



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ
ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
“ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ & ΥΠΗΡΕΣΙΕΣ”

**Πρόβλεψη της Τιμής του Bitcoin με Χρήση Μηχανικής Μάθησης,
Βαθιάς Μάθησης και Ανάλυσης Συναισθήματος**

Από
Γιώργος
Φουτζόπουλος
Υποβάλλεται
για την εκπλήρωση των προϋποθέσεων λήψης
Μεταπτυχιακού Διπλώματος
στην ειδίκευση «ΜΔΑ»
του ΠΜΣ “Πληροφοριακά Συστήματα & Υπηρεσίες”
στο
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Ιούλιος 2026

Επιβλέπων/Επιβλέπουσα: Μιχαήλ Φιλιππάκης
Ακαδημαϊκή Θέση: Καθηγητής

Πανεπιστήμιο Πειραιώς. Κάτοχος όλων των δικαιωμάτων
University of Piraeus,. All rights reserved.

Συγγραφέας / Author

ΣΕΛΙΔΑ ΕΓΚΥΡΟΤΗΤΑΣ

Ονοματεπώνυμο Φοιτητή/Φοιτήτριας:

Τίτλος Μεταπτυχιακής Διπλωματικής Εργασίας:

Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία υποβάλλεται ως μερική εκπλήρωση των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών “Πληροφοριακά Συστήματα & Υπηρεσίες” του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς και εγκρίθηκε στις [ημερομηνία έγκρισης] από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή

Επιβλέπων/ουσα (Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο

Πειραιώς).....[ονοματεπώνυμο, βαθμίδα]

Μέλος Εξεταστικής Επιτροπής:[ονοματεπώνυμο, βαθμίδα]

Μέλος Εξεταστικής Επιτροπής:[ονοματεπώνυμο, βαθμίδα]

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΑΥΘΕΝΤΙΚΟΤΗΤΑΣ

Ο/Η....., γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «.....», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Επιπλέον δηλώνω υπεύθυνα ότι η συγκεκριμένη Μεταπτυχιακή Διπλωματική Εργασία έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει αξιολογηθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου. Σε κάθε περίπτωση, αναληθούς ή ανακριβούς δηλώσεως, υπόκειμαι στις συνέπειες που προβλέπονται τις διατάξεις που προβλέπει η Ελληνική και Κοινοτική Νομοθεσία περί πνευματικής ιδιοκτησίας.

Ο/Η ΔΗΛΩΝ/ΟΥΣΑ

Ονοματεπώνυμο:

Αριθμός Μητρώου:

Υπογραφή:

Περιεχόμενα

Λίστα διαγραμμάτων	5
Λίστα πινάκων.....	6
Λίστα εξισώσεων	7
Περίληψη	8
Abstract.....	8
1. Εισαγωγή.....	10
2. Βιβλιογραφική ανασκόπηση.....	12
2.1 Πρόβλεψη χρηματοοικονομικών χρονοσειρών και κρυπτονομισμάτων.....	12
2.2 Βαθιά μάθηση και προηγμένες αρχιτεκτονικές στη πρόβλεψη bitcoin και κρυπτονομισμάτων.....	14
2.3 Sentiment analysis, ειδήσεις και εξωτερική πληροφορία στη πρόβλεψη κρυπτονομισμάτων	16
2.4 Μακροοικονομική, συστημικοί και εξωτερικοί παράγοντες.....	18
2.5 Συμπέρασμα βιβλιογραφικής ανασκόπησης και ερευνητικό κενό.....	19
3. Μεθοδολογία και μοντέλα πρόβλεψης.....	22
3.1 Μοντέλα πρόβλεψης.....	23
3.2 Στατιστικά Μοντέλα Αναφοράς.....	23
3.3 Παλινδρόμηση Ridge (Ridge Regression)	24
3.4 Παλινδρόμηση Τυχαίων Δασών (Random Forest Regression)	25
3.5 Η παλινδρόμηση XGBoost (XGBoost Regression).....	27
3.6 Δίκτυο Μακράς Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory - LSTM) και Δίκτυο Πυλών Επαναλαμβανόμενης Μονάδας (Gated Recurrent Unit - GRU).....	28
3.7 Ανάλυση συναισθήματος από κείμενα (sentiment analysis).....	30
4. Μεθοδολογία έρευνας	32
4.1 Δεδομένα.....	32
4.2 Έλεγχος ποιότητας δεδομένων.....	35
4.3 Μετασχηματισμός μεταβλητής στόχου	43

4.4	Διαχωρισμός δεδομένων σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου	46
4.5	Συλλογή δεδομένων συναισθήματος από κοινωνικά δίκτυα και ειδησεογραφικές πηγές 48	
5.	Αποτελέσματα έρευνας	56
5.1	Αποτελέσματα χωρίς χρήση μεταβλητών συναισθήματος.....	56
5.1.1	Πρόβλεψη χρονικού διαστήματος 1 λεπτού.....	56
5.1.2	Πρόβλεψη χρονικού διαστήματος 15 λεπτών	60
5.1.3	Πρόβλεψη χρονικού διαστήματος 30 λεπτών	64
5.2	Αποτελέσματα ανάλυσης συναισθήματος.....	68
5.3	Αποτελέσματα με χρήση μεταβλητών συναισθήματος.....	71
5.3.1	Πρόβλεψη χρονικού διαστήματος 1 λεπτού.....	72
5.3.2	Πρόβλεψη χρονικού διαστήματος 15 λεπτών	75
5.3.3	Πρόβλεψη χρονικού διαστήματος 30 λεπτών	78
6.	Επίλογος – Συμπεράσματα	82
7.	Μελλοντική έρευνα	88
	Βιβλιογραφία	90
	Παράρτημα	94

Λίστα διαγραμμάτων

Διάγραμμα 1. Απεικόνιση της τιμής κλεισίματος του Bitcoin κάθε λεπτό (2022-2026)	33
Διάγραμμα 2. Εξέλιξη της τιμής κλεισίματος του Bitcoin με επιλεγμένα γεγονότα	35
Διάγραμμα 3.Αποσύνθεση της χρονοσειράς.....	37
Διάγραμμα 4.Αποσύνθεση της χρονοσειράς με λογαριθμική απόδοση λεπτού	38
Διάγραμμα 5.Αυτοσυσχέτισης και μερικής αυτοσυσχέτισης στη τιμή κλεισίματος του bitcoin.	40
Διάγραμμα 6.Αυτοσυσχέτισης και μερικής αυτοσυσχέτισης της λογαριθμικής απόδοσης λεπτού... ..	41
Διάγραμμα 7. Αυτοσυσχέτιση και μερική αυτοσυσχέτιση της λογαριθμικής απόδοσης 15 λεπτών ..	42
Διάγραμμα 8. Αυτοσυσχέτιση και μερική αυτοσυσχέτιση της λογαριθμικής απόδοσης 30 λεπτών ..	42
Διάγραμμα 9.Απεικόνιση της λογαριθμικής απόδοσης του Bitcoin κάθε λεπτό (2022-2026).....	45
Διάγραμμα 10. Πλήθος tweets ανά ώρα ημέρας	50
Διάγραμμα 11. Δημοσιεύσεις και σχόλια ανά υποκατηγορία (subreddit).....	51

Διάγραμμα 12. Πλήθος δημοσιεύσεων και σχολίων ανά ώρα ημέρας.....	52
Διάγραμμα 13. Ειδησεογραφικές πηγές.....	53
Διάγραμμα 14. Πλήθος ειδήσεων ανά ώρα ημέρας	54
Διάγραμμα 15. Γραφική παράσταση προβλεπόμενης και πραγματικής τιμής BTC για χρονικό ορίζοντα 1 λεπτού για το ARIMA	59
Διάγραμμα 16. Ακρίβεια πρόβλεψης κατεύθυνσης ανά μοντέλο για χρονικό ορίζοντα 1 λεπτού	60
Διάγραμμα 17. Γραφική παράσταση προβλεπόμενης και πραγματικής τιμής BTC για χρονικό ορίζοντα 15 λεπτών για το ARIMA.....	63
Διάγραμμα 18. Ακρίβεια πρόβλεψης κατεύθυνσης ανά μοντέλο για χρονικό ορίζοντα 15 λεπτών ...	64
Διάγραμμα 19. Γραφική παράσταση προβλεπόμενης και πραγματικής τιμής BTC για χρονικό ορίζοντα 30 λεπτών για το ARIMA.....	67
Διάγραμμα 20. Ακρίβεια πρόβλεψης κατεύθυνσης ανά μοντέλο για χρονικό ορίζοντα 30 λεπτών ...	68
Διάγραμμα 21. Μέσες τιμές VADER scores ανά κατηγορία συναισθήματος	69
Διάγραμμα 22. Κατανομή compound score ανά κατηγορία συναισθήματος.....	70
Διάγραμμα 23. Γραφική παράσταση προβλεπόμενης και πραγματικής τιμής BTC για χρονικό ορίζοντα 1 λεπτού με χρήση μεταβλητών συναισθήματος για το Ridge Regression	74
Διάγραμμα 24. Ακρίβεια πρόβλεψης κατεύθυνσης ανά μοντέλο για χρονικό ορίζοντα 1 λεπτού με χρήση sentiment	75
Διάγραμμα 25. Γραφική παράσταση προβλεπόμενης και πραγματικής τιμής BTC για χρονικό ορίζοντα 15 λεπτών με χρήση μεταβλητών συναισθήματος για το Ridge Regression.....	77
Διάγραμμα 26. Ακρίβεια πρόβλεψης κατεύθυνσης ανά μοντέλο για χρονικό ορίζοντα 15 λεπτών με χρήση sentiment	78
Διάγραμμα 27. Γραφική παράσταση προβλεπόμενης και πραγματικής τιμής BTC για χρονικό ορίζοντα 30 λεπτών με χρήση μεταβλητών συναισθήματος για το LSTM	80
Διάγραμμα 28. Ακρίβεια πρόβλεψης κατεύθυνσης ανά μοντέλο για χρονικό ορίζοντα 30 λεπτών με χρήση sentiment	81

Λίστα πινάκων

Πίνακας 1. Αποτελέσματα ελέγχου ADF για την τιμή κλεισίματος και τις λογαριθμικές αποδόσεις. .	40
Πίνακας 2. Διαχωρισμός συνόλου δεδομένων	47
Πίνακας 3. Παράδειγμα ενοποιημένου πίνακα tweets	49
Πίνακας 4. Παράδειγμα ενοποιημένου πίνακα reddit δημοσιεύσεων και σχολίων	50
Πίνακας 5. Παράδειγμα ενοποιημένου πίνακα ειδήσεων	52

Πίνακας 6. Αποτελέσματα μοντέλων χωρίς ανάλυση συναισθήματος για χρονικό ορίζοντα 1 λεπτού	58
Πίνακας 7. Αποτελέσματα μοντέλων χωρίς ανάλυση συναισθήματος για χρονικό ορίζοντα 15 λεπτών	62
Πίνακας 8. Αποτελέσματα μοντέλων χωρίς ανάλυση συναισθήματος για χρονικό ορίζοντα 30 λεπτών	66
Πίνακας 9. Κατανομή κατηγοριών συναισθήματος.....	68
Πίνακας 10. Αποτελέσματα sentiment ανά πλατφόρμα	69
Πίνακας 11. Ενδεικτικά παραδείγματα θετικού, ουδέτερου και αρνητικού συναισθήματος βάσει VADER	71
Πίνακας 12. Αποτελέσματα μοντέλων με ανάλυση συναισθήματος για χρονικό ορίζοντα 1 λεπτού .	73
Πίνακας 13. Αποτελέσματα μοντέλων με ανάλυση συναισθήματος για χρονικό ορίζοντα 15 λεπτών	76
Πίνακας 14. Αποτελέσματα μοντέλων με ανάλυση συναισθήματος για χρονικό ορίζοντα 30 λεπτών	79
Πίνακας 15. Σύγκριση των αποτελεσμάτων πριν και μετά την προσθήκη μεταβλητών συναισθήματος	82
Πίνακας 16. Σύγκριση του Directional Accuracy πριν και μετά την προσθήκη μεταβλητών sentiment στο test set.....	83

Λίστα εξισώσεων

Εξίσωση 1. Μετατροπή προβλεπόμενης λογαριθμικής απόδοσης σε προβλεπόμενη τιμή κλεισίματος	23
Εξίσωση 2. Πρόβλεψη του μοντέλου μηδενικής απόδοσης	24
Εξίσωση 3. Γραμμική συνάρτηση πρόβλεψης παλινδρόμησης Ridge.....	25
Εξίσωση 4. λ2 ποινή κανονικοποίησης	25
Εξίσωση 5. Τελική πρόβλεψη του Random Forest ως μέσος όρος των δέντρων.....	25
Εξίσωση 6. Διαδοχική ενημέρωση του μοντέλου στη βαθμιαία ενίσχυση/XGBoost.....	27
Εξίσωση 7. Πρόβλεψη με ακολουθιακό παράθυρο εισόδου στα LSTM και GRU	29
Εξίσωση 8. Υπολογισμός απλής πρώτης διαφοράς της τιμής κλεισίματος.....	43
Εξίσωση 9. Υπολογισμός λογαριθμικής απόδοσης ενός λεπτού	43
Εξίσωση 10. Υπολογισμός μελλοντικής λογαριθμικής απόδοσης για χρονικό ορίζοντα h	45

Περίληψη

Η παρούσα μεταπτυχιακή διπλωματική εργασία εξετάζει την πρόβλεψη της τιμής του Bitcoin σε βραχυχρόνιους χρονικούς ορίζοντες, αξιοποιώντας στατιστικά μοντέλα, μοντέλα μηχανικής μάθησης, μοντέλα βαθιάς μάθησης και μεταβλητές ανάλυσης συναισθήματος. Η ανάλυση βασίστηκε σε δεδομένα υψηλής συχνότητας για την ισοτιμία BTC/USD, με χρονικό βήμα ενός λεπτού, για την περίοδο Ιανουαρίου 2022 έως Μαΐου 2026. Η μεταβλητή στόχος ορίστηκε ως η μελλοντική λογαριθμική απόδοση για χρονικούς ορίζοντες 1, 15 και 30 λεπτών, καθώς η λογαριθμική απόδοση κρίθηκε καταλληλότερη από την απόλυτη τιμή κλεισίματος λόγω της μη στασιμότητας της αρχικής χρονοσειράς. Για την πρόβλεψη εφαρμόστηκαν και συγκρίθηκαν τα μοντέλα αυτοπαλινδρομού κινητού μέσου (ARIMA), η παλινδρόμηση Ridge (Ridge Regression), η παλινδρόμηση τυχαίων δασών (Random Forest Regression), η παλινδρόμηση XGBoost (XGBoost Regression), το δίκτυο Μακράς Βραχυπρόθεσμης Μνήμης (Long Short Term Memory - LSTM) και το δίκτυο Πυλών Επαναλαμβανόμενης Μονάδας (Gated Recurrent Unit - GRU), ενώ η αξιολόγηση πραγματοποιήθηκε με μετρικές όπως η ρίζα μέσου τετραγωνικού σφάλματος (Root Mean Squared Error - RMSE), μέσου απόλυτου σφάλματος (Mean Absolute Error - MAE) και η Ακρίβεια κατεύθυνσης (Directional Accuracy), τόσο σε επίπεδο λογαριθμικών αποδόσεων όσο και σε επίπεδο προβλεπόμενης τιμής. Επιπλέον, συλλέχθηκαν και επεξεργάστηκαν δεδομένα κειμένου από Twitter, Reddit και ειδησεογραφικές πηγές, από τα οποία εξήχθησαν μεταβλητές θετικού, αρνητικού, ουδέτερου και συνολικού συναισθήματος με χρήση του εργαλείου VADER (Valence Aware Dictionary and sEntiment Reasoner), μια λεξικογραφικής μεθόδου ανάλυσης συναισθήματος. Οι μεταβλητές αυτές συγχωνεύθηκαν χρονικά με τα δεδομένα τιμών, ώστε να εξεταστεί αν η εξωτερική πληροφορία από τη δημόσια συζήτηση και την ειδησεογραφική ροή μπορεί να βελτιώσει την προβλεπτική ικανότητα των μοντέλων. Τα αποτελέσματα έδειξαν ότι η προσθήκη μεταβλητών συναισθήματος δεν οδηγεί σε καθολική βελτίωση όλων των μοντέλων και όλων των χρονικών οριζόντων, αλλά η επίδρασή διαφοροποιείται ανάλογα με το μοντέλο και τον ορίζοντα πρόβλεψης.

Abstract

This postgraduate dissertation examines the prediction of Bitcoin prices over short-term forecasting horizons by employing statistical models, machine learning models, deep learning models, and sentiment analysis variables. The analysis is based on high-frequency BTC/USD data with one-minute intervals, covering the period from January 2022 to May 2026. The target variable is defined as the future logarithmic return for forecasting horizons of 1, 15, and 30 minutes, as logarithmic

returns were considered more suitable than the absolute closing price due to the non-stationarity of the original price series. For the forecasting task, ARIMA, Ridge Regression, Random Forest, XGBoost, LSTM, and GRU models were implemented and compared, while model evaluation was conducted using RMSE, MAE, and Directional Accuracy, both at the level of logarithmic returns and predicted prices. In addition, textual data were collected and processed from Twitter, Reddit, and news sources, from which positive, negative, neutral, and compound sentiment variables were extracted using the VADER tool. These sentiment variables were temporally merged with the Bitcoin price data in order to examine whether external information from public discussion and news flow could improve the predictive performance of the models. The results showed that the inclusion of sentiment variables does not lead to a universal improvement across all models and forecasting horizons; rather, its effect varies depending on the model and the prediction horizon.

1. Εισαγωγή

Τα κρυπτονομίσματα τα τελευταία χρόνια αποτελούν ένα από τα πιο δυναμικά και ταυτόχρονα ασταθή τμήματα των χρηματοοικονομικών αγορών. Το Bitcoin, ως το πρώτο και πιο γνωστό από τα κρυπτονομίσματα, έχει προσελκύσει έντονο ενδιαφέρον τόσο από επενδυτές όσο και από την ακαδημαϊκή κοινότητα, κυρίως λόγω της υψηλής μεταβλητότητας, της συνεχούς διαπραγμάτευσης και της ευαισθησίας του σε ειδήσεις, ρυθμιστικές εξελίξεις, μακροοικονομικές συνθήκες και μεταβολές στο επενδυτικό συναίσθημα. Η πρόβλεψη της τιμής του αποτελεί σύνθετο πρόβλημα, καθώς η αγορά χαρακτηρίζεται από έντονο θόρυβο, μη στασιμότητα και μη γραμμικές σχέσεις.

Η παρούσα μεταπτυχιακή διπλωματική εργασία εστιάζει στην ανάπτυξη και αξιολόγηση μοντέλων πρόβλεψης της βραχυχρόνιας μεταβολής της τιμής του Bitcoin, αξιοποιώντας δεδομένα υψηλής συχνότητας για την ισοτιμία BTC/USD. Ειδικότερα, αξιοποιούνται δεδομένα ανά λεπτό, τα οποία, πέραν της τιμής κλεισίματος (close), περιλαμβάνουν πληροφορίες σχετικά με την τιμή ανοίγματος (open), την υψηλότερη τιμή (high), τη χαμηλότερη τιμή (low) και τον όγκο συναλλαγών (volume). Το πρόβλημα πρόβλεψης δεν διατυπώνεται ως άμεση εκτίμηση της μελλοντικής απόλυτης τιμής κλεισίματος, αλλά ως πρόβλεψη της μελλοντικής λογαριθμικής απόδοσης για χρονικούς ορίζοντες 1, 15 και 30 λεπτών. Παράλληλα, εξετάζεται η ικανότητα των μοντέλων να προβλέπουν ορθά την κατεύθυνση της μεταβολής της τιμής.

Η προσέγγιση βασίζεται στη μετατροπή των πρωτογενών δεδομένων αγοράς σε ένα σύνολο επεξεργασμένων και εμπλουτισμένων χαρακτηριστικών που αποτυπώνουν τη βραχυχρόνια συμπεριφορά της αγοράς. Σε αυτό το πλαίσιο, εξετάζονται στατιστικά μοντέλα, όπως το μοντέλο αυτοπαλίνδρομου κινητού μέσου (ARIMA), μοντέλα μηχανικής μάθησης, όπως η παλινδρόμηση Ridge (Ridge Regression), η παλινδρόμηση τυχαίων δασών (Random Forest Regression), η παλινδρόμηση XGBoost (XGBoost Regression), καθώς και μοντέλα βαθιάς μάθησης, το δίκτυο Μακράς Βραχυπρόθεσμης Μνήμης (Long Short Term Memory - LSTM) και το δίκτυο Πυλών Επαναλαμβανόμενης Μονάδας (Gated Recurrent Unit - GRU). Η αξιολόγηση πραγματοποιείται τόσο σε επίπεδο λογαριθμικών αποδόσεων όσο και σε επίπεδο προβλεπόμενης τιμής, με χρήση μετρικών όπως η ρίζα μέσου τετραγωνικού σφάλματος (RMSE), μέσου απόλυτου σφάλματος (MAE) και Ακρίβεια Κατεύθυνσης (Directional Accuracy). Με τον τρόπο αυτό εξετάζεται όχι μόνο το μέγεθος του σφάλματος πρόβλεψης, αλλά και η ικανότητα των μοντέλων να προβλέπουν σωστά την κατεύθυνση της μελλοντικής μεταβολής.

Παράλληλα, η εργασία διερευνά αν η προσθήκη μεταβλητών ανάλυσης συναισθήματος μπορεί να βελτιώσει την προβλεπτική απόδοση των μοντέλων. Για τον σκοπό αυτό συλλέχθηκαν και επεξεργάστηκαν δεδομένα κειμένου από Twitter, Reddit και ειδησεογραφικές πηγές. Μέσω του εργαλείου VADER εξήχθησαν μεταβλητές θετικού, αρνητικού, ουδέτερου και συνολικού συναισθήματος, οι οποίες ενώθηκαν χρονικά με τη χρονοσειρά του Bitcoin. Έτσι, συγκρίνονται μοντέλα που βασίζονται αποκλειστικά σε ιστορικά χαρακτηριστικά με μοντέλα που ενσωματώνουν και εξωτερική πληροφορία από τη δημόσια συζήτηση και την ειδησεογραφική ροή.

Τα βασικά ερευνητικά ερωτήματα της εργασίας αφορούν το ποια κατηγορία μοντέλων αποδίδει καλύτερα στη βραχυχρόνια πρόβλεψη της τιμής του Bitcoin, πώς μεταβάλλεται η απόδοση ανάλογα με τον χρονικό ορίζοντα και αν οι μεταβλητές συναισθήματος προσφέρουν ουσιαστική πρόσθετη αξία στη πρόβλεψη. Η συνεισφορά της παρούσας διπλωματικής έγκειται στη συγκριτική αξιολόγηση μοντέλων με και χωρίς μεταβλητές συναισθήματος σε δεδομένα υψηλής συχνότητας, καθώς και στη διερεύνηση του κατά πόσο η εξωτερική πληροφορία μπορεί να ενισχύσει την πρόβλεψη σε μία αγορά υψηλής μεταβλητότητας.

Η δομή οργανώνεται ως εξής. Στο δεύτερο κεφάλαιο παρουσιάζεται η σχετική βιβλιογραφία για την πρόβλεψη χρηματοοικονομικών χρονοσειρών, τη μηχανική και βαθιά μάθηση και την ανάλυση συναισθήματος. Στο τρίτο κεφάλαιο παρουσιάζονται τα μοντέλα και οι τεχνικές που χρησιμοποιούνται. Στο τέταρτο κεφάλαιο αναλύονται τα δεδομένα, η προεπεξεργασία, η δημιουργία της μεταβλητής στόχου και η συλλογή δεδομένων συναισθήματος. Στο πέμπτο κεφάλαιο παρουσιάζονται τα αποτελέσματα των μοντέλων με και χωρίς μεταβλητές συναισθήματος. Τέλος, στο έκτο κεφάλαιο διατυπώνονται τα βασικά συμπεράσματα και στο έβδομο παρουσιάζονται προτάσεις για μελλοντική έρευνα.

2. Βιβλιογραφική ανασκόπηση

Τα κρυπτονομίσματα αποτελούν μία ιδιαίτερη κατηγορία χρηματοοικονομικών περιουσιακών στοιχείων, καθώς χαρακτηρίζονται από υψηλή μεταβλητότητα, αποκεντρωμένη λειτουργία, συνεχή διαπραγμάτευση και έντονη ευαισθησία σε εξωτερικά γεγονότα. Σε αντίθεση με παραδοσιακές αγορές, η τιμή τους δεν επηρεάζεται μόνο από ιστορικά μοτίβα και τεχνικούς δείκτες, αλλά και από παράγοντες όπως οι ειδήσεις, οι ρυθμιστικές παρεμβάσεις, το επενδυτικό συναίσθημα, οι μακροοικονομικές συνθήκες και οι μεταβολές στη γενικότερη χρηματοοικονομική αβεβαιότητα. [1]

Για τον λόγο αυτό, η πρόβλεψη της τιμής του Bitcoin αποτελεί σύνθετο πρόβλημα, στο οποίο οι γραμμικές σχέσεις και οι απλές στατιστικές υποθέσεις συχνά δεν επαρκούν.

2.1 Πρόβλεψη χρηματοοικονομικών χρονοσειρών και κρυπτονομισμάτων

Οι πρώτες προσεγγίσεις στην πρόβλεψη χρηματοοικονομικών χρονοσειρών βασίστηκαν κυρίως, σε παραδοσιακά στατιστικά και οικονομετρικά μοντέλα, όπως το αυτοπαλίνδρομο μοντέλο κινητού μέσου (ARIMA), το γενικευμένο αυτοπαλίνδρομο υπόδειγμα υπό συνθήκη ετεροσκεδαστικότητας (GARCH) και παραλλαγές τους. Τα μοντέλα αυτά είναι χρήσιμα για τη μοντελοποίηση τάσεων, αυτοσυσχέτισης και μεταβλητότητας, ειδικά όταν τα δεδομένα παρουσιάζουν σχετικά σταθερή και γραμμική συμπεριφορά. Ωστόσο, σε αγορές όπως αυτή των κρυπτονομισμάτων, όπου οι τιμές παρουσιάζουν απότομες μεταβολές, μη γραμμικές σχέσεις και συχνές περιόδους ακραίας μεταβλητότητας, οι παραδοσιακές μέθοδοι συχνά εμφανίζουν περιορισμένη προβλεπτική ικανότητα.

Η μελέτη των Sunki, SatyaKumar, Surya Narayana, Koppera, & Hakeem [2] εξετάζει μεθόδους όπως ARIMA, δίκτυο μακράς βραχυπρόθεσμης μνήμης (LSTM) και το μοντέλο Facebook Prophet, δείχνοντας ότι το ARIMA είναι πιο κατάλληλο για σταθερές και γραμμικές χρονοσειρές, ενώ το LSTM αποδίδει καλύτερα όταν υπάρχουν σύνθετες χρονικές εξαρτήσεις και μη γραμμικά μοτίβα. Αντίστοιχα, οι Panchal, Ferdouse, & Sultana [3] συγκρίνουν ARIMA και LSTM σε χρηματοοικονομικά δεδομένα μετοχών. Το ARIMA αξιοποιείται ως κλασικό στατιστικό μοντέλο χρονοσειρών, ενώ το LSTM ως μοντέλο βαθιάς μάθησης ικανό να αποτυπώνει σύνθετες χρονικές εξαρτήσεις και μη γραμμικά μοτίβα. Καταλήγουν δε στο ότι το LSTM παρουσιάζει χαμηλότερα σφάλματα

πρόβλεψης, γεγονός που ενισχύει την άποψη ότι τα μοντέλα βαθιάς μάθησης μπορούν να είναι καταλληλότερα για αγορές με έντονη μη γραμμικότητα και μεταβλητότητα, όπως το Bitcoin. Παράλληλα, οι Tripathy, Hota, Singh, Acharya, & Nayak [4] εστιάζουν στη μεταβλητότητα του Bitcoin και δείχνουν ότι πιο σύνθετα οικονομετρικά μοντέλα, όπως το μοντέλο TARCH με επαναδειγματοληψία Bootstrap (Bootstrap TARCH), μπορούν να αποτυπώσουν χαρακτηριστικά όπως ο κίνδυνος ακραίων ζημιών (tail risk), οι επιδράσεις μόχλευσης (leverage effects) και η συμπεριφορά μακράς μνήμης (long-memory behavior).

Η μετάβαση από τα παραδοσιακά στατιστικά μοντέλα σε τεχνικές μηχανικής μάθησης οφείλεται κυρίως στην ανάγκη αποτύπωσης πιο σύνθετων σχέσεων στα δεδομένα. Οι Saberironaghi, Ren, & Saberironaghi [5] παρουσιάζουν μια εκτενή βιβλιογραφική ανασκόπηση σχετικά με την πρόβλεψη χρηματιστηριακών αγορών μέσω τεχνικών μηχανικής μάθησης και βαθιάς μάθησης. Η μελέτη ταξινομεί τις βασικές προσεγγίσεις πρόβλεψης σε τεχνική ανάλυση, θεμελιώδη ανάλυση, ανάλυση συναισθήματος και υβριδικές μεθόδους, ενώ παράλληλα εξετάζει παραδοσιακές στατιστικές τεχνικές, αλγορίθμους μηχανικής μάθησης και νευρωνικά δίκτυα. Επισημαίνουν ότι μοντέλα όπως μηχανή διανυσμάτων υποστήριξης (SVM), Random Forest, συνελκτικό νευρωνικό δίκτυο (Convolutional Neural Network – CNN), αναδρομικό νευρωνικό δίκτυο (Recurrent Neural Network – RNN), LSTM και μονάδα ελεγχόμενης επαναφοράς (Gated Recurrent Unit – GRU) έχουν χρησιμοποιηθεί ευρέως, καθώς μπορούν να αναγνωρίσουν μη γραμμικά μοτίβα και χρονικές εξαρτήσεις που δεν αποτυπώνονται εύκολα από κλασικά στατιστικά υποδείγματα. Στο ευρύτερο πεδίο της πρόβλεψης της κατεύθυνσης ή της ακριβούς τιμής των κρυπτονομισμάτων μια συγκεκριμένη χρονική στιγμή, αρκετές μελέτες έχουν εξετάσει βασικά μοντέλα μηχανικής μάθησης για την πρόβλεψη της κατεύθυνσης των τιμών.

Το άρθρο των Islam, et al. [6] εξετάζει την εφαρμογή Logistic Regression, Random Forest και XGBoost για την ταξινόμηση ανοδικών και καθοδικών κινήσεων σε μεγάλα κρυπτονομίσματα, χρησιμοποιώντας ιστορικά δεδομένα, όγκους συναλλαγών και τεχνικούς δείκτες. Παρότι τα αποτελέσματα είναι σχετικά περιορισμένα, η μελέτη είναι χρήσιμη καθώς δείχνει ότι οι απλές τεχνικές επιβλεπόμενης μάθησης μπορούν να εφαρμοστούν στη πρόβλεψη κρυπτονομισμάτων, αλλά δεν επαρκούν πάντα για την αποτύπωση της πολυπλοκότητας της αγοράς. Επίσης, οι Akyildirim, Goncu, & Sensoy [7] εξετάζουν δώδεκα πολύ ρευστά κρυπτονομίσματα σε ημερήσια και ενδοημερήσια συχνότητα και εφαρμόζει SVM, Logistic Regression, τεχνητό νευρωνικό δίκτυο (Artificial Neural Network – ANN) και Random Forest classification για την πρόβλεψη της κατεύθυνσης των αποδόσεων. Τα αποτελέσματα δείχνουν predictive accuracies περίπου 55–65%, με το SVM να παρουσιάζει τη μεγαλύτερη σταθερότητα.

2.2 Βαθιά μάθηση και προηγμένες αρχιτεκτονικές στη πρόβλεψη bitcoin και κρυπτονομισμάτων

Η χρήση μοντέλων βαθιάς μάθησης έχει αποκτήσει ιδιαίτερη σημασία στην πρόβλεψη τιμών κρυπτονομισμάτων, καθώς τα νευρωνικά δίκτυα μπορούν να διαχειριστούν ακολουθιακά δεδομένα και να εντοπίσουν πολύπλοκες χρονικές εξαρτήσεις. Αρχιτεκτονικές όπως LSTM, GRU, υβριδικά μοντέλα συνελκτικών νευρωνικών δικτύων και δικτύων μακράς βραχυπρόθεσμης μνήμης (CNN-LSTM), υβριδικά μοντέλα συνελκτικών νευρωνικών δικτύων και μονάδων ελεγχόμενης επαναφοράς (CNN-GRU), καθώς και μοντέλα βασισμένα σε αρχιτεκτονικές Μετασχηματιστών (Transformer-based models) έχουν εφαρμοστεί σε προβλήματα πρόβλεψης τιμών, αποδόσεων και κατεύθυνσης της αγοράς. Τα LSTM και GRU είναι ιδιαίτερα χρήσιμα για χρονοσειρές, επειδή μπορούν να διατηρούν πληροφορία από προηγούμενες χρονικές στιγμές, ενώ τα CNN μπορούν να εξάγουν τοπικά μοτίβα από ακολουθιακά δεδομένα.

Οι Dubey & Enke [8] εξετάζουν την πρόβλεψη της κατεύθυνσης της τιμής του Bitcoin με χρήση δεδομένων εντός αλυσίδας (on-chain data) και τεχνικών επιλογής χαρακτηριστικών. Η μελέτη αξιοποιεί 196 χαρακτηριστικά από το blockchain του Bitcoin, και εφαρμόζει μεθόδους όπως L1 regression, Boruta και ανάλυση κύριων συνιστωσών (Principal Component Analysis – PCA) για τη μείωση της διαστατικότητας. Στη συνέχεια συγκρίνει Random Forest, TCN και CNN-LSTM, με τον συνδυασμό Boruta για επιλογή μεταβλητών και CNN-LSTM να επιτυγχάνει την καλύτερη απόδοση. Η μελέτη είναι σημαντική διότι δείχνει ότι πέρα από τις ιστορικές τιμές, α χαρακτηριστικά εντός αλυσίδας (on-chain features), όπως η πραγματοποιηθείσα αξία (realized value) και η μη πραγματοποιηθείσα αξία (unrealized value), μπορούν να προσφέρουν ουσιαστική προβλεπτική πληροφορία. Επιπλέον, οι Islam, et al. [9] εξετάζουν την πρόβλεψη της τιμής του XRP/USDT, ενσωματώνοντας εκτός από ιστορικά δεδομένα και δύο δείκτες ρευστότητας, το VVR και τη σταθμισμένη ως προς τον όγκο μέση τιμή (Volume-Weighted Average Price – VWAP). Η μελέτη συγκρίνει τη γραμμική παλινδρόμηση (Linear Regression), το τυχαίο δάσος (Random Forest), την ακραία ενίσχυση κλίσης (Extreme Gradient Boosting – XGBoost) και το δίκτυο μακράς βραχυπρόθεσμης μνήμης (Long Short-Term Memory – LSTM), καταλήγοντας ότι το LSTM παρουσιάζει την καλύτερη προβλεπτική απόδοση τόσο χωρίς όσο και με τις μεταβλητές ρευστότητας. Το εύρημα αυτό είναι χρήσιμο για τη βιβλιογραφία, καθώς δείχνει ότι η ρευστότητα μπορεί να αποτελέσει σημαντική πρόσθετη πληροφορία για την πρόβλεψη τιμών κρυπτονομισμάτων και ότι τα αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Networks – RNN) μπορούν να αξιοποιήσουν αποτελεσματικά τέτοιου τύπου χαρακτηριστικά.

Αντίστοιχα οι Li, Liu, Liu, Zhu, & Yan [10] εξετάζουν την προβλεψιμότητα των αποδόσεων κρυπτονομισμάτων μέσω ενός πλαισίου υψηλής διαστατικότητας, το οποίο συνδυάζει μακροοικονομικούς και ειδικούς για τα κρυπτονομίσματα παράγοντες. Η μελέτη αξιοποιεί δεδομένα από το CoinMarketCap και άλλες πηγές και κατασκευάζει ένα σύνολο μεταβλητών που περιλαμβάνει δείκτες δικτύου, όπως new addresses και active addresses, χρηματοοικονομικούς δείκτες, όπως το market-to-realized-value ratio, δείκτες ροών ανταλλακτηρίων, Google Trends, καθώς και μακροοικονομικούς παράγοντες. Για την πρόβλεψη των εβδομαδιαίων αποδόσεων εφαρμόζονται διαφορετικά μοντέλα, η μέθοδος των ελαχίστων τετραγώνων (Ordinary Least Squares – OLS), η παλινδρόμηση ελάχιστης απόλυτης συρρίκνωσης και επιλογής (Least Absolute Shrinkage and Selection Operator – LASSO), η παλινδρόμηση Ridge (Ridge Regression), το τυχαίο δάσος (Random Forest), τα δέντρα παλινδρόμησης ενίσχυσης κλίσης (Gradient Boosting Regression Trees – GBRT), ο αλγόριθμος των k πλησιέστερων γειτόνων (k-Nearest Neighbors – KNN) και τα νευρωνικά δίκτυα. Τα αποτελέσματα δείχνουν ότι μη γραμμικά μοντέλα, κυρίως το Random Forest, παρουσιάζουν καλύτερη εκτός δείγματος απόδοση σε σχέση με τα γραμμικά μοντέλα. Επιπλέον, μέσω της ανάλυσης SHAP, η μελέτη αναδεικνύει ότι χαρακτηριστικά όπως τα active addresses, τα new addresses και το market-to-realized-value ratio είναι από τους σημαντικότερους παράγοντες πρόβλεψης.

Πιο σύνθετες αρχιτεκτονικές έχουν επίσης χρησιμοποιηθεί για την καλύτερη αποτύπωση βραχυπρόθεσμων και μακροπρόθεσμων μοτίβων στη χρονοσειρά των κρυπτονομισμάτων. Οι Azizi, et al. [11] προτείνουν ένα μοντέλο πρόβλεψης τιμών κρυπτονομισμάτων που συνδυάζει νευρωνικά δίκτυα με δυναμικά συστήματα, βασισμένο στη μέθοδο 2DSL. Η προσέγγιση αυτή αντιμετωπίζει τη χρονοσειρά τιμών ως σήμα και την αποσυνθέτει σε επιμέρους συνιστώσες μέσω της εμπειρικής αποσύνθεσης τρόπων (Empirical Mode Decomposition – EMD), ώστε να διαχωριστούν οι υψηλής συχνότητας διακυμάνσεις από τις πιο σταθερές μακροχρόνιες τάσεις. Η μέθοδος εφαρμόζεται σε ιστορικά δεδομένα Bitcoin και συγκρίνεται με κλασικές αρχιτεκτονικές βαθιάς μάθησης, όπως LSTM και GRU. Τα αποτελέσματα δείχνουν ότι η αποσύνθεση της χρονοσειράς μπορεί να βελτιώσει την ακρίβεια και τη σταθερότητα της πρόβλεψης σε σχέση με κλασικές αρχιτεκτονικές LSTM και GRU.

Παράλληλα, οι Izadi & Hajizadeh [12] με το paper “Time Series Prediction for Cryptocurrency Markets with Transformer and Parallel Convolutional Neural Networks” προτείνουν ένα μοντέλο που συνδυάζει δομικά τμήματα Μετασχηματιστών (Transformer blocks) και παράλληλα συνελκτικά νευρωνικά δίκτυα (parallel Convolutional Neural Networks – CNNs) για την πρόβλεψη της μελλοντικής τάσης χρονοσειρών κρυπτονομισμάτων. Η μελέτη εισάγει μια νέα μέθοδο

διανυσματοποίησης για τη μετατροπή των ακατέργαστων δεδομένων σε κατάλληλη είσοδο για transformer encoders. Αν και τα αποτελέσματα στην πρόβλεψη κατεύθυνσης του Bitcoin παραμένουν μέτρια, η μελέτη δείχνει ότι οι πιο πρόσφατες αρχιτεκτονικές βαθιάς μάθησης μπορούν να προσφέρουν εναλλακτικές λύσεις σε σχέση με τα παραδοσιακά αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Networks – RNNs).

2.3 Sentiment analysis, ειδήσεις και εξωτερική πληροφορία στη πρόβλεψη κρυπτονομισμάτων

Ένα σημαντικό τμήμα της πρόσφατης βιβλιογραφίας υποστηρίζει ότι η πρόβλεψη τιμών κρυπτονομισμάτων δεν πρέπει να βασίζεται αποκλειστικά σε ιστορικές τιμές και τεχνικούς δείκτες. Η αγορά των κρυπτονομισμάτων είναι ιδιαίτερα ευαίσθητη σε ειδήσεις, ανακοινώσεις, δημοσιεύσεις στα κοινωνικά δίκτυα και γενικότερα στο επενδυτικό συναίσθημα. Για τον λόγο αυτό, η ανάλυση συναισθήματος έχει ενσωματωθεί σε αρκετές μελέτες ως πρόσθετη πηγή πληροφορίας, με σκοπό τη βελτίωση της προβλεπτικής απόδοσης των μοντέλων.

Οι Zubair, et al. [13], προτείνουν ένα βελτιωμένο πλαίσιο πρόβλεψης τιμών κρυπτονομισμάτων, το οποίο συνδυάζει τεχνική ανάλυση ιστορικών τιμών με ανάλυση συναισθήματος. Εστιάζουν στην υψηλή μεταβλητότητα της αγοράς κρυπτονομισμάτων και υποστηρίζουν ότι η πρόβλεψη των τιμών δεν πρέπει να βασίζεται αποκλειστικά σε ιστορικά δεδομένα, καθώς εξωτερικοί παράγοντες, όπως ειδήσεις, κοινωνικά δίκτυα και επενδυτικό κλίμα, μπορούν να επηρεάσουν σημαντικά την αξιοπιστία των προβλέψεων. Για την πρόβλεψη των τιμών, αναπτύσσουν ένα υβριδικό μοντέλο που συνδυάζει αμφίδρομα δίκτυα μακράς βραχυπρόθεσμης μνήμης (Bidirectional Long Short-Term Memory – Bi-LSTM) και μονάδες ελεγχόμενης επαναφοράς (Gated Recurrent Units – GRU), το οποίο αξιοποιεί τη δυνατότητα των αναδρομικών νευρωνικών δικτύων να εντοπίζουν χρονικές εξαρτήσεις σε δεδομένα χρονοσειρών. Παράλληλα, ενσωματώνουν ένα υβριδικό μοντέλο ανάλυσης συναισθήματος που βασίζεται στο μοντέλο αμφίδρομων αναπαραστάσεων κωδικοποιητών από Μετασχηματιστές (Bidirectional Encoder Representations from Transformers – BERT) και στο εργαλείο λεξικογραφικής ανάλυσης συναισθήματος VADER (Valence Aware Dictionary and sEntiment Reasoner), με στόχο την αξιολόγηση του αν οι προβλεπόμενες τιμές μπορούν να θεωρηθούν αξιόπιστες ή αν πρέπει να αντιμετωπιστούν με επιφύλαξη λόγω αρνητικού κλίματος στην αγορά. Η μελέτη εστιάζει σε Bitcoin, Ethereum και Dogecoin και δείχνει ότι η ενσωμάτωση μεταβλητών συναισθήματος μπορεί να βελτιώσει την αξιοπιστία των προβλέψεων.

Οι Hamadou, et al. [14] εξετάζουν την πρόβλεψη των αποδόσεων του Bitcoin με χρήση τεχνικών μηχανικής και βαθιάς μάθησης, δίνοντας ιδιαίτερη έμφαση στον ρόλο του επενδυτικού συναισθήματος. Η μελέτη κατασκευάζει έναν νέο δείκτη sentiment βασισμένο σε δεδομένα Google Trends και σε όρους σχετικούς με τα κρυπτονομίσματα, με στόχο να αποτυπωθεί η ψυχολογία των επενδυτών στην αγορά Bitcoin. Παράλληλα, λαμβάνουν υπόψη δείκτες χρηματοοικονομικής αβεβαιότητας και παραδοσιακές μεταβλητές αγοράς, όπως ο S&P 500, ο χρυσός και το πετρέλαιο Brent. Για την πρόβλεψη συγκρίνονται το μονοδιάστατο συνελκτικό νευρωνικό δίκτυο (One-Dimensional Convolutional Neural Network – 1D-CNN), το αμφίδρομο δίκτυο μακράς βραχυπρόθεσμης μνήμης (Bidirectional Long Short-Term Memory – BLSTM) και η παλινδρόμηση διανυσμάτων υποστήριξης (Support Vector Regression – SVR), με το 1D-CNN να παρουσιάζει την καλύτερη προβλεπτική απόδοση. Ένα βασικό εύρημα της μελέτης είναι ότι το επενδυτικό συναίσθημα αποτελεί έναν από τους σημαντικότερους παράγοντες πρόβλεψης των αποδόσεων του Bitcoin, ιδιαίτερα κατά την περίοδο της πανδημίας COVID-19.

Ακόμη πιο εξειδικευμένη προσέγγιση ακολουθούν οι Gür & Ünal [15] οι οποίοι εξετάζουν την επίδραση του ειδησεογραφικού συναισθήματος στην πρόβλεψη της τιμής του Bitcoin, αξιοποιώντας μοντέλα μηχανικής και βαθιάς μάθησης σε συνδυασμό με τεχνικές επεξεργασίας φυσικής γλώσσας. Η μελέτη ενσωματώνει FinBERT βαθμολογίες συναισθήματος και σημασιολογικά χαρακτηριστικά από περισσότερους από 650.000 οικονομικούς τίτλους ειδήσεων, μαζί με υστερήσεις της τιμής του Bitcoin, χρηματοοικονομικούς δείκτες, μακροοικονομικές μεταβλητές και χρονικά χαρακτηριστικά. Για την πρόβλεψη χρησιμοποιούν μοντέλα όπως XGBoost, Random Forest, LightGBM, SVR, LSTM, GRU και CNN, τόσο με όσο και χωρίς μεταβλητές συναισθήματος. Τα αποτελέσματα δείχνουν ότι η προσθήκη ειδησεογραφικού συναισθήματος βελτιώνει την προβλεπτική ακρίβεια, με το FinBERT-LSTM να επιτυγχάνει την καλύτερη απόδοση. Επιπλέον, η μελέτη αναδεικνύει ότι τα αρνητικά νέα έχουν ισχυρότερη επίδραση στην τιμή και τη μεταβλητότητα του Bitcoin από ό,τι τα θετικά νέα, στοιχείο που υποδηλώνει ασύμμετρη σχέση μεταξύ ειδήσεων και αγοράς.

Οι Tabe, Mansourabady, Rasekh, & Tanoori [16] προτείνουν ένα υβριδικό μοντέλο CNN-GRU για την πρόβλεψη τιμών Bitcoin, Ethereum και Monero, συνδυάζοντας ιστορικά δεδομένα από το CoinMarketCap και ειδησεογραφικά δεδομένα από το Google News. Για την εξαγωγή του συναισθήματος εφαρμόζονται τρία διαφορετικά εργαλεία, VADER, TextBlob και Flair, τα οποία συγκρίνονται ως προς τη συμβολή τους στην ακρίβεια πρόβλεψης. Η αρχιτεκτονική του προτεινόμενου μοντέλου περιλαμβάνει συνελκτικά επίπεδα για την εξαγωγή τοπικών χαρακτηριστικών και GRU επίπεδα για την αποτύπωση χρονικών εξαρτήσεων. Τα αποτελέσματα

δείχνουν ότι ο συνδυασμός ιστορικών τιμών και ειδησεογραφικού sentiment βελτιώνει την προβλεπτική απόδοση σε σχέση με τη χρήση μόνο ιστορικών δεδομένων, ενώ το TextBlob παρουσιάζει συνολικά την καλύτερη συνεισφορά σε αρκετές περιπτώσεις.

Η σημασία των ειδήσεων και των ανακοινώσεων επιβεβαιώνεται και από μελέτες που δεν εστιάζουν αποκλειστικά στην πρόβλεψη με τη χρήση βαθιάς μάθησης, αλλά στην επίδραση γεγονότων στην αγορά. Ο Sidoron [17] εξετάζει την επίδραση δημόσιων ανακοινώσεων, ρυθμιστικών παρεμβάσεων και σημαντικών γεγονότων στις τιμές των Bitcoin, Ethereum, Ripple και Dogecoin. Η μελέτη χρησιμοποιεί ημερήσια δεδομένα τιμών για την περίοδο 2017–2024 και αναλύει γεγονότα όπως tweets επιδραστικών προσώπων, αποφάσεις ρυθμιστικών αρχών, το FTX collapse και εξωγενείς κρίσεις όπως η πανδημία COVID-19 και η ρωσοουκρανική σύγκρουση. Τα αποτελέσματα δείχνουν ότι τα πιο ασταθή κρυπτονομίσματα, όπως το Dogecoin και το Ripple, αντιδρούν εντονότερα σε θετικές ανακοινώσεις, ενώ τα πιο ώριμα και σταθερά κρυπτονομίσματα, όπως το Bitcoin και το Ethereum, παρουσιάζουν πιο περιορισμένες και βραχυπρόθεσμες αντιδράσεις. Παράλληλα, τα αρνητικά γεγονότα, ιδίως όσα σχετίζονται με ρυθμιστικές παρεμβάσεις ή διαταραχές της αγοράς, έχουν πιο παρατεταμένη και επιζήμια επίδραση στις τιμές. Αντίστοιχα, οι Almaqableh, Khasawneh, & Sahiner [18] εξετάζουν την επίδραση περιβαλλοντικών και κλιματικών γεγονότων στη συμπεριφορά της αγοράς Bitcoin, και υποστηρίζουν ότι τέτοιες πληροφοριακές διαταραχές μπορούν να επηρεάσουν τις αποδόσεις και τη μεταβλητότητα του Bitcoin, ιδιαίτερα όταν σχετίζονται με πολιτικές αποφάσεις ή ακραία κλιματικά γεγονότα.

2.4 Μακροοικονομική, συστημικοί και εξωτερικοί παράγοντες

Πέρα από την ανάλυση συναισθήματος, αρκετές μελέτες δείχνουν ότι η αγορά κρυπτονομισμάτων επηρεάζεται και από ευρύτερους μακροοικονομικούς και συστημικούς παράγοντες. Το Bitcoin, παρότι αρχικά θεωρήθηκε από πολλούς ως ένα αποκεντρωμένο περιουσιακό στοιχείο ανεξάρτητο από τις παραδοσιακές αγορές, φαίνεται πλέον να συνδέεται σε σημαντικό βαθμό με το ευρύτερο χρηματοοικονομικό περιβάλλον.

Η μελέτη των Mahfooz & Phillips [19] εξετάζει αν εξωγενείς μεταβλητές, όπως τα επιτόκια και η πιθανότητα ύφεσης, βελτιώνουν την πρόβλεψη της τιμής του Bitcoin. Χρησιμοποιώντας LSTM και Facebook Prophet και καταλήγει ότι το LSTM αποδίδει καλύτερα, ενώ η ενσωμάτωση των δύο εξωγενών μεταβλητών ενισχύει σημαντικά την προβλεπτική ικανότητα των μοντέλων. Η μελέτη αυτή υποστηρίζει ότι η αποκλειστική χρήση ιστορικών τιμών δεν είναι πάντα επαρκής και ότι η

πρόβλεψη του Bitcoin μπορεί να βελτιωθεί όταν λαμβάνονται υπόψη εξωτερικές οικονομικές μεταβλητές.

Οι Mgadmi, Sadraoui, & Abidi [20] μελετούν τη σχέση μεταξύ χρηματιστηριακών δεικτών και κρυπτονομισμάτων πριν και κατά τη διάρκεια του πολέμου Ρωσίας–Ουκρανίας. Με τη χρήση του ελέγχου αιτιότητας κατά Granger (Granger causality test) και μοντέλων γενικευμένης αυτοπαλίνδρομης υπό συνθήκη ετεροσκεδαστικότητας (Generalized Autoregressive Conditional Heteroskedasticity – GARCH), καθώς και κλασματικά ολοκληρωμένων μοντέλων GARCH (Fractionally Integrated GARCH – FIGARCH/FGARCH), δείχνουν ότι οι σχέσεις μεταξύ παραδοσιακών και crypto αγορών μεταβάλλονται έντονα σε περιόδους γεωπολιτικής αστάθειας. Αντίστοιχα, οι Gaies, Chaâbane, & Bouzouita [21] εξετάζουν τη σχέση ανάμεσα στη χρηματοοικονομική αστάθεια και τη μεταβλητότητα της crypto αγοράς από μακροοικονομική σκοπιά, χρησιμοποιώντας δείκτες όπως το Cryptocurrency VIX και μεθόδους ανάλυσης χρόνου–συχνότητας. Τα αποτελέσματα δείχνουν ότι η αστάθεια του παραδοσιακού χρηματοπιστωτικού συστήματος συνδέεται με τη μεταβλητότητα της αγοράς κρυπτονομισμάτων, ιδιαίτερα σε ακραίες συνθήκες και σε πιο μακροχρόνιους ορίζοντες.

Οι μελέτες αυτές δείχνουν ότι το Bitcoin δεν λειτουργεί πλήρως απομονωμένα από το υπόλοιπο χρηματοοικονομικό σύστημα. Αντίθετα, η τιμή και η μεταβλητότητά του μπορούν να επηρεαστούν από επιτόκια, δείκτες αβεβαιότητας, χρηματιστηριακές αγορές, γεωπολιτικές κρίσεις και περιβαλλοντικά γεγονότα. Επομένως, η ενσωμάτωση εξωτερικών μεταβλητών στα μοντέλα πρόβλεψης μπορεί να προσφέρει πιο ολοκληρωμένη εικόνα της αγοράς.

2.5 Συμπέρασμα βιβλιογραφικής ανασκόπησης και ερευνητικό κενό

Η διαθέσιμη βιβλιογραφία δείχνει πως η πρόβλεψη της τιμής του Bitcoin και των κρυπτονομισμάτων γενικότερα έχει εξελιχθεί σημαντικά τα τελευταία χρόνια. Οι παραδοσιακές στατιστικές μέθοδοι, όπως ARIMA και GARCH, παραμένουν χρήσιμες για την ανάλυση τάσης και μεταβλητότητας, αλλά συχνά δεν επαρκούν για την αποτύπωση των μη γραμμικών σχέσεων που χαρακτηρίζουν την αγορά. Τα μοντέλα μηχανικής μάθησης, όπως Random Forest, SVM και XGBoost, προσφέρουν μεγαλύτερη ευελιξία, ενώ τα μοντέλα βαθιάς μάθησης, όπως LSTM, GRU, CNN-LSTM και Transformer-based αρχιτεκτονικές, έχουν τη δυνατότητα να αξιοποιούν πιο αποτελεσματικά τη χρονική δομή των δεδομένων.

Παράλληλα, η πρόσφατη βιβλιογραφία αναδεικνύει τη σημασία της εξωτερικής πληροφορίας. Ειδήσεις, tweets, Google Trends, ρυθμιστικές ανακοινώσεις, μακροοικονομικοί δείκτες και περιβαλλοντικά γεγονότα μπορούν να επηρεάσουν τις προσδοκίες των επενδυτών και, κατά συνέπεια, την τιμή και τη μεταβλητότητα του Bitcoin. Ιδιαίτερα, η ανάλυση συναισθήματος φαίνεται να συμβάλλει σημαντικά στη προβλεπτική ικανότητα, καθώς επιτρέπει στα μοντέλα να ενσωματώνουν πληροφορία που δεν περιλαμβάνεται άμεσα στις ιστορικές τιμές.

Με βάση τα παραπάνω, η παρούσα διπλωματική εργασία τοποθετείται στο σημείο τομής μεταξύ χρηματοοικονομικής πρόβλεψης, μηχανικής/βαθιάς μάθησης και ανάλυσης συναισθήματος. Στόχος της είναι να εξετάσει την απόδοση διαφορετικών μοντέλων πρόβλεψης της τιμής του Bitcoin σε εναλλακτικούς χρονικούς ορίζοντες, συγκρίνοντας προσεγγίσεις που βασίζονται αποκλειστικά σε ιστορικά και τεχνικά δεδομένα με προσεγγίσεις που ενσωματώνουν εξωτερική πληροφορία συναισθήματος. Με αυτόν τον τρόπο, επιχειρείται να αξιολογηθεί αν και σε ποιο βαθμό η ανάλυση συναισθήματος μπορεί να βελτιώσει την προβλεπτική ικανότητα των μοντέλων σε μία αγορά υψηλής μεταβλητότητας και αβεβαιότητας όπως το Bitcoin.

3. Μεθοδολογία και μοντέλα πρόβλεψης

Στο παρόν κεφάλαιο παρουσιάζονται τα μοντέλα που παρουσιάστηκαν προηγουμένως και χρησιμοποιήθηκαν για την πρόβλεψη της μελλοντικής λογαριθμικής απόδοσης του Bitcoin. Συγκεκριμένα, τα στατιστικά μοντέλα, όπως το μοντέλο μηδενικής απόδοσης (Naive Zero Return) και το αυτοπαλίνδρομο κινητού μέσου (ARIMA) με στόχο τη δημιουργία ενός βασικού επιπέδου σύγκρισης. Στη συνέχεια, εφαρμόστηκαν μοντέλα μηχανικής μάθησης, όπως η παλινδρόμηση Ridge (Ridge Regression), η παλινδρόμηση τυχαίου δάσους (Random Forest Regression) και η παλινδρόμηση XGBoost (XGBoost Regression). Τέλος, εφαρμόστηκαν επαναλαμβανόμενα μοντέλα βαθιάς μάθησης (recurrent deep learning models), και συγκεκριμένα το δίκτυο μακράς βραχυπρόθεσμης μνήμης (Long Short-Term Memory – LSTM) και η επαναλαμβανόμενη μονάδα με πύλες (Gated Recurrent Unit – GRU), ώστε να εξεταστεί αν η ακολουθιακή δομή των δεδομένων μπορεί να βελτιώσει την πρόβλεψη.

Για την υλοποίηση χρησιμοποιήθηκε το περιβάλλον Google Colab, καθώς η ανάλυση βασίστηκε σε μεγάλο όγκο δεδομένων υψηλής συχνότητας και απαιτούσε την εκτέλεση πολλών πειραμάτων για διαφορετικούς χρονικούς ορίζοντες πρόβλεψης και συνδυασμούς υπερπαραμέτρων. Η χρήση του Google Colab διευκόλυνε την πειραματική διαδικασία, παρέχοντας πρόσβαση σε υπολογιστικούς πόρους και, όπου ήταν απαραίτητο, δυνατότητα αξιοποίησης GPU για την εκπαίδευση των μοντέλων βαθιάς μάθησης.

Τα μοντέλα υλοποιήθηκαν σε γλώσσα Python, αξιοποιώντας βιβλιοθήκες για ανάλυση δεδομένων, στατιστική επεξεργασία, μηχανική μάθηση και βαθιά μάθηση. Η διαδικασία ξεκίνησε με την προεπεξεργασία των δεδομένων, όπου πραγματοποιήθηκε χρονική ταξινόμηση των παρατηρήσεων, έλεγχος και αφαίρεση διπλότυπων timestamps, μετατροπή της χρονοσειράς σε σταθερή συχνότητα ενός λεπτού και διαχείριση ελλিপών τιμών. Στη συνέχεια, δημιουργήθηκε η μεταβλητή στόχος, δηλαδή η μελλοντική λογαριθμική απόδοση για κάθε χρονικό ορίζοντα πρόβλεψης, καθώς και τα πρόσθετα χαρακτηριστικά που χρησιμοποιήθηκαν ως είσοδοι στα μοντέλα.

Η επιλογή των υπερπαραμέτρων πραγματοποιήθηκε μέσω Αναζήτηση Πλέγματος Υπερπαραμέτρων (Grid Search of Hyperparameters) στο σύνολο επικύρωσης (validation set), ενώ η τελική αξιολόγηση των μοντέλων έγινε στο ανεξάρτητο σύνολο ελέγχου (test set). Η ίδια διαδικασία εφαρμόστηκε ξεχωριστά για κάθε χρονικό ορίζοντα, ώστε να διασφαλιστεί ότι τα

αποτελέσματα είναι συγκρίσιμα και ότι κάθε μοντέλο αξιολογείται στο αντίστοιχο βραχυχρόνιο πρόβλημα πρόβλεψης.

3.1 Μοντέλα πρόβλεψης

Για την πρόβλεψη της μελλοντικής λογαριθμικής απόδοσης του Bitcoin σε διαφορετικούς χρονικούς ορίζοντες εφαρμόστηκε σύνολο στατιστικών μοντέλων, μοντέλων μηχανικής μάθησης και μοντέλων βαθιάς μάθησης. Η πρόβλεψη πραγματοποιήθηκε αρχικά σε επίπεδο λογαριθμικής απόδοσης, καθώς οι αποδόσεις αποτελούν καταλληλότερη μορφή στόχου για χρηματοοικονομικές χρονοσειρές υψηλής συχνότητας σε σχέση με τις απόλυτες τιμές. Στη συνέχεια, οι προβλεπόμενες αποδόσεις μετατράπηκαν σε προβλεπόμενες τιμές κλεισίματος μέσω της σχέσης:

$$\text{Price}_{t+h} = \text{Price}_t \cdot e^{\hat{r}_{t+h}}$$

Εξίσωση 1. Μετατροπή προβλεπόμενης λογαριθμικής απόδοσης σε προβλεπόμενη τιμή κλεισίματος

όπου Price_t είναι η τιμή κλεισίματος τη χρονική στιγμή t και $\hat{r}_{t+h}$ η προβλεπόμενη λογαριθμική απόδοση για χρονικό ορίζοντα h . Όλα τα μοντέλα εκπαιδεύτηκαν και αξιολογήθηκαν ξεχωριστά για κάθε χρονικό ορίζοντα πρόβλεψης. Η επιλογή υπερπαραμέτρων έγινε με βάση το χαμηλότερο μέσο τετραγωνικό σφάλμα ρίζας (RMSE) στο σύνολο επικύρωσης (validation set), ενώ η τελική αξιολόγηση πραγματοποιήθηκε στο σύνολο ελέγχου (test set). Ως μετρικές αξιολόγησης χρησιμοποιήθηκαν το μέσο τετραγωνικό σφάλμα ρίζας και η ακρίβεια κατεύθυνσης (Directional Accuracy), τόσο σε επίπεδο αποδόσεων όσο και σε επίπεδο προβλεπόμενων τιμών κλεισίματος. Η ακρίβεια κατεύθυνσης χρησιμοποιήθηκε για να αξιολογηθεί αν το εκάστοτε μοντέλο προβλέπει σωστά την κατεύθυνση της μελλοντικής μεταβολής, δηλαδή αν η τιμή του Bitcoin αναμένεται να κινηθεί ανοδικά ή καθοδικά στον αντίστοιχο χρονικό ορίζοντα.

3.2 Στατιστικά Μοντέλα Αναφοράς

Πριν την εφαρμογή και αξιολόγηση των πιο σύνθετων μοντέλων μηχανικής και βαθιάς μάθησης, χρησιμοποιήθηκαν απλούστερα στατιστικά μοντέλα ως σημεία αναφοράς. Με αυτόν τον τρόπο μπορεί να αξιολογηθεί αν πιο σύνθετες μέθοδοι, προσφέρουν ουσιαστική βελτίωση σε σχέση με την επιπλέον πολυπλοκότητα που εισάγουν.

Ως βασικό αποτέλεσμα σύγκρισης επιλέχθηκε το μοντέλο μηδενικής απόδοσης (Naive Zero Return), σύμφωνα με το οποίο η προβλεπόμενη μελλοντική λογαριθμική απόδοση του Bitcoin

θεωρείται ίση με μηδέν. Η υπόθεση αυτή συνεπάγεται ότι η προβλεπόμενη μελλοντική τιμή είναι ίση με την τρέχουσα τιμή.

$$\hat{r}_{t+h} = 0$$

Εξίσωση 2. Πρόβλεψη του μοντέλου μηδενικής απόδοσης

Το συγκεκριμένο μοντέλο αποτελεί ένα πολύ απλό benchmark, ειδικά σε χρηματοοικονομικές χρονοσειρές υψηλής συχνότητας, όπου οι βραχυχρόνιες αποδόσεις είναι συχνά πολύ κοντά στο μηδέν. Ωστόσο, λόγω της υψηλής μεταβλητότητας του Bitcoin, η υπόθεση αυτή μπορεί να αποτύχει σε περιόδους έντονων ανοδικών ή καθοδικών κινήσεων (spikes). Η επιλογή του συγκεκριμένου μοντέλου έναντι άλλων αφελών (naive) προσεγγίσεων έγινε επειδή παρέχει πιο σταθερό και άμεσα ερμηνεύσιμο σημείο αναφοράς.

Εκτός από το αφελές μοντέλο, εφαρμόστηκαν και κλασικά στατιστικά μοντέλα χρονοσειρών. Χρησιμοποιήθηκε το μοντέλο αυτοπαλίνδρομο κινητού μέσου (ARIMA), το οποίο συνδυάζει αυτοπαλίνδρομο μέρος, διαφοροποίηση και κινητό μέσο όρο. Η γενική μορφή του συμβολίζεται ως ARIMA(p,d,q), όπου p είναι η τάξη του αυτοπαλίνδρομου (AR) μέρους, d είναι ο βαθμός διαφοροποίησης και q είναι η τάξη του κινητού μέσου (MA) μέρους. Το ARIMA εφαρμόστηκε στις λογαριθμικές αποδόσεις του Bitcoin, με στόχο να εξεταστεί αν ένας κλασικός στατιστικός μηχανισμός χρονοσειρών μπορεί να αποτυπώσει τη δυναμική των βραχυχρόνιων αποδόσεων με υψηλή μεταβλητότητα.

Για τις παραμέτρους του μοντέλου δοκιμάστηκαν διαφορετικοί συνδυασμοί μέσω αναζήτησης πλέγματος (grid search) και για κάθε χρονικό ορίζοντα επιλέχθηκε ο συνδυασμός παραμέτρων που παρουσίασε το χαμηλότερο μέσο τετραγωνικό σφάλμα ρίζας στο σύνολο ελέγχου.

3.3 Παλινδρόμηση Ridge (Ridge Regression)

Η παλινδρόμηση Ridge (Ridge Regression) [22] χρησιμοποιήθηκε ως κανονικοποιημένη γραμμική μέθοδος πρόβλεψης. Σε αντίθεση με την απλή γραμμική παλινδρόμηση, το Ridge εφαρμόζει λ_2 κανονικοποίηση στους συντελεστές, περιορίζοντας το μέγεθός τους και μειώνοντας τον κίνδυνο υπερπροσαρμογής. Η μέθοδος είναι χρήσιμη όταν υπάρχουν πολλά χαρακτηριστικά καθώς μειώνει την επίδραση της πολυσυγγραμμικότητας.

Η γραμμική πρόβλεψη δίνεται από τη σχέση:

$$\hat{y} = \beta_0 + \sum_{j=1}^p \beta_j x_j$$

Εξίσωση 3. Γραμμική συνάρτηση πρόβλεψης παλινδρόμησης Ridge

όπου $\hat{y}$ είναι η προβλεπόμενη λογαριθμική απόδοση του Bitcoin, x_j είναι τα χαρακτηριστικά εισόδου και β_j οι συντελεστές του μοντέλου. Στο Ridge, η πρόβλεψη παραμένει γραμμική, αλλά κατά την εκπαίδευση εφαρμόζεται λ_2 ποινή στους συντελεστές:

$$\lambda_2 \text{ ποινή} = \alpha \sum_{j=1}^p \beta_j^2$$

Εξίσωση 4. λ_2 ποινή κανονικοποίησης

Η παράμετρος α καθορίζει την ένταση της κανονικοποίησης. Μικρές τιμές του α οδηγούν σε συμπεριφορά κοντά στην απλή γραμμική παλινδρόμηση, ενώ μεγαλύτερες τιμές συρρικνώνουν περισσότερο τους συντελεστές προς το μηδέν, περιορίζοντας το overfitting. Σε περιπτώσεις όπου το α είναι υπερβολικά μεγάλο, το μοντέλο μπορεί να γίνει πολύ απλό και να οδηγηθεί σε underfitting.

Στην υλοποίηση χρησιμοποιήθηκε μέθοδος τυποποίησης χαρακτηριστικών (StandardScaler) στις εξωτερικές μεταβλητές του μοντέλου. Η τυποποίηση εφαρμόστηκε καθώς το Ridge επηρεάζεται από την κλίμακα των χαρακτηριστικών. Για κάθε χρονικό ορίζοντα πρόβλεψης δοκιμάστηκαν 17 τιμές του α σε λογαριθμική κλίμακα, από 10^{-4} έως 10^4 . Μετά την επιλογή της βέλτιστης τιμής, το μοντέλο εκπαιδεύτηκε στο ενιαίο σύνολο εκπαίδευσης και επικύρωσης και αξιολογήθηκε στο σύνολο ελέγχου. Επιπλέον, εξετάστηκαν οι συντελεστές του Ridge με βάση την απόλυτη τιμή τους, ώστε να εντοπιστούν τα χαρακτηριστικά με τη μεγαλύτερη επίδραση στην πρόβλεψη.

3.4 Παλινδρόμηση Τυχαίων Δασών (Random Forest Regression)

Η παλινδρόμηση τυχαίων δασών (Random Forest Regression) [23] χρησιμοποιήθηκε ως μη γραμμικό συνδυαστικό (ensemble) μοντέλο για την πρόβλεψη της μελλοντικής λογαριθμικής απόδοσης του Bitcoin. Το μοντέλο αποτελείται από πολλά δέντρα απόφασης, τα οποία εκπαιδεύονται σε διαφορετικά τυχαία υποσύνολα των δεδομένων και των χαρακτηριστικών. Η τελική πρόβλεψη προκύπτει ως ο μέσος όρος των προβλέψεων όλων των δέντρων:

$$\hat{y} = (1 / B) \sum_{b=1}^B T_b(x)$$

Εξίσωση 5. Τελική πρόβλεψη του Random Forest ως μέσος όρος των δέντρων

όπου B είναι ο αριθμός των δέντρων και $T_B(x)$ η πρόβλεψη του κάθε δέντρου για τα χαρακτηριστικά x . Η χρήση πολλών δέντρων μειώνει τη διακύμανση του μοντέλου και το καθιστά πιο σταθερό σε σχέση με ένα μεμονωμένο δέντρο απόφασης (decision tree).

Το μοντέλο κρίθηκε κατάλληλο, καθώς μπορεί να αποτυπώσει μη γραμμικές σχέσεις και αλληλεπιδράσεις μεταξύ των χαρακτηριστικών χωρίς να απαιτεί αυστηρές παραδοχές για τη μορφή της σχέσης μεταξύ εισόδων και στόχου. Οι συνδυασμοί υπερπαραμέτρων αξιολογήθηκαν στο σύνολο επικύρωσης και επιλέχθηκε ο συνδιασμός με το χαμηλότερο μέσο τετραγωνικό σφάλμα ρίζας.

Οι υπερπαραμέτροι που ελέγχθηκαν ήταν οι εξής:

- `n_estimators`: καθορίζει τον αριθμό των δέντρων των τυχαίων δασών. Στη παρούσα διπλωματική δοκιμάστηκαν οι τιμές 100, 200, 300 και 500. Περισσότερα δέντρα μπορούν να βελτιώσουν τη σταθερότητα του μοντέλου, αλλά αυξάνουν τον χρόνο εκπαίδευσης.
- `max_depth`: ορίζει το μέγιστο βάθος κάθε δέντρου. Δοκιμάστηκαν οι τιμές 4, 6 και 8. Μικρότερο βάθος περιορίζει την πολυπλοκότητα του μοντέλου, ενώ μεγαλύτερο βάθος επιτρέπει στο μοντέλο να μάθει πιο σύνθετες σχέσεις, με αυξημένο όμως κίνδυνο υπερπροσαρμογής (overfitting).
- `min_samples_leaf`: καθορίζει τον ελάχιστο αριθμό παρατηρήσεων που πρέπει να υπάρχουν σε κάθε τελικό φύλλο του δέντρου. Δοκιμάστηκαν οι τιμές 10 και 30. Μεγαλύτερη τιμή οδηγεί σε πιο ομαλές προβλέψεις και περιορίζει την υπερπροσαρμογή σε θορυβώδεις κινήσεις της αγοράς.
- `min_samples_split`: καθορίζει τον ελάχιστο αριθμό παρατηρήσεων που απαιτούνται για να διασπαστεί ένας κόμβος. Δοκιμάστηκαν οι τιμές 20 και 40. Μεγαλύτερη τιμή περιορίζει τον αριθμό των διασπάσεων και κάνει τα δέντρα λιγότερο σύνθετα.
- `max_features`: ορίζει πόσα χαρακτηριστικά εξετάζονται σε κάθε διάσπαση του δέντρου. Δοκιμάστηκαν οι επιλογές `sqrt` και 0.5. Η επιλογή `sqrt` χρησιμοποιεί περίπου την τετραγωνική ρίζα του αριθμού των χαρακτηριστικών, ενώ το 0.5 χρησιμοποιεί το 50% των χαρακτηριστικών. Η τυχαία επιλογή χαρακτηριστικών μειώνει τη συσχέτιση μεταξύ των δέντρων.
- `bootstrap`: καθορίζει αν κάθε δέντρο εκπαιδεύεται σε bootstrap δείγμα των δεδομένων. Στην εργασία χρησιμοποιήθηκε η τιμή `True`, δηλαδή κάθε δέντρο εκπαιδεύτηκε σε τυχαίο δείγμα των παρατηρήσεων με επανατοποθέτηση.

Επιπλέον, στην υλοποίηση χρησιμοποιήθηκε $\text{max_samples} = 0.5$, ώστε κάθε δέντρο να εκπαιδεύεται με το 50% των διαθέσιμων παρατηρήσεων. Αυτό μειώνει το υπολογιστικό κόστος και εισάγει επιπλέον τυχαιότητα στο μοντέλο, βοηθώντας στον περιορισμό του overfitting.

Μετά την επιλογή των βέλτιστων υπερπαραμέτρων, το τελικό μοντέλο εκπαιδεύτηκε στο ενιαίο σύνολο εκπαίδευσης και επικύρωσης. Επιπλέον, υπολογίστηκε η σημαντικότητα των εξωτερικών μεταβλητών του μοντέλου, ώστε να εντοπιστούν τα χαρακτηριστικά με τη μεγαλύτερη συμβολή στην πρόβλεψη της μελλοντικής απόδοσης.

3.5 Η παλινδρόμηση XGBoost (XGBoost Regression)

Η παλινδρόμηση XGBoost (XGBoost Regression) [24] εφαρμόστηκε ως δεύτερο μη γραμμικό συνδυαστικό μοντέλο. Σε αντίθεση με τα τυχαία δάση, όπου τα δέντρα εκπαιδεύονται σε μεγάλο βαθμό ανεξάρτητα, το XGBoost βασίζεται στη βαθμιαία ενίσχυση (gradient boosting). Τα δέντρα δημιουργούνται διαδοχικά, με κάθε νέο δέντρο να προσπαθεί να διορθώσει τα σφάλματα των προηγούμενων.

Η γενική λογική της ενίσχυσης μπορεί να αποδοθεί ως:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x)$$

Εξίσωση 6. Διαδοχική ενημέρωση του μοντέλου στη βαθμιαία ενίσχυση/XGBoost

όπου $F_m(x)$ είναι το μοντέλο μετά το m -οστό βήμα, $h_m(x)$ είναι το νέο δέντρο που προστίθεται και η είναι το learning rate. Στην περίπτωση της παρούσας διπλωματικής, η έξοδος του μοντέλου αντιστοιχεί στην προβλεπόμενη λογαριθμική απόδοση του Bitcoin.

Το μοντέλο κρίθηκε κατάλληλο επειδή μπορεί να αποτυπώσει μη γραμμικές σχέσεις και σύνθετες αλληλεπιδράσεις μεταξύ των χαρακτηριστικών. Αυτό είναι σημαντικό για την αγορά του Bitcoin, η οποία χαρακτηρίζεται από έντονη μεταβλητότητα και θόρυβο.

Για κάθε χρονικό ορίζοντα πρόβλεψης, το XGBoost εκπαιδεύτηκε και αξιολογήθηκε ξεχωριστά. Οι υπερπαραμέτροι επιλέχθηκαν μέσω αναζήτησης πλέγματος, όπου κάθε συνδυασμός εκπαιδεύτηκε στο σύνολο εκπαίδευσης και αξιολογήθηκε στο σύνολο επικύρωσης. Ως βέλτιστος συνδυασμός επιλέχθηκε εκείνος που πέτυχε το χαμηλότερο μέσω τετραγωνικό σφάλμα στις λογαριθμικές αποδόσεις.

Οι υπερπαραμέτροι που εξετάστηκαν περιλάμβαναν τον αριθμό των δέντρων ($n_estimators$), το μέγιστο βάθος (max_depth) με τιμές ίδιες με το μοντέλο τυχαίων δασών (Random Forest). Επιπλέον ελέγχθηκαν τα:

- `learning_rate`: καθορίζει πόσο έντονα συνεισφέρει κάθε νέο δέντρο στο συνολικό μοντέλο. Δοκιμάστηκαν οι τιμές 0.01 και 0.05. Μικρότερο `learning rate` οδηγεί σε πιο σταδιακή και σταθερή μάθηση.
- `subsample`: καθορίζει το ποσοστό των παρατηρήσεων που χρησιμοποιείται για την εκπαίδευση κάθε δέντρου. Δοκιμάστηκαν οι τιμές 0.6 και 0.8. Η χρήση υποσυνόλου παρατηρήσεων εισάγει τυχαιότητα και μειώνει το `overfitting`.
- `colsample_bytree`: καθορίζει το ποσοστό των χαρακτηριστικών που χρησιμοποιείται σε κάθε δέντρο. Δοκιμάστηκαν οι τιμές 0.6 και 0.8.
- `reg_alpha`: αντιστοιχεί στην $L1$ κανονικοποίηση. Δοκιμάστηκαν οι τιμές 0.0 και 0.01. Η $L1$ ποινή περιορίζει την πολυπλοκότητα του μοντέλου και μπορεί να μειώσει την επίδραση λιγότερο χρήσιμων χαρακτηριστικών.
- `reg_lambda`: αντιστοιχεί στην λ_2 κανονικοποίηση. Δοκιμάστηκαν οι τιμές 1.0 και 5.0. Μεγαλύτερη τιμή αυξάνει τον περιορισμό στους συντελεστές/βάρη του μοντέλου και βοηθά στη μείωση της υπερπροσαρμογής στα δεδομένα.

Στην υλοποίηση χρησιμοποιήθηκε ως `objective` το `reg:squarederror`, δηλαδή το μοντέλο εκπαιδεύτηκε με στόχο την ελαχιστοποίηση του τετραγωνικού σφάλματος. Μετά την επιλογή των βέλτιστων υπερπαραμέτρων, το τελικό XGBoost εκπαιδεύτηκε στο ενιαίο σύνολο εκπαίδευσης και επικύρωσης και αξιολογήθηκε στο σύνολο ελέγχου. Επιπλέον, υπολογίστηκε η σημαντικότητα των εξωτερικών μεταβλητών, ώστε να εξεταστεί ποια χαρακτηριστικά συνέβαλαν περισσότερο στην πρόβλεψη.

3.6 Δίκτυο Μακράς Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory- LSTM) και Δίκτυο Πυλών Επαναλαμβανόμενης Μονάδας (Gated Recurrent Unit- GRU)

Για την αξιοποίηση της ακολουθιακής δομής των δεδομένων εφαρμόστηκαν αρχιτεκτονικές επαναλαμβανόμενων νευρωνικών δικτύων (Recurrent Neural Networks – RNN): το Δίκτυο Μακράς

Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory - LSTM) [25] και το δίκτυο πυλών επαναλαμβανόμενης μονάδας (Gated Recurrent Unit - GRU) [26]. Σε αντίθεση με τα μη ακολουθιακά μοντέλα, τα οποία λαμβάνουν ως είσοδο μία μεμονωμένη γραμμή χαρακτηριστικών, τα LSTM και GRU λαμβάνουν ως είσοδο ακολουθίες πρόσφατων παρατηρήσεων. Έτσι, μπορούν να αξιοποιούν τη χρονική εξέλιξη της αγοράς και όχι μόνο την πληροφορία της τρέχουσας χρονικής στιγμής.

Η γενική ιδέα μπορεί να αποδοθεί ως:

$$\hat{r}_{t+h} = f(X_{t-l+1}, X_{t-l+2}, \dots, X_t)$$

Εξίσωση 7. Πρόβλεψη με ακολουθιακό παράθυρο εισόδου στα LSTM και GRU

όπου $\hat{r}_{t+h}$ είναι η προβλεπόμενη λογαριθμική απόδοση του Bitcoin για χρονικό ορίζοντα h , X_t είναι το διάνυσμα χαρακτηριστικών τη χρονική στιγμή t και l είναι το μήκος του παραθύρου αναδρομής (lookback window). Με αυτόν τον τρόπο, το μοντέλο δεν βασίζεται μόνο στην τρέχουσα παρατήρηση, αλλά σε μια πρόσφατη ακολουθία τιμών και τεχνικών χαρακτηριστικών.

Το LSTM είναι ειδική μορφή επαναλαμβανόμενου νευρωνικού δικτύου (Recurrent Neural Network – RNN) σχεδιασμένη για την αντιμετώπιση του προβλήματος εξαφάνισης της βαθμίδας (vanishing gradients) και για την αποτύπωση χρονικών εξαρτήσεων σε ακολουθιακά δεδομένα. Το GRU έχει παρόμοια λογική, αλλά απλούστερη αρχιτεκτονική και συνήθως λιγότερες παραμέτρους, καθώς δεν διαθέτει ξεχωριστή πύλη εξόδου (output gate), διατηρώντας όμως την ικανότητα να αποτυπώνει χρονικές εξαρτήσεις σε ακολουθιακά δεδομένα. Για τον λόγο αυτό χρησιμοποιήθηκε ως εναλλακτική του LSTM, ώστε να εξεταστεί αν μια πιο απλή επαναλαμβανόμενη αρχιτεκτονική μπορεί να αποδώσει ικανοποιητικά στο ίδιο πρόβλημα πρόβλεψης.

Πριν την εκπαίδευση, τα χαρακτηριστικά κανονικοποιήθηκαν με StandardScaler και στη συνέχεια μετατράπηκαν σε ακολουθίες κατάλληλες για είσοδο στα μοντέλα. Η αρχιτεκτονική και των δύο μοντέλων είχε κοινή λογική: ένα LSTM ή GRU layer, ένα επίπεδο εγκατάλειψης νευρώνων (Dropout layer) για περιορισμό της υπερπροσαρμογής, ένα πλήρως συνδεδεμένο επίπεδο (Dense layer) με 16 νευρώνες και συνάρτηση ενεργοποίησης ReLU και ένα τελικό συνδεδεμένο επίπεδο (Dense layer) με έναν νευρώνα για την πρόβλεψη της λογαριθμικής απόδοσης.

Για την εκπαίδευση χρησιμοποιήθηκε ο βελτιστοποιητής (optimizer) Adam και ως συνάρτηση απώλειας το μέσο τετραγωνικό σφάλμα ρίζας (MSE). Επιπλέον, χρησιμοποιήθηκε πρόωρη

διακοπή της εκπαίδευσης (Early Stopping) με τιμή υπομονής (patience) ίση με 5, ώστε η εκπαίδευση να σταματά όταν δεν υπάρχει βελτίωση στο σφάλμα επικύρωσης.

Οι υπερπαραμέτροι που ελέγχθηκαν ήταν οι εξής:

- **lookback:** καθορίζει πόσες προηγούμενες χρονικές στιγμές χρησιμοποιεί το μοντέλο ως είσοδο. Δοκιμάστηκαν οι τιμές 15, 30, 60 και 120. Μεγαλύτερο lookback επιτρέπει στο LSTM να αξιοποιεί μεγαλύτερο ιστορικό, αλλά αυξάνει την πολυπλοκότητα και τις απαιτήσεις μνήμης.
- **units:** καθορίζει τον αριθμό των LSTM νευρώνων. Δοκιμάστηκαν οι τιμές 32 και 64. Περισσότερα units αυξάνουν την ικανότητα του μοντέλου να μαθαίνει σύνθετα μοτίβα, αλλά αυξάνουν και τον κίνδυνο υπερπροσαρμογής.
- **dropout:** καθορίζει το ποσοστό των νευρώνων που απενεργοποιούνται τυχαία κατά την εκπαίδευση. Δοκιμάστηκαν οι τιμές 0.10, 0.20 και 0.30. Μεγαλύτερο dropout αυξάνει την κανονικοποίηση και βοηθά στον περιορισμό του overfitting.
- **batch_size:** καθορίζει πόσες ακολουθίες χρησιμοποιούνται σε κάθε βήμα ενημέρωσης των βαρών. Δοκιμάστηκαν οι τιμές 64 και 128. Μικρότερο batch size μπορεί να προσφέρει πιο συχνές ενημερώσεις, ενώ μεγαλύτερο batch size είναι συνήθως πιο αποδοτικό υπολογιστικά.
- **learning_rate:** καθορίζει το μέγεθος των βημάτων με τα οποία ενημερώνονται τα βάρη του μοντέλου. Δοκιμάστηκαν οι τιμές 0.001 και 0.0005. Μικρότερο learning rate οδηγεί σε πιο σταδιακή και σταθερή εκπαίδευση.

Για κάθε χρονικό ορίζοντα πρόβλεψης του Bitcoin, δοκιμάστηκαν 11 διαφορετικοί συνδυασμοί των παραπάνω υπερπαραμέτρων. Κάθε μοντέλο εκπαιδεύτηκε στο σύνολο εκπαίδευσης και αξιολογήθηκε στο επικύρωσης. Ως βέλτιστος συνδυασμός επιλέχθηκε εκείνος που πέτυχε το χαμηλότερο μέσο τετραγωνικό σφάλμα ρίζας στις λογαριθμικές αποδόσεις.

Μετά την επιλογή των βέλτιστων υπερπαραμέτρων, τα τελικά μοντέλα εκπαιδεύτηκαν στο ενιαίο σύνολο εκπαίδευσης και επικύρωσης και αξιολογήθηκαν στο σύνολο ελέγχου.

3.7 Ανάλυση συναισθήματος από κείμενα (sentiment analysis)

Για την εξαγωγή πληροφορίας συναισθήματος από τα κείμενα που συλλέχθηκαν από tweets, Reddit και ειδησεογραφικές πηγές, χρησιμοποιήθηκε το εργαλείο VADER (Valence Aware

Dictionary and sEntiment Reasoner). Το VADER [27] είναι λεξικογραφική μέθοδος ανάλυσης συναισθήματος, είναι κατάλληλη για σύντομα και άτυπα κείμενα, καθώς λαμβάνει υπόψη χαρακτηριστικά που εμφανίζονται συχνά σε περιεχόμενο μεσών κοινωνικής δικτύωσης, όπως η ένταση των λέξεων, η στίξη, η χρήση κεφαλαίων και οι σύντομες εκφράσεις συναισθήματος. Για τον λόγο αυτό κρίθηκε κατάλληλη η συγκεκριμένη προσέγγιση, καθώς μεγάλο μέρος των δεδομένων προέρχεται από σύντομα κείμενα χρηστών, όπως tweets και Reddit comments.

Δεν εφαρμόστηκε κάποιο εξειδικευμένο εργαλείο όπως το FinBERT που χρησιμοποιούν οι Gür & Ünal [15] για την εξαγωγή συναισθήματος από χρηματοοικονομικά κείμενα και ειδησεογραφικό περιεχόμενο. Το FinBERT έχει σχεδιαστεί για ανάλυση συναισθήματος σε χρηματοοικονομικό λόγο, αξιοποιώντας προεκπαιδευμένα γλωσσικά μοντέλα προσαρμοσμένα στο χρηματοοικονομικό πεδίο [28]. Ωστόσο, το σύνολο κειμένων της παρούσας διπλωματικής δεν αποτελείται αποκλειστικά από ειδησεογραφικά άρθρα ή τυπικό χρηματοοικονομικό λόγο, αλλά περιλαμβάνει και σύντομα, άτυπα και συχνά θορυβώδη κείμενα από κοινωνικά δίκτυα, όπως tweets και Reddit comments.

Για κάθε γραμμή κειμένου υπολογίστηκαν τέσσερις βασικές μεταβλητές: το `sentiment_neg`, το οποίο εκφράζει το ποσοστό αρνητικού συναισθήματος, το `sentiment_neu`, το οποίο εκφράζει το ουδέτερο συναίσθημα, το `sentiment_pos`, το οποίο εκφράζει το θετικό συναίσθημα, και το `sentiment_compound`, το οποίο αποτελεί τη συνολική κανονικοποιημένη βαθμολογία συναισθήματος του κειμένου. Η τιμή κυμαίνεται από -1 έως 1, όπου αρνητικές τιμές αντιστοιχούν σε αρνητικό συναίσθημα, θετικές τιμές σε θετικό συναίσθημα και τιμές κοντά στο μηδέν σε ουδέτερο συναίσθημα.

Στη συνέχεια, με βάση τη μεταβλητή συνολικού συναισθήματος `sentiment_compound`, κάθε κείμενο ταξινομήθηκε σε μία από τις τρεις κατηγορίες συναισθήματος: θετικό, αρνητικό ή ουδέτερο. Συγκεκριμένα, κείμενα με τιμή `compound` μεγαλύτερη ή ίση του 0.05 χαρακτηρίστηκαν ως θετικά, κείμενα με τιμή `compound` μικρότερη ή ίση του -0.05 χαρακτηρίστηκαν ως αρνητικά, ενώ οι ενδιάμεσες τιμές ταξινομήθηκαν ως ουδέτερες. Με αυτόν τον τρόπο, η μη δομημένη πληροφορία κειμένου μετατράπηκε σε αριθμητικά και κατηγορικά χαρακτηριστικά, τα οποία μπορούσαν να χρησιμοποιηθούν στη συνέχεια για διερευνητική ανάλυση και για συγχώνευση με τα δεδομένα τιμών του Bitcoin.

4. Μεθοδολογία έρευνας

Στο παρόν κεφάλαιο παρουσιάζεται ο τρόπος συλλογής, περιγραφής και προετοιμασίας των δεδομένων που χρησιμοποιήθηκαν στην εμπειρική ανάλυση και στην ανάπτυξη των μοντέλων πρόβλεψης.

4.1 Δεδομένα

Τα δεδομένα που χρησιμοποιήθηκαν στην ανάλυση αντλήθηκαν από την σελίδα Bitstamp by Robinhood [29] και αφορούν την ισοτιμία Bitcoin/Δολαρίου Αμερικής (BTC/USD). Από την πηγή αυτή αντλήθηκαν δεδομένα διαφορετικής χρονικής συχνότητας, και συγκεκριμένα ημερήσια, ωριαία και δεδομένα ανά λεπτό.

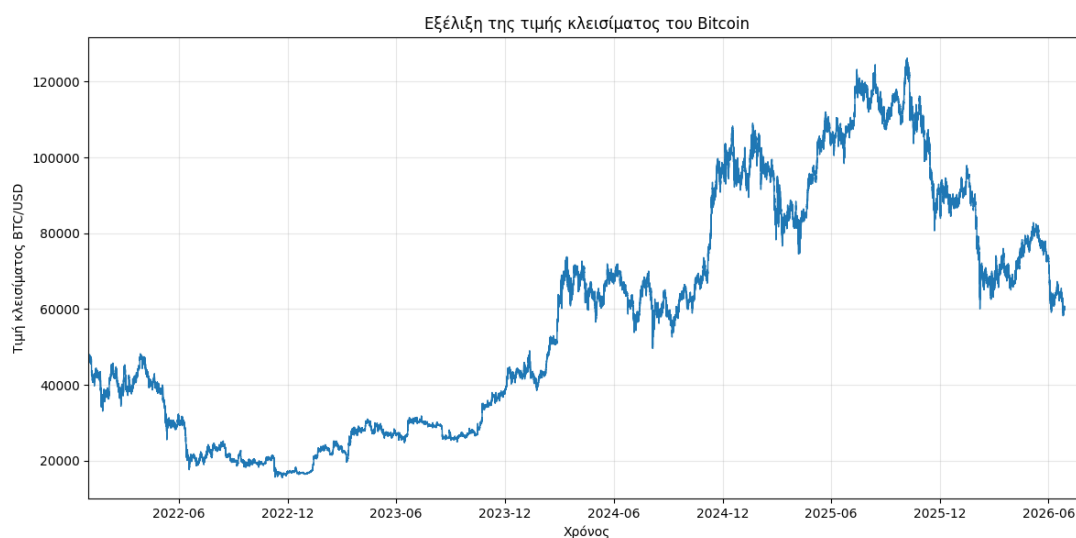
Η χρονική περίοδος που καλύπτουν τα δεδομένα εκτείνεται από τον Ιανουάριο του 2022 έως και τον Ιούνιο του 2026, με συνολικό αριθμό 2.361.527 παρατηρήσεων. Κάθε εγγραφή αντιστοιχεί σε μία χρονική περίοδο της αντίστοιχης συχνότητας και περιλαμβάνει το timestamp της εγγραφής, δηλαδή την ημερομηνία και ώρα της παρατήρησης, καθώς και τις βασικές μεταβλητές OHLCV τιμή ανοίγματος και κλεισίματος, τη μέγιστη και την ελάχιστη τιμή καθώς και τον όγκο συναλλαγών.

Στην περίπτωση των δεδομένων λεπτού, η τιμή ανοίγματος αντιστοιχεί στην πρώτη διαθέσιμη τιμή του Bitcoin στην αρχή του εκάστοτε λεπτού, ενώ η τιμή κλεισίματος αντιστοιχεί στην τελευταία διαθέσιμη τιμή στο τέλος του ίδιου λεπτού. Η μέγιστη και η ελάχιστη τιμή εκφράζουν αντίστοιχα τη μεγαλύτερη και τη μικρότερη τιμή που καταγράφηκε εντός του λεπτού. Τέλος, ο όγκος συναλλαγών αποτυπώνει τη συνολική ποσότητα συναλλαγών που πραγματοποιήθηκε στο συγκεκριμένο χρονικό διάστημα. Οι μεταβλητές αυτές αποτελούν τη βασική πληροφορία της αγοράς και χρησιμοποιούνται τόσο για την περιγραφική ανάλυση όσο και για τη δημιουργία πρόσθετων χαρακτηριστικών, όπως τεχνικοί δείκτες, μεταβλητότητα και λογαριθμικές αποδόσεις.

Η βασική μεταβλητή ενδιαφέροντος είναι η τιμή κλεισίματος του Bitcoin. Η επιλογή της τιμής κλεισίματος αποτελεί μία από τις πιο συχνά χρησιμοποιούμενες μεταβλητές στη χρηματοοικονομική ανάλυση και στην πρόβλεψη χρονοσειρών, καθώς συνοψίζει την τελική αποτίμηση του περιουσιακού στοιχείου στο τέλος κάθε χρονικής περιόδου. Παράλληλα, οι υπόλοιπες μεταβλητές, όπως οι τιμές ανοίγματος, μέγιστης και ελάχιστης τιμής, καθώς και ο όγκος συναλλαγών, αξιοποιούνται ως εξωτερικές μεταβλητές που περιγράφουν τη δυναμική, τη μεταβλητότητα και την ένταση της αγοράς.

Αν και αντλήθηκαν δεδομένα σε ημερήσια, ωριαία και λεπτού συχνότητα, η βασική ανάλυση της παρούσας εργασίας επικεντρώνεται κυρίως στα δεδομένα λεπτού, καθώς αυτά επιτρέπουν τη μελέτη της βραχυχρόνιας συμπεριφοράς της τιμής του Bitcoin και την ανάπτυξη μοντέλων πρόβλεψης για μικρούς χρονικούς ορίζοντες, όπως 1, 15 και 30 λεπτά.

Στο Διάγραμμα 1 παρουσιάζεται η εξέλιξη της τιμής κλεισίματος του Bitcoin σε δολάρια Αμερικής για την περίοδο 2022–2026, με βάση δεδομένα συχνότητας ενός λεπτού. Η χρονοσειρά εμφανίζει έντονες διακυμάνσεις, γεγονός που επιβεβαιώνει τη μεταβλητή φύση της αγοράς κρυπτονομισμάτων. Κατά τη διάρκεια του 2022 παρατηρείται σημαντική πτωτική πορεία, ενώ από το 2023 και κυρίως μέσα στο 2024 η τιμή παρουσιάζει έντονη ανοδική τάση. Στη συνέχεια, η περίοδος 2024–2025 χαρακτηρίζεται από υψηλότερα επίπεδα τιμών αλλά και αυξημένη μεταβλητότητα, με απότομες ανόδους και διορθώσεις με τη περίοδο 2025–2026 να παρουσιάζει ξανά πτωτική τάση.



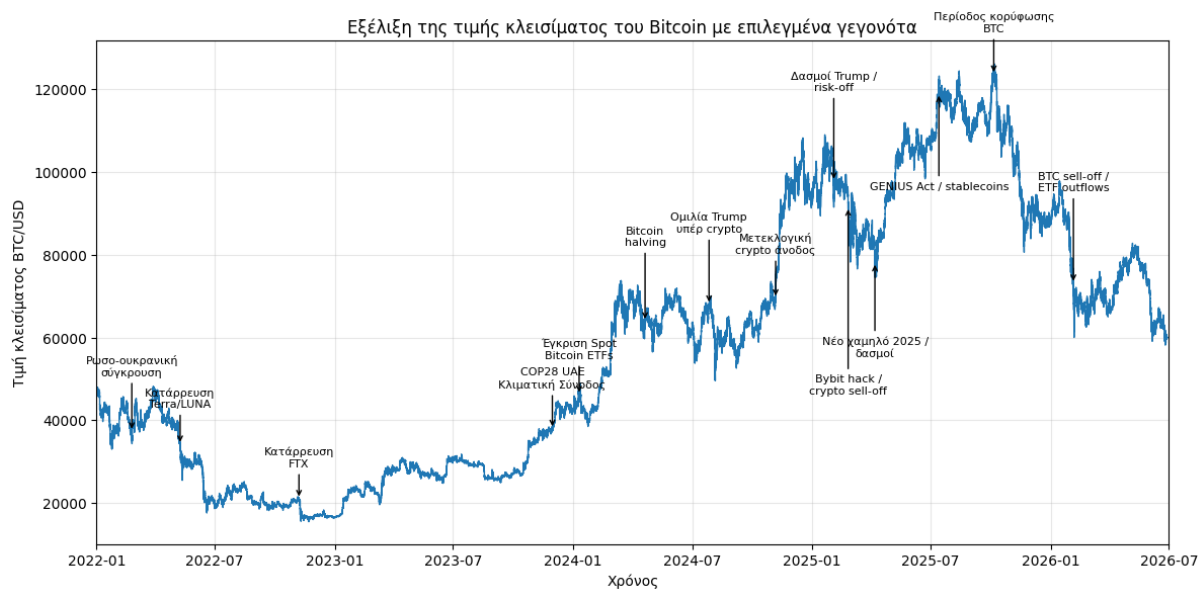
Διάγραμμα 1. Απεικόνιση της τιμής κλεισίματος του Bitcoin κάθε λεπτό (2022–2026)

Επιπλέον, στο Διάγραμμα 2 παρουσιάζεται η ίδια χρονοσειρά με την προσθήκη επιλεγμένων σημαντικών γεγονότων, τα οποία συμπίπτουν χρονικά με περιόδους έντονων μεταβολών στην τιμή του Bitcoin. Η επιλογή των γεγονότων βασίστηκε σε ιστορικές αναφορές και αναλύσεις αγοράς, όπως οι Sweeney [30], OANDA [31] και Bitcoin Magazine Pro [32], οι οποίες καταγράφουν γεγονότα όπως η κατάρρευση Terra/LUNA, η κατάρρευση της FTX, η έγκριση των spot Bitcoin ETFs, το Bitcoin halving και οι πολιτικές εξελίξεις γύρω από τα κρυπτονομίσματα ως σημαντικά σημεία στην ιστορική πορεία της αγοράς Bitcoin. Επιπλέον, υπάρχουν γεγονότα όπως η ανακοίνωση δασμών από τη κυβέρνηση Trump στις αρχές του 2025 που συνδέθηκε με επιδείνωση του επενδυτικού κλίματος και απομάκρυνση των επενδυτών από περιουσιακά στοιχεία υψηλού

κινδύνου, οδηγώντας το Bitcoin κάτω από τα 100.000 δολάρια [33]. Παράλληλα, η ανακοίνωση για δημιουργία στρατηγικού αποθέματος κρυπτονομισμάτων στις ΗΠΑ αποτέλεσε θετική πολιτική εξέλιξη για την αγορά, καθώς συνδέθηκε με σημαντική άνοδο των τιμών των βασικών κρυπτονομισμάτων. Επίσης, η ψήφιση του GENIUS Act τον Ιούλιο του 2025 αποτέλεσε σημαντικό ρυθμιστικό γεγονός, καθώς δημιούργησε ομοσπονδιακό πλαίσιο για τα stablecoins στις ΗΠΑ και θεωρήθηκε βήμα προς τη θεσμική ενσωμάτωση των ψηφιακών περιουσιακών στοιχείων. [34]

Για το 2026, η σημαντική πτώση του Φεβρουαρίου συνδέθηκε με εκτεταμένες ρευστοποιήσεις θέσεων, καθώς και με ευρύτερη αποστροφή κινδύνου στις αγορές. [35]

Τα γεγονότα αυτά περιλαμβάνουν γεωπολιτικές εξελίξεις, κρίσεις στην αγορά κρυπτονομισμάτων, ρυθμιστικές αποφάσεις και πολιτικοοικονομικές εξελίξεις που επηρέασαν το ευρύτερο επενδυτικό κλίμα. Η απεικόνιση αυτή είναι χρήσιμη, καθώς δίνει στοιχεία ως προς ότι η αγορά του Bitcoin δεν διαμορφώνεται αποκλειστικά από την προηγούμενη πορεία της τιμής, αλλά φαίνεται να συνδέεται και με εξωτερικές πληροφορίες, ειδήσεις και μεταβολές στο επενδυτικό συναίσθημα. Ειδικότερα, γεγονότα που συνδέονται με κρίσεις στην αγορά κρυπτονομισμάτων, όπως η κατάρρευση Terra/LUNA και η κατάρρευση της FTX και το hack στο ανταλλακτήριο Bybit εμφανίζονται σε περιόδους πτωτικής πίεσης και αυξημένης αβεβαιότητας. Αντίθετα, γεγονότα που συνδέονται με θετικές προσδοκίες για την αγορά, όπως η έγκριση των Spot Bitcoin ETFs, το Bitcoin halving, η αυξημένη πολιτική υποστήριξη προς τα κρυπτονομίσματα και η ψήφιση του GENIUS Act, συμπίπτουν με περιόδους ανοδικής δυναμικής. Ωστόσο, η παρατήρηση αυτή πρέπει να ερμηνευθεί με προσοχή, καθώς η οπτική συσχέτιση γεγονότων και μεταβολών της τιμής δεν αποδεικνύει σχέση αιτιώδους συνάφειας. Παρ' όλα αυτά, λειτουργεί ως ένδειξη της ευαισθησίας της αγοράς του Bitcoin σε σημαντικά γεγονότα και υποστηρίζει την ανάγκη διερεύνησης εξωτερικών παραγόντων, όπως οι ειδήσεις και το επενδυτικό συναίσθημα, στα μοντέλα πρόβλεψης.



Διάγραμμα 2. Εξέλιξη της τιμής κλεισίματος του Bitcoin με επιλεγμένα γεγονότα

4.2 Έλεγχος ποιότητας δεδομένων

Πριν από τη διαδικασία εφαρμογής και δοκιμής των διαφόρων μοντέλων πραγματοποιήθηκε έλεγχος ποιότητας των δεδομένων, ώστε να διασφαλιστεί πως η χρονοσειρά είναι σωστά δομημένη και κατάλληλη για ανάλυση υψηλής συχνότητας. Αρχικά, η στήλη timestamp (Ημερομηνία και ώρα της τιμής κλεισίματος) μετατράπηκε σε κατάλληλη μορφή ημερομηνίας και ώρας και χρησιμοποιήθηκε ως χρονικός δείκτης του συνόλου δεδομένων. Στη συνέχεια, οι παρατηρήσεις ταξινομήθηκαν χρονολογικά και ελέγχθηκε η ύπαρξη διπλότυπων χρονικών στιγμών. Σε περιπτώσεις όπου υπήρχαν περισσότερες από μία εγγραφές για το ίδιο timestamp, διατηρήθηκε η τελευταία διαθέσιμη παρατήρηση.

Ακολούθως, ελέγχθηκε αν η χρονοσειρά ακολουθεί σταθερή συχνότητα ενός λεπτού. Η σταθερή χρονική συχνότητα είναι απαραίτητη, καθώς τα μοντέλα που χρησιμοποιούνται στη συνέχεια βασίζονται σε υστερήσεις, τεχνικούς δείκτες και χαρακτηριστικά που προϋποθέτουν ομοιόμορφη χρονική απόσταση μεταξύ των παρατηρήσεων. Έτσι, η χρονοσειρά επαναδειγματοληπτήθηκε (resample) σε συχνότητα ενός λεπτού. Για κάθε λεπτό, διατηρήθηκε η τελευταία διαθέσιμη εγγραφή, ενώ στις περιπτώσεις ελλειπουσών χρονικών στιγμών εφαρμόστηκε συμπλήρωση (ffill) με την αμέσως προηγούμενη διαθέσιμη τιμή. Με αυτόν τον τρόπο δημιουργήθηκε ένα πλήρες χρονικό πλέγμα ενός λεπτού και διασφαλίστηκε η συνέπεια της χρονοσειράς.

Η επιλογή της μεθόδου forward fill έγινε με στόχο να διατηρηθεί η συνέχεια της σειράς χωρίς να εισαχθούν τεχνητές τιμές μέσω παρεμβολής. Η μέθοδος αυτή θεωρήθηκε κατάλληλη για

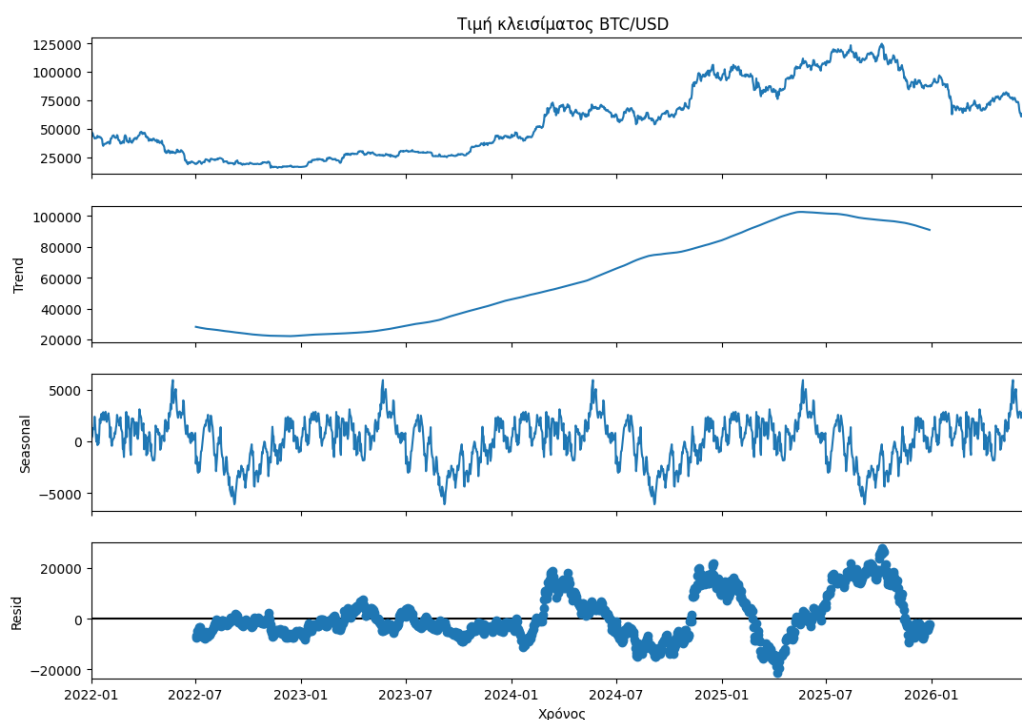
μεμονωμένες ελλείπουσες χρονικές στιγμές, καθώς υποθέτει ότι, ελλείψει νέας διαθέσιμης παρατήρησης, η τελευταία καταγεγραμμένη τιμή παραμένει η πιο αντιπροσωπευτική εκτίμηση για το αμέσως επόμενο λεπτό, δεδομένης της μεταβλητότητας που υπήρχε στη χρονοσειρά στις τελευταίες παρατηρήσεις πριν το κενό. Τέλος, ελέγχθηκε εκ νέου η χρονική συχνότητα της σειράς, ώστε να επιβεβαιωθεί ότι το τελικό σύνολο δεδομένων ακολουθεί σταθερό βήμα ενός λεπτού.

Επιπλέον, εξετάστηκε η ύπαρξη ελλিপών τιμών και ασυνεπειών στις βασικές μεταβλητές OHLCV, δηλαδή στην τιμή ανοίγματος, τη μέγιστη τιμή, την ελάχιστη τιμή, την τιμή κλεισίματος και τον όγκο συναλλαγών.

Για τη διερευνητική ανάλυση της δομής της χρονοσειράς εφαρμόστηκε κλασική αποσύνθεση χρονοσειράς (additive seasonal decomposition) με περίοδο 365 ημερών. Η αποσύνθεση εφαρμόστηκε σε ημερήσια δεδομένα τιμής κλεισίματος, τα οποία προέκυψαν από τη συμπύκνωση των αρχικών δεδομένων λεπτού. Η επιλογή αυτή έγινε επειδή η εφαρμογή κλασικής εποχικής αποσύνθεσης απευθείας σε δεδομένα λεπτού οδηγεί σε δυσκολότερα ερμηνεύσιμη εποχική συνιστώσα, λόγω της πολύ υψηλής συχνότητας των παρατηρήσεων και του μεγάλου αριθμού σημείων ανά πιθανό εποχικό κύκλο. Αντίθετα, η ημερήσια συχνότητα επιτρέπει πιο καθαρή αποτύπωση της μακροχρόνιας τάσης, των πιθανών επαναλαμβανόμενων μοτίβων και των υπολειμματικών διακυμάνσεων στη τιμή κλεισίματος του bitcoin.

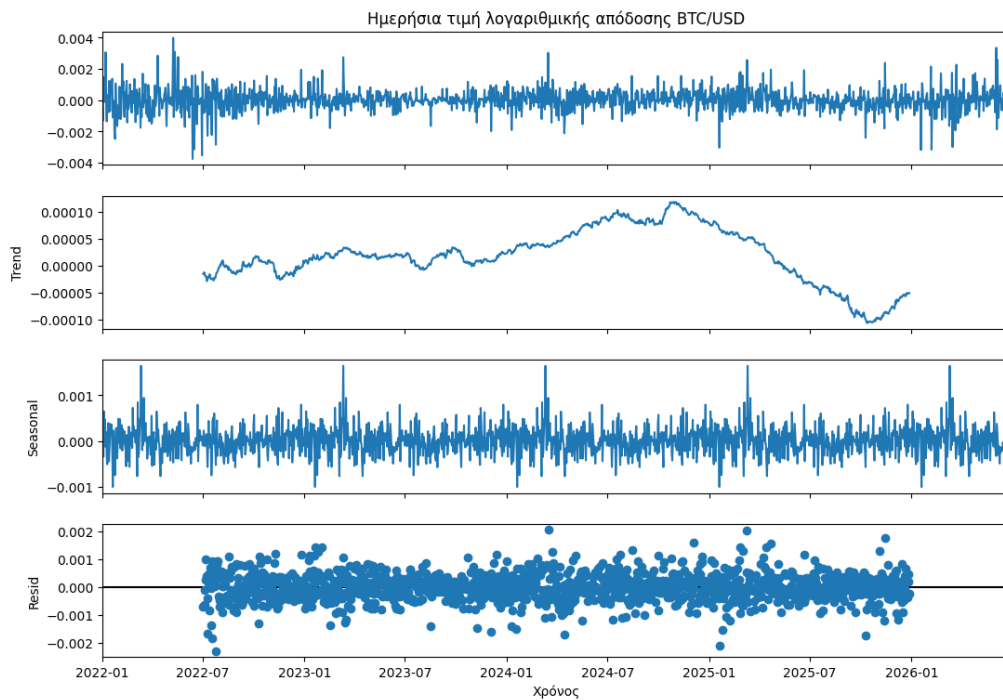
Στο Διάγραμμα 3 παρουσιάζεται η αποσύνθεση της αρχικής χρονοσειράς της τιμής κλεισίματος. Η συνιστώσα της τάσης (trend) αποτυπώνει τη γενικότερη ανοδική πορεία της τιμής κατά την εξεταζόμενη περίοδο, καθώς και τις ενδιάμεσες φάσεις διόρθωσης. Η εποχική συνιστώσα (seasonal) εμφανίζει ορισμένα επαναλαμβανόμενα μοτίβα, ωστόσο η ερμηνεία τους απαιτεί προσοχή, καθώς η αγορά κρυπτονομισμάτων δεν χαρακτηρίζεται από θεσμοθετημένη εποχικότητα, όπως συμβαίνει σε άλλες οικονομικές δραστηριότητες. Συνεπώς, η παρατηρούμενη εποχικότητα μπορεί να αποτελεί κυρίως αποτέλεσμα της μη στασιμότητας της σειράς και της ίδιας της μεθόδου αποσύνθεσης, παρά ένδειξη σταθερών οικονομικών κύκλων. [8]

Τέλος, η υπολειμματική συνιστώσα (residuals) παρουσιάζει μεγάλες και ακανόνιστες αποκλίσεις γύρω από το μηδέν, γεγονός που υποδηλώνει έντονη μεταβλητότητα και παρουσία απρόβλεπτων διακυμάνσεων. Το χαρακτηριστικό αυτό είναι σύμφωνο με τη φύση των αγορών κρυπτονομισμάτων, όπου οι τιμές μπορεί να συνδέονται με ειδήσεις, μεταβολές στο επενδυτικό συναίσθημα, ρυθμιστικές εξελίξεις και μακροοικονομικά γεγονότα.



Διάγραμμα 3.Αποσύνθεση της χρονοσειράς

Στη συνέχεια, για περαιτέρω διερεύνηση της στασιμότητας της σειράς, εφαρμόστηκε αντίστοιχη αποσύνθεση μετά τον υπολογισμό των λογαριθμικών αποδόσεων της τιμής κλεισίματος. Όπως παρουσιάζεται στο Διάγραμμα 4, μετά την αφαίρεση της μακροχρόνιας τάσης, η εποχική συνιστώσα αποδυναμώνεται σημαντικά και δεν παρουσιάζει σταθερά επαναλαμβανόμενα μοτίβα. Το εύρημα αυτό υποδηλώνει ότι η εποχικότητα που παρατηρείται στο επίπεδο των τιμών σχετίζεται κυρίως με τη μη στασιμότητα της αρχικής χρονοσειράς και όχι με εγγενείς περιοδικές συμπεριφορές στις μεταβολές της τιμής του Bitcoin.



Διάγραμμα 4. Αποσύνθεση της χρονοσειράς με λογαριθμική απόδοση λεπτού

Η παραπάνω ανάλυση ενισχύει την ανάγκη μετασχηματισμού της αρχικής χρονοσειράς πριν από τη μοντελοποίηση. Η τιμή κλεισίματος του Bitcoin παρουσιάζει έντονη τάση και μη στασιμότητα, γεγονός που μπορεί να δυσχεράνει την εκπαίδευση των μοντέλων και να οδηγήσει σε παραπλανητικές σχέσεις. Για τον λόγο αυτό, στη συνέχεια η πρόβλεψη δεν πραγματοποιείται απευθείας στο επίπεδο της τιμής, αλλά στη μελλοντική λογαριθμική απόδοση, η οποία αποτελεί πιο κατάλληλη μεταβλητή στόχο για χρηματοοικονομικές χρονοσειρές υψηλής συχνότητας.

Για την περαιτέρω διερεύνηση της στασιμότητας και της εξάρτησης της χρονοσειράς από προηγούμενες τιμές, εφαρμόστηκε ο έλεγχος Augmented Dickey-Fuller (ADF) [36], καθώς και τα διαγράμματα αυτοσυσχέτισης και μερικής αυτοσυσχέτισης. Ο έλεγχος ADF χρησιμοποιήθηκε για να εξεταστεί η ύπαρξη μοναδιαίας ρίζας, ενώ τα διαγράμματα αυτοσυσχέτισης (ACF) και μερικής αυτοσυσχέτισης (PACF) αξιοποιήθηκαν για τη διερεύνηση της αυτοσυσχέτισης και της μερικής αυτοσυσχέτισης της σειράς.

Ο έλεγχος ADF αποτελεί έναν από τους πιο διαδεδομένους ελέγχους στασιμότητας και χρησιμοποιείται για να εξεταστεί αν μία χρονοσειρά παρουσιάζει μοναδιαία ρίζα. Η ύπαρξη μοναδιαίας ρίζας σημαίνει ότι η σειρά είναι μη στάσιμη, δηλαδή τα στατιστικά της χαρακτηριστικά, όπως ο μέσος όρος και η διακύμανση, δεν παραμένουν σταθερά στον χρόνο. Σε τέτοιες περιπτώσεις, οι μεταβολές της σειράς μπορεί να έχουν μόνιμη επίδραση στο επίπεδό της,

με αποτέλεσμα η χρονοσειρά να παρουσιάζει έντονη εξάρτηση από τις προηγούμενες τιμές της και να μην επιστρέφει εύκολα γύρω από έναν σταθερό μακροχρόνιο μέσο.

Η μη στασιμότητα έχει σημαντική επίπτωση στη διαδικασία πρόβλεψης. Όταν ένα μοντέλο εκπαιδεύεται απευθείας πάνω σε μη στάσιμες τιμές, υπάρχει κίνδυνος να αποτυπώσει κυρίως τη γενική τάση της περιόδου εκπαίδευσης και όχι σταθερές σχέσεις που μπορούν να γενικευθούν σε μελλοντικά δεδομένα. Επιπλέον, οι σχέσεις μεταξύ των μεταβλητών μπορεί να είναι παραπλανητικές, καθώς δύο σειρές με έντονη τάση μπορούν να εμφανίζουν υψηλή συσχέτιση χωρίς να υπάρχει ουσιαστική προβλεπτική σχέση. Για τον λόγο αυτό, σε χρηματοοικονομικές χρονοσειρές χρησιμοποιούνται συχνά μετασχηματισμοί όπως οι πρώτες διαφορές ή οι λογαριθμικές αποδόσεις, οι οποίοι αφαιρούν σε μεγάλο βαθμό την τάση και οδηγούν σε πιο στάσιμες σειρές. Η μηδενική υπόθεση του ελέγχου είναι ότι η χρονοσειρά διαθέτει μοναδιαία ρίζα και επομένως είναι μη στάσιμη. Αντίθετα, η εναλλακτική υπόθεση είναι ότι η χρονοσειρά δεν διαθέτει μοναδιαία ρίζα και μπορεί να θεωρηθεί στάσιμη. Συνεπώς, όταν η τιμή πιθανότητας p (p -value) του ελέγχου είναι μεγαλύτερη από το επιλεγμένο επίπεδο σημαντικότητας, δεν υπάρχουν στατιστικά σημαντικές ενδείξεις ως προς την απόρριψη της μηδενικής υπόθεσης και η σειρά δεν μπορεί να θεωρηθεί στάσιμη. Αντίθετα, όταν το p -value είναι μικρότερο από το επίπεδο σημαντικότητας, η μηδενική υπόθεση απορρίπτεται και υπάρχουν ενδείξεις ότι η σειρά είναι στάσιμη.

Αρχικά, ο έλεγχος ADF εφαρμόστηκε στην χρονοσειρά της τιμής κλεισίματος του Bitcoin. Τα αποτελέσματα όπως φαίνονται στο πίνακα 1 έδειξαν τιμή στατιστικού ADF ίση με -0.834 και p -value ίσο με 0.808. Δεδομένου ότι το p -value είναι σημαντικά μεγαλύτερο από το επίπεδο σημαντικότητας του 0,05, δεν υπάρχουν στατιστικά σημαντικές ενδείξεις ως προς την απόρριψη της μηδενικής υπόθεσης, δηλαδή της ύπαρξης μοναδιαίας ρίζας. Συνεπώς, η αρχική χρονοσειρά της τιμής κλεισίματος δεν μπορεί να θεωρηθεί στάσιμη.

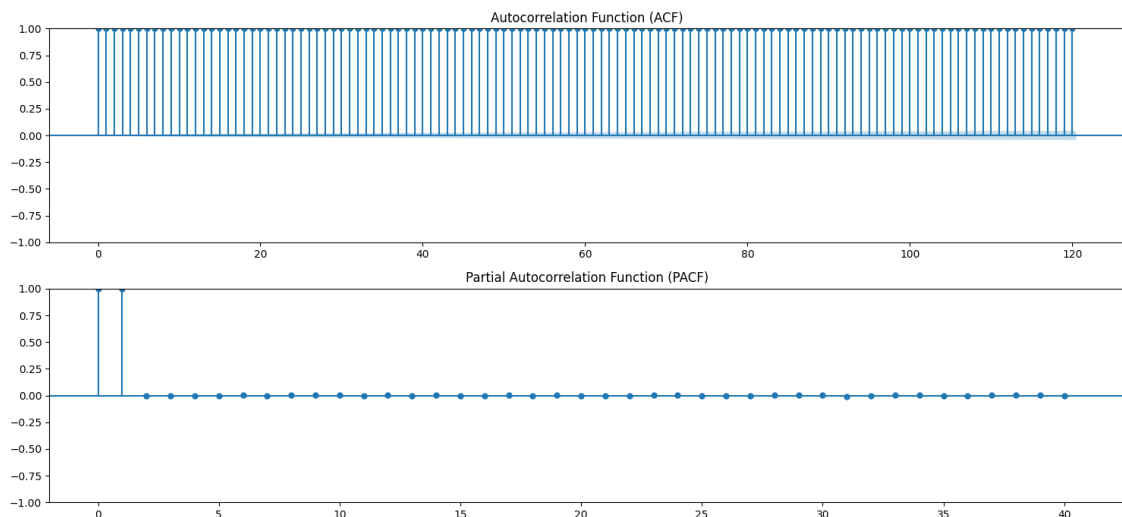
Αντίθετα, για τις λογαριθμικές αποδόσεις των εξεταζόμενων χρονικών οριζόντων, τα αποτελέσματα του ελέγχου ADF οδηγούν σε διαφορετικό συμπέρασμα. Συγκεκριμένα, για τη λογαριθμική απόδοση ενός λεπτού το στατιστικό ADF είναι -71.863 και το p -value είναι 0.000. Αντίστοιχα, για τη λογαριθμική απόδοση 15 λεπτών το στατιστικό ADF είναι -69.478 με p -value 0.000, ενώ για τη λογαριθμική απόδοση 30 λεπτών το στατιστικό ADF είναι -60.562 με p -value 0.000. Σε όλες τις περιπτώσεις, το p -value είναι μικρότερο από το επίπεδο σημαντικότητας 0.05. Επομένως, υπάρχουν στατιστικά σημαντικές ενδείξεις ως προς την απόρριψη της μηδενικής υπόθεσης δηλαδή της ύπαρξης μοναδιαίας ρίζας. Το αποτέλεσμα αυτό διαφοροποιείται

σημαντικά από την αρχική τιμή κλεισίματος και επιβεβαιώνει ότι ο μετασχηματισμός της τιμής σε λογαριθμική απόδοση περιορίζει τη μη στασιμότητα της αρχικής χρονοσειράς.

Πίνακας 1. Αποτελέσματα ελέγχου ADF για την τιμή κλεισίματος και τις λογαριθμικές αποδόσεις.

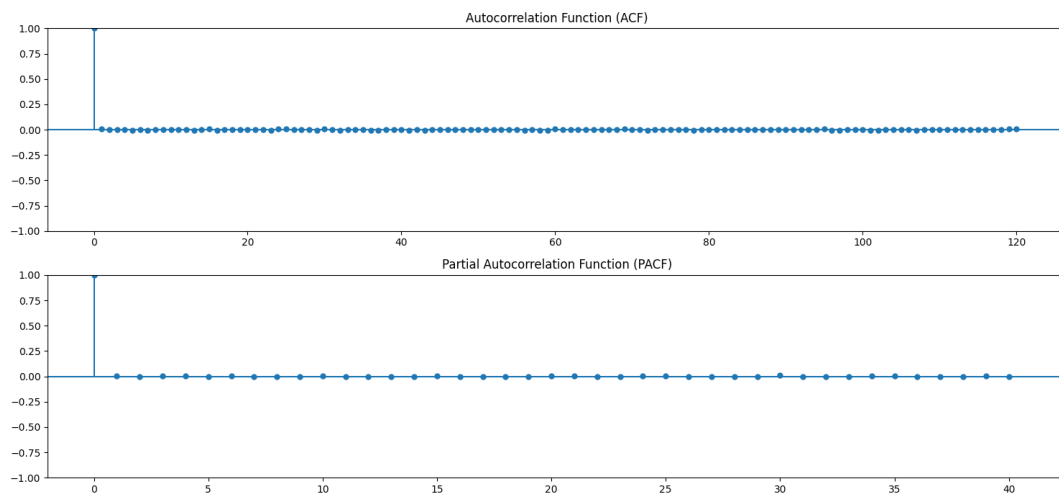
Χρονοσειρά	ADF statistic	p-value	Συμπέρασμα
Τιμή κλεισίματος Bitcoin	-0.834	0.808	Μη στάσιμη
Λογαριθμική απόδοση 1 λεπτού	-71.863	0.000	Στάσιμη
Λογαριθμική απόδοση 15 λεπτών	-69.478	0.000	Στάσιμη
Λογαριθμική απόδοση 30 λεπτών	-60.562	0.000	Στάσιμη

Τα συμπεράσματα του ελέγχου ADF ενισχύονται και από τα διαγράμματα αυτοσυσχέτισης και μερικής αυτοσυσχέτισης. Στο Διάγραμμα 5, που αφορά την αρχική τιμή κλεισίματος, η συνάρτηση αυτοσυσχέτισης παρουσιάζει πολύ υψηλές τιμές σε μεγάλο αριθμό υστερήσεων και φθίνει αργά. Η συμπεριφορά αυτή αποτελεί τυπική ένδειξη μη στασιμότητας και δείχνει ότι η αρχική σειρά εξαρτάται έντονα από τις προηγούμενες τιμές της. Αντίστοιχα, στο διάγραμμα PACF παρατηρείται ισχυρή μερική αυτοσυσχέτιση στις πρώτες υστερήσεις, ενώ στη συνέχεια η επίδραση περιορίζεται.



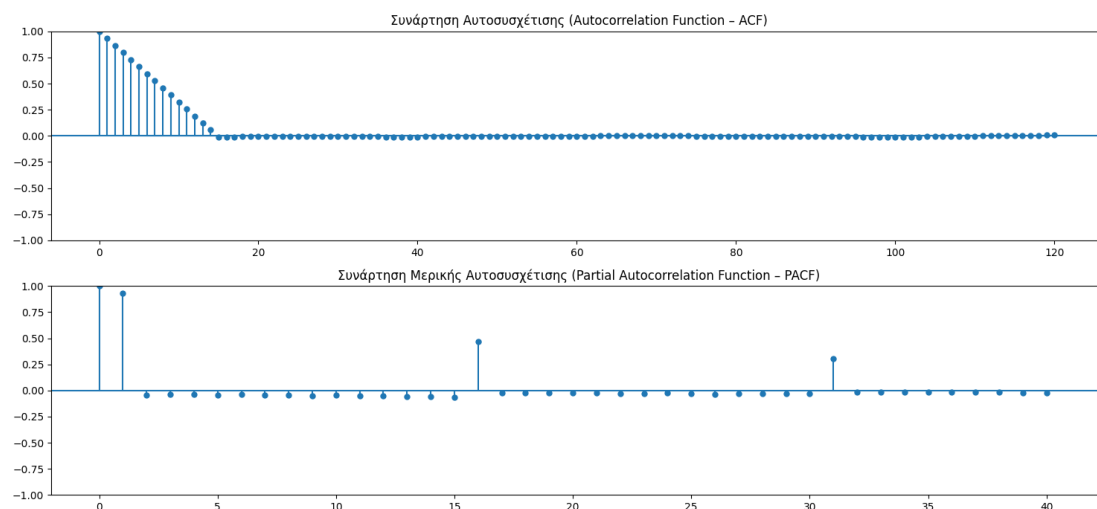
Διάγραμμα 5. Αυτοσυσχέτισης και μερικής αυτοσυσχέτισης στη τιμή κλεισίματος του bitcoin.

Αντίθετα, στο Διάγραμμα 6, που αφορά τη λογαριθμική απόδοση ενός λεπτού, οι τιμές της αυτοσυσχέτισης και της μερικής αυτοσυσχέτισης είναι ιδιαίτερα περιορισμένες και συγκεντρώνονται κοντά στο μηδέν μετά την αρχική υστέρηση. Η εικόνα αυτή δείχνει ότι οι λογαριθμικές αποδόσεις δεν παρουσιάζουν την έντονη επίμονη εξάρτηση που παρατηρείται στην αρχική τιμή κλεισίματος. Επομένως, ο μετασχηματισμός σε λογαριθμικές αποδόσεις αφαιρεί σε μεγάλο βαθμό τη μακροχρόνια τάση και οδηγεί σε σειρά καταλληλότερη για μοντελοποίηση.



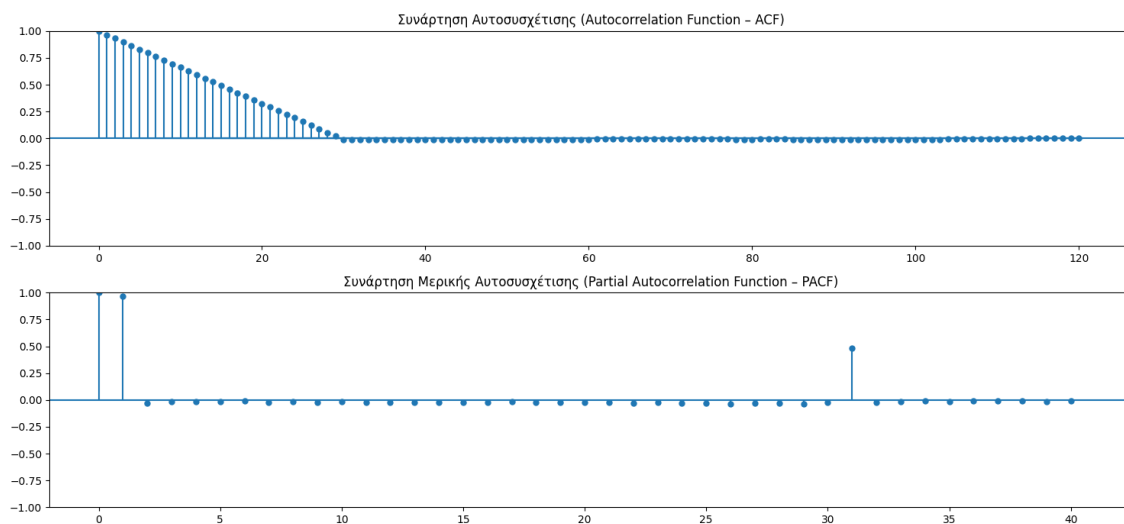
Διάγραμμα 6. Αυτοσυσχέτισης και μερικής αυτοσυσχέτισης της λογαριθμικής απόδοσης λεπτού

Στο Διάγραμμα 7 παρουσιάζονται η ACF και η PACF για τη λογαριθμική απόδοση 15 λεπτών. Παρότι ο έλεγχος ADF υποδεικνύει στάσιμη συμπεριφορά, η συνάρτηση αυτοσυσχέτισης εμφανίζει θετικές τιμές στις πρώτες υστερήσεις, οι οποίες μειώνονται σταδιακά και πλησιάζουν το μηδέν μετά από περίπου 15 υστερήσεις. Η συμπεριφορά αυτή είναι αναμενόμενη, καθώς η λογαριθμική απόδοση 15 λεπτών υπολογίζεται με βάση επικαλυπτόμενα χρονικά παράθυρα σε δεδομένα συχνότητας ενός λεπτού. Κατά συνέπεια, διαδοχικές παρατηρήσεις της μεταβλητής μοιράζονται μεγάλο μέρος της ίδιας πληροφορίας, γεγονός που δημιουργεί βραχυχρόνια αυτοσυσχέτιση. Στο PACF παρατηρείται έντονη επίδραση στις πρώτες υστερήσεις, ενώ στη συνέχεια οι τιμές περιορίζονται σημαντικά, με ορισμένες μεμονωμένες αιχμές.



Διάγραμμα 7. Αυτοσυσχέτιση και μερική αυτοσυσχέτιση της λογαριθμικής απόδοσης 15 λεπτών

Αντίστοιχη εικόνα παρατηρείται και στο Διάγραμμα 8 για τη λογαριθμική απόδοση 30 λεπτών. Η ACF παρουσιάζει υψηλές θετικές τιμές στις αρχικές υστερήσεις και φθίνει σταδιακά μέχρι να προσεγγίσει το μηδέν περίπου μετά την τριακοστή υστέρηση. Η μορφή αυτή συνδέεται με τον τρόπο υπολογισμού της απόδοσης 30 λεπτών, καθώς οι διαδοχικές τιμές βασίζονται σε έντονα επικαλυπτόμενα διαστήματα. Επομένως, η ύπαρξη βραχυχρόνιας αυτοσυσχέτισης στους μεγαλύτερους ορίζοντες δεν αναιρεί το συμπέρασμα περί στασιμότητας, αλλά δείχνει ότι οι σειρές αυτές περιέχουν χρονική εξάρτηση που μπορεί να αξιοποιηθεί από τα μοντέλα πρόβλεψης.



Διάγραμμα 8. Αυτοσυσχέτιση και μερική αυτοσυσχέτιση της λογαριθμικής απόδοσης 30 λεπτών

Συνεπώς, τα αποτελέσματα του ADF, του ACF και του PACF υποστηρίζουν την επιλογή της λογαριθμικής απόδοσης ως βασικής μεταβλητής στόχου. Η χρήση των λογαριθμικών αποδόσεων επιτρέπει τη μοντελοποίηση μιας πιο στάσιμης σειράς, η οποία είναι καταλληλότερη για

χρηματοοικονομικές χρονοσειρές υψηλής συχνότητας και για την ανάπτυξη στατιστικών μοντέλων, μοντέλων μηχανικής μάθησης και μοντέλων βαθιάς μάθησης.

4.3 Μετασχηματισμός μεταβλητής στόχου

Με βάση τα αποτελέσματα του ελέγχου στασιμότητας και της ανάλυσης αυτοσυσχέτισης καθώς και τη σχετική βιβλιογραφία, η μοντελοποίηση δεν πραγματοποιήθηκε απευθείας στην τιμή κλεισίματος του Bitcoin, αλλά στη μελλοντική λογαριθμική απόδοση. Η επιλογή αυτή έγινε καθώς η αρχική τιμή κλεισίματος παρουσιάζει έντονη τάση και μη στασιμότητα. Με τις στατιστικές ιδιότητες της χρονοσειράς να μεταβάλλονται στο χρόνο, η απευθείας χρήση της τιμής κλεισίματος ως μεταβλητής στόχου μπορεί να οδηγήσει τα μοντέλα στο να μαθαίνουν κυρίως τη μακροχρόνια ανοδική ή πτωτική τάση της περιόδου εκπαίδευσης, αντί για πιο γενικεύσιμα μοτίβα βραχυχρόνιας μεταβολής.

Μία εναλλακτική προσέγγιση για την αφαίρεση της τάσης από τη χρονοσειρά θα ήταν η χρήση της απλής πρώτης διαφοράς της τιμής κλεισίματος, δηλαδή:

$$\text{price_diff}_t = \text{Close}_t - \text{Close}_{t-1}$$

Εξίσωση 8. Υπολογισμός απλής πρώτης διαφοράς της τιμής κλεισίματος

Η απλή διαφορά δείχνει πόσα δολάρια μεταβλήθηκε η τιμή μεταξύ δύο διαδοχικών χρονικών στιγμών. Ωστόσο, το βασικό μειονέκτημά της είναι ότι εξαρτάται άμεσα από το επίπεδο της τιμής. Για παράδειγμα, μία αύξηση 100 δολαρίων όταν το Bitcoin βρίσκεται στις 20.000 δολάρια αντιστοιχεί σε μεταβολή 0,50%, ενώ η ίδια αύξηση 100 δολαρίων όταν η τιμή βρίσκεται στις 80.000 δολάρια αντιστοιχεί σε μεταβολή μόλις 0,125%. Επομένως, η απλή διαφορά αντιμετωπίζει τις δύο μεταβολές ως ίσες, παρότι η σχετική οικονομική τους σημασία είναι διαφορετική.

Αντίθετα, η λογαριθμική απόδοση υπολογίζεται ως:

$$r_t = \log (\text{Close}_t / \text{Close}_{t-1})$$

Εξίσωση 9. Υπολογισμός λογαριθμικής απόδοσης ενός λεπτού

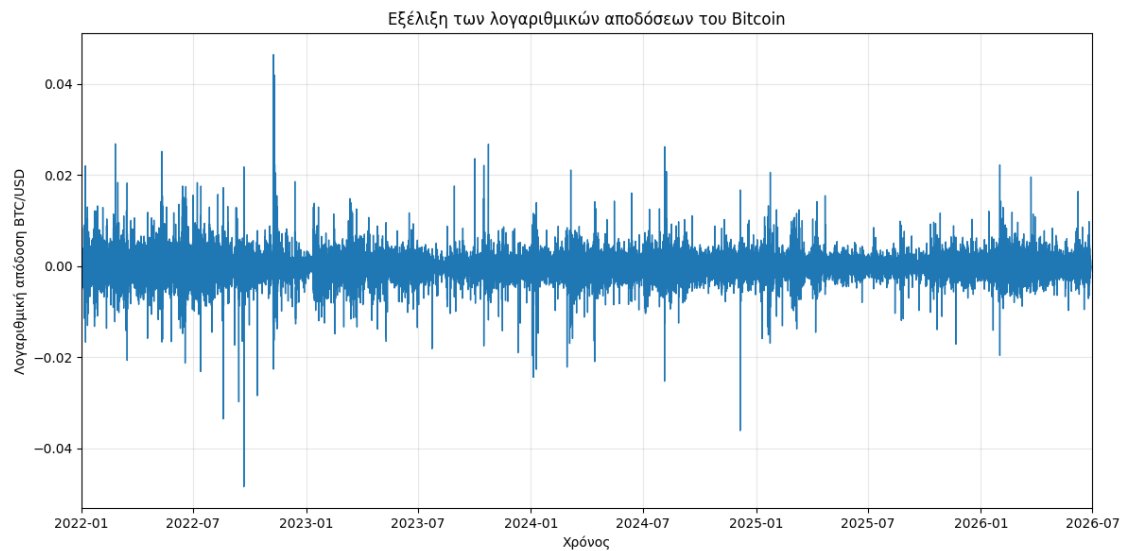
Σε αντίθεση με την απλή διαφορά, μετρά τη σχετική μεταβολή της τιμής και όχι την απόλυτη μεταβολή σε δολάρια. Για τον λόγο αυτό είναι πιο κατάλληλη σε χρονοσειρές με έντονες μεταβολές. Επιπλέον, η χρήση της λογαριθμικής απόδοσης κάνει τη σύγκριση και με άλλα κρυπτονομίσματα της αγοράς πιο άμεση και εύκολη.

Για τον λόγο αυτό, η μεταβλητή στόχος μετασχηματίστηκε από απόλυτη τιμή κλεισίματος σε μελλοντική λογαριθμική απόδοση. Η λογαριθμική απόδοση εκφράζει τη σχετική μεταβολή της τιμής μεταξύ δύο χρονικών στιγμών και όχι το ίδιο το επίπεδο της τιμής. Έτσι, το πρόβλημα πρόβλεψης μετατρέπεται σε “πόσο θα μεταβληθεί η τιμή” σε ένα συγκεκριμένο χρονικό ορίζοντα παρά του πόσο θα είναι η τιμή.

Ενδεικτικά, οι AlMadany, et al. [37] χρησιμοποιούν λογαριθμική απόδοση ώστε να αντιμετωπιστεί το πρόβλημα της μη στασιμότητας και να βελτιωθεί η σταθερότητα των δεδομένων. Επιπλέον, επισημαίνουν ότι οι αρχικές σειρές τιμών παρουσιάζουν τάση και σημεία καμπής, ενώ οι σειρές αποδόσεων δεν εμφανίζουν αντίστοιχη τάση και παρουσιάζουν πιο σταθερή συμπεριφορά ως προς τον μέσο και την τυπική απόκλιση, καθιστώντας τη χρονοσειρά κατάλληλη για πρόβλεψη.

Οι Sung, et al. [38] χρησιμοποιούν λογαριθμικές αποδόσεις για την πρόβλεψη βασικών κρυπτονομισμάτων, όπως το Bitcoin, το Ethereum και το Binance Coin. Η επιλογή αυτή συνδέεται με την υψηλή μεταβλητότητα της αγοράς κρυπτονομισμάτων και με την ανάγκη μετασχηματισμού των τιμών σε μορφή που αποτυπώνει καλύτερα τις σχετικές μεταβολές. Οι συγγραφείς τονίζουν επίσης ότι οι τιμές διαφορετικών κρυπτονομισμάτων έχουν πολύ διαφορετικές κλίμακες, γεγονός που καθιστά αναγκαία την προεπεξεργασία τους. Τέλος, οι Liu, Li, Nekhili, & Sultan [39] εστιάζουν στην πρόβλεψη αποδόσεων κρυπτονομισμάτων και όχι στην πρόβλεψη των απόλυτων τιμών τους. Η προσέγγιση αυτή είναι σύμφωνη με τη χρηματοοικονομική λογική, στην οποία η επενδυτική σημασία βρίσκεται κυρίως στη μεταβολή της αξίας ενός περιουσιακού στοιχείου και όχι μόνο στο επίπεδο της τιμής του. Η χρήση αποδόσεων επιτρέπει την ανάλυση της προβλεψιμότητας της αγοράς με όρους μεταβολών και τη δυναμική της στιγμής.

Συνεπώς, προτιμήθηκε η λογαριθμική απόδοση έναντι της απλής πρώτης διαφοράς της τιμής κλεισίματος. Η απλή διαφορά αφαιρεί μεν μέρος της τάσης, αλλά παραμένει εκφρασμένη σε απόλυτες μονάδες τιμής και επηρεάζεται από το επίπεδο της αγοράς. Αντίθετα, η λογαριθμική απόδοση αποτυπώνει τη σχετική μεταβολή, είναι πιο συγκρίσιμη στον χρόνο, μειώνει την επίδραση της κλίμακας και οδηγεί σε πιο κατάλληλη μορφή χρονοσειράς.



Διάγραμμα 9. Απεικόνιση της λογαριθμικής απόδοσης του Bitcoin κάθε λεπτό (2022-2026)

Στο Διάγραμμα 9 παρουσιάζεται η λογαριθμική απόδοση του Bitcoin ανά λεπτό. Σε αντίθεση με την αρχική τιμή κλεισίματος, η οποία εμφανίζει έντονη μακροχρόνια τάση, η σειρά των λογαριθμικών αποδόσεων κινείται γύρω από το μηδέν και δεν παρουσιάζει αντίστοιχη τάση. Παράλληλα, είναι εμφανής η ύπαρξη περιόδων αυξημένης μεταβλητότητας και απότομων διακυμάνσεων, χαρακτηριστικό που είναι σύνηθες στις χρηματοοικονομικές χρονοσειρές υψηλής συχνότητας και ειδικά στην αγορά κρυπτονομισμάτων.

Η λογαριθμική απόδοση ενός λεπτού ορίζεται ως:

$$r_t = \log(\text{Close}_t / \text{Close}_{t-1})$$

όπου Close_t είναι η τιμή κλεισίματος του Bitcoin τη χρονική στιγμή t και Close_{t-1} η τιμή κλεισίματος της προηγούμενης χρονικής στιγμής. Για κάθε χρονικό ορίζοντα πρόβλεψης h , η μεταβλητή στόχος ορίστηκε ως η μελλοντική λογαριθμική απόδοση:

$$r_{t+h} = \log(\text{Close}_{t+h} / \text{Close}_t)$$

Εξίσωση 10. Υπολογισμός μελλοντικής λογαριθμικής απόδοσης για χρονικό ορίζοντα h

Εξετάστηκαν οι χρονικοί ορίζοντες 1, 15 και 30 λεπτών, προκειμένου να αξιολογηθεί κατά πόσο η προβλεπτική ικανότητα των μοντέλων διαφοροποιείται σε διαφορετικές βραχυχρόνιες περιόδους πρόβλεψης. Η επιλογή βραχυχρόνιων ενδοημερήσιων οριζόντων πρόβλεψης συνδέεται με τη φύση των αγορών κρυπτονομισμάτων και τη σημασία της υψηλής συχνότητας ανάλυσης. Οι Akyildirim, Goncu, & Sensoy [7] επισημαίνουν ότι μεγάλο μέρος της σχετικής βιβλιογραφίας

επικεντρώνεται σε ημερήσιες αποδόσεις, ενώ η ενδοημερήσια ανάλυση έχει μελετηθεί λιγότερο. Για τον λόγο αυτό εξετάζουν την προβλεψιμότητα των αποδόσεων κρυπτονομισμάτων σε διαφορετικές χρονικές κλίμακες, χρησιμοποιώντας ορίζοντες 15, 30 και 60 λεπτών, καθώς και ημερήσια συχνότητα. Επιπλέον, οι Jaquart, Dann, & Weinhardt [40] εξετάζουν την προβλεψιμότητα της αγοράς σε ορίζοντες από 1 έως 60 λεπτά και δείχνουν ότι η προβλεπτική ακρίβεια τείνει να αυξάνεται όσο μεγαλώνει ο χρονικός ορίζοντας πρόβλεψης.

Παρότι η εκπαίδευση των μοντέλων πραγματοποιήθηκε σε επίπεδο λογαριθμικών αποδόσεων, οι προβλέψεις μετατράπηκαν στη συνέχεια και σε προβλεπόμενες τιμές κλεισίματος, ώστε να είναι δυνατή η αξιολόγηση τόσο σε επίπεδο απόδοσης όσο και σε επίπεδο τιμής.

Τέλος, πέρα από την ακρίβεια της αριθμητικής πρόβλεψης, αξιολογήθηκε και η ικανότητα των μοντέλων να προβλέπουν σωστά την κατεύθυνση της μελλοντικής μεταβολής της τιμής. Για τον σκοπό αυτό χρησιμοποιήθηκε η ακρίβεια κατεύθυνσης, η οποία συγκρίνει το πρόσημο της πραγματικής λογαριθμικής απόδοσης με το πρόσημο της προβλεπόμενης λογαριθμικής απόδοσης. Με αυτόν τον τρόπο εξετάζεται και αν το μοντέλο προβλέπει σωστά αν η τιμή του Bitcoin θα κινηθεί ανοδικά ή καθοδικά στον εκάστοτε χρονικό ορίζοντα, ανεξάρτητα από το ακριβές μέγεθος της μεταβολής.

4.4 Διαχωρισμός δεδομένων σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου

Παρότι στην περιγραφική ανάλυση και στα αρχικά διαγράμματα παρουσιάστηκε ολόκληρη η διαθέσιμη χρονοσειρά της τιμής του Bitcoin από το 2022 έως το 2026, η εκπαίδευση και η αξιολόγηση των μοντέλων πραγματοποιήθηκαν στο υποσύνολο των δεδομένων από τις αρχές Μαρτίου 2025 και έπειτα. Η επιλογή αυτή έγινε καθώς από το συγκεκριμένο χρονικό σημείο ήταν διαθέσιμα και τα δεδομένα συναισθήματος που αξιοποιήθηκαν στην παρούσα διπλωματική. Με αυτόν τον τρόπο, τα μοντέλα με και χωρίς μεταβλητές συναισθήματος αξιολογήθηκαν στην ίδια χρονική περίοδο και κάτω από συγκρίσιμες συνθήκες, ώστε να εξεταστεί με πιο δίκαιο τρόπο αν η προσθήκη των sentiment χαρακτηριστικών βελτιώνει ουσιαστικά την προβλεπτική απόδοση.

Η παρουσίαση ολόκληρης της περιόδου 2022–2026 στα πρώτα διαγράμματα κρίθηκε απαραίτητη για τη συνολική κατανόηση της πορείας της τιμής του Bitcoin, της μακροχρόνιας τάσης και των περιόδων αυξημένης μεταβλητότητας. Επιπλέον, η πλήρης απεικόνιση της χρονοσειράς επέτρεψε τη σύνδεση σημαντικών γεγονότων της αγοράς με περιόδους έντονων μεταβολών στην τιμή.

Μετά τη δημιουργία της μεταβλητής στόχου και των αντίστοιχων χαρακτηριστικών, το τελικό σύνολο δεδομένων διαχωρίστηκε σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου. Ο διαχωρισμός πραγματοποιήθηκε χρονολογικά και όχι τυχαία, καθώς σε προβλήματα χρονοσειρών είναι απαραίτητο να διατηρείται η χρονική σειρά των παρατηρήσεων. Με αυτόν τον τρόπο αποφεύγεται η διαρροή πληροφορίας από μελλοντικές παρατηρήσεις προς το παρελθόν και η αξιολόγηση των μοντέλων προσομοιάζει περισσότερο μια πραγματική διαδικασία πρόβλεψης.

Συγκεκριμένα, το αρχικό τμήμα της χρονοσειράς χρησιμοποιήθηκε ως σύνολο εκπαίδευσης, το επόμενο τμήμα ως σύνολο επικύρωσης και το τελικό τμήμα ως σύνολο ελέγχου. Ο διαχωρισμός έγινε με αναλογία 80% για το σύνολο εκπαίδευσης, 10% για το σύνολο επικύρωσης και 10% για το σύνολο ελέγχου. Το σύνολο εκπαίδευσης χρησιμοποιήθηκε για την εκμάθηση των παραμέτρων των μοντέλων, το σύνολο επικύρωσης για την επιλογή υπερπαραμέτρων και τη σύγκριση εναλλακτικών διαμορφώσεων, ενώ το σύνολο ελέγχου διατηρήθηκε αποκλειστικά για την τελική αξιολόγηση της προβλεπτικής απόδοσης.

Η ίδια διαδικασία εφαρμόστηκε ξεχωριστά για κάθε χρονικό ορίζοντα πρόβλεψης. Αυτό κρίθηκε απαραίτητο, καθώς ο ορισμός της μεταβλητής στόχου μεταβάλλεται ανάλογα με τον χρονικό ορίζοντα h . Για κάθε ορίζοντα, αφαιρέθηκαν οι παρατηρήσεις για τις οποίες δεν ήταν διαθέσιμη η αντίστοιχη μελλοντική τιμή, ώστε το σύνολο δεδομένων να περιλαμβάνει μόνο πλήρεις γραμμές με διαθέσιμα χαρακτηριστικά και μεταβλητή στόχο.

Πίνακας 2. Διαχωρισμός συνόλου δεδομένων

Σύνολο	Περίοδος	Πλήθος	Ποσοστό
Εκπαίδευσης	2025-03-06 14:43 – 2026-03-24 13:54	551.472	80%
Επικύρωσης	2026-03-24 13:55 – 2026-05-11 10:48	68.934	10%
Ελέγχου	2026-05-11 10:49 – 2026-06-28 07:42	68.934	10%

4.5 Συλλογή δεδομένων συναισθήματος από κοινωνικά δίκτυα και ειδησεογραφικές πηγές

Πέρα από τα αριθμητικά δεδομένα της αγοράς του Bitcoin, στην παρούσα εργασία αξιοποιήθηκαν και δεδομένα κειμένου με στόχο την ενσωμάτωση πληροφορίας που σχετίζεται με το επενδυτικό συναίσθημα γύρω από το Bitcoin και την αγορά κρυπτονομισμάτων. Η χρήση τέτοιων δεδομένων βασίζεται στην υπόθεση ότι οι ειδήσεις, οι δημόσιες συζητήσεις και η δραστηριότητα στα κοινωνικά δίκτυα μπορούν να περιέχουν πληροφορία για τη βραχυχρόνια συμπεριφορά της αγοράς, καθώς επηρεάζουν τις προσδοκίες, την αβεβαιότητα και τη στάση των επενδυτών. [14]

Για τη δημιουργία του συνόλου δεδομένων συναισθήματος χρησιμοποιήθηκαν δύο κατηγορίες πηγών. Η πρώτη κατηγορία αφορά έτοιμο σύνολο δεδομένων με tweets σχετικά με το Bitcoin, ενώ η δεύτερη αφορά πρόσθετα δεδομένα που συλλέχθηκαν από Tweeter, Reddit και ειδησεογραφικές πηγές. Με τον τρόπο αυτό, επιχειρήθηκε να δημιουργηθεί ένα ευρύτερο σύνολο πληροφορίας που δεν περιορίζεται αποκλειστικά στις ιστορικές τιμές, αλλά αποτυπώνει και τη δημόσια συζήτηση γύρω από την αγορά.

Αρχικά, χρησιμοποιήθηκε το διαθέσιμο σύνολο δεδομένων Bitcoin Tweets Dataset (2025–2026) από το Kaggle. Το συγκεκριμένο dataset περιλαμβάνει tweets σχετικά με το Bitcoin και μεταδεδομένα των χρηστών, όπως όνομα χρήστη, τοποθεσία χρήστη, περιγραφή προφίλ, ημερομηνία δημιουργίας χρήστη, αριθμό followers, αριθμό friends, αριθμό favourites, ένδειξη επαλήθευσης χρήστη, ημερομηνία δημοσίευσης, κείμενο tweet, hashtags, πηγή δημοσίευσης και ένδειξη retweet. Σύμφωνα με το αντίστοιχο Kaggle notebook, το dataset περιλαμβάνει 4.689.355 γραμμές και 13 στήλες, γεγονός που το καθιστά ιδιαίτερα εκτεταμένη πηγή δεδομένων για ανάλυση συναισθήματος σε σχέση με το Bitcoin. Το dataset αυτό χρησιμοποιήθηκε ως βασική πηγή ιστορικών tweets για την περίοδο 2025–2026.

Παράλληλα, αναπτύχθηκε διεργασία συλλογής νέων δεδομένων από το X/Twitter μέσω της βιβλιοθήκης Tweepy και χρησιμοποιώντας το API key του X Developer Console [41]. Η διεργασία βασίστηκε σε αναζήτηση πρόσφατων δημοσιεύσεων με όρους σχετικούς με το Bitcoin και τα κρυπτονομίσματα, όπως “bitcoin”, “crypto” και συναφείς λέξεις-κλειδιά, περιορίζοντας τα αποτελέσματα σε αγγλόφωνες δημοσιεύσεις και αποκλείοντας retweets. Για κάθε tweet αποθηκεύτηκαν το μοναδικό αναγνωριστικό, η ημερομηνία δημοσίευσης, το κείμενο και βασικά engagement metrics, όπως likes, retweets, replies και quotes. Για την αποφυγή διπλότυπων εγγραφών χρησιμοποιήθηκε το μοναδικό tweet id, ενώ εφαρμόστηκε έλεγχος μέσω του since_id,

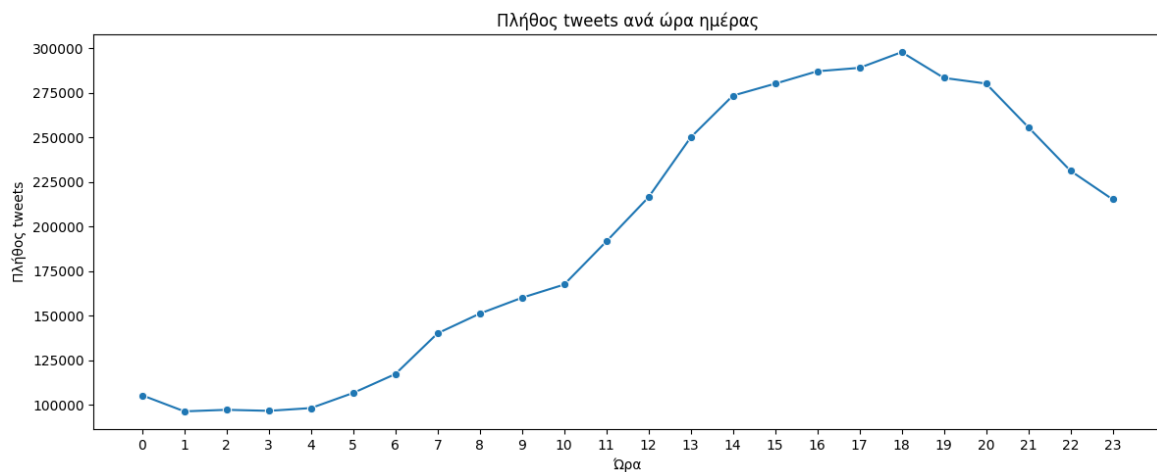
ώστε κάθε εκτέλεση να συλλέγει μόνο νέες δημοσιεύσεις και να μη γίνεται επανάληψη ήδη αποθηκευμένων δεδομένων.

Στο πίνακα 3 παρουσιάζονται ενδεικτικές εγγραφές από το ενοποιημένο σύνολο δεδομένων των tweets. Στον πίνακα εμφανίζονται η ημερομηνία δημοσίευσης, το κείμενο του tweet και ο αριθμός λέξεων, ο οποίος χρησιμοποιήθηκε ως βασικό χαρακτηριστικό για την περιγραφική ανάλυση του περιεχομένου.

Πίνακας 3. Παράδειγμα ενοποιημένου πίνακα tweets

Ημερομηνία	Κείμενο	Πλήθος λέξεων
2026-01-06 17:46:35	Bitcoin squeeze is SUPER TIGHT, which way will we break? I am BULLISH 🙌🚀🚀	21
2026-01-06 17:46:29	Closed #BTC short at 16725. Missed my long place by a hair. Will wait to see how this reacts before choosing a direction (freaking crabwalk)	25
2026-01-06 17:46:22	1฿ = \$16814.31 -0.09% ▼ Details: Change: ▼ - 14.33\$ 24H Low = \$16812.25 ▼ 24H High = \$16867.95 📈 #bitcoin #btc	17

Επιπλέον, όπως είναι εμφανές στο διάγραμμα 10 η δραστηριότητα είναι χαμηλότερη τις πρώτες πρωινές ώρες, ενώ αυξάνεται σταδιακά από το πρωί και κορυφώνεται κατά τις απογευματινές ώρες. Το μοτίβο αυτό δείχνει ότι η δραστηριότητα στο X/Twitter δεν είναι ομοιόμορφα κατανεμημένη μέσα στην ημέρα, γεγονός που είναι σημαντικό για τη μετέπειτα χρονική ευθυγράμμιση των δεδομένων sentiment με τις τιμές του Bitcoin.



Διάγραμμα 10. Πλήθος tweets ανά ώρα ημέρας

Επιπλέον, συλλέχθηκαν δεδομένα από το Reddit μέσω της βιβλιοθήκης PRAW. Η συλλογή επικεντρώθηκε σε subreddits σχετικά με την αγορά κρυπτονομισμάτων, όπως CryptoCurrency, Bitcoin, ethereum, CryptoMarkets, btc και ethtrader. Για τον εντοπισμό σχετικών δημοσιεύσεων και σχολίων χρησιμοποιήθηκαν λέξεις-κλειδιά όπως bitcoin, btc, ethereum, eth, crypto, cryptocurrency, altcoin και defi. Από το Reddit συλλέχθηκαν τόσο δημοσιεύσεις (posts) όσο και σχόλια (comments). Για κάθε post αποθηκεύτηκαν πληροφορίες όπως id, ημερομηνία δημιουργίας, subreddit, τίτλος, κείμενο, score, αριθμός σχολίων, permalink και url. Για κάθε comment αποθηκεύτηκαν αντίστοιχα το id, η ημερομηνία δημιουργίας, το subreddit, το κείμενο του σχολίου, το score και το permalink. Επιπλέον, εφαρμόστηκε καθαρισμός ώστε να αφαιρεθούν πολύ σύντομα κείμενα και εγγραφές τύπου [deleted] ή [removed], ενώ χρησιμοποιήθηκε έλεγχος με βάση το created_utc, ώστε να συλλέγονται μόνο νέες εγγραφές σε επόμενες εκτελέσεις.

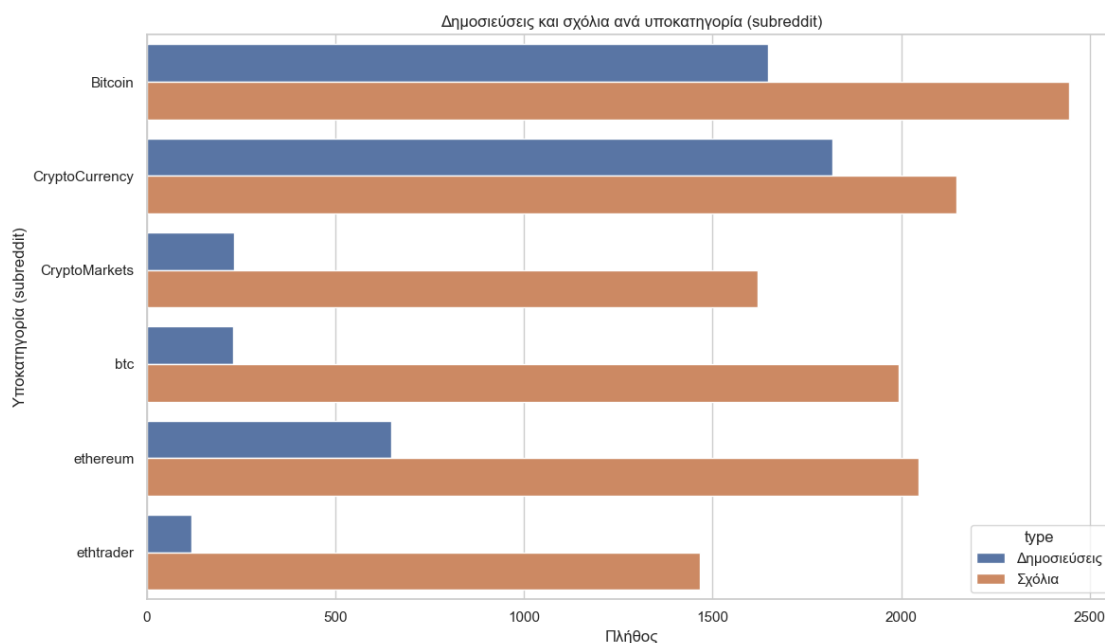
Στο πίνακα 4 παρουσιάζονται ενδεικτικές εγγραφές από το ενοποιημένο σύνολο Reddit. Για την ενοποίηση των δημοσιεύσεων και των σχολίων δημιουργήθηκε κοινό πεδίο κειμένου, ενώ διατηρήθηκε το score ως μέτρο αλληλεπίδρασης της κοινότητας με κάθε εγγραφή.

Πίνακας 4. Παράδειγμα ενοποιημένου πίνακα reddit δημοσιεύσεων και σχολίων

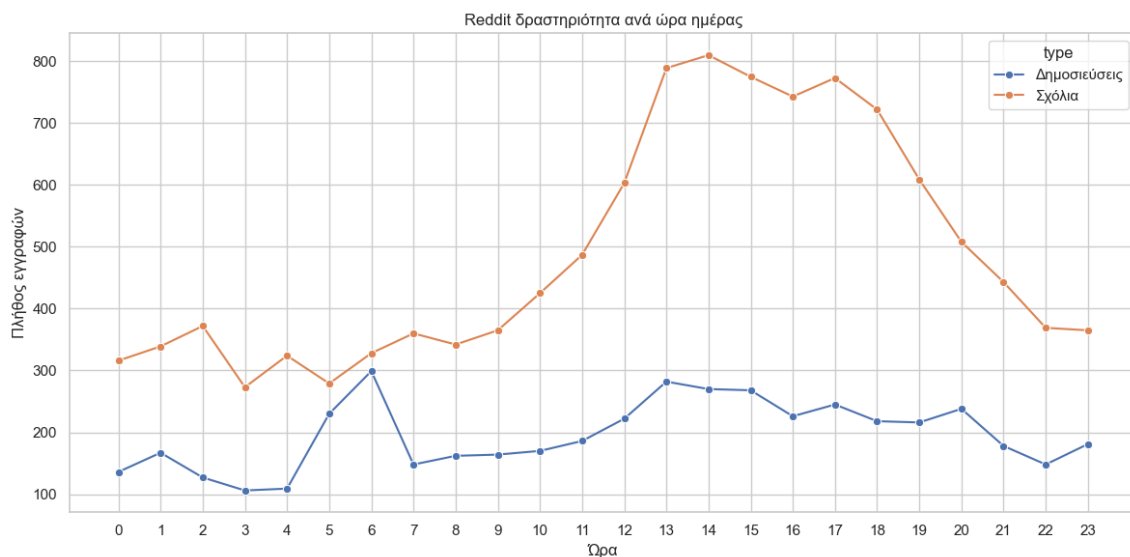
Ημερομηνία	Κείμενο	Score
2026-03-03 02:40:18	South Korean tax officials accidentally leaked a tax evasion suspect's crypto wallet seed phrases in a press release, leading to the theft of up to \$5 million worth of cryptocurrency, specifically Pre-Retogeum. The National Tax Service (NTS) admitted the error	105

	and is working with police to recover the assets. The incident highlights a lack of understanding of cryptocurrency security by officials.	
2026-03-03 07:18:17	That's like saying there are 1000 billionaire and 200 vintage rolls Royce cars in the world. Why are the billionaire not fighting over them. The reality is not everyone wants a rolls Royce and not everyone wants a bitcoin"	351

Στο διάγραμμα 11, παρουσιάζεται η κατανομή των δημοσιεύσεων (posts) και σχολίων (comments) ανά υποκατηγορία (subreddit). Παρατηρείται ότι τα subreddits Bitcoin και CryptoCurrency συγκεντρώνουν σημαντικό αριθμό δημοσιεύσεων, ενώ σε άλλα subreddits, όπως CryptoMarkets, btc και ethtrader, το σύνολο δεδομένων αποτελείται κυρίως από σχόλια. Επιπλέον, στο διάγραμμα 12 φαίνεται πως τα σχόλια εμφανίζουν υψηλότερη συχνότητα σε σχέση με τις δημοσιεύσεις στις περισσότερες ώρες, ενώ η συνολική δραστηριότητα αυξάνεται κυρίως από τις μεσημεριανές έως τις απογευματινές ώρες.



Διάγραμμα 11. Δημοσιεύσεις και σχόλια ανά υποκατηγορία (subreddit)



Διάγραμμα 12. Πλήθος δημοσιεύσεων και σχολίων ανά ώρα ημέρας

Τέλος, συλλέχθηκαν ειδησεογραφικά δεδομένα από RSS feeds σημαντικών πηγών του χώρου των κρυπτονομισμάτων. Συγκεκριμένα, χρησιμοποιήθηκαν ειδησεογραφικά site όπως coindesk [42], cointelegraph [43], cryptoslate [44], cryptopotato [45], cryptonews [46] και defiant [47]. Για κάθε άρθρο αποθηκεύτηκαν το id, η ημερομηνία δημοσίευσης ή ενημέρωσης, η πηγή, ο τίτλος, η περίληψη, το url και το αρχικό HTML summary όπου ήταν διαθέσιμο. Η περίληψη καθαρίστηκε από HTML tags και περιττά κενά, ώστε να προκύπτει καθαρό κείμενο κατάλληλο για ανάλυση συναισθήματος. Για την αποφυγή διπλότυπων εγγραφών χρησιμοποιήθηκαν τόσο το id όσο και το url του άρθρου, ενώ εφαρμόστηκε χρονικό παράθυρο συλλογής πρόσφατων άρθρων και έλεγχος με βάση την τελευταία ημερομηνία που είχε καταγραφεί.

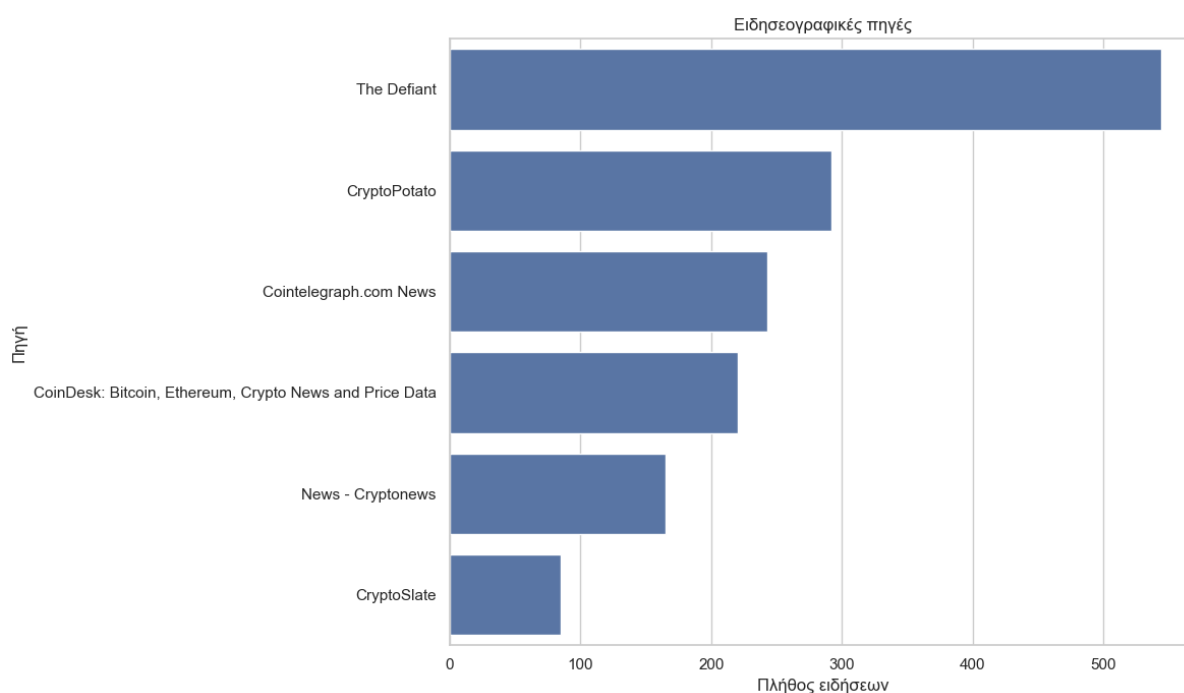
Στο πίνακα 5 παρουσιάζονται ενδεικτικές εγγραφές από το ειδησεογραφικό dataset. Σε κάθε εγγραφή διατηρείται η ημερομηνία, το καθαρισμένο κείμενο και η πηγή προέλευσης του άρθρου. Τα δεδομένα αυτά αποτελούν πιο επίσημη και δομημένη μορφή πληροφόρησης σε σχέση με τα δεδομένα από κοινωνικά δίκτυα.

Πίνακας 5. Παράδειγμα ενοποιημένου πίνακα ειδήσεων

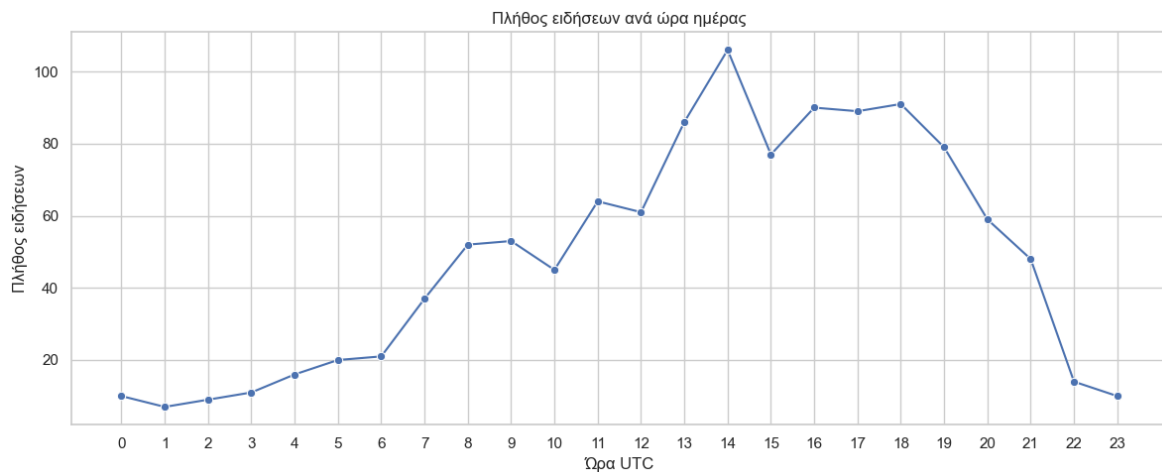
Ημερομηνία	Κείμενο	Πηγή
2026-02-05 15:05:11	Tether's \$100 million investment in Anchorage Digital underscores a commitment to secure, regulated financial systems, reinforcing Anchorage's	The Defiant

	status as the first federally chartered crypto bank in the U.S.	
2026-02-05 19:53:53	Bitcoin plunged under \$66,000 while most altcoins cratered.	The Defiant
2026-02-20 18:02:33	Bitcoin's mining difficulty climbed to 144.4 trillion after January storms briefly slashed hash rate, while some US miners offset downtime by selling electricity back to the grid.	Cointelegraph.com News

Το Διάγραμμα 13 παρουσιάζει το πλήθος των άρθρων ανά ειδησεογραφική πηγή. Η πηγή The Defiant εμφανίζεται με τον μεγαλύτερο αριθμό ειδήσεων, ενώ ακολουθούν πηγές όπως CryptoPotato, Cointelegraph και CoinDesk. Τέλος, το διάγραμμα 14 παρουσιάζει το πλήθος ειδήσεων ανά ώρα ημέρας. Η ειδησεογραφική δραστηριότητα είναι χαμηλή κατά τις πρώτες πρωινές ώρες και αυξάνεται σημαντικά από το πρωί έως το απόγευμα, με κορύφωση γύρω στις μεσημεριανές και απογευματινές ώρες. Το μοτίβο αυτό διαφέρει από τα κοινωνικά δίκτυα, καθώς οι ειδήσεις επηρεάζονται περισσότερο από τα ωράρια δημοσίευσης των ειδησεογραφικών οργανισμών.



Διάγραμμα 13. Ειδησεογραφικές πηγές



Διάγραμμα 14. Πλήθος ειδήσεων ανά ώρα ημέρας

Με βάση τα παραπάνω, τα τρία σύνολα δεδομένων παρουσιάζουν διαφορετικά χαρακτηριστικά ως προς τη δομή, τη συχνότητα και τη χρονική κατανομή τους. Τα tweets παρέχουν μεγάλο όγκο σύντομων και άμεσων αντιδράσεων, τα Reddit posts και comments αποτυπώνουν πιο αναλυτικές συζητήσεις εντός θεματικών κοινοτήτων, ενώ τα ειδησεογραφικά άρθρα παρέχουν πιο επίσημη και δομημένη πληροφόρηση. Για τον λόγο αυτό, η ξεχωριστή διερεύνηση κάθε πηγής κρίθηκε απαραίτητη πριν από την ενοποίηση των δεδομένων και την εφαρμογή των τεχνικών ανάλυσης συναισθήματος.

Μετά τη συλλογή και τον αρχικό καθαρισμό, τα διαφορετικά σύνολα κειμένων ενοποιήθηκαν σε κοινή μορφή. Κάθε εγγραφή αντιστοιχίστηκε σε συγκεκριμένη χρονική στιγμή μέσω της ημερομηνίας δημοσίευσης, ενώ διατηρήθηκε το βασικό κείμενο προς ανάλυση. Συγκεκριμένα, για τα tweets χρησιμοποιήθηκε το κείμενο της δημοσίευσης, για το Reddit χρησιμοποιήθηκε το κείμενο των posts ή των comments, ενώ για τα ειδησεογραφικά άρθρα χρησιμοποιήθηκε ο συνδυασμός τίτλου και περίληψης.

Στη συνέχεια, τα ενοποιημένα κείμενα χρησιμοποιήθηκαν για την εξαγωγή σκορ συναισθήματος (sentiment scores). Τα scores αυτά ομαδοποιήθηκαν χρονικά, ώστε να προκύπτουν συγκεντρωτικά χαρακτηριστικά συναισθήματος ανά χρονική περίοδο. Έπειτα, τα χρονικά ομαδοποιημένα sentiment χαρακτηριστικά συγχωνεύθηκαν με τα δεδομένα τιμών του Bitcoin με βάση το αντίστοιχο χρονικό σημείο (timestamp). Με τον τρόπο αυτό, για κάθε χρονική περίοδο της αγοράς ήταν διαθέσιμη, όπου υπήρχε σχετική πληροφορία, η αντίστοιχη ένδειξη συναισθήματος από κοινωνικά δίκτυα και ειδησεογραφικές πηγές.

Ως αποτέλεσμα, δημιουργήθηκε ένα εμπλουτισμένο σύνολο δεδομένων, το οποίο συνδυάζει αριθμητικά χαρακτηριστικά αγοράς από τις μεταβλητές OHLCV με πληροφορία συναισθήματος από εξωτερικές πηγές κειμένου. Η προσθήκη των sentiment μεταβλητών επιτρέπει την αξιολόγηση των μοντέλων τόσο χωρίς εξωτερική πληροφορία όσο και με ενσωματωμένο sentiment, προκειμένου να εξεταστεί αν η δημόσια συζήτηση και η ειδησεογραφική ροή βελτιώνουν την προβλεπτική ικανότητα των μοντέλων.

5. Αποτελέσματα έρευνας

Στο παρόν κεφάλαιο παρουσιάζονται και αξιολογούνται τα αποτελέσματα των μοντέλων πρόβλεψης της μελλοντικής λογαριθμικής απόδοσης του Bitcoin για χρονικούς ορίζοντες 1, 15 και 30 λεπτών. Η αξιολόγηση πραγματοποιήθηκε τόσο σε επίπεδο λογαριθμικής απόδοσης όσο και σε επίπεδο προβλεπόμενης τιμής κλεισίματος, με βασικές μετρικές το μέσο τετραγωνικό σφάλμα ρίζας (RMSE) και την Ακρίβεια Κατεύθυνσης (Directional Accuracy).

Η παρουσίαση των αποτελεσμάτων οργανώνεται σε δύο βασικά μέρη. Αρχικά εξετάζονται τα μοντέλα που βασίζονται αποκλειστικά στα ιστορικά δεδομένα αγοράς και στα τεχνικά χαρακτηριστικά. Στη συνέχεια παρουσιάζονται τα αντίστοιχα αποτελέσματα μετά την προσθήκη μεταβλητών ανάλυσης συναισθήματος, ώστε να αξιολογηθεί αν η πληροφορία από κοινωνικά δίκτυα και ειδησεογραφικές πηγές προσφέρει πρόσθετη προβλεπτική αξία. Με αυτόν τον τρόπο, η σύγκριση δεν περιορίζεται στην απόδοση μεμονωμένων μοντέλων, αλλά εξετάζει και το κατά πόσο η εξωτερική πληροφορία μπορεί να βελτιώσει τη βραχυχρόνια πρόβλεψη της τιμής του Bitcoin.

5.1 Αποτελέσματα χωρίς χρήση μεταβλητών συναισθήματος

Στην ενότητα που ακολουθεί παρουσιάζονται τα αποτελέσματα των μοντέλων πρόβλεψης χωρίς την ενσωμάτωση ανάλυσης συναισθήματος. Τα αποτελέσματα αυτά λειτουργούν ως βασικό σημείο αναφοράς, με βάση το οποίο θα αξιολογηθεί στη συνέχεια η επίδραση της προσθήκης χαρακτηριστικών sentiment.

Η αξιολόγηση πραγματοποιήθηκε τόσο στο σύνολο επικύρωσης όσο και στο σύνολο ελέγχου, με χρήση μετρικών σε επίπεδο λογαριθμικών αποδόσεων και σε επίπεδο προβλεπόμενης τιμής κλεισίματος.

5.1.1 Πρόβλεψη χρονικού διαστήματος 1 λεπτού

Για το αυτοπαλίνδρομο μοντέλο κινητού μέσου (ARIMA), ο καλύτερος συνδυασμός παραμέτρων που προέκυψε από την αναζήτηση πλέγματος (grid search) είναι το (5,0,4), δηλαδή μοντέλο με 5 αυτοπαλίνδρομους όρους (AR), μηδενική διαφοροποίηση (lags) και 4 όρους κινητού μέσου (MA). Η μηδενική διαφοροποίηση είναι αναμενόμενη, καθώς το μοντέλο εφαρμόστηκε στις

λογαριθμικές αποδόσεις, οι οποίες έχουν ήδη αποδειχθεί στάσιμες. Επομένως, δεν ήταν απαραίτητη επιπλέον διαφοροποίηση της σειράς.

Στη παλινδρόμηση Ridge (Ridge Regression), η τιμή της παραμέτρου κανονικοποίησης α που επιλέχθηκε είναι η 0.0001, δηλαδή η μικρότερη τιμή από το εύρος τιμών που εξετάστηκε. Το αποτέλεσμα αυτό δείχνει ότι η εφαρμογή ισχυρής κανονικοποίησης δεν βελτίωσε την απόδοση του μοντέλου στον χρονικό ορίζοντα του 1 λεπτού. Αντίθετα, το μοντέλο απέδωσε καλύτερα όταν η ποινή στους συντελεστές ήταν πολύ μικρή, γεγονός που υποδηλώνει ότι τα διαθέσιμα χαρακτηριστικά (open, high, low, volume) δεν οδήγησαν σε έντονη υπερπροσαρμογή για το συγκεκριμένο γραμμικό μοντέλο.

Για τα τυχαία δάση (Random forest), ο καλύτερος συνδυασμός παραμέτρων είναι ο `bootstrap=True`, `max_depth=8`, `max_features=sqrt`, `min_samples_leaf=30`, `min_samples_split=20` και `n_estimators=500`. Ο συγκεκριμένος συνδυασμός δείχνει ότι το μοντέλο επωφελήθηκε από τη χρήση μεγάλου αριθμού δέντρων, καθώς τα 500 δέντρα συμβάλλουν στη σταθεροποίηση της τελικής πρόβλεψης μέσω του μέσου όρου των επιμέρους δέντρων. Παράλληλα, το μέγιστο βάθος 8 και οι σχετικά αυξημένες τιμές των `min_samples_leaf` και `min_samples_split` περιορίζουν την πολυπλοκότητα κάθε δέντρου, μειώνοντας τον κίνδυνο υπερπροσαρμογής. Στο XGBoost, οι βέλτιστες παράμετροι δείχνουν παρόμοια λογική περιορισμού της πολυπλοκότητας. Το μοντέλο χρησιμοποίησε μεγάλο αριθμό δέντρων (`n_estimators = 500`), αλλά με μικρό μέγιστο βάθος (`max_depth = 3`) και σχετικά χαμηλό ρυθμό μάθησης (`learning_rate = 0.05`). Επιπλέον, οι τιμές (`subsample = 0.6`) και (`colsample_bytree = 0.6`) σημαίνουν ότι κάθε δέντρο εκπαιδεύτηκε με μέρος των παρατηρήσεων και των χαρακτηριστικών, εισάγοντας τυχαιότητα στη διαδικασία εκπαίδευσης. Η επιλογή αυτή λειτουργεί ως μορφή κανονικοποίησης και περιορίζει την εξάρτηση του μοντέλου από συγκεκριμένες παρατηρήσεις ή μεταβλητές. Τέλος, η επιλογή `reg_alpha=0.0` δείχνει ότι δεν εφαρμόστηκε L1 κανονικοποίηση, ενώ η τιμή `reg_lambda=1.0` αντιστοιχεί σε ήπια λ_2 κανονικοποίηση.

Για το δίκτυο μακράς βραχυπρόθεσμης μνήμης (LSTM), Ο καλύτερος συνδυασμός παραμέτρων είναι `lookback=30`, `units=64`, `dropout=0.30`, `batch_size=128` και `learning_rate=0.0005`. Η επιλογή `lookback=30` δείχνει ότι το μοντέλο αξιοποίησε τις προηγούμενες 30 παρατηρήσεις, δηλαδή πληροφορία από τα τελευταία 30 λεπτά, για την πρόβλεψη της μελλοντικής λογαριθμικής απόδοσης. Παράλληλα, η χρήση 64 LSTM units επιτρέπει στο μοντέλο να αποτυπώσει χρονικές εξαρτήσεις στη σειρά, ενώ το σχετικά υψηλό dropout 0.30 λειτουργεί ως μηχανισμός κανονικοποίησης, περιορίζοντας τον κίνδυνο υπερπροσαρμογής. Αντίθετα, στο δίκτυο

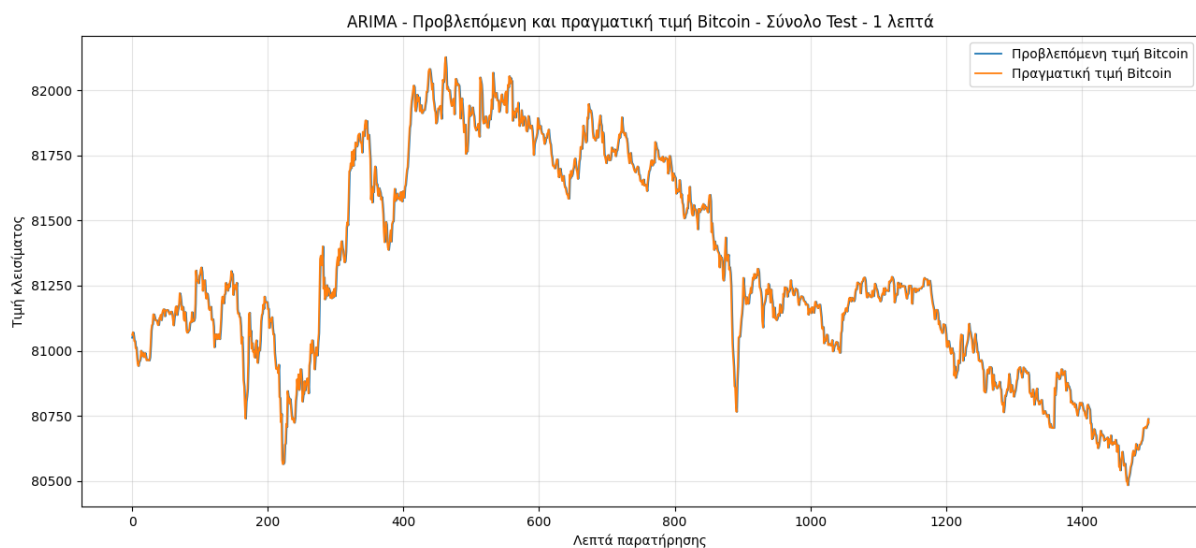
επαναλαμβανόμενης μονάδας με πύλες (GRU) επιλέχθηκε μικρότερο παράθυρο εισόδου, (lookback = 15), αλλά μεγαλύτερος αριθμός μονάδων, (units = 128). Αυτό υποδηλώνει ότι το GRU απέδωσε καλύτερα όταν εστίασε σε πιο πρόσφατη βραχυχρόνια πληροφορία, αξιοποιώντας παράλληλα μεγαλύτερη εσωτερική αναπαραστατική ικανότητα.

Πίνακας 6. Αποτελέσματα μοντέλων χωρίς ανάλυση συναισθήματος για χρονικό ορίζοντα 1 λεπτού

Μοντέλο	Σύνολο δεδομένων	RMSE λογαριθμικής απόδοσης	Ακρίβεια κατεύθυνσης	RMSE τιμής κλεισίματος
ARIMA	Validation	0.052%	0.463	38.196
ARIMA	Test	0.062%	0.486	41.067
Ridge Regression	Validation	0.052%	0.463	38.196
Ridge Regression	Test	0.062%	0.484	41.067
Random Forest	Validation	0.052%	0.465	38.205
Random Forest	Test	0.070%	0.484	45.338
XGBoost	Validation	0.052%	0.466	38.199
XGBoost	Test	0.062%	0.483	41.081
LSTM	Validation	0.052%	0.471	38.405
LSTM	Test	0.062%	0.484	41.138
GRU	Validation	0.053%	0.463	38.734
GRU	Test	0.063%	0.484	41.521

Το χαμηλότερο μέσο τετραγωνικό σφάλμα ρίζας της τιμής κλεισίματος στο σύνολο ελέγχου εμφανίζεται από το ARIMA, με τιμή 41.067. Πολύ κοντά ακολουθούν το Ridge Regression και το XGBoost, με Price RMSE 41.081 και 41.067 αντίστοιχα. Οι διαφορές μεταξύ αυτών των μοντέλων είναι εξαιρετικά μικρές, γεγονός που δείχνει ότι δεν υπάρχει σαφής υπεροχή ενός μοντέλου ως προς το αριθμητικό σφάλμα πρόβλεψης στον ορίζοντα του 1 λεπτού.

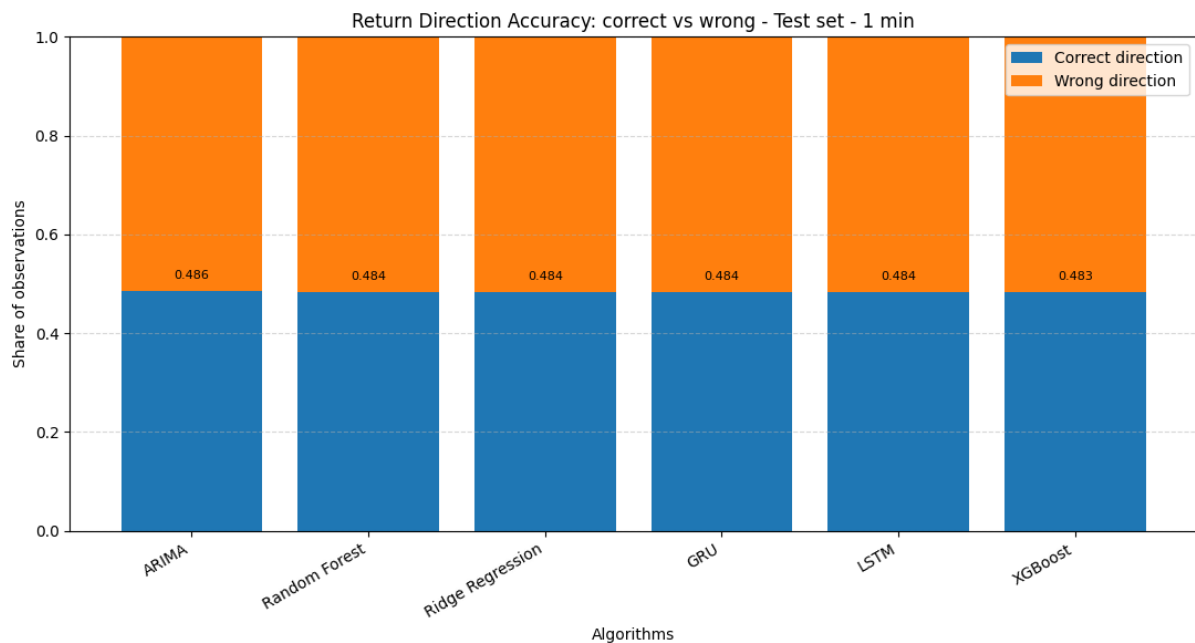
Στο Διάγραμμα 15 παρουσιάζεται η σύγκριση της προβλεπόμενης και της πραγματικής τιμής κλεισίματος του Bitcoin στο σύνολο ελέγχου για το συγκεκριμένο μοντέλο. Οι δύο χρονοσειρές εμφανίζονται σχεδόν ταυτισμένες, γεγονός που δείχνει ότι το ARIMA ακολουθεί πολύ κοντά το επίπεδο της πραγματικής τιμής. Η εικόνα αυτή είναι αναμενόμενη σε δεδομένα υψηλής συχνότητας, καθώς οι μεταβολές της τιμής από λεπτό σε λεπτό είναι συνήθως μικρές σε σχέση με το συνολικό επίπεδο της τιμής. Για τον λόγο αυτό, ακόμη και σχετικά απλά μοντέλα μπορούν να επιτύχουν χαμηλό σφάλμα σε επίπεδο τιμής, χωρίς αυτό να σημαίνει απαραίτητα ότι εντοπίζουν με ακρίβεια τη βραχυχρόνια δυναμική της αγοράς.



Διάγραμμα 15. Γραφική παράσταση προβλεπόμενης και πραγματικής τιμής BTC για χρονικό ορίζοντα 1 λεπτού για το ARIMA

Η ερμηνεία αυτή ενισχύεται από την Ακρίβεια Κατεύθυνσης (Directional Accuracy). Όπως φαίνεται στο Διάγραμμα 16, κανένα μοντέλο δεν υπερβαίνει το 50% στο test set. Το υψηλότερο Directional Accuracy εμφανίζεται από το ARIMA, με τιμή 48,6%, ενώ τα υπόλοιπα μοντέλα κινούνται πολύ κοντά σε αυτό το επίπεδο. Το αποτέλεσμα αυτό δείχνει ότι, παρότι τα μοντέλα προσεγγίζουν ικανοποιητικά την τιμή κλεισίματος, δεν μπορούν να προβλέψουν συστηματικά αν η επόμενη μεταβολή θα είναι θετική ή αρνητική.

Η διαφορά αυτή μεταξύ χαμηλού RMSE και χαμηλού Directional Accuracy είναι σημαντική. Σε ορίζοντα 1 λεπτού, η προβλεπόμενη τιμή μπορεί να είναι πολύ κοντά στην πραγματική τιμή, αλλά ακόμη και μία πολύ μικρή απόκλιση στη λογαριθμική απόδοση γύρω από το μηδέν μπορεί να αλλάξει το πρόσημο της πρόβλεψης. Συνεπώς, το χαμηλό αριθμητικό σφάλμα δεν συνεπάγεται απαραίτητα επιτυχημένη πρόβλεψη της κατεύθυνσης.



Διάγραμμα 16. Ακρίβεια πρόβλεψης κατεύθυνσης ανά μοντέλο για χρονικό ορίζοντα 1 λεπτού

Συνολικά, στον ορίζοντα του 1 λεπτού τα μοντέλα εμφανίζουν πολύ παρόμοια συμπεριφορά ως προς το RMSE, ενώ η πρόβλεψη της κατεύθυνσης παραμένει χαμηλότερη από το επίπεδο του 50%. Επομένως, δεν προκύπτει σαφής υπεροχή κάποιου πιο σύνθετου μοντέλου έναντι των απλούστερων προσεγγίσεων.

5.1.2 Πρόβλεψη χρονικού διαστήματος 15 λεπτών

Για το αυτοπαλίνδρομο μοντέλο κινητού μέσου (ARIMA), ο καλύτερος συνδυασμός παραμέτρων που προέκυψε μέσω της αναζήτησης πλέγματος είναι το (4,0,2), δηλαδή μοντέλο με 4 αυτοπαλίνδρομους όρους (AR), μηδενική διαφοροποίηση (lags) και 2 όρους κινητού μέσου (MA).

Στη παλινδρόμηση Ridge, η τιμή της παραμέτρου κανονικοποίησης (α) που επιλέχθηκε ήταν 10000.0. Σε αντίθεση με τον χρονικό ορίζοντα του 1 λεπτού, όπου επιλέχθηκε πολύ μικρή τιμή κανονικοποίησης, στον ορίζοντα των 15 λεπτών το μοντέλο απέδωσε καλύτερα με ισχυρή λ_2 κανονικοποίηση. Το αποτέλεσμα αυτό δείχνει ότι, για μεγαλύτερο χρονικό ορίζοντα, η έντονη

συρρίκνωση των συντελεστών βοήθησε στη μείωση της υπερπροσαρμογής και στη σταθεροποίηση της πρόβλεψης.

Για τη παλινδρόμηση τυχαίων δασών (Random Forest), ο καλύτερος συνδυασμός παραμέτρων ήταν `bootstrap=True`, `max_depth=4`, `max_features=sqrt`, `min_samples_leaf=10`, `min_samples_split=40` και `n_estimators=500`. Ο μεγάλος αριθμός δέντρων συμβάλλει στη σταθεροποίηση της τελικής πρόβλεψης, ενώ το μικρό μέγιστο βάθος περιορίζει σημαντικά την πολυπλοκότητα κάθε δέντρου. Παράλληλα, η αυξημένη τιμή του `min_samples_split` δείχνει ότι το μοντέλο απέδωσε καλύτερα όταν αποφεύχθηκαν πολύ λεπτομερείς διαχωρισμοί, οι οποίοι θα μπορούσαν να οδηγήσουν σε προσαρμογή στον θόρυβο των δεδομένων. Στο XGBoost, οι βέλτιστες παράμετροι ήταν `colsample_bytree=0.6`, `learning_rate=0.05`, `max_depth=3`, `n_estimators=500`, `reg_alpha=0.0`, `reg_lambda=1.0` και `subsample=0.6`. Ο συγκεκριμένος συνδυασμός δείχνει ότι το μοντέλο απέδωσε καλύτερα με μεγάλο αριθμό σχετικά απλών δέντρων και χαμηλό ρυθμό μάθησης. Οι τιμές `subsample=0.6` και `colsample_bytree=0.6` σημαίνουν ότι κάθε δέντρο εκπαιδεύτηκε με μέρος των παρατηρήσεων και των χαρακτηριστικών, εισάγοντας τυχαιότητα στη διαδικασία εκπαίδευσης. Η επιλογή αυτή λειτουργεί ως μηχανισμός κανονικοποίησης και μειώνει την εξάρτηση του μοντέλου από συγκεκριμένες παρατηρήσεις ή μεταβλητές.

Για το δίκτυο μακράς βραχυπρόθεσμης μνήμης (LSTM), ο καλύτερος συνδυασμός παραμέτρων ήταν `lookback=60`, `units=64`, `dropout=0.30`, `batch_size=128` και `learning_rate=0.0005`. Η επιλογή `lookback=60` δείχνει ότι το μοντέλο αξιοποίησε τις προηγούμενες 60 παρατηρήσεις, δηλαδή πληροφορία από την τελευταία μία ώρα, για την πρόβλεψη της λογαριθμικής απόδοσης σε ορίζοντα 15 λεπτών. Για το δίκτυο επαναλαμβανόμενης μονάδας με πύλες (GRU), ο καλύτερος συνδυασμός παραμέτρων ήταν `lookback=30`, `units=128`, `dropout=0.20`, `batch_size=128` και `learning_rate=0.001`. Το GRU αξιοποίησε μικρότερο παράθυρο εισόδου σε σχέση με το LSTM, δηλαδή τις προηγούμενες 30 παρατηρήσεις, αλλά με μεγαλύτερο αριθμό μονάδων. Αυτό υποδηλώνει ότι το GRU απέδωσε καλύτερα όταν βασίστηκε σε πιο πρόσφατη βραχυχρόνια πληροφορία, διατηρώντας παράλληλα μεγαλύτερη εσωτερική αναπαραστατική ικανότητα.

Στον Πίνακα 7 παρουσιάζονται συγκεντρωτικά τα αποτελέσματα όλων των μοντέλων για τον χρονικό ορίζοντα 15 λεπτών.

Πίνακας 7. Αποτελέσματα μοντέλων χωρίς ανάλυση συναισθήματος για χρονικό ορίζοντα 15 λεπτών

Μοντέλο	Σύνολο δεδομένων	RMSE λογαριθμικής απόδοσης	Ακρίβεια κατεύθυνσης	RMSE τιμής κλεισίματος
ARIMA	Validation	0.202%	0.493	147.995
ARIMA	Test	0.234%	0.511	155.037
Ridge Regression	Validation	0.202%	0.503	147.982
Ridge Regression	Test	0.235%	0.488	155.094
Random Forest	Validation	0.202%	0.500	147.991
Random Forest	Test	0.621%	0.504	383.939
XGBoost	Validation	0.202%	0.497	148.070
XGBoost	Test	0.237%	0.499	156.413
LSTM	Validation	0.202%	0.503	147.880
LSTM	Test	0.235%	0.487	155.141
GRU	Validation	0.230%	0.493	168.654
GRU	Test	0.235%	0.511	155.393

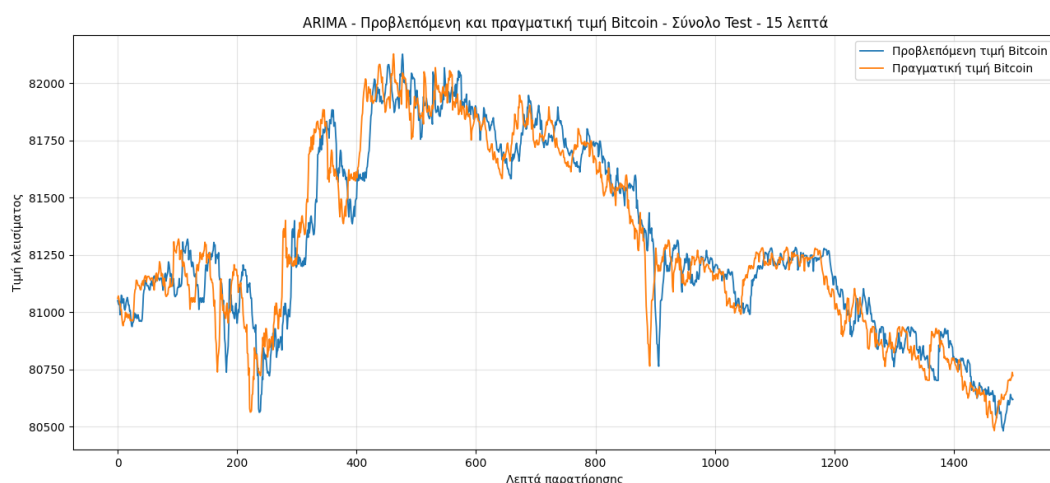
Σε σχέση με τον ορίζοντα του 1 λεπτού, παρατηρείται αύξηση του RMSE της λογαριθμικής απόδοσης και διαμορφώνεται περίπου στο 0,218%–0,240% για τα περισσότερα μοντέλα, με εξαίρεση το Random Forest, το οποίο εμφανίζει αισθητά υψηλότερο σφάλμα στο σύνολο ελέγχου.

Αντίστοιχα, το μέσο τετραγωνικό σφάλμα ρίζας της τιμής αυξάνεται σε περίπου 150-160 μονάδες. Η αύξηση αυτή είναι αναμενόμενη, καθώς όσο μεγαλώνει ο χρονικός ορίζοντας πρόβλεψης, αυξάνεται και η αβεβαιότητα σχετικά με τη μελλοντική μεταβολή της τιμής.

Ως προς το σφάλμα τιμής, το χαμηλότερο μέσο τετραγωνικό σφάλμα ρίζας της τιμής στο σύνολο ελέγχου εμφανίζεται από το ARIMA, με τιμή 155.037. Πολύ κοντά ακολουθεί το Ridge Regression, με σφάλμα 155.094, ενώ το XGBoost και το LSTM εμφανίζουν επίσης συγκρίσιμες επιδόσεις. Αντίθετα, το Random Forest παρουσιάζει σημαντικά υψηλότερο σφάλμα, γεγονός που δείχνει ότι η καλή συμπεριφορά του στο validation set δεν μεταφέρθηκε με την ίδια επιτυχία στο test set.

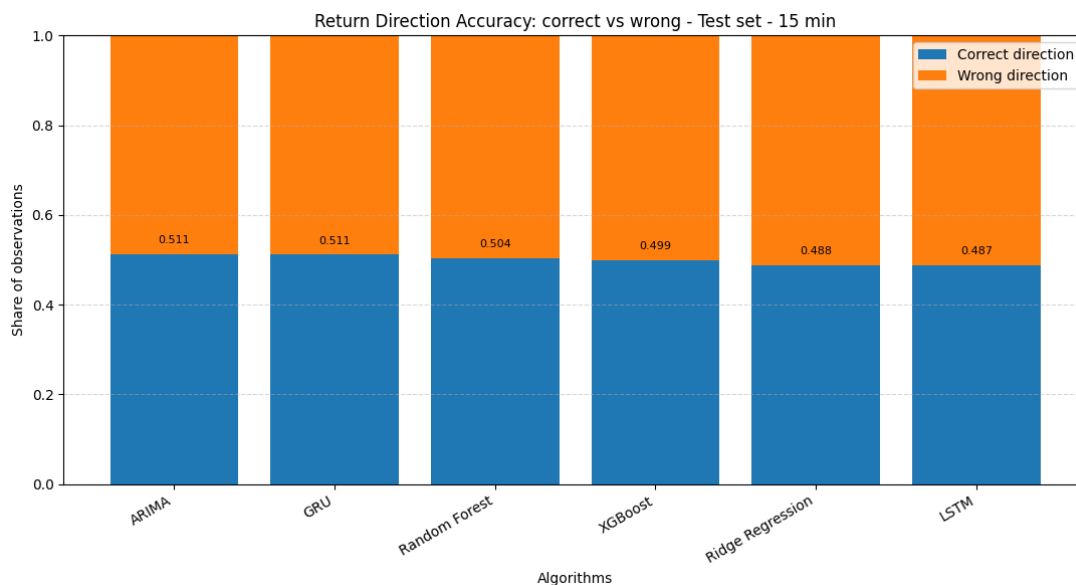
Ως προς την Ακρίβεια Κατεύθυνσης, το ARIMA και το GRU εμφανίζουν τις υψηλότερες τιμές στο test set, με αποτέλεσμα περίπου 51,1%. Παρότι οι τιμές αυτές υπερβαίνουν οριακά το 50%, η διαφορά είναι μικρή και δεν επαρκεί για να υποστηριχθεί ισχυρή ικανότητα πρόβλεψης ως προς τη φορά της μεταβολής. Ωστόσο, σε σχέση με τον ορίζοντα του 1 λεπτού, παρατηρείται ότι ο μεγαλύτερος χρονικός ορίζοντας επιτρέπει σε ορισμένα μοντέλα να αποδώσουν ελαφρώς καλύτερα ως προς την κατεύθυνση.

Τα αποτελέσματα για τον χρονικό ορίζοντα των 15 λεπτών δείχνουν ότι το ARIMA εμφανίζει την καλύτερη συνολική εικόνα, καθώς συνδυάζει το χαμηλότερο μέσο τετραγωνικό σφάλμα ρίζας (RMSE) της τιμής με την υψηλότερη ακρίβεια κατεύθυνσης στο σύνολο ελέγχου. Επομένως, με βάση το βασικό κριτήριο σφάλματος τιμής, το ARIMA επιλέγεται ως το καλύτερο μοντέλο για τον χρονικό ορίζοντα των 15 λεπτών.



Διάγραμμα 17. Γραφική παράσταση προβλεπόμενης και πραγματικής τιμής BTC για χρονικό ορίζοντα 15 λεπτών για το ARIMA

Τα αποτελέσματα για τον χρονικό ορίζοντα των 15 λεπτών δείχνουν ότι η αύξηση του ορίζοντα πρόβλεψης οδηγεί σε μεγαλύτερο αριθμητικό σφάλμα, αλλά η ακρίβεια κατεύθυνσης πλησιάζει περισσότερο το επίπεδο του 50%. Στο Διάγραμμα 18 παρουσιάζεται συγκριτικά η ακρίβεια πρόβλεψης κατεύθυνσης όλων των μοντέλων στο test set για χρονικό ορίζοντα 15 λεπτών. Παρατηρείται ότι ορισμένα μοντέλα υπερβαίνουν οριακά το επίπεδο του 50%, με το ARIMA να εμφανίζει την υψηλότερη τιμή. Ωστόσο, οι διαφορές παραμένουν περιορισμένες, γεγονός που δείχνει ότι η πρόβλεψη της κατεύθυνσης εξακολουθεί να είναι δύσκολη.



Διάγραμμα 18. Ακρίβεια πρόβλεψης κατεύθυνσης ανά μοντέλο για χρονικό ορίζοντα 15 λεπτών

5.1.3 Πρόβλεψη χρονικού διαστήματος 30 λεπτών

Το αυτοπαλίνδρομο μοντέλο κινητού μέσου, είναι το $ARIMA(3,0,6)$, δηλαδή μοντέλο με τρεις αυτοπαλίνδρομους όρους, μηδενική διαφοροποίηση και έξι όρους κινητού μέσου.

Στη παλινδρόμηση Ridge, η τιμή της παραμέτρου κανονικοποίησης (α) που επιλέχθηκε ήταν 10000.0. Όπως και στον χρονικό ορίζοντα των 15 λεπτών, η υψηλή τιμή του (α) δείχνει ότι το μοντέλο απέδωσε καλύτερα με ισχυρή λ_2 κανονικοποίηση. Το αποτέλεσμα αυτό υποδηλώνει ότι, όσο αυξάνεται ο χρονικός ορίζοντας πρόβλεψης, η συρρίκνωση των συντελεστών συμβάλλει στη σταθεροποίηση του γραμμικού μοντέλου και στη μείωση της πιθανής υπερπροσαρμογής.

Για το Random Forest, ο καλύτερος συνδυασμός παραμέτρων ήταν `bootstrap=True`, `max_depth=4`, `max_features=sqrt`, `min_samples_leaf=10`, `min_samples_split=20` και `n_estimators=300`. Σε σχέση με τον ορίζοντα του 1 λεπτού, το μοντέλο επέλεξε μικρότερο μέγιστο βάθος, γεγονός που δείχνει

ότι η περιορισμένη πολυπλοκότητα των δέντρων ήταν πιο κατάλληλη για τον μεγαλύτερο χρονικό ορίζοντα. Παράλληλα, η χρήση 300 δέντρων εξακολουθεί να παρέχει σταθεροποίηση της τελικής πρόβλεψης μέσω του μέσου όρου των επιμέρους δέντρων. Στο XGBoost, οι βέλτιστες παράμετροι ήταν `colsample_bytree=0.7`, `learning_rate=0.01`, `max_depth=2`, `n_estimators=300`, `reg_alpha=0.0`, `reg_lambda=1.0` και `subsample=1.0`. Ο συγκεκριμένος συνδυασμός δείχνει ότι το μοντέλο απέδωσε καλύτερα με πολύ χαμηλό ρυθμό μάθησης και ιδιαίτερα ρηχά δέντρα. Το `max_depth=2` περιορίζει σημαντικά την πολυπλοκότητα κάθε δέντρου, ενώ το `learning_rate=0.01` οδηγεί σε πιο σταδιακή μάθηση. Η τιμή `reg_lambda=1.0` αντιστοιχεί σε ήπια λ_2 κανονικοποίηση, ενώ η τιμή `reg_alpha=0.0` δείχνει ότι δεν εφαρμόστηκε L1 κανονικοποίηση.

Για το δίκτυο μακράς βραχυπρόθεσμης μνήμης (LSTM), ο καλύτερος συνδυασμός παραμέτρων ήταν `lookback=15`, `units=32`, `dropout=0.10`, `batch_size=64` και `learning_rate=0.001`. Σε αντίθεση με τον ορίζοντα των 15 λεπτών, όπου επιλέχθηκε μεγαλύτερο παράθυρο εισόδου, στον ορίζοντα των 30 λεπτών το LSTM απέδωσε καλύτερα αξιοποιώντας τις προηγούμενες 15 παρατηρήσεις. Η επιλογή μικρότερου αριθμού μονάδων και χαμηλότερου dropout δείχνει ότι ένα πιο απλό ακολουθιακό μοντέλο ήταν επαρκές για τον συγκεκριμένο ορίζοντα. Τέλος, για το δίκτυο επαναλαμβανόμενης μονάδας με πύλες (GRU), ο καλύτερος συνδυασμός παραμέτρων ήταν `lookback=15`, `units=32`, `dropout=0.10`, `batch_size=64` και `learning_rate=0.001`. Η επιλογή αυτή είναι ίδια με εκείνη του LSTM για τον ορίζοντα των 30 λεπτών, γεγονός που δείχνει ότι και τα δύο ακολουθιακά μοντέλα απέδωσαν καλύτερα όταν βασίστηκαν σε μικρότερο παράθυρο πρόσφατης πληροφορίας και σε πιο περιορισμένη εσωτερική αρχιτεκτονική. Το αποτέλεσμα αυτό υποδηλώνει ότι η αύξηση της πολυπλοκότητας δεν οδήγησε σε καλύτερη απόδοση στον συγκεκριμένο χρονικό ορίζοντα. Στον Πίνακα 8 παρουσιάζονται συγκεντρωτικά τα αποτελέσματα όλων των μοντέλων για τον χρονικό ορίζοντα των 30 λεπτών

Πίνακας 8. Αποτελέσματα μοντέλων χωρίς ανάλυση συναισθήματος για χρονικό ορίζοντα 30 λεπτών

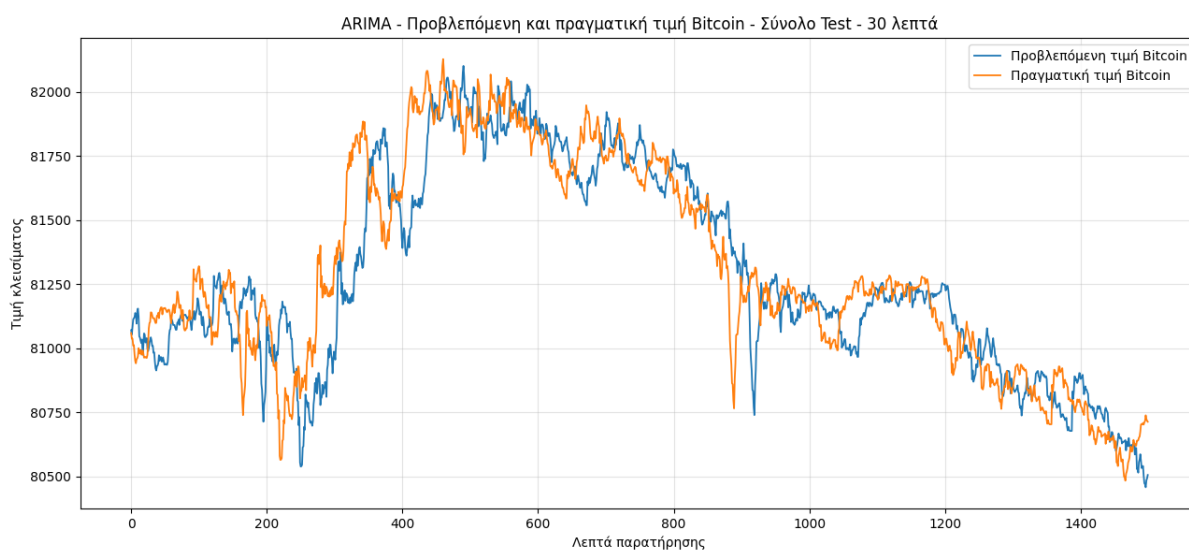
Μοντέλο	Σύνολο δεδομένων	RMSE λογαριθμικής απόδοσης	Ακρίβεια κατεύθυνσης	RMSE τιμής κλεισίματος
ARIMA	Validation	0.280%	0.488	205.620
ARIMA	Test	0.331%	0.511	218.307
Ridge Regression	Validation	0.280%	0.507	205.693
Ridge Regression	Test	0.331%	0.488	218.230
Random Forest	Validation	0.280%	0.493	205.723
Random Forest	Test	1.174%	0.505	728.847
XGBoost	Validation	0.281%	0.494	205.880
XGBoost	Test	0.337%	0.503	221.760
LSTM	Validation	0.296%	0.488	217.391
LSTM	Test	0.334%	0.511	220.486
GRU	Validation	0.296%	0.488	217.390
GRU	Test	0.334%	0.511	220.486

Σε σχέση με τους ορίζοντες του 1 και των 15 λεπτών, παρατηρείται περαιτέρω αύξηση του RMSE της λογαριθμικής απόδοσης, το οποίο στο test set διαμορφώνεται περίπου στο 0,300% - 0,330%. Αντίστοιχα, το η ρίζα του μέσου τετραγωνικού σφάλματος της τιμής αυξάνεται σε περίπου 220

μονάδες, με εξαίρεση το random forest. Η αύξηση αυτή παραμένει αναμενόμενη, καθώς όσο μεγαλώνει ο χρονικός ορίζοντας πρόβλεψης, αυξάνεται και η αβεβαιότητα σχετικά με τη μελλοντική εξέλιξη της τιμής.

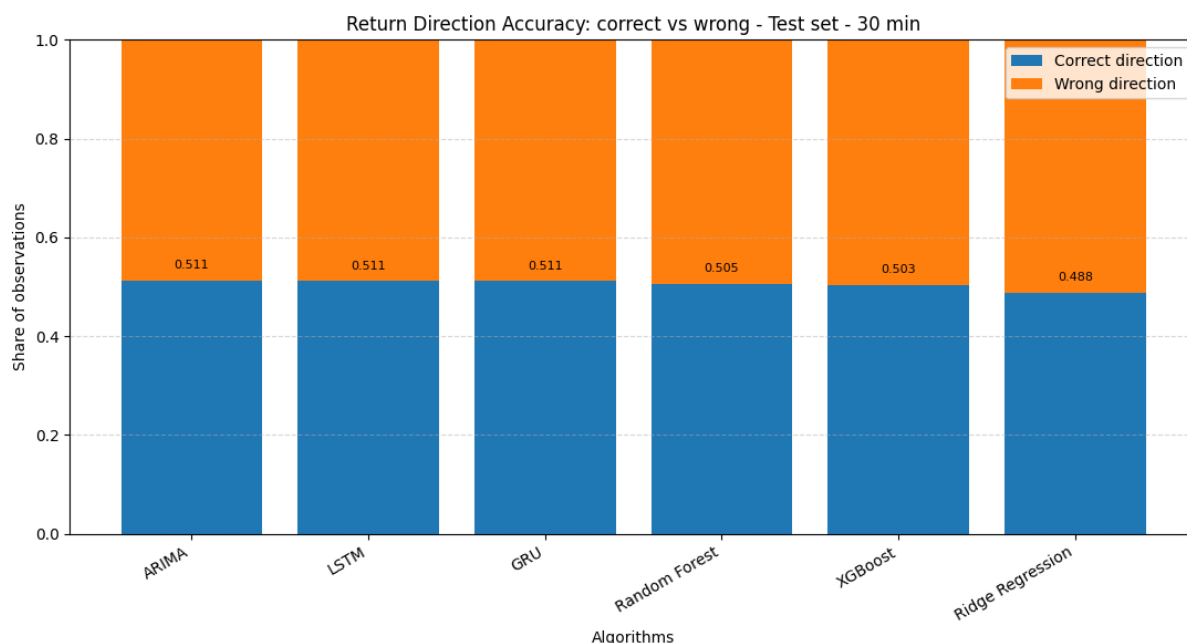
Ως προς το σφάλμα τιμής, το χαμηλότερο Price RMSE στο test set εμφανίζεται από το Ridge Regression, με τιμή 218,230. Πολύ κοντά ακολουθεί το ARIMA, με Price RMSE 218.307, ενώ τα δίκτυα βαθιάς μάθησης LSTM και GRU εμφανίζουν ελαφρώς υψηλότερη τιμή, ίση με 220.486. Οι διαφορές μεταξύ Ridge Regression και ARIMA είναι εξαιρετικά μικρές, επομένως δεν προκύπτει καθαρή υπεροχή ενός μοντέλου ως προς το αριθμητικό σφάλμα πρόβλεψης.

Σε επίπεδο Directional Accuracy, το ARIMA μαζί με τα LSTM και GRU εμφανίζουν την υψηλότερη τιμή στο test set, με 51,1%, ενώ ακολουθεί το XGBoost με 50,3%. Με βάση το βασικό κριτήριο σφάλματος τιμής, το ARIMA επιλέγεται ως το καλύτερο μοντέλο για τον χρονικό ορίζοντα των 30 λεπτών, καθώς η πολυπλοκότητα των υπόλοιπων μοντέλων δεν οδήγησε σε ουσιαστική βελτίωση της προβλεπτικής ικανότητάς τους.



Διάγραμμα 19. Γραφική παράσταση προβλεπόμενης και πραγματικής τιμής BTC για χρονικό ορίζοντα 30 λεπτών για το ARIMA

Στο Διάγραμμα 20 παρουσιάζεται συγκριτικά η ακρίβεια πρόβλεψης κατεύθυνσης όλων των μοντέλων στο test set για χρονικό ορίζοντα 30 λεπτών. Παρατηρείται ότι τα ARIMA, LSTM και GRU υπερβαίνουν οριακά το 50%, ενώ τα υπόλοιπα μοντέλα κινούνται ελαφρώς χαμηλότερα. Παρόλα αυτά, οι τιμές παραμένουν κοντά στο επίπεδο μιας ισορροπημένης τυχαίας πρόβλεψης.



Διάγραμμα 20. Ακρίβεια πρόβλεψης κατεύθυνσης ανά μοντέλο για χρονικό ορίζοντα 30 λεπτών

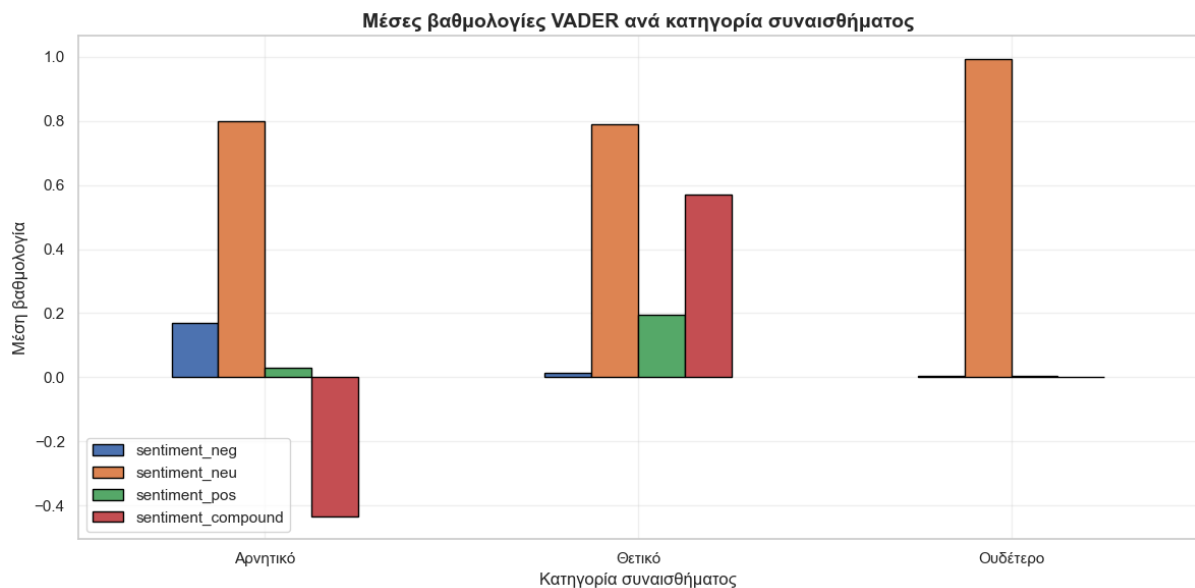
5.2 Αποτελέσματα ανάλυσης συναισθήματος

Μετά την εφαρμογή της ανάλυσης συναισθήματος στο ενοποιημένο σύνολο δεδομένων, προέκυψαν συνολικά 4.706.819 εγγραφές με διαθέσιμες βαθμολογίες συναισθήματος. Η κατανομή των κατηγοριών συναισθήματος έδειξε ότι το μεγαλύτερο ποσοστό των κειμένων ταξινομήθηκε ως θετικό. Συγκεκριμένα, οι θετικές εγγραφές ήταν 2.319.809, ποσοστό 49,29% του συνόλου. Οι ουδέτερες εγγραφές ανήλθαν σε 1.736.859, δηλαδή 36,90%, ενώ οι αρνητικές εγγραφές ήταν 650.151, ποσοστό 13,81%. Επομένως, το περιεχόμενο που σχετίζεται με το Bitcoin εμφανίζει συνολικά περισσότερο θετικό και ουδέτερο προσανατολισμό, ενώ το αρνητικό συναίσθημα αντιπροσωπεύει μικρότερο μέρος του δείγματος.

Πίνακας 9. Κατανομή κατηγοριών συναισθήματος

Κατηγορία συναισθήματος	Πλήθος γραμμών	Ποσοστό
Θετικό	2.319.809	49,29%
Ουδέτερο	1.736.859	36,90%
Αρνητικό	650.151	13,81%

Η μέση τιμή του sentiment_compound ανέρχεται σε 0,2205, γεγονός που υποδηλώνει οριακά θετική συνολική τάση. Ωστόσο, η διάμεσος ήταν 0,0000, κάτι που δείχνει ότι μεγάλο μέρος των κειμένων είχε ουδέτερο ή ασθενές συναισθηματικό περιεχόμενο. Παράλληλα, η τυπική απόκλιση του sentiment_compound ήταν 0,4144, ένδειξη σημαντικής διακύμανσης μεταξύ των επιμέρους εγγραφών.



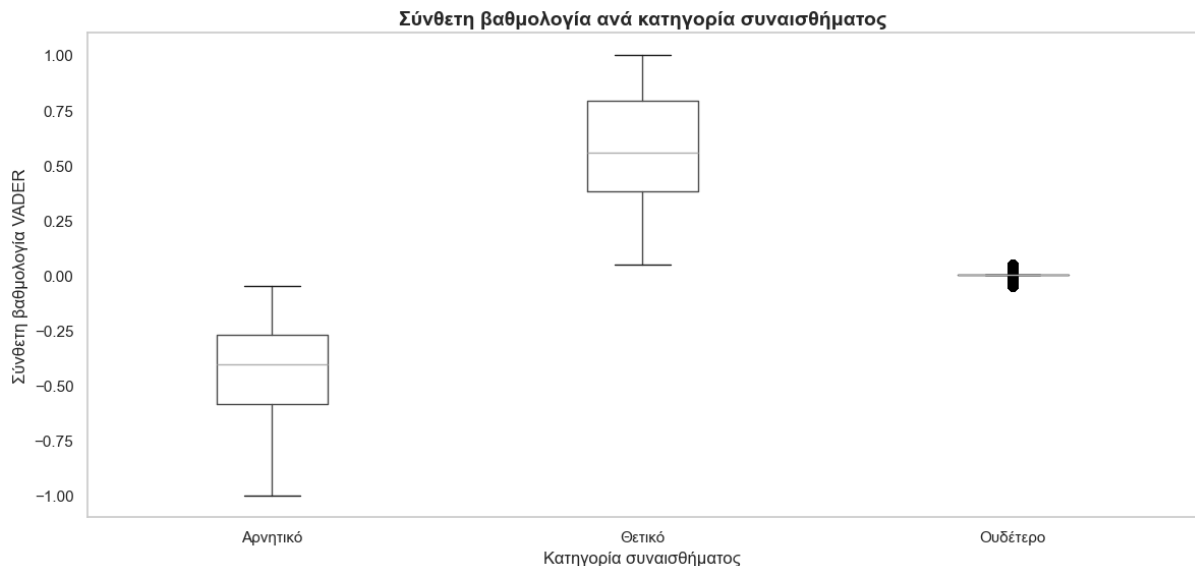
Διάγραμμα 21. Μέσες τιμές VADER scores ανά κατηγορία συναισθήματος

Η ανάλυση ανά πλατφόρμα ανέδειξε διαφοροποιήσεις μεταξύ των πηγών. Το Twitter παρουσίασε τη μεγαλύτερη μέση τιμή sentiment_compound, ίση με 0,2208, ενώ το Reddit είχε χαμηλότερη μέση τιμή, ίση με 0,1562. Αντίθετα, τα ειδησεογραφικά άρθρα είχαν μέση τιμή 0,0488, γεγονός που δείχνει ότι το ύφος τους ήταν σαφώς πιο ουδέτερο σε σχέση με τα κοινωνικά δίκτυα.

Πίνακας 10. Αποτελέσματα sentiment ανά πλατφόρμα

Πλατφόρμα	Μέσο compound
News	0,0488
Reddit	0,1562
Twitter	0,2208

Τα παραπάνω αποτελέσματα δείχνουν ότι το Twitter εμφανίζει την εντονότερη θετική κλίση. Το Reddit παρουσιάζει επίσης θετική μέση τιμή, αλλά με μεγαλύτερη σχετική παρουσία αρνητικών εγγραφών σε σχέση με το Twitter. Αντίθετα, οι ειδησεογραφικές πηγές εμφανίζονται πιο ισορροπημένες, με χαμηλότερο μέσο compound και πιο ουδέτερο ύφος, στοιχείο που είναι συμβατό με τον πληροφοριακό χαρακτήρα των ειδησεογραφικών άρθρων.



Διάγραμμα 22. Κατανομή compound score ανά κατηγορία συναισθήματος

Τέλος, εξετάστηκε η μέση τιμή του sentiment_compound ανά πλατφόρμα. Όπως φαίνεται στο box plot που προηγείται, το Twitter παρουσιάζει την υψηλότερη μέση τιμή, περίπου 0,22, ακολουθούμενο από το Reddit με μέση τιμή περίπου 0,16. Αντίθετα, οι ειδησεογραφικές πηγές εμφανίζουν αρκετά χαμηλότερη μέση τιμή, περίπου 0,05. Η διαφοροποίηση αυτή δείχνει ότι το sentiment δεν είναι ομοιόμορφο μεταξύ των πηγών. Τα κοινωνικά δίκτυα, και ιδιαίτερα το Twitter, εμφανίζουν πιο θετικό μέσο συναισθήμα, ενώ οι ειδησεογραφικές πηγές παραμένουν πιο κοντά σε ουδέτερο τόνο.

Στον παρακάτω πίνακα παρουσιάζονται ενδεικτικά παραδείγματα κειμένων που ταξινομήθηκαν σε θετική, ουδέτερη και αρνητική κατηγορία συναισθήματος. Για κάθε παράδειγμα δίνεται η αντίστοιχη βαθμολογία συναισθήματος, όπως υπολογίστηκε από το VADER.

Πίνακας 11. Ενδεικτικά παραδείγματα θετικού, ουδέτερου και αρνητικού συναισθήματος βάσει VADER

Κατηγορία συναισθήματος	Βαθμολογία	Κείμενο
Θετική	0.4588	'#Bitcoin will benefit from these regulatory developments. @saylor DON'T SELL YOUR #BITCOIN
Ουδέτερη	0,0000	#BTC is now at \$35138.34
Αρνητική	-0.8228	The reason why bitcoin maxi's hate any other #crypto is because they could be much more prosperous right now if only \$btc were alone in crypto space

Όπως φαίνεται, η θετική κατηγορία αντιστοιχεί σε κείμενο με ευνοϊκή στάση απέναντι στο Bitcoin, η ουδέτερη σε καθαρά πληροφοριακή αναφορά τιμής, ενώ η αρνητική σε κείμενο με επικριτικό ή αρνητικά φορτισμένο περιεχόμενο.

5.3 Αποτελέσματα με χρήση μεταβλητών συναισθήματος

Στην ενότητα που ακολουθεί παρουσιάζονται τα αποτελέσματα των μοντέλων πρόβλεψης μετά την ενσωμάτωση των μεταβλητών συναισθήματος που αναλύθηκαν προηγουμένως. Η αξιολόγηση πραγματοποιήθηκε για τους ίδιους χρονικούς ορίζοντες με τα προηγούμενα αποτελέσματα 1, 15 και 30 λεπτών, ώστε να εξεταστεί αν η πληροφορία που προέρχεται από τα δεδομένα συναισθήματος μπορεί να βελτιώσει την πρόβλεψη της μελλοντικής λογαριθμικής απόδοσης και της αντίστοιχης τιμής κλεισίματος του Bitcoin. Η αξιολόγηση πραγματοποιήθηκε τόσο στο σύνολο

επικύρωσης όσο και στο ελέγχου με χρήση της ρίζας του μέσου τετραγωνικού σφάλματος (RMSE) της λογαριθμικής απόδοσης, της Ακρίβειας Κατεύθυνσης (Directional Accuracy) και του RMSE της τιμής.

Η αναζήτηση πλέγματος στις παραμέτρους κάθε μοντέλου για κάθε χρονικό ορίζοντα, πραγματοποιήθηκε από την αρχή με τη προσθήκη αποκλειστικά των μεταβλητών συναισθήματος. Επιπλέον, το μοντέλο ARIMA δεν εφαρμόστηκε με χρήση μεταβλητών συναισθήματος οδηγώντας σε ένα μοντέλο π.χ. ARIMAX παρά μόνο με τη τιμή του close, έτσι δεν θα σχολιαστεί στη παρούσα ενότητα αλλά στα συμπεράσματα.

5.3.1 Πρόβλεψη χρονικού διαστήματος 1 λεπτού

Στη παλινδρόμηση Ridge, η τιμή της παραμέτρου κανονικοποίησης (α) που επιλέχθηκε ήταν 10000.0. Η υψηλή αυτή τιμή δείχνει ότι το μοντέλο απέδωσε καλύτερα με ισχυρή λ_2 κανονικοποίηση. Χωρίς να υπάρχει κάποια αλλαγή στις παραμέτρους σε σχέση με τα προηγούμενα αποτελέσματα.

Για τα τυχαία δάση (Random Forest), ο καλύτερος συνδυασμός παραμέτρων ήταν `bootstrap=True`, `max_depth=4`, `max_features=sqrt`, `min_samples_leaf=10`, `min_samples_split=20` και `n_estimators=500`. Η επιλογή μικρού μέγιστου βάθους δείχνει ότι το μοντέλο απέδωσε καλύτερα όταν περιορίστηκε η πολυπλοκότητα των επιμέρους δέντρων. Στο XGBoost, οι βέλτιστες παράμετροι ήταν `colsample_bytree=0.6`, `learning_rate=0.01`, `max_depth=2`, `n_estimators=300`, `reg_alpha=0.0`, `reg_lambda=5.0` και `subsample=0.6`. Ο συγκεκριμένος συνδυασμός δείχνει ότι το μοντέλο απέδωσε καλύτερα με πολύ ρηχά δέντρα, χαμηλό ρυθμό μάθησης και αυξημένη λ_2 κανονικοποίηση.

Για το δίκτυο μακράς βραχυπρόθεσμης μνήμης (LSTM), ο καλύτερος συνδυασμός παραμέτρων ήταν `lookback=120`, `units=64`, `dropout=0.30`, `batch_size=128` και `learning_rate=0.0005`. Η επιλογή `lookback=120` δείχνει ότι το μοντέλο αξιοποίησε πληροφορία από τις προηγούμενες 120 παρατηρήσεις, δηλαδή από ένα αρκετά εκτεταμένο βραχυχρόνιο ιστορικό. Το ίδιο παράθυρο εισόδου επιλέχθηκε και στο GRU, με βέλτιστες παραμέτρους `lookback=120`, `units=64`, `dropout=0.30`, `batch_size=128` και `learning_rate=0.0005`. Η επιλογή αυτή δείχνει πώς μετά την προσθήκη των μεταβλητών sentiment, τα μοντέλα απέδωσαν καλύτερα όταν είχαν πρόσβαση σε μεγαλύτερο χρονικό ιστορικό, πιθανώς επειδή η επίδραση του sentiment δεν αποτυπώνεται απαραίτητα άμεσα στην επόμενη τιμή.

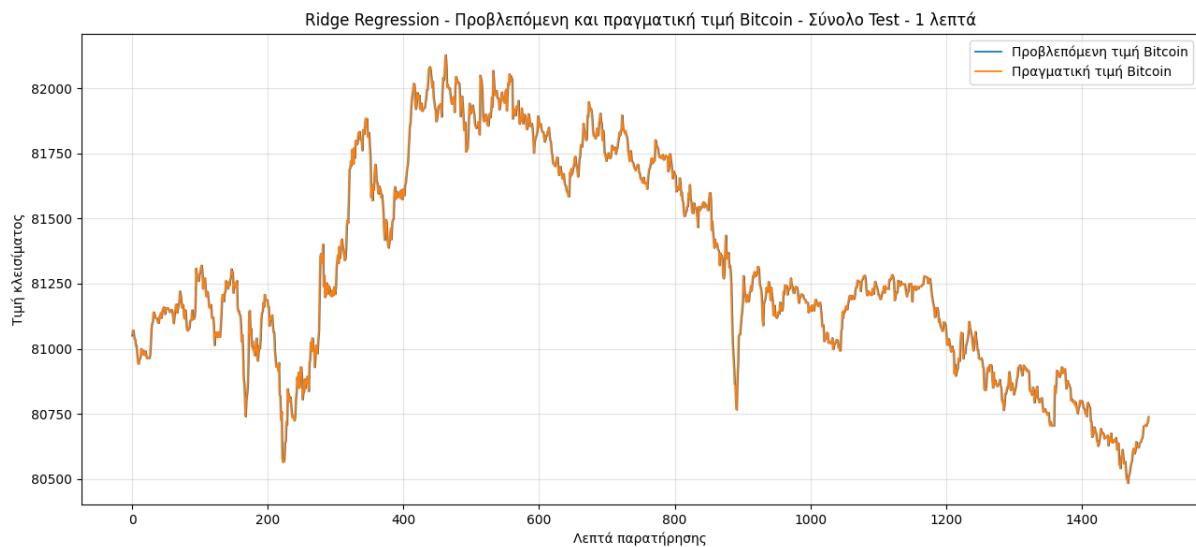
Στον Πίνακα 12 παρουσιάζονται συγκεντρωτικά τα αποτελέσματα όλων των μοντέλων με χρήση μεταβλητών συναισθήματος για τον χρονικό ορίζοντα 1 λεπτού.

Πίνακας 12. Αποτελέσματα μοντέλων με ανάλυση συναισθήματος για χρονικό ορίζοντα 1 λεπτού

Μοντέλο	Σύνολο δεδομένων	RMSE λογαριθμικής απόδοσης	Ακρίβεια κατεύθυνσης	RMSE τιμής κλεισίματος
Ridge Regression	Validation	0.052%	0.465	38.197
Ridge Regression	Test	0.062%	0.487	41.067
Random Forest	Validation	0.052%	0.463	38.202
Random Forest	Test	0.062%	0.485	41.070
XGBoost	Validation	0.052%	0.464	38.197
XGBoost	Test	0.062%	0.487	41.068
LSTM	Validation	0.052%	0.468	38.105
LSTM	Test	0.062%	0.484	41.097
GRU	Validation	0.052%	0.471	38.105
GRU	Test	0.062%	0.484	41.096

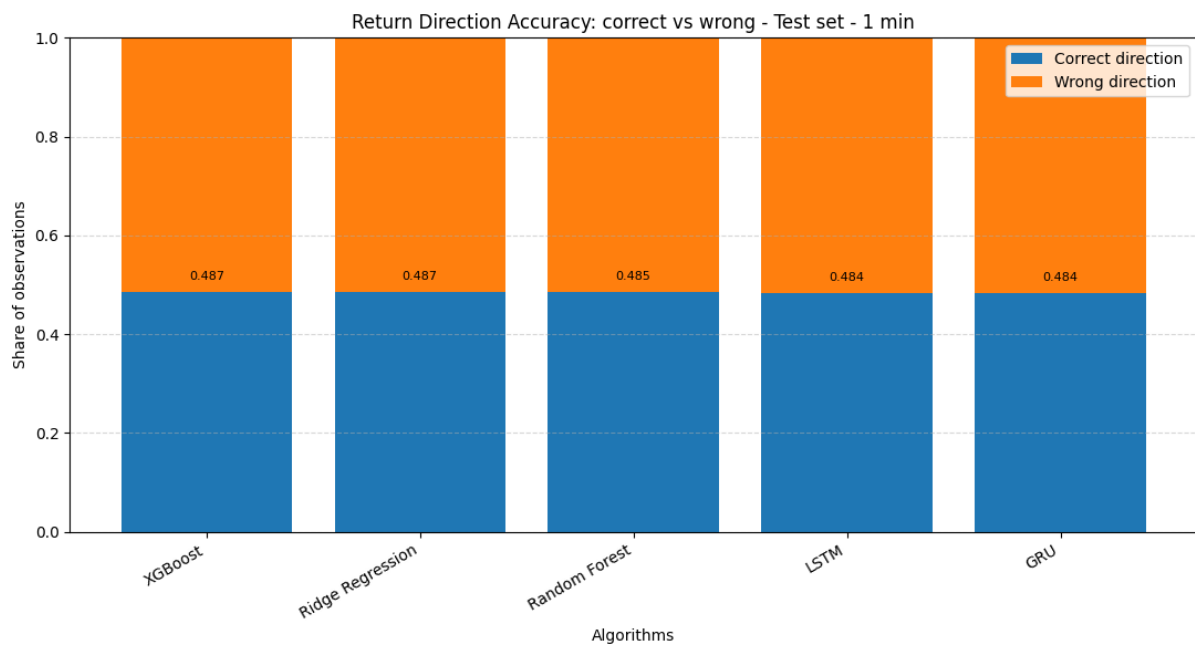
Ως προς το σφάλμα τιμής, το χαμηλότερο Price RMSE στο test set εμφανίζεται από το Ridge Regression, με τιμή 41,067. Πολύ κοντά ακολουθούν και τα υπόλοιπα μοντέλα με μέγιστη διαφορά 0,030 δολάρια γεγονός που δείχνει ότι δεν υπάρχει καθαρή υπεροχή ενός μοντέλου ως προς το αριθμητικό σφάλμα πρόβλεψης στον ορίζοντα του 1 λεπτού.

Με βάση τη χαμηλότερη ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) στο σύνολο ελέγχου, το Ridge Regression επιλέγεται ως το καλύτερο μοντέλο για τον χρονικό ορίζοντα του 1 λεπτού με χρήση μεταβλητών συναισθήματος. Στο Διάγραμμα 23 παρουσιάζεται η σύγκριση της προβλεπόμενης και της πραγματικής τιμής κλεισίματος του Bitcoin στο σύνολο ελέγχου για το συγκεκριμένο μοντέλο. Η οπτική σύγκλιση πρέπει να ερμηνεύεται σε συνδυασμό με τις μετρικές αξιολόγησης, καθώς η ακρίβεια κατεύθυνσης παραμένει χαμηλότερο από το 50%.



Διάγραμμα 23. Γραφική παράσταση προβλεπόμενης και πραγματικής τιμής BTC για χρονικό ορίζοντα 1 λεπτού με χρήση μεταβλητών συναισθήματος για το Ridge Regression

Στο Διάγραμμα 24 παρουσιάζεται συγκριτικά η ακρίβεια πρόβλεψης κατεύθυνσης όλων των μοντέλων στο test set για χρονικό ορίζοντα 1 λεπτού με χρήση sentiment. Παρατηρείται ότι όλα τα μοντέλα εμφανίζουν ακρίβεια κατεύθυνσης κοντά στο 48%, χωρίς κάποιο από αυτά να υπερβαίνει το 50%. Το αποτέλεσμα αυτό δείχνει ότι, παρότι η προσθήκη μεταβλητών συναισθήματος οδηγεί σε μικρή βελτίωση σε σχέση με τα μοντέλα χωρίς, η πρόβλεψη της κατεύθυνσης στο επόμενο λεπτό παραμένει ιδιαίτερα απαιτητική.



Διάγραμμα 24. Ακρίβεια πρόβλεψης κατεύθυνσης ανά μοντέλο για χρονικό ορίζοντα 1 λεπτού με χρήση sentiment

5.3.2 Πρόβλεψη χρονικού διαστήματος 15 λεπτών

Στη παλινδρόμηση Ridge (Ridge Regression), η παράμετρος (alpha) παρέμεινε ίση με 10000.0, γεγονός που δείχνει ότι η ισχυρή λ_2 κανονικοποίηση εξακολουθεί να είναι χρήσιμη και στον χρονικό ορίζοντα των 15 λεπτών.

Για το Random Forest, ο καλύτερος συνδυασμός παραμέτρων ήταν bootstrap=True, max_depth=4, max_features=sqrt, min_samples_leaf=10, min_samples_split=20 και n_estimators=100. Σε σχέση με τον ορίζοντα του 1 λεπτού, ο αριθμός των δέντρων που επιλέχθηκε είναι μικρότερος, ενώ το μέγιστο βάθος παραμένει περιορισμένο. Αυτό δείχνει ότι στον ορίζοντα των 15 λεπτών το μοντέλο δεν χρειάστηκε πολύ μεγάλο ensemble για να επιτύχει την καλύτερη απόδοση. Στο XGBoost, οι βέλτιστες παράμετροι ήταν colsample_bytree=0.6, learning_rate=0.01, max_depth=2, n_estimators=300, reg_alpha=0.0, reg_lambda=1.0 και subsample=0.8.

Για το LSTM, ο καλύτερος συνδυασμός παραμέτρων ήταν lookback=15, units=32, dropout=0.10, batch_size=64 και learning_rate=0.001. Σε αντίθεση με τον ορίζοντα του 1 λεπτού, όπου επιλέχθηκε μεγάλο παράθυρο εισόδου, στον ορίζοντα των 15 λεπτών το LSTM απέδωσε καλύτερα με πιο σύντομο ιστορικό και πιο απλή αρχιτεκτονική. Για το GRU, ο καλύτερος συνδυασμός παραμέτρων ήταν lookback=60, units=64, dropout=0.20, batch_size=64 και learning_rate=0.001. Το GRU αξιοποίησε μεγαλύτερο παράθυρο εισόδου σε σχέση με το LSTM, δηλαδή τις προηγούμενες 60 παρατηρήσεις.

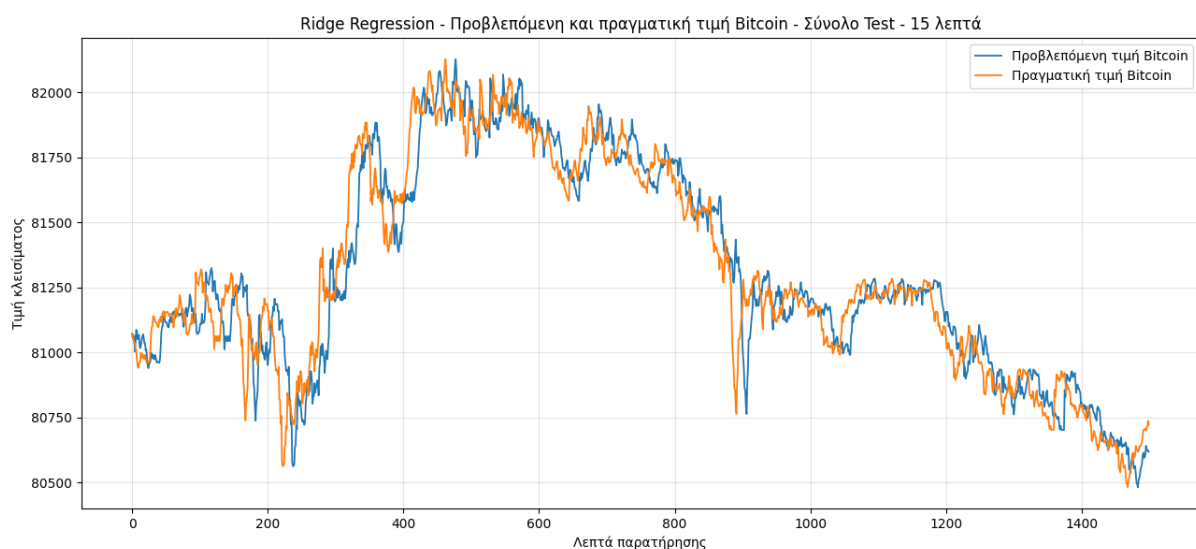
Πίνακας 13. Αποτελέσματα μοντέλων με ανάλυση συναισθήματος για χρονικό ορίζοντα 15 λεπτών

Μοντέλο	Σύνολο δεδομένων	RMSE λογαριθμικής απόδοσης	Ακρίβεια κατεύθυνσης	RMSE τιμής κλεισίματος
Ridge Regression	Validation	0.202%	0.495	148.016
Ridge Regression	Test	0.235%	0.499	155.123
Random Forest	Validation	0.203%	0.494	148.358
Random Forest	Test	0.235%	0.506	155.290
XGBoost	Validation	0.202%	0.498	148.050
XGBoost	Test	0.235%	0.508	155.139
LSTM	Validation	0.203%	0.503	148.741
LSTM	Test	0.236%	0.511	156.221
GRU	Validation	0.202%	0.503	148.188
GRU	Test	0.289%	0.512	185.628

Στον χρονικό ορίζοντα των 15 λεπτών παρατηρείται αύξηση της ρίζας του μέσου τετραγωνικού σφάλματος (RMSE) της λογαριθμικής απόδοσης σε σχέση με τον ορίζοντα του 1 λεπτού, κάτι που είναι αναμενόμενο καθώς η αβεβαιότητα αυξάνεται όσο μεγαλώνει το διάστημα πρόβλεψης. Παρόλα αυτά, τα αποτελέσματα με χρήση μεταβλητών συναισθήματος δείχνουν βελτιωμένη επίδοση σε επίπεδο σφάλματος της τιμής σε σχέση με τα αντίστοιχα αποτελέσματα χωρίς, καθώς οι τιμές του RMSE κινούνται περίπου στις 150 μονάδες για τα καλύτερα μοντέλα.

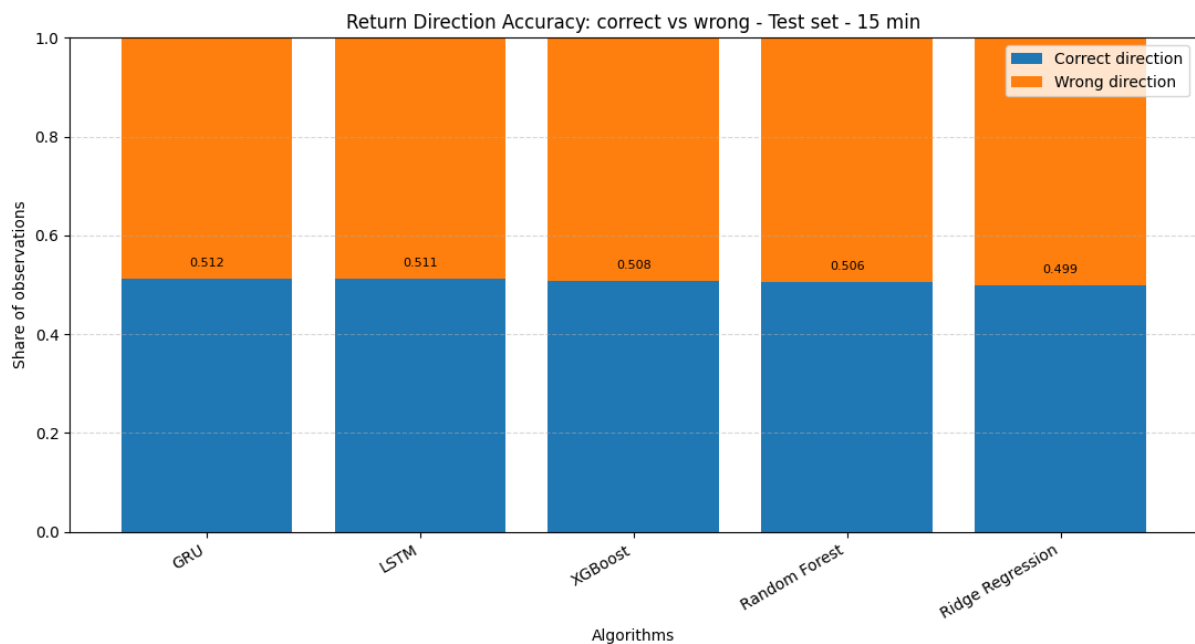
Σε επίπεδο Ακρίβειας Κατεύθυνσης, το GRU με το LSTM εμφανίζουν την υψηλότερη επίδοση στο σύνολο ελέγχου, με τιμή 51,2%, ενώ ακολουθεί το XGBoost με 50,8% . Σε αντίθεση με τον ορίζοντα του 1 λεπτού, αρκετά μοντέλα υπερβαίνουν οριακά το 50%, γεγονός που δείχνει ότι η πληροφορία συναισθήματος ενδέχεται να προσφέρει μικρή βελτίωση στην πρόβλεψη της κατεύθυνσης για τον χρονικό ορίζοντα των 15 λεπτών. Ωστόσο, οι τιμές παραμένουν πολύ κοντά στο 50% και δεν διαφοροποιούνται πολύ από τη αντίστοιχη πρόβλεψη χωρίς μεταβλητές συναισθήματος.

Με βάση το χαμηλότερο RMSE της τιμής στο σύνολο ελέγχου, το Ridge Regression επιλέγεται ως το καλύτερο μοντέλο για τον χρονικό ορίζοντα των 15 λεπτών με χρήση μεταβλητών συναισθήματος. Στο Διάγραμμα 25 μπορεί να παρουσιαστεί η σύγκριση της προβλεπόμενης και της πραγματικής τιμής κλεισίματος του Bitcoin στο σύνολο ελέγχου για το συγκεκριμένο μοντέλο.



Διάγραμμα 25. Γραφική παράσταση προβλεπόμενης και πραγματικής τιμής BTC για χρονικό ορίζοντα 15 λεπτών με χρήση μεταβλητών συναισθήματος για το Ridge Regression

Στο Διάγραμμα 26 παρουσιάζεται συγκριτικά η ακρίβεια πρόβλεψης κατεύθυνσης όλων των μοντέλων στο σύνολο ελέγχου. Παρατηρείται ότι το XGBoost και το Random Forest εμφανίζουν τις υψηλότερες τιμές, υπερβαίνοντας οριακά το 50%. Το αποτέλεσμα αυτό δείχνει ότι η ενσωμάτωση πληροφορίας συναισθήματος μπορεί να προσφέρει μικρή βελτίωση στην πρόβλεψη της κατεύθυνσης.



Διάγραμμα 26. Ακρίβεια πρόβλεψης κατεύθυνσης ανά μοντέλο για χρονικό ορίζοντα 15 λεπτών με χρήση sentiment

5.3.3 Πρόβλεψη χρονικού διαστήματος 30 λεπτών

Στην παρούσα ενότητα παρουσιάζονται τα αποτελέσματα των μοντέλων πρόβλεψης για χρονικό ορίζοντα 30 λεπτών, με την ενσωμάτωση μεταβλητών συναισθήματος. Ο χρονικός αυτός ορίζοντας επιτρέπει την εξέταση της πιθανής επίδρασης των μεταβλητών συναισθήματος σε ένα ελαφρώς μεγαλύτερο βραχυχρόνιο διάστημα, όπου οι πληροφορίες από ειδήσεις και κοινωνικά δίκτυα ενδέχεται να έχουν περισσότερο χρόνο να ενσωματωθούν στην τιμή.

Στη παλινδρόμηση Ridge, η τιμή της παραμέτρου (alpha) παρέμεινε ίση με 10000.0, όπως και στους προηγούμενους ορίζοντες με sentiment. Η σταθερή επιλογή υψηλής κανονικοποίησης δείχνει ότι, μετά την προσθήκη των sentiment χαρακτηριστικών, το γραμμικό μοντέλο αποδίδει καλύτερα όταν περιορίζεται έντονα το μέγεθος των συντελεστών.

Για το Random Forest, ο καλύτερος συνδυασμός παραμέτρων ήταν bootstrap=True, max_depth=4, max_features=sqrt, min_samples_leaf=10, min_samples_split=40 και n_estimators=200. Στο XGBoost, οι βέλτιστες παράμετροι ήταν colsample_bytree=0.6, learning_rate=0.01, max_depth=2, n_estimators=300, reg_alpha=0.0, reg_lambda=1.0 και subsample=0.8. Όπως και στον ορίζοντα των 15 λεπτών, το μοντέλο βασίστηκε σε ρηχά δέντρα και χαμηλό ρυθμό μάθησης.

Για το LSTM, ο καλύτερος συνδυασμός παραμέτρων ήταν lookback=60, units=64, dropout=0.30, batch_size=128 και learning_rate=0.0005. Το μοντέλο αξιοποίησε τις προηγούμενες 60

παρατηρήσεις, δηλαδή πληροφορία από την τελευταία μία ώρα, γεγονός που δείχνει ότι η πρόβλεψη σε ορίζοντα 30 λεπτών επωφελήθηκε από μεγαλύτερο ιστορικό. Για το GRU, ο καλύτερος συνδυασμός παραμέτρων ήταν $\text{lookback}=120$, $\text{units}=64$, $\text{dropout}=0.30$, $\text{batch_size}=128$ και $\text{learning_rate}=0.0005$. Το GRU αξιοποίησε ακόμη μεγαλύτερο παράθυρο εισόδου, δηλαδή τις προηγούμενες 120 παρατηρήσεις.

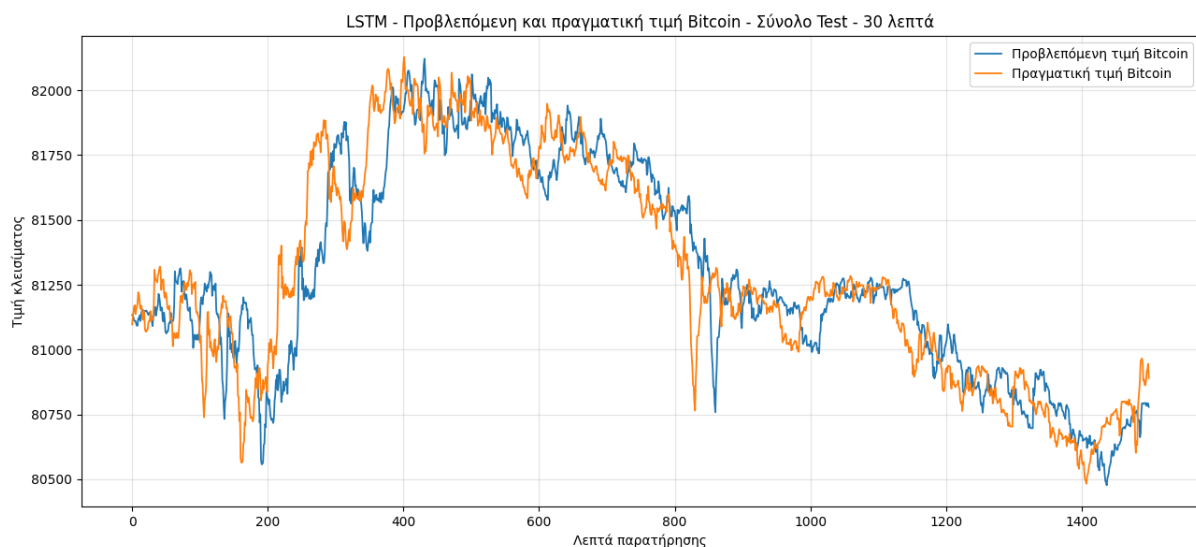
Πίνακας 14. Αποτελέσματα μοντέλων με ανάλυση συναισθήματος για χρονικό ορίζοντα 30 λεπτών

Μοντέλο	Σύνολο δεδομένων	RMSE λογαριθμικής απόδοσης	Ακρίβεια κατεύθυνσης	RMSE τιμής κλεισίματος
Ridge Regression	Validation	0.280%	0.499	205.708
Ridge Regression	Test	0.331%	0.497	218.432
Random Forest	Validation	0.281%	0.492	206.390
Random Forest	Test	0.331%	0.505	218.380
XGBoost	Validation	0.281%	0.489	206.005
XGBoost	Test	0.331%	0.505	218.168
LSTM	Validation	0.293%	0.510	215.176
LSTM	Test	0.331%	0.511	218.031
GRU	Validation	0.282%	0.510	207.210
GRU	Test	0.331%	0.512	218.093

Στον χρονικό ορίζοντα των 30 λεπτών, το RMSE της λογαριθμικής απόδοσης αυξάνεται σε σχέση με τους μικρότερους χρονικούς ορίζοντες, ενώ το RMSE της τιμής διαμορφώνεται περίπου στις 218 μονάδες για τα καλύτερα μοντέλα. Η αύξηση του σφάλματος είναι αναμενόμενη, καθώς η πρόβλεψη γίνεται δυσκολότερη όσο αυξάνεται το χρονικό διάστημα μεταξύ της τρέχουσας και της μελλοντικής τιμής.

Ως προς το σφάλμα τιμής, το χαμηλότερο εμφανίζεται από το δίκτυο μακράς βραχυπρόθεσμης μνήμης (LSTM), με τιμή 218.031. Πολύ κοντά ακολουθούν το GRU και το XGBoost, με τιμές 218.093 και 218.168 αντίστοιχα. Οι διαφορές μεταξύ των τριών αυτών μοντέλων είναι πολύ μικρές, ωστόσο το LSTM εμφανίζει οριακά την καλύτερη επίδοση με βάση το RMSE της τιμής. Σε επίπεδο ακρίβειας κατεύθυνσης, τα δίκτυα βαθιάς μάθησης εμφανίζουν τα καλύτερα αποτελέσματα με τιμή 51,1 – 51,2%.

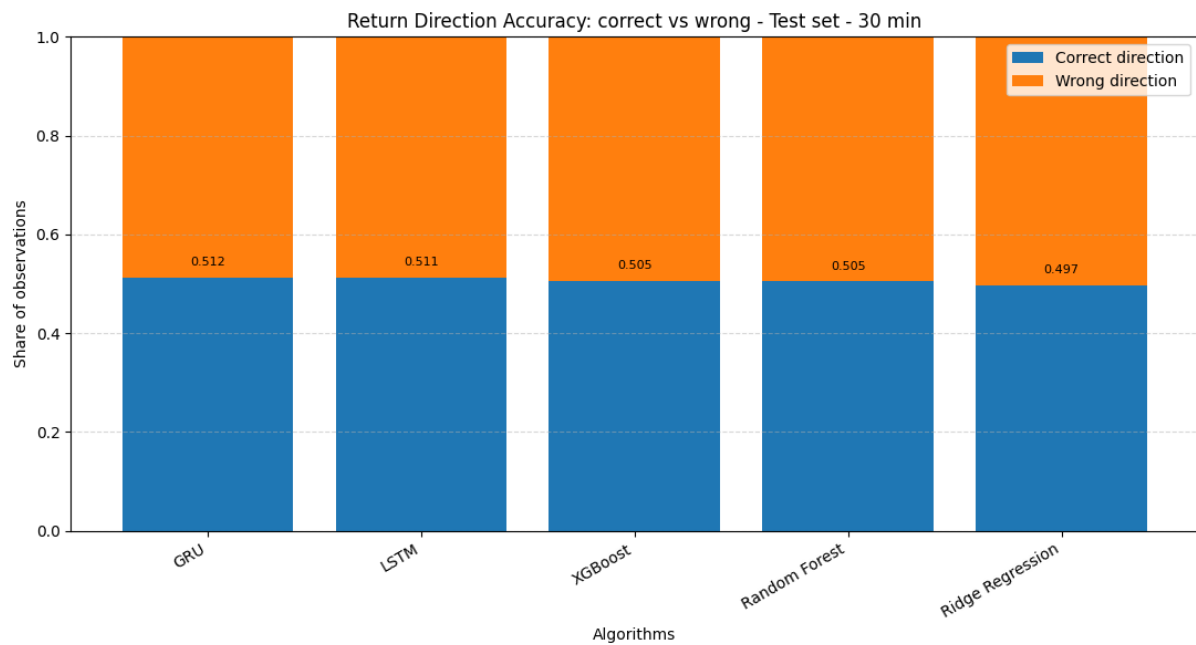
Με βάση το χαμηλότερο σφάλμα στο σύνολο ελέγχου, το LSTM επιλέγεται ως το καλύτερο μοντέλο για τον χρονικό ορίζοντα των 30 λεπτών με χρήση μεταβλητών συναισθήματος. Στο Διάγραμμα 27 παρουσιάζεται η σύγκριση της προβλεπόμενης και της πραγματικής τιμής κλεισίματος του Bitcoin στο test set για το συγκεκριμένο μοντέλο.



Διάγραμμα 27. Γραφική παράσταση προβλεπόμενης και πραγματικής τιμής BTC για χρονικό ορίζοντα 30 λεπτών με χρήση μεταβλητών συναισθήματος για το LSTM

Στο Διάγραμμα 28 παρουσιάζεται συγκριτικά η ακρίβεια πρόβλεψης κατεύθυνσης όλων των μοντέλων στο test set για χρονικό ορίζοντα 30 λεπτών με χρήση sentiment. Το αποτέλεσμα αυτό δείχνει ότι, στον ορίζοντα των 30 λεπτών, η ενσωμάτωση μεταβλητών συναισθήματος φαίνεται να ενισχύει οριακά την πρόβλεψη της κατεύθυνσης της τιμής σε ορισμένα μοντέλα, κυρίως στα

δενδροειδή μοντέλα και στα μοντέλα βαθιάς μάθησης. Ωστόσο, οι τιμές παραμένουν κοντά στο 50%, επομένως η βελτίωση πρέπει να ερμηνεύεται ως περιορισμένη και όχι ως ένδειξη σταθερής προβλεπτικής υπεροχής.



Διάγραμμα 28. Ακρίβεια πρόβλεψης κατεύθυνσης ανά μοντέλο για χρονικό ορίζοντα 30 λεπτών με χρήση sentiment

6. Επίλογος – Συμπεράσματα

Στην παρούσα ενότητα εξετάζεται κατά πόσο η προσθήκη μεταβλητών συναισθήματος βελτίωσε την απόδοση των μοντέλων πρόβλεψης. Η σύγκριση πραγματοποιείται στο σύνολο ελέγχου όπως και στις προηγούμενες ενότητες, καθώς αυτό αποτελεί το τελικό σύνολο αξιολόγησης και αποτυπώνει καλύτερα το πόσο καλά μπορούν να γενικεύσουν τα εξεταζόμενα μοντέλα. Για τη σύγκριση χρησιμοποιούνται δύο βασικές διαστάσεις: η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) της προβλεπόμενης τιμής, και η Ακρίβεια κατεύθυνσης, το οποίο δείχνει την ικανότητα πρόβλεψης της σωστής κατεύθυνσης της μεταβολής.

Πίνακας 15. Σύγκριση των αποτελεσμάτων πριν και μετά την προσθήκη μεταβλητών συναισθήματος

Χρονικός ορίζοντας (λεπτά)	Μοντέλο	RMSE τιμ. κλεισίματος χωρίς sentiment	RMSE τιμ. κλεισίματος με sentiment	Μεταβολή RMSE	Βελτίωση (%) RMSE
1	ARIMA	41.067	-	-	-
1	Ridge Regression	41.067	41.067	0.000	0.00%
1	Random Forest	45.338	41.070	-4.268	+9.41%
1	XGBoost	41.081	41.068	-0.013	+0.03%
1	LSTM	41.138	41.097	-0.041	+0.10%
1	GRU	41.521	41.096	-0.425	+1.02%
15	ARIMA	155.037	-	-	-
15	Ridge Regression	155.094	155.123	+0.029	-0.02%
15	Random Forest	383.939	155.290	-228.649	+59.55%

15	XGBoost	156.413	155.139	-1.274	+0.81%
15	LSTM	155.141	156.221	+1.080	-0.70%
15	GRU	155.393	185.628	+30.235	-19.46%
30	ARIMA	218.307	-	-	-
30	Ridge Regression	218.230	218.432	+0.202	-0.09%
30	Random Forest	728.847	218.380	-510.467	+70.04%
30	XGBoost	221.760	218.168	-3.592	+1.62%
30	LSTM	220.486	218.031	-2.455	+1.11%
30	GRU	220.486	218.093	-2.393	+1.09%

Πίνακας 16. Σύγκριση του *Directional Accuracy* πριν και μετά την προσθήκη μεταβλητών *sentiment* στο *test set*

Χρονικός ορίζοντας (λεπτά)	Μοντέλο	DA χωρίς sentiment	DA με sentiment	Βελτίωση DA
1	ARIMA	0.486	-	-
1	Ridge Regression	0.484	0.487	+0.003
1	Random Forest	0.484	0.485	+0.001
1	XGBoost	0.483	0.487	+0.004
1	LSTM	0.484	0.484	0.000

1	GRU	0.484	0.484	0.000
15	ARIMA	0.511	-	
15	Ridge Regression	0.488	0.499	+0.011
15	Random Forest	0.504	0.506	+0.002
15	XGBoost	0.499	0.508	+0.009
15	LSTM	0.487	0.511	+0.024
15	GRU	0.511	0.512	+0.001
30	ARIMA	0.511	-	-
30	Ridge Regression	0.488	0.497	+0.009
30	Random Forest	0.505	0.505	0.000
30	XGBoost	0.503	0.505	+0.002
30	LSTM	0.511	0.511	0.000
30	GRU	0.511	0.512	+0.001

Από τους δύο συγκριτικούς πίνακες προκύπτει ότι η προσθήκη μεταβλητών sentiment δεν οδήγησε σε ομοιόμορφη βελτίωση όλων των μοντέλων. Η επίδρασή της διαφοροποιείται ανάλογα με το μοντέλο και τον χρονικό ορίζοντα πρόβλεψης.

Σε ορίζοντα 1 λεπτού, οι μεταβολές είναι μικρές για τα περισσότερα μοντέλα. Η παλινδρόμηση Ridge (Ridge Regression) και το XGBoost παραμένουν πρακτικά αμετάβλητα ως προς το σφάλμα της προβλεπόμενης τιμής. Το LSTM εμφανίζει επίσης πολύ μικρή βελτίωση, καθώς η ρίζα του

μέσου τετραγωνικού σφάλματος (RMSE) μειώνεται από 41.138 σε 41.097. Αντίθετα, μεγαλύτερη βελτίωση παρουσιάζουν το Random Forest και το GRU. Συγκεκριμένα, στο Random Forest το RMSE μειώνεται από 45.338 σε 41.070, δηλαδή παρουσιάζει βελτίωση 9.41%, ενώ στο GRU μειώνεται από 41.521 σε 41.096, με βελτίωση 1.02%. Ως προς την ακρίβεια κατεύθυνσης, οι μεταβολές είναι επίσης περιορισμένες. Το Ridge Regression βελτιώνεται από 0.484 σε 0.487, το XGBoost από 0.483 σε 0.487, ενώ το Random Forest αυξάνεται οριακά σε 0.485. Τα LSTM και GRU παραμένουν αμετάβλητα στο 0.484.

Στον ορίζοντα των 15 λεπτών, η μεγαλύτερη βελτίωση παρατηρείται στο Random Forest. Το σφάλμα της τιμής μειώνεται από 278.839 σε 149.272, δηλαδή παρουσιάζει βελτίωση περίπου 46.47%. Η μεταβολή αυτή είναι ιδιαίτερα σημαντική και δείχνει ότι η προσθήκη μεταβλητών συναισθήματος βοήθησε το Random Forest να περιορίσει σημαντικά το σφάλμα πρόβλεψης. Ωστόσο, πρέπει να σημειωθεί ότι το Random Forest χωρίς τη προθήκη μεταβλητών συναισθήματος είχε εμφανίσει πολύ υψηλό σφάλμα σε σχέση με τα υπόλοιπα μοντέλα. Επομένως, η μεγάλη ποσοστιαία βελτίωση δεν σημαίνει απαραίτητα ότι το sentiment οδηγεί γενικά σε μεγάλη αύξηση της προβλεπτικής ικανότητας, αλλά ότι στη συγκεκριμένη περίπτωση βοήθησε το Random Forest να γενικεύσει καλύτερα στο σύνολο ελέγχου.

Εκτός από την μεγάλη μεταβολή στο Random Forest, το XGBoost παρουσιάζει μικρή βελτίωση, καθώς το RMSE μειώνεται από 156.413 σε 155.139, δηλαδή κατά 0.81%. Αντίθετα, το Ridge Regression παρουσιάζει πρακτικά αμελητέα επιδείνωση, από 155.094 σε 155.123, ενώ το LSTM αυξάνει το σφάλμα από 155.141 σε 156.221. Η μεγαλύτερη επιδείνωση παρατηρείται στο GRU, όπου το RMSE αυξάνεται από 155.393 σε 185.628. Ως προς την ακρίβεια κατεύθυνσης, οι περισσότερες μεταβολές είναι θετικές. Το Ridge Regression αυξάνεται από 0.488 σε 0.499, το XGBoost από 0.499 σε 0.508, το LSTM από 0.487 σε 0.511, το Random Forest από 0.504 σε 0.506 και το GRU από 0.511 σε 0.512.

Στον ορίζοντα των 30 λεπτών, το Random Forest παρουσιάζει ξανά τη μεγαλύτερη βελτίωση. Βελτίωση στο RMSE εμφανίζουν επίσης το XGBoost, το LSTM και το GRU. Το XGBoost μειώνει το RMSE από 221.760 σε 218.168, παρουσιάζοντας βελτίωση 1.62%. Το LSTM μειώνει το RMSE από 220.486 σε 218.031, με βελτίωση 1.11%, ενώ το GRU μειώνει το RMSE από 220.486 σε 218.093, με βελτίωση 1.09%. Αντίθετα, το Ridge Regression παρουσιάζει οριακή επιδείνωση, καθώς το RMSE αυξάνεται από 218.230 σε 218.432. Ως προς την ακρίβεια κατεύθυνσης, οι μεταβολές είναι μικρές. Το Ridge Regression βελτιώνεται από 0.488 σε 0.497, το XGBoost από 0.503 σε 0.505 και το

GRU από 0.511 σε 0.512. Το Random Forest παραμένει σταθερό στο 0.505, ενώ το LSTM παραμένει επίσης αμετάβλητο στο 0.511.

Συνολικά, τα αποτελέσματα δείχνουν ότι η προσθήκη μεταβλητών ανάλυσης συναισθήματος βελτιώνει τα μοντέλα αλλά όχι συστηματικά και όχι καθολικά. Η πιο έντονη θετική επίδραση παρατηρείται στο Random Forest. Τα Ridge Regression και XGBoost εμφανίζουν γενικά σταθερή συμπεριφορά, με πολύ μικρές μεταβολές στο ρίζα του μέσου τετραγωνικού σφάλματος (RMSE), ενώ το XGBoost παρουσιάζει ελαφρώς θετικότερη εικόνα στους μεγαλύτερους ορίζοντες. Τα LSTM και GRU παρουσιάζουν μεικτή συμπεριφορά, καθώς σε ορισμένες περιπτώσεις βελτιώνουν το RMSE, ενώ σε άλλες, ιδιαίτερα στον ορίζοντα των 15 λεπτών, παρουσιάζουν επιδείνωση.

Ως προς την ακρίβεια κατεύθυνσης, η προσθήκη μεταβλητών συναισθήματος οδηγεί κυρίως σε μικρές θετικές μεταβολές, ιδιαίτερα στους ορίζοντες 15 και 30 λεπτών. Ωστόσο, οι τιμές παραμένουν κοντά στο 50%, γεγονός που δείχνει ότι, παρά τις επιμέρους βελτιώσεις, η πρόβλεψη της κατεύθυνσης της τιμής παραμένει δύσκολη. Το γεγονός ότι η ακρίβεια κατεύθυνσης παραμένει κοντά στο 50% δείχνει ότι η πρόβλεψη της κατεύθυνσης της τιμής είναι δυσκολότερη από την προσέγγιση του επιπέδου της τιμής. Σε δεδομένα υψηλής συχνότητας, οι διαδοχικές μεταβολές είναι συχνά πολύ μικρές και κινούνται γύρω από το μηδέν. Έτσι, ακόμη και μικρό σφάλμα στην πρόβλεψη της λογαριθμικής απόδοσης μπορεί να οδηγήσει σε λανθασμένο πρόσημο και συνεπώς σε λάθος πρόβλεψη κατεύθυνσης. Παράλληλα, η αγορά του Bitcoin αντιδρά γρήγορα σε νέα πληροφορία και χαρακτηρίζεται από έντονο θόρυβο, γεγονός που περιορίζει τη δυνατότητα σταθερής πρόβλεψης ανοδικών ή καθοδικών κινήσεων σε τόσο μικρούς χρονικούς ορίζοντες.

Επομένως, τα χαρακτηριστικά συναισθήματος φαίνεται να προσφέρουν κάποια πρόσθετη πληροφορία, αλλά η επίδρασή τους είναι περιορισμένη και εξαρτάται σημαντικά από το μοντέλο και τον χρονικό ορίζοντα πρόβλεψης.

7. Μελλοντική έρευνα

Η παρούσα διπλωματική εργασία επικεντρώθηκε στη βραχυχρόνια πρόβλεψη της τιμής του Bitcoin, εξετάζοντας διαφορετικές κατηγορίες μοντέλων και αξιολογώντας την επίδραση της προσθήκης μεταβλητών ανάλυσης συναισθήματος. Παρότι τα αποτελέσματα έδειξαν ότι οι μεταβλητές συναισθήματος μπορούν να προσφέρουν πρόσθετη πληροφορία σε ορισμένα μοντέλα και χρονικούς ορίζοντες, η βελτίωση δεν ήταν καθολική. Επομένως, υπάρχουν αρκετές κατευθύνσεις στις οποίες θα μπορούσε να επεκταθεί η παρούσα έρευνα.

Μία πρώτη κατεύθυνση αφορά τον εμπλουτισμό του συνόλου δεδομένων με περισσότερες εξωτερικές μεταβλητές. Στην παρούσα εργασία χρησιμοποιήθηκαν κυρίως ιστορικά δεδομένα αγοράς, τεχνικά χαρακτηριστικά και μεταβλητές συναισθήματος από κοινωνικά δίκτυα και ειδησεογραφικές πηγές. Μελλοντικά, θα μπορούσαν να ενσωματωθούν επιπλέον μεταβλητές που αποτυπώνουν τη ρευστότητα της αγοράς. Ακόμα, θα μπορούσαν να αξιοποιηθούν τιμές άλλων περιουσιακών στοιχείων, όπως Ethereum, S&P 500, Nasdaq, χρυσός ή δολαριακοί δείκτες, ώστε να εξεταστεί αν οι σχέσεις μεταξύ του Bitcoin και άλλων αγορών μπορούν να βελτιώσουν την προβλεπτική ικανότητα των μοντέλων.

Επιπλέον, μελλοντική έρευνα θα μπορούσε να δώσει μεγαλύτερη έμφαση στην πρόβλεψη της κατεύθυνσης της μεταβολής και όχι αποκλειστικά στην ελαχιστοποίηση του αριθμητικού σφάλματος της τιμής. Στην παρούσα διπλωματική τα μοντέλα εκπαιδεύτηκαν κυρίως ως μοντέλα παλινδρόμησης, με στόχο την πρόβλεψη της μελλοντικής λογαριθμικής απόδοσης. Ωστόσο, από πρακτική και επενδυτική σκοπιά, η σωστή πρόβλεψη της κατεύθυνσης της τιμής μπορεί να είναι εξίσου ή και περισσότερο σημαντική από την ακριβή εκτίμηση του μεγέθους της μεταβολής. Για τον λόγο αυτό, θα μπορούσαν να εξεταστούν μοντέλα ταξινόμησης, τα οποία θα προβλέπουν αν η τιμή θα κινηθεί ανοδικά ή καθοδικά σε κάθε χρονικό ορίζοντα. Παράλληλα, η επιλογή υπερπαραμέτρων θα μπορούσε να γίνεται με βάση την ακρίβεια κατεύθυνσης (directional accuracy) ή άλλες σχετικές μετρικές ταξινόμησης, αντί για τη ρίζα του μέσου τετραγωνικού σφάλματος (RMSE).

Μία ακόμη σημαντική κατεύθυνση αφορά τη βελτίωση της ανάλυσης συναισθήματος. Στην παρούσα εργασία χρησιμοποιήθηκε το VADER, το οποίο είναι κατάλληλο για σύντομα και άτυπα κείμενα, όπως tweets και Reddit comments. Ωστόσο, σε μελλοντική εργασία θα μπορούσαν να συγκριθούν διαφορετικές μέθοδοι εξαγωγής sentiment, όπως FinBERT, BERT-based μοντέλα ή άλλα γλωσσικά μοντέλα προσαρμοσμένα στο χρηματοοικονομικό λεξιλόγιο. Επίσης, θα μπορούσε να εξεταστεί ξεχωριστά η επίδραση κάθε πηγής, δηλαδή Twitter, Reddit και ειδήσεων, ώστε να διαπιστωθεί αν κάποια πηγή παρέχει πιο χρήσιμη πληροφορία για συγκεκριμένους χρονικούς ορίζοντες.

Τέλος, μία πρακτική επέκταση της παρούσας εργασίας θα ήταν η ανάπτυξη ενός live pipeline πρόβλεψης και ανάλυσης. Ένα τέτοιο σύστημα θα μπορούσε να συλλέγει σε πραγματικό χρόνο δεδομένα τιμών, όγκου συναλλαγών, ειδήσεων και δημοσιεύσεων από κοινωνικά δίκτυα, να υπολογίζει αυτόματα τεχνικούς δείκτες και sentiment scores και να εκτελεί προβλέψεις με βάση τον χρονικό ορίζοντα που επιλέγει ο χρήστης. Με αυτόν τον τρόπο, η μεθοδολογία της διπλωματικής θα μπορούσε να μετατραπεί από ένα στατικό ερευνητικό πείραμα σε ένα δυναμικό εργαλείο υποστήριξης αποφάσεων. Το σύστημα αυτό θα μπορούσε να παρουσιάζει στον χρήστη την προβλεπόμενη μεταβολή, την πιθανή κατεύθυνση της τιμής, τις πρόσφατες μεταβολές στο sentiment και την αβεβαιότητα της πρόβλεψης, παρέχοντας μία πιο ολοκληρωμένη εικόνα της αγοράς σε πραγματικό χρόνο.

Συνολικά, οι παραπάνω επεκτάσεις θα μπορούσαν να ενισχύσουν την ερμηνευτική και πρακτική αξία της ανάλυσης. Η ενσωμάτωση περισσότερων εξωτερικών μεταβλητών, η μεγαλύτερη έμφαση στην πρόβλεψη κατεύθυνσης και η ανάπτυξη ενός live συστήματος πρόβλεψης αποτελούν φυσικά επόμενα βήματα για τη συνέχιση της έρευνας γύρω από την πρόβλεψη της τιμής του Bitcoin.

Βιβλιογραφία

- [1] M. Zubair, J. Ali and M. Alhussein, "An Improved Machine Learning-Driven Framework for Cryptocurrencies Price Prediction with Sentimental Cautioning," *IEEE Access*, pp. 51395-51418, 2024.
- [2] A. Sunki, C. SatyaKumar, G. Surya Narayana, V. Koppera and M. Hakeem, "Time series forecasting of stock market using ARIMA, LSTM and FB prophet," *MATEC Web of Conferences*, 2024.
- [3] S. A. Panchal, L. Ferdouse and A. Sultana, "Comparative Analysis of ARIMA and LSTM Models for Stock Price Prediction," in *27th IEEE/ACIS International Summer Conference on Software Engineering Artificial Intelligence Networking and Parallel/Distributed Computing, SNPD 2024 - Proceedings*, Beijing, 2024.
- [4] N. Tripathy, S. Hota, D. Singh, B. M. Acharya and S. K. Nayak, "A comprehensive analysis of Bitcoin volatility forecasting using time-series econometric models," *Applied Soft Computing*, 2025.
- [5] M. Saberironaghi, J. Ren and A. Saberironaghi, "Stock Market Prediction Using Machine Learning and Deep Learning Techniques: A Review," *Applied Math*, 2025.
- [6] M. S. Islam, M. Bashir, S. Rahman, M. A. Al Montaser, J. C. Bortty, A. Nishan and M. R. Haque, "Machine Learning-Based Cryptocurrency Prediction: Enhancing Market Forecasting with Advanced Predictive Models," *Journal of Ecohumanism*, 2025.
- [7] E. Akyildirim, A. Goncu and A. Sensoy, "Prediction of cryptocurrency returns using machine learning," *Annals of Operations Research*, pp. 3-36, 2021.
- [8] R. Dubey and D. Enke, "Bitcoin price direction prediction using on-chain data and feature selection," *Bitcoin price direction prediction using on-chain data and feature selection*, 2025.
- [9] M. Z. Islam, M. S. Rahman, M. Sumsuzoha, B. Sarker, M. R. Islam, M. Alam and S. K. Shil, "Cryptocurrency Price Forecasting Using Machine Learning: Building Intelligent Financial Prediction Models," *International Journal of Accounting and Economics Studies*, pp. 255-268, 2025.

- [10] X. Li, Z. Liu, Y. Liu, S. Zhu and J. Yan, "Predicting cryptocurrency returns with machine learning: Evidence from high-dimensional factor modeling," *Pacific-Basin Finance Journal*, 2026.
- [11] S. P. Azizi, S. P. Azizi, C. Y. Huang, A. Nafei and Y. Kord, "A neural network-driven dynamical system model for cryptocurrency price prediction," *Physica A: Statistical Mechanics and its Applications*, 2026.
- [12] M. A. Izadi and E. Hajizadeh, "Time series prediction for cryptocurrency markets with transformer and parallel convolutional neural networks," *Applied Soft Computing*, 2025.
- [13] M. Zubair, J. Ali, M. Alhussein, S. Hassan, K. Aurangzeb and M. Umair, "An Improved Machine Learning-Driven Framework for Cryptocurrencies Price Prediction with Sentimental Cautioning," *IEEE Access*, pp. 51395-51418, 2024.
- [14] F. Ben Hamadou, T. Mezghani, R. Zouari and M. Boujelbène-Abbes, "Forecasting Bitcoin returns using machine learning algorithms: impact of investor sentiment," *EuroMed Journal of Business*, pp. 179-200, 2025.
- [15] Y. Gür and E. Ünal, "The Impact of News Sentiment on the Bitcoin Price via Machine Learning and Deep Learning-Based NLP Models," *Journal of Forecasting*, 2025.
- [16] F. Tabe, A. Mansourabady, A. H. Rasekh and B. Tanoori, "Proposing a hybrid deep learning model with sentiment and historical data for cryptocurrencies price prediction," *Expert Systems with Applications*, 2026.
- [17] A. Sidorov, "THE IMPACT OF ANNOUNCEMENTS ON CRYPTOCURRENCY PRICES," *Revista Economica*, pp. 69-94, 2024.
- [18] L. Almaqableh, M. Khasawneh and M. Sahiner, "Environmental News and Bitcoin Market Dynamics: An Event Study of Global Climate-Related Shocks," *FinTech*, 2025.
- [19] A. Mahfooz and J. L. Phillips, "Conditional Forecasting of Bitcoin Prices Using Exogenous Variables," *IEEE Access*, pp. 44510-44526, 2024.
- [20] N. Mgadmi, T. Sadraoui and A. Abidi, "Causality between stock indices and cryptocurrencies before and during the Russo-Ukrainian war," *International Review of Economics*, pp. 301-323, 2024.

- [21] B. Gaies, N. Chaâbane and N. Bouzouita, "Navigating the storm: Time-frequency quantile dependence and non-linear causality between crypto-currency market volatility and financial instability," *Quarterly Review of Economics and Finance*, pp. 43-70, 2024.
- [22] A. E. Hoerl and R. W. Kennard, "Technometrics This content downloaded from 79.167.136.2 on Wed," American Statistical Association and American Society for Quality, 1970.
- [23] L. Breiman, "Random Forests," *Machine Learning*, pp. 5-32, 2001.
- [24] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *ArXiv, Machine Learning*, 2016.
- [25] S. Hochreiter and J. Schmidhuber, "LONG SHORT-TERM MEMORY," *Neural Computation*, pp. 1735-1780, 1997.
- [26] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," 2014.
- [27] C. J. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.
- [28] D. T. Araci and Z. Genc, "Financial Sentiment Analysis with Pre-trained Language Models," 2020.
- [29] "Bitstamp by Robinhood," 10 5 2026. [Online]. Available: <https://www.bitstamp.net/api/>.
- [30] T. Sweeney, "CoinTracker," 26 3 2026. [Online]. Available: <https://www.cointracker.io/blog/bitcoin-price-history>.
- [31] OANDA, "OANDA," 4 6 2026. [Online]. Available: <https://www.oanda.com/us-en/trade-tap-blog/asset-classes/crypto/oanda-bitcoin-price-history-key-market-events-data-charts-insights-volatility/>.
- [32] Bitcoin_Magazine_Pro, "Bitcoin Magazine Pro," 27 12 2024. [Online]. Available: <https://www.bitcoinmagazinepro.com/blog/10-catalytic-bitcoin-events-in-2024-that-could-define-2025/>.
- [33] Reuters, "Reuters," 3 February 2025. [Online]. [Accessed July 2026].

- [34] H. Lang, "Reuters," 19 July 2025. [Online]. Available: <https://www.reuters.com/legal/government/trump-signs-stablecoin-law-crypto-industry-aims-mainstream-adoption-2025-07-18>. [Accessed July 2026].
- [35] H. Lang, "Reuters," February 2026. [Online]. Available: <https://www.reuters.com/markets/wealth/crypto-market-volatility-triggers-25-billion-bitcoin-liquidations-2026-02-02>. [Accessed July 2026].
- [36] D. A. Dickey and W. A. Fuller, "Distribution of the Estimators for Autoregressive Time Series With a Unit Root," *Journal of the American Statistical Association*, pp. 427-431, 1979.
- [37] N. N. AlMadany, O. Hujran, G. A. Naymat and A. Maghyereh, "Forecasting cryptocurrency returns using classical statistical and deep learning techniques," *International Journal of Information Management Data Insights*, 2024.
- [38] S. H. Sung, J. M. Kim, B. K. Park and S. Kim, "A Study on Cryptocurrency Log-Return Price Prediction Using Multivariate Time-Series Model," *Axioms*, 2022.
- [39] Y. Liu, Z. Li, R. Nekhili and J. Sultan, "Forecasting cryptocurrency returns with machine learning," *Research in International Business and Finance*, 2023.
- [40] P. Jaquart, D. Dann and C. Weinhardt, "Short-term bitcoin market prediction via machine learning," *Journal of Finance and Data Science*, pp. 45-66, 2021.
- [41] X. D. Console, "X Developer Console," 2 2 2026. [Online]. Available: <https://console.x.com/accounts>.
- [42] coindesk, "coindesk," 2 2 2026. [Online]. Available: <https://www.coindesk.com/arc/outboundfeeds/rss/?outputType=xml>.
- [43] cointelegraph, "cointelegraph," 2 2 2026. [Online]. Available: <https://cointelegraph.com/rss>.
- [44] cryptoslate, "cryptoslate," 2 2 2026. [Online]. Available: <https://cryptoslate.com/feed/>.
- [45] cryptopotato, "cryptopotato," 2 2 2026. [Online]. Available: <https://cryptopotato.com/feed/>.
- [46] cryptonews, "cryptonews," 2 2 2026. [Online]. Available: <https://cryptonews.com/news/feed/>.
- [47] t. defiant, "the defiant," 2 2 2026. [Online]. Available: <https://thedefiant.io/feed/>.

Παράρτημα

Στον ακόλουθο πίνακα παρουσιάζονται τα ακρωνύμια, οι συντομογραφίες και τα σύμβολα που χρησιμοποιούνται στην παρούσα εργασία, μαζί με την πλήρη αγγλική ονομασία και την αντίστοιχη ελληνική απόδοση ή επεξήγηση.

Ακρωνύμιο / συντομογραφία	Πλήρης αγγλική ονομασία	Ελληνική απόδοση / επεξήγηση
1D-CNN	One-Dimensional Convolutional Neural Network	Μονοδιάστατο συνελκτικό νευρωνικό δίκτυο
ACF	Autocorrelation Function	Συνάρτηση αυτοσυσχέτισης
ADF	Augmented Dickey–Fuller Test	Επαυξημένος έλεγχος Dickey–Fuller για τη στασιμότητα χρονοσειρών
ANN	Artificial Neural Network	Τεχνητό νευρωνικό δίκτυο
API	Application Programming Interface	Διεπαφή προγραμματισμού εφαρμογών
AR	Autoregressive	Αυτοπαλίνδρομο μέρος μοντέλου χρονοσειράς
ARIMA	Autoregressive Integrated Moving Average	Αυτοπαλίνδρομο ολοκληρωμένο μοντέλο κινητού μέσου
ARIMAX	Autoregressive Integrated Moving Average with Exogenous Variables	ARIMA με εξωγενείς μεταβλητές
BERT	Bidirectional Encoder Representations from Transformers	Αμφίδρομες αναπαραστάσεις κωδικοποιητών από Μετασχηματιστές
Bi-LSTM / BLSTM	Bidirectional Long Short-Term Memory	Αμφίδρομο δίκτυο μακράς βραχυπρόθεσμης μνήμης
BTC	Bitcoin	Σύμβολο διαπραγμάτευσης του Bitcoin
BTC/USD	Bitcoin / United States Dollar	Ισοτιμία Bitcoin προς δολάριο ΗΠΑ
CNN	Convolutional Neural Network	Συνελκτικό νευρωνικό δίκτυο
CNN-GRU	Convolutional Neural Network – Gated Recurrent Unit	Υβριδικό μοντέλο CNN και GRU
CNN-LSTM	Convolutional Neural Network – Long Short-Term Memory	Υβριδικό μοντέλο CNN και LSTM
DA	Directional Accuracy	Ακρίβεια κατεύθυνσης
EMD	Empirical Mode Decomposition	Εμπειρική αποσύνθεση τρόπων
ETF	Exchange-Traded Fund	Διαπραγματεύσιμο αμοιβαίο κεφάλαιο

FGARCH / FIGARCH	Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity	Κλασματικά ολοκληρωμένο μοντέλο GARCH
FTX	FTX Cryptocurrency Exchange	Ονομασία ανταλλακτηρίου κρυπτονομισμάτων
GARCH	Generalized Autoregressive Conditional Heteroskedasticity	Γενικευμένο αυτοπαλίνδρομο υπόδειγμα υπό συνθήκη ετεροσκεδαστικότητας
GBRT	Gradient Boosting Regression Trees	Δέντρα παλινδρόμησης βαθμιαίας ενίσχυσης
GENIUS Act	Guiding and Establishing National Innovation for U.S. Stablecoins Act	Νομοθετική πράξη των ΗΠΑ για τη ρύθμιση των stablecoins
GRU	Gated Recurrent Unit	Επαναλαμβανόμενη μονάδα με πύλες
KNN	k-Nearest Neighbors	Αλγόριθμος k πλησιέστερων γειτόνων
LASSO	Least Absolute Shrinkage and Selection Operator	Τελεστής ελάχιστης απόλυτης συρρίκνωσης και επιλογής
LSTM	Long Short-Term Memory	Δίκτυο μακράς βραχυπρόθεσμης μνήμης
MA	Moving Average	Μέρος κινητού μέσου σε μοντέλο χρονοσειράς
MAE	Mean Absolute Error	Μέσο απόλυτο σφάλμα
MSE	Mean Squared Error	Μέσο τετραγωνικό σφάλμα
OHLCV	Open, High, Low, Close, Volume	Τιμή ανοίγματος, μέγιστη, ελάχιστη, κλεισίματος και όγκος συναλλαγών
OLS	Ordinary Least Squares	Μέθοδος ελαχίστων τετραγώνων
PACF	Partial Autocorrelation Function	Συνάρτηση μερικής αυτοσυσχέτισης
PCA	Principal Component Analysis	Ανάλυση κύριων συνιστωσών
PRAW	Python Reddit API Wrapper	Βιβλιοθήκη Python για πρόσβαση στο API του Reddit
ReLU	Rectified Linear Unit	Διορθωμένη γραμμική μονάδα ενεργοποίησης
RMSE	Root Mean Squared Error	Ρίζα μέσου τετραγωνικού σφάλματος
RNN	Recurrent Neural Network	Επαναλαμβανόμενο νευρωνικό δίκτυο
RSS	Really Simple Syndication	Μορφή ροής για αυτόματη διανομή ενημερώσεων περιεχομένου
SHAP	SHapley Additive exPlanations	Μέθοδος ερμηνείας της συμβολής των χαρακτηριστικών στις προβλέψεις
SVM	Support Vector Machine	Μηχανή διανυσμάτων υποστήριξης
SVR	Support Vector Regression	Παλινδρόμηση διανυσμάτων υποστήριξης
TARCH	Threshold Autoregressive Conditional Heteroskedasticity	Κατωφλικό αυτοπαλίνδρομο υπόδειγμα υπό συνθήκη ετεροσκεδαστικότητας

TCN	Temporal Convolutional Network	Χρονικό συνελκτικό δίκτυο
USDT	Tether USD	Σταθερό κρυπτονόμισμα συνδεδεμένο με το δολάριο ΗΠΑ
VADER	Valence Aware Dictionary and sEntiment Reasoner	Λεξικογραφικό εργαλείο ανάλυσης συναισθήματος
VIX	Volatility Index	Δείκτης μεταβλητότητας
VVR	Volume-to-Volatility Ratio	Λόγος όγκου συναλλαγών προς μεταβλητότητα
VWAP	Volume-Weighted Average Price	Σταθμισμένη ως προς τον όγκο μέση τιμή
XGBoost	eXtreme Gradient Boosting	Ακραία βαθμιαία ενίσχυση