ΕΡΓΑΣΙΑ 1

ΣΤΑΤΙΣΤΙΚΕΣ ΜΕΘΟΔΟΙ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΜΠΑΝΤΖΗΣ ΣΤΑΥΡΟΣ 1097449

Πίνακας περιεχομένων

Σημειώσεις

Στην παρούσα εργασία για την συλλογή ιστορικών δεδομένων χρησιμοποιήθηκε η βιβλιοθήκη yfinance της python η οποία έχει ένα API και μου έδωσε όσα δεδομένα χρειάστηκα, έναντι της Alpha Vantage η οποία έδινε μόνο τις τελευταίες 100 μέρες.

Η μετοχή που χρησιμοποίησα για την εργασία ήταν αυτή της 3M Company (MMM) και πήρα δεδομένα για 3 χρόνια για να έχω καλύτερη εκπαίδευση και να έχω μία καλύτερη απόδοση.

Όλος ο κώδικας της εργασίας υπάρχει μέσα στο αρχείο «data_acquisition» με τα κατάλληλα σχόλια.

Όσα περιέχονται στον κώδικα (csv, json, αποτελέσματα) είναι από το τελευταίο test που έκανα στις 28/11. Αν γίνει κάποιο τεστ προφανώς οι τιμές μπορεί να διαφέρουν.
Στην συγκεκριμένη μετοχή θα υπάρξει μέρισμα τον Δεκέμβρη, πράγμα που θα επηρεάσει τις μελλοντικές τιμές λιγάκι.

Προεπεξεργασία Δεδομένων

1)Στόχος Πρόβλεψης: Στην εργασία επικεντρωθήκαμε στην τιμή κλεισίματος των μηνιαίων τιμών. Ο στόχος ήταν να προβλέψουμε την μέση τιμή κλεισίματος κάθε μήνα χρησιμοποιώντας ως είσοδο τις τιμές κλεισίματος και τον αντίστοιχο όγκο συναλλαγών των προηγούμενων μηνών.

2)Δημιουργία Χαρακτηριστικών με καθυστέρηση: Για να αναγνωρίσει το μοντέλο κάποια μοτίβα, δημιουργήσαμε χαρακτηριστικά καθυστέρησης όπως close_t-1, close_t-2 όπου είναι η μέση τιμή κλεισίματος για 1,2 μήνες πριν αντίστοιχα, και volume_t-1, volume_t-2 όπου είναι ο μέσος όγκος συναλλαγών 1,2 μήνες πριν αντίστοιχα. Αυτά τα χαρακτηριστικά χρησιμοποιήθηκαν ως είσοδοι για τα μοντέλα παλινδρόμησης, ενώ η τρέχουσα τιμή κλεισίματος ήταν ο στόχος.
Αυτά φαίνονται στον κώδικα στο κελί 3.

3)Διαχωρισμός σε εκπαίδευση και επικύρωση: Διαχωρίσαμε τα δεδομένα ως εξής:
Training set: όλα τα δεδομένα πριν το 2024

Validation set: δεδομένα του 2024 και 2025

Αυτό επιτρέπει την αξιολόγηση του μοντέλου σε απρόβλεπτα δεδομένα και εξασφαλίζει ότι δεν υπάρχει πληροφορία από το μέλλον.
Αυτό γίνεται στα κελία 3 και 7 του κώδικα

4)Μέθοδοι αξιολόγισης:

Μέθοδος Α: Προβλέπουμε τον επόμενο μήνα χρησιμοποιώντας τις πραγματικές τιμές του προηγούμενου μήνα

Μέθοδος Β(recursive): Χρησιμοποιούμε τις προβλέψεις των προηγούμενων μηνών ως είσοδο για να προβλέψουμε τον επόμενο μήνα. Αυτή η μέθοδος είναι πιο δύσκολη, αλλά επιτρέπει μακροπρόθεσμες προβλέψεις.
Αυτά γίνονται στα κελιά 4 και 11 του κώδικα.

5) Προεπεξεργασία / Μείωση θορύβου: Για να μειώσουμε τον θόρυβο στις τιμές, εφαρμόσαμε Gaussian smoothing στις μηνιαίες τιμές κλεισίματος με σ=1. Αυτό βοηθά το μοντέλο να εντοπίζει γενικά trends αντί για τυχαίες διακυμάνσεις.

Αυτό γίνεται στο κελί 3 του κώδικα.

Απαντήσεις στα Ερωτήματα

Α. *Χρησιμοποιήστε ένα γραμμικό μοντέλο παλινδρόμησης για να βρείτε τη σχέση μεταξύ των παρελθόντων τιμών κλεισίματος (χαρακτηριστικά καθυστέρησης) και του στόχου (επόμενη τιμή κλεισίματος). Επιλέξτε πόσο πριν στο χρόνο θέλετε να πάτε – κάντε περισσότερες δοκιμές. Προσοχή να μην ανεβάσετε πολύ τον αριθμό παραμέτρων. Αναφέρετε τις παραμέτρους του μοντέλου που υπολογίσατε. Αναφέρετε τις κατάλληλες μετρικές σφάλματος για το σύνολο εκπαίδευσης και το σύνολο επικύρωσης*.

Απάντηση:

Για να προβλεφθεί η μέση τιμή κλεισίματος του μήνα δημιουργήθηκαν χαρακτηριστικά καθυστέρησης (close_t-1, close_t-2, close_t-3, volume_t-1, volume_t-2, volume_t-3) όπου αναπαριστούν τις τιμές κλεισίματος και τον όγκο των 3 προηγούμενων μηνών. Αυτά τα χαρακτηριστικά χρησιμοποιούνται ως είσοδος (X) στο μοντέλο.

Το training set περιλαμβάνει όλα τα δεδομένα πριν το 2024, ενώ το validation set περιλαμβάνει τα έτη 2024-2025.

Για το μοντέλο χρησιμοποιούμε γραμμική παλινδρόμηση (Linear Regression) και χρησιμοποιήθηκαν όλα τα χαρακτηριστικά lag για να μην αυξηθεί πολύ ο αριθμός των παραμέτρων και να αποφευχθεί το overfitting.

Τα αποτελέσματα του μοντέλου είναι (Ιανουάριος 2024) 82.70$ με σφάλματα στο validation set MSE = 17.4381, MAE = 3.2985

Σχόλια: Η πρόβλεψη χρησιμοποιεί τις πραγματικές τιμές του προηγούμενου μήνα για το πρώτο βήμα (μέθοδος Α) και μπορεί να επεκταθεί recursive για μακροπρόθεσμη πρόβλεψη (μέθοδος Β)

Β. *Χρησιμοποιήστε ένα πολυωνυμικό μοντέλο παλινδρόμησης με L1, L2 νόρμες κανονικοποίησης. Επιλέξτε κατάλληλες υπερπαραμέτρους. Αναφέρετε τις παραμέτρους του μοντέλου που υπολογίσατε. Αναφέρετε τις κατάλληλες μετρικές σφάλματος για το σύνολο εκπαίδευσης και το σύνολο επικύρωσης.*

Απάντηση: Χρησιμοποιήθηκε το πολυωνυμικό μοντέλο Ridge (L2 regularization) $2^{ου}$ βαθμού. Οι είσοδοι του μοντέλου ήταν τα χαρακτηριστικά close_t-1 και volume_t-1 τα οποία επεκτάθηκαν σε πολυωνυμικούς όρους $2^{ου}$ βαθμού μέσω Polynomial Features. Η υπερπαράμετρος του Ridge ήταν alpha=1.0 . Για το Lasso(L1) χρησιμοποιήθηκε alpha=0.1 . Το μοντέλο εκπαιδεύτηκε με τα δεδομένα όλων των μηνών εκτός της τελευταίας γραμμής (Δεκέμβριος 2025). Για

το validation χρησιμοποιήθηκε η τελευταία γραμμή του dataset (Δεκέμβριος 2025) ενώ η πρόβλεψη για τον επόμενο μήνα (Ιανουάριος 2026) έγινε με recursive forecasting χρησιμοποιώντας την τελευταία διαθέσιμη τιμή.

Τα αποτελέσματα που έχουμε από τον κώδικα είναι:
Recursive πρόβλεψη (Ridge Poly) για τον Ιανουάριο 2026: 158.7419.

Το L2 βοήθησε στην μείωση overfitting σε σχέση με απλή πολυωνυμική παλινδρόμηση.

Γ. *Μειώστε τη διάσταση ακολουθώντας μεθοδολογία PCA, CFA και μια μέθοδο wrapper της επιλογής σας. Συγκρίνετε τα αποτελέσματα.*

1)PCA (Principal Component Analysis): Χρησιμοποιήθηκαν τα χαρακτηριστικά close_t-1 και volume_t-1. Οι διαστάσεις μειώθηκαν σε 2 κύριες συνιστόσες, ώστε να μειωθεί η πολυπλοκότητα και να αποφευχθεί overfitting. Ridge Regression (alpha = 1.0) εφαρμόστηκε πάνω στις συνιστώσες.

```
Recursive Prediction (Ridge + PCA) (Ιανουάριος 2026): 163.47967554950532
```

2)RFE(Recursive Feature Elimination) – Wrapper Method: Χρησιμοποιήθηκε Ridge ως Estimator. Το Ridge εκπαιδεύτηκε ξανά με το επιλεγμένο χαρακτηριστικό

```
Features selected by RFE: ['close_t-1']
Recursive Prediction (Ridge + RFE)(Ιανουάριος 2026): 163.5455388212434
```

3)CFA (Confirmatory Factor Analysis): Factor Analysis με παράγοντα (n_factors=1) και χωρίς επιστροφή. Το αποτέλεσμα των παραγόντων τα βάλαμε σε Ridge Regression.

```
Καλύτερο μοντέλο: Ridge + RFE
Πρόβλεψη επόμενου μήνα (Ιανουάριος 2026): 159.39442606142416
```

Παρατηρήσεις: PCA και CFA συνοψίζουν ή μειώνουν την διάσταση, αλλά μπορεί να χάνεται κάποιο σήμα με σχετικό με τον στόχο. Η RFE, με επιλογή του σημαντικότερου χαρακτηριστικού (close_t-1), απέδωσε καλύτερα στο dataset μας. Όλα τα μοντέλα χρησιμοποιούν Ridge Regression για σταθερότητα και αποφυγή overfitting. Η επιλογή της μεθόδου μείωσης διάστασης επηρεάζει άμεσα την ακρίβεια πρόβλεψης ειδικά σε δεδομένα με περιορισμένο πλήθος παραμέτρων.

Δ. *Δώστε πρόβλεψη τιμής για Δεκέμβριο 2025 και Γενάρη 2025*.

Κελί 7 του κώδικα:

```
Training set: (11, 11)
Validation set: (23, 11)
Τιμή κλεισίματος Δεκεμβρίου 2025: 166.52103333333332
```

Κελί 11(σύγκριση μοντέλων):

```
Καλύτερο μοντέλο: Ridge + RFE
Πρόβλεψη επόμενου μήνα (Ιανουάριος 2026): 159.39442606142416
```