Дипломная работа

на тему:

Ежемесячное производство молока

Разработал: Кузьмин Антон Леонидович

Руководитель: Шестакова Екатерина Андреевна

2022 г.

СОДЕРЖАНИЕ

Введение

В качестве исследования для дипломной работы был выбран датасет с показателями ежемесячного производства молока с 1962 по 1975г.г. Целью дипломного проекта является проведение исследования данных и построение прогноза дальнейшего увеличения производства молока.

Для достижения поставленной цели необходимо решить следующие задачи:

– провести анализ данных о ежемесячном объеме производства продукции;

– построить прогнозы производства молока, используя различные методы прогнозирования и привести их сравнительную характеристику.

Для выполнения работы были выбраны и использованы следующие инструменты:

– датасет, размещенный по следующей ссылке: https://raw.githubusercontent.com/AnToxa0887/innopolis/main/monthly_milk_production.csv

– программа, находящаяся в свободном доступе Google Colab. Ссылка на дипломную работу, выполненную в данной программе: https://github.com/AnToxa0887/innopolis_2/blob/main/Diplom_Kuzmin_Anton_.ipynb

1 Загрузка данных

Для выполнения поставленной задачи загрузили библиотеки обработки данных, а также функции, модели и метрики. Перечень указан в листинге программы. Далее была загружена сама таблица данных.

2 Знакомство с данными

При вызове таблицы на экране обращаем внимание, что датасет состоит из двух столбцов. В первом перечислены даты учета молока (ежемесячно), во втором – объем произведенного продукта за текущий месяц. Количество строк 168, что соответствует количеству месяцев в промежутке с 1962 по 1975года строк.

3 Предобработка данных

При дальнейшей работе с таблицей наличие пустых строк не было обнаружено. Проверили типы данных и посмотрели общую информацию о датасете. В качестве прогнозируемой метрики был выбран показатель объема производства молока.

4 EDA (exploratory data analysis) или разведочный анализ данных

В данном разделе были поставлены следующие задачи:

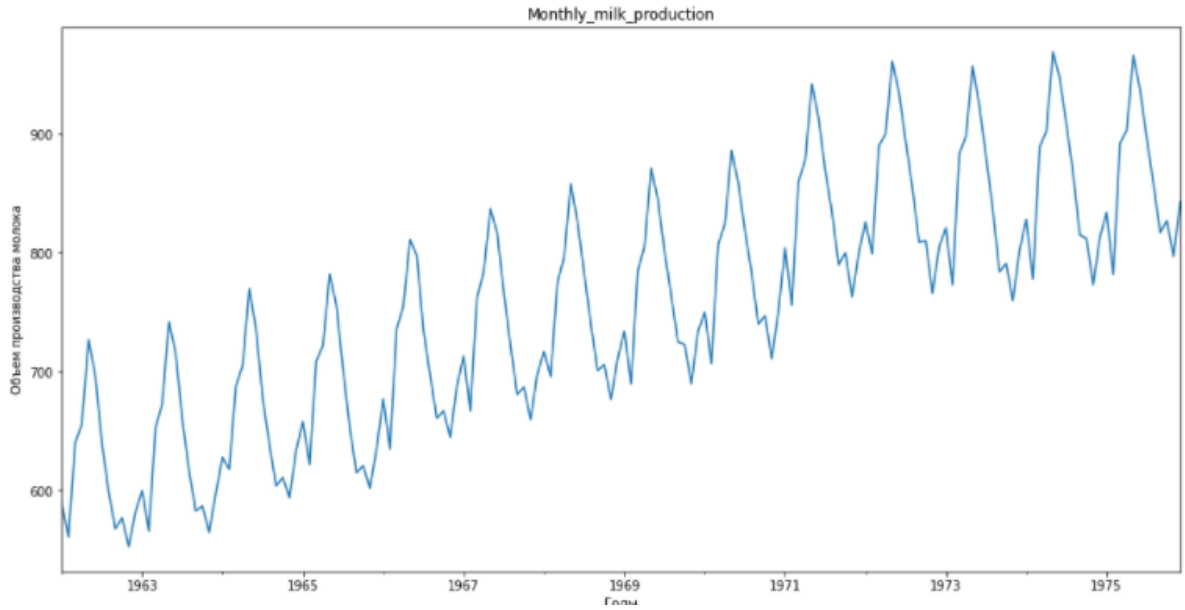
- Сделать столбец с датами индексом;
- Вывести статистику по нужным столбцам;
- Построить графическое отображение столбцов;
- Выявить связи между признаками.

Индексом был выбран столбец с указанием месяца. По показателю «production» был выполнен расчет основных статистических метрик

	production
count	168.000000
mean	754.708333
std	102.204524
min	553.000000
25%	677.750000

	production
50%	761.000000
75%	824.500000
max	969.00000

Также был построен общий график для метрик



Сделали следующие выводы:

1. Наблюдается общий восходящий тренд: объем производства молока с каждым годом увеличивается;

2. Наблюдаются сезонные колебания объема продукции с годовой периодичностью и пиками в середине года;

Была выдвинута гипотеза: производство объемов молока в последующие годы будет также увеличиваться с сохранением сезонности.

5 Построение моделей, анализ результатов

Для выполнения поставленной задачи необходимо спрогнозировать поведение моделей. За основу был взят следующий алгоритм:

1. Описать модель;

2. Подобрать оптимальные параметры;

3. Создать модель;

4. Обучить модель;

5. Сделать прогноз на период тестовой выборки;

6. Сравнить прогноз с тестовой выборкой (построить график);

7. Оценить качество прогноза;

8. Сделать прогноз на год;

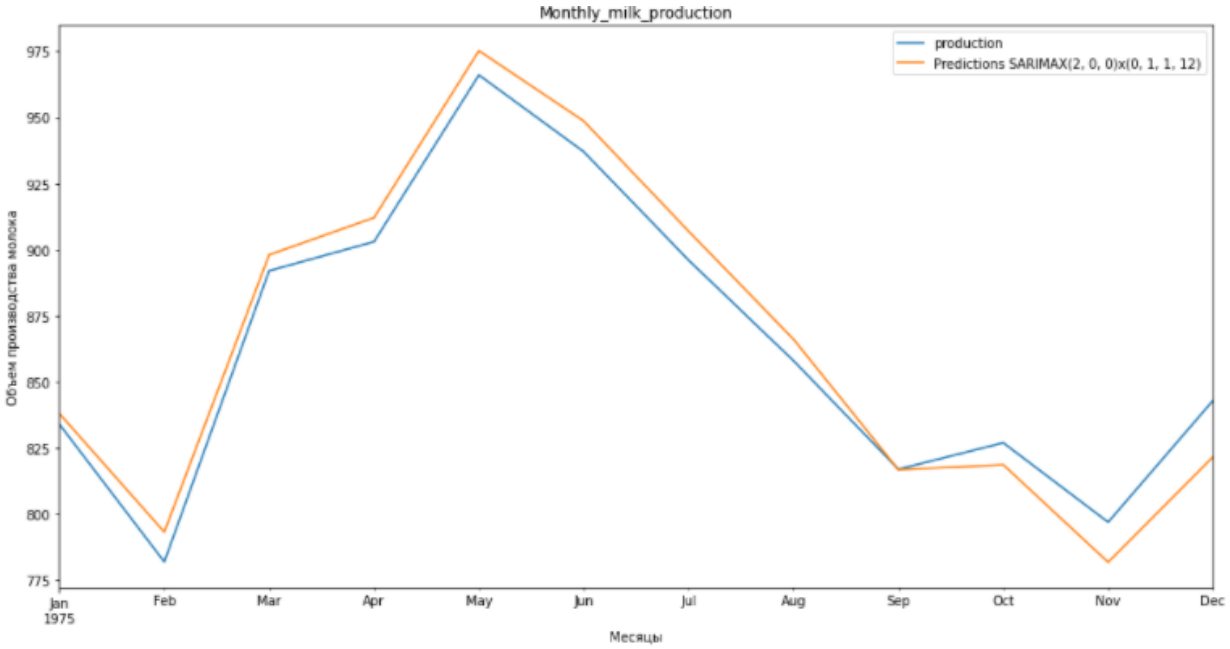
9. Сделать выводы о работе данного метода прогнозирования.

5.1 Модель Sarimax

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(2, 0, 0)x(0, 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

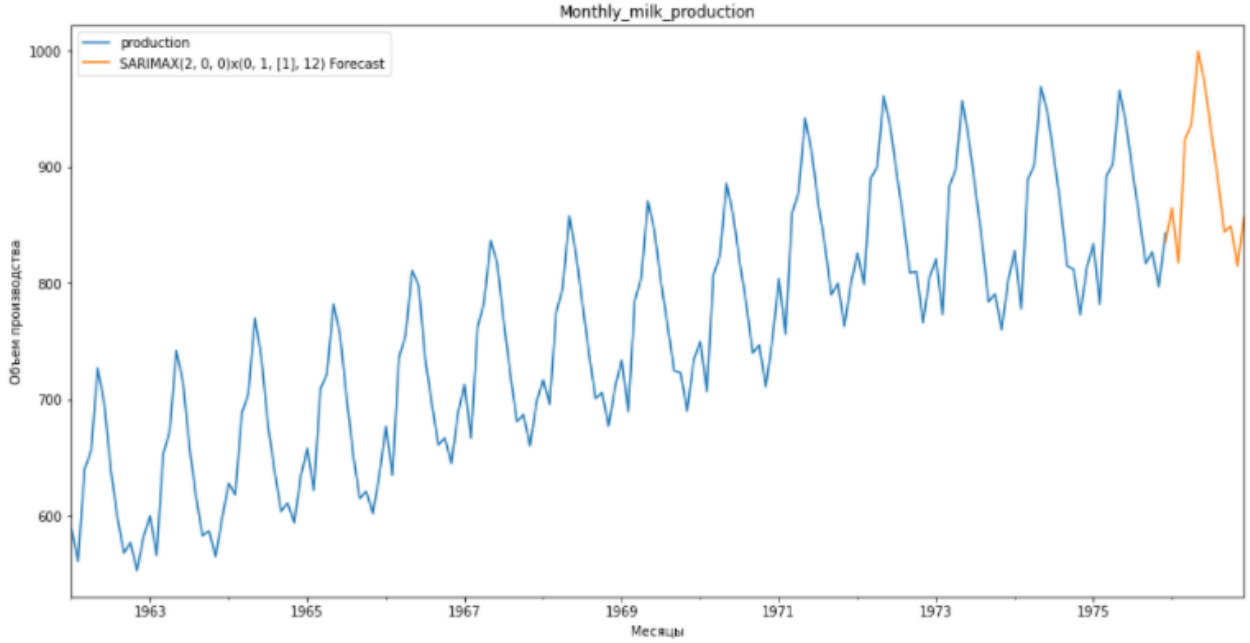


4. Рассчитаны значения критериев оценки качества модели:
   a. MAE: 9.60195423
   b. MSE: 118.256919
   c. RMSE: 10.87459972

d. MAPE: 1.118331538

5. Указанные выше значения добавлены в структуру сравнительного анализа качества моделей.

6. Построен и визуализирован прогноз на год вперед.



Выводы по работе модели

Модель показала себя хорошо:

– RMSE=10.87 - это очень хороший показатель.

– MAPE=1.12% - это хороший результат.

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

5.2 Модель Prophet

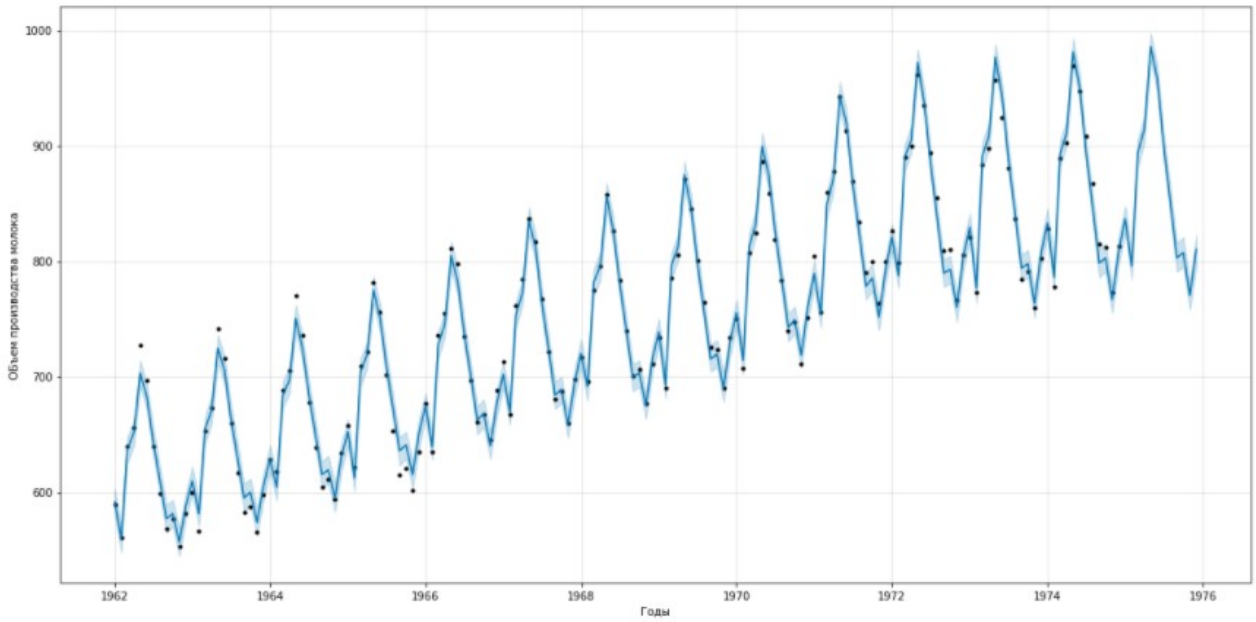
5.1 Построение модели

1. Подготовлены данные для построения модели;

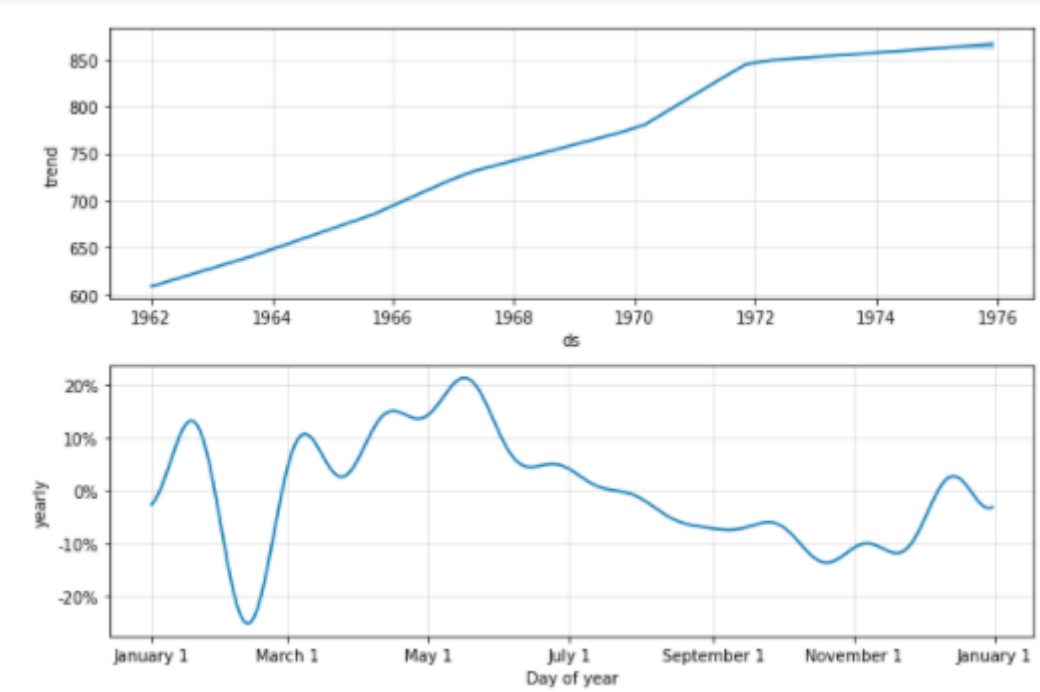
2. Выполнен автоматический подбор параметров модели с входными настройками мультипликативной сезонности. В результате алгоритм проигнорировал недельную и дневную сезонность, но обнаружил годовую сезонность и использовал её при настройке модели;

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.



5. Временной ряд разложен на основные компоненты – тренд и сезонность



Наблюдается возрастающий тренд продаж и годовая сезонность.

6. Рассчитаны значения критериев оценки качества модели:

   a. MAE: 14.37304065

   b. MSE: 297.4948351

   c. RMSE: 17.24803859

   d. MAPE: 1.682529777

7. Указанные выше значения добавлены в структуру сравнительного анализа качества моделей.

8. Построен и визуализирован прогноз на год вперед

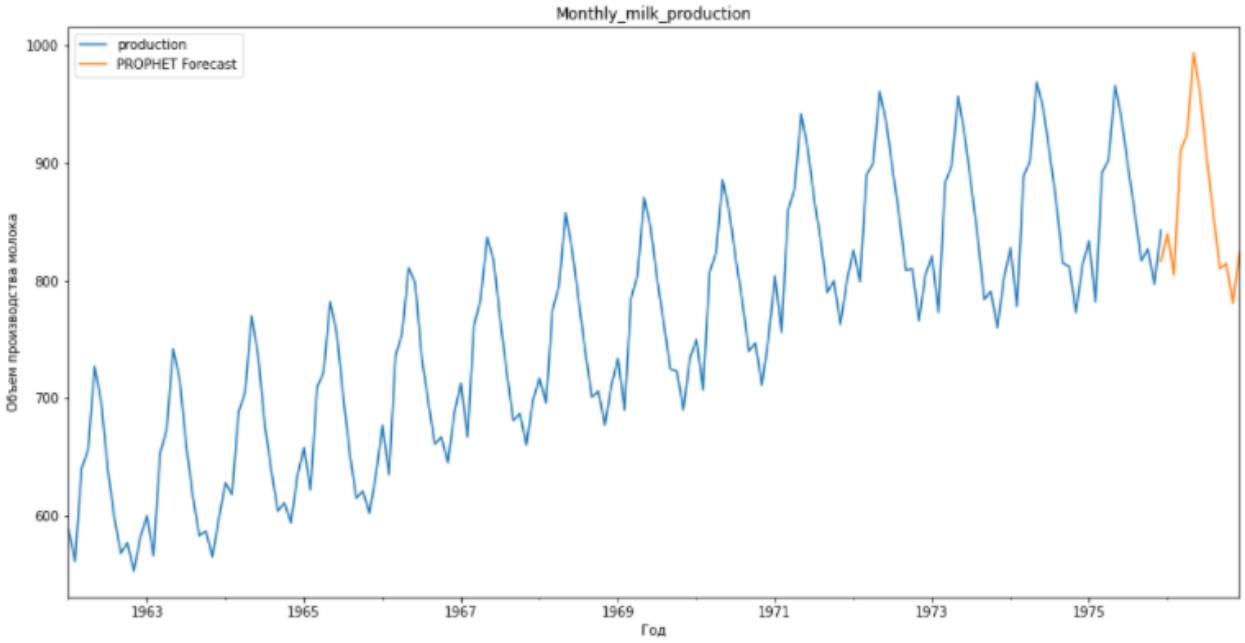


Рисунок 7 – График прогноза на год вперед

Выводы по работе модели

Модель показала себя хорошо:

– RMSE=17.25 - хороший показатель.

– MAPE=1.68 % - хороший результат.

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

5.3 Модель Exponential Smoothing

Метод также известен как метод простого экспоненциального сглаживания, или метод Брауна

1. Рассмотрена модель Holt-Winters

2. Построен график для визуального сравнения прогнозных данных с тестовой выборкой

3. Рассчитаны значения критериев оценки качества модели:

MAE: 13.76465315

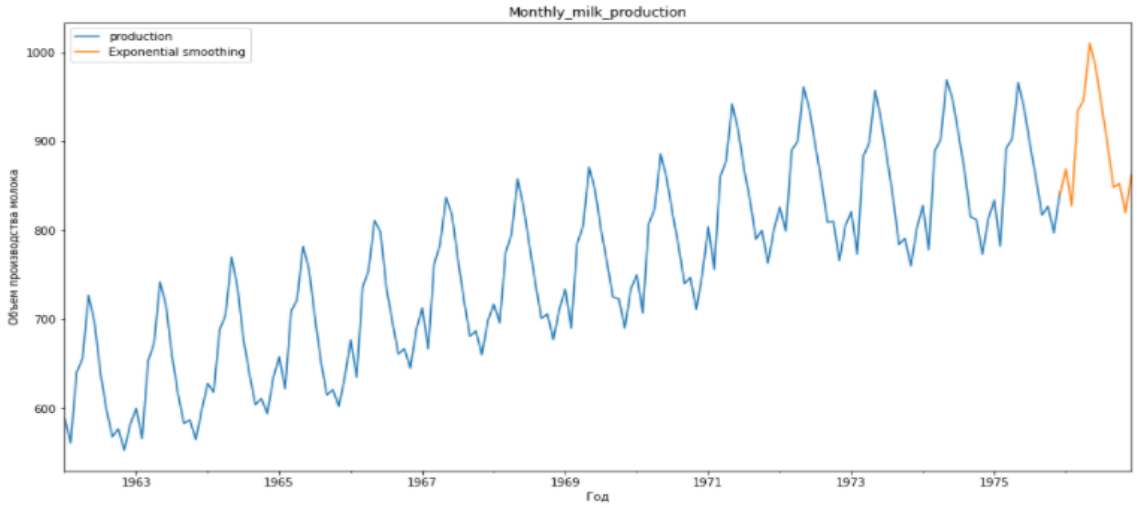
MSE: 245.1180418

RMSE: 15.6562461

MAPE: 1.566134973

4. Указанные выше значения добавлены в структуру сравнительного анализа качества моделей.

5. Построены и визуализированы прогнозы на год вперед,



Выводы по работе модели

Модель показала себя хорошо:

– RMSE=15.66 - хороший показатель.

– MAPE=1.57 % - хороший результат.

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

6. Сравнение качества моделей

Построены данные для сравнения качества построенных моделей, таблица

	model	mae_error	mse_error	rmse_error	mape_error
0	SARIMAX(2, 0, 0)x(0, 1, [1], 12)	9.601954	118.256919	10.874600	1.118332
1	PROPHET	14.373041	297.494835	17.248039	1.682530
2	prediction_exps	13.764653	245.118042	15.656246	1.566135

- MAE - средняя абсолютная ошибка
- MSE - средняя квадратичная ошибка
- RMSE - корень из средней квадратичной ошибки
- MAPE - средняя абсолютная процентная ошибка

Исходя из показателей rmse и mape делаем вывод, что модель SARIMAX(2, 0, 0)x(0, 1, 1, 12) являетя наиболее качественной, т.к. выдаёт наименьшие ошибки по каждому из критериев.

ИТОГ

1. Проведен анализ данных с использованием различных методов обработки стат истической информации (рассмотрели три варианта)

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

3. Изначальный прогноз оправдался. Мы заметили, что в каждом из методов оцен ки ожидался дальнейший рост производства молока в последующий год.

4. Исходя из значений расчитанных метрик пришли к выводу, что наиболее качес твенной из построенных является модель SARIMAX.