ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Μεταπτυχιακό Πρόγραμμα Σπουδών

Διοίκηση Επιχειρήσεων

Master in Business Administration

Μεταπτυχιακή Διπλωματική Εργασία

Συγκριτική αξιολόγηση μεθόδων μηχανικής μάθησης για την εκτίμηση του κινδύνου πτώχευσης επιχειρήσεων

Comparative evaluation of machine learning methods for distance to default estimation

Νικόλαος Ανδρεουλάκης

Χανιά, 2025

Η παρούσα διπλωματική εργασία εκπονήθηκε για την απόκτηση του Διπλώματος Μεταπτυχιακών Σπουδών στη «Διοίκηση Επιχειρήσεων - Master in Business Administration» (ειδίκευση «Χρηματοοικονομική»), που απονέμει η Σχολή Μηχανικών Παραγωγής και Διοίκησης του Πολυτεχνείου Κρήτης.

Εγκρίθηκε την 25-9-2025 από την εξεταστική επιτροπή:

1. Μιχάλης Δούμπος, Καθηγητής
2. Κωνσταντίνος Ζοπουνίδης, Καθηγητής
3. Γεώργιος Ατσαλάκης, Αναπληρωτής Καθηγητής

Περίληψη

Στην παρούσα διπλωματική εργασία αναλύεται το θεωρητικό πλαίσιο του μοντέλου αγοράς Black-Scholes-Merton, το οποίο χρησιμοποιείται ευρέως για την εκτίμηση της απόστασης από ασυνέπεια (distance to default) μιας επιχείρησης, ως μέτρο του πιστωτικού τής κινδύνου. Αρχικά παρουσιάζονται τα βασικά χαρακτηριστικά του μοντέλου, καθώς και οι κύριες παραδοχές που το συνοδεύουν. Στη συνέχεια εξετάζονται επεκτάσεις και εναλλακτικά μοντέλα που έχουν προταθεί στη διεθνή βιβλιογραφία, με στόχο τη βελτίωση της ακρίβειας και της ρεαλιστικότητας της εκτίμησης του κινδύνου αθέτησης. Σε δεύτερο στάδιο, η εργασία εστιάζει στη χρήση μοντέλων μηχανικής μάθησης για την εκτίμηση της απόστασης από ασυνέπεια. Για τους σκοπούς της εμπειρικής ανάλυσης χρησιμοποιείται δείγμα δεδομένων εισηγμένων επιχειρήσεων από την ευρωπαϊκή αγορά. Περιγράφεται η διαδικασία προεπεξεργασίας των δεδομένων, η επιλογή χαρακτηριστικών και η μέθοδολογία εκπαίδευσης και αξιολόγησης των αλγορίθμων. Δίνεται έμφαση όχι μόνο στην προβλεπτική ικανότητα των μοντέλων, αλλά και στη διαφάνεια και ερμηνευσιμότητά τους, μέσω της εφαρμογής του αλγόριθμου SHAP, ο οποίος επιτρέπει την αποσαφήνιση της συμβολής κάθε μεταβλητής στην τελική πρόβλεψη. Η μελέτη καταλήγει στο συμπέρασμα ότι ο συνδυασμός παραδοσιακών χρηματοοικονομικών μοντέλων με μεθόδους μηχανικής μάθησης μπορεί να προσφέρει πιο ολοκληρωμένα εργαλεία εκτίμησης πιστωτικού κινδύνου, τα οποία είναι ταυτόχρονα ακριβή και ερμηνεύσιμα, συμβάλλοντας έτσι στη λήψη καλύτερα τεκμηριωμένων χρηματοοικονομικών αποφάσεων.

Περιεχόμενα

Κατάλογος πινάκων

Κατάλογος σχημάτων

1. Εισαγωγή

Η ανάπτυξη των χρηματοοικονομικών αγορών κατά τον 20ό αιώνα συνοδεύτηκε από τη ραγδαία αύξηση της σημασίας των παράγωγων προϊόντων. Τα δικαιώματα προαίρεσης (options), ως ένα από τα πιο διαδεδομένα είδη παραγώγων, προσέλκυσαν ιδιαίτερο ενδιαφέρον από τους επενδυτές θεσμικούς και μη, όσο και από την ακαδημαϊκή έρευνα, καθώς παρέχουν τη δυνατότητα αποτελεσματικής διαχείρισης του χαρτοφυλακίου, της αντιστάθμισης κινδύνου ενώ μπορούν να συνεισφέρουν στην εύρεση επενδυτικών ευκαιριών. Ωστόσο, η εύλογη αποτίμηση αυτών των δικαιωμάτων αποτελούσε για πολλά χρόνια ένα από τα σημαντικότερα προβλήματα της χρηματοοικονομικής θεωρίας, αφού δεν υπήρχε ένα καθολικά αποδεκτό και μαθηματικά αυστηρό πλαίσιο που να επιτρέπει την συνεπή τιμολόγηση τους. Η ανάγκη αυτή καλύφθηκε με την εργασία των Black και Scholes (1973), η οποία εμπλουτίστηκε θεωρητικά από τον Merton (1973), οδηγώντας στη διαμόρφωση του γνωστού πλέον ως μοντέλου Black–Scholes–Merton (BSM). Το μοντέλο αυτό εισήγαγε μια ριζοσπαστική καινοτομία: τη χρήση στοχαστικών διαφορικών εξισώσεων για την περιγραφή της πορείας των τιμών των υποκείμενων αξιών, με την υπόθεση ότι ακολουθούν γεωμετρική κίνηση Brown (geometric Brownian motion). Με αυτόν τον τρόπο, κατέστη δυνατή η ανάπτυξη μιας λύσης κλειστού τύπου (closed-form solution) για την τιμολόγηση των δικαιωμάτων προαίρεσης, κάτι που μέχρι τότε θεωρούνταν εξαιρετικά δύσκολο. Η συνεισφορά του Merton (1973) στο μοντέλο υπήρξε καθοριστική για την αποτίμηση του εταιρικού πιστωτικού κινδύνου. Η προσέγγιση της αναγνώρισης των κεφαλαίων των επιχειρήσεων μέσα από το μοντέλο αυτό ως ένα δικαίωμα προαίρεσης έδωσε διαφορετική δυναμική στην μέχρι εκείνη την εποχή θεωρία αποτίμησης του πιστωτικού κινδύνου. Η έρευνα στο συγκεκριμένο πεδίο βασίζονταν στις πιστοληπτικές διαβαθμίσεις, μίξη μοντέλων και γνώμης ειδικών (expert systems), καθώς και στατιστικά μοντέλα με βάση τα ιστορικά δεδομένα. Η απόρροια του μοντέλου BSM ήταν η δημιουργία της απόστασης από ασυνέπεια (distance to default, DtD), ένα μέτρο που ενσωματώνει μεγάλη πληροφόρηση για τον υποκείμενο πιστωτικό κίνδυνο. Αυτό βασίζεται στις υποθέσεις των Black–Scholes–Merton σχετικά με την υπόθεση της αποτελεσματικής αγοράς (efficient market hypothesis) αλλά και την ενσωμάτωση στην απόσταση από ασυνέπεια την ιδιαιτερότητα της αγοράς να σκεφτεται τα μελλοντικά αποτελέσματα (forward looking prices). Η δημοσίευση του μοντέλου αποτέλεσε σημείο καμπής

για τη χρηματοοικονομική επιστήμη και οδήγησε στη θεμελίωση της σύγχρονης χρηματοοικονομικής μηχανικής. Η συμβολή του αναγνωρίστηκε με την απονομή του Βραβείου Νόμπελ Οικονομικών στον Scholes και τον Merton το 1997. Από την εποχή εκείνη μέχρι σήμερα, το μοντέλο Black–Scholes–Merton συνεχίζει να αποτελεί σημείο αναφοράς, παρά την ύπαρξη πιο σύνθετων και προσαρμοσμένων υποδειγμάτων.

Η παρούσα διπλωματική εργασία επικεντρώνεται στη παρουσίαση του μοντέλου Black–Scholes–Merton για την εκτίμηση της απόστασης από ασυνέπεια καθώς επίσης και την πρόβλεψη αυτής μέσων των μοντέρνων μεθόδων μηχανικής μάθησης. Τα τελευταία χρόνια, η ραγδαία ανάπτυξη της επιστήμης δεδομένων και των μεθόδων μηχανικής μάθησης έχει ανοίξει νέους δρόμους για την εκτίμηση χρηματοοικονομικών δεδομένων. Τα μοντέλα μηχανικής μάθησης αξιοποιούν μεγάλα σύνολα δεδομένων και αλγορίθμους πρόβλεψης προκειμένου να ανιχνεύσουν μοτίβα και μη γραμμικές σχέσεις που δεν είναι εύκολο να περιγραφούν με παραδοσιακές μεθόδους. Η χρήση τεχνικών όπως η support vector regression, τα νευρωνικά δίκτυα, ή οι αλγόριθμοι ensemble έχουν χρησιμοποιηθεί εκτενώς στη βιβλιογραφία αποδεικνύοντας ότι βελτιώνουν την ακρίβεια πρόβλεψης ανταποκρινόμενα στις δυναμικές συνθήκες των σύγχρονων αγορών. Μέσω συγκριτικής ανάλυσης επιδιώκεται να αποτυπωθεί η προσφορά των σύγχρονων αλγορίθμων, ώστε να αναδειχθεί κατά πόσο οι μέθοδοι μηχανικής μάθησης μπορούν να ξεπεράσουν το όριο της κλασικής γραμμικής παλινδρόμησης και να προσφέρουν πιο αξιόπιστα αποτελέσματα σε πραγματικά δεδομένα αγοράς. Στη δεύτερη ενότητα αναλύονται το βασικό μοντέλο BSM και ο ορισμός της απόστασης από ασυνέπεια καθώς και επεκτάσεις τους. Στη τρίτη ενότητα αναπτύσονται τα μοντέλα μηχανικής μάθησης που χρησιμοποιήθηκαν για την πρόβλεψη της απόστασης από ασυνέπεια. Στη τέταρτη ενότητα αναλύονται τα δεδομένα και οι μεταβλητές επιλογής καθώς και η μεθοδολογία που προτιμήθηκε για την ανάλυση. Στην πέμπτη ενότητα παρουσιάζονται τα μέτρα επίδοσης, τα αποτελέσμα της συγκριτικής ανάλυσης και μια μεθοδολογία για την ερμηνεία των αποτελεσμάτων μηχανικής μάθησης.

2. Θεωρία

2.1 Black Scholes Merton

Οι Black και Scholes (1973) και Merton (1973) έγραψαν για την τιμολόγηση των δικαιωμάτων προαίρεσης (options) που αργότερα θα τιμηθεί με το βραβείο Nobel (1997). Ο Merton (1974) αργότερα επέκτεινε την ιδέα εφαρμόζοντας το ίδιο μοντέλο για την αξιολόγηση του επιχειρηματικού χρέους (corporate debt). Η ιδέα είναι ότι το ενεργητικό της επιχείρησης ακολουθεί τη γεωμετρική κίνηση Brown της μορφής:

$$dV_A = \mu V_A dt + \sigma_A V_A dW,$$

Η σχέση αυτή είναι μια διαφορική εξίσωση σύμφωνα με την οποία η οριακή μεταβολή του ενεργητικού εξαρτάται από την οριακή αναμενόμενη απόδοση(drift, μ) του στοιχείου αυτού και την οριακή μεταβλητότητα (σ) συν τη διαδικασία Wiener.

Αν ορίσουμε E και V_A την αξία των ιδίων κεφαλαίων και ενεργητικού αντίστοιχα και X_t το χρέος που παρουσιάζεται στον ισολογισμό στο χρόνο t και έχει την έννοια της τιμής άσκησης του δικαιώματος στα πλαίσια τιμολόγησης ενός δικαιώματος αγοράς, τότε παίρνουμε την εξίσωση των Black-Scholes-Merton για την αποτίμηση της αξίας E με βάση τον παρακάτω τύπο:

$$E = V_A N(d_1) - X e^{-rT} N(d_2), \tag{2.1}$$

όπου

$$d_1 = \frac{\ln(V_A/X) + \left(r + \frac{1}{2}\sigma_A^2\right) T}{\sigma_A \sqrt{T}},$$
$$d_2 = d_1 - \sigma_A \sqrt{T},$$

και $\mathcal{N}(\cdot)$ η τυπική κανονική κατανομή.

Κάποιες παρατηρήσεις για την εξίσωση (2.1). Η σχέση αναδεικνύει την υπολλειματική αξία που λαμβάνουν οι μέτοχοι της επιχείρησης αν αφαιρεθεί από τον ενεργητικό της το χρέος που χρειάστηκε για να χρηματοδοτηθεί. Επιπλέον ως ένα δικαίωμα αγοράς αυτό το χρέος πρέπει να προεξοφληθεί στο χρόνο εκείνο για τον οποίο γίνεται η αποτίμηση. Περαιτέρω έχουμε τα $N(d_1)$ και $N(d_2)$ τα οποία αποτελούν τις πιθανότητες οι οποίες αυξάνουν τα ποσά εκείνα που κρίνουν αν θα ασκηθεί το συμβόλαιο. Σημαντικός παράγοντας στην παραπάνω σχέση είναι ο λόγος μεταξύ ενεργητικού και υποχρεώσεων V_A/X καθώς αυτός αυξάνει τον αριθμητή του κλάσματος d_1 ενώ το d_2 αυξάνει σε μικρότερο βαθμό. Επιπλέον η άυξηση της μεταβλητότητας αυξάνει τον παρονομαστή στο d_1 και συνολικά το κλάσμα είναι μεγαλύτερο απ'ότι στο d_2. Ο χρόνος, T, οπου προεξοφλείται το X, μειώνει σε μεγαλύτερο βαθμό το δεύτερο μέρος της εξίσωσης (2.1) καθώς αυξάνεται. Τέλος η εξίσωση των Black-Scholes-Merton δεν εξαρτάται από την αναμενόμενη απόδοση των στοιχείων του ενεργητικού, αλλά μόνο από το επιτόκιο μηδενικού κινδύνου (risk free rate).

2.1.1 Απόσταση από την ασυνέπεια

Η πιθανότητα πτώχευσης μιας επιχείρησης συμβαίνει όταν η μελλοντική αξία του ενεργητικού $(t+T)$ είναι μικρότερη από την αξία του χρέους τη στιγμή που εξετάζουμε. Έτσι μπορούμε να την ορίσουμε ως:

$$P_{\text{def},t} = \text{Prob}\left(V_{A,t+T} \leq X_t \mid V_{A,t}\right) = \text{Prob}\left(\ln(V_{A,t+T}) \leq \ln(X_t) \mid V_{A,t}\right). \tag{2.2}$$

Όπως αναφέρθηκε, η αξία του ενεργητικού ακολουθεί τη γεωμετρική κίνηση Brown, επομένως μια οποιαδήποτε στιγμή $t+T$, η αξία του ενεργητικού δίνεται από:

$$\ln(V_{A,t+T}) = \ln(V_{A,t}) + \left(\mu - \frac{\sigma_A^2}{2}\right)T + \sigma_A\sqrt{T}\varepsilon_{t+T}, \tag{2.3}$$

Στη σχέση αυτή, η αξία του ενεργητικού κατά τη χρονική στιγμή $t+T$ εξαρτάται από την αξία τη στιγμή t, την αναμενόμενη απόδοση, τη μεταβλητότητα και την τιμή ε_{t+T} της διαδικασίας Wiener:

$$\varepsilon_{t+T} = \frac{W(t+T) - W(t)}{\sqrt{T}}, \quad \text{και} \quad \varepsilon_{t+T} \sim \mathcal{N}(0,1).$$

Με βάση τη σχέση (2.3), η πιθανότητα πτώχευσης της σχέσης (2.2), γράφεται ως εξής:

$$P_{\text{def},t} = \text{Prob}\left(\ln(V_{A,t}) - \ln(X_t) + \left(\mu - \frac{\sigma_A^2}{2}\right)T + \sigma_A\sqrt{T}\varepsilon_{t+T} \leq 0\right)$$

$$P_{\text{def},t} = \text{Prob}\left(\frac{\ln\left(\frac{V_{A,t}}{X_t}\right) + \left(\mu - \frac{\sigma_A^2}{2}\right)T}{\sigma_A\sqrt{T}} \leq -\varepsilon_{t+T}\right).$$

Σύμφωνα με τη σχέση αυτή, η απόσταση από την ασυνέπεια (distance to default, DtD), ορίζεται ως:

$$DtD_t = \frac{\ln(V_{A,t}/X_t) + \left(\mu - \frac{1}{2}\sigma_A^2\right)T}{\sigma_A\sqrt{T}}. \tag{2.4}$$

Η σχέση (2.4) μας λέει πόσες τυπικές αποκλίσεις πρέπει να αποκλίνει το ενεργητικό από το μέσο του για να προκληθεί πιστωτικό γεγονός. Πτώχευση προκαλείται όταν ο λόγος ενεργητικού προς χρέος είναι μικρότερος από 1 ή όταν ο λογάριθμός τους είναι αρνητικός καθώς οι άλλες παράμετροι δεν είναι ικανές να επηρεάσουν το αποτέλεσμα δραματικά. Το DtD σε αυτή τη περίπτωση λαμβάνει αρνητικές τιμές χωρίς όμως να σημαίνει εμπειρικά ότι η επιχείρηση έχει πτωχεύσει καθώς ειναι δυνατόν να μπορούν να αναχρηματοδοτήσουν τα χρέη τους. Αξίζει να σημειωθεί ότι ενώ για την αποτίμηση της σχέσης (2.1) λαμβάνεται υπ'όψιν το επιτόκιο μηδενικού κινδύνου, το DtD εξαρτάται από το μ καθώς η μελλοντική αξία του ενεργητικού εξαρτάται από την αναμενόμενη απόδοσή του.

2.1.2 Υπολογισμός απόστασης από την ασυνέπεια

Από την κεντρική εξίσωση (2.1) των Black-Scholes-Merton μπορούμε να κάνουμε δύο παραδοχές. Αρχικά, η αξία των ιδίων κεφαλαίων της επιχείρησης σχετίζεται άμεσα από την αξία του ενεργητικού της. Επιπλέον είναι σαφές ότι και η μεταβλητότητα θα εξαρτάται και αυτή από τη μεταβλητότητα του ενεργητικού το οποίο μπορούμε να δούμε μέσα από τους όρους d_1 και d_2. Το ενδιαφερόν στρέφεται στο να δούμε πως εξελίσσεται το E μέσα από την εξίσωση $E = E(V,t)$. Λόγω της υπόθεσης του Merton ότι το V_A ακολουθεί τη γεωμετρική κίνηση Brown, μπορούμε να εφαρμόσουμε το λήμμα του Itô. Εφαρμόζοντάς το στη μορφή του E, το λήμμα του Itô σε ανάπτυγμα Taylor παίρνουμε την εξίσωση:

$$dE = \left(\frac{\partial E}{\partial t} + \mu V_A\frac{\partial E}{\partial V_A} + \frac{1}{2}\sigma_A^2 V_A^2\frac{\partial^2 E}{\partial V_A^2}\right)dt + \sigma_A V_A\frac{\partial E}{\partial V_A}dW$$

Το πρώτο μέρος αυτής της στοχαστικής διαφορικής εξίσωσης είναι το λεγόμενο drift term που εξηγεί πως μεταβάλλεται η απόδοση στον χρόνο και το δεύτερο μέρος λέγεται όρος διάχυσης που εξηγεί την στιγμιαία μεταβλητότητα του υποκείμενου. Επειδή αυτός ο όρος εξηγεί τη μεταβλητότητα το οποίο είναι το ζητούμενο μπορούμε να θέσουμε:

$$\sigma_E E = \sigma_A V_A \frac{\partial E}{\partial V_A}$$

και αν λύσουμε ως προς σ_E,

$$\sigma_E = \frac{V_A}{E} \frac{\partial E}{\partial V_A} \sigma_A$$

Ο όρος $\frac{\partial E}{\partial V_A}$ μπορεί να βρεθεί παίρνοντας τη μερική παράγωγο της (2.1) και έτσι καταλήγουμε:

$$\sigma_E = \frac{V_A}{E} N(d_1) \sigma_A, \tag{2.5}$$

η οποία μας δείχνει τη σχέση μεταξύ σ_E και σ_A. Για το λήμμα του Itô αλλά και περισσότερες λεπτομέρειες του μοντέλου Black Scholes Merton μπορεί να βρεθεί στον Hull (2009).

Για τον υπολογισμό της απόστασης από την ασυνέπεια (2.4), κάποια στοιχεία είναι δημοσίως διαθέσιμα όπως το χρέος και ο χρονικός ορίζοντας της εκτίμησης, ενώ το ενεργητικό, η μεταβλητότητά του και η μέση απόδοση, δεν είναι εύκολα παρατηρήσιμα. Μέσα από τον συσχετισμό των (2.1) και (2.5) μπορούμε μέσα από μια επαναληπτική διαδικασία να βρούμε τα στοιχεία που μας λείπουν, όπως έχουν πραγματοποιήσει στη βιβλιογραφία για παράδειγμα οι Vassalou και Xing (2004) και Bharath και Shumway (2008).

2.2 Το μοντέλο των Vassalou & Xing

Για τον υπολογισμό της απόστασης από την ασυνέπεια, οι Vassalou και Xing (2004) χρησιμοποίησαν την παρακάτω επαναληπτική μέθοδο:

1. Εκτίμηση του σ_E των τελευταίων 12 μηνών ως αρχική τιμή του σ_A.
2. Για κάθε χρηματιστηριακή ημέρα του προηγούμενου έτους υπολογίζουμε το V_A με τη σχέση (2.1) χρησιμοποιώντας το V_E εκείνης της ημέρας.

3. Έχοντας αποκτήσει ημερήσια δεδομένα για το V_A υπολογίζεται η τυπική απόκλιση η οποία θα χρησιμοποιηθεί ως νέο σ_A σε νέα επανάληψη.

4. Η επανάληψη σταματάει όταν δύο συνεχόμενες τιμές σ_A συγκλίνουν με μια διαφορά μικρότερη του 10E-4.

5. Όταν βρεθεί το τελικό σ_A λύνουμε ξανά την (2.1) για να βρεθεί το τελικό V_A.

Η παραπάνω διαδικασία επαναλαμβάνεται μηνιαία για τον υπολογισμό της μέσης απόδοσης (μ) του $\ln V_A$. Χρησιμοποιείται το επιτόκιο μηδενικού κινδύνου (r) στο τέλος κάθε μήνα. Τέλος υπολογίζεται το DtD αντικαθιστώντας τα εκτιμημένα στοιχεία στην (2.4).

2.3 Το μοντέλο των Bharath & Shumway

Για τον υπολογισμό του DtD οι Bharath και Shumway (2008) ακολούθησαν την ίδια επαναληπτική διαδικασία με μικρές διαφορές:

1. Αρχική τιμή του $\sigma_A = \sigma_E[E/(E+F)]$.

2. Αντικατάσταση στην (2.1) για την εύρεση του V_A κάθε ημέρα του προηγούμενου έτους.

3. Υπολογισμός της λογαριθμικής απόδοσης του ενεργητικού κάθε ημέρα και εύρεση της μέση απόδοσης αυτής, μ και μεταβλητότητας σ_A.

4. Η επανάληψη σταματάει όταν δύο συνεχόμενες τιμές σ_A συγκλίνουν με μια διαφορά μικρότερη του 10E-3.

Οι διαφορές στη μέθοδο επίλυσης δεν είναι γνωστό αν αυξάνουν το προβλεπτικό αποτέλεσμα της πιθανότητα πτώχευσης καθώς δεν ανέφεραν γιατί επέλεξαν να μην ακολουθήσουν τη μέθοδο των Vassalou και Xing (2004). Αντιθέτως επιχειρηματολόγησαν κατασκευάζοντας ένα μοντέλο που δεν χρειάζεται επίλυση όπως ήταν μέχρι εκείνη τη στιγμή γνωστό στη βιβλιογραφία για τα μοντέλα DtD. Το ονόμασαν 'Naive DD'. Συγκεκριμένα εάν ορίσουμε το ονομαστικό χρέος της επιχείρησης ότι εξισώνεται με την τιμή του χρέους της αγοράς (market value of debt) τότε naïve $D = X_t$. Υποθέτοντας ότι οι επιχειρήσεις που έχουν πολυ μικρή "απόσταση" από την πτώχευση έχουν χρέος με μεγαλύτερο κίνδυνο, καθώς και ότι το ρίσκο της κεφαλαιοποίησής τους συσχετίζεται με το ρίσκο του χρέους τότε μπορούμε να πούμε:

$$\text{naïve } \sigma_D = 0.05 + 0.25\,\sigma_E$$

Όπου η μεταβλητότητα του χρέους $naive\ \sigma_D$ εξαρτάται 25% από τον κίνδυνο της κεφαλαιοποίησης συν 5% η αναμενόμενη μεταβλητότητα της αγοράς (term structure volatility). Με αυτές τις υποθέσεις υπολογίζουν τη μεταβλητότητα του ενεργητικού ως εξής:

$$\begin{aligned}\text{naïve } \sigma_A &= \frac{E}{E + \text{naïve } D}\,\sigma_E + \frac{\text{naïve } D}{E + \text{naïve } D}\,\text{naïve } \sigma_D \\ &= \frac{E}{E+F}\,\sigma_E + \frac{F}{E+F}\,(0.05 + 0.25\,\sigma_E).\end{aligned}$$

Σύμφωνα με τη σχέση αυτή η μεταβλητότητα εξαρτάται από το ποσοστό μεταξύ χρέους και ενεργητικού προς το συνολικό μείγμα χρηματοδότησης καθώς και τη μεταβλητότητα που παρουσιάζουν ξεχωριστά. Περαιτέρω οι Bharath και Shumway (2008) θεώρησαν ότι για τον υπολογισμό της μέσης απόδοσης του ενεργητικού μπορεί να θεωρηθεί η μέση απόδοση της μετοχής το προηγούμενο έτος, naive $\mu = r_{it-1}$. Με αυτό το τρόπο μπορούν να αποκτήσουν σημαντική πληροφορία που αλλιώς θα έπρεπε να υπολογιστεί από την παραπάνω επαναληπτική διαδικασία που αναφέραμε.

$$\text{naive DD} = \frac{\ln(E + X_t/X_t) + \left(r_{it-1} - \frac{1}{2}\,\text{naive } \sigma_A^2\right) T}{\text{naive } \sigma_A \sqrt{T}} \tag{2.6}$$

Με τις παραπάνω τροποποιήσεις το naive DD μπορεί να υπολογιστεί άμεσα χωρίς καμμία περίπλοκη επαναληπτική διαδικασία. Οι τροποποιήσεις αυτές μπορεί να είναι απλοϊκές, η μέθοδος όμως, πράγματι, είχε πολύ καλά αποτελέσματα. Στις παλινδρομήσεις που υπολόγισαν το naive DD κατέφερε να βγει στατιστικά σημαντικό σε σχέση με το αντίστοιχο DtD από την επαναληπτική μέθοδο και προσέφερε μεγαλύτερη προβλεπτική ικανότητα. Το συμπέρασμα που βγήκε από αυτή την έρευνα ήταν ότι κυρίαρχη σημασία στη προβλεπτική ικανότητα των μοντέλων DtD ήταν όχι ο ακριβής υπολογισμός των μεταβλητών μέσω επαναληπτικών μεθόδων αλλά η συναρτησιακή σχέση του μοντέλου του Merton.

2.4 Η μεθοδολογία της Moody's KMV

Η KMV Corporation ξεκίνησε αρχικά στον κλάδο της χρηματοοικονομικής τεχνολογίας παρουσιάζοντας διάφορα μοντέλα εκτίμησης πιστωτικού κινδύνου και ανάλυσης χαρτοφυλακίου ώσπου εξαγοράστηκε στις αρχές του τις δεκαετίας του 2000 από τη Moody's, ένα από τους μεγαλύτερους οίκους πιστοληπτικής αξιολόγησης. Ανάμεσα στα μοντέλα που δημιούργησε ήταν και το «VK model», μία επέκταση του μοντέλου του Merton. Πλέον είναι γνωστό ως Moody's KMV και είναι διαθέσιμο στο Moody's Analytics (ως Credit Monitor), σε όλους τους πελάτες τής παγκοσμίως, που

περιλαμβάνουν τράπεζες και άλλα χρηματοπιστωτικά ιδρύματα. Παρ΄ότι το μοντέλο έχει συγκεκριμένες παραμετροποιήσεις που το διαφοροποιεί από το μοντέλο του Merton έχουν γίνει αρκετές προσπάθειες για την αποκρυπτογράφηση του.

Οι κύριες διαφοροποιήσεις του όπως παρουσιάζεται από τους Bohn και Crosbie (2002) και Kealhofer (2003) στο άρθρα που έχουν εκδόσει υπό την αιγίδα της Moody's:

1. Διαφορετική μοντελοποίηση που περιλαμβάνει πέντε διαφορετικά είδη χρέους όπως long-term & short-term, convertibles, προνομιούχες και κοινές μετοχές.
2. Μπεϋζιανή μεθοδολογία στην εκτίμηση της μεταβλητότητας του ενεργητικού λαμβάνοντας υπόψιν τη χώρα, κλάδο δραστηριότητας και μέγεθος της επιχείρησης.
3. Χρησιμοποίησαν την πραγματική κατανόμη των πτωχεύσεων που κατέχουν από μια μεγάλη βάση δεδομένων αντί για την κανονική κατανομή που αναφέρεται στη θεωρία του Merton.

Παρ΄όλα αυτά η επαναληπτική μεθοδολογία εκτίμησης της μεταβλητότητας του ενεργητικού φαίνεται να είναι παρόμοια με αυτή που χρησιμοποιούν οι Vassalou και Xing (2004) και Bharath και Shumway (2008) χωρίς όμως η Moody's να δίνει περισσότερες λεπτομέρειες. Μία ακόμα κριτική που έκαναν στο μοντέλο του Merton ήταν ότι είναι αδύνατη η σταθερή εκτίμηση του DtD βασιζόμενοι στην ταυτόχρονη επίλυση των (2.1) και (2.5). Όπως αναφέρουν η μόχλευση στην αγορά είναι τόσο μεγάλη που είναι αδύνατη η σταθερή εκτίμηση της μεταβλητότητας μέσα από αυτές τις δύο σχέσεις. Η σχέση (2.5) μπορεί να γραφεί αλλιώς ως:

$$\sigma_E = \frac{V_A}{E} N(d_1)\sigma_A \Rightarrow$$
$$\sigma_A = \frac{E}{V_A}\frac{\sigma_E}{N(d_1)}$$

Το πρώτο κλάσμα είναι η αντίστροφη μόχλευση της επιχείρησης. Σε μια δυσμενή κατάσταση με υψηλή μεταβλητότητα θα είχαμε, με έναν απότομο ρυθμό αύξησης της μόχλευσης ($\uparrow V_A$) υποεκτίμηση του σ_A. Με έναν απότομο ρυθμό μείωσης της μόχλευσης ($\downarrow V_A$) υπερεκτίμηση του σ_A. Αυτό έχει τα αντίθετα αποτελέσματα με βάση τη λογική για τον πιστωτικό κίνδυνο της επιχείρησης. Με αυτό το τρόπο έχουμε λανθασμένες εκτιμήσεις τόσο για το DtD όσο και για τη πιθανότητα πτώχευσης.

2.5 Credit Research Initiative

Στην παρούσα εργασία χρησιμοποιήθηκε η εκτίμηση της απόστασης από ασυνέπεια (DtD) από το Credit Research Initiative (CRI) του Asian Institute of Digital Finance του National University of Singapore. Πρόκειται για ένα μη κερδοσκοπικό ερευνητικό ινστιτούτο που επικεντρώνεται στην έρευνα πιστωτικού κινδύνου και παρέχει αναλυτικά δεδομένα για εισηγμένες επιχειρήσεις σε παγκόσμια κλίμακα. Το CRI ερευνά την πιθανότητα πτώχευσης των επιχειρήσεων αναπτύσσοντας ένα μοντέλο forward intensity. Η απόσταση από ασυνέπεια σε αυτό το μοντέλο εισάγεται ως ανεξάρτητη μεταβλητή μαζί με άλλες για την αληθοφανή εκτίμηση της πιθανότητας πτώχευσης σε αντίθεση με ότι είδαμε μέχρι στιγμής, όπου η πιθανότητα πτώχευσης μπορούσε να προβλεφθεί με το DtD και την κατανομή των πτωχεύσεων (KMV) ή την κανονική κατανομή. Σύμφωνα με το Credit Research Initiative (2023) ο υπολογισμός βασίζεται σε αυτή την απλούστερη σχέση:

$$DtD_t = \frac{\ln(V_{A,t}/X_t)}{\sigma_A\sqrt{T}} \tag{2.7}$$

όπου $X_t =$ βραχυπρόθεσμες υποχρεώσεις $+\frac{1}{2}$ μακροπρόθεσμες υποχρεώσεις $+\delta$ άλλες υποχρεώσεις. Αρχικά υπολογίζονται η αγοραία αξία του ενεργητικού μέσα από την (2.1) και στη συνέχεια για την εύρεση των παραμέτρων, βελτιστοποιείται η παρακάτω συνάρτηση πιθανοφάνειας:

$$\begin{aligned}\tilde{\mathcal{L}}(\sigma,\delta) = &-\frac{n-1}{2}\log(2\pi) - \frac{1}{2}\sum_{t=2}^{n}\log(\sigma^2 h_t) - \sum_{t=2}^{n}\log\left(\frac{\hat{V}_{A,t}(\sigma,\delta)}{A_t}\right)\\ &-\sum_{t=2}^{n}\log N(d_+) - \frac{1}{2\sigma^2}\left\{\sum_{t=2}^{n}\frac{1}{h_t}\left[\log\left(\frac{\hat{V}_{A,t}(\sigma,\delta)}{A_t}\times\frac{A_{t-1}}{\hat{V}_{A,t-1}(\sigma,\delta)}\right)\right]^2\right.\\ &\left.-\frac{1}{\sum_{t=2}^{n}h_t}\left[\log\left(\frac{\hat{V}_{A,n}(\hat{\sigma},\hat{\delta})}{A_n}\times\frac{A_1}{\hat{V}_{A,1}(\hat{\sigma},\hat{\delta})}\right)\right]^2\right\}.\end{aligned}$$

όπου h_t είναι το $T-t$, και A_t η ονομαστική αξία του ενεργητικού το χρόνο t και εισάγεται για την κανονικοποίηση του εκτιμημένου ενεργητικού $\hat{V}_{A,t}$. Η βελτιστοποίηση γίνεται σε δύο στάδια. Στο πρώτο εκτιμούνται οι παραμέτροι σ, δ. Στο δεύτερο στάδιο γίνεται επανεκτίμηση του σ με σταθερό το δ ως μέση τιμή των παρατηρήσεων ανάμεσα στους κλάδους δραστηριότητας. Οι αστάθειες στους υπολογισμούς του CRI οδήγησαν στο συμπέρασμα ότι πρέπει να οριστεί το $\mu = \frac{\sigma^2}{2}$ και έτσι μπορεί να εξηγηθεί η απλούστερη σχέση της απόστασης από ασυνέπεια (2.7).

2.6 Εμπειρικές συγκριτικές έρευνες

Οι Afik κ.ά. (2016) πραγματοποίησαν μια συγκριτική ανάλυση πολλαπλών μεθόδων που έχουν διερευνηθεί στη βιβλιογραφία. Εξέτασαν τα μοντέλα των Merton, την επαναληπτική μέθοδο της KMV, Bharath & Shumway naive model, μοντέλο DaO, CDLT (Charitou κ.ά., 2013) καθώς και τη δική τους naive μεθοδολογίας. Πέρα από αυτά εξέτασαν πως η επιλογή των παραμέτρων r, σ, μ και διαφορετικά ποσοστά μακροχρόνιου χρέους στο επίπεδο ασυνέπειας (X_t), μπορούν να επηρεάσουν τη προβλεπτική ικανότητα των μοντέλων στην εκτίμηση της πιθανότητας πτώχευσης. Οι Agarwal και Taffler (2008) σύγκριναν δύο μοντέλα που βασίζονται στη θεωρία των Black-Scholes-Merton και συγκεκριμένα των Hillegeist κ.ά. (2004) και Bharath και Shumway (2008) καθώς και το μοντέλο Z-Score του Altman (1968) που βασίζεται σε ιστορικά δεδομένα χρηματοοικονομικών δεικτών. Με βάση την ανάλυση τους σε ένα δείγμα επιχειρήσεων του Ηνωμένου Βασιλείου έκριναν ότι το Z-score του Altman είχε μεγαλύτερη προβλεπτική ικανότητα στο μέτρο επίδοσης Area under the Curve (AUC). Αντιθέτως οι Hillegeist κ.ά. (2004) συγκρίνοντας το μοντέλο των Black-Scholes-Merton και των Z-Score (Altman, 1968), O-Score (Ohlson, 1980) έκριναν ότι το μοντέλο Black-Scholes-Merton παρέχει περισσότερη πληροφορία σε σχέση με τα παραδοσιακά μοντέλα των χρηματοοικονομικών δεικτών. Η θέση των ερευνητών στη βιβλιογραφία, δεν είναι ξεκάθαρη για το εάν τα μοντέλα της αγοράς (BSM) υπερτερούν έναντι των παραδοσιακών μοντέλων που βασίζονται σε ιστορικά δεδομένα λογιστικών καταστάσεων, για την εκτίμηση της πιθανότητας πτώχευσης. Η πιθανότητα πτώχευσης μας αφορά καθώς συνδέεται άμεσα με την απόσταση από ασυνέπεια (DtD) στα μοντέλα της αγοράς. Οι Charitou κ.ά. (2013) επέκτειναν το Naive μοντέλο των Bharath & Shumway εισάγοντας τη διανομή μερισμάτων στον υπολογισμό της αναμενόμενης απόδοσης (r) καθώς επίσης και τη δικιά τους εκδοχή για τον υπολογισμό της μεταβλητότητας του ενεργητικού ($\sigma_{V[CDLT]}$), και τα σύγκριναν εφαρμόζοντας ένα Cox hazard μοντέλο.

Σε άλλα άρθρα στη βιβλιογραφία εξετάστηκε η χρησιμότητα της απόστασης από ασυνέπεια ως ανεξάρτητη μεταβλητή σε συνδυασμό με άλλους χρηματοοικονομικούς δείκτες. Τέτοια άρθρα είναι η εκτίμηση της πιθανότητας πτώχευσης των Campbell κ.ά. (2008) με το μοντέλο Logit και στην περίπτωση πρόβλεψης των πιστοληπτικών διαβαθμίσεων των Doumpos κ.ά. (2015) με ένα μοντέλο πολυκριτήριας ανάλυσης. Οι Duffie κ.ά. (2007) δημιούργησαν ένα πολυ-περιοδικό οικονομετρικό μοντέλο για την πρόβλεψη της πιθανότητας πτώχευσης χρησιμοποιώντας την απόσταση από ασυνέπεια ως ανεξάρτητη μεταβλητή μαζί με άλλες, εξερευνώντας τη δυναμική της μέσα στο χρόνο.

3. Μοντέλα Πρόβλεψης

Τα τελευταία χρόνια, η πρόβλεψη της πτώχευσης επιχειρήσεων έχει αναδειχθεί σε ένα πεδίο αυξανόμενου ενδιαφέροντος, τόσο στον ακαδημαϊκό χώρο όσο και στην πράξη, λόγω της ανάπτυξης νέων μεθόδων πρόβλεψης του πιστωτικού κινδύνου. Η ανάγκη για την εύρεση νέων, καινοτόμων και πιο αξιόπιστων εργαλείων αξιολόγησης πιστωτικού κινδύνου δείχνει τη σοβαρότητα με την οποία αντιμετωπίζουν το ζήτημα τα χρηματοπιστωτικά ιδρύματα και άλλοι ενδιαφερόμενοι. Παραδοσιακά, η πρόβλεψη του πιστωτικού κινδύνου, βασιζόταν σε στατιστικές μεθόδους, όπως η λογιστική παλινδρόμηση και η διακριτική ανάλυση. Πλέον οι παραδοσιακές στατιστικές μεθόδοι έχουν ξεπεραστεί καθώς οι ανασκοπήσεις της βιβλιογραφίας συστηματικά αναδεικνύουν την αποτελεσματικότητα των μεθόδων μηχανικής μάθησης αφού προσφέρουν υψηλότερα επίπεδα ακρίβειας των προβλέψεων συνεισφέροντας στη λήψη των χρηματοοικονομικών αποφάσεων. Πέρα από τη πρόβλεψη του πιστωτικού κινδύνου, η μηχανική μάθηση χρησιμοποιείται ευρέως και σε προβλήματα πρόβλεψης τιμών μετοχών, διαχείρισης χαρτοφυλακίου και ανίχνευσης απάτης(fraud detection & qualified audits). Σε αυτή την ενότητα θα παρουσιαστούν τα μοντέλα που χρησιμοποιήθηκαν στη συγκεκριμένη ανάλυση για την πρόβλεψη της απόστασης από ασυνέπεια.

3.1 Γραμμική Παλινδρόμηση

Η γραμμικής παλινδρόμηση είναι από τις θεμελιώδεις μεθόδους στη στατιστική και οικονομετρία για ερμηνεία και πρόβλεψη. Βασίζεται στη μέθοδο των ελαχίστων τετραγώνων για την εκτίμηση,

$$\min_{\mathbf{w},b} \frac{1}{2m} \sum_{i=1}^{m} \left(y^{(i)} - \mathbf{w}^\top \mathbf{x}^{(i)} - b\right)^2$$

όπου $y^{(i)}$ η εξαρτημένη μεταβλητή,$x^{(i)}$ οι επεξηγηματικές μεταβλητές, $w^\top$ το διάνυσμα συντελεστών και b ο σταθερός όρος. Ελαχιστοποιώντας τη σχέση προσαρμόζει τη γραμμή παλινδρόμησης στα σημεία παρατηρήσεων, δημιουργώντας ένα γραμμικό μοντέλο. Λόγω της απλής μορφής της είναι η καταλληλότερη για ερμηνεία των μεταβλητών καθώς υπάρχουν διαστήματα εμπιστοσύνης

και έλεγχοι υποθέσεων για τη σημαντικότητα των συντελεστών. Η γραμμική παλινδρόμηση θα χρησιμοποιηθεί ως μοντέλο βάσης για την σύγκριση με τις μεθόδους μηχανικής μάθησης, για να διερευνηθεί αν και πόσο μεγαλύτερη προβλεπτική ικανότητα μπορούν να προσφέρουν.

3.2 Random Forest

Ο αλγόριθμος Random Forest βασίζεται στα δένδρα αποφάσεων και τη μέθοδο bagging. Τα δένδρα είναι μια αρκετά απλή μέθοδος ταξινόμησης και παλινδρόμησης και αποτελούνται από διαδοχικά φύλλα (κόμβους). Προσφέρουν χαμηλή προβλεπτική ικανότητα και μεγάλη διακύμανση και για αυτό ονομάζονται weak learners. Η μέθοδος bagging προέρχεται από το bootstrap aggregating και είναι η ιδέα του συνδυασμού πολλών μοντέλων (weak learners) εκτίμησης, που έχουν εκπαιδευτεί σε υποσύνολο του δείγματος με επανατοποθέτηση (bootstrap). Ο συνδύασμός πολλών μοντέλων (aggregating) βελτιώνει την αποδοτικότητα, μειώνει την διακύμανση της εκτίμησης και ενισχύει την ευστάθεια των αποτελεσμάτων. Το πρόβλημα εμφανίζεται στο τρόπο με τον οποίο τα δένδρα διαχωρίζουν τις παρατηρήσεις από φύλλο σε φύλλο. Συγκεκριμένα για να επιτευχθεί ο στόχος της καλύτερης πρόβλεψης δοκιμάζονται όλες οι μεταβλητές σε διάφορα όρια. Επιλέγεται εκείνη η μεταβλητή και το όριο που σε ένα οποιοδήποτε κόμβο θα διαχωρίσει καλύτερα με την έννοια ότι θα μειώσει τη μεταβλητότητα των προβλέψεων στα δύο επόμενα φύλλα (child nodes). Εάν η μείωση της μεταβλητότητας των προβλέψεων είναι αυτοσκοπός στα δέντρα τότε με τη μέθοδο bagging θα έχουμε συσχετισμένα δένδρα και συνεπώς συσχετισμένες εκτιμήσεις. Αυτό θα συμβαίνει διότι στο δείγμα υπάρχουν ισχυρές και λιγότερο ισχυρές μεταβλητές, επομένως τα δένδρα πάντα θα ξεκινούν τους κόμβους και τα διαδοχικά φύλλα με μια μορφή ιεραρχίας ανάμεσα στις πιο ισχυρές μεταβλητές. Συνεπώς θα έχουμε πολλά weak learners που είναι παρόμοια, το οποίο δε βοηθά στη γενίκευση των εκτιμήσεων. Ο Breiman (2001) πρότεινε τα Random Forests για τη λύση αυτού του προβλήματος. Τα Random Forests πέρα από την ιδέα τη συγκρότησης πολλών δένδρων μαζί με τη μέθοδο bagging εισάγουν τη έννοια της τυχαιότητας στην επιλογή των μεταβλητών για να αποφευχθούν τα συσχετισμένα δένδρα. Συγκεκριμένα κατά το διαχωρισμό ενός κόμβου στο δένδρο ένα συγκεκριμένο υποσύνολο των μεταβλητών θα κρίνεται προς έλεγχο. Στην έρευνα του δοκίμασε τυχαίες υποψήφιες μεταβλητές από το σύνολο όπως $m = 1$ ή $m = \log_2(F) + 1$ με πολύ καλά αποτελέσματα, ενώ σημείωσε ότι σε προβλήματα παλινδρόμησης ο αριθμός των τυχαίων μεταβλητών θα πρέπει να είναι μεγαλύτερος.

3.3 XGBoost

Ο αλγόριθμος eXtreme Gradient Boosting (Chen και Guestrin (2016)) έχει πολλά κοινά με τα Random Forests με το ποιο σημαντικό ότι πρόκειται για εκτιμήσεις που προέρχονται από πολλά διαφορετικά δέντρα αποφάσεων. Η διαφορά τους είναι στο τρόπο με τον οποίο εκπαιδεύονται στο δείγμα. Στις μεθόδους Boosting έχουμε τη λεγόμενη προσθετική εκμάθηση (additive training). Διαδοχικά δέντρα κατασκευάζονται τα οποία ελαχιστοποιούν τις λάθος εκτίμησεις (κατάλοιπα) του προηγούμενου δέντρου. Όπως προαναφέρθηκε, στα δάση τα δέντρα ελαχιστοποιούν την αντικειμενική συνάρτηση από φύλλο σε φύλλο, αλλά τα δέντρα είναι ανεξάρτητα μεταξύ τους. Αν υποθέσουμε $\hat{y}_i = 0$ η αρχική τιμή εκτίμησης, κατασκευάζονται δέντρα έτσι ώστε:

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\vdots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}$$

όπου $f_k(x) : 1, \dots t$ οι συναρτήσεις που περιγράφουν τη δομή των δέντρων. Για την εύρεση του δέντρου σε κάθε διαδοχικό βήμα ορίζεται μια αντικειμενική συνάρτηση που πρέπει να βελτιστοποιηθεί.

$$\begin{aligned}
\text{obj}^{(t)} &= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \omega(f_k) \\
&= \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \omega(f_t) + \text{C}
\end{aligned}$$

Το αριστερό μέρος τη συνάρτησης είναι η συνάρτηση απώλειας (loss function) και το δεξί μέρος λέγεται regulatization. Στην περίπτωση παλινδρόμησης η σύνηθης συνάρτηση απώλειας είναι το μέσο τετραγωνικό σφάλμα (mean squared error, MSE), και η αντικειμενική γίνεται:

$$\begin{aligned}
\text{obj}^{(t)} &= \sum_{i=1}^{n} \left(y_i - \left(\hat{y}_i^{(t-1)} + f_t(x_i)\right)\right)^2 + \sum_{k=1}^{t} \omega(f_k) \\
&= \sum_{i=1}^{n} \left[2\left(\hat{y}_i^{(t-1)} - y_i\right) f_t(x_i) + f_t(x_i)^2\right] + \omega(f_t) + \text{C}
\end{aligned}$$

Γενικότερα μπορούμε να πάρουμε το ανάπτυγμα Taylor της αντικειμενικής ως προς την loss function:

$$\text{obj}^{(t)} = \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}_i^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \omega(f_t) + \text{C}$$

όπου τα g_i(gradient) and h_i(hessian) ορίζονται ως:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$$
$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l\left(y_i, \hat{y}_i^{(t-1)}\right)$$

Αγνοώντας τις σταθερές, διότι στο δέντρο t μας ενδιαφέρει η μεταβολή στην αντικειμενική από το προηγούμενο δέντρο, θέλουμε να βελτιστοποιήσουμε την:

$$\sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \omega(f_t)$$

Στο XGBoost το regularization term ορίζεται ως:

$$\omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2$$

όπου w_j περιγράφει τη δομή των φύλλων του δέντρου. Μετά από κάποιες πράξεις και αντικαταστάσεις η αντικειμενική μπορεί να γραφεί:

$$\text{obj}^{(t)} = \sum_{j=1}^{T} \left[G_j w_j + \frac{1}{2} (H_j + \lambda) w_j^2 \right] + \gamma T$$

Τα φύλλα w_j είναι ανεξάρτητα το ένα από το άλλο και αν βελτιστοποιήσουμε την αντικειμενική παίρνουμε

$$w^\star = -\frac{G_j}{H_j + \lambda}$$
$$\text{obj}^\star = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$$

Όπου $w^\star$ μας δείχνει τη βέλτιστη δομή των φύλλων του δέντρου που εξαρτάται από τη πρώτη και

τη δεύτερη μερική παράγωγο της loss function και τη παράμετρο λ. Η obj* δείχνει πόσο καλό είναι συνολικά το δέντρο με μικρότερα score να αυξάνουν την προβλεπτική ικανότητα. Η παράμετρος γ συνεισφέρει στις τεχνικές pruning του δέντρου (κοψίματος φύλλων που δε συνεισφέρουν στην απόδοση).

3.4 Support Vector Regression (SVR)

Έστω ένα σύνολο εκπαίδευσης αποτελούμενο από n δείγματα της μορφής $\{(x_i, y_i)\}_{i=1}^n$, όπου $x_i \in \mathbb{R}^d$ είναι τα διανύσματα εισόδου και $y_i \in \mathbb{R}$ οι αντίστοιχες τιμές στόχου. Σκοπός των μοντέλων SVR είναι να προσδιοριστεί μια συνάρτηση $f(x)$ που προσεγγίζει τις τιμές y_i με ακρίβεια τουλάχιστον ε, ενώ ταυτόχρονα διατηρεί τη μορφή της όσο το δυνατόν πιο απλή.

Η γενική μορφή του υπό εκτίμηση μοντέλου είναι:

$$f(x) = \langle w, \phi(x) \rangle + b,$$

όπου $\phi(x)$ είναι ένας (πιθανώς μη γραμμικός) μετασχηματισμός του x σε έναν χώρο χαρακτηριστικών υψηλότερης διάστασης, και w, b είναι παράμετροι που προκύπτουν μέσω εκπαίδευσης.

Για την εκτίμηση του μοντέλου χρησιμοποιείται η λεγόμενη *ε-μη ευαίσθητη συνάρτηση απώλειας*:

$$L_\varepsilon(y, f(x)) = \begin{cases} 0, & \text{αν } |y - f(x)| \leq \varepsilon, \\ |y - f(x)| - \varepsilon, & \text{διαφορετικά.} \end{cases}$$

Αυτό σημαίνει πως το μοντέλο δεν τιμωρεί σφάλματα όταν βρίσκονται εντός του περιθωρίου ε, εστιάζοντας μόνο στις μεγαλύτερες αποκλίσεις.

Για να επιτευχθεί επιπεδότητα και ταυτόχρονα να επιτραπεί κάποια ανοχή σε αποκλίσεις, διατυπώνεται το εξής πρόβλημα βελτιστοποίησης:

$$\begin{aligned} \min_{w,b,\xi_i,\xi_i^*} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{υπό τις συνθήκες} \quad & y_i - \langle w, \phi(x_i) \rangle - b \leq \varepsilon + \xi_i, \\ & \langle w, \phi(x_i) \rangle + b - y_i \leq \varepsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \geq 0, \quad i = 1, \ldots, n. \end{aligned}$$

Όπου:

- ξ_i, ξ_i^* είναι οι μεταβλητές σφάλματος (slack variables),
- $C > 0$ είναι παράμετρος κανονικοποίησης που ελέγχει τον συμβιβασμό μεταξύ της πολυπλοκότητας του μοντέλου και της ανοχής στα σφάλματα.

Η παραπάνω βελτιστοποίηση μπορεί να λυθεί πιο αποδοτικά μέσω της δυϊκής της μορφής. Εισάγοντας πολλαπλασιαστές Lagrange, προκύπτει το εξής πρόβλημα:

$$\begin{aligned} \max_{\alpha_i, \alpha_i^*} \quad & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(x_i, x_j) \\ & -\varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y_i(\alpha_i - \alpha_i^*) \\ \text{υπό τους περιορισμούς} \quad & \sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0, \\ & 0 \le \alpha_i, \alpha_i^* \le C. \end{aligned}$$

Η συνάρτηση $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ είναι η συνάρτηση πυρήνα (kernel), η οποία επιτρέπει τον υπολογισμό στο χώρο χαρακτηριστικών χωρίς να απαιτείται ρητός μετασχηματισμός μέσω του $\phi(x)$.Μια δημοφιλής επιλογή πυρήνα είναι ο RBF (Radial Basis Function), ο οποίος δίνεται από τη σχέση:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right), \quad \gamma > 0.$$

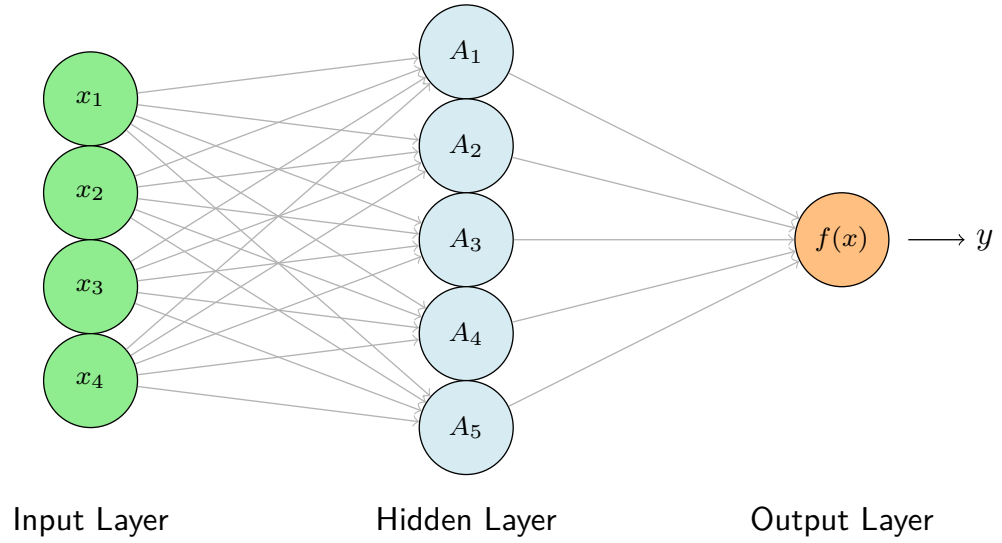
Ο RBF επιτρέπει την αποτύπωση πολύπλοκων, μη γραμμικών σχέσεων μεταξύ των παρατηρήσεων, καθιστώντας την SVR ιδιαίτερα ευέλικτη. Αφού λυθεί το δυϊκό πρόβλημα, η τελική συνάρτηση παλινδρόμησης είναι:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b.$$

Σημειώνεται ότι μόνο τα σημεία εκπαίδευσης για τα οποία $(\alpha_i - \alpha_i^*) \neq 0$ συνεισφέρουν στο τελικό μοντέλο. Τα σημεία αυτά ονομάζονται **διανύσματα στήριξης** (support vectors).

3.5 Νευρωνικά δίκτυα

Τα νευρωνικά δίκτυα είναι μία πολύ δημοφιλής κατηγορία μεθόδων τα τελευταία χρόνια για τη δημιουργία προβλεπτικών μοντέλων. Ιδιαίτερα τα τελευταία 15 χρόνια μετα την εξάπλωση της μηχανικής μάθησης, βαθιά μάθησης και τεχνητής νοημοσύνης έχουν γίνει ένα κυρίαρχο μοντέλο για την ανάλυση δεδομένων, κειμένων, εικόνων ακόμα και ηχητικοακουστικών δεδομένων. Φυσικά σε αυτό βοήθησε η μεγέθυνση τη υπολογιστικής ισχύς λόγω των μεγάλων απαιτήσεων για την επεξεργασία τους. Ονομάστηκαν έτσι γιατί προσομοιάζουν τους νευρώνες του εγκεφάλου που συνομιλούν και στέλνουν σήματα σε παρακείμενους νευρώνες. Το μοντέλο που θα χρησιμοποιηθεί στην ανάλυση είναι το multi-layer perceptron (MLP). Το MLP είναι ένα είδος νευρωνικού δικτύου το οποίο αποτελείται από στρώματα νευρώνων (nodes) που είναι πλήρως διασυνδεδεμένοι μεταξύ τους. Στην απλότερη του μορφή, όπως φαίνεται στο Σχήμα 3.1, έχουμε το πρώτο στρώμα το οποίο τροφοδοτείται με τα δεδομένα εισαγωγής, το δεύτερο στρώμα που ονομάζεται κρυφό, και το στρώμα εξαγωγής το οποίο παράγει προβλέψεις. Η κατεύθυνση που στέλνουν δεδομένα είναι από την εισαγωγή προς την εξαγωγή και έτσι ονομάζεται νευρωνικό δίκτυο εμπρόσθιας τροφοδότησης (feed-forward neural network).



Σχήμα 3.1: Αρχιτεκτονική Νευρωνικού δικτύου

Στην πιο απλή μορφή του MLP με ένα κρυφό στρώμα, το στρώμα εισαγωγής περιλαμβάνει τις τιμές κάθε μεταβλητής σε μορφή διανυσμάτων και είναι ίσο με τον αριθμό των μεταβλητών. Συνδέεται με το κρυφό στρώμα που το οποίο επεξεργάζεται τις παρατηρήσεις μέσα από μία μηγραμμική συνάρτηση ενεργοποίησης. Στη συνέχεια στέλνει τα δεδομένα στο στρώμα εξαγωγής παράγωντας το διάνυσμα προβλέψεων (y). Η συναρτησιακή μορφή του νευρωνικού δικτύου ενός

κρυφού στρώματος του σχήματος αναπαρίσταται μαθηματικά:

$$f(X) = \beta_0 + \sum_{k=1}^{K} \beta_k h_k(X) \tag{3.1}$$

$$= \beta_0 + \sum_{k=1}^{K} \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j\right). \tag{3.2}$$

και A_k οι νευρώνες του κρυφού στρώματος:

$$A_k = h_k(X) = g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j\right).$$

Η συνάρτηση g είναι η συνάρτηση ενεργοποίησης ReLU (rectified linear unit). Μέσω αυτής μπορεί και εξερευνεί πολύπλοκες αλληλεπιδράσεις και μη-γραμμικότητα ανάμεσα στις μεταβλητές. Η ReLU είναι ευρέως διαδεδομένη στα σύγχρονα δίκτυα και χρησιμοποιήθηκε σε αυτό το μοντέλο:

$$g(z) = (z)_+ = \begin{cases} 0 & \text{εάν } z < 0 \\ z & \text{εάν } z \geq 0 \end{cases}$$

Η συναρτησιακή μορφή (3.2) θυμίζει αρκετά τη μορφή ενός γραμμικού μοντέλου. Λόγω της μη-γραμμικής συνάρτησης ενεργοποίησης και τα φωλευμένα βάρη των νευρώνων δεν υπάρχει μοναδική λύση. Για την εκπαίδευση και την επίλυση του μοντέλου χρησιμοποιείται η μέθοδος οπισθοδιάδοσης (backpropagation). Λόγω ότι έχουμε ένα πρόβλημα παλινδρόμησης μπορούμε να γράψουμε τη συνάρτηση απώλειας του μοντέλου ως:

$$R(\theta^m) = \frac{1}{2} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2,$$

όπου θ το διάνυσμα που περιλαμβάνει όλες τις παραμέτρους και τα βάρη του μοντέλου και $f(x_i)$ εξαρτάται από το θ. Υπολογίζοντας τη διανυσματική παράγωγο στο σημείο $\theta = \theta^m$:

$$\nabla R(\theta^m) = \left.\frac{\partial R(\theta)}{\partial \theta}\right|_{\theta=\theta^m}$$

και

$$\theta^{m+1} \leftarrow \theta^m - \rho \nabla R(\theta^m)$$

όπου ρ ο ρυθμός εκμάθησης (learning rate), μπορούμε να βρούμε προς ποια κατεύθυνση το διάνυσμα των παραμέτρων μειώνει με μεγαλύτερο ρυθμό τη συνάρτηση απώλειας $R(\theta)$ έτσι ώστε $R(\theta^{m+1}) \leq R(\theta^m)$. Παρουσιάστηκε η μέθοδος επίλυσης για ένα MLP ενός κρυφού στρώματος αλλά μπορεί να επεκταθεί και σε δίκτυα περισσότερων κρυφών στρωμάτων.

Εφαρμόζοντας έναν αρκετά μικρό ρυθμό εκμάθησης μπορούμε να πετύχουμε μεγαλύτερη ακρίβεια. Κατά την εκπαίδευση χρησιμοποιούμε και τη διαδικασία mini-batch training όπου κομμάτια των δεδομένων τροφοδοτούνται τμηματικά στον αλγόριθμο και εκπαιδεύονται σε ορισμένα Epochs. Ένα epoch εκπαίδευσης ολοκληρώνεται όταν θα περάσει ολόκληρο το σετ δεδομένων σε μορφή mini-batches από τον αλγόριθμο. Ο ορισμός των τιμών epochs, μέγεθος των mini-batch και ρυθμού εκμάθησης είναι τόσο σημαντικός στην ακρίβεια του μοντέλου όσο και στη γενίκευσή του στα δεδομένα ελέγχου. Ένας τρόπος περιορισμού του overfitting είναι η εισαγωγή dropout layers στο δίκτυο. Στη πραγματικότητα δεν εισάγουμε περαιτέρω στρώματα στο δίκτυο αλλά κατά τη διάρκεια της εκπαίδευσης τυχαίοι νευρώνες στο κρυφό στρώμα απενεργοποιούνται με μια πιθανότητα p και δεν παράγουν προβλέψεις ή αλλιώς τα βάρη τους μηδενίζονται. Αυτό έχει ως αποτέλεσμα το δίκτυο να μη δίνει τόση σημασία σε κυρίαρχους νευρώνες που επηρεάζουν σε σημαντικό βαθμό τις τελικές προβλέψεις. Όλοι αυτοί οι υπερ-παραμέτροι του δικτύου μπορούν να υπολογιστούν σε σχετικά ιδανικές τιμές μέσω cross validation για βελτιστοποίηση της γενίκευσης σε δεδομένα ελέγχου.

4. Δεδομένα και Μεθοδολογία

4.1 Το πλαίσιο της ανάλυσης

Το γενικό πλαίσιο της ανάλυσης αφορά την εκτίμηση της απόστασης από ασυνέπεια όπως αυτή προκύπτει από το εξειδικευμένο μοντέλο τύπου Black-Scholes-Merton. Η εκτίμηση αυτού του δεδομένου φαίνεται χρήσιμη σε περιπτώσεις που δεν μπορούν να βρεθούν δεδομένα για τον υπολογισμό ενός μοντέλου αγοράς ή λόγω πολυπλοκότητας του υπολογισμού. Σε άλλη περίπτωση, η εκτίμηση της απόστασης από ασυνέπεια μέσα από το μοντέλο αγοράς, θα ήταν αδύνατη για μη εισηγμένη εταιρία. Επομένως το σκεπτικό της ανάλυσης είναι διττό, πρώτον, η ανάπτυξη ενός μοντέλου που θα μπορεί να χρησιμοποιηθεί εύκολα και γρήγορα για μεγάλο φάσμα επιχειρήσεων καθώς θα έχει ως ανεξάρτητες μεταβλητές βασικούς χρηματοοικονομικούς δείκτες, και δεύτερον η συγκριτική ανάλυση των μοντέλων που θα αναπτυχθούν να δώσει μια εικόνα για την προβλεπτική ικανότητά τούς, και αν θα καταφέρουν να πλησιάσουν το μοντέλο της αγοράς. Πρέπει να επισημανθεί ότι επειδή η ανάπτυξη των μοντέλων βασίστηκε σε κοινούς χρηματοοικονομικούς δείκτες η ανάπτυξη ενός μοντέλου μπορεί να πραγματοποιηθεί χωρίς να υπάρχει πρόσβαση σε δεδομένα αγοράς ή πτωχεύσεων.

Με αυτό το σκεπτικό κρίθηκε αναγκαίο να αναπτυχθούν μοντέλα που είναι σύγχρονα και έχουν αναφερθεί εκτενώς στη βιβλιογραφία. Για την αμεροληψία της ανάλυσης υϊοθετήθηκαν καλές πρακτικές στην προετοιμασία των δεδομένων (επιλογή μεταβλητών, αφαίρεση ακραίων τιμών) και στην ανάπτυξη των μοντέλων (cross-validation, παραμετροποίηση).

4.2 Δεδομένα

Το μέρος των δεδομένων που παρέχουν την εκτίμηση του Distance-to-default προέρχονται από το Credit Research Initiative (CRI) του Asian Institute of Digital Finance του National University of Singapore. Πρόκειται για ένα μη κερδοσκοπικό ερευνητικό ινστιτούτο που επικεντρώνεται στην

έρευνα πιστωτικού κινδύνου και παρέχει αναλυτικά δεδομένα για εισηγμένες επιχειρήσεις σε παγκόσμια κλίμακα.

Το δεύτερο μέρος των δεδομένων προέρχεται από τη βάση δεδομένων της Refinitiv και περιλαμβάνει 5 έτη παρατηρήσεων εισηγμένων Ευρωπαϊκών επιχειρήσεων τα έτη 2019-2023. Τα δεδομένα περιλαμβάνουν χρηματοοικονομικούς δείκτες και ποιοτικές μεταβλητές όπως η χώρα αναφοράς και η κατηγορία οικονομικής δραστηριότητας (TRBC Economic Sector). Δεν περιλαμβάνουν Χρηματοπιστωτικά ιδρύματα. Τα δεδομένα θα πρέπει να αντιστοιχηθούν με τις παρατηρήσεις για κάθε οντότητα και κάθε έτος του DtD από το CRI. Το CRI αναφέρει ημερήσια δεδομένα DtD τα οποία ανάγουμε σε ετήσια βάση. Έτσι καταλήγουμε σε ένα αρχείο δεδομένων με **9067** έτη-παρατηρήσεις από **2016** ξεχωριστές οντότητες.

4.2.1 Χρηματοοικονομικοί δείκτες επιλογής

Δείκτης ανακύκλωσης απαιτήσεων: Είναι ο λόγος καθαρές πιστωτικές πωλήσεις προς τον μέσο του λογαριασμού απαιτήσεις εισπρακτέες. Μας δείχνει πόσο συχνά μετατρέπει τις πιστώσεις σε ρευστά διαθέσιμα κατά τη διάρκεια του έτους. Είναι ένας δείκτης ρευστότητας.

$$\text{Δείκτης ανακύκλωσης απαιτήσεων} = \frac{\text{Καθαρές πιστωτικές πωλήσεις}}{\text{Μέσες απαιτήσεις εισπρακτέες}}$$

EBIT / Σύνολο ενεργητικού: Είναι τα κέρδη προ τόκων και φόρων (EBIT) προς το σύνολο του ενεργητικού. Μας δείχνει πόσο καλά χρησιμοποιεί η επιχείρηση το ενεργητικό της για να παράγει κερδοφορία. Ένας λόγος επιλογής του EBIT σε αντίθεση με το EBITDA είναι ότι το EBITDA αυξάνει εφόσον περιλαμβάνει τις αποσβέσεις κάτι που είναι πλεονέκτημα για επιχειρήσεις έντασης κεφαλαίου. Η αύξηση του EBITDA μέσω των αποσβέσεων δεν είναι αποτέλεσμα αποδοτικότητας της οικονομικής δραστηριότητας της επιχείρησης. Η επιλογή του EBIT σε σύγκριση με το ROA (Καθαρό αποτέλεσμα / Σύνολο ενεργητικού) συνέβη διότι το ROA είναι το καθαρό αποτέλεσμα χρήσης που σημαίνει ότι χρηματοοικονομικά έξοδα(τόκοι από δανεισμό) και φορολογία έχουν αφαιρεθεί. Η χώρα και ο κλάδος δραστηριότητας επηρεάζουν τόσο το δανεισμό όσο και τη φορολογία και κρίθηκε απαραίτητο αυτή τη διακύμανση να την εξηγούν οι δύο ποιοτικές μεταβλητές που βρίσκονται στα δεδομένα (Country, Sector). Ένας άλλος λόγος επιλογής του EBIT είναι η υψηλότερη γραμμική συσχέτιση με την εξαρτημένη μεταβλητή (απόσταση από ασυνέπεια).

Δείκτης συνολικής ικανότητα δανεισμού: Είναι ένας δείκτης μόχλευσης της επιχείρησης και περιλαμβάνει στον αριθμητή τις βραχυπρόθεσμες και μακροπρόθεσμες υποχρεώσεις ενώ στον πα-

ρονομαστή το συνολικό ενεργητικό. Δείχνει πόσο χρέος χρησιμοποιεί για να χρηματοδοτήσει το ενεργητικό της. Χρησιμοποιείται για την σύγκριση του βαθμού μόχλευσης επιχειρήσεων στον ίδιο κλάδο.

$$\text{Δείκτης συνολικής ικανότητα δανεισμού} = \frac{\text{Βραχυπρόθεσμες + Μακροπρόθεσμες υποχρεώσεις}}{\text{Σύνολο ενεργητικού}}$$

Κεφάλαιο κίνησης / Σύνολο ενεργητικού: Ο λόγος κεφαλαίου κίνησης προς το σύνολο του ενεργητικού μας δείχνει σε τι ποσοστό του σύνολικού ενεργητικού της η επιχείρηση μπορεί να χρηματοδοτήσει τις βραχυπρόσθεσμες ανάγκες της.

$$\text{Κεφάλαιο κίνησης / Σύνολο ενεργητικού} = \frac{\text{Κυκλοφορούν ενεργητικό - Βραχυπρόθεσμες υποχρεώσεις}}{\text{Σύνολο ενεργητικού}}$$

Εναλλακτικά:

$$\frac{\text{(Ιδία Κεφάλαια + Μακροπρόθεσμες υποχρεώσεις) - Πάγιο ενεργητικό}}{\text{Σύνολο ενεργητικού}}$$

Δείκτης άμεσης ρευστότητας: Ο δείκτης άμεσης ρευστότητας είναι το κυκλοφορούν ενεργητικό μείον τα αποθέματα προς τις βραχυπρόθεσμες υποχρεώσεις. Μας δείχνει σε τι βαθμό τα γρήγορα ρευστοποιήσιμα στοιχεία του ενεργητικού υπερκαλύπτουν τον βραχυπρόθεσμο δανεισμό της επιχείρησης. Αυτά τα στοιχεία περιλαμβάνουν τα ταμειακά διαθέσιμα και ισοδύναμα, γραμμάτια και επιταγές εισπρακτέες κ.ά.

$$\text{Δείκτης άμεσης ρευστότητας} = \frac{\text{Ταμειακά διαθέσιμα και ισοδύναμα}}{\text{Βραχυπρόθεσμες υποχρεώσεις}}$$

Σύνολο ενεργητικού (Φυσικός Λογάριθμος): Το σύνολο του ενεργητικού μας δείχνει το συνολικό μέγεθος της επιχείρησης και είναι από τις πιο σημαντικές μεταβλητές σε έρευνες παρόμοιου αντικειμένου. Η προσθήκη αυτού του μεγέθους σε φυσικό λογάριθμο βελτιώνει την κατανομή της μεταβλητής και αποφεύγονται υπολογιστικά προβλήματα στην εκτίμηση των παραμέτρων όταν υπάρχουν παρατηρήσεις μεγάλης διαφοράς μεγέθους.

4.2.2 Αφαίρεση Ακραίων τιμών

Οι ακραίες τιμές (outliers) επηρεάζουν τη διακύμανση μιας μεταβλητής και μεταβάλλουν το μέσο συνεπώς δημιουργούν προβλήματα στην εκτίμηση των παραμέτρων των μοντέλων. Οι ακραίες αυτές τιμές μπορεί να είναι είτε λάθος καταχώρηση στα δεδομένα είτε εσκεμμένες για υπερεκτίμηση ή υποεκτίμηση των ανάλογων μεγεθών. Όπως έχουν δείξει πολλές έρευνες η λανθασμένη διατύπωση των μεγεθών στις χρηματοοικονομικές καταστάσεις αυξάνει την πιθανότητα μια επιχείρηση να λάβει qualified report κάτι που συνδέετε με αύξηση του πιστωτικού κινδύνου (Pasiouras κ.ά., 2007). Χωρίς να γνωρίζουμε ότι αυτό έχει συμβεί σε αυτή τη περίπτωση οι μεταβλητές πρέπει να χαρακτηρίζονται από ορθολογικές τιμές με βάση τις διεθνείς νόρμες. Για αυτό το λόγο γίνεται εφαρμογή τις μεθόδου winsorization. Με βάση αυτή οι ακραίες τιμές αντικαθιστούνται με το άνω και κάτω όριο τιμών στα οποία σταματάμε να θεωρούμε ότι υπάρχουν ακραίες τιμές. Αυτή η προσέγγιση είναι εμπειρική και μπορεί να χαρακτηριστεί υποκειμενική. Περισσότερα για την επίδραση των ακραίων τιμών στους χρηματοοικονομικούς δείκτες μπορεί κανείς να δει τους Lev και Sunder (1979), Frecka και Hopwood (1983).

Πίνακας 4.1: Επιλεγμένα όρια για αντικατάσταση στις μεταβλητές

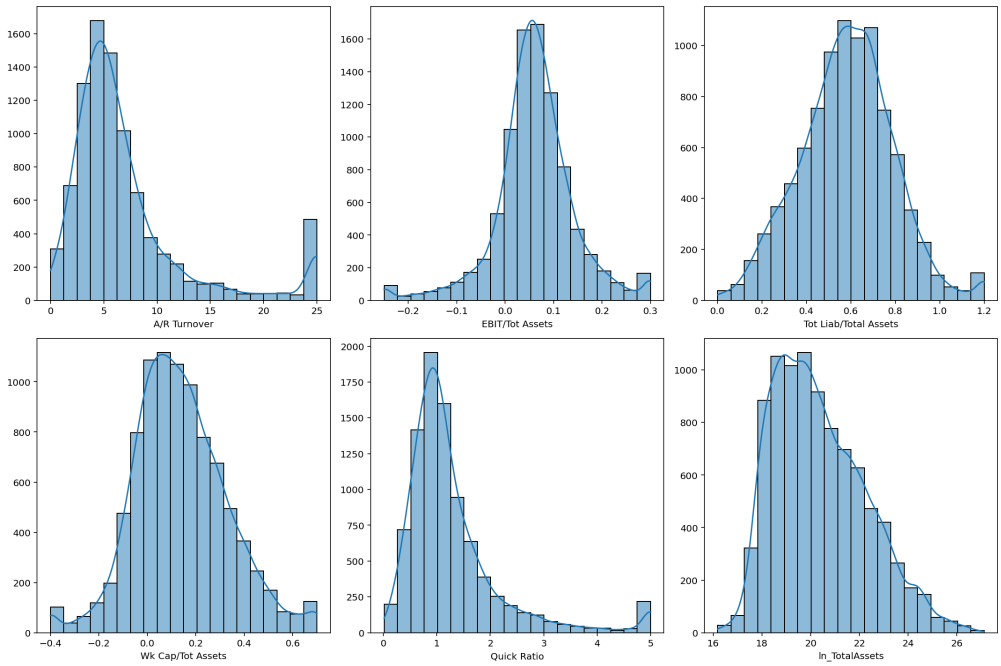
Μεταβλητή	Κάτω Όριο	Άνω Όριο
Accounts Receivable Turnover Ratio	0	25
EBIT / Total Assets	-0.25	0.3
Debt to Assets Ratio	0	1.2
Working Capital / Total Assets	-0.4	0.7
Quick Ratio	0	5
Total Assets (Φυσικός Λογάριθμος)	--	99th pctl.

Στον Πίνακα 4.1 φαίνονται οι τιμές επιλογής για κάθε μεταβλητή. Επιλέχθηκε το 0 ως αντικατάσταση για το κάτω όριο για τον δείκτη ανακύκλωσης απαιτήσεων, τον δείκτη ικανότητας δανεισμού και άμεσης ρευστότητας καθώς οι αριθμητές τους δε μπορούν να περιέχουν αρνητικές τιμές ενώ το άνω όριο έχει κρατηθεί σε φυσιολογικά επίπεδα. Στους δείκτες Κεφαλαίου κίνησης/ΤΑ και EBIT/ΤΑ έχουμε αφήσει ένα ορθολογικό επίπεδο αρνητικού Κ.Κ. και ζημιών(αρνητικό EBIT) ως ποσοστό του Ενεργητικού. Στο άνω όριο ορίστηκε ένα 30% ως ποσοστό του Ενεργητικού στο EBIT και ένα 70% στο Κ.Κ. Η μεταβλητή σύνολο ενεργητικού δεν έχει κάτω όριο καθώς ήδη έχουμε φιλτράρει τις παρατηρήσεις σε επιχειρήσεις $\geq$10m Ενεργητικού ενώ στο άνω όριο αντικαταστήσαμε τις ακραίες τιμές με το 99στό ποσοστημόριο. Στον Πίνακα 4.2 παρουσιάζονται περιγραφικά στατιστικά για τις μεταβλητές.

Πίνακας 4.2: Περιγραφικα στατιστικά μεταβλητών επιλογής

Μεταβλητή	Mean	Std.Dev.	Min	25%	Median	75%	Max
A/R Turnover	7.11	5.7	0.0	3.72	5.42	8.02	25.0
EBIT/TA	0.06	0.08	-0.25	0.02	0.06	0.1	0.3
Tot Liab/TA	0.58	0.21	0.0	0.45	0.59	0.72	1.2
Wk Cap/TA	0.14	0.19	-0.4	0.01	0.12	0.26	0.7
Quick Ratio	1.29	0.92	0.01	0.75	1.04	1.5	5.0
ln_TA	20.37	1.89	16.19	18.88	20.07	21.63	27.12

Στο Σχήμα 4.1 παρουσιάζονται τα ιστογράμματα με τις κατανομές των μεταβλητών. Με βάση τις τροποποιήσεις που έχουμε κάνει οι κατανομές προσεγγίζουν την κανονική ενώ σε κάποιες μεταβλητές έχουμε αυξημένη συχνότητα στην ουρά της κατανομής κάτι που θεωρούμε ότι δε θα επηρεάσει τη μοντελοποίηση.



Σχήμα 4.1: Ιστογράμματα Ανεξάρτητων Μεταβλητών

4.3 Cross-Validation

Για την αξιολόγηση της προβλεπτικής ικανότητας των μοντέλων χρησιμοποιούμε κάποια μορφή της μεθοδολογίας Cross-Validation. Τα σύγχρονα μοντέλα μηχανικής μάθησης έχουν την ικανότητα

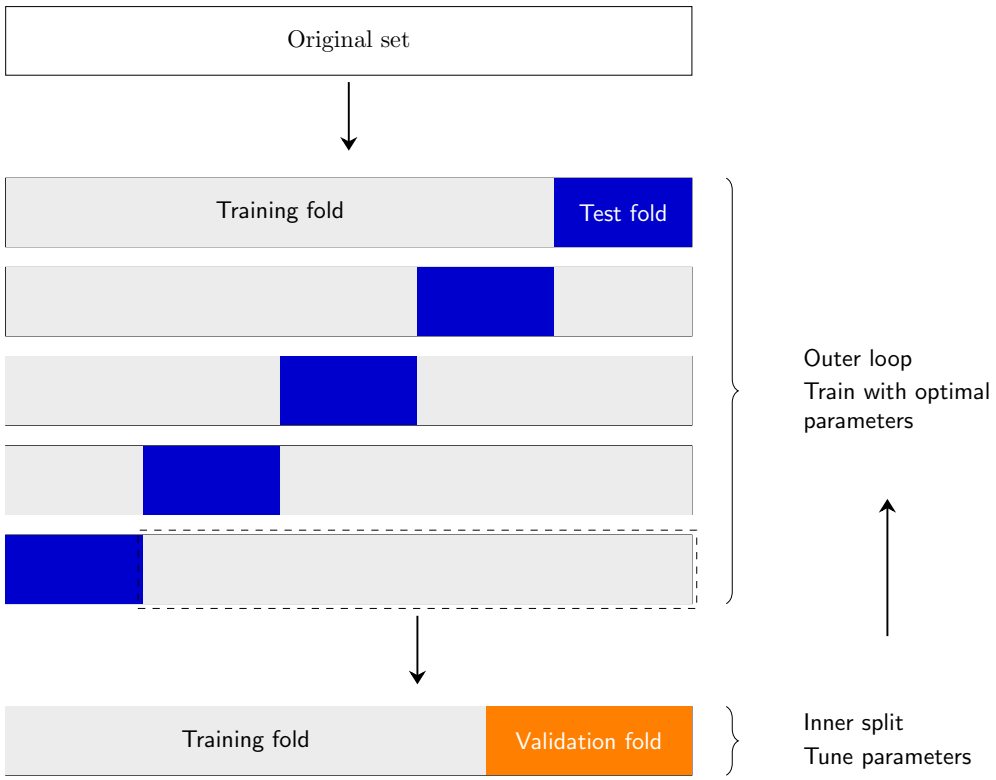
να προσαρμόζονται πολύ καλά στα δεδομένα στα οποία εκπαιδεύονται. Η εκπαίδευση (training) του μοντέλου δεν γίνεται σε ολόκληρο το αρχείο δεδομένων. Αρχικά, αυτό θα προκαλέσει μεγάλη προσαρμογή στα δεδομένα εκπαίδευσης (overfitting) και δεύτερον, κακή προσαρμογή σε δεδομένα εκτός δείγματος (out-of-sample). Δεδομένου αυτών, δημιουργήθηκε η λογική του train-test split. Η διαδικασία δηλαδή που το δείγμα μας χωρίζεται σε δεδομένα εκπαίδευσης και αξιολόγησης. Τα δεδομένα που προορίζονται για αξιολόγηση δεν έρχονται σε καμμία επαφή με τη διαδικασία της εκπαίδευσης και θεωρούνται δεδομένα out-of-sample. Τα μειονεκτήμα της μεθόδου train-test split είναι ότι η προβλεπτική ικανότητα καθορίζεται απο τον τρόπο που έχει διαχωριστεί το δείγμα. Έτσι σε ένα τυχαίο χώρισμα μπορούμε να έχουμε διαφορετικά αποτελέσματα από κάποιο άλλο τυχαίο διαχωρισμό του δείγματος. Έτσι έχουμε υψηλή μεταβλητότητα και αστάθεια σε κάθε δοκιμή που κάνουμε. Ένα άλλο μειονέκτημα είναι ότι χρησιμοποιούμε μόνο ένα μέρος των δεδομένων προς εκπαίδευση. Η λύση αυτού του προβλήματος είναι η επαναληπτική αξιολόγηση πολλών διαδικασιών train-test split το οποίο ονομάζεται k-fold Cross Validation. Στο k-fold CV, όπου συνηθέστερα k=5 ή k=10, έχουμε την επαναληπτική (k) εκπαίδευση και αξιολόγηση του δείγματος, όπου σε κάθε επανάληψη επιλέγεται διαφορετικό κομμάτι για την εκπαίδευση και αξιολόγηση. Με αυτό το τρόπο μειώνουμε τη μεταβλητότητα και έχουμε ένα πιο ισχυρό στατιστικά αποτέλεσμα.

4.3.1 Βελτιστοποίηση υπερ-παραμέτρων

Στην παρούσα εργασία έχουμε μοντέλα μηχανικής μάθησης τα οποία πρέπει να εκπαιδευτούν με τις σωστές υπέρ-παραμέτρους (hyper-parameters). Η εύρεση των σχετικά βέλτιστων υπερ-παραμέτρων οδηγεί σε αύξηση της προβλεπτικής ικανότητας, επομένως κρίνεται σκόπιμη η εύρεση αυτών για να υπάρξει μια αμερόληπτη συγκριτική αξιολόγηση μεταξύ των μοντέλων. Για παράδειγμα, η αρχιτεκτονική ενός νευρωνικού δικτύου είναι πολύ σημαντική για την απόδοση του. Η εύρεση των παραμέτρων αυτών πρέπει να γίνει με την ίδια λογική του cross-validation. Δηλαδή, η εύρεση των σωστών παραμέτρων θα πρέπει να αξιολογηθεί σε κάποιο δείγμα εκτός δείγματος εκπαίδευσης για να περιοριστεί το ενδεχόμενο υπερ-προσαρμογής. Επιπλέον, η αξιολόγηση για την εύρεση των ιδανικών παραμέτρων και της προβλεπτικής ικανότητας του μοντέλου δε μπορεί να γίνει στο ίδιο δείγμα ελέγχου. Αυτό θα προκαλέσει το ενδεχόμενο διαρροής δεδομένων (data leakage).

Για αυτό το λόγο στην ανάλυση εφαρμόστηκε η μέθοδος 10-fold Cross-Validation με εσωτερικό διαχωρισμό σε κάθε fold. Η διαδικασία έχει ως εξής: Για κάθε τμήμα εκμάθησης από τα 10-folds που έχουν δημιουργηθεί στον εξωτερικό βρόγχο, το τμήμα αυτό χωρίζεται εκ νέου σε εκμάθησης και επικύρωσης. Το δείγμα επικύρωσης θα χρησιμοποιηθεί για την αξιολόγηση της αποδοτικότητας των παραμέτρων. Η ολοκλήρωση της διαδικασίας θα δώσει ιδανικές υπερ-παράμετρους για κάθε fold για να εκπαιδευτεί το τελικό μοντέλο με ένα σύνολο τιμών και να το αξιολογηθεί στο δείγμα

ελέγχου. Το Σχήμα 4.2 απεικονίζει τη διαδικασία σε ένα CV με k=5 (για εξοικονόμηση χώρου). Αξίζει να σημειωθεί ότι στη διάρκεια της διαδικασίας, για τον περιορισμό της διαρροής δεδομένων (data leakage), σε κάθε δείγμα εκπαίδευσης, επικύρωσης και ελέγχου πραγματοποιήθηκε κλίμακα κανονικοποίησης (scaling) στις αριθμητικές μεταβλητές (μηδενικός μέσος, μοναδιαία τυπική απόκλιση) και one-hot encoding στις κατηγορικές μεταβλητές, όπως ορίζουν οι καλές πρακτικές.



Σχήμα 4.2: k-fold CV με εσωτερικό διαχωρισμό βελτιστοποίησης

Η διαδικασία της βελτιστοποίησης απαιτεί την ύπαρξη αρκετών υπερπαραμέτρων σε εύλογες τιμές για να έχει νόημα η διαδικασία άλλα όχι σε τέτοιο βαθμό έως ότου να μην είναι υπολογιστικά εφικτή σε χρόνο και υπολογιστική ισχύ. Για την εύρεση των κατάλληλων υπερ-παραμέτρων η εύρεση επικεντρώθηκε στις πιο σημαντικές που επηρεάζουν την απόδοση των μοντέλων σε μεγαλύτερο βαθμό και αναφέρονται στη θεωρία των μοντέλων. Στο Πίνακα 4.3 παρουσιάζεται ο χώρος αναζήτησης των υπερ-παραμέτρων κάθε μοντέλου.

Αρχικά πραγματοποιήθηκε μια διαδικασία δοκιμής και σφάλματος. Για παράδειγμα, στα random forest και XGBoost βρέθηκε ένα κατάλληλο εύρος όπου ο αριθμός των εκτιμητών απέδιδαν καλύτερα. Στα SVR και νευρωνικά δίκτυα οι δοκιμές έδειξαν ότι οι kernel RBF και βελτιστοποιητής Adam αντίστοιχα είχαν τα καλύτερα αποτελέσματα. Στη συνέχεια και εφόσον ορίστηκε το εύρος των υπερ-παραμέτρων σε λογικές τιμές, χρησιμοποιήθηκε στον εσωτερικό χώρισμα της διαδικασίας CV ένας αλγόριθμος επιλογής παραμέτρων. Ο αλγόριθμος Optuna (Akiba κ.ά., 2019) που χρησιμο-

Πίνακας 4.3: Χώροι αναζήτησης υπερ-παραμέτρων για κάθε μοντέλο.

Μοντέλα	Υπερ-παράμετροι	Υποψήφιες τιμές
Linear Regression		
	--	--
Ensemble methods		
Random forest	No. of trees	$\{500, 600, \ldots, 1500\}$
	Max tree depth	$\{2, \ldots, 30\}$
	Min samples leaf	$\{1, \ldots, 15\}$
	Max features	$[0.1, \ldots, 0.7]$
	Max samples	$[0.5, \ldots, 0.9]$
XGBoost	No. of trees	$\{600, 700, \ldots, 2000\}$
	Max tree depth	$\{5, \ldots, 15\}$
	Learning rate	$[0.01, \ldots, 0.2]$ (log scale)
	γ	$[0, \ldots, 1]$
	λ	$[1, \ldots, 7]$
	Subsample	$[0.5, \ldots, 0.8]$
Support Vector Regression		
	C	$[0.1, \ldots, 100.0]$ (log scale)
	ϵ	$[0.01, \ldots, 1]$
	γ	$[0.01, \ldots, 1]$ (log scale)
	Kernel	RBF (fixed)
Neural Network		
	Number of hidden layers	$\{1, 2, 3\}$
	Hidden layer dimensions	$\{8, \ldots, 128\}$ (constrained by previous layer)
	Dropout layers	None, Layer 1, Layer 2, Layer 3, Layers 1&2, Layers 1&3, Layers 2&3, All layers
	Dropout probability	$[0.01, \ldots, 0.4]$ (if dropout applied)
	Learning rate	$[1 \times 10^{-4}, \ldots, 1 \times 10^{-2}]$ (log scale)
	Batch size	$\{16, \ldots, 256\}$
	Epochs	$\{20, \ldots, 80\}$
	Optimizer	Adam (fixed)

ποιήθηκε προσφέρει αποδοτικότερη και ταχύτερη εύρεση των κατάλληλων παραμέτρων από έναν αλγόριθμο Random search ή Grid Search. Αυτό το πετυχαίνει μέσα από τεχνικές succesive halving και ανεξάρτητης δειγματοληψίας των υποσχόμενων υπερ-παραμέτρων στο χώρο αναζήτησης. Φυσικά για να επιτευχθεί ένα αμερόληπτο αποτέλεσμα αλλά και για να υπάρξουν σταθερές επιλογές ανάμεσα στα folds χρειάζεται να πραγματοποιηθούν αρκετές δοκιμές για την εύρεση των κατάλληλων συνδυασμών. Έτσι σε κάθε εσωτερικό χώρισμα, ο αλγόριθμος Optuna διετέλεσε 100 δοκιμές διαφορετικών συνδυασμών στο χώρο αναζήτησης. Συνολικά, για τα τέσσερα μοντέλα μηχανικής μάθησης που αναζητήθηκαν οι κατάλληλες παραμέτροι, έγιναν 4×10 folds $\times$ 100 δοκιμές $=$ 4000 δοκιμές. Με την επιλογή του αριθμού 100 δοκιμών υπήρξαν σταθερά αποτελέσματα σε κάθε fold, συνέπεια μεταξύ των 10-fold και ικανοποιητική απόδοση των μοντέλων. Για να βρεθεί ένα τελικό σύνολο υπερ-παραμέτρων αποφασίσθηκε η χρήση των διάμεσων τιμών αυτών ανάμεσα στα 10-fold .Παρ'όλα αυτά εάν η υπολογιστική ισχύς και ο χρόνος είναι μεγαλύτερος η επιλογή περισσότερων δοκιμών θα ενισχύσει ακόμη περισσότερο την αποδοτικότητα των μοντέλων. Άλλωστε η διαδικασία της βελτιστοποίησης έγκειται στο να βρεθεί όχι ο τέλειος συνδυασμός παραμέτρων,

αυτό ειναι αδύνατον, αλλά ένας ικανοποιητικός συνδυασμός που παρουσιάζει σταθερότητα και αποδοτικότητα ανάμεσα στο δείγμα.

5. Αποτελέσματα

5.1 Μέτρα επίδοσης

Για την συγκριτική αξιολόγηση των μοντέλων χρειάζονται κάποια μέτρα επίδοσης που είναι κατάλληλα για μοντέλα παλινδρόμησης. Έτσι έγινε η επιλογή 2 μέτρων επίδοσης που χρησιμοποιούνται συνεχώς στη βιβλιογραφία. Ακόμη, χρησιμοποιήθηκε ένας δείκτης που μετράει το σφάλμα ως ποσοστό και είναι κατάλληλος για την συγκεκριμένη περίπτωση εξαρτημένης μεταβλητής.

Mean Absolute Error: Είναι το ποιο διαδεδομένο μέτρο για την σύγκριση μοντέλων. Μετράει τη μέση απόλυτη διαφορά των προβλέψεων από την πραγματική τιμή. Τα σφάλματα αυξάνουν το συγκεκριμένο μέτρο όσο και η απόλυτη διαφοράς της εκτίμησης από την πραγματική τιμή. Είναι δημοφιλές λόγω της εύκολης ερμηνείας του.

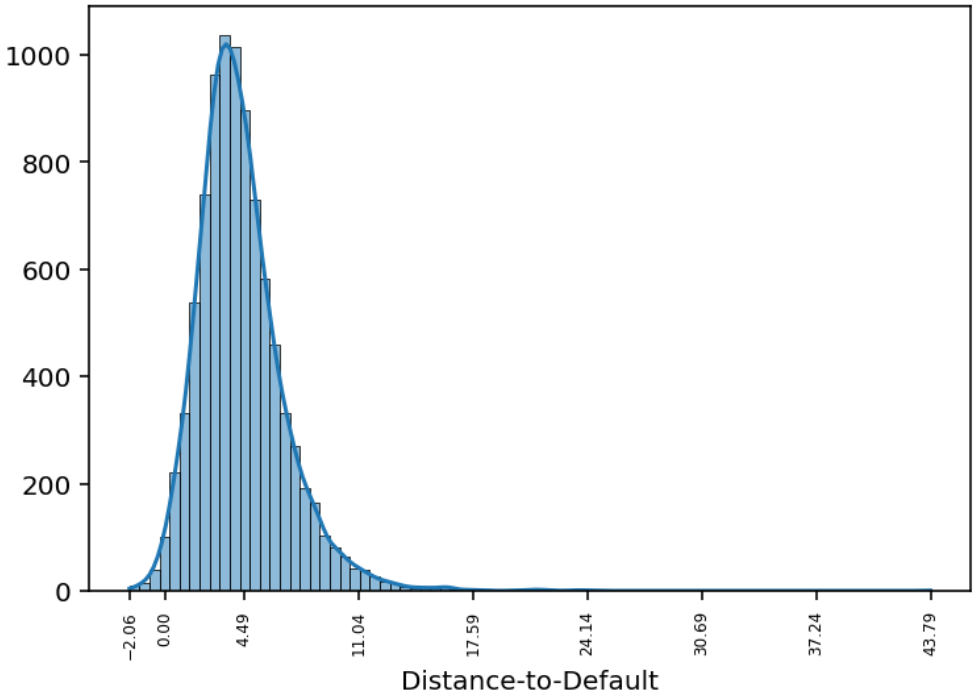
$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$

Root Mean Squared Error: Στο συγκεκριμένο μέτρο η διαφορά των προβλέψεων από την πραγματική τιμή υψώνεται στο τετράγωνο και έτσι έχουμε μια υπερεκτίμηση των σφαλμάτων. Τα υψηλά σφάλματα μεγενθύνουν περισσότερο το συγκεκριμένο μέτρο και ετσι μπορούμε να συμπεράνουμε, κατα μέσο όρο, ποιο μοντέλο είναι ικανό να κάνει μικρότερης ή μεγαλύτερης σημασίας σφάλματα.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

Οι δύο παραπάνω δείκτες μετρούν σε όρους μεγέθους της κλίμακας της εξαρτημένης μεταβλητής. Συνήθης πρακτική σε μοντέλα παλινδρόμησης είναι σύγκριση των σφαλμάτων σε όρους αναλογίας, δηλαδή ποσοστό σφάλματος ανά μονάδα πραγματικής τιμής (ή αντιθέτως το ποσοστό σωστής πρόβλεψης προς την πραγματική τιμή). Στη συγκεκριμένη περίπτωση η εξαρτημένη μας μεταβλητή έχει

αρκετές τιμές οι οποίες βρίσκονται κοντά στο μηδέν. Αυτό αλλοιώνει τα αποτελέσματα και όταν τα σφάλματα σε συγκεκριμένες επαναλήψεις είναι μεγάλα σε σύγκριση με την πραγματική τιμή, το ποσοστό αυξάνεται σε απαγορευτικό βαθμό. Για αυτό το λόγο δεν ήταν δυνατή η χρήση του κλασικού για παράδειγμα MAPE $= \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t-\hat{y}_t}{y_t}\right|$. Στο Σχήμα 5.1 μπορούμε να δούμε πως κατανέμεται το DTD.



Σχήμα 5.1: Ιστόγραμμα εξαρτημένης μεταβλητής DTD

Λύση στην αντιμετώπιση του προβλήματος που αναφέρθηκε είναι το ***Symmetric Mean Absolute Percentage Error***. Στο συγκεκριμένο μέτρο ο παρονομαστής του κλάσματος ειναι η απόλυτη τιμή του αθροίσματος της πρόβλεψης και της πραγματικής τιμής με αποτέλεσμα το μέτρο να παρουσιάζει φυσιολογικά ποσοστά σφάλματος. Η ερμηνεία του μέτρου αυτού δε θα πρέπει να είναι τόσο ευθύς όσο το MAPE καθώς όπως είναι φυσικό εισάγεται μια αμεροληψία, αν μη τι άλλο όμως αυτό το συμβαίνει για να μπορεί να γίνει μια σύγκριση μεταξύ των μοντέλων σε μέγεθος ποσοστού.

$$\text{sMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|\hat{y}_t - y_t|}{|y_t| + |\hat{y}_t|}$$

5.2 Συγκριτική Αξιολόγηση

Στον Πίνακα 5.1 παρουσιάζονται τα αποτελέσματα στα τρία μέτρα επίδοσης που αναλύθηκαν παραπάνω. Τα αποτελέσματα αυτά, είναι οι μέσες τιμές των μέτρων επίδοσης στα 10-folds της Nested CV. Στη σύγκριση του ισχυρότερου μοντέλου με κάθε ένα από τα υπόλοιπα μοντέλα, πραγματοποιήθηκε ένα two sample t-test για το μέσο του δείγματος των 10 folds, για να ελεγχθεί η στατιστική σημαντικότητα των αποτελεσμάτων. Η σημαντικότητα του ελέγχου παρουσιάζεται στα επίπεδα 1%(***), 5%(**) και 10%(*).

Πίνακας 5.1: Σύγκριση των μοντέλων

Μοντέλο	MAE	RMSE	sMAPE
SVR	**1.142**	1.665	**15.67**%
XGBoost	1.157*	**1.647**	15.80%
Random Forest	1.173***	1.691***	15.90%*
Neural Network	1.277***	1.830***	17.15%***
Linear Regression	1.413***	2.013***	18.92%***

Το βασικό μοντέλο γραμμικής παλινδρόμησης είχε τη χειρότερη απόδοση σε κάθε μέτρο επίδοσης και χρησιμοποιήθηκε ως μοντέλο βάσης για να δειχθεί η διαφορά σε σχέση με τα μοντέλα μηχανικής μάθησης. Το νευρώνικό δίκτυο είχε καλύτερα αποτελέσματα από την γραμμική παλινδρόμηση αλλά ήταν δεύτερο χειρότερο σε απόδοση σε σχέση με τα υπόλοιπα. Αξίζει να αναφερθεί ότι ήταν το μοντέλο με τις περισσότερες υπερ-παραμέτρους και τον μεγαλύτερο χρόνο εκπαίδευσης σε σχέση με τα υπόλοιπα. Στις μεθόδους δένδρων το XGBoost είναι ξεκάθαρα ισχυρότερο από το Random Forest καθώς είναι σε τουλάχιστον ένα μέτρο επίδοσης (RMSE) καλύτερο και στατιστικά σημαντικό ενώ και στα άλλα μέτρα παρουσιάζει καλύτερα αποτελέσματα. Παρ'ολα αυτά το πιο ισχυρό μοντέλο απ'όλα είναι το Support Vector Regression καθώς είναι στατιστικά σημαντικά ισχυρότερο από τις 3 άλλες μεθόδους σε επίπεδο σημαντικότητας 1% στα μέτρα MAE, RMSE ενώ στο sMAPE είναι σε επίπεδο 10% από το Random Forest. Σε σχέση με το XGBoost δεν υπάρχουν στατιστικά σημαντικές διαφορές στα μέτρα RMSE και sMAPE, ενώ στο MAE υπάρχει στατιστικά σημαντική διαφορά σε επίπεδο 10%. Αν και οι διαφορές είναι αμελητέες, μπορούμε να πούμε ότι το SVR είναι συνολικά ισχυρότερο μοντέλο αφού τουλάχιστον σε ένα μέτρο επίδοσης είναι στατιστικά σημαντικά καλύτερο από το XGBoost.

Κατά την ανάπτυξη των μοντέλων υπήρξαν μεγάλες διαφορές στην ταχύτητα εκπαίδευσης. Ο Πίνακας 5.2 δείχνει τον χρόνο εκπαίδευσης και την πολλαπλασιαστική βελτίωση της ταχύτητας σε σχέση με το Random Forest. Οι χρόνοι περιλαμβάνουν την πλήρη επανάληψη της 10-fold cross-validation εκπαίδευσης και εκτίμησης των μέτρων επίδοσης χρησιμοποιώντας τις τελικές

υπερ-παραμέτρους. Δε συμπεριλήφθηκε η γραμμική παλινδρόμηση, καθώς χρειάζεται μόλις κάποια milisecond για την εκτέλεση της. Το νευρωνικό δίκτυο ήταν το πιο απαιτητικό στην εκπαίδευση παρ΄ότι χρησιμοποήθηκε η ενσωματωμένη νευρωνική μηχανή του υπολογιστή. Κατά τη διάρκεια της εκπαίδευσης παρατηρήθηκε η χρησιμοποίησή τής στο περίπου 20% χωρίς να είναι γνωστό για ποιο λόγο είχε αυτή τη συμφόρηση καθώς πρόκειται για ποιο τεχνικό ζητήμα. Παρ'όλα αυτά τα μοντέλα SVR, XGBoost και Random Forest ορίστηκαν να δεσμεύσουν για την εκπαίδευση 4-πυρήνες CPU παράλληλα, με το SVR να επιτυγχάνει για άλλη μια φορά.

Πίνακας 5.2: Χρόνος εκπαίδευσης

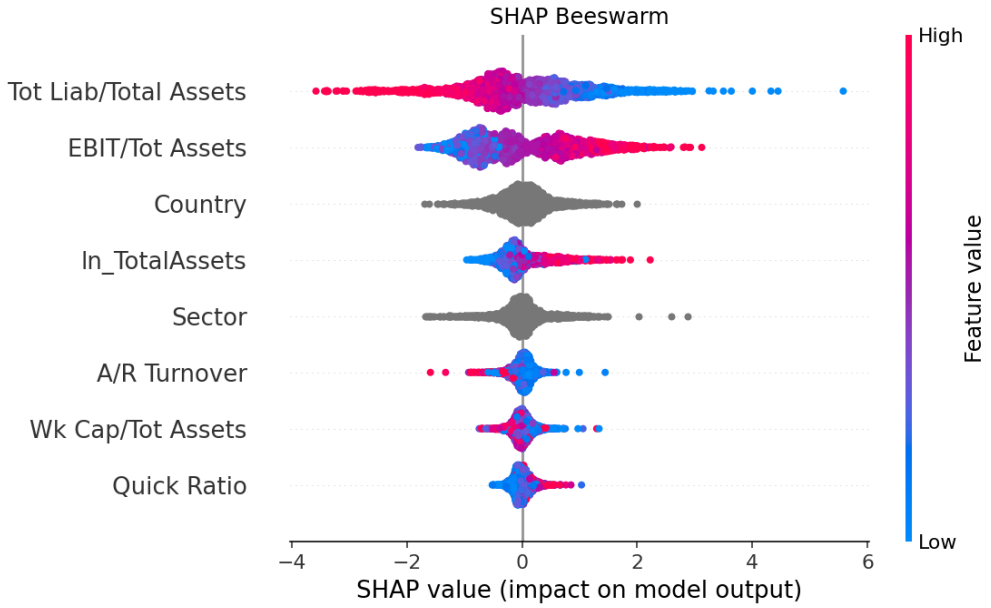
Μοντέλο	Χρόνος (sec.)	Βελτίωση ταχύτητας
SVR	4.59	17.86x
XGBoost	18.32	4.47x
Random Forest	82.00	--
Neural Network	280.21	--

5.3 Οι τιμές SHAP για την ερμηνεία μεθόδων μηχανικής μάθησης

Οι Lundberg και Lee (2017) παρουσίασαν τη μέθοδο SHapley Additive exPlanations για την ερμηνέια των μεθόδων μηχανικής μάθησης. Η ιδέα προέρχεται από τις τιμές Shapley στη θεωρία παιγνίων, όπου ο στόχος είναι η δίκαιη κατανομή της «πληρωμής» μεταξύ των παικτών ανάλογα με το πόσο συνέβαλε ο καθένας στο αποτέλεσμα του παιχνιδιού. Οι τιμές SHAP είναι μια μέθοδος για την ερμηνεία μοντέλων μηχανικής μάθησης ποσοτικοποϊώντας τη επίδραση κάθε μεταβλητής σε μια δεδομένη πρόβλεψη. Μια τιμή SHAP μας λέει πόσο μια μεμονωμένη μεταβλητή ώθησε την πρόβλεψη υψηλότερα ή χαμηλότερα σε σύγκριση με τη μέση πρόβλεψη (baseline rate). Οι θετικές τιμές SHAP αυξάνουν την πρόβλεψη, ενώ οι αρνητικές τιμές SHAP τη μειώνουν. Μέσω των τιμών SHAP μπορούν να παρουσιαστούν πολλά γραφήματα που θα βοηθήσουν να η ερμηνευτεί η σημαντικότητα των μεταβλητών στο πλαίσιο της ανάλυσης. Λόγω της αργής υπολογιστικά διαδικασίας, για την ανάλυση των τιμών SHAP και την ανάπτυξη των γραφημάτων, χρησιμοποιήθηκε το μοντέλο XGBoost. Στη βιβλιοθήκη του μοντέλου υπάρχει ενσωματωμένη η μέθοδος SHAP για μεγαλύτερη συμβατότητα, η οποία χρησιμοποιεί τον αλγόριθμο tree-explainer για ταχύτερες εκτιμήσεις των οριακών επιδράσεων.

Το Σχήμα 5.2 απεικονίζει το Beeswarm plot των μεταβλητών. Στον οριζόντιο άξονα φαίνεται η διακύμανση των τιμών SHAP κάθε μεταβλητής. Με λίγα λόγια μας εξηγεί πρώτον πόσο κυμαίνονται οι τιμές κάθε μεταβλητής και δεύτερον σε ποια σημεία συγκεντρώνεται μεγάλο πλήθος των τιμών SHAP. Η κατάταξη βασίζεται στην σημαντικότητα αυτών των μεταβλητών. Οι κατηγορικές

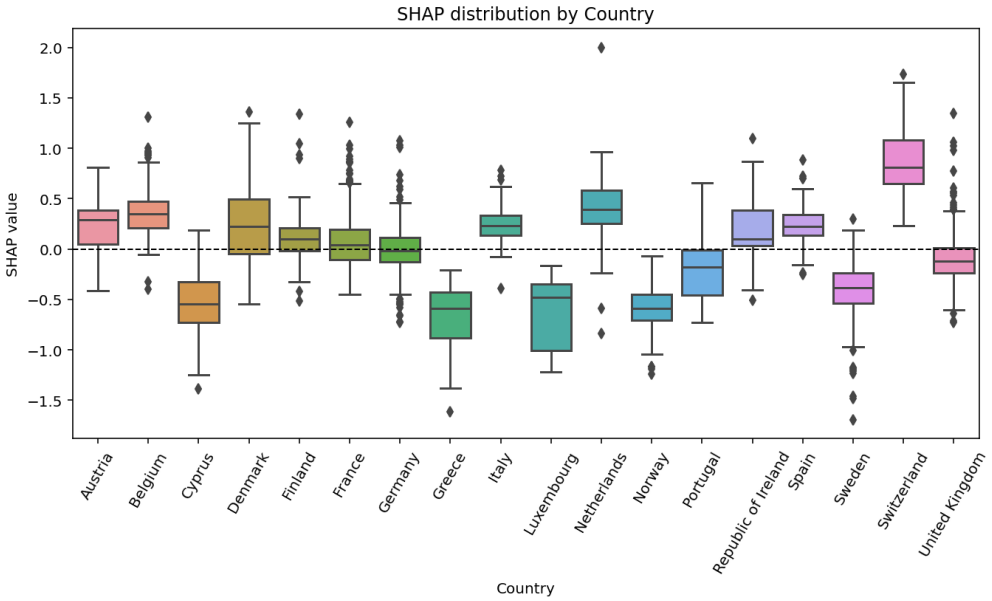
μεταβλητές «Country» και «Sector» παρουσιάζονται συνολικά για κάθε υποκατηγορία τους. Στο σχήμα είναι ξεκάθαρο ότι η ποιο σημαντική μεταβλητή που προβλέπει το μοντέλο ότι επιδρά στην απόσταση από ασυνέπεια είναι η συνολικές υποχρεώσεις προς σύνολο του ενεργητικού. Συγχρόνως οι παρατηρήσεις είναι χρωματισμένες με βάση την τιμή κάθε μεταβλητής, όπου για μεγάλες τιμές πάμε σε ποιο κόκκινα χρώματα ενώ οι μπλε τιμές το αντίθετο (οι κατηγορικές μεταβλητές δε μπορούν να χρωματιστούν). Για παράδειγμα, με βάση τη λογική οι υψηλότερες τιμές συνολικών υποχρεώσεων προς ενεργητικό, έχουν αρνητική οριακή επίδραση (αρνητικές τιμές SHAP) όπως παρουσιάζει και το γράφημα.



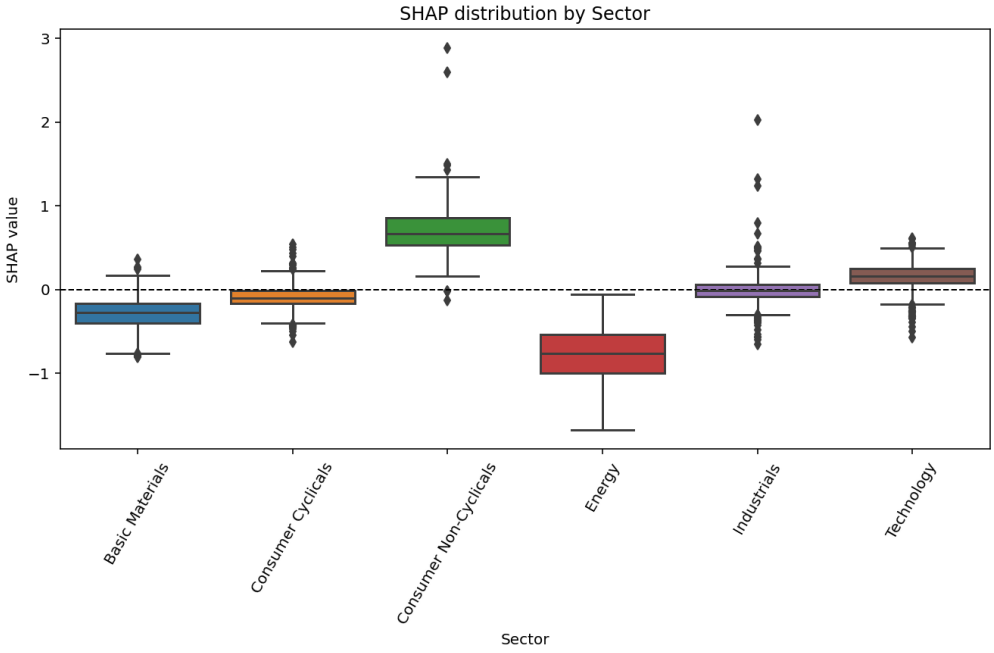
Σχήμα 5.2: Επίδραση των μεταβλητών σύμφωνα με τον δείκτη SHAP

Στα Σχήματα 5.3, 5.4 παρουσιάζονται το θηκόγραμμα για τις χώρες και τους τομείς δραστηριότητας στις κατηγορικές μεταβλητές. Οι τιμές SHAP για την ερμηνεία κατηγορικών μεταβλητών έχουν ως εξής: Είναι οι οριακή επίδραση που έχει μια παρατήρηση με το να ανήκει σε συγκεκριμένη κατηγορία σε σχέση με την τιμή βάσης (baseline) που υπολογίζει ο αλγόριθμος. Στο θηκόγραμμα εμφανίζονται το ελάχιστο και μέγιστο σημείο καθώς και το διατεταρτημοριακό εύρος. Στην περίπτωση των χωρών παρατηρούνται ότι οι παρατηρήσεις που ανήκουν στις χώρες Ελλάδα, Λουξεμβούργο και Νορβηγία το μοντέλο προβλέπει αρνητική οριακή επίδραση για το σύνολο των παρατηρήσεων ενώ το ίδιο συμβαίνει σε μεγάλο βαθμό και για τις χώρες Κύπρος και Σουηδία. Η ισχυρότερη χώρα για να ανήκει κάποια επιχείρηση είναι η Ελβετία με ισχυρά θετική οριακή επίδραση. Στους τομείς δραστηριότητας αρνητική οριακή επίδραση σε σχέση με την βάση παρουσιάζουν οι επιχειρήσεις στον κλάδο ενέργειας. Ο κλάδος ενέργειας έχει μεγάλη ευαισθησία στις τιμές πετρελαίου, φυσικού αερίου και στις γεωπολιτικές αλλαγές καθώς επίσης διακατέχεται από μεγάλες κεφαλαιακές ανάγκες. Άγνωστη είναι η πιθανότητα το δείγμα μας να υποτίμησε τις επιχει-

ρήσεις ενέργειας καθώς χρονολογικά περιλαμβάνει την περίοδο της πανδημίας, μια περίοδο που η βιομηχανική δραστηριότητα είχε μειωθεί σε χαμηλά επίπεδα.



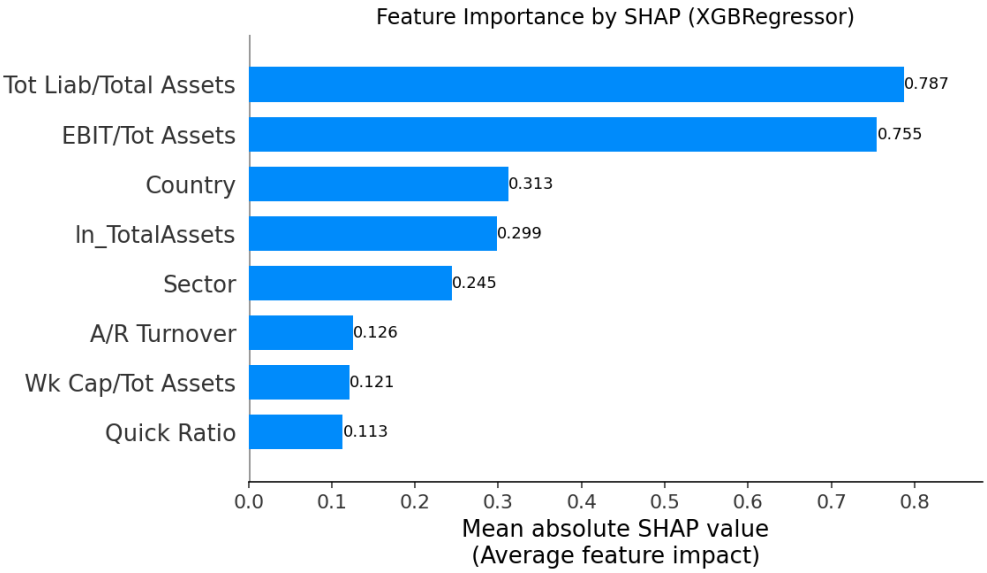
Σχήμα 5.3: Θηκόγραμμα τιμών SHAP ανά χώρα



Σχήμα 5.4: Θηκόγραμμα τιμών SHAP ανά κλάδο δραστηριότητας

Στο Σχήμα 5.5 μπορούμε να δούμε συνολικά τη σημαντικότητα των μεταβλητών σύμφωνα με τη μέση απόλυτη τιμή SHAP που παρουσιάζουν. Η κυρίαρχη μεταβλητή με βάση το μέσο απόλυτο SHAP είναι ο δείκτης μόχλευσης συνολικές υποχρεώσεις προς σύνολο ενεργητικού. Είναι η αντί-

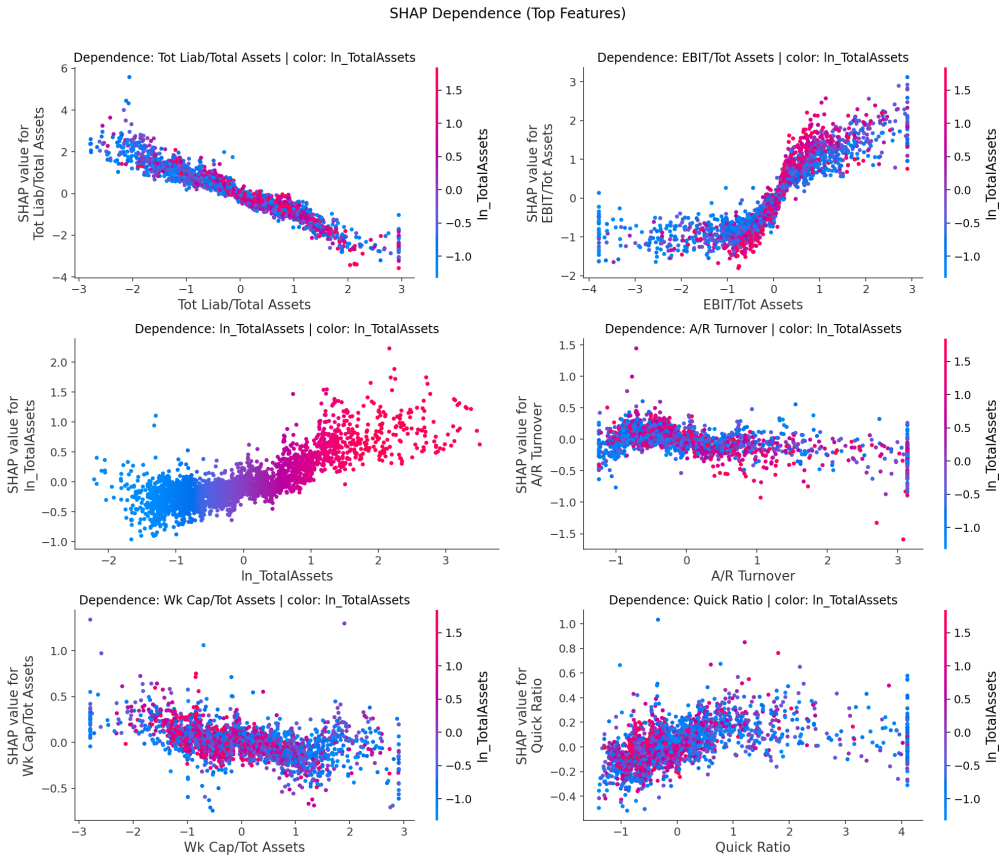
στροφη μεταβλητή που βρίσκεται στον αριθμητή του κλάσματος της απόστασης από ασυνέπεια τόσο στη θεωρία όσο και στη μοντελοποίηση του Credit Research Initiative. Ακολουθεί δεύτερο και πολύ κοντά ένας δείκτης κερδοφορίας, ο λόγος κέρδη προ φόρων προς σύνολο ενεργητικού. Η μεταβλητή αυτή δεν υπάρχει στο τύπο της απόστασης από ασυνέπεια παρ'όλα αυτά ενισχύει τη προβλεπτικότητα του μοντέλου μέσα από τη σχέση αιτίου αποτελέσματος, οι περισσότερο κερδοφόρες επιχειρήσεις έχουν μεγαλύτερη απόσταση απο πτώχευση. Πολύ κοντά στην 3η και 4η θέση είναι η κατηγορική μεταβλητή χώρα και σύνολο ενεργητικού μια μεταβλητή που δείχνει το μέγεθος της επιχείρησης. Μέτρια επίδραση έχει η μεταβλητή κλάδος δραστηριότητας, ενώ στις τρεις τελευταίες θέσεις είναι οι συμπληρωματικές μεταβλητές που αφορούν τις πωλήσεις, το κεφάλαιο κίνησης και τη ρευστότητα δείχνοντας μας ότι αυτά τα μεγέθη μπορούν να χρησιμοποιηθούν μόνο συμπληρωματικά για την πρόβλεψη της απόστασης από πτώχευση.



Σχήμα 5.5: Κατάταξη των μεταβλητών με βάση τη σημαντικότητά τους

Στο Σχήμα 5.6 παρουσιάζονται τα γραφήματα εξάρτησης (dependence plots). Μοιαζουν αρκετά με γραφήματα γραμμικής συσχέτισης μεταβλητών μόνο που στην περίπτωση τους παρουσιάζουν γραμμικά πως μεταβάλλεται η τιμή SHAP με βάση την τιμή της υποκείμενης μεταβλητής. Οι παρατηρήσεις είναι χρωματισμένες με βάση το μέγεθος της επιχείρησης (μέγεθος συνολικού ενεργητικού, κόκκινα μεγαλύτερα μεγέθη). Στον οριζόντιο άξονα παρουσιάζονται οι τιμές των μεταβλητών σε κανονικοποιημένη μορφή (μηδενικός μέσος, μοναδιαία τυπική απόκλιση). Η ισχυρότερη μεταβλητή του δείγματος (Total liabilities / Total Assets) παρουσιάζει γραμμικά φθίνον τιμή SHAP καθώς αυξάνονται οι συνολικές υποχρεώσεις προς το ενεργητικό δείχνοντας ότι υπάρχει ξεκάθαρη αρνητική επίδραση της μεταβλητής σε όλο το εύρος τιμών της. Η διασπορά των παρατηρήσεων ειναι επίσης μικρή κάτι που ενισχύει την προβλεπτική ικανότητά τής. Το ποιο ενδιαφέρον γράφημα

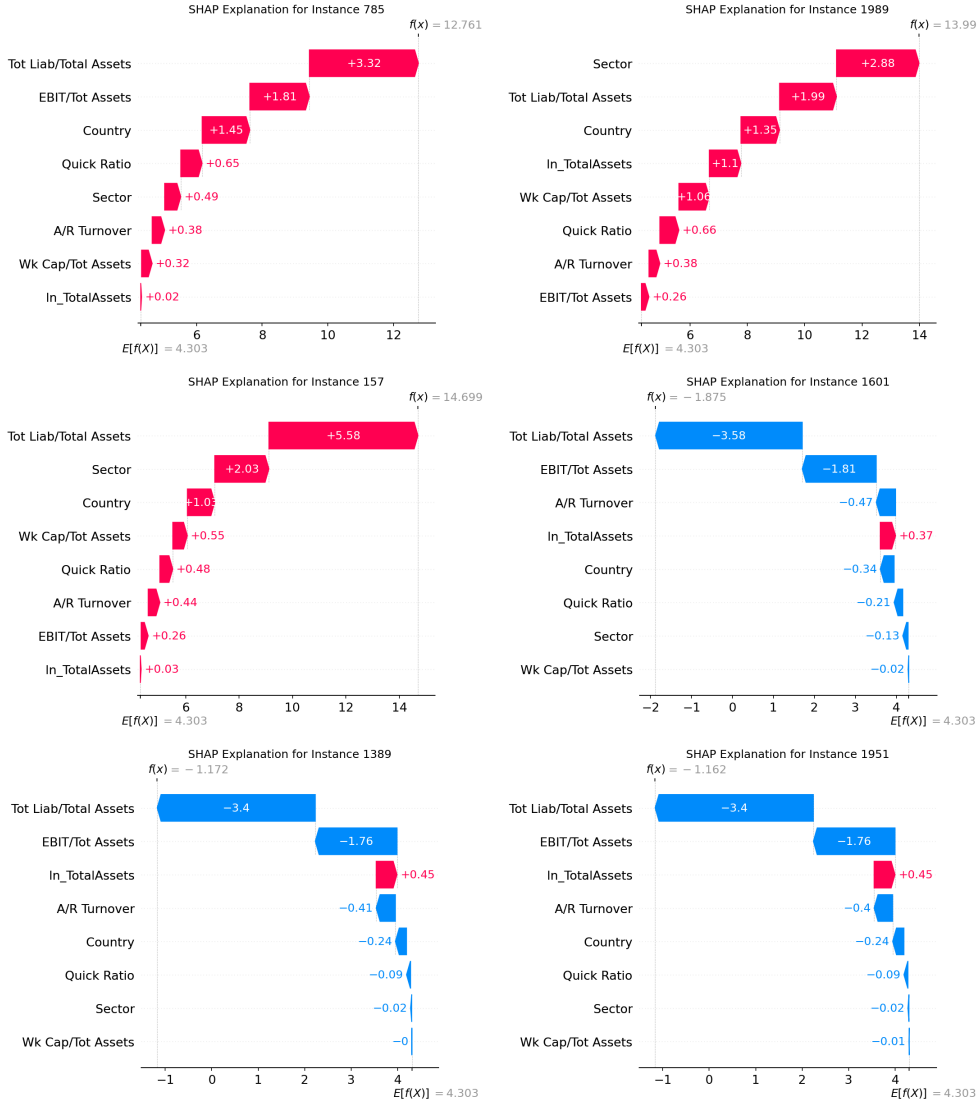
παρουσιάζει η κερδοφορία (EBIT/Total Assets). Σύμφωνα με αυτό στο εύρος της μεταβλητής από -1 έως 1 (σε κανονικοποιημένη μορφή) παρατηρούμε μια εκθετική αύξηση των τιμών SHAP και ύστερα αυξάνονται με μικρότερο ρυθμό. Με βάση αυτό μπορούμε να πούμε ότι αν ο δείκτης της κερδοφορίας είναι τουλάχιστον -1 και μεγαλύτερος (σε κανονικοποιημένη μορφή) τόσο μεγαλύτερη είναι η οριακή επίδραση στην απόσταση από ασυνέπεια και βελτιώνεται ο πιστωτικός κίνδυνος. Ενδιαφέρον είναι επίσης ότι το μεγαλύτερο ποσοστό των μεγαλύτερων επιχειρήσεων (κόκκινες παρατηρήσεις) συγκεντρώνεται σε αυτό το έυρος. Στο γράφημα του μεγέθους της επιχείρησης έχουμε αύξον τιμή SHAP αλλά με μεγαλύτερη διασπορά των παρατηρήσεων. Οι υπόλοιπες τρεις μεταβλητές παρουσιάζουν αρκετά ασθενή αποτελέσματα ως προς την εξάρτηση της τιμής SHAP και του εύρους των μεταβλητών.



Σχήμα 5.6: Γραφήματα εξάρτησης τιμών SHAP και μεταβλητών

Στο Σχήμα 5.7 παρουσιάζονται τα γραφήματα καταρράκτης (Waterfall plots). Πολλές φορές ενδιαφερόμαστε να αναλύσουμε συγκεκριμένα πως επηρεάζεται μια παρατήρηση με βάση της μεταβλητές της. Αυτό μπορούμε πολύ απλά να το εφαρμόσουμε με τις τιμές SHAP και να δούμε τις οριακές επιδράσεις των μεταβλητών πάνω στη συγκεκριμένη παρατήρηση σε σχέση με τη τιμή βάσης. Για την καλύτερη ερμηνεία επιλέχθηκαν οι τρεις παρατηρήσεις με τη μικρότερη και μεγα-

λύτερη τιμή απόστασης απο ασυνέπεια για να δούμε τι είναι αυτό που προβλέπει το μοντέλο ότι τις φέρνει στη συγκεκριμένη τιμή.



Σχήμα 5.7: Γραφήματα μεμονωμένων επιδράσεων των μεταβλητών σε παρατηρήσεις

6. Συμπεράσματα

Σκοπός τη παρούσας εργασίας ήταν η διερεύνηση των μοντέλων αγοράς τύπου Black-Scholes-Merton και η σύνδεση τους με της σύγχρονες μεθόδους πρόβλεψης. Τα μοντέλα αυτά χρησιμοποιούνται σε διάφορες παραλλαγές ακόμα και σήμερα για την εκτίμηση της πιθανότητας πτώχευσης μιας συγκεκριμένης οντότητας, από ερευνητικά κέντρα (CRI), χρηματοπιστωτικά ιδρύματα και οίκους αξιολόγησης (Moody's). Η χρησιμοποίηση κοινών μεγεθών της επιχείρησης και στοιχεία από τις μελλοντικές βλέψεις της αγοράς είναι δύο λόγοι που έχουν γίνει δημοφιλή, από την άλλη, η δυσκολία έγκειτε στον τρόπο υπολογισμού τους. Το ερώτημα που τέθηκε στην εργασία ήταν, μπορούν τα μοντέλα αυτά να αναπαρασταθούν ως ένα σημείο από τις σύγχρονες μεθόδους μηχανικής μάθησης; Η ανεξάρτητες μεταβλητές που χρησιμοποιήθηκαν είναι δημοσίως διαθέσιμες ενώ η αναπαραγωγή ενός τέτοιου εκπαιδευμένου μοντέλου μηχανικής μάθησης από κάποιον τρίτο είναι σχετικά εύκολη διαδικασία.

Τα αποτελέσματα έδειξαν ότι με ένα ισχυρό sMAPE της τάξης του 15.7% (SVR) στο πρόβλημα παλινδρόμησης που αναλύθηκε, τα μοντέλα μηχανικής μάθησης αντιπροσωπεύουν σε καλό βαθμό τον πιστωτικό κίνδυνο που προβλέπει το μοντέλο Black-Scholes-Merton . Συγχρόνως, μεθόδοι όπως οι XGBoost και SVR παρουσιάζουν στατιστικά σημαντικά καλύτερα αποτελέσματα από το μοντέλο βάσης, την γραμμική παλινδρόμηση. Σε περίπτωση που ο ενδιαφερόμενος χρήστης των μοντέλων θελήσει να αναλύσει τις ανεξάρτητες μεταβλητές σε ένα ερμηνευτικό πλαίσιο, το πρόβλημα της μη-διαφάνειας (black box methods) έχει ξεπεραστεί μέσω του αλγόριθμου SHAP, για την ερμηνεία της συμβολής των μεταβλητών στην τελική πρόβλεψη .

Περαιτέρω έρευνα θα μπορούσε να υπάρξει σε διαφορετικά γεωγραφικά μέρη του κόσμου. Για παράδειγμα οι ΗΠΑ παρουσιάζουν τη μεγαλύτερη συγκέντρωση κεφαλαίων, χαλαρότερη ρυθμιστική πολιτική για ανάδειξη της καινοτομίας, διαφορετική κουλτούρα ηγεσίας και management ή η Ασιατική ήπειρος που παρουσιάζει διαφορετική κουλτούρα επενδύσεων, αμφισβητούμενα πολιτικά συστήματα και αναδυόμενες χώρες. Σε μια άλλη διάσταση, η έρευνα θα μπορούσε να επεκταθεί στη χρήση ενός μοντέλου μηχανικής μάθησης σε μη εισηγμένες στις χρηματαγορές επιχειρήσεις, για την πρόβλεψη της απόστασης από ασυνέπεια. Με τα κατάλληλα δεδομένα θα μπορούσε να είναι

εφικτή η αναπαράσταση ενός μοντέλου αγοράς από μία τέτοια μέθοδο χρησιμοποιώντας χρηματο-οικονομικούς δείκτες.

Βιβλιογραφία

Afik, Z., Arad, O., & Galil, K. (2016). Using Merton model for default prediction: An empirical assessment of selected alternatives. *Journal of Empirical Finance*, *35*, 43–67.

Agarwal, V., & Taffler, R. (2008). Comparing the performance of market-based and accounting-based bankruptcy prediction models. *Journal of banking & finance*, *32*(8), 1541–1551.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631.

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, *23*(4), 589–609.

Bharath, S. T., & Shumway, T. (2008). Forecasting default with the Merton distance to default model. *The Review of Financial Studies*, *21*(3), 1339–1369.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of political economy*, *81*(3), 637–654.

Bohn, J., & Crosbie, P. (2002). Modeling default risk. *Moody's KMV*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Campbell, J. Y., Hilscher, J., & Szilagyi, J. (2008). In search of distress risk. *The Journal of finance*, *63*(6), 2899–2939.

Charitou, A., Dionysiou, D., Lambertides, N., & Trigeorgis, L. (2013). Alternative bankruptcy prediction models using option-pricing theory. *Journal of Banking & Finance*, *37*(7), 2329–2341.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Credit Research Initiative. (2023). NUS Credit Research Initiative Technical Report.

Credit Research Initiative of the National University of Singapore. (2022). *Probability of Default (PD) White Paper*. https://nuscri.org/en/white_paper/

Doumpos, M., Niklis, D., Zopounidis, C., & Andriosopoulos, K. (2015). Combining accounting data and a structural model for predicting credit ratings: Empirical evidence from European listed firms. *Journal of Banking & Finance*, *50*, 599–607.

Duan, J.-C., & Wang, T. (2012). Measuring distance-to-default for financial and non-financial firms. *World Scientific Book Chapters*, 95–108.

Duffie, D., Saita, L., & Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, *83*(3), 635–665.

Frecka, T. J., & Hopwood, W. S. (1983). The effects of outliers on the cross-sectional distributional properties of financial ratios. *Accounting Review*, 115–128.

Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of accounting studies*, *9*(1), 5–34.

Hull, J. C. (2009). *Options, Futures, and Other Derivatives* (7η έκδοση). Pearson.

Kealhofer, S. (2003). Quantifying credit risk I: default prediction. *Financial Analysts Journal*, *59*(1), 30–44.

Lev, B., & Sunder, S. (1979). Methodological issues in the use of financial ratios. *Journal of accounting and economics*, *1*(3), 187–210.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Merton, R. C. (1973). Rational theory of option pricing. *Bell Journal of Economics and Management Science*, *4*(1), 141–183.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance*, *29*(2), 449–470.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109–131.

Pasiouras, F., Gaganis, C., & Zopounidis, C. (2007). Multicriteria decision support methodologies for auditing decisions: The case of qualified audit reports in the UK. *European Journal of Operational Research*, *180*(3), 1317–1330.

Vassalou, M., & Xing, Y. (2004). Default risk in equity returns. *The journal of finance*, *59*(2), 831–868.

A. Παράρτημα

A.1 Περιγραφικά δεδομένα δείγματος

Πίνακας A.1: Παρατηρήσεις ανά χώρα και έτος

Country	2019	2020	2021	2022	2023	All
Austria	25	35	35	34	34	163
Belgium	40	42	43	44	44	213
Cyprus	13	13	13	15	11	65
Denmark	30	33	33	37	38	171
Finland	80	79	82	90	97	428
France	207	230	240	249	257	1183
Germany	246	258	259	271	264	1298
Greece	68	71	70	70	61	340
Italy	126	155	163	174	187	805
Luxembourg	11	14	14	16	17	72
Netherlands	45	48	51	55	55	254
Norway	72	75	84	110	119	460
Portugal	28	28	27	27	25	135
Republic of Ireland	14	18	21	22	22	97
Spain	59	62	67	70	73	331
Sweden	102	109	112	138	150	611
Switzerland	97	109	109	116	115	546
United Kingdom	279	382	391	415	428	1895
All	1542	1761	1814	1953	1997	9067

Πίνακας Α.2: Παρατηρήσεις ανά χώρα και κλάδο δραστηριότητας

Country	Basic Materials	Consumer Cyclicals	Consumer Non-Cyclicals	Energy	Industrials	Technology	All
Austria	31	24	14	10	53	31	163
Belgium	45	40	48	15	22	43	213
Cyprus	15	18	10	5	11	6	65
Denmark	15	41	20	7	48	40	171
Finland	42	92	36	5	166	87	428
France	97	389	118	47	268	264	1183
Germany	127	294	79	44	354	400	1298
Greece	57	58	40	34	116	35	340
Italy	43	304	59	38	222	139	805
Luxembourg	18	4	4	10	10	26	72
Netherlands	33	28	44	20	61	68	254
Norway	36	44	59	115	133	73	460
Portugal	30	49	10	10	15	21	135
Republic of Ireland	9	20	14	0	27	27	97
Spain	58	78	20	13	113	49	331
Sweden	46	135	30	12	180	208	611
Switzerland	74	113	58	5	203	93	546
United Kingdom	195	529	183	141	513	334	1895
All	971	2260	846	531	2515	1944	9067

Πίνακας Α.3: Μέση απόσταση από ασυνέπεια ανά χώρα και έτος

Country	2019	2020	2021	2022	2023
Austria	4,56	3,68	4,58	3,68	3,92
Belgium	5,49	4,39	5,57	5,07	5,01
Cyprus	4,09	3,44	3,96	3,98	3,46
Denmark	6,05	4,44	5,49	4,14	4,09
Finland	4,98	4,33	5,44	4,16	4,29
France	4,53	3,67	4,50	4,14	4,19
Germany	4,64	3,79	4,74	4,18	4,26
Greece	3,26	2,53	3,41	3,31	3,58
Italy	4,32	3,49	4,15	3,69	3,86
Luxembourg	4,29	3,17	4,22	3,21	3,31
Netherlands	5,31	4,33	5,51	4,87	4,80
Norway	3,44	2,90	3,84	3,62	3,36
Portugal	3,38	2,55	3,16	3,32	3,75
Republic of Ireland	6,33	4,44	5,55	5,04	4,81
Spain	4,36	3,30	4,01	3,84	4,26
Sweden	4,58	4,09	4,81	3,67	3,55
Switzerland	6,10	5,42	6,86	5,45	5,29
United Kingdom	5,61	3,82	4,90	4,58	4,16

Πίνακας Α.4: Μέση απόσταση από ασυνέπεια ανά κλάδο δραστηριότητας και έτος

Sector	2019	2020	2021	2022	2023
Basic Materials	4,66	3,82	5,01	4,27	4,24
Consumer Cyclicals	4,55	3,33	4,28	3,8	3,77
Consumer Non-Cyclicals	5,96	4,86	5,97	5,32	4,82
Energy	3,81	2,55	3,14	3,14	3,05
Industrials	4,64	3,59	4,56	4,06	4,08
Technology	5,21	4,54	5,35	4,59	4,57

Πίνακας Α.5: Μέση απόσταση από ασυνέπεια ανά χώρα και κλάδο δραστηριότητας

Country	Basic Materials	Consumer Cyclicals	Consumer Non-Cyclicals	Energy	Industrials	Technology
Austria	4,68	3,71	4,91	3,53	3,76	4,00
Belgium	4,86	4,64	6,06	4,54	4,50	5,23
Cyprus	6,24	2,99	4,26	4,59	2,08	1,87
Denmark	4,15	5,5	5,92	3,19	4,27	4,61
Finland	4,60	4,43	5,93	6,75	3,91	5,52
France	4,28	3,73	4,36	3,81	4,40	4,67
Germany	3,92	4,05	4,88	3,81	4,15	4,74
Greece	3,02	4,14	4,90	2,65	2,04	4,46
Italy	4,74	3,50	5,64	3,09	3,61	4,37
Luxembourg	3,38	1,64	3,90	4,18	4,59	3,36
Netherlands	4,58	4,26	5,45	4,42	5,05	5,20
Norway	3,89	3,48	5,10	2,32	3,09	4,29
Portugal	4,57	2,57	4,90	3,56	1,48	3,08
Republic of Ireland	6,07	4,84	3,50	--	5,15	6,02
Spain	4,97	3,31	4,25	4,48	3,61	4,35
Sweden	3,39	3,55	4,64	4,17	4,46	4,15
Switzerland	5,80	5,02	6,94	3,15	5,86	6,08
United Kingdom	4,34	4,09	5,94	2,62	4,5	5,52

A.2 Υπερ-παραμέτροι επιλογής μέσω διαδικασίας βελτιστοποίησης

Πίνακας Α.6: Υπερ-παραμέτροι επιλογής

Μοντέλα	Υπερ-παράμετροι	Τελικές τιμές
Linear Regression		
	--	--
Ensemble methods		
Random forest	No. of trees	1050
	Max tree depth	28
	Min samples leaf	1
	Max features	0.14
	Max samples	0.81
XGBoost	No. of trees	1550
	Max tree depth	10
	Learning rate	0.017
	γ	0.10
	λ	3.77
	Subsample	0.64
Support Vector Regression		
	C	5.97
	ϵ	0.58
	γ	0.456
	Kernel	RBF (fixed)
Neural Network		
	Number of hidden layers	3
	Hidden layers dimensions	26, 22, 17
	Dropout layers	Layer 2 & Layer 3
	Dropout probability	0.1407
	Learning rate	0.0041
	Batch size	93
	Epochs	64
	Optimizer	Adam (fixed)