

## Выпускная квалификационная работа

# «Прогнозирование стоимости фьючерса на золото с помощью методов машинного обучения»

Программа профессиональной переподготовки  
«Анализ данных на языке Python»

Выполнила: Паникарова Т.С.  
Руководитель: к.э.н., доц. Заграновская А. В.

Санкт - Петербург  
2024

# Объект и предмет исследования

## **Объект исследования:**

Данные стоимости золота по дням

## **Предмет исследования:**

Методы прогнозирования

## **Цель работы:**

Анализ стоимости золота, разработка модели прогнозирования стоимости для оптимизации пользовательского портфеля. Построение и выбор оптимальной модели.

- первичный анализ данных, выявление аномалий, влияние факторов на стоимость золота
- прогноз стоимости с использованием моделей: модель взвешенная скользящая средняя, модель Брауна, модель Хольта, модель на основе кривых роста, ARIMA
- оценка полученных прогнозных значений
- выбор подходящей модели

# Исходные данные

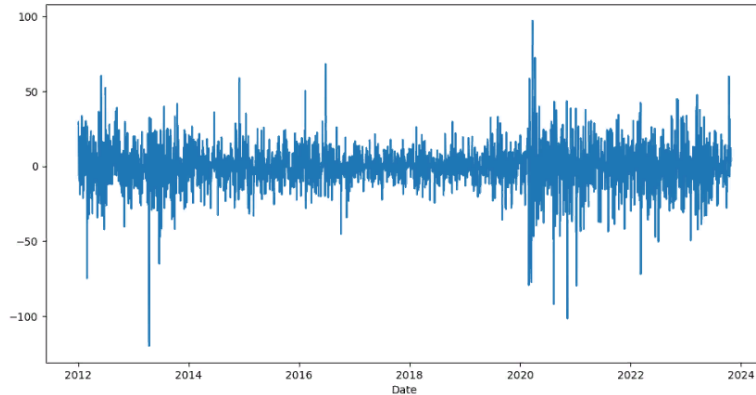
Для анализа были использованы данные с сайта investing.com за 2012-2023 годы

	Date	Price	Open	High	Low
0	10/27/2023	1998.5	1995.0	2019.7	1986.4
1	10/26/2023	1997.4	1991.2	2003.7	1981.6
2	10/25/2023	1994.9	1982.7	1998.6	1973.6
3	10/24/2023	1986.1	1984.1	1992.0	1964.6
4	10/23/2023	1987.8	1987.7	1994.3	1971.0

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 3060 entries, 0 to 3059  
Data columns (total 5 columns):  
#   Column  Non-Null Count  Dtype  
---  ---      -  
0   Date    3060 non-null   object  
1   Price   3060 non-null   float64  
2   Open    3060 non-null   float64  
3   High    3060 non-null   float64  
4   Low     3060 non-null   float64  
dtypes: float64(4), object(1)
```

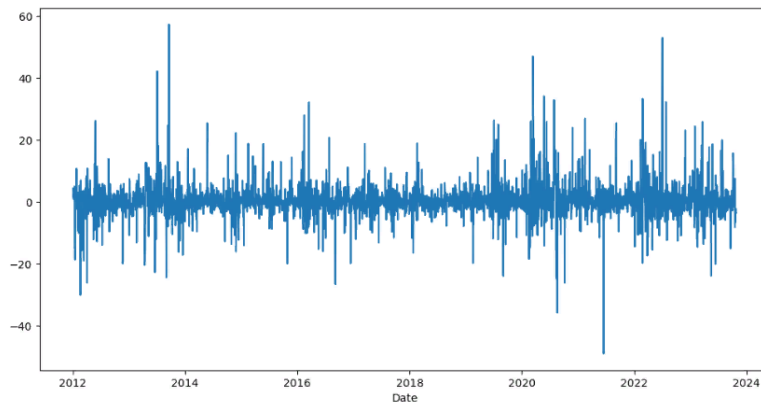


Разница между ценой закрытия и открытия



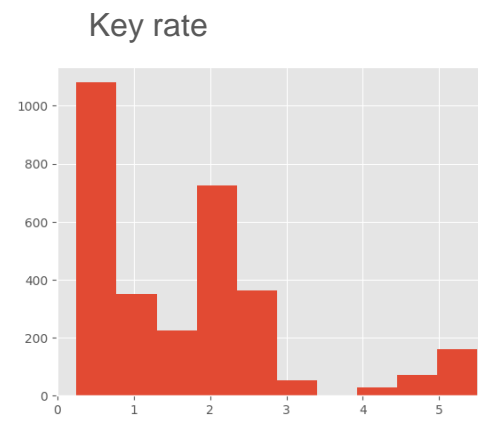
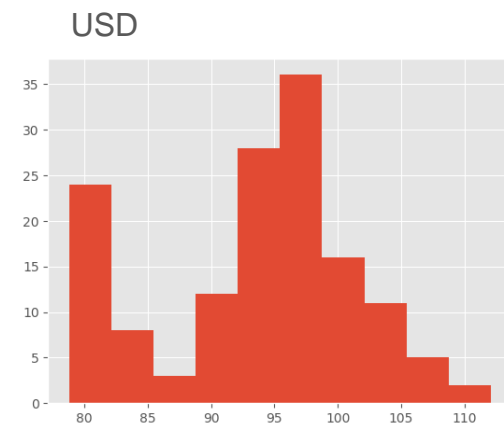
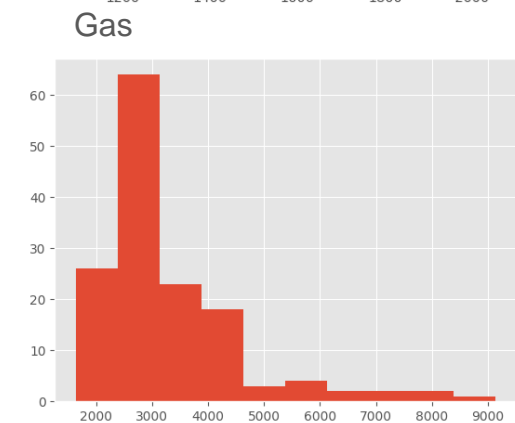
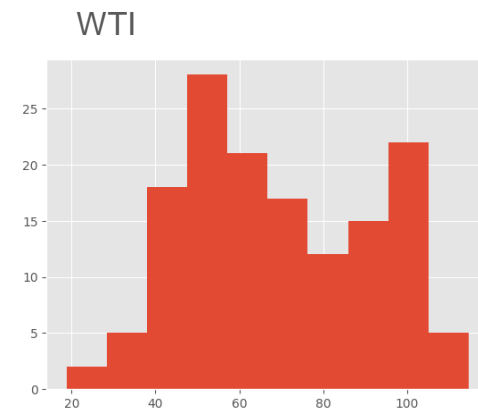
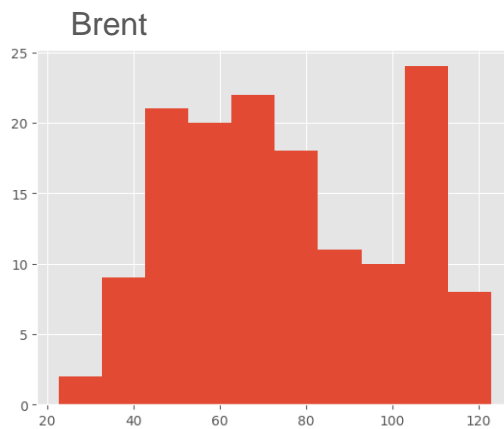
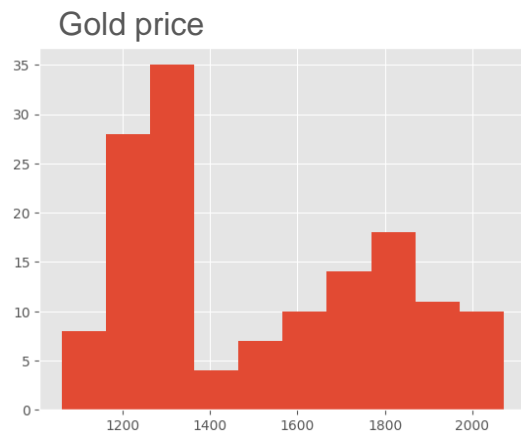
Разница между ценой закрытия и открытия без существенных изменений на протяжении длительного периода, аномальные изменения в начале 2020 года скорее всего связаны с началом эпидемии Covid 19

Разница между ценой открытия и закрытия предыдущего дня



# Влияние различных характеристик на стоимость золота

Для проверки влияния на стоимость взяла факторы: стоимость нефти, газа, курса доллара, ставка ФРС США.



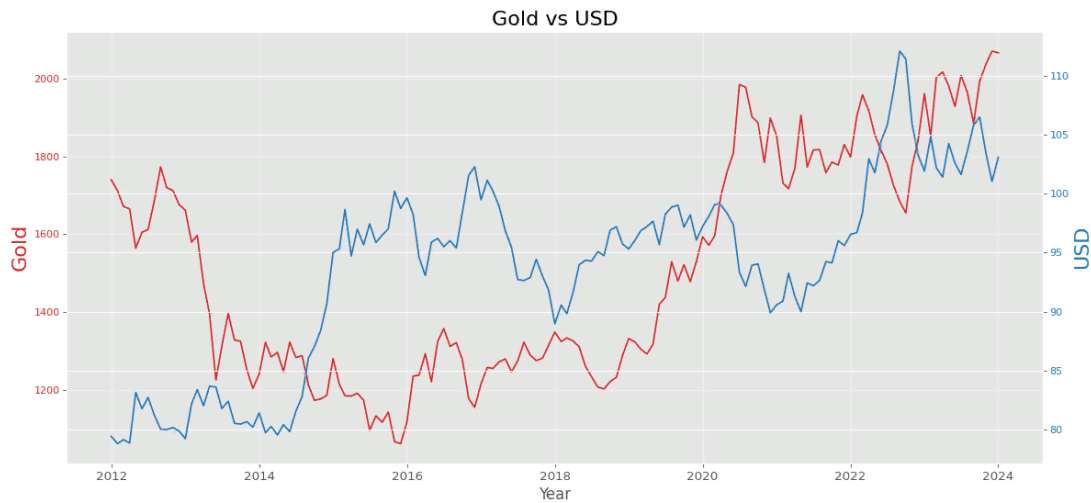
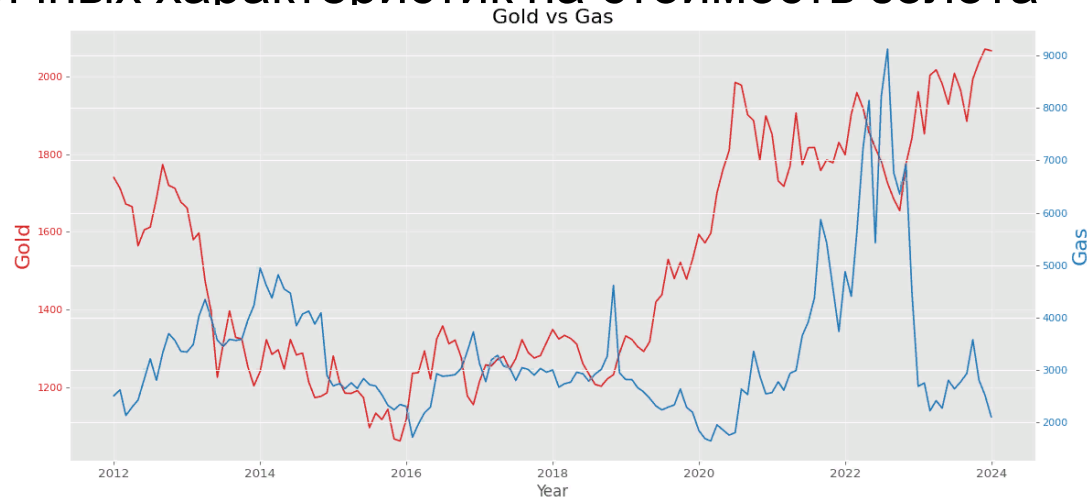
# Влияние различных характеристик на стоимость золота

Прологарифмировала данные и на основе критерия Шапиро-Уилка, получили вывод, что распределения переменных отличается от нормального, Для проверки гипотезы о влиянии факторов на стоимость золота применялся коэффициент корреляции Кенделла.

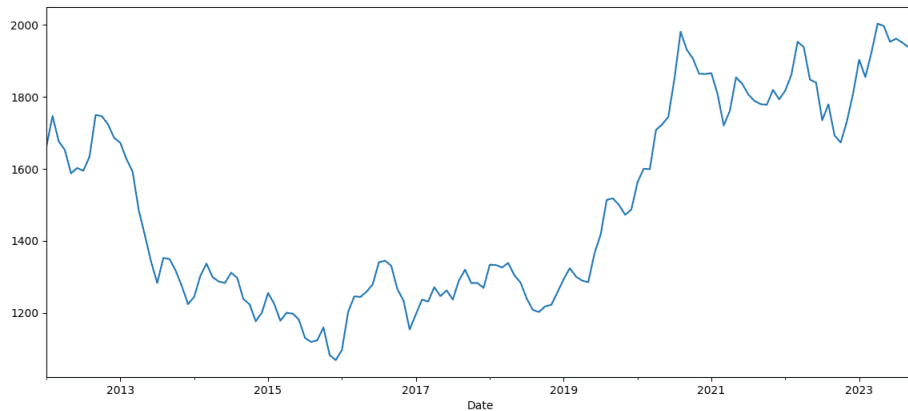
Факторы	коэффициент корреляции Спирмена/ статистическая значимость	коэффициент корреляции Кенделла/ статистическая значимость
Gold - Brent	statistic=0.285, pvalue=0.0005	statistic=0.191, pvalue=0.0006
Gold - WTI	statistic=0.286, pvalue=0.0004	statistic=0.190, pvalue=0.0006
Gold - Gas	statistic=0.007, pvalue=0.9240	statistic=0.005, pvalue=0.9278
Gold - USD	statistic=0.159, pvalue=0.0560	statistic=0.084, pvalue=0.1360
Gold - Key rate	statistic=0.372, pvalue=5.302e-101	statistic=0.280, pvalue=1.194e-103

Обнаружена слабая связь между стоимостью золота и стоимостью газа, а также между стоимостью золота и индексом доллара США.

# Влияние различных характеристик на стоимость золота

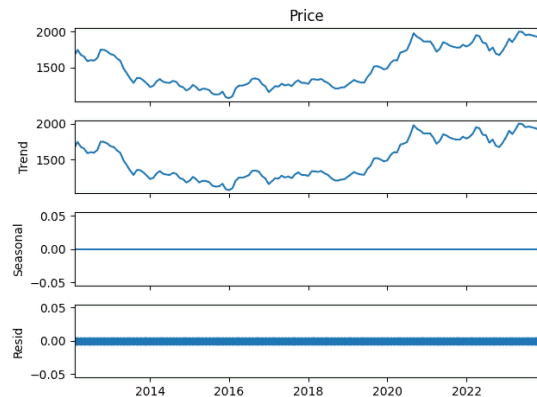


Перейдем от дней к месяцам



Визуально ряд не имеет тренда, провели тест Дики Фуллера. Получили  $p$ -значение 0,73, не смогли отвергнуть нулевую гипотезу.

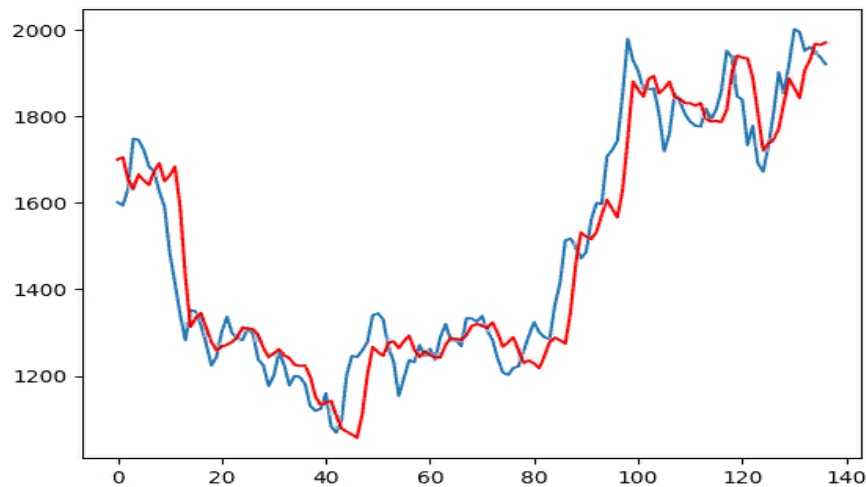
Результат декомпозиции временного ряда



После дифференцирования ряда проверили полученный новый ряд на стационарность с помощью расширенного теста Дики-Фуллера. Ряд стационарен

# Модель взвешенная скользящая средняя с фиксированным окном

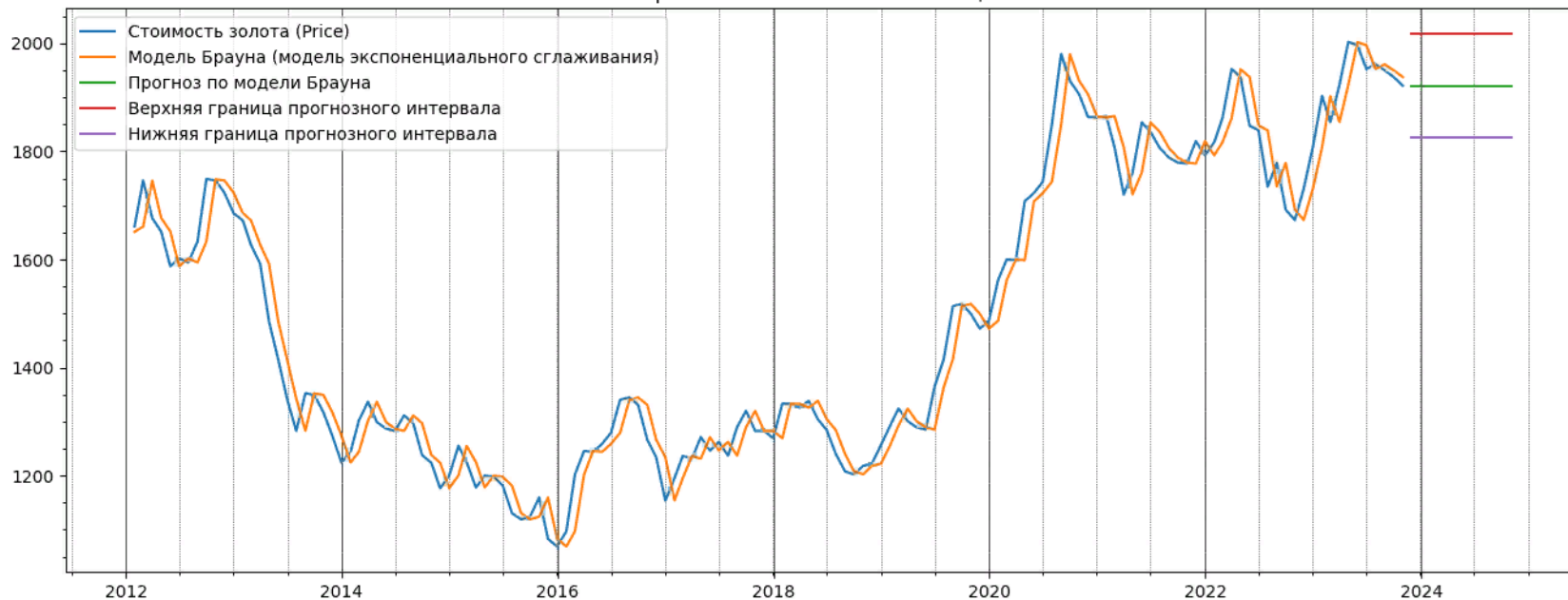
В случае взвешенной выбрали весами для сглаживания по полиномам 2-го/3-го порядка (ширина окна 5).



MAE = 64.593  
 RMSE = 68.200  
 MAPE = 0.044

# Модель Брауна

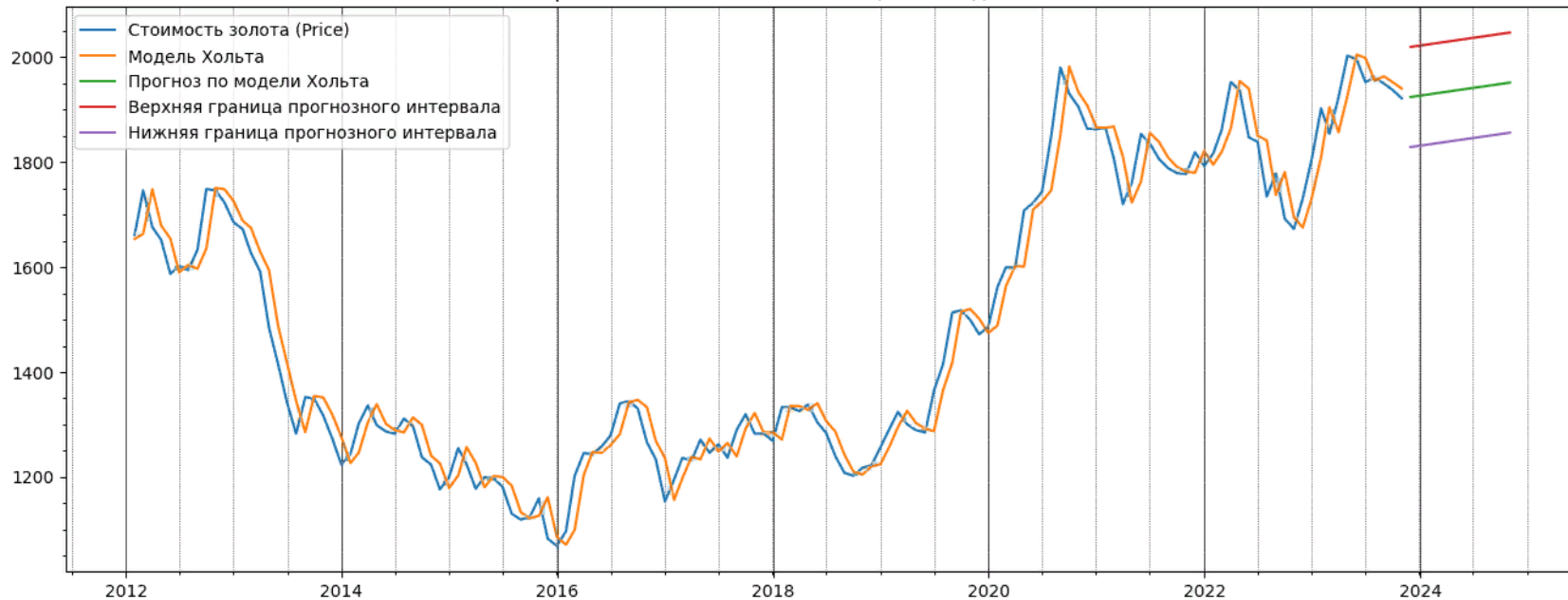
Прогноз стоимости на 12 месяцев



MAE = 38.280  
RMSE = 48.361  
MAPE = 0.025

# Модель Хольта

Прогноз стоимости на 12 месяцев по модели Хольта



MAE = 64.593

RMSE = 48.329

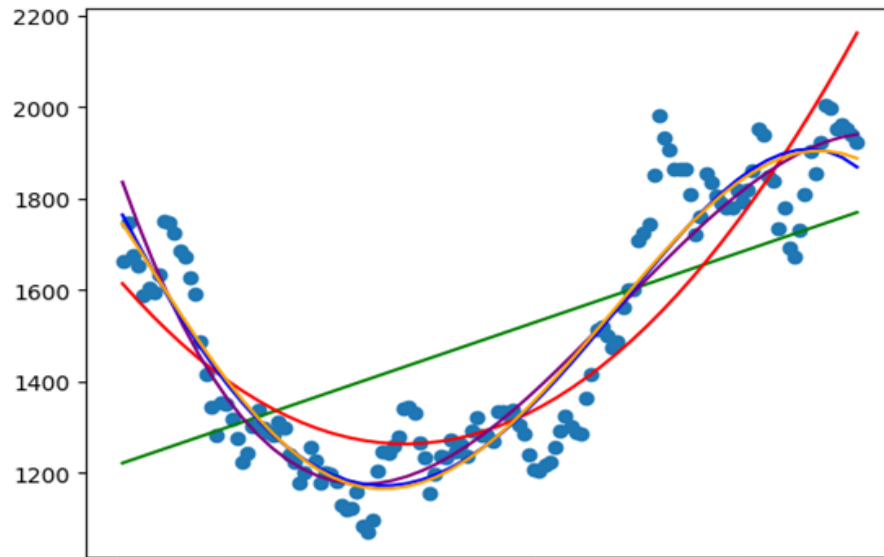
MAPE = 0.025

# Моделирование тренда на основе кривых роста

Сопоставим несколько кривых:

```
1 model1 = np.poly1d (np.polyfit (df.t , df.Price , 1))
2 model2 = np.poly1d (np.polyfit (df.t , df.Price , 2))
3 model3 = np.poly1d (np.polyfit (df.t , df.Price , 3))
4 model4 = np.poly1d (np.polyfit (df.t , df.Price , 4))
5 model5 = np.poly1d (np.polyfit (df.t , df.Price , 5))
6
7 polyline = np.linspace (1, 142)
8 plt.scatter (df.t , df.Price)
```

Модель по полиному 1 степени - зеленый,  
Модель по полиному 2 степени - красный,  
Модель по полиному 3 степени - фиолетовый,  
Модель по полиному 4 степени - голубой,  
Модель по полиному 5 степени - оранжевый



# Моделирование тренда на основе кривых роста.

Исходя из значений скорректированного R-квадрата каждой модели выберем степень полинома

```
{'r_squared': 0.3340781097469614}  
{'r_squared': 0.7642741103582319}  
{'r_squared': 0.8668604877401711}  
{'r_squared': 0.8748782141989346}  
{'r_squared': 0.8745345450436344}
```

Модель на основе полинома 4-й степени:

$$-1.322e-05 x^4 + 0.002165 x^3 + 0.1176 x^2 - 21.91 x + 1785$$

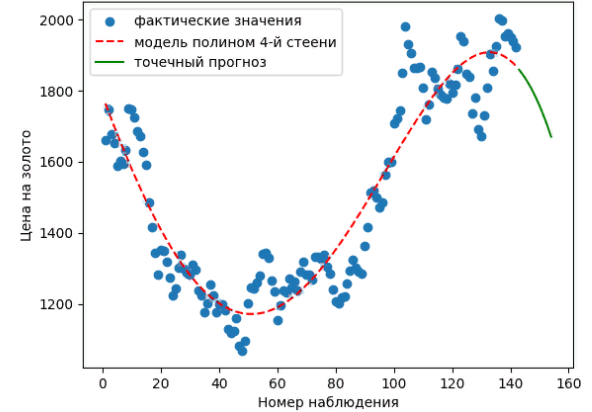
MAE = 72.130  
RMSE = 92.042  
MAPE = 0.039

Модель на основе полинома 3-й степени:

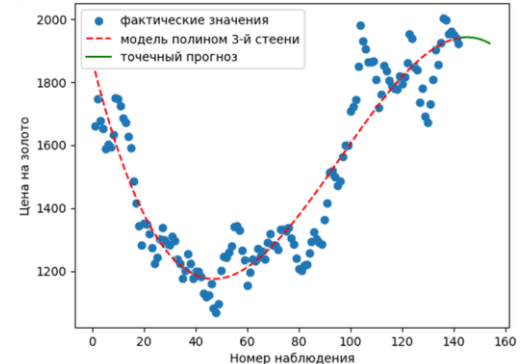
$$-0.001617 x^3 + 0.4661 x^2 - 33.08 x + 1868$$

MAE = 57.891  
RMSE = 82.822  
MAPE = 0.032

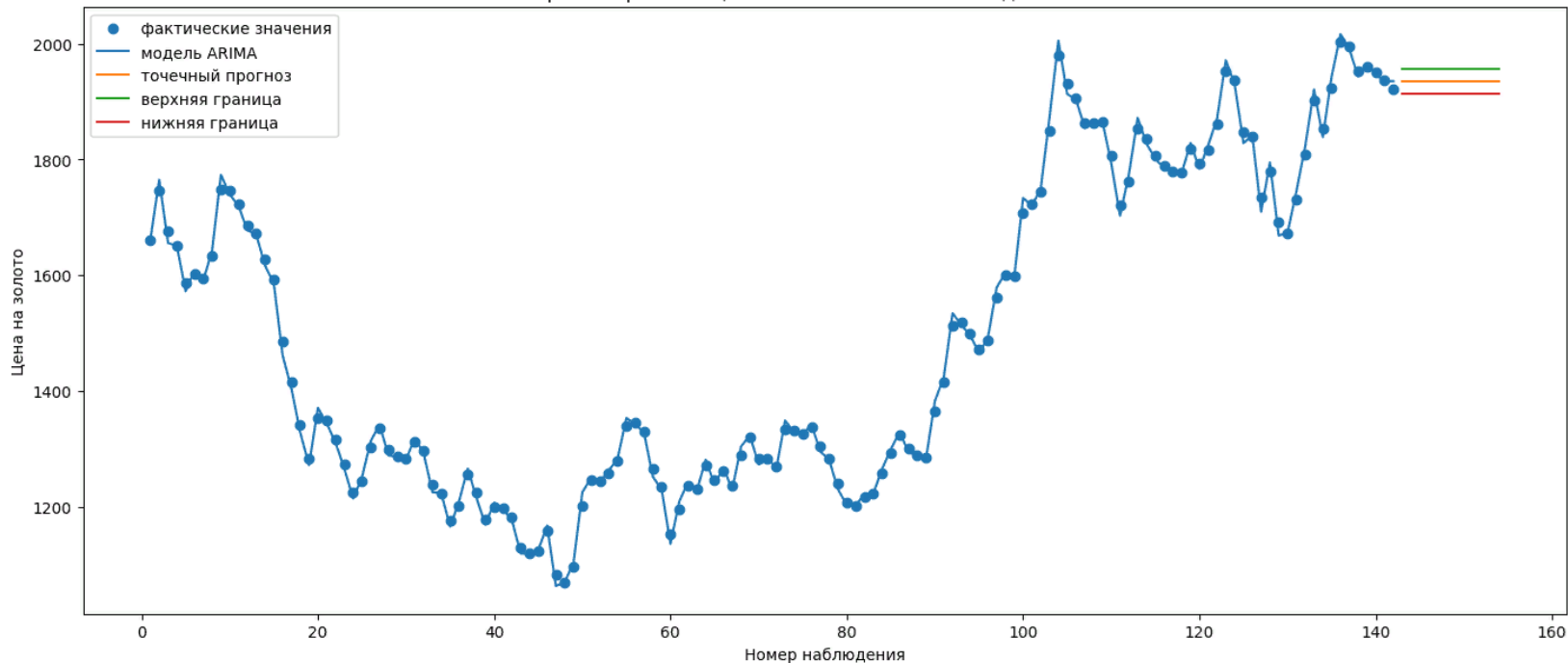
Прогнозирование цены на золото на основе модели полинома 4-й степени



Прогнозирование цены на золото на основе модели полинома 3-й степени



Прогнозирование цены на золото на основе модели ARIMA



MAE = 0.686  
RMSE = 10,956  
MAPE = 0.006

# Сравнение моделей(оценка точности прогноза)

Модель	MAE	MAPE	RMSE
Модель взвешенная скользящая средняя с фиксированным окном	64,59	4,36%	68,20
Модель Брауна	38,28	2,55%	48,36
Модель Хольта	64,59	2,58%	48,32
Кривая роста полином 3-й степени	57,89	3,22%	82,82
ARIMA	0,69	0,006	10,96

Модель ARIMA показала на исследуемых данных самый точный прогноз. MAE - Средняя абсолютная ошибка – степень несоответствия между фактическим и прогнозируемым значением по модели составила всего 0,69 \$, MAPE - средняя абсолютная ошибка 0,6%, RMSE – среднеквадратичная ошибка 10.96.