Задание №2

Работа выполнена командой №4: задача состояла в проведении анализа некоторого временного ряда и предсказания значений для последующих месяцев. В рамках данного задания мы:

1. научились правильно считывать данные и их визуализировать;
2. поняли как определять свойства временных рядов;
3. познакомились с различными моделями и оценили качество полученных моделей.

Цель работы заключалась в том, чтобы познакомиться с библиотеками pandas (для работы с данными) и statsmodels (для работы с различными статистическими моделями, в том числе и временными рядами).

Содержание

1 Основная теория

1.1 Используемые понятия

В данной работе использовались следующие термины:

Временной ряд — совокупность наблюдений определенной величины (например, экономической) в различные моменты времени. Они задаются на фиксированном временном промежутке. Начало временного промежутка примем за 0, конец за T. Мы будем обозначать тестируемый временной ряд символом Y. Временной ряд имеет хотя бы один единичный корень (или порядок интеграции один), если его первые разности образуют стационарный ряд.

Временной ряд называется **интегрированным порядка k**, если разности ряда k-го порядка являются стационарными, а разности меньшего порядка (и сам временной ряд) не являются стационарными рядами.

Другими словами, ряд называется интегрируемым порядка k, если его разности порядка $k-1$ включительно нестационарны, а k-я разность — стационарна.

Если ряд y является интегрированным порядка k, то данный факт обозначают $Y(k) \sim I(k)$.

Временной ряд называется **строго стационарным** (стационарным в узком смысле), если сдвиг во времени не меняет ни одной из функций плотности распределения. Следствием из определения будут независимость матожидания (обозначение $E()$) и дисперсии (обозначение $D()$) от времени. Слабая стационарность (стационарность в широком смысле) — случайный процесс, у которого математическое ожидание и дисперсия существуют и не зависят от времени, а автокорреляционная (автоковариационная) функция зависит только от разности значений $(t_1 - t_2)$.

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

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

Ковариация (корреляционный момент, ковариационный момент) — в теории вероятностей и математической статистике мера линейной зависимости двух случайных величин.

Простое скользящее среднее, или арифметическое скользящее среднее (англ. simple moving average, англ. SMA) численно равно среднему арифметическому значений исходной функции за установленный период.

Стандартное отклонение — показывает, на сколько в среднем отклонился ряд от средней вариации ряда (от среднего арифметического, в нашем случае). Скользящая средняя вместе со стандартным отклонением составляют **скользящие статистики**.

Преобразование Бокса-Кокса применяется для приближения имеющегося распределения к нормальному и выглядит следующим образом (логарифмирование – это частный случай трансформации Бокса-Кокса):

$$y_{new} = \begin{cases} \frac{y^p - 1}{p} & , p \neq 0 \\ \ln y & , p = 0 \end{cases}$$

1.2 Тест Дики-Фуллера

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.

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 $\Delta\, y_t$ является стационарным $y_t \sim I(0)$.

При помощи теста Дики-Фуллера проверяют значение коэффициента a в авторегрессионном уравнении первого порядка $AR(1)$.

Если a = 1, то процесс имеет единичный корень, в этом случае ряд y_t не стационарен, является интегрированным временным рядом первого порядка — $I(1)$. Если |a| < 1, то ряд стационарный — $I(0)$. Значение |a| > 1 не свойственно для временных рядов, которые встречаются в реальной жизни - для него требуется более сложный анализ.

Приведенное авторегрессионное уравнение AR(1) можно переписать в виде: $\Delta\, y_t = by_{t-1} + \epsilon_t$, где b = a - 1, а Δ - оператор разности первого порядка.

Основная (нулевая) гипотеза: H_0: b = 0 — нестационарный процесс.

Альтернативная гипотеза H_1: b < 0 — стационарный процесс.

Для того чтобы проверить временной ряд y на порядок интегрируемости, следует рассчитать значение t-статистики Стьюдента для параметра b и сравнить его с верхним и нижним критическими значениями DF-статистики из таблицы теста Дики-Фуллера. Если для n наблюдений значение расчетной t-статистики меньше, чем нижнее критическое значение, нулевую гипотезу H_0 отклоняют и принимают альтернативную о стационарности процесса y_t. В противном случае, нулевая гипотеза не может быть отклонена. Если нулевая гипотеза не отклоняется, можно лишь утверждать, что процесс y_t нестационарен, т.е. либо он интегрируем более высокого порядка, либо неинтегрируем вообще.

1.3 Тренд, сезонность, остаток

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.

Сезонность — периодические колебания, наблюдаемые на временных рядах.

Остаток — величина, показывающая нерегулярную (не описываемую трендом и сезонностью) составляющую исходного ряда в определенном временном интервале. Фактически, остатком называется разница между предсказанным и наблюдаемым значением.

Приступим к разложению ряда на составляющую тренда, сезонную компоненту и оставшуюся нерегулярную составляющую. Этот процесс называется сезонной декомпозицией.

Функцию исходного ряда можно разложить на следующие компоненты:

- T – тренд, устойчивая долговременная тенденция изменения значений временного ряда, закономерно изменяющаяся во времени;
- S – сезонная составляющая, периодически повторяющаяся компонента временного ряда, на которую влияют погодные условия, социальные привычки, религиозные традиции и прочее;
- E – остаток – величина, показывающая нерегулярную (не описываемую трендом или сезонностью) составляющую исходного ряда в определённом временном интервале.

Общий вид аддитивной модели: Y = T + S + E;
Общий вид мультипликативной модели: Y = T * S * E;

2 Модель ARIMA и дополнительные понятия

2.1 Интегрированная модель авторегрессии ARIMA

$ARIMA$ (англ. autoregressive integrated moving average) — интегрированная модель авторегрессии — скользящего среднего — модель и методология анализа временных рядов.

Это расширение моделей $ARMA$ для нестационарных временных рядов.

Модель $ARIMA(p, d, q)$ для нестационарного временного ряда X_t имеет вид: $\Delta^d X_t = c + \sum_{i=1}^{p} a_i \Delta^d X_{t-i} + \sum_{j=1}^{q} b_j \Delta^d \epsilon_{t-j} + \epsilon_t$, где ϵ_t - стационарный временной ряд, c, a_i, b_j - параметры модели, а 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 порядка и т.д.)

$ARMA$ (англ. autoregressive moving average) — математическая модель, используемая для анализа и прогнозирования стационарных временных рядов. Объединяет 2 более простые модели: авторегрессии (AR) и скользящего среднего (MA).

AR (англ. autoregressive model) — авторегрессионная модель временных рядов, в которой значение временного ряда в данный момент зависит от предыдущих значений этого же ряда.

MA (англ. moving average model) — модель скользящего среднего, в которой моделируемый уровень временного ряда можно представить как линейную функцию прошлых ошибок, т.е. разностей между прошлыми фактическими и теоретическими уровнями.

Алгоритм построения модели $ARMA$ заключается в поиске коэффициентов p, q — порядков для моделей $AR(p)$ и $MA(q)$. Это позволит построить функцию автокорреляции и функцию частичной автокорреляции.

Таким образом, построение $ARIMA$ зависит от 3 параметров: $ARIMA(p, d, q)$, где p — порядок $AR(p)$, d — порядок интегрированности, q — порядок $MA(q)$.

2.2 Информационный критерий Акаике AIC

Информационный критерий — применяемая в эконометрике мера относительного качества эконометрических моделей, учитывающая степень "подгонки" модели под данные с корректировкой на используемое количество оцениваемых параметров, т.е. критерии основаны на некотором компромиссе между точностью и сложностью модели. Критерии различаются тем, как они обеспечивают этот баланс. Информационные модели используются исключительно для сравнения моделей между собой, без содержательной интерпретации значений этих критериев. Обычно чем меньше значения критериев, тем выше относительное качество модели.

AIC (an information criterion) — информационный критерий Акаике - критерий, применяющийся исключительно для выбора из нескольких статистических моделей. Связан с концепцией информационной энтропии и расстоянии Кульбака-Лейблера, на основе которой был разработан этот критерий.

В общем случае AIC: $AIC = 2k - 2\ln L$, где k — число параметров в статистической модели, и L — максимизированное значение функции правдоподобия модели.

2.3 Коэффициент детерминации R2-score

Коэффициент детерминации R^2 — это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными. Истинный коэффициент детерминации модели зависимости случайной величины y от факторов x определяется следующим образом: $R^2 = 1 - \frac{V(y|x)}{V(y)} = 1 - \frac{\sigma^2}{\sigma_y^2}$, где $V(y|x) = \sigma^2$ —

условная (по факторам x) дисперсия зависимой переменной (дисперсия случайной ошибки модели).

Коэффициент детерминации для модели с константой принимает значения от 0 до 1. Чем ближе значение коэффициента к 1, тем сильнее зависимость. При оценке регрессионных моделей это интерпретируется как соответствие модели данным. Для приемлемых моделей предполагается, что коэффициент детерминации должен быть хотя бы не меньше 50% (в этом случае коэффициент множественной корреляции превышает по модулю 70%). Модели с коэффициентом детерминации выше 80% можно признать достаточно хорошими (коэффициент корреляции превышает 90%). Значение коэффициента детерминации 1 означает функциональную зависимость между переменными.

3 Результаты и выводы

В данном задании информационный критерий Акаике (AIC) оценивает относительный объем информации, потерянной данной моделью. Чем меньше, тем лучше, следовательно в нашем случае модель ARIMA с параметрами (12,1,4) является лучшей. По итогам критерия Акаике и наглядной интерпретации работы трёх моделей нами была получена оптимальная модель — третья. Её параметры (p, d, q) = 12, 1, 4.

С помощью этого задания были получены навыки и инструменты для работы с временными рядами. Подобные задания имеют ярко выраженный прикладной характер.

4 Требуемое ПО

4.1 Необходимые библиотеки

- pandas — программная библиотека для обработки и анализа данных. Мы используем её для работы с временными рядами.
- statsmodels — представляет собой пакет, который позволяет исследовать данные, оценивать статистические модели и выполнять статистические тесты. Мы используем его для проведения теста Дики-Фуллера, а так же для построения модели ARIMA.
- sklearn.metrics — пакет для машинного обучения. Мы используем его для оценки метрики r2 score.
- matplotlib — пакет для визуализации данных двумерной графикой. Мы используем его для отрисовки графиков.

4.2 Необходимые программы

- Python 3.6
- Anaconda3
- Jupyter

5 Список участников и их вклад

- Максим Приходько — разложение временного ряда на тренд, сезональность, остаток, их визуализация и оценка стационарности рядов

- Антон Тарасов — написание теста Дики-Фуллера, анализ интегрированности временного ряда порядка k и применение модели ARIMA
- Кристина Светличная — написание функции с подсчётом r2-score для каждой модели, написание readme и верстка данного файла в LaTeX
- Александра Коробельникова — визуализация временных рядов и скользящей статистики
- Екатерина Маслихина — отбор модели по критерию Акаике

Также общими усилиями была проведена проверка кода на PEP8.