Τελική αναφορά για πρόβλεψη της τιμής του bitcoin με μεθόδους βαθιάς μάθησης

# Πρόλογος

Τα μοντέλα βαθιά μάθησης δείχνουν να υπερέχουν έναντι κλασικών αλγορίθμων ως προς την απόδοση τους στο πρόβλημα της πρόβλεψης χρονοσειρών. Μέχρι τώρα δεν έχουν γίνει μεγάλες έρευνες για την πρόβλεψη της τιμής του bitcoin χρησιμοποιώντας καινοτόμες αρχιτεκτονικές νευρωνικών δικτυών. Στο παρόν έγγραφο παρουσιάζουμε μια μέθοδο για την πρόβλεψη τις τιμής του Bitcoin σε βάθος μιας μέρας με μεθόδους βαθιάς μάθησης επιτυγχάνοντας αρκετά ικανοποιητικά αποτελέσματα.

# Φορμαλισμός του προβλήματος

Η προσπάθεια πρόβλεψης [+++]

Σε αυτή την ερευνα επικενρωνομαστε στις εξης καινοτομιες:

1. Εφαρμογη νεων πιο περιπλοκων αρχιτεκτονικων νευρωνικων δικτυωνς χεδιασμενων για προβλεψη χρονοσειρων για την προβλεψη της τιμης του Bitcoin
2. Συλλογη και αναλυση εταιρογενων χαρακτηριστικων για την προβλεψη της τιμης του Bitcoin
3. Backtesting των μοντελων μας σε δεδομενα πραγματικου χρονου διαρκειας 1 μηνα για την εξεταση των περιθωριων κερδοφοριας

Τα προβλήματα στα οποία τα δεδομένα είναι διατεταγμένα από τον χρόνο (χρονοσειρές) μπορεί να σχετίζονται με πρόβλεψη μελλοντικών τιμών (Παλινδρόμηση-Regression) η ταξινόμηση (Classification). Η τιμή του κρυπτονομίσματος Bitcoin είναι χρονικά εξαρτώμενη και συνεπώς μπορεί να αντιμετωπιστεί ως χρονοσειρά. Η πρόβλεψη δε της τιμής του bitcoin είναι ένα κλασσικό πρόβλημα παλίνδρομης χρονοσειράς . Επειδή η τιμή του bitcoin δεν εξαρτάται μόνο από τον χρόνο άλλα και από άλλες μεταβλητές έχουμε να κάνουμε με μια πολύ-μεταβλητή χρονοσειρά (Multi-variable Timeseries).

Η ανάπτυξη τις μοντέλου για την πρόβλεψη χρονοσειρών σημαίνει ότι θεωρούμε ως δεδομένων ότι η χρονοσειρά έχει μια εσωτερική δομή και δεν είναι τυχαίος θόρυβος. Βάσει αυτού θεωρούμε ότι υπάρχουν μοτίβα εξελισσόμενα στο χρόνο των οποίων της αναλυτική μορφή προσπαθούμε να προσεγγίζουμε με μοντέλο . Η εύρεση τις της συνάρτησης είναι η διαδικασία μοντελοποίηση της χρονοσειρά. Ως χρονοσειρές ορίζοντες ακολουθίες τιμών που εξελίσσονται στο χρόνο. Συγκεκριμένα για μια ακολουθίας εισόδου που εξελίξετε στο χρόνο όπου θέλουμε να προβλέψουμε μια ακολουθία εξόδου το μοντέλο ορίζετε ως:

Όπου T το μήκος της χρονοσειράς. Βασικός περιορισμός είναι ότι θέλουμε να προβλέψουμε την τιμή ως δεδομένα εισόδου μόνο την ακολουθία μέχρι αυτήν την χρονική στιγμή Η διαδικασία μάθησης στην πρόβλεψη χρονοσειρών είναι να βρεθεί το κατάλληλο μοντέλο που ελαχιστοποιεί το σφάλμα μεταξύ των τιμών που έχουν προβλεφθεί από την συνάρτηση και των πραγματικών τιμών.

Υπαρχουν 3 ειδων χαρακτηριστικα που μπορουν να χρησιμοποιηθουν για την προβλεψη της τιμης του Bitcoin: Τεχνικα χαρακτηριστικα (Technical features), Blockchain Χαρακτηριστικα (Blockchain features), Κοινωνικα χαρακτηριστικα (Social features). Παρακατω αναλυονται ξεχωριστα οι τρεις κατηγοριες χαρακτηριστικων:

1. **Τεχνικα χαρακτηριστικα:** Περιλαμβανουν χαρακτηριστικα που εχουν να κανουν με την κινηση της τιμης του Bitcoin (Υψηλο, Χαμηλο, Ανοιγμα, Κλεισιμο, Ογκος συναλλαγων), όπως επισης και στατιστικα χαρακτηριστικα που μπορει να προκυψουν από την αναλυση της κινησης της τιμης σε παρελθοντικους χρονους.
2. **Blockchain χαρακτηριστικα:** [+++]
3. **Κοινωνικα χαρακτηριστικα**: Περολαμβανουν τιμες που προλυπτουν από την αναλυση του συναισθηματος της αγορας γυρω από το Bitcoin. Είναι κοινως αποδεκτο οπως προκυπτει από οικονομικες μελετες, ότι οι κινησεις των αξιων συμπεριλαμβανομαινου και του Bitcoin εξαρτονται από το γενικο συναισθημα (θετικο η αρνητικο) που υπαρχει.

Παρακατω θα συλλεχθουν αρκετα χαρακτηριστικα και θα αξιολογηθει η σημαντικοτητα τους στη δυνατοτητα προβλεψης της τιμης του Bitcoin.

# Έρευνα

Για την τελική πρόβλεψη τις τιμής του bitcoin χρησιμοποιήθηκε η αρχιτεκτονική Temporal Convolution Network (TCN), περισσότερες πληροφορίες για αυτή την αρχιτεκτονική υπάρχουν στη μελέτη Bai, S et al. []. Το μοντέλο αυτό ήταν State of the art για το έτος 2018 για πολλά προβλήματα πρόβλεψης χρονοσειρών σε πολλά διαφορετικά dataset ξεπερνώντας κατά πολύ σε απόδοση μοντέλα RNN και παραλλαγές τις LSTM η GRU. Τις βάσει τις μελέτης Liu, M et all [] από τα τρία διαφορετικά είδη αρχιτεκτονικών που υπάρχουν για πρόβλεψη χρονοσειρών οι οποίες είναι αναδρομικές αρχιτεκτονικές (RNN, LSTM, GRU κλπ.), Αρχιτεκτονικές που βασίζονται τις Transformers [], και αρχιτεκτονικές που βασίζονται σε συνέλιξη, οι τελευταίες φαίνεται να έχουν την καλύτερη απόδοση. Η ερευνά αυτή παρουσιάζει ένα νέο μοντέλο που βασίζετε στη συνέλιξη και παρουσιάζει αποτελέσματα σε διαφορά dataset καλυτέρα από τις αρχιτεκτονικές TCN. Εφαρμογή τις μεθόδου για την πρόβλεψη της τιμής του Bitcoin αναμένεται να γίνει.

# Συλλογή Δεδομένων και προεπεξεργασία

## 3.1 Συλλογή δεδομένων

Τα δεδομένα τα οποία συλλέχθηκαν ήταν με συχνότητα μέρας από τις ημερομηνίες 18-02-01 έως την ημερομηνία 22-03-13. Τα **χαρακτηριστικά** (features) για κάθε ημέρα ήταν οι τιμές: Open, High, Low, Adjustment close, Close, Volume, Fear and Greed (FG) value (σύνολο 7). Η τιμή fear and greed index προστέθηκε ως χαρακτηριστικό ώστε να προστεθεί επιπλέον πληροφορία η οποία θα μπορούσε να λαμβάνει υπόψιν το μοντέλο για την τελική πρόβλεψη τις τιμής. Η τιμή fear and greed index πρακτικά συμπεριλαμβάνει ως πληροφορία εξωγενής παράγοντες που μπορεί αν επηρεάσουν την τιμή του Bitcoin και προκύπτει “αναλύοντας” τα συναισθήματα τις αγοράς. Περισσότερες πληροφορίες για το fear and greed index μπορούν να βρεθούν στον ισότοπο [].

## Προεπεξεργασία δεδομένων

Για την προεπεξεργασία των δεδομένων έγινε Standard normalization μόνο. Δεν εφαρμόστηκε κάποια μέθοδος για μείωση τις διαστατικότητα. Το παράθυρο (Temporal Window) ορίστηκε τις 24 (24 μέρες). Το βήμα μετάβασης (Shift) του παραθύρου ορίστηκε 1 (μια μέρα). Ο ορίζοντας πρόβλεψης (Horizon) ορίστηκε 1 (μια μέρα). Πρακτικά αυτό σημαίνει ότι το μοντέλο θα προβλέπει την τιμή του bitcoin “κοιτώντας” 24 μέρες πίσω. Η τιμή στόχος για κάθε μέρα (label data) είναι η τιμή Close τις επομένης μέρας. Το μοντέλο για την πρόβλεψή τις τιμής στόχου λαμβάνει υπόψιν τα 7 χαρακτηριστικά που αναφέρθηκαν πάνω για τις προηγούμενες 24 μέρες. **Πρακτικά το μοντέλο προβλέπει την τιμή τις επομένης μέρας “κοιτώντας” τα χαρακτηριστικά των 24 προηγουμένων ημέρων.**

Έγινε διαχωρισμός σε σύνολο εκπαίδευσης και σύνολο αξιολόγησης με το 80% του συνόλου των δεδομένων να χρησιμοποιείται για εκπαίδευσή και το 20% για αξιολόγηση. Δεν έγινε ανακάτεμα των δεδομένων καθώς είναι διατεταγμένα στο χρόνο οπότε οι πρώτες 1178 ημέρες χρησιμοποιήθηκαν ως σύνολο εκπαίδευσης και το υπόλοιπες 295 ως σύνολο αξιολόγησης.

# Επιλογή βέλτιστων υπερπαραμέτρων του μοντέλου

Τις αναφέραμε το μοντέλο που χρησιμοποιήθηκε είναι το TCN [] (Temporal Convolution Network) Η διαδικασία επιλογής βέλτιστων υπερπαραμέτρων (Fine tuning) του μοντέλου είναι η διαδικασία επιλογής των βέλτιστων υπερπαραμέτρων του μοντέλου ώστε να έχει την καλύτερη απόδοση σε ένα σύνολο δεδομένων. Η απόδοση του μοντέλου εξαρτάται σημαντικά από τις υπερπαραμέτρους που θα επιλεχθούν. Για το συγκεκριμένο μοντέλο επιλέχθηκαν 5 υπερπαράμετροι για βελτιστοποίηση με τις τιμές που παρουσιάζονται στο παρακάτω πίνακα:

|  |  |  |
| --- | --- | --- |
| **Υπερπαράμετρος** | **Τιμές** | **Βέλτιστη τιμή** |
| nb\_filters | [8,16,32,64,128] | 8 |
| Kernel size | [2,3,4,8,12] | 2 |
| nb\_stacks | [1,2,3,4] | 1 |
| dropout\_rate | [Νο\_dropout, dropout = 0.05] | dropout = 0.05 |
| dilations | [1,2,3,4,8]  [1,2,3,4,8,16]  [1,2,3,4,8,16,32]  [1,2,3,4,8,16,32,64]  [1,2,3,4,8,16,32,64,128] | 128 |
| Learning rate | [min = 1e-4, max = 1e-2] | 0.01 |

Η μεθοδολογία για την επιλογή βέλτιστων υπερπαραμέτρων ορίστηκε η Bayesian optimization. Έτρεξαν συνολικά 50 δοκιμές και σε κάθε δοκιμή έγιναν δυο πειράματα. Για κάθε πειράματα το Batch\_size 4 και για 100 εποχές (epochs). Το Loss ορίστηκε ως Μέσο Τετραγωνικό Σφάλμα.

# Αποτελέσματα

## 5.1 Μετρικές αξιολόγησης

Για την αξιολόγηση του μοντέλου χρησιμοποιήθηκε το Μέσο Τετραγωνικό Σφάλμα (Mean Squared Error) και το Μέσο Απολυτό Σφάλμα (Mean Absolute Error). To μέσο τετραγωνικό σφάλμα εκφράζει την διασπορά των τιμών που έχει προβλέψει το μοντέλο από τις πραγματικές τιμές.

Για οι πραγματικές τιμές και οι τιμές που πρόβλεψε το μοντέλο και το σύνολο των προβλέψεων το Μέσο Τετραγωνικό Σφάλμα ορίζετε ως:

Αντίστοιχα το Μέσο Απόλυτο Σφάλμα ορίζετε ως:

## Αποτελέσματα πειραμάτων

Στην αρχή έγιναν δυο πειράματα χωρίς να γίνει επιλογή βέλτιστων υπερπαραμέτρων. Το πρώτο πείραμα έγινε στο σύνολο δεδομένων το οποίο δεν περιείχε το ως χαρακτηριστικό την τιμή του Fear and Greed index. Το δεύτερο πείραμα έγινε προσθέτοντας την τιμή του Fear and Greed Index. Και τα δυο πειράματα έτρεξαν για 100 εποχές και με batch size 4.

|  |  |  |
| --- | --- | --- |
| Με τιμή Fear and greed index | Μέσο Τετραγωνικό σφάλμα | 0.027 |
| Μέσο απόλυτο σφάλμα | 0.117 |
| Χωρίς τιμή Fear and greed index | Μέσο Τετραγωνικό σφάλμα | 0.012 |
| Μέσο απόλυτο σφάλμα | 0.086 |

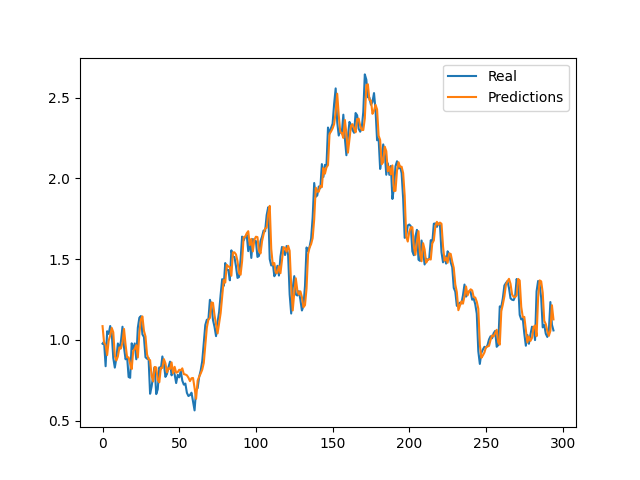
Από τα πειράματα φαίνεται καθαρά ότι το μοντέλο αποδίδει καλυτέρα χωρίς την τιμή Fear and greed index ωστόσο επιλέξαμε να την κρατήσουμε για τα επόμενα πειράματα. Αυτό έγινε επειδή η τιμή Fear and greed index είναι μια επιπλέον πηγή πληροφορίας η οποία ξέρουμε ότι καθορίζει την τιμή του Bitcoin. Έτσι θεωρούμε ότι οι προβλέψεις που περιλαμβάνουν την τιμή fear and greed index είναι πιο αντικειμενικές καθώς λαμβάνουν υπόψιν και εξωγενής (κοινωνικούς -συναίσθημα της αγοράς) που καθορίζουν την τιμή του Bitcoin.

Στην συνέχεια παρουσιάζονται τα αποτελέσματα από το καλύτερο μοντέλο που προέκυψε από την διαδικασία επιλογής των υπερπαραμέτρων.

Παρακάτω παρουσιάζονται τα αποτελέσματα από το καλύτερο πειράματα στο σύνολο αξιολόγησης

|  |  |
| --- | --- |
| **Μετρική** | **Τιμή** |
| Μέσο τετραγωνικό σφάλμα | 0.009 |
| Μέσο απόλυτο σφάλμα | 0.071 |

Στο παρακάτω σχήμα παρουσιάζονται οι πραγματικές τιμές σε σχέση με τις τιμές που πρόβλεψε το μοντέλο για το bitcoin στο σύνολο αξιολόγησης.



\

# Συμπεράσματα και μελλοντικές βελτιώσεις

Μέσω πειραμάτων για την εύρεση των βέλτιστων υπερπαραμέτρων προέκυψε ένα μοντέλο με αρκετά καλή απόδοση (0.9% Μέσο Τετραγωνικό Σφάλμα). Στην συνέχεια μπορούν να γίνουν και νέα πειράματα προσθέτοντας νέες μεθοδολογίες για την βελτίωση αυτής της απόδοσης στην πρόβλεψη της τιμής του Bitcoin. Κάποιές από αυτές τις μεθοδολογίες μπορούν να είναι:

* Πειραματισμός με νέα μοντέλα που έχουν καλύτερη απόδοση σε προβλήματα πρόβλεψης χρονοσειρών.
* Καλύτερη προεπεξεργασία των δεδομένων η οποία μπορεί να περιλαμβάνει μείωση της διαστατικότητας, δημιουργία νέων χαρακτηριστικών, εφαρμογή καλύτερων μεθόδων κανονικοποίηση των δεδομένων.