ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Ισμίρογλου Βασιλικής

Πρόβλεψη Έκβασης Τηλεφωνικής Προώθησης Τραπεζικών Υπηρεσιών Με Τη Χρήση Μεθόδων Αναγνώρισης Προτύπων

Επιβλέπων
Κουσουρής Κωνσταντίνος
Αναπλ. Καθηγητής Ε.Μ.Π

Αθήνα, Οκτώβριος 2022

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Ισμίρογλου Βασιλικής

Πρόβλεψη Έκβασης Τηλεφωνικής Προώθησης Τραπεζικών Υπηρεσιών Με Τη Χρήση Μεθόδων Αναγνώρισης Προτύπων

Επιβλέπων
Κουσουρής Κωνσταντίνος
Αναπλ. Καθηγητής Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 13η Οκτωβρίου 2022

.....	.....	.....
Κουσουρής Κωνσταντίνος	Τσιπολίτης Γεώργιος	Δρακοπούλου Ευαγγελία
Αναπλ. Καθηγητής Ε.Μ.Π	Καθηγητής Ε.Μ.Π	Ερευνήτρια, ΕΚΕΦΕ «Δημόκριτος»

Αθήνα, Οκτώβριος 2022

.....
Ισμίρογλου Βασιλική

Ευχαριστίες

Με την εργασία αυτή ολοκληρώνεται ο προπτυχιακός κύκλος σπουδών μου στο Ε.Μ.Π. Θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή, Κουσουρή Κωνσταντίνο, για την καθοδήγησή του στο μάθημα αναγνώρισης προτύπων καθώς και την εμπιστοσύνη που μου έδειξε καθ΄ όλη τη διάρκεια εκπόνησης της διπλωματικής εργασίας.

Θα ήθελα επιπλέον να ευχαριστήσω την οικογένειά μου, που με στήριξε αδιάκοπα τα χρόνια των σπουδών μου και δεν έπαψαν ποτέ να πιστεύουν σε εμένα.

Τέλος, είμαι ευγνώμων για όλους τους φίλους μου, συμφοιτητές και μη, που έμειναν δίπλα μου στα δύσκολα, στις ανησυχίες και στις στενοχώριες αλλά μοιράστηκαν μαζί μου και όλες τις ευχάριστες στιγμές. Δεν θα τα είχα καταφέρει χωρίς εσάς.

Περίληψη

Η παρούσα εργασία πραγματεύεται την υλοποίηση μεθόδων μηχανικής μάθησης για ανάλυση δεδομένων, που αφορούν την τηλεφωνική προώθηση τραπεζικών υπηρεσιών. Αρχικά επιδιώκεται η ανάπτυξη μοντέλων, ικανών να ταξινομήσουν το σύνολο δεδομένων, με έμφαση στις δυνατότητές τους για γενίκευση και κατ΄ επέκταση πρόβλεψη. Συγκεκριμένα υλοποιήθηκαν ταξινομητές λογιστικής παλινδρόμησης, νευρωνικού δικτύου, τυχαίου δάσους και ενισχυμένων δέντρων με τους αλγορίθμους gradient boost και adaptive boost. Σε δεύτερη φάση, εξάγεται η σπουδαιότητα των χαρακτηριστικών για κάθε μέθοδο που μπορεί να οδηγήσει σε ερμηνεύσημα συμπεράσματα για τα δεδομένα.

Από τα αποτελέσματα αναδεικνύεται ο ουσιαστικός ρόλος των συγκεκριμένων εργαλείων σε διαδικασίες πρόβλεψης, καθώς όλα τα μοντέλα επιτυγχάνουν την αναγνώριση εξαρτήσεων μεταξύ της κλάσης και των χαρακτηριστικών. Κατά την σύγκριση των μεθόδων αναδύονται οι άμεσοι περιορισμοί που τίθενται από τη μορφή του συνόλου δεδομένων. Προκύπτει, ότι οι αμερόληπτες ως προς αυτή μέθοδοι έχουν μικρά περιθώρια βελτίωσης, σε σχέση με τα πιο απλά μοντέλα, προτού υπερεκπαιδευτούν. Τέλος, σε ότι αφορά τη σπουδαιότητα των χαρακτηριστικών, παρά τις μικρές διαφορές μεταξύ προσεγγίσεων, παρατηρείται συνέπεια ως προς τα κυρίαρχα χαρακτηριστικά. Συνεπάγεται ότι η συμβολή τους δεν περιορίζεται στις ιδιαίτερες διαδικασίες που αφορούν τη κάθε μέθοδο, αλλά δύναται να αποτυπώνει μια πραγματική επιρροή στην απόφαση του πελάτη.

Λέξεις κλειδιά: αναγνώριση προτύπων, μηχανική μάθηση, ταξινόμηση, πρόβλεψη, νευρωνικά δίκτυα, δέντρα απόφασης, λογιστική παλινδρόμηση, προώθηση τραπεζικών υπηρεσιών

Abstract

This work deals with the implementation of machine learning methods in order to analyze bank marketing data. Initially, we attempt the development of models capable of classifying the data, with an emphasis on their potential for generalization and, by extension, prediction. In particular, logistic regression, neural network, random forest and boosted tree classifiers were implemented using the gradient boost and adaptive boost algorithms. Furthermore, the feature importance is extracted for each method aiming at interpretable conclusions regarding the data.

The results highlight the essential role of the aforementioned methods, since they all succeed in recognizing the dependencies between the class and features. When comparing the models, the immediate limitations set by the shape of the data emerge. It becomes apparent that low bias models have small margins for improvement, over simpler ones, before they overfit. Finally, despite the minimal differences between approaches, the observed results show consistency in regards to feature importance. It follows that the their contribution is not limited to the particular procedures within each method, but may reflect a real influence on the customer's decision.

Keywords: pattern recognition, machine learning, classification, prediction, neural networks, decision trees, logistic regression, bank marketing

Περιεχόμενα

Κατάλογος Σχημάτων

Εισαγωγή

1.1 Επιχειρηματική ευφυΐα και μηχανική μάθηση

Τα συστήματα υποστήριξης λήψης αποφάσεων (Decision Support Systems/DSS)[1] αποτελούν αναπόσπαστο κομμάτι των λειτουργιών μεγάλων επιχειρήσεων, με στόχο αυτές να παραμείνουν ανταγωνιστικές στην σύγχρονη οικονομία. Βασίζονται στην ανάλυση μεγάλων συνόλων δεδομένων, για την επεξήγηση και την πρόβλεψη φαινομένων και συμπεριφορών που επηρεάζουν τις λειτουργίες τους, καθώς και για τη ρύθμιση των λειτουργιών αυτών. Κάτω από τον όρο «Επιχειρηματική Ευφυΐα» θα βρει κανείς πλήθος διαδικασιών που εφαρμόζουν τις επιστήμες της στατιστικής, των υπολογιστών και πιο πρόσφατα της μηχανικής μάθησης ως επέκταση αυτών, για να επιτύχουν τους παραπάνω στόχους.

Εστιάζοντας στις ανάγκες πρόβλεψης, η ταξινόμηση μέσω μηχανικής μάθησης αποτελεί αντικείμενο ύψιστης σημασίας. Στην πλειοψηφία τους, οι βάσεις δεδομένων περιέχουν υπέρογκες ποσότητες πληροφορίας, χωρίς αυτό να συνεπάγεται την άμεση χρησιμότητά της. Το γεγονός αυτό κάνει την εξαγωγή ουσιωδών συμπερασμάτων ένα αρκετά πολυσύνθετο ζήτημα. Ένα καλά κατασκευασμένο μοντέλο ταξινόμησης δύναται να προβλέψει την κλάση ενός πελάτη, δηλαδή μία αποτύπωση της συμπεριφοράς του, αξιοποιώντας σε κατάλληλο βαθμό τις διαθέσιμες πληροφορίες, με τρόπο μη εμφανή με «γυμνό μάτι». Αντίστοιχες διαδικασίες μπορεί να αφορούν την επιτυχία ενός προϊόντος, ή το ρίσκο μίας επένδυσης.

Ένα επιπλέον εργαλείο που προκύπτει άμεσα από τις μεθόδους ταξινόμησης είναι και αυτό της εξόρυξης δεδομένων (Data Mining). Το όνομα θεωρείται από κάποιους παραπλανητικό, καθώς δεν αναφέρεται τόσο στην εύρεση δεδομένων, όσο στην εξόρυξη «κρυφής» πληροφορίας από αυτά. Η μελέτη λειτουργικών ταξινομητών, μπορεί να οδηγήσει στην εξαγωγή των χαρακτηριστικών, που συμβάλουν κυρίαρχα στην επιτυχία του. Κατ΄ επέκταση, δύναται να ερμηνευτούν και ως εκείνα που επηρεάζουν πιο άμεσα τη συμπεριφορά του πελάτη ή του προϊόντος.

Από τα παραπάνω, γίνεται εμφανές ότι με τη χρήση εργαλείων μηχανικής μάθησης, οι επιχειρήσεις μπορούν να οδηγηθούν σε συμπεράσματα για τους πελάτες τους, τα προϊόντα καθώς και τη γενικότερη συμπεριφορά της αγοράς και της οικονομίας. Δύναται να

ρυθμιστούν έτσι, οι στοχευμένες προωθητικές ενέργειες, ο σχεδιασμός και η ανάπτυξη μελλοντικών προϊόντων και υπηρεσιών με καλά ορισμένο κοινό, καθώς και το ρίσκο που αυτά συνεπάγονται.

1.2 Αντικείμενο της εργασίας

Αντικείμενο της παρούσας διπλωματικής είναι η αξιοποίηση των εργαλείων μηχανικής μάθησης για την υλοποίηση μοντέλων, ικανών να προβλέψουν τη συμπεριφορά πελατών. Η προσέγγιση του ζητήματος μπορεί να αναλυθεί σε δύο στάδια. Αρχικά επιδιώκεται η ανάπτυξη ταξινομητών που δύναται να συμπεράνουν, με βάση τα διαθέσιμα δεδομένα, εάν ένας πελάτης θα προβεί σε κάποια ενέργεια και πιο συγκεκριμένα αν θα αποδεχθεί να πραγματοποιήσει προθεσμιακή κατάθεση ως αποτέλεσμα τηλεφωνικής προώθησης από πλευράς της τράπεζας. Ένα τέτοιο μοντέλο μπορεί άμεσα να εφαρμοστεί με σκοπό το στοχευμένο marketing. Το δεύτερο στάδιο αφορά τη σπουδαιότητα των διαθέσιμων χαρακτηριστικών και άρα το βαθμό στον οποίο αυτά συμβάλλουν στην τελική απόφαση.

1.3 Δομή της εργασίας

- **Κεφάλαιο 2 - Θεωρητικό Υπόβαθρο**: Περιγράφεται συνοπτικά το αντικείμενο της μηχανικής μάθησης ενώ αναλύεται η θεωρία και η δομή των συγκεκριμένων μοντέλων και εργαλείων που θα υλοποιηθούν στην παρούσα εργασία.

- **Κεφάλαιο 3 - Πρακτικό Μέρος**: Γίνεται πρακτική εφαρμογή των μεθόδων που έχουν αναλυθεί στη θεωρία, σε σύνολο δεδομένων, προερχόμενο από μία Πορτογαλική τράπεζα, που αφορά προωθητικές ενέργειες μέσω τηλεφώνου. Αρχικά αναφέρονται τα βασικά χαρακτηριστικά των δεδομένων ενώ επιπλέον αναλύονται οι διαδικασίες που πραγματοποιήθηκαν στο στάδιο προεπεξεργασίας. Στη συνέχεια παρουσιάζονται τα βήματα που ακολουθήθηκαν για την επιλογή των παραμέτρων του κάθε μοντέλου καθώς και τα αποτελέσματα αυτών. Τέλος, οι διάφορες προσεγγίσεις συγκρίνονται με ιδιαίτερη έμφαση στη σπουδαιότητα των χαρακτηριστικών.

- **Κεφάλαιο 4 - Συμπεράσματα**: Παρατίθενται τα συμπεράσματα της εργασίας καθώς και τα πιθανά πλαίσια μελλοντικής μελέτης.

Θεωρητικό Υπόβαθρο

2.1 Μηχανική Μάθηση

Η μηχανική μάθηση αποτελεί έναν σύγχρονο και συνεχώς αναπτυσσόμενο κλάδο της επιστήμης υπολογιστών με κυρίαρχο στόχο την «εκπαίδευση» υπολογιστικών συστημάτων. Συγκεκριμένα αφορά τις μεθόδους και τους αλγορίθμους που επιτρέπουν στον υπολογιστή να μάθει από δεδομένα, χωρίς να έχει ρητά προγραμματιστεί και να εξάγει αποτελέσματα με τη μορφή προβλέψεων και αποφάσεων.

Εντοπίζονται τρεις βασικές προσεγγίσεις στο ζήτημα της μάθησης:

- **Μάθηση με επίβλεψη (supervised learning)**: Επιδιώκεται η αναγνώριση προτύπων σε σύνολα δεδομένων τα οποία περιέχουν τα επιθυμητά αποτελέσματα. Συνεπώς ο στόχος είναι να βρεθούν οι σχέσεις εξάρτησης μεταξύ των εισόδων και του αποτελέσματος. Σε αυτή την κατηγορία ανήκουν οι διαδικασίες παλινδρόμησης και ταξινόμησης.

- **Μάθηση χωρίς επίβλεψη (unsupervised learning)**: Και εδώ τα μοντέλα εκπαιδεύονται με αξιοποίηση ενός συνόλου δεδομένων, με ουσιαστική διαφορά ότι αυτό δεν είναι ήδη κατηγοριοποιημένο. Οι μέθοδοι μηχανικής μάθησης αναζητούν ομοιότητες στα διαθέσιμα παραδείγματα και στοχεύουν να τα κατηγοριοποιήσουν με βάση αυτές. Η διαδικασία είναι γνωστή ως clustering.

- **Ενισχυτική μάθηση (reinforcement learning)**: Σε αυτή τη προσέγγιση αντί ενός συνόλου δεδομένων κατασκευάζεται ένα δυναμικό περιβάλλον. Το υπολογιστικό πρόγραμμα αφήνεται να αλληλεπιδράσει με αυτό και η επίδοσή του αξιολογείται ανάλογα, με κάποια μορφή επιβράβευσης. Λόγω της γενικότητας της συγκεκριμένης μεθόδου μελετάται ως αντικείμενο πολλών κλάδων, ενώ συνήθεις εφαρμογές είναι τα συστήματα αυτόματης πλοήγησης ή η εκπαίδευση του υπολογιστή στο να παίζει παιχνίδια (σκάκι, go, κ.α.)

Μαζί με τα παραπάνω θα βρει κανείς και άλλες, πιο στοχευμένες προσεγγίσεις (self-learning, semi-supervised learning, κ.α.) καθώς και ένα σύνολο εργαλείων που αξιοποιούνται για την υποβοήθηση του κυρίαρχου μοντέλου (μείωση διαστάσεων, εκμάθηση χαρακτηριστικών κ.α.)

Εστιάζοντας στις διαδικασίες μάθησης με επίβλεψη, στόχος είναι η ταξινόμηση αντικειμένων σε κλάσεις. Τα αντικείμενα αυτά προκύπτουν από τις άμεσες ανάγκες της σύγχρονης πραγματικότητας και μπορεί να είναι κείμενα, εικόνες, ήχοι καθώς και οτιδήποτε άλλο δύναται να προσφέρει χρήσιμη πληροφορία. Αντίστοιχα προκύπτουν και οι πολλαπλές εφαρμογές της συγκεκριμένης επιστημονικής περιοχής, κάποιες από τις οποίες είναι οι παρακάτω:

- Αναγνώριση και επεξεργασία φωνής (Speech processing applications)
- Επεξεργασία Φυσικής Γλώσσας (Natural Language Processing)
- Υπολογιστική όραση (Computer vision)
- Εξόρυξη δεδομένων (Data mining)
- Εφαρμογές στην υπολογιστική Βιολογία
- Υποβοηθούμενη από τον υπολογιστή διάγνωση
- Διάφορα άλλα πρακτικά προβλήματα όπως ο εντοπισμός οικονομικής απάτης και ηλεκτρονικών επιθέσεων.

Η λίστα αυτή σαφώς δεν είναι πλήρης και ούτε δύναται να γίνει. Στη μεγάλη πλειοψηφία τους τα διάφορα πρακτικά προβλήματα που παρουσιάζονται, μπορούν να αντιμετωπιστούν ως προβλήματα εκμάθησης για τον κλάδο της μηχανικής μάθησης, ο οποίος συνεχώς θα επεκτείνεται.

2.2 Μοντέλα ταξινόμησης

Όπως έχει προαναφερθεί, η ταξινόμηση πρόκειται για διαδικασία μηχανικής μάθησης με επίβλεψη. Αυτό συνεπάγεται ότι στο διαθέσιμο σύνολο δεδομένων εμπεριέχεται το αποτέλεσμα που αντιστοιχεί στο κάθε παράδειγμα.

- Το κάθε δείγμα στο σύνολο χαρακτηρίζεται από τις *ανεξάρτητες μεταβλητές* (independent variables) ή απλώς *χαρακτηριστικά* (features). Αυτά περιγράφονται μέσω διανυσμάτων $\boldsymbol{x}_i$, $i = 1, 2, \ldots, N$ όπου N το μέγεθος του συνόλου.
- Η κατηγορία στην οποία ανήκει το κάθε δείγμα, εξαρτάται από τα χαρακτηριστικά του. Ονομάζεται κλάση (class), στόχος (target) ή εξαρτημένη μεταβλητή (dependent variable) και στη συνέχεια της παρούσας εργασίας θα συμβολίζεται ως ω_i, $i = 1, 2, \ldots, N$. Αν και το πλήθος των κλάσεων δεν έχει συγκεκριμένα όρια και εξαρτάται από τα δεδομένα, θα γίνει εστιασμένη ανάλυση στα προβλήματα δυαδικής ταξινόμησης. Κάποια από τα μοντέλα που θα μελετηθούν και υλοποιηθούν έχουν σχετικά απλές επεκτάσεις στο ζήτημα τις ταξινόμησης πολλών κλάσεων.

Συνεπάγεται από τα παραπάνω ότι, το αντικείμενο των ταξινομητών είναι να εξάγουν τις εξαρτήσεις των κλάσεων από τα χαρακτηριστικά. Εντοπίζουν δηλαδή, τα πρότυπα που πιθανώς υπάρχουν στα δεδομένα και τα συσχετίζουν με κάποια κλάση. Αυτό επιτυγχάνεται με την εύρεση των βέλτιστων ορίων διαχωρισμού, των ορίων δηλαδή, στον χώρο των χαρακτηριστικών, που καθορίζουν τις περιοχές της κάθε εξαρτημένης μεταβλητής. Για την επίλυση του προβλήματος υιοθετείται μία από τρεις διαφορετικές προσεγγίσεις[2]:

- Η πιο απλή μέθοδος είναι αυτή της διακριτικής συνάρτησης. Πρόκειται για την κατασκευή μίας συνάρτησης $y(\boldsymbol{x})$, η οποία επιστρέφει απευθείας, σαν αποτέλεσμα, την κλάση στην οποία ταξινομείται το κάθε χαρακτηριστικό.

- Η δεύτερη κατηγορία είναι αυτή των πιθανολογικών διακριτικών μοντέλων. Εδώ, επιδιώκεται η μοντελοποίηση των εκ των υστέρων πιθανοτήτων $p(\omega_i|\boldsymbol{x})$, από τις οποίες γίνεται τελικά η ανάθεση σε κλάσεις.

- Τέλος υπάρχουν τα παραγωγικά (generative) μοντέλα, τα οποία, μέσω εύλογων υποθέσεων, μοντελοποιούν τις δεσμευμένες πιθανότητες $p(\boldsymbol{x}|\omega_i)$ και υπολογίζουν τις εκ των υστέρων πιθανότητες με τον κανόνα του Bayes.

$$p(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i) \cdot p(\omega_i)}{p(\boldsymbol{x})} \tag{2.1}$$

Μία διαφορετική κατηγοριοποίηση που αφορά τα μοντέλα ταξινόμησης είναι αυτή της γραμμικότητας. Οι γραμμικοί ταξινομητές, υπολογίζουν τα όρια διαχωρισμού ως γραμμικές συναρτήσεις των χαρακτηριστικών. Προκύπτει εύλογα, ότι ελλείψει αυτού του περιορισμού, τα μη γραμμικά μοντέλα είναι ισχυρότερα. Αν και τα περισσότερα ρεαλιστικά προβλήματα ταξινόμησης δεν είναι γραμμικώς διαχωρίσιμα, οι γραμμικοί ταξινομητές αποτελούν ένα πολύ αποδοτικό εργαλείο και συνδυάζουν ικανοποιητικά αποτελέσματα με μικρούς χρόνους εκπαίδευσης.

2.2.1 Γραμμικοί Ταξινομητές

Ο αλγόριθμος perceptron

Ο αλγόριθμος perceptron [3][4] ανήκει στις μεθόδους υπολογισμού μίας συνάρτησης διάκρισης και εφαρμόζεται στην απλούστερη περίπτωση των δύο κλάσεων. Ορίζεται λοιπόν η συνάρτηση:

$$y(\boldsymbol{x}) = sign[\boldsymbol{w}^T\boldsymbol{x} + w_0] \tag{2.2}$$

έτσι ώστε:

- Αν $\boldsymbol{w}^T\boldsymbol{x} + w_0 > 0$, τότε το $\boldsymbol{x_n}$ ανήκει στην κλάση ω_1
- Αν $\boldsymbol{w}^T\boldsymbol{x} + w_0 < 0$, τότε το $\boldsymbol{x_n}$ ανήκει στην κλάση ω_2

Το όριο διαχωρισμού θα είναι το υπερεπίπεδο $y(\boldsymbol{x}) = 0$. Το διάνυσμα $\boldsymbol{w}$ είναι κάθετο στο υπερεπίπεδο διαχωρισμού και ονομάζεται διάνυσμα βαρών, ενώ το w_0 αναφέρεται στη διεθνή βιβλιογραφία ως bias ή threshold. Η βηματική συνάρτηση, αξιοποιείται για τη συσχέτισή του αποτελέσματος με την κάθε κλάση και είναι γνωστή ως συνάρτηση ενεργοποίησης (activation function)

Ο στόχος του μοντέλου είναι να υπολογίσει τις ποσότητες $\boldsymbol{w}, w_0$ ελαχιστοποιώντας το κόστος perceptron:

$$E_p(\boldsymbol{w}) = -\sum_{n \in Y} \boldsymbol{w}^T x_n t_n \tag{2.3}$$

όπου Y είναι το υποσύνολο των λανθασμένα ταξινομημένων διανυσμάτων, ενώ η ποσότητα $t_n \in \{-1, 1\}$ αντιπροσωπεύει την πραγματική κλάση. Σημειώνεται επιπλέον ότι στη σχέση 2.3 αξιοποιούνται τα επαυξημένα διανύσματα χαρακτηριστικών, στα οποία υπάρχει ο όρος $x_{0,n} = 1$ αντί για τον όρο κατωφλίου w_0. Εύκολα προκύπτει ότι η ποσότητα $\boldsymbol{w}^T x_n t_n$ είναι πάντα θετική. Για ένα διάνυσμα x_n που ανήκει στην κλάση ω_1, η λανθασμένη ταξινόμηση θα προκύψει από τη σχέση $\boldsymbol{w}^T x_n < 0$, ενώ ταυτόχρονα $t_n = -1$.

Από τα παραπάνω, συνεπάγεται, ότι η ταξινόμηση μετατρέπεται σε ένα πρόβλημα βελτιστοποίησης. Θα μπορούσε κανείς να προσπαθήσει να ελαχιστοποιήσει το κόστος, υπολογίζοντας τα σημεία μηδενισμού της παραγώγου, αλλά αυτή παρουσιάζει ασυνέχειες, όταν ένα διάνυσμα περνά από την περιοχή λανθασμένης ταξινόμησης στην άλλη. Για τον λόγο αυτό, αξιοποιείται η μέθοδος καθόδου προς την κατεύθυνση της παραγώγου (gradient descent).

Πρόκειται για μια επαναληπτική διαδικασία, κατά την οποία, η παράγωγος υπολογίζεται στα σημεία στα οποία ορίζεται ως:

$$\frac{\partial E_p(\boldsymbol{w})}{\partial \boldsymbol{w}} = -\sum_{n \in Y} x_n t_n \tag{2.4}$$

Στη συνέχεια, τα καινούργια βάρη κάθε φορά προκύπτουν από τη σχέση:

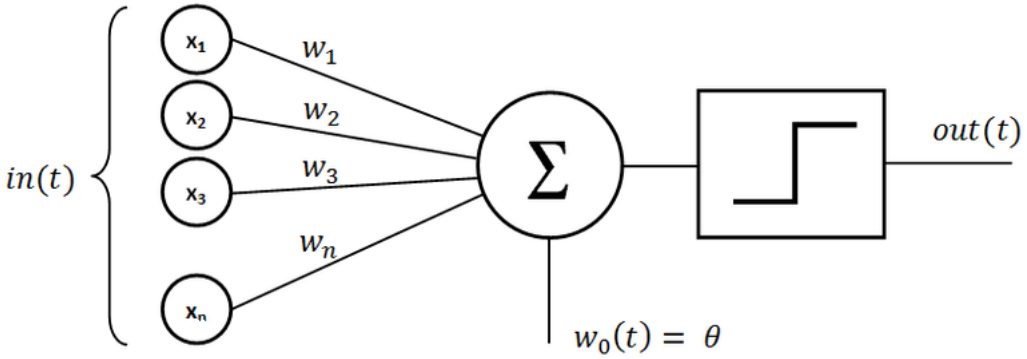
$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} - \eta \sum_{n \in Y} x_n t_n \tag{2.5}$$

όπου το η αντιπροσωπεύει τον ρυθμό εκμάθησης.

Αποδεικνύεται [3] ότι ο αλγόριθμος συγκλίνει μετά από πεπερασμένο αριθμό βημάτων αν οι κλάσεις είναι γραμμικά διαχωρίσιμες. Σε αντίθετη περίπτωση πρέπει να οριστεί μέγιστος αριθμός βημάτων. Μία γραφική αναπαράσταση του perceptron παρουσιάζεται στο Σχήμα 2.1. Το αντικείμενο του αλγορίθμου είναι γνωστό και ως νευρώνας λόγω της ομοιότητας του με τους βιολογικούς νευρώνες του εγκεφάλου.

Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση ή λογιστική διάκριση (logistic regression), ανήκει στην κατηγορία των πιθανολογικών διακριτικών μοντέλων. Στόχος είναι να προσδιορισθούν οι εκ των υστέρων πιθανότητες των κλάσεων $p(\omega_i|\boldsymbol{x})$ και με βάση αυτές να ταξινομηθούν



Σχήμα 2.1: Δομή του perceptron[5]

τα διανύσματα χαρακτηριστικών[2]. Στην περίπτωση των δύο κλάσεων ισχύει ότι:

$$\begin{aligned} p(\omega_i|\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|\omega_1)p(\omega_1)}{p(\boldsymbol{x}|\omega_1)p(\omega_1) + p(\boldsymbol{x}|\omega_2)p(\omega_2)} \\ &= \frac{1}{1+\exp(-\alpha)} = \sigma(\alpha) \end{aligned} \tag{2.6}$$

όπου:

$$\alpha = \ln \frac{p(\boldsymbol{x}|\omega_1)p(\omega_1)}{p(\boldsymbol{x}|\omega_2)p(\omega_2)} = \ln \frac{p(\omega_1|\boldsymbol{x})}{p(\omega_2|\boldsymbol{x})} \tag{2.7}$$

Ο λογάριθμος του λόγου πιθανοτήτων μοντελοποιείται ως γραμμική συνάρτηση των χαρακτηριστικών:

$$\ln \frac{p(\omega_1|\boldsymbol{x})}{p(\omega_2|\boldsymbol{x})} = w_0 + \boldsymbol{w}^T\boldsymbol{x} \tag{2.8}$$

και οι εκ των υστέρων πιθανότητες των κλάσεων δίνονται τελικά από τις σχέσεις:

$$p(\omega_1|\boldsymbol{x}) = \frac{\exp(w_0 + \boldsymbol{w}^T x)}{1+\exp(w_0 + \boldsymbol{w}^T x)} \tag{2.9}$$

$$p(\omega_2|\boldsymbol{x}) = \frac{1}{1+\exp(w_0 + \boldsymbol{w}^T x)} \tag{2.10}$$

Η εκπαίδευση του μοντέλου βασίζεται, για ακόμα μία φορά, στην ελαχιστοποίηση μίας συνάρτησης κόστους, το ρόλο της οποίας έχει η αρνητική συνάρτηση λογαριθμικής πιθανοφάνειας:

$$\begin{aligned} L(\boldsymbol{w}) &= -\ln p(\boldsymbol{\omega}|\boldsymbol{w}) \\ &= -\sum_{n=1}^{N}[\omega_n \ln y_n + (1-\omega_n)\ln(1-y_n)] + r(\boldsymbol{w}) \end{aligned} \tag{2.11}$$

Η βελτιστοποίηση ως προς τις παραμέτρους $\boldsymbol{w}, w_0$, γίνεται με τη χρήση αριθμητικών μεθόδων και πιο συγκεκριμένα, στην πρακτική εφαρμογή χρησιμοποιήθηκε ο αλγόριθμος

Broyden-Fletcher-Goldfarb-Shanno[6]. Ο όρος $r(\boldsymbol{w})$ έχει τη μορφή $\lambda\|\boldsymbol{w}\|_2$ και αποσκοπεί στην οριοθέτηση του μέτρου του διανύσματος βαρών, καθώς η βελτιστοποίηση τείνει να το απειρίσει.

2.2.2 Μη Γραμμικοί Ταξινομητές

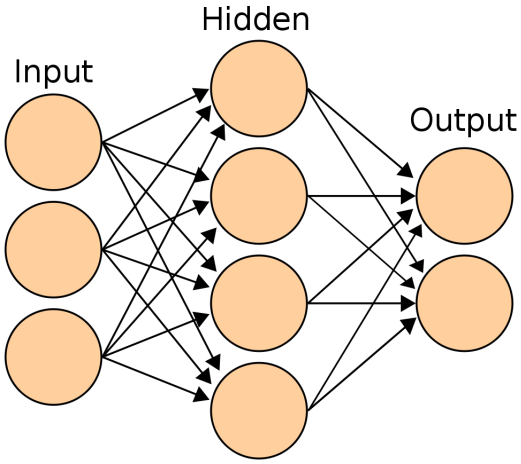
Νευρωνικά Δίκτυα

Γνωστά και ως perceprton πολλών επιπέδων, τα νευρωνικά δίκτυα αποτελούν μία επέκταση του αλγορίθμου perceptron για την αντιμετώπιση μη γραμμικά επιλύσιμων προβλημάτων ταξινόμησης. Στην πιο απλή περίπτωση των δύο επιπέδων, κατασκευάζεται ένα «κρυφό» επίπεδο με $k = 1, 2, \ldots, K$ νευρώνες της μορφής:

$$y_k(\boldsymbol{x}) = f(\boldsymbol{w}_k^T \boldsymbol{x} + w_{k0}) \tag{2.12}$$

όπου $f(\cdot)$ κάποια συνάρτηση ενεργοποίησης. Το δεύτερο επίπεδο ή επίπεδο εξόδου, αποτελείται από $m = 1, 2, \ldots, M$ νευρώνες, όπου M το πλήθος των κλάσεων, που έχουν τη μορφή:

$$g_m(\boldsymbol{y}) = f(\boldsymbol{w}_m^T \boldsymbol{y} + w_{m0}) \tag{2.13}$$



Σχήμα 2.2: Νευρωνικό δίκτυο δύο επιπέδων [7].

Η συνάρτηση ενεργοποίησης στην περίπτωση του μοντέλου perceptron ήταν η βηματική. Για λόγους που θα προκύψουν από την διαδικασία ελαχιστοποίησης της συνάρτησης κόστους, στα νευρωνικά δίκτυα χρησιμοποιείται η λογιστική συνάρτηση:

$$\sigma(x) = \frac{1}{1 + \exp(-\alpha x)} \tag{2.14}$$

ή η συνάρτηση tanh:

$$f(x) = c\,\frac{1 - \exp(-\alpha x)}{1 + \exp(-\alpha x)} = c\ \tanh(\frac{\alpha x}{2}) \tag{2.15}$$

Η διαδικασία ταξινόμησης μπορεί εύκολα να περιγραφεί.

- Το κάθε διάνυσμα χαρακτηριστικών του συνόλου δεδομένων εισέρχεται στο επίπεδο εισόδου. Το πλήθος των νευρώνων του επιπέδου αυτού είναι ίσο με το πλήθος των χαρακτηριστικών, ενώ σε αυτό το στάδιο δεν εκτελείται καμία επεξεργασία.
- Η μετάβαση από το επίπεδο εισόδου στο κρυφό επίπεδο, γίνεται με μετασχηματισμούς που περιγράφονται από τη Σχέση 2.12.
- Η τελική ταξινόμηση πραγματοποιείται στο επίπεδο εξόδου και αφορά τον μετασχηματισμένο χώρο όπως προκύπτει από το κρυφό επίπεδο.

Ένα τέτοιο νευρωνικό δίκτυο, χαρακτηρίζεται ως *εμπρόσθιας διάδοσης* (feedforward neural network) λόγω της ροής της πληροφορίας. Η αρχιτεκτονική του δικτύου μπορεί να επεκταθεί με περισσότερα κρυφά επίπεδα, ακολουθώντας την ίδια μεθοδολογία.

Η εκπαίδευση του μοντέλου συνίσταται στον υπολογισμό των παραμέτρων όλων των επιπέδων, ενώ επιδιώκεται η έξοδος να έχει πιθανολογική ερμηνεία. Στην δυαδική ταξινόμηση, ορίζεται, για τον λόγο αυτό, μία και μοναδική έξοδος της μορφής:

$$y = \sigma(\alpha) \tag{2.16}$$

έτσι ώστε $0 \leq y(\boldsymbol{x}, \boldsymbol{w}) \leq 1$. Μπορεί λοιπόν κανείς να ερμηνεύσει το αποτέλεσμα ως την εκ των υστέρων πιθανότητα $y(\boldsymbol{x}, \boldsymbol{w}) = p(\omega_1|\boldsymbol{x})$ ενώ ισχύει ότι $p(\omega_2|\boldsymbol{x}) = 1 - p(\omega_1|\boldsymbol{x})$. Η συνάρτηση κόστους ορίζεται, όπως και στη λογιστική παλινδρόμηση, ως η αρνητική συνάρτηση λογαριθμικής πιθανοφάνειας:

$$L(\boldsymbol{w}) = -\sum_{n=1}^{N}[\omega_n \ln y_n + (1 - \omega_n)\ln(1 - y_n)] + r(\boldsymbol{w}) \tag{2.17}$$

Για την ελαχιστοποίηση της συνάρτησης κόστους, αξιοποιούνται και εδώ αριθμητικές μέθοδοι όπως η gradient descent, όπου γίνεται φανερό το ζήτημα υπολογισμού των μερικών παραγώγων ως προς τα βάρη των κρυφών επιπέδων. Ο αλγόριθμος που επιλύει τον προβληματισμό ονομάζεται αλγόριθμος οπισθοδιάδοσης (backpropagation). Γνωρίζουμε ότι κάθε νευρώνας υπολογίζει ένα άθροισμα των εισόδων του:

$$\alpha_j = \sum_i w_{ji} z_i \tag{2.18}$$

Με χρήση του κανόνα της αλυσίδας

$$\frac{\partial E_n}{\partial w_{ji}} = \frac{\partial E_n}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}} \tag{2.19}$$

οδηγούμαστε στην παρακάτω σχέση για τις μερικές παραγώγους:

$$\frac{\partial E_n}{\partial w_{ji}} = \delta_j z_i, \quad \delta_j = \frac{\partial E_n}{\partial a_j} \tag{2.20}$$

Η ποσότητα δ είναι γνωστή για το επίπεδο εξόδου και ίση με $\delta_m = y_m - \omega_m$ ενώ για το αμέσως προηγούμενο επίπεδο υπολογίζεται ως εξής:

$$\delta_j = h'(\alpha_j) \sum w_{mj} \delta_m \tag{2.21}$$

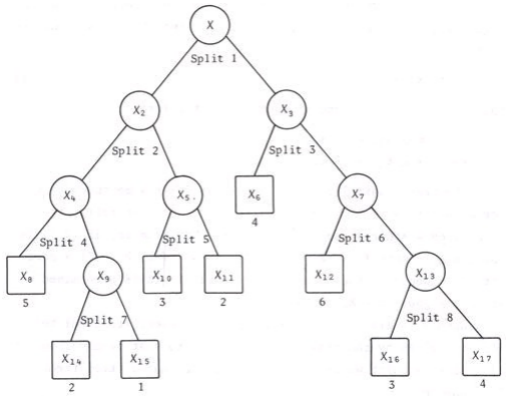
όπου $h(\cdot)$ η συνάρτηση ενεργοποίησης.

Προκύπτει συνεπώς, ότι οι μερικές παράγωγοι ως προς τις παραμέτρους των κρυφών επιπέδων υπολογίζονται ξεκινώντας από το επίπεδο εξόδου και κινούμενοι «προς τα πίσω».

Δέντρα Απόφασης

Τα δέντρα απόφασης αποτελούν μία από τις πιο διαδεδομένες μη γραμμικές μεθόδους ταξινόμησης. Πρόκειται για μια διαδικασία που διαδοχικά απορρίπτει κλάσεις, μέσω κατάλληλων «ερωτήσεων», έως ότου να καταλήξει σε μία μοναδική κλάση που αντιστοιχεί σε ένα τμήμα του αρχικού συνόλου δεδομένων. Ένα δυαδικό δέντρο απόφασης αποτελείται από:

- Τη ρίζα (root node) που σχετίζεται με το αρχικό, πλήρες σύνολο εκπαίδευσης.
- Τους κόμβους t (nodes) που σχετίζονται με υποσύνολα X_t του συνόλου εκπαίδευσης.
- Τα φύλλα ή τερματικούς κόμβους (leaves/ terminal nodes) στα οποία ανατίθεται μία κλάση.



Σχήμα 2.3: Δυαδικό δέντρο απόφασης [8]

Σύνολο ερωτήσεων

Βασικό συστατικό των δέντρων απόφασης είναι το σύνολο ερωτήσεων με βάση τις οποίες διαιρούνται οι κόμβοι. Οι ερωτήσεις αυτές αφορούν τα χαρακτηριστικά x_k του συνόλου δεδομένων και είναι της μορφής $x_k \leq a$;. Η απάντηση μπορεί να είναι «ναι» ή «όχι» και οδηγεί στη διαίρεση του κόμβου σε δύο κόμβους απογόνους X_{t_Y},X_{t_N} για τους οποίους ισχύουν οι σχέσεις:

$$X_{t_Y} \cap X_{t_N} = \emptyset$$
$$X_{t_Y} \cup X_{t_N} = X_t$$

Οι πιθανές ερωτήσεις φαίνεται να είναι άπειρες αν, παραδείγματος χάριν, τα χαρακτηριστικά παίρνουν πραγματικές τιμές $x_k \in \mathbb{R}$. Επειδή πρακτικά αυτό δεν γίνεται να υλοποιηθεί, οι ερωτήσεις περιορίζονται από το πεπερασμένο σύνολο δεδομένων. Το χαρακτηριστικό x_k μπορεί να πάρει το πολύ $N_t \leq N$ διαφορετικές τιμές σε ένα υποσύνολο X_t και πάνω σε αυτές ορίζονται οι τιμές της ποσότητας α.

Κριτήριο Διαίρεσης

Η επιλογή των κατάλληλων ερωτήσεων κατά την εκπαίδευση του μοντέλου, γίνεται με τη χρήση ενός κριτηρίου διαίρεσης. Πρόκειται για μια ποσότητα που μετρά την «καθαρότητα» (purity) των κόμβων.

- Αρχικά ορίζουμε τις αναλογίες $p(\omega_i|t), i = 1, 2, \ldots, M$ ως το ποσοστό των γεγονότων που ανήκουν στην κλάση ω_i στον κόμβο t, έτσι ώστε:

$$\sum_i p(\omega_i|t) = 1$$

- Στη συνέχεια μπορούμε να ορίσουμε την ποσότητα $i(t)$ ώστε να μας δείχνει την «μη καθαρότητα» (impurity) του κόμβου t συναρτήσει των παραπάνω αναλογιών έτσι ώστε:

$$\begin{aligned} i(t) = \max, \text{αν } p(\omega_i|t) = \frac{1}{M}, \forall i = 1, 2, \ldots, M \\ i(t) = 0, \text{αν } p(\omega_i|t) = 1 \text{ και } p(\omega_j|t) = 0, \forall\, j \neq i \end{aligned} \tag{2.22}$$

  Ή αλλιώς, η συνάρτηση $i(t)$ μεγιστοποιείται όταν γεγονότα από όλες τις κλάσεις βρίσκονται ισοπίθανα στον συγκεκριμένο κόμβο ενώ μηδενίζεται όταν ο κόμβος περιέχει στοιχεία μόνο από μία κλάση.

- Τέλος ορίζεται η μείωση της ποσότητας $i(t)$ κατά τη διαίρεση s του κόμβου από τη σχέση

$$\Delta_i(s, t) = i(t) - p_Y\, i(t_Y) - p_N\, i(t_N)$$

Η ερώτηση που τελικά καθορίζει τα δύο υποσύνολα - απογόνους είναι και αυτή που προκαλεί τη μεγιστοποίηση της ποσότητας Δ_i.

Ανάθεση κλάσης σε τερματικούς κόμβους

Κατά την ολοκλήρωση της διαδικασίας διαίρεσης, οι τελευταίοι κόμβοι ονομάζονται τερματικοί ή φύλλα και τους ανατίθεται μία κλάση. Ο πιο συνήθης κανόνας για την ανάθεση είναι ο κανόνας της πλειοψηφίας, ορίζεται δηλαδή η κλάση ω_j για την οποία ισχύει:

$$j = \arg\max_i p(\omega_i|t) \tag{2.23}$$

ή αλλιώς η κλάση που χαρακτηρίζει την πλειοψηφία των γεγονότων στο συγκεκριμένο φύλλο. Ορίζεται επιπλέον η ποσότητα $r(t)$ ως εκτιμητής λανθασμένης ταξινόμησης, ως δηλαδή η πιθανότητα ένα γεγονός στο φύλλο t να ταξινομηθεί λανθασμένα

$$r(t) = 1 - \max_i p(\omega_i|t) \tag{2.24}$$

ενώ τελικά, μπορεί να οριστεί ο εκτιμητής λανθασμένης ταξινόμησης ολόκληρου του δέντρου ως

$$R(T) = \sum_{t \in \tilde{T}} r(t)p(t) \tag{2.25}$$

όπου $\tilde{T}$ το σύνολο των τερματικών κόμβων. Ο ορισμός της ποσότητας ως εκτιμητή σφάλματος βασίζεται στο γεγονός ότι το σφάλμα αυτό αφορά μόνο το σύνολο εκπαίδευσης. Το πραγματικό σφάλμα ταξινόμησης προκύπτει από την εφαρμογή του μοντέλου σε κάποιο σύνολο ελέγχου και αναμένεται να είναι μεγαλύτερο.

Κριτήρια διακοπής και η μέθοδος «κλαδέματος»

Η διαδικασία ανάπτυξης του δέντρου απόφασης μπορεί να συνεχιστεί έως ότου το κάθε φύλλο να περιέχει ένα μοναδικό στοιχείο. Η καθαρότητα των τερματικών κόμβων σε αυτή την περίπτωση θα είναι μέγιστη, όμως, το μοντέλο βρίσκεται σε κατάσταση υπερεκπαίδευσης. Είναι πλέον εξειδικευμένο στο δεδομένο σύνολο εκπαίδευσης και αδυνατεί να γενικεύσει.

Το πιο απλό κριτήριο διακοπής έχει τη μορφή ενός κατωφλίου. Ορίζεται δηλαδή ένα κατώφλι δ, τέτοιο ώστε αν $\Delta_{i(t)\max} \leq \delta$ η διαδικασία σταματάει. Πρακτικά το συγκεκριμένο κριτήριο δεν έχει ικανοποιητικά αποτελέσματα. Αν το κατώφλι αυτό οριστεί πολύ μικρό, το μοντέλο καταλήγει πάλι σε κατάσταση υπερεκπαίδευσης. Αν όμως, το κατώφλι οριστεί πολύ μεγάλο, τότε μπορεί να χαθεί ουσιώδης πληροφορία. Μια διαίρεση μπορεί να οδηγεί σε μικρό Δ_i και να διακόψει τη διαδικασία, ενώ οι επόμενες διαιρέσεις των κόμβων απογόνων θα μπορούσαν να έχουν πολύ μεγαλύτερη αύξηση της καθαρότητας.

Μια εναλλακτική μέθοδος κατάλληλης διακοπής της διαδικασίας είναι εκείνη του κλαδέματος. Το δέντρο αφήνεται να μεγαλώσει αρκετά, συνήθως μέχρι τα φύλλα να έχουν ένα μόνο στοιχείο ή να είναι καθαρά και στη συνέχεια «κλαδεύονται» τμήματα έως ότου να φτάσουμε στο βέλτιστο μέγεθος. Ο ρυθμός λανθασμένης ταξινόμησης στο σύνολο εκπαίδευσης θα συνεχίσει να μειώνεται όσο το δέντρο μεγαλώνει. Αντιθέτως, η επίδοση του μοντέλου στο σύνολο ελέγχου θα βελτιώνεται μέχρι ένα σημείο και καθώς το δέντρο μεγαλώνει περαιτέρω η επίδοση θα ελαττώνεται αφού το μοντέλο θα υπερεκπαιδεύτεται. Έχοντας αυτή την πληροφορία, μπορεί κανείς να ορίσει ένα κατάλληλο μέγεθος δέντρου αφού πρώτα έχει μελετήσει τη συμπεριφορά του μέγιστου δέντρου.

2.2.3 Συνδυαστικές μέθοδοι

Μια συχνή προσέγγιση στο ζήτημα ταξινόμησης είναι αυτή του συνδυασμού μοντέλων. Αδύναμα μοντέλα (weak learners), με αποτελέσματα ελαφρώς καλύτερα από την τυχαία

ταξινόμηση, συνεισφέρουν με κατάλληλη στάθμιση στο τελικό αποτέλεσμα. Οι ταξινομητές δέντρων αποτελούν ιδιαίτερα καλή επιλογή, λόγω της ευκολίας και του μικρού υπολογιστικού κόστους που απαιτούν και συνεπώς η συνέχεια της ενότητας εστιάζει σε αυτούς.

AdaBoost

Ο αλγόριθμος AdaBoost (Adaptive Boost) είναι μια επαναληπτική μέθοδος και αξιοποιεί ρηχά δέντρα, συνήθως μιας μοναδικής διαίρεσης, εφαρμόζοντας κάθε φορά μια διαφορετική στάθμιση στα δείγματα του συνόλου εκπαίδευσης. Πιο συγκεκριμένα, σε κάθε βήμα της επαναληπτικής διαδικασίας, συνάπτονται βάρη $w_1, w_2, \ldots, w_N$ στα δείγματα του συνόλου. Χωρίς πρότερη γνώση των εύκολα ή δύσκολα ταξινομήσιμων δειγμάτων, τα βάρη ορίζονται αρχικά ως $w_i = 1/N$ όπου N το πλήθος των δειγμάτων.

Με κάθε επανάληψη, γίνεται προσαρμογή ενός δέντρου στο σταθμισμένο σύνολο και τα βάρη των λανθασμένα ταξινομημένων γεγονότων μεταβάλλονται, έτσι ώστε οι επόμενοι ταξινομητές να εστιάσουν περισσότερο σε αυτά. Η μεταβολή των βαρών εξαρτάται από την απόδοση του ταξινομητή, όπως αυτή προκύπτει από την εκθετική συνάρτηση κόστους. Το αποτέλεσμα απορρέει από τον συνδυασμό των ρηχών δέντρων, στον οποίο συμμετέχουν με μεγαλύτερο βάρος εκείνα που είχαν τα καλύτερα αποτελέσματα.

ΨΕΥΔΟΚΩΔΙΚΑΣ AdaBoost

- Αρχικοποίησε: $w_i = 1/N, i = 1, 2, \ldots, N$
- Για $m = 1$ έως M:
  - Εκπαίδευσε έναν ταξινομητή $G_m(x)$ στο σύνολο εκπαίδευσης με βάρη w_i.
  - Υπολόγισε:
$$err_m = \frac{\sum_{i=1}^{N} w_i I(\omega_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}$$
  - Υπολόγισε:
$$\alpha_m = \log(\frac{1 - err_m}{err_m})$$
  - Υπολόγισε
$$w_i^{(m+1)} = w_i^{(m)} \exp[\alpha_m \cdot I(\omega_i \neq G_m(x_i))], \quad i = 1, 2, \ldots, N$$
- Υπολόγισε
$$G(x) = sign[\sum_{m=1}^{M} \alpha_m G_m(x)]$$

Η συνάρτηση $I(\omega_i \neq G_m(x_i))$ επιστρέφει 1 στην περίπτωση της λανθασμένης ταξινόμησης, αλλιώς επιστρέφει 0.

Gradient Boost

Ο αλγόριθμος gradient boost επιδιώκει να ελαχιστοποιήσει τη συνάρτηση κόστους $L(\omega_i, G(\boldsymbol{x}_i))$ μοντελοποιώντας τη $G(\cdot)$ ως ένα άθροισμα της μορφής:

$$G(\boldsymbol{x}) = \sum_{m=1}^{M} \phi(\boldsymbol{x}) \tag{2.26}$$

όπου οι $\phi(\boldsymbol{x})$ είναι weak learners. Επειδή ο απευθείας υπολογισμός είναι δύσκολος, ακολουθείται μία διαδικασία παρόμοια του gradient descent. Θεωρείται δηλαδή, ότι η ελαχιστοποίηση μπορεί να επιτευχθεί με βήματα στη διεύθυνση της παραγώγου:

$$G_m = G_{m-1} - \nu \left[\frac{\partial L(\omega_i, G(x_i))}{\partial G(x_i)} \right]_{G(x)=G_{m-1}(x)} \tag{2.27}$$

Τα $\phi(\boldsymbol{x})$, συνεπώς, πρέπει να προσαρμοστούν σε αυτή την παράγωγο, η οποία ονομάζεται και υπόλειμμα (residual). Στην περίπτωση ενδυναμωμένων δέντρων, το αδύναμο μοντέλο είναι δέντρο παλινδρόμησης με μικρό αριθμό φύλλων και επιδιώκει βελτιστοποίηση ως προς τα residuals, στα υποσύνολα που αντιστοιχούν σε κάθε φύλλο.

Χρησιμοποιώντας την αρνητική συνάρτηση λογαριθμικής πιθανοφάνειας ως συνάρτηση κόστους, μπορούμε εύκολα να δώσουμε πιθανολογική ερμηνεία στα αποτελέσματα. Η προβλεπόμενη πιθανότητα προκύπτει από το υπόλειμμα σε σχέση με τη δοθείσα πιθανότητα (1 ή 0), όπως αυτό υπολογίζεται από το μοντέλο. Ακολουθεί η υλοποίηση της πιο γενικής μορφής του αλγορίθμου για οποιαδήποτε παραγωγίσιμη συνάρτηση κόστους.

ΨΕΥΔΟΚΩΔΙΚΑΣ gradient boost

- Αρχικοποίησε:

$$G_0(x) = \arg\min_{\gamma} \sum_{i=1}^{N} L(\omega_i, \gamma) \tag{2.28}$$

- Για $m = 1$ έως M:
  - Υπολόγισε:

$$r_{i,m} = - \left[\frac{\partial L(\omega_i, G(x_i))}{\partial G(x_i)} \right]_{G(x)=G_{m-1}(x)} \tag{2.29}$$

  - Προσάρμοσε έναν αδύναμο ταξινομητή δέντρου στα $r_{im},\ i = 1, 2, \ldots, N$ με φύλλα $R_{jm},\ j = 1, 2, \ldots, J_m$
  - Υπολόγισε:

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{x_i \in R_{ij}} L(\omega_i, G_{m-1}(x_i) + \gamma) \tag{2.30}$$

– Ανανέωσε:

$$G_m(x) = G_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_m I(x \in R_{jm}) \tag{2.31}$$

Ταξινομητής τυχαίου δάσους (Random Forest Classifier)

Το όνομα του ταξινομητή δάσους προκύπτει από το γεγονός ότι δεν είναι τίποτα άλλο, παρά ένα σύνολο δέντρων. Σε αντίθεση με τις μεθόδους ενίσχυσης AdaBoost, Gradient Boost όμως, δεν αποτελείται εν γένει από αδύναμους ταξινομητές. Αναπτύχθηκε[8][9]με στόχο τη συνεχή βελτίωση της ικανότητας γενίκευσης χωρίς αυτό να επιτυγχάνεται με κόστος στις επιδόσεις στο σύνολο εκπαίδευσης.

Η αρχική προσέγγιση στο ζήτημα ήταν αυτή του bootstrap. Πρόκειται για τη διαδικασία κατά την οποία το μοντέλο εκπαιδεύεται σε διαφορετικά τμήματα του συνόλου δεδομένων, ενώ το τελικό αποτέλεσμα προκύπτει ως μέσος όρος ή πλειοψηφικά. Θεωρώντας B τυχαίες μεταβλητές, η κάθε μία με διακύμανση σ^2 και συσχέτιση ανά ζεύγη ρ, ο μέσος όρος θα έχει διακύμανση [10]:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{2.32}$$

Για αρκετά μεγάλο πλήθος δέντρων ο δεύτερος όρος «χάνεται».

Η πρώτη ιδιότητα της συγκεκριμένης συνδυαστικής μεθόδου είναι ότι η διακύμανση μειώνεται με αύξηση του πλήθους των δέντρων. Το δεύτερο χαρακτηριστικό είναι πως η πραγματική βελτιστοποίηση επιτυγχάνεται με την κατασκευή ασυσχέτιστων εκτιμητών για ελάττωση της ποσότητας ρ. Για τον λόγο αυτό οι πιο σύγχρονοι αλγόριθμοι τυχαίου δάσους εκτελούν bootstrapping όχι μόνο στο σύνολο δεδομένων αλλά και στο σύνολο των χαρακτηριστικών. Το κάθε δέντρο μπορεί να επεκταθεί στο μέγιστο δυνατό μέγεθός του, επιτυγχάνοντας πάντα άριστη ταξινόμηση στο σύνολο εκπαίδευσης, ενώ η αξιοποίηση διαφορετικών υποσυνόλων χαρακτηριστικών οδηγεί σε ασυσχέτιστα δέντρα.

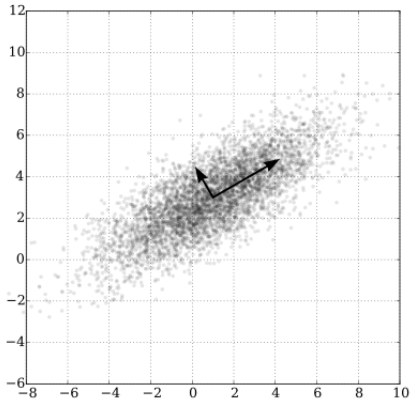
ΨΕΥΔΟΚΩΔΙΚΑΣ random forest

- Για $b = 1$ έως B
  - Επέλεξε ένα υποσύνολο $\boldsymbol{Z}^*$ μεγέθους N από το σύνολο εκπαίδευσης
  - Υλοποίησε ένα δέντρο T_b με τερματικούς κόμβους $n_{\min}$ ως εξής:
    - Επέλεξε m τυχαία χαρακτηριστικά.
    - Επέλεξε και εκτέλεσε την καλύτερη διαίρεση με βάση τα m.
- Επέστρεψε το σύνολο των δέντρων $\{T_b\}_1^B$

Για κάθε νέο παράδειγμα η κλάση επιλέγεται πλειοψηφικά από τα δέντρα $\{T_b\}_1^B$

2.3 Ανάλυση Κύριων Συνιστωσών

Πολλές φορές, ο χώρος των χαρακτηριστικών στα προβλήματα ταξινόμησης είναι μεγάλης διάστασης και απαιτεί υπέρογκους υπολογιστικούς πόρους. Επιπλέον καθίσταται δύσκολο να απεικονιστεί. Τα δύο αυτά ζητήματα μπορούν να επιλυθούν με την ανάλυση κυρίων συνιστωσών (principal component analysis/PCA)[11]. Πρόκειται για μια διαδικασία προβολής του αρχικού χώρου, διάστασης D, σε έναν νέο με διάσταση $M \leq D$. Το κριτήριο που χρησιμοποιείται για την επιλογή της βάσης του νέου χώρου είναι αυτό της μέγιστης διασποράς.



Σχήμα 2.4: Ανάλυση PCA μίας Gaussian κατανομής. Τα διανύσματα είναι τα ιδιοδιανύσματα του πίνακα συνδιασποράς κανονικοποιημένα ως προς την τετραγωνική ρίζα της αντίστοιχης ιδιοτιμής[12].

Επιλέγοντας την προβολή σε έναν μονοδιάστατο χώρο, που περιγράφεται από το διάνυσμα $\boldsymbol{u}$, η διασπορά των προβαλλόμενων δειγμάτων δίνεται από τον τύπο:

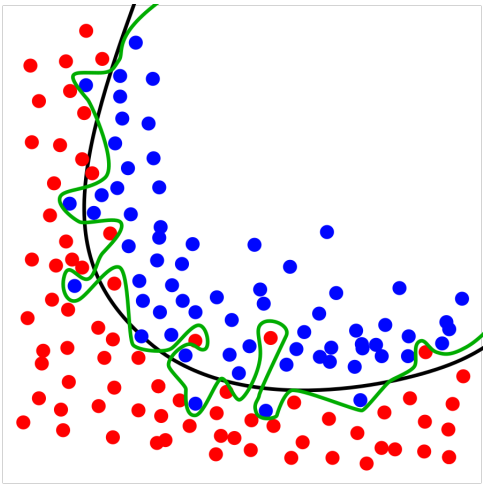
$$\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{u}^T\boldsymbol{x}_n - \boldsymbol{u}^T\bar{\boldsymbol{x}})^2 = \boldsymbol{u}^T\boldsymbol{S}\boldsymbol{u} \tag{2.33}$$

όπου $\boldsymbol{S}$ ο πίνακας συνδιασποράς. Η μεγιστοποίηση της διασποράς επιτυγχάνεται όταν το $\boldsymbol{u}$ είναι ιδιοδιάνυσμα του πίνακα $\boldsymbol{S}$ και συγκεκριμένα εκείνο που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή. Για την προβολή σε περισσότερες από μία διαστάσεις αρκεί να βρεθεί κάθε φορά η αμέσως επόμενη διεύθυνση μέγιστης διασποράς, ορθοκανονική στα προγηούμενα διανύσματα.

Στις περισσότερες πρακτικές εφαρμογές, το ιδανικό μέγεθος διάστασης δεν είναι γνωστό εκ των προτέρων. Συνεπώς, η ανάλυση PCA εφαρμόζεται για προβολή σε έναν χώρο ίσης διάστασης με την αρχική. Στη συνέχεια επιλέγονται οι διαστάσεις ανάλογα με το ποσοστό απώλειας πληροφορίας και τους διαθέσιμους υπολογιστικούς πόρους.

2.4 Το φαινόμενο της υπερεκπαίδευσης

Κατά την ανάλυση των ταξινομητών και ιδιαίτερα των μη γραμμικών, έγινε αναφορά στο πρόβλημα της υπερεκπαίδευσης. Γνωστό και ως overfitting ή overtraining, πρόκειται για το στάδιο στη διαδικασία εκπαίδευσης, κατά το οποίο, το μοντέλο σταματάει να εντοπίζει γενικές συμπεριφορές ή «τάσεις» στα δεδομένα και εξειδικεύεται πλήρως στο σύνολο εκπαίδευσης αδυνατώντας να γενικεύσει.



Σχήμα 2.5: Κατάλληλο όριο διαχωρισμού (μαύρο) και από υπερεκπαιδευμένο ταξινομητή (πράσινο)[13].

Στην περίπτωση των γραμμικών ταξινομητών, η συμπεριφορά παρατηρείται κυρίως όταν το πλήθος των χαρακτηριστικών είναι πολύ μεγάλο συγκριτικά με το πλήθος των δειγμάτων. Οι μη γραμμικοί ταξινομητές είναι πιο ευαίσθητοι, καθώς έχουν τη δυνατότητα να επιτύχουν τέλεια ταξινόμηση «αποστηθίζοντας» τα δεδομένα.

Συνήθως υιοθετούνται δύο είδη προσεγγίσεων για την αποφυγή του φαινομένου.

- Εάν το σύνολο δεδομένων είναι αρκετά μεγάλο, διαιρείται τυχαία σε σύνολο εκπαίδευσης και σύνολο ελέγχου. Τα μοντέλα εκπαιδεύονται στο πρώτο ενώ η επίδοσή τους ελέγχεται στο δεύτερο. Τα διάφορα εργαλεία που ποσοτικοποιούν την απόδοση του ταξινομητή, δύναται να εντοπίσουν το στάδιο της διαδικασίας, κατά το οποίο, κάθε περαιτέρω βελτίωση στο σύνολο εκπαίδευσης συνεπάγεται επιδείνωση στο σύνολο ελέγχου.

- Διαφορετικά αξιοποιείται η μέθοδος cross-validation. Το σύνολο δεδομένων χωρίζεται σε k υποσύνολα και η διαδικασία εκπαίδευσης επαναλαμβάνεται, αξιοποιώντας κάθε φορά ένα διαφορετικό υποσύνολο για έλεγχο και τα υπόλοιπα για εκπαίδευση. Πρόκειται για μια πιο υπολογιστικά απαιτητική προσέγγιση και για το λόγο αυτό επιλέγεται όταν το πλήθος των δεδομένων είναι μικρό.

Πρακτικό μέρος

3.1 Εισαγωγικά

Το πρακτικό μέρος της εργασίας αποσκοπεί αρχικά στην εφαρμογή μοντέλων μηχανικής μάθησης με επίβλεψη, για ταξινόμηση πραγματικών δεδομένων από τον επιχειρηματικό χώρο. Στη συνέχεια επιδιώκεται να εξαχθούν, από αυτά, ερμηνεύσιμα συμπεράσματα. Το σύνολο δεδομένων που χρησιμοποιήθηκε προέρχεται από την έρευνα των S.Moro, P.Cortez και P.Rita[14] με στόχο την πρόβλεψη της επιτυχίας τηλεφωνικών προωθητικών ενεργειών από τράπεζες. Αξιοποιήθηκαν δεδομένα 52,944 κλήσεων μεταξύ μίας Πορτογαλικής τράπεζας και των πελατών της, που αφορούσαν την προώθηση προθεσμιακών καταθέσεων από το 2008 έως το 2013. Οι διαθέσιμες πληροφορίες περιλάμβαναν χαρακτηριστικά των πελατών (ηλικία, επάγγελμα κλπ.) καθώς και χαρακτηριστικά της επικοινωνίας με αυτούς (πλήθος κλήσεων, εμπειρία υπαλλήλου κλπ.). Σε αυτά προστέθηκαν δείκτες της οικονομικής περιόδου που έλαβε χώρα η επικοινωνία (EURIBOR 3 μηνών, μέσα ποσοστά ανεργείας κλπ.).

Το συνολικό πλήθος των χαρακτηριστικών της αρχικής έρευνας ήταν 150, από τα οποία επιλέχθηκαν τελικά τα 22 πιο σημαντικά. Το διαθέσιμο σύνολο δεδομένων που μελετάται στη παρούσα εργασία, αποτελεί ένα τμήμα του αρχικού, με 41,188 δείγματα πελατών και 20 χαρακτηριστικά. Για κάθε ένα από αυτά τα δείγματα, είναι επιπλέον γνωστό το αποτέλεσμα της προωθητικής ενέργειας. Η πλήρης λίστα των διαθέσιμων χαρακτηριστικών φαίνεται στον Πίνακα 3.1

Στις επόμενες ενότητες γίνεται μια αρχική παρουσίαση και ανάλυση των δεδομένων καθώς και η κατάλληλη προετοιμασία του συνόλου για να είναι συμβατό με τις μεθόδους μηχανικής μάθησης. Στη συνέχεια υλοποιούνται και αξιολογούνται ταξινομητές λογιστικής παλινδρόμησης, ενισχυμένων δέντρων απόφασης και νευρωνικού δικτύου. Τέλος επιδιώκεται να εξαχθούν από αυτούς τα σημαντικότερα χαρακτηριστικά που οδήγησαν στην επιτυχή ταξινόμηση, με στόχο να «μεταφραστούν» σε αξιοποιήσιμη, για πρόβλεψη, πληροφορία.

Όνομα	Περιγραφή	Τύπος	
age	Ηλικία (αριθμητικές τιμές)	Αριθμητική	
job	Επάγγελμα	Κατηγορική	admin, blue- collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed, unknown
marital	Οικογενειακή κατάσταση	Κατηγορική	divorced, married, single, unknown
education	Εκπαίδευση	Κατηγορική	basic.4y, basic.6y, basic.9y, high.school, illiterate, professional course, universitydegree, unknown
default	Μη εξυπηρετούμενο δάνειο	Κατηγορική	no, yes, unknown
housing	Στεγαστικό δάνειο	Κατηγορική	no, yes, unknown
loan	Καταναλωτικό δάνειο	Κατηγορική	no, yes, unknown
contact	Τρόπος πιο πρόσφατης επικοινωνίας	Κατηγορική	cellular, telephone
month	Μήνας πιο πρόσφατης επικοινωνίας	Κατηγορική	jan, feb, mar, ..., nov, dec
day_of_week	Ημέρα πιο πρόσφατης επικοινωνίας	Κατηγορική	mon, tue, wed, thu, fri
duration	Διάρκεια κλήσης	Αριθμητική	
capaign	Συνολικές κλήσεις κατά τη διάρκεια αυτής της καμπάνιας με αυτόν τον πελάτη	Αριθμητική	
pdays	Ημέρες από την τελευταία επικοινωνία που αφορούσε παλαιότερη καμπάνια	Αριθμητική	999 αν δεν έχει υπάρχει προηγούμενη επικοινωνία
previous	Συνολικές κλήσεις πριν την παρούσα καμπάνια με αυτόν τον πελάτη	Αριθμητική	
poutcome	Αποτέλεσμα προηγούμενης καμπάνιας	Κατηγορική	failure, nonexistent, success
emp.var.rate	Τρίμηνη μεταβολή δείκτη απασχόλησης	Αριθμητική	
cons.price.idx	Δείκτης τιμών καταναλωτή	Αριθμητική	
cons.conf.idx	Δείκτης εμπιστοσύνης καταναλωτή	Αριθμητική	
euribor3m	EURIBOR τριών μηνών	Αριθμητική	
nr.employed	Αριθμός απασχολούμενων	Αριθμητική	
y	Αποτέλεσμα	Κατηγορική	no, yes

Πίνακας 3.1: Πίνακας χαρακτηριστικών

3.2 Εργαλεία, περιβάλλον, βιβλιοθήκες

Η μελέτη έγινε με τη χρήση της γλώσσας προγραμματισμού Python[15] σε περιβάλλον Jupyter[16]. Ακολουθούν αναφορές των πακέτων που χρησιμοποιήθηκαν καθώς και σύντομες περιγραφές τους

NumPy
Προσφέρει υποστήριξη πολυδιάστατων πινάκων καθώς και ένα μεγάλο πλήθος μαθηματικών λειτουργιών. Αξιοποιήθηκε κυρίαρχα κατά την προετοιμασία και αρχική ανάλυση του συνόλου δεδομένων[17].

pandas
Περιέχει μεθόδους χειραγώγησης και ανάλυσης δεδομένων. Συγκεκριμένα, το αντικείμενο pandas.DataFrame είχε ουσιαστικό ρόλο, καθώς αποτέλεσε το κύριο εργαλείο διαχείρισης του συνόλου δεδομένων[18].

scikit-learn
Περιέχει όλους τους αλγορίθμους που χρησιμοποιήθηκαν κατά την υλοποίηση των μοντέλων ταξινόμησης καθώς και κατά την ανάλυση κυρίων συνιστωσών[19].

imblearn
Περιλαμβάνει μεθόδους και εργαλεία για την αντιμετώπιση άνισων κλάσεων κατά τη διαδικασία ταξινόμησης. Αξιοποιήθηκε για την εξισορρόπησή τους[20].

matplotlib
Πρόκειται για την βιβλιοθήκη γραφημάτων που χρησιμοποιήθηκε για το υπόλοιπο της εργασίας[21].

Εργαλεία Ελέγχου

Για την αξιολόγηση και σύγκριση των μοντέλων χρησιμοποιήθηκαν διάφορα ποσοτικά και γραφικά εργαλεία:

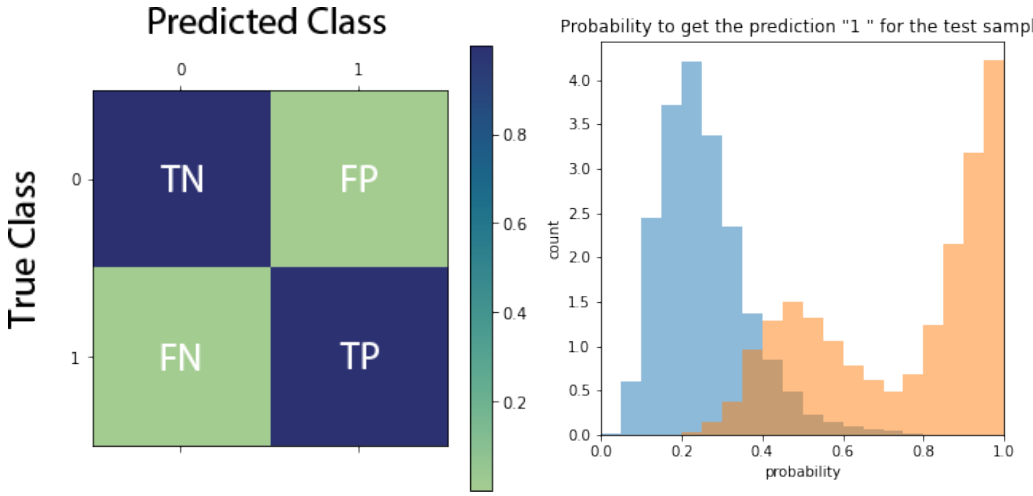
Καμπύλη ROC
Πρόκειται για τη γραφική παράσταση των ρυθμών σωστής και λανθασμένης ταξινόμησης που ορίζονται ως:

$$TruePositiveRate(TPR) = \frac{TruePositive(TP)}{TP + FalseNegative(FN)}$$
$$FalsePositiveRate(FPR) = \frac{FalsePositive(FP)}{FP + TrueNegative(TN)}$$

Στο γράφημα παρουσιάζεται συνήθως και η ευθεία $x = y$ ως το αποτέλεσμα μίας τελείως τυχαίας (50-50) ταξινόμησης. Το εμβαδόν κάτω από την καμπύλη ROC είναι ένα καλό μέτρο της επίδοσης του μοντέλου.

Πίνακας Σύγχυσης (Confusion Matrix)

Στην περίπτωση της δυαδικής ταξινόμησης, είναι ένας 2×2 πίνακας που αναπαριστά τα αποτελέσματα ως TP, TN, FP, FN όπως παρακάτω:



«α΄» *«β΄»*

Σχήμα 3.1: α) Μορφή του πίνακα σύγχυσης. β) Μορφή γραφήματος εκ των υστέρων πιθανοτήτων

Κατανομή των εκ των υστέρων πιθανοτήτων

Στη δυαδική ταξινόμηση η έξοδος των μοντέλων είναι ορισμένες φορές η εκ των υστέρων πιθανότητα της «επιθυμητής» κλάσης $p(\omega_1|\boldsymbol{x})$. Αναπαριστώντας γραφικά αυτές τις κατανομές, για κάθε κλάση ξεχωριστά, περιμένουμε να δούμε έναν διαχωρισμό. Για ένα αποδοτικό μοντέλο τα στοιχεία της επιθυμητής κλάσης τείνουν στη μονάδα ενώ τα στοιχεία της άλλης τείνουν στο 0. Ένα παράδειγμα τέτοιου γραφήματος φαίνεται στο Σχήμα 3.1 β΄.

3.2.1 Ποσοτικά Εργαλεία

Για την αξιολόγηση των αποτελεσμάτων αξιοποιούνται επιπλέον οι παρακάτω ποσότητες

- $Accuracy = \frac{TN+TP}{TN+FP+FN+TP}$
- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $F1score = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$

3.3 Προετοιμασία του συνόλου δεδομένων

Η κατάλληλη επεξεργασία του συνόλου δεδομένων είναι εξίσου σημαντική με την επιλογή και την ρύθμιση των μοντέλων ταξινόμησης, ενώ επιπλέον απαιτεί μια βασική κατανόηση της πληροφορίας που είναι διαθέσιμη. Το πρώτο βήμα είναι να αφαιρεθεί από το σύνολο, το χαρακτηριστικό της διάρκειας κλήσης. Όπως αναφέρεται στην ιστοσελίδα στην οποία είναι διαθέσιμο [22], η διάρκεια κλήσης έχει πολύ μεγάλη συσχέτιση με το τελικό αποτέλεσμα της προωθητικής ενέργειας, ενώ παράλληλα δεν είναι γνωστή εκ των προτέρων. Συνεπώς, δεν μπορεί να συμπεριληφθεί για την κατασκευή ενός ρεαλιστικού μοντέλου πρόβλεψης.

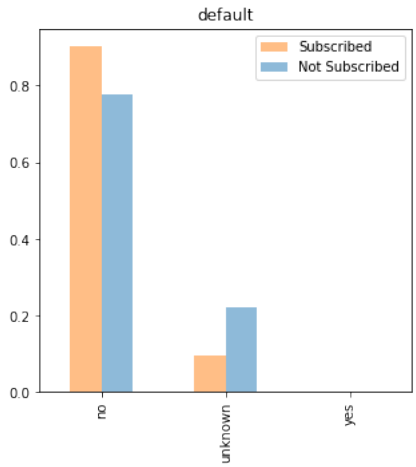
Στη συνέχεια πρέπει να αντιμετωπιστεί το ζήτημα των αγνώστων χαρακτηριστικών. Πιο συγκεκριμένα, στα χαρακτηριστικά εκείνα που αφορούν τους πελάτες, υπάρχει πλήθος αγνώστων τιμών που στο σύνολο δεδομένων είναι καταγεγραμμένα ως "unknown". Τα κελιά αυτά δεν είναι κενά, συνεπώς δεν τίθεται ζήτημα συμβατότητας. Παρά ταύτα μπορεί να οδηγήσουν το μοντέλο σε λανθασμένα συμπεράσματα, καθώς θα τα εκλάβει ως κοινά. Π.χ. Οι πελάτες με άγνωστη οικογενειακή κατάσταση θα έχουν την ίδια οικογενειακή κατάσταση, κάτι το οποίο δεν ισχύει κατά κανόνα.

Feature	**No. unknown**
job	330
marital	80
education	1731
default	8597
housing	990
loan	990
clients	10700

Πίνακας 3.2: Πλήθος αγνώστων δεδομένων ανά κατηγορία

Η μελέτη της συχνότητας των αγνώστων γίνεται σε δύο άξονες. Πρώτα υπολογίζεται η συχνότητά τους ανά χαρακτηριστικό και στη συνέχεια ανά πελάτη (Πίνακας 3.2). Φαίνεται ότι το κυρίαρχα προβληματικό χαρακτηριστικό είναι αυτό του μη εξυπηρετούμενου δανείου (default/defaulted loan). Εστιάζοντας παραπάνω, εμφανίζεται μια ακόμα ιδιαιτερότητα, ότι οι πελάτες που έχουν θετικό το συγκεκριμένο χαρακτηριστικό είναι μόνο 3 σε όλο το δείγμα.

Αξίζει σε αυτό το σημείο να αναφερθεί ότι εμφανίζεται σημαντική διαχωρισημότητα στο δεδομένο χαρακτηριστικό, γεγονός που υποδεικνύει πως η απάντηση unknown δύναται να εμπεριέχει πληροφορία. Το υποσύνολο θα μπορούσε να είναι προκατειλημμένο λόγω της ευαισθησίας του ζητήματος ή σε μία διαφορετική προσέγγιση, μπορεί να εξαχθούν συμπεράσματα που αφορούν την επιφυλακτικότητα των πελατών να απαντήσουν



Σχήμα 3.2: Κατανομή του χαρακτηριστικού default

σε ερωτήσεις και πως αυτή σχετίζεται με την επιτυχία της επικοινωνίας. Τα προαναφερθέντα ερωτήματα βέβαια, ξεφεύγουν από το πλαίσιο της παρούσας εργασίας, ενώ επιπλέον, οι προβλεπτικές ικανότητες του μοντέλου θα ήταν βασισμένες σε υποθέσεις που η συγγραφέας δεν έχει το υπόβαθρο να υποστηρίξει. Συνεπώς θεωρείται πως όλες οι unknown τιμές δεν μπορούν να μας δώσουν με ασφάλεια ουσιαστική πληροφορία και το χαρακτηριστικό αφαιρείται. Καθώς τα εναπομείναντα δείγματα με άγνωστες τιμές είναι πλέον λίγα σε σχέση με το μέγεθος του συνόλου, προτιμάται να αφαιρεθούν αυτά έναντυ επιπλέον χαρακτηριστικών. Το τελικό σύνολο έχει μορφή όπως φαίνεται στον πίνακα 3.3.

Class	No. Instances
no	33987
yes	4258

Πίνακας 3.3: Μέγεθος συνόλου με αφαίρεση των αγνώστων

Εξισορρόπηση κλάσεων, διαίρεση συνόλου σε σύνολα εκπαίδευσης και ελέγχου

Από την πρώτη ανάλυση που έγινε στα δεδομένα, είναι εμφανές ότι οι κλάσεις είναι άνισες. Λαμβάνοντας υπόψιν ότι τα περισσότερα μοντέλα ελαχιστοποιούν μια συνάρτηση κόστους, που βασίζεται στις λανθασμένες ταξινομήσεις, γίνεται προφανές ότι ένα μοντέλο μπορεί να μάθει να επιλέγει πάντα την κυρίαρχη κλάση και να έχει πολύ καλά αποτελέσματα χωρίς να έχει αναγνωρίσει κάποιο πρότυπο. Για να αποφευχθεί αυτό, οι κλάσεις πρέπει να εξισορροπηθούν. Το διαθέσιμο σύνολο δεδομένων έχει αρκετά μεγάλο πλήθος, ώστε η εξισορρόπηση να μπορεί να επιτευχθεί αφαιρώντας δεδομένα από την κυρίαρχη κλάση. Η διαδικασία πραγματοποιείται με τυχαία αφαίρεση δειγμάτων καθώς έτσι διατηρούνται τα χαρακτηριστικά της αρχικής κατανομής.

Στη συνέχεια το σύνολο χωρίζεται σε ένα σύνολο εκπαίδευσης (80% του πλήρους συνόλου) και ένα σύνολο ελέγχου (20% του πλήρους συνόλου). Στο πρώτο θα εκπαιδευτούν τα διάφορα μοντέλα, ενώ το δεύτερο είναι απαραίτητο για έλεγχο των αποτελεσμάτων και αποφυγή υπερεκπαίδευσης. Η βελτιστοποίηση των μοντέλων βασίζεται στην επίδοση αυτών, στο σύνολο ελέγχου.

Κωδικοποίηση κατηγορικών χαρακτηριστικών

Πολλά από τα χαρακτηριστικά, καθώς και η στήλη που περιέχει την κλάση, δεν παίρνουν αριθμητικές τιμές αλλά απαντήσεις πολλαπλής επιλογής τύπου `string`. Όταν αυτές είναι δυαδικές, όπως οι απαντήσεις «ναι» και «όχι», τότε μπορούν εύκολα να μετατραπούν σε 1 και 0 αντίστοιχα. Όταν οι πιθανές απαντήσεις είναι περισσότερες, αυτές μπορούν να ανήκουν σε μία από δύο κατηγορίες:

- Διατάξιμες μεταβλητές. Είναι αυτές που μπορούν να ιεραρχηθούν και συνεπώς ικανοποιούν σχέσεις της μορφής $a < b < c$. Η κωδικοποίησή τους μπορεί να γίνει με απλή μετατροπή σε ακεραίους στην σωστή, φυσικά, διάταξη.
- Μη Διατάξιμες μεταβλητές. Είναι εκείνες που δεν ικανοποιούν σχέσεις ιεραρχίας και κωδικοποιούνται με τη μέθοδο One-Hot Encoding. Πρακτικά, οι διάφορες κατηγορίες μετατρέπονται σε χαρακτηριστικά με δυαδικές απαντήσεις (0,1).

Ακολουθεί ο πίνακας με τα χαρακτηριστικά ανά είδος και τον τρόπο με τον οποίο κωδικοποιήθηκαν.

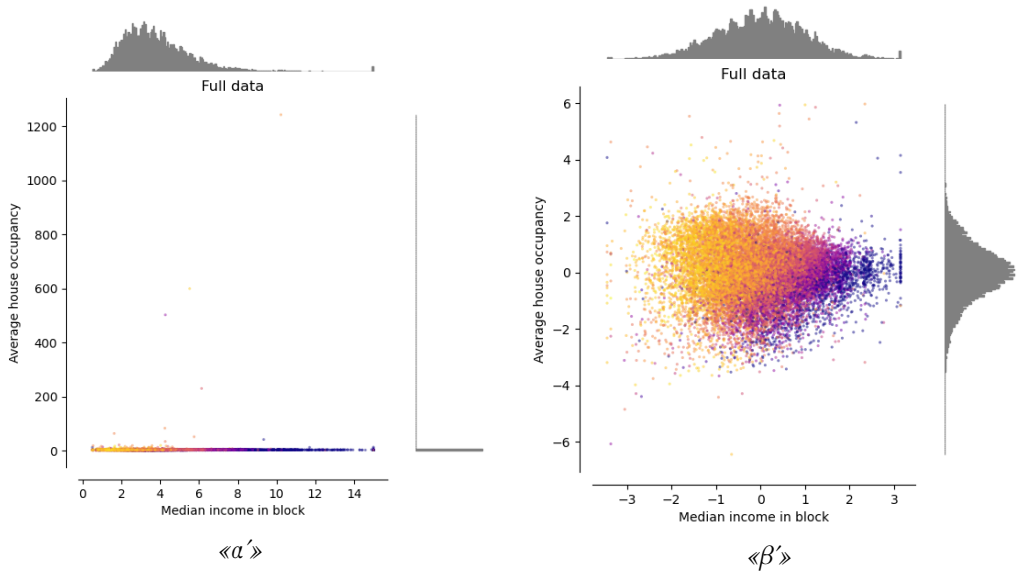
Features	**Type**	**Encoding**
age, campaign, pdays, previous, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed	Numeric	No Encoding
education, housing, loan, contact, outcome(y)	Ordinal/Binary	Ordinal/Binary
job, marital, month, day of week	Not Ordinal	One-Hot

Πίνακας 3.4: Κωδικοποίηση Χαρακτηριστικών

Κανονικοποίηση Δεδομένων

Ένα πολύ σημαντικό στάδιο της προεπεξεργασίας των δεδομένων είναι αυτό της κανονικοποίησης. Καθώς το κάθε χαρακτηριστικό περιγράφει διαφορετικές ποσότητες, η τάξη μεγέθους μεταξύ τους μπορεί να διαφέρει δραστικά. Μοντέλα τα οποία υπολογίζουν αποστάσεις στον χώρο των χαρακτηριστικών ή χρησιμοποιούν όρους κανονικοποίησης στις συναρτήσεις κόστους, επηρεάζονται άμεσα από τέτοιες δυσανάλογες ποσότητες. Για το λόγο αυτό τα δεδομένα προσαρμόζονται στην ίδια τάξη μεγέθους διατηρώντας τις «εσωτερικές» τους σχέσεις.

Για την κανονικοποίηση έγινε χρήση της μεθόδου Yeo-Johnson PowerTransformer[23] του πακέτου sklearn. Πρόκειται για έναν μετασχηματισμό που πέρα από την απλή κανονικοποίηση, επιδιώκει να δώσει μια Gaussian μορφή στα δεδομένα.



Σχήμα 3.3: Yeo-Johnson power transformation. α) Πριν το μετασχηματισμό. β) Μετά.[24]

3.4 Κατανομές των χαρακτηριστικών του συνόλου εκπαίδευσης και πίνακες συσχέτισης

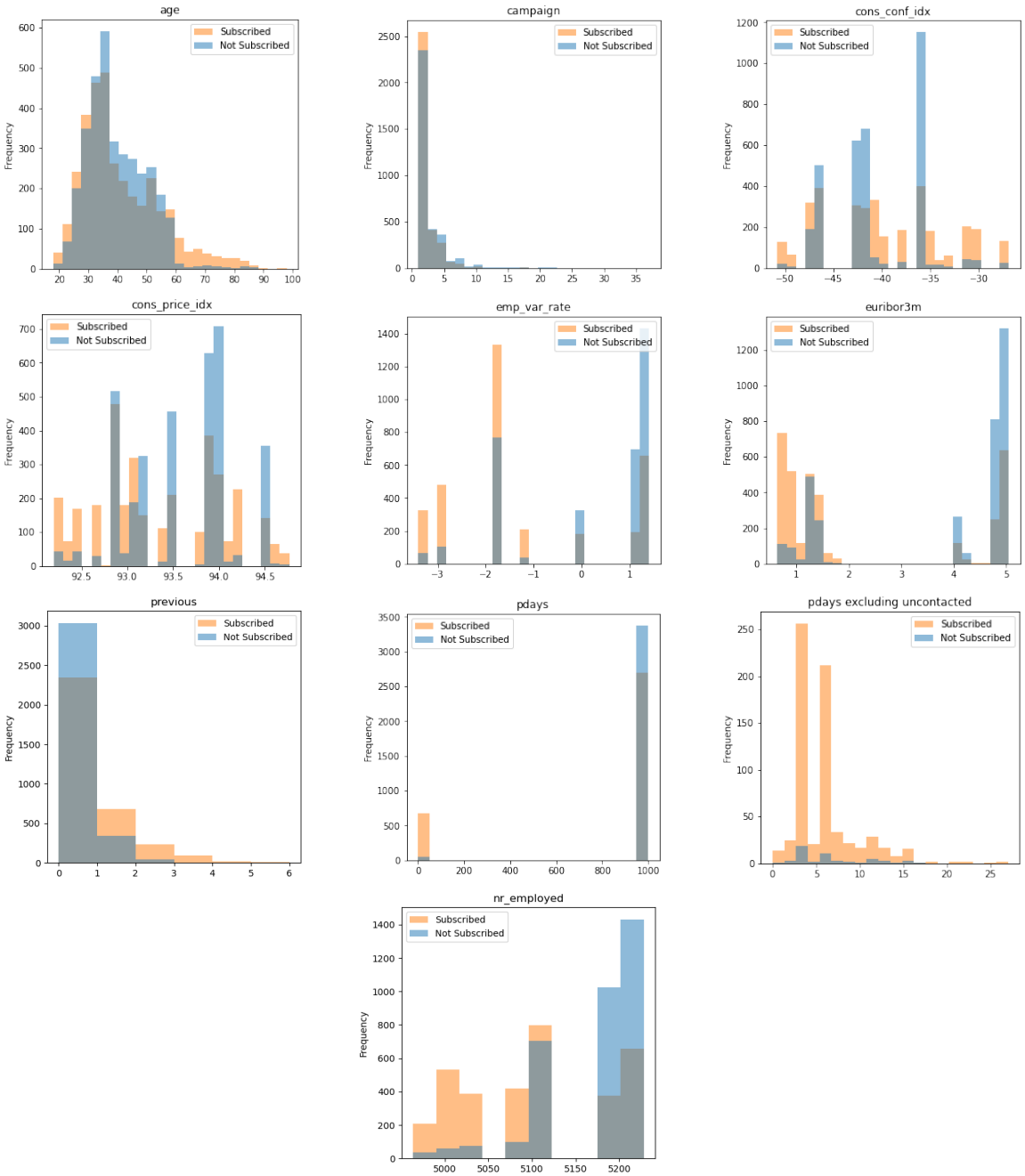
Από μία πρώτη ματιά στα δεδομένα (Σχήμα 3.4) προκύπτει πως τα περισσότερα χαρακτηριστικά έχουν κάποιο βαθμό διαχωριστικής ικανότητας. Εκείνα της ημέρας, του καταναλωτικού και του στεγαστικού δανείου είναι τα μοναδικά, που τουλάχιστον μεμονωμένα, φαίνεται να μην εμπεριέχουν σημαντική πληροφορία. Βέβαια, σε κανένα χαρακτηριστικό δεν εμφανίζεται κάποιο καλά ορισμένο όριο που θα μπορούσε να μας οδηγήσει σε ασφαλή συμπεράσματα χωρίς περαιτέρω μελέτη. Οι κατανομές παραμένουν ένα πολύτιμο εργαλείο φυσικά, καθώς δύναται να επανεξεταστούν μετά το πέρας της ταξινόμησης για να εξαχθούν συμπεράσματα ως προς τα σπουδαιότερα χαρακτηριστικά.

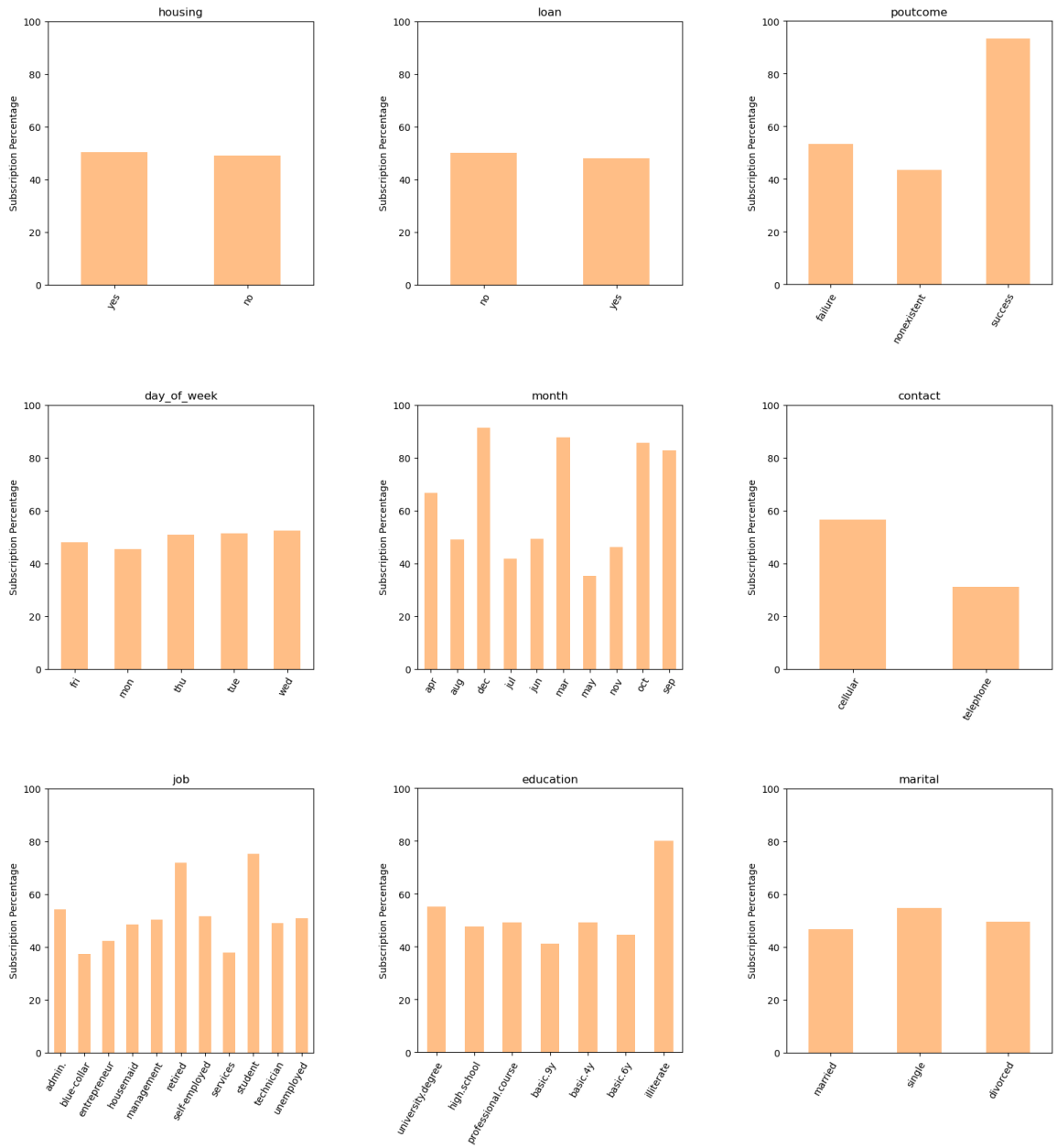
Μία δεύτερη προσέγγιση στο ζήτημα της εύρεσης προτύπων είναι εκείνη της μελέτης συσχέτισης των χαρακτηριστικών. Χρησιμοποιείται ο συντελεστής Pearson που ορίζεται ως:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.1}$$

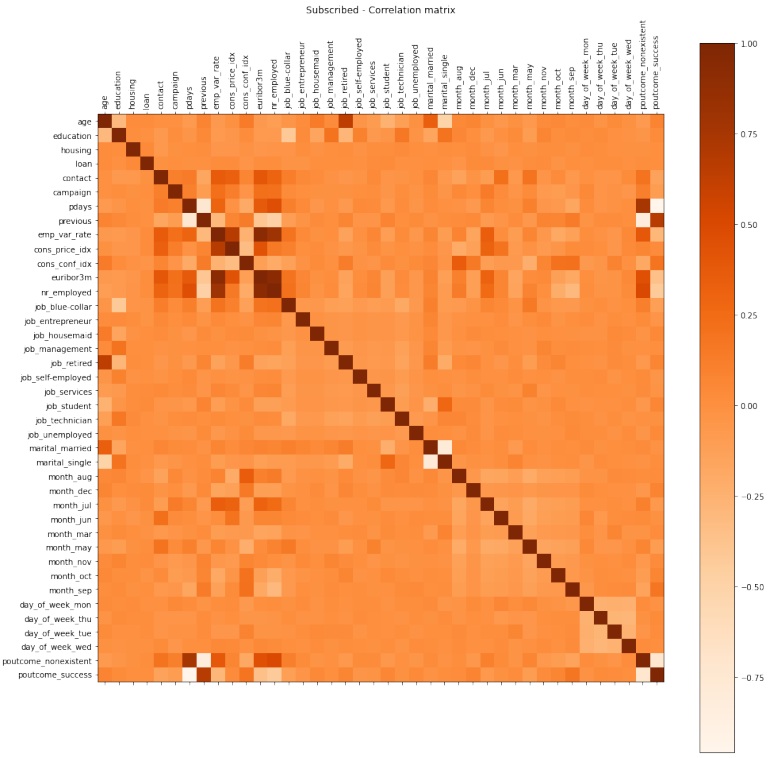
Παίρνει τιμές από -1 έως 1 και αναπαριστά τη γραμμική συσχέτιση των χαρακτηριστικών. Τα αποτελέσματα παρουσιάζονται σε πίνακες, διαφορετικούς για κάθε κλάση

(Σχήμα 3.5). Χαρακτηριστικά με ισχυρή συσχέτιση μπορούν να θεωρηθούν πως εμπεριέχουν «ίδια» πληροφορία και συνεπώς να αφαιρεθούν διατηρώντας μόνο ένα ανά ζεύγος. Καθώς κάτι τέτοιο επιλέγεται να μην γίνει στην παρούσα μελέτη, είναι αναγκαίο να ληφθούν υπόψιν οι συσχετίσεις κατά την εξαγωγή της σπουδαιότητας των χαρακτηριστικών.

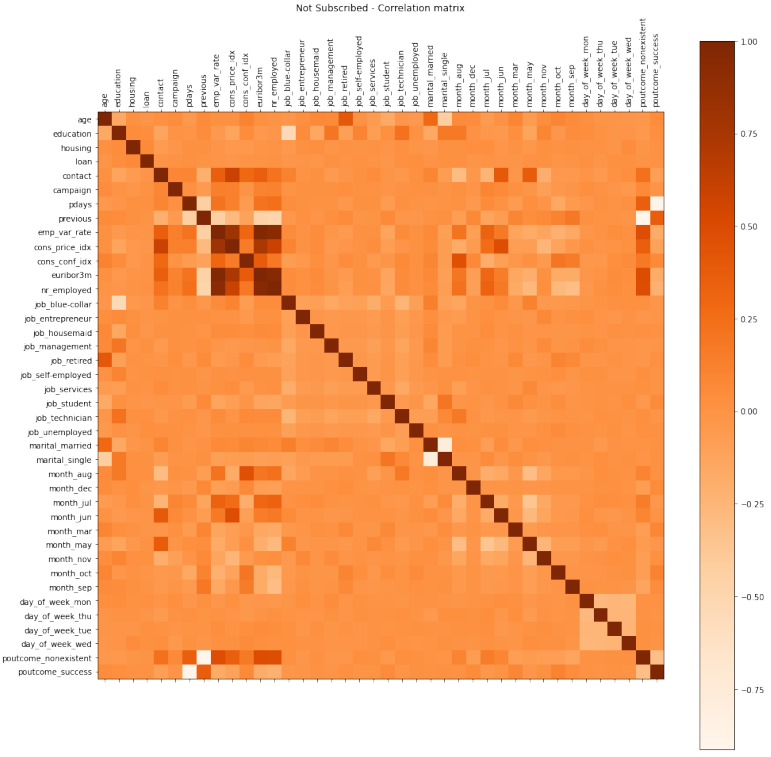




Σχήμα 3.4: Κατανομές των χαρακτηριστικών του συνόλου εκπαίδευσης



«α΄»



«β΄»

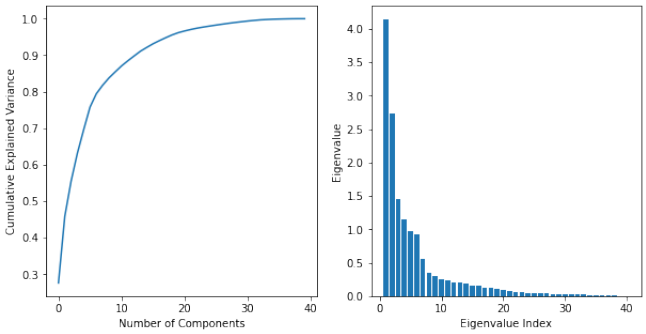
Σχήμα 3.5: Πίνακας συσχέτισης μεταβλητών. α) Για τη θετική κλάση. β) Για την αρνητική κλάση

3.5 Ανάλυση PCA

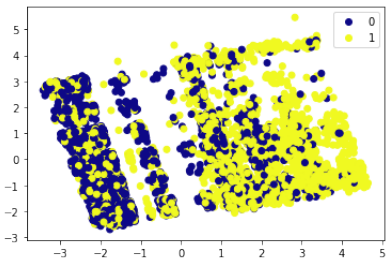
Το τελικό βήμα προτού υλοποιηθούν τα μοντέλα, είναι η ανάλυση κυρίων συνιστωσών. Όπως αναδεικνύεται στο σχήμα 3.6, το μεγαλύτερο ποσοστό της διαθέσιμης πληροφορίας περιέχεται στα πρώτα έξι νέα χαρακτηριστικά, με διαφορές στην ίδια τάξη μεγέθους. Λόγω του σχετικά μικρού δείγματος, αναλογικά πάντα με τις υπολογιστικές ικανότητες του διαθέσιμου συστήματος και τις ανάγκες των υλοποιούμενων μοντέλων, δεν είναι απαραίτητη η μείωση των χαρακτηριστικών. Παρά ταύτα, η ανάλυση μπορεί να αξιοποιηθεί για μια πιο διαισθητική απεικόνιση των δεδομένων.

Στο σχήμα 3.7 φαίνονται τα δεδομένα εκπαίδευσης στον δισδιάστατο χώρο, που ορίζεται από τα πρώτα δύο χαρακτηριστικά όπως προέκυψαν από την ανάλυση PCA. Σαφώς δεν περιέχουν όλη τη διαθέσιμη πληροφορία αλλά αποτελούν μια αρκετά ικανοποιητική προσέγγιση για την επιλογή των ταξινομητών. Υπάρχει σημαντική εισχώρηση σημείων της κάθε κλάσης στον χώρο της άλλης, ενώ αυτός φαίνεται να είναι γενικά γραμμικώς διαχωρίσιμος.

Συμπερασματικά, αναμένεται ένας γραμμικός ταξινομητής να οδηγήσει σε καλά αποτελέσματα, ενώ παράλληλα, τα μη γραμμικά μοντέλα θα έχουν μικρά περιθώρια βελτίωσης προτού αρχίσουν να υπερεκπαιδεύονται στα διάφορα μεμονωμένα σημεία που έχουν εισχωρήσει στην αντίθετη κλάση.



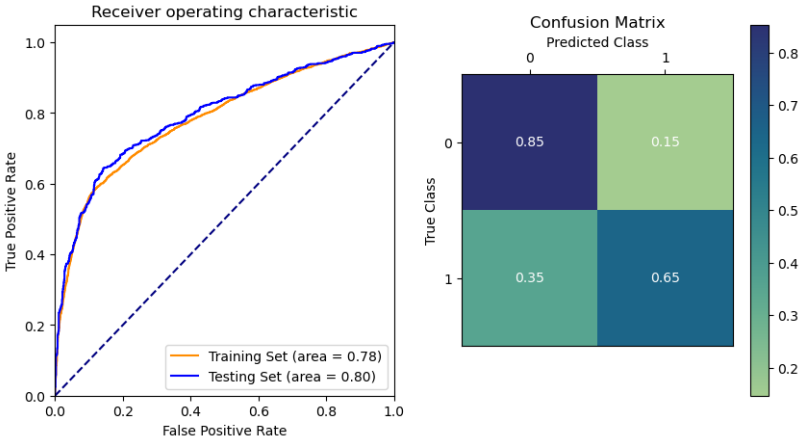
Σχήμα 3.6: Ανάλυση PCA



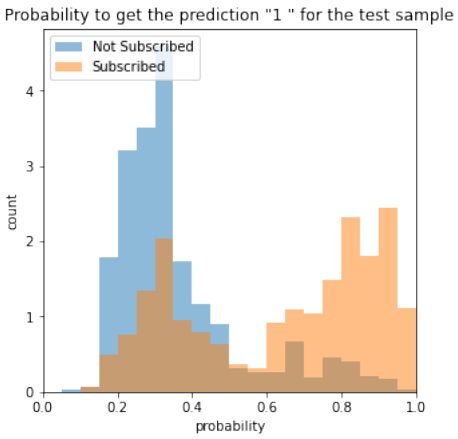
Σχήμα 3.7: Αναπαράσταση των χαρακτηριστικών στον δισδιάστατο χώρο

3.6 Λογιστική Παλινδρόμηση

Παρά την απλότητά τους, οι γραμμικοί ταξινομητές αξιοποιούνται στις περισσότερες εφαρμογές. Όπως αναδείχθηκε από την απεικόνιση στον δισδιάστατο χώρο, οι κλάσεις παρουσιάζουν γραμμική διαχωρισιμότητα. Συνεπώς υλοποιήθηκε ο γραμμικός ταξινομητής λογιστικής παλινδρόμησης ως ένα βασικό μοντέλο, τα αποτελέσματα του οποίου θα κληθούν να βελτιώσουν τα πιο σύνθετα μοντέλα στη συνέχεια.



«α΄»



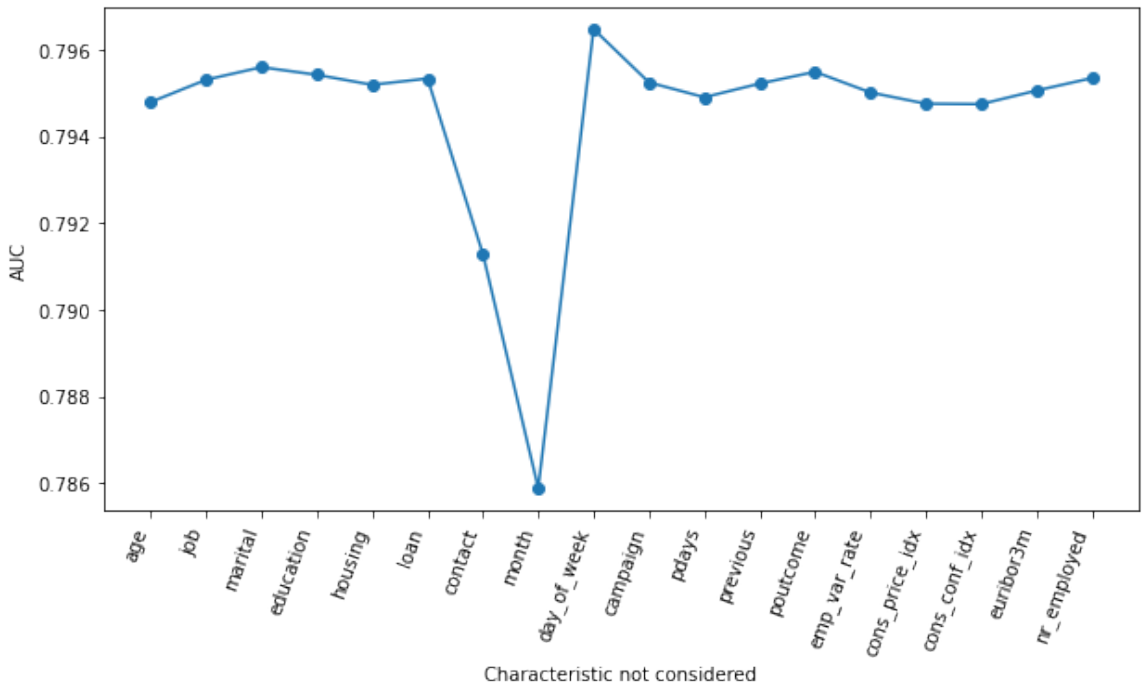
«β΄»

Σχήμα 3.8: Αποτελέσματα ταξινομητή λογιστικής παλινδρόμησης.

	Σύνολο Εκπαίδευσης	Σύνολο Ελέγχου
Accuracy	0.73	0.75
Precision	0.68	0.71
Recall	0.62	0.65
F1score	0.70	0.72
Time Elapsed:	$0.09s$	

Πίνακας 3.5: Αποτελέσματα του ταξινομητή λογιστικής παλινδρόμησης

Για τη βελτιστοποίηση της συνάρτησης κόστους επιλέχθηκε η μέθοδος lbfgs [6] ενώ για την οριοθέτηση της χρησιμοποιήθηκε η ευκλείδεια νόρμα. Η διαδικασία συγκλίνει πολύ γρήγορα, με όριο τις 100 επαναλήψεις και δεν δύναται να αποδώσει καλύτερα αποτελέσματα στο σύνολο εκπαίδευσης χωρίς κάποιον πρότερο μετασχηματισμό του χώρου των χαρακτηριστικών. Από το Σχήμα 3.8 το μοντέλο τείνει να έχει περισσότερα FN αλλά επιτυγχάνει ικανοποιητική ταξινόμηση στην κλάση 0. Αυτό αντικατοπτρίζεται και στην ποσότητα recall. Με βάση το Σχήμα 3.9 το πιο σημαντικό χαρακτηριστικό ήταν ο μήνας που έλαβε χώρα η επικοινωνία αν και η διαφορά που προκύπτει στο εμβαδόν είναι αρκετά μικρή.



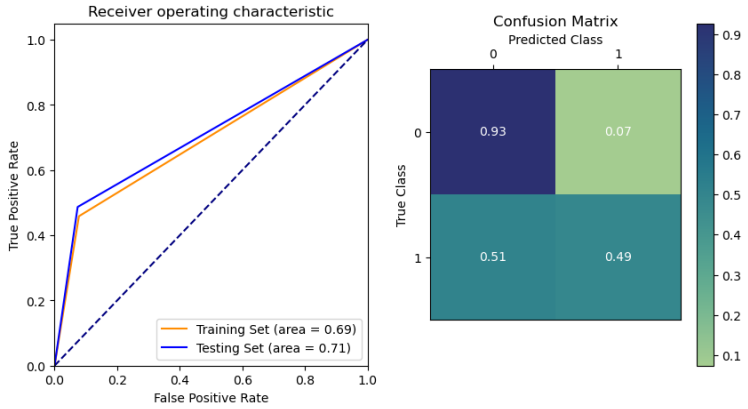
Σχήμα 3.9: Σπουδαιότητα χαρακτηριστικών για τη λογιστική παλινδρόμηση με αφαίρεση ενός χαρακτηριστικού κάθε φορά.

3.7 Ενισχυμένα Δέντρα απόφασης

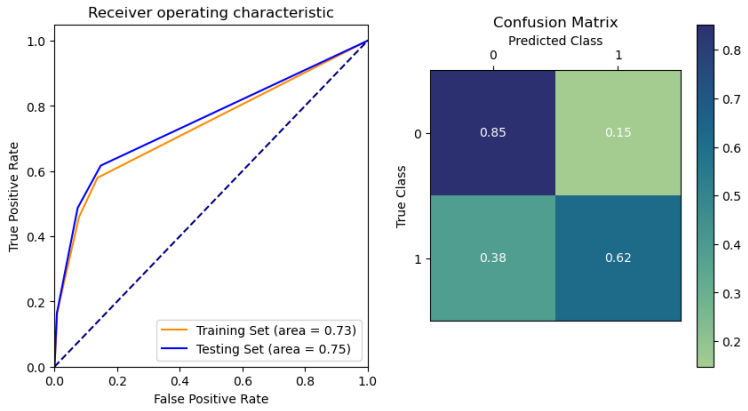
Ρηχά δέντρα

Όπως έχει προαναφερθεί, τα ενισχυμένα δέντρα απόφασης βασίζονται στην κατάλληλη βελτιστοποίηση και σύνθεση αδύναμων ταξινομητών. Για να γίνει εμφανής η βελτίωση των αποτελεσμάτων, παρουσιάζονται αρχικά οι επιδόσεις ρηχών δέντρων, που χρησιμοποιήθηκαν στη συνέχεια στις μεθόδους ενίσχυσης. Αν και έχουν πραγματοποιηθεί μία και τρεις διαιρέσεις αντίστοιχα, είναι εμφανές ότι τα αποτελέσματα υπερτερούν μιας τυχαίας ταξινόμησης. Πληρούται δηλαδή η αναγκαία προϋπόθεση για έναν weak learner που μπορεί τουλάχιστον να αναγνωρίσει κάποια πρότυπα.

Στο υπόλοιπο της ενότητας θα υλοποιηθούν οι μέθοδοι random forest, adaptive boost, gradient boost και θα μελετηθεί η επίδραση που έχει η αρχιτεκτονική των αδύναμων ταξινομητών σε αυτές. Επιπλέον θα αξιολογηθεί η επίδοσή τους ενώ τελικά θα εξαχθεί ο τρόπος με τον οποίο αξιοποιούν τα δεδομένα.



«α΄»



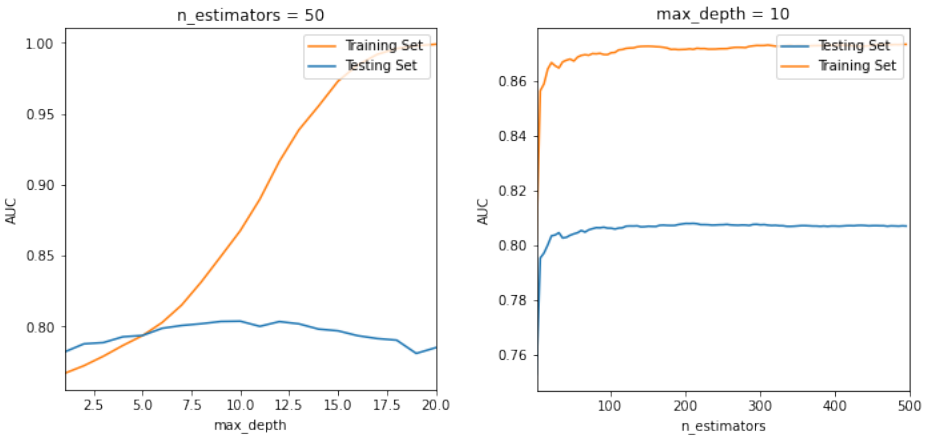
«β΄»

Σχήμα 3.10: Ταξινόμηση με ρηχά δέντρα. α) Δέντρο βάθους 1. β) Δέντρο βάθους 2.

3.7.1 Random Forest

Για τον ταξινομητή random forest έγινε έλεγχος των υπερ-παραμέτρων μέγιστου βάθους και μεγέθους δάσους. Η ποσότητα max_featues ορίστηκε ως 'sqrt' που συνεπάγεται ότι για κάθε διαίρεση λαμβάνεται υπόψιν μόνο ένα υποσύνολο των χαρακτηριστικών. Όπως είναι γνωστό από τη θεωρία, η συγκεκριμένη προσέγγιση έχει την ικανότητα να μειώνει σχεδόν μονοτονικά την διακύμανση και άρα στόχος είναι να βρεθεί το κατάλληλο «σημείο ισορροπίας», για το οποίο επιτυγχάνονται τα καλύτερα αποτελέσματα χωρίς μεγάλο πλήθος επαναλήψεων. Με ένα αρχικό μέγεθος δάσους $n\_estimators = 50$ (Σχήμα 3.11 α΄) φαίνεται πως μια καλή επιλογή είναι δέντρα βάθους 10. Επιπλέον, από το Σχήμα 3.11 β΄ δεν παρατηρείται σοβαρή βελτίωση για περισσότερους από 200 εκτιμητές. Οι τελικές τιμές των υπερπαραμέτρων είναι οι παρακάτω:

- `max_depth = 10`
- `n_estimators= 200`



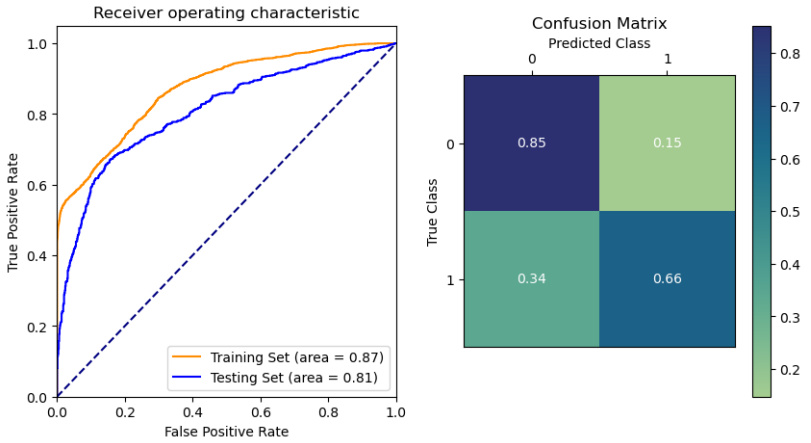
«α΄» *«β΄»*

Σχήμα 3.11: Συμπεριφορά της μεθόδου random forest για διάφορες τιμές των υπερ-παραμέτρων. α) AUC ως συνάρτηση του μέγιστου βάθους του weak learner β) AUC ως συνάρτηση του μεγέθους του δάσους.

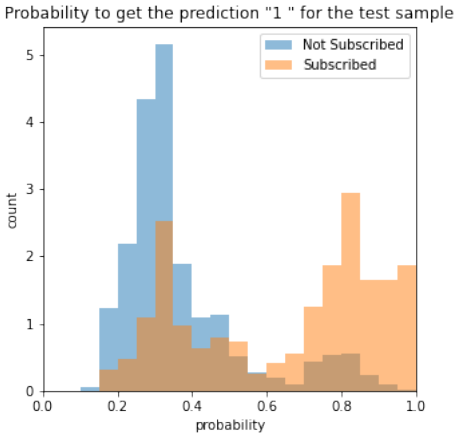
Η σπουδαιότητα των χαρακτηριστικών για το μοντέλο επιδιώχθηκε να εξαχθεί με δύο τρόπους. Αρχικά έγινε χρήση της ιδιότητας .feature_importances που υπολογίζεται αυτόματα στο πακέτο scklearn και επιλέγει τα χαρακτηριστικά εκείνα που οδήγησαν στις περισσότερες διαιρέσεις. Με αυτή τη προσέγγιση δύναται να «χαθούν» χαρακτηριστικά καθώς η διαίρεση γίνεται πάντα με βάση το καλύτερο. Συνεπώς υιοθετήθηκε και μία δεύτερη μέθοδος, κατά την οποία υπολογίζεται το εμβαδόν κάτω από την καμπύλη ROC με αφαίρεση ενός χαρακτηριστικού τη φορά. Η προσέγγιση αυτή θα αξιοποιηθεί σε όλα τα μοντέλα, καθώς οτιδήποτε δεν βασίζεται σε ταξινομήσεις δέντρων δεν έχει τη δυνατότητα να εξάγει αυτόματα τις σπουδαιότητες.

Τα αποτελέσματα φαίνονται στο Σχήμα 3.13. Παρατηρείται ότι η μεμονωμένη αφαίρεση χαρακτηριστικών έχει πολύ μικρή επίδραση στο εμβαδόν της καμπύλης ROC. Βέβαια, οι οικονομικοί δείκτες και κυρίαρχα ο EURIBOR3M, φαίνεται να συνεισφέρουν στο μεγαλύτερο βαθμό λαμβάνοντας υπόψιν και τις δύο προσεγγίσεις.

Είναι εμφανές ότι με τη σχετικά απλή επαναληπτική διαδικασία και με τυχαίες επιλογές στα δείγματα και τα χαρακτηριστικά, όλες οι ποσότητες (Σχήμα 3.12, Πίνακας 3.6) που χαρακτηρίζουν την ποιότητα του ταξινομητή βλετιώθηκαν. Σημειώνεται επιπλέον ότι ο χρόνος εκπαίδευσης είναι ιδιαίτερα μικρός.



«α΄»

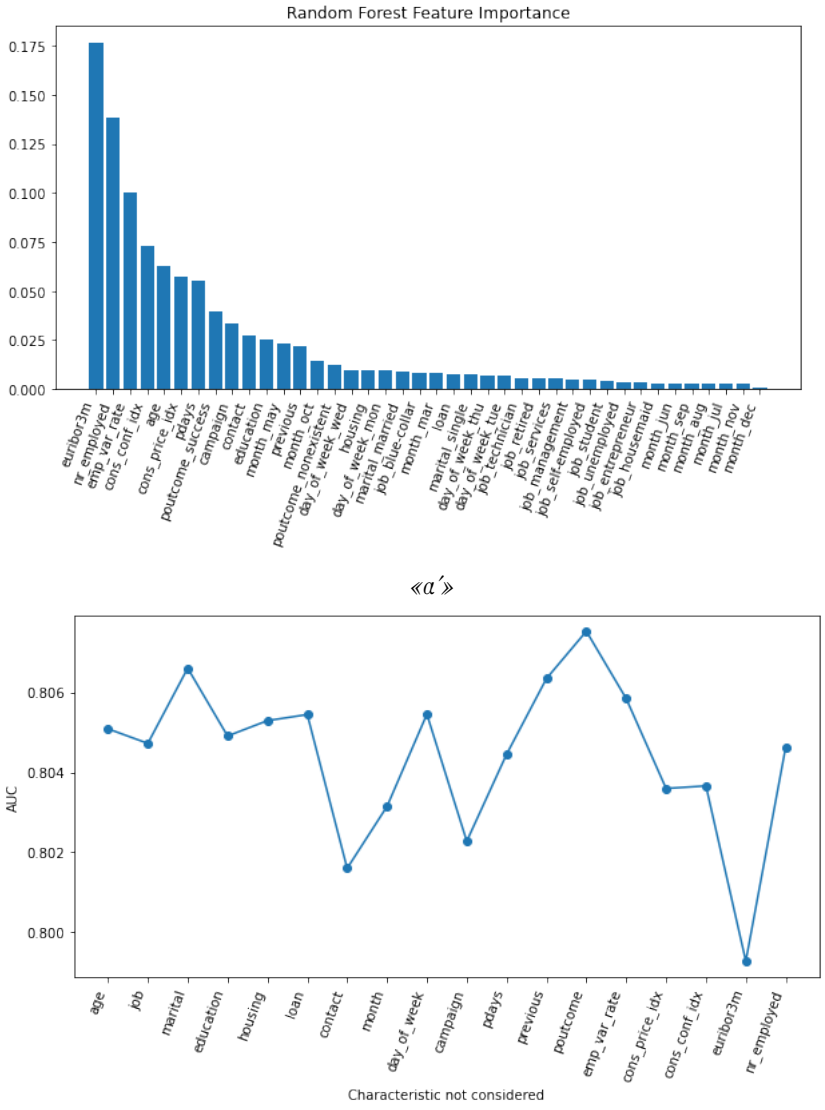


«β΄»

Σχήμα 3.12: Ταξινόμηση με Random Forest

	Σύνολο Εκπαίδευσης	Σύνολο Ελέγχου
Accuracy:	0.77	0.75
Precision:	0.73	0.72
Recall:	0.64	0.66
F1score:	0.73	0.73
Time Elapsed:	$0.71s$	

Πίνακας 3.6: Αποτελέσματα ταξινομητή Random Forest



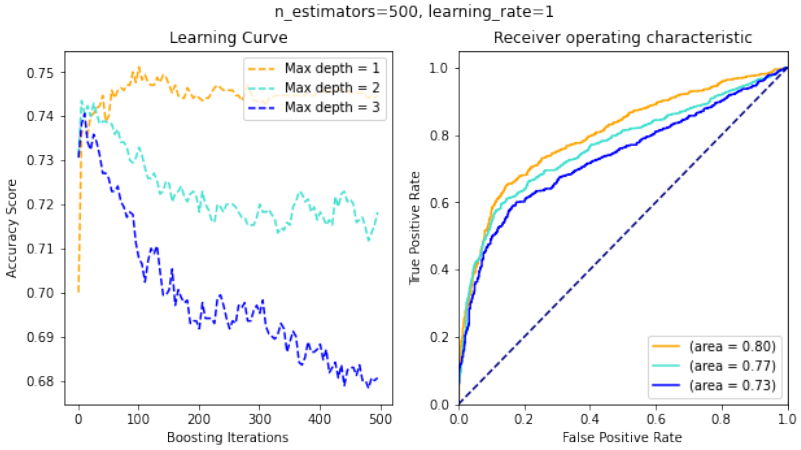
«α΄»

«β΄»

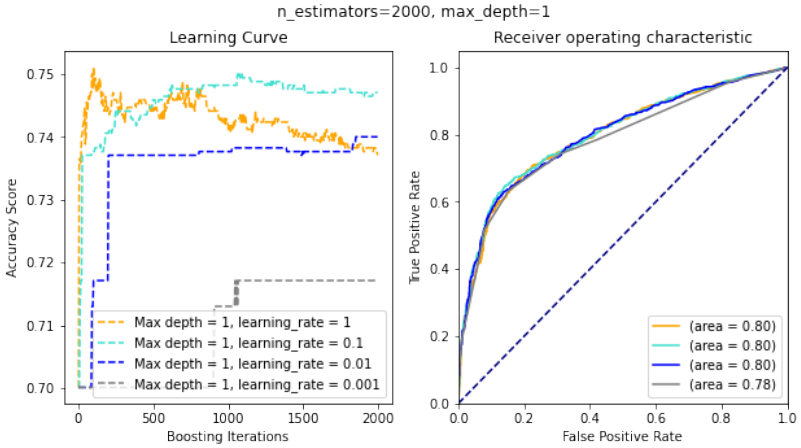
Σχήμα 3.13: Σπουδαιότητα χαρακτηριστικών για τη μέθοδο Random Forests. α) Όπως προκύπτει από τα splits. β)Αφαιρώντας πλήρως ένα χαρακτηριστικό κάθε φορά.

3.7.2 Ενίσχυση με τον αλγόριθμο AdaBoost

Αντίστοιχα με το μοντέλο random forest, και για τον αλγόριθμο AdaBoost πραγματοποιείται βελτιστοποίηση των υπέρ-παραμέτρων. Το βάθος του αδύναμου ταξινομητή επηρεάζει έντονα τα αποτελέσματα και πιο συγκεκριμένα, οδηγεί πολύ γρήγορα σε υπερεκπαίδευση του μοντέλου. Αυτό παρουσιάζεται ως απότομη πτώση στο γράφημα ακρίβειας, όπως αυτή υπολογίζεται στο σύνολο ελέγχου και εύκολα φαίνεται ο λόγος που συνήθως υλοποιείται με μία διαίρεση σε κάθε επανάληψη.



«α΄»



«β΄»

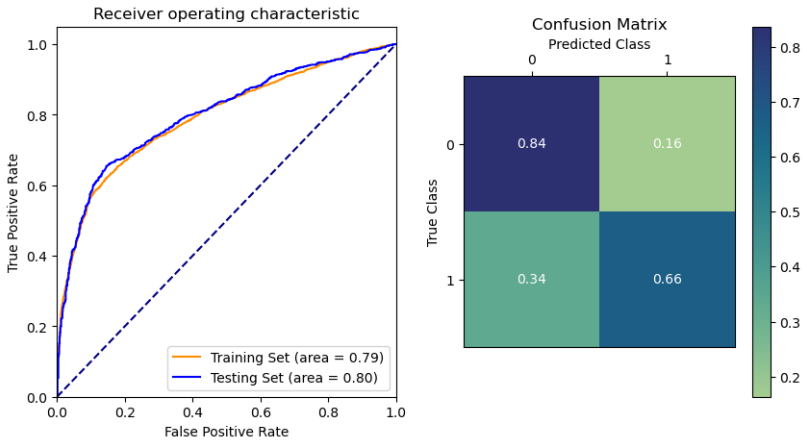
Σχήμα 3.14: Συμπεριφορά του μοντέλου AdaBoost στο σύνολο ελέγχου για διάφορες τιμές α) του μέγιστου βάθους β) της παραμέτρου learning rate.

Βελτιστοποιείται επιπλέον η παράμετρος learning rate η οποία μεταβάλλει τα βάρη κατά τον τελικό συνυπολογισμό των αδύναμων ταξινομητών. Όσο μικρότερη είναι η τιμή της, τόσο περισσότεροι εκτιμητές απαιτούνται, αλλά παράλληλα γίνεται πιο εύκολη η βελτίωση του μοντέλου με αποφυγή της υπερεκπαίδευσης. Για πολύ μικρές τιμές η σύγκλιση στα επιθυμητά αποτελέσματα καθυστερεί αρκετά, ενώ επιπλέον αυτά δεν βελτιώνονται. Όπως φαίνεται από το Σχήμα 3.14, οι ιδανικές παράμετροι για το μοντέλο

είναι:

- `max_depth = 1`
- `learning_rate = 0.1`
- `n_estimators = 1000`

Για την κατασκευή των γραφημάτων χρησιμοποιήθηκε η μέθοδος staged_predict που είναι διαθέσιμη στο πακέτο sklearn, καθώς επιτρέπει τον προσεγγιστικό υπολογισμό της ποσότητας accuracy κατά τη διάρκεια της εκπαίδευσης. Με τον τρόπο αυτό αποφεύγεται η υλοποίηση διαφορετικού μοντέλου για κάθε πλήθος ενισχυτικών επαναλήψεων. Λόγω της δομής του αλγορίθμου, δεν είναι δυνατή η εξαγωγή εκ των υστέρων πιθανοτήτων και τα αποτελέσματα έχουν δυαδική μορφή (0, 1 για την κλαση). Βέβαια από τον πίνακα σύγχυσης και τα ποσοτικά αποτελέσματα (Σχήμα 3.15, Πίνακας 3.7) βλέπουμε αντίστοιχη συμπεριφορά με τα προηγούμενα μοντέλα. Τα αντικείμενα της κλάσης 0 ταξινομούνται σωστά ενώ υπάρχουν αρκετά FN.

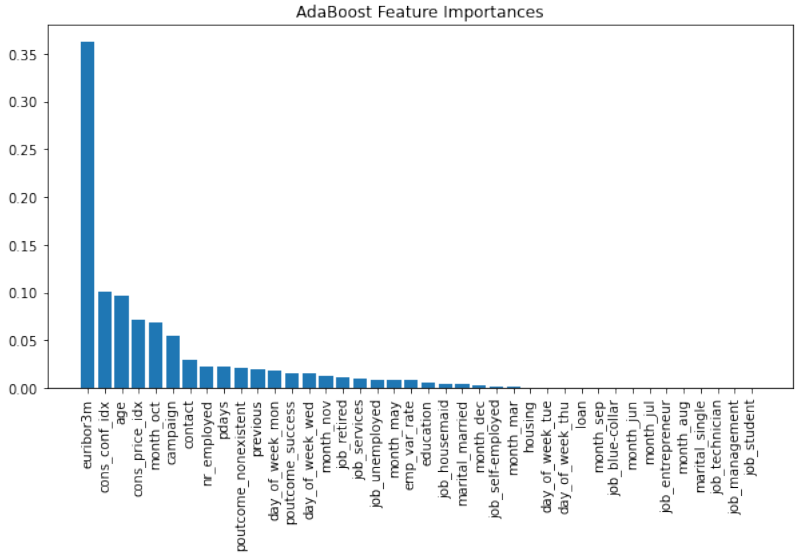


Σχήμα 3.15: Ταξινόμηση με τον αλγόριθμο AdaBoost.

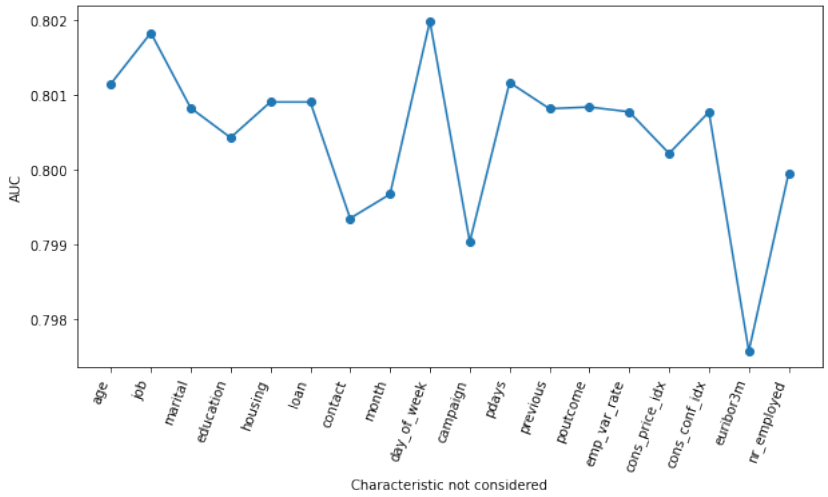
	Σύνολο Εκπαίδευσης	Σύνολο Ελέγχου
Accuracy:	0.74	0.75
Precision:	0.69	0.71
Recall:	0.63	0.67
F1score:	0.70	0.73
Time Elapsed:	$4.92s$	

Πίνακας 3.7: Αποτελέσματα ταξινομητή AdaBoost

Και οι δύο προσεγγίσεις (Σχήμα 3.16) στο ζήτημα σπουδαιότητας χαρακτηριστικών, αναδεικνύουν το EURIBOR3M ως εκείνο με τη μεγαλύτερη διαχωριστική ικανότητα. Στα υπόλοιπα παρατηρείται διαφοροποίηση αλλά η επίδραση στην επίδοση είναι ελάχιστη.



«α΄»

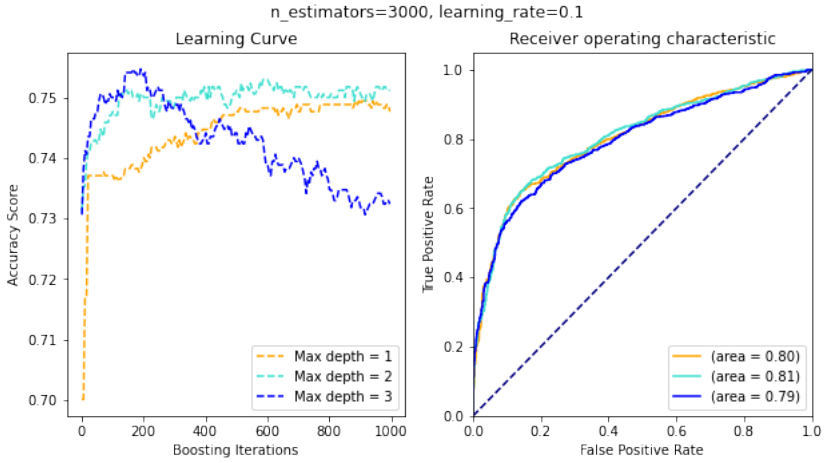


«β΄»

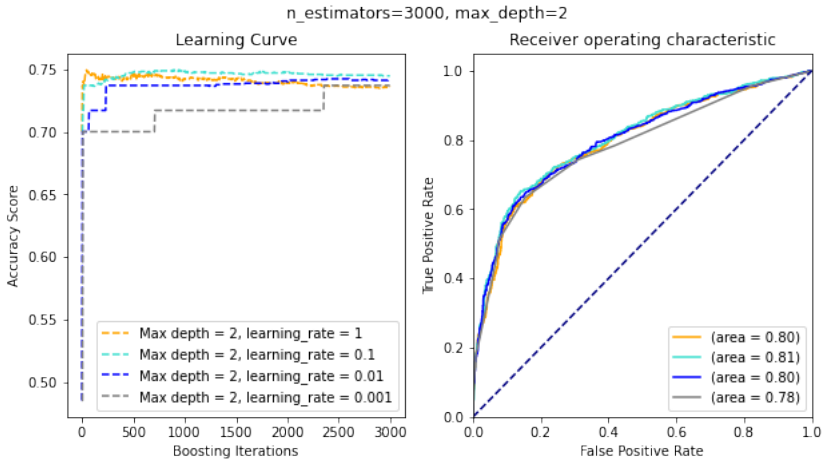
Σχήμα 3.16: Σπουδαιότητα χαρακτηριστικών για τη μέθοδο AdaBoost. α) Όπως προκύπτει από τα splits. β)Αφαιρώντας πλήρως ένα χαρακτηριστικό κάθε φορά.

3.7.3 Ενίσχυση με τη μέθοδο Gradient Boost

Και εδώ απαιτείται η ρύθμιση των υπέρ-παραμέτρων βάθους και ρυθμού εκμάθησης. Σε αντίθεση με τη μέθοδο AdaBoost, ο αλγόριθμος Gradient Boost (GBT), φαίνεται να έχει καλύτερα αποτελέσματα με τη χρήση δέντρου μέγιστου βάθους 2. Περαιτέρω αύξηση της πολυπλοκότητας του αδύναμου ταξινομητή οδηγεί και πάλι σε υπερεκπαίδευση.



«α΄»



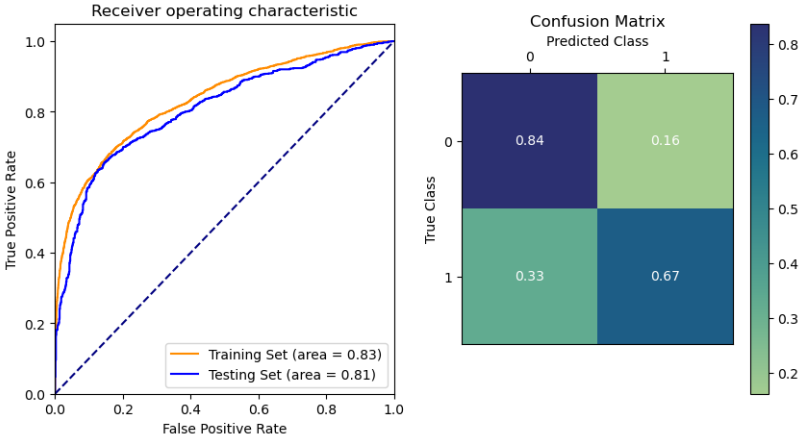
«β΄»

Σχήμα 3.17: Συμπεριφορά του μοντέλου GradientBoost στο σύνολο ελέγχου για διάφορες τιμές α) του μέγιστου βάθους β) της παραμέτρου learning rate.

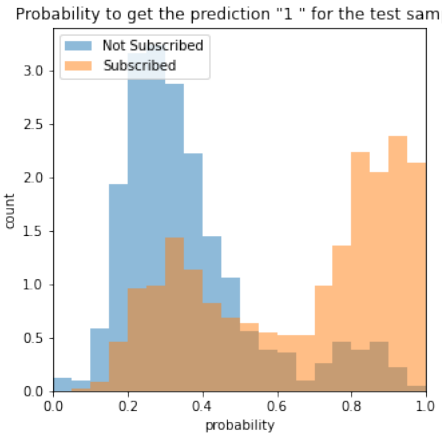
Οι πολύ μικροί ρυθμοί εκμάθησης φαίνεται να χρειάζονται περισσότερες από 3000 επαναλήψεις για να επιτύχουν καλή ακρίβεια κάτι το οποίο απαιτεί μεγάλο υπολογιστικό χρόνο. Λαμβάνοντας τα παραπάνω υπόψη, και με βάση το Σχήμα 3.17, οι παράμετροι για το συγκεκριμένο μοντέλο επιλέχθηκαν ως:

- `max_depth = 2`

- `learning_rate = 0.1`
- `n_estimators = 700`



«α΄»



«β΄»

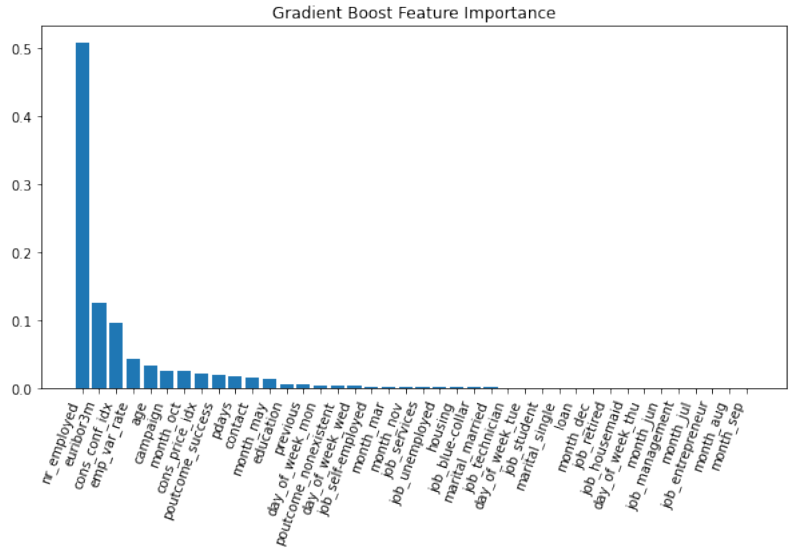
Σχήμα 3.18: Ταξινόμηση με GBT.

	Σύνολο Εκπαίδευσης	Σύνολο Ελέγχου
Accuracy:	0.76	0.75
Precision:	0.71	0.72
Recall:	0.65	0.67
F1score:	0.73	0.74
Time Elapsed:	$3.41s$	

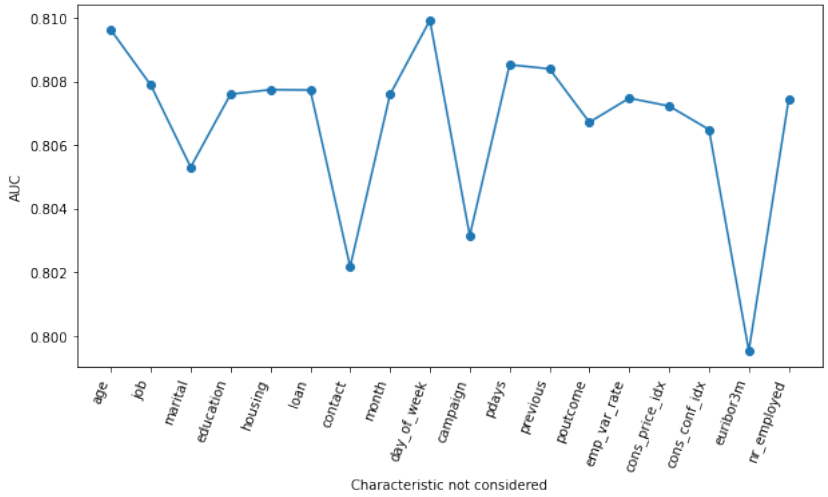
Πίνακας 3.8: Αποτελέσματα ταξινομητή GBT

Κατά την ανάλυση σπουδαιότητας των χαρακτηριστικών φανερώνεται το φαινόμενο της «καταστολής» ορισμένων από αυτά στις διαιρέσεις. Αν και το μοντέλο πραγματοποίησε

καλύτερα splits με βάση το nr_employed, η μεγαλύτερη πτώση της ποσότητας AUC παρουσιάζεται κατά την αφαίρεση του χαρακτηριστικού EURIBOR3M.



«α΄»



«β΄»

Σχήμα 3.19: Σπουδαιότητα χαρακτηριστικών για τη μέθοδο GBT. α) Όπως προκύπτει από τα splits. β)Αφαιρώντας πλήρως ένα χαρακτηριστικό κάθε φορά.

3.8 Νευρωνικό Δίκτυο

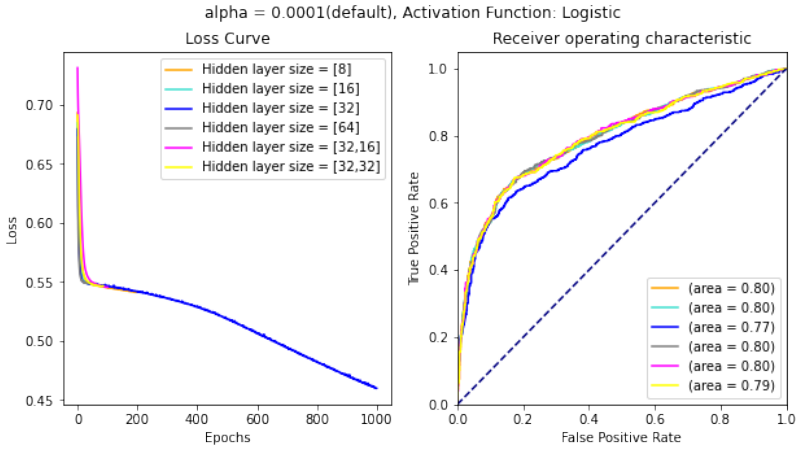
Το νευρωνικό δίκτυο (Neural Network/NN) έχει τη μεγαλύτερη δυσκολία ρύθμισης λόγω του μεγάλου πλήθους υπέρ-παραμέτρων. Σαν συνάρτηση ενεργοποίησης επιλέχθηκε η σιγμοειδής με βάση την οποία έγιναν οι υπόλοιπες ρυθμίσεις. Αρχικά μελετήθηκε η αρχιτεκτονική, δοκιμάζοντας διάφορα μεγέθη και πλήθη κρυφών επιπέδων. Είναι εμπειρικά γνωστό [25][26], ότι τα περισσότερα προβλήματα ταξινόμησης μπορούν να επιλυθούν ικανοποιητικά με μόνο ένα κρυφό επίπεδο, κάτι το οποίο φαίνεται και στο σχήμα 3.20. Επιλέχθηκε η αρχιτεκτονική ενός μόνο κρυφού επιπέδου με 32 νευρώνες, καθώς αυτή επιτυγχάνει συνεχή βελτίωση στο σύνολο εκπαίδευσης.

Για αποφυγή overfitting, ελέγχθηκαν διάφορες τιμές της ποσότητας *alpha* η οποία ρυθμίζει την ισχύ της νόρμας L^2. Η αύξησή της οδηγεί σε πιο απλές λύσεις που μπορούν να γενικεύσουν με μεγαλύτερη ευκολία. Τέλος επιλέχθηκε το κατάλληλο πλήθος εποχών. Οι παράμετροι του δικτύου είναι:

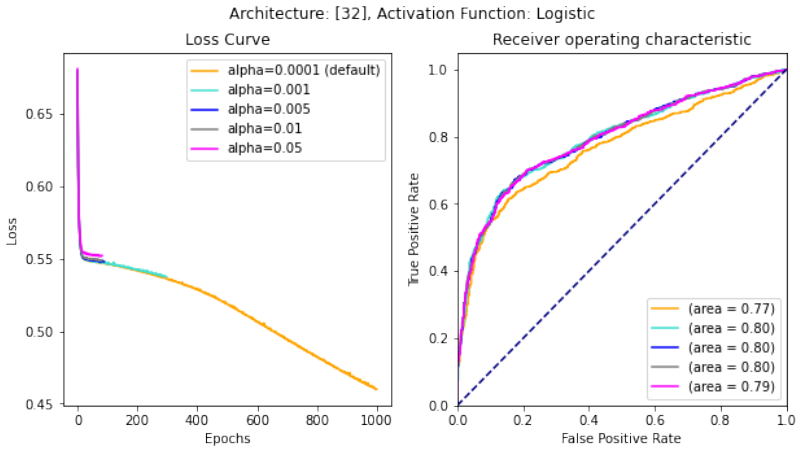
- `architecture:[32]`
- `alpha=0.001`
- `max_iter=100`

Ιδιαίτερο ενδιαφέρον παρουσιάζει το γεγονός, ότι κάποιες ρυθμίσεις αδυνατούν να βελτιώσουν περαιτέρω τα αποτελέσματα στο σύνολο εκπαίδευσης ανεξαιρέτως του πλήθους των εποχών. Φαίνεται δηλαδή ότι οι ικανότητες του μοντέλου καθώς και η ευαισθησία του στην υπερεκπαίδευση είναι κάποιες φορές ζήτημα αρχιτεκτονικής και όχι επαναλήψεων.

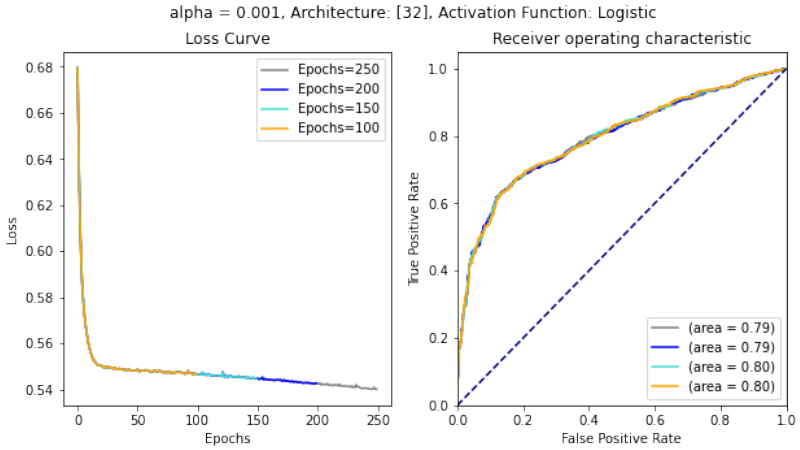
Από το Σχήμα 3.22 προκύπτει ότι σαν μεμονωμένο χαρακτηριστικό, τη μεγαλύτερη διαχωριστική ικανότητα την έχει εκείνο του μήνα. Παρόλα αυτά επιδιώχθηκε να μελετηθεί η σπουδαιότητα ανά κατηγορία, από την οποία προκύπτει ότι το υποσύνολο των χαρακτηριστικών που επηρεάζουν πιο άμεσα την τελική ταξινόμηση, είναι εκείνο που αφορά τους οικονομικούς δείκτες της περιόδου.



«α΄»

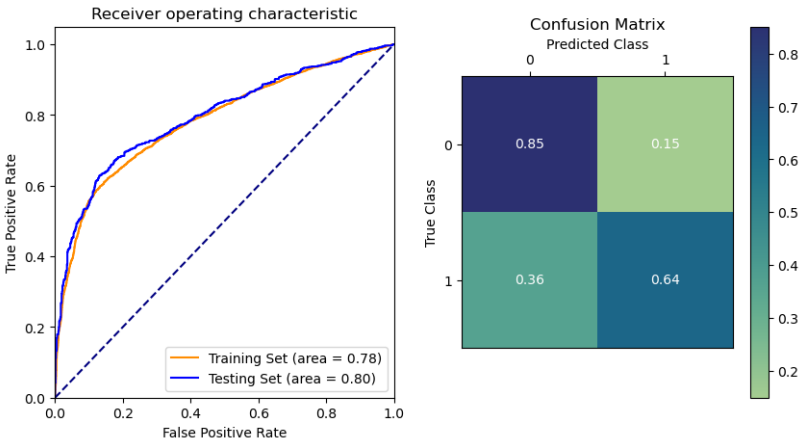


«β΄»

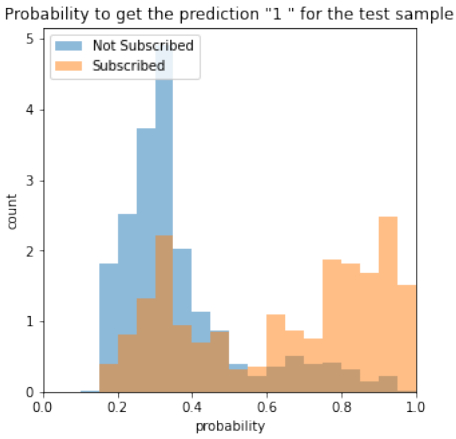


«γ΄»

Σχήμα 3.20: Επίδοση του NN στο σύνολο ελέγχου.α) Αρχιτεκτονική. β) Παράμετρος alpha. γ) Εποχές



«α΄»

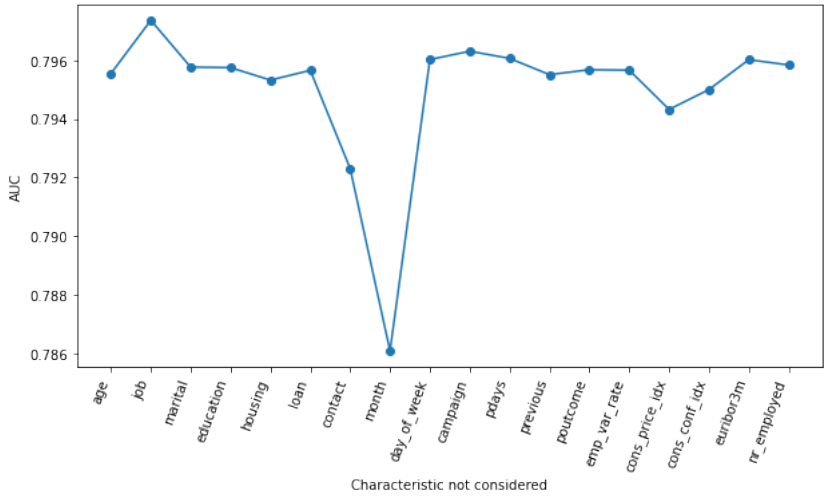


«β΄»

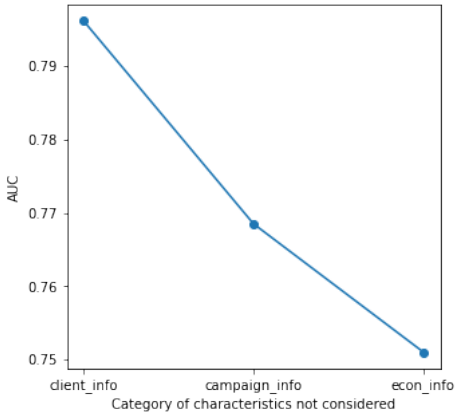
Σχήμα 3.21: Ταξινόμηση με NN.

	Σύνολο Εκπαίδευσης	Σύνολο Ελέγχου
Accuracy:	0.73	0.74
Precision:	0.68	0.71
Recall:	0.62	0.64
F1score:	0.70	0.72
Time Elapsed:	1.53s	

Πίνακας 3.9: Αποτελέσματα ταξινομητή NN



«α΄»

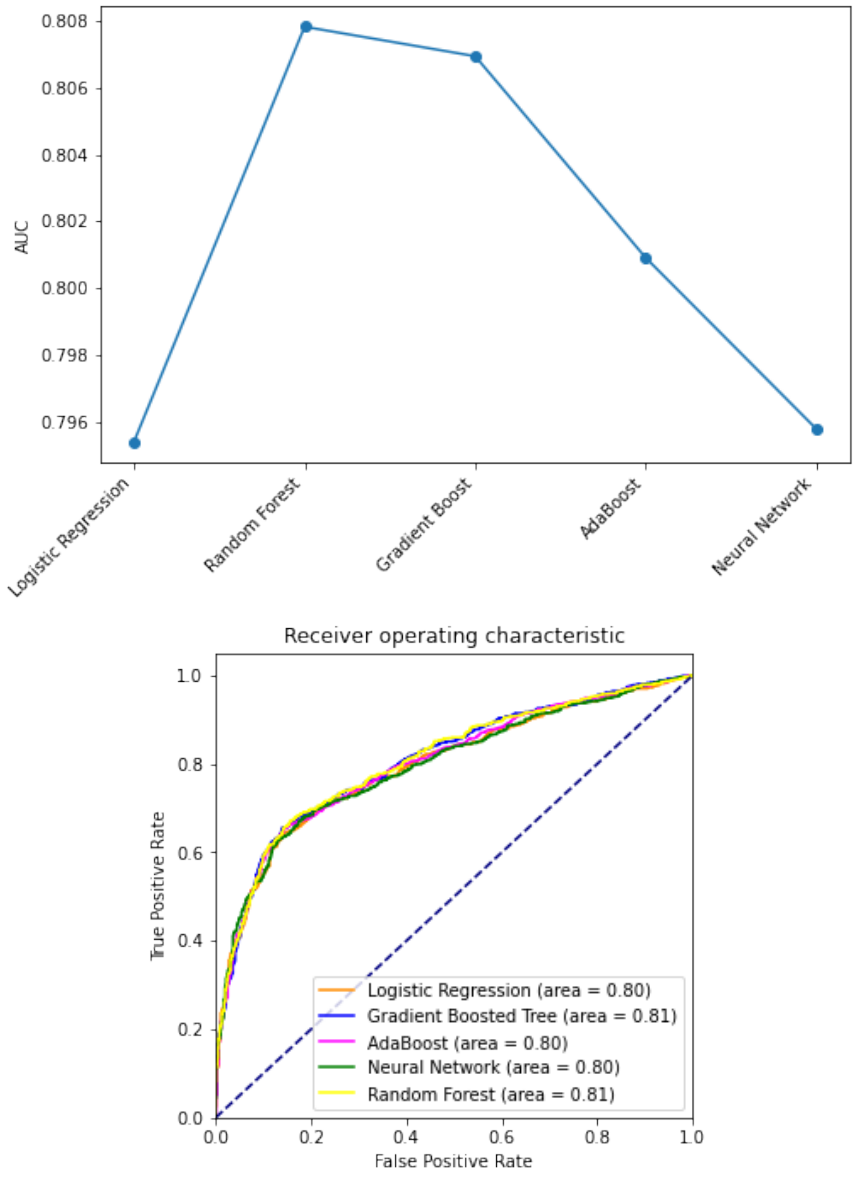


«β΄»

Σχήμα 3.22: Σπουδαιότητα χαρακτηριστικών για NN. α)Αφαιρώντας πλήρως ένα χαρακτηριστικό κάθε φορά. β) Αφαιρώντας ένα υποσύνολο χαρακτηριστικών κάθε φορά.

3.9 Σύγκριση μεθόδων

Σε ότι αφορά τις επιδόσεις των τεσσάρων μοντέλων, υπάρχουν πολύ μικρές διαφορές. Η μέθοδος random forest έχει τα καλύτερα αποτελέσματα ως προς το εμβαδόν της καμπύλης ROC με ελάχιστες υπολογιστικές απαιτήσεις. Το νευρωνικό δίκτυο, παρά την πολυπλοκότητά του, είχε αντίστοιχη διαχωριστική ικανότητα με εκείνη του ταξινομητή λογιστικής παλινδρόμησης και αισθητά μεγαλύτερο χρόνο επεξεργασίας. Αυτό μπορεί να οφείλεται σε μη ιδανική αρχιτεκτονική καθώς υφίστανται μεγάλα περιθώρια για κατασκευαστικές αλλαγές, αλλά η δυσκολία στη ρύθμιση και υλοποίηση ενός μοντέλου δεν γίνεται να αγνοηθεί σαν παράμετρος κατά τη σύγκριση.



Σχήμα 3.23: Καμπύλες ROC των μοντέλων που έχουν μελετηθεί

Οι αλγόριθμοι AdaBoost και gradient boost είχαν τις μεγαλύτερες ανάγκες υπολογιστικών πόρων όπως προκύπτει και από τον χρόνο εκπαίδευσης. Φυσικά τα αποτελέσματά τους είναι συγκρίσιμα με των υπολοίπων, αλλά ο πρώτος έρχεται με ένα σοβαρό μειονέκτημα καθώς αυτά δεν έχουν πιθανολογική ερμηνεία.

Εστιάζοντας στις υπόλοιπες ποσότητες (Πίνακας 3.10) προκύπτει πως με ακρίβεια δύο δεκαδικών ψηφίων, ο πιο αποτελεσματικός ταξινομητής είναι ο GBT. Αυτό γίνεται εμφανές και στις κατανομές των εκ των υστέρων πιθανοτήτων. Όλα τα μοντέλα εντοπίζουν με ικανοποιητική ακρίβεια την αρνητική κλάση αλλά δυσκολεύονται στην θετική, με έναν σημαντικό αριθμό ψευδώς αρνητικών εκτιμήσεων, από όπου προκύπτει και η χαμηλή τιμή της ποσότητας Recall. Η μέθοδος GBT επιτυγχάνει τις καλύτερες εκτιμήσεις για τη θετική κλάση.

	L.Reg.	RF	GBT	AdaB.	NN
Accuracy:	0.75	0.75	0.75	0.75	0.74
Precision:	0.71	0.72	0.72	0.71	0.71
Recall:	0.65	0.66	0.67	0.67	0.64
F1score:	0.72	0.73	0.74	0.73	0.71
Time (s):	0.09	0.71	3.41	4.90	1.32

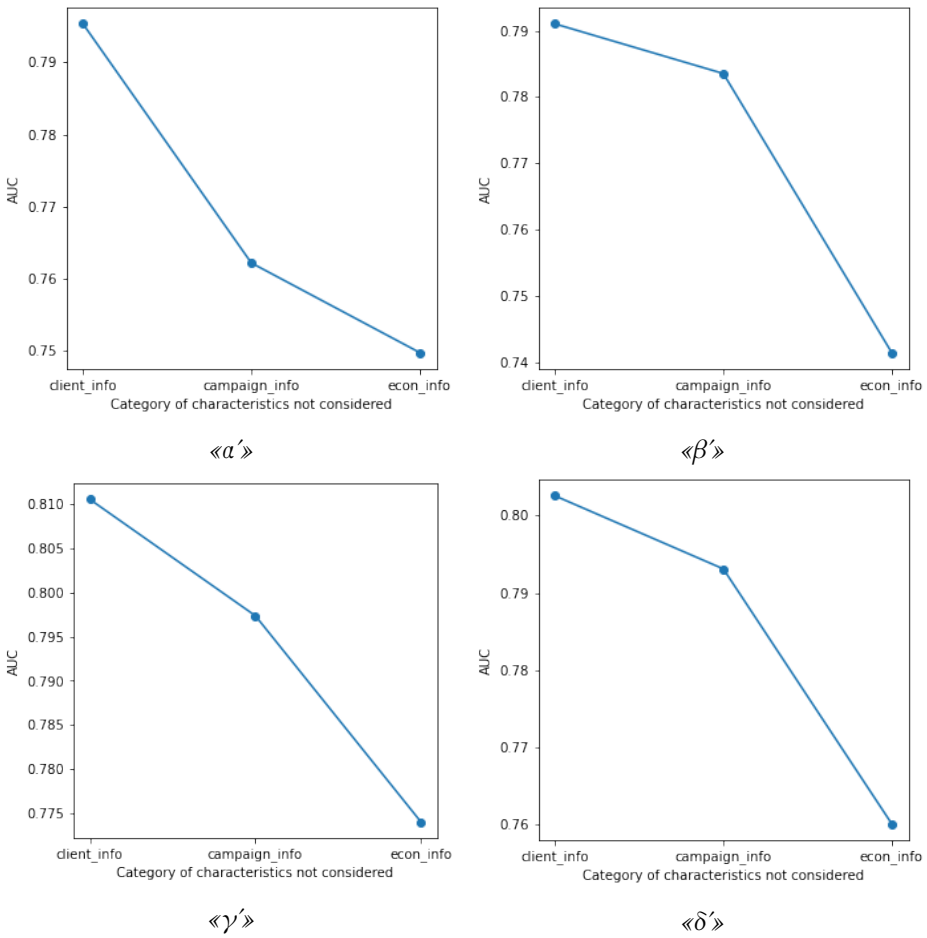
Πίνακας 3.10: Τελικά αποτελέσματα

Όσον αφορά την ευαισθησία στην υπερεκπαίδευση, τα μη γραμμικά μοντέλα χρειάστηκαν ειδική μεταχείριση, με εφαρμογή περιορισμών, καθώς ήταν πολύ στενά τα περιθώρια βελτίωσης πριν καταλήξουν στην εξειδίκευση στο σύνολο εκπαίδευσης. Αντίστοιχη συμπεριφορά δεν αναδείχθηκε στον γραμμικό ταξινομητή λογιστικής παλινδρόμησης.

Ιδιαίτερο ενδιαφέρον παρουσιάζει η σπουδαιότητα των χαρακτηριστικών ανά μέθοδο. Όλοι οι ταξινομητές που βασίζονται σε δέντρα φαίνεται να εξορύσσουν μεγάλο μέρος της πληροφορίας από τον δείκτη EURIBOR. Αντιθέτως, η λογιστική παλινδρόμηση και τα νευρωνικά δίκτυα υπέστησαν μεγαλύτερη πτώση με την αφαίρεση του μήνα. Αξιοσημείωτο βέβαια είναι, ότι παρουσιάζεται ελάχιστη επιδείνωση των αποτελεσμάτων σε όλες τις περιπτώσεις και η ταξινόμηση είναι συγκρίσιμη με την αρχική παρά την απουσία των χαρακτηριστικών.

Classifier	Feature
L.Reg.	month
RF	euribor3m
GBT	euribor3m
AdaB	euribor3m
NN	month

Πίνακας 3.11: Σπουδαιότερο χαρακτηριστικό ανά ταξινομητή



Σχήμα 3.24: Σπουδαιότητα ομάδων χαρακτηριστικών. α)Logistic Regression. β) Random Forest. γ) GBT. δ) AdaBoost. Νευρωνικό δίκτυο στο Σχήμα 3.22 β

Για τον λόγο αυτό, αξιοποιείται μία επιπλέον μέθοδος, αντίστοιχη με αυτή που υλοποιήθηκε κατά την ανάλυση του νευρωνικού δικτύου. Ελέγχεται δηλαδή η συνεισφορά συνόλων χαρακτηριστικών και μάλιστα με τέτοιο τρόπο ώστε να ληφθούν ερμηνεύσιμα συμπεράσματα. Εφαρμόστηκε συνεπώς η αρχική κατηγοριοποίηση των χαρακτηριστικών ως εκείνα που αφορούν τον πελάτη, εκείνα που αφορούν την τωρινή ή παλαιότερες προσπάθειες επικοινωνίας για προωθητικούς λόγους και εκείνα που αφορούν τους οικονομικούς δείκτες. Τα αποτελέσματα παρουσιάζονται στο Σχήμα3.24.

Παρατηρείται ότι οι ίδιοι ταξινομητές που ανέδειξαν τον μήνα ως σπουδαιότερο χαρακτηριστικό, παρουσιάζουν εμφανή πτώση αφαιρώντας τις πληροφορίες επικοινωνίας. Παρόλα αυτά και λαμβάνοντας υπόψιν όλα τα μοντέλα, φαίνεται πως το σύνολο των χαρακτηριστικών που επηρεάζει πιο άμεσα την τελική απάντηση του πελάτη και άρα την κλάση, είναι εκείνο που αφορά τους οικονομικούς δείκτες της περιόδου.

Συμπεράσματα

Στην παρούσα εργασία επιδιώκεται η μελέτη και κατ΄ επέκταση η υλοποίηση ταξινομητών σε πραγματικά δεδομένα. Αρχικά αναλύθηκε το θεωρητικό υπόβαθρο, από όπου και έγινε εμφανές ότι τα μοντέλα μηχανικής μάθησης δεν είναι τίποτα άλλο παρά πολυσύνθετα μαθηματικά κατασκευάσματα. Επιπλέον δόθηκε ιδιαίτερη έμφαση στους στόχους πίσω από κάθε αλγόριθμο, όπως είναι η δυνατότητα πιθανολογικής ερμηνείας, η δυνατότητα γενίκευσης, η προσπάθεια αντιγραφής της ανθρώπινης σκέψης.

Με την εισαγωγή στο πρακτικό μέρος, παρουσιάστηκε ένας από τους πολυπληθείς κλάδους εφαρμογής της συγκεκριμένης επιστήμης. Η διαδικασία προετοιμασίας ανέδειξε την σημασία της «κλασικής», όχι πλήρως αυτοματοποιημένης ανάλυσης δεδομένων, καθώς ένα μεγάλο μέρος της λειτουργικότητας των μεθόδων βασίζεται στις κατάλληλες επιλογές και στην επαρκή κατανόηση τους.

Κατά την υλοποίηση των ταξινομητών αναδείχθηκε η σπουδαιότητά τους σαν εργαλεία, καθώς οδήγησαν πολύ γρήγορα σε αναγνώριση προτύπων, μη προφανή από τις διάφορες κατανομές των χαρακτηριστικών. Εξερευνήθηκε η συμπεριφορά για διαφοροποιήσεις στις υπέρ-παραμέτρους, στην αρχιτεκτονική καθώς και στο μεγάλο πλήθος επαναλήψεων που γίνονται εφικτές με τα σύγχρονα υπολογιστικά μέσα.

Ιδιαίτερη σημασία δόθηκε στην αποφυγή υπερεκπαίδευσης και στην ικανότητα των μοντέλων να γενικεύσουν σε δείγματα τα οποία δεν έχουν «συναντήσει» πρωτύτερα. Έγινε εμφανές ότι ένας μεγάλος περιορισμός έρχεται από τα ίδια τα δεδομένα, καθώς η μορφή τους καθορίζει άμεσα τη βέλτιστη επίδοση που μπορεί να επιτύχουν οι ταξινομητές.

Τέλος, σε μία προσέγγιση data mining, εξήχθησαν από τα μοντέλα, τα πιο ουσιώδη χαρακτηριστικά, είτε μεμονωμένα είτε σε σύνολα, με στόχο την ερμηνεία τους, αφού σε πρώτη προσέγγιση οι ταξινομητές είναι «μαύρα κουτιά».

Περαιτέρω μελέτη στο διαθέσιμο σύνολο δεδομένων, μπορεί να πραγματοποιηθεί μέσω της αποκρυπτογράφησης πιθανής κρυφής πληροφορίας στις άγνωστες τιμές. Επιπλέον θα μπορούσε να επεκταθεί η έρευνα σε δεδομένα από άλλες αντίστοιχες επιχειρήσεις, που αφορούν την ίδια χρονική περίοδο, για σύγκριση και αξιολόγηση των συμπερασμάτων.

Βιβλιογραφία

[1] Efraim Turban, Ramesh Sharda, and Dursun Delen. *Business Intelligence and Analytics: Systems for Decision Support*. English. PRENTICE HALL.

[2] Christopher M. Bishop. *Pattern recognition and machine learning*. English. Springer, 2016.

[3] F. Rosenblatt. "The perceptron: A probabilistic model for information storage and organization in the brain.". English. In: *Psychological Review* 65.6 (1958), pp. 386–408. ISSN: 0033-295X. DOI: 10.1037/h0042519. URL: http://dx.doi.org/10.1037/h0042519.

[4] S. Theodoridis and K. Koutroumbas. *Pattern recognition*. English. Academic Press, 1999.

[5] *Perceptron-unit*. English. URL: https://commons.wikimedia.org/wiki/File:Perceptron-unit.svg.

[6] R.H. Byrd, L. Peihuang, and J. Nocedal. "A limited-memory algorithm for bound-constrained optimization". English. In: *SIAM Journal on Scientific and Statistical Computing* (1996). DOI: 10.2172/204262.

[7] *artificial neural network*. English. URL: https://commons.wikimedia.org/wiki/File:Artificial_neural_network.svg.

[8] Leo Breiman. "Random Forests". English. In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/a:1010933404324.

[9] Tin Kam Ho. "Random decision forests". English. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition* (1995). DOI: 10.1109/icdar.1995.598994.

[10] Trevor Hastie, Jerome Friedman, and Robert Tisbshirani. *The elements of Statistical Learning: Data Mining, Inference, and prediction*. English. Springer, 2017.

[11] I. T. Jolliffe. *Principal component analysis. 2nd ed*. English. Springer-Verlag, 2002.

[12] *gaussianscatterpca*. English. URL: https://en.wikipedia.org/wiki/File:GaussianScatterPCA.svg.

[13] *overfitting*. English. URL: https://en.wikipedia.org/wiki/File:Overfitting.svg.

[14] Sérgio Moro, Paulo Cortez, and Paulo Rita. "A data-driven approach to predict the success of bank telemarketing". English. In: *Decision Support Systems* 62 (2014), pp. 22–31. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2014.03.001. URL: https://www.sciencedirect.com/science/article/pii/S016792361400061X.

[15] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. English. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[16] Thomas Kluyver et al. "Jupyter Notebooks – a publishing format for reproducible computational workflows". English. In: *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. Ed. by F. Loizides and B. Schmidt. IOS Press. 2016, pp. 87–90.

[17] Charles R. Harris et al. "Array programming with NumPy". English. In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[18] Wes McKinney. "Data Structures for Statistical Computing in Python". English. In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[19] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". English. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[20] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". English. In: *Journal of Machine Learning Research* 18.17 (2017), pp. 1–5. URL: http://jmlr.org/papers/v18/16-365.html.

[21] J. D. Hunter. "Matplotlib: A 2D graphics environment". English. In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[22] *Bank Marketing Data Set*. English. URL: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#.

[23] I.-K. Yeo. "A new family of power transformations to improve normality or symmetry". English. In: *Biometrika* 87.4 (2000), pp. 954–959. DOI: 10.1093/biomet/87.4.954.

[24] *Compare the effect of different scalers on data with outliers*. English. URL: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#sphx-glr-auto-examples-preprocessing-plot-all-scaling-py.

[25] Alan J. Thomas et al. "Two hidden layers are usually better than one". English. In: *Engineering Applications of Neural Networks* (2017), pp. 279–290. DOI: 10.1007/978-3-319-65172-9_24.

[26] Guang-Bin Huang and H.A. Babri. “Upper Bounds on the number of hidden neurons in feedforward networks with arbitrary bounded nonlinear activation functions”. English. In: *IEEE Transactions on Neural Networks* 9.1 (1998), pp. 224–229. DOI: 10.1109/72.655045.