|  |
| --- |
| **Previzionarea acțiunilor Google** |
| Analiza avansată a seriilor de timp Master Statistică Aplicată și Data Science |
| Cercetarea vizează combinarea analizei tradiționale a seriilor de timp cu utilizarea algoritmilor de Deep Learning pentru a prognoza prețul acțiunilor companiei Google, firmă aflată pe locul 5 în lume în ceea ce privește Market Capitalisation. |
|  |

Cuprins

[Introducere 2](#_Toc74523935)

[Scop și obiective 3](#_Toc74523936)

[Revizia literaturii 4](#_Toc74523937)

[Analiza univariată a seriei de timp 5](#_Toc74523938)

[Testarea staționarității 5](#_Toc74523939)

[Tehnici de netezire exponențială 23](#_Toc74523940)

[Analiza multivariată a seriei de timp 31](#_Toc74523941)

[Deep Learning – LSTM 31](#_Toc74523942)

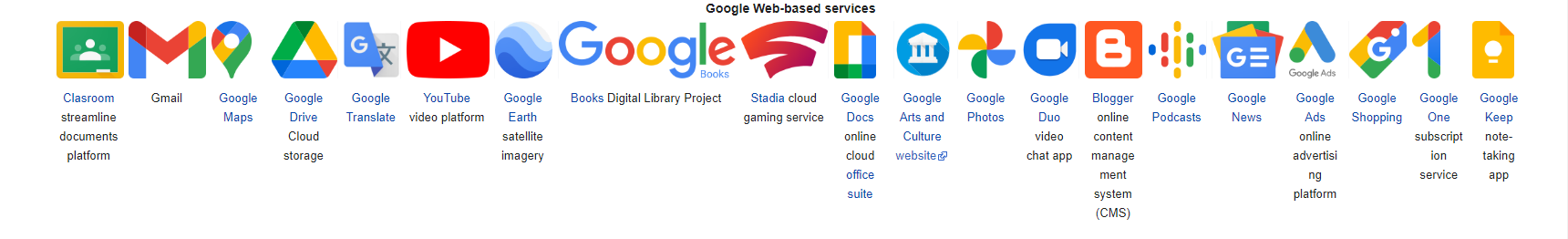
[Conluzii 35](#_Toc74523943)

[Bibliografie 36](#_Toc74523944)

# Introducere

Google LLC este una dintre cele mai cunoscute companii din lume. Compania este o multinațională americană, specializată în servicii și produse oferite prin intermediul internetului, ce includ: tehnologii de publicitate online, un motor de căutare (conform Alexa.com, motorul de căutare Google.com este în prezent cel mai vizitat website din lume, urmat la câteva locuri distanță de Youtube, Facebook sau Amazon), cloud computing, software și hardware. Este considerată una dintre cele cinci companii Big Tech împreună cu Amazon, Facebook, Apple și Microsoft.

Fondată în 1998 de doi studenți la Universitatea Stanford din California, Google a devenit o companie publică printr-o ofertă publică inițială (IPO) pe 19 august 2004. În octombrie 2015, Google s-a reorganizat ca filială a unui conglomerat numit Alphabet Inc. (rămânând totodată și cea mai mare sucursală). Creșterea rapidă a companiei de la încorporare a inclus produse, achiziții și parteneriate dincolo de motorul principal de căutare. Firma oferă în prezent zeci de servicii ce servesc unor scopuri diverse.

Figura 1: Produse oferite de Google  
Sursă: https://en.wikipedia.org/wiki/Google

PwC a clasat cele mai mari companii cotate la bursă după capitalizarea lor de piață în dolari, iar Alphabet se află pe locul 5 în lume. De asemenea, pe lista celor mai valoroase branduri, Google este clasat pe locul al doilea de Forbes și pe locul al patrulea de Interbrand. Cu toate acestea, compania a primit critici semnificative care implică probleme precum confidențialitatea, evitarea plăților impozitelor, cenzură, neutralitatea căutărilor, antitrust și abuzul poziției sale de monopol.

# Scop și obiective

Proiectul are ca scop previzionarea prețului acțiunilor firmei Google LLC prin modelarea datelor și utilizarea modelelor econometrice, folosind date de pe o perioadă de peste 5 ani, din intervalul 4/1/2016 – 30/4/2021. Acest set de date, care a fost preluat de pe Yahoo Finance, are 1341 de observații (ce corespund zilelor de tranzacționare la bursă) și conține informații despre:

* **Open** (prețul de deschidere);
* **High** (prețul maxim dintr-o zi de tranzacționare);
* **Low** (prețul minim dintr-o zi de tranzacționare);
* **Close** (prețul de închidere);
* **Adjusted close** (prețul de închidere ajustat);
* **Volume** (volumul de tranzacționare).

În realizarea proiectului s-a avut în vedere ilustrarea concepte de analiză a seriilor de timp atât pentru seriile univariate, cât și pentru cele multivariate.

Ținem să menționăm că am folosit ticker-ul GOOGL, adică acțiunile de clasă A ale companiei Alphabet (acțiuni care oferă și drepturi de vot, spre deosebire de GOOG, ce se încadrează în clasa C și nu oferă acest privilegiu).

# Revizia literaturii

Modelele econometrice se folosesc deseori pentru previzionarea prețurilor acțiunilor în viitor, întrucât piața bursieră este considerată „haotică, complexă, volatilă și dinamică”. (Ritika & Shashi, 2016) Analiștii încearcă să găsească cel mai bun model cu putință, care să asigure totodată și cea mai mare acuratețe.

De exemplu, Arkaprava et. al. au încercat să afle dacă previzionarea seriilor de timp și modelele LTSM (Long short-term memory) pot fi utilizate pentru a observa evoluția prețurilor. Autorii au combinat analiza tradițională a seriilor de timp cu text mining, știrile fiind extrase de pe New York Times pentru o perioadă de 10 ani. Așa cum menționează și articolul, predicția prețurilor bursiere se bazează pe Efficient Market Hypothesis, adică pe faptul că piața reacționează instant la știri, iar prețurile se ajustează întocmai. (Arkaprava, et al., 2018)

Arkaprava et. al aplică diverse teste (Holt-Winters, Netezire Exponențială, ARIMA) pentru a previziona prețul acțiunilor. S-a descoperit că variabile precum sentimentul pozitiv, sentimentul negativ, sentimentul neutru și inflația sunt semnificative pentru procesul de previzionare. De asemenea, evenimente ca uraganul Sandy, impactul trimestrial, Lunea Neagră și alegerile prezidențiale ajută la îmbunătățirea acurateței modelului. (Arkaprava, et al., 2018)

S-au implementat de asemenea și algoritmi de Deep Learning, pentru care s-au extras headline-urile de la secția „Business and Technology”. În final s-a concluzionat că LTSM a avut cel mai mic RMSE, iar modelul RNN a avut cea mai mare acuratețe dintre toate modelele de clasificare folosite (78%). (Arkaprava, et al., 2018)

# Analiza univariată a seriei de timp

## Testarea staționarității

Primul pas în analiza oricărei serii de timp este reprezentarea grafică. Utilizând Eviews, am realizat graficul ce arată evoluția prețului acțiunilor (Adjusted Close) în perioada 2016-2021 și am realizat statistica descriptivă.

Figura 1: Prețul ajustat de închidere



Sursă: Prelucrare proprie în Eviews

Din grafic, prima impresie pe care o avem este că seria are o tendință crescătoare în timp. Se poate spune că media, varianța și autocovarianța nu par a fi par a fi invariate în raport cu timpul. Seria este nestaționară, iar trendul are o tendință de tip stochastic (în sensul că se modifică pe diverse secvențe de timp).

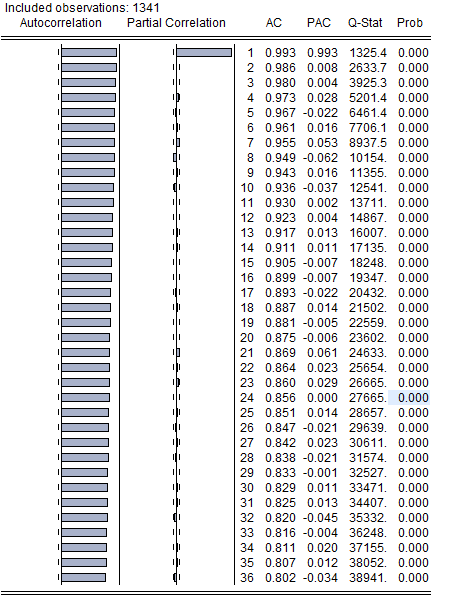
Figura 2: Statistici descriptive



Sursă: Prelucrare proprie în Eviews

Mai departe, realizăm un test simplu al staționarității seriei cu ajutorul funcției de autocorelație (ACF), printr-o corelogramă ce reprezintă graficul funcţiei de autocorelaţie în raport cu decalajul k. Din coloana AC se poate observa că toți coeficienții 𝜌̂𝑘 sunt semnificativi statistic. De asemenea, AC începe cu valori foarte mari și scade treptat, deci este clar că seria este nestaționară.

Figura 3: Graficul funcției de autocorelație



Sursă: Prelucrare proprie în Eviews

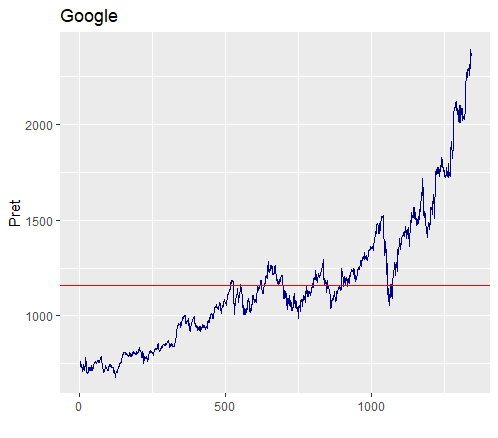
Conform testului Bartlett, varianța estimatorului coeficientului de corelație este: 𝑉̑𝑎𝑟(𝜌̂𝑘) = 1𝑛[1+2(𝜌̂12+𝜌̂22+⋯+𝜌̂𝑘−12)], iar pentru n suficient de mare (aproximarea repartiției Student prin cea normală), se utilizează: 𝑉̑𝑎𝑟(𝜌̂𝑘)=1/𝑛. Rezultă că 𝜌̂𝑘~𝑁(0,1/𝑛)şi 𝑠𝑒(𝜌̂𝑘)=1/√𝑛. Așadar, un interval de încredere 95% pentru 𝜌𝑘 are forma (−1,96/√𝑛;+1,96/√𝑛).

Testarea semnificației unui singur coeficient de autocorelaţie 𝜌𝑘:  
𝐻0:𝜌𝑘=0 (𝜌𝑘 nu este semnificativ statistic)  
𝐻1:𝜌𝑘≠0 (𝜌𝑘 este semnificativ statistic)

În cazul nostru, deoarece n = 1341, varianța lui 𝜌̂𝑘 este 1/1341, iar eroarea standard este √1/1341 (0.0273). Intervalul de încredere 95% pentru orice 𝜌𝑘 va fi ±1,96(0,0273) = ± 0.0535. Așadar, dacă 𝜌̂𝑘 se află în intervalul (-0.0535; 0.0535), nu respingem ipoteza că 𝜌𝑘 real este zero, iar în caz contrar respingem ipoteza.

Scopul acestei analize este de a construi un proces pentru a estima în mod corespunzător VaR, având în vedere volatilitatea în timp. VaR este utilizată pe scară largă pentru a măsura riscul de piață pentru instituțiile financiare. Datele noastre din seria cronologică constau în randamente de stocuri pentru 1340 de zile. Cu scopul de a explica o mică proporție a varianței randamentelor zilnice, efectuăm metodologia Box-Jenkins pentru a găsi un model de medie mobilă integrată autoregresivă (ARIMA). Mai târziu, verificăm normalitatea randamentelor şi încercăm să găsim cele mai potrivite distribuții. Estimăm varianța condițională a reziduurilor cu metoda Heteroscedasticității Generalizate Autoregresive (GARCH), comparând-o cu abordarea delta-normală. În cele din urmă, efectuăm cu un pas înainte prognoza VaR și efectuăm testarea înapoi pentru a verifica dacă modelul nostru este adecvat.

În scopul modelării, am colectat date (ianuarie 2016 - aprilie 2021) corespunzătoare preţurilor acţiunilor Google la o frecvență zilnică (un total de 1341 observații).



Sursă: Prelucrare proprie în R Studio

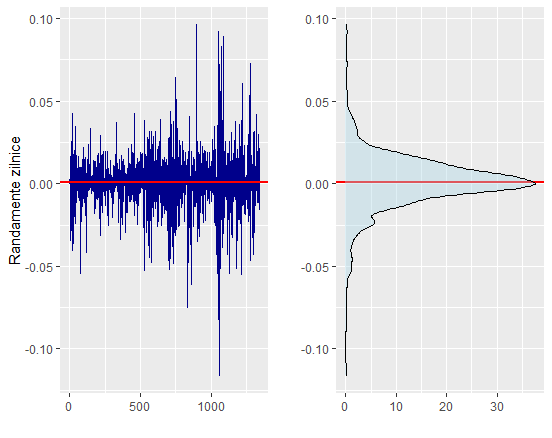
Linia roșie indică prețul mediu de închidere pentru intervalul de timp menţionat.

Procesele non-staționare au medii, varianțe și covarianțe care se schimbă în timp. Folosirea datelor nestaționare duce la o prognoză nesigură. Un proces staționar fluctuează în jurul unei medii constante cu varianță constantă.

Pentru a rezolva această problemă a staţionarităţii preferăm să calculăm randamentele zilnice.

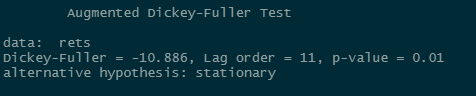


Sursă: <https://www.kaggle.com/ionaskel/value-at-risk-estimation-using-garch-model>



Sursă: Prelucrare proprie în R Studio

Pentru a verifica staționaritatea randamentelor, utilizăm testul Augmented Dickey-Fuller în care ipoteza nulă indică serii de timp nestaționare.



Sursă: Prelucrare proprie în R Studio

P-value (<0,01) sugerează că există suficiente dovezi pentru a respinge ipoteza nulă, prin urmare seria de timp devine staționară.

Metodologia Box-Jenkins

Această metodă utilizează o abordare de modelare în trei etape:

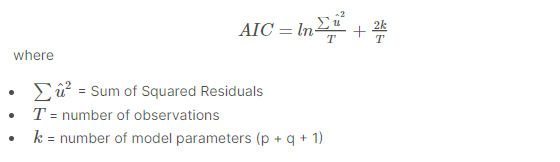
a) identificare,

b) estimare,

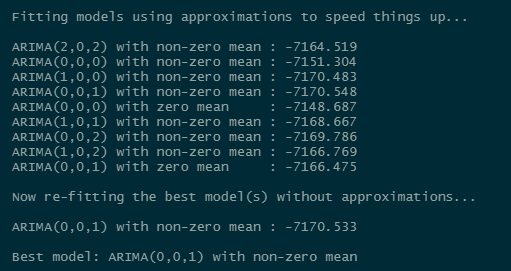
c) verificare diagnostic

**Identificare**

Pentru a utiliza metodologia Box-Jenkins trebuie să ne asigurăm că seriile cronologice sunt staționare. În cazul nostru, folosim returnările stocului pe care l-am verificat deja pentru staționaritate în partea anterioară. Mai mult, pe baza funcției de autocorelație (ACF) și a funcției de autocorelație parțială (PACF) este posibil să se determine ordinea p, d și q a modelului ARIMA. O altă modalitate de identificare a modelului este Criteriul informațional Akaike (AICc). AIC estimează calitatea fiecărui model în raport cu fiecare dintre celelalte modele.



Sursă: <https://www.kaggle.com/ionaskel/value-at-risk-estimation-using-garch-model>

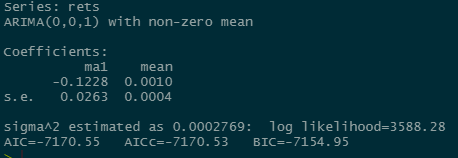


Sursă: Prelucrare proprie în R Studio

Se poate observa cu procesul de mai sus că am calculat scorurile AIC pentru diferite modele ARIMA și deducem că modelul adecvat este un (MA (1)).

**Estimare**

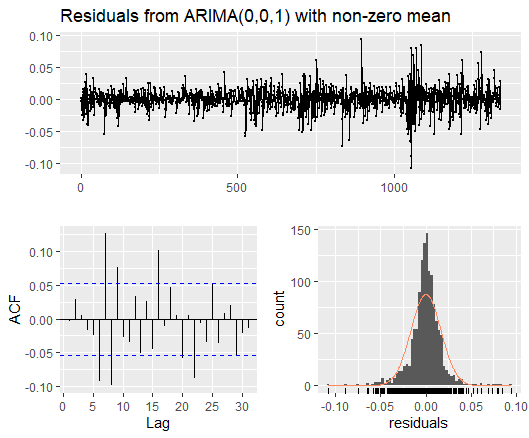
Pentru a estima coeficienții parametrilor folosim Maximum Likelihood. Folosind ARIMA (0, 0, 1) ca model selectat, rezultatele sunt după cum urmează:



Sursă: Prelucrare proprie în R Studio

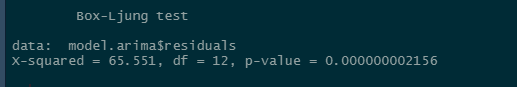
**Verificare diagnostic**

Procedura include observarea graficului reziduurilor și a diagramelor ACF și PACF și verificarea rezultatului testului Ljung-Box. Dacă reziduurile ACF și PACF ale modelului nu prezintă laguri semnificative, modelul selectat este adecvat.



Sursă: Prelucrare proprie în R Studio

Conform ACF reziduurile par corelate. Pentru o mai mare acurateţe vom aplica testul Ljung-Box.

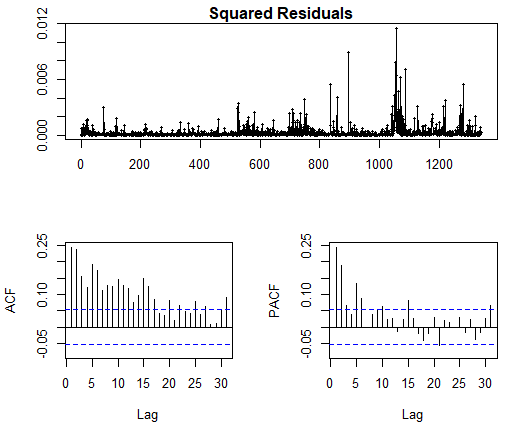


Sursă: Prelucrare proprie în R Studio

Potrivit testului Ljung-Box, p-value < 0.05, ceea ce înseamnă că reziduurile modelului nu sunt independente şi sunt autocorelate.

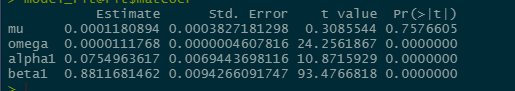
**Implementare GARCH**

Procesul GARCH este valabil atunci când squared residuals sunt corelate. Graficele ACF și PACF indică în mod clar o corelație semnificativă.



Sursă: Prelucrare proprie în R Studio

O altă modalitate de a testa heteroscedasticitatea squared residuals este efectuarea testelor de semnificație pe parametrii a1 și β1.



Sursă: Prelucrare proprie în R Studio

Atât a1 cât și β1 sunt semnificativ diferite de zero, prin urmare este rezonabil să presupunem volatilitatea variabilă în timp a reziduurilor.

**VaR**

Estimează cât de mult ar putea pierde un set de investiții în condițiile normale de piață într-o perioadă de timp stabilită. O statistică VaR are trei componente:

a) perioada de timp

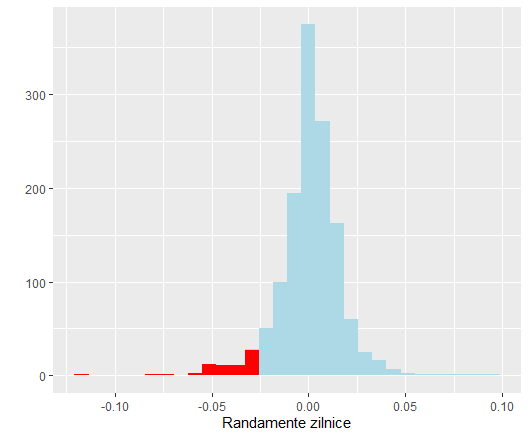
b) nivelul de încredere

c) cantitatea de pierdere (sau procentul de pierdere)

Pentru un nivel de încredere de 95%, putem spune că cea mai gravă pierdere zilnică nu va depăși estimarea VaR. Dacă folosim date istorice, putem estima VaR luând valoarea cuantilă de 5%. Pentru datele noastre, această estimare este:



Sursă: Prelucrare proprie în R Studio

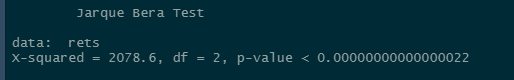


Sursă: Prelucrare proprie în R Studio

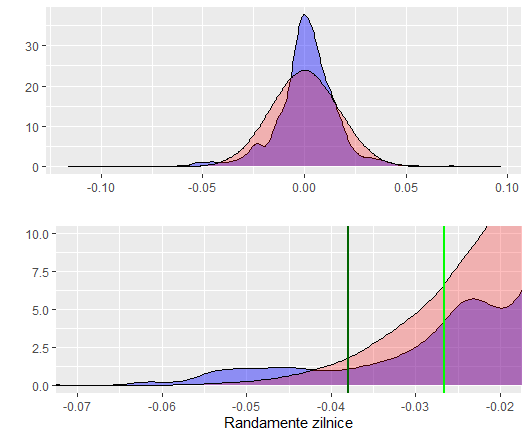
Barele roșii se referă la returnări mai mici de 5% cuantile.

**Proprietăți de distribuție**

Pentru a estima VaR, trebuie să definim corect cuantila corespunzătoare a distribuției asumate. Pentru distribuția normală, cuantila corespunzătoare a = 5% este -1,645. Dovezile empirice sugerează că presupunerea normalității produce adesea rezultate slabe. Testul Jarque-Bera poate testa ipoteza că randamentele stocului urmează o distribuție normală.



Sursă: Prelucrare proprie în R Studio



Sursă: Prelucrare proprie în R Studio

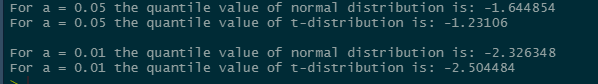
În figura de mai sus, graficele de densitate sunt afișate pentru randamentele stocului (albastru) și datele normal distribuite (roșu). Liniile verticale ale graficului inferior reprezintă cuantila normală corespunzătoare pentru a = 0,05 (verde deschis) și a = 0,01 (verde închis). Graficul inferior indică faptul că, pentru semnificația de 95%, utilizarea normală a distribuției poate supraestima VaR. Cu toate acestea, pentru un nivel de semnificație de 99%, o distribuție normală ar subestima riscul.

**Distribuţia t-student**

Pentru a modela mai adecvat grosimea cozilor, putem folosi alte ipoteze de distribuție pentru rentabilitatea stocului. Distribuția t este simetrică și în formă de clopot, la fel ca distribuția normală, dar are cozi mai grele, ceea ce înseamnă că este mai predispusă la producerea unor valori care sunt departe de media sa. Folosim funcția fitdist din pachetul rugarch pentru a obține parametrii de potrivire ai distribuției t.



Sursă: Prelucrare proprie în R Studio



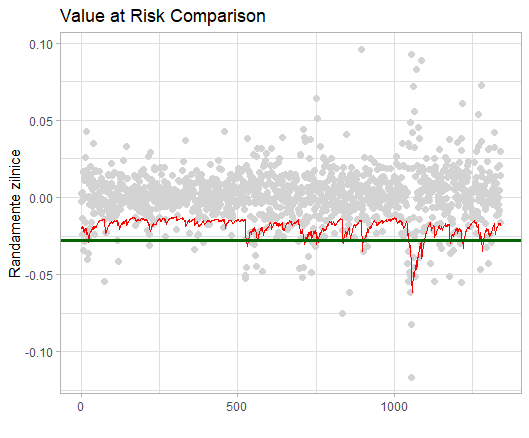
Sursă: Prelucrare proprie în R Studio

După cum observăm mai sus, cuantilele pentru nivelul de semnificație de 95% indică faptul că distribuția normală supraestimează riscul, dar pentru 99% nu reușește să capteze existența valorilor aberante, de aceea apare subestimarea riscului.

**Garch VaR vs Delta-normal**

Abordarea Delta-normală presupune că toate randamentele stocului sunt distribuite în mod normal. Această metodă constă în a reveni în timp și a calcula varianța randamentelor.

Anterior am observat că randamentele prezintă o volatilitate care variază în timp. Prin urmare, pentru estimarea VaR utilizăm varianța condițională dată de modelul GARCH (1,1). Pentru proprietățile de distribuție ale activului avut în vedere, folosim distribuția t student.



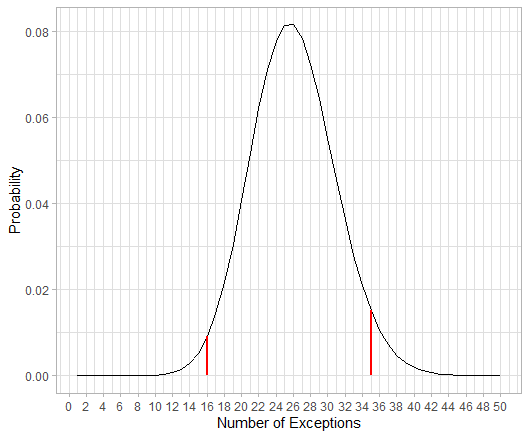
Sursă: Prelucrare proprie în R Studio

Linia roșie indică VaR produs de modelul GARCH, iar linia verde se referă la VaR delta-normal.

**VaR forecasting**

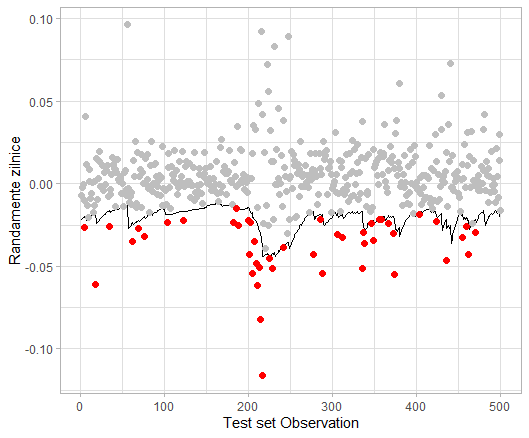
Am setat ultimele 500 de observații ca set de testare și efectuăm o prognoză în mișcare cu 1 pas înainte a deviației standard condiționate, σ ^ t + 1 | t. Reestimăm parametrii GARCH la fiecare 50 de observații.

**Backtesting**



Sursă: Prelucrare proprie în R Studio

Graficul de mai sus reprezintă distribuția probabilităților pentru excepții date de distribuția binomială. Numărul așteptat este 25 (= 500obs. X 5%). Cele două linii roșii denotă nivelul de încredere de 95%, cel inferior fiind 16 și cel superior 35. Prin urmare, atunci când verificăm excepțiile de pe setul de testare, ne așteptăm ca un număr între 16 și 35 să afirme că modelul GARCH este previzionat cu succes.



Sursă: Prelucrare proprie în R Studio

Linia neagră reprezintă VaR-ul prognozat zilnic dat de modelul GARCH, iar punctele roșii se referă la randamente mai mici decât VaR. Pasul final este de a număra numărul de excepții și de a-l compara cu cel generat cu abordare delta-normală.



Sursă: Prelucrare proprie în R Studio

După cum am afirmat mai devreme, ne așteptam ca abordarea delta-normală să supraestimeze riscul. La backtesting, de 46 de ori randamentele au fost mai mici decât VaR, situându-se în afara nivelului de semnificație de 95% (<16). Pe de altă parte, nici abordarea GARCH (46 de excepții) nu pare a fi un instrument predictiv eficient în acest caz.

## Tehnici de netezire exponențială

Netezirea exponențială, metodă apărută pentru prima dată în 1950, este o tehnică de ajustare care se poate aplica unei serii de timp cu o pronunțată componentă aleatoare. Modelul folosește medii ponderate, în care observațiile cele mai recente reprezintă ponderile cele mai mari – iar cele mai îndepărtate observații cântăresc tot mai puțin în medie. Această metodă este destul de intuitivă, în general putând fi aplicată pe o serie mare de date.

Printre tehnicile de netezire exponențială se numără simple exponential smoothing (SES), metoda Holt (care permite utilizarea seriilor de timp cu trend) și extensia ei, metoda Holt-Winters, care a fost dezvoltată pentru a cuprinde și sezonalitatea.

Am început analiza cu metoda simplă de netezire exponențială, care este potrivită prognozării datelor care nu au un trend sau o sezonalitate bine definită. Ecuația este: yt+1|t=αyt+(1−α)^yt|t−1, unde α, parametrul de netezire, ia valori între 0 și 1. În practică, dacă alfa ia valori între 0,1 și 0,2, atunci netezirea exponențială va funcționa destul de bine.

Când alfa este mai aproape de 0, atunci se consideră a fi învățare lentă, deoarece algoritmul acordă mai multă importanță datelor istorice. Dacă valoarea alfa este mai aproape de 1, atunci se numește învățare rapidă, deoarece algoritmul conferă datelor recente o pondere mai mare.

Folosind pachetele *tidyverse* și *fpp2*, am inițiat procedeul de netezire exponențială în R. Am ales α ca 0.2 și am ales h (perioada de prognozare) 21 de zile cu ajutorul funcției ses(). Așadar, 1320 de observații au fost clasate în setul de training, iar 21 în setul de testare. Alternativ, am testat procedeele și pentru o perioadă de prognozare de 63 de zile (pentru lunile februarie-aprilie).

Din graficul rezultat (figura 16), se observă că estimarea proiectată spre viitor este liniară. Așadar, putem spune că deocamdată nu este surprinsă tendința actuală. Pentru a corecta, am folosit funcția diff(), ce elimină componenta trend din seria de timp, și am reaplicat SES pentru aceste date. Rezultatele se pot vedea în figura 17.

Figura 16: Estimare prin Simple Exponential Smoothing

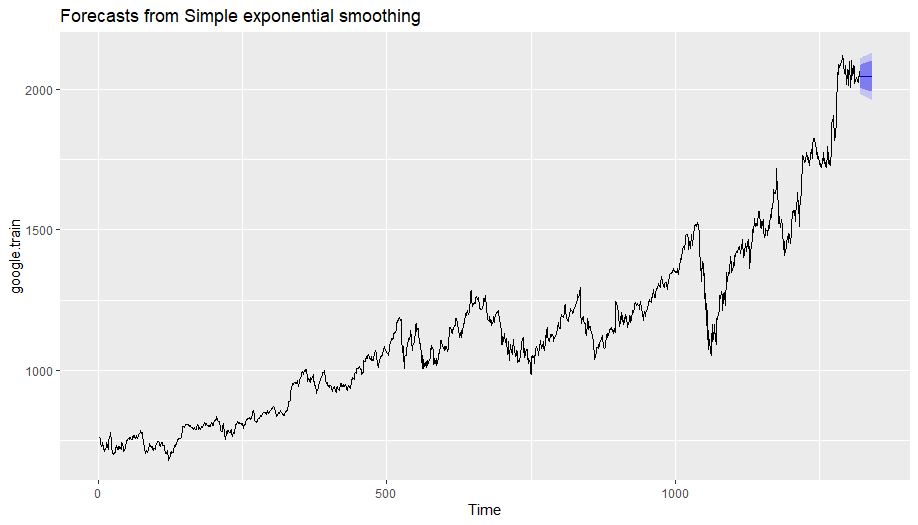
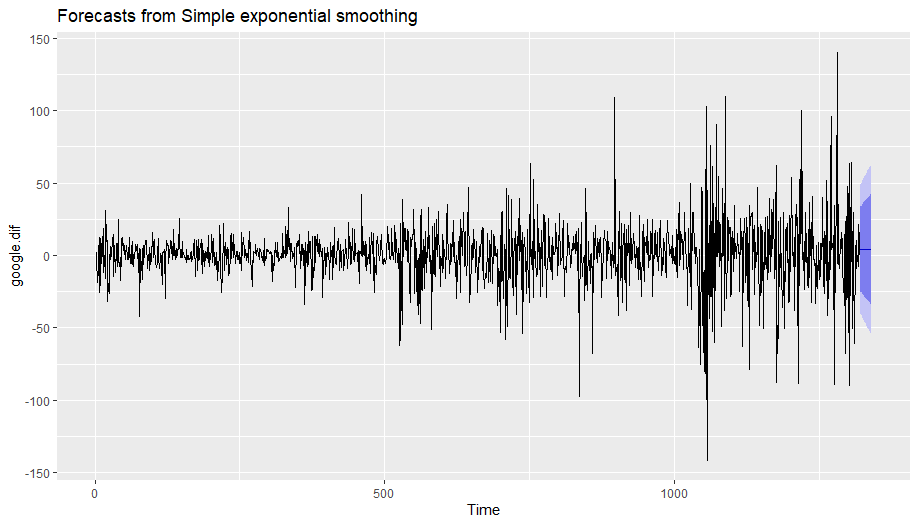
  
Sursă: Prelucrare proprie în R Studio

Figura 17: Reaplicarea SES după eliminarea trendului



Sursă: Prelucrare proprie în R Studio

Pentru a înțelege cât mai bine performanța modelului, am comparat prognoza noastră cu setul de testare. Deoarece setul de date a fost diferențiat, am diferențiat în continuare și setul de testare. Am setat valoarea lui α, folosind un loop, între 0.01 și 0.99. Mai departe, am aflat valoarea lui α care minimează cel mai mult valoarea RMSE (0.57). Rezultate sunt prezentate în figurile 18 și 19.

Figura 18: Indicatori de acuratețe ai modelului optim

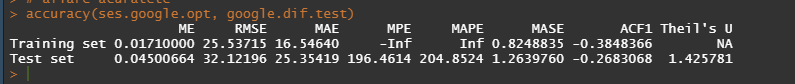
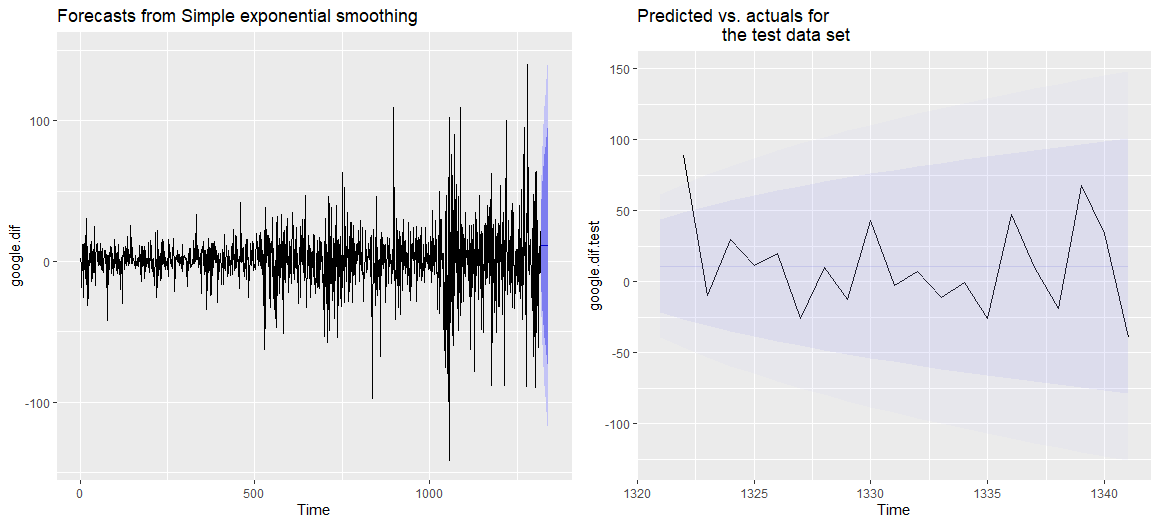
Sursă: Prelucrare proprie în R Studio

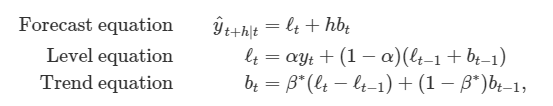
Figura 19: Ploturi pentru modelul optim (α = 0.57)

  
Sursă: Prelucrare proprie în R Studio

În continuare, vom aplica metoda lui Holt. Această tehnică funcționează la seriile de timp care arată un trend, dar nu și sezonalitate. Pentru a face predicții, Metoda Holt folosește 2 parametri de netezire, alpha și beta.

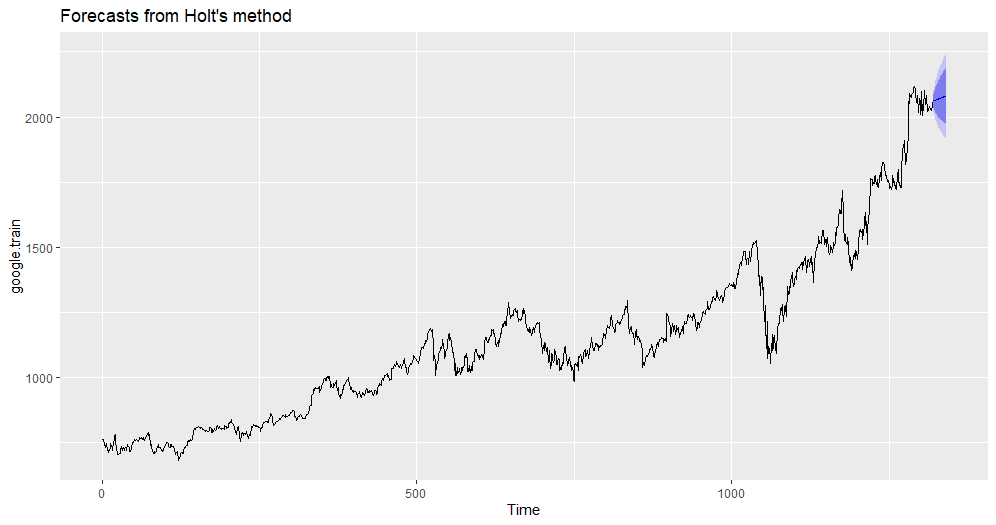
Ecuațiile modelului sunt:

Figura 20: Ecuațiile modelului Holt

  
Sursă: Forecasting: principles and practice

ℓt denotă o estimare a nivelului seriei la momentul t, *bt* denotă o estimare a tendinței (pantei) seriei la momentul t, iar α și β sunt coeficienți de netezire cuprinși între 0 și 1. Vom aplica metoda cu ajutorul funcției Holt().

Figura 21: Predicțiile metodei Holt

  
Sursă: Prelucrare proprie în R Studio

Mai departe, am aflat valorile lui alpha și beta și am observat și acuratețea modelului.

Smoothing parameters: alpha = 0.8762 beta = 1e-04   
Initial states: l = 764.9415 b = 0.9868 sigma: 20.7646  
 AIC AICc BIC   
17498.49 17498.54 17524.42

Valoarea optimă, adică beta = 0.0001, este utilizată pentru a elimina erorile din setul de training, așadar putem regla beta-ul la această valoare optimă. În continuare, aflăm valoarea optimă a lui beta printr-un loop cuprins între 0.0001 to 0.5 care va minimiza testul RMSE (0.499). Cu toate acestea, modelul Holt are un RMSE mare pentru setul de testare.

Figura 22: Indicatori de performanță pentru modelul optim Holt (α = 0.499)

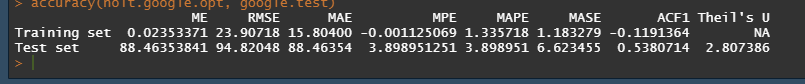
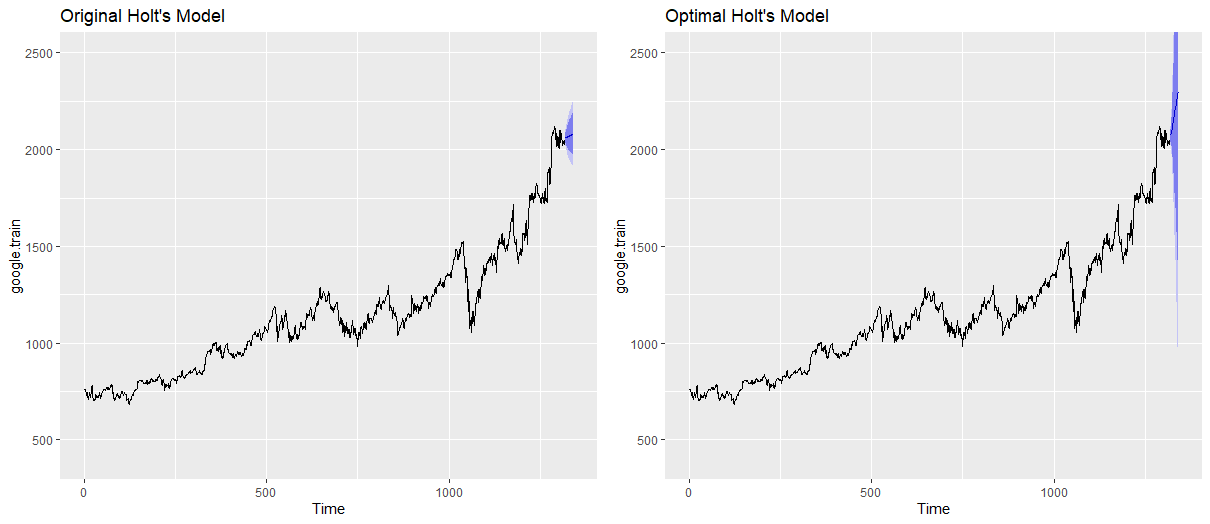
Sursă: Prelucrare proprie în R Studio

Figura 23: Modelul original vs. modelul optim

Sursă: Prelucrare proprie în R Studio

Pentru prognozarea pe 63 de zile (3 luni calendaristice) am rulat aceleași proceduri de netezire exponențială simplă și metoda Holt. Rezultatele obținute sunt prezentate în figurile de mai jos.

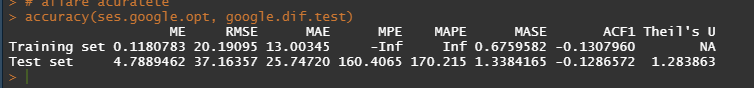
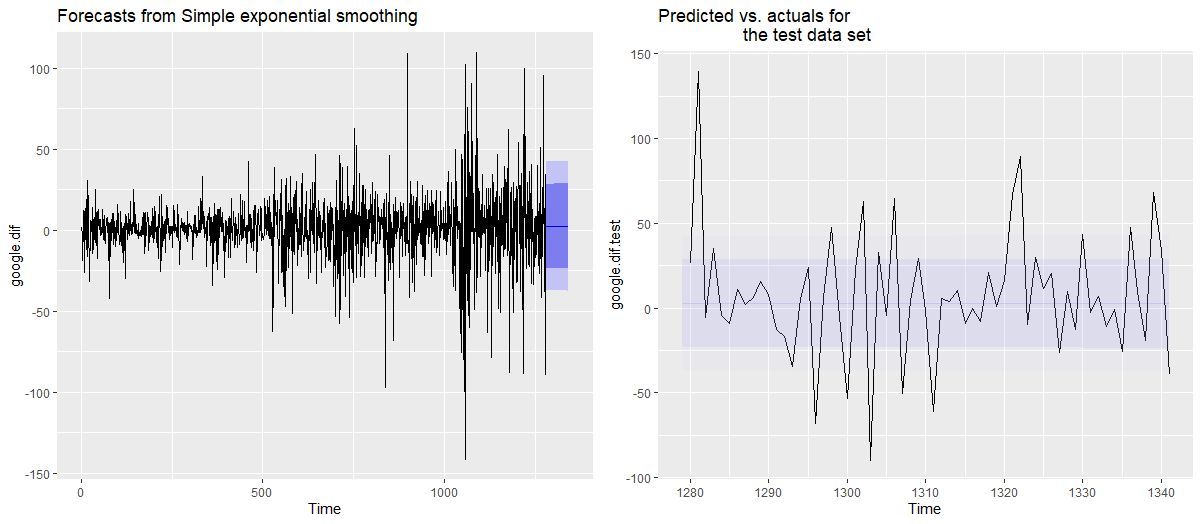
Figura 24: Indicatori de performanță ai modelului de netezire exponențială (α = 0.02)Sursă: Prelucrare proprie în R Studio

Figura 25: Ploturi pentru modelul optim



Sursă: Prelucrare proprie în R Studio

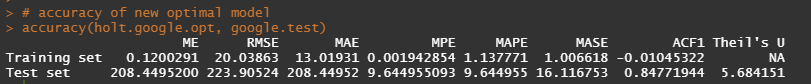
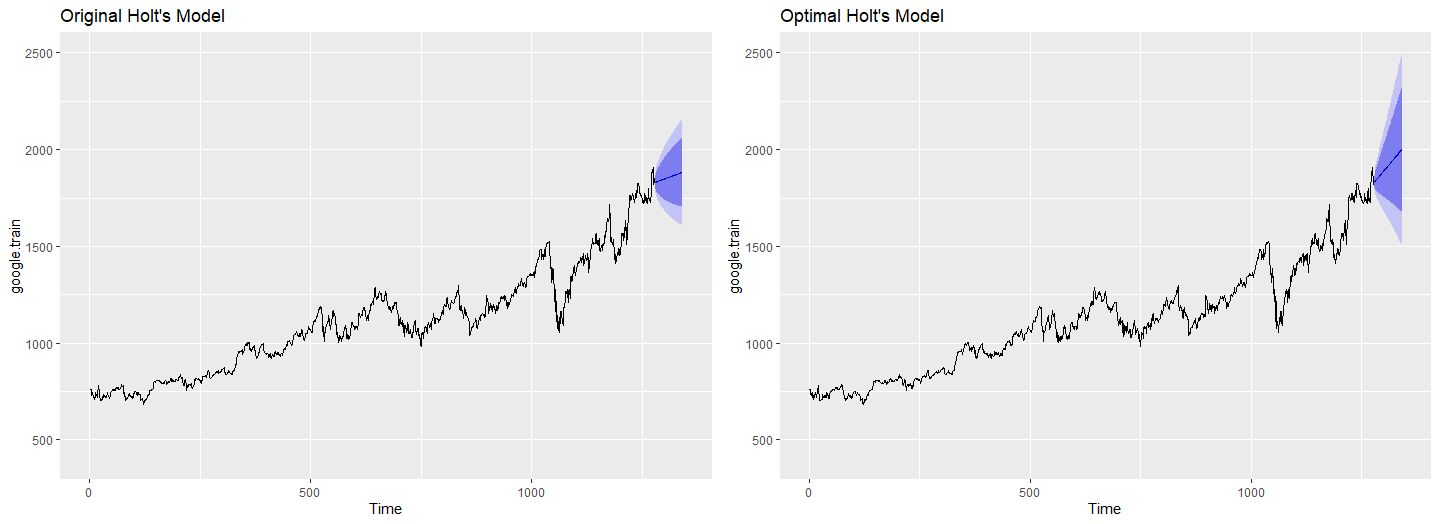
Figura 26: Indicatori de performanță ai modelului Holt (α = 0.0211)Sursă: Prelucrare proprie în R Studio

Figura 27: Ploturi pentru modelul optim Holt



Sursă: Prelucrare proprie în R Studio

Comparând cele 2 orizonturi de prognozare, putem observa că ambele metode funcționează mult mai bine pentru o perioadă de prognozare mai scurtă, întrucât indicatorii de performanță sunt mai buni. De asemenea, tehnica de netezire exponențială simplă este mai potrivită, pentru că RMSE-ul este minimizat (comparativ cu metoda Holt). Nu vom continua analiza cu Holt-Winters, întrucât datele nu au sezonalitate, iar acest tip de model nu este recomandat în prognozarea prețurilor acțiunilor.

# Analiza multivariată a seriei de timp

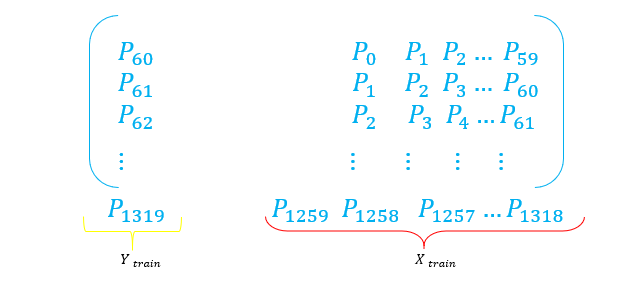
## Deep Learning – LSTM

Pentru analiza seriilor multivariate, vom aplica Deep Learning pentru Google și alte două companii considerate Big Tech. Mai exact, vom utiliza modelul Long Short-Term Memory (LTSM) pentru a previzona prețul acțiunilor în luna Aprilie 2021. Vom folosi datele corespunzătoare 4/1/2016 – 31/3/2021 pentru seria de training, iar luna aprilie va constitui seria de testare. Previzionăm în continuare prețul ajustat de închidere.

În primul rând, vom normaliza datele cu ajutorul librăriei Scikit-Learn din Python, după formula:

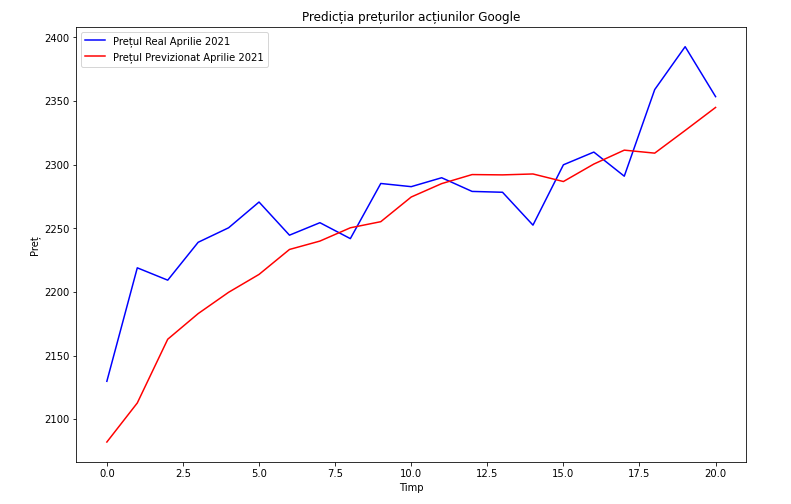
, unde și reprezintă valoare minimă și valoarea maximă a vectorului X.

Deoarece o lună are, în medie, 20 de zile de tranzacționare, alegem 60 ca număr de timestamps. Vom avea 2 parametri: (va conține prețul acțiunilor pentru ultimele 60 de zile) și (va conține prețul acțiunilor din acea zi încolo), acesta menținându-se pentru toate perioadele de timp. Astfel, se va crea un data frame în care lui îi corespund prețurile pentru 60 de zile precedente. Data frame-ul sumarizat este prezentat mai jos.



Pentru obținearea modelelor, folosim Tensorflow și Kensas. Am folosit 50 de neuroni pentru fiecare layer LTSM. Rezultatele obținute pentru toate cele 3 companii, cât și indicatorii de performanță ai modelelor sunt prezentate în continuare.

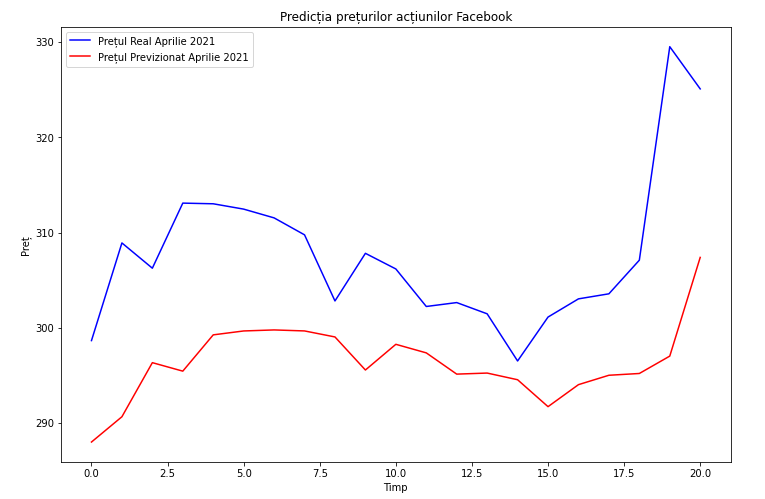
Figura : Preț real vs. preț previzionat - Google

  
Sursa: Prelucrare proprie în Jupyter Notebooks

Pentru acest model, am obținut:

Eroarea medie de prognoză: 23.01  
Eroarea medie absolută: 32.16  
MSE: 1696.21  
RMSE: 41.185

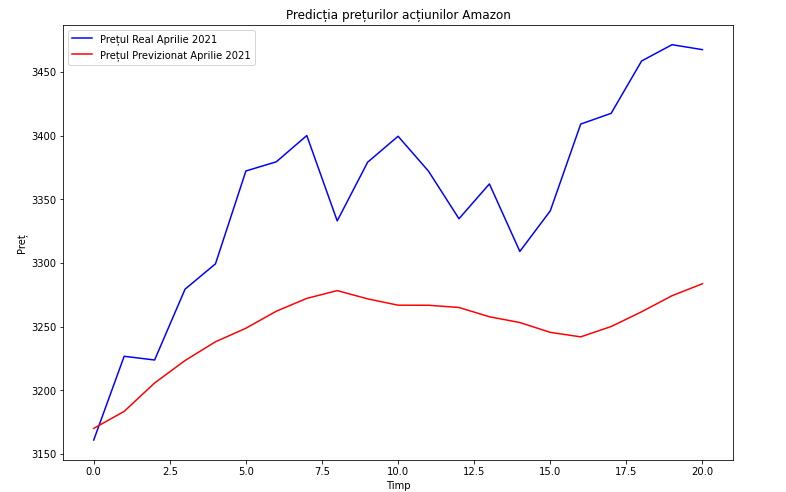
Figura : Preț real vs. preț previzionat - Facebook

  
Sursa: Prelucrare proprie în Jupyter Notebooks

Pentru acest model, am obținut:

Eroarea medie de prognoză: 11.35  
Eroarea medie absolută: 11.35  
MSE: 169.04  
RMSE: 13.001

Figura : Preț real vs. preț previzionat - Amazon

  
Sursa: Prelucrare proprie în Jupyter Notebooks

Pentru acest model, am obținut:

Eroarea medie de prognoză: 103.55  
Eroarea medie absolută: 104.42  
MSE: 13942.79  
RMSE: 118.07

Din toate cele 3 companii, observăm că algoritmul a funcționat cel mai bine pentru Facebook, deoarece are cel mai mic RMSE – astfel, LSTM a fost capabil să prezică corect tendințele ascendente și descendente, linia roșie corespunzătoare prețurilor previzionate urmează același model cu linia albastră, ce corespunde prețurilor reale. Pentru Google, LTSM a urmat tendința generală de creștere a prețului acțiunilor, dar nu a surprins toate mișcările pieței. În ceea ce privește Amazon, algoritmul a funcționat cel mai puțin bine – trendul pentru prețurile prognozate este mai degrabă liniar și nu include tendințele ascendente și descendente.

# Conluzii

Proiectul a vizat analiza seriilor de timp univariate și multivariate pentru prețul acțiunilor Google. Pe parcursul cercetării, am rulat mai multe teste pentru a obține o acuratețe cât mai bună. În ceea ce privește ARIMA și GARCH, nu am obținut un model eficient. Netezirea exponențială simplă a fost de ajutor într-o oarecare măsură, în timp ce metoda Holt nu a prezentat indicatori de performanță la fel de buni. De asemenea, am observat că Deep Learning este un algoritm bun, dar care a prezis cel mai bine mișcările pieței pentru Facebook.

# Bibliografie

1. Arkaprava, K., Arkaprava, K., Shanmugha, P. A. & Sanjay, V., 2018. GOOGLE STOCK PRICE FORECASTING, s.l.: s.n.
2. Ritika, S. & Shashi, S., 2016. Stock prediction using deep learning. Springer Science+Business.
3. Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.
4. <https://en.wikipedia.org/wiki/Google>
5. <https://www.investopedia.com/ask/answers/052615/whats-difference-between-googles-goog-and-googl-stock-tickers.asp>
6. <https://www.visualcapitalist.com/the-biggest-companies-in-the-world-in-2021/>
7. <https://en.wikipedia.org/wiki/Long_short-term_memory>
8. https://www.geeksforgeeks.org/exponential-smoothing-in-r-programming/
9. Tatev, K., A. Recurrent Neural Networks Case Study. Predicting the Stock Price of Google
10. <https://www.kaggle.com/ionaskel/value-at-risk-estimation-using-garch-model>
11. <https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=1029&context=etd_coll>
12. <https://www.econstor.eu/bitstream/10419/155283/1/880703547.pdf>