Алгоритм ARIMA (Харь Александра)

Целью работы является прогноз временного ряда методом ARIMA. Для эксперимента используются данные о цене закрытия акций компании Газпром и Сбербанк с 15 декабря 2017 года по 15 декабря 2020 года с промежутком раз в сутки.

Постановка задачи: решается задача восстановления регрессии. X - пространство объектов, $Y = \mathbb{R}$ - пространство ответов. Есть неизвестная зависимость $y^* : X \to Y$, значения которой известны только на train выборке: $X^l = (x_i, y_i)_{i=1}^l$. Нужно построить зависимость $a : X \to Y$, аппроксимирующую исходную зависимость y^*

ARIMA (autoregressive integrated moving) - обобщением модели авторегрессионного скользящего среднего. Эти модели используются при работе с временными рядами для более глубокого понимания данных или предсказания будущих точек ряда. Модель использует зависимую связь между наблюдением и некоторым количеством запаздывающих наблюдений (авторегрессия : AR), использует разность необработанных наблюдений (например, вычитание наблюдения из наблюдения на предыдущем временном шаге) для того, чтобы сделать временной ряд стационарным (интегрированная : I), использует зависимость между наблюдением и остаточной ошибкой от модели скользящего среднего, примененной к лаговым наблюдениям (скользящая средняя : MA).

Модель ARIMA(p, q, d) означает, что разности временного ряда порядка d подчиняются модели ARMA(p, q).

В этой модели ARIMA(p, d, q) есть три параметра:
• p - число наблюдений отставания, включенных в модель, также называемое порядком отставания;
• d - количество раз, когда исходные наблюдения различаются, также называется степенью различия;
• q - размер окна скользящей средней, также называемый порядком скользящей средней.

Модель ARIMA(p, d, q) для нестационарного временного ряда X_t имеет вид:

$$\Delta^d X_t = c + \sum_{i=1}^{p} a_i \Delta^d X_{t-i} + \sum_{j=1}^{q} b_j \epsilon_{t-j} + \epsilon_t, \qquad (1)$$

где ϵ_t - стационарный временной ряд, c, a_i, b_j - параметры модели, Δ^d - оператор разности временного ряда порядка d (последовательное взятие d раз разностей первого порядка - сначала от временного ряда, затем от полученных разностей первого порядка, затем от второго порядка и т.д.)

Можно записать с помощью лагового оператора $Lx_t = x_{t-1}$:

$$(1-L)^d X_t = c + (\sum_{i=1}^{p} a_i L^i)(1-L)^d X_t + (1 + \sum_{j=1}^{q} b_j L^j)\epsilon_t \tag{2}$$

Пример: $x_t = x_{t-1} + \epsilon_t \Rightarrow_t = (1-L)x_t = \epsilon_t \Rightarrow$ это модель ARIMA(0, 1, 0)

В эксперименте я в итоге использовала следующие значения параметров: (p, d, q) = (5, 2, 0) для данных Сбербанка и (p, d, q) = (6, 2, 0) для данных Газпрома, именно такой набор хорошо подошел под мои данные.

Данные на train и test делю в пропорции 7:3.

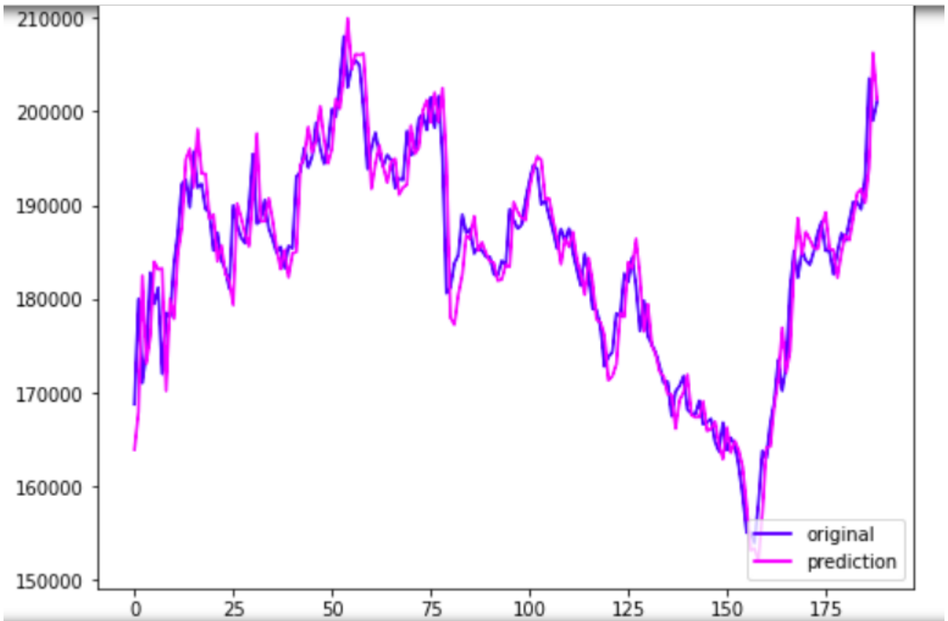
В качестве метрики качества я использовала MSE, и (для наглядности) средний модуль отклонения.

Приведу результат работы алгоритма с данными Сбербанка:

Здесь Test MSE: 23648939.295, что является неплохим результатом, так как значения достаточно большие (порядка 200000)

В среднем модуль отклонения равен 17843.62 что составляет около 8%.

И результат работы алгоритма с данными Газпрома:



Здесь Test MSE: 14000186.809

В среднем модуль отклонения равен 12766.24 что составляет около 7%.