Дипломная работа

на тему:

Анализ суммы продаж алкогольной продукции в США

Автор: Полусмак Вячеслав Иванович

Руководитель: Шестакова Екатерина Андреевна

2022 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

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.

Целью дипломного проекта является проведение исследования данных и построение прогноза суммы продаж алкогольной продукции.

Для достижения поставленной цели необходимо решить следующие задачи:

– провести анализ данных о суммах продаж алкогольной продукции;

– построить прогнозы суммы продаж алкогольной продукции, используя различные методы прогнозирования и привести их сравнительную характеристику.

Для выполнения работы были выбраны и использованы следующие инструменты:

– выборка с данными по суммам продаж алкогольной продукции в формате csv, ссылка на файл:
https://github.com/poluslavik13/innopolis/blob/main/Retail_Sales_Beer_Liquor_2018-12-01.csv

– язык программирования Phyton на базе инструмента Google Colab, ссылка на файл:
https://github.com/poluslavik13/innopolis/blob/main/%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%9F%D0%BE%D0%BB%D1%83%D1%81%D0%BC%D0%B0%D0%BA_%D0%92_%D0%98_.ipynb

1 Знакомство с данными

1.1 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».

1.2 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

1.3 Заключение

Выполнена первоначальная обработка данных, в качестве прогнозируемой метрики выбрана сумма розничных продаж. Возможно переходить к следующему этапу.

2 EDA (exploratory data analysis) или разведочный анализ данных

2.1 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).

Таблица 1 – Расчет основных статистических метрик

	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

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

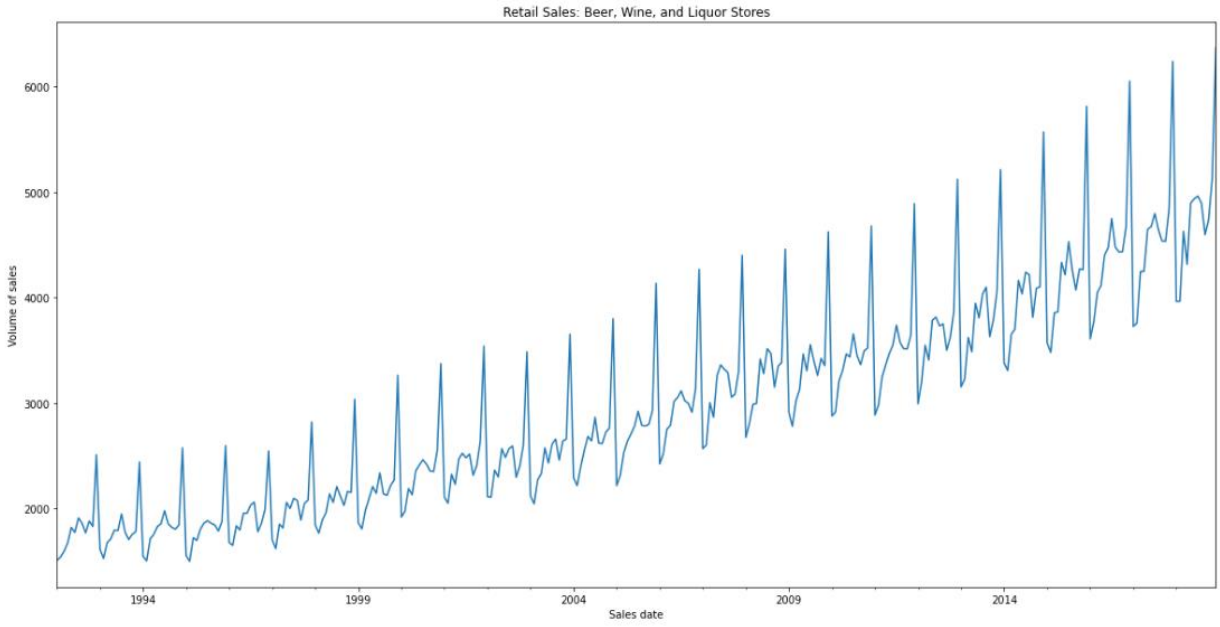

Рисунок 1 – Общий график сумм продаж

2.3 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.

3 Построение моделей

3.1 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. 2.

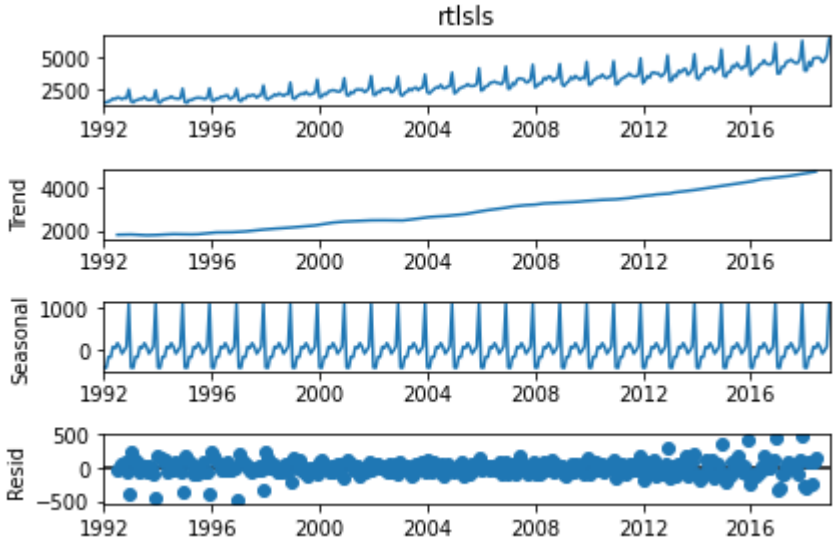

Рисунок 2 – Модели

a. Наблюдается положительный (восходящий) тренд;

b. Наблюдается годовая сезонность.

4 Модель 1. Sarimax

4.1 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.3.

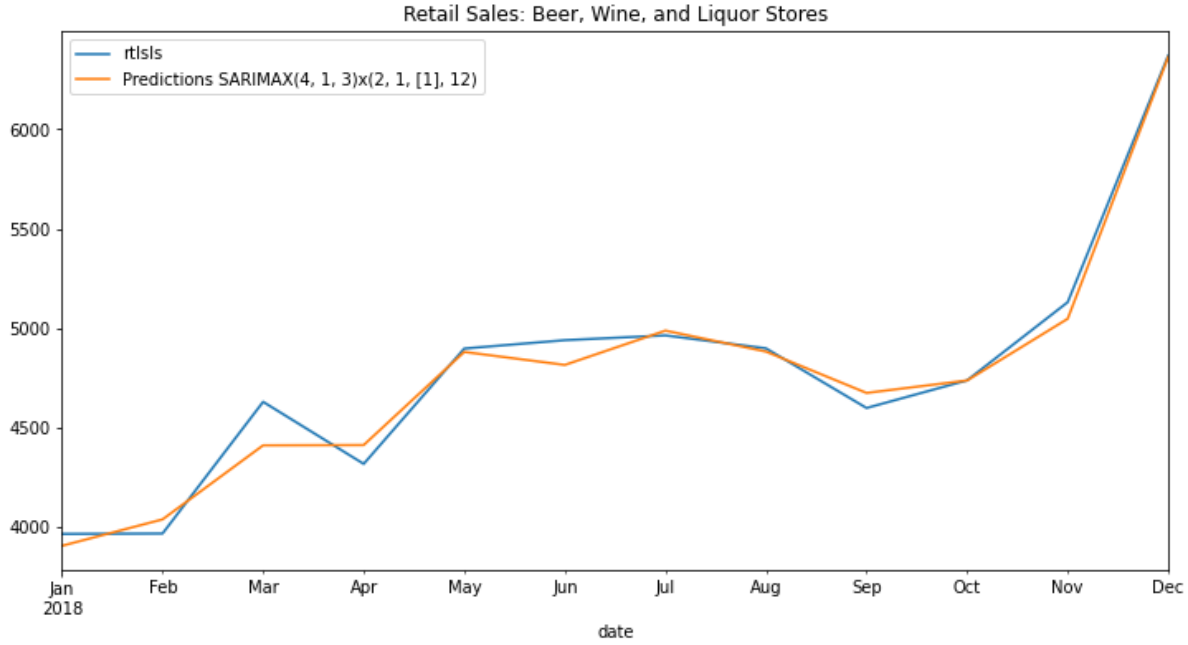

Рисунок 3 – Графики визуального сравнения прогнозных данных

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.4.

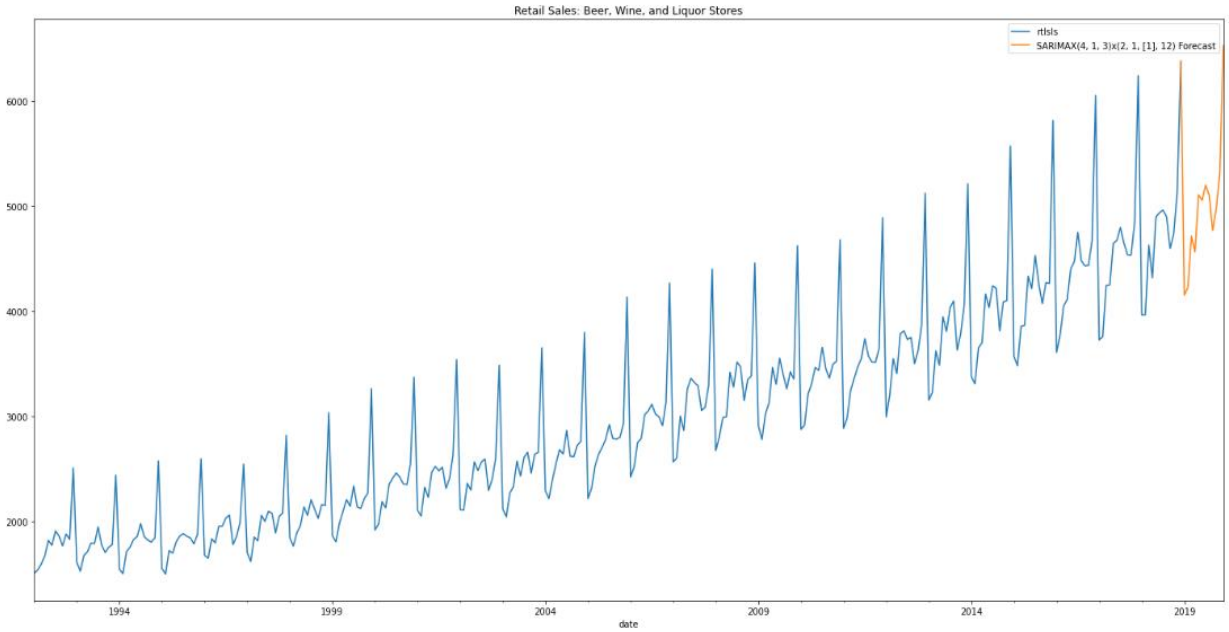

Рисунок 4 – Визуализация прогноза на год вперед

4.2 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.

5 Модель 2. Prophet

5.1 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.5.

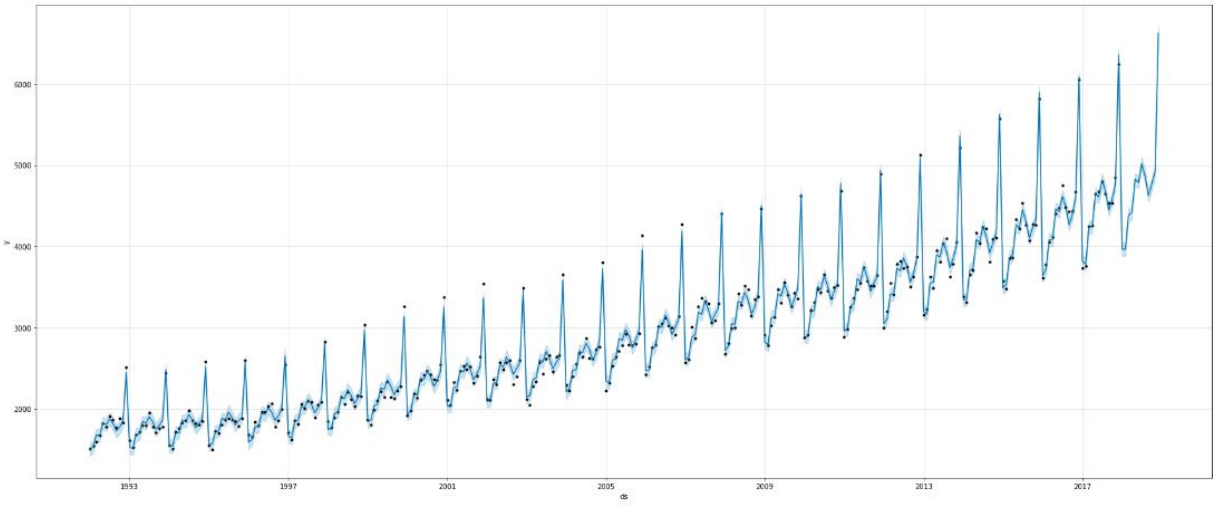

Рисунок 5 – Графики для визуального сравнения прогнозных данных

5. Временной ряд разложен на основные компоненты – тренд и сезонность (рис.6)

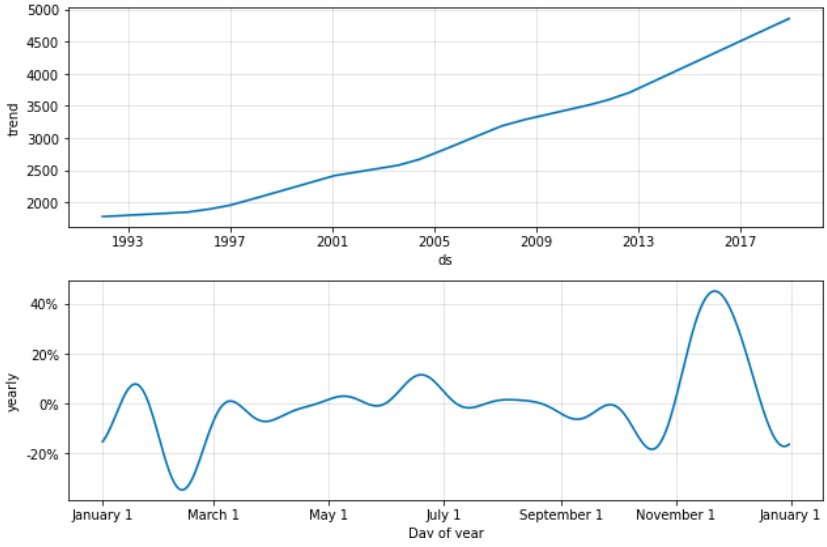

Рисунок 6 – Разложение временного ряда на компоненты

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.7).

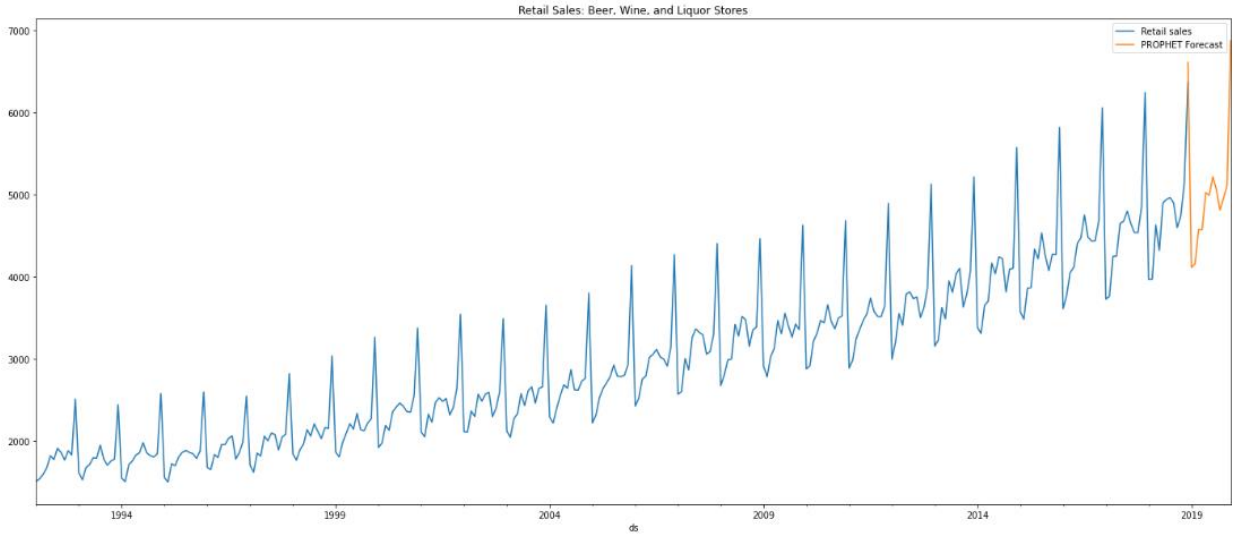

Рисунок 7 – График прогноза на год вперед

5.2 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.

6 Модель 3. Exponential Smoothing

6.1 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.8.

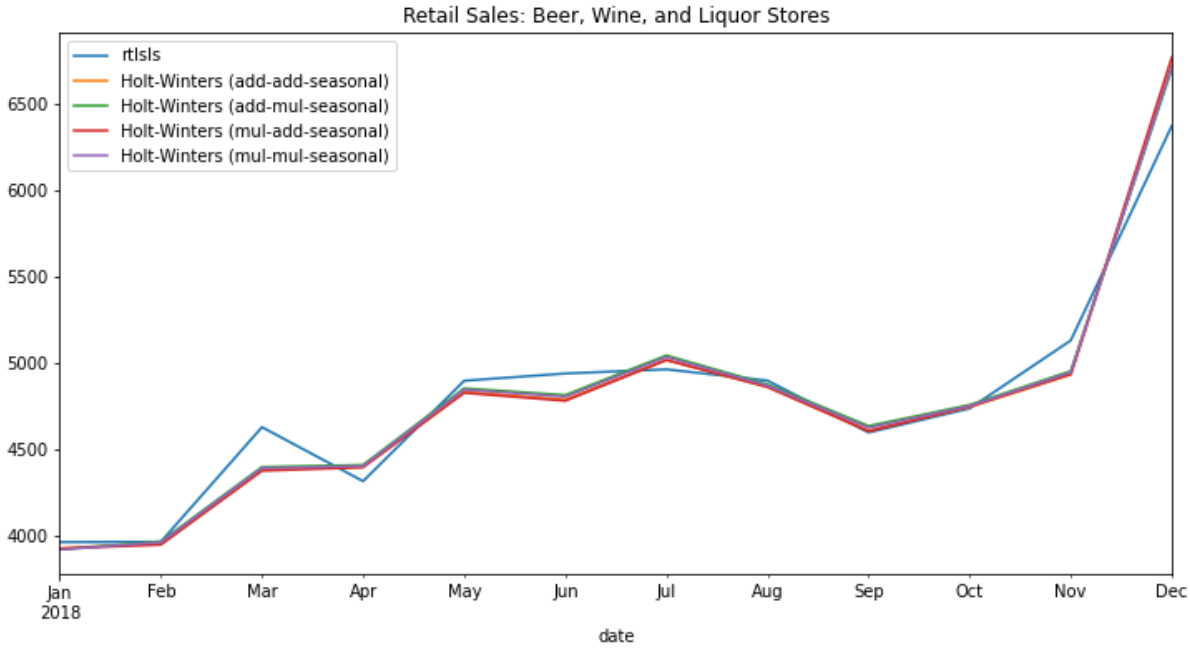

Рисунок 8 – Графики визуального сравнения прогнозных данных

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.9.

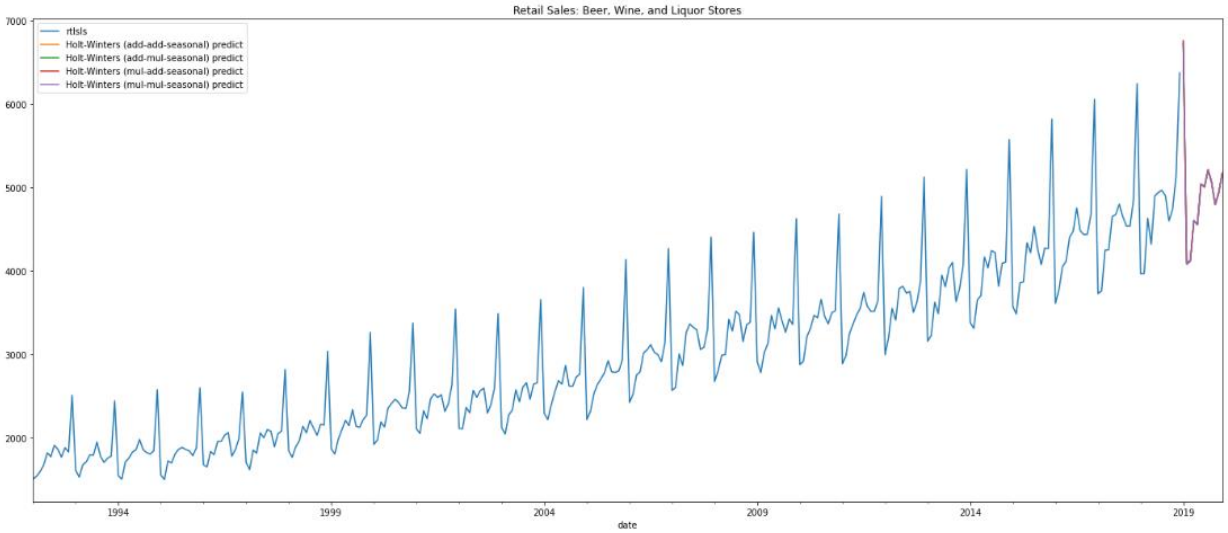

Рисунок 9 – Визуализация прогноза на год вперед

6.2 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.

7 Сравнение качества моделей

1. Построены данные для сравнения качества построенных моделей, таблица 2.

Таблица 2 – Сравнение качества моделей

model	mae_error	mse_error	rmse_error	mape_error
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

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

1. Проведен анализ данных с использованием различных методов обработки статистической информации.

2. Рассчитаны основные статистические метрики, позволяющие судить о характере исследуемого явления.

3. 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.

4. Сравнительный анализ значений критериев качества построенных моделей показал, что наиболее качественной из построенных является модель SARIMAX.