ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Πρόβλεψη τιμής Ether με χρήση τεχνικών μηχανικής μάθησης

ΠΟΛΙΤΗΣ Σ. ΑΓΗΣΙΛΑΟΣ

Επιβλέπων: Νεκτάριος Κοζύρης

Καθηγητής, Ε.Μ.Π

Αθήνα, Οκτώβριος 2020

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΥΠΟΛΟΓΙΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Πρόβλεψη τιμής Ether με χρήση τεχνικών μηχανικής μάθησης

ΠΟΛΙΤΗΣ Σ. ΑΓΗΣΙΛΑΟΣ

Επιβλέπων: Νεκτάριος Κοζύρης

Καθηγητής, Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 13/11/2020

……	……	……
Νεκτάριος Κοζύρης	Γεώργιος Γκούμας	Ιωάννης Κωνσταντίνου
Καθηγητής Ε.Μ.Π	Επίκουρος Καθηγητής Ε.Μ.Π	Επίκουρος Καθηγητής Π.Θ.

…….

Πολίτης Αγησίλαος

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π

Περίληψη

Η ραγδαία ανάπτυξη των κρυπτονομισμάτων τα τελευταία χρόνια τα έχει καταστήσει ως ένα από τα πλέον σημαντικά επενδυτικά κεφάλαια που ανταλλάσσονται καθημερινά. Ωστόσο οι συνεχείς και αστάθμητες διακυμάνσεις στην τιμή τους καθιστούν τη διαδικασία πρόβλεψης της αξίας τους ένα ιδιαίτερα απαιτητικό πρόβλημα που εξαρτάται από πολλούς παράγοντες και η επίλυση του οποίου έχει τη δυνατότητα να παρέχει σημαντικά περιθώρια οικονομικού κέρδους. Στην παρούσα εργασία εξετάζουμε τους παράγοντες που μπορούν να επηρεάσουν την τιμή του Ether χρησιμοποιώντας ένα σύνολο τεχνικών επιλογής χαρακτηριστικών και λαμβάνοντας υπόψιν ένα ευρύ φάσμα χαρακτηριστικών που σχετίζονται με την αγορά και το χρηματιστήριο, την δημοφιλία του Ether, τεχνικούς δείκτες της τιμής του και παραμέτρους του Ethereum blockchain. Παράλληλα αναπτύσσουμε ένα σύνολο μοντέλων μηχανικής μάθησης: LSTM, GRU, XGBoost και Ensemble που αποτελεί των συνδυασμό των τριών αυτών μοντέλων με στόχο την πρόβλεψη της τιμής του Ether σε χρονικό παράθυρο μίας και εφτά ημερών. Οι προβλέψεις της τιμής πραγματοποιούνται γύρω από δύο άξονες 1) πρόβλεψη της ακριβούς τιμής του Ether (πρόβλημα regression) και 2) πρόβλεψη αύξησης ή μείωσης της τιμής του (πρόβλημα classification). Τα μοντέλα GRU και XGBoost έχουν τις καλύτερες επιδόσεις στο regression πρόβλημα για χρονικό πλαίσιο μίας και εφτά ημερών αντίστοιχα ενώ το Ensemble μοντέλο αναδεικνύεται το καλύτερο στο classification και στις δύο περιπτώσεις σημειώνοντας προβλέψεις με ακρίβεια 82.46% και 77.19%. Οι προτεινόμενες μέθοδοι μπορούν να εφαρμοστούν σε ένα μεγάλο σύνολο κρυπτονομισμάτων και αναδεικνύουν τη δυνατότητα για προβλέψεις της τιμής τους με μεγάλη ακρίβεια.

Λέξεις κλειδιά: Μηχανική μάθηση, νευρωνικά δίκτυα, επιλογή χαρακτηριστικών, κρυπτονομίσματα, LSTM, GRU, XGBoost, Ethereum.

Abstract

The rapid growth of cryptocurrencies in recent years has made them one of the most important investment funds traded daily. However, the unstable nature of their price makes the process of predicting their value a very demanding problem that depends on many factors and the solution of which has the potential to provide significant financial profit margins. In this paper we examine the factors that can affect the price of Ether by using a set of feature selection techniques and taking into account a wide range of features that have to do with trading and market, Ether popularity, technical price indicators and parameters of the Ethereum blockchain. We develop a set of machine learning models: LSTM, GRU, XGBoost and Ensemble which is a combination of these three models in order to predict the price of Ether in a time window of one and seven days. Price predictions are made around two axes 1) forecast of the exact price of Ether (regression problem) and 2) forecast increase or decrease of its price (classification problem). GRU and XGBoost models have the best performance in the regression problem for a period of one and seven days respectively, while the Ensemble model emerges as the best in the classification in both cases, achieving forecasts with an accuracy of 82.46% and 77.19%. The proposed methods can be applied to a large set of cryptocurrencies and highlight the ability to predict their price with great accuracy.

Keywords: machine learning, deep learning, recurrent neural networks, feature selection, LSTM, GRU, XGBoost, Ensemble learning, cryptocurrencies, time series, Random Forest, regression, classification, price prediction, Ethereum.

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Πρόβλεψη τιμής κρυπτονομισμάτων

Η υλοποίηση της ιδέας των κρυπτονομισμάτων πραγματοποιήθηκε για πρώτη φορά το 2008 από έναν ανώνυμο ερευνητή με το ψευδώνυμο Satoshi Nakamoto, ο οποίος μέσα από την εργασία του «Bitcoin: A Peer-to-Peer (P2P) Electronic Cash System» [1] δημιούργησε το πρώτο κρυπτονόμισμα με όνομα Bitcoin. Οι αρχές της αποκεντροποίησης της δύναμης, της εξάλειψης των μεσαζόντων και της ελευθερίας των συναλλαγών που πρέσβευαν τα κρυπτονομίσματα ως ανταλλακτικό μέσο οδήγησε στην ραγδαία αύξηση της δημοφιλίας τους και στην άμεση δημιουργία πολλών νέων κρυπτονομισμάτων, με το σημαντικότερο από αυτά να είναι το Ether.

Ωστόσο με το πέρασμα των χρόνων πέρα από την χρήση τους ως ανταλλακτικό μέσο αναδείχθηκε και η φύση τους ως ένα επενδυτικό κεφάλαιο, καθώς παρόμοια με τις παραδοσιακές μετοχές τα κρυπτονομίσματα μπορούν να αγοραστούν και να ανταλλαχθούν. Μάλιστα τα τελευταία χρόνια σημειώνονται πολύ υψηλά ποσά επενδύσεων στην αξία των κρυπτονομισμάτων γεγονός που αναδεικνύεται και από την Εικόνα 1. στην οποία παρουσιάζονται τα χρηματικά ποσά που επενδύονται σε δολάρια στο σύνολο των κρυπτονομισμάτων για την χρονική περίοδο 2014 έως 2019.

Εικόνα 1. Επενδύσεις σε κρυπτονομίσματα (2014-2019)

Είναι λοιπόν εμφανές ότι η πρόβλεψη της τιμής τους με υψηλά επίπεδα ακρίβειας είναι δυνατόν να παρέχει σημαντικά περιθώρια οικονομικού κέρδους. Ωστόσο η τιμή των κρυπτονομισμάτων επηρεάζεται από ένα ευρύ φάσμα παραγόντων όπως τεχνικά χαρακτηριστικά του blockchain, δημοφιλία και τάσεις της αγοράς [2] γεγονός που οδηγεί σε συνεχείς και αστάθμητες διακυμάνσεις της καθιστώντας την πρόβλεψή της ένα ιδιαίτερα απαιτητικό πρόβλημα. Για τον λόγο αυτό η τιμή των κρυπτονομισμάτων έχει συγκεντρώσει το ενδιαφέρον ενός μεγάλου πλήθους ερευνητών με τις προσπάθειες να εστιάζουν στους παράγοντες που την επηρεάζουν και στις μεθόδους με τις οποίες μπορεί να πραγματοποιηθεί επιτυχώς η πρόβλεψή της.

1.2 Στόχος εργασίας

Στα πλαίσια της παρούσας εργασίας θα ερευνηθούν οι παράγοντες που μπορούν να επηρεάσουν την πορεία της τιμής του Ether και θα αναπτυχθεί ένα σύνολο μοντέλων μηχανικής και βαθιάς μάθησης, τα οποία θα αξιοποιούν τα παραπάνω δεδομένα στοχεύοντας σε όσο το δυνατόν ακριβέστερες προβλέψεις. Πιο συγκεκριμένα θα αναπτυχθούν τα επαναληπτικά νευρωνικά δίκτυα LSTM και GRU, το μοντέλο ενίσχυσης κλίσης XGBoost καθώς και ένας συνδυασμός τους με στόχο την περαιτέρω αύξηση της απόδοσης τους. Επιπλέον για λόγους σύγκρισης με τα παραδοσιακά στατιστικά μοντέλα θα αναπτυχθεί και ένα μοντέλο ARIMA. Θα πραγματοποιηθούν προβλέψεις για χρονικά παράθυρα μίας και εφτά ημερών στο μέλλον και οι οποίες θα κινούνται γύρω από δύο άξονες:

- Πρόβλεψη της ακριβούς τιμής του Ether (regression)
- Πρόβλεψη αύξησης ή μείωσης της τιμής του Ether (classification)

Επιλέχθηκε να μελετηθεί και το classification πρόβλημα καθώς η ακριβής πρόβλεψη της τιμής ενός κρυπτονομίσματος είναι ένα ιδιαίτερα δύσκολο εγχείρημα, ιδιαίτερα για πιο μακροχρόνιες προβλέψεις, με αποτέλεσμα ορισμένες φορές να μην είναι δυνατή η ακριβής προσέγγιση της. Στις περιπτώσεις αυτές η γνώση της πορείας της τιμής (αύξηση ή μείωση) που θα μας δώσει το classification πρόβλημα είναι ιδιαίτερα σημαντική και μπορεί να βοηθήσει τους επενδυτές στην λήψη χρήσιμων αποφάσεων . Επίσης αξίζει να σημειωθεί ότι η επιλογή των μοντέλων μηχανικής μάθησης έναντι των παραδοσιακών στατιστικών μοντέλων έγινε διότι οι τιμές των κρυπτονομισμάτων αποτελούν χρονοσειρές με μεγάλη μεταβλητότητα και δεν παρουσιάζουν εποχικότητα, φαινόμενο απαραίτητο για την υψηλή απόδοση των στατιστικών μοντέλων [3] .

Η επιλογή του Ether ως το κρυπτονόμισμα του οποίου η τιμή θα προβλεφθεί έγινε για το γεγονός ότι η πρόβλεψη της αξίας του Ether και των παραγόντων που την

επηρεάζουν δεν έχει μελετηθεί αρκετά στην διεθνή βιβλιογραφία καθώς η πλειοψηφία των ερευνών εστιάζει στο Bitcoin. Παράλληλα η φύση του διαφέρει από αυτήν του Bitcoin καθώς αποτελεί το κρυπτονόμισμα του Ethereum blockchain το οποίο αποτελεί μια πλατφόρμα για την ανάπτυξη κατανεμημένων εφαρμογών και όχι απλά ένα σύστημα πληρωμών όπως αυτό του Bitcoin. Πιο συγκεκριμένα, οι τρόποι με τους οποίους ανταλλάσσεται το Ether είναι διαφορετικοί από αυτούς του Bitcoin καθώς μία μεταφορά ενός ποσού από Ether μπορεί να μεταβιβαστεί μεταξύ δύο λογαριασμών μέσω ενός έξυπνου συμβολαίου γεγονός που δεν μπορεί να συμβεί στο σύστημα του Bitcoin [4].

1.3 Συναφείς έρευνες

Η πρόβλεψη της τιμής κρυπτονομισμάτων και ιδιαίτερα του Bitcoin αποτελεί ένα πρόβλημα το οποίο όπως αναφέρθηκε και προηγουμένως έχει κεντρίσει το ενδιαφέρον πολλών ερευνητών. Ερευνητές από διάφορα επιστημονικά παιδία έχουν χρησιμοποιήσει ποικίλες τεχνικές για να εξερευνήσουν τους παράγοντες που επηρεάζουν την τιμή των κρυπτονομισμάτων. Παράλληλα έχουν χρησιμοποιηθεί μοντέλα τόσο από τον χώρο της στατιστικής όσο και της μηχανικής και βαθιάς μάθησης με στόχο την πρόβλεψη της πορείας της τιμής τους. Οι Chen κ.ά. [4] ανέπτυξαν ένα σύνολο μοντέλων μηχανικής μάθησης όπως Random Forest, ANN, RNN, Naive Bayes, SVM και Logistic Regression καθώς και το στατιστικό μοντέλο ARIMA με στόχο να μελετήσουν την πορεία της τιμής του Ether ανά ώρα. Την καλύτερη απόδοση παρουσίασε το μοντέλο ARIMA σημειώνοντας accuracy 61.17%. Οι Wu κ.ά. [5] ανέπτυξαν ένα LSTM χρησιμοποιώντας τους ACF και PACF ελέγχους για να καθορίσουν τον κατάλληλο αριθμό από παρελθοντικές ημέρες που θα χρησιμοποιούσε το μοντέλο και κατάφεραν να προβλέψουν την τιμή του Bitcoin σημειώνοντας σφάλμα RMSE ίσο με 247.33. Οι McNally κ.ά. [6] σύγκριναν την απόδοση των μοντέλων LSTM, RNN και ARIMA για την πρόβλεψη της τιμής του Bitcoin την επόμενη μέρα και κατέληξαν στο LSTM που παρουσίασε accuracy 52.78%. Οι Glenski κ.ά. [7] μελέτησαν την επιρροή των κοινωνικών δικτύων στην πορεία της τιμής των Bitcoin, Ether και Monero. Πιο συγκεκριμένα χρησιμοποίησαν δεδομένα από τις πλατφόρμες Reddit και Github σε συνδυασμό με τα ημερήσια ιστορικά δεδομένα της τιμής των παραπάνω κρυπτονομισμάτων τα οποία εισήγαγαν σε δίκτυα LSTM. Κατέληξαν στο ότι τα δεδομένα αυτά επιδρούν θετικά στην πρόβλεψη της τιμής των κρυπτονομισμάτων πετυχαίνοντας σφάλμα MAPE 4.71± 0.96 για την πρόβλεψη της τιμής του Ether την επόμενη μέρα. Οι Chen κ.ά. [8] μελέτησαν την απόδοση πολλών στατιστικών και μοντέλων μηχανικής μάθησης για την πρόβλεψη της ημερήσιας τιμής αλλά και της τιμής υψηλής συχνότητας (ανά 5 λεπτά) του Bitcoin. Κατέληξαν στο ότι τα στατιστικά μοντέλα αποδίδουν καλύτερα στις ημερήσιες προβλέψεις ενώ τα μοντέλα μηχανικής μάθησης στις προβλέψεις υψηλής συχνότητας με τα Logistic Regression και LSTM να πετυχαίνουν αντίστοιχα τις καλύτερες αποδόσεις με accuracy 66% και 67.2%. Οι Kumar κ.ά.[9] σύγκριναν

την επίδοση των MLP και LSTM στην πρόβλεψη της τιμής του Ether ανά μέρα, ώρα και λεπτό και κατέληξαν στο ότι το LSTM παρουσιάζει τα καλύτερα αποτελέσματα με RMSE 20.53, 7.12 και 18.16 αντίστοιχα. Τέλος, οι Patel κ.ά. [10] χρησιμοποίησαν ένα υβριδικό δίκτυο LSTM – GRU δίκτυο για την πρόβλεψη της τιμής των Monero και Litecoin σημειώνοντας MAPE 4.94% και 19.35% αντίστοιχα.

1.4 Οργάνωση της εργασίας

Στο Κεφάλαιο 2 περιγράφεται το θεωρητικό υπόβαθρο της εργασίας η γνώση του οποίου κρίνεται ιδιαίτερα σημαντική στην κατανόηση της συνολικής εργασίας. Περιγράφονται εκτενώς τα μοντέλα βαθιάς και μηχανικής μάθησης που αναπτύσσονται, οι μετρικές αξιολόγησης που χρησιμοποιούνται και η θεωρία από διάφορα στάδια της προεπεξεργασίας των δεδομένων. Επιπλέον παρέχονται κάποιες βασικές πληροφορίες για το Ethereum blockchain που θα βοηθήσουν στην βαθύτερη κατανόηση της φύσης του κρυπτονομίσματος που μελετάμε.

Στο Κεφάλαιο 3 αναφέρονται αρχικά οι πηγές από τις οποίες συλλέχθηκαν τα δεδομένα καθώς και γίνεται μια συνοπτικής περιγραφή της σημασίας τους. Έπειτα αναπτύσσονται οι διάφορες τεχνικές επιλογής χαρακτηριστικών στα σύνολα δεδομένων για να καθοριστούν τα χαρακτηριστικά που θα χρησιμοποιηθούν στις ημερήσιες και εβδομαδιαίες προβλέψεις.

Στο Κεφάλαιο 4 αρχικά παρουσιάζονται τα στάδια της προεπεξεργασίας των δεδομένων που προέκυψαν από το προηγούμενο κεφάλαιο ώστε να αξιοποιηθούν με όσο πιο αποδοτικό τρόπο από τους αλγορίθμους μηχανικής μάθησης στη συνέχεια. Έπειτα αναπτύσσονται τα διάφορα μοντέλα για της ημερήσιες και εβδομαδιαίες προβλέψεις, γίνεται η βελτιστοποίηση των υπερπαραμέτρων τους και παρουσιάζονται τα αποτελέσματα των προβλέψεων τους για το regression και το classification πρόβλημα.

Τέλος στο Κεφάλαιο 5 γίνεται μία σύνοψη της μεθοδολογίας που ακολουθήθηκε και των συμπερασμάτων που παράχθηκαν και προτείνονται πιθανές κατευθύνσεις μελλοντικής έρευνας.

Κεφάλαιο 2

Θεωρητικό υπόβαθρο

Στο κεφάλαιο αυτό θα αναπτυχθεί το απαραίτητο θεωρητικό υπόβαθρο που απαιτείται για την πλήρη κατανόηση και πληρότητα της παρούσας διπλωματικής εργασίας.

2.1 Ethereum

Το Ethereum αποτελεί ένα blockchain ανοιχτού κώδικα που δημιουργήθηκε από τον Vitalik Buterin το 2015 [11] και ενώ βασίστηκε σε μεγάλο βαθμό σε λειτουργίες του Bitcoin παρουσιάζει σημαντικές διαφορές από αυτό. Η σημαντικότερη διαφοροποίηση που αποτελεί και τον βασικό λόγο για την ανάπτυξη του Ethereum είναι ότι εκτός από την λειτουργία του ως ένα σύστημα πληρωμών μέσω του κρυπτονομίσματος του (Ether) αποτελεί ένα blockchain που περιλαμβάνει μία Turing-complete γλώσσα προγραμματισμού μέσω της οποίας μπορούν να αναπτυχθούν κάθε ειδών κατανεμημένες εφαρμογές. Επιπλέον παρέχει τη δυνατότητα ανάπτυξης έξυπνων συμβολαίων (smart contracts) τα οποία αποτελούν κώδικα που εκτελείται αυτόματα όταν πληρούνται συγκεκριμένες συνθήκες στο Ethereum blockchain.

Όπως ήδη αναφέρθηκε το Ether αποτελεί το κρυπτονόμισμα του Ethereum συστήματος μέσω του οποίου πραγματοποιούνται όλες οι εσωτερικές πληρωμές. Μία επιπλέον σημαντική διαφοροποίηση του Ethereum από το Bitcoin είναι το γεγονός ότι περιλαμβάνει δύο είδη λογαριασμών: τους εξωτερικά ελεγχόμενους λογαριασμούς (externally owned accounts), που ελέγχονται από ένα ιδιωτικό κλειδί και τους λογαριασμούς συμβολαίου (contract accounts) που ελέγχονται από τον κωδικό του συμβολαίου. Ένας εξωτερικός λογαριασμός δεν περιλαμβάνει κώδικα και μπορεί να στέλνει μηνύματα σε άλλους λογαριασμούς δημιουργώντας και υπογράφοντας μία συναλλαγή. Ένας λογαριασμός συμβολαίου περιλαμβάνει κώδικα ο οποίος ενεργοποιείται σε περίπτωση λήψης μηνύματος και μπορεί να στείλει μηνύματα ή να δημιουργήσει άλλα έξυπνα συμβόλαια ως απάντηση. Αναδεικνύεται λοιπόν η σημαντική διαφορά που αναφέρθηκε και στην εισαγωγή σχετικά με τους διαφορετικούς τρόπους που διακινούνται τα Ether σε σχέση με τα Bitcoin. Μία μεταβίβαση Ether μπορεί να προέλθει είτε από έναν εξωτερικό λογαριασμό είτε μέσω

ενός έξυπνου συμβολαίου σε αντίθεση με το Bitcoin που διαθέτει μόνο εξωτερικούς λογαριασμούς.

Τέλος, όπως αναφέρθηκε και πριν η λειτουργικότητα του Ethereum δεν περιορίζεται απλά στα κρυπτονομίσματα αλλά επεκτείνεται σε ένα πλήθος κατανεμημένων εφαρμογών όπως κατανεμημένη αποθήκευση αρχείων και εφαρμογές κατανεμημένων υπολογισμών.

2.2 Χρονοσειρές - Στατιστικά μοντέλα

Μία χρονοσειρά είναι μια ακολουθία αριθμητικών τιμών σε διαδοχική σειρά. Σαν χρονοσειρά μπορεί να θεωρηθεί οποιαδήποτε τιμή μεταβάλλεται με τον χρόνο με χαρακτηριστικά παραδείγματα να αποτελούν οι τιμές μετοχών και κρυπτονομισμάτων, η ζήτηση και οι πωλήσεις προϊόντων κ.ά. Το πρόβλημα της πρόβλεψης χρονοσειρών αναφέρεται στην πρόβλεψη της μελλοντικής τιμής μιας χρονοσειράς κάποια χρονική στιγμή $t + h$ χρησιμοποιώντας μόνο δεδομένα έως και την παροντική χρονική στιγμή t.

Για την επίλυση προβλημάτων πρόβλεψης χρονοσειρών χρησιμοποιούνται σε μεγάλο βαθμό στατιστικές μέθοδοι, όπως οι SMA (Simple Moving Average), Holt's exponential smoothing και ARIMA (Autoregressive Integrated Moving Average). Πιο συγκεκριμένα η μέθοδος ARIMA αποτελεί μία από τις πιο ευρέως χρησιμοποιούμενες στατιστικές μεθόδους για την πρόβλεψη χρονοσειρών λόγω της ευκολίας χρήσης της και των υψηλών της αποδόσεων. Ένα μοντέλο ARIMA αποτελείται από τα εξής τρία μέρη:

- Autoregression (AR) : εκφράζει την εξάρτηση της τιμής της χρονοσειράς από τις τιμές της σε προηγούμενες χρονικές περιόδους.

- Integrated (I) : αναφέρεται στο πόσες αφαιρέσεις διαδοχικών τιμών πρέπει να πραγματοποιηθούν για να γίνει η χρονοσειρά στατική[1].

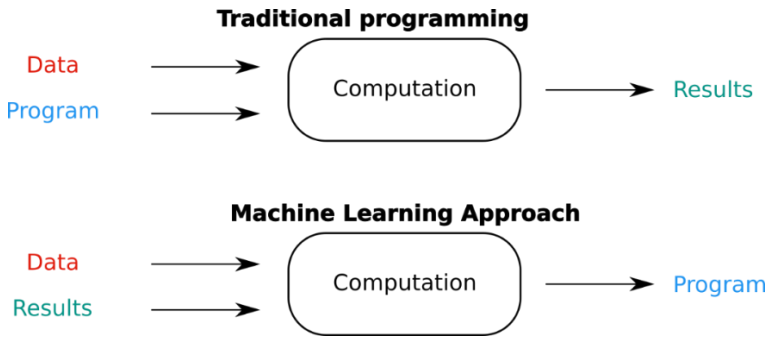
- Moving Average (MA) : ενσωματώνει την εξάρτηση μεταξύ μιας παρατήρησης και ενός υπολειπόμενου σφάλματος από ένα μοντέλο κινητού μέσου όρου που εφαρμόζεται σε προηγούμενες παρατηρήσεις.

Για την ανάπτυξη του μοντέλου ARIMA στη συνέχεια έγινε χρήση της βιβλιοθήκης statsmodels[2] της Python. Οι παράμετροι p, d, q του μοντέλου ARIMA αναφέρονται στους όρους AR, I και MA αντίστοιχα.

[1] Στατική είναι η χρονοσειρά της οποίας οι ιδιότητες όπως η μέση τιμή και η διακύμανση δεν μεταβάλλονται με τον χρόνο.
[2] https://www.statsmodels.org/dev/generated/statsmodels.tsa.arima_model.ARIMA.html

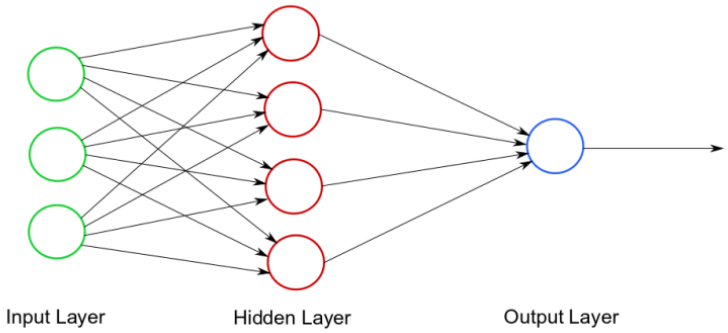
2.3 Μοντέλα μηχανικής και βαθιάς μάθησης

Η έννοια της μηχανικής μάθησης, η οποία αποτελεί υπόκειται στο ευρύτερο πεδίο της τεχνητής νοημοσύνης, και εκφράζει το σύνολο των αλγορίθμων που έχουν την δυνατότητα να «μαθαίνουν» κατά τη διάρκεια της εκτέλεσής τους, να αναγνωρίζουν δηλαδή μοτίβα και εξαρτήσεις στα δεδομένα που χρησιμοποιούν χωρίς να έχουν προγραμματιστεί ρητά γι' αυτό. Η ιδιότητα αυτή των αλγορίθμων μηχανικής μάθησης φαίνεται και στην Εικόνα 2. σε αντιπαράθεση τους συμβατικά προγράμματα που απαιτούν ένα σύνολο κανόνων και δεδομένα εισόδου για να καταλήξουν σε αποτέλεσμα.



Εικόνα 2. Μηχανική μάθηση vs Συμβατικά προγράμματα

Η έννοια της βαθιάς μάθησης προτάθηκε αρχικά από τον Hinton το 1986 [12] και βασίζεται στην δημιουργία υπολογιστικών συστημάτων που θα προσομοιώνουν τη λειτουργία εκμάθησης του ανθρωπίνου εγκεφάλου, τα λεγόμενα νευρωνικά δίκτυα. Ωστόσο η ανάπτυξη και ευρεία χρήση των νευρωνικών δικτύων σημειώθηκε κυρίως τα τελευταία 10 χρόνια λόγω των πολύ υψηλών τους απαιτήσεων σε υπολογιστική ισχύ. Στην Εικόνα 3. παρουσιάζεται η μια από τις πιο απλές μορφές ενός νευρωνικού δικτύου εμπρόσθιας τροφοδότησης.

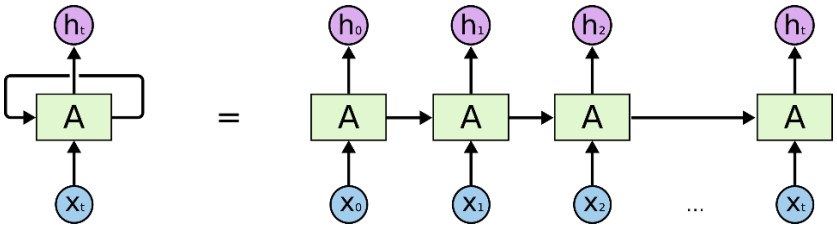


Εικόνα 3. Νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης

Στην παρούσα εργασία θα αναπτυχθεί ένας ειδικός τύπος νευρωνικών δικτύων που ονομάζονται επαναληπτικά νευρωνικά δίκτυα και πιο συγκεκριμένα τα δίκτυα LSTM (Long-Short Term Memory) και GRU (Gated Recurrent Unit) τα οποία ειδικεύονται στην πρόβλεψη χρονοσειρών.

2.3.1 Επαναληπτικά νευρωνικά δίκτυα (RNN)

Ένα επαναληπτικό νευρωνικό δίκτυο (recurrent neural network) αποτελεί έναν τύπο νευρωνικού δικτύου εμπρόσθιας τροφοδότησης το οποίο σε αντίθεση με τα συμβατικά νευρωνικά δίκτυα χαρακτηρίζεται από την ιδιότητα της εσωτερικής μνήμης. Τα RNN είναι ουσιαστικά νευρωνικά δίκτυα που έχουν αναδρομικές συνδέσεις γεγονός που επιτρέπει σε πληροφορίες από προηγούμενα στάδια να παραμένουν στο δίκτυο. Οι αναδρομικές αυτές συνδέσεις μπορούν να γίνουν ευκολότερα κατανοητές αν σκεφτεί κανείς ένα RNN ως πολλά αντίγραφα ενός δικτύου τα οποία περνάνε μηνύματα κάθε φορά στο επόμενο τους, όπως φαίνεται και στην Εικόνα 4.



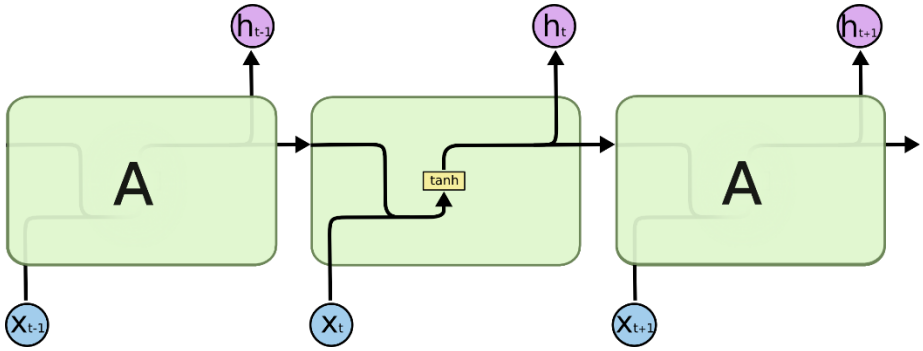
Εικόνα 4. Ξεδιπλωμένο RNN

Λόγω αυτής τους της φύσης έχουν την δυνατότητα να εξάγουν χρονικά χαρακτηριστικά από τα δεδομένα γεγονός που τα καθιστά ιδιαίτερα αποτελεσματικά σε προβλήματα όπου τα δεδομένα παρουσιάζουν χρονικές εξαρτήσεις όπως αναγνώριση ομιλίας, πρόβλεψη χρονοσειρών και επισήμανση εικόνων. Ωστόσο τα RNN παρουσιάζουν το vanishing gradient πρόβλημα [13] με αποτέλεσμα να μην μπορούν να μάθουν εύκολα χρονικές εξαρτήσεις μεγάλου εύρους. Τη λύση στο πρόβλημα αυτό παρέχουν τα δίκτυα LSTM και GRU που θα αναλυθούν στην συνέχεια.

2.3.2 Long Short Term Memory (LSTM)

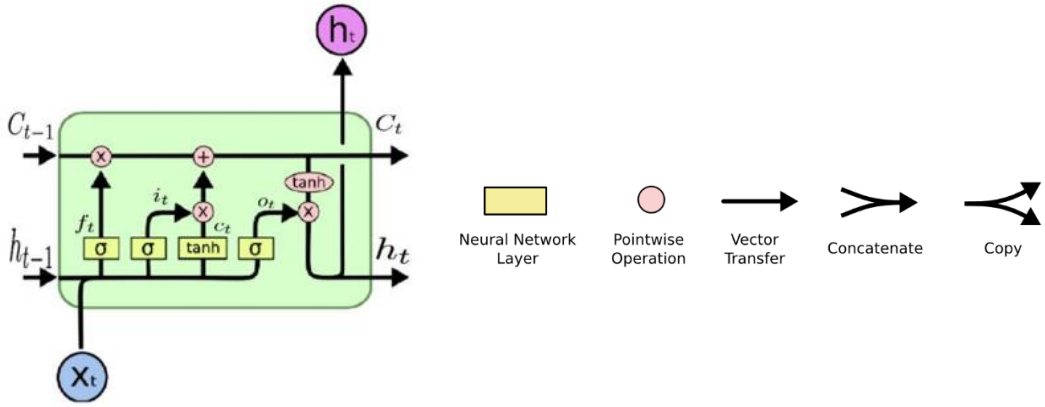
Τα LSTM είναι ένας ειδικός τύπος RNN που προτάθηκαν για πρώτη φορά από τους Hochreiter και Schmidhuber το 1997 [14] και έχουν την δυνατότητα να μαθαίνουν εξαρτήσεις μεγάλου χρονικού εύρους. Όπως όλα τα RNN, έτσι και τα LSTM έχουν τη μορφή μιας αλυσίδας επαναλαμβανόμενων μονάδων από νευρωνικά δίκτυα με τη

διαφορά όμως ότι η δομή αυτών των μονάδων τους είναι αρκετά πιο περίπλοκή από αυτή των απλών RNN η οποία αποτελείται από ένα επίπεδο με μία tanh συνάρτηση ενεργοποίησης (Εικόνα 5.)



Εικόνα 5. Δομή μονάδας RNN

Αντίθετα η δομής κάθε LSTM μονάδας αποτελείται από τέσσερα επίπεδα νευρωνικών δικτύων αντί για ένα τα οποία οφείλονται και για τις ικανότητες των LSTM στην αναγνώριση μακροχρόνιων αλλά και βραχυχρόνιων εξαρτήσεων. Η δομή μιας LSTM μονάδας παρουσιάζεται στην Εικόνα 6.μαζί με την επεξήγηση των συμβόλων που χρησιμοποιούνται.



Εικόνα 6. Δομή μονάδας LSTM

Η βασική ιδέα λειτουργίας του LSTM είναι η κατάσταση του κυττάρου (cell state) που απεικονίζεται με την μεταβλητή *C* και η οποία αναπαριστά τη ροή της πληροφορίας. Η ροή αυτής της πληροφορίας ελέγχεται από πύλες οι οποίες έχουν τη δυνατότητα να προσθέσουν ή να αφαιρέσουν πληροφορία από την κατάσταση κυττάρου. Όλες οι μεταβλητές που θα παρουσιαστούν είναι διανύσματα και οι

αντίστοιχες πράξεις μεταξύ τους είναι πράξεις μεταξύ διανυσμάτων. Η ακριβής λειτουργία μιας μονάδας LSTM παρουσιάζεται στην συνέχεια.

- Αρχικά καθορίζεται ποιο μέρος της υπάρχουσας πληροφορίας στο cell state θα απορριφθεί μέσω της πρώτης πύλης forget gate (f_t) η οποία είναι μια πύλη με σιγμοειδή συνάρτηση ενεργοποίησης και λαμβάνει υπόψιν την είσοδο τη δεδομένη χρονική στιγμή x_t και την έξοδο της προηγούμενης μονάδας h_t

- Στη συνέχεια καθορίζεται η πληροφορία που θα προστεθεί στο cell state, διαδικασία που περιλαμβάνει δύο μέρη. Αρχικά μέσω της σιγμοειδούς input gate (i_t) καθορίζονται ποιες τιμές του cell state θα ανανεωθούν και σε τι βαθμό. Στη συνέχεια παράγονται οι νέες υποψήφιες τιμές για προσθήκη στο cell state $\hat{C}_t$ οι οποίες πολλαπλασιάζονται με την έξοδο του input gate και προστίθενται εν τέλη στο cell state.

- Τέλος, υπολογίζεται η έξοδος της μονάδας (hidden state) η οποία αποτελεί ένα τμήμα του cell state.

Οι εξισώσεις που περιγράφουν όλα τα στάδια που αναφέρθηκαν σχετικά με τη λειτουργία μίας μονάδας LSTM παρατίθενται στη συνέχεια.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\hat{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

$$C_t = C_{t-1} * f_t + i_t * \hat{C}_t \quad (4)$$

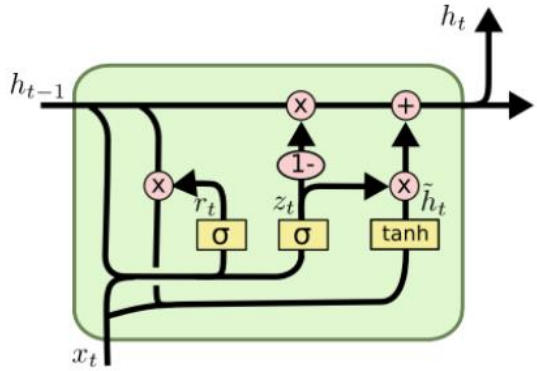
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * tanh(C_t) \quad (6)$$

Εκτός από την κλασική μορφή του LSTM που περιγράφηκε έχει αναπτυχθεί ένα μεγάλο πλήθος διαφοροποιήσεων του οι οποίες βασίζονται κυρίως σε μικρές παραλλαγές στις πύλες του LSTM. Μία από αυτές που διαφοροποιούνται στον μεγαλύτερο βαθμό και παρουσιάζουν υψηλές αποδόσεις είναι τα δίκτυα GRU που θα μελετηθούν στη συνέχεια.

2.3.3 Gated Recurrent Unit (GRU)

Τα δίκτυα GRU προτάθηκαν για πρώτη φορά από τους Cho κ.ά. [15] και αποτελούν μια απλοποιημένη εκδοχή των LSTM καθώς περιέχει μικρότερο αριθμό πυλών. Επιπλέον δεν υπάρχει η έννοια του cell state καθώς έχει ενσωματωθεί στο hidden state και η ροή της πληροφορίας ελέγχεται από τις πύλες reset και update οι οποίες καθορίζουν τι πληροφορία πρέπει να ξεχαστεί και τι πληροφορία πρέπει να περάσει στην επόμενη μονάδα αντίστοιχα. Η δομή της μονάδας GRU και οι εξισώσεις που καθορίζουν τη λειτουργία της παρουσιάζονται στην Εικόνα 7.



$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (7)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (8)$$

$$\hat{h}_t = tanh(W \cdot [r_t * h_{t-1}, x_t]) \quad (9)$$

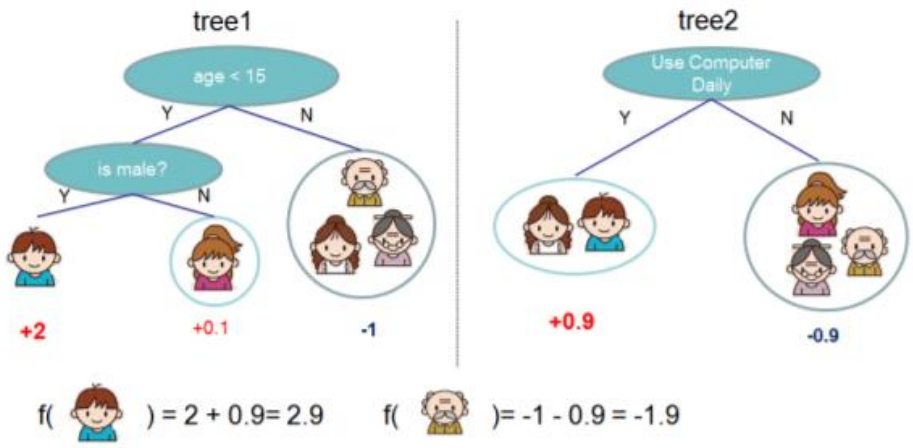
$$h_t = (1 - z_t) * h_{t-1} + z_t * \hat{h}_t \quad (10)$$

Εικόνα 7. Δομή και εξισώσεις μονάδας GRU

Λόγω της απλούστερης μορφής του το GRU έχει μικρότερο αριθμό παραμέτρων σε σχέση με το LSTM με αποτέλεσμα να έχει γενικά μικρότερο χρόνο εκπαίδευσης. Έχει πραγματοποιηθεί ένα μεγάλο πλήθος ερευνών που συγκρίνουν την απόδοση αυτών των δικτύων όπως αυτή των Karpathy κ.ά. [16] με τα αποτελέσματα να δείχνουν ότι καμία μορφή δεν είναι ξεκάθαρα καλύτερη από την άλλη με τα δίκτυα να παρουσιάζουν πολύ κοντινά αποτελέσματα.

2.3.4 XGBoost

Το XGBoost αποτελεί μία υλοποίηση των αλγορίθμων ενίσχυσης κλίσης (Gradient Boosting algorithms) [17] που προτάθηκε για πρώτη φορά από τους Chen κ.ά. [18] και στοχεύει σε μειωμένη ταχύτητα εκπαίδευσης και αύξηση της απόδοσης των παραπάνω αλγορίθμων. Ανήκει στην ευρύτερη κατηγορία των αλγορίθμων που βασίζονται στη μέθοδο Ensemble learning και τα τελευταία χρόνια έχει αποτελέσει την state-of-the-art μέθοδο στην επίλυση μεγάλου πλήθους classification και regression προβλημάτων. Όπως όλοι οι αλγόριθμοι ενίσχυσης κλίσης έτσι και ο XGBoost βασίζει την λειτουργία του στην δημιουργία απλοϊκών μοντέλων (“weak learners”), που στην συγκεκριμένη περίπτωση είναι τα δέντρα αποφάσεων. Αρχικά ξεκινάει από ένα δέντρο το οποίο παράγει τις αρχικές προβλέψεις και σταδιακά προσπαθεί να μειώσει το σφάλμα πρόβλεψης δημιουργώντας ένα ακόμα δέντρο τη φορά και αθροίζοντας της επιμέρους προβλέψεις, μέχρι έναν καθορισμένο αριθμό δέντρων ή έως ότου φτάσει σε ένα επιθυμητό επίπεδο σφάλματος. Το τελικό μοντέλο αποτελεί έναν συνδυασμό των δέντρων που κατασκευάστηκαν και οι προβλέψεις του είναι το άθροισμα όλων των επιμέρους προβλέψεων. Στην Εικόνα 8. που προέρχεται από την εργασία των Chen κ.ά. φαίνεται ένα απλό παράδειγμα συνδυασμού δύο δέντρων αποφάσεων.



Εικόνα 8. Συνδυασμός προβλέψεων των επιμέρους δέντρων

2.4 Μετρικές απόδοσης

Η αξιολόγηση των μοντέλων που αναπτύχθηκαν πραγματοποιήθηκε με τη χρήση ενός συνόλου μετρικών απόδοσης ώστε να υπάρχει πιο ολοκληρωμένη εικόνα για την απόδοση κάθε μοντέλου. Από τη στιγμή που οι προβλέψεις πραγματοποιήθηκαν και σε επίπεδα πρόβλεψης της πραγματικής τιμής αλλά και της πορείας της λήφθηκαν υπόψιν μετρικές και για τους δύο τύπους προβλημάτων.

2.4.1 Μετρικές regression

Μέσο τετραγωνικό σφάλμα - Mean Squared Error (MSE)

Εκφράζει τη μέση τετραγωνική απόκλιση των προβλέψεων του μοντέλου από τις πραγματικές τιμές. Εξαιτίας του γεγονότος ότι τα σφάλματα τετραγωνίζονται ο δείκτης αυτός επηρεάζεται σημαντικά από outliers και δεν πρέπει να επιλέγεται σε περιπτώσεις όπου το σύνολο δεδομένων χαρακτηρίζεται από πολλές τέτοιες τιμές. Υπολογίζεται από τον τύπο:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (11)$$

Μέσο τετραγωνικό ριζικό σφάλμα - Root Mean Squared Error (RMSE)

Αποτελεί μία από τις πιο συχνά χρησιμοποιούμενες μετρικές σε regression προβλήματα και όπως και το MSE εξαρτάται σημαντικά από outlier τιμές, ωστόσο σε μικρότερο βαθμό λόγω της ρίζας. Υπολογίζεται από τον τύπο:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (12)$$

Μέσο απόλυτο ποσοστιαίο σφάλμα - Mean Absolute Percentage Error (MAPE)

Οι παραπάνω μετρικές αποτελούν πολύ καλούς δείκτες απόδοσης ενός μοντέλου για μία συγκεκριμένη μεταβλητή ωστόσο δεν μπορούν να χρησιμοποιηθούν για τη σύγκριση των προβλέψεων του μοντέλου για διαφορετικές μεταβλητές καθώς εξαρτώνται από το εύρος τιμών των μεταβλητών. Σε αυτή την περίπτωση μπορεί να χρησιμοποιηθεί η μετρική MAPE η οποία εκφράζει την μέση ποσοστιαία απόκλιση των προβλέψεων από τις πραγματικές τιμές. Με τη μετρική MAPE εκτός από

συγκρίσεις μεταξύ προβλέψεων για διαφορετικές μεταβλητές παρέχεται και μια πιο διαισθητική εικόνα του εύρους του σφάλματος καθώς δίνεται σε ποσοστό επί της πραγματικής τιμής. Υπολογίζεται από τον τύπο:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i - \hat{y}_i}{y_i} \quad (13)$$

2.4.2 Μετρικές classification

Accuracy

Αποτελεί την πιο συχνά χρησιμοποιούμενη μετρική απόδοσης σε προβλήματα classification καθώς παρέχει τα πιο διαισθητικά αποτελέσματα. Εκφράζει το ποσοστό των προβλέψεων που πραγματοποιήθηκαν επιτυχημένα. Ωστόσο είναι καλύτερο να μην χρησιμοποιείται ως η αποκλειστική μέθοδος αξιολόγησης της απόδοσης ενός μοντέλου αλλά σε συνδυασμό με παραμέτρους όπως το Precision και το Recall που θα αναλυθούν στη συνέχεια για την απόκτηση μιας πιο σφαιρικής εικόνας σχετικά με την αποτελεσματικότητα του μοντέλου.

Precision

Πριν την εξήγηση αυτής της μετρικής θα γίνει μία αναφορά στους όρους που χρησιμοποιούνται για τον χαρακτηρισμό μιας πρόβλεψης σε ένα binary classification πρόβλημα όπου η μεταβλητή στόχος ανήκει σε δύο κλάσεις Positive (P) και Negative (N).

- True Positive (T_P): Πλήθος σημείων που κατατάχθηκαν σωστά στην κλάση P.
- True Negative (T_N): Πλήθος σημείων που κατατάχθηκαν σωστά στην κλάση N.
- False Positive (F_P): Πλήθος σημείων που κατατάχθηκαν λάθος στην κλάση P.
- False Negative (F_N): Πλήθος σημείων που κατατάχθηκαν λάθος στην κλάση N.

Η μετρική Precision εκφράζει το ποσοστό των επιτυχημένων προβλέψεων στην κλάση P και υπολογίζεται από τον τύπο:

$$Precision = \frac{T_P}{T_P + F_P} \quad (14)$$

Recall

Εκφράζει το ποσοστό των σημείων που ανήκουν στην κλάση P και κατατάχθηκαν επιτυχημένα και υπολογίζεται από τον τύπο:

$$Recall = \frac{T_P}{T_P + F_N} \quad (15)$$

F1 – score

Η μετρική αυτή χρησιμοποιείται πολύ συχνά σε προβλήματα binary classification και αποτελεί τον αρμονικό μέσο των μετρικών Precision και Recall. Ο αρμονικός μέσος χρησιμοποιείται έναντι του απλού αριθμητικού μέσου για να δίνεται μεγαλύτερη βαρύτητα στην χειρότερη εκ των δύο μετρικών. Με άλλα λόγια για να είναι υψηλός ο δείκτης F1 πρέπει και το Precision και το Recall να είναι υψηλά. Υπολογίζεται από τον τύπο:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (16)$$

Κεφάλαιο 3

Συλλογή δεδομένων και επιλογή κατάλληλων χαρακτηριστικών

Σε αυτό το κεφάλαιο θα ασχοληθούμε με την δημιουργία του αρχικού συνόλου δεδομένων περιγράφοντας τα διαφορετικά δεδομένα και τους τρόπους με τους οποίους αυτά συλλέχθηκαν καθώς και με την επιλογή των κατάλληλων χαρακτηριστικών (Feature selection) που θα χρησιμοποιηθούν εν τέλη από τους αλγορίθμους μηχανικής μάθησης για την πρόβλεψη της τιμής του Ether.

3.1 Συλλογή δεδομένων

Αρχικά έγινε η συλλογή των δεδομένων της τιμής του Ether, τα οποία παρέχονται δωρεάν από ένα μεγάλο πλήθος πηγών στο διαδίκτυο. Στη συγκεκριμένη εργασία χρησιμοποιήθηκαν τα ιστορικά ημερήσια δεδομένα της τιμής του Ether από τις 16/9/2018 έως τις 16/4/2020 τα οποία αντλήθηκαν από το cryptodatadownload.com[3].

Όπως αναφέρθηκε και στην εισαγωγή ένας από τους βασικούς στόχους της εργασίας είναι η εύρεση των χαρακτηριστικών που μπορεί να έχουν σημαντική επιρροή στις μεταβολές τις αξίας του Ether. Επιλέχθηκαν συνολικά 13 χαρακτηριστικά από ένα ευρύ φάσμα πεδίων, τα οποία μπορούν να διακριθούν στις εξής κατηγορίες:

- Αγορά και χρηματιστήριο
- Κοινωνική δημοφιλία
- Τεχνικοί δείκτες
- Χαρακτηριστικά δικτύου

[3] Διαδικτυακή πλατφόρμα που παρέχει ιστορικά οικονομικά δεδομένα

Στον Πίνακα 1. παρουσιάζονται μια συγκεντρωτική εικόνα των δεδομένων που συλλέχθηκαν.

Γνώρισμα	Περιγραφή	Είδος
Volume_ETH	Συνολικό ποσό που ανταλλάχθηκε για αγορά Ether σε δολάρια	Αγορά και χρηματιστήριο
Volume_USD	Συνολικό ποσό που ανταλλάχθηκε στο χρηματιστήριο σε δολάρια	Αγορά και χρηματιστήριο
Tx_per_day	Αριθμός συναλλαγών ανά ημέρα στο Ethereum blockchain	Αγορά και χρηματιστήριο
Amount_per_day_Wei	Ποσό που δαπανήθηκε σε συναλλαγές στο Ethereum blockchain σε Wei	Αγορά και χρηματιστήριο
BTC_Open	Τιμή Bitcoin	Αγορά και χρηματιστήριο
Google_trends	Δημοτικότητα αναζήτησης λέξης Ethereum στο Google (κλίμακα 1-100)	Κοινωνική δημοφιλία
Google_trends_Coinbase	Δημοτικότητα αναζήτησης λέξης Coinbase στο Google (κλίμακα 1-100)	Κοινωνική δημοφιλία
Google_trends_Exodus	Δημοτικότητα αναζήτησης λέξης Exodus στο Google (κλίμακα 1-100)	Κοινωνική δημοφιλία
14ma	Κυλιόμενος μέσος όρος 14^{ων} ημερών της τιμής του Ether	Τεχνικοί δείκτες
MACD	Τιμή οικονομικού δείκτη MACD	Τεχνικοί δείκτες
14ema	Κυλιόμενος εκθετικός μέσος όρος 14^{ων} ημερών της τιμής του Ether	Τεχνικοί δείκτες
block_size	Μέση τιμή μεγέθους block ανά ημέρα σε MB	Χαρακτηριστικά δικτύου
mining_difficulty	Μέση τιμή mining difficulty ανά ημέρα	Χαρακτηριστικά δικτύου

Πίνακας 1. Χαρακτηριστικά ανά ημέρα

3.1.1 Δεδομένα αγοράς και χρηματιστηρίου

Η τιμή του Ether, όπως και η τιμή κάθε κρυπτονομίσματος και μετοχής, είναι πιθανό να επηρεάζεται από την κατάσταση στις διεθνείς αγορές και στο χρηματιστήριο. Δύο χαρακτηριστικά ικανά να αποτυπώσουν την κατάσταση στις αγορές είναι τα συνολικά ποσά που διακινούνται καθημερινά στο χρηματιστήριο αλλά και πιο συγκεκριμένα τα ποσά που ξοδεύονται για αγορές Ether [8]. Επιπλέον χρησιμοποιήθηκε η ημερήσια τιμή του Bitcoin, το οποίο ως το πρώτο και πιο ευρέως χρησιμοποιούμενο κρυπτονόμισμα μπορεί να έχει τη δυνατότητα να επηρεάσει σε μεγάλο βαθμό την αξία του Ether. Τα τρία ανωτέρω χαρακτηριστικά συλλέχθηκαν παρόμοια με την τιμή του Ether από το cryptodatadownload.com.

Δύο επιπλέον παράγοντες που ανήκουν σε αυτή την κατηγορία και κρίθηκαν σημαντικοί είναι ο αριθμός των ημερήσιων συναλλαγών στο Ethereum δίκτυο καθώς και το ποσό που αντιστοιχεί σε αυτές τις συναλλαγές. Τα δύο αυτά χαρακτηριστικά μπορούν να αναδείξουν τον βαθμό εμπιστοσύνης των χρηστών στο δίκτυο γεγονός

που τα καθιστά χρήσιμα στην επίλυση του προβλήματος που αντιμετωπίζουμε. Η απόκτηση τους πραγματοποιήθηκε με τη δημιουργία κατάλληλων SQL ερωτημάτων, όπως φαίνεται στις Εικόνες 9 και 10, στη βάση δεδομένων Google BigQuery στην οποία παρέχονται δωρεάν όλα τα δεδομένα του Ethereum blockchain.

```
select Date, sum(Tx) as Tx_per_day
from (select DATE(block_timestamp) as Date, count(block_timestamp) as Tx
      from `bigquery-public-data.crypto_ethereum.transactions`
      where DATE(block_timestamp) between "2018-9-16" and "2020-4-16"
      group by block_timestamp
      order by block_timestamp)
group by Date
order by Date
```

Εικόνα 9. SQL κώδικας για αριθμό συναλλαγών ανά ημέρα

```
select Date, sum(Amount) as Amount_per_day
from (select DATE(block_timestamp) as Date, sum(value) as Amount
      from `bigquery-public-data.crypto_ethereum.transactions`
      where DATE(block_timestamp) between "2016-9-16" and "2020-4-16"
      group by block_timestamp
      order by block_timestamp)
group by Date
order by Date
```

Εικόνα 10. SQL κώδικας συνολικό ποσό συναλλαγών ανά ημέρα

3.1.2 Δεδομένα σχετικά με κοινωνική δημοφιλία

Η χρήση δεδομένων σχετικών με τη δημοφιλία των κρυπτονομισμάτων στα κοινωνικά δίκτυα και σε μηχανές αναζήτησης έχει αποδειχθεί ότι μπορεί να συνεισφέρει θετικά στην πρόβλεψη της τιμής τους [7] [19] καθώς ανάλογα με το ενδιαφέρον του κόσμου για το συγκεκριμένο κρυπτονόμισμα μπορεί να διακυμαίνεται και η αξία του. Για τον λόγο αυτό το σύνολο των δεδομένων εμπλουτίστηκε με τον δείκτη δημοτικότητας από το Google trends[4] για τους όρους Ethereum, Coinbase[5] και Exodus[3], ο οποίος είναι ένας ακέραιος στην κλίμακα 1-100. Αξίζει να σημειωθεί ότι η επίδραση της δημοφιλίας πορτοφολιών κρυπτονομισμάτων στην πρόβλεψη της τιμής τους δεν έχει εξερευνηθεί μέχρι στιγμής και εξετάζεται για πρώτη φορά στην παρούσα εργασία.

Ένα πρόβλημα που αντιμετωπίστηκε κατά τη συλλογή των δεδομένων ήταν ότι το Google trends παρέχει τα δεδομένα σε εβδομαδιαία και όχι ημερήσια διαστήματα. Η

[4] https://trends.google.com/trends/?geo=US
[5] Δύο από τα πιο ευρέως χρησιμοποιούμενα πορτοφόλια κρυπτονομισμάτων

επίλυση του ανωτέρω προβλήματος έγινε με την παραδοχή ότι όλες οι μέρες μιας εβδομάδας έχουν το δείκτη δημοτικότητας της εβδομάδας που ανήκουν.

3.1.3 Δεδομένα του Ethereum δικτύου

Παρόμοια με τα δεδομένα των ημερήσιων συναλλαγών και των αντίστοιχων χρηματικών ποσών τους αντλήθηκαν μέσω του Google BigQuery το μέσο μέγεθος block αλλά και η μέση δυσκολία εξόρυξης, χαρακτηριστικά που αποτελούν δύο από τα σημαντικότερα του Ethereum blockchain. Στις παρακάτω εικόνες φαίνονται τα SQL ερωτήματα για την απόκτηση των ανωτέρω χαρακτηριστικών.

```
select Date, avg(block_size) as block_size
from (select Date(timestamp) as Date, size as block_size
      from `bigquery-public-data.crypto_ethereum.blocks`
      where DATE(timestamp) between "2016-9-16" and "2020-4-16"
      order by timestamp)
group by Date
order by Date
```

Εικόνα 11. SQL κώδικας για μέσο μέγεθος block ανά ημέρα

```
select Date, avg(mining_difficulty) as mining_difficulty
from (select Date(timestamp) as Date, difficulty as mining_difficulty
      from `bigquery-public-data.crypto_ethereum.blocks`
      where DATE(timestamp) between "2016-9-16" and "2020-4-16"
      order by timestamp)
group by Date
order by Date
```

Εικόνα 12. SQL κώδικας για μέση δυσκολία εξόρυξης ανά ημέρα

3.1.4 Δεδομένα σχετικά με τεχνικούς δείκτες

Τέλος, στο σύνολο δεδομένων ενσωματώθηκαν οι τεχνικοί δείκτες 14ma, 14ema και MACD οι οποίοι έχουν χρησιμοποιηθεί σε σημαντικό βαθμό σε έρευνες σχετικές με την πρόβλεψη της τιμής κρυπτονομισμάτων και μετοχών [20] [21] και κατασκευάστηκαν αποκλειστικά από τα ιστορικά δεδομένα της τιμής του Ether. Η επιλογή του πλαισίου των 14 ημερών έγινε καθώς αποτελεί ένα από τα πιο συνηθισμένα χρονικά πλαίσια σε συναφείς μελέτες και επειδή ενσωματώνει τόσο βραχυχρόνια όσο και πιο μακροχρόνια χαρακτηριστικά της τιμής του Ether.

3.2 Επιλογή χαρακτηριστικών (Feature selection)

Η επιλογή χαρακτηριστικών αποτελεί ένα από τα σημαντικότερα στάδια της προεπεξεργασίας των δεδομένων καθώς μπορεί να προσφέρει τα εξής σημαντικά οφέλη κατά την ανάπτυξη του μοντέλου μηχανικής μάθησης:

- Αύξηση της απόδοσης του μοντέλου λόγω μείωσης της πολυπλοκότητας του
- Μείωση του χρόνου εκπαίδευσης του μοντέλου
- Μείωση του overfitting[6]

Στα πλαίσια της παρούσας εργασίας αναπτύχθηκαν διάφορες τεχνικές επιλογής χαρακτηριστικών με στόχο την εύρεση του βέλτιστου συνόλου δεδομένων. Η πρώτη αφορά στην εξάλειψη αχρείαστων χαρακτηριστικών ενώ οι επόμενες θα αναδείξουν αυτά που είναι πιθανό να επηρεάζουν σε σημαντικό βαθμό την τελική πρόβλεψη. Τα χαρακτηριστικά αυτά τέλος θα ελεγχθούν και σε κανονικές συνθήκες πρόβλεψης για να διαπιστωθεί αν όντως έχουν θετική συνεισφορά καθώς οι συντελεστές σημαντικότητας εκφράζουν ένα σχετικό μέτρο επιρροής πάνω στην τελική πρόβλεψη.

3.2.1 Έλεγχος γραμμικής συσχέτισης

Όπως είναι γνωστό δύο μεταβλητές x, y είναι γραμμικά συσχετιζόμενες αν συνδέονται με μία σχέση της μορφής

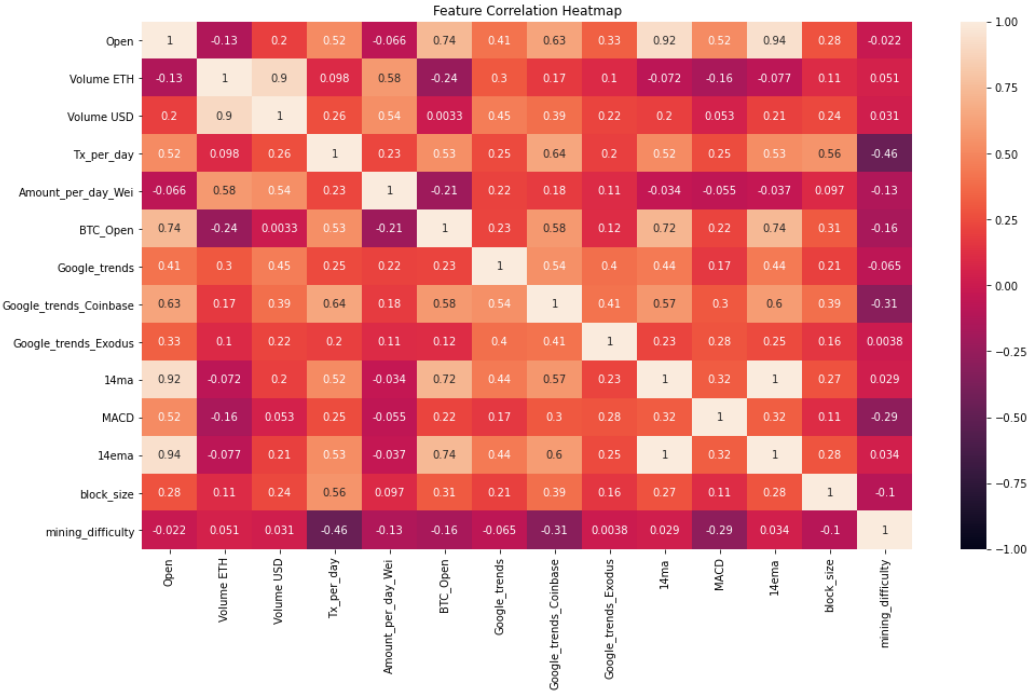
$$y = ax + b$$

όπου a,b σταθεροί συντελεστές. Η ύπαρξη γραμμικά συσχετιζόμενων γνωρισμάτων σε ένα σύνολο δεδομένων δεν συνεισφέρει θετικά στην απόδοση του αλγορίθμου οδηγώντας πολλές φορές και σε χειρότερες αποδόσεις. Για τον λόγο αυτό υπολογίστηκε ο βαθμός της συσχέτισης των χαρακτηριστικών μέσω του συντελεστή συσχέτισης Pearson[7] (Pearson correlation) και απομακρύνθηκαν τα χαρακτηριστικά που παρουσίαζαν μεταξύ τους συσχέτιση μεγαλύτερη του 0.8. Στην Εικόνα 13. παρουσιάζονται οι συσχετίσεις κατά Pearson όλων των γνωρισμάτων του συνόλου δεδομένων. Όπως φαίνεται υπάρχει πολλή υψηλή συσχέτιση μεταξύ των Volume ETH – Volume USD (0.9) και 14ma – 14ema (1) γεγονός που οδήγησε στην απομάκρυνση των Volume USD και 14ma. Τα χαρακτηριστικά Volume USD και 14ema επιλέχθηκαν αντί των άλλων δύο καθώς παρουσιάζουν υψηλότερη συσχέτιση

[6] Η προσκόλληση ενός αλγορίθμου μηχανικής μάθησης στα δεδομένα εκπαίδευσης με αποτέλεσμα να μην έχει τη δυνατότητα να γενικεύσει τις προβλέψεις του νέα δεδομένα

[7] https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

με την τιμή του Ether, η οποία είναι και το ζητούμενο της πρόβλεψης, όπως διακρίνεται από την πρώτη στήλη.



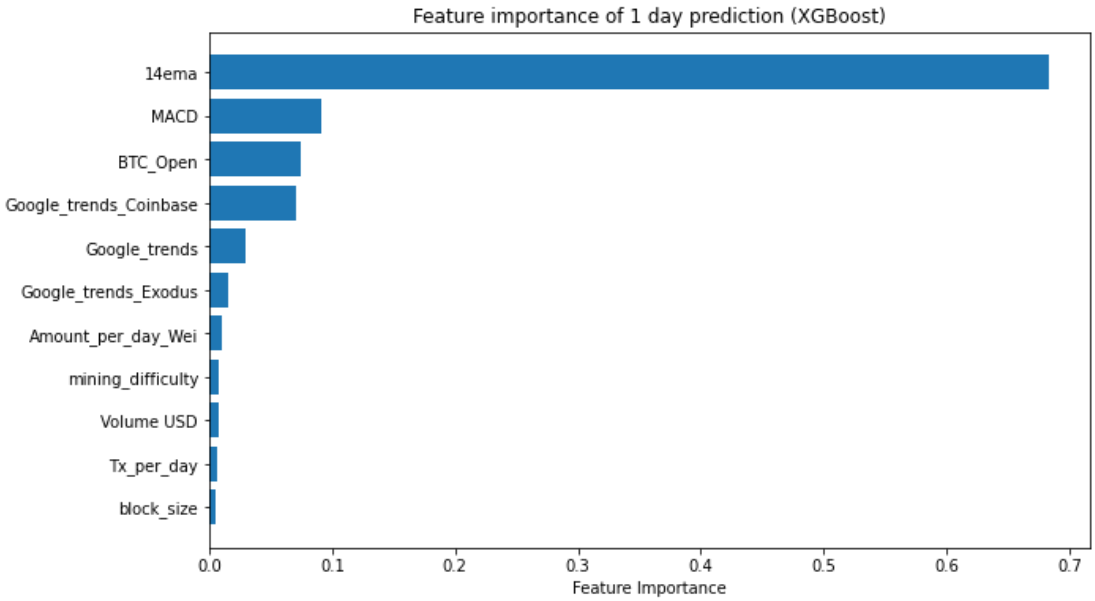
Εικόνα 13. Πίνακας συσχέτισης Pearson

3.2.2 Σημαντικότητα χαρακτηριστικών (Feature importance)

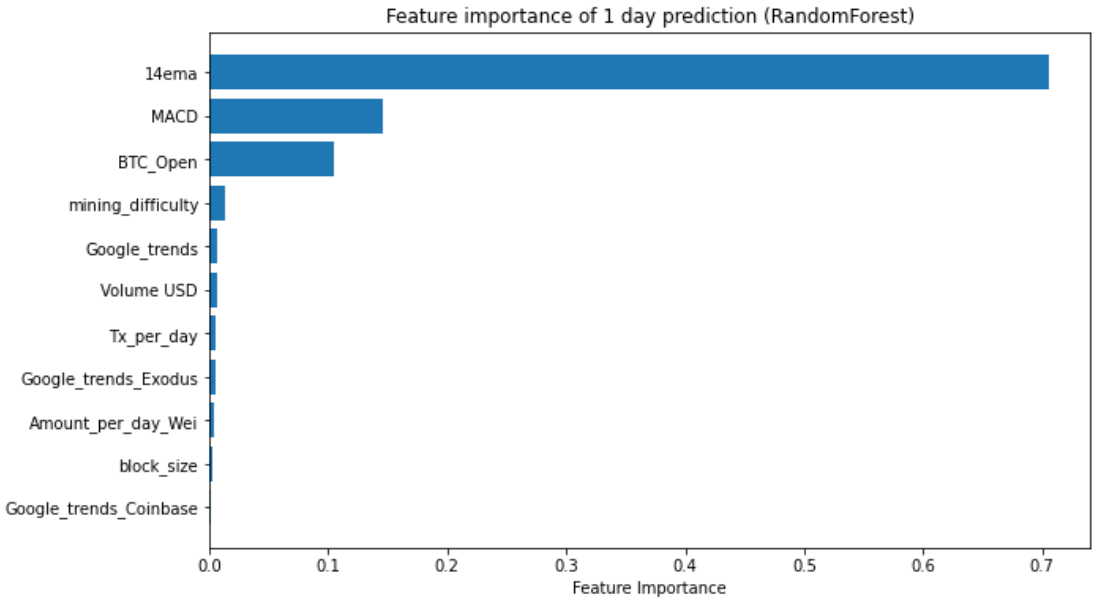
Οι αλγόριθμοι που βασίζονται στην κατασκευή δέντρων αποφάσεων παρέχουν την δυνατότητα υπολογισμού ενός συντελεστή σημαντικότητας για κάθε χαρακτηριστικό του συνόλου δεδομένων στο οποίο εκπαιδεύτηκαν. Ο συντελεστής αυτός σχετίζεται με το πόσο σημαντικό ήταν το κάθε χαρακτηριστικό κατά τον σχηματισμό των δέντρων αποφάσεων που χρησιμοποιούν οι ανωτέρω αλγόριθμοι και μπορεί να παρέχει σημαντικές πληροφορίες για την επιρροή που μπορεί να έχει ένα γνώρισμα στην τελική πρόβλεψη. Όσο περισσότερο χρησιμοποιήθηκε το χαρακτηριστικό σε σημαντικές αποφάσεις των αλγορίθμων τόσο υψηλότερος είναι και ο συντελεστής του.

Οι συντελεστές σημαντικότητας υπολογίστηκαν με τη χρήση των αλγορίθμων XGBoost και Random Forest καθώς αποτελούν δύο από τους πιο αποδοτικούς και ευρέως διαδεδομένους αλγορίθμους κατασκευής δέντρων αποφάσεων. Πραγματοποιήθηκαν υπολογισμοί σε επίπεδο ημερήσιας αλλά και εβδομαδιαίας πρόβλεψης της τιμής του Ether καθώς η επιρροή κάθε χαρακτηριστικού μπορεί να είναι διαφορετική στα δύο αυτά χρονικά επίπεδα και επιλέχθηκαν τα χαρακτηριστικά με συντελεστή μεγαλύτερο του 0.1 (επιλέχθηκαν και αυτά με οριακά μικρότερο του 0.1).

Πρόβλεψη μίας μέρας



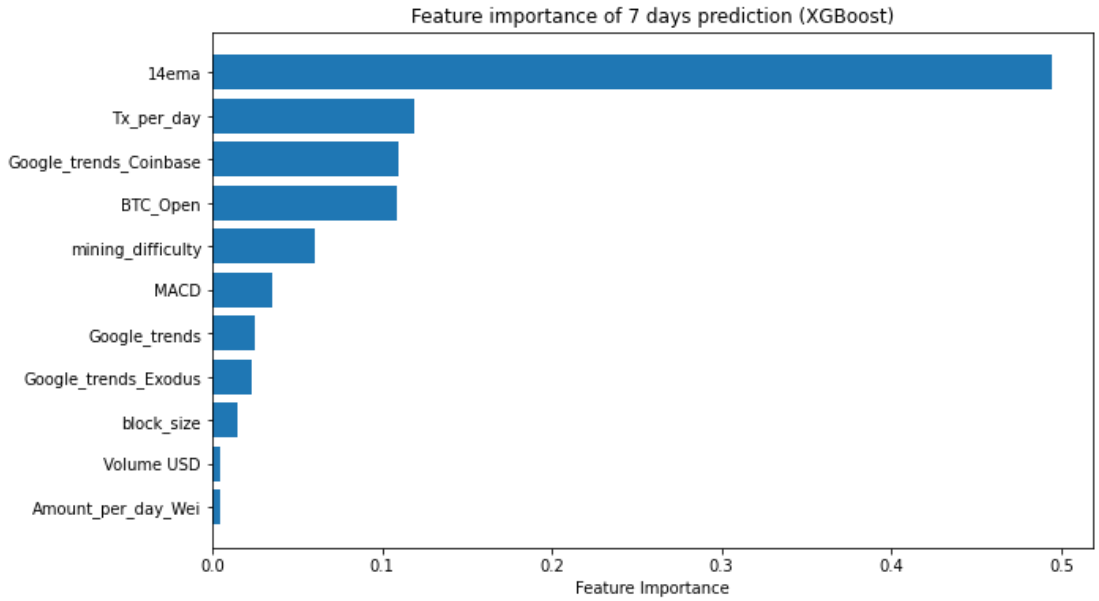
Εικόνα 14. Σημαντικότητα χαρακτηριστικών - XGBoost 1 ημέρα



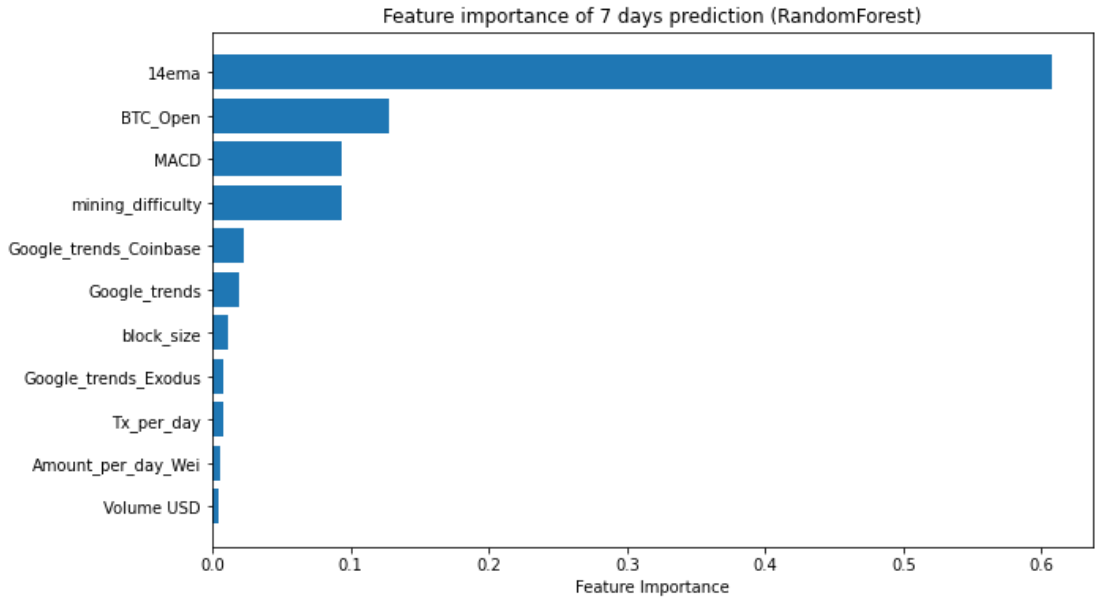
Εικόνα 15. Σημαντικότητα χαρακτηριστικών - Random Forest 1 ημέρα

Όπως φαίνεται οι μεγαλύτερες επιρροές στον αλγόριθμο συνέβησαν από τα χαρακτηριστικά 14ema και MACD και στα δύο είδη πρόβλεψης. Σημαντικά αναδείχθηκαν επίσης τα BTC_Open και Google_trends_Coinbase.

Πρόβλεψη μίας εβδομάδας



Εικόνα 16. Σημαντικότητα χαρακτηριστικών – XGBoost 7 ημέρες



Εικόνα 17. Σημαντικότητα χαρακτηριστικών – Random Forest 7 ημέρες.

Και σε αυτή την περίπτωση οι μεγαλύτερες επιρροές συνέβησαν από το χαρακτηριστικό 14ema κ και στα δύο είδη πρόβλεψης. Σημαντικά αναδείχθηκαν επίσης τα BTC_Open, mining_difficulty, Tx_per_day και MACD.

3.2.3 Αναδρομική απαλοιφή χαρακτηριστικών (Recursive feature elimination)

Επιπλέον σαν συμπληρωματική μέθοδος επιλογής των καταλληλότερων χαρακτηριστικών χρησιμοποιήθηκε η τεχνική Recursive feature elimination (Rfe) [22] κατά την οποία ένας εκτιμητής ο οποίος ξεκινά την εκπαίδευση του από όλο το σύνολο των δεδομένων απαλείφει σταδιακά χαρακτηριστικά που δεν επιδρούν σημαντικά στην πρόβλεψη μέχρι έναν προκαθορισμένο αριθμό χαρακτηριστικών. Τα βήματα της μεθόδου Rfe είναι τα εξής:

1. Εκπαίδευσε το μοντέλο στο σύνολο των χαρακτηριστικών

2. Υπολόγισε την απόδοση του και κατέταξε τα χαρακτηριστικά

3. Εκπαίδευσε ξανά το μοντέλο βγάζοντας ένα διαφορετικό χαρακτηριστικό κάθε φορά και υπολόγισε την απόδοση του μοντέλου

4. Αφαίρεσε το χαρακτηριστικό η αφαίρεση του οποίου είχε τη μικρότερη επίδραση στην απόδοση

5. Επανέλαβε από το βήμα 3 έως τον απαιτούμενο αριθμό χαρακτηριστικών

Για την υλοποίηση του αλγορίθμου rfe χρησιμοποιήθηκε η κλάση RFE της βιβλιοθήκης sklearn.feature_selection και σαν εκτιμητής χρησιμοποιήθηκε ο αλγόριθμος Decision tree. Ο αριθμός των τελικών χαρακτηριστικών του rfe επιλέχθηκε να είναι 4 για την ημερήσια και 6 για την εβδομαδιαία πρόβλεψη σύμφωνα με τα αποτελέσματα που προέκυψαν και από την προηγούμενη ενότητα.

Τα χαρακτηριστικά που αναδείχθηκαν σημαντικότερα, σύμφωνα με τη μέθοδο rfe, για την ημερήσια πρόβλεψη ήταν τα BTC_Open, Google_trends, 14ema και MACD ενώ για την εβδομαδιαία τα Tx_per_day, BTC_Open, Google_trends, Google_trends_Coinbase, 14ema και MACD.

3.2.4 Σύνολο δεδομένων έπειτα από επιλογή χαρακτηριστικών

Για την τελική επιλογή χαρακτηριστικών λήφθηκαν υπόψιν αυτά που κρίθηκαν σημαντικά έστω και σε μία από τις μεθόδους που αναπτύχθηκαν στις ενότητες 3.2.2 και 3.2.3 για να μην παραλειφθεί κάποια σημαντική πληροφορία. Έτσι καταλήξαμε στα εξής σύνολα δεδομένων:

Σύνολο χαρακτηριστικών ημερήσιας πρόβλεψης

- BTC_Open
- Google_trends
- Google_trends_Coinbase
- MACD
- 14ema

Date	Open	BTC_Open	Google_trends	Google_trends_Coinbase	MACD	14ema
9/16/2018	222.07	6530.08	11	6	-29.603667	223.965662
9/17/2018	220.37	6498.37	11	6	-27.370100	223.486241
9/18/2018	196.50	6251.56	11	6	-27.212403	219.888075
9/19/2018	208.35	6339.92	11	6	-25.833440	218.349665
9/20/2018	209.71	6383.42	11	6	-24.350167	217.197710

Εικόνα 18. Χαρακτηριστικά ημερήσιας πρόβλεψης

Σύνολο χαρακτηριστικών εβδομαδιαίας πρόβλεψης

- BTC_Open
- Google_trends
- Google_trends_Coinbase
- MACD
- 14ema
- Tx_per_day
- mining_difficulty

Date	Open	Tx_per_day	BTC_Open	Google_trends	Google_trends_Coinbase	MACD	14ema	mining_difficulty
9/16/2018	222.07	458432	6530.08	11	6	-29.603667	223.965662	3.129490e+15
9/17/2018	220.37	553720	6498.37	11	6	-27.370100	223.486241	3.140090e+15
9/18/2018	196.50	500513	6251.56	11	6	-27.212403	219.888075	3.178890e+15
9/19/2018	208.35	507255	6339.92	11	6	-25.833440	218.349665	3.170660e+15
9/20/2018	209.71	498146	6383.42	11	6	-24.350167	217.197710	3.176540e+15

Εικόνα 19. Χαρακτηριστικά εβδομαδιαίας πρόβλεψης

Τέλος, όπως αναφέρθηκε και παραπάνω τα χαρακτηριστικά αυτά θα εξεταστούν και πάνω στο validation set των αλγορίθμων για να εξεταστεί εάν όντως έχουν θετική συνεισφορά στις προβλέψεις.

Κεφάλαιο 4

Προεπεξεργασία δεδομένων και εφαρμογή αλγορίθμων μηχανικής μάθησης

Σε αυτό το κεφάλαιο θα ασχοληθούμε με τις διάφορες μεθόδους προεπεξεργασίας του συνόλου δεδομένων που προέκυψε από το προηγούμενο κεφάλαιο με στόχο την βελτίωση της απόδοσης των αλγορίθμων μηχανικής μάθησης που θα εφαρμοστούν. Στη συνέχεια θα αναπτυχθούν οι αλγόριθμοι LSTM, GRU και XGBoost για ημερήσιες και εβδομαδιαίες προβλέψεις και θα γίνει βελτιστοποίηση των υπερπαραμέτρων τους. Τέλος θα παρουσιαστούν τα αποτελέσματα των προβλέψεων για τους τρεις διαφορετικούς αλγορίθμους τόσο για το regression όσο και για το classification πρόβλημα.

4.1 Προεπεξεργασία δεδομένων

Ένας από τους σημαντικότερους παράγοντες που επιδρούν στην τελική απόδοση των αλγορίθμων μηχανικής μάθησης είναι η κατάλληλη προεπεξεργασία των δεδομένων που χρησιμοποιούν καθώς η απόδοση τους εξαρτάται σημαντικά από την ποιότητα των δεδομένων που χρησιμοποιούν (αναφορά). Στο προηγούμενο κεφάλαιο αναπτύχθηκαν τεχνικές προεπεξεργασίας με στόχο την απαλοιφή πλεονάζουσας ή και λανθασμένης πληροφορίας στα δεδομένα. Σε αυτό θα εστιάσουμε στον μετασχηματισμό των δεδομένων σε μορφή που θα μπορεί να αξιοποιηθεί με βέλτιστο τρόπο από τους αλγορίθμους μηχανικής μάθησης. Θα πραγματοποιηθούν τα εξής βήματα προεπεξεργασίας των δεδομένων:

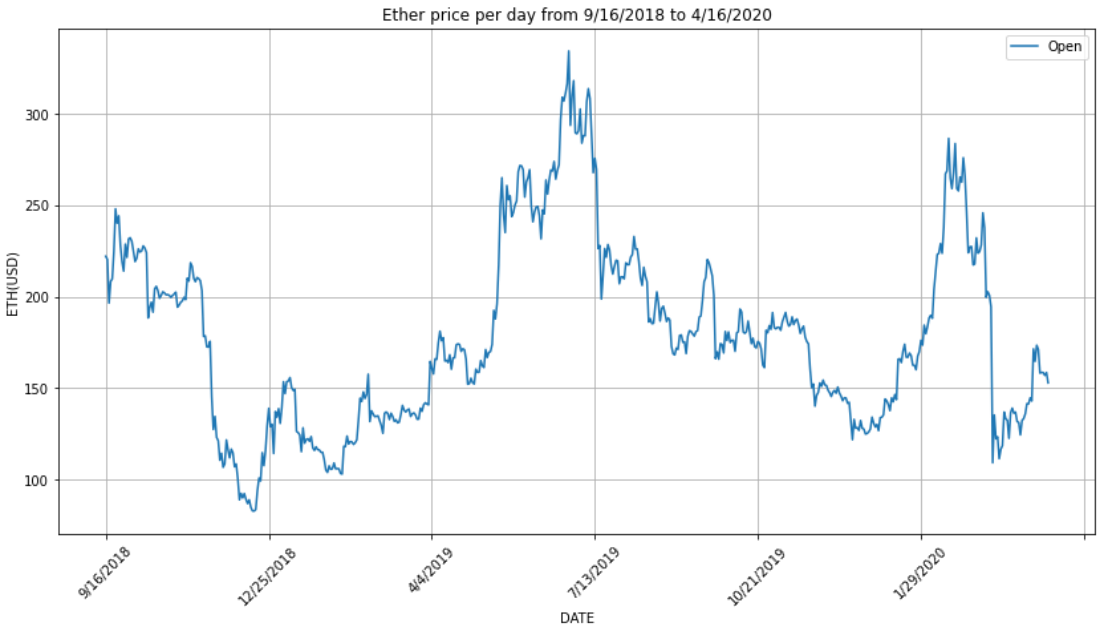
- Μείωση θορύβου
- Χωρισμός σε train, validation και test set
- Κανονικοποίηση δεδομένων
- Καθορισμός παρελθοντικού χρονικού πλαισίου που θα χρησιμοποιούν οι αλγόριθμοι
- Μετασχηματισμός δεδομένων σε συνεπή μορφή για τα LSTM, GRU και XGBoost

Τα δύο πρώτα βήματα είναι κοινά και για τους τρεις αλγορίθμους ενώ τα δύο επόμενα είναι κοινά για τα δίκτυα LSTM, GRU και διαφορετικό για τον XGBoost.

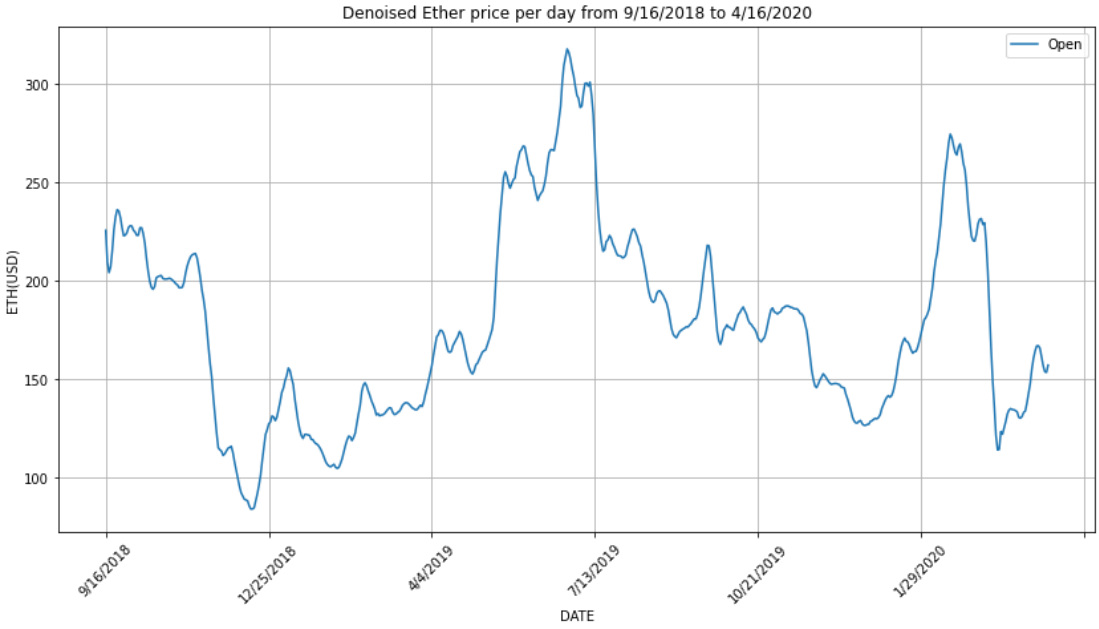
4.1.1 Μείωση θορύβου (Noise reduction)

Η μείωση του θορύβου είναι μια από τις πιο διαδεδομένες τεχνικές προεπεξεργασίας που χρησιμοποιούνται για την εξομάλυνση χρονοσειρών που χαρακτηρίζονται από υψηλά επίπεδα θορύβου [23]. Απαλείφοντας τον θόρυβο σε μία χρονοσειρά διατηρείται όλη η σημαντική πληροφορία που περιέχει ενώ παράλληλα απλουστεύεται η μορφή της γεγονός που συνεισφέρει θετικά στην πρόβλεψη του αλγορίθμου μηχανικής μάθησης καθώς μπορεί πλέον να αναγνωρίζει ευκολότερα πιθανά μοτίβα στα δεδομένα και να αντιμετωπίζει το φαινόμενο του overfitting.

Η τιμή του Ether χαρακτηρίζεται από ένα είδος θορύβου καθώς όπως σε όλα τα κρυπτονομίσματα συμβαίνουν ξαφνικές διακυμάνσεις που μπορεί να μην ακολουθούν την γενικότερη τάση της πορείας της τιμής τους. Για την εξομάλυνση των χρονοσειρών χρησιμοποιήθηκε η μέθοδος Savitzky – Golay [24] κατά την οποία η χρονοσειρά που μας ενδιαφέρει προσεγγίζεται βέλτιστα από ένα σύνολο πολυωνυμικών συναρτήσεων με τη μέθοδο των ελαχίστων τετραγώνων. Η μέθοδος Savitzky – Golay υλοποιήθηκε με τη χρήση της συνάρτησης scipy.signal.savgol_filter οι παράμετροι της οποίας ορίστηκαν με τρόπο που να μειώνεται ο θόρυβος ενώ παράλληλα διατηρείται η μορφή της χρονοσειράς. Για τον λόγο αυτό χρησιμοποιήθηκαν οι τιμές window_length=11 και polyorder=3 που αντιστοιχούν στον αριθμό των συντελεστών και στον βαθμό της πολυωνυμικής. Στις Εικόνες 20 και 21 φαίνεται η τιμή του Ether πριν και μετά την εξομάλυνση.



Εικόνα 20. Τιμή Ether



Εικόνα 21. Τιμή Ether έπειτα από εξομάλυνση

4.1.2 Χωρισμός σε train, validation και test set

Το 80% των δεδομένων (435 εγγραφές) χρησιμοποιήθηκε ως train set και τα υπόλοιπα δεδομένα χωρίστηκαν στα σύνολα validation 10% (58 εγγραφές) και test 10% (57 εγγραφές). Ένα σημείο που αξίζει να τονιστεί είναι ότι ο διαχωρισμός έγινε διατηρώντας την χρονική σειρά των δεδομένων το οποίο είναι απαραίτητο λόγω της φύσης του προβλήματος. Πιο συγκεκριμένα οι πρώτες 435 εγγραφές αποτέλεσαν το train set, οι επόμενες 58 το validation set και οι τελευταίες 57 το test set.

4.1.3 Κανονικοποίηση χαρακτηριστικών (Feature scaling)

Η κανονικοποίηση των δεδομένων είναι ένα από τα απαραίτητα στάδια κατά την προεπεξεργασία τους καθώς οδηγεί σε αυξημένη ταχύτητα σύγκλισης των αλγορίθμων. Επιπλέον σε περιπτώσεις όπου τα διαφορετικά χαρακτηριστικά κυμαίνονται σε πολύ διαφορετικά εύρη τιμών ένα μεγάλο πλήθος των αλγορίθμων μηχανικής μάθησης (κυρίως αυτοί που υπολογίζουν αποστάσεις μεταξύ των χαρακτηριστικών) δεν παράγει σωστά αποτελέσματα καθώς θεωρεί ότι οι μεταβλητές που έχουν μεγαλύτερο διάστημα τιμών έχουν και μεγαλύτερη ισχύ πάνω στην πρόβλεψη, γεγονός που είναι αναληθές,

Δύο από τις πιο γνωστές μεθόδους κανονικοποίησης αποτελούν οι Min-Max scaler και Standard scaler. Η πρώτη μέθοδος μετασχηματίζει τα δεδομένα στο

διάστημα [0, 1] ενώ η δεύτερη μετασχηματίζει την κατανομή κάθε χαρακτηριστικού με τρόπο που να έχει μηδενική μέση τιμή και διακύμανση ίση με 1.

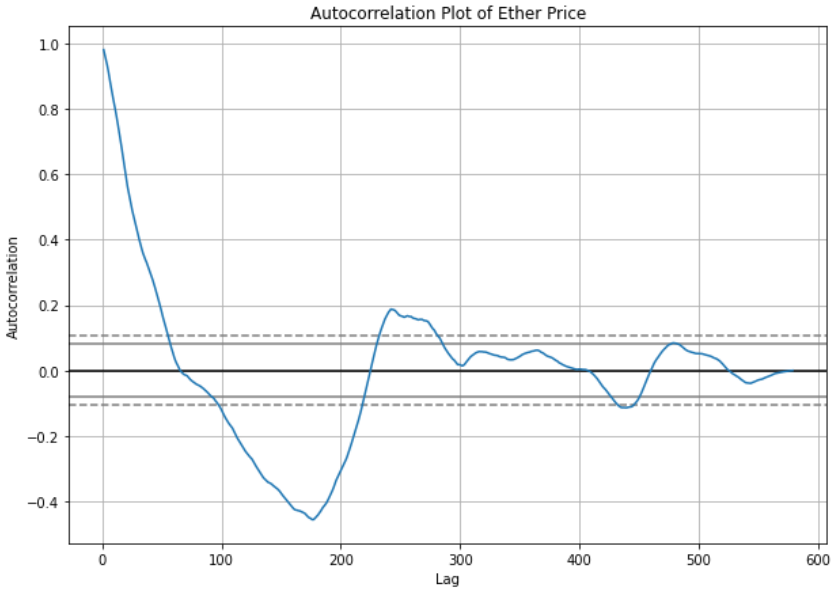
- Min-Max scaler: $x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$
- Standard scaler: $x_{new} = \frac{x - \mu}{\sigma}$

Δοκιμάστηκαν και οι δύο μέθοδοι κανονικοποίησης και παρατηρήθηκε ότι η πρώτη παρουσίαζε ελάχιστα καλύτερα αποτελέσματα γι' αυτό και επιλέχθηκε.

4.1.4 Καθορισμός παρελθοντικού χρονικού πλαισίου (timesteps)

Ένα από τα σημαντικότερα χαρακτηριστικά των επαναληπτικών δικτύων LSTM και GRU είναι η δυνατότητα τους να ανακαλύπτουν τόσο μακροχρόνιες όσο και βραχυχρόνιες εξαρτήσεις στα δεδομένα γεγονός που συμβάλει στην βελτίωση της απόδοσης των μελλοντικών τους προβλέψεων. Είναι επομένως ιδιαίτερα σημαντικός ο καθορισμός του χρονικού παραθύρου που χρησιμοποιούν τα δίκτυα αυτά για να βελτιστοποιήσουν τις προβλέψεις τους.

Για την εύρεση του κατάλληλου αριθμού των timesteps που θα χρησιμοποιηθούν μελετήθηκε η αυτοσυσχέτιση των τιμών του Ether [5], το κατά πόσο δηλαδή η τιμή σε μία δεδομένη χρονική στιγμή σχετίζεται με τιμές σε παρελθοντικές χρονικές στιγμές. Για την εύρεση της αυτοσυσχέτισης χρησιμοποιήθηκε η συνάρτηση autocorrelation_plot της βιβλιοθήκης pandas και το αποτέλεσμα παρουσιάζεται στην Εικόνα 22.



Εικόνα 22. Αυτοσυσχέτιση της τιμής του Ether

Οι στατιστικά σημαντικές παρελθοντικές τιμές είναι αυτές των οποίων ο συντελεστής αυτοσυσχέτισης βρίσκεται πάνω από την οριζόντια διακεκομμένη γραμμή. Στη συγκεκριμένη περίπτωση βλέπουμε ότι μέχρι και περίπου 60 ημέρες πριν οι τιμές του Ether είναι στατιστικά σημαντικές. Για τον σκοπό αυτό δοκιμάστηκαν οι τιμές 30, 45, 60 ως timesteps, οι οποίες παρουσίασαν πολύ κοντινά αποτελέσματα με την τιμή 30 να έχει λίγο καλύτερα. Για τον λόγο αυτό η τιμή timesteps τέθηκε στο 30.

Σε αντίθεση με τα δίκτυα LSTM και GRU ο αλγόριθμος XGBoost δεν χρησιμοποιεί κάποιο χρονικό παράθυρο στις προβλέψεις του. Ωστόσο μπορεί να ενσωματωθεί παρελθοντική πληροφορία μέσω της χρήσης χαρακτηριστικών καθυστέρησης (lag features). Ο αριθμός lag features που οδηγούσε στις καλύτερες προβλέψεις ήταν 15, δηλαδή ο αλγόριθμος είχε καλύτερα αποτελέσματα χρησιμοποιώντας πληροφορία για τις προηγούμενες 15 ημέρες.

4.1.5 Μετασχηματισμός των δεδομένων στην τελική τους μορφή

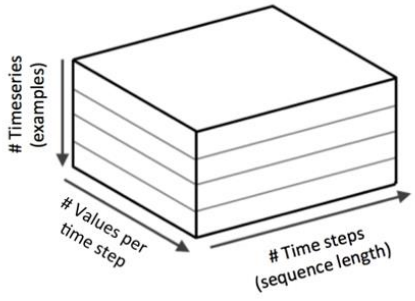
Αφού καθορίστηκαν και τα απαραίτητα παρελθοντικά χρονικά παράθυρα τα δεδομένα μετασχηματίστηκαν στην τελική τους μορφή ώστε να συμβαδίζουν με τα πρότυπα των αλγορίθμων.

LSTM και GRU

Τα δίκτυα αυτά απαιτούν τα δεδομένα να είναι στην τρισδιάστατη μορφή της Εικόνας 23. Έτσι καταλήξαμε στα εξής σύνολα δεδομένων για την ημερήσια πρόβλεψη:

- X_train (435, 30, 6) y_train (435, 1)
- X_validation (58, 30, 6) y_validation (58, 1)
- X_test (57, 30, 6) y_test (57, 1)

Τα σύνολα για την εβδομαδιαία πρόβλεψη έχουν αντίστοιχες διαστάσεις με τη διαφορά ότι έχουν 2 παραπάνω χαρακτηριστικά.



Εικόνα 23. Μορφή δεδομένων εισόδου LSTM και GRU

XGBoost

Ο αλγόριθμος αυτό χρησιμοποιεί την κλασική μορφή δεδομένων που χρησιμοποιείται από την πλειοψηφία των αλγορίθμων μηχανικής μάθησης, ότι δηλαδή κάθε γραμμή αντιστοιχεί σε μία παρατήρηση και κάθε στήλη σε ένα χαρακτηριστικό. Έτσι θα είχαμε πχ για X_train διαστάσεις (435, 6). Ωστόσο για την βελτίωση των προβλέψεων του αλγορίθμου το σύνολο των χαρακτηριστικών ενισχύθηκε και με παρελθοντικές τιμές (τις 15 τελευταίες) και χρησιμοποιήθηκαν συνολικά 6 + 15*6 = 96 χαρακτηριστικά. Καταλήξαμε λοιπόν στα εξής σύνολα δεδομένων (για ημερήσιες προβλέψεις).

- X_train (435, 96) y_train (435, 1)
- X_validation (58, 96) y_validation (58, 1)
- X_test (57, 96) y_test (57, 1)

Τα σύνολα για την εβδομαδιαία πρόβλεψη έχουν αντίστοιχες διαστάσεις με τη διαφορά ότι έχουν 32 παραπάνω χαρακτηριστικά.

4.2 Εφαρμογή αλγορίθμων μηχανική μάθησης

Στην ενότητα αυτή θα παρουσιαστεί η ανάπτυξη των αλγορίθμων LSTM, GRU και XGBoost η οποία έγινε με τη χρήση των βιβλιοθηκών tensorflow.keras και xgboost αντίστοιχα καθώς και τα αποτελέσματα των προβλέψεων τους για το regression και το classification πρόβλημα. Παράλληλα θα παρουσιαστούν και τα αποτελέσματα ενός μοντέλου ARIMA που αναπτύχθηκε για να γίνει και μία σύγκριση με τα κλασικά στατιστικά μοντέλα που χρησιμοποιούνται συχνά για την πρόβλεψη χρονοσειρών [3].

4.2.1 Ημερήσιες προβλέψεις

Η πρόβλεψη της τιμής της επόμενης ημέρας για μια χρονοσειρά είναι ένα ιδιαίτερα σημαντικό πρόβλημα που έχει μελετηθεί εκτενώς στη διεθνή βιβλιογραφία καθώς η γνώση της μπορεί να προσφέρει σημαντικά οφέλη τόσο σε οικονομικό επίπεδο (πχ για πρόβλεψη αξίας μετοχών ή κρυπτονομισμάτων) όσο και στην βέλτιστη οργάνωση ενεργειών και λήψη αποφάσεων. Ωστόσο αρκετές φορές η ακριβής πρόβλεψη της τιμής είναι ένα αρκετά δύσκολο πρόβλημα και γι΄αυτό μελετάται και το classification πρόβλημα δηλαδή εάν η τιμή μιας χρονοσειρά αυξάνεται ή μειώνεται την επόμενη ημέρα.

4.2.1.1 ARIMA

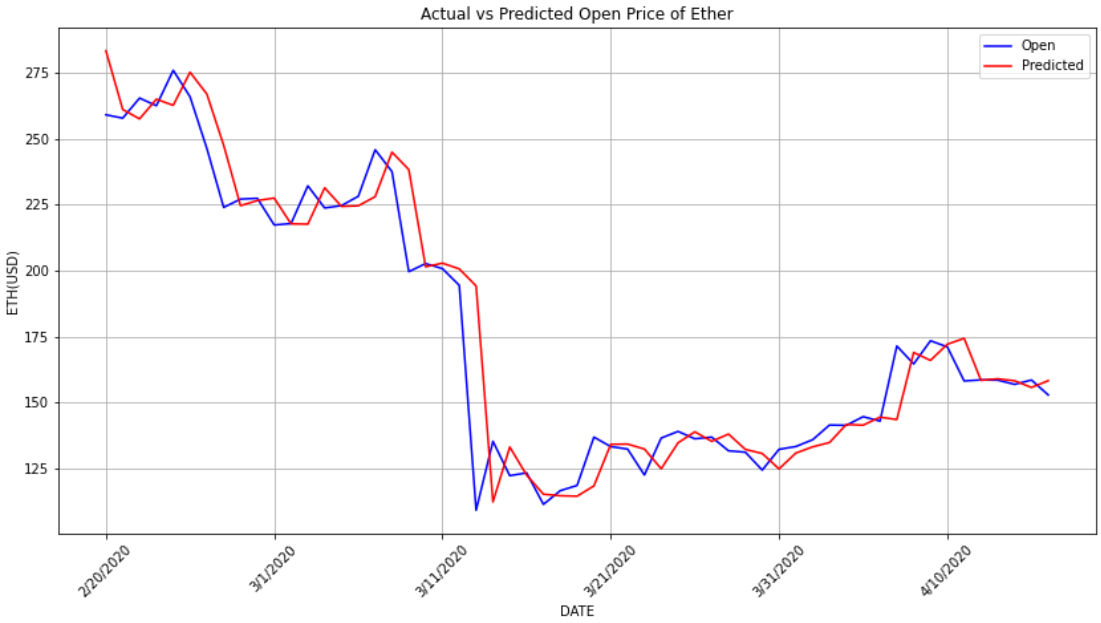
Το μοντέλο αυτό είναι το πιο διαδεδομένο στατιστικό μοντέλο που χρησιμοποιείται στην πρόβλεψη χρονοσειρών λόγω της αυξημένης του απόδοσης και ευκολίας χρήσης. Το ARIMA δεν χρησιμοποιεί επιπρόσθετα χαρακτηριστικά όπως οι υπόλοιποι αλγόριθμοι μηχανικής μάθησης και η μόνη απαίτηση που έχει από την χρονοσειρά εισόδου είναι να είναι στατική γι΄αυτό και έγινε διαφορετική προεπεξεργασία στα δεδομένα σε σχέση με τους υπόλοιπους αλγορίθμους.

Όπως φαίνεται στην Εικόνα 24. η τιμή του Ether είναι μη – στατική χρονοσειρά και γι΄αυτό μετατράπηκε σε στατική με τη μέθοδο differencing πρώτου βαθμού, αφαιρώντας δηλαδή από κάθε χρονική στιγμή την προηγούμενη της.

Εικόνα 24. Έλεγχος στατικότητας της τιμής του Ether

Στη συνέχεια αναπτύχθηκε ένα μοντέλο ARIMA κυλιόμενου παραθύρου και έπειτα από grid search οι παράμετροι του ορίστηκαν στις τιμές:

- p = 4, έλεγχος στο διάστημα [0 - 5]
- d = 1, εξετάστηκε μόνο η τιμή 1
- q = 0, έλεγχος στο διάστημα [0 - 2]



Εικόνα 25. Ημερήσιες προβλέψεις της τιμής του Ether (ARIMA)

Όσον αφορά το regression πρόβλημα βλέπουμε ότι η προβλεπόμενη τιμή είναι αρκετά κοντά στην πραγματική. Πιο συγκεκριμένα το μοντέλο παρουσιάζει MSE = 246.8, RMSE = 15.71 και MAPE = 5.5% έχει δηλαδή πολύ μικρές αποκλίσεις από την πραγματική τιμή. Ωστόσο η προβλεπόμενη χρονοσειρά φαίνεται απλά να ακολουθεί την πραγματική και το μοντέλο να μην έχει κάποια σημαντική γνώση σχετικά με την κατανομή των τιμών του Ether. Το φαινόμενο αυτό φαίνεται και από το γεγονός ότι παρουσιάζει μικρό accuracy ίσο με 54.39%.

4.2.1.2 LSTM

Πέρα από την κατάλληλη προεπεξεργασία των δεδομένων που χρησιμοποιούν, η απόδοση των νευρωνικών δικτύων και συνεπώς και του LSTM εξαρτάται σε καίριο βαθμό από τις τιμές των υπερπαραμέτρων τους. Για την κατασκευή του βέλτιστου μοντέλου εξετάστηκαν οι εξής υπερπαράμετροι:

- αριθμός LSTM επιπέδων
- μονάδες LSTM ανά επίπεδο
- dropout ανά επίπεδο
- batch size
- αριθμός κρυφών επιπέδων
- αριθμός εποχών
- learning rate
- optimizer

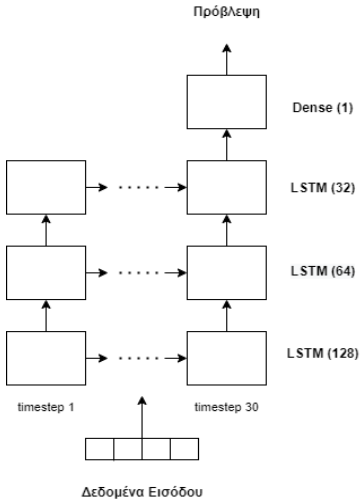
Σαν loss function χρησιμοποιήθηκε το mean squared error (mse) και σαν activation functions οι tanh και relu για τα LSTM και Dense επίπεδα αντίστοιχα. Επίσης όπως ήδη αναφέρθηκε αποδείχθηκε ότι κάποια χαρακτηριστικά δεν βοηθούσαν στις προβλέψεις όπως το Google_trends_Coinbase και γι΄αυτό αφαιρέθηκαν.

Η εύρεση του κατάλληλου συνόλου υπερπαραμέτρων έγινε με τη μέθοδο grid search υπολογίζοντας την απόδοση του δικτύου πάνω στο validation set και το βέλτιστο σύνολο παρουσιάζεται Πίνακα 2.

Υπερπαράμετρος	Σύνολο αναζήτησης	Βέλτιστη
Επίπεδα LSTM	[1, 2, 3]	3 επίπεδα
Επίπεδα Dense	[1, 2]	1 επίπεδο με 1 μονάδα
Μονάδες LSTM ανά επίπεδο	[16, 32, 64, 128, 256]	128, 64, 32 αντίστοιχα
Dropout ανά επίπεδο	[0.1, 0.2, 0.3]	0.2 σε όλα τα επίπεδα
Εποχές	[80, 100, 120, 140]	120
Learning rate	[0.0001, 0.001, 0.01, 0.1]	0.001
Batch size	[8, 16, 32]	16
optimizer	[“rmsprop”, “adam”]	rmsprop

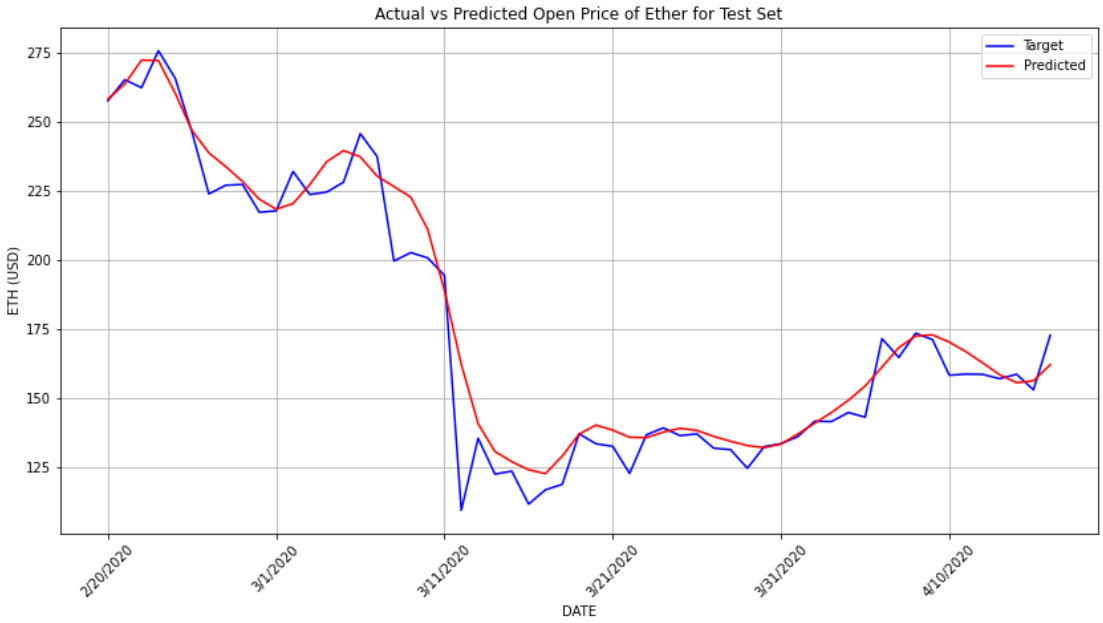
Πίνακας 2. Βέλτιστες υπερπαράμετροι LSTM – (1 ημέρα)

Στην Εικόνα 26. παρουσιάζεται η αρχιτεκτονική του δικτύου LSTM που αναπτύχθηκε. Την ίδια αρχιτεκτονική ακολουθούν όλα τα επαναληπτικά δίκτυα που αναπτύχθηκαν με τις εκάστοτε διαφοροποιήσεις στον τύπο των κελίων (στο δίκτυο GRU) και στον αριθμό των μονάδων του τελευταίου επαναληπτικού επιπέδου που είναι 64 αντί για 32 στις εβδομαδιαίες προβλέψεις όπως θα δούμε και στη συνέχεια.



Εικόνα 26. Αρχιτεκτονική δικτύου LSTM

Οι προβλέψεις του δικτύου παρουσιάζονται στην Εικόνα 27.



Εικόνα 27. Ημερήσιες προβλέψεις της τιμής του Ether (LSTM)

Όπως είναι εμφανές το δίκτυο καταφέρνει να προβλέψει σε πολύ καλά επίπεδα την τιμή του Ether και σε αντίθεση με το ARIMA φαίνεται να έχει αποκτήσει γνώση σχετικά με την κατανομή της τιμής του. Τα αποτελέσματα των μετρικών του είναι ιδιαίτερα καλά τόσο για το regression όσο και για το classification πρόβλημα.

Οι μετρικές απόδοσης για το LSTM παρουσιάζονται στον Πίνακα 3. τόσο για το regression όσο και για το classification πρόβλημα.

regression			classification			
MSE	RMSE	MAPE	Accuracy	Precision	Recall	F1 score
112.8	10.62	4.38%	78.95%	71.05%	96.43%	0.8182

Πίνακας 3. Μετρικές απόδοσης LSTM

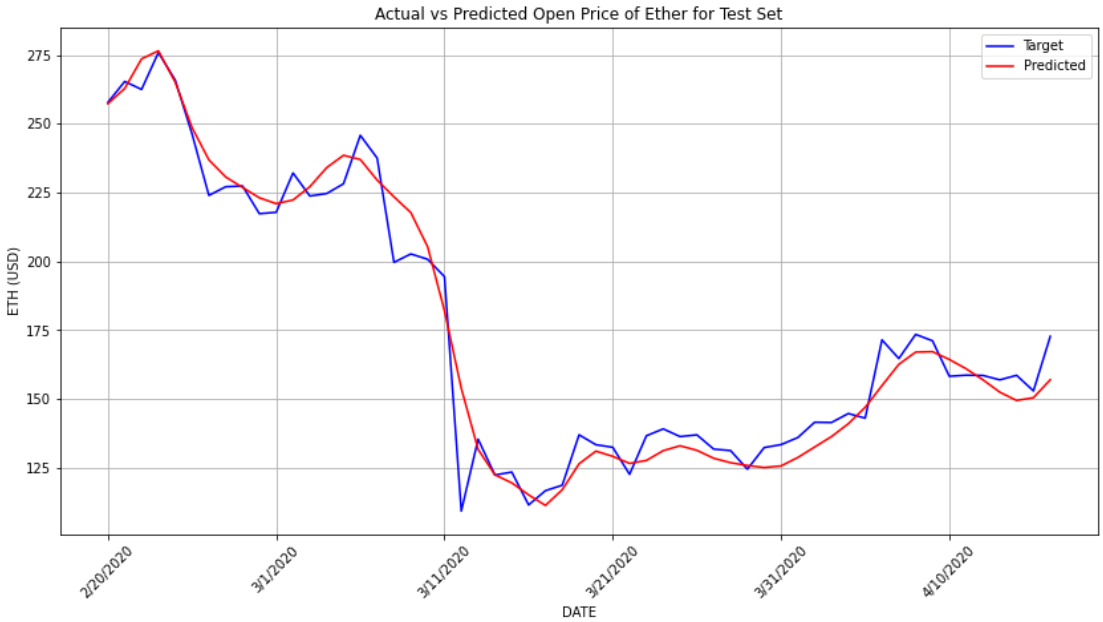
4.2.1.3 GRU

Η βελτιστοποίηση των υπερπαραμέτρων του δικτύου GRU έγινε με παρόμοιο τρόπο με αυτήν του LSTM και το βέλτιστο σύνολο όπως φαίνεται και στον Πίνακα 4. προέκυψε το ίδιο με τη διαφορά ότι η εκπαίδευση του GRU έγινε για 140 αντί για 120 εποχές.

Υπερπαράμετρος	Σύνολο αναζήτησης	Βέλτιστη
Επίπεδα GRU	[1, 2, 3]	3 επίπεδα
Επίπεδα Dense	[1, 2]	1 επίπεδο με 1 μονάδα
Μονάδες GRU ανά επίπεδο	[16, 32, 64, 128, 256]	128, 64, 32 αντίστοιχα
Dropout ανά επίπεδο	[0.1, 0.2, 0.3]	0.2 σε όλα τα επίπεδα
Εποχές	[80, 100, 120, 140]	140
Learning rate	[0.0001, 0.001, 0.01, 0.1]	0.001
Batch size	[8, 16, 32]	16
optimizer	[“rmsprop”, “adam”]	rmsprop

Πίνακας 4. Βέλτιστες υπερπαράμετροι GRU– (1 ημέρα)

Οι προβλέψεις του δικτύου GRU για το test set παρουσιάζονται στην Εικόνα 28.



Εικόνα 28. Ημερήσιες προβλέψεις της τιμής του Ether (GRU)

Και σε αυτήν την περίπτωση βλέπουμε ότι το δίκτυο πραγματοποιεί πολύ ακριβείς προβλέψεις κατορθώνοντας να προσεγγίσει σε μεγάλο βαθμό τις πραγματικές τιμές του Ether. Η μορφή της προβλεπόμενης χρονοσειράς είναι σε σημαντικό βαθμό παρόμοια με αυτή του LSTM γεγονός που είναι λογικό λόγω των πολλών ομοιοτήτων των δικτύων. Επίσης όπως βλέπουμε και στον Πίνακα 5. και το GRU παρουσιάζει υψηλή απόδοση στο regression και το classification πρόβλημα.

regression			classification			
MSE	RMSE	MAPE	Accuracy	Precision	Recall	F1 score
91.9	9.58	4.20%	77.19%	85.71%	64.29%	0.7347

Πίνακας 5. Μετρικές απόδοσης GRU

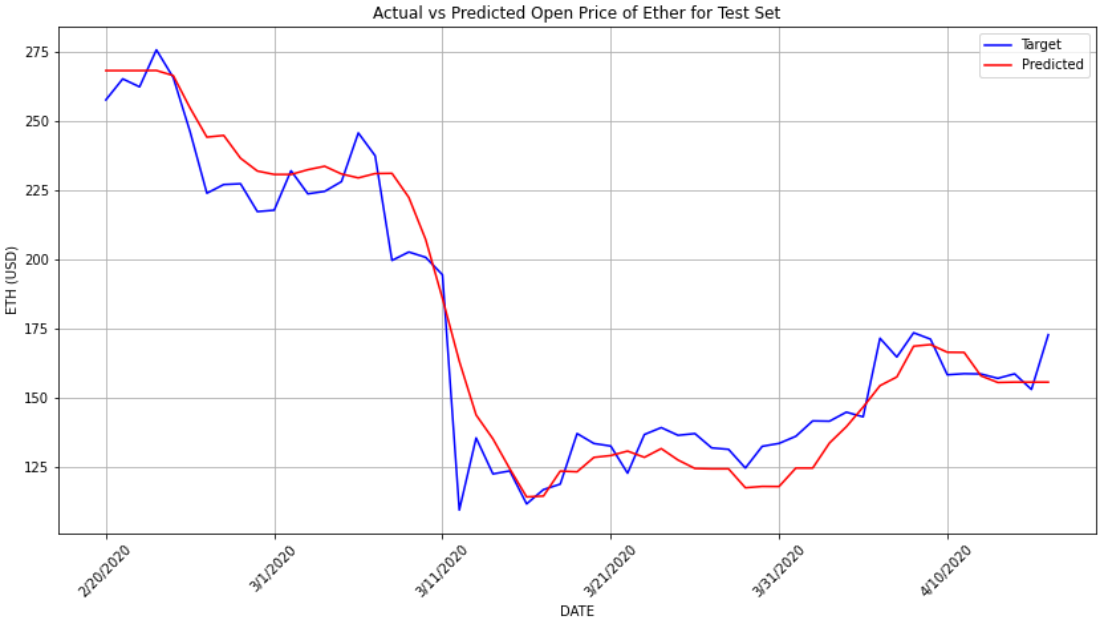
4.2.1.4 XGBoost

Ο XGBoost είναι ένας από τους πιο ευρέως χρησιμοποιούμενους αλγορίθμους που χρησιμοποιούνται τα τελευταία χρόνια στο πεδίο της μηχανικής μάθησης γεγονός που οφείλεται στην πολύ υψηλή απόδοση των προβλέψεων του και στην ταχύτητα της εκτέλεσης του. Επιπλέον ένα σημαντικό χαρακτηριστικό του είναι ότι είναι ιδιαίτερα ακριβής και με της προκαθορισμένες παραμέτρους του χωρίς να απαιτεί χρονοβόρες προσπάθειες στην βελτιστοποίηση των υπερπαραμέτρων του. Ωστόσο για την απόκτηση όσο το δυνατόν πιο ακριβών προβλέψεων γίνεται έγινε και σε αυτή την περίπτωση grid search για την εύρεση του καλύτερου δυνατού συνόλου υπερπαραμέτρων. Οι παράμετροι που επιλέχθηκαν για βελτιστοποίηση και οι τελικές τιμές τους παρουσιάζονται στον Πίνακα 6. Ένα σημείο που αξίζει να τονιστεί είναι ότι η διαδικασία εύρεσης των βέλτιστων υπερπαραμέτρων στον XGBoost ήταν ταχύτερη σε μεγάλο βαθμό από την αντίστοιχη στα δίκτυα LSTM και GRU.

Υπερπαράμετρος	Σύνολο αναζήτησης	Βέλτιστη
n_estimators	[50, 100, 150, 200]	150
learning_rate	[0.001, 0.01, 0.05, 0.1]	0.05
max_depth	[3, 4, 5, 6]	3
min_child_weight	[1, 2, 3, 4]	1

Πίνακας 6. Βέλτιστες υπερπαράμετροι XGBoost– (1 ημέρα)

Οι προβλέψεις του δικτύου XGBoost για το test set παρουσιάζονται στην Εικόνα 29.



Εικόνα 29. Ημερήσιες προβλέψεις της τιμής του Ether (XGBoost)

regression			classification			
MSE	RMSE	MAPE	Accuracy	Precision	Recall	F1 score
162.1	12.73	5.96%	75.44%	76.92%	71.42%	0.7407

Πίνακας 7. Μετρικές απόδοσης XGBoost

Όπως είναι διακριτό και από τα παραπάνω αποτελέσματα οι προβλέψεις του XGBoost έχουν πολύ υψηλά επίπεδα ακρίβειας και για τα δύο προβλήματα. Ωστόσο ενώ στο classification πρόβλημα παρουσιάζει παρόμοια απόδοση με τα LSTM και GRU υπολείπεται στο regression παρουσιάζοντας ναι μεν καλά αποτελέσματα αισθητά χειρότερα όμως από τα επαναληπτικά δίκτυα που μελετήθηκαν.

4.2.1.5 LSTM + GRU + XGBoost (Ensemble)

Η τεχνική ensemble learning αναφέρεται στον συνδυασμό των προβλέψεων πολλών διαφορετικών ή και ίδιων μοντέλων μηχανικής μάθησης με στόχο την παραγωγή ενός μοντέλου με αυξημένες αποδόσεις. Στην ενότητα αυτή παρουσιάζονται τα αποτελέσματα του μοντέλου LSTM + GRU + XGBoost χρησιμοποιώντας τα μοντέλα μηχανικής μάθησης που αναπτύχθηκαν στις προηγούμενες ενότητες. Πιο συγκεκριμένα το μοντέλο αυτό αναπτύχθηκε συνδυάζοντας τις αποφάσεις των τριών άλλων μοντέλων ως εξής:

- regression πρόβλημα: η πρόβλεψη του μοντέλου για κάθε χρονική στιγμή είναι ο μέσος όρος των προβλέψεων των τριών άλλων μοντέλων.

- classification πρόβλημα: η πρόβλεψη του μοντέλου για κάθε χρονική στιγμή είναι η πλειοψηφία των προβλέψεων των τριών άλλων μοντέλων (πχ αν τα LSTM, GRU προέβλεψαν αύξηση στην τιμή και το GRU μείωση το Ensemble μοντέλο προέβλεψε αύξηση.

Τα αποτελέσματα του μοντέλου LSTM + GRU + XGBoost παρουσιάζονται στον ακόλουθο πίνακα.

regression			classification			
MSE	RMSE	MAPE	Accuracy	Precision	Recall	F1 score
105.7	10.28	4.25%	82.46%	87.50%	75.00%	0.8077

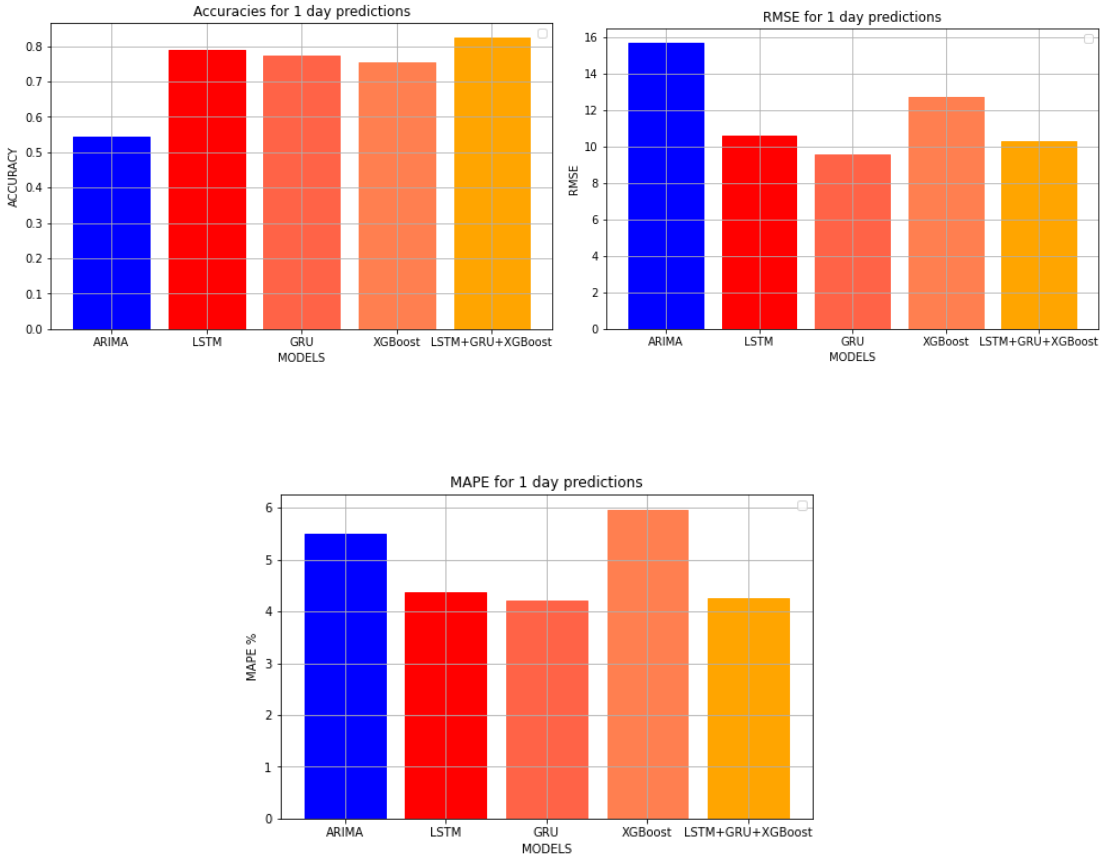
Πίνακας 8. Μετρικές απόδοσης LSTM+GRU+XGBoost

4.2.1.6 Συγκεντρωτικά αποτελέσματα ημερήσιων προβλέψεων

Σε αυτή την ενότητα παρουσιάζονται συγκεντρωτικά τα αποτελέσματα των προβλέψεων από όλες τις μεθόδους που χρησιμοποιήθηκαν για την πρόβλεψη της τιμής του Ether ώστε να είναι ευκολότερος ο σχολιασμός τους και η εξαγωγή συμπερασμάτων. Πριν την παρουσίαση των αποτελεσμάτων αξίζει να αναφερθεί ότι τα δίκτυα LSTM και GRU εκπαιδεύτηκαν σε μία Tesla T4 GPU που παρέχει το Google colab σε χρόνους 84s και 94s αντίστοιχα. Η εκπαίδευση των μοντέλων σε GPU παρουσιάζει 4.8 επιτάχυνση σε σχέση με την εκπαίδευση σε μία Intel(R) Xeon(R) CPU 2.30GHz στην οποία τα μοντέλα εκπαιδεύτηκαν σε χρόνους 405s και 409s αντίστοιχα.

	regression			classification			
Μοντέλο	MSE	RMSE	MAPE	Accuracy	Precision	Recall	F1
ARIMA	246.8	15.71	5.50%	54.38%	51.85%	52.74%	0.5229
LSTM	112.8	10.62	4.38%	78.95%	71.05%	96.43%	0.8182
GRU	91.9	9.58	4.20%	77.19%	85.71%	64.29%	0.7347
XGBoost	162.1	12.73	5.96%	75.44%	76.92%	71.42%	0.7407
Ensemble	105.7	10.28	4.25%	82.46%	87.50%	75.00%	0.8077

Πίνακας 9. Συγκεντρωτικά αποτελέσματα για ημερήσιες προβλέψεις



Εικόνα 30. Accuracy, RMSE και MAPE για ημερήσιες προβλέψεις

Όπως γίνεται εμφανές από τα αποτελέσματα Πίνακα 9. και οι τρεις μέθοδοι μηχανικής μάθησης παρουσιάζουν πολύ καλύτερα αποτελέσματα από τη μέθοδο ARIMA ιδιαίτερα για το classification πρόβλημα. Στο regression το μοντέλο ARIMA έχει σχετικά καλή απόδοση αλλά τα πολύ χαμηλά του αποτελέσματα στην πρόβλεψη σχετικά με τον εάν θα αυξηθεί ή θα μειωθεί η τιμή του Ether σε χρονικό παράθυρο μίας ημέρας δείχνει ότι δεν έχει αποκτήσει κάποια γνώση σχετικά με τις διακυμάνσεις της τιμής του Ether. Αντίθετα τα μοντέλα LSTM, GRU και XGBoost φαίνεται να έχουν αποκτήσει γνώσεις σχετικά με την πορεία της προβλεπόμενης χρονοσειράς και

γι΄αυτό παρουσιάζουν υψηλές αποδόσεις και στο classification. Πιο συγκεκριμένα το LSTM έχει την μεγαλύτερη ακρίβεια στο classification ενώ το GRU στο regression πρόβλημα. Οι αποδόσεις του XGBoost είναι εξίσου καλές ωστόσο όχι τόσο υψηλές όσο των επαναληπτικών τύπων δικτύων που ειδικεύονται στην πρόβλεψη χρονοσειρών. Τέλος το ensemble μοντέλο βελτιώνει τις προβλέψεις των τριών άλλων μοντέλων μηχανικής μάθησης παρουσιάζοντας το καλύτερο Accuracy. Ωστόσο δεν καταφέρνει να παρουσιάσει καλύτερες αποδόσεις από το GRU στο regression πρόβλημα γεγονός που μάλλον οφείλεται σε κάποιες όχι τόσο ακριβείς προβλέψεις των XGBoost και LSTM.

4.2.2 Εβδομαδιαίες προβλέψεις

Εκτός από τις βραχυχρόνιες προβλέψεις (μία ημέρα) πραγματοποιήθηκαν προβλέψεις και σε χρονικό επίπεδο μίας εβδομάδας, για να εξερευνηθούν οι δυνατότητες των αλγορίθμων που αναπτύχθηκαν και σε πιο μακροχρόνιες προβλέψεις. Οι μακροχρόνιες προβλέψεις της τιμής κρυπτονομισμάτων είναι ένα ιδιαίτερα απαιτητικό πρόβλημα (ιδιαίτερα το regression) καθώς όσο αυξάνεται το χρονικό παράθυρο της πρόβλεψης τόσο πιο αστάθμητες γίνονται και οι μεταβολές της αξίας τους.

Για την ανάπτυξη των μοντέλων για την εβδομαδιαία πρόβλεψη πραγματοποιήθηκε η ίδια μεθοδολογία με αυτή των ημερήσιων προβλέψεων με τη διαφορά ότι δεν αναπτύχθηκε μοντέλο ARIMA εξαιτίας των κακών επιδόσεων του. Οι βέλτιστες υπερπαράμετροι του κάθε μοντέλου υπολογίστηκαν με τη μέθοδο grid search και οι τιμές τους παρουσιάζονται στους ακόλουθους πίνακες.

Υπερπαράμετρος	Σύνολο αναζήτησης	Βέλτιστη
Επίπεδα LSTM	[1, 2, 3]	3 επίπεδα
Επίπεδα Dense	[1, 2]	1 επίπεδο με 1 μονάδα
Μονάδες LSTM ανά επίπεδο	[16, 32, 64, 128, 256]	128, 64, 64 αντίστοιχα
Dropout ανά επίπεδο	[0.1, 0.2, 0.3]	0.2 σε όλα τα επίπεδα
Εποχές	[80, 100, 120, 140]	120
Learning rate	[0.0001, 0.001, 0.01, 0.1]	0.001
Batch size	[8, 16, 32]	16
optimizer	[“rmsprop”, “adam”]	rmsprop

Πίνακας 10. Βέλτιστες υπερπαράμετροι LSTM – (7 ημέρες)

Υπερπαράμετρος	Σύνολο αναζήτησης	Βέλτιστη
Επίπεδα GRU	[1, 2, 3]	3 επίπεδα
Επίπεδα Dense	[1, 2]	1 επίπεδο με 1 μονάδα
Μονάδες GRU ανά επίπεδο	[16, 32, 64, 128, 256]	128, 64, 64 αντίστοιχα
Dropout ανά επίπεδο	[0.1, 0.2, 0.3]	0.2 σε όλα τα επίπεδα
Εποχές	[80, 100, 120, 140]	140
Learning rate	[0.0001, 0.001, 0.01, 0.1]	0.001
Batch size	[8, 16, 32]	16
Optimizer	[“rmsprop”, “adam”]	rmsprop

Πίνακας 11. Βέλτιστες υπερπαράμετροι GRU – (7 ημέρες)

Υπερπαράμετρος	Σύνολο αναζήτησης	Βέλτιστη
n_estimators	[50, 100, 150, 200]	150
learning_rate	[0.001, 0.01, 0.05, 0.1]	0.05
max_depth	[3, 4, 5, 6]	3
min_child_weight	[1, 2, 3, 4]	1

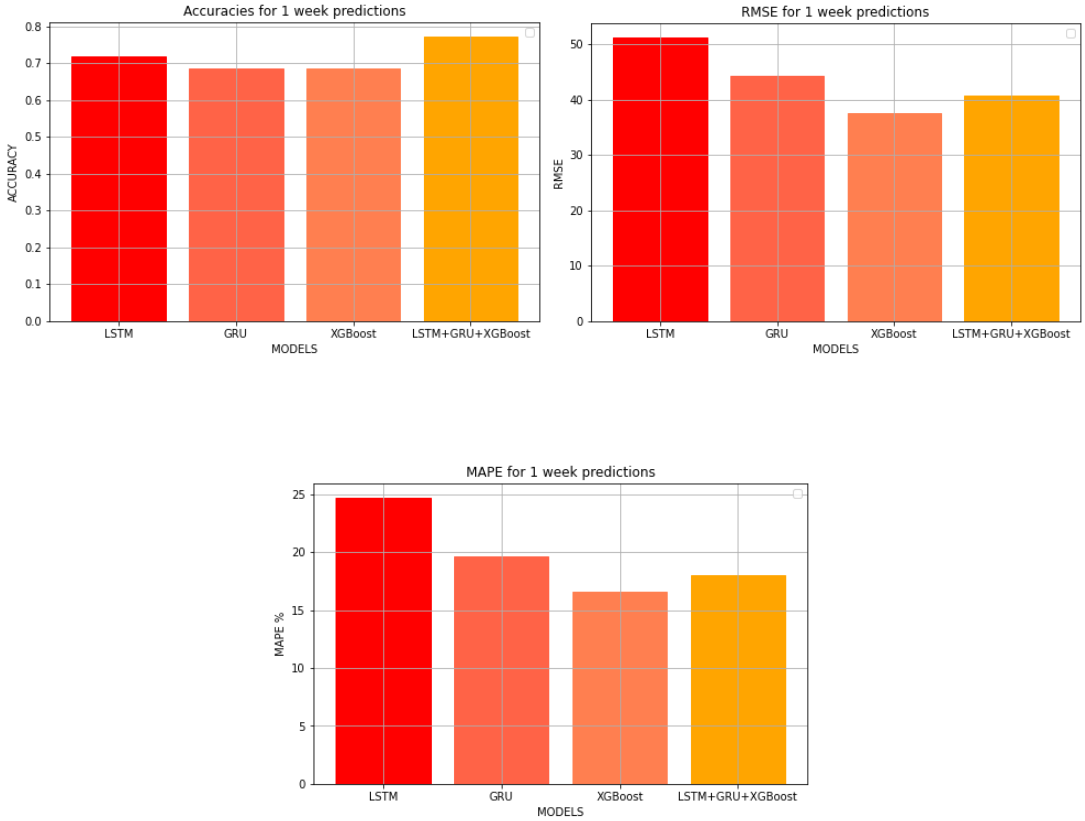
Πίνακας 12. Βέλτιστες υπερπαράμετροι XGBoost – (7 ημέρες)

Όπως φαίνεται το σύνολο των βέλτιστων υπερπαραμέτρων είναι και για τους τρεις αλγορίθμους σχεδόν το ίδιο με αυτό των ημερησίων προβλέψεων (για τα LSTM και GRU αυξήθηκε μόνο ο αριθμός των νευρώνων του τελευταίου επαναληπτικού επιπέδου από 32 σε 64, ενώ οι υπερπαράμετροι του XGBoost παρέμειναν οι ίδιες).

Επιπλέον, παρόμοια με τις ημερήσιες προβλέψεις πραγματοποιήθηκε και έλεγχος πάνω στο validation set για το αν τα χαρακτηριστικά που επιλέχθηκαν στην ενότητα 3.2.4 είναι όντως τα βέλτιστα. Αποδείχθηκε ότι τα χαρακτηριστικά Google_trends_Coinbase, Tx_per_day και mining_difficulty δεν επιδρούσαν θετικά στην απόδοση του μοντέλου και γι΄αυτό δεν χρησιμοποιήθηκαν. Έτσι το τελικό σύνολο δεδομένων που χρησιμοποίησαν οι αλγόριθμοι ήταν το ίδιο για τις ημερήσιες και τις εβδομαδιαίες προβλέψεις αναδεικνύοντας έτσι το γεγονός ότι η τιμή του Ether επηρεάζεται από τους ίδιους παράγοντες για βραχυχρόνιο (μία ημέρα) και για λίγο πιο μακροχρόνιο χρονικό πλαίσιο (7 ημέρες). Έπειτα από την εκπαίδευση και των τριών μοντέλων μηχανικής μάθησης προέκυψαν τα εξής αποτελέσματα πάνω στο test set. Όπως και στην περίπτωση των ημερησίων προβλέψεων έτσι και εδώ τα μοντέλα εκπαιδεύτηκαν σε GPU με αντίστοιχη βελτίωση χρόνου εκπαίδευσης.

	regression			classification			
Μοντέλο	MSE	RMSE	MAPE	Accuracy	Precision	Recall	F1
LSTM	2624	51.2	24.7%	71.93%	61.29%	82.61%	0.7037
GRU	1954	44.2	19.6%	68.42%	60.00%	65.22%	0.6250
XGBoost	1411	37.6	16.6%	68.43%	56.41%	96.65%	0.7097
Ensemble	1664	40.8	18.0%	77.19%	64.71%	95.66%	0.7719

Πίνακας 13. Συγκεντρωτικά αποτελέσματα για εβδομαδιαίες προβλέψεις



Εικόνα 31. Accuracy, RMSE και MAPE για εβδομαδιαίες προβλέψεις

Είναι εμφανές ότι τα αποτελέσματα στο regression πρόβλημα είναι αισθητά χειρότερα από αυτά που είχαν προκύψει για τις ημερήσιες προβλέψεις γεγονός που επιβεβαιώνει την δυσκολία της προσέγγισης της τιμής κρυπτονομισμάτων για μεγάλο χρονικό παράθυρο πρόβλεψης. Ωστόσο τα αποτελέσματα του classification παραμένουν σε καλά επίπεδα σημειώνοντας κατά μέσο όρο μείωση περίπου 7% στο accuracy κατά μέσο όρο συγκριτικά με της ημερήσιες προβλέψεις. Στο regression η καλύτερη απόδοση σημειώνεται από τον XGBoost με MAPE 16.6% ενώ στο classification από το LSTM με 71.93% accuracy. Τέλος βλέπουμε ότι η χρήση της ensemble μεθόδου αυξάνει σημαντικά την ακρίβεια της πρόβλεψης στο classification παρουσιάζοντας 5.26% υψηλότερη απόδοση από το LSTM ενώ στο regression, όπως και στις ημερήσιες προβλέψεις δεν αυξάνει την απόδοση.

Κεφάλαιο 5

Σύνοψη - Συμπεράσματα

5.1 Σύνοψη

Η πρόβλεψη της τιμής των κρυπτονομισμάτων αποτελεί ένα ιδιαίτερα απαιτητικό πρόβλημα για τους ερευνητές εξαιτίας της υψηλής αστάθειας της τιμής τους γεγονός που οφείλεται στην επιρροή τους από ένα μεγάλο πλήθος παραγόντων. Στα πλαίσια της παρούσας διπλωματικής εργασίας εξετάστηκαν οι πιθανοί παράγοντες που μπορούν να επηρεάζουν την τιμή του Ether για βραχυχρόνιες (μία ημέρα) αλλά και πιο μακροχρόνιες (εφτά ημέρες) προβλέψεις καθώς και αναπτύχθηκαν μοντέλα μηχανικής μάθησης που αξιοποιώντας τους ανωτέρω παράγοντες προβλέπουν την μελλοντική πορεία της τιμής του Ether. Πιο συγκεκριμένα λήφθηκαν υπόψιν 13 χαρακτηριστικά που μπορούν να διαχωριστούν στις εξής κατηγορίες: αγορά και χρηματιστήριο, κοινωνική δημοφιλία, τεχνικοί δείκτες και χαρακτηριστικά του Ethereum blockchain. Από τα χαρακτηριστικά αυτά επιλέχθηκαν τα σημαντικότερα για τις ημερήσιες και εβδομαδιαίες προβλέψεις ξεχωριστά μέσα από ένα σύνολο τεχνικών επιλογής χαρακτηριστικών, τα οποία στη συνέχεια προεπεξεργάστηκαν κατάλληλα ώστε να μπορούν να αξιοποιηθούν από τους αλγορίθμους μηχανικής μάθησης. Τέλος αναπτύχθηκαν τα μοντέλα LSTM, GRU, XGBoost και το μοντέλο Ensemble (LSTM + GRU + XGBoost) που συνδυάζει τις προβλέψεις των επιμέρους μοντέλων. Το μοντέλο Ensemble παρουσίασε πολύ υψηλές αποδόσεις στο classification πρόβλημα σημειώνοντας accuracy 82.46% και 77.19% για ημερήσιες και εβδομαδιαίες προβλέψεις αντίστοιχα ενώ τις καλύτερες αποδόσεις στο regression είχαν τα μοντέλα GRU και XGBoost.

5.2 Συμπεράσματα – Συνεισφορά εργασίας

Η πλειοψηφία των ερευνών στο πεδίο της πρόβλεψης της τιμής κρυπτονομισμάτων έχει ασχοληθεί εκτεταμένα με την εξερεύνηση των μοντέλων που μπορούν να πετύχουν τις υψηλότερες αποδόσεις καθώς και με τους παράγοντες που μπορούν να επηρεάζουν την τιμή τους. Ωστόσο οι παράγοντες αυτοί είτε χρησιμοποιούνται μεμονωμένα είτε όλοι μαζί χωρίς να εξετάζεται ποιοι όντως βελτιώνουν τις αποδόσεις

των μοντέλων. Παράλληλα χρησιμοποιείται το ίδιο σύνολο παραγόντων ανεξάρτητα από το χρονικό παράθυρο των μελλοντικών προβλέψεων γεγονός που δεν είναι σωστό καθώς η μακροχρόνια πορεία της τιμής των κρυπτονομισμάτων μπορεί να εξαρτάται από διαφορετικούς παράγοντες σε σχέση με την βραχυχρόνια. Η παρούσα εργασία, στην οποία λήφθηκαν υπόψιν τα δύο ανωτέρω φαινόμενα, μπορεί να συνεισφέρει θετικά σε μελλοντικές έρευνες τόσο μέσω της μεθοδολογίας που ακολούθησε αλλά και των συμπερασμάτων στα οποία κατέληξε. Πιο συγκεκριμένα παρέχει τις εξής συνεισφορές:

- Εξετάστηκε η επιρροή της τιμής του Ether από παράγοντες που ανήκουν σε ένα ευρύ φάσμα πεδίων κάποιοι από τους οποίους εξετάζονταν για πρώτη φορά (δημοτικότητα πορτοφολιών κρυπτονομισμάτων).

- Προτάθηκε ένα πλαίσιο για επιλογή των σημαντικότερων από τους ανωτέρω παράγοντες, η σημαντικότητα των οποίων εξετάστηκε ξεχωριστά ανάλογα με το χρονικό παράθυρο της πρόβλεψης.

- Διαπιστώθηκε ότι οι σημαντικότεροι παράγοντες για την τιμή του Ether είναι η τιμή του bitcoin, η δημοτικότητα του Ether στο Google αλλά και οι τεχνικοί δείκτες 14ema και MACD τόσο για τις ημερήσιες όσο και για τις εβδομαδιαίες προβλέψεις.

- Στο classification πρόβλημα έχουμε προβλέψεις με πολύ υψηλή απόδοση και στις βραχυχρόνιες αλλά και στις μακροχρόνιες προβλέψεις σε αντίθεση με το regression όπου οι μακροχρόνιες προβλέψεις παρουσιάζουν λιγότερο καλές επιδόσεις.

- Το μοντέλο Ensemble που προτάθηκε παρουσιάζει καλύτερες αποδόσεις στο classification σε σχέση με το LSTM που χρησιμοποιείται σαν το state-of-the-art μοντέλο.

5.3 Μελλοντικές επεκτάσεις

Στα πλαίσια της παρούσας εργασίας μελετήθηκε η επιρροή ενός μεγάλου συνόλου παραγόντων στην επιρροή του Ether και αναπτύχθηκαν μοντέλα που προβλέπουν την τιμή του με υψηλή ακρίβεια. Ωστόσο τα μοντέλα που αναπτύχθηκαν δεν μπορούν να χρησιμοποιηθούν απευθείας για εμπορικούς σκοπούς καθώς η μελέτη έγινε για ένα συγκεκριμένο χρονικό διάστημα. Σαν μελλοντική κατεύθυνση έρευνας το σύστημα που αναπτύχθηκε μπορεί να τροποποιηθεί ώστε να είναι διαδραστικό και να πραγματοποιεί real-time προβλέψεις χρησιμοποιώντας επικαιροποιημένα δεδομένα. Ένα τέτοιο σενάριο είναι εφικτό καθώς όπως είδαμε η εκπαίδευση των μοντέλων είναι πολύ γρήγορη με αποτέλεσμα να μπορούμε να επανεκπαιδεύουμε το μοντέλο σε τακτά χρονικά διαστήματα με στόχο να αξιοποιεί και τα πιο πρόσφατα δεδομένα. Επιπλέον μπορεί να εξερευνηθεί και η επιρροή επιπρόσθετων παραγόντων όπως η δημοτικότητα του Ether σε κοινωνικά δίκτυα όπως το twitter και το github καθώς και να πραγματοποιηθούν προβλέψεις και σε μεγαλύτερο χρονικό εύρος. Τέλος, μια ενδιαφέρουσα κατεύθυνση είναι και η εκπαίδευση των μοντέλων με δεδομένα ανά ώρα και ανά λεπτό και η εξερεύνηση των παραγόντων που επηρεάζουν τις προβλέψεις σε αυτές τις περιπτώσεις.

Αναφορές

[1] S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash System,” 2008.

[2] Y. Sovbetov, “Munich Personal RePEc Archive Factors Influencing Cryptocurrency Prices: Evidence from Bitcoin, Ethereum, Dash, Litcoin, and Monero Factors Influencing Cryptocurrency Prices: Evidence from Bitcoin, Ethereum, Dash, Litcoin, and Monero,” *J. Econ. Financ. Anal.*, vol. 2, no. 2, pp. 1–27, 2018.

[3] E. Ghysels, D. R. Osborn, and P. M. M. Rodrigues, “Chapter 13 Forecasting Seasonal Time Series,” *Handb. Econ. Forecast.*, vol. 1, no. 05, pp. 659–711, 2006, doi: 10.1016/S1574-0706(05)01013-X.

[4] M. Chen, N. Narwal, and M. Schultz, “Predicting Price Changes in Ethereum,” *Cs229.Stanford.Edu*, no. 2016, pp. 1–6, 2017, [Online]. Available: http://cs229.stanford.edu/proj2017/final-reports/5244039.pdf.

[5] C. H. Wu, C. C. Lu, Y. F. Ma, and R. S. Lu, “A new forecasting framework for bitcoin price with LSTM,” *IEEE Int. Conf. Data Min. Work. ICDMW*, vol. 2018-Novem, pp. 168–175, 2019, doi: 10.1109/ICDMW.2018.00032.

[6] S. McNally, J. Roche, and S. Caton, “Predicting the Price of Bitcoin Using Machine Learning,” *Proc. - 26th Euromicro Int. Conf. Parallel, Distrib. Network-Based Process. PDP 2018*, pp. 339–343, 2018, doi: 10.1109/PDP2018.2018.00060.

[7] M. Glenski, T. Weninger, and S. Volkova, “Improved Forecasting of Cryptocurrency Price using Social Signals,” 2019, [Online]. Available: http://arxiv.org/abs/1907.00558.

[8] Z. Chen, C. Li, and W. Sun, “Bitcoin price prediction using machine learning: An approach to sample dimension engineering,” *J. Comput. Appl. Math.*, vol. 365, p. 112395, 2020, doi: 10.1016/j.cam.2019.112395.

[9] D. Kumar and S. K. Rath, “Predicting the Trends of Price for Ethereum Using Deep Learning Techniques,” *Adv. Intell. Syst. Comput.*, vol. 1056, pp. 103–114, 2020, doi: 10.1007/978-981-15-0199-9_9.

[10] M. M. Patel, S. Tanwar, R. Gupta, and N. Kumar, “A Deep Learning-based Cryptocurrency Price Prediction Scheme for Financial Institutions,” *J. Inf.*

Secur. Appl., vol. 55, no. May, p. 102583, 2020, doi: 10.1016/j.jisa.2020.102583.

[11] V. Buterin, "A next-generation smart contract and decentralized application platform," *Etherum*, no. January, pp. 1–36, 2014, [Online]. Available: http://buyxpr.com/build/pdfs/EthereumWhitePaper.pdf.

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986, doi: 10.1038/323533a0.

[13] S. Hochreiter, "The vanishing gradient problem during learning Recurrent Neural Networks and problem solutions," 1997.

[14] S. Hochreiter and J. Schmidhuber, "LSTM can solve hard long time lag problems," *Adv. Neural Inf. Process. Syst.*, pp. 473–479, 1997.

[15] K. Cho *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1724–1734, 2014, doi: 10.3115/v1/d14-1179.

[16] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and Understanding Recurrent Networks," pp. 1–12, 2015, [Online]. Available: http://arxiv.org/abs/1506.02078.

[17] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[18] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.

[19] N. Smuts, "What drives cryptocurrency prices? An investigation of Google trends and telegram sentiment," *Perform. Eval. Rev.*, vol. 46, no. 3, pp. 131–134, 2019, doi: 10.1145/3308897.3308955.

[20] Y. Zhai, A. Hsu, and S. K. Halgamuge, "Combining news and technical indicators in daily stock price trends prediction," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4493 LNCS, no. PART 3, pp. 1087–1096, 2007, doi: 10.1007/978-3-540-72395-0_132.

[21] M. Moein Aldin, H. Dehghan Dehnavi, and S. Entezari, "Evaluating the Employment of Technical Indicators in Predicting Stock Price Index Variations Using Artificial Neural Networks (Case Study: Tehran Stock Exchange)," *Int. J. Bus. Manag.*, vol. 7, no. 15, pp. 25–34, 2012, doi: 10.5539/ijbm.v7n15p25.

[22] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, “Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products,” *Chemom. Intell. Lab. Syst.*, vol. 83, no. 2, pp. 83–90, 2006, doi: 10.1016/j.chemolab.2006.01.007.

[23] Z. Shi and M. Han, “Support Vector Echo - State Machine for Chaotic Time Series Prediction,” pp. 1–31.

[24] J. Luo, K. Ying, and J. Bai, “Savitzky-Golay smoothing and differentiation filter for even number data,” *Signal Processing*, vol. 85, no. 7, pp. 1429–1434, 2005, doi: 10.1016/j.sigpro.2005.02.002.