2022

Дипломная работа

НА ТЕМУ «АНАЛИЗ СУММЫ ПРОДАЖ АЛКОГОЛЬНОЙ ПРОДУКЦИИ В США»

АВТОР: ЛАЗАРЕНКО А.Б.

Оглавление

Ведение.

Для анализа была выбрана выборка с сумами розничных продаж 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 алкогольной продукции.

Задачи:

1. Провести анализ данных о суммах продаж алкогольной продукции;
2. Построить прогнозы суммы продаж алкогольной продукции, используя различные методы прогнозирования и привести их сравнительную характеристику.

Выбор инструментов для выполнения работы:

1. Выборка с данными по суммам продаж алкогольной продукции в формате csv
   Файл:
   https://github.com/LazarenkoAB/innopolis_2/blob/384dcf11c065dbd07e45b5ee379a400684686013/Retail_Sales_Beer_Liquor_2018-12-01.csv
2. Язык программирования Phyton на базе инструмента Google Colab
   Файл:
   https://github.com/LazarenkoAB/innopolis_2/blob/649a129054073095ca5f49d4a68cc22082590fb7/%D0%94%D0%B8%D0%BF%D0%BB%D0%BE%D0%BC%D0%BD%D0%B0%D1%8F_%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%B0_%D0%9B%D0%B0%D0%B7%D0%B0%D1%80%D0%B5%D0%BD%D0%BA%D0%BE_%D0%90_%D0%91.ipynb

Знакомство с данными

Загрузка данных

1. Загрузка выполнялась с помощью методов pandas, файл расположен на github.com, при запуске не требуется дополнительно его подгружать в Google Colab;
2. Выполнено проверка формата данных – в датасете существует два поля:
    a. **«DATE»**:
        i. При загрузке определился формат object;
        ii. В поле указаны даты 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 int64;
        ii. В поле указано значение суммы продаж в миллионах долларах за месяц, соответствующий полю «DATE».

Предобработка данных

1. Поля переименованы в целях удобства дальнейшего использования:
    a. «MRTSSM4453USN» переименовано в «**rtlsls**» (Retail sales).
    b. «DATE» переименовано в «**date**»
2. Выполнена проверка на наличие пропусков в данных – пропуски отсутствуют
3. Изменен формат данных поля «**date**» из object на datetime64[ns] – для корректного считывания и отображения

Заключение

Выполнена первоначальная обработка данных, в качестве прогнозируемой метрики выбрана сумма розничных продаж.

Возможно переходить к следующему этапу.

EDA (exploratory data analysis) или разведочный анализ данных

Выполнение расчёта основных статистических метрик

1. Индексом анализируемого pandas dataframe решено сделать поле «**date**»;
2. По полю «**rtlsls**» выполнен расчёт основных статистических метрик (таблица 1):

	RTLSLS
COUNT	324.000000
MEAN	2972.895062
STD	1010.218574
MIN	1501.000000
25%	2109.000000
50%	2791.000000
75%	3627.250000
MAX	6370.000000

Таблица 1

3. Построен общий график сумм продаж алкогольной продукции по годам (рис.1)

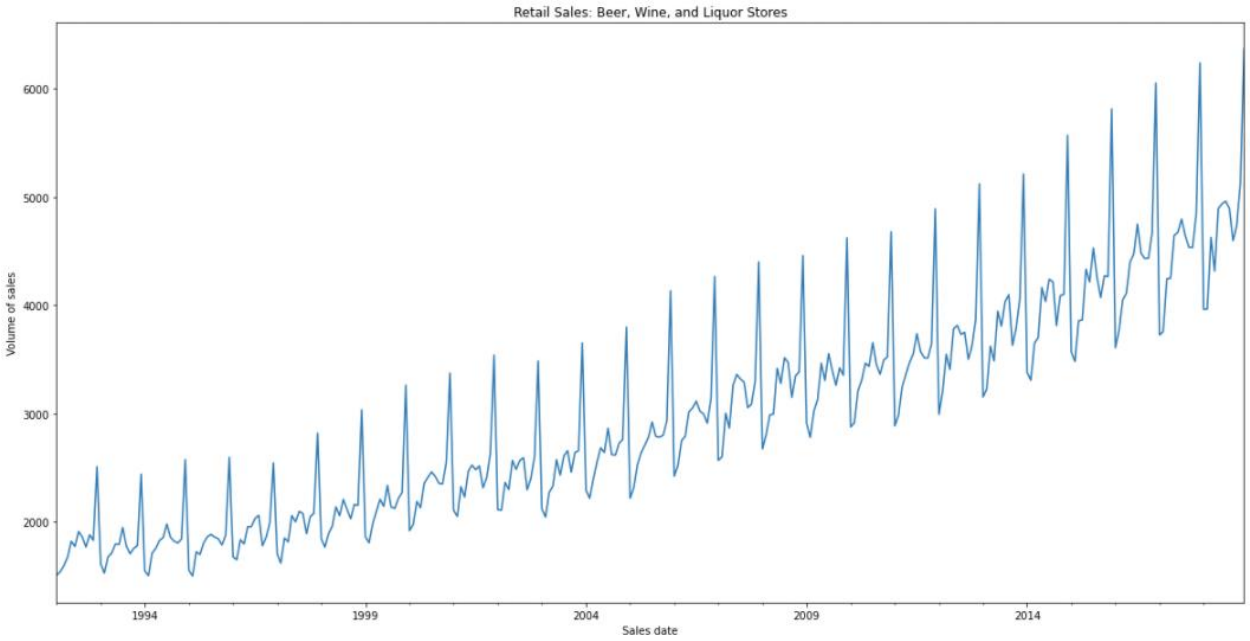

Рисунок 1

Заключение

1. Наблюдается общий восходящий тренд: сумма продаж с каждым годом увеличивается;
2. Наблюдаются сезонные колебания суммы продаж с годовой периодичностью и пиками продаж в конце каждого года;

Выдвинута гипотеза:

Увеличение суммы продаж в будущем с сохранением сезонности.

Построение моделей

Подготовка данных для моделей

1. Сформированы тестовая и обучающая выборки:
    a. Тестовая: 1 год;
    b. Обучающая выборка: остальные 26 лет.
2. Создана структура для будущего сравнительного анализа качества моделей, заполняемая в ходе построения моделей.
3. Выполнена декомпозиция временного ряда с использованием аддитивной модели (рис.3)

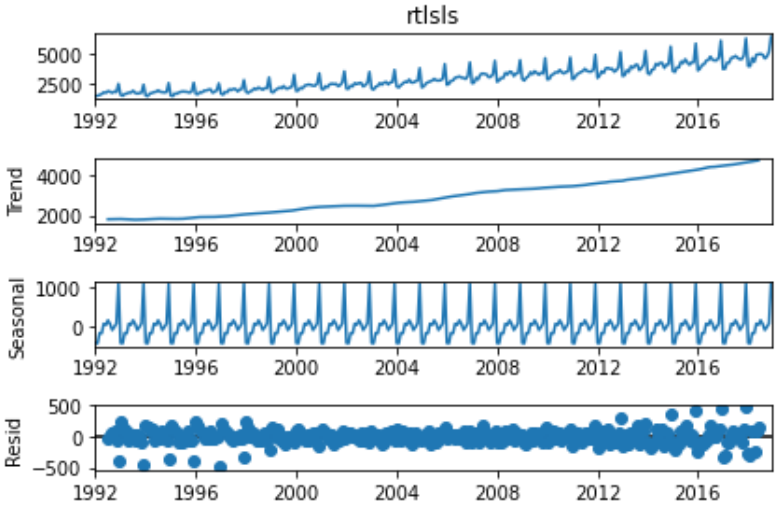

Рисунок 3

    a. Наблюдается положительный (восходящий) тренд;
    b. Наблюдается годовая сезонность.

Модель 1. Sarimax

Построение модели

1. Выполнен автоматический подбор параметров модели с входными настройками подбора на всем датасете с включением сезонности перидом в 1 год.
   В результате определена модель: **SARIMAX(4, 1, 3)x(2, 1, [1], 12)**;
2. Модель обучена на обучающей выборке и построен прогноз на период, соответствующий тестовой выборке.
3. Построены графики для визуального сравнения прогнозных данных с тестовой выборкой (рис.4)

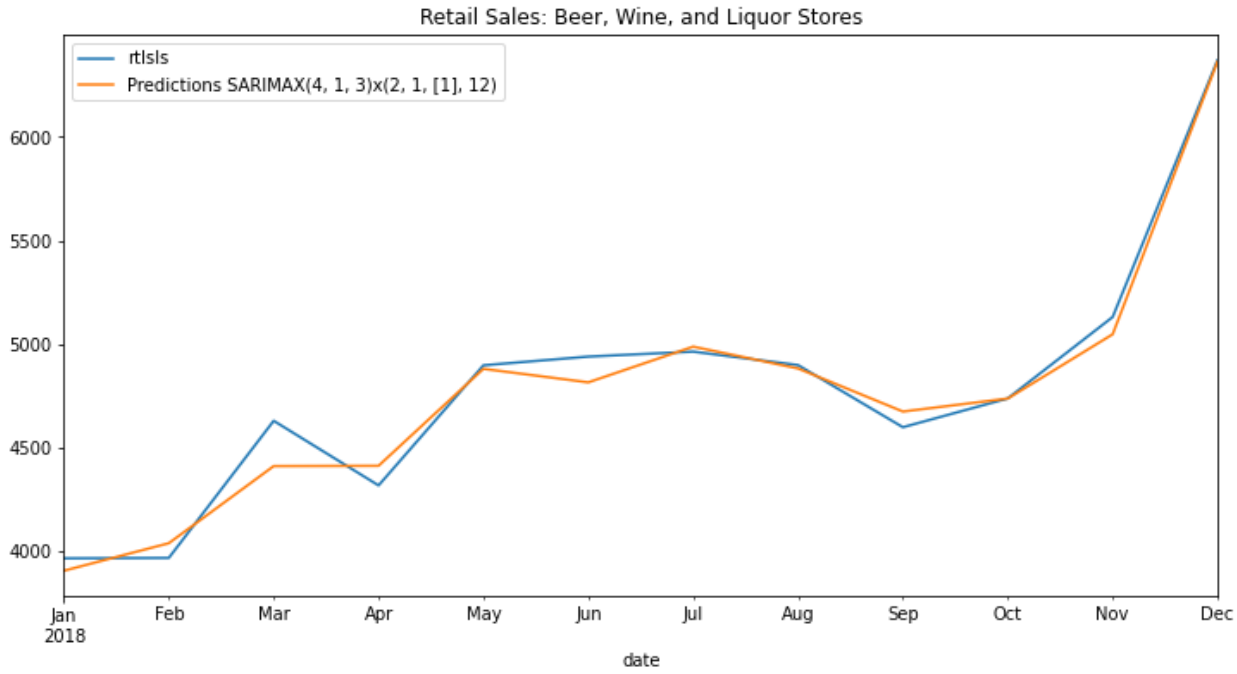

Рисунок 4

4. Рассчитаны значения критериев оценки качества модели:
   a. **MAE**: 66.06013915
   b. **MSE**: 7896.543616
   c. **RMSE**: 88.86249837
   d. **MAPE**: 1.441353299
5. Указанные выше значения добавлены в структуру сравнительного анализа качества моделей.

6. Построен и визуализирован прогноз на год вперед (рис.5)

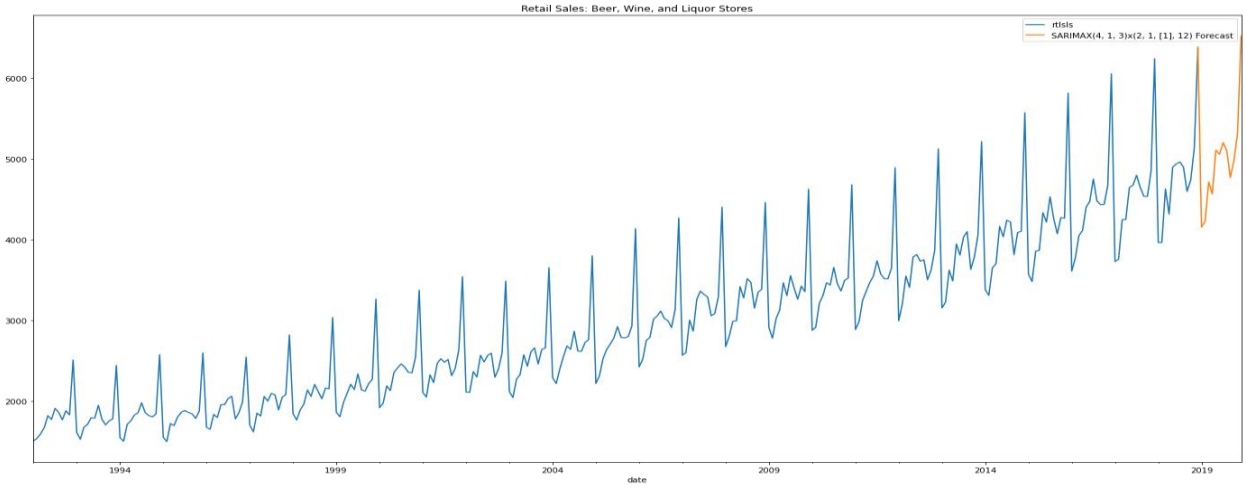

Рисунок 5

Выводы по работе модели

Модель показала себя хорошо:

- **RMSE**=88.86 - это очень хороший показатель.
- **MAPE**=1.44% - это хороший результат.

Согласно графику, на будущее видим, что тренд и высота амплитуда были отображены корректно, общая динамика прослеживается.

Модель 2. Prophet

Построение модели

1. Подготовлены данные для построения модели;
2. Выполнен автоматический подбор параметров модели с входными настройками мультипликативной сезонности.
   В результате алгоритм проигнорировал недельную и дневную сезонность, но обнаружил годовую сезонность и использовал её при настройке модели;
3. Модель обучена на обучающей выборке и построен прогноз на период, соответствующий тестовой выборке.
4. Построены графики для визуального сравнения прогнозных данных с тестовой выборкой (рис.6).

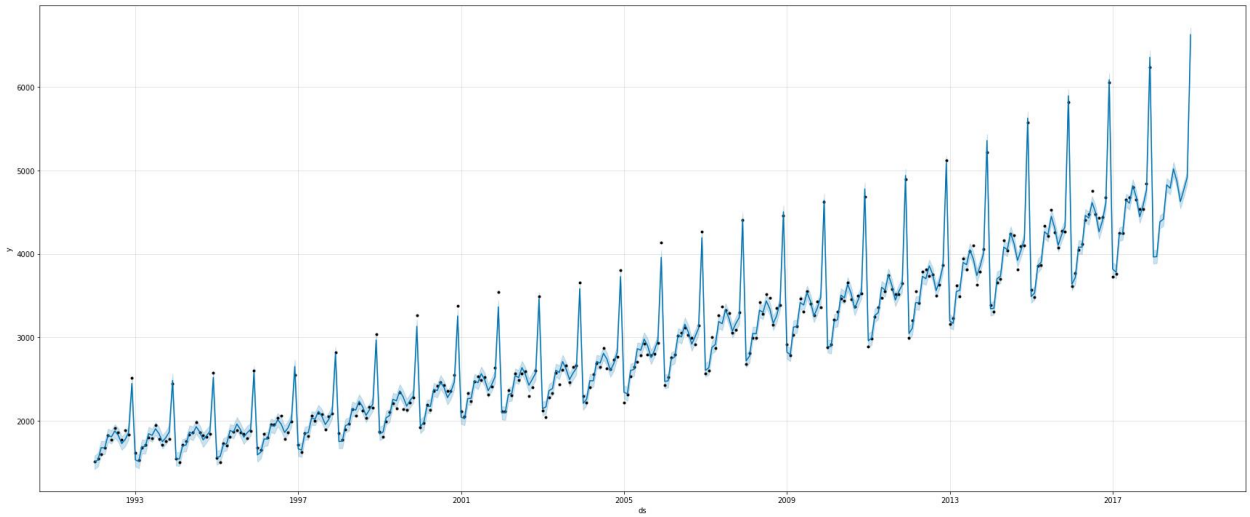

Рисунок 6

5. Временной ряд разложен на основные компоненты – тренд и сезонность (рис.7).

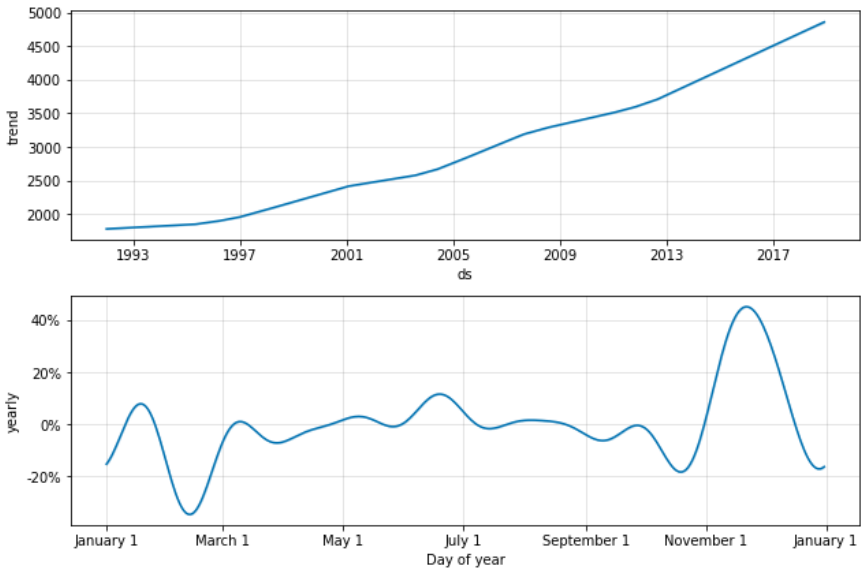

Рисунок 7

Наблюдается возрастающий тренд продаж и годовая сезонность.

6. Рассчитаны значения критериев оценки качества модели:
   a. **MAE**: 98.73289647
   b. **MSE**: 17973.33688

c. **RMSE**: 134.0646743
d. **MAPE**: 1.947700413

7. Указанные выше значения добавлены в структуру сравнительного анализа качества моделей.
8. Построен и визуализирован прогноз на год вперед (рис.8).

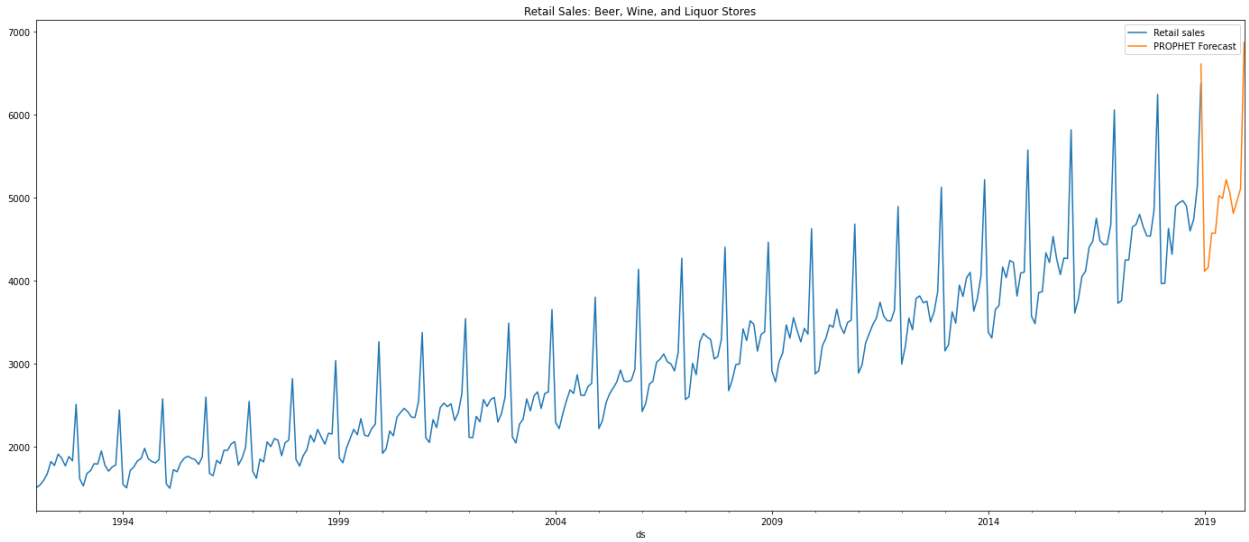

Рисунок 8

Выводы по работе модели

Модель показала себя хорошо:

- **RMSE**=134.06 - хороший показатель.
- **MAPE**=1.95% - хороший результат.

Согласно графику, на будущее видим, что тренд и высота амплитуда были отображены корректно, общая динамика прослеживается.

Модель 3. Exponential Smoothing

Построение модели

1. Рассмотрено 4 модели Хольта-Винтерса (т.к. они позволяют учесть тренд и сезонность) со следующими настройками:
    a. **Holt-Winters (add-add-seasonal):**
        i. Период сезонности = 12 месяцев,
        ii. Тренд - аддитивный,
        iii. Сезонность - аддитивная,
        iv. Использование преобразование Боса-Кокса
    b. **Holt-Winters (add-mul-seasonal) RMSE:**
        i. Период сезонности = 12 месяцев,
        ii. Тренд - аддитивный,
        iii. Сезонность - мультипликативная,
        iv. Использование преобразование Боса-Кокса
    c. **Holt-Winters (mul-add-seasonal) RMSE:**
        i. Период сезонности = 12 месяцев,
        ii. Тренд - мультипликативный,
        iii. Сезонность - аддитивная,
        iv. Использование преобразование Боса-Кокса
    d. **Holt-Winters (mul-mul-seasonal) RMSE:**

i. Период сезонности = 12 месяцев,
ii. Тренд - мультипликативный,
iii. Сезонность - мультипликативная,
iv. Использование преобразование Боса-Кокса

2. Каждая из моделей обучена на обучающей выборке и для каждой построен прогноз на период, соответствующий тестовой выборке.
3. Построены графики для визуального сравнения прогнозных данных с тестовой выборкой (рис.9).

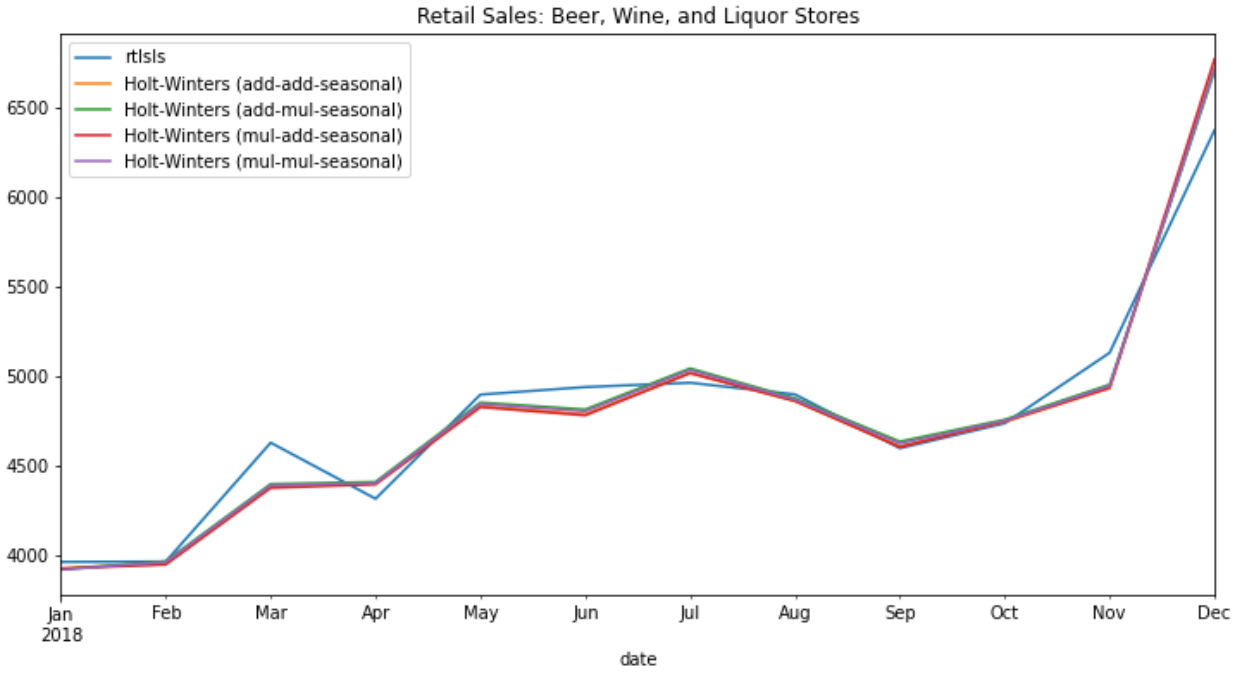

Рисунок 9

4. Рассчитаны значения критериев оценки качества модели:
   a. **Holt-Winters (add-add-seasonal)**:
      i. **MAE**: 103.56
      ii. **MSE**: 21686.45
      iii. **RMSE**: 147.26
      iv. **MAPE**: 2.02
   b. **Holt-Winters (add-mul-seasonal)**:
      i. **MAE**: 100.31
      ii. **MSE**: 19156.13
      iii. **RMSE**: 138.41
      iv. **MAPE**: 1.97
   c. **Holt-Winters (mul-add-seasonal)**:
      i. **MAE**: 109.37
      ii. **MSE**: 25021.60
      iii. **RMSE**: 158.18
      iv. **MAPE**: 2.13
   d. **Holt-Winters (mul-mul-seasonal)**:
      i. **MAE**: 101.94
      ii. **MSE**: 20312.10
      iii. **RMSE**: 142.52
      iv. **MAPE**: 2.00

5. Указанные выше значения добавлены в структуру сравнительного анализа качества моделей.
6. Построены и визуализированы прогнозы на год вперед (рис.10).

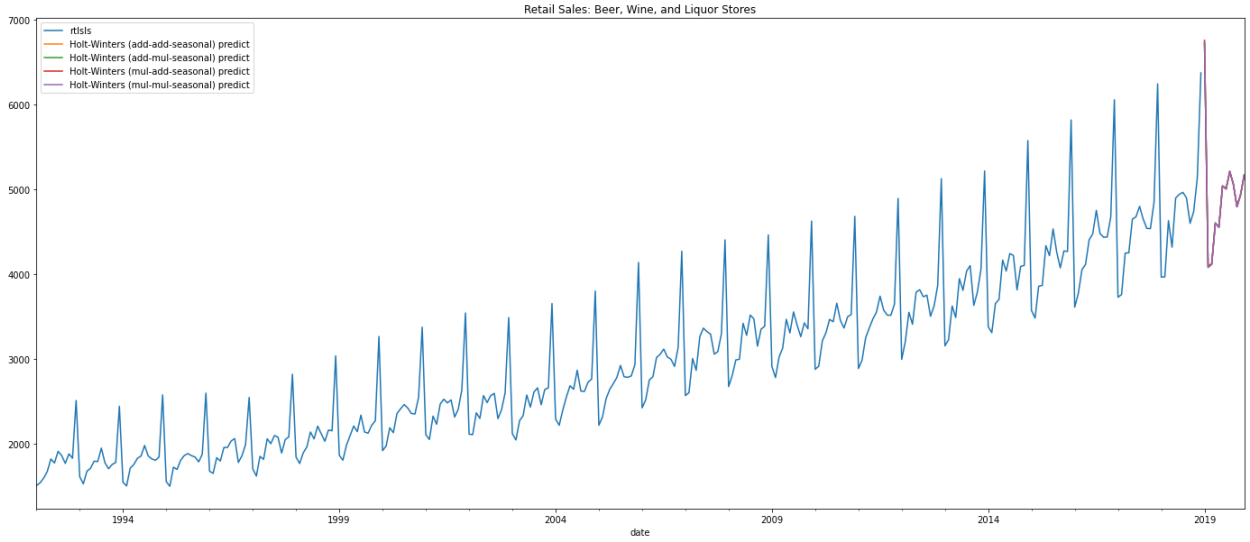

Рисунок 10

Выводы по работе модели

Все 4 модели экспоненциального сглаживания показали себя неплохо:

1. Хорошие показатели **RMSE**:
    a. **Holt-Winters (add-add-seasonal) RMSE**: 147.26
    b. **Holt-Winters (add-mul-seasonal) RMSE:** 138.41
    c. **Holt-Winters (mul-add-seasonal) RMSE**: 158.18
    d. **Holt-Winters (mul-mul-seasonal) RMSE**: 142.52
2. Не высокие проценты рассчитанной ошибки **MAPE**:
    a. **Holt-Winters (add-add-seasonal) MAPE**: 2.02
    b. **Holt-Winters (add-mul-seasonal) MAPE**: 1.97
    c. **Holt-Winters (mul-add-seasonal) MAPE**: 2.13
    d. **Holt-Winters (mul-mul-seasonal) MAPE**: 2.00

Согласно графикам на будущее видим, что тренд и высота амплитуда были отображены корректно, общая динамика прослеживается.

Сравнение качества моделей

1. Построены данные для сравнения качества построенных моделей (табл.2)

Модель	*MAE*	*MSE*	*RMSE*	*MAPE*
SARIMAX(4, 1, 3)x(2, 1, [1], 12)	66.060139	7896.543616	88.862498	1.441353
PROPHET	98.732896	17973.336882	134.064674	1.947700
Holt-Winters (add-add-seasonal)	103.564704	21686.456703	147.263223	2.020403
Holt-Winters (add-mul-seasonal)	100.308709	19156.133642	138.405685	1.968399
Holt-Winters (mul-add-seasonal)	109.370491	25021.595428	158.182159	2.125332
Holt-Winters (mul-mul-seasonal)	101.944204	20312.101779	142.520531	1.995114

Таблица 2

2. На основании указанных выше данных сделан вывод, что модель SARIMAX является наиболее качественной, т.к. дает наилучшие показатели по каждому из оценочных критериев.

Выводы

- Проведен анализ данных с использованием различных методов обработки статистической информации.
- Рассчитаны основные статистические метрики, позволяющие судить о характере исследуемого явления.
- Прогнозные модели позволили выявить тенденцию роста суммы розничных продаж по сравнению с предыдущим годом, а также сохранение характера амплитудных колебаний в разрезе каждого года с пиками продаж в период новогодних праздников.
- Сравнительный анализ значений критериев качества построенных моделей показал, что наиболее качественной из построенных является модель SARIMAX.