ΠΑΝΕΠΙΣΤΗΜΙΟ
ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
UNIVERSITY OF WEST ATTICA

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΒΙΟΜΗΧΑΝΙΚΗΣ ΣΧΕΔΙΑΣΗΣ ΚΑΙ ΠΑΡΑΓΩΓΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΝΙΧΝΕΥΣΗ ΑΝΩΜΑΛΙΩΝ ΣΕ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΕΣ ΧΡΟΝΟΣΕΙΡΕΣ ΜΕ ΜΕΘΟΔΟΥΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

ΔΑΜΟΡΑΚΗΣ ΙΩΑΝΝΗΣ

71446178

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ:

ΚΑΝΤΖΟΣ ΔΗΜΗΤΡΙΟΣ

Αθήνα, Οκτώβριος 2024

ΠΑΝΕΠΙΣΤΗΜΙΟ
ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
UNIVERSITY OF WEST ATTICA

SCHOOL OF ENGINEERING

DEPARTMENT OF INDUSTRIAL DESIGN AND PRODUCTION ENGINEERING

DIPLOMA THESIS

ANOMALY DETECTION IN FINANCIAL TIMESERIES USING MACHINE LEARNING METHODS

IOANNIS DAMORAKIS

71446178

SUPERVISOR

DIMITRIOS KANTZOS

Athens, October 2024

ΑΝΙΝΧΕΥΣΗ ΑΝΩΜΑΛΙΩΝ ΣΕ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΕΣ ΧΡΟΝΟΣΕΙΡΕΣ ΜΕ ΜΕΘΟΔΟΥΣ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ

Μέλη Εξεταστικής Επιτροπής συμπεριλαμβανομένου και του Εισηγητή

Η διπλωματική εργασία εξετάστηκε επιτυχώς από την κάτωθι Εξεταστική Επιτροπή:

Α/α	ΟΝΟΜΑ ΕΠΩΝΥΜΟ	ΒΑΘΜΙΔΑ/ΙΔΙΟΤΗΤΑ	ΨΗΦΙΑΚΗ ΥΠΟΓΡΑΦΗ
1	ΔΗΜΗΤΡΙΟΣ ΚΑΝΤΖΟΣ	ΚΑΘΗΓΗΤΗΣ	DIMITRIOS KANTZOS Digitally signed by DIMITRIOS KANTZOS Date: 2024.10.14 11:16:11 +03'00'
2	ΧΡΗΣΤΟΣ ΔΡΟΣΟΣ	ΕΠΙΚΟΥΡΟΣ ΚΑΘΗΓΗΤΗΣ	Digitally signed by CHRISTOS DROSOS Date: 2024.10.14 13:26:22 +03'00'
3	ΓΡΗΓΟΡΗΣ ΝΙΚΟΛΑΟΥ	ΛΕΚΤΟΡΑΣ	Grigorios Nikolaou Digitally signed by Grigorios Nikolaou Date: 2024.10.14 17:01:57 +03'00'

ΔΗΛΩΣΗ ΣΥΓΓΡΑΦΕΑ ΔΙΠΛΩΜΑΤΙΚΗΣ ΕΡΓΑΣΙΑΣ

Ο κάτωθι υπογεγραμμένος Δαμοράκης Ιωάννης του Γεωργίου, με αριθμό μητρώου 71446178 φοιτητής του Πανεπιστημίου Δυτικής Αττικής της Σχολής Πανεπιστήμιο Δυτικής Αττικής του Τμήματος Βιομηχανικής Σχεδίασης και Παραγωγής, δηλώνω υπεύθυνα ότι:

«Είμαι συγγραφέας αυτής της πτυχιακής/διπλωματικής εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, οι όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών ή λέξεων, είτε ακριβώς είτε παραφρασμένες, αναφέρονται στο σύνολό τους, με πλήρη αναφορά στους συγγραφείς, τον εκδοτικό οίκο ή το περιοδικό, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Επίσης, βεβαιώνω ότι αυτή η εργασία έχει συγγραφεί από μένα αποκλειστικά και αποτελεί προϊόν πνευματικής ιδιοκτησίας τόσο δικής μου, όσο και του Ιδρύματος.

Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου»

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα καθηγητή της πτυχιακής μου εργασίας, κ. Δημήτριο Κάντζο, για την καθοδήγησή του και τις πολύτιμες παρατηρήσεις του καθ' όλη τη διάρκεια της εργασίας μου. Επίσης, ευχαριστώ από καρδιάς τους φίλους μου την οικογένειά μου και την κοπέλα μου για την αγάπη και την αμέριστη υποστήριξη που μου προσέφεραν καθ' όλη τη διάρκεια των σπουδών μου.

ΠΕΡΙΛΗΨΗ

Η μηχανική μάθηση έχει εισέλθει δυναμικά στην καθημερινή μας ζωή τα τελευταία χρόνια και αποδεικνύεται ως ένα ισχυρό εργαλείο με πολλές εφαρμογές. Καθώς οι χρηματοοικονομικές αγορές είναι εκτεθειμένες σε διάφορους παράγοντες κινδύνου που μπορούν να προκαλέσουν ανωμαλίες στα δεδομένα, η ανίχνευση ανωμαλιών αποτελεί κρίσιμο εργαλείο για τους επενδυτές, τους τραπεζίτες και άλλους φορείς της αγοράς. Μέσω της μηχανικής μάθησης, μπορούμε να εκπαιδεύσουμε μοντέλα που είναι σε θέση να αναγνωρίζουν μοτίβα και τάσεις στις χρηματοοικονομικές χρονοσειρές. Όταν οι χρονοσειρές παρουσιάζουν ανωμαλίες ή απρόβλεπτες συμπεριφορές, αυτά τα μοντέλα μπορούν να εντοπίσουν τα σημεία αυτά και να εκδώσουν συναφείς προειδοποιήσεις ή σήματα.

Περιεχόμενα

Κατάλογος Εικόνων

1. ΕΙΣΑΓΩΓΗ

Ο αυτοματισμός εισέρχεται σε όλα τα πεδία με έναν πρωτόγνωρο τρόπο, και ο χώρος των οικονομικών δεν αποτελεί εξαίρεση. Στη σύγχρονη εποχή, στον οικονομικό τομέα χρησιμοποιούνται πολλά μοντέλα για την τιμολόγηση χιλιάδων κεφαλαίων. Στις συναλλαγές, μια ανωμαλία μπορεί να οφείλεται σε ασυνήθιστη δραστηριότητα κάποιου πελάτη, όπως μεγάλος όγκος συναλλαγών ή απροσδόκητες κινήσεις κεφαλαίων. Στην περίπτωση τιμών μετοχών ή κρυπτονομισμάτων, οι ανωμαλίες μπορεί να προκληθούν από μοχλεύσεις μεγάλων παικτών της αγοράς ή από ξαφνικά γεγονότα όπως πόλεμοι, σκάνδαλα μεγάλων εταιρειών ή άλλες απρόβλεπτες οικονομικές εξελίξεις. Οι συγκεκριμένοι παράγοντες δεν μπορούν πάντα να ενσωματωθούν πλήρως στα μαθηματικά μοντέλα ή τους αλγορίθμους μηχανικής μάθησης, καθώς είναι συχνά απρόβλεπτοι και εξαρτώνται από την ανθρώπινη συμπεριφορά ή τις παγκόσμιες συνθήκες. Ως εκ τούτου, οι ανωμαλίες αυτές αντικατοπτρίζουν τη δυναμική και πολύπλοκη φύση των χρηματοοικονομικών αγορών.

Ως ανωμαλία ορίζουμε ένα σημείο δεδομένων που βρίσκεται εκτός της αναμενόμενης κατανομής. Η αναμενόμενη κατανομή είναι είτε μια συνάρτηση των προηγούμενων τιμών για το ίδιο σημείο δεδομένων, είτε μια συνάρτηση άλλων μεταβλητών με αιτιώδεις σχέσεις. Αυτές οι ανωμαλίες μπορούν να έχουν σημαντικές επιπτώσεις στην επιχείρηση.

Επομένως υπάρχει ανάγκη για ένα σύστημα ανίχνευσης ανωμαλιών που θα εντοπίζει πιθανά σφάλματα στους υπολογισμούς ή ακόμα και ένα αυτοματοποιημένο σύστημα που θα μπορούσε ενδεχομένως να παγώσει τις συναλλαγές σε περίπτωση ανίχνευσης ανωμαλιών. Σήμερα, υπάρχει πρόσβαση σε στατιστικές μεθόδους που φιλτράρουν τις πιθανώς ανώμαλες τιμές. Ωστόσο, η αποτελεσματικότητα αυτών των μεθόδων είναι πολύ χαμηλή.

Η ανίχνευση ανωμαλιών σε χρονοσειρές είναι σημαντική επειδή επιτρέπει την αναγνώριση μοτίβων και τάσεων που δεν είναι εμφανή στο μάτι και μπορεί να υποδεικνύουν προβλήματα ή κινδύνους, ιδίως στην οικονομική πτυχή, όπου μπορεί να σηματοδοτούν σοβαρές οικονομικές ανισορροπίες. Η ανίχνευση ανωμαλιών συνήθως πραγματοποιείται μέσω στατιστικών μοντέλων που αποτυπώνουν τάσεις, εποχικότητα και επίπεδα σε δεδομένα χρονοσειρών. Μια τιμή στη χρονοσειρά θεωρείται μη φυσιολογική όταν η συμπεριφορά της διαφέρει σημαντικά από τις υπόλοιπες τιμές της χρονοσειράς. Σε αυτήν τη μελέτη, επικεντρωνόμαστε στην

ανίχνευση ανωμαλιών στις χρηματοοικονομικές χρονοσειρές με μεθόδους μηχανικής μάθησης.

Σε αυτή την εργασία, διερευνούμε την εφαρμογή μεθόδων μηχανικής μάθησης για ανίχνευση ανωμαλιών σε οικονομικές χρονοσειρές, με στόχο τη βελτίωση της ακρίβειας και της αποτελεσματικότητας ανίχνευσης. Διερευνούμε διάφορες τεχνικές μηχανικής εκμάθησης, συμπεριλαμβανομένων εποπτευόμενων, μη εποπτευόμενων και ημι-εποπτευόμενων προσεγγίσεων, όπως μηχανές υποστήριξης διανυσμάτων, νευρωνικά δίκτυα, αλγόριθμους ομαδοποίησης και μεθόδους συνόλου. Επιπλέον, διερευνούμε την ενοποίηση αρχιτεκτονικών βαθιάς μάθησης, όπως τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN) και τα συνελικτικά νευρωνικά δίκτυα (CNN), για την καταγραφή χρονικών εξαρτήσεων και χωρικών χαρακτηριστικών σε δεδομένα οικονομικών χρονοσειρών. Συγκεκριμένα, η ανάλυσή μας επικεντρώνεται σε χρονοσειρές τιμών αγαθών όπως μετοχές ή κρυπτονομίσματα, όπως το Bitcoin, με στόχο την ανίχνευση ανωμαλιών που μπορεί να υποδεικνύουν σημαντικές οικονομικές αλλαγές ή δραστηριότητες στην αγορά.

Οι προκλήσεις που σχετίζονται με την προεπεξεργασία δεδομένων, την εξαγωγή επιλεγμένων χαρακτηριστικών από τα δεδομένα, την επιλογή μοντέλου και τις μεθόδους αξιολόγησης της επίδοσης εξετάζονται με στόχο την ισχυρή και αξιόπιστη ανίχνευση ανωμαλιών. Μέσα από εμπειρικές μελέτες και μελέτες περιπτώσεων, καταδεικνύουμε την αποτελεσματικότητα και τους περιορισμούς των μεθόδων ανίχνευσης ανωμαλιών που βασίζονται στη μηχανική μάθηση. Αυτή η εργασία ασχολείται με την εφαρμογή τεχνικών μηχανικής μάθησης για τον εντοπισμό ανωμαλιών σε χρηματοοικονομικές χρονοσειρές, παρέχοντας πληροφορίες για τις βέλτιστες πρακτικές, τις μελλοντικές κατευθύνσεις έρευνας και τις πιθανές εφαρμογές στον χρηματοοικονομικό κλάδο.

1.1 Χρηματοοικονομικές Χρονοσειρές

Μια χρονοσειρά είναι ένα σύνολο παρατηρήσεων μιας διαδικασίας που πραγματοποιούνται σειριακά στο χρόνο. Η ανάλυσή της είναι απαραίτητη σε μια ευρεία γκάμα θεμάτων έρευνας στη μηχανική, την ιατρική, τις οικονομικές και άλλες επιστημονικές περιοχές. Ο κύριος στόχος που επιδιώκεται με την ανάλυση των χρονοσειρών είναι η πρόβλεψη μελλοντικών τιμών βασισμένη σε προηγούμενες

παρατηρήσεις. Μεταξύ άλλων εφικτών στόχων περιλαμβάνονται η περιγραφή, η εξήγηση, ο έλεγχος ή η ανίχνευση ανωμαλιών. Η ειδική περίπτωση ανίχνευσης ανωμαλιών αναφέρεται στην αναζήτηση προτύπων με ασυνήθιστη συμπεριφορά, τα οποία μπορούν να ερμηνευτούν ως μη έγκυρες ή ανώμαλες ενέργειες στα δεδομένα.
Η κλασική στατιστική ανάλυση των χρονοσειρών περιλαμβάνει την εκτίμηση σημαντικών στατιστικών μεγεθών από μοτίβα που τυπικά εμφανίζουν: μη σταθερότητα, αυτοσυσχέτιση και εποχικότητα. Στην περίπτωση της ανίχνευσης ανωμαλιών, η χρήση παραδοσιακών στατιστικών τεχνικών έχει παράγει ικανοποιητικά αποτελέσματα. Ωστόσο, τα τελευταία χρόνια έχουν αναφερθεί αποτελέσματα μεγαλύτερης σημασίας μέσω της χρήσης Τεχνητής Νοημοσύνης, και ειδικότερα μεθόδων μηχανικής μάθησης (Machine Learning). Εφαρμοσμένη στα χρηματοοικονομικά συστήματα, η ανίχνευση ανωμαλιών με χρήση μηχανικής μάθησης έχει επιτρέψει την αναγνώριση και πρόληψη κακόβουλων δραστηριοτήτων όπως απάτες και εισβολές, μεταξύ άλλων και παράνομων δραστηριοτήτων.

1.2 Απαιτήσεις Δεδομένων για Ανάλυση Χρονοσειρών

Για να πραγματοποιηθεί η ανάλυση με επιτυχία, είναι ζωτικής σημασίας να διαθέτουμε δεδομένα που πληρούν συγκεκριμένα κριτήρια. Αρχικά, πρέπει να υπάρχει ένας επαρκής όγκος δεδομένων που να καταγράφουν τη μεταβολή της μεταβλητής με την πάροδο του χρόνου. Η ποσότητα αυτών των δεδομένων εξαρτάται από την προβλεπόμενη ανάλυση και τη συχνότητα συλλογής. Για να εξεταστούν αποτελεσματικά οι τάσεις, απαιτείται ένα μεγάλο σύνολο δεδομένων που καλύπτει ένα εκτεταμένο χρονικό διάστημα. Επίσης, η ανάλυση της εποχικότητας απαιτεί τουλάχιστον έναν πλήρη κύκλο παρατηρήσεων που να περιλαμβάνει όλες τις περιόδους. Ακόμα, για την εξέταση του θορύβου, απαιτούνται τουλάχιστον είκοσι δείγματα δεδομένων που να παρουσιάζουν επαρκή μεταβλητότητα.
Δεύτερον, είναι απαραίτητο να διαθέτουμε δεδομένα που καταγράφουν με ακρίβεια τις χρονικές αλλαγές της μεταβλητής. Αυτό περιλαμβάνει τη συλλογή δεδομένων σε σταθερά χρονικά διαστήματα, χωρίς παραλείψεις ή επικαλύψεις παρατηρήσεων, καθώς και το χρονικό ευθυγράμμισή τους.

Τέλος, για την ανάλυση των πρωταρχικών στοιχείων μιας χρονοσειράς (τάση, εποχικότητα, θόρυβος), τα δεδομένα πρέπει να πληρούν τα κριτήρια του επιλεγμένου αναλυτικού μοντέλου.

Με την τήρηση αυτών των απαιτήσεων και την εφαρμογή των κατάλληλων αλγορίθμων και μεθόδων ανάλυσης, μπορεί να πραγματοποιηθεί μια αξιόπιστη ανάλυση χρονοσειρών και ένας αποτελεσματικός εντοπισμός ανωμαλιών.

Οι χρονοσειρές έχουν τυπικά χαρακτηριστικά που περιγράφουν με ακρίβεια τη φύση της χρονοσειράς:

Περίοδος: ένα χρονικό διάστημα σταθερού μήκους για ολόκληρη τη σειρά, στα άκρα του οποίου η σειρά παίρνει κοντινές τιμές,

Εποχικότητα: η ιδιότητα της περιοδικότητας,

Κύκλος: χαρακτηριστικές αλλαγές σε μια σειρά που σχετίζονται με παγκόσμιες αιτίες (για παράδειγμα, κύκλοι στην οικονομία), στις οποίες δεν υπάρχει σταθερή περίοδος,

Τάση: μια τάση προς μια μακροπρόθεσμη αύξηση ή μείωση των τιμών μιας σειράς.

Οι αλγόριθμοι μηχανικής μάθησης για ανίχνευση ανωμαλιών χρησιμοποιούν δεδομένα σχετικά με τη λειτουργία της διαδικασίας (σύνολα δεδομένων). Ανάλογα με την θεματική περιοχή, το σύνολο δεδομένων μπορεί να περιέχει διαφορετικούς τύπους ανωμαλιών.

Οι χρονοσειρές συχνά χωρίζονται σε μονομεταβλητές (μονοδιάστατες) και πολυμεταβλητές (πολυδιάστατες). Αυτά τα δύο είδη ορίζονται στις ακόλουθες υποενότητες. Στη συνέχεια, παρουσιάζονται οι συνιστώσες που βρίσκονται εκτός της χρονοσειράς. Ακολουθεί μια ταξινόμηση των τύπων ανωμαλιών βασισμένη στις συνιστώσες και τα χαρακτηριστικά της χρονοσειράς.

1.3 Μονοδιάστατη Χρονοσειρά (Μιας Μεταβλητής)

Όπως υποδηλώνει το όνομά της, μια μονοδιάστατη χρονοσειρά είναι μια σειρά δεδομένων που βασίζεται σε μια μόνο μεταβλητή που μεταβάλλεται με την πάροδο του χρόνου. Το **X** με χρονικά σημεία **t** μπορεί να αναπαρασταθεί ως μια ταξινομημένη ακολουθία δεδομένων ως εξής: $X = (x_1, x_2, \ldots, x_t)$

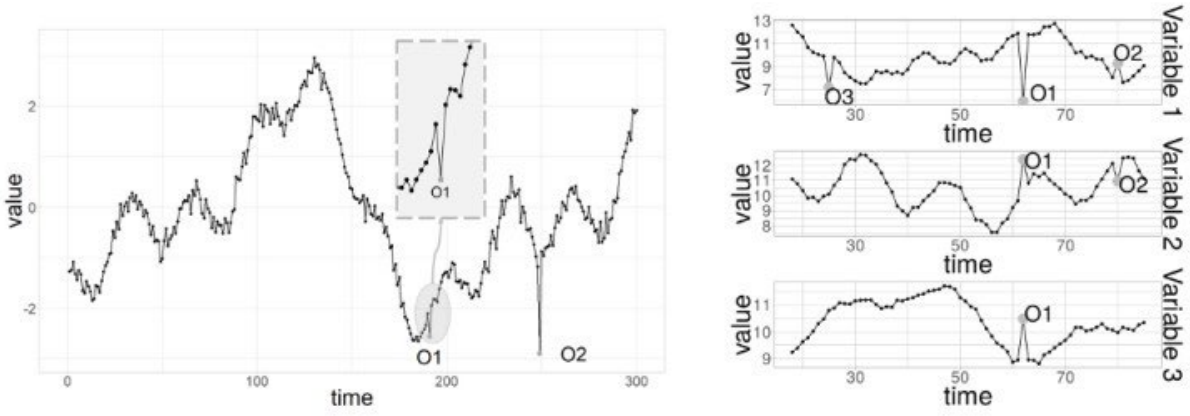
όπου $\mathbf{x}_i$ αντιπροσωπεύει τα δεδομένα στα χρονικά σημεία $i \in T$ και $T = \{1, 2, \ldots, t\}$.

1.4 Πολυδιάστατη Χρονοσειρά (Πολλών Μεταβλητών)

Επιπλέον, μια πολυδιάστατη χρονοσειρά αντιπροσωπεύει πολλές μεταβλητές που εξαρτώνται από τον χρόνο, καθεμία από τις οποίες επηρεάζεται τόσο από προηγούμενες τιμές (όπως αναφέρεται ως "στιγμιαία" εξάρτηση) όσο και από άλλες μεταβλητές (διαστάσεις) με βάση τη συσχέτισή τους. Οι συσχετίσεις μεταξύ διαφορετικών μεταβλητών αναφέρονται ως χωρικές ή διαμετρικές. Η πολυδιάστατη χρονοσειρά αναπαρίσταται ως ένα διάνυσμα $\mathbf{X_t}$ με διαστάσεις **d** ως εξής:

$Xt = (x_{1t}, x_{2t}, \ldots, x_{dt})$

όπου η **j**-οστή σειρά του **Xt** είναι $\mathbf{x_{jt}}$ που αντιπροσωπεύει τα δεδομένα για το χρονικό σημείο **t** για την **j**-οστή διάσταση, και $j = \{1, 2, ..., d\}$, όπου d είναι ο αριθμός των διαστάσεων.



Σχήμα 1.1 Πολυδιάστατη Χρονοσειρά [1]

Η εικόνα παρουσιάζει δύο διαφορετικές προσεγγίσεις ανίχνευσης ανωμαλιών:

Αριστερή Εικόνα - Μονοδιάστατη Χρονοσειρά:

- Η καμπύλη δείχνει τη συμπεριφορά μιας μονοδιάστατης χρονοσειράς με σημεία ανωμαλίας (**O1**, **O2**).
- Οι ανωμαλίες επισημαίνονται σε περιοχές όπου η συμπεριφορά της χρονοσειράς αποκλίνει από την κανονική τάση.
- Το **O1** μπορεί να αντιπροσωπεύει μια σημαντική απόκλιση, ενώ το **O2** εμφανίζεται ως μεμονωμένη ανωμαλία.

Δεξιά Εικόνα - Πολυδιάστατη Χρονοσειρά:

- Η εικόνα δείχνει τρεις διαφορετικές μεταβλητές (Variable 1, 2, 3), οι οποίες αναλύονται παράλληλα για την ανίχνευση ανωμαλιών.
- Οι ανωμαλίες (**O1**, **O2**, **O3**) εμφανίζονται σε διαφορετικές μεταβλητές, υποδεικνύοντας ότι η πολυδιάστατη ανάλυση επιτρέπει την ανίχνευση πιο σύνθετων αποκλίσεων.
- Το **O1** ανιχνεύεται σε δύο μεταβλητές, ενώ το **O2** είναι παρόν σε όλες, δείχνοντας συνδυαστική ανωμαλία που δεν θα ήταν εμφανής αν αναλυόταν μόνο μία μεταβλητή.

1.5 Ανάλυση Χρονοσειράς

Η ανάλυση της χρονοσειράς είναι ένας τρόπος ανάλυσης μιας χρονοσειράς ώστε να αναδείξει τις διάφορες συνιστώσες που συνθέτουν την κίνησή της. Κατά τη διάρκεια αυτής της διαδικασίας, η χρονοσειρά διασπάται σε τέσσερις βασικές συνιστώσες:

• **Κύρια τάση**: Οι τάσεις δεδομένων συμβαίνουν όταν υπάρχει μακροπρόθεσμη άνοδος ή πτώση. Η κύρια τάση αντιπροσωπεύει το γενικό πρότυπο των δεδομένων με την πάροδο του χρόνου και δεν χρειάζεται να είναι γραμμική.

• **Εποχικές παραλλαγές**: Ανάλογα με το μήνα, την ημέρα της εβδομάδας ή τη διάρκεια, μια χρονοσειρά μπορεί να εμφανίζει εποχιακό πρότυπο. Η εποχικότητα συμβαίνει πάντα σε σταθερή συχνότητα.

• **Κυκλικές διακυμάνσεις**: Ένας κύκλος ορίζεται ως αύξηση ή μείωση των δεδομένων χωρίς σταθερή συχνότητα. Επίσης, είναι γνωστό ως «η μορφή της χρονοσειράς».

• **Τυχαίες παραλλαγές**: Αναφέρεται σε τυχαία, ανεπίσημα γεγονότα. Είναι το υπόλοιπο μετά από την αφαίρεση όλων των άλλων συνιστωσών.

Η αναάλυση των χρονοσειρών επιτρέπει στους αναλυτές να κατανοήσουν και να μοντελοποιήσουν καλύτερα την κίνηση των δεδομένων και να προβλέπουν μελλοντικές τάσεις.

1.6 ΑΝΩΜΑΛΙΕΣ ΣΤΙΣ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΕΣ ΧΡΟΝΟΣΕΙΡΕΣ

Σε πολλούς τομείς όπως ο τομέας των χρηματοοικονομικών, όπου τα δεδομένα συλλέγονται ως χρονοσειρές, ο χρόνος αποτελεί ανεξάρτητη μεταβλητή και οι φυσικές ποσότητες που μετρούνται έναντι αυτού είναι εξαρτημένες μεταβλητές. Τα τελευταία χρόνια, οι ερευνητές ενδιαφέρονται όλο και περισσότερο για την ανάλυση ασυνήθιστων αλλά ενδιαφερόντων φαινομένων στα δεδομένα χρονοσειρών, όπως οι ανωμαλίες.

Στις χρηματοοικονομικές χρονοσειρές, η έννοια της ανωμαλίας αφορά οποιοδήποτε σημείο δεδομένων που ξεφεύγει σημαντικά από το αναμενόμενο πρότυπο συμπεριφοράς. Η διαδικασία ανίχνευσης ανωμαλιών αποτελεί έναν τρόπο για να κατανοήσουμε εάν ένα σύστημα συμπεριφέρεται με τον αναμενόμενο τρόπο ή εάν υπάρχουν παράξενες συμπεριφορές που απαιτούν περαιτέρω εξέταση από ειδικούς στον χώρο.

Στον χρηματοοικονομικό τομέα, είναι δύσκολο να οριστεί ακριβώς τι αποτελεί μια ανωμαλία. Δεν υπάρχουν συγκεκριμένες οδηγίες για το πώς φαίνεται μια ανωμαλία, καθώς υπάρχουν πολλοί παράγοντες που μπορούν να την προκαλέσουν. Μια συναλλαγή μπορεί να θεωρηθεί ανώμαλη εάν ξεπερνά το αναμενόμενο όγκο συναλλαγών ή εάν πραγματοποιείται εκτός των συνήθων χρονικών πλαισίων. Υπάρχουν πολλά σενάρια στα οποία μια συναλλαγή θα μπορούσε να θεωρηθεί ανώμαλη, αλλά δεν υπάρχει ένα καθολικά αποδεκτό όριο ή κατευθυντήριες γραμμές που ορίζουν τις ανωμαλίες.

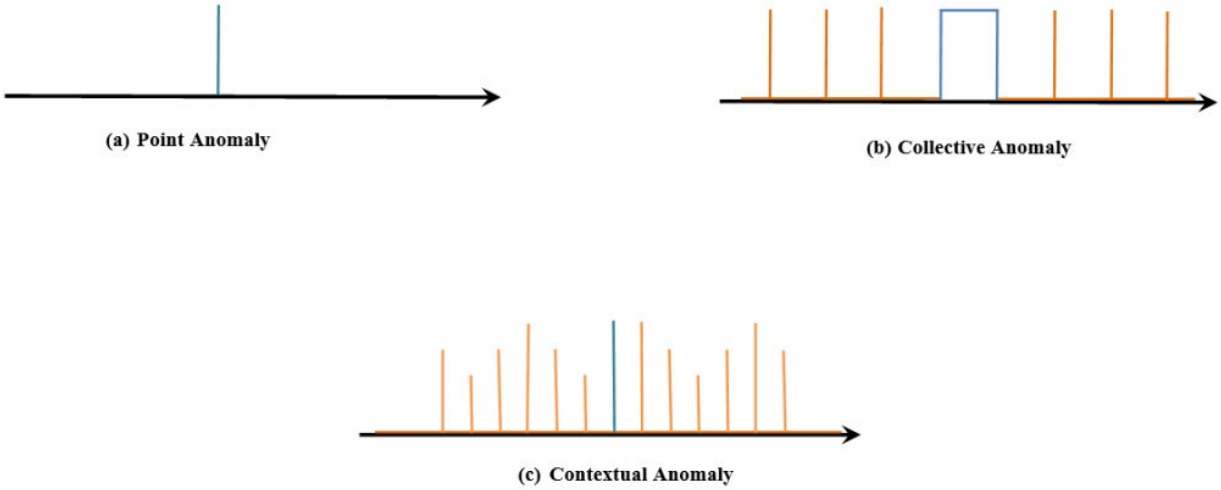
Τα τραπεζικά ιδρύματα που παρέχουν υπηρεσίες βασιζόμενες σε αυτόματα συστήματα τιμολόγησης απαιτούν αναλυτές κινδύνου για να παρακολουθούν τις τιμές και τις κινήσεις της αγοράς ενώ παράλληλα διασφαλίζουν ότι οι ανωμαλίες διορθώνονται το συντομότερο δυνατόν. Ωστόσο, η αύξηση του αριθμού των αυτόματων συστημάτων τιμολόγησης υπερβαίνει γρήγορα τον ρυθμό με τον οποίο οι αναλυτές μπορούν να επιβεβαιώνουν τα συστήματα. Σε ορισμένες περιπτώσεις, οι αναλυτές μπορεί να είναι υπεύθυνοι για την επιβεβαίωση έως και 100.000 σημείων τιμολόγησης ημερησίως, περίπου 4 σημεία δεδομένων ανά δευτερόλεπτο. Η χειροκίνητη επιβεβαίωση των τιμών εκτός από το γεγονός ότι είναι χρονοβόρα, απαιτεί και έναν ολοένα αυξανόμενο αριθμό πόρων.

1.7 Κίνητρα για τον εντοπισμό ανωμαλιών

Για να κατανοήσουν τα δεδομένα και να αντιμετωπίσουν τον μεγάλο όγκο πληροφοριών, οι εταιρείες συχνά αντιμετωπίζουν δυσκολίες στην πρόσληψη επαρκούς αριθμού αναλυτών λόγω των οικονομικών περιορισμών. Η αυξανόμενη πολυπλοκότητα των δεδομένων δημιουργεί ένα χάσμα μεταξύ του όγκου των πληροφοριών και του αριθμού των ανθρώπων που είναι διαθέσιμοι για να τα αναλύσουν. Αυτό μπορεί να δημιουργήσει ένα πρόβλημα συγχρονισμού, καθώς οι εταιρείες αναζητούν τρόπους για τη διαχείριση του όγκου των δεδομένων.
Μια λύση για αυτό το πρόβλημα είναι η αξιοποίηση των υπολογιστικών συστημάτων. Τα προγράμματα μηχανικής μάθησης μπορούν να αναλύουν τα δεδομένα και να ανιχνεύουν ανωμαλίες με πολύ μεγαλύτερη αποτελεσματικότητα από ότι η ανθρώπινη επιθεώρηση. Αυτά τα συστήματα μπορούν να παράγουν προειδοποιήσεις και σήμανση ανωμαλιών, ενώ οι ειδικοί μπορούν στη συνέχεια να διερευνήσουν και να αντιμετωπίσουν τα προβλήματα ανάλογα με την προτεραιότητά τους. Με αυτόν τον τρόπο μειώνεται η ανάγκη για ανθρώπινη επέμβαση σε κάθε σημείο και η διαχείριση του όγκου των δεδομένων γίνεται αποτελεσματικότερη.
Για να επιτευχθεί αυτό, εκπαιδεύονται μοντέλα μηχανικής μάθησης με στόχο την ανίχνευση ανωμαλιών στις χρηματοοικονομικές χρονοσειρές. Αυτά τα μοντέλα εκπαιδεύονται να αναγνωρίζουν τα πρότυπα και τις αποκλίσεις που μπορούν να υποδείξουν πιθανές ανωμαλίες. Η χρήση μηχανικής μάθησης σε αυτό το πλαίσιο μπορεί να είναι πολύ αποτελεσματική και να βοηθήσει τις εταιρείες να αντιμετωπίσουν την πρόκληση του ραγδαίου ρυθμού αύξησης των δεδομένων.

1.8 Ανωμαλίες

Οι βασικότερες ανωμαλίες που εντοπίζονται στις χρονοσειρές εντάσσονται στις εξής κατηγορίες: (point, collective and contextual anomalies).



Σχήμα 1.2 Ανωμαλίες [2]

1.8.1 Ανωμαλία σημείου

Η ανωμαλία σημείων είναι ένα σημείο δεδομένων ή μια ακολουθία που αποκλίνει απότομα από τον κανόνα (Εικ. 1(α)). Τέτοιες ανωμαλίες τείνουν να φαίνονται ως προσωρινός θόρυβος και συχνά προκαλούνται από σφάλματα αισθητήρα ή μη φυσιολογικές λειτουργίες του συστήματος. Για την ανίχνευση, οι χειριστές ορίζουν παραδοσιακά ανώτατα και κατώτερα όρια ελέγχου, που συνήθως αναφέρονται ως UCL και LCL, αντίστοιχα, με βάση προηγούμενα δεδομένα. Οι τιμές που υπάρχουν εκτός αυτών των ορίων θεωρούνται ως ανωμαλίες σημείου.

1.8.2 Ανωμαλία πλαισίου

Παρόμοια με μια ανωμαλία σημείου, μια ομαδική ανωμαλία αντιπροσωπεύει ένα σημείο δεδομένων ή μια ακολουθία που παρατηρείται σε σύντομο χρονικό διάστημα, αλλά δεν αποκλίνει από το κανονικό εύρος με τον ίδιο τρόπο όπως οι προκαθορισμένες ανωμαλίες που οριοθετούνται με **UCL** και **LCL**. Ωστόσο, λαμβάνοντας υπόψη το δεδομένο πλαίσιο (Εικ. 1(β)), τα σημεία δεδομένων είναι εκτός του αναμενόμενου σχεδίου ή σχήματος, γεγονός το οποίο καθιστά δύσκολη την ανίχνευση των συγκεκριμενών ανωμαλιών.

1.8.3 Ανωμαλία Ομάδας

Αυτός ο τύπος ανωμαλίας αναφέρεται σε ένα σύνολο σημείων δεδομένων που θα πρέπει να θεωρηθούν ως ανωμαλία επειδή σταδιακά εμφανίζουν διαφορετικό μοτίβο από τα κανονικά δεδομένα με την πάροδο του χρόνου (Εικ. 1(γ)). Οι μεμονωμένες αξίες σε αυτόν τον τύπο ανωμαλίας μπορεί να φαίνονται χωρίς προβλήματα, αλλά συλλογικά, εγείρουν υποψίες. Δεδομένου ότι δεν είναι εύκολα αναγνωρίσιμα αμέσως, τα περιβάλλοντα μακροπρόθεσμα έχουν ιδιαίτερη σημασία για την ανίχνευσή τους. Εφόσον η ανωμαλία δεν συμπίπτει με την κανονική κατάσταση, το τι ορίζουμε ως μη φυσιολογικό απορρέει από το τι ορίζουμε ως φυσιολογικό. Σε γενικές γραμμές, οι ανωμαλίες μπορούν να ταξινομηθούν σε έναν από τους τρεις προαναφερθέντες τύπους, χωρίς όμως να περιορίζονται μόνο σε αυτές τις κατηγορίες.

Η κύρια εργασία στην ανίχνευση ανωμαλιών είναι ο διαχωρισμός της κανονικής συμπεριφοράς από την ανώμαλη. Ωστόσο, οι ανώμαλες περιπτώσεις δεδομένων συνήθως αποτελούν μόνο ένα πολύ μικρό κομμάτι του συνόλου των δεδομένων. Επομένως, αντίθετα με ένα τυπικό πρόβλημα ταξινόμησης, όπου αναμένεται να είναι ισορροπημένες και οι δύο κατηγορίες, η ανίχνευση ανωμαλιών αντιπροσωπεύει ένα πρόβλημα, στο οποίο οι κατηγορίες δεν είναι ισόποσες.

Οι αλγόριθμοι ανίχνευσης ανωμαλιών λειτουργούν με μια εναλλακτική προσέγγιση, όπου τα μοντέλα εκπαιδεύονται πρώτα για να υπολογίζουν τους βαθμούς κάθε δείγματος δεδομένων, και στη συνέχεια τα δείγματα δεδομένων που λαμβάνουν τους υψηλότερους βαθμούς αναφέρονται ως ανωμαλίες. Ωστόσο, πολλές από αυτές τις ανωμαλίες που αναφέρονται από τους αλγόριθμους ανίχνευσης ανωμαλιών μπορεί να είναι ψευδής, διότι προκύπτουν από δείγματα δεδομένων που δεν ταιριάζουν σε ένα κανονικό μοντέλο. Επιπλέον, τα δεδομένα συχνά συσσωρεύουν θόρυβο λόγω της μεταβλητότητας που εμπλέκεται στη δημιουργία, συλλογή και επεξεργασία τους, γεγονός το οποίο δυσκολεύει περαιτέρω το πρόβλημα της ανίχνευσης των πραγματικών ανωμαλιών, προκαλώντας συχνά περισσότερες ψευδείς ανωμαλίες.

Οι περισσότερες τεχνικές ανίχνευσης ανωμαλιών είναι εξειδικευμένες για συγκεκριμένους τομείς. Για παράδειγμα, τεχνικές που αναπτύσσονται ειδικά για την ανίχνευση απάτης με πιστωτικές κάρτες μπορεί να μην ανιχνεύουν ανωμαλίες στην αγορά μετοχών. Σε ορισμένους τομείς, η κανονική συμπεριφορά εξελίσσεται συνεχώς, με αποτέλεσμα μια τρέχουσα προδιαγραφή της κανονικής συμπεριφοράς να μην είναι

ισχύουσα για τη μελλοντική ανίχνευση ανωμαλιών. Η σειρά των ενεργειών, το χρονικό διάστημα και η σειρά μεταξύ τους έχουν σημασία. Επομένως, οι πληροφορίες σειριακής φύσης που ενσωματώνονται στην τάξη και το χρονικό διάστημα των ενεργειών της αγοράς πρέπει να αιχμαλωτίζονται και να λαμβάνονται υπόψη από το σύστημα προκειμένου να κατηγοριοποιήσει τη συμπεριφορά της αγοράς ως ανωμαλία.

2. Ανίχνευση ανωμαλιών στις χρηματοοικονομικές χρονοσειρές

Ο προσδιορισμός του ορίου για την ανίχνευση ανωμαλιών βασίζεται σε διάφορους παράγοντες όπως τα επίπεδα εμπιστοσύνης, η κατανομή σφαλμάτων και η συχνότητα δεδομένων. Τυπικά, η έννοια του διαστήματος εμπιστοσύνης χρησιμοποιείται για τον καθορισμό του ορίου. Το διάστημα εμπιστοσύνης αντιπροσωπεύει ένα εύρος εντός του οποίου η προβλεπόμενη τιμή βρίσκεται με μια συγκεκριμένη πιθανότητα π.χ., ένα διάστημα εμπιστοσύνης σε ποσοστό 95% υποδεικνύει μια πιθανότητα σε ποσοστό 95% ότι η προβλεπόμενη τιμή εμπίπτει σε αυτό το εύρος. Οποιαδήποτε παρατηρούμενη τιμή εκτός αυτού του διαστήματος θεωρείται ανωμαλία.

Ωστόσο, η απλή ανίχνευση ανωμαλιών δεν αρκεί. Η κατανόηση των αιτιών και των συνεπειών τους είναι ζωτικής σημασίας. Ορισμένες ανωμαλίες, που κατηγοριοποιούνται ως θόρυβος, μπορεί να προέρχονται από σφάλματα μέτρησης ή δυσλειτουργίες στην επεξεργασία δεδομένων και μπορούν να αγνοηθούν ή να διορθωθούν. Από την άλλη πλευρά, οι ανωμαλίες που οφείλονται σε δομικές αλλαγές, κακόβουλες δραστηριότητες ή έκτακτα γεγονότα, τα οποία ονομάζονται σήματα, απαιτούν πιο προσεκτικό έλεγχο και ανάλυση.

Για την κατανόηση των αιτιών και των συνεπειών των ανωμαλιών, η γνώση του τομέα των χρηματοοικονομικών παίζει καθοριστικό ρόλο. Η εξοικείωση με το πλαίσιο των δεδομένων και τη σημασία των μεταβλητών τους είναι απαραίτητη. Επιπλέον, η αξιοποίηση συμπληρωματικών πηγών πληροφοριών, όπως οι σχετικές χρονοσειρές, τα ιστορικά δεδομένα, οι ειδήσεις και οι αναφορές, βοηθά στην ερμηνεία των αποτελεσμάτων ανίχνευσης ανωμαλιών και στην επινόηση κατάλληλων ενεργειών.

Ορισμένες από τις κύριες μεθόδους μηχανικής μάθησης που χρησιμοποιούνται για την ανίχνευση ανωμαλιών σε χρονοσειρές χρηματοοικονομικών δεδομένων περιλαμβάνουν:

Αυτόματη Ανίχνευση Εκτίμησης (Automatic Estimation Detection - AED): Αυτή η μέθοδος χρησιμοποιεί τεχνικές όπως οι χρονοσειρές ή οι εκτιμήσεις για να ανιχνεύσει ανωμαλίες στα δεδομένα.

Μοντέλα Μηχανικής Μάθησης όπως τα Νευρωνικά Δίκτυα (NN), τα αναδρομικά νευρωνικά δίκτυα (RNNs) και η Μακρά Βραχυπρόθεσμη Μνήμη (LSTM), είναι ισχυρά

εργαλεία για την ανάλυση χρονοσειρών. Είναι ικανά να καταγράφουν μακροπρόθεσμες εξαρτήσεις σε δεδομένα και χρησιμοποιούνται για πρόβλεψη, ανίχνευση ανωμαλιών, ταξινόμηση και δημιουργία χρονοσειρών.

Αλγόριθμοι Ανίχνευσης Ανωμαλιών: Οι αλγόριθμοι Isolation Forest και οι αλγόριθμοι βασισμένοι στην αρχή της παραγωγής (GANs) χρησιμοποιούνται για τη δημιουργία συνθετικών δεδομένων και για την ανάλυση και την πρόβλεψη μακροπρόθεσμων τάσεων της αγοράς.

Συχνά, ο συνδυασμός των προαναφερθέντων μεθόδων μπορεί να οδηγήσει σε βελτιωμένες επιδόσεις στην ανίχνευση ανωμαλιών σε χρονοσειρές χρηματοοικονομικών δεδομένων. Επιπλέον, η επιτυχία αυτών των μεθόδων εξαρτάται σε μεγάλο βαθμό από την ποιότητα και την κατάλληλη προεπεξεργασία των δεδομένων.

2.1 Ιστορική αναδρομή

Η ανίχνευση ανωμαλιών στις χρηματοοικονομικές χρονοσειρές είναι κρίσιμο ζήτημα για τους οικονομολόγους, τους αναλυτές και τους επενδυτές, καθώς οι ανωμαλίες μπορούν να υποδηλώνουν σημαντικά γεγονότα όπως κρίσεις, απάτες ή μεγάλες αλλαγές στις συνθήκες της αγοράς. Η χρήση μεθόδων μηχανικής μάθησης για την ανίχνευση αυτών των ανωμαλιών έχει εξελιχθεί σημαντικά τις τελευταίες δεκαετίες.

2.1.1 Πρώτες Προσεγγίσεις και ARIMA

Οι πρώτες προσπάθειες για την ανίχνευση ανωμαλιών στις χρηματοοικονομικές χρονοσειρές επικεντρώνονταν κυρίως σε στατιστικές μεθόδους. Η μέθοδος ARIMA (AutoRegressive Integrated Moving Average) ήταν μία από τις πρώτες και πιο διαδεδομένες προσεγγίσεις. Το μοντέλο ARIMA, το οποίο αναπτύχθηκε από τους Box και Jenkins στη δεκαετία του 1970, είναι ένα από τα πιο γνωστά και χρησιμοποιούμενα μοντέλα για την ανάλυση χρονοσειρών.

Το μοντέλο ARIMA μπορεί να μοντελοποιήσει στατικές και μη στατικές χρονοσειρές χρησιμοποιώντας συνδυασμό αυτοπαλινδρόμησης (AR), ολοκλήρωσης (I) και κινούμενου μέσου (MA). Αν και το ARIMA είναι ισχυρό στην ανάλυση χρονοσειρών, η ικανότητά του να ανιχνεύει ανωμαλίες εξαρτάται από τη σωστή επιλογή των

παραμέτρων και την υποκειμενική κρίση του αναλυτή για τον καθορισμό των ορίων ανωμαλίας. Στις πρώτες εφαρμογές, τα υπολείμματα του μοντέλου (residuals) χρησιμοποιούνταν για την ανίχνευση ανωμαλιών, με τις σημαντικές αποκλίσεις από το μοντέλο να θεωρούνται ανωμαλίες.

2.1.2 Ανάπτυξη Μεθόδων Μηχανικής Μάθησης

Με την ανάπτυξη της τεχνολογίας και την αύξηση της υπολογιστικής ισχύος, η μηχανική μάθηση άρχισε να εφαρμόζεται ευρέως στην ανίχνευση ανωμαλιών. Οι μέθοδοι μηχανικής μάθησης μπορούν να επεξεργάζονται μεγάλα σύνολα δεδομένων και να ανακαλύπτουν περίπλοκα μοτίβα που δεν είναι εμφανή με παραδοσιακές στατιστικές μεθόδους.

ARIMA: Χρησιμοποιείται συχνά για την ανάλυση χρονοσειρών που εμφανίζουν γραμμικά πρότυπα και εποχικότητα. Έχει χρησιμοποιηθεί εκτενώς για την πρόβλεψη χρηματοοικονομικών δεδομένων, όπως η τιμή των μετοχών ή του Bitcoin, λόγω της ικανότητάς του να εντοπίζει γραμμικές σχέσεις μεταξύ παρελθοντικών και μελλοντικών τιμών.

SVM (Support Vector Machines): Το SVM χρησιμοποιείται συχνά για την ανίχνευση ανωμαλιών, καθώς μπορεί να κατηγοριοποιήσει δεδομένα σε μη γραμμικές κατηγορίες με υψηλή ακρίβεια. Στην περίπτωση της ανάλυσης χρονοσειρών, το SVM μπορεί να βοηθήσει στον εντοπισμό περιόδων που αποκλίνουν σημαντικά από τα φυσιολογικά πρότυπα, ενισχύοντας την ικανότητα του συστήματος να ανιχνεύει και να διαχωρίζει ανωμαλίες από τις κανονικές τάσεις.

LSTM: Το LSTM είναι πολύ αποτελεσματικό στην ανάλυση μακροχρόνιων εξαρτήσεων και στην ανίχνευση μη γραμμικών προτύπων. Έχει χρησιμοποιηθεί ευρέως για την πρόβλεψη τιμών σε δυναμικά περιβάλλοντα, όπως η αγορά κρυπτονομισμάτων και άλλες αγορές με υψηλή μεταβλητότητα. Για παράδειγμα, υπάρχουν αρκετές μελέτες που χρησιμοποιούν LSTM για την πρόβλεψη της τιμής του Bitcoin ή των μετοχών, καθώς το μοντέλο είναι εξαιρετικά ικανό στην ανίχνευση απότομων αλλαγών και ανωμαλιών.

2.1.3 Συνδυασμός Στατιστικών και Μηχανικής Μάθησης

Η ενσωμάτωση μεθόδων μηχανικής μάθησης όπως το k-NN με παραδοσιακά στατιστικά μοντέλα όπως το ARIMA έχει οδηγήσει σε πιο αποτελεσματικές προσεγγίσεις ανίχνευσης ανωμαλιών. Αυτές οι υβριδικές μέθοδοι εκμεταλλεύονται τα πλεονεκτήματα και των δύο προσεγγίσεων. Για παράδειγμα, οι Goh και Lee (2018) προτείνουν τη χρήση ενός υβριδικού μοντέλου που συνδυάζει ARIMA με k-NN για την ανίχνευση ανωμαλιών σε χρηματοοικονομικές χρονοσειρές, όπου το ARIMA χρησιμοποιείται για τη μοντελοποίηση της χρονοσειράς και τα υπολείμματα του μοντέλου αναλύονται με το k-NN για την ανίχνευση ανωμαλιών.

2.1.4 Σύγχρονες Εξελίξεις και Εφαρμογές

Η σύγχρονη έρευνα στην ανίχνευση ανωμαλιών συνεχίζει να εξελίσσεται με τη χρήση πιο προχωρημένων μεθόδων μηχανικής μάθησης και βαθιάς μάθησης (deep learning). Μέθοδοι όπως οι Autoencoders, οι Long Short-Term Memory (LSTM) και οι Convolutional Neural Networks (CNN) έχουν αρχίσει να χρησιμοποιούνται για την ανάλυση χρηματοοικονομικών χρονοσειρών και την ανίχνευση ανωμαλιών. Αυτές οι μέθοδοι μπορούν να χειριστούν πολύπλοκες και μη γραμμικές σχέσεις στα δεδομένα, προσφέροντας υψηλότερη ακρίβεια στην ανίχνευση ανωμαλιών.
Για παράδειγμα, οι Malhotra et al. (2015) ανέπτυξαν ένα μοντέλο LSTM για την ανίχνευση ανωμαλιών σε χρονοσειρές, το οποίο έχει αποδειχθεί πολύ αποτελεσματικό στην ανίχνευση ανωμαλιών σε πραγματικό χρόνο. Παρόμοια, οι Zhao et al. (2019) πρότειναν τη χρήση των Autoencoders για την ανίχνευση ανωμαλιών στις χρονοσειρές με βάση τη δυνατότητά τους να αναπαριστούν τα δεδομένα σε χαμηλότερη διάσταση και να εντοπίζουν αποκλίσεις από την κανονική συμπεριφορά.

2.2 Διαφόριση(differencing) στην Ανάλυση Χρονοσειρών

Μία πρωταρχική πρόκληση στην ανάλυση χρονοσειρών είναι το γεγονός ότι τα δεδομένα παρουσιάζουν αστάθεια, όπου οι στατιστικές ιδιότητες όπως ο μέσος όρος και η διακύμανση υφίστανται διακυμάνσεις με την πάροδο του χρόνου. Αυτό το χαρακτηριστικό περιπλέκει την εφαρμογή των παραδοσιακών στατιστικών μεθόδων,

οι οποίες συνήθως προϋποθέτουν τη σταθερότητα των δεδομένων. Για την αποτελεσματική χρήση αυτών των μεθόδων, είναι επιτακτική ανάγκη να μετασχηματιστούν τα δεδομένα για να επιτευχθεί σταθερότητα ή τουλάχιστον να γίνει κατά προσέγγιση. Μια τεχνική που χρησιμοποιείται συνήθως για το σκοπό αυτό είναι η διαφόριση.

Η διαφόριση περιλαμβάνει την αφαίρεση της προηγούμενης τιμής από κάθε τιμή της χρονοσειράς, δίνοντας μια νέα σειρά που απεικονίζει την αλλαγή στα δεδομένα με την πάροδο του χρόνου. Για παράδειγμα, με δεδομένη μια χρονοσειρά $\{x_1, x_2, x_3, \ldots\}$, η πρώτη της διαφορά θα ήταν $\{x_2 - x_1, x_3 - x_2, \ldots\}$. Αυτή η διαδικασία μπορεί να επαναληφθεί πολλές φορές για να ληφθεί η δεύτερη διαφορά, η τρίτη διαφορά και ούτω καθεξής. Ο στόχος της διαφόριση είναι να εξαλειφθούν οι συνιστώσες της τάσης και της εποχικότητας από τις χρονοσειρές, καθώς αυτές είναι οι κύριοι παράγοντες που συνεισφέρουν στη μη σταθερότητα. Συγκεκριμένα, εάν τα δεδομένα παρουσιάζουν τάση ή εποχικότητα, οι τιμές τους θα καταδεικνύουν συσχέτιση με προηγούμενες ή μεταγενέστερες τιμές. Με την αφαίρεση αυτών των τιμών, αυτή η συσχέτιση μειώνεται ή εξαλείφεται.

Για παράδειγμα, μια χρονοσειρά που δείχνει μια αυξητική τάση και ετήσια εποχικότητα. Η πρώτη της διαφορά θα εξαλείφει αποτελεσματικά την τάση, αφήνοντας ανέπαφη την εποχικότητα. Στη συνέχεια, η δεύτερη διαφορά θα εξαλείψει τόσο την τάση όσο και την εποχικότητα.

Η διαφόριση αποτελεί τη βάση ενός από τα πιο ευρέως χρησιμοποιούμενα μοντέλα για ανάλυση χρονοσειρών και ανίχνευση ανωμαλιών: το μοντέλο ARIMA, το οποίο θα αναλύσουμε στη συνέχεια.

2.3 Μοντέλο ARIMA

Τα μοντέλα ARIMA παρέχουν μια διαφορετική προσέγγιση στην πρόβλεψη χρονοσειρών από τα μοντέλα εκθετικής εξομάλυνσης. Και τα δύο είναι ευρέως χρησιμοποιούμενες προσεγγίσεις, αλλά λειτουργούν με διαφορετικούς τρόπους και εστιάζουν σε διαφορετικά χαρακτηριστικά των δεδομένων.

Τα μοντέλα εκθετικής εξομάλυνσης βασίζονται στην πρόβλεψη της τάσης και της εποχικότητας των δεδομένων. Αυτή η προσέγγιση είναι καλή για δεδομένα που έχουν έντονη τάση και εποχικότητα.

Από την άλλη πλευρά, τα μοντέλα ARIMA στοχεύουν στην περιγραφή της αυτοσυσχέτισης των δεδομένων, δηλαδή της συσχέτισης μεταξύ των τιμών της χρονοσειράς και των προηγούμενων τιμών. Αυτή η προσέγγιση είναι πιο κατάλληλη για χρονοσειρές που δεν έχουν έντονη τάση ή εποχικότητα, αλλά εμφανίζουν συσχέτιση μεταξύ των παρατηρήσεων τους.

2.3.1 Θεωρητικό υπόβαθρο

Το μοντέλο ARIMA (Autoregressive Integrated Moving Average) ξεχωρίζει ως ένα από τα κορυφαία εργαλεία ανάλυσης χρονοσειρών και ανίχνευσης ανωμαλιών. Αυτό το μοντέλο ενσωματώνει τρία βασικά στοιχεία:
1. Η συνιστώσα αυτό-παλίνδρομης (AR) καταγράφει τη συσχέτιση μεταξύ τιμών χρονοσειρών και προηγούμενων τιμών. Για παράδειγμα, στα κυκλικά δεδομένα, οι τρέχουσες τιμές επηρεάζονται από προηγούμενες.
2. Το ενσωματωμένο στοιχείο (I) αντιμετωπίζει τη μη σταθερότητα της χρονοσειράς διαφοροποιώντας την, αφαιρώντας τις τάσεις ή την εποχικότητα.
3. Το στοιχείο κινούμενου μέσου όρου (MA) δημιουργεί μοντέλα για τη συσχέτιση μεταξύ σφαλμάτων χρονοσειρών και προηγούμενων σφαλμάτων. Είναι ιδιαίτερα χρήσιμο στο χειρισμό του θορύβου στα δεδομένα.
Το μοντέλο ARIMA χαρακτηρίζεται από τρεις κύριες παραμέτρους: p, d και q. Η παράμετρος p υποδηλώνει τον αριθμό των αυτό-παλινδρομικών όρων, το d δηλώνει τον αριθμό των διαφοροποιήσεων για την επίτευξη σταθερότητας και το q αντιπροσωπεύει τον αριθμό των όρων κινητού μέσου όρου. Η γενική μορφή του μοντέλου ARIMA(p,d,q) είναι:

$$(1 – \varphi L - \varphi_2 L^2 - \ldots - \varphi_p L^p)(1 – L)dy_t = (1 + \theta L + \theta_2 L^2 + \ldots + \theta_q L^q)et$$

΄Οπου:

- **Yt** είναι η τιμή της χρονοσειράς στη στιγμή **t**
- **L** είναι ο τελεστής υστέρησης (lag operator), δηλαδή *LYt=Yt-1*
- Φ_i είναι οι παράμετροι του μέρους **AR**
- Θ_j είναι οι παράμετροι του μέρους **MA**
- **Et** είναι ο λευκός θόρυβος (white noise)

Για παράδειγμα, ένα μοντέλο ARIMA(1,1,1) ενσωματώνει έναν αυτό-παλινδρομικό όρο, μία διαφόριση και έναν όρο κινούμενου μέσου όρου.

2.3.2 Ανάλυση χρονοσειρών και ανίχνευση ανωμαλιών

Η χρήση του μοντέλου ARIMA για ανάλυση χρονοσειρών και ανίχνευση ανωμαλιών περιλαμβάνει τα εξής βήματα:

1. **Εκτίμηση της σταθερότητα**ς των χρονοσειρών χρησιμοποιώντας στατιστικές δοκιμές όπως η επαυξημένη δοκιμή Dickey-Fuller για να εξεταστεί εάν ο μέσος όρος και η διακύμανση παραμένουν σταθερές με την πάροδο του χρόνου.
2. **Επαναληπτική διαφόριση** της χρονοσειράς μέχρι να επιτύχει αστάθεια. Γραφικά εργαλεία όπως η αυτό-συσχέτιση και οι συναρτήσεις μερικής αυτό-συσχέτισης βοηθούν στον προσδιορισμό του αριθμού των απαιτούμενων διαφορών.
3. **Εκτίμηση των παραμέτρων του μοντέλου ARIMA** χρησιμοποιώντας τεχνικές βελτιστοποίησης όπως η εκτίμηση μέγιστης πιθανότητας. Τα κριτήρια επιλογής μοντέλου όπως το κριτήριο πληροφοριών Akaike ή το κριτήριο πληροφοριών Bayesian βοηθούν στην επιλογή βέλτιστων τιμών για τα p, d και q.
4. **Επικύρωση του μοντέλου ARIMA** μέσω μεθόδων επαλήθευσης όπως η δοκιμή Ljung-Box ή η δοκιμή Jarque-Bera. Η γραφική ανάλυση των υπολειμμάτων ή η ακρίβεια πρόβλεψης βοηθά στην αξιολόγηση της προσαρμογής του μοντέλου και στην ανίχνευση ανωμαλιών.
5. **Χρήση του μοντέλου ARIMA** για την περιγραφή χαρακτηριστικών χρονοσειρών, την πρόβλεψη μελλοντικών τιμών και τον εντοπισμό ανωμαλιών. Μετρήσεις αξιολόγησης όπως το μέσο τετράγωνο σφάλμα ή το μέσο απόλυτο μετρητή σφάλματος ακρίβεια πρόβλεψης και αποτελεσματικότητα ανίχνευσης ανωμαλιών.

2.4 Προβλήματα αναζήτησης ανωμαλιών σε χρονοσειρές

Ο εντοπισμός ανωμαλιών σε χρονοσειρές αντιμετωπίζει ποικίλες προκλήσεις λόγω της πολυπλοκότητας των δεδομένων και των ποικίλων παραγόντων που μπορεί να επηρεάσουν τις χρονοσειρές. Μερικά σημαντικά προβλήματα που σχετίζονται με τον εντοπισμό ανωμαλιών σε χρονοσειρές:

Υπάρχουν πολλοί τρόποι με τους οποίους μπορεί να εντοπιστεί μια ανωμαλία που εμφανίζεται σε μια χρονοσειρά. Ένα συμβάν μέσα σε μια χρονοσειρά μπορεί να είναι ανώμαλο, μια ακολουθία μέσα σε μια χρονοσειρά μπορεί να είναι ανώμαλη, ή ακόμα και μια ολόκληρη χρονοσειρά μπορεί να είναι ανώμαλη σε σχέση με ένα σύνολο κανονικών χρονοσειρών.

Για την ανίχνευση ανώμαλης ακολουθίας, το ακριβές μήκος της ακολουθίας είναι συχνά άγνωστο.

Οι χρονοσειρές για εκπαίδευση και δοκιμές μπορεί να έχουν διαφορετικά μήκη.

Οι καλύτερες μετρήσεις ομοιότητας/διαφοράς που μπορούν να χρησιμοποιηθούν για διαφορετικούς τύπους χρονοσειρών δεν είναι εύκολο να προσδιοριστούν. Απλά μέτρα όπως η Ευκλείδεια απόσταση δεν είναι πάντα αποτελεσματικά, καθώς είναι ευαίσθητα σε ακραίες τιμές και δεν ισχύουν όταν οι χρονοσειρές είναι διαφορετικού μήκους.

Η ακρίβεια πολλών αλγορίθμων ανίχνευσης ανωμαλιών επηρεάζεται σε μεγάλο βαθμό από το θόρυβο στα δεδομένα, καθώς ο διαχωρισμός των ανωμαλιών από τον θόρυβο είναι μια δύσκολη εργασία.

Οι χρονοσειρές σε πραγματικές εφαρμογές είναι συνήθως μεγάλες σε μέγεθος (μήκος), και όσο αυξάνεται το μήκος, αυξάνεται και η υπολογιστική πολυπλοκότητα.

Η ακρίβεια πολλών αλγορίθμων ανίχνευσης ανωμαλιών εξαρτάται σε μεγάλο βαθμό από τη συνέπεια των κλιμάκων χρονοσειρών, κάτι που δεν ισχύει για τα περισσότερα δεδομένα.

Τα παραπάνω παραδείγματα προσεγγίσεων ανίχνευσης ανωμαλιών καθιστούν σαφές πώς η επιλογή μεθόδου επηρεάζει την απόδοση του συστήματος ανίχνευσης ανωμαλιών. Η κατάλληλη επιλογή της μεθόδου εξαρτάται από τα χαρακτηριστικά των δεδομένων, τη διαθεσιμότητα ετικετών και το περιβάλλον εφαρμογής.

3. Μέθοδοι μηχανικής μάθησης για ανίχνευση ανωμαλιών σε χρηματοοικονομικές χρονοσειρές

Η μηχανική μάθηση είναι ένας τομέας της τεχνητής νοημοσύνης που εστιάζει στην ανάπτυξη αλγορίθμων και μοντέλων που επιτρέπουν στα συστήματα να μαθαίνουν από δεδομένα και εμπειρίες, χωρίς να απαιτείται ρητή προγραμματιστική παρέμβαση. Η βασική ιδέα είναι να επιτρέπεται στο σύστημα να βελτιώνει την απόδοσή του με την επεξεργασία των δεδομένων, την εξαγωγή προτύπων και την λήψη αποφάσεων, χωρίς να απαιτείται προγραμματισμός για κάθε συγκεκριμένη εργασία. Για τις προγνωστικές αναλύσεις η αποτελεσματικότητα των μοντέλων μηχανικής μάθησης εξαρτάται από την ποιοτική οργάνωση της συλλογής δεδομένων, της επεξεργασίας και της προκαταρκτικής ανάλυσης.

Ο εντοπισμός ανωμαλιών στα δεδομένα χρηματοοικονομικών χρονοσειρών διαδραματίζει κρίσιμο ρόλο στον εντοπισμό ακανόνιστων προτύπων, ακραίων τιμών και απροσδόκητων συμπεριφορών που αποκλίνουν από τις κανονικές συνθήκες της αγοράς. Οι παραδοσιακές στατιστικές μέθοδοι έχουν περιορισμούς στην αποτύπωση πολύπλοκων προτύπων και μη γραμμικών σχέσεων που υπάρχουν στα οικονομικά δεδομένα.

3.1 Μηχανική Μάθηση για Ανίχνευση Ανωμαλιών σε Χρηματοοικονομικές Χρονοσειρές

Η μηχανική μάθηση αποτελεί ένα πεδίο στα στατιστικά στοιχεία όπου τα προγράμματα χρησιμοποιούν ιστορικά δεδομένα για να κάνουν προβλέψεις για μελλοντικά σημεία δεδομένων ή άγνωστες ετικέτες. Ένα μοντέλο μαθαίνει όταν είναι σε θέση να παράγει εκτιμήσεις ενός σημείου δεδομένων με βάση τη δομή που έχει συμπεράνει από προηγούμενα δεδομένα. Στο πλαίσιο των χρονοσειρών, η μηχανική μάθηση μπορεί να εφαρμοστεί για την πρόβλεψη της τιμής στο επόμενο χρονικό βήμα, σε προβλήματα ταξινόμησης όπου ο στόχος είναι να προβλεφθεί η ετικέτα μιας εισόδου, και στην ομαδοποίηση για τον προσδιορισμό ποια σημεία μοιάζουν μεταξύ τους σε χώρο υψηλών διαστάσεων. Ο στόχος της μηχανικής μάθησης είναι να χαρτογραφήσει δεδομένα υψηλών διαστάσεων και να προσπαθήσει να διαχωρίσει σημεία δεδομένων

με βάση τις ετικέτες τους. Τα μοντέλα προσπαθούν να ελαχιστοποιήσουν τις αποστάσεις ανάμεσα σε παρόμοια σημεία και να τα ταξινομήσουν σωστά. Στον πυρήνα της μηχανικής μάθησης είναι συναρτήσεις δεδομένων και απώλειας. Τα μοντέλα προσπαθούν να μάθουν χαρακτηριστικά από τα δεδομένα με τη βοήθεια συναρτήσεων απώλειας που γενικεύουν τα μοντέλα σε δεδομένα που δεν είχαν δει προηγουμένως. Η ανίχνευση ανωμαλιών σε χρηματοοικονομικές χρονοσειρές είναι ένα πεδίο που έχει χρησιμοποιήσει ιστορικά πολλές στατιστικές μεθόδους για να προσπαθήσει να μαντέψει ανωμαλίες σε διαδοχικά δεδομένα. Συχνά, τα δεδομένα χρονοσειρών εμφανίζουν μοτίβα όπως η περιοδικότητα, η κυκλική ανάπτυξη, η εκθετική αποσύνθεση. Τα δεδομένα μπορεί επίσης μερικές φορές να υποστούν κραδασμούς ή αλλαγές που αντιβαίνουν στον προβλέψιμο χαρακτήρα του. Για παράδειγμα, η τιμή μιας μετοχής μπορεί να μειωθεί αν διαρρεύσουν νέα για μεγάλη απώλεια της εταιρείας.

Για την ανίχνευση αυτών των ανωμαλιών, θα ήταν χρήσιμο να υπάρχει ένα σύστημα που μπορεί να τις εντοπίζει και να τις ανιχνεύει, δίνοντας στους διαχειριστές τη δυνατότητα να μειώνουν τις απώλειες.

3.2 Παράγοντες για ανίχνευση ανωμαλιών

Η ανίχνευση ανωμαλιών μέσω μεθόδων μηχανικής μάθησης εξαρτάται από πολλούς παράγοντες που επηρεάζουν την αναγνώριση μη συνήθων προτύπων ή συμπεριφορών στα δεδομένα. Ανάμεσα σε αυτούς τους παράγοντες είναι:

1. Τύπος Δεδομένων: Η φύση των δεδομένων που χρησιμοποιούνται είναι κρίσιμης σημασίας. Για παράδειγμα, στις χρηματοοικονομικές χρονοσειρές, οι ανωμαλίες μπορεί να εκδηλωθούν ως απρόβλεπτες αλλαγές στις τιμές ή στους όγκους των συναλλαγών.
2. Χαρακτηριστικά Δεδομένων: Η επιλογή και η επεξεργασία των χαρακτηριστικών δεδομένων επηρεάζει την απόδοση της ανίχνευσης ανωμαλιών. Επομένως, τα σωστά επιλεγμένα χαρακτηριστικά μπορούν να βελτιώσουν τη δυνατότητα αναγνώρισης ανωμαλιών.
3. Μοντέλο Μηχανικής Μάθησης: Η επιλογή του κατάλληλου μοντέλου μηχανικής μάθησης είναι σημαντική. Ορισμένα μοντέλα είναι καταλληλότερα για συγκεκριμένους τύπους ανωμαλιών από άλλα.

Επομένως, η επιτυχημένη ανίχνευση ανωμαλιών με χρήση μηχανικής μάθησης εξαρτάται από την ορθή χρήση και των τριών παραγόντων.

3.3 Κίνδυνοι στη διαδικασία εκπαίδευσης του μοντέλου

Κατά τη διάρκεια της διαδικασίας εκπαίδευσης ενός μοντέλου μηχανικής μάθησης, υπάρχουν ορισμένοι κίνδυνοι που μπορεί να επηρεάσουν την απόδοση και την αξιοπιστία του μοντέλου. Ορισμένοι από αυτούς τους κινδύνους περιλαμβάνουν:

1. Υπερ-εκπαίδευση (Overfitting): Η υπερ-εκπαίδευση συμβαίνει όταν το μοντέλο εκπαιδεύεται να μάθει τα δεδομένα εκπαίδευσης και αποτυγχάνει να γενικεύσει σωστά σε νέα δεδομένα. Αυτό μπορεί να συμβεί όταν το μοντέλο είναι πολύπλοκο ή όταν έχουμε λίγα δεδομένα εκπαίδευσης σε σχέση με τον αριθμό των παραμέτρων του μοντέλου.

2. Υπο-εκπαίδευση (Underfitting): Η υπο-εκπαίδευση συμβαίνει όταν το μοντέλο είναι πολύ απλό για να μάθει την πολυπλοκότητα των δεδομένων εκπαίδευσης. Αυτό μπορεί να συμβεί όταν το μοντέλο έχει πολύ λίγες παραμέτρους ή όταν δεν εκπαιδεύεται για αρκετά μεγάλο χρονικό διάστημα.

3. Διαφόριση (Covariate Shift): Η διαφόριση συμβαίνει όταν η κατανομή των δεδομένων εκπαίδευσης διαφέρει από την κατανομή των δεδομένων εφαρμογής. Αυτό μπορεί να οδηγήσει σε μοντέλα με ανακριβή δεδομένα.

4. Ανισορροπία Δεδομένων (Data Imbalance): Η ανισορροπία στις κλάσεις των δεδομένων μπορεί να οδηγήσει σε μοντέλα που είναι προκατειλημμένα προς τις συχνότερες κλάσεις και αγνοούν τις λιγότερο συχνές.

Η αξιοπιστία του μοντέλου βασίζεται στην καλή εκπαίδευσή του με στόχο την αντιμετώπιση τέτοιων κινδύνων.

Ορισμένες από τις κύριες μεθόδους μηχανικής μάθησης που χρησιμοποιούνται για την ανίχνευση ανωμαλιών σε χρονοσειρές χρηματοοικονομικών δεδομένων περιλαμβάνουν:

Αυτόματη Ανίχνευση Εκτίμησης (Automatic Estimation Detection - AED): Αυτή η μέθοδος χρησιμοποιεί τεχνικές όπως οι χρονοσειρές ή οι εκτιμήσεις για να ανιχνεύσει ανωμαλίες στα δεδομένα.

Μοντέλα Μηχανικής Μάθησης όπως τα Νευρωνικά Δίκτυα, τα αναδρομικά νευρωνικά δίκτυα (RNNs) και η Μακρά Βραχυπρόθεσμη Μνήμη (LSTM), είναι ισχυρά εργαλεία

για την ανάλυση χρονοσειρών. Είναι ικανά να καταγράφουν μακροπρόθεσμες εξαρτήσεις σε δεδομένα και χρησιμοποιούνται για πρόβλεψη, ανίχνευση ανωμαλιών, ταξινόμηση και δημιουργία χρονοσειρών.

Αλγόριθμοι Ανίχνευσης Ανωμαλιών: Οι αλγόριθμοι όπως οι αλγόριθμοι Isolation Forest και οι αλγόριθμοι βασισμένοι στην αρχή της παραγωγής (GANs) χρησιμοποιούνται για τη δημιουργία συνθετικών δεδομένων και για την ανάλυση και την πρόβλεψη μακροπρόθεσμων τάσεων της αγοράς.

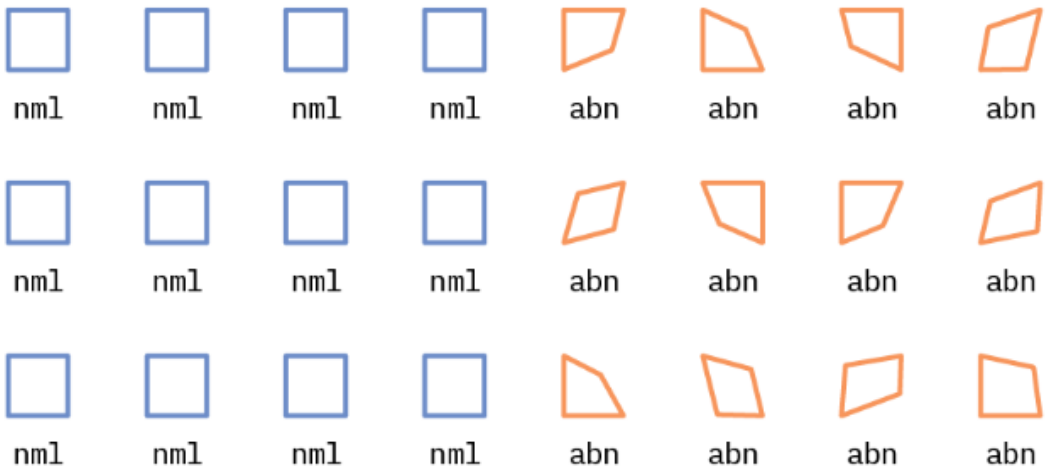
Συχνά, ο συνδυασμός αυτών των μεθόδων μπορεί να οδηγήσει σε βελτιωμένες επιδόσεις στην ανίχνευση ανωμαλιών σε χρονοσειρές χρηματοοικονομικών δεδομένων. Επιπλέον, η επιτυχία αυτών των μεθόδων εξαρτάται σε μεγάλο βαθμό από την ποιότητα και την κατάλληλη προεπεξεργασία των δεδομένων.

4. Κατηγορίες μηχανικής μάθησης

Οι βασικές κατηγορίες μηχανικής μάθησης είναι οι εξής:
Supervised Learning
Unsupervised Learning
Semi-Supervised Learning

4.1 Supervised Learning

Η πρώτη κατηγορία περιλαμβάνει μεθόδους επιβλεπόμενης μάθησης που χρησιμοποιούν δεδομένα με ετικέτα, δηλαδή παραδείγματα με κανονικές και ανώμαλες καταστάσεις, για να εκπαιδεύσουν ταξινομητές που προβλέπουν ανωμαλίες. Ένας ταξινομητής επιδιώκει να μάθει μια συνάρτηση που απεικονίζει τα χαρακτηριστικά εισόδου (X) στις ετικέτες εξόδου (Y), δηλαδή Y = f(X), όπου X είναι ο χώρος των εισόδων και Y είναι ο χώρος των εξόδων. Καθώς η επιβλεπόμενη ανίχνευση ανωμαλιών είναι παρόμοια με την τυπική διάταξη ταξινόμησης, μπορούν να εφαρμοστούν διάφορες μέθοδοι ταξινόμησης με καλή εμπειρική απόδοση. Έρευνες έχουν δείξει ότι η λογιστική παλινδρόμηση(Logistic Regression) σε συνδυασμό με τεχνητά νευρωνικά δίκτυα(ANN) και η μέθοδος των Μηχανών Διανυσμάτων Υποστήριξης(SVM) παρουσιάζουν καλύτερες επιδόσεις σε σύγκριση με στατιστικές τεχνικές, όσον αφορά το συνολικό ποσοστό ταξινόμησης και την ευαισθησία.



Σχήμα 4.1 Επιβλεπόμενη Μάθηση [3]

Η εικόνα αντιπροσωπεύει μια διαδικασία **επιβλεπόμενης μάθησης**, όπου όλα τα δεδομένα είναι κατηγοριοποιημένα.

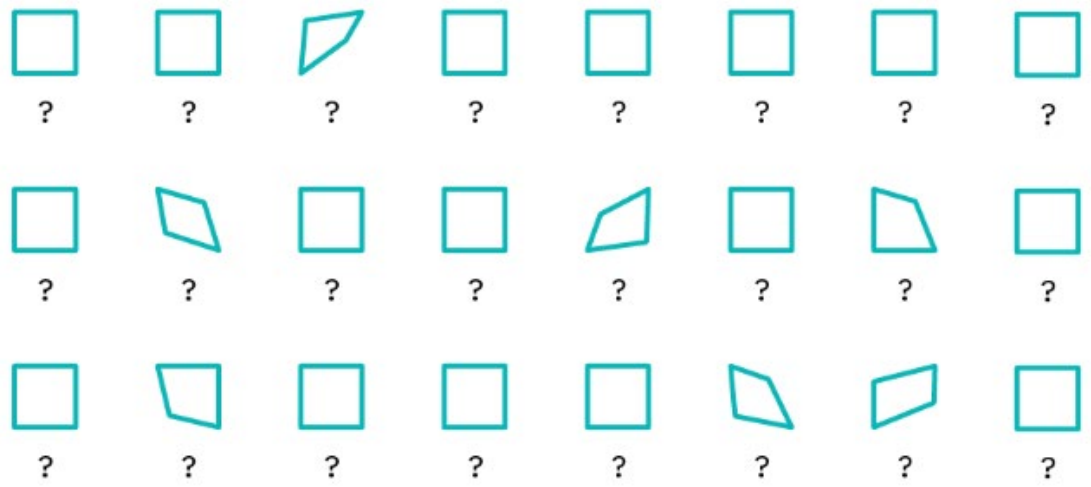
- **Κανονικά Σχήματα (nm1)**: Τα μπλε τετράγωνα αναφέρονται ως "nm1", που υποδηλώνει ότι είναι κανονικά δεδομένα (χωρίς ανωμαλίες).
- **Ανώμαλα Σχήματα (abn)**: Τα πορτοκαλί παραμορφωμένα πολύγωνα αναφέρονται ως "abn", υποδηλώνοντας ανώμαλα δεδομένα (αποκλίσεις από το κανονικό).

Η επιβλεπόμενη μάθηση βασίζεται σε δεδομένα που είναι πλήρως κατηγοριοποιημένα, όπως φαίνεται στην εικόνα. Κάθε σχήμα έχει ήδη ταξινομηθεί ως είτε κανονικό ("nm1") είτε ανώμαλο ("abn"). Σε αυτό το πλαίσιο, τα δεδομένα που έχουν ήδη ετικέτες χρησιμοποιούνται για την εκπαίδευση ενός μοντέλου, το οποίο στη συνέχεια μπορεί να μάθει να ταξινομεί νέα δεδομένα βασισμένο σε αυτές τις ετικέτες. Η διαδικασία αυτή βασίζεται στη χρήση ετικετών ως επιβλεπόμενη πληροφορία για την εκπαίδευση αλγορίθμων, όπως νευρωνικά δίκτυα ή SVM, ώστε να αναγνωρίζουν μοτίβα και να προβλέπουν τις κατηγορίες νέων παρατηρήσεων.

4.2 Unsupervised Learning

Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning): Στην αντίθετη περίπτωση, κατά την μη επιβλεπόμενη μάθηση, το σύστημα εκπαιδεύεται σε ένα σύνολο δεδομένων χωρίς ετικέτες εξόδου. Ο βασικός στόχος της μάθησης χωρίς επίβλεψη είναι η εξαγωγή χρήσιμης πληροφορίας από τα δεδομένα, όπως η ανακάλυψη κρυμμένων δομών, ομάδων ή περιοχών ομοιότητας.

Η μάθηση χωρίς επίβλεψη είναι ιδιαίτερα χρήσιμη όταν έχουμε μεγάλα σύνολα δεδομένων για τα οποία δεν υπάρχουν ετικέτες εξόδου, ή όταν θέλουμε να ανακαλύψουμε κρυμμένα πρότυπα ή δομές στα δεδομένα, για αυτό και χρησιμοποιείται για την ανίχνευση ανωμαλιών.



Σχήμα 4.2 Μη Επιβλεπόμενη Μάθηση [4]

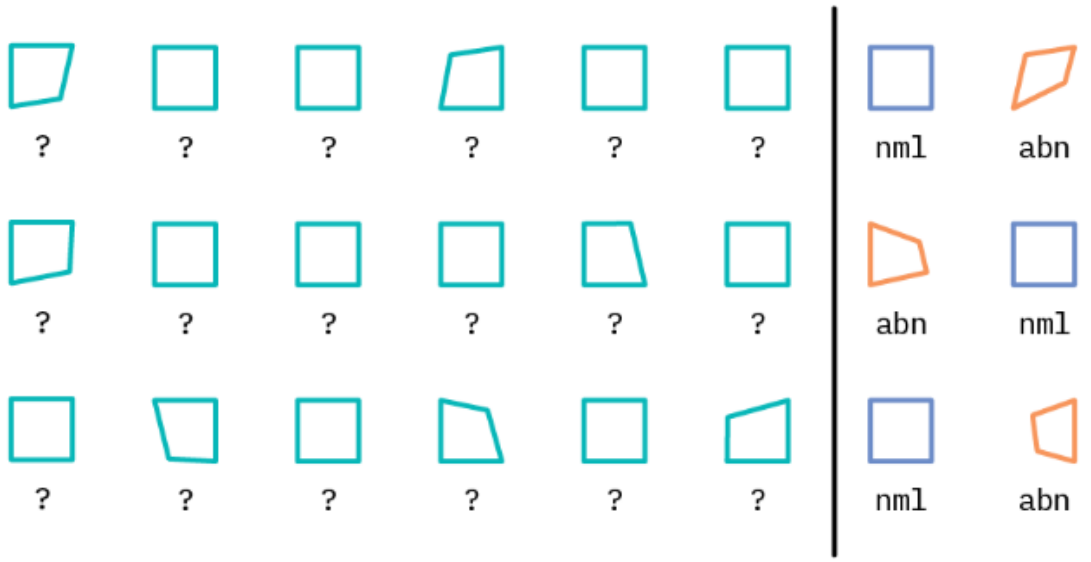
Η εικόνα αντιπροσωπεύει μια διαδικασία **μη επιβλεπόμενης μάθησης**, όπου τα δεδομένα δεν έχουν ετικέτες και ο αλγόριθμος πρέπει να ανακαλύψει μόνος του τα μοτίβα ή τις κατηγορίες.

- **Μη κατηγοριοποιημένα σχήματα**: Κάθε σχήμα (είτε κανονικό τετράγωνο είτε παραμορφωμένο πολύγωνο) συνοδεύεται από ένα ερωτηματικό, που υποδηλώνει ότι δεν υπάρχει κάποια προκαθορισμένη ετικέτα ή κατηγορία.
- **Αβεβαιότητα Κατηγοριοποίησης**: Ο στόχος στη μη επιβλεπόμενη μάθηση είναι να εντοπιστούν μοτίβα ή ομάδες (clusters) στα δεδομένα χωρίς τη βοήθεια ετικετών. Στην προκειμένη περίπτωση, ο αλγόριθμος θα πρέπει να αναγνωρίσει ότι τα τετράγωνα είναι κανονικά και τα παραμορφωμένα πολύγωνα είναι ανώμαλα, χωρίς να του έχει δοθεί αυτή η πληροφορία εκ των προτέρων.

Στη μη επιβλεπόμενη μάθηση, όπως υποδεικνύεται στην εικόνα, τα δεδομένα δεν έχουν προκαθορισμένες ετικέτες και ο αλγόριθμος πρέπει να ανακαλύψει τη δομή τους. Έτσι, μέσω τεχνικών όπως το **clustering** ή η **ανίχνευση ανωμαλιών**, μπορεί να καταλήξει σε ομάδες ή κατηγορίες (π.χ., κανονικά τετράγωνα και ανώμαλα πολύγωνα) με βάση τις ομοιότητες και τις διαφορές στα δεδομένα.

4.3 Semi-Supervised Learning

Η ημι-εποπτευόμενη μάθηση είναι ένα κλάδος της μηχανικής μάθησης που συνδυάζει την εποπτευόμενη και τη μη εποπτευόμενη μάθηση χρησιμοποιώντας ταυτόχρονα



Σχήμα 4.3 Ημι-επιβλεπόμενη Μάθηση [5]

επισημασμένα και μη επισημασμένα δεδομένα για την εκπαίδευση των μοντέλων τεχνητής νοημοσύνης για κατηγοριοποίηση και προβλήματα παλινδρόμησης.

Η εικόνα απεικονίζει μια διαδικασία **ημι-επιβλεπόμενης μάθησης**, όπου υπάρχει ένας συνδυασμός κατηγοριοποιημένων και μη κατηγοριοποιημένων δεδομένων.

- **Αριστερή Πλευρά (ερωτηματικά)**: Τα σχήματα με τα ερωτηματικά δεν έχουν ετικέτες ("?" για κατηγοριοποίηση), γεγονός που δείχνει ότι αυτά τα δεδομένα είναι μη κατηγοριοποιημένα. Αυτά αντιπροσωπεύουν το τμήμα των δεδομένων που δεν είναι γνωστό σε σχέση με το αν είναι "κανονικά" ή "ανώμαλα".
- **Δεξιά Πλευρά (nm1, abn)**: Αυτά τα σχήματα έχουν ήδη κατηγοριοποιηθεί ως "nm1" (κανονικά) ή "abn" (ανώμαλα). Αυτά τα δεδομένα είναι γνωστά και χρησιμεύουν ως η επιβλεπόμενη πλευρά της εκπαίδευσης.

Η εικόνα δείχνει μια τυπική περίπτωση **ημι-επιβλεπόμενης μάθησης** όπου ένας αλγόριθμος χρησιμοποιεί τόσο κατηγοριοποιημένα (δεξιά πλευρά) όσο και μη κατηγοριοποιημένα (αριστερή πλευρά) δεδομένα για να κάνει προβλέψεις ή να μάθει τα μοτίβα. Τα κατηγοριοποιημένα δεδομένα βοηθούν στην εκπαίδευση του μοντέλου, ενώ τα μη κατηγοριοποιημένα χρησιμοποιούνται για την περαιτέρω βελτίωση των προβλέψεων ή την ανίχνευση νέων κατηγοριών με βάση τα μοτίβα που αναγνωρίζονται.

4.4 Αλγόριθμοι μηχανική μάθησης

4.4.1 Αλγόριθμος k-NN

Ο αλγόριθμος K-Nearest Neighbors (kNN) αναζητά τα πλησιέστερα δεδομένα σε ένα σύνολο εκπαίδευσης για να προβλέψει την τιμή ενός νέου σημείου δεδομένων. Ο αριθμός K αναπαριστά πόσα από τα κοντινότερα δεδομένα θα ληφθούν υπόψη. Αντί να χρησιμοποιεί τον μέσο όρο των απαντήσεων αυτών των γειτόνων, προτείνεται η χρήση της μέσης τιμής για καλύτερη ανθεκτικότητα σε ανωμαλίες. Επίσης, μπορεί να χρησιμοποιηθεί για τον υπολογισμό ενός επιπέδου εμπιστοσύνης, που δείχνει την πιθανή αξιοπιστία της πρόβλεψης βασιζόμενο στα δεδομένα των κοντινότερων γειτόνων. Αυτό είναι ιδιαίτερα χρήσιμο για την ανίχνευση ανωμαλιών και για να καθοριστεί ο βαθμός εμπιστοσύνης στην πρόβλεψη. Συνολικά, αυτές οι προσαρμογές βοηθούν τον αλγόριθμο να είναι πιο αξιόπιστος και αποτελεσματικός σε περιπτώσεις με ανωμαλίες στα δεδομένα.

4.4.2 Logistic Regression

Η λογιστική παλινδρόμηση(Logistic Regression) είναι ένας αλγόριθμος με επιβλεπόμενη μάθηση που χρησιμοποιείται για προβλήματα ταξινόμησης. Προσπαθεί να προβλέψει την πιθανότητα ότι ένα δεδομένο δείγμα ανήκει σε μία από δύο ή περισσότερες κατηγορίες, με βάση τα χαρακτηριστικά του. Αν και η λογιστική παλινδρόμηση είναι ένας απλός αλγόριθμος, αποτελεσματικός και ευέλικτος.

4.4.3 GAN

Generative Adversarial Network (GAN) είναι ένας αλγόριθμος μηχανική μάθησης που αποτελείται από δύο μέρη, μια γεννήτρια και έναν διακριτή. Η γεννήτρια παράγει δείγματα που μοιάζουν με αυτά από ένα συγκεκριμένο σύνολο δεδομένων, ενώ ο διακριτής προσπαθεί να διακρίνει μεταξύ πραγματικών και ψευδών δειγμάτων. Χρησιμοποιώντας GAN για ανίχνευση ανωμαλιών σε χρονοσειρές, μπορούμε να δημιουργήσουμε ψευδείς χρονοσειρές που αντιπροσωπεύουν την κανονική

συμπεριφορά. Έπειτα, συγκρίνοντας πραγματικές και ψευδείς χρονοσειρές, μπορούμε να εντοπίσουμε ανωμαλίες, όπως σε χρηματοοικονομικές χρονοσειρές.

4.4.4 Isolation Forest

Το Isolation Forest (IF) είναι ένας αλγόριθμος που χρησιμοποιείται για την ανίχνευση ανωμαλιών σε σύνολα δεδομένων, συμπεριλαμβανομένων και των χρονοσειρών. Η βασική του ιδέα είναι ότι οι ανωμαλίες πρέπει να είναι συνήθως πιο "μακριά" από τα κανονικά δεδομένα σε έναν χώρο χαρακτηριστικών. Για να επιτευχθεί αυτό, το IF δημιουργεί τυχαία δέντρα απομόνωσης, χωρίζοντας τα δεδομένα σε υποσύνολα. Κάθε δέντρο ορίζει μια διαδρομή από τη ρίζα μέχρι το φύλλο για κάθε σημείο δεδομένων, και ο αριθμός των διαδρομών που χρειάζονται για να φτάσει ένα σημείο στο φύλλο χρησιμοποιείται ως μέτρο ανωμαλίας. Συνήθως, οι ανωμαλίες έχουν συντομότερες διαδρομές σε σύγκριση με τα κανονικά δεδομένα, καθώς είναι πιο απομονωμένες.

Για την ανίχνευση ανωμαλιών στις χρονοσειρές, το IF μπορεί να εφαρμοστεί εισάγοντας τα χαρακτηριστικά της χρονοσειράς σε έναν χώρο χαρακτηριστικών και στη συνέχεια εφαρμόζοντας τον αλγόριθμο. Με τη χρήση του IF, μπορούν να εντοπιστούν αυτές οι ανωμαλίες στις χρονοσειρές με αποτελεσματικό και αυτόματο τρόπο, καθιστώντας το ένα χρήσιμο εργαλείο για την παρακολούθηση και τη διαχείριση χρονοσειρών σε ποικίλες εφαρμογές, όπως οι χρηματοοικονομικές προβλέψεις.

4.4.5 ANN & RNN

Τα τεχνητά νευρωνικά δίκτυα (ANN) και τα αναδρομικά νευρωνικά δίκτυα (RNN) αντιπροσωπεύουν ένα πιο πολύπλοκο μοντέλο μάθησης, εμπνευσμένο από τη λειτουργία του ανθρώπινου εγκεφάλου. Τα ANN αποτελούνται από διάφορα επίπεδα νευρώνων και χρησιμοποιούνται για πολλαπλές εφαρμογές, συμπεριλαμβανομένης της ταξινόμησης και της παλινδρόμησης. Τα RNN είναι ένα είδος ANN που έχει τη δυνατότητα να διατηρεί πληροφορία για μεγαλύτερο χρονικό διάστημα.

4.4.6 Support Vector Machines (SVM)

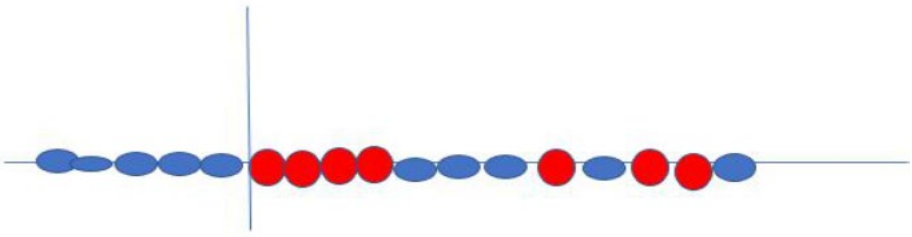
Το Support Vector Machine (SVM) είναι ένας ισχυρός και ευέλικτος αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για διάφορες εφαρμογές. Οι κύριες χρήσεις του περιλαμβάνουν τη γραμμική ή μη γραμμική ταξινόμηση, την παλινδρόμηση, και την ανίχνευση ακραίων τιμών. Το SVM μπορεί να εφαρμοστεί σε πληθώρα εργασιών και προβλημάτων, καθιστώντας το έναν από τους πιο χρήσιμους αλγόριθμους στον τομέα της μηχανικής μάθησης.

Ανάλογα με τη φύση του ορίου απόφασης, οι Support Vector Machines (SVM) μπορούν να χωριστούν σε δύο κύριες κατηγορίες:

- Γραμμικό SVM
- Μη γραμμικό SVM

4.4.6.1 Γραμμικό SVM

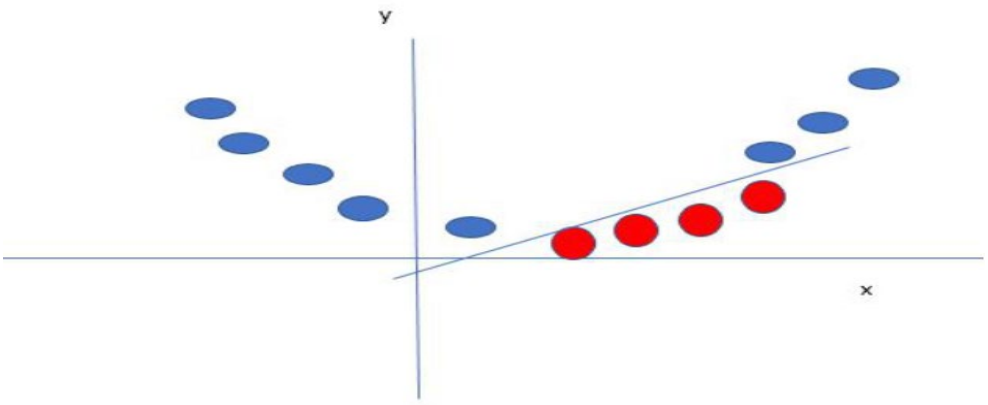
Τα γραμμικά SVM χρησιμοποιούν ένα γραμμικό όριο απόφασης για να διαχωρίσουν τα σημεία δεδομένων διαφορετικών κλάσεων. Όταν τα δεδομένα μπορούν να διαχωριστούν ακριβώς γραμμικά, τα γραμμικά SVM είναι ιδιαίτερα κατάλληλα. Αυτό σημαίνει ότι μια απλή ευθεία γραμμή (σε 2D) ή ένα υπερεπίπεδο (σε υψηλότερες διαστάσεις) μπορεί να διαχωρίσει πλήρως τα σημεία δεδομένων στις αντίστοιχες κατηγορίες τους. Το υπερεπίπεδο που μεγιστοποιεί το περιθώριο μεταξύ των κλάσεων είναι το όριο απόφασης.



Σχήμα 4.1 Γραμμικό SVM [6]

4.4.6.2 Μη Γραμμικό SVM

Τα μη γραμμικά SVM χρησιμοποιούνται για την ταξινόμηση δεδομένων όταν δεν μπορούν να διαχωριστούν σε δύο κλάσεις με μια ευθεία γραμμή (στην περίπτωση 2D). Χρησιμοποιώντας συναρτήσεις πυρήνα, τα μη γραμμικά SVM μπορούν να διαχειριστούν μη γραμμικά διαχωρίσιμα δεδομένα. Οι συναρτήσεις πυρήνα μετασχηματίζουν τα αρχικά δεδομένα εισόδου σε έναν χώρο χαρακτηριστικών υψηλότερης διάστασης, όπου τα σημεία δεδομένων μπορούν να διαχωριστούν γραμμικά. Ένα γραμμικό SVM χρησιμοποιείται στη συνέχεια σε αυτόν τον τροποποιημένο χώρο για να εντοπίσει ένα μη γραμμικό όριο απόφασης.



Σχήμα 4.2 Μη Γραμμικό SVM [7]

4.4.6.3 One-Class SVM

Το One-Class SVM είναι ένας αλγόριθμος μη επιβλεπόμενης μηχανικής μάθησης που χρησιμοποιείται για την ανίχνευση ανωμαλιών. Βασίζεται στον συνδυασμό της μεθόδου One-Class Classification (OCC) και των Support Vector Machines (SVM). Η λειτουργία του One-Class SVM επικεντρώνεται στην εκμάθηση των χαρακτηριστικών μιας συγκεκριμένης κλάσης δεδομένων και στην αναγνώριση τυχόν εξαιρέσεων από αυτή την κλάση. Ουσιαστικά, το μοντέλο προσπαθεί να προσδιορίσει μια σφαίρα που περικλείει τα δεδομένα εκπαίδευσης, θεωρώντας ότι οτιδήποτε βρίσκεται εκτός αυτής της σφαίρας αποτελεί ανωμαλία.

4.4.6.4 One-Class Classification (OCC)

Η One-Class Classification (OCC) είναι μια τεχνική μηχανικής μάθησης που στοχεύει στην αναγνώριση κανονικών δεδομένων και ανωμαλιών χωρίς να έχει πρόσβαση σε παραδείγματα ανωμαλιών κατά τη διάρκεια της εκπαίδευσης. Δηλαδή, το μοντέλο μαθαίνει από τα κανονικά δεδομένα και προσπαθεί να ανιχνεύσει οποιαδήποτε δεδομένα που αποκλίνουν από αυτά. Αυτή η μέθοδος είναι ευρέως χρησιμοποιούμενη για την επίλυση προβλημάτων ανίχνευσης ανωμαλιών και είναι επίσης γνωστή ως unary classification ή class-modelling.

4.4.6.5 Λειτουργίες των SVM

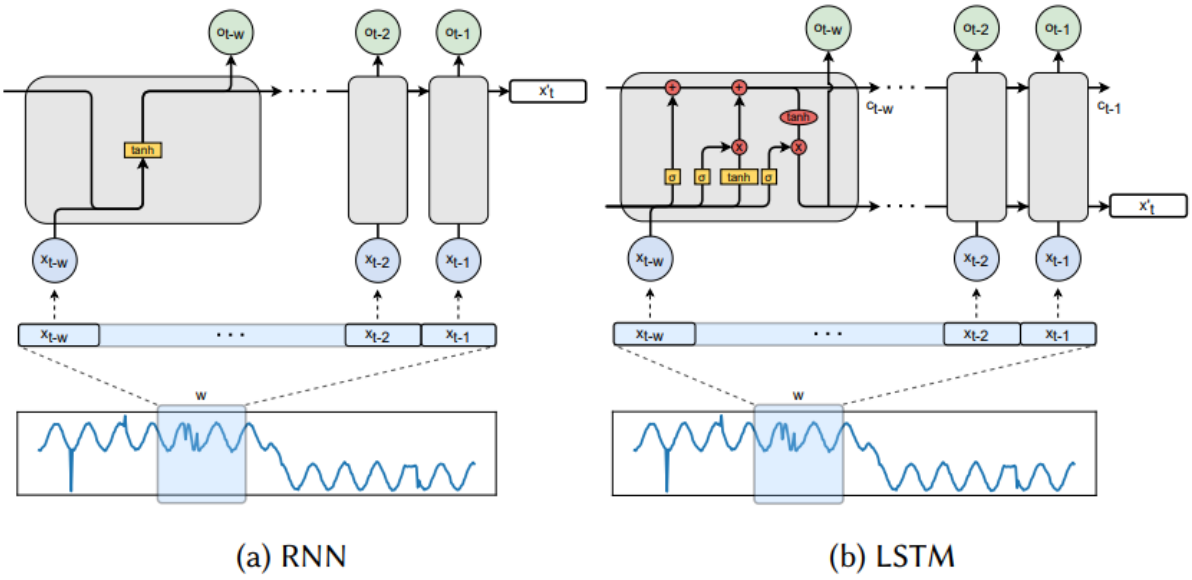
Οι αλγόριθμοι SVM είναι ιδιαίτερα αποτελεσματικοί διότι επιδιώκουν να βρουν το μέγιστο διαχωριστικό επίπεδο μεταξύ των διαφορετικών κλάσεων στα δεδομένα στόχου. Αυτό σημαίνει ότι οι SVM προσπαθούν να βρουν το επίπεδο που διαχωρίζει τις κλάσεις με τη μέγιστη απόσταση από τα πλησιέστερα σημεία δεδομένων κάθε κλάσης, γνωστά ως υποστηρικτικά διανύσματα (support vectors). Αυτή η ιδιότητα εξασφαλίζει τη βέλτιστη ταξινόμηση και τη μέγιστη γενίκευση στα νέα δεδομένα.
Για να επιτύχουν αυτό το στόχο, οι SVM χρησιμοποιούν συναρτήσεις πυρήνα (kernels) για να μετασχηματίσουν τα δεδομένα σε υψηλότερες διαστάσεις, όπου οι κλάσεις γίνονται πιο εύκολα διαχωρίσιμες. Το γεγονός αυτό καθιστά τους SVM ιδιαίτερα αποτελεσματικούς στην επίλυση προβλημάτων με μη γραμμικές σχέσεις και στην επεξεργασία δεδομένων υψηλών διαστάσεων.

4.4.7 LSTM

Η Μακρά Βραχυπρόθεσμη Μνήμη(Long Short-Term Memory), αντιπροσωπεύει μια εξειδικευμένη αρχιτεκτονική εντός του πεδίου των επαναλαμβανόμενων νευρωνικών δικτύων (RNNs), που χρησιμοποιείται ευρέως στον τομέα της βαθιάς μάθησης(Deep Learning). Ξεχωρίζει για την εξαιρετική του ικανότητα να καταγράφει και να δημιουργεί μοντέλα με μακροπρόθεσμες εξαρτήσεις, καθιστώντας το ιδιαίτερα κατάλληλο για εργασίες που περιλαμβάνουν διαδοχική πρόβλεψη δεδομένων. Σε αντίθεση με τα συμβατικά νευρωνικά δίκτυα, τα δίκτυα LSTM ενσωματώνουν συνδέσεις ανάδρασης,

επιτρέποντάς τους να αναλύουν ολόκληρες ακολουθίες δεδομένων αντί να επεξεργάζονται μεμονωμένα σημεία δεδομένων. Αυτή η ικανότητα καθιστά τα LSTM ιδιαίτερα ικανά στην αναγνώριση και πρόβλεψη μοτίβων εντός διαδοχικών τύπων δεδομένων, όπως χρονοσειρές, δεδομένα κειμένου και σήματα ομιλίας.

Στην ανάπτυξη των δικτύων μακράς βραχυπρόθεσμης μνήμης, διαπιστώθηκε ότι αντιμετωπίζουν αποτελεσματικά το πρόβλημα της εξαφάνισης της κλίσης που παρουσιάζεται στα απλά αναδρομικά νευρωνικά δίκτυα (RNN). Σε ένα θεμελιώδες επίπεδο, ένα LSTM λειτουργεί παρόμοια με ένα κελί RNN. Ωστόσο, οι εσωτερικοί μηχανισμοί του είναι δομημένοι για να αντιμετωπίζουν τις προκλήσεις που σχετίζονται με τη διατήρηση μακροπρόθεσμων εξαρτήσεων.



Σχήμα 4.3 Δομή RNN (αριστερά) LSTM (δεξιά) [8]

Η ανίχνευση ανωμαλιών στις χρηματοοικονομικές χρονοσειρές έχει διανύσει μεγάλη πορεία από τις πρώτες στατιστικές μεθόδους έως τις σύγχρονες προσεγγίσεις μηχανικής μάθησης. Έχουν διαδραματίσει σημαντικό ρόλο στην ανάπτυξη αυτής της περιοχής, και η ενσωμάτωσή τους με πιο προχωρημένες τεχνικές μηχανικής μάθησης συνεχίζει να βελτιώνει την ακρίβεια και την αποτελεσματικότητα της ανίχνευσης ανωμαλιών. Με την πρόοδο της τεχνολογίας και τη διαθεσιμότητα μεγάλων δεδομένων, οι μελλοντικές εξελίξεις στην ανίχνευση ανωμαλιών αναμένονται να προσφέρουν ακόμα πιο προηγμένες και ακριβείς λύσεις.

5. ΠΡΑΚΤΙΚΟ ΜΕΡΟΣ

Το πρακτικό μέρος της πτυχιακής εργασίας εστιάζει στην ανίχνευση ανωμαλιών στις χρηματοοικονομικές χρονοσειρές χρησιμοποιώντας μεθόδους μηχανικής μάθησης, με την ιστορική τιμή του Bitcoin ως αντικείμενο μελέτης. Αρχικά, θα γίνει εισαγωγή και προετοιμασία των δεδομένων ιστορικών τιμών του Bitcoin. Αυτό περιλαμβάνει τη συλλογή, το φιλτράρισμα και την επεξεργασία των δεδομένων ώστε να είναι έτοιμα για ανάλυση. Η σωστή προετοιμασία των δεδομένων είναι κρίσιμη για την ακρίβεια των αποτελεσμάτων που θα παραχθούν. Στη συνέχεια, θα κατασκευαστούν και θα εφαρμοστούν κατάλληλα μοντέλα μηχανικής μάθησης για την ανίχνευση ανωμαλιών. Η διαδικασία αυτή θα περιλαμβάνει την επιλογή των σωστών μοντέλων, την εκπαίδευσή τους και την αξιολόγηση της αποτελεσματικότητάς τους. Θα γίνει σύγκριση διαφορετικών μοντέλων για να εντοπιστεί ποιο από αυτά είναι το πιο αποδοτικό στην ανίχνευση ανωμαλιών στις τιμές του Bitcoin. Μετά την εφαρμογή των μοντέλων, θα ακολουθήσει η ανάλυση των αποτελεσμάτων. Αυτό περιλαμβάνει την ερμηνεία των ανωμαλιών που ανιχνεύθηκαν και την κατανόηση των επιπτώσεών τους στην αγορά και στη χρονοσειρά των τιμών του Bitcoin. Η ανάλυση αυτή είναι σημαντική για να κατανοηθεί πώς οι ανωμαλίες επηρεάζουν τη χρηματοοικονομική χρονοσειρά. Τέλος, θα δημιουργηθούν γραφικές παραστάσεις για την οπτική αναπαράσταση των πραγματικών τιμών, των προβλέψεων και των ανωμαλιών. Αυτές οι παραστάσεις θα βοηθήσουν στην καλύτερη κατανόηση των δεδομένων και των αποτελεσμάτων. Επιπλέον, θα συνταχθούν αναφορές και θα παρουσιαστούν τα αποτελέσματα με τρόπο κατανοητό και περιεκτικό, ώστε να είναι προσβάσιμα τόσο στους ακαδημαϊκούς όσο και στους μη ειδικούς.

Στο παρόν πρακτικό μέρος της πτυχιακής εργασίας, παρουσιάζουμε την κατασκευή ενός προγράμματος στην Python για την ανίχνευση ανωμαλιών στις τιμές του Bitcoin χρησιμοποιώντας το μοντέλο ARIMA (Autoregressive Integrated Moving Average) καθώς και τους αλγορίθμους SVM και LSTM. Η ανάλυση αυτή έχει στόχο να εντοπίσει ασυνήθιστες τιμές που μπορεί να υποδηλώνουν σημαντικά γεγονότα στην αγορά ή να αποκαλύπτουν προβλήματα στα δεδομένα.

Το πρακτικό μέρος της πτυχιακής εργασίας στοχεύει στην εφαρμογή και αξιολόγηση μεθόδων μηχανικής μάθησης για την ανίχνευση ανωμαλιών στις χρηματοοικονομικές χρονοσειρές. Η ανάλυση των δεδομένων του Bitcoin θα μας επιτρέψει να αναδείξουμε

την αποτελεσματικότητα αυτών των μεθόδων και να κατανοήσουμε καλύτερα τις δυναμικές της αγοράς. Μέσα από την προσεκτική εφαρμογή και αξιολόγηση των μοντέλων, επιδιώκουμε να συμβάλουμε στην υπάρχουσα βιβλιογραφία και να παράσχουμε πρακτικά εργαλεία για την ανίχνευση ανωμαλιών σε πραγματικά δεδομένα.

5.1 Dataset

Το dataset που χρησιμοποιήθηκε είναι ιστορικά δεδομένα τιμών του Bitcoin από το αρχείο BTC-USDD.csv. Περιλαμβάνει διάφορα χαρακτηριστικά για το Bitcoin, όπως η τιμή κλεισίματος (**Close**), για την περίοδο από τον Ιούλιο του 2019 έως τον Ιανουάριο του 2024, που καλύπτει περίπου 5 χρόνια. Τα συγκεκριμένα πεδία που χρησιμοποιήθηκαν στην ανάλυσή σου ήταν οι τιμές κλεισίματος του Bitcoin, και οι ανωμαλίες εντοπίστηκαν με βάση τις αποκλίσεις σε αυτές τις τιμές.
Το dataset περιλαμβάνει 1828 εγγραφές και 7 στήλες. Οι στήλες είναι οι εξής: η στήλη "Date", η οποία αναφέρεται στην ημερομηνία της καταγραφής και αποθηκεύεται σε μορφή object/string, η στήλη "Open" που αναπαριστά την τιμή ανοίγματος του Bitcoin για τη συγκεκριμένη ημέρα, η στήλη "High", η οποία περιέχει την υψηλότερη τιμή που έφτασε το Bitcoin κατά τη διάρκεια της ημέρας, και η στήλη "Low", που δείχνει τη χαμηλότερη τιμή του Bitcoin την ίδια ημέρα. Επίσης, υπάρχει η στήλη "Close", που δείχνει την τιμή κλεισίματος του Bitcoin για την ημέρα, καθώς και η στήλη "Adj Close", η οποία είναι η προσαρμοσμένη τιμή κλεισίματος που λαμβάνει υπόψη μερίσματα ή άλλα εταιρικά γεγονότα (αν και στο Bitcoin δεν υπάρχουν τέτοια γεγονότα, αυτή η στήλη συνήθως είναι ίδια με την τιμή κλεισίματος). Τέλος, η στήλη "Volume" αναπαριστά τον συνολικό όγκο συναλλαγών του Bitcoin για την ημέρα.
Το dataset περιέχει μια πλήρη χρονική σειρά των τιμών του Bitcoin σε καθημερινή βάση, η οποία μπορεί να χρησιμοποιηθεί για ανάλυση χρονοσειρών, όπως αυτή που πραγματοποιήθηκε με το μοντέλο ARIMA. Τα δεδομένα είναι συνεπή και κατάλληλα για περαιτέρω ανάλυση χωρίς να χρειάζονται σημαντικές διορθώσεις.

5.1.1 Χρονική Κάλυψη

Τα δεδομένα καλύπτουν μια περίοδο από την 24η Ιουλίου 2019 έως τον Ιανουάριο του 2024. Ουσιαστικά, έχουμε ημερήσιες τιμές για περίπου 5 χρόνια, οι οποίες είναι συνεπείς χωρίς κενές εγγραφές στις στήλες.
Για τις πρώτες εγγραφές από την 24η Ιουλίου 2019, οι τιμές είναι οι εξής:

- Τιμή κλεισίματος στις 24 Ιουλίου 2019: 9811.93 USD.
- Ο όγκος συναλλαγών για την ίδια ημέρα: 17,398,734,322 (σε δολάρια).

5.1.2 Στατιστική Ανάλυση Δεδομένων

Στην υπάρχουσα στατιστική ανάλυση των δεδομένων οι τιμές του Bitcoin εμφανίζονται σε μορφή float64, ενώ ο όγκος συναλλαγών είναι τύπου int64. Αυτό εξασφαλίζει ακρίβεια στην καταγραφή των αριθμητικών δεδομένων. Επιπλέον, οι ημερήσιες εγγραφές είναι πλήρεις, χωρίς ελλείψεις ή κενά στις τιμές, κάτι που είναι σημαντικό για την ανάλυση χρονοσειρών, καθώς μια συνεπής και αδιάλειπτη σειρά δεδομένων επιτρέπει την εφαρμογή αλγορίθμων ανίχνευσης ανωμαλιών και μοντέλων πρόβλεψης, όπως το ARIMA, χωρίς την ανάγκη περαιτέρω καθαρισμού των δεδομένων. Αυτό ενισχύει την αξιοπιστία της ανάλυσης, καθώς δεν χρειάζεται να αντιμετωπιστούν ελλείψεις ή να γίνουν προσαρμογές σε ελλιπή δεδομένα.
Οι στήλες που αφορούν τις τιμές του Bitcoin (Open, High, Low, Close) δείχνουν τη διακύμανση της τιμής του Bitcoin σε καθημερινή βάση, με τον όγκο συναλλαγών να δείχνει τον συνολικό αριθμό των συναλλαγών για την κάθε ημέρα.
Οι τύποι δεδομένων float64 και int64 αναφέρονται στον τρόπο με τον οποίο αποθηκεύονται και αναπαρίστανται οι αριθμητικές τιμές στον υπολογιστή. Ο τύπος float64 χρησιμοποιείται για την αποθήκευση δεκαδικών αριθμών, όπως οι τιμές των μετοχών ή του Bitcoin (Open, High, Low, Close), καθώς αυτές οι τιμές δεν είναι ακέραιες και χρειάζονται δεκαδικά ψηφία. Το "float" σημαίνει "floating point" (κινητή υποδιαστολή), και το 64 υποδηλώνει ότι χρησιμοποιούνται 64 bits για την αποθήκευση της τιμής, επιτρέποντας μεγάλη ακρίβεια στους υπολογισμούς με δεκαδικά ψηφία και πολύ μεγάλους ή πολύ μικρούς αριθμούς. Από την άλλη πλευρά, ο τύπος **int64** χρησιμοποιείται για την αποθήκευση ακέραιων αριθμών, όπως ο όγκος συναλλαγών (Volume), καθώς ο αριθμός των συναλλαγών είναι πάντα ακέραιος. Το "int" σημαίνει

"integer" (ακέραιος), και το 64 σημαίνει ότι χρησιμοποιούνται 64 bits για την αποθήκευση του αριθμού, επιτρέποντας την αποθήκευση πολύ μεγάλων ακέραιων αριθμών, είτε θετικών είτε αρνητικών. Συνολικά, οι τύποι δεδομένων float64 και int64 εξασφαλίζουν ακρίβεια και επαρκή χώρο για την αποθήκευση και την επεξεργασία των αριθμητικών τιμών στο dataset.

5.2 Ανωμαλίες

Οι ανωμαλίες που ανιχνεύονται αφορούν απρόσμενες και μεγάλες αποκλίσεις στην τιμή κλεισίματος του Bitcoin από την προβλεπόμενη τάση. Συγκεκριμένα, χρησιμοποιείται ένα μοντέλο ARIMA (AutoRegressive Integrated Moving Average), το οποίο προβλέπει την αναμενόμενη τιμή κλεισίματος κάθε ημέρας. Οι ανωμαλίες εντοπίζονται συγκρίνοντας την πραγματική τιμή με την προβλεπόμενη τιμή και ανιχνεύονται όταν το σφάλμα (ή υπόλοιπο) μεταξύ τους υπερβαίνει ένα καθορισμένο κατώφλι.
Το κατώφλι αυτό έχει οριστεί σε 2,5 φορές την τυπική απόκλιση των υπολοίπων (residuals). Αν η διαφορά είναι μεγαλύτερη από αυτό το όριο, τότε θεωρείται ότι έχουμε ανωμαλία, δηλαδή μια σημαντική απόκλιση από την αναμενόμενη πορεία των τιμών του Bitcoin. Αυτές οι ανωμαλίες μπορεί να αντιστοιχούν σε απότομες αυξήσεις ή πτώσεις της τιμής, οι οποίες πιθανόν να προκλήθηκαν από εξωτερικά γεγονότα, όπως οικονομικές ανακοινώσεις ή ακραίες κινήσεις της αγοράς.

5.3 Αλγόριθμος

Η ανίχνευση ανωμαλιών πραγματοποιήθηκε υπολογίζοντας τα υπολείμματα (residuals), δηλαδή τη διαφορά μεταξύ των προβλεπόμενων τιμών και των πραγματικών τιμών. Οι ανωμαλίες εντοπίστηκαν όταν τα υπολείμματα ξεπερνούσαν ένα όριο που καθορίστηκε με βάση τη σταθερή απόκλιση των υπολοίπων και αντιπροσωπεύουν σημεία όπου η τιμή του Bitcoin αποκλίνει σημαντικά από τις αναμενόμενες τάσεις.
Ο αλγόριθμος SVM προσαρμόστηκε στα δεδομένα, και μέσω της κατηγοριοποίησης των παρατηρήσεων ως κανονικές ή ανώμαλες. Ο αλγόριθμος θεωρεί ανωμαλίες τις τιμές που αποκλίνουν σημαντικά από το γενικό μοτίβο και τις μαρκάρει ως -1. Αυτή η

μέθοδος είναι ιδιαίτερα κατάλληλη για την ανίχνευση απότομων αλλαγών στην τιμή, που συχνά οφείλονται σε εξωτερικούς παράγοντες ή συναισθηματικές αντιδράσεις της αγοράς.

Η προσέγγιση LSTM χρησιμοποιήθηκε για να ανιχνεύσει ανωμαλίες στη χρονοσειρά του Bitcoin. Το μοντέλο εκπαιδεύτηκε με δεδομένα προηγούμενων τιμών για να προβλέψει τις μελλοντικές τιμές, και στη συνέχεια υπολογίστηκαν τα σφάλματα πρόβλεψης. Οι ανωμαλίες εντοπίστηκαν όταν το σφάλμα πρόβλεψης ξεπερνούσε ένα όριο που καθορίστηκε από τη σταθερή απόκλιση των σφαλμάτων.

Ο συνδυασμός του στατιστικού μοντέλου ARIMA και των τεχνικών μηχανικής μάθησης SVM και LSTM προσφέρει μια ολιστική προσέγγιση για την ανίχνευση ανωμαλιών στην τιμή του Bitcoin. Το ARIMA παρέχει ένα ισχυρό εργαλείο για την ανάλυση γραμμικών σχέσεων και εποχικών τάσεων, ενώ οι αλγόριθμοι μηχανικής μάθησης (SVM και LSTM) αναγνωρίζουν πιο σύνθετα και μη γραμμικά μοτίβα που μπορεί να μην είναι άμεσα εμφανή μέσω παραδοσιακών στατιστικών μεθόδων. Η πολυπλοκότητα της τιμής του Bitcoin, λόγω της αβεβαιότητας και της μεταβλητότητας της αγοράς, καθιστά τον συνδυασμό αυτών των τεχνικών ένα εξαιρετικό εργαλείο για την ανάλυση και πρόβλεψη ανωμαλιών.

Με αυτό τον τρόπο, η παρούσα μελέτη επιτυγχάνει μια σφαιρική κατανόηση των παραγόντων που επηρεάζουν την τιμή του Bitcoin, ενώ παρέχει και μια ισχυρή προσέγγιση για την ανίχνευση ανωμαλιών σε δυναμικά περιβάλλοντα.

5.4 Παράμετροι υλοποίησης αλγορίθμου

Οι αλγόριθμοι ARIMA, SVM, και LSTM υλοποιούνται με συγκεκριμένες παραμέτρους για την ανίχνευση ανωμαλιών στην τιμή του Bitcoin.

Για το μοντέλο ARIMA, χρησιμοποιείται η συνάρτηση auto_arima, η οποία αυτόματα επιλέγει τις βέλτιστες παραμέτρους για την προσαρμογή του στα δεδομένα. Οι παράμετροι που χρησιμοποιούνται είναι οι εξής: Το μοντέλο ξεκινά από τιμές p = 1 και q = 1, ενώ οι μέγιστες τιμές για p και q ορίζονται στο 5. Η παράμετρος d ορίζεται σε 1, που σημαίνει ότι πραγματοποιείται μία διαφορά στα δεδομένα για να καταστούν στάσιμα. Το μοντέλο είναι εποχικό, με εποχικότητα m = 12 και μία εποχική διαφορά (D = 1). Η επιλογή των παραμέτρων γίνεται με βηματική αναζήτηση (stepwise), και το

σύστημα επιλέγει τον βέλτιστο συνδυασμό των παραμέτρων με βάση το κριτήριο AIC (Akaike Information Criterion).

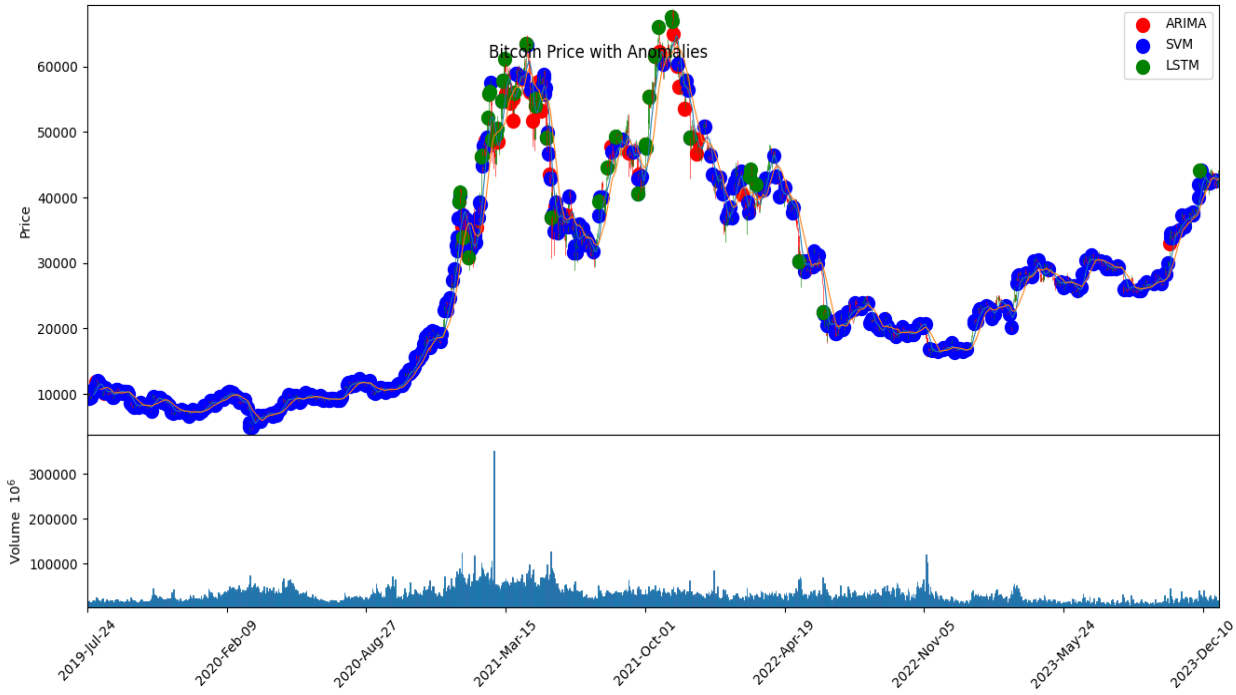
Για την ανίχνευση ανωμαλιών με τον αλγόριθμο SVM (Support Vector Machine), χρησιμοποιείται το μοντέλο One-Class SVM. Ο πυρήνας που χρησιμοποιείται είναι ο Radial Basis Function (RBF), ο οποίος είναι κατάλληλος για την ανίχνευση μη γραμμικών σχέσεων. Η παράμετρος nu έχει οριστεί σε 0.01, καθορίζοντας ότι το 1% των δεδομένων θα θεωρηθεί ανωμαλία. Η τιμή gamma είναι 0.1, καθορίζοντας την ευαισθησία του μοντέλου όσον αφορά την απόσταση μεταξύ των δεδομένων.

Τέλος, για την υλοποίηση του LSTM (Long Short-Term Memory) χρησιμοποιείται ένα νευρωνικό δίκτυο με δύο επίπεδα LSTM. Το μήκος της ακολουθίας των δεδομένων είναι 50 χρονικά σημεία. Το πρώτο επίπεδο LSTM περιέχει 50 νευρώνες και επιστρέφει μια ακολουθία δεδομένων, ενώ το δεύτερο επίπεδο LSTM έχει επίσης 50 νευρώνες και επιστρέφει την τελευταία χρονική στιγμή της ακολουθίας. Ακολουθούν δύο επίπεδα Dense (πλήρως συνδεδεμένα επίπεδα), με την τελική έξοδο να παράγεται από ένα επίπεδο. Το μοντέλο εκπαιδεύεται χρησιμοποιώντας τον βελτιστοποιητή Adam και τη συνάρτηση απώλειας Mean Squared Error. Η εκπαίδευση γίνεται για 20 εποχές με μέγεθος δέσμης 32 παρατηρήσεις.

Συνολικά, αυτά τα μοντέλα συνδυάζονται για την ανίχνευση ανωμαλιών στην τιμή του Bitcoin. Το ARIMA αναλύει γραμμικές τάσεις και εποχικές διακυμάνσεις, το SVM ανιχνεύει μη γραμμικές ανωμαλίες, ενώ το LSTM αναγνωρίζει σύνθετα μοτίβα και εξαρτήσεις σε μακροχρόνιες χρονοσειρές.

5.5 Ανάλυση Γραφημάτων

Αυτό το γράφημα παρουσιάζει την τιμή του Bitcoin με ανιχνευμένες ανωμαλίες χρησιμοποιώντας τρεις διαφορετικές τεχνικές: **ARIMA**, **SVM**, και **LSTM**. Η ανάλυση βασίζεται σε ιστορικά δεδομένα και χρησιμοποιεί συνδυασμό στατιστικών μοντέλων και μηχανικής μάθησης για να ανιχνεύσει αποκλίσεις από τις αναμενόμενες τιμές, υποδηλώνοντας πιθανές ανωμαλίες στην αγορά.



Σχήμα 5.1 Τιμές Bitcoin, όγκος συναλλαγών και ανωμαλίες Arima, SVM, LSTM.

Τιμή του Bitcoin

Το πάνω μέρος του γραφήματος απεικονίζει την τιμή του Bitcoin σε μια περίοδο από τον Ιούλιο του 2019 έως τον Δεκέμβριο του 2023. Το Bitcoin εμφανίζει διακυμάνσεις με αυξημένες και μειωμένες τιμές, υποδεικνύοντας την υψηλή μεταβλητότητα της αγοράς κατά την εν λόγω περίοδο. Καθώς η τιμή αυξάνεται και μειώνεται με έντονους ρυθμούς, αυτό καθιστά την ανίχνευση ανωμαλιών ιδιαίτερα σημαντική για τη μελέτη ασυνήθιστων συμβάντων που μπορεί να επηρεάσουν την αγορά.

Όγκος Συναλλαγών

Στο κάτω μέρος του γραφήματος, απεικονίζεται ο όγκος συναλλαγών, που δείχνει τον συνολικό αριθμό των συναλλαγών που πραγματοποιήθηκαν σε συγκεκριμένες χρονικές περιόδους. Ο όγκος των συναλλαγών παρουσιάζει αυξομειώσεις ανάλογα με τις διακυμάνσεις της τιμής, και οι μεγάλες αυξήσεις στον όγκο συναλλαγών συνήθως σχετίζονται με σημαντικές κινήσεις στην τιμή του Bitcoin. Οι κορυφώσεις στον όγκο συναλλαγών συχνά συμπίπτουν με τις ανωμαλίες, υποδεικνύοντας έντονη δραστηριότητα στην αγορά.

Ανωμαλίες με ARIMA (Κόκκινες Κουκκίδες)

Οι κόκκινες κουκκίδες στο γράφημα υποδεικνύουν ανωμαλίες που ανιχνεύθηκαν από το στατιστικό μοντέλο ARIMA. Στο γράφημα, οι ανωμαλίες που εντοπίστηκαν με ARIMA αντιπροσωπεύουν τιμές του Bitcoin που αποκλίνουν σημαντικά από τις προβλέψεις του μοντέλου, υποδεικνύοντας ότι υπήρξαν γεγονότα ή παράγοντες που προκάλεσαν απότομες αλλαγές στην αγορά. Συνολικά, το ARIMA ανίχνευσε 58 ανωμαλίες.

Ανωμαλίες με SVM (Μπλε Κουκκίδες)

Οι μπλε κουκκίδες δείχνουν τις ανωμαλίες που εντοπίστηκαν με τη χρήση του Support Vector Machine (SVM). Το SVM χρησιμοποιεί τον RBF πυρήνα για να ανιχνεύσει μη γραμμικές σχέσεις και είναι σε θέση να εντοπίζει ανωμαλίες βασισμένο σε διαφορετικά μοτίβα από το ARIMA. Το μοντέλο SVM είναι σχεδιασμένο για να διακρίνει τις ανώμαλες τιμές από τις κανονικές με βάση τα δεδομένα που έχει μάθει ως "κανονικά". Οι μπλε κουκκίδες είναι διασκορπισμένες σε όλη τη διάρκεια της περιόδου, και όπως αναφέρεται στο PDF, το SVM ανίχνευσε 585 ανωμαλίες.

Ανωμαλίες με LSTM (Πράσινες Κουκκίδες)

Οι πράσινες κουκκίδες αντιπροσωπεύουν τις ανωμαλίες που εντοπίστηκαν με το LSTM (Long Short-Term Memory), ένα νευρωνικό δίκτυο που είναι ικανό να ανιχνεύει πιο σύνθετα μοτίβα και μακροχρόνιες εξαρτήσεις στα δεδομένα. Οι ανωμαλίες που εντοπίζονται με LSTM είναι λιγότερες σε αριθμό σε σύγκριση με τα άλλα δύο μοντέλα, καθώς το LSTM επικεντρώνεται σε πιο μακροχρόνιες εξαρτήσεις και χρησιμοποιείται για την ανίχνευση μη γραμμικών σχέσεων που μπορεί να παραβλέπονται από άλλες τεχνικές.

Συνολική Σύγκριση

- **ARIMA**: Κατάλληλο για την ανίχνευση γραμμικών και εποχικών ανωμαλιών. Είναι διαφανές και εύκολο να ερμηνευτεί, αλλά περιορίζεται στην ανίχνευση μόνο γραμμικών αποκλίσεων.
- **SVM**: Πολύ ευαίσθητο στις αλλαγές, ανιχνεύει πολλές μη γραμμικές ανωμαλίες, αλλά μπορεί να έχει υπερευαισθησία και να παράγει πολλά ψευδώς θετικά αποτελέσματα.

- **LSTM**: Ικανό να ανιχνεύει ανωμαλίες που βασίζονται σε μακροχρόνιες εξαρτήσεις και πιο σύνθετες μη γραμμικές σχέσεις. Παρέχει λιγότερες ανιχνεύσεις, αλλά με μεγαλύτερη ακρίβεια, αν και η εκπαίδευσή του είναι πιο περίπλοκη.

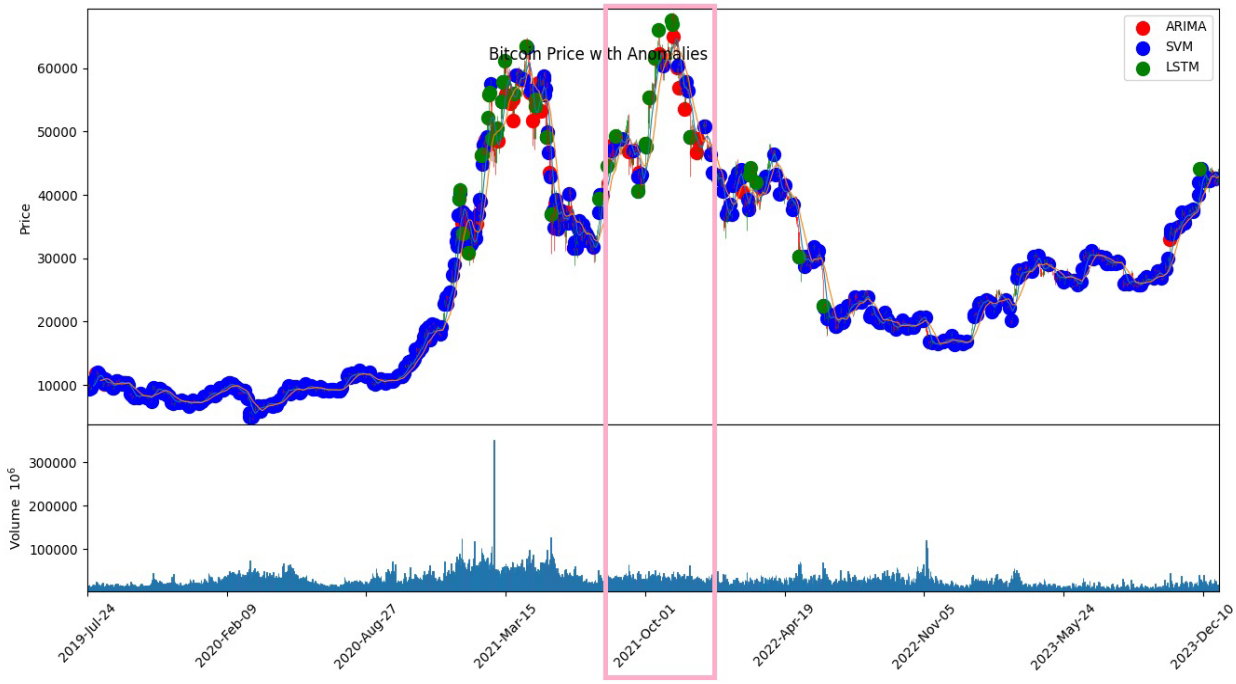
Ο συνδυασμός των τριών αυτών αλγορίθμων επιτρέπει μια πολυδιάστατη ανάλυση της τιμής του Bitcoin, με τον κάθε αλγόριθμο να συνεισφέρει στη διαφορετική πτυχή της ανίχνευσης ανωμαλιών.

Η τιμή του Bitcoin παρουσιάζει μεγάλες διακυμάνσεις, με κορυφές στα τέλη του 2021 και μικρότερες ανόδους και καθόδους καθ' όλη τη διάρκεια της περιόδου.

Οι ανωμαλίες που εντοπίστηκαν από το ARIMA σημειώνονται συχνότερα κατά τη διάρκεια απότομων αυξομειώσεων, κάτι που υποδεικνύει την ευαισθησία του μοντέλου σε αιφνίδιες αλλαγές στην τιμή.

Οι ανωμαλίες είναι πιο έντονες κατά την άνοδο του 2021 και στις περιόδους μεγάλης πτώσης, υποδεικνύοντας σημεία που το μοντέλο εντόπισε ως απρόβλεπτα ή μη κανονικά. Χωρίς τις ανωμαλίες, μπορούμε να δούμε καθαρά την τάση και τη διακύμανση της τιμής του Bitcoin με μεγαλύτερη ακρίβεια.

Οι κύριες περίοδοι έντονης ανόδου και πτώσης συμπίπτουν με τα οικονομικά γεγονότα και τις επενδυτικές τάσεις που επηρεάζουν το Bitcoin και την αγορά των κρυπτονομισμάτων.



Σχήμα 5.2 Ανίχνευση Ανωμαλιών στην Τιμή του Bitcoin τον Οκτώβριο 2021 με ARIMA, SVM, και LSTM

Σημαντικά Σημεία

Η εικόνα δείχνει μια περιοχή όπου η τιμή του Bitcoin βρίσκεται κοντά σε μια κορύφωση και ακολουθεί μια σημαντική πτώση. Αυτές οι έντονες αλλαγές στην τιμή συνδέονται με ανωμαλίες που ανιχνεύθηκαν και από τους τρεις αλγόριθμους.

Οι κόκκινες κουκκίδες (ARIMA), οι μπλε κουκκίδες (SVM), και οι πράσινες κουκκίδες (LSTM) βρίσκονται πυκνά διασκορπισμένες σε αυτό το τμήμα, υποδεικνύοντας ότι όλοι οι αλγόριθμοι εντόπισαν ανωμαλίες σε αυτή την περίοδο.

Συγκέντρωση Ανωμαλιών κατά την Πτώση:

- Κατά την κορύφωση της τιμής του Bitcoin και αμέσως μετά την έναρξη της πτώσης, παρατηρείται συγκέντρωση ανωμαλιών από όλους τους αλγόριθμους. Αυτό σημαίνει ότι αυτή η περιοχή ήταν μια κρίσιμη περίοδος για την αγορά του Bitcoin, όπου οι προβλέψεις απέκλιναν από τις πραγματικές τιμές.
- Το γεγονός ότι οι ανωμαλίες ανιχνεύθηκαν και από το ARIMA, το SVM, και το LSTM υποδηλώνει ότι υπήρχε σημαντική απόκλιση από τις αναμενόμενες τάσεις, τόσο σε γραμμικό όσο και σε μη γραμμικό επίπεδο.

Αλληλοεπικαλυπτόμενες Ανωμαλίες:

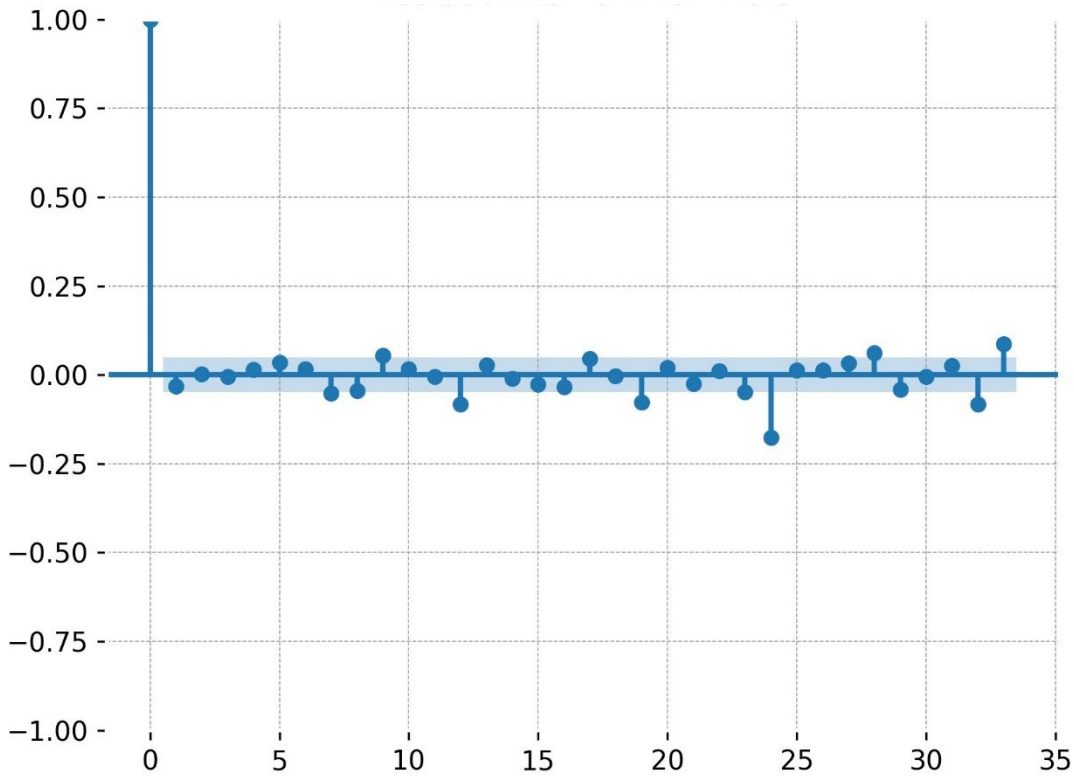
- Παρατηρείται ότι οι κουκκίδες (ανωμαλίες) από τους τρεις αλγόριθμους συχνά αλληλεπικαλύπτονται στην ίδια περιοχή. Αυτό σημαίνει ότι οι αλγόριθμοι δεν ανιχνεύουν ανωμαλίες σε διαφορετικά σημεία, αλλά σε κοινές περιοχές, γεγονός που υποδεικνύει ότι αυτές οι αλλαγές ήταν πραγματικά ασυνήθιστες για την αγορά.
- Οι ανωμαλίες από το SVM φαίνονται να είναι πιο πυκνές, γεγονός που υποδηλώνει ότι το SVM ανιχνεύει περισσότερες αποκλίσεις από την κανονική συμπεριφορά σε αυτήν την περίοδο.

Συμπεριφορά της Αγοράς:

- Η εικόνα αντιπροσωπεύει μια περίοδο έντονης μεταβλητότητας στην αγορά του Bitcoin. Η ύπαρξη τόσων πολλών ανωμαλιών σε μια μικρή χρονική περίοδο δείχνει ότι υπήρξαν εξωτερικοί παράγοντες ή γεγονότα που προκάλεσαν μεγάλες αλλαγές στην τιμή του Bitcoin, κάτι που ενδέχεται να οφείλεται σε γεγονότα όπως ανακοινώσεις, θεσμικές κινήσεις ή άλλα σημαντικά οικονομικά γεγονότα.

Από την εικόνα προκύπτει ότι γύρω από την περίοδο του Οκτωβρίου 2021 η τιμή του Bitcoin παρουσίασε έντονη μεταβλητότητα, με σημαντικές αλλαγές και αποκλίσεις από τις αναμενόμενες τιμές. Οι αλγόριθμοι ARIMA, SVM, και LSTM ανίχνευσαν πλήθος ανωμαλιών, ιδιαίτερα γύρω από την κορύφωση και την αρχή της πτώσης της τιμής, υποδεικνύοντας την παρουσία ισχυρών ασυνήθιστων γεγονότων στην αγορά. Η παρουσία τόσων ανωμαλιών σε αυτή την περίοδο υπογραμμίζει την αστάθεια της αγοράς του Bitcoin και τη σημασία της χρήσης πολλαπλών μοντέλων για την ακριβή ανίχνευση ανωμαλιών.

Υπόλοιπα (Residuals)



Σχήμα 5.3 Υπόλοιπα (residuals) ARIMA

Αυτό το διάγραμμα απεικονίζει τα ACF (Autocorrelation Function) και PACF (Partial Autocorrelation Function) των υπολειμμάτων (residuals) από το μοντέλο ARIMA που χρησιμοποιήθηκε. Τα διαγράμματα αυτά είναι σημαντικά εργαλεία για την ανάλυση της συμπεριφοράς των υπολειμμάτων μετά την εφαρμογή ενός μοντέλου ARIMA.

Ανάλυση του Διαγράμματος

ACF (Autocorrelation Function):

- Το ACF δείχνει την αυτοσυσχέτιση των υπολειμμάτων με τις καθυστερημένες τιμές (lags).

- Η κορυφή στην αρχή του διαγράμματος (στο lag 0) είναι πάντα 1, καθώς κάθε σειρά είναι 100% συσχετισμένη με τον εαυτό της.

- Οι υπόλοιπες κορυφές είναι αρκετά κοντά στο μηδέν, και η πλειονότητα των σημείων πέφτει εντός του διαστήματος εμπιστοσύνης (που φαίνεται με την μπλε σκίαση).

- Αυτό υποδηλώνει ότι δεν υπάρχει σημαντική αυτοσυσχέτιση στα υπολείμματα, κάτι που είναι επιθυμητό, καθώς δείχνει ότι το μοντέλο ARIMA έχει αφαιρέσει την προβλεψιμότητα από τα δεδομένα και τα υπολείμματα συμπεριφέρονται σαν "λευκός θόρυβος" (white noise).

PACF (Partial Autocorrelation Function):

- Το PACF δείχνει την αυτοσυσχέτιση των υπολειμμάτων με τις καθυστερημένες τιμές, αφαιρώντας την επίδραση των ενδιάμεσων καθυστερήσεων.
- Και σε αυτό το διάγραμμα, παρατηρείται ότι οι τιμές των καθυστερημένων αυτοσυσχετίσεων είναι αρκετά κοντά στο μηδέν και δεν ξεπερνούν το διάστημα εμπιστοσύνης, κάτι που υποδηλώνει ότι δεν υπάρχει σημαντική άμεση συσχέτιση ανάμεσα στις τιμές.

Το γεγονός ότι τόσο το ACF όσο και το PACF δεν παρουσιάζουν σημαντικές κορυφές πέραν του διαστήματος εμπιστοσύνης υποδηλώνει ότι το μοντέλο ARIMA έχει κάνει καλή δουλειά στο να μοντελοποιήσει τα δεδομένα και να αφαιρέσει οποιαδήποτε συστηματική τάση ή αυτοσυσχέτιση.

- Τα υπολείμματα συμπεριφέρονται σαν λευκός θόρυβος, κάτι που είναι επιθυμητό σε ένα καλά εκπαιδευμένο μοντέλο ARIMA.
- Εάν υπήρχαν σημαντικές κορυφές εκτός των ορίων εμπιστοσύνης, θα υπήρχε η ένδειξη ότι το μοντέλο δεν έχει καταφέρει να συλλάβει πλήρως τη δομή των δεδομένων και θα ήταν απαραίτητη η αναθεώρησή του.

Συνολικά, η εικόνα αυτή υποδηλώνει ότι το μοντέλο ARIMA είναι επαρκές για τη συγκεκριμένη χρονοσειρά, καθώς τα υπολείμματα δεν παρουσιάζουν έντονη αυτοσυσχέτιση.

6. Συμπεράσματα

Η ανάλυση των δύο γραφημάτων, που παρουσιάζουν την ανίχνευση ανωμαλιών στην τιμή του Bitcoin καθώς και τη μερική αυτοσυσχέτιση των υπολειμμάτων (residuals) του μοντέλου ARIMA, μας οδηγεί σε μια βαθύτερη κατανόηση της συμπεριφοράς της αγοράς του Bitcoin και της αποτελεσματικότητας των διαφόρων αλγορίθμων ανίχνευσης ανωμαλιών. Ας προχωρήσουμε σε μια αναλυτική διερεύνηση των παρατηρήσεων που μπορούμε να εξάγουμε από αυτά τα γραφήματα. Αυτό που παρατηρούμε άμεσα είναι ότι το SVM εντοπίζει πολύ περισσότερες ανωμαλίες σε σχέση με τα άλλα δύο μοντέλα, κάτι που υποδηλώνει την αυξημένη ευαισθησία του στην ανίχνευση μη γραμμικών αποκλίσεων και μικρότερων διακυμάνσεων. Το ARIMA, από την άλλη, ανιχνεύει λιγότερες ανωμαλίες, κυρίως σε σημεία όπου παρατηρούνται μεγάλες διακυμάνσεις στην τιμή, όπως κοντά στις κορυφώσεις και κατά τη διάρκεια απότομων πτώσεων. Το LSTM εντοπίζει επίσης αρκετές ανωμαλίες, αλλά αυτές φαίνεται να είναι πιο επιλεκτικές, επικεντρωμένες σε σημεία με μακροχρόνιες εξαρτήσεις και πιο σύνθετα μοτίβα στις διακυμάνσεις της τιμής.

Η σύγκριση των τριών μοντέλων δείχνει ότι κάθε μοντέλο ανιχνεύει ανωμαλίες με διαφορετικό τρόπο, ανάλογα με την ευαισθησία και την προσέγγιση που χρησιμοποιεί. Το ARIMA, ως παραδοσιακό στατιστικό μοντέλο, επικεντρώνεται κυρίως σε μεγάλες γραμμικές αποκλίσεις από τις προβλέψεις του, οι οποίες συχνά συνδέονται με εποχικές τάσεις και πιο εμφανείς μεταβολές στην αγορά. Το SVM, χάρη στην ευαισθησία του σε μικρότερες μεταβολές και τη χρήση του RBF πυρήνα, εντοπίζει πολλές μικρότερες ανωμαλίες, που σε πολλές περιπτώσεις μπορεί να θεωρηθούν ψευδώς θετικές. Το LSTM, από την άλλη, με την ικανότητά του να ανιχνεύει μακροχρόνιες εξαρτήσεις και πιο σύνθετα μοτίβα, επιλέγει λιγότερες αλλά πιο στοχευμένες ανωμαλίες, οι οποίες ενδέχεται να προέρχονται από σύνθετες αλληλεπιδράσεις στο χρόνο.

Ένα σημαντικό στοιχείο που μπορούμε να εξάγουμε από το πρώτο γράφημα είναι ότι η τιμή του Bitcoin χαρακτηρίζεται από υψηλή μεταβλητότητα, ιδιαίτερα κατά τη διάρκεια περιόδων απότομων αλλαγών, όπως μεταξύ των ετών 2020 και 2021. Αυτή η υψηλή μεταβλητότητα συνδέεται άμεσα με την εμφάνιση πολλών ανωμαλιών, οι οποίες υποδηλώνουν απότομες διακυμάνσεις που δεν ακολουθούν τις αναμενόμενες τάσεις. Αυτό γίνεται ιδιαίτερα εμφανές στην περιοχή γύρω από τον Μάρτιο του 2021, όπου

παρατηρείται κορύφωση στην τιμή του Bitcoin, με όλα τα μοντέλα να ανιχνεύουν ανωμαλίες. Οι ανωμαλίες αυτές συνδέονται με μια ξαφνική άνοδο και πτώση της τιμής, η οποία πιθανόν προκλήθηκε από κάποιο εξωτερικό γεγονός ή απότομα μεταβαλλόμενες συνθήκες στην αγορά. Το γεγονός ότι και τα τρία μοντέλα εντόπισαν ανωμαλίες στην ίδια χρονική περίοδο υποδηλώνει ότι οι συγκεκριμένες αποκλίσεις ήταν πολύ έντονες και σημαντικές για την αγορά. Το δεύτερο γράφημα, το οποίο παρουσιάζει το διάγραμμα της μερικής αυτοσυσχέτισης των υπολειμμάτων (PACF) του μοντέλου ARIMA, μας δίνει μια σαφέστερη εικόνα για την αποτελεσματικότητα του μοντέλου και την ικανότητά του να απομακρύνει τις τάσεις και τις συσχετίσεις από τα δεδομένα. Στο γράφημα αυτό, η γραμμή της αυτοσυσχέτισης για τις περισσότερες καθυστερήσεις (lags) είναι κοντά στο μηδέν, υποδηλώνοντας ότι τα υπολείμματα του ARIMA μοντέλου είναι τυχαία και δεν παρουσιάζουν σημαντική συσχέτιση με τις προηγούμενες τιμές. Αυτό είναι ένδειξη ότι το μοντέλο ARIMA έχει καταφέρει να εξαλείψει τις συστηματικές τάσεις και να δημιουργήσει ένα σύνολο υπολειμμάτων που δεν επηρεάζονται από παρελθούσες τιμές.

Η μοναδική καθυστέρηση που δείχνει υψηλή αυτοσυσχέτιση είναι η καθυστέρηση στο μηδέν, η οποία είναι αναμενόμενη, καθώς κάθε δεδομένο έχει τέλεια συσχέτιση με τον εαυτό του. Η απουσία αυτοσυσχέτισης στις επόμενες καθυστερήσεις υποδηλώνει ότι το ARIMA μοντέλο είναι κατάλληλο για τη χρονοσειρά του Bitcoin και ότι έχει επιτύχει καλή προσαρμογή στα δεδομένα. Το γεγονός ότι δεν παρατηρούνται σημαντικές συσχετίσεις πέραν της πρώτης καθυστέρησης ενισχύει την υπόθεση ότι τα δεδομένα μετά την αφαίρεση των τάσεων είναι στατικά και δεν παρουσιάζουν περαιτέρω εξαρτήσεις που δεν έχουν ληφθεί υπόψη από το μοντέλο.

Σε συνδυασμό, τα δύο γραφήματα προσφέρουν μια ολοκληρωμένη κατανόηση τόσο της φύσης της τιμής του Bitcoin όσο και της αποτελεσματικότητας των μοντέλων ανίχνευσης ανωμαλιών. Το πρώτο γράφημα δείχνει την υψηλή μεταβλητότητα της αγοράς και τον τρόπο με τον οποίο τα τρία διαφορετικά μοντέλα ανιχνεύουν ανωμαλίες με διαφορετικές μεθόδους. Το δεύτερο γράφημα, από την άλλη, αποδεικνύει την αποτελεσματικότητα του ARIMA μοντέλου στη διαχείριση της τάσης και των εξαρτήσεων της χρονοσειράς, επιβεβαιώνοντας ότι τα υπολείμματα του μοντέλου είναι τυχαία και χωρίς αυτοσυσχέτιση.

Συνολικά, η ανάλυση αυτή υποδεικνύει ότι η χρήση πολλαπλών μοντέλων ανίχνευσης ανωμαλιών είναι απαραίτητη για την πλήρη κατανόηση της αγοράς του Bitcoin. Κάθε

μοντέλο έχει τις δικές του δυνάμεις και αδυναμίες, και ο συνδυασμός τους προσφέρει μια πιο ολοκληρωμένη εικόνα των αποκλίσεων και της αστάθειας της αγοράς. Τα στατιστικά μοντέλα όπως το ARIMA είναι κατάλληλα για τη διαχείριση γραμμικών τάσεων και εποχικών αλλαγών, ενώ τα μη γραμμικά μοντέλα μηχανικής μάθησης όπως το SVM και το LSTM είναι πιο ικανά να ανιχνεύσουν σύνθετες και ασυνήθιστες ανωμαλίες.

7. Παράρτημα

7.1 Κώδικας Python

Ακολουθεί η αναφορά στις βιβλιοθήκες που χρησιμοποιούνται στο πρόγραμμα, μαζί με μια σύντομη περιγραφή των λειτουργιών τους:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import mplfinance as mpf
from pmdarima.arima import auto_arima
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from fpdf import FPDF
from PIL import Image
from sklearn.svm import OneClassSVM
import os
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense
from sklearn.preprocessing import MinMaxScaler
```

Σχήμα 7.1 Βιβλιοθήκες Python

Χρησιμοποιούμενες Βιβλιοθήκες:

1. **pandas**
   - Βιβλιοθήκη για την επεξεργασία και ανάλυση δεδομένων σε δομές πινάκων (DataFrames).
   - Χρησιμοποιείται για την φόρτωση, επεξεργασία και φιλτράρισμα των δεδομένων που προέρχονται από το αρχείο CSV.

2. **numpy**
   - Παρέχει υποστήριξη για αριθμητικούς υπολογισμούς και τη διαχείριση πολυδιάστατων πινάκων.

- Χρησιμοποιείται για υπολογισμούς όπως η τυπική απόκλιση, καθώς και για τη διαχείριση αριθμητικών πράξεων.

3. **matplotlib**
   - Βιβλιοθήκη για τη δημιουργία γραφημάτων και διαγραμμάτων.
   - Χρησιμοποιείται για τη σχεδίαση των διαγραμμάτων υπολειμμάτων και την οπτικοποίηση των δεδομένων σε συνδυασμό με άλλες βιβλιοθήκες.

4. **mplfinance**
   - Παρέχει δυνατότητα δημιουργίας ειδικών διαγραμμάτων για χρηματοοικονομικά δεδομένα, όπως τα candlestick charts.
   - Χρησιμοποιείται για τη δημιουργία candlestick διαγραμμάτων που απεικονίζουν τις τιμές του Bitcoin, καθώς και για τη σχεδίαση ανωμαλιών πάνω σε αυτά τα διαγράμματα.

5. **pmdarima**
   - Βιβλιοθήκη για την αυτοματοποίηση της διαδικασίας επιλογής παραμέτρων σε μοντέλα ARIMA.
   - Χρησιμοποιείται για την κατασκευή του μοντέλου ARIMA, το οποίο προβλέπει τιμές βασισμένες σε ιστορικά δεδομένα και βοηθά στην ανίχνευση ανωμαλιών μέσω υπολειμμάτων.

6. **statsmodels**
   - Παρέχει στατιστικά εργαλεία για την ανάλυση χρονοσειρών.
   - Χρησιμοποιείται για τη σχεδίαση των διαγραμμάτων ACF (Autocorrelation Function) και PACF (Partial Autocorrelation Function) για τα υπολείμματα του ARIMA, βοηθώντας στην ανάλυση της αυτό-συσχέτισης στα δεδομένα.

7. **fpdf**
   - Βιβλιοθήκη για τη δημιουργία αρχείων PDF.
   - Χρησιμοποιείται για τη δημιουργία της τελικής αναφοράς PDF, η οποία περιλαμβάνει τις αναλυτικές πληροφορίες σχετικά με τις ανωμαλίες που ανιχνεύθηκαν, καθώς και τα διαγράμματα.

8. **PIL**

- Παρέχει δυνατότητες επεξεργασίας εικόνων.
- Χρησιμοποιείται για την επεξεργασία και προβολή των εικόνων που δημιουργούνται από τα διαγράμματα, καθώς και για την προσθήκη ανωμαλιών στα candlestick charts.

9. **sklearn**

- Παρέχει αλγορίθμους μηχανικής μάθησης, συμπεριλαμβανομένου του SVM.
- Χρησιμοποιείται για την ανίχνευση ανωμαλιών μέσω του αλγορίθμου One-Class SVM, ο οποίος εκπαιδεύεται σε δεδομένα για να εντοπίζει σημεία που αποκλίνουν από το φυσιολογικό πρότυπο συμπεριφοράς.

10. **os**

- Παρέχει δυνατότητες για την αλληλεπίδραση με το σύστημα αρχείων.
- Χρησιμοποιείται για να ελέγχεται αν το αρχείο δεδομένων υπάρχει στη διαδρομή που καθορίζεται και για την αποθήκευση ή φόρτωση των αρχείων εικόνας και αναφοράς.

11-12. **TensorFlow και Keras**

- Οι βιβλιοθήκες TensorFlow και Keras χρησιμοποιούνται για την ανάπτυξη και την εκπαίδευση του μοντέλου LSTM (Long Short-Term Memory).
- Παρέχουν εργαλεία για την κατασκευή νευρωνικών δικτύων, τα οποία είναι κρίσιμα για την ανίχνευση ανωμαλιών βασισμένων σε σειρές χρόνου.

13. **MinMaxScaler**

- Η MinMaxScaler από τη βιβλιοθήκη scikit-learn χρησιμοποιείται για την κανονικοποίηση των δεδομένων, μετασχηματίζοντας τις τιμές ώστε να βρίσκονται σε ένα καθορισμένο εύρος.

Η χρήση των παραπάνω βιβλιοθηκών επιτρέπει την ανάπτυξη ενός πολύπλευρου συστήματος που συνδυάζει την ανάλυση δεδομένων, την πρόβλεψη με χρονοσειρές, την ανίχνευση ανωμαλιών με αλγορίθμους μηχανικής μάθησης και τη δημιουργία

αναφορών PDF, δίνοντας ολοκληρωμένα αποτελέσματα για χρηματοοικονομικές χρονοσειρές.

Το κομμάτι αυτό του κώδικα είναι η εισαγωγή των απαραίτητων βιβλιοθηκών που θα χρησιμοποιήσω στην πτυχιακή εργασία μου για την ανίχνευση ανωμαλιών σε χρηματοοικονομικές χρονοσειρές με μεθόδους μηχανικής μάθησης. Αναλυτικότερα, κάθε βιβλιοθήκη που εισάγεται έχει συγκεκριμένο ρόλο και σκοπό.

```python
def load_data(filepath):
    if not os.path.exists(filepath):
        print(f"File not found: {filepath}")
        return None
    try:
        df = pd.read_csv(filepath)
        df['Date'] = pd.to_datetime(df['Date'], format='%Y-%m-%d')
        df.set_index('Date', inplace=True)
        return df
    except Exception as e:
        print(f"Σφάλμα κατά τη φόρτωση δεδομένων: {e}")
        return None
```

Σχήμα 7.2 Φόρτωση Δεδομένων

- Η συνάρτηση ελέγχει αν το αρχείο υπάρχει.
- Διαβάζει τα δεδομένα από το αρχείο CSV.
- Μετατρέπει την στήλη Date σε αντικείμενα ημερομηνίας και την ορίζει ως δείκτη του DataFrame.
- Αν παρουσιαστεί σφάλμα, το πρόγραμμα το αναφέρει και επιστρέφει None.

```python
def prepare_arima_model(data):
    try:
        model = auto_arima(
            data,
            start_p=1, start_q=1,
            max_p=5, max_q=5, m=12,
            start_P=0, seasonal=True,
            d=1, D=1,
            trace=True,
            error_action='ignore',
            suppress_warnings=True,
            stepwise=True
        )
        return model
    except Exception as e:
        print(f"Σφάλμα κατά την προετοιμασία του ARIMA μοντέλου: {e}")
        return None
```

Σχήμα 7.3 Προετοιμασία Arima

- Αυτή η συνάρτηση δημιουργεί ένα μοντέλο ARIMA για τα δεδομένα χρονοσειράς.
- Χρησιμοποιεί την auto_arima για την αυτόματη επιλογή των παραμέτρων με εποχικότητα.
- Αν υπάρξει κάποιο σφάλμα κατά τη διαδικασία, το σφάλμα αναφέρεται και η συνάρτηση επιστρέφει None.

```python
def detect_anomalies_arima(data, model):
    predictions = model.predict_in_sample()
    residuals = data - predictions
    threshold = 2.5 * np.std(residuals)
    anomalies = data[(residuals < -threshold) | (residuals > threshold)]
    return anomalies, residuals
```

Σχήμα 7.4 Ανίχνευση ανωμαλιών με Arima

- Η συνάρτηση χρησιμοποιεί το μοντέλο ARIMA για να προβλέψει τιμές και στη συνέχεια υπολογίζει τα υπολείμματα.
- Χρησιμοποιώντας το όριο που ορίζεται ως 2.5 φορές την τυπική απόκλιση, εντοπίζει τις ανωμαλίες στα δεδομένα.

```python
def detect_anomalies_svm(data):
    try:
        # Προετοιμασία δεδομένων για SVM
        data_for_svm = data.values.reshape(-1, 1)
        svm_model = OneClassSVM(nu=0.01, kernel="rbf", gamma=0.1)
        svm_model.fit(data_for_svm)

        # Προβλέψεις ανωμαλιών (-1 = ανωμαλίες)
        svm_predictions = svm_model.predict(data_for_svm)
        anomalies = data[svm_predictions == -1]
        return anomalies
    except Exception as e:
        print(f"Σφάλμα κατά την ανίχνευση ανωμαλιών με SVM: {e}")
        return pd.Series()
```

Σχήμα 7.5 Ανίχνευση ανωμαλιών με SVM

- Επιστρέφει τα σημεία που θεωρούνται ανωμαλίες καθώς και τα υπολείμματα που δείχνουν τις αποκλίσεις από τις προβλεπόμενες τιμές.
- Η συνάρτηση χρησιμοποιεί τον αλγόριθμο One-Class SVM για την ανίχνευση ανωμαλιών σε δεδομένα χρονοσειράς.
- Προετοιμάζει τα δεδομένα σε μορφή κατάλληλη για SVM και εκπαιδεύει το μοντέλο.
- Κάνει προβλέψεις ανωμαλιών και επιστρέφει τα σημεία όπου εντοπίζονται ανωμαλίες.
- Διαχειρίζεται τυχόν σφάλματα επιστρέφοντας ένα άδειο αποτέλεσμα αν η διαδικασία αποτύχει.

```python
def detect_anomalies_lstm(data, scaler=None, sequence_length=50, threshold_multiplier=3):
    try:
        if scaler is None:
            scaler = MinMaxScaler()
            scaled_data = scaler.fit_transform(data.values.reshape(-1, 1))
        else:
            scaled_data = scaler.transform(data.values.reshape(-1, 1))

        def create_sequences(data, seq_length):
            X = []
            for i in range(seq_length, len(data)):
                X.append(data[i-seq_length:i, 0])
            return np.array(X)

        X = create_sequences(scaled_data, sequence_length)
        X = X.reshape((X.shape[0], X.shape[1], 1))
```

Σχήμα 7.6 Ανίχνευση ανωμαλιών με LSTM

- Η συνάρτηση προσαρμόζει τις τιμές των δεδομένων ώστε να βρίσκονται σε ένα συγκεκριμένο εύρος, συνήθως μεταξύ 0 και 1. Αυτή η διαδικασία διευκολύνει την ανάλυση και την επεξεργασία των δεδομένων από το μοντέλο, καθιστώντας την εκπαίδευση πιο αποτελεσματική.
- Δημιουργούνται ακολουθίες δεδομένων με καθορισμένο μήκος (π.χ., 50 χρονικές στιγμές). Κάθε σειρά αποτελείται από τις προηγούμενες τιμές, επιτρέποντας στο μοντέλο να κατανοήσει τα χρονικά μοτίβα και τις τάσεις που υπάρχουν στα δεδομένα.
- Οι δημιουργημένες σειρές παρατηρήσεων οργανώνονται με τέτοιο τρόπο ώστε να μπορούν να αναλυθούν αποτελεσματικά από το μοντέλο LSTM. Αυτή η διαμόρφωση εξασφαλίζει ότι το μοντέλο λαμβάνει τα δεδομένα σε μορφή που του επιτρέπει να κάνει ακριβείς προβλέψεις και να εντοπίζει ανωμαλίες.

```
# Define LSTM model
model = Sequential()
model.add(LSTM(50, return_sequences=True, input_shape=(X.shape[1], 1)))
model.add(LSTM(50, return_sequences=False))
model.add(Dense(25))
model.add(Dense(1))

# Compile and fit the model
model.compile(optimizer='adam', loss='mean_squared_error')
model.fit(X, scaled_data[sequence_length:], epochs=20, batch_size=32, verbose=0)

# Make predictions
predictions = model.predict(X)
predictions = scaler.inverse_transform(predictions)
```

Σχήμα 7.7 Κατασκευή και Εκπαίδευση Μοντέλου LSTM για Ανίχνευση Ανωμαλιών

- Δημιουργείται ένα μοντέλο που μαθαίνει από τις ιστορικές τιμές του Bitcoin, χρησιμοποιώντας δύο επίπεδα ανάλυσης για να κατανοήσει τα περίπλοκα μοτίβα και τις τάσεις των δεδομένων.
- Το μοντέλο εκπαιδεύεται με τα παρελθόντα δεδομένα τιμών, επεξεργαζόμενο τις πληροφορίες μέσα σε πολλαπλές φάσεις για να βελτιώσει την ακρίβεια των προβλέψεων του.

```
    # Calculate errors
    error = data.values[sequence_length:] - predictions.flatten()
    error = pd.Series(error, index=data.index[sequence_length:])

    # Determine threshold
    threshold = threshold_multiplier * error.std()
    anomalies = data.iloc[sequence_length:][error.abs() > threshold]

    return anomalies, error, model, scaler
except Exception as e:
    print(f"Σφάλμα κατά την ανίχνευση ανωμαλιών με LSTM: {e}")
    return pd.Series(), pd.Series(), None, None
```

Σχήμα 7.8 Υπολογισμός σφαλμάτων και Καθορισμός ορίου

```
def plot_candlestick(df_filtered, anomalies_dict, filepath, title):
    try:
        # Prepare additional plots for anomalies
        apds = []
        colors = {'ARIMA': 'red', 'SVM': 'blue', 'LSTM': 'green'}
        for method, anomalies in anomalies_dict.items():
            if not anomalies.empty:
                # Create a Series with NaN and set anomalies
                anomaly_plot = pd.Series(np.nan, index=df_filtered.index)
                anomaly_plot[anomalies.index] = anomalies
                apds.append(mpf.make_addplot(anomaly_plot, type='scatter', markersize=100, marker='o', color=colors.get(method, 'red'), label=method))

        # Define style
        mc = mpf.make_marketcolors(up='g', down='r', inherit=True)
        s = mpf.make_mpf_style(marketcolors=mc)
```

Σχήμα 7.9 Σχεδίαση γραφημάτων

- Αφού ολοκληρωθεί η εκπαίδευση, το μοντέλο χρησιμοποιείται για να προβλέψει μελλοντικές τιμές. Οι προβλέψεις αυτές μετατρέπονται πίσω στην αρχική κλίμακα για να μπορούν να συγκριθούν με τις πραγματικές τιμές και να εντοπιστούν πιθανές ανωμαλίες.
- Υπολογίζει τη διαφορά μεταξύ των πραγματικών τιμών δεδομένων και των προβλέψεων του μοντέλου LSTM.
- Ορίζει ένα όριο ανίχνευσης ανωμαλιών βασισμένο στην τυπική απόκλιση των σφαλμάτων και έναν πολλαπλασιαστή ορίου.
- Εντοπίζει και επισήμανση των δεδομένων που τα σφάλματα τους υπερβαίνουν το καθορισμένο όριο, επισημαίνοντας πιθανές ανωμαλίες.
- Δημιουργεί επιπρόσθετα scatter plots για κάθε μέθοδο ανίχνευσης ανωμαλιών (ARIMA, SVM, LSTM) χρησιμοποιώντας διαφορετικά χρώματα για την απεικόνιση.

```
    mpf.plot(
        df_filtered,
        type='candle',
        mav=(10, 20),
        style=s,
        title=title,
        volume=True,
        addplot=apds,
        figsize=(14, 7),
        savefig=filepath,
        tight_layout=True,
        warn_too_much_data=1000  # Adjust as needed to suppress warnings
    )
    print(f"Candlestick chart saved to {filepath}")
except Exception as e:
    print(f"Σφάλμα κατά τη σχεδίαση candlestick: {e}")
```

Σχήμα 7.10 Δημιουργία γραφημάτων και αποθήκευση με Κινητούς Μέσους Όρους και Ανωμαλίες

- Ρυθμίζει τα χρώματα των αγορών (ανόδους σε πράσινο, πτώσεις σε κόκκινο) και καθορίζει το συνολικό στυλ του candlestick γραφήματος.
- Συνδυάζει τα candlestick δεδομένα με τα ανώμαλα σημεία και αποθηκεύει το τελικό γράφημα στο καθορισμένο αρχείο.

- Χρησιμοποιεί τη βιβλιοθήκη mplfinance για να σχεδιάσει ένα candlestick γράφημα των φιλτραρισμένων δεδομένων με κινητούς μέσους όρους (10 και 20 περιόδους).
- Προσθέτει επιπλέον scatter plots που αντιπροσωπεύουν ανωμαλίες από διάφορες μεθόδους (π.χ., ARIMA, SVM, LSTM) στο κύριο γράφημα, χρησιμοποιώντας διαφορετικά χρώματα για κάθε μέθοδο.
- Αποθηκεύει το τελικό γράφημα στο καθορισμένο αρχείο με καθορισμένες διαστάσεις και στυλ, ενώ διαχειρίζεται πιθανά σφάλματα κατά τη διαδικασία σχεδίασης και αποθήκευσης του γραφήματος.

```python
def plot_residuals(residuals, filepath):
    try:
        plt.figure(figsize=(12, 6))
        plot_acf(residuals)
        plot_pacf(residuals)
        plt.suptitle('Residual ACF & PACF Plots')
        plt.tight_layout()
        plt.savefig(filepath, format='png', dpi=300)
        plt.close()
        print(f"Residuals plot saved to {filepath}")
    except Exception as e:
        print(f"Σφάλμα κατά τη σχεδίαση υπολειπόμενων: {e}")
```

Σχήμα 7.11 Σχεδίαση υπολοιπόμενων

- Η συνάρτηση δημιουργεί και αποθηκεύει τα γραφήματα ACF και PACF για τα υπολείμματα του ARIMA μοντέλου, τα οποία βοηθούν στην ανάλυση της αυτοσυσχέτισης των δεδομένων.
- Τα γραφήματα αποθηκεύονται ως εικόνα PNG σε υψηλή ανάλυση και το διάγραμμα κλείνει για να απελευθερωθούν οι πόροι.
- Η συνάρτηση χειρίζεται σφάλματα και εμφανίζει σχετικό μήνυμα αν κάτι δεν πάει καλά κατά τη διαδικασία.

```python
def generate_pdf_report(model_summary_arima, arima_anomalies, svm_anomalies, lstm_anomalies, img_path_candlestick, residuals_img_path, output_path):
    try:
        print("Starting PDF generation...")
        pdf = CustomPDF()
        pdf.add_page()

        # Περιγραφή μοντέλου ARIMA
        pdf.set_font("Arial", 'B', 12)
        pdf.cell(0, 10, 'ARIMA Model Summary:', ln=True)
        pdf.set_font("Arial", size=12)
        for line in str(model_summary_arima).split('\n'):
            pdf.multi_cell(0, 10, txt=line, align='L')

        # Ανιχνευμένες Ανωμαλίες
        pdf.set_font("Arial", 'B', 12)
        pdf.cell(0, 10, 'Anomalies Detected:', ln=True)
        pdf.set_font("Arial", size=12)
        pdf.cell(0, 10, f"ARIMA Anomalies Detected: {len(arima_anomalies)}", ln=True)
        pdf.cell(0, 10, f"SVM Anomalies Detected: {len(svm_anomalies)}", ln=True)
        pdf.cell(0, 10, f"LSTM Anomalies Detected: {len(lstm_anomalies)}", ln=True)
```

Σχήμα 7.12 Δημιουργία PDF αναφοράς

- Εισάγει την περίληψη του μοντέλου ARIMA στο PDF, παρουσιάζοντας τα βασικά χαρακτηριστικά και αποτελέσματα του μοντέλου.
- Αναφέρει τον αριθμό των ανιχνευμένων ανωμαλιών για κάθε μοντέλο (ARIMA, SVM, LSTM), παρέχοντας μια συγκριτική ανάλυση.

```
# Πίνακας ανωμαλιών
pdf.set_font("Arial", 'B', 12)
pdf.cell(0, 10, 'ARIMA Anomalies:', ln=True)
pdf.set_font("Arial", size=12)
for index, value in arima_anomalies.items():
    pdf.cell(0, 10, txt=f"{index.strftime('%Y-%m-%d')}: {value:.2f}", ln=True, align='L')

pdf.set_font("Arial", 'B', 12)
pdf.cell(0, 10, 'SVM Anomalies (first 50):', ln=True)
pdf.set_font("Arial", size=12)
for index, value in svm_anomalies.head(50).items():
    pdf.cell(0, 10, txt=f"{index.strftime('%Y-%m-%d')}: {value:.2f}", ln=True, align='L')

pdf.set_font("Arial", 'B', 12)
pdf.cell(0, 10, 'LSTM Anomalies (first 50):', ln=True)
pdf.set_font("Arial", size=12)
for index, value in lstm_anomalies.head(50).items():
    pdf.cell(0, 10, txt=f"{index.strftime('%Y-%m-%d')}: {value:.2f}", ln=True, align='L')

# Προσθήκη γραφημάτων
pdf.add_page()
pdf.set_font("Arial", 'B', 12)
pdf.cell(0, 10, 'Candlestick Chart with Anomalies:', ln=True)
pdf.image(img_path_candlestick, x=10, y=20, w=190)

pdf.add_page()
pdf.set_font("Arial", 'B', 12)
pdf.cell(0, 10, 'Residual ACF & PACF Plots:', ln=True)
pdf.image(residuals_img_path, x=10, y=20, w=190)
```

Σχήμα 7.13 Προσθήκη πινάκων ανωμαλιών και γραφημάτων στην PDF αναφορά

- Προσθέτει γραφήματα candlestick και εικόνες υπολοίπων σφαλμάτων στο PDF και αποθηκεύει το ολοκληρωμένο έγγραφο στο καθορισμένο μονοπάτι.
- Εισάγει πίνακες με τις ανιχνευμένες ανωμαλίες από τα μοντέλα ARIMA, SVM και LSTM, παρουσιάζοντας τις ημερομηνίες και τις αντίστοιχες τιμές των ανωμαλιών.
- Προβάλλει τις πρώτες 50 ανωμαλίες για τα μοντέλα SVM και LSTM, εξασφαλίζοντας την ευαναγνωσία του PDF και την αποτελεσματική παρουσίαση των δεδομένων.
- Προσθέτει στο PDF γραφήματα candlestick με ανωμαλίες καθώς και διαγράμματα Residual ACF & PACF, παρέχοντας οπτική ανάλυση και συμπλήρωση των περιγραφικών δεδομένων.

```
    # Save the PDF and check for file generation
    pdf.output(output_path)
    print(f"Report successfully saved to {output_path}")

    # Check if the file was generated
    if not os.path.exists(output_path):
        print(f"Error: The PDF report was not saved successfully. Please check the file path and permissions.")
    else:
        print(f"PDF report saved successfully at {output_path}")

except PermissionError:
    print(f"Permission denied: Unable to save the report to {output_path}. Please check the file path and permissions.")
except Exception as e:
    print(f"Σφάλμα κατά την αποθήκευση της αναφοράς PDF: {e}")
```

Σχήμα 7.14 Αποθήκευση PDF αναφοράς και έλεγχος επιτυχίας δημιουργίας αρχείου

- Χρησιμοποιεί τη μέθοδο output για να αποθηκεύσει το PDF στο καθορισμένο μονοπάτι (output_path), και εμφανίζει μήνυμα επιτυχίας.
- Επαληθεύει αν το αρχείο PDF δημιουργήθηκε επιτυχώς ελέγχοντας την ύπαρξή του στο σύστημα αρχείων, και εκτυπώνει αντίστοιχα μηνύματα.
- Αντιμετωπίζει συγκεκριμένα σφάλματα όπως το PermissionError για προβλήματα δικαιωμάτων πρόσβασης και γενικά σφάλματα, παρέχοντας κατάλληλα μηνύματα σφάλματος για ευκολότερη διάγνωση.

Αυτό το τμήμα του κώδικα είναι σημαντικό για την οπτικοποίηση των δεδομένων, επιτρέποντας στον αναλυτή να δει τις τιμές του Bitcoin με επισημασμένες τις ανωμαλίες.

Κύρια Λειτουργία

- Διαβάζει το αρχείο δεδομένων BTC-USDD.csv, ελέγχει την επιτυχή φόρτωσή του και φιλτράρει τα δεδομένα για την περίοδο από Ιανουάριο 2019 έως Ιανουάριο 2024.
- Εκτελεί ανάλυση με τα μοντέλα ARIMA, SVM και LSTM για την ανίχνευση ανωμαλιών στις τιμές κλεισίματος του Bitcoin, συλλέγοντας τα αποτελέσματα από κάθε μοντέλο.
- Ελέγχει την επιτυχία κάθε βήματος της διαδικασίας, διακόπτει την εκτέλεση σε περίπτωση αποτυχίας και προετοιμάζει τα ανιχνευμένα ανωμαλίες για περαιτέρω επεξεργασία και αναφορά.

```python
if __name__ == "__main__":
    filepath = r'C:\Users\johnny\Desktop\Trade2\BTC-USDD.csv'
    output_path = r'C:\Users\johnny\Desktop\Trade2\Bitcoin_Anomaly_Report.pdf'
    candlestick_img_path = r'C:\Users\johnny\Desktop\Trade2\Bitcoin_Candlestick_Anomalies.png'
    residuals_img_path = r'C:\Users\johnny\Desktop\Trade2\Bitcoin_Residuals.png'

    # Φόρτωση δεδομένων
    df = load_data(filepath)
    if df is None:
        exit()

    # Φιλτράρισμα δεδομένων (Τελευταία 5 χρόνια)
    df_filtered = df[(df.index >= '2019-01-01') & (df.index <= '2024-01-01')]

    # Προετοιμασία δεδομένων
    data = df_filtered['Close']

    # Ανάλυση με ARIMA
    arima_model = prepare_arima_model(data)
    if arima_model is None:
        exit()

    arima_anomalies, residuals = detect_anomalies_arima(data, arima_model)

    # Ανάλυση με SVM
    svm_anomalies = detect_anomalies_svm(data)

    # Ανάλυση με LSTM
    lstm_anomalies, lstm_errors, lstm_model, lstm_scaler = detect_anomalies_lstm(data)
```

Σχήμα 7.15 Φόρτωση δεδομένων και ανίχνευση ανωμαλιών με ARIMA, SVM και LSTM

```python
# Συλλογή όλων των ανωμαλιών
anomalies_dict = {
    'ARIMA': arima_anomalies,
    'SVM': svm_anomalies,
    'LSTM': lstm_anomalies
}

# Σχεδίαση διαγραμμάτων με όλες τις ανωμαλίες
plot_candlestick(df_filtered, anomalies_dict, candlestick_img_path, "Bitcoin Price with Anomalies")

# Σχεδίαση υπολειπόμενων (ARIMA residuals)
plot_residuals(residuals, residuals_img_path)

# Δημιουργία αναφοράς PDF
generate_pdf_report(
    arima_model.summary(),
    arima_anomalies,
    svm_anomalies,
    lstm_anomalies,
    candlestick_img_path,
    residuals_img_path,
    output_path
)
```

Σχήμα 7.16 Συλλογή ανωμαλιών, οπτικοποίηση και δημιουργία PDF αναφοράς

- Δημιουργεί ένα λεξικό (anomalies_dict) που συγκεντρώνει τις ανιχνευμένες ανωμαλίες από τα μοντέλα ARIMA, SVM και LSTM.
- Χρησιμοποιεί τη συνάρτηση plot_candlestick για να σχεδιάσει candlestick γραφήματα με όλες τις ανωμαλίες και τη συνάρτηση plot_residuals για την απεικόνιση των υπολειπόμενων σφαλμάτων του μοντέλου ARIMA.
- Καλεί τη συνάρτηση generate_pdf_report για να δημιουργήσει μια ολοκληρωμένη PDF αναφορά που περιλαμβάνει την περίληψη του μοντέλου ARIMA, τις ανιχνευμένες ανωμαλίες και τα γραφήματα που δημιουργήθηκαν.

7.2 Έκθεση Ανίχνευσης Ανωμαλιών στην Τιμή του Bitcoin

Η ανάλυση της ανίχνευσης ανωμαλιών στην τιμή του Bitcoin πραγματοποιήθηκε με τη χρήση τριών μοντέλων μηχανικής μάθησης: ARIMA, SVM και LSTM. Η μελέτη καλύπτει την περίοδο από τις 24 Ιουλίου 2019 έως τις 23 Οκτωβρίου 2023, με στόχο την ανίχνευση αποκλίσεων από τα αναμενόμενα πρότυπα της τιμής του Bitcoin.

Λογάριθμος Πιθανοφάνειας (Log Likelihood): -13786.74

Ο λογάριθμος πιθανοφάνειας (log likelihood) είναι ένα μέτρο του πόσο καλά προσαρμόζεται το μοντέλο στα δεδομένα. Η τιμή του λογαρίθμου πιθανοφάνειας υποδεικνύει πόσο πιθανό είναι τα δεδομένα που παρατηρούμε να προέρχονται από το συγκεκριμένο μοντέλο. Στη συγκεκριμένη περίπτωση, η τιμή -13786.74 είναι αρκετά χαμηλή, κάτι που δείχνει ότι το μοντέλο έχει κάποιες αποκλίσεις από τα πραγματικά δεδομένα, αλλά είναι αναμενόμενο σε μια τόσο σύνθετη και δυναμική χρονοσειρά όπως η τιμή του Bitcoin.

AIC (Κριτήριο Πληροφορίας Akaike): 27579.48

Το κριτήριο πληροφορίας του Akaike (AIC) είναι ένα μέτρο που χρησιμοποιείται για να αξιολογηθεί η ποιότητα ενός μοντέλου, λαμβάνοντας υπόψη τόσο την καλή προσαρμογή του στα δεδομένα όσο και την πολυπλοκότητά του. Μια χαμηλότερη τιμή AIC είναι συνήθως προτιμότερη, καθώς υποδεικνύει ότι το μοντέλο έχει καλή προσαρμογή χωρίς να είναι υπερβολικά περίπλοκο. Η τιμή 27579.48 είναι ενδεικτική για το πόσο καλά το συγκεκριμένο μοντέλο εξισορροπεί μεταξύ πολυπλοκότητας και ακρίβειας.

BIC (Κριτήριο Πληροφορίας Bayesian): 27595.63

Το κριτήριο πληροφορίας Bayesian (BIC) είναι παρόμοιο με το AIC, αλλά δίνει μεγαλύτερη βαρύτητα στην πολυπλοκότητα του μοντέλου. Όσο χαμηλότερη είναι η τιμή του BIC, τόσο καλύτερη θεωρείται η προσαρμογή του μοντέλου. Σε αυτή την περίπτωση, η τιμή BIC είναι 27595.63, ελαφρώς υψηλότερη από την τιμή AIC, που δείχνει ότι το μοντέλο ενδεχομένως να έχει μια πολυπλοκότητα που πρέπει να λαμβάνεται υπόψη, αλλά εξακολουθεί να είναι αποδεκτό με βάση τα δεδομένα.

Συντελεστές εποχικών AR (Αυτοπαλίνδρομα)

Το ARIMA μοντέλο περιλαμβάνει αυτοπαλινδρομικούς συντελεστές (AR), οι οποίοι χρησιμοποιούν προηγούμενες τιμές της χρονοσειράς για την πρόβλεψη μελλοντικών τιμών. Στην περίπτωσή μας, εξετάζονται δύο συγκεκριμένες χρονικές υστερήσεις:

- **AR με χρονική υστέρηση 12 περιόδων**: Ο συντελεστής αυτός είναι -0.7133, πράγμα που σημαίνει ότι υπάρχει μια ισχυρή αρνητική σχέση μεταξύ της τιμής του Bitcoin πριν από 12 περιόδους και της τρέχουσας τιμής. Με άλλα λόγια, όταν η τιμή του Bitcoin πριν από 12 περιόδους ήταν υψηλή, είναι πιθανό η τρέχουσα τιμή να είναι χαμηλότερη (και το αντίστροφο). Η στατιστική σημαντικότητα ($p<0.001$) δείχνει ότι αυτή η σχέση είναι πολύ ισχυρή και δεν είναι αποτέλεσμα τυχαίας παρατήρησης.
- **AR με χρονική υστέρηση 24 περιόδων**: Ο συντελεστής είναι -0.2744, που δείχνει μια πιο ήπια αρνητική σχέση μεταξύ της τιμής του Bitcoin πριν από 24 περιόδους και της τρέχουσας τιμής. Και εδώ, η στατιστική σημαντικότητα ($p<0.001$) επιβεβαιώνει ότι αυτή η σχέση είναι σημαντική.

Διακύμανση Σφάλματος (sigma2): 1.613 x 10^6

Η διακύμανση σφάλματος (συχνά αναφέρεται και ως σφάλμα του μοντέλου) δείχνει την έκταση των αποκλίσεων των προβλεπόμενων τιμών από τις πραγματικές τιμές. Στην περίπτωση αυτή, η διακύμανση είναι 1.613×10^6, που υποδηλώνει ότι υπάρχουν σημαντικές διακυμάνσεις στις προβλέψεις του μοντέλου. Αυτό είναι λογικό, δεδομένης της υψηλής μεταβλητότητας της τιμής του Bitcoin, η οποία συχνά υπόκειται σε ξαφνικές αλλαγές λόγω οικονομικών και κοινωνικών παραγόντων.

Η ανάλυση των παραπάνω στατιστικών στοιχείων παρέχει μια εικόνα της απόδοσης του μοντέλου ARIMA και το πώς αυτό προσαρμόζεται στις ιστορικές τιμές του Bitcoin. Παρά το γεγονός ότι το μοντέλο παρουσιάζει κάποια αβεβαιότητα, οι στατιστικοί δείκτες επιβεβαιώνουν ότι έχει νόημα και ότι μπορεί να ανιχνεύσει σημαντικές τάσεις και ανωμαλίες στην αγορά του Bitcoin.

Άλλα σημαντικά στατιστικά στοιχεία περιλαμβάνουν το τεστ Ljung-Box (L1) με τιμή *Q = 1.83*, το οποίο δείχνει ότι τα υπολείμματα του μοντέλου δεν παρουσιάζουν σημαντική αυτοσυσχέτιση. Η ασυμμετρία (skew) είναι κοντά στο μηδέν (-0.00) και η κύρτωση (kurtosis) ανέρχεται σε 9.63, υποδεικνύοντας την ύπαρξη ακραίων τιμών στην κατανομή των υπολοίπων.

Σύγκριση με Άλλα Μοντέλα

Πέρα από το ARIMA, χρησιμοποιήθηκαν και τα μοντέλα SVM (Support Vector Machines) και LSTM (Long Short-Term Memory) για την ανίχνευση ανωμαλιών. Συγκεκριμένα, το SVM ανίχνευσε 585 ανωμαλίες, ενώ το LSTM εντόπισε 39 ανωμαλίες. Ορισμένες από αυτές τις ανωμαλίες παρατηρούνται και στα τρία μοντέλα, ενώ άλλες είναι μοναδικές για το κάθε μοντέλο.

Βιβλιογραφία

Aggarwal, C. (2017). *Outlier Analysis .* Springer.

Alfayoumi, S. (2021). *Anomaly Detection*. Retrieved from Knowledge Sharing Platform: https://ksp-windmill-itn.eu/research/anomaly-detection/

Bajaj, A. (2023). *Anomaly detection in time series* . Retrieved from neptune.ai: https://neptune.ai/blog/anomaly-detection-in-time-series

Bakumennko A. & Elragal A. (2022). Detecting Anomalies in FInancial Data Using Machine Learning Algorithms. *Systems*.

Box G. E. P. & Jenkins G. M. . (1970). *TIme Series Analysis: Forecasting and Control.* Holden-Day.

Breunig, M. M. (2000). LOF: Identifying Density-Based Local Outliers.

Chalapathy, R. &. (2019). Deep Learning for Anomaly Detection: A Survey.

Chandola, V. B. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 1-58.

Chandola, V. B. (2012). Anomaly Detection for Discrete Sequences: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 823-839.

Chollet, F. (2018). *Deep Learning with Python.* NY, USA : Manning Publications Co, Shelter Island.

Cover, T. &. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 21-27.

Crépey S., L. N. (2022). *Anomaly Detection on Financial Time Series by Principal Component Analysis and Neural Networks.*

Darban, Z. Z. (2024). *Deep learning for time Series Anomaly Detection: A survey.* ACM Computing Surveys.

Deep learning for anomaly detection. (n.d.). Retrieved from Fastforward labs: https://ff12.fastforwardlabs.com/

Fan, J. L. (2021). A transfer learning architecture based on a support vector machine for histopathology image classification. *Applied Sciences*.

Goh, C. K. (2018). *Hybrid Models for Financial Time Series Prediction.* Springer.

Gupta, M. G. (2014). Outlier Detection for Temporal Data: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 2250-2267.

Hawkins, D. M. (1980). *Identification of Outliers.* Chapman and Hall.

Hodge, V. J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 85-126.

Hyndman, R. J. (2018). Forecasting: Principles and Practice. *OTexts*.

Introduction to recurrent neural network. (2024). Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/

Karageorgiou, N. (2019). *Application of k-NN and ARIMA methods in anomaly detection in financial data.* Aristotle University of Thessaloniki: Thesis.

Korompilia, A. (2017). *Detection of financial anomalies using neural network and machine learning methods.* Athens University of Economics and Business: Thesis.

Liu, F. T. (2008). Isolation Forest. *Proceedings of the 2008 IEEE International Conference on Data Mining.*

Malhotra, P. V. (2015). Long Short-Term Memory Networks for Anomaly Detection in Time Series. *Proceedings of the 23rd European Symposium on Artificial Neural Networks (ESANN).*

Manolakou, E. (2018). *Methodology of anomaly detection in timeseries data.* University of Patras: Thesis.

Mendes-Neves, T. S. (2024). Estimating the Likelihood of Financial Behaviours Using Nearest Neighbors. *Computational Economics*, 1477–1491 .

Navon, A. &. (2015). Financial Time Series Prediction using Deep Learning. *Journal of Latex Class Files*.

Pang, G. S. (2021). Deep Learning for Anomaly Detection: A Review. *ACM Computing Surveys*, 1-38.

Saxena, S. (2024). *What is LSTM? Introduction to Long Short-Term Memory*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/

Support Vector Machine (SVM) algorithm. (2024). Retrieved from GeeksforGeeks.: https://www.geeksforgeeks.org/support-vector-machine-algorithm/

Support vector machine algorithm . (n.d.). Retrieved from Acervo Lima: https://acervolima.com/algoritmo-de-maquina-de-vetor-de-suporte/#google_vignette

Team, G. (2023). *Deep learning*. Retrieved from Greeco: https://greeco.gr/business/techniti-noimosyni/vathiamathisi/#%CE%9A%CE%B1%CF%84%CE%B1%CE%BD%CF%8C%CE%B7%CF%83%CE%B7_%CF%84%CE%B7%CF%82_%CE%BC%CE%B7%CF%87%CE%B1%CE%BD%CE%B9%CE%BA%CE%AE%CF%82_%CE%B

Tiwari S., R. H. (2021). "Machine Learning in Financial Market Surveillance: A Survey," . *in IEEE Access*.

What is LSTM Long Short Term Memory? . (2024). Retrieved from GeeksforGeeks.: https://www.geeksforgeeks.org/deep-learning-introduction-to-long-short-term-memory/

Xu, H. L. (2020). Anomaly Detection in Time Series: A Comprehensive Survey.

Zhao, Y. N. (2019). PyOD: A Python Toolbox for Scalable Outlier Detection. Journal of Machine Learning Research. 1-7.