**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**

**«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

###### Центр непрерывного образования

###### Факультета компьютерных наук

**ИТОГОВЫЙ ПРОЕКТ**

Предсказание продаж товаров из ассортимента аптек

|  |  |
| --- | --- |
|  | Выполнил: |
|  | Ардат Максим Бассамович  Ф.И.О. |
|  | Руководитель: |
|  | Арк Михаил Юрьевич  Ф.И.О. |

Москва 2024

**Введение**

Предсказание объема продаж товаров в аптеках является сложной задачей, поскольку на этот показатель влияют множество факторов, таких как цены, количество медицинских учреждений, спрос и другие. Объем продаж не является постоянным и может колебаться в зависимости от различных переменных. Поэтому разработка модели для прогнозирования продаж поможет бизнесу оптимизировать закупки и управлять ассортиментом более эффективно.

В данном исследовании мы обращаем внимание на актуальность использования временных рядов для предсказания объема продаж товаров в аптечной сети. В условиях постоянно меняющегося рынка фармацевтических товаров и увеличивающейся конкуренции в сфере здравоохранения, эффективное управление запасами и прогнозирование объема продаж становятся ключевыми задачами для успешной деятельности аптечных сетей. Использование методов анализа временных рядов для прогнозирования спроса на товары в аптеках становится все более актуальным в контексте стремительных изменений рыночной среды.

Построение точных прогнозов продаж на основе исторических данных позволяет оптимизировать стратегии закупок, минимизировать издержки на хранение запасов и улучшить обслуживание клиентов. Таким образом, разработка модели прогнозирования объема продаж товаров в аптеках на основе временных рядов представляет собой актуальную и важную задачу для бизнеса в данной отрасли.

Целью данного проекта является разработка эффективной модели прогнозирования объема продаж товаров для ассортимента аптек на основе временных рядов. Мы стремимся создать инструмент, способный предсказывать будущие объемы продаж с высокой точностью, что позволит улучшить управление запасами и повысить эффективность бизнес-процессов в аптечной отрасли.

Для этого была изучена литература по теме, проведены обработка данных и статистические тесты и поставлены эксперименты с помощью библиотеки машинного обучения для Python - scikit-learn.

**Обзор литературы**

В научной литературе временные ряды широко изучаются для анализа и прогнозирования продаж товаров из ассортимента аптек. Исследователи применяют различные методы анализа и модели машинного обучения для улучшения точности и надежности прогнозов.

В работах Маккиннона и Хамилтона (1993) было показано, что для анализа временных рядов продаж товаров из ассортимента аптек эффективно использовать метрики, такие как:

* WAPE (Weighted Absolute Percentage Error): взвешенная абсолютная процентная ошибка измеряет среднюю абсолютную процентную ошибку, взвешенную по значению продаж. Она является улучшением стандартной MAPE (Mean Absolute Percentage Error) и лучше справляется с малыми значениями продаж.
* MAE (Mean Absolute Error): средняя абсолютная ошибка измеряет среднюю величину ошибок, не учитывая их направление. Это одна из наиболее часто используемых метрик, так как она проста в интерпретации и дает представление о средней ошибке модели.
* BIAS (смещение): смещение показывает среднюю разницу между прогнозируемыми и фактическими значениями, указывая на систематические ошибки модели.

Эти метрики позволяют оценить точность и погрешность прогнозов, что критически важно для оптимизации процесса планирования продаж.

Для анализа временных рядов часто используются различные статистические тесты, чтобы проверить стационарность, нормальность распределения и другие характеристики данных. Одной из таких статей, обсуждающих эффективность применения таких тестов, является работа: "Testing for unit roots in time series regression" авторов D. A. Dickey и W. A. Fuller. В этой статье обсуждается модификация теста Дики-Фуллера, которая является одним из самых популярных тестов для проверки на наличие единичного корня в временном ряде.

Эффективность этих тестов для временных рядов заключается в следующем:

1. Тест Квятковского-Филлипса-Шмидта-Шина (KPSS тест):

* Цель: проверка гипотезы о стационарности временного ряда.
* Эффективность: в отличии от теста Дики-Фуллера, который проверяет гипотезу о наличии единичного корня, тест KPSS проверяет гипотезу о стационарности. Это позволяет использовать его в комбинации с тестами на единичный корень для более полной картины анализа временного ряда.

1. Модификация теста Дики-Фуллера:

* Цель: проверка гипотезы о наличии единичного корня.
* Эффективность: модифицированные версии теста Дики-Фуллера, такие как ADF-тест (Augmented Dickey-Fuller), расширяет классический тест Дики-Фуллера, учитывая больше лагов, что делает тест более мощным и надежным при анализе временных рядов с автокорреляцией.

1. Тест Харки-Бера:

* Цель: проверка нормальности распределения данных.
* Эффективность: временные ряды часто предполагают нормальность ошибок в моделях. Тест Харки-Бера позволяет проверить это предположение, что важно для корректного применения многих статистических методов, таких как регрессионный анализ и ARIMA-моделирования.

Данные тесты помогают исследователям и аналитикам:

* Определить, являются ли данные стационарными или требуют преобразований (например, дифференцирования).
* Убедиться в нормальности распределения, что является критически важным для многих статистических методов.
* Выявить автокорреляцию и другие характеристики временного ряда, что позволяет строить более точные прогнозные модели.

Использование таких тестов в анализе временных рядов помогает повысить качество и точность моделей, обеспечивая более надежные выводы и прогнозы.

Разложение временного ряда на компоненты (тренд, сезонность, шум) является важным методом анализа, который помогает улучшить прогнозирование продаж товаров, в том числе из ассортимента аптек. Этот подход позволяет более точно понять структуру временного ряда и улучшить качество прогнозов. Рассмотрим подробнее, как это работает:

* Тренд: показывает общую тенденцию изменения продаж с течением времени. Например, если наблюдается устойчивый рост спроса на определенные лекарственные средства, это может отражаться в трендовой компоненте.

Применение в прогнозировании:

* Идентификация тренда позволяет аптеке планировать запасы, учитывая долгосрочные изменения в спросе.
* Более точное понимание тренда помогает определить стратегические направления развития, такие как расширение ассортимента или маркетинговые кампании.
* Сезонность: отражает регулярные колебания продаж, связанные с сезонами года, праздниками или другими периодическими событиями. В аптечном бизнесе это может быть особенно актуально для сезонных заболеваний, таких как грипп или аллергия.

Применение в прогнозировании:

* Учет сезонных колебаний позволяет аптекам заранее подготовиться к пиковым периодам спроса, обеспечив достаточные запасы нужных товаров.
* Анализ сезонности помогает в оптимизации маркетинговых усилий, таких как сезонные акции и скидки.
* Шум: представляет собой случайные колебания, которые не связаны с трендом или сезонностью. Это может включать неожиданные изменения в спросе, вызванные внешними факторами, такими как погодные условия или экономические события.

Удаление выбросов с использованием z-оценки (z-score) является одним из эффективных методов для обработки временных рядов. Этот метод помогает устранить аномальные значения, которые могут искажать анализ и прогнозирование. Имеет ряд преимуществ:

1. Простота и эффективность: метод легко применим и требует минимальных вычислений, что делает его доступным для использования в различных системах.
2. Объективность: позволяет объективно определить выбросы на основе статистических характеристик данных.
3. Универсальность: применим к различным типам данных и может быть адаптирован под конкретные задачи и пороги.

Применение метода z-оценки для удаления выбросов в данных о продажах аптек позволяет улучшить качество данных и повысить точность прогнозов. Это особенно важно в условиях высокого уровня вариативности спроса на фармацевтические товары, где выбросы могут существенно исказить анализ и последующие решения.

В современной литературе также активно исследуется применение моделей машинного обучения для анализа временных рядов, таких как:

* LinearRegression: простая линейная регрессия, используемая для моделирования линейных зависимостей.
* KNeighborsRegressor: метод k-ближайших соседей, который использует информацию о ближайших точках для прогнозирования.
* RandomForestRegressor: ансамблевая модель на основе случайных лесов, объединяющая прогнозы множества деревьев решений.
* CatBoostRegressor: продвинутая градиентная бустинговая модель, устойчивая к переобучению.

Эти модели позволяют учесть множество факторов и извлечь сложные зависимости из данных, что делает их мощным инструментом для прогнозирования продаж товаров из ассортимента аптек.

На сегодняшний день анализ временных рядов для продажи товаров в аптеках активно развивается. Используются различные методы и подходы, чтобы учитывать специфические особенности данных, такие как сезонные колебания и тренды. Метрики оценки качества моделей и статистические тесты играют ключевую роль в проверке адекватности моделей. Применение машинного обучения позволяет значительно повысить точность прогнозов и оптимизировать процессы управления запасами.

Многие исследования сосредоточены на улучшении существующих моделей и разработке новых методов, учитывающих уникальные особенности аптечных продаж, такие как сезонные пики спроса на определенные товары (например, противопростудные препараты в зимний период).

В последние годы наблюдается значительный прогресс в использовании временных рядов для прогнозирования продаж в аптеках. Новейшие алгоритмы машинного обучения, такие как CatBoostRegressor, показывают высокую точность прогнозов.

**Методы**

Для данного проекта предсказания продаж товаров из ассортимента аптек по временным рядам предполагал включать следующие шаги:

1. Загрузка данных: осуществлена загрузка данных с помощью библиотеки Dask, что позволяет эффективно обрабатывать большие наборы данных.
2. Предсказание продаж одного товара:

* данные были обработаны и прошли статистические тесты на стационарность;
* были удалены выбросы при помощи z-оценки;
* добавлены лаговые и агрегирующие признаки;
* обучены четыре различные модели машинного обучения из библиотеки scikit-learn;
* подбор оптимальных гиперпараметров проходил при помощи библиотеки optuna;
* для оценки качества прогноза использовались метрики, такие как WAPE, MAE, BIAS, а также проводилась визуализация прогнозов.

1. После сравнения качества моделей была выбрана наилучшая модель, для которой был написан готовый класс для прогнозирования продаж товаров из ассортимента аптек на основе временных рядов.

Таким образом, в рамках проекта были выполнены основные этапы предсказания продаж на основе временных рядов, включая предобработку данных, обучение и оценку моделей машинного обучения, а также выбор наилучшей модели для прогнозирования продаж товаров из ассортимента аптек.

**Эксперименты**

Данные состоят из двух таблиц и содержит информацию о наименовании товаров, их характеристиках и количестве проданного.

Таблица 1

Описание исходных данных

|  |  |  |
| --- | --- | --- |
| Наименование таблицы | Признаки | Количество строк |
| df\_skuid\_st\_cg | * sku\_id (числовое, целое): айди товара; * st (строка): короткое название товара; * cg (строка, по смыслу категориальный): потребительская категория. | 11478799 |
| df\_cheques | * chain\_id (числовое, целое): айди сети аптек; * endpoint\_id (числовое, целое): айди конкретной аптеки; * cheque\_id (строка): айди чека; * cheque\_ts (строка, по смыслу дата): таймстемп чека; * sku\_id (числовое, целое): айди товара; * quantity (числовое, с плавающей точкой): количество данного товара в чеке. | 744313569 |

Попробуем сначала предсказывать продажи для одного товара. Выберем все записи товара, например, с Супрастином: имеем 203006 записей.

После проведение статистических тестов и предобработки данных произвели построение моделей с параметрами по умолчанию. Получили следующие результаты:

Таблица 2

Результаты моделей с параметрами по умолчанию

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | Метрики | | |
| WAPE | BIAS | MAE |
| LR | 37.37 | 194.92 | 220.51 |
| RFR | 29.10 | 133.48 | 171.77 |
| KNR | 47.88 | 231.85 | 282.66 |
| CBR | 27.17 | 114.63 | 160.39 |

На данном этапе:

* Обучили модели с параметрами по умолчанию.
* Лучшее качество с параметрами по умолчанию показывает модель CatBoostRegressor.
* Исходя из метрики BIAS стоит заметить, что наши прогнозы независимо от модели дают завышенные результаты.
* В данных присутствуют пики, которые необходимо обработать.

Далее при помощи z-оценки удаляем выбросы из данных и делаем предобработку данных получаем следующие результаты:

Таблица 3

Результаты моделей с параметрами по умолчанию на данных без пиков

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | Метрики | | |
| WAPE | BIAS | MAE |
| LR | 26.21 | 144.76 | 149.36 |
| RFR | 13.65 | 41.93 | 77.77 |
| KNR | 16.25 | 57.36 | 92.61 |
| CBR | 11.56 | 34.14 | 65.88 |

После обработки пиков, результаты всех моделей улучшились. Заметим, что:

* Лучшее качество показала модель CatBoostRegressor

Далее подберем оптимальные гиперпараметры для данной при помощи библиотека optuna. Получили следующие результаты:

Таблица 4

Результаты CatBoostRegressor на подобранных оптимальных гиперпараметрах

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | WAPE | BIAS | MAE |
| CBR | 9.94 | 0.81 | 56.65 |

Обученная модель CatBoostRegressor на лучших гиперпараметрах, подобранных через *optuna* показала следующие результаты:

1. WAPE равен 9.94%, что означает, что модель ошибается в прогнозировании результатов на 9.94% в среднем.
2. MAE равен 56.65, что указывает на среднюю абсолютную ошибку в прогнозах.
3. BIAS равен 0.81, что указывает на небольшое смещение модели в предсказаниях.

С учетом полученных значений, можно сделать вывод, что модель имеет среднюю ошибку (MAE) в размере 56.65, но при этом имеет небольшое смещение (BIAS) и сравнительно приемлемую абсолютную процентную ошибку (WAPE). Модель, скорее всего, хорошо предсказывает данные, однако возможно требуется дополнительный анализ для определения причин возможных ошибок и возможных улучшений.

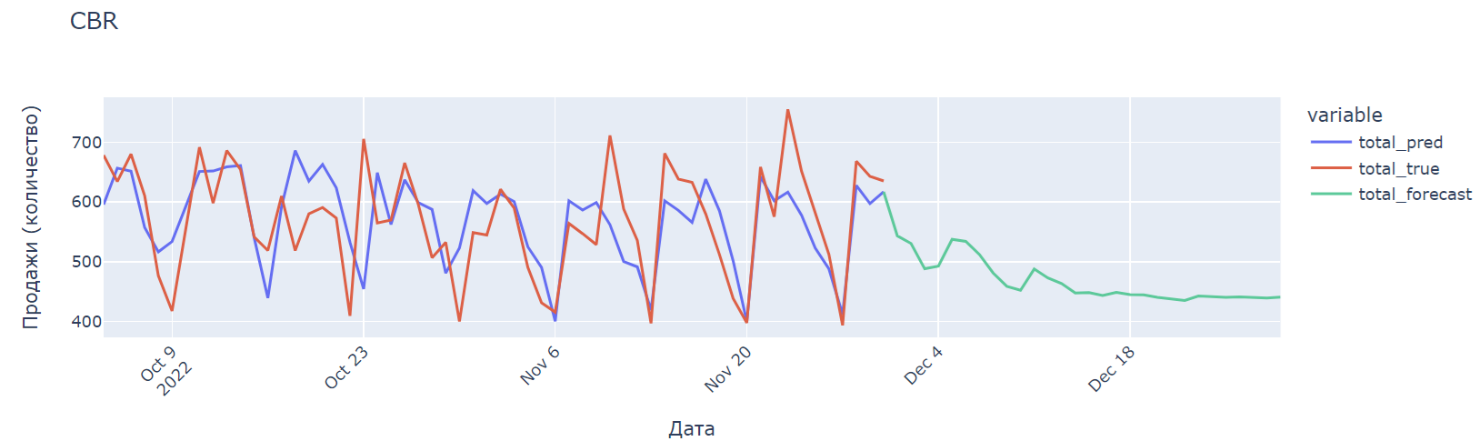


Рис. 1. Визуализация прогноза продаж товара.

Для данной модели написан класс для прогнозирования продаж товаров из ассортимента аптек на основе временных рядов, в который на вход принимает название препарата и количество дней прогноза, которые хотим предсказать. На выходе имеем данный ряд прогнозирования с метриками качества.

Производим проверку построенной модели на других товаров аптек. Результаты приведены в таблице.

Таблица 5

Результаты по препаратам

|  |  |  |  |
| --- | --- | --- | --- |
| Препарат | Метрики | | |
| MAE | WAPE | BIAS |
| Супрастин | 56.65 | 9.94 | 0.81 |
| ТЕРАФЛЮ | 213.65 | 7.31 | 55.60 |
| ИНГАВИРИН | 48.67 | 16.41 | 17.19 |
| НУРОФЕН | 162.07 | 20.71 | 53.33 |
| АРБИДОЛ | 118.88 | 15.30 | -70.47 |
| МИРАМИСТИН | 103.36 | 17.89 | -90.75 |
| НИМЕСИЛ | 170.07 | 8.13 | -118.32 |
| КАГОЦЕЛ | 82.87 | 24.31 | -78.66 |
| МАКСИКОЛД | 171.73 | 10.35 | -51.92 |
| КЕТАНОВ | 58.08 | 10.14 | -19.70 |

Из полученных результатов можно сделать следующий вывод:

1. Общий уровень точности модели:

* MAE для большинства лекарственных препаратов находится в диапазоне от 50 до 200. Это означает, что модель имеет значительную ошибку (на порядок) при прогнозировании спроса. Однако, стоит отметить, что и продажи некоторых товаров отличаются в объёме на порядок.
* WAPE в основном находится в диапазоне от 7 до 25. Это указывает на средний уровень точности, но некоторые модели (например, "КАГОЦЕЛ", "НУРОФЕН") демонстрируют высокую ошибку.
* Значения BIAS показывают, что модели имеют как положительную, так и отрицательное смещение. Это означает, что в некоторых случаях модели завышают прогноз, а в других – занижают.

1. Сравнение показателей по препаратам:

* "НИМЕСИЛ", "КЕТАНОВ": данные препараты относятся к группе *НПВП*, демонстрируют приемлемый уровень WAPE и MAE (при соответствующим уровне продаж), что свидетельствует о лучшей точности прогнозирования. У "НИМЕСИЛ" наблюдается сильно отрицательное смещение, что может привести к потери недополученной прибыли и снижению трафика, но это не приведет к тяжелым последствиям, так как аналог - препарат "КЕТАНОВ" показывает небольшое смещение относительно нуля.
* "ТЕРАФЛЮ", "МАКСИКОЛД": данные препараты относятся к группе *Комбинированных препаратов*, демонстрируют приемлемый уровень WAPE и MAE, что свидетельствует о лучшей точности прогнозирования. Смещение данных товаров по модулю находится на одном уровне, так что препарат всегда можно заменить аналогом.
* "КАГОЦЕЛ": данный препарат относится к группе *интерфероногены*, показывает высокий WAPE, что указывает на наихудшую точность модели.
* "АРБИДОЛ", "ИНГАВИРИН": данные препараты относятся к группе *ППС*, точность прогнозирования находится в среднем диапазоне.
* "СУПРАСТИН": данный товар относится к группе *ПСh1*, демонстрируют приемлемый уровень WAPE и MAE, модель хорошо предсказывает данные, показывая практически нулевое смещение.
* "НУРОФЕН": данный товар относится к группе *НПС*, показывает высокий WAPE, что указывает на неприемлемую точность модели.
* "МИРАМИСТИН": данный товар относится к группе *КА*, показывает высокий WAPE, что указывает на неприемлемую точность модели.

1. Важные наблюдения по BIAS:

* "КАГОЦЕЛ", "МИРАМИСТИН", "НИМЕСИЛ": у данных препаратов наблюдается отрицательное смещение, что означает, что модели, скорее всего, занижают прогноз спроса.
* "ТЕРАФЛЮ", "ИНГАВИРИН", "НУРОФЕН": положительное смещение для этих препаратов указывает на то, что модель, скорее всего, завышает прогноз спроса.

1. Необходимо дальнейшее исследование:

* Причины высокого MAE и WAPE: необходимо выяснить, какие факторы влияют на точность прогнозирования для препаратов с высокой ошибкой. Возможно, это связано с сезонностью спроса, недостаточным количеством данных или другими факторами.
* Поведение BIAS: важно понять, почему модели имеют смещение. Возможно, это связано с особенностями данных, алгоритмами модели или другими факторами.

Полученные результаты указывают на то, что модель при прогнозировании спроса на лекарства не универсальна и на ряде препаратов имеет значительную ошибку. По данной выборке можно предположить, что модель дает приемлемое качество на препараты, относящиеся к группам *НПВП*, *Комбинированных препаратов*, *ПСh1*.

Для товаров группы *интерфероногены*, *ППС*, *НПС*, *КА* необходимо провести дальнейший анализ, чтобы понять причины высокой ошибки и разработать более точную систему прогнозирования.

**Заключение**

В процессе работы над проектом была изучена литература, проведены эксперименты и получены следующие результаты:

1. Построена модель CatBoostRegressor, обученная на оптимальных гиперпараметрах, демонстрирующая оптимальную точность прогнозирования спроса на лекарственные препараты. Её ключевые показатели:

* В среднем WAPE составляет 9.94%. Это указывает на приемлемый уровень точности, но для некоторых препаратов ошибка может быть значительно выше.
* MAE равна 56.65, что указывает на среднюю абсолютную ошибку в прогнозах. Важно учитывать масштаб продаж каждого препарата при интерпретации MAE.
* BIAS равен 0.81, что указывает на небольшое смещение модели. Смещение может быть как положительным, так и отрицательным, что влияет на принятие решений по запасам и сбыту.

1. Проведен показатель по препаратам для данной модели, который разделен по категориям (*НПВП, комбинированные препараты, интерфероногены, ППС, ПСh1, НПС, КА), что позволяет выявить тенденции в точности модели для разных групп препаратов:*

* Высокая точность: препараты "НИМЕСИЛ", "КЕТАНОВ", "ТЕРАФЛЮ", "МАКСИКОЛД", "СУПРАСТИН" демонстрируют приемлемый уровень WAPE и MAE, что свидетельствует о хорошей точности прогнозирования.
* Средняя точность: для препаратов "АРБИДОЛ" и "ИНГАВИРИН" точность прогнозирования находится в среднем диапазоне.
* Низкая точность: "КАГОЦЕЛ", "НУРОФЕН" и "МИРАМИСТИН" имеют высокий WAPE, что указывает на неприемлемую точность модели.

1. Полученная модель может быть интересна:

* Управление запасами: прогнозирование общих продаж по названию товара может быть полезным для управления запасами и планирования поставок. Это позволяет оптимизировать уровень запасов и ресурсов в магазинах или на складах.
* Партнерство с поставщиками: прогнозирование общих продаж может быть полезным для руководства взаимоотношениями с поставщиками, помогая им лучше понять спрос на их товары и соответственно управлять производством и поставками.

Модель имеет средний уровень точности, но для некоторых препаратов может быть значительной. Необходимо провести дополнительные исследования для улучшения точности прогнозирования препаратов с высокой ошибкой, также необходимо выявить причины смещения модели и устранить их, например:

* провести дополнительный анализ на учёт сезонности спроса на препараты с высокой ошибкой – включить в модель сезонные компоненты, используя сезонные индексы, тригонометрические функции или другие методы;
* провести анализ тренда: выявить тенденции спроса и включить её в модель;
* рассмотреть возможность использования других моделей машинного обучения или статистических методов.

Всю информацию о проделанной работе можно найти в открытом доступе в репозитории проекта[[1]](#footnote-1).

Модель CatBoostRegressor представляет собой хороший инструмент для прогнозирования спроса на лекарственные препараты, но ее точность может быть улучшена. Дополнительные исследования и оптимизация модели позволят повысить ее точность и эффективность.

**Список Литературы**

1. Russell Davidson, James G. MacKinnon; [Estimation and Inference in Econometrics](https://www.researchgate.net/publication/227466701_Estimation_and_Inference_in_Econometrics); 1993.
2. James D. Hamilton; [Time Series Analysis](http://mayoral.iae-csic.org/timeseries2021/hamilton.pdf); 1994.
3. Peter C.B. Phillips, Pierre Perron; [Testing for unit roots in time series regression](https://www.researchgate.net/publication/4744135_Testing_for_a_Unit_Root_in_Time_Series_Regression); 1986.
4. Hyndman, R.J., & Athanasopoulos, G.; [Forecasting: principles and practice](https://otexts.com/fpp3/); 2018.
5. Chatfield, C.; [The Analysis of Time Series: An Introduction](https://www.researchgate.net/publication/327207744_The_Analysis_of_Time_Series_An_Introduction_Sixth_Edition); 2003.
6. Makridakis, S., Wheelwright, S.C., & Hyndman, R.J.; [Forecasting: methods and applications](https://www.researchgate.net/publication/52008212_Forecasting_Methods_and_Applications); 1998.
7. Maheshkumar R Sabhnani, Daniel B Neill, Andrew W Moore; [Anomaly detection in pharmacy sales data using statistical methods](https://www.researchgate.net/publication/228337469_Detecting_anomalous_patterns_in_pharmacy_retail_data); 2005.
8. Петрова С. В., Кононова С. В., Дадус Н. Н., Чеснокова Н. Н., Жукова Е. В.; [Факторы эффективных продаж в аптечном бизнесе](https://cyberleninka.ru/article/n/faktory-effektivnyh-prodazh-v-aptechnom-biznese); 2012.

1. <https://github.com/m-ardat/Time-Series---Sales-prediction> [↑](#footnote-ref-1)