Vol. 3, No. 1, 2021

https://doi.org/10.23939/cds2021.01.061

УДК 62-522.7

УДК 004.9

І. Фармага, В. Гадомський,
Національний університет “Львівська політехніка”

СИСТЕМА ПРОВЕДЕННЯ АНАЛІЗУ, ДОСЛІДЖЕННЯ ТА ПЕРЕДБАЧЕННЯ ПОДІЙ У ПОСЛІДОВНОСТЯХ ДАНИХ ДИСКРЕТНОГО ЧАСУ

У роботі розроблено програмне забезпечення для передбачення за допомогою часових рядів за допомогою мови програмування Python. Для розроблення системи було використано модель SARIMA.

Ключові слова – Python, SARIMA модель, часові ряди, передбачення з використанням часових рядів, AWS, AWS Lambda

Вступ

У світі існує надзвичайно багато процесів, які залежать від часу. Через це, часові передбачення, які реалізовані з використанням часових рядів, є актуальними на сьогодні. Всі ми могли чути про розмови людей, як можуть змінюватися ціни предметів з плином часу. Ціни то збільшуються, то зменшуються, а ці предмети можуть бути чим завгодно, нехай то продукти харчування, паливо, чи ювелірна продукція. Відсоток кредитів та депозитів також різниться з плином часу. На скільки ці всі дані корисні, і які вони взагалі? Всі, описані в цій роботі, дані відповідають вимогам часових рядів. Їх можна аналізувати для подальших прогнозів. Часові ряди можуть використовуватися як природою, так і людьми для обміну даними між собою, для описів і візуалізації певних даних. Взагалі, час є фізичною величиною, але елементи, коефіцієнти, параметри ти характеристики часових рядів – це математичні величини. Тому, інтерпретація в реальному світі можлива і для часових рядів.

Актуальність використання часових рядів

Станом на сьогодні, в суспільстві виникає багато різнопланових задач, для розв’язку яких потрібне передбачення у майбутньому. Для прогнозування тенденцій змін, час для нас – це важливий параметр. Наприклад, дуже цікавим було б прогнозування навантаження на метрополітен: яка кількість пасажирів приходиться в день, о котрій годині пік пасажиропотоку і чи змінюється показник у вихідні дні? Ось для таких подій ми можемо використати прогнозування на основі методу часових рядів.

Важливість прогнозування подій за допомогою часових рядів

Передбачення майбутнього – дуже цікава тема, чи не так? Про це мріяв кожен, будучи дитиною, і прогнозування за допомогою часових рядів зможе приблизити нас до здійснення «дитячої» мрії. Прогнозування дає змогу нам "побачити" і досягти успіху у багатьох сферах. Саме цій задачі присвячена дана робота - дослідженням питань про передбачення на основі часових рядів, їх корисність зокрема.

Передбачення за допомогою часових рядів

Поки ми накопичуємо дані, час залишається вирішальним фактором. По суті, в аналізі часових рядів час є важливим елементом даних.

Часовий ряд - це сукупність спостережень за конкретним процесом у певному часовому порядку, де інтервал часу між кожним спостереженням постійний, наприклад, тижні, місяці та роки, у деяких випадках допустимі невеликі відхилення у часових інтервалах. Дані тимчасових рядів корисні для вивчення того як, цікавий для нас актив, цінний папір чи економічна змінна, змінюється з часом.

Аналіз часових рядів

Аналіз часових рядів – це процес аналізу спостережень за точками даних, зібраними за певний період часу, тобто даних часових рядів. При аналізі часових рядів аналітики даних записують спостереження даних із постійними інтервалами для набору періодів часу замість випадкового запису спостережень даних. Швидкість спостереження (тимчасовий інтервал) може становити від мілісекунд до кількох років.

Щоб перевірити, «як змінні змінюються з часом», дані часового ряду описують досліджуване явище у певні моменти часу для аналізу коливань змінних з плином часу. Параметри, що цікавлять, можуть змінюватись в залежності від домену, наприклад:

- значення, зареєстровані науковими приладами за день;
- кількість відвідувань деяких веб-сайтів за день;
- щотижнева вартість акцій на фондовому ринку;
- кількість дощових днів на рік.

Крім того, аналіз часових рядів має справу з великою кількістю точок даних для забезпечення узгодженості та надійності. Величезний обсяг даних відображає відносно великий розмір вибірки, який гарантує, що будь-яка виявлена тенденція чи закономірність не є викидом. Крім того, дані часових рядів можна використовувати для прогнозування майбутніх результатів на основі попередніх даних.

Прогнозування часових рядів

Аналіз часових рядів [1] допомагає компаніям зрозуміти причини коливань тенденцій чи основних закономірностей з плином часом. Використовуючи різні методи візуалізації даних, організації можуть вивчати сезонні тенденції та проводити додаткові дослідження, щоб зрозуміти причини цих тенденцій. Коли організації аналізують дані часових рядів через регулярні проміжки часу, вони використовують прогнозування часових рядів, щоб передбачити майбутні події.

Простіше кажучи, прогнозування часових рядів – це метод прогнозування майбутніх подій шляхом аналізу поведінки чи тенденцій минулих даних з урахуванням припущень про те, що майбутні тенденції матимуть схожість із минулими тенденціями.

При прогнозуванні даних часових рядів мета полягає у тому, щоб передбачити, як спостереження даних буде тривати чи зміниться у майбутньому. Загалом часовий ряд дозволяє проаналізувати основні закономірності, такі як: сезонність, циклічність тенденції та нерегулярність. Аналіз часових рядів використовується в різних галузях. Наприклад, розпізнавання образів, аналіз фондового ринку, прогнозування землетрусів, економічне прогнозування, аналіз перепису тощо.

Закономірності у часових рядах

Часові ряди включають в себе тенденції, цикли та сезонність. На превеликий жаль, виникає багато плутанини між сезонністю і циклами, саме щоб її уникнути, потрібно розібрати кожен з цих факторів.

- Тенденції: коливання кількості даних на певному проміжку часу.
- Сезонність: Зазвичай, сезонність – це відома та фіксована частота. Сезонна закономірність виникає, коли на часовий ряд впливають сезонні фактори, наприклад, пора року або день тижня.
- Цикл: При виникненні коливання даних виникає цикл, але якщо порівнювати з сезонними, то це не фіксована частота.

Вступ до SARIMA. Прогнозування часових рядів мовою програмування Python

Авторегресивне інтегроване змінне середнє або ARIMA - один з методів прогнозування, що найбільш широко використовуються, для одновимірного прогнозування даних часових рядів.

Хоча цей метод може обробляти дані з трендом, він не підтримує часові ряди із сезонною складовою.

Розширення ARIMA, яке підтримує пряме моделювання сезонної складової ряду, називається SARIMA.

Ми відкриємо для себе метод сезонного авторегресійного інтегрованого змінного середнього або SARIMA, метод передбачення часових рядів, що містить тенденції та сезонність.

Що не так з ARIMA?

Авторегресійне інтегроване середнє змінне або ARIMA - це метод прогнозування одновимірних даних часових рядів.

Як випливає з назви, він підтримує елементи як авторегресії, так і змінного середнього. Інтегрований елемент відноситься до різниці, даючи змогу методу підтримувати дані часових рядів із трендом.

Проблема з ARIMA у тому, що вона не підтримує сезонні дані. Це тимчасовий ряд з циклом, що повторюється.

ARIMA очікує дані, які або не є сезонними, або в них видалено сезонний компонент, наприклад, з урахуванням сезонних коливань за допомогою таких методів, як сезонна різниця.

Що таке SARIMA?

Сезонне авторегресійне інтегроване середнє змінне, SARIMA (також відомий як Seasonal ARIMA), є додатком до ARIMA, яке підтримує дані одновимірних часових рядів із сезонною складовою.

Він додає три нові гіперпараметри, щоб вказати авторегресію (AR), різницю (I) та ковзаюче середнє (MA) для сезонного компонента ряду, а також додаткові параметри для періодів сезонності.

SARIMA складається методом додавання додаткових членів до ARIMA. Сезонна частина моделі складається з компонентів, які мають походження від несезонних компонентів моделі, але включають зрушення назад сезонного періоду.[2]

Як налаштувати SARIMA?

Для налаштування SARIMA необхідно вибрати гіперпараметри як трендових, так сезонних елементів ряду.

Елементи тренду

Є три елементи тенденції, які вимагають налаштування.

Вони такі самі, як модель ARIMA, а саме:

p: Порядок авторегресії тренду.

d: Порядок різниці трендів.

q: Порядок ковзного середнього тренду.

Сезонні елементи

Необхідно налаштувати чотири сезонні елементи, що не є частиною ARIMA, а саме:

P: Сезонний авторегресійний порядок.

D: Порядок сезонної різниці.

Q: Порядок сезонної змінної середньої.

m: Кількість тимчасових кроків одного сезонного періоду.

У сукупності позначення для моделі SARIMA визначаються як:

SARIMA(p,d,q)(P,D,Q)m

Де вказані спеціально вибрані гіперпараметри для моделі. Наприклад:

SARIMA (3,1,0)(1,1,0)12

Важливо, що параметр m впливає на параметри P, D і Q. Наприклад, значення m, що дорівнює 12 для місячних даних означає річний сезонний цикл.

При P=1 буде використано перше спостереження за сезонним зміщенням у моделі, напр. t-(m*1) або t-12. А P=2, використає два останніх сезонних зміщення спостережень t-(m * 1), t-(m * 2).

Аналогічно, D, що дорівнює 1, обчислює сезонну різницю першого порядку, а Q=1 використовує помилки першого порядку в моделі (наприклад, змінне середнє).

Сезонна модель ARIMA використовує диференцію з лагом, що дорівнює кількості сезонів, щоб усунути додаткові адитивні сезонні ефекти. Як і при диференціації лаг 1 для усунення тенденції, диференціація лаг s вводить змінне середнє. Сезонна модель ARIMA включає терміни авторегресії та змінного середнього з лагом s. [3]

Елементи тенденції можна вибрати шляхом ретельного аналізу графіків ACF та PACF з огляду на кореляції останніх кроків часу (наприклад, 1, 2, 3).

Аналогічно, графіки ACF і PACF можна проаналізувати для визначення значень для сезонної моделі, подивившись на кореляцію на сезонних кроках часу затримки.

Сезонні моделі ARIMA потенційно можуть мати велику кількість параметрів і комбінацій термінів. Тому доцільно випробувати широкий спектр моделей під час пристосування до даних і вибрати найкращу модель за відповідним критерієм. [4]

Крім того, можна використовувати пошук у сітці за тенденцією та сезонними гіперпараметрами.

Алгоритм використання SARIMA в Python:

Метод прогнозування часових рядів SARIMA підтримується в Python через бібліотеку Statsmodels [5].

Щоб використовувати SARIMA, потрібно виконати три кроки:

1. Визначити модель.
2. Підігнати визначену модель.
3. Зробити прогноз за допомогою відповідної моделі.

Алгоритм передбачення за допомогою часових рядів

Як приклад, для демонстрації працездатності та тестування розробленої системи використаємо дані про погоду. Для аналізу та прогнозування було використано інформацію про погоду в місті Київ з 2020 по 2021 рік. Ці дані враховують новітні фактори зміни погоди: сезонність, тенденції та цикли. Не зважаючи на те, що обрані нами дані є представленням погоди, використати розроблену систему можна і для даних пов’язаних з продажами, навантаженням, тощо. Ми можемо це зробити, адже передбачення насамперед залежать від часу, від часових рядів. На основі яких ми і будуємо передбачення.

Щоб отримати більше інформації про вхідні дані, бібліотека Python Pandas надає функцію «опис», щоб показати кількість, середнє, стандартне відхилення, мінімальне/максимальне значення та квантилі нашого набору даних. Деякі відмінні закономірності з’являються, коли ми візуалізуємо дані (рис. 1). Часовий ряд має сезонність, наприклад температура завжди низька на початку року і висока в середині року.

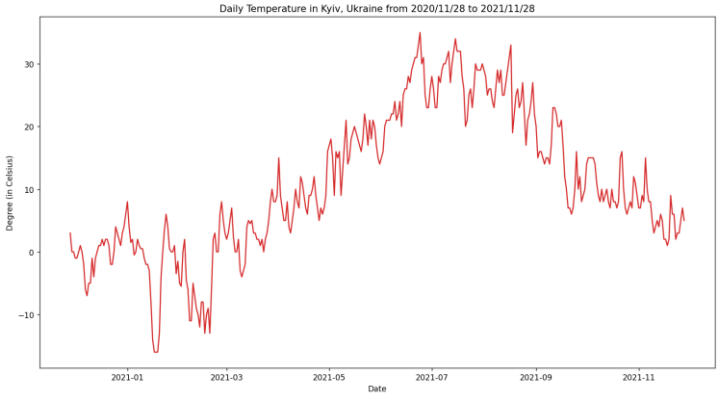

Рис.1 Візуалізація вхідних даних

Наведений вище графік трішки «зашумлений», оскільки містить в собі всі добові температури. Однак, уважно переглянувши точки даних, ми бачимо, що є лише незначна зміна температури між поточною датою та наступною датою.

Як ми можемо побудувати графік, що містить лише зміну температури, що перевищує добову, і, отже, виглядає акуратніше?

Ось інструмент: змінне середнє (рис. 2) в основному використовується з даними часових рядів, щоб фіксувати короткострокові коливання з фокусуванням на більш тривалих тенденціях.

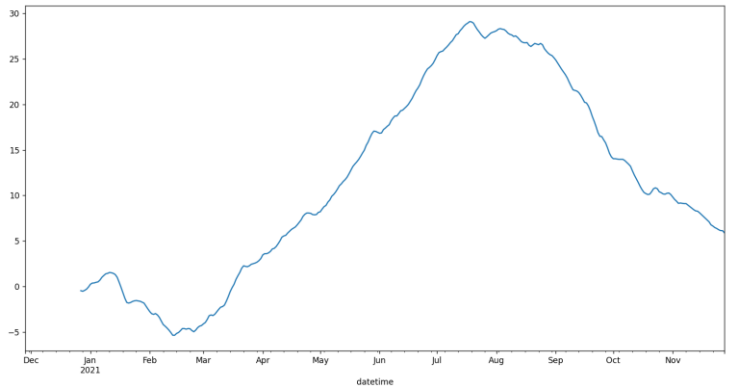

Рис.2 Змінне середнє часового ряду

Ми також можемо візуалізувати наші дані за допомогою методу, який називається декомпозицією часових рядів, який дозволяє нам розкласти наш часовий ряд на три окремі компоненти: тенденцію, сезонність і шум.

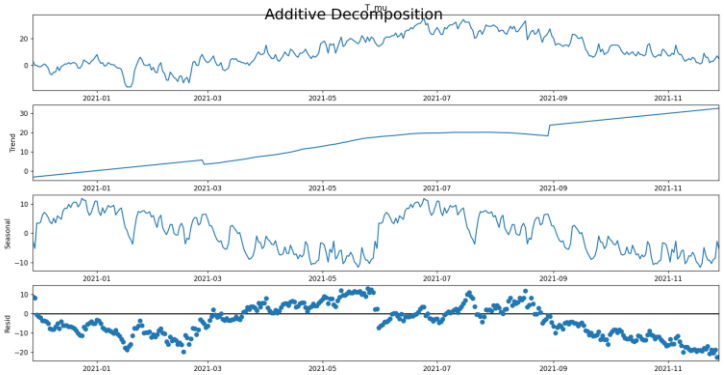

Рис.3 Графік тенденції, сезонності та шуму.

З наведеного вище графіка (рис. 3) чітко видно, що температура нестабільна, а також очевидна її сезонність.

Проаналізувавши ряд, тепер ми можемо побудувати реальні та прогнозовані значення середньої добової температури (рис. 4), щоб оцінити, наскільки добре ми зробили. Зверніть увагу, як ми збільшили кінець часового ряду, розрізавши індекс дати.

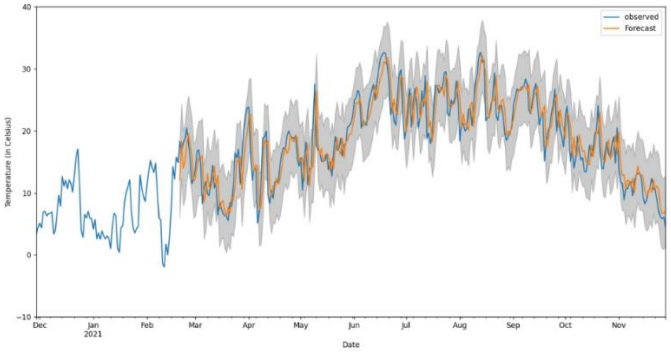

Рис.4 Графік прогнозу.

Загалом, наші прогнози дуже добре відповідають істинним значенням, показуючи сезонний цикл тривалістю 365 днів. Також корисно кількісно оцінити точність наших прогнозів. Ми будемо використовувати MSE (середня квадратична помилка), в якій для кожного передбаченого значення ми обчислюємо його відстань до справжнього значення та зводимо результат у квадрат.

У області прогнозу погоди похибка прогнозу в 2,97 градусів здається багатообіцяючою та достатньою, оскільки існує багато інших факторів, які сприяють зміні температури, включаючи, крім всього іншого, швидкість вітру, тиск повітря тощо.

Розгортання системи передбачення у хмарному середовищі AWS

Головною ціллю диплому було розробити систему передбачення за допомогою часових. Відповідно основна логіка (передбачення) – має виконуватись в окремому компоненті. Для цього було обрано сервіс AWS Lambda. AWS Lambda – це сервіс для обчислення, який може запускати програмний код майже для будь яких типів програм або серверних служб. І все це може відбуватися без необхідного адміністрування. AWS Lambda дозволяє виконати всі процедури адміністрування за користувача. Нам всього лиш потрібно надати свій код на одній з мов, яку підтримує Lambda.

Створивши функцію, виникає проблема викликати її. Наразі є можливість лише зробити це вручну. На допомогу нам можуть “прийти” тригери. AWS Lambda виконується при кожному запуску виконання у відповідь на тригер повідомлення про подію. Відповідно, створивши тригер, ми створимо точку доступу до самої функції. Для цього було використано сервіс Amazon API Gateway. Від допомагає створити інтерфейс доступу до AWS Lambda функції. Також варто відзначити, що було додано приватний ключ, яким функція є захищена. Таким чином лише авторизовані користувачі зможуть виконувати її.

Висновки

Для передбачення за допомогою часових рядів було розроблено систему за допомогою мови програмування Python. Система базується на моделі SARIMA, яку було детально описано. Також було здійснено аналіз та візуалізацію вхідних даних для наочності виконання алгоритму.

Література

1. Часовий ряд [Електронний ресурс] - Режим доступу до ресурсу https://uk.wikipedia.org/wiki/Часовий_ряд

2. R. J. Hyndman, G. Athanasopoulos. Forecasting: principles and practice. Otexts.com. 2013. -p. 242.

3. P. S.P Cowpertwait, A. V. Metcalfe. Intoductory Time Series with R (Use R!). Springer. 2009. -p. 142.

4. P. S.P Cowpertwait, A. V. Metcalfe. Intoductory Time Series with R (Use R!). Springer. 2009. -p. 143-144

5. Statsmodels [Електронний ресурс] - Режим доступу до ресурсу: https://www.statsmodels.org/stable/index.html

I. Farmaha, V. Hadomskyi
Lviv Polytechnic National University

SYSTEM FOR ANALYSIS, RESEARCH AND FORECAST OF EVENTS IN DISCRETE TIME DATA SEQUENCES

This paper is devoted develop software for time series forecasting using Python programming language. SARIMA model was used to develop the system.

Keywords – Python, SARIMA model, time series, time series forecast, AWS, AWS Lambda.