Проведение анализа временного ряда

1 Постановка задачи

Необходимо:

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:

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

ii. 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. Надо сделать, чтобы было

25 баллов - оформление файла readme.pd

Реализация:

Будем называть **временным рядом** совокупность значений какого-либо показателя за несколь- ко последовательных периодов времени. Отличительная

особенность статистического анали- за временных рядов состоит в том, что последовательность наблюдений $y(t_1), y(t_2), ..., y(t_r)$ рассматривается как реализация последовательности, вообще говоря, статистически зависи- мых случайных величин. Чтобы сделать задачу статистического анализа временных рядов доступной для практического решения, приходится ограничивать класс рассматриваемых мо- делей временных рядов, вводя те или иные предположения относительно структуры ряда и структуры его вероятностных характеристик. Одно из таких ограничений предполагает ста- ционарность временного ряда. Под стационарностью мы будем понимать, что у временного ряда некоторые свойства не меняются с течением времени. В соответствии с этим рассмотрим
2 типа стационарности:

- **Строгая стационарность**, или стационарность в узком смысле.

Будем называть случайный процесс строго стационарным, если сдвиг во времени не меняет ни одну из функций плотности распределения.То есть, если ко всем моментам времени добавить некоторую целочисленную величину, то сама функция плотности не изменится, $f_n(x_{t1}, ..., x_{tn}) = f_n(x_{t1+\delta}, ..., x_{tn+\delta})$ для всех n, моментов времени $t1, ..., tn$ и це- лочисленных δ. Если плотности распределения не существует, можно сформулировать это определение в терминах функции распределения. Поскольку при каждом фиксиро- ванном t случайный процесс есть случайная величина, то для каждой из них можно рассматривать такие характеристики, как мат.ожидание, дисперсию, автоковариацион- ную и автокорреляционную функции. **Автоковариационной функцией** случайного процесса называется совокупность значений ковариаций при всевозможных значениях расстояния между моментами времени. Если случайный процесс стационарен в узком смысле, то у него:

- математическое ожидание не зависит от времени;
- дисперсия не зависит от времени;
- автоковариационная и автокорреляционные функции: а) зависят только от сдвига, от разности моментов времени; б) являются четными функциями.

Коэффициент корреляции $\rho(\tau) = \gamma(\tau)/\gamma(o)$ между разделёнными на τ значениями временного ряда, где функцию $\gamma(\tau)$ можно рассматривать как всевозможные значения автоковариаций, где τ пробегает целочисленные значения от до + , а $\gamma(0) = Cov(X_t, X_t) = \sigma^2$ - это ковариация, разделенная на корень из произведения двух дисперсий, но так как дисперсия постоянна, то мы получаем просто σ^2. Это выражение определяет **автокорреляционную функцию** временного ряда.

- **Слабая стационарность**, или стационарность в широком смысле.

Если случайный процесс таков, что у него математическое ожидание и дисперсия суще- ствуют и не зависят от времени, а автокорреляционная (автоковариационная) функция зависит только от разности значений (t_1 - t_2), то

такой процесс мы назовем стационар- ным в широком смысле, или слабо стационарным.

Следовательно, построив графики ряда и скользящей статистики, мы сможем определить является ли ряд стационарным в широком смысле. Также это можно сделать с помощью теста Дики-Фуллера. Для этого введём следующие понятия.

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.

Авторегрессионная (AR-) модель - модель временных рядов, в которой значения вре- менного ряда в данный момент линейно зависят от предыдущих значений этого же ряда. Авторегрессионный процесс порядка p (AR(p)-процесс) определяется следующим образом

$$X_t = c + \sum_{i=1}^{p} a_i X_{t-i} + \varepsilon_t,$$

где $a_1, \ldots, a_p$ — параметры модели (коэффициенты авторегрессии), c — постоянная (часто для упрощения предполагается равной нулю), а ε_t — белый шум. Процессом **белого шума** называют стационарный случайный процесс X_t, $t = 0, \pm1, \pm2 \ldots$, для которого верно:

$$(X_t) = 0$$

$$\mathrm{D}(X_t) = \sigma^2$$

и при t != 0 $\rho(\tau) = 0$

Временной ряд имеет **единичный корень**, или порядок интеграции один, если его первые разности образуют стационарный ряд. Это условие записывается как $y_t \sim I(1)$ если ряд первых разностей $\Delta y_t = y_t - y_{t-1}$ является стационарным $\Delta y_t \ I(0)$. При помощи этого теста проверяют значение коэффициента a в авторегрессионном уравнении первого порядка AR(1)

$$y_t = a \cdot y_{t-1} + \varepsilon_t,$$

где y_t - временной ряд, а ε - белый шум. Если $a = 1$, то процесс имеет единичный корень, тогда ряд y_t не стационарен, является **интегрированным временным рядом** первого порядка - $I(1)$. Если $|a| < 1$, то ряд стационарный - $I(0)$. Приведенное авторегрессионное уравнение AR(1) можно переписать в виде:

$$\Delta y_t = b \cdot y_{t-1} + \varepsilon_t,$$

где $b = a - 1$, а Δ — оператор разности первого порядка $\Delta y_t = y_t - y_{t-1}$. Поэтому проверка гипотезы о единичном корне в данном представлении означает проверку нулевой гипотезы о равенстве нулю коэффициента b (существует единичный корень, ряд нестационарный). **Статистика теста (DF-статистика)** — это обычная t-статистика для проверки значимости коэффициентов линейной регрессии. Однако,

распределение данной статистики отличает- ся от классического распределения *t*-статистики. Распределение DF-статистики называется **распределением Дики — Фуллера**.

Существует три версии теста (тестовых регрессий):

- Без константы и тренда: $\Delta y_t = b \cdot y_{t-1} + \varepsilon_t$
- С константой, но без тренда: $\Delta y_t = b_0 + b \cdot y_{t-1} + \varepsilon_t$
- С константой и линейным трендом: $\Delta y_t = b_0 + b_1 \cdot t + b \cdot y_{t-1} + \varepsilon_t$

Для каждой из трёх тестовых регрессий существуют свои критические значения DF-статистики, которые берутся из специальной таблицы Дики — Фуллера (МакКиннона). Если значение статистики лежит левее критического значения (критические значения — отрицательные) при данном уровне значимости, то нулевая гипотеза о единичном корне отклоняется и процесс признается стационарным (в смысле данного теста). В противном случае гипотеза не отвергается и процесс может содержать единичные корни, то есть быть нестационарным (интегрированным) временным рядом.

Модель с распределённым лагом — это модель временного ряда, в которой в урав- нение регрессии включено как текущее значение объясняющей переменной, так и значения этой переменной в предыдущих периодах.

Если в тестовые регрессии добавить лаги первых разностей временного ряда, то распределение DF-статистики (а значит, критические значения) не изменится. Такой тест называют **расширенным тестом Дики — Фуллера (ADF)**. Необходимость включения лагов пер- вых разностей связана с тем, что процесс может быть авторегрессией не первого, а более высокого порядка. Рассмотрим на примере модели AR(2):

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + \varepsilon_t.$$

Данную модель можно представить в виде:

$$\Delta y_t = (a_1 + a_2 - 1) y_{t-1} - a_2 \Delta y_{t-1} + \varepsilon_t.$$

Если временной ряд имеет один единичный корень, то первые разности по определению ста- ционарны. А поскольку y_{t-1} по предположению нестационарен, то если коэффициент при нём не равен нулю, уравнение противоречиво. Таким образом, из предположения об интегри- рованности первого порядка для такого ряда следует, что $a_1 + a_2 - 1 = 0$. Таким образом, для проверки наличия единичных корней в данной модели следует провести стандартный DF-тест для коэффициента при y_{t-1}, причем в тестовую регрессию должен быть добавлен лаг первой разности зависимой переменной.

Приступим к разложению ряда на составляющую тренда, сезонную компоненту и остав- шуюся нерегулярную составляющую. Этот процесс называется **сезонной декомпозицией**.

Общая модель. Функцию исходного ряда можно разложить на следующие компоненты:

- x(t) – **тренд**, устойчивая долговременная тенденция изменения значений временного ряда, закономерно изменяющаяся во времени;

- s(t) – **сезонная** составляющая, периодически повторяющаяся компонента временного ряда, на которую влияют погодные условия, социальные привычки, религиозные тра- диции и прочее;

- z(t) – **остаток** – величина, показывающая нерегулярную (не описываемую трендом или сезонностью) составляющую исходного ряда в определённом временном интервале.

Конкретные функциональные взаимосвязи между этими компонентами могут иметь самый разный вид. Однако можно выделить два основных способа, с помощью которых они могут взаимодействовать:
Аддитивная модель: $y(t) = x(t) + s(t) + z(t)$
Мультипликативная модель: $y(t) = x(t) * s(t) * z(t)$
Здесь $y(t)$ обозначает значение временного ряда в момент времени t. Процесс построения модели включает в себя следующие шаги:

- Выравнивание исходного ряда **методом скользящей средней**.

- Расчет значений сезонной компоненты s(t).

- Устранение сезонной компоненты из исходных уровней ряда и получение выровненных данных $y(t) - s(t) = x(t) + z(t)$ в аддитивной или $y(t) / s(t) = x(t) * z(t)$ в мультипликативной
модели.

- Аналитическое выравнивание уровней $x(t) + z(t)$ или $x(t)*z(t)$ и расчет значений $x(t)$ с использованием полученного уравнения тренда.

- Расчет полученных по модели значений $x(t) + z(t)$ или $x(t) * z(t)$.

- Расчет абсолютных и/или относительных ошибок. Если из временного ряда удалить тренд и сезонную составляющую, то останется нерегулярная компонента, так называ- емая, ошибка. Если полученные значения ошибок не содержат автокорреляции, ими можно заменить исходные уровни ряда и в дальнейшем использовать временной ряд ошибок для анализа взаимосвязи исходного ряда и других временных рядов.

Модель скользящего среднего q-го порядка $MA(q)$ — модель временного ряда вида:

$$X_t = \sum_{j=0}^{q} b_j \varepsilon_{t-j},$$

где ε_t — белый шум, b_j — параметры модели (b_0 можно считать равным 1 без ограничения общности).

Также в модель иногда добавляют константу. Тем не менее, поскольку чаще всего модели скользящего среднего используются для моделирования случайных ошибок временных рядов, то константу можно считать параметром основной модели.
Процесс белого шума формально можно считать процессом скользящего среднего нуле- вого порядка — $MA(0)$.
Чаще всего на практике используют процесс скользящего среднего первого порядка MA(1):

$$X_t = \varepsilon_t + b\varepsilon_{t-1}$$

Метод скользящих средних

Прежде, чем рассчитывать сезонную компоненту, исходный временной ряд необходимо выровнять методом скользящих средних;
Вычисляя скользящее среднее для временного ряда, **интервал сглаживания** (ширина окна) берется равным периоду сезонности. Если период сезонности — четное число, можно в случае простого скользящего среднего, провести процедуру центрирования, которая заключается в повторном скольжении с шагом, равным двум. Число уровней сглаженного ряда будет меньше на величину шага скользящей средней.
После определения скользящих средних вся сезонная (т.е. внутри сезона) изменчивость будет исключена и поэтому разность (в случае аддитивной модели) или отношение (для мультипликативной модели) между
наблюдаемым Y_i и сглаженным рядом $\hat{Y}_t$ будет выделять сезонную составляющую плюс нерегулярную компоненту.
Таким образом, результатом процедуры сглаживания будет временной ряд выровненных значений $\hat{Y}_t$, не содержащий сезонной компоненты. То есть: ряд скользящих средних вычи- тается из наблюдаемого ряда (в аддитивной модели) или же значения наблюдаемого ряда делятся на значения скользящих средних (в мультипликативной модели).

Модель ARIMA

Если ряд после взятия d последовательных разностей приводится к стационарному, то назовем этот ряд ARIMA(p, d, q). ARIMA - процесс авторегрессии - интегрированного сколь- зящего среднего. При этом p - параметр AR-части, d - степень интеграции, и q - это параметр MA-части. В операторном виде ARIMA (p, d, q) записывается как:

$$\alpha_p(L)\Delta^d x_t = \beta_q(L)\varepsilon_t$$

Этот процесс нестационарный, потому что здесь не выполняется условие, что все корни характеристического уравнения по модулю меньше единицы. Но, если обозначить $(1 - L)^d x_t = y_t$, то y_t - стационарный процесс. Задачу построения модели типа ARIMA по реализации случайного процесса можно разбить на несколько этапов.

I этап

1)Установить порядок интеграции d, то есть добиться стационарности ряда, взяв достаточное количество последовательных разностей.

2)После этого мы получаем временной ряд Y к которому нужно подобрать уже ARMA(p,q). Исходя из поведения автокорреляционной и частной автокорреляционной функций, установить параметры p и q.

I этап принято называть идентификацией модели ARIMA(p,d,q). Это всего лишь определение величин p, d, q, но именно в такой последовательности: сначала d, а потом p и q.

II этап Оценка коэффициентов $\alpha_1, \alpha_2, \ldots, \alpha_p, \beta_1, \beta_2, \ldots, \beta_q$ при условии, что мы уже знаем p и q.

III этап По остаткам осуществляется тестирование или диагностика построенной модели.

IV этап Использование модели, в основном, для прогнозирования будущих значений временного ряда. Если исследуемый ряд нестационарный, то его автокорреляционная функция не будет убывать. Если ряд стационарен, то мы знаем, что, начиная с какого-то номера, теоретические автокорреляции будут убывать. Поэтому можно рассчитать их оценки - выборочные автокорреляции, посмотреть, убывают они или нет. Если ряд окажется стационарным, перейти к определению параметров p и q. Если нет, то надо построить ряд первых разностей и проверить на стационарность его. Сначала выбираем значение порядков дифференцирования d так, чтобы ряд стал стационарным. Выбор параметра q проводим с помощью графика функции автокорреляции. q - номер последнего несезонного лага, при котором автокорреляция значима. Параметр p выбираем по графику функции частичной автокорреляции.
Частичная автокорреляция - автокорреляция после снятия регрессии предыдущего порядка. p задается как номер последнего несезонного лага, при котором частичная автокорреляция значима. Далее будем перебирать разные модели в классе ARIMA с выбранным значением параметра d, начиная с тех начальных приближений, которые мы получили из автокорреляционных функций. Сравнивать разные модели будем по **информационному критерию Акаике (AIC)**:

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

где k — число параметров в статистической модели, и L — максимизированное значение функции правдоподобия модели. Оптимальной по критерию Акаике будет модель, у которой значение этого критерия самое маленькое из всех возможных, потому что такая модель будет достаточно хорошо описывать данные, с одной стороны, и содержать не слишком большое количество параметров, с другой.

2 Программный подход к решению задачи

Первым шагом мы считываем данные из таблиц training.xlsx, testing.xlsx с помощью метода read_excel() из библиотеки Pandas. Затем проводим необходимую агрегацию данных. Именно удаляем столбик ‘Date’, седлав его индексом датасета.

Проводим визуальную оценку. Строим график с помощью библиотеки seaborn, на котором отображен сам ряд и различные скользящие. Для расчета скользящего делим ряд на временные интервалы, равные window и считаем для каждого среднее и стандартное отклонение.

Затем проводим расширенный тест Дики - Фуллера. Находим, что у ряда есть единичные корни по значениям p-value и adf, значит он нестационарен. Проводим тест Дики-Фуллера для datadiff, получаем, что он стационарен, следовательно, исходный ряд - интегрированный ряд первого порядка.

Далее раскладываем ряд на тренд, сезональность и отсатки в соответствии с аддитивной и мультипликативной моделями. С помощью seaborn рисуем их графики. При помощи «нашего» теста Дики - Фуллера проверяем все компоненты на стационарность.

С помощью функции plot_acf рисуем функции автокорреляции и частичной автокорреляции. Используя функцию ARIMA из библиотеки statsmodels, создаём model.

Первый параметр функции ARIMA - значения таблицы, а второй-тройка параметров p,d,q. Параметры p и q находим с помощью графиков функции частичной и полной автокорреляции. По коррелограмме ACF определяем q = количество автокорреляционных коэффициен- тов сильно отличных от нуля в модели MA - оно равно 2. По коррелограмме PACF можно определить p = максимальный номер коэффициента, сильно отличного от нуля в модели AR
- оно равно 2. Используем функции plot_acf, autocorrelation_plot нарисуем лаги ряда и автокорреляцию. Берем тестовые данные из testing.xlsx и прогнозируем значения с помощью полученных моделей. Функция r2_score библиотеки sclearn пвозвращает араметр r2, характеризующий точность предсказаний.

Далее используем метод predict и получаем предсказание для продолжения таблицы, нарисуем его. Теперь точно так же сделаем другое предсказание, с помощью метода поиска по сетке или оптимизации гиперпараметров, были найдены 3 лучшие модели по критерию Акаике, и визуализированы, а также посчитаны их r2_score.

3 Выводы

Для данного нам временного ряда мы построили графики скользящей статистики и по нем определили, что ряд нестационарен, так как его мат.ожидание и дисперсия зависят от време- ни. Нестационарность данного ряда подтверждает и тест Дики-Фуллера. Значение статистики лежит правее критического значения при уровне значимости 5%, поэтому нулевая гипотеза о единичном корне не отклоняется, то есть процесс может содержать единичные корни. До- стоверность статистики больше уровня значимости.

После разложения временных рядов на тренд, сезональность и остаток, мы заметили, что для обоих моделей оригинальный ряд и трендовая составляющая

нестационарны, так как их скользящая статистика зависит от времени. Однако так же для обоих моделей справедлива стационарность сезонной составляющей и остатка, их скользящая статистика не зависит от времени.

Наилучшее значение *r2score* = 1. Если оно будет отклоняться вправо или влево, то велика вероятность неточности построения модели. Более удачной моделью будет модель с меньшим AIC.

PEP8 был соблюден автокорректором кода (autoformatted).

4 Работу выполнили

- София Серебрякова - написание кода (пункт 1, 2)
- Чингиз Турсуналиев - написание кода (пункт 3)
- Сарджаев Меред - написание README