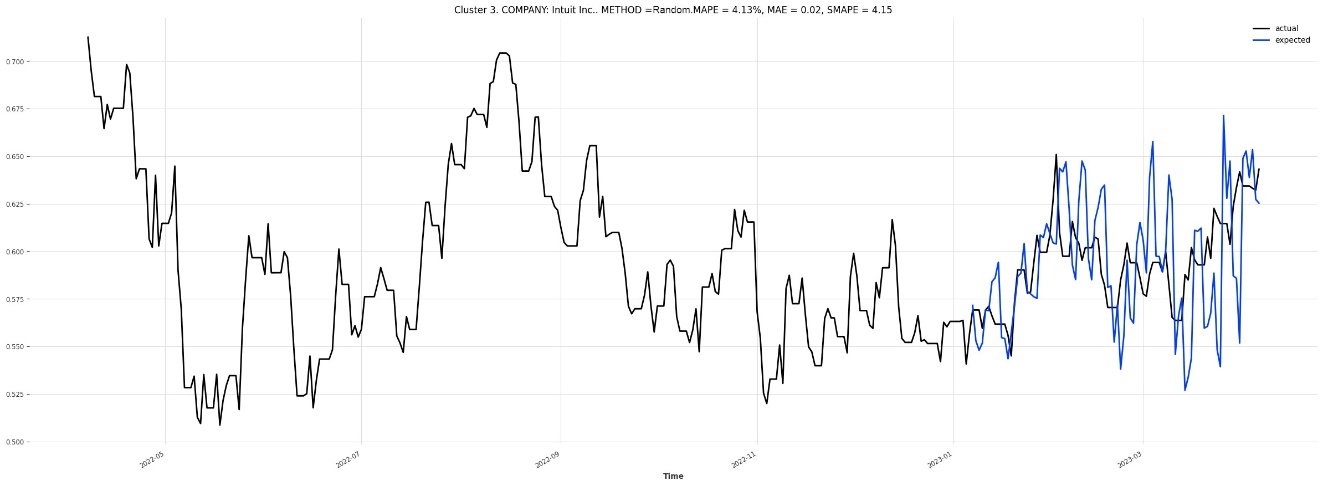


Рисунок 32 – Кластер 3: Intuit Inc. [Модель, обученная на временных рядах из рандомных индексов]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 33 – Кластер 3: Adobe Inc. [Модель, обученная на одном временном ряде]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 34 – Кластер 3: Adobe Inc. [Модель, обученная на временных рядах из одного индекса]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 35 – Кластер 3: Adobe Inc. [Модель, обученная на временных рядах из рандомных индексов]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 36 – Кластер 3: Skyworks Solutions, Inc. [Модель, обученная на одном временном ряде]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 37 – Кластер 3: Skyworks Solutions, Inc. [Модель, обученная на временных рядах из одного индекса]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 38 – Кластер 3: Skyworks Solutions, Inc. [Модель, обученная на временных рядах из рандомных индексов]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 39 – Кластер 4: Marvell Technology, Inc. [Модель, обученная на одном временном ряде]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 40 – Кластер 4: Marvell Technology, Inc. [Модель, обученная на временных рядах из одного индекса]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 41 – Кластер 4: Marvell Technology, Inc. [Модель, обученная на временных рядах из рандомных индексов]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 42 – Кластер 4: Baidu, Inc. [Модель, обученная на одном временном ряде]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 43 – Кластер 4: Baidu, Inc. [Модель, обученная на временных рядах из одного индекса]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 44 – Кластер 4: Baidu, Inc. [Модель, обученная на временных рядах из рандомных индексов]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 45 – Кластер 4: Match Group, Inc. [Модель, обученная на одном временном ряде]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 46 – Кластер 4: Match Group, Inc. [Модель, обученная на временных рядах из одного индекса]

Изображение выглядит как диаграмма

Автоматически созданное описание

Рисунок 47 – Кластер 4: Match Group, Inc. [Модель, обученная на временных рядах из рандомных индексов]

Таким образом, проведенная работа показала, что кластеризация компаний на основе их временных рядов может иметь большое практическое значение для анализа финансовых рынков и прогнозирования тенденций на них в будущем.

# ЗАКЛЮЧЕНИЕ

В этой статье было проведено исследование на тему кластеризации временных рядов с целью улучшения прогнозов. Были рассмотрены различные методы кластеризации временных рядов, а также алгоритмы прогнозирования.

Среди рассмотренных алгоритмов кластеризации метод иерархической кластеризации на основе DTW оказался особенно эффективным. Несравненным преимуществом алгоритма является то, что параметром для регулирования количества кластеров в данном случае является среднее расстояние между кластерами. Изменяя этот параметр, гораздо проще добиться оптимального количества кластеров. Метрика расстояния также играет огромную роль в любом алгоритме кластеризации, поскольку важно, чтобы при расчете расстояния учитывался сдвиг в двух временных рядах.

Гипотеза этого исследования была подтверждена. Обучение на данных компаний из одного и того же индекса действительно может улучшить прогноз. В этом случае дополнительные временные ряды становятся аугментациями к исходному ряду.

Подводя итог, в рамках этого исследования мною была проделана следующая работа:

1. Провела обзор литературы по алгоритмам кластеризации для анализа временных рядов и методам их прогнозирования;
2. Создала синтетический индекс на основе оценки результатов кластеризации;
3. Применила TFT-модель для прогнозирования данных финансовых временных рядов;
4. Применила синтетический индекс для улучшения качества прогнозирования.

В целом, результаты этого исследования показывают важность использования кластеризации для повышения точности прогнозирования в машинном обучении. Дальнейшие исследования могут быть проведены на большом наборе данных с использованием других методов кластеризации и метрик расстояния между строками для достижения более высокой точности прогнозирования.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Time series similarity measures and time series indexing. / Gunopulos, Dimitrios & Das, Gautam. // (2001) ACM SIGMOD Record. 30. 624. 10.1145/376284.375808.

2. Recent Techniques of Clustering of Time Series Data: A Survey. / Rani, Sangeeta & Sikka, Geeta. // (2012) International Journal of Computer Applications. 52. 1-9. 10.5120/8282-1278.

3. Clustering Time Series Data — A Survey. / Liao, T. // (2005) Pattern Recognition. 38. 1857-1874. 10.1016/j.patcog.2005.01.025.3.

4. Time-series clustering - A decade review. / Aghabozorgi, Sr & Shirkhorshidi, Ali Seyed & Wah, Teh. // (2015) Information Systems. 53. 10.1016/j.is.2015.04.007.

5. Stock price prediction using LSTM, RNN and CNN-sliding window model. / Selvin, Sreelekshmy & Ravi, Vinayakumar & Gopalakrishnan, E. A & Menon, Vijay & Kp, Soman. // (2017). 1643-1647. 10.1109/ICACCI.2017.8126078.

6. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. / Lim, Bryan & Arık, Sercan & Loeff, Nicolas & Pfister, Tomas. // (2021). International Journal of Forecasting. 37. 10.1016/j.ijforecast.2021.03.012.

7. Yahoo Finance [Электронный ресурс] / Режим доступа: https://finance.yahoo.com/, свободный (дата обращения: 01.11.2022 – 30.11.2022)

# ПРИЛОЖЕНИЕ А

**Ссылка на гитхаб с дополнительными материалами**

На гитхабе, расположенном по ссылке <https://github.com/ANasykhova/Course-Work-3-year>, можно найти исходный код всех приведенных эксперимпентов, а также препринт этой статьи на англйском языке.