GeekBrains

Прогнозирование спроса: Использование исторических данных о продажах и других факторов для прогнозирования спроса на товары и услуги

Программа: Разработчик
Специализация: Инженер баз данных
Чертыковцев Дмитрий

Москва

2024

ОГЛАВЛЕНИЕ

Введение.....3

Методы сбора и подготовки данных о продажах к обработке......5

Исследование данных и определение взаимосвязи между спросом и различными факторами......9

Выбор и обучение модели прогнозирования спроса, используя алгоритм градиентного бустинга деревьев решений......13

Оценка точности модели и внесение корректировок по мере необходимости......16

Заключение.....18

Ход выполнения проекта.....19

Список литературы.....25

Введение

Что из себя представляет проект

Проект предназначен для того, чтобы спрогнозировать общее количество проданных товаров в каждом магазине. В данном проекте мы будем использовать тестовую выборку, данные взяты с сайта Kaggle.com.

Обоснование актуальности темы

Осуществление прогнозирования спроса на товары является важным и актуальным направлением в современной экономике. Такой прогноз позволит компаниям держать уровень запасов товаров на нужном уровне и управлять логистикой более эффективно. Это поможет избежать излишних запасов или недостатка товаров, их пересортицы, что, в свою очередь, приведет к снижению издержек и повышению эффективности бизнес-процессов. Кроме того, имея такие прогнозы, компании увеличивают конкурентоспособности за счет большей гибкости и адаптивности к изменениям рыночной ситуации. Также прогнозирование спроса помогает улучшить стратегическое планирование бизнеса, выявить тенденции и паттерны потребительского спроса. На основе этих данных компании могут разрабатывать новые продукты и услуги, оптимизировать ассортимент и улучшать взаимодействие с клиентами.

Цель проекта заключается в том, чтобы помочь компаниям оптимизировать свои запасы, планировать производство, улучшить маркетинговые стратегии и обеспечить более точное прогнозирование спроса на свои продукты или услуги. В результате правильного прогнозирования спроса компании могут сэкономить ресурсы, увеличить свою конкурентоспособность и улучшить обслуживание клиентов.

План работы.

1. Собрать и подготовить данные о продажах к обработке.

2 Исследовать данные и определить взаимосвязи между спросом и различными факторами.

3. Выбрать и обучить модель прогнозирования спроса, используя алгоритм градиентного бустинга деревьев решений.

4. Оценить точность модели и внести корректировки по мере необходимости.

Специализация дипломного проекта:

Анализ больших данных в рамках специализации Разработчик: инженер баз данных.

Инструменты и технологии:

Jupiter notebook, Pandas, NumPy, matplotlib, seaborn, sklearn, xgboost

Методы сбора и подготовки данных о продажах к обработке.

Для сбора и подготовки данных о продажах для прогнозирования спроса в нашем проекте используются тестовые данные с сайта Kaggle.com в виде csv файлов высокой очистки данных.

В реальных условиях, как правило, используются следующие методы сбора данных:

1. Импорт данных из базы данных о продажах из внутренних систем предприятия, таких как CRM (Customer Relationship Management) и ERP (Enterprise Resource Planning).

В данном методе мы идентифицируем необходимые данные, то есть определяем, какие именно данные о продажах нам необходимы для анализа. Это могут быть данные о продуктах, клиентах, заказах, продажах и т.д. Данные о продажах мы извлекаем напрямую из баз данных с помощью SQL-запросов или других инструментов для работы с базами данных

2. API-интеграция.

Многие CRM и ERP сторонние системы предоставляют API для интеграции с другими системами. Мы можем использовать API для извлечения данных о продажах из этих систем, осуществляя, таким образом, интеграцию с онлайн-платформами и маркетплейсами, чтобы получать данные о продажах в реальном времени.

3. Экспорт данных в формате CSV или Excel.

Многие системы CRM и ERP позволяют экспортировать данные в формате CSV или Excel. Мы можем регулярно экспортировать данные и использовать их для анализа.

4. Использование инструментов интеграции данных

Мы можем использовать специальные инструменты интеграции данных, такие как ETL (Extract, Transform, Load), для автоматизации процесса сбора данных о продажах из различных источников.

5. Использование отчетов и аналитики

Многие CRM и ERP системы предоставляют возможность создания отчетов и аналитики, которые мы можем использовать для анализа данных о продажах.

Вышеперечисленные методы можно комбинировать и настраивать так, как требует та или иная задача и зависит от конкретных потребностей и целей по анализу данных.

Методы подготовки данных к обработке:

1. Очистка данных от ошибок, дубликатов и пропущенных значений.

Это важный этап в обработке данных, который позволяет обеспечить их качество и достоверность. Для идентификации и удаления дубликатов мы проводим поиск и определение повторяющихся записей в данных. Дубликаты могут возникать из-за ошибок ввода данных или технических проблем. После идентификации дубликатов их можно удалить из набора данных. После этого обрабатываем ошибки. Ошибки в данных могут включать в себя неправильно введенные значения, некорректные форматы данных и другие неточности. Для исправления ошибок можно использовать автоматизированные методы, такие как проверка на соответствие формату и диапазону значений. Далее заполняем пропущенные значения, которые могут возникать из-за различных причин, например, из-за ошибок ввода или отсутствия информации. Для заполнения пропущенных значений можно использовать различные методы, такие как заполнение средними значениями, медианой или модой, или использование алгоритмов машинного обучения для предсказания пропущенных значений. В конце проверяем качество данных, чтобы убедиться, что данные соответствуют заявленным требованиям и готовы к дальнейшей обработке и анализу.

2. Преобразование данных в удобный формат для анализа, например, в формате таблицы или базы данных.

При преобразовании данных в удобный формат мы должны определить, какие данные нам необходимы для анализа и какие типы данных

они представляют. После этого создаем схему данных - разрабатываем структуру таблиц или базы данных, которая будет отражать нашу модель данных. Далее импортируем данные из источников в таблицу или базу данных. После загрузки выполним необходимые преобразования данных, такие как объединение таблиц, преобразование типов данных, создание новых переменных и т. д. Проверим данные: устраним ошибки, дубликаты и пропущенные значения в данных. Потом осуществим индексирование данных, то есть создадим индексы для ускорения доступа к данным при выполнении запросов. Проведем оптимизацию базы данных для улучшения производительности запросов. После всех произведенных шагов подготовим данные в удобном формате для проведения анализа, например, с помощью SQL запросов или инструментов для визуализации данных.

3. Обогащение данных путем добавления дополнительных переменных, таких как информация о клиентах, продуктах, времени и т. д.

Обогащение или добавлении данных о продажах товаров может осуществляться путем интеграции данных из различных источников. Например, мы можем добавить данные о погоде, праздниках, акциях и других факторах, которые могут влиять на спрос на товары. Это позволит создать более полную картину и улучшить точность прогнозирования спроса. Для обогащения данных о продажах мы можем использовать различные методы, такие как слияние данных из различных источников, использование внешних API для получения дополнительной информации, а также создание новых переменных на основе существующих данных. Важно учитывать, что при обогащении данных необходимо следить за качеством и достоверностью добавляемых переменных, чтобы избежать искажений в анализе и прогнозировании спроса на товары.

4. Применение методов агрегации данных для создания сводных таблиц или отчетов.

Данные методы используются для суммирования, усреднения или иной обработки данных с целью создания сводных таблиц или отчетов. В

контексте данных о продажах товаров, методы агрегации могут включать в себя следующие шаги:

- Группировка данных: данные о продажах товаров могут быть сгруппированы по различным параметрам, таким как дата продажи, категория товара, регион продажи и т.д.

- Суммирование данных: с помощью методов агрегации можно вычислить суммарные значения, например, общая сумма продаж за определенный период времени или общее количество проданных товаров.

- Усреднение данных: для анализа средних показателей, таких как средняя цена продукта, средний объем продаж и т.д., можно использовать методы усреднения.

- Создание сводных таблиц: после агрегации данных можно создать сводные таблицы, которые позволяют визуализировать и анализировать данные более удобным способом.

- Создание отчетов: на основе сводных таблиц можно создавать различные отчеты и дашборды для анализа и принятия управленческих решений.

Эти методы агрегации данных позволяют обобщить большие объемы информации о продажах товаров и выделить ключевые тренды и показатели для принятия бизнес-решений.

Исследование данных и определение взаимосвязи между спросом и различными факторами.

Для исследования данных и определения взаимосвязей между спросом и различными факторами можно использовать различные статистические методы и аналитические инструменты. Для процесса исследования данных можно применить следующие шаги:

1. Анализ корреляции. В данном шаге мы оцениваем корреляцию между спросом на товары и различными факторами, такими как цена, сезонность, маркетинговые активности и т.д. Это позволит определить, какие факторы имеют наибольшее влияние на спрос.

Для оценки корреляции можно использовать статистические методы, такие как коэффициент корреляции Пирсона или коэффициент корреляции Спирмена.

Коэффициент корреляции Пирсона измеряет линейную зависимость между двумя непрерывными переменными. Значение коэффициента корреляции Пирсона находится в диапазоне от -1 до 1. Значение ближе к 1 указывает на положительную корреляцию, ближе к -1 - на отрицательную корреляцию, а значение около 0 - на отсутствие корреляции.

Коэффициент корреляции Спирмена также используется для измерения степени связи между двумя переменными, но оценивает не только линейную зависимость, но и монотонную зависимость. Этот метод подходит для оценки корреляции между несколькими переменными, включая категориальные переменные.

После вычисления корреляции можно провести статистический анализ и интерпретацию результатов, чтобы определить, какие факторы имеют наибольшее влияние на спрос на товары. Такой анализ поможет выявить ключевые переменные, учитывающиеся при прогнозировании спроса и принятии управленческих решений.

2. Регрессионный анализ. Проведя регрессионный анализ, мы сможем определить, какие факторы являются статистически значимыми предикторами спроса на товары. Это позволит построить модель прогнозирования спроса на основе этих переменных.

Для проведения регрессионного анализа, как правило, используются такие библиотеки Python, как pandas, numpy, statsmodels, matplotlib или scikit-learn.

Сначала необходимо подготовить данные, включая переменные, которые могут влиять на спрос на товары, такие как цена, сезонность, маркетинговые активности и другие. Затем необходимо разделить данные на обучающий и тестовый наборы, чтобы оценить качество модели. После этого можно построить регрессионную модель, используя выбранный метод регрессии (например, линейная регрессия, логистическая регрессия и т. д.). Далее провести анализ значимости коэффициентов регрессии и определить, какие факторы статистически значимы для прогнозирования спроса на товары. После этого можно оценить качество модели с помощью различных метрик, таких как коэффициент детерминации (R-квадрат), средняя квадратичная ошибка (MSE) и другие. В конце мы сможем интерпретировать результаты и сделать выводы о влиянии различных факторов на спрос на товары. Важно понимать, что это лишь общий принципы процесса регрессионного анализа.

3. Кластерный анализ. Мы можем использовать кластерный анализ для группировки потребителей или товаров по схожим характеристикам. Это поможет выделить различные сегменты рынка и понять, какие факторы влияют на спрос в каждом из них.

Для проведения кластерного анализа можно использовать различные методы, такие как метод k-средних, иерархический кластерный анализ, методы DBSCAN и другие.

Метод k-средних. Этот метод разделяет данные на k кластеров, где k - заранее заданное количество кластеров. Он минимизирует сумму квадратов расстояний между точками в кластерах и их центроидами.

Иерархический кластерный анализ. Этот метод строит дерево кластеров, иерархически объединяя близкие кластеры. Можно использовать агломеративный или дивизионный подход.

Метод DBSCAN. Этот метод основан на плотности данных. Он может автоматически определять количество кластеров и обнаруживать выбросы. После применения алгоритма кластеризации можно изучить полученные группы потребителей или товаров, их характеристики и поведение, что поможет лучше понять спрос и адаптировать стратегии продаж и маркетинга.

4. Визуализация данных. Используем графики и диаграммы для визуализации взаимосвязей между спросом и нужными факторами.

Визуализацию можно построить с помощью линейных графиков, отображающих изменение спроса в зависимости от времени, цены или других факторов. Это позволит визуально оценить тенденции и взаимосвязи. Таже можно использовать диаграммы рассеяния. Такие диаграммы позволяют оценить корреляцию между спросом и другими переменными. На них можно увидеть, есть ли какая-либо зависимость между этими переменными. Кроме того, можно использовать столбчатые диаграммы, которые могут быть использованы для сравнения уровня спроса на разные товары или услуги в разные периоды времени или при разных условиях, Круговые диаграммы, полезные для визуализации доли спроса на различные товары или услуги в общем объеме спроса, а также Тепловые карты. Используя тепловые карты, можно визуализировать интенсивность спроса на товары в различных категориях или сегментах рынка.

Наглядное представление данных поможет лучше понять их структуру и взаимосвязи.

5. Машинное обучение. Можно применить методы машинного обучения, такие как случайный лес или нейронные сети, для построения

более сложных моделей прогнозирования спроса на основе большого объема данных.

При применении методов машинного обучения для анализа данных сначала выполняются обычные шаги: подготовка данных, их очистка, заполнение пропущенными значениями и преобразование данные в удобный формат для обучения модели. После этого нужно разделить данные на обучающий и тестовый наборы для оценки производительности модели, выбрать подходящую модель машинного обучения для прогнозирования спроса. Например, случайный лес и нейронные сети являются популярными моделями для таких задач. Далее нужно обучить выбранную модель на обучающем наборе данных, настроив параметры модели для достижения наилучших результатов. В конце мы оценим производительность модели на тестовом наборе данных с использованием метрик, таких как среднеквадратичная ошибка или коэффициент детерминации и, используем обученную модель для прогнозирования спроса на основе новых данных. После анализа результатов мы сможем оценить точность прогнозов, и, при необходимости, провести дополнительную настройку модели для улучшения результатов.

Выбор и обучение модели прогнозирования спроса, используя алгоритм градиентного бустинга деревьев решений.

При выборе модели машинного обучения для прогнозирования спроса на товары, важно учитывать несколько факторов:

1. Тип задачи. Нужно определить, является ли задача прогнозирования спроса на товары задачей регрессии (например, прогнозирование количества продаж) или классификации (например, прогнозирование категории спроса).

2. Количество данных. Убедитесь, что у имеется достаточно данных для обучения модели. Чем больше данных, тем сложнее и точнее может быть модель.

3. Характеристики данных. Изучив данные, мы сможем понять их структуру, наличие выбросов, пропущенных значений и т.д. Это поможет выбрать подходящую модель.

4. Интерпретируемость. Если важно понимать, как работает модель и какие признаки влияют на прогноз, то стоит выбирать модели с хорошей интерпретируемостью.

Градиентный бустинг деревьев решений (Gradient Boosting Decision Trees) - это техника машинного обучения для задач классификации и регрессии, которая строит модель предсказания в форме ансамбля слабых предсказывающих моделей, обычно деревьев решений. Считается одной из самых эффективных реализаций градиентного бустинга.

В основе XGBoost, библиотеки машинного обучения, лежит алгоритм градиентного бустинга деревьев решений. Такой алгоритм обладает несколькими преимуществами для задач прогнозирования спроса:

- Высокая точность: обычно обеспечивает высокую точность прогнозов за счет комбинирования нескольких деревьев решений.

- Высокая скорость и эффективность: оптимизирован для работы с большими наборами данных и обладает высокой скоростью обучения и предсказания

- Способность работать с различными типами данных: хорошо справляется с категориальными и числовыми признаками, а также с пропущенными значениями.

- Регуляризация: имеет механизмы регуляризации, которые помогают избежать переобучения модели.

- Поддержка различных функций потерь и критериев разделения: позволяет выбирать различные функции потерь и критерии разделения, что делает его гибким для различных задач.

- Параллельное обучение: поддерживает параллельное обучение на многих процессорах, что ускоряет процесс обучения модели.

- Поддержка кросс-валидации: позволяет проводить кросс-валидацию для оценки качества модели и подбора оптимальных гиперпараметров.

Важно помнить, что выбор модели всегда зависит от конкретной задачи, данных и предпочтений. Рекомендуется провести сравнительный анализ различных моделей, включая градиентный бустинг, чтобы выбрать наиболее подходящую для конкретной ситуации.

Обучение модели прогнозирования спроса с использованием XGBoost включает в себя несколько шагов:

1. Подготовка данных: подготовьте данные, включая признаки (features) и целевую переменную (спрос на товары).

2. Разделите данные на обучающий и тестовый наборы, чтобы оценить качество модели.

3. Определите параметры модели, такие как глубина деревьев, скорость обучения и количество деревьев.

4. Обучите модель на обучающем наборе данных.

5. Оцените качество модели на тестовом наборе данных, используя метрики оценки качества, такие как средняя абсолютная ошибка (MAE), средняя квадратичная ошибка (MSE) и другие.

6. Проведите оптимизацию модели, изменяя параметры и проводя кросс-валидацию, чтобы улучшить качество прогнозирования спроса.

Повторимся, что XGBoost - это мощный алгоритм градиентного бустинга, который обладает высокой точностью и способностью обрабатывать большие объемы данных. Он хорошо подходит для задач прогнозирования спроса на товары из-за своей способности улавливать сложные зависимости в данных и эффективно обрабатывать категориальные признаки.

Оценка точности модели и внесение корректировок по мере необходимости.

Точность прогноза зависит от используемой модели и самих данных. Когда лежащие в основе механизмы прогнозирования неизвестны, слишком сложны для понимания или известны не полностью, как розничные продажи, применяют простую статистическую модель. Популярными классическими методам, относящимися к этой категории, являются ARIMA, методы экспоненциального сглаживания, такие как Holt-Winters, и метод Theta, который используется менее широко.

Подходы к компьютерному обучению, в том числе алгоритмы случайного леса, стали частью инструментария прогнозиста. Свою эффективность также показали рекуррентные нейронные сети в случае достаточного количества данных. Невозможно изначально знать, какая модель приведет к оптимальной производительности и выполнит поставленную задачу. Поэтому необходимо применить различные методы прогнозирования спроса, чтобы изучить их особенности и выявить наиболее эффективные среди них для решения текущей задачи.

Для оценки точности модели прогнозирования спроса можно использовать различные метрики, такие как средняя абсолютная ошибка (MAE), среднеквадратичная ошибка (MSE), коэффициент детерминации (R ** 2) и другие.

Чтобы оценить точность модели, нужно:

- разделить данные на обучающий и тестовый наборы;

- обучить модель на обучающем наборе;

- оценить ее производительность на тестовом наборе, используя выбранную метрику.

Если точность модели не удовлетворяет требованиям, можно внести корректировки, такие как изменение гиперпараметров модели, добавление или удаление признаков, увеличение объема данных для обучения и т. д. При

внесении корректировок важно следить за изменениями в производительности модели и выбирать те, которые приводят к улучшению точности прогнозов.

Важно понимать, что существуют ограничения прогнозирования спроса и продаж с помощью машинного обучения:

1. Нужен качественный набор данных: для достижения точных прогнозов требуется наличие большого количества высококачественных данных, включающих различные факторы, которые могут влиять на спрос и продажи. Недостаток данных или их низкое качество могут снизить точность прогнозов.

2. Неучтенность экономических и социальных факторов: Модели машинного обучения могут не учитывать экономические и социальные факторы, которые могут значительно влиять на спрос и продажи, например, изменение политической ситуации, экономического климата или моды.

3. Проблема предсказания новых товаров: Предсказание спроса на новые товары достаточно сложная задача, так как отсутствуют данные о предыдущих продажах. Это ограничение может затруднить точное прогнозирование спроса и продаж.

4. Переразметка данных: Изменение бизнес-логики или условий спроса требует переразметки данных и повторного обучения модели, что может занимать значительное время и ресурсы.

Заключение

В заключение необходимо подвести итоги нашего проекта. По результатам проведенных исследований на тестовом наборе мы получили файл с результатами прогноза, что позволит нам планировать дальнейшие действия по оптимизации бизнес-процессов и увеличении прибыли.

Укажем на основные моменты проекта:

1. Методы прогнозирования спроса:

- Регрессионный анализ. Строит математическую модель, которая связывает спрос с независимыми переменными, такими как цена, сезонность и промо-акции.

- Методы скользящего среднего. Учитывают средние значения спроса за определенный период времени, чтобы предсказать будущий спрос.

- Методы экспоненциального сглаживания. Придают больший вес недавним данным, чем историческим, для прогнозирования спроса.

- Нейронные сети. Мощные алгоритмы машинного обучения, которые могут изучать сложные взаимосвязи в данных.

2. Факторы, влияющие на спрос:

- Цена
- Сезонность
- Промо-акции
- Экономические условия
- Конкурентная среда
- Демографические факторы

Ход выполнения проекта

Проект выполнен в файле «analis.ipynb» и будет приложен к проекту

Основные выводы:

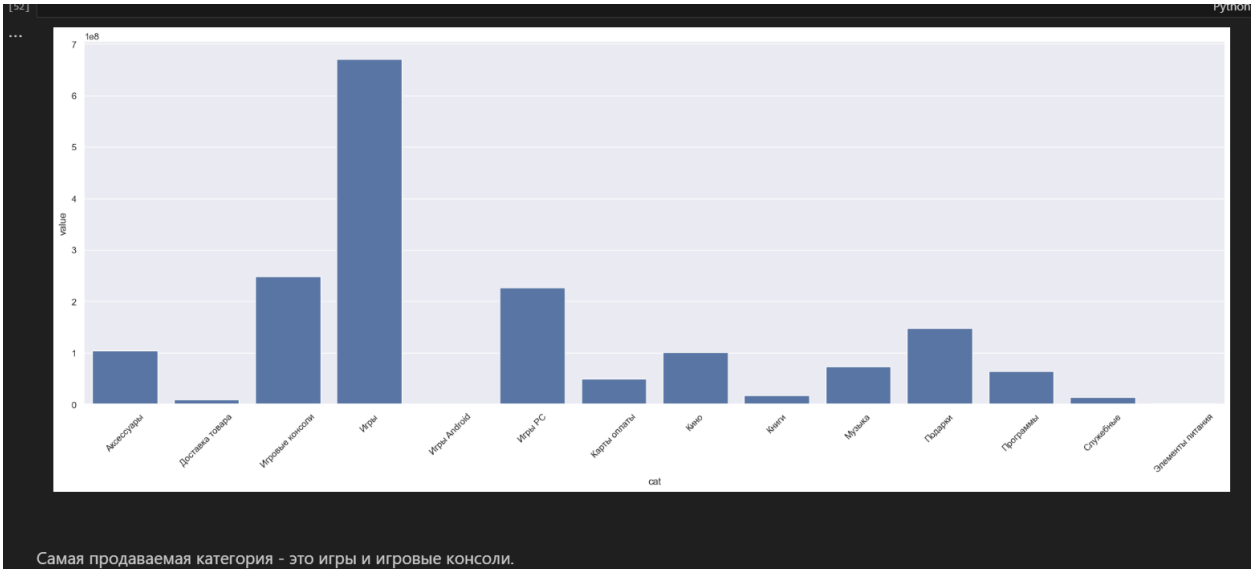
Взяли тестовый датасет

```
# Загрузка данных для анализа
test = pd.read_csv('test.csv')
item_cat = pd.read_csv('item_categories.csv')
items = pd.read_csv('items.csv')
shops = pd.read_csv('shops.csv')
sales = pd.read_csv('sales_train.csv')
```

[8] ✓ 0.9s Python

```
# Для экономии ресурсов уменьшим размер данных
sales = sales.loc[sales["shop_id"].isin(test["shop_id"].unique()), :]
sales = sales.loc[sales["item_id"].isin(test["item_id"].unique()), :]
```

[9] ✓ 0.2s Python



Самая продаваемая категория - это игры и игровые консоли.

Топ 20 самых продаваемых товаров:

```
comp_sales.groupby(by='item_name').sum()['item_category_id'].sort_values(ascending=False).head(20)
```

[53]

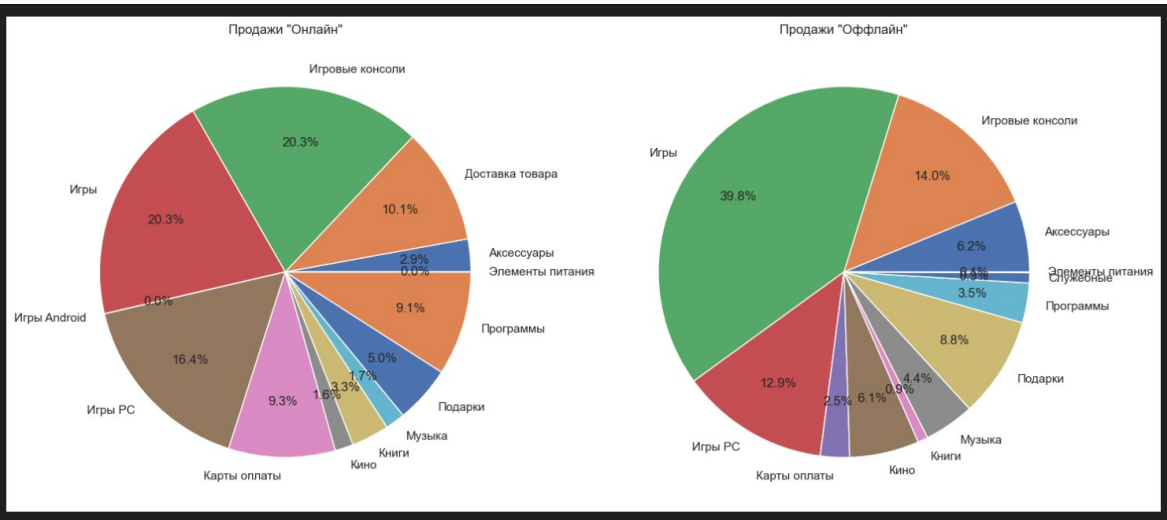
```
item_name
Фирменный пакет майка 1С Интерес белый (34*42) 45 мкм                                  79804
Прием денежных средств для 1С-Онлайн                                                   74497
Kaspersky Internet Security Multi-Device Russian Edition. 2-Device 1 year Renewal Box  69900
Kaspersky Internet Security Multi-Device Russian Edition. 2-Device 1 year Base Box     68025
Элемент питания DURACELL LR06-BC2                                                      56274
Настольная игра Уно                                                                    55185
Настольная игра Манчкин Цветная версия, арт. 1031                                      52800
Настольная игра Мафия  Вся семья в сборе (карточная игра)  арт. 1070                   52390
Настольная игра World of Tanks Rush арт.1123                                           50624
1С:Бухгалтерия 8. Базовая версия                                                       50516
Office Home and Student 2013 32/64 Russian Russia Only EM DVD No Skype                 50175
Головоломка Кубик Рубика 3х3 без наклеек, мягкий механизм                              49714
Dr.Web Security Space КЗ 2 ПК/2 года (картонная упаковка)                              49125
Магический шар 8 оригинальный                                                          48921
Элемент питания DURACELL LR03-BC2                                                      48804
Фигурка Minecraft Series 1 Player Survival Pack 3"                                     48456
ЛЕПС ГРИГОРИЙ  The Best  3CD (фирм.)                                                   47190
Настольная игра Свинтус Правила этикета (новая версия), арт. 1059                      46280
DEL REY LANA  Born To Die  The Paradise Edition  2CD                                    44440
Настольная игра Свинтус (новая версия), арт. 1058                                      44135
Name: item_category_id, dtype: int64
```

Топ 20 самых непопулярных товаров:

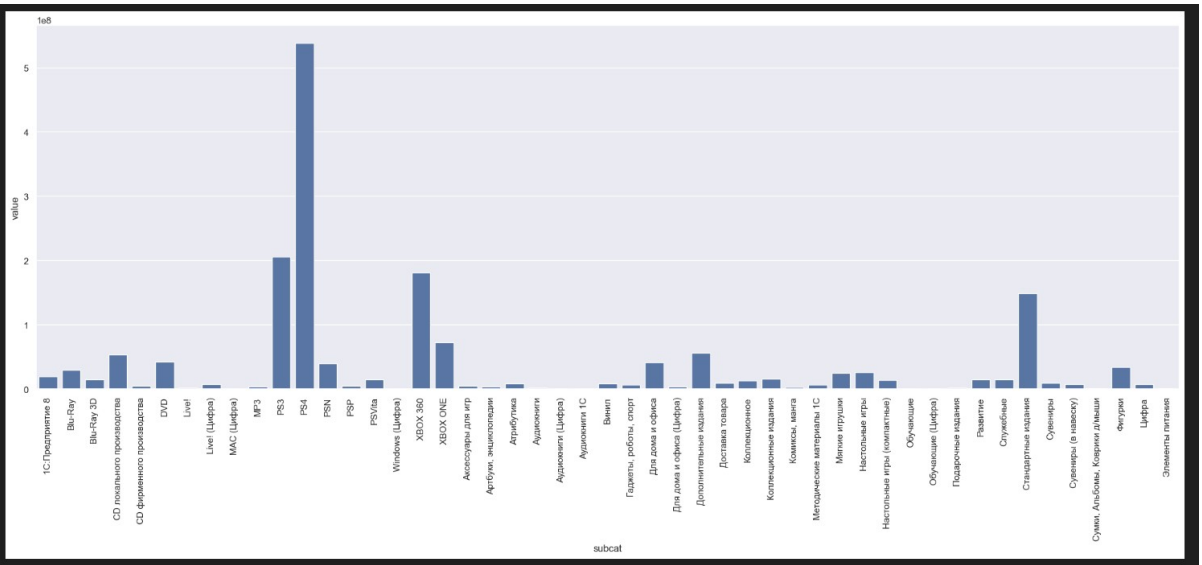
```
comp_sales.groupby(by='item_name').sum()['item_category_id'].sort_values(ascending=False).tail(20)
```

[54]

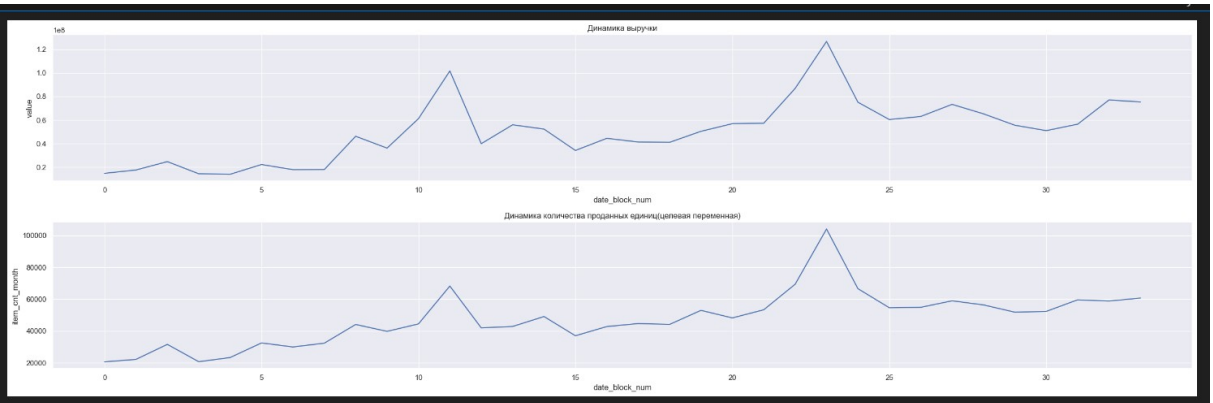
```
item_name
Карта оплаты Windows: 2500 рублей [Цифровая версия]                                                    36
Комплект «Sony PlayStation 4 (1Tb) Black (CUH-1208B)» + игра «Assassin's Creed: Синдикат» + игра «Watch_Dogs»  36
Комплект предзаказа на консоль Microsoft Xbox One 1TB гибридной памяти с геймпадом "Elite" [Цифровая версия]  32
Europa Universalis IV. The Art of War Collection [PC, Цифровая версия]                                 31
Lord of the Rings: War in the North [PC, Цифровая версия]                                              31
Fallout 4. Season Pass [PC, Цифровая версия]                                                           31
WRC 5 [PC, Цифровая версия]                                                                            31
Europa Universalis IV. DLC Collection [PC, Цифровая версия]                                            31
Комплект Grand Theft Auto V + Great White Shark Cash Card (активация только в России) [PC, Цифровая версия]  31
Ведьмак 3: Дикая Охота - Дополнение "Каменные Сердца" [PC, Цифровая версия]                            31
Batman: Рыцарь Аркхема. Premium Edition [PC, Цифровая версия]                                          31
Средиземье: Тени Мордора. Game of the Year Edition (Upgrade) [PC, Цифровая версия]                     31
Just Cause 3 [PC, Цифровая версия]                                                                     31
Universal: Кабель HDMI Gioteck XC-4 высокоскоростной, Ethernet, 1.8м, 3D, 1080p, v1.4, плоский, серы  30
PS4: Контроллер игровой беспроводной камуфляжный (Dualshock 4 Cont Urban Cammo: CUH-ZCT1: SCEE)       27
Transformers: Devastation [Xbox One, английская версия]                                                24
Guitar Hero Live. Контроллер "Гитара" [PS4, английская версия]                                         20
Комплект предзаказа на Беспроводной геймпад для Xbox One "Elite" [Цифровая версия]                    14
Комплект силиконовых чехлов для PS Move Motion & Navigation контроллеров - Graphite                    10
Комплект силиконовых чехлов для PS Move Motion & Navigation контроллеров - Red                          2
Name: item_category_id, dtype: int64
```



В сфере онлайн покупок доминируют Игры, Игровые консоли и Игры PC, похожая ситуация и с Оффлайн продажами, однако, доля продаж игровых консолей меньше, при этом весомую часть занимают Подарки, в Онлайн сегменте также можно выделить большую долю продаж Карт оплат и Программ, а также доставку. При этом доля оплаты за доставку превышает стоимость продаж большинства групп товаров.



Лидерами продаж выступают консоли от Sony и игры к ним.

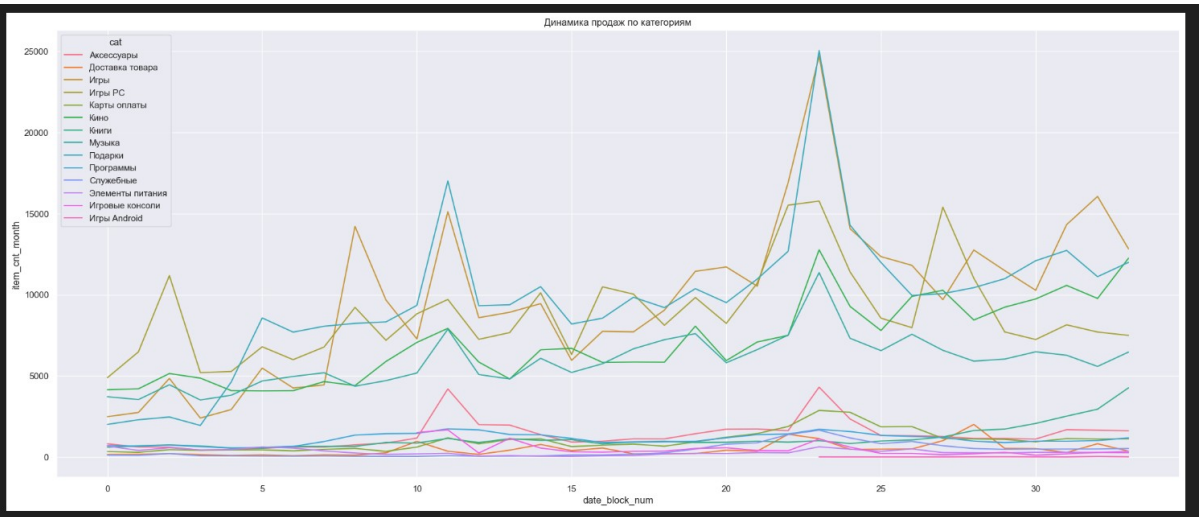


Наблюдается восходящий тренд, а также явные пиковые точки продаж. Предположительно это связано с новогодними праздниками. Проверим:

```
[58]  comp_sales.groupby(by='date_block_num').sum().sort_values(by='value', ascending=False)['value'].head(2)
```

```
date_block_num
23    1.267673e+08
11    1.016969e+08
Name: value, dtype: float64
```

Учитывая, что нумерация месяцев начинается с 0 и с 1-го января 2013 года, то 23 и 11 месяц это декабрь 2014 и декабрь 2013 соотвественно.



Наблюдается восходящий тренд на самые популярные товары, также наиболее яркие пики продаж заметны в декабре. Стоит обратить внимание и на периодические большие скачки продаж по играм, причиной могут быть сезонные распродажи от издателей или премьеры новых продуктов.

	shop_id	item_id	date_block_num	item_cnt_month	item_price
0	2	33	0	1.0	499.0
1	2	482	0	1.0	3300.0
2	2	491	0	1.0	600.0
3	2	839	0	1.0	3300.0
4	2	1007	0	3.0	449.0

Как мы видели выше, распределение продаж по точкам не однородно, следовательно вероятны ситуации, когда какой-то товар не продавался в магазине в определённом периоде. Для повышения качества прогноза, необходимо явно выделить такие ситуации, для этого необходимо расширить наш датасет и включить в него все возможные комбинации 'item_id' и 'shop_id' помесячно

```
# предсказания для оценочного набора данных
predictions = model.predict(x_test).clip(0,20)

# предсказание для валидационного набора данных
pred_val = model.predict(x_valid)
```

[82] Python

```
model.best_score
```

[83] Python

```
0.7390607675769325
```

Ошибка довольно низкая.

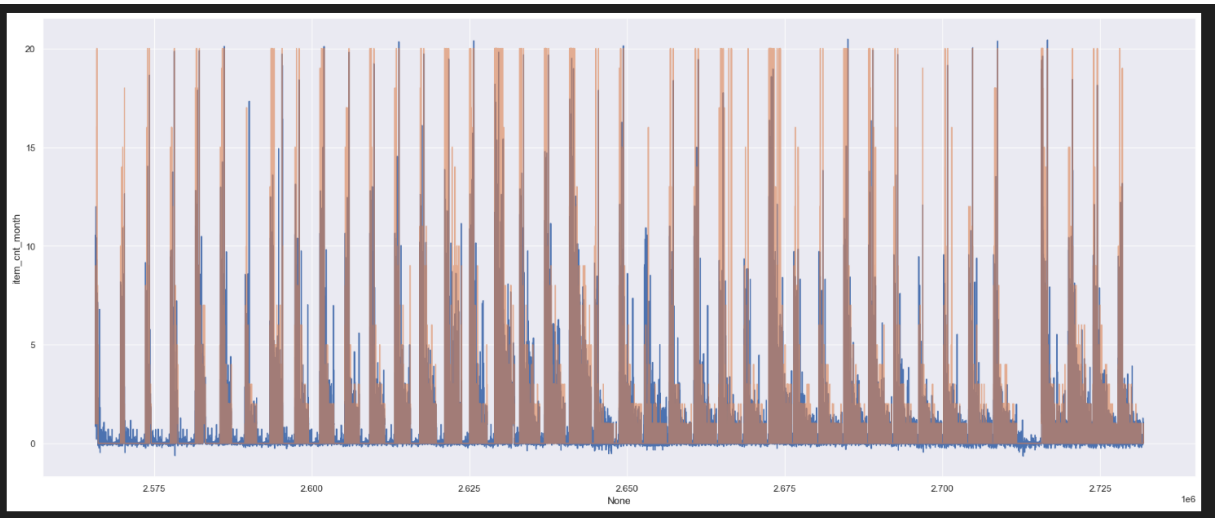
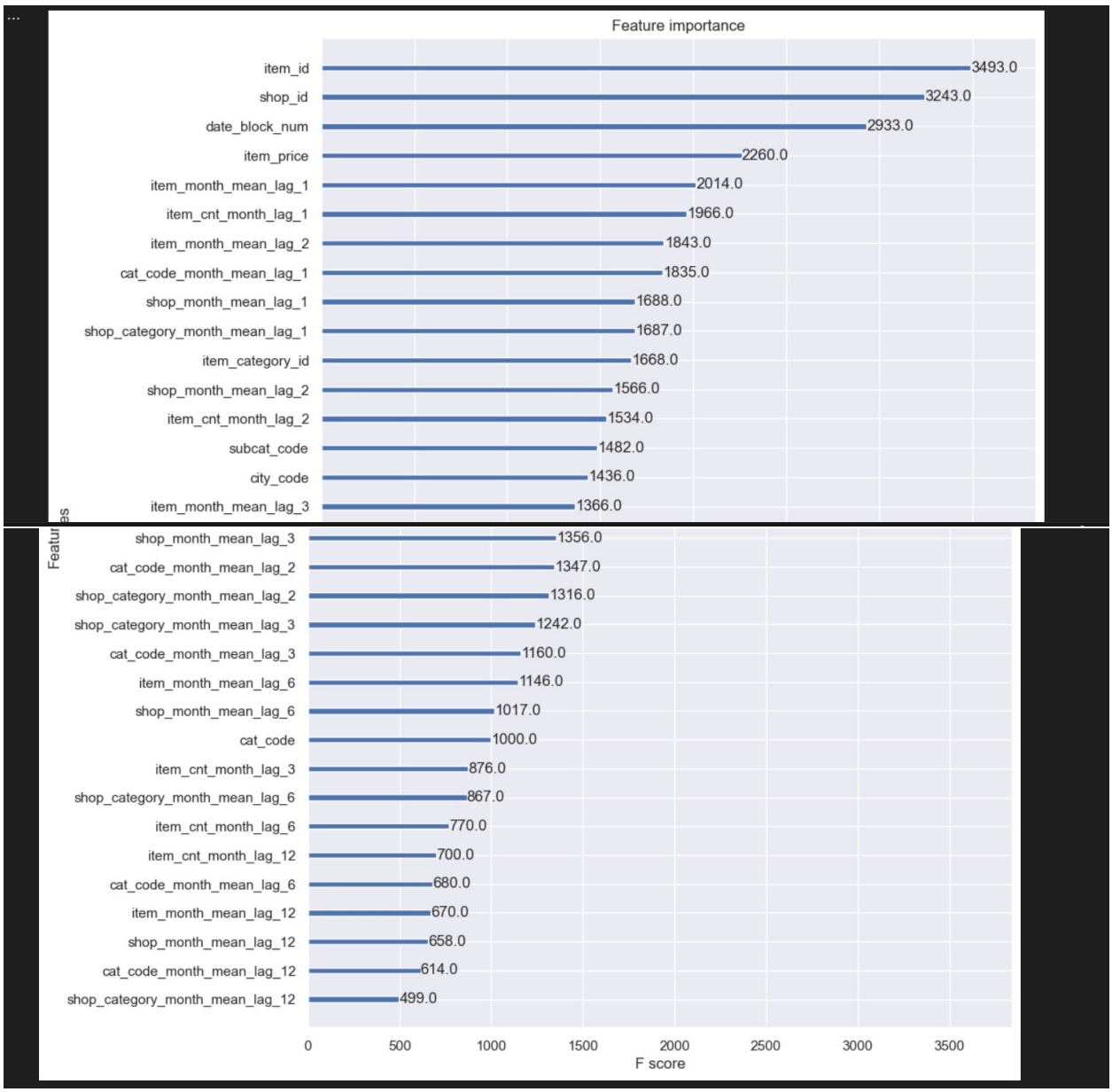


График прогноза довольно близко повторяет истинные значения. К тому же ошибка на валидации довольно низка - модель можно использовать для прогноза.



- модель можно использовать для прогноза;

- признаки со сдвигами довольно сильно влияют на итоговый результат.

Список литературы

1. Статья https://intelad.ru/iskusstvennyj-intellekt-v-marketinge/prognozirovanie_sprosa_i_prodazh_s_pomoschju_mashinnogo_obuchenija/
2. Статья https://habr.com/ru/companies/ozontech/articles/431950/
3. Уэс Маккинни, Python для анализа данных
4. Герон Орельен, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow
5. Научная статья Пилипенко А.Ю. Прогнозирование спроса на товары средствами машинного обучения
6. Научная статья Рогулин Р.С., Обзор прикладных основ использования аналитики данных и машинного обучения в прогнозировании спроса
7. Научная статья, Грошева Е.В., Особенности прогнозирования спроса на рынке продукции промышленного назначения.