***Отчет по Task1***

Исходный .txt файл я решил преобразовать в csv-таблцу для дальшнейшего удобства в работе с данными, а все прогнозы записываются в файл *task1\_report.txt*.

**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. В связи с этим принял решение использовать *auto\_arima*, которая берет подбор вссех параметров на себя, а также предлагает оптимальную модель (*ARMA, ARIMA или SARIMAX*). Но так как даже сами сезонные части не обладали сильным подобием, а значения временного ряда — это 0, …, 6, то даже этот инструмент не помог уловить происходящий процесс, и в итоге поргноз на 1, 3 или 10 шагов вперед — просто прямая линия, равная среднему выборки. Будь значения ряда более разнообразными, или сезонные части сильно похожи друг на друга, результат прогноза оказался бы куда лучше и правдоподобнее. Также было выяснено, что конкретно на этих данных значение параметра *seasonal* не влияет на результат.

**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 или одного из видов регрессии. 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). Испробовав данных подход, понял, что он не дал ожидаемого результата, поэтому изменил концепцию. Так как сезоннность имеется с самого начала выборки, то будем находить, сколько времени прошло с последнего периода, и составим матрицу X, у которой кажадя строка — тот же самый временной период, но в предыдущие сезоны, а столбец y будет составляться как и в самом начале. Данный подход показал чуть более правдободобный результат, но из-за особенностей ряда, описанных во втором этапе, прогноз всё же не до конца точный, особенно если посмотреть на метрики.

**Вывод**: рассмотрев 3 различных подхода, я понял, что лучше всего на именно этом временном ряду себя показывает мой собственный алгоритм. Но, как бы то ни было, 2 других подхода тоже имееют место быть, а выбор наиболее подходящего зависит от данных с которыми приходится работать.