Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών
και Φυσικών Επιστημών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

Πρόβλεψη ευστοχίας καλαθοσφαιριστών με χρήση τεχνικών Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΔΗΜΗΤΡΙΟΣ Γ. ΠΕΤΡΟΓΙΑΝΝΟΣ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Γεώργιος Αλεξανδρίδης
Εργαστηριακό και Διδακτικό Προσωπικό Ε.Μ.Π.

Αθήνα, Ιούλιος 2020

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών
και Φυσικών Επιστημών
Τομέας Τεχνολογίας Πληροφορικής και
Υπολογιστών

Πρόβλεψη ευστοχίας καλαθοσφαιριστών με χρήση τεχνικών Μηχανικής Μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΔΗΜΗΤΡΙΟΣ Γ. ΠΕΤΡΟΓΙΑΝΝΟΣ

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Συνεπιβλέπων : Γεώργιος Αλεξανδρίδης
Εργαστηριακό και Διδακτικό Προσωπικό Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 15η Ιουλίου 2020.

Ανδρέας Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Αναπληρωτής Καθηγητής Ε.Μ.Π.	Γεώργιος Αλεξανδρίδης Ε.ΔΙ.Π. Ε.Μ.Π.

Αθήνα, Ιούλιος 2020

Δημήτριος Γ. Πετρόγιαννος

Δίπλωμα Μεταπτυχιακών Σπουδών στην περιοχή της Επιστήμης Δεδομένων και Μηχανικής Μάθησης (Data Science and Machine Learning)

Περίληψη

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η επεξεργασία των δεδομένων που συλλέγονται κατά τη διάρκεια ενός αγώνα καλαθοσφαίρισης και η συνεπακόλουθη ανάλυσή τους μέσω στατιστικών τεχνικών και μεθόδων μηχανικής μάθησης.

Η βιομηχανία της ανάλυσης αθλητικών δεδομένων εκτιμάται στα 780 εκατομμύρια δολάρια και αναμένεται να ξεπεράσει τα 3 δισεκατομμύρια το 2024. Οι ομάδες δεν επιθυμούν μόνο να μεγιστοποιήσουν την απόδοση τους, αλλά ακόμη και τη φυσική κατάσταση των παικτών τους, τη βέλτιστη απόδοση στην αγορά παικτών, το μάρκετινγκ και την παρουσία τους στα μέσα κοινωνικής δικτύωσης.

Η ανάπτυξη της μηχανικής μάθησης μας επιτρέπει την επεξεργασία και εξόρυξη γνώσης από μεγάλο πλήθος δεδομένων και σε συνδυασμό με τη στατιστική αποκτούμε κατανόηση των παραγόντων που οδηγούν στην ατομική και ομαδική επιτυχία. Η εφαρμογή αναλυτικών εργαλείων έχει οδηγήσει στην ανάπτυξη σύγχρονων μεθόδων για την επεξεργασία μεγάλου όγκου δεδομένων με σκοπό την πρόβλεψη αποτελεσμάτων, ρεκόρ ομάδων μέσα στη σεζόν και τη συμπεριφορά των αντίπαλων ομάδων με στόχο την πρόβλεψη ενός αποτελέσματος στο μέλλον. Όσον αφορά την αναλυτική καλαθοσφαίρισης, εφαρμόζουμε μεθόδους παλινδρόμησης για την πρόβλεψη του αριθμού των νικών μιας ομάδας κατά τη διάρκεια μιας σεζόν και τεχνικές βελτιστοποίησης για τη βαθμολόγηση της απόδοσης μιας ομάδας. Επίσης εξετάζουμε τεχνικές μηχανικής μάθησης για την πρόβλεψη της έκβασης μιας επιθετικής προσπάθειας από έναν παίκτη που επιχειρεί να ευστοχήσει στο καλάθι της αντίπαλης ομάδας και για τη δυναμική πρόβλεψη νικητή.

Τα δεδομένα που χρησιμοποιήθηκαν στην πειραματική διαδικασία έχουν συλλεχθεί από το πρωτάθλημα του NBA, τα οποία έχουν εμπλουτιστεί και από άλλες πηγές. Παρότι η ακριβής πρόβλεψη του αποτελέσματος μιας επιθετικής προσπάθειας παραμένει ιδιαίτερα δύσκολη, στην παρούσα εργασία χρησιμοποιώντας σύγχρονα μοντέλα μηχανικής μάθησης και ανάλυσης δεδομένων καλαθοσφαίρισης θέτουμε ένα αρχικό σημείο περαιτέρω επεξεργασίας και πειραματισμού στο μέλλον.

Λέξεις κλειδιά

Αναλυτική Αθλητικών Δεδομένων, Παλινδρόμηση, Μηχανές Διανυσμάτων Υποστήριξης, Δένδρα Απόφασης, Τυχαία Δάση, Αναδρομικά Νευρωνικά Δίκτυα

Abstract

The objective of the current thesis is the processing of the data collecting during a basketball game and their consequent analysis though statistical techniques and machine learning methods.

The sport analytics industry is currently valued at $780 million and is expected to surpass $3 billion by 2024. Sport teams are not only interested in maximizing their performance, but also to assess their players' fitness, to make optimal decisions in player transfers, to enhance their marketing profile and to boost their presence in online social networks.

The development of relevant machine learning techniques allows us to handle and extract knowledge from an increased volume of data and along with the statistical analysis offers a deeper insight into the factors that lead to individual and team success. The application of analytical tools has lead to the development of novel methodologies for the processing of large volumes of data with the intent of predicting outcomes, team records within seasons and the behavior of opponent teams in order to determine the outcome of a feature match. In particular, for basketball analytics, we apply regression methods for predicting the number of team wins during a full season and optimization techniques for rating a team's performance. Additionally, we examine machine learning techniques for predicting shot outcome and dynamic winner prediction.

The data used in the experimental procedure have been collected from the NBA championship and they have been enriched from other sources. Even though the exact prediction of the outcome of an attack is still a difficult task, in the current thesis, using machine learning models and data analytics procedures, a starting point for further analysis and experimentation is set.

Key words

Game Analytics, Regression, Support Vector Machines, Decision Trees, Random Forests, Recurrent Neural Networks

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον Καθηγητή Ε.Μ.Π κ. Σταφυλοπάτη Ανδρέα-Γεώργιο, που μου έδωσε τη δυνατότητα να ασχοληθώ με το θέμα της αναλυτικής αθλητικών δεδομένων και πιο συγκεκριμένα, δεδομένων καλαθοσφαίρισης. Πρόκειται για έναν ερευνητικό τομέα που συνδυάζει τρία πολύ αγαπημένα αντικείμενα που απασχολούν την καθημερινότητα μου: το μπάσκετ, τη μηχανική μάθηση και την ποσοτική ανάλυση.

Στη συνέχεια, θα ήθελα να ευχαριστήσω τον Αναπληρωτή Καθηγητή ΕΜΠ, κ. Στάμου Γεώργιο, για την τιμή που μου έκανε να είναι μέλος της τριμελούς επιτροπής εξέτασης της παρούσας διπλωματικής εργασίας, καθώς και για τις συμβουλές του και την καθοδήγησή του κατά τη διάρκεια των μεταπτυχιακών μου σπουδών.

Επίσης, οφείλω ένα μεγάλο ευχαριστώ στο μέλος Εργαστηριακού και Διδακτικού Προσωπικού (Ε.ΔΙ.Π.) κ. Αλεξανδρίδη Γεώργιο, για την πολύτιμη βοήθεια και τις συμβουλές κατά την εκπόνηση της παρούσας διπλωματικής εργασίας, όπως επίσης και για τη συμμετοχή του στην τριμελή εξεταστική επιτροπή.

Τέλος, δε μπορώ να αγνοήσω τη βοήθεια της οικογένειας η οποία με υποστήριξε και έκανε δυνατή την απερίσπαστη ενασχόλησή μου με όλα τα πρότζεκτ που προέκυψαν τα τελευταία δύο χρόνια.

Η ολοκλήρωση της διπλωματικής εργασίας χρηματοδοτήθηκε από το ΙΚΥ στο πλαίσιο του «προγράμματος χορήγησης υποτροφιών για μεταπτυχιακές σπουδές πρώτου κύκλου (Master) στην Ελλάδα με ένταξη στην αγορά εργασίας, στο πλαίσιο συνεργασίας του Ιδρύματος Κρατικών Υποτροφιών (ΙΚΥ) και της Εθνικής Τράπεζας της Ελλάδος (ΕΤΕ), ακαδημαϊκού έτους 2018-2019».

Δημήτριος Γ. Πετρόγιαννος,

Αθήνα, 15η Ιουλίου 2020

Περιεχόμενα

Περίληψη 5

Abstract 7

Ευχαριστίες 9

Περιεχόμενα 11

Κατάλογος πινάκων 13

Κατάλογος σχημάτων 15

1. Εισαγωγή 17
1.1 Σκοπός της εργασίας 18

2. Αναλυτική Αθλητικών Δεδομένων 21
2.1 Το μοντέλο των τεσσάρων παραγόντων του Dean Oliver 21
2.2 Βαθμολογώντας τις ομάδες 25
2.3 Έλεγχος για σερί στα αποτελέσματα 30
2.4 Το σύστημα βαθμολόγησης Elo 33
2.4.1 Ο παράγοντας E 34
2.4.2 Ο παράγοντας K 34

3. Μηχανική Μάθηση 37
3.1 Επιβλεπόμενη μάθηση 37
3.2 Επιλογή χαρακτηριστικών 37
3.2.1 Η μέθοδος των πυρήνων για την εκτίμηση της συνάρτηση πυκνότητας πιθανότητας 38
3.3 Μετρικές αξιολόγησης 39
3.4 Διασταυρούμενη Επικύρωση 41
3.4.1 Έλεγχος Προσαρμογής 41
3.4.2 Διασταυρούμενη Επικύρωση Εξαίρεσης Ενός 42
3.4.3 Διασταυρούμενη Επικύρωση Εξαίρεσης k 42
3.4.4 Διασταυρούμενη Επικύρωση k μερών 42
3.5 Μοντέλα παλινδρόμησης 43
3.5.1 Γραμμική 43
3.5.2 Λογιστική 44
3.5.3 Ridge και Lasso παλινδρόμηση 45
3.6 Ταξινομητές 47
3.6.1 Λογιστική παλινδρόμηση 47

3.6.2 Support Vector Machines . . . 47
3.6.3 Δένδρα Απόφασης . . . 50
3.6.4 Τυχαία Δάση . . . 51
3.6.5 XGBoost . . . 51
3.6.6 Πολυεπίπεδα Perceptron . . . 54
3.6.7 Αναδρομικά Νευρωνικά Δίκτυα . . . 56

4. Πειραματική διαδικασία και Αποτελέσματα . . . 59
4.1 Το σύνολο των δεδομένων . . . 59
4.2 Βαθμολογία Draymond . . . 64
4.2.1 Επιλογή μεταβλητών . . . 66
4.3 Επιλογή χαρακτηριστικών . . . 68
4.4 Αποτελέσματα ταξινόμησης . . . 68
4.4.1 Λογιστική παλινδρόμηση . . . 68
4.4.2 Τυχαία Δάση . . . 69
4.4.3 Μηχανές Διανυσμάτων Υποστήριξης . . . 70
4.4.4 XGBoost . . . 71
4.4.5 Πολυεπίπεδα Perceptron . . . 71
4.4.6 Αναδρομικά νευρωνικά δίκτυα . . . 72
4.5 Δυναμική πρόβλεψη νικητή . . . 73
4.5.1 Εφαρμογή ταξινομητών . . . 74

5. Συμπεράσματα και μελλοντικές κατευθύνσεις . . . 79
5.1 Πρόβλεψη ευστοχίας . . . 79
5.2 Δυναμική πρόβλεψη νικητή . . . 80

Βιβλιογραφία . . . 81

Παράρτημα 85

Α. Ευρετήριο Ακρωνυμίων και Συντμήσεων . . . 85

Κατάλογος πινάκων

2.1 Προσαρμογή του μοντέλου των 4 παραγόντων για τη σεζόν 2018-19 23
2.2 Προσαρμογή του μοντέλου των 4 παραγόντων με Lasso παλινδρόμηση για τη σεζόν 2017-2018 24
2.3 Προσαρμογή του μοντέλου των 4 παραγόντων με Lasso παλινδρόμηση και το κανόνα της μιας τυπικής απόκλισης τη σεζόν 2017-2018 24
2.4 Αποτελέσματα της βελτιστοποίησης για τη σεζόν 2018-2019 27

3.1 Πίνακας Σύγχυσης 40

4.1 Βέλτιστες τιμές για τις υπερπαραμέτρους του ταξινομητή τυχαίων δασών . 70
4.2 Βέλτιστες τιμές για τις υπερπαραμέτρους του ταξινομητή SVM 70
4.3 Βέλτιστες τιμές για τις υπερπαραμέτρους του ταξινομητή XGBoost 71
4.4 Βέλτιστες τιμές για τις υπερπαραμέτρους του ταξινομητή τυχαίων δασών . 76
4.5 Βέλτιστες τιμές για τις υπερπαραμέτρους του ταξινομητή μηχανών διανυσμάτων υποστήριξης 76
4.6 Βέλτιστες τιμές για τις υπερπαραμέτρους του ταξινομητή XGBoost 77

Κατάλογος σχημάτων

1.1 Ανάπτηξη αναλυτικής αθλητικών δεδομένων τα προσεχή έτη 17

2.1 Πίνακας συσχέτισης των μεταβλητών του μοντέλου για το έτος 2018-2019 23
2.2 Η τιμή των συντελεστών για κάθε τιμή του βαθμού ομαλοποίησης 24
2.3 Επιθετική και Αμυντική Αποτελεσματικότητα των ομάδων για το έτος 2018-2019 25
2.4 Ιστόγραμμα των υπολοίπων από τα αποτελέσματα της βελτιστοποίησης . . 28
2.5 Ιστόγραμμα των υπολοίπων από τις βαθμολογίες με ξεχωριστές εγγραφές για εντός και εκτός έδρας 29
2.6 Η μέθοδος των πυρήνων για τα υπόλοιπα 29
2.7 Βαθμολογίες για απόδοση εντός και εκτός έδρας 30
2.8 Αμυντική και επιθετική βαθμολογία ομάδων για τη σεζόν 2018-2019 31

3.1 Παράδειγμα εφαρμογής της μεθόδου των πυρήνων 39
3.2 Παράδειγμα ταξινομητή με μηδενική ακρίβεια 40
3.3 Παράδειγμα τέλειου ταξινομητή 41
3.4 Παράδειγμα τυχαίου ταξινομητή 41
3.5 Η σιγμοειδής συνάρτηση 45
3.6 Σχηματική αναπαράσταση του προβλήματος ελαχιστοποίησης της Ridge και της Lasso παλινδρόμησης στο $\mathbb{R}^2$ 47
3.7 Γραφική απεικόνιση του προβλήματος βελτιστοποίησης των SVM 48
3.8 Ένα παράδειγμα SVM με τη μέθοδο των πυρήνων 49
3.9 Ένα παράδειγμα ενός δένδρου απόφασης 50
3.10 Σχηματικά η εκπαίδευση δύο συλλογικών δένδρων απόφασης 52
3.11 Ένα παράδειγμα σε ένα δέντρο με 3 φύλλα 52
3.12 Οι δύο κύριες στρατηγικές ανάπτυξης των δέντρων 54
3.13 Απο το Perceptron σε πιο βαθιές αρχιτεκτονικές 55
3.14 Σχηματική αναπαράσταση της σύγκλισης με και χωρίς ορμή 56
3.15 Ισοδυναμία μεταξύ αναδρομικών και κλασικών νευρωνικών δικτύων 57
3.16 Σχηματική απεικόνιση ενός μοντέλου LSTM 58

4.1 Οι διαφορετικές ζώνες 60
4.2 Οι διαφορετικές διευθύνσεις επίθεσης 60
4.3 Οι ζώνες που δημιουργήθηκαν στο πλαίσιο της διπλωματικής 61
4.4 Συχνότητα εκδήλωσης επίθεσης ανά ζώνη 61
4.5 Ποσοστό ευστοχίας ανά ζώνη 61
4.6 Ποσοστό ευστοχίας ανά διεύθυνση 62
4.7 Εκτιμώμενος αριθμός πόντων ανά ζώνη 62
4.8 Χρόνος εκτέλεσης της επιθετικής προσπάθειας 63
4.9 Γραφήματα των επιθέσεων 3 ομάδων 63

4.10 Συνολικά ποσοστά ευστοχίας 64
4.11 Μέσος αριθμός αμυντικών προσπαθειών ανά θέση (2013-14) 65
4.12 Μέσος αριθμός αμυντικών προσπαθειών ανά θέση για τη σεζόν 2014-15 . . 65
4.13 Μέση βαθμολογία ανά θέση (2013-14) 66
4.14 Μέση βαθμολογία ανά θέση (2014-15) 66
4.15 Θηκόγραμμα προσαρμοσμένης βαθμολογίας ανά θέση (2013-14) 67
4.16 Θηκόγραμμα προσαρμοσμένης βαθμολογίας ανά θέση (2014-15) 67
4.17 Παλινδρόμηση Lasso για όλες τις διαθέσιμες μεταβλητές 68
4.18 z_score για έλεγχο σερί στις προσπάθειες ανά θέση 68
4.19 Ποσοστό διακύμανσης των δεδομένων που περιέχεται στα k πρώτα χαρακτηριστικά της PCA 69
4.20 Αποτελέσματα για τη λογιστική παλινδρόμηση 69
4.21 Αποτελέσματα για τον ταξινομητή τυχαίων δασών 70
4.22 Πίνακας σύγχυσης για τα SVM 71
4.23 Αποτελέσματα για τον ταξινομητή XGBoost 72
4.24 Αποτελέσματα για το πολυεπίπεδο Perceptron 72
4.25 Διαγραμματική απεικόνιση του τελικού μοντέλου του αναδρομικού νευρωνικού δικτύου 73
4.26 Παλινδρόμηση Lasso για όλες τις υποψήφιες μεταβλητές 75
4.27 Αποτελέσματα πρόβλεψης νικητή για τη λογιστική ταξινόμηση 75
4.28 Αποτελέσματα πρόβλεψης νικητή για τον ταξινομητή τυχαίων δασών . . . 76
4.29 Πίνακας σύγχυσης για το SVM 77
4.30 Αποτελέσματα πρόβλεψης νικητή για τον ταξινομητή XGBoost 77
4.31 Αποτελέσματα πρόβλεψης νικητή για το πολυεπίπεδο Perceptron 78

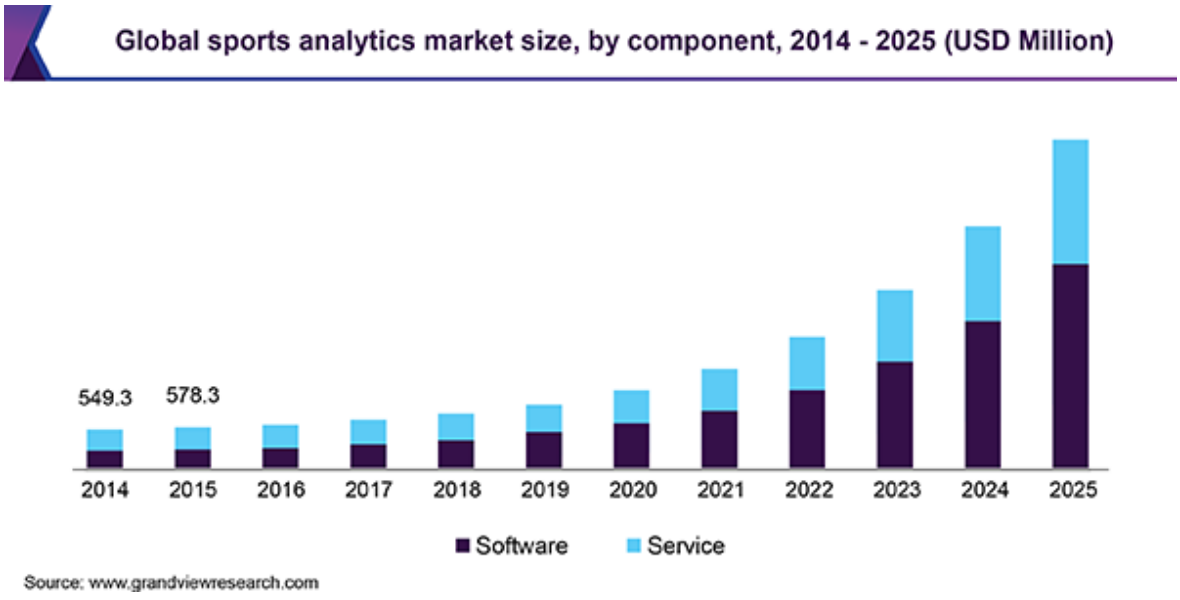
5.1 Διάγραμμα απόστασης από κοντινότερο αμυντικό ανά ζώνη. 79

Κεφάλαιο 1

Εισαγωγή

Μετά το πέρας ενός αθλητικού γεγονότος πληθώρα στατιστικών δεδομένων καταγράφονται και παρουσιάζονται στο κοινό. Οι πιο απλές αναλύσεις μπορεί να περιλαμβάνουν την καταγραφή των βασικών στοιχείων που οδήγησαν στην έκβαση του αγώνα, που σε συνδυασμό με κάποια στιγμιότυπα και κάποια γραφήματα, δίνουν μια απτή εικόνα του αποτελέσματος στο θεατή. Αν επιθυμούμε μια πιο λεπτομερή ανάλυση, απαιτείται να μην περιοριστούμε μόνο σε μια γραφική αναπαράσταση, αλλά να καταφύγουμε σε μια αναλυτική επεξεργασία των στατιστικών, ώστε να αποκτήσουν νόημα και να μας βοηθήσουν στην βαθύτερη κατανόηση του αθλήματος και στην πρόβλεψη του μέλλοντος.

Ο κλάδος της *αναλυτικής αθλητικών δεδομένων* (sports analytics) [Alam13], δηλαδή η συλλογή και επεξεργασία δεδομένων με σκοπό την εξαγωγή γνώσης προς όφελος των ομάδων, αποτελεί έναν ταχύτατα αναπτυσσόμενο κλάδο, όπως φαίνεται και στο Σχήμα 1.1.



Σχήμα 1.1: Εκτιμήσεις για την ανάπτυξη της αγοράς της αναλυτικής αθλητικών δεδομένων τα προσεχή έτη (Πηγή: https://www.gradientviewresearch.com)

Πλέον όλες οι ομάδες που αγωνίζονται στην *Εθνική Καλαθοσφαιρική Ομοσπονδία των Ηνωμένων Πολιτειών* (National Basketball Association - NBA) περιλαμβάνουν τόσο στο τεχνικό επιτελείο όσο και στα τμήματα στρατηγικής και ανάπτυξης άτομα που χρησιμοποιούν αναλυτικές μεθόδους με σκοπό τη βελτίωση της απόδοσης της ομάδας [Troi16]. Η βελτίωση αφορά όλες τις πτυχές και όλους τους παράγοντες που απασχολούν την ομάδα. Ειδικότερα, οι ομάδες επιθυμούν να προλαμβάνουν τραυματισμούς, να βελτιστοποιούν την απόδοση των συνθέσεων τους τόσο σε ομαδικό όσο και σε ατομικό επίπεδο καθώς και τις τακτικές των

αντιπάλων που, αν τις περιορίσουν, μπορούν στα μεταξύ τους παιχνίδια να φτάσουν στη νίκη. Πολύ σημαντικό είναι ακόμη όχι μόνο να εντοπίσουν τους καλύτερους παίκτες των αντίπαλων ομάδων αλλά και τους μελλοντικούς παίκτες που η προσθήκη στο ρόστερ τους θα εκτοξεύσει την παρουσία τους στις διοργανώσεις με το μικρότερο δυνατό κόστος.

Η πρώτη γνωστή απόπειρα χρήσης στατιστικών στοιχείων για τη βελτιστοποίηση της απόδοσης μιας ομάδας έγινε από τον Billy Beane, παράγοντα της ομάδας baseball, Oakland Athletics. Αν και βρισκόταν σε μια ομάδα μέτριας δυναμικότητας, κατάφερε να επιλέγει και να προσαρμόζει τη στρατηγική της ομάδας διαλέγοντας παίκτες με βάση όχι τα χαρακτηριστικά που αφορούν τη σωματοδομή τους, αλλά το τρόπο με τον οποίο αγωνίζονται. Η συλλογή δεδομένων και η επεξεργασία τους τον οδήγησε στο να βελτιώσει σημαντικά την απόδοση της ομάδας και να γίνει γνωστός διεθνώς. Η ιστορία του έγινε βιβλίο το 2003 με τίτλο *Moneyball* [Lewi04] και ταινία το 2011. Η έκδοση του βιβλίου και η δημοφιλία που απέκτησε οδήγησε στην περαιτέρω ανάπτυξη του κλάδου της αναλυτικής αθλητικών δεδομένων.

Το πρωτάθλημα του NBA αποτελεί τη σημαντικότερη διοργάνωση καλαθοσφαίρισης παγκοσμίως, καταφέρνοντας να προσελκύει τους καλύτερους παίκτες παγκοσμίως και κάθε χρόνο να βελτιώνεται το θέαμα. Το πιο γνωστό σύστημα συλλογής δεδομένων προς επεξεργασία αποτελεί το πρόγραμμα εφαρμογής της SportsVU [Tami08]. Πρόκειται για 6 κάμερες οι οποίες 25 φορές ανά δευτερόλεπτο καταγράφουν τις κινήσεις και μετρήσεις των παικτών. Προκύπτει η ανάγκη να βρεθεί ένας αποτελεσματικός τρόπος διαχείρισης των δεδομένων ώστε να εξαχθεί η περισσότερη δυνατή γνώση. Με τα αποτελέσματα που προκύπτουν από την επεξεργασία των δεδομένων το προπονητικό τμήμα είναι σε θέση να εντοπίζει τα δυνατά και τα αδύνατα σημεία τόσο της ομάδας τους όσο και της αντίπαλης. Συνολικά, η ανάπτυξη και βελτίωση των μεθόδων μηχανικής μάθησης βοηθά: (i) στη γρήγορη συλλογή δεδομένων, μεγάλου όγκου και πολυπλοκότητας, (ii) στη βέλτιστη ανάλυση και ταχύτατη λήψη απόφασης ή εντοπισμού των αδυναμιών των ομάδων.

1.1 Σκοπός της εργασίας

Παρακολουθώντας έναν αγώνα καλαθοσφαίρισης θα δούμε πως οι ομάδες προσπαθούν μέσα από την καλή κυκλοφορία της μπάλας να οδηγηθούν στην καλύτερη δυνατή επιθετική προσπάθεια. Αυτό σημαίνει πως επιθυμούν ο πιο εύστοχος παίκτης της ομάδας να σουτάρει υπό τις καλύτερες δυνατές συνθήκες, ώστε η ομάδα να σκοράρει και να φτάσει συνολικά στο μεγαλύτερο αριθμό πόντων σε σύγκριση με την αντίπαλη ομάδα. Στην παρούσα διπλωματική εργασία θα προσπαθήσουμε, κυρίως, να εξηγήσουμε ποιοι είναι αυτοί οι παράγοντες που επηρεάζουν την ευστοχία και στη συνέχεια να προβλέψουμε την έκβαση της προσπάθειας ενός παίκτη πάνω σε ένα πραγματικό σύνολο δεδομένων.

Το πρώτο στοιχείο που μας βοηθάει να προβλέψουμε το αποτέλεσμα είναι το ποσοστό ευστοχίας του επιτιθέμενου παίκτη. Ωστόσο τα ποσοστά ευστοχίας ποικίλουν ανά ζώνη, οπότε πρέπει καταρχήν να μπορούν να εντοπιστούν οι ζώνες αυτές. Επίσης, η ευστοχία εξαρτάται και από άλλους παράγοντες, όπως λ.χ. το πόσο κοντά ή μακρυά από τη μπασκέτα είναι ο παίκτης κλπ. Για να μην καταλήξουμε σε ένα μοντέλο που στηρίζεται απλά στα ποσοστά ευστοχίας, θα θέλαμε να προσομοιώσουμε ένα πιο ρεαλιστικό μοντέλο ως προς την πρόβλεψη της ευστοχίας. Συνεπώς, στο σύνολο δεδομένων θα θέλαμε να υπάρχει και η πληροφορία για την απόσταση από τον κοντινότερο αμυντικό. Στην πορεία της ανάλυσης θα προκύψει η σημασία της και ο βαθμός που επηρεάζει τον επιτιθέμενο. Τέλος, πέρα από την εξαγωγή γνώσης για την ανάλυση αμυντικών και επιθετικών χαρακτηριστικών, θα προσπαθήσουμε να εξάγουμε μετρικές για την αποτελεσματικότητα της αμυντικής συμπεριφοράς των παικτών.

Στο Κεφάλαιο 2 θα προσπαθήσουμε να χρησιμοποιήσουμε ήδη γνωστά μοντέλα, είτε θα τροποποιήσουμε κάποια, με σκοπό την εξαγωγή γνώσης από το σύνολο δεδομένων με τα χαρακτηριστικά από τις επιθετικές προσπάθειες. Αρχικά, θα παρουσιάσουμε το γνωστό *μοντέλο των 4 παραγόντων* του Dean Oliver [Kuba07b]. Πρόκειται για ένα ιδιαίτερα ενδιαφέρον μοντέλο μιας και παρουσιάζει τους παράγοντες που περιγράφουν με μεγάλη ακρίβεια τον αριθμό των νικών που θα κάνει μια ομάδα κατά τη διάρκεια μιας σεζόν, με τις επεξηγηματικές μεταβλητές να έχουν φυσική σημασία. Στη συνέχεια, θα παρουσιάσουμε ένα τρόπο βελτιστοποίησης με τον οποίο θα εκτιμήσουμε την επιθετική αποτελεσματικότητα, χωρίς να επηρεάζονται τα αποτελέσματα από το πλεονέκτημα έδρας. Θα δείξουμε πως τα υπόλοιπα της διαδικασίας από τη βελτιστοποίηση ακολουθούν κανονική κατανομή. Τέλος, θα αναλύσουμε τη βαθμολογία Elo [Elo78], η οποία αποτελεί μια δυναμική διαδικασία αποτύπωσης της δυναμικότητας των ομάδων και θα μας βοηθήσει στη δυναμική πρόβλεψη του νικητή. Θα αναλυθούν διεξοδικά όλες οι παράγοντες που χρησιμοποιούνται στη βαθμολογία Elo καθώς και η χρησιμότητα τους.

Τα εργαλεία που θα χρησιμοποιηθούν περιγράφονται στο Κεφάλαιο 3 και εντάσσονται στον κλάδο της στατιστικής και της μηχανικής μάθησης. Για την επιλογή των μεταβλητών χρησιμοποιούνται οι τεχνικές παλινδρόμησης Lasso [Tibs96] και Ridge [Hoer70]. Πρόκειται για μεθόδους γραμμικής παλινδρόμησης στις οποίες προστίθενται όροι ομαλοποίησης και οι οποίες χρησιμοποιούνται σημαντικά σε ερευνητικό επίπεδο. Η προσθήκη όρων ομαλοποίησης στη μηχανική μάθηση δεν περιορίζεται μόνο στην επιλογή μεταβλητών αλλά μπορεί να βοηθήσει στη βελτίωση όλων των μεθόδων μάθησης, μειώνοντας το σφάλμα γενίκευσης.

Σε θεωρητικό και πρακτικό επίπεδο, θα χρησιμοποιήσουμε τις πιο κλασικές μεθόδους και θα αναλύσουμε μια από τις πιο σύγχρονες και ευρέως χρησιμοποιούμενες τεχνικές μηχανικής μάθησης, τη μέθοδο *extreme gradient boosting* ή XGBoost [Frie01]. Πρόκειται για μια μέθοδο ταξινόμησης που στηρίζεται στη μέθοδο *συλλογικής* (ensemble) εκπαίδευσης δέντρων, σε συνδυασμό με μεθόδους *ενίσχυσης της κλίσης* (gradient boosting) [Frie02]. Πιο συγκεκριμένα, εκπαιδεύονται συνεχόμενα k δέντρα απόφασης, όπου σε κάθε επανάληψη μειώνεται το συνολικό σφάλμα και η τελική απόφαση λαμβάνεται από όλα τα k εκπαιδευμένα δέντρα. Η μέθοδος δεν εξαρτάται από τη συνάρτηση ωφέλειας που καλείται να βελτιστοποιήσει και εμφανίζει καλά αποτελέσματα στα προβλήματα που πρόκειται να τη χρησιμοποιήσουμε. Ήδη στην πλατφόρμα του Kaggle[1], η οποία φιλοξενεί διαγωνισμούς μηχανικής μάθησης καθώς και συλλογές δεδομένων, αποτελεί τη μεθοδολογία που δίνει τα καλύτερα αποτελέσματα στα προβλήματα που εξετάζουμε στην παρούσα διπλωματική εργασία, είτε όπως την περιγράφεται από τους συγγραφείς στο αρχικό paper [Frie01] ή με κάποιες τροποποιήσεις.

Το κυριότερο ζήτημα που εγείρεται κατά τη μελέτη της πρόβλεψης ευστοχίας είναι κατά πόσο επηρεάζεται η έκβαση του αποτελέσματος από τις προηγούμενες εκβάσεις των προσπαθειών του παίκτη. Στη βιβλιογραφία, αυτό το ζήτημα είναι γνωστό και ως *the hot hand hypothesis* [Tver89]. Αρχικά, θα χρησιμοποιήσουμε ένα πολύ γνωστό στατιστικό έλεγχο, τον έλεγχο Wald–Wolfowitz για σερί στα αποτελέσματα [Wald43], ώστε να αποκτήσουμε μια διαίσθηση για το βαθμό στον οποίο επηρεάζει το παρελθόν την πιθανότητα ευστοχίας.

Τέλος, θα παρουσιαστούν όλες οι μέθοδοι που χρησιμοποιήθηκαν καθώς και το εύρος των υπερπαραμέτρων που εξετάστηκαν προς βέλτιστη ακρίβεια στο αποτέλεσμα των εκάστοτε ταξινομητών. Για να εξετάσουμε την αποτελεσματικότητα της επίθεσης χρησιμοποιώντας τα αποτελέσματα από το παρελθόν, θα χρησιμοποιήσουμε τα αναδρομικά νευρωνικά δίκτυα και θα κάνουμε προτάσεις για βελτίωση. Παράλληλα, έχοντας αναλύσει σε βάθος το σύνολο δεδομένων μπορούμε να οδηγηθούμε στη δυναμική πρόβλεψη νικητή. Δυναμική, υπό την έννοια πως για τη χρονική στιγμή που θέλουμε να προβλέψουμε το νικητή σε ένα ζευγάρι

[1] https://www.kaggle.com/

αντίπαλων ομάδων θα χρησιμοποιήσουμε μόνο τη στατιστική τους απόδοση στα προηγούμενα παιχνίδια.

Κεφάλαιο 2

Αναλυτική Αθλητικών Δεδομένων

2.1 Το μοντέλο των τεσσάρων παραγόντων του Dean Oliver

Οι ομάδες προσπαθούν μέσω της βέλτιστης απόδοσης τόσο στην άμυνα όσο και στην επίθεση να φτάσουν στο μεγαλύτερο πλήθος νικών που μπορούν να πετύχουν. Μετά το τέλος οποιοδήποτε αγώνα μπορούμε εύκολα να έχουμε πρόσβαση σε ένα μεγάλο πλήθος στατιστικών που μας περιγράφουν αναλυτικά τα όσα συνέβησαν κατά τη διάρκεια του αγώνα μεταξύ των δύο ομάδων. Μπορούμε να δούμε μια έστω και απλή στατιστική εικόνα των επιδόσεων των ομάδων. Το ζητούμενο ωστόσο είναι να προσδιορίσουμε τα κυριότερα χαρακτηριστικά με τα οποία μπορούμε να περιγράψουμε την αγωνιστική ταυτότητα της ομάδας και παράλληλα να εκτιμήσουμε αποτελεσματικά το πλήθος των νικών που μπορεί να επιτύχει μέσα στην αγωνιστική περίοδο. Κατά συνέπεια επιθυμούμε να εντοπίσουμε τους παράγοντες εκείνους που οδηγούν σε υψηλή πιθανότητα νίκης σε οποιοδήποτε παιχνίδι.

Την απάντηση μπορεί να μας δώσει το *μοντέλο των τεσσάρων παραγόντων* (four factor model) του Dean Oliver [Kuba07b]. Ο Dean Oliver, σύμβουλος διάφορων ομάδων στο NBA, κατάφερε με τέσσερα στατιστικά νούμερα που αφορούν ξεχωριστά την άμυνα και την επίθεση να προβλέψει, μέσω απλής γραμμικής παλινδρόμησης, τον αριθμό των νικών μιας ομάδας.

Οι τέσσερις παράγοντες που αφορούν την επίθεση είναι οι εξής:

1. **Η αποτελεσματικότητα της ευστοχίας** (Effective Field Goal percentage - EFG). Η μαθηματική παράσταση που περιγράφει το στατιστικό αυτό δίνεται στην Εξίσωση 2.1 παρακάτω:

$$\frac{\text{FG's made} + 0.5 * \text{3PTS made}}{\text{FG's attempted}} \tag{2.1}$$

   Θυμίζουμε πως field goal είναι οποιαδήποτε προσπάθεια σε σουτ πλην των βολών. Αυτό το στατιστικό επιβραβεύει τις ομάδες που έχουν καλύτερο ποσοστό τριπόντων προσθέτοντας μισή μονάδα στον αριθμητή για κάθε εύστοχο τρίποντο. Θεωρούμε για παράδειγμα ότι δυο διαφορετικές ομάδες A, B έχουν συνολικά 1000 προσπάθειες σε σουτ με τις 500 να είναι εύστοχες. Ειδικότερα,η ομάδα A έχει 350 δίποντα και 150 τρίποντα εύστοχα και η ομάδα B έχει 400 δίποντα και 100 τρίποντα εύστοχα.

   Σύμφωνα με τον παραπάνω τύπο η ομάδα A έχει αποτελεσματικότητα ευστοχίας ίση με

$$\frac{500 + 0.5 * 150}{1000} = 57.5\%$$

   ενώ η ομάδα B αποτελεσματικότητα ευστοχίας ίση με

$$\frac{500 + 0.5 * 100}{1000} = 55\%$$

2. **Αριθμός λαθών ανά κατοχή** (Turnovers Committed per Possession - TPP). Ως κατοχή ορίζεται η πρώτη στιγμή που η ομάδα αποκτά την μπάλα και η αντίπαλη ομάδα βρίσκεται σε θέση άμυνας. Η κατοχή διακόπτεται όταν η μπάλα περάσει στην αντίπαλο. Ως λάθος ορίζεται η απώλεια της κατοχής χωρίς να έχει προηγηθεί επιθετική προσπάθεια προς το αντίπαλο καλάθι. Αν η ομάδα επιτεθεί και χάσει τη διεκδίκηση rebound τότε δε προσμετράται στα λάθη. Αν όμως μια λάθος πάσα ενός παίκτη της ομάδας οδηγήσει τη μπάλα εκτός των ορίων του γηπέδου τότε η κατοχή αλλάζει και το λάθος καταγράφεται στο στατιστικό που περιγράψαμε.

3. **Ποσοστό επιθετικών ριμπάουντ** (Offensive Rebounding Percentage - ORP). Αφορά τον αριθμό των άστοχων επιθέσεων από τις οποίες η ομάδα ανακτά ξανά την κατοχή μέσω της διεκδίκησης του rebound και έχει την ευκαιρία να ξαναεπιτεθεί.

4. **Ποσοστό επιθέσεων που οδηγούν σε ελεύθερες βολές** (Free Throw Rate - FTR). Περιγράφει τη συχνότητα με την οποία η ομάδα εκτελεί βολές. Οι βολές στο σύγχρονο μπάσκετ είναι πολύ σημαντικές αφού δίνουν εύκολα πόντους και προκαλούν και φθορά στην αντίπαλη ομάδα.

Ομοίως, οι τέσσερις παράγοντες που αφορούν την άμυνα της ομάδας είναι οι αντίστοιχες που αφορούν τα στατιστικά των ομάδων που αντιμετωπίζει ως αντιπάλους.

1. **Αποτελεσματικότητα ευστοχίας του αντιπάλου** (Opponent's Effective Field Goal Percentage - OEFG)

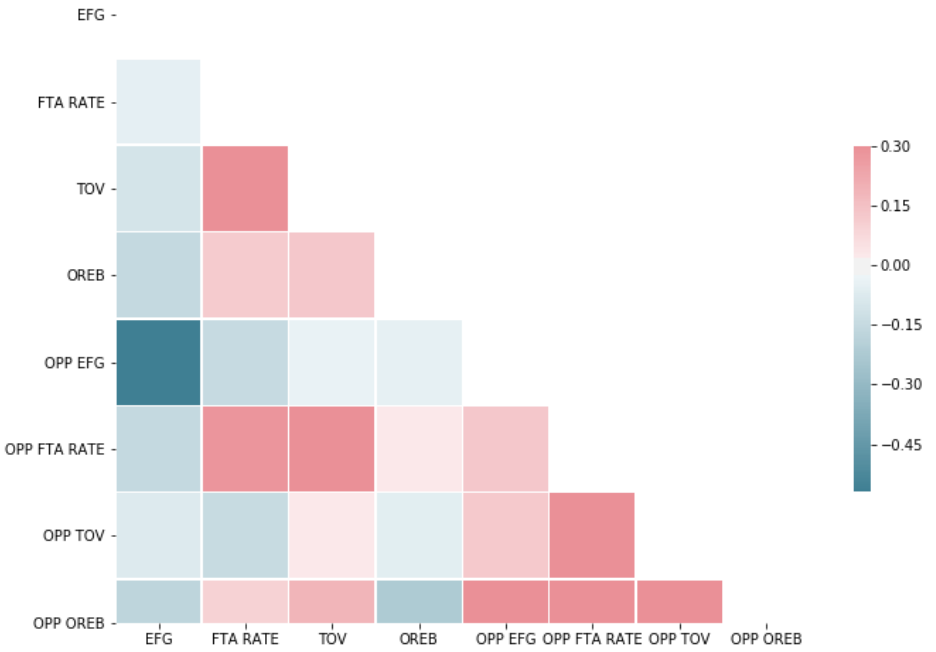
2. **Αριθμός λαθών ανά κατοχή αντιπάλου** (Defensive Turnovers Caused per Possession - DTPP)

3. **Ποσοστό αμυντικών ριμπάουντ** (Defensive Rebounding Percentage - DRP)

4. **Ποσοστό ελεύθερων βολών αντιπάλου** (Opponent's Free Throw Rate - OFTR)

Ουσιαστικά, περιγράφουν πόσο εύκολα η ομάδα μας καταφέρνει η ίδια αλλά ταυτόχρονα αποτρέπει την αντίπαλη από το: (i) να σκοράρει, (ii) να μην πραγματοποιεί λάθη, (iii) να ανανεώνει τις κατοχές μετά από άστοχες προσπάθειες, (iv) να πηγαίνει σε ελεύθερες βολές Ένα πρώτο χαρακτηριστικό είναι πως οι 8 αυτοί παράγοντες φαίνονται ασυσχέτιστοι (Σχήμα 2.1). Η μικρότερη τιμή στον πίνακα συσχέτισης είναι η τιμή -0.45 μεταξύ της αποτελεσματικής ευστοχίας της ομάδας και του αντιπάλου, που δηλώνει πως όσο περισσότερη εύστοχη είναι μια ομάδα κατά μέσο όρο, τόσο λιγότερη αποτελεσματική είναι η αντίπαλη ομάδα. Προσαρμόζοντας το μοντέλο παλινδρόμησης βλέπουμε πως σχεδόν όλες οι συντελεστές έχουν το πρόσημο που αναμένεται εκτός του συντελεστή του ποσοστού των βολών της αντίπαλης ομάδας.

Στο Σχήμα 2.2 και στον Πίνακα 2.1 φαίνονται τα αποτελέσματα απο την παλινδρόμηση Lasso και οι τιμές της απλής παλινδρόμησης στις 8 μεταβλητές του μοντέλου. Παρατηρούμε πως μόνο τα στατιστικά που αφορούν τις βολές μηδενίζονται πρώτα. Όσο πιο καλή είναι μια ομάδα επιθετικά τόσο πιο καλά είναι τα στατιστικά νούμερα της και οι συντελεστές είναι θετικοί, οδηγώντας σε αναμενόμενη αύξηση του πλήθους των νικών. Η αύξηση των λαθών οδηγεί σε μειωμένες πιθανότητες νίκης εξού και ο αρνητικός συντελεστής. Αντίστοιχη ερμηνεία επιδέχονται και οι τέσσερις συντελεστές που αφορούν τα στατιστικά της αντίπαλης ομάδας. Το μόνο προβληματικό σημείο είναι ο θετικός συντελεστής στο ποσοστό των βολών με αποτέλεσμα να χρειάζεται να τρέξουμε το μοντέλο της παλινδρόμησης για παλιότερες χρονιές και να παρατηρήσουμε τα πρόσημα των συντελεστών.



Σχήμα 2.1: Πίνακας συσχέτισης των μεταβλητών του μοντέλου για το έτος 2018-2019

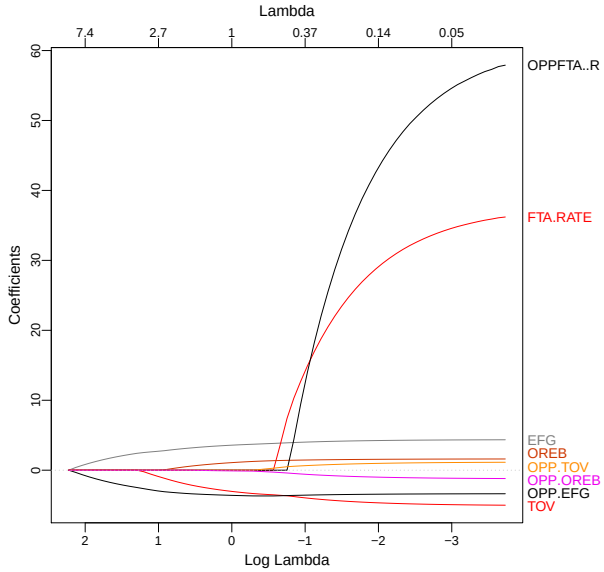
Πίνακας 2.1: Προσαρμογή του μοντέλου των 4 παραγόντων για τη σεζόν 2018-19

Team	**Coef**	**p_value**
EFG	6.1	0.1
FTA.RATE	4.4	$\leq 10^{-8}$
TOV	+37.8	0.17
OREB	-5.1	$\leq 10^{-5}$
OPP.EFG	+1.6	$\leq 10^{-3}$
OPP.FTA.RATE	+61.3	0.14
OPP.TOV	+1.17	0.1
OPP.OREB	-1.23	0.035
R2	0.95	-
F_statistic	49	0

Πολλοί από τους συντελεστές μπορεί να είναι και περιττοί. Πολλές από τις p_{value} προκύπτουν μεγαλύτερες από 5%. Χρειάζεται να ελέγξουμε αν με την αφαίρεση κάποιων μεταβλητών μπορούμε να καταλήξουμε σε ένα απλούστερο μοντέλο, αλλά με παρόμοια προβλεπτική ικανότητα. Θα χρησιμοποιήσουμε τα εργαλεία των ομαλοποιημένων μορφών της παλινδρόμησης.

Προσπαθώντας να ελαχιστοποιήσουμε το μέσο τετραγωνικό σφάλμα παρατηρούμε πως οι συντελεστές παραμένουν μη μηδενικοί (Πίνακας 2.2). Ακόμη, και στον κανόνα της μιας τυπικής απόκλισης οι συντελεστές παραμένουν διάφοροι του μηδενός (Πίνακας 2.3).

Ενδιαφέρον παρουσιάζεται στον κανόνα της μιας τυπικής απόκλισης όπου οι συντελεστές που περιγράφουν όμοια στατιστικά αλλά για διαφορετικές ομάδες πλησιάζουν κατά απόλυτη τιμή στο ίδιο σχεδόν νούμερο. Προσαρμόζοντας το μοντέλο για τη σεζόν 2017-2018 εντοπίζουμε πως ο συντελεστής είναι αρνητικός όπως είναι λογικό (Πίνακες 2.3-2.2)



Σχήμα 2.2: Η τιμή των συντελεστών για κάθε τιμή του βαθμού ομαλοποίησης

Πίνακας 2.2: Προσαρμογή του μοντέλου των 4 παραγόντων με Lasso παλινδρόμηση για τη σεζόν 2017-2018

4factor statistic	Coef
EFG	4
FTA.RATE	9.5
TOV	-5
OREB	+1.9
OPP.EFG	-3.5
OPP.FTA.RATE	-12,7
OPP.TOV	+2.56
OPP.OREB	-0.9

Πίνακας 2.3: Προσαρμογή του μοντέλου των 4 παραγόντων με Lasso παλινδρόμηση και το κανόνα της μιας τυπικής απόκλισης τη σεζόν 2017-2018

4factor statistic	Coef
EFG	4
FTA.RATE	14.9
TOV	-4
OREB	+1.5
OPP.EFG	-3.5
OPP.FTA.RATE	-14
OPP.TOV	+0.66
OPP.OREB	-0.6

2.2 Βαθμολογώντας τις ομάδες

Αρχικά, μπορούμε να παρατηρήσουμε πως η γηπεδούχος ομάδα έχει περισσότερες πιθανότητες νίκης σε σχέση με τη φιλοξενούμενη. Άρα το γεγονός πως μια ομάδα είναι γηπεδούχος της δίνει και κάποιο πλεονέκτημα που θα επιθυμούσαμε να το εκτιμήσουμε.

Ένας εύκολος τρόπος να εκτιμήσουμε την αποτελεσματικότητα των ομάδων σε άμυνα και επίθεση είναι να τις συνδέσουμε με τον αριθμό των κατοχών μπάλας. Εάν γνωρίζουμε για μια ομάδα τους πόντους που ευστόχησε και δέχτηκε καθώς και το πλήθος των κατοχών που παρατηρήθηκαν για τα δύο αυτά κύρια χαρακτηριστικά τότε οι σχέσεις των Εξισώσεων 2.2-2.3

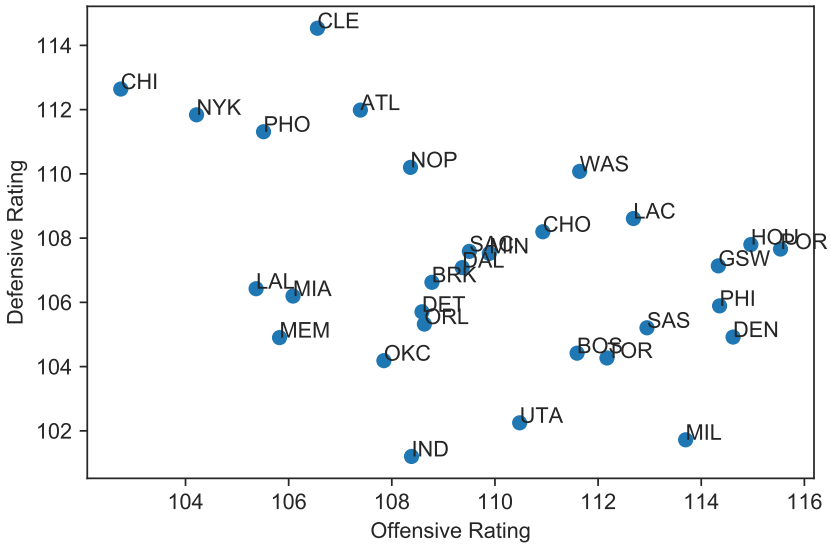
$$\text{Offensive Rating} = 100 \cdot \frac{\text{Points Scored}}{\text{Possesions made}} \tag{2.2}$$

$$\text{Defensive Rating} = 100 \cdot \frac{\text{Points Conceded}}{\text{Possesions Conceded}} \tag{2.3}$$

μας δίνουν μια άμεση εικόνα των πόσων πόντων μια ομάδα κατά μέσο όρο σκοράρει και δέχεται ανά 100 κατοχές.

Ας φανταστούμε ένα γράφημα όπου παρουσιάζονται για κάθε ομάδα τα παραπάνω νούμερα με την επιθετική αποτελεσματικότητα να βρίσκεται τον x-άξονα και την αμυντική αποτελεσματικότητα στον y-άξονα. Τότε ομάδες με μεγάλη επιθετική αποτελεσματικότητα, δηλαδή υψηλό μέσο αριθμό πόντων ανά 100 κατοχές, βρίσκονται όλο και δεξιότερα στο γράφημα. Ενώ ομάδες που έχουν βασικό γνώρισμα την αποτελεσματική άμυνα έχουν όλο και μικρότερο δείκτη αμυντικής αποτελεσματικότητας μιας και δέχονται λιγότερους πόντους ανά κατοχή σε σχέση με τις υπόλοιπες ομάδες οπότε θα βρίσκονται όλο και χαμηλότερα στο γράφημα.

Στη συνέχεια, παρουσιάζεται το γράφημα της αμυντικής και της επιθετικής αποτελεσματικότητας για όλες τις ομάδες του NBA κατά την αγωνιστική χρονιά 2018-2019 (Σχήμα 2.3).



Σχήμα 2.3: Επιθετική και Αμυντική Αποτελεσματικότητα των ομάδων για το έτος 2018-2019

Στο Σχήμα 2.3 παρατηρούμε πως οριακά οι Portland Blazers έχουν την καλύτερη επιθετική αποτελεσματικότητα, με άλλες 4 ομάδες να βρίσκονται πολύ κοντά τους. Αντίστοιχα,

οι Indiana Pacers φαίνεται να έχουν την καλύτερη με διαφορά άμυνα και κοντά τους να βρίσκονται άλλες δύο ομάδες, οι οποίες όμως έχουν αρκετά καλύτερη επίθεση.

Αν και οι παραπάνω φόρμουλες είναι εύκολες τόσο στον υπολογισμό όσο και στην ερμηνεία δε μας βοηθούν στην εκτίμηση του πλεονεκτήματος έδρας. Για να βρούμε τη λύση πιο εύκολα χρειάζεται να σκεφτούμε για μια στιγμή αντίστροφα. Σε κάθε αγώνα δύο ομάδες θα έχουν περίπου τον ίδιο αριθμό κατοχών και θα προσπαθήσουν μέσα από τη βελτιστοποίηση της άμυνας και της επίθεσης να φτάσουν στη νίκη ή, ισοδύναμα, σε περισσότερους εύστοχους πόντους σε σχέση με την αντίπαλη ομάδα. Αν θεωρήσουμε ως δεδομένο το πλεονέκτημα έδρας ως home advantage τότε θα θέλαμε η ποσότητα

$$\textbf{home advantage+home team rating-away team rating}$$

να είναι πάρα πολύ κοντά στη παρατηρούμενη διαφορά πόντων μεταξύ της γηπεδούχου αλλά και της φιλοξενούμενης ομάδας. Δηλαδή

$$\textbf{home team score-away team score} \approx$$
$$\textbf{(home advantage+home team rating-away team rating)}$$

Αυτό επιτυγχάνεται με την ελαχιστοποίηση της ποσότητας

$$\underset{\textbf{team ratings}}{\textbf{minimize}} \sum_{i=1,..,\,\textbf{total games}} (\textbf{home team score-away team score} - \textbf{(home advantage+home team rating-away team rating)})^2$$

υπό τον περιορισμό

$$\sum_i \textbf{team ratings(i)=0}$$

δηλαδή το άθροισμα όλων των βαθμολογιών των ομάδων να είναι μηδέν.

Για την επίλυση του προβλήματος εμείς χρείαζεται να καλέσουμε μια ρουτίνα που υλοποιεί μια επαναληπτική μέθοδο για την εύρεση ελαχίστου της συνάρτησης που θέλουμε να ελαχιστοποιήσουμε υπό περιορισμούς. Ως γνωστόν, επειδή οι περιορισμοί αφορούν ισότητες η επαναληπτική μέθοδος στηρίζεται στη μεθοδολογία Karush–Kuhn–Tucker [Kuhn51].

Η λύση αυτού του προβλήματος μας επιστρέφει βαθμολογίες οι οποίες δεν επηρεάζονται από το πλεονέκτημα έδρας και έχουν πλέον φυσική ερμηνεία. Για παράδειγμα, αν η γηπεδούχος ομάδα Los Angeles Lakers έχει βαθμολογία +2 ενώ η φιλοξενούμενη ομάδα Denver Nuggets -1 με το πλεονέκτημα έδρας στο +3 τότε τα αποτελέσματα μας έχουν την εξής ερμηνεία:

1. Οι Los Angeles Lakers σκοράρουν κατά μέσο όρο 2 πόντους παραπάνω σε σχέση με τη μέση ομάδα.

2. Οι Denver Nuggets σκοράρουν κατά μέσο όρο -1 πόντους παραπάνω σε σχέση με τη μέση ομάδα.

3. Η γηπεδούχος ομάδα έχει πλεονέκτημα 3 πόντων.

4. Εάν οι δύο παραπάνω ομάδες έρθουν αντιμέτοποι με τους Lakers ως γηπεδούχους τότε η εκτιμώμενη διαφορά πόντων είναι $+3+2-(-1)=6$ πόντους. Ενώ με τους Nuggets ως γηπεδούχους η εκτιμώμενη διαφορά πόντων είναι $+3+(-1)-(2)=0$ πόντους.

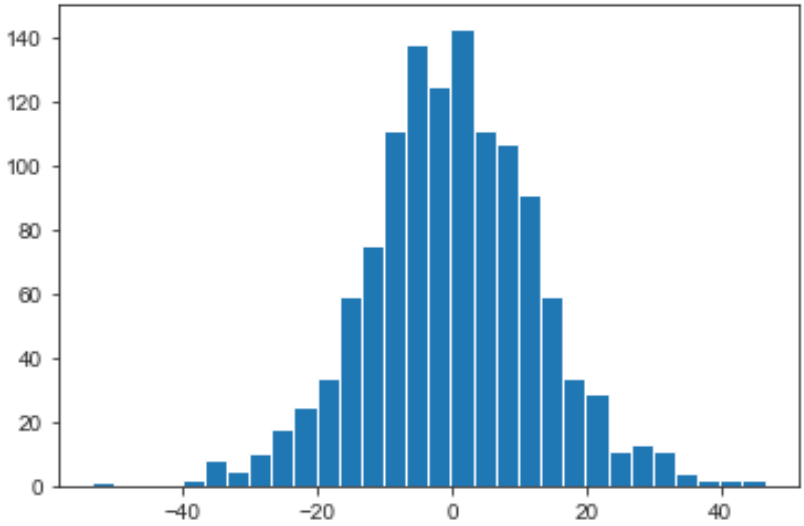
Στη συνέχεια παρουσιάζονται τα αποτελέσματα της βελτιστοποίησης για τη σεζόν 2018-2019 (Πίνακας 2.4). Το πλεονέκτημα έδρας είναι 2.7 πόντοι υπέρ της γηπεδούχου. Τη μεγαλύτερη βαθμολογία την έχουν οι Milwaukee Bucks (8) και ακολουθούν οι Golden State Wariors με 6.4. Στη τρίτη θέση ακολουθούν οι πρωταθλητές για τη σεζόν 2018-2019 Toronto Raptors με 5.5. Οι Indiana Pacers που ξεχώρισαν για την αποτελεσματική τους άμυνα βρίσκονται 2.8 βαθμούς πάνω από τη μέση ομάδα.

Πίνακας 2.4: Αποτελέσματα της βελτιστοποίησης για τη σεζόν 2018-2019

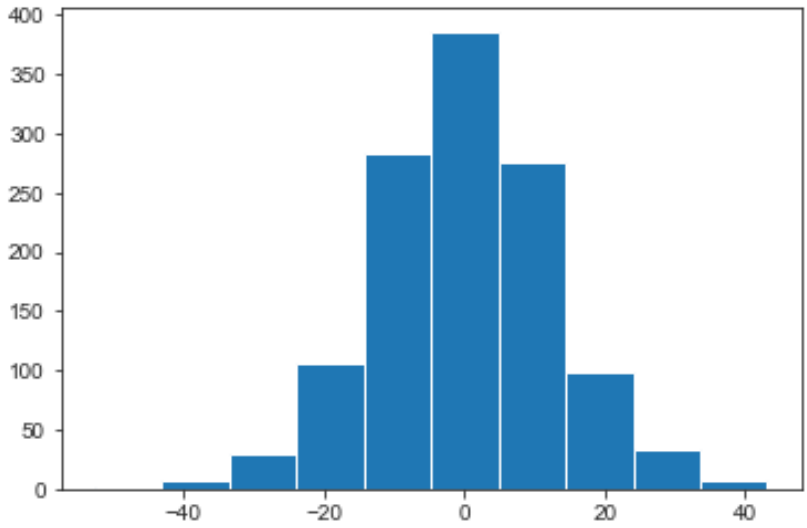
Team	**Rating**
Home edge points	2.7
ATL	-6.1
BOS	3.9
BRK	-0.4
CHI	-8.3
CHO	-1.3
CLE	-9.4
DAL	-0.9
DEN	4.2
DET	-0.6
GSW	6.4
HOU	5.0
IND	2.8
LAC	1.1
LAL	-1.3
MEM	-2.1
MIA	-0.5
MIL	8.0
MIN	-1.0
NOP	-1.1
NYK	-8.9
OKC	3.6
ORL	0.3
PHI	2.3
PHO	-8.6
POR	4.4
SAC	-0.8
SAS	1.8
TOR	5.5
UTA	5.3
WAS	-3.3

Προκύπτει το ερώτημα τι συμβαίνει όμως με τα υπόλοιπα που δημιουργούνται από τη βελτιστοποίηση των βαθμολογιών και τι πληροφορία μπορούμε να εξάγουμε. Όπως είδαμε νωρίτερα στα παραδείγματα, οι βαθμολογίες που προκύπτουν μας δίνουν μια μικρή πρόβλεψη-εκτίμηση για το νικητή του αγώνα. Εμείς θα δημιουργήσουμε ένα διάνυσμα όπου σε κάθε αγώνα υπολογίζουμε τη διαφορά *home edge* + *home team rating* - *away team rating*. Εάν αυτή η διαφορά είναι θετική, δηλαδή προβλέπουμε πως ο γηπεδούχος θα νικήσει τότε αφαιρούμε απο την προηγούμενη διαφορά τη διαφορά που προέκυψε (*home team points scored*

- *away team points scored*). Σε αντίθετη, περίπτωση πολλαπλασιάζουμε με -1 το προηγούμενο νούμερο. Αν αναπαραστήσουμε σε ένα γράφημα (Σχήμα 2.4) το αποτέλεσμα που προκύπτει είναι πως τα υπόλοιπα όπως τα ορίσαμε ακολουθούν την κανονική κατανομή με μέση τιμή και τυπική απόκλιση ίση με τη δειγματική διασπορά (12.8 πόντους). Εάν θέλουμε να μειώσουμε τη διασπορά θα προσπαθήσουμε να ελαχιστοποιήσουμε στο σύνολο διαχωρίζοντας την απόδοση των ομάδων σε εντός και εκτός έδρας. Το αποτέλεσμα που προκύπτει παρουσιάζεται στο Σχήμα 2.5, χωρίς να μειώνεται σημαντικά η διασπορά.

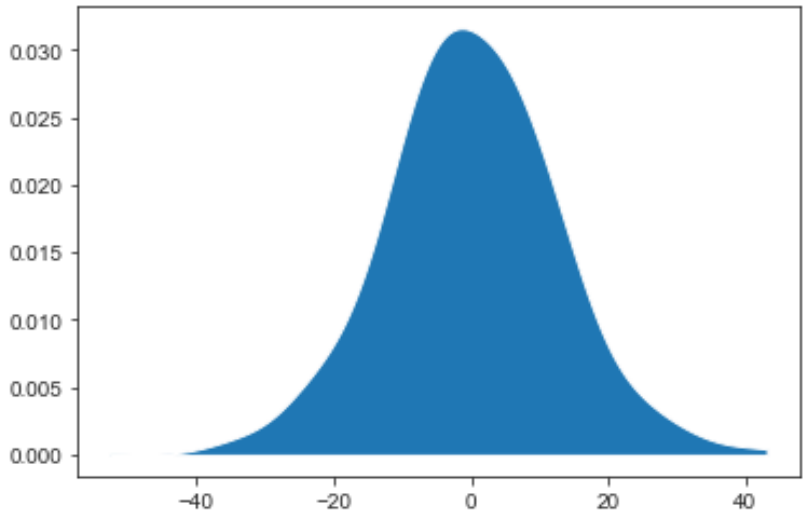


Σχήμα 2.4: Ιστόγραμμα των υπολοίπων από τα αποτελέσματα της βελτιστοποίησης



Σχήμα 2.5: Ιστόγραμμα των υπολοίπων από τις βαθμολογίες με ξεχωριστές εγγραφές για εντός και εκτός έδρας

Για την εύρεση του βέλτιστου bandwidth εφαρμόζουμε τη μέθοδο των πυρήνων με cv leave one out και επιλέγουμε εκείνο που μεγιστοποιεί την πιθανοφάνεια.



Σχήμα 2.6: Η μέθοδος των πυρήνων για τα υπόλοιπα

Για να μπορέσουμε να βαθμολογήσουμε τις ομάδες και στην άμυνα χρειάζεται να εκφράσουμε κατάλληλα τη συνάρτηση και τους συντελεστές ώστε να έχουν νόημα. Και πάλι χρειάζεται να σκεφτούμε αντίστροφα. Οι ομάδες κερδίζουν πόντους και από την άμυνα της αντίπαλης ομάδας. Αυτό, θα γίνει σε κάθε εγγραφή προσθέτοντας τους χαρακτήρες Η ή Α. Θα χρειαστούμε να τα εκφράσουμε όλα σε σχέση με το μέσο όρο πόντων κάθε ομάδας. Η γηπεδούχος ομάδα θα σκοράρει

leage average +0.5*home edge+home offence rating+away defence rating

ενώ η φιλοξενούμενη ομάδα θα σκοράρει

leage average -0.5*home edge+away offence rating+home defence rating

Για να δούμε γιατί αυτό είναι σωστό θα παραθέσουμε ένα παράδειγμα

Σχήμα 2.7: Βαθμολογίες για απόδοση εντός και εκτός έδρας

Team	Offence Rating	Defence Rating
Golden State Warriors	+9	-2
Dallas Mavericks	+3	+2

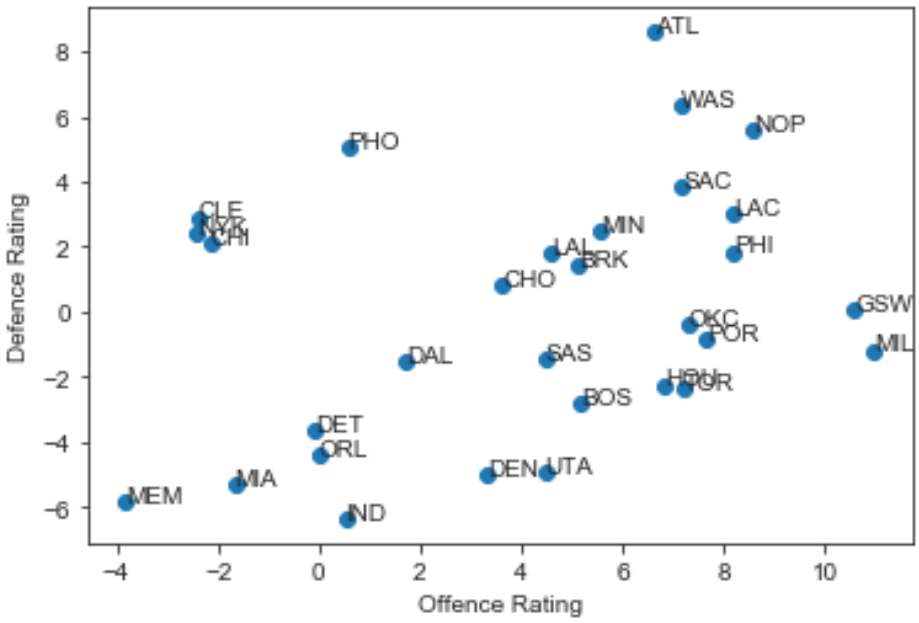
και ο μέσος όρος πόντων στους οποίους ευστοχεί η ομάδα είναι 100 και το πλεονέκτημα έδρας 3 πόντοι. Τότε, αν το παιχνίδι πραγματοποιείται στην έδρα των Golden State Warriors: (i) η γηπεδούχος θα σκοράρει $100 + 0.5 * 3 + 9 + 2 = 112.5$ και (ii) η φιλοξενούμενη ομάδα θα σκοράρει $100 - 0.5 * 3 + 3 + (-2) = 99.5$, με την εκτιμώμενη διαφορά να είναι 13 πόντοι. Παρατηρούμε πως η αρνητική βαθμολογία σε αντίθεση με τα προηγούμενα παραδείγματα δείχνει πως η ομάδα έχει καλύτερη αμυντική συμπεριφορά σε σχέση με το μέσο όρο.
Η συνάρτηση που θέλουμε να ελαχιστοποιήσουμε τώρα είναι η :

$$\underset{\text{team ratings}}{\textbf{minimize}} \sum_{i=1,..,\ \text{total games}} (\textbf{home team score-league average-}$$
$$\textbf{0.5}\cdot\textbf{home advantage-home offence team rating-away team defence rating})^2$$
$$+(\textbf{home team score} - \textbf{league average+0.5*home edge}$$
$$-\textbf{home offence team rating-away team defence rating})^2$$

με τα αποτελέσματα να παρουσιάζονται στο Σχήμα 2.8

2.3 Έλεγχος για σερί στα αποτελέσματα

Παρακολουθώντας αγώνες καλαθοσφαίρισης μπορεί να εντοπίσουμε παίχτες να επιτυγχάνουν διαδοχικά καλάθια και να πιστεύουμε πως βρίσκονται σε βέλτιστη κατάσταση. Σε μεγαλύτερα χρονικά διάστημα, οι ομάδες καταγράφουν μεγάλα επιτυχημένα ή αποτυχημένα αποτελέσματα και στα αθλητικά μέσα ενημέρωσης διαβάζουμε πως η απόδοση της ομάδας παρουσιάζει συστηματική άνοδο ή πτώση αντίστοιχα. Θα θέλαμε να βρούμε ένα τρόπο να ελέγξουμε εάν πραγματικά οι ακολουθίες που περιγράφουν σε διαδοχικές στιγμές την έκβαση των προσπαθειών σε ατομικό η ομαδικό επίπεδο εξαρτώνται από το ταλέντο-ικανότητα ή περισσότερο σε τύχη.



Σχήμα 2.8: Αμυντική και επιθετική βαθμολογία ομάδων για τη σεζόν 2018-2019

Αρχικά, χρειάζεται να ορίσουμε τι είναι μια συνεχόμενη ακολουθία αποτελεσμάτων (*run*). Έστω πως έχουμε 10 παρατηρήσεις από τις προσπάθειες ενός παίκτη σε 10 σουτ π.χ WLWLWWLLWW, όπου με W συμβολίζεται η ευστοχία και με L η αστοχία. Για να μετρήσουμε τα runs αρχίζουμε από το πρώτο στοιχείο, από το οποίο ξεκινά μια ακολουθία διαδοχικών αποτελεσμάτων, την οποία και καταγράφουμε. Στη συνέχεια, πηγαίνουμε στο επόμενο στοιχείο. Αν το επόμενο είναι ίδιο τότε ο μετρητής για τα runs δεν αυξάνεται. Σε αντίθετη περίπτωση, προσθέτουμε ακόμη μια παρατήρηση στο μετρητή των runs. Η διαδικασία αυτή συνεχίζεται μέχρι να εξαντλήσουμε όλα τα στοιχεία της ακολουθίας. Στο παράδειγμα μας μπορούμε να μετρήσουμε 7 διαδοχικές ακολουθίες συνεχόμενων ίδιων αριθμών runs.

Ας υποθέσουμε πως ένας παίκτης σε 10 προσπάθειες είχε 5 εύστοχα και 5 άστοχα σουτ. Από τη θεωρία των διακριτών μαθηματικών υπάρχουν $\binom{10}{5}$=252 διαφορετικοί τρόποι να συμβεί αυτό. Οι ακολουθίες με τις περισσότερες εναλλαγές εκβάσεων είναι η WLWLWLWLWL και η LWLWLWLWLW, όπου και στις δύο περιπτώσεις μια θετική έκβαση μιας προσπάθειας ακολουθείται από μια αρνητική έκβαση και το αντίστροφο. Σε μια τέτοια περίπτωση μπορούμε να ισχυριστούμε πως η ευστοχία του παίκτη φαίνεται να εξαρτάται από τις ικανότητές του και όχι από κάποιο είδος τυχαιότητας λόγω ορμής από προηγούμενα θετικά ή αρνητικά αποτελέσματα.

Αντίστοιχα, οι ακολουθίες με τις λιγότερες εναλλαγές διαφορετικών εκβάσεων είναι η WWWWW-LLLLL και η LLLLLWWWWW. Εδώ, ο παίκτης ξεκινά με 5 συνεχόμενες εύστοχες ή άστοχες αντίστοιχα και στη συνέχεια καταλήγει σε 5 συνεχόμενες αντίθετες εκβάσεις σε σχέση με τις αρχικές του προσπάθειες. Στην περίπτωση μας, η μοναδική εναλλαγή εκβάσεων στην ακολουθία αποτελεί ένδειξη πως το συνολικό αποτέλεσμα οφείλεται περισσότερο σε τύχη και λόγω ορμής απο τα προηγούμενα αποτελέσματα μέσα στην ακολουθία παρά στην ικανότητα του.

Έστω ότι σε μία ακολουθία μήκους έχουμε S θετικές εκβάσεις και F αποτυχημένες εκβάσεις. Προφανώς $N = S + F$, με την Εξίσωση 2.4 να απεικονίζει τον αναμενόμενο

αριθμό διαδοχικών αποτελεσμάτων και την Εξίσωση 2.5 την τυπική τους απόκλιση

$$\mu = \frac{2FS}{N} + 1 \tag{2.4}$$

$$\sigma = \sqrt{\frac{(\mu-1)(\mu-2)}{N-1}} \tag{2.5}$$

Αν αντικαταστήσουμε στις παραπάνω σχέσεις τα νούμερα του αρχικού παραδείγματος προκύπτει $\mu = 6$ και $\sigma = 1.49$.

Αφού μπορούμε να εκφράσουμε τα πάντα πλέον με πιθανότητες μπορούμε να καταλήξουμε σε κάποιο έλεγχο υποθέσεων. Αναζητώντας στη βιβλιογραφία αυτός ο έλεγχος είναι γνωστός ως *Wald Wolfowitz runs test* [Wald43]. Στον απλό έλεγχο για μια ακολουθία εξετάζουμε τη μηδενική υπόθεση

Τα αποτελέσματα στην ακολουθία με 0 και 1 ως παρατηρήσεις κατανέμονται τυχαία

Η μηδενική υπόθεση απορρίπτεται όταν οι διαδοχικές ακολουθίες είναι πολύ λίγες ή πάρα πολλές ως προς το πλήθος. Στην περίπτωση μας, το αποτέλεσμα του ελέγχου μπορεί να ερμηνευτεί ως εξής:

Μηδενική υπόθεση: Οι αποτυχημένες και οι επιτυχημένες εκβάσεις είναι τυχαία κατανεμημένες στην παρατηρούμενη ακολουθία. Δηλαδή, οι προηγούμενες εκβάσεις δε φαίνεται να επηρεάζουν το αποτέλεσμα των επόμενων προσπαθειών και η ικανότητα του παίκτη προσδιορίζει τα αποτελέσματα συνολικά
Εναλλακτική υπόθεση: Οι αποτυχημένες και οι επιτυχημένες εκβάσεις δεν είναι τυχαία κατανεμημένες στην παρατηρούμενη ακολουθία. Οι προηγούμενες εκβάσεις βοηθούν στην πρόβλεψη των μελλοντικών εκβάσεων.

Για να μπορέσουμε να πραγματοποιήσουμε έλεγχο υποθέσεων θα δημιουργήσουμε τις κανονικοποιημένες παρατηρήσεις γνωστές και ως z_{score} (Εξίσωση 2.6)

$$z_{score} = \frac{observed - \mu}{\sigma} \tag{2.6}$$

όπου σε κάθε παρατήρηση αφαιρούμε την αναμενόμενη μέση τιμή και διαιρούμε με τη τυπική απόκλιση. Αυτές οι κανονικοποιημένες μεταβλητές ακολουθούν στο όριο την κανονική κατανομή με μέση τιμή 0 και διασπορά 1.

Στο πρώτο παράδειγμα με την ακολουθία με τις περισσότερες εναλλαγές το πλήθος των διαδοχικών ίδιων αποτελεσμάτων είναι 10 και ο έλεγχος σύμφωνα με τα προηγούμενα είναι

$$z_{score} = \frac{10-6}{1.49} = 2.68$$

ενώ στο δεύτερο παράδειγμα με μόνο δύο συνεχόμενες διαδοχικές ακολουθίες αποτελεσμάτων ο έλεγχος είναι

$$z_{score} = \frac{2-6}{1.49} = -2.68$$

Για να ελέγξουμε την υπόθεση σε επίπεδο σημαντικότητας $(1-\alpha)\%$ θα χρησιμοποιήσουμε τα z_{score} και θα πραγματοποιήσουμε αμφίπλευρο έλεγχο. Στην περίπτωση που η παρατηρούμενη τιμή βρίσκεται ανάμεσα στα όρια

$$-z_{-\alpha/2} \leq z_{score} \leq z_{\alpha/2}$$

,όπου με $z_{-\alpha/2}$ συμβολίζουμε το ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής που στην περιοχή (-∞,$z_{-\alpha/2}$] το εμβαδόν κάτω από την καμπύλη είναι $\alpha/2$ ενώ αντίστοιχα με $z_{\alpha/2}$ συμβολίζουμε το ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής που στην περιοχή [$z_{\alpha/2}$,∞) το εμβαδόν κάτω από την καμπύλη είναι $\alpha/2$, τότε μπορούμε να απορρίψουμε τη μηδενική υπόθεση με βεβαιότητα α% και να αποδεχτούμε την εναλλακτική υπόθεση. Αντίθετα, αν η τιμή του z_{score} δεν είναι στο διάστημα [-$z_{-\alpha/2}$,$z_{-\alpha/2}$] τότε δε μπορούμε να απορρίψουμε τη μηδενική υπόθεση.

2.4 Το σύστημα βαθμολόγησης Elo

Ο καθηγητής Φυσικής και λάτρης του σκακιού, Arpad Elo, δημιούργησε ένα σύστημα αξιολόγησης των παικτών του σκακιού [Elo78] που αποδεικνύεται ιδιαίτερα δημοφιλές σχεδόν σε όλα τα αθλήματα. Ας φανταστούμε μια διοργάνωση σε ένα οποιοδήποτε άθλημα όπου συμμετέχουν 4 διαγωνιζόμενοι. Εμείς δεν έχουμε καμία πληροφορία και θα θέλαμε να δημιουργήσουμε έναν αλγόριθμο γρήγορο και αποτελεσματικό, υπό την έννοια πως παρατηρώντας τα αποτελέσματα μετά την λήξη ενός αγώνα να μπορούμε να ανανεώσουμε τις βαθμολογίες των διαγωνιζόμενων κατά τέτοιο τρόπο ώστε να αντανακλούν τη δυναμική τους. Αρχικά, όλοι οι διαγωνιζόμενοι ξεκινούν με την ίδια βαθμολογία. Μια πρώτη λογική απαίτηση στο σύστημα αξιολόγησης μας είναι πως εάν ο παίκτης που αναδεικνύεται σε έναν αγώνα νικητής κερδίζει κ πόντους στο σύστημα αξιολόγησης τότε ο αντίπαλος, αυτομάτως, χάνει κ πόντους.Θέλουμε να είναι ένα παιχνίδι μηδενικού αθροίσματος. Θα παρουσιάσουμε ένα μικρό παράδειγμα για να εξηγήσουμε αναλυτικότερα τις επιλογές στον αλγόριθμο μας.

Match	Score	Elo Change	Team1	Team2	Team3	Team4
-	-	-	1300	1300	1300	1300
Team1 vs Team2	100-90	15	1315	1285	1200	1200
Team3 vs Team4	99-96	5	1315	1285	1305	1295
Team1 vs Team3	103-87	20	1315	1285	1200	1200
Team4 vs Team2	93-95	1	1315	1285	1200	1200

Για να ανανεώσουμε τις βαθμολογίες των διαγωνιζόμενων μετά από κάθε αγωνιστική θα χρησιμοποιήσουμε βεβαίως τις παλιές βαθμολογίες και ένα κομμάτι αύξησης ή μείωσης ανάλογα με την έκβαση του αγώνα για τον παίκτη μας (Εξίσωση 2.7)

$$R_{i+1} = K \cdot (S_{team} - E_{team}) + R_i \tag{2.7}$$

όπου οι μεταβλητές που εμφανίζονται στην Εξίσωση 2.7 έχουν την ακόλουθη σημασία

1. S_{team}: είναι η μεταβλητή που περιγράφει την έκβαση του αγώνα που προηγήθηκε για την ομάδα μας. Συνήθεις επιλογές είναι 0 για την ήττα, 1 για τη νίκη και 0.5 για την ισοπαλία. Φυσικά, για αθλήματα όπου δεν υπάρχει δυνατότητα ισοπαλίας ,όπως η καλαθοσφαίριση, επιλέγεται η δυαδική αναπαράσταση των αποτελεσμάτων.

2. E_{team}: είναι η μεταβλητή που περιγράφει την εκτιμώμενη πιθανότητα νίκης για την ομάδα που μελετάμε. Είναι συνάρτηση των βαθμολογιών των ομάδων που κοντράρονται τη χρονική στιγμή i και έχει συνήθως τη μορφή μιας συνάρτησης κατανομής.

3. K: είναι ο παράγοντας που καθορίζει την ευαισθησία στις μεταβολές των βαθμολογιών. Δεν είναι απαραίτητο να υπολογίζει τη παρατηρούμενη διαφορά στα σκορ των αγώνων και τις περισσότερες φορές είναι μια σταθερά.

2.4.1 Ο παράγοντας E

Όπως περιγράψαμε νωρίτερα ο όρος αυτός θα εκφράζει την πιθανότητα νίκης της ομάδας που μελετάμε και θα είναι συνάρτηση των παλιών βαθμολογιών του ζευγαριού που αγωνίζεται. Μπορεί να είναι οποιαδήποτε μορφή συνάρτησης κατανομής. Ως γνωστόν μια συνάρτηση κατανομής έχει τις παρακάτω ιδιότητες: (i) είναι αύξουσα συνάρτηση και (ii) έχει πεδίο τιμών στο $[0, 1]$ Μια τρίτη ιδιότητα που θα θέλαμε να ικανοποιεί ο παράγοντας E είναι να έχει πεδίο ορισμού όλο το R γιατί θέλουμε να ικανοποιείται η παρακάτω σχέση (Εξίσωση 2.8)

$$E(-x) = 1 - E(x) \tag{2.8}$$

Στη πράξη χρησιμοποιούνται κυρίως η κανονική ή η λογιστική κατανομή. Εμείς θα εξετάσουμε τη δεύτερη (Εξίσωση 2.9)

$$\frac{1}{1 + e^{\frac{x-\mu}{\sigma}}} \tag{2.9}$$

όπου στη θέση του $x - \mu$ θα εισάγουμε τη διαφορά δυναμικότητας των ομάδων σε σχέση με τις παλιές βαθμολογίες, ενώ θα θεωρήσουμε το σ ως παράγοντα κανονικοποίησης, που συνήθως εκφράζει τη τυπική απόκλιση του συστήματος βαθμολόγησης Όλο και μεγαλύτερες τιμές έχουν ως αποτέλεσμα οι δυνατές ομάδες να επιβραβεύονται όχι μόνο συχνότερα αλλά και περισσότερο.

Η πιο σημαντική παρατήρηση αφορά τη συνήθη μετατροπή του παράγοντα κανονικοποίησης (Εξίσωση 2.10)

$$\sigma = \frac{n}{\ln(10)} \tag{2.10}$$

όπου n οποιοσδήποτε ακέραιος, σε μια μορφή που να έχει την ακόλουθη σημασία: εάν ένας παίκτης έχει n βαθμούς παραπάνω σε σχέση με τον αντίπαλο του τότε είναι 10 φορές καλύτερος και αναμένεται να νικήσει σε 10/11 φορές που θα αναμετρηθεί. Έτσι καταλήγουμε σε μια σχέση όπως αυτή την Εξίσωσης 2.11

$$E_{team} = \frac{1}{1 + 10^{\frac{opp\_elo - team\_elo}{\sigma}}} \tag{2.11}$$

όπου n κάποιος θετικός ακέραιος που συνήθως είναι ίσος με 400.

2.4.2 Ο παράγοντας K

Η φυσική σημασία του παράγοντα K είναι πως εκφράζει τον μέγιστο αριθμό μονάδων που μπορεί να κερδίσει μια ομάδα και εύκολα γίνεται αντιληπτό πως είναι πιο σημαντικός παράγοντας του μοντέλου μας, αφού καθορίζει τη συχνότητα και τον όγκο των μεταβολών στις βαθμολογίες μας. Αν υποθέσουμε πως τη χρονική στιγμή i θα αναμετρηθούν δυο ισοδύναμες ομάδες (*home team elo = away team elo*), τότε μετά τον αγώνα στην κάθε ομάδα θα παρατηρηθεί μεταβολή ίση με $\frac{K}{2}$. Αν αντίθετα έχουμε δύο ομάδες με πολύ μεγάλη διαφορά δυναμικότητας (400 μονάδες για παράδειγμα) και ο παράγοντας K είναι πολύ μεγάλος τότε σε περίπτωση νίκης της πιο αδύναμης ομάδας, το αουτσάιντερ θα επιβραβευτεί με K μονάδες ενώ το φαβορί θα χάσει K μονάδες με αποτέλεσμα τυχαία αποτελέσματα να δίνουν μεγάλες και απροσδόκητες μεταβολές. Αντίθετα, αν το K είναι πολύ μικρό θα έχουμε ανεπαίσθητες μεταβολές στο μοντέλο μας και τα δεδομένα μας δεν θα έχουν καμία χρήσιμη πληροφορία.

Οι πιο έξυπνοι αλγόριθμοι λαμβάνουν υπόψη τόσο το πλεονέκτημα έδρας όσο και τη διαφορά στο σκορ που παρατηρήθηκε. Στην παρούσα εργασία κατασκευάστηκες μια ανεπτυγμένη φόρμουλα για τον υπολογισμό του K, που βασίστηκε στο [Silv08b]. Αν συμφωνήσουμε

πως υπάρχει το πλεονέκτημα έδρας, τότε στη βαθμολογία της γηπεδούχου θα προσθέτουμε μια ποσότητα ίση με 100.

$$\text{Home team elo} = \text{team elo} + 100$$

Επίσης, η δυναμική εκδοχή του παράγοντα K λαμβάνει υπόψη τη διαφορά πόντων που παρατηρήθηκε σε σχέση με τη διαφορά δυναμικότητας. Έτσι, στην απόλυτη τιμή της τελικής διαφοράς προσθέτουμε 3 πόντους και διαιρούμε το νούμερο που προκύπτει με την ποσότητα $7.5 + 0.006\Delta E$, όπου ΔE η διαφορά της βαθμολογίας Elo της γηπεδούχου από τη φιλοξενούμενη.

Για να γίνει κατανοητή η παραπάνω διαδικασία, θα χρησιμοποιήσουμε ένα παράδειγμα για να αντιληφθούμε τις επιβραβεύσεις του αλγόριθμου. Στους ημιτελικούς της δυτικής περιφέρειας το 2018, συναντήθηκαν οι Houston Rockets με τους Golden State Warriors . Στο πρώτο παιχνίδι της σειράς στην έδρα των GSW η γηπεδούχος είχε 118 βαθμούς Elo παραπάνω σε σχέση με τη φιλοξενούμενη. Άρα

$$\Delta E = 118 + 100 = 218$$

. Το τελικό αποτέλεσμα ήταν +4 υπέρ των GSW

$$\frac{(4+3)^{0.8}}{7.5 + .006 * 218} = 0.54$$

Αν όμως το τελικό αποτέλεσμα ήταν -4 τότε

$$\frac{(4+3)^{0.8}}{7.5 - .006 * 218} = 0.77$$

δηλαδή οι Houston Rockets θα επιβραβευτούν περισσότερο σε περίπτωση νίκης.

Η φόρμουλα που μόλις παρουσιάσαμε λαμβάνει υπόψη πως οι πιο ικανές ομάδες επικρατούν με μεγάλη διαφορά έναντι των αδύναμων αντιπάλων τους. Σε περίπτωση νίκης της πιο ισχυρής ομάδας ο παρανομαστής αυξάνεται περιορίζοντας τον αριθμητή που μπορεί να πάρει μεγάλες τιμές. Σε περίπτωση νίκης της λιγότερο ισχυρής ομάδας, ο αριθμητής δεν αλλάζει αλλά ο παρανομαστής θα μειωθεί επιβραβεύοντας με μεγαλύτερη βαθμολογία τη συγκεκριμένη νίκη.

Κεφάλαιο 3

Μηχανική Μάθηση

3.1 Επιβλεπόμενη μάθηση

Τα περισσότερα προβλήματα μηχανικής μάθησης ανήκουν σε δύο κατηγορίες, στην (i) *επιβλεπόμενη* και (ii) στην *μη-επιβλεπόμενη* μάθηση [Hast09]. Στην πρώτη περίπτωση πρόκειται για προβλήματα στα οποία ο ερευνητής έχει διαθέσιμη πληροφορία για ένα σύνολο επεξηγηματικών μεταβλητών $X_i, i = 1, \ldots, n$ και για την τιμή της μεταβλητής απόκρισης Y_i που περιγράφει την κατάσταση του συστήματος για τη δεδομένη τιμή των χαρακτηριστικών x_i. Το ζητούμενο είναι να εκτιμηθεί μια συνάρτηση $\hat{f}$ η οποία προσεγγίζει βέλτιστα τα ζευγάρια παρατηρήσεων (x_i, Y_i), υπό την έννοια $Y_i \approx \hat{f}(x_i)$.

Αντίθετα, στα προβλήματα μη-επιβλεπόμενης μάθησης ενώ έχουμε πληροφορία μόνο για κάποιες παρατηρήσεις $X_i, i = 1, \ldots, n$, χωρίς να συνοδεύονται από μεταβλητές $_i$ που να περιγράφουν την κατάσταση του συστήματος. Στην παρούσα εργασία θα ασχοληθούμε αποκλειστικά με προβλήματα που αφορούν επιβλεπόμενη μάθηση.

3.2 Επιλογή χαρακτηριστικών

Στο σύνολο δεδομένων από το οποίο καλούμαστε να εξάγουμε όλη τη δυνατή γνώση αποτελείται από δείγματα, καθένα από το οποία χαρακτηρίζεται από το ίδιο πλήθος χαρακτηριστικών. Μια τεχνική που εκτιμά τη σημασία των χαρακτηριστικών αυτών και συνεπακόλουθα μας επιτρέπει να διατηρήσουμε τα πιο "σημαντικά" από αυτά είναι η *ανάλυση κυρίων συνιστωσών* (principal component analysis - PCA) [Shle14].

Η PCA βασίζεται στον υπολογισμό της *διασποράς* και της *συνδιασποράς*. Η πρώτη εκτιμά την απόκλιση γύρω από τη μέση τιμή τυχαίας μεταβλητής ($Var(X) = \frac{\sum_i (X_i - X)^2}{n-1}$) ενώ η δεύτερη εκτιμά την απόκλιση ταυτόχρονα δύο τυχαίων μεταβλητών ως προς τις μέσες τιμές τους ($Cov(X, Y) = \frac{\sum_i (X_i - X)(Y_i - Y)}{n-1}$) και κυμαίνεται στο $[-1, 1]$ (στις ακραίες τιμές ισχύει $Y = -X$ και $Y = X$, αντίστοιχα). Στις ακραίες περιπτώσεις, οι δύο αυτές τυχαίες μεταβλητές παρουσιάζουν τέλεια γραμμική εξάρτηση, οπότε μπορούμε να κρατήσουμε τη μια από τις δυο. Όταν η συνδιασπορά είναι 0 λέμε πως οι μεταβλητές είναι ασυσχέτιστες, πράγμα που αποτελεί ισχυρή ένδειξη πως είναι ανεξάρτητες, οπότε χρειάζεται να τις χρησιμοποιήσουμε και τις δύο στην προσπάθεια εξόρυξης γνώσης. Συμπερασματικά, από τον πίνακα διασποράς μπορούμε να πληροφορηθούμε για το πως συνδέονται δυο τυχαίες μεταβλητές. Στην περίπτωση που η μέση τιμή των τυχαίων μεταβλητών A, B είναι 0 έχουμε $Cov(A, B) = A^\top B$ και ο πίνακας διασποράς του συνόλου δεδομένων παίρνει τη μορφή της Εξίσωσης 3.1

$$X^\top X = \begin{bmatrix} Var(X_1) & Cov(X_1, X_2) & .... & Cov(X_1, X_n) \\ Cov(X_2, X_1) & Var(X_2) & .... & Cov(X_2, X_n) \\ .. & ... & .. & \end{bmatrix} \tag{3.1}$$

Υποθέτοντας πως τα πιο σημαντικά χαρακτηριστικά των δεδομένων μας εμφανίζουν τη μεγαλύτερη διασπορά ενώ εκείνα με πολύ μικρή αποτελούν "θόρυβο", και κατ' επέκταση δε βοηθούν στην εξαγωγή γνώσης, επιθυμούμε να βρούμε ένα ορθοκανονικό μετασχηματισμό P τέτοιο ώστε (Εξίσωση 3.2)

$$Y = P \cdot X \tag{3.2}$$

Οι στήλες του P αποτελούν το σύνολο των διανυσμάτων (χώρο) στον οποίο προβάλλονται τα δεδομένα μας.

Κάθε πίνακας X μπορεί να αναλυθεί στην ισοδύναμη μορφή της Εξίσωσης 3.3

$$X = U\Sigma V^\top \tag{3.3}$$

που είναι γνωστή ως *ανάλυση ιδιαζουσών τιμών* (singular value decomposition - SVD). Ο πίνακας Σ είναι διαγώνιος με τα διαγώνια στοιχεία να είναι επιπρόσθετα μη-αρνητικά. Οι πίνακες U, V είναι ορθογώνιοι δηλαδή $U^\top = U^{-1}, V^\top = V^{-1}$. Για να υπολογιστούν οι πίνακες U, V παρατηρούμε πως $X^\top X = V\Sigma^\top \Sigma V^\top$ και ο πίνακας V έχει τα ιδιοδιανύσματα του συμμετρικού πίνακα ενώ ο $\Sigma^\top\Sigma$ τις ιδιοτιμές. Αν διαιρέσουμε με $n-1$ προκύπτει ο γνωστός πίνακας διασποράς.

Οι ιδιοτιμές (s_i) του πίνακα διασποράς με τις ιδιοτιμές του αρχικού πίνακα συνδέονται με τη σχέση $s_i = \lambda_i^2$. Παραγοντοποιώντας τον πίνακα διασποράς (η παραγοντοποίηση δεν είναι μοναδική) και θέτοντας τον περιορισμό των ιδιοτιμών να βρίσκονται σε αύξουσα σειρά στον πίνακα Σ, τότε η παραγοντοποίηση SVD καθίσταται μοναδική μόνο για τον πίνακα X. Εάν θέλουμε να μειώσουμε τις διαστάσεις του X, επιλέγουμε τις k κύριες (μεγαλύτερες) ιδιοτιμές του καθώς και τα ιδιοδιανύσματα που αντιστοιχούν σε αυτές (Εξίσωση 3.4)

$$X_k = U_k \Sigma_k {V_k}^\top \tag{3.4}$$

Οι άξονες με τις k κυριότερες συνιστώσες βρίσκονται στον V_k. Πολλές φορές για να αποφασίσουμε ποιες διαστάσεις να χρησιμοποιήσουμε, σχεδιάζουμε τη γραφική παράσταση της εκτιμώμενης διασποράς υπό εξέταση.

3.2.1 Η μέθοδος των πυρήνων για την εκτίμηση της συνάρτηση πυκνότητας πιθανότητας

Έστω πως έχουμε $\mathcal{X}_1, \mathcal{X}_2, ..., \mathcal{X}_n$, ανεξάρτητες παρατηρήσεις από μια άγνωστη κατανομή f. Για να μπορέσουμε να εκτιμήσουμε την f αρχικά τα μόνα μας εργαλεία είναι τα μέτρα θέσης και διασποράς, αλλά και η γραφική αναπαράσταση συνήθως με κάποιο ιστόγραμμα συχνοτήτων. Ακόμα και η επιλογή διαφορετικού πλήθους αριθμού κλάσεων μας δίνει διαφορετική διαίσθηση για τη ζητούμενη συνάρτηση. Μια πολύ χρήσιμη τεχνική για τη γραφική αναπαράσταση της f είναι η μέθοδος των πυρήνων [Chen17]. Πρόκειται για μια μη-παραμετρική μέθοδο, αφού δεν χρειάζεται να κάνουμε υποθέσεις για τη ζητούμενη συνάρτηση, αλλά πρέπει να εφαρμόσουμε σε κάθε σημείο στο δείγμα μας μια συνάρτηση που θα λαμβάνει υπόψη και όλα τα σημεία γύρω από το σημείο από το οποίο θέλουμε να εκτιμήσουμε τη συνάρτηση πιθανότητας.

Γενικότερα, η συνάρτηση θα πρέπει να ικανοποιεί τις ιδιότητες της Εξίσωσης 3.5

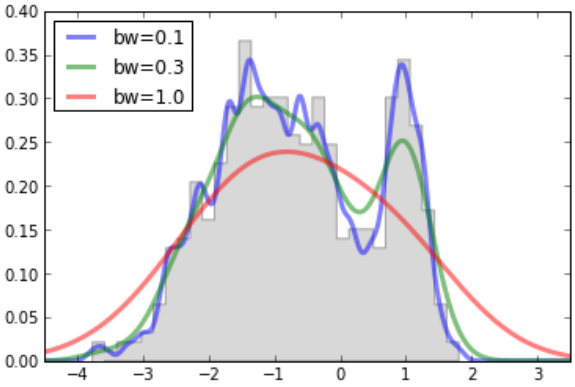
$$K(x) > 0, \quad \int_{-\infty}^{\infty} K(x) = 1, \quad E[K(x)] = 0, \quad \int_{-\infty}^{\infty} x^2 K(x) = 1 \tag{3.5}$$

Μια από τις πιο συνηθισμένες επιλογές συνάρτησης πυρήνα είναι η συνάρτηση $K(x) = \frac{1}{2\pi}\exp\left(-x^2/2\right)$, ενώ η εκτιμώμενη συνάρτηση πυκνότητας πιθανότητας ($\hat{f}$) για

οποιαδήποτε επιτρεπτή συνάρτηση πυρήνα είναι ίση με (Εξίσωση 3.6)

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{3.6}$$

Ο παράγοντας h είναι γνωστός ως εύρος (bandwidth) και είναι ο πιο σημαντικός παράγοντας που πρέπει να εκτιμηθεί για να έχουμε το καλύτερο δυνατό αποτέλεσμα. Ήδη από την Εξίσωση 3.6 βλέπουμε πως καθορίζει πόσο βάρος θα δώσουμε στις παρατηρήσεις γύρω από το οποίο θέλουμε να εκτιμήσουμε την πιθανότητα. Η σημασία της επιλογής του εύρους φαίνεται στο Σχήμα 3.1. Μεγάλη τιμή της παραμέτρου h μας δίνει εκτιμήσεις πολύ ομαλές, μιας και δίνεται το ίδιο σχεδόν βάρος σε όλες τις παρατηρήσεις. Αντίθετα, μικρές τιμές της παραμέτρου h μας δίνουν εκτιμήσεις λιγότερο ομαλές, αφού δίνεται μεγαλύτερο βάρος σε κοντινές παρατηρήσεις με αποτέλεσμα να δημιουργούνται αρκετές κορυφές στο γράφημα.



Σχήμα 3.1: Παράδειγμα εφαρμογής της μεθόδου των πυρήνων

Για την εκτίμηση της παραμέτρου h έχουν αναπτυχθεί αρκετές μέθοδοι αναλυτικές και προσεγγιστικές, ανάλογα με την περίπτωση, αλλά μια μέθοδος που λειτουργεί πάντα με πολύ ικανοποιητικά αποτελέσματα και είναι συμβατή με τη θεωρία είναι η μέθοδος της διασταυρούμενης επικύρωσης.

3.3 Μετρικές αξιολόγησης

Για να συγκρίνουμε τις διαφορετικές μεθοδολογίες που μπορούν να εφαρμοστούν σε ένα πρόβλημα μηχανικής μάθησης, χρειάζομαστε τις κατάλληλες μετρικές. Μια από αυτές είναι η ακρίβεια (precision), η οποία ορίζεται όπως στην Εξίσωση 3.7 παρακάτω

$$\text{Ακρίβεια} = \frac{\text{Πλήθος σωστών προβλέψεων}}{\text{Πλήθος συνολικών προβλέψεων}} \tag{3.7}$$

Για να ελέγξουμε την αποδοτικότητα των αλγορίθμων πρόβλεψης ανάμεσα στις διαφορετικές κλάσεις μπορούμε καταρχήν χρησιμοποιούμε τον παρακάτω *πίνακα σύγχυσης* (confusion matrix)

Στη βάση του Πίνακα 3.1 ορίζονται οι παρακάτω μετρικές (Εξισώσεις 3.8-3.10)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.8}$$

$$\text{True Positive Rate} = \frac{TP}{\text{TP+FN}} = \text{Sensitivity} \tag{3.9}$$

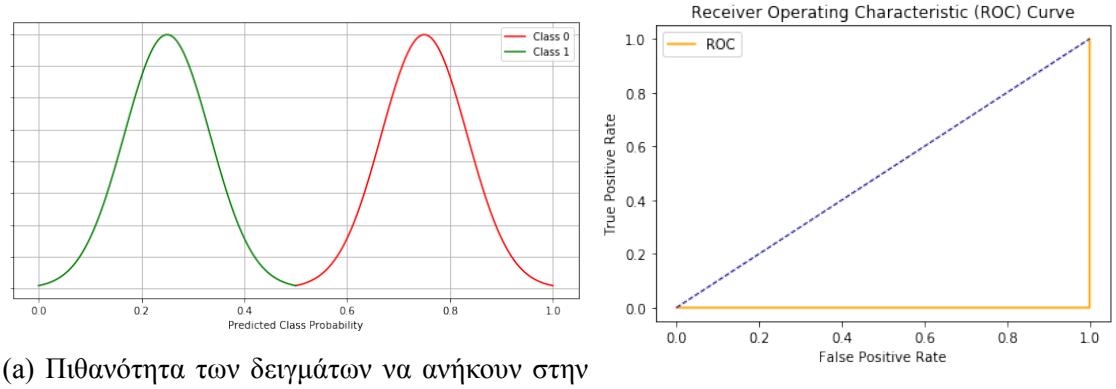
$$\text{False Positive Rate} = \frac{FP}{\text{TN+FP}} = 1 - \text{Specificity} \tag{3.10}$$

Πίνακας 3.1: Πίνακας Σύγχυσης

	Πρόβλεψη Αρνητικού	**Πρόβλεψη Θετικού**
Πραγματικό Αρνητικό	Αληθώς Αρνητικό (TN)	Ψευδώς Θετικό (FP)
Πραγματικό Θετικό	Ψευδώς Αρνητικό (FN)	Αληθώς Θετικό (TP)

Για τη ταξινόμηση των δειγμάτων συνήθως χρησιμοποιείται το κατώφλι του 0,5. Δηλαδή, αν ο ταξινομητής μετά την εκπαίδευση δίνει πιθανότητα 0,6 και 0,4 το δείγμα να ανήκει στην κλάση 1 και 0 αντίστοιχα τότε μπορούμε να αποφανθούμε πως για το κατώφλι του 0,5 το δείγμα μας ανήκει στην κατηγορία 1. Αν το κατώφλι ήταν 0,7 τότε το δείγμα θα είχε ταξινομηθεί στην κλάση 0. Παρατηρούμε πως όσο πιο μεγάλη ικανότητα έχει ο ταξινομητής να διαχωρίσει τα δείγματα που ανήκουν στην κλάση 1 από την κλάση 0, τότε όσο αυξάνουμε βλέπουμε πως όλο και λιγότερο θα μειώνεται η *ευαισθησία* (sensitivity) και αντίστοιχα όλο και λιγότερο θα αυξάνεται η *εξειδίκευση* (specificity).

Στην συνέχεια θα παρουσιαστούν παραδείγματα για να δούμε πως προκύπτει η καμπύλη και πως αποτελεί μέτρο της διακριτικής ικανότητας ενός δίτιμου ταξινομητή. Στο Σχήμα 3.2

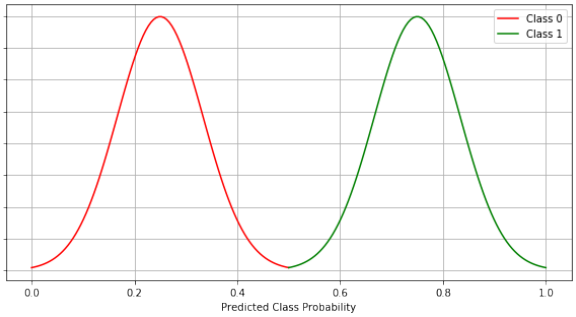


(a) Πιθανότητα των δειγμάτων να ανήκουν στην κλάση 1

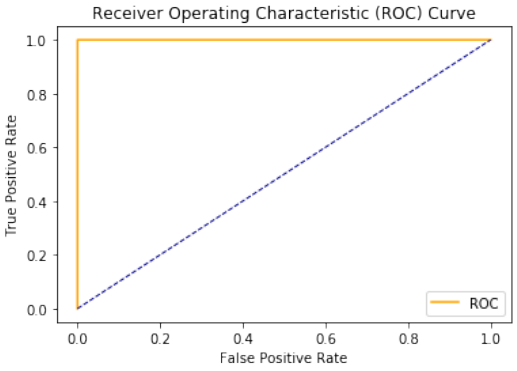
(b) ROC καμπύλη

Σχήμα 3.2: Παράδειγμα ταξινομητή με μηδενική ακρίβεια

ο ταξινομητής παρουσιάζει μηδενική ακρίβεια στο δείγμα μας, αφού όλα τα δείγματα που ανήκουν στην κλάση 1 δείχνουν να έχουν πιθανότητα μικρότερη από 0,5 να ανήκουν σε αυτή. Ακριβώς ίδια εικόνα για τα δείγματα που ανήκουν στην κλάση 0. Συνεπώς, ο συγκεκριμένος ταξινομητής δε μπορεί να διακρίνει τις κλάσεις μεταξύ τους και το *εμβαδόν κάτω από την καμπύλη* (area under the curve - AUC) είναι 0. Στο Σχήμα 3.3 ο ταξινομητής ταξινομεί όλα τα δείγματα σωστά στην κλάση που ανήκουν και το εμβαδόν κάτω την *καμπύλη λειτουργικών χαρακτηριστικών* (receiver operating characteristic - ROC) λαμβάνει τη μέγιστη τιμή του, που είναι ίση με 1. Στο Σχήμα 3.4 ο ταξινομητής δίνει σε κάθε δείγμα την ίδια πιθανότητα να ανήκει και στις δύο κλάσεις. Η τυχαία αυτή συμπεριφορά έχει σαν αποτέλεσμα το εμβαδόν κάτω από την καμπύλη ROC να είναι 0,5. Στην πράξη θα επιθυμούμε σίγουρα ο ταξινομητής να μην έχει τυχαία συμπεριφορά και το εμβαδόν κάτω από την καμπύλη ROC να είναι μεγαλύτερο του 0,5.

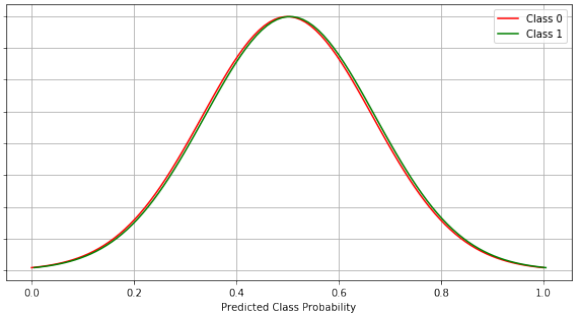


(a) Πιθανότητα των δειγμάτων να ανήκουν στην κλάση 1

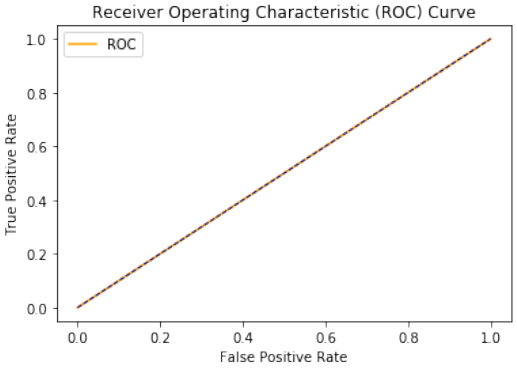


(b) ROC καμπύλη

Σχήμα 3.3: Παράδειγμα τέλειου ταξινομητή



(a) Πιθανότητα των δειγμάτων να ανήκουν στην κλάση 1



(b) ROC καμπύλη

Σχήμα 3.4: Παράδειγμα τυχαίου ταξινομητή

3.4 Διασταυρούμενη Επικύρωση

3.4.1 Έλεγχος Προσαρμογής

Στις τεχνικές της επιβλεπόμενης μάθησης γίνεται η θεώρηση πως η ετικέτα y των δεδομένων x προκύπτει ως συνάρτησή τους (δηλαδή ισχύει $y = f(x)$), η εύρεση της οποίας είναι και το ζητούμενο της μάθησης. Στην πράξη, αναζητείται συνάρτηση $\hat{f}$ που προσεγγίζει βέλτιστα την f. Η αναζήτηση αυτή γίνεται βάσει ενός κριτηρίου, το οποίο επιθυμούμε να βελτιστοποιήσουμε και το οποίο είναι γνωστό ως *συνάρτηση σφάλματος*.

Επειδή αφενός χρειάζεται να εξετάσουμε τη λειτουργία και τη συμπεριφορά του ταξινομητή σε "νέα" δεδομένα (στα οποία δεν έχει εκπαιδευτεί), αφετέρου οι η απόδοση των περισσότερων ταξινομητών επηρεάζεται από παραμέτρους θ, το σύνολο των δεδομένων χωρίζεται σε δύο ξεχωριστά υποσύνολα

1. Το *σύνολο εκπαίδευσης*, στο οποίο εκτιμώνται οι παράμετροι του μοντέλου

2. Το *σύνολο επικύρωσης*, στο οποίο υπολογίζεται η συνάρτηση σφάλματος σύμφωνα με τις παραμέτρους που προέκυψαν κατά την εκπαίδευση του μοντέλου.

Έτσι πλέον μπορεί να ελαχιστοποιηθεί ταυτόχρονα το σφάλμα στο σύνολο εκπαίδευσης και επικύρωσης χωρίς να υπάρχει μεγάλη απόκλιση μεταξύ τους. Όταν το σφάλμα εκπαίδευσης δε μπορεί να μειωθεί, λέμε πως έχουμε *υποεκπαίδευση* (underfitting) και όταν το σφάλμα

επικύρωσης είναι πολύ μεγαλύτερο σε σχέση με το σφάλμα εκπαίδευσης τότε έχουμε *υπερεκπαίδευση* (overfitting). Η τεχνική με την οποία πραγματοποιείται ολόκληρη η παραπάνω διαδικασία ονομάζεται *διασταυρούμενη επικύρωση* (cross validation) [Hast09]. Ως υπερπαράμετροι επιλέγονται εκείνες που μειώνουν το σφάλμα επικύρωσης, σύμφωνα με την ακολουθούμενη κάθε φορά τακτική.

3.4.2 Διασταυρούμενη Επικύρωση Εξαίρεσης Ενός

Στη *διασταυρούμενη επικύρωση εξαίρεσης ενός* (leave-one-out cross validation) επιλέγεται η i-οστή παρατήρηση ως σύνολο επικύρωσης. Το μοντέλο εκπαιδεύεται στα στα υπόλοιπα στοιχεία πλην της $i-$οστής παρατήρησης και η διαδικασία επαναλαμβάνεται, μέχρι κάθε στοιχείο να βρεθεί στο σύνολο ελέγχου μια φορά. Η συνολική εκτίμηση σφάλματος δίνεται στην Εξίσωση 3.11

$$\mathcal{E} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{y}_{-i}) \tag{3.11}$$

όπου $\hat{y}_{-i}$ είναι η εκτίμηση για την $i-$οστη παρατήρηση μετά την εκπαίδευση από όλο το δείγμα χωρίς αυτήν.

3.4.3 Διασταυρούμενη Επικύρωση Εξαίρεσης k

Σε αντίθεση με την προηγούμενη περίπτωση, στη *διασταυρούμενη επικύρωση εξαίρεσης* k (leave k-out cross validation), επιλέγονται k τυχαίες παρατηρήσεις ως σύνολο επικύρωσης και οι υπόλοιπες αποτελούν το σύνολο εκπαίδευσης.

3.4.4 Διασταυρούμενη Επικύρωση k μερών

Στη διασταυρούμενη επικύρωση k μερών (k-fold cross validation) το σύνολο δεδομένων χωρίζεται σε k τυχαία, μη-επικαλυπτόμενα μέρη ίσου μεγέθους. Σε κάθε επανάληψη της διαδικασίας, επιλέγεται ένα μέρος ως σύνολο επικύρωσης και τα υπόλοιπα $k-1$ συνενώνονται στο σύνολο εκπαίδευσης. Το συνολικό σφάλμα δίνεται στην Εξίσωση 3.12

$$\mathcal{E} = \frac{1}{k}\sum_{i=1}^{K} e_i \tag{3.12}$$

όπου e_i το σφάλμα επικύρωσης στην i-οστή επανάληψη της διαδικασίας (Εξίσωση 3.13)

$$e_i = \frac{1}{|J_i|}\sum_{j\in J_i} L(y_i, \hat{y}_{-i}) \tag{3.13}$$

όπου με J_i συμβολίζεται το σύνολο των δεικτών που ανήκουν στην πτυχή i.

Γενικότερα, υπάρχουν και άλλες τεχνικές διασταυρούμενης επικύρωσης. Σε κάθε περίπτωση, η τεχνική που επιλέγεται εξαρτάται από τον όγκο των δεδομένων, την υπολογιστική ισχύ καθώς και την πολυπλοκότητα του προβλήματος.

3.5 Μοντέλα παλινδρόμησης

3.5.1 Γραμμική

Πολλές φορές η σχέση που συνδέει τη μεταβλητή απόκρισης Y με τις επεξηγηματικές μεταβλητές που τις συμβολίζουμε με X_i είναι γραμμική, όπως φαίνεται στην Εξίσωση 3.14

$$Y_i = \beta_0 + \beta_1 x_{i1} + +\beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \ldots, n \tag{3.14}$$

όπου $\beta_i,\ i = 1, \ldots, p$ είναι οι συντελεστές του μοντέλου που περιγράφουν το βαθμό επίδρασης της $i-$επεξηγηματικής μεταβλητής και β_0 η σταθερά του μοντέλου. Το μοντέλο της *γραμμικής παλινδρόμησης* έχει προταθεί στη βιβλιογραφία για τη μοντελοποίηση τέτοιων προβλημάτων επιβλεπόμενης μάθησης [Tibs94, Κα17]. Ο όρος ϵ_i αφορά τα τυχαία σφάλματα και δεν πρέπει να συγχέεται με τις ποσότητες $e_i = y_i - \hat{y}_i$, οι οποίες εκτιμούν το πόσο μακριά είναι οι εκτιμώμενες τιμές της μεταβλητής απόκρισης $\hat{y}_i = \beta^\top x_i$ από τις πραγματικές τιμές της μεταβλητής απόκρισης (y_i). Για να τον υπολογισμό των συντελεστών γράφουμε τη σχέση υπό μορφή πινάκων (Εξίσωση 3.15)

$$Y = \mathrm{X}\beta + \epsilon \Rightarrow X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{d1} & x_{n2} & x_{n3} & \ldots & x_{np} \end{bmatrix} \tag{3.15}$$

και θα προσπαθήσουμε να ελαχιστοποιήσουμε την εκτιμήτρια των ελαχίστων τετραγώνων με τους εκτιμώμενους συντελεστές του μοντέλου να είναι (Εξίσωση 3.16)

$$\hat{\beta} = \arg \min_{\beta \in R^{p+1}} \|Y - \mathrm{X}\beta\|_2^2 \tag{3.16}$$

Ο έλεγχος και η αξιολόγηση του εκτιμώμενου ή προσαρμοσμένου μοντέλου μας με τη μέθοδο των ελαχίστων τετραγώνων βασίζεται στις ακόλουθες υποθέσεις για τα τυχαία σφάλματα ϵ_i: (i) $E[\epsilon_i] = 0$, (ii) $V[\epsilon_i] = \sigma^2$ και (iii) Cov(ϵ_i, ϵ_j) = 0. Αν συμβολίσουμε με $RSS = \|Y - \mathrm{X}\beta\|_2^2$, τότε αναπτύσσεται σε μορφή πινάκων σύμφωνα με την Εξίσωση 3.17

$$RSS(\beta) = \|Y - \mathrm{X}\beta\|_2^2 = (Y - \mathrm{X}\beta)^\top \cdot (Y - \mathrm{X}\beta) \tag{3.17}$$

και μπορούμε να υπολογίσουμε τους συντελεστές εύκολα σε μορφή πινάκων (Εξίσωση 3.18)

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2X^\top (Y - X\beta) = 0 \tag{3.18}$$

ενώ αν αντιστρέφεται ο πίνακας ($\mathrm{X}^\top\mathrm{X}$) τότε (Εξίσωση 3.19)

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y \tag{3.19}$$

Αν επιπλέον υποθέσουμε πως τα τυχαία σφάλματα ακολουθούν την κανονική κατανομή ($\epsilon_i \sim N(0, \sigma^2)$) και είναι και ανεξάρτητα και ισόνομα τότε (Εξίσωση 3.20)

$$\hat{\beta} \sim N(\beta, \sigma^2(\mathrm{X}^\top\mathrm{X})^{-1}) \tag{3.20}$$

Οι συντελεστές φαίνονται να είναι αμερόληπτοι εκτιμητές των πραγματικών συντελεστών. Η φυσική ερμηνεία κάθε συντελεστή β_i είναι:

Ερμηνεία 1: *Αν η μεταβλητή απόκρισης* X_i *αυξηθεί κατά μια μονάδα διατηρώντας τις υπόλοιπες μεταβλητές σταθερές τότε η μεταβολή της μεταβλητής απόκρισης Y είναι ίση με* β_i

Συνεπώς, αν μια μεταβλητή που έχουμε συμπεριλάβει στο πρόβλημα έχει θετικό συντελεστή μετά την προσαρμογή του μοντέλου γραμμικής παλινδρόμησης αυτό σημαίνει πως η αύξηση της οδηγεί σε αύξηση της μεταβλητής απόκρισης κρατώντας τις υπόλοιπες μεταβλητές σταθερές. Ακριβώς, αντίστοιχα αποτελέσματα έχουμε στην περίπτωση που ο β_i είναι αρνητικός.

Ο συντελεστής προσδιορισμού R^2 (Εξίσωση 3.21) αποτελεί κριτήριο εκτίμησης του ποσοστού της μεταβλητότητας της μεταβλητής απόκρισης που εξηγείται από την X. Οι τιμές που παίρνει βρίσκονται ανάμεσα στο $[0, 1]$ και όσο μεγαλύτερη είναι τόσο ισχυρότερη είναι η γραμμική εξάρτηση μεταξύ των X και Y.

$$R^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{3.21}$$

Κάθε στατιστικός έλεγχος έχει μια p-τιμή (p-value) που μας δηλώνει την πιθανότητα ο συντελεστής να είναι στατιστικά διάφορος του μηδενός. Ο στατιστικός έλεγχος F εξετάζει την υπόθεση όλοι οι συντελεστές του μοντέλου να είναι ίσοι με μηδέν.

3.5.2 Λογιστική

Η *λογιστική παλινδρόμηση* [Κα17] αποτελεί ειδική μορφή παλινδρόμησης που μελετά φαινόμενα στα οποία η εξαρτημένη μεταβλητή μπορεί να πάρει δύο μόνο δυνατές τιμές, την τιμή 0 ή την τιμή 1. Το 1 ως έξοδος καλείται συνήθως ως επιτυχία και το 0 ως αποτυχία. Ενώ στο απλό γραμμικό μοντέλο είχαμε $E[y_i] = \hat{y}_i = \beta^\top x_i$, στη λογιστική παλινδρόμηση επιθυμούμε να βρούμε ένα τρόπο να εκφράσουμε την πιθανότητα το i-οστό δείγμα να ανήκει σε ένα από τα δύο ενδεχόμενα. Έτσι αναζητούμε μια συνάρτηση h τέτοια ώστε (Εξίσωση 3.22)

$$h(p) = \beta^\top x_i, \quad h \text{ αντιστρέψιμη} \tag{3.22}$$

και η αντίστροφη απεικόνιση της h να έχει σύνολο τιμών το $[0, 1]$. Η πιο συνηθισμένη επιλογή *συνάρτησης σύνδεσης* (link function) είναι η συνάρτηση logit (Εξίσωση 3.23)

$$h(p) = \ln \frac{p}{1-p} \tag{3.23}$$

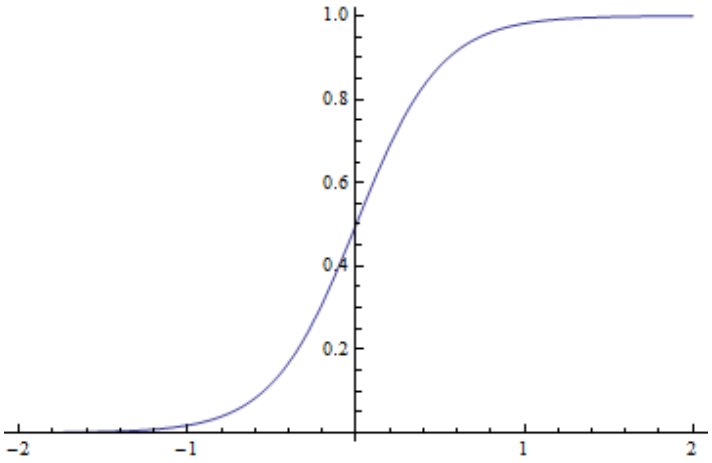
με αντίστροφη συνάρτηση (Εξίσωση 3.24)

$$h^{-1}(\beta^\top x_i) = \frac{e^{\beta^\top x_i}}{1 + e^{\beta^\top x_i}} = p_i \tag{3.24}$$

Η συγκεκριμένη αντίστροφη συνάρτηση που βρήκαμε είναι η γνωστή σιγμοειδής συνάρτηση (Σχήμα 3.5) που χρησιμοποιείται ευρέως στη μηχανική μάθηση. Το μοντέλο μπορεί να γραφτεί με τη μορφή της Εξίσωσης 3.25

$$\text{logit}(p_i) = g(p_i) = \beta_0 + \beta_1 x_{i1} + +\beta_2 x_{i2} + ... + \beta_p x_{ip} \tag{3.25}$$

Για να ερμηνεύσουμε τους συντελεστές θα χρειαστεί να γράψουμε τις προηγούμενες εξισώσεις υπό τη μορφή $\frac{p}{1-p} = e^{x^\top \beta}$ και να θυμηθούμε πως η ποσότητα p εκφράζει την πιθανότητα το δείγμα μας να ανήκει στην κλάση που συμβολίζουμε με 1. Άρα, με $1 - p$ δηλώνεται η πιθανότητα να ανήκει στην έτερη κλάση. Συνεπώς, είναι εύκολο να παρατηρήσουμε πως αν ο



Σχήμα 3.5: Η σιγμοειδής συνάρτηση

συντελεστής β_j είναι θετικός τότε η αύξηση του παράγοντα x_i αυξάνει τη σχετική πιθανότητα της Εξίσωσης 3.26 (odds ratio)

$$\frac{p}{1-p} = \frac{P[X \in 1|X = x]}{P[X \in 0|X = x]} \tag{3.26}$$

κρατώντας τους υπόλοιπους παράγοντες σταθερούς.

3.5.3 Ridge και Lasso παλινδρόμηση

Στη γραμμική παλινδρόμηση οι εκτιμήτριες που προκύπτουν έχουν την ιδιότητα της αμεροληψίας. Ωστόσο, υπάρχει ο κίνδυνος το μοντέλο που προκύπτει να μην μπορεί να γενικεύσει ικανοποιητικά στην εφαρμογή του σε δεδομένα που δεν έχει δει και στην πράξη να είναι άχρηστο. Για να αντιμετωπιστεί αυτό το ενδεχόμενο, μια λύση είναι η προσθήκη όρων ποινής στις εκτιμήτριες, καθιστώντας τες πλέον μη-αμερόληπτες. Γενικότερα, η αύξηση της αμεροληψίας επιφέρει μείωση της διασποράς, βοηθάει στην αντιμετώπιση της πολυσυγγραμμικότητας και μπορεί να δώσει απάντηση στο ερώτημα του ποιες μεταβλητές να επιλεγούν.

Η πιο γνωστή προσέγγιση είναι η προσθήκη κάποιας νόρμας των εκτιμητριών μαζί με τα τετράγωνα των υπολοίπων στη συνάρτηση προς ελαχιστοποίηση. Η εύρεση των συντελεστών πλέον μεταφράζεται σε ένα πρόβλημα ελαχιστοποίησης, στο οποίο πρέπει να βρεθεί ο βέλτιστος βαθμός ομαλοποίησης της νόρμας του διανύσματος των συντελεστών. Έτσι, οι συντελεστές πλέον δεν μπορούν να πάρουν απεριόριστα μεγάλες τιμές και υπό κατάλληλες συνθήκες μπορούν να μηδενιστούν, λαμβάνοντας έτσι μια πολύ ισχυρή μεθοδολογία για την επιλογή μεταβλητών για τους αλγορίθμους μηχανικής μάθησης [Hast09, Wier15, Tibs94].

Παλινδρόμηση Ridge

Στη μεθοδολογία αυτή, για την εκτίμηση των εκτιμητριών καλούμαστε να ελαχιστοποιήσουμε την ποσότητα της Εξίσωσης 3.27

$$\min_{\beta \in R^{p+1}} \|Y - \mathrm{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \tag{3.27}$$

όπου λ είναι ο παράγοντας συρρίκνωσης. Αποδεικνύεται ότι (Εξίσωση 3.28)

$$\min_{\beta \in R^{p+1}} \|Y - \mathrm{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \Leftrightarrow \min_{\beta \in R^{p+1}} \|Y - \mathrm{X}\beta\|_2^2 \text{ με} \sum_i \beta_i^2 \leq t \tag{3.28}$$

όπου t είναι άνω φράγμα της l_2 νόρμας των εκτιμητριών. Άρα, υπάρχει ένα-προς-ένα αντιστοιχία μεταξύ της παραμέτρου ποινής και της της l_2 νόρμας των εκτιμητριών. Για $\lambda = 0$ λαμβάνουμε τους εκτιμητές της απλής γραμμικής παλινδρόμησης ενώ για $\lambda \to \infty$ οι συντελεστές τείνουν να μηδενιστούν. Στην περίπτωση της Ridge παλινδρόμησης μπορούμε να υπολογίσουμε αναλυτικά τις εκτιμήτριες της μεθοδολογίας (Εξίσωση 3.29)

$$\frac{\partial RSS(\beta)}{\partial \beta} = \frac{\partial (Y - \mathrm{X}\beta)^\top (Y - \mathrm{X}\beta) + \lambda \beta^\top \beta}{\partial \beta} = 2X^\top (Y - X\beta) + 2\lambda\beta = 0 \qquad (3.29)$$

Συνεπώς

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top y$$

και

$$E[\hat{\beta}] = E[(X^\top X + \lambda I)^{-1} X^\top y] = (X^\top X + \lambda I)^{-1} X^\top [y] = (X^\top X + \lambda I)^{-1} X^\top X\beta$$

Η μέση τιμή είναι διαφορετική από β όταν $\lambda \neq 0$ και κατά συνέπεια οι συντελεστές που προκύπτουν δεν ικανοποιούν την ιδιότητα της αμεροληψίας.

Αντίστοιχα, για να υπολογίσουμε τη διασπορά της εκτιμήτριας (Εξίσωση 3.30)

$$\begin{aligned} V[\hat{\beta}] &= V[(X^\top X + \lambda I)^{-1} X^\top y] = (X^\top X + \lambda I)^{-1} V[y]((X^\top X + \lambda I)^{-1})^\top \\ &= \sigma_\epsilon^2 (X^\top X + \lambda I)^{-1} [(X^\top X)(X^\top X + \lambda I)^{-1}]^\top \end{aligned} \qquad (3.30)$$

και η διασπορά είναι μικρότερη από τους εκτιμητές που λαμβάνουμε από τη γραμμική παλινδρόμηση [Wier15].

Παλινδρόμηση Lasso

Στην παλινδρόμηση Lasso προσπαθούμε να ελαχιστοποιήσουμε τη L_1 νόρμα των βαρών (Εξίσωση 3.31)

$$\min_{\beta \in R^{p+1}} \|Y - \mathrm{X}\beta\|_2^2 + \lambda \|\beta\|_1^2 \qquad (3.31)$$

και όπως και στην περίπτωση της παλινδρόμησης Ridge, έχουμε (Εξίσωση 3.32)

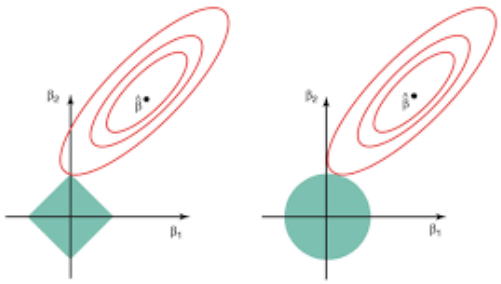
$$\min_{\beta \in R^{p+1}} \|Y - \mathrm{X}\beta\|_2^2 + \lambda \|\beta\|_1^2 \Leftrightarrow \min_{\beta \in R^{p+1}} \|Y - \mathrm{X}\beta\|_2^2 \qquad (3.32)$$

με $\sum_i |\beta_i| \leq t$. Η διαφορά με την προηγούμενη περίπτωση είναι πως πλέον δεν έχουμε κλειστή λύση στο πρόβλημα ελαχιστοποίησης που προκύπτει. Η πιο σημαντική διαφορά, όμως, είναι πως μπορεί να δώσει λύσεις όπου οι συντελεστές κάποιων βαρών είναι μηδενικοί, επιστρέφοντας αραιές λύσεις ως προς το διάνυσμα των συντελεστών. Συνεπώς, μπορεί να αποτελέσει ένα σημαντικό εργαλείο στη διαδικασία επιλογής μεταβλητών.

Σε κάθε περίπτωση προσπαθούμε να λύσουμε το σύστημα της μορφής (Εξίσωση 3.33)

$$\min_{\beta \in R^{p}} \|Y - \mathrm{X}\beta\|_2^2 \qquad (3.33)$$

με $\|\beta\|_j \leq t$, όπου $j \in \{1, 2\}$. Στο παράδειγμα του Σχήματος 3.6 βλέπουμε την απλή περίπτωση που το διάνυσμα των μεταβλητών είναι δισδιάστατο. Οι κόκκινες καμπύλες είναι διαφορετικές και ισοϋψείς, ενώ στην πράσινη επιφάνεια είναι ο χώρος μέσα στον οποίο βρίσκονται όλες οι επιτρεπτές λύσεις. Στην περίπτωση της Lasso, οι εκτιμήτριες βρίσκονται εντός ενός ρόμβου ενώ στη Ridge, εντός ενός κύκλου. Στην πράξη, για τον υπολογισμό της



Σχήμα 3.6: Σχηματική αναπαράσταση του προβλήματος ελαχιστοποίησης της Ridge και της Lasso παλινδρόμησης στο $\mathbb{R}^2$

παραμέτρου ποινής εφαρμόζουμε διασταυρούμενη επικύρωσης ως προς οποιοδήποτε κριτήριο επιθυμούμε. Συχνά, χρησιμοποιείται και ο κανόνας γνωστός ως *1 standard error*, όπου: (i) εντοπίζουμε τον παράγοντα $\hat{\lambda}$ με την ιδιότητα $\hat{\lambda} = \arg\min CV$ και (ii) ψάχνουμε το μεγαλύτερο $\tilde{\lambda}$ που ικανοποιεί τη σχέση $CV(\tilde{\lambda}) \leq CV(\hat{\lambda}) + S.E(\hat{\lambda})$

Οι συντελεστές που επιστρέφονται έχουν παρόμοια προβλεπτική ικανότητα αλλά είναι περισσότερο ομαλοποιημένοι, δηλαδή το διάνυσμα που επιστρέφεται έχει μικρότερους συντελεστές σε σχέση με το μοντέλο που αντιστοιχεί στο $\hat{\lambda} = \arg\min CV$.

3.6 Ταξινομητές

3.6.1 Λογιστική παλινδρόμηση

Η Εξίσωση 3.34 περιγράφει πως ένα μοντέλο λογιστικής παλινδρόμησης μπορεί να λειτουργήσει ως ταξινομητής

$$logit(p) = \sum_{i=0}^{n} w_i x_i = \mathbf{w}^\top x \tag{3.34}$$

Ο όρος p δηλώνει τη πιθανότητα το δείγμα με χαρακτηριστικά $x = (x_1, x_2, ..., x_n)$ και ετικέτα y να ανήκει στην κατηγορία που καλούμε *επιτυχία*. Συνεπώς η έξοδος του ταξινομητή είναι (Εξίσωση 3.35)

$$\hat{y} = \begin{cases} 1 & \text{εάν } logit \geq 0 \\ 0 & \text{αλλιώς} \end{cases} \tag{3.35}$$

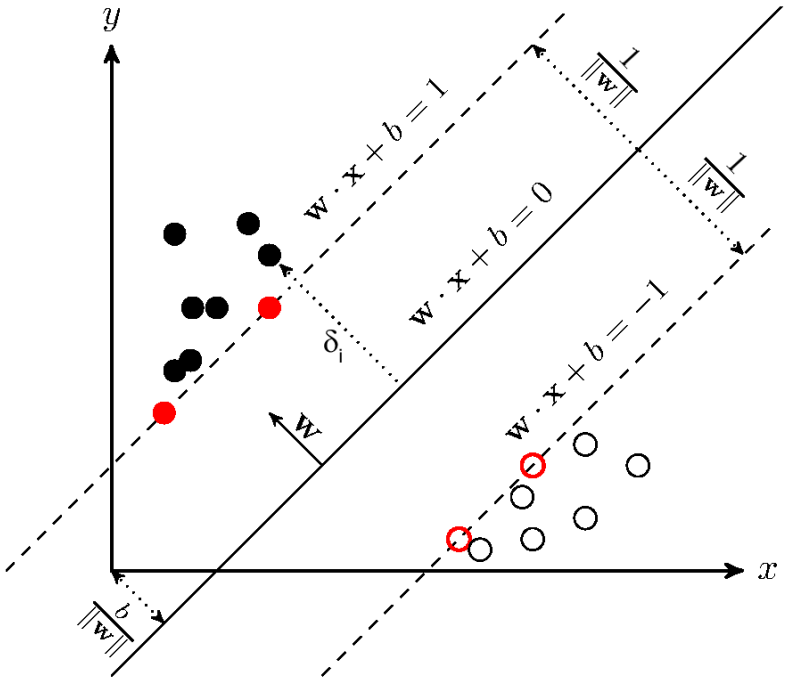
3.6.2 Support Vector Machines

Οι *μηχανές διανυσμάτων υποστήριξης* (support vector machines - SVM) είναι μια μέθοδος ταξινόμησης που προσπαθεί να βρει ένα υπερεπίπεδο ανάμεσα στα σημεία του χώρου, το οποίο θα χωρίζει όλα τα δείγματα σε δύο μη-επικαλυπτόμενες περιοχές [Evge01]. Για το σκοπό αυτό, πρέπει να βρεθεί διάνυσμα $\mathbf{w}$ τέτοιο ώστε να ισχύει $\mathbf{w}^\top x + b = 0$. Σε αυτή την περίπτωση, το i-οστό δείγμα των δεδομένων ταξινομείται στις δύο κλάσεις σύμφωνα με την Εξίσωση 3.36

$$\hat{y} = \begin{cases} +1 & \text{εάν } \mathbf{w}^\top x^{(i)} \geq 1 \\ -1 & \text{εάν } \mathbf{w}^\top x^{(i)} \leq -1 \end{cases} \tag{3.36}$$

Στη γενική περίπτωση υπάρχουν πολλά υπερεπίπεδα που μπορούν να διαχωρίσουν το χώρο, όταν τα δεδομένα είναι γραμμικώς διαχωρίσιμα. Τα SVM αναζητούν το υπερεπίπεδο

που επιτυγχάνει το βέλτιστο δυνατό *περιθώριο* (margin). Ως περιθώριο ορίζεται η απόσταση του πιο κοντινού σημείου μιας κλάσης από το υπερεπίπεδο που διαχωρίζει τις δύο αυτές κλάσεις.



Σχήμα 3.7: Γραφική απεικόνιση του προβλήματος βελτιστοποίησης των SVM

Στο Σχήμα 3.7 παρατηρούμε πως το διάνυσμα των βαρών w είναι κάθετο στο υπερεπίπεδο που ψάχνουμε. Η βελτιστοποίηση έγκειται στον εντοπισμό των κοντινότερων δειγμάτων από τις δύο κλάσεις σε σχέση με το διαχωριστικό επίπεδο και στη συνέχεια στην μεγιστοποίηση της απόστασης τους (Εξίσωση 3.37)

$$\begin{aligned} w_0 + w^\top x_{\{+1\}} &= 1 \\ w_0 + w^\top x_{\{-1\}} &= -1 \end{aligned} \tag{3.37}$$

Αφαιρώντας κατά μέλη και κανονικοποιώντας με τη νόρμα 2 προκύπτει (Εξίσωση 3.38)

$$\frac{w^\top (x_{\{+1\}} - x_{\{-1\}})}{\|w\|} = \frac{2}{\|w\|} \tag{3.38}$$

Άρα η μεγιστοποίηση του περιθωρίου ισοδυναμεί με την ελαχιστοποίηση της νόρμας των βαρών, οπότε το τελικό πρόβλημα ορίζεται ως εξής (Εξίσωση 3.39)

$$\begin{aligned} &\min \|w\|_2 \\ s.t. \quad &(w_0 + w^\top x^{(i)})\hat{y}^{(i)} \geq 1 \end{aligned} \tag{3.39}$$

δηλαδή όλα τα δείγματα να κατανέμονται σωστά, ισοδύναμα στο σωστό υπερεπίπεδο.

Στην περίπτωση που τα δεδομένα μας δεν είναι γραμμικώς διαχωρίσιμα, τότε η μέθοδος προσπαθεί να βρει μια κατάλληλη απεικόνιση από το χώρο στον οποίο βρίσκονται τα δεδομένα σε ένα χώρο μεγαλύτερης διάστασης, στον οποίο τα δεδομένα μας θα είναι πλέον

γραμμικά διαχωρίσιμα ή εναλλακτικά μπορεί να "ανέχετε" ορισμένα σφάλματα στην ταξινόμηση.

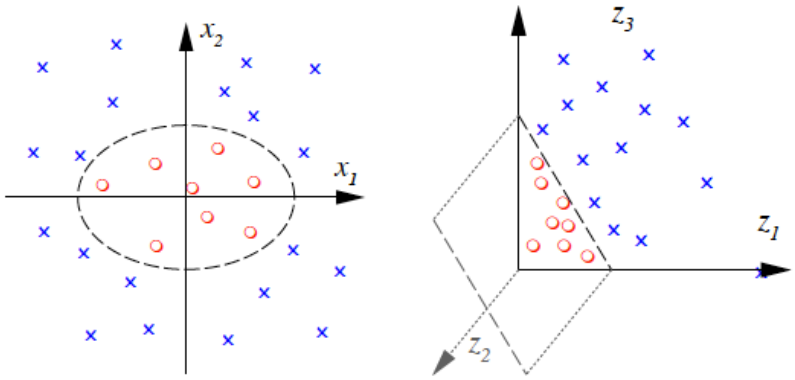
Οι *μεταβλητές χαλάρωσης* (ξ_i) είναι θετικοί αριθμοί και επιτρέπουν πλέον στο γραμμικό πρόβλημα, όπως το ορίσαμε, να συγκλίνει, κάνοντας ανεκτά σφάλματα στην ταξινόμηση. Οι μεταβλητές χαλάρωσης προστίθενται στο πρόβλημα ελαχιστοποίησης, όπως φαίνεται στην Εξίσωση 3.40

$$\begin{aligned} &\min_{w,\xi_i} \|w\|_2 + C\sum_i \xi_i \\ s.t. \quad &(w_0 + w^\top x^{(i)})\hat{y}^{(i)} \geq 1 - \xi_i \end{aligned} \tag{3.40}$$

όπου η μεταβλητή C αποτελεί υπερπαράμετρο της μεθόδου και ελέγχει το βάρος της ποινής από τη λάθος ταξινόμηση. Μεγάλες τιμές της μεταβλητής C δεν επιτρέπουν πολλές λάθος ταξινομήσεις και κατ΄επέκταση έχουμε μεγαλύτερη μεροληψία ενω μικρές τιμές είναι λιγότερες αυστηρές σε σφάλματα, οδηγώντας σε μεγαλύτερη διασπορά.

Ένας εναλλακτικός τρόπος που ενισχύει την προσπάθεια εύρεσης του βέλτιστου υπερεπιπέδου είναι η εφαρμογή μιας απεικόνισης ϕ στο διάνυσμα χαρακτηριστικών, ο οποίος απεικονίζει τα δεδομένα σε ένα χώρο μεγαλύτερων διαστάσεων που πιθανότατα θα είναι γραμμικώς διαχωρίσιμα Στο παράδειγμα του Σχήματος 3.8 βλέπουμε πως με ένα μη-γραμμικό μετασχη-

$$\Phi : R^2 \to R^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{(2)}x_1x_2, x_2^2)$$



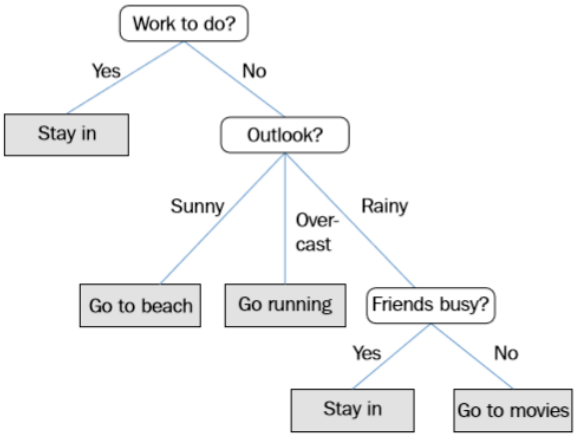
Σχήμα 3.8: Ένα παράδειγμα SVM με τη μέθοδο των πυρήνων

ματισμό τα δεδομένα από χώρο δύο διαστάσεων προβάλλονται σε χώρο τριών διαστάσεων, όπου είναι γραμμικώς διαχωρίσιμα. Οι πιο γνωστές απεικονίσεις που χρησιμοποιούνται είναι ο ακτινικός (Εξίσωση 3.41), ο γραμμικός (Εξίσωση 3.42) και ο πολυωνυμικός πυρήνας (Εξίσωση 3.43).

$$K(x^{(i)}, x^{(j)}) = \exp -\frac{\left\|x^{(i)} - x^{(j)}\right\|^2}{2\sigma^2} \tag{3.41}$$
$$K(x^{(i)}, x^{(j)}) = x^{(i)}x^{(j)} + \gamma \tag{3.42}$$
$$K(x^{(i)}, x^{(j)}) = (\alpha x^{(i)}x^{(j)} + \gamma)^d \tag{3.43}$$

3.6.3 Δένδρα Απόφασης

Ένα *δένδρο απόφασης* [Rasc15] είναι ένας απλός ταξινομητής, ο οποίος μπορεί να δώσει πολύ ικανοποιητικά αποτελέσματα. Αρχικά, το δέντρο απόφασης τροφοδοτείται με όλο το σύνολο δεδομένων. Σε κάθε επίπεδο, κατά τη φάση της εκπαίδευσης τα δεδομένα διασπώνται σε διαφορετικά υποσύνολα, μέχρι να καταλήξουμε στον αντικειμενικό σκοπό του προβλήματος. Δηλαδή θέλουμε, διαχωρίζοντας τα δεδομένα σε κόμβους, να φτάσουμε στα φύλλα όπου λαμβάνεται η τελική απόφαση για την ταξινόμηση του κάθε δείγματος σε μια κλάση.



Σχήμα 3.9: Ένα παράδειγμα ενός δένδρου απόφασης

Στο Σχήμα 3.9 βλέπουμε ένα παράδειγμα εκπαίδευσης ενός δένδρου απόφασης. Σε κάθε επίπεδο, με διαδοχικές ερωτήσεις, το δέντρο προσπαθεί να καταλήξει στα φύλλα, όπου ο άνθρωπός αποφασίζει ποια μορφή δραστηριότητας να επιλέξει ανάλογα με τις συνθήκες που επικρατούν.

Σε κάθε προσπάθεια ο αλγόριθμος ταξινόμησης ψάχνει να βρει το χαρακτηριστικό που μπορεί να προσδώσει τη μεγαλύτερη δυνατή πληροφόρηση, δηλαδή θέλει να μεγιστοποιήσει την Εξίσωση 3.44

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} \cdot I(D_j) \tag{3.44}$$

όπου το σύμβολο f συμβολίζει το χαρακτηριστικό για το οποίο γίνεται η διάσπαση στις m συνολικά κατηγορίες της μεταβλητής απόκρισης. D_p είναι το σύνολο δεδομένων του κόμβου που είναι ο γονέας με N_p το πλήθος στοιχεία και D_j είναι το σύνολο δεδομένων στο j-οστό κόμβο που έχει N_j και είναι παιδί του προηγούμενου κόμβου. Με το I αποτιμάται η εντροπία (Εξίσωση 3.45) ή ο δείκτης Gini (Εξίσωση 3.46)

$$I_{entropy} = -\sum_{i=1}^{m} p(i|t) \cdot \log_2(p(i|t)) \tag{3.45}$$

$$I_{gini} = 1 - \sum_{i=1}^{m} p(i|t)^2 \tag{3.46}$$

όπου $p(i|t)$ είναι η πιθανότητα το i-οστό δείγμα του κόμβου t να ανήκει στην κλάση c.

3.6.4 Τυχαία Δάση

Ο ταξινομητής *τυχαίου δάσους* (random forest) [Rasc15] ανήκει στη γενική κατηγορία των *bagging* (bootstrap aggregating) μεθόδων. Στη μεθοδολογία αυτή έχουμε ένα απλό ταξινομητή με τον οποίο εκπαιδεύουμε το σύνολο δεδομένων μας

1. Διαλέγουμε ένα δείγμα με επανατοποθέτηση, μεγέθους ίσο με το αρχικό σύνολο δεδομένων.
2. Διαλέγουμε d χαρακτηριστικά χωρίς επανατοποθέτηση.
3. Εκπαιδεύουμε τα δέντρα που δημιουργούνται όπως περιγράφεται στη γενική θεωρία.
4. Επαναλαμβάνουμε τη διαδικασία για n γύρους.
5. Λαμβάνουμε τη εκτίμηση για κάθε δείγμα συνολικά από όλα τα δέντρα.

Η διαδικασία που μόλις περιγράψαμε πολλές φορές οδηγεί σε υπερεκπαίδευση, ειδικά αν επιλέξουμε πολύ μεγάλο πλήθος επαναλήψεων ή χαρακτηριστικών σε κάθε γύρο. Η επαναληπτική εκπαίδευση δέντρων με τη στοχαστικότητα όπως την ορίσαμε, μειώνει την πιθανότητα υπερεκπαίδευσης, δίνοντας καλύτερα αποτελέσματα. Όμως, πιο μοντέρνες τεχνικές που θα παρουσιάσουμε στη συνέχεια, οι οποίες στηρίζονται στην εκπαίδευση δέντρων, δίνουν ικανοποιητικότερα αποτελέσματα.

3.6.5 XGBoost

Ο ταξινομητής αυτός ανήκει στην κατηγορία των gradient boosting μεθόδων [Chen16]. Πρόκειται για μια επαναληπτική μεθοδολογία, όπου σε κάθε γύρο εκπαιδεύουμε έναν απλό ταξινομητή. Η διαφορά είναι πως εδώ σε κάθε βήμα εντός μιας επανάληψης κάθε ταξινομητής που είναι να εκπαιδευτεί προσπαθεί να μειώσει το σφάλμα, τροφοδοτώντας με λανθασμένες εκτιμήσεις από την ακριβώς προηγούμενη εκπαίδευση που προέκυψε από τον απλό ταξινομητή.

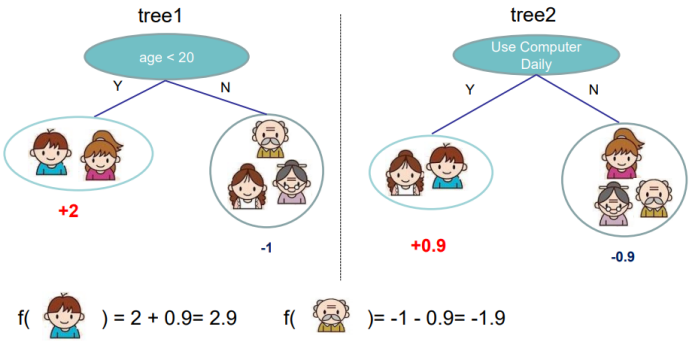
Πιο συγκεκριμένα, έχουμε μια συνάρτηση που εκτιμά την απώλεια και προσπαθούμε να την ελαχιστοποιήσουμε. Ο τρόπος που αυτό επιτυγχάνεται είναι μέσω της κίνησης προς την αντίθετη κατεύθυνση της κλίσης, δηλαδή της παραγώγου της συνάρτησης απώλειας ως προς τις παραμέτρους του μοντέλου. Σε ένα πρόβλημα ταξινόμησης, η συνάρτηση απώλειας μπορεί να έχει τη μορφή της Εξίσωσης 3.47

$$\mathcal{L} = \sum_{i=1}^{n} y_i \ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i) \ln\left(1 + e^{+\hat{y}_i}\right) \tag{3.47}$$

όπου $\hat{y}_i$ η πρόβλεψη του ταξινομητή για το i-οστό δείγμα εισόδου και y_i η πραγματική ετικέτα. Στην περίπτωση *συλλογικών* (ensemble) μεθόδων, η ολική έξοδος του συστήματος προκύπτει από το συνδυασμό των k επιμέρους απλών ταξινομητών (Εξίσωση 3.48)

$$\hat{y}_i = \sum_{i=1}^{K} \hat{y}_i^{(k)} \tag{3.48}$$

Στο Σχήμα 3.10 παρουσιάζεται το παράδειγμα της εκπαίδευσης ενός δένδρου και τις τιμές που παίρνουν δύο δείγματα. Βλέπουμε ξεκάθαρα πλέον πως αντί για βάρη εκπαιδεύουμε δένδρα απόφασης και λαμβάνουμε υπόψη όλες τις προηγούμενες εκτιμήσεις. Στην περίπτωση



Σχήμα 3.10: Σχηματικά η εκπαίδευση δύο συλλογικών δένδρων απόφασης

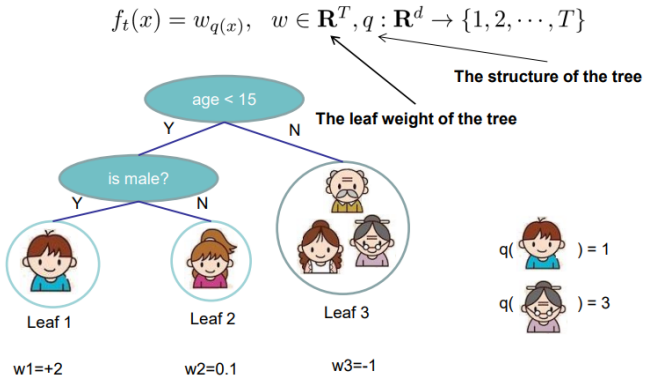
των gradient boosting αλγορίθμων έχουμε έναν απλό ταξινομητή και σε κάθε επανάληψη εκπαιδεύουμε έναν επόμενο ώστε να μειώσουμε το σφάλμα (Εξισώσεις 3.49-3.52).

$$\hat{y}_i = 0 \tag{3.49}$$
$$\hat{y}_i^1 = f_1(x_i) = \hat{y}_i + f_1(x_i) \tag{3.50}$$
$$\hat{y}_i^2 = f_2(x_i) = \hat{y}_i^1 + f_2(x_i) \tag{3.51}$$
$$\cdots\cdots\cdots$$
$$\hat{y}_i^K = \sum_{j=1}^{K} f_j(x_i) = \hat{y}_i^{K-1} + f_K(x_i) \tag{3.52}$$

Οι συναρτήσεις f_i ανήκουν στο χώρο που περιγράφεται από την Εξίσωση 3.53

$$F = \{f(x) = w_{q(x)}\}\, q : \mathbb{R}^m \to T, w \in \mathbb{R}^T \tag{3.53}$$

όπου w είναι τα βάρη των φύλλων, q μια συνάρτηση απεικόνισης των δειγμάτων στα φύλλα και T το πλήθος των φύλλων. Στο Σχήμα 3.11 βλέπουμε ένα δέντρο με 3 φύλλα. Για να



Σχήμα 3.11: Ένα παράδειγμα σε ένα δέντρο με 3 φύλλα

διαλέξουμε τη συνάρτηση απεικόνισης εργαζόμαστε ως εξής: έστω $\hat{y}_i^t$ η πρόβλεψη για την i-οστή παρατήρηση στην t-οστή επανάληψη. Η εκτίμηση του σφάλματος δίνεται από την Εξίσωση 3.54

$$\sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i^t) = \sum_{i=1}^{n} \mathcal{L}(y_i, \hat{y}_i^{t-1} + f_t(x_i)) \tag{3.54}$$

Συμβολίζοντας με g_i, h_i την πρώτη και τη δεύτερη μερική παράγωγό ως προς $\hat{y}^{t-1}$ αντίστοιχα (Εξίσωση 3.55)

$$g_i = \frac{\partial \mathcal{L}(y, \hat{y}^{t-1})}{\partial \hat{y}^{t-1}}, \quad h_i = \frac{\partial^2 \mathcal{L}(y, \hat{y}^{t-1})}{\partial (\hat{y}^{t-1})^2} \tag{3.55}$$

η καινούργια εκτίμηση σφάλματος τη χρονική στιγμή t χρησιμοποιώντας τον τύπο του Taylor και αγνοώντας τις υπόλοιπες σταθερές είναι (Εξίσωση 3.56)

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] \tag{3.56}$$

Ορίζουμε την πολυπλοκότητα του δένδρου σύμφωνα με την Εξίσωση 3.57

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \tag{3.57}$$

όπου γ είναι μια σταθερά και T είναι το πλήθος των φύλλων όπως αναφέραμε νωρίτερα. Παρατηρούμε πως στην πολυπλοκότητα προστίθεται και η L_2 νόρμα των βαρών όπως στην παλινδρόμηση Ridge. Ορίσουμε ως I_j το σύνολο των δειγμάτων που ανήκουν στο φύλλο j. Σύμφωνα με τα παραπάνω, η προσεγγιστική πλέον εκτίμηση για τη συνάρτηση σφάλματος μπορεί να γραφεί ως (Εξίσωση 3.58)

$$\begin{aligned} \mathcal{L}^{(t)} &= \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \\ &= \sum_{i=1}^{n} [w_{q(x_i)} g_i + \frac{1}{2} h_i w_{q(x_i)}^2(x_i)] + \gamma T + \lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) {w_j}^2] + \gamma T \end{aligned} \tag{3.58}$$

που είναι το άθροισμα T τετραγωνικών συναρτήσεων. Στην περίπτωση των μονοδιάστατων τετραγωνικών συναρτήσεων $Gx + \frac{1}{2} H x^2$, η ελάχιστη τιμή της συνάρτησης είναι -$\frac{1}{2} \frac{G^2}{H}$ για $H > 0$ και το ελάχιστο επιτυγχάνεται όταν $x = -\frac{G}{H}$. Θέτοντας $G_i = \sum_{i \in I_j} g_i$ και $H_j = \sum_{i \in I_j} h_i$ και γνωρίζοντας πλέον τη δομή του δέντρου που θα εκπαιδεύσουμε, τα βέλτιστα βάρη είναι (Εξίσωση 3.59)

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{3.59}$$

και η εκτιμώμενη συνάρτηση σφάλματος για το δένδρο (Εξίσωση 3.60)

$$\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{3.60}$$

Υπάρχουν πολλά δένδρα που μπορούν να δημιουργηθούν. Για την εύρεση του καλύτερου δυνατού, μπορούμε να ξεκινήσουμε μια εξαντλητική διαδικασία αναζήτησης. Αρχικά, παίρνουμε ένα δέντρο χωρίς φύλλα και αρχίζουμε να προσθέτουμε βάθος. Σε κάθε καινούργιο διαχωρισμό, η μεταβολή της συνάρτησης απώλειας είναι (Εξίσωση 3.61)

$$\frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_R + G_L)^2}{H_R H_L + \lambda} \right] - \gamma \tag{3.61}$$

και αναζητούμε το διαχωρισμό με τη μεγαλύτερη απώλεια. Μπορούμε να σταματήσουμε στο επίπεδο που η προσθήκη και άλλου επιπέδου δε μειώνει τη συνάρτηση σφάλματος ή μπορούμε να αναπτύξουμε το δέντρο μέχρι το μέγιστο δυνατό βάθος και ύστερα να μειώσουμε

Σχήμα 3.12: Οι δύο κύριες στρατηγικές ανάπτυξης των δέντρων

το βάθος διαγράφοντας τα φύλλα με αρνητικό κέρδος (Σχήμα 3.12). Οι περισσότεροι αλγόριθμοι πλέον χρησιμοποιούν τη μέθοδο του ιστογράμματος, όπου για κάθε χαρακτηριστικό δημιουργείται ένα ιστόγραμμα με σταθερό αριθμό κάδων k (bins) και αναζητείται σε αυτό το υποσύνολο ο καλύτερος διαχωρισμός.

Μια σημαντική βελτίωση είναι επίσης η χρήση του αλγορίθμου LightGBM [Ke17], ο οποίος κρατά κατά την εκπαίδευση τα δείγματα που έχουν τη μεγαλύτερη κλίση και δειγματοληπτεί τυχαία τα δείγματα με μικρή κλίση. Η βασική θεώρηση πίσω από αυτή τη τεχνική είναι πως δείγματα με μικρή κλίση είναι πιθανότατα ταξινομημένα σωστά ενώ αντίθετα αυτά με μεγάλη κλίση παρουσιάζουν μεταβλητότητα και μπορούν να οδηγήσουν συνεχώς σε καλύτερη εκπαίδευση.

3.6.6 Πολυεπίπεδα Perceptron

Θα περιγράψουμε αρχικά τον αλγοριθμό Perceptron και θα καταλήξουμε στη λογική των σύγχρονων νευρωνικών δικτύων [Good16, Rasc15]. Ας υποθέσουμε πως θέλουμε να ταξινομήσουμε ένα δείγμα παρατηρήσεων όπου παρατηρούνται δύο κλάσεις τις οποίες θα συμβολίζουμε με $\{-1, 1\}$. Ξεκινάμε με ένα διάνυσμα βαρών w που το αρχικοποιούμε με μικρές τυχαίες τιμές. Η είσοδος για τη ταξινόμηση θα αποτελέσει ο γραμμικός συνδυασμός $z = w_1x_1 + \ldots + x_n$, όπου $x_i, i = 1, \ldots, n$ είναι το διάνυσμα των χαρακτηριστικών ενός τυχαίου δείγματος από τις παρατηρήσεις. Η έξοδος του ταξινομητή δίνεται στην Εξίσωση 3.62

$$\phi(z) = \begin{cases} +1 & \text{εάν } z \geq \theta \\ -1 & \text{διαφορετικά} \end{cases} \tag{3.62}$$

Η ποσότητα θ είναι γνωστή και ως *πόλωση* (bias) και εισάγεται στην είσοδο του Perceptron σύμφωνα με την Εξίσωση 3.63

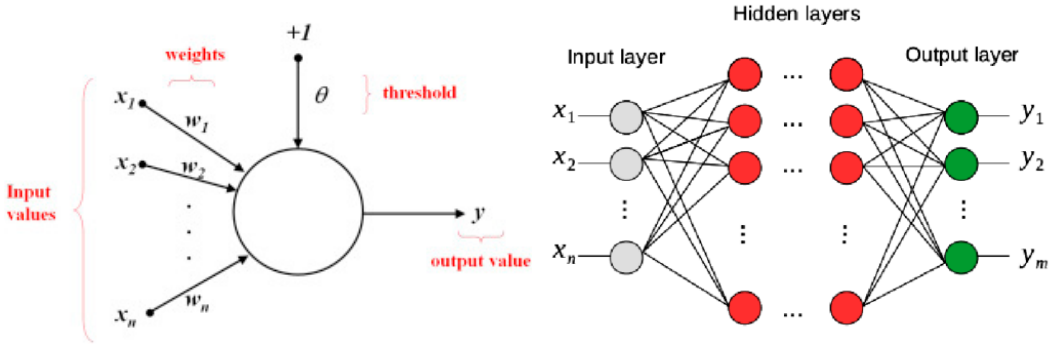
$$z = -\theta + w_1x_1 + ... + x_n = w_0x_0 + w_1x_1 + ... + x_n \tag{3.63}$$

Η εκπαίδευση του ταξινομητή ή ισοδύναμα η εύρεση των βέλτιστων βαρών w γίνεται μέσω μιας επαναληπτικής διαδικασίας. Αρχικά, θέτουμε στα βάρη τυχαίες τιμές και για κάθε δείγμα $x^{(i)}$ του συνόλου δεδομένων μας, εκτιμούμε την κλάση $\hat{y}$ στην οποία ανήκει. Στη συνέχεια μεταβάλλουμε το διάνυσμα των βαρών κατά παράγοντα Δw_i (Εξίσωση 3.64)

$$\Delta w_i := \eta(y^{(i)} - \hat{y}^{(i)})x_j^{(i)} \tag{3.64}$$

Με η συμβολίζεται ο *ρυθμός εκπαίδευσης* (learning rate), ο οποίος κυμαίνεται στο $(0, 1)$. Αποτελεί βασική παράμετρο εκπαίδευσης ενός νευρωνικού δικτύου και ρυθμίζει την ευαισθησία στις μεταβολές των βαρών. Μεγάλες τιμές κάνουν τα βάρη να παρουσιάζουν μεγάλες μεταβολές και τη διαδικασία εκπαίδευσης μη σταθερή, ενώ πολύ μικρές τιμές δε βοηθούν

στη γρήγορη σύγκλιση. Ο αλγόριθμος Perceptron τελικά συγκλίνει μόνο στην περίπτωση που οι κλάσεις των δειγμάτων μας είναι γραμμικά διαχωρίσιμες, συνθήκη που δεν ισχύει απαραίτητα στην πράξη.



(a) Perceptron (b) Βαθύτερες αρχιτεκτονικές

Σχήμα 3.13: Απο το Perceptron σε πιο βαθιές αρχιτεκτονικές

Για να ξεπεραστεί η απαίτηση της γραμμικής διαχωρισιμότητας τον κλάσεων των δεδομένων εισόδων, μπορούν να προστεθούν στην αρχιτεκτονική του νευρωνικού δικτύου επιπρόσθετα επίπεδα μεταξύ εισόδου και εξόδου, τα οποία είναι γνωστά ως *"κρυφά"* (hidden) επίπεδα (Σχήμα 3.13). Οι *συναρτήσεις ενεργοποίησης* (activation functions) των κρυφών επιπέδων συνήθως είναι η σιγμοειδής, η υπερβολική εφαπτομένη και η ημι-γραμμική. Ως συνάρτηση απώλειας συνήθως επιλέγεται αυτή της Εξίσωσης 3.65

$$\mathcal{L} = \mathbb{1}_{y=1} \cdot \log(p) + (1 - \mathbb{1}_{y=1}) \cdot \log(1 - p) \tag{3.65}$$

όπου p η πιθανότητα το δείγμα εισόδου να ανήκει στην κλάση $+1$.

Η ενημέρωση των βαρών γίνεται με τον κανόνα της *προς τα πίσω διάδοσης του σφάλματος* (back propagation), η γενική μορφή του οποίου αναγράφεται στην Εξίσωση 3.66

$$\Delta w_{ij}^{(l)} = -\eta \frac{\partial J}{\partial w_{ij}^{(l)}} \tag{3.66}$$

όπου $w_{ij}^{(l)}$ το βάρος σύνδεσης του νευρώνα j του στρώματος $l-1$ με το νευρώνα i του στρώματος l. Ο κανόνας Δέλτα είναι γνωστός και ως μέθοδος *κατάβασης κλίσης* (gradient descent), που περιγράφεται στην Εξίσωση 3.67

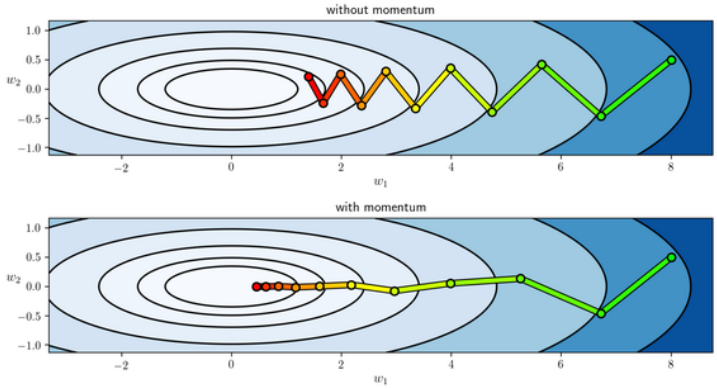
$$w(t+1) = w(t) - \eta \nabla J(w) \tag{3.67}$$

Πιο αποτελεσματικός αλγόριθμος ενημέρωσης των βαρών, ωστόσο, είναι ο Adam [King14], ο οποίος χρησιμοποιήθηκε και στο πλαίσιο της παρούσας διπλωματικής εργασίας.

Για να μειωθούν οι ταλαντώσεις και να συγκλίνει ευσταθώς στο ελάχιστο (Σχήμα 3.14) απαιτείται η εισαγωγή της έννοιας της ορμής (Εξίσωση 3.68)

$$u_t = \gamma u_{t-1} + \eta \nabla J(w) \tag{3.68}$$

όπου με η συμβολίζεται ο ρυθμός μάθησης και με γ είναι το ποσοστό της πληροφορίας που κρατάμε από τις προηγούμενες ενεργοποιήσεις του νευρώνα. Ο αλγόριθμος Adam προσαρμόζει την πρώτη και τη δεύτερη παράγωγο της κλίσης κάθε χρονική στιγμή. Έστω m_t, v_t οι



Σχήμα 3.14: Σχηματική αναπαράσταση της σύγκλισης με και χωρίς ορμή

πρώτες και οι δεύτερες παράγωγοι, οι οποίες υπολογίζονται με κινούμενο μέσο και ανανεώνονται με τα στατιστικά που αφορούν το δείγμα τη χρονική στιγμή t (Εξισώσεις 3.69-3.70)

$$m_t = \beta_1 * m_{t-1} + (1 - \beta_1) * \nabla J(w_t) \tag{3.69}$$

$$u_t = \beta_2 * u_{t-1} + (1 - \beta_2) * \nabla J(w_t)^2 \tag{3.70}$$

Επειδή τα διανύσματα u_t, v_t τείνουν προς το μηδέν, εισάγονται οι διορθώσεις της Εξίσωσης 3.71

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{u}_t = \frac{u_t}{1 - \beta_2^t} \tag{3.71}$$

Οι σταθερές β_1, β_2 είναι συνήθως πολύ κοντά στη μονάδα και έτσι η παραγόμενη έξοδος γίνεται (Εξίσωση 3.72)

$$w_t = w_{t-1} - \frac{\eta}{\sqrt{\hat{u}_{t-1}} + \epsilon} \hat{m}_t \tag{3.72}$$

με τη σταθερά ϵ να λαμβάνει πολύ μικρή τιμή.

3.6.7 Αναδρομικά Νευρωνικά Δίκτυα

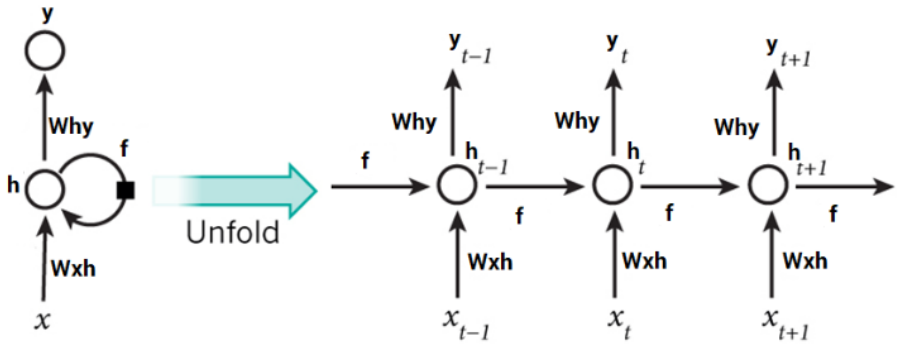
Στα πολυεπίπεδα Perceptron γίνεται η υπόθεση πως δεν υπάρχουν εξαρτήσεις μεταξύ των δειγμάτων που χρησιμοποιούνται στην εκπαίδευση. Το δίκτυο τροφοδοτείται με τα δείγματα από τα οποία υπολογίζεται το σφάλμα και διαδίδεται προς τα πίσω για τη διόρθωση των βαρών. Ωστόσο αυτή η μορφή εκπαίδευσης δεν είναι ιδανική, αν τα δεδομένα έχουν ακολουθιακή δομή. Για αυτές τις περιπτώσεις καταλληλότερα είναι τα *αναδρομικά νευρωνικά δίκτυα* (recurrent neural networks) [Good16] (Σχήμα 3.15), τα οποία αποτελούν επέκταση των Perceptrons με την προσθήκη *βρόχων ανάδρασης* (feedback loops). Σε αυτά τα δίκτυα, τα δείγμα εισάγονται σειριακά, και για κάθε είσοδο παράγεται μια έξοδος που χρησιμοποιείται ως είσοδος για την επόμενη πρόβλεψη.

Την χρονική στιγμή t η είσοδος στο αναδρομικό δίκτυο είναι x_t. Η κρυφή κατάσταση θα συμβολίζεται με h_t. Η είσοδος στο κρυφό επίπεδο εξαρτάται και από την κρυφή κατάσταση του προηγούμενου βήματος αλλά και από την είσοδο την ίδια χρονική στιγμή (Εξίσωση 3.73)

$$h_t = f h_{t-1} + W_{xh} * x_t \tag{3.73}$$

ενώ η έξοδος υπολογίζεται σύμφωνα με την Εξίσωση 3.74

$$y_t = W_{hy} * h_t \tag{3.74}$$



Σχήμα 3.15: Ισοδυναμία μεταξύ αναδρομικών και κλασικών νευρωνικών δικτύων

Τα W_{xh}, W_{hy}, f αποτελούν παραμέτρους του μοντέλου, οι οποίες "μαθαίνονται" κατά τη διάρκεια της εκπαίδευσης, όπως και στα κλασικά δίκτυα πρόσθιας τροφοδότησης. Μια πρώτη σημαντική παρατήρηση έγκειται στο γεγονός πως αυτές οι μεταβλητές δεν φαίνεται να έχουν εξάρτηση από το χρόνο. Αυτό συμβαίνει επειδή θέλουμε το μοντέλο μας να γενικεύει και να μη μαθαίνει μόνο το συγκεκριμένο σύνολο δεδομένων. Για παράδειγμα, αν το σύνολο δεδομένων είναι ένα κείμενο όπου υπάρχει η φράση *χτες έφαγα μια σοκολάτα* θα θέλαμε το νευρωνικό να αναπαράγει και τη φράση *έφαγα μια σοκολάτα χθες*. Υπάρχει όμως κάποιο πρόβλημα με το δίκτυο που μόλις παρουσιάσαμε. Επειδή τα βάρη θα ανανεώνονται και πάλι με τη διάδοση σφάλματος, χρειάζεται να μελετήσουμε τις παραγώγους και να εντοπίσουμε τυχόν προβλήματα. Το συνολικό σφάλμα μέχρι τη χρονική στιγμή t είναι $E = \sum_{i=1}^{t} E_t$, δηλαδή το άθροισμα των σφαλμάτων όλων των προηγούμενων επαναλήψεων. Αν συμβολίσουμε με θ οποιαδήποτε από τις παραμέτρους που εμφανίστηκαν πριν στο μοντέλο, τότε

$$\frac{\partial E}{\partial \theta} = \sum_{i=1}^{t} \frac{\partial E_i}{\partial \theta}, \quad \frac{\partial E_t}{\partial \theta} = \sum_{1 \le k \le t}^{t} \frac{\partial E_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \frac{\partial h_t}{\partial h_i} \frac{\partial h_i}{\partial \theta}, \quad \frac{\partial h_{t+1}}{\partial h_t} = f, \quad \frac{\partial h_t}{\partial h_i} = \prod_{1 \le k \le t-1} \frac{\partial x_{k+1}}{\partial x_k}$$

και για μεγάλες χρονικές εξαρτήσεις παρατηρούμε πως μπορεί η παράγωγος να τείνει στο άπειρο ή ακόμα και να μηδενίζεται [Pasc12].

Λύση σε αυτό το πρόβλημα δίνουν τα δίκτυα LSTM (long short-term memory networks) [Stau19], που αποτελούνται από υποδίκτυα που συνδέονται μεταξύ τους αναδρομικά (Σχήμα 3.16) και επιτελούν τρεις λειτουργίες: (i) Εισόδου (ii) Εξόδου (iii) Επιλεκτικής συγκράτησης

Το δίκτυο LSTM εισάγει, μια νέα παράμετρο την *κατάσταση του κυττάρου* (cell state) C_t και κάθε λειτουργία του περιγράφεται από ένα πίνακα. Αρχικά, το δίκτυο δέχεται ως είσοδο το ζεύγος x_t, h_{t-1}, όπως και τα αναδρομικά δίκτυα που παρουσιάστηκαν παραπάνω, με το επίπεδο της επιλεκτικής συγκράτησης αποφασίζει τι ποσοστό της συνολικής πληροφορίας θα απορρίψει (Εξίσωση 3.75)

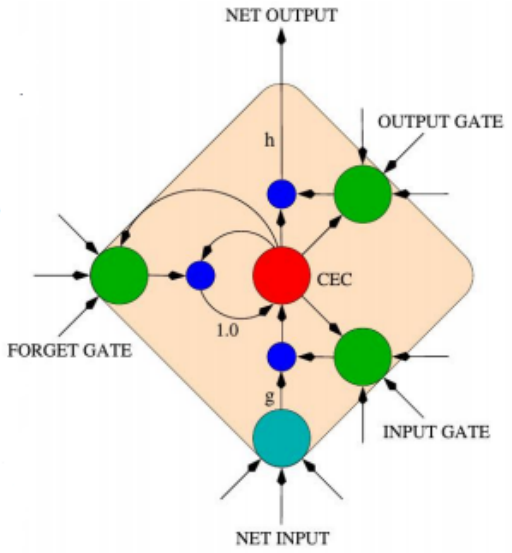
$$f_t = \sigma(W_f(h_{t-1}, x_t) + b_f) \tag{3.75}$$

Στο επόμενο βήμα, αποφασίζεται ποια νέα πληροφορία πρέπει να αποθηκευτεί και αυτό γίνεται στο στάδιο της εισόδου. Αρχικά, αποφασίζει το ποσοστό της πληροφορίας που πρέπει να αποθηκευτεί (Εξίσωση 3.76)

$$i_t = \sigma(W_i(h_{t-1}, x_t) + b_i) \tag{3.76}$$

και στη συνέχεια αποφασίζονται οι υποψήφιες τιμές προς ενημέρωση της κατάστασης του κυττάρου (Εξίσωση 3.77)

$$\tilde{C}_t = \tanh(W_c(h_{t-1}, x_t) + b_c) \tag{3.77}$$



Σχήμα 3.16: Σχηματική απεικόνιση ενός μοντέλου LSTM

Τέλος, καθορίζεται ποια θα είναι η παραγόμενη έξοδος. Η καινούργια κατάσταση του κυττάρου είναι (Εξίσωση 3.78)

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{3.78}$$

και οι εκτιμήσεις τη χρονική στιγμή t είναι (Εξίσωση 3.79)

$$o_t = \sigma(W_0(h_{t-1}, x_t) + b_0), \quad h_t = o_t \cdot \tanh C_t \tag{3.79}$$

Εδώ πλέον αντιμετωπίζεται το πρόβλημα της έκρηξης ή στασιμότητας των παραγώγων. Πιο συγκεκριμένα, αντικαθιστώντας στις προηγούμενες σχέσεις

$$\frac{\partial C_t}{\partial C_{t-1}} = \frac{\partial C_t}{\partial f_t}\frac{\partial f_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial C_{t-1}} + \frac{\partial C_t}{\partial i_t}\frac{\partial i_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial C_{t-1}} \tag{3.80}$$

$$+\frac{\partial C_t}{\partial \tilde{C}_t}\frac{\partial \tilde{C}_t}{\partial h_{t-1}}\frac{\partial h_{t-1}}{\partial C_{t-1}} + \frac{\partial C_t}{\partial C_{t-1}} \tag{3.81}$$

$$\frac{\partial C_t}{\partial C_{t-1}} = C_{t-1} \cdot \sigma'() \cdot W_f \cdot o_{t-1} \cdot \tanh' C_{t-1} \tag{3.82}$$

$$+\tilde{C}_t \cdot \sigma'() \cdot W_i \cdot o_{t-1} \cdot \tanh' C_{t-1} \tag{3.83}$$

$$+i_t \cdot \tanh'()W_c \cdot o_{t-1} \cdot \tanh' C_{t-1} + f_t \tag{3.84}$$

η διάδοση του σφάλματος γίνεται άθροισμα πολλών παραγόντων (Εξισώσεις 3.80-3.84).

Κεφάλαιο 4

Πειραματική διαδικασία και Αποτελέσματα

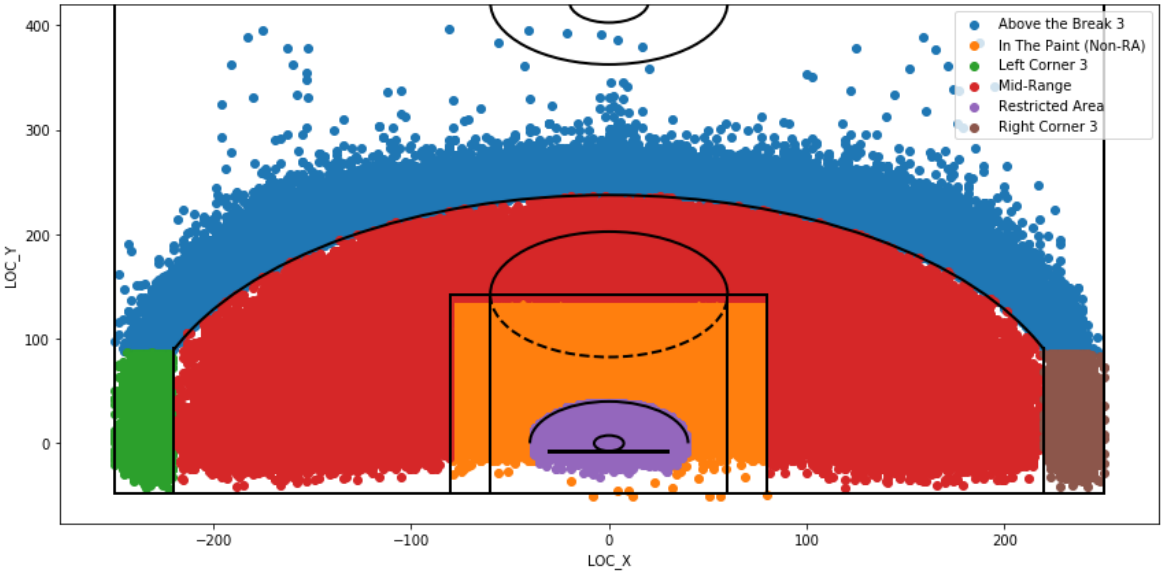
4.1 Το σύνολο των δεδομένων

Το σύνολο δεδομένων που χρησιμοποιήθηκε στο πειραματικό μέρος της παρούσας διπλωματικής προέρχεται από τις μετρήσεις που λήφθηκαν μέσω του συστήματος SportsVu [hwch15] για τις επιθετικές προσπάθειες στο NBA, τις σεζόν 2013-2014 και 2014-2015. Τα χαρακτηριστικά των δειγμάτων που περιέχονται στη συλλογή δεδομένων συνοψίζονται στα παρακάτω σημεία:

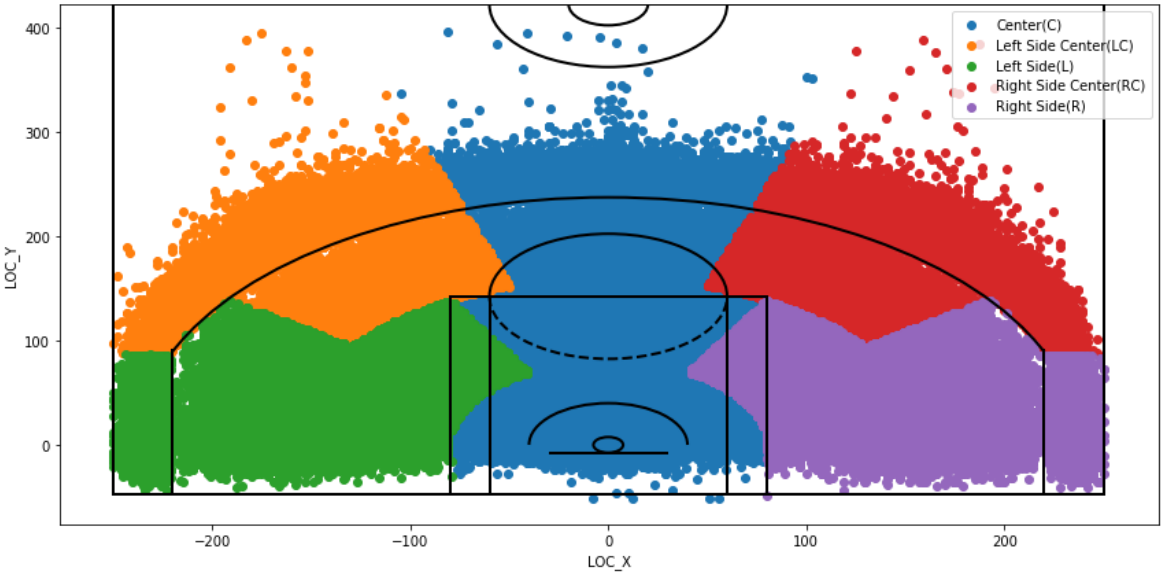
1. Ο παίκτης που αμύνεται στον παίκτη που πραγματοποιεί την επιθετική του προσπάθεια.
2. Η απόσταση μεταξύ του επιτιθέμενου και του αμυνόμενου. Η απόσταση μετριέται από τη στιγμή που ο παίκτης ο οποίος επιτίθεται απελευθερώνει τη μπάλα.
3. Η ζώνη που πραγματοποιείται η προσπάθεια.
4. Το είδος της επιθετικής προσπάθειας.
5. Η τοποθεσία της επιθετικής προσπάθειας σε όρους συντεταγμένων μέσα στο γήπεδο.
6. Ο χρόνος που έχει τη μπάλα στην κατοχή του και οι ντρίμπλες που προηγήθηκαν από τον επιτιθέμενο.
7. Ο χρόνος για τη λήξη της περιόδου.
8. Ο αριθμός της ατομικής επιθετικής προσπάθειας.
9. Η τελική διαφορά μεταξύ των ομάδων.

Συνολικά, υπάρχουν 336.351 παρατηρήσεις που αφορούν σχεδόν όλους τους αγώνες που πραγματοποιήθηκαν στην κανονική διάρκεια κατά τις σεζόν 2013-14 και 2014-15. Στη συνέχεια, θα προσπαθήσουμε να εξηγήσουμε τη διαθέσιμη πληροφορία με γραφήματα. Για να κατανοήσουμε τις διαφορετικές ζώνες του αγωνιστικού χώρου χρειάζεται να τις αναπαραστήσουμε σε ένα διάγραμμα (Σχήμα 4.1), όπου αναπαρίσταται κάθε προσπάθεια που έχει καταγραφεί στο σύνολο δεδομένων με διαφορετικό χρώμα ανάλογα με τη τοποθεσία του σουτ.

Στο Σχήμα 4.2 βλέπουμε τις περιοχές στις οποίες χωρίζεται ο χώρος επίθεσης με βάση τις διευθύνσεις επίθεσης. Για να μειώσουμε τον αριθμό των διαθέσιμων μεταβλητών, ενώνουμε τις πληροφοίες από τα παραπάνω γραφήματα, καταλήγοντας σε 12 μεταβλητές που περιγράφουν την τοποθεσία του σουτ, που απεικονίζονται με διαφορετικά χρώματα στο Σχήμα 4.3



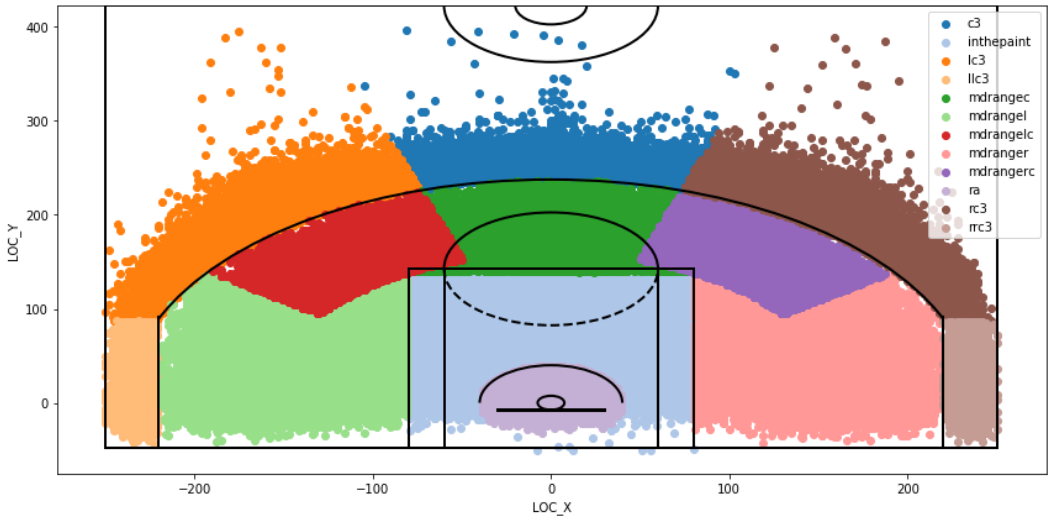
Σχήμα 4.1: Οι διαφορετικές ζώνες



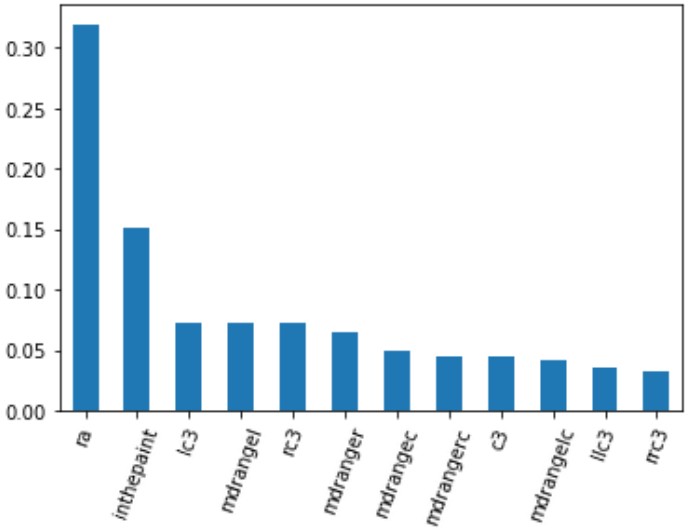
Σχήμα 4.2: Οι διαφορετικές διευθύνσεις επίθεσης

Σε αυτές τις ζώνες θα καταγράψουμε το ποσοστό ευστοχίας κάθε παίκτη. Σε κάθε προσπάθεια θα θεωρούμε πως το ποσοστό που εκφράζει την πιθανότητα ευστοχίας είναι το ποσοστό που έχει ο επιτιθέμενος στη ζώνη αυτή. Στο Σχήμα 4.4 απεικονίζεται η συχνότητα εκδήλωσης επίθεσης ανά ζώνη και στο Σχήμα 4.5 το ποσοστό ευστοχίας ανά ζώνη. Οι επιθέσεις που γίνονται από την κεντρικό άξονα έχουν κατά μέσο όρο μεγαλύτερο ποσοστό ευστοχίας (Σχήμα 4.6).

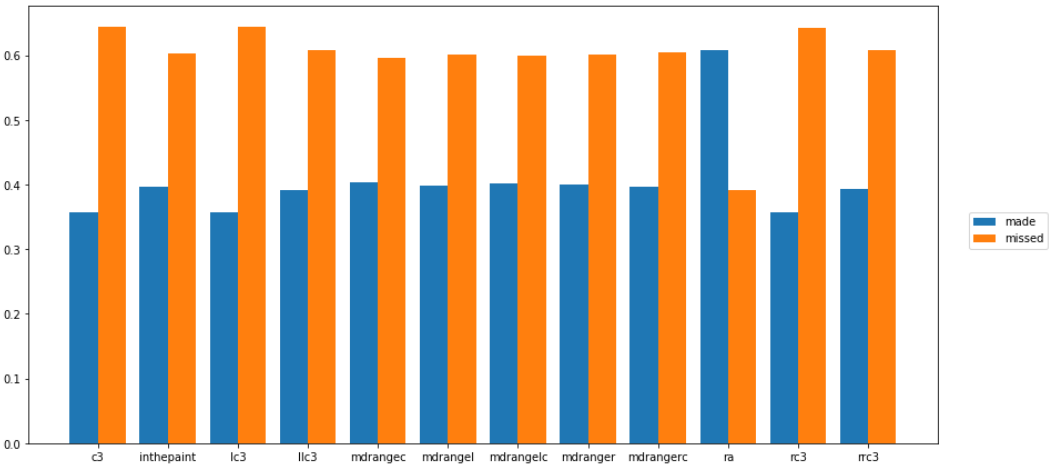
Για να δημιουργήσουμε το γράφημα που θα μας δείξει τον αναμενόμενο αριθμό πόντων για ένα τυχαίο σουτ θα πολλαπλασιάσουμε το ποσοστό ευστοχίας επί το συντελεστή των πόντων που αντιστοιχούν στην επιτυχημένη έκβαση. Τα δίποντα θα πολλαπλασιαστούν με συντελεστή 2 ενώ τα τρίποντα με 3. Από το Σχήμα 4.7 βλέπουμε πως οι προσπάθειες μέσα στη ρακέτα (restricted area) και όλες οι προσπάθειες από το τρίποντο έχουν αναμενόμενο αριθμό πόντων μεγαλύτερο του 1. Αντίθετα, όλα τα δίποντα πλην αυτών που πραγματοποιούνται



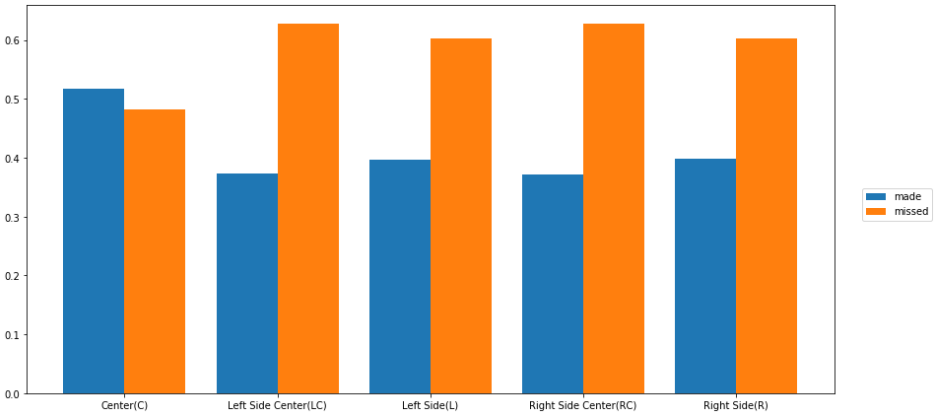
Σχήμα 4.3: Οι ζώνες που δημιουργήθηκαν στο πλαίσιο της διπλωματικής



Σχήμα 4.4: Συχνότητα εκδήλωσης επίθεσης ανά ζώνη

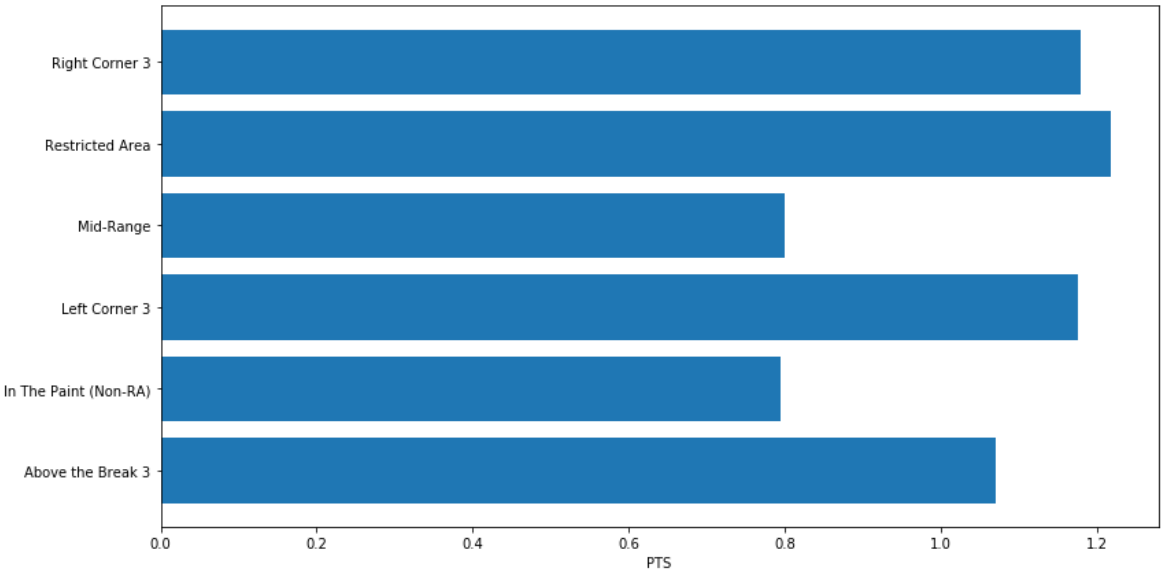


Σχήμα 4.5: Ποσοστό ευστοχίας ανά ζώνη



Σχήμα 4.6: Ποσοστό ευστοχίας ανά διεύθυνση

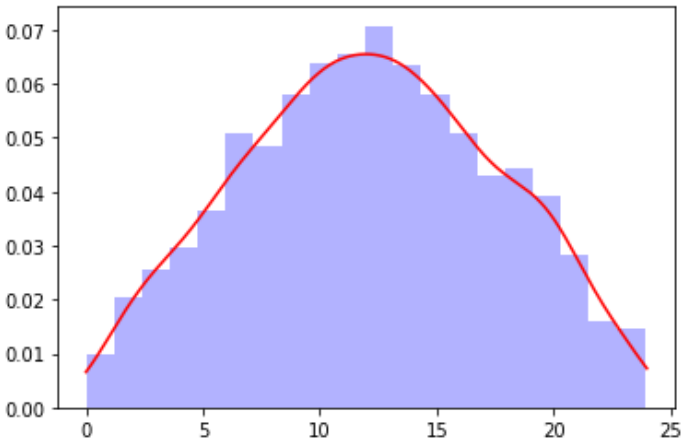
κοντά στη ρακέτα έχουν περίπου 0.8 εκτιμώμενο αριθμό πόντων. Αυτό μας δικαιολογεί την ολοένα και αυξανόμενη τάση των ομάδων να μην επιλέγουν τις προσπάθειες από μέση απόσταση αλλά να επιτίθενται από το τρίποντο και από κοντινές αποστάσεις. Θυμίζουμε πως τα δεδομένα μας είναι από τις σεζόν 2013-2014 και 2014-2015 όπου ξεκινάει η πρακτική αυτή.



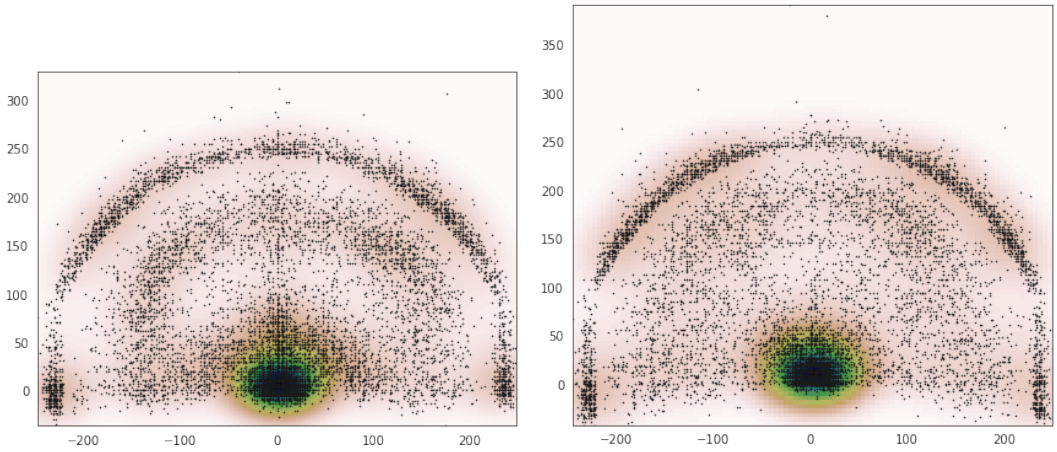
Σχήμα 4.7: Εκτιμώμενος αριθμός πόντων ανά ζώνη

Σημαντικό είναι να παρατηρήσουμε πως, αν και ο μέσος χρόνος εκδήλωσης της επιθετικής προσπάθειας είναι κοντά στα 12 δευτερόλεπτα, βλέπουμε μια ελαφρά τάση να εκδηλώνεται η προσπάθεια όλο και πιο νωρίς, στοιχείο που κορυφώνεται τα τελευταία χρόνια δημιουργώντας ένα έντονο ρυθμό στο παιχνίδι (Σχήμα 4.8). Η κόκκινη γραμμή συμβολίζει την εκτιμώμενη πυκνότητα πιθανότητας για την κατανομή της χρονικής εκδήλωσης της επιθετικής προσπάθειας. Επειδή η συλλογή δεδομένων έχει συνολικά περίπου 340 χιλιάδες εγγραφές και για τη μέθοδο των πυρήνων στους υπολογισμούς χρειάζεται να φορτωθούν στη μνήμη του υπολογιστή όλα τα δείγματα, θα κρατήσουμε μόνο το 1/5 του συνόλου και θα εφαρμόσουμε διασταυρούμενη επικύρωση εξαίρεσης ενός για την αναζήτηση της τιμής του εύρους που μεγιστοποιεί την πιθανοφάνεια.

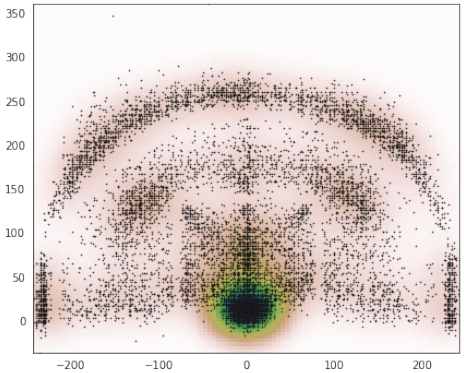
Στα γραφήματα που παρουσιάζονται στο Σχήμα 4.9 βλέπουμε τις επιθέσεις τριών ομάδων με την μέθοδο των πυρήνων σε δύο διαστάσεις. Των πρωταθλητών Golden State Warriors, οι



Σχήμα 4.8: Χρόνος εκτέλεσης της επιθετικής προσπάθειας



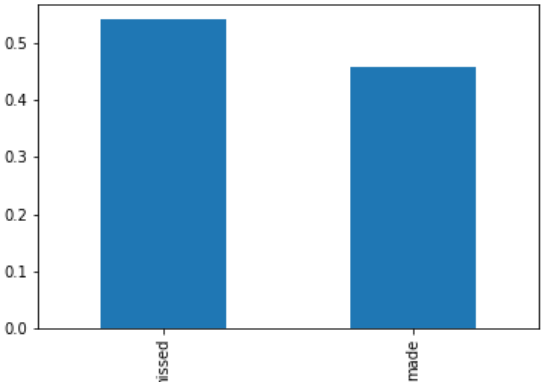
(a) San Antonio Spurs (b) New York Knicks



(c) Golden State Warriors

Σχήμα 4.9: Γραφήματα των επιθέσεων 3 ομάδων

οποίοι έχουν ένα πολύ γρήγορο τέμπο, μιας παραδοσιακής ομάδας (San Antonio Spurs), που στηρίζεται σε πιο αργό τέμπο παιχνιδιού και των ουραγών του πρωταθλήματος, New York Knicks. Άμεσα βλέπουμε πως οι ουραγοί έχουν τις λιγότερες προσπάθειες κοντά στη ρακέτα, επιτίθεται με ομοιόμορφο τρόπο μέσα στη ζώνη του διπόντου και στα τρίποντα βλέπουμε αρκετές προσπάθειες. Οι Golden State Warriors, σε αντίθεση με τους San Antonio Spurs, εκτελούν τρίποντα κατά μέσο όρο σε μεγαλύτερη απόσταση από τη ρακέτα και εκμεταλλεύονται περισσότερο το χώρο γύρω από τη ρακέτα. Συνολικά το ποσοστό ευστοχίας είναι της τάξης του 45% (Σχήμα 4.10). Η μεταβλητή απόκρισης που μελετάμε είναι η έκβαση της προσπάθειας και είναι δυαδική.



Σχήμα 4.10: Συνολικά ποσοστά ευστοχίας

4.2 Βαθμολογία Draymond

Θα θέλαμε να βρούμε ένα τρόπο για να μετρήσουμε την αποτελεσματικότητα ενός παίκτη στην άμυνα. Πέρα από τις κλασικές στατιστικές όπως κλεψίματα και τάπες (blocks), οι οποίες μας περιγράφουν ένα μέρος της αμυντικής ταυτότητας του παίκτη, θα θέλαμε να μετρήσουμε την αποτελεσματικότητα του και σε καταστάσεις ένας εναντίον ενός, με τον ίδιο να βρίσκεται στη θέση του αμυνόμενου. Υπενθυμίζουμε πως κατά τη διάρκεια του παιχνιδιού η επιτιθέμενη ομάδα προσπαθεί να αλλάξει πολλές φορές τη διάταξη της αμυνόμενης ομάδας, με στόχο η επίθεση να καταλήξει σε ένα σουτ υπό καλές προϋποθέσεις. Το ζητούμενο μπορεί να είναι είτε ένα σουτ στο οποίο ο αμυντικός βρίσκεται αρκετά μακριά από τον επιτιθέμενο και ο παίκτης που σουτάρει έχει αρκετό χρόνο να εκτελέσει, είτε ο επιτιθέμενος βρίσκεται σε πλεονεκτικότερη θέση είτε στοχεύοντας το χειρότερο αμυντικό της αντίπαλης ομάδας.

Για την μέτρηση της αποτελεσματικότητας του παίκτη, βασιζόμαστε στην παρακάτω διαδικασία

1. **Κάθε παίκτης ξεκινα με raw draymond 0.**

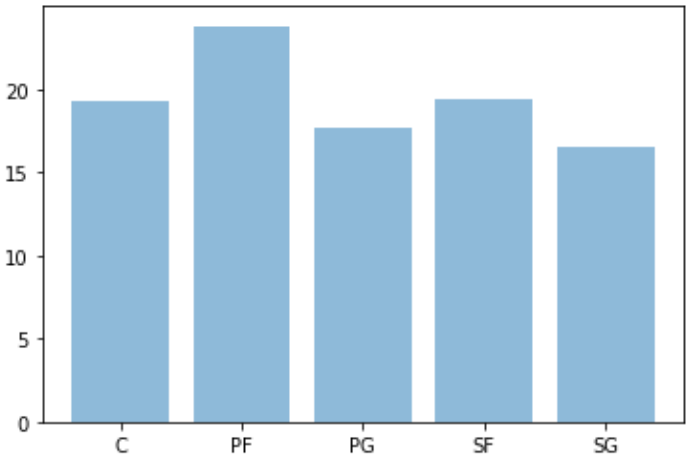
2. **Προσθέτουμε στη βαθμολογία τον εκτιμώμενο αριθμό πόντων που απέτρεψε.**
   Σε κάθε παίκτη όταν αμύνεται σε ένα σουτ που καταλήγει σε άστοχη προσπάθεια τότε προσθέτουμε στο raw draymond το ποσοστό ευστοχίας του επιτιθέμενου πολλαπλασιασμένο με συντελεστή 2 ή 3 ανάλογα αν είναι δίποντο ή τρίποντο. Έτσι, πλέον το διάνυσμα raw draymond εκτιμά για κάθε παίκτη πόσους πόντους απέτρεψε να δεχτεί η ομάδα του.

3. **Διαιρούμε με τον αριθμό των αμυνών στις οποίες εμφανίζεται ο παίκτης ως κοντινότερος αμυντικός.**

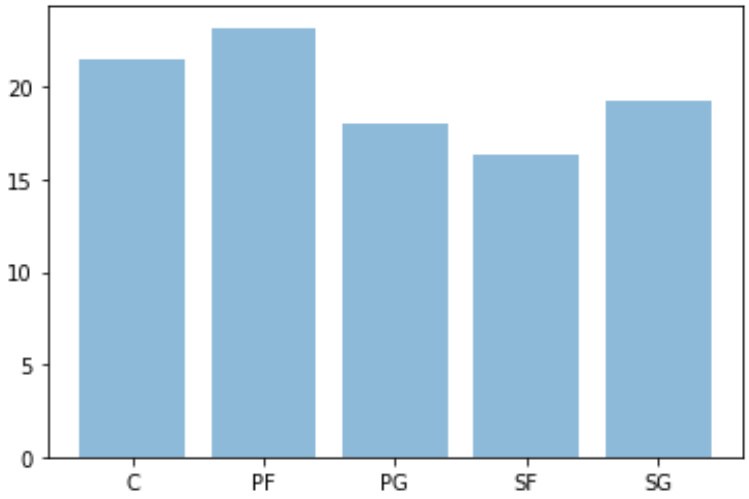
4. **Προσαρμόζουμε το στατιστικό με βάση τη θέση του παίκτη**.

5. **Αφαιρούμε από κάθε παίκτη τη μέση τιμή του δείκτη ανάλογα με τη θέση στην οποία αγωνίζεται**

Για τη σεζόν 2013-2014, ο κάθε παίκτης αμύνθηκε σε 19,3 επιθέσεις (Σχήμα 4.11). Αντίστοιχα, για τη σεζόν 2014-2015 ο μέσος όρος αμυνών είναι 19,6 (Σχήμα 4.12). Στα Σχήματα 4.13- 4.14 φαίνεται η μέση βαθμολογία ανά θέση για τις σεζόν 2013-14 και 2014-15, αντίστοιχα.

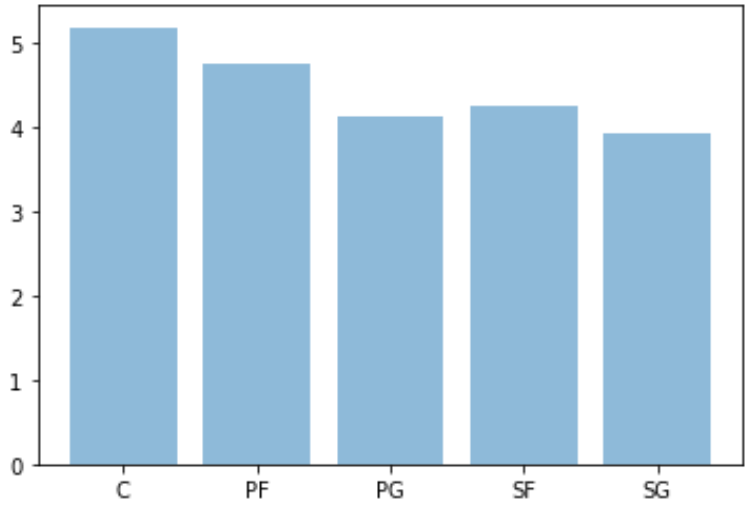


Σχήμα 4.11: Μέσος αριθμός αμυντικών προσπαθειών ανά θέση (2013-14)

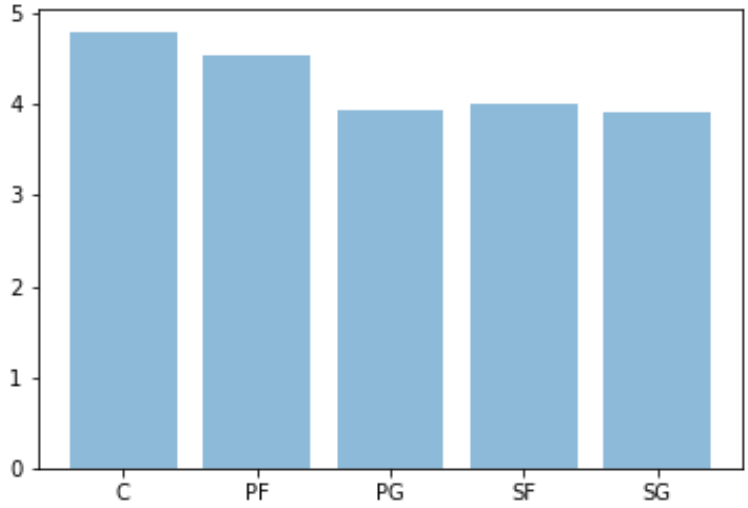


Σχήμα 4.12: Μέσος αριθμός αμυντικών προσπαθειών ανά θέση για τη σεζόν 2014-15

Παρατηρούμε πως το μεγαλύτερο όγκο επιθέσεων για αναχαίτιση τον αναλαμβάνουν οι ψηλοί παίκτες, δηλαδή αυτοί που παίζουν στις θέσεις 4 και 5 (PF,C). Είναι λογικό οι ψηλοί να έχουν τον υψηλότερο δείκτη μιας και προστατεύουν σουτ με μεγαλύτερο ποσοστό ευστοχίας. Οι SG έχουν ως κύριο στόχο την προσφορά πόντων στην επίθεση για αυτό και τα αμυντικά τους νούμερα είναι σχετικά πιο χαμηλά.



Σχήμα 4.13: Μέση βαθμολογία ανά θέση (2013-14)



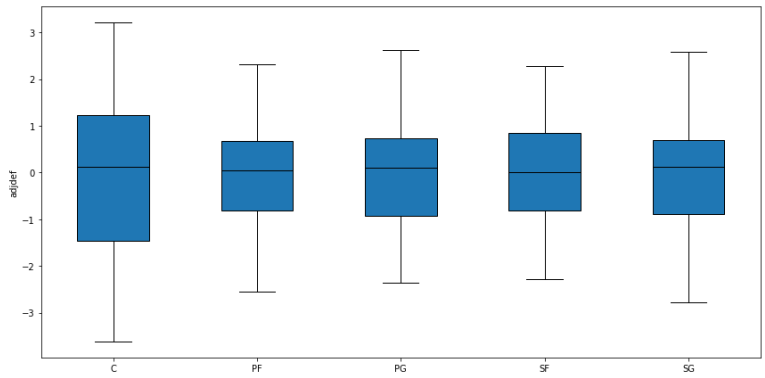
Σχήμα 4.14: Μέση βαθμολογία ανά θέση (2014-15)

Σε κάθε θηκόγραμμα (Σχήματα 4.16-4.15) βλέπουμε τη μεταβλητότητα της τελικής βαθμολογίας για κάθε θέση. Οι κατανομές δεν είναι πάντα σταθερές και ούτε συμμετρικές γύρω από το μηδέν που φαίνεται να είναι και η διάμεσος. Η μεταβλητότητα που παρουσιάζουν δεν είναι σταθερή μεταξύ των διαφορετικών χρονολογιών.

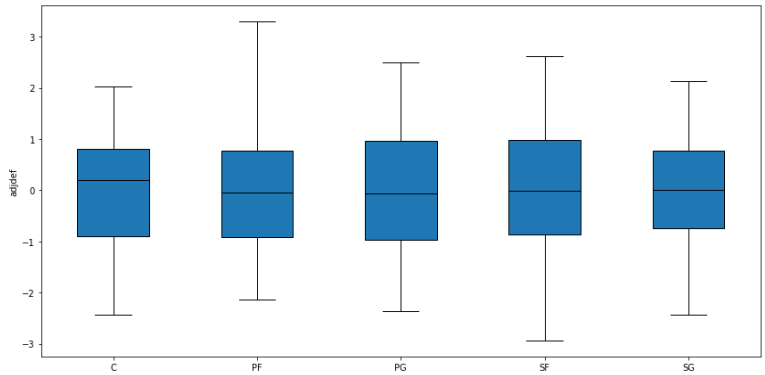
4.2.1 Επιλογή μεταβλητών

Για την επιλογή μεταβλητών χρησιμοποιούμε αρχικά όλες τις διαθέσιμες μεταβλητές καθώς και το δείκτη Draymond. Αν προσαρμόσουμε τη παλινδρόμηση Lasso θα δούμε πως μόνο τα ποσοστά που αφορούν την ευστοχία από μέση απόσταση κρίνονται μη σημαντικά. Χρησιμοποιώντας όμως τον κανόνα της μιας τυπικής απόκλισης θα καταλήξουμε σε ένα πιο φειδωλό μοντέλο.

Πολύ σημαντικό είναι να παρατηρήσουμε πως κάποιες προσπάθειες έχουν μεγαλύτερη πιθανότητα ευστοχίας. Αυτή την πληροφορία μπορούμε να τη βρούμε από τα διάφορα είδη σουτ, φιλτράροντας με βάση την πιθανότητα ευστοχίας. Αυτά είναι 25 περίπου διαφορε-



Σχήμα 4.15: Θηκόγραμμα προσαρμοσμένης βαθμολογίας ανά θέση (2013-14)

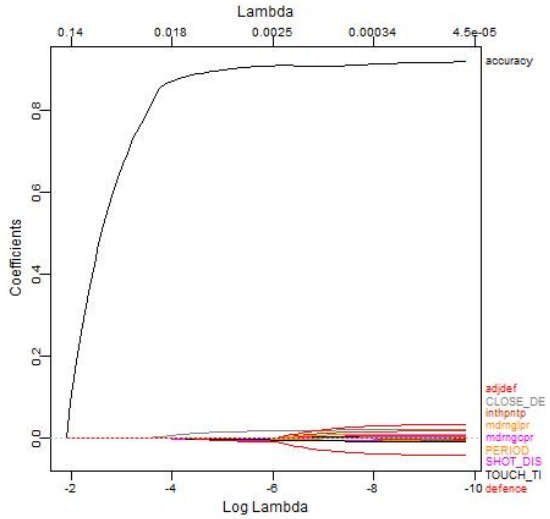


Σχήμα 4.16: Θηκόγραμμα προσαρμοσμένης βαθμολογίας ανά θέση (2014-15)

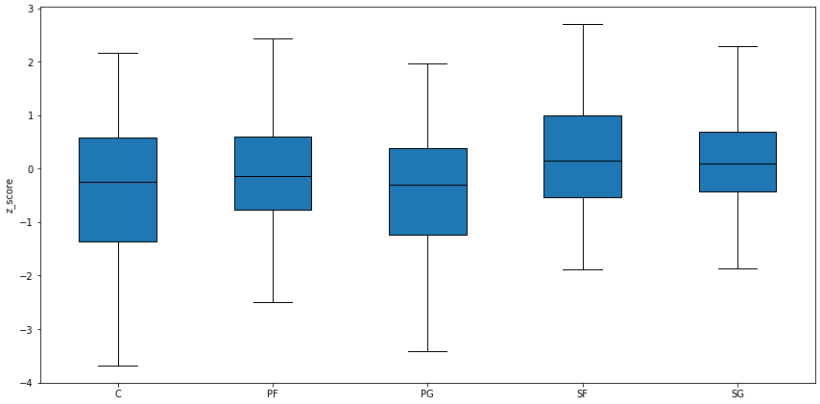
τικά είδη που αντιπροσωπεύουν το 15% των προσπαθειών. Άρα, για να έχουμε καλύτερα αποτελέσματα, χρειάζεται να δημιουργήσουμε έναν τρόπο για να αποκωδικοποιήσουμε τον παραπάνω πίνακα, έτσι ώστε να εξάγουμε τη μέγιστη δυνατή γνώση. Αυτό επιτυγχάνεται μειώνοντας τις διαστάσεις των δεδομένων, ομαδοποιώντας τις και βρίσκοντας κοινά χαρακτηριστικά. Ειδικότερα, παρατηρούμε στον προηγούμενο πίνακα τα καρφώματα (dunks), τα σουτ με ταμπλό (bank) και όσα έχουν κίνηση (Driving, Running, Reverse, Putback) έχουν πολύ μεγάλο ποσοστό ευστοχίας. Κάθε μία κατηγορία θα αναπαρασταθεί με 1 εάν ανήκει σε μια από τις παραπάνω κατηγορίες. Επιπλέον, θα θεωρήσουμε 2 επιπλέον κατηγορίες για κάθε layup και για τα jumpshot. Όσες προσπάθειες δεν ανήκουν σε καμιά από τις κατηγορίες που αναφέραμε, θα παίρνουν μια μονάδα και αυτές. Άρα οι μεταβλητές που θα εισάγουμε στα μοντέλα μας είναι:

1. Απόσταση από τον κοντινότερο αμυντικό
2. Χρόνος που απομένει για τη λήξη της επίθεσης
3. Απόσταση από τη μπασκέτα
4. Βαθμολογία Draymond του αμυνόμενου
5. Διάνυσμα σε μορφή one-hot για το εάν είναι σουτ με υψηλή ευστοχία ή όχι

Από το Σχήμα 4.18 βλέπουμε πως τα στατιστικά στην πλειονότητα των παικτών για τον έλεγχο για σερί στις επιθετικές προσπάθειες οδηγούν στην απόρριψη της μηδενικής υπόθεσης. Ενώ από το Σχήμα 4.17 βλέπουμε πως η πληροφορία για την ευστοχία είναι βεβαίως και η πιο σημαντική για την πρόβλεψη της έκβασης.



Σχήμα 4.17: Παλινδρόμηση Lasso για όλες τις διαθέσιμες μεταβλητές



Σχήμα 4.18: z_score για έλεγχο σερί στις προσπάθειες ανά θέση

4.3 Επιλογή χαρακτηριστικών

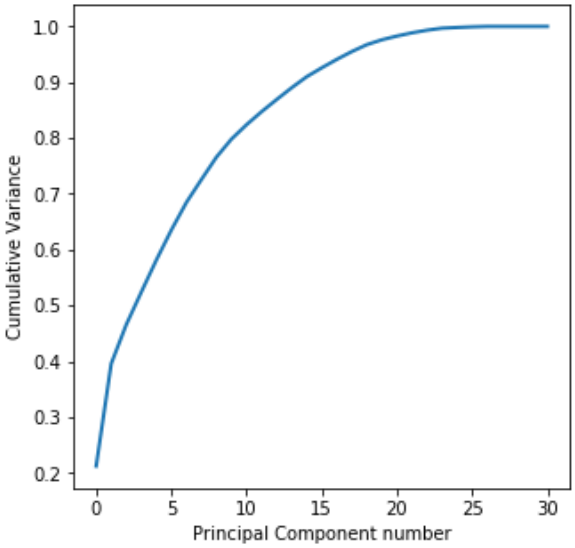
Στο Σχήμα 4.19 απεικονίζεται το αθροιστικό ποσοστό της διακύμανσης των δεδομένων που περιέχεται στα k πρώτα χαρακτηριστικά της ανάλυσης PCA. Φαίνεται πως 15 χαρακτηριστικά μπορούν να ερμηνεύσουν το 80% της διακύμανσης, ενώ 20 μεταβλητές το 100%.

4.4 Αποτελέσματα ταξινόμησης

Το σύνολο δεδομένων αποτελείται από 336.361, το οποίο τεμαχίζεται σε σύνολο εκπαίδευσης (60%), σύνολο επικύρωσης (20%) και σύνολο ελέγχου (20%).

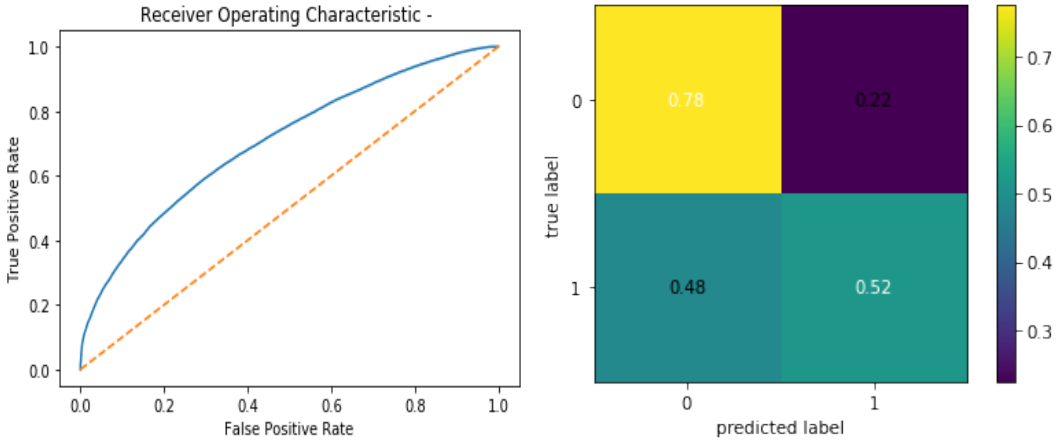
4.4.1 Λογιστική παλινδρόμηση

Με εξαντλητική αναζήτηση, ψάξαμε την καλύτερη μορφή ομαλοποίησης καθώς και τον βαθμό της C, χρησιμοποιώντας διασταυρούμενη επικύρωση. Η προσθήκη της L_1 νόρμας στη συνάρτηση ελαχιστοποίησης και ο παράγοντας $C = 3$ φαίνεται να δίνουν το βέλτιστο



Σχήμα 4.19: Ποσοστό διακύμανσης των δεδομένων που περιέχεται στα k πρώτα χαρακτηριστικά της PCA

αποτελέσματα (65, 7% ακρίβεια). Στον πίνακα σύγχυσης του Σχήματος 4.20 βλέπουμε πως εύκολα η λογιστική παλινδρόμηση μπορεί να προβλέψει τα άστοχα σουτ με μεγάλη ακρίβεια (true negatives 78%), ενώ είναι δύσκολο να ξεχωρίσει τα σουτ που τελικά ήταν εύστοχα (true positives 52%).



(a) Διάγραμμα ROC (b) Πίνακας σύγχυσης

Σχήμα 4.20: Αποτελέσματα για τη λογιστική παλινδρόμηση

4.4.2 Τυχαία Δάση

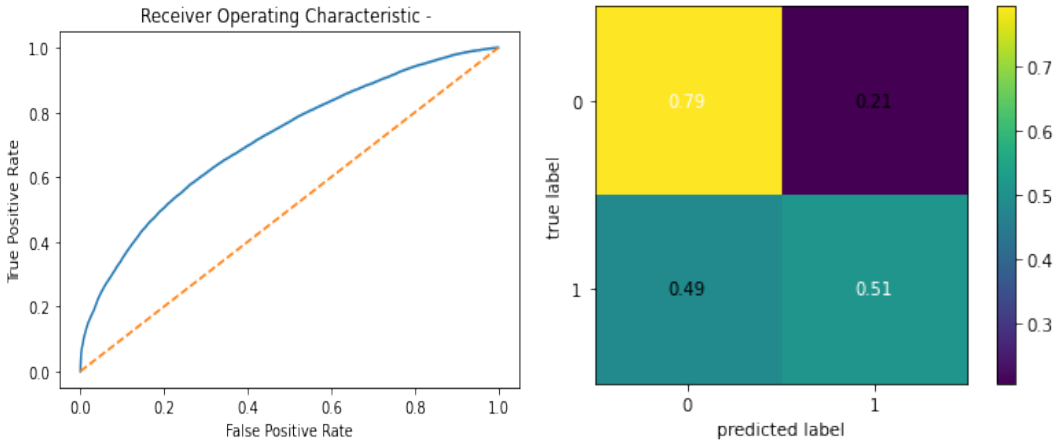
Στην περίπτωση του ταξινομητή τυχαίων δασών, ο Πίνακας 4.1 συνοψίζει τις βέλτιστες υπερπαραμέτρους, οι οποίες βρέθηκαν κατόπιν εξαντλητικής αναζήτησης.

Το μέγιστο βάθος κυμάνθηκε από 2 μέχρι 16 με βήμα 1. Το βάθος ενός δέντρου αφορά την απόσταση από τη ρίζα στα τελικά φύλλα. Επίσης εξετάστηκε εάν χρειάζεται ο ταξινομητής να παίρνει δείγματα με επανατοποθέτηση ή όχι (bootstrap δείγματα). Για το πλήθος των δέντρων που χρειάζεται να εκπαιδεύσουμε, ξεκινήσαμε από 100 και φτάσαμε μέχρι και 400

Πίνακας 4.1: Βέλτιστες τιμές για τις υπερπαραμέτρους του ταξινομητή τυχαίων δασών

Παράμετρος	Βέλτιστη τιμή
`Μέγιστο βάθος`	10
`Bootstrap Δείγματα`	Ναι
`Min samples leaf`	0,008
`Min samples split`	0,01

(βήμα 50). Σημαντικό είναι να εξετάσουμε το μικρότερο αριθμό δειγμάτων για τον οποίο μπορούμε να προχωρήσουμε ένα διαχωρισμό (min samples split) ή το μικρότερο αριθμό δειγμάτων ώστε να θεωρηθεί ένας κόμβος φύλλο (min samples leaf). Η τελευταία υπερπαράμετρος εξετάστηκε από 0,01 ως 0,5 με βήμα 0,01. Τα αποτελέσματα συνοψίζονται στο Σχήμα 4.21, με τη βέλτιστη τιμή του AUC να είναι $0,715$ και της ακρίβειας $66,5\%$.



(a) Διάγραμμα ROC

(b) Πίνακας σύγχυσης

Σχήμα 4.21: Αποτελέσματα για τον ταξινομητή τυχαίων δασών

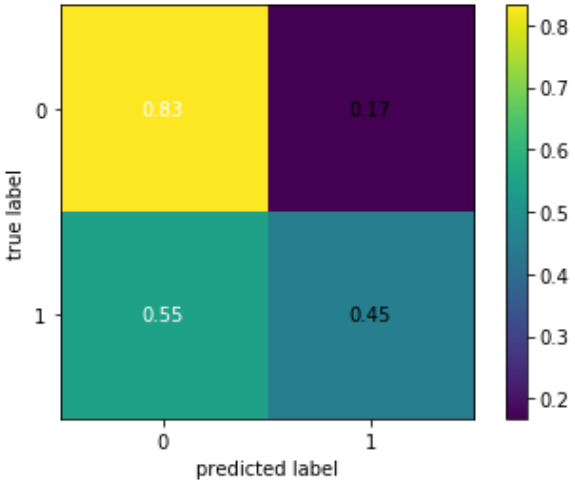
4.4.3 Μηχανές Διανυσμάτων Υποστήριξης

Στα SVM θα εξεταστεί ποιος πυρήνας ανάμεσα στο γραμμικό, πολυωνυμικό και τις ακτινικές συναρτήσεις βάσης δίνει τα βέλτιστα αποτελέσματα. Επίσης θα εξεταστεί ποιος όρος ποινής, σε συνδυασμό με τους τρεις παραπάνω πυρήνες, οδηγεί στη μέγιστη ακρίβεια. Ο βέλτιστος όρος ποινής αναζητείται στο εύρος $[100, 600]$ με αρχικό βήμα 50, το οποίο μειώνεται διαδοχικά. Ο Πίνακας 4.2 συνοψίζει τις βέλτιστες υπερπαραμέτρους, οι οποίες βρέθηκαν κατόπιν εξαντλητικής αναζήτησης.

Πίνακας 4.2: Βέλτιστες τιμές για τις υπερπαραμέτρους του ταξινομητή SVM

Παράμετρος	Βέλτιστη τιμή
`Πυρήνας`	Ακτινική συνάρτηση βάσης
`C`	510

Η μέγιστη τιμή της ακρίβειας φτάνει το $65,8\%$. Ο ταξινομητής φαίνεται να παρουσιάζει τη μεγαλύτερη ακρίβεια στην πρόβλεψη της αστοχίας (83% true negatives) αλλά και τη μικρότερη ακρίβεια στην πρόβλεψη του εύστοχου σουτ (45% true positives).



Σχήμα 4.22: Πίνακας σύγχυσης για τα SVM

4.4.4 XGBoost

Στην περίπτωση του ταξινομητή XGBoost, ο Πίνακας 4.3 περιέχει τις βέλτιστες τιμές για τις υπερπαραμέτρους που εξετάστηκαν, δηλαδή την αντικειμενική συνάρτηση προς ελαχιστοποίηση, το μέγιστο βάθος ενός δένδρου, το ρυθμό μάθησης και το ποσοστό της δειγματοληψίας κατά στήλες και γραμμές. Η εκπαίδευση μπορεί να επιταχυνθεί θέτοντας ένα πολύ μεγάλο αριθμό επαναλήψεων (εμείς ορίσαμε 1000) και ορίζοντας το πλήθος των συνεχόμενων επαναλήψεων για τις οποίες επιτρέπεται το σφάλμα να μην μειώνεται (εμείς ορίσαμε μέχρι 10).

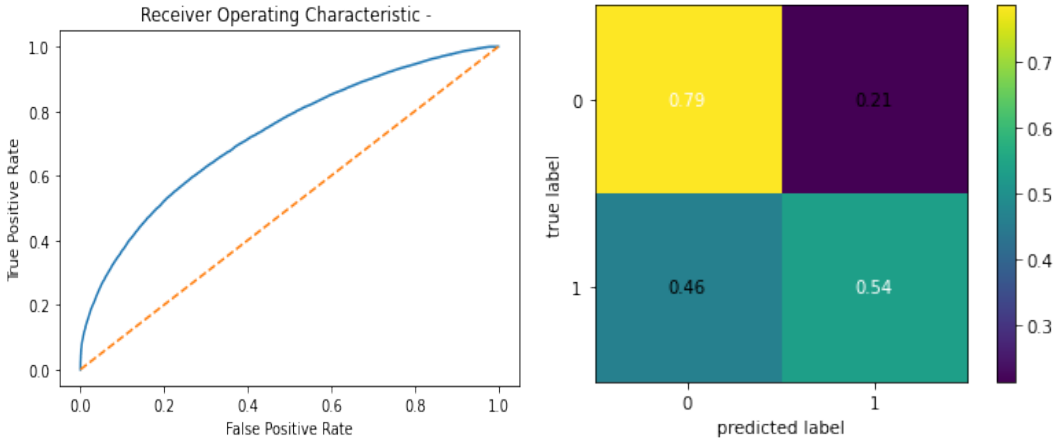
Πίνακας 4.3: Βέλτιστες τιμές για τις υπερπαραμέτρους του ταξινομητή XGBoost

Παράμετρος	**Βέλτιστη τιμή**
`Μέγιστο βάθος`	6
`Ρυθμός Μάθησης`	0,13
`Δειγματοληψία κατά γραμμές`	0,85
`Δειγματοληψία κατά στήλες`	0,85

Στο Σχήμα 4.23 συνοψίζεται η καμπύλη ROC και ο πίνακας σύγχυσης για τον XGBoost. Το AUC είναι 0,728 και η μέγιστη ακρίβεια λαμβάνει την τιμή $67,1\%$. Παρατηρούμε πως η μέθοδος XGBoost παρουσιάζει το μεγαλύτερο AUC με χαρακτηριστικό το μεγαλύτερο αριθμό true positives και η ακρίβεια λαμβάνει τη μεγαλύτερη, μέχρι στιγμής, τιμή.

4.4.5 Πολυεπίπεδα Perceptron

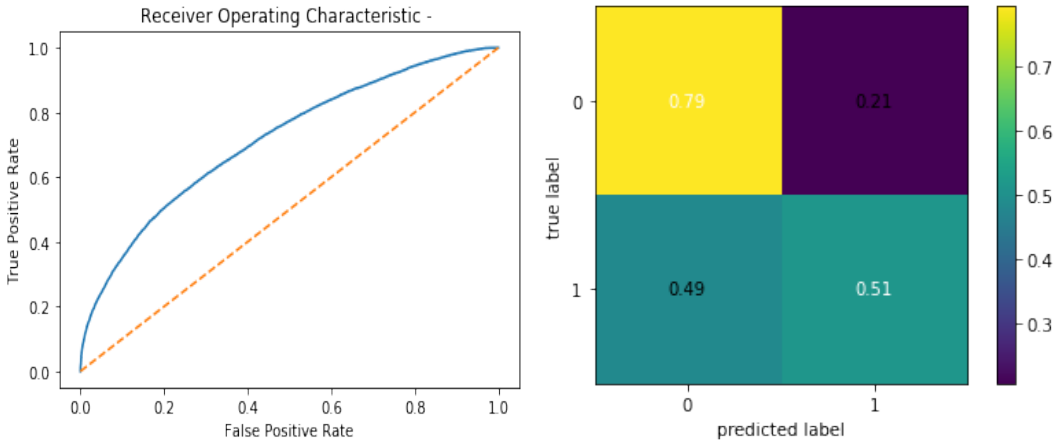
Στα νευρωνικά δίκτυα, οι πολύ βαθιές αρχιτεκτονικές δεν έδωσαν ικανοποιητικά αποτελέσματα, για αυτό εξετάστηκαν πολυεπίπεδα Perceptrons μέχρι 2 κρυφά επίπεδα. Σε αυτή την περίπτωση των ταξινομητών δώσαμε επιπρόσθετη είσοδο τη θέση του αμυνόμενου. Συνολικά η είσοδος είχε 17 διαστάσεις, ενώ το μέγεθος των κρυφών επιπέδων δεν ξεπέρασε τους 5 νευρώνες. Η συνάρτηση ενεργοποίησης που επιλέχθηκε είναι η υπερβολική εφαπτομένη. Η αρχικοποίηση των βαρών έγινε με δειγματοληψία από την ημι-κανονική κατανομή με μέση τιμή 0. Στο επίπεδο εξόδου η συνάρτηση ενεργοποίησης ήταν η σιγμοειδής, ενώ για την εκπαίδευση χρησιμοποιήθηκε ο βελτιστοποιητής Adam. Η συνάρτηση απώλειας είναι η δυαδική διασταυρούμενη εντροπία.



(a) Διάγραμμα ROC (b) Πίνακας σύγχυσης

Σχήμα 4.23: Αποτελέσματα για τον ταξινομητή XGBoost

Εκπαιδεύσαμε το δίκτυο για 200 εποχές και αποθηκεύσαμε τα βάρη στην εποχή εκείνη που η ακρίβεια στο σύνολο επικύρωσης έλαβε τη μέγιστη τιμή. Μετά από διάφορους πειραματισμούς, παρατηρήσαμε πως θέτοντας ένα ρυθμό εκπαίδευσης μικρό και σταθερό για τις 10 πρώτες εποχές και αργότερα εφαρμόζοντας εκθετική μείωση, προσθέτοντας παράλληλα την επιλογή batch normalization, το δίκτυο συγκλίνει ομαλά σε ένα τοπικό μέγιστο χωρίς να ταλαντώνεται. Στο Σχήμα 4.24 παρουσιάζονται η καμπύλη ROC και ο πίνακας σύγχυσης για το πολυεπίπεδο Perceptron. Η ακρίβεια φτάνει σε ένα μέγιστο της τάξης του 67%, ενώ το εμβαδόν κάτω από την καμπύλη είναι $0,73$.



(a) Διάγραμμα ROC (b) Πίνακας σύγχυσης

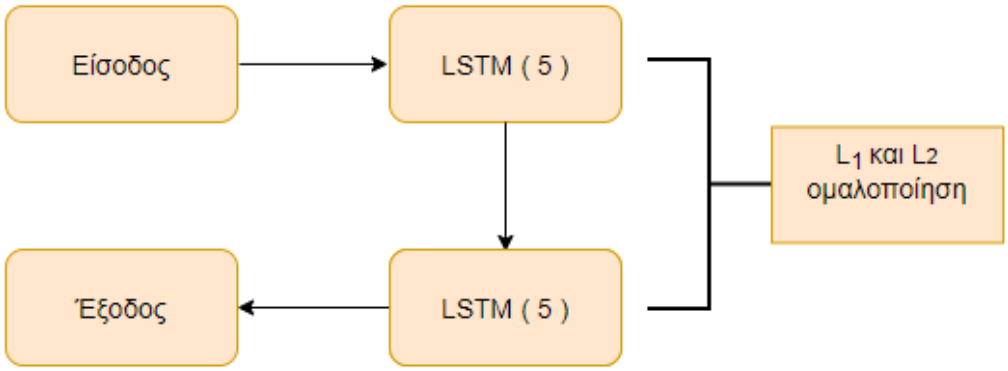
Σχήμα 4.24: Αποτελέσματα για το πολυεπίπεδο Perceptron

4.4.6 Αναδρομικά νευρωνικά δίκτυα

Για να μελετήσουμε τα αναδρομικά νευρωνικά δίκτυα και την εξάρτηση του παρελθόντος θα χρειαστεί να εξετάσουμε τις αρχιτεκτονικές LSTM, GRU, σε μορφές stacked και bi-directional. Λόγω ύπαρξης κάποιων ελλειμματικών παρατηρήσεων, δε μπορούμε να εξετάσουμε πολύ παρελθοντικές τιμές, οπότε περιοριζόμαστε μέχρι 3 προσπάθειες πίσω. Για τις προσπάθειες κάθε παίκτη εκπαιδεύεται και διαφορετικό αναδρομικό νευρωνικό δίκτυο,

του οποίου η είσοδος είναι ταξινομημένη ως προς το χρόνο που πραγματοποιείται η επίθεση και ως προς το χρόνο που εκτελείται η προσπάθεια. Δεδομένου ότι λαμβάνουμε υπόψη μόνο 3 παρελθοντικές προσπάθειες, δεν θα χρειαστεί να καταφύγουμε σε ιδιαίτερα περίπλοκες αρχιτεκτονικές αναδρομικών δικτύων.

Το κυριότερο ζήτημα που προκύπτει είναι αν θα λάβουμε υπόψη μας τις παρελθοντικές τιμές που αφορούν μόνο το αποτέλεσμα ή θα συμπεριλάβουμε πλήρως όλες τις πληροφορίες για την επιθετική προσπάθεια. Πειραματιζόμενοι με διάφορα συστήματα και αρχιτεκτονικές καταλήξαμε ότι τελικά χρειάζεται να λάβουμε υπόψη μας όλη την πληροφορία για τις παρελθοντικές προσπάθειες.



Σχήμα 4.25: Διαγραμματική απεικόνιση του τελικού μοντέλου του αναδρομικού νευρωνικού δικτύου

Τα αποτελέσματα ποικίλουν ανά παίκτη, χωρίς να υπάρχει κάποιο συγκεκριμένο πρότυπο. Η μέγιστη ακρίβεια κυμαίνεται από 62% μέχρι 71% στο σύνολο επικύρωσης. Μεγάλη βελτίωση προέκυψε προσθέτοντας στα επίπεδα με τα LSTM τη L_1 και L_2 ποινικοποίηση αθροιστικά (Σχήμα 4.25). Με αυτό το τρόπο παρατηρήσαμε πως, ενώ το σφάλμα παραμένει ίδιο στο σύνολο εκπαίδευσης και επικύρωσης, η ακρίβεια βελτιώνεται σημαντικά μέχρι και 3% παραπάνω σε σχέση με τα αποτελέσματα που είχαμε χωρίς καμία μορφή ομαλοποίησης. Αν και τα δεδομένα ανά παίκτη είναι λίγα (κυμαίνονται από 200 μέχρι 1000 δείγματα), η απόδοσης κρίνεται ικανοποιητική. Μετά από 20 εποχές από την επίτευξη της μέγιστης ακρίβειας στο σύνολο δεδομένων, οι παράγωγοι αρχίζουν να παίρνουν μεγάλες τιμές ή να εξαφανίζονται. Η προσπάθεια περιορισμού της κλίσης (gradient clipping) δεν βοήθησε στην αντιμετώπιση του προβλήματος που πιθανότατα να οφείλεται στο μικρό δείγμα.

4.5 Δυναμική πρόβλεψη νικητή

Θα θέλαμε να μελετήσουμε ποιοι παράγοντες οδηγούν στη νίκη και εάν μπορούμε να προβλέψουμε με ένα πιο δυναμικό τρόπο το νικητή μεταξύ δύο διαγωνιζόμενων ομάδων. Θα θέλαμε, σε αντίθεση με την προηγούμενη περίπτωση όπου μελετήσαμε μια πιο στατική περίπτωση στην πρόβλεψη για την έκβαση ενός σουτ, να μπορέσουμε, έχοντας παρατηρήσει τη χρονική στιγμή t τι έχει κάνει μια ομάδα, να προβλέψουμε με ικανοποιητική ακρίβεια τον νικητή για την επόμενη χρονική στιγμή $t+1$. Αυτόματα εγείρονται πολλά ζητήματα σχετικά με ποιες μεταβλητές να εισάγουμε στο μοντέλο και με ποιο τρόπο. Στην περίπτωση

που έχουμε καταλήξει σε ποιες μεταβλητές θα χρησιμοποιήσουμε, το ερώτημα είναι πως θα τις εισάγουμε στα διάφορα μοντέλα μάθησης.

Πιο συγκεκριμένα, τα κυριότερα σχεδιαστικά ζητήματα που καλούμαστε να απαντήσουμε είναι τα ακόλουθα

1. αν το σύνολο δεδομένων θα χωρίζεται σε δύο μεγάλες κατηγορίες, μία για κάθε διαγωνιζόμενη ομάδα.

2. αν θα εξετάσουμε τις τιμές των χαρακτηριστικών για κάθε ομάδα ή τις διαφορές τους

Επιπρόσθετα θα ήταν επιθυμητό να προσθέταμε κάποιες μεταβλητές που να περιγράφουν τη δυναμική της ομάδας, η οποία δεν θα είναι απλώς η θέση της στο βαθμολογικό πίνακα, αλλά θα αντικατοπτρίζει τη φόρμα και τη δυναμική της σε σχέση με τις υπόλοιπες. Για το σκοπό αυτό, χρησιμοποιούμε το δείκτη Elo (Ενότητα 2.4). Η βαθμολογία Elo των ομάδων ξεκινάει από το 1200 και χρησιμοποιείται ο τύπος που κρατά το 75% της παλιάς βαθμολογίας και ανανεώνεται κατά 25% από την επιβράβευση που προκύπτει από το πρώτο παιχνίδι της καινούργιας αγωνιστικής χρονιάς.

Η βαθμολογία Elo έχει το χαρακτηριστικό του ότι δεν μπορεί να προσαρμοστεί πολύ γρήγορα σε ξαφνικές αλλαγές όπως μεταγραφές και τραυματισμοί παικτών. Για το λόγο αυτό, θα λάβουμε υπόψη μας μια επιπλέον μεταβλητή, το σερί νικών στα 4 προηγούμενα παιχνίδια.

Οι χρονοσειρές που θα χρησιμοποιήσουμε θα ενημερώνονται με κινούμενο μέσο και δε θα αφαιρέσουμε μεταξύ τους όσες αναφέρονται στον ίδιο παράγοντα αλλά για διαφορετικές ομάδες)θα διατηρηθούν όπως είναι). Συνολικά, οι παράγοντες που θα εξεταστούν για κάθε μοντέλο είναι:

1. Πλήθος και ευστοχία σε Field goals, δίποντα, τρίποντα και βολές

2. Πλήθος αμυντικών, επιθετικών και συνολικών ριμπάουντ

3. Οι 4 παράγοντες του μοντέλου του Dean Oliver

4. Βαθμολογία Elo

5. Σερί νικών στα 4 τελευταία παιχνίδια

και θα αφορούν τόσο τη γηπεδούχο όσο και την φιλοξενούμενη ομάδα. Παρότι φαίνεται να μην είναι όλες στατιστικά σημαντικές, εντούτοις, για αρχή, θα τις συμπεριλάβουμε όλες στο μοντέλο (Σχήμα 4.26).

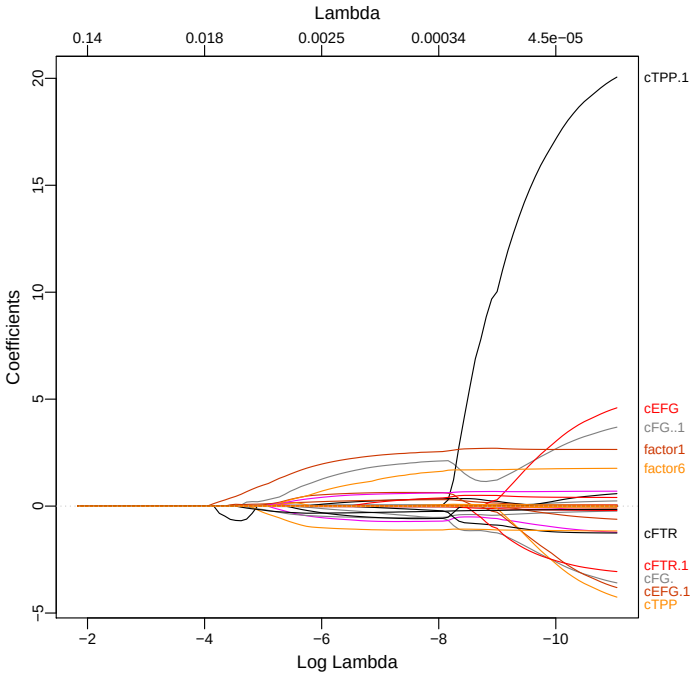
Το σύνολο δεδομένων που θα χρησιμοποιηθεί αφορά όλους τους αγώνες της κανονικής διάρκεια από τη σεζόν 2014-15 ως και την 2018-19[1]. Το ποσοστό νίκης της γηπεδούχου ομάδας ανέρχεται στο $58,2\%$. Καθώς για τη συγκεκριμένη εργασία το δείγμα είναι μικρό, θα το διαχωρίσουμε μόνο σε δύο μέρη (σύνολο εκπαίδευσης και σύνολο ελέγχου) και όχι σε 3, όπως προηγουμένως.

4.5.1 Εφαρμογή ταξινομητών

Λογιστική παλινδρόμηση

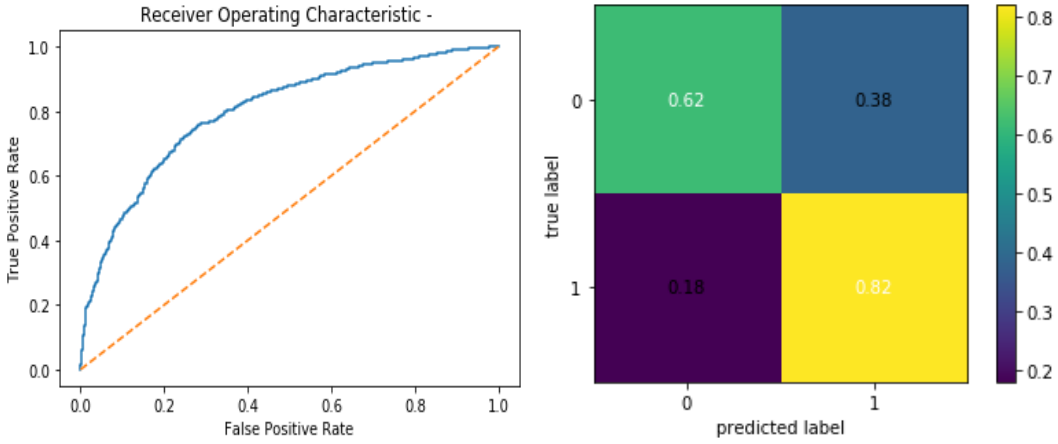
Με εξαντλητική αναζήτηση και διασταυρούμενη επικύρωση 5 πτυχών βρίσκουμε πως με L_1 κανονικοποίηση και $C = 10$ παίρνουμε ένα βέλτιστο αποτέλεσμα στην ακρίβεια που

[1] Διαθέσιμα στο https://www.basketball-reference.com/



Σχήμα 4.26: Παλινδρόμηση Lasso για όλες τις υποψήφιες μεταβλητές

είναι 73, 8%. Το AUC είναι 0,81 και στο πίνακα σύγχυσης βλέπουμε πως με πιθανότητα 0,8 προβλέπει τη νίκη του γηπεδούχου με ακρίβεια ενώ με πιθανότητα 0,6 προβλέπει τη νίκη της φιλοξενούμενης ομάδας σωστά (Σχήμα 4.27).



(a) Διάγραμμα ROC (b) Πίνακας σύγχυσης

Σχήμα 4.27: Αποτελέσματα πρόβλεψης νικητή για τη λογιστική ταξινόμηση

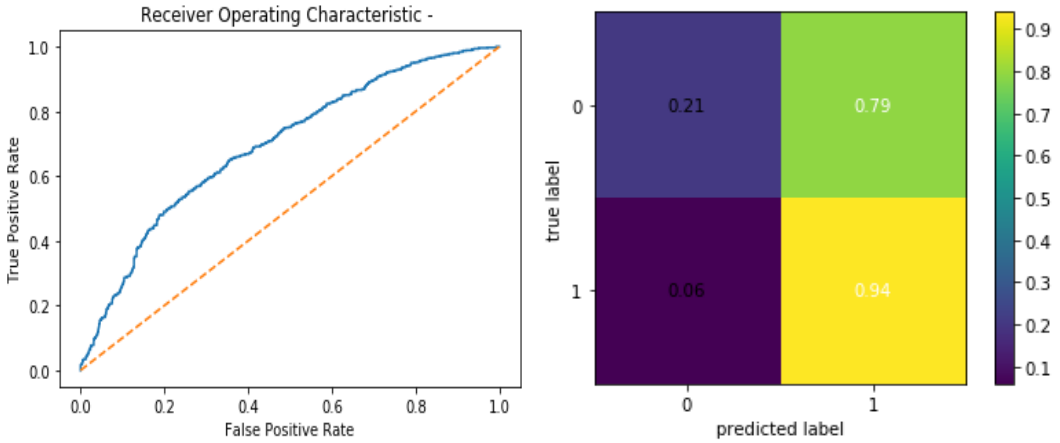
Τυχαία Δάση

Με εξαντλητική αναζήτηση βρίσκουμε πως οι βέλτιστοι υπερπαράμετροι παίρνουν τις τιμές του Πίνακα 4.4. Η ακρίβεια δεν ξεπερνά το 67%, ενώ το AUC είναι 0,70, το μικρότερο από όλα (Σχήμα 4.28). Το χαρακτηριστικό των τυχαίων δασών είναι πως οι περισσότερες

προβλέψεις τους είναι για νίκη της γηπεδούχου ομάδας.

Πίνακας 4.4: Βέλτιστες τιμές για τις υπερπαραμέτρους του ταξινομητή τυχαίων δασών

Παράμετρος	**Βέλτιστη τιμή**
`Μέγιστο βάθος`	9
`Bootstrap Δείγματα`	Ναι
`Min samples leaf`	0,1
`Min samples split`	0,1



(a) Διάγραμμα ROC (b) Πίνακας σύγχυσης

Σχήμα 4.28: Αποτελέσματα πρόβλεψης νικητή για τον ταξινομητή τυχαίων δασών

Μηχανές Διανυσμάτων Υποστήριξης

Οι βέλτιστες τιμές των υπερπαραμέτρων για τα SVM δίνονται στον Πίνακα 4.5. Το SVM καταφέρνει να παρουσιάσει ακρίβεια 73.4% και να προβλέψει ικανοποιητικά την νίκη της γηπεδούχου. (Σχήμα 4.29)

Πίνακας 4.5: Βέλτιστες τιμές για τις υπερπαραμέτρους του ταξινομητή μηχανών διανυσμάτων υποστήριξης

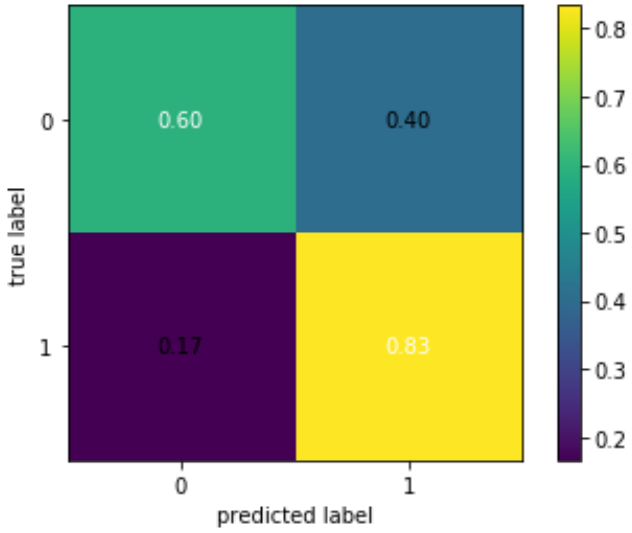
Παράμετρος	**Βέλτιστη τιμή**
`Πυρήνας`	Συνάρτηση ακτινικής βάσης
`Συντελεστής Ποινικοποίησης`	210

XGBoost

Οι βέλτιστες τιμές των υπερπαραμέτρων για τον XGBoost δίνονται στον Πίνακα 4.6. Η βέλτιστη ακρίβεια που παρουσιάζει ο XGBoost φτάνει στο 72, 5% (Σχήμα 4.30).

Πολυεπίπεδο Perceptron

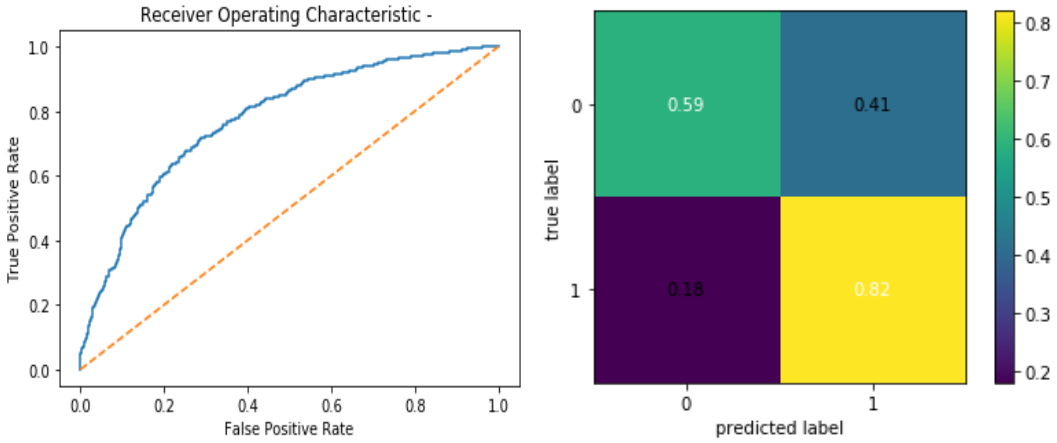
Και στην περίπτωση της δυναμική πρόβλεψης του νικητή, οι πολύ βαθιές αρχιτεκτονικές δεν παρουσίασαν ικανοποιητικά αποτελέσματα. Για αυτό εξετάσαμε νευρωνικά δίκτυα με 3



Σχήμα 4.29: Πίνακας σύγχυσης για το SVM

Πίνακας 4.6: Βέλτιστες τιμές για τις υπερπαραμέτρους του ταξινομητή XGBoost

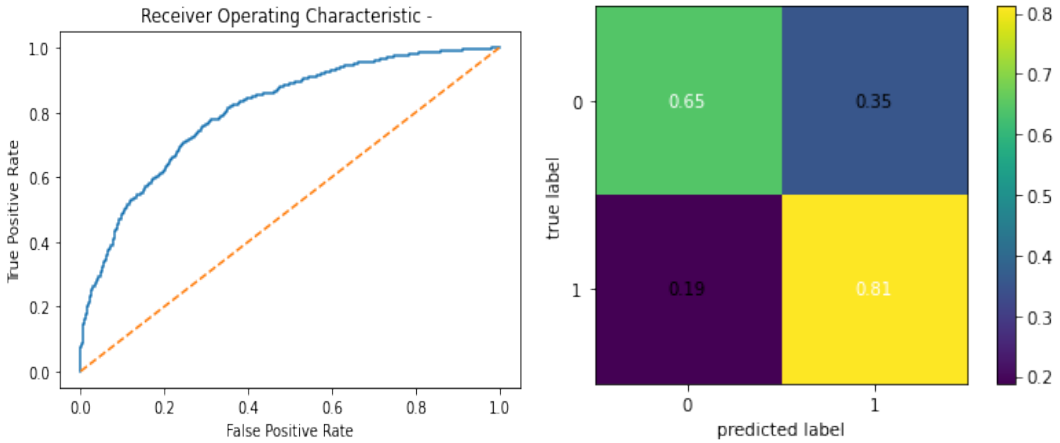
Παράμετρος	**Βέλτιστη τιμή**
`Μέγιστο βάθος`	3
`Bootstrap Δείγματα`	Ναι
`Δειγματοληψία κατά στήλες`	0.8
`Δειγματοληψία κατά γραμμές`	0.8



(a) Διάγραμμα ROC (b) Πίνακας σύγχυσης

Σχήμα 4.30: Αποτελέσματα πρόβλεψης νικητή για τον ταξινομητή XGBoost

το πολύ κρυφά επίπεδα. Τελικά, το βέλτιστο δίκτυο είχε 2 κρυφά επίπεδα με το πρώτο να έχει 6 νευρώνες και το δεύτερο 2. Και εδώ, η αρχικοποίηση των βαρών έγινε μέσω δειγματοληψίας από την ημι-κανονική κατανομή με μέση τιμή μηδέν. Τέλος οι συναρτήσεις ενεργοποίησης ήταν η γραμμική για τα κρυφά επίπεδα και η σιγμοειδής για το επίπεδο εξόδου.



(a) Διάγραμμα ROC (b) Πίνακας σύγχυσης

Σχήμα 4.31: Αποτελέσματα πρόβλεψης νικητή για το πολυεπίπεδο Perceptron

Η μέγιστη τιμή της ακρίβειας είναι ίση με $74,4\%$. Η AUC εμφανίζει τη μέγιστη τιμή σε σχέση με όλους τους ταξινομητές (0,82) Το χαρακτηριστικό που παρατηρούμε από τον πίνακα σύγχυσης είναι (Σχήμα 4.31) είναι πως τα true negatives λαμβάνουν τη μέγιστη τιμή που είναι ίση με 65%.

Κεφάλαιο 5

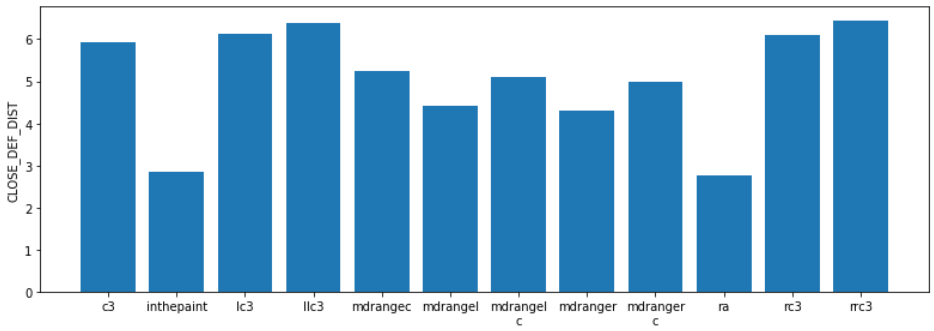
Συμπεράσματα και μελλοντικές κατευθύνσεις

5.1 Πρόβλεψη ευστοχίας

Έχοντας εφαρμόσει μια πληθώρα ταξινομητών σε συνδυασμό με την εξαντλητική αναζήτηση προς εύρεση των βέλτιστων δυνατών παραμέτρων συμπεραίνουμε πως τα πολυεπίπεδα perceptrons, η λογιστική παλινδρόμηση και ο XGBoost έχουν την καλύτερη αποτελεσματικότητα. Όμως δε παρατηρείται σημαντική βελτίωση στα αποτελέσματα παρά την προσπάθεια μας να εξάγουμε όλη τη δυνατή γνώση και να χρησιμοποιήσουμε και εξωτερική πληροφορία.

Μια σημαντική έλλειψη στα δεδομένα μας ήταν ότι δεν γνωρίζαμε πόσο ήταν το σκορ τη στιγμή της εκδήλωσης της εκάστοτε επίθεσης, έστι ώστε να μπορούμε να προβλέψουμε την ψυχολογική κατάσταση του παίκτη. Για παράδειγμα, αν η διαφορά είναι 10 πόντοι υπερ του επιτιθέμενου, είναι αναμενόμενο ο παίκτης περισσότερη αυτοπεποίθηση και κατ' επέκταση μεγαλύτερη πιθανότητα ευστοχίας.

Επίσης φαίνεται πως είναι δύσκολο να προβλέψουμε το αποτελέσματα μιας δυναμικής διαδικασίας (θα είναι επιτυχημένη η συγκεκριμένη απόπειρα του παίκτη ή όχι) μέσω μιας στατικής εικόνας. Αν φιλτράρουμε τις προσπάθεις στα σουτ όπου οι επιτιθέμενοι έχουν ευστοχία πάνω από 45% δημιουργείται το γράφημα της μέσης απόστασης από τον κοντινότερο αμυντικό ανά ζώνη προσπάθειας (Σχήμα 5.1).



Σχήμα 5.1: Διάγραμμα απόστασης από κοντινότερο αμυντικό ανά ζώνη.

Παρατηρούμε πως όσο πιο πολύ απομακρυνόμαστε από την μπασκέτα, τα σουτ φαίνεται να γίνονται υπό καλύτερες προϋποθέσεις, δηλαδή ο επιτιθέμενος επιχειρεί την προσπάθεια του με λιγότερη πίεση. Με μια απλή αναζήτηση παρατηρούμε πως η ελάχιστη μέση εκτίναξη του αμυνόμενου χωρίς άλμα είναι πάνω από 3 μέτρα, σχεδόν διπλάσια από τη μέση μέγιστη απόσταση που παρατηρήσαμε στο παραπάνω γράφημα. Επίσης, οι πολύ αποτελεσματικοί επιτιθέμενοι μαρκάρονται στενά από τις αντίπαλες άμυνες με αποτέλεσμα να είναι δύσκολο να προβλεφθεί η έκβαση των προσπαθειών τους.

Τέλος, καταγράφοντας μια μονάδα σε κάθε αμυνόμενο για στη ζώνη που εμφανίζεται να προσπαθεί να δυσκολέψει τον επιτιθέμενο, δημιουργούμε έναν πίνακα V. Για να μειώσουμε

τις διαστάσεις του, μπορούμε να εφαρμόσουμε τεχνικές όπως η *μη-αρνητική παραγοποίηση πινάκων* (non negative matrix factorization) [Lee01]. Με τη διαδικασία αυτή προσπαθούμε να βρούμε δύο πίνακες W, H τέτοιους ώστε $V \approx WH$. όπου η διάσταση f των στηλών του W και των γραμμών του H είναι πολύ μικρότερη από τις αρχικές διαστάσεις του V.

Στο παράδειγμα που εξετάζουμε, το εσωτερικό γινόμενο των γραμμών του πίνακα W εκφράζει πλέον την ομοιότητα των χαρακτηριστικών των παικτών ή ισοδύναμα την αμυντική τους συμπεριφορά. Κάποια αρχικά πειράματα που έχουμε κάνει εμφανίζουν τον πιο ψηλό και ταυτόχρονα πιο αργό παίκτη να έχει ομοιότητα 50% με τον πιο κοντό, ένδειξη των πολλών αλλαγών θέσεων που μπορεί να πραγματοποιηθούν κατά τη διάρκεια μιας επίθεσης.

5.2 Δυναμική πρόβλεψη νικητή

Παρατηρούμε πως με η λογιστική παλινδρόμηση και τα πολυεπίπεδα Perceptron εμφανίζουν τα βέλτιστα αποτελέσματα για την ακρίβεια (της τάξης του 74%) και το AUC (80%). Στο NBA οι καλές ομάδες σε σύνολο 81 αγώνων κάνουν 20 με 30 ήττες συνολικά, είτε από ισοδύναμες ομάδες είτε από χαμηλότερης δυναμικότητας. Στο Κεφάλαιο 2 είδαμε πως οι διαφορές στα παιχνίδια, εφόσον γνωρίζουμε τη βαθμολογία όπως την υπολογίσαμε, ακολουθούν μια τυπική κατανομή με μέσο που διαφέρει ανάλογα τις διαγωνιζόμενες ομάδες, αλλά πολύ μεγάλη τυπική απόκλιση (θυμίζουμε πως είχαμε δείξει ότι η τυπική απόκλιση είναι 12 μονάδες). Επίσης, κατά τη διάρκεια του παιχνιδιού μπορούν να γίνουν κάποιοι τραυματισμοί ή αποβολή σημαντικών παικτών, οι οποίοι με την παρουσία τους μπορεί να επηρέαζαν καθοριστικά το τελικό αποτέλεσμα.

Μια άμεση βελτίωση στη διαδικασία πρόβλεψης είναι να προσθέσουμε τις βαθμολογίες των παικτών που αγωνίζονται στις εκάστοτε ομάδες, κατ' αντίστοιχο τρόπο με την πρόσθεση των βαθμολογιών των ομάδων. Χρειαζόμαστε την πληροφορία για τους παίκτες που βρίσκονται στο παρκέ, καθώς και την παρατηρούμενη διαφορά μεταξύ των ομάδων σε κάθε χρονική στιγμή. Θα πρέπει επίσης να λάβουμε υπόψη μας το πλεονέκτημα έδρας, το οποίο το υπολογίσαμε σε 48 λεπτά καθαρού αγώνα. Με αυτή την πληροφορία θα μπορούσαμε να προσθέσουμε στο μοντέλο μας την απουσία ενός σημαντικού παίκτη και πως αυτή επηρεάζει την τελική διαφορά στο σκορ.

Βιβλιογραφία

[Alam13] Benjamin C Alamar, *Sports analytics: A guide for coaches, managers, and other decision makers*, Columbia University Press, 2013.

[Ali12] Jehad Ali, Rehanullah Khan, Nasir Ahmad and Imran Maqsood, "Random Forests and Decision Trees", *International Journal of Computer Science Issues(IJCSI)*, vol. 9, 09 2012.

[Anal] Agile Sports Analytics, "Sports Analytics Methods – Machine Learning Analysis", https://www.agilesportsanalytics.com/sports-analytics-machine-learning-analysis/.

[Brad97] Andrew P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", *Pattern Recognition*, vol. 30, no. 7, pp. 1145 – 1159, 1997.

[Chen16] Tianqi Chen and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", *CoRR*, vol. abs/1603.02754, 2016.

[Chen17] Yen-Chi Chen, "A Tutorial on Kernel Density Estimation and Recent Advances", 2017.

[Elo78] Arpad E Elo, *The rating of chessplayers, past and present*, Arco Pub., 1978.

[Evge01] Theodoros Evgeniou and Massimiliano Pontil, "Support Vector Machines: Theory and Applications", vol. 2049, pp. 249–257, 01 2001.

[Frie01] Jerome H Friedman, "Greedy function approximation: a gradient boosting machine", *Annals of statistics*, pp. 1189–1232, 2001.

[Frie02] Jerome H Friedman, "Stochastic gradient boosting", *Computational statistics & data analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[Good16] Ian J. Goodfellow, Yoshua Bengio and Aaron Courville, *Deep Learning*, MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[Hast09] Trevor Hastie, Robert Tibshirani and Jerome Friedman, *The elements of statistical learning: data mining, inference and prediction*, Springer, 2 edition, 2009.

[Hayk09] S. Haykin and S.S. Haykin, *Neural Networks and Learning Machines*, no. v. 10 in Neural networks and learning machines, Prentice Hall, 2009.

[Hoer70] Arthur E Hoerl and Robert W Kennard, "Ridge regression: Biased estimation for nonorthogonal problems", *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[hwch15] hwchase17, "Pulling and working with NBA SportVu data", https://github.com/hwchase17/sportvu, 2015.

[Ke17] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye and Tie-Yan Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", in I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pp. 3146–3154, Curran Associates, Inc., 2017.

[King14] Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization", 2014.

[Kons18] Pelechrinis Konstantinos, "Moneyball 2.0: Winning in Sports w Data", 2018.

[Kraf88] D. Kraft, *A Software Package for Sequential Quadratic Programming*, Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt Köln: Forschungsbericht, Wiss. Berichtswesen d. DFVLR, 1988.

[Kuba07a] Justin Kubatko, Dean Oliver, Kevin Pelton and Dan Rosenbaum, "A Starting Point for Analyzing Basketball Statistics", *Journal of Quantitative Analysis in Sports*, vol. 3, pp. 1–1, 02 2007.

[Kuba07b] Justin Kubatko, Dean Oliver, Kevin Pelton and Dan T Rosenbaum, "A starting point for analyzing basketball statistics", *Journal of Quantitative Analysis in Sports*, vol. 3, no. 3, 2007.

[Kuhn51] H. W. Kuhn and A. W. Tucker, "Nonlinear Programming", in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 481–492, Berkeley, Calif., 1951, University of California Press.

[Lee01] Daniel D. Lee and H. Sebastian Seung, "Algorithms for Non-negative Matrix Factorization", in T. K. Leen, T. G. Dietterich and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pp. 556–562, MIT Press, 2001.

[Lewi04] Michael Lewis, *Moneyball: The art of winning an unfair game*, WW Norton & Company, 2004.

[Oliv11] Dean Oliver, Kevin Pelton and Dan T. Rosenbaum, "Journal of Quantitative Analysis in Sports A Starting Point for Analyzing Basketball Statistics", 2011.

[Pasc12] Razvan Pascanu, Tomas Mikolov and Yoshua Bengio, "On the difficulty of training Recurrent Neural Networks", 2012.

[Peng02] Joanne Peng, Kuk Lee and Gary Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting", *Journal of Educational Research - J EDUC RES*, vol. 96, pp. 3–14, 09 2002.

[Rasc15] Sebastian Raschka, *Python Machine Learning*, Packt Publishing, 2015.

[Shle14] Jonathon Shlens, "A Tutorial on Principal Component Analysis", 2014.

[Silv08a] Nate Silver, "A better way to evaluate nba defense", https://fivethirtyeight.com/features/a-better-way-to-evaluate-nba-defense/, 2008. [Online; accessed 19-July-2008].

[Silv08b] Nate Silver and Reuben Fischer-Baum, "how we calculate nba elo ratings", https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/, 2008. [Online; accessed 19-July-2008].

[Stau19] Ralf C. Staudemeyer and Eric Rothstein Morris, "Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks", 2019.

[Tami08] Michael Tamir and Gal Oz, "Real-Time Objects Tracking and Motion Capture in Sports Events", August 14 2008. US Patent App. 11/909,080.

[Tibs94] Robert Tibshirani, "Regression Shrinkage and Selection Via the Lasso", *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 58, pp. 267–288, 1994.

[Tibs96] Robert Tibshirani, "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[Troi16] Michael Troilo, Adrien Bouchet, Timothy L. Urban and William A. Sutton, "Perception, reality, and the adoption of business analytics: Evidence from North American professional sport organizations", *Omega*, vol. 59, pp. 72 – 83, 2016. Business Analytics.

[Tver89] Amos Tversky and Thomas Gilovich, "The "hot hand": Statistical reality or cognitive illusion?", *Chance*, vol. 2, no. 4, pp. 31–34, 1989.

[Wald43] A. Wald and J. Wolfowitz, "An Exact Test for Randomness in the Non-Parametric Case Based on Serial Correlation", *Ann. Math. Statist.*, vol. 14, no. 4, pp. 378–388, 12 1943.

[Wier15] Wessel Wieringen, "Lecture notes on ridge regression", 09 2015.

[Wins09] Wayne L. Winston, *Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football*, Princeton University Press, 2009.

[Κα17] Χρυσηΐς Καρώνη, *Στατιστικά μοντέλα παλινδρόμησης*, Συμεών, 2017.

Παράρτημα Α

Ευρετήριο Ακρωνυμίων και Συντμήσεων

Με αλφαβητική σειρά, ως προς τη σύντμηση

DRP: Area Under the Curve (Εμβαδόν κάτω από την καμπύλη)
DRP: Defensive Rebounding Percentage (Ποσοστό αμυντικών ριμπάουντ) στο μοντέλο των 4 παραγόντων του Dean Oliver
DTPP: Defensive Turnovers Caused per Possession (Αριθμός λαθών ανά κατοχή αντιπάλου) στο μοντέλο των 4 παραγόντων του Dean Oliver
EFG: Effective Field Goal percentage (Αποτελεσματικότητα της ευστοχίας) στο μοντέλο των 4 παραγόντων του Dean Oliver
FTR: Free Throw Rate (Ποσοστό επιθέσεων που οδηγούν σε ελεύθερες βολές) στο μοντέλο των 4 παραγόντων του Dean Oliver
GRU: Gated Array Unit
LSTM: Long Short-Term Memory networks
NBA: National Basketball Association (Εθνική Καλαθοσφαιρική Ομοσπονδία των Ηνωμένων Πολιτειών)
OEFG: (Opponent's Effective Field Goal Percentage (Αποτελεσματικότητα ευστοχίας του αντιπάλου) στο μοντέλο των 4 παραγόντων του Dean Oliver
OFTR: Opponent's Free Throw Rate (Ποσοστό ελεύθερων βολών αντιπάλου) στο μοντέλο των 4 παραγόντων του Dean Oliver
ORP: Offensive Rebounding Percentage (Ποσοστό επιθετικών ριμπάουντ) στο μοντέλο των 4 παραγόντων του Dean Oliver
PCA: Principal Component Analysis (Ανάλυση Κυρίων Συνιστωσών)
ROC: Receiver Operating Characteristic (Καμπύλη Λειτουργικών Χαρακτηριστικών)
SVD: Singular Value Decomposition (Ανάλυση Ιδιαζουσών Τιμών)
SVM: Support Vector Machines (Μηχανές διανυσμάτων υποστήριξης)
TPP: Turnovers Committed per Possession (Αριθμός λαθών ανά κατοχή) στο μοντέλο των 4 παραγόντων του Dean Oliver
XGBoost: Τεχνική Extreme Gradient Boosting [Frie01]