

**Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»**

**Факультет компьютерных наук
Основная образовательная программа
«Прикладная математика и информатика»**

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

**Исследовательский проект на тему
Сегментация generalized
set для обучения
нескольких моделей ML
прогнозирования спроса.**

**Выполнил студент группы 194, 4 курса,
Соколов Ян Олегович**

Руководитель ВКР:

Руководитель департамента, Доцент, Соколов Евгений Андреевич

Соруководитель:

Приглашенный преподаватель, Романенко Алексей Александрович

Москва 2023

1. Введение.....	4
1.1. Описание предметной области.....	4
1.2. Актуальность и значимость работы	4
1.3. Цель и задачи работы	4
1.4. Неформальная постановка задачи.....	5
1.5. Формальная постановка задачи.....	5
1.6. Структура работы	6
2. Обзор литературы	7
3. Главы	9
3.1. Анализ данных.....	9
3.2. Создание признаков.....	12
3.3. Описание метода	15
3.4. Экспериментальный анализ применимости метода	16
3.5. Настройка моделей	21
4. Заключение	22

Аннотация

Данная курсовая работа посвящена разработке и исследованию метода обобщенной сегментации множества данных внутри ML Training Pipeline для прогнозирования спроса. Целью работы является улучшение точности моделей прогнозирования спроса путем создания умной сегментации обучающей выборки и обучении нескольких моделей. Предполагается, что такая сегментация позволит повысить эффективность ML Training Pipeline для прогнозирования спроса и предоставит новые методы по улучшению точности моделей прогнозирования спроса. В рамках работы была проведена серия экспериментов, в которых сравнивалась точность предсказаний моделей, обученных на сегментированных данных, с точностью моделей, обученных на всем наборе данных. Эксперименты показали, что использование предложенного метода сегментации обучающей выборки может значительно улучшить точность моделей прогнозирования спроса. Результаты работы могут быть полезны для компаний и организаций, занимающихся прогнозированием спроса, и способствовать более точному планированию производства, снабжения и оценке потребностей рынка.

Abstract

This research project focuses on the development and investigation of a generalized set segmentation method within the ML Training Pipeline for demand forecasting. This research proposes the use of advanced clustering techniques to create intelligent data segmentations. It is anticipated that such segmentations will enhance the effectiveness of the ML Training Pipeline for demand forecasting and provide insights for improving forecasting model accuracy. A series of experiments were conducted to compare the predictive performance of models trained on segmented data versus models trained on the entire dataset. The experimental results demonstrate that the proposed method of data segmentation improves the accuracy of demand forecasting models. The findings have practical implications for companies and organizations involved in demand forecasting, enabling more precise production planning, supply management, and market demand assessment.

Ключевые слова: Demand forecasting; Machine learning; Data segmentation; Generalized set segmentation; Statistical analysis; Accuracy; Forecasting models; Decision-making; Comparative analysis

1. Введение

1.1. Описание предметной области

Прогнозирование спроса - один из важных аспектов для компаний и организаций, которые занимаются производством и продажей товаров или услуг. Точность прогнозирования спроса позволяет оптимизировать планирование производства, управлять запасами и повышать качество обслуживания клиентов. Однако точность прогнозирования зависит от качества данных и используемых методов.

Существует множество методов прогнозирования спроса, таких как экспоненциальное сглаживание, регрессионный анализ, временные ряды и машинное обучение. Каждый метод имеет свои преимущества и недостатки, и выбор метода зависит от характеристик данных и целей прогнозирования.

1.2. Актуальность и значимость работы

Область прогнозирования спроса становится все более востребованной в последние годы в связи с необходимостью точного прогнозирования потребительского спроса в различных отраслях. На данный момент не существует универсального решения в таких задачах, каждая из задач рассматривается с точки зрения существующих методов, выбирая лучший. Мы хотим представить новый подход в решении подобных задач - обобщенная сегментация обучающей выборки. Исследуя этот метод, мы стремимся улучшить такие аспекты как точность и надежность моделей прогнозирования спроса, которые имеет решающее значение для принятия решений в таких отраслях, как розничная торговля, логистика и управление поставками.

1.3. Цель и задачи работы

Цель данной работы состоит в том, чтобы разработать и применить метод Generalized Set Segmentation внутри ML Training Pipeline для прогнозирования спроса на основе данных о продажах. В частности, работа направлена на решение следующих задач:

- Разработка и реализация метода GSS внутри ML Training Pipeline для прогнозирования спроса.

- Исследование качества прогнозирования спроса при использовании метода GSS в сравнении с другими методами.
- Оптимизация выбора методов прогнозирования для различных типов групп данных на основе результатов GSS.

Важной частью работы является то, что внутри пайплайнов для получения предсказаний будем использовать регрессионную модель, основанную на идее ансамблей из деревьев решений - CatBoostRegressor.

1.4. Неформальная постановка задачи

Данная работа направлена на исследование методов решения следующей задачи: выделение сегментов обучающей выборки, обучение различных моделей на этих сегментах данных и получение предсказаний на тестовой выборке с использованием обученных моделей.

1.5. Формальная постановка задачи

Входные данные:

- Обучающая выборка D , состоящая из множества пар товар-магазин $P = \{(p_1, s_1), (p_2, s_2), \dots, (p_n, s_n)\}$, где каждая пара характеризуется временным рядом спроса $y_i = \{y_1, y_2, \dots, y_m\}$, где $y_i(t)$ – величина спроса на пару товар-магазин (p_i, s_i) в момент времени t .
- Признаковое пространство $X = \{X_1, X_2, \dots, X_i\}$ для всех пар товар-магазин.

Выходные данные:

- Прогноз спроса на каждую пару товар-магазин из множества P для будущих периодов времени.

Процесс решения задачи:

- Предобработка данных о продажах для каждого товара в множестве D .
- Применение метода Generalized Set Segmentation (GSS) для разбиения обучающей выборки на обобщенные наборы данных S_1 ,

S_2, \dots, S_k , где $S_1 \cup S_2 \cup \dots \cup S_k = D$ и $S_1 \cap S_2 \cap \dots \cap S_k = \emptyset$.

- Обучение множества моделей $M = \{M_1, M_2, \dots, M_k\}$ прогнозирования спроса на основе данных из полученных сегментов S_1, S_2, \dots, S_k .
- Вычисление прогнозов спроса для каждого товара из множества P на будущие периоды времени с помощью обученных моделей.
- Оценка качества прогнозирования с помощью метрики средней абсолютной ошибки (MAE).

1.6. Структура работы

Данная работа состоит из следующих разделов:

- Введение. В этом разделе дается описание предметной области, объясняется актуальность и значимость работы, формулируется цель и задачи, заявляются неформальная и формальная постановки задачи.
- Обзор литературы. В данной главе будет проведено краткое описание и сравнение релевантных работ.
- Основная часть. В основной части работы подробно описаны различные методы сегментации обучающей выборки, проведенные с ними эксперименты и полученные результаты.
- Заключение. В этом разделе дана характеристика результатов работы, а также описаны перспективы развития работы.
- Список источников. В этой части будут указаны все использованные источники литературы в ходе данной работы.

2. Обзор литературы

В литературе по прогнозированию спроса и машинному обучению нет достаточно глубоких исследований по использованию различных методов сегментации обучающей выборки, в имеющихся исследованиях авторы работ лишь частично затрагивают этот метод.

В статье [S. Mazzanti, "What Is Better: One General Model or Many Specialized Models?"](#) делается вывод, что нет веских причин для использования специализированных моделей, поскольку мощные алгоритмы машинного обучения, основанные на идее деревянных ансамблей при должной настройке, самостоятельно могут разобраться со сложными закономерностями в данных, а использование специализированных моделей связано с рядом практических сложностей. Автор сравнил пайплайны с использованием сегментации и без нее, используя 12 реальных наборов данных, и обнаружил, что общая модель превосходит пайплайн с отдельными моделями под каждый сегмент. Важно заметить, что в этом исследовании не была произведена настройка модели, учтем это и проведем ее в нашей работе. Также, автор данной статьи использовал самую примитивную сегментацию обучающей выборки, в нашем исследовании мы хотим использовать более продвинутые методы разделения данных для создания более разумной сегментации.

В следующих статьях [C. Short, "Segmentation: Does it Impact Predictive Modelling Results for People Analytics?"](#) и [Y. Hadar, "Should I Train a Model for Each Customer or Use One Model for All of My Customers?"](#) также исследуется влияние сегментации на результаты прогнозирования. Авторы делают вывод, что результаты сегментации зависят от типа используемой модели и специфики конкретных данных. В статье подчеркивается, что сегментация полезна для моделей логистической регрессии, поскольку они могут не учитывать эффекты взаимодействия между переменными. Сегментация обучающей выборки может учесть эти факторы и помочь в получении более точных прогнозов. Однако для более сложных методов машинного обучения, основанных на деревьях решений, исследование показало, что сегментация не

привела к значительному повышению точности. В статье делается вывод о том, что, хотя влияние сегментации и может быть разным, оно по-прежнему является полезным инструментом для бизнеса в отдельных случаях. В нашем исследовании мы проверим данные выводы, а также сделаем свои с использованием новых методов сегментации.

3. Главы

3.1. Анализ данных

В качестве данных был предоставлен набор реальных данных, состоящий из 3.5 млн строк со следующей структурой:

- Колонки-ключи: `product_rk`, `store_location_rk` – id товара и магазина
- Колонка с датой: `period_start_dt` – дата фиксации уровня спроса
- Колонки-драйверы: `promo1_fl`, `promo2_fl` – флаг акционного дня, `regular_price`, `promo_price` – обычная цена, цена со скидкой.
- Геолокационные колонки: `store_location_lvl{1, 2, 3, 4, 5}` – иерархические данные о местоположении магазина.
- Продуктовые колонки: `product_lvl_rk{1, 2, 3, 4, 5}` – иерархическая группировка товаров.
- Целевая переменная: `demand`

Период, за который предоставлены данные – с 2016 по 2019 год. Количество уникальных магазинов – 10, количество уникальных товаров > 25000. Данные не содержат пропусков. Гранулярность данных – 1 неделя, то есть мы имеем дело с понедельным спросом и наш горизонт прогнозирования будет составлять несколько недель.

Рассмотрим средние значения спроса на разные товары, для примера возьмем 15 случайных объектов:

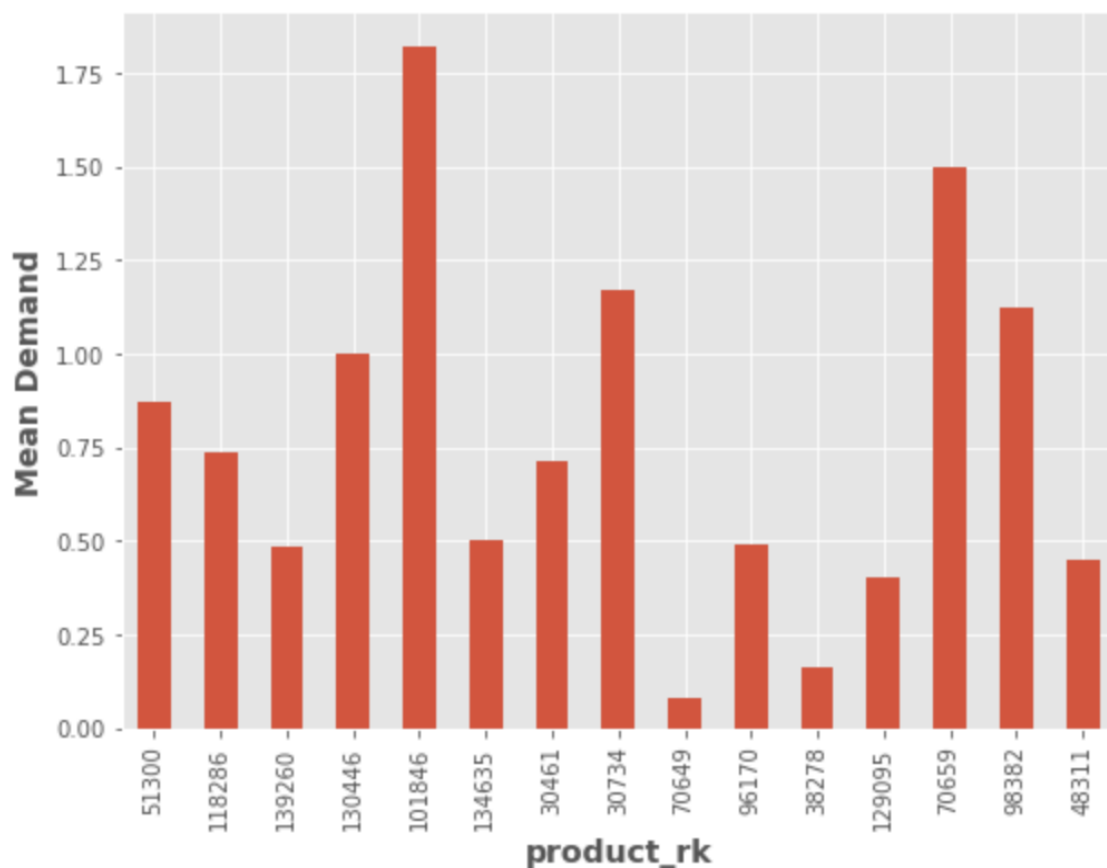


Рисунок 3.1- Средний спрос по товарам

По графику видим, что товары различаются между собой по популярности спроса.

Аналогично посмотрим на графики среднего спроса по другим полям:

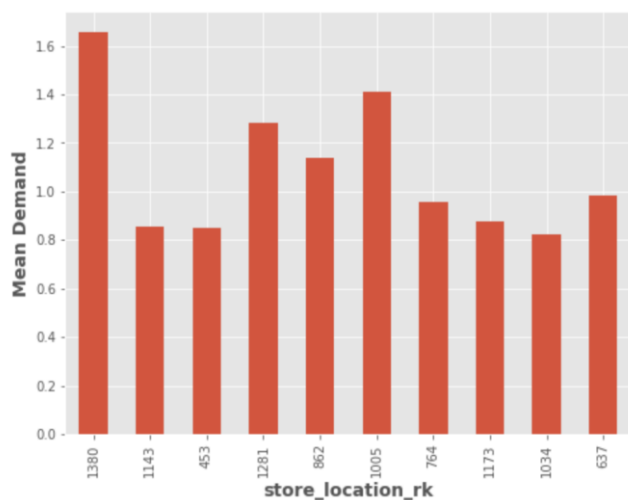


Рисунок 3.2 - Средний спрос по магазинам

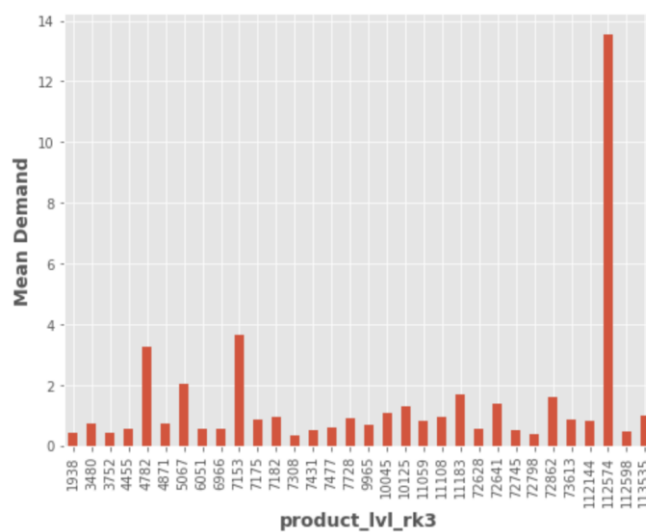


Рисунок 3.3 - Средний спрос по группам товаров

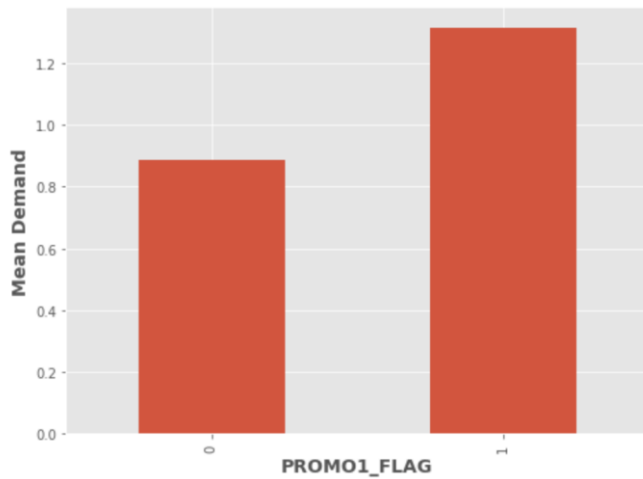


Рисунок 3.4 - Средний спрос по промо флагу

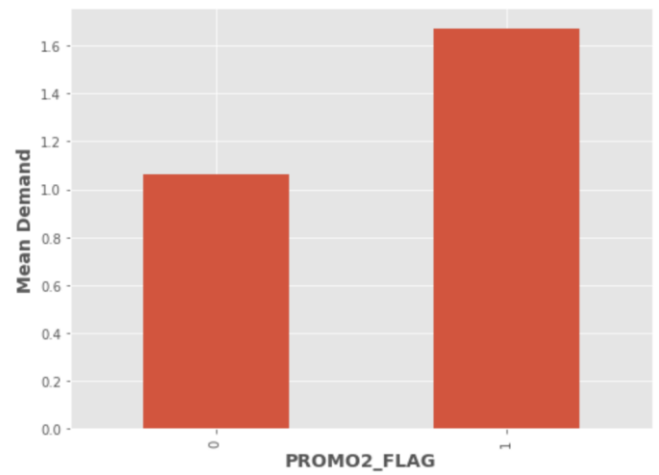


Рисунок 3.5 - Средний спрос по Promo2

Также видим, что присутствует различие в средних значениях спроса в зависимости от локации магазина, группы товара, промо флага. Например, средний спрос в магазине 1143 в 2 раза ниже, чем в 1380, а средний спрос на группу товаров 112574 выше более чем в 5 раз по сравнению с остальными группами.

Рассмотрим средние значения спроса по товарам в различных магазинах:

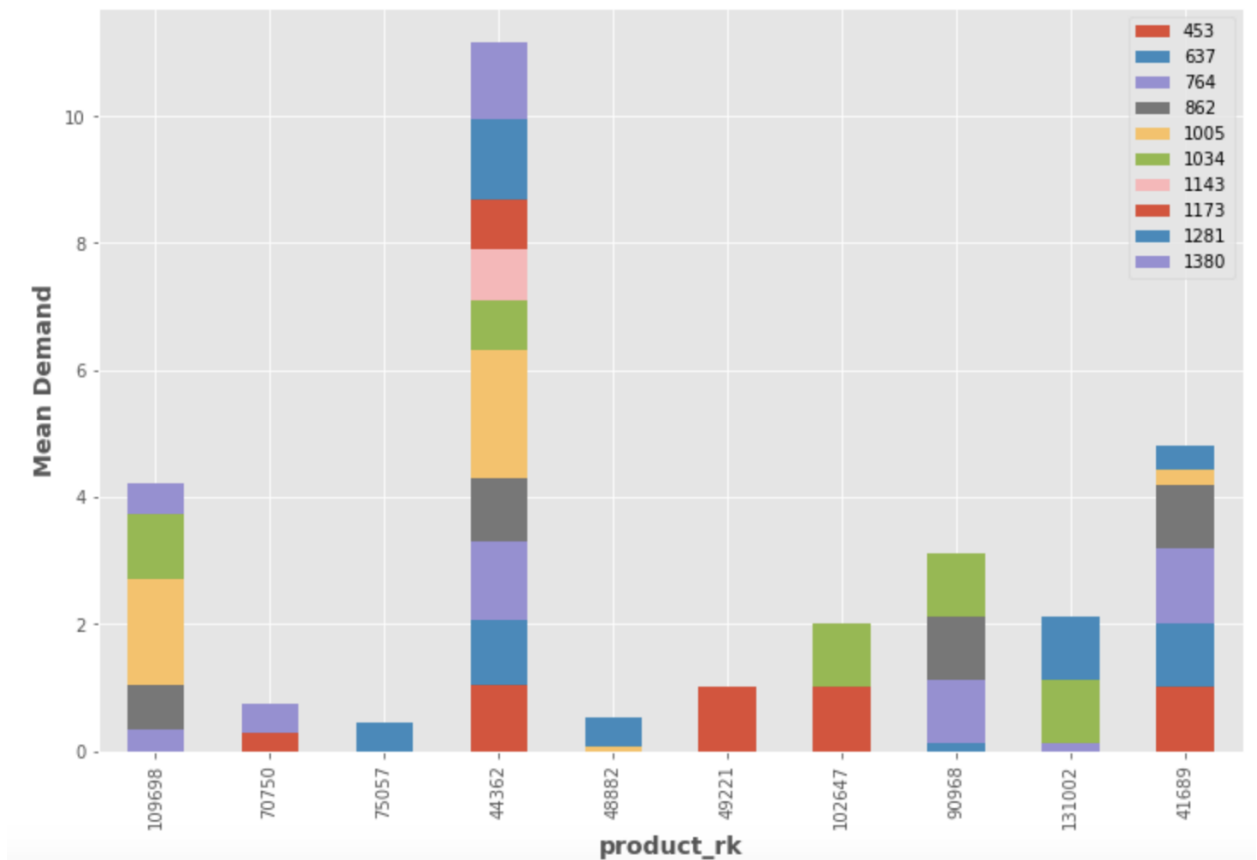


Рисунок 3.6 - Средний спрос по товарам в разбивке на магазины

Можно отметить не менее интересное наблюдение: в разных магазинах спрос на один и тот же товар может сильно отличаться или его может вовсе не быть, если товар отсутствует в продаже. Так, например, наблюдаем минимальный спрос на товар 41689 в магазине 1005, в то время как в этом же магазине видим высокий спрос на товар 44362.

Из проведенного анализа можно сделать несколько выводов:

- Целевая переменная сильно зависит от различных факторов: акций, магазинов, типа товаров и т.д. Для получения качественных прогнозов нужно учесть эти факторы при построении признаков для обучения моделей.
- Наши данные могут не содержать ту самую переменную, по которой модели, основанные на идее деревьев решений оптимально могли бы разбить плоскость, именно с этим и может помочь метод сегментации обучающей выборки.

3.2. Создание признаков

Для построения признаков, необходимых для обучения моделей, была использована идея характеристик на основе скользящего среднего. На данном рисунке можно увидеть схематичную иллюстрацию данного метода:

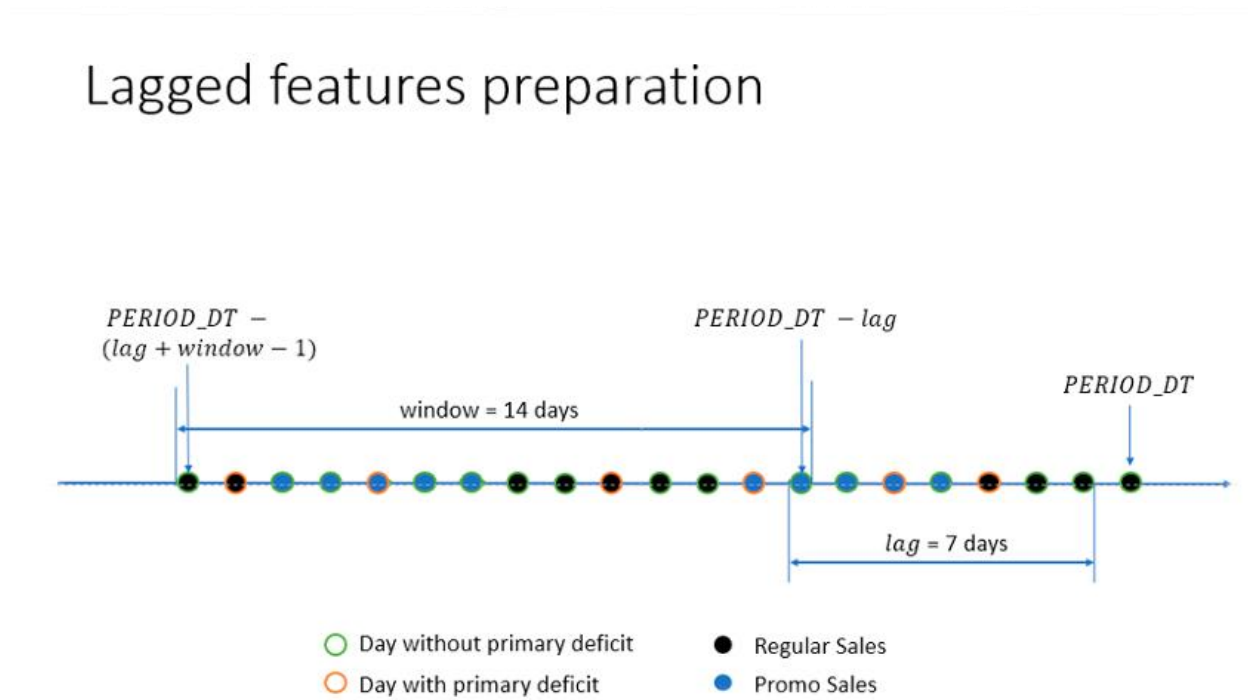


Рисунок 3.7 - Схема построения лаговых признаков

Пояснение к иллюстрации: для расчета признака на текущую дату по целевой переменной используется агрегация значений внутри окна заданного размера (window), фильтрация используемых расчетных значений и отступ (лаг), определяющийся по заданному в задаче горизонту прогнозирования.

Введем более формальную формулировку:

$$FeatureLag_t = \left(\sum_{j=t-wSize-lag+1}^{j \neq t-lag} y_j * [x_j = filter] \right).apply(agggregation)$$

Формула 3.1 - Формула лагового признака

Данная идея была реализована в виде функции, принимающей ряд параметров:

#	Column	Dtype	Description
0	data	pd.DataFrame	Данные, для которых осуществляется генерация признаков
1	target_cols	list	Список целевых переменных, по которым производится расчет признаков
2	id_cols	list	Ключи объектов, по которым производится группировка данных
3	date_col	str	Колонка с датой
4	lags	list	Значения лага(отступ от текущей даты, измеряется в днях)
5	windows	list	Размер окна, внутри которого применяется функция(дни/недели/месяцы)
6	preagg_methods	list	Применяемые методы для получения уникального ключа каждого объекта
7	agg_methods	list	Методы агрегации, применяемые внутри заданного окна
8	dynamic_filters	list	Названия колонок, которые используются в качестве фильтра
9	ewm_params	dict	Параметры pandas.ewm метода для каждого dynamic_filter

Рисунок 8 - Входные параметры функции

С помощью данной функции были построены 128 признаков с различными значениями параметров lag, window, dynamic_filters, agg_methods.

Выберем конкретные значения id товара и магазина и рассмотрим

примеры построенных признаков:

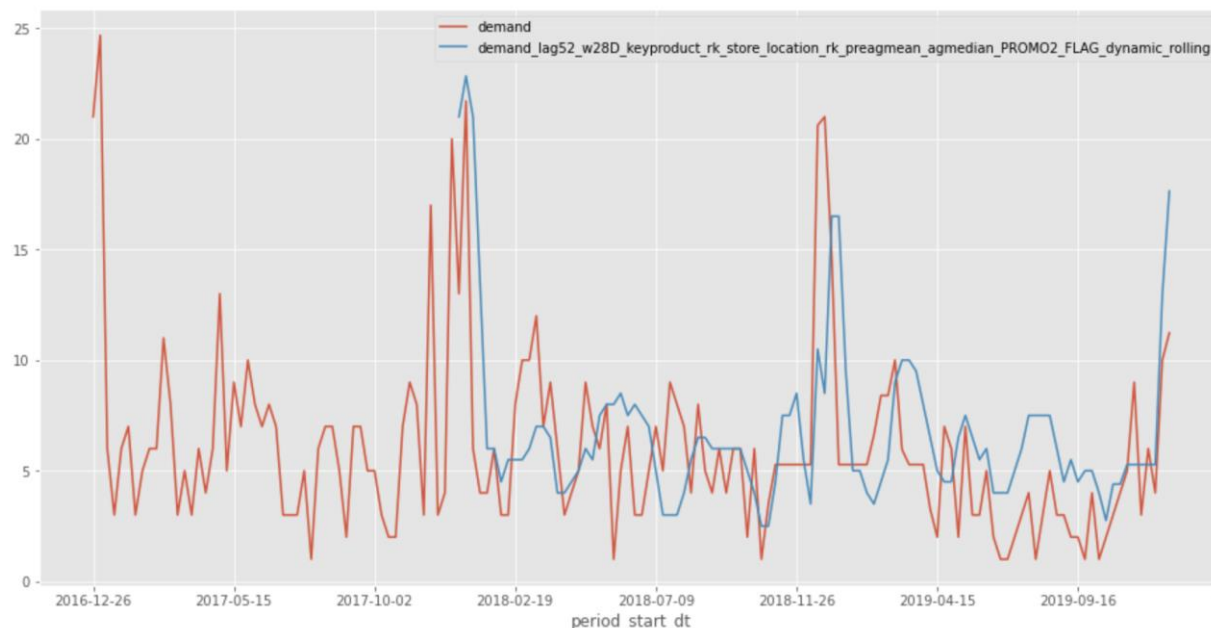


Рисунок 3.9 - Пример лагового признака

Синим цветом показан признак с отступом в 52 недели и окнов в 28 дней, красным - целевая переменная. Видим, что пиковые значения спроса хорошо повторяются на основе данных предыдущего года с помощью нашего построенного признака.

Еще один пример характеристики спроса, но уже с меньшим отступом:

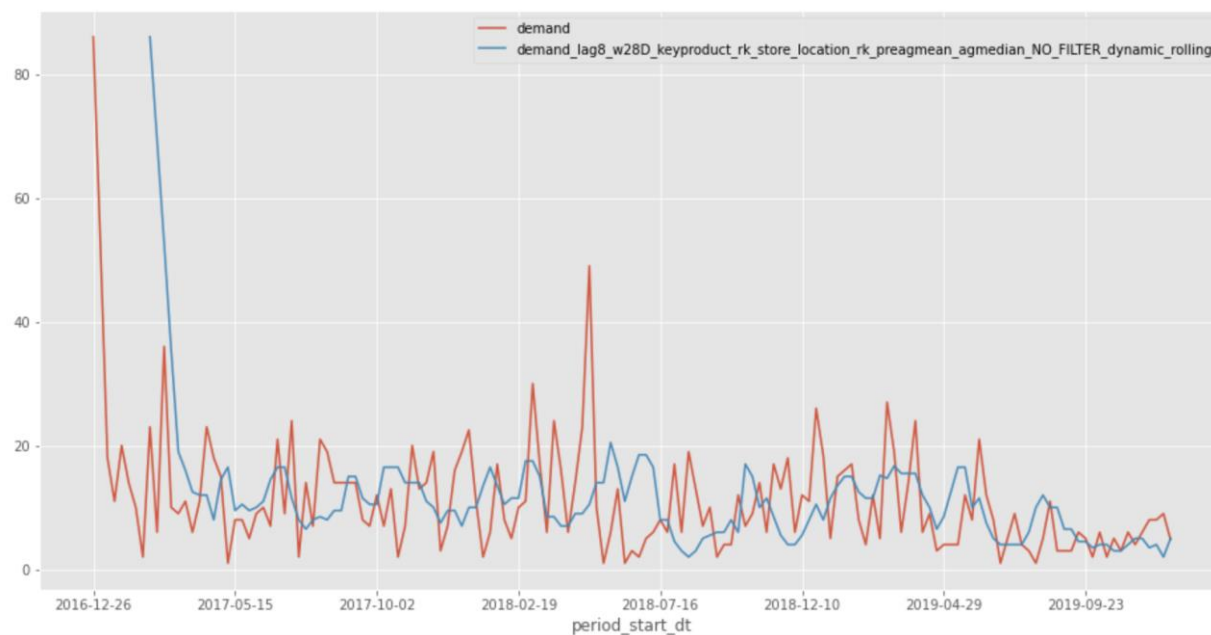


Рисунок 3.10 - Пример лагового признака

Здесь значение лага составляет всего 8 недель, а также нет фильтрации по

промо флагу, то есть мы будем брать все значения в окне 28 дней, которые были 28 дней назад для данной пары товар-магазин.

Проанализировав построенные признаки, можно сделать вывод, что для пар товар-магазин с несильно колеблющимся спросом хорошо подойдут признаки, не учитывающие промо флаг при построении, а для пар, спрос которых сильно коррелирует со значением промо флага - признаки, которые учитывают этот драйвер, поэтому добавим все построенные признаки в наш датасет.

3.3. Описание метода

В данной главе будет представлено описание метода сегментации обучающей выборки внутри ML Training Pipeline для прогнозирования спроса.

Выбор признаков для сегментации

Перед началом сегментации необходимо определить набор признаков, по которым будет происходить разделение данных. Эти признаки могут включать характеристики товаров, сезонность, географическую информацию или любые другие факторы, которые могут оказывать влияние на спрос.

Алгоритмы сегментации

Для сегментации обучающей выборки мы можем использовать 2 подхода:

- Сегментация по бизнес-правилу. Например, отдельная модель для различных значений промо флага.
- Сегментация с использованием обучения без учителя - кластеризация. В нашем исследовании в качестве такого алгоритма была выбрана иерархическая кластеризация - AgglomerativeClustering.

В работе будут продемонстрированы результаты обоих подходов к разделению обучающей выборки.

Оценка качества сегментации

После применения алгоритма сегментации необходимо оценить качество полученных сегментов. В эту оценку входит анализ структуры сегментов, оценку схожести товаров внутри каждого сегмента, а также проверку,

насколько сегменты отражают реальные различия в поведении спроса. Важно учитывать, что оценка качества сегментации является итеративным процессом, и возможно потребуются настроить параметры алгоритма или применить дополнительные методы для улучшения результатов.

Применение сегментации на обучающей выборке

После успешной сегментации обучающей выборки каждый сегмент представляет собой подмножество данных, которые имеют похожие характеристики спроса. Эти сегменты можно использовать для обучения моделей прогнозирования спроса. Полученные прогнозы сравниваются с прогнозами, полученными общей моделью (обученной на всем датасете без сегментации) и делаются выводы посредством подсчета скоринговых метрик.

3.4. Экспериментальный анализ применимости метода

Подготовка к сегментации

В данной части работы проводится экспериментальный анализ для оценки применимости предложенного метода сегментации обучающей выборки внутри ML Training Pipeline для прогнозирования спроса. Для измерения качества предсказаний используется метрика MAE (Mean Absolute Error).

Для проведения эксперимента была выделена тестовая выборка, состоящая из последних двух месяцев - ноября и декабря 2019 года. Это позволяет оценить качество предсказаний на реальных данных, которые модель ранее не видела. В качестве тренировочной выборки мы взяли данные с сентября 2018 по октябрь 2019, так как этот период имеет наименьшее количество пропусков в построенных признаках. Итоговый размер обучающей выборки - 1.5 млн строк.

В качестве базового сравнения для проверки адекватности предсказаний использовалась дамми модель, которая всегда предсказывает средние значения спроса по каждой паре товар-магазин. Метрика MAE данной модели составила 0.884.

Построение общей модели

Для начала была построена общая модель, которая использовала всю обучающую выборку для обучения. С помощью нее были получены предсказания на тестовой выборке, MAE данной модели - 0.8408. При проведении экспериментов с сегментацией будем учитывать это значение и стараться получить меньшую ошибку.

Эксперименты с использованием бизнес правил для сегментации

Первой построенной сегментацией стало разделение обучающей выборки по промо флагу. Обучающая выборка разделилась на 2 части - со значением промо флага равному 0 (1 млн строк) и 1 (0.5 млн строк). Важно упомянуть, что при обучении общей модели мы также добавляем признак промо флага. Результат такой сегментации - MAE = 0.86. Такой результат можно объяснить простотой разбиения и хорошо построенными лаговыми фичами, которые были сформированы с использованием фильтрации по промо флагу. В таком случае ухудшение MAE связано с сокращением объема тренировочных выбор для моделей.

Следующее разбиение, полученное с помощью сегментации по бизнес-правилу, стало разбиение по второму уровню из иерархии товаров - `product_lvl_rk2`. Для каждого уникального значения этой колонки была построена своя модель, с помощью которой были сделаны предсказания на тестовой выборке с теми же значениями `product_lvl_rk2`. Данный эксперимент также не улучшил нашу предсказательную способность пайплайна, MAE = 0.85.

Простые разбиения не дают прироста в качестве предсказаний, как и упоминалось выше, скорее всего модели, основанные на идее деревьев решений, сами хорошо находят оптимальные значения предикатов и для улучшения предсказательной способности пайплайна нужны более хитрые разбиения.

Эксперименты с использованием алгоритмов кластеризации

В качестве алгоритма машинного обучения без учителя была выбрана

широко применяемая для данных с временными рядами иерархическая кластеризация, идея которой заключается в последовательном объединении меньших кластеров. Данный тип кластеризации предусматривает 2 метода получения количества кластеров:

- Заранее указать желаемое количество кластеров
- При помощи параметра порогового значения. Слияние кластеров происходит до тех пор, пока расстояние между ними меньше порогового значения.

В нашей работе мы будем использовать первый вариант и оценивать качество кластеризации с помощью одновременного учета двух метрик: Silhouette score и Davies-Bouldin index. Распишем как считаются данные метрики.

Silhouette score:

$$\frac{b - a}{\max(a, b)}$$

Формула 2 - Silhouette score

где a - среднее внутрикластерное расстояние, а b - среднее межкластерное расстояние. Таким образом, область определения метрики - $[-1, 1]$, чем ближе значение к 1, тем компактней лежат точки внутри кластера и тем дальше расположены кластеры друг от друга, значение около 0 - наложение кластеров друг на друга, а близкое значение к -1 - сигнал плохой кластеризации.

Davies-Bouldin index:

$$DB = \frac{1}{N} \sum_i^N D_i ; D_i = \max(R_{i,j}) , i \neq j ; R_{i,j} = \frac{r_i + r_j}{d(C_i, C_j)} ,$$

Формула 3 - Davies-Bouldin score

где r_i - среднее расстояние от точек кластера до его центра. Таким образом, метрика принимает значения от 0 до бесконечности и значения, лежащие ближе к нулю, свидетельствуют о сгруппированности точек внутри кластера и отдаленности кластеров друг от друга.

Первой сегментацией, полученной с помощью обучения без учителя стало разбиение с использованием в качестве признаков среднего спроса на товары по иерархической группе `product_lvl_rk3`. Для подбора оптимального числа кластеров использовали 2 вышеупомянутые метрики:

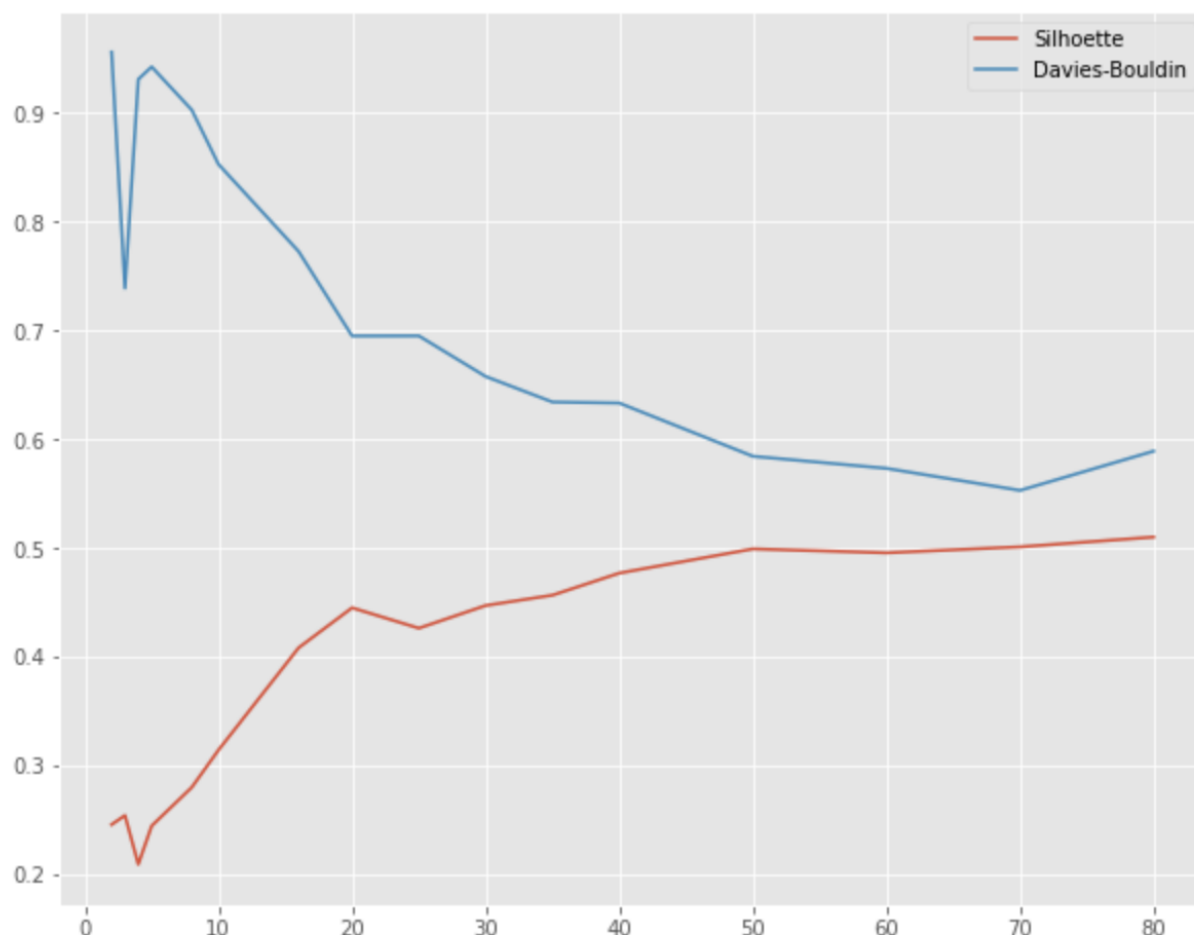


Рисунок 3.11 - График метрик кластеризации

Как видно по графику, оптимальным числом кластеров является - 70. Построим кластеризацию с полученным параметром, добавив метки кластеров в тренировочную и тестовую выборку и построим отдельные модели для каждого кластера. Важно отметить, что каждая модель так же как и общая, подвергалась настройке. Также стоит заметить, что в тестовой выборке присутствуют объекты без метки кластера - это связано с тем, что кластеризацию мы проводили по тренировочной выборке во избежание утечки целевой переменной. Получим предсказания для таких объектов общей моделью. Полученное значение метрики MAE: 0.825, улучшение составило 1.9%.

Для следующего варианта кластеризации в качестве признаков был выбран средний спрос по городам и произведен поиск оптимального числа разбиений:

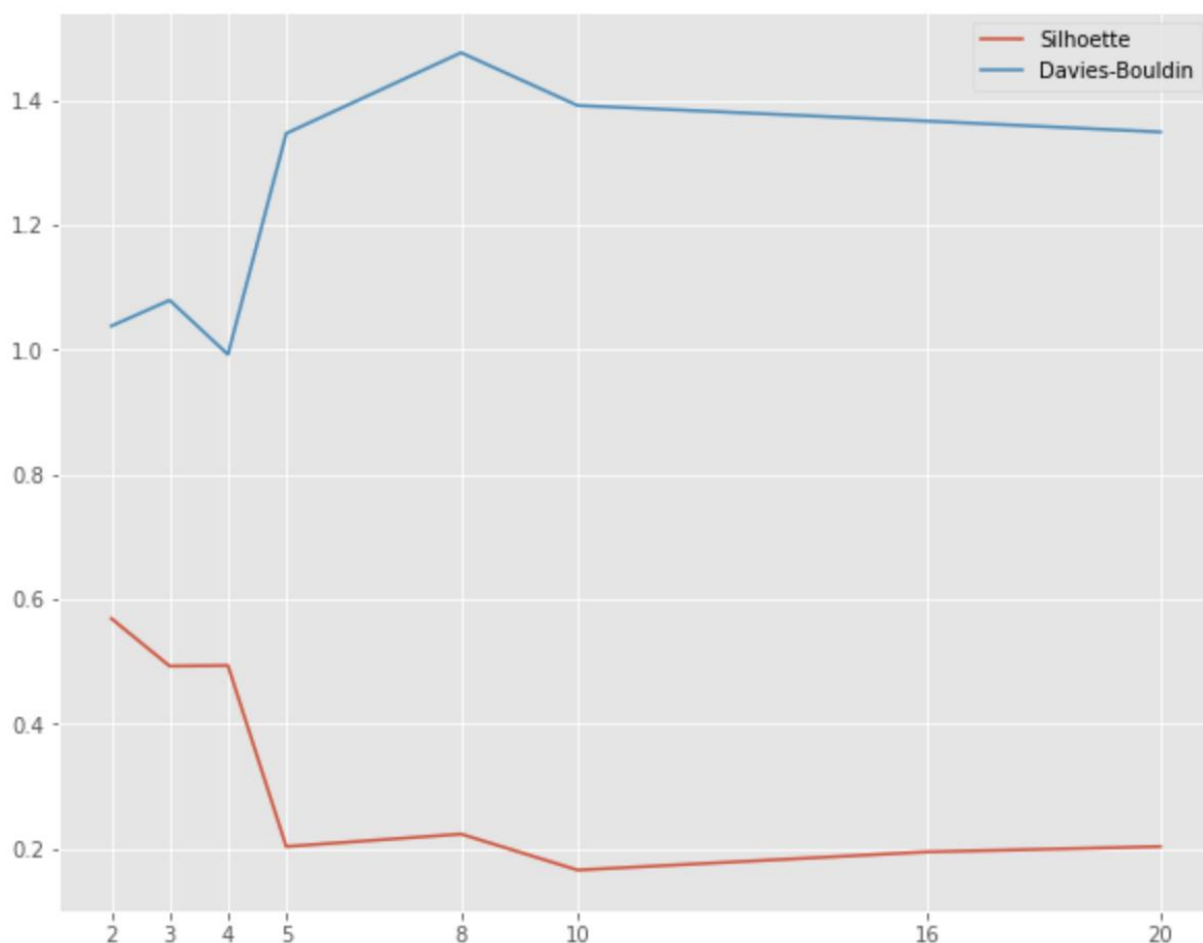


Рисунок 3.12 - График метрик кластеризации

Для такого варианта кластеризации оптимальное число кластеров оказалось меньше - 2. Это может быть связано, например, с тем, что города в наших данных распределены на 2 региона. С помощью данного варианта кластеризации итоговое значение метрики MAE получилось равным - 0.822, что лучше предыдущего варианта кластеризации, и превосходит пайплайн с общей моделью на 2.3%.

Также был проведен эксперимент с сегментацией обучающей выборки по парам товар-магазин. Для каждой такой пары была построена и настроена отдельная модель. Такой способ разделения обучающей выборки не дал улучшения с точки зрения MAE, скорее всего это связано с небольшим

количеством данных для каждого кластера, для каждой группы в среднем выходит ~150 объектов для обучения, поэтому модели недообучаются и имеют плохую предсказательную способность.

3.5. Настройка моделей

Для настройки моделей был использован перебор гиперпараметров по заданной сетке с использованием кросс-валидации. Каждый набор гиперпараметров из сетки тестировался с использованием пяти фолдов. В качестве параметров для перебора были выбраны `learning_rate` и `max_depth`. Сетка параметров выглядела следующим образом:

- `learning_rate` [0.01, 0.05, 0.1]
- `max_depth` [2, 4, 6, 8]

4. Заключение

В рамках данного исследования были выполнены все поставленные задачи:

- Был проведен анализ имеющихся данных, выявлены зависимости целевой переменной от различных факторов
- Были построены лаговые признаки, с помощью которых можно улавливать тренды и обучать модели
- Был реализован метод сегментации обучающей выборки внутри ML Pipeline по прогнозированию спроса
- Были задействованы различные методы сегментации: как бизнес правила, так и обучение без учителя
- Был реализован подбор оптимального числа кластеров для метода обучения без учителя с использованием учета двух кластерных метрик
- Были проведены эксперименты по сравнению пайплайна, содержащего одну общую модель, с пайплайнами, состоящими из нескольких обученных моделей под каждый сегмент обучающей выборки
- Были сделаны выводы по каждому эксперименту

Лучшим пайплайном с сегментацией обучающей выборки, оказался пайплайн с использованием метода обучения без учителя для кластеризации данных. Он превзошел пайплайн с одной общей моделью на 1.9% - в рамках бизнес-задач это может быть большим приростом в прибыли компании.

В качестве дальнейших идей для развития данного исследования могут выступать следующие предложения:

- Создание более “хитрой” сегментации обучающей выборки
- Тестирование других методов обучения без учителя для улучшения качества кластеризации

Построенный пайплайн с сегментацией обучающей выборки имеет преимущество в сравнении с общей моделью для всех датасета, несмотря на

такой результат, нужно принимать во внимание структуру и особенность данных, с которыми приходится работать, для каждого отдельного случая стоит рассматривать несколько различных пайплайнов и выбирать лучший.

Список источников

1. P. Lalou, S. T. Ponis, and O. K. Efthymiou "Demand Forecasting of Retail Sales Using Data Analytics and Statistical Programming"
2. R. Fildes, S. Ma, and S. Kolassa, "Retail forecasting: Research and practice"
3. O. B. Yu'zbasoglu and H. Ku'cu'kaydin "Forecasting with Ensemble Methods: An Application Using Fashion Retail Sales Data"
4. S. Mazzanti, "What Is Better: One General Model or Many Specialized Models?"
5. C. Short, "Segmentation: Does it Impact Predictive Modelling Results for People Analytics?"
6. Y. Hadar, "Should I Train a Model for Each Customer or Use One Model for All of My Customers?"
7. D. Rade'ic', "Time Series From Scratch — Moving Averages (MA) Theory and Implementation"