ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΑΛΕΞΑΚΗ ΒΑΣΙΛΙΚΗ Α.Μ :1097464

ΕΡΓΑΣΙΑ 1: Πρόβλεψη Τιμών Μετοχών με Γραμμική Παλινδρόμηση

Πρόβλεψη Τιμής Μετοχής AMGN - Amgen Inc.

Στόχος της εργασίας είναι η πρόβλεψη της μηνιαίας τιμής κλεισίματος της μετοχής **AMGN** της εταιρείας **Amgen Inc**., με γραμμική και πολυωνυμική παλινδρόμηση, καθώς και με μεθόδους μείωσης διάστασης. Τα μοντέλα αξιολογήθηκαν τόσο με βραχυπρόθεσμη όσο και μακροπρόθεσμη πρόβλεψη ώστε να διερευνηθεί η δυνατότητα γενίκευσης του μοντέλου στον χρόνο.

1. ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ - monthly_data.py

Τα δεδομένα ανακτήθηκαν από το **API** της **Alpha Vantage** και αφορούν ιστορικές τιμές μετοχής σε **ημερήσια** και **μηνιαία** συχνότητα. Αρχικά έγινε λήψη των ημερήσιων δεδομένων της μετοχής **AMGN** και στη συνέχεια πραγματοποιήθηκε μετατροπή της συχνότητας από ημερήσια σε μηνιαία, ώστε η ανάλυση να πραγματοποιηθεί σε μεγαλύτερη χρονική κλίμακα και να μειωθεί ο θόρυβος της χρονοσειράς. Παράλληλα, ανακτήθηκαν απευθείας και τα **μηνιαία** δεδομένα από το **API**. Τα τελικά αυτά δεδομένα **ταξινομήθηκαν χρονικά** και αποθηκεύτηκαν στο αρχείο **AMGN_monthly.csv**, το οποίο χρησιμοποιήθηκε ως είσοδος σε όλα τα επόμενα στάδια της ανάλυσης.

Για κάθε μήνα καταγράφηκαν οι παρακάτω μεταβλητές: **open**: τιμή ανοίγματος, **high**: μέγιστη τιμή, **low** :ελάχιστη τιμή, **close**: τιμή κλεισίματος και **volume**: όγκος συναλλαγών.

2. ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ – preprocessing.py

Στο αρχείο **preprocessing.py** πραγματοποιήθηκε προεπεξεργασία των δεδομένων για τη βελτίωση της ποιότητάς τους και την κατάλληλη προετοιμασία τους για εκπαίδευση μοντέλων μηχανικής μάθησης.

Αρχικά, εφαρμόστηκε **φίλτρο Gaussian** με παράμετρο **σ = 1** στην τιμή κλεισίματος, με στόχο τη μείωση του θορύβου και των τυχαίων διακυμάνσεων στη χρονοσειρά.

Στη συνέχεια, δημιουργήθηκαν χαρακτηριστικά καθυστέρησης - **lag features** ώστε να ενσωματωθεί **χρονική πληροφορία** στα δεδομένα. Επιλέχθηκε αριθμός καθυστερήσεων **N = 3**, μετά από πειραματισμούς με μεγαλύτερες τιμές, οι οποίες δεν οδήγησαν σε ουσιαστική βελτίωση των αποτελεσμάτων. Για κάθε μήνα δημιουργήθηκαν τα εξής χαρακτηριστικά:

close_t-1, **close_t-2**, **close_t-3** και **volume_t-1**, **volume_t-2**, **volume_t-3**

Με τον τρόπο αυτό, κάθε δείγμα περιλάμβανε πληροφορία για τους **τρεις προηγούμενους μήνες**.

Ο **διαχωρισμός των δεδομένων** πραγματοποιήθηκε χρονικά ως εξής:

Σύνολο εκπαίδευσης - **Training set** : Έτη πριν το 2024 (**train.csv)**

Σύνολο επικύρωσης - **Validation set** : Έτη 2024–2025 (**validation.csv)**

Για την **αξιολόγηση της απόδοσης** των μοντέλων χρησιμοποιήθηκαν οι μετρικές σφάλματος **MAE** (Mean Absolute Error) και **RMSE** (Root Mean Squared Error). Οι συγκεκριμένες μετρικές επιλέχθηκαν ώστε να παρέχουν πλήρη εικόνα τόσο για το μέσο σφάλμα πρόβλεψης όσο και για την επίδραση μεγάλων αποκλίσεων.

3. ΕΡΩΤΗΜΑ Α : Γραμμική Παλινδρόμηση - A_linearModel.py

Για τη μοντελοποίηση χρησιμοποιήθηκε γραμμικό μοντέλο παλινδρόμησης με είσοδο τα χαρακτηριστικά καθυστέρησης τριών μηνών (N = 3), τόσο για την τιμή κλεισίματος όσο και για τον όγκο συναλλαγών.

Αξιολόγηση

```
TRAINING
MAE: 1.1577931620425572
RMSE: 1.6484090444218102

VALIDATION
MAE: 2.654208319604572
RMSE: 3.440099041011889

Συντελεστές μοντέλου:
open          -4.626442e-01
high          -4.814080e-02
low           -3.748803e-02
volume        -1.930024e-09
close_t-1      2.746712e+00
volume_t-1    -1.587841e-09
close_t-2     -1.710502e+00
volume_t-2    -2.658072e-10
close_t-3      5.126273e-01
volume_t-3     1.923492e-10
dtype: float64
```

Η απόδοση του μοντέλου αξιολογήθηκε στο σύνολο εκπαίδευσης και στο σύνολο επικύρωσης χρησιμοποιώντας τις μετρικές **MAE** και **RMSE**.

Τα αποτελέσματα δείχνουν μικρό σφάλμα στο σύνολο εκπαίδευσης και ελαφρώς αυξημένο στο σύνολο επικύρωσης, αναμενόμενο για χρονοσειρές, δεν υποδηλώνει σημαντικό overfitting.

Από την **ανάλυση των συντελεστών** προκύπτει ότι οι **προηγούμενες τιμές κλεισίματος close_t-3** και **close_t-1** έχουν τη **μεγαλύτερη επιρροή** στην πρόβλεψη της τρέχουσας τιμής, καθώς σε αυτές αντιστοιχούν οι μεγαλύτεροι θετικοί συντελεστές. **Αντίθετα**, οι συντελεστές που σχετίζονται με τον **όγκο συναλλαγών** είναι κατά πολύ μικρότεροι σε μέγεθος. Επομένως, ο όγκος διαδραματίζει **σημαντικά μικρότερο ρόλο** στην πρόβλεψη σε σύγκριση με την ιστορική τιμή **κλεισίματος**.

Τέλος, οι τιμές **ανοίγματος, υψηλού και χαμηλού** (open, high, low) εμφανίζουν μικρή αρνητική συνεισφορά. Έτσι, φαίνεται να έχουν **περιορισμένη προγνωστική τους ισχύ** σε σχέση με τις τιμές κλεισίματος των προηγούμενων χρονικών περιόδων.

Επιπλέον εφαρμόστηκαν δύο στρατηγικές αξιολόγησης - **linear_AB.py**

- **Μέθοδος Α**: Πρόβλεψη του επόμενου μήνα με χρήση των πραγματικών δεδομένων.
- **Μέθοδος Β**: Μακροπρόθεσμη πρόβλεψη - Recursive forecasting, όπου κάθε πρόβλεψη χρησιμοποιείται ως είσοδος για την επόμενη χρονική στιγμή.

```
LINEAR MODEL RESULTS
METHOD A MAE: 6.909
METHOD B MAE: 20.749
Reliable horizon (MAE ≤ 10): 1 months
```

Η **μέθοδος Α** παρουσιάζει σημαντικά μικρότερο σφάλμα σε σύγκριση με τη **μέθοδο Β**, γεγονός που δείχνει ότι το **γραμμικό μοντέλο** είναι αρκετά αποτελεσματικό για βραχυπρόθεσμη πρόβλεψη, όταν χρησιμοποιούνται πραγματικές τιμές προηγούμενων μηνών.

Αντίθετα, στη **μέθοδο Β** παρατηρείται σημαντική αύξηση του σφάλματος, καθώς τα λάθη των προηγούμενων προβλέψεων συσσωρεύονται και επηρεάζουν τις μελλοντικές εκτιμήσεις. Έτσι, ο **αξιόπιστος χρονικός ορίζοντας** περιορίζεται στον **ένα μήνα** είναι φανερή η δυσκολία της μακροπρόθεσμης πρόβλεψης χρηματιστηριακών δεδομένων με απλό γραμμικό μοντέλο.

Εικόνα 1

```
actual,predicted
314.26,308.5125224176855
273.83,294.13491204694293
284.32,275.1739021098272
273.94,268.4210881531962
305.85,308.9276543417389
312.45,312.8812341178103
332.47,330.5675222879897
333.83,315.67967786291007
322.21,319.9842549881776
320.16,316.6080798420599
282.87,279.9832766594532
260.64,258.4538872182057
285.42,276.90553334223796
308.06,308.74900037938164
311.55,324.3672919414701
290.92,279.0410328729643
288.18,269.3562043024895
279.21,282.6023941761946
295.1,301.5555247787395
287.71,288.6508052053842
282.2,272.706244318387
298.43,296.86126810053
345.46,329.85932670711395
337.49,336.98273803187993
```

Εικόνα 2

```
horizon,actual,predicted,abs_error
1,314.26,308.51252241768555,5.7474775823144455
2,273.83,310.2858293807366,36.45582938073659
3,284.32,309.69029208418874,25.370292084188748
4,273.94,310.4839749872879,36.543974987287925
5,305.85,310.72508562733293,4.875085627332908
6,312.45,310.6590931797895,1.7909068202104663
7,332.47,310.66966726524976,21.80033273475027
8,333.83,310.6886428332033,23.141357166796695
9,322.21,310.68690128300824,11.52309871699174
10,320.16,310.68559298419495,9.47440701580507
11,282.87,310.68651767345625,27.816517673456246
12,260.64,310.68666105496675,50.04666105496676
13,285.42,310.6865358469329,25.266535846932868
14,308.06,310.68655361770703,2.6265536177070317
15,311.55,310.68657367805844,0.8634263219415743
16,290.92,310.6865685152509,19.766568515250867
17,288.18,310.6865670327976,22.50656703279759
18,279.21,310.68656827952617,31.47656827952619
19,295.1,310.68656830872186,15.586568308721837
20,287.71,310.6865681327951,22.976568132795137
21,282.2,310.68656816977085,28.486568169770862
22,298.43,310.6865681909213,12.256568190921314
23,345.46,310.68656818174185,34.77343181825813
24,337.49,310.68656818060634,26.803431819393666
```

Εικόνα 3

```
horizon,mae
1,5.7474775823144455
2,21.101653481525517
3,22.524533015746595
4,26.029393508631927
5,21.79853193237212
6,18.463927747011848
7,18.940557030974478
8,19.465657047952256
9,18.58315056673442
10,17.672276211641485
11,18.594479980897372
12,21.215495070403154
13,21.52711359167467
14,20.177073593534125
15,18.88949710876129
16,18.944314071666888
17,19.153858363498106
18,19.838453358833
19,19.61466993514294
20,19.782764845025547
21,20.19723167001342
22,19.83629242096378
23,20.447037612150492
```

Εικόνα 1: Σύγκριση πραγματικών και προβλεπόμενων μηνιαίων τιμών κλεισίματος για τη μέθοδο Α. Κάθε γραμμή αντιστοιχεί σε έναν μήνα του συνόλου επικύρωσης. Παρατηρείται ότι

οι προβλέψεις βρίσκονται κοντά στις πραγματικές τιμές, γεγονός που επιβεβαιώνει την ικανοποιητική απόδοση του γραμμικού μοντέλου σε βραχυπρόθεσμη πρόβλεψη.

Εικόνα 2: Αποτελέσματα της μεθόδου Β με αναδρομική πρόβλεψη. Εκτυπώνονται οι πραγματικές τιμές, οι προβλεπόμενες τιμές, ο χρονικός ορίζοντας πρόβλεψης -horizon και το απόλυτο σφάλμα ανά μήνα. Παρατηρείται ότι το σφάλμα αυξάνεται σημαντικά καθώς αυξάνεται ο χρονικός ορίζοντας, ένδειξη συσσώρευσης λάθους στη μακροπρόθεσμη πρόβλεψη.

Εικόνα 3: Μεταβολή της μέσης απόλυτης απόκλισης MAE σε συνάρτηση με τον χρονικό ορίζοντα πρόβλεψης - horizon. Παρατηρείται απότομη αύξηση του σφάλματος ήδη από τον δεύτερο μήνα, γεγονός που δείχνει ότι το γραμμικό μοντέλο είναι αξιόπιστο μόνο για βραχυπρόθεσμες προβλέψεις.

4. ΕΡΩΤΗΜΑ Β: Πολυωνυμικό Μοντέλο με L1 και L2 Κανονικοποίηση - B_polynomialModel.py

Για τη μοντελοποίηση των μη γραμμικών σχέσεων μεταξύ των χαρακτηριστικών εφαρμόστηκε πολυωνυμικός μετασχηματισμός δεύτερου βαθμού, αυξάνοντας τον αριθμό των χαρακτηριστικών.

Ως είσοδοι χρησιμοποιήθηκαν καθυστερημένες τιμές των μεταβλητών **close** και **volume**. Αρχικά δημιουργήθηκαν **10 χαρακτηριστικά**, τα οποία μετά την πολυωνυμική επέκταση αυξήθηκαν σε **65 χαρακτηριστικά** μέσω συνδυασμών δεύτερου βαθμού.

Στη συνέχεια εκπαιδεύτηκαν δύο πολυωνυμικά μοντέλα:

Ridge Regression (L2 regularization): η οποία περιορίζει το μέγεθος των συντελεστών και μειώνει τη μεταβλητότητα του μοντέλου και **Lasso Regression** (L1 regularization): η οποία επιπλέον μηδενίζει ορισμένους συντελεστές, κάνοντας ταυτόχρονα και feature selection.

Τα μοντέλα εκπαιδεύτηκαν με τις ίδιες εισόδους με το γραμμικό μοντέλο, για άμεση σύγκριση.

Οι υπερπαράμετροι των μοντέλων επιλέχθηκαν έπειτα από πειραματισμούς:

Για την **Lasso** δοκιμάστηκαν τιμές [0.001, 0.01, 0.05, 0.1, 0.5] και επιλέχθηκε **α = 0.05**, καθώς παρείχε την καλύτερη ισορροπία μεταξύ σφάλματος εκπαίδευσης και επικύρωσης.
Για την Ridge δοκιμάστηκαν τιμές [0.1, 1, 10, 50, 100] και επιλέχθηκε **α = 10**, που έδωσε το μικρότερο **MAE** και **RMSE** στο σύνολο επικύρωσης χωρίς έντονο overfitting.

* Κατά την εκπαίδευση των γραμμικών μοντέλων εμφανίστηκαν αριθμητικές αστάθειες λόγω διαφορετικής κλίμακας χαρακτηριστικών και ισχυρής συσχέτισης μεταξύ τους. Έτσι, πριν την εκπαίδευση, για την αποφυγή overfitting, εφαρμόστηκε κανονικοποίηση: **StandardScaler**, ώστε όλα τα χαρακτηριστικά να βρίσκονται στην ίδια κλίμακα.

```
inear_model\_coordinate_descent.py:695: ConvergenceWarning: Objective did not converge. You might want to increase the number of iterations, check t
he scale of the features or consider increasing regularisation. Duality gap: 2.618e+03, tolerance: 1.458e+02
  model = cd_fast.enet_coordinate_descent(
C:\Users\vassi\AppData\Local\Packages\PythonSoftwareFoundation.Python.3.11_qbz5n2kfra8p0\LocalCache\local-packages\Python311\site-packages\sklearn\l
inear_model\_ridge.py:215: LinAlgWarning: Ill-conditioned matrix (rcond=3.03626e-17): result may not be accurate.
  return linalg.solve(A, Xy, assume_a="pos", overwrite_a=True).T
```

Αξιολόγηση

```
Feature Count
Original features: 10
Polynomial features: 65

POLYNOMIAL MODEL RESULTS

METHOD A
LASSO MAE: 8.22
RIDGE MAE: 9.96
LASSO RMSE: 10.43
RIDGE RMSE: 12.72

METHOD B
LASSO MAE: 19.51
RIDGE MAE: 21.01
LASSO RMSE: 22.58
RIDGE RMSE: 24.59

Reliable Horizon (MAE ≤ 10)
LASSO: 1
RIDGE: 1
```

Από την αξιολόγηση προκύπτει ότι το μοντέλο **Lasso** υποδεικνύει καλύτερη γενίκευση στα δεδομένα επικύρωσης, έναντι της **Ridge**. Η διαφορά αυτή οφείλεται στην ικανότητα της **Lasso** να πραγματοποιεί ταυτόχρονα επιλογή χαρακτηριστικών, μηδενίζοντας τους μη σημαντικούς πολυωνυμικούς όρους και περιορίζοντας την πολυπλοκότητα και αποφεύγοντας έτσι το **overfitting**.

Και εδώ εφαρμόστηκαν οι δύο στρατηγικές αξιολόγησης :

- **Μέθοδος Α**: Πρόβλεψη του επόμενου μήνα με χρήση των πραγματικών δεδομένων του validation set.

Το μοντέλο **Lasso** παρουσίασε μικρότερο σφάλμα και στις δύο μετρικές, γεγονός που υποδεικνύει καλύτερη γενίκευση.

- **Μέθοδος Β**: Μακροπρόθεσμη πρόβλεψη - Recursive forecasting, όπου κάθε πρόβλεψη χρησιμοποιείται ως είσοδος για την επόμενη χρονική στιγμή.

Παρατηρείται σημαντική αύξηση του σφάλματος και για τα δύο μοντέλα λόγω της συσσώρευσης σφάλματος.

Εικόνα 1 - Προβλέψεις με πραγματικά δεδομένα

actual	predicted_lasso	predicted_ridge
314.26	307.0474699696042	309.280582571962
273.83	290.96823103992125	307.47336799960385
284.32	273.46765203740205	283.7606864237581
273.94	266.4998180362544	280.16677116587266
305.85	305.29567958491816	311.6465393953687
312.45	311.56217174896386	322.8852083798565
332.47	328.12864619958555	337.24958659446673
333.83	314.2860739971027	335.2791115303979
322.21	317.37902805981633	334.1838223758422
320.16	313.7960966635462	328.0946290544931
282.87	273.82961282503055	299.4517187597084
260.64	255.21806366118864	273.2894698638565
285.42	273.1518081321062	284.41399977430297
308.06	306.80065486945136	320.69968984054395
311.55	323.471722523287	337.8430538565123
290.92	276.57714792374554	298.4419080367919
288.18	266.60011559015004	284.14256538445176
279.21	279.81382398806653	295.1400260731615
295.1	298.76205022797905	307.0549663310308
287.71	288.46962953124165	299.38743373007503
282.2	271.95390290398865	281.3877254299323
298.43	294.73470835360285	299.44418968004607
345.46	325.32507162870314	331.48979745668436
337.49	340.58545996870424	352.54696356846125

Εικόνα 2 - MAE ανά χρονικό ορίζοντα

horizon	mae_lasso	mae_ridge
1	7.212530030395726	4.979417428038005
2	19.49189877535457	20.227672581348685
3	19.582405247198494	22.041906135981247
4	22.335382941837892	26.038748563359817
5	18.10654539745085	22.03543524161654
6	16.398762617800525	18.46644976102897
7	18.03960357258242	18.749876917569605
8	19.439144052281748	19.13402124012277
9	19.236636609016728	18.14329093033343
10	18.86972352062454	17.143812584888423
11	19.129084984550204	18.23442377490504
12	21.197726202894227	20.995568354842888
13	21.041958376587584	21.425913068788358
14	19.78662277791187	20.17763605004529
15	18.931331610448478	18.863107495492727
16	18.602669922070334	19.002281701375498
17	18.473850753871467	19.286259596739086
18	18.857678164940054	20.037015431169984
19	18.36478690378054	19.87242949764181
20	18.290684767861674	20.0938022969992
21	18.48602093044058	20.55647268936688
22	17.92587198744495	20.239354940641444
23	18.923323458226402	20.813735912677526
24	19.50557063978859	21.008168494135592

Εικόνα 3 - Αναδρομική πρόβλεψη

horizon	actual	predicted_lasso	predicted_ridge	abs_error_lasso	abs_error_ridge
1	314.26	307.04746996960426	309.280582571962	7.212530030395726	4.979417428038005
2	273.83	305.6012675203134	309.30592773465935	31.771267520313415	35.475927734659365
3	284.32	304.08341819088633	309.99037324524636	19.76341819088634	25.670373245246367
4	273.94	304.5343160257561	311.96927584549553	30.594316025756086	38.02927584549553
5	305.85	304.65880478009734	311.87218195464345	1.1911952199026814	6.022181954643429
6	312.45	304.5901512804511	311.8284776419089	7.859848719548893	0.6215223580911129
7	332.47	304.58535069872624	312.0195601431866	27.884649301273782	20.450439856813432
8	333.83	304.59407258982293	312.00696850200507	29.235927410177055	21.82303149799492
9	322.21	304.5934229371034	311.9925515479813	17.616577062896567	10.217448452018687
10	320.16	304.59249427490516	312.01149252411665	15.567505725094861	8.148507475883378
11	282.87	304.5926996238068	312.0105356750712	21.722699623806818	29.14053567507119
12	260.64	304.59277960467847	312.0081587341592	43.95277960467848	51.36815873415924
13	285.42	304.5927444609079	312.010049636134	19.172744460907893	26.590049636133983
14	308.06	304.59274000487244	312.0100348063854	3.467259995127563	3.9500348063854176
15	311.55	304.592744734039	312.00970773175686	6.95725526596101	0.45970773175685054
16	290.92	304.5927445963982	312.0098947896171	13.67274459639816	21.089894789617063
17	288.18	304.59274406268963	312.00990592255647	16.412744062689626	23.829905922556463
18	279.21	304.592744153106	312.00986461649524	25.382744153106046	32.799864616495256
19	295.1	304.5927442029093	312.0098826941347	9.49274420290925	16.909882694134694
20	287.71	304.5927441854032	312.0098854847896	16.882744185403226	24.299885484789627
21	282.2	304.59274418201875	312.0098805367205	22.39274418201876	29.809880536720527
22	298.43	304.5927441845367	312.00988221740727	6.1627441845366775	13.57988221740726
23	345.46	304.59274418458165	312.00988270252867	40.86725581541833	33.45011729747131
24	337.49	304.59274418428106	312.0098821323289	32.89725581571895	25.480117867671083

Εικόνα 1: Περιλαμβάνει τις πραγματικές τιμές και τις προβλέψεις των μοντέλων **Lasso** και **Ridge** στο **validation set** με χρήση πραγματικών εισόδων - **Μοντέλο Α**. Παρατηρείται ότι το

Lasso προσεγγίζει καλύτερα τις πραγματικές τιμές σε σύγκριση με το **Ridge**, επιβεβαιώνοντας τα χαμηλότερα σφάλματα **MAE** και **RMSE**.

Εικόνα 2: Η εξέλιξη του **MAE** ως προς τον χρονικό ορίζοντα δείχνει ότι και τα δύο μοντέλα ξεπερνούν το κατώφλι **MAE ≤ 10** από τον δεύτερο μήνα. Ο αξιόπιστος χρονικός ορίζοντας αποδεικνύεται **1 μήνας** για **Lasso** και **Ridge**.

Εικόνα 3: Παρουσιάζει τις αναδρομικές προβλέψεις - **Μοντέλο Β** . Παρατηρείται συσσώρευση σφάλματος με την αύξηση του χρονικού ορίζοντα, με τις απόλυτες αποκλίσεις να αυξάνονται γρήγορα μετά τον πρώτο μήνα.

```
Feature,Ridge_Coefficient,Lasso_Coefficient
open,5.6892516194380125,-0.0
high,10.04363436216387,47.64210592752925
low,9.636792551729206,25.905971897594128
volume,-4.526612996323567,-0.0
close_t-1,5.750831049277387,-0.0
volume_t-1,-4.632205628330105,-0.0
close_t-2,5.254279218635809,-0.0
volume_t-2,-4.491836013861832,-0.0
close_t-3,6.040513660632336,0.0
volume_t-3,-4.90892700014337,-0.21047596266832422
open^2,-1.258526005185207,-12.382107943871725
open high,1.8208823453524376,-0.0
open low,1.9559831477418153,-0.0
open volume,-1.8824351145948959,-4.921146287124595
open close_t-1,-1.220889131718561,-0.0
open volume_t-1,-1.8878570550859917,-0.33202777266301453
open close_t-2,-1.698649807877788,-0.0
open volume_t-2,-1.4835241216380877,-0.0055273060822553675
open close_t-3,-0.8717630632983742,-0.0
open volume_t-3,-1.1474948921184205,-0.0
high^2,5.046119265585597,0.0
high low,5.017318033638776,0.0
high volume,3.9125419862739608,4.193090214879228
high close_t-1,1.865037301023677,-0.0
high volume_t-1,3.82705419677014,0.0
high close_t-2,1.4546881587421756,-0.0
high volume_t-2,3.4832619177537305,0.0
high close_t-3,2.31625963106096,0.0
```

Παράμετροι μοντέλων

Απόσπασμα από το αρχείο **B_coefficients.csv.**

Οι συντελεστές των μοντέλων εξάγονται σε κοινό αρχείο για σύγκριση. Το **Lasso** μηδενίζει μεγάλο αριθμό συντελεστών, ενώ το **Ridge** διατηρεί όλους τους συντελεστές με μικρότερες τιμές λόγω **L2** κανονικοποίησης.

5. ΕΡΩΤΗΜΑ Γ: ΜΕΙΩΣΗ ΔΙΑΣΤΑΣΗΣ

Συγκρίθηκαν τρεις διαφορετικές μέθοδοι μείωσης διάστασης/επιλογής χαρακτηριστικών:

1. PCA :Principal Component Analysis
2. CFA : Correlation-based Feature Analysis
3. RFE : Recursive Feature Elimination

Στόχος είναι να μειωθεί η πολυπλοκότητα του μοντέλου, να αποφευχθεί το **overfitting** και να βελτιωθεί η γενίκευση στο **validation set**.

Σε όλες τις μεθόδους χρησιμοποιήθηκε ίδιο μοντέλο πρόβλεψης: **Linear Regression**.

1. Principal Component Analysis - C_PCA.py

Η **Principal Component Analysis** είναι μέθοδος μείωσης διάστασης που βασίζεται σε γραμμικό μετασχηματισμό των αρχικών χαρακτηριστικών σε νέο σύνολο ορθογώνιων μεταβλητών, ονομαζόμενων κύριες συνιστώσες, οι οποίες μεγιστοποιούν τη διακύμανση των δεδομένων.

Στην παρούσα υλοποίηση εφαρμόστηκε κανονικοποίηση μέσω **StandardScaler**, καθώς η **PCA** είναι ευαίσθητη στην κλίμακα των χαρακτηριστικών. Ο αριθμός των τελικών συνιστωσών επιλέχθηκε δυναμικά ώστε να διατηρείται τουλάχιστον το **95%** της συνολικής διακύμανσης, βάσει του αθροιστικού λόγου εξηγούμενης διακύμανσης. Οι τελικές συνιστώσες χρησιμοποιήθηκαν ως είσοδος σε γραμμικό μοντέλο παλινδρόμησης, το οποίο αξιολογήθηκε μέσω **MAE** και **RMSE** στο **validation set**, επιτρέποντας τη συγκριτική αξιολόγηση της **PCA** με τις υπόλοιπες μεθόδους.

Αξιολόγηση

```
PCA RESULTS

TRAIN METRICS
MAE : 3.4863
RMSE: 4.8972

VALIDATION METRICS
MAE : 8.4762
RMSE: 10.4725
```

Παρατηρείται χαμηλό σφάλμα στο σύνολο εκπαίδευσης, αλλά υψηλότερο σφάλμα στο σύνολο επικύρωσης, γεγονός που υποδηλώνει ότι το μοντέλο δεν γενικεύει επαρκώς σε νέα δεδομένα. Το αποτέλεσμα δείχνει ότι η PCA διατηρεί τη γενική πληροφορία, αλλά δεν αποτύπωσε επαρκώς τη χρονική εξάρτηση της χρονοσειράς.

```
PCA COMPONENTS INFO
Number of components: 3

Explained variance ratio per component:
PC1: 0.8482
PC2: 0.0927
PC3: 0.0246

Cumulative explained variance: 0.9655
```

Εδώ εμφανίζεται το ποσοστό διακύμανσης που εξηγεί κάθε κύρια συνιστώσα. Η πρώτη συνιστώσα PC1 καλύπτει το μεγαλύτερο ποσοστό της πληροφορίας, ενώ οι υπόλοιπες συνεισφέρουν μικρότερα ποσά. Συνολικά, διατηρείται το μεγαλύτερο μέρος της πληροφορίας (96.55%) με μόλις τρεις συνιστώσες, γεγονός που καταδεικνύει την ισχυρή συμπίεση δεδομένων.

```
PCA COMPONENT MATRIX (Eigenvectors):
      open   high    low  volume  close_t-1  volume_t-1  close_t-2  volume_t-2  close_t-3  volume_t-3
PC1  0.334  0.333  0.334  -0.286      0.334      -0.289      0.334      -0.289      0.332      -0.288
PC2  0.234  0.239  0.223   0.403      0.234       0.427      0.239       0.420      0.242       0.383
PC3  0.010 -0.004  0.040  -0.609      0.004      -0.342     -0.008       0.353     -0.018       0.621
```

Παραπάνω απεικονίζεται ο πίνακας βαρύτητας των αρχικών χαρακτηριστικών στις κύριες συνιστώσες. Η πρώτη συνιστώσα συνδυάζει σχεδόν εξίσου όλα τα χαρακτηριστικά τιμών και καθυστερημένων τιμών, υποδεικνύοντας γενική τάση αγοράς. Οι επόμενες συνιστώσες σχετίζονται περισσότερο με τη μεταβλητότητα του όγκου συναλλαγών.

2. Correlation-based Feature Analysis – C_CFA.py

Η **Correlation-based Feature Analysis** (CFA) είναι τεχνική επιλογής χαρακτηριστικών που βασίζεται στη μέτρηση της γραμμικής συσχέτισης Pearson μεταξύ κάθε μεταβλητής εισόδου και της μεταβλητής στόχου.

Στην παρούσα υλοποίηση, ο πίνακας συσχέτισης υπολογίστηκε αποκλειστικά στο **training set** ώστε να αποφευχθεί το **data leakage**. Στη συνέχεια εφαρμόστηκε φίλτρο επιλογής με βάση την απόλυτη τιμή του συντελεστή συσχέτισης, με προκαθορισμένο κατώφλι **|ρ| ≥ 0.3**. Τα χαρακτηριστικά που ικανοποιούσαν το κριτήριο χρησιμοποιήθηκαν ως είσοδος σε γραμμικό μοντέλο παλινδρόμησης, το οποίο εκπαιδεύτηκε και αξιολογήθηκε μέσω των μετρικών **MAE** και **RMSE** στο **validation set**, επιτρέποντας την αντικειμενική σύγκριση με άλλες μεθόδους μείωσης διάστασης.

Αξιολόγηση

```
TRAIN SET
MAE : 1.1578
RMSE: 1.6484

VALIDATION SET
MAE : 2.6542
RMSE: 3.4401
```

Το χαμηλό σφάλμα στο **training set** και η μικρή απόκλιση από το **validation set** υποδεικνύουν πολύ καλή γενίκευση και απουσία **overfitting**. Η **CFA** παρουσιάζει τη μικρότερη απόδοση σφάλματος σε σύγκριση με άλλες μεθόδους.

```
close          1.000000
close_t-1      0.998213
high           0.997669
low            0.997175
open           0.995203
close_t-2      0.994140
close_t-3      0.989910
volume_t-3    -0.733627
volume_t-2    -0.731829
volume_t-1    -0.729886
volume        -0.727359
```

Το τελικό σύνολο χαρακτηριστικών που επιλέχθηκε με βάση κατώφλι συσχέτισης **|corr| ≥ 0.3**. Επιλέχθηκαν συνολικά **10** χαρακτηριστικά, τα οποία περιλαμβάνουν καθυστερημένες τιμές close, τιμές αγοράς και μεταβλητές όγκου, γεγονός που δείχνει ότι το μοντέλο διατηρεί τόσο χρονική όσο και χρηματοοικονομική πληροφορία.

```
CFA MODEL COEFFICIENTS (sc
   Feature   Coefficient
 close_t-1  2.746712e+00
 close_t-2 -1.710502e+00
 close_t-3  5.126273e-01
      open -4.626442e-01
      high -4.814081e-02
       low -3.748803e-02
    volume -1.930024e-09
volume_t-1 -1.587841e-09
volume_t-2 -2.658073e-10
volume_t-3  1.923492e-10
```

Οι **συντελεστές του γραμμικού μοντέλου** που προέκυψε από την **CFA**. Διαπιστώνεται ότι οι καθυστερημένες τιμές του **close** έχουν τους υψηλότερους συντελεστές, γεγονός που δείχνει την ισχυρή επίδρασή τους στην πρόβλεψη, ενώ οι μεταβλητές όγκου εμφανίζουν μικρότερη συνεισφορά.

3. Recursive Feature Elimination – C_RFE_Wrapper.py

Η **Recursive Feature Elimination** είναι μέθοδος επιλογής χαρακτηριστικών τύπου **wrapper**, η οποία βασίζεται στην επαναληπτική εκπαίδευση ενός μοντέλου πρόβλεψης και στην αξιολόγηση της σχετικής σημασίας των χαρακτηριστικών του.

Στην παρούσα υλοποίηση χρησιμοποιήθηκε ως **εκτιμητής γραμμικό μοντέλο παλινδρόμησης**, το οποίο επανεκπαιδεύεται σε κάθε επανάληψη αφαιρώντας το λιγότερο σημαντικό χαρακτηριστικό σύμφωνα με τους συντελεστές του μοντέλου. Η διαδικασία συνεχίζεται μέχρι να παραμείνουν **4** χαρακτηριστικά. Το τελικό σύνολο χαρακτηριστικών χρησιμοποιήθηκε για την εκπαίδευση νέου γραμμικού μοντέλου, το οποίο αξιολογήθηκε μέσω **MAE** και **RMSE** στο **validation set**, επιτρέποντας άμεση σύγκριση με **PCA** και **CFA**.

Αξιολόγηση

```
RFE MODEL RESULTS

TRAIN SET
MAE : 1.1672
RMSE: 1.6754

VALIDATION SET
MAE : 2.8537
RMSE: 3.625
```

Παρατηρείται σημαντικά χαμηλό σφάλμα εκπαίδευσης και μικρή απόκλιση στο **validation set**, γεγονός που καταδεικνύει ικανοποιητική γενίκευση και όχι overfitting. Το **RFE** παρουσιάζει επίδοση παρόμοια με την **CFA**, με σαφώς καλύτερα αποτελέσματα από την **PCA**.

```
RFE SELECTED FEATURES:
1. open
2. close_t-1
3. close_t-2
4. close_t-3

Feature ranking (1 = selected):
   Feature  Rank
      open     1
 close_t-1     1
 close_t-2     1
 close_t-3     1
      high     2
       low     3
    volume     4
volume_t-1     5
volume_t-2     6
volume_t-3     7
```

Το σύνολο των **4** χαρακτηριστικών που επιλέχθηκαν από τη μέθοδο **RFE**, καθώς και η κατάταξη όλων των υπολοίπων χαρακτηριστικών βάσει της σημασίας τους. Οι μεταβλητές **open**, **close_t-1**, **close_t-2** και **close_t-3** διατηρήθηκαν στο τελικό μοντέλο, ενώ τα χαρακτηριστικά που σχετίζονται με τον όγκο βρίσκονται στις τελευταίες θέσεις, έχοντας μικρότερη συνεισφορά στην πρόβλεψη της τιμής.

```
RFE MODEL COEFFICIENTS
  Feature  Coefficient
close_t-1     2.591368
close_t-2    -1.605404
close_t-3     0.481283
     open    -0.464598
```

Οι συντελεστές του τελικού γραμμικού μοντέλου που προέκυψε μετά την εφαρμογή του **RFE**. Η μεταβλητή **close_t-1** έχει τον μεγαλύτερο θετικό συντελεστή, επιβεβαιώνοντας τη σημασία της πιο πρόσφατης τιμής στην πρόβλεψη της επόμενης. Αντίθετα, οι **close_t-2** και **open** παρουσιάζουν αρνητική επίδραση, οπότε έχουν διορθωτικό ρόλο στη μοντελοποίηση της χρονικής εξέλιξης.

Σύγκριση αποτελεσμάτων

Η συγκριτική αξιολόγηση των τριών μεθόδων δείχνει ότι η καλύτερη μέθοδος είναι η **CFA** ως προς την ακρίβεια πρόβλεψης, καθώς παρουσίασε το μικρότερο σφάλμα στο **validation set**. Η **RFE** είναι πάρα πολύ κοντά σε απόδοση, διατηρεί σημαντικά λιγότερα χαρακτηριστικά και επιτυγχάνει παρόμοιο επίπεδο γενίκευσης. Αντίθετα, η **PCA**, παρότι πέτυχε μεγάλη μείωση της διάστασης, εμφάνισε σημαντικά υψηλότερο σφάλμα, γεγονός που καταδεικνύει ότι η γραμμική αναπαράσταση μέσω κύριων συνιστωσών δεν αποτύπωσε αποτελεσματικά την χρονοσειρά. Επομένως, επιλέγεται η **CFA** ως βέλτιστη προσέγγιση, με τη **RFE** να αποτελεί αξιόπιστη εναλλακτική.

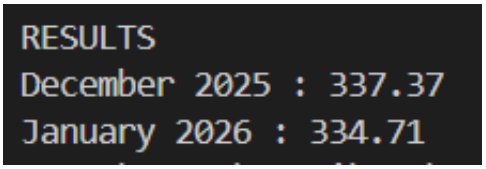
6. ΕΡΩΤΗΜΑ Δ: ΠΡΟΒΛΕΨΗ ΤΙΜΗΣ - D_meCFA.py

Πρόβλεψη Τιμών Μετοχής για Δεκέμβριο 2025 και Ιανουάριο 2026

Για την πρόβλεψη μελλοντικών τιμών επιλέχθηκε ως τελικό μοντέλο το **CFA** σε συνδυασμό με **γραμμική παλινδρόμηση**, καθώς παρουσίασε την καλύτερη απόδοση στο **validation set** στο Ερώτημα Γ.

Η πρόβλεψη πραγματοποιήθηκε με τη μέθοδο της **αναδρομικής πρόβλεψης**, όπου η εκτιμώμενη τιμή ενός μήνα χρησιμοποιείται ως είσοδος για την πρόβλεψη του επόμενου.

Το μοντέλο εκπαιδεύτηκε χρησιμοποιώντας όλα τα διαθέσιμα ιστορικά δεδομένα **έως το 2025**. Τα χαρακτηριστικά εισόδου που χρησιμοποιήθηκαν προέκυψαν από τη **διαδικασία CFA** και περιλαμβάνουν: **close_t-1, close_t-2, close_t-3**, **volume, volume_t-1, volume_t-2, volume_t-3**,**open, high, low**



Το μοντέλο **CFA** προβλέπει ήπια ανοδική τάση της τιμής της μετοχής από τον **Δεκέμβριο 2025** προς τον **Ιανουάριο 2026**. Η μικρή μεταβολή μεταξύ των δύο μηνών υποδηλώνει σταθεροποίηση της τιμής σε υψηλά επίπεδα. Δεδομένης της αναδρομικής φύσης της πρόβλεψης, η εκτίμηση του πρώτου μήνα θεωρείται πιο αξιόπιστη.