ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ & ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Ολοκληρωμένες μεθοδολογίες κατάταξης επενδυτικών αγαθών

Διπλωματική εργασία

Κουτούγερα Άννα

Επιβλέπων : Βασίλειος Ασημακόπουλος

Καθηγητής Ε.Μ.Π.

Υπεύθυνος : Ευάγγελος Σπηλιώτης

Διδάκτωρ Ε.Μ.Π.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ & ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Ολοκληρωμένες μεθοδολογίες κατάταξης επενδυτικών αγαθών

Διπλωματική εργασία

Κουτούγερα Άννα

Επιβλέπων : Βασίλειος Ασημακόπουλος
Καθηγητής Ε.Μ.Π.

Υπεύθυνος : Ευάγγελος Σπηλιώτης
Διδάκτωρ Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 10 Μαρτίου 2023.

(Υπογραφή)	(Υπογραφή)	(Υπογραφή)
……	……	……
Βασίλειος Ασημακόπουλος	Ιωάννης Ψαρράς	Δημήτριος Ασκούνης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

(Υπογραφή)

.....

Κουτούγερα Άννα

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π

Περίληψη

Σύμφωνα με την υπόθεση της αποτελεσματικής αγοράς, η φύση του χρηματιστηρίου καθιστά αδύνατη την πρόβλεψη των τιμών των επενδυτικών αγαθών. Οποιαδήποτε αλλαγή στην τιμή των αγαθών προκύπτει από την άμεση αντίδραση της αγοράς σε ένα στιγμιαίο γεγονός ειδήσεων ή σε μία αναπάντεχη μεταβολή της προσφοράς και της ζήτησης. Έτσι, πρακτικά δεν δύναται να αναπτυχθεί μια επενδυτική στρατηγική που να αποδίδει συστηματικά καλύτερα από την αγορά με βάση τα κλασικά κριτήρια κινδύνου και απόδοσης. Ωστόσο, υπάρχουν άφθονα αντιπαραδείγματα αυτής της υπόθεσης, πολλά από τα οποία βασίζονται στην τεχνική ανάλυση. Σε αυτήν τη λογική βασίστηκε και ο διαγωνισμός M6, σκοπός του οποίου ήταν η εμπειρική μελέτη του εν λόγω παραδόξου έτσι ώστε να μπορέσει να εξηγηθεί από τη μία η περιορισμένη απόδοση πολλών ενεργών επενδυτών και από την άλλη η εξαιρετική κερδοφορία που έχει επιτευχθεί από γνωστούς επενδυτές όπως ο Warren Buffet, Peter Lynch και George Soros, οι οποίες είναι αδύνατο να δικαιολογηθούν από απλή τύχη.

Η παρούσα διπλωματική εργασία ασχολείται με το πρώτο κεφάλαιο του διαγωνισμού M6 στο οποίο αξιολογείται η ικανότητα πρόβλεψης της σχετικής θέσης ενδεικτικών επενδυτικών αγαθών, της κατάταξής τους δηλαδή σε πέντε κλάσεις βάσει απόδοσης. Για το λόγο αυτό, ακολουθώντας τη δομή του διαγωνισμού, επιλέγονται τα ιστορικά δεδομένα πενήντα μετοχών του S&P 500 για τα οποία παράγονται προβλέψεις ως προς τη σχετική τους κατάταξη, τόσο σε βραχυπρόθεσμο, όσο και σε μεσοπρόθεσμο ορίζοντα, χρησιμοποιώντας στατιστικές μεθόδους και τεχνικές μηχανικής μάθησης.

Αρχικά, γίνεται μια αναφορά στις βασικές έννοιες του χρηματιστηρίου, τη μεγάλη πρόκληση που αποτελεί για τους επενδυτές η ανάλυση των κινήσεων και των συμπεριφορών της αγοράς, καθώς και οι βασικές προσεγγίσεις που χρησιμοποιούνται, όπως η τεχνική ανάλυση, στη χρήση της οποίας βασίστηκε και η παρούσα μελέτη.

Έπειτα, αναλύονται οι μέθοδοι πρόβλεψης που χρησιμοποιήθηκαν για την παραγωγή των προβλέψεων, πρώτα σε θεωρητικό επίπεδο και εν συνεχεία σε πειραματικό με επεξήγηση της μεθοδολογίας που εφαρμόστηκε. Γίνεται διεξοδική επισκόπηση τόσο των στατιστικών μεθόδων που χρησιμοποιήθηκαν ως σημεία αναφοράς, όσο και των τεχνικών μηχανικής μάθησης.

Τέλος, αξιολογείται το σύνολο των μεθόδων ως προς την απόδοσή τους, σύμφωνα με τις παραδοχές και τις μετρικές του διαγωνισμού M6. Παρατηρώντας και αξιολογώντας τα προβλεπόμενα αποτελέσματα, γίνεται επιλογή των βέλτιστων μεθόδων και προσαρμογή τους στις συνθήκες του διαγωνισμού M6 για την παραγωγή αποτελεσμάτων για το πρώτο και δεύτερο τρίμηνο υποβολών.

Λέξεις Κλειδιά: Τεχνικές Προβλέψεων, Πρόβλεψη Χρονοσειρών, Τεχνική Ανάλυση, Μηχανική Μάθηση, Διαγωνισμός M6, Στατιστικές τεχνικές, Χρηματιστήριο, Μετοχές, Συσταδοποίηση, Μέθοδοι Ταξινόμησης

Abstract

According to the Efficient Market Hypothesis, it is not possible to beat the market by developing a strategy based on a historical price series. Any change in price represents the immediate reaction to an instantaneous news event or to new and unexpected changes in supply/demand figures, which implies that it is not possible to develop an investment strategy that can beat the market under the classical criteria of risk and return. However, there are many critics against this hypothesis, much of which is based on the use of technical analysis. The M6 competition was built on this logic, the purpose of which was to empirically investigate this paradox in order to explain on the one hand the poor performance of active funds, and on the other hand the excellent returns achieved by well-known investors such as Warren Buffet, Peter Lynch and George Soros which cannot be justified by mere chance.

This thesis attempts to deal with the first part of the competition in which the ability to accurately predict the relative position of individual shares (assets), i.e. their classification in five basic classes in terms of their expected returns, is evaluated. For this reason, following the structure of the competition, predictions are made using the historical data of 50 S&P 500 stocks considering both short- and medium-term forecast horizons, while using statistical methods and machine learning techniques.

First of all, a reference is made to the basic concepts of the stock market, the great challenge that the investors face when analysing market movements and behaviors, as well as the main approaches used for said tasks, such as the technical analysis on which the present study is based on.

Afterwards, the forecasting methods used to produce the forecasts are analyzed, first at a theoretical and then at an experimental level, while an explanation of the methodology applied is provided. Both statistical methods used as benchmarks and machine learning techniques are thoroughly reviewed.

Finally, all the methods are evaluated in terms of their performance, using the metric and the assumptions of the M6 competition. By observing and evaluating the results, the most effective methods are selected and adapted to the conditions of the M6 competition to produce results for the first and second quarter submissions.

Keywords: Forecasting Techniques, Time Series Forecasting, Technical Analysis, Machine Learning, M6 Competition, Statistical Techniques, Stock Market, Stocks, Clustering, Classification Methods

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια των ερευνητικών δραστηριοτήτων της Μονάδας Προβλέψεων και Στρατηγικής κατά το ακαδημαϊκό έτος 2022-2023. Η μονάδα υπάγεται στον Τομέα Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων της Σχολής Ηλεκτρολόγων Μηχανικών και Μηχανικών Ηλεκτρονικών Υπολογιστών, του Εθνικού Μετσόβιου Πολυτεχνείου.

Αρχικά, θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Βασίλειο Ασημακόπουλο για την ευκαιρία που μου έδωσε να συμμετάσχω στην ομάδα της Μονάδας και με την ανάθεση αυτής της διπλωματικής εργασίας να ασχοληθώ σε βάθος τόσο με τον τομέα των προβλέψεων, όσο και με τον τομέα της μηχανικής μάθησης, κλάδοι που μέσω της ευκαιρίας αυτής, τροφοδότησαν το ενδιαφέρον μου και για μελλοντική επαγγελματική ενασχόληση.

Θερμές ευχαριστίες οφείλω και στον υπεύθυνο της παρούσας εργασίας, Ευάγγελο Σπηλιώτη, ερευνητικό συνεργάτη και συντονιστή της Μονάδας Προβλέψεων και Στρατηγικής. Η καθοδήγησή του για την εκπόνηση της παρούσας εργασίας, καθώς και οι συμβουλές του και η συνεισφορά του, αποτέλεσαν μεγάλη πηγή γνώσης και εμπειρίας.

Επιπλέον, θα ήθελα ακόμα να ευχαριστήσω τους καθηγητές Ιωάννη Ψαρρά και Δημήτριο Ασκούνη, τόσο για τη συμμετοχή τους στην επιτροπή εξέτασης, όσο κυρίως για τη γνώση και την εμπειρία που έχω αποκτήσει, ακούγοντας τους από έδρας και συμμετέχοντας στα μαθήματά τους.

Τέλος, προσωπικά θα ήθελα να κάνω ιδιαίτερη αναφορά και στα υπόλοιπα μέλη της Μονάδας Προβλέψεων και Στρατηγικής, Αναστάσιος Καλτσούνης, Αρτέμιος-Ανάργυρος Σεμένογλου, Ευάγγελος Θεοδώρου, Κωνσταντίνος Γκρέζιος για την υποστήριξη τους σε κάθε επίπεδο καθώς και τους γονείς μου Μαρία Ακουμιανάκη και Ανδρέα Κουτούγερα για την ψυχολογική και οικονομική υποστήριξη που μου παρείχαν στα χρόνια των σπουδών μου.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΕΧΟΜΕΝΑ ΣΧΗΜΑΤΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ ΠΙΝΑΚΩΝ

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΤΩΝ ΑΓΟΡΩΝ

Αυτό το κεφάλαιο παρέχει μια σύντομη εισαγωγή στις βασικές έννοιες του χρηματιστηρίου, τη μεγάλη πρόκληση που αποτελεί για τους επενδυτές η ανάλυση των κινήσεων και των συμπεριφορών της αγοράς λόγω της θορυβώδους και μη γραμμικής φύσης των δεδομένων καθώς και στα διαφορετικά διαθέσιμα είδη αναλύσεων και προβλέψεων που έχουν μελετηθεί τις τελευταίες δεκαετίες.

1.1 ΕΙΣΑΓΩΓΗ

Οι χρηματοοικονομικές αγορές παίζουν ζωτικό ρόλο στην οικονομική και κοινωνική οργάνωση της σύγχρονης κοινωνίας επηρεάζοντας την οικονομική ανάπτυξη πολλών χωρών παγκοσμίως. Σε αυτά τα είδη αγορών, η επιτυχία ενός επενδυτή εξαρτάται από την ποιότητα των πληροφοριών που χρησιμοποιεί για να υποστηρίξει τη λήψη αποφάσεων καθώς και από το πόσο γρήγορα είναι σε θέση να λάβει αυτές τις αποφάσεις. Για αυτό το λόγο, η ανάλυση των κινήσεων της αγοράς έχει μελετηθεί ευρέως στους τομείς των χρηματοοικονομικών, της μηχανικής και των μαθηματικών τις τελευταίες δεκαετίες [1] [2]. Με την έλευση των χρηματιστηρίων, δημιουργήθηκαν πολλές πλατφόρμες διαθέσιμες για συναλλαγές και επενδύσεις μέσω του Διαδικτύου αυξάνοντας την προσβασιμότητα σε μεγάλο πλήθος ανθρώπων, φέρνοντας την επανάσταση στον τρόπο με τον οποίο οι άνθρωποι αγοράζουν και πωλούν μετοχές καθώς και οδηγώντας στη δημιουργία διαφορετικών τύπων κεφαλαίων ανάλογα με την ανάληψη ρίσκου που άνθρωποι ή ιδρύματα είναι διατεθειμένοι να αναλάβουν. Οι κυβερνήσεις των περισσότερων χωρών επενδύουν ένα μέρος των ταμείων υγειονομικής περίθαλψης, απασχόλησης ή συνταξιοδότησης στα χρηματιστήρια για να επιτύχουν καλύτερες αποδόσεις. Οι χρηματοοικονομικές αγορές έχουν εξελιχθεί γρήγορα σε μια ισχυρή και διασυνδεδεμένη παγκόσμια αγορά. Αυτές οι εξελίξεις και ο εκσυγχρονισμός των χρηματοοικονομικών συναλλαγών και των πληροφοριακών συστημάτων, έχουν δημιουργήσει νέες ευκαιρίες αλλά θέτουν επίσης και μια ολόκληρη σειρά νέων προκλήσεων. Αν και πολλοί επενδυτές και ερευνητές ενδιαφέρθηκαν να αναπτύξουν μοντέλα προβλέψεων [3], η ανάλυση των κινήσεων της αγοράς και των διακυμάνσεων των τιμών είναι εξαιρετικά δύσκολη λόγω της δυναμικής, μη γραμμικής, μη στάσιμης και θορυβώδους φύσης των αγορών [4]. Σύμφωνα με τους Zhong και Enke [5], οι χρηματιστηριακές αγορές επηρεάζονται από πολλούς αλληλένδετους οικονομικούς, πολιτικούς και ψυχολογικούς παράγοντες. Η αστάθεια της αγοράς αυξάνεται με την πάροδο των χρόνων, οι προβλέψεις πολλές φορές είναι ασαφείς, πρέπει να λαμβάνεται υπόψη όχι μόνο η δυναμική της αγοράς αλλά και η δυναμική των μεμονωμένων μετοχών και όσο καλή και να είναι η ανάλυση στο τέλος, το μέλλον παραμένει πάντα αβέβαιο. Παρόλα αυτά, η ακριβής πρόβλεψη της χρηματιστηριακής αγοράς μπορεί να βοηθήσει τους επενδυτές να λάβουν καλύτερες

αποφάσεις και ακόμα και μικρές βελτιώσεις στην απόδοση των προβλέψεων μπορεί να είναι πολύ κερδοφόρες.

1.2 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

1.2.1 Επενδυτικό αγαθό

Ο όρος επενδυτικό αγαθό γνωστός και ως «security» αφορά ένα πιστοποιητικό ή άλλο χρηματοπιστωτικό μέσο που έχει νομισματική αξία και μπορεί να διαπραγματευτεί. Στην ουσία μπορεί να είναι ένα εμπορεύσιμο και ανταλλάξιμο χρηματοοικονομικό περιουσιακό στοιχείο που χρησιμοποιείται για την άντληση κεφαλαίων σε δημόσιες και ιδιωτικές αγορές. Συνήθως ταξινομούνται είτε σαν ίδια κεφάλαια (equity securities) όπως οι μετοχές παρέχοντας δικαιώματα ιδιοκτησίας στους κατόχους, είτε σαν χρεόγραφα (debt securities) όπως τα ομόλογα, δηλαδή δάνεια που αποπληρώνονται με περιοδικές πληρωμές, είτε σαν ένας συνδυασμός τους (hybrid securities).

1.2.2 Μετοχές

Μετοχή (stock) είναι ένα μερίδιο κυριότητας επί μιας εταιρείας. Οι εταιρείες πωλούν μετοχές προκειμένου να συγκεντρώσουν κεφάλαια. Σε αντάλλαγμα, οι μέτοχοι μπορούν να εισπράξουν μερίσματα, να τους διανεμηθούν κέρδη επί των μετοχών τους ή να εισπράξουν την απόδοση της επένδυσής τους σε περίπτωση που ανέβει η τιμή των μετοχών. Οι τιμές των μετοχών αυτών ορίζονται με βάση την προσφορά και τη ζήτηση και η αγοραπωλησία τους πραγματοποιείται κυρίως σε χρηματιστήρια σε συμμόρφωση πάντα και με τους κυβερνητικούς κανονισμούς. Οι μετοχές αποτελούν τη βάση πολλών χαρτοφυλακίων μεμονωμένων επενδυτών και ανάλογα με τον τύπο της μετοχής, κοινό ή προνομιούχο, που κατέχει ένας μέτοχος, καθορίζονται τα δικαιώματα και τα οφέλη της ιδιοκτησίας.

1.2.3 Ομόλογα

Το ομόλογο (bond) είναι ένα χρηματοπιστωτικό μέσο το οποίο παρέχει σε έναν επενδυτή τη δυνατότητα να δανείζει χρήματα. Σε αντάλλαγμα για την κατοχή του ομολόγου, οι δανειοδότες εισπράττουν τόκο, που ονομάζεται επίσης τοκομερίδιο. Με απλά λόγια, τα χρήματα που μια εταιρεία λαμβάνει από τα ομόλογα που εκδίδονται θεωρούνται δάνειο. Επομένως, σε αντίθεση με τις μετοχές, οι κάτοχοι των ομολόγων δικαιούνται και τόκους αντί μόνο για αποπληρωμή του κεφαλαίου που επενδύθηκε. Οι δανειοδότες αυτοί, έχουν νομική προτεραιότητα έναντι των υπολοίπων μετόχων σε περίπτωση χρεοκοπίας και πρέπει να αποπληρωθούν πρώτοι εάν μια εταιρεία αναγκαστεί να πουλήσει τα περιουσιακά της στοιχεία.

1.2.4 Αμοιβαία κεφάλαια

Αμοιβαίο κεφάλαιο (mutual funds) είναι ένα χαρτοφυλάκιο αξιών με διασπορά, δηλαδή ένας χρηματοοικονομικός διαμεσολαβητής που επιτρέπει σε μια ομάδα επενδυτών να τοποθετήσουν τα χρήματά τους σύμφωνα με ένα προκαθορισμένο επενδυτικό σκοπό. Τα αμοιβαία κεφάλαια έχουν τις ρίζες τους στην ανάγκη του ανθρώπου να μετριάσει τους οικονομικούς κινδύνους που απορρέουν από τις διαρκείς

μεταβολές του οικονομικού περιβάλλοντος στο οποίο ζει και δραστηριοποιείται. Πρόκειται στην ουσία για μια εταιρεία που συγκεντρώνει χρήματα από πολλούς επενδυτές και επενδύει τα χρήματα σε τίτλους όπως μετοχές, ομόλογα και βραχυπρόθεσμο χρέος. Οι συνδυασμένες συμμετοχές του αμοιβαίου κεφαλαίου είναι γνωστές ως το χαρτοφυλάκιό του. Οι επενδυτές αγοράζουν μετοχές σε αμοιβαία κεφάλαια. Κάθε μετοχή αντιπροσωπεύει τη μερική ιδιοκτησία ενός επενδυτή στο αμοιβαίο κεφάλαιο και το εισόδημα που δημιουργεί.

1.2.5 Διαπραγματεύσιμα αμοιβαία κεφάλαια

Πρόκειται για ένα προϊόν το οποίο λειτουργεί όπως και ένα αμοιβαίο κεφάλαιο και ακολουθεί έναν δείκτη, ένα ομόλογο ή έναν συνδυασμό προϊόντων, αλλά σε αντίθεση με αυτό, η αγοραπωλησία του πραγματοποιείται κυρίως σε χρηματιστήρια όπως και με μία απλή μετοχή. Ένα διαπραγματεύσιμο αμοιβαίο κεφάλαιο (ETF) είναι ένας τύπος αμοιβαίου κεφαλαίου που κατέχει πολλαπλά υποκείμενα περιουσιακά στοιχεία, και όχι μόνο ένα όπως μια μετοχή. Για παράδειγμα, ένα κεφάλαιο ETF που παρακολουθεί τον δείκτη S&P 500 θα αποτελείται από κλάσματα μετοχών εταιρειών που περιλαμβάνονται στον συγκεκριμένο δείκτη. Τα ETF μπορούν επομένως να περιέχουν πολλούς τύπους επενδύσεων, συμπεριλαμβανομένων μετοχών, εμπορευμάτων, ομολόγων ή ενός μίγματος διαφορετικών τύπων επενδύσεων και για αυτό αποτελούν δημοφιλή επιλογή για τη στρατηγική διαφοροποίησης χαρτοφυλακίων.

1.2.6 Μερίσματα

Το μέρισμα (dividend) είναι η διανομή των κερδών μιας εταιρείας στους μετόχους της και καθορίζεται από το διοικητικό συμβούλιο της εταιρείας. Τα μερίσματα συχνά διανέμονται ανά τρίμηνο και μπορούν να πληρωθούν ως μετρητά ή με τη μορφή μετοχών ή άλλων περιουσιακών στοιχείων.

1.2.7 Τάξεις περιουσιακών στοιχείων

Μια κατηγορία περιουσιακών στοιχείων (asset class) είναι μια ομάδα επενδύσεων που παρουσιάζουν παρόμοια χαρακτηριστικά και υπόκεινται στους ίδιους νόμους και κανονισμούς. Επομένως, οι κατηγορίες περιουσιακών στοιχείων αποτελούνται από εργαλεία που συχνά συμπεριφέρονται παρόμοια μεταξύ τους στην αγορά. Οι επενδύσεις μπορούν να ταξινομηθούν στις εξής βασικές κατηγορίες :

- Μετρητά, δηλαδή οποιοδήποτε ανταλλακτικό μέσο το οποίο είναι αποδεκτό από όλα τα μέλη μιας κοινωνίας και χρησιμοποιείται για το αντάλλαγμα οποιουδήποτε αγαθού (συνήθως τα χαρτονομίσματα και νομίσματα). Ωστόσο, πρόκειται για περιουσιακό στοιχείο χαμηλής απόδοσης και σε περιόδους πληθωρισμού χάνει γρήγορα μέρος της αξίας του.
- Ομόλογα, τα οποία προσφέρουν χαμηλό αλλά σταθερό εισόδημα.
- Μετοχές, οι οποίες προσφέρουν υψηλές αποδόσεις έχουν όμως και υψηλό κίνδυνο.
- Ακίνητα, τα οποία αποτελούν μια ασφαλή επένδυση που προσφέρει σταθερές αποδόσεις και θεωρητικά καλές μακροχρόνιες προοπτικές μεγέθυνσης της αξίας.

1.2.8 Χαρτοφυλάκιο

Χαρτοφυλάκιο (portfolio) είναι το σύνολο των περιουσιακών στοιχείων που έχει ένας επενδυτής στην κατοχή του. Οι μετοχές και τα ομόλογα θεωρούνται γενικά τα βασικά δομικά στοιχεία ενός χαρτοφυλακίου αλλά μπορούν να χρησιμοποιηθούν και άλλα περιουσιακά στοιχεία συμπεριλαμβανομένων ακινήτων, χρυσού, πίνακες ζωγραφικής και άλλα συλλεκτικά αντικείμενα τέχνης. Η επιλογή των περιουσιακών αυτών στοιχείων και της ποσότητάς τους γίνεται με βάση τρία βασικά χαρακτηριστικά:

- Προσδοκώμενη απόδοση
- Ρίσκο
- Ρευστότητα

1.2.9 Όγκος συναλλαγών

Όγκος συναλλαγών (trading volume) είναι ο αριθμός των μετοχών που διακινούνται σε μια συγκεκριμένη χρονική περίοδο. Οι επενδυτές χρησιμοποιούν συχνά τον όγκο συναλλαγών για να επιβεβαιώσουν την ύπαρξη ή τη συνέχιση μιας τάσης ή την αντιστροφή της. Όταν ένα επενδυτικό αγαθό είναι πιο ενεργό, ο όγκος των συναλλαγών του είναι υψηλός και όταν είναι λιγότερο ενεργό, ο όγκος των συναλλαγών του είναι χαμηλός.

1.2.10 Μεταβλητότητα

Η μεταβλητότητα (volatility) είναι ο ρυθμός με τον οποίο η τιμή μιας μετοχής αυξάνεται ή μειώνεται σε μια συγκεκριμένη περίοδο. Υψηλότερη αστάθεια των τιμών των μετοχών συχνά σημαίνει και υψηλότερο κίνδυνο και βοηθά έναν επενδυτή να εκτιμήσει τις διακυμάνσεις που μπορεί να συμβούν στο μέλλον. Η τυπική απόκλιση είναι το στατιστικό μέτρο που χρησιμοποιείται συνήθως για την αναπαράσταση της μεταβλητότητας. Ο όρος αυτός επηρεάζεται συνήθως από πολιτικούς και οικονομικούς παράγοντες αλλά και από την επίδοση μιας εταιρείας.

1.2.11 Συνθήκες της αγοράς

Οι όροι «bull» και «bear» υποδηλώνουν τις τάσεις στις χρηματιστηριακές αγορές καθώς και ποια είναι η προοπτική των επενδυτών για την αγορά γενικά. Όταν η αγορά χαρακτηρίζεται ως «bullish», αυτό υποδηλώνει μια περίοδο ισχυρής οικονομικής ανάπτυξης όπου τα επίπεδα απασχόλησης είναι γενικά υψηλά και η εμπιστοσύνη των επενδυτών αυξάνεται. Κατά την περίοδο αυτή, οι επενδυτές συχνά πιστεύουν ότι η ανοδική τάση θα συνεχιστεί μακροπρόθεσμα κι έτσι τείνουν να αγοράζουν και να κρατάνε τις μετοχές επηρεάζοντας έτσι και τις ίδιες τις τιμές των μετοχών στο τέλος.

Από την άλλη, μια αγορά «bearish» χαρακτηρίζεται από μια διαρκή πτώση των τιμών των μετοχών, τουλάχιστον 20% ή περισσότερο από τις πρόσφατες υψηλές τιμές. Οι επενδυτές αντί να αγοράσουν, θέλουν να πουλήσουν στην αγορά, συχνά καταφεύγοντας στην ασφάλεια των μετρητών ή των επενδύσεων σταθερού εισοδήματος. Κατά τη διάρκεια αυτής της καθοδικής περιόδου, η ανεργία αυξάνεται

καθώς οι εταιρείες αρχίζουν να απολύουν εργαζομένους. Το πρώτο και πιο γνωστό παράδειγμα αυτού του είδους αγοράς ήταν η μεγάλη οικονομική ύφεση γνωστή και ως «The great depression» καθώς και επίσης γνωστή ήταν και η στεγαστική κρίση το 2007-2008.

1.2.12 Τεχνικοί δείκτες

Οι τεχνικοί δείκτες (technical indicators), εστιάζονται κυρίως σε ιστορικά δεδομένα συναλλαγών, όπως η τιμή και ο όγκος συναλλαγών, αντί των θεμελιωδών στοιχείων μιας επιχείρησης, όπως τα κέρδη και τα έσοδα. Συνήθως χρησιμοποιούνται για την ανάλυση βραχυπρόθεσμων κινήσεων τιμών αλλά οι επενδυτές μπορούν επίσης να χρησιμοποιήσουν τους τεχνικούς δείκτες και για να προσδιορίσουν σημεία εισόδου και εξόδου από την αγορά. Χωρίζονται σε 4 βασικές κατηγορίες:

- Trend indicators: Η τάση αναφέρεται στην κατεύθυνση της τιμής για μια χρονική περίοδο.
- Volume indicators: Ο όγκος αναφέρεται στον αριθμό των συναλλαγών που πραγματοποιήθηκαν για μια χρονική περίοδο.
- Volatility indicators: Μετριέται ο ρυθμός κίνησης των τιμών, ανεξάρτητα από την κατεύθυνση.
- Momentum indicators: Προσδιορίζεται η ταχύτητα της κίνησης των τιμών, συγκρίνοντας για παράδειγμα την τρέχουσα τιμή κλεισίματος με τα προηγούμενα κλεισίματα.

1.3 ΒΑΣΙΚΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ ΑΝΑΛΥΣΗΣ ΤΗΣ ΧΡΗΜΑΤΙΣΤΗΡΙΑΚΗΣ ΑΓΟΡΑΣ

Η θεμελιώδης και η τεχνική ανάλυση είναι οι δύο κύριες προσεγγίσεις που χρησιμοποιούνται για την ανάλυση των χρηματοοικονομικών αγορών με στόχο την επένδυση σε μετοχές με υψηλά κέρδη και χαμηλό ρίσκο καθώς και τη λήψη αποφάσεων (Park and Irwin 2007, Nguyen et al. 2015). Η πρώτη προσέγγιση μελετά τους οικονομικούς παράγοντες που μπορεί να επηρεάσουν τις κινήσεις της αγοράς και είναι η πλέον κατάλληλη για ένα μακροπρόθεσμο φάσμα προβλέψεων. Η δεύτερη προσέγγιση, από την άλλη, υποστηρίζει ότι η τιμή ενός αγαθού περιλαμβάνει ήδη όλα τα θεμελιώδη στοιχεία που την επηρεάζουν. Στην ουσία, η τεχνική ανάλυση βασίζεται στην ιστορική συμπεριφορά ενός χρηματοοικονομικού περιουσιακού στοιχείου, πιστεύοντας ότι η ιστορία τείνει να επαναλαμβάνεται [6].

Πιο αναλυτικά, η θεμελιώδης ανάλυση (fundamental analysis) περιλαμβάνει την αξιολόγηση μιας επιχείρησης και χρησιμοποιείται από τους επενδυτές για να καθοριστεί κατά πόσο η τρέχουσα τιμή της μετοχής μιας εταιρείας αντικατοπτρίζει τη μελλοντική αξία της. Στην ουσία εξετάζει την εγγενή αξία (intrinsic value) μιας εταιρείας, χρησιμοποιώντας οικονομικούς και χρηματοοικονομικούς παράγοντες και ειδησεογραφικά γεγονότα, δηλαδή την αξία μιας επένδυσης με βάση την οικονομική κατάσταση της εταιρείας και τις τρέχουσες συνθήκες της αγοράς και τις γενικότερες οικονομικές συνθήκες. Για παράδειγμα, υποθέτοντας ότι η οικονομία είναι ισχυρή και

το ΑΕΠ αυξάνεται, τότε μπορεί να αυξηθεί και η ζήτηση για αυτοκίνητα, με αποτέλεσμα να αυξηθεί και η ανάγκη για προϊόντα όπως ο χάλυβας και το αλουμίνιο που χρησιμοποιούνται συχνά στην κατασκευή αυτοκινήτων. Άρα η ανάλυση αυτή βασίζεται στη μελέτη της προσφοράς και της ζήτησης. Η αύξηση της προσφοράς ή η μείωση της ζήτησης τείνει να μειώσει την τιμή ενός εμπορεύματος. Αντίθετα, μια μείωση της προσφοράς ή μια αύξηση της ζήτησης θα αυξήσει την τιμή.

Από την άλλη, η τεχνική ανάλυση (technical analysis) επιχειρεί να προβλέψει την κατεύθυνση των τιμών αναλύοντας τα ιστορικά δεδομένα της αγοράς, όπως την τιμή και τον όγκο συναλλαγών. Για την ανάλυση αυτή, δε χρησιμοποιούνται εξωτερικά οικονομικά δεδομένα και σχετικές ειδήσεις καθώς γίνεται η παραδοχή ότι αυτοί οι εξωτερικοί παράγοντες αντανακλώνται στο ιστορικό μοτίβο τιμών των επενδυτικών αγαθών κι επομένως χρησιμοποιούνται με έμμεσο τρόπο. Αντιθέτως, οι αναλυτές χρησιμοποιούν τεχνικούς δείκτες, γραφήματα και διάφορα ακόμα τεχνικά εργαλεία με τα πιο βασικά να χωρίζονται στις εξής κατηγορίες:

- «Oscillators»,
- «Volume and momentum indicators»
- «Moving averages»
- «Support and resistance levels»
- «Price trends»
- «Chart patterns»

για τον εντοπισμό μοτίβων που υποδεικνύουν μελλοντικές τάσεις ή κατευθύνσεις τιμών. Πιστεύουν ότι η ιστορική δραστηριότητα των συναλλαγών και οι προηγούμενες αλλαγές των τιμών ενός επενδυτικού αγαθού μπορούν σε κάποιο βαθμό να καθορίσουν τη μελλοντική κατεύθυνση των τιμών. Συνήθως χρησιμοποιείται για την παραγωγή βραχυπρόθεσμων προβλέψεων και την αξιολόγηση της ισχύος ή αδυναμίας ενός επενδυτικού αγαθού σε σχέση με την ευρύτερη αγορά. Στην ουσία, βασίζεται στις εξής παραδοχές:

- Η τιμή ενός επενδυτικού αγαθού αντανακλά ήδη όλα όσα γεγονότα έχουν ή μπορούν να την επηρεάσουν.
- Η τιμή ακολουθεί μια τάση. Μια τέτοια υπόθεση είναι η βάση για πολλές στρατηγικές τεχνικών συναλλαγών.
- Η ιστορία τείνει να επαναληφθεί, δηλαδή τα μοτίβα κίνησης των τιμών είναι συχνά επαναλαμβανόμενα και μπορούν να χρησιμοποιηθούν για την ανάλυση των κινήσεων στην αγορά.

1.4 ΒΑΣΙΚΕΣ ΘΕΩΡΙΕΣ

Πολλές θεωρίες σχετικά με την πρόβλεψη της χρηματιστηριακής τιμής έχουν διαμορφωθεί με τα χρόνια και προσπαθούν είτε να εξηγήσουν τη φύση του χρηματιστηρίου είτε να εξηγήσουν κατά πόσο η αγορά μπορεί να «νικηθεί». Στην προσπάθεια τους να προβλέψουν τις αγορές, πολλοί επενδυτές υποθέτουν ότι τα

μελλοντικά συμβάντα βασίζονται εν μέρει σε σημερινά και παρελθοντικά γεγονότα. Ωστόσο, οι οικονομικές χρονοσειρές αποτελούν από τα πιο δύσκολα δεδομένα για πρόβλεψη λόγω της θορυβώδης τους φύσης.

Αυτό έχει οδηγήσει πολλούς οικονομολόγους να υιοθετήσουν την υπόθεση της αποτελεσματικής αγοράς γνωστή και ως «Efficient Market Hypothesis» που αναφέρθηκε για πρώτη φορά από τον Eugene Fama [7] [8]. Σύμφωνα με τη θεωρία αυτή, οι αλλαγές των τιμών είναι ανεξάρτητες από το παρελθόν καθιστώντας τις μεταβολές των τιμών απρόβλεπτες. Στην ουσία, η υπόθεση της αποτελεσματικής αγοράς βασίζεται στην ιδέα ενός «τυχαίου περιπάτου» (random walk), ένας όρος που χρησιμοποιείται για να χαρακτηρίσει ότι οι αλλαγές των τιμών που προκύπτουν αντιπροσωπεύουν τυχαίες αποκλίσεις από τις προηγούμενες τιμές. Η λογική είναι ότι εάν η ροή των πληροφοριών είναι ανεμπόδιστη, οι πληροφορίες αυτές αντικατροπτίζονται αμέσως στην αγοραία τιμή των μετοχών κι επομένως, η αυριανή αλλαγή της τιμής θα αντικατοπτρίζει μόνο τα αυριανά νέα και θα είναι ανεξάρτητη από τις αλλαγές των τιμών σήμερα. Με άλλα λόγια, οποιαδήποτε αλλαγή στην τιμή αντιπροσωπεύει την άμεση αντίδραση σε ένα στιγμιαίο γεγονός ειδήσεων ή σε νέες αλλαγές στα στοιχεία προσφοράς και ζήτησης.

Αν και έχει γίνει πολλή συζήτηση σχετικά με αυτήν τη θεωρία, είναι δύσκολο να αποδειχθεί ή να διαψευθεί. Με την πάροδο του χρόνου, τείνει η κυριαρχία της στο χώρο να γίνεται όλο και λιγότερο καθολική. Πολλοί οικονομολόγοι και στατιστικολόγοι αρχίζουν να πιστεύουν ότι οι χρηματοπιστωτικές αγορές είναι «κάπως» προβλέψιμες [9]. Η ύπαρξη τάσεων και προτύπων στις τιμές των μετοχών καθώς και οι υψηλές συσχετίσεις ανάμεσα σε θεμελιώδη γεγονότα και οικονομικούς δείκτες που επηρεάζουν την αγορά, δημιουργούν ορισμένα προβλέψιμα μοτίβα δίνοντας τη δυνατότητα στους επενδυτές να κερδίσουν υψηλότερα ποσοστά απόδοσης ενάντια όμως στην υπόθεση της αποτελεσματικής αγοράς. Όλο και περισσότερες μελέτες αμφισβητούν την εγκυρότητα της υπόθεσης αυτής και εισάγουν νέες, επιτυχημένες προσεγγίσεις συνδυάζοντας την τεχνική ανάλυση καθώς και μοτίβα διαγραμμάτων με μεθοδολογίες από την οικονομετρία, τη στατιστική και την τεχνητή νοημοσύνη με τέτοιο τρόπο ώστε να μπορούν να σχεδιαστούν κερδοφόρες στρατηγικές [10] [11] [12].

Πάνω σε αυτή τη λογική βασίστηκε και ο διαγωνισμός M6, σκοπός του οποίου ήταν η εμπειρική μελέτη αυτού του παραδόξου έτσι ώστε να μπορέσει να εξηγηθεί από τη μία η κακή απόδοση πολλών ενεργών κεφαλαίων και από την άλλη οι εξαιρετικές αποδόσεις που έχουν πετύχει γνωστοί επενδυτές όπως ο Warren Buffet, Peter Lynch και George Soros οι οποίες είναι αδύνατο να δικαιολογηθούν από απλή τύχη, δημιουργώντας έτσι αμφιβολίες για την εγκυρότητα της υπόθεσης της αποτελεσματικής αγοράς.

1.5 ΜΕΘΟΔΟΙ ΠΡΟΒΛΕΨΗΣ ΧΡΗΜΑΤΙΣΤΗΡΙΟΥ

Πολλές νέες τεχνολογίες και μέθοδοι έχουν αναπτυχθεί τα τελευταία χρόνια για την πρόβλεψη της συμπεριφοράς του χρηματιστηρίου [13] [14]. Οι πρόσφατες εξελίξεις στην ανάλυση και την πρόβλεψη μετοχών οδήγησαν στην δημιουργία 4 βασικών κατηγοριών: στατιστική, αναγνώριση προτύπων, μηχανική μάθηση (ML) και ανάλυση

συναισθήματος (sentiment analysis). Αυτές οι κατηγορίες εμπίπτουν ως επί το πλείστον στην ευρύτερη κατηγορία της τεχνικής ανάλυσης, ωστόσο, υπάρχουν ορισμένες τεχνικές μηχανικής μάθησης που συνδυάζουν επίσης τις ευρύτερες κατηγορίες τεχνικής ανάλυσης με προσεγγίσεις θεμελιώδους ανάλυσης για την πρόβλεψη των χρηματιστηριακών αγορών. Οι διάφορες τεχνικές προβλέψεων που έχουν εφαρμοστεί και μελετηθεί ευρέως, αναφέρονται τόσο στην πρόβλεψη των τιμών των μετοχών, των γενικών δεικτών του χρηματιστηρίου, στην πρόβλεψη μεταβλητότητας [15] καθώς και στη βέλτιστη επιλογή χαρτοφυλακίων.

Πριν την έλευση των τεχνικών μηχανικής μάθησης, γινόταν ευρεία χρήση στατιστικών μεθόδων για την πρόβλεψη των επενδυτικών αγαθών. Σύμφωνα με τους Zhong και Enke [5], οι στατιστικές προσεγγίσεις εμπίπτουν στην κατηγορία της ανάλυσης μίας μεταβλητής (univariate analysis) λόγω της χρήσης των οικονομικών χρονοσειρών ως μεταβλητές εισόδου, με τις πιο γνωστές να είναι οι αυτοπαλινδρομικές μέθοδοι κινητού μέσου όρου (Autoregressive Moving Average) γνωστότερες με τη συντομογραφία ARMA. Συγκεκριμένα, το μοντέλο ARIMA είναι μια ευρέως χρησιμοποιούμενη τεχνική για ανάλυση του χρηματιστηρίου [16]. Άλλοι τύποι στατιστικών προσεγγίσεων περιλαμβάνουν την γραμμική διακριτή ανάλυση (LDA), την τετραγωνική διακριτή ανάλυση (QDA), τη γραμμική παλινδρόμηση (LR) καθώς και μηχανές διανύσματος υποστήριξης (SVM) καθεμία από τις οποίες συνήθως περιλαμβάνει πολλαπλές μεταβλητές εισόδου (multivariate analysis). Για παράδειγμα, η στατιστική μέθοδος SVM έχει χρησιμοποιηθεί τόσο για την πρόβλεψη της κατεύθυνσης της κίνησης του χρηματιστηρίου σε συνδυασμό και με άλλες μεθόδους ταξινόμησης [17], όσο και ως μέθοδος αναγνώρισης προτύπων για να ανακαλύψει τη μεσοπρόθεσμη αστάθεια της τιμής μιας μετοχής [18]. Το μοντέλο εκθετικής εξομάλυνσης (ESM) είναι μια δημοφιλής τεχνική εξομάλυνσης που εφαρμόζεται σε δεδομένα χρονοσειρών για την ανάλυση τους [19]. Η κατασκευή μοντέλων ARIMA έχει χρησιμοποιηθεί για την ικανοποιητική πρόβλεψη των τιμών των μετοχών της Nokia και της τράπεζας Zenith [20].

Με την πάροδο του χρόνου, και την ανάπτυξη του τομέα τεχνητής νοημοσύνης, όλο και πιο διαδεδομένη γίνεται η χρήση τεχνικών μηχανικής μάθησης και νευρωνικών δικτύων για την πρόβλεψη του χρηματιστηρίου. Αυτό το φαινόμενο οφείλεται κυρίως στο γεγονός ότι με την πληθώρα διαθέσιμων δεδομένων είναι εύκολο να ληφθούν χρηματοοικονομικά δεδομένα από πολλαπλές πηγές συμπεριλαμβανομένων πληροφοριών τεχνικών δεικτών, ειδησεογραφικών γεγονότων και κοινωνικών δικτύων. Επιπλέον, η έρευνα των «ευφυών» αλγορίθμων έχει εμβαθυνθεί κι εξελιχθεί από τα απλά γραμμικά μοντέλα και τα ρηχά νευρωνικά δίκτυα μέχρι και σε αλγορίθμους ενισχυμένης μάθησης που χρησιμοποιούνται αποτελεσματικά στα πεδία αναγνώρισης εικόνων και ανάλυσης κειμένου. Πιστεύεται ότι αυτοί οι προηγμένοι αλγόριθμοι θα μπορούσαν μέχρι και να αποτυπώσουν τις δυναμικές αλλαγές της αγοράς και να λάβουν αυτόματες επενδυτικές αποφάσεις. Τέλος, η ταχεία ανάπτυξη των πληροφοριακών συστημάτων και η όλο και πιο ευρεία χρήση Μονάδων Επεξεργασίας Γραφικών (GPU) παρέχει μεγάλο αποθηκευτικό χώρο και υπολογιστική ισχύ για τη χρήση υψηλού όγκου οικονομικών δεδομένων.

Η μηχανική μάθηση λοιπόν, έχει μελετηθεί εκτενώς όσον αφορά την πρόβλεψη του χρηματιστηρίου [21] και συνήθως κατηγοριοποιείται σε δύο κύριες ομάδες, την επιβλεπόμενη και μη επιβλεπόμενη μάθηση. Στην πρώτη περίπτωση, τα διαθέσιμα δεδομένα αποτελούνται από παραδείγματα με ετικέτα, που σημαίνει ότι κάθε γραμμή δεδομένων περιέχει χαρακτηριστικά και μια σχετική ετικέτα (label). Στόχος της είναι η επιτυχημένη αντιστοίχηση των δεδομένων εισόδου με τα δεδομένα εξόδου και μετά την εκπαίδευση, το μοντέλο μαθαίνει από μία γραμμή χαρακτηριστικών να προβλέψει την αναμενόμενη έξοδο. Από την άλλη, η μη επιβλεπόμενη μάθηση χρησιμοποιεί αλγόριθμους μηχανικής μάθησης για την ανάλυση και την ομαδοποίηση συνόλων δεδομένων χωρίς ετικέτα. Αυτοί οι αλγόριθμοι ανακαλύπτουν κρυφά μοτίβα ή ομαδοποιήσεις δεδομένων. Γνωστές μέθοδοι που χρησιμοποιούνται συνήθως είναι τα δέντρα αποφάσεων, τυχαία δάση, λογιστική παλινδρόμηση, η ομαδοποίηση K-means και τα νευρωνικά δίκτυα. Τα τυχαία δάση έχουν αποδειχθεί έναν από τους καλύτερους αλγορίθμους για την μακροπρόθεσμη πρόβλεψη της κατεύθυνσης της τιμής μιας μετοχής σε σύγκριση και με άλλες μεθόδους ταξινόμησης [22]. Πέρα από την πρόβλεψη της τιμής ενός επενδυτικού αγαθού ή της κατεύθυνσής της, μία ακόμα δυνατότητα είναι η δημιουργία ενός νέου συστήματος υποστήριξης αποφάσεων για την ανάπτυξη αποτελεσματικών στρατηγικών συναλλαγών που μπορεί να προσφέρει κερδοφόρες αποδόσεις για τους επενδυτές [23]. Το δίκτυο αυτό παράγει ένα σύνολο συνεχών σημάτων συναλλαγών που κυμαίνονται στο πεδίο από 0 έως 1 αναλύοντας τη μη γραμμική σχέση μεταξύ μερικών γνωστών τεχνικών δεικτών. Μία ακόμα γνωστή τεχνική που έχει τραβήξει το ενδιαφέρον είναι μια νέα παραλλαγή των δέντρων ενίσχυσης κλίσης γνωστή και ως Extreme Gradient Boosting (XGBoost). Η μέθοδος αυτή χρησιμοποιήθηκε για την πρόβλεψη της κατεύθυνσης των μετοχών χρησιμοποιώντας τεχνικούς δείκτες ως χαρακτηριστικά και αποδείχθηκε ότι ξεπερνά άλλες τεχνικές στην απόδοση επιτυγχάνοντας ακρίβεια γύρω στο 87–99% για τη μακροπρόθεσμη πρόβλεψη των μετοχών της Apple και της Yahoo [24]. Τέλος, οι τοπολογίες των νευρωνικών δικτύων όπως για παράδειγμα τα δίκτυα μακράς βραχυπρόθεσμης μνήμης (LSTM) αρχίζουν να προσελκύουν την προσοχή στην πρόβλεψη χρονοσειρών. Η ανάπτυξη δικτύων LSTM για την πρόβλεψη των κινήσεων της αγοράς για τις μετοχές του S&P 500 από το 1992 έως το 2015 [25] παρουσιάζει υψηλότερη ακρίβεια πρόβλεψης και απόδοσης σε σύγκριση με άλλες μεθόδους όπως τα βαθιά δίκτυα (deep nets), τα τυχαία δάση (random forest) και τη λογιστική παλινδρόμηση (logistic regression). Η χρήση τεχνητών νευρωνικών δικτύων (ANN) για την πρόβλεψη της ημερήσιας κατεύθυνσης του δείκτη S&P 500 και η επιλογή των πιο σημαντικών χαρακτηριστικών που προτείνονται από το μοντέλο αποδίδει σημαντικά καλύτερα αποτελέσματα από τα παραδοσιακά μοντέλα τόσο στην πρόβλεψη του δείκτη, όσο και στη βελτίωση αποδόσεων των στρατηγικών συναλλαγών [26].

Η αναγνώριση προτύπων (pattern recognition) είναι συνώνυμη με τη μηχανική μάθηση, ωστόσο εφαρμόζεται με πολύ διαφορετικούς τρόπους, όσον αφορά την ανάλυση μετοχών. Η αναγνώριση προτύπων εστιάζει στον εντοπισμό μοτίβων και τάσεων στα δεδομένα και στις χρηματιστηριακές αγορές και συνήθως αναφέρεται στις ιστορικές τιμές Open-High-Low-Close (OHLC) οι οποίες χρησιμοποιώντουσαν ιστορικά και ως σήματα αγοράς και πώλησης. Για παράδειγμα, η ανάπτυξη ενός έξυπνου μοντέλου αναγνώρισης προτύπων για την υποστήριξη επενδυτικών αποφάσεων στο

χρηματιστήριο φαίνεται να ξεπερνά κλασικούς αλγορίθμους όπως τη θεωρία συνόλων (rough set theory), τους γενετικούς αλγόριθμους και το υβριδικό τους μοντέλο, παρέχοντας υψηλό επίπεδο κερδοφορίας [27].

Η ανάλυση συναισθήματος (sentiment analysis) είναι μια άλλη προσέγγιση που χρησιμοποιείται για την ανάλυση της συμπεριφοράς της αγοράς και είναι η διαδικασία πρόβλεψης των τάσεων των μετοχών μέσω αυτόματης ανάλυσης κειμένων και ειδησεογραφικών γεγονότων. Πιο συγκεκριμένα, πολυάριθμες συζητήσεις έχουν γίνει τα τελευταία χρόνια που επιχειρούν να εξηγήσουν τον ρόλο των δεδομένων της πλατφόρμας Twitter στις κινήσεις των τιμών των μετοχών. Λαμβάνοντας μια διαφορετική προσέγγιση από τα παραδοσιακά μοντέλα πρόβλεψης μετοχών που βασίζονται στην υπόθεση αποτελεσματικής αγοράς, οι Mittal και Goel [28] θέτουν την προϋπόθεση ότι υπάρχει άμεση συσχέτιση μεταξύ των συναισθημάτων του κοινού και της αγοράς. Χρησιμοποιώντας πάνω από 476 εκατομμύρια δημόσια διαθέσιμα tweets από τον Ιούνιο του 2009 έως τον Δεκέμβριο του 2009, κατηγοριοποιούν τα tweets σε «ήρεμα, χαρούμενα, σε εγρήγορση και ευγενικά» χαρτογραφώντας ένα προφίλ καταστάσεων διάθεσης. Από τα αποτελέσματα της μελέτης του αλγορίθμου επιτυγχάνεται ακρίβεια 75,56% στις κατευθύνσεις των μετοχών.

ΚΕΦΑΛΑΙΟ 2: ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ ΕΡΓΑΣΙΑΣ

Αυτό το κεφάλαιο ασχολείται με την επεξήγηση ορισμένων βασικών εργαλείων, αλγορίθμων και μεθόδων που χρησιμοποιούνται για πολλές διαφορετικές περιπτώσεις προβλέψεων με επίκεντρο κυρίως στον οικονομικό τομέα. Κάθε ένα από τα μοντέλα που θα επεξηγηθεί στο κεφάλαιο αυτό, θα χρησιμοποιηθεί αναλυτικά και στο πειραματικό στάδιο της μελέτης.

2.1 ΜΕΘΟΔΟΙ ΠΡΟΒΛΕΨΗΣ ΧΡΟΝΟΣΕΙΡΩΝ

2.1.1 Εισαγωγή

Οι μέθοδοι πρόβλεψης χρονοσειρών αφορούν τεχνικές που χρησιμοποιούν ιστορικά δεδομένα για την πρόβλεψη μελλοντικών τιμών για μια χρονική περίοδο στο μέλλον και την πραγματοποίηση παρατηρήσεων για την καθοδήγηση μελλοντικών στρατηγικών αποφάσεων. Κατά την πρόβλεψη δεδομένων χρονοσειρών, ο στόχος είναι να εκτιμηθεί πώς θα συνεχιστεί η ακολουθία των παρατηρήσεων στο μέλλον. Έχουν τη δυνατότητα να εντοπίσουν μοτίβα (patterns) χρονικής αποσύνθεσης όπως τάση, εποχιακότητα, κυκλικότητα καθώς και ασυνέχειες στα δεδομένα για εκπαίδευση, γεγονός που πολλές μέθοδοι μηχανικής μάθησης δεν μπορούν να κάνουν από προεπιλογή. Πολλοί τομείς συμπεριλαμβανομένου του μάρκετινγκ, των οικονομικών και των πωλήσεων, χρησιμοποιούν κάποια μορφή πρόβλεψης χρονοσειρών για την αξιολόγηση του πιθανού τεχνικού κόστους και της ζήτησης των καταναλωτών. Παραδείγματα δεδομένων χρονοσειρών περιλαμβάνουν:

- Καθημερινές τιμές μετοχών της IBM
- Μηνιαίες βροχοπτώσεις
- Τριμηνιαία αποτελέσματα πωλήσεων για την Amazon

Στις πιο συνηθισμένες μεθόδους συγκαταλέγονται η αφελής πρόβλεψη (Naïve), η γραμμική παλινδρόμηση, η εκθετική εξομάλυνση (Exponential Smoothing) και οι αυτοπαλινδρομικές μέθοδοι κινητού μέσου όρου (Autoregressive Integrated Moving Average) γνωστότερες και με τη συντομογραφία ARIMA. Παρ'όλα αυτά δεν αποφέρουν όλα τα μοντέλα τα ίδια αποτελέσματα για το ίδιο σύνολο δεδομένων κι επομένως είναι σημαντικό να καθοριστεί ποιο από αυτά λειτουργεί καλύτερα με βάση τις επιμέρους χρονοσειρές. Το μοντέλο που θα χρησιμοποιηθεί στην πρόβλεψη εξαρτάται από τους πόρους και τα διαθέσιμα δεδομένα, την ακρίβεια των υπόλοιπων μοντέλων και τον τρόπο με τον οποίο θα χρησιμοποιηθεί το μοντέλο πρόβλεψης. Τα μοντέλα ARIMA είναι στοχαστικά μαθηματικά μοντέλα τα οποία βοηθάνε στην ανάλυση και πρόβλεψη της εξέλιξης μεγεθών. Σε αντίθεση με τα ντετερμινιστικά μοντέλα (γραμμική παλινδρόμηση), η εφαρμογή των μοντέλων ARIMA βασίζεται στον υπολογισμό της πιθανότητας για την οποία η τιμή του μεγέθους βρίσκεται εντός

κάποιου διαστήματος. Καθώς όμως σκοπός της παρούσας διπλωματικής είναι η πρόβλεψη της σχετικής θέσης των ατομικών μετοχών (assets) δηλαδή της κατάταξης τους σε πέντε βασικές κλάσεις σύμφωνα με τις αναμενόμενες αποδόσεις τους και το μοντέλο ARIMA είναι μια μορφή παλινδρομικής ανάλυσης (regression analysis), απορρίπτεται ως μέθοδος για περαιτέρω έρευνα και εξετάζονται μόνο η αφελής μέθοδος καθώς και μέθοδοι εκθετικής εξομάλυνσης ως σημείο αναφοράς για σύγκριση και με πιο εξελιγμένες τεχνικές.

2.1.2 Αφελής μέθοδος

Πρόκειται για μια απλοϊκή (Naïve) τεχνική πρόβλεψης όπου η πρόγνωση της επόμενης περιόδου αντιστοιχεί στην πραγματική παρατήρηση της προηγούμενης [29] [30]. Στην ουσία η αφελής πρόβλεψη ισούται πάντα με την πιο πρόσφατα παρατηρούμενη τιμή χωρίς μεταβολές και περιγράφεται και από τον παρακάτω τύπο:

$$F_t = Y_{t-1}$$

όπου F_t : η πρόβλεψη τη χρονική περίοδο t

Y_{t-1}: η πραγματική τιμή την προηγούμενη περίοδο

Αν και πολύ απλή, λειτουργεί καλά συνήθως σε χρηματοοικονομικές χρονοσειρές και πολλές φορές χρησιμοποιείται ως σημείο αναφοράς για σύγκριση και με πιο εξελιγμένες μεθόδους για να εξεταστεί κατά πόσο αξίζουν περαιτέρω εξερεύνηση. Είναι σημαντικό να αναφερθεί ότι εάν η χρονοσειρά εμφανίσει απροσδόκητες αλλαγές, αν και η μέθοδος Naïve θα ανανεωθεί πολύ γρήγορα, επειδή κρατάει στη μνήμη της πάντα μόνο την τελευταία παρατήρηση δε θα φιλτράρει πιθανό θόρυβο από τα δεδομένα αλλά θα τον μεταφέρει στο μέλλον.

2.1.3 Απλή εκθετική εξομάλυνση

Μία από τις πιο αποτελεσματικές μεθόδους πρόβλεψης για χρονοσειρές, εμφανίστηκε για πρώτη φορά το 1956 από τον Robert Goodell Brown και στη συνέχεια επεκτάθηκε και σε άλλες παραλλαγές από τον Charles C. Holt [31] [32], [33]. Η μέθοδος της απλής εκθετικής εξομάλυνσης (Simple Exponential Smoothing - SES) χρησιμοποιείται πολύ συχνά και συνήθως είναι κατάλληλη για την πρόβλεψη δεδομένων χωρίς κάποιο μοτίβο εποχιακότητας ή τάσης καθώς όπως υποδηλώνει και η ονομασία της εξομαλύνει τα δεδομένα αφαιρώντας οποιοδήποτε θόρυβο. Περιγράφεται από τις παρακάτω εξισώσεις:

$$e_t = Y_t - F_t$$

$$S_t = S_{t-1} + a * e_t$$

$$F_{t+1} = S_t$$

όπου e_t : σφάλμα, δηλαδή απόκλιση της πραγματικής τιμής από την πρόβλεψη

S_t : το επίπεδο το οποίο συνήθως είναι ο απλός σταθμισμένος μέσος όρος της τρέχουσας παρατήρησης και του προηγούμενου εξομαλυμένου επιπέδου

F_t : η πρόβλεψη τη χρονική περίοδο t

Y_t : η πραγματική τιμή τη χρονική περίοδο t

a : ο συντελεστής εξομάλυνσης με τις τιμές του να κυμαίνονται σε [0,1]

Οι προβλέψεις υπολογίζονται χρησιμοποιώντας σταθμισμένους μέσους όρους όπου τα βάρη μειώνονται εκθετικά όσο πιο παλιές γίνονται οι παρατηρήσεις. Ο ρυθμός με τον οποίο μειώνονται τα βάρη αυτά ορίζεται από τον συντελεστή εξομάλυνσης. Όσο πιο κοντά η τιμή βρίσκεται στο 0, τόσο μεγαλύτερη βαρύτητα δίνεται στις πιο μακρινές παρατηρήσεις, ενώ για μεγαλύτερες τιμές κοντά στη μονάδα λαμβάνονται υπόψη περισσότερο οι πιο πρόσφατες παρατηρήσεις. Για α = 1, η μέθοδος μετατρέπεται στη Naïve που αναφέρθηκε και προηγουμένως.

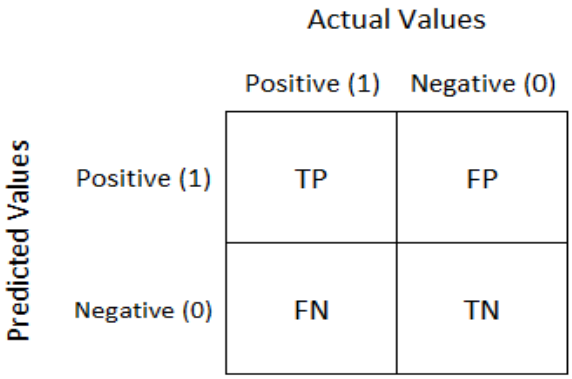
2.2 ΜΕΘΟΔΟΙ ΤΑΞΙΝΟΜΗΣΗΣ

2.2.1 Εισαγωγή

Η ταξινόμηση είναι μια μορφή επιβλεπόμενης μηχανικής μάθησης στην οποία το μοντέλο εκπαιδεύεται σε γνωστά δεδομένα ώστε να προβλέψει σε ποια κατηγορία ανήκει ένα αντικείμενο. Για παράδειγμα, μια κλινική μπορεί να χρησιμοποιήσει διαγνωστικά δεδομένα ενός ασθενή όπως το ύψος, το βάρος και η πίεση του αίματός του για να προβλέψει εάν ο ασθενής έχει διαβήτη. Υπάρχουν δύο βασικά είδη, η δυαδική και η πολυταξική ταξινόμηση. Στην πιο απλή μορφή, τη δυαδική ταξινόμηση, η πρόβλεψη της κλάσης γίνεται με προσδιορισμό της πιθανότητας για κάθε πιθανή κλάση ως μια τιμή που κυμαίνεται από 0 (αδύνατο) έως και 1 (βέβαιο). Η συνολική πιθανότητα για όλες τις κατηγορίες είναι πάντα 1 και μια τιμή κατωφλίου, συνήθως 0.5, χρησιμοποιείται για τον προσδιορισμό της προβλεπόμενης κλάσης. Είναι επίσης δυνατό να δημιουργηθούν μοντέλα ταξινόμησης πολλαπλών τάξεων (multiclass classification), στα οποία υπάρχουν περισσότερες από δύο πιθανές κατηγορίες. Δίνεται ένα σύνολο δειγμάτων εκπαίδευσης χωρισμένα σε K διακριτές κλάσεις και δημιουργείται το μοντέλο για να προβλεφθεί σε ποια από αυτές τις κατηγορίες ανήκουν κάποια προηγουμένως άγνωστα δεδομένα. Το μοντέλο μαθαίνει μοτίβα ειδικά για κάθε τάξη από το σύνολο δεδομένων εκπαίδευσης και χρησιμοποιεί αυτά τα μοτίβα για να προβλέψει την ταξινόμηση των μελλοντικών δεδομένων. Σε αυτήν την κατηγορία συγκαταλέγεται και το αντικείμενο της παρούσας διπλωματικής όπου σκοπός της είναι η κατάταξη των σχετικών θέσεων των μετοχών σε πέντε βασικές κλάσεις ως προς τις αναμενόμενες αποδόσεις τους.

Γενικότερα, ένα πρόβλημα ταξινόμησης αξιολογείται με τη χρήση ορισμένων μετρικών απόδοσης. Υπολογίζοντας απλά πόσες προβλέψεις είναι σωστές είναι μερικές φορές παραπλανητικό κι επομένως για τη λήψη πιο λεπτομερών πληροφοριών,

τα αποτελέσματα μπορούν να καταγραφούν σε μια δομή γνωστή και ως πίνακας σύγχυσης όπως φαίνεται και παρακάτω:



Εικόνα 1: Confusion matrix

Όπου:

- TP (true positive): Ο αριθμός των σωστών προβλέψεων ότι ένα παράδειγμα που είναι θετικό προσδιορίζεται σωστά ως θετικό.
- FP (false positive): Ο αριθμός των λάθος προβλέψεων ότι ένα παράδειγμα που είναι αρνητικό προσδιορίζεται λάθος ως θετικό.
- FN (false negative): Ο αριθμός των λάθος προβλέψεων ότι ένα παράδειγμα που είναι θετικό προσδιορίζεται λάθος ως αρνητικό.
- TN (true negative): Ο αριθμός των σωστών προβλέψεων ότι ένα παράδειγμα που είναι αρνητικό προσδιορίζεται σωστά ως αρνητικό.

Εκτός από τον πίνακα αυτό, χρησιμοποιούνται και άλλες μετρικές απόδοσης όπως:

- Accuracy: Η μετρική αυτή μετράει το σύνολο των προβλέψεων που βγήκαν σωστές και περιγράφεται από τον παρακάτω τύπο.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- Recall: Η μετρική αυτή μετράει από το σύνολο των παρατηρήσεων που ήταν 1 (positive), πόσες κατάφερε το μοντέλο να προβλέψει σωστά και περιγράφεται από τον παρακάτω τύπο.

$$\frac{TP}{TP + FN}$$

- Precision: Η μετρική αυτή μετράει από το σύνολο των προβλέψεων που βγήκαν ότι είναι για παράδειγμα 1 (positive), πόσες πραγματικά πρατηρήσεις ήταν 1, δηλαδή δείχνει την ορθότητα που επιτεύχθηκε στη θετική πρόβλεψη και περιγράφεται από τον παρακάτω τύπο.

$$\frac{TP}{TP + FP}$$

Στα πλαίσια της παρούσας διπλωματικής, προκειμένου να συμβαδίσει η έρευνα και με τις παραδοχές του διαγωνισμού, αντί για τις παραπάνω κλασικές μετρικές απόδοσης προτιμήθηκε η μετρική που χρησιμοποιήθηκε και στα πλαίσια του διαγωνισμού M6 γνωστή και ως RPS (Ranked Probability Score) η οποία θα εξηγηθεί στη συνέχεια.

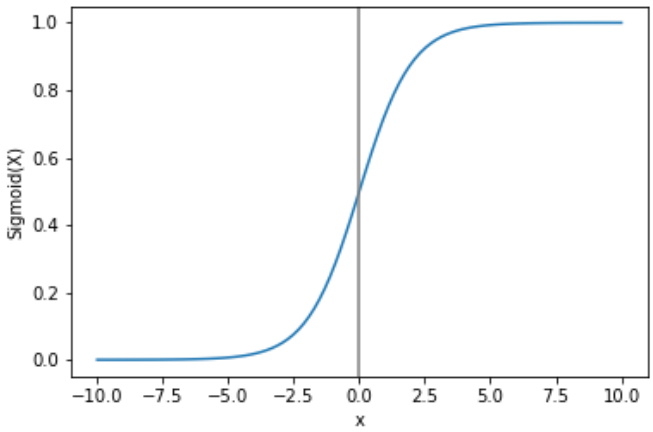
2.2.2 Λογιστική Παλινδρόμηση

Πρόκειται για μια μέθοδο επιβλεπόμενης μάθησης η οποία εστιάζεται στην εύρεση της συσχέτισης μεταξύ μιας εξαρτημένης και μίας ή περισσότερων ανεξάρτητων μεταβλητών υπολογίζοντας τις πιθανότητες με τη χρήση μιας σιγμοειδούς συνάρτησης. Σε αντίθεση με την ονομασία της χρησιμοποιείται σε προβλήματα ταξινόμησης (classification) όπου υπολογίζει την πιθανότητα ενός γεγονότος να πραγματοποιηθεί ή μιας ετικέτας να είναι true (1) ή false (0) δεδομένου ενός συνόλου ανεξάρτητων μεταβλητών [34], [35].

Η λογιστική παλινδρόμηση (Logistic Regression) είναι παρόμοια με τη γραμμική παλινδρόμηση η οποία προβλέπει μια πραγματική τιμή για έξοδο βασιζόμενη στο σταθμισμένο άθροισμα των μεταβλητών της εισόδου με τη διαφορά ότι στη συνέχεια περνάει το αποτέλεσμα αυτό από μία μη γραμμική συνάρτηση, τη σιγμοειδή ώστε να προκύψει το τελικό αποτέλεσμα σε δυαδική μορφή.

Η σιγμοειδής συνάρτηση δίνεται από την παρακάτω εξίσωση :

$$y = \frac{1}{1 + e^{-z}}$$

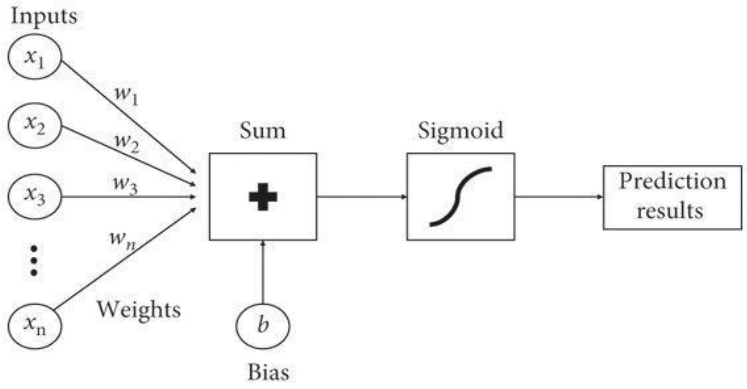


Εικόνα 2: Σιγμοειδής συνάρτηση

Όπου στην περίπτωση της λογιστικής παλινδρόμησης $z = logit(p(x))$

Η συνάρτηση logit περιγράφεται από την εξίσωση $logit(p(x)) = ln\left(\frac{p(x)}{1-p(x)}\right)$

Ενδεικτικά παρατίθεται παρακάτω και γραφικά ολοκληρωμένη η διαδικασία :



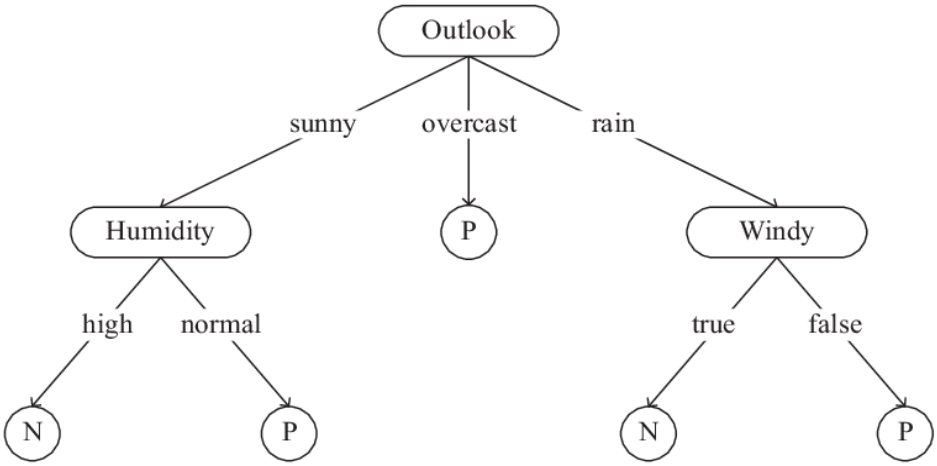
Εικόνα 3: Λειτουργία Logistic Regression

Υπάρχουν τρία είδη λογιστικής παλινδρόμησης :

- Binary logistic regression: Είναι η πιο συχνά χρησιμοποιημένη εκδοχή και διαθέτει δύο πιθανές εξόδους 0 ή 1 και χρησιμοποιείται σε προβλήματα πρόβλεψης για παράδειγμα αγοράς ή πώλησης μιας μετοχής.
- Multinomial logistic regression: Σε αυτήν την περίπτωση η εξαρτημένη μεταβλητή μπορεί να έχει πάνω από τρεις πιθανές κλάσεις χωρίς κάποια καθορισμένη σειρά προτεραιότητας. Η παραλλαγή αυτή είναι χρήσιμη όταν για παράδειγμα είναι επιθυμητή η πρόβλεψη του είδους της ταινίας που το κοινό επιθυμεί περισσότερο να παρακολουθήσει από μια λίστα επιλογών όπως περιπέτεια, φαντασίας, θρίλερ κλπ.
- Ordinal logistic regression: Σε αυτήν την περίπτωση η εξαρτημένη μεταβλητή μπορεί να έχει πάνω από τρεις πιθανές κλάσεις αλλά σε αυτήν την περίπτωση οι τιμές έχουν μία προκαθορισμένη σειρά όπως για παράδειγμα η βαθμολόγηση από 1 έως 10.

2.2.3 Δέντρα αποφάσεων

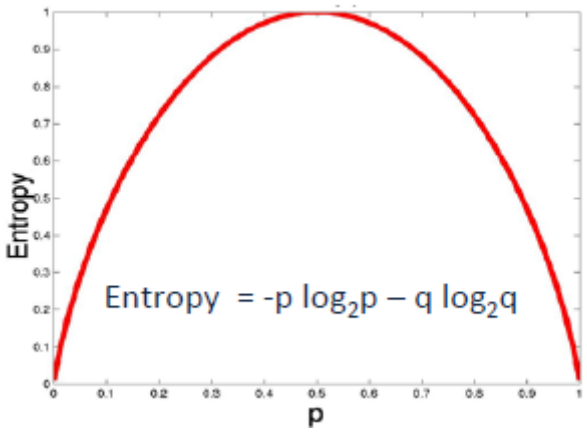
Τα δέντρα αποφάσεων (Decision trees) είναι μια τεχνική επιβλεπόμενης μάθησης η οποία χρησιμοποιείται για την κατηγοριοποίηση και την παραγωγή προβλέψεων ανάλογα με το πως απαντήθηκε ένα προηγούμενο σύνολο ερωτήσεων. Στην ουσία, χωρίζει το αρχικό σύνολο δεδομένων σε μικρότερα υποσύνολα έως ότου καταλήξει στην κατασκευή ενός δέντρου με κόμβους αποφάσεων και «φύλλα» τα οποία και αναπαριστούν την πρόβλεψη (είτε πρόκειται για πρόβλημα ταξινόμησης είτε για παλινδρόμηση). Η διαδικασία ξεκινάει από τη ρίζα του δέντρου όπου συγκρίνονται οι τιμές των χαρακτηριστικών και με βάση τη σύγκριση αυτή, επιλέγεται ο κατάλληλος κλάδος. Στη συνέχεια, ακολουθεί ένα σύνολο από κόμβους αποφάσεων που αναπαριστούν ο καθένας από μία ερώτηση δηλαδή ένα σημείο διαχωρισμού και οι κόμβοι που προκύπτουν από αυτά τα σημεία αντιστοιχούν στις πιθανές απαντήσεις.



Εικόνα 4: Λειτουργία Decision tree

Τα δέντρα αποφάσεων χρησιμοποιούν πολλαπλούς αλγορίθμους προκειμένου να αποφασίσουν σε πόσους κλάδους να χωρίσουν ένα σημείο διαχωρισμού με τον πιο βασικό να είναι ο ID3. Ο ID3 ακολουθεί μια top-down, άπληστη προσέγγιση και χρησιμοποιεί τις παρακάτω έννοιες για την κατασκευή του δέντρου:

- Entropy: Ξεκινώντας από την ρίζα του δέντρου, τα αρχικά χαρακτηριστικά διαχωρίζονται σε υποσύνολα τα οποία περιλαμβάνουν περιπτώσεις με παρόμοιες τιμές. Αυτή η ομογένεια του δείγματος υπολογίζεται μέσω της εντροπίας. Εάν το δείγμα είναι τελείως ομογενές παίρνει τιμή 0 και αντιστοιχεί σε «φύλλο» ενώ εάν το δείγμα είναι ισόποσα διαχωρισμένο παίρνει τιμή 1. Οποιοδήποτε δείγμα λαμβάνει τιμή πάνω από 0, χρειάζεται επιπρόσθετο διαχωρισμό.



Εικόνα 5: Entropy

Στο παραπάνω διάγραμμα παρατηρείται ότι η τιμή της εντροπίας προκύπτει 0 όταν η πιθανότητα είναι 0 ή 1 ενώ παίρνει τη μέγιστη τιμή της 1 όταν η πιθανότητα είναι 0.5 και ο υπολογισμός προκύπτει μέσω της εξίσωσης :

$$-0.5log_20.5 - 0.5log_25 = 1$$

Εκτός από τον τύπο αυτό υπολογισμού της εντροπίας για ένα χαρακτηριστικό υπάρχει και ακόμα ένας τύπος για τον υπολογισμό της εντροπίας για πολλαπλά χαρακτηριστικά.

$$Entropy(T, X) = \sum_{c \in X} P(c) * Entropy(c)$$

όπου T: η τωρινή χρονική στιγμή T

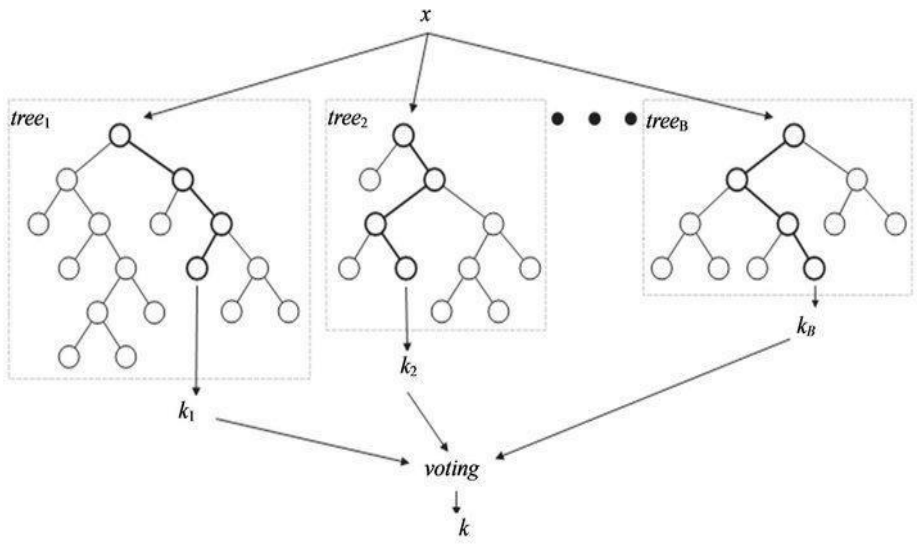
X: το επιλεγμένο χαρακτηριστικό

- Information gain: Πρόκειται για μια στατιστική μετρική που υπολογίζει πόσο καλά έχει πραγματοποιηθεί ένας διαχωρισμός δεδομένων σύμφωνα με ένα χαρακτηριστικό. Σκοπός είναι η κατασκευή ενός δέντρου αποφάσεων με το υψηλότερο IG και τη μικρότερη τιμή εντροπίας. Ο υπολογισμός της μετρικής δίνεται και από τον παρακάτω τύπο:

$$Information\ Gain(T, X) = Entropy(T) - Entropy(T, X)$$

2.2.4 Τυχαία Δάση

Τα τυχαία δάση (Random Forest) χρησιμοποιούν μια ensemble τεχνική, δηλαδή πολλαπλά μοντέλα για την παραγωγή προβλέψεων [36] [37]. Πιο συγκεκριμένα, λειτουργούν σύμφωνα με τη λογική της συνάθροισης δειγμάτων όπου για κάθε μοντέλο δημιουργείται ένα διαφορετικό υποσύνολο δεδομένων και εκπαιδεύονται ανεξάρτητα παράγοντας αποτελέσματα. Από αυτά, η τελική πρόβλεψη προκύπτει συνδυάζοντας όλες τις διαθέσιμες επιλογές και επιλέγοντας την πλειοψηφία (majority voting). Η μεθοδολογία αυτή χρησιμοποιείται για προβλήματα παλινδρόμησης (regression problems) και ταξινόμησης (classification problems) αντίστοιχα.



Εικόνα 6: Λειτουργία Random Forest

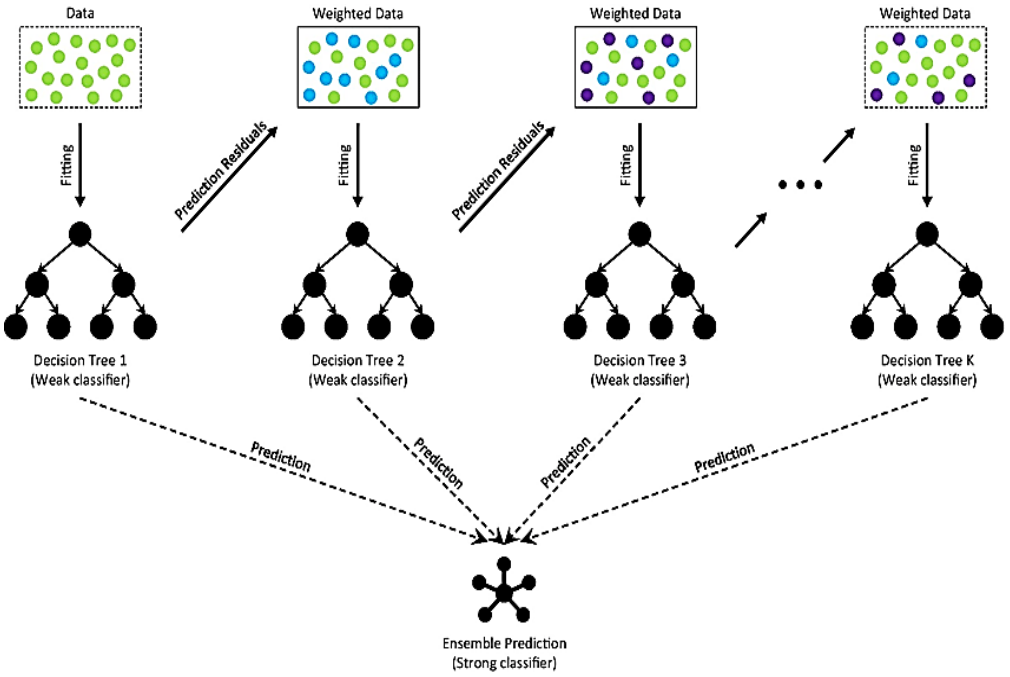
Πρόκειται για έναν μεγάλο αριθμό ατομικών δέντρων αποφάσεων τα οποία λειτουργούν ως ένα σύνολο. Η χαμηλή συσχέτιση ανάμεσα στα μοντέλα προστατεύει τα δέντρα μεταξύ τους από τα ατομικά τους σφάλματα. Ως αποτέλεσμα, η μέθοδος αυτή δεν εμφανίζει πρόβλημα υπερπροσαρμογής (overfitting). Επιπλέον, η μέθοδος διαθέτει ενσωματωμένη την εντολή feature_importances η οποία βοηθάει στην επιλογή των χαρακτηριστικών που συνεισφέρουν περισσότερο.

Τέλος, αποτελείται από ένα πλήθος υπερπαραμέτρων οι οποίες είτε ενισχύουν την απόδοση του μοντέλου είτε το καθιστούν πιο γρήγορο. Οι πιο βασικοί υπερπαράμετροι που αξίζει να αναφερθούν καθώς χρησιμοποιήθηκαν και στα πλαίσια της παρούσας μελέτης συνοψίζονται στα εξής:

- n_estimators: Ο αριθμός των δέντρων που χτίζει ο αλγόριθμος.
- max_features: Ο αριθμός των χαρακτηριστικών που λαμβάνονται υπόψη για διαχωρισμό σε κάθε κόμβο φύλλου.
- bootstrap: Εάν θα χρησιμοποιηθεί το σύνολο των δεδομένων για το φτιάξιμο κάθε δέντρου ή θα δημιουργηθούν δείγματα αυτού.

2.2.5 Δέντρα Ενίσχυσης Κλίσης

Τα δέντρα ενίσχυσης κλίσης (Gradient Boosting) χρησιμοποιούν και αυτά ένα σύνολο δέντρων αποφάσεων λειτουργώντας όμως σύμφωνα με την αρχή ενίσχυσης (boosting) [38]. Δημιουργείται ένα σύνολο από μεμονωμένα «αδύναμα» μοντέλα με χαμηλή απόδοση τα οποία στη συνέχεια αθροίζουν τις προβλέψεις τους, ενισχύοντας έτσι το μοντέλο και παρέχοντας βελτιωμένη ακρίβεια. Η λογική πίσω από τον αλγόριθμο είναι να αξιοποιήσει το μοτίβο στα σφάλματα και να ενισχύσει ένα αδύναμο μοντέλο έως ότου τα σφάλματα να κατανέμονται τυχαία, δηλαδή να ελαχιστοποιήσει τη συνάρτηση κόστους (loss function). Σε οποιαδήποτε στιγμή t, τα αποτελέσματα του μοντέλου ζυγίζονται με βάση τα αποτελέσματα της προηγούμενης στιγμής t-1. Στα αποτελέσματα που προβλέπονται σωστά δίνεται μικρότερο βάρος ενώ σε εκείνα που δεν έχουν ταξινομηθεί με σωστό τρόπο σταθμίζονται υψηλότερα. Αυτή η τεχνική ακολουθείται για πρόβλημα ταξινόμησης ενώ παρόμοια τεχνική χρησιμοποιείται για παλινδρόμηση.



Εικόνα 7: Λειτουργία Gradient Boosting

Μέσα από πολλούς διαγωνισμούς ML και κυρίως μέσω της πλατφόρμας Kaggle φαίνεται ότι συχνά ο αλγόριθμος αυτός σε τυπικά προβλήματα (εκτός από εικόνα, ήχο και πολύ αραιά δεδομένα) είναι ο πιο αποτελεσματικός. Επιπλέον, τα τελευταία χρόνια έχουν δημιουργηθεί πολλές πετυχημένες παραλλαγές που βασίζονται στην ίδια αρχή ενίσχυσης (boosting) όπως το lightGBM το 2017 από τη Microsoft και το XGBoost το 2014 από τον Tianqi Chen [39].

Όπως και στην περίπτωση με το random forest, ο αλγόριθμος αυτός αποτελείται από ένα πλήθος υπερπαραμέτρων που χωρίζονται σε 3 βασικές κατηγορίες:

1. Tree-Specific Parameters: Παράμετροι που αφορούν το κάθε μεμονωμένο δέντρο του μοντέλου.
2. Boosting Parameters: Παράμετροι που βοηθούν στην ενίσχυση του μοντέλου.
3. Miscellaneous Parameters: Γενικοί παράμετροι για την λειτουργία του μοντέλου.

Από τις παραπάνω, οι πιο σημαντικοί που αξίζει να αναφερθούν καθώς χρησιμοποιήθηκαν και στα πλαίσια της παρούσας μελέτης συνοψίζονται στα εξής:

- n_estimators: Ο αριθμός των δέντρων που χτίζει ο αλγόριθμος .
- max_depth: Το μέγιστο βάθος ενός δέντρου.
- learning_rate: Ο ρυθμός εκμάθησης του μοντέλου.

Και οι 3 παράμετροι είναι χρήσιμοι καθώς η ρύθμισή τους με μια μέθοδο cross-validation βοηθάει στην αντιμετώπιση του προβλήματος υπερπροσαρμογής (overfitting).

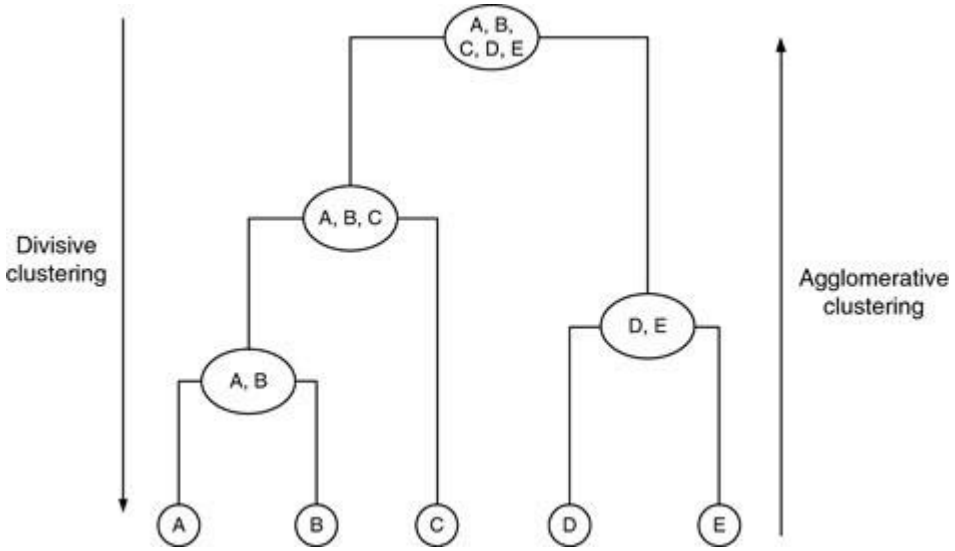
2.3 ΜΕΘΟΔΟΙ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ ΚΑΙ ΜΕΙΩΣΗΣ ΔΙΑΣΤΑΣΕΩΝ

2.3.1 Εισαγωγή

Η συσταδοποίηση πρόκειται για μία μέθοδο μη επιβλεπόμενης μάθησης, στην οποία οι παρατηρήσεις ομαδοποιούνται σε «συμπλέγματα» (clusters) με βάση τις ομοιότητες στις τιμές των δεδομένων ή ορισμένων χαρακτηριστικών τους. Όταν κάθε παράδειγμα ορίζεται από ένα ή δύο χαρακτηριστικά, είναι εύκολο να μετρηθεί η ομοιότητα, καθώς όμως ο αριθμός των χαρακτηριστικών αυξάνεται, η δημιουργία ενός μέτρου ομοιότητας γίνεται πιο περίπλοκη [40]. Η τεχνική αυτή χρησιμοποιείται όχι μόνο στο χρηματιστήριο αλλά εφαρμόζεται και σε πολλούς ακόμα κλάδους όπως για παράδειγμα την ανάλυση κοινωνικών δικτύων, την τμηματοποίηση της αγοράς (market segmentation) και την ομαδοποίηση αποτελεσμάτων αναζήτησης (search result grouping).

Υπάρχουν 4 βασικές προσεγγίσεις [41] από τις οποίες οι 2 είναι οι πιο σημαντικές στην παρούσα διπλωματική:

- Centroid-based clustering: Οργανώνει τα δεδομένα σε μη ιεραρχικά συμπλέγματα και λειτουργεί σύμφωνα με την εγγύτητα των «σημείων» δεδομένων με την επιλεγμένη κεντρική τιμή (centroid). Τα σύνολα δεδομένων χωρίζονται σε έναν δεδομένο αριθμό συστάδων και ένα διάνυσμα τιμών αναφέρεται σε κάθε σύμπλεγμα. Το k-means είναι ο πιο ευρέως χρησιμοποιούμενος αλγόριθμος ομαδοποίησης σε αυτήν την κατηγορία. Οι αλγόριθμοι αυτοί είναι αποτελεσματικοί αλλά ευαίσθητοι σε αρχικές συνθήκες και ακραίες τιμές.
- Hierarchical Clustering: Η ιεραρχική ομαδοποίηση δημιουργεί ένα δέντρο συστάδων και βασίζεται στην αρχή ότι κάθε «σημείο» συνδέεται με τους γείτονές του ανάλογα με την απόσταση εγγύτητάς τους. Οι αλγόριθμοι ιεραρχικής ομαδοποίησης χωρίζονται σε 2 κατηγορίες: από πάνω προς τα κάτω (top-down) ή από κάτω προς τα πάνω (bottom-up). Οι αλγόριθμοι από κάτω προς τα πάνω αντιμετωπίζουν κάθε σημείο δεδομένων ως ένα ενιαίο σύμπλεγμα στην αρχή και στη συνέχεια συγχωνεύοντας διαδοχικά ζεύγη συμπλεγμάτων καταλήγουν σε ένα ενιαίο σύμπλεγμα που περιέχει όλα τα σημεία δεδομένων. Αυτή η ομαδοποίηση είναι γνωστή και ως hierarchical agglomerative clustering ή HAC. Αυτή η ιεραρχία των συστάδων αναπαρίσταται με ένα δεντρογράφημα. Η ρίζα του δέντρου είναι η μοναδική συστάδα που συγκεντρώνει όλα τα δείγματα, τα φύλλα είναι οι συστάδες με ένα μόνο δείγμα.



Εικόνα 8: Παράδειγμα ιεραρχικής συσταδοποίησης

Ένα πλεονέκτημα της μεθόδου είναι ότι μπορεί να επιλεγεί οποιοσδήποτε αριθμός συστάδων κόβοντας το δέντρο στο σωστό επίπεδο. Παρόλα αυτά, διαθέτει κόστος χαμηλότερης απόδοσης, καθώς έχει χρονική πολυπλοκότητα $O(n^3)$, σε αντίθεση με τη γραμμική πολυπλοκότητα του K-Means. Καθώς λοιπόν, ο αλγόριθμος συσταδοποίησης K-Means είναι εύκολος να κατανοηθεί και να εφαρμοστεί σε κώδικα καθώς και επειδή έχει το πλεονέκτημα ότι είναι αρκετά γρήγορος με $O(n)$ πολυπλοκότητα, επιλέγεται ως την επικρατέστερη μέθοδο μελέτης στην παρούσα διπλωματική.

2.3.2 Ανάλυση σε κύριες συνιστώσες

Η ανάλυση σε κύριες συνιστώσες (PCA analysis) είναι μια τεχνική μείωσης της διαστατικότητας που χρησιμοποιείται για την εξαγωγή πληροφορίας από ένα χώρο υψηλών διαστάσεων και προβολή σε ένα χώρο χαμηλότερων διαστάσεων [42]. Προσπαθεί να κρατήσει τα πιο σημαντικά σημεία, διατηρώντας έτσι τη μέγιστη διακύμανση δεδομένων. Ένα σημαντικό πράγμα που πρέπει να σημειωθεί σχετικά με το PCA είναι ότι πρόκειται για μια μη επιβλεπόμενη τεχνική μείωσης διαστάσεων που προσδιορίζει τον μικρότερο αριθμό χαρακτηριστικών που απαιτούνται για να γίνει μια ακριβής πρόβλεψη καθώς πολλές φορές δεν είναι όλα τα χαρακτηριστικά χρήσιμα για να πραγματοποιηθεί μια πρόβλεψη.

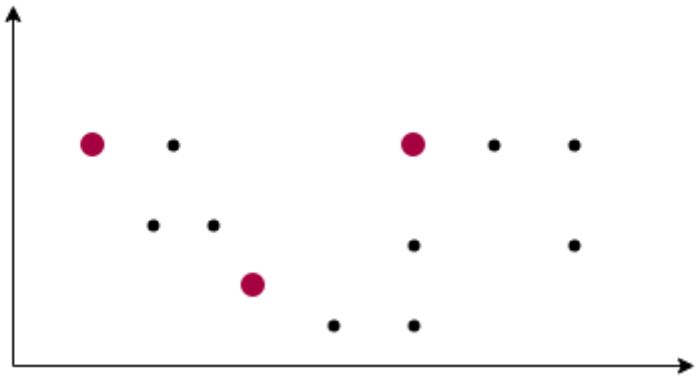
Η ανάλυση σε κύριες συνιστώσες μπορεί να βοηθήσει:

- Στο πρόβλημα υπερπροσαρμογής (overfitting) ενός μοντέλου με δεδομένα με θόρυβο.
- Στην επιτάχυνση της εκπαίδευσης ενός αλγορίθμου μηχανικής μάθησης.
- Στην ευκολότερη οπτικοποίηση των δεδομένων κυρίως όταν μειώνεται σε 2 διαστάσεις.

2.3.3 Συσταδοποίηση K-means

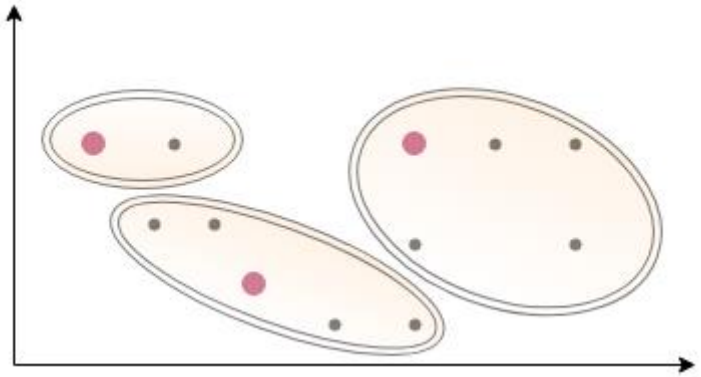
Πρόκειται για έναν αλγόριθμο μη επιβλεπόμενης μάθησης που αποσκοπεί στο διαχωρισμό n παρατηρήσεων σε k συστάδες (Clusters) όπου κάθε παρατήρηση θα ανήκει στη συστάδα με το πιο κοντινό μέσο/κέντρο. Η διαδικασία συνοψίζεται στα εξής βήματα:

1. Ορισμός του αριθμού k των συστάδων προκειμένου να ομαδοποιηθούν τα δεδομένα.
2. Αρχικοποίηση των κέντρων (centroids): Εφόσον η τοποθεσία των κέντρων είναι ακόμα άγνωστη, ανάλογα με την επιλογή του αριθμού k των ομάδων, γίνεται μια τυχαία επιλογή των σημείων αυτών και ορίζονται ως τα κέντρα για κάθε ομάδα. Στο παρακάτω ενδεικτικό παράδειγμα, επιλέχθηκε ο διαχωρισμός των δεδομένων σε τρεις ομάδες (k = 3) εξού και ορίζονται στο διάγραμμα 3 σημεία με κόκκινο χρώμα τυχαία.



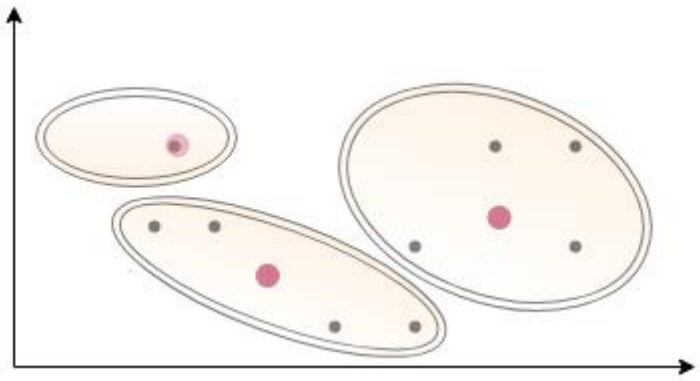
Εικόνα 9: Αρχικοποίηση κέντρων

3. Κάθε σημείο-δεδομένο (μαύρη κουκίδα) αντιστοιχίζεται στο πλησιέστερο κέντρο του.



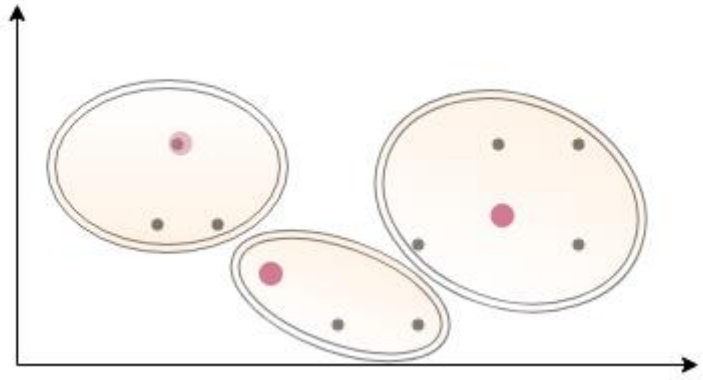
Εικόνα 10: Αντιστοίχιση σημείων στο πλησιέστερο κέντρο

4. Κάθε κέντρο μετακινείται στη μέση της ομάδας, δηλαδή στο κέντρο των σημείων που του αναλογούν σύμφωνα με τη μέση απόσταση μεταξύ τους.



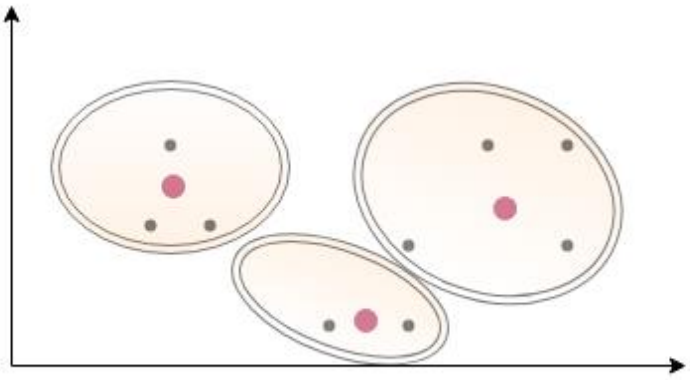
Εικόνα 11: Μετακίνηση κέντρων

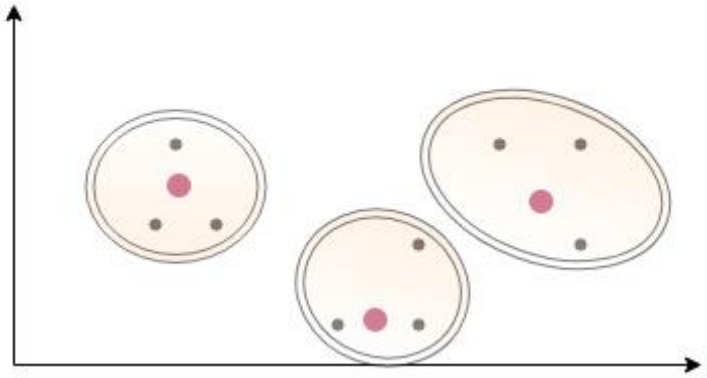
5. Μετά τη μετακίνηση του κέντρου, τα σημεία μπορεί τώρα να είναι πιο κοντά σε ένα διαφορετικό κέντρο. Επομένως τα σημεία-δεδομένα επανατοποθετούνται σε συμπλέγματα με βάση το νέο πλησιέστερο κέντρο.



Εικόνα 12: Επανατοποθέτηση σημείων σε συστάδες

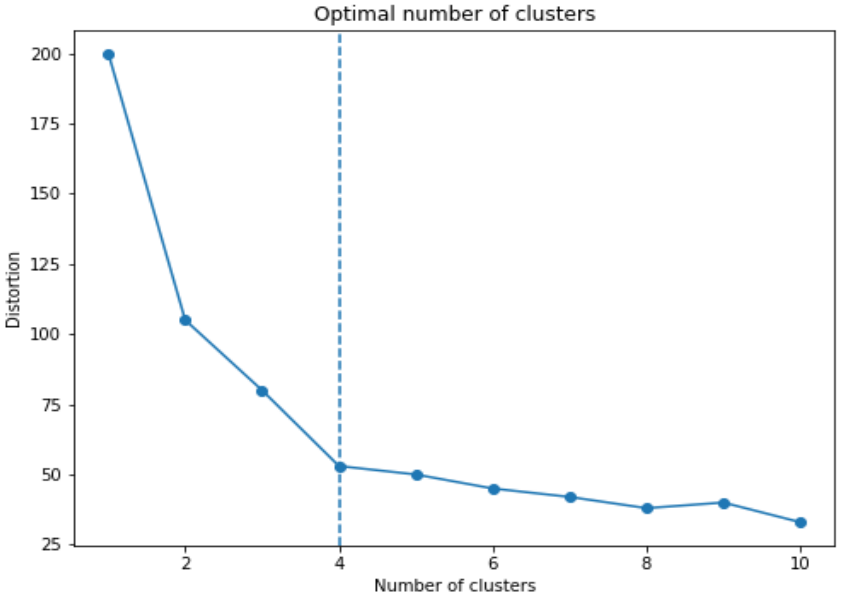
6. Τα βήματα μετακίνησης του κέντρου και της ανακατανομής συμπλέγματος επαναλαμβάνονται έως ότου τα συμπλέγματα γίνουν σταθερά ή μέχρι να επιτευχθεί ένας προκαθορισμένος μέγιστος αριθμός επαναλήψεων.





Εικόνα 13: Επανάληψη βημάτων 4 και 5

Για την επιλογή του βέλτιστου αριθμού ομάδων (k), μία από τις πιο γνωστές μεθόδους είναι η «Elbow curve» η οποία χρησιμοποιείται για να επισημανθεί η σχέση ανάμεσα στον αριθμό των επιλεγμένων συστάδων σε κάθε περίπτωση και το άθροισμα των τετραγωνικών σφαλμάτων τους (sum of squared errors-SSE). Πρόκειται για μια εμπειρική μέθοδο που βοηθάει στην βέλτιστη επιλογή της τιμής k (αριθμού συστάδων) όταν η προσθήκη μίας ακόμα ομάδας δεν βελτιώνει τα αποτελέσματα της μοντελοποίησης.



Εικόνα 14: Παράδειγμα χρήσης Elbow method

Στο παράδειγμα της εικόνας παρατηρείται ότι ο βέλτιστος αριθμός ομάδων για τα δεδομένα είναι k=4 καθώς στη συνέχεια το σφάλμα μειώνεται με πολύ πιο αργό ρυθμό.

ΚΕΦΑΛΑΙΟ 3: ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΤΑΞΗ

Αυτό το κεφάλαιο ασχολείται με την επεξήγηση και ανάλυση των δεδομένων και μεθόδων που χρησιμοποιήθηκαν, είτε πρόκειται για στατιστικές είτε για τεχνικές μηχανικής μάθησης. Τα δεδομένα που συλλέχθηκαν, τα απαραίτητα εργαλεία και χαρακτηριστικά, η γλώσσα προγραμματισμού και οι βιβλιοθήκες που χρησιμοποιήθηκαν θα εξηγηθούν περαιτέρω σε αυτό το κεφάλαιο καθώς και η μεθοδολογική προσέγγιση που επιλέχθηκε για την ανάπτυξη και αξιολόγηση των αποτελεσμάτων.

3.1 ΠΑΡΟΥΣΙΑΣΗ ΔΕΔΟΜΕΝΩΝ

3.1.1 Εισαγωγή

Αυτή η διπλωματική έχει εμπνευστεί από τον διαγωνισμό χρηματοοικονομικών προβλέψεων M6, τον πιο πρόσφατο από τους γνωστούς διαγωνισμούς Μακριδάκη (M – competitions) οι οποίοι έχουν ως σκοπό να ρίξουν φως στις αιτίες και συνέπειες του παραδόξου EMH (Efficient Market Hypothesis) γνωστό και ως υπόθεση αποτελεσματικής αγοράς. Συγκεκριμένα, ο διαγωνισμός προβλέψεων M6 επικεντρώνεται σε δύο βασικές διστάσεις, την ακρίβεια των προβλέψεων και την απόδοση των επενδύσεων καθώς και στη σημαντικότητα της πρόβλεψης όταν αυτή χρησιμοποιείται για να υποστηρίξει επενδυτικές αποφάσεις. Η διπλωματική αυτή ασχολείται με τη πρώτη διάσταση του διαγωνισμού στην οποία κρίνεται η ικανότητα πρόβλεψης με ακρίβεια της σχετικής θέσης των ατομικών μετοχών (assets) δηλαδή της κατάταξης τους σε πέντε βασικές κλάσεις ως προς τις αναμενόμενες αποδόσεις τους. Για το λόγο αυτό, όπως θα εξηγηθεί και παρακάτω αναλυτικά, ακολουθώντας τη δομή του διαγωνισμού, επιλέγονται τα ιστορικά δεδομένα πενήντα μετοχών SP500 για τα οποία θα διεξαχθούν προβλέψεις ως προς την κατάταξή τους.

Η κατάταξη προκύπτει από τις ποσοστιαίες αποδόσεις των προσαρμοσμένων τιμών των μετοχών σε πέντε κλάσεις, με τις πιο κερδοφόρες να ανήκουν στην τάξη 5 (υψηλότερες προβλεπόμενες ποσοστιαίες αποδόσεις) και αυτές με τη χειρότερη απόδοση στην τάξη 1. Η κατάταξη αφορά τη συγκριτική απόδοση του συνόλου των μετοχών μεταξύ τους από το διαθέσιμο σύμπαν επενδυτικών αγαθών του διαγωνισμού. Δεδομένου ότι στα πλαίσια της παρούσας μελέτης θα χρησιμοποιηθεί ένα σύνολο από 50 μετοχές του SP500, 10 από αυτές θα ανήκουν πάντα στην κλάση 1, 10 στην κλάση 2, 10 στην κλάση 3 κοκ.

3.1.2 Κατανόηση δεδομένων

Σε κάθε μελέτη πρόβλεψης, η κατάλληλη επιλογή και οργάνωση των δεδομένων καθώς και η συλλογή τους από αξιόπιστες πηγές, πολλές φορές είναι το κλειδί που θα κάνει τη διαφορά στα αποτελέσματα της πρόβλεψης, όσο καλή και να είναι η μέθοδος που επιλέξαμε. Για το λόγο αυτό, προκειμένου να υπάρχει επαρκής αριθμός παρατηρήσεων, συλλέχθηκαν τα ημερήσια δεδομένα του έτους 2021 καθώς και οι τρεις πρώτοι μήνες του έτους 2022 τα οποία χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων. Το χρονικό αυτό εύρος ορίστηκε έχοντας ως κριτήριο την ύπαρξη δεδομένων για όλες τις διαθέσιμες εργάσιμες ημέρες καθώς και το ότι αυτά τα στοιχεία θα είναι αντιπροσωπευτικά της καθημερινής κίνησης της αγοράς. Δεδομένου ότι κατά τη διάρκεια του έτους 2020 το ξέσπασμα της πανδημίας προκάλεσε μια ελεύθερη πτώση στις τιμές των μετοχών επηρεάζοντας το χρηματιστήριο και προκαλώντας μια αβεβαιότητα και μεταβλητότητα καθώς και από τέλη Φεβρουαρίου του έτους 2022, ο πόλεμος της Ρωσίας με την Ουκρανία έχει προκαλέσει αστάθεια στις οικονομικές αγορές κυρίως στον τομέα της ενέργειας, τα κριτήρια αυτά λήφθηκαν υπόψη για την επιλογή του χρονικού εύρους.

Ως διάστημα πρόβλεψης ορίστηκαν 3 περίοδοι (μήνες), με το διάστημα στο οποίο πραγματοποιούνται πάντα οι προβλέψεις να κυμαίνεται από 4/01/2022 έως και 30/03/2022 το οποίο ήταν κοινό για όλες τις μεθόδους έτσι ώστε να εξασφαλιστεί ότι θα είναι ομοιόμορφη η σύγκριση των αποτελεσμάτων. Σε αυτό το χρονικό διάστημα για την παραγωγή των προβλέψεων εφαρμόστηκε σε όλες τις μεθόδους κυλιόμενη πρόβλεψη ανά ημέρα, ενώ και στις τεχνικές μηχανικής μάθησης υλοποιήθηκε και κυλιόμενη πρόβλεψη ανά μήνα, μέθοδος που προτιμάται σε συνεχώς μεταβαλλόμενες συνθήκες όπου υπάρχει η ανάγκη για γρήγορη προσαρμογή και είναι και σύμφωνη με τις παραδοχές παραγωγής προβλέψεων του διαγωνισμού. Για παράδειγμα, στις 4/01/2022 έχει γίνει συλλογή δεδομένων μέχρι και την προηγούμενη μέρα, 3/01/2022 και πραγματοποιείται πρόβλεψη για τις 5/01/2022 (την επόμενη μέρα) στη μία περίπτωση και για τις 3/02/2022 (τον επόμενο μήνα) στην άλλη έτσι ώστε να μελετηθεί ποια μεθοδολογία παράγει καλύτερα αποτελέσματα.

3.1.3 Συλλογή δεδομένων

Τα βασικά ιστορικά δεδομένα που χρησιμοποιήθηκαν στην παρούσα διπλωματική, αντλήθηκαν από την οικονομική σελίδα Yahoo Finance μέσω της οποίας παρέχονται ειδήσεις του χρηματιστηρίου, ιστορικά δεδομένα και οικονομικοί δείκτες εταιρειών, πληροφορίες και σχολιασμοί σχετικά με επενδύσεις και προσωπική διαχείριση οικονομικών. Η σελίδα αυτή διαθέτει τη δυνατότητα της εύκολης εξαγωγής δεδομένων και οικονομικών δεικτών με τη βοήθεια της γλώσσας προγραμματισμού Python, μια γλώσσα που τα τελευταία χρόνια χρησιμοποιείται όλο και περισσότερο στον οικονομικό τομέα. Η εξαγωγή των δεδομένων πραγματοποιήθηκε με τη βοήθεια του πακέτου Pandas Datareader της python το οποίο χρησιμοποιώντας πηγές από το διαδίκτυο συμπεριλαμβανομένου και του Yahoo Finance, μετατρέπει τις πληροφορίες

σε dataframes αντικείμενα. Για την οπτικοποίηση των χαρακτηριστικών αυτών χρησιμοποιήθηκε πληθώρα βιβλιοθηκών της python συμπεριλαμβανομένων της matplotlib, seaborn, plotly express καθώς και με χρήση excel.

Συγκεκριμένα, τα δεδομένα που αντλήθηκαν δωρεάν από τη σελίδα αυτή είναι :

- **Ημερομηνία** : Η αναλυτική ημερομηνία σε μορφή yyyy/mm/dd για την οποία αντλήθηκαν τα ιστορικά δεδομένα ανά ημέρα.
- **Τιμή ανοίγματος**
- **Τιμή κλεισίματος**
- **Υψηλότερη τιμή ημέρας**
- **Χαμηλότερη τιμή ημέρας**
- **Προσαρμοσμένη τιμή κλεισίματος** : Πρόκειται για την τιμή κλεισίματος μετά από προσαρμογές όπως οι διανομές μερισμάτων.
- **Volume**: Μετράει τον αριθμό των συναλλαγών που πραγματοποιούνται σε ένα χρονικό διάστημα, στην περίπτωση αυτή ανά ημέρα.

Ενδεικτικά, παρακάτω φαίνεται μια αρχική εμφάνιση των δεδομένων αμέσως μετά την εξαγωγή τους από την οικονομική σελίδα για την πρώτη διαθέσιμη μετοχή, την ABBV. Το εργαλείο στο οποίο έτρεξε ο κώδικας και εφαρμόστηκε το σύνολο των πειραμάτων είναι τα Jupiter Notebooks.

Date	Open	High	Low	Close	Adj Close	Volume
2021-01-04	107.180000	107.349998	103.860001	105.410004	97.798592	9523400
2021-01-05	105.410004	107.019997	104.629997	106.500000	98.809883	6823800
2021-01-06	104.750000	107.190002	104.180000	105.580002	97.956322	11017500
2021-01-07	106.110001	107.059998	105.570000	106.709999	99.004723	8196000
2021-01-08	106.839996	107.529999	105.760002	107.269997	99.524277	5345900

Εικόνα 15: Πρώτες σειρές των ιστορικών δεδομένων της μετοχής ABBV

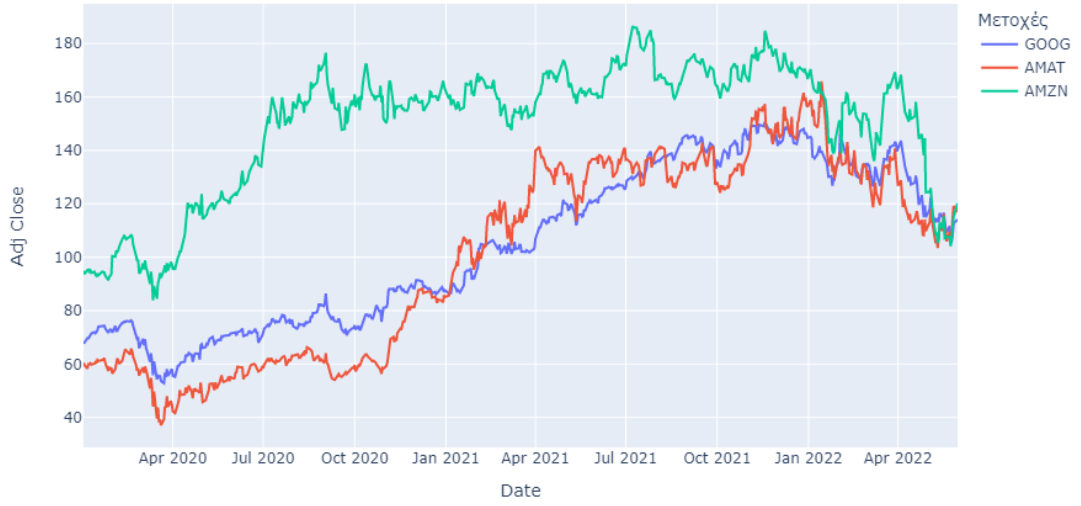
Από τα παραπάνω στοιχεία, για τη δημιουργία των χαρακτηριστικών που θα χρησιμοποιηθούν στη συνέχεια για την εκπαίδευση των μοντέλων απορρίφθηκε η στήλη «Volume» καθώς κύριος στόχος της τεχνικής ανάλυσης που θα πραγματοποιηθεί είναι ο προσδιορισμός μοτίβων που υποδεικνύουν μελλοντικές τάσεις ή κατευθύνσεις τιμών μετοχών έτσι ώστε να γίνει μια επιτυχημένη κατάταξη κι επομένως προτιμώνται τεχνικοί δείκτες που περιλαμβάνουν ως παραμέτρους τις υπόλοιπες στήλες όπως θα εξηγηθεί και αναλυτικά παρακάτω. Παρ' όλα αυτά , η στήλη αυτή δεν θα παραληφθεί κατά την συσταδοποίηση αφού καθιστά ευκολότερο το διαχωρισμό των μετοχών μεταξύ τους με βάση ορισμένα χαρακτηριστικά τους.

Χρησιμοποιώντας την ίδια μετοχή με προηγουμένως, φαίνεται και γραφικά πώς μεταβάλλεται η προσαρμοσμένη τιμή κατά τη διάρκεια του έτους 2021 μέχρι και τον Μάρτιο του 2022.

Εικόνα 16: ABBV προσαρμοσμένη τιμή κλεισίματος με την πάροδο του χρόνου

Εικόνα 17: Candlestick αναπαράσταση της μετοχής ABBV

Για καλύτερη κατανόηση των χρονοσειρών και το κριτήριο επιλογής του εύρους για τη διεξαγωγή της μελέτης, επιλέγονται και ορισμένες γνωστές μετοχές, η GOOG, AMAT και AMZN για οπτικοποίηση για ένα μεγαλύτερο διάστημα από αυτό που χρησιμοποιήθηκε από το 2020 έως και το Σεπτέμβριο του 2022.



Εικόνα 18: Αναπαράσταση προσαρμοσμένων τιμών κλεισίματος για την περίοδο 2020 με μέσα 2022

Παρατηρείτε ότι το Μάρτιο του 2020 οι τιμές και των τριών μετοχών έπεσαν δραματικά εξαιτίας της έξαρσης της πανδημίας καθώς και το Μάρτιο με Απρίλιο του 2022, πιθανότατα λόγω του πολέμου με τη Ρωσία.

3.1.4 Προεπεξεργασία δεδομένων

Τα δεδομένα αντλήθηκαν με χρήση Jupiter Notebooks και επεξεργάστηκαν με τη χρήση των γνωστών βιβλιοθηκών pandas, numpy, plotly, matplotlib. Κατά τη διαδικασία εξαγωγής των δεδομένων μέσω της οικονομικής σελίδας, παρατηρήθηκαν κενές τιμές για μία από τις διαθέσιμες μετοχές. Σκοπός του m6-competition είναι η κατάταξη των στοιχείων αυτών σύμφωνα με τις ποσοστιαίες συνολικές αποδόσεις τους ξεκινώντας από το χειρότερο (κλάση 1) μέχρι και το καλύτερο (κλάση 5). Προκειμένου στην περίπτωση αυτή να υπάρχει ομοιόμορφη κατανομή των μετοχών αυτών στις πέντε διαθέσιμες κλάσεις, επιλέγονται να αφαιρεθούν συνολικά πέντε επενδυτικά στοιχεία, με τυχαίο τρόπο και συμπεριλαμβανομένου της μετοχής με τις κενές τιμές. Με αυτόν τον τρόπο, από τις πλέον διαθέσιμες 45 μετοχές , οι 9 από αυτές θα αναλογούν στην κλάση 1, οι 9 στην κλάση 2 και αντίστοιχα σε κάθε κατηγορία θα αναλογούν ακριβώς 9 από τα διαθέσιμα επενδυτικά στοιχεία.

3.2 ΕΠΕΞΗΓΗΜΑΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

3.2.1 Εισαγωγή

Σε αυτό το κομμάτι θα αναλυθούν οι τεχνικές έννοιες που χρησιμοποιήθηκαν για την δημιουργία των χαρακτηριστικών που τροφοδότησαν τα μοντέλα μηχανικής μάθησης. Η επεξηγηματική ανάλυση (Exploratory Data Analysis) είναι πολύ χρήσιμη για την καλύτερη κατανόηση του συνόλου δεδομένων καθώς συνοψίζονται τα κύρια χαρακτηριστικά και οπτικοποιούνται μέσω αντιπροσωπευτικών γραφικών παραστάσεων έτσι ώστε να εντοπιστούν μοτίβα, συσχετίσεις και να επιβεβαιωθεί ότι επιλέχθηκαν τα κατάλληλα χαρακτηριστικά.

Τα αρχικά ιστορικά δεδομένα εξάγονται στη μορφή ενός Dataframe, ενός δισδιάστατου πίνακα του οποίου οι στήλες έχουν τη δυνατότητα να διαθέτουν και διαφορετικούς τύπους δεδομένων, μέσω του οποίου μπορούν να ληφθούν και περισσότερες στατιστικές πληροφορίες για κάθε στήλη όπως η μέση τιμή, η τυπική απόκλιση, μέγιστη και ελάχιστη αξία καθώς και πόσο συχνά μία τιμή εμφανίζεται. Ενδεικτικά, οι παραπάνω πληροφορίες φαίνονται αναλυτικά και για την πρώτη διαθέσιμη μετοχή, την ABBV.

	Open	High	Low	Close	Adj Close	Volume
count	314.000000	314.000000	314.000000	314.000000	314.000000	3.140000e+02
mean	119.225669	120.365191	118.273376	119.424172	114.768040	7.019393e+06
std	14.509064	14.743716	14.518299	14.745561	15.846335	3.558215e+06
min	102.879997	103.889999	101.809998	102.300003	96.023247	2.740300e+06
25%	108.609997	109.394999	107.547501	108.580000	103.648844	5.250300e+06
50%	115.165001	116.000000	114.190002	115.145000	109.959518	6.312450e+06
75%	121.975002	122.447498	120.937500	121.780003	118.405987	8.025275e+06
max	162.990005	164.660004	162.100006	163.750000	160.865295	5.094320e+07

Εικόνα 19: Στατιστική περίληψη ιστορικών δεδομένων

Από τα διαθέσιμα ιστορικά δεδομένα, ο υπολογισμός και η ανάλυση των χαρακτηριστικών βασίζονται κυρίως στην προσαρμοσμένη τιμή κλεισίματος καθώς αυτή αντικατοπτρίζει την πραγματική αξία της τιμής της μετοχής λαμβάνοντας υπόψη οποιουσδήποτε παράγοντες ή ενέργειες που μπορεί να την επηρεάσουν όπως η διανομή μερισμάτων ή η έκδοση νέων μετοχών [43] [44]. Ως αποτέλεσμα, συνήθως προτιμάται από άλλες εναλλακτικές όπως η τιμή κλεισίματος για την ανάλυση επίδοσης των μετοχών ή τον υπολογισμό γνωστών μετρικών.

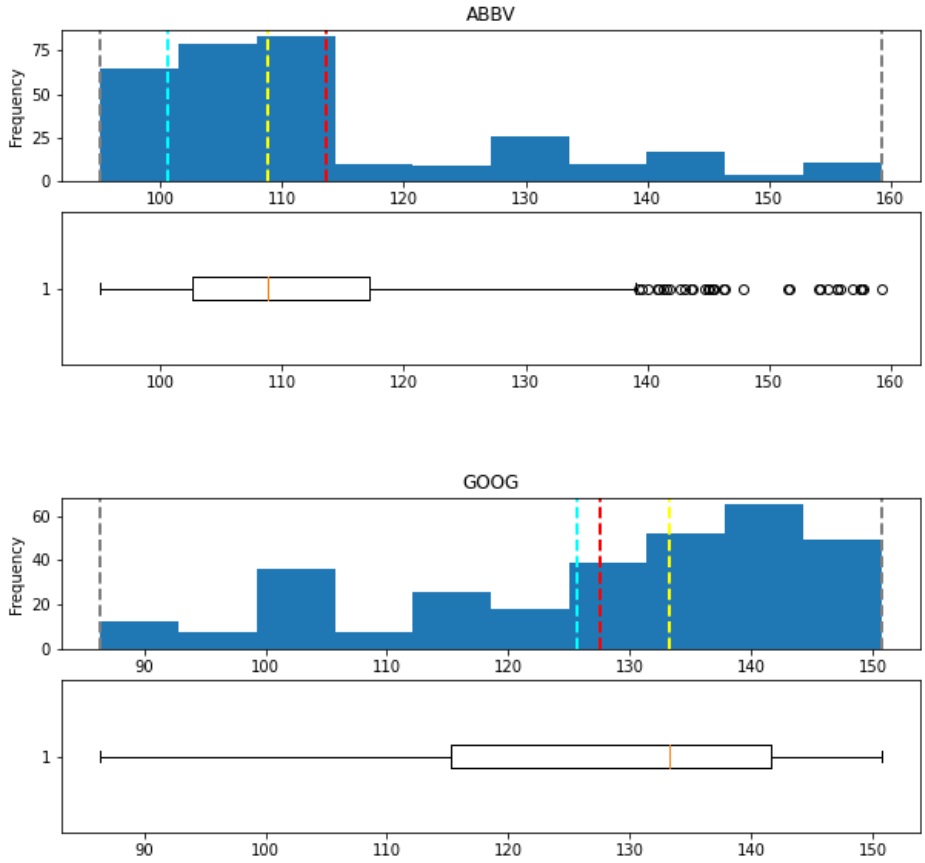
3.2.2 Ανάλυση χαρακτηριστικών επενδυτικών αγαθών

Για την τροφοδότηση των μοντέλων μηχανικής μάθησης δημιουργούνται ορισμένα χαρακτηριστικά και συγκεκριμένοι τεχνικοί δείκτες οι οποίοι βοηθάνε στην πρόβλεψη

μεταβολών εξετάζοντας ιστορικά δεδομένα. Αυτοί οι τεχνικοί δείκτες παρέχουν πληροφορίες σχετικά με τον όγκο, την ταχύτητα, την τάση, την ορμή και τις αποδόσεις των καθημερινών συναλλαγών. Για την κατασκευή τους επιλέχθηκε η βιβλιοθήκη TA-Lib η οποία χρησιμοποιείται ευρέως για την πραγματοποίηση τεχνικών αναλύσεων και περιλαμβάνει πάνω από 150 χρηματοοικονομικούς δείκτες.

Είναι σημαντικό να αναφερθεί ότι οι τιμές των μετοχών ποικίλλουν σημαντικά ανάλογα με την εταιρία. Για παράδειγμα, η τιμή της μετοχής της Domino's (DPZ) άνοιξε στα $330.11 και έκλεισε στα $320.14 έχοντας δηλαδή μία μείωση $9.97 ή 3.02%. Την ίδια μέρα, η τιμή της μετοχής PayPal (PYPL) πήγε από $90.8 σε $87.66, μείωση $3.2 ή 3.5%. Αν και η μείωση σε δολάρια της Domino's είναι η τριπλάσια από αυτή της PayPal, οι καθημερινές ποσοστιαίες μεταβολές τους είναι παρόμοιες [45]. Λαμβάνοντας υπόψη την πληροφορία αυτή για τις μεταβολές των τιμών, θα δημιουργηθούν ορισμένα χαρακτηριστικά για όλες τις μετοχές ανά ημέρα τα οποία είναι αντιπροσωπευτικά των πραγματικών μεταβολών σε σύγκριση πάντα και με τα υπόλοιπα επενδυτικά αγαθά.

Πολλά από τα χαρακτηριστικά που θα σχηματιστούν, χρησιμοποιούν τις προσαρμοσμένες τιμές κλεισίματος για αυτό και σε πρώτο βήμα είναι σημαντική η οπτικοποίηση του μεγέθους αυτού και της κατανομής του.



Εικόνα 20: Κατανομή τιμών για την ABBV και την GOOG

Τα χαρακτηριστικά που επιλέχθηκαν εξηγούνται αναλυτικά και παρακάτω:

- **Ποσοστιαία μεταβολή της τιμής ανοίγματος**

$$= \frac{Τιμή\ ανοίγματος\ μία\ μέρα\ πριν - Τιμή\ ανοίγματος\ δύο\ μέρες\ πριν}{Τιμή\ ανοίγματος\ δύο\ μέρες\ πριν}$$

Ο λόγος που χρησιμοποιούνται οι τιμές μία μέρα και δύο μέρες πριν είναι επειδή έχει γίνει η παραδοχή όπως θα εξηγηθεί και παρακάτω ότι την ημέρα που πραγματοποιείται η πρόβλεψη δεν είναι γνωστά ακόμα τα ημερήσια ιστορικά δεδομένα καθώς το χρηματιστήριο δεν έχει κλείσει ακόμα (επομένως η τιμή κλεισίματος για παράδειγμα δεν είναι ακόμα διαθέσιμη).

- **Ποσοστιαία μεταβολή της τιμής κλεισίματος**

$$= \frac{Τιμή\ κλεισίματος\ μία\ μέρα\ πριν - Τιμή\ κλεισίματος\ δύο\ μέρες\ πριν}{Τιμή\ κλεισίματος\ δύο\ μέρες\ πριν}$$

- **Ποσοστιαία μεταβολή της μέγιστης τιμή**

$$= \frac{Μέγιστη\ τιμή\ μία\ μέρα\ πριν - Μέγιστη\ τιμή\ δύο\ μέρες\ πριν}{Μέγιστη\ \ τιμή\ δύο\ μέρες\ πριν}$$

- **Ποσοστιαία μεταβολή της ελάχιστης τιμής**

$$= \frac{Ελάχιστη\ τιμή\ μία\ μέρα\ πριν - Ελάχιστη\ τιμή\ δύο\ μέρες\ πριν}{Ελάχιστη\ τιμή\ δύο\ μέρες\ πριν}$$

- **Ποσοστιαία μεταβολή της προσαρμοσμένης τιμής κλεισίματος**

$$= \frac{Προσαρμοσμένη\ τιμή\ κλεισίματος\ μία\ μέρα\ πριν - Προσαρμοσμένη\ τιμή\ κλεισίματος\ δύο\ μέρες\ πριν}{Προσαρμοσμένη\ τιμή\ κλεισίματος\ δύο\ μέρες\ πριν}$$

- **Ποσοστιαία μεταβολή της τιμής την προηγούμενη μέρα**

$$= \frac{Τιμή\ κλεισίματος\ μία\ μέρα\ πριν - Τιμή\ ανοίγματος\ μία\ μέρα\ πριν}{Τιμή\ κλεισίματος\ μία\ μέρα\ πριν}$$

Η διαφορά με τις προηγούμενες μεταβολές είναι ότι σε αυτήν την περίπτωση ελέγχεται η διακύμανση της τιμής κατά τη διάρκεια μίας μόνο μέρας και όχι σε σύγκριση με την προηγούμενη. Αυτό έχει ως σκοπό να εξεταστεί η ημερήσια κίνηση των τιμών σε αντίθεση με τις προηγούμενες που σκοπός τους είναι η σύγκριση των μεταβολών για ένα μόνο είδος τιμής κάθε φορά.

- **Ποσοστιαία μεταβολή των ακραίων τιμών (outliers) την προηγούμενη μέρα**

$$= \frac{Μέγιστη\ τιμή\ μία\ μέρα\ πριν - Ελάχιστη\ τιμή\ μία\ μέρα\ πριν}{Μέγιστη\ τιμή\ μία\ μέρα\ πριν}$$

- **Τυπική απόκλιση της προσαρμοσμένης τιμής κλεισίματος**
  Η τυπική απόκλιση είναι ένα στατιστικό μέγεθος που μετράει τη διακύμανση της τιμής δηλαδή τη διασπορά της πραγματικής γύρω από μια μέση τιμή. Υψηλές τιμές υποδεικνύουν μεγαλύτερη αστάθεια ενώ χαμηλές υποδεικνύουν μικρότερη διασπορά και επομένως μεγαλύτερη σταθερότητα. Ο υπολογισμός όπως φαίνεται και παρακάτω, βασίζεται στις τιμές κλεισίματος και στην παρούσα διπλωματική υπολογίζεται για μία περίοδο δέκα ημερών.

$$\sigma = \sqrt{\frac{\sum(xi - \mu)^2}{n}}$$

όπου σ = τυπική απόκλιση,
n = περίοδος / μέγεθος παρατηρήσεων,
xi = ημερήσια τιμή κλεισίματος,
μ = μέση τιμή παρατηρήσεων

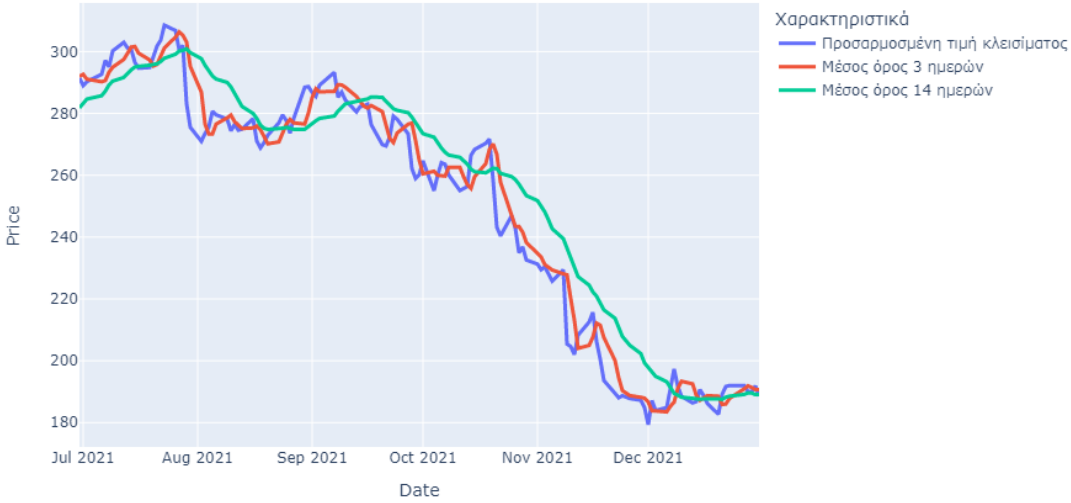
Εικόνα 21: Ανάλυση της τυπικής απόκλισης με βάση τα ιστορικά δεδομένα της XOM

Στην παραπάνω περίπτωση φαίνεται ένα παράδειγμα μιας υγιούς αγοράς. Κατά τη διάρκεια του Νοεμβρίου παρατηρούνται υψηλές τιμές κλεισίματος με σχετικά χαμηλή τυπική απόκλιση , ένα θετικό δείγμα για την ανοδική πορεία της μετοχής.

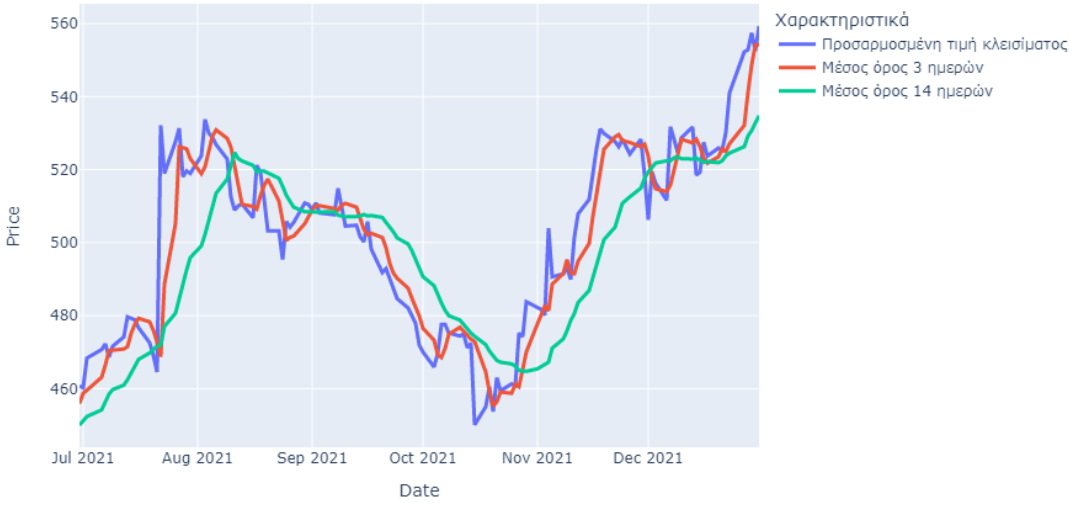
- **Κινητός μέσος όρος των προηγούμενων 3,6,10 και 14 ημερών**
  Σε κάθε περίπτωση χρησιμοποιούνται οι τιμές προσαρμοσμένου κλεισίματος των προηγούμενων ημερών και υπολογίζεται ο μέσος όρος τους. Έπειτα, ανά

ημέρα αυτό που χρησιμοποιείται είναι η ποσοστιαία μεταβολή αυτών των μέσων όρων για να υπάρχει μια ομοιόμορφη κλίμακα στο σύνολο των δεδομένων. Χρησιμοποιήθηκε μεγάλη ποικιλία, από τους οποίους στην πορεία δεν παραμένουν πάντα όλοι, για πειραματικούς λόγους έτσι ώστε να εξεταστεί ποιος από αυτούς συμβάλλει περισσότερο καθώς και για να μελετηθεί τόσο η βραχυπρόθεσμη όσο και η μεσοπρόθεσμη κίνηση των τιμών. Μεγαλύτερος ορίζοντας μέσων όρων δε θα ήταν επιθυμητός καθώς δεν υπάρχουν αρκετά δεδομένα διαθέσιμα.

Παρακάτω φαίνεται μια απεικόνιση των χαρακτηριστικών αυτών για δύο μετοχές που συζητήθηκαν και στην αρχή του υποκεφαλαίου, την PayPal και Domino’s.

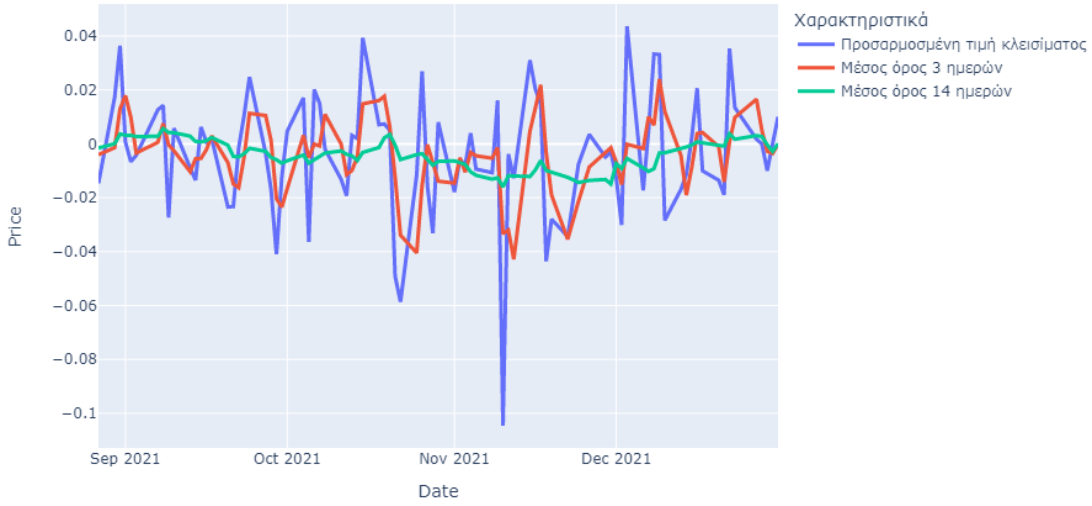


Εικόνα 22: PayPal (PYPL) Κινητοί Μέσοι όροι

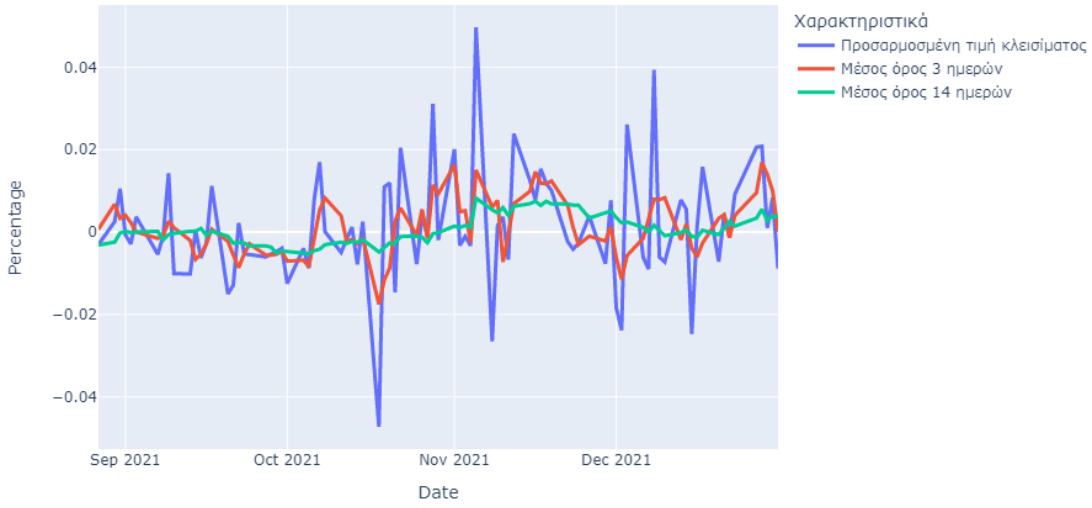


Εικόνα 23: Domino's (DPZ) Κινητοί Μέσοι όροι

Παρατηρείται ότι οι μέσοι όροι για περισσότερες μέρες έχουν πιο εξομαλυμένη απεικόνιση καθώς βασίζονται λιγότερο σε καθημερινές διακυμάνσεις σε αντίθεση με τον πιο βραχυπρόθεσμο μέσο όρο. Αυτό επιβεβαιώνεται και από τις ποσοστιαίες μεταβολές των τιμών αυτών με τις οποίες θα τροφοδοτηθούν τα μοντέλα.



Εικόνα 24: PayPal (PYPL) Ποσοστιαίες μεταβολές μέσων όρων



Εικόνα 25: Domino's (DPZ) Ποσοστιαίες μεταβολές μέσων όρων

- **Απόσταση από τη μέση τιμή**
  Παρόμοιας λογικής με την τυπική απόκλιση, με τη μόνη διαφορά ότι ο υπολογισμός είναι πιο ευθύς και απλός. Στην ουσία κάθε μέρα υπολογίζεται η απόσταση της τωρινής τιμής από τη μέση τιμή των τελευταίων έξι και δέκα ημερών αντίστοιχα δηλαδή :

$$\text{Distance from mean} = \frac{current\ closing\ price - average\ of\ 6/10\ days\ closing\ prices}{average\ of\ 6/10\ days\ closing\ prices}$$

- **Relative Strength Index (RSI) :**
  Πρόκειται για ένα δείκτη ο οποίος μετράει την ταχύτητα και την κίνηση των τιμών των μετοχών. Έχει κλίμακα από 0 έως 100 και χρησιμοποιείται όχι μόνο για την εύρεση υπερεκτιμημένων και υποτιμημένων μετοχών αλλά και για τον εντοπισμό γενικότερων τάσεων [46]. Συνήθως συνιστάται η χρήση του δείκτη αυτού για μια περίοδο 14 ημερών αλλά εξίσου γνωστή είναι και η χρήση για περιόδους από 9 μέχρι και 25 ημερών. Στην περίπτωση αυτή, λόγω και του μικρού πλήθους δεδομένων προτιμήθηκε η χρήση του δείκτη για την τυπική περίοδο των 14 ημερών καθώς και 6 ημερών, έτσι ώστε να υπάρχει μια μεσοπρόθεσμη και βραχυπρόθεσμη εικόνα της μετρικής αυτής.

  Για τον υπολογισμό του χρησιμοποιείται ο βασικός τύπος :

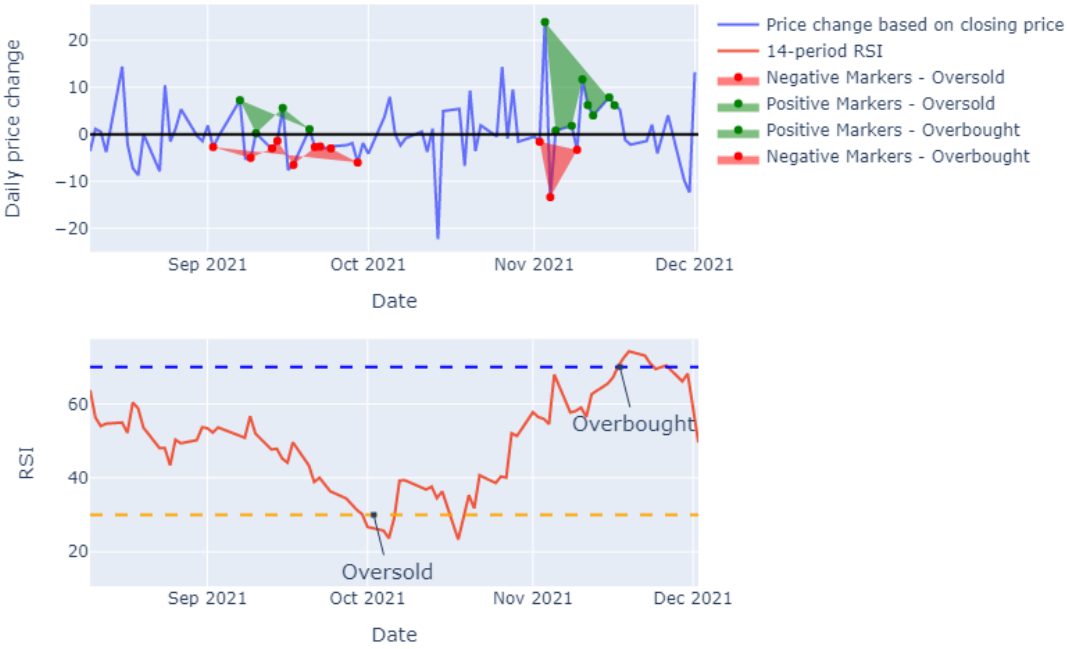
$$RSI = 100 - \frac{100}{1 + \frac{U}{D}}$$

  Όπου U = Μέση ανοδική μεταβολή τιμής και

  D = Μέση καθοδική μεταβολή τιμής

  Στην ουσία, ο δείκτης αυτός αναλύει πόσο καλή ήταν η απόδοση της μετοχής σε σύγκριση με τον εαυτό της χρησιμοποιώντας την προσαρμοσμένη τιμή κλεισίματος όπως φαίνεται και παρακάτω για τη μετοχή DPZ.

Εικόνα 26: Ανάλυση RSI με βάση τα ιστορικα δεδομένα της Domino's (DPZ)



Εικόνα 27: Ανάλυση RSI με βάση την τιμή προσαρμοσμένου κλεισίματος της Domino's(DPZ)

Παρατηρείται ότι ο δείκτης ανεβαίνει όταν ο αριθμός των ημερών όπου η τιμή κλεισίματος είναι μεγαλύτερη από την τιμή κλεισίματος την προηγούμενη μέρα αυξάνεται και το αντίθετο όταν είναι περισσότερες οι ημέρες όπου η τιμή κλεισίματος είναι μικρότερη από την ίδια τιμή την προηγούμενη ημέρα. Στην παραπάνω περίπτωση, κατά τη διάρκεια του Σεπτεμβρίου εμφανίζεται μια καθοδική πορεία γεγονός που δηλώνει ότι η μετοχή κάθε μέρα έκλεινε σε πιο χαμηλή τιμή από την προηγούμενη. Ως αποτέλεσμα το RSI μειώθηκε δραματικά ακόμα κι αν οι τιμές κλεισίματος τον Σεπτέμβριο θεωρούνταν ακόμα υψηλές συγκριτικά με τις υπόλοιπες, διότι σκοπός της ήταν να εντοπίσει την καθοδική τάση της τιμής και την ταχύτητα με την οποία μεταβλήθηκε. Φτάνοντας τιμές από 30 και κάτω, η μετοχή μπορεί να χαρακτηριστεί ότι έχει περισσότερους πωλητές απότι αγοραστές, εξού και η ονομασία της «oversold» ενώ σε αντίθετη περίπτωση, για τιμές πάνω από 80 θεωρείται «overbought» δηλαδή έχει περισσότερους αγοραστές απότι πωλητές.

- **Average Directional Index (ADX)** :

  Ο δείκτης αυτός χρησιμοποιείται για ποσοτικοποίηση της δύναμης της τάσης. Παρόμοια και με το προηγούμενο χαρακτηριστικό που αναλύθηκε, και οι δύο τεχνικοί δείκτες δημιουργήθηκαν από τον Welles Wilder και οι τιμές τους κυμαίνονται σε μία κλίμακα από 0 έως 100 και για περίοδο συνήθως 14 ημερών, με τη μόνη διαφορά ότι η μετρική αυτή είναι non-directional δηλαδή μετράει τη δύναμη της τάσης είτε αυτή είναι καθοδική, είτε ανοδική [47].

Εικόνα 28: : Ανάλυση ADX με βάση τα ιστορικά δεδομένα της PYPL

Αυτό επιβεβαιώνεται και από το παραπάνω γράφημα όπου παρατηρείται ότι ενώ η τιμή της μετοχής φαίνεται να μειώνεται παρουσιάζοντας μια καθοδική τάση, η τιμή ADX αυξάνεται εφόσον εντόπισε μια οποιαδήποτε τάση χωρίς να έχει σημασία το είδος της. Η μείωση της μετρικής ADX αφού φτάσει στο peak της δε σημαίνει ότι η τάση άλλαξε πορεία, αλλά ότι η δύναμη της τάσης αποδυναμώθηκε δηλαδή ότι συνεχίζεται η καθοδική πορεία λιγότερο έντονα.

Η μετρική αυτή μπορεί επίσης να μετρήσει την αλλαγή στο κλίμα της αγοράς παρακολουθώντας τις αλλαγές εντός του εύρους τιμών. Διαβάζοντας τη γραμμή ADX, μπορούμε να μετρήσουμε την υποκείμενη ισχύ της τάσης η οποία συνήθως χωρίζεται σε 2 βασικές κατηγορίες:

- Εάν η τιμή ADX είναι 0-25, τότε η ισχύς της τάσης θεωρείται αδύναμη (weak).
- Εάν η τιμή ADX είναι πάνω από 25, τότε η ισχύς της τάσης θεωρείται δυνατή (strong).

Ο δείκτης συνήθως απεικονίζεται στο ίδιο παράθυρο με τις δύο γραμμές δείκτη κίνησης κατεύθυνσης (DMI), από τις οποίες και προκύπτει η δημιουργία του.

Εικόνα 29: Ανάλυση ADX ως προς τους δείκτες κατεύθυνσης κίνησης της Google

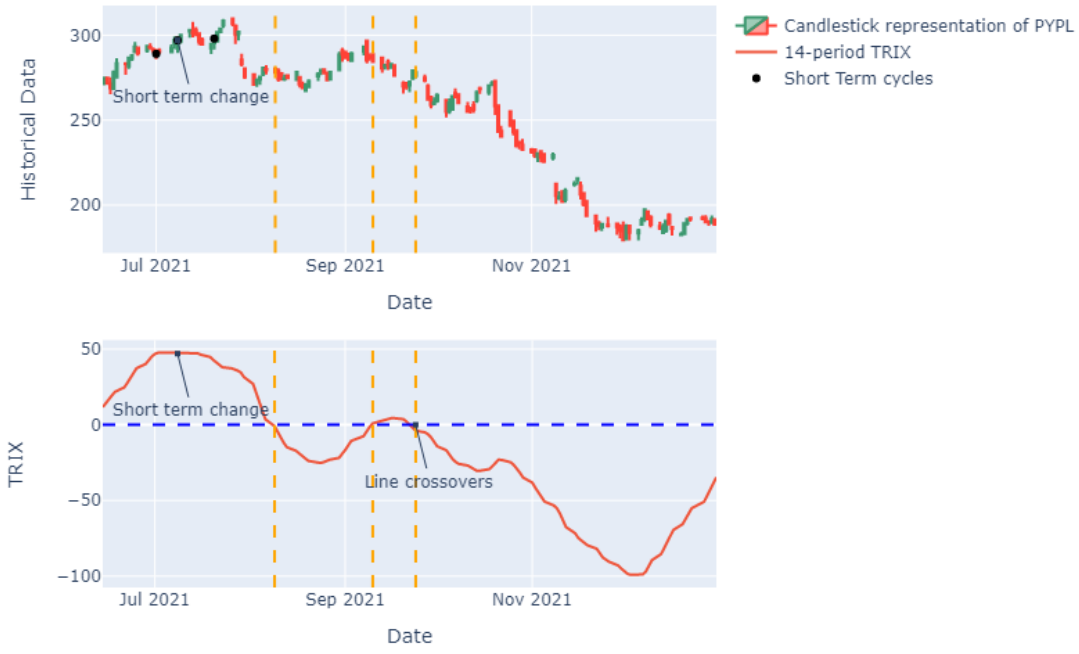
Στο παραπάνω παράδειγμα, όταν η θετική κατεύθυνση (+DI) ξεπερνά την αρνητική κατεύθυνση (-DI), τότε η γραμμή υποδηλώνει μια ανοδική τάση και σε αντίθετη περίπτωση υποδηλώνεται καθοδική τάση. Ο τρόπος λοιπόν που

ερμηνεύεται το παραπάνω διάγραμμα είναι ότι ενώ υπάρχει μια ανοδική τάση της τιμής της Google, αυτή δεν φαίνεται να έχει μεγάλη ισχύ καθώς η γραμμή ADX κατεβαίνει κάτω από 25.

Όσον αφορά την παρούσα διπλωματική χρησιμοποιήθηκε τόσο η μετρική για την τυπική περίοδο των 14 ημερών, όσο και για την πιο βραχυπρόθεσμη περίοδο των 6 ημερών ώστε να υπάρχει μια ομοιομορφία και με τα υπόλοιπα χαρακτηριστικά.

- **Triple exponential average (TRIX)**

Αυτός ο δείκτης που αναπτύχθηκε τη δεκαετία του 1980 από τον Jack Hutson, υπολογίζει την ποσοστιαία μεταβολή του τριπλού εκθετικού εξομαλυμένου μέσου όρου (EMA) και διαφέρει από τις προηγούμενες μετρικές καθώς διαθέτει όχι μόνο τη δυνατότητα να εντοπίζει τάσεις και σημάδια υπερτιμημένων ή υποτιμημένων αγορών, αλλά το πιο σημαντικό καταφέρνει να φιλτράρει πιθανό θόρυβο στις χρονοσειρές δηλαδή να εξαλείψει βραχυπρόθεσμες αλλαγές στην κατεύθυνση της αγοράς που δεν έχουν σχέση με τη γενική, κυρίαρχη τάση. Είναι γνωστός και ως «δείκτης παρόρμησης» και λειτουργεί υπολογίζοντας τη διαφορά της «εξομαλυμένης» εκδοχής της τιμής κλεισίματος.



Εικόνα 30: Ανάλυση TPIX με βάση τα ιστορικά δεδομένα της PYPL

Παρατηρείται ότι οι βραχυπρόθεσμες αλλαγές δεν επηρεάζουν τη γενική τάση όπως και αναμενόταν. Όταν χρησιμοποιείται ως δείκτης της ταχύτητας κίνησης του χρηματιστηρίου, μια θετική τιμή υποδηλώνει ότι η ταχύτητα αυξάνεται ενώ μια αρνητική τιμή ότι μειώνεται. Πολλοί αναλυτές πιστεύουν ότι όταν ο δείκτης περνά πάνω από τη μηδενική γραμμή δίνει σήμα αγοράς και όταν κλείνει κάτω

από το μηδέν δίνει σήμα πώλησης. Οποιαδήποτε απόκλιση παρατηρείται ανάμεσα στην τιμή και το δείκτη είναι σημάδι σημείων καμπής (turning points) της αγοράς. Γενικά, είναι ικανός να υποδείξει πότε υπάρχει μια αυξανόμενη ή «βυθισμένη» κίνηση στην αγορά και παράγει σήματα συναλλαγών παρόμοια με τον δείκτη MACD.

Williams %R (WILLR) :

Ο WILLR βασίζεται όπως και οι υπόλοιποι δείκτες στην τιμή κλεισίματος για μια περίοδο τυπικά 10 με 14 ημερών και υπολογίζει τη σχέση της τωρινής τιμής με βάση την μεγαλύτερη από τις υψηλές τιμές των προηγούμενων ημερών για την περίοδο που επιλέχθηκε (στην περίπτωση αυτή 14 ημέρες). Στην κλίμακα που έχει από 0 έως -100, τιμές του δείκτη μεγαλύτερες από -20 δηλώνουν ότι η τιμή είναι πιο κοντά στις υψηλότερες από το εύρος τιμών ενώ για τιμές μικρότερες από -80 ισχύει το αντίθετο. Ο υπολογισμός του είναι σχετικά απλός, όπως φαίνεται και παρακάτω:

$$WILLIAMS\%R = \frac{Highest\ High - Close}{Highest\ High - Lowest\ Low}$$

Εικόνα 31: Ανάλυση WILLR με βάση τα ιστορικά δεδομένα της META

Είναι σημαντικό να αναφερθεί παρατηρώντας και τη γραφική παράσταση ότι κατά τη διάρκεια του Σεπτέμβρη, όταν η τιμή του δείκτη περάσει το κατώφλι του -20 και συνεχίζει να έχει μια πτώση, αυτό είναι δείγμα μια καθοδικής τάσης, ενώ το αντίθετο ισχύει όταν η τιμή περάσει το κατώφλι του -80 και ακολουθεί ανοδική πορεία.

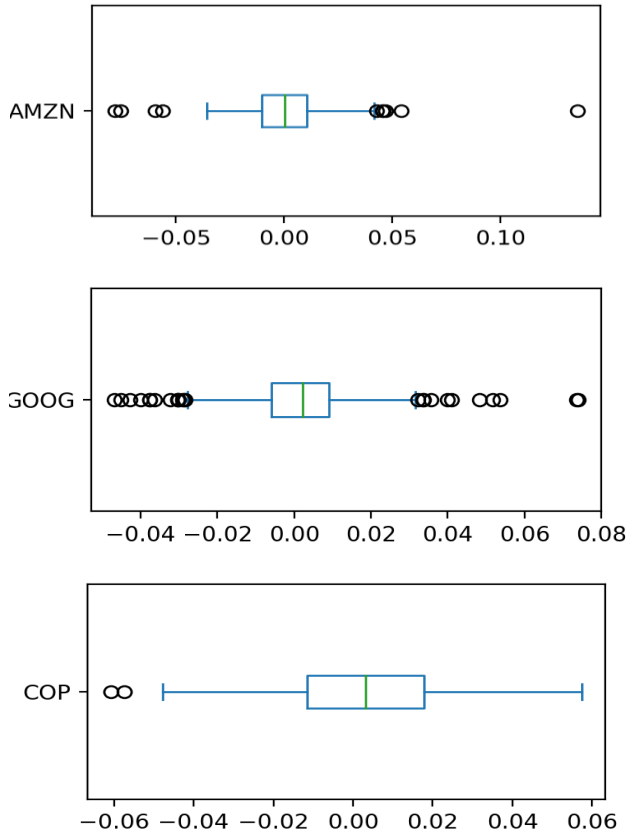
- **Moving Average Convergence Divergence (MACD):**
  Όπως δηλώνει και το όνομα του δείκτη αυτού, χρησιμοποιεί δύο κινητούς μέσους όρους, από τους οποίους αφαιρεί το μέσο όρο με το μεγαλύτερο ορίζοντα από το μέσο όρο με τον πιο βραχυπρόθεσμο. Λειτουργεί παρόμοια με τον δείκτη TRIX που επεξηγήθηκε παραπάνω με τη μόνη διαφορά ότι η αναπαράστασή του δεν είναι το ίδιο εξομαλυμένη. Αν και συνηθίζεται να χρησιμοποιείται για μακροπρόθεσμο ορίζοντα 26 ημερών και βραχυπρόθεσμο ορίζοντα 12 ημερών, για να εξασφαλιστεί μια ομοιομορφία στην κατασκευή των χαρακτηριστικών επιλέχθηκαν ορίζοντες 14 και 6 ημέρες αντίστοιχα. Συνηθίζεται να χρησιμοποιείται συνδυαστικά και με τη μετρική RSI καθώς η μία μετράει τη σχέση μεταξύ δύο εκθετικών κινητών μέσων όρων, ενώ η άλλη υπολογίζει την αλλαγή της τιμής σε σχέση με τις πρόσφατες υψηλές και χαμηλές τιμές.

Εικόνα 32: Ανάλυση MACD με βάση τα ιστορικά δεδομένα της META

- **Percentage Price Oscillator (PPO):**
  Η μετρική αυτή είναι πανομοιότυπη με την MACD με τη διαφορά ότι μετράει ποσοστιαίες διαφορές ανάμεσα σε δύο κινητούς εκθετικούς μέσους όρους, Συνήθως, αυτή η μετρική προτιμάται και κατά την επιλογή των χαρακτηριστικών για την εκπαίδευση των μοντέλων όπως θα εξηγηθεί και σε αργότερο υποκεφάλαιο, αλλά και γενικότερα από αναλυτές γιατί είναι πιο εύκολη στην ερμηνεία της, δηλαδή στη σύγκριση των τιμών μεταξύ τους. Με παρόμοια λογική, κατασκευάστηκε για μακρινό και κοντινό ορίζοντα 14 και 6 ημέρες αντίστοιχα. Ο υπολογισμός του είναι σχετικά απλός, όπως φαίνεται και παρακάτω :

$$PPO = \frac{Fast\ EMA - Slow\ EMA}{Fast\ EMA}$$

- **Ημερήσιες ποσοστιαίες αποδόσεις (Daily Percentage Returns) :**
  Αυτό το χαρακτηριστικό είναι ιδιαίτερα σημαντικό καθώς σύμφωνα με αυτό γίνεται και η κατηγοριοποίηση των μετοχών. Πρόκειται για τη διαφορά της τιμής προσαρμοσμένου κλεισίματος την προηγούμενη μέρα με την τιμή προσαρμοσμένου κλεισίματος δύο μέρες πριν. Ενδεικτικά, παρακάτω φαίνεται η κατανομή των ποσοστιαίων αποδόσεων για τρεις γνωστές μετοχές, την Amazon, την Google και την ConocoPhillips.



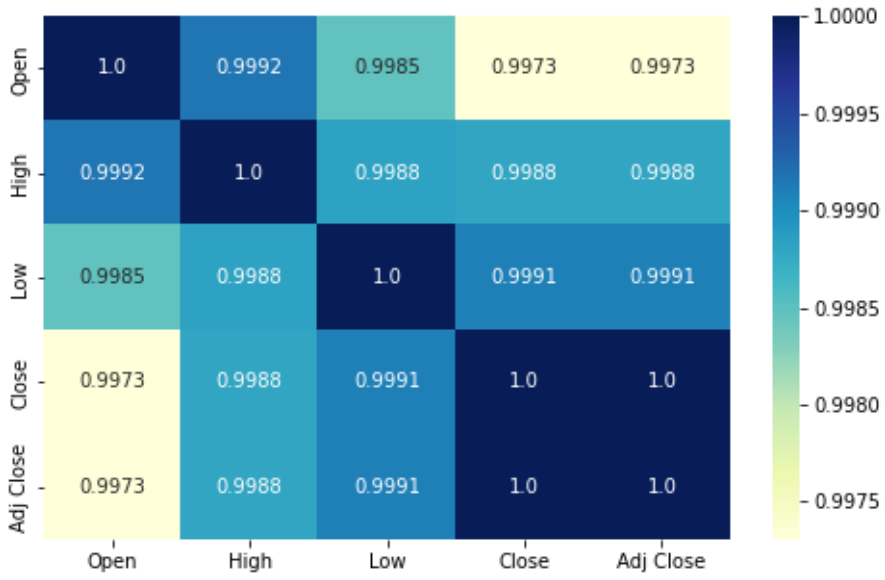
Εικόνα 33: Κατανομή των ποσοστιαίων αποδόσεων για την AMZN,GOOG,COP

- **Absolute rank :**
  Με βάση τις ημερήσιες ποσοστιαίες αποδόσεις που επεξηγήθηκαν παραπάνω, γίνεται μια κατάταξη των τιμών αυτών ανά ημέρα για όλες τις μετοχές. Για παράδειγμα, αν σε μία μέρα η μετοχή COP ήταν η καλύτερη με το μεγαλύτερο percentage return και η μετοχή CNC ήταν συγκριτικά με τις υπόλοιπες η χειρότερη με το μικρότερο percentage return , τότε σύμφωνα και με τις παραδοχές του m6-competition, η καλύτερη μετοχή θα έχει απόλυτη κατάταξη 45 ενώ η χειρότερη θα έχει απόλυτη κατάταξη 1. Κατά την εκπαίδευση του μοντέλου, εφόσον δεν είναι γνωστή η κατάταξη της τωρινής μέρας, το μοντέλο πάντα τροφοδοτείται με τους υπολογισμούς που έγιναν την προηγούμενη ημέρα.

3.2.3 Ανάλυση συσχετίσεων χαρακτηριστικών επενδυτικών αγαθών

Σε αυτό το υποκεφάλαιο, θα αναλυθούν σύντομα οι σχέσεις των μεταβλητών μεταξύ τους ώστε να επεξηγηθεί η επιλογή και σημαντικότητα τους. Σε κάθε μέθοδο, όπως και θα εξηγηθεί σε μετέπειτα κεφάλαιο, γίνεται στο τέλος επιλογή από αυτή την πληθώρα χαρακτηριστικών, καθώς ορισμένα από αυτά εμφανίζουν υψηλές συσχετίσεις μεταξύ τους και επομένως η επιλογή ορισμένων μόνο είναι η επιθυμητή μεθοδολογία.

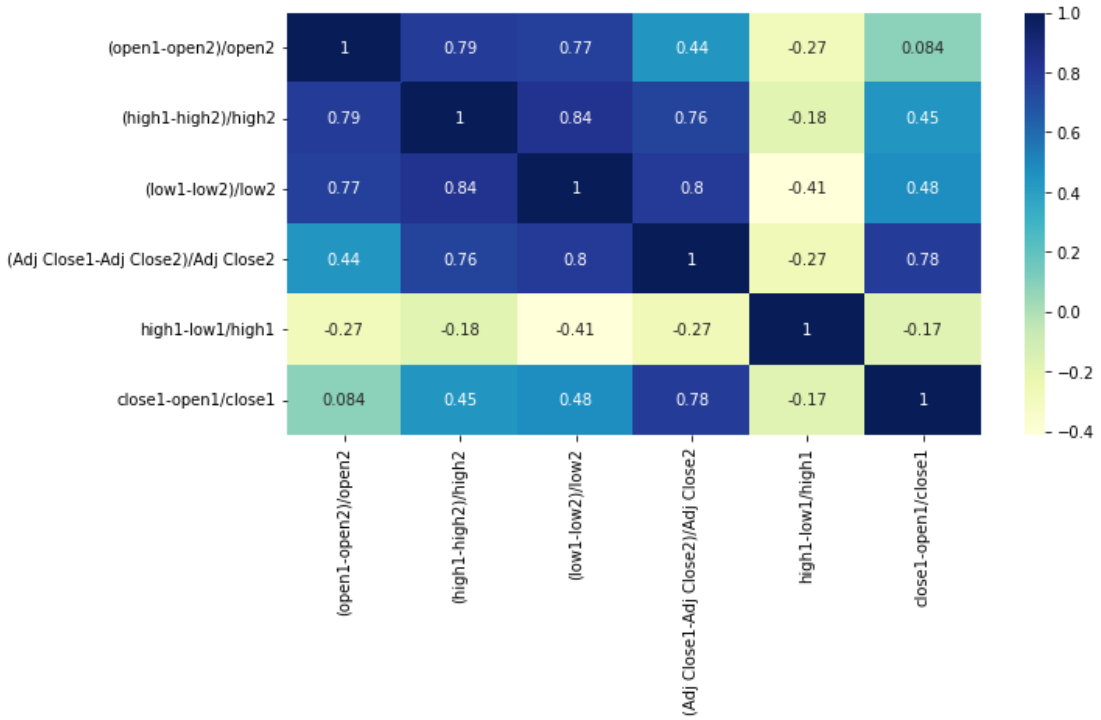
Καταρχάς, είναι σημαντικό να εξεταστεί η συσχέτιση των αρχικών ιστορικών δεδομένων, όπως αυτά αντλήθηκαν από την οικονομική σελίδα Yahoo Finance. Ενδεικτικά, θα εξεταστούν τα δεδομένα αυτά για τη γνωστή μετοχή PayPal.



Εικόνα 34: Συσχετίσεις ιστορικών δεδομένων της μετοχής PayPal

Παρατηρείται ότι από τα αρχικά χαρακτηριστικά εμφανίζεται μια υψηλή συσχέτιση αφού το σύνολο των τιμών τους είναι πάντα κοντά στη μονάδα. Αυτό είναι λογικό αφού οι ημερήσιες αυτές τιμές έχουν μικρές διαφορές μεταξύ τους.

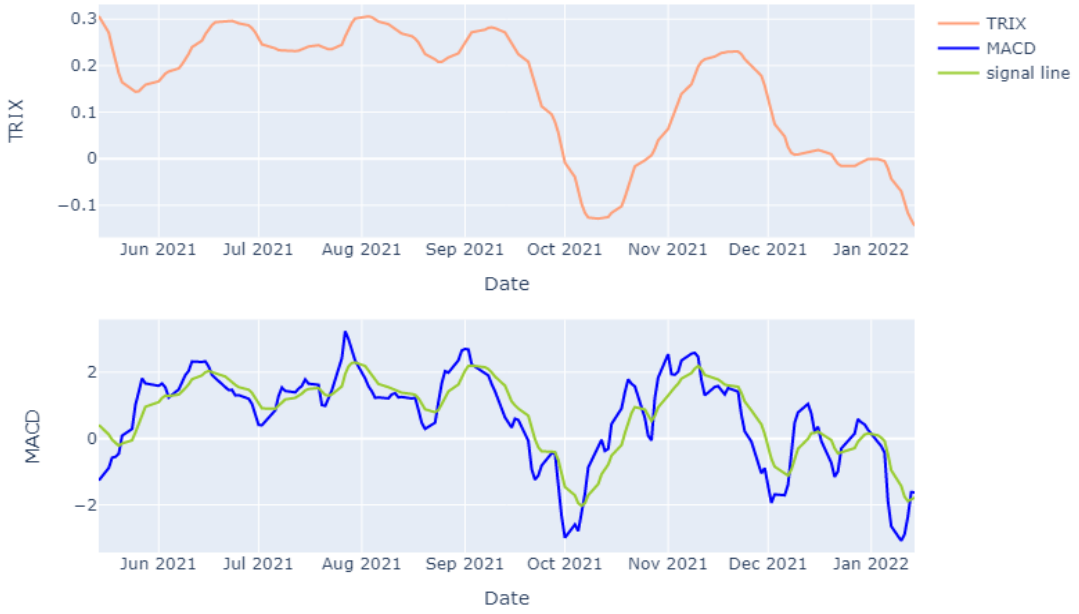
Χρησιμοποιώντας όμως τα παραπάνω δεδομένα, υπάρχει η δυνατότητα να αξιοποιηθούν οι ημερήσιες μεταβολές τους τόσο σε σύγκριση με την προηγούμενη μέρα, όσο και σε σύγκριση με τη διακύμανση των ακραίων τιμών τους όπως φαίνεται και παρακάτω :



Εικόνα 35: Συσχετίσεις νέων χαρακτηριστικών αξιοποιώντας τα ιστορικά δεδομένα

Αυτά τα έξι καινούρια χαρακτηριστικά που προέκυψαν από τα ιστορικά δεδομένα, προσφέρουν μεγαλύτερη ποικιλία πληροφοριών που μπορεί να αξιοποιηθεί στη συνέχεια.

Μια ακόμα σημαντική σχέση παρατηρείται και για τις μετρικές TRIX, MACD και RSI. Όπως εξηγήθηκε και παραπάνω, οι δείκτες TRIX και MACD λειτουργούν με παρόμοιο τρόπο και οι γραφικές τους παραστάσεις ερμηνεύουν τις ίδιες πληροφορίες με τη διαφορά ότι ο πρώτος δείκτης έχει πιο εξομαλυμένη αναπαράσταση. Παρ' όλα αυτά, πολλές φορές προτιμάται ο συνδυασμός τους καθώς παρέχουν πληροφορίες εκ των προτέρων για πιθανή είσοδο ή έξοδο από μία νέα τάση, γεγονός που βοηθάει τόσο στις προβλέψεις όσο και στην ανάπτυξη στρατηγικών συναλλαγών.



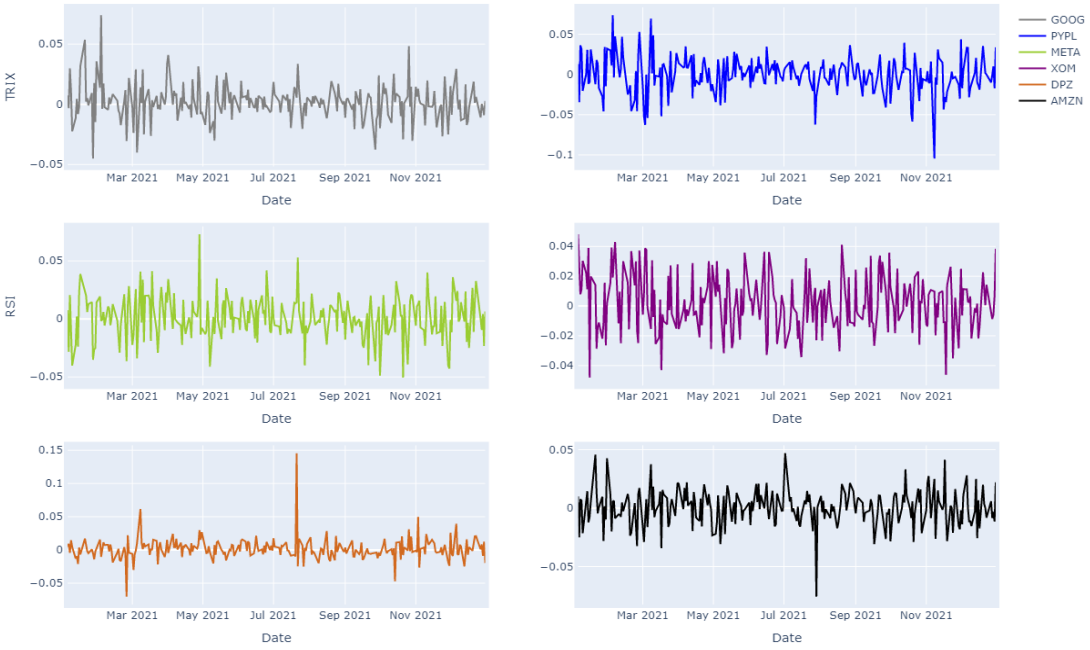
Εικόνα 36: Συσχέτιση μετρικών MACD και TRIX της Google

Και οι δύο γραφικές παραστάσεις έχουν παρόμοια σχήματα με τη μεγαλύτερη διαφορά να είναι ότι η TRIX εμφανίζει λιγότερη τραχύτητα και τείνει να αλλάξει πορεία λίγο πιο αργά.

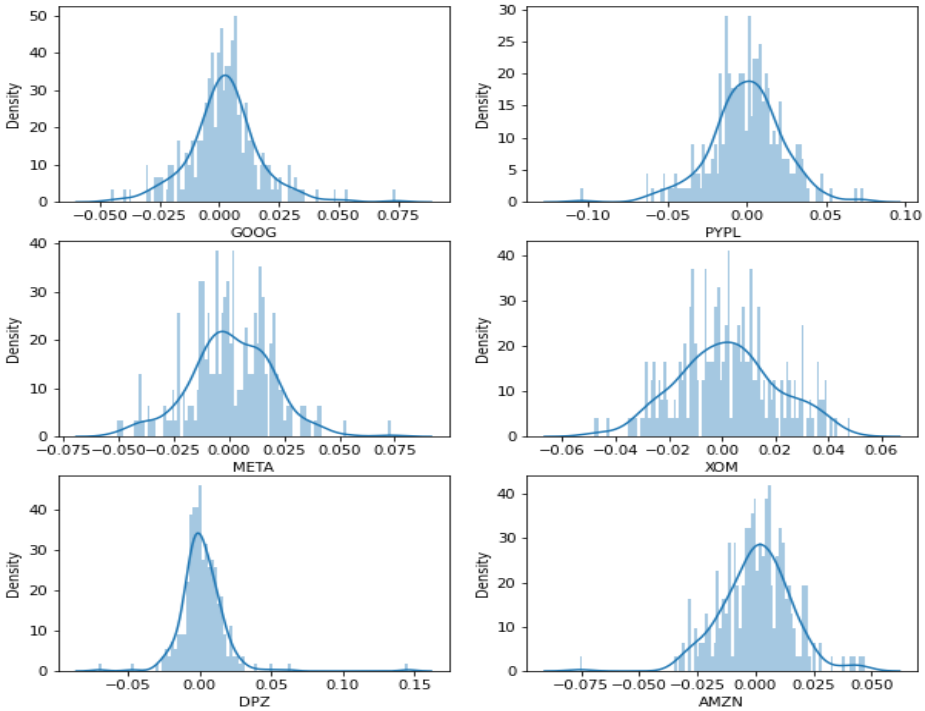
Το ίδιο σημαντικός είναι και ο συνδυασμός του TRIX ή του MACD με το RSI καθώς είναι χρήσιμα συνδυαστικά στο να προσφέρουν πληροφορίες για στρατηγικές συναλλαγών και αντιστροφές τάσεων δηλαδή κινήσεων των τιμών. Ενώ οι δείκτες RSI λαμβάνουν υπόψη την αναλογία των κερδών και των ζημιών σε σύγκριση με την προηγούμενη ημέρα, το MACD είναι ουσιαστικά ένας κινητός μέσος όρος της τιμής. Επομένως, δεν είναι απαραίτητο ότι τόσο ο RSI όσο και ο MACD θα δώσουν το ίδιο σήμα σε μια συγκεκριμένη χρονική στιγμή για αυτό και άλλωστε χρησιμοποιούνται συνδυαστικά πολλές φορές.

Είναι ιδιαίτερα σημαντικό εκτός από τις σχέσεις μεταξύ των χαρακτηριστικών να εξεταστούν και οι συσχετίσεις ανάμεσα στις μετοχές. Επειδή το σύνολο των διαθέσιμων μετοχών είναι μεγάλο, επιλέγονται από αυτές ορισμένες γνωστές μετοχές όπως η Google, PayPal, META, XOM, Domino’s και Amazon.

Αρχικά, σχεδιάζονται οι γραφικές παραστάσεις των ημερήσιων ποσοστιαίων αποδόσεών τους κατά τη διάρκεια του έτους 2021.

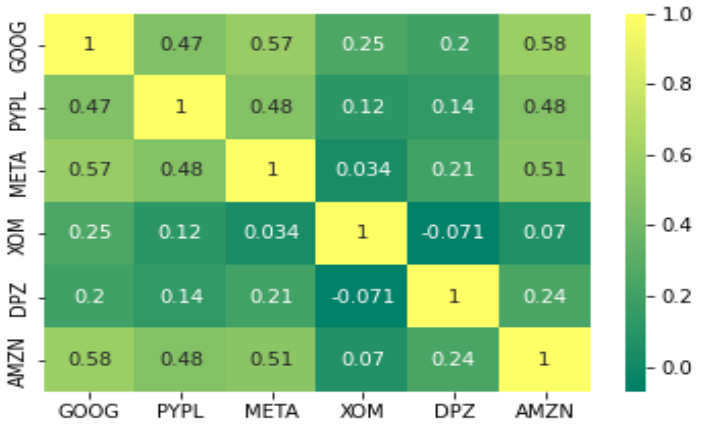


Εικόνα 37: Ποσοστιαίες αποδόσεις μετοχών (2021)



Εικόνα 38: Κατανομή ποσοστιαίων αποδόσεων

Για καλύτερη κατανόηση των αποτελεσμάτων, χρησιμοποιείται ο συντελεστής συσχέτισης Pearson που προτιμήθηκε και παραπάνω έτσι ώστε να πραγματοποιηθεί μια σύγκριση των καθημερινών αυτών ποσοστιαίων αποδόσεων μεταξύ των μετοχών.



Εικόνα 39: Συσχέτιση ημερήσιων ποσοστιαίων αποδόσεων μεταξύ των μετοχών

Από το γράφημα συμπεραίνεται ότι ορισμένες μετοχές έχουν μεγαλύτερη συσχέτιση μεταξύ τους ως προς τα daily returns, όπως για παράδειγμα η Amazon με τη Meta καθώς ο συντελεστής τους ξεπερνάει το 0.5 ενώ σε άλλες περιπτώσεις ισχύει το αντίθετο με ισχυρό παράδειγμα τη μετοχή της Amazon με την XOM όπου εμφανίζεται πολύ χαμηλή συσχέτιση, σχεδόν κοντά στο 0. Από αυτήν την παρατήρηση δημιουργείται η ανάγκη για περαιτέρω ανάλυση και διαχωρισμό των διαθέσιμων μετοχών σε ομάδες ανάλογα με ορισμένα αντιπροσωπευτικά χαρακτηριστικά, έτσι ώστε να ελεγχθεί εάν η ομαδοποίησή τους εναλλακτικά θα μπορούσε να οδηγήσει σε πιο αποδοτική εκπαίδευση των μοντέλων.

3.2.4 Συσταδοποίηση μετοχών και ανάλυση σε κύριες συνιστώσες

Για την επιτυχημένη συσταδοποίηση των μετοχών, ένα μόνο δεδομένο όπως οι ποσοστιαίες αποδόσεις δεν πληρεί τις προϋποθέσεις αυτές, οπότε για αυτό το λόγο χρησιμοποιούνται και περισσότερα χαρακτηριστικά για την ανάλυση, ορισμένα προαναφερθέντα καθώς και μερικά καινούρια που προστίθενται στη συνέχεια. Είναι σημαντικό να αναφερθεί ότι στην περίπτωση αυτή όπου σκοπός είναι η ομαδοποίηση των μετοχών ανάλογα με τα μεγέθη τους, θα μπορούσαν να μας ενδιαφέρουν τόσο τα ποσοστιαία όσο και τα απόλυτα μεγέθη καθώς έχει σημασία οι ομάδες που θα δημιουργηθούν να είναι αντιπροσωπευτικές. Για παράδειγμα, στην περίπτωση αυτή παρουσιάζει ενδιαφέρον το ότι η τιμή της μετοχής Domino's κυμαίνεται περίπου στα 300$ ενώ της Google στα 100$ ασχέτως των ποσοστιαίων μεταβολών τους, υποδεικνύοντας μια σημαντική διαφορά στα μεγέθη που αξίζει να ληφθεί υπόψη κατά τη διάρκεια της ανάλυσης.

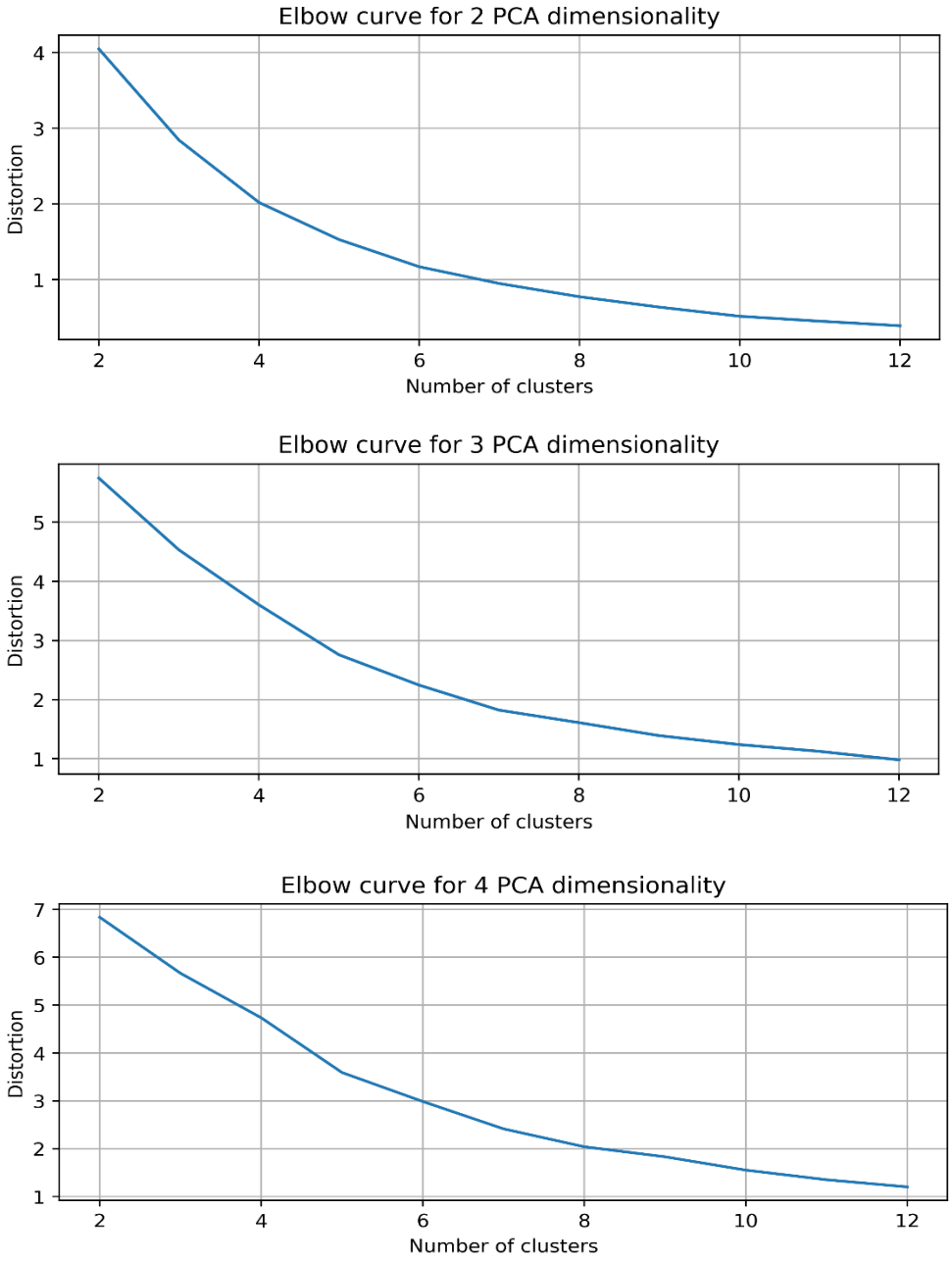
Για την ανάλυση σε ομάδες, χρησιμοποιήθηκαν ιστορικά δεδομένα 15 μηνών, ξεκινώντας από τον Ιανουάριο του 2021 έως και το Μάρτιο του 2022 με τα οποία για

κάθε μία από τις διαθέσιμες μετοχές, δημιουργήθηκε ένα σύνολο χαρακτηριστικών [48]. Από αυτά, έπειτα από πειραματισμούς κι επαναλήψεις της διαδικασίας ομαδοποίησης και με κριτήριο την ελαχιστοποίηση του σφάλματος από το σύνολο των χαρακτηριστικών που κατασκευάστηκαν αρχικά, παρέμειναν τα παρακάτω :

- **Price**: Ο μέσος όρος των τιμών προσαρμοσμένου κλεισίματος.
- **Returns_mean:** Η μέση τιμή των ημερήσιων ποσοστιαίων αποδόσεων κάθε μετοχής.
- **Daily_volatility**: Η μέση τιμή της ποσοστιαίας τυπικής απόκλισης των προσαρμοσμένων τιμών κλεισίματος.
- **Cumulative_returns**: Πρόκειται για τη συνολική, αθροιστική μεταβολή της τιμής κλεισίματος της επένδυσης για το καθορισμένο χρονικό διάστημα.
- **rolling_rets_1** : Πρόκειται για τον υπολογισμό της συνολικής μεταβολής της τιμής κλεισίματος (cumulative return) κυλιόμενα ανά μήνα. Από τα αποτελέσματα αυτά προκύπτει ένας μέσος όρος των τιμών αυτών έτσι ώστε να υπάρχει και μια βραχυπρόθεσμη εικόνα των μηνιαίων συνολικών αποδόσεων.
- **Volume :** Η μέση τιμή του όγκου των ημερήσιων συναλλαγών.

Τα δεδομένα αυτά στη συνέχεια κανονικοποιήθηκαν με τη μέθοδο του min-max scaling χρησιμοποιώντας τη βιβλιοθήκη mlxtend (machine learning extensions) της python. Για την ομαδοποίηση χρησιμοποιήθηκε ο μη επιβλεπόμενος αλγόριθμος K-Means σκοπός του οποίου είναι να διαχωρίσει τις μετοχές σε Κ ομάδες στα οποία κάθε μετοχή ανήκει στο σύμπλεγμα με τον πλησιέστερο μέσο όρο/κέντρο. Επειδή ο αλγόριθμος προϋποθέτει να είναι γνωστός ο αριθμός των συστάδων εκ των προτέρων προτού προσαρμοστεί το μοντέλο στα δεδομένα, αντί να επιλεγεί αυθαίρετα, χρησιμοποιήθηκε η μέθοδος «Elbow curve» για να επισημανθεί η σχέση ανάμεσα στον αριθμό των επιλεγμένων συστάδων σε κάθε περίπτωση και το άθροισμα των τετραγωνικών σφαλμάτων τους (sum of squared errors-SSE). Πρόκειται για μια εμπειρική μέθοδο που βοηθάει στην βέλτιστη επιλογή της τιμής k.

Στη συνέχεια, εφαρμόστηκε στα αρχικά δεδομένα μια στατιστική διαδικασία γνωστή και ως Ανάλυση σε Κύριες Συνιστώσες με σκοπό τη μείωση της διαστατικότητας καθώς και την ερμηνεία και οπτικοποίηση των δεδομένων. Χρησιμοποιώντας την προκαθορισμένη συνάρτηση PCA της βιβλιοθήκης scikit learn, πραγματοποιήθηκαν δοκιμές τόσο για την εύρεση του βέλτιστου αριθμού συστάδων όσο και για την εύρεση των καλύτερων κύριων συνιστωσών, έχοντας ως κριτήριο επιλογής την ελαχιστοποίηση του αθροίσματος των τετραγωνικών σφαλμάτων τους.

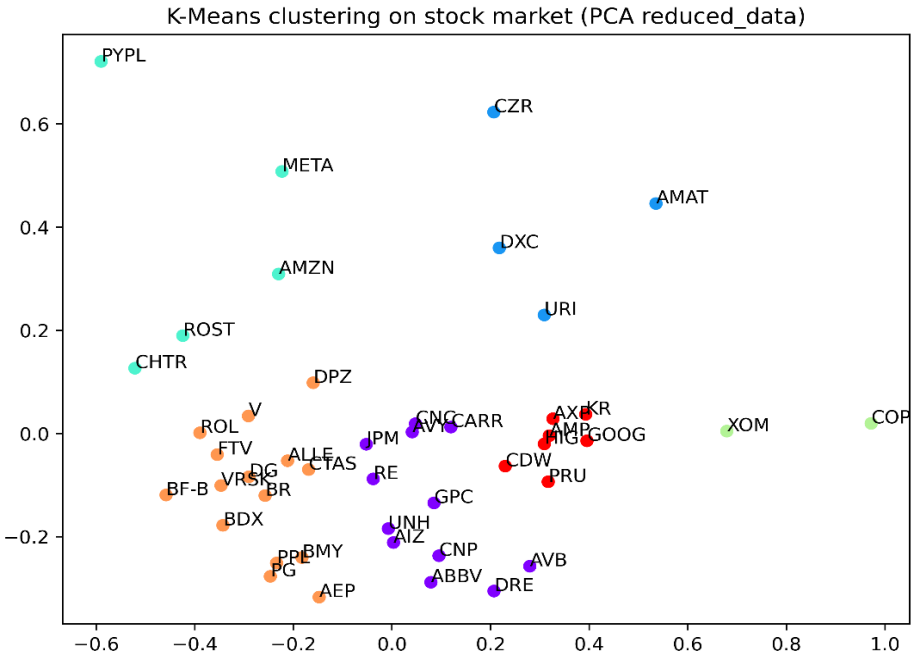


Εικόνα 40: Elbow curve for different PCA components

Παρατηρείται ότι η μείωση σε δύο διαστάσεις παράγει τις χαμηλότερες μέσες τετραγωνικές αποστάσεις μεταξύ κάθε σημείου δεδομένων και του αντίστοιχου κέντρου του (distortion) καθώς και είναι εύκολη η οπτικοποίησή της με τη χρήση ενός scatterplot. Όσον αφορά την επιλογή του βέλτιστου αριθμού συστάδων, από τα παραπάνω επιλέγεται ο διαχωρισμός των μετοχών σε 6 ομάδες καθώς σε εκείνη την περιοχή αρχίζει να ελαχιστοποιείται το σφάλμα με πιο αργούς ρυθμούς. Στον παρακάτω πίνακα παρουσιάζεται η τελική οργάνωση των μετοχών σε συστάδες καθώς και η οπτικοποίηση των αποτελεσμάτων τους (βλέπε πίνακα 1).

Clusters	*Stocks belonging in the cluster*
1	ABBV, AIZ, AVB, AVY, CARR, CNC, CNP, DRE, GPC, JPM, RE, UNH
2	AMAT, CZR,DXC, URI
3	AMZN, CHTR, META, PYPL, ROST
4	COP, XOM
5	AEP, ALLE, BDX, BF-B, BMY, BR, CTAS, DG, DPZ, FTV, PG, PPL, ROL, V, VRSK
6	AMP, AXP, CDW, GOOG, HIG, KR, PRU

Πίνακας 1: Διαχωρισμός μετοχών σε ομάδες (clusters)



Εικόνα 41: K-means clustering with k=6

Ελαχιστοποιώντας το σφάλμα σε 1.164, δημιουργούνται συνολικά 6 ομάδες με τα εξής χαρακτηριστικά :

- Cluster 1 (μωβ): Η ομάδα αυτή αποτελείται από μετοχές με χαμηλή διακύμανση και χαμηλό όγκο συναλλαγών.

ID	*Returns mean*	*Price*	*Daily volatility*	*Cumulative returns*	*rolling_rets*	*Volume*	*PC1*	*PC2*
ABBV	0.666	0.139	0.087	0.323	0.667	0.097	0.079	-0.288
AIZ	0.577	0.199	0.136	0.34	0.576	0.002	0.003	-0.211
AVB	0.676	0.288	0.096	0.639	0.678	0.006	0.28	-0.257
AVY	0.47	0.267	0.289	0.516	0.467	0.003	0.041	0.003

Πίνακας 2: Ενδεικτικές τιμές και συντελεστές της πρώτης συστάδας

- Cluster 2 (σκούρο μπλε): Η ομάδα αυτή αποτελείται από μετοχές με μεσαίες προς υψηλές ποσοστιαίες ημερήσιες και αθροιστικές αποδόσεις με πολύ μεγάλη όμως διακύμανση.

ID	*Returns mean*	*Price*	*Daily volatility*	*Cumulative returns*	*rolling_rets*	*Volume*	*PC1*	*PC2*
AMAT	0.694	0.167	0.835	0.791	0.598	0.114	0.535	0.446
CZR	0.499	0.107	1	0.485	0.457	0.038	0.207	0.623
DXC	0.563	0.014	0.756	0.471	0.508	0.029	0.218	0.36
URI	0.672	0.469	0.581	0.638	0.598	0.006	0.309	0.23

Πίνακας 3: Ενδεικτικές τιμές και συντελεστές της δεύτερης συστάδας

- Cluster 3 (ανοιχτό μπλε): Η ομάδα αυτή αποτελείται από γνωστές μετοχές όπως η Paypal, η Meta και η Amazon με χαμηλές αθροιστικές αποδόσεις και μεσαία διακύμανση.

ID	*Returns mean*	*Price*	*Daily volatility*	*Cumulative returns*	*rolling_rets*	*Volume*	*PC1*	*PC2*
AMZN	0.398	0.218	0.402	0.141	0.388	1	-0.23	0.309
CHTR	0.267	1	0.255	0.135	0.333	0.012	-0.521	0.127
META	0.297	0.441	0.681	0.282	0.3	0.319	-0.223	0.508
ROST	0.234	0.137	0.434	0.071	0.253	0.027	-0.424	0.19

Πίνακας 4: Ενδεικτικές τιμές και συντελεστές της τρίτης συστάδας

- Cluster 4 (πράσινο): Πρόκειται για την ομάδα που περιλαμβάνει μετοχές του ενεργειακού τομέα. Αυτό είναι λογικό καθώς το τελευταίο διάστημα λόγω της ενεργειακής κρίσης, οι εταιρείες αυτές αν και με σχετικά χαμηλές τιμές προσαρμοσμένου κλεισίματος σε σύγκριση και με τις υπόλοιπες διαθέσιμες μετοχές, έχουν αποδώσει πολύ μεγάλες ημερήσιες και αθροιστικές ποσοστιαίες αποδόσεις καθώς και υψηλές μεταβολές στην τιμή επένδυσης.

ID	*Returns mean*	*Price*	*Daily volatility*	*Cumulative returns*	*rolling_rets*	*Volume*	*PC1*	*PC2*
COP	1	0.059	0.548	1	1	0.128	0.972	0.02
XOM	0.857	0.053	0.412	0.816	0.834	0.369	0.679	0.005

Πίνακας 5: Ενδεικτικές τιμές και συντελεστές της τέταρτης συστάδας

- Cluster 5 (πορτοκαλί): Η ομάδα αυτή αποτελείται από μετοχές με μεσαίες ημερήσιες ποσοστιαίες αποδόσεις αλλά χαμηλές συνολικές μεταβολές της τιμής της επένδυσης για το προκαθορισμένο χρονικό διάστημα (cumulative returns) καθώς και γενικότερα χαμηλή τιμή μετοχών.

ID	*Returns mean*	*Price*	*Daily volatility*	*Cumulative returns*	*rolling_rets*	*Volume*	*PC1*	*PC2*
AEP	0.512	0.091	0.014	0.2	0.539	0.038	-0.148	-0.317
ALLE	0.354	0.158	0.223	0.263	0.408	0.006	-0.212	-0.053
BDX	0.41	0.338	0.084	0.111	0.422	0.016	-0.343	-0.178
BF-B	0.288	0.071	0.149	0.001	0.332	0.01	-0.458	-0.119

Πίνακας 6: Ενδεικτικές τιμές και συντελεστές της πέμπτης συστάδας

- Cluster 6 (κόκκινο): Αποτελείται από έναν πυρήνα μετοχών πολύ κοντά συνυφασμένων μεταξύ τους όπως φαίνεται και στο παραπάνω σχήμα οι οποίες έχουν μεσαίες ποσοστιαίες αποδόσεις και χαμηλές τιμές και όγκο συναλλαγών.

ID	*Returns mean*	*Price*	*Daily volatility*	*Cumulative returns*	*rolling_rets*	*Volume*	*PC1*	*PC2*
AMP	0.684	0.369	0.353	0.656	0.644	0.005	0.319	-0.004
AXP	0.68	0.21	0.417	0.592	0.667	0.05	0.327	0.029
CDW	0.594	0.234	0.275	0.616	0.588	0.008	0.23	-0.063
GOOG	0.674	0.161	0.287	0.715	0.66	0.379	0.396	-0.014

Πίνακας 7: Ενδεικτικές τιμές και συντελεστές της έκτης συστάδας

3.3 ΜΕΘΟΔΟΛΟΓΙΚΗ ΠΡΟΣΕΓΓΙΣΗ

3.3.1 Εισαγωγή

Όπως έχει ήδη αναφερθεί, σκοπός της παρούσας διπλωματικής είναι πρόβλεψη της σχετικής θέσης των μετοχών σε μία κλίμακα από 1 έως και 5 σύμφωνα με τις ποσοστιαίες αποδόσεις τους. Για παράδειγμα, η κλάση 1 αφορά τα επενδυτικά στοιχεία με τις χαμηλότερες προβλεπόμενες ποσοστιαίες αποδόσεις ενώ η κλάση 5 περιλαμβάνει τις πιο κερδοφόρες μετοχές. Στο πρώτο μέρος της μελέτης, εφαρμόζονται

τόσο στατιστικές όσο και τεχνικές μηχανικής μάθησης για βραχυπρόθεσμο ορίζοντα πρόβλεψης μιας περιόδου δηλαδή οι κυλιόμενες προβλέψεις που πραγματοποιούνται αφορούν εκτιμήσεις κάθε φορά για την ακόλουθη μέρα. Στη συνέχεια, εφαρμόζονται οι ίδιες μεθοδολογίες μεταβάλλοντας αυτή τη φορά τον ορίζοντα πρόβλεψης από μία μέρα σε ένα μήνα. Τέλος, παρατηρώντας και αξιολογώντας τα προβλεπόμενα αποτελέσματα, γίνεται επιλογή των βέλτιστων μεθόδων και προσαρμογή τους στις συνθήκες του m6 διαγωνισμού για την παραγωγή αποτελεσμάτων για το πρώτο και δεύτερο τρίμηνο υποβολών ώστε να υπάρχει μια συνολική συγκριτική εικόνα.

3.3.2 Καθορισμός προβλήματος

Σε κάθε μεθοδολογία που χρησιμοποιείται, παράγονται προβλέψεις αναφορικά με την κατάταξη στην οποία ανήκουν οι μετοχές (1 έως 5). Συγκεκριμένα παράγεται ένας πίνακας πέντε θέσεων, κάθε θέση να αντιστοιχεί σε μία από τις κλάσεις κατάταξης, και σε κάθε θέση/κλάση δίνεται η πιθανότητα η μετοχή να ανήκει σε αυτή. Υπάρχει η παραδοχή σύμφωνα και με τη θεωρία πιθανοτήτων ότι το άθροισμα των βαρών αυτών πρέπει να αθροίζει σε ένα. Για παράδειγμα ο πίνακας [0.1 , 0.3 , 0.5 , 0.1 , 0] δηλώνει ότι προβλέπεται ότι η μετοχή μπορεί να ανήκει στην τάξη 1 με πιθανότητα 0.1 ,στην τάξη 2 με πιθανότητα 0.3, στην τάξη 3 με πιθανότητα 0.5, στην τάξη 4 με πιθανότητα 0.1 και δεν προβλέπεται ότι υπάρχει περίπτωση να ανήκει στην τάξη 5. Σε κάθε περίπτωση δημιουργείται με χρήση της βιβλιοθήκης pandas της python ένα dataframe με γραμμές τις ημερομηνίες που πραγματοποιούνται οι προβλέψεις και στήλες το σύνολο των διαθέσιμων μετοχών που χρησιμοποιήθηκαν με κάθε παρατήρηση να είναι ένας τέτοιος πίνακας πέντε θέσεων που αντιστοιχεί στην πρόβλεψη που πραγματοποιήθηκε εκείνη τη μέρα για τη συγκεκριμένη μετοχή.

date	ABBV	AEP	AIZ	ALLE	AMAT	AMP	AMZN
2022-01-04	[0.12, 0.19, 0.07, 0.06, 0.56]	[0.26, 0.21, 0.14, 0.3, 0.09]	[0.03, 0.33, 0.18, 0.2, 0.26]	[0.08, 0.26, 0.26, 0.31, 0.09]	[0.29, 0.26, 0.05, 0.22, 0.18]	[0.08, 0.33, 0.24, 0.2, 0.15]	[0.32, 0.11, 0.09, 0.17, 0.31]
2022-01-05	[0.19, 0.2, 0.24, 0.15, 0.22]	[0.23, 0.25, 0.18, 0.24, 0.1]	[0.08, 0.29, 0.22, 0.2, 0.21]	[0.3, 0.32, 0.25, 0.07, 0.06]	[0.38, 0.1, 0.03, 0.19, 0.3]	[0.32, 0.09, 0.19, 0.25, 0.15]	[0.38, 0.1, 0.17, 0.12, 0.23]
2022-01-06	[0.15, 0.39, 0.1, 0.12, 0.24]	[0.35, 0.2, 0.04, 0.18, 0.23]	[0.04, 0.18, 0.5, 0.18, 0.1]	[0.1, 0.15, 0.18, 0.45, 0.12]	[0.21, 0.06, 0.1, 0.03, 0.6]	[0.11, 0.22, 0.11, 0.52, 0.04]	[0.23, 0.23, 0.16, 0.26, 0.12]
2022-01-07	[0.11, 0.17, 0.24, 0.19, 0.29]	[0.29, 0.4, 0.1, 0.17, 0.04]	[0.09, 0.23, 0.38, 0.21, 0.09]	[0.25, 0.19, 0.09, 0.28, 0.19]	[0.29, 0.21, 0.09, 0.15, 0.26]	[0.15, 0.18, 0.14, 0.49, 0.04]	[0.22, 0.28, 0.07, 0.26, 0.17]

Εικόνα 42: Παράδειγμα πίνακα προβλέψεων (Random Forest)

Το παραπάνω dataframe με τις προβλέψεις αξιολογείται με βάση τη μετρική RPS (Ranked Probability Score) σύμφωνα και με τις παραδοχές του διαγωνισμού m6. Ο υπολογισμός της μετρικής για το i επενδυτικό αγαθό την T περίοδο είναι σχετικά απλός και δίνεται από τον παρακάτω τύπο :

$$\mathbf{RPS_{i,T}} = \frac{1}{5}\sum_{j=1}^{5}(\sum_{k=1}^{j} pi,T,k - \sum_{k=1}^{j} ei,T,k)^2$$

όπου p = ο πίνακας με τις προβλεπόμενες πιθανότητες

και e = η πραγματική κατάταξη

Προκειμένου να γίνει σωστά η αξιολόγηση, η πραγματική κατάταξη παίρνει και αυτή τη μορφή ενός πίνακα πέντε θέσεων. Για παράδειγμα, εάν η μετοχή την προβλεπόμενη μέρα έχει τελικά κατάταξη 3, αυτόματα δημιουργείται ένας πίνακας πέντε θέσεων [0 , 0 , 1 , 0 , 0] όπου όπως φαίνεται, η μετοχή βρίσκεται με 100 % πιθανότητα στη θέση 3. Τόσο με τα πραγματικά αυτά δεδομένα, όσο και με τις προβλέψεις τροφοδοτείται η συνάρτηση που χρησιμοποιείται από το διαγωνισμό για τον υπολογισμό της μετρικής RPS και δημιουργείται ένα dataframe που περιλαμβάνει τις τιμές RPS για κάθε μετοχή ανά ημέρα.

	ABBV	AEP	AIZ	ALLE	AMAT	AMP	AMZN
date							
2022-01-04	0.166287	0.193668	0.071257	0.184080	0.343066	0.064911	0.075843
2022-01-05	0.197988	0.170089	0.180125	0.356346	0.165196	0.064689	0.168639
2022-01-06	0.079820	0.192892	0.189395	0.182239	0.342773	0.064469	0.075487
2022-01-07	0.165921	0.069464	0.357728	0.181685	0.185835	0.080436	0.168247

Εικόνα 43: Παράδειγμα ημερήσιων τιμών RPS (SES)

Για παράδειγμα, εάν η πρόβλεψη τη μία μέρα ήταν [0.1 , 0.15 , 0.3 , 0.25 , 0.2] και την προβλεπόμενη μέρα τελικά η κατάταξη προέκυψε ότι ήταν [0 , 1 , 0 , 0 , 0] τότε η υπολογιζόμενη τιμή RPS προκύπτει ότι είναι :

$$= \frac{(0-0.1)^2 + (1-0.25)^2 + (1-0.55)^2 + (1-0.8)^2 + (1-1)^2}{5}$$

$$= \frac{0.01+0.5625+0.2025+0.04+0}{5}$$

$$= \frac{0.815}{5} = 0.163$$

Για τις τιμές RPS που προκύπτουν υπολογίζεται ο μέσος όρος τους ανά ημέρα κι έπειτα ο μέσος όρος όλων των ημερών έτσι ώστε να υπάρχει μια γενική εικόνα για την αξιολόγηση των μοντέλων. Η τιμή αυτή στη συνέχεια συγκρίνεται με τη benchmark αξία 0.16 η οποία προκύπτει εάν όλες οι θέσεις ήταν ισόβαρες δηλαδή εάν ο πίνακας προβλέψεων ήταν [0.2 , 0.2 , 0.2 , 0.2 , 0.2]. Στην ουσία,η μετρική RPS κάθε μεθόδου διαιρείται με την benchmark τιμή 0.16 και παρατηρείται συγκριτικά εάν είναι καλύτερη (κάτω από τη μονάδα) ή χειρότερη (πάνω από τη μονάδα). Για παράδειγμα, εάν κατά την αξιολόγηση ενός μοντέλου προκύπτει RPS 0.17 τότε συγκριτικά με τη benchmark τιμή θα είναι 0.17/0.16 = 1.0625 δηλαδή θα παράγει λίγο χειρότερα αποτελέσματα ενώ σε μια άλλη περίπτωση με RPS 0.15 θα προέκυπτε 0.15/0.16 = 0.9375 όπου τα

αποτελέσματα θα ήταν λίγο καλύτερα από μια τυχαία πρόβλεψη. Γενικά, επιθυμητές είναι χαμηλότερες τιμές από το benchmark υποδηλώνοντας καλύτερη μέθοδο πρόβλεψης.

Στην πλειοψηφία, η πρόβλεψη κάθε φορά πραγματοποιείται στα ranks δηλαδή τις ποσοστιαίες αποδόσεις των προσαρμοσμένων τιμών κλεισίματος κάθε μετοχής. Για το λόγο αυτό μόλις αντληθούν οι τιμές αυτές κλεισίματος υπολογίζονται αυτόματα και οι κλάσεις τους ώστε να υπάρχουν διαθέσιμα τόσο για εκπαίδευση των μοντέλων όσο και για επαλήθευση των αποτελεσμάτων. Η διαδικασία που ακολουθείται περιλαμβάνει τον υπολογισμό της ποσοστιαίας ημερήσιας μεταβολής της τιμής ανά ημέρα για κάθε μετοχή κι έπειτα την ταξινόμηση των αποτελεσμάτων κατά φθίνουσα σειρά. Από αυτά, κάθε μέρα οι αριθμοί ταξινόμησης από 1 έως 9 αντιστοιχούν στο rank 1, οι αριθμοί 10-18 στο rank 2 κοκ.

Date	ABBV	AEP	AIZ	ALLE	AMAT	AMP	AMZN
2021-01-04	0.010341	-0.001349	0.000303	0.001922	0.031426	-0.005848	0.010004
2021-01-05	-0.008638	0.007982	0.043889	0.022755	0.013728	0.057377	-0.024897
2021-01-06	0.010703	-0.026681	0.010076	0.026341	0.041066	0.012087	0.007577
2021-01-07	0.005248	-0.008887	-0.005239	-0.003488	0.010575	0.007095	0.006496

Εικόνα 44: 1) Υπολογισμός ποσοστιαίων μεταβολών τιμών ανά ημέρα

Date	ABBV	AEP	AIZ	ALLE	AMAT	AMP	AMZN
2021-01-04	34	14	15	21	41	9	33
2021-01-05	8	14	37	25	20	40	3
2021-01-06	25	2	23	36	44	28	20
2021-01-07	28	8	13	17	37	31	30

Εικόνα 45: 2) Ταξινόμηση αποτελεσμάτων σε φθίνουσα σειρά

Date	ABBV	AEP	AIZ	ALLE	AMAT	AMP	AMZN
2021-01-04	4	2	2	3	5	1	4
2021-01-05	1	2	5	3	3	5	1
2021-01-06	3	1	3	4	5	4	3
2021-01-07	4	1	2	2	5	4	4

Εικόνα 46: 3) Αντιστοίχηση απόλυτης κατάταξης σε πέντε κλάσεις (ranks)

Σε όλες τις περιπτώσεις πραγματοποιείται μια κυλιόμενη ημερήσια πρόβλεψη για ένα διάστημα τριών μηνών από τον Ιανουάριο μέχρι και το Μάρτιο του 2022 δηλαδή αυτό είναι το διάστημα κατά το οποίο πραγματοποιούνται οι προβλέψεις και είναι κοινό για όλες τις μεθόδους. Παρ' όλα αυτά ο ορίζοντας πρόβλεψης, το πλήθος των δεδομένων για εκπαίδευση και η μεθοδολογία τροποποιείται ανάλογα με τη μέθοδο και τον σκοπό της πρόβλεψης (βραχυπρόθεσμο ή μεσοπρόθεσμο) όπως θα εξηγηθεί αναλυτικά και στη συνέχεια.

3.3.3 Ολοκληρωμένες μεθοδολογίες με ορίζοντα πρόβλεψης την ακόλουθη μέρα

Σε πρώτη φάση εφαρμόζονται τόσο στατιστικές όσο και τεχνικές μηχανικής μάθησης για βραχυπρόθεσμο ορίζοντα πρόβλεψης μιας περιόδου, δηλαδή γίνεται η παραδοχή ότι κάθε μέρα χρησιμοποιούνται δεδομένα όλων των προηγούμενων ημερών για παραγωγή πρόβλεψης για την ακόλουθη ημέρα. Η επιλογή του ορίζοντα αυτού βασίστηκε στο πλήθος των διαθέσιμων ιστορικών δεδομένων που αντλήθηκαν καθώς και στα χαρακτηριστικά που τυπικά υπολογίζονται για βραχυπρόθεσμες και μεσοπρόθεσμες περιόδους.

Στατιστικές Μέθοδοι

1. Αφελής μέθοδος στις κλάσεις κατάταξης

Πρόκειται για την πιο απλή στατιστική μέθοδο πρόβλεψης καθώς χρησιμοποιεί ως πρόβλεψη την τελευταία ιστορική παρατήρηση. Έτσι, δεν έχει κάποια ιδιαίτερη πολυπλοκότητα, έχοντας παράλληλα ως πλεονέκτημα το γεγονός ότι οι προβλέψεις της εκφράζονται άμεσα στις ζητούμενες κλάσεις. Η μέθοδος υλοποιείται από τον παρακάτω τύπο :

$$\mathbf{Ft = Y_{t-1}}$$

Όπου Ft = η πρόβλεψη

Y_{t-1} = η παρατήρηση

Σε πρώτη φάση δοκιμάστηκε η μέθοδος κυλιόμενης πρόβλεψης χρησιμοποιώντας την τελευταία πιο πρόσφατη παρατήρηση, με αποτέλεσμα οι προβλέψεις να αντικατοπτρίζουν την πραγματική χρονοσειρά με καθυστέρηση μίας παρατήρησης. Χρησιμοποιήθηκαν δεδομένα από τον Ιανουάριο του 2021 μέχρι και το Μάρτιο 2022. Ξεκινώντας από τις αρχές Ιανουαρίου του 2022 για ένα διάστημα τριών μηνών κάθε μέρα προβλεπόταν η κλάση της επόμενης μέρας χρησιμοποιώντας την κλάση της προηγούμενης. Πέρα όμως από την τυπική λειτουργία δοκιμάστηκαν και κάποιες εναλλαγές της μεθόδου ως προς την τιμή που προβλεπόταν. Συνολικά δοκιμάστηκαν τρεις παραλλαγές της Naive :

- Πρόβλεψη της κλάσης της επόμενης μέρας με την κλάση της προηγούμενης (κλασική μέθοδος)

- Πρόβλεψη της κλάσης της επόμενης μέρας με την κλάση του προηγούμενου μήνα.
- Πρόβλεψη της κλάσης της επόμενης μέρας υπολογίζοντας τον κυλιόμενο μέσο όρο των τιμών προσαρμοσμένου κλεισίματος των προηγούμενων τριών ημερών.

Η τελευταία περίπτωση διαφέρει από τις προηγούμενες καθώς πραγματοποιεί πρόβλεψη στα ranks χρησιμοποιώντας τιμές προσαρμοσμένου κλεισίματος. Προκειμένου οι προβλεπόμενοι μέσοι όροι ανά ημέρα για κάθε μετοχή να μεταφραστούν σε ranks , ταξινομούνται οι τιμές αυτές για όλες τις μετοχές ανά ημέρα πρόβλεψης σε φθίνουσα σειρά κι έπειτα αντιστοιχίζεται η απόλυτη κατάταξή τους στις 5 κλάσεις όπως εξηγήθηκε και στο προηγούμενο υποκεφάλαιο.

Τα αποτελέσματα αυτά στη συνέχεια συγκρίνονται με τα πραγματικά δεδομένα και αξιολογούνται με τη μετρική RPS που προαναφέρθηκε όπως φαίνεται και στον παρακάτω πίνακα:

	RPS	*Comparison with benchmark*
Naïve 1	0.3178	1.9861
Naïve 2	0.3095	1.9343
Naïve 3	0.3098	1.9363

Πίνακας 8: Αποτελέσματα RPS για τη μέθοδο Naive

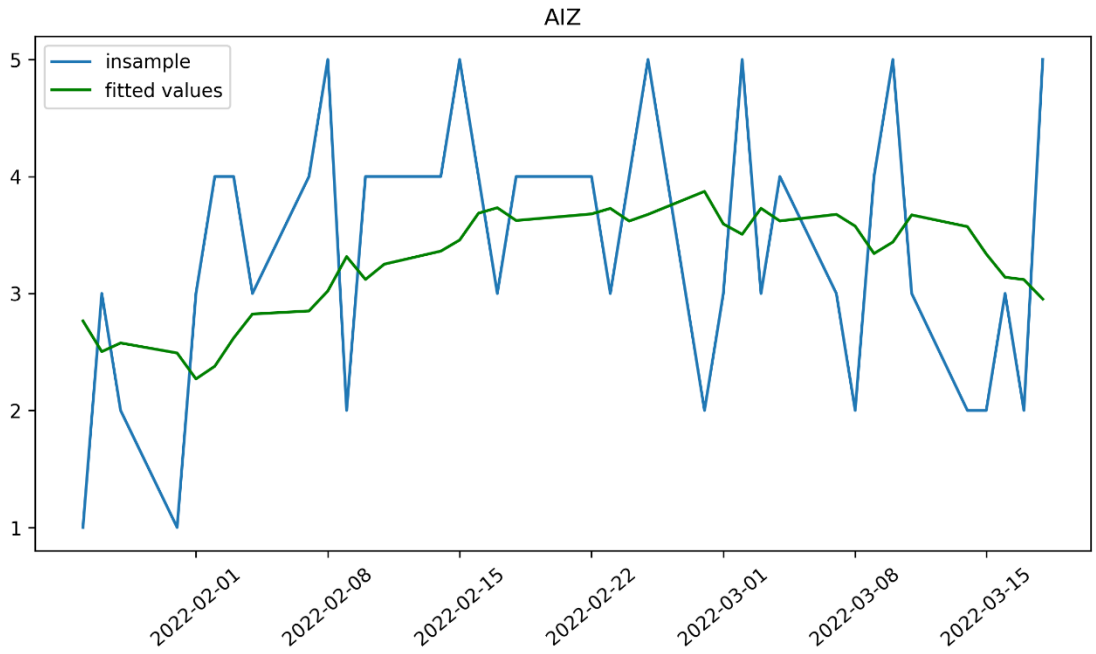
2. Απλή εκθετική εξομάλυνση στις κλάσεις κατάταξης

Αυτή η μέθοδος (SES on ranks) είναι κατάλληλη για την πρόβλεψη δεδομένων υψηλής διακύμανσης που δεν παρουσιάζουν κάποιο μοτίβο εποχιακότητας ή τάσης. Για την υλοποίηση της χρησιμοποιήθηκε η βιβλιοθήκη statsmodels της python η οποία προσφέρει τη δυνατότητα είτε να επιλεγεί συγκεκριμένη τιμή για την παράμετρο εξομάλυνσης α, είτε αυτοματοποιημένα να υπολογίσει τη βέλτιστη τιμή ανάλογα με τη χρονοσειρά.

Στην περίπτωση αυτή οι προβλέψεις πραγματοποιούνται για κάθε μετοχή ξεχωριστά, δηλαδή ανά χρονοσειρά έτσι ώστε η κάθε μετοχή να διαθέτει τη βέλτιστη παράμετρο εξομάλυνσης. Κάθε μέρα ξεκινώντας από τις αρχές Ιανουαρίου του 2022 ως δεδομένα εκπαίδευσης χρησιμοποιούνται όλες οι προηγούμενες παρατηρήσεις δηλαδή οι κατατάξεις που είχε η μετοχή μέχρι και την προηγούμενη ημέρα της πρόβλεψης. Παράγεται κυλιόμενη ημερήσια πρόβλεψη με ορίζοντα πρόβλεψης πάντα την ακόλουθη μέρα με τη βοήθεια της εντολής forecast(1) της βιβλιοθήκης statsmodels που χρησιμοποιείται για την υλοποίηση της μεθόδου.

Καθώς για την αξιολόγηση χρειάζεται ένας πίνακας πιθανοτήτων πέντε θέσεων όπου σε κάθε θέση αναφέρεται ποια είναι η πιθανότητα να ανήκει η μετοχή στην αντίστοιχη κλάση, δε χρησιμοποιείται η απόλυτη πρόβλεψη όπως δίνεται. Υπολογίζονται οι προσαρμοσμένες προβλεπόμενες τιμές και στη συνέχεια το σφάλμα των

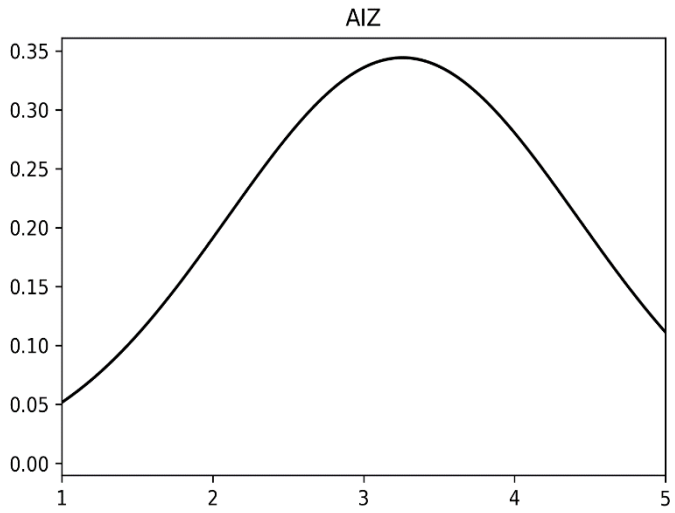
προβλεπόμενων αυτών τιμών με τις πραγματικές κατατάξεις. Η μέση τιμή και η τυπική απόκλιση αυτών των σφαλμάτων χρησιμοποιούνται μαζί με την απόλυτη πρόβλεψη (forecast) και σχεδιάζεται μια κανονική κατανομή με τη βοήθεια της εντολής norm της βιβλιοθήκης scipy από την οποία και προκύπτουν οι πιθανότητες των 5 κλάσεων που χρειάζονται για την αξιολόγηση των μοντέλων.



Εικόνα 47: 1) Μελέτη πραγματικών (insample) και προσαρμοσμένων τιμών (fitted values)

-1.57529887	0.7733522	-0.39780901	-0.30976442	0.75879381	1.5908547
1.23876073	-0.0354064	0.97242983	1.75720799	-1.63170392	1.72943095
0.34666677	1.26994117	0.988873	-0.22998811	-1.17908627	0.08187355
0.06375299	-0.9503571	0.25997955	1.20243989	-2.06368864	-0.60694528 ...

Πίνακας 9: 2) Υπολογισμός σφάλματος



Εικόνα 48: 3) Δημιουργία κανονικής κατανομής

Από το παραπάνω σχήμα, αναμένεται η μετοχή AIZ να ανήκει με μικρή πιθανότητα στην κλάση 1, λίγο μεγαλύτερη στην κλάση 2 και 5 καθώς και με τη μεγαλύτερη πιθανότητα να ανήκει στην κατάταξη 3 και 4. Πράγματι μετά τους υπολογισμούς προκύπτει:

AIZ	[0.0357, 0.1132, 0.2829, 0.3469, 0.1946]

Πίνακας 10: Υπολογισμός πιθανοτήτων

Τα αποτελέσματα αυτής της μεθόδου συγκρίνονται στη συνέχεια και με τα point forecasts, δηλαδή επιλέγεται κάθε φορά μόνο η κλάση με τη μεγαλύτερη πιθανότητα και παίρνει πιθανότητα 1 (100%) ενώ οι άλλες αυτομάτως μηδενίζονται, για να εξεταστεί ποια προσέγγιση είναι η πιο αποτελεσματική.

	RPS	*Comparison with benchmark*
SES in ranks	0.1714	1.0713
Comparison with point forecast	0.2581	1.6131

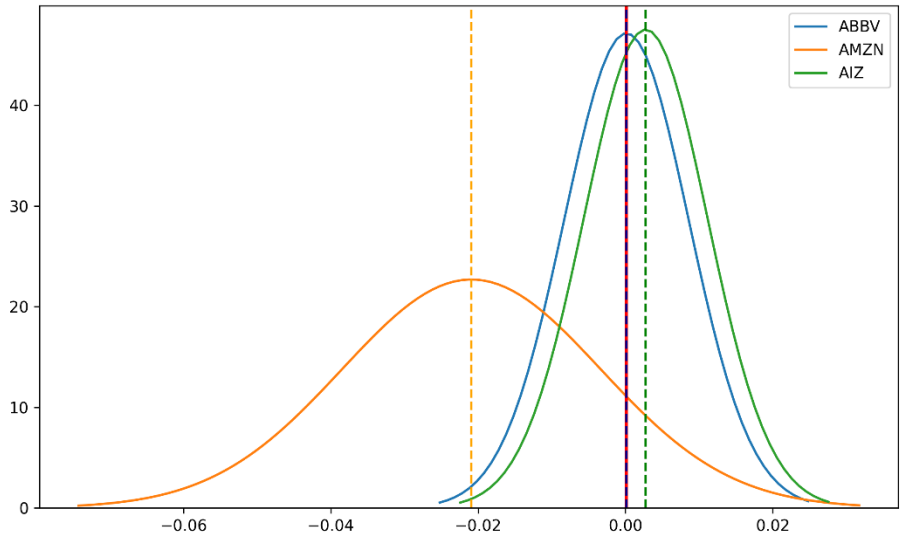
Πίνακας 11: Αποτελέσματα RPS για τη μέθοδο SES στις κατατάξεις

3. Απλή εκθετική εξομάλυνση στις ποσοστιαίες αποδόσεις

Σε αυτήν την προσέγγιση (SES on percentage returns) η πρόβλεψη πραγματοποιείται για κάθε μετοχή ξεχωριστά, δηλαδή ανά χρονοσειρά, αλλά αφορά την ποσοστιαία απόδοσή της αντί για τη θέση κατάταξης. Προκειμένου να υπολογιστεί ο πίνακας πιθανοτήτων, για κάθε μετοχή σχεδιάζεται μια κανονική κατανομή με τον ίδιο τρόπο που επεξηγήθηκε και προηγουμένως. Πιο συγκεκριμένα, υπολογίζεται η διαφορά των προσαρμοσμένων τιμών του μοντέλου και των πραγματικών ποσοστιαίων αποδόσεων και χρησιμοποιώντας τη μέση τιμή και την τυπική απόκλιση των αποτελεσμάτων, μαζί

με την πρόβλεψη του μοντέλου σχεδιάζεται ανά ημέρα και για κάθε μετοχή η κανονική κατανομή του.

Στη συνέχεια, ανά ημέρα σχεδιάζονται όλες οι κανονικές κατανομές στο ίδιο σχήμα. Οι μέσες τιμές τους αποθηκεύονται με σκοπό το διάστημα αυτό να χωριστεί σε 5 σημεία. Αφού χωριστεί το σχήμα σε διαστήματα, ελέγχεται κάθε περίπτωση ανάμεσα σε ποια περιοχή (κλάση) ανήκει και με πόση πιθανότητα. Παρακάτω επεξηγείται η μέθοδος χρησιμοποιώντας μια απλή εκδοχή 3 μετοχών οι οποίες χωρίζονται σε δύο περιοχές /κλάσεις, 1 και 2.



Εικόνα 49: Μέθοδος για 3 μετοχές με χωρισμό σε 2 περιοχές

Για τις τρεις αυτές μετοχές εκείνη τη μέρα οι μέσες τιμές που προέκυψαν από την κανονική κατανομή ήταν:

ABBV	0.00013584443527257106
AMZN	-0.020920142232155093
AIZ	0.0027460982459838797

Πίνακας 12: Μέσες τιμές κανονικής κατανομής μετοχών παραδείγματος

Αυτές οι τιμές ταξινομούνται και προκειμένου να χωριστεί το σχήμα σε δύο περιοχές επιλέγεται η μεσαία τιμή έτσι ώστε να διεξαχθεί μια δίκαιη κατανομή.

Πίνακας με τις ταξινομημένες τιμές :

[-0.020920142232155093 , 0.00013584443527257106 , 0.0027460982459838797]

Η τιμή που επιλέχθηκε από τον πίνακα φαίνεται και στο σχήμα με κόκκινο χρώμα και αντιστοιχεί στη μέση τιμή της κανονικής κατανομής της μετοχής ABBV χωρίζοντας έτσι το σχήμα στη μέση. Για αυτό κι όλας η μετοχή ABBV αναμένεται να ανήκει κατά το ήμισυ στην κατάταξη 1 και 2 εξίσου. Η μετοχή AMZN φαίνεται να ανήκει περισσότερο στην κατάταξη 1 και λιγότερο στη 2 καθώς και η μετοχή AIZ αναμένεται να ανήκει λίγο περισσότερο στην κατάταξη 2 όπως φαίνεται και από το σχήμα. Πράγματι μετά από υπολογισμούς προκύπτουν οι πιθανότητες:

ABBV	[0.5 , 0.5]
AMZN	[0.8845, 0.1155]
AIZ	[0.3779 , 0.6221]

Πίνακας 13: Υπολογισμός πιθανοτήτων ανά περιοχή παραδείγματος

Με την ίδια λογική πραγματοποιούνται οι προβλέψεις για όλες τις μετοχές ανά ημέρα χωρίζοντας την περιοχή σε 5 διαστήματα και προκύπτουν ορισμένα αποτελέσματα τα οποία στη συνέχεια συγκρίνονται και με το point forecast εκδοχή της μεθόδου για να εξεταστεί ποια προσέγγιση είναι η πιο αποτελεσματική.

	RPS	*Comparison with benchmark*
SES in ranks	0.1946	1.2163
Comparison with point forecast	0.3903	2.4394

Πίνακας 14: Αποτελέσματα RPS για τη μέθοδο SES στις ποσοστιαίες αποδόσεις

4. Συχνότητα εμφάνισης κλάσης

Η Συχνότητα εμφάνισης κλάσης (Frequency of ranks) πρόκειται για μία προσέγγιση κατά την οποία κάθε μέρα προβλέπεται η κατάταξη της επόμενης μέρας επιλέγοντας την πιο «δημοφιλή» κλάση που είχε η μετοχή από τις προηγούμενες παρατηρήσεις της. Πιο συγκεκριμένα, κάθε μέρα και για κάθε διαθέσιμη μετοχή με τη βοήθεια ενός μετρητή (defaultdict) από τη βιβλιοθήκη collections της python, μετριέται η συχνότητα εμφάνισης κάθε κλάσης και στη συνέχεια συγκρίνεται με το σύνολο των παρατηρήσεων μέχρι και την ημέρα της πρόβλεψης, δηλαδή υπολογίζεται ο μέσος όρος εμφάνισης κάθε κλάσης για μία μετοχή. Με αυτόν τον τρόπο διαμορφώνεται και ο πίνακας των πιθανοτικών προβλέψεων που χρησιμοποιείται για την αξιολόγηση του μοντέλου. Ενδεικτικά, τα βήματα επεξηγούνται αναλυτικά για τη μετοχή ABBV για την ημέρα 4/01/2022.

- Βήμα 1: Μέτρηση συχνότητας εμφάνισης κλάσης για τη μετοχή ABBV
  Η ημέρα που πραγματοποιείται η πρόβλεψη είναι η 4/01/2022.
  Το insample ξεκινάει από την 4/01/2021 μέχρι και την ημέρα 3/01/2022.
  Μετρώντας τη συχνότητα εμφάνισης κάθε κλάσης προκύπτει ο πίνακας :

[1: 42, 2: 45, 3: 58, 4: 61, 5: 47]

Παρατηρείται ότι η κλάση 1 εμφανίστηκε 42 φορές, η κλάση 2 εμφανίστηκε 45 φορές και συνεχίζοντας είναι εμφανές ότι η πιο δημοφιλής κλάση για τη μετοχή ABBV μέχρι και εκείνη τη μέρα ήταν η 4.

- Βήμα 2: Υπολογισμός μέσων όρων που χρησιμοποιούνται ως οι πιθανότητες πρόβλεψης. Για παράδειγμα, η πιθανότητα να ανήκει στην κατάταξη 1 προκύπτει διαιρώντας τον αριθμό εμφανίσεων με το συνολικό αριθμό παρατηρήσεων μέχρι και εκείνη την ημέρα δηλαδή 42 / 253 = 0.1660079 ≈ 0.166008. Με αντίστοιχο τρόπο υπολογίζονται και οι υπόλοιπες πιθανότητες και προκύπτει ο τελικός πίνακας:

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Prediction	0.166008	0.177866	0.229249	0.241107	0.185771

Πίνακας 15: Υπολογισμός πιθανοτήτων με βάση τη μεθοδολογία

Στη συνέχεια η διαδικασία συνεχίζεται κατά τα γνωστά και ο μέσος όρος των τιμών RPS που προέκυψαν συγκρίνεται με την point forecast εκδοχή της μεθόδου. Ενδεικτικά, ο πίνακας στη σύγκριση με το point forecast θα έπαιρνε τιμή μόνο στη θέση / rank με τη μεγαλύτερη πιθανότητα δηλαδή θα είχε τη μορφή :

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Prediction	0	0	0	1	0

Πίνακας 16: Ενδεικτικός πίνακας πιθανοτήτων για την point forecast πρόβλεψη

Τα τελικά αποτελέσματα της μεθόδου φαίνονται και στον παρακάτω πίνακα:

	RPS	*Comparison with benchmark*
Frequency of ranks	0.1587	0.9919
Comparison with point forecast	0.2878	1.7988

Πίνακας 17: Αποτελέσματα RPS για τη μέθοδο Frequency of ranks

Εξετάστηκαν με τον ίδιο τρόπο διαφορετικοί περίοδοι εκπαίδευσης έτσι ώστε να διαπιστωθεί από ποια περίοδο και μετά παράγονται καλύτερα αποτελέσματα. Για παράδειγμα, κατά την επιλογή περιόδου insample 3 μηνών, κάθε ημέρα πρόβλεψης χρησιμοποιούνται για υπολογισμό μόνο οι παρατηρήσεις των προηγούμενων 3 μηνών αντί για τις συνολικές παρατηρήσεις από τις αρχές του 2021, επηρεάζοντας έτσι τη συχνότητα εμφάνισης και κατ' επέκταση και τον υπολογισμό πιθανοτήτων. Στον παρακάτω πίνακα φαίνονται οι μέσοι όροι των μετρικών για κάθε διαφορετική περίοδο:

	10 days	*15 days*	*1 month*	*2 months*	*3 months*	*6 months*	*10 months*	*1 year*
Frequency of ranks	0.1743	0.1691	0.1652	0.1613	0.1608	0.1590	0.1588	0.1586
Comparison with benchmark	1.0894	1.0569	1.0325	1.0081	1.005	0.9938	0.9925	0.9913

Πίνακας 18: Αποτελέσματα RPS για τη μέθοδο Frequency of ranks για διαφορετικές περιόδους εκπαίδευσης

Τεχνικές μηχανικής μάθησης

Εφόσον σκοπός είναι η πρόβλεψη της πιθανότητας να ανήκει ένα επενδυτικό στις πέντε κλάσεις, είναι απαραίτητη η δημιουργία multiclass classification μοντέλων των οποίων οι ατομικές πιθανότητες τους πρέπει να αθροίζουν στο 1, καθώς αναγκαστικά κάθε μετοχή πρέπει να ανήκει σε μία κλάση και η πιο πιθανή θα προβλεφθεί από τα μοντέλα.

Η βασική μεθοδολογία που ακολουθείται στην πλειοψηφία των περιπτώσεων είναι η ακόλουθη:

- Εισαγωγή βιβλιοθήκης
- Εισαγωγή δεδομένων
- Δημιουργία τεχνικών δεικτών και χαρακτηριστικών
- Δημιουργία δείκτη πρόβλεψης
- Χωρισμός του συνόλου δεδομένων σε σύνολο εκπαίδευσης (train set) και αξιολόγησης (test set)
- Προσαρμογή των δεδομένων στο μοντέλο (fitting)
- Πρόβλεψη των πιθανοτήτων κάθε κλάσης
- Αξιολόγηση του μοντέλου με τη μετρική RPS
- Υπολογισμός μέσου όρου αποτελεσμάτων

Ανάλογα το μοντέλο χρησιμοποιούνται και διαδικασίες επιλογής των διαθέσιμων χαρακτηριστικών (feature selection), αντιμετώπισης προβλημάτων overfitting, καθώς και προσαρμογής των υπερπαραμέτρων (tuning hyperparameters) για παραγωγή καλύτερων αποτελεσμάτων.

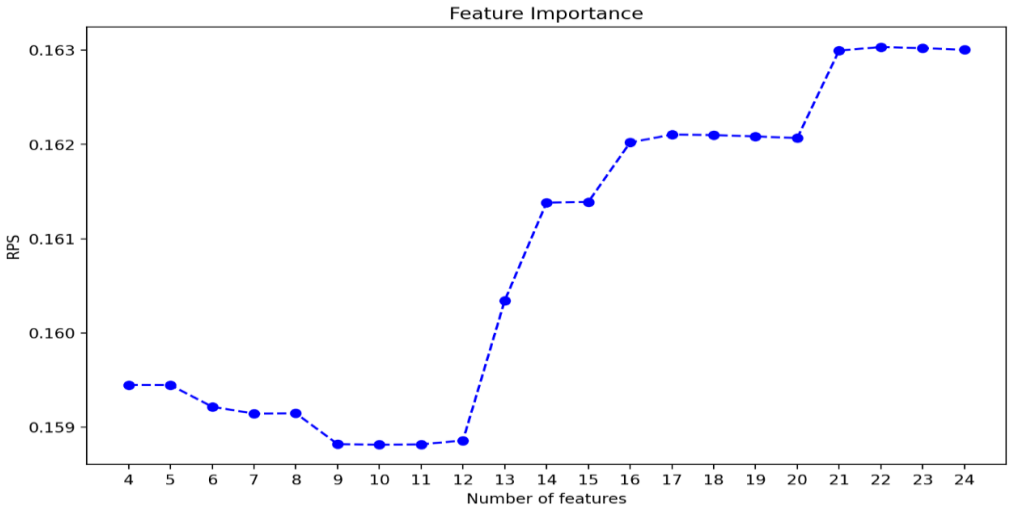
5. Λογιστική Παλινδρόμηση

Για τη μέθοδο αυτή, χρησιμοποιήθηκαν ιστορικά δεδομένα από αρχές του 2021 μέχρι και το Μάρτιο του 2022. Σύμφωνα με αυτά υπολογίστηκαν οι απαραίτητοι τεχνικοί δείκτες και στη συνέχεια χωρίστηκε το σύνολό τους στα δεδομένα που τροφοδοτούν το μοντέλο (X) και στις ετικέτες με τις πραγματικές κατατάξεις (y). Επειδή την ημέρα που πραγματοποιείται η πρόβλεψη γίνεται η παραδοχή ότι δεν μπορεί να είναι γνωστά κάποια χαρακτηριστικά όπως για παράδειγμα η τιμή κλεισίματος, χρησιμοποιούνται κάθε φορά οι παρατηρήσεις τις προηγούμενης ημέρας καθώς και επειδή σκοπός του μοντέλου είναι η πρόβλεψη της κλάσης για την ακόλουθη ημέρα, στις ετικέτες που χρησιμοποιούνται για εκπαίδευση δίνονται οι πραγματικές κλάσεις της επόμενης

ημέρας. Ως δεδομένα εκπαίδευσης επιλέχθηκαν τα δεδομένα του έτους 2021 , ένα σύνολο δηλαδή 253 παρατηρήσεων πάνω στα οποία εκπαιδεύτηκε το μοντέλο μία φορά κι έπειτα για κάθε νέα γραμμή δεδομένων από τον Ιανουάριο του 2022 μέχρι και τον Μάρτιο της ίδια χρονιάς παράγεται και μία καινούρια πρόβλεψη των πιθανοτήτων της κάθε κλάσης με τη βοήθεια της εντολής predict_proba που προσφέρει η μέθοδος logistic regression της scikit-learn.

Η βιβλιοθήκη αυτή της python προσφέρει τη δυνατότητα εξέτασης του μοντέλου και με τη χρήση διαφορετικών solvers πέραν της προκαθορισμένης επιλογής *lbfgs*. Βάση αποτελεσμάτων προτιμήθηκε και η προσέγγιση *newton-cg* κι εξετάστηκε παράλληλα με την προκαθορισμένη επιλογή.

Στην πορεία, έγινε μια προσπάθεια βελτιστοποίησης του μοντέλου κάνοντας μια επιλογή από τα διαθέσιμα χαρακτηριστικά (feature selection). Καθώς ορισμένα από αυτά μπορεί να μην ήταν αντιπροσωπευτικά ή να μη συνδράμανε στην απόδοση του μοντέλου έγινε μια προσπάθεια διαλογής των πιο σημαντικών. Σε αντίθεση με τα υπόλοιπα μοντέλα που θα αναλυθούν στη συνέχεια, η μέθοδος αυτή δε διαθέτει ενσωματωμένη εντολή που να προσφέρει αυτήν την πληροφορία (feature importances) οπότε δημιουργήθηκε ένα Decision Tree Classifier από την scikit-learn το οποίο χρησιμοποιήθηκε ώστε να γίνει μια ανάλυση των πιο σημαντικών στοιχείων. Για κάθε στοιχείο που αποχωρούσε, ξαναγινόταν αξιολόγηση του μοντέλου ως προς τη μετρική RPS μέχρι να βρεθεί ο βέλτιστος αριθμός.



Εικόνα 50: Επιλογή βέλτιστου αριθμού χαρακτηριστικών σύμφωνα με το RPS

Από την παραπάνω γραφική, παρατηρείται ότι ο βέλτιστος αριθμός χαρακτηριστικών για την επίτευξη μιας χαμηλής τιμής RPS είναι γύρω στα 9 με 11 χαρακτηριστικά κι έπειτα από δοκιμές με βάση και τα feature importances που παράγονται από το Decision Tree Classifier προκύπτουν ότι παραμένουν τα παρακάτω στοιχεία :

- Ποσοστιαία μεταβολή της τιμής ανοίγματος
- Ποσοστιαία μεταβολή της μέγιστης τιμή
- Ποσοστιαία μεταβολή της ελάχιστης τιμή
- Ποσοστιαία μεταβολή της τιμής την προηγούμενη μέρα
- Ποσοστιαία μεταβολή των ακραίων τιμών (outliers) την προηγούμενη μέρα

- Κινητός μέσος όρος των προηγούμενων 3,6,10 και 14 ημερών
- Average Directional Index (ADX)

Έπειτα από εκπαίδευση του μοντέλου εκ νέου με τα τελικά χαρακτηριστικά προκύπτει ο παρακάτω συγκεντρωτικός πίνακας αποτελεσμάτων :

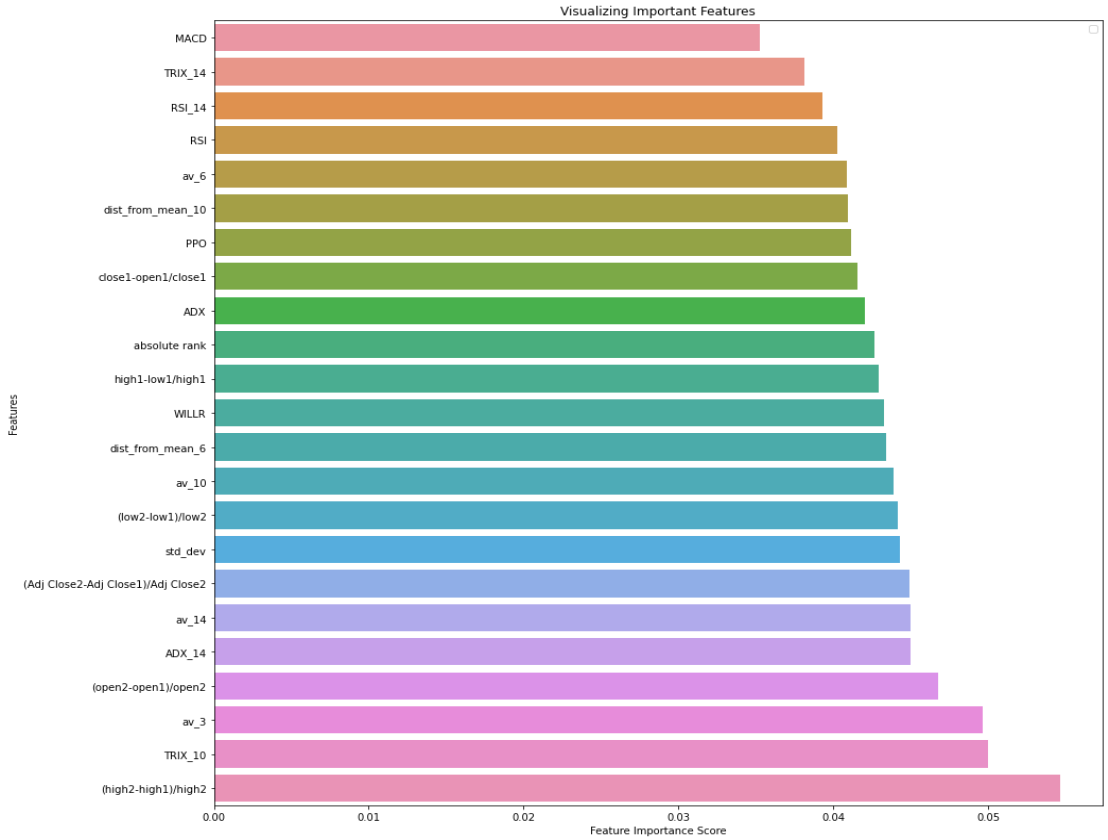
Logistic Regression (next day prediction)	*RPS*	*Comparison with benchmark*
Baseline	0.1709	1.068
After feature selection	0.1590	0.9938

Πίνακας 19: Αποτελέσματα RPS για τη μέθοδο Logistic Regression

6. Τυχαία Δάση

Η μέθοδος αυτή λειτουργεί με δέντρα αποφάσεων δημιουργώντας τυχαία «δάση» από υποσύνολα δεδομένων ανεξάρτητα το ένα από το άλλο και παίρνοντας το μέσο αποτέλεσμα τους, αντιμετωπίζοντας έτσι και το πρόβλημα overfitting. Με παρόμοιο τρόπο με την προηγούμενη μέθοδο έγινε η εκπαίδευση του μοντέλου, προσαρμόστηκαν τα χαρακτηριστικά και πραγματοποιήθηκαν προβλέψεις την ίδια χρονική περίοδο (Ιανουάριος 2022 – Μάρτιος 2022) με την ίδια εντολή predict_proba που διαθέτει και το random forest του scikit-learn.

Στη συνέχεια έγινε μια προσπάθεια βελτιστοποίησης του βασικού μοντέλου πραγματοποιώντας μια διαλογή των διαθέσιμων χαρακτηριστικών (feature selection). Η μέθοδος αυτή διαθέτει την ενσωματωμένη εντολή feature_importances η οποία δηλώνει για κάθε χαρακτηριστικό πόσο μεγάλη ήταν η συνεισφορά του στο τελικό αποτέλεσμα.



Εικόνα 51: Συνεισφορά χαρακτηριστικών στο μοντέλο Random Forest

Από αυτά, αφαιρούνται με τη σειρά τα χαρακτηριστικά με τη χαμηλότερη βαθμολογία και αξιολογείται το μοντέλο από την αρχή ως προς τη μετρική RPS. Στο τέλος, τα στοιχεία που αποχωρούν είναι τα πρώτα 6 που φαίνονται και στην εικόνα δηλαδή:

- Moving Average Convergence Divergence (MACD)
- Triple exponential average (TRIX) για περίοδο 14 ημερών
- Relative Strength Index (RSI) για περίοδο 6 και 14 ημερών
- Κινητός μέσος όρος των προηγούμενων 6 ημερών
- Απόσταση από τη μέση τιμή των προηγούμενων 10 ημερών

Ένας άλλος τρόπος που εφαρμόστηκε για βελτιστοποίηση του μοντέλου είναι η προσαρμογή ορισμένων υπερπαραμέτρων έτσι ώστε να ελεγχθεί αν το μοντέλο λειτουργεί καλύτερα χωρίς τις προκαθορισμένες τιμές. Η καλύτερη και πιο γρήγορη προσέγγιση είναι η αξιολόγηση ενός μεγάλου εύρους τιμών χρησιμοποιώντας τη μέθοδο RandomizedSearchCV του scikit-learn η οποία ελέγχει ένα εύρος τιμών για κάθε παράμετρο. Σε κάθε επανάληψη, επιλέγει ένα τυχαίο δείγμα από το πλέγμα, εκτελώντας K-Fold cross validation με κάθε συνδυασμό τιμών.

Οι πιο σημαντικές παράμετροι που δοκιμάστηκαν ήταν ο αριθμός των δέντρων στο δάσος (n_estimators), ο αριθμός των χαρακτηριστικών που λαμβάνονται υπόψη για διαχωρισμό σε κάθε κόμβο φύλλου (max_features) καθώς και εάν θα χρησιμοποιηθεί το σύνολο των δεδομένων για το φτιάξιμο κάθε δέντρου ή θα δημιουργηθούν δείγματα αυτού (bootstrap). Εφαρμόστηκαν δύο περιπτώσεις, η μία με 3-fold και η άλλη με 5-fold cross validation για να υπάρχει μια πλήρη εικόνα. Εναλλακτικά, θα μπορούσε να

χρησιμοποιηθεί και η μέθοδος Grid Search του scikit- learn , αλλά για εξοικονόμηση χρόνου και λόγω έλλειψης πόρων προτιμήθηκε η πρώτη επιλογή με εξίσου καλά αποτελέσματα.

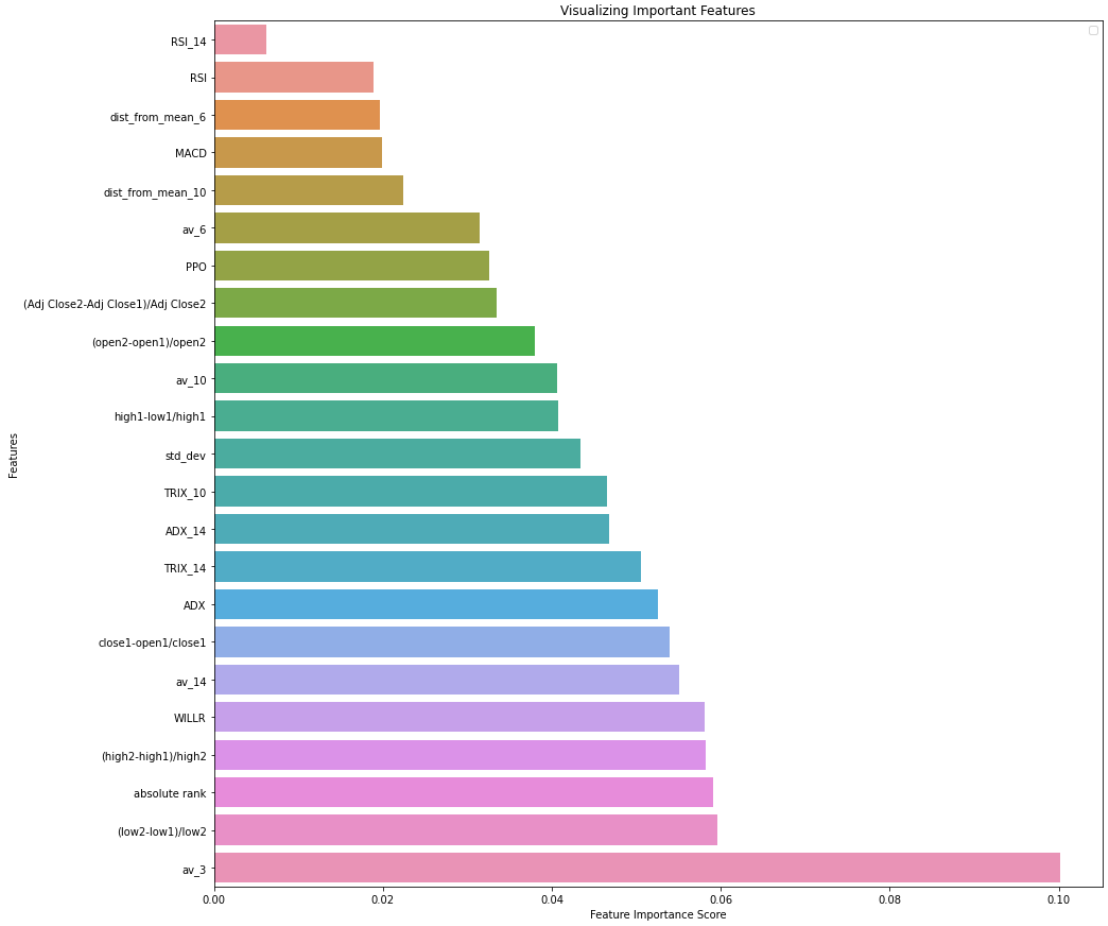
Τέλος, έγινε και μια τελική αξιολόγηση του μοντέλου αυτή τη φορά συνδυάζοντας τις δύο προηγούμενες βελτιστοποιήσεις και παράγοντας τα τελικά αποτελέσματα όπως φαίνονται και στον παρακάτω πίνακα:

Random Forest (next day prediction)	*RPS*	*Comparison with benchmark*
Baseline	0.1667	1.0419
Feature selection only	0.1665	1.0406
3-fold tuning only	0.1660	1.0375
5-fold tuning only	0.1656	1.035
Feature selection & 3-fold tuning	0.1674	1.0462
Feature selection & 5-fold tuning	0.1653	1.0331

Πίνακας 20: Αποτελέσματα RPS για τη μέθοδο Random Forest

7. Δέντρα Ενίσχυσης Κλίσης

Η μέθοδος αυτή συνδυάζει ένα σετ από δέντρα αποφάσεων και στόχος της κάθε φορά είναι να ενδυναμώσει ένα αδύναμο μοντέλο πρόβλεψης μέχρι να ελαχιστοποιηθεί η συνάρτηση κόστους δηλαδή το σφάλμα. Ακολουθούνται τα ίδια βήματα όπου αρχικά δημιουργείται ένα βασικό μοντέλο με όλα τα διαθέσιμα χαρακτηριστικά του και παράγονται προβλέψεις την περίοδο από τον Ιανουάριο του 2022 μέχρι και το Μάρτιο του 2022. Στη συνέχεια δοκιμάζεται μια μέθοδος βελτιστοποίησης κάνοντας διαλογή των πιο κατάλληλων χαρακτηριστικών (feature selection) χρησιμοποιώντας την ενσωματωμένη εντολή feature_importances και οπτικοποιώντας τα αποτελέσματα.



Εικόνα 52: Συνεισφορά χαρακτηριστικών στο μοντέλο Gradient Boosting

Από το παραπάνω, αφαιρούνται με τη σειρά τα χαρακτηριστικά με τη χαμηλότερη βαθμολογία και αξιολογείται το μοντέλο από την αρχή ως προς τη μετρική RPS. Στο τέλος, τα στοιχεία που αποχωρούν είναι τα πρώτα 5 που φαίνονται και στην εικόνα δηλαδή:

- Απόσταση από τη μέση τιμή
- Relative Strength Index (RSI) για περίοδο 6 και 14 ημερών
- Moving Average Convergence Divergence (MACD)

Ένας ακόμα τρόπος βελτιστοποίησης που εφαρμόστηκε είναι η προσαρμογή ορισμένων υπερπαραμέτρων (tuning) χρησιμοποιώντας RandomizedSearchCV όπως και με τη μέθοδο Random Forest. Ακολουθώντας την ίδια διαδικασία, η μόνη διαφορά που παρατηρείται είναι η επιλογή των υπερπαραμέτρων. Δοκιμάστηκαν οι πιο σημαντικές οι οποίες είναι ο αριθμός των δέντρων (n_estimators), το μέγιστο βάθος ενός δέντρου (max_depth) καθώς και ο ρυθμός εκμάθησης του μοντέλου (learning_rate). Εφαρμόστηκαν δύο περιπτώσεις, η μία με 3-fold και η άλλη με 5-fold cross validation για να υπάρχει μια πλήρη εικόνα.

Είναι σημαντικό να αναφερθεί ότι σε αντίθεση με τα άλλα μοντέλα, το Gradient Boosting μπορεί να παρουσιάσει πρόβλημα overfitting πράγμα που διαπιστώθηκε και από τις υψηλές τιμές που προέκυπταν κατά την αξιολόγηση του μοντέλου. Για το λόγο αυτό, μειώθηκε το εύρος τιμών για τις υπερπαραμέτρους καθώς παρατηρείται ότι

ελαττώνοντας τον αριθμό των επαναλήψεων , το μέγιστο βάθος των δέντρων και το δείκτη εκμάθησης, βελτιώνεται το πρόβλημα overfitting σημαντικά όπως παρατηρείται και από τα αποτελέσματα.

Gradient Boosting (next day prediction)	*RPS*	*Minimize overfitting*	*Comparison with benchmark*
Baseline	0.2111		1.3194
Feature selection only	0.2106		1.3160
3-fold tuning only	0.2093	0.1662	1.0389
5-fold tuning only	0.1986	0.1746	1.0915
Feature selection & 3-fold tuning	0.1766	0.1591	0.9945
Feature selection & 5-fold tuning	0.1974	0.1610	1.0065

Πίνακας 21: Αποτελέσματα RPS για τη μέθοδο Gradient Boosting

Επισκόπηση αποτελεσμάτων

Στο σχήμα παρατίθεται μια τελική επισκόπηση των καλύτερων αποτελεσμάτων για κάθε μέθοδο για βραχυπρόθεσμο ορίζοντα πρόβλεψης.

Methods	*RPS*	*Comparison with benchmark*
Naïve	0.3095	1.9344
SES on ranks	0.1714	1.0713
SES on percentage returns	0.1946	1.2163
Frequency of ranks	0.1587	0.9919
Logistic Regression next day	0.1590	0.9938
Random Forest next day	0.1654	1.0338
Gradient Boosting next day	0.1591	0.9944

Πίνακας 22: Τελικά αποτελέσματα RPS των μεθόδων με ορίζοντα πρόβλεψης την ακόλουθη μέρα

3.3.4 Επίδραση του ορίζοντα πρόβλεψης

Για την εύρεση του αποτελεσματικότερου μοντέλου, εξετάστηκαν οι ίδιες μεθοδολογίες τεχνικών μηχανικής μάθησης για διαφορετικό ορίζοντα πρόβλεψης. Πιο συγκεκριμένα, δοκιμάστηκε η παραγωγή προβλέψεων για το μεσοπρόθεσμο ορίζοντα του επόμενου μήνα αντί για τη βραχυπρόθεσμη πρόβλεψη της επόμενης ημέρας. Γίνεται η παραδοχή ότι ο επόμενος μήνας αντιστοιχεί γενικά σε 22 εργάσιμες ημέρες καθώς οι εργάσιμες μέρες των μηνών που πραγματοποιούνται οι προβλέψεις (Ιανουάριος-Μάρτιος) κυμαίνονται σε ένα εύρος 20-23.

Λογιστική παλινδρόμηση, Τυχαία Δάση, Δέντρα Ενίσχυσης Κλίσης - πρόβλεψη επόμενου μήνα

Για κάθε μέθοδο ακολουθείται η αντίστοιχη διαδικασία που εφαρμόστηκε και προηγουμένως και προκύπτουν τα τελικά αποτελέσματα :

Logistic Regression (next month prediction)	*RPS*	*Comparison with benchmark*
Baseline	0.1688	1.055
After feature selection	0.1586	0.9913

Πίνακας 23: Αποτελέσματα RPS για τη μέθοδο Logistic Regression με ορίζοντα πρόβλεψης τον επόμενο μήνα

Random Forest (next month prediction)	*RPS*	*Comparison with benchmark*
Baseline	0.1677	1.0483
Feature selection only	0.1673	1.0457
3-fold tuning only	0.1676	1.0474
5-fold tuning only	0.1672	1.0451
Feature selection & 3-fold tuning	0.1667	1.0424
Feature selection & 5-fold tuning	0.1670	1.0443

Πίνακας 24: Αποτελέσματα RPS για τη μέθοδο Random Forest με ορίζοντα πρόβλεψης τον επόμενο μήνα

Gradient Boosting (next month prediction)	*RPS*	*Minimize overfitting*	*Comparison with benchmark*
Baseline	0.2125		1.3194
Feature selection only	0.2119		1.3160
3-fold tuning only	0.2075	0.1684	1.0389
5-fold tuning only	0.2080	0.1686	1.0915
Feature selection & 3-fold tuning	0.2062	0.1586	0.9945
Feature selection & 5-fold tuning	0.2062	0.1586	1.0065

Πίνακας 25: Αποτελέσματα RPS για τη μέθοδο Gradient Boosting με ορίζοντα πρόβλεψης τον επόμενο μήνα

Επισκόπηση αποτελεσμάτων

Στο σχήμα παρατίθεται μια τελική επισκόπηση των καλύτερων αποτελεσμάτων για κάθε μέθοδο για το μεσοπρόθεσμο ορίζοντα πρόβλεψης.

Methods	*RPS*	*Comparison with benchmark*
Logistic Regression next month	0.1586	0.9913
Random Forest next month	0.1668	1.0425
Gradient Boosting next month	0.1586	0.9913

Πίνακας 26: Τελικά αποτελέσματα RPS των μεθόδων με ορίζοντα πρόβλεψης τον ακόλουθο μήνα

3.3.5 Επίδραση συσταδοποίησης των μετοχών στην εκπαίδευση των μοντέλων

Σε αυτήν την τελική δοκιμή, για λόγους πληρότητας εξετάζεται εάν η ομαδοποίηση των μετοχών που πραγματοποιήθηκε προηγουμένως, θα μπορούσε να οδηγήσει σε πιο αποδοτική εκπαίδευση των μοντέλων. Για το λόγο αυτό με τη χρήση του αλγορίθμου k-means κι έπειτα από ανάλυση σε κύριες συνιστώσες, χρησιμοποιήθηκαν οι 6 ομάδες που δημιουργήθηκαν. Πιο συγκεκριμένα, η κάθε ομάδα εκπαιδευόταν σε ένα διαφορετικό μοντέλο σε κάθε περίπτωση. Επομένως, συνολικά για κάθε μέθοδο εκπαιδεύτηκαν 6 διαφορετικά μοντέλα, ένα για κάθε συστάδα, διαδικασία αρκετά πιο χρονοβόρα σε σύγκριση με τις προηγούμενες μεθοδολογίες. Για το λόγο αυτό εξετάστηκε για κάθε μέθοδο ένα βασικό μοντέλο χωρίς τα επιπλέον βήματα βελτιστοποίησης που εφαρμόστηκαν και στη συνέχεια τα αποτελέσματα αυτά συγκρίθηκαν με τα αντίστοιχα στις προηγούμενες μεθόδους έτσι ώστε να διαπιστωθεί εάν αξίζει περαιτέρω διερεύνηση του τρόπου αυτού εκπαίδευσης.

Methods	*Baseline model*	*Baseline model with clustering*
Logistic Regression next day	0.1709	0.1711
Random Forest next day	0.1661	0.1664
Gradient Boosting next day	0.2111	0.2111
Logistic Regression next month	0.1688	0.1694
Random Forest next month	0.1677	0.1672
Gradient Boosting next month	0.2125	0.2078

Πίνακας 27: Σύγκριση αποτελεσμάτων RPS μετά από τη συσταδοποίηση μετοχών

Όπως διαπιστώνεται, δεν υπάρχουν μεγάλες διαφορές στα αποτελέσματα μεταξύ τους και για αυτό η μέθοδος αυτή εκπαίδευσης των μοντέλων απορρίπτεται από περαιτέρω εξερεύνηση.

3.3.6 Εξέταση βέλτιστων μεθόδων σε σχέση με το διαγωνισμό M6

Σε αυτό το μέρος της μελέτης, έχει ολοκληρωθεί η εξέταση των μοντέλων στο σύνολό τους και προκειμένου να υπάρξει μια ολοκληρωμένη εικόνα, γίνεται επιλογή των βέλτιστων μεθόδων οι οποίες αξιολογούνται για το χρονικό διάστημα διεξαγωγής του διαγωνισμού και αντικατοπτρίζουν τις πραγματικές συνθήκες του. Παράγονται προβλέψεις που αντιστοιχούν στις 6 υποβολές που πραγματοποιήθηκαν για το πρώτο και δεύτερο τρίμηνο (Q1 και Q2) οι οποίες και θα συγκριθούν στη συνέχεια με τις πραγματικές κατατάξεις υποβολών στο επόμενο υποκεφάλαιο. Τα καλύτερα μοντέλα που θα εξεταστούν είναι οι μέθοδοι Frequency of ranks, Logistic Regression και Gradient Boosting. Η εξέταση των μοντέλων πραγματοποιείται προσομοιώνοντας την κάθε μέθοδο στις πραγματικές συνθήκες του διαγωνισμού.

- Συχνότητα εμφάνισης κλάσης

Με στόχο την προσομοίωση των πραγματικών συνθηκών του διαγωνισμού, η συλλογή των δεδομένων για εκπαίδευση και η πραγματοποίηση της πρόβλεψης γινόταν μέχρι και την ημερομηνία λήξης της προθεσμίας υποβολής. Για παράδειγμα, για την 1η υποβολή συλλέγονταν δεδομένα μέχρι και την Παρασκευή 4 Μαρτίου και η πρόβλεψη εκείνης της ημέρας αντιστοιχούσε στην υποβολή για την 1η Απριλίου. Όπως και προηγουμένως, δοκιμάζεται η μέθοδος για διαφορετικές περιόδους εκπαίδευσης έτσι ώστε να διαπιστωθεί από ποια περίοδο και μετά παράγονται καλύτερα αποτελέσματα.

Ακολουθώντας τα ίδια βήματα για όλες τις υποβολές ξεχωριστά και για διαφορετικές περιόδους εκπαίδευσης, τα αποτελέσματα φαίνονται στον παρακάτω πίνακα.

Q1- Frequency of ranks	*15 days*	*1 month*	*2 months*	*3 months*	*6 months*	*All the samples*
1st April	0.1651	0.1697	0.1655	0.1619	0.1619	0.1620
29 April	0.1749	0.1664	0.1562	0.1518	0.1600	0.1619
27 May	0.1924	0.1943	0.1688	0.1715	0.1669	0.1584
Average	0.1774	0.1768	0.1635	0.1619	0.1629	0.1608
Comparison with benchmark	1.1089	1.1051	1.0219	1.0111	1.0185	1.0049

Πίνακας 28: Αποτελέσματα RPS μεθόδου frequency of ranks για το διαγωνισμό m6 (Q1)

Q2- Frequency of ranks	*15 days*	*1 month*	*2 months*	*3 months*	*6 months*	*All the samples*
24 June	0.1898	0.1801	0.1825	0.1738	0.1724	0.1639
22 July	0.1646	0.1664	0.1632	0.1490	0.1492	0.1526
19 August	0.1825	0.1662	0.1579	0.1499	0.1476	0.1504
Average	0.18	0.17	0.17	0.16	0.16	0.16
Comparison with benchmark	1.11	1.09	1.04	1.00	1.00	0.99

Πίνακας 29: Αποτελέσματα RPS μεθόδου frequency of ranks για το διαγωνισμό m6 (Q2)

- Λογιστική Παλινδρόμηση

Για αυτήν την περίπτωση και την επόμενη, τη μέθοδο Gradient Boosting δοκιμάστηκαν δύο μέθοδοι εκπαίδευσης των μοντέλων: στη μία περίπτωση η εκπαίδευση του μοντέλου γινόταν ανά υποβολή δηλαδή ανά μήνα και στην άλλη περίπτωση ανά τρίμηνο. Η μελέτη και στις δύο περιπτώσεις συνεχίστηκε με τη μέθοδο με το μεσοπρόθεσμο ορίζοντα πρόβλεψης, καθώς σε κάθε περίπτωση υποβολής η διορία έληγε σχεδόν ένα μήνα νωρίτερα κι επομένως η ημερήσια παραγωγή πρόβλεψης δε θα ήταν εφικτή λόγω έλλειψης δεδομένων.

Είναι σημαντικό να αναφερθεί ότι η εκπαίδευση απαιτείται να σταματήσει πριν τη διορία υποβολής του διαγωνισμού καθώς δεν υπάρχουν διαθέσιμα δεδομένα για κατασκευή χαρακτηριστικών τις ακόλουθες μέρες (τιμή κλεισίματος, ανοίγματος κλπ). Προκειμένου βέβαια να προσομοιωθούν οι πραγματικές συνθήκες του διαγωνισμού, πρέπει επίσης να ληφθούν υπόψη και οι ετικέτες (labels) που αφορούν τις πραγματικές κλάσεις των μετοχών. Εφόσον η μέθοδος πραγματοποιεί προβλέψεις με ορίζοντα τον επόμενο μήνα, ως ετικέτες χρησιμοποιούνται οι κλάσεις του επόμενου μήνα. Αυτό βέβαια αποτελεί πρόβλημα καθώς για παράδειγμα εάν η εκπαίδευση λήγει την 1η Μαρτίου και η διορία υποβολής είναι στις 6 Μαρτίου γεγονός που σημαίνει ότι δεν είναι γνωστά τα πραγματικά στοιχεία μετά την ημερομηνία αυτή, δε θα ήταν σωστό στην εκπαίδευση στις 25 Φεβρουαρίου να δίνονται ως ετικέτα η πραγματική κλάση της 25ης Μαρτίου, ένα μήνα αργότερα, καθώς αυτή η πληροφορία δεν είναι διαθέσιμη. Για το λόγο αυτό, προκειμένου να προσομοιωθούν πλήρως οι πραγματικές συνθήκες του διαγωνισμού, η εκπαίδευση του μοντέλου επιλέγεται να σταματήσει ένα μήνα νωρίτερα από ότι ήταν για κάθε υποβολή έτσι ώστε να αξιολογηθεί το μοντέλο με τα ίδια δεδομένα που ήταν διαθέσιμα και για τους συμμετέχοντες του διαγωνισμού. Δηλαδή, για την πρώτη υποβολή η εκπαίδευση του μοντέλου σταματάει στις 31 Ιανουαρίου, για τη δεύτερη υποβολή στις 28 Φεβρουαρίου κοκ. Τα τελικά αποτελέσματα μετά και από διαλογή χαρακτηριστικών (feature selection) φαίνονται και στον παρακάτω πίνακα και μπορούν να συγκριθούν και με το βαθμολογικό πίνακα του διαγωνισμού.

Q1- Logistic Regression	*RPS*	*Comparison with benchmark*
1st April	0.1617	1.0108
29 April	0.1609	1.0055
27 May	0.1603	1.0020
Average	0.1610	1.0061

Πίνακας 30: Αποτελέσματα RPS μεθόδου Logistic Regression για το διαγωνισμό m6 (Q1) - Εκπαίδευση ανά υποβολή

Q2- Logistic Regression	*RPS*	*Comparison with benchmark*
24 June	0.1658	1.0363
22 July	0.1495	0.9345
19 August	0.1450	0.9064
Average	0.1535	0.9591

Πίνακας 31: Αποτελέσματα RPS μεθόδου Logistic Regression για το διαγωνισμό m6 (Q2) - Εκπαίδευση ανά υποβολή

Στη συνέχεια, η ίδια μέθοδος δοκιμάζεται με τη διαφορά ότι πραγματοποιείται μία μόνο φορά εκπαίδευση του μοντέλου μέχρι τις 31 Ιανουαρίου για το πρώτο τρίμηνο και μέχρι τις 21 Απριλίου για το δεύτερο τρίμηνο και πραγματοποιείται σε κάθε περίπτωση κατευθείαν πρόβλεψη και για τις 3 υποβολές του τριμήνου. Στην ουσία η διαφορά είναι στο ότι η εκπαίδευση του μοντέλου εδώ γίνεται ανά τρίμηνο ενώ στην προηγούμενη περίπτωση ανά υποβολή δηλαδή ανά μήνα. Τα αποτελέσματα με μικρές διαφορές φαίνονται παρακάτω :

Q1- Logistic Regression	*RPS*	*Comparison with benchmark*
1st April	0.1617	1.0108
29 April	0.1617	1.0107
27 May	0.1594	0.9964
Average	0.1610	1.006

Πίνακας 32: Αποτελέσματα RPS μεθόδου Logistic Regression για το διαγωνισμό m6 (Q1) - Εκπαίδευση ανά τρίμηνο

Q2- Logistic Regression	*RPS*	*Comparison with benchmark*
24 June	0.1658	1.0363
22 July	0.1510	0.9438
19 August	0.1496	0.9349
Average	0.1555	0.9717

Πίνακας 33: Αποτελέσματα RPS μεθόδου Logistic Regression για το διαγωνισμό m6 (Q2) - Εκπαίδευση ανά τρίμηνο

- Δέντρα Ενίσχυσης Κλίσης

Ακολουθώντας την ίδια διαδικασία που αναλύθηκε παραπάνω, για τη μέθοδο με το μεσοπρόθεσμο ορίζοντα πρόβλεψης (τον ακόλουθο μήνα) πραγματοποιήθηκαν οι τελικές προβλέψεις σύμφωνα και με τα διαθέσιμα δεδομένα του διαγωνισμού με δύο τρόπους, εκπαιδεύοντας στη μία το μοντέλο ανά υποβολή και στην άλλη ανά τρίμηνο.

Q1- Gradient Boosting	*RPS*	*Comparison with benchmark*
1st April	0.1629	1.0182
29 April	0.1620	1.0127
27 May	0.1604	1.0027
Average	0.1618	1.0112

Πίνακας 34: Αποτελέσματα RPS μεθόδου Gradient Boosting για το διαγωνισμό m6 (Q1) - Εκπαίδευση ανά υποβολή

Q2- Gradient Boosting	*RPS*	*Comparison with benchmark*
24 June	0.1667	1.0417
22 July	0.1496	0.9348
19 August	0.1446	0.9037
Average	0.1536	0.9601

Πίνακας 35: Αποτελέσματα RPS μεθόδου Gradient Boosting για το διαγωνισμό m6 (Q2) - Εκπαίδευση ανά υποβολή

Q1- Gradient Boosting	*RPS*	*Comparison with benchmark*
1st April	0.1632	1.0203
29 April	0.1596	0.9975
27 May	0.1629	1.018
Average	0.1619	1.012

Πίνακας 36: Αποτελέσματα RPS μεθόδου Gradient Boosting για το διαγωνισμό m6 (Q1) - Εκπαίδευση ανά τρίμηνο

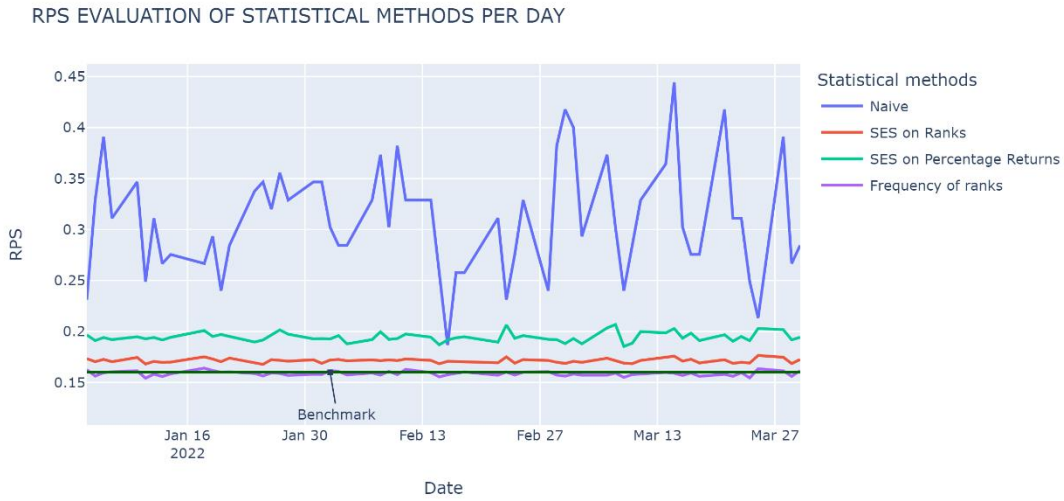
Q2- Gradient Boosting	*RPS*	*Comparison with benchmark*
24 June	0.1658	1.0363
22 July	0.1510	0.9438
19 August	0.1496	0.9349
Average	0.1555	0.9717

Πίνακας 37: Αποτελέσματα RPS μεθόδου Gradient Boosting για το διαγωνισμό m6 (Q2) - Εκπαίδευση ανά τρίμηνο

3.4 ΑΞΙΟΛΟΓΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

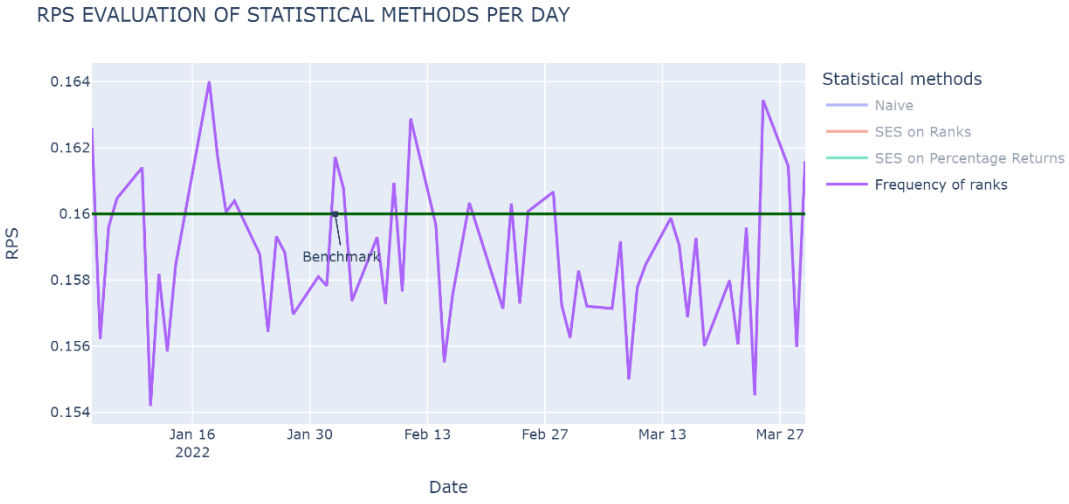
Όπως προαναφέρθηκε, σκοπός της παρούσας διπλωματικής είναι η ανάπτυξη ολοκληρωμένων μεθοδολογιών κατάταξης επενδυτικών αγαθών προκειμένου να αξιολογηθεί η απόδοση της κάθε μεθοδολογίας σε πραγματικά δεδομένα και να εντοπιστεί η πλέον κατάλληλη με τη βοήθεια της μετρικής RPS (Ranked Probability Score) που αναλύθηκε και παραπάνω. Τα αποτελέσματα της κάθε μεθόδου συγκρίνονται στη συνέχεια με την τιμή benchmark 0.16 καθώς και οι βέλτιστες μέθοδοι πραγματοποιούν προβλέψεις για το πρώτο διάστημα διεξαγωγής του διαγωνισμού m6 προκειμένου να συγκριθούν και με το βαθμολογικό πίνακα του διαγωνισμού.

Στο παρακάτω σχήμα, παρουσιάζεται η απόδοση των στατιστικών μοντέλων για το κοινό διάστημα από τον Ιανουάριο μέχρι και τον Μάρτιο του 2022. Έπειτα, στο επόμενο σχήμα παρουσιάζεται η απόδοση των τεχνικών μηχανικής μάθησης τόσο σε βραχυπρόθεσμο όσο και σε μεσοπρόθεσμο ορίζοντα πρόβλεψης καθώς και ο τελικός πίνακας επιδόσεων για όλες τις μεθόδους που εξετάστηκαν κατά το πρώτο τρίμηνο του 2022. Για κάθε περίπτωση επιλέχθηκε η μεθοδολογική προσέγγιση που προσέφερε τα καλύτερα αποτελέσματα και ο τελικός μέσος όρος των τιμών για τις μεθόδους που εμφανίζονται, προέκυψε μετά από την προεπεξεργασία, εκπαίδευση και προσαρμογή των παραμέτρων και δεδομένων κάθε μοντέλου. Για την οπτικοποίηση των αποτελεσμάτων χρησιμοποιήθηκε η βιβλιοθήκη plotly express της python.

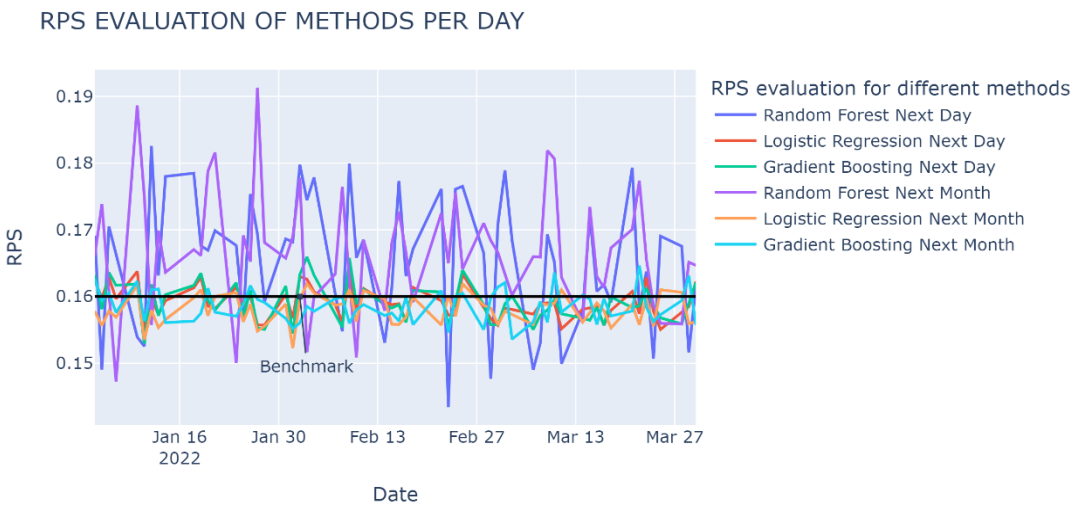


Εικόνα 53: Στατιστικές μέθοδοι

Όσον αφορά τις στατιστικές μεθόδους παρατηρείται ότι με διαφορά η μόνη που συγκρίνεται με την τιμή benchmark είναι η μέθοδος frequency of ranks ενώ οι υπόλοιπες παράγουν χειρότερα αποτελέσματα κι επομένως απορρίπτονται από περαιτέρω ανάλυση.



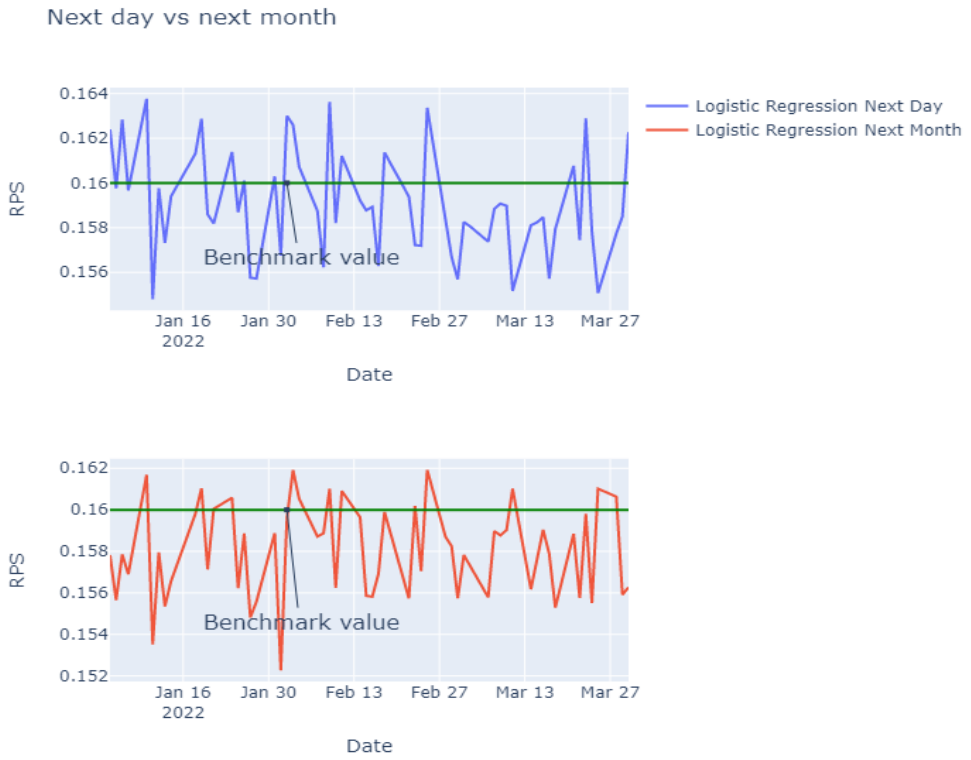
Εικόνα 54: Σύγκριση της μεθόδου frequency of ranks με τη benchmark τιμή



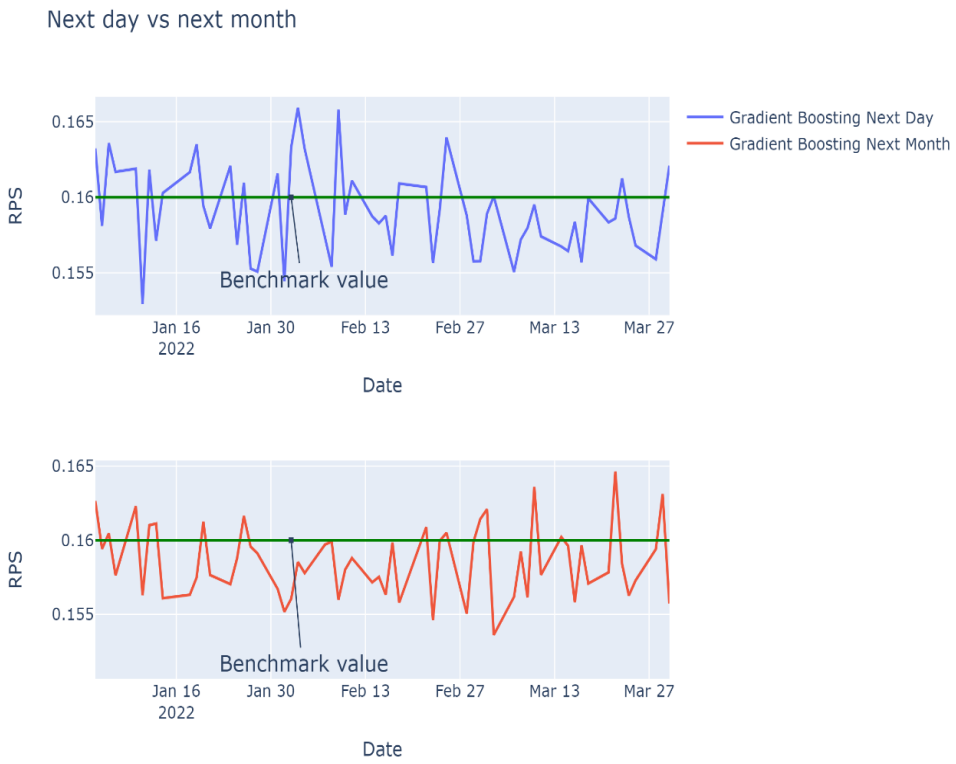
Εικόνα 55: Τεχνικές Μηχανικής μάθησης

Σχετικά με τις τεχνικές μηχανικής μάθησης παρατηρείται ότι λειτουργούν καλύτερα κατά το πέρας των ημερών οι μέθοδοι Logistic Regression και Gradient Boosting γεγονός που επιβεβαιώνει την επιλογή τους και για την παραγωγή προβλέψεων κατά το πρώτο διάστημα του διαγωνισμού M6. Σε αντίθεση, η μέθοδος Random Forest τόσο σε βραχυπρόθεσμο όσο και σε μεσοπρόθεσμο ορίζοντα αν και όχι άστοχα, δεν προσφέρει βέλτιστα αποτελέσματα και γι αυτό και αυτή απορρίπτεται από περαιτέρω ανάλυση.

Από τις τεχνικές μηχανικής μάθησης που απέμειναν αξίζει να οπτικοποιηθούν τα αποτελέσματα των μεθόδων για το βραχυπρόθεσμο και μεσοπρόθεσμο ορίζοντά τους.



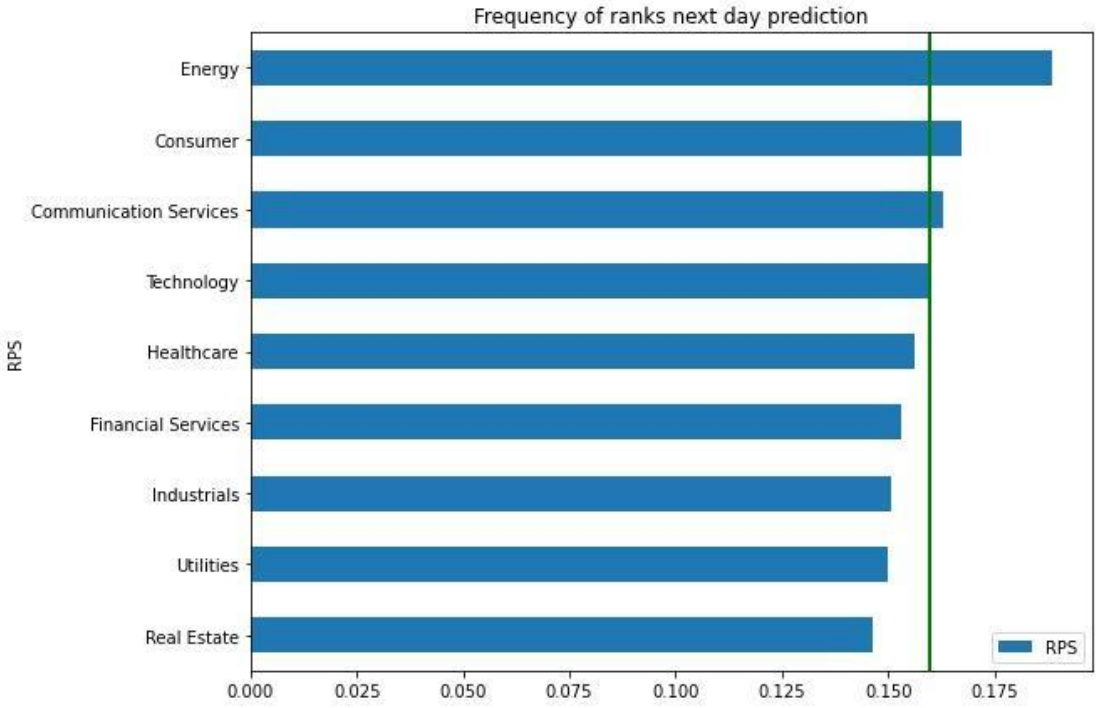
Εικόνα 56: Αξιολόγηση μεθόδου Logistic Regression για βραχυπρόθεσμο και μεσοπρόθεσμο ορίζοντα πρόβλεψης

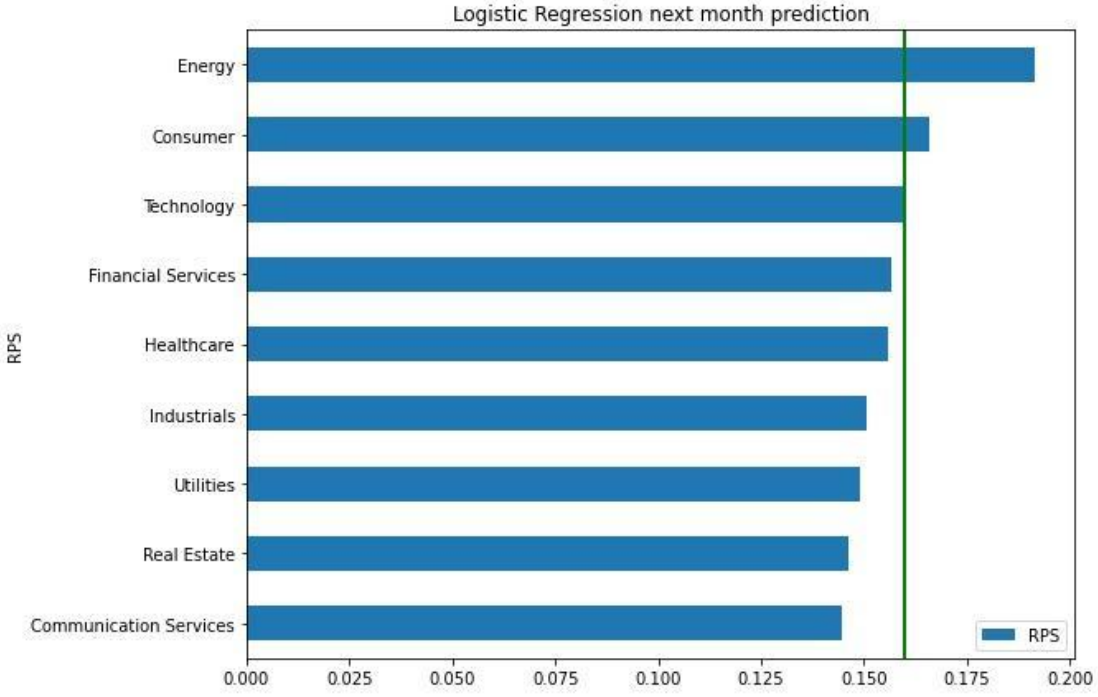


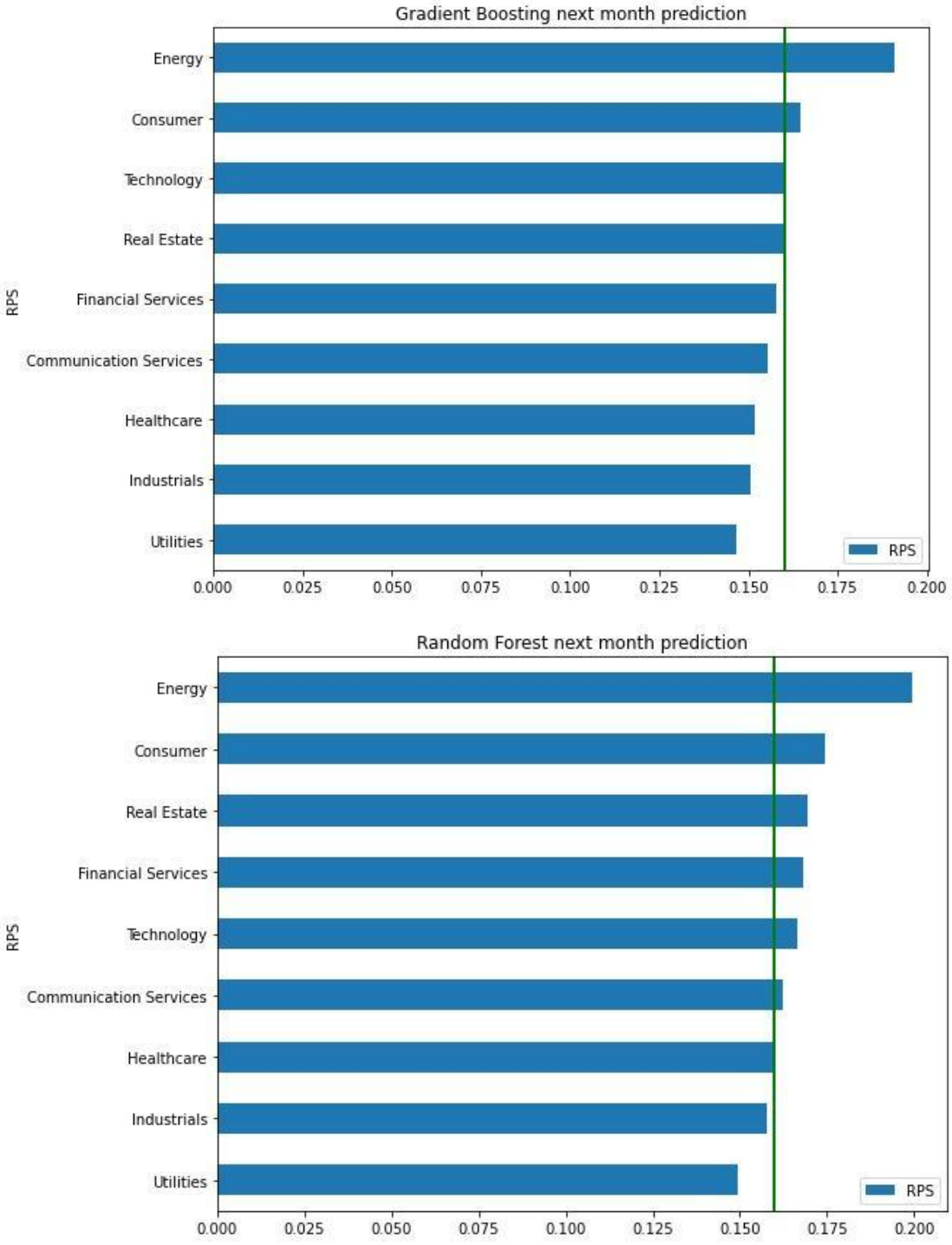
Εικόνα 57: Αξιολόγηση μεθόδου Gradient Boosting για βραχυπρόθεσμο και μεσοπρόθεσμο ορίζοντα πρόβλεψης

Και στις δύο περιπτώσεις φαίνεται ότι παράγονται καλύτερες προβλέψεις στο μεσοπρόθεσμο ορίζοντα καθώς η πλειοψηφία των ημερήσιων μέσων τιμών RPS βρίσκονται κάτω από την τιμή benchmark. Επομένως, καθημερινά παράγονται καλύτερα αποτελέσματα για προβλέψεις που αφορούν τον επόμενο μήνα.

Η αξιολόγηση των αποτελεσμάτων μπορεί να γίνει όχι μόνο ανά ημέρα αλλά και ανά κλάδο. Συνολικά, οι μετοχές χωρίζονται σε 9 διαφορετικούς τομείς όπως υγεία, οικονομικές υπηρεσίες, ενέργεια κλπ. με το μεγαλύτερο αριθμό μετοχών να ανήκουν στις οικονομικές υπηρεσίες (financial services) και τον τομέα καταναλωτή (consumer). Ενδεικτικά, φαίνονται τα αποτελέσματα για τις καλύτερες μεθόδους που αναφέρθηκαν και παραπάνω και η οπτικοποίησή τους πραγματοποιείται με τη χρήση της βιβλιοθήκης matplotlib της python.


Frequency of ranks next day prediction
Energy
Consumer
Communication Services
Technology
Healthcare
Financial Services
Industrials
Utilities
Real Estate
RPS
RPS
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175


Logistic Regression next month prediction
Energy
Consumer
Technology
Financial Services
Healthcare
Industrials
Utilities
Real Estate
Communication Services
RPS
RPS
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200



Εικόνα 58: Αξιολόγηση βέλτιστων μοντέλων ανά βιομηχανία

Παρατηρείται ότι σε όλες τις μεθόδους παράγονται χειρότερα αποτελέσματα στον τομέα της ενέργειας και του καταναλωτή. Βέβαια είναι σημαντικό να αναφερθεί ότι το αποτέλεσμα της πρώτης κατηγορίας μπορεί να μην είναι αντιπροσωπευτικό του συνόλου των μετοχών που ανήκουν σε αυτή γενικότερα καθώς αποτελείται από 2 μόνο μετοχές. Επίσης, οι συνθήκες με τον πόλεμο με τη Ρωσία και την ενεργειακή κρίση, καθιστούν τις πρόβλεψεις στον τομέα της ενέργειας εξαιρετικά δύσκολες. Οι μέσοι όροι των αποδόσεων κάθε μεθόδου φαίνονται αναλυτικά και στον παρακάτω πίνακα επιδόσεων.

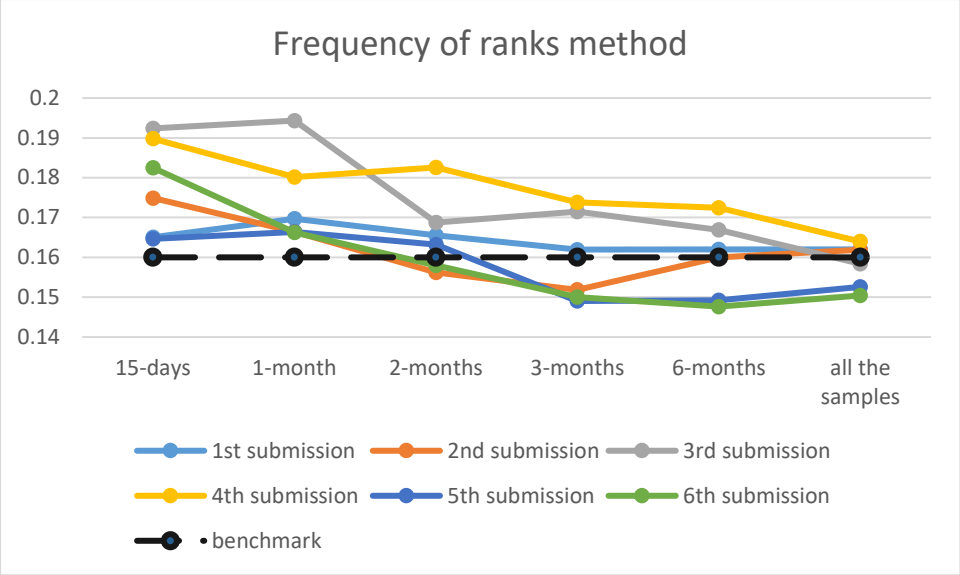
Methods	*RPS*	*Comparison with benchmark*
Next day prediction:		
Naïve	0.3095	1.9344
SES on ranks	0.1714	1.0713
SES on percentage returns	0.1946	1.2163
Frequency of ranks	0.1587	0.9919
Logistic Regression	0.1590	0.9938
Random Forest	0.1654	1.0338
Gradient Boosting	0.1591	0.9944
Next month prediction:		
Logistic Regression	0.1586	0.9913
Random Forest	0.1668	1.0425
Gradient Boosting	0.1586	0.9913

Πίνακας 38: Τελικά αποτελέσματα

Όπως φαίνεται, οι 3 επικρατέστερες μέθοδοι είναι η frequency of ranks, το logistic regression και το gradient boosting οι οποίες και χρησιμοποιούνται και για την παραγωγή προβλέψεων για το διαγωνισμό m6. Αξίζει να σημειωθεί βέβαια, καθώς το RPS πρόκειται για τετραγωνική μετρική, όταν για παράδειγμα η καλύτερη μέθοδος αξιολογείται με RPS 0.1586 το οποίο συγκριτικά με τη benchmark τιμή αναλογεί σε 0.9913 δηλαδή βελτίωση 1%, στην πραγματικότητα ποσοστιαία υπάρχει βελτίωση :

$$\sqrt{\frac{(0.16 - 0.1586)}{0.16}} \approx 9\%$$

Ξεκινώντας λοιπόν με τη μέθοδο frequency of ranks, παράγονται προβλέψεις για διαφορετικά διαστήματα εκπαίδευσης του μοντέλου όπως εξηγήθηκε και στο προηγούμενο υποκεφάλαιο έτσι ώστε να διαπιστωθεί ποια περίοδος εκπαίδευσης είναι βέλτιστη για την παραγωγή προβλέψεων. Η οπτικοποίηση των αποτελεσμάτων πραγματοποιήθηκε με χρήση excel.



Εικόνα 59: Αξιολόγηση μεθόδου frequency of ranks για τον διαγωνισμό m6

Είναι πολύ δύσκολο να φανεί μια συγκεκριμένη περίοδος κατά την οποία παράγονται σε όλες τις περιπτώσεις τα καλύτερα αποτελέσματα και καθώς δεν είναι ρεαλιστικό να είναι γνωστό από πριν ποια περίοδος εκπαίδευσης είναι βέλτιστη κάθε φορά γιατί αυτό θα αναιρούσε το σκοπό της πρόβλεψης, η μέθοδος αυτή αν και δίνοντας καλά αποτελέσματα απορρίπτεται.

Στη συνέχεια εξετάζονται οι δύο επικρατέστερες μέθοδοι, το Logistic Regression και το Gradient Boosting για παραγωγή πρόβλεψης στο μεθοπρόθεσμο ορίζοντα του επόμενου μήνα , προσομοιώνοντας τις πραγματικές συνθήκες του διαγωνισμού M6. Στον παρακάτω πίνακα, φαίνεται για κάθε περίπτωση η αξιολόγηση της μεθόδου για κάθε υποβολή ξεχωριστά, καθώς και η σύγκρισή της με τα αποτελέσματα του διαγωνισμού. Ρεαλιστικά, αναφέρεται και η κατάταξη στην οποία θα ανήκε εάν έπαιρνε μέρος στο διαγωνισμό.

Logistic Regression	*RPS*	*Comparison with benchmark*	*Best RPS from the competition*	*Comparison with benchmark*	*Rank*
1st submission	0.1617	1.0108	0.14458	0.9036	89/163
2nd submission	0.1609	1.0055	0.14900	0.9313	72/177
3rd submission	0.1603	1.0020	0.15289	0.9556	95/187
4th submission	0.1658	1.0363	0.14350	0.8968	134/199
5th submission	0.1495	0.9345	0.14473	0.9046	5/201
6th submission	0.1450	0.9064	0.12778	0.7986	5/206

Πίνακας 39: Σύγκριση αποτελεσμάτων RPS ανά υποβολή της μεθόδου Logistic Regression σε σχέση με τον διαγωνισμό m6

Gradient Boosting	*RPS*	*Comparison with benchmark*	*Best RPS from the competition*	*Comparison with benchmark*	*Rank*
1st submission	0.1629	1.0181	0.14458	0.9036	95/163
2nd submission	0.1596	0.9975	0.14900	0.9313	25/177
3rd submission	0.1604	1.0025	0.15289	0.9556	98/187
4th submission	0.1658	1.0363	0.14350	0.8968	134/199
5th submission	0.1510	0.9438	0.14473	0.9046	5/201
6th submission	0.1496	0.9349	0.12778	0.7986	13/206

Πίνακας 40: Σύγκριση αποτελεσμάτων RPS ανά υποβολή της μεθόδου Gradient Boosting σε σχέση με τον διαγωνισμό m6

Οι ίδιες πληροφορίες παρουσιάζονται και στους παρακάτω πίνακες αυτήν τη φορά ανά τρίμηνο Q1 και Q2 ως ο μέσος όρος των προβλέψεων :

Logistic Regression	*RPS*	*Comparison with benchmark*	*Best RPS from the competition*	*Comparison with benchmark*	*Rank*
Q1	0.1609	1.0059	0.15577	0.9736	71/163
Q2	0.1535	0.9591	0.14937	0.9336	8/199

Πίνακας 41: Σύγκριση αποτελεσμάτων RPS ανά τρίμηνο της μεθόδου Logistic Regression σε σχέση με τον διαγωνισμό m6

Gradient Boosting	*RPS*	*Comparison with benchmark*	*Best RPS from the competition*	*Comparison with benchmark*	*Rank*
Q1	0.1618	1.0112	0.15577	0.9736	75/163
Q2	0.1536	0.9600	0.14937	0.9336	10/199

Πίνακας 42: Σύγκριση αποτελεσμάτων RPS ανά τρίμηνο της μεθόδου Gradient Boosting σε σχέση με τον διαγωνισμό m6

Παρατηρείται ότι τα μοντέλα λειτουργούν καλύτερα κατά τη διάρκεια του δεύτερου τριμήνου με λίγο καλύτερα αποτελέσματα να παράγονται από τη μέθοδο Logistic Regression προσφέροντας συνολική βελτίωση στο δεύτερο τρίμηνο σε σχέση με τη benchmark τιμή περίπου 20%. Παρόλα αυτά, είναι φανερό ότι στην πλειοψηφία των περιπτώσεων οι προβλέψεις δεν απέχουν τόσο πολύ ακόμα κι από τις καλύτερες μετρικές του διαγωνισμού. Η πλειοψηφία των συμμετεχόντων έχει πιάσει τη benchmark τιμή σε πολλές περιπτώσεις με αποτέλεσμα ακόμα και αν η κατάταξη (rank) της πρόβλεψης φαίνεται να είναι πχ 72/177 με 0,1609 αυτό οφείλεται σε μικρές διαφορές καθώς μπορεί δεκάδες διαγωνιζόμενοι να έχουν πιάσει την τιμή 0,16. Επομένως, επαληθεύεται το συμπέρασμα ότι η πρόβλεψη της κατάταξης των μετοχών είναι μια δύσκολη και περίπλοκη διαδικασία και ακόμα και οι καλύτερες μέθοδοι δεν αποδίδουν πάντα τις καλύτερες προβλέψεις.

ΚΕΦΑΛΑΙΟ 4: ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ

Στο τελευταίο κεφάλαιο της παρούσας εργασίας, γίνεται μία σύνοψη των συμπερασμάτων που εξήχθησαν από την ανάλυση που προηγήθηκε. Η διπλωματική αυτή βασίστηκε στο πρώτο κεφάλαιο του M6 διαγωνισμού, του οποίου σκοπός ήταν να βελτιώσει την ακρίβεια των προβλέψεων στον οικονομικό τομέα, να αντιμετωπίσει το πρόβλημα της αβεβαιότητας στις προβλέψεις και να εκμεταλλευτεί τα ευρήματα αυτά προσομοιώνοντας όσο γίνεται την πραγματικότητα χρησιμοποιώντας μεθόδους πέρα από την κριτική πρόβλεψη. Εκτός από μία ανακεφαλαίωση της μελέτης και την εξαγωγή συμπερασμάτων όσον αφορά το κομμάτι των προβλέψεων, προτείνονται και ορισμένες μελλοντικές προεκτάσεις που θα μπορούσαν να διερευνηθούν περαιτέρω.

4.1 ΣΥΜΠΕΡΑΣΜΑΤΑ

Μέσω της εργασίας αυτής πραγματοποιήθηκε μια ολοκληρωμένη μελέτη όσον αφορά την πρόβλεψη της κατάταξης των 50 διαθέσιμων μετοχών που επιλέχθηκαν και για το διαγωνισμό m6. Η κατάταξη των μετοχών αυτών ορίστηκε σύμφωνα με τις ποσοστιαίες αποδόσεις τους σε 5 βασικές κλάσεις. Σε πρώτο μέρος επεξηγήθηκε αναλυτικά το θεωρητικό υπόβαθρο στο οποίο βασίστηκε η υλοποίηση των προβλέψεων δηλαδή τα μοντέλα, οι αλγόριθμοι και οι χρηματοοικονομικοί δείκτες που χρησιμοποιήθηκαν. Στη συνέχεια έγινε λόγος για το πιο σημαντικό κομμάτι, το πειραματικό στάδιο της εργασίας κατά το οποίο αναφέρθηκαν τα δεδομένα που συλλέχθηκαν καθώς και οι μεθοδολογίες που εφαρμόστηκαν τόσο για τις στατιστικές όσο και για τις τεχνικές μηχανικής μάθησης. Διερευνήθηκε κατά πόσο η αλλαγή του ορίζοντα πρόβλεψης ή η συσταδοποίηση των μετοχών μπορεί να οδηγήσουν σε καλύτερες προβλέψεις. Τέλος, παρουσιάστηκαν κι οπτικοποιήθηκαν τα αποτελέσματα για όλες τις μεθόδους ενώ οι καλύτερες προσαρμόστηκαν και συγκρίθηκαν και με τις κατατάξεις του διαγωνισμού.

Από τη μελέτη αυτή προέκυψαν ορισμένες αξιοσημείωτες παρατηρήσεις. Καταρχάς, παρατηρήθηκε ότι ανάμεσα σε όλα τα μοντέλα, μια απλή μέθοδος όπως η συχνότητα εμφάνισης κλάσης μπορεί πολλές φορές να αποδώσει τα ίδια ίσως και καλύτερα αποτελέσματα από μια πιο περίπλοκη και χρονοβόρα μέθοδο όπως η λογιστική παλινδρόμηση (logistic regression) ή τα δέντρα ενίσχυσης κλίσης (gradient boosting) που απαιτούν και περισσότερα βήματα βελτιστοποίησης. Αυτό βέβαια εξαρτάται όμως και από την επιλογή της βέλτιστης περιόδου εκπαίδευσης, έτσι ώστε να εξασφαλιστεί ότι αυτό θα ισχύει σε κάθε περίπτωση. Και σε γενικότερο πλαίσιο όμως, αν και κάποια μοντέλα εμφανίζουν σταθερά καλύτερα αποτελέσματα από τα υπόλοιπα, η απόδοση τους, ανεξαρτήτως μεθόδου φαίνεται να αλλάζει και από την περίοδο που θέλουμε να προβλέψουμε, γεγονός που επιβεβαιώνεται παρατηρώντας και τα αποτελέσματα των επικρατέστερων μεθόδων κατά τη διάρκεια του πρώτου και δεύτερου τριμήνου του διαγωνισμού, όπου οι ίδιες μέθοδοι ενώ τα πήγαν μέτρια το πρώτο τρίμηνο, στο δεύτερο είχαν αισθητά καλύτερη απόδοση χωρίς κάποια αλλαγή.

Επιπλέον, αν και κατασκευάστηκε πληθώρα χρηματοοικονομικών και στατιστικών δεικτών για την εκπαίδευση των μοντέλων στις τεχνικές μηχανικής μάθησης, στην πλειοψηφία των περιπτώσεων μεγαλύτερη χρησιμότητα παρουσίασαν για τις προβλέψεις οι κινητοί μέσοι όροι βραχυπρόθεσμου και μεσοπρόθεσμου ορίζοντα ή οι ποσοστιαίες μεταβολές των ιστορικών τιμών. Ακόμα κι αυτές όμως πάντα επιλέγονταν σε συνδυασμό με τουλάχιστον έναν τεχνικό δείκτη, συνήθως τον ADX.

Από τις ολοκληρωμένες μεθοδολογίες που εξετάστηκαν τόσο στα πλαίσια της μελέτης όσο και κατά τη διάρκεια του διαγωνισμού, παρατηρήθηκε ότι δουλεύει καλύτερα η μέθοδος της λογιστικής παλινδρόμησης (logistic regression) προσφέροντας έως και 9% βελτίωση στην παραγωγή προβλέψεων. Το επόμενο στη σειρά καλύτερο μοντέλο φαίνεται να είναι τα δεντρα ενίσχυσης κλίσης (gradient boosting) με πολύ κοντινές τιμές.

Ένα ακόμα από τα σκέλη της μελέτης, ήταν να εντοπίσει το καταλληλότερο μοντέλο και ως προς τον ορίζοντα πρόβλεψης ο οποίος φαίνεται ότι έπαιξε ρόλο στην παραγωγή προβλέψεων, με το μεσοπρόθεσμο ορίζοντα πρόβλεψης να παράγει σταθερά λίγο καλύτερες κυλιόμενες προβλέψεις. Από την άλλη βέβαια η συσταδοποίηση των μετοχών δεν επηρέασε καθόλου την εκπαίδευση των μοντέλων.

Τέλος, όσον αφορά τα αποτελέσματα ανά κλάδο για τις καλύτερες μεθόδους, βέλτιστες προβλέψεις πραγματοποιούνται στον τομέα ακινήτων, τηλεπικοινωνιών, βιομηχανίας και παροχών και διανομής ηλεκτρισμού. Δυσκολότερη παρουσιάζεται η πρόβλεψη των εταιριών στον τομέα της ενέργειας και του καταναλωτή γεγονός που μπορεί να οφείλεται στην ενεργειακή κρίση καθώς και σε μια πληθώρα προβλημάτων όπως η συνεχώς μεταβαλλόμενη ζήτηση των καταναλωτών και το καίριο πρόβλημα της εφοδιαστικής αλυσίδας (supply chain disruption).

4.2 ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ

Η παρούσα διπλωματική εξέτασε στατιστικές μεθόδους και τεχνικές μηχανικής μάθησης για την επιτυχημένη κατάταξη μετοχών τόσο σε βραχυπρόθεσμο όσο και σε μεσοπρόθεσμο ορίζοντα πρόβλεψης. Παρόλα αυτά, στο μέλλον η μελέτη αυτή θα μπορούσε να επεκταθεί και να εμπλουτιστεί έτσι ώστε να υπάρχει και μια πιο ολοκληρωμένη εικόνα για την πρόβλεψη της κατάταξης των μετοχών. Ενδεικτικά, ορισμένα σκέλη που θα μπορούσαν να διερευνηθούν συνοψίζονται παρακάτω στα εξής:

- Sentiment analysis: Στον οικονομικό τομέα χρησιμοποιείται για να εξάγει πληροφορίες από ειδήσεις, μέσα κοινωνικής δικτύωσης και οικονομικές αναφορές για τη διεξαγωγή επενδύσεων, συναλλαγών ή οτιδήποτε άλλο σχετικό με τα οικονομικά. Σύμφωνα με μελέτες, η κίνηση των τιμών του χρηματιστηρίου συσχετίζεται με το δημόσιο συναίσθημα, και όταν αυτό λαμβάνεται υπόψη στην παραγωγή προβλέψεων, η ακρίβεια τους βελτιώνεται

[49], [50]. Συνήθως, η μεθοδολογία που ακολουθείται περιλαμβάνει την εξαγωγή των ειδήσεων ή σχολίων με τη βοήθεια βιβλιοθηκών όπως το Beautiful soup, Snscrape ή πιο συγκεκριμένα τις Twint, Tweepy, εργαλεία που εξειδικεύονται στην εξαγωγή σχολίων από την πλατφόρμα Twitter, που χρησιμοποιείται ευρέως από επενδυτές. Στη συνέχεια, μετά από καθαρισμό των κειμένων από σημεία στίξης και κοινές λέξεις και με τη βοήθεια ενός αναλυτή όπως ο VADER, σε κάθε λέξη αντιστοιχεί μία τιμή που κυμαίνεται από -1 έως και 1 με τη χαμηλότερη τιμή να υποδηλώνει αρνητικό συναίσθημα, την υψηλότερη θετικό και το μηδέν ουδετερότητα. Αυτές οι πληροφορίες θα μπορούσαν είτε από μόνες τους, είτε σε συνδυασμό και με τα προηγούμενα χαρακτηριστικά να οδηγήσουν σε καλύτερη εκπαίδευση των μοντέλων.

Παρ'όλα αυτά υπάρχουν ορισμένοι περιορισμοί που καθιστούν δύσκολη την πετυχημένη ανάλυση. Καταρχάς, ορισμένες βιβλιοθήκες δεν επιτρέπουν την εξαγωγή μεγάλου αριθμού κειμένων και για αυτό τα αποτελέσματα προκύπτουν πολλές φορές να μην είναι αντιπροσωπευτικά. Στις περιπτώσεις που δεν υπάρχει περιορισμός εξαγωγής, υπάρχει πρόβλημα αποθήκευσης του μεγάλου αυτού όγκου δεδομένων. Επιπλέον, καθώς η ροή των ειδήσεων συνεχώς ανανεώνεται, παίζει σημαντικό ρόλο η κατάλληλη στιγμή εξαγωγής των ειδήσεων ώστε οι πληροφορίες που θα χρησιμοποιηθούν να μην έχουν ήδη ληφθεί υπόψη από την αγορά.

- Χρήση μοντέλων νευρωνικών δικτύων: Εφόσον οι διάφορες τιμές των μετοχών είναι διαθέσιμες σε μορφή χρονοσειρών δηλαδή διαδοχικών δεδομένων (sequential data), η χρήση νευρωνικών όπως το Long Short Term Memory (LSTM) έχουν αποδειχθεί πιο αποδοτικές για την καλύτερη αξιοποίηση των διαθέσιμων χαρακτηριστικών. Η εξέταση της μεθόδου αυτής από μόνη της ή σε συνδυασμό και με τη χρήση Convolutional neural networks (CNN) θα μπορούσαν να οδηγήσουν σε πιο αξιόπιστες προβλέψεις. Η χρήση του CNN θα βοηθούσε στην εξαγωγή των πιο χρήσιμων χρονικών χαρακτηριστικών ενώ στη συνέχεια το LSTM θα έβρισκε την αλληλεξάρτηση των δεδομένων και θα πραγματοποιούσε την πρόβλεψη. Φυσικά υπάρχουν πολλές ακόμα διαθέσιμες μέθοδοι που έχουν εφαρμοστεί όσον αφορά τη βαθιά μάθηση για την πρόβλεψη της τιμής μιας μετοχής οι οποίες θα μπορούσαν και να προσαρμοστούν ώστε να προβλέπουν την κατάταξη των μετοχών όπως και στα πλαίσια της παρούσας διπλωματικής [51], [52]. Αυτό βέβαια, λόγω του μεγάλου όγκου των δεδομένων που απαιτεί, θα λειτουργούσε καλύτερα για ενδοημερήσιες προβλέψεις και μπορεί να μην είχε τόσο καλή απόδοση με τη χρήση των διαθέσιμων δεδομένων στα πλαίσια αυτής της εργασίας.
- Εξέταση μεγαλύτερου αριθμού μετοχών: Για την εκπαίδευση των μοντέλων χρησιμοποιήθηκαν οι 50 μετοχές που ήταν διαθέσιμες και στα πλαίσια του διαγωνισμού M6. Μια ενδιαφέρουσα επέκταση της μελέτης θα ήταν η εξέταση των μοντέλων σε μεγαλύτερο πλήθος μετοχών, ισάριθμα διαχωρισμένων για κάθε βιομηχανία και η ομαδοποίησή τους ανά τομέα για εξέταση μεταβολών κατά την εκπαίδευση, έτσι ώστε να είναι πιο αντιπροσωπευτικά τα αποτελέσματα με μια πιο δίκαια κατανομή ανά κλάδο.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] C. S. Lin, S. H. Chiu, and T. Y. Lin, “Empirical mode decomposition–based least squares support vector regression for foreign exchange rate forecasting,” *Econ Model*, vol. 29, no. 6, pp. 2583–2590, Nov. 2012, doi: 10.1016/J.ECONMOD.2012.07.018.

[2] P. D. Yoo, M. H. Kim, and T. Jan, “Machine learning techniques and use of event information for stock market prediction: A survey and evaluation,” *Proceedings - International Conference on Computational Intelligence for Modelling, Control and Automation, CIMCA 2005 and International Conference on Intelligent Agents, Web Technologies and Internet*, vol. 2, pp. 835–841, 2005, doi: 10.1109/CIMCA.2005.1631572.

[3] E. F. Fama, “Random Walks in Stock Market Prices,” *https://doi.org/10.2469/faj.v51.n1.1861*, vol. 51, no. 1, pp. 75–80, Jan. 2019, doi: 10.2469/FAJ.V51.N1.1861.

[4] Y. S. Abu-Mostafa and A. F. Atiya, “Introduction to financial forecasting,” *Applied Intelligence*, vol. 6, no. 3, pp. 205–213, 1996, doi: 10.1007/BF00126626/METRICS.

[5] X. Zhong and D. Enke, “Forecasting daily stock market return using dimensionality reduction,” *Expert Syst Appl*, vol. 67, pp. 126–139, Jan. 2017, doi: 10.1016/J.ESWA.2016.09.027.

[6] J. J. Murphy, *Murphy, John J. Technical analysis of the financial markets: A comprehensive guide to trading methods and applications.* . Penguin, 1999.

[7] E. F. Fama, “Efficient Capital Markets: A Review of Theory and Empirical Work,” *J Finance*, vol. 25, no. 2, p. 383, May 1970, doi: 10.2307/2325486.

[8] E. F. Fama *et al.*, “Efficient Capital Markets: II,” *J Finance*, vol. 46, no. 5, pp. 1575–1617, Dec. 1991, doi: 10.1111/J.1540-6261.1991.TB04636.X.

[9] B. G. Malkiel, “The Efficient Market Hypothesis and Its Critics,” *Journal of Economic Perspectives*, vol. 17, no. 1, pp. 59–82, Dec. 2003, doi: 10.1257/089533003321164958.

[10] R. Arévalo, J. García, F. Guijarro, and A. Peris, “A dynamic trading rule based on filtered flag pattern recognition for stock market price forecasting,” *Expert Syst Appl*, vol. 81, pp. 177–192, Sep. 2017, doi: 10.1016/J.ESWA.2017.03.028.

[11] C. H. Park and S. H. Irwin, “WHAT DO WE KNOW ABOUT THE PROFITABILITY OF TECHNICAL ANALYSIS?,” *J Econ Surv*, vol. 21, no. 4, pp. 786–826, Sep. 2007, doi: 10.1111/J.1467-6419.2007.00519.X.

[12] L. A. Teixeira and A. L. I. de Oliveira, “A method for automatic stock trading combining technical analysis and nearest neighbor classification,” *Expert Syst Appl*, vol. 37, no. 10, pp. 6885–6890, Oct. 2010, doi: 10.1016/J.ESWA.2010.03.033.

[13] D. Shah, H. Isah, and F. Zulkernine, “Stock Market Analysis: A Review and Taxonomy of Prediction Techniques,” *International Journal of Financial Studies 2019, Vol. 7, Page 26*, vol. 7, no. 2, p. 26, May 2019, doi: 10.3390/IJFS7020026.

[14] D. Lv, S. Yuan, M. Li, and Y. Xiang, “An Empirical Study of Machine Learning Algorithms for Stock Daily Trading Strategy,” *Math Probl Eng*, vol. 2019, 2019, doi: 10.1155/2019/7816154.

[15] M. N. Vafopoulos, “Financial Volatility Forecasting.” Sep. 17, 2000. Accessed: Jan. 23, 2023. [Online]. Available: https://papers.ssrn.com/abstract=1887544

[16] M. Hiransha, E. A. Gopalakrishnan, V. K. Menon, and K. P. Soman, “NSE Stock Market Prediction Using Deep-Learning Models,” *Procedia Comput Sci*, vol. 132, pp. 1351–1362, Jan. 2018, doi: 10.1016/J.PROCS.2018.05.050.

[17] W. Huang, Y. Nakamori, and S. Y. Wang, “Forecasting stock market movement direction with support vector machine,” *Comput Oper Res*, vol. 32, no. 10, pp. 2513–2522, Oct. 2005, doi: 10.1016/J.COR.2004.03.016.

[18] J. Chen, “SVM application of financial time series forecasting using empirical technical indicators,” *ICINA 2010 - 2010 International Conference on Information, Networking and Automation, Proceedings*, vol. 1, 2010, doi: 10.1109/ICINA.2010.5636430.

[19] B. Billah, M. L. King, R. D. Snyder, and A. B. Koehler, “Exponential smoothing model selection for forecasting,” *Int J Forecast*, vol. 22, no. 2, pp. 239–247, Apr. 2006, doi: 10.1016/J.IJFORECAST.2005.08.002.

[20] A. A. Adebiyi, A. O. Adewumi, and C. K. Ayo, “Stock price prediction using the ARIMA model,” *Proceedings - UKSim-AMSS 16th International Conference on Computer Modelling and Simulation, UKSim 2014*, pp. 106–112, 2014, doi: 10.1109/UKSIM.2014.67.

[21] S. Shen, H. Jiang, and T. Zhang, “Stock Market Forecasting Using Machine Learning Algorithms”.

[22] M. Ballings, D. van den Poel, N. Hespeels, and R. Gryp, “Evaluating multiple classifiers for stock price direction prediction,” *Expert Syst Appl*, vol. 42, no. 20, pp. 7046–7056, Nov. 2015, doi: 10.1016/J.ESWA.2015.05.013.

[23] R. Dash and P. K. Dash, “A hybrid stock trading framework integrating technical analysis with machine learning techniques,” *The Journal of Finance and Data Science*, vol. 2, no. 1, pp. 42–57, Mar. 2016, doi: 10.1016/J.JFDS.2016.03.002.

[24] S. Saha, B. Pilani, B. Goa, S. Basak, S. Dey, and Y. Kumar, “Forecasting to Classification: Predicting the direction of stock market price using Xtreme Gradient Boosting”, doi: 10.13140/RG.2.2.15294.48968.

[25] T. Fischer and C. Krauss, “Deep learning with long short-term memory networks for financial market predictions,” *Eur J Oper Res*, vol. 270, no. 2, pp. 654–669, Oct. 2018, doi: 10.1016/J.EJOR.2017.11.054.

[26] S. T. A. Niaki and S. Hoseinzade, “Forecasting S&P 500 index using artificial neural networks and design of experiments,” *Journal of Industrial Engineering International*, vol. 9, no. 1, pp. 1–9, Dec. 2013, doi: 10.1186/2251-712X-9-1/TABLES/4.

[27] T. L. Chen and F. Y. Chen, “An intelligent pattern recognition model for supporting investment decisions in stock market,” *Inf Sci (N Y)*, vol. 346–347, pp. 261–274, Jun. 2016, doi: 10.1016/J.INS.2016.01.079.

[28] A. Mittal and A. Goel, “Stock Prediction Using Twitter Sentiment Analysis”.

[29] I. Svetunkov, *Forecasting and Analytics with ADAM*. Lancaster, UK, 2022. Accessed: Jan. 23, 2023. [Online]. Available: https://openforecast.org/adam/

[30] I. Svetunkov and F. Petropoulos, “Old dog, new tricks: a modelling view of simple moving averages,” *Int J Prod Res*, vol. 56, no. 18, pp. 6034–6047, Sep. 2018, doi: 10.1080/00207543.2017.1380326.

[31] R. J. Hyndman and G. Athanasopoulos, “Forecasting: Principles and Practice.” OTexts, 2018. Accessed: Jan. 23, 2023. [Online]. Available: https://research.monash.edu/en/publications/forecasting-principles-and-practice-2

[32] C. C. Holt, “Forecasting seasonals and trends by exponentially weighted moving averages,” *Int J Forecast*, vol. 20, no. 1, pp. 5–10, Jan. 2004, doi: 10.1016/J.IJFORECAST.2003.09.015.

[33] P. R. Winters, “Forecasting Sales by Exponentially Weighted Moving Averages,” *Manage Sci*, vol. 6, no. 3, pp. 324–342, Apr. 1960, doi: 10.1287/MNSC.6.3.324.

[34] M. P. LaValley, “Logistic Regression,” *Circulation*, vol. 117, no. 18, pp. 2395–2399, May 2008, doi: 10.1161/CIRCULATIONAHA.106.682658.

[35] “What is Logistic regression? | IBM.” https://www.ibm.com/topics/logistic-regression

[36] G. Biau and G. B. Fr, “Analysis of a Random Forests Model,” *Journal of Machine Learning Research*, vol. 13, pp. 1063–1095, 2012.

[37] “Random Forest | Introduction to Random Forest Algorithm.” https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[38] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Front Neurorobot*, vol. 7, no. DEC, p. 21, Dec. 2013, doi: 10.3389/FNBOT.2013.00021/BIBTEX.

[39] “How CatBoost Algorithm Works In Machine Learning.” https://dataaspirant.com/catboost-algorithm/

[40] D. Xu and Y. Tian, “A Comprehensive Survey of Clustering Algorithms,” *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, Jun. 2015, doi: 10.1007/S40745-015-0040-1.

[41] “Clustering Algorithms | Machine Learning | Google Developers.” https://developers.google.com/machine-learning/clustering/clustering-algorithms

[42] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, Aug. 1987, doi: 10.1016/0169-7439(87)80084-9.

[43] “Adjusted Closing Price Definition.” https://www.investopedia.com/terms/a/adjusted_closing_price.asp

[44] "Adjusted Closing Price vs Closing Price: Know the Difference." https://www.angelone.in/knowledge-center/share-market/difference-between-closing-price-and-adjusted-closing-price#:~:text=While%20closing%20price%20merely%20refers,accurate%20measure%20of%20stocks%27%20value

[45] N. N. Taleb, "K-Means Stock Clustering Analysis Based on Historical Price Movements and Financial Ratios," p. 519, 2020.

[46] "What is RSI? - Relative Strength Index - Fidelity." https://www.fidelity.com/learning-center/trading-investing/technical-analysis/technical-indicator-guide/RSI

[47] "Average Directional Index (ADX) [ChartSchool]." https://school.stockcharts.com/doku.php?id=technical_indicators:average_directional_index_adx

[48] R. Peachavanish, "Stock Selection and Trading Based on Cluster Analysis of Trend and Momentum Indicators".

[49] D. Yan, G. Zhou, X. Zhao, Y. Tian, and F. Yang, "Predicting stock using microblog moods," *China Communications*, vol. 13, no. 8, pp. 244–257, Aug. 2016, doi: 10.1109/CC.2016.7563727.

[50] Z. Jin, Y. Yang, and Y. Liu, "Stock closing price prediction based on sentiment analysis and LSTM," *Neural Comput Appl*, vol. 32, no. 13, pp. 9713–9729, Jul. 2020, doi: 10.1007/S00521-019-04504-2/FIGURES/11.

[51] W. Lu, J. Li, Y. Li, A. Sun, and J. Wang, "A CNN-LSTM-based model to forecast stock prices," *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/6622927.

[52] S. Borovkova and I. Tsiamas, "An ensemble of LSTM neural networks for high-frequency stock market classification," *J Forecast*, vol. 38, no. 6, pp. 600–619, Sep. 2019, doi: 10.1002/FOR.2585.