МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ М. В. ЛОМОНОСОВА
Факультет вычислительной математики и кибернетики

Кафедра исследования операций

Отчет по заданию №2

Выполнили студенты: Битиев Алексей
Кулакова Мария

Преподаватель: Гусева Юлия

Москва
2019

Содержание

Постановка задачи:

1) Считать данные из training.xlsx. Ответы на тестовой выборке testing.xlsx не следует использовать ни в каких экспериментах, кроме финального. Проверить является ли ряд стационарным в широком смысле.
   Это можно сделать двумя способами:

   - Провести визуальную оценку, отрисовав ряд и скользящую статистику(среднее, стандартное отклонение). Постройте график на котором будет отображен сам ряд и различные скользящие
   - Провести тест Дики - Фуллера.

   Сделать выводы из полученных результатов. Оценить достоверность статистики. (25 баллов)

2) Разложить временной ряд на тренд, сезональность, остаток в соответствии с аддитивной, мультипликативной моделями. Визуализировать их, оценить стационарность получившихся рядов, сделать выводы.
   (15 баллов)

3) Проверить является ли временной ряд интегрированным порядка k. Если является, применить к нему модель ARIMA, подобрав необходимые параметры с помощью функции автокорреляции и функции частичной автокорреляции. Выбор параметров обосновать. Отобрать несколько моделей. Предсказать значения для тестовой выборки. Визуализировать их, посчитать r2 score для каждой из моделей. Произвести отбор наилучшей модели с помощью информационного критерия Акаике. Провести анализ получившихся результатов. (40 баллов)

Помимо этого обязательные пункты к выполнению:

- 10 баллов - соблюдение PEP8
- 10 баллов - использование для визуализации библиотек bokeh или seaborn.
- 25 баллов - оформление файла readme.pdf
- 25 баллов - прохождение ревью

Решение:

Вводные понятия

Временной ряд (Y) - это последовательность значений, описывающих протекающий во времени процесс, измеренных в последовательные моменты времени, обычно через равные промежутки.

Анализ временных рядов — совокупность математико-статистических методов анализа, предназначенных для выявления структуры временных рядов и для их прогнозирования.

Ряд называется ***стационарным в широком смысле)***, если он имееет независимые *матожидания* (обозначение E(Y)) и *дисперсии* (обозначение D(Y)) от времени, а ковариационная функция зависит только от сдвига.

Математическое ожидание - среднее значение случайной величины (распределение стационарной случайной величины) при стремлении количества выборок или количества измерений её к бесконечности.

Дисперсия - мера разброса значений случайной величины относительно её математического ожидания.

При анализе временных рядов можно рассматривать ковариацию значений ряда в различные моменты времени $Cov[X(t_1), X(t_2)]$. Она не зависит от сдвига времени вперед, а зависит только от разности моментов времени $t_2 - t_1 = T$

Ковариация - мера линейной зависимости двух случайных величин.

Автоковариационная функция - совокупность значений ковариаций при всевозможных значениях T (Обозначим за y).

Коэффициент корреляции - $Corr(T) = \dfrac{y(T)}{y(0)}$. График коэффициента корреляции называется ***коррелограмма***.

При проверки ряда на стационарность можно пользоваться визуальной оценкой. Для этого нужно будет воспользоваться *скользящей статистикой.*

Скользящая средняя - общее название для семейства функций, значения которых в каждой точке определения равны среднему значению исходной функции за предыдущий период. Обычно используется для сглаживания краткосрочных колебаний и выделения основных тенденций.

Стандартное отклонение оценивается по формуле:

$SD = \sqrt{\dfrac{\Sigma_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$, где $\overline{X}$ - среднее значение выборки.

Другим способом является проведение ***теста Дики-Фуллера***, который является одним из тестом на *единичные корни.*

Временной ряд имеет ***единичный корень (порядок интеграции один)***, если его первые разности образуют стационарный ряд.

Тренд (T) временного ряда - изменение, определяющее общее направление развития ряда - рост, падение, неизменность.
Сезонность (S) - периодические колебания уровней временного ряда внутри периода.
Остаток (E) - величина, показывающая нерегулярную (т.е. не описываемую трендом или сезонностью) составляющую исходного ряда в определенном временном интервале.
Существует две модели временных рядов - аддитивная и мультипликативная.
Аддитивная модель имеет вид: $Y = T + S + E$
Мультипликативная модель имеет вид: $Y = T * S * E$
Интегрированный временной ряд - нестационарный временной ряд, разности некоторого порядка от которого являются стационарным рядом. Временной ряд называется *интегрированным порядка* k, если разности k-го порядка являются стационарными, а разности меньшего порядка (и сам временной ряд соответственно) не является стационарными рядами.

Суть метода

1) **Скользящая статистика.**
Скользящая средняя (простое)
$SMA_t = \frac{1}{n}\Sigma_{i=0}^{n-1} p_{t-i}$, где SMA_t -значение простого скользящего среднего в точке t, n - количество значений исходной функции, p_{t-i} значение исходной функции в точке $t-i$
Стандартное отклонение оценивается по формуле:
$SD = \sqrt{\frac{\Sigma_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$, где $\overline{X}$ - среднее значение выборки.

2) **Тест Дики-Фуллера.**
Рассмотрим авторегрессионное уравнение:
$y_t = a * y_{t-1} + \varepsilon_t$, где y_t -временной ряд, а ε - ошибка.
Если $a = 1$, то процесс имеет единичный корень, тогда рассматриваемый ряд нестационарен и является интегрированным временным рядом первого порядка. Если $|a| < 1$, то ряд стационарный.
Для удобства проведением преобразования:
$y_t = a * y_{t-1} + \varepsilon_t$
$y_t - y_{t-1} = a * y_{t-1} - y_{t-1} + \varepsilon_t$
$\Delta y_t = (a-1) * y_{t-1} + \varepsilon_t$, далее

3) **Автокорреляционная функция** - зависимость связи между функцией (сигналом) и её сдвинутой копией от величины временного сдвига. Для детерминированных сигналов автокорреляционная функция сигнала f(t) определяется интегралом:
$\int_{-\infty}^{+\infty} f(t)f(t-\tau)dt$

4) **Модель** $ARIMA(p,d,q)$ для нестационарного временного ряда X_t имеет вид:
$\Delta^d * X_t = c + \Sigma_{i=1}^{p} a_i * \Delta^d * X_{t-1} + \Sigma_{j=1}^{q} b_i * \varepsilon_{t-j} + \varepsilon_t$, 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 и т.д)

Описание программы

1) Чтение данных из "training.xlsx"с помощью функции pd.read_excel('training.xlsx', index_col='Date') из библиотеки "pandas"

2) Нахождение скользящей статистики и стандартного отклонение с помощью функций train.rolling().mean() и train.rolling().std() из библиотеки "pandas". Строим графики и по ним с помощью визуальной оценки определяем стационарность ряда.

3) Раскладываем временной ряд на тренд, сезонности и остаток в соответствии с аддитивной и мультипликативной моделями.
Чтобы найти тренд будем пользоваться методом наименьших квадратов, который основан на минимизации суммы квадратов отклонений некоторых функций от искомых переменных.
Тренд линеен, поэтому он будет иметь вид: $Y = a_0t + a_1$, и система уравнений МНК будет выглядеть следующим образом:

$$\begin{cases} a_0n + a_1\Sigma t = \Sigma Y \\ a_0\Sigma t + a_1\Sigma t^2 = \Sigma Y * t \end{cases}$$, где n - число промежутков.

Чтобы найти сезональность найдем скользящее среднее, после от него еще раз скользящее среднее и получим центрированное скользящее среднее ($\overline{MA}$).
Сезональность найдем по формуле: $S = Y - \overline{MA}$
И остаток $E = Y - T - S$.
В мультипликативной модели будем похожие формулы:

$$S = \frac{Y}{\overline{MA}} \; E = \frac{Y}{T * S}$$

4) Для предсказания реализовывалась функция predict_it(train, test, supposition), использующая модель ARIMA

Библиотеки

- Numpy - для работы с матрицами.
- Pandas - библиотека для хранения таблиц, содержащая функции для их обработки.
- Matplotlib - пакет, для для отрисовки графиков
- seaborn - для красивых графиков.
- statsmodels - пакет для исследования стат. данных.

Необходимые компоненты
Для работы будет необходима программа *Jupyter Notebook*

Вклад участников

Битиев Алексей - задания 2 и 3
Кулакова Мария - задание 1, readme