



Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Υπολογιστική Νοημοσύνη

Κωνσταντίνος Παπακώστας

AEM: 3064

Εργασία στα Βαθιά Νευρωνικά Δίκτυα (DNNs)

01/06/2021

Περιεχόμενα

1. Εισαγωγή	2
2. Υπόβαθρο	3
2.1. Περιγραφή χαρακτηριστικών	3
2.2. Αρχιτεκτονικές βαθιών νευρωνικών δικτύων	5
2.2.1. Multi Layer Perceptron (MLP)	5
2.2.2. Convolutional Neural Network (CNN)	6
2.2.3. Recurrent Neural Network (RNN)	8
2.2.4. Long Short-Term Memory (LSTM)	8
3. Σχετικές Μελέτες	9
4. Αλγόριθμος & Μεθοδολογία	10
4.1. Συλλογή δεδομένων	10
4.1.1. Αξία Bitcoin και όγκος συναλλαγών ανά ημέρα	11
4.1.2. Google Trends	11
4.1.3. Δείκτης Standard and Poor's 500	11
4.1.4. Ισοτιμίες νομισμάτων	12
4.1.5. Στατιστικά του Bitcoin blockchain	12
4.2. Προεπεξεργασία δεδομένων	12
4.2.1. Συμπλήρωση ελλειπών τιμών	12
4.2.2. Επιλογή χαρακτηριστικών	12
4.2.3. Υπολογισμός τεχνικών δεικτών	14
4.2.4. Υπολογισμός τιμής - στόχου	15
4.2.5. Κλιμάκωση χαρακτηριστικών	15
4.2.6. Χωρισμός σε ακολουθίες	15
4.3. Πρόβλεψη κατεύθυνσης κίνησης για την αξία του Bitcoin	17
4.3.1. Support Vector Machine (SVM)	19
4.3.2. Deep Neural Network (DNN)	19
4.3.3. Long Short-Term Memory (LSTM)	20
4.3.4. Συνδυασμός CNN - LSTM	20
4.4. Έκθεση αποτελεσμάτων	20
5. Συμπεράσματα & Μελλοντικές Επεκτάσεις	22
Αναφορές	23

1. Εισαγωγή

Η παρούσα εργασία αποτελεί μέρος του μαθήματος “Υπολογιστική Νοημοσύνη” του Τμήματος Πληροφορικής ΑΠΘ για το ακαδημαϊκό έτος 2020 - 2021. Το αντικείμενο που πραγματεύεται είναι η χρήση βαθιών νευρωνικών δικτύων, όπως το μοντέλο Long Short-Term Memory (LSTM), για την πρόβλεψη της κατεύθυνσης κίνησης της τιμής του κρυπτονομίσματος Bitcoin (Nakamoto 2009).

Το Bitcoin είναι ένα σύστημα ηλεκτρονικών πληρωμών το οποίο έχει ως στόχο την αποκέντρωση (decentralization) του ελέγχου όλων των συναλλαγών, έτσι ώστε οι συμμετέχοντες να μη χρειάζεται να εμπιστεύονται κάποιον τρίτο οργανισμό όπως τα τραπεζικά συστήματα για να επικυρωθεί η εγκυρότητά τους. Ο τρόπος με τον οποίο επιτυγχάνεται αυτό είναι η χρήση κρυπτογραφικών αλγορίθμων που επαληθεύουν τις νέες συναλλαγές, λαμβάνοντας υπόψη ταυτόχρονα εμμέσως όλες αυτές που έχουν προηγηθεί. Η αλυσιδωτή αυτή εξάρτηση των συναλλαγών αποτελεί το υπόβαθρο της τεχνολογίας “blockchain”, που εξασφαλίζει ότι κανένας παράγοντας δεν μπορεί να παραποιήσει το ιστορικό προς όφελός του.

Ωστόσο, η τεχνική θεμελίωση του δικτύου δεν καθορίζει αυτομάτως την αξία του νομίσματος, παρά μόνο θέτει ένα άνω όριο στο πλήθος των Bitcoins που θα υπάρχουν συνολικά σε κυκλοφορία όταν ο αλγόριθμος “εξόρυξης” σταματήσει να παράγει νέα νομίσματα. Αν και αυτό το απόλυτο όριο προσδίδει μια σπανιότητα στο κάθε Bitcoin, η ονομαστική αξία με την οποία ανταλλάσσεται το καθένα καθορίζεται εξ ολοκλήρου από την ελεύθερη αγορά με βάση τον κανόνα της προσφοράς και της ζήτησης.

Στο χώρο του χρηματιστηρίου, διάφοροι οικονομικοί αναλυτές μελετούν τις μετοχές εταιρειών έτσι ώστε να προβλέψουν την κατεύθυνση κίνησής τους, δηλαδή αν θα αυξηθεί ή θα μειωθεί η αξία της κάθε μετοχής, για να επενδύσουν ανάλογα και να μεγιστοποιήσουν το κέρδος τους. Πλέον, η αγορά των κρυπτονομισμάτων φαίνεται να ακολουθεί μία αντίστοιχη λογική αγοραπωλησιών (trading), κατά την οποία χρήστες που ενδιαφέρονται για ένα κρυπτονόμισμα όπως το Bitcoin μπορούν να καταθέσουν “παραδοσιακά” νομίσματα όπως ευρώ και δολάρια (fiat money) και να λάβουν το αντίστοιχο ποσό σε κρυπτονομίσματα. Επομένως υπάρχει ένα προφανές ενδιαφέρον στην πρόβλεψη της αξίας τέτοιων χρηματοοικονομικών περιουσιακών στοιχείων, καθώς πολλοί είναι αυτοί που ενδιαφέρονται για το κέρδος που μπορεί να τους αποφέρει ένα σύνολο από αγοραπωλησίες, παρά για το ίδιο το νόμισμα και τον σκοπό για τον οποίο είχε δημιουργηθεί.

2. Υπόβαθρο

Οι προβλέψεις για την εξέλιξη τιμής τέτοιων δεικτών συχνά συνοδεύονται από μεγάλες και απρόβλεπτες διακυμάνσεις, οι οποίες δεν είναι εύκολα επεξηγήσιμες ως προς την αιτία που τις προκαλεί. Ειδικότερα στα κρυπτονομίσματα, τα οποία δεν έχουν εδραιωθεί στις χρηματοοικονομικές αγορές όσο και οι μετοχές, οποιαδήποτε βραχυπρόθεσμη εκτίμηση έχει σχετικά μεγάλο διάστημα εμπιστοσύνης, που σημαίνει ότι η πιθανότητα οικονομικής απώλειας είναι επίσης μεγάλη.

2.1. Περιγραφή χαρακτηριστικών

Υποθέτοντας ότι είναι εφικτή μια τέτοια πρόβλεψη, ο μόνος τρόπος με τον οποίο μπορούμε να αναπτύξουμε μια κερδοφόρα στρατηγική είναι να αναλύσουμε πολλά και σύνθετα σήματα ώστε να εκμεταλλευτούμε την όποια συσχέτισή τους με την πορεία του Bitcoin. Επομένως, ένα σημαντικό μέρος της έρευνας σε αυτό το θέμα αφορά το κομμάτι των χαρακτηριστικών τα οποία θα εκμεταλλεύεται η τεχνική μας ώστε να αποφανθεί για το αν η τιμή θα κινηθεί ανοδικά ή καθοδικά.

Αρχικά, τα πιο συνηθισμένα στοιχεία που χρησιμοποιούνται για την επίλυση του προβλήματος είναι αυτά που αφορούν το ίδιο το blockchain του Bitcoin. Καθώς τα δεδομένα είναι δημόσια και ανοιχτά προς όλους, μπορούμε να εκμεταλλευτούμε μετρικές όπως το πλήθος των συναλλαγών ανά ημέρα αλλά και το πλήθος των διευθύνσεων που χρησιμοποιούνται για τις συναλλαγές. Ταυτόχρονα, ιδιαίτερα χρήσιμα είναι και τα στοιχεία που αφορούν τη διαδικασία του υπολογισμού του νέου συνόλου συναλλαγών (block of transactions), το οποίο είναι γνωστό και ως “mining”. Ο αλγόριθμος αυτός χαρακτηρίζεται από μία αυτορυθμιζόμενη δυσκολία έτσι ώστε το κάθε νέο block να υπολογίζεται περίπου κάθε 10 λεπτά. Επίσης, η συνεισφορά των “miners”, δηλαδή των μηχανημάτων που συμμετέχουν στη διαδικασία του mining που ανταμείβονται με ένα ποσοστό του νέου Bitcoin που θα δημιουργηθεί, μετριέται με βάση την υπολογιστική ισχύ που αφιερώνεται για την εύρεση ενός νέου block.

Έχει παρατηρηθεί (Vassiliadis et al. 2017) ότι, πέρα από τα χαρακτηριστικά του blockchain που αναφέρθηκαν παραπάνω, ορισμένοι άλλοι χρηματιστηριακοί δείκτες όπως ο S&P 500 εμφανίζουν μία συσχέτιση με την τιμή του Bitcoin. Αυτό εξηγείται αν αναλογιστούμε ότι όταν η αγορά έχει γενικότερα θετικό πρόσημο, με τις μετοχές μεγάλων εταιριών να σημειώνουν θετική πρόοδο, τότε γενικότερα είναι πιθανό να υπάρχει και ανάπτυξη και σε άλλα χρηματοοικονομικά περιουσιακά στοιχεία όπως τα κρυπτονομίσματα.

Επίσης, ένας τρόπος βελτίωσης της ακρίβειας των μοντέλων είναι η εκμετάλλευση τεχνικών δεικτών υψηλών διαστάσεων (Huang et al. 2019). Οι δείκτες αυτοί μπορούν να διαχωριστούν σε πέντε κύριες κατηγορίες:

i. “Overlap study” indicators

Οι δείκτες αυτοί χρησιμοποιούνται για να τονίσουν μία τάση ανόδου ή καθόδου μέσω εξομάλυνσης της ταχείας μεταβλητότητας κάποιας τιμής. Ένα από τα πιο διαδεδομένα παραδείγματα ενός τέτοιου δείκτη είναι ο κινούμενος μέσος όρος (Moving Average - MA). Κατά κανόνα, τέτοιες μετρικές αποτυπώνουν την παρελθοντική αξία ενός στοιχείου και για το λόγο αυτό θεωρούνται ότι “υστερούν” της τρέχουσας τιμής (lagging indicators).

ii. “Cycle” indicators

Στην περίπτωση των κυκλικών δεικτών, έμφαση δίνεται στην ανίχνευση επαναλαμβανόμενων μοτίβων σε μία αγορά, οι οποίοι υποδεικνύουν ότι η αξία ενός στοιχείου μπορεί σύντομα είτε να υποτιμηθεί είτε να υπερεκτιμηθεί. Ένα χαρακτηριστικό τέτοιο παράδειγμα είναι ο δείκτης Schaff Trend Cycle (STC).

iii. “Momentum” indicators

Στις τεχνικές αναλύσεις, πολύτιμη γνώση αποτελεί η ταχύτητα με την οποία αλλάζει η τιμή ενός στοιχείου, κάτι το οποίο επιτυγχάνεται με τους δείκτες ορμής. Σε αυτή την κατηγορία οι τιμές “προηγούνται” της τρέχουσας τιμής καθώς συνεπάγονται εμμέσως είτε τη συνέχεια είτε την ανατροπή της πορείας που παρατηρείται ήδη (leading indicators). Ένα από τα συνηθισμένα παραδείγματα σε αυτή την κατηγορία δεικτών είναι ο δείκτης σχετικής ισχύος (Relative Strength Index - RSI) (Wilder 1978).

iv. “Volatility” indicators

Οι δείκτες πτητικότητας χρησιμοποιούνται για να τονίσουν πόσο γρήγορα αλλάζει η τάση ενός στοιχείου και άρα πόσο επικίνδυνη είναι μια επένδυση σε αυτό. Σε αυτή την κατηγορία υπάρχουν τόσο απλές μετρικές, όπως η τυπική απόκλιση, όσο και σύνθετες, όπως το μέσο πραγματικό εύρος (Average True Range - ATR) (Wilder 1978) και οι Bollinger Bands (Bollinger 1992).

v. “Pattern recognition” indicators

Μελετώντας τις γραφικές αναπαραστάσεις της εξέλιξης της τιμής ενός χρηματοοικονομικού στοιχείου, εντοπίζονται ορισμένα σχηματικά μοτίβα τα οποία συνήθως συνοδεύονται από αύξηση ή πτώση στην αξία του στοιχείου. Μερικά από αυτά είναι τα “three black crows”, τα “doji stars” και το “dark cloud”.

Παράλληλα, αντίστοιχες έρευνες (Dyhrberg 2016) έχουν δείξει ότι η τιμή του Bitcoin μπορεί επίσης να επηρεαστεί από συναλλαγματικές ισοτιμίες διαφόρων fiat νομισμάτων, όπως είναι το αμερικανικό δολάριο (USD), το ευρώ (EUR), το ελβετικό φράγκο (CHF), η αγγλική λίρα (GBP) και το κινεζικό γουάν (CNY).

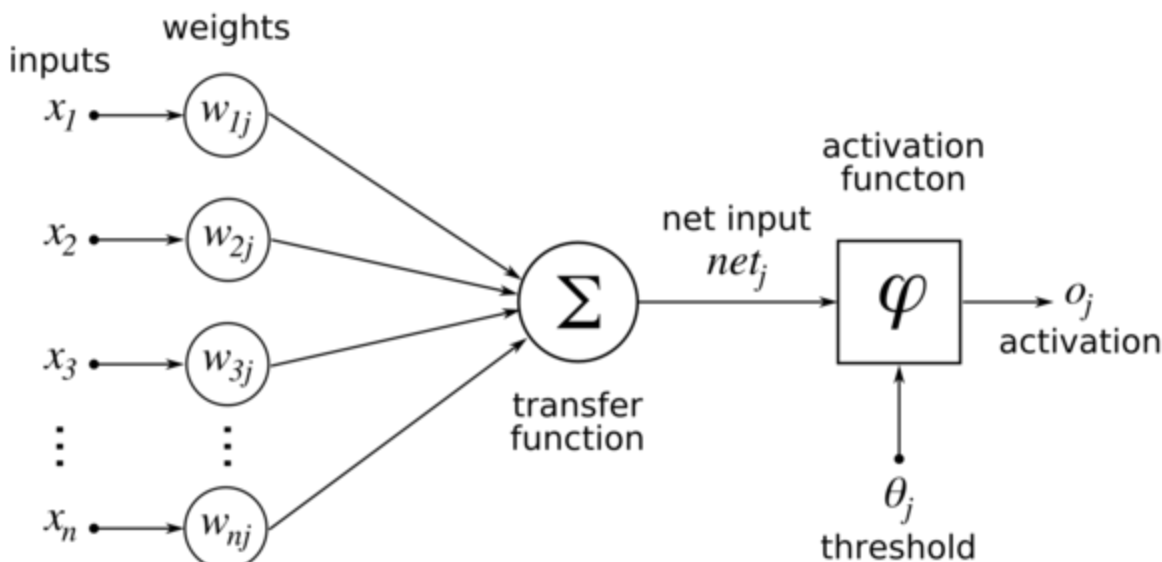
Τέλος, είναι άξιο ενδιαφέροντος να μελετηθεί πώς μπορούν να χρησιμοποιηθούν μετρικές όπως το πλήθος αναζητήσεων του όρου “Bitcoin” στις μηχανές αναζήτησης του διαδικτύου, καθώς κάτι τέτοιο φανερώνει το ενδιαφέρον του ευρύτερου κοινού σε μια ορισμένη χρονική περίοδο. Σύμφωνα με προηγούμενη έρευνα σε αυτό το ζήτημα (Garcia et al. 2014), όταν ο γενικός πληθυσμός που ενώ δεν έχει προηγούμενη επαφή με το αντικείμενο ξαφνικά θέλει να ασχοληθεί με αυτό, τότε παρατηρείται μία διόγκωση στην τιμή. Αυτές οι “φούσκες” (bubbles), παρόλο που δεν αποτελούν ακριβή και έγκυρη ένδειξη της πραγματικής αξίας ενός στοιχείου, μπορούν να ληφθούν υπόψη με σκοπό να προσαρμοστεί κατά έναν παράγοντα η πρόβλεψη, είτε επηρεάζοντάς τη θετικά είτε αρνητικά.

2.2. Αρχιτεκτονικές βαθιών νευρωνικών δικτύων

2.2.1. Multi Layer Perceptron (MLP)

Οι πιο κλασικές αρχιτεκτονικές νευρωνικών δικτύων βασίζονται στη δομή του ανθρώπινου εγκεφάλου όπου πολλοί βιολογικοί νευρώνες ενώνονται μεταξύ τους και σχηματίζουν μια σύνθετη δομή η οποία λαμβάνει ερεθίσματα και αντιδρά σε αυτά εκτελώντας κάποια ενέργεια. Με βάση αυτή τη λογική, ο Rosenblatt (1958) επινόησε το μοντέλο του Perceptron, ενός τεχνητού νευρώνα ο οποίος δέχεται ως είσοδο αριθμητικές τιμές, τις αθροίζει συνυπολογίζοντας το βάρος κάθε εισόδου - το ποσοστό, δηλαδή, κατά το οποίο επηρεάζεται η συμπεριφορά του νευρώνα από την εκάστοτε είσοδο - και τελικά εφαρμόζει μια συνάρτηση ενεργοποίησης ώστε να υπολογιστεί η έξοδος που θα προωθήσει στον επόμενο νευρώνα. Μία γραφική απεικόνιση του τρόπου λειτουργίας του Perceptron φαίνεται στην Εικόνα 2.1.

Ωστόσο, ένας νευρώνας Perceptron μπορεί να διακρίνει μόνο μεταξύ γραμμικά διαχωρίσιμων κλάσεων, με χαρακτηριστικό παράδειγμα την αδυναμία επίλυσης του προβλήματος XOR. Έτσι, το επόμενο λογικό βήμα είναι η διαδοχική παράταξη πολλαπλών επιπέδων με νευρώνες Perceptron, η οποία σχηματίζει την αρχιτεκτονική νευρωνικών δικτύων πολυστρωματικού Perceptron (Multi Layer Perceptron - MLP). Το γεγονός ότι τα επόμενα επίπεδα νευρώνων δέχονται ως είσοδο την έξοδο ενός προηγούμενου επιπέδου Perceptron προσδίδει στο μοντέλο την ικανότητα να επιλύει οποιοδήποτε πρόβλημα κατηγοριοποίησης ή παλινδρόμησης, δεδομένου ότι υπάρχει ένα ικανοποιητικό πλήθος νευρώνων και επιπέδων.



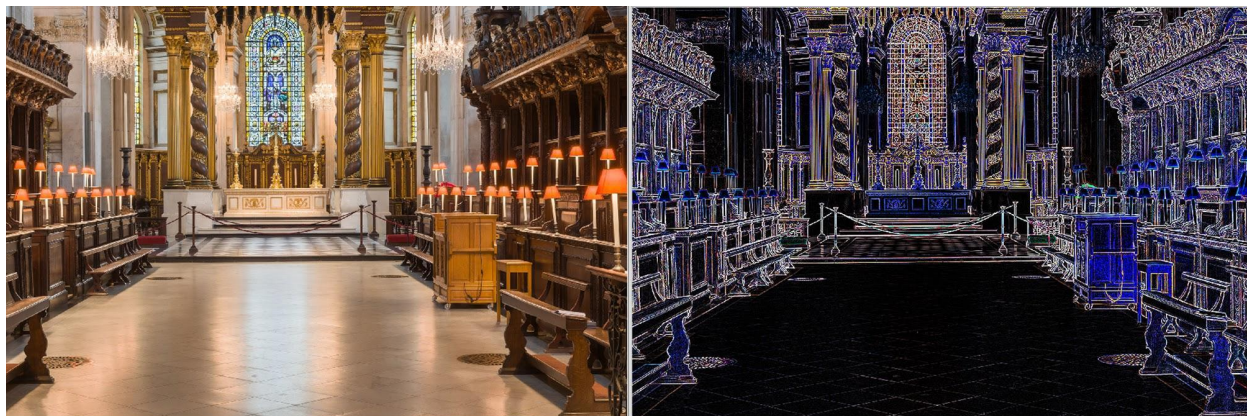
Εικόνα 2.1: Γραφική απεικόνιση ενός νευρώνα Perceptron
(Perceptron. Mitchell, Machine Learning, p87., CC BY-SA 3.0)

Χάρη σε αυτή τη δυνατότητα, τα νευρωνικά δίκτυα MLP έχουν πλέον καθιερωθεί ως ένα από τα πιο ισχυρά μοντέλα μηχανικής μάθησης, βρίσκοντας εφαρμογή τόσο στην ακαδημαϊκή κοινότητα όσο και στη βιομηχανία, καθώς είναι ικανά να εντοπίσουν μη γραμμικές συσχετίσεις μεταξύ των δεδομένων εισόδου και του στόχου που πρέπει να προβλέψουν, πετυχαίνοντας ιδιαίτερα υψηλή ακρίβεια στα αποτελέσματά τους.

2.2.2. Convolutional Neural Network (CNN)

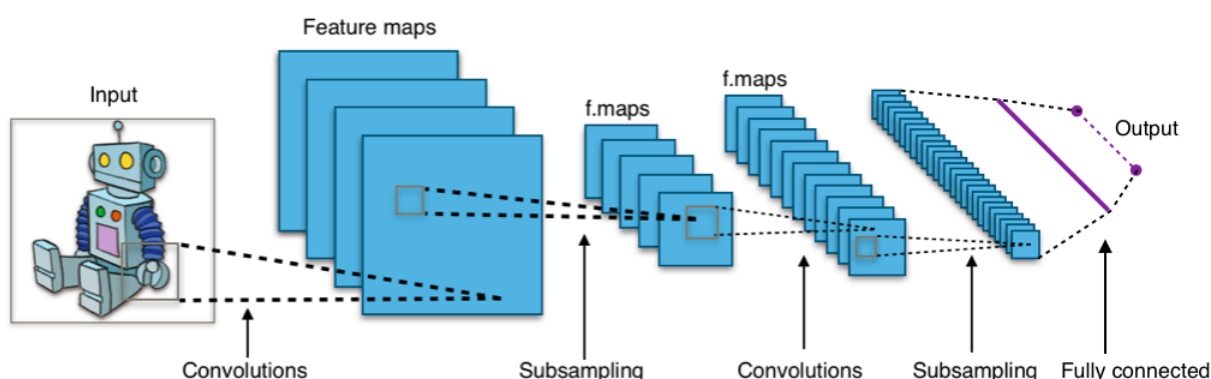
Σε συγκεκριμένες περιπτώσεις, τα δεδομένα μπορεί να είναι δομημένα με τέτοιο τρόπο που οι “γειτονικές” τιμές ενός στοιχείου έχουν κάποια φυσική σημασία με περισσότερη πληροφορία από την αρχική αναπαράσταση. Σαφές παράδειγμα μιας τέτοιας κατηγορίας είναι οι εικόνες, όπου η ανάλυση ενός συνόλου γειτονικών pixels μπορεί να καταλήξει σε ένα διαφορετικό αποτέλεσμα απ’ ότι το καθένα ξεχωριστά.

Μία συγκεκριμένη επεξεργασία που μπορούμε να υποβάλουμε σε τέτοιες ομάδες γειτονικών στοιχείων είναι η πράξη της συνέλιξης. Εφαρμόζοντας πολλά και διαφορετικά φίλτρα (kernels) σε αυτές τις ομάδες, προκύπτουν νέα δεδομένα τα οποία στη συνέχεια μπορούν να αναλυθούν εκ νέου. Στην περίπτωση των εικόνων που αναφέρθηκε προηγουμένως, εφαρμόζοντας ένα κατάλληλο φίλτρο μπορούμε να εντοπίσουμε τις ακμές που τις αποτελούν, κάτι που είναι χρήσιμο για παράδειγμα στην αναγνώριση αντικειμένων σε μία φωτογραφία, όπως φαίνεται στην Εικόνα 2.2.



Εικόνα 2.2: Παράδειγμα εντοπισμού ακμών σε μια φωτογραφία
(Uqam, CC BY-SA 4.0, via Wikimedia Commons)

Τα συνελκτικά νευρωνικά δίκτυα (Convolutional Neural Networks - CNN) είναι μια εφαρμογή που βασίζεται στην μεθοδολογία που περιγράφηκε προηγουμένως, τα οποία συνήθως αποτελούνται από (i) συνελκτικά επίπεδα που συνελίσσουν την είσοδο με ένα φίλτρο, (ii) επίπεδα pooling που εφαρμόζουν μία συγκεντρωτική πράξη στα δεδομένα που προέκυψαν από τη συνέλιξη ώστε να μειωθεί η διάστασή τους, όπως η επιλογή του μεγίστου ή της μέσης τιμής, και (iii) πλήρως συνδεδεμένα επίπεδα (dense) όπως αυτά που συναντώνται στα δίκτυα MLP. Οι δύο πρώτες κατηγορίες είναι τυπικά υπεύθυνες για την εξαγωγή χαρακτηριστικών από τα δεδομένα, ενώ η τελευταία είναι αυτή που κάνει την τελική κατηγοριοποίηση ή παλινδρόμηση για την πρόβλεψη του μοντέλου. Η οπτικοποίηση ενός τέτοιου δικτύου φαίνεται στην Εικόνα 2.3.



Εικόνα 2.3: Παράδειγμα αρχιτεκτονικής ενός συνελκτικού νευρωνικού δικτύου
(Aphex34, CC BY-SA 4.0, via Wikimedia Commons)

2.2.3. Recurrent Neural Network (RNN)

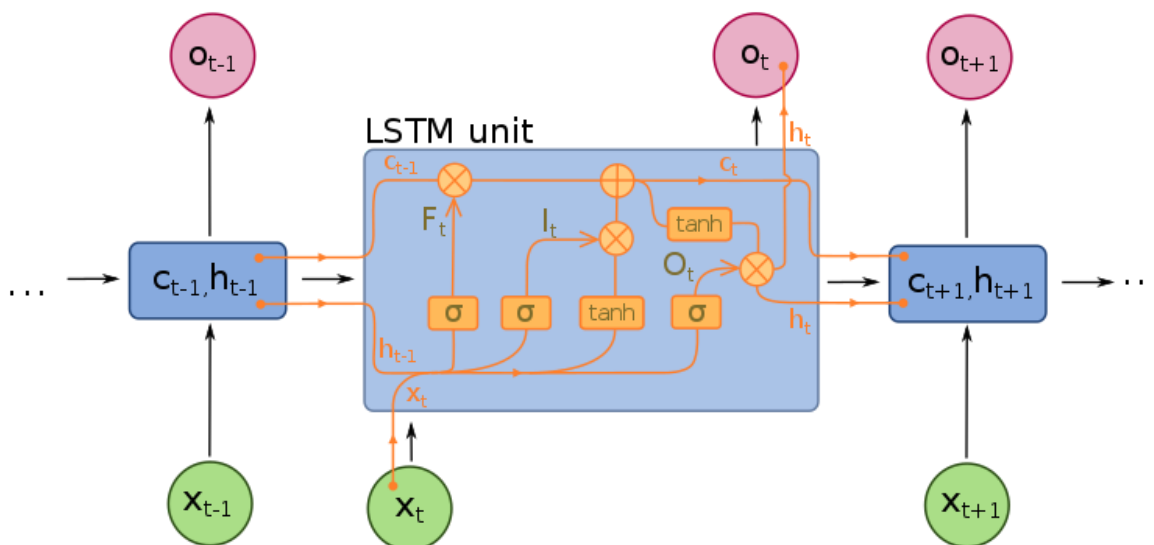
Επαναφέροντας τη διαδικασία της βιολογικής μάθησης στο προσκήνιο, παρατηρούμε πως στον πραγματικό κόσμο τα γεγονότα συχνά σχετίζονται μεταξύ τους, και αυτό βοηθάει στην αναγνώριση μοτίβων που βοηθούν στην κατηγοριοποίησή τους. Για αυτό το λόγο, οι αναμνήσεις και οι υπάρχουσες γνώσεις γενικότερα κατέχουν ένα βασικό ρόλο στον τρόπο με τον οποίο ένας άνθρωπος, για παράδειγμα, μαθαίνει κάτι καινούργιο. Όμως, η αρχιτεκτονική των MLP δεν διαθέτει κάποιο μηχανισμό που να εκμεταλλεύεται την πληροφορία που είναι κωδικοποιημένη στους νευρώνες ώστε να επεκτείνει τη νοημοσύνη του μοντέλου.

Ο Hopfield (1982) πρότεινε ένα τέτοιο μοντέλο στο οποίο, μεταξύ άλλων καινοτόμων στοιχείων, οι νευρώνες κάθε επιπέδου μπορούν να συνδέονται μεταξύ τους, κάτι που προσδίδει στο δίκτυο ένα μηχανισμό ανατροφοδότησης. Ο σχεδιασμός αυτός εντάσσεται στη γενικότερη κατηγορία των επαναλαμβανόμενων νευρωνικών δικτύων (Recurrent Neural Networks - RNNs), τα οποία χρησιμοποιούν εσωτερικές καταστάσεις “μνήμης” για να αξιοποιήσουν την υπάρχουσα γνώση του μοντέλου. Όταν ένας νευρώνας, λοιπόν, έχει πρόσβαση στο περιεχόμενο της μνήμης, δηλαδή της τιμής εξόδου του για μία διαφορετική είσοδο που έλαβε στο παρελθόν, τότε είναι ικανός να εξαγάγει πληροφορίες που σχετίζονται, παραδείγματος χάριν, με τη χρονική αλληλεξάρτηση των δεδομένων ή τα συμφραζόμενα ενός αποσπάσματος κειμένου.

2.2.4. Long Short-Term Memory (LSTM)

Ένα από τα ζητήματα που προέκυψε με τα RNN (Bengio et al. 1994) ήταν η δυσκολία αξιοποίησης της μνήμης όταν αυτή περιείχε στοιχεία που δεν ήταν απαραίτητα για την έξοδο του νευρώνα, ωστόσο έπρεπε να συμπεριληφθούν στους υπολογισμούς του κάθε βήματος. Σε αυτή την περίπτωση, το μοντέλο εμφάνιζε δυσκολία κατά την εκπαίδευση στην αφομοίωση των εννοιών και άρα μειωνόταν η αποδοτικότητά του. Για παράδειγμα, όταν σε μία πρόταση υπάρχουν πολλές συμπληρωματικές λέξεις που δεν συνεισφέρουν στην πρόβλεψη που καλείται να κάνει το μοντέλο, τότε αυτό θα δυσκολευτεί να αξιοποιήσει τα στοιχεία που πραγματικά χρησιμεύουν στην απόφαση.

Η λύση σε αυτό το πρόβλημα προτάθηκε από τους Hochreiter και Schmidhuber (1997) με την μορφή των δικτύων που βασίζονται στη μονάδα Long Short-Term Memory (LSTM) που επεκτείνει το “απλό” μοντέλο των RNN. Η κύρια διαφοροποίηση των συγκεκριμένων μοντέλων είναι ότι παρέχουν τη δυνατότητα στο δίκτυο να ξεχάσει κάτι που υπάρχει ήδη στη μνήμη, έτσι ώστε να απαλλαχθεί από τα πλεονάζοντα στοιχεία. Η εσωτερική δομή μιας τέτοιας μονάδας φαίνεται στην Εικόνα 2.4.



Εικόνα 2.4: Εσωτερική δομή ενός LSTM νευρώνα (fdeloche, CC BY-SA 4.0, via Wikimedia Commons)

3. Σχετικές Μελέτες

Μία από τις έρευνες σε αυτό το θέμα με σημαντικό πλήθος αναφορών είναι η μελέτη των McNally et al. (2018), στην οποία έδειξαν ότι με τη χρήση μοντέλων LSTM μπορούμε να πετύχουμε ακρίβεια ως και 52%, όταν το ζητούμενο είναι η πρόβλεψη της κατεύθυνσης κίνησης του Bitcoin, ενώ το μέσο σφάλμα RMSE (Root Mean Squared Error) ήταν της τάξης του 8% για την πρόβλεψη της επόμενης τιμής. Τα μοντέλα που δοκιμάστηκαν δεχόντουσαν ως είσοδο στοιχεία όπως την προηγούμενη τιμή του Bitcoin, σταθμισμένη με τη μέθοδο του κυλιόμενου μέσου, αλλά και ενδείξεις που συζητήθηκαν προηγουμένως όπως η υπολογιστική ισχύς του δικτύου (total hashrate) και η δυσκολία εύρεσης νέου block (difficulty).

Οι Mallqui και Fernandes (2019), χρησιμοποιώντας επιπλέον τις τιμές από άλλα οικονομικά προϊόντα ως χαρακτηριστικά (όπως το πετρέλαιο, τον χρυσό αλλά και ορισμένους δείκτες χρηματιστηριακών ιδρυμάτων), συνέθεσαν ένα σύνολο μοντέλων που απαρτίζονταν από πολλαπλές αρχιτεκτονικές βαθιών νευρωνικών δικτύων, μοντέλα συσταδοποίησης k-means, όπως και δενδρικά μοντέλα. Η μεθοδολογία αυτή κατάφερε να αυξήσει την ακρίβεια συγκριτικά με το προηγούμενο state-of-the-art μοντέλο των McNally κατά 10%.

Η μελέτη του Smuts (2019) εστίασε στο αν η πρόβλεψη της πορείας στην οποία θα κινηθούν κρυπτονομίσματα όπως το Bitcoin και το Ethereum (Wood 2014) μπορεί να

βασιστεί εν μέρει στη συναισθηματική ανάλυση των χρηστών που τα αναφέρουν σε μέσα κοινωνικής δικτύωσης αλλά και μηχανές αναζήτησης. Ως πηγή δεδομένων χρησιμοποιήθηκαν στοιχεία από το Telegram (<https://telegram.org>) και τη Google (<https://google.com>) αντίστοιχα, ενώ τα αποτελέσματα φαίνεται να είναι αρκετά ενθαρρυντικά για προβλέψεις σε βραχυπρόθεσμο χρονικό διάστημα.

Οι Li και Dai (2020) ακολούθησαν μια νέα προσέγγιση συνδυάζοντας συνελικτικά νευρωνικά δίκτυα (CNN) με μοντέλα LSTM ώστε να προβλέψουν την κατεύθυνση κίνησης του Bitcoin, πετυχαίνοντας precision της τάξης του 64% και recall του 81%. Ενώ χρησιμοποίησαν όλες τις μετρικές οι οποίες έχουν αναφερθεί προηγουμένως, τις εμπλούτισαν εφαρμόζοντας τους τεχνικούς δείκτες Relative Strength Index (RSI), Money Flow Index (MFI) και On Balance Volume (OBV) ώστε να βελτιώσουν στο μέγιστο τις δυνατότητες του μοντέλου.

Τέλος, οι Cavalli και Amoretti (2021) έδειξαν ότι η πρόβλεψη της κατεύθυνσης της αξίας του Bitcoin είναι εφικτή και με τη χρήση μονοδιάστατων συνελικτικών νευρωνικών δικτύων (CNNs). Τα δίκτυα αυτά σχεδιάστηκαν έτσι ώστε να λαμβάνουν ως είσοδο (i) τις προηγούμενες τιμές του Bitcoin μαζί με διάφορους τεχνικούς δείκτες που περιγράφουν ορισμένες τάσεις, (ii) πληροφορίες σχετικά με το ίδιο το blockchain, καθώς και (iii) συναισθηματική ανάλυση των δημοσιεύσεων στο Twitter (<https://twitter.com>) από χρήστες που συζητάνε για το Bitcoin. Η ακρίβεια που πέτυχαν ξεπερνάει το 74%, ενώ ιδιαίτερη σημασία δόθηκε στην αποδοτικότητα του μοντέλου όσον αφορά τις επενδύσεις, έτσι ώστε το κέρδος να είναι μεγαλύτερο σε περιόδους κατά τις οποίες η αγορά έχει θετικό πρόσημο (bullish market), ενώ η ζημία να είναι η μικρότερη δυνατή στην αντίθετη περίπτωση (bearish market).

4. Αλγόριθμος & Μεθοδολογία

4.1. Συλλογή δεδομένων

Ένα σημαντικό κομμάτι με το οποίο πρέπει να ασχοληθούμε αρχικά είναι η συλλογή των δεδομένων που θα χρησιμοποιηθούν για την εκπαίδευση των μοντέλων. Η διαδικασία αυτή χρήζει προσοχής καθώς ένα dataset με ανεπαρκή πληροφορία θα βλάψει την ικανότητα του δικτύου στην πρόβλεψη της κατεύθυνσης κίνησης. Τα δεδομένα που υπάρχουν σχετικά με το Bitcoin μπορούν να αντληθούν από διάφορες πηγές, από τις πιο προφανείς όπως το ίδιο το blockchain, ως και τις πιο σύνθετες όπως τα μέσα κοινωνικής δικτύωσης που φιλοξενούν τους επενδυτές οι οποίοι αλληλεπιδρούν και έμμεσα φανερώνουν τη σιγουριά που έχουν για την εξέλιξή του.

Καθώς το επίπεδο δυσκολίας της συλλογής και επεξεργασίας δεδομένων ποικίλλει ανάλογα με την πλατφόρμα στην οποία βρίσκονται, τίθεται το ζήτημα ως προς το ποιες πηγές αποτελούν τις σημαντικότερες ενδείξεις για την πορεία της τιμής και ποιες ενδέχεται να είναι θορυβώδεις. Έτσι, μπορούμε να αναλογιστούμε αν ο χρόνος και το κόστος απόκτησης των δεδομένων αξίζει το όφελος που θα έχουμε στην ισχύ του μοντέλου, ή αν μπορούμε να τα παραλείψουμε. Για παράδειγμα, η συλλογή δεδομένων από ένα ιδιωτικό API ή μια κλειστή ομάδα επενδυτών είναι δυσκολότερη από τη χρήση δημοσίων δεδομένων.

Με το παραπάνω σκεπτικό, καταλήξαμε στις εξής πηγές για αυτή την μελέτη:

4.1.1. Αξία Bitcoin και όγκος συναλλαγών ανά ημέρα

Ένα από τα πρώτα σύνολα χαρακτηριστικών που πρέπει να συμπεριληφθούν στα δεδομένα μας είναι η τιμή στην οποία συναλλάσσεται το Bitcoin ημερησίως, καθώς και ο όγκος που κυκλοφόρησε συνολικά στην αγορά. Για τα στοιχεία αυτά χρησιμοποιήθηκε ο ιστότοπος **Bitcoincharts** (<https://bitcoincharts.com>), από τον οποίο εξήχθησαν οι τιμές Open, Close, High και Low για κάθε μέρα, αλλά και ο όγκος σε USD και BTC. Οι ημερομηνίες για τις οποίες συλλέχθηκαν δεδομένα είναι μεταξύ 18 Δεκεμβρίου 2011 και 7 Μαΐου 2021, οι οποίες θα χρησιμοποιηθούν και για τις υπόλοιπες πηγές ώστε να υπάρχει πληρότητα και ομοιομορφία.

4.1.2. Google Trends

Ένα δημοφιλές μέσο, στο οποίο η πρόσβαση είναι σχετικά εύκολη, είναι οι τάσεις που καταγράφονται από την Google για κάθε όρο που χρησιμοποιείται μέσω της μηχανής αναζήτησής της. Ενώ τα δεδομένα αυτά είναι ανοιχτά προς όλους μέσω της ιστοσελίδας τους (<https://trends.google.com/trends>), δεν υπάρχει κάποιο διαθέσιμο API ή ένας τρόπος ώστε να αποθηκευτούν ως CSV σε, και γι' αυτό χρησιμοποιήθηκε η ανεπίσημη βιβλιοθήκη της Python **pytrends** (<https://pypi.org/project/pytrends>). Με αυτόν τον τρόπο, συλλέχθηκαν ημερήσια δεδομένα για το ποσοστό των αναζητούσαν που αφορούν τον όρο "Bitcoin".

4.1.3. Δείκτης Standard and Poor's 500

Όπως αναφέρθηκε προηγουμένως, ιδιαίτερο ενδιαφέρον έχει η εξερεύνηση της ιδέας πως η τιμή των κρυπτονομισμάτων ενδέχεται να κινείται κατά κάποιο βαθμό μαζί με αυτή μεγάλων τεχνολογικών μετοχών και άλλους χρηματιστηριακούς δείκτες. Για το λόγο αυτό, συμπεριλήφθηκε στην έρευνα ο δείκτης S&P 500 από τον ιστότοπο **Stooq** (<https://stooq.com>). Από τον δείκτη αυτό χρησιμοποιήθηκε μόνο η τιμή κλεισίματος κάθε μέρα (S&P 500 Close) για λόγους απλότητας.

4.1.4. Ισοτιμίες νομισμάτων

Ομοίως με την προηγούμενη κατηγορία δεδομένων, μελετήθηκε η συσχέτιση της εξέλιξης της τιμής διαφόρων νομισμάτων που χρησιμοποιούνται ανά τον κόσμο, και συγκεκριμένα η ισοτιμία του αμερικανικού δολαρίου, συνάρτηση του οποίου εκφράζονται τα υπόλοιπα δεδομένα, με το ευρώ (USD/EUR), την αγγλική λίρα (USD/GBP), το κινεζικό γουάν (USD/CNY) και το ελβετικό φράγκο (USD/CHF). Ως πηγή για τα δεδομένα των ισοτιμιών χρησιμοποιήθηκε ο ιστότοπος **Investing.com** (<https://www.investing.com/currencies>).

4.1.5. Στατιστικά του Bitcoin blockchain

Καθώς τα δεδομένα του blockchain είναι ανοιχτά και προσβάσιμα από όλους, η χρήση τους είναι μία από τις πρώτες σκέψεις στην επιλογή χαρακτηριστικών. Για τον σκοπό αυτό, χρησιμοποιήθηκαν κάποια από τα δεδομένα που παρέχει το ανταλλακτήριο **Blockchain.com** (<https://blockchain.com/charts>). Αναλυτικότερα, τα στοιχεία τα οποία αντλήθηκαν φαίνονται στον *Πίνακα 4.1*.

4.2. Προεπεξεργασία δεδομένων

4.2.1. Συμπλήρωση ελλειπών τιμών

Αφού συλλεχθούν τα δεδομένα που αναφέρθηκαν προηγουμένως, το πρώτο βήμα προεπεξεργασίας που ακολουθεί είναι η συμπλήρωση των ελλειπών τιμών που οφείλονται είτε στη φύση των δεδομένων είτε σε κάποιο άγνωστο τεχνικό σφάλμα της πηγής από την οποία προήλθαν. Στην πρώτη κατηγορία εντάσσονται οι τιμές των χρηματιστηριακών δεικτών και των νομισματικών ισοτιμιών, οι οποίες δεν ορίζονται για τις ημέρες που αντιστοιχούν σε Σάββατο ή Κυριακή, ενώ στη δεύτερη και για το συγκεκριμένο dataset που συλλέχθηκε ήταν ο όγκος συναλλαγών του Bitcoin σε USD και το πλήθος των μοναδικών διευθύνσεων που καταγράφησαν για κάθε ημέρα.

4.2.2. Επιλογή χαρακτηριστικών

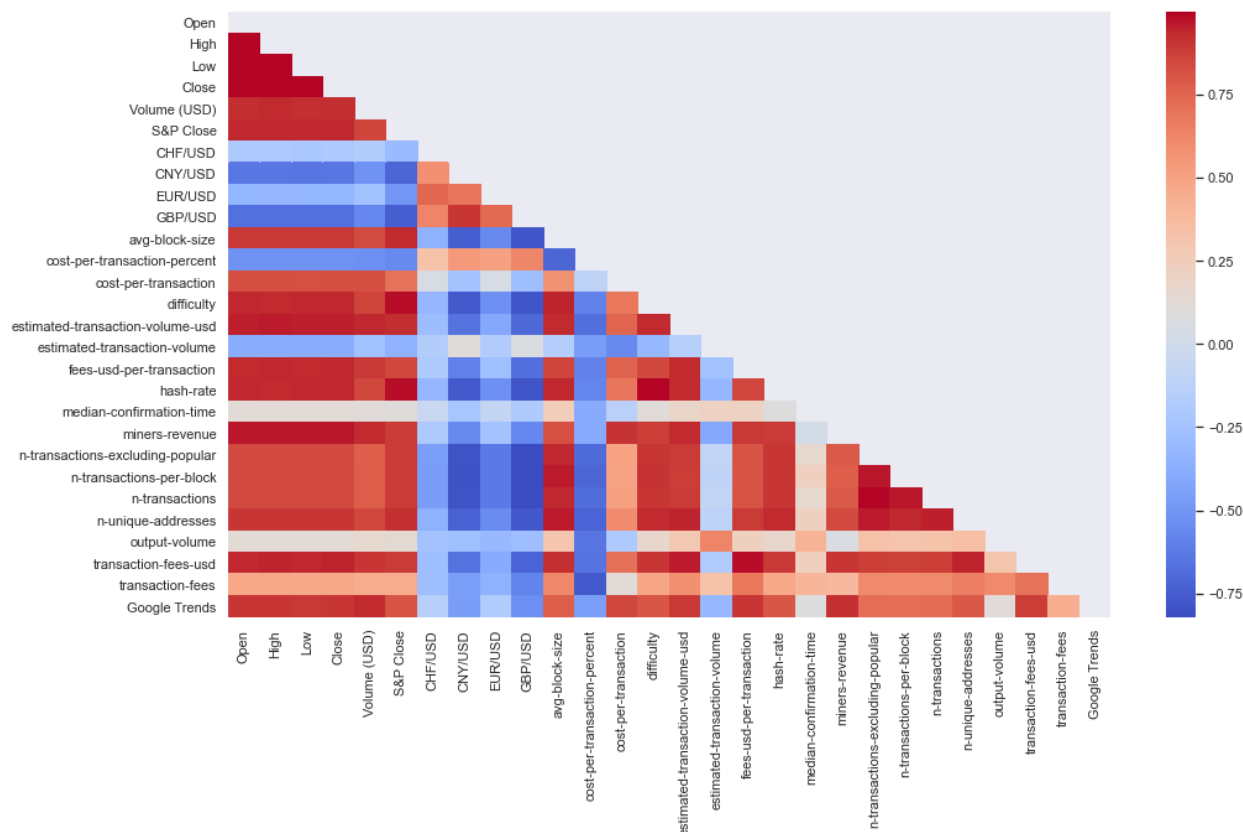
Μέχρι αυτό το σημείο, έχουμε δημιουργήσει ένα σύνολο δεδομένων που αποτελείται από διάφορα χαρακτηριστικά, τα οποία ενδέχεται να μην είναι όλα χρήσιμα για το πρόβλημα που θέλουμε να επιλύσουμε. Για να μειώσουμε τις διαστάσεις των δεδομένων και να διατηρήσουμε μόνο όσες μεταβλητές θα βοηθήσουν στην πρόβλεψη της κατεύθυνσης της τιμής, χρησιμοποιήθηκε ο συντελεστής συσχέτισης του Spearman. Σε αυτή τη μετρική, ελέγχεται η εξάρτηση δύο μεταβλητών και βαθμολογείται στο διάστημα $[-1, 1]$, με το -1 να υποδεικνύει μια αντιστρόφως ανάλογη σχέση, το $+1$ μια ανάλογη σχέση, ενώ το 0 δείχνει μηδενική συσχέτιση των μεταβλητών.

Βαθμολογώντας όλα τα χαρακτηριστικά ως προς τη συσχέτισή τους με την τιμή κλεισίματος του Bitcoin για κάθε ημέρα, προκύπτουν κατ' απόλυτη τιμή οι βαθμοί συσχέτισης που φαίνονται στον Πίνακα 4.1, ενώ στην Εικόνα 4.1 παρουσιάζεται η συσχέτιση όλων των μεταβλητών ως ένας χάρτης θερμότητας (heatmap).

Με βάση αυτή τη μετρική, επιλέχθηκαν τα χαρακτηριστικά τα οποία είχαν βαθμό συσχέτισης κατ' απόλυτη τιμή μεταξύ του 0.6 και του 0.95, έτσι ώστε να συμπεριληφθούν μόνο όσα έχουν πιθανότητα να συμβάλουν στην πρόβλεψη κατά έναν βαθμό, αλλά όχι να κυριαρχήσουν απολύτως, υπονομεύοντας τα υπόλοιπα. Αυτά τα χαρακτηριστικά εμφανίζονται με έντονη γραμματοσειρά στην κατάταξη.

Χαρακτηριστικό	Συσχέτιση Spearman με το "BTC Close" κατ' απόλυτη τιμή	Χαρακτηριστικό	Συσχέτιση Spearman με το "BTC Close" κατ' απόλυτη τιμή
BTC Close	1.000000	avg-block-size	0.896855
BTC High	0.999642	n-tx	0.848586
BTC Open	0.999353	n-tx-incl-popular	0.845546
BTC Low	0.999205	n-tx-per-block	0.845355
miners-revenue	0.968044	cost-per-tx	0.831109
est-tx-volume-usd	0.956597	GBP/USD	0.668510
tx-fees-usd	0.942834	CNY/USD	0.642996
S&P 500 Close	0.938270	cost-per-tx-percent	0.511847
network-difficulty	0.936627	tx-fees	0.479493
network-hashrate	0.936136	estimated-tx-volume	0.379336
fees-usd-per-tx	0.935825	EUR/USD	0.325682
BTC Volume (USD)	0.925129	CHF/USD	0.200396
Google Trends	0.902132	output-volume	0.129563
n-unique-addresses	0.901492	median-conf-time	0.120142

Πίνακας 4.1: Συσχέτιση Spearman κατ' απόλυτη τιμή με το χαρακτηριστικό "BTC Close"



Εικόνα 4.1: Heatmap της συσχέτισης Spearman μεταξύ όλων των μεταβλητών

4.2.3. Υπολογισμός τεχνικών δεικτών

Σε μία προσπάθεια να ενισχυθεί το σύνολο δεδομένων με κάποια πιο σύνθετα στοιχεία που συναντώνται σε περισσότερο προηγμένα περιβάλλοντα επενδύσεων, υπολογίστηκαν ορισμένοι τεχνικοί δείκτες (technical indicators) που σύμφωνα με τη σχετική βιβλιογραφία έχουν ιδιαίτερη σημασία για την πρόβλεψη της πορείας ενός χρηματοοικονομικού στοιχείου, ακόμη και στην περίπτωση του Bitcoin.

Έτσι, χρησιμοποιώντας τη βιβλιοθήκη **pandas-ta** (<https://pypi.org/project/pandas-ta>), η οποία έχει τη δυνατότητα να επεκτείνει ένα πίνακα δεδομένων που έχει στήλες Open/Close/High/Low με ένα πλήθος διαφορετικών δεικτών, παρήχθησαν τα εξής:

1. Bitcoin 10-day Simple Moving Average (SMA-10)
2. Bitcoin 2-week Commodity Channel Index (CCI-14)
3. Bitcoin 2-week Relative Strength Index (RSI-14)

Οι παραπάνω δείκτες συμπεριλήφθηκαν μαζί με τα υπόλοιπα χαρακτηριστικά που προέκυψαν από τη συσχέτιση Spearman και έτσι σχηματίστηκε το τελικό σύνολο χαρακτηριστικών βάση του οποίου θα γίνονται οι προβλέψεις των μοντέλων.

4.2.4. Υπολογισμός τιμής - στόχου

Το πρόβλημα που επιδιώκουμε να μελετήσουμε στην τρέχουσα έρευνα είναι η κατηγοριοποίηση της κατεύθυνσης προς την οποία αναμένεται να κινηθεί η τιμή του Bitcoin, δηλαδή αν αυτή θα είναι είτε ανοδική είτε καθοδική, και όχι η πρόβλεψη της επόμενης τιμής. Γι' αυτό το λόγο, προτού σχεδιάσουμε τα μοντέλα, πρέπει να μετατρέψουμε το χαρακτηριστικό που έχουμε, δηλαδή το "BTC Close", σε ένα νέο χαρακτηριστικό το οποίο αποτυπώνει τις δύο κλάσεις που επιθυμούμε.

Ο τρόπος που γίνεται αυτό είναι πολύ απλός, καθώς το μόνο που χρειάζεται είναι μια διαδοχική σύγκριση όλων των τιμών του αρχικού χαρακτηριστικού, ώστε να συμπεράνουμε αν η τιμή αυξάνεται την επόμενη μέρα ή όχι. Στην περίπτωση, λοιπόν, που η επόμενη τιμή "BTC Close" είναι μεγαλύτερη από την τρέχουσα, τότε η κλάση του συγκεκριμένου παραδείγματος είναι θετική και έχει τιμή + 1 (δηλαδή ανοδική), ενώ σε αντίθετη περίπτωση είναι αρνητική και έχει τιμή 0 (δηλαδή καθοδική).

4.2.5. Κλιμάκωση χαρακτηριστικών

Ένα επίσης σημαντικό βήμα στην προεπεξεργασία των δεδομένων μας είναι η κλιμάκωση των χαρακτηριστικών έτσι ώστε οι αλγόριθμοι που χρησιμοποιούνται από τα νευρωνικά δίκτυα για τη βελτιστοποίησή τους να μην έχουν θέματα με τη σύγκλισή τους λόγω διαφορετικών εκτάσεων στις τιμές.

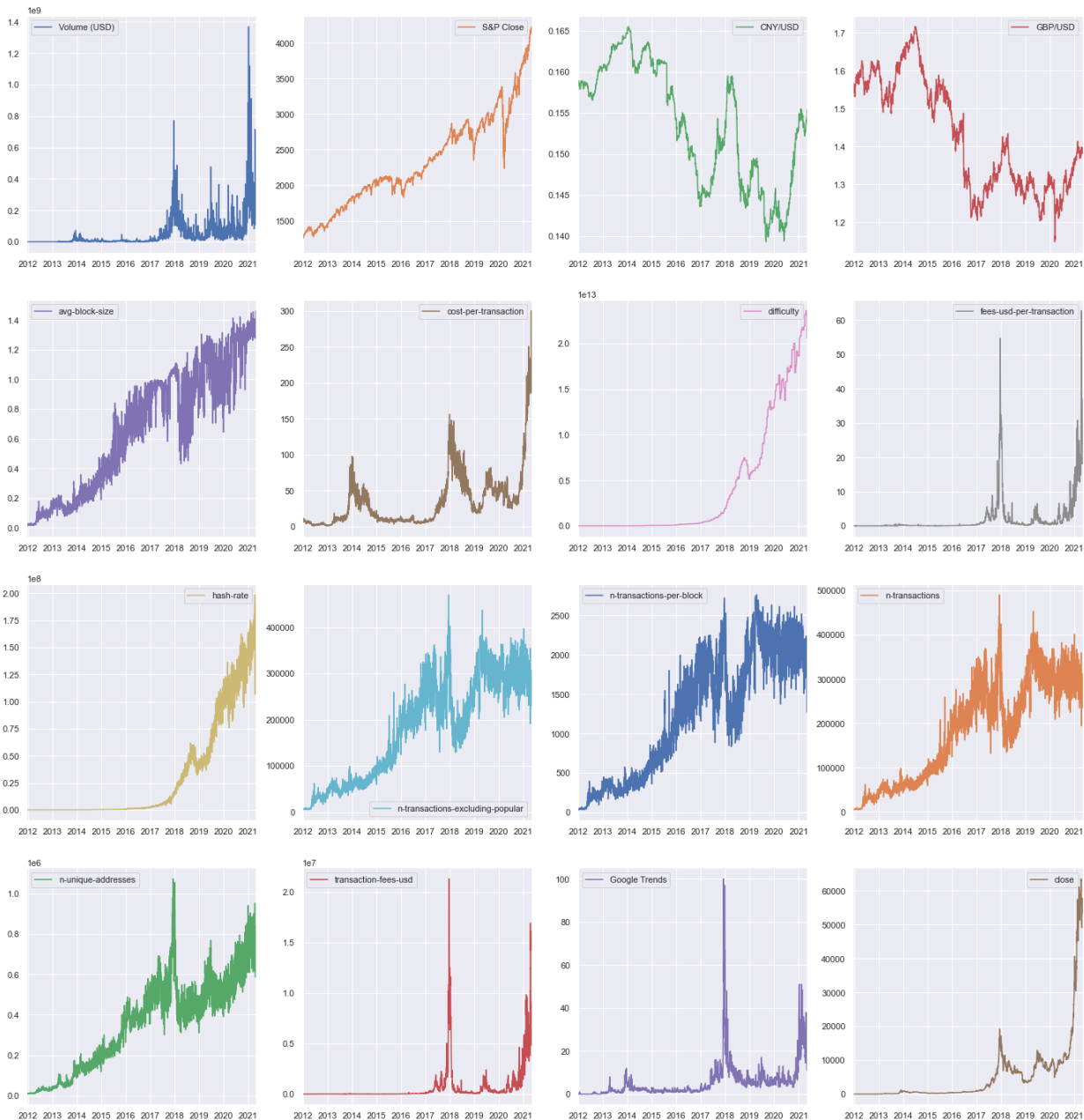
Όπως φαίνεται και στην Εικόνα 4.2, ορισμένα χαρακτηριστικά είναι πολλαπλές τάξεις μεγέθους μεγαλύτερα από τα υπόλοιπα. Για αυτό τον λόγο, το πρώτο βήμα στην κλιμάκωσή τους είναι να υπολογίσουμε τη λογαριθμική τιμή τους, μειώνοντας το εύρος στο οποίο κυμαίνονται. Κατόπιν, μπορούμε να εφαρμόσουμε σε όλα τα χαρακτηριστικά εισόδου την τεχνική της τυποποίησης (standardization), έτσι ώστε όλα να έχουν μέση τιμή το μηδέν (0) και τυπική απόκλιση τη μονάδα (1).

4.2.6. Χωρισμός σε ακολουθίες

Αφού σκοπεύουμε να χρησιμοποιήσουμε μοντέλα τα οποία δέχονται ακολουθίες δεδομένων και προσπαθούν να μάθουν με τη χρήση εσωτερικών μνημών, πρέπει επίσης να χωρίσουμε το dataset σε ακολουθίες (sequences). Για να γίνει αυτό, αρκεί ενώ το διατρέχουμε να επιλέγουμε κάθε φορά τις N προηγούμενες ημέρες και να δημιουργούμε μια ομάδα παραδειγμάτων που πλέον θα αντιστοιχούν σε μία έξοδο, αυτή που είχε η τελευταία ημέρα της ομάδας.

Σύμφωνα με τη βιβλιογραφία, για προβλήματα classification στο Bitcoin χρειάζεται το μέγεθος των ακολουθιών να είναι σχετικά μεγάλο, και γι' αυτό επιλέχθηκε η τιμή

$N = 100$. Αντιθέτως, εάν θέλαμε να προβλέψουμε αυτούσια την επόμενη τιμή που θα έχει το Bitcoin, τότε ενδείκνυται να χρησιμοποιηθεί μια μικρή τιμή όπως $N = 10$. Ο λόγος γι' αυτό είναι ότι όταν μας ενδιαφέρει η κατεύθυνση κίνησης, έχει μεγαλύτερη σημασία να υπάρχουν αρκετά ιστορικά στοιχεία διαθέσιμα στη μνήμη, για παράδειγμα των προηγούμενων 2-3 μηνών, ενώ στην παλινδρόμηση θέλουμε να εκμεταλλευτούμε περισσότερο τις αλλαγές που υπήρξαν στο πρόσφατο παρελθόν της τιμής και άρα αρκούν τα στοιχεία από μία ή δύο εβδομάδες.



Εικόνα 4.2: Γραφική αναπαράσταση των χρονοσειρών (χωρίς κλιμάκωση)

4.3. Πρόβλεψη κατεύθυνσης κίνησης για την αξία του Bitcoin

Έχοντας ως σκοπό να συγκρίνουμε την αποδοτικότητα των διάφορων μοντέλων βαθιάς μάθησης στην πρόβλεψη της κατεύθυνσης κίνησης του Bitcoin, θα μελετήσουμε την ακρίβεια που πετυχαίνουν τα διαφορετικά μοντέλα που έχουν αναφερθεί προηγουμένως. Ωστόσο, είναι σημαντικό να έχουμε υπόψη δύο παράγοντες όταν ερμηνεύουμε τα αποτελέσματα:

- Χρηματιστήριο και φούσκες (bubbles)
Καθώς το Bitcoin θεωρείται από πολλούς όμοιο με αρκετά χρηματοοικονομικά περιουσιακά στοιχεία, είναι πιθανό κατά περιόδους να βρίσκεται σε μια κατάσταση “φούσκας” (bubble), κατά την οποία η τιμή του διαρκώς αυξάνεται, ενδεχομένως πολύ περισσότερο από την πραγματική αξία που θα ανέμενε κάποιος να έχει με βάση τα στοιχεία που υπάρχουν. Η διάρκεια ενός τέτοιου φαινομένου δεν μπορεί να είναι απεριόριστη και συνήθως όταν ξεκινήσει η ανατροπή, ή αλλιώς το “ξεφούσκωμα”, τότε η αξία θα υποβαθμιστεί πολύ πιο γρήγορα από ότι θα γινόταν υπό κανονικές συνθήκες.
- Ανατιμητικές και υποτιμητικές αγορές (bull and bear markets)
Ανάλογα με διάφορα γεγονότα στην παγκόσμια οικονομία και πολιτική, οι χρηματιστηριακές αγορές παρουσιάζουν περιόδους όπου γενικώς οι τιμές των μετοχών και λοιπών στοιχείων έχουν την τάση να μειώνονται μαζικά (bear markets), ή αντιστρόφως να αυξάνονται (bull markets). Τέτοια γεγονότα μπορεί να είναι η κρίση που επιφέρει μια παγκόσμια πανδημία η οποία δυσχεραίνει τις αγορές, ή από την άλλη η οικονομική ενίσχυση των πολιτών με επιδόματα για την ανάκαμψη της οικονομίας και του εμπορίου. Αυτά τα δύο είδη αγορών εναλλάσσονται συχνά μεταξύ τους, αλλά δεν είναι πάντα εύκολο να προβλέψει κανείς πότε θα προκύψει η επόμενη ανατροπή, λόγω των πολλαπλών παραγόντων που τα καθορίζουν.

Συγκεκριμένα, ο αντίκτυπος των δύο παραπάνω φαινομένων στην αγορά των κρυπτονομισμάτων μπορεί να φανεί στην Εικόνα 4.3 στη χρονική περίοδο του Μαρτίου 2020, όπου είναι εμφανής η αρνητική επίδραση στην παγκόσμια οικονομία, και κατ’ επέκταση στην αξία του Bitcoin, όταν η πλειοψηφία του πλανήτη βρέθηκε σε lockdown ως μέτρο πρόληψης έναντι του κορονοϊού SARS-CoV-2. Αντιθέτως, από τα μέσα του Δεκεμβρίου 2020 μέχρι τα μέσα του Μαΐου 2021, παρατηρούμε ότι η τιμή του Bitcoin τριπλασιάστηκε, αγγίζοντας τα \$60,000 από τα \$20,000 που ήταν στην αρχή της περιόδου.



Εικόνα 4.3: Εξέλιξη τιμής του Bitcoin για το διάστημα Ιανουάριος 2020 - Μάιος 2021
(Bitcoincharts.com, CC BY-SA 3.0)

Όλοι οι παράγοντες που αναφέρθηκαν παραπάνω δείχνουν ότι οι αγορές όπως αυτές των κρυπτονομισμάτων χαρακτηρίζονται ως εξαιρετικά πτητικές (volatile), με αυξομειώσεις της τάξης των 10% να μην είναι απίθανες σε περιόδους λίγων μόνο ημερών. Όταν όμως συμβαίνει κάτι τέτοιο, οποιοδήποτε επενδυτικό μοντέλο έρχεται σε πολύ δύσκολη θέση καθώς βρίσκεται σε ασταθή κατάσταση και τα σήματα που λαμβάνει ως είσοδο δεν μπορούν να χρησιμοποιηθούν με την ίδια αξιοπιστία σε σύγκριση με αυτά που δέχεται σε περιόδους “κανονικής λειτουργίας”.

Για τον παραπάνω λόγο, δημιουργήθηκε ένα μικρότερο σύνολο δεδομένων από το αρχικό, για να χρησιμοποιηθεί και να εκτιμηθεί ξεχωριστά η ακρίβεια των μοντέλων, το οποίο δε λαμβάνει υπόψη τα δεδομένα από το 2021. Έτσι, μπορούμε να έχουμε μια εκτίμηση για το πόσο επηρεάζονται οι αλγόριθμοι βαθιάς μάθησης από τα φαινόμενα αστάθειας που προαναφέρθηκαν, και επομένως πόσο συνετό είναι να εφαρμόζονται στην πράξη για την λήψη επενδυτικών αποφάσεων με βάση τις προβλέψεις τους σε τέτοιες περιόδους.

Χωρίζοντας τα δύο, πλέον, σύνολα δεδομένων σε υποσύνολα εκπαίδευσης και ελέγχου, καταλήγουμε στα ποσοστά θετικότητας που φαίνονται στον Πίνακα 4.2. Ο διαχωρισμός έγινε έτσι ώστε το 80% των δεδομένων να βρίσκεται στο σύνολο εκπαίδευσης και το 20% στο σύνολο ελέγχου. Η ερμηνεία των ποσοστών κάθε κλάσης είναι ιδιαίτερα σημαντική, καθώς αν βασιστούμε σε μία απλή επενδυτική στρατηγική που αναμένει την αύξηση της αξίας κάθε μέρα, η ακρίβεια θα ήταν όσο και η θετικότητα στο κάθε σύνολο. Στη συνέχεια, για ευκολία σύγκρισης, θα γίνεται αναφορά σε αυτή τη στρατηγική ως “θετικό επενδυτή”.

Σύνολο δεδομένων	Σύνολο εκπαίδευσης (θετική κλάση)	Σύνολο ελέγχου (θετική κλάση)
Πλήρες (3429 τιμές)	54.50%	54.34%
Μειωμένο (3224 τιμές)	54.13%	52.93%

Πίνακας 4.2: Ποσοστά θετικότητας σε κάθε υποσύνολο δεδομένων

Για την εκπαίδευση των παρακάτω μοντέλων βαθιάς μάθησης, χρησιμοποιήθηκαν οι βιβλιοθήκες scikit-learn (<https://scikit-learn.org>) (Pedregosa et al. 2011) και Keras (<https://keras.io>) (Chollet 2015), ενώ η εκπαίδευση έγινε με έναν επεξεργαστή Intel i7 7700k @ 4.20 GHz και μία κάρτα γραφικών NVIDIA GTX 1080 8 GB.

Ο τρόπος με τον οποίο χωρίζεται το σύνολο δεδομένων σε υποσύνολα εκπαίδευσης και επαλήθευσης κατά το Cross Validation είναι με τη χρήση της κλάσης [TimeSeriesSplit](#) του sklearn, που χωρίζει το σύνολο εκπαίδευσης σε $n = 5$ υποσύνολα και διαδοχικά εκπαιδεύει το μοντέλο με τα $k = 1, 2, 3, 4$ πρώτα σύνολα και το επαληθεύει με το αμέσως επόμενο. Έτσι επιλέγεται ποια αρχιτεκτονική από κάθε είδος δικτύου είναι η καλύτερη για να χρησιμοποιηθεί και να συγκριθεί με τους υπόλοιπους τύπους μοντέλων.

4.3.1. Support Vector Machine (SVM)

Ως βασικό μέτρο σύγκρισης (baseline model), εκπαιδεύτηκε ένα μοντέλο SVM.

Στο μειωμένο σύνολο δεδομένων, το μοντέλο δεν κατάφερε να ξεπεράσει την απόδοση του “θετικού επενδυτή”, τόσο κατά την εκπαίδευση όσο και στον έλεγχο με το σύνολο επαλήθευσης, σημειώνοντας ακρίβεια 53.46% και 52.93% αντίστοιχα. Στο πλήρες σύνολο δεδομένων, αν και κατά τη διαδικασία του Cross Validation το μοντέλο φαινόταν να αποδίδει καλύτερα από το ποσοστό θετικότητας του συνόλου εκπαίδευσης με ακρίβεια 54.9%, στην επαλήθευση βρέθηκε οριακά κάτω από αυτό έχοντας ακρίβεια 53.15%.

4.3.2. Deep Neural Network (DNN)

Στο μειωμένο σύνολο δεδομένων, η μέση ακρίβεια ενός μοντέλου με αρχιτεκτονική MLP κατά την επαλήθευση είναι 52.93%, που είναι ανάλογη με αυτή του “θετικού επενδυτή”, ενώ στο πλήρες σύνολο έχουμε ακρίβεια της τάξης του 54.85%. Ωστόσο, μέχρι αυτό το σημείο δεν έχουμε εκμεταλλευτεί ακόμη τη χρονική συσχέτιση των δεδομένων, καθώς το δίκτυο MLP ισοπεδώνει τις ακολουθίες στην είσοδό του, χάνοντας έτσι το στοιχείο της χρονικής διαφοράς.

4.3.3. Long Short-Term Memory (LSTM)

Αν εφαρμόσουμε έναν αλγόριθμο με καταστάσεις μνήμης, αναμένουμε το μοντέλο να μπορέσει να ανακαλύψει τα χρονικά μοτίβα και να εξάγει την “κρυφή” γνώση που εμπεριέχεται σε αυτά τα δεδομένα για να βελτιωθεί η ακρίβειά του. Εκπαιδεύοντας ένα δίκτυο με μονάδες LSTM, φαίνεται πως η ακρίβεια που πετυχαίνει είναι 55.31% για το μειωμένο σύνολο δεδομένων, και 55.36% για το πλήρες.

4.3.4. Συνδυασμός CNN - LSTM

Επιλέγοντας μια σύνθετη αρχιτεκτονική, όπου τα πρώτα επίπεδα είναι συνελκτικά, τα οποία ενεργούν ως μονάδες εξαγωγής χαρακτηριστικών (feature extraction), και στη συνέχεια η έξοδός τους τροφοδοτείται σε μονάδες LSTM, αυξάνουμε την πολυπλοκότητα του δικτύου έτσι ώστε να μπορέσει να μάθει πιο σύνθετες έννοιες από τη χρονοσειρά των δεδομένων και να αυξηθεί η απόδοσή του. Η τελική ακρίβεια ενός τέτοιου μοντέλου είναι 54.76% στο μειωμένο σύνολο και 55.53% στο πλήρες.

4.4. Έκθεση αποτελεσμάτων

Παρακάτω φαίνονται οι μετρικές αξιολόγησης από την εκπαίδευση των παραπάνω μοντέλων βαθιάς μάθησης (Εικόνα 4.4), τα αποτελέσματα της αναζήτησης των υπερπαραμέτρων (Πίνακας 4.4, Πίνακας 4.5) μέσω Grid Search, καθώς και οι συγκρίσεις των τελικών αποδόσεων στο σύνολο ελέγχου (Πίνακας 4.3).

Από τα αποτελέσματα, φαίνεται πως η αρχιτεκτονική **LSTM** που σχεδιάστηκε ήταν η πιο αποτελεσματική για τη διεργασία που ορίσαμε ως στόχο, πετυχαίνοντας ακρίβεια μεγαλύτερη του 55% και στα δύο σύνολα δεδομένων που εξετάστηκαν.

Αρχιτεκτονική	Μειωμένο σύνολο δεδομένων		Πλήρες σύνολο δεδομένων	
	Σύνολο Επαλήθευσης	Σύνολο Ελέγχου	Σύνολο Επαλήθευσης	Σύνολο Ελέγχου
SVM	53.46%	52.93%	54.90%	53.15%
DNN	53.80%	52.93%	54.40%	54.85%
LSTM	52.49%	55.31%	54.50%	55.36%
CNN - LSTM	53.02%	54.76%	54.72%	55.53%

Πίνακας 4.3: Ακρίβεια μοντέλων στην πρόβλεψη της κατεύθυνσης κίνησης του Bitcoin



(α) Ακρίβεια εκπαίδευσης DNN



(β) Ακρίβεια εκπαίδευσης LSTM

Εικόνα 4.4: Μετρικές αξιολόγησης των μοντέλων

Νευρώνες / Επίπεδα	Ποσοστό dropout	Χρόνος εκπαίδευσης	Ακρίβεια (val)
4 - 8 - 4	0.4	2.27s	54.40%
4 - 8 - 4	0.2	2.49s	54.18%
8 - 16 - 16 - 8	0.4	2.82s	54.18%
4 - 8 - 4	0.1	2.31s	53.90%

Πίνακας 4.4: Αποτελέσματα Grid Search για το μοντέλο DNN (πλήρες dataset)

Νευρώνες / Επίπεδα	Ποσοστό dropout	Χρόνος εκπαίδευσης	Ακρίβεια (val)
8 - 16 - 8	0.2	35.71s	54.50%
4 - 8 - 4	0.6	35.40s	54.13%
4 - 4	0.6	23.48s	53.40%
4 - 4	0.4	26.40s	53.00%

Πίνακας 4.5: Αποτελέσματα Grid Search για το μοντέλο LSTM (πλήρες dataset)

5. Συμπεράσματα & Μελλοντικές Επεκτάσεις

Από τα παραπάνω αποτελέσματα, είναι εμφανές ότι το πρόβλημα προς επίλυση δεν είναι εύκολο και αυτό οφείλεται εν μέρει στην μεγάλη αβεβαιότητα που κυριαρχεί στις αγορές των κρυπτονομισμάτων. Όπως δείχνει και η πρόσφατη βιβλιογραφία, τα ιστορικά δεδομένα του blockchain από μόνα τους δεν αρκούν για να αιτιολογήσουν και πόσο μάλλον να προβλέψουν με σιγουριά την τάση που θα ακολουθήσει.

Πράγματι, αν αναλογιστούμε τους παράγοντες που οδήγησαν στην εκρηκτική αύξηση της τιμής από την αρχή του 2021 μέχρι και το Μάιο, καταλήγουμε στο γεγονός ότι η μεγαλύτερη ώθηση πιθανόν να προέρχεται από τον ενθουσιασμό που δημιουργήθηκε γύρω από τα κρυπτονομίσματα. Για αυτό το λόγο, η έρευνα τους τελευταίους μήνες φαίνεται να στρέφεται προς την ανάλυση συναισθημάτων σε κοινωνικές πλατφόρμες και ειδησεογραφικές σελίδες που καλύπτουν την επικαιρότητα, έτσι ώστε να ερμηνεύσουν αν το κλίμα που επικρατεί ευνοεί ή δυσχεραίνει την ανάπτυξη χρηματοοικονομικών στοιχείων όπως το Bitcoin.

Επομένως, μία επέκταση της παρούσας μελέτης θα μπορούσε να συμπεριλάβει ροές δεδομένων από μέσα δικτύωσης και ανταλλαγής μηνυμάτων, από τις οποίες θα εξαγεται πληροφορία σχετικά με την εμπιστοσύνη του ευρύτερου κοινού για την πορεία των κρυπτονομισμάτων. Η δυσκολία σε αυτό το κομμάτι έγκειται αφενός στον τρόπο με τον οποίο ο αλγόριθμος θα αποκτήσει πρόσβαση σε αυτά τα στοιχεία, πιθανώς μέσω κάποιου περιορισμένου API, και αφετέρου στην αποτελεσματική ανάλυση των δεδομένων για την εξαγωγή των επιθυμητών χαρακτηριστικών.

Παράλληλα, ένα σημείο που χρήζει περαιτέρω διερεύνησης είναι η ανάδειξη των τεχνικών δεικτών των οποίων η εφαρμογή στα ιστορικά δεδομένα του Bitcoin παράγει χρήσιμα δεδομένα για τα μοντέλα μηχανικής μάθησης, καθώς δεν υπάρχει σχετική έρευνα που να καλύπτει σε βάθος το συγκεκριμένο θέμα, και ως αποτέλεσμα οι δείκτες που χρησιμοποιούνται ποικίλλουν χωρίς κάποιο κοινώς αποδεκτό σύνολο.

Τέλος, ένα κομμάτι το οποίο δε θίχτηκε από τη συγκεκριμένη εργασία είναι η πρόβλεψη της αξίας του Bitcoin για την επόμενη ημέρα. Κάτι τέτοιο ενδεχομένως να είναι δυσκολότερο από την κατηγοριοποίηση δύο κλάσεων, αλλά σε περίπτωση που βρεθεί μια ικανοποιητική λύση, θα μπορούσε να χρησιμοποιηθεί ως παράμετρος για τη βελτίωση της ακρίβειας του τρέχοντος προβλήματος, σε ένα ensemble μοντέλων.

Αναφορές

1. Bengio, Yoshua, et al. "Learning long-term dependencies with gradient descent is difficult." *IEEE transactions on neural networks*, vol. 5, no. 2, 1994, 157–166.
2. Bollinger, John. "Using bollinger bands." *Stocks & Commodities*, vol. 10, no. 2, 1992, 47–51.
3. Cavalli, Stefano, and Michele Amoretti. "CNN-based multivariate data analysis for bitcoin trend prediction." *Applied Soft Computing*, vol. 101, 2021, p. 107065.
4. Chollet, François. *Keras*. 2015, <https://keras.io>.
5. Dyhrberg, Anne Haubo. "Bitcoin, gold and the dollar--A GARCH volatility analysis." *Finance Research Letters*, vol. 16, 2016, 85–92.
6. Garcia, David, et al. "The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy." *Journal of the Royal Society Interface*, vol. 11, no. 99, 2014, p. 20140623.
7. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation*, vol. 9, no. 8, 1997, 1735–1780.
8. Hopfield, John J. "Neural networks and physical systems with emergent collective computational abilities." *Proceedings of the national academy of sciences*, vol. 79, no. 8, 1982, 2554–2558.
9. Huang, Jing-Zhi, et al. "Predicting Bitcoin returns using high-dimensional technical indicators." *The Journal of Finance and Data Science*, vol. 5, 2019, 140–155.
10. LeCun, Yann, et al. "Backpropagation applied to handwritten zip code recognition." *Neural computation*, vol. 1, no. 4, 1989, 541–551.
11. Li, Yan, and Wei Dai. "Bitcoin price forecasting method based on CNN-LSTM hybrid neural network model." *The Journal of Engineering*, no. 13, 2020, 344–347.
12. Mallqui, Dennys CA, and Ricardo AS Fernandes. "Predicting the direction, maximum, minimum and closing prices of daily Bitcoin exchange rate using machine learning techniques." *Applied Soft Computing*, vol. 75, 2019, 596–606.
13. McNally, Sean, et al. "Predicting the price of bitcoin using machine learning." *2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP)*, IEEE, 2018, 339–343.

14. Nakamoto, Satoshi. "Bitcoin: A Peer-to-Peer Electronic Cash System." *Cryptography Mailing list* at <https://metzdowd.com>, 2009.
15. Pedregosa, Fabian, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, 2011, 2825–2830.
16. Rosenblatt, Frank. "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, 1958, p. 386.
17. Smuts, Nico. "What drives cryptocurrency prices? an investigation of google trends and telegram sentiment." *ACM SIGMETRICS Performance Evaluation Review*, vol. 46, no. 3, 2019, 131–134.
18. Vassiliadis, Savvas, et al. "Bitcoin value analysis based on cross-correlations." *Journal of Internet Banking and Commerce*, vol. 22, no. S7, 2017, p. 1.
19. Wilder, J Welles. *New concepts in technical trading systems*. Trend Research, 1978.
20. Wood, Gavin. "Ethereum: A secure decentralised generalised transaction ledger." *Ethereum project yellow paper*, vol. 151, 2014, 1–32.