Modelul SARIMAX pentru previziune, utilizând Python

Achim Andrei-Cătălin

ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI
Facultatea: Contabilitate și Informatică de Gestiune

Profesor coordonator: Prof. univ. dr. Mihai Florin

Rezumat: Modelul SARIMAX (S-Seasonal, AR-Auto-Regressive, IMA-Integrated Moving Average, X-eXogenous factor) apartine familiei modelelor de previziune de tipul serie-timp precum AR, ARIMA, Auto ARIMA, SARIMA etc. Fiecare model poate fi diferențiat prin capacitatea sa de previziune. Alegerea modelului se face în funcție de scop, plecând de la datele stocate care se schimbă în functie de timp si formeaza un set de date de tipul serie-timp. Previziunea se poate face prin metoda AR auto-regresiva si pe masura ce adaugam factori sezoniei S si factori exogeni X ajungem la modele mai complexe. Implementarea modelului in limbajul informatic se realizeaza optim folosind limbajul python prin intermediul librariilor de care acesta dispune. Astfel cel mai potrivit mediu de utilizare al modelului este Google Colab datorita sinergiei dintre acestea.Ideea principala a modelelor de previziune este de a creea o prognoza bazata pe datele din trecut intre care exista o legatura de tip tendinta, ciclica, iregulata si/sau sezoniera. Modelul SARIMAX este unl din cele mai complexe, deoarece cuprinde toate variantele de caracterizare a unui set de date.

Cuvinte-cheie: SARIMAX, ARIMA, Python, Colab.

Abstract: The SARIMAX model, also known as the S-Seasonal, AR-Auto-Regressive, IMA-Integrated Moving Average, and X-eXogenous factor model, is a member of the family of time-series forecasting models that also includes the AR, ARIMA, Auto ARIMA, SARIMA, and others. The ability of each model to predict can be used to distinguish them. Starting with the data that is stored and changes over time to create a data set of the time-series type, the model is chosen based on the goal. The auto-regressive AR technique may be used to forecast, and when we include exogenous parameters X and seasonal components S, our models become more complex. Python and the libraries that are accessible to it provide for the best model implementation in the computer language. Due to their interdependence, Google Colab is the most suitable environment for using the concept. Creating a forecast based on historical data with a trend, cyclical, irregular, and/or seasonal relationship is the major goal of forecasting models. Due to the fact that it incorporates all possible characterizations of a data set, the SARIMAX model is one of the most complicated.

Keywords: SARIMAX, ARIMA, Python, Colab.

*Obiectele inserate în lucrare fac parte din din proiecul meu și sunt create de mine pe baza unui set de date descărcat[1]

1. Obiectivul lucrării

Obiectivul îl reprezintă dezvoltarea științifică și a literaturii de specialitate din domeiul statistic al modelelor de previziune din familia SARIMAX în limba română. Totodată, aceasta include și posibilitățile de integrare a modelelor statistice cu instrumente informatice colaborative și de programare în limbajul Python. Această lucrare este, în același timp, o îndrumare spre aprofundare, un punct de pornire pentru pasionații de statistică și de informatică.

2. Modalitatea de cercetare

Cercetarea pentru modelele de previziune, în special SARIMAX și ARIMA, am efectuat-o, în mare parte, prin articole și secvențe din cărți de specialitate din limba engleză, (deoarece literatura de specialitate în acest domeniu s-a dezvoltat mai mult în partea vestică); pentru cercetarea pentru Python am studiat atât cărți de specialitate, dar și, mai ales, articole care descriu librăriile și instrucțiunile folosite, iar Google Colab este un API online bazat pe cloud, despre care am aflat online.

3. Modelele de previziune

3.1. ARIMA

Pentru dezvoltarea prognozelor bazate pe date din trecut, modelul ARIMA (Autoregressive Integrated Moving Average) este o metodă populară de analiză a seriilor de timp. Combină trei elemente esențiale: diferențierea, mediile mobile și autoregresia (I).

Seria temporală este modelată ca o funcție liniară a propriilor valori anterioare în componenta autoregresivă. Această parte ia în considerare modul în care observațiile curente se raportează la observațiile anterioare din aceeași serie, ordinea relației fiind determinată de parametrul autoregresiv. *Ordinea este determinată de parametrul mediei mobile*[2], iar componenta mediei mobile simulează dependența observațiilor curente de erorile istorice. Având în vedere diferența dintre observațiile succesive, componenta de diferențiere elimină tendința sau sezonalitatea din serie, creând o serie temporală staționară.

Modelul ARIMA este utilizat, de obicei, pentru a prognoza variabile cu modele și tendințe aleatoare, și este adesea potrivit pentru procese cu autocorelare sau heteroscedasticitate. Modelul ARIMA este, de asemenea, util atunci când datele arată o tendință sau o componentă sezonieră.

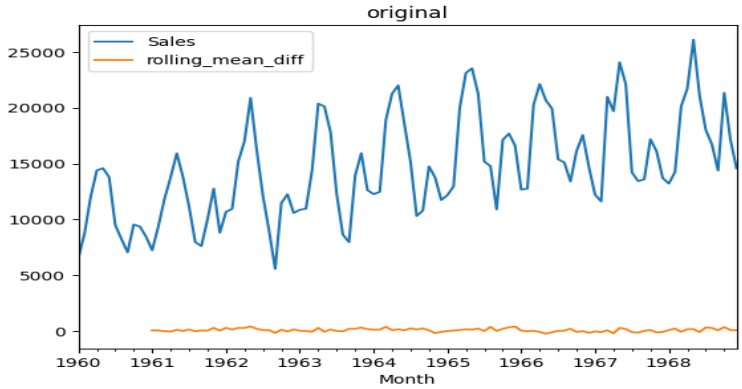


Figura 1. Diferența dintre un set de date original (linia albastră), care este sezonier nestaționar și după ce este trasformat (linia portocalie) într-un set de date staționar

3.2. SARIMAX

O variație a modelului ARIMA, care ia în considerare atât variabilele sezoniere, cât și cele exogene în date se numește SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous Variables). Toate elementele modelului ARIMA sunt incluse în modelul SARIMAX, dar permite și inserarea de predictori sau regresori externi care nu fac parte din seria temporală modelată.

În timp ce modelul SARIMAX este deosebit de util atunci când datele demonstrează caracterul sezonier, care este un model repetat de comportament care se întâmplă la intervale regulate, modelul ARIMA este adecvat pentru datele din seria temporală care nu sunt sezoniere. În aceste situații, modelul SARIMAX permite introducerea de variabile sezoniere pentru a capta periodicitatea datelor, cum ar fi numărul de zile dintr-o săptămână, luni dintr-un an sau trimestre dintr-un an fiscal.

Modelul SARIMAX permite, de asemenea, adăugarea de factori externi care ar putea afecta seria temporală studiată. Variabilele exogene sunt lucruri din afara seriei temporale care au un impact asupra acesteia, dar nu fac de fapt parte din seria temporală. Variabilele exogene pot fi orice, de la indicatori economici la modele meteorologice până la informații demografice. Modelul SARIMAX poate surprinde mai bine impactul acestor evenimente exterioare asupra seriilor temporale care sunt modelate prin includerea unor astfel de variabile.

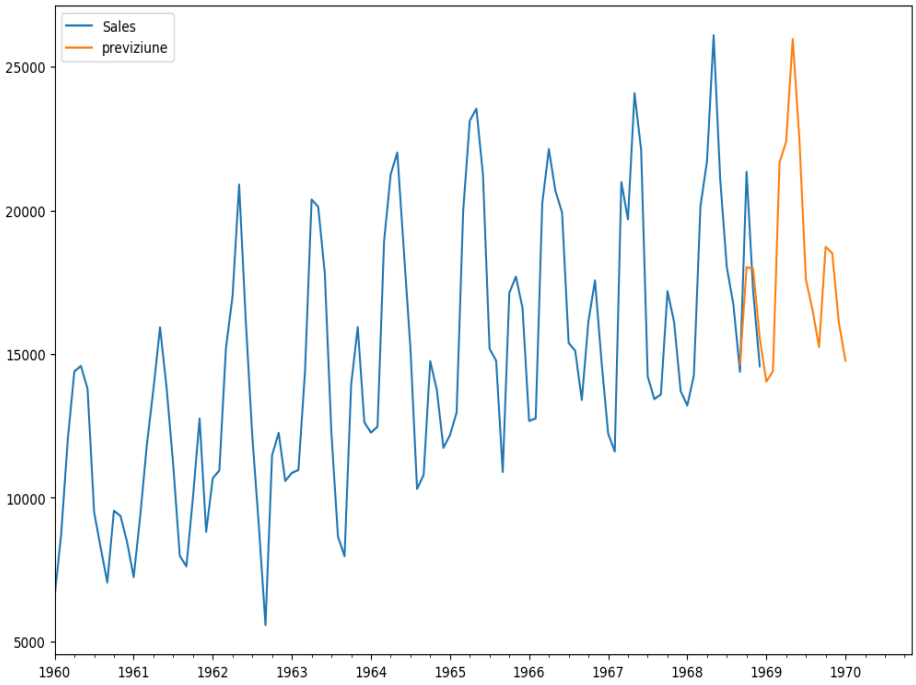


Figura 2. Set de date pe care se face previziunea utilizând metoda SARIMAX

4. Integrarea Colab,Python si SARIMAX

Modelul SARIMAX pentru prognoza serii temporale poate fi implementat cu ușurință folosind Python și *Google Colab*[6]. Librariile Python precum *Pandas*[8], NumPy și *Statsmodels*[7] pot fi integrate perfect cu platforma bazată pe cloud Colab.

Importul bibliotecilor adecvate și introducerea datelor din seria temporală într-un DataFrame sunt primii pași în utilizarea modelului SARIMAX în Colab. Pentru a elimina orice valori nevalide sau valori aberante, datele ar trebui să fie preprocesate și curățate după caz.

Prognozele pentru perioadele de timp viitoare pot fi create folosind modelul SARIMAX după ce acesta a fost calculat și validat. Instrumentele Python precum Matplotlib sau Seaborn pot fi folosite pentru a ilustra rezultatele, iar valori precum eroarea medie absolută sau eroarea medie pătrată pot fi folosite pentru a evalua cât de precise au fost prognozele.

Astfel acest trio poate fi utilizat pe o gama larga de set-uri de date usurand munca si eliminand greseala umana.

5. Analiza datelor

O tehnică numită prognoză în serie de timp este utilizată pentru a prezice valorile viitoare ale unei variabile pe baza valorilor sale trecute. Mai multe industrii, inclusiv buisness, economie, finanțe, asistență medicală și meteorologie, pot folosi această tehnică.

Este nevoie de un set de date cu o variabilă de tip serie temporală pentru a aplica previziunile serii temporale. Setul de date ar trebui să includă o serie de observații efectuate de-a lungul timpului, care pot fi la intervale neregulate sau regulate (cum ar fi ora, zi, saptamana sau luna). Setul de date ar trebui să fie suficient de lung pentru a identifica modele în date și pentru a produce predicții precise. În multe domenii diferite, prognoza poate oferi informații aprofundate pentru luarea deciziilor.

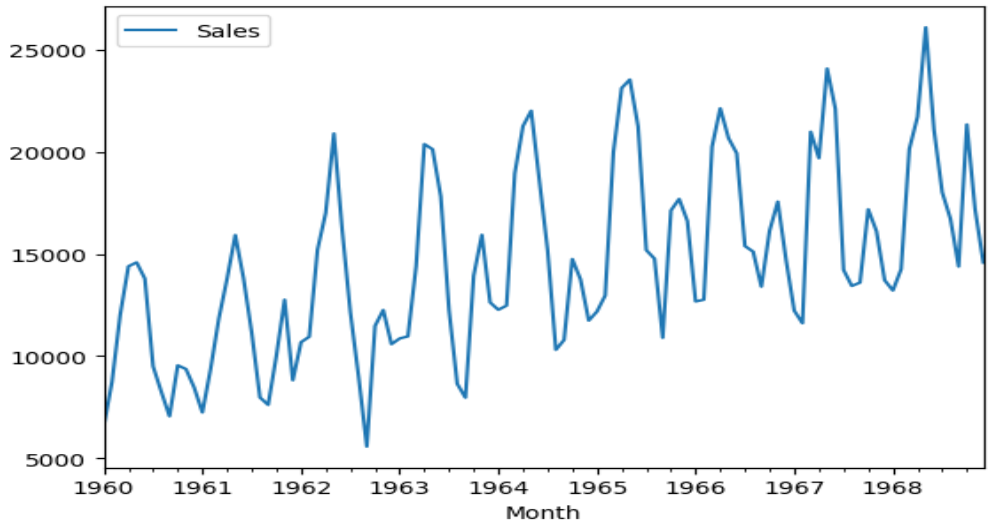
	Month	Sales
0	1960-01	6550
1	1960-02	8728
2	1960-03	12026
3	1960-04	14395
4	1960-05	14587

Figură 3. Verificare prin data.head() a primelor date din setul de date

	Month	Sales
103	1968-08	16722
104	1968-09	14385
105	1968-10	21342
106	1968-11	17180
107	1968-12	14577

Figură 4. Verificare prin data.tail() a ultimeleor date din setul de date

Trasarea datelor din seria temporală și analizarea tiparelor acestora este prima etapă în prognozarea seriilor temporale. Pe parcursul datelor, ar trebui să existe tendințe, sezonalitate, ciclicitate și alte modele. Sezonalitatea se referă la modelele recurente ale datelor care au loc la intervale regulate, în timp ce tendințele sunt modificări pe termen lung ale mediei seriilor temporale (de exemplu, zilnic, săptămânal, anual). Termenul „ciclicitate” descrie modele, cum ar fi ciclurile economice sau ciclurile economice, care au loc la intervale neregulate, dar au o perioadă previzibilă.



Figură 5. Setul de date dispus prin comanda data.plot()

În concluzie, o variabilă de serie cronologică cu o succesiune de observații colectate de-a lungul timpului constituie adesea setul de date necesar pentru prognozarea seriilor temporale. O metodă adecvată de prognoză ar trebui aleasă pe baza caracteristicilor datelor și a obiectivelor de prognoză după ce datele au fost evaluate pentru tendințe. Setul de date este folosit pentru a antrena modelul, iar diferiti metrici sunt utilizati pentru a evalua acuratețea prognozelor. Pentru luarea deciziilor într-o varietate de industrii, prognoza serii cronologice poate oferi informații profunde.

Concluzii

Pe scurt, analiza seriilor temporale este o abilitate esențială într-o gamă largă de discipline, cum ar fi finanțe, economie și studii de mediu. Capacitatea de a prezice valorile viitoare cu acuratețe folosind tendințele istorice ar putea oferi informații utile pentru luarea deciziilor.

Două modele statistice populare pentru prognoză și analiza serii de timp sunt ARIMA și SARIMAX. Un model popular pentru datele serii temporale non-sezoniere este ARIMA, iar o variantă numită SARIMAX poate lua în considerare variabilele exogene și sezonalitatea. Ambele modele necesită preprocesarea datelor și estimarea parametrilor, care pot fi realizate cu ajutorul unor programe precum Python și biblioteci precum Statsmodels.

SARIMAX este mai complicat decât ARIMA, dar pentru că ia în considerare influențele sezoniere și exogene care afectează seria temporală analizată, poate genera prognoze mai precise. În schimb, ARIMA poate fi mai ușor de implementat și este mai potrivită pentru date non-sezoniere. În orice situație, este esențial să se evalueze acuratețea prognozei pentru a se asigura că modelele oferă informații valoroase în ambele scenarii.

Aceste modele pot fi implementate cu ajutorul unor instrumente robuste precum Python și Google Colab, iar acuratețea prognozelor poate fi evaluată folosind o varietate de metrici, cum ar fi eroarea medie absolută sau eroarea medie pătrată. Aceste modele pot fi utilizate pentru a îmbunătăți alocarea resurselor, luarea deciziilor și eficiența într-o serie de industrii.

Bibliografie online:

[1] https://www.kaggle.com/datasets/hugoherrera11/monthly-car-sales

[2] https://en.wikipedia.org/wiki/Moving_average

[3] https://www.expressanalytics.com/blog/time-series-analysis/

[4] https://analyticsindiamag.com/general-overview-of-time-series-data-analysis/

[5] https://docs.oracle.com/cd/E57185_01/CBREG/ch06s03s04.html

[6] https://colab.research.google.com/

[7] https://www.geeksforgeeks.org/linear-regression-in-python-using-statsmodels/

[8] https://www.geeksforgeeks.org/pandas-tutorial/