2022

Дипломная работа

НА ТЕМУ «АНАЛИЗ СУММЫ ПРОДАЖ АЛКОГОЛЬНОЙ ПРОДУКЦИИ В США»

АВТОР: ВИКТОР НИКИТЕНКО

Оглавление

Ведение 2

Цель 2

Задачи 2

Выбор инструментов для выполнения работы 2

Знакомство с данными 3

Загрузка данных 3

Предобработка данных 3

Заключение 3

EDA (exploratory data analysis) или разведочный анализ данных 4

Выполнение расчёта основных статистических метрик 4

Заключение 5

Построение моделей 6

Подготовка данных для моделей 6

Модель 1. Sarimax 6

Модель 2: PROPHET 8

Модель 3: «Экспоненциальное сглаживание» (Exponential smoothing) 10

Заключение 11

Ведение

Для анализа была выбрана выборка с сумами розничной продажи алкогольной продукции в США в период с 1992 года по 2018 год. Суммы указаны в миллионах долларах.

Цель

Проведение исследования данных и построение прогноза суммы продаж алк огольной продукции.

Задачи

1. Провести анализ данных о суммах продаж алкогольной продукции;
2. Построить прогноз суммы продаж алкогольной продукции
   Гипотеза: увеличение суммы продаж в дальнейшем с сохранением сезонности.

Выбор инструментов для выполнения работы

1. Выборка сданными по суммам продаж алкогольной продукции в формате csv
   Файл:https://github.com/Viktor193/Diplom_innopolis/blob/de798092f5829ad2e9f46a907165a52283e07bc6/Retail_Sales_Beer_Liquor.csv

2. Язык программирования Phyton при использовании среды Google Colab
   Файл:https://github.com/Viktor193/Diplom_innopolis/blob/07fe611608b7e5316535901cd74217cc4fbdb77d/%D0%98%D1%82%D0%BE%D0%B3%D0%BE%D0%B2%D0%B0%D1%8F_%D0%B0%D1%82%D1%82%D0%B5%D1%81%D1%82%D0%B0%D1%86%D0%B8%D1%8F_%D0%9D%D0%B8%D0%BA%D0%B8%D1%82%D0%B5%D0%BD%D0%BA%D0%BE_%D0%92_%D0%92.ipynb

Знакомство с данными

Загрузка данных

1. Загрузка выполнялась с помощью spark, файл расположен на github.com, при запуске не требуется дополнительно его подгружать в Google Colab;
2. Выполнено проверка формата данных – в выборке существует два поля:
    a. **«DATA»**:
        i. При загрузке определился формат string;
        ii. В поле указана дата в формате ГГГГ-ММ-ДД, при этом для каждого значения указан день=01, т.е. фактически поле обозначает месяц конкретного года.
    b. **«MRTSSM4453USN»**:
        i. При загрузке определился формат string;
        ii. В поле указано значение суммы продаж в миллионах долларах за месяц из поля «DATA».

Предобработка данных

1. Поле «MRTSSM4453USN» переименовано в более информативное «**Volume of sales**».
2. Изменен формат данных:
    a. «**DATA**» - Date;
    b. «**Volume of sales**» - Double.
3. При проверке пустых значений не обнаружено, дополнительных преобразований не потребуется.

Заключение

Выполнена первоначальная обработка данных, возможно переходить в следующему этапу.

EDA (exploratory data analysis) или разведочный анализ данных

Выполнение расчёта основных статистических метрик

1. Выполнена преобразование данных в dataframe pandas для построения графиков.
2. Индексом было принято сделать поле DATA.
3. Расчёт основных статистических метрик (таблица 1):

	VOLUME OF SALES	DAY	MONTH
COUNT	324.000000	324.0	324.000000
MEAN	2972.895062	1.0	6.500000
STD	1010.218574	0.0	3.457392
MIN	1501.000000	1.0	1.000000
25%	2109.000000	1.0	3.750000
50%	2791.000000	1.0	6.500000
75%	3627.250000	1.0	9.250000
MAX	6370.000000	1.0	12.000000

Таблица 1.

4. График сумм продаж алкогольной продукции по годам (рис.1):

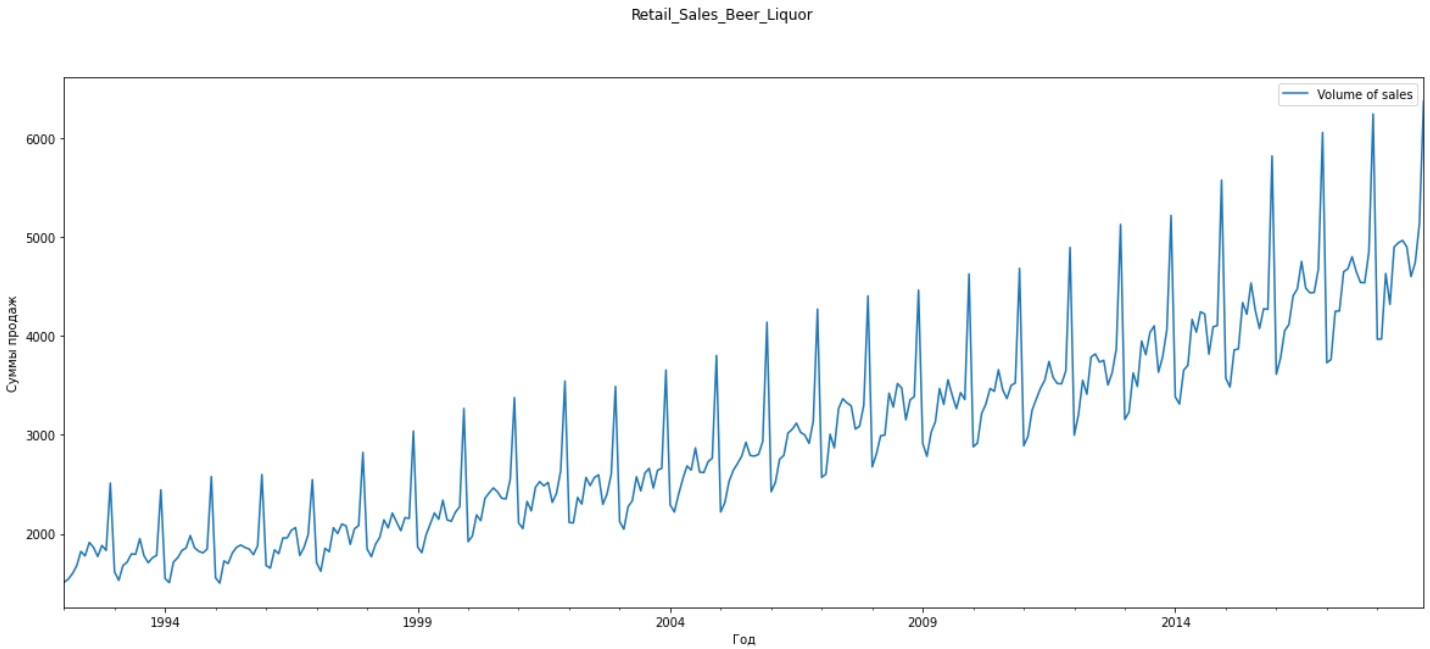

Рисунок 1.

5. Гистограмма для определения распределения данных (рис.2):

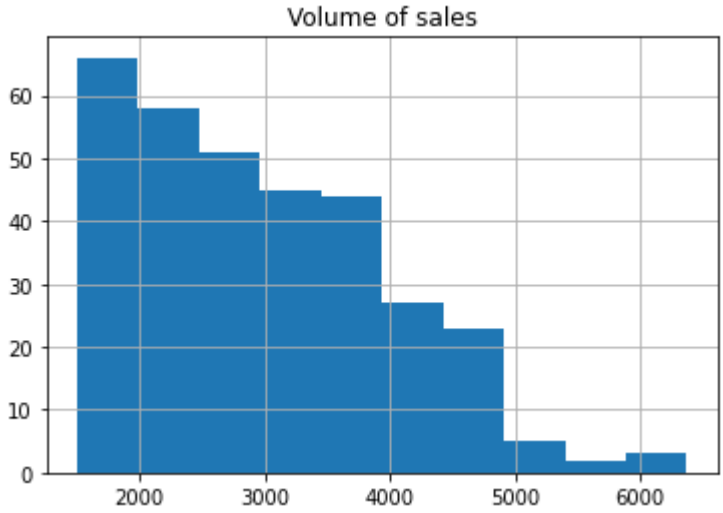

Рисунок 2.

Заключение

1. Общий тренд восходящий: сумма продаж с каждым годом увеличивается.
2. Присутствует сезонное увеличение суммы продаж.
3. Больше всего месяцев с наименьшими суммами продаж, чем выше су мма продаж тем меньше месяцев с такими суммами.

Гипотеза: увеличение суммы продаж в дальнейшем с сохранением сезон ности.

Построение моделей

Подготовка данных для моделей

1. Сформировали тестовую и обучающую выборки:
    a. Тестовая: 1 год;
    b. Обучающая выборка: остальные 26 лет.
2. Декомпозиция временного ряда (рис.3):

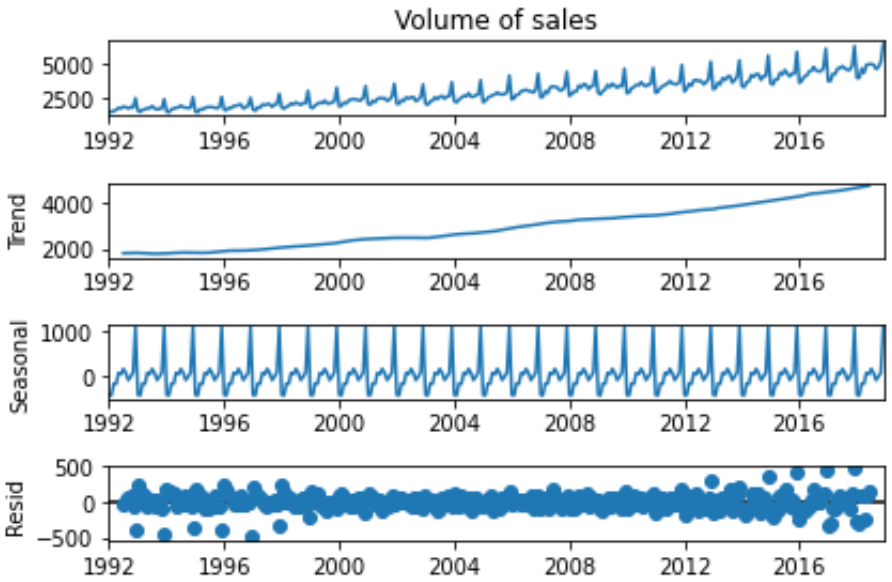

Рисунок 3.

    a. Положительный (восходящий) тренд;
    b. Сезонность в течении года.

Модель 1. Sarimax

1. Для построения сезонности выбран 1 год (12 месяцев), в результате показатели модели: **SARIMAX(4, 1, 3)x(2, 1, [1], 12)**.
2. После создания модели с параметра SARIMAX(4, 1, 3)x(2, 1, [1], 12) и обучении на обучающей выборке параметры оценки модели:
    a. Средняя абсолютная ошибка (MAE): **66.06013915**;
    b. Средняя квадратическая ошибка (MSE): **7896.543616**;
    c. Среднеквадратическая ошибка (RMSE): **88.86249837**;
    d. Средняя абсолютная ошибка в процентах (MAPE): **1.441353299**.

3. Результаты прогноза на год (рис.4):

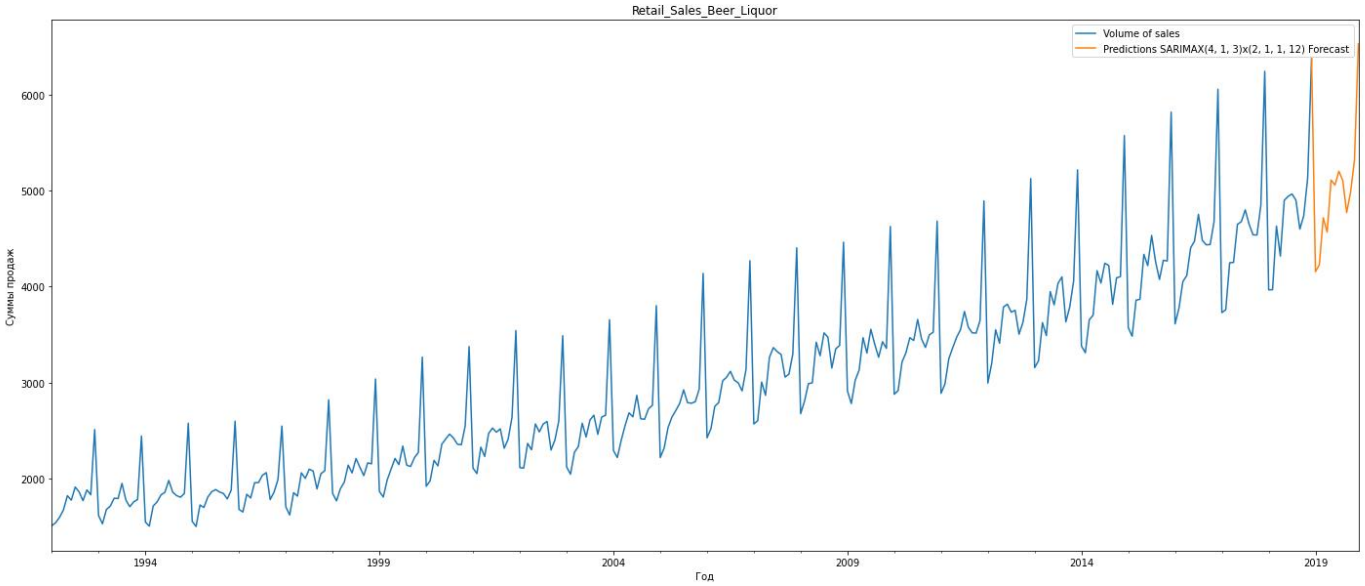

Рисунок 4.

4. Выводы работы метода SARIMAX:
    a. Модель показала себя хорошо: RMSE=88.86 – это хороший показатель;
    b. Процент рассчитанной ошибки MAPE=1.44%, это хороший результат;
    c. Согласно графику, на будущее видим, что тренд и высота амплитуда были отображены корректно, общая динамика прослеживается.

Модель 2: PROPHET

1. Подготовили данные для обучения модели, создали модель, обучили её.
2. При анализе модели было определено:
    a. Восходящий тренд (рис. 5):

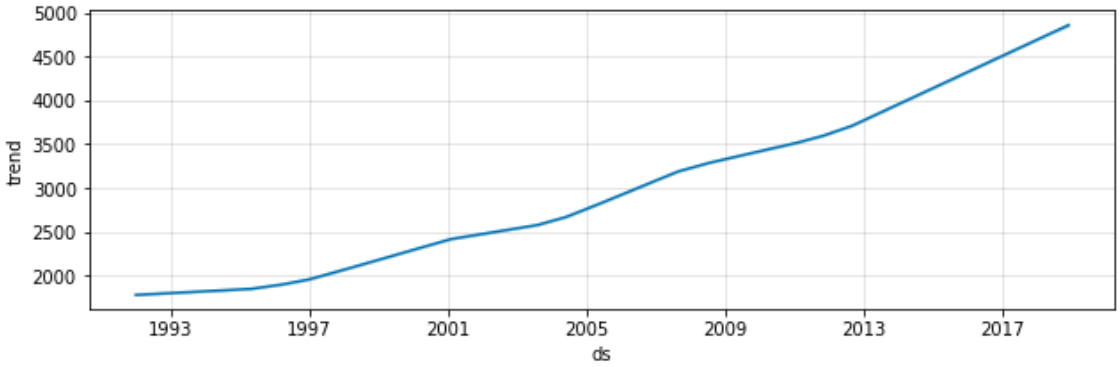

Рисунок 5

    b. Годовую сезонность (рис.6):

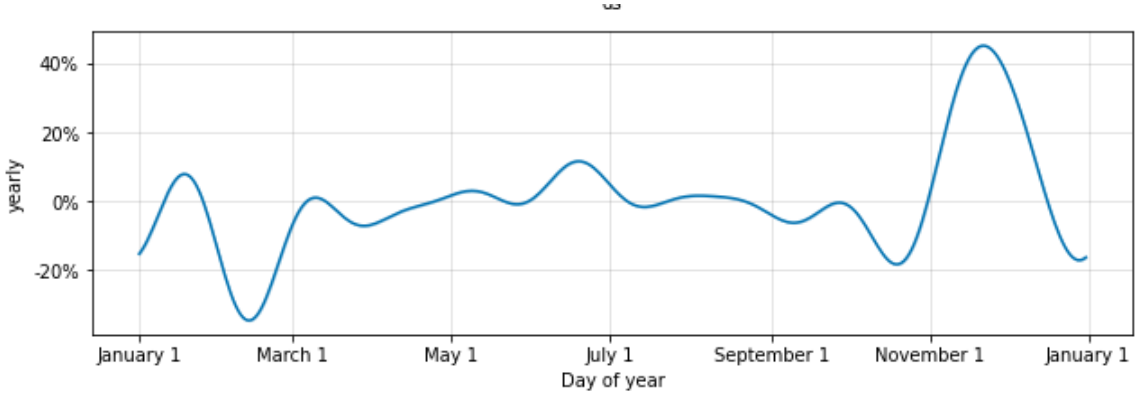

Рисунок 6

3. Результаты прогноза на год (рис.7):

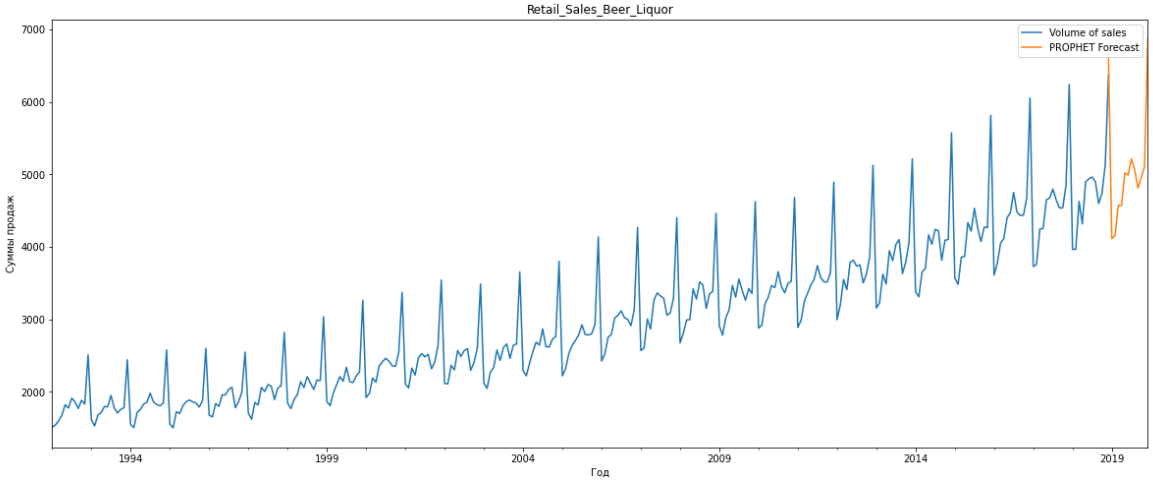

Рисунок 7.

4. Параметры оценки модели:
    a. Средняя абсолютная ошибка (MAE): **98.73289647**;
    b. Средняя квадратическая ошибка (MSE): **17973.33688**;
    c. Среднеквадратическая ошибка (RMSE): **134.0646743**;
    d. Средняя абсолютная ошибка в процентах (MAPE): **1.947700413**.
5. Выводы работы метода PROPHET:
    a. Модель показала себя хорошо: RMSE=134.06- это хороший показатель.
    b. Процент рассчитанной ошибки MAPE=1.94%, это хороший результат.
    c. Согласно графику, на будущее видим, что тренд и высота амплитуда были отображены корректно, общая динамика прослеживается.

Модель 3: «Экспоненциальное сглаживание» (Exponential smoothing)

1. Был выбран метод Хольта-Винтера, так как данный метод учитывает тренд, сезонность.
2. Создали модель с параметрами:
    a. Тренд: положительный;
    b. Период сезонности: 12 месяце;
    c. Аддитивный сезонный период;
    d. Использование преобразования Бокса-Кокса.
3. Обучили модель на обучающей выборке.
4. Результаты прогноза на год (рис.8):

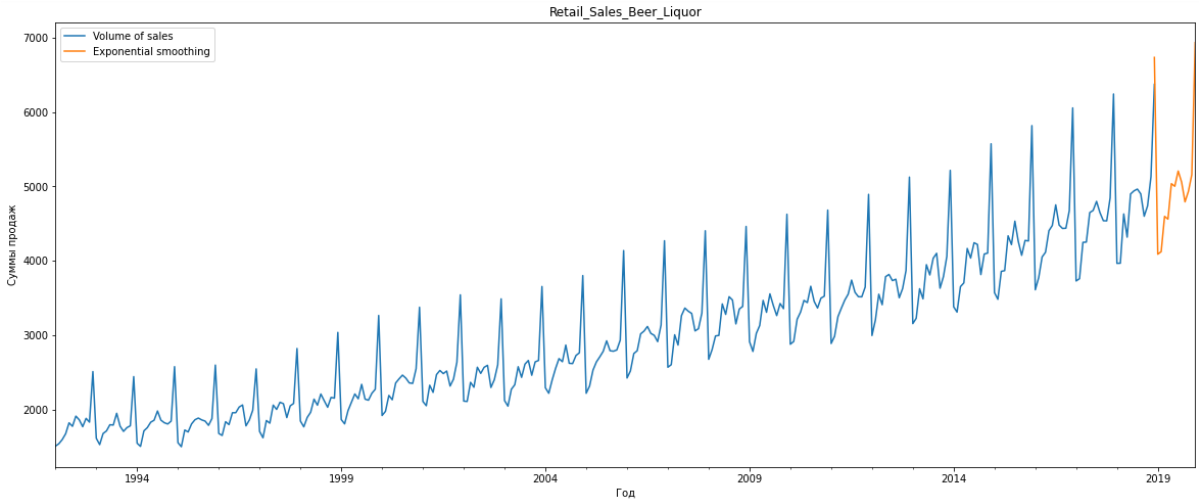

Рисунок 8

5. Параметры оценки модели:
    a. Средняя абсолютная ошибка (MAE): **103.5647038**;
    b. Средняя квадратическая ошибка (MSE): **21686.4567**;
    c. Среднеквадратическая ошибка (RMSE): **147.2632225**;
    d. Средняя абсолютная ошибка в процентах (MAPE): **2.020402968**.
6. Выводы работы метода Exponential smoothing:
    a. Модель показала себя хорошо: RMSE=147.26- это хороший показатель.
    b. Процент рассчитанной ошибки MAPE=2.02%, это хороший результат.
    c. Согласно графику, на будущее видим, что тренд и высота амплитуда были отображены корректно, общая динамика прослеживается.

Заключение

1. Проведен анализ данных с использованием современных методов обработки информации.
2. Рассчитаны основные статистически метрики, позволяющие судить о характере суммы продажи алкогольной продукции.
3. Подтверждена гипотеза: увеличение суммы продаж в дальнейшем с сохранением сезонности.

   Прогнозная модель позволила зафиксировать сохранение тенденции роста суммы продаж по сравнению с предыдущим годом, а также сохранение амплитудных значений в период новогодних праздников.
4. Сравнение моделей:
   a. Все модели используются для прогнозирования одномерных данных временных рядов и имеют возможность настройки сезонности, тренда.
   b. Ниже приведена таблица сравнения показателей оценки:

Модель / **Показатель оценки**	**Sarimax**	**PROPHET**	**Exponential smoothing**
Средняя абсолютная ошибка (MAE)	66.06013915	98.73289647	103.5647038
Средняя квадратическая ошибка (MSE)	7896.543616	17973.33688	21686.4567
Среднеквадратическая ошибка (RMSE)	88.86249837	134.0646743	147.2632225
Средняя абсолютная ошибка в процентах (MAPE)	1.441353299	1.947700413	2.020402968

   c. Наиболее эффективна показала себя модель **Sarimax** по всем показателям.