



Universitat Autònoma de Barcelona

Treball Final de Grau

Stock Machine Learning Predictions

Autor:
Marc Bosom Saperas

Tutor:
Antonio Lozano Bagen

GRAU EN MATEMÀTIQUES COMPUTACIONALS I
ANALÍTICA DE DADES

FACULTAT DE CIÈNCIES

29 de juny del 2024

Índex

1	Introducció	2
2	Motivació	2
3	Objectius	3
3.1	Objectius Específics	3
4	Plantejament del projecte	3
5	Anàlisi Avançada del Mercat de Valors	4
5.1	Oferta i Demanda	4
5.2	Volum	5
5.3	Estructures de Wyckoff	5
5.3.1	Esforç i Resultat	5
5.3.2	Causa Efecte	5
5.3.3	Com es mou el mercat	6
5.4	VSA	7
5.4.1	Barres	7
6	Mercats Utilitzats	8
7	Estratègies plantejades	9
7.1	Predictió del valor de tancament	9
7.2	Predictió per classificació d'Up Bars	9
8	Resultats	10
8.1	Primer enfocament	10
8.1.1	Estudi amb les dades originals	10
8.1.2	Diferencies	11
8.1.3	Identificació de les NS	11
8.1.4	Identificació de les ND	12
8.1.5	Interpretació genèrica dels resultats	12
8.2	Segon Enfoc	13
8.2.1	Anàlisi de les dades	13
8.2.2	Classificacions	14
9	Interpretació dels Resultats	17
10	Bibliografia	19
11	Annex	20
11.1	Predictions del primer enfoc amb les dades originals	20
11.2	Predictions del primer enfoc amb les diferencies	23
11.3	Predictions del primer enfoc amb les veles NS	26
11.4	Predictions del primer enfoc amb les veles ND	29

1 Introducció

En l'actualitat, les intel·ligències artificials (IA) i l'aprenentatge automàtic (AA) s'han transformat de tecnologies emergents a pilars essencials en el nostre entorn quotidià i en sectors clau de l'economia global. Aquestes potents eines es troben des de sistemes d'assistència virtual fins a sofisticats algorismes predictius utilitzats en àmbits tan diversos com finances, salut, comerç electrònic i manufactura. La seva capacitat per automatitzar tasques complexament humanes, millorar la presa de decisions basades en dades massives i accelerar el progrés científic i tecnològic les fa imprescindibles per al futur dels negocis i la investigació.

Un dels aspectes més destacats de l'IA i l'AA és la seva capacitat per aprendre i adaptar-se a partir de grans volums de dades. Aquesta habilitat permet que els sistemes es tornin cada vegada més precisos i eficients a mesura que es processen més dades, una característica que ha revolucionat sectors com la medicina, on els models d'IA poden ajudar a diagnosticar malalties amb una precisió comparable o superior a la dels professionals mèdics. En el sector financer, els algorismes d'AA són capaços d'analitzar tendències del mercat en temps real, facilitant la presa de decisions estratègiques i la gestió de riscos.

Els avantatges de l'IA i l'AA es manifesten en la seva capacitat per analitzar i processar enormes quantitats de dades amb una precisió i velocitat impossibles pels humans. Això no només ha revolucionat la manera en què es gestionen les operacions diàries, sinó que també ha obert noves fronteres en la personalització dels serveis, la millora de la qualitat de vida i la innovació en productes. Per exemple, en el comerç electrònic, els sistemes de recomanació basats en AA poden oferir experiències de compra altament personalitzades, augmentant la satisfacció del client i els ingressos per a les empreses. En la indústria manufacturera, l'ús de robots intel·ligents i sistemes de manteniment predictiu ha millorat l'eficiència operativa i reduït els costos.

No obstant això, mentre que l'impacte positiu de l'IA i l'AA és evident, també s'ha incrementat la complexitat i la necessitat de comprendre millor com les dades que alimenten aquestes tecnologies afecten les seves pròpies prediccions i sortides. La qualitat de les dades, la seva quantitat i la seva naturalesa són factors crucials que poden influir en gran mesura en el rendiment dels models d'AA. Per exemple, dades esbiaixades o incompltes poden portar a prediccions errònies, la qual cosa pot tenir conseqüències greus en aplicacions sensibles com la salut o les finances.

Aquest treball de recerca s'endinsa en com la qualitat i naturalesa de les dades utilitzades per a l'entrenament dels models d'AA poden influir de manera significativa en la fiabilitat i la precisió de les seves prediccions. A través d'una anàlisi detallada i empírica, s'examina com les decisions de disseny de dades impacten en la robustesa dels models, amb un enfocament particular en sectors com ara la inversió financer, on les prediccions exactes poden tenir conseqüències significatives.

2 Motivació

El món de la borsa és un món molt interessant, a la vegada que complicat. Cada mercat es resumeix en un gràfic que mostra l'evolució del preu per intervals de temps, formant estructures i patrons que molts intenten estudiar per mirar de treure'n el màxim profit

possible. Aquests patrons segueixen una base teòrica que permet a les persones que s’hi dediquen a estudiar el comportament d’aquests gràfics i mirar de predir com evolucionaran aquests valors de cara al futur, però per poder fer això, cal dedicar-hi molt de temps d’estudi i pràctica per poder entendre la situació que s’està plantejant i, sobretot, agilitzar la presa de decisions per treure’n el màxim benefici.

Aprofitant aquest moment d’esplendor de l’aprenentatge automàtic, es pregunta si un model és capaç d’entendre la situació del mercat que l’està plantejant, i si és capaç de predir l’evolució dels valors al llarg del temps. A més, es vol estudiar si un model és capaç de treure’n profit a la base teòrica que fan servir els professionals per determinar les tendències del mercat, o si ignorarà aquestes noves dades.

3 Objectius

Els objectius d’aquest estudi és determinar fins a quin punt un model d’aprenentatge automàtic és capaç d’entendre el comportament d’un mercat i fer prediccions valuoses dels diferents valors dels gràfics. A més, es pretén estudiar si un model és capaç de millorar aquestes prediccions inicials fent servir una base teòrica sobre els diferents patrons que poden tenir els gràfics de mercat.

3.1 Objectius Específics

Els objectius més específics d’aquest projecte són:

1. **Estudiar i entendre les bases teòriques** que la gent professional i inversors utilitzen per estudiar els gràfics dels mercats on treballa.
2. **Saber transmetre aquests coneixements apresos** en el punt anterior en un llenguatge de programació, en aquest cas Python.
3. **Poder provar un model d’aprenentatge automàtic** en diferents mercats, fer un estudi sobre en quin d’ells s’adapta millor i entendre el perquè.
4. **Poder validar els resultats obtinguts.**

4 Plantejament del projecte

El projecte no només busca poder entrenar un model que faci prediccions sobre l’evolució del mercat, sinó també poder solidificar coneixements bàsics sobre l’anàlisi avançada del mercat de valors. Així doncs, el pla de treball d’aquesta recerca és el següent:

1. **Establir uns coneixements teòrics** bàsics sobre borsa per poder establir estratègies basades en conceptes ja provats, i agilitzar el procés d’entendre les dades amb les quals es treballaran.
2. **Seleccionar un mercat** i un període d’aquest que permeti posar en pràctica les estratègies estudiades, així com un període amb una volatilitat baixa que faciliti el màxim possible un aprenentatge per part del model.
3. **Entrenar diferents models** basats en estratègies diferents, de més simples a més complexes, per tal de comprovar si existeix una millora. Es partirà des d’un model

molt simple i s'aniran afegint components més avançats, amb models més complexes o considerant les bases teòriques estudiades, per estudiar si el rendiment del model és millor que la versió anterior.

4. **Validar els resultats** fent servir les dades reals del mercat seleccionat, o fent servir estratègies pròpies d'Aprenentatge automàtic que permetin avaluar la qualitat dels resultats obtinguts.

5 Anàlisi Avançada del Mercat de Valors

El mercat de valors es caracteritza per un flux continu i dinàmic de dades que canvia constantment a causa de la influència de múltiples variables. Comprendre l'evolució d'aquests valors i els patrons que segueixen és fonamental per a aquelles persones que desitgen dedicar-se a la borsa i obtenir-ne un benefici. Per aquest motiu, economistes i analistes han estudiat durant dècades els gràfics i les dades del mercat amb l'objectiu de consolidar una base teòrica sòlida que els permeti realitzar prediccions amb el màxim nivell de precisió possible.

Per analitzar el mercat de valors i prendre decisions informades, existeixen diverses tècniques avançades que els analistes utilitzen per comprendre els patrons i tendències dels preus. Aquestes tècniques es divideixen principalment en dues categories: l'anàlisi tècnica i l'anàlisi fonamental.

L'anàlisi tècnica se centra en l'estudi dels moviments històrics dels preus i els volums de negociació per predir els futurs moviments del mercat. Aquesta tècnica utilitza gràfics i indicadors tècnics per identificar patrons repetitius i tendències del mercat. Entre les metodologies més avançades, destaca l'Anàlisi de la Distribució del Volum (VSA) i les estructures de Wyckoff. La VSA analitza la relació entre el volum de transaccions i l'acció del preu per detectar els moviments dels grans operadors del mercat, mentre que les estructures de Wyckoff, ofereixen un marc per comprendre les fases del mercat, com l'acumulació i la distribució, i identificar els patrons de comportament dels preus.

Dins del món de la borsa és important destacar als anomenats professionals, els quals són aquells inversors experimentats i ben informats que operen amb grans volums de diners i que sovint tenen accés a informació privilegiada i recursos analítics avançats. Aquests professionals inclouen gestors de fons, traders institucionals, analistes financers i altres operadors de mercat que influeixen significativament en les dinàmiques del mercat de valors a través de les seves decisions d'inversió i operacions de trading.

5.1 Oferta i Demanda

El preu es mou a causa de la llei d'oferta i demanda, en la qual s'assumeix que a major oferta que demanda, el preu baixarà i viceversa. També és important destacar que el preu pujarà si hi ha falta d'oferta, i baixarà si hi ha falta de demanda. En un moviment ascendent en el gràfic del preu, el més normal és que continuï pujant, perquè les barres probablement indiquen que hi ha poca oferta. Només girarà en el moment en què les barres indiquin que hi ha un augment de l'oferta, i és doncs quan començarà a baixar.

5.2 Volum

El volum és el nombre d'operacions tancades en aquell *time frame* determinat. També es pot entendre com els diners que s'han mogut en aquell moment, i és el que provoca el canvi en les barres.

Si es detecta que el valor és molt alt, però el preu no puja en concordança, es pot suposar que part del volum són compres, i que l'oferta és molt alta. En canvi, si el preu sí que puja amb relació al volum, aleshores es pot suposar que hi ha molta demanda.

5.3 Estructures de Wyckoff

Les estructures de Wyckoff són un conjunt de patrons i principis desenvolupats per Richard D. Wyckoff a principis del segle XX per analitzar el comportament dels preus en els mercats financers. Wyckoff defensa que totes les fluctuacions en el mercat i en totes les accions haurien d'estudiar-se com si fossin el resultat de les operacions d'un únic operador. Es treballen amb tres regles fonamentals:

1. **Llei d'oferta i demanda:** si hi ha més oferta que demanda, el preu baixarà. I al revés.
2. **Causa i efecte:** el lateral és el que marca la tendència. Les estructures de Wyckoff mostren com evolucionarà el preu més endavant. La mida del lateral és directament proporcional al canvi de tendència posterior.
3. **Esforç i resultat:** l'esforç realitzat en un moviment ha d'estar en concordança amb el resultat obtingut. Un volum molt elevat hauria de provocar un canvi en el preu.

5.3.1 Esforç i Resultat

El moviment del preu ha d'estar en harmonia amb el volum que l'ha fet moure. Es diu que hi ha harmonia si en un moviment on el volum és molt elevat, el preu continua pujant o baixant. No hi ha harmonia en cas contrari.

A més, un *Up Bar* petit amb molt volum indica molta oferta. Això és perquè s'ha fet molt esforç per pujar el valor del preu, però no s'ha assolit l'objectiu. De la mateixa manera, una *Down Bar* petita amb un volum molt elevat, implica molta demanda.

5.3.2 Causa Efecte

Un lateral és la causa i la tendència posterior a l'efecte. A major duració i volum en la causa, major serà l'efecte.



Figura 1: Exemple de dues estructures de Wyckoff amb volums diferents

En aquest exemple es pot veure dues estructures de Wyckoff diferents. La diferència entre totes les franges està en el volum i la longitud de l'estructura. Això es tradueix en un canvi molt més pronunciat en la segona que en la primera estructura.

5.3.3 Com es mou el mercat

La cotització dels actius financers es mou en funció dels interessos dels diners professionals. Aquest realitza processos d'acumulació per comprar la quantitat més gran al menor preu possible, i la distribució per vendre aquests actius que s'han comprat en l'acumulació al major preu possible. Aquests processos es poden resumir en el següent gràfic.

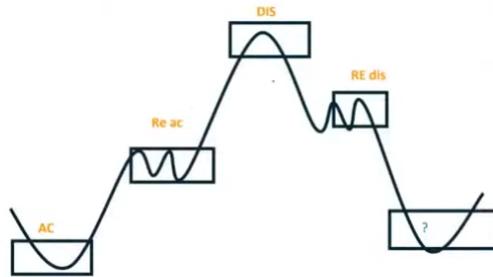


Figura 2: Processos d'acumulació, distribució, re-acumulació i redistribució

Aquestes franges consisteixen en:

- **Acumulació:** comprar el màxim possible de l'actiu sense que el preu vagi en contra.
- **Distribució:** vendre el màxim possible de l'actiu sense que el preu vagi en contra.
- **Re-acumulació:** procés d'acumulació entre l'anterior i el de distribució.
- **Redistribució:** procés de distribució posterior a l'anterior.

El mercat es mou sempre amb aquests processos. La dificultat està a saber amb certesa si es tracta d'una acumulació o d'una distribució.

5.4 VSA

L'Anàlisi de la Distribució del Volum (Volume Spread Analysis, VSA) és una metodologia que examina la relació entre el volum de transaccions i la variació del preu per identificar les accions dels grans operadors del mercat. Mitjançant l'estudi d'aquests patrons, els analistes poden anticipar els moviments futurs dels preus, entenent millor les fases d'acumulació i distribució. El que defensa aquesta estratègia és que la realitat de la companyia no importa, ni els seus resultats econòmics. Només és important el que el conjunt dels inversors, i especialment els professionals, opinen de la seva cotització. Només importa el preu de l'actiu.

VSA consisteix a llegir barra a barra i detectar l'oferta i la demanda d'un actiu en zones concretes, per així saber quina és la direcció més probable d'un actiu. Així doncs, aquesta estratègia permet preveure les maniobres dels diners professionals abans que comenci o continui una tendència, o hi hagi un gir de mercat. La informació rellevant doncs és:

- El volum corresponent a la barra en la qual estem.
- El rang de la barra.
- Relació amb les barres anteriors.
- Zona del gràfic on està la barra.

Això és per la segona llei de WYCKOFF, la qual diu que un volum molt gran implica un gran desplaçament vertical en les dades, mentre que un volum molt petit, es tradueix en un desplaçament més petit.

5.4.1 Barres

Una barra és un element gràfic que representa el preu d'una acció durant un període de temps específic. Aquesta barra mostra el preu d'obertura, el preu de tancament, el màxim i el mínim del període, oferint una visió completa de la fluctuació del preu en aquest interval. D'aquesta informació és interessant:

- **La relació entre el volum i el rang de la barra.** L'efecte que té el volum sobre la barra on està.
- **El tancament de la barra respecte a l'anterior.** Es diferencia entre *Up Bar* (barra que tanca per sobre de l'anterior) i *Down Bar* (barra que tanca per sota de l'anterior).
- **Zona del gràfic i el context en què es troba la barra.** Entendre el context de la zona on es troba la barra que s'està estudiant.

VSA classifica les veles segons aquestes característiques anteriors. Aquest projecte se centra en dues:

- *No Supply Candles* (NS): es tracta d'una *Down Bar* amb menys volum que les dues barres anteriors. En aquest cas, el preu hauria de pujar després d'aquesta barra, ja que la demanda supera l'oferta durant aquest període.
- *No Demand Candles* (ND): es tracta d'una *Up Bar* amb menys volum que les dues barres anteriors. En aquest cas, el preu hauria de baixar després d'aquesta barra, perquè l'oferta supera la demanda en durant aquest període.

Les ND i les NS en certs punts del gràfic ens indiquen que els diners professionals volen continuar amb el moviment, i serveixen per comprovar que la tendència dels valors segueix el moviment que volen generar.

6 Mercats Utilitzats

La selecció dels mercats d'aquest projecte ha estat un punt clau en el desenvolupament de l'estudi. El que es buscava era el següent:

- **Baixa volatilitat en les dades:** tenir un conjunt de dades amb baixa volatilitat era clau pel desenvolupament del projecte, perquè facilita al model el seu aprenentatge.
- **Menor nombre de NULLS possible.**
- **Trobar estructures de Wyckoff:** seleccionar un mercat amb estructures de Wyckoff és important perquè permet identificar fases clau d'acumulació i distribució que indiquen moviments potencials de preu.

Després de fer un estudi, s'ha dut a terme la selecció de quatre mercats diferents per poder fer l'estudi posterior i permetre així la comparació dels resultats entre l'un i l'altre.



Figura 3: Amazon



Figura 4: Apple



Figura 5: Netflix



Figura 6: SPX 500

Cada dataset està format per les mateixes mateixes *features*, i cada fila correspon a un *timeframe* de 24 hores. Les característiques amb les que es treballarà inicialment són:

- **Open:** és el preu que té l'*stock* a l'inici del període.

- **Close:** és el preu que té l'*stock* al final del període
- **High:** és el valor màxim de preu al que arriba l'*stock* durant el període.
- **Low:** és el valor mínim de preu al que arriba l'*stock* durant el període.

7 Estratègies plantejades

L'objectiu d'aquest treball és comprovar si un model és capaç de predir l'evolució d'un mercat de valors, i si pot donar-li un valor a estratègies emprades per gent que viu de la compra i venda d'accions com a professió.

Aquesta investigació s'ha dut a terme a partir de dos plantejaments diferents, els quals s'han realitzat a partir de diferents models per facilitar l'extracció de resultats i de conclusions.

7.1 Predicció del valor de tancament

El propòsit d'aquest enfocament ha estat utilitzar models de regressió per predir el valor de tancament (*close*) en el següent *time frame*. Inicialment, s'ha aplicat aquesta estratègia utilitzant les dades originals, i s'ha explorat la possibilitat de millorar aquesta predicció mitjançant l'addició progressiva de noves característiques. Aquestes noves *features* s'han seleccionat basant-se en els fonaments teòrics discutits anteriorment, les quals són:

1. Calcular la diferència absoluta entre l'instant actual i l'anterior. Si c_t és el valor del *close* a l'instant t , es considera la variable $\Delta c = c_t - c_{t-1}$.
2. Identificar les veles NS del conjunt de dades. Després d'aquestes veles, el valor de tancament hauria de ser superior.
3. Identificar les veles NS del conjunt de dades. Després d'aquestes veles, el valor del tancament hauria de ser inferior.

Aquestes noves característiques s'han anat afegint de manera progressiva per poder estudiar si hi ha hagut millora en les prediccions.

7.2 Predicció per classificació d'Up Bars

L'objectiu d'aquest segon plantejament ha estat dur a terme una classificació binària, segons si la vela següent serà una *up bar* o no. Per fer aquest estudi s'ha fet servir la informació predeterminada del conjunt de dades i la informació de si la vela és una NS o una ND, fent les permutacions següents:

1. Utilitzar únicament les dades originals.
2. Incorporar l'etiqueta NS juntament amb les dades originals.
3. Incloure l'etiqueta ND en combinació amb les dades originals.
4. Combinar les dades originals amb les etiquetes NS i ND.

Fer l'estudi amb les quatre permutacions permet avaluar si el model de classificació realment és capaç de donar-li un valor a la nova informació.

8 Resultats

En aquesta secció es durà a terme l'explicació del procediment i de l'evolució dels resultats pels diferents enfocaments que ha plantejat aquest projecte. Abans de dur a terme cap estudi, però, és molt important normalitzar les dades per millorar el rendiment dels diferents models que s'implementaran durant l'estudi.

8.1 Primer enfocament

Per aquest primer enfocament s'han utilitzat models de regressió per fer la predicción del valor de tancament del següent interval de temps. Els models de regressió que s'han fet servir han estat:

- Random Forest amb 1000 arbres de decisió.
- Decision Tree.
- XGBoost.
- Xarxa Neuronal LSTM.

Fer servir quatre models per cada etapa de l'enfocament permetrà estudiar millor com les diferents *features* van afectant la qualitat de les prediccions. Aquesta qualitat es mesura quantitativament amb les mètriques MAE i RMSE, i qualitativament comparant les prediccions generades amb el conjunt de validació de les dades.

8.1.1 Estudi amb les dades originals

El primer que s'ha fet és dur a terme la predicción amb les dades originals per poder anar comparant si existeix la millora en anar afegint les dades. Així doncs, s'apliquen els quatre models mencionats amb les característiques *open*, *close*, *high* i *low*, i s'intentarà predir qui és el valor de tancament del següent interval de temps. Els valors del *MAE* i *RMSE* són:

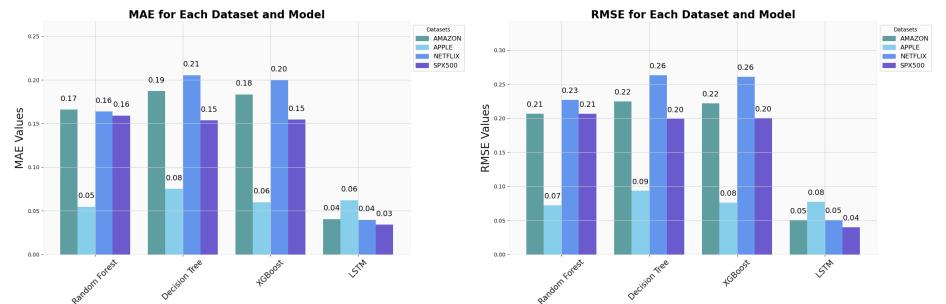


Figura 7: Resultats del MAE i RMSE amb les dades originals

D'aquests gràfics es destaca que el conjunt de dades d'Apple és el que millor s'ajusta a tots els models, i el model que millor rendiment dona és l'LSTM, fet que es veu reflectit també en com es dibuixa la corba de les prediccions sobre les dades de validació, com es pot veure en la figura 44. Així com el model LSTM les prediccions i les dades de validació són molt similars, la resta de models no s'ajusta bé als canvis bruscos que presenten les dades, dibuixant en molts casos una línia totalment horitzontal respecte a l'eix de les x. Un exemple d'això podrien ser les imatges 30 i 32. En la primera imatge es pot veure

com el model no ajusta gens bé la volatilitat de les dades i acaba dibuixant una línia horitzontal, mentre que en la segona el model és capaç de predir el moviment del mercat.

8.1.2 Diferencies

Un cop obtinguts els resultats amb les dades originals, s'intenta millorar el rendiment dels models afegint la diferència del valor de close entre el moment actual i l'anterior. Els resultats són els següents:

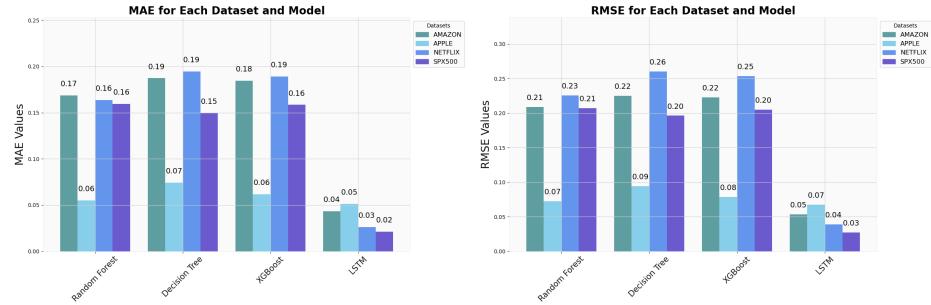


Figura 8: Resultats del MAE i RMSE afegint la característica de diferencies

Els resultats després d'afegir la característica de les diferències no millora respecte a els resultats anteriors 7. Destaca, però, una millora tant en l'MAE i el RMSE en el Decision Tree en el conjunt de dades de Netflix. Però, si es compara com afecta aquest canvi en el plot de les prediccions en les imatges 38 i 54 es pot veure com, així i tot, els resultats obtinguts no són gens bons. Tot i que el rendiment de la LSTM sobre el conjunt de dades Apple sigui més dolent que el cas anterior, aquest continua destacant per tenir un millor ajust en totes les dades, malgrat que en el cas d'Apple els models Random Forest i XGBoost donin millors resultats. Si es comprova com es dibuixen les dades i les prediccions a les imatges 49, 51 i 52 es pot veure com el model que realitza un millor ajust a les dades és el model de Random Forest, però en el cas de la LSTM es veu com el model ha entès millor la volatilitat, ja que en l'àmbit visual els resultats són més similars tot i que estiguin dibuixats per sota de la corba.

8.1.3 Identificació de les NS

La següent característica que s'ha afegit al model és un valor boleà que indica si la vela corresponent és una NS o no, és a dir, si es tracta d'una *down bar* el volum de la qual és inferior a les dues veles anteriors. Afegint aquesta *feature*, s'han obtingut els resultats següents:

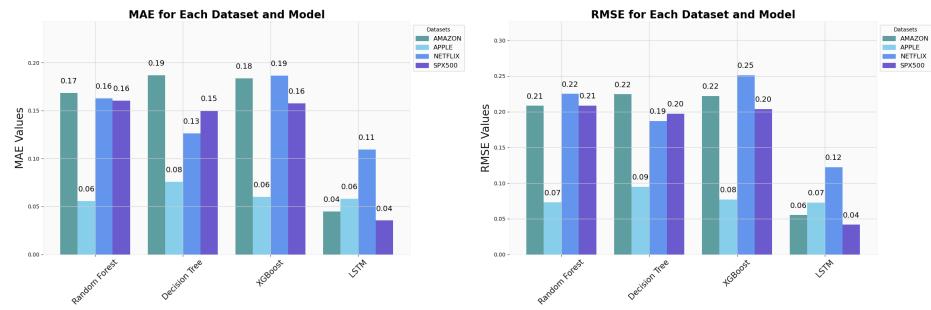


Figura 9: Resultats del MAE i RMSE afegint la característica NS

Els resultats després d'afegir la nova característica, com a normal general, no milloren el rendiment del model. En el cas del Decision Tree amb el dataset de Neftlix es perd la millora que s'obté en els resultats anteriors 8. La resta de models continuen comportant-se de la mateixa manera que el cas anterior, sent l'LSTM qui millor s'adapta a la volatilitat de les dades, i sent el conjunt de dades Apple qui millors resultats presenta en tots quatre models.

8.1.4 Identificació de les ND

Finalment s'afegeix una última característica per valorar si hi ha o no millora en el model. Per acabar amb aquest plantejament es indiquen les veles ND, és a dir es marca una *up bar* amb menys volum que les dues veles anteriors. Els resultats són els següents:

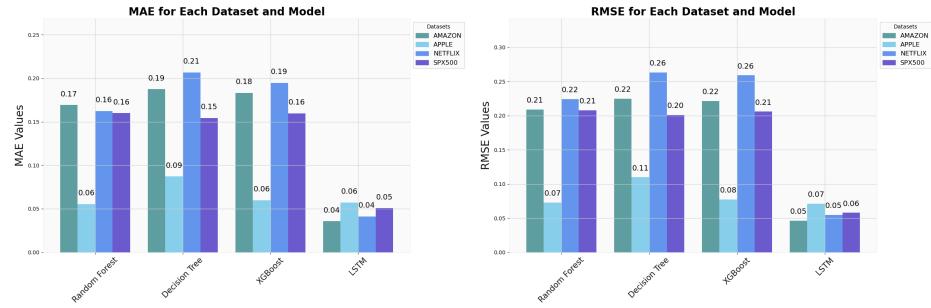


Figura 10: Resultats del MAE i RMSE afegint la característica ND

Els resultats tampoc canvien en general després d'aplicar aquesta característica. Destaca el fet que Decision Tree amb el dataset de Netflix torna a corregir l'increment de l'error que s'obté en el grafic anterior 9. El dataset de Apple torna a donar millor rendiment amb els models Random Forest i XGBoost que amb la xarxa LSTM, però com a norma general no es pot detectar cap millora significativa respecte els resultats anteriors.

8.1.5 Interpretació genèrica dels resultats

El model que ha donat un millor resultat ha estat el model LSTM, mentre que el conjunt de dades que ha estat més fàcils pels models ha estat l'Apple. En el cas d'aquest últim dataset, es veu com tots els models donen un rendiment acceptable i que, més o menys, prediuen com les dades estan evolucionant al llarg del temps. En canvi, amb la resta de conjunts de dades, els models Random Forest, Decision Tree i XGBoost no han estat

capaços de dur a terme bones prediccions en les dades.

Si fixem el model LSTM, es pot veure com els resultats segueixen prou bé l'evolució de les dades, però en tots els casos tendeix a dibuixar la línia per sota de les dades originals. Això podria ser a causa del soroll que pot haver-hi en les dades.

8.2 Segon Enfoc

En aquest segon enfocament s'han fet servir models de classificació per predir si la vela posterior a l'actual té creixement positiu o no. Els models que s'han implementat han estat els següents:

- Classificador Random Forest amb 1000 arbres de decisió.
- Logistic Regression.
- Support Vector Machine.
- K-Nearest Neighbors.
- Gradient Boosting.
- Naive Bayes.
- Xarxa Neuronal LSTM.

Altra vegada s'ha estudiat quina és la millora en anar variant les *features* fent servir diferents models. Per cada conjunt de dades i per cada model, s'ha calculat l'àrea per sota la corba ROC per poder validar si el model està aprenent a partir de les noves característiques, i també per extreure el valor del *threshold* òptim. Amb aquest valor, es calcula la proporció de veles NS i ND que tenen un comportament esperat segons les estratègies VSA.

L'experiment s'ha realitzat duent a terme les permutacions mencionades amb anterioritat.

8.2.1 Anàlisi de les dades

Abans de realitzar provar de classificar les veles amb els models, s'ha dut a terme una anàlisi de les dades per poder estudiar quina és la proporció de veles NS i ND que hi ha per cada conjunt de dades; i, per aquestes veles en concret, quantes d'elles segueixen el comportament esperat que marca VSA. Aquesta informació es recull en la taula següent:

Dataset	Total	NS	ND	Proporció Upbar NS	Proporció Upbar ND
Amazon	251	60	33	0.55	0.545
Apple	293	68	37	0.382	0.649
Netflix	293	67	40	0.485	0.525
SPX500	293	25	83	0.68	0.530

Taula 1: Conteig de les veles NS i ND sobre les veles totals i càcul de la proporció de UpBars posteriors a elles

Aquesta taula mostra el nombre total de veles que hi ha per cada conjunt de dades d'aquest

projecte i quantes d'elles han estat classificades com a NS i ND. Les columnes "Proporció UpBar NS", "Proporció UpBar ND" mostren quantes d'aquestes barres tenen a continuació una barra ascendent. La teoria d'VSA marca que després d'una barra NS el preu hauria de pujar, mentre que després d'una ND el preu hauria de baixar. Així doncs, si es calcula la proporció d'UpBars, s'esperaria que per les NS els valors fos molt proper a 1, mentre que d'una ND valors siguin propers al 0. Però el recompte indica tot el contrari, portant a valors propers al 0.5 per ambdues veles. Així doncs, conèixer si es tracta d'una NS o d'una ND no és suficient per determinar si el preu pujarà o no, si no cal entendre i estudiar molt bé el context que envolta aquesta vela per determinar quina serà l'evolució del preu.

8.2.2 Classificacions

Un cop realitzat el recull de les dades de la taula 1, es comença a fer la classificació de les veles amb els models esmentats a cadascuna de les permutacions.

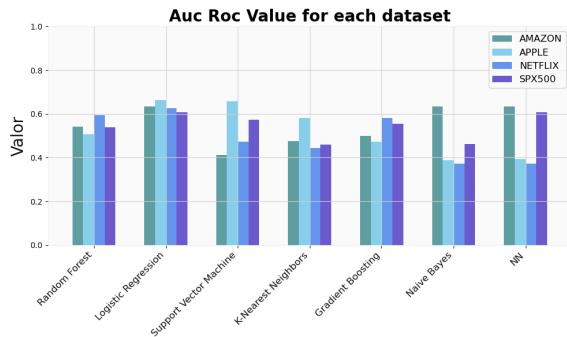


Figura 11: Dades originals

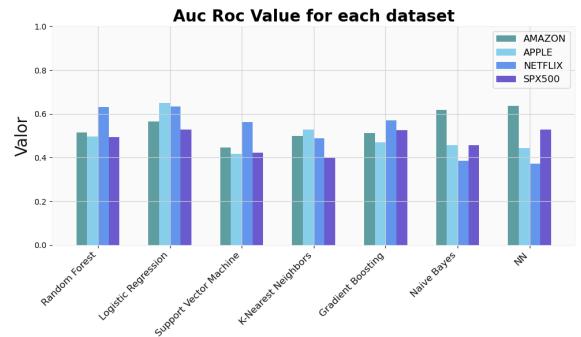


Figura 12: Veles NS

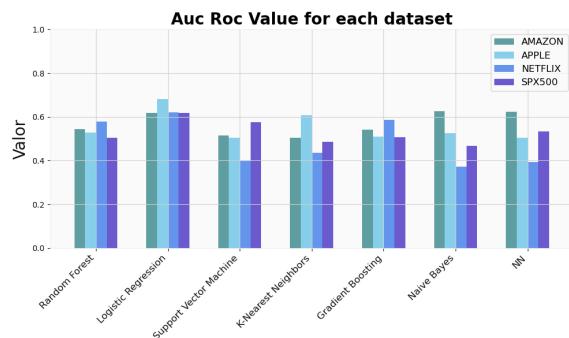


Figura 13: Veles ND

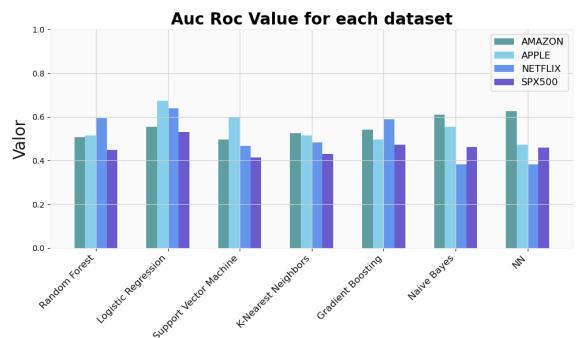


Figura 14: Veles NS i ND

Després d'entrenar els models amb totes les permutacions, es pot veure com cap de les àrees per sota de la corba millora respecte a les altres. Això indica que cap dels models és capaç de predir millor a partir de les noves característiques que s'ofereixen per fer l'entrenament.

El següent que es vol estudiar és quins resultats s'han obtingut en les NS i en les ND. És a dir, quina és la *accuracy* del model mirant únicament les prediccions sobre aquestes veles.

Per dur a terme aquest estudi, s'ha filtrat les dades dels conjunts de dades únicament per aquests valors, i s'han comparat les prediccions obtingudes pels valors reals. Els resultats són els següents:

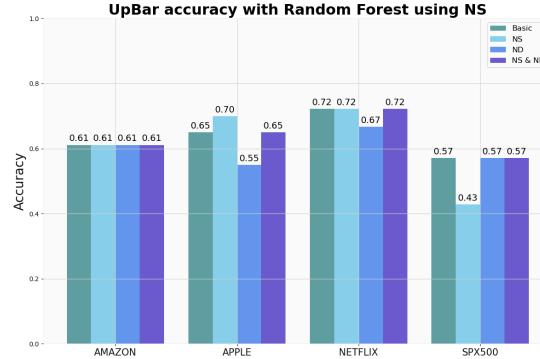


Figura 15: Random Forest sobre NS

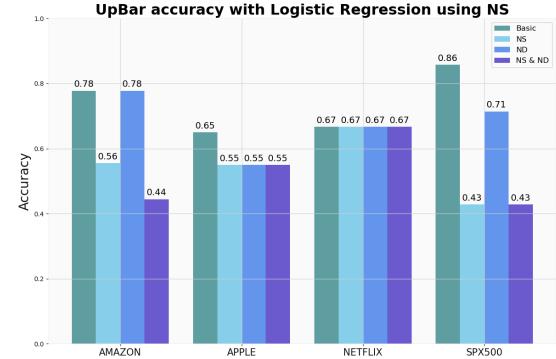


Figura 16: Logistic Regression sobre NS

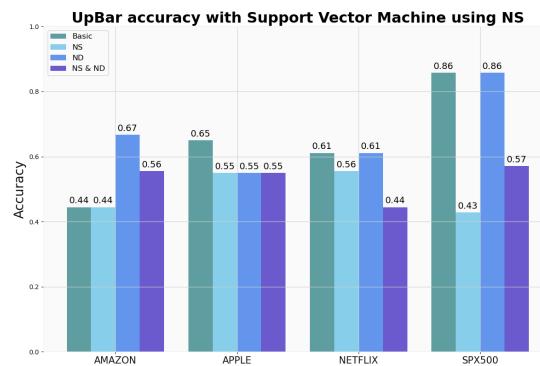


Figura 17: SVM sobre NS

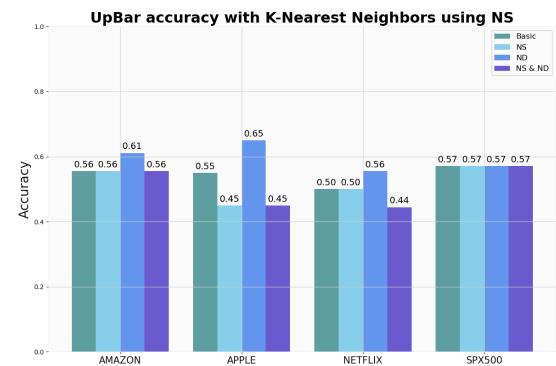


Figura 18: K-Nearest Neighbors sobre NS

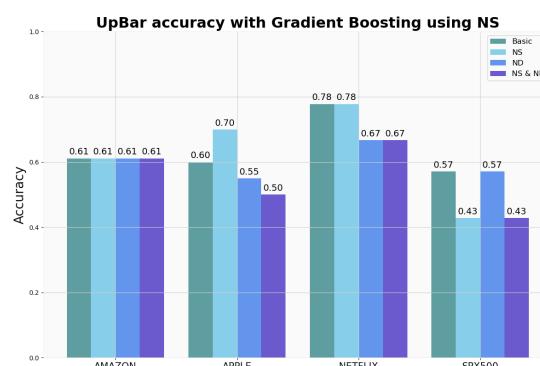


Figura 19: Gradient Boosting sobre NS

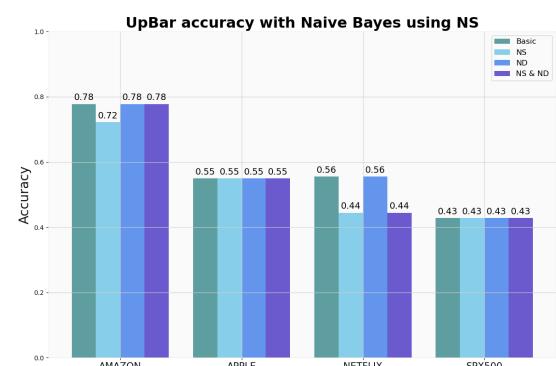


Figura 20: Naive Bayes sobre NS

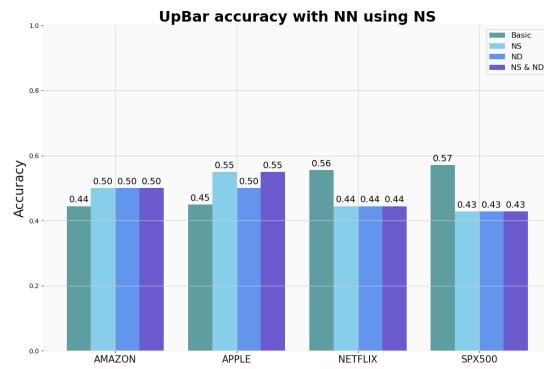


Figura 21: LSTM sobre NS

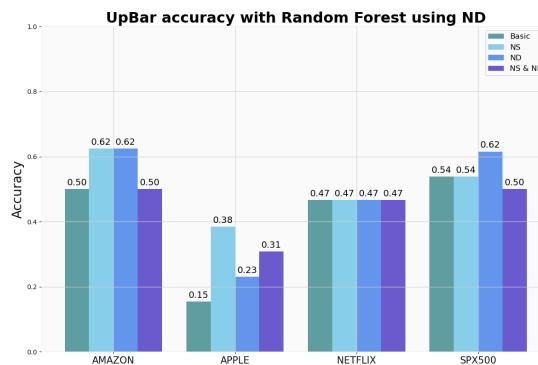


Figura 22: Random Forest sobre ND

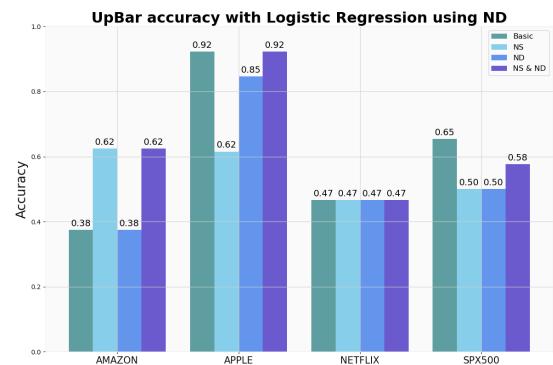


Figura 23: Logistic Regression sobre ND

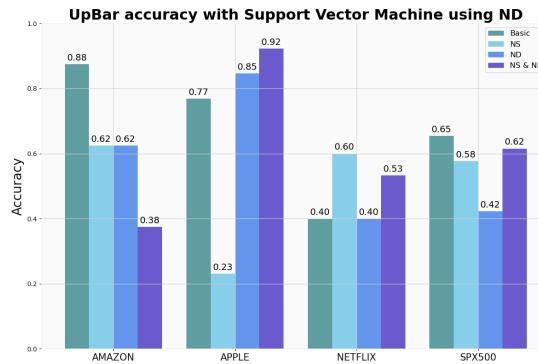


Figura 24: SVM sobre ND

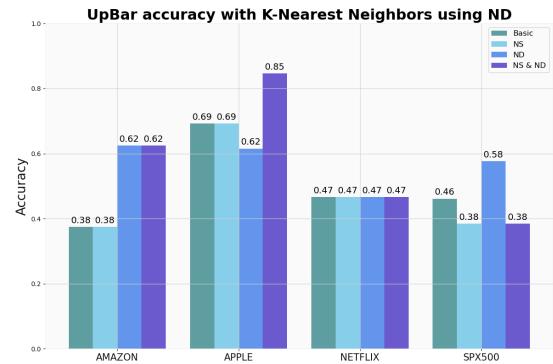


Figura 25: K-Nearest Neighbors sobre ND

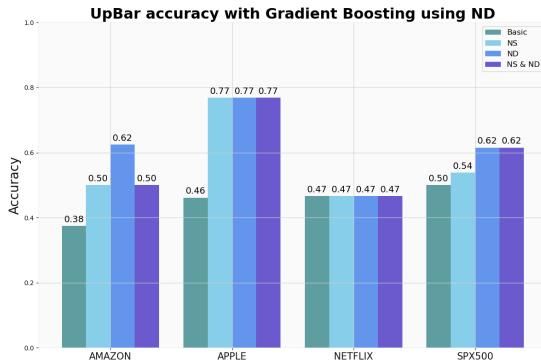


Figura 26: Gradient Boosting sobre ND

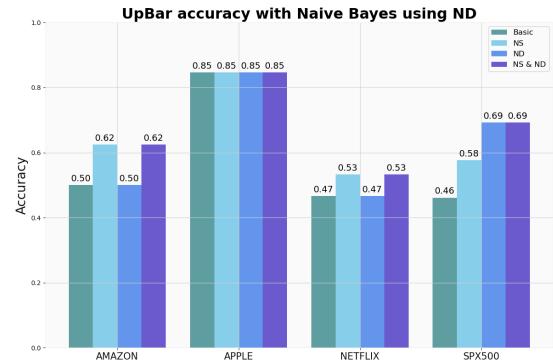


Figura 27: Naive Bayes sobre ND

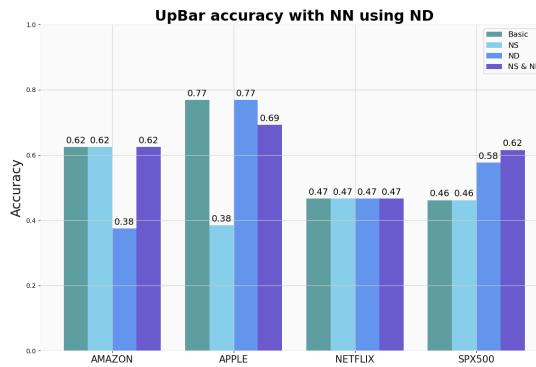


Figura 28: LSTM sobre ND

Les *accuracies* són molt variades en tots dos casos. Així com en l'anterior enfocament hi havia un model que predominava sobre la resta, en aquest es pot veure com el rendiment de tots els models és similar. Els resultats varien molt en funció del model i del conjunt de dades amb el qual s'estigui treballant. Tot i que es poden detectar diferents pics d'*accuracies* superiors al 70% que podrien mostrar que el model està interpretant correctament el context del mercat i analitzant les veles segons les bases teòriques que indica VSA; també hi ha casos on l'*accuracy* és del 15%. Encara que el rang d'*accuracies* sigui tan gran, sembla que tots els models interpreten millor les NS que no pas les ND, ja que quan es té en compte només el primer tipus de veles, els valors de precisió no són tan baixos com el segon tipus.

9 Interpretació dels Resultats

Els resultats que s'han obtingut en aquest treball són molt variats. S'han plantejat dos enfocaments diferents per poder comparar quin dels dos dona millor rendiment o quin dels dos es beneficia més de les noves *features* que es generen.

Pel que fa al primer enfocament es pot afirmar que no hi ha cap millora significativa respecte al model inicial i el final. El model no li dona gens d'importància ni cap valor a les noves característiques. La diferència de l'error entre les característiques inicials i finals massa baixa per a determinar que el model està traient un profit a aquesta informació, i per tant es pot concloure que aquesta informació ha resultat irrelevants pel model.

En el segon enfocament tampoc s'ha pogut detectar cap millora en el rendiment del model fent servir les característiques de les veles que ofereix VSA. L'*accuracy* de totes quatre permutacions varia en funció del model i del conjunt de dades, però cap d'elles segueix un patró de millora a mesura que va disposant de més característiques, és per això que es conclou que el model ha classificat tenint en compte altres paràmetres diferents. Això mateix es pot veure quan s'analitza en detall com el model prediu les veles NS i les ND. Les prediccions sobre aquestes veles varien molt, des de valors de 0.86 a 0.15. Tenint els resultats anteriors, no es creu que el model hagi estat capaç de beneficiar-se del que indica VSA, i ha fet servir altres valors per dur a terme les seves prediccions.

Per tant, considerant tots dos enfocaments, es pot concloure que cap dels models ni cap dels enfocaments ha estat capaç de predir els seus resultats fent servir les bases teòriques que fan servir els professionals i els inversors per prendre les seves decisions. Cada model ha actuat de manera diferent segons les dades i el dataset amb el qual estava treballant, però la tendència a mesura que s'anaven afegint més característiques no ha estat positiva en tots els casos.

10 Bibliografia

- [1] Nasdaq. Descàrrega del dataset SPX500.
- [2] Yahoo! finance historical data. Descàrrega del dataset AMAZON.
- [3] Yahoo! finance historical data. Descàrrega del dataset APPLE.
- [4] Yahoo! finance historical data. Descàrrega del dataset NETFLIX.
- [5] Jordi Martí. Qué es wyckoff y para qué se utiliza. invertir en bolsa de 0 a 50, 15 noviembre 2022.
- [6] Jordi Martí. Qué es vsa y para qué se utiliza en trading. cómo invertir en bolsa, 22 noviembre 2022.

11 Annex

11.1 Prediccions del primer enfoc amb les dades originals

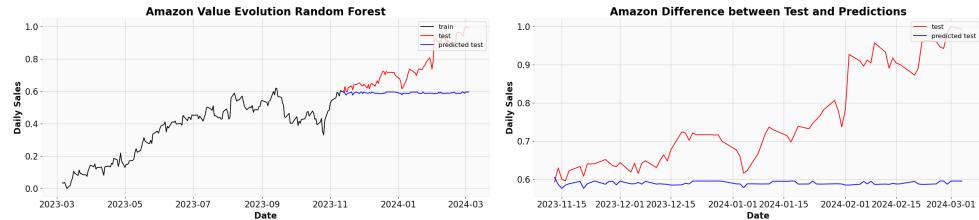


Figura 29: Prediccions del dataset Amazon amb Random Forest i Comparació ampliada

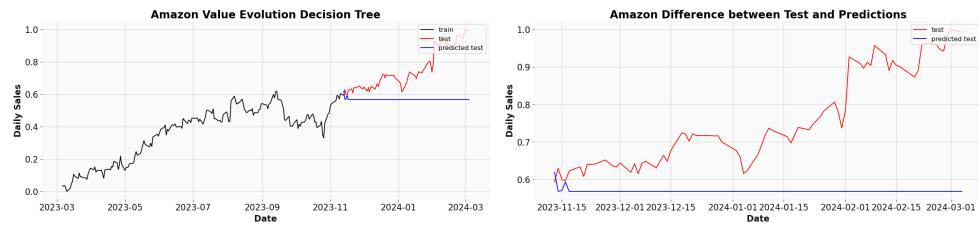


Figura 30: Prediccions del dataset Amazon amb Decision Tree i Comparació ampliada

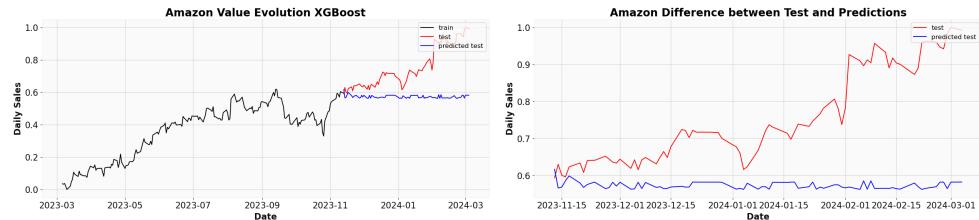


Figura 31: Prediccions del dataset Amazon amb XGBoost i Comparació ampliada

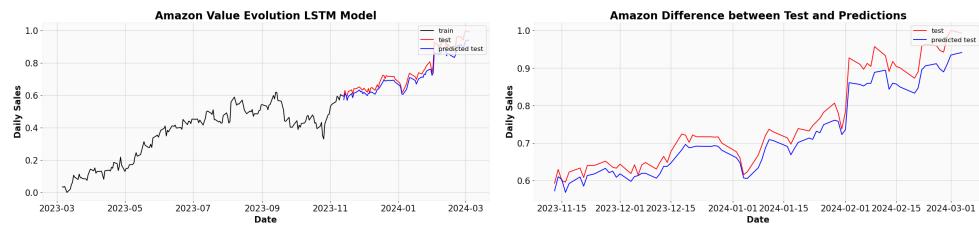


Figura 32: Prediccions del dataset Amazon amb LSTM i Comparació ampliada

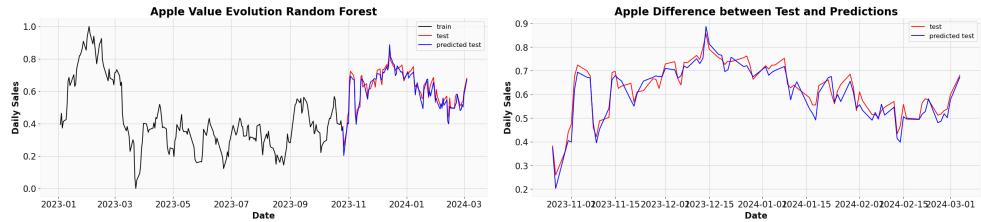


Figura 33: Prediccions del dataset Apple amb Random Forest i Comparació ampliada

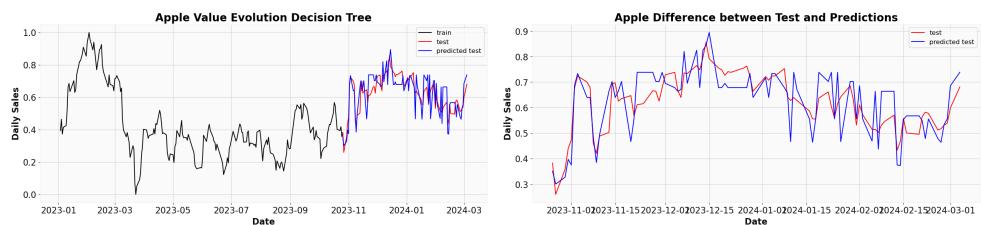


Figura 34: Prediccions del dataset Apple amb Decision Tree i Comparació ampliada

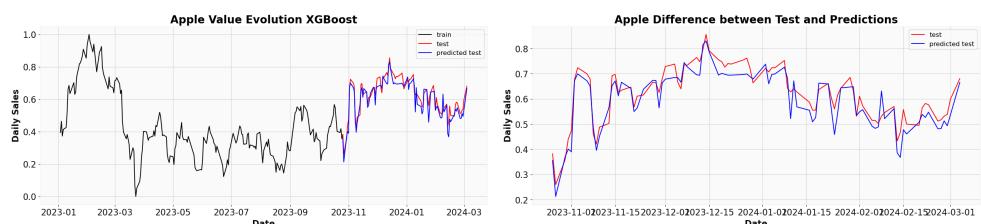


Figura 35: Prediccions del dataset Apple amb XGBoost i Comparació ampliada

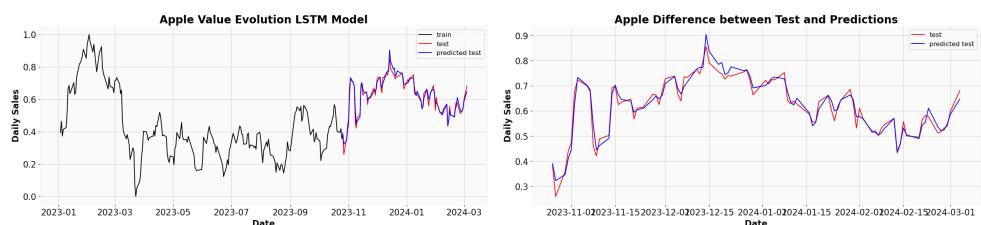


Figura 36: Prediccions del dataset Apple amb LSTM i Comparació ampliada

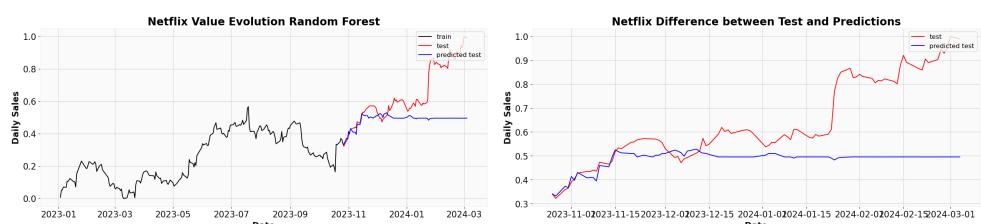


Figura 37: Prediccions del dataset Netflix amb Random Forest i Comparació ampliada

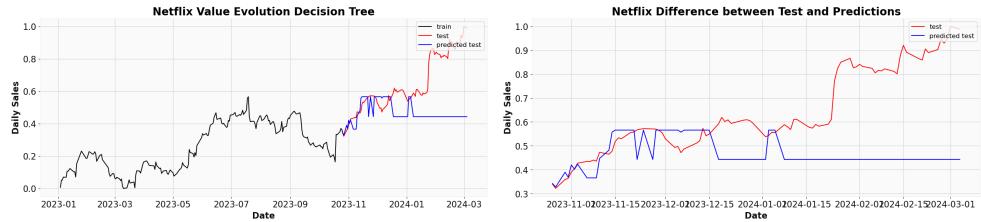


Figura 38: Prediccions del dataset Netflix amb Decision Tree i Comparació ampliada

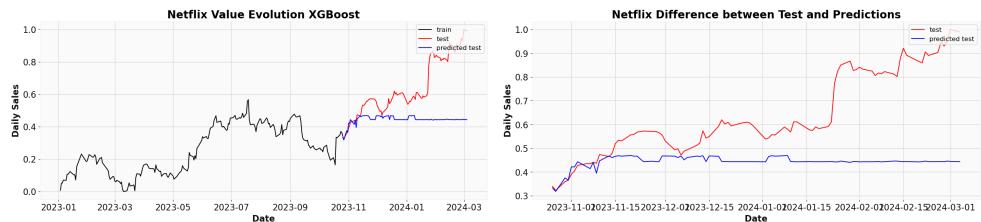


Figura 39: Prediccions del dataset Netflix amb XGBoost i Comparació ampliada

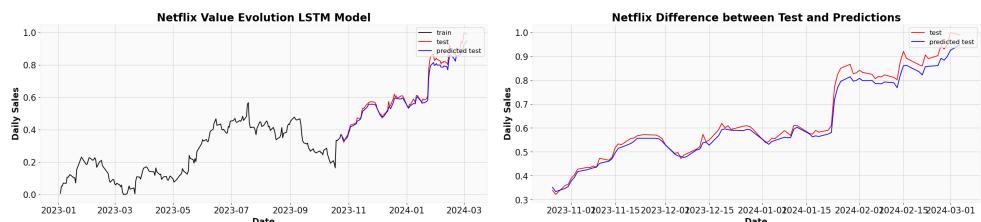


Figura 40: Prediccions del dataset Netflix amb LSTM i Comparació ampliada

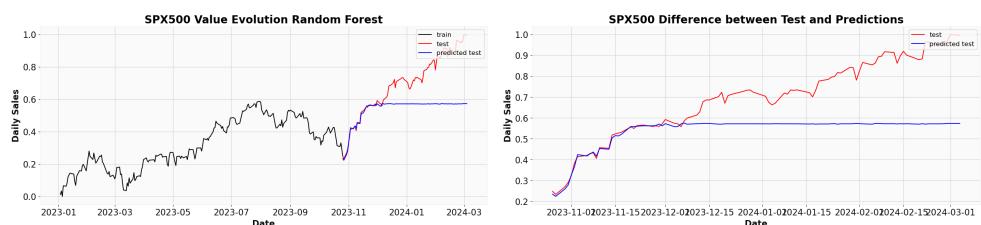


Figura 41: Prediccions del dataset SPX500 amb Random Forest i Comparació ampliada

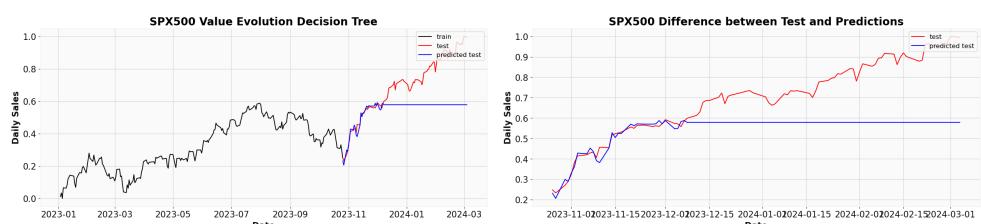


Figura 42: Prediccions del dataset SPX500 amb Decision Tree i Comparació ampliada

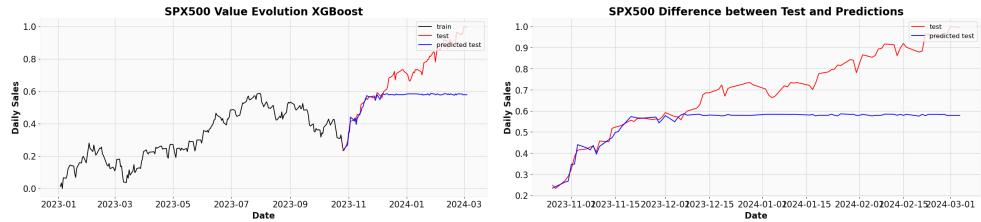


Figura 43: Prediccions del dataset SPX500 amb XGBoost i Comparació ampliada

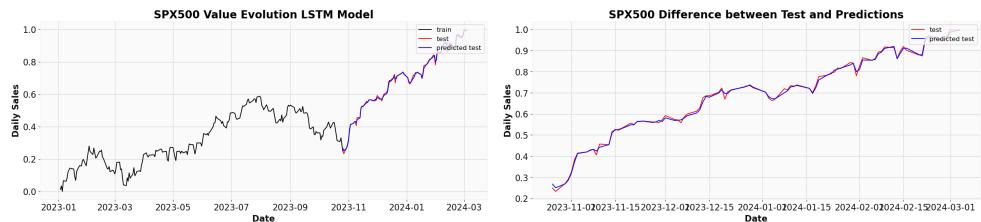


Figura 44: Prediccions del dataset SPX500 amb LSTM i Comparació ampliada

11.2 Prediccions del primer enfoc amb les difències

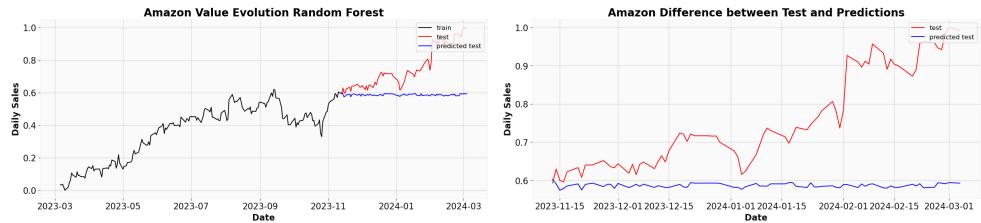


Figura 45: Prediccions del dataset Amazon amb Random Forest i Comparació ampliada

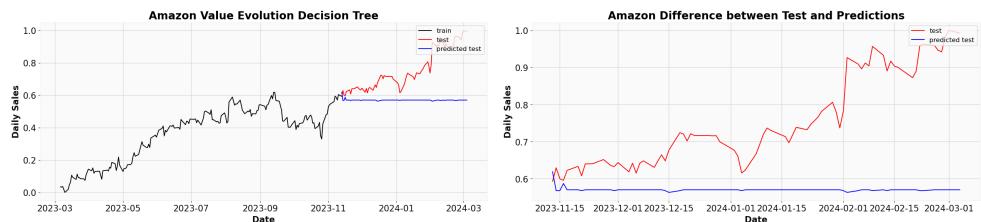


Figura 46: Prediccions del dataset Amazon amb Decision Tree i Comparació ampliada

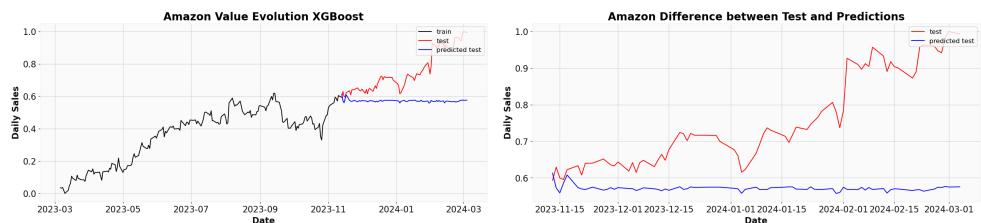


Figura 47: Prediccions del dataset Amazon amb XGBoost i Comparació ampliada

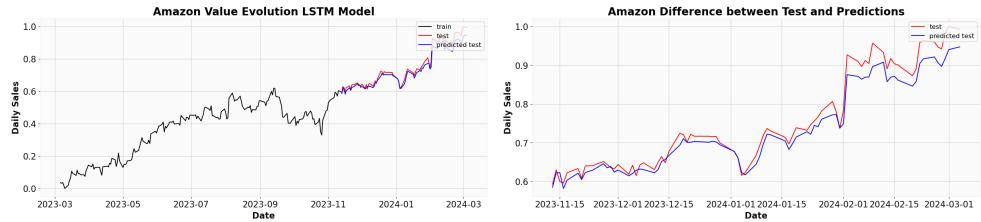


Figura 48: Prediccions del dataset Amazon amb LSTM i Comparació ampliada

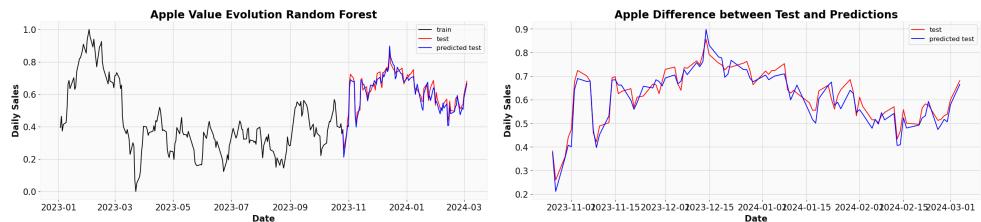


Figura 49: Prediccions del dataset Apple amb Random Forest i Comparació ampliada

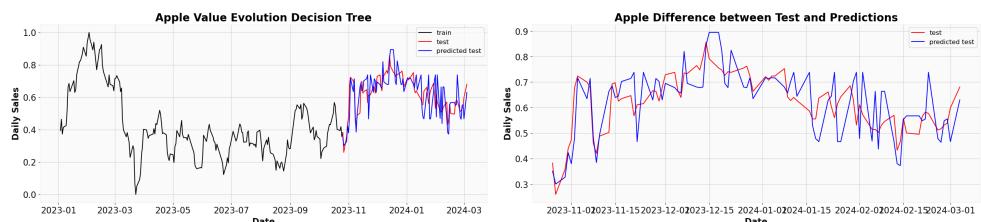


Figura 50: Prediccions del dataset Apple amb Decision Tree i Comparació ampliada

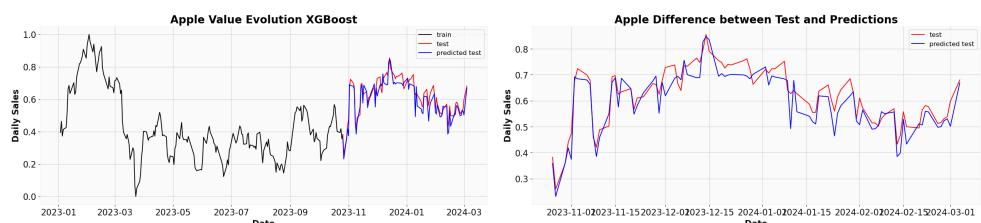


Figura 51: Prediccions del dataset Apple amb XGBoost i Comparació ampliada

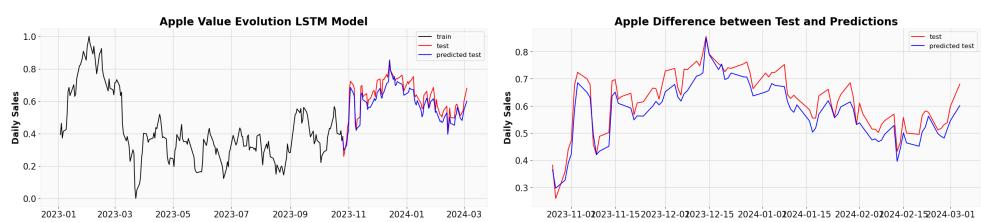


Figura 52: Prediccions del dataset Apple amb LSTM i Comparació ampliada

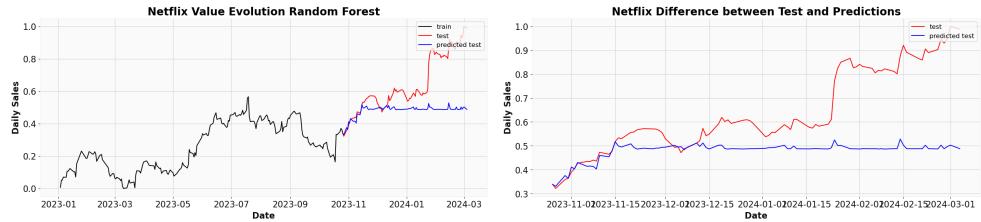


Figura 53: Prediccions del dataset Netflix amb Random Forest i Comparació ampliada

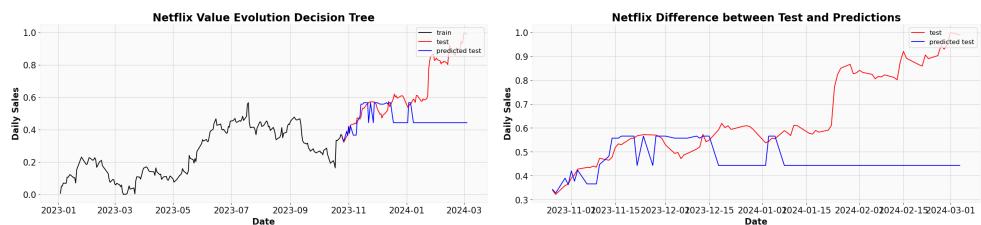


Figura 54: Prediccions del dataset Netflix amb Decision Tree i Comparació ampliada

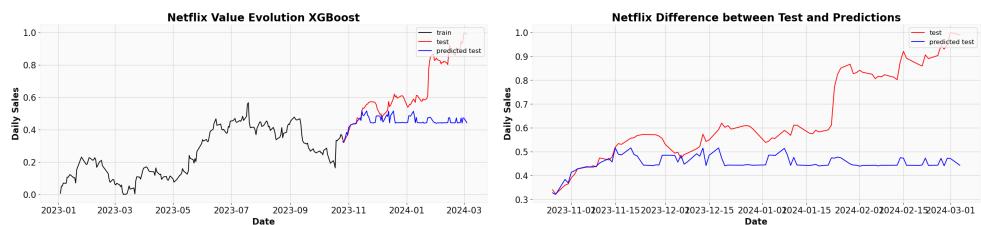


Figura 55: Prediccions del dataset Netflix amb XGBoost i Comparació ampliada

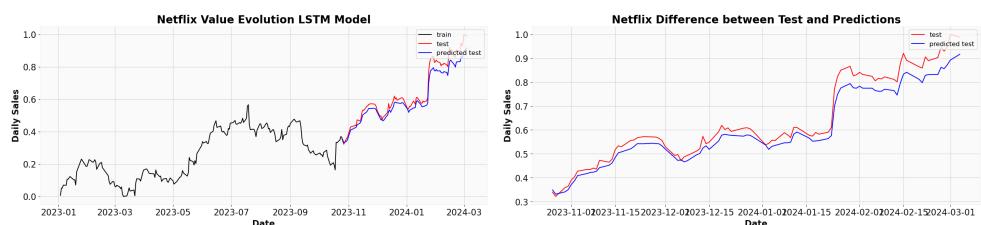


Figura 56: Prediccions del dataset Netflix amb LSTM i Comparació ampliada

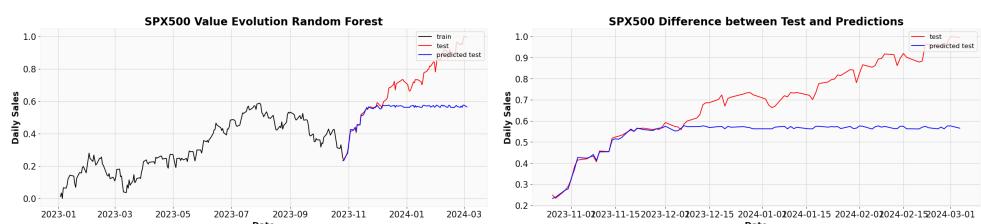


Figura 57: Prediccions del dataset SPX500 amb Random Forest i Comparació ampliada

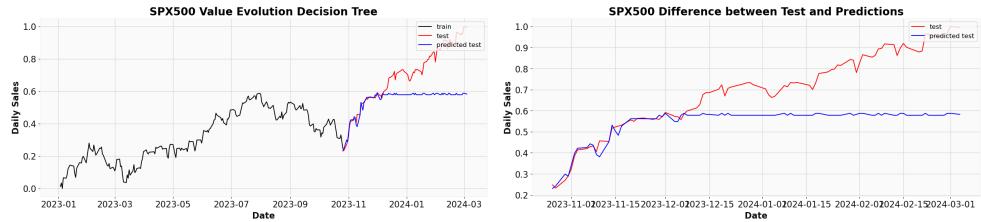


Figura 58: Prediccions del dataset SPX500 amb Decision Tree i Comparació ampliada

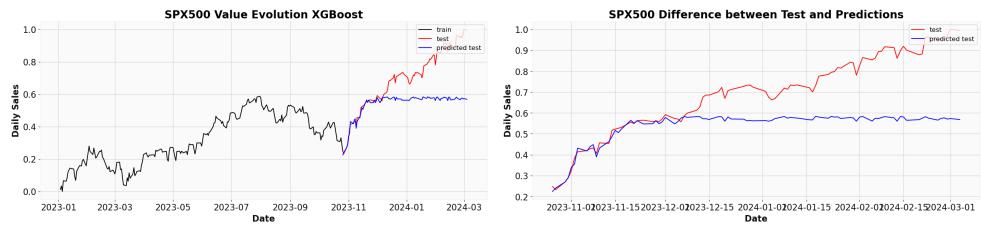


Figura 59: Prediccions del dataset SPX500 amb XGBoost i Comparació ampliada

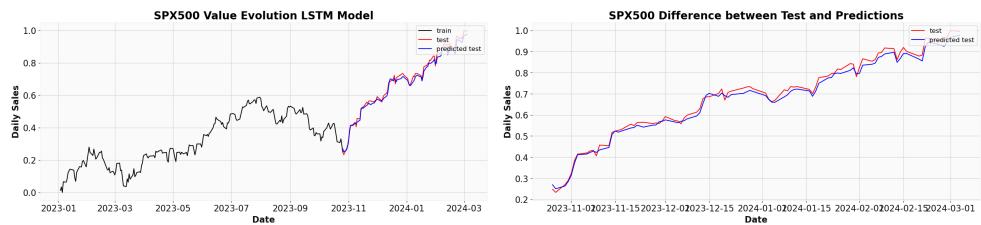


Figura 60: Prediccions del dataset SPX500 amb LSTM i Comparació ampliada

11.3 Prediccions del primer enfoc amb les veles NS

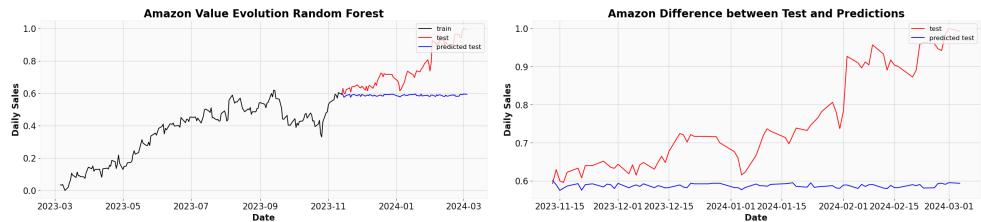


Figura 61: Prediccions del dataset d'Amazon amb Random Forest i Comparació ampliada

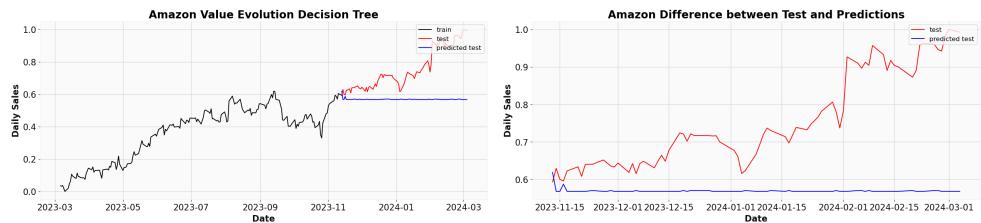


Figura 62: Prediccions del dataset d'Amazon amb Decision Tree i Comparació ampliada

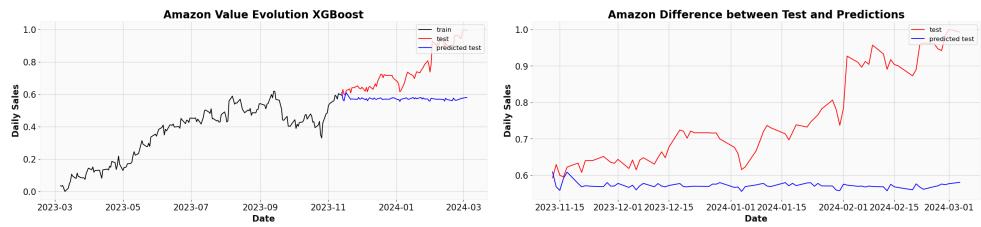


Figura 63: Prediccions del dataset d'Amazon amb XGBoost i Comparació ampliada

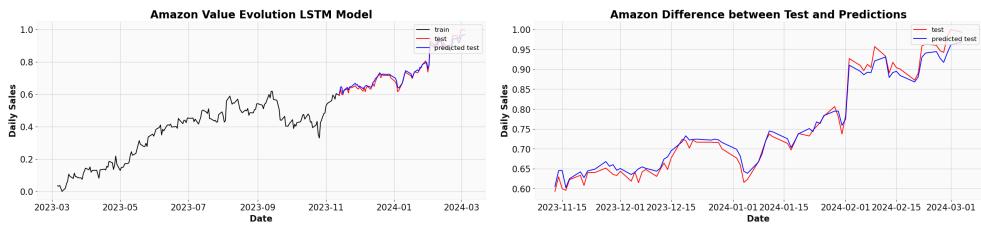


Figura 64: Prediccions del dataset d'Amazon amb LSTM i Comparació ampliada

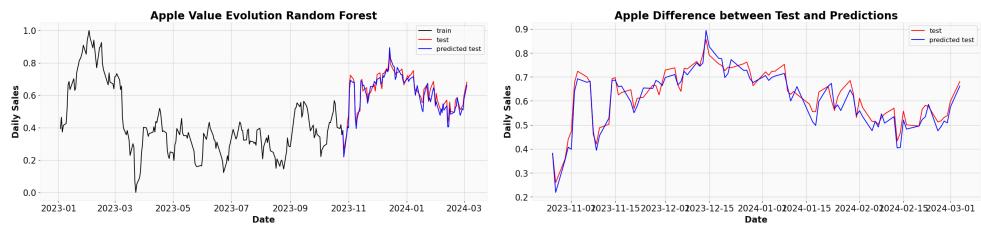


Figura 65: Prediccions del dataset d'Apple amb Random Forest i Comparació ampliada

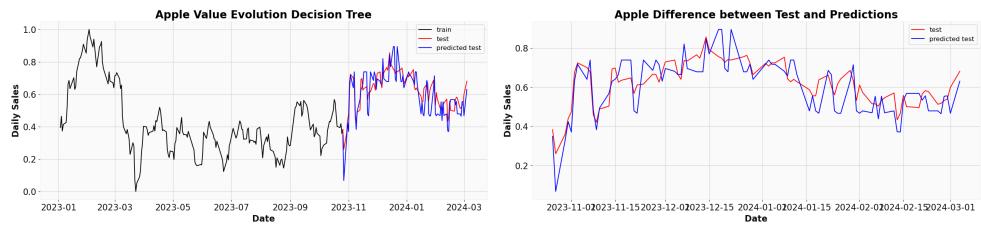


Figura 66: Prediccions del dataset d'Apple amb Decision Tree i Comparació ampliada

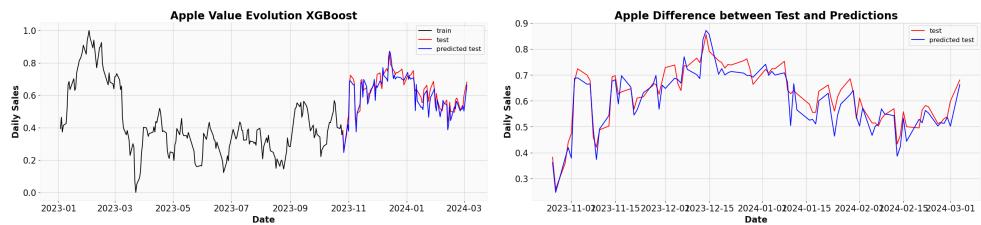


Figura 67: Prediccions del dataset d'Apple amb XGBoost i Comparació ampliada

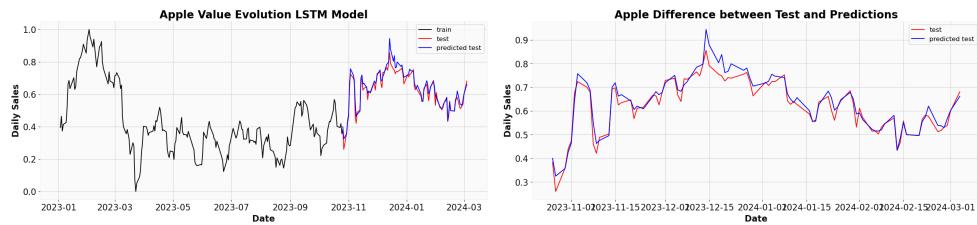


Figura 68: Prediccions del dataset d'Apple amb LSTM i Comparació ampliada

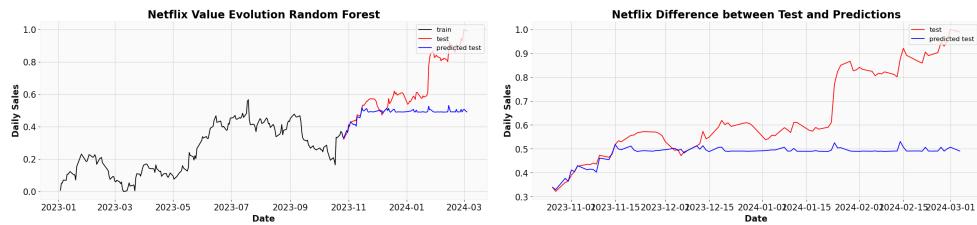


Figura 69: Prediccions del dataset de Netflix amb Random Forest i Comparació ampliada

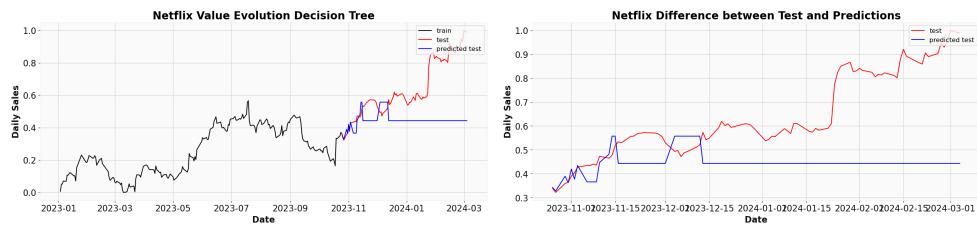


Figura 70: Prediccions del dataset de Netflix amb Decision Tree i Comparació ampliada

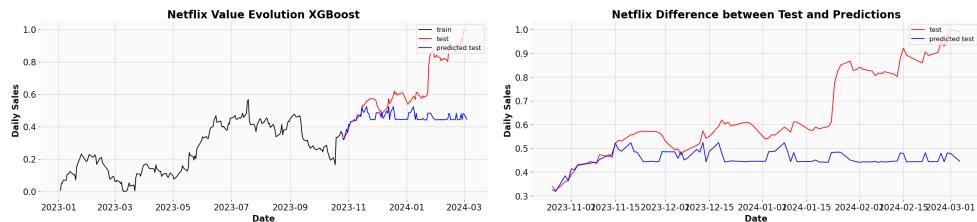


Figura 71: Prediccions del dataset de Netflix amb XGBoost i Comparació ampliada

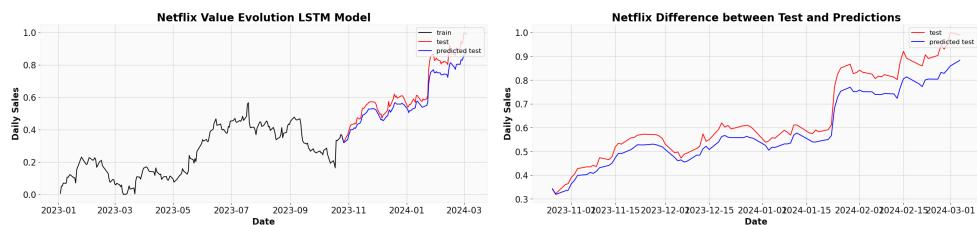


Figura 72: Prediccions del dataset de Netflix amb LSTM i Comparació ampliada

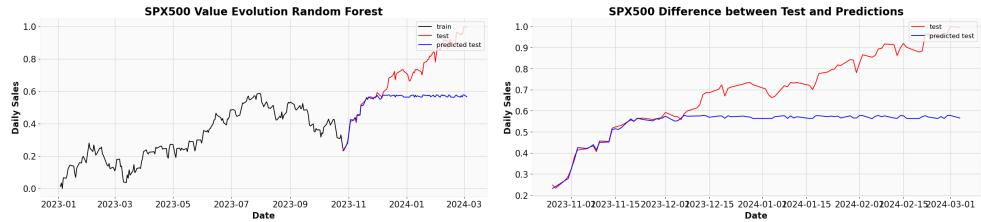


Figura 73: Prediccions del dataset d'SPX500 amb Random Forest i Comparació ampliada

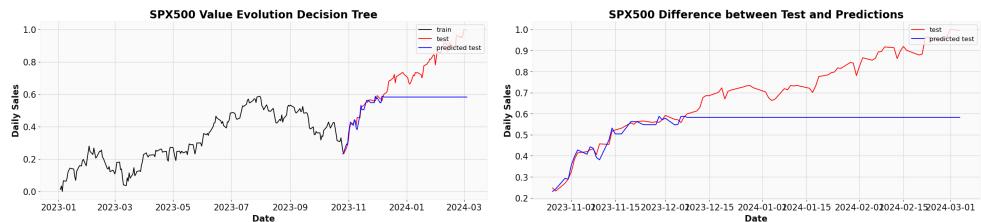


Figura 74: Prediccions del dataset d'SPX500 amb Decision Tree i Comparació ampliada

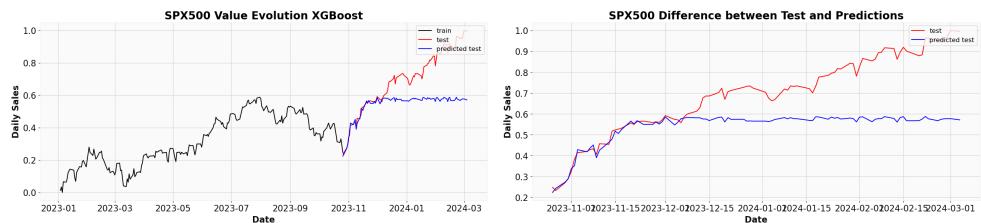


Figura 75: Prediccions del dataset d'SPX500 amb XGBoost i Comparació ampliada

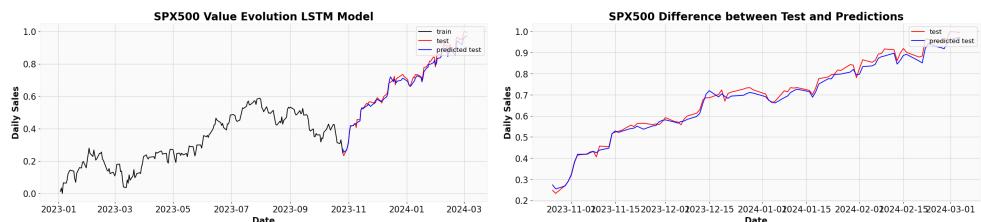


Figura 76: Prediccions del dataset d'SPX500 amb LSTM i Comparació ampliada

11.4 Prediccions del primer enfoc amb les veles ND

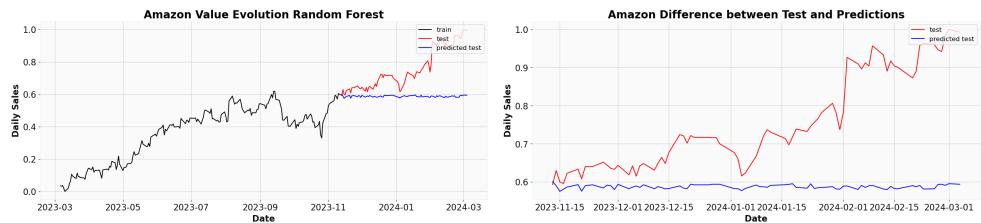


Figura 77: Prediccions del dataset d'Amazon amb Random Forest i Comparació ampliada

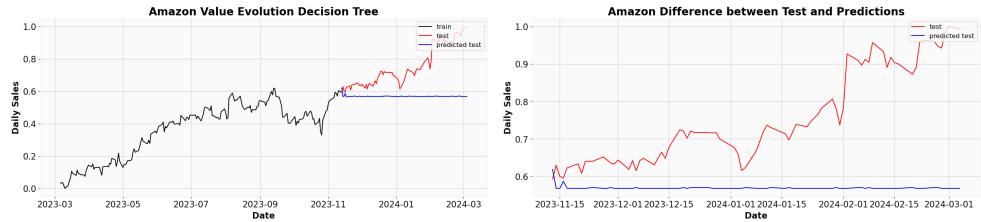


Figura 78: Prediccions del dataset d'Amazon amb Decision Tree i Comparació ampliada

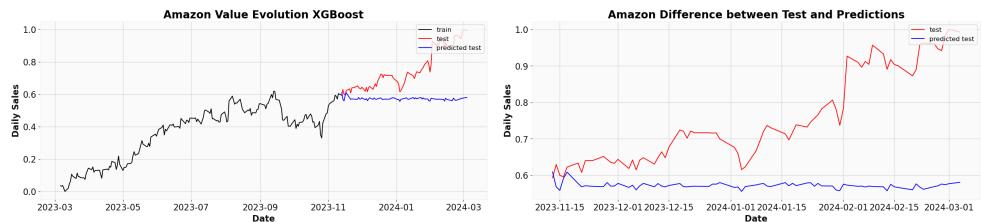


Figura 79: Prediccions del dataset d'Amazon amb XGBoost i Comparació ampliada

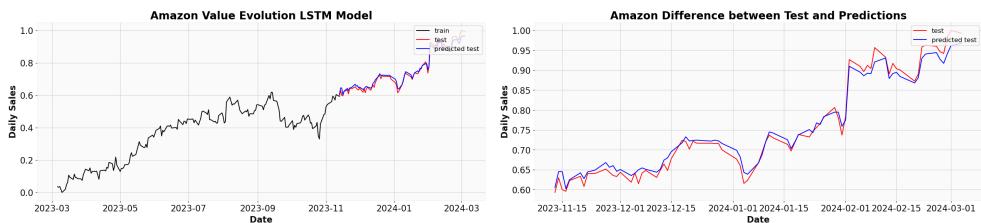


Figura 80: Prediccions del dataset d'Amazon amb LSTM i Comparació ampliada

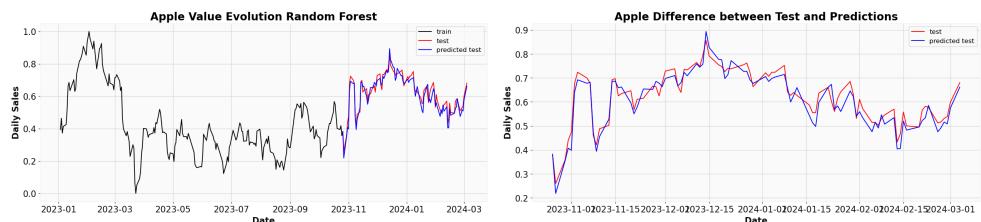


Figura 81: Prediccions del dataset d'Apple amb Random Forest i Comparació ampliada

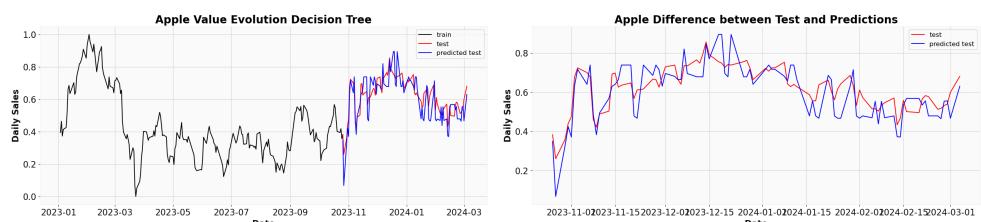


Figura 82: Prediccions del dataset d'Apple amb Decision Tree i Comparació ampliada

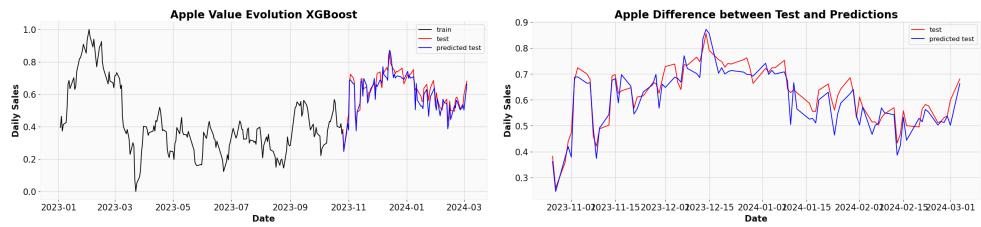


Figura 83: Prediccions del dataset d'Apple amb XGBoost i Comparació ampliada

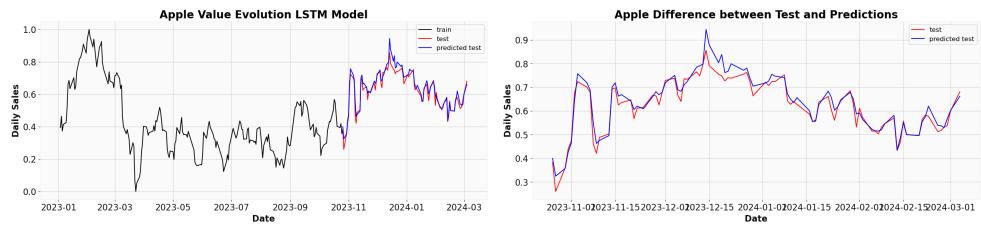


Figura 84: Prediccions del dataset d'Apple amb LSTM i Comparació ampliada

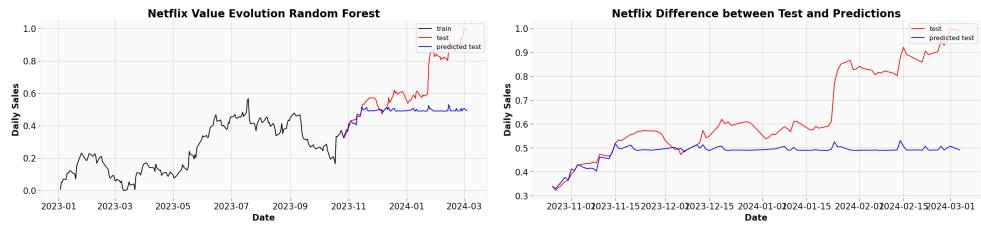


Figura 85: Prediccions del dataset de Netflix amb Random Forest i Comparació ampliada

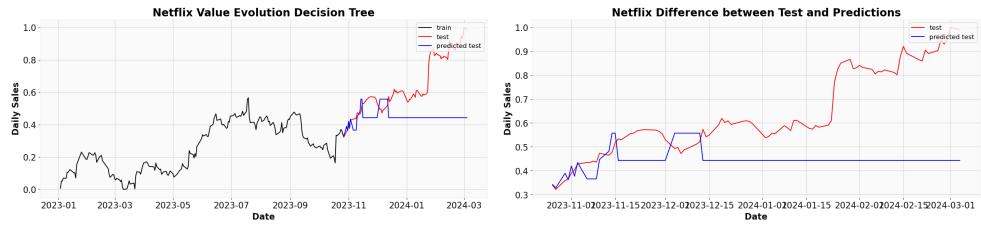


Figura 86: Prediccions del dataset de Netflix amb Decision Tree i Comparació ampliada

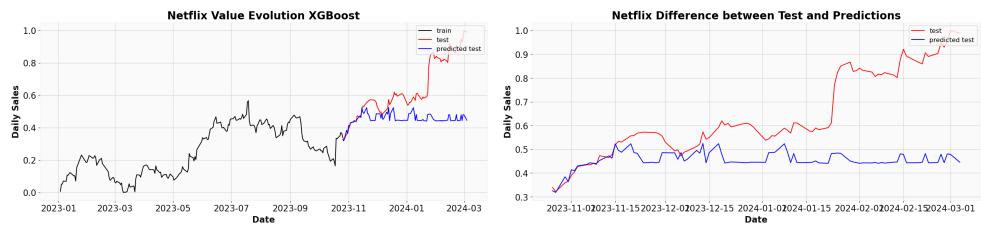


Figura 87: Prediccions del dataset de Netflix amb XGBoost i Comparació ampliada

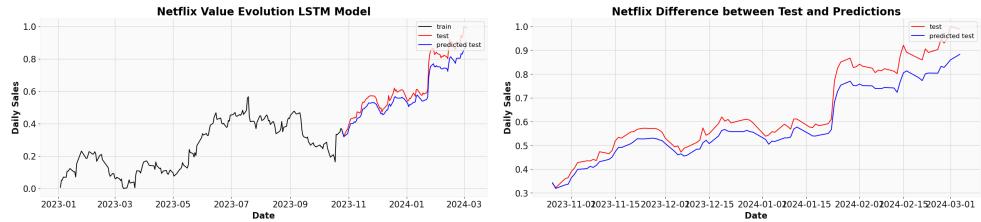


Figura 88: Prediccions del dataset de Netflix amb LSTM i Comparació ampliada

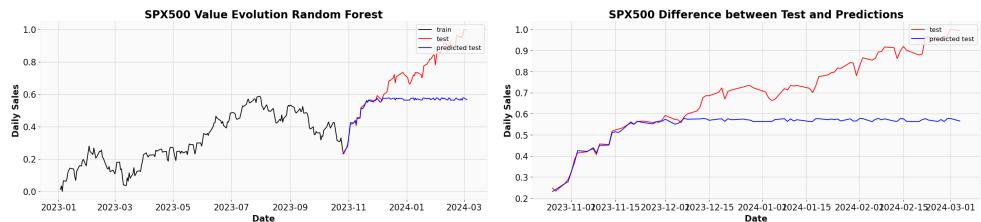


Figura 89: Prediccions del dataset d'SPX500 amb Random Forest i Comparació ampliada

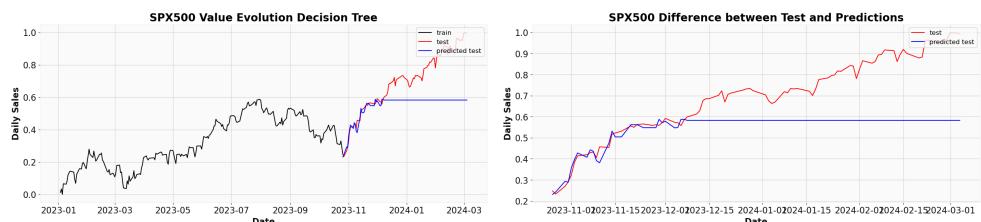


Figura 90: Prediccions del dataset d'SPX500 amb Decision Tree i Comparació ampliada

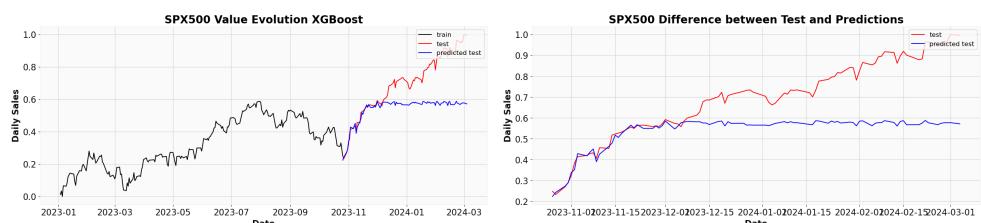


Figura 91: Prediccions del dataset d'SPX500 amb XGBoost i Comparació ampliada

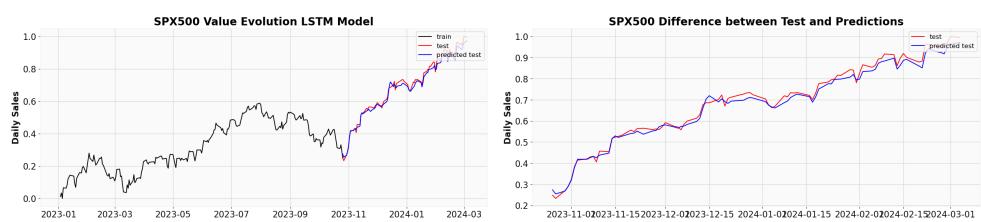


Figura 92: Prediccions del dataset d'SPX500 amb LSTM i Comparació ampliada