ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Σύγκριση μοντέλων μηχανικής μάθησης

στη πρόβλεψη κρυπτονομισμάτων.

Χρανιώτης Παναγιώτης-Ιάσων

Α.Μ. 141002

Επιβλέπων καθηγητής: Φατούρος Σταύρος

ΑΘΗΝΑ-ΑΙΓΑΛΕΩ, 09/2021

Η Τριμελής Επιτροπή

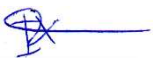
Stavros Fatouros Digitally signed by Stavros Fatouros Date: 2022.09.23 11:50:54 +03'00'	Dimitrios Karolidis Digitally signed by Dimitrios Karolidis Date: 2022.09.24 13:54:09 +03'00'	Georgios Meletiou Digitally signed by Georgios Meletiou Date: 2022.09.26 20:39:11 +03'00'
Σταύρος Φατούρος	Δημήτριος Καρολίδης	Γεώργιος Μελετίου
Αναπληρωτής Καθηγητής	Λέκτορας Εφαρμογών	Ε.ΔΙ.Π.

Δήλωση Συγγραφέα Διπλωματικής Εργασίας

Ο κάτωθι υπογεγραμμένος Χρανιώτης Παναγιώτης-Ιάσων του Δημητρίου, με αριθμό μητρώου 141002 φοιτητής του Πανεπιστημίου Δυτικής Αττικής της Σχολής Μηχανικών του Τμήματος Μηχανικών Πληροφορικής και Υπολογιστών, δηλώνω υπεύθυνα ότι:

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

Ο Δηλών



Ευχαριστίες

Στο σημείο αυτό θα ήθελα να ευχαριστήσω την οικογένεια μου που με στήριξε όλα αυτά τα χρόνια της φοίτησής μου, καθώς και τον επιβλέποντα καθηγητή μου για την καθοδήγηση που μου παρείχε καθ' όλη την διάρκεια της εκπόνησης της διπλωματικής εργασίας.

Περίληψη

Η πρόβλεψη πριν από κάθε απόφαση στο χρηματοοικονομικό τομέα είναι ζωτικής σημασίας. Μέχρι πρόσφατα το ζήτημα αυτό είχαν αναλάβει εξ' ολοκλήρου οι οικονομολόγοι. Τελευταία όμως, η τεχνολογία είναι ένας σοβαρός διεκδικητής του έργου αυτού. Με την πρόσφατη πρόοδο στην υπολογιστική ισχύ των υπολογιστών και πιο σημαντικό την ανάπτυξη πιο προηγμένων μοντέλων μηχανικής μάθησης, δημιουργούνται νέοι αλγόριθμοι για την πρόβλεψη δεδομένων χρονοσειρών. Γενικά, υπάρχουν αρκετές τεχνικές για την αποτελεσματική πρόβλεψη των μελλοντικών δεδομένων χρονοσειρών, με τα στατιστικά μοντέλα, να χρησιμοποιούν, μεταξύ άλλων χαρακτηριστικών , την περιοδικότητα των χρονοσειρών για να κάνουν αξιόλογες προβλέψεις. Τί γίνεται όμως στην περίπτωση που οι χρονοσειρές δεν έχουν περιοδικότητα; Η φύση των χρονοσειρών των κρυπτονομισμάτων είναι τέτοια που δεν επηρεάζονται από δεδομένα του παρελθόντος αλλά από εξωτερικούς παράγοντες όπως η προσφορά και η ζήτηση. Για το λόγο αυτό, πέρα από στατιστικά μοντέλα θα γίνει και χρήση μοντέλου νευρωνικών δικτύων, ένα ισχυρό εργαλείο γενικού σκοπού της μηχανικής μάθησης. Στη μελέτη αυτή θα γίνει η ανάλυση των δύο αυτών διαφορετικών φιλοσοφιών προσέγγισης, καθώς και η σύγκριση των αποτελεσμάτων αυτών.

Λέξεις κλειδιά: Μηχανική μάθηση, Προβλεπτική ανάλυση, Κρυπτονομίσματα, Πρόβλεψη κρυπτονομισμάτων, Μοντέλα μηχανικής μάθησης

Abstract

Forecasting before any decision in financials is vital. Until recently, this issue was taken up entirely by economists. Lately, technology is a serious contender for this project. With recent advances in computing power and more importantly the development of more advanced machine learning models, new algorithms for predicting time series data are being created. In general, there are several techniques for effectively predicting future time series data, with statistical models using, among other features, time series seasonality to make remarkable predictions. But what if the time series do not have seasonality? The nature of cryptocurrency time series is such that they are not influenced by past data but by external factors such as supply and demand. For this reason, in addition to statistical models, a neural network model will be used, a powerful tool of general purpose in machine learning. This study will analyze these two different approaches, as well as compare the results they give.

Keywords: Machine Learning, Predictive Analysis, Cryptocurrencies, Cryptocurrency Prediction, Machine Learning Models

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΙΚΕΣ ΣΗΜΕΙΩΣΕΙΣ

1.1 Αντικείμενο και Στόχοι της Διπλωματικής Εργασίας

Η εκπόνηση της συγκεκριμένης μελέτης έχει ως σκοπό την σύγκριση μοντέλων μηχανικής μάθησης στη πρόβλεψη χρονοσειρών κρυπτονομισμάτων με στόχο την εύρεση του αποδοτικότερου από αυτούς. Επίσης διαπραγματεύεται το ερώτημα εάν συνιστάται η χρήση τέτοιων μοντέλων για την πρόβλεψη κρυπτονομισμάτων έπειτα από τα ποσοστά επιτυχίας των μετρικών αξιολογήσεων.

1.2 Περιγραφή και βασικά χαρακτηριστικά

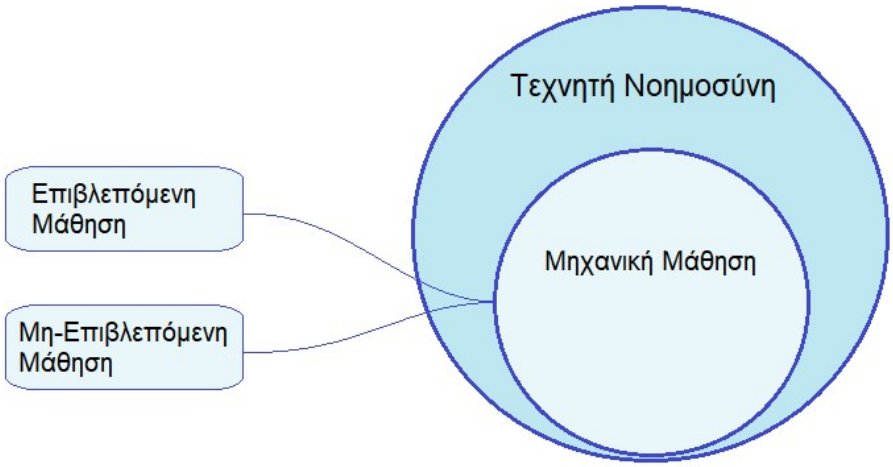
1.2.1 Μηχανική μάθηση

1.2.1.1 Εισαγωγή

Η μηχανική μάθηση είναι ένας κλάδος της τεχνητής νοημοσύνης (AI) και της επιστήμης των υπολογιστών που επικεντρώνεται στη χρήση δεδομένων και αλγορίθμων για να μιμηθεί τον τρόπο με τον οποίο μαθαίνουν οι άνθρωποι, βελτιώνοντας σταδιακά την ακρίβειά του. Η κλασική μηχανική μάθηση συχνά κατηγοριοποιείται με βάση το πώς ένας αλγόριθμος μαθαίνει να γίνεται πιο ακριβής στις προβλέψεις του. Υπάρχουν τρείς βασικές κατηγορίες: Η επιβλεπόμενη μάθηση και η μη-επιβλεπόμενη μάθηση.

Επιβλεπόμενη μάθηση: Σε αυτόν τον τύπο μηχανικής μάθησης, ο αλγόριθμος εκπαιδεύεται σε επισημασμένα δεδομένα εκπαίδευσης, οπού και η είσοδος και η έξοδος είναι προκαθορισμένες.

Μη-επιβλεπόμενη μάθηση: Αντίθετα, στην μη-επιβλεπόμενη μάθηση, τα δεδομένα που δίνονται στον αλγόριθμο δεν είναι επισημασμένα, αλλά προσπαθεί από μόνος του να τα κατηγοριοποιήσει βάσει μοτίβων που αναγνωρίζει σε αυτά.



Σχήμα 1.1: Μηχανική Μάθηση.

Ο τύπος αλγορίθμων που επιλέγει να χρησιμοποιήσει κάποιος εξαρτάται από το είδος των δεδομένων που θέλει να προβλέψει. Στη συγκεκριμένη μελέτη, οπού τα δεδομένα μας είναι χρονοσειρές, συνιστάται η χρήση αλγορίθμων επιβλεπόμενης μάθησης.

1.2.1.2 State-of-the-Art στη μηχανική μάθηση

Τα τελευταία χρόνια η μηχανική μάθηση έχει αρχίσει να μπαίνει στην καθημερινότητα του μέσου ανθρώπου και να χρησιμοποιείται όλο και περισσότερο. Αυτό έχει ως αποτέλεσμα καινούργιοι αλγόριθμοι να δημιουργούνται συνέχεια με τον έναν να είναι καλύτερος από τον άλλον. Φυσικά εδώ ισχύει το "έκαστος στο είδος του" εννοώντας πως δεν υπάρχει αλγόριθμος που να είναι ο καλύτερος σε όλους τους τομείς αλλά μόνο για το σκοπό για τον οποίο κατασκευάστηκε.

Παρακάτω παρουσιάζονται κάποιοι από τους καλύτερους αλγόριθμους οι οποίοι θεωρούνται οι αποτελεσματικότεροι σε κάποια βασικές εφαρμογές της μηχανικής μάθησης σύμφωνα με τα αποτελέσματα που παρουσιάζονται στη ιστοσελίδα https://paperswithcode.com/sota.

Όραση υπολογιστών

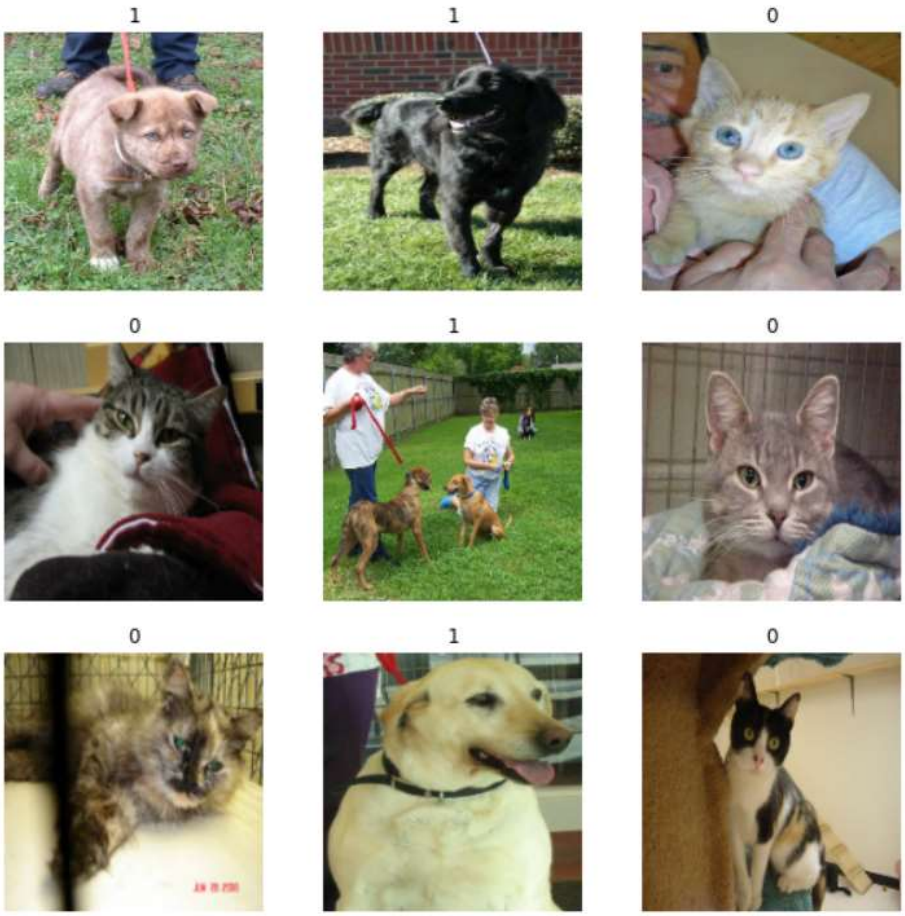
Ίσως το πιο διαδεδομένο πεδίο της μηχανικής μάθησης, η όραση υπολογιστών χρησιμοποιείται για απλές καθημερινές εφαρμογές, όπως η αναγνώριση προσώπου σε μια κάμερα κινητού, μέχρι πολύπλοκες εφαρμογές, όπως η αυτόνομη οδήγηση. Επιτρέπει στους υπολογιστές και τα

συστήματα να αντλούν σημαντικές πληροφορίες από ψηφιακές εικόνες, βίντεο και άλλες οπτικές εισόδους, τις οποίες χρησιμοποιούν στη συνέχεια για να μιμηθούν αλγοριθμικά την αίσθηση της όρασης.

Κάποιες από τις πιο σημαντικές τεχνικές της όρασης υπολογιστών είναι: η Ταξινόμηση εικόνας, η Σημασιολογική κατάτμηση και η Ανίχνευση αντικειμένων.

Ταξινόμηση εικόνας

Η ταξινόμηση εικόνας είναι μια εργασία που προσπαθεί να κατανοήσει μια ολόκληρη εικόνα στο σύνολό της. Ο στόχος είναι να ταξινομηθεί μια εικόνα, αναθέτοντάς της μια συγκεκριμένη ετικέτα. Συνήθως, η ταξινόμηση εικόνας αναφέρεται σε εικόνες στις οποίες εμφανίζεται και αναλύεται μόνο ένα αντικείμενο.



Σχήμα 1.2: Ταξινόμηση εικόνας όπου 0 είναι γάτα και 1 είναι σκύλος.[1]

Η αναγνώριση προσώπου με σκοπό την ταυτοποίηση είναι ίσως η πιο διαδεδομένη εφαρμογή που βλέπει η ταξινόμηση εικόνας, αλλά χρησιμοποιείται και σε άλλες διεργασίες όπως ο εντοπισμός ακατάλληλων ή προσβλητικών εικόνων.

Ο αλγόριθμος ViT-G/14 με ακρίβεια 90.45% θεωρείται ο καλύτερος αλγόριθμος στη ταξινόμηση εικόνας. Ένας άλλος αλγόριθμος που αξίζει να αναφερθεί είναι ο EfficientNet-L2 με ακρίβεια 90.20%.

Σημασιολογική κατάτμηση

Στη σημασιολογική κατάτμηση ο αλγόριθμος προσπαθεί να ξεχωρίσει τις δομές και τα στοιχεία που αποτελούν μια ψηφιακή εικόνα βάζοντας "ταμπέλες" στα pixel της. Θα μπορούσε κανείς να πει πως είναι σαν την ταξινόμηση εικόνας σε επίπεδο pixel.

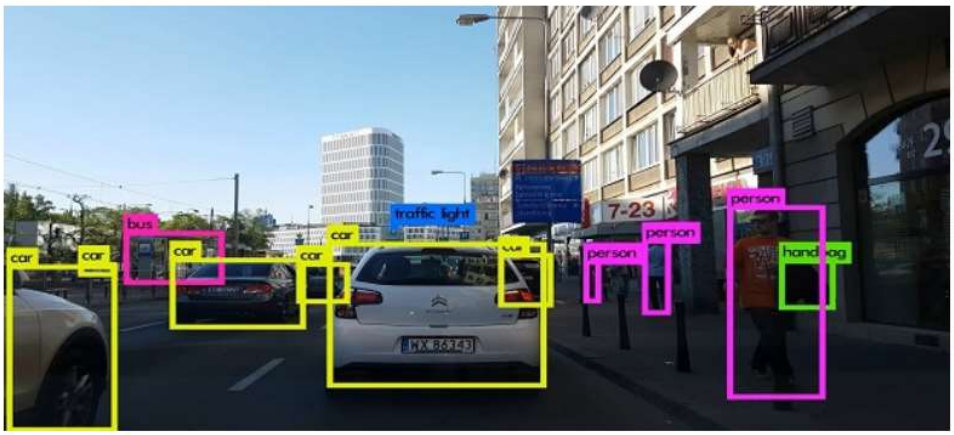
Σχήμα 1.3: Σημασιολογική κατάτμηση στην αυτόνομη οδήγηση.[2]

Χρησιμοποιείται σε διαδικασίες όπως η αναγνώριση χειρόγραφων γραμμάτων ή συμβόλων αλλά πιο σημαντικά στην αυτόνομη οδήγηση, οπού ένα αυτοκινούμενο όχημα χρειάζεται πλήρη κατανόηση του περιβάλλοντός του σε επίπεδο pixel. Ως εκ τούτου, η σημασιολογική κατάτμηση χρησιμοποιείται για τον προσδιορισμό λωρίδων και άλλων απαραίτητων πληροφοριών.

Η μετρική που χρησιμοποιείται στη σημασιολογική κατάτμηση για να αξιολογήσει τα μοντέλα είναι η Mean Intersection Over Union (Mean IOU), στην οποία ο HRNet-OCR, με βαθμολογία 85.1%, κατέχει την θέση ως ο αποδοτικότερος αλγόριθμος στο τομέα αυτό. Άλλοι αλγόριθμοι με κορυφαίες επιδόσεις είναι οι EfficientPS με βαθμολογία 84.21% και ο Panoptic-DeepLab με 84.20%.

Ανίχνευση αντικειμένου

Σε αντίθεση με την σημασιολογική κατάτμηση, όπου όλα τα στιγμιότυπα μιας κατηγορίας αντικειμένου κατηγοριοποιούνται στην ίδια ομάδα, στην ανίχνευση αντικειμένων τα μοντέλα προσπαθούν να εντοπίσουν και να ξεχωρίσουν όλα τα προς μελέτη αντικείμενα που περιέχει μια εικόνα ή ένα βίντεο. Χρησιμοποιείται για την καταμέτρηση αντικειμένων, τον προσδιορισμό και την παρακολούθηση των ακριβών τοποθεσιών τους, ενώ παράλληλα προσθέτει στο κάθε ένα την αντίστοιχη ετικέτα.



Σχήμα 1.4: Ανίχνευση Αντικειμένου στην αυτόνομη οδήγηση.[3]

Εφαρμογές που χρησιμοποιούν τέτοια τεχνική, μεταξύ άλλων, είναι η αναγνώριση προσώπου από μια φωτογραφική μηχανή σε πραγματικό χρόνο, η αναγνώριση κινούμενων αντικειμένων από μια κάμερα ασφαλείας, η αυτόνομη οδήγηση κ.α..

Ο αλγόριθμος που θεωρείται ο καλύτερος στην ανίχνευση αντικειμένου είναι ο Soft Teacher +Swin-L με βαθμολογία 61.3 σύμφωνα με την μετρική box Average Precision, η οποία χρησιμοποιείται για να αξιολογήσει τέτοιου είδους μοντέλα.

Επεξεργασία φυσικής γλώσσας

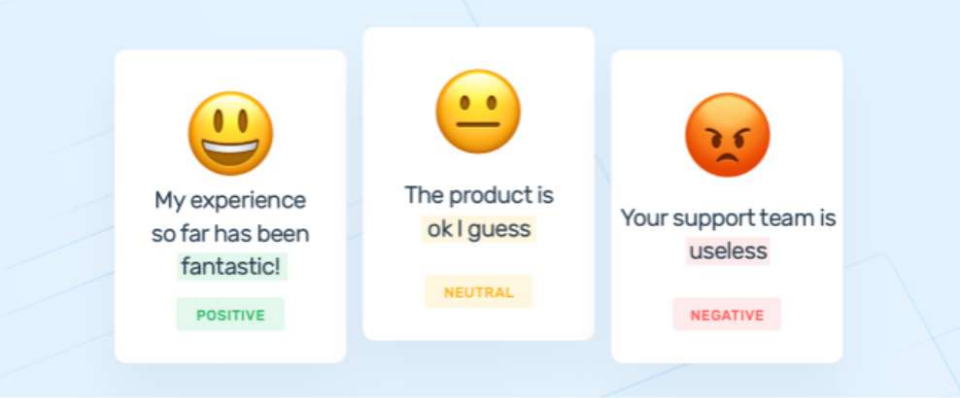
Η επεξεργασία φυσικής γλώσσας είναι ένα πεδίο της μηχανικής μάθησης το οποίο δίνει σε μια μηχανή την ικανότητα να διαβάσει, να κατανοήσει και να παραγάγει νόημα από ανθρώπινη γλώσσα. Πρακτικά, η φυσική γλώσσα χωρίζεται σε κομμάτια, έτσι ώστε η γραμματική δομή των προτάσεων και το νόημα των λέξεων να μπορούν να αναλυθούν και να γίνουν κατανοητά

στο σύνολό τους. Αυτό βοηθά τους υπολογιστές να διαβάζουν και να κατανοούν προφορικό ή γραπτό κείμενο με τον ίδιο τρόπο όπως οι άνθρωποι.

Κάποιες από τις πιο σημαντικές τεχνικές της επεξεργασίας φυσικής γλώσσας είναι η Ανάλυση συναισθήματος, η Γλωσσική μοντελοποίηση, η Μηχανική μετάφραση, και Αυτόματη ερωταπόκριση.

Ανάλυση συναισθήματος

Η ανάλυση συναισθήματος, η οποία αναφέρεται και ως εξόρυξη γνώμης, είναι η προσπάθεια ενός συστήματος να κατηγοριοποιήσει ένα κείμενο ως θετικό, αρνητικό ή ουδέτερο, δίνοντας έμφαση κυρίως σε λέξεις οι οποίες εκφράζουν κάποιο συναίσθημα (π.χ. "ικανοποιητική εξυπηρέτηση" - θετικό, "άνοστο φαγητό" - αρνητικό).



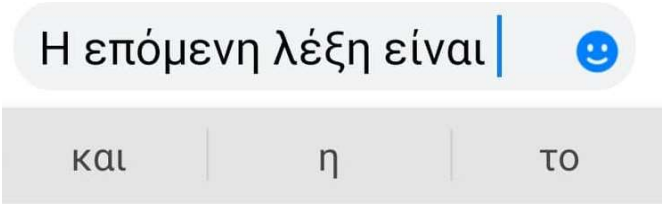
Σχήμα 1.5: Ανάλυση συναισθήματος σε κριτικές.[4]

Η ανάλυση συναισθήματος χρησιμοποιείται κατά κόρον από εταιρίες, οι οποίες επιθυμούν να έχουν ένα αξιόπιστο "feedback" όσον αφορά την θέση τους στην αγορά σύμφωνα με την γνώμη των καταναλωτών. Πιο συγκεκριμένα χρησιμοποιείται κυρίως σε μέσα κοινωνικής δικτύωσης ή και σε σχόλια προϊόντων ή υπηρεσιών από χρήστες, για την αποτελεσματικότερη μονόδρομη επικοινωνία πελάτη-εταιρίας που αποσκοπεί στην βελτίωση των προϊόντων ή υπηρεσιών.

Ο αποτελεσματικότερος αλγόριθμος στην ανάλυση συναισθήματος με ποσοστό ακρίβειας 97.5% είναι ο SMART-RoBERTa Large. Με μικρή διαφορά συνεχίζουν οι αλγόριθμοι T5-3B και ALBERT με 97.4 και 97.1 αντίστοιχα.

Γλωσσική μοντελοποίηση

Η Γλωσσική μοντελοποίηση είναι η τεχνική κατά την οποία ένα σύστημα προσπαθεί να κάνει προβλέψεις όσων αφορά την λέξη ή γράμματα μιας λέξης σε ένα κείμενο, βάση της προηγούμενης λέξης ή και ολόκληρης πρότασης. Στην ουσία, ένα γλωσσικό μοντέλο δίνει την πιθανότητα μια σειρά από λέξεις να είναι έγκυρη. Εγκυρότητα στην συγκεκριμένη περίπτωση δεν σημαίνει γραμματική ή συντακτική, αλλά σημαίνει ότι μοιάζει περισσότερο με τον τρόπο που μιλούν, ή πιο συγκεκριμένα γράφουν, οι άνθρωποι.



Σχήμα 1.6: Γλωσσική μοντελοποίηση στο Κοινωνικό δίκτυο.

Το υπό-πεδίο αυτό της επεξεργασίας φυσικής γλώσσας χρησιμοποιείται σε πολλές εφαρμογές με τις πιο κοινές να είναι οι μηχανές αναζήτησης και η χρήση “chat” στα μέσα κοινωνικής δικτύωσης.

Ο αλγόριθμος που θεωρείται καλύτερος στη Γλωσσική Μοντελοποίηση είναι ο Megatron-LM. Σύμφωνα με την αρνητικού προσανατολισμού μετρική Test Perplexity που χρησιμοποιείται για την αξιολόγηση σε γλωσσικά μοντέλα, ο Megatron-LM με βαθμολογία 10.81, κατέχει την πρώτη θέση με σημαντική διαφορά από τα υπόλοιπα.

Μηχανική μετάφραση

Η Μηχανική μετάφραση, όπως ο τίτλος υπονοεί, είναι το υπό-πεδίο το οποίο ασχολείται με την μετάφραση λέξεων ή φράσεων από μία φυσική γλώσσα σε μία άλλη. Είναι αρκετά αποδοτικό στο να δώσει μια γενική ιδέα του προς μετάφραση κειμένου, αλλά παρουσιάζει μια σχετική αδυναμία στην λέξη-προς-λέξη μετάφραση.

Γενικού σκοπού μοντέλα μηχανικής μετάφρασης, όπως το Google Translator, χρησιμοποιούνται καθημερινά από τον μέσο άνθρωπο σε μια πληθώρα ξεχωριστών θεμάτων. Εταιρίες όμως που χρησιμοποιούν συγκεκριμένο λεξιλόγιο με ειδικούς όρους κάνουν χρήση ειδικευμένων μοντέλων μηχανικής μετάφρασης.

Η αξιολόγηση των μοντέλων του υπό-πεδίου αυτού γίνεται με την μετρική BLEU, η οποία μετράει την αντιστοιχία μεταξύ μηχανικής μετάφρασης και ανθρώπινης μετάφρασης. Με βαθμολογία 35.14 ο αλγόριθμος Transformer Cycle (Rev) θεωρείται ο καλύτερος στη μηχανική μετάφραση, με ελάχιστη διαφορά από τον δεύτερο, Noisy back-translation, με βαθμολογία 35.0.

Αυτόματη ερωταπόκριση

Η αυτόματη ερωταπόκριση είναι το υπό-πεδίο το οποίο ασχολείται με την ανάπτυξη συστημάτων που απαντούν αυτόματα σε ερωτήσεις που θέτουν οι άνθρωποι σε μια φυσική γλώσσα. Τα μοντέλα αυτόματης ερωταπόκρισης λειτουργούν είτε ψάχνοντας πληροφορίες στο διαδίκτυο, είτε ανατρέχοντας σε ήδη αποθηκευμένη γνώση.

Ίσως η πιο κοινή χρήση των μοντέλων αυτών γίνεται από τις εταιρίες με σκοπό να μειώσουν των φόρτο εργασίας από το προσωπικό εξυπηρέτησης πελατών. Οι μηχανές αναζήτησης χρησιμοποιούν επίσης την αυτόματη ερωταπόκριση με στόχο την αποτελεσματικότερη περιήγηση στο διαδίκτυο.

Σύμφωνα με την μετρική F1, ο αλγόριθμος XLNet+DSC φαίνεται να είναι πιο αποδοτικός με βαθμολογία 95.7. Άλλοι αλγόριθμοι με κορυφαίες επιδόσεις είναι οι T5-11B και ο LUKE με βαθμολογίες 90.06 και 89.80 αντίστοιχα.

Σύστημα συστάσεων

Τα συστήματα συστάσεων είναι συστήματα μηχανικής μάθησης που βοηθούν τους χρήστες να ανακαλύψουν νέα προϊόντα και υπηρεσίες. Είναι σαν πωλητές που γνωρίζουν, με βάση το ιστορικό και τις προτιμήσεις των χρηστών, τι είναι αυτό το οποίο πιθανώς να τους ενδιαφέρει.

Τα συστήματα συστάσεων χρησιμοποιούνται σε διάφορους τομείς, όπως είναι η δημιουργία λίστας αναπαραγωγής για υπηρεσίες βίντεο και μουσικής, προτάσεις προϊόντων σε διαδικτυακά καταστήματα ή προτάσεις περιεχομένου για πλατφόρμες κοινωνικής δικτύωσης.

Σύμφωνα με τον αρνητικού προσανατολισμού δείκτη RMSE, ο οποίος θα αναλυθεί σε άλλο κεφάλαιο, ο αλγόριθμος GLocal-K με βαθμολογία 0.822 θεωρείται ο καλύτερος στα

συστήματα συστάσεων. Άλλοι αλγόριθμοι που αξίζει να αναφερθούν είναι οι Sparse FC και CF-NADE βαθμολογία 0.824 και 0.829 αντίστοιχα.

Αναγνώριση ομιλίας

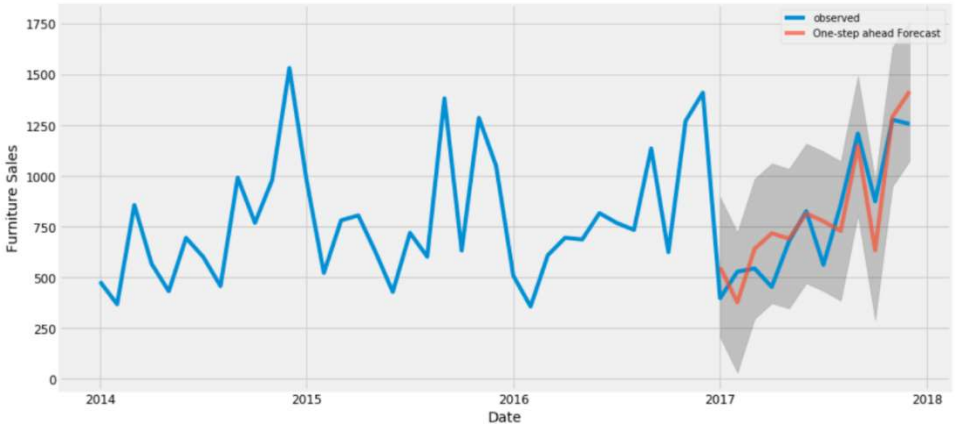
Η αναγνώριση ομιλίας, γνωστή και ως speech-to-text, αναφέρεται στην ικανότητα των συστημάτων να επεξεργάζονται ανθρώπινη ομιλία με σκοπό την καταγραφή της σε γραπτή μορφή. Ένα τυπικό λογισμικό αναγνώρισης ομιλίας έχει περιορισμένο λεξιλόγιο και μπορεί να προσδιορίζει λέξεις και φράσεις μόνο όταν μιλούνται καθαρά.

Το πεδίο αυτό της μηχανικής μάθησης χρησιμοποιείται σε μια πληθώρα εφαρμογών. Η πιο διαδεδομένη χρήση τους είναι σε εικονικούς βοηθούς, όπως είναι η Siri της Apple, για να προσφέρει μια πιο διαδραστική εμπειρία στους χρήστες. Σημαντική επίσης είναι και η χρήση τους σε συστήματα πλοήγησης στα οχήματα, ώστε να προστατέψουν τους οδηγούς από επικίνδυνες ενέργειες. Επιπρόσθετα, χρησιμοποιείται σε αυτόματους τηλεφωνητές με σκοπό την διευθέτηση απλών ζητημάτων.

Σύμφωνα με τον αρνητικού προσανατολισμού δείκτη WER (Word Error Rate), ο οποίος χρησιμοποιείται για να αξιολογήσει μοντέλα αναγνώρισης ομιλίας, ο αλγόριθμος Conformer + Wav2vec 2.0 + SpecAugment-based Noisy Student Training with Libri-Light με βαθμολογία 1.4 είναι ο πιο αποδοτικός.

Πρόβλεψη Χρονοσειρών

Η πρόβλεψη χρονοσειρών είναι μια τεχνική για την πρόβλεψη γεγονότων μέσα από μια ακολουθία χρόνου. Προβλέπει μελλοντικά γεγονότα αναλύοντας τις τάσεις του παρελθόντος, κάνοντας την υπόθεση πως οι μελλοντικές τάσεις θα είναι παρόμοιες με τις ιστορικές.



Σχήμα 1.7: Πρόβλεψη χρονοσειράς σε πωλήσεις επίπλων. Η μπλέ γραμμή αντιπροσωπεύει τις πραγματικές τιμές, ενώ η πορτοκαλί τις προβλεπόμενες. Η γκρίζα περιοχή δείχνει το "confidence level" του μοντέλου.[5]

Χρησιμοποιείται κυρίως από εταιρίες για να προβλέψουν μελλοντικές αλλαγές σε παράγοντες που δύναται να τις επηρεάσουν. Χρήση γίνεται επίσης και στο χρηματοοικονομικό τομέα όπου η πρόβλεψη των τιμών χρηματιστηρίου και κρυπτονομισμάτων είναι ύψιστης σημασίας για πολλούς οργανισμούς.

Σύμφωνα με τον αρνητικού προσανατολισμού δείκτη MAE, ο οποίος θα αναλυθεί σε επόμενο κεφάλαιο, ο αλγόριθμος με την καλύτερη βαθμολογία (0.284) είναι ο ARIMA. Ένας ακόμα αλγόριθμος που με κορυφαίες επιδόσεις θεωρείται ο Prophet της Facebook.

1.2.2 Κρυπτονομίσματα

1.2.2.1 Εισαγωγή

Το κρυπτονόμισμα είναι μια μορφή ψηφιακού περιουσιακού στοιχείου που βασίζεται σε ένα δίκτυο που διανέμεται σε μεγάλο αριθμό υπολογιστών και διασφαλίζεται με κρυπτογραφία, γεγονός που καθιστά σχεδόν αδύνατη την παραποίηση ή τη διπλή δαπάνη. Αυτή η αποκεντρωμένη δομή του επιτρέπει να υπάρχει εκτός του ελέγχου των κυβερνήσεων και των κεντρικών αρχών.[6]

Ένα καθοριστικό χαρακτηριστικό των κρυπτονομισμάτων είναι ότι δεν εκδίδονται από καμία κεντρική αρχή, δηλαδή είναι αποκεντρωμένα, καθιστώντας τα θεωρητικά απρόσβλητα από κυβερνητικές παρεμβάσεις ή χειρισμούς. Αντίθετα, ελέγχονται από χρήστες και αλγόριθμους υπολογιστών.

Οι συναλλαγές κρυπτονομισμάτων καταγράφονται σε ένα αποκεντρωμένο δημόσιο “βιβλίο”. Αυτό το "βιβλίο” ονομάζεται blockchain. Κάθε φορά που ένα κρυπτονόμισμα αγοράζεται ή πωλείται, η συναλλαγή προστίθεται στο blockchain - μια δημόσια βάση δεδομένων των συναλλαγών, η οποία είναι διαθέσιμη σε άλλους κατόχους κρυπτογράφησης. Οποιοσδήποτε μπορεί να συμμετάσχει στο blockchain, αλλά τα δεδομένα για μεμονωμένες συναλλαγές, και τα άτομα που συμμετέχουν σε αυτά προστατεύονται χρησιμοποιώντας κρυπτογραφία. Για κάθε συναλλαγή που προστίθεται στο blockchain, υπάρχει μια διαδικασία ψηφιακής επικύρωσης για την επαλήθευση και την πρόληψη της απάτης.[7]

Το κρυπτονόμισμα είναι μια μέθοδος πληρωμής που μπορεί να ανταλλαχθεί μέσω Διαδικτύου με αγαθά και υπηρεσίες. Πολλές εταιρείες έχουν εκδώσει τα δικά τους κρυπτονομίσματα, και αυτά μπορούν να ανταλλαχθούν ειδικά για το αγαθό ή την υπηρεσία που παρέχει η εταιρεία. Μοιάζει πολύ με την πολιτική που ακολουθεί ένα καζίνο, δηλαδή για να χρησιμοποιήσει κάποιος τις υπηρεσίες που παρέχει αυτό, θα πρέπει πρώτα να αγοράσει τις ειδικές μάρκες του συγκεκριμένου καζίνου. Η ανταλλαγή πραγματικού νομίσματος με κρυπτονόμισμα είναι απαραίτητη για την πρόσβαση στο αγαθό ή την υπηρεσία.

Στον πυρήνα του, το κρυπτονόμισμα είναι ένα σύστημα αξίας. Όταν οι επενδυτές αγοράζουν ένα κρυπτονόμισμα, ποντάρουν ότι η αξία αυτού του περιουσιακού στοιχείου θα αυξηθεί στο μέλλον, ακριβώς όπως οι επενδυτές του χρηματιστηρίου αγοράζουν μετοχές όταν πιστεύουν ότι η εταιρεία θα αναπτυχθεί και οι τιμές των μετοχών θα αυξηθούν.[8]

1.2.2.2 Ιστορική Αναδρομή

Πολλοί πιστεύουν ότι το κρυπτονόμισμα είναι μια έννοια που αναπτύχθηκε και ξεκίνησε την τελευταία δεκαετία περίπου, αλλά η ιστορία του κρυπτονομίσματος φτάνει μέχρι το 1983. Αυτά τα εικονικά νομίσματα μπορούν να εντοπιστούν σε έναν άνθρωπο: τον κρυπτογράφο David Chaum.

Ο Chaum ανέπτυξε το πρώτο ψηφιακό νόμισμα με ένα σύστημα συναλλαγών που ονομάζεται eCash. Αργότερα, το 1995, το εφάρμοσε μέσω Digicash, μια πρώιμη μορφή κρυπτογραφικών ηλεκτρονικών πληρωμών που απαιτούσε λογισμικό χρήστη για να αποσύρει σημειώσεις από μια τράπεζα και να ορίσει συγκεκριμένα κρυπτογραφημένα κλειδιά προτού αποσταλεί σε παραλήπτη. Το θεμελιώδες στοιχείο για το DigiCash (και ένα που παρέμεινε θεμελιώδες στα κρυπτονομίσματα που ακολούθησαν τα βήματα του DigiCash) είναι ότι οι συναλλαγές ήταν

ανώνυμες και πραγματοποιήθηκαν σε δημόσιο δίκτυο, δηλαδή δεν ήταν ανιχνεύσιμο από την εκδότρια τράπεζα, την κυβέρνηση ή οποιοδήποτε τρίτο μέρος.

Το 1998, ο Wei Dai δημοσίευσε μια περιγραφή του "b-money", ενός ανώνυμου, διανεμημένου ηλεκτρονικού συστήματος μετρητών. Λίγο αργότερα, ο Nick Szabo δημιούργησε το "bit gold". Όπως το bitcoin και τα άλλα νομίσματα που θα το ακολουθούσαν, το bit gold ήταν ένα ηλεκτρονικό σύστημα νομισμάτων που απαιτούσε από τους χρήστες να ολοκληρώσουν μια απόδειξη εργασίας (proof-of-work) με τις λύσεις να συγκεντρώνονται κρυπτογραφικά και στη συνέχεια να δημοσιεύονται. Το proof-of-wrok (POW) είναι ουσιαστικά η απόδειξη σε άλλους χρήστες ότι έχει δαπανηθεί το ποσό μίας συγκεκριμένης υπολογιστικής προσπάθειας.

Το πρώτο αποκεντρωμένο κρυπτονόμισμα, το bitcoin, δημιουργήθηκε το 2009 από τον προγραμματιστή με το ψευδώνυμο Satoshi Nakamoto. Το bitcoin χρησιμοποιεί το SHA-256, μια κρυπτογραφική συνάρτηση κατακερματισμού, ως σχέδιο απόδειξης εργασίας.

Τον Απρίλιο του 2011, το Namecoin δημιουργήθηκε ως μια προσπάθεια σχηματισμού ενός αποκεντρωμένου DNS, κάτι που θα έκανε πολύ δύσκολη τη λογοκρισία στο διαδίκτυο.

Λίγο αργότερα, τον Οκτώβριο του 2011, το Litecoin κυκλοφόρησε. Ήταν το πρώτο επιτυχημένο κρυπτονόμισμα που χρησιμοποίησε το scrypt ως λειτουργία hash αντί του SHA-256.

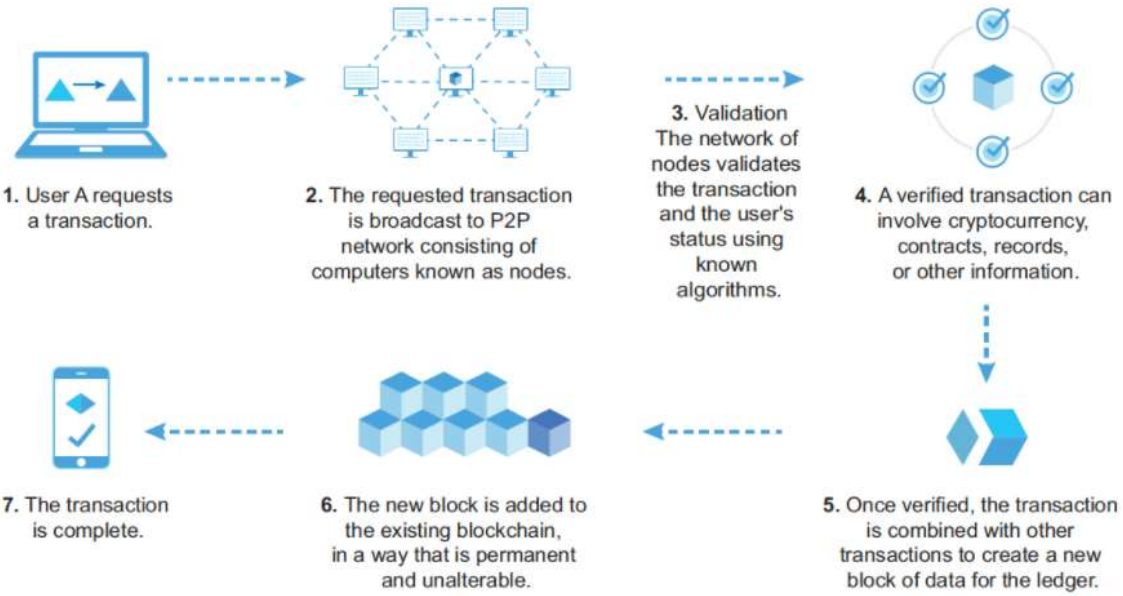
Δημιουργήθηκαν πολλά άλλα νομίσματα, αν και λίγα ήταν επιτυχημένα, καθώς δεν έφεραν πολλά στον τρόπο της τεχνικής καινοτομίας. Στις 6 Αυγούστου 2014, το Ηνωμένο Βασίλειο ανακοίνωσε ότι το Υπουργείο Οικονομικών του είχε αναλάβει να μελετήσει τα κρυπτονομίσματα και ποιος ρόλος, εάν υπάρχει, μπορούν να παίξουν στην οικονομία του Ηνωμένου Βασιλείου. Η μελέτη θα έπρεπε επίσης να εξετάσει εάν θα ήταν απαραίτητη η λήψη ρυθμίσεων και κανονισμών όσων αφορά τα κρυπτονομίσματα.

Εκπρόσωποι των κεντρικών τραπεζών έχουν δηλώσει ότι η υιοθέτηση νομισμάτων όπως το bitcoin αποτελεί σημαντική πρόκληση για την ικανότητα των κεντρικών τραπεζών να επηρεάζουν την τιμή της πίστωσης για ολόκληρη την οικονομία. Έχουν επίσης δηλώσει ότι καθώς το εμπόριο χρησιμοποιώντας νομίσματα γίνεται πιο δημοφιλές, να είναι απώλεια της εμπιστοσύνης των καταναλωτών στα παραστατικά νομίσματα. Ο Gareth Murphy, ανώτερος αξιωματούχος της κεντρικής τράπεζας δήλωσε ότι «η ευρεία χρήση [κρυπτονομίσματος] θα καθιστούσε επίσης πιο δύσκολο για τις στατιστικές υπηρεσίες να συλλέξουν δεδομένα σχετικά με την οικονομική δραστηριότητα, τα οποία χρησιμοποιούνται από τις κυβερνήσεις για να

κατευθύνουν την οικονομία». Προειδοποίησε ότι τα εικονικά νομίσματα θέτουν μια νέα πρόκληση στον έλεγχο των κεντρικών τραπεζών στις σημαντικές λειτουργίες της νομισματικής και συναλλαγματικής πολιτικής.[9]

1.2.2.3 Η Τεχνολογία Blockchain

Το blockchain είναι ένα αποκεντρωμένο, δημόσιο "βιβλίο" ή αλλιώς μια λίστα συναλλαγών κρυπτονομισμάτων. Τα ολοκληρωμένα block, που αποτελούνται από τις πιο πρόσφατες συναλλαγές, καταγράφονται και προστίθενται στο blockchain. Αποθηκεύονται με χρονολογική σειρά ως ανοιχτή, μόνιμη και επαληθεύσιμη εγγραφή. Κάθε block περιέχει συνήθως έναν δείκτη κατακερματισμού ως σύνδεσμο προς ένα προηγούμενο block, μια χρονική σήμανση και δεδομένα συναλλαγών. Από σχεδίαση, τα blockchains είναι εγγενώς ανθεκτικά στην τροποποίηση των δεδομένων.



Σχήμα 1.8: Λειτουργία μίας συναλλαγής που χρησιμοποιεί την τεχνολογία blockchain.[10]

Όταν πραγματοποιείται μια συναλλαγή κρυπτονομίσματος, η συναλλαγή αυτή αποστέλλεται σε όλους τους χρήστες που φιλοξενούν ένα αντίγραφο του blockchain. Συγκεκριμένοι τύποι χρηστών που ονομάζονται "miners" προσπαθούν στη συνέχεια να λύσουν ένα κρυπτογραφικό παζλ (χρησιμοποιώντας λογισμικό) που τους επιτρέπει να προσθέσουν ένα block συναλλαγών στο "βιβλίο". Όποιος λύσει το παζλ παίρνει πρώτα μερικά νομίσματα ως ανταμοιβή (λαμβάνουν επίσης τέλη συναλλαγής που πληρώνονται από αυτούς που δημιούργησαν τις συναλλαγές).

Μερικές φορές οι miners συγκεντρώνουν υπολογιστική ισχύ και μοιράζονται τα νέα νομίσματα. Ο αλγόριθμος βασίζεται στη συναίνεση. Εάν η πλειοψηφία των χρηστών που προσπαθούν να λύσουν το παζλ υποβάλλουν όλοι τα ίδια δεδομένα συναλλαγών, τότε επιβεβαιώνεται ότι οι συναλλαγές είναι σωστές. Επιπλέον, η ασφάλεια του blockchain βασίζεται στην κρυπτογραφία. Κάθε block συνδέεται με τα δεδομένα του τελευταίου block μέσω μονόδρομων κρυπτογραφικών κωδικών που ονομάζονται hash και έχουν σχεδιαστεί για να κάνουν πολύ δύσκολη την επέμβαση στο blockchain.

Η προσφορά νέων νομισμάτων ως ανταμοιβή, η δυσκολία στο σπάσιμο των κρυπτογραφικών παζλ και το ποσό της προσπάθειας που θα χρειαστεί για να προσθεθούν εσφαλμένα δεδομένα στο blockchain, συμβάλλει στην ασφάλεια απέναντι σε κακόβουλους χρήστες.

Η έννοια του blockchain αποδίδεται στον ιδρυτή του bitcoin, Satoshi Nakamoto. Αυτή η ιδέα υπήρξε η έμπνευση για άλλες εφαρμογές πέρα από τα ψηφιακά μετρητά και το κρυπτονόμισμα.[11]

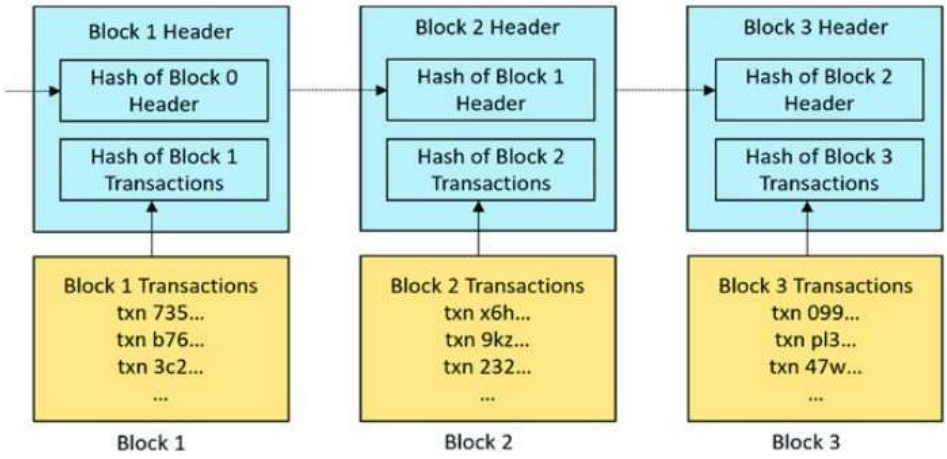
1.2.2.4 Τα συστατικά του Blockchain

Το Blockchain αποτελείται από τρεις σημαντικές έννοιες: Τα **block**, τους **κόμβους** και τους **miners**.

Block:

Κάθε αλυσίδα αποτελείται από πολλά blocks και κάθε block έχει δύο βασικά στοιχεία:

- Ένα block header, το οποίο περιέχει το hash, το οποίο είναι ένας αριθμός 256-bit. Πρέπει να ξεκινά με έναν τεράστιο αριθμό μηδενικών (δηλαδή να είναι εξαιρετικά μικρός). Έναν ακέραιο αριθμό 32-bit που ονομάζεται nonce. Το nonce δημιουργείται τυχαία όταν δημιουργείται ένα block και είναι συνδεδεμένο με το hash. Το hash του προηγούμενου block και διάφορα χαρακτηριστικά που περιγράφουν το συγκεκριμένο block.
- Μια λίστα που περιέχει όλες τις συναλλαγές οι οποίες πραγματοποιούνται κατα την διάρκεια εξόρυξης του συγκεκριμένου block



Σχήμα 1.9: Δομή ενός τυπικού block.[12]

Όταν δημιουργείται το πρώτο block μιας αλυσίδας, ένα nonce δημιουργεί το κρυπτογραφικό hash. Τα δεδομένα στο block θεωρούνται υπογεγραμμένα και για πάντα συνδεδεμένα με το nonce και το hash, εκτός αν εξορύσσονται.

Miners:

Οι miners δημιουργούν νέα block στην αλυσίδα μέσω μιας διαδικασίας που ονομάζεται εξόρυξη.

Σε ένα blockchain κάθε block έχει το δικό του μοναδικό nonce και hash, αλλά αναφέρεται και το hash του προηγούμενου block στην αλυσίδα, οπότε η εξόρυξη ενός block δεν είναι εύκολη, ειδικά σε μεγάλες αλυσίδες.

Οι miners χρησιμοποιούν ειδικό λογισμικό για την επίλυση του απίστευτα πολύπλοκου μαθηματικού προβλήματος της εύρεσης ενός nonce που δημιουργεί ένα αποδεκτό hash. Επειδή το nonce είναι μόνο 32 bit και το hash είναι 256, υπάρχουν περίπου τέσσερα δισεκατομμύρια πιθανοί συνδυασμοί nonce-hash που πρέπει να εξορυχθούν πριν βρεθεί το σωστό. Όταν συμβαίνει αυτό, οι miners λένε ότι βρήκαν το "golden nonce" και το block τους προστίθεται στην αλυσίδα.

Όταν ένα block εξορύσσεται επιτυχώς, η αλλαγή γίνεται αποδεκτή από όλους τους κόμβους του δικτύου και ο miner ανταμείβεται οικονομικά.

Κόμβοι:

Μια από τις πιο σημαντικές έννοιες στην τεχνολογία του blockchain είναι η αποκέντρωση. Κανένας υπολογιστής ή οργανισμός δεν μπορεί να κατέχει την αλυσίδα. Αντ 'αυτού, είναι ένα κατανεμημένο δημόσιο “βιβλίο” μέσω των κόμβων που συνδέονται με την αλυσίδα. Οι κόμβοι μπορεί να είναι κάθε είδους ηλεκτρονική συσκευή που διατηρεί αντίγραφα του blockchain και διατηρεί τη λειτουργία του δικτύου.

Κάθε κόμβος έχει το δικό του αντίγραφο του blockchain και το δίκτυο πρέπει να εγκρίνει αλγοριθμικά κάθε πρόσφατα αποκλεισμένο block για να ενημερώνεται, να εμπιστεύεται και να επαληθεύεται η αλυσίδα. Δεδομένου ότι τα blockchains είναι “διαφανή”, κάθε ενέργεια στο “βιβλίο” μπορεί εύκολα να ελεγχθεί και να προβληθεί. Σε κάθε συμμετέχοντα δίνεται ένας μοναδικός αριθμός αναγνώρισης που δείχνει τις συναλλαγές του.

Ο συνδυασμός δημόσιων πληροφοριών με ένα σύστημα ελέγχου και ισορροπίας βοηθά το blockchain να διατηρήσει την ακεραιότητα και δημιουργεί εμπιστοσύνη στους χρήστες.[13][14]

1.2.2.5 Η λειτουργία του Blockchain:

Το πώς λειτουργεί το blockchain, για να τεθεί απλά, είναι μέσω μιας σειράς αρχείων δεδομένων με χρονοσφραγίδα, που διαχειρίζεται μια ομάδα υπολογιστών που δεν ανήκουν σε καμία οντότητα, άτομο ή εταιρεία. Τα blocks δεδομένων συνδέονται μεταξύ τους με τη χρήση κρυπτογραφίας, σχηματίζοντας την ομώνυμη "αλυσίδα".

Τα blockchain διαχειρίζονται κυρίως αυτόνομα και χρησιμοποιούνται σε δίκτυα peer-to-peer για την ανταλλαγή δεδομένων μεταξύ συνδεδεμένων ομάδων. Όπως είναι η φύση του blockchain, δεν υπάρχει ανάγκη για διαχειριστή. Οι χρήστες συνεργάζονται ως συλλογικός διαχειριστής. Μια άλλη μορφή blockchain, γενικά γνωστή ως private blockchain, επιτρέπει σε έναν οργανισμό να δημιουργεί και να διαχειρίζεται δίκτυα συναλλαγών που μπορούν να χρησιμοποιηθούν με συνεργάτες, είτε εσωτερικά είτε από τη μία εταιρεία στην άλλη.

Κάθε συναλλαγή blockchain περνά από τα ίδια βήματα ανεξάρτητα από το αν χρησιμοποιείται για οικονομικές συναλλαγές. Η βασική αρχή της λειτουργίας οποιουδήποτε blockchain μπορεί να χωριστεί σε τέσσερα διαφορετικά, συνεχόμενα βήματα:

1. **Γίνεται εγγραφή για κάθε συναλλαγή.** Αυτή η εγγραφή, η οποία περιέχει ορισμένα στοιχεία των ατόμων που πραγματοποιούν τη συναλλαγή πιστοποιείται με την ψηφιακή υπογραφή του καθενός.
2. **Κάθε συναλλαγή επαληθεύεται για να διασφαλιστεί η εγκυρότητά της.** Αυτή η διαδικασία επαλήθευσης ολοκληρώνεται από τους υπολογιστές που είναι συνδεδεμένοι στο δίκτυο, καθένας από τους οποίους ελέγχει ανεξάρτητα για να διασφαλίσει ότι η συναλλαγή είναι νόμιμη. Επειδή αυτή είναι μια αποκεντρωμένη διαδικασία, σημαίνει ότι κάθε κόμβος στο δίκτυο πρέπει να συμφωνήσει πριν ολοκληρωθεί η διαδικασία.
3. **Μόλις επαληθευτεί, κάθε συναλλαγή προστίθεται σε ένα block που κατακερματίζεται.** Ο κατακερματισμός διασφαλίζει επίσης την ακεραιότητα των δεδομένων για να δείξει ότι δεν έχουν τροποποιηθεί από τότε που καταγράφηκαν στο block.
4. **Μόλις ολοκληρωθεί, το block προστίθεται στο τέλος του blockchain.** Αυτό μας οδηγεί στο τέλος της διαδικασίας δημιουργίας και επαλήθευσης blockchain. Μόλις ολοκληρωθεί ένα block, σύντομα θα ακολουθήσει ένα άλλο block, συνήθως μέσα σε λίγα λεπτά.

ΚΕΦΑΛΑΙΟ 2

ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΗΣΗ

2.1 Προβλεπτική Ανάλυση

2.1.1 Εισαγωγή

Μέχρι πρόσφατα η διαδικασία λήψης αποφάσεων και ιδίως για στρατηγικές που χρησιμοποιούνται για επιχειρηματικές αποφάσεις ήταν καθαρά και μόνο ευθύνη οικονομολόγων και άλλων ειδικών οι οποίοι ναι μεν βασίζονταν σε κάποιο ιστορικό και άλλα στοιχεία, αλλά αναπόφευκτα έμπαινε και η διαίσθηση στην μέση, κάτι πολύ υποκειμενικό που οδηγούσε πολλές φορές σε εσφαλμένες αποφάσεις. Η ανάγκη λοιπόν για μία μέθοδο βασισμένη σε στοιχεία του παρελθόντος η οποία θα απείχε από οποιοδήποτε προαίσθημα ήταν ξεκάθαρη.

Η προβλεπτική ανάλυση είναι ένας όρος που χρησιμοποιείται κυρίως σε στατιστικές και αναλυτικές τεχνικές. Αυτός ο όρος αντλείται από στατιστικές, μηχανική μάθηση, τεχνικές βάσεων δεδομένων και τεχνικές βελτιστοποίησης. Έχει ρίζες στην κλασική στατιστική. Προβλέπει το μέλλον αναλύοντας τρέχοντα και ιστορικά δεδομένα. Τα μελλοντικά γεγονότα και η συμπεριφορά των μεταβλητών μπορούν να προβλεφθούν χρησιμοποιώντας τα μοντέλα προβλεπτικών αναλύσεων. Τα μοτίβα ιστορικών δεδομένων αξιοποιούνται από αυτά τα μοντέλα για να βρουν τη λύση για πολλά επιχειρηματικά και επιστημονικά προβλήματα. Αυτά τα μοντέλα είναι χρήσιμα για τον εντοπισμό του κινδύνου και των ευκαιριών για κάθε μεμονωμένο πελάτη, υπάλληλο ή διευθυντή ενός οργανισμού. Το κεφάλαιο αυτό εξετάζει την **διαδικασία**, τις **τεχνικές** και τις **κατηγορίες** των μοντέλων προβλεπτικής ανάλυσης.

2.1.2 Κατανοώντας την προβλεπτική ανάλυση

Η προβλεπτική ανάλυση δεν είναι πρωτοφανές φαινόμενο. Η ιδέα υπάρχει εδώ και αρκετό καιρό και έχει χρησιμοποιηθεί με επιτυχία από μεγάλες εταιρείες που δραστηριοποιούνται σε μικρό αριθμό βιομηχανικών τομέων. Ωστόσο, τα οφέλη και οι δυνατότητες της προβλεπτικής ανάλυσης εκτιμήθηκαν πρόσφατα λόγω του φαινομένου των μεγάλων δεδομένων. Αυτή η νέα

εκτίμηση της προβλεπτικής ανάλυσης συνδυάζεται με την επιθυμία πολλών εταιρικών οργανώσεων για να προβλέπουν μελλοντικά αποτελέσματα ή γεγονότα με υψηλό επίπεδο εμπιστοσύνης.[15]

2.1.3 Διαδικασία προβλεπτικής ανάλυσης

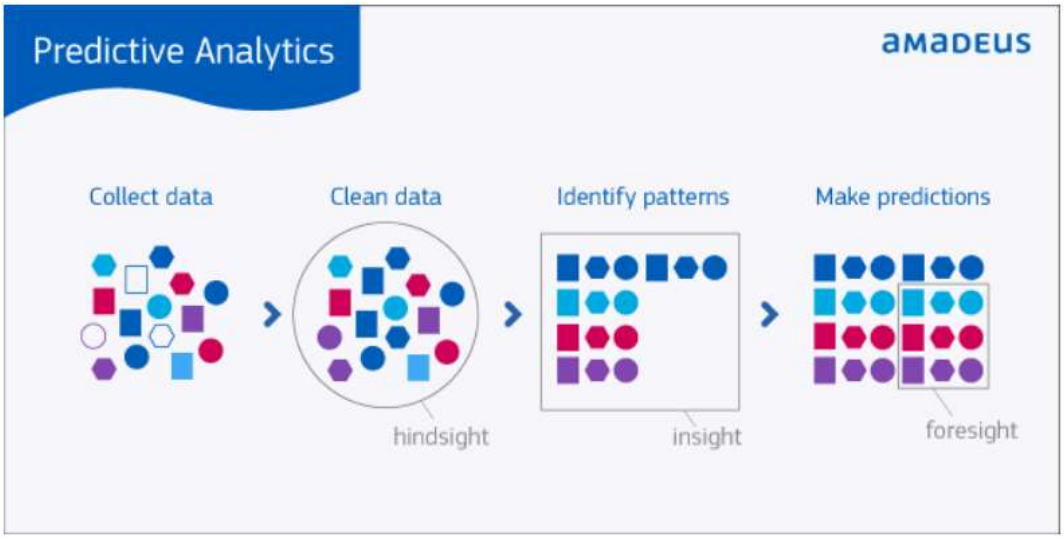
Για να αναπτυχθεί ένα μοντέλο προβλεπτικής ανάλυσης θα πρέπει να γίνει σαφές ποιος είναι ο στόχος της πρόβλεψης. Άρα αρχικά θα πρέπει να γίνει μια σωστή συλλογή απαιτήσεων και να προσδιοριστεί ποια δεδομένα του πελάτη θα απαιτηθούν για την ανάπτυξη του μοντέλου.

Αφού γνωστοποιηθούν οι απαιτήσεις του πελάτη, ο αναλυτής θα συλλέξει τα σύνολα δεδομένων, που μπορεί να προέρχονται από διαφορετικές πηγές, που απαιτούνται για την ανάπτυξη του μοντέλου πρόβλεψης.

Έπειτα, οι αναλυτές δεδομένων αναλύουν τα συλλεγμένα δεδομένα και τα προετοιμάζουν για ανάλυση και για χρήση στο μοντέλο. Τα μη δομημένα δεδομένα μετατρέπονται σε δομημένη μορφή σε αυτό το βήμα. Όταν όλα τα δεδομένα είναι διαθέσιμα σε δομημένη μορφή, τότε ελέγχεται η ποιότητά τους. Υπάρχει πιθανότητα να υπάρχουν λανθασμένα δεδομένα στο κύριο σύνολο δεδομένων, κάτι που θα πρέπει να αντιμετωπιστεί. Η αποτελεσματικότητα του μοντέλου πρόβλεψης εξαρτάται σε μεγάλο βαθμό από την ποιότητα των δεδομένων.

Η διαδικασία της προβλεπτικής ανάλυσης χρησιμοποιεί πολλές στατιστικές τεχνικές και τεχνικές μηχανικής μάθησης. Η θεωρία πιθανοτήτων και η ανάλυση παλινδρόμησης είναι οι πιο σημαντικές τεχνικές που χρησιμοποιούνται ευρέως στην ανάλυση. Ομοίως, τα τεχνητά νευρωνικά δίκτυα και το δέντρο αποφάσεων είναι εργαλεία της μηχανικής μάθησης που χρησιμοποιούνται ευρέως σε πολλές εργασίες προβλεπτικής ανάλυσης. Όλα τα μοντέλα προβλεπτικής ανάλυσης βασίζονται σε στατιστικές και/ή τεχνικές μηχανικής μάθησης. Ως εκ τούτου, οι αναλυτές εφαρμόζουν τις έννοιες της στατιστικής και της μηχανικής μάθησης προκειμένου να αναπτύξουν προβλεπτικά μοντέλα.

Τέλος, αναπτύσσεται ένα μοντέλο βασισμένο σε στατιστικές τεχνικές και τεχνικές μηχανικής μάθησης και φυσικά το σύνολο δεδομένων που έχει συλλεχθεί. Μετά την ανάπτυξη, δοκιμάζεται στα δεδομένα δοκιμής, τα οποία είναι μέρος του κύριου συνόλου δεδομένων και χρησιμοποιείται για να ελεγχθεί η εγκυρότητα του μοντέλου. Εάν είναι επιτυχές, το μοντέλο λέγεται ότι είναι κατάλληλο, και πλέον μπορεί να κάνει προβλέψεις για τα νέα δεδομένα που εισάγονται ως είσοδος στο σύστημα.[16]



Σχήμα 2.1: Διαδικασία προβλεπτικής ανάλυσης.[17]

2.1.4 Κατηγορίες μοντέλων προβλεπτικής ανάλυσης

Η γενική έννοια της προβλεπτικής ανάλυσης είναι η προβλεπτική μοντελοποίηση, η οποία ουσιαστικά είναι η βαθμολόγηση δεδομένων χρησιμοποιώντας μοντέλα πρόβλεψης. Γενικά όμως, χρησιμοποιείται ως όρος που αναφέρεται στους κλάδους που σχετίζονται με την ανάλυση. Αυτοί οι κλάδοι περιλαμβάνουν τη διαδικασία ανάλυσης δεδομένων και χρησιμοποιούνται στη λήψη επιχειρηματικών αποφάσεων. Υπάρχουν διαφορετικά μοντέλα που έχουν αναπτυχθεί για συγκεκριμένες λειτουργίες σχεδιασμού και μπορούν να ενταχθούν στις εξής κατηγορίες:

Μοντέλα ταξινόμησης:

Ένα από τα πιο κοινά μοντέλα προβλεπτικής ανάλυσης είναι τα μοντέλα ταξινόμησης. Αυτά τα μοντέλα λειτουργούν κατηγοριοποιώντας πληροφορίες με βάση ιστορικά δεδομένα. Τα μοντέλα ταξινόμησης χρησιμοποιούνται σε διαφορετικούς κλάδους επειδή μπορούν εύκολα να επανεκπαιδευτούν με νέα δεδομένα, γεγονός που εξηγεί γιατί είναι τόσο συνηθισμένα σε σύγκριση με άλλα μοντέλα.[18]

Μοντέλα Outlier:

Ενώ τα μοντέλα ταξινόμησης λειτουργούν με ιστορικά δεδομένα, το μοντέλα outlier λειτουργούν με ανώμαλες καταχωρήσεις δεδομένων σε ένα σύνολο δεδομένων. Όπως υποδηλώνει το όνομα, τα ανώμαλα δεδομένα αναφέρονται σε δεδομένα που αποκλίνουν από τον κανόνα. Λειτουργεί με τον εντοπισμό ασυνήθιστων δεδομένων, είτε μεμονωμένα είτε σε σχέση με διαφορετικές κατηγορίες και αριθμούς. Τα Outlier μοντέλα είναι χρήσιμα σε βιομηχανίες όπου ο εντοπισμός ανωμαλιών μπορεί να εξοικονομήσει στους οργανισμούς εκατομμύρια δολάρια, συγκεκριμένα στη λιανική και τη χρηματοοικονομική. Ένας λόγος για τον οποίο τα μοντέλα προβλεπτικής ανάλυσης είναι τόσο αποτελεσματικά στην ανίχνευση απάτης είναι επειδή μπορούν να χρησιμοποιηθούν Outlier μοντέλα για τον εντοπισμό ανωμαλιών. Δεδομένου ότι μια απάτη είναι μια απόκλιση από τον κανόνα, ένα Outlier μοντέλο είναι πιο πιθανό να το προβλέψει πριν συμβεί.[19]

Μοντέλα Χρονοσειρών:

Ενώ τα μοντέλα ταξινόμησης εστιάζουν σε ιστορικά δεδομένα, τα Outlier μοντέλα επικεντρώνονται σε δεδομένα ανωμαλιών. Το μοντέλο Χρονοσειράς εστιάζει σε δεδομένα όπου ο χρόνος είναι η παράμετρος εισόδου. Το μοντέλο Χρονοσειρών λειτουργεί χρησιμοποιώντας διαφορετικά σημεία δεδομένων (λαμβανόμενα από τα δεδομένα του παρελθόντος) για να αναπτύξει μια αριθμητική μέτρηση που θα προβλέπει τις τάσεις εντός συγκεκριμένης μελλοντικής περιόδου.

Εάν μία επιχείρηση έχει την ανάγκη να δει πώς αλλάζει μια συγκεκριμένη μεταβλητή με την πάροδο του χρόνου, τότε χρειάζονται ένα μοντέλο προβλεπτικής ανάλυσης χρονοσειρών. Για παράδειγμα, εάν ένας ιδιοκτήτης μιας μικρής επιχείρησης θέλει να μετρήσει τις πωλήσεις τα τελευταία τέσσερα τρίμηνα, τότε χρειάζεται ένα μοντέλο χρονοσειράς. Ένα μοντέλο χρονοσειράς είναι ανώτερο από τις συμβατικές μεθόδους υπολογισμού της προόδου μιας μεταβλητής, επειδή μπορεί να λάβει υπόψη παράγοντες που θα μπορούσαν να επηρεάσουν τις μεταβλητές, όπως οι εποχές.[20]

Μοντέλα Συσταδοποίησης:

Το μοντέλο Συσταδοποίησης λαμβάνει δεδομένα και τα ταξινομεί σε διαφορετικές ομάδες ή συστάδες με βάση κοινά χαρακτηριστικά. Η δυνατότητα διαίρεσης δεδομένων σε διαφορετικά σύνολα δεδομένων με βάση συγκεκριμένα χαρακτηριστικά είναι ιδιαίτερα χρήσιμη σε

ορισμένες εφαρμογές, όπως το μάρκετινγκ. Για παράδειγμα, οι έμποροι μπορούν να διαιρέσουν μια δυνητική βάση πελατών με βάση κοινά χαρακτηριστικά. Στη συσταδοποίηση οι συστάδες δεν είναι προκαθορισμένες αλλά προσδιορίζονται από τα δεδομένα.[21]

2.1.5 Τεχνικές προβλεπτικής ανάλυσης

Όλα τα μοντέλα προβλεπτικής ανάλυσης ομαδοποιούνται σε μοντέλα ταξινόμησης και μοντέλα παλινδρόμησης.

Η ταξινόμηση είναι μια διαδικασία εύρεσης μιας συνάρτησης που βοηθά στη διαίρεση του συνόλου δεδομένων σε κλάσεις με βάση διαφορετικές παραμέτρους. Στην ταξινόμηση, ένα πρόγραμμα υπολογιστή εκπαιδεύεται στο σύνολο δεδομένων εκπαίδευσης και κατηγοριοποιεί τα δεδομένα σε διαφορετικές κατηγορίες.[22]

Οι αλγόριθμοι παλινδρόμησης προβλέπουν μια συνεχή τιμή με βάση τις μεταβλητές εισόδου. Ο κύριος στόχος των μοντέλων παλινδρόμησης είναι η εκτίμηση μιας συνάρτησης χαρτογράφησης με βάση τις μεταβλητές εισόδου και εξόδου.[23]

Παρακάτω παρουσιάζονται τέσσερις σημαντικές τεχνικές που χρησιμοποιούνται ευρέως στην ανάπτυξη των προβλεπτικών μοντέλων.

Δέντρο αποφάσεων (Decision Tree)

Ένα δέντρο αποφάσεων είναι ένα μοντέλο ταξινόμησης, αλλά μπορεί να χρησιμοποιηθεί και σε μοντέλα παλινδρόμησης. Είναι ένα μοντέλο που μοιάζει με δέντρο και σχετίζει τις αποφάσεις και τις πιθανές συνέπειές τους. Οι συνέπειες μπορεί να είναι το αποτέλεσμα γεγονότων, το κόστος των πόρων ή η χρησιμότητα. Στη δομή που μοιάζει με δέντρο, κάθε κλαδί αντιπροσωπεύει μια επιλογή μεταξύ πολλών εναλλακτικών και κάθε φύλλο αντιπροσωπεύει μια απόφαση. Με βάση τις κατηγορίες μεταβλητών εισόδου, χωρίζει τα δεδομένα σε υποσύνολα. Η ευκολία κατανόησης και ερμηνείας καθιστά τα δέντρα αποφάσεων δημοφιλή στη χρήση.[24]

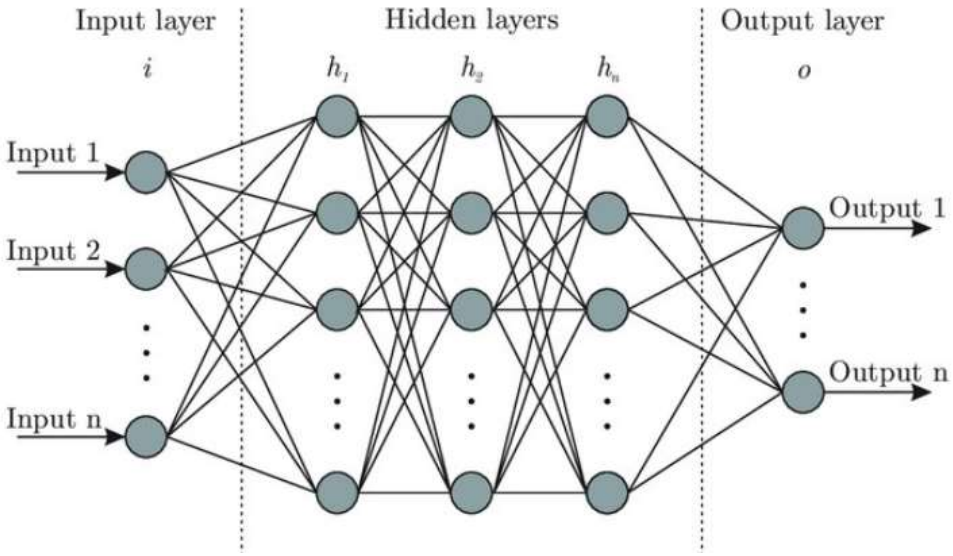
Μοντέλο παλινδρόμησης (Regression Model)

Η παλινδρόμηση είναι μια από τις πιο δημοφιλείς στατιστικές τεχνικές που εκτιμούν τη σχέση μεταξύ των μεταβλητών. Μοντελοποιεί τη σχέση μεταξύ εξαρτημένης μεταβλητής και μίας ή περισσοτέρων ανεξάρτητων μεταβλητών. Αναλύει πώς επηρεάζεται η τιμή της εξαρτημένης μεταβλητής στην αλλαγή των τιμών των ανεξάρτητων μεταβλητών στη σχέση που διαμορφώνεται.[25]

Τεχνητό νευρωνικό δίκτυο (Artificial Neural Network)

Ένα τεχνητό νευρωνικό δίκτυο βασίζεται σε μια συλλογή συνδεδεμένων μονάδων ή κόμβων που ονομάζονται τεχνητοί νευρώνες, οι οποίοι προσομοιάζουν τους νευρώνες σε έναν βιολογικό εγκέφαλο. Κάθε σύνδεση, όπως οι συνάψεις σε έναν βιολογικό εγκέφαλο, μπορεί να μεταδώσει ένα σήμα σε άλλους νευρώνες. Ένας τεχνητός νευρώνας που λαμβάνει ένα σήμα το επεξεργάζεται και μπορεί να το μεταφέρει σε νευρώνες που συνδέονται με αυτό. Το "σήμα" σε μια σύνδεση είναι ένας πραγματικός αριθμός και η έξοδος κάθε νευρώνα υπολογίζεται από κάποια μη γραμμική συνάρτηση του αθροίσματος των εισόδων του. Οι συνδέσεις ονομάζονται ακμές.

Οι νευρώνες έχουν συνήθως ένα βάρος που προσαρμόζεται καθώς προχωρά η μάθηση. Το βάρος αυξάνει ή μειώνει την ισχύ του σήματος σε μια σύνδεση. Οι νευρώνες μπορεί να έχουν ένα κατώφλι τέτοιο ώστε ένα σήμα να αποστέλλεται μόνο εάν το συνολικό σήμα ξεπεράσει αυτό το όριο. Τυπικά, οι νευρώνες συγκεντρώνονται σε στρώματα. Διαφορετικά στρώματα μπορούν να πραγματοποιήσουν διαφορετικούς μετασχηματισμούς στις εισόδους τους. Τα σήματα ταξιδεύουν από το πρώτο στρώμα (το στρώμα εισόδου), στο τελευταίο στρώμα (το στρώμα εξόδου).[26]



Σχήμα 2.2: Μορφή ενός τυπικού τεχνητού νευρωνικού δικτύου.[27]

Ανάλυση χρονοσειρών (Time Series Analysis)

Η ανάλυση χρονοσειρών είναι μια στατιστική τεχνική που χρησιμοποιεί δεδομένα χρονοσειρών που συλλέγονται για μια χρονική περίοδο σε ένα συγκεκριμένο διάστημα. Συνδυάζει τις παραδοσιακές τεχνικές εξόρυξης δεδομένων και την πρόβλεψη. Προβλέπει το μέλλον μιας μεταβλητής σε μελλοντικά χρονικά διαστήματα με βάση την ανάλυση των τιμών στα προηγούμενα χρονικά διαστήματα.[28] Χρησιμοποιείται πολύ δημοφιλώς στη πρόβλεψη του χρηματιστηρίου, την πρόγνωση του καιρού και όπως θα φανεί στην συνέχεια στη πρόβλεψη τιμών των κρυπτονομισμάτων.

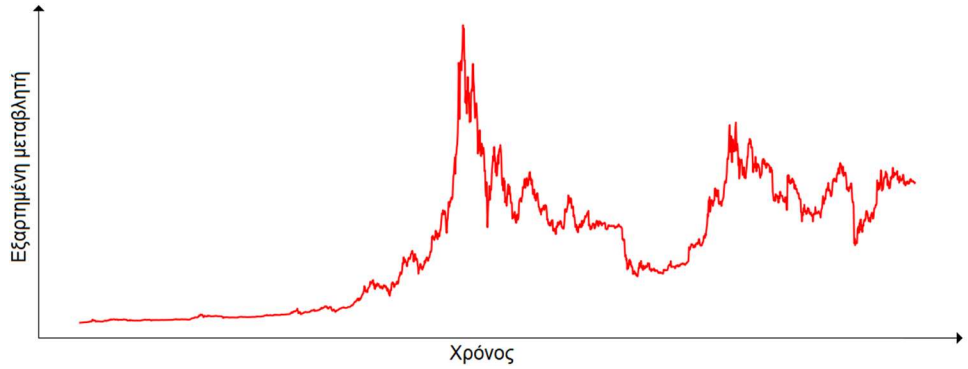
Ανακεφαλαίωση

Η προβλεπτική ανάλυση είναι η εφαρμογή δεξιοτήτων, εμπειρογνωμοσύνης και δυνατοτήτων λογισμικού για την εξαγωγή, την ανάκριση, την ανάλυση και τη μετατροπή δεδομένων σε σαφή, εύπεπτη μορφή που τροφοδοτείται σε μια διαδικασία λήψης αποφάσεων. Με τον τρόπο αυτό, η προβλεπτική ανάλυση συνδυάζει ανθρώπινες δεξιότητες και τεχνογνωσία με τεχνολογία όπως η μηχανική μάθηση προτύπων σε τρέχοντα και ιστορικά δεδομένα και η εφαρμογή αλγορίθμων όχι μόνο για τον εντοπισμό προτύπων στα δεδομένα αλλά και για την πρόβλεψη μελλοντικών πιθανοτήτων για τα δεδομένα αυτά.[29]

2.2 Χρονοσειρές

2.2.1 Εισαγωγή

Σε πολλούς τομείς, όπως ο χρηματοοικονομικός, τα δεδομένα που μελετώνται είναι συνδεδεμένα με τον χρόνο. Τι νόημα έχει η ανάλυση ή η μελέτη μίας τιμής η οποία καθημερινά αυξομειώνεται, για παράδειγμα κάποιας εταιρείας στο χρηματιστήριο , εάν δεν ληφθεί υπ' όψη ο χρόνος. Γεννιέται λοιπόν έτσι μια συσχέτιση ανάμεσα σε δεδομένα, όπως είναι και οι τιμές κρυπτονομισμάτων, και στον χρόνο. Η συσχέτιση αυτή ονομάζεται χρονοσειρά.



Σχήμα 2.3: Απεικόνιση μίας χρονοσειράς.

Συγκεκριμένα, με τον όρο χρονοσειρά θεωρούμε ένα σύνολο παρατηρήσεων σε συνάρτηση με τον χρόνο με σταθερό χρονικό βήμα. Ουσιαστικά μια χρονοσειρά μας δείχνει την εξάρτηση που έχει μία μεταβλητή x σε κάποια χρονική στιγμή t, x_t, από την ίδια μεταβλητή σε προηγούμενες χρονικές στιγμές, x_{t-1}, x_{t-2}, ...

2.2.2 Βασικά χαρακτηριστικά Χρονοσειρών

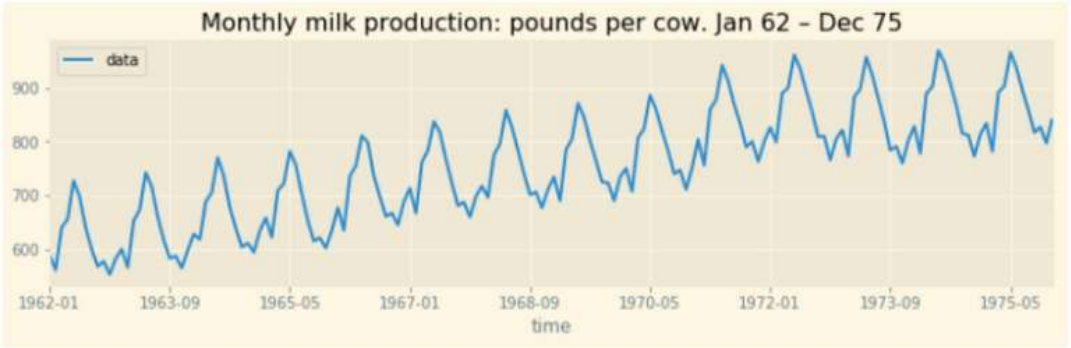
Για κάποιον ο οποίος αναλύει μια χρονοσειρά, είναι σημαντικό να έχει κατανοήσει τις έννοιες οι οποίες την συνθέτουν. Τα βασικά χαρακτηριστικά που θα αναλυθούν σε αυτό το κεφάλαιο είναι η **τάση**, η **περιοδικότητα**, οι **ακραίες τιμές**, ο **λευκός θόρυβος**, η **στασιμότητα** και η **μέση τιμή**.

<u>Ταση</u>

Η τάση είναι ένα μοτίβο στα δεδομένα που δείχνει την κίνηση μιας χρονοσειράς σε σχετικά υψηλότερες ή χαμηλότερες τιμές για μεγάλο χρονικό διάστημα. Με άλλα λόγια, παρατηρείται μια τάση όταν υπάρχει μια αυξανόμενη ή μειούμενη κλίση στις χρονοσειρές. Η τάση συνήθως συμβαίνει για κάποιο χρονικό διάστημα και μετά εξαφανίζεται, δεν επαναλαμβάνεται, δηλαδή η χρονοσειρά επιστρέφει σε μία "κανονικότητα".

Περιοδικότητα

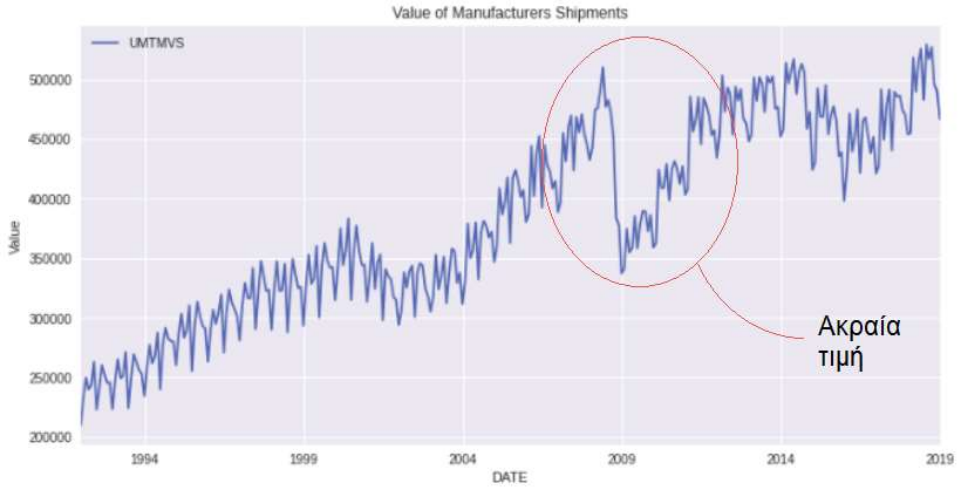
Με τον όρο περιοδικότητα εννοούμε διακριτές και προβλέψιμες διακυμάνσεις που συμβαίνουν στα δεδομένα μίας χρονοσειράς οι οποίες επαναλαμβάνονται κάθε ένα συγκεκριμένο διάστημα, όπως είναι ένα ημερολογιακό έτος. Για παράδειγμα, η ζήτηση του αρνιού βλέπει μια μεγάλη αύξηση την περίοδο του Πάσχα, όπου στην συνέχεια επιστρέφει σε μία "κανονικότητα" μέχρι να ξαναδεί μία παρόμοια αύξηση την ίδια περίοδο τον αμέσως επόμενο χρόνο.



Σχήμα 2.4: Τυπική περιοδική χρονοσειρά.[10]

Ακραίες τιμές

Οι ακραίες τιμές στα δεδομένα μίας χρονοσειρά λέγεται ότι είναι μία μεγάλη, απότομη και μη προβλέψιμη διακύμανση, η οποία δεν έχει περιοδικότητα. Αυτό που τις ξεχωρίζει από την τάση είναι ότι οι ακραίες τιμές αναφέρονται σε μία μεγαλύτερη διακύμανση και κρατάει για ένα μικρό χρονικό διάστημα. Είναι αυτό που αποκαλούμε μια ασυνέχεια στα δεδομένα. Εάν η διακύμανση αυτή παραμείνει, δηλαδή δεν είναι παροδική, τότε μιλάμε για level-shift, που είναι μια απότομη αλλαγή στη χρονοσειρά.[31]



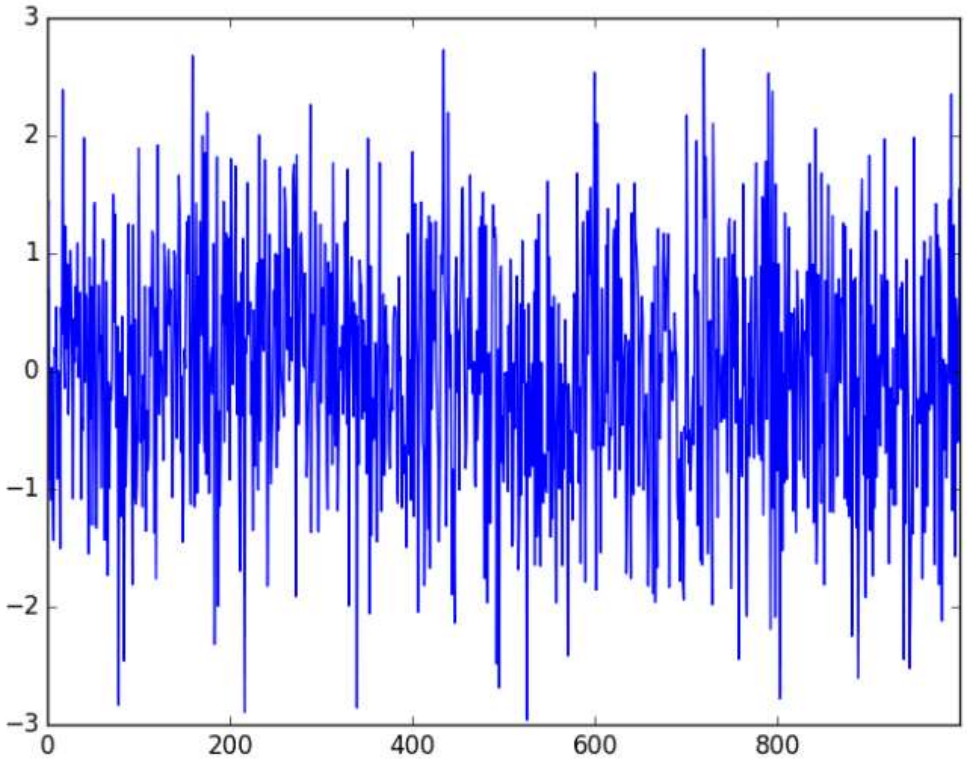
Σχήμα 2.5: Η ακραία τιμή που παρουσιάζεται στην αξία των αποστολών των κατασκευαστικών εταιριών.[32]

Λευκός Θόρυβος

Μια χρονοσειρά λέγεται ότι είναι λευκός θόρυβος εάν οι μεταβλητές είναι ανεξάρτητες και πανομοιότυπα κατανεμημένες και δίνεται από τον τύπο:

$$y_t = \varepsilon_t$$

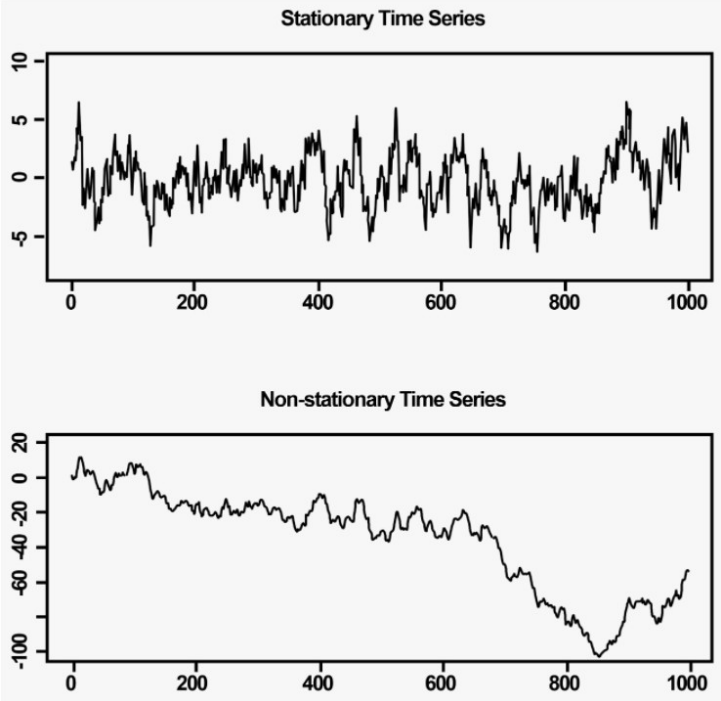
Αυτό σημαίνει ότι μια τέτοια χρονοσειρά έχει μέση τιμή 0 και όλες οι μεταβλητές έχουν την ίδια διακύμανση (σ^2) και κάθε τιμή έχει μηδενικό συσχετισμό με όλες τις άλλες τιμές της χρονοσειράς. Μια χρονοσειρά λευκού θορύβου είναι μια ακολουθία τυχαίων αριθμών και δεν μπορεί να προβλεφθεί.[33]



Σχήμα 2.6: μια τυπική χρονοσειρά λευκού θορύβου.[34]

Στασιμότητα

Στάσιμη χρονοσειρά είναι αυτή της οποίας οι ιδιότητες δεν εξαρτώνται από το χρόνο . Έτσι, οι χρονοσειρές με τάσεις ή με εποχικότητα δεν είναι στάσιμες επειδή η τάση και η εποχικότητα θα επηρεάσουν την αξία των χρονοσειρών σε διαφορετικές χρονικές στιγμές. Από την άλλη πλευρά, μια χρονοσειρά λευκού θορύβου θα είναι πάντα στάσιμη επειδή δεν έχει σημασία το πότε παρατηρείται, θα πρέπει να μοιάζει σχεδόν ίδια σε οποιαδήποτε χρονική στιγμή. Γενικά, μια στάσιμη χρονοσειρά δεν θα έχει προβλέψιμα πρότυπα μακροπρόθεσμα.[35]



Σχήμα 2.7: Διαφορά στάσιμης, από μη-στάσιμης χρονοσειράς.[36]

Μέση Τιμή

Η μέση τιμή μίας χρονοσειράς αναφέρεται στο μέσο όρο των τιμών της. Εάν οι παρατηρήσεις στη χρονοσειρά είναι y_1, y_2,…, y_T, τότε η μέση τιμή της θα είναι:

$$\bar{y} = \hat{\mu}_y = \frac{1}{T}\sum_{t=1}^{T} y_t$$

Και η διακύμανση της:

$$s^2 = \hat{\sigma}_y^2 = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \bar{y}\right)^2$$

Το μέγεθος αυτό χρησιμοποιείται κυρίως σε τεχνικές διαφοροποίησης δεδομένων όπως θα φανεί στη συνέχεια.

2.2.3 Στατιστικά μεγέθη χρονοσειράς

Για μια χρονοσειρά οι τυχαίες μεταβλητές X_t και X_s σχετίζονται με την ίδια ποσότητα που μετριέται σε διαφορετικά χρονικά σημεία. Επομένως, η εξάρτηση μεταξύ τους περιγράφεται

από τις συναρτήσεις αυτοσυνδιακύμανσης και αυτοσυσχέτισης, με το πρόθεμα «αυτο» να προστίθεται για να δηλώσει το γεγονός ότι και οι δύο τυχαίες μεταβλητές μετρούν την ίδια ποσότητα (αν και σε διαφορετικά χρονικά σημεία)

Αυτοσυνδιακύμανση (autocovariance)

Η συνδιακύμανση είναι το μέτρο του βαθμού συσχετίσεως δύο διαφορετικών μεταβλητών. Η αυτοσυνδιακύμανση είναι η συνδιακύμανση μιας μεταβλητής με τον εαυτό της σε κάποια άλλη στιγμή, μετρούμενη με χρονική υστέρηση. Η συνάρτηση αυτοσυνδιακύμανσης ορίζεται για όλα τα s, t που ανήκουν στο T ως:

$$\gamma_{s,t}=\mathrm{Cov}[X_s,X_t]=E[X_sX_t]-E[X_t]E[X_s]$$

Για s=t: $\gamma_{t,t}=\mathrm{Cov}[X_t,X_t]=\mathrm{Var}[X_t]=\sigma_t^2$ δηλαδή η διακύμανση της χρονοσειράς την χρονική στιγμή t.

Για παράδειγμα, εάν βρέχει σήμερα, τα δεδομένα υποδηλώνουν ότι είναι πιο πιθανό να βρέξει αύριο, παρά εάν ήταν καθαρός ο ουρανός σήμερα.[37]

Αυτοσυσχέτιση (autocorrelation)

Η αυτοσυσχέτιση αντιπροσωπεύει τον βαθμό ομοιότητας μεταξύ μιας δεδομένης χρονοσειράς και μιας έκδοσής της σε κάποια άλλη χρονική στιγμή. Η αυτοσυσχέτιση μετρά τη σχέση μεταξύ της τρέχουσας τιμής μιας μεταβλητής και των προηγούμενων τιμών της.

Η συνάρτηση αυτοσυσχέτισης είναι η κανονικοποιημένη μορφή της συνάρτησης και ορίζεται για όλα τα s, t που ανήκουν στο T ως:

$$\rho_{s,t}=Corr\left[X_s,X_t\right]=\frac{Cov\left[X_s,X_t\right]}{\sqrt{Var\left[X_s\right]Var\left[X_t\right]}}=\frac{\gamma_{s,t}}{\sigma_s\sigma_t}$$

Για s=t: $\gamma_{t,t}=\mathrm{Corr}[X_t,X_t]=1$ (δηλαδή υπάρχει τέλεια συσχέτιση).

Για τον υπολογισμό των συναρτήσεων αυτοσυνδιακύμανσης και αυτοσυσχέτισης για πραγματικά δεδομένα, γίνεται η υπόθεση ότι η δομή εξάρτησης των δεδομένων δεν αλλάζει με την πάροδο του χρόνου. Δηλαδή ότι ισχύει:

$$\gamma_{s,t}=\mathrm{Cov}[X_s,X_t]=\mathrm{Cov}[X_{s+r},X_{t+r}]=\gamma_{s+r,t+r}$$

για οποιαδήποτε χρονικά σημεία (s, t) και διάνυσμα αύξησης r. Υπό αυτήν την υπόθεση, ο μόνος παράγοντας που επηρεάζει τη συνδιακύμανση είναι η απόσταση τ = | s − t | μεταξύ των παρατηρήσεων, η οποία ονομάζεται υστέρηση(lag). Επομένως, οι μόνες αυτοσυνδιακυμάνσεις που πρέπει να υπολογιστούν είναι το σύνολο

$$\gamma_\tau = \mathrm{Cov}[X_t, X_{t+\tau}]$$

με το τ να παίρνει μηδενικές ή θετικές ακέραιες τιμές. [38]

ΚΕΦΑΛΑΙΟ 3

ΠΡΟΕΙΝΟΜΕΝΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ

3.1 Προτεινόμενα Μοντέλα Πρόβλεψης

3.1.1 Εισαγωγή

Τα μοντέλα μηχανικής μάθησης έχουν καθιερωθεί την τελευταία δεκαετία ως ένα σημαντικό εργαλείο στον τομέα της πρόβλεψης. Τόσο μοντέλα νευρωνικών δικτύων όσο και στατιστικά μοντέλα είναι σε θέση να δώσουν σχετικά αξιόπιστες εκτιμήσεις όσον αφορά προβλέψεις σε μελλοντικά δεδομένα.

Υπήρξαν εντυπωσιακές πρόοδοι σε αυτόν τον τομέα τα τελευταία χρόνια, τόσο στη θεωρητική κατανόηση των μοντέλων όσο και στην ποσότητα και τις παραλλαγές των μοντέλων που αναπτύχθηκαν. Εκτός από την ανάπτυξη και ανάλυση μοντέλων, πρέπει παράλληλα να γίνει και μία προσπάθεια σύγκρισης των υφιστάμενων μοντέλων και των παραλλαγών τους. Αυτό θα είναι τεράστιας σημασίας για οποιονδήποτε θελήσει να αναπτύξει κάποιο προβλεπτικό μοντέλο, καθώς θα περιορίσει τις πιθανές επιλογές του και θα του δώσει εικόνα για τα δυνατά και αδύνατα σημεία των διαθέσιμων μοντέλων.

Στο κεφάλαιο αυτό θα αναλυθούν τα πιο δημοφιλή μοντέλα που χρησιμοποιούνται σε δεδομένα χρονοσειρών, τα πλεονεκτήματα αλλά και μειονεκτήματα τους. Τα μοντέλα αυτά είναι: **LSTM , ARIMA** και **FBProphet** .

3.1.2 Μοντέλο LSTM

Το Long Short-Term Memory (LSTM) είναι ένα είδος αναδρομικών νευρωνικών δικτύων - Recurrent Neural Network (RNN) με την ικανότητα να “θυμάται” τις τιμές από προηγούμενα στάδια με σκοπό τη μελλοντική χρήση. Πριν εμβαθύνουμε στο LSTM, είναι απαραίτητο να έχουμε μια γνώση στο τι είναι ένα νευρωνικό δίκτυο.

Τεχνητά Νευρωνικά Δύκτια - Artificial Neural Network (ANN)

Ένα νευρωνικό δίκτυο αποτελείται από τουλάχιστον τρία επίπεδα: ένα στρώμα εισόδου, κρυφά στρώματα και ένα στρώμα εξόδου. Ο αριθμός των χαρακτηριστικών του συνόλου δεδομένων καθορίζει τη διάσταση ή τον αριθμό των κόμβων στο επίπεδο εισόδου. Αυτοί οι κόμβοι συνδέονται μέσω συνδέσμων που ονομάζονται "συνάψεις" με τους κόμβους που δημιουργούνται στο κρυφό επίπεδο. Οι σύνδεσμοι συνάψεων φέρουν κάποια βάρη για κάθε κόμβο στο επίπεδο εισόδου. Τα βάρη ουσιαστικά αποφασίζουν για το ποιο σήμα μπορεί να περάσει και ποιο όχι. Ένα νευρωνικό δίκτυο μαθαίνει προσαρμόζοντας το βάρος για κάθε σύνοψη.

Στα κρυφά στρώματα, οι κόμβοι εφαρμόζουν μια συνάρτηση ενεργοποίησης (π.χ. σιγμοειδές ή υπερβολικής εφαπτομένης - tangent hyperbolic(tanh)) στο σταθμισμένο άθροισμα εισόδων για να μετατρέψουν τις εισόδους σε εξόδους ή στις προβλεπόμενες τιμές.[39]

Το RNN διατηρεί ένα κρυφό διάνυσμα h που ενημερώνεται στο χρονικό βήμα t,

$$h_t = \tanh(Wh_{t-1} + Ix_t)$$

Όπου το tanh αντιπροσωπεύει τη συνάρτηση υπερβολικής εφαπτομένης, το διάνυσμα εισόδου στο χρονικό βήμα t συμβολίζεται ως x_t, ο πίνακας βάρους συμβολίζεται με W και ο πίνακας προβολής συμβολίζεται με I.[40]

Το στρώμα εξόδου δημιουργεί ένα διάνυσμα πιθανοτήτων για τις διάφορες εξόδους και επιλέγει αυτό με το ελάχιστο ποσοστό σφάλματος, ελαχιστοποιώντας τις διαφορές μεταξύ των αναμενόμενων και των προβλεπόμενων τιμών, γνωστό και ως κόστος, χρησιμοποιώντας μια συνάρτηση που ονομάζεται SoftMax. Η συνάρτηση Softmax ομαλοποιεί τις προβλέψεις εξόδου του μοντέλου ως έγκυρη κατανομή πιθανότητας. Μια πρόβλεψη y_t μπορεί να γίνει χρησιμοποιώντας μια κρυφή κατάσταση h και έναν πίνακα βάρους W,

$$y_t = \text{softmax}(Wh_{t-1}).[40]$$

Οι αναθέσεις στο διάνυσμα βαρών και κατ' επέκταση τα λάθη που προέκυψαν μέσω της εκπαίδευσης δικτύου για πρώτη φορά μπορεί να μην είναι τα καλύτερα. Για να βρεθούν οι βέλτιστες τιμές για τα σφάλματα, γίνεται back propagation στα σφάλματα, διαδίδονται δηλαδή στο δίκτυο από το στρώμα εξόδου προς τα κρυμμένα στρώματα και ως αποτέλεσμα τα βάρη προσαρμόζονται. Η διαδικασία επαναλαμβάνεται αρκετές φορές με τις ίδιες παρατηρήσεις και τα βάρη προσαρμόζονται εκ νέου μέχρι να υπάρξει βελτίωση στις προβλεπόμενες τιμές.[39]

Αναδρομικά Νευρωνικά Δύκτια - Recurrent Neural Network (RNN)

Ένα RNN είναι μια ειδική περίπτωση νευρωνικού δικτύου όπου ο στόχος είναι η πρόβλεψη του επόμενου βήματος στην ακολουθία των παρατηρήσεων σε σχέση με τα προηγούμενα βήματα που παρατηρήθηκαν στην ακολουθία. Στην πραγματικότητα, η ιδέα πίσω από τα RNN είναι να κάνουν χρήση διαδοχικών παρατηρήσεων και να μάθουν από τα προηγούμενα στάδια για να προβλέψουν τις μελλοντικές τάσεις. Στα RNN, τα κρυμμένα στρώματα λειτουργούν ως μια "αποθήκη" για την αποθήκευση των πληροφοριών που συλλέγονται σε προηγούμενα στάδια. Τα RNN ονομάζονται "αναδρομικά" επειδή εκτελούν την ίδια διεργασία για κάθε στοιχείο της ακολουθίας, με την ιδιαιτερότητα ότι κάνουν χρήση των πληροφοριών που συλλέχθηκαν νωρίτερα για την πρόβλεψη διαφορετικών μελλοντικών δεδομένων.[39]

Τα RNN μπορούν να στοιβαχτούν για να δημιουργήσουν βαθύτερα δίκτυα χρησιμοποιώντας την κρυφή κατάσταση, h^{l-1} ενός στρώματος RNN l−1 ως είσοδο στην κρυφή κατάσταση, h^{l} ενός άλλου στρώματος RNN l,

$$h_t^l = \sigma\left(Wh_{t-1}^l + Ih_t^{l-1} \right).$$

όπου σ είναι η σιγμοειδής συνάρτηση, η οποία παίρνει έναν πραγματικό αριθμό και τον μετατρέπει σε ένα ίδιο στο εύρος από 0 έως και 1.[40]

Η κύρια πρόκληση με ένα τυπικό RNN είναι ότι αυτά τα δίκτυα θυμούνται μόνο μερικά προηγούμενα βήματα στην ακολουθία και έτσι δεν είναι κατάλληλα για να θυμούνται μεγαλύτερες ακολουθίες δεδομένων.[39]

Long Short-Term Memory (LSTM)

Το LSTM είναι ένα ειδικό είδος RNN με την ιδιαιτερότητα ότι απομνημονεύει μεγάλου μήκους ακολουθίες δεδομένων με μεγαλύτερη ακρίβεια από άλλους τύπους RNN. Η απομνημόνευση της προηγούμενης τάσης των δεδομένων είναι δυνατή μέσω ορισμένων πυλών οι οποίες είναι ενσωματωμένες σε ένα τυπικό LSTM. Αυτό καθιστά τα μοντέλα LSTM ιδιαίτερα ικανά στη πρόβλεψη χρονοσειρών μεγάλου μήκους.

Κάθε LSTM είναι ένα σύνολο κελιών, όπου καταγράφονται και αποθηκεύονται τα δεδομένα. Τα κελιά μοιάζουν με μια γραμμή μεταφοράς δεδομένων. Λόγω της χρήσης ορισμένων πυλών σε κάθε κελί, τα δεδομένα φιλτράρονται πριν περάσουν στα επόμενα κελιά. Ως εκ τούτου, οι πύλες, οι οποίες βασίζονται σε ένα σιγμοειδές στρώμα, επιτρέπουν στα κελιά να απορρίπτουν ή να αφήνουν τα δεδομένα να περάσουν.

Κάθε σιγμοειδές στρώμα αποδίδει αριθμούς στην περιοχή του μηδέν και του ενός, απεικονίζοντας το ποσοστό κάθε τμήματος δεδομένων που πρέπει να περάσει στο επόμενο κελί. Πιο συγκεκριμένα, μια εκτίμηση μηδενικής αξίας συνεπάγεται ότι δεν θα αφήσει τίποτα να περάσει, ενώ μια εκτίμηση μοναδιαίας αξίας υποδεικνύει ότι θα επιτρέψει σε όλα τα δεδομένα να περάσουν. Τρεις τύποι πυλών εμπλέκονται σε κάθε LSTM με στόχο τον έλεγχο της κατάστασης κάθε κελιού:

- Η πύλη forget, η οποία αποφασίζει το ποσοστό της προηγούμενης κατάστασης που θα περάσει.
- Η πύλη εισόδου καθορίζει το ποσοστό της νέας υπολογισμένης κατάστασης που θα αποθηκευτεί στο κελί.
- Και τέλος η πύλη εξόδου, η οποία καθορίζει το ποσοστό της εσωτερικής κατάστασης που θα περάσει στα επόμενα κελιά.

Όλες οι πύλες έχουν το ίδιο μέγεθος κρυφών καταστάσεων.[39]

Για να γίνει κατανοητό το πως δουλεύει ένα LSTM αρκεί να καταλάβει κανείς πως υπολογίζεται η κρυφή κατάσταση h_t. Ο υπολογισμός σε κάθε χρονικό βήμα απεικονίζεται ως εξής:

$$g^u = \sigma(W^u h_{t-1} + I^u x_t)$$

$$g^f = \sigma(W^f h_{t-1} + I^f x_t)$$

$$g^o = \sigma(W^o h_{t-1} + I^o x_t)$$

$$g^c = \tanh(W^c h_{t-1} + I^c x_t)$$

$$m_t = g^f * m_{t-1} + g^u * g^c$$

$$h_t = \tanh(g^o * m_t)$$

όπου g^u, g^f, g^o είναι τα διανύσματα των πυλών εισόδου, forget και εξόδου και υπολογίζονται με την σιγμοειδή συνάρτηση σ, ενώ το διάνυσμα g^c είναι η κατάσταση του κελιού και υπολογίζεται με την συνάρτηση υπερβολικής εφαπτομένης. Είναι φανερό πως για τον υπολογισμό των διανυσμάτων αυτών είναι απαραίτητη η είσοδος x την χρονική στιγμή t και η κρυφή κατάσταση h την χρονική στιγμή t-1. Το διάνυσμα m_t αναφέρεται στην εσωτερική μνήμη της κρυφής κατάστασης. Τα αναδρομικά βάρη απεικονίζονται ως W^u ,W^f ,W^o ,W^c ενώ οι πίνακες προβολής ως I^u , I^f , I^o , I^c. Τέλος, υπολογίζεται η κρυφή κατάσταση h_t μέσω της συνάρτηση υπερβολικής εφαπτομένης πολλαπλασιάζοντας το διάνυσμα της πύλης εξόδου με το διάνυσμα εσωτερικής μνήμης.[40] Το σύμβολο (*) αναφέρεται στο εσωτερικό γινόμενο.

Πλεονεκτήματα και μειονεκτήματα μοντέλων LSTM

Όπως αναφέρθηκε παραπάνω τα μοντέλα LSTM είναι ένα είδος νευρωνικών δικτύων, η φύση των οποίων τα καθιστά ιδιαίτερα χρήσιμα σε ένα μεγάλο εύρος προβλημάτων. Όσων αφορά όμως την πρόβλεψη χρονοσειρών, τα μοντέλα αυτά παρουσιάζουν και ορισμένα μειονεκτήματα.

Το κύριο πλεονέκτημα των μοντέλων LSTM είναι ότι μπορούν να χρησιμοποιηθούν σε μη γραμμικά δεδομένα χωρίς την βοήθεια πρόσθετων συναρτήσεων ή την προσαρμογή παραμέτρων. Επίσης το back propagation που γίνεται στα σφάλματα μεταξύ των κελιών του δικτύου, επιτρέπει στο μοντέλο να κάνει προβλέψεις με μεγάλη υστέρηση(lag) με σχετικά καλή ακρίβεια.

Ενώ τα μοντέλα αυτά είναι ικανά στην διαχείριση μη γραμμικών δεδομένων, για να το πετύχουν αυτό χρειάζονται αρκετά μεγάλο όγκο δεδομένων, κάτι που πολλές φορές δεν είναι διαθέσιμο. Ένα ακόμα μειονέκτημα είναι ότι απαιτούν πολλούς πόρους και χρόνο για να εκπαιδευτούν ώστε να είναι έτοιμα για ρεαλιστικές εφαρμογές. Τέλος, η εκπαίδευσή τους απαιτεί ιδιαίτερη γνώση σχετικά με την παραμετροποίηση των μοντέλων.

3.1.3 Μοντέλο ARIMA

Τα τελευταία χρόνια πολλά μοντέλα και τεχνικές έχουν αναπτυχθεί για την πρόβλεψη των τιμών των μετοχών και κρυπτονομισμάτων. Μεταξύ αυτών και μοντέλα τεχνητών νευρωνικών δικτύων (ANN) τα οποία είναι τα πιο δημοφιλή λόγω της ικανότητάς τους να μαθαίνουν μοτίβα από δεδομένα που πολλές φορές είναι άγνωστα.

Ενώ τα ANN είναι μοντέλα βασισμένα στην λογική της τεχνητής νοημοσύνης, τα μοντέλα ARIMA προέρχονται από στατιστικά μοντέλα. Γενικά, η διαδικασία της πρόβλεψης σε ένα υπολογιστικό μοντέλο μπορεί να γίνει βάση δύο τεχνικών: Αυτή της στατιστικής και αυτή της τεχνητής νοημοσύνης [41]. Τα μοντέλα ARIMA είναι γνωστά ως ισχυρά και αποτελεσματικά στη πρόβλεψη χρονοσειρών στην χρηματοοικονομική, ιδιαίτερα σε βραχυπρόθεσμες προβλέψεις, ακόμη και από τις πιο δημοφιλείς τεχνικές ANN [42].

Τι είναι το μοντέλο ARIMA:

Για να κατανοήσει κάποιος τι είναι το μοντέλο ARIMA θα πρέπει πρώτα να κατανοήσει έννοιες όπως το αυτοπαλίνδρομο μοντέλο - AutoRegressive model (AR), η Στασιμότητα - Integrated (I) και η Χρονοσειρές Κινητού Μέσου - Moving Average (MA). Από τα προαναφερθέντα καταλαβαίνει κανείς ότι το ARIMA είναι ένα μικτό ολοκληρωμένο μοντέλο που χρησιμοποιεί τις εν λόγω τεχνολογίες.

Αυτοπαλίνδρομο μοντέλο (AR)

Ένα αυτοπαλίνδρομο μοντέλο είναι ένα μοντέλο χρονοσειράς που χρησιμοποιεί παρατηρήσεις από προηγούμενα χρονικά βήματα ως είσοδο σε μια εξίσωση παλινδρόμησης για να προβλέψει την τιμή στο επόμενο χρονικό βήμα. Έστω ένα μοντέλο AR(p), το μοντέλο αυτό ορίζεται από την σχέση:

$$y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

όπου y_t είναι η εξαρτημένη μεταβλητή που παλινδρομείται στις προηγούμενες τιμές τις ίδιας της μεταβλητής y_t, δ είναι μια σταθερά, τα $\varphi_1....\varphi_p$ είναι οι αυτοπαλινδρούμενοι παράμετροι και το p δηλώνει την τάξη του μοντέλου(δηλαδή την υστέρηση). Στην ουσία δηλαδή η τιμή της παρατήρησης y_t εξαρτάται κατά παράγοντα φ_1 από την προηγούμενη παρατήρηση, κατά παράγοντα φ_2 από την προ-προηγούμενη παρατήρηση … και κατά παράγοντα φ_p από την παρατήρηση που βρίσκεται p περιόδους πίσω.[43] Το ε_t είναι ο λευκός θόρυβος που χρησιμοποιεί το μοντέλο.

Μια άλλη αναπαράσταση της παραπάνω εξίσωσης χρησιμοποιώντας τον τελεστή B είναι:

$$\Phi(B)y_t = \delta + \varepsilon_t$$

όπου $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$.

Η χρονοσειρά AR τάξης p είναι στάσιμη εάν η ρίζα του πολυώνυμου αυτού είναι μεγαλύτερη της μονάδας σε απόλυτη τιμή.

Είναι προφανές ότι εάν η παράμετρος φ_1 έχει την τιμή 0, τότε το μοντέλο είναι ισοδύναμο με ένα μοντέλο λευκού θορύβου.[44]

Στασιμότητα και Διαφοροποίηση (I)

Όπως έχει ήδη αναφερθεί σε προηγούμενο κεφάλαιο, μια χρονοσειρά είναι στάσιμη εάν οι ιδιότητές της δεν επηρεάζονται από μεταβολή του χρόνου. Δηλαδή, εάν η κατανομή πιθανότητας των παρατηρήσεων y_t, y_t+1,…, y_t+n είναι ακριβώς η ίδια με την κατανομή πιθανότητας των παρατηρήσεων y_t+k, yt+k+1,…, y_t+k+ n τότε η χρονική σειρά είναι αυστηρά στάσιμη. Όταν n = 0 τότε η κατανομή πιθανότητας του y_t είναι η ίδια για όλες τις χρονικές περιόδους και μπορεί να γραφτεί ως f(y).[44]

Η στασιμότητα υπονοεί έναν τύπο στατιστικής ισορροπίας ή σταθερότητας στα δεδομένα. Κατά συνέπεια, η χρονοσειρά έχει μια σταθερή μέση τιμή που ορίζεται συνήθως ως:

$$\mu_y = E(y) = \int_{-\infty}^{\infty} y f(y)\, dy$$

και σταθερή διακύμανση η οποία ορίζεται ως:[43]

$$\sigma_y^2 = Var(y) = \int_{-\infty}^{\infty} \left(y - \mu_y\right)^2 f(y)\, dy$$

Η μέση τιμή και η διακύμανση του δείγματος χρησιμοποιούνται για την εκτίμηση αυτών των παραμέτρων. Ένα μοντέλο ARIMA για να λειτουργήσει αποδοτικά θα πρέπει να εισαχθεί σε αυτό μια στάσιμη χρονοσειρά. Εάν η χρονοσειρά δεν είναι στάσιμη τότε γίνεται χρήση μίας τεχνικής που ονομάζεται διαφοροποίηση, με σκοπό την μετατροπή της σε στάσιμη. Κατά την διαφοροποίηση σταθεροποιείται ο μέσος όρος μιας χρονοσειράς, εξαλείφοντας (ή μειώνοντας) την τάση και την εποχικότητα.[45]

Για μία χρονοσειρά Z_t, μέσω της διαφοροποίησης, δημιουργούμε μία καινούργια χρονοσειρά Y_t:

$$Y_i = Z_i - Z_{i-1}$$

Τα διαφοροποιημένα δεδομένα θα περιέχουν μία λιγότερη εγγραφή από τα αρχικά. Αν και είναι εφικτό να διαφοροποιηθούν τα δεδομένα περισσότερες από μία φορές, μια διαφοροποίηση είναι συνήθως επαρκής.[45]

Κινητός μέσος όρος(MA)

Μερικές φορές είναι χρήσιμο να παράγεται και να προβάλλεται μια εξομαλυμένη έκδοση των αρχικών δεδομένων στο διάγραμμα χρονοσειρών. Αυτή η τεχνική βοηθάει ιδιαίτερα στο να

αποκαλυφθούν μοτίβα που αλλιώς θα ήταν δύσκολο να αναγνωριστούν. Υπάρχουν διάφοροι τύποι εξομάλυνσης δεδομένων που μπορούν να χρησιμοποιηθούν. Ένας από τους απλούστερους και πιο ευρέως χρησιμοποιούμενους είναι ο απλός κινητός μέσος όρος.

Αν M_T είναι ο κινητός μέσος όρος, τότε ο κινητός μέσος όρος N-εύρους στη χρονική περίοδο T είναι:

$$M_T = \frac{y_T + y_{T-1} + \ldots + y_{T-N+1}}{N} = \frac{1}{N} \sum_{t=T-N+1}^{T} y_t \quad [44]$$

Μία γενική διατύπωση μίας χρονοσειράς Y κινητού μέσου είναι MA(q), όπου το q συμβολίζει την τάξη του κινητού μέσου (μήκος της υστέρησης της μεταβλητής ε_t) και δίνεται από τον τύπο:

$$y_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

όπου ε_t μια χρονοσειρά λευκού θορύβου και θ είναι τα μη μηδενικά βάρη, δηλαδή το σφάλμα που παράχθηκε σε κάθε περίοδο, τα οποία μαζί με την σταθερά μ αποτελούν τις παραμέτρους του μοντέλου που μπορούν να πάρουν οποιαδήποτε πραγματική τιμή.

Μια άλλη αναπαράσταση της παραπάνω εξίσωσης χρησιμοποιώντας τον τελεστή B είναι:

$$y_t = \mu + (1 - \theta_1 B - \cdots - \theta_q B^q) \varepsilon_t$$

Και επειδή το ε_t είναι λευκός θόρυβος, η αναμενόμενη τιμή μίας διαδικασίας MA(q) είναι απλώς:

$$E(y_t) = E (\mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}) = \mu$$

Μία διαδικασία MA(q) είναι πάντα στάσιμη ανεξαρτήτως των τιμών που παίρνουν τα βάρη, καθώς είναι ένα άθροισμα στάσιμων διεργασιών. Επίσης, μια διαδικασία MA(q) λέγεται ότι είναι αντιστρέψιμη εάν οι ρίζες του πολυωνύμου $\theta_q(B) = 0$ είναι σε απόλυτη τιμή μεγαλύτερες από την μονάδα.[46]

Ολοκληρωμένα αυτοπαλίνδρομα μοντέλα κινητού μέσου (ARIMA)

Από τα προαναφερθέντα γίνεται φανερό ότι τα αυτοπαλίνδρομα μοντέλα AR είναι μοντέλα που μπορούν να σταθούν από μόνα τους και να κάνουν προβλέψεις σε χρονοσειρές. Οι προβλέψεις αυτές βέβαια δεν θα είναι τόσο αποδοτικές όσο εάν συνεργαστούν με μοντέλα

κινητού μέσου MA. Ο συνδυασμός αυτός των δύο μοντέλων (ARMA) παράγει αποδοτικές αναλύσεις και προβλέψεις αλλά μόνο σε στάσιμες χρονοσειρές. Στην μελέτη αυτή, τα δεδομένα που χρησιμοποιούνται, δηλαδή οι τιμές των κρυπτονομισμάτων, δεν έχουν στασιμότητα, μίας και είναι κοινό τα κρυπτονομίσματα να έχουν τάσεις ή να φτάνουν σε ακραίες τιμές. Σύνολα δεδομένων σαν αυτό δημιουργούν την ανάγκη για μοντέλα τα οποία θα μπορούν να κάνουν προβλέψεις σε πιο ρεαλιστικές χρονοσειρές, δηλαδή σε χρονοσειρές που επηρεάζονται από τον χρόνο. Την ανάγκη αυτή έρχονται να καλύψουν τα μοντέλα ARIMA(p,d,q) τα οποία εισάγουν τη διαφοροποίηση για την διασφάλιση της στασιμότητας.[47]

Άρα τα μοντέλα ARIMA είναι στατιστικά μοντέλα τα οποία κάνουν προβλέψεις σε μη-στάσιμες χρονοσειρές, εξομαλύνοντας τα δεδομένα με σκοπό την αύξηση της ακρίβειας.

Πλεονεκτήματα και μειονεκτήματα του μοντέλου ARIMA

Σύμφωνα με τους Box και Jenkins, δύο στατιστικολόγους οι οποίοι μελέτησαν εκτεταμένα τα μοντέλα αυτά, υπάρχουν αρκετά πλεονεκτήματα που καθιστούν τα μοντέλα ARIMA ως ένα κατάλληλο τρόπο για προβλέψεις, κυρίως βραχυπρόθεσμων, χρονοσειρών.

Το κύριο πλεονέκτημα που έχουν αυτά τα μοντέλα είναι ότι εκμεταλλευόμενα την αυστηρά στατιστική τους προσέγγιση, απαιτούν μόνο τα προηγούμενα δεδομένα μιας χρονοσειράς για τη γενίκευση της πρόβλεψης. Ως εκ τούτου, η μέθοδος ARIMA μπορεί να αυξήσει την ακρίβεια πρόβλεψης διατηρώντας τον αριθμό των παραμέτρων στο ελάχιστο.

Ένα από τα μειονεκτήματα των μοντέλων ARIMA είναι ότι επειδή κάνουν χρήση παραμέτρων οι οποίες θα πρέπει να γραφτούν από τον χειριστή του μοντέλου, η απόδοση των αποτελεσμάτων του μοντέλου είναι άρρηκτα συνδεδεμένη με τις ικανότητες του χειριστή. Επιπροσθέτως, τα μοντέλα ARIMA είναι ουσιαστικά « backward looking ». Ως εκ τούτου, δεν είναι πολύ ικανά στο να προβλέπουν ακραίες τιμές, εκτός εάν το σημείο καμπής αντιπροσωπεύει μια επιστροφή στην ισορροπία “κανονικότητα” των δεδομένων.[48]

3.1.4 Μοντέλο FBProphet

Ενώ τα μοντέλα ARIMA είναι πολύ ικανά στη πρόβλεψη χρονοσειρών, χρειάζεται να υπάρχει έμπειρο προσωπικό για την αποδοτική τους χρήση. Το 2017 η ομάδα του Facebook

δημιούργησε το FBProphet, ένα μοντέλο ανοιχτού κώδικα για πρόβλεψη χρονοσειρών το οποίο χρησιμοποιεί προκαθορισμένες παραμέτρους αφαιρώντας έτσι τον παράγοντα "χειριστή" από την απόδοση του μοντέλου. Σε περιπτώσεις βέβαια που είναι απαραίτητο, η μη-αυτόματη ρύθμιση παραμέτρων είναι δυνατή.

Τι είναι το μοντέλο FBProphet

Το FBProphet είναι ένα μοντέλο για την πρόβλεψη δεδομένων χρονοσειρών που βασίζεται σε ένα γενικευμένο προσθετικό μοντέλο (το οποίο θα αναλυθεί παρακάτω).[49] Αποτελεί ένα μοντέλο παλινδρόμησης με ερμηνεύσιμες παραμέτρους που μπορούν να προσαρμοστούν είτε αυτόματα ως προεπιλεγμένες παράμετροι, είτε επιτρέπει στους αναλυτές να επιλέξουν τα στοιχεία που σχετίζονται με το συγκεκριμένο πρόβλημα πρόβλεψης και να κάνουν εύκολα προσαρμογές ανάλογα με τις ανάγκες του προβλήματος.[50]

Εάν y(t) ένα μοντέλο FBProphet, οι συνιστώσες του είναι:

$$y(t)=g(t)+s(t)+h(t)+\epsilon_t$$

g(t) - Growth

Με τον όρο g(t) εννοούμε μια συνάρτηση τάσης που χρησιμοποιείται για την ανάλυση των μη περιοδικών αλλαγών της χρονοσειράς. Για την πρόβλεψη του Growth, το βασικό συστατικό της διαδικασίας δημιουργίας δεδομένων είναι ένα μοντέλο για το πώς αυξήθηκε ο πληθυσμός και πώς αναμένεται να συνεχίσει να αυξάνεται. Η μοντελοποίηση του Growth είναι συχνά παρόμοια με την αύξηση του πληθυσμού στα φυσικά οικοσυστήματα , όπου υπάρχει μη γραμμική ανάπτυξη μέχρι να φτάσει σε ένα carrying capacity, το οποίο είναι το μέγιστο μέγεθος πληθυσμού που μπορεί να υποστηρίξει. Για παράδειγμα, το carrying capacity για τον αριθμό των χρηστών του Facebook σε μια συγκεκριμένη περιοχή μπορεί να αντιστοιχεί στον αριθμό των ατόμων που έχουν πρόσβαση στο διαδίκτυο. Αυτό το είδος ανάπτυξης τυπικά μοντελοποιείται χρησιμοποιώντας το μοντέλο logistic growth, το οποίο στην πιο βασική του μορφή είναι:

$$g(t) = \frac{C}{1 + exp(-k(t-m))}$$

όπου C είναι το carrying capacity, k ο ρυθμός αύξησης και m μία offset παράμετρος, δηλαδή η τιμή από την οποία θα αρχίζουν τα δεδομένα προς μέτρηση.[50]

s(t) - Seasonality

Ο όρος s(t) αντιπροσωπεύει τους περιοδικούς παράγοντες οι οποίοι επηρεάζουν τα δεδομένα. Για να μπορέσει να μοντελοποιήσει και στη συνέχεια να προβλέψει τέτοιου είδους μοτίβα, το μοντέλο FBProphet χρησιμοποιεί σειρές Fourier. Εάν P είναι ο χρόνος της περιόδου που μελετάμε (π.χ. σε ένα χρόνο μία περίοδος είναι 365,25, όταν εκτιμούμε τη μεταβλητή του χρόνου σε ημέρες), τότε το μοντέλο μπορεί να προσεγγίσει αυτά τα περιοδικά μοτίβα με την συνάρτηση:

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2\pi n t}{P}\right) + b_n \left(\frac{2\pi n t}{P}\right)\right)$$

η οποία είναι μία τυπική σειρά Fourier. Τα a και b είναι οι παράμετροι που πρέπει να υπολογιστούν κατασκευάζοντας έναν πίνακα διανυσμάτων περιοδικότητας για κάθε τιμή του t στα ιστορικά και μελλοντικά δεδομένα:[46]

$$X(t) = \left[\cos\left(\frac{2\pi n_1 t}{P}\right), \ldots, \sin\left(\frac{2\pi n_N t}{P}\right)\right]$$

Το συστατικό s(t) του μοντέλου μπορεί τότε να γραφτεί ως:

$$s(t) = X(t)\beta$$

Όπου $\beta \sim \text{Normal}(0,\sigma^2)$, που σημαίνει ότι το μοντέλο αρχικά εξομαλύνει την περιοδικότητα.[50]

h(t) - Holidays

Οι διακοπές και άλλα ειδικά γεγονότα όπως αργίες παρέχουν μεγάλα, κάπως προβλέψιμα σοκ σε πολλές χρονοσειρές και συχνά δεν ακολουθούν περιοδικό πρότυπο, επομένως οι επιδράσεις τους δεν διαμορφώνονται σωστά από έναν ομαλό κύκλο.

Ο αναλυτής έχει την δυνατότητα να παρέχει στο μοντέλο μια προσαρμοσμένη λίστα με προηγούμενα και μελλοντικά ειδικά γεγονότα. Η ενσωμάτωση αυτής της λίστας γεγονότων στο μοντέλο είναι απλή υποθέτοντας ότι οι επιπτώσεις των ειδικών αυτών γεγονότων είναι ανεξάρτητες. Για κάθε ειδικό γεγονός i, ορίζεται το D_i σαν το σύνολο των προηγούμενων και μελλοντικών ημερομηνιών για αυτό το γεγονός. Επίσης ορίζεται σε κάθε γεγονός μια

παράμετρο k_i που είναι η αντίστοιχη αλλαγή στη πρόβλεψη. Αυτό γίνεται με παρόμοιο τρόπο όπως στη περιοδικότητα δημιουργώντας πίνακα διανυσμάτων παλινδρόμησης:

$$Z(t) = [1(t \in D_i),\ldots,(1(t \in D_L)]$$

Και όπως με την περιοδικότητα, χρησιμοποιείται η παράμετρος $\kappa \sim (0,v^2)$ για να σχηματιστεί η συνάρτηση των ειδικών γεγονότων:[49]

$$h(t) = Z(t)\kappa$$

Είναι συχνά σημαντικό να ληφθούν υπόψη τα δεδομένα για ένα παράθυρο ημερών γύρω από ένα συγκεκριμένο γεγονός. Για να γίνει αυτό, χρησιμοποιούνται πρόσθετες παράμετροι για τις ημέρες που περιβάλλουν το γεγονός, ουσιαστικά αντιμετωπίζοντας τις ημέρες γύρω από τ γεγονός ως το ίδιο το γεγονός.[50]

ϵ_t - Error

Ο τελευταίος όρος της συνάρτησης αντιπροσωπεύει τυχόν αλλαγές που δεν προσαρμόζονται στο μοντέλο.

Γενικευμένο προσθετικό μοντέλο - generalized additive model (GAM)

Το γενικευμένο προσθετικό μοντέλο είναι μια κατηγορία μοντέλων παλινδρόμησης με μη γραμμικούς εξομαλυντές που εφαρμόζονται στα δεδομένα. Εδώ χρησιμοποιούμε μόνο τον χρόνο ως παλινδρόμηση αλλά επίσης και αρκετές γραμμικές και μη γραμμικές συναρτήσεις του χρόνου ως συστατικά. Η μοντελοποίηση της εποχικότητας ως πρόσθετου συστατικού είναι η ίδια προσέγγιση που ακολουθείται από την εκθετική εξομάλυνση. Το GAM έχει το πλεονέκτημα ότι αποσυντίθεται εύκολα και φιλοξενεί νέα συστατικά όταν είναι απαραίτητο, για παράδειγμα όταν εντοπίζεται μια νέα πηγή εποχικότητας. Τα GAM κάνουν fit πολύ γρήγορα, έτσι ώστε ο χρήστης να μπορεί να αλλάξει διαδραστικά τις παραμέτρους του μοντέλου.[51]

Πλεονεκτήματα και μειονεκτήματα του μοντέλου FBProphet

Το FBProphet αναπτύχθηκε για τυπικά θέματα της Facebook, όπως πρόβλεψη δραστηριότητας χρηστών σε διάφορα τμήματα της. Είναι αρκετά προσαρμοστικό χάρη στις αυτόματες δυνατότητες φιλτραρίσματος και ρύθμιση παραμέτρων και η διεπαφή του είναι απλή. Ωστόσο, το FBProphet δεν προοριζόταν ποτέ ως ένας αλγόριθμος πρόβλεψης για όλες τις χρήσεις. Έχει αξιοσημείωτα μειονεκτήματα, τα οποία αντικατοπτρίζουν την εξειδίκευσή του.

Μερικά από τα πλεονεκτήματά του είναι:

- Είναι αρκετά ικανό στην αναγνώριση των ακραίων τιμών και στη διαχείριση ελλιπών δεδομένων στη χρονοσειρά.
- Είναι πολύ ευέλικτο. Μπορεί να προσαρμόσει την εποχικότητα με πολλές περιόδους και να επιτρέψει στον αναλυτή να κάνει διάφορες υποθέσεις σχετικά με τις τάσεις.
- Το fitting είναι πολύ γρήγορο, επιτρέποντας στον αναλυτή να διερευνήσει διαδραστικά σε μικρό χρονικό διάστημα πολλές προδιαγραφές του μοντέλου.

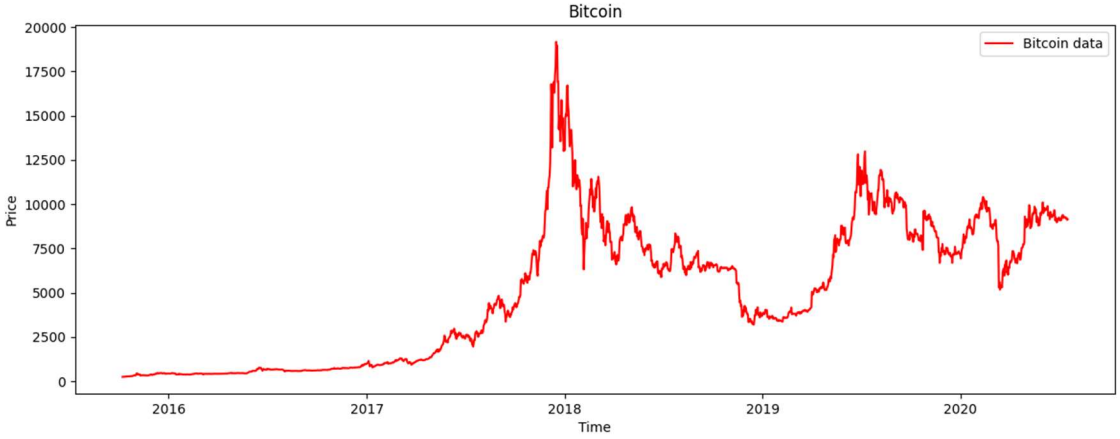
Μερικά από τα μειονεκτήματά του είναι:

- Το μοντέλο αυτό δεν μπορεί να συμπεριλάβει άλλα ουσιαστικά χαρακτηριστικά πέρα από την εποχικότητα ή τα ειδικά γεγονότα.
- Λειτουργεί καλύτερα με χρονοσειρές που έχουν έντονη περιοδικότητα, κάτι που σε πολλές χρονοσειρές, όπως αυτές της μελέτης αυτής, δεν έχουν.

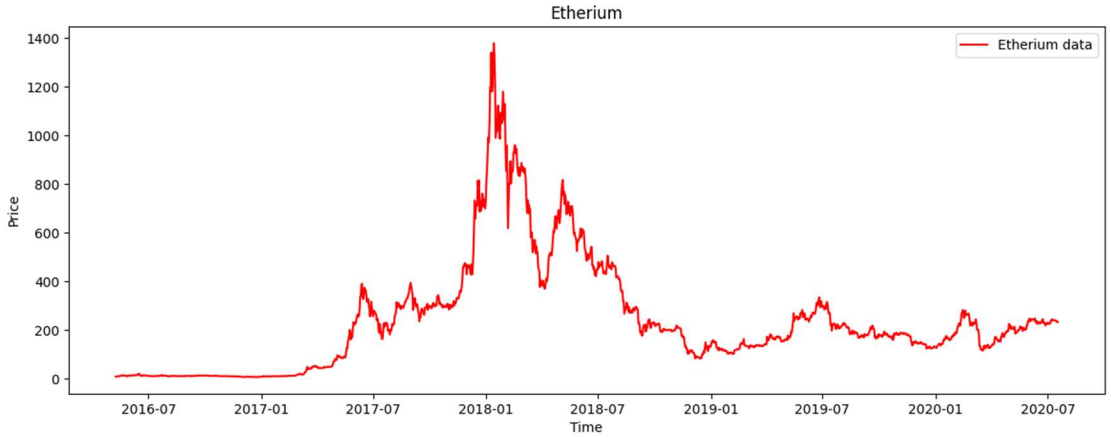
3.2 Δεδομένα Εισόδου

Στα πλαίσια αυτής της μελέτης θα χρησιμοποιηθούν τρία διαφορετικά κρυπτονομίσματα με σκοπό την καλύτερη δυνατή σύγκριση των μοντέλων. Τα κρυπτονομίσματα αυτά είναι τα Bitcoin, Ethereum και Litecoin τα δεδομένα των οποίων αντλήθηκαν από την ιστοσελίδα https://www.cryptodatadownload.com/data/gemini/ και φαίνονται στα παρακάτω σχήματα.

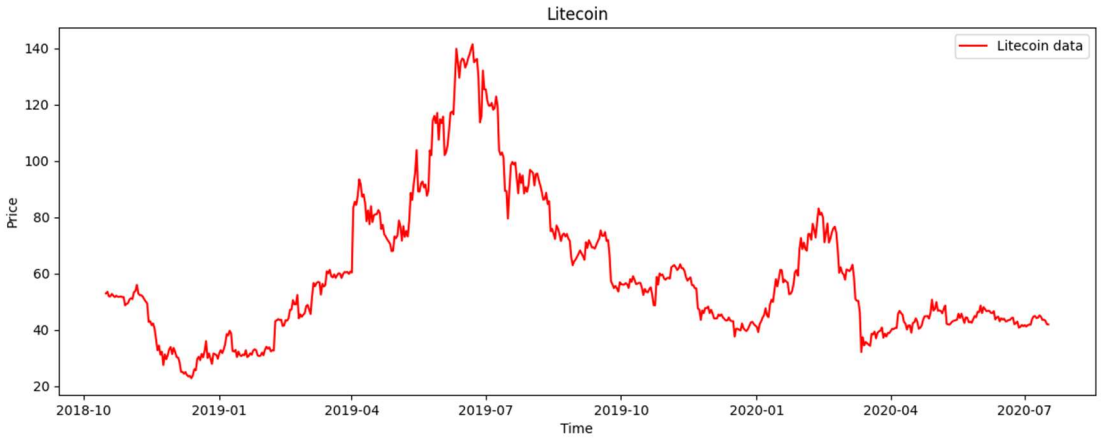
Η επιλογή των κρυπτονομισμάτων δεν είναι τυχαία. Ενώ η καταλυτική ημερομηνία των δεδομένων είναι ίδια σε όλα τα κρυπτονομίσματα (17/07/2020), η αρχική ημερομηνία διαφέρει. Συγκεκριμένα, τα δεδομένα που είναι διαθέσιμα για το Bitcoin ξεκινούν από τις 08/05/2015, για το Ethereum από τις 09/05/2016 και για το Litecoin από τις 16/10/2018. Αντίστοιχα, τα δεδομένα για το καθένα είναι 1745, 1531 και 641. Αυτό προσδίδει μια ποικιλία στα δεδομένα από την άποψη ότι τα μοντέλα θα καλεστούν να κάνουν προβλέψεις με διαφορετικού μεγέθους δεδομένα εκπαίδευσης.



Σχήμα 3.1: Το σύνολο των δεδομένων του Bitcoin.



Σχήμα 3.2: Το σύνολο των δεδομένων του Ethereum.



Σχήμα 3.3: Το σύνολο των δεδομένων του Litecoin.

Τα δεδομένα εισόδου είναι αρχεία τύπου .csv, τα οποία αποθηκεύονται στον ίδιο φάκελο με το πρόγραμμα, και εισάγονται στον κώδικα με την μορφή Dataframes. Πέρα από τις

ημερομηνίες τα δεδομένα αυτά συνοδεύονται από πολλά είδη τιμών, όπως η πρώτη τιμή της κάθε ημέρας, η τελευταία, η ψηλότερη κ.α.. Στην μελέτη αυτή θα χρησιμοποιηθεί η τελευταία τιμή (Close) της κάθε ημέρας.

	Date	Close
0	2015-10-08 04:00:00	243.60
1	2015-10-09 04:00:00	245.51
2	2015-10-10 04:00:00	246.30
3	2015-10-11 04:00:00	248.98
4	2015-10-12 04:00:00	245.75

Σχήμα 3.4: Η μορφή των δεδομένων του Bitcoin ως Dataframe.

Ιδιαίτερη είναι η περίπτωση του μοντέλου FBProphet, του οποίου το Dataframe πρέπει να έχει συγκεκριμένη μορφή. Η στήλη με της ημερομηνίες θα πρέπει να ονομάζεται 'ds' και η στήλη με της τιμές 'y'.

Όλα τα μοντέλα έχουν εκπαιδευτεί στο 90% των δεδομένων σε κάθε κρυπτονόμισμα και δοκιμάζονται στο υπολειπόμενο 10%.

3.2.1 Η λειτουργία των κρυπτονομισμάτων της μελέτης

Η λειτουργία των τριών αυτών κρυπτονομισμάτων, στον πυρήνα της, δεν διαφέρει πολύ. Και τα τρία χρησιμοποιούν την τεχνολογία Blockchain για να αποθηκεύουν τις συναλλαγές που γίνονται μεταξύ των χρηστών. Υπάρχουν όμως κάποιες διαφορές που έχουν επιτρέψει στα κρυπτονομίσματα αυτά να ξεχωρίσουν έναντι άλλων όμοιών τους. Η κύρια διαφορά των τριών αυτών κρυπτονομισμάτων έγκειται στο ότι το κάθε ένα από αυτά χρησιμοποιεί διαφορετικό αλγόριθμο κρυπτογράφησης.

Ο αλγόριθμος κρυπτογράφησης που χρησιμοποιεί το Bitcoin είναι ο SHA-256. Ο αλγόριθμος αυτός είναι ουσιαστικά μια συνάρτηση hash, που δημοφιλώς χρησιμοποιείται σε πρωτόκολλα κρυπτογράφησης. Στη συγκεκριμένη περίπτωση, ο Satoshi Nakamoto, ο δημιουργός του Bitcoin, χρησιμοποιεί τον SHA-256 με σκοπό την επιβεβαίωση της κάθε συναλλαγής(proof-of-work). Η χρήση 256 bit, όπως προσδίδει και το όνομα του, στη τιμή του hash, καθιστά τον αλγόριθμο αυτό πρακτικά απαραβίαστο απέναντι σε κακόβουλες ενέργειες.[52]

Παρόμοια με το SHA-256, και ο αλγόριθμος κρυπτογράφησης του Ethereum, ο Ethash, χρησιμοποιεί 256 bit στη τιμή του hash και γενικά είναι μια αναβαθμισμένη έκδοση του

πρώτου. Η ιδιαιτερότητά του και αυτό που το ξεχωρίζει από τον SHA-256 είναι ότι κατασκευάστηκε για να είναι ανθεκτικό σε ένα πρόβλημα που αντιμετώπιζαν κρυπτονομίσματα όπως το Bitcoin. Το πρόβλημα αυτό είναι ότι ήταν πολύ εύκολο να κατασκευαστεί υλικός εξοπλισμός ο οποίος ειδικευόταν στην επίλυση των περίπλοκων αλγορίθμων του Blockchain. Για να καταπολεμήσει τη πρόκληση αυτή, ο Vitalik Buterin, ο δημιουργός του Ethereum, χρησιμοποίησε στον αλγόριθμο κρυπτογράφησής του τον αλγόριθμο Hashimoto, του οποίου ο σκοπός ήταν να αυξήσει την κατανάλωση της RAM κατά την διαδικασία της εξόρυξης, κάτι που θα περιόριζε σημαντικά τις επιδόσεις ενός πιθανού εξειδικευμένου εξοπλισμού.[52]

Παρά την προσπάθεια του Ethash να εναντιωθεί στο πρόβλημα αυτό, δύο χρόνια μετά την δημιουργία του ο πρώτος υλικός εξοπλισμός κατασκευασμένος για αποτελεσματικότερη εξόρυξη του Ethereum ήταν γεγονός. Αλλά εκεί που απέτυχε το Ethereum θα πετύχαινε το Litecoin. Ο αλγόριθμος κρυπτογράφησης του κρυπτονομίσματος αυτού, ο scrypt, ο οποίος χρησιμοποιεί και αυτός 256 bit στη τιμή του hash του, βρήκε ένα τρόπο να αντιμετωπίσει την πρόκληση αυτή κάνοντας την παραγωγή κλειδιού για το hash μια αργή και μεγάλης κατανάλωσης μνήμης διαδικασία.[53]

3.3 Μετρικές αξιολόγησης

3.3.1 Εισαγωγή

Σε ρεαλιστικά προβλήματα προβλεπτικής ανάλυσης είναι απίθανο οι τιμές της πρόβλεψης που δίνει κάποιο μοντέλο να συμβαδίζουν πλήρως με τις πραγματικές τιμές. Πάντα θα υπάρχει κάποια διαφορά σε αυτές τις τιμές η οποία ονομάζεται σφάλμα.

Το σφάλμα πρόβλεψης είναι η διαφορά μεταξύ της πραγματικής τιμής και της προβλεπόμενης τιμής για την αντίστοιχη περίοδο:

$$E_t = Y_t - F_t$$

όπου E_t είναι το σφάλμα την περίοδο t, Y_t είναι η πραγματική τιμή την περίοδο t και F_t είναι η προβλεπόμενη τιμή την περίοδο t.

Για να γίνει η σύγκριση των μοντέλων, θα πρέπει πρώτα να μετρηθεί το σφάλμα που δίνει το καθένα με την βοήθεια έτοιμων συναρτήσεων. Μια τέτοια ενέργεια επιτρέπει την λήψη μια

ξεκάθαρης εικόνας για τις προβλεπτικές ικανότητες του κάθε μοντέλου με σκοπό την εξαγωγή, όσο το δυνατόν, σωστών συμπερασμάτων.

Υπάρχουν πολλών ειδών δείκτες οι οποίοι εκφράζουν το σφάλμα με διαφορετικούς τρόπους ο καθένας. Στη μελέτη αυτή θα χρησιμοποιηθούν οι δείκτες MAE (Mean Absolute Error), RMSE (Root Mean Squared Error) και MAPE (Mean Absolute Percentage Error) οι οποίοι θα αναλυθούν παρακάτω.

3.3.2 MAE (Mean Absolute Error)

Το μέσο απόλυτο σφάλμα (MAE) είναι η τυπική απόκλιση των σφαλμάτων. Τα σφάλματα είναι ένα μέτρο του πόσο μακριά είναι τα προβλεπόμενα σημεία δεδομένων από τα πραγματικά. Το MAE είναι ένα μέτρο για το πόσο διασκορπισμένα είναι αυτά τα σφάλματα. Είναι βαθμολογία με αρνητικό προσανατολισμό, πράγμα που σημαίνει ότι οι χαμηλότερες τιμές είναι καλύτερες.

Ο δείκτης αυτός είναι ο μέσος όρος των απόλυτων τιμών της απόκλισης και δίνεται από τον μαθηματικό τύπο:

$$MAE = \frac{\sum_{t=1}^{N} |E_t|}{N}$$

3.3.3 RMSE (Root Mean Squared Error)

Η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) δείχνει τη ρίζα των μέσων όρων των τετραγώνων των σφαλμάτων. Επειδή υπολογίζεται το τετράγωνο των σφαλμάτων, το αποτέλεσμα είναι ότι τα μεγαλύτερα σφάλματα έχουν μεγαλύτερο βάρος στη μέτρηση. Αυτό σημαίνει ότι ο δείκτης αυτός είναι “αυστηρός” με χρονοσειρές που παρουσιάζουν ακραίες τιμές κατά την δοκιμή, μίας και σε αυτές τις περιπτώσεις παρουσιάζονται μεγαλύτερα σφάλματα. Παρόμοια με τον δείκτη MAE έχει και αυτός αρνητικό προσανατολισμό.

Το RMSE διατυπώνεται από τον ακόλουθο τύπο:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N} E_t^2}{N}}$$

3.3.4 MAPE (Mean Absolute Percentage Error)

Από την στιγμή που η μελέτη αυτή συγκρίνει δεδομένα από διαφορετικές χρονοσειρές και άρα διαφορετικές κλίμακες, η ύπαρξη ενός μετρικού το οποίο είναι ανεξάρτητο από την κλίμακα είναι απαραίτητη.

Το μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE) εκφράζει σε μορφή ποσοστού τη τυπική απόκλιση των υπολοίπων. Είναι ο μέσος όρος του αθροίσματος των σφαλμάτων προς την πραγματική τιμή πολλαπλασιασμένα με το εκατό. Ένα μειονέκτημα αυτού του μετρικού είναι ότι εάν σε κάποια περίοδο η πραγματική τιμή είναι μηδενική, τότε το αποτέλεσμα θα είναι το άπειρο αφού θα γίνει διαίρεση με το μηδέν. Η φύση των δεδομένων μας όμως είναι τέτοια που η τιμή μηδέν δεν είναι εφικτό να υπάρξει.

Το MAPE διατυπώνεται από τον ακόλουθο τύπο:

$$MAPE = 100\frac{\sum_{t=1}^{N} \left| \frac{E_t}{Y_t} \right|}{N}$$

ΚΕΦΑΛΑΙΟ 4

ΕΦΑΡΜΟΓΗ ΤΩΝ ΜΟΝΤΕΛΩΝ

4.1 Υλοποίηση

4.1.1 Εισαγωγή

Τα μοντέλα που χρησιμοποιούνται σε αυτή τη μελέτη έχουν υλοποιηθεί στη γλώσσα προγραμματισμού Python. Υπάρχουν πολλά κοινά μεταξύ των μοντέλων αλλά οι ιδιαιτερότητές τους οδηγούν σε ξεχωριστές προσεγγίσεις όσον αφορά τον κώδικα. Παρακάτω παρουσιάζεται το κομμάτι κώδικα, αυτό το οποίο υλοποιεί ουσιαστικά το μοντέλο, και θα αναλυθεί σύντομα η λειτουργία του σε ένα ρεαλιστικό πλέον παράδειγμα.

Γιατί Python;

Μια από τις πτυχές που καθιστά την Python τόσο δημοφιλής επιλογή γενικά, είναι η πληθώρα βιβλιοθηκών που διευκολύνουν την κωδικοποίηση και εξοικονομούν χρόνο ανάπτυξης.

Η μηχανική μάθηση συγκεκριμένα έχει επωφεληθεί σε μεγάλο βαθμό από αυτό, με βιβλιοθήκες όπως οι numpy και pandas να παρέχουν εξαιρετική υποστήριξη στη διαχείριση των δεδομένων. Η βιβλιοθήκη Keras είναι πολύ σημαντικό εργαλείο για την δημιουργία τεχνιτών νευρωνικών δικτύων, όπως είναι το μοντέλο LSTM. Η βιβλιοθήκη sklearn επιτρέπει την χρήση έτοιμων συναρτήσεων για τον υπολογισμό των μετρικών αξιολογήσεων. Τέλος, μια ακόμη βαρυσήμαντη βιβλιοθήκη είναι η matplotlib, η οποία επιτρέπει την αναπαράσταση των αποτελεσμάτων σε μορφή γραφημάτων και όχι μόνο.

Συγκεκριμένα, για την υλοποίηση όλων των προγραμμάτων χρησιμοποιήθηκαν οι εξής βιβλιοθήκες:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import sklearn
from sklearn.metrics import mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.metrics import mean_absolute_percentage_error
import math
```

Για την δημιουργία του μοντέλου LSTM έγινε χρήση των συναρτήσεων:

```
from keras.models import Sequential
from keras.layers import LSTM, Dense, Dropout
from keras.preprocessing.sequence import TimeseriesGenerator
```

Για την δημιουργία του μοντέλου ARIMA έγινε χρήση της συνάρτησης:

```
from statsmodels.tsa.arima_model import ARIMA
```

Για την δημιουργία του μοντέλου FBProphet έγινε χρήση της συνάρτησης:

```
from fbprophet import Prophet
```

4.1.2 Υλοποίηση LSTM μοντέλου

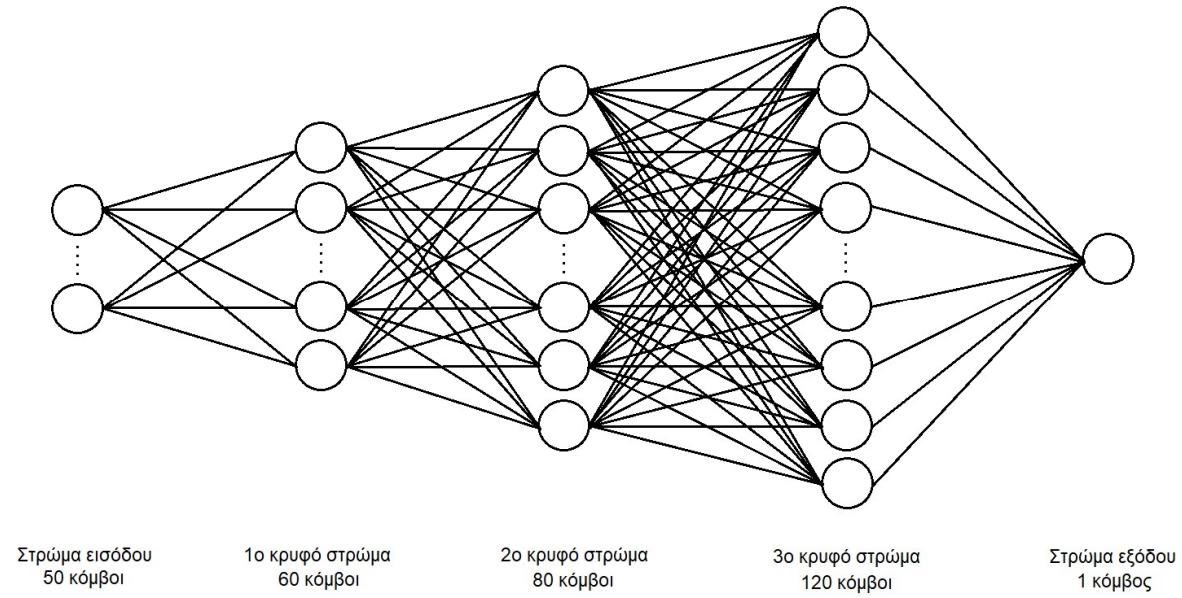
Τα μοντέλα LSTM είναι μοντέλα νευρωνικών δικτύων επιβλεπόμενης μάθησης. Αυτό σημαίνει ότι κατά την εκπαίδευσή του, το μοντέλο για κάθε δεδομένο εισόδου χρειάζεται και ένα δεδομένο εξόδου. Λειτουργεί δηλαδή σαν μια σχέση δασκάλου-μαθητή, όπου του δίνονται τα δεδομένα και μαζί με αυτά και οι αναμενόμενες "απαντήσεις", π.χ. εάν η είσοδος είναι μια εικόνα, η έξοδος θα ήταν μια ετικέτα της οποίας το όνομα θα ήταν το αντικείμενο που παρουσιάζει η εικόνα. Σε προβλήματα χρονοσειρών όμως, η φύση των δεδομένων είναι τέτοια που δεν μπορεί να υπάρξει μια αντίστοιχη ετικέτα. Αντ' αυτού, χρησιμοποιείται μια υστέρηση(lag) στα δεδομένα εισόδου για να μετατραπούν σε δεδομένα εξόδου. Αυτό υλοποιείται αυτόματα με την βοήθεια της συνάρτησης TimeseriesGenerator():

```
train_generator = TimeseriesGenerator(train, train, length=look_back, batch_size=batch_size)
```

Όπου η μεταβλητη "train" περιέχει τα δεδομένα προς εκπαίδευση, η μεταβλητή "look_back" περιέχει την υστέρηση και η μεταβλητη "batch_size" δηλώνει το μέγεθος του "batch", το οποίο αναφέρεται στον αριθμό παραδειγμάτων εκπαίδευσης που χρησιμοποιούνται σε μία επανάληψη.

Έπειτα μοντελοποιείται το LSTM ορίζοντας τα στρώματα του νευρωνικού δικτύου μαζί με τις αντίστοιχες παραμέτρους αυτών.

```
model = Sequential()
model.add(LSTM(units=50, activation = 'relu', return_sequences=True, input_shape=(look_back, 1)))
model.add(Dropout(0.1))
model.add(LSTM(units=60, activation = 'relu', return_sequences=True))
model.add(Dropout(0.2))
model.add(LSTM(units=80, activation = 'relu', return_sequences=True))
model.add(Dropout(0.3))
model.add(LSTM(units=120, activation = 'relu'))
model.add(Dropout(0.4))
model.add(Dense(units=1))
model.compile(optimizer='adam', loss='mse')
```



Σχήμα 4.1: Αναπαράσταση του μοντέλου LSTM της μελέτης.

Πρωτίστως, αρχικοποιείται το μοντέλο ως sequencial(). Το sequencial μοντέλο είναι μια γραμμική στοίβα στρωμάτων, δηλαδή επιτρέπει την προσθήκη στρωμάτων το ένα μετά το άλλο, με την έξοδο του ενός να είναι η είσοδος του επόμενου.

Συγκεκριμένα, το παραπάνω τεχνητό νευρωνικό δίκτυο αποτελείται από πέντε στρώματα όπου η μεταβλητή "units" αναφέρεται στους νευρώνες που περιέχει το καθένα, με το στρώμα εξόδου να ορίζεται συνήθως ως ένα.

Το "input_shape" δείχνει τις διαστάσεις των δεδομένων που θα εισαχθούν στο στρώμα εισόδου.

Η παράμετρος "Dropout" είναι μια μέθοδος κανονικοποίησης όπου οι είσοδοι και οι επαναλαμβανόμενες συνδέσεις με τους νευρώνες αποκλείονται βάση πιθανότητας από την ενεργοποίηση και τις ενημερώσεις βάρους κατά την εκπαίδευση ενός δικτύου, το οποίο έχει ως αποτέλεσμα τη βελτίωση της απόδοσης του μοντέλου.

Η παράμετρος “activation” δηλώνει την συνάρτηση ενεργοποίησης, όπου στη συγκεκριμένη περίπτωση χρησιμοποιείται η “ReLU” (Rectified Linear Unit). Σε σύγκριση με την συνάρτηση υπερβολικής εφαπτομένης (tanh) η οποία είναι η “default” συνάρτηση ενεργοποίησης, η “ReLU” επιτρέπει στα μοντέλα να εκπαιδεύονται γρηγορότερα και να αποδίδουν καλύτερα.[54]

Σε περιπτώσεις που χρησιμοποιούνται παραπάνω από ένα στρώμα νευρωνικού δικτύου, όπως γίνεται εδώ, είναι σημαντικό η παράμετρος “return_sequence” να είναι True. Αυτό γίνεται με σκοπό η έξοδος του στρώματος να είναι ίδιας διάστασης με την είσοδό του, μιας και το επόμενο στρώμα θα λάβει ως είσοδο την έξοδο αυτή.

Όσον αφορά το “compile”, οι παράμετροι **“**optimizer**”** και “loss” εμφανίζουν στοιχεία στην κονσόλα του προγράμματος κατά την εκπαίδευση.

Μετά από πολλές δοκιμές αποδείχθηκε ότι αυτές οι παράμετροι δίνουν τα καλύτερα αποτελέσματα για τις χρονοσειρές που χρησιμοποιούνται στην μελέτη αυτή.

Στη συνέχεια το μοντέλο κάνει χρήση της μεθόδου fit_generator() η οποία παίρνει ως ορίσματα το “train_generator” που αναλύθηκε παραπάνω, των αριθμό των εποχών (μια εποχή σημαίνει ότι θα χρησιμοποιήσει τα δεδομένα εκπαίδευσης μια φορά) και τη μεταβλητή “steps_per_epoch”, η οποία δηλώνει πόσα “batch” θα χρησιμοποιηθούν για κάθε εποχή, ώστε να γίνει η εκπαίδευση.

```
model.fit_generator(train_generator, steps_per_epoch=len(train_generator), epochs=num_epochs)
```

Τέλος αρκεί να γίνει χρήση της συνάρτησης “predict” του μοντέλου στα δεδομένα δοκιμής. Ένα κοινό λάθος που γίνεται στη συγκεκριμένη φάση είναι να χρησιμοποιήσει κάποιος την συνάρτηση TimeseriesGenerator() στα δεδομένα δοκιμής όπως στα δεδομένα εκπαίδευσης και να κάνει πρόβλεψη στο αποτέλεσμα αυτής. Σε αυτή την περίπτωση τα αποτελέσματα της πρόβλεψης θα είναι σε ένα μη-ρεαλιστικό βαθμό επηρεασμένα από τα δεδομένα δοκιμής, θα έχουν δηλαδή μια προκατάληψη και δεν θα ανταποκρίνονται στις πραγματικές ικανότητες του μοντέλου. Για να γίνει σωστά η πρόβλεψη θα πρέπει να γίνεται επαναλαμβανόμενα η μέθοδος predict στα αποτελέσματα που έχει δώσει η ίδια.

```
def predict(num_prediction, model):
    prediction_list = train[-look_back:]

    for _ in range(num_prediction):
        x = prediction_list[-look_back:]
        x = x.reshape((1, look_back, 1))
        out = model.predict(x)[0][0]
        prediction_list = np.append(prediction_list, out)
    prediction_list = prediction_list[look_back - 1:]

    return prediction_list
```

Στη συνάρτηση αυτή, η λίστα prediction_list αρχικοποιείται με τα τελευταία δεδομένα (όση και η υστέρηση) εκπαίδευσης, στην οποία γίνεται η πρόβλεψη και προστίθενται σε αυτή οι προβλεπόμενες τιμές σε κάθε επανάληψη. Γίνεται ουσιαστικά μια ολίσθηση προς τα δεξιά της λίστας αυτής, όπου σε κάθε βήμα γίνεται η πρόβλεψη για την επόμενη τιμή.

4.1.3 Υλοποίηση ARIMA μοντέλου

Σε αντίθεση με το μοντέλο LSTM που είναι ένα μοντέλο γενικού σκοπού, το μοντέλο ARIMA ειδικεύεται σε δεδομένα χρονοσειρών. Αυτό, όπως φαίνεται παρακάτω, έχει ως αποτέλεσμα την απλή και εύκολη υλοποίηση του, μιας και οι περισσότερες λειτουργίες γίνονται από έτοιμες συναρτήσεις στο παρασκήνιο.

Αρχικά, αφού έχουν χωριστεί τα δεδομένα, χρησιμοποιείται η συνάρτηση ARIMA() με τα δεδομένα εκπαίδευσης και τις παραμέτρους p,d,q για να δημιουργηθεί το μοντέλο.

```
model = ARIMA(train, order=(2, 1, 1))
fitted = model.fit()
```

Οι παράμετροι (2, 1, 1) σημαίνουν ότι χρησιμοποιείται υστέρηση ίση με δύο, γίνεται μια φορά διαφοροποίηση στα δεδομένα για να γίνει στάσιμη η χρονοσειρά, και μπαίνει ένα υστερημένο σφάλμα πρόβλεψης στο μοντέλο αντίστοιχα.

Στη συνέχεια, αφού γίνει fit το μοντέλο αρκεί να γίνει η πρόβλεψη, χρησιμοποιώντας την έτοιμη συνάρτηση forecast() δίνοντας ως όρισμα τον αριθμό των ημερών που θα γίνει η πρόβλεψη.

```
fc, se, conf = model_fit.forecast(test_data_number)
```

Όπου η μεταβλητή fc περιέχει τα προβλεπόμενα δεδομένα.

4.1.4 Υλοποίηση FBProphet μοντέλου

Παρόμοια με το μοντέλο ARIMA, και το μοντέλο FBProphet είναι ένα στατιστικό μοντέλο που κατασκευάστηκε για να χρησιμοποιείται σε χρονοσειρές.

Σε αντίθεση με τα άλλα μοντέλα, στο μοντέλο αυτό τα δεδομένα εισόδου πρέπει να έχουν μια συγκεκριμένη μορφή. Συγκεκριμένα, δέχεται αποκλειστικά και μόνο dataframes τα οποία θα πρέπει να έχουν δύο στήλες: η στήλη με τις ημερομηνίες με όνομα “ds” και η στήλη με τις τιμές με όνομα “y”.

Αρχικά δημιουργείται το μοντέλο, και έπειτα, όταν γίνεται fit, εκπαιδεύεται δίνοντας ως όρισμα τα δεδομένα εκπαίδευσης.

```
model = Prophet()
model.fit(train)
```

Στη συνέχεια, χρησιμοποιώντας την συνάρτηση make_future_dataframe(), ορίζεται η τιμή των μελλοντικών ημερών στις οποίες θα γίνει η πρόβλεψη.

```
future = model.make_future_dataframe(periods=days)
```

Τέλος, αρκεί να γίνει χρήση της μεταβλητής future στη έτοιμη συνάρτηση predict του μοντέλου για να γίνει η πρόβλεψη.

```
prediction = model.predict(future)
```

Τα προβλεπόμενα δεδομένα θα βρίσκονται στη στήλη “yhat” του dataframe prediction που δημιουργήθηκε.

4.2 Σύγκριση αποτελεσμάτων

4.2.1 Εισαγωγή

Αφού αναλύθηκε το πως πρακτικά δομείται ένα μοντέλο, στην ενότητα αυτή παρουσιάζονται κα συγκρίνονται, με χρήση γραφημάτων, τα αποτελέσματα αυτών. Πέρα από την δοκιμή τριών διαφορετικών χρονοσειρών κρυπτονομισμάτων, ένας άλλος παράγοντας που θα βοηθήσει να

εξεταστούν πιο σφαιρικά οι επιδόσεις των μοντέλων είναι η χρονική περίοδος για την οποία θα γίνει η πρόβλεψη.

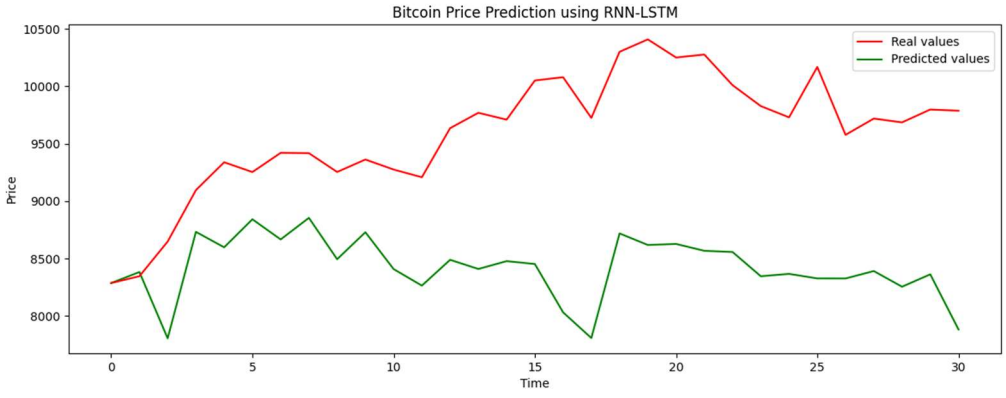
Στις βραχυπρόθεσμες προβλέψεις τα μοντέλα "νιώθουν" πιο σίγουρα, αφού ένα μεγάλο κομμάτι των δεδομένων που χρησιμοποιούν, ώστε να δημιουργήσουν τα καινούργια προβλεπόμενα δεδομένα, είναι τα υπάρχοντα δεδομένα του ιστορικού.

Αυτό όμως δεν συμβαίνει στην μακροπρόθεσμη πρόβλεψη. Εδώ, σχεδόν όλα τα προβλεπόμενα δεδομένα έχουν δημιουργηθεί από ήδη προηγούμενα προβλεπόμενα δεδομένα. Σε τέτοιου είδος προβλέψεις είναι σημαντικό να υπάρχει μια σταθερότητα ή και περιοδικότητα στα δεδομένα ώστε τα μοντέλα να μπορούν να "πατήσουν" πάνω σε αυτά τα χαρακτηρίστηκα. Όπως έχει ήδη αναφερθεί, οι χρονοσειρές των κρυπτονομισμάτων είναι τυχαίες και επηρεάζονται από άλλους παράγοντες. Αξίζει όμως να δούμε πως συμπεριφέρονται τα μοντέλα σε προβλέψεις πολλών μελλοντικών δεδομένων.

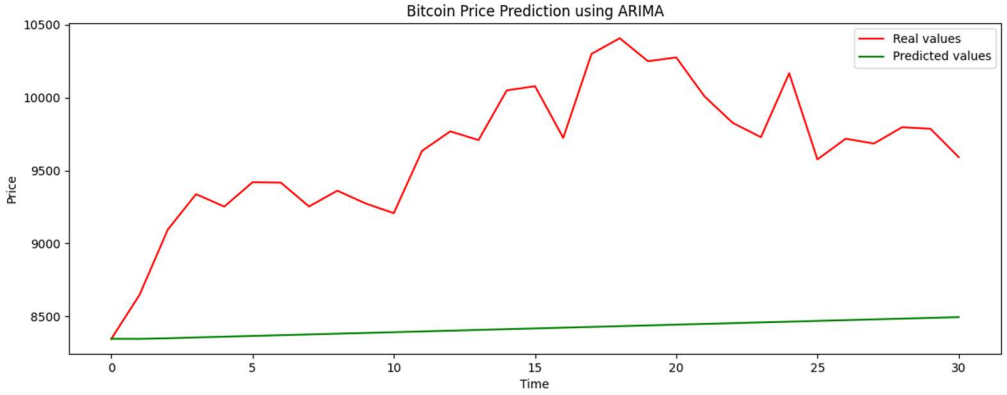
Στη μελέτη αυτή, ορίζεται ως βραχυπρόθεσμη πρόβλεψη το διάστημα των 30 ημερών, ενώ ως μακροπρόθεσμη το διάστημα των 150 ημερών.

4.2.2 Βραχυπρόθεσμη πρόβλεψη

Αρχικά θα αναλυθεί η πρόβλεψη των 30 ημερών για κάθε κρυπτονόμισμα ξεχωριστά, ξεκινώντας με το Bitcoin, του οποίου τα δεδομένα δοκιμής παρουσιάζουν μια απότομη αύξηση.

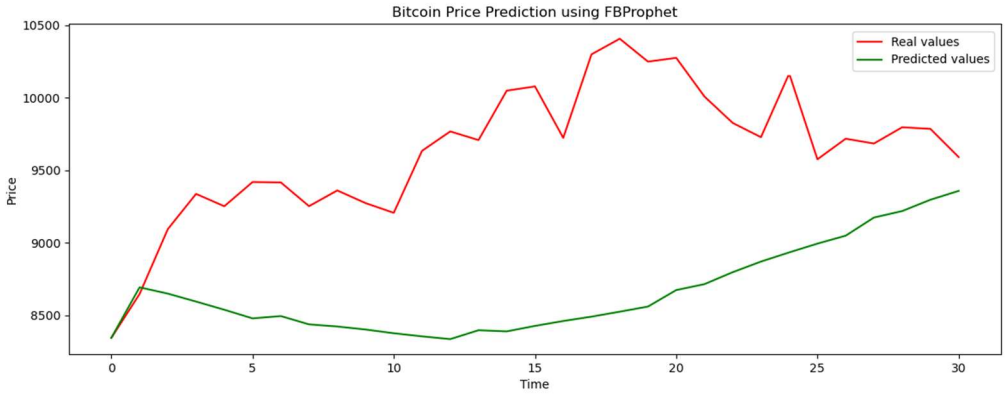


Σχήμα 4.2: Βραχυπρόθεσμη πρόβλεψη του Bitcoin με μοντέλο LSTM.



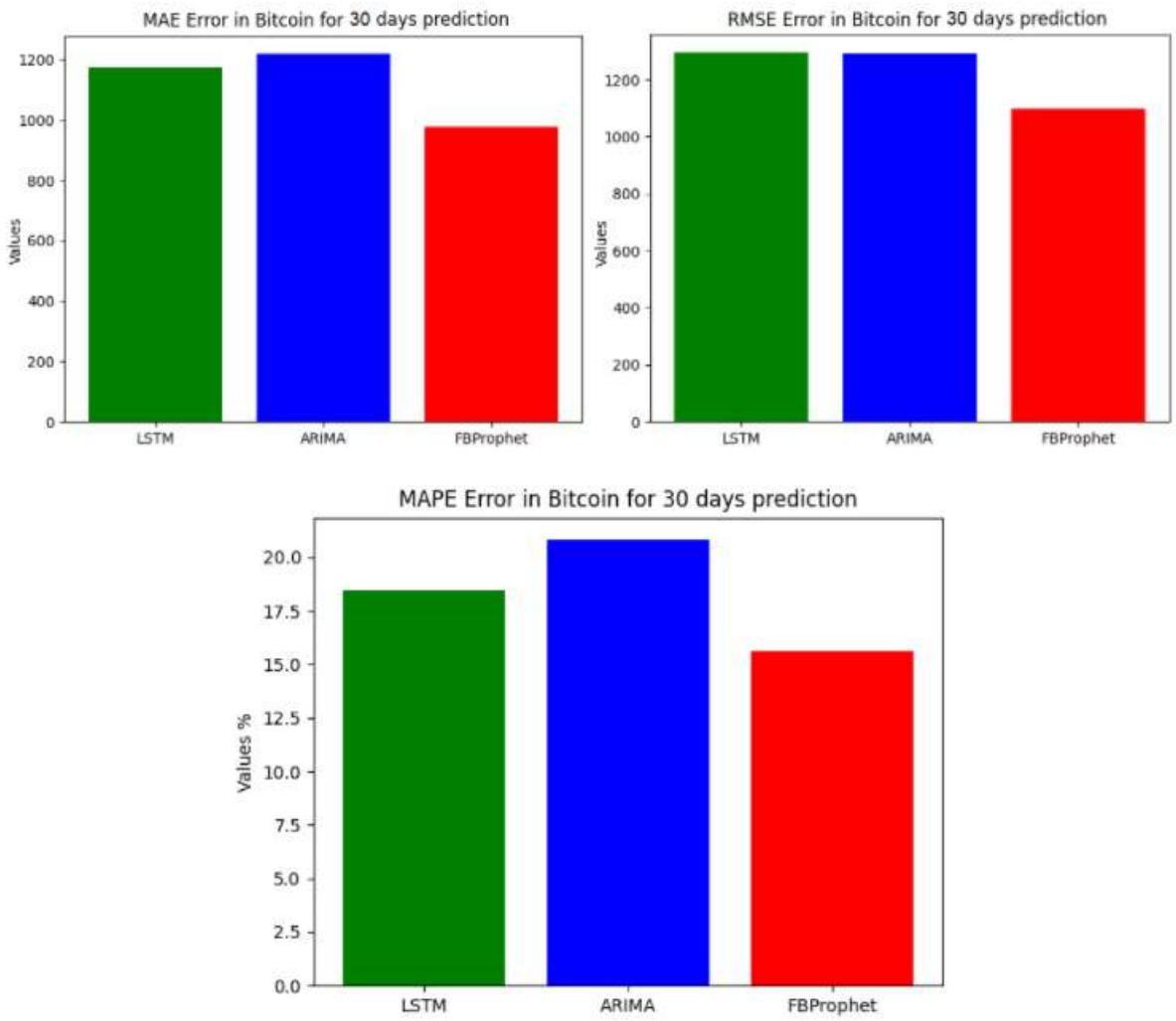
Σχήμα 4.3: Βραχυπρόθεσμη πρόβλεψη του Bitcoin με μοντέλο ARIMA.

Σε αυτό το σημείο αξίζει να σημειωθεί ότι από τη στιγμή που δεν χρησιμοποιείται κάποιου είδους εποχικότητα στο μοντέλο ARIMA, τα προβλεπόμενα δεδομένα θα έχουν τη μορφή μιας ευθείας γραμμής.



Σχήμα 4.4: Βραχυπρόθεσμη πρόβλεψη του Bitcoin με μοντέλο FBProphet.

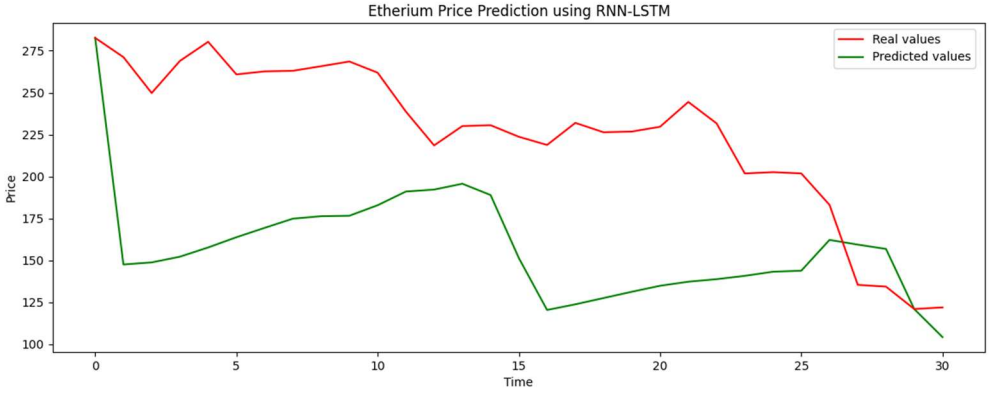
Ήδη εποπτικά φαίνεται η αδυναμία των μοντέλων να προβλέψουν τις ακραίες τιμές του κρυπτονομίσματος. Ιδιαίτερη είναι η περίπτωση του μοντέλου FBProphet, το οποίο αν και καθυστερημένα, έδειξε να έχει μια ανοδική πορεία. Για να γίνουν καλύτερα αντιληπτές όμως οι επιδόσεις των μοντέλων θα πρέπει να ληφθούν υπόψη οι δείκτες αξιολόγησης.



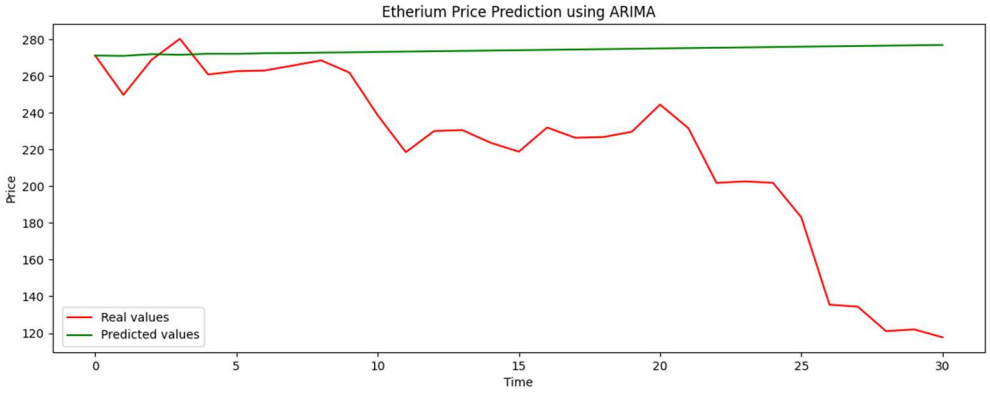
Σχήμα 4.5: Μετρικές αξιολογήσεις της βραχυπρόθεσμης πρόβλεψης των μοντέλων στο Bitcoin.

Όπως ήταν αναμενόμενο, το μοντέλο FBProphet, αν και όχι ιδιαίτερα αξιόπιστο, έκανε την καλύτερη πρόβλεψη.

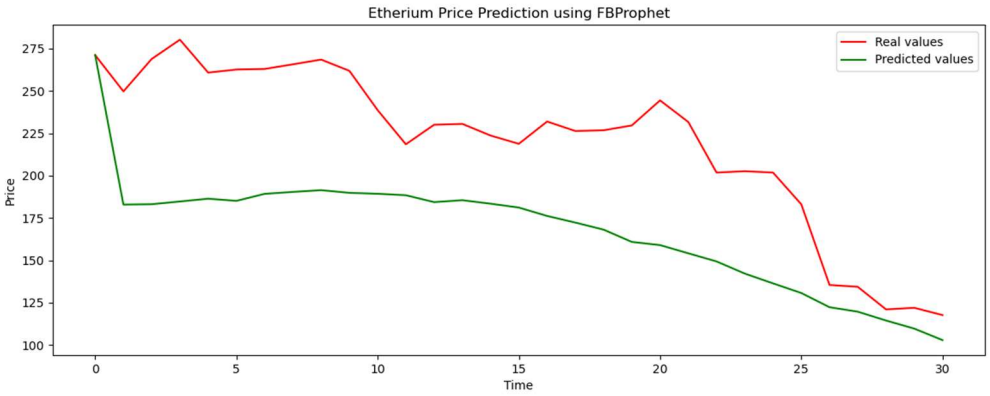
Το επόμενο κρυπτονόμισμα είναι το Ethereum. Όπως στο Bitcoin, και εδώ τα δεδομένα δοκιμής παρουσιάζουν ακραίες τιμές, αλλά σε αυτή τη περίπτωση φθίνουν.



Σχήμα 4.6: Βραχυπρόθεσμη πρόβλεψη του Ethereum με μοντέλο LSTM.



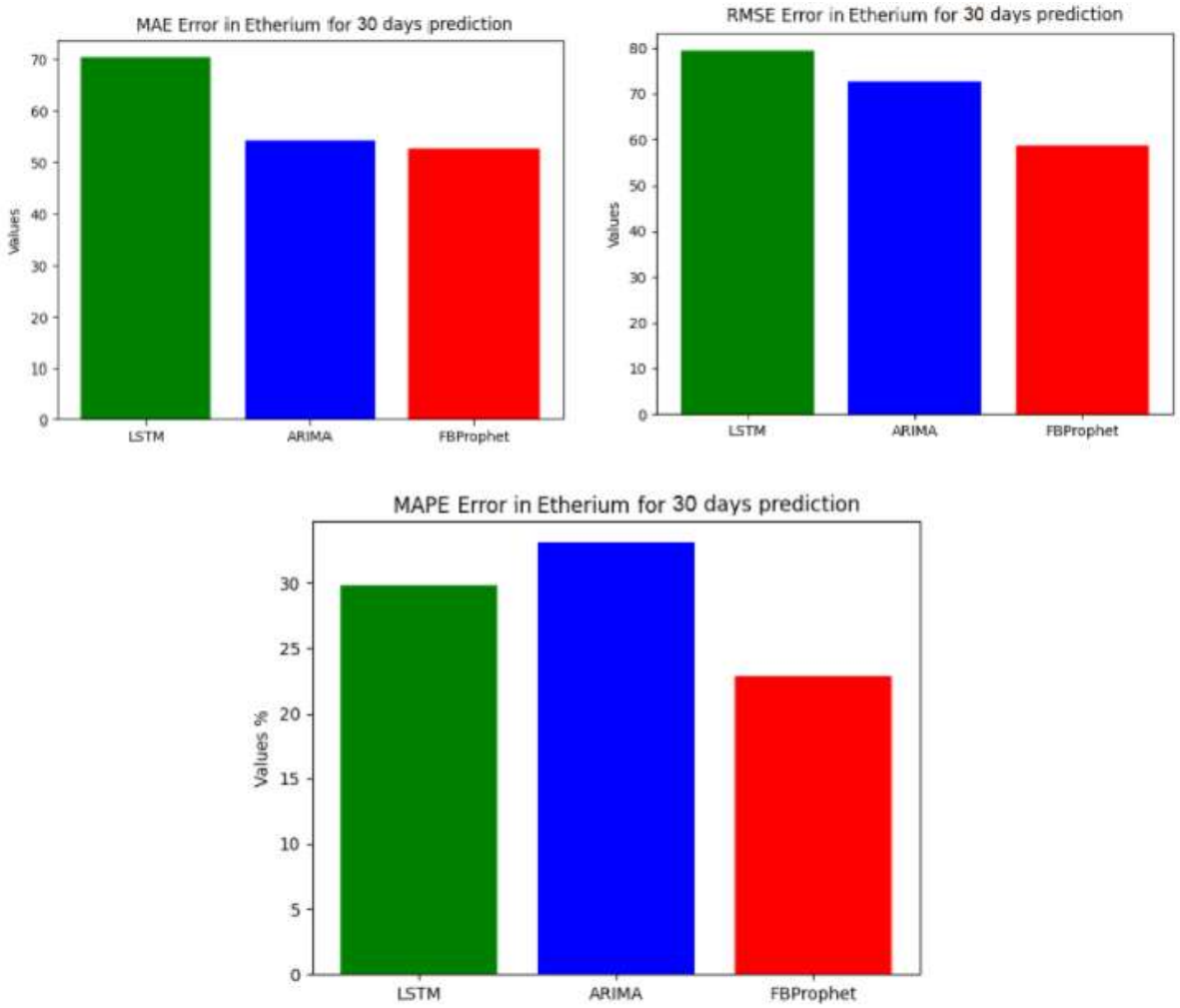
Σχήμα 4.7: Βραχυπρόθεσμη πρόβλεψη του Ethereum με μοντέλο ARIMA.



Σχήμα 4.8: Βραχυπρόθεσμη πρόβλεψη του Ethereum με μοντέλο FBProphet.

Από τα παραπάνω διαγράμματα είναι φανερό πως τα μοντέλα LSTM και FBProphet κατάφεραν, με μια σχετική επιτυχία, να προβλέψουν την μείωση των τιμών του κρυπτονομίσματος.

Οι μετρικές αξιολογήσεις των γραφημάτων αυτών φαίνονται στη συνέχεια.



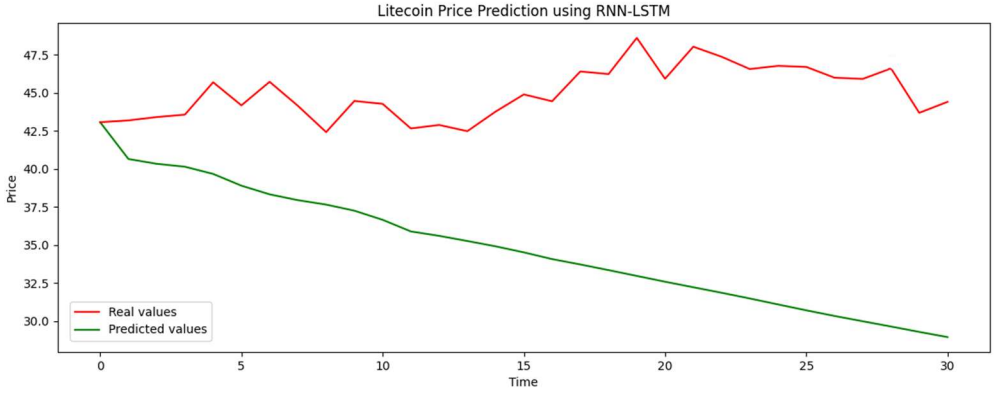
Σχήμα 4.9: Μετρικές αξιολογήσεις της βραχυπρόθεσμης πρόβλεψης των μοντέλων στο Ethereum.

Ιδιαίτερη περίπτωση εδώ είναι το πώς επηρεάζεται ο δείκτης RMSE από τις ακραίες τιμές. Αυτό φαίνεται ξεκάθαρα στο μοντέλο ARIMA, ο οποίος σε σύγκριση με τον δείκτη MAE έχει μια σημαντική διαφορά.

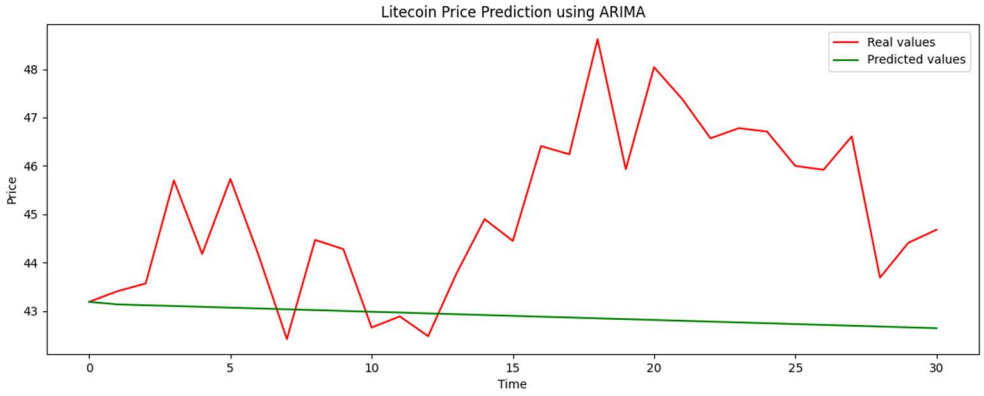
Αξιοσημείωτο επίσης είναι πως ενώ οι δείκτες MAE και RMSE θέλουν το ARIMA με καλύτερες επιδόσεις έναντι του μοντέλου LSTM, ενώ στο δείκτη MAPE το δεύτερο έχει μικρότερο ποσοστό σφάλματος.

Και σε αυτή τη περίπτωση, το βέλτιστο μοντέλο φαίνεται να είναι το FBProphet.

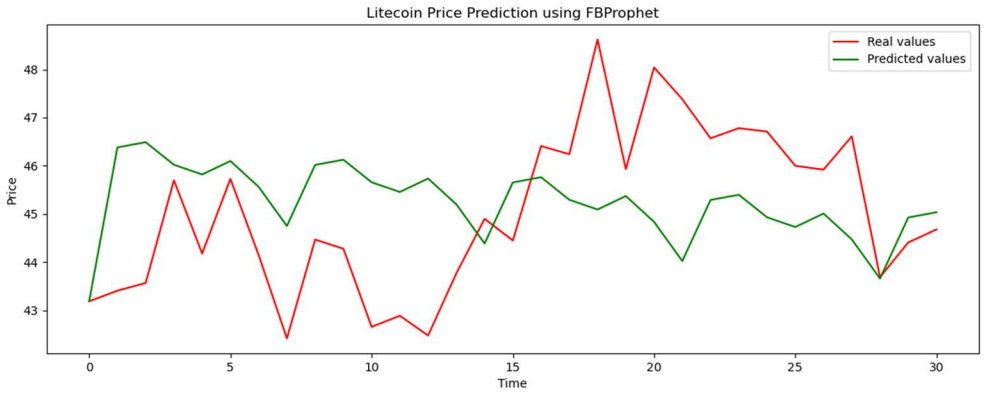
Το τελευταίο προς ανάλυση κρυπτονόμισμα είναι το Litecoin. Όπως θα φανεί και στα διαγράμματα, τα δεδομένα δοκιμής του κρυπτονομίσματος αυτού παρουσιάζουν μια σχετική σταθερότητα έναντι των άλλων δύο.



Σχήμα 4.10: Βραχυπρόθεσμη πρόβλεψη του Litecoin με μοντέλο LSTM.

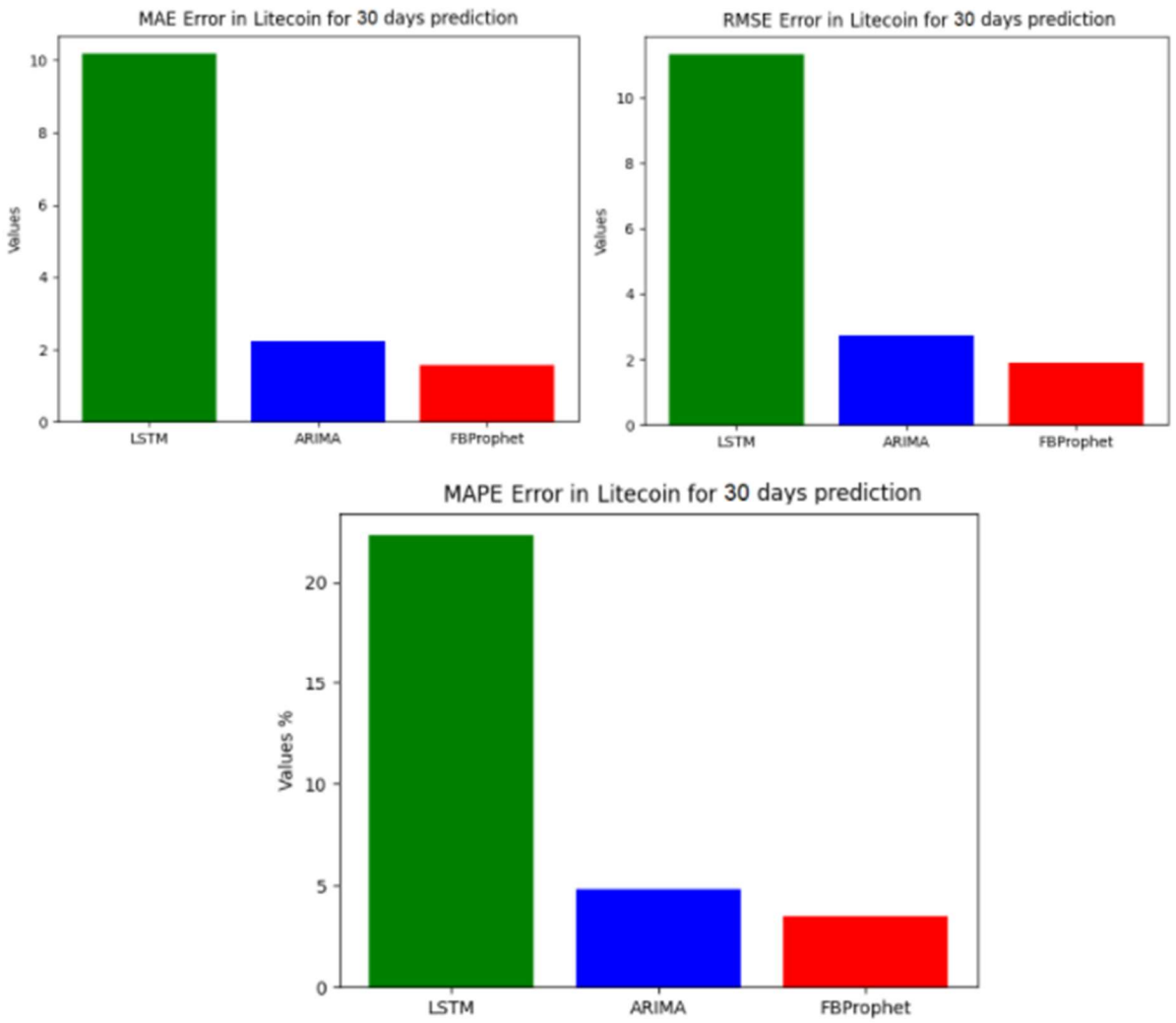


Σχήμα 4.11: Βραχυπρόθεσμη πρόβλεψη του Litecoin με μοντέλο ARIMA.



Σχήμα 4.12: Βραχυπρόθεσμη πρόβλεψη του Litecoin με μοντέλο FBProphet.

Στο συγκεκριμένο παράδειγμα, τα λίγα δεδομένα εκπαίδευσης κάνουν αισθητή την παρουσία τους στο διάγραμμα του μοντέλου LSTM, το οποίο λανθασμένα προβλέπει μια καθοδική πορεία των μελλοντικών τιμών. Αντιθέτως, τα άλλα δύο μοντέλα φαίνεται να μην επηρεάζονται από το χαρακτηριστικό αυτό, κάτι που γίνεται καλύτερα αντιληπτό στο παρακάτω σχήμα:



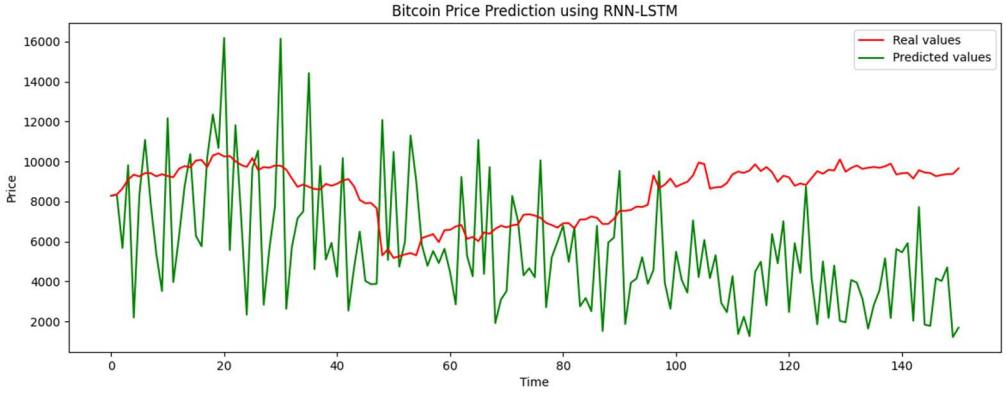
Σχήμα 4.13: Μετρικές αξιολογήσεις της βραχυπρόθεσμης πρόβλεψης των μοντέλων στο Litecoin.

Πράγματι, οι επιδόσεις των μοντέλων ARIMA και FBProphet είναι κατά πολύ καλύτερες από αυτή του LSTM, με το FBProphet να έχει ξανά ένα ελάχιστο προβάδισμα.

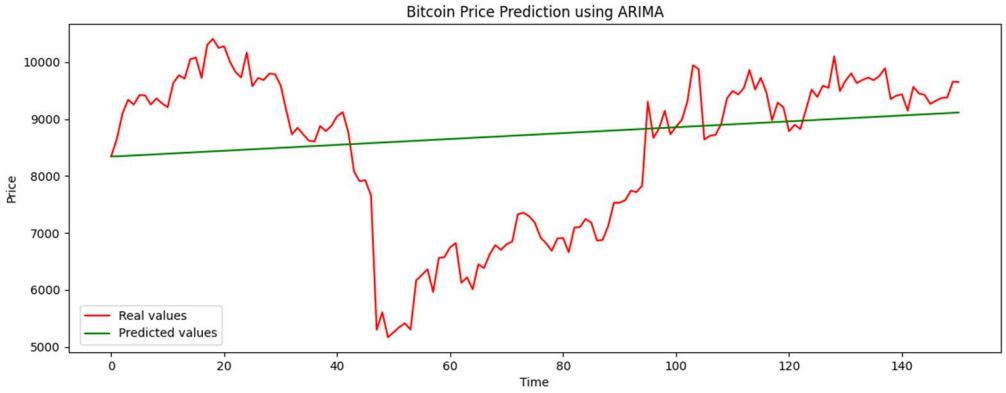
4.2.3 Μακροπρόθεσμη πρόβλεψη

Στα πλαίσια της παρούσας μελέτης έγινε επίσης απόπειρα υλοποίησης μακροπρόθεσμης πρόβλεψης στα κρυπτονομίσματα. Όπως έχει αναφερθεί, το εγχείρημα αυτό δυσκολεύει πολύ τα μοντέλα, αλλά είναι μια σημαντική διαδικασία ώστε αντληθούν πιο ολοκληρωμένα συμπεράσματα.

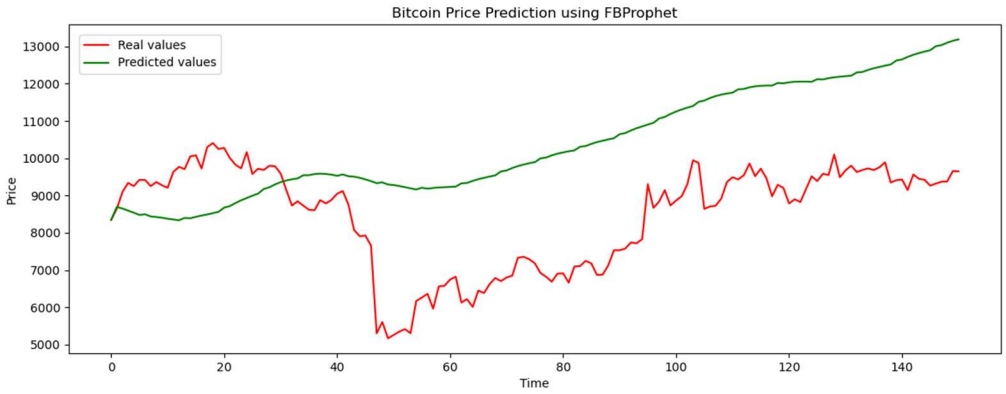
Παρακάτω φαίνονται τα γραφήματα των αποτελεσμάτων των μοντέλων στα δεδομένα του Bitcoin για την περίοδο 150 ημερών:



Σχήμα 4.14: Μακροπρόθεσμη πρόβλεψη του Bitcoin με μοντέλο LSTM.

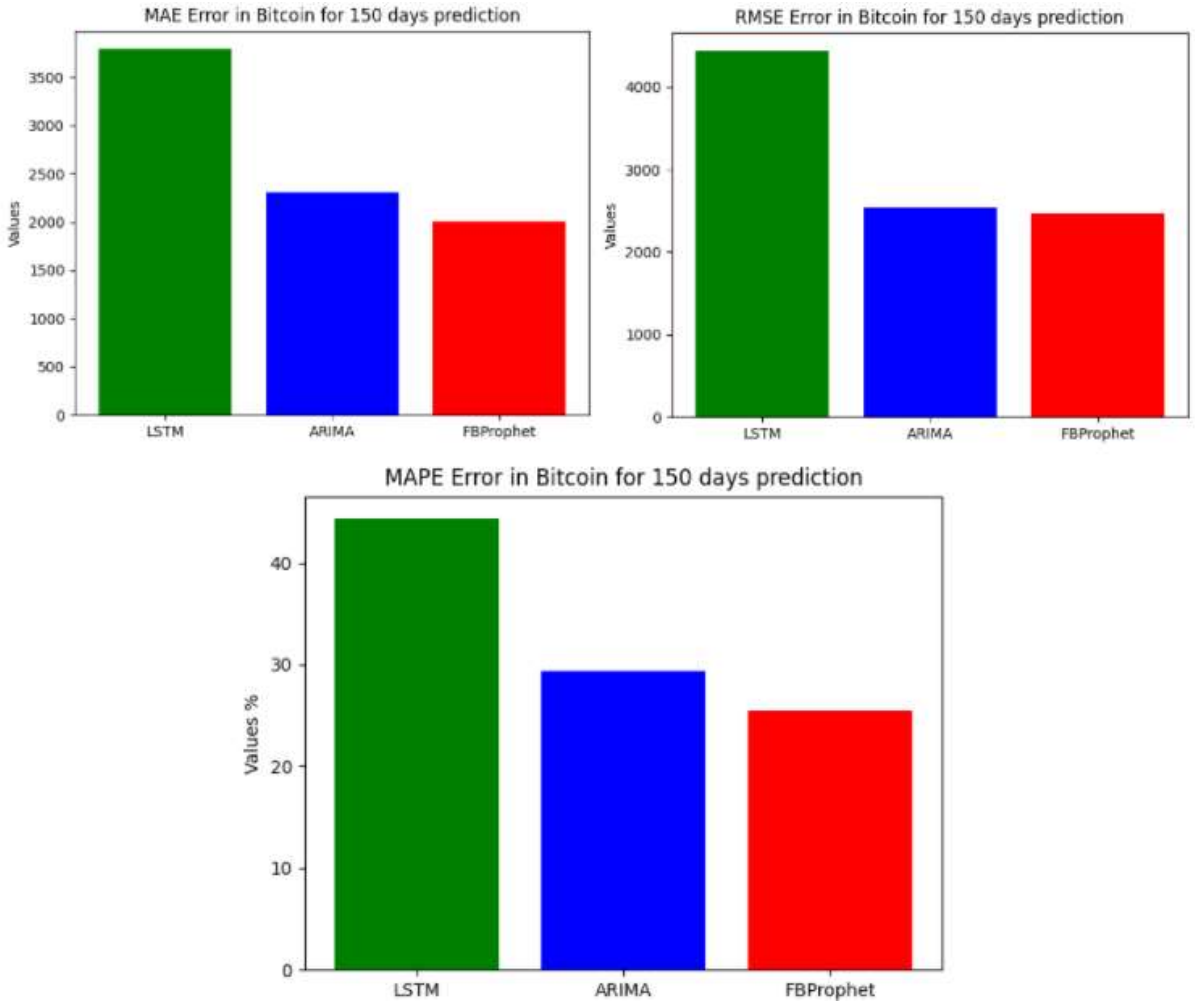


Σχήμα 4.15: Μακροπρόθεσμη πρόβλεψη του Bitcoin με μοντέλο ARIMA.



Σχήμα 4.16: Μακροπρόθεσμη πρόβλεψη του Bitcoin με μοντέλο FBProphet.

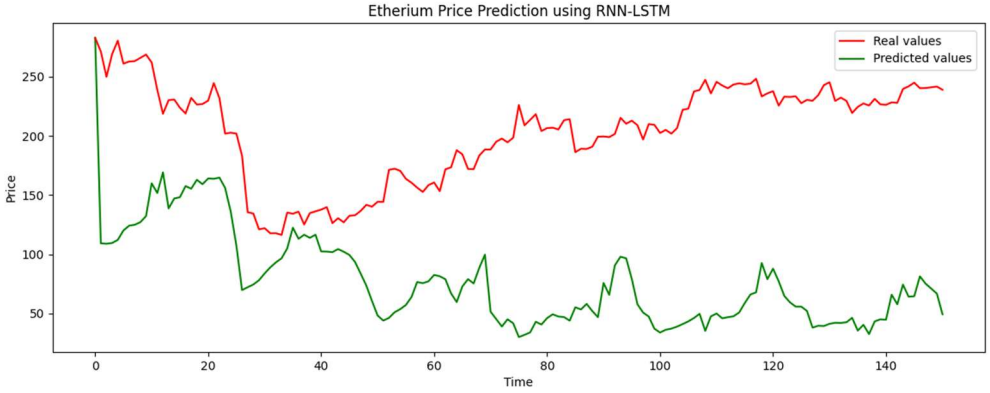
Και τα μετρικά αυτών:



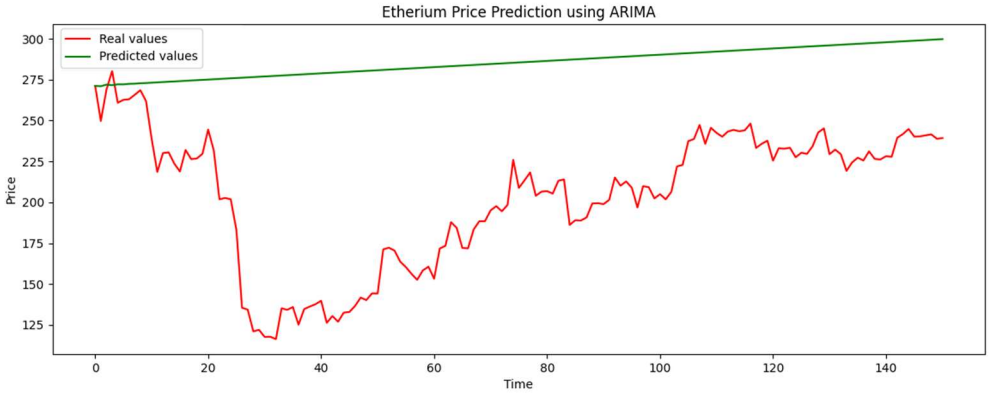
Σχήμα 4.17: Μετρικές αξιολογήσεις της μακροπρόθεσμης πρόβλεψης των μοντέλων στο Bitcoin.

Συγκριτικά με τα αντίστοιχα αποτελέσματα της βραχυπρόθεσμης πρόβλεψης, τα μοντέλα ARIMA και FBProphet παρουσίασαν μια μικρή αύξηση του ποσοστού σφάλματος, ενώ το μοντέλο LSTM υπερδιπλασίασε το ποσοστό αυτό.

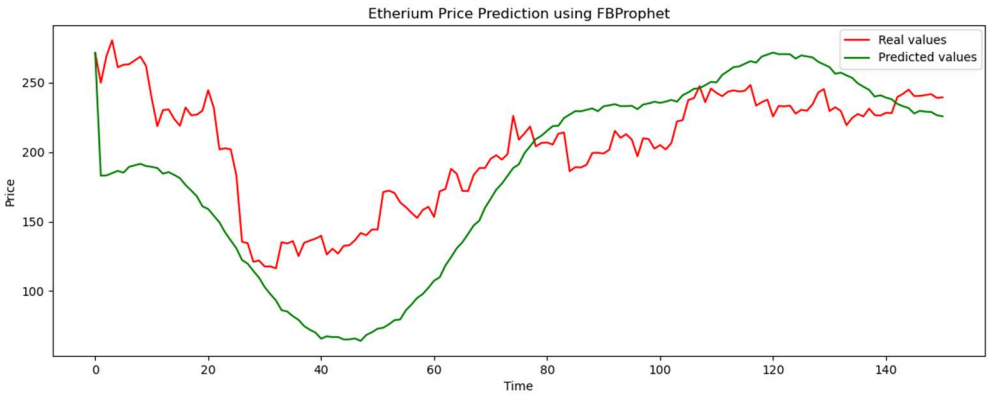
Παρακάτω φαίνονται τα γραφήματα των αποτελεσμάτων των μοντέλων στα δεδομένα του Ethereum για την περίοδο 150 ημερών:



Σχήμα 4.18: Μακροπρόθεσμη πρόβλεψη του Ethereum με μοντέλο LSTM.

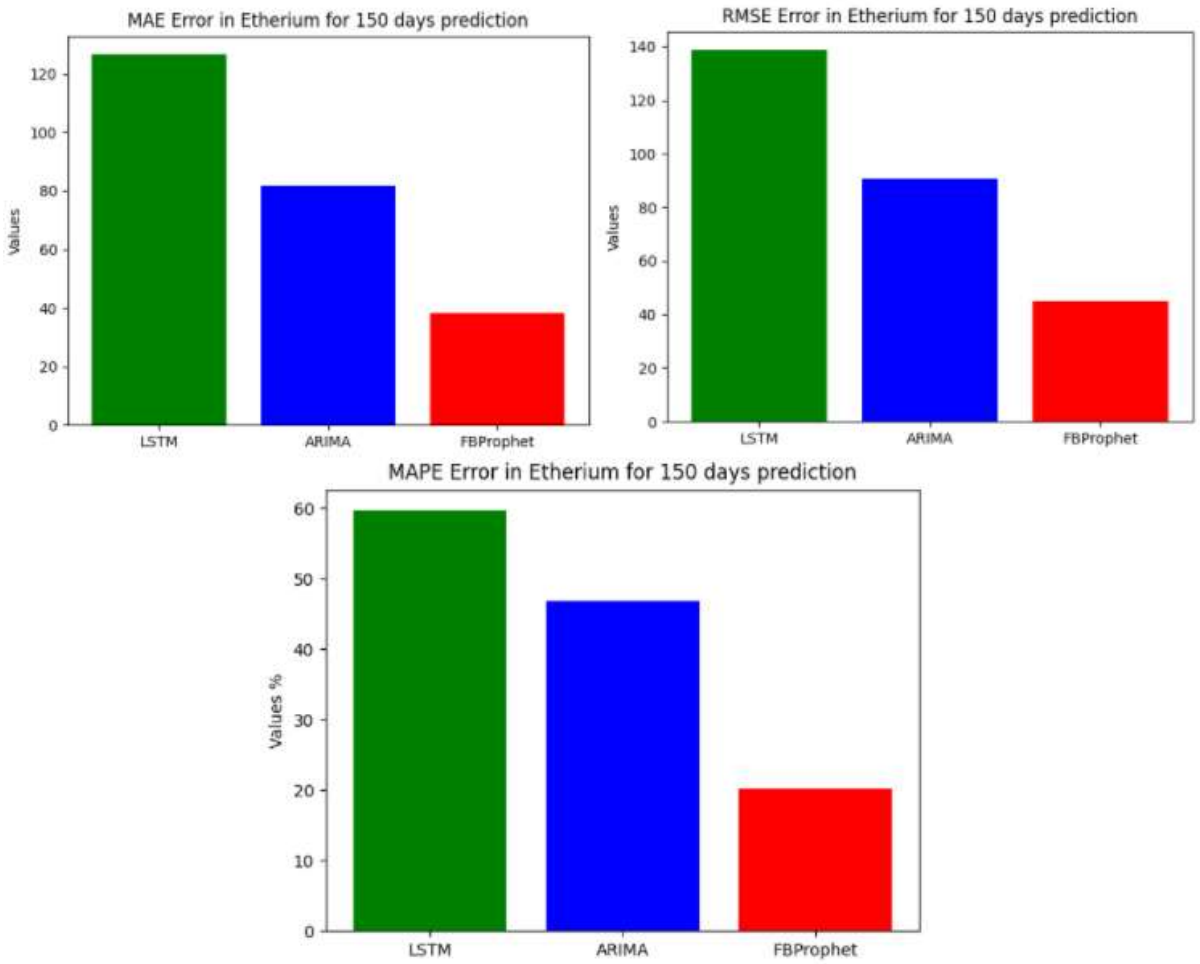


Σχήμα 4.19: Μακροπρόθεσμη πρόβλεψη του Ethereum με μοντέλο ARIMA.



Σχήμα 4.20: Μακροπρόθεσμη πρόβλεψη του Ethereum με μοντέλο FBProphet.

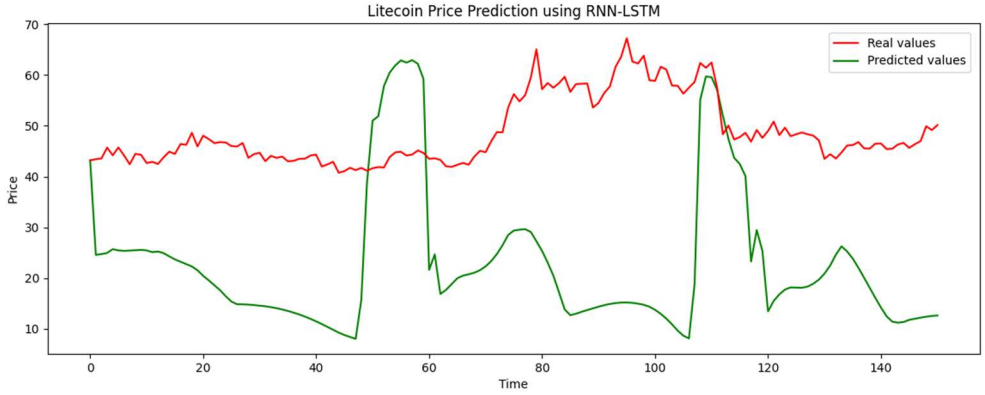
Και τα μετρικά αυτών:



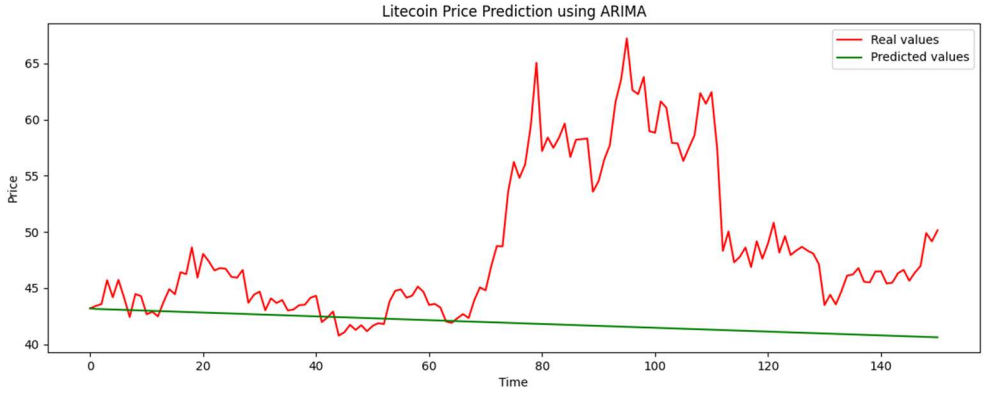
Σχήμα 4.21: Μετρικές αξιολογήσεις της μακροπρόθεσμης πρόβλεψης των μοντέλων στο Ethereum.

Στο παράδειγμα αυτό φαίνεται για άλλη μια φορά η αδυναμία του μοντέλου LSTM να κάνει μακροπρόθεσμες προβλέψεις. Το μοντέλο FBProphet από την άλλη, κατάφερε να μειώσει ελάχιστα το ποσοστό σφάλματος.

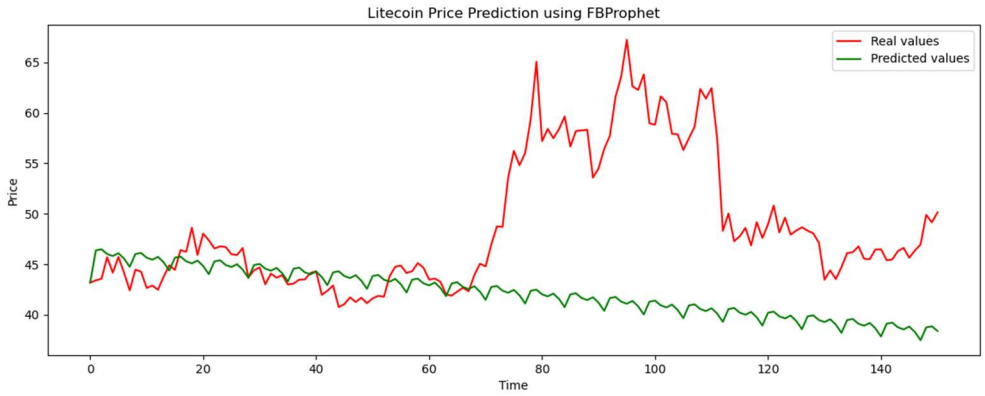
Τέλος, ακολουθούν τα αποτελέσματα των μοντέλων στα δεδομένα του Litecoin για την περίοδο 150 ημερών:



Σχήμα 4.22: Μακροπρόθεσμη πρόβλεψη του Litecoin με μοντέλο LSTM.

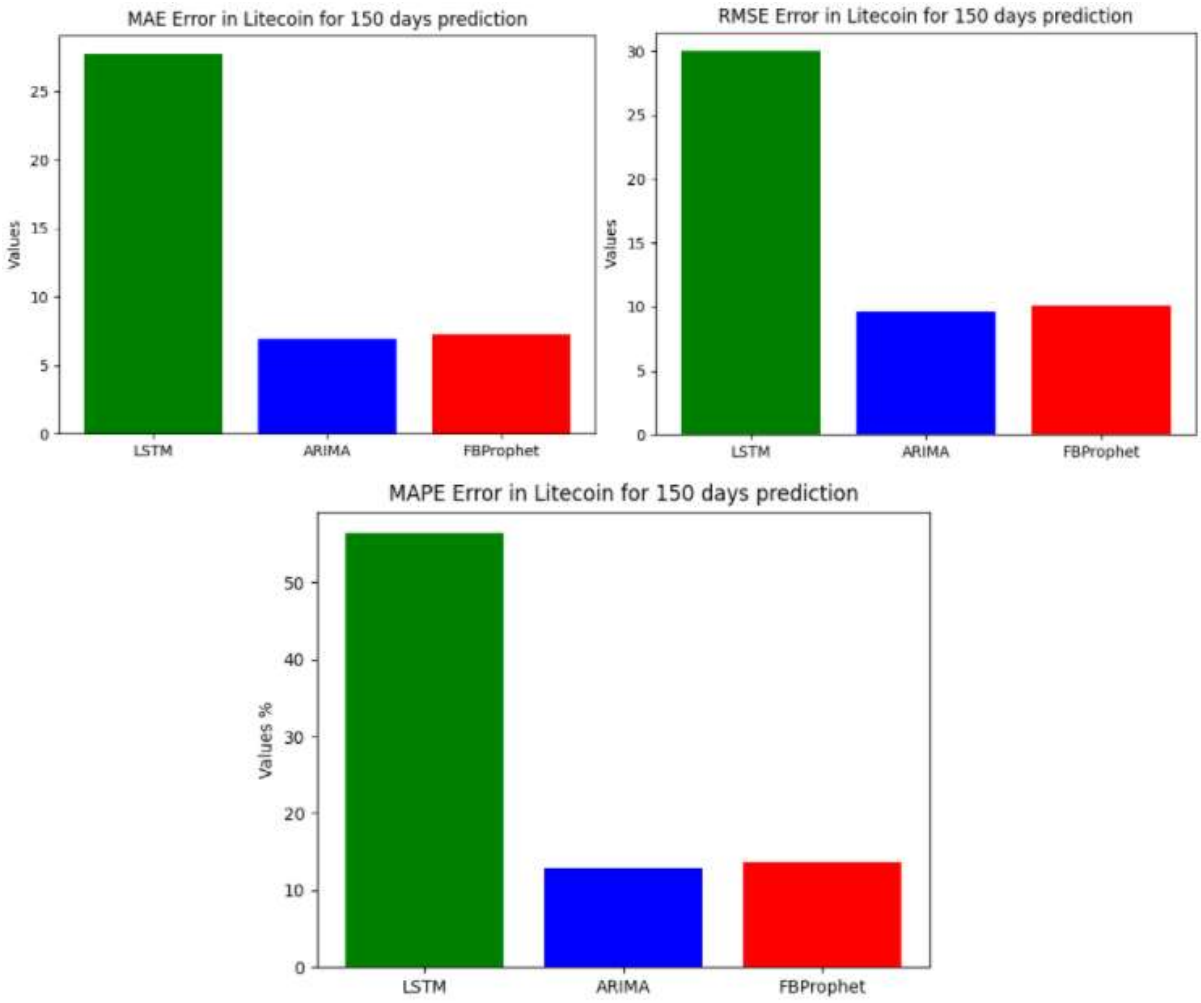


Σχήμα 4.23: Μακροπρόθεσμη πρόβλεψη του Litecoin με μοντέλο ARIMA.



Σχήμα 4.24: Μακροπρόθεσμη πρόβλεψη του Litecoin με μοντέλο FBProphet.

Εδώ φαίνεται η προσπάθεια του μοντέλου LSTM να προβλέψει δύο περιπτώσεις ακραίων τιμών, και μια ιδιαίτερη συμπεριφορά από το μοντέλο FBProphet, το οποίο δείχνει να προβλέπει μια φθίνουσα περιοδική πορεία. Πως αντικατοπτρίζεται αυτό όμως στις μετρικές αξιολογήσεις;



Σχήμα 4.25: Μετρικές αξιολογήσεις της μακροπρόθεσμης πρόβλεψης των μοντέλων στο Litecoin.

Πέρα από την προσπάθεια του LSTM, κανένα άλλο μοντέλο δεν κατάφερε να προβλέψει τις ακραίες τιμές. Παρ' όλα αυτά, οι δείκτες των μετρικών παρουσιάζουν ένα μικρό ποσοστό σφάλματος στα μοντέλα ARIMA και FBProphet, πιθανότατα επειδή πρόβλεψαν με μία σχετική επιτυχία τα δεδομένα των πρώτων ημερών.

Για να γίνει πιο αντιληπτή η σύγκριση ανάμεσα στη βραχυπρόθεσμη και τη μακροπρόθεσμη πρόβλεψη, στους παρακάτω πίνακες φαίνονται οι επιδόσεις των μοντέλων σύμφωνα με τον δείκτη MAPE.

Μοντέλο LSTM

	Βραχυπρόθεσμη πρόβλεψη	Μακροπρόθεσμη πρόβλεψη
Bitcoin	18.23%	44.33%
Ethereum	29.85%	59.71%
Litecoin	22.36%	56.43%

Πίνακας 4.1: Μετρικές αξιολογήσεις μοντέλου LSTM.

Μοντέλο ARIMA

	Βραχυπρόθεσμη πρόβλεψη	Μακροπρόθεσμη πρόβλεψη
Bitcoin	20.45%	29.41%
Ethereum	33.19%	46.87%
Litecoin	4.84%	12.81%

Πίνακας 4.2: Μετρικές αξιολογήσεις μοντέλου ARIMA.

Μοντέλο FBProphet

	Βραχυπρόθεσμη πρόβλεψη	Μακροπρόθεσμη πρόβλεψη
Bitcoin	15.66%	25.41%
Ethereum	22.90%	20.40%
Litecoin	3.52%	13.59%

Πίνακας 4.3: Μετρικές αξιολογήσεις μοντέλου FBProphet.

ΚΕΦΑΛΑΙΟ 5

ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

5.1 Συμπεράσματα

Μία απλή ανάγνωση στους παραπάνω πίνακες είναι αρκετή για να καταλάβει κανείς τη δυσκολία των μοντέλων να προβλέψουν δεδομένα για μεγάλα χρονικά διαστήματα, όπως άλλωστε ήταν αναμενόμενο.

Το μοντέλο το οποίο είχε την μεγαλύτερη αδυναμία σε ένα τέτοιο εγχείρημα ήταν το LSTM. Δυσκολεύτηκε ιδιαίτερα να προβλέψει μελλοντικές τιμές, ακόμα και όταν τα δεδομένα παρουσίαζαν μια σχετική σταθερότητα. Είχε συνολικά την μεγαλύτερη απόκλιση από την βραχυπρόθεσμη στη μακροπρόθεσμη πρόβλεψη, δίνοντας τουλάχιστον το διπλάσιο σφάλμα και στις τρεις περιπτώσεις. Όντας ένα μοντέλο νευρωνικών δικτύων, ο σχετικά μικρός αριθμός των δεδομένων που χρησιμοποιήθηκαν για να το εκπαιδεύσουν ενδεχομένως να έπαιξαν ένα σημαντικό ρόλο στην κακή απόδοσή του σε σύγκριση με τα άλλα μοντέλα.

Το μοντέλο ARIMA φάνηκε να διαχειρίζεται επίδοξα χρονοσειρές οι οποίες παρουσίαζαν μια σταθερότητα στα δεδομένα τους, αλλά είχε αυξημένα ποσοστά σφάλματος σε χρονοσειρές οι οποίες είχαν ακραίες τιμές. Συγκεκριμένα, στο κρυπτονόμισμα Litecoin έκανε μια αρκετά αποτελεσματική βραχυπρόθεσμη πρόβλεψη με μόλις 4.84% σφάλμα. Αντιθέτως, βλέποντας κάποιος τις μετρικές του στο Ethereum, γίνεται εύκολα αντιληπτή αδυναμία του να προβλέψει μεγάλες διακυμάνσεις στα δεδομένα, και σε συνδυασμό με την δυσκολία που προσθέτει η μακροπρόθεσμη πρόβλεψη έφτασε σε ποσοστά σφάλματος 46.87%.

Ποιο ήταν όμως το πιο αποδοτικό μοντέλο σύμφωνα με τις μετρικές; Ύστερα από δοκιμές προέκυψε ότι το μοντέλο FBProphet μπορεί να προσεγγίσει ικανοποιητικώς μια μελλοντική πρόβλεψη, ακόμα και για περιόδους πολλών ημερών. Σε ορισμένες περιπτώσεις πρόβλεψε ακραίες τιμές στα δεδομένα δοκιμής, αυξομειώσεις οι οποίες δεν είχαν βάση σε κάποια εποχικότητα. Μάλιστα, όπως και το μοντέλο ARIMA, εκμεταλλευόμενο τις μικρές διακυμάνσεις στα δεδομένα του κρυπτονομίσματος Litecoin, στη βραχυπρόθεσμη πρόβλεψη είχε το συγκριτικά μικρότερο ποσοστό σφάλματος ίσο με 3.52%.

Θα μπορούσε κάποιος να πει πως το ποσοστό αυτό δικαιολογεί το ρίσκο που εμπεριέχεται σε μια επένδυση χρημάτων για την αγορά κρυπτονομισμάτων. Η σημασία όμως της πρόβλεψης

σε χρονοσειρές όπως αυτές των κρυπτονομισμάτων είναι η επιτυχής πρόγνωση τον ακραίων τιμών ή έστω μιας ανώμαλης διακύμανσης. Θα θεωρούταν δηλαδή αποτελεσματικό ένα μοντέλο εάν για παράδειγμα μπορούσε να προειδοποιήσει μια μελλοντική καθοδική πορεία των τιμών ώστε να αποφευχθεί η απώλεια κεφαλαίου, ή αν κατάφερνε επιτυχώς να προβλέψει μια αύξηση στις τιμές με σκοπό την έγκυρη επένδυση. Γίνεται κατανοητό λοιπόν ότι, αν και η πρόβλεψη μιας μελλοντικής σταθερότητας στα δεδομένα είναι κάτι διαχειρίσιμο από τα στατιστικά μοντέλα, μία τέτοια πρόβλεψη θα ήταν πρακτικά ανώφελη.

Γενικά η χρήση αλγορίθμων μηχανικής μάθησης για την πρόβλεψη των τιμών κρυπτονομισμάτων δεν συνιστάται. Η φύση των παραγόντων που συμβάλλουν στην διακύμανση των τιμών είναι τέτοια που δεν επιτρέπει στα μοντέλα να τους χρησιμοποιήσουν. Για παράδειγμα, η τιμή ενός κρυπτονομίσματος μπορεί να αλλάξει ραγδαία σε πάροδο λίγων ημερών, ή ακόμα και ωρών, απλά και μόνο από μια φήμη ή από μια δημοσίευση της άποψης κάποιου προσώπου που έχει επιρροή σε μεγάλο κοινό ανθρώπων. Άλλωστε όπως έχει ήδη αναφερθεί, οι άνθρωποι, σε ατομικό επίπεδο, είναι αυτοί που διαχειρίζονται και κατά συνέπεια επηρεάζουν τα κρυπτονομίσματα, και ο παράγοντας “άνθρωπος” τείνει να είναι ιδιαίτερα απρόβλεπτος.

5.2 Μελλοντική εργασία

Η μελέτη αυτή μπορεί να χρησιμοποιηθεί ως βάση για περαιτέρω έρευνα και εργασίες. Μια προσθήκη που θα μπορούσε να γίνει είναι η χρήση της τεχνολογίας Blockchain από τα μοντέλα όπως παρουσιάζει στη μελέτη του ο Sudheer Palakurla[55]. Επίσης, μια πιο αποδοτική παραμετροποίηση, κυρίως στο μοντέλο LSTM, είναι ενδεχομένως δυνατή. Τέλος, με την συνεχή ανάπτυξη μοντέλων μηχανικής μάθησης, η μελέτη αυτή θα μπορούσε να χρησιμοποιηθεί ως μέτρο σύγκρισης της απόδοσης των καινούργιων μοντέλων σε χρονοσειρές κρυπτονομισμάτων.

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

[1] https://keras.io/examples/vision/image_classification_from_scratch/ [Accessed September 2021]

[2] https://nanonets.com/blog/semantic-image-segmentation-2020/ [Accessed September 2021]

[3] https://www.pixelsolutionz.com/application-object-detection-real-life/ [Accessed September 2021]

[4] https://www.revechat.com/blog/customer-sentiment-analysis/ [Accessed September 2021]

[5] https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b [Accessed September 2021]

[6] Dr Garrick Hileman & Michel Rauchs, "GLOBAL CRYPTOCURRENCY BENCHMARKING" STUDY (2017). https://www.crowdfundinsider.com/wp-content/uploads/2017/04/Global-Cryptocurrency-Benchmarking-Study.pdf

[7] Yukun Liu & Aleh Tsyvinski, "Risks and Returns of Cryptocurrency" (2018). https://academic.oup.com/rfs/article/34/6/2689/5912024?login=true

[8] Ryan Farell, "An Analysis of the Cryptocurrency Industry" (2015).

https://repository.upenn.edu/cgi/viewcontent.cgi?article=1133&context=wharton_research_scholars

[9] https://www.brightfinance.co/the-history-of-cryptocurrency.html [Accessed June 2021]

[10] https://www.cloudcredential.org/blog/understanding-and-working-with-blockchain-101/ [Accessed September 2021]

[11] Dylan Yaga, Peter Mell, Nik Roby, Karen Scarfone, "Blockchain Technology Overview" (2018). https://arxiv.org/ftp/arxiv/papers/1906/1906.11078.pdf

[12] https://www.researchgate.net/publication/332215097_Blockchain_Technology_in_Healthcare_A_Systematic_Review/figures?lo=1 [Accessed September 2021]

[13] Xiaoqi Li, Peng Jiang, Ting Chen, Xiapu Luo, Qiaoyan Wen, "A Survey on the Security of Blockchain Systems" (2017). https://www.sciencedirect.com/science/article/pii/S0167739X17318332

[14] Mandrita Banerjee, Junghee Lee, Kim-Kwang Raymond Choo, "A blockchain future for internet of things security" (2018).

[15] James Ogunleye, "The Concepts of Predictive Analytics" (2014). http://www.ijkie.org/IJKIE_December2014_JAMES%20OGUNLEYE.pdf

[16] Donald Brown, Ahmed Abbasi, Raymond Y. K. Lau, "Predictive Analytics INTRODUCTION" (2015). https://www.researchgate.net/publication/292526586_Predictive_Analytics_INTRODUCTION

[17] https://bigdata-madesimple.com/5-examples-predictive-analytics-travel-industry/ [Accessed September 2021]

[18] Sidath Asiri, "Machine Learning Classifiers" (2018). https://towardsdatascience.com/machine-learning-classifiers-a5cc4e1b0623

[19] Mayank Tripathi, "Knowing all about Outliers in Machine Learning" (2020). https://datascience.foundation/sciencewhitepaper/knowing-all-about-outliers-in-machine-learning

[20] Oleksandr Gerasymov, "MACHINE LEARNING FOR TIME SERIES FORECASTING" (2021). https://codeit.us/blog/machine-learning-time-series-forecasting#modeling-time-series

[21] https://www.geeksforgeeks.org/clustering-in-machine-learning/ [Accessed June 2021]

[22] https://www.javatpoint.com/regression-vs-classification-in-machine-learning [Accessed June 2021]

[23] https://in.springboard.com/blog/regression-vs-classification-in-machine-learning/ [Accessed June 2021]

[24] B. Kaminski, M. Jakubczyk, P. Szufel, "A framework for sensitivity analysis of decision trees", Central European Journal of Operations Research (2018). https://www.researchgate.net/publication/317381136_A_framework_for_sensitivity_analysis_of_decision_trees

[25] J. S. Armstrong, "Illusions in regression analysis", International Journal of Forecasting, Vol28 (2012). https://repository.upenn.edu/cgi/viewcontent.cgi?article=1190&context=marketing_papers

[26] A.D.Dongare, R.R.Kharde, Amit D.Kachare, I"ntroduction to Artificial Neural Network" (2012). https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1082.1323&rep=rep1&type=pdf

[27] https://www.researchgate.net/publication/321259051_Prediction_of_wind_pressure_coefficients_on_building_surfaces_using_Artificial_Neural_Networks/figures?lo=1&utm_source=google&utm_medium=organic [Accessed September 2021]

[28] Jessica Lin Eamonn Keogh Stefano Lonardi Bill Chiu, "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms" (2003). https://www.cs.ucr.edu/~eamonn/SAX.pdf

[29] Vaibhav Kumar, M. L. Garg, “Predictive Analytics: A Review of Trends and Techniques” (2018). https://www.researchgate.net/publication/326435728_Predictive_Analytics_A_Review_of_Trends_and_Techniques

[30] https://towardsdatascience.com/time-series-in-python-part-2-dealing-with-seasonal-data-397a65b74051 [Accessed September 2021]

[31] Σπυρίδων Ι. Αθανασόπουλος “Πειραματική μελέτη, αξιολόγηση και σύγκριση μεθόδων πρόβλεψης της Οριακής Τιμής Συστήματος της ηλεκτρικής ενέργειας στην Ελληνική Αγορά Ενέργειας.” (2021).

[32] https://www.kdnuggets.com/2020/09/introduction-time-series-analysis-python.html [Accessed September 2021]

[33] E. E. Holmes, M. D. Scheuerell, and E. J. Ward, “Applied Time Series Analysis for Fisheries and Environmental Sciences” (2021). https://nwfsc-timeseries.github.io/atsa-labs/

[34] https://machinelearningmastery.com/white-noise-time-series-python/ [Accessed September 2021]

[35] Rob J Hyndman, George Athanasopoulos “Forecasting: Principles and Practice” (2018).

[36] https://www.oreilly.com/library/view/hands-on-machine-learning/9781788992282/15c9cc40-bea2-4b75-902f-2e9739fec4ae.xhtml [Accessed September 2021]

[37] Tim Smith, “Autocorrelation” (2021).

[38] Gary Napier, “Time Series” (2020). https://bookdown.org/gary_a_napier/time_series_lecture_notes/

[39] Sima Siami-Namini, Neda Tavakoli, Akbar Siami Namin, “A Comparison of ARIMA and LSTM in Forecasting Time Series” (2018). https://par.nsf.gov/servlets/purl/10186768

[40] Fazle Karima , Somshubra Majumdarb , Houshang Darabia , Samuel Harforda, “Multivariate LSTM-FCNs for Time Series Classification” (2019). https://arxiv.org/pdf/1801.04503.pdf

[41] J.J. Wang, J.Z. Wang, Z.G. Zhang and S.P. Guo, “Stock index forecasting based on a hybrid model” (2012).

[42] L.C. Kyungjoo, Y. Sehwan and J. John, “Neural Network Model vs. SARIMA Model In Forecasting Korean Stock Price Index", Is(2007). https://www.researchgate.net/publication/237820911_Neural_Network_Model_vs_SARIMA_Model_In_Forecasting_Korean_Stock_Price_Index_KOSPI

[43] https://eclass.unipi.gr/modules/document/file.php/DES103/%CE%94%CE%B1%CE%B3%CE%BF%CF%8D%CE%BC%CE%B1%CF%82/ARIMA/ [Accessed September 2021]

[44] Douglas C. Montgomery, Cheryl L. Jennings, Murat Kulahci, " Time Series Analysis and Forecasting" (2015). http://ndl.ethernet.edu.et/bitstream/123456789/28722/1/Douglas%20C.%20Montgomery_2015.pdf

[45] Rob J Hyndman, George Athanasopoulos "Forecasting: Principles and Practice" (2018).

[46] Andr´es M. Alonso, Carolina Garc´ıa-Martos, "Time Series Analysis" (2012).

[47] Sio-Iong Ao, " Applied Time Series Analysis and Innovative Computing" (2009)

[48] Meyler, Aidan and Kenny, Geoff and Quinn, Terry, "Forecasting irish inflation using ARIMA models" (1998) https://mpra.ub.uni- muenchen.de/11359/1/cbi_3RT98_inflationarima.pdf

[49] Peipei Wang, Xinqi Zheng, Jiayang Li and Bangren Zhua, "Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics" (2020). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7328553/

[50] Sean J. Taylor, Benjamin Letham, "Forecasting at Scale" (2017). https://peerj.com/preprints/3190/

[51] Traiani Stari, "Seasonal Stability in Time Series of Zooplankton Abundance" (2010) https://www.strath.ac.uk/media/departments/mathematics/research/groups/marinegroup/phdtheses/traiani_stari.pdf

[52] https://www.investopedia.com/articles/investing/031416/bitcoin-vs-ethereum-driven-different-purposes.asp [Accessed September 2021]

[53] https://www.investopedia.com/articles/investing/042015/bitcoin-vs-litecoin-whats-difference.asp [Accessed September 2021]

[54] https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/ [Accessed September 2021]

[55] Sudheer Palakurla, "Predictive Analysis of Cryptocurrency using Machine Learning with Blockchain technology" 2020. https://www.researchgate.net/publication/347453695_Predictive_Analysis_of_Cryptocurrency_using_Machine_Learning_with_Blockchain_technology