ОТЧЁТ ПО КОНКУРСУ GPN-CUP

Выполнил: Цыплов Алексей

2020 г.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В далеком 2148 году мир переживает последствия кризиса и глобальной войны. Постапокалиптическую пустошь населяют безжалостные воины, но все еще есть место для честных предпринимателей. Вы работаете в Компании, управляющей сетью магазинов, которая торгует различными товарами, пользующимися спросом в данной реальности.

Для лучшего управления магазинами, в частности, для более оптимального планирования промо-кампаний и прогнозирования спроса, вам необходимо разбить магазины на кластеры похожих. Единственный способ, которым пользовалась компания в прошлом – это разбитие по географическому признаку, то есть по городам. Но вы верите, что прочие характеристики магазинов, а самое главное, профили продаж магазинов, помогут сделать это гораздо точнее.

1. ПРЕДОБРАБОТКА ДАННЫХ

1.1. Описание данных

Подробное знакомство с данными (пошагово) описано в файле preprocessing.ipynb. Всего имеется три файла в формате PARQUET, а именно:

- данные о локациях;
- данные о магазинах;
- данные о продажах.

В файле с информацией о магазинах содержатся пропуски.

Всего имеется пятнадцать городов в трёх локациях. К сожалению, из-за маскировки коммерческой информации, мы можем использовать информацию о локации только для сравнения магазинов между собой и никак не можем использовать информацию о самой локации. Например, компания могла бы предоставить такую информацию о городах, как типы и количество клиентов, наличие и количество конкурентов по каждому типу товара и т. п.

У каждого магазина в компании есть владелец. Большинство магазинов принадлежат рейдерам (примерно 87.5%). Всего же есть пять различных владельцев.

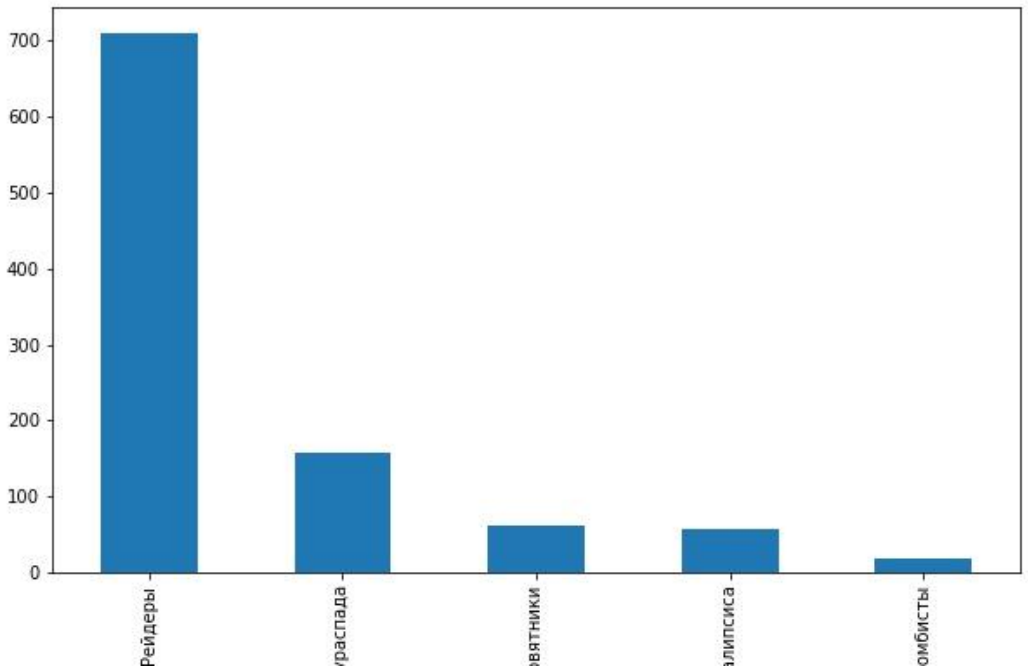

Рисунок 1 – Гистограмма владельцев магазинов

К сожалению, у нас нет информации о владельцах, которую можно было бы использовать при кластеризации (хотя технически, если бы было больше времени), можно было бы попытаться получить её из данных о продажах.

Отметим также, что всего магазинов 845, а уникальных записей в таблице «владелец», «магазин» – 1002, что означает, что с течением времени магазины меняли владельцев, но не более одного раза. Посмотрим теперь на продажи различных товаров.

Всего в продаже одиннадцать типов товаров. Наиболее популярным типом товара является «Бензак», а на втором месте находится «Солярка».

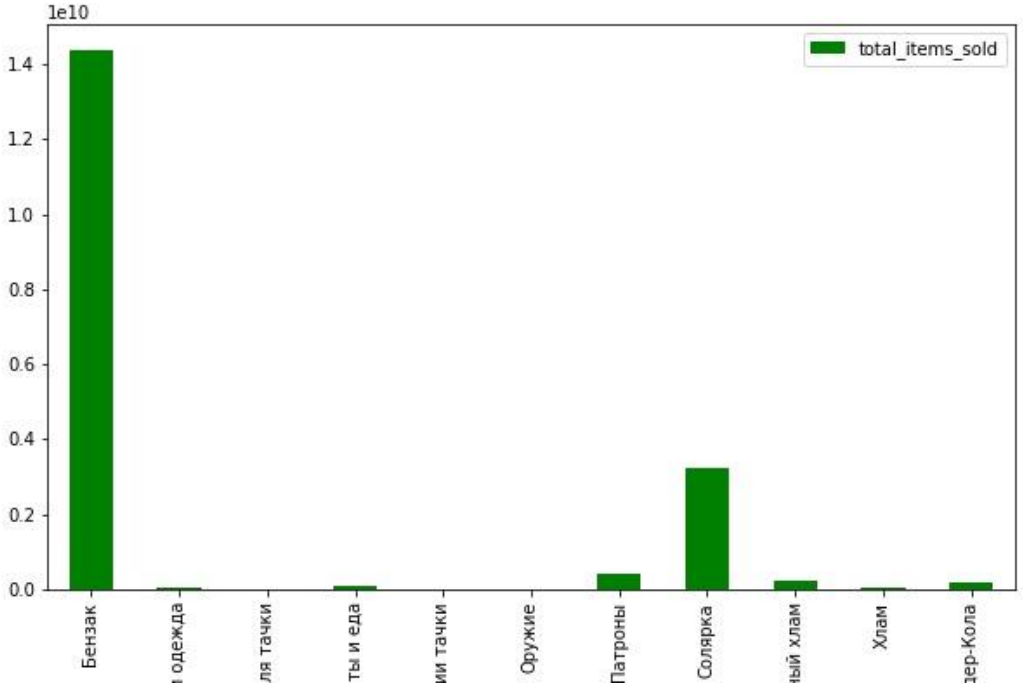

Рисунок 2 – Распределение продаж по типам товаров

Были также построены гистограммы продаж для каждого типа товара. Визуально переменная total_items_sold имеет пуассоновское распределение. Это наталкивает на мысль, что в дальнейшем при прогнозировании разумно использовать обобщённые линейные модели.

Общий вывод после знакомства с данными: можно построить несколько различных вариантов кластеризации, используя как данные о магазинах (расположение, наличие сервисов и т. д.), так и информацию о продажах. С

этой целью было построено несколько таблиц, по которым в дальнейшем будет проведена кластеризация.

1.2. Препроцессинг

Для каждого типа товара была создана таблица: имена строк – date, имена столбцов – shop_id, на пересечении – total_items_sold.

Была создана таблица encoded_shops, которая содержит в себе информацию о магазинах из файла shops.parquet с заполненными пропусками (Таблица 1) и информацию о локации магазинов.

Таблица 1 – Заполнение пропущенной информации о магазинах

Столбец	Перевод	Чем заполняются пропуски
city	Город	Неизвестно
location	Локация	Неизвестно
is_on_the_road	Рядом с дорогой	нет
is_with_the_well	Есть колодец	нет
is_with_additional_services	Есть дополнительные сервисы	нет
shop_type	Тип магазина	0

Информация в таблице encoded_shops закодирована методом one hot encoding. Признаки переобозначены через x_i.

Эти данные присоединены по ключу shop_id к данным о продажах. Полученная таблица записана в forecast_data.parquet. Информация в ней тоже закодирована методом one hot encoding.

2. КЛАСТЕРИЗАЦИЯ

2.1. Цель

Главной целью кластеризации в приведённой работе является упрощения построения прогнозной модели. Ключевая идея состоит в том, чтобы разбить магазины на кластеры так, чтобы динамика продаж в магазинах одного кластера была похожей, а динамика магазинов из разных кластеров отличалась. При таком подходе можно добавить информацию о кластере в прогнозную модель в качестве фиктивной переменной либо дополнительно к уже имеющимся факторам, либо взамен их.

Альтернативный вариант состоит в построении своей прогнозной модели для каждого кластера.

В работе приведено несколько вариантов разбиения на кластеры по факторам и разбиение на кластеры по рядам динамики для каждого типа товара.

2.2. Средства

Во всех случаях использовалась агломеративная кластеризация. Для визуальной оценки качества строились дендрограммы.

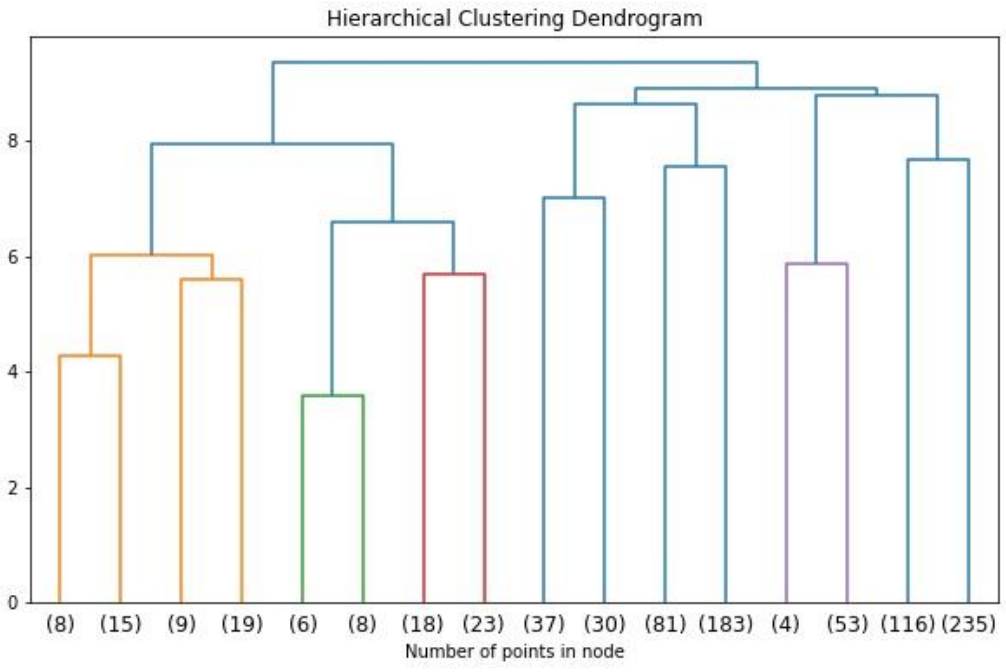

Рисунок 3 – Пример дендрограммы

Для численной оценки качества – силуэт и индекс Дэвиса-Боулдина. Достаточно знать, что чем больше силуэт, тем лучше. Чем меньше индекс Дэвиса-Боулдина, тем лучше.

2.2.1. Кластеризация по факторам

Первая попытка кластеризировать объекты состояла в разбиении по факторам из таблицы encoded_shops. При этом, поскольку все факторы бинарные, то в качестве расстояния между магазинами использовалось количество несовпадений. Расстояние между кластерами считалось как расстояние между центрами масс.

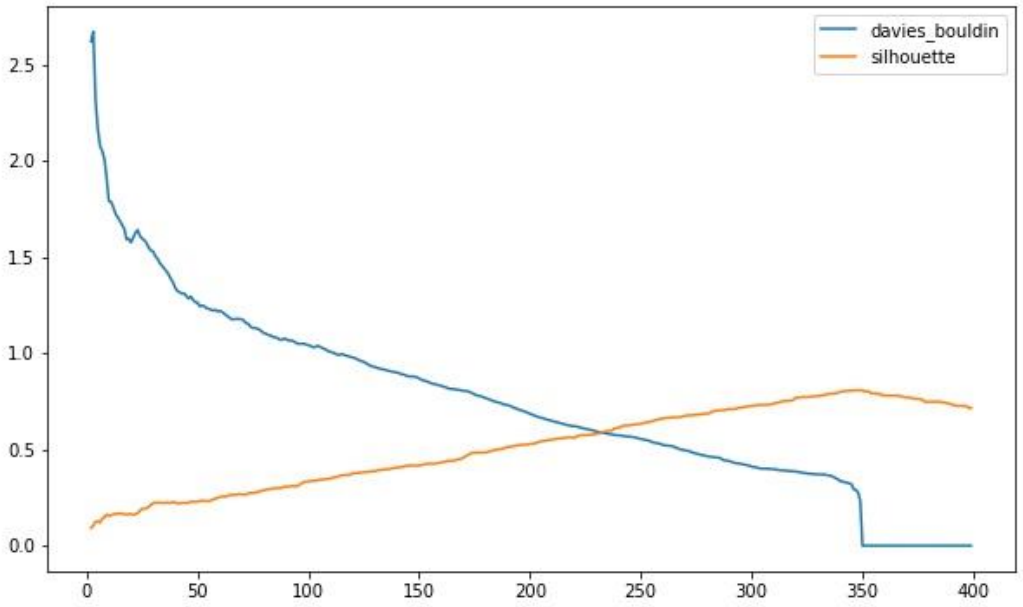

Рисунок 4 – Метрики качества кластеризации от количества кластеров

Наилучшее разбиение достигается при количестве кластеров $n = 350$. Поскольку это достаточно большое количество кластеров, был выполнен дополнительный анализ. А именно, к таблице был применён метод главных компонент. По критерию каменистой осыпи было определено оптимальное количество компонент для описания данных (восемь компонент, доля объясняющей дисперсии 0.679).

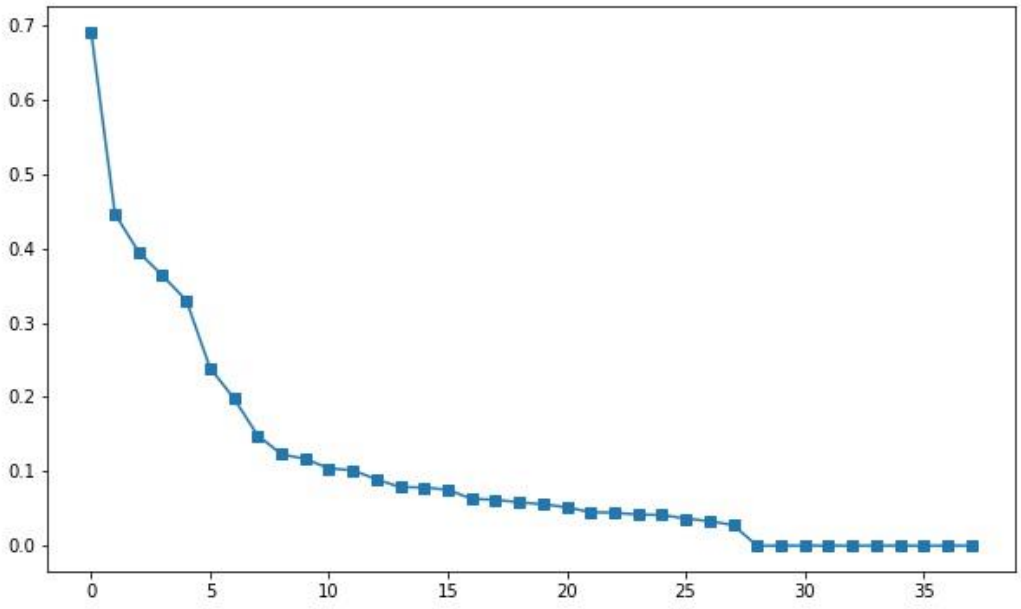

Рисунок 5 – Дисперсия компонент

Поскольку после проведения анализа новые факторы утратили свойство бинарности, в качестве метрики было использовано расстояние городских кварталов.

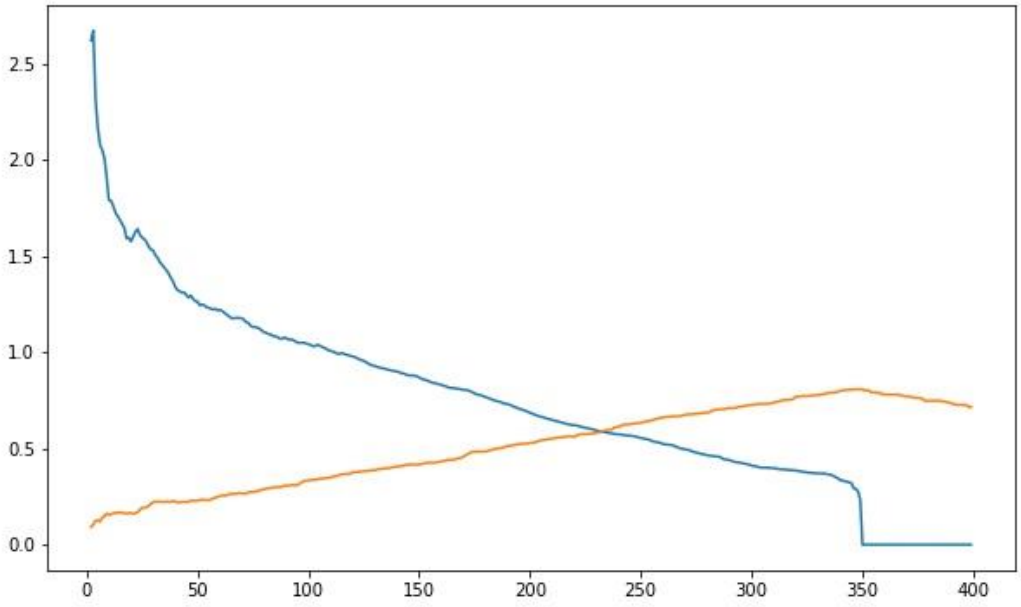

Рисунок 6 – Метрики качества кластеризации от количества кластеров

Визуально было выделено три разбиения в точках локального минимума индекса Дэвиса-Боулдина: на 30, на 55 и на 350 кластеров.

Разбиение на 55 кластеров выбран как основной результат работы.

2.2.2. Кластеризация по рядам динамики

Как уже было оговорено ранее, можно провести кластеризацию по рядам динамики. В качестве метрики снова использовалось шахматное расстояние. При таком подходе оптимальное количество кластеров значительно меньше, чем при предыдущем. Например, для типа товара «Бензак» минимум индекса Дэвиса-Боулдина достигается при $n = 4$.

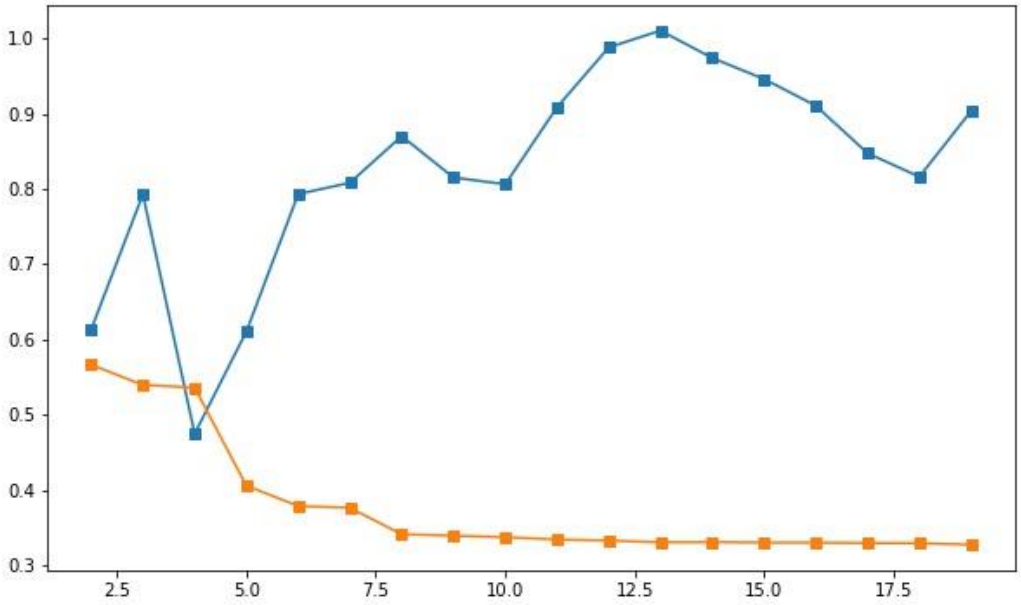

Рисунок 7 – Метрики качества кластеризации от количества кластеров

Аналогичная кластеризация была проведена для каждого типа товара. Результаты сохранены в файл clustering.tsv.

Основной результат сохранён в файл submission.tsv.

2.3. Результаты

Для проверки результатов был построен пример прогнозной модели. А именно, был выбран пример кластеризации. Была взята таблица, построенная при препроцессинге.

По этой таблице построена пуассоновская регрессия (обобщённая линейная модель без регуляризации). Выбран отдельный кластер. Внутри кластера по остаткам пуассоновской регрессии построен усредняющий временной ряд. Данный временной ряд спрогнозирован моделью SARIMA.

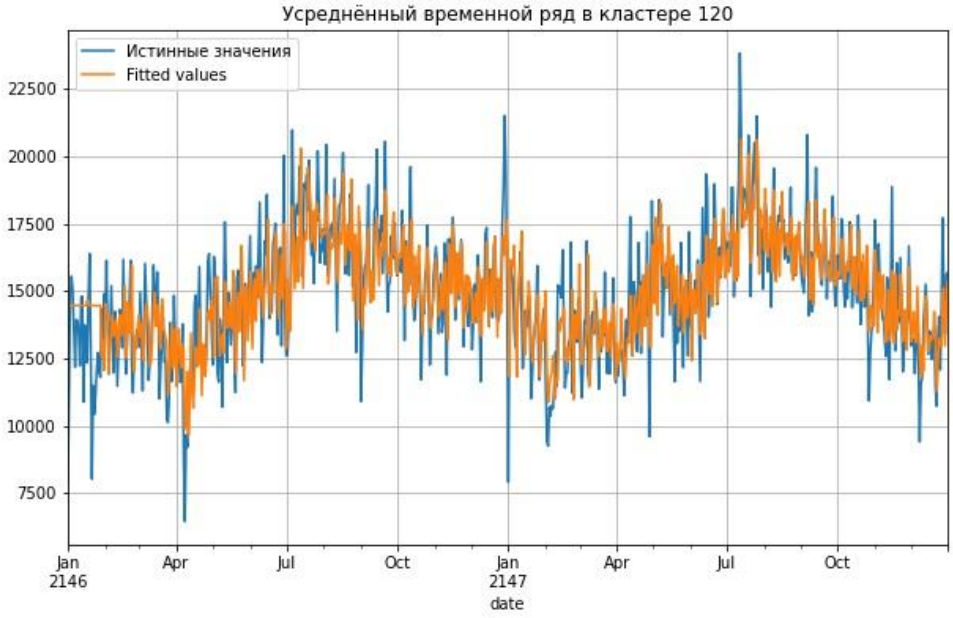

Рисунок 8 – Fitted values

Для усредняющего временного ряда в выбранном кластере был получен прогноз на неделю вперёд.

Рисунок 9 – Прогноз для усредняющего ряда

Для получения прогноза для каждого магазина были использованы мультипликаторы, рассчитанные как среднее отношение между временным рядом магазина и усредняющим рядом динамики.

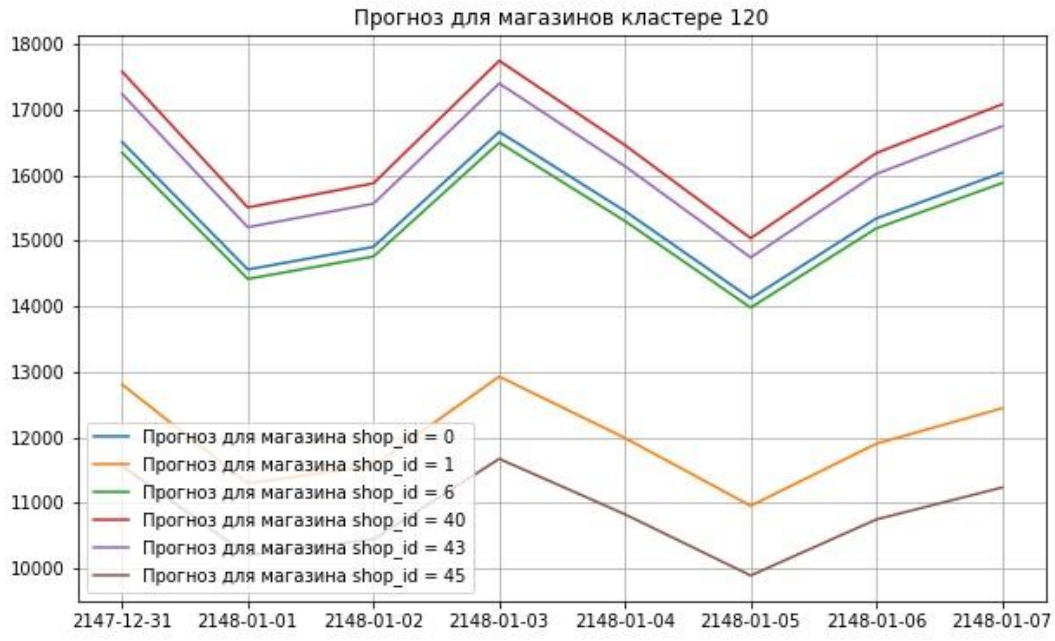

Рисунок 10 – Прогноз для магазинов в кластере

Предполагается построения аналогичных прогнозных моделей для каждого кластера.

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы были изучены предоставленные данные, построены гистограммы и графики. Проведён визуальный анализ данных, выполнена задача препроцессинга, в ходе которой построены таблицы для построения кластеризации и прогнозных моделей.

С использованием иерархического кластерного анализа и различных метрик было получено несколько разбиений магазинов на кластеры по факторам и по рядам динамики продаж для каждого типа товара.

Для полученного разбиения был приведён пример прогнозной модели с использованием пуассоновской регрессии и SARIMA.

Можно было бы улучшить результат кластеризации, если бы имелась более подробная информация о локациях, в которых расположены магазины, и информация о типах клиентов.