ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
Ιανουάριος 2022

Μηχανική Μάθηση

ΕΞΑΜΗΝΙΑΙΑ ΕΡΓΑΣΙΑ

Άγγελος Τσελές (Α.Μ : 3170160)

Εισαγωγικά

Αρχικά, φορτώνω το dataset κατευθείαν από την ιστοσελίδα του Yahoo Finance με τα δεδομένα των τελευταίων 5 ετών (από 1/1/2017 έως και 17/1/2022 που γράφεται η παρούσα αναφορά).

Παρατηρούμε ότι οι τιμές των πεδίων Adj Close και Close είναι ακριβώς οι ίδιες για κάθε μέρα, οπότε και διαγράφουμε το πεδίο Adj Close για δική μας ευκολία. Χωρίζουμε το dataset με αναλογία 80/20 στις μεταβλητές X και y.

Τέλος, οπτικοποιούμε την διαμόρφωση της τιμής κατά την διάρκεια αυτών των 5 ετών και παρατηρούμε την ραγδαία αύξηση της τιμής κλεισίματος από το τελευταίο τρίμηνο του 2020 και έπειτα.

Προεπεξεργασία Δεδομένων

Πρώτα, κάνουμε reshape τα X και y προκειμένου να μετατραπούν και τα δύο σε πίνακες δύο διαστάσεων. Έπειτα, κάνουμε κανονικοποίηση χρησιμοποιώντας την συνάρτηση **MinMaxScaler** για να ομαλοποιήσουμε τις τιμές στα δεδομένα σε ένα εύρος μεταξύ 0 και 1.

Δημιουργούμε το σετ με τα δεδομένα εκπαίδευσης σε μορφή data frame με τις τιμές των προηγούμενων 50 ημερών πριν την κάθε δεδομένη μέρα και στην συνέχεια μετατρέπουμε τις αντίστοιχες μεταβλητές σε πίνακες.

Με τον ίδιο τρόπο δημιουργούμε και το σετ με τα δεδομένα ελέγχου και πλέον είμαστε έτοιμοι για να εφαρμόσουμε την Γραμμική Παλινδρόμηση.

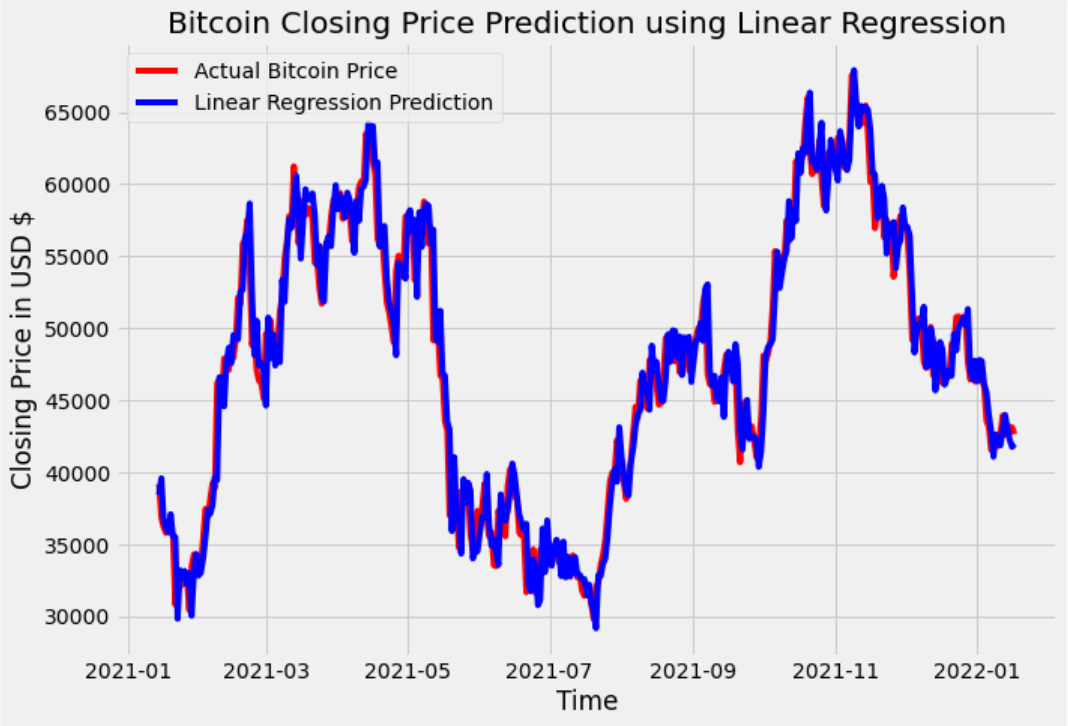
Γραμμική Παλινδρόμηση

Χρησιμοποιώντας την βιβλιοθήκη sklearn , εκπαιδεύουμε το μοντέλο μας με τα δεδομένα εκπαίδευσης που δημιουργήσαμε προηγουμένως. Με την μέθοδο predict , παίρνουμε τις προβλέψεις και κάνοντας χρήση της μεθόδου inverse_transform επαναφέρουμε τις προβλέψεις στην μορφή πίνακα μιας διάστασης.

Μετέπειτα, εκτελούμε μετρήσεις cross validation όπου βλέπουμε ότι επιτυγχάνουμε ακρίβεια περίπου 95% στα δεδομένα εκπαίδευσης, το οποίο επιβεβαιώνει ότι κάναμε καλή δουλειά όσον αφορά τα δεδομένα εκπαίδευσης.

Εκτελούμε και μετρήσεις για την ακρίβεια και τις αποκλίσεις μεταξύ των πραγματικών τιμών και των τιμών που πρόβλεψε το μοντέλο μας.

Επιτυγχάνουμε ακρίβεια με ποσοστό περίπου 96% και οπτικοποιώντας τις προβλέψεις μας με τις πραγματικές τιμές, βλέπουμε ότι το μοντέλο μας τα πήγε εξαιρετικά στην πρόβλεψη των τιμών κλεισίματος με την μέθοδο αυτή.



Λογιστική Παλινδρόμηση

Μιας και το μοντέλο της Λογιστικής Παλινδρόμησης αφορά προβλήματα κατηγοριοποίησης, θα πρέπει να αναγάγουμε το πρόβλημα μας σε κάτι διαφορετικό από την πρόβλεψη της τιμής κλεισίματος που είδαμε στην Γραμμική Παλινδρόμηση.

Μία ιδέα είναι να λογαριθμήσουμε την τιμή κλεισίματος για να καθαρίσουμε τον θόρυβο από τα δεδομένα μας και να συγκρίνουμε την λογαριθμημένη τιμή κλεισίματος μιας μέρας με την αντίστοιχη λογαριθμημένη τιμή κλεισίματος της προηγούμενης μέρας. Αν η διαφορά αυτή είναι μικρότερη από 0 τότε την κατηγοριοποιούμε με την τιμή 1 αλλιώς αν η διαφορά είναι μεγαλύτερη ή ίση του 0,τοτε την κατηγοριοποιούμε με την τιμή 0.

Αφού δημιουργήσουμε την μεταβλητή return_var με τιμές 0 ή 1, κάνουμε τα ίδια βήματα με πριν , όπου δημιουργούμε τα X και y, χωρίζουμε τα δεδομένα μας με αναλογία 80/20 και δημιουργούμε τα δεδομένα εκπαίδευσης και δεδομένα ελέγχου.

Χρησιμοποιώντας την βιβλιοθήκη sklearn , εκπαιδεύουμε το μοντέλο μας με τα δεδομένα εκπαίδευσης που δημιουργήσαμε προηγουμένως. Με την μέθοδο predict , παίρνουμε τις προβλέψεις και μετατρέπουμε τον πίνακα δύο διαστάσεων με τα δεδομένα ελέγχου του y σε πίνακα μίας διάστασης.

Μετέπειτα, εκτελούμε μετρήσεις cross validation όπου βλέπουμε ότι επιτυγχάνουμε ακρίβεια περίπου 55% στα δεδομένα εκπαίδευσης, το οποίο αποδεικνύει ότι η μέθοδος μας δεν είναι και καλύτερη δυνατή.

Εκτελούμε και μετρήσεις για την ακρίβεια και τις αποκλίσεις μεταξύ των πραγματικών τιμών και των τιμών που πρόβλεψε το μοντέλο μας.

Επιτυγχάνουμε ακρίβεια με ποσοστό περίπου 52% και οπτικοποιώντας τις προβλέψεις μας με τις πραγματικές τιμές, βλέπουμε ότι το μοντέλο μας δεν μπορεί να κάνει προβλέψεις με καλή ακρίβεια.

Βεβαίως, αυτό δεν είναι κάτι που μας ξαφνιάζει μιας και η χρήση της μεθόδου της λογιστικής παλινδρόμησης σε προβλήματα πρόβλεψης μετοχών και κρυπτονομισμάτων δεν επιτυγχάνει υψηλή ακρίβεια και ακριβείς προβλέψεις. Ένας τρόπος για να βελτιώσουμε το υπάρχον μοντέλο είναι η χρήση ενός κινητού μέσου, το οποίο μπορεί να ανεβάσει την ακρίβεια τουλάχιστον κατά 20-30%.

Νευρωνικό Δίκτυο

Αρχικά να τονίσουμε ότι επειδή έχουμε χρησιμοποιήσει την ίδια λογική που αναφέρει η εκφώνηση και στην Γραμμική Παλινδρόμηση, δεν χρειάζεται να κάνουμε κάποιες παραπάνω ενέργειες πάνω στα δεδομένα. Επομένως, θα χρησιμοποιήσουμε τα ίδια δεδομένα εκπαίδευσης και ελέγχου στο νευρωνικό μας δίκτυο.

Δημιουργούμε το μοντέλο μας με δύο κρυφά επίπεδα με 50 νευρώνες έκαστος καθώς και το επίπεδο output με ένα νευρώνα. Κάνουμε compile το μοντέλο χρησιμοποιώντας ως μετρική αξιολόγησης το loss και τον optimizer Adam του Keras.

Στο σημείο αυτό να τονίσουμε ότι έγιναν 8 διαφορετικές μετρήσεις πειράζοντας διαφορετικές παραμέτρους κάθε φορά.

Οι παράμετροι οι οποίοι άλλαζαν κάθε αφορά ήταν το learning rate του optimizer Adam καθώς και το πλήθος των epochs κατά την εκπαίδευση του μοντέλου. Το Mean Squared Error ήταν η μετρική που χρησιμοποιήθηκε για να γίνει η σύγκριση μεταξύ των μετρήσεων.

No	**epochs**	**Learning rate**	**Mean Squared Error**
1	20	0.001 (default)	8.4 * 10^ (-4)
2	50	0.001 (default)	7.0* 10^ (-4)
3	100	0.001 (default)	5.5* 10^ (-4)
4	200	0.001 (default)	5.8* 10^ (-4)
5	20	0.01	6.2* 10^ (-4)
6	50	0.01	4.8* 10^ (-4)
7	100	0.01	4.1* 10^ (-4)
8	200	0.01	4.5* 10^ (-4)

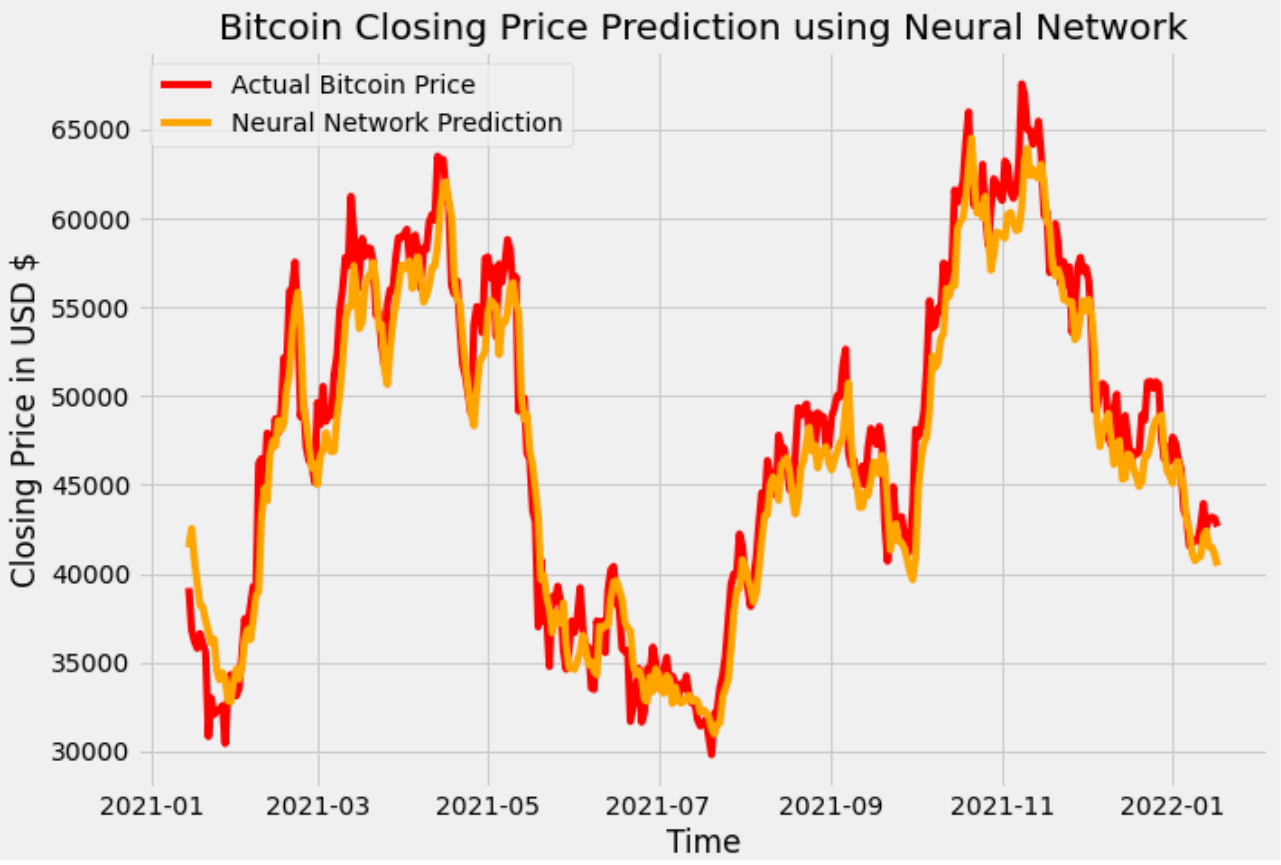
Παρατηρούμε ότι για learning rate = 0.01 (δηλαδή πιο αργό διάβασμα) επιτυγχάνουμε αρκετά μικρότερο MSE σε σύγκριση με την περίπτωση που είναι ίσο με 0.001 που είναι η default τιμή του optimizer.

Ως προς το πλήθος των epochs, παρατηρούμε ότι για epochs = 100 επιτυγχάνουμε το μικρότερο δυνατό MSE.

Εκπαιδεύουμε το μοντέλο και παρατηρούμε ότι το validation loss είναι της τάξης του 0,003-0,004% , το οποίο επιβεβαιώνει ότι κάναμε καλή δουλειά όσον αφορά τα δεδομένα εκπαίδευσης.

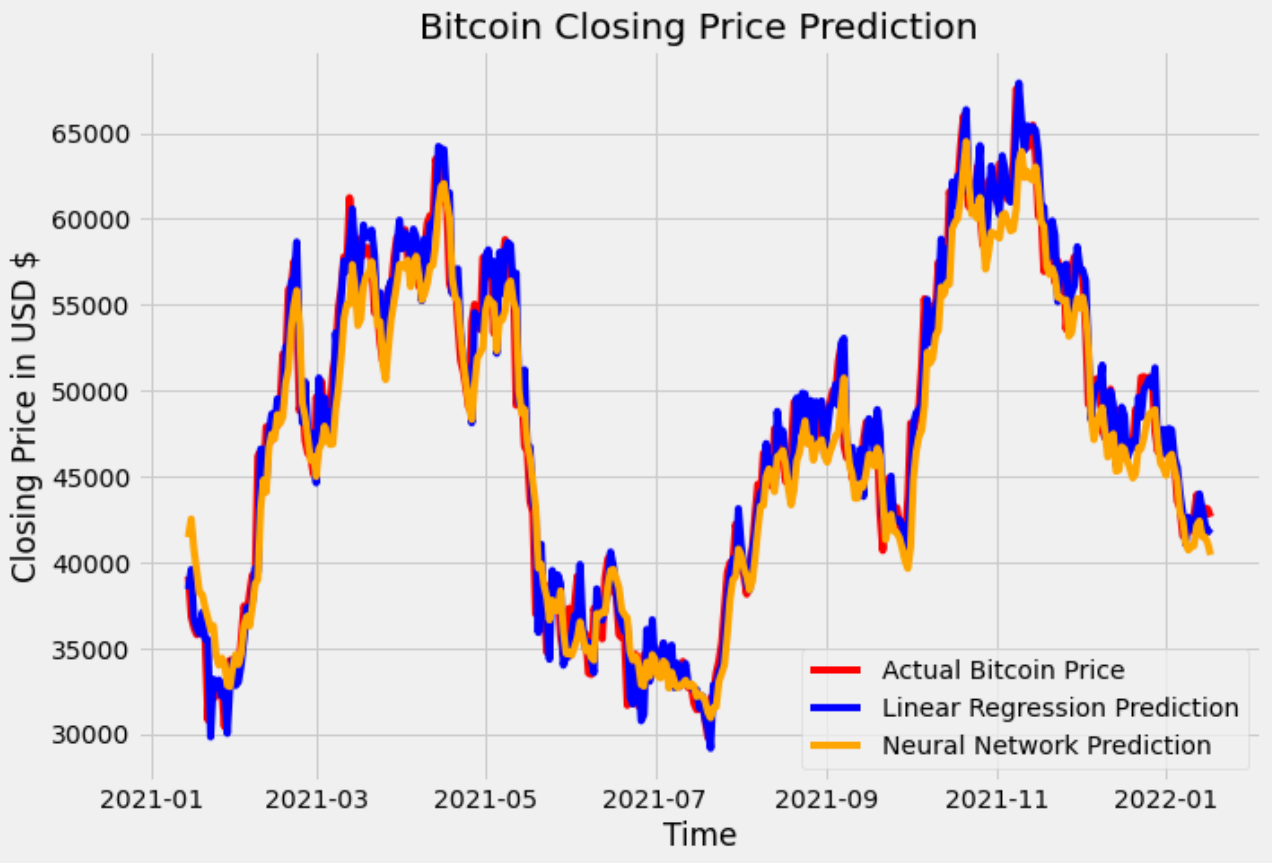
Εκτελούμε και μετρήσεις για την ακρίβεια και τις αποκλίσεις μεταξύ των πραγματικών τιμών και των τιμών που πρόβλεψε το μοντέλο μας.

Επιτυγχάνουμε ακρίβεια με ποσοστό περίπου 93% και οπτικοποιώντας τις προβλέψεις μας με τις πραγματικές τιμές, βλέπουμε ότι το μοντέλο μας τα πήγε εξαιρετικά στην πρόβλεψη των τιμών κλεισίματος με την μέθοδο αυτή.



Συμπέρασμα

Συγκρίνοντας τις μεθόδους της Γραμμικής Παλινδρόμησης και του Νευρωνικού Δικτύου, παρατηρούμε ότι και τα δύο μοντέλα που φτιάξαμε επιτυγχάνουν πολύ υψηλά ποσοστά ακρίβειας στις προβλέψεις τους, με μια ελαφριά υπεροχή της Γραμμικής Παλινδρόμησης κατά περίπου 3%.



Το μοντέλο της Λογιστικής Παλινδρόμησης μπορεί να χρησιμοποιηθεί για να προβλέψουμε αν η τιμή κλεισίματος αυξάνεται ή μειώνεται σε σύγκριση με την προηγούμενη μέρα. Η απλή υλοποίηση ,όμως, δεν αρκεί για ακριβείς προβλέψεις και χρειάζεται να υλοποιηθεί με την βοήθεια της μεθόδου του κινητού μέσου προκειμένου να πιάνει υψηλότερα ποσοστά ακρίβειας σε σύγκριση με το μοντέλο που φτιάξαμε εμείς.