Colab link:
https://colab.research.google.com/drive/1sGGgx23q1P7Bkv_oxuOLD1o4IX2S0OhQ?usp=sharing

GitHub:
https://github.com/chrisaMel/Machine-Learning.git

Χρυσούλα Μελαδάκη

lis21102

Μηχανική Μάθηση

17 January 2025

Αναφορά εργασίας

Εισαγωγή

Το Πρόβλημα

Τα κρυπτονομίσματα, με κύριο εκπρόσωπο το Bitcoin, χαρακτηρίζονται από έντονη αστάθεια στις τιμές τους και πολύπλοκα μοτίβα αγοράς. Αυτή η απρόβλεπτη φύση δημιουργεί προκλήσεις για τους επενδυτές που επιδιώκουν να βελτιστοποιήσουν τις στρατηγικές τους και να μειώσουν τους κινδύνους. Η αξιόπιστη πρόβλεψη των τιμών κρυπτονομισμάτων είναι κρίσιμη για την αξιοποίηση ευκαιριών και τη μείωση των οικονομικών απωλειών.

Η Σημασία της Μελέτης

Το Bitcoin και τα κρυπτονομίσματα γενικότερα δεν αποτελούν μόνο οικονομικά περιουσιακά στοιχεία, αλλά και τεχνολογικές καινοτομίες που μετασχηματίζουν τον χρηματοοικονομικό τομέα. Η αποκεντρωμένη φύση τους και η χρήση της τεχνολογίας blockchain τα καθιστούν μοναδικά και ελκυστικά τόσο για μεμονωμένους επενδυτές όσο και για θεσμικούς παίκτες. Αξιόπιστα μοντέλα πρόβλεψης τιμών μπορούν να βελτιώσουν την αποτελεσματικότητα της αγοράς, να καθοδηγήσουν επενδυτικές αποφάσεις και να υποστηρίξουν την ανάπτυξη κανονιστικών πλαισίων.

Θεμελιώδης Έρευνα

Η παρούσα αναφορά βασίζεται σε μεθοδολογίες και ευρήματα προηγούμενων μελετών

σχετικά με τη χρήση της μηχανικής μάθησης στις χρηματοοικονομικές αγορές. Σημαντικές πηγές περιλαμβάνουν:

- **Κουτσαβδής (2022):** Αναλύει τη χρήση αλγορίθμων μηχανικής μάθησης και τεχνητών νευρωνικών δικτύων για την πρόβλεψη της τιμής του Bitcoin, υπογραμμίζοντας την αποτελεσματικότητα των μοντέλων LSTM και ARIMA.
- **Ζενούνη (2022):** Συγκρίνει διάφορους αλγορίθμους μηχανικής μάθησης σε ασταθείς συνθήκες της αγοράς κρυπτονομισμάτων, συμπεριλαμβάνοντας Monte Carlo προσομοιώσεις για ενίσχυση της ακρίβειας και χρησιμοποιεί ως δεδομένα τεχνικούς δείκτες.

Οι μελέτες αυτές προσφέρουν κρίσιμες γνώσεις για την εφαρμογή προχωρημένων αλγορίθμων σε σύνθετα και δυναμικά δεδομένα.

Συγκριτική Ανάλυση Τεχνικών

Στην ανάλυση εφαρμόστηκαν παραδοσιακά στατιστικά μοντέλα όπως το ARIMA, καθώς και σύγχρονοι αλγόριθμοι μηχανικής μάθησης όπως Random Forest και LSTM. Τα κύρια ευρήματα περιλαμβάνουν:

- **Random Forest:** Ανέδειξε εξαιρετική ακρίβεια στην κατανόηση πολύπλοκων σχέσεων αγοράς, αν και απαιτεί σημαντικούς υπολογιστικούς πόρους.
- **Linear Regression:** Παρείχε αξιόπιστα αποτελέσματα για πιο σταθερά δεδομένα, προσφέροντας γρήγορη εναλλακτική με αποδεκτή ακρίβεια.
- **LSTM:** Ξεχώρισε στην αναγνώριση σύνθετων μοτίβων σε χρονοσειρές, αν και με μεγαλύτερο υπολογιστικό κόστος.
- **ARIMA:** Ήταν αποτελεσματικό για στατικές χρονοσειρές, αλλά υστερούσε σε πιο δυναμικές και μη γραμμικές συνθήκες.

Αυτή η πολυδιάστατη προσέγγιση προσφέρει μια λεπτομερή κατανόηση των δυνατοτήτων και περιορισμών κάθε μεθόδου στο πλαίσιο πρόβλεψης των τιμών κρυπτονομισμάτων.

Αναλυτική Αναφορά του κώδικα

1. Δεδομένα

Αρχικά, έγινε εισαγωγή του αρχείου δεδομένων με την ονομασία BTC_USD Bitfinex Historical Data LargeDataset.csv. Το αρχείο περιέχει λεπτομερή ιστορικά δεδομένα της ισοτιμίας BTC/USD από την ιστοσελίδα investing.com, συμπεριλαμβανομένων των τιμών ανοίγματος, υψηλών, χαμηλών, κλεισίματος, του όγκου συναλλαγών και της ποσοστιαίας μεταβολής ανά ημέρα. Η ανάλυση αυτών των δεδομένων παρέχει σημαντικές πληροφορίες για τη δυναμική της αγοράς και αποτελεί τη βάση για την ανάπτυξη προηγμένων μοντέλων πρόβλεψης.

2. Προκαταρκτική Επεξεργασία Δεδομένων

Για τη σωστή ανάλυση των δεδομένων, πραγματοποιήθηκαν οι ακόλουθες ενέργειες:

- Αντικατάσταση του κόμματος με τελεία στις στήλες αριθμητικών τιμών και μετατροπή τους σε τύπο float. Αυτό εξασφαλίζει τη συμβατότητα των δεδομένων με μαθηματικές πράξεις και μοντέλα.
- Μετατροπή της στήλης ημερομηνίας (Date) σε τύπο datetime και ορισμός της ως δείκτη του DataFrame. Αυτό επιτρέπει την αποτελεσματική διαχείριση και οπτικοποίηση των χρονοσειρών.
- Αφαίρεση του συμβόλου ποσοστού (%) από τη στήλη Change % και μετατροπή της σε αριθμητική μορφή, διασφαλίζοντας ακρίβεια στις μετέπειτα αναλύσεις.

3. Υπολογισμός Τεχνικών Δεικτών

Για την ανάλυση της χρονοσειράς, υπολογίστηκαν βασικοί τεχνικοί δείκτες που

χρησιμοποιούνται συχνά στη χρηματοοικονομική ανάλυση, οι οποίοι παρέχουν πληροφορίες σχετικά με τη συμπεριφορά της αγοράς και διευκολύνουν τη λήψη αποφάσεων:

- **Κινητός Μέσος Όρος 10 ημερών (MA_10):** Υπολογίστηκε ο μέσος όρος της τιμής κλεισίματος για τα τελευταία 10 ημερήσια δεδομένα, παρέχοντας ενδείξεις για τις βραχυπρόθεσμες τάσεις.
- **Δείκτης Σχετικής Ισχύος (RSI):** Υπολογίστηκε για χρονικό παράθυρο 14 ημερών, καταγράφοντας τη δυναμική της τιμής σε σχέση με προηγούμενες μεταβολές.
- **Ρυθμός Αλλαγής (ROC):** Υπολογίστηκε για περίοδο 10 ημερών, μετρώντας τη σχετική μεταβολή της τιμής σε συγκεκριμένο χρονικό διάστημα.
- **Όγκος Ισορροπίας (OBV):** Υπολογίστηκε με βάση τη διακύμανση της τιμής και τον όγκο συναλλαγών, παρέχοντας πληροφορίες για τις σχέσεις όγκου και τιμής.

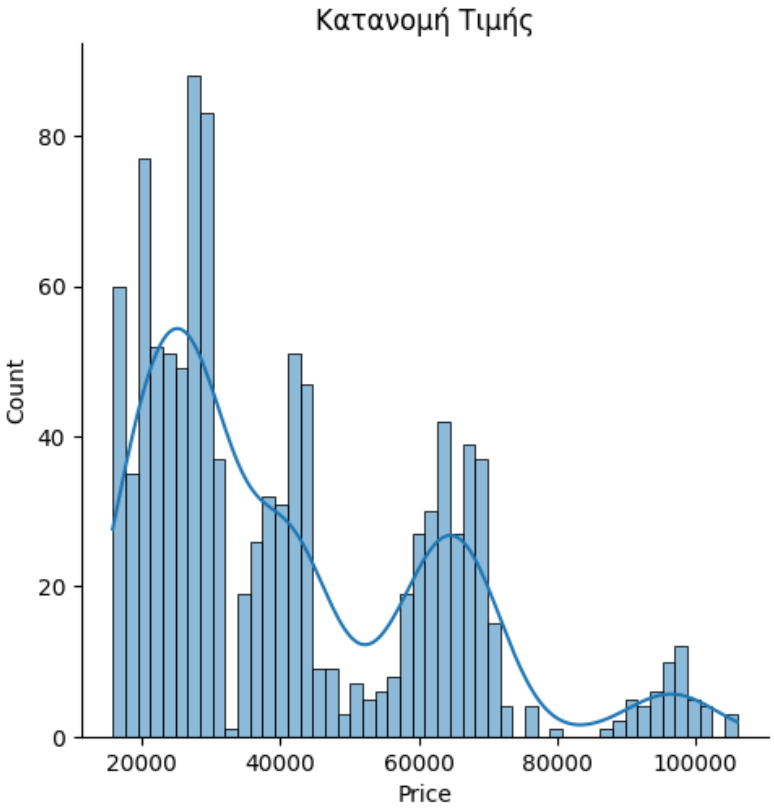
4. Διαχείριση Ελλιπών Τιμών

Λόγω της φύσης των τεχνικών δεικτών, ορισμένες τιμές ήταν ελλιπείς (NaN) στα αρχικά δείγματα. Αυτές οι τιμές προέκυψαν κυρίως από την εφαρμογή δεικτών που απαιτούν ιστορικά δεδομένα. Οι ελλιπείς τιμές αφαιρέθηκαν προσεκτικά από το σύνολο δεδομένων για να διασφαλιστεί η ορθότητα και η αξιοπιστία των μετέπειτα αναλύσεων.

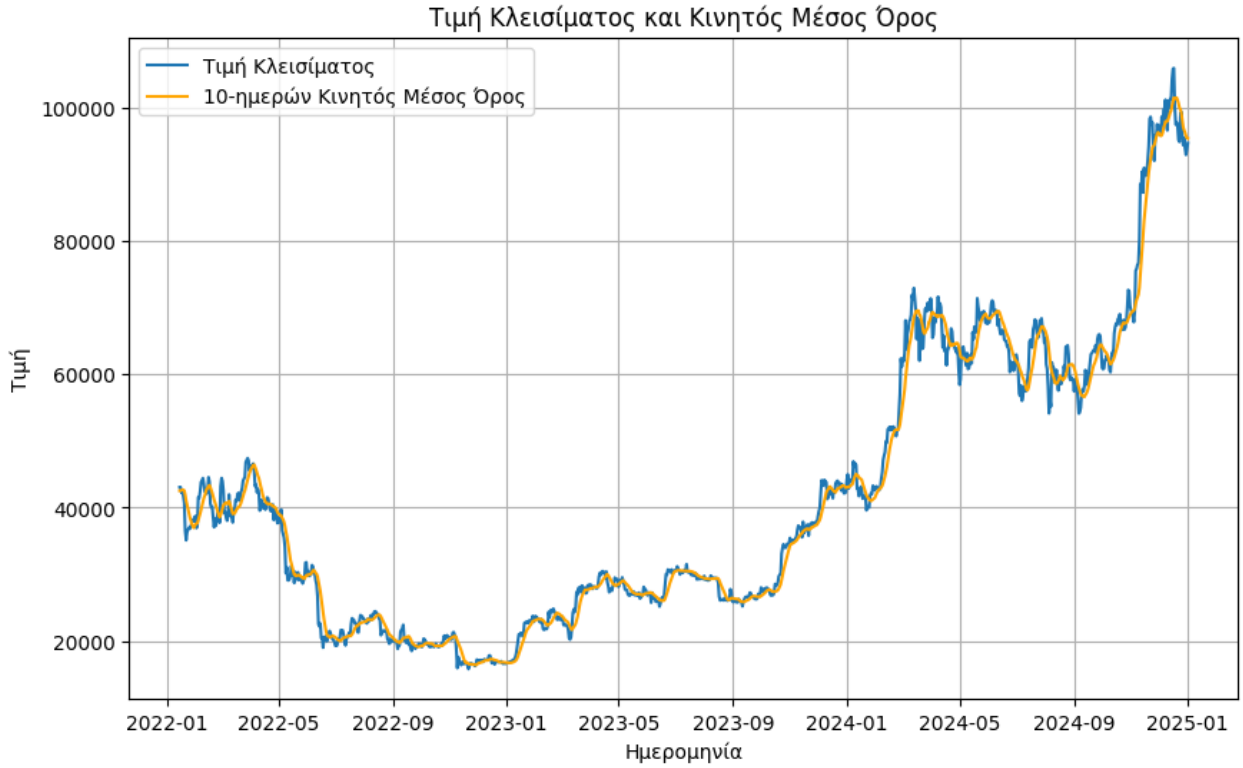
5. Οπτικοποίηση Δεδομένων

Πραγματοποιήθηκε οπτικοποίηση των δεδομένων με στόχο την καλύτερη κατανόηση της κατανομής των τιμών και του όγκου συναλλαγών, καθώς και της χρονοσειράς της τιμής κλεισίματος. Τα παρακάτω γραφήματα αποτέλεσαν βασικά εργαλεία κατανόησης:

- Ιστόγραμμα της τιμής κλεισίματος: Παρουσιάζει τη συχνότητα εμφάνισης τιμών, επιτρέποντας την αξιολόγηση της κατανομής.



- Γραφική απεικόνιση της χρονοσειράς της τιμής κλεισίματος και του κινητού μέσου όρου 10 ημερών: Δίνει έμφαση στις βραχυπρόθεσμες και μακροπρόθεσμες τάσεις.

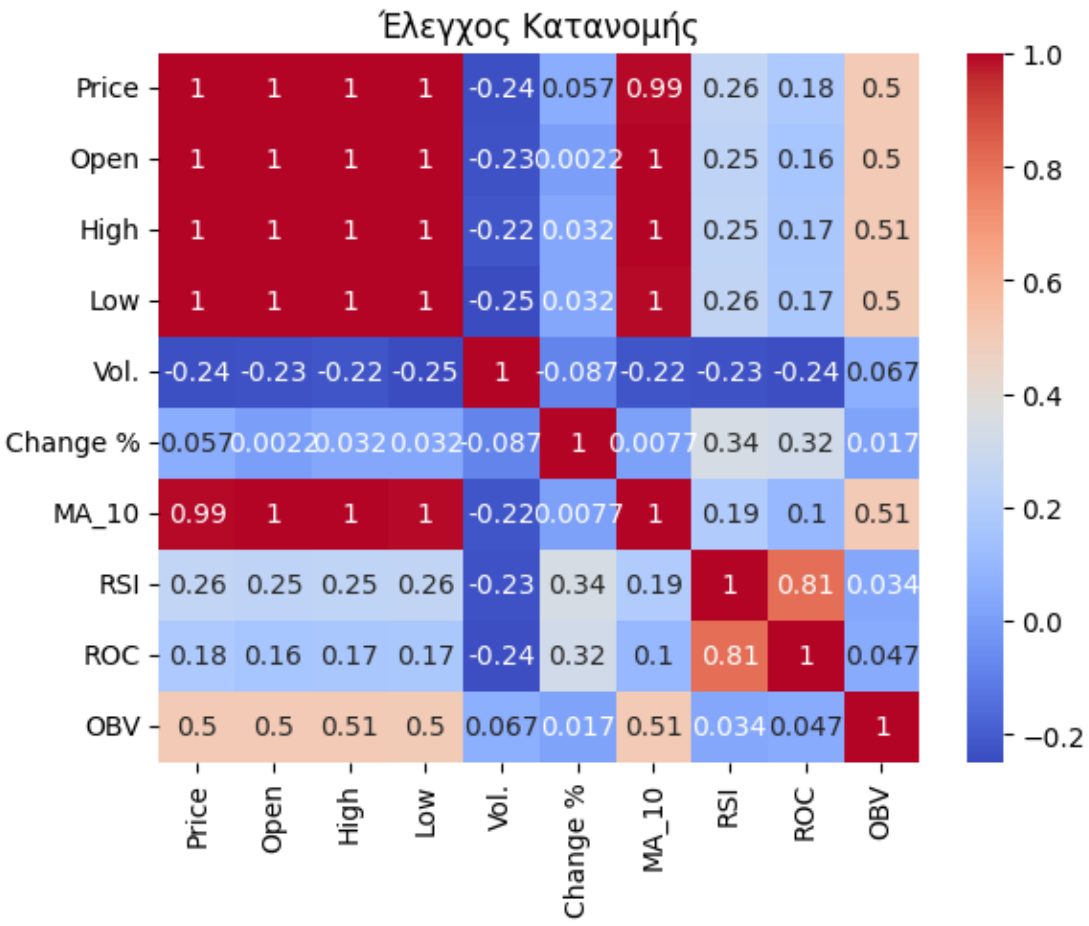


6. Κανονικοποίηση Δεδομένων

Για τη βελτιστοποίηση των αλγορίθμων μηχανικής μάθησης που θα ακολουθήσουν, τα δεδομένα κανονικοποιήθηκαν με τη χρήση της μεθόδου Min-Max Scaling. Αυτή η διαδικασία μείωσε τις διαφορές μεγέθους μεταξύ των χαρακτηριστικών, βελτιώνοντας τη σύγκλιση των αλγορίθμων και την αξιοπιστία των αποτελεσμάτων.

7. Έλεγχος Συσχέτισης Μεταβλητών

Έγινε ενδελεχής έλεγχος συσχέτισης μεταξύ των χαρακτηριστικών με τη χρήση διαγράμματος θερμότητας (heatmap). Από την ανάλυση προέκυψε ισχυρή συσχέτιση μεταξύ ορισμένων χαρακτηριστικών (π.χ. Price, Open, High, Low). Αυτό οδήγησε στην απόφαση να αφαιρεθούν αυτές οι στήλες από το σύνολο εκπαίδευσης, διατηρώντας μόνο τα πιο ανεξάρτητα χαρακτηριστικά για την πρόβλεψη.



Το τελικό σύνολο δεδομένων εκπαίδευσης αποτελείται από τα εξής χαρακτηριστικά:

- Όγκος συναλλαγών (Vol.): Παρέχει πληροφορίες για τη δραστηριότητα της αγοράς.
- Ποσοστιαία μεταβολή τιμής (Change %): Αναδεικνύει τις ημερήσιες διακυμάνσεις.
- Κινητός μέσος όρος 10 ημερών (MA_10): Αντικατοπτρίζει τη βραχυπρόθεσμη τάση.
- Δείκτης σχετικής ισχύος (RSI): Εντοπίζει υπεραγορασμένες ή υπερπουλημένες συνθήκες.
- Ρυθμός αλλαγής (ROC): Εντοπίζει ταχύτητα αλλαγής στην αγορά.
- Όγκος ισορροπίας (OBV): Αναδεικνύει τη σχέση μεταξύ όγκου και τάσης τιμών.

Εκπαίδευση Μοντέλων

8.Προετοιμασία για Εκπαίδευση

Αρχικά, ορίστηκε η μεταβλητή “Target”, η οποία περιλαμβάνει την τιμή της επόμενης ημέρας. Ακολούθως, το dataset διαχωρίστηκε σε σύνολα εκπαίδευσης και δοκιμής (80%-20%) και δημιουργήθηκαν 5 K-folds για τη διαδικασία cross-validation. Η διαδικασία αυτή εγγυάται πιο αξιόπιστες εκτιμήσεις απόδοσης, μειώνοντας τον κίνδυνο υπερεκπαίδευσης.

9.Εκπαίδευση Απλών Μοντέλων

Η εκπαίδευση πραγματοποιήθηκε για τα εξής μοντέλα:

- **Linear Regression**
- **Random Forest**
- **SVM**

Για κάθε μοντέλο εφαρμόστηκε cross-validation, και υπολογίστηκαν μετρικές όπως το Μέσο Απόλυτο Σφάλμα (MAE), το Μέσο Τετραγωνικό Σφάλμα (MSE), η Ρίζα Μέσου Τετραγωνικού Σφάλματος (RMSE) και ο Συντελεστής R^2, παρέχοντας σαφή εικόνα της απόδοσής τους.

10. Εκπαίδευση Σύνθετων Μοντέλων

LSTM

Για την ανάλυση χρονοσειρών, χρησιμοποιήθηκε το μοντέλο Long Short-Term Memory (LSTM). Τα δεδομένα αναδιαμορφώθηκαν για να ταιριάζουν στις απαιτήσεις του LSTM, και η εκπαίδευση πραγματοποιήθηκε με ενισχυμένες τεχνικές cross-validation. Το LSTM καταφέρνει να αναγνωρίζει πολύπλοκα μοτίβα στις χρονοσειρές, αν και απαιτεί μεγαλύτερη υπολογιστική ισχύ.

ARIMA

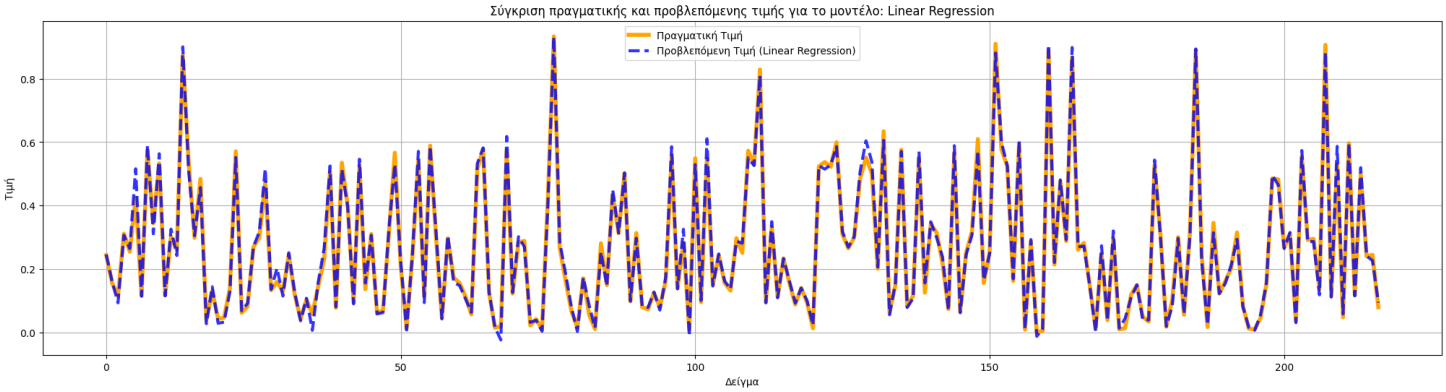
Για τη στατιστική ανάλυση των χρονοσειρών, εφαρμόστηκε το μοντέλο ARIMA, το οποίο αποτελεί ένα από τα πιο διαδεδομένα εργαλεία για την πρόβλεψη χρονοσειρών. Το ARIMA (AutoRegressive Integrated Moving Average) είναι ένα μοντέλο που συνδυάζει αυτοπαλινδρομήσεις, ολοκληρώσεις και κινητούς μέσους όρους για τη μοντελοποίηση στατικών χρονοσειρών. Προηγουμένως, εφαρμόστηκε ο έλεγχος Dick-ey-Fuller (ADF), ο οποίος είναι μια στατιστική μέθοδος που εξετάζει αν μια χρονοσειρά είναι στατική ή αν απαιτείται μετασχηματισμός για τη σταθεροποίησή της. Ο έλεγχος έδειξε ότι η χρονοσειρά απαιτούσε περαιτέρω σταθεροποίηση για την αποτελεσματική εφαρμογή του ARIMA. Παρόλο που το ARIMA αποδείχθηκε αποτελεσματικό σε συγκεκριμένα σενάρια, οι περιορισμοί του, όπως η εξάρτηση από την υπόθεση της στασιμότητας, έγιναν εμφανείς στις δυναμικές και πολύπλοκες αλλαγές της συγκεκριμένης χρονοσειράς.

Αποτελέσματα

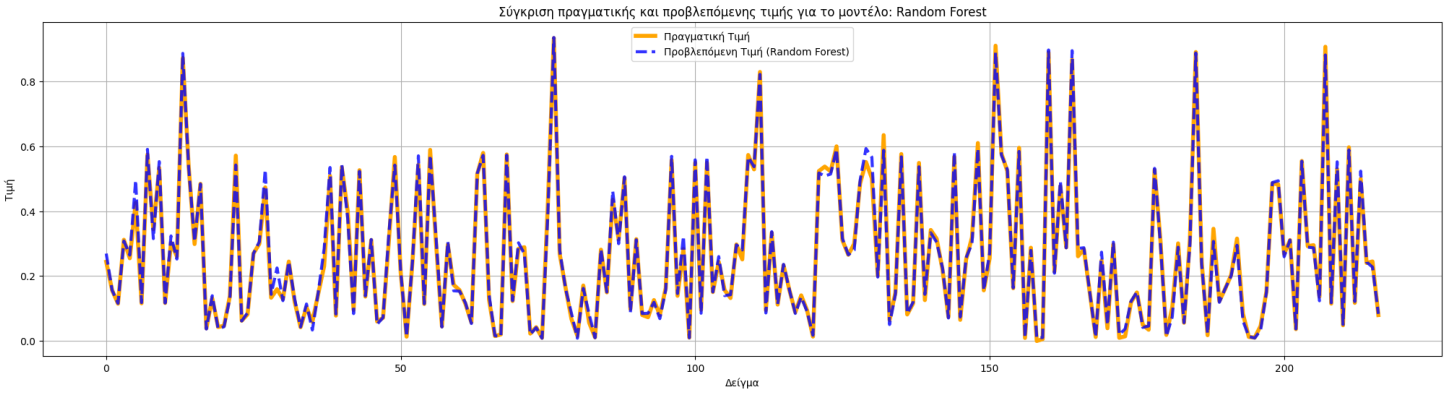
Σύγκριση Πραγματικής και Προβλεπόμενης Τιμής

Οι παρακάτω παρατηρήσεις αφορούν την απόδοση των μοντέλων:

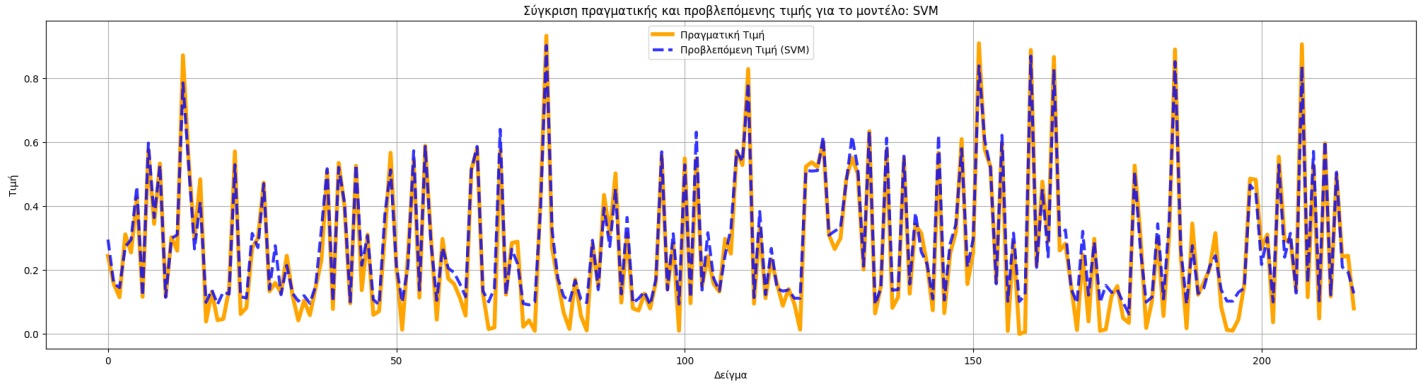
- **Linear Regression:** Αξιόπιστο στις σταθερές χρονοσειρές, με μικρές αποκλίσεις στις προβλέψεις.



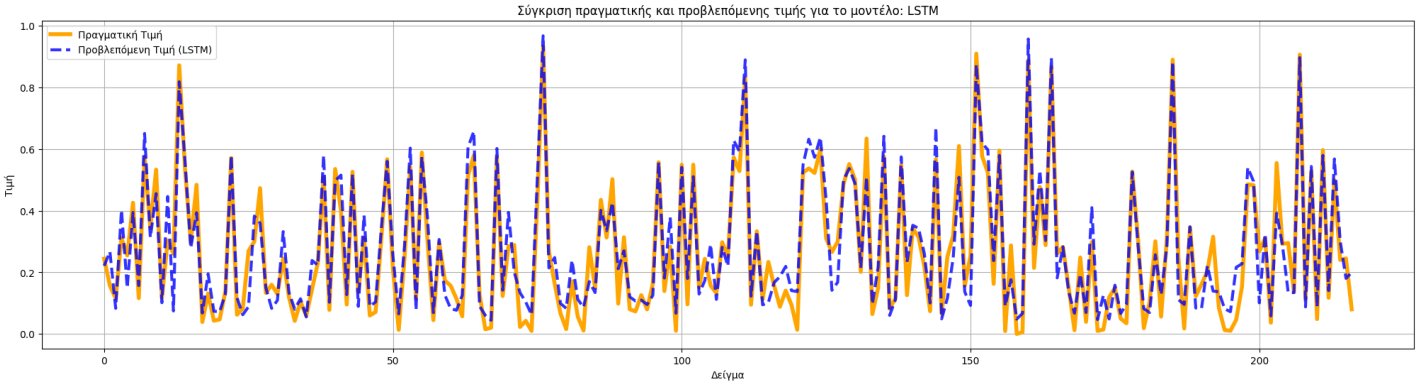
- **Random Forest:** Εξαιρετική ακρίβεια σε διακυμάνσεις της αγοράς, αποδεικνύοντας την αξία του σε σύνθετα δεδομένα.



- **SVM:** Καλή απόδοση, αν και αντιμετωπίζει προκλήσεις σε ακραίες τιμές.

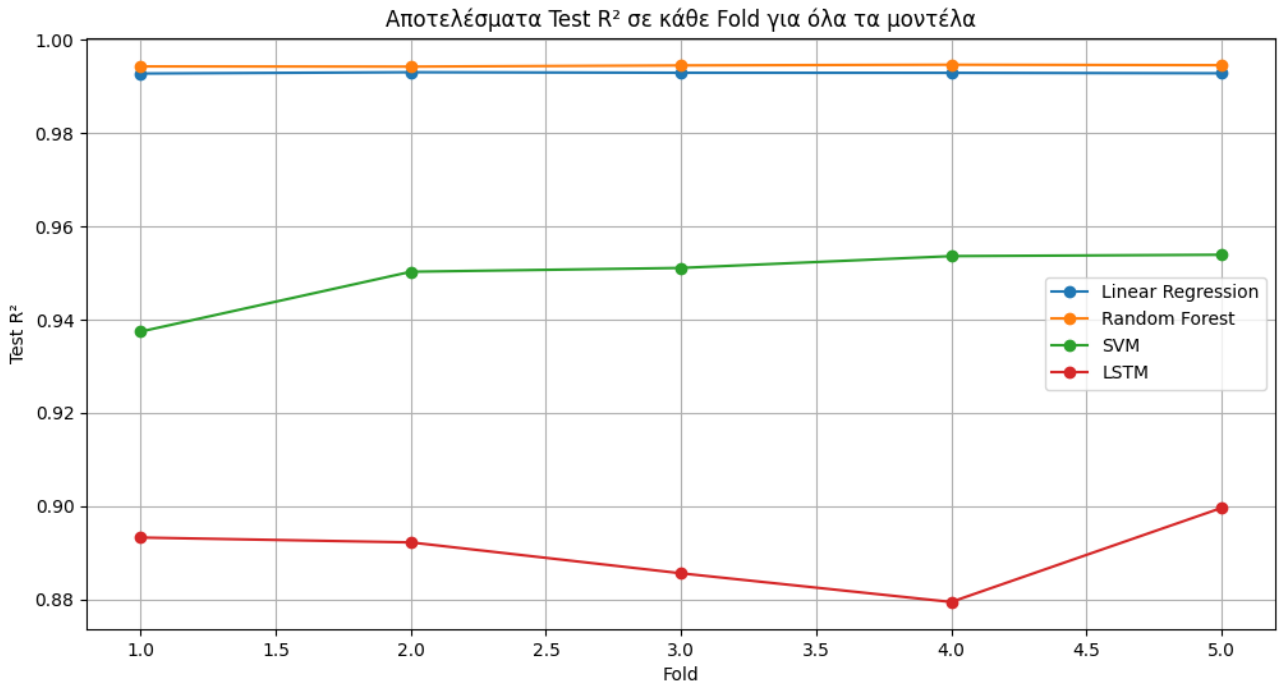


- **LSTM:** Καταγράφει λεπτομερώς μοτίβα, αλλά εμφανίζει αποκλίσεις σε μη αναμενόμενες συνθήκες.



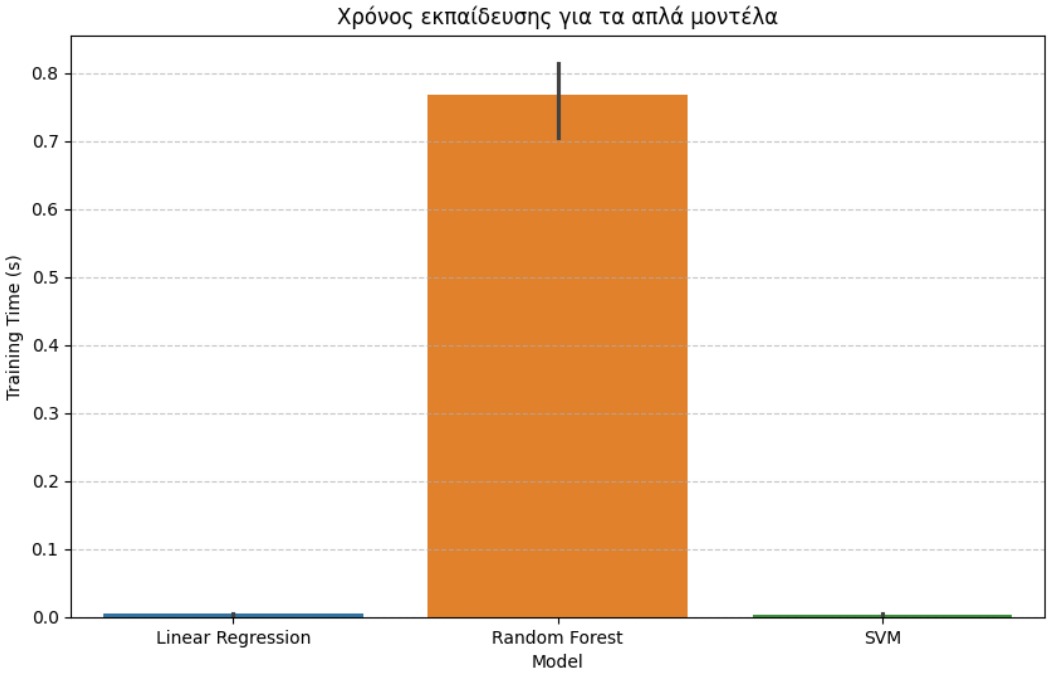
Ανάλυση Μετρικών Απόδοσης για K-Folds

- **Test MAE:** Το Random Forest υπερέχει, αποδεικνύοντας τη χαμηλότερη απόκλιση στις προβλέψεις.
- **Test MSE:** Παρόμοια με το MAE, το Random Forest παραμένει κυρίαρχο.
- **Test RMSE:** Οι χαμηλές τιμές RMSE επιβεβαιώνουν την ακρίβεια του Random Forest.
- **Test R^2:** Οι υψηλές τιμές του Random Forest αποδεικνύουν την εξαιρετική εξήγηση της μεταβλητότητας.

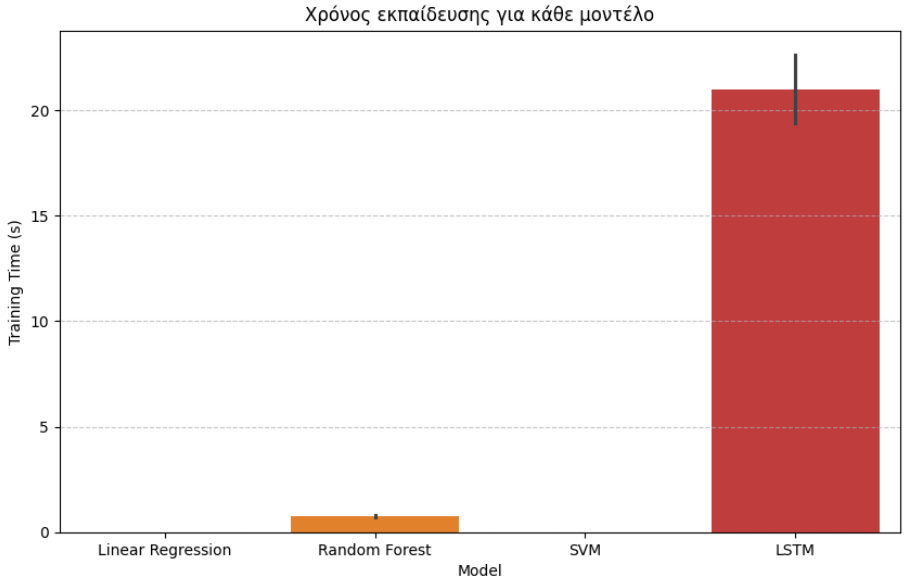


Ανάλυση Χρόνου Εκπαίδευσης

- **Χρόνος Εκπαίδευσης για Απλά Μοντέλα:** Το Random Forest απαιτεί περισσότερο χρόνο, λόγω της πολυπλοκότητάς του.

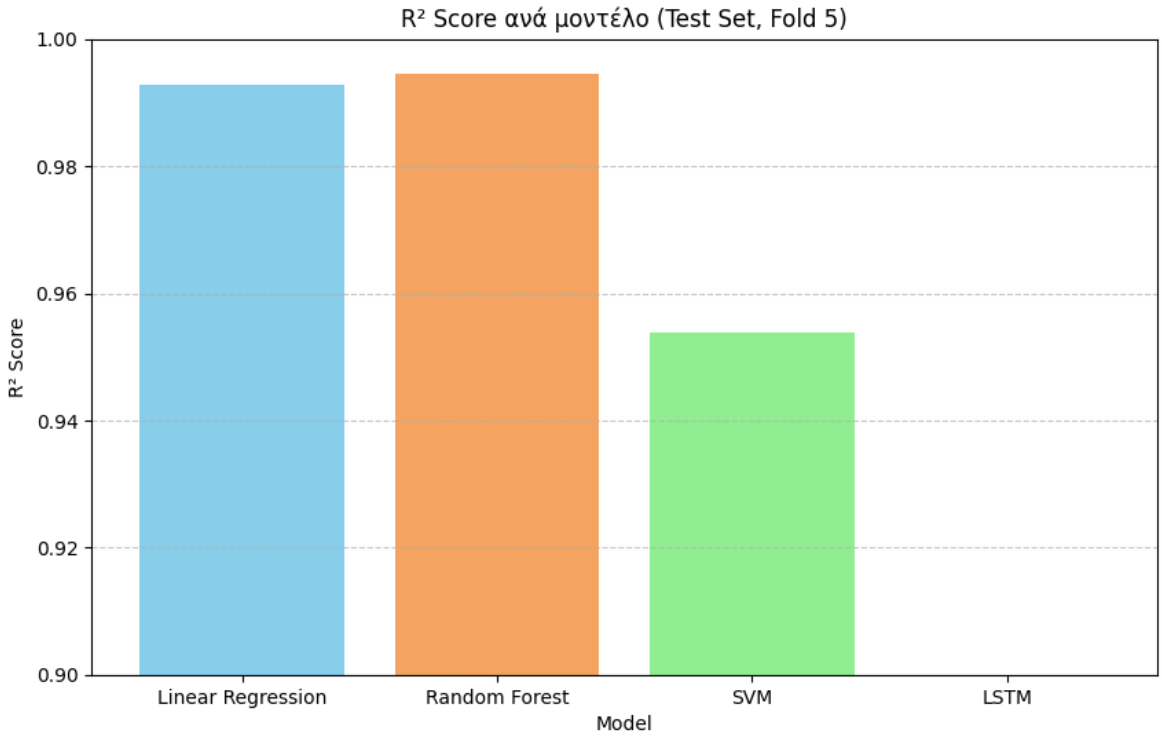


- **Χρόνος Εκπαίδευσης για Σύνθετα Μοντέλα:** Το LSTM καταγράφει τη μεγαλύτερη καθυστέρηση, αντανακλώντας την πολυπλοκότητά του.

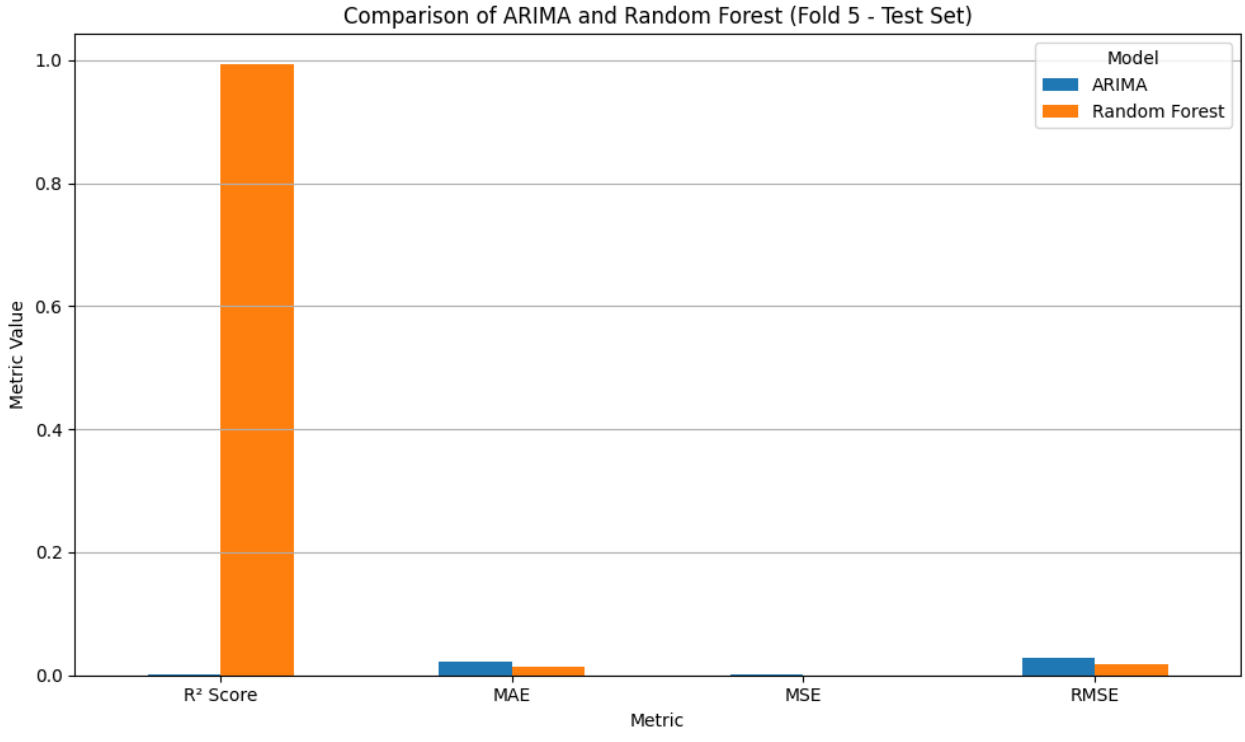


Συγκριτική Ανάλυση Μετρικών

- **MAE, MSE, RMSE για το Test Set:** Το Random Forest αποδεικνύει την ανωτερότητά του.



- Συγκριτική Απόδοση μεταξύ ARIMA και Random Forest: Η απόδοση του ARIMA είναι αισθητά μικρότερη.



Συνολικά, τα αποτελέσματα της ανάλυσης καταδεικνύουν ότι το Random Forest είναι ένα εξαιρετικά αποδοτικό μοντέλο, ιδανικό για την πρόβλεψη διακυμάνσεων σε δεδομένα με σύνθετες σχέσεις. Η δυνατότητά του να ενσωματώνει πληροφορίες από πολλαπλά δέντρα απόφασης το καθιστά ανθεκτικό και αξιόπιστο, ιδιαίτερα σε δυναμικά περιβάλλοντα. Παρά το γεγονός ότι το Random Forest απαιτεί περισσότερο χρόνο εκπαίδευσης σε σχέση με άλλα μοντέλα, η ανώτερη ακρίβειά του και η ικανότητά του να διαχειρίζεται πολύπλοκα δεδομένα το καθιστούν την καλύτερη επιλογή για πολλές περιπτώσεις. Παράλληλα, το LSTM αναδεικνύεται σε απαραίτητο εργαλείο για την ανάλυση χρονοσειρών με πολύπλοκα μοτίβα, παρέχοντας λεπτομερείς προβλέψεις, αν και απαιτεί σημαντικούς υπολογιστικούς πόρους. Από την άλλη πλευρά, η Γραμμική Παλινδρόμηση παρείχε αξιόπιστα αποτελέσματα, ειδικά σε δεδομένα με πιο σταθερές σχέσεις, καθιστώντας την μια γρήγορη και εύχρηστη επιλογή για βασικές προβλέψεις. Το ARIMA συνεχίζει να αποτελεί μια ισχυρή επιλογή για στατιστική μοντελοποίηση χρονοσειρών, με την ακρίβειά του να εξαρτάται από τη σωστή προσαρμογή των παραμέτρων και την εξασφάλιση της στασιμότητας. Ωστόσο, η ανάγκη για περαιτέρω προσαρμογές στο ARIMA επισημαίνει τους περιορισμούς του σε πιο περίπλοκες και μη γραμμικές χρονοσειρές, καθιστώντας το περισσότερο κατάλληλο για απλούστερα σενάρια. Με βάση αυτά τα ευρήματα, η επιλογή του κατάλληλου μοντέλου θα πρέπει να καθοδηγείται από τη φύση και την πολυπλοκότητα των δεδομένων, καθώς και από τους διαθέσιμους υπολογιστικούς πόρους.