


**ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

Факультет компьютерных наук
Образовательная программа «Программная инженерия»

УДК 339.133.017

СОГЛАСОВАНО

Приглашенный преподаватель факультета
компьютерных наук



_____ А. А. Романенко
«18» _05_____ 2023 г.

УТВЕРЖДАЮ

Академический руководитель
образовательной программы

«Программная инженерия»,

профессор департамента программной
инженерии, канд. техн. наук


_____ В. В. Шилов
«_18_» _05_____ 2023 г.

**Выпускная квалификационная работа
(академическая)**


на тему: **Исследование методов сегментации данных о товарах для
обучения моделей прогнозирования спроса**

по направлению подготовки 09.03.04 «Программная инженерия»

ВЫПОЛНИЛ

студент группы БПИ194
образовательной программы
09.03.04 «Программная инженерия»

_____ Савинов М.Г.
И.О. Фамилия


_____ 18.05.2023
Подпись, Дата

Москва 2023

РЕФЕРАТ

Ключевые слова: *спрос; цена; сегментация; модель прогнозирования спроса; лаговые переменные;*

Работа содержит 33 страницы, 3 главы, 4 рисунка, 8 таблиц, 8 источников.

В отчете представлены результаты преддипломной практики на тему «Исследование методов сегментации данных о товарах для обучения моделей прогнозирования спроса», выполненного на основании учебного плана подготовки бакалавров по направлению 09.03.04 "Программная инженерия" и приказа декана факультета компьютерных наук И.В. Аржанцева о закреплении тем ВКР за студентами и назначении руководителей и консультантов ВКР № 2.3-02/151222-7 от декабря 2022 года

Объект исследования – модель прогнозирования спроса, построенная на данных о продажах ритейл-сети за 2016-2019 года.

Предмет исследования – сегментация данных по различным признакам для улучшения качества работы модели.

Цель исследования – изучение возможность применения сегментации в моделях прогнозирования спроса для повышения точности прогнозов.

Задачи исследования:

- Составление обзора статей о данных по продажам и переменных, которые используются для его предсказания.
- Генерация лаговых переменных для модели прогнозирования спроса.
- Разработка модели прогнозирования спроса на всех имеющихся данных на языке Python в среде разработки Jupyter Notebook.
- Выбор признаков, на основе которых можно реализовать методы сегментации.
- Разбитие данных на базовые сегменты по имеющимся полям и на сегменты с использованием метода агломеративной кластеризации.
- Построение моделей прогнозирования спроса для каждого набора сегментов, полученных различными методами кластеризации на языке Python в среде разработки Jupyter Notebook.
- Проведение сравнительного анализа моделей с использованием различных способов сегментации и без нее.
- Анализ результатов, полученных в ходе проведения сравнительного анализа, расчет метрик качества модели, на основе которого можно сделать вывод о необходимости

использования алгоритмов сегментации в моделях прогнозирования спроса, а также определить наилучший из методов сегментации.

Методы исследования:

- Изучение публикаций и статей.
- Реализация базовых методов сегментации, а также сегментации с использованием алгоритмов кластеризации.
- Проведение вычислительного эксперимента.
- Расчёт статистических ошибок

Научная новизна:

- В настоящее время при прогнозировании спроса не используются методы сегментации, в исследовании проведен анализ о возможности применения этих методов в моделях прогнозирования спроса.
- Проведен сравнительный анализ эффективности работы моделей прогнозирования с использованием сегментации и без него.

Достоверность научных результатов подтверждена результатами экспериментальных исследований на базе разработанной программной реализации.

Практическая значимость. В отличие от стандартного подхода построения одной модели на всех данных о продажах предложен вариант построения нескольких моделей на отдельных выделенных сегментах, что улучшает качество моделей прогнозирования спроса.

Результаты работы

- Изучены данные о продажах ритейл сети и на их основе сгенерированы переменные для моделей прогнозирования спроса.
- Проанализированы поля на основе, которых возможно применение сегментации в моделях прогнозирования спроса.
- Проведены вычислительные эксперименты, рассчитаны статистические ошибки для сравнения эффективности работы модели прогнозирования спроса на всех данных и с построением моделей на каждой группе сегментов отдельно.
- На основании вычислительного эксперимента сделан вывод о возможности применения сегментации для прогнозирования спроса.

ABSTRACT

Demand forecasting models for products are very widespread among various retail companies. However, most of them use the concept of segmentation in a very narrow form at the level of a specific product. This paper describes various approaches of data segmentation by different parameters, which allows to build demand forecasting models for each segment separately to improve prediction accuracy. The article also describes a comparative analysis between models for demand forecasting for retail network data using segmentation methods and without using them. Thus, it will be possible to gain an understanding of how effective the segmentation method is in demand forecasting.

The work contains 33 pages, 3 chapters, 4 figures, 8 tables, 8 bibliography items.

Keywords— demand; price; segmentation; demand forecasting model; lagged features.

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

Спрос – количественный признак, характеризующийся количеством приобретенного товара в конкретную дату.

Сегментация – процесс разбивки пар товар-магазин на различные группы (или сегменты, кластеры), в рамках которых все пары имеют схожие или аналогичные качества, или свойства.

Лаговые переменные – это значения на предыдущих временных шагах, которые считаются полезными, поскольку они созданы на основе событий произошедших в прошлом и влияют на появление такого события в будущем. Являются основными переменными для прогнозирования спроса, так как включают в себя полную агрегированную картину истории продаж.

Оглавление

ВВЕДЕНИЕ	7
ГЛАВА 1. ПОСТРОЕНИЕ МОДЕЛИ ДЛЯ ПРОГНОЗИРОВАНИЯ СПРОСА	10
1.1. Подготовка данных для построения модели	10
1.2. Генерация лаговых переменных	11
1.3. Обучение модели на всех данных	12
ГЛАВА 2. Методы обучения модели с использованием сегментации объектов	16
2.1 Сегментация данных по базовым переменным	16
2.2 Сегментация с использованием методов машинного обучения	17
2.2.1 Агломеративная кластеризация на основе базовых полей	19
2.2.2 Агломеративная кластеризация магазинов на основе данных о товарах	22
ГЛАВА 3. Сравнительный анализ	28
ЗАКЛЮЧЕНИЕ	31
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	33

ВВЕДЕНИЕ

Спрос на продаваемые товары является одним из главных основополагающих факторов для выставления корректных цен и производства нужного объема товаров, как следствие спрос напрямую влияет на прибыльность ритейл-компаний. Правильное предсказание спроса является важнейшим элементом успешной стратегии бизнеса. Если спрос на продукцию или услуги недооценен, то компания может потерять потенциальных клиентов и доходы. Если спрос переоценен, то это может привести к излишнему производству, что увеличит затраты на производство и складирование товаров. Кроме того, правильное предсказание спроса помогает оптимизировать производственные процессы, планировать закупки и управлять запасами. В целом, правильное предсказание спроса позволяет компании удерживать конкурентные позиции на рынке, увеличивать выручку и максимизировать прибыль. Актуальность моделей для прогнозирования спроса растет с огромной скоростью, большинство крупных ритейл-сетей уже используют методы машинного обучения для корректной оценки потребления товаров.

В большинстве случаев из-за нехватки ресурсов компании используют одну модель прогнозирования спроса для всех товаров и магазинов. Такая модель имеет достаточно высокое качество из-за больших объемов, накопленных данных и показывает приемлемые результаты по различным метрикам.

Однако использование одной и той же модели для различных типов товара, полностью отличающихся друг от друга является не совсем корректным. Например, в некоторых магазинах гораздо большая проходимость, что несомненно напрямую влияет на спрос. Для некоторых товаров данные о продажах в корне отличаются друг от друга, так одни товары являются продуктами повседневного спроса, а другие более завязаны под конкретные сегменты людей.

В дополнение, зачастую огромный объем входных данных требует использования огромного количества вычислительных ресурсов, в свою очередь использование нескольких моделей, построенных на определенных сегментах позволяет рассчитывать прогнозы параллельно.

В данной работе будет представлен подход использования нескольких моделей для прогнозирования спроса на различных сегментах. Будут проанализированы различные базовые переменные, такие как тип товара или расположение магазина, в котором он продается. Также будет проанализирована возможность применения кластеризации с использованием алгоритмов машинного обучения. На каждом из наборов выделенных

сегментов будет построена своя модель предсказания спроса, что позволит более четко учитывать влияние второстепенных признаков при определении спроса.

Сегментация в моделях прогнозирования спроса нужна для более точного прогнозирования спроса для конкретных товаров в различных группах потребителей. Путем разбиения данных на сегменты можно выявить различия в зависимости спроса от различных категорий, типов товаров, расположении магазинов, цен в определенной период времени, что позволяет более точно определить факторы, влияющие на спрос в каждом сегменте. Подобные различия можно будет выявить при помощи метода построения отдельной модели прогнозирования спроса для каждого из выделенных сегментов.

Качество проведенных исследований будет проанализировано с помощью сравнительного анализа с базовой моделью прогнозирования спроса. Будет рассчитана метрика RMSE (Root Mean Square Error) для модели, обученной без использования сегментации, а также для каждого из объединения моделей с использованием методов сегментации. В завершении эксперимента можно будет установить является ли использование сегментации актуальным в рамках настоящей задачи и насколько использованием сегментации улучшает качество итоговых моделей прогнозирования спроса.

Таким образом, цель данной работы – исследование и разработка подходов к прогнозированию спроса на основе сегментации объектов ритейл-сети, применение этих методов и дальнейшее сравнение с базовой моделью.

Для анализа использовалась база данных о продажах продуктов ритейл-сети, состоящая из 40 различных магазинов, расположенных в нескольких городах и 36000 продаваемых в них продуктов. Исследование проводилось в разрезе Товар | Магазин | День на временном промежутке с 2016 по 2019 год, таким образом всего выборка включает в себя 12 миллионов различных значений о продажах.

Отчет организован следующим образом. В главе 1 представлено построение модели прогнозирования спроса, большое внимание уделено преобразованию сырых данных о продажах к переменным, на основе которых будет производится прогнозирование, также обоснован выбор используемой модели – градиентного бустинга и временных промежутков для построения и тестирования модели. Глава 2 содержит описание алгоритмов прогнозирования с помощью кластеризации на основе базовых переменных, а также, алгоритмов, использующих методы машинного обучения для определения сегментов, приведено применение этих алгоритмов для обучения модели и рассчитаны статистические

ошибки. В главе 3 представлены результаты вычислительных экспериментов и сравнение с существующим алгоритмом прогнозирования. В заключении перечислены выполненные задачи, обсуждаются результаты экспериментов и пути дальнейшей работы.

ГЛАВА 1. ПОСТРОЕНИЕ МОДЕЛИ ДЛЯ ПРОГНОЗИРОВАНИЯ СПРОСА

1.1. Подготовка данных для построения модели

Для построения модели прогнозирования спроса прежде всего необходимо подготовить данные для обучения. Исходные данные, которые были предложены являются сырыми данными о продажах, в их состав входит множество незначимых и бесполезных переменных, поэтому первым шагом нужно привести их к агрегированному виду.

Каждая пара товар-магазин в исходных данных имеет несколько категориальных переменных описывающих эту пару, основные из них представлены в таблице.

Таблица 1. Категориальные переменные исходных данных

LOCATION		PRODUCTS	
Поле	Описание	Поле	Описание
STORE_LOCATION_RK	Номер магазина	PRODUCT_RK	Номер товара
STORE_LOCATION_RK4	Населённый пункт	PRODUCT_LVL_RK2	Категория товара
STORE_LOCATION_RK3	Город	PRODUCT_LVL_RK3	Подкатегория товара
STORE_LOCATION_RK2	Экономический регион	PRODUCT_LVL_RK5	Коллекция товара
		PRODUCT_LVL_RK6	Линейка товара

Эти переменные являются необходимыми для будущей сегментации, но также включаются в модель предсказания на всех данных.

Исходные данные также включают в себя цены на товар в различные промежутки времени, флаг продавался ли товар по скидке в текущий момент, и если да, то цену на этот товар, а также данные о спросе, который будет являться целевой переменной для будущей модели.

Однако, подавляющее большинство сырых исходных данных не являются информативными и на их основе нельзя строить необходимые в нашей задаче прогнозы, поэтому как уже было описано выше основными признаками, которые используются при построении моделей прогнозирования спроса будут являться лаговые переменные [1], они строятся на основе исходных данных о продажах и позволяют агрегировать сырые данные к необходимому для построения модели виду.

1.2. Генерация лаговых переменных

Лаговые переменные - это значения переменных, которые были измерены в прошлом и используются для прогнозирования будущих значений. В контексте моделей прогнозирования спроса, лаговые переменные могут быть предыдущими значениями спроса на товары, которые используются для прогнозирования будущих значений спроса. Лаговые переменные могут также включать данные о ценах, доходах, погоде и других факторах, которые могут влиять на спрос. Использование лаговых переменных позволяет учитывать исторические тенденции и изменения в поведении потребителей при прогнозировании будущих значений спроса. Другими словами, они применяются в моделях прогнозирования спроса для учета прошлых значений признаков и позволяют учитывать динамику изменения признаков во времени для использования этой информации для прогнозирования будущих значений. Поэтому для предсказания о продаже товара на будущий временной промежуток, используются именно лаговые переменные для учета продажи за предыдущие месяцы.

Таким образом, можно увидеть, как меняется спрос на товар со временем и использовать эту информацию для более точного прогнозирования будущих продаж.

При генерации лаговых переменных есть 2 основных фактора, на которые стоит обращать внимание: сдвиг и скользящее окно [2]. Сдвиг в лаговых переменных означает использование значений переменных, измеренных в прошлом, но с определенным временным отставанием. Например, если мы хотим прогнозировать спрос на товар в следующем месяце, мы можем использовать значения спроса за предыдущие месяцы как лаговые переменные. Однако, чтобы учесть сезонность или другие временные факторы, мы можем использовать значения спроса за тот же месяц в прошлых годах, сдвигая лаговые переменные на определенное количество месяцев или лет. Сдвиг в лаговых переменных позволяет учитывать изменения в поведении потребителей и внешних факторов во времени при прогнозировании будущих значений. Скользящее окно в лаговых переменных означает использование определенного количества последних значений переменной в качестве лаговых переменных. При этом каждый новый месяц добавляется в скользящее окно, а самое старое значение удаляется. Скользящее окно позволяет учитывать последние изменения в поведении потребителей и внешних факторов при прогнозировании будущих значений, и может помочь улучшить точность прогнозов. Таким образом, при значении сдвига равным 4 неделям, и значению окна равна 4 неделям, лаговая переменная будет рассчитана на данных за предыдущие 4 недели от даты, которая была 4 недели назад.

При генерации лаговых переменных использовались такие агрегаты, как среднее значение за промежуток, заданный окном, и медианное значение за этот промежуток. Дополнительно использовался фильтр по флагу акции на товар, то есть отдельная лаговая переменная без фильтров и отдельная переменная с использованием фильтра.

Таким образом при заданных значениях сдвига равных 4, 8, 26 и 52 неделям, значениям скользящего окна равных 4, 8, 26, 52 неделям, двум агрегатам – среднему значению и медиане и 4 различными фильтрами – 3 флага акции и без фильтров было сгенерировано 128 лаговых переменных.

1.3. Обучение модели на всех данных

Таким образом для построения итоговой модели прогнозирования спроса были получены 149 признаков (лаговые переменные, категориальные переменные, цены и флаги акций), количество уникальных пар товар-магазин в гранулярности Товар | Магазин | День равно 12 миллионам.

Для моделей прогнозирования спроса чаще всего используется градиентный бустинг, так как набор признаков для обучения достаточно большой, а также данная модель является логической и может находить нелинейные зависимости в данных.

В настоящем исследовании был применен CatBoost [3], который имеет несколько преимуществ перед другими алгоритмами градиентного бустинга:

- Может работать с категориальными признаками без необходимости преобразовывать их в числовые значения.
- Автоматически обрабатывает пропущенные значения в данных.
- Использует специальный алгоритм для обработки выбросов (аномалий), что позволяет улучшить качество модели.
- Обучение модели CatBoost происходит быстрее, чем у многих других алгоритмов градиентного бустинга, с учетом выборки в 12 миллионов данных – это очень существенное преимущество.

В качестве основной метрики оценки качества предсказания была использована статистическая ошибка RMSE [4] - Root Mean Square Error:
$$\sqrt{\sum_{i=1}^n \frac{(x_i - y_i)^2}{n}}.$$

Временной промежуток для обучения модели был выбран с начала данных (2016 год) до 01.10.2019. Таким образом для тестирования качества модели использованы 3 последних месяца выборки октябрь, ноябрь и декабрь. Такой выбор обусловлен тем, что основными переменными для обучения являются лаговые переменные, для которых очень важно соблюдать глубокую историчность данных, кроме того спрос является очень нестабильной целевой переменной и резко реагирует на различные изменения. Таким образом в обучающей выборке использовалось порядка 11 миллионов данных. Для проверки необходимости использования такого количества данных был проведен небольшой вычислительный эксперимент зависимости качества обучения модели от размера обучающей выборки. На рисунках ниже можно наблюдать зависимость метрики RMSE на тестовой выборке от количества обучающих данных (по шкале X – доля данных от общего количества, по шкале Y – метрика RMSE и MAE соответственно).

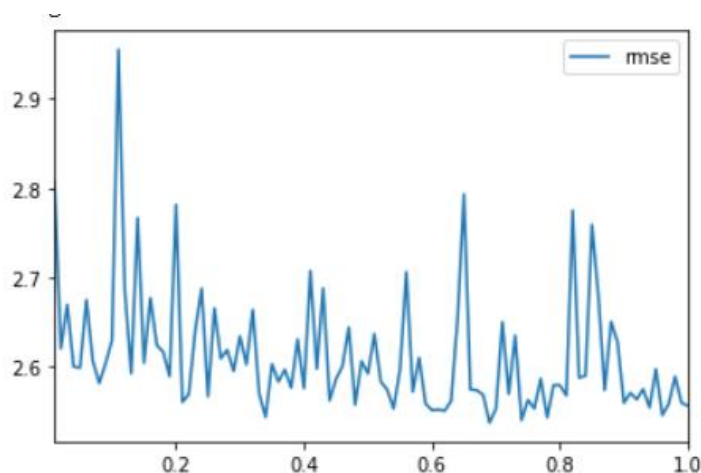


Рисунок 1. Зависимость качества модели по метрике RMSE от объема обучающей выборки

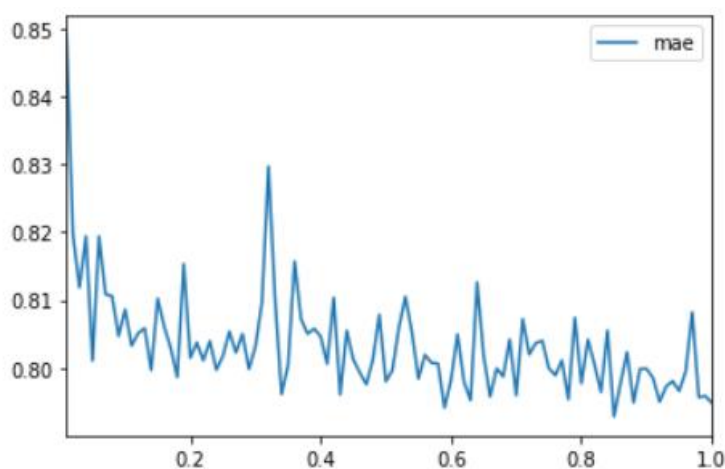


Рисунок 2. Зависимость качества модели по метрике MAE от объема обучающей выборки

Для подбора оптимальных параметров для построения модели был использован алгоритм GridSearch [5] - метод оптимизации параметров модели машинного обучения путем перебора различных комбинаций значений параметров и выбора тех, которые дают наилучшие результаты на заданном наборе данных. Gridsearch позволяет автоматизировать процесс подбора оптимальных параметров, что может значительно ускорить и улучшить процесс обучения модели. В качестве основной метрики оценивания также была использована метрика RMSE. Благодаря использованию данного метода удалось подобрать оптимальные значения для параметров depth и learning rate, которые составляют 8 и 0.1 соответственно, а также улучшить качество основного прогноза тестовой выборке на 3%.

На рисунке ниже представлены основные значимые признаки, вошедшие в модель на всех переменных. Как можно заметить большое влияние имеют также именно категориальные переменные, такие как PRODUCT_RK, PRODUCT_LVL_RK2, PRODUCT_LVL_RK3, STORE_LOCATION_RK. Именно по ним будет происходить дальнейшее разделение по сегментам.

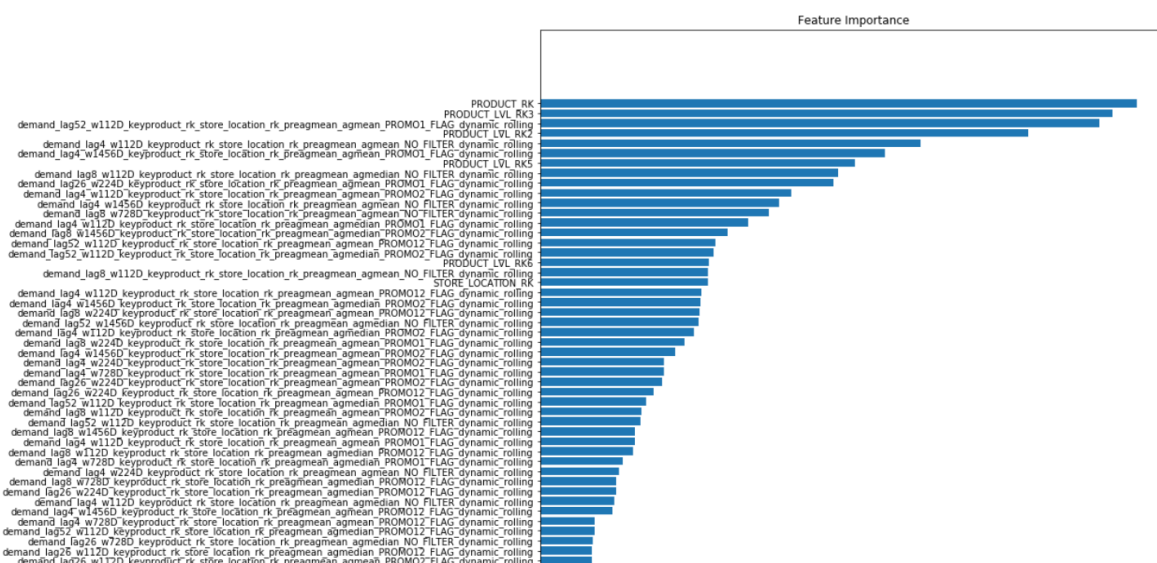


Рисунок 3. Значимость признаков в модели на всех данных

Остальные переменные такие как цены, флаги акции и более детальные данные о расположении магазинов не имеют большой значимости при обучении основной модели, поэтому их в качестве возможных переменных для сегментации в данной работе мы не будем. Такие переменные как PRODUCT_LVL_RK5, PRODUCT_LVL_RK6 являются расширением данных о переменных PRODUCT_RK, PRODUCT_LVL_RK2, PRODUCT_LVL_RK3, однако так как основная суть этих признаков одинакова, то для сегментации были использованы 3 наиболее важных переменные PRODUCT_RK – номер товара, PRODUCT_LVL_RK2 – категория товара, PRODUCT_LVL_RK3 – подкатегория товара.

ГЛАВА 2. Методы обучения модели с использованием сегментации объектов

2.1 Сегментация данных по базовым переменным

Первым подходом, который был проанализирован - это сегментация всех данных на основе, готовых категориальных переменных, а именно `STORE_LOCATION_RK` (номер магазина), `PRODUCT_LVL_RK3` (подкатегория товара), `PRODUCT_LVL_RK2` (категория товара). Эти переменные имеют высокую значимость в модели, обученной на всех данных, поэтому гипотеза об их использовании при алгоритмах сегментации является полностью обоснованной.

Такая сегментация является интуитивно-понятной, так как для большинства категорий, подкатегорий товаров зависимость спроса от времени может очень сильно различаться. Аналогичная ситуация для магазинов, если магазин находится не в очень людном месте, то спрос в нем логично будет меньше, поэтому его возможно можно выделить в отдельный сегмент.

Самым значимым признаков в модели прогнозирования спроса является `PRODUCT_RK` (наименование магазина), однако количество различных значений по этому признаку около 36 тысяч, что делает бессмысленным сегментацию на основе этого поля, ведь при обучении модели на каждом из таких сегментов просто не будет хватать данных для качественного обучения.

Таким образом выделим 3 основных поля для сегментации по базовому принципу:

- `STORE_LOCATION_RK` – 40 уникальных значений
- `PRODUCT_LVL_RK2` – 22 уникальных значения
- `PRODUCT_LVL_RK3` – 33 уникальных значений

Для каждого кластера обучим свою модель прогнозирования спроса (период обучения и тестирования, метод построения модели и входные переменные аналогичны модели на всех продуктах). Итого для первого случая получается 40 моделей, для второго 22 модели и для третьего 33 модели соответственно.

Далее необходимо провести расчет метрики `RMSE` для проверки качества для этого добавим предсказанные значения в изначальную выборку.

После всех этих операций можно посчитать метрику `RMSE` на тестовой выборке относительно целевой переменной спроса. Для этого объединим спрогнозированный спрос моделью на каждом из сегментов в одну общую выборку.

Итоговое качество моделей с использованием сегментации на аналогичном модели на всех данных тестовом промежутке составляет:

- По STORE_LOCATION_RK: 2.701105
- PRODUCT_LVL_RK2 – 2.135943
- PRODUCT_LVL_RK3 – 2.590263

Таким образом сегментация по наименованию магазина и подкатегориям не дает прироста в качестве модели, однако модель с использованием сегментации по категориям дает значительный прирост в метрике RMSE более чем 15%. Однако сегментация по базовым признакам исходных данных является достаточно примитивным подходом, она не учитывает, что продукты из разных категорий могут иметь похожий спрос, хотя такое на практике бывает достаточно часто. Именно поэтому было принято решение попробовать методы сегментации на основе методов машинного обучения, а именно агломеративную кластеризацию.

2.2 Сегментация с использованием методов машинного обучения

Существует огромное множество способов кластеризации данных, такие как например алгоритм K-Means [6], EM-алгоритм [7], алгоритм DBSCAN [8], однако проанализировав все варианты было принято решение использовать именно метод агломеративной кластеризации.

Агломеративная кластеризация - это метод кластеризации, который основывается на объединении более мелких кластеров в более крупные, пока не останется только один кластер, содержащий все объекты.

Алгоритм начинает работу с создания отдельных кластеров, содержащих только один объект. Затем на каждой итерации два ближайших кластера объединяются в один, пока не останется единственный кластер, содержащий все объекты. В качестве меры близости между кластерами используются различные расстояния, например, евклидово расстояние или косинусное расстояние.

Одним из основных преимуществ агломеративной кластеризации является ее простота и возможность визуализации результатов. Кроме того, агломеративная кластеризация является одним из наиболее точных методов кластеризации, поскольку использует все имеющиеся данные в процессе кластеризации.

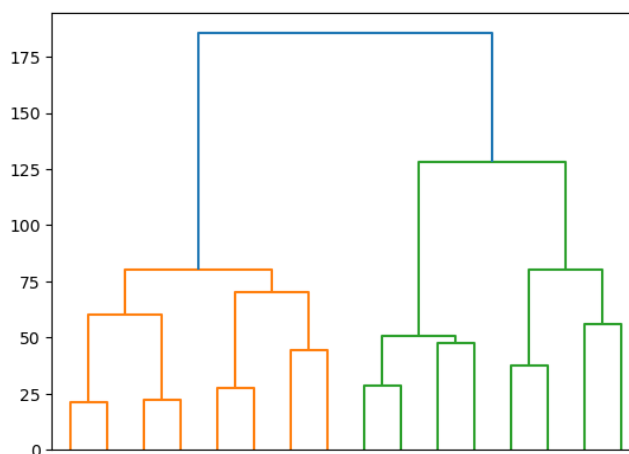


Рисунок 4. Пример работы агломеративной кластеризации для трех сегментов

В качестве значения, на основе которого выборка будет разделяться на сегменты было использовано среднее значение спроса по полю, по которому происходило сама кластеризация.

Для использования агломеративной кластеризации данные также необходимо разделить на обучающую и тестовую выборки. Аналогично основной модели, определение кластеров тестовой выборки (октябрь – ноябрь – декабрь 2019 года) будет сделано на основе предсказаний модели кластеризации, обученной на обучающей выборке (до октября 2019 года), чтобы не использовать данные из будущего при разделении выборки на сегменты.

Так как модель кластеризации не использует в обучении данные тестовой выборки, то если в тестовой выборке находятся пары товар-магазин, которых не было до октября 2019 года, например, если товар начал продаваться в магазине только с октября 2019 года, то такие пары не будут определены ни в один кластер, а значит итоговая модель прогнозирования спроса не сможет рассчитать спрос для них. Для решения этой проблемы в задачах сегментации обычно считают прогноз на основе модели, обученной на всех данных, однако в нашем случае ни для одной пары товар-магазин такого не наблюдается, а значит тестовые выборки моделей на сегментах и модели на всех данных без сегментации полностью совпадают, что позволяет проводить сравнение качества построенных моделей.

Идея использования алгоритмов сегментации на основе машинного обучения, относительно среднего спроса заключается в поиске зависимостей между товарами и магазинами, которые относятся к разным категориям или подкатегориям, однако имеют схожий характер поведения спроса во времени. При использовании метода

агломеративной кластеризации в один сегмент могут попасть товары из разных категорий, чего никак не может произойти при сегментации по базовым категориальным полям, описанным выше. Такой подход расширяет возможности применения сегментации и позволяет уловить многие зависимости, которые не учитываются в предыдущем подходе, описанном в предыдущем разделе.

Также при использовании кластеризации на основе методов машинного обучения мы в праве сами задавать и количество необходимых для разбиения на сегменты кластеров, что также является некоторым преимуществом.

2.2.1 Агломеративная кластеризация на основе базовых полей

Первой итерацией было использование агломеративной кластеризации для базовых полей аналогичных пункту 2.1., то есть относительно, `PRODUCT_LVL_RK2` (категория товара), а также `PRODUCT_RK` (наименование магазина). Сегментация относительно `STORE_LOCATION_RK` и `PRODUCT_LVL_RK3` не имеет смысла, так как из пункта того же пункта 2.1 настоящего документа можно сделать вывод, что сегментация относительно этих полей не дает добавки к модели, более того сегментация по этим полям ухудшает итоговое качество прогнозирования.

1. Кластеризация по `PRODUCT_RK`.

Как уже было озвучено выше `PRODUCT_RK` является одним из самых значимых признаков в базовой модели, поэтому сегментация по этому полю может значительно повлиять на качество исходной модели. Так как различных значений поля `PRODUCT_RK` более 36000, то оптимальное количество кластеров необходимых для наибольших добавок по этому признаку нельзя просто выбрать случайно. Для решения этой проблемы был проведен вычислительный эксперимент с перебором различного количества кластеров для сегментации по этому полю. Обучение моделей на каждом кластере происходило аналогично пункту 2.1. настоящего отчета. Таким образом, было рассмотрено различное приемлемое количество сегментов для разделения исходных данных от 10 до 90. Результаты этого вычислительного эксперимента можно наблюдать в таблице ниже. Сравнение полученных результатов происходила также по метрике RMSE.

Таблица 2. Результаты сравнения количества кластеров для поля PRODUCT_RK

Количество кластеров	Итоговое RMSE
10	2.236958
20	2.229610
30	2.141435
40	2.167921
50	2.196824
60	2.198132
70	2.201140
80	2.215952
90	2.232142

Из таблицы следует вывод, что оптимальным количеством кластеров при сегментации по полю PRODUCT_RK является значение равное 30. При меньшем количестве кластеров деление на сегменты получается слишком грубым и многие товары попадают в один сегмент, хотя имеют не сильно схожую тенденцию спроса, а при большем количестве кластеров внутри сегментов не хватает данных для качественного обучения модели, поэтому наблюдается зависимость увеличения статистической ошибки RMSE при увеличении количества сегментов.

2. Кластеризация по PRODUCT_LVL_RK2.

Вторым наиболее значимым признаком при построении модели является PRODUCT_LVL_RK2 – категория товара. Аналогично предыдущему пункту был также проведен сравнительный анализ для поиска количества необходимых кластеров для лучшего качества модели, однако по

сравнению с полем PRODUCT_RK для данного признака имеется всего 22 различных значения, поэтому количество кластеров в сравнительном анализе будет варьироваться от 3 до 14. Результаты этого эксперимента представлены в таблице ниже.

Таблица 3. Результаты сравнения количества кластеров для поля PRODUCT_LVL_RK2

Количество кластеров	Качество RMSE
3	2.134697
4	2.144947
5	2.144781
6	2.158346
7	2.155040
8	2.155099
9	2.157659
10	2.157705
11	2.157716
12	2.159149
13	2.159260
14	2.162857

По результатам проведенного эксперимента можно сделать вывод, что оптимальным количеством при сегментации по этому полю является значением, равное трем, это показывает нам, что при сегментации по этому

полно достаточным является деление на малое количество сегментов, при этом качество итоговой модели существенно возрастает.

2.2.2 Агломеративная кластеризация магазинов на основе данных о товарах

Еще один способ сегментации с использованием агломеративной кластеризации заключается в сегментации магазинов по данным товаров, так как пара товар-магазин является уникальным идентификатором, и сегментация по базовым полям данных о товарах дает значительную добавку к модели, а именно по полям `PRODUCT_RK`, `PRODUCT_LVL_RK2`, `PRODUCT_LVL_RK3`.

Идея такого подхода сегментации основывается на различии специфики продаваемых товаров в некоторых магазинах, так, например, в магазинах, которые специализируются на определенных товарах, которые можно выделить по полям `PRODUCT_RK`, `PRODUCT_LVL_RK2`, `PRODUCT_LVL_RK3` могут иметь больший спрос, нежели магазины, специализирующиеся на других категориях товаров. Для проверки этой гипотезы попробуем использовать алгоритм агломеративной кластеризации для разбиения товаров на сегменты на основе продаваемых в них товаров.

Для реализации такой сегментации посчитаем средний спрос по каждому из описанных выше полей для каждого магазина и далее обучим модель агломеративной кластеризации на полученных значениях, которая будет разделять не все пары товар-магазин, а только непосредственно магазины.

1. Кластеризация магазинов по `PRODUCT_RK`.

Всего количество различных магазинов в выборке равно 40, а различных номеров продуктов более 36000, как уже упоминалось ранее, поэтому оптимальное количество кластеров было рассчитано аналогично предыдущим пунктам с помощью вычислительного эксперимента. Так как сегментируются именно магазины, то наибольшее количество кластеров может быть не более 40, однако, как было выяснено ранее сегментация просто по полю `STORE_LOCATION_RK` не дает прибавки в качестве модели, поэтому для подбор оптимального количества кластеров будем использовать значения от 3 до 14. Сравнительный анализ при построении различного количества моделей на основе полученных кластеров можно видеть в таблице, представленной ниже.

Таблица 4. Результаты сравнения количества кластеров для магазинов по данным поля
PRODUCT_RK

Качество RMSE	Количество кластеров
2.261293	3
2.269420	4
2.347375	5
2.306709	6
2.254698	7
2.254037	8
2.258017	9
2.192159	10
2.206513	11
2.232264	12
2.234779	13
2.239860	14

Таким образом, наиболее оптимальным количеством кластеров является значение равное 10, причем при делении на кластеры не наблюдается четкой зависимости метрики RMSE от количества кластеров, как было в предыдущих пунктах, однако при использовании количество кластеров более 10 все-таки видна деградация качества модели, поэтому проведение эксперимента более, чем для 14 кластеров не имеет практического смысла.

2. Кластеризация магазинов по PRODUCT_LVL_RK2.

Проведем аналогичное разбиение магазинов только теперь для поля PRODUCT_LVL_RK2 – категория товара. Также проведем вычислительный эксперимент по подбору оптимального количества кластеров и сравнению метрики RMSE полученных моделей с использованием сегментации на все той же тестовой выборке за октябрь-декабрь 2019 года. Результаты этого эксперимента представлены в таблице ниже.

Таблица 5. Результаты сравнения количества кластеров для магазинов по данным поля PRODUCT_LVL_RK2

Количество кластеров	Качество RMSE
3	2.190658
4	2.167583
5	2.177938
6	2.177188
7	2.187741
8	2.171807
9	2.167583
10	2.173996
11	2.171165
12	2.179994
13	2.189158
14	2.207280

Сопоставимо с предыдущим разбиением магазинов по полю PRODUCT_RK наилучшая метрика RMSE при разбиении магазинов по полю PRODUCT_LVL_RK2 достигается при значении равном 9 кластерам.

3. Кластеризация магазинов по PRODUCT_LVL_RK3.

Последним из проанализированных возможных методов сегментации магазинов по продуктовым полям является кластеризация по полю PRODUCT_LVL_RK3 – подкатегория товара. Хотя, как было выяснено в пункте 2.1 сегментации по полю PRODUCT_LVL_RK3 и STORE_LOCATION_RK по отдельности не дают улучшения качества модели, возникает гипотеза будут ли они давать улучшение при совместном использовании, потому что, делая вывод на основе важности всех переменных модели переменная PRODUCT_LVL_RK3 все же является достаточно значимым признаком. Для проверки этой гипотезы аналогично предыдущим пунктам был проведен сравнительный эксперимент с подбором оптимального количества кластеров, результаты которого представлены в таблице ниже.

Таблица 6. Результаты сравнения количества кластеров для магазинов по данным поля
PRODUCT_LVL_RK3

Качество RMSE	Количество кластеров
2.620539	3
2.633903	4
2.647374	5
2.647785	6
2.642615	7
2.642664	8
2.646475	9

2.669471	10
2.671975	11
2.688001	12
2.662180	13
2.662959	14
2.641206	15
2.644229	16
2.642205	17
2.629733	18
2.652366	19

Однако, как можно видеть из таблицы ни одна из сегментаций не дает прирост к модели, обученной на всех данных. Лучшим показателем является значение количества кластеров равное трем, но даже оно значительно хуже, чем прогнозирование спроса на всех данных сразу. Значение в три кластера является весьма обоснованным, ведь при таком разбиении прогноз максимально близок к модели на всех данных, а значит влияние такой сегментации меньше, чем при большем количестве кластеров. Таким образом, наша гипотеза о возможном улучшении модели при сегментации магазинов по подкатегории товаров не оправдалась.

Исходя из проведенных в этом пункте экспериментов можно сделать вывод, что гипотеза о применении сегментация магазинов на основе данных о продуктах все же является оправданной. Качество полученных моделей значительно превосходит качество модели, обученной на всех данных, а также значительно лучше по сравнению с базовой сегментацией по магазинам, описанной в пункте 2.1 настоящего документа. Такой вывод весьма обоснован, потому что во многих магазинах спрос действительно является схожим

именно из-за товаров, которые там продаются, а при сегментации по базовому полю `STORE_LOCATION_RK` для некоторых магазинов просто не хватает количества данных для обучения, чего уже не наблюдается в сегментации магазинов на основе данных о товарах, потому что количество кластеров при таком разбиении значительно меньше. Этот вывод доказывают и проведенные вычислительные эксперименты, в ходе которых оптимальное количество разных сегментов по магазинам было определено как 9-10 кластеров.

ГЛАВА 3. Сравнительный анализ

Для проведения сравнительного анализа агрегируем рассчитанные метрики RMSE для моделей с оптимальным количеством кластеров по каждому из методов сегментации в одну таблицу:

Таблица 7. Сравнительный анализ всех исследованных моделей по метрики RMSE

Название метода	Качество RMSE
Baseline_model	2.521415
Segmentation_by_store_location_rk	2.701105
Segmentation_by_product_lvl_rk2	2.135943
Segmentation_by_product_lvl_rk3	2.590263
Segmentation_by_product_lvl_rk2_aglomerative	2.134697
Segmentation_by_product_rk_aglomerative	2.141435
Segmentation_locations_by_product_rk_aglomerative	2.192159
Segmentation_locations_by_product_lvl_rk2_aglomerative	2.167583
Segmentation_locations_by_product_lvl_rk3_aglomerative	2.620539

Из таблицы видно, что большинство из предложенных методов сегментации дают значительное улучшение предсказательной способности модели на тестовой выборке. Приведем ниже еще одну таблицу, в которой будет рассчитан процент уменьшения метрики RMSE от каждого метода сегментации относительно качества модели на всех данных без сегментации.

Таблица 8. Процент уменьшения метрики RMSE относительно основной модели без сегментации

Название метода	Процент уменьшения RMSE
Segmentation_by_store_location_rk	+ 7.1%
Segmentation_by_product_lvl_rk2	- 15.3%
Segmentation_by_product_lvl_rk3	+ 2.7%
Segmentation_by_product_lvl_rk2_aglomerative	-15.4%
Segmentation_by_product_rk_aglomerative	-15.1%
Segmentation_locations_by_product_rk_aglomerative	-13.1%
Segmentation_locations_by_product_lvl_rk2_aglomerative	-14.1%
Segmentation_locations_by_product_lvl_rk3_aglomerative	+3.9%

В таблице выше, знак минус перед рассчитанным процентом означает уменьшение ошибки RMSE, а значит улучшение качества модели, а знак плюс наоборот обозначает увеличение ошибки RMSE и ухудшение качества прогноза спроса.

Из проведенного сравнительного анализа можно сделать следующие выводы:

- Алгоритм сегментации по базовому полю STORE_LOCATION_RK – номер магазина не дает прироста к качеству итоговой модели относительно модели прогнозирования спроса на всех данных, и даже ухудшает качество прогноза.
- Алгоритм сегментации по базовому полю PRODUCT_LVL_RK3 – подкатегория магазина не дает прироста к качеству модели и ухудшает ее качество. Разбиение магазинов на сегменты с использованием метода агломеративной кластеризации по полю PRODUCT_LVL_RK3 тоже не улучшает качество прогноза, таким образом разбиение по подкатегории товаров не имеет никакого практического смысла.
- Алгоритм сегментации по базовому полю PRODUCT_LVL_RK2 – категория товара дает значительные прибавки к качеству итоговой модели относительно модели обученной на всех данных без сегментации. Сегментация всех данных по полю PRODUCT_LVL_RK2 с использованием метода агломеративной кластеризации уменьшает статистическую ошибку по метрике RMSE на еще большее значение. Итоговое улучшение качество прогноза спроса с использованием этих методов более 15%, что является существенной разницей при построении моделей прогнозирования спроса. Разбиение магазинов по категории товара также улучшает итоговый прогноз, но всего лишь на 14 процентов, это вероятно связано с недостаточным количеством магазинов (около 40 различных значений), однако улучшение также очень существенное.
- Алгоритм сегментации на основе поля PRODUCT_RK – номер товара с использованием агломеративной кластеризации тоже дает большой прирост к итоговому качеству модели – ошибка RMSE уменьшается на 15.1%, кластеризация магазинов по полю PRODUCT_RK с использованием агломеративной кластеризации также уменьшает статистическую ошибку RMSE, но на 14.1%.

Таким образом, на основе всех исследованных методов можно сделать вывод, что использование сегментации по некоторым полям в моделях прогнозирования спроса значительно улучшает качество модели относительно модели, обученной на всех данных. Сегментация по полям PRODUCT_RK и PRODUCT_LVL_RK2 (номер товара и категория

товара) показывает сильный прирост при построении моделей, обученных на выделенных сегментах. Наибольший прирост качества составляет 15.4% уменьшения метрики RMSE при использовании метода агломеративной кластеризации по полю PRODUCT_LVL_RK2.

ЗАКЛЮЧЕНИЕ

По итогам проведенного исследования можно сделать вывод, что использование сегментации в моделях предсказания спроса улучшает качество моделей прогнозирования спроса. Благодаря применению сегментации в моделях прогнозирования спроса можно не только улучшить точность прогноза модели, но и ускорить построение моделей, а также сократить нагрузки на вычислительные системы, потому что построение модели на каждом сегменте отдельно требует гораздо меньше вычислительных ресурсов. Также кластеризация помогает выделить группы товаров и магазинов с общими характеристиками. Это может помочь в построении более точных моделей прогнозирования спроса для каждого сегмента и быстрому реагированию при изменении спроса в одном конкретном сегменте.

В настоящем документе было описано выполнение всех поставленных для исследования задач, таких как:

- Составление обзора статей о данных по продажам и переменных, которые используются для его предсказания.
- Генерация лаговых переменных для модели прогнозирования спроса.
- Разработка модели прогнозирования спроса на всех имеющихся данных
- Выбор признаков, на основе которых можно реализовать методы сегментации.
- Разбиение данных на базовые сегменты по имеющимся полям и на сегменты с использованием метода агломеративной кластеризации.
- Построение моделей прогнозирования спроса для каждого набора сегментов, полученных различными методами кластеризации
- Проведение сравнительного анализа моделей с использованием различных способов сегментации и без нее.
- Анализ результатов, полученных в ходе проведения сравнительного анализа, расчет метрик качества модели, на основе которого можно сделать вывод о необходимости.

Проведенное исследование подтверждает гипотезу о возможности и даже необходимости применения сегментации в моделях прогнозирования спроса, и этот вывод подтверждается сравнительным анализом.

Возможные пути для дальнейшей работы:

- Использование для построения основной модели других методов градиентного бустинга, помимо CatBoost для обобщения сделанных выводов.

- Использование других методов сегментации, помимо агломеративной или разбиение на сегменты методом агломеративной кластеризации, но не по среднему значению спроса, а, например, по медиане.
- Подтверждение сделанных выводов на других источниках данных и на более свежих периодах обучения.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Introduction to Time Series Analysis and Forecasting, Douglas C. Montgomery, Cheryl L. Jennings, and Murat Kulahci, 2015. – С. 200 – 230.
2. Forecasting: Principles and Practice, Rob J Hyndman and George Athanasopoulos, 2013. – С. 73 – 113.
3. CatBoost | Yandex. [Электронный ресурс]. URL: <https://catboost.ai/en/docs/>. (дата обращения 11.04.2023).
4. Betascript Publishing, Lambert M. Surhone, Miriam T. Timpledon, Susan F. Marseken, 2010.
5. Python Machine Learning, Sebastian Raschka, Vahid Mirjalili, 2019. – С. 207 – 241.
6. Encyclopedia of Machine Learning, Claude Sammut, Geoffrey I. Webb, 2019. – С. 563-564.
7. Hastie T; Tibshirani R; Friedman J. The Elements of Statistical Learning. – New York: Springer, 2001. – С. 236 – 243.
8. A density-based algorithm for discovering clusters in large spatial databases with noise, Ester Martin, Kriegel Hans-Peter, Sander Jörg, Xu Xiaowei, Simoudis Evangelos, Han Jiawei, Fayyad Usama, 1983. – С. 226 – 231.