Прогнозирование цены золота с помощью ARIMA-модели

В теории существует множество фундаментальных, макро- и чувствительных факторов, которые могут влиять на цену золота. С точки зрения фундаментальных факторов, существует спрос со стороны центрального банка, покупателей ювелирных изделий и ETF-фондов на физическое золото, в то время как в макропространстве существует инфляция, процентная ставка, денежная масса и устойчивость валюты. Названные фундаментальные факторы увеличивают цену доллара: чем больше золота правительство и люди покупают, тем меньше золота на рынке, и цена должна вырасти.

В макропространстве уровень инфляции (дефляции) является показателем роста (снижения) цен на корзину товаров и услуг. Если уровень инфляции повысится, то поднимутся и цены на золото. И наоборот, если процентная ставка повышается, золото, как инвестиционный инструмент, становится непривлекательным по отношению к казначейским облигациям.

И последнее, но не менее важное, сила валюты может повлиять на цену золота и любых других товаров, выраженных в ней. Если рубль становится слабым, другие страны могут купить больше рублей, а затем больше золота, что приводит к росту цены на золото и другие товары, если они котируются в рублях.

Мы будем анализировать исторические данные цены на золото. Совокупность цен в конкретный момент времени является временным рядом. **Временной ряд** - это совокупность значений какого-либо показателя за несколько последовательных периодов времени. Отличительная особенность статистического анализа временных рядов состоит в том, что последовательность наблюдений $y(t_1), y(t_2), ..., y(t_r)$ рассматривается как реализация последовательности, вообще говоря, статистически зависимых случайных величин. Чтобы сделать задачу статистического анализа временных рядов доступной для практического решения, приходится ограничивать класс рассматриваемых моделей временных рядов, вводя те или иные предположения относительно структуры ряда и структуры его вероятностных характеристик. Одно из таких ограничений предполагает **стационарность** временного ряда. Под стационарностью мы будем понимать, что у временного ряда некоторые свойства не зависят от времени. Существует 2 типа стационарности строгая стационарность, или стационарность в узком смысле, и слабая стационарность, или стационарность в широком смысле. Мы будем использовать вторую из них. Дадим её определение:

Слабая стационарность, или стационарность в широком смысле

Если случайный процесс таков, что у него математическое ожидание и дисперсия существуют и не зависят от времени, а автокорреляционная (автоковариационная) функция зависит только от разности значений (t1 t2), то такой процесс мы назовем стационарным в широком смысле, или слабо стационарным. Следовательно, построив графики ряда и скользящей статистики, мы сможем определить является ли ряд стационарным в широком смысле. Также это можно сделать с помощью теста Дики-Фуллера, которым мы м будем пользоваться. Автоковариационной функцией случайного процесса называется совокупность значений ковариаций при всевозможных значениях расстояния между моментами времени.

Интегрированная модель авторегрессии — скользящего среднего (ARIMA)

Если ряд после взятия d последовательных разностей приводится к стационарному, то назовем этот ряд ARIMA(p, d, q). ARIMA - процесс авторегрессии - интегрированного скользящего среднего. При этом p - параметр AR-части, d - степень интеграции, и q - это параметр MA-части.

Задачу построения модели типа ARIMA по реализации случайного процесса можно разбить на несколько этапов.

I этап

1. Установить порядок интеграции d, то есть добиться стационарности ряда, взяв достаточное количество последовательных разностей.

2. После этого мы получаем временной ряд Y к которому нужно подобрать уже ARMA(p,q). Исходя из поведения автокорреляционной и частной автокорреляционной функций, установить параметры p и q.

Модель ARMA обобщает две более простые модели временных рядов — модель авторегрессии (AR) и модель скользящего среднего (MA). ARMA(p, q), где p и q — целые числа, задающие порядок модели, имеет вид:

$$X_t = c + \varepsilon_t + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i},$$

где c — константа, ε_t — белый шум, то есть последовательность независимых и одинаково распределённых случайных величин (как правило, нормальных), с нулевым средним, а $\alpha_1, \ldots, \alpha_p$ и $\beta_1, \ldots, \beta_q$ — действительные числа, авторегрессионные коэффициенты и коэффициенты скользящего среднего, соответственно.

Такая модель может интерпретироваться как линейная модель множественной регрессии, в которой в качестве объясняющих переменных выступают прошлые значения самой зависимой переменной, а в качестве регрессионного остатка — скользящие средние из элементов белого шума. ARMA-процессы имеют более сложную структуру по сравнению со схожими по поведению AR- или MA-процессами в чистом виде, но при этом они характеризуются меньшим количеством параметров, что является одним из их преимуществ.

ARMA-процессы можно считать MA-процессами бесконечного порядка с определенными ограничениями на структуру коэффициентов. Все стационарные процессы можно сколь угодно приблизить ARMA-моделью некоторого порядка.

I этап принято называть идентификацией модели ARIMA(p,d,q). Это всего лишь определение величин p, d, q, но именно в такой последовательности: сначала d, а потом p и q.

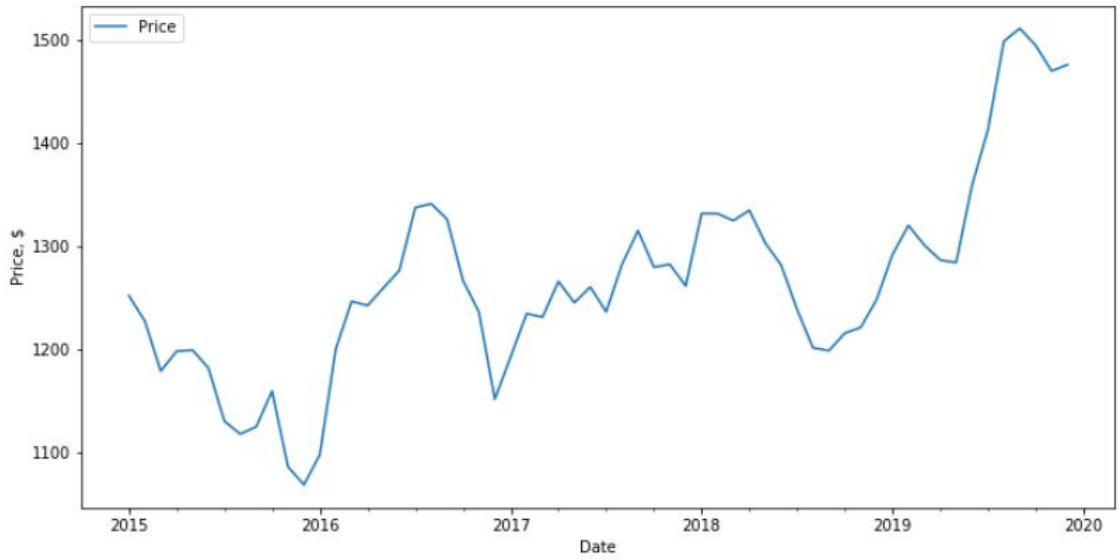
II этап Оценка коэффициентов $\alpha_1, \alpha_2, \ldots, \alpha_p, \beta_1, \beta_2, \ldots, \beta_q$ при условии, что мы уже знаем p и q.

III этап По остаткам осуществляется тестирование или диагностика построенной модели.

IV этап Использование модели для прогнозирования будущих значений цены золота

Построение модели

Прогноз цены на золото в этой работе основан на исторических данных с сайта https://www.gold.org/. Мы рассматриваем средние месячные цены за последние 5 лет. Так выгдядит дикамика цены золота:



Проверка на стационарность

Визуализируем скользящую среднюю (англ. Simple Moving Average, SMA), а также проверим есть ли у наших наблюдений тренд и сезональность. Затем, подтвердим наше предположение о стационарности ряда тестом Дики-Фуллера.

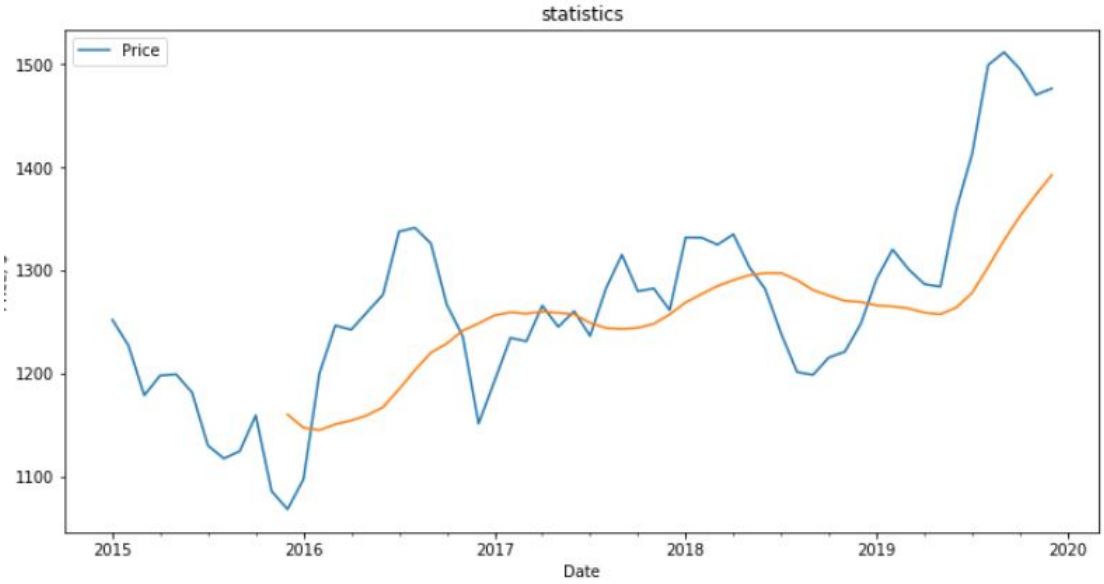
Прежде, чем рассчитывать сезонную компоненту, исходный временной ряд необходимо выровнять методом скользящих средних. **Скользящая средняя** представляет собой последовательную серию средних значений с определенным периодом сглаживания и рассчитывается с целью определения тенденции изменения случайной величины. При этом чем больше будет период сглаживания, тем более плавным будет график полученной линии. Вычисляя скользящее среднее для временного ряда, интервал сглаживания (ширина окна) берется равным периоду сезонности. После определения скользящих средних вся сезонная (т.е. внутри сезона) изменчивость будет исключена и поэтому разность (в случае аддитивной модели) или отношение (для мультипликативной модели) между наблюдаемым и сглаженным рядом будет выделять сезонную составляющую плюс нерегулярную компоненту.

Итак, добавим скользящую среднюю с интервалом сглаживания равным 12 месяцев. Её значения в каждой точке вычисляются по формуле:

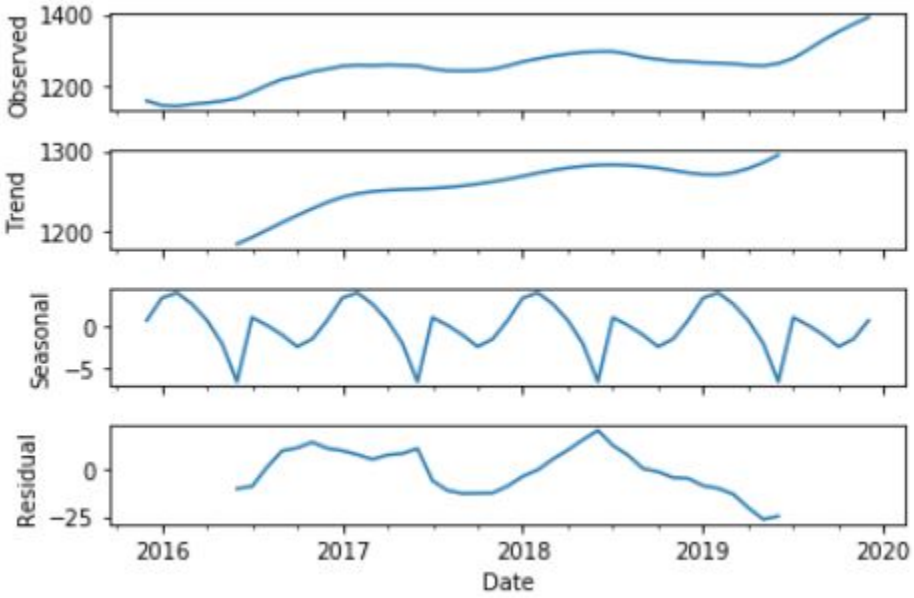
$$SMA_t = 1/12 * \sum_{i=0}^{11} P_{t-i},$$

где P_{t-i} - значение случайной величины на момент (t-i).

Получим:



Далее необходимо выяснить есть ли у полученной выборки тренд и каков период сезональности. **Тренд** — тенденция изменения показателей временного ряда. Тренды могут быть описаны различными функциями — линейными, степенными, экспоненциальными и т. д. Тип тренда устанавливают на основе данных временного ряда, путем осреднения показателей динамики ряда, на основе статистической проверки гипотезы о постоянстве параметров графика. При иследовании были рассмотрены мультикативная и аддитивная модели, они дали похожие результаты. Поэтому в данной работе продемонстрируем результаты одной из них, аддитивной модели:



Как видно из графиков, временной ряд не является стационарным, так как величины математического ожидания и дисперсии будут зависеть от момента t. Для посторения про-

гноза нам нужно определить его степень интегрируемости d. Для этого мы проведём тест Дики-Фуллера.

Тест Дики-Фуллера

Тест Дики — Фуллера (DF-тест, Dickey — Fuller test) — это методика, которая используется в прикладной статистике и эконометрике для анализа временных рядов для проверки на стационарность. Является одним из тестов на единичные корни (*Unit root test*).

Понятие единичного корня

Временной ряд имеет единичный корень, или порядок интегрирования один, если его первые разности образуют стационарный ряд. Это условие записывается как $y_t \sim I(1)$ если ряд первых разностей $\triangle y_t = y_t - y_{t-1}$ является стационарным $\triangle y_t \sim I(0)$. При помощи этого теста проверяют значение коэффициента a авторегрессионном уравнении первого порядка AR(1):

$$y_t = a \cdot y_{t-1} + \varepsilon_t,$$

где y_t — временной ряд, а ε — ошибка. Если $a = 1$, то процесс имеет единичный корень, в этом случае ряд y_t не стационарен, и является интегрированным временным рядом первого порядка — $I(1)$. Если $|a| < 1$, то ряд стационарный — $I(0)$.

Для финансово-экономических процессов значение $|a| > 1$ не свойственно, так как в этом случае процесс является «взрывным». Возникновение таких процессов маловероятно, так как финансово-экономическая среда достаточно инерционная, что не позволяет принимать бесконечно большие значения за малые промежутки времени.

Сущность DF-теста

Приведенное авторегрессионное уравнение AR(1) можно переписать в виде: $\triangle y_t = b \cdot y_{t-1} + \varepsilon_t$, где $b = a - 1$, а $\triangle$ — оператор разности первого порядка $\triangle y_t = y_t - y_{t-1}$. Поэтому проверка гипотезы о единичном корне в данном представлении означает проверку нулевой гипотезы о равенстве нулю коэффициента b. Поскольку случай «взрывных» процессов исключается, то тест является односторонним, то есть альтернативной гипотезой является гипотеза о том, что коэффициент b меньше нуля. Статистика теста (DF-статистика) —это обычная t-статистика для проверки значимости коэффициентов линейной регрессии. Однако, распределение данной статистики отличается от классического распределения t-статистики (распределение Стьюдента или асимптотическое нормальное распределение). Распределение DF-статистики выражается через винеровский процесс и называется распределением Дики — Фуллера.

Существует три версии теста (тестовых регрессий):

1. Без константы и тренда
$$\triangle y_t = b \cdot y_{t-1} + \varepsilon_t.$$
2. С константой, но без тренда
$$\triangle y_t = b_0 + b \cdot y_{t-1} + \varepsilon_t.$$

3. С константой и линейным трендом

$$\triangle y_t = b_0 + b_1 \cdot t + b \cdot y_{t-1} + \varepsilon_t.$$

Для каждой из трёх тестовых регрессий существуют свои критические значения *DF*-статистики, которые берутся из специальной таблицы Дики — Фуллера (МакКиннона). Если значение статистики лежит левее критического значения (критические значения — отрицательные) при данном уровне значимости, то нулевая гипотеза о единичном корне отклоняется и процесс признается стационарным (в смысле данного теста). В противном случае гипотеза не отвергается и процесс может содержать единичные корни, то есть быть нестационарным (интегрированным) временным рядом.

Расширенный тест Дики — Фуллера (ADF)

Если в тестовые регрессии добавить лаги первых разностей временного ряда, то распределение DF-статистики (а значит, критические значения) не изменится. Такой тест называют *расширенным тестом Дики —Фуллера* (Augmented DF, ADF). Необходимость включения лагов первых разностей связана с тем, что процесс может быть авторегрессией не первого, а более высокого порядка.

Замечание

Тест Дики — Фуллера, как и многие другие тесты, проверяют наличие лишь одного единичного корня. Однако, процесс может иметь теоретически несколько единичных корней. В этом случае тест может быть некорректным. Поскольку обычно предполагается, что больше трёх единичных корней вряд ли могут встречаться в реальных экономических временных рядах, то теоретически обоснованным является тестирование в первую очередь вторых разностей ряда. Если гипотеза единичного корня для этого ряда отвергается, то тогда тестируется единичный корень в первых разностях. Если на этом этапе гипотеза не отвергается, то исходный ряд имеет два единичных корня. Если отвергается, то проверяется единичный корень в самом временном ряде, как описано выше. На практике часто все делают в обратной последовательности, что не совсем корректно. Для корректных выводов необходимы результаты тестов для вторых и первых разностей наряду с самим временным рядом.

Анализ наблюдений

1 шаг

Проведём тест Дики-Фуллера для исходного ряда, с учётом константы и тренда, при уровне значимости 5% получим

ADF-статистика: -3.317167

Критические значения:

1%: -4.137

5%: -3.495

10%: -3.176

Значение статистики лежит правее критического значения при данном уровне значимости, следовательно, нулевая гипотеза о единичном корне не отвергается. Процесс может содержать единичные корни, то есть является нестационарным (интегрированным) временным рядом.

2 шаг

Проверим с помощью теста Дики-Фуллера является ли временной ряд интегрированным порядка один, то есть образуют ли его первые разности $\triangle y_t = y_t - y_{t-1}$ стационарный ряд. Повторив тест для новых данных имеем:

ADF-статистика: -5.217997836785988

Критические значения:

1%: -4.130261462053571

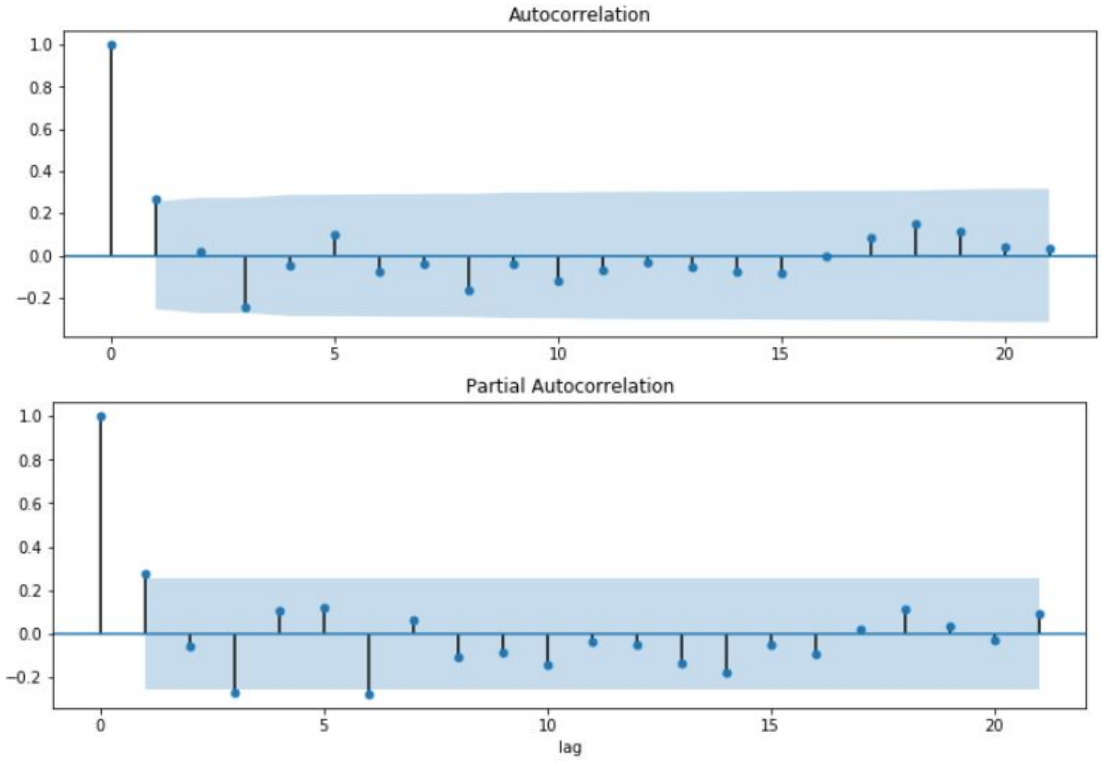
5%: -3.4920297480867344

10%: -3.1746004600947524

Значение статистики лежит левее критического значения при данном уровне значимости(5%), следовательно, нулевая гипотеза о единичном корне отклоняется и процесс признается стационарным.

Построение сезонной модели ARIMA (SARIMA)

Для прогнозирования цены на ближайшие 5 лет с помощью модели ARIMA, построенную для ряда первых разностей, нам нужны 3 параметра: p — порядок компоненты AR, d — порядок интегрированного ряда, q — порядок компонетны MA. Параметр d мы уже нашли, он равен 1. По коррелограмме ACF определяем q = количество автокорреляционных коэффициентов сильно отличных от 0 в модели MA. По коррелограмме PACF можно определить p = максимальный номер коэффициента сильно отличный от 0 в модели AR. Так же при выборе параметров следует руководствоваться **информационным критерием Акаике (AIC)**.



Теперь необходимо определить сезонные параметры P,D,Q,S. Последний из них мы уже знаем, S = 12. Мы берём D = 1, так как предполагаем явно выраженную сезональность. P = 2, Q = 0 мы определили из автокоррелограмы и по информационному критерию AIC.

Информационный критерий Акаике (AIC)

Критерий для выбора лучшей из нескольких статистических моделей, построенных на одном и том же наборе данных. Предложен Хироцугу Акаикэ в 1974 году. Критерий является не статистическим, а информационным, поскольку основан на оценке потери информации при уменьшении числа параметров модели. Критерий позволяет найти компромисс между сложностью модели (числом параметров) и ее точностью. В общем случае AIC вычисляется по формуле:

$$AIC = 2k - 2ln(L),$$

где k — число параметров модели, L — максимизированное значение функции правдоподобия модели. Лучшей признается та модель, для которой значение AIC минимально.

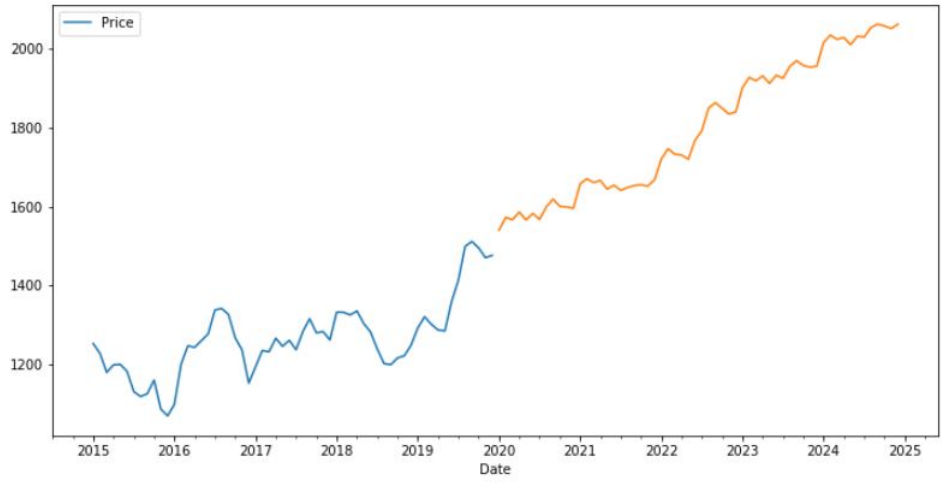
Из выражения видно, что при фиксированном размере выборки рост критерия обусловлен в основном увеличением числа параметров модели, а не ее ошибкой. Т.е. за увеличение числа параметров модель «штрафуется» сильнее, чем за долю необъясненной дисперсии ошибки. Таким образом, задача заключается в том, чтобы выбрать модель с минимальным числом параметров, которые объясняют наибольшую долю дисперсии ошибки. На практике это делается следующим образом. Берется «нулевая модель», которая содержит только свободный член, и для нее вычисляется значение критерия. Затем в нулевую модель поочередно добавляются параметры, и каждый раз AIC вычисляется вновь. Выбирается модель, для которой значение критерия окажется минимальным. Некоторые из них:

AIC (SARIMA(0,1,0)x(2,1,0,12)) = 493.525
AIC (SARIMA(0,1,0)x(0,1,0,12)) = 511.374
AIC (SARIMA(0,1,0)x(1,1,0,12)) = 503.019
AIC (SARIMA(1,1,0)x(2,1,0,12)) = 492.225
AIC (SARIMA(1,1,1)x(2,1,0,12)) = 494.082
AIC (SARIMA(2,1,0)x(2,1,0,12)) = 493.776
AIC (SARIMA(6,1,3)x(2,1,0,12)) = 499.203
AIC (SARIMA(6,1,0)x(2,1,0,12)) = 493.563

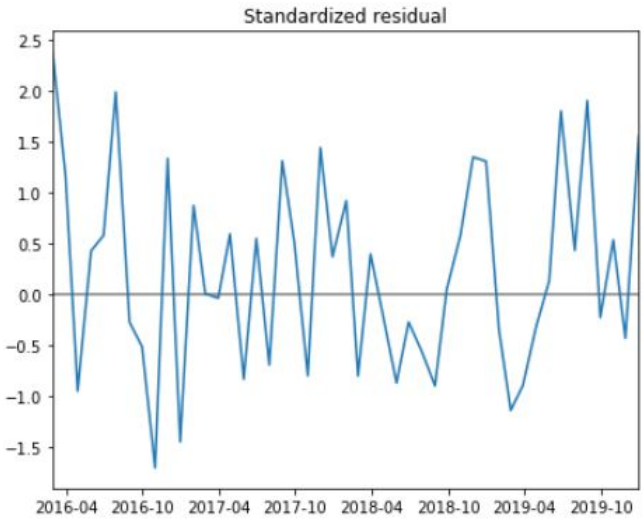
Мы выбираем модель с наименьшим AIC = 492.225 – SARIMA(1,1,0)x(2,1,0,12)

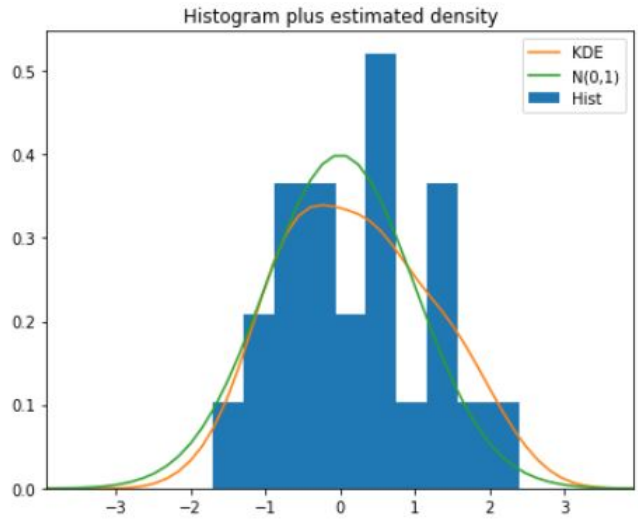
Посмотрим какой прогноз соответствует этой модели:

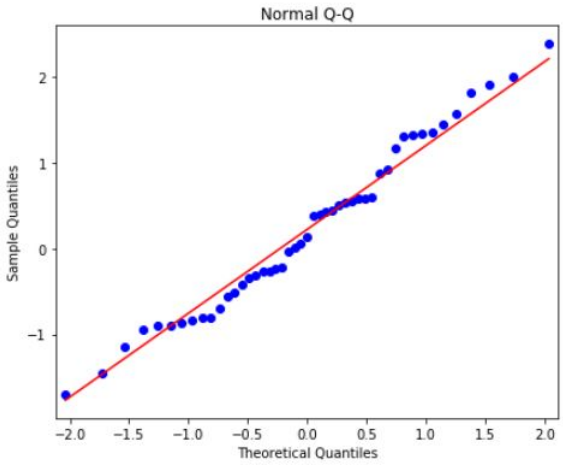
	coef	std err
ar.L1	0.2695	0.163
ar.S.L12	-0.8213	0.161
ar.S.L24	-0.5631	0.123
sigma2	1322.6314	368.161

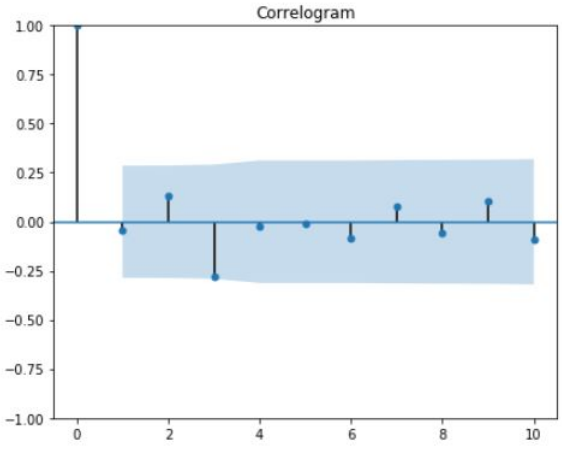


Проверим теперь нашу гипотезу. Воспользуемся инструментами того же класса, с помощью которого мы смоделировали динамику цен. Диагностика выдаёт следующее:









Согласно нашей диагностике нормальный график квантиль-квантиль и гистограмма показывают, что остатки нормально распределены. Коррелограмма также показывает, что они не связаны между собой.**Числовые данные**:

Дата:	**Цена:**
2020-01-31	1539.883985
2020-02-28	1572.350489
2020-03-31	1566.375864
2020-04-30	1586.130297
2020-05-29	1565.959144
2020-06-30	1582.453579
2020-07-31	1567.448505
2020-08-31	1599.126780
2020-09-30	1619.037037
2020-10-30	1600.480286
2020-11-30	1599.046752
2020-12-31	1595.248111
2021-01-29	1657.622950
2021-02-26	1670.649248
2021-03-31	1660.772221
2021-04-30	1666.246221
2021-05-31	1644.141662
2021-06-30	1654.389508
2021-07-30	1641.389888
2021-08-31	1648.218374
2021-09-30	1653.440772
2021-10-29	1655.390270
2021-11-30	1651.868049
2021-12-31	1667.930724
2022-01-31	1720.241746
2022-02-28	1746.893212
2022-03-31	1732.796114
2022-04-29	1730.731278
2022-05-31	1720.171297
2022-06-30	1768.546006
2022-07-29	1792.724108
2022-08-31	1850.449773
2022-09-30	1863.567701
2022-10-31	1849.828458
2022-11-30	1834.874629
2022-12-30	1840.154784
2023-01-31	1901.558306
2023-02-28	1927.967012
2023-03-31	1919.533307
2023-04-28	1931.701629
2023-05-31	1912.749097

2023-06-30 1933.328689	2024-01-31 2016.430664	2024-08-30 2054.435667
2023-07-31 1925.844244	2024-02-29 2035.366059	2024-09-30 2063.426740
2023-08-31 1955.762764	2024-03-29 2024.657568	2024-10-31 2058.761223
2023-09-29 1970.667385	2024-04-30 2029.381833	2024-11-29 2052.132982
2023-10-31 1958.265351	2024-05-31 2010.820862	2024-12-31 2063.068812
2023-11-30 1953.876238	2024-06-28 2032.757733	
2023-12-29 1956.827412	2024-07-31 2030.341286	