README
Задание №2

Выполнила
студентка 311 учебной группы факультета ВМК МГУ
Чернобай Анна Александровна

Москва
2021

Оглавление

Глава 1

Постановка задачи

В этом задании необходимо провести анализ некоторого временного ряда и попробовать предсказать значения для последующих месяцев. Что предстоит сделать:

1. Считать данные из Данные.xlsx. Ответы на тестовой выборке Ответы.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).
   - 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. (50 баллов)

Глава 2

Описание алгоритма

Краткие теоретические сведения

Временной ряд - последовательно измеренные через некоторые (зачастую равные) промежутки времени данные.

Интегрированный временной ряд - нестационарный временной ряд, разности некоторого порядка от которого являются стационарным рядом.

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).

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

Скользящее среднее — общее название для семейства функций, значения которых в каждой точке определения равны некоторому среднему значению исходной функции за предыдущий период.

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. (T)

Сезонность - периодически колебания, наблюдаемые на временных рядах.(S)

Остаток - величина, показывающая нерегулярную (не описываемую трендом и се- зонностью) составляющую исходного ряда в определенном временном интервале.(E)

Исходный ряд можно разложить на вышеперечисленные компоненты. Этот процесс называется сезонной декомпо- зицией.

Аддитивная модель - модель, в которой временной ряд представлен как сумма вышеперечисленных компонент. Общий вид аддитивной модели: $Y = T + S + E$

Мультипликативная модель - модель, в которой временной ряд представлен как произведение вышеперечисленных компонент. Общий вид мультипликативной модели: $Y = T * S * E$

Модель ARIMA (autoregressive integrated moving average model) - интегрированная модель авторегрессии скользящего среднего - модель анализа временных рядов. Это расширение моделей ARMA для нестационарных временных рядов.

ARIMA зависит от 3-х параметров: ARIMA(p, d, q), где p - порядок $AR(p)$ - модель авторегрессии, d - порядок интегрированности, q - порядок $MA(q)$ - модель скользящего среднего.

$$\text{ARIMA}(p, d, q) = \text{AR}(p) + \text{MA}(q)\ \text{I}(d)$$

AIC (an information criterion), информационный критерий Акаике - критерий выбора из класса параметризованных регрессионных моделей.

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

где k число параметров в статистической модели, L максимизированное значение функции правдоподобия модели

Алгоритм решения

1. Читаем данные из файла Данные.xlsx с помощью read_excel(pandas)
2. Проверяем ряд на стационарность:
   - Строим график ряда, скользящего среднего (rolling().mean()), скользящего стандартного отклонения (rolling().std()) с помощью Matplotlib
   - Проводим тест Дики-Фуллера с помощью sm.tsa.adfuller(y) (Statsmodels.api). На вход функции подаем временной ряд. На выходе получаем массив с данными о ряде. Нас будут интересовать нулевой и четвертый элемент этого массива. Проанализировав их, делаем вывод о стационарности/ не стационарности ряда. Реализация - Aug_Dickey_Fuller(y - временной ряд)
3. Раскладываем временной ряда на тренд, сезональность и остаток в соответствии с аддитивной и мультипликативной моделями.
   - Для разложения в соотвестии с аддитивной моделью используем sm.tsa.seasonal_decompose(..., model = 'additive') (Statsmodels.api)
   - Для разложения в соотвестии с мультипликативной моделью используем sm.tsa.seasonal_decompose(..., model = 'multiplicate') (Statsmodels.api)
4. Ищем порядок интегрированности ряда в функции Order(y - временной ряд) с помощью теста Дики-Фуллера и Aug_Dickey_Fuller(y - временной ряд). Порядок интегрированности будем использовать в качестве параметра d в модели ARIMA
5. Строим графики автокорреляции и частичной корреляции для подбора параметров q и p модели ARIMA соответсвенно. Для построения графиков используем sm.graphics.tsa.plot_acf и sm.graphics.tsa.plot_pacf соответсвенно (Statsmodels.api).
6. Читаем данные из файла Ответы.xlsx с помощью read_excel(pandas). Далее используем данные из этого файла для отрисовки графиков и сравнения с результатами работы модели ARIMA
7. Пишем функцию arima(data, order, test). На вход данная фукнция получает исходный ряд, (data) параметры для модели ARIMA (order) и продолжение исходного ряда из файла Ответы.xlsx(test).

   В функции вызываем sm.tsa.ARIMA(data, order=order, freq='MS').fit() и с помощью полученного результата (сохранен в model) "предсказываем"продолжение временного ряда data используя model.predict.

Далее на основе полученного временного ряда строим сравнительный график предсказания и временного ряда из файла Ответы.xlsx.

Также считаем для этих двух рядов r2_score (Sklearn.metrics).

Затем выводим значение критерия Акаике для построенной модели (model.aic) и общую информацию о ней (model.summary())

8. Запускаем функцию arima(data, order, test) на различных значениях параметров. Всего тестировалось 9 разных наборов(см. ноутбук).

9. Результаты работы функций на различных наборах параметров, а также и реальное продолжение временного ряда изображаем на одном графике для визуального сравнения.

Глава 3

Инструкция по запуску

Необходимые программы

- Python 3
- Jupyter

Необходимые библиотеки Python

- NumPy
- Pandas
- Matplotlib
- Sklearn.metrics
- Statsmodels.api

Запуск

Далее считаем, что все выше перечисленные программы и библиотеки установлены.

- Для запуска программы необходимо скачать файл time_series.py. Далее в директории, в которую был помещен скачанный файл, открываем терминал и пишем python3 time_series.py, предварительно убедившись, что excel-файлы Данные.xlsx и Ответы.xlsx лежат в той же директории, что и time_series.py.
- Для запуска ноутбука необходимо запустить Jupyter и в нем открыть файл Task_2.ipynb, предварительно убедившись, что excel-файлы Данные.xlsx и Ответы.xlsx лежат в той же директории, что и файл Task_2.ipynb.

Литература

[1] https://habr.com/ru/post/207160/ - Анализ временных рядов с помощью python

[2] https://www.statsmodels.org/ - документация по библиотеке statsmodels

[3] https://pythonpip.ru/examples/model-arima-v-python - Модель ARIMA в Python для прогнозирования временных рядов

[4] http://www.machinelearning.ru/wiki/