***Отчет по Task1***

Исходный .txt файл я решил преобразовать в csv-таблцу для дальшнейшего удобства в работе с данными.

Первый этап — отрисовка временного ряда и получение его основных характеристик — среднего, дисперсии и квантилей — с помощью функции describe(). Из нее можно видеть, что дисперсия временного ряда близка к 1, что говорит о его возможной стационарности, так как подобным свойством обладает белый шум — ряд со стандартным нормальным распределением *N(0, 1)*. Но отрисовав только первые 200 значений, я заметил, что в данных имеется сезонность размером в 75 шагов, которая начинается с самого начала. Поэтому стационарность оказалась под вопросом, хотя даже тест Дики-Фуллера посчитал исходный ряд стационарным. Отмечу, что количество операций дифференцирования ряда, необходимое для приведения его к стационарным называется порядком интегрированности временного ряда, а это один из параметров модели ARIMA, которую я решил использовать для прогнозирования.

Второй этап — построение и настройка модели ARIMA. Для определения параметров p и q я воспользовался соотвествующими графиками — PACF и ACF (частичная и обычная автокорреляции). Число сильно выделяющихся точек на этих графиках и определяет соотвествующие параметры. Изначальная выборка была разделена на обучающую и тестовую. Для настройки модели используется цикл по всем возможным тройкам параметров *(p, d, q),* а отбор параметров, соответвующих наиулчшей модели, осуществляется с помощью метрики R2. Чтобы понять, насколько хорошо каждая модель делает прогноз, обучающая выборка пошагово дополняется одним значением из выборки, спрогнозированное значение записывется в массив, и после данного цикла высчитывается отклонение тестовой выборки от предсказанных значений. В процессе обучения данной модели возникла ошибка с параметрами, и как я понял, вызвана она была тем, что хоть дисперсия ряда и была близка к 1, но в ряде присутствовала сезонность, поэтому исходные данные я продифференцировал один раз и обучал ARIMA с d равным уже 1, а не 0. В конце осуществляется отрисовка тестовой выборки и спрогнозированных значений, чтобы можно было визуально оценить качество предсказаний. Я заведомо решил не использовать информационный критерий Акаике в качестве метрики, потому что он наклыдвает штраф и ошибку в прогнозе, и на количество параметров в модели, а их не так уж и много, поэтому приоритетным стало только качество прогнозов. Вполне вероянто, что данную модель можно расширить до SARIMA, чтобы лучше учитывать сезонность, но свои идеи на этот счет я решил воплотить в третьем этапе.

Третий этап — реализация собственного алгоритма прогноза. Так как в исходных данных явно прослеживается сезонность, то я решил, что наилучший способ прогноза на 1 (или же несоколько) шаг — это посмотреть, какие значение были в соотвествующие моменты, но в предыдущие периоды. Всего значений в выборке — 1000, а длина периода — 75, поэтому для любого момента из периода имеется около 13 предыдущих значений, поэтому, применив одно из множества сглаживаний к подвыборке, можно получить прогноз на необходимое число шагов вперед. Для каждой позиции в периоде набор предыдущих значений маленький, поэтому, на мой взгляд, использование экспоненциального сглаживания вполне достаточно. При коэффициенте *a = 0.85* прогнозируемое значение близко к среднему по этой маленькой выборке, поэтому можно сделать вывод, что скользящее среднее не только делает нормальный прогноз, но и учитывает несколько предыдущих значений.

Четвертый этап — прогнозирование с помощью алгоритма градиентного бустинга. Будем считать, что текущее значение данного мне ряда зависит от 15 предыдущих. Основываясь на этом, составим матрицу X и столбец y, где строка матрицы — 15 предыдущих значений, а соотвествующий строке элемент столбца — зависимое значение (концепция матрицы признаков и столбца ответов).