



**ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΛΟΓΙΣΜΙΚΟΥ ΚΑΙ ΑΝΑΠΤΥΞΗΣ ΕΦΑΡΜΟΓΩΝ**

**Ανάλυση Δεδομένων για την Προγνωστική
Μοντελοποίηση του Μισθού των Παικτών του NBA**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παναγιώτης Αραπάκης, ΑΕΜ:57910

Επιβλέπων Καθηγητής: Αυγερινός, Αραμπατζής, Καθηγητής, Τμήμα
Ηλεκτρολόγων Μηχανικών και Μηχανικών
Υπολογιστών, Δ.Π.Θ.

Ξάνθη, 2025



**ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ
ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΛΟΓΙΣΜΙΚΟΥ ΚΑΙ ΑΝΑΠΤΥΞΗΣ ΕΦΑΡΜΟΓΩΝ**

**Ανάλυση Δεδομένων για την Προγνωστική Μοντελοποίηση
του Μισθού των Παικτών του NBA**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Παναγιώτης Αραπάκης, ΑΕΜ:57910

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Επιβλέπων Καθηγητής: Αυγερινός Αραμπατζής, Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Δημοκρίτειο Πανεπιστήμιο Θράκης

2ο Μέλος: Παύλος, Εφραιμίδης, Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Δημοκρίτειο Πανεπιστήμιο Θράκης

3ο Μέλος: Ιωάννης, Ανδρεάδης, Καθηγητής, Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Δημοκρίτειο Πανεπιστήμιο Θράκης

Ξάνθη, 2025



**DEMOCRITUS UNIVERSITY OF THRACE
SCHOOL OF ENGINEERING
DEPARTMENT OF ELECTRICAL AND COMPUTER
ENGINEERING
SECTOR OF SOFTWARE AND APPLICATION
DEVELOPMENT**

Data Analytics for Predictive Modeling of NBA Players' Salary

DIPLOMA THESIS

Panagiotis Arapakis, Registration Number:57910

COMMITTEE OF EXAMINERS

Supervisor: Avgerinos, Arampatzis, Professor, Department of Electrical and Computer Engineering, Democritus University of Thrace

Member 2: Pavlos, Efraimidis, Professor, Department of Electrical and Computer Engineering, Democritus University of Thrace

Member 3: Ioannis, Andreadis, Professor, Department of Electrical and Computer Engineering, Democritus University of Thrace

Xanthi, 2025

ΑΝΑΦΟΡΑ ΣΤΗΝ ΤΗΡΗΣΗ ΤΩΝ ΑΚΑΔΗΜΑΪΚΩΝ ΑΡΧΩΝ ΔΕΟΝΤΟΛΟΓΙΑΣ

Η παρούσα εργασία με τίτλο «Ανάλυση Δεδομένων για την Προγνωστική Μοντελοποίηση του Μισθού των Παικτών του NBA» είναι πρωτότυπη και πραγματοποιήθηκε από τον φοιτητή του Τμήματος Αραπάκη Παναγιώτη με Αρ. Μητρώου ΑΕΜ 57910 στο Εργαστήριο Προγραμματισμού και Επεξεργασίας Πληροφοριών του τομέα Λογισμικού και Ανάπτυξης Εφαρμογών, υπό την επίβλεψη του Καθηγητή Αραμπατζή Αυγερινού. Η συγγραφή της εργασίας πραγματοποιήθηκε εξολοκλήρου από τον φοιτητή Αραπάκη Παναγιώτη, υπό την καθοδήγηση και τις υποδείξεις του επιβλέποντά του Αραμπατζή Αυγερινό, και με την αναγνώριση της συμβολής της Υποψήφιου Διδάκτορα Σαφούρη Κωνσταντίνα.

Βεβαιώνεται ότι, κατά την εκπόνηση και τη συγγραφή της εργασίας του, ο φοιτητής τήρησε τα προβλεπόμενα από τον νόμο και τον αντίστοιχο εσωτερικό κανονισμό του Τμήματος, σεβάστηκε πλήρως τις Αρχές της Ακαδημαϊκής Ήθικής και του Κώδικα Δεοντολογίας, οι οποίες απαγορεύουν την παραποίηση των ερευνητικών/πειραματικών αποτελεσμάτων, την αναφορά ψευδών στοιχείων, την κατάχρηση της διανοητικής ιδιοκτησίας τρίτων και τη λογοκλοπή και ότι έγινε με σεβασμό στις αρχές.

Αφιερώσεις

Αφιερώνω αυτή τη δουλειά στην οικογένεια μου

Σε εσάς,

που στέκετε πίσω μου

όπως πάντα.

Ευχαριστίες

Θα ήθελα να εκφράσω την ευχαριστία μου σε όλους όσοι συνέβαλαν, άμεσα ή έμμεσα, στην ολοκλήρωση αυτής της εργασίας.

Κατ' αρχάς, ευχαριστώ θερμά τον Καθηγητή Δρ. Αυγερινό Αραμπατζή για τη συνεχή καθοδήγηση και την υποστήριξη κατά τη διάρκεια της έρευνας.

Έπειτα, θα ήθελα να ευχαριστήσω τα μέλη της τριμελούς επιτροπής αξιολόγησης της εν λόγω διπλωματικής, Δρ. Παύλο Εφραιμίδη Καθηγητή του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Δημοκρίτειου Πανεπιστήμιου Θράκης και Δρ. Ανδρεάδη Ιωάννη Καθηγητή του τμήματος Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Δημοκρίτειου Πανεπιστήμιου Θράκης για τον χρόνο τους.

Επίσης, εκτιμώ ιδιαίτερα τη βοήθεια της υποψήφιου διδάκτορα του τμήματος των Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Δημοκρίτειου Πανεπιστημίου Θράκης Σαφούρη Κωνσταντίνα για τις χρήσιμες πληροφορίες και τις εποικοδομητικές συμβουλές.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου για την ατέρμονη αγάπη, την κατανόηση και την ενθάρρυνση σε κάθε βήμα μου.

Περίληψη

Το άθλημα της καλαθοσφαίρισης αποτελεί ένα από τα πιο δημοφιλή αθλήματα σε παγκόσμιο επίπεδο. Αγώνες μπάσκετ πραγματοποιούνται και μεταδίδονται ζωντανά κάθε μέρα. Το πρωτάθλημα NBA (National Basketball Association), το οποίο λαμβάνει χώρα στην Αμερική, θεωρείται το κορυφαίο πρωτάθλημα μπάσκετ στον κόσμο και παίκτες κάθε εθνικότητας ονειρεύονται να συμμετέχουν σε αυτό.

Η μεγάλη προβολή και το παγκόσμιο ενδιαφέρων για τους αγώνες αυτού του αθλήματος οδηγεί στην συλλογή και δημοσιοποίηση πολλών δεδομένων και στατιστικών σχετικά με την έκβαση των αγώνων, τους παίκτες και τις ομάδες τους. Αυτός ο εκτεταμένος όγκος δεδομένων μπορεί να εκμεταλλευτεί μέσω των μεθόδων Μηχανικής Μάθησης για την εκπαίδευση κατάλληλων μοντέλων υπεύθυνων για την πρόβλεψη, την λήψη αποφάσεων και εξαγωγή συμπερασμάτων.

Η παρούσα εργασία βασίζεται σε ατομικά στατιστικά παικτών του NBA που έχουν καταγραφεί κατά τις αγωνιστικές περιόδους 2000-01 έως 2022-23. Στόχος της εργασίας είναι η πρόβλεψη του μισθού των παικτών βάση αυτών των στατιστικών μέσω μεθόδων Μηχανικής Μάθησης. Αυτό επιτυγχάνεται, συνοπτικά, μέσω της Συλλογής των δεδομένων (Data collection), την Επεξεργασία τους (Data preprocessing), την Διερευνητική Ανάλυση (Exploratory Data Analysis) αυτών, την εκπαίδευση και αξιολόγηση μοντέλων παλινδρόμησης και την βελτιστοποίηση (Optimization) τους. Αυτά τα βήματα πραγματοποιούνται με στόχο την επιλογή των μοντέλων που εμφανίζουν τις ακριβέστερες προβλέψεις καθώς και την επισήμανση των χαρακτηριστικών που επηρεάζουν σημαντικά την τιμή της τελικής πρόβλεψης.

Σκοπός της εργασίας είναι η συνεισφορά της στο χώρο του αθλητισμού και γενικά στον τομέα των Sport Analytics (Ανάλυση δεδομένων για τον αθλητισμό) διερευνώντας τη σχέση μεταξύ των στατιστικών απόδοσης των παικτών του NBA και των αντίστοιχων μισθών τους και παρέχοντας μελλοντικά ένα μέσο ανάλυσης δεδομένων που μπορεί να παρέχει πρακτικές

πληροφορίες, χρήσιμες στους παράγοντες των ομάδων (πχ. ιδιοκτήτες ομάδας, παίκτες, μάνατζερ).

Τα δεδομένα της εργασίας λήφθηκαν σε συνδυασμό από το NBA api και τον ιστότοπο HoopsHype (<https://hoopshype.com>).

Λέξεις κλειδιά: Διερευνητική Ανάλυση δεδομένων, Συλλογή δεδομένων, Επεξεργασία δεδομένων, Μοντέλα παλινδρόμησης, Βελτιστοποίηση, Μηχανική Μάθηση, Στατιστικά απόδοσης, Πρόβλεψη μισθού.

Abstract

The sport of basketball is one of the most popular sports worldwide. Basketball matches are held and broadcast live every day. The NBA (National Basketball Association) championship, which takes place in America, is considered the top basketball championship in the world and players of all nationalities dream of participating in it.

The great visibility and global interest in the matches of this sport leads to the collection and publication of a lot of data and statistics regarding the outcome of the matches, the players and their teams. This extensive volume of data can be exploited through Machine Learning methods to train appropriate models responsible for prediction, decision-making and drawing conclusions.

This work is based on individual statistics of NBA players recorded during the 2000-01 to 2022-23 seasons. The aim of the work is to predict the salary of players based on these statistics through Machine Learning methods. This is achieved, in short, through the Collection of data, Preprocessing, Exploratory Data Analysis, the training and evaluation of regression models and their optimization. These steps are carried out with the aim of selecting the models that show the most accurate predictions as well as highlighting the characteristics that significantly affect the value of the final prediction.

The purpose of the work is its contribution to the field of sports and in general to the field of Sport Analytics by investigating the relationship between the performance statistics of NBA players and their corresponding salaries and providing a future means of data analysis that can provide practical information, useful to team factors (e.g. owners, players, managers).

The data was obtained in combination from the NBA api and the HoopsHype website (<https://hoopshype.com>).

Keywords: Exploratory Data Analysis, Data Collection, Data Preprocessing, Regression Models, Machine Learning, Optimization, Performance Statistics, Salary Prediction.

Πίνακας περιεχομένων

1.	Εισαγωγή	12
1.1	Η εξέλιξη της ανάλυσης δεδομένων και η Μηχανική Μάθηση.....	12
1.2	Το NBA	13
1.3	Το Πρόβλημα.....	13
1.4	Στόχος της Εργασίας	14
1.5	Δομή της εργασίας.....	14
2.	Βιβλιογραφική Ανασκόπηση	15
2.1	Θεωρητικό Υπόβαθρο.....	15
2.2	Σχετική Έρευνα.....	16
3.	Μεθοδολογία	19
3.1	Συλλογή Δεδομένων – Data Collection	19
3.2	Προ-επεξεργασία Δεδομένων – Data Preprocessing	24
3.2.1	Καθαρισμός Δεδομένων	24
3.2.2	Μετασχηματισμός δεδομένων.....	25
3.2.3	Αποκοπή παικτών με περιορισμένα παιχνίδια.....	37
3.3	Διερευνητική Ανάλυση Δεδομένων – Exploratory Data Analysis	38
3.3.1	Μελέτη της Συσχέτισης μεταξύ των Χαρακτηριστικών	38
3.3.2	Επιλογή χαρακτηριστικών – Feature selection	41
3.4	Διαχωρισμός δεδομένων σε δεδομένα εκπαίδευσης και δοκιμής	41
3.5	Αξιολόγηση μοντέλων	42
3.6	Μετρικές.....	44
3.7	Μοντέλα.....	45
3.8	Αναζήτηση πλέγματος – Grid search	57
4.	Πειραματικά αποτελέσματα	58
4.1	Κυλιόμενο Παράθυρο - Rolling Window	58
4.2	Επεκτεινόμενο παράθυρο - Expanding Window.....	62
4.3	Τεχνική αναζήτησης πλέγματος - Grid search.....	65
4.4	Σημαντικά χαρακτηριστικά - Feature importance.....	67
4.5	Μέθοδοι συνόλου – Ensembles	71
4.6	Ακραίες τιμές - Outliers	76
5.	Συμπεράσματα.....	79
6.	Μελλοντικές Επεκτάσεις.....	80
7.	Παράρτημα: Βιβλιοθήκες	81
8.	Βιβλιογραφικές Αναφορές	83

1. Εισαγωγή

1.1 Η εξέλιξη της ανάλυσης δεδομένων και η Μηχανική Μάθηση

Η ανάλυση δεδομένων έχει καταστεί θεμελιώδες εργαλείο σε πολλούς κλάδους, επαναστατικοποιώντας τον τρόπο λήψης αποφάσεων και επίλυσης προβλημάτων. Στην πιο βασική της μορφή, η ανάλυση δεδομένων περιλαμβάνει την επιθεώρηση, τον καθαρισμό, τη μετατροπή και τη μοντελοποίηση των δεδομένων με σκοπό την ανακάλυψη χρήσιμων πληροφοριών, την υποστήριξη συμπερασμάτων και τη λήψη αποφάσεων. Αυτό που ξεκίνησε ως μια καθαρά στατιστική άσκηση, έχει εξελιχθεί πλέον σε έναν δυναμικό τομέα που ενσωματώνει πιο εξελιγμένες τεχνικές, όπως η προγνωστική ανάλυση, η μηχανική μάθηση και η τεχνητή νοημοσύνη. Αυτές οι εξελίξεις έχουν αυξήσει σημαντικά τη δύναμη των δεδομένων και των εφαρμογών τους σε διάφορους τομείς, όπως τα οικονομικά, την υγειονομική περίθαλψη και τον αθλητισμό. Η εξέλιξη της ανάλυσης δεδομένων μπορεί να αποδοθεί στην άνοδο της υπολογιστικής ισχύος και στην ανάπτυξη στατιστικών μοντέλων που επέτρεψαν την εις βάθος εξερεύνηση σύνθετων συνόλων δεδομένων. Καθώς τα σύνολα δεδομένων αυξάνονταν σε μέγεθος και πολυπλοκότητα, οι παραδοσιακές μέθοδοι στατιστικής ανάλυσης έφτασαν στα όριά τους. Αυτό άνοιξε το δρόμο για την ανάπτυξη της μηχανικής μάθησης (Machine Learning), η οποία επιτρέπει στους υπολογιστές να «μαθαίνουν» από τα δεδομένα και να κάνουν προβλέψεις χωρίς να έχουν προγραμματιστεί ρητά για συγκεκριμένους τρόπους απόφασης. Η μηχανική μάθηση, ένας υποτομέας της τεχνητής νοημοσύνης, περιλαμβάνει διάφορους αλγορίθμους που μπορούν να αναγνωρίζουν μοτίβα, να προβλέπουν αποτελέσματα και να βελτιστοποιούν αποφάσεις με βάση τα δεδομένα που τροφοδοτούμε. Από απλές γραμμικές παλινδρομήσεις μέχρι πολύπλοκα νευρωνικά δίκτυα, τα μοντέλα μηχανικής μάθησης έχουν βελτιώσει σημαντικά την ικανότητα ανάλυσης μεγάλων συνόλων δεδομένων, προσφέροντας πιο ακριβείς και χρήσιμες πληροφορίες. Η άνοδος των μεγάλων δεδομένων (Big Data), σε συνδυασμό με τις προόδους στην υπολογιστική ισχύ, έχουν επεκτείνει το εύρος και την αποτελεσματικότητα αυτών των τεχνικών σε

πολλούς τομείς, συμπεριλαμβανομένης της ανάλυσης δεδομένων στον αθλητισμό (sport analytics).

1.2 Το NBA

Στον επαγγελματικό αθλητισμό, και ειδικά στην Εθνική Ένωση Καλαθοσφαίρισης της Αμερικής (NBA), η ανάλυση δεδομένων έχει επιφέρει σημαντικές αλλαγές. Το NBA υπήρξε πρωτοπόρο στην υιοθέτηση προηγμένων στατιστικών, καθώς το πρωτάθλημα ενδιαφέρεται ιδιαίτερα για τη βελτίωση της απόδοσης των παικτών, την ανάπτυξη στρατηγικών και την ενίσχυση της αφοσίωσης των φιλάθλων. Αυτό που ξεκίνησε ως απλή καταγραφή στατιστικών αγώνων (BoxScore) έχει εξελιχθεί σε μια πολύπλοκη ανάλυση που χρησιμοποιείται από τις ομάδες για τη λήψη κρίσιμων αποφάσεων σε πραγματικό χρόνο κατά τη διάρκεια των αγώνων, αλλά και για μακροπρόθεσμες αποφάσεις που σχετίζονται με την διαχείριση της ομάδας και τις οικονομικές αποφάσεις.

Οι ομάδες του NBA χρησιμοποιούν τεράστιο όγκο δεδομένων για να παρακολουθούν την απόδοση των παικτών, να αξιολογούν τις ατομικές τους συνεισφορές και να προβλέπουν τη μελλοντική επιτυχία της ομάδας. Τα στατιστικά των παικτών, όπως οι πόντοι, τα ριμπάουντ και οι ασσίστ είναι μόνο η αρχή. Οι ομάδες αναλύουν πλέον προηγμένους δείκτες όπως ο ατομικός δείκτης απόδοσης παίκτη (Player Efficiency Rating - PER), το ποσοστό αληθινών σουτ (True Shooting Percentage - TS%) και το ποσοστό χρήσης (Usage Rate - USG%) για να αποκτήσουν μια βαθύτερη κατανόηση της αξίας των παικτών στο γήπεδο. Αυτές οι πληροφορίες καθοδηγούν τις αποφάσεις της ομάδας σχετικά με την απόκτηση παικτών, τις ανταλλαγές και τις διαπραγματεύσεις συμβολαίων, καθιστώντας τελικά την ανάλυση δεδομένων βασικό συστατικό της διαχείρισης ομάδων.

1.3 Το Πρόβλημα

Παρά την αυξανόμενη χρήση της ανάλυσης δεδομένων στο NBA, ο καθορισμός ενός δίκαιου και βασισμένου σε δεδομένα μισθού για έναν παίκτη παραμένει πρόκληση. Οι μισθοί επηρεάζονται από πολλούς παράγοντες,

όπως η απόδοση του παίκτη, το δυναμικό του, η εμπορευσιμότητα και οι ανάγκες της ομάδας. Ωστόσο, οι υποκειμενικές προκαταλήψεις μπορεί να οδηγήσουν σε διαφορές στην εκτίμηση της αξίας των παικτών. Με την αυξανόμενη διαθεσιμότητα λεπτομερών στατιστικών δεδομένων ανά σεζόν, σε συνδυασμό με τη δύναμη της μηχανικής μάθησης, υπάρχει η δυνατότητα να δημιουργηθούν προγνωστικά μοντέλα που εκτιμούν τους μισθούς πιο αντικειμενικά.

1.4 Στόχος της Εργασίας

Αυτή η εργασία στοχεύει να διερευνήσει τη σχέση μεταξύ των στατιστικών απόδοσης των παικτών του NBA και των αντίστοιχων μισθών τους, αξιοποιώντας αλγόριθμους μηχανικής μάθησης. Εστιάζοντας στα προσωπικά στατιστικά ανά σεζόν, στόχος είναι να προσδιοριστούν οι μετρικές απόδοσης που επηρεάζουν περισσότερο τους μισθούς και να εντοπιστούν τα πιο αποτελεσματικά προγνωστικά μοντέλα. Αυτή η έρευνα μπορεί συμβάλλει στον τομέα των Sport Analytics, καθώς και να παρέχει πρακτικές πληροφορίες για τις διαπραγματεύσεις συμβολαίων σε ομάδες, μάνατζερ και παίκτες.

1.5 Δομή της εργασίας

Η εργασία παρουσιάζει την ακόλουθη Δομή:

- Στο κεφάλαιο 1 βρίσκεται η εισαγωγή της εργασίας.
- Στο κεφάλαιο 2 θα γίνει η βιβλιογραφική ανασκόπηση, θα παρουσιαστεί το θεωρητικό υπόβαθρο και θα αναφερθούν σχετικά επιστημονικά άρθρα.
- Στο κεφάλαιο 3 θα γίνει ανάλυση επιστημονικής Μεθοδολογίας.
- Στο κεφάλαιο 4 θα παρουσιαστούν τα πειραματικά αποτελέσματα.
- Στο κεφάλαιο 5 θα καταγραφούν τα συμπεράσματα.
- Στο κεφάλαιο 6 θα αναφερθούν πιθανές μελλοντικές επεκτάσεις.
- Στο κεφάλαιο 7 (Παράρτημα) θα παρουσιαστούν οι βιβλιοθήκες που αξιοποιήθηκαν.
- Στο κεφάλαιο 8 (Βιβλιογραφικές Αναφορές) θα παρουσιαστεί η βιβλιογραφία στην οποία βασίστηκε η παρούσα εργασία.

2. Βιβλιογραφική Ανασκόπηση

2.1 Θεωρητικό Υπόβαθρο

Μηχανική Μάθηση

Ο όρος Μηχανική Μάθηση (Machine Learning) αναφέρεται για πρώτη φορά το 1959 από τον Αμερικανό πρωτοπόρο Άρθουρ Σάμουελ στο άρθρο του με τίτλο “Some Studies in Machine Learning Using the Game of Checkers” [1].

Η μηχανική μάθηση αποτελεί ένα σύνολο αλγορίθμων που μαθαίνει και βελτιώνεται από τα δεδομένα εισόδου με σκοπό να πραγματοποιεί μελλοντικές προβλέψεις. Στόχος της μηχανικής μάθησης είναι η επίλυση προβλημάτων πιο αποτελεσματικά και πιο αποδοτικά. Η μηχανική μάθηση συχνά χωρίζεται σε επιβλεπόμενη, μη επιβλεπόμενη και ενισχυτική μάθηση [2].

Στην επιβλεπόμενη μηχανική μάθηση, ο αλγόριθμος μαθαίνει από ετικέτες ή στόχους που περιέχονται στα δεδομένα εκπαίδευσης. Στη μη επιβλεπόμενη μηχανική μάθηση, τα δεδομένα εισόδου είναι αταξινόμητα και χωρίς ετικέτες, και ο αλγόριθμος πρέπει να ανακαλύψει μοτίβα στα δεδομένα. Η ενισχυτική μάθηση είναι ένας αλγόριθμος προσανατολισμένος σε στόχους. Η μέθοδος αυτή περιλαμβάνει έναν παράγοντα (agent) με έναν στόχο, και όταν πραγματοποιείται μια ενέργεια, ο παράγοντας λαμβάνει ανταμοιβή ή τιμωρία, ανάλογα με την ενέργεια σε σχέση με τον στόχο. Ο αλγόριθμος μαθαίνει μια συμπεριφορά που στοχεύει στη μεγιστοποίηση της ανταμοιβής.

Αυτή η εργασία επικεντρώνεται στην επιβλεπόμενη μηχανική μάθηση και πιο συγκεκριμένα σε μεθόδους παλινδρόμησης (regression).

2.2 Σχετική Έρευνα

Η πρόβλεψη των μισθών των παικτών του NBA έχει αποτελέσει θέμα σημαντικού ενδιαφέροντος στον τομέα των sport analytics με διάφορες μελέτες να διερευνούν τους παράγοντες που επηρεάζουν τις αμοιβές των παικτών. Σε αυτή την ενότητα εξετάζεται η σχετική βιβλιογραφία που έχει συμβάλει στην κατανόηση των καθοριστικών παραγόντων των μισθών των παικτών του NBA και των μεθοδολογιών που χρησιμοποιούνται για την πρόβλεψή τους.

Αρκετές μελέτες έχουν ταυτοποιήσει βασικές μετρικές απόδοσης (στατιστικά) που επηρεάζουν τους μισθούς των παικτών του NBA. Οι Lyons, Jackson, Livingston (2015) [3] χρησιμοποιώντας δεδομένα 243 παικτών από τη σεζόν 2013-2014 διαπίστωσαν μέσω μεθόδου πολλαπλής γραμμικής παλινδρόμησης ότι οι πόντοι ανά αγώνα (PPG), τα ριμπάουντ και τα προσωπικά φάουλ συμβάλλουν σημαντικά στις αποδοχές των παικτών, τονίζοντας τη σημασία των στατιστικών σκοραρίσματος και άμυνας.

Ομοίως, ο Zhao (2022) [4] με την χρήση μεθόδων πολλαπλής γραμμικής παλινδρόμησης εντόπισε ότι ο χρόνος συμμετοχής, η αξία αντικατάστασης (Value over Replacement Player - VORP) και η επιλογή ενός παίκτη ως All-Star είναι σημαντικοί παράγοντες πρόβλεψης των μισθών, τονίζοντας τον ρόλο των ατομικών μετρικών απόδοσης.

Οι Παπαδάκη και Τσαγκρής (2022) [5] υποστήριξαν περαιτέρω αυτά τα ευρήματα χρησιμοποιώντας έναν αλγόριθμο Random Forest για να προβλέψουν το μερίδιο μισθών (ποσοστό επί της μισθοδοσίας της ομάδας), επιτυγχάνοντας υψηλή ακρίβεια δίνοντας προτεραιότητα στα χρόνια εμπειρίας και στα παιχνίδια που ξεκίνησαν ως μέλος της αρχικής πεντάδας.

Ο Cao (2024) [6] ανέπτυξε ένα πολλαπλό γραμμικό μοντέλο παλινδρόμησης χρησιμοποιώντας πόντους, ασίστ και λάθη για την άμεση πρόβλεψη μισθών, διαπιστώνοντας ότι κάθε πόντος αύξησε τον ετήσιο μισθό κατά 946.567 δολάρια, ενώ τα λάθη των μείωσαν κατά 478.153 δολάρια για τη σεζόν 2022-2023.

Ο Zhang (2024) [7] μελετώντας τους 50 κορυφαίους ενεργούς παίκτες από το έτος 2000 έως το έτος 2022 συνέκρινε την Απλή Γραμμική Παλινδρόμηση (simple Linear regression) με την Πολλαπλή Γραμμική Παλινδρόμηση (multiple Linear regression). Διαπίστωσε ότι η Πολλαπλή Γραμμική Παλινδρόμηση απόδωσε καλύτερα και η «προς τα πίσω» εξάλειψη χαρακτηριστικών βελτίωσε την ερμηνευσιμότητα (interpretability) του μοντέλου, αλλά αύξησε ελαφρώς το σφάλμα.

Ο Xiong (2024) [8] χρησιμοποιώντας δεδομένα από τη σεζόν 2022-2023 του NBA, πραγματοποίησε γραμμική ανάλυση παλινδρόμησης με τα στατιστικά στοιχεία ανά αγώνα ως ανεξάρτητες μεταβλητές και τους μισθούς παίκτων ως εξαρτημένη μεταβλητή. Συμπέρανε ότι η αποτελεσματικότητα στο σουτ (PTS, 3PA, 2PA) και οι αμυντικές ικανότητες (REB, STL, BLK) έχουν σημαντική επίδραση στους μισθούς των παίκτων.

Μια μελέτη του 2023 (Feng, Wang, Xiong) [9] χρησιμοποιεί δεδομένα από 331 παίκτες του NBA κατά τη σεζόν 2020–2021 για την ανάπτυξη πολλαπλών μοντέλων γραμμικής παλινδρόμησης, συμπεριλαμβανομένων των OLS, Ridge, Lasso και Elastic Net. Αυτή η εργασία ενσωματώνει τόσο βασικά όσο και προηγμένα στατιστικά στοιχεία παίκτων για την πρόβλεψη των κατάλληλων μισθών των παίκτων.

Ο Wang (2024) [10] χρησιμοποιεί αλγόριθμο που εισάγει την τεχνική του gradient boosting για την πρόβλεψη μισθών της σεζόν 2022–23 εμφανίζοντας βελτίωση απόδοσης σε σχέση με προηγούμενες μεθόδους.

Εκτός από τις επιδόσεις στο γήπεδο, έχει αποδειχθεί ότι παράγοντες εκτός γηπέδου, όπως η παρακολούθηση στα μέσα κοινωνικής δικτύωσης, επηρεάζουν τους μισθούς των παίκτων. Οι Lee, Zhang, Lomotey, Watkins, Huang (2023) [11] διαπίστωσαν ότι οι παίκτες με μεγαλύτερο κοινό στα μέσα κοινωνικής δικτύωσης, ιδίως σε πλατφόρμες όπως το Instagram, τείνουν να κερδίζουν υψηλότερους μισθούς. Αυτό υποδηλώνει ότι η εμπορευσιμότητα και η δημόσια επιρροή ενός παίκτη μπορεί να επηρεάσει σημαντικά την αμοιβή του, ανεξάρτητα από την απόδοσή του στο γήπεδο.

Ενώ οι υπάρχουσες μελέτες έχουν συμβάλει σημαντικά στην κατανόηση των καθοριστικών παραγόντων του μισθού στο NBA, παραμένουν αρκετοί περιορισμοί. Ειδικότερα, πολλές μελέτες επικεντρώνονται σε μία μόνο σεζόν ή σε ένα περιορισμένο σύνολο παικτών, το οποίο μπορεί να μην αποτυπώνει τις μακροπρόθεσμες τάσεις (όπως την μεταβολή του ανώτατου ορίου μισθών - salary cap) ή την πλήρη ποικιλομορφία του συνόλου των παικτών του NBA. Παράλληλα μικρά σύνολα δεδομένων αυξάνουν τον κίνδυνο υπερπροσαρμογής (overfitting) των μοντέλων παλινδρόμησης. Επίσης πολλές μελέτες βασίζονται σε γραμμικά μοντέλα (linear models), τα οποία ενδέχεται να απλοποιούν υπερβολικά τις σχέσεις μεταξύ των μετρικών απόδοσης των παικτών και του μισθού και να μην καταφέρουν να αποτυπώσουν τις μη γραμμικές σχέσεις.

Στην παρούσα εργασία η χρήση πολλαπλών σεζόν και μεγαλύτερου συνόλου δεδομένων στοχεύει στη βελτίωση της αξιοπιστίας των μοντέλων και στην αποφυγή της υπερπροσαρμογής (overfitting) λαμβάνοντας υπόψη παράλληλα την εξελισσόμενη δυναμική του πρωταθλήματος. Συμπληρωματικά ο συνδυασμός γραμμικών και μη γραμμικών προσεγγίσεων μέσω των μεθόδων συνόλου (ensemble methods) στοχεύει να αποτυπώσει καλύτερα πολύπλοκες (μη-γραμμικές) σχέσεις.

3. Μεθοδολογία

3.1 Συλλογή Δεδομένων – Data Collection

Στα πλαίσια της εργασίας, για την επιλογή κατάλληλων δεδομένων, χρησιμοποιήθηκαν και δοκιμάστηκαν πολλαπλές πηγές. Τα δεδομένα που συλλέγονται αφορούν τα ετήσια στατιστικά επίδοσης (ανά παίκτη και ανά σεζόν) και τους αντίστοιχους ετήσιους μισθούς. Αρχικά ελέγχθηκε η καταλληλότητα των ήδη έτοιμων βάσεων δεδομένων που ήταν ανεβασμένα στο ιστότοπο Kaggle [12]. Στο Kaggle βρέθηκαν ορισμένες βάσεις δεδομένων που περιείχαν μόνες τους ή σε συνδυασμό δεδομένα που ήταν μερικώς κατάλληλα. Αυτές όμως απορρίφθηκαν λόγω περιορισμένου αριθμού μετρικών και ατομικών στατιστικών ή μη επαρκούς χρονολογικής κάλυψης. Δηλαδή θεωρήθηκε ότι ένας μεγαλύτερος και ποικιλότερος όγκος δεδομένων μπορεί να οδηγήσει σε καλύτερα τελικά αποτελέσματα στην πειραματική φάση της εργασίας. Η συλλογή των μισθολογικών δεδομένων των παικτών συνεχίστηκε με χρήση μεθόδων data scraping από ιστότοπους που συλλέγουν και διαμοιράζουν στατιστικά του NBA. Οι ιστότοποι που ελέγχθηκαν σε αυτή την φάση ήταν το hoopshype [13] και το ESPN [14]. Ο πρώτος (hoopshype) επικράτησε διότι, μεταγενέστερα, αποδείχθηκε ότι το ESPN είχε σημαντικά κενά στις καταγραφές του. Παράλληλα για την απόκτηση εκτενών ατομικών δεδομένων των παικτών έγινε χρήση του NBA API. Πιο συγκεκριμένα μέσω του API κλήθηκαν κοινά και προχωρημένα (Basic και Advanced) στατιστικά μέσω του endpoint με το όνομα «leaguedashplayerstats». Σε αυτές τις μετρικές προστέθηκαν πληροφορίες σχετικές με την χώρα προέλευσης, το σχολείο φοίτησης και την επιλογή του κάθε παίκτη στο NBA draft.

Στην συνέχεια παρατίθενται τα στατιστικά που αποτελούν την βάση δεδομένων της εργασίας.

ID	Όνομα Μεταβλητής	Σύντομη περιγραφή
0	Player	Όνομα του παίκτη του NBA.
1	salary	Ο μισθός του παίκτη για τη δεδομένη σεζόν (σε

		δολάρια).
2	season	Η σεζόν του NBA στη μορφή "YYYY-YY" (π.χ., 2023-24).
3	TEAM_ABBREVIATION	Συντομογραφία της ομάδας του παίκτη (π.χ., LAL για Los Angeles Lakers).
4	AGE	Η ηλικία του παίκτη
5	GP	Αγώνες στους οποίους συμμετείχε ο παίκτης κατά τη σεζόν.
	W	Συνολικοί αγώνες που κέρδισε η ομάδα του παίκτη στους οποίους συμμετείχε.
7	L	Συνολικοί αγώνες που έχασε η ομάδα του παίκτη στους οποίους συμμετείχε.
8	W_PCT	Ποσοστό νικών στους αγώνες που συμμετείχε ο παίκτης.
9	MIN	Μέσος όρος λεπτών συμμετοχής ανά αγώνα.
10	FGM	Εύστοχα σουτ ανά αγώνα.
11	FGA	Προσπάθειες για σουτ ανά αγώνα.
12	FG_PCT	Ποσοστό ευστοχίας σουτ (FGM/FGA).
13	FG3M	Εύστοχα τρίποντα ανά αγώνα.
14	FG3A	Προσπάθειες για τρίποντα ανά αγώνα.
15	FG3_PCT	Ποσοστό ευστοχίας στα τρίποντα (FG3M/FG3A).
16	FTM	Εύστοχες ελεύθερες βολές ανά αγώνα.
17	FTA	Προσπάθειες για ελεύθερες βολές ανά αγώνα.
18	FT_PCT	Ποσοστό ευστοχίας στις ελεύθερες βολές (FTM/FTA).
19	OREB	Μέσος όρος επιθετικών ριμπάουντ ανά αγώνα.
20	DREB	Μέσος όρος αμυντικών ριμπάουντ ανά αγώνα.
21	REB	Σύνολο ριμπάουντ ανά αγώνα (OREB + DREB).
22	AST	Ασίστ ανά αγώνα.
23	TOV	Λάθη ανά αγώνα.
24	STL	Κλεψίματα ανά αγώνα.

25	BLK	Κοψίματα ανά αγώνα.
26	BLKA	Αποτυχημένες προσπάθειες λόγω κοψίματος ανά αγώνα.
27	PF	Προσωπικά φάουλ ανά αγώνα.
28	PFD	Φάουλ που κέρδισε ο παίκτης ανά αγώνα.
29	PTS	Πόντοι που σημείωσε ανά αγώνα.
30	PLUS_MINUS	Μέσος όρος διαφοράς πόντων όταν ο παίκτης βρίσκεται στο παρκέ.
31	DD2	Αριθμός double-doubles που πέτυχε κατά τη σεζόν.
32	TD3	Αριθμός triple-double που πέτυχε κατά τη σεζόν.
33	OFF_RATING	Επιθετική αξιολόγηση του παίκτη (πόντοι ανά 100 κατοχές).
34	DEF_RATING	Αμυντική αξιολόγηση του παίκτη (πόντοι που δέχτηκε ανά 100 κατοχές).
35	NET_RATING	Καθαρή αξιολόγηση (OFF_RATING - DEF_RATING).
36	AST_PCT	Ποσοστό ασίστ (ποσοστό των σουτ των συμπαικτών που δημιούργησε ο παίκτης όταν ήταν στο παρκέ).
37	AST_TO	Αναλογία ασίστ προς λάθη.
38	AST_RATIO	Ασίστ ανά 100 κατοχές.
39	OREB_PCT	Ποσοστό επιθετικών ριμπάουντ (ποσοστό των επιθετικών ριμπάουντ που πήρε ο παίκτης).
40	DREB_PCT	Ποσοστό αμυντικών ριμπάουντ (ποσοστό των αμυντικών ριμπάουντ που πήρε ο παίκτης).
41	REB_PCT	Ποσοστό συνολικών ριμπάουντ.
42	TM_TOV_PCT	Ποσοστό λαθών της ομάδας (λάθη ομάδας ανά 100 κατοχές).
43	TS_PCT	Πραγματικό ποσοστό ευστοχίας (συνυπολογίζονται τα σουτ, τα τρίποντα και οι ελεύθερες βολές).

44	USG_PCT	Ποσοστό χρήσης (ποσοστό των φάσεων της ομάδας που χρησιμοποιεί ο παίκτης όταν είναι στο παρκέ).
45	PACE	Κατοχές της ομάδας ανά 48 λεπτά όταν ο παίκτης είναι στο παρκέ.
46	PIE	Εκτίμηση επιρροής του παίκτη (επίδραση στην επιτυχία της ομάδας).
47	POSS	Συνολικές κατοχές που χρησιμοποίησε ο παίκτης κατά τη σεζόν.
48	college	Κολλέγιο που φοίτησε ο παίκτης
49	country	Χώρα καταγωγής του παίκτη.
50	Draft_number	Θέση στο draft (π.χ., 1 για πρώτη επιλογή, "Undrafted" αν δεν επιλέχθηκε).

Πρόσθετες επεξηγήσεις σχετικά με τον προηγούμενο πίνακα

TS_PCT

Πραγματικό ποσοστό ευστοχίας, που λαμβάνει υπόψη σουτ, τρίποντα και ελεύθερες βολές. Υπολογίζεται ως: $TS_PCT = PTS / (2 \times (FGA + 0.44 \times FTA))$.

USG_PCT

Ποσοστό χρήσης. Υπολογίζεται ως: $USG_PCT = 100 \times (FGA + 0.44 \times FTA + TOV) / Team_Possessions$.

PACE

Κατοχές ανά 48 λεπτά. Υπολογίζεται ως: $PACE = 48 \times (Team_Possessions + Opponent_Possessions) / (2 * (Tm MP / 5))$.

PIE

Εκτίμηση επιρροής του παίκτη. Υπολογίζεται ως: $PIE = (PTS + FGM + FTM - FGA - FTA + DREB + (.5 * OREB) + AST + STL + (.5 * BLK) - PF - TO) /$

(GmPTS + GmFGM + GmFTM - GmFGA - GmFTA + GmDREB + (.5 * GmOREB) + GmAST + GmSTL + (.5 * GmBLK) - GmPF - GmTO).

Προσθήκες στο σύνολο των στατιστικών

Συμπληρωματικά στα προηγούμενα στατιστικά προστέθηκε το inflated salary, δηλαδή ο μισθός των παικτών προσαρμοσμένος στον πληθωρισμό. Ο πληθωρισμός επηρεάζει την αξία των χρημάτων με την πάροδο του χρόνου, και οι μισθοί του NBA δεν αποτελούν εξαίρεση. Ένα συμβόλαιο 20 εκατομμυρίων το 2000 δεν ισοδυναμεί με ένα συμβόλαιο 20 εκατομμυρίων το 2023 λόγω της μειωμένης αγοραστικής δύναμης του χρήματος. Προσαρμόζοντας τους μισθούς στον πληθωρισμό, μπορούμε να συγκρίνουμε τα συμβόλαια σε διαφορετικές εποχές και να δυνητικά να παρέχουμε ακριβέστερες προβλέψεις. Αυτή η προσαρμογή των μισθών στην σημερινή τους αξία γίνεται με την χρήση του δείκτη CPI (consumer price index).

Επιπλέον έγινε η χρήση του salary cap το οποίο πρόκειται για το όριο που επιβάλλει το πρωτάθλημα στο συνολικό ποσό που μπορούν να ξοδέψουν οι ομάδες για τους μισθούς των παικτών και παίζει καθοριστικό ρόλο στον καθορισμό του πόσο αμείβονται οι παίκτες. Το salary cap έχει σχεδιαστεί για να προωθήσει την ανταγωνιστική ισορροπία, αποτρέποντας τις πιο πλούσιες ομάδες από το να ελέγχουν όλο το ταλέντο και προσαρμόζεται ετησίως. Το salary cap μπορεί να χρησιμοποιηθεί τόσο ως χαρακτηριστικό (feature) όσο και ως παράγοντας κανονικοποίησης στα μοντέλο μηχανικής μάθησης.

Στα πλαίσια της διπλωματικής εργασίας έγιναν οι εξής δοκιμές αναφορικά με το Salary Cap:

- Χρήση του salary cap για κάθε σεζόν ως ένα αυτόνομο χαρακτηριστικό. Αυτό επιτρέπει στο μοντέλο να μάθει πώς οι αλλαγές στο salary cap επηρεάζουν τους μισθούς. Για παράδειγμα, εάν το salary cap αυξηθεί σημαντικά σε μια συγκεκριμένη χρονιά, οι μισθοί των παικτών μπορεί επίσης να αυξηθούν αναλογικά.
- Κανονικοποίηση των μισθών των παικτών διαιρώντας τους με το salary cap για την αντίστοιχη σεζόν. Αυτό δημιουργεί έναν

κανονικοποιημένο μισθό ως προς το salary cap, που αντιπροσωπεύει τον μισθό του παίκτη ως ποσοστό του συνολικού salary cap. Η κανονικοποίηση αυτή βοηθά στην τυποποίηση των μισθών σε διαφορετικές εποχές, διευκολύνοντας τη σύγκριση παικτών από σεζόν με πολύ διαφορετικά επίπεδα salary cap.

3.2 Προ-επεξεργασία Δεδομένων – Data Preprocessing

3.2.1 Καθαρισμός Δεδομένων

Ο καθαρισμός των δεδομένων (Data cleaning) είναι μια από τις πιο κρίσιμες φάσεις σε οποιαδήποτε έρευνα ή ανάλυση δεδομένων. Βελτιώνει την ποιότητα των δεδομένων, διασφαλίζει την ακεραιότητα τους και εξαλείφει πιθανές πηγές ‘θορύβου’ που εισέρχονται κατά την συλλογή των δεδομένων.

Στο στάδιο του Καθαρισμού Δεδομένων της εργασίας πραγματοποιήθηκε:

- Διόρθωση των διπλότυπων εγγραφών που προήλθαν κατά την συγχώνευση των βάσεων δεδομένων των χαρακτηριστικών και των μισθών σε μια ενιαία συλλογή. Τέτοιες διπλότυπες εγγραφές ήταν τα χαρακτηριστικά τα οποία χρησιμοποιηθήκαν σαν σημεία αγκύρωσης (anchor points) κατά την συγχώνευση των πινάκων δεδομένων.
- Αφαίρεση του χαρακτήρα «\$» από τον μισθό των παικτών και διαχείριση των διαχωριστικών κομμάτων που αναπαριστούν τις χιλιάδες. Για την πρόβλεψη του μισθού μέσω μοντέλων μηχανικής μάθησης είναι αναγκαία η μετατροπή του σε αριθμητική μορφή. Πριν γίνει αυτό είναι αναγκαία η απομάκρυνση του χαρακτήρα δολαρίου «\$» και των κομμάτων μέσω χρήσης βιβλιοθηκών της python όπως η «pandas» η οποία είναι μια ανοιχτού κώδικα βιβλιοθήκη που διευκολύνει την επεξεργασία και ανάλυση δεδομένων. Δηλαδή με την βοήθεια της βιβλιοθήκης «pandas» καταγραφές δεδομένων μισθών της μορφής «\$40,000,000» καθαρίζονται και μετατρέπονται σε «40000000».

- Μετατροπή χαρακτηριστικών όπως ο μισθός σε αριθμητική μορφή.
- Επεξεργασία της στήλης «draft number» για αντιμετώπιση των τιμών «undrafted». Η στήλη του «draft number» αποτελείται από τιμές από το ένα έως στο εξήντα [1, 60] οι οποίες υποδηλώνουν τον αριθμό κλήσης του κάθε παίκτη στην λοταρία του NBA (Draft). Παίκτες μπορούν να εισέλθουν στο πρωτάθλημα του NBA ακόμα και στην περίπτωση που δεν επιλεγούν μέσω της διαδικασίας του “draft”, αυτοί είναι οι λεγόμενοι «Undrafted» παίκτες. Επομένως η στήλη «draft number» αποτελείται από αριθμητικές τιμές και κείμενο. Για ορθή χρήση αυτού του δεδομένου από τα μοντέλα μηχανικής μάθησης μετατρέπουμε την εγγραφή «Undrafted» στον αριθμό 100 που υποδηλώνει ότι ο παίκτης εισήλθε στο πρωτάθλημα όμως με μικρότερη προβλεπόμενη αξία. Ο αριθμός 100 επιλέχτηκε εμπειρικά.
- Διαγραφή των στηλών draft round, draft year που εισήλθαν κατά την συγχώνευση των δεδομένων καθώς δεν εισάγουν νέες πληροφορίες.
- Έλεγχος για ύπαρξη κενών τιμών (nan values). Υπήρχαν κενές τιμές στον αριθμό επιλογής του draft ορισμένων παικτών που δεν επιλέχθηκαν. Αυτές οι περιπτώσεις χειρίστηκαν όπως η προηγούμενη περίπτωση του «Undrafted».

3.2.2 Μετασχηματισμός δεδομένων

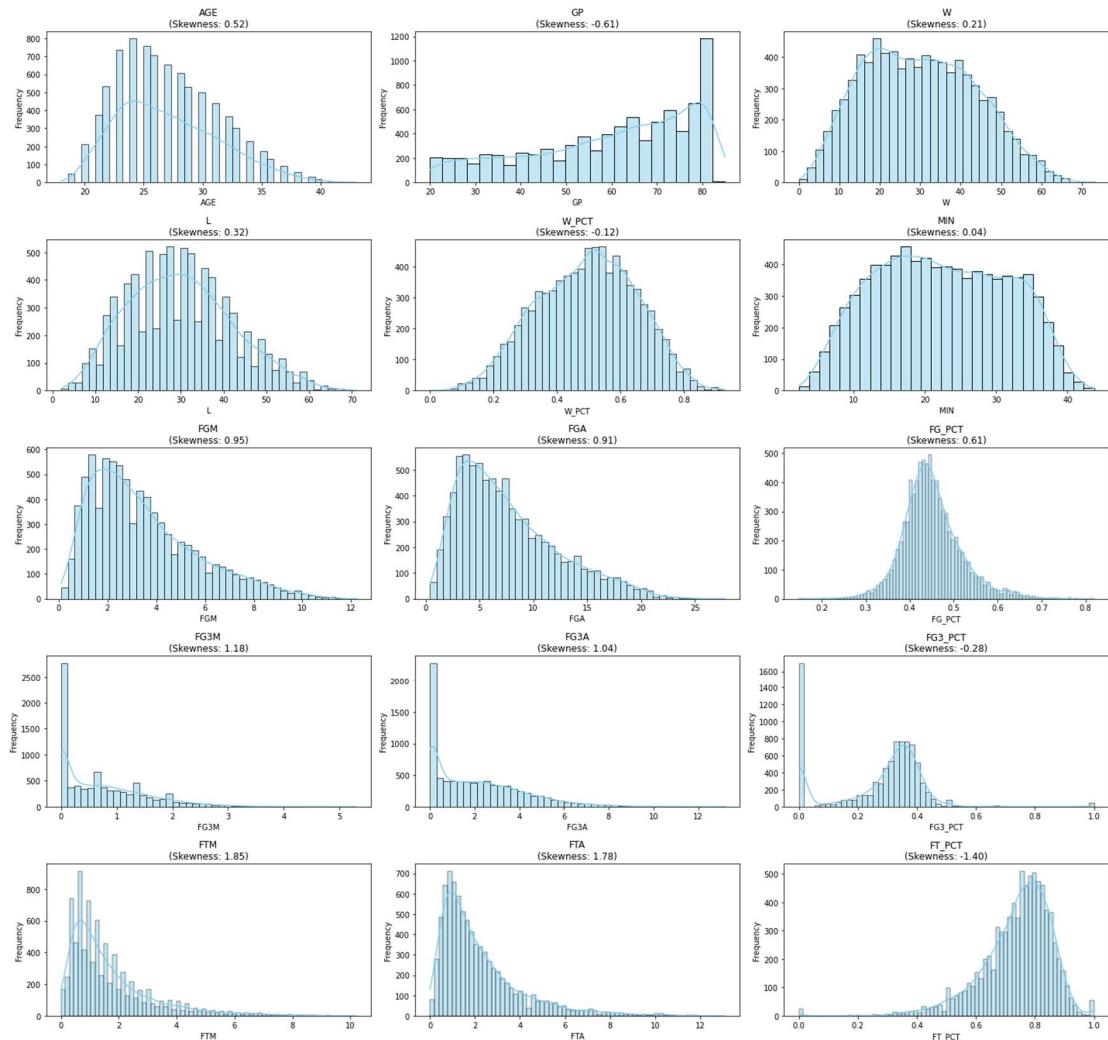
Κλιμάκωση και κανονικοποίηση

Στα πλαίσια της εργασίας εφαρμόστηκαν τεχνικές τροποποίησης και μετασχηματισμού των δεδομένων. Τέτοιες τεχνικές περιλαμβάνουν την κανονικοποίηση(normalization) και την κλιμάκωση (scaling) των δεδομένων. Αυτές διασφαλίζουν ότι οι μετρήσεις με υψηλότερα μεγέθη (πχ πόντοι) δεν επισκιάζουν άλλες με μικρότερα (πχ λάθη).

Ανάλυση κατανομών δεδομένων

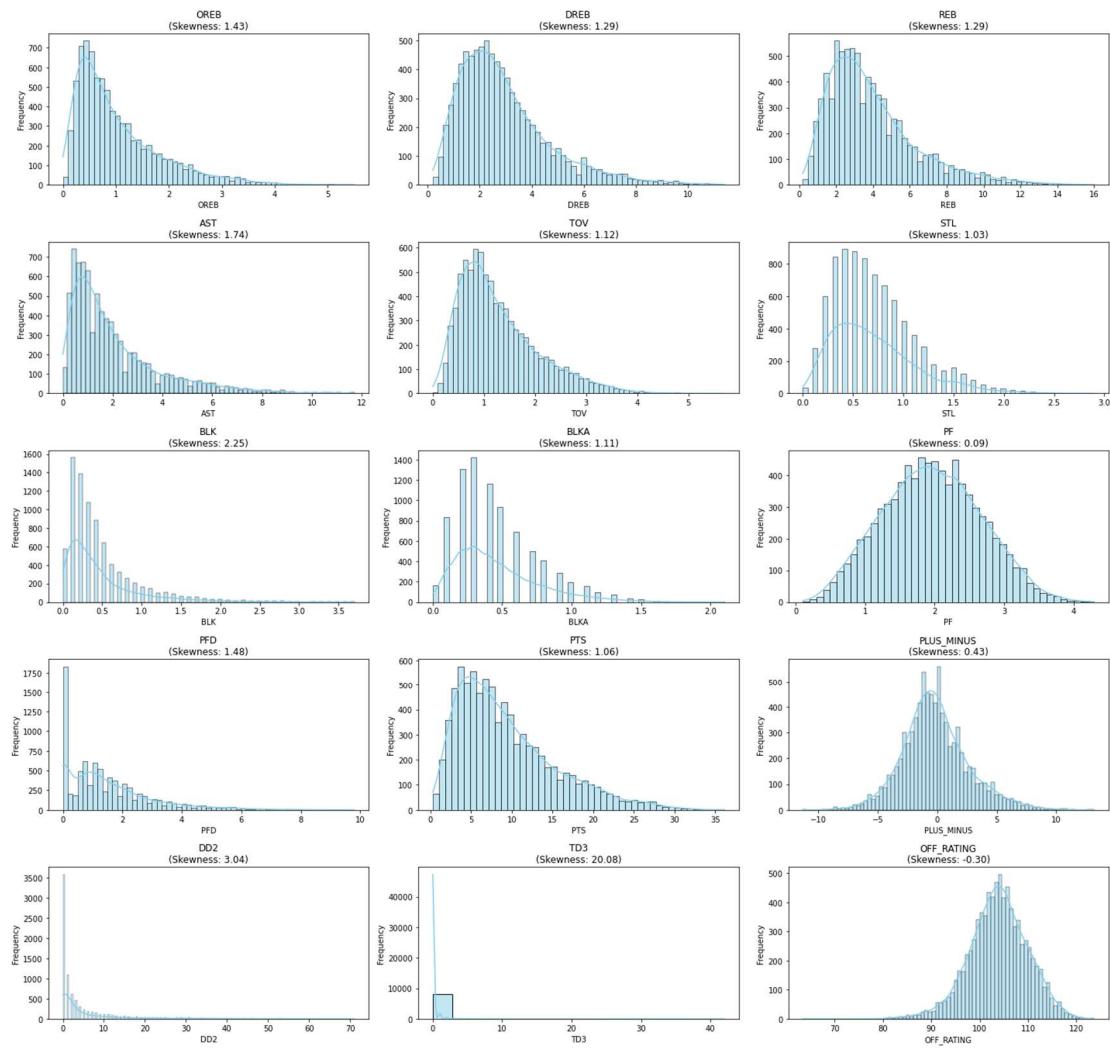
Επιπλέον σε αυτό το στάδιο έγινε η ανάλυση των κατανομών των δεδομένων και παρατηρήθηκε ποιες από αυτές τείνουν να ακολουθήσουν κανονική κατανομή (normal distribution) ή παρουσιάζουν κάποια λοξότητα. Ακολουθούν τα διαγράμματα των κατανομών των αριθμητικών χαρακτηριστικών των δεδομένων.

Figure 1: Feature Distributions



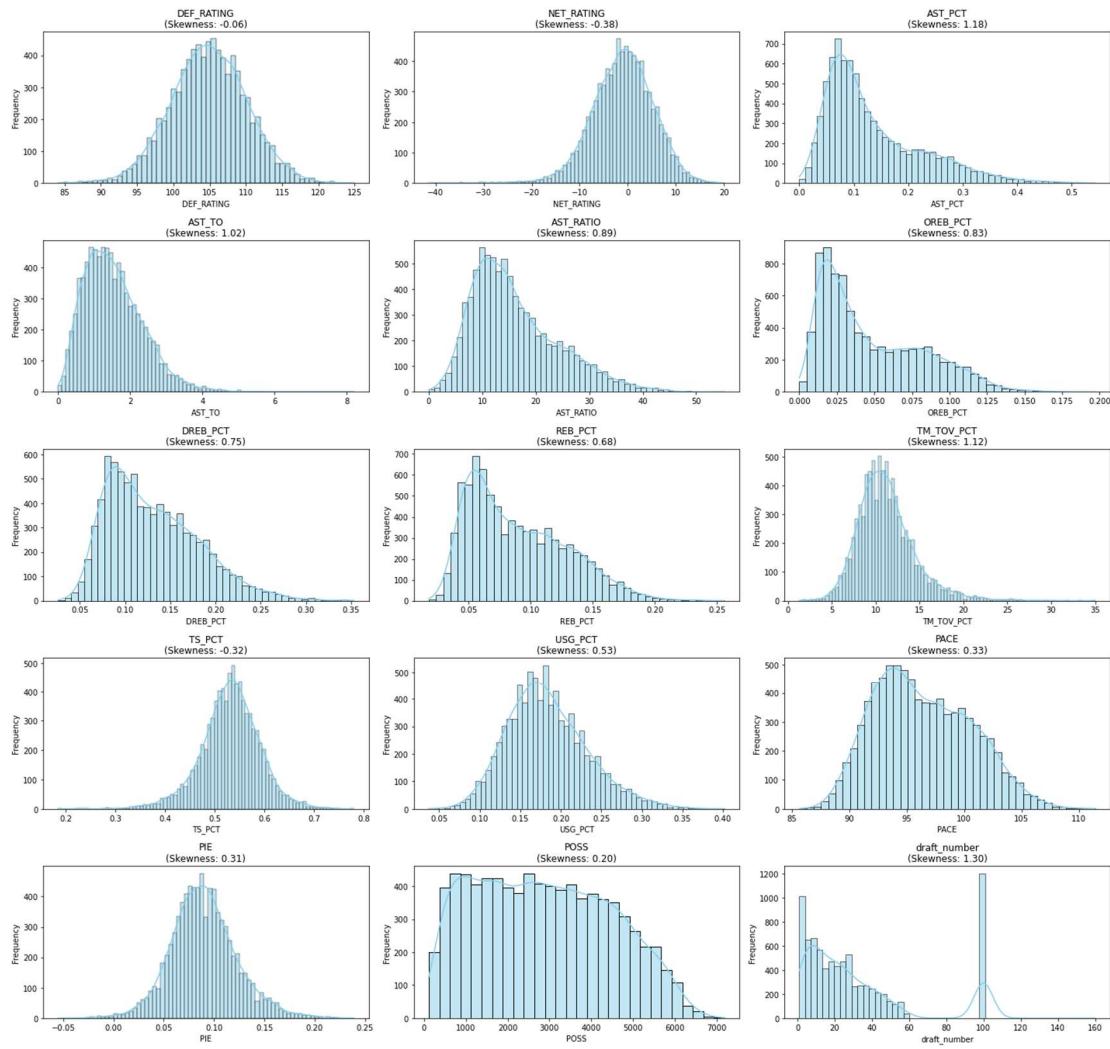
Εικόνα 3-1 Κατανομές δεδομένων πριν το μετασχηματισμό

Figure 2: Feature Distributions



Εικόνα 3-2 Κατανομές δεδομένων πριν το μετασχηματισμό

Figure 3: Feature Distributions



Εικόνα 3-3 Κατανομές δεδομένων πριν το μετασχηματισμό

Στα διαγράμματα παρατηρείται ότι υπάρχουν χαρακτηριστικά τα οποία ακολουθούν κατανομές πολύ κοντά στην κανονική κατανομή όπως τα PF (personal foul-προσωπικά φάουλ) και το DEF_rating (αμυντική αξιολόγηση). Ταυτόχρονα υπάρχουν πολλά χαρακτηριστικά τα οποία παρουσιάζουν λοξότητα (skewness) είτε προς τα αριστερά όπως πχ GP (games played-παιχνίδια που παίχτηκαν) και net_rating (καθαρή αξιολόγηση) είτε προς τα δεξιά όπως πχ age (ηλικία), rebounds (ριμπάουντ), steals (κλεψίματα). Τα χαρακτηριστικά τα οποία παρουσιάζουν λοξότητα προς τα δεξιά φαίνονται να είναι περισσότερα.

Τα λοξά δεδομένα μπορούν να προκαλέσουν προβλήματα σε εφαρμογές στατιστικής ανάλυσης και σε ορισμένα μοντέλα μηχανικής μάθησης. Πολλοί αλγόριθμοι μηχανικής μάθησης (π.χ. linear regression, logistic regression) υποθέτουν ότι τα χαρακτηριστικά εισόδου ακολουθούν την κανονική κατανομή ή είναι τουλάχιστον συμμετρικά. Τα λοξά (skewed) δεδομένα παραβιάζουν αυτή την υπόθεση, οδηγώντας σε κακή απόδοση του μοντέλου ή σε λανθασμένα συμπεράσματα. Για αυτό είναι χρήσιμος ο μετασχηματισμός των δεδομένων που παρουσιάζουν μεγάλη λοξότητα ο οποίος έχει το επιπλέον πλεονέκτημα της μείωσης της επιρροής των ακραίων τιμών το οποίο οδηγεί σε βελτίωση της απόδοσης του μοντέλου.

Μία μέθοδος μετασχηματισμού των δεδομένων με λοξή κατανομή είναι ο μετασχηματισμός Box-Cox. Ο μετασχηματισμός Box Cox είναι ένας μετασχηματισμός μη κανονικών εξαρτημένων μεταβλητών σε κανονικό σχήμα. Πήρε το όνομα του από τους στατιστικολόγους George Box και Sir David Roxbee Cox, οι οποίοι συνεργάστηκαν σε μια εργασία του 1964 και ανέπτυξαν την τεχνική. Ο μετασχηματισμός Box-Cox για μη αρνητικές αποκρίσεις είναι συνάρτηση της παραμέτρου λ . Το λάμδα (λ) συνήθως κυμαίνεται από -5 έως 5. Η «βέλτιστη τιμή» είναι αυτή που οδηγεί στην καλύτερη προσέγγιση της καμπύλης κανονικής κατανομής.

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(y), & \lambda = 0 \end{cases}$$

Με $\lambda = 1$ αντιστοιχεί σε μη μετασχηματισμό, με $\lambda = 1/2$ σε μετασχηματισμό τετραγωνικής ρίζας, με $\lambda = 0$ στο λογαριθμικό μετασχηματισμό και με $\lambda = -1$ στον αντίστροφο μετασχηματισμό, αποφεύγοντας έτσι την ασυνέχεια στο μηδέν. [15]

Ένα αρνητικό του μετασχηματισμού Box Cox είναι ότι λειτουργεί μόνο για θετικές τιμές δεδομένων. Στο σύνολο των δεδομένων αυτής της εργασίας περιέχονται και αρνητικές τιμές σε ορισμένα στατιστικά όπως το plus_minus (Μέσος όρος διαφοράς πόντων όταν ο παίκτης βρίσκεται στο παρκέ). Μια λύση αυτού του προβλήματος είναι η πρόσθεση μιας σταθεράς στα δεδομένα που περιέχουν αρνητικές τιμές. Στα πλαίσια αυτής της εργασίας όμως

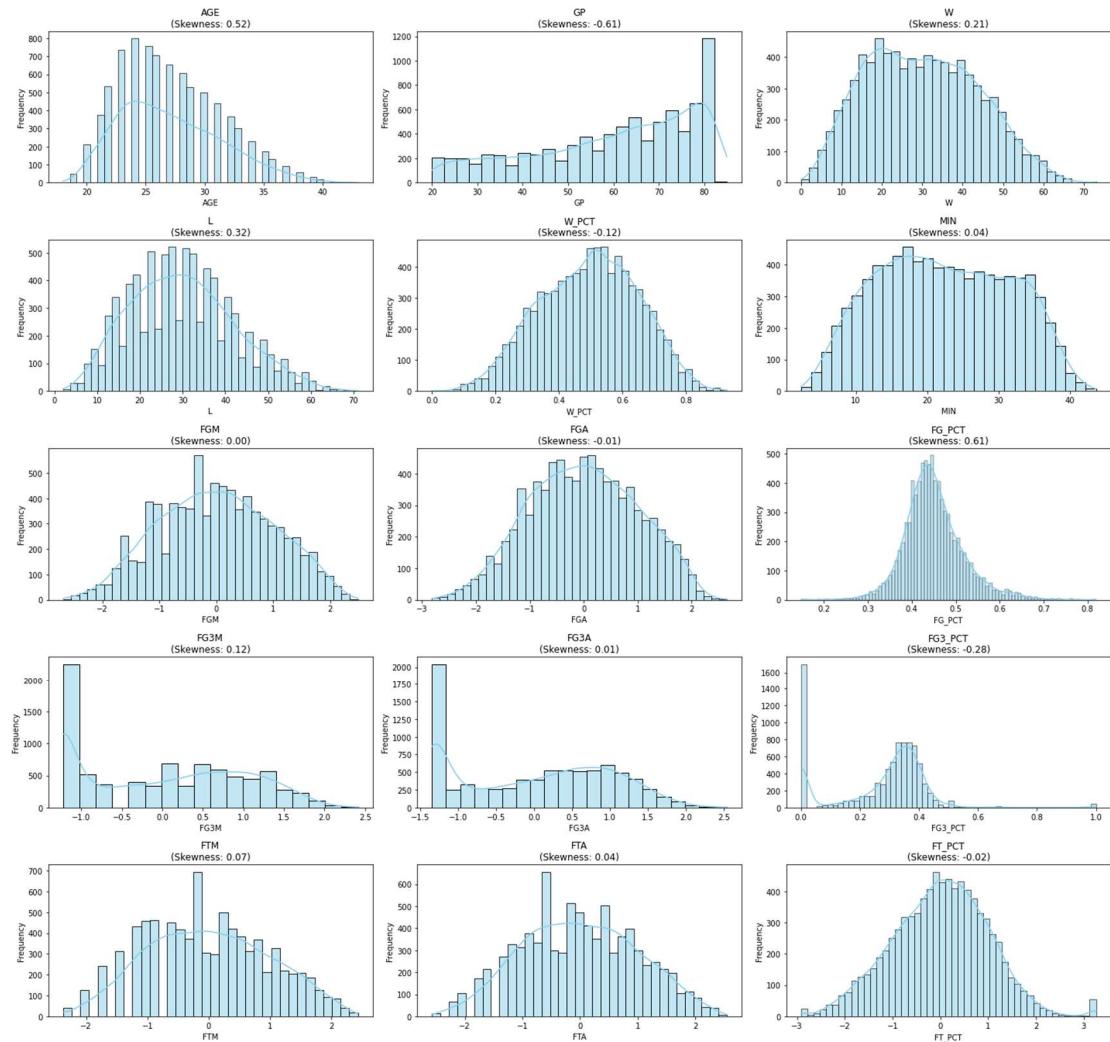
επιλέχθηκε η χρήση της μεθόδου Yeo-Johnson. Οι Yeo και Johnson (2000) [15], [16] επέκτειναν τον μετασχηματισμό Box-Cox σε παρατηρήσεις που μπορεί να είναι θετικές ή αρνητικές.

$$y_i^{(\lambda)} = \begin{cases} ((y_i + 1)^\lambda - 1)/\lambda & \text{if } \lambda \neq 0, y \geq 0 \\ \ln(y_i + 1) & \text{if } \lambda = 0, y \geq 0 \\ -((-y_i + 1)^{(2-\lambda)} - 1)/(2 - \lambda) & \text{if } \lambda \neq 2, y < 0 \\ -\log(-y_i + 1) & \text{if } \lambda = 2, y < 0 \end{cases}$$

Για μετασχηματισμό επιλέχθηκαν τα χαρακτηριστικά τα οποία παρουσίασαν λοξότητα (skewness) μεγαλύτερη από 0.75.

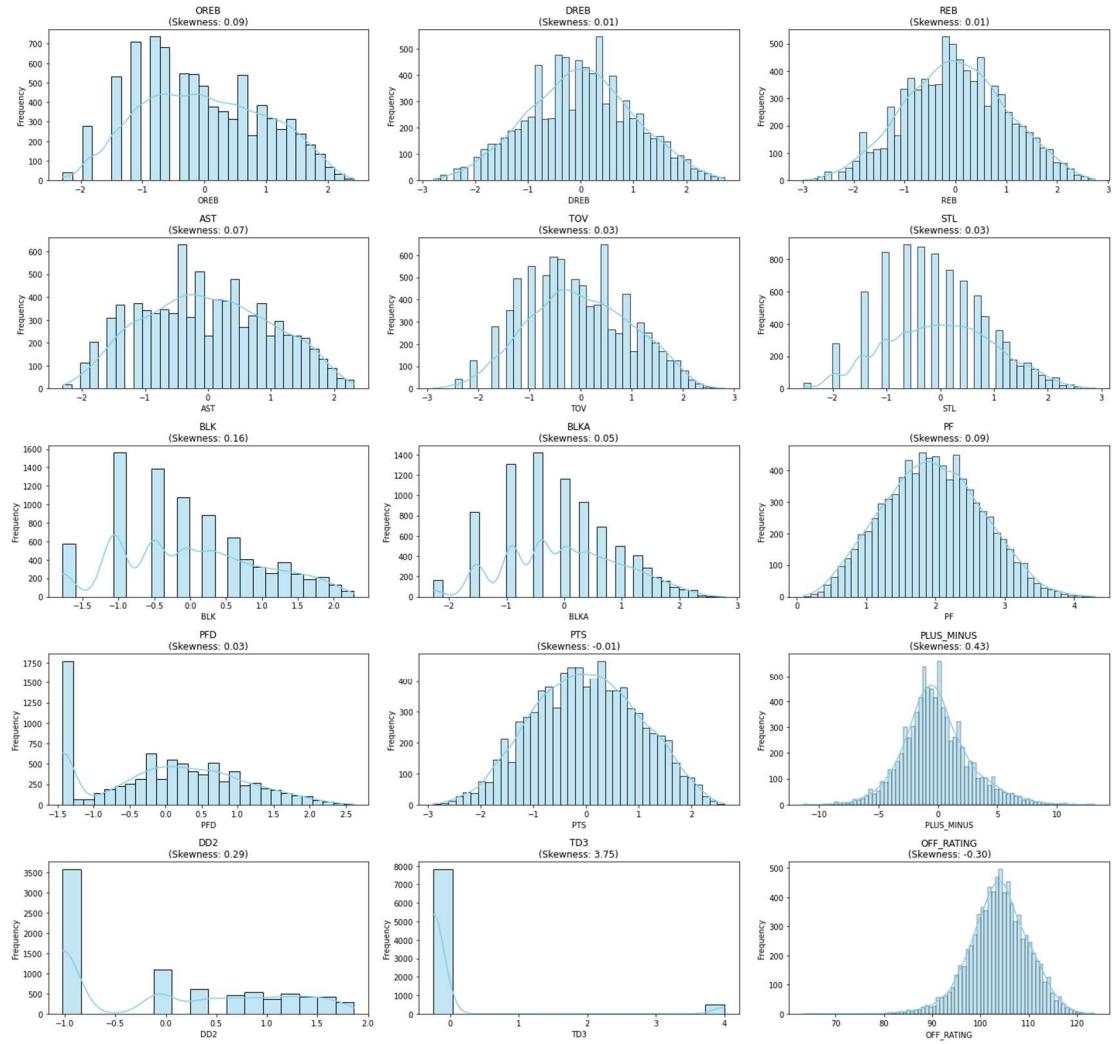
Ακολουθούν τα διαγράμματα των κατανομών των αριθμητικών χαρακτηριστικών των δεδομένων μετά την χρήση του μετασχηματισμού Yeo-Johnson.

Figure 1: Feature Distributions



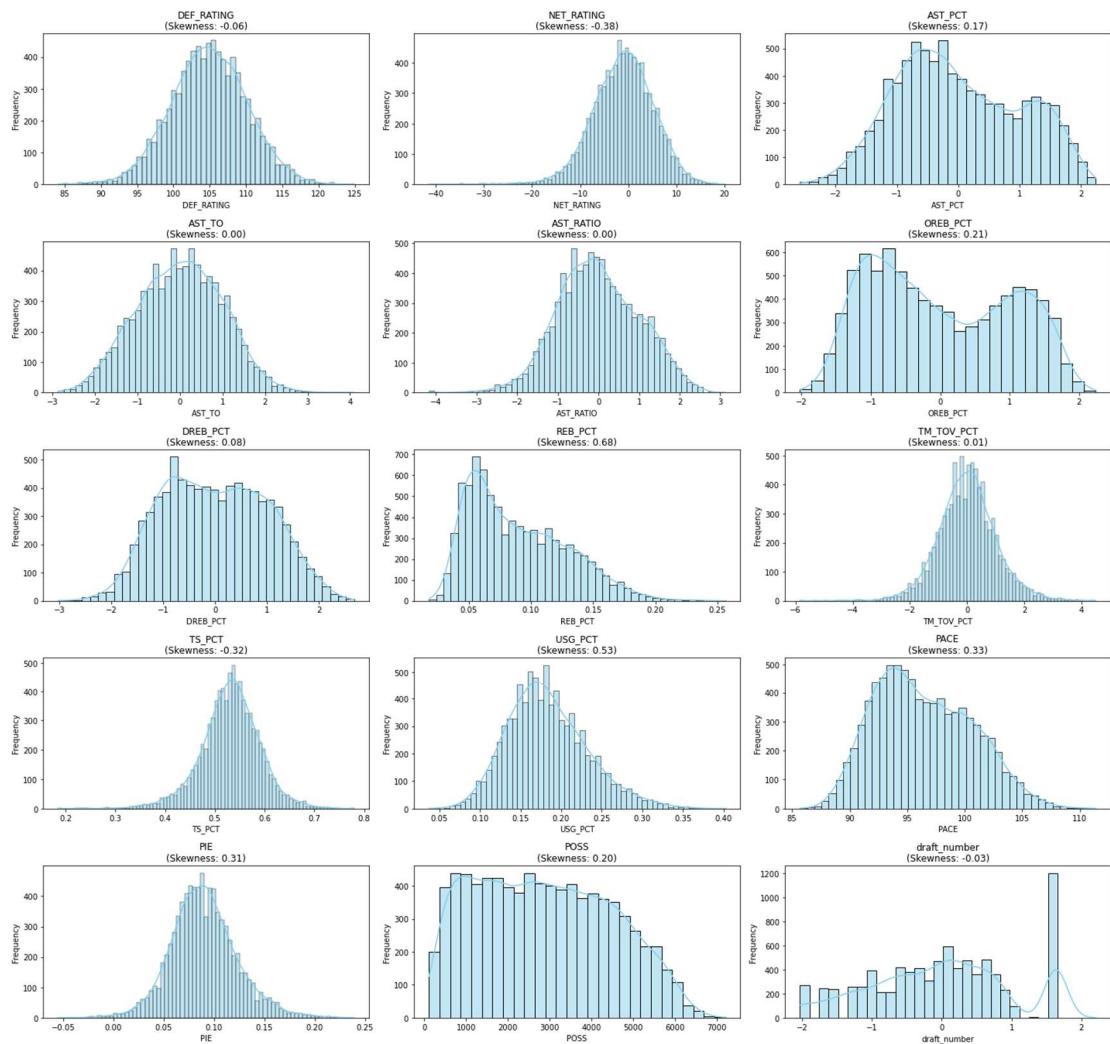
Εικόνα 3-4 Κατανομές δεδομένων μετά το μετασχηματισμό

Figure 2: Feature Distributions



Εικόνα 3-5 Κατανομές δεδομένων μετά το μετασχηματισμό

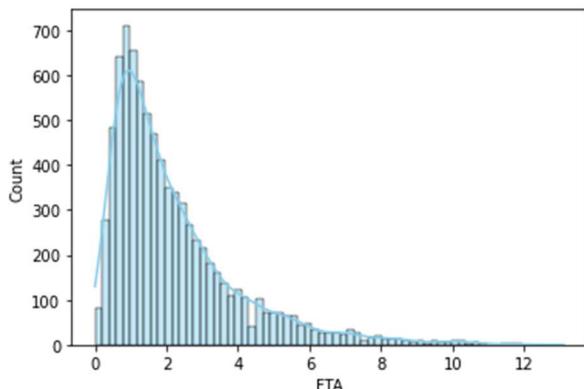
Figure 3: Feature Distributions



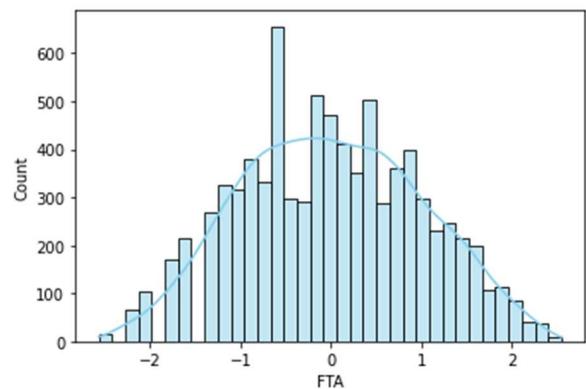
Εικόνα 3-6 Κατανομές δεδομένων μετά το μετασχηματισμό

Είναι φανερό ότι οι κατανομές που προηγουμένως εμφάνιζαν μεγάλη λοξότητα (skewness), πλέον έχουν «κεντραριστεί» και έχουν έρθει πιο κοντά στην χαρακτηριστική μορφή της κανονικής κατανομής (normal distribution).

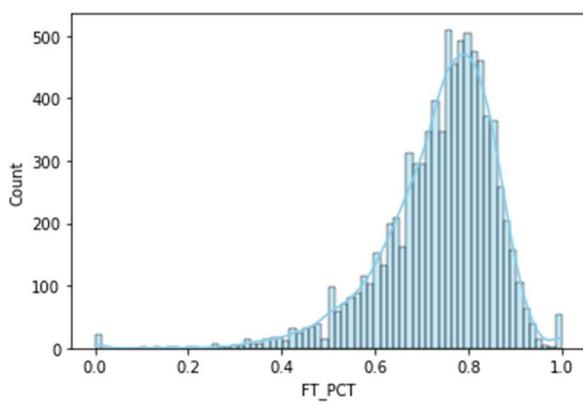
Για παράδειγμα ακολουθούν τα συγκριτικά παραδείγματα των FTA και FT_PCT πριν (αριστερά) και μετά (δεξιά) τον μετασχηματισμό Yeo-Johnson.



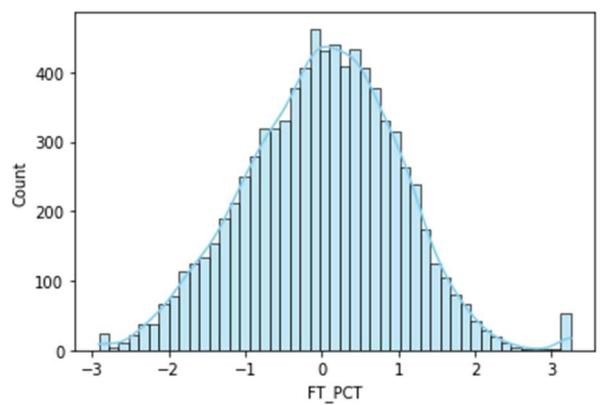
Εικόνα 3-9 3-10 Προσπάθειες ελευθέρων βολών πριν το μετασχηματισμό Yeo-Johnson



Εικόνα 3-7 3-8 Προσπάθειες ελευθέρων βολών μετά το μετασχηματισμό Yeo-Johnson



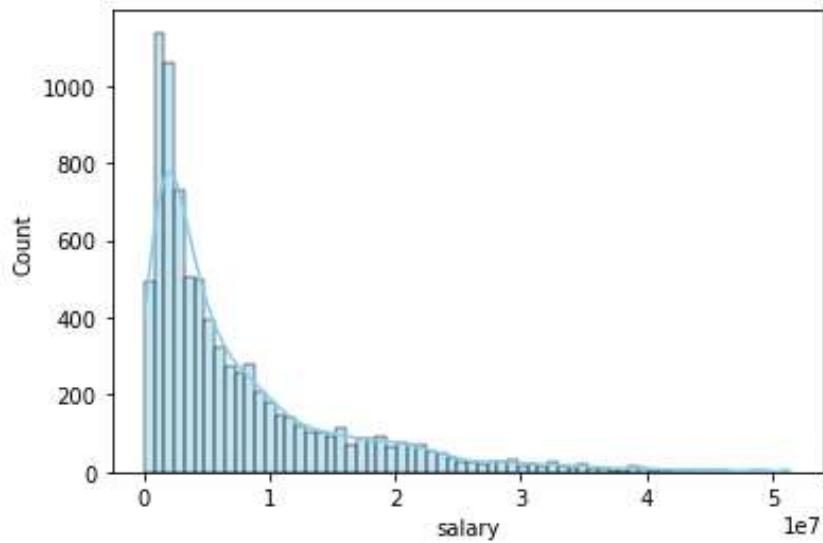
Εικόνα 3-12 3-13 Ποσοστών ελευθέρων βολών πριν το μετασχηματισμό Yeo-Johnson



Εικόνα 3-11 Ποσοστών ελευθέρων βολών μετά το μετασχηματισμό Yeo-Johnson

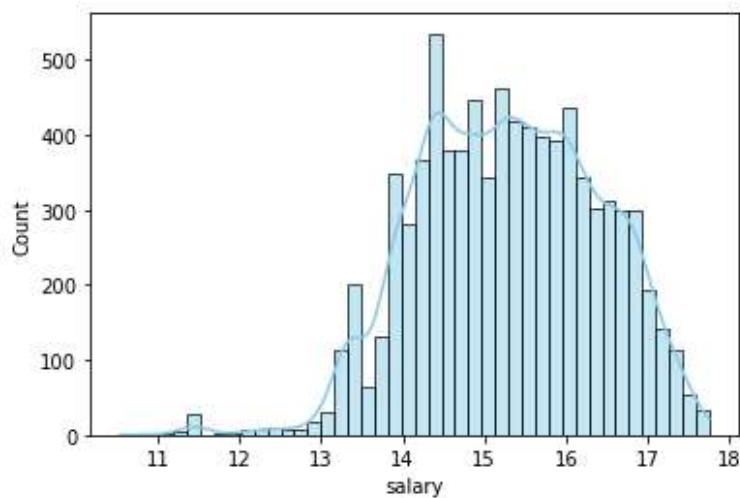
Μέχρι τώρα οι μετασχηματισμοί και τα διαγράμματα κατανομών αφορούν τα χαρακτηριστικά των δεδομένων και όχι την τιμή-στόχο (target value) που αποτελεί τον στόχο των μοντέλων παλινδρόμησης. Στη συνέχεια αναλύεται η κατανομή της τιμής-στόχου που είναι ο μισθός των παικτών του πρωταθλήματος.

Η κατανομή του μισθού των παικτών είναι αναμενόμενα λοξή προς τα δεξιά. Αυτή η λοξότητα είναι κοινό χαρακτηριστικό των μισθολογικών δεδομένων σε πολλούς κλάδους, αλλά είναι ιδιαίτερα έντονη στο NBA. Αυτό συμβαίνει διότι ένας περιορισμένος αριθμός ταλαντούχων παικτών «super star» κερδίζει σημαντικά υψηλότερους μισθούς σε σύγκριση με την πλειονότητα των αθλητών που έχουν υποστηρικτικούς ρόλους. Παραδείγματα τέτοιων παικτών είναι παίκτες όπως ο Λεμπρόν Τζέιμς, ο Στίβεν Κάρι και ο Κέβιν Ντουράντ, ονόματα που είναι γνωστά και εκτός του μπασκετικού χώρου.



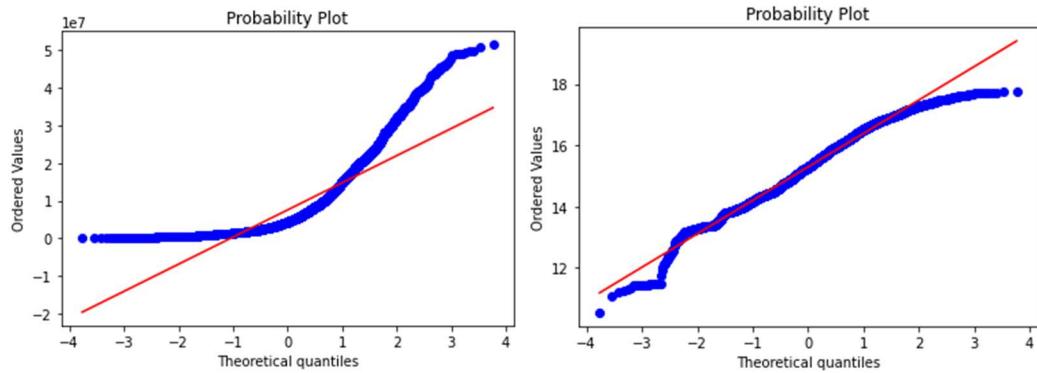
Εικόνα 3-14 Κατανομή του μισθού (τιμή-στόχος)

Εφόσον η κατανομή του μισθού στο NBA παρουσιάζει λοξότητα προς τα δεξιά εφαρμόζεται λογαριθμικός μετασχηματισμός. Ο λογαριθμικός μετασχηματισμός είναι απλούστερος του Yeo-Johnson και καθιστά πιο εύκολη την εφαρμογή του αντίστροφου μετασχηματισμού στη περίπτωση που χρειάζεται σύγκριση της προβλεπόμενης τιμής με την αρχική στην πειραματική διαδικασία. Ακολουθεί το διάγραμμα της κατανομής μετά τον μετασχηματισμό.



Εικόνα 3-15 Κατανομή του μισθού (τιμή-στόχος) μετά τον Λογαριθμικό μετασχηματισμό

Σύγκριση Q-Q διαγράμματος πριν (αριστερά) και μετά τον μετασχηματισμό (δεξιά)



Εικόνα 3-16 και 3-17 Διάγραμμα Q-Q για σύγκριση της κατανομής του μισθού πριν και μετά τον μετασχηματισμό

Το διάγραμμα Q-Q (quantile-quantile plot) είναι ένα γραφικό εργαλείο που χρησιμοποιείται για τη σύγκριση της κατανομής ενός συνόλου δεδομένων με μια θεωρητική κατανομή (εδώ με την κανονική κατανομή).

Παρατηρήσεις: Τα σημεία ευθυγραμίζονται κοντά στη γραμμή, ειδικά στα μεσαία quantile. Εμφανίζονται μικρές αποκλίσεις στα άκρα (πολύ χαμηλοί/υψηλοί μισθοί). Επιβεβαιώνεται ότι ο λογαριθμικός μετασχηματισμός μειώνει την ασυμμετρία και κάνει την κατανομή πιο συμμετρική καθιστώντας τη πιο κατάλληλη για στατιστικές προκλήσεις όπως η παλινδρόμηση.

Κωδικοποίηση κατηγορικών τιμών

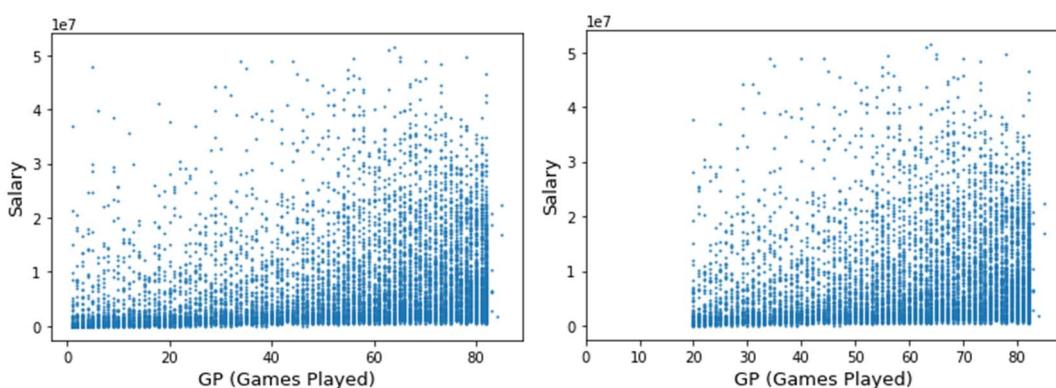
Στο σύνολο των δεδομένων περιέχονται κατηγορικές τιμές (categorical values) όπως η χώρα προέλευσης και το όνομα του κολεγίου ή των ομάδων των παικτών. Η κωδικοποίηση (encoding) αυτών είναι η διαδικασία μετατροπής τους σε αριθμητική μορφή κατάλληλη για χρήση σε αλγορίθμους μηχανικής μάθησης. Η τεχνική κωδικοποίησης που χρησιμοποιείται σε αυτή την εργασία είναι η κωδικοποίηση ετικέτας (label encoding) η οποία αντιστοιχεί έναν μοναδικό ακέραιο σε κάθε κατηγορία.

3.2.3 Αποκοπή παικτών με περιορισμένα παιχνίδια

Η ύπαρξη παικτών στο σύνολο των στατιστικών δεδομένων με περιορισμένο αριθμό παιχνιδιών σε μια συγκεκριμένη σεζόν μπορεί να οδηγήσει σε ανεπιθύμητα αποτελέσματα. Στα πλαίσια αυτής της εργασίας επιλέχθηκε η αποκοπή των παικτών που έχουν αγωνιστεί σε λιγότερα από 25 παιχνίδια κατά την διάρκεια μίας σεζόν.

Οι παίκτες που έχουν αγωνιστεί σε λιγότερα από 25 παιχνίδια έχουν συχνά περιορισμένα λεπτά συμμετοχής και ευκαιρίες να συμβάλουν ουσιαστικά στην απόδοση της ομάδας τους. Αυτό μπορεί να οδηγήσει στην ύπαρξη αναξιόπιστων στατιστικών στοιχείων όπως οι πόντοι ανά παιχνίδι ή ριμπάουντ τα οποία μπορεί να μην αντικατοπτρίζουν τη πραγματική τους απόδοση. Παράλληλα παίκτες με λιγότερα από 25 παιχνίδια μπορεί να έχουν χάσει σημαντικό χρόνο λόγω τραυματισμών, γεγονός που μπορεί να αλλοιώσει τα στατιστικά τους στοιχεία και να τους καταστήσει outliers (ακραίες τιμές). Συμπερασματικά η αποκοπή των παικτών που βρίσκονται σε αυτή την κατηγορία αποσκοπεί στην αποφυγή της υποβάθμισης της απόδοσης των μοντέλων μηχανικής μάθησης λόγω εισαγωγής θορύβου και αναξιόπιστων δεδομένων.

Διαγράμματα Παιχνιδιών GP(Games Played) – Μισθού (Salary) πριν (αριστερά) και μετά (δεξιά) την αποκοπή.



Εικόνα 3-18 και Εικόνα 3-19 Διάγραμμα Παιχνιδιών – Μισθού προ και μετά αποκοπής

3.3 Διερευνητική Ανάλυση Δεδομένων – Exploratory Data Analysis

Ένα σημαντικό στάδιο στην ανάλυση των δεδομένων που αποτελούν δεδομένα εκπαίδευσης μοντέλων μηχανικής μάθησης είναι η μελέτη της συσχέτισης μεταξύ αυτών. Ο εντοπισμός αυτών των συσχετίσεων μπορεί να οδηγήσει στην ορθή επιλογή ή απόρριψη χαρακτηριστικών.

3.3.1 Μελέτη της Συσχέτισης μεταξύ των Χαρακτηριστικών

Στα πλαίσια της εργασίας χρησιμοποιείται η συνηθής μέθοδος υπολογισμού της συσχέτισης (correlation) δύο μεταβλητών που ονομάζεται συντελεστής Pearson. Είναι ο λόγος μεταξύ της συνδιακύμανσης δύο μεταβλητών και του γινομένου των τυπικών αποκλίσεων τους.

Ο συντελεστής συσχέτισης Pearson (r) είναι ένα στατιστικό μέτρο που πιστοποιεί τη γραμμική σχέση μεταξύ δύο συνεχών μεταβλητών. Κυμαίνεται από -1 έως +1, όπου:

- +1 υποδεικνύει μια τέλεια θετική γραμμική σχέση,
- -1 υποδεικνύει μια τέλεια αρνητική γραμμική σχέση,
- 0 υποδηλώνει απουσία γραμμικής συσχέτισης, δηλαδή οι μεταβλητές είναι πλήρως ανεξάρτητες.

Αν $r > 0$, τότε οι μεταβλητές είναι θετικά συσχετισμένες, δηλαδή η αύξηση της μίας συνεπάγεται και την αύξηση της άλλης.

Αν $r < 0$, τότε οι μεταβλητές είναι αρνητικά συσχετισμένες, δηλαδή η αύξηση της μίας συνεπάγεται και τη μείωση της άλλης.

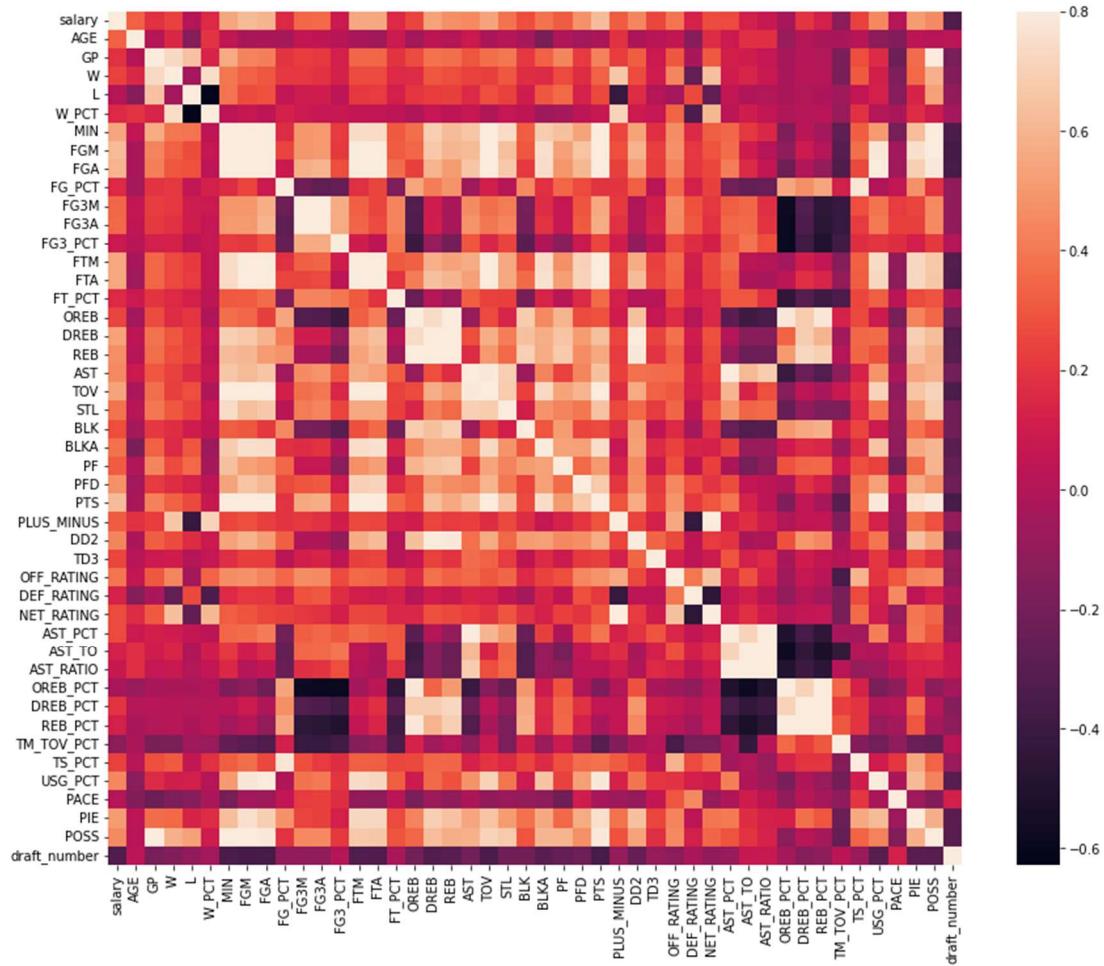
Ο συντελεστής r υπολογίζεται ως εξής:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

\bar{X}, \bar{Y} Μέσες τιμές των X, Y αντίστοιχα.

Ακολουθεί γραφική αναπαράσταση της συσχέτισης των δεδομένων τις παρούσας εργασίας.

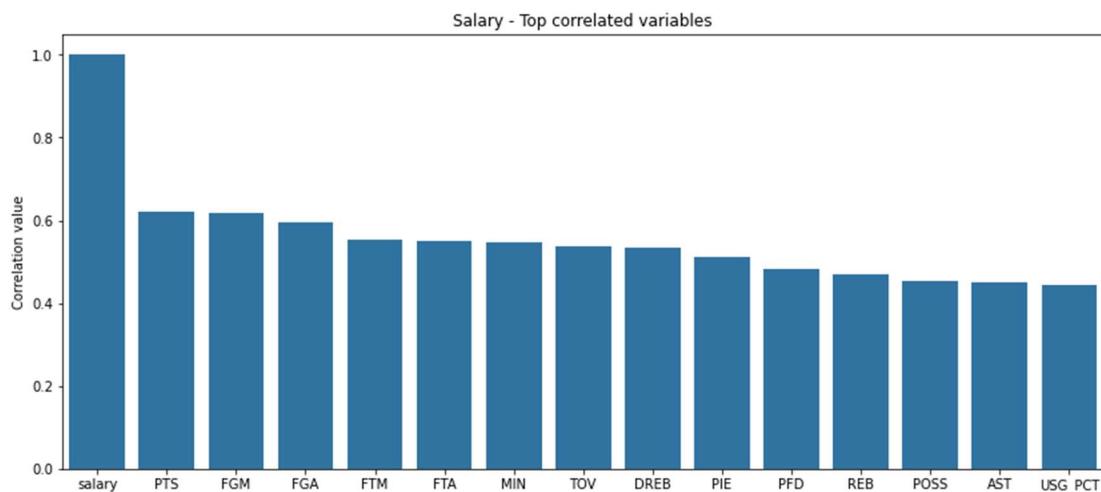
Το σκούρο χρώμα υποδηλώνει ότι οι δύο μεταβλητές έχουν χαμηλότερη συσχέτιση και αντίθετα όσο πιο ανοιχτό χρώμα, τόσο υψηλότερη συσχέτιση.



Εικόνα 3-20 πίνακας συσχέτισης (correlation matrix)

Με τη βοήθεια του παραπάνω πίνακα μπορούν εποπτικά να προκύψουν συμπεράσματα για το κάθε χαρακτηριστικό ξεχωριστά. Αρχικά έχει ενδιαφέρον η συσχέτιση του κάθε χαρακτηριστικού με την τιμή του μισθού που αποτελεί το στόχο πρόβλεψης.

Τα χαρακτηριστικά με την μεγαλύτερη συσχέτιση αναφορικά με τον μισθό των παικτών είναι:



Εικόνα 3-21 Κατάταξη συσχετιζόμενων μεταβλητών

Παρατηρήσεις: Με βάση την ανάλυση της συσχέτισης (correlation) παρατηρείται ότι στατιστικά όπως οι πόντοι, οι προσπάθειες σουτ, τα λεπτά και ο αριθμός των βιολών εμφανίζουν σημαντικό βαθμό συσχέτισης με τον μισθό. Επομένως μπορούμε να οδηγηθούμε στην υπόθεση ότι αυτοί οι δείκτες θα αποτελέσουν σημαντικά χαρακτηριστικά για την πρόβλεψη του μισθού.

Από την ανάλυση προκύπτει ότι κύρια στατιστικά παικτών (π.χ. πόντοι ανά παιχνίδι και προσπάθειες σουτ) παρουσιάζουν υψηλή συσχέτιση μεταξύ τους. Αυτό το γεγονός υποδηλώνει την εμφάνιση πολυσυγγραμμικότητας (multicollinearity). Η ύπαρξη πολυσυγγραμμικότητας μπορεί να οδηγήσει στην δυσκολία ερμηνείας των μοντέλων και πιθανόν τη μείωση απόδοσης.

Βέβαια ορισμένα μοντέλα παλινδρόμησης έχουν μεθόδους ρύθμισης (regularization) που καταπολεμούν την πολυσυγγραμμικότητα. Το μοντέλο Ridge (L2) και το Lasso (L1), αντιμετωπίζουν την πολυσυγγραμμικότητα εφαρμόζοντας ποινές στους συντελεστές. Το Ridge συρρικνώνει ομοιόμορφα τους συντελεστές, ενώ το Lasso μηδενίζει ασήμαντες μεταβλητές, προσφέροντας ταυτόχρονα επιλογή χαρακτηριστικών. Τα μοντέλα δέντρων (Random Forest, XGBoost, κλπ.) αντιμετωπίζουν φυσικά την πολυσυγγραμμικότητα, αφού διαχωρίζουν τα χαρακτηριστικά με βάση τη σημαντικότητά τους και χρησιμοποιούν υποσύνολα χαρακτηριστικών.

3.3.2 Επιλογή χαρακτηριστικών – Feature selection

Για την μείωση της εμφάνισης της πολυσυγγραμμικότητας (multicollinearity), στα πλαίσια της εργασίας χρησιμοποιώντας τον συντελεστή Pearson προσδιορίστηκαν τα ζεύγη χαρακτηριστικών που εμφανίζουν ιδιαίτερα υψηλή μεταξύ τους συσχέτιση (π.χ. πόντοι ανά παιχνίδι PPG και προσπάθειες σου FGA).

Από κάθε τέτοιο ζεύγος, διατηρήθηκε μόνο μία μεταβλητή (αυτή με καλύτερη ερμηνευσιμότητα - interpretability). Έτσι διατηρείται η σταθερότητα του μοντέλου, απλουστεύεται η ερμήνευση του, εξαλείφονται περιπτοί προγνωστικοί δείκτες και μειώνεται το υπολογιστικό κόστος εκπαίδευσης.

3.4 Διαχωρισμός δεδομένων σε δεδομένα εκπαίδευσης και δοκιμής

Στα πλαίσια της εργασίας τα δεδομένα χωρίζονται σε δεδομένα εκπαίδευσης (train) και σε δεδομένα δοκιμής (test). Πιο συγκεκριμένα τα δεδομένα που αφορούν την αγωνιστική σεζόν πριν το έτος 2016 είναι τα δεδομένα εκπαίδευσης και τα νεότερα δεδομένα είναι τα δεδομένα δοκιμής. Ο διαχωρισμός των δεδομένων σε δεδομένα εκπαίδευσης (train) και δεδομένα δοκιμής (test) με βάση τη χρονική τους σχέση είναι συνηθισμένη μέθοδος σε προβλήματα πρόβλεψης όπου η χρονική διάσταση έχει κάποια σημασία. Αυτή η προσέγγιση επιλέχθηκε για να προσομοιωθεί μια ρεαλιστική περίπτωση πρόβλεψης, όπου το μοντέλο εκπαιδεύεται σε ιστορικά δεδομένα και ελέγχεται σε νεότερα. Ο χρονικός διαχωρισμός αποφεύγει την πιθανότητα διαρροίς πληροφοριών (data leakage). Αυτή η διαρροή πληροφοριών θα μπορούσε να συμβεί εάν ένα μοντέλο παλινδρόμησες «μάθαινε» μοτίβα που δεν ήταν διαθέσιμα στην πραγματικότητα στην περίπτωση που το σύνολο δεδομένων εκπαίδευσης περιείχε και μελλοντικά δεδομένα.

3.5 Αξιολόγηση μοντέλων

Τα μοντέλα αξιολογούνται πάνω στα δεδομένα εκπαίδευσης με τις μεθόδους rolling window (κυλιόμενο παράθυρο) ή αλλιώς sliding window και expanding window (επεκτεινόμενο παράθυρο) ή growing window. Αυτές οι μέθοδοι σε αντίθεση με την παραδοσιακή διασταυρούμενη επικύρωση (cross validation) δεν ανακατεύουν τα δεδομένα. Έτσι πετυχαίνεται η διατήρηση της χρονικής εξάρτησης [17].

- **Rolling window (κυλιόμενο παράθυρο)**

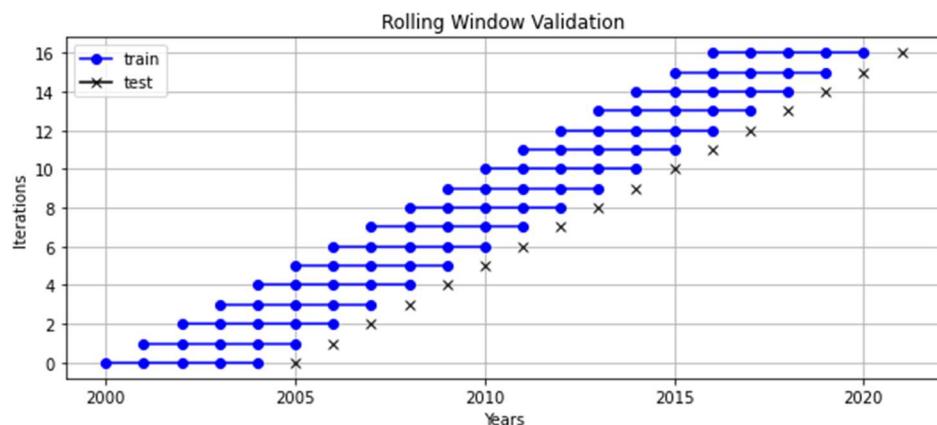
Στην επικύρωση κυλιόμενου παραθύρου, το σύνολο εκπαίδευσης «ολισθαίνει» προς τα εμπρός στο χρόνο, διατηρώντας ένα σταθερό μέγεθος.

- **Expanding window (επεκτεινόμενο παράθυρο)**

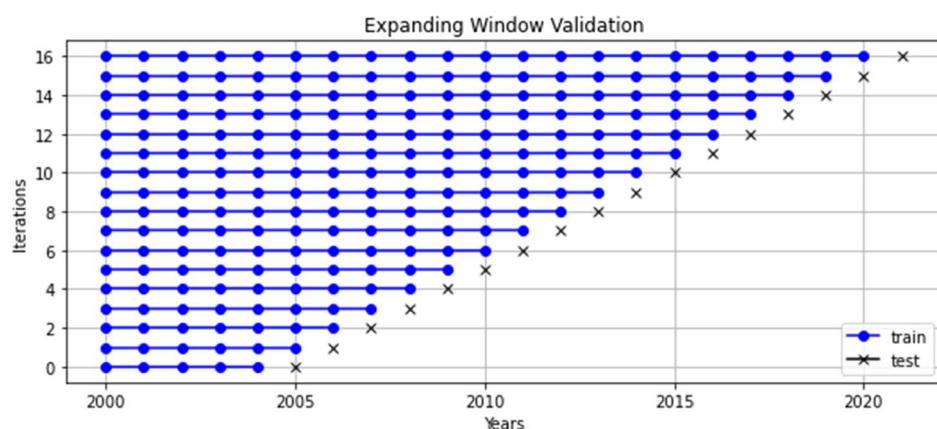
Στην επικύρωση με επεκτεινόμενο παράθυρο, το σύνολο εκπαίδευσης αυξάνεται με την πάροδο του χρόνου. Το αρχικό παράθυρο επεκτείνεται ώστε να περιλαμβάνει περισσότερα δεδομένα σε κάθε επανάληψη.

Η μέθοδος επεκτεινόμενου παραθύρου τείνει να αποτυπώνει την επίδοση του μοντέλου σε μακροπρόθεσμες τάσεις ενώ παράλληλα η μέθοδος κυλιόμενου παραθύρου τείνει να αποτυπώνει την ικανότητα του μοντέλου να ανταπεξέλθει σε πιο ξαφνικές αλλαγές (βραχυπρόθεσμες τάσεις).

Ακολουθούν παραδείγματα κυλιόμενου και επεκτεινόμενου παραθύρου όπου το αρχικό παράθυρο εκπαίδευσης είναι ίσο με 5 έτη και το παράθυρο ελέγχου είναι 1 έτος.



Εικόνα 3-22 Παράδειγμα επικύρωσης με την μέθοδο κυλιόμενου παραθύρου



Εικόνα 3-23 Παράδειγμα επικύρωσης με την μέθοδο επεκτεινόμενου παραθύρου

3.6 Μετρικές

Για την αξιολόγηση των μοντέλων που αναφέρθηκε πρωτύτερα χρησιμοποιούνται οι εξής μετρικές:

- **RMSE (root mean squared error – ρίζα μέσου τετραγωνικού σφάλματος):**

Η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) μετρά τη μέση διαφορά μεταξύ των προβλεπόμενων τιμών ενός στατιστικού μοντέλου και των πραγματικών τιμών. Όσο μικρότερη είναι η τιμή του σφάλματος, τόσο πιο ακριβές είναι το μοντέλο στην πρόβλεψη [18].

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

\hat{y}_i : Είναι η πραγματική τιμή

\hat{y}_i : Είναι η προβλεπόμενη τιμή

- **R-squared (R² ή Coefficient of Determination):**

Το R² μετρά πόσο καλά το μοντέλο εξηγεί τη διακύμανση της μεταβλητής-στόχου. Συνήθως παίρνει τιμές από το μηδέν έως το ένα (0,1). Όσο η τιμή του πλησιάζει τη μονάδα τόσο καλύτερα το μοντέλο «ταιριάζει» με τα δεδομένα [19].

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

\hat{y}_i : Είναι η πραγματική τιμή

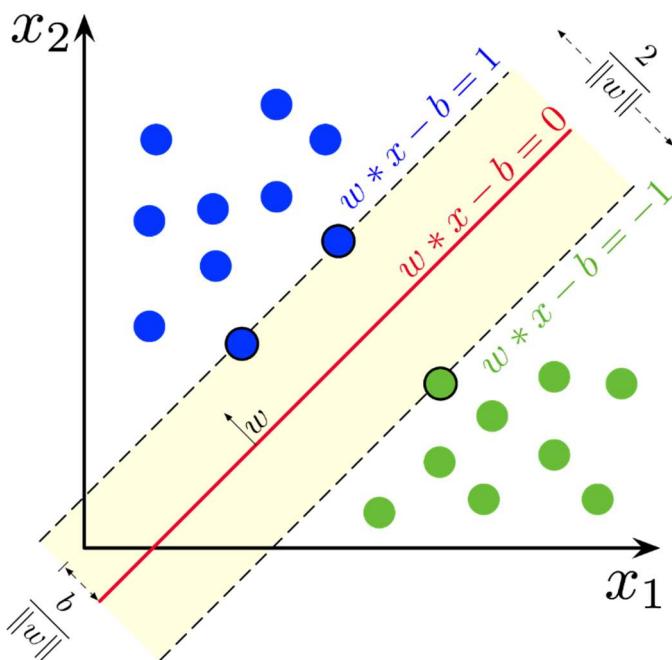
\hat{y}_i : Είναι η προβλεπόμενη τιμή

\bar{y} : Είναι ο μέσος όρος των πραγματικών τιμών

3.7 Μοντέλα

SVR

Ο μηχανισμός ενός SVM (Support Vector Machine) κατασκευάζει ένα υπερ-επίπεδο ή ένα σύνολο υπερ-επιπέδων σε έναν χώρο υψηλών ή άπειρων διαστάσεων, το οποίο μπορεί να χρησιμοποιηθεί για ταξινόμηση ή παλινδρόμηση. Γενικά, ένας καλός διαχωρισμός επιτυγχάνεται από το υπερ-επίπεδο που έχει τη μεγαλύτερη απόσταση από τα πλησιέστερα σημεία δεδομένων εκπαίδευσης οποιασδήποτε κλάσης (το λεγόμενο functional margin περιθώριο), αφού γενικά όσο μεγαλύτερο είναι το περιθώριο τόσο μικρότερο είναι το σφάλμα γενίκευσης του ταξινομητή [20].

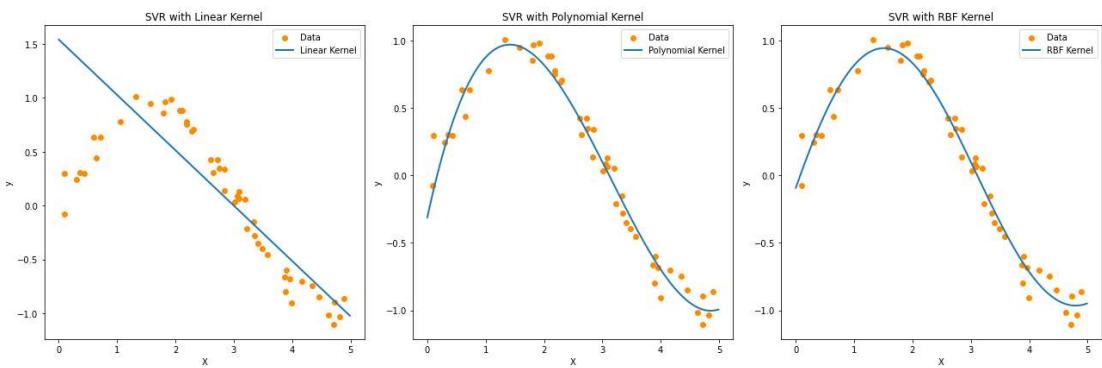


*Eikόνα 3-24 By Larhmam - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=73710028>*

Η μέθοδος ταξινόμησης μέσω του μηχανισμού SVM μπορεί να επεκταθεί για την επίλυση προβλημάτων παλινδρόμησης. Η μέθοδος αυτή ονομάζεται SVR (Support Vector Regression).

Βασικές Υπερπαράμετροι SVR [21]

Kernel: Καθορίζει τον τύπο πυρήνα (Kernel) που θα χρησιμοποιηθεί στον αλγόριθμο. Στην εργασία θα χρησιμοποιηθεί ο τύπος 'rbf'. Οι πυρήνες χρησιμοποιούνται για τον υπολογισμό μη γραμμικά διαχωρίσιμων συναρτήσεων. Ακολουθεί παράδειγμα τυχαίων δεδομένων.



Εικόνα 3-25 Παραδείγματα χρήσης διαφορετικών πυρήνων

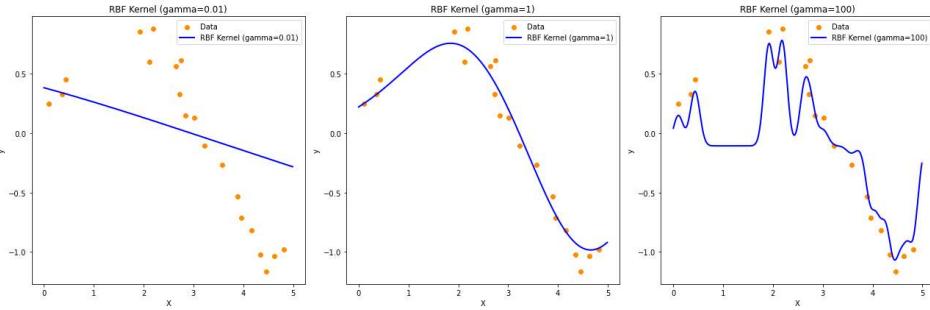
Ο πυρήνας rbf υπολογίζεται από:

$$k(x_i, x_j) = \exp\left(-\frac{d(x_i, x_j)^2}{2l^2}\right)$$

Όπου d είναι η ευκλείδεια απόσταση μεταξύ δύο σημείων και l είναι η παράμετρος μήκους.

Gamma: Είναι μια υπερπαράμετρος που καθορίζει πόσο μακριά φτάνει η επιρροή ενός μεμονωμένου παραδείγματος εκπαίδευσης. Όταν το gamma είναι μεγάλο, η επιρροή κάθε παραδείγματος εκπαίδευσης περιορίζεται σε μια μικρή περιοχή. Αυτό σημαίνει ότι το όριο απόφασης θα είναι πολύ στενό γύρω από μεμονωμένα σημεία δεδομένων. Αυτό μπορεί να οδηγήσει σε υπερπροσαρμογή (overfitting), με αποτέλεσμα την κακή γενίκευση σε νέα δεδομένα. Όταν το gamma είναι μικρό, η επιρροή κάθε παραδείγματος εκπαίδευσης φτάνει πιο μακριά, με αποτέλεσμα πιο ομαλά και λιγότερο πολύπλοκα όρια απόφασης. Αυτό μπορεί να οδηγήσει σε υποπροσαρμογή

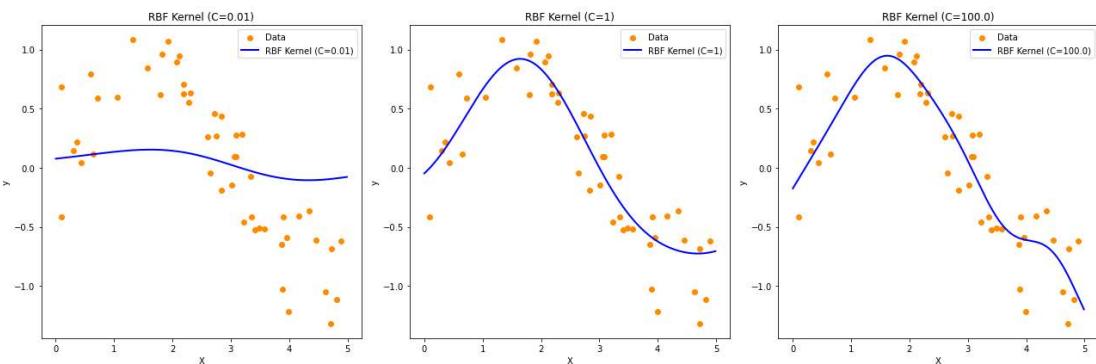
(underfitting), όπου το μοντέλο είναι πολύ απλό και αποτυγχάνει να αναγνωρίσει μοτίβα στα δεδομένα. Ακολουθεί παράδειγμα τυχαίων δεδομένων με διαφορετικές τιμές gamma.



Εικόνα 3-26 Παραδείγματα χρήσης διαφορετικών τιμών gamma

C: Είναι μια παράμετρος κανονικοποίησης που ελέγχει το βάρος του σφάλματος κατά την εκπαίδευση. Το C είναι υπεύθυνο για την ισορροπία μεταξύ της πολυπλοκότητας του μοντέλου και του σφάλματος εκπαίδευσης.

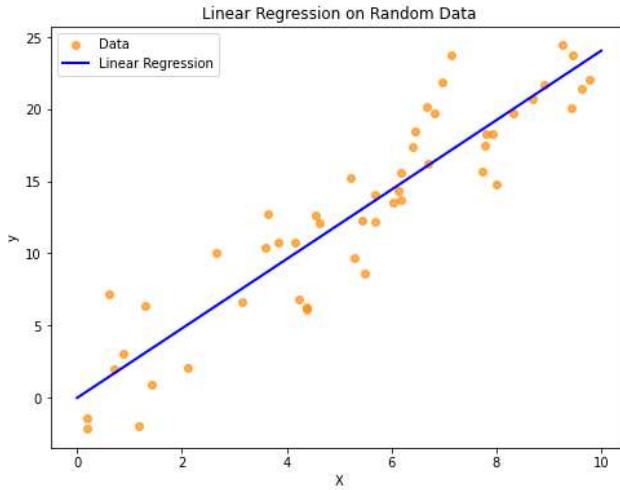
Ένα μικρότερο C έχει ως αποτέλεσμα ένα απλούστερο μοντέλο (μεγαλύτερο περιθώριο), ενώ ένα μεγαλύτερο C έχει ως αποτέλεσμα ένα πιο πολύπλοκο μοντέλο (μικρότερο περιθώριο). Πολύ μικρό C μπορεί να οδηγήσει σε υποπροσαρμογή (underfitting) ενώ πολύ μεγάλο C μπορεί να οδηγήσει σε υπερπροσαρμογή (overfitting).



Εικόνα 3-27 Παραδείγματα χρήσης διαφορετικών τιμών C

Linear regression

Η γραμμική παλινδρόμηση (Linear regression) υπολογίζει τη γραμμική σχέση μεταξύ της εξαρτημένης μεταβλητής και ενός ή περισσότερων ανεξάρτητων χαρακτηριστικών.



Εικόνα 3-28 Παράδειγμα: Γραμμική Παλινδρόμηση σε τυχαία δεδομένα

Ο κώδικας για τη δημιουργία των διαγραμμάτων (Εικόνες 3-25 έως 3-28) προέκυψε από τροποποίηση του κώδικα του παραδείγματος SVR του scikit-learn [22]

Η εξίσωση της γραμμικής παλινδρόμησης είναι

$$y = w_0 + w_1 X_1 + w_2 X_2 + \dots + w_p X_p + \epsilon$$

όπου w_0 είναι το bias, w_1, w_2, \dots, w_p είναι οι συντελεστές/βάρη (coefficients - weights) και ϵ τυχαίο λάθος ή θόρυβος. Στόχος είναι η ελαχιστοποίηση της διαφοράς μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Αυτό πετυχαίνεται με την μέθοδο ελαχίστων τετραγώνων.

$$\min_w \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

(y_i) Είναι η πραγματική τιμή

(\hat{y}_i) Είναι η προβλεπόμενη τιμή

Τροποποιήσεις της γραμμικής παλινδρόμησης έχουν οδηγήσει στη δημιουργία των μοντέλων Ridge, Lasso και Elastic Net.

Lasso

Το μοντέλο Lasso (Least Absolute Shrinkage and Selection Operator) προσθέτει την ποινή κανονικοποίησης L1 που μειώνει το μέγεθος της τιμής

των συντελεστών W . Σε ορισμένες περιπτώσεις η τιμή ενός συντελεστή φθάνει το μηδέν και ουσιαστικά το αντίστοιχο δείγμα θεωρείται μη σημαντικό και απορρίπτεται. Αυτό είναι μια μορφή επιλογής χαρακτηριστικών (feature selection).

$$\min_w \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |w_j|$$

- (y_i) Είναι η πραγματική τιμή
- (\hat{y}_i) Είναι η προβλεπόμενη τιμή
- (w_j) Είναι οι συντελεστές
- (λ) Είναι η παράμετρος κανονικοποίησης

Ridge

Η παλινδρόμηση Ridge είναι μια τροποποίηση προσθέτει την ποινή κανονικοποίησης L2 για τη μείωση του μεγέθους των συντελεστών παλινδρόμησης. Βοηθά στη μείωση της υπερπροσαρμογής (overfitting) με την τιμωρία των μεγάλων βαρών, καθιστώντας το μοντέλο πιο σταθερό σε περιπτώσεις που εμφανίζεται πολυσυγγραμμικότητα (multicollinearity). Πολυσυγγραμμικότητα εμφανίζεται όταν υπάρχουν χαρακτηριστικά που είναι έντονα συσχετισμένα. Η παλινδρόμηση Ridge δεν αναγκάζει τους συντελεστές να μηδενιστούν, αλλά τους συρρικνώνει προς το μηδέν. Αυτό σημαίνει ότι όλα τα χαρακτηριστικά εξακολουθούν να χρησιμοποιούνται στο μοντέλο.

$$\min_w \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p w_j^2$$

- (y_i) Είναι η πραγματική τιμή
- (\hat{y}_i) Είναι η προβλεπόμενη τιμή
- (w_j) Είναι οι συντελεστές
- (λ) Είναι η παράμετρος κανονικοποίησης

Kernel Ridge

Η παλινδρόμηση kernel Ridge συνδυάζει την παλινδρόμηση Ridge με την χρήση πυρήνα (kernel) για χρήση σε δεδομένα που δεν είναι γραμμικά διαχωρίσιμα. Ο πυρήνας μπορεί να είναι πολυώνυμο κάποιου βαθμού ή τύπου RBF όπως αναλύθηκε προηγουμένως στο SVR.

Elastic Net

Το Elastic Net είναι ένα υβρίδιο των μοντέλων Lasso και Ridge, το οποίο συνδυάζει τις ποινές L1 και L2.

$$\min_w \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^p |w_j| + \lambda_2 \sum_{j=1}^p w_j^2$$

(y_i) Είναι η πραγματική τιμή

(\hat{y}_i) Είναι η προβλεπόμενη τιμή

(w_j) Είναι οι συντελεστές

(**L1**) παράμετρος κανονικοποίησης L1

(**L2**) παράμετρος κανονικοποίησης L2

Παλινδρόμηση με Δέντρα αποφάσεων (tree regressors)

Τα μοντέλα παλινδρόμησης με βάση τα Δέντρα βασίζονται σε μια δομή απόφασης, η οποία χωρίζει τα δεδομένα σε υποσύνολα χρησιμοποιώντας μια σειρά από κανόνες αποφάσεων (if-then-else) με βάση τις τιμές των χαρακτηριστικών. Το δέντρο αποτελείται από κόμβους, κλάδους και φύλλα. Στους κόμβους γίνονται οι διαχωρισμοί, οι κλάδοι είναι τα αποτελέσματα των διαχωρισμών και τα φύλλα αποτελούν τους τελικούς κόμβους όπου παρέχονται οι προβλέψεις του μοντέλου. Κάθε κόμβος στο εσωτερικό του δέντρου αντιπροσωπεύει μια απόφαση που πάρθηκε βάσει ενός χαρακτηριστικού και κάθε φύλλο αντιπροσωπεύει μια προβλεπόμενη τιμή. Το δέντρο κατασκευάζεται/εκπαιδεύεται αναδρομικά διασπώντας το σύνολο των δεδομένων σε υποσύνολα με στόχο την ελαχιστοποίηση του σφάλματος μεταξύ της προβλεπόμενης τιμής και της τιμής-στόχου.

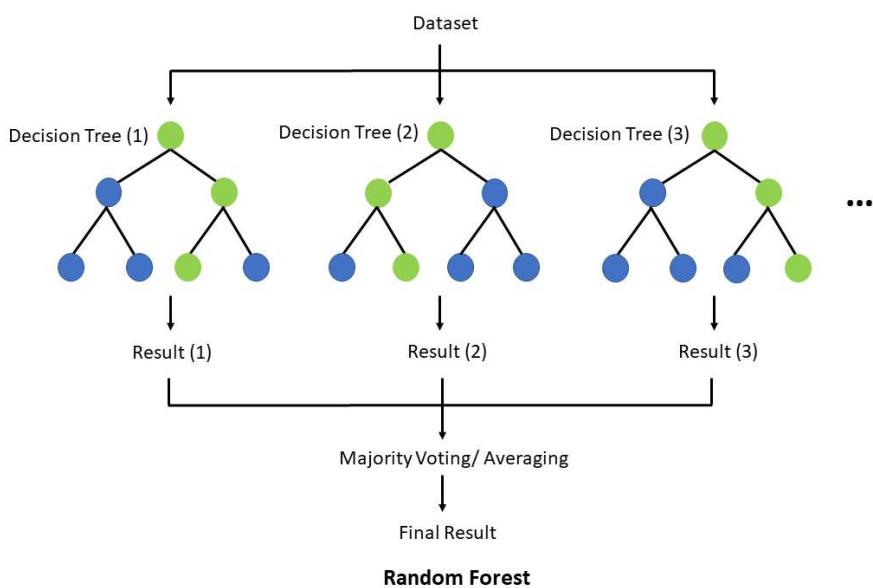
Επεκτάσεις που προέκυψαν με βάση τα Δέντρα αποφάσεων

❖ Random Forest

Ένα μειονέκτημα του Δέντρου απόφασης είναι ότι όσο αυξάνεται το βάθος του μπορεί να οδηγηθεί εύκολα σε υπερπροσαρμογή (overfitting). Το μοντέλο Random Forest ακολουθεί την μέθοδο του συνόλου (ensemble) εκπαιδεύοντας πολλαπλά δέντρα αποφάσεων (ή αλλιώς ένα «δάσος») συνδυάζοντας τις προβλέψεις τους με στόχο τη δημιουργία ενός μοντέλου που είναι πιο ακριβές και πιο ανθεκτικό στο φαινόμενο της υπερπροσαρμογής (overfitting). Το Random Forest μοντέλο είναι επίσης χρήσιμο στο χειρισμό μεγάλων συνόλων δεδομένων με υψηλή διαστατικότητα και ετερογενείς τύπους χαρακτηριστικών (πχ κατηγορηματικά και αριθμητικά). Τα μειονεκτήματα του μοντέλου Random Forest, δηλαδή τα μειονεκτήματα του «Δάσους» σε σχέση με το «Δέντρο», που προέρχονται από την πολυπλοκότητα του είναι η υψηλότερη υπολογιστική δαπάνη και η δυσκολία ερμηνείας του τρόπου λήψης των αποφάσεων [23].

Βασικές Υπερπαράμετροι Random Forest [24]

- **n_estimators:** Η παράμετρος που ορίζει το πλήθος των Δέντρων Αποφάσεων που θα δημιουργηθούν.
- **max depth:** Η παράμετρος που ορίζει το μέγιστο βάθος κάθε δέντρου απόφασης. Μεγαλύτερο βάθος συνεπάγεται και μεγαλύτερο όγκο πληροφορίας για την εκπαίδευση του μοντέλου.
- **min samples split:** Η παράμετρος που ορίζει το ελάχιστο πλήθος δειγμάτων που απαιτούνται για τον διαχωρισμό ενός εσωτερικού κόμβου στο Δέντρο Αποφάσεων.
- **min samples leaf:** Η παράμετρος που ορίζει το ελάχιστο πλήθος δειγμάτων που οφείλουν να υπάρχουν σε έναν κόμβο-φύλλο.
- **Bootstrap:** Αν η παράμετρος bootstrap τεθεί «Ψευδής», τότε κάθε Δέντρο Απόφασης θα χρησιμοποιήσει το σύνολο των δεδομένων εκπαίδευσης για την δημιουργία του. Αν η τιμή της είναι «Αληθής» τότε κάθε Δέντρο εκπαιδεύεται σε ένα τυχαίο υποσύνολο των δεδομένων.



Eikόνα 3-29 By TseKiChun - Own work, CC BY-SA 4.0,
<https://commons.wikimedia.org/w/index.php?curid=112433454>

❖ Gradient Boosting Regressors

Τα μοντέλα που χρησιμοποιούν την τεχνική του Gradient Boosting είναι μοντέλα τύπου «συνόλου» (ensemble) τα οποία κατασκευάζουν δέντρα διαδοχικά. Κάθε νέο δέντρο διορθώνει τα σφάλματα των προηγούμενων. Το Gradient Boosting είναι ένας αλγόριθμος ενίσχυσης που συνδυάζει διάφορα αδύναμα μοντέλα σε ισχυρότερα. Κάθε νέο μοντέλο εκπαιδεύεται με στόχο να ελαχιστοποιήσει τη συνάρτηση απώλειας - loss function (πχ. το μέσο τετραγωνικό σφάλμα του προηγούμενου μοντέλου) χρησιμοποιώντας την μέθοδο gradient descent. Σε κάθε επανάληψη, ο αλγόριθμος υπολογίζει την κλίση της συνάρτησης απώλειας σε σχέση με τις προβλέψεις του τρέχοντος συνόλου και στη συνέχεια εκπαιδεύει ένα νέο αδύναμο μοντέλο για να ελαχιστοποιήσει αυτή την κλίση. Οι προβλέψεις του νέου μοντέλου προστίθενται στη συνέχεια στο σύνολο και η διαδικασία επαναλαμβάνεται μέχρι να ικανοποιηθεί το κριτήριο διακοπής (πχ ο αριθμός επαναλήψεων ή η μικρή τιμή σφάλματος).

Στην εργασία χρησιμοποιήθηκαν οι ακόλουθες υλοποιήσεις Gradient Boosting μοντέλων:

➤ XGBoost (eXtreme Gradient Boosting)

Ο XGBoost είναι μία από τις πρώτες και πιο ευρέως χρησιμοποιούμενες Gradient Boosting βιβλιοθήκες. Χρησιμοποιεί μια στρατηγική ανάπτυξης δέντρων κατά επίπεδο (κατά βάθος), δηλαδή χτίζει τα δέντρα επίπεδο προς επίπεδο [25].

➤ LightGBM (Light Gradient Boosting Machine)

Ο LightGBM είναι νεότερη υλοποίηση συγκριτικά με τον XGBoost η οποία περιλαμβάνει βελτιστοποιήσεις που οδηγούν σε υψηλότερη ταχύτητα και συχνά σε υψηλότερη ακρίβεια. Μια τέτοια βελτιστοποίηση είναι η ανάπτυξη των δέντρων κατά φύλλο σε αντίθεση με τον XGBoost που χρησιμοποιεί ανάπτυξη κατά βάθος [26], [27].

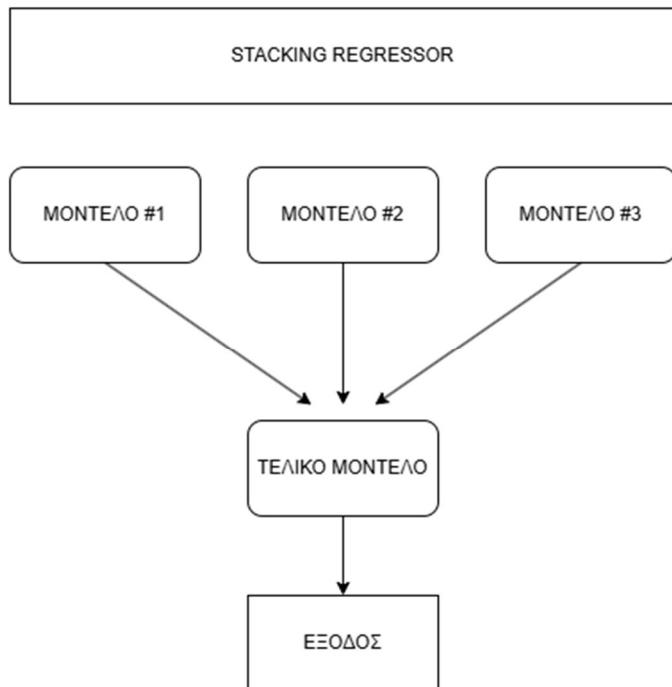
➤ CatBoost (Categorical Boosting)

Ο CatBoost είναι συγκριτικά η πιο πρόσφατη υλοποίηση. Είναι ειδικά σχεδιασμένη για να αντιμετωπίζει κατηγορικά δεδομένα χωρίς την ανάγκη προεπεξεργασίας. Σε αντίθεση με τις δυο προηγούμενες υλοποιήσεις χρησιμοποιεί συμμετρικά δέντρα επιταχύνοντας τη διαδικασία εκπαίδευσης και μειώνοντας την κατανάλωση μνήμης. Επιπλέον για την καταπολέμηση του φαινομένου data leakage που παρατηρείται σε Boosting αλγορίθμους και οδηγεί σε υπερπροσαρμογή (overfitting) χρησιμοποιεί την τεχνική του Ordered Boosting η οποία είναι εμπνευσμένη από την έννοια της «online» μάθησης, όπου το μοντέλο ενημερώνεται σταδιακά καθώς φθάνουν νέα δεδομένα.

Το CatBoost ανακατεύει τυχαία (permutes) το σύνολο δεδομένων εκπαίδευσης πριν από την εκπαίδευση. Αυτή η αντιμετάθεση εξασφαλίζει ότι το μοντέλο δεν βασίζεται στη σειρά των δεδομένων, η οποία μπορεί να εισάγει μεροληψία (bias). Για κάθε παράδειγμα εκπαίδευσης ο CatBoost το αντιμετωπίζει σαν να έφτανε με διαδοχική σειρά. Το μοντέλο εκπαιδεύεται χρησιμοποιώντας μόνο τα δεδομένα που «προηγήθηκαν» του τρέχοντος παραδείγματος στη μετάθεση. Αυτό προσομοιώνει ένα πραγματικό σενάριο όπου το μοντέλο δεν έχει πρόσβαση σε μελλοντικά δεδομένα κατά τη διάρκεια της εκπαίδευσης [28], [29].

Stacking regressors

Ο όρος Stacking generalization (γενίκευση «στοίβας») περιγράφει μια μέθοδο συνδυασμού εκτιμητών-μοντέλων με σκοπό τη μείωση του bias (μεροληψίας) τους [30]. Πιο συγκεκριμένα, οι προβλέψεις κάθε μεμονωμένου μοντέλου στοιβάζονται μαζί και χρησιμοποιούνται ως είσοδος σε έναν τελικό εκτιμητή-μοντέλο για τον υπολογισμό της πρόβλεψης. Αυτή είναι η μορφή ενός Stacking Regressor (παλινδρόμηση «στοίβας»).



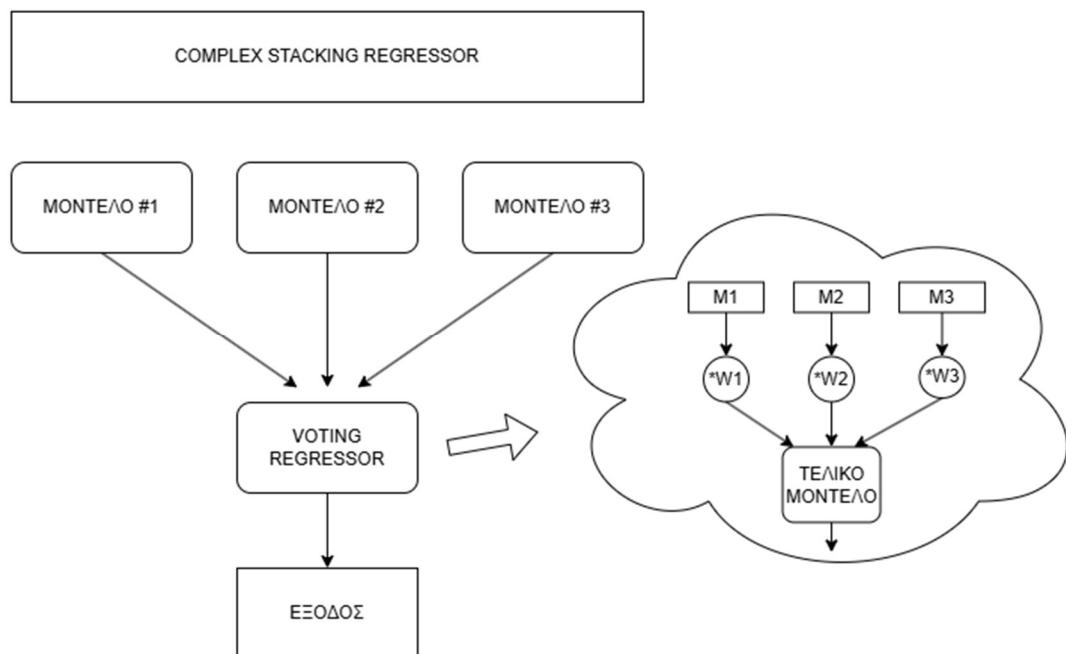
Εικόνα 3-30 Παλινδρόμηση "Στοίβας"

Voting regressor

Μια απλούστερη μέθοδος συνδυασμού εκτιμητών-μοντέλων είναι ο Voting regressor (παλινδρόμηση «ψήφου»). Σε αυτή την μέθοδο συνδυάζονται οι προβλέψεις πολλαπλών βασικών μοντέλων (π.χ. γραμμική παλινδρόμηση, δέντρα απόφασης, SVM) με μέσο όρο ή με σταθμισμένο μέσο όρο των προβλέψεών τους.

Σύνθεση Stacking και Voting regressor

Συνδυάζοντας τις μεθόδους της παλινδρόμησης «στοίβας» (stacking regressor) και τις παλινδρόμησης «ψήφου» (voting regressor) μπορεί να προκύψουν πιο σύνθετες μορφές μοντέλων όπως το παρακάτω:



Εικόνα 3-31 Συνδυασμός μεθόδων παλινδρόμησης "Στοίβας" και "Ψήφου" (Combination of stacking and voting regressors)

3.8 Αναζήτηση πλέγματος – Grid search

Η αναζήτηση πλέγματος (Grid search) είναι μια τεχνική που χρησιμοποιείται στη μηχανική μάθηση για την εύρεση των βέλτιστων υπερπαραμέτρων ενός μοντέλου. Οι υπερπαράμετροι (hyperparameters) είναι ρυθμίσεις που ελέγχουν τη συμπεριφορά ενός αλγορίθμου μηχανικής μάθησης (π.χ. ο αριθμός των δέντρων σε ένα μοντέλο Random Forest ή ο ρυθμός μάθησης σε ένα νευρωνικό δίκτυο). Σε αντίθεση με τις παραμέτρους του μοντέλου, οι οποίες μαθαίνονται κατά τη διάρκεια της εκπαίδευσης, οι υπερπαράμετροι ορίζονται πριν από την έναρξη της εκπαίδευσης.

Βήματα αναζήτησης πλέγματος:

- Καθορισμός ενός συνόλου πιθανών τιμών (ενός πλέγματος - grid) για κάθε υπερπαράμετρο.
- Αξιολόγηση κάθε πιθανού συνδυασμού υπερπαραμέτρων στο πλέγμα.
- Για κάθε συνδυασμό, εκπαίδευση του μοντέλου και αξιολόγηση της απόδοσης του, χρησιμοποιώντας μια μετρική όπως η ρίζα του μέσου τετραγωνικού σφάλματος (RMSE) ή το R^2 .
- Επιλογή του καλύτερου συνδυασμού.

Η αναζήτηση πλέγματος πρόκειται για μια εξονυχιστική αναζήτηση η οποία μπορεί να καταλήξει εξαιρετικά δαπανηρή υπολογιστικά. Για αυτό το λόγο υπάρχουν και υλοποιήσεις οι οποίες εισάγουν τυχαιότητα στην επιλογή των τιμών των υπερπαραμέτρων με σκοπό την μείωση της υπολογιστικής ακρίβειας.

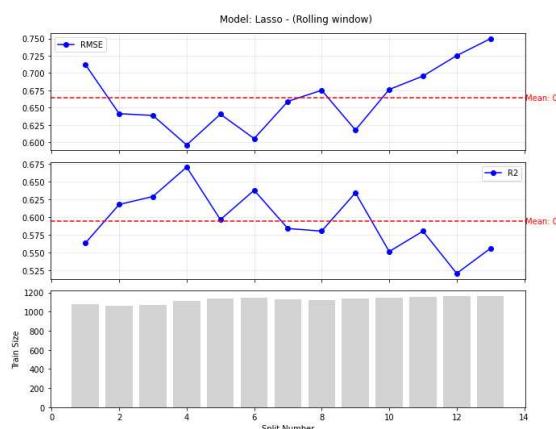
4. Πειραματικά αποτελέσματα

Η επιλογή των μοντέλων που αποδίδουν καλύτερα στην προκειμένη πρόκληση πραγματοποιείται μέσω επικύρωσης (validation) με τις μεθόδους κυλιόμενου παραθύρου (rolling window) και επεκτεινόμενου παραθύρου (expanding window). Στη συνέχεια αφού επιλεγούν τα κατάλληλα μοντέλα ακολουθεί η τεχνική αναζήτησης πλέγματος (grid search) που αποσκοπεί στη βελτίωση της απόδοσης τους, μέσω της εύρεσης βέλτιστων υπερπαραμέτρων. Τελικά θα ελεγχθεί η χρήση μεθόδων συνόλου (ensembles) όπως η παλινδρόμηση «στοίβας» (stacking regressor) και η παλινδρόμηση «ψήφου» (voting regressor).

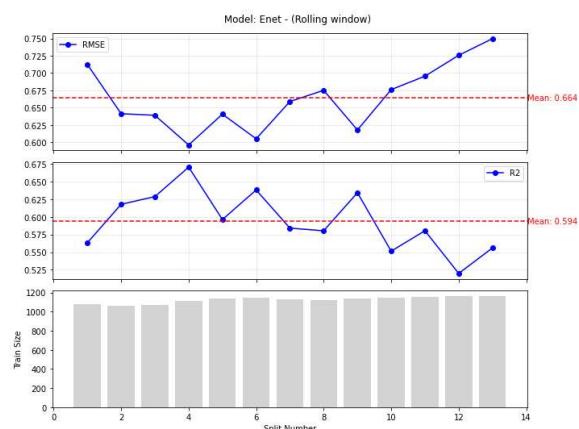
4.1 Κυλιόμενο Παράθυρο - Rolling Window

Ακολουθούν διαγράμματα των μετρικών RMSE (root mean squared error – ρίζα μέσου τετραγωνικού σφάλματος) και R2 (Coefficient of Determination) για το κάθε μοντέλο τα οποία αποτυπώνουν την απόδοση του μοντέλου κατά την διαδικασία επικύρωσης με την μέθοδο Κυλιόμενου παραθύρου.

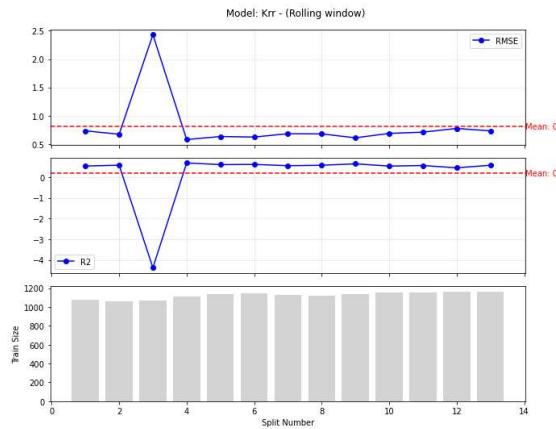
Lasso



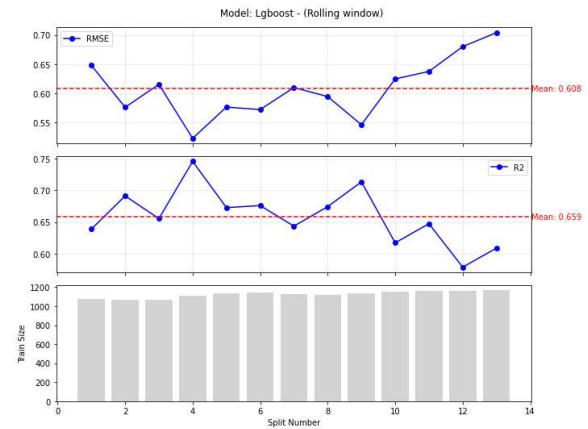
Enet



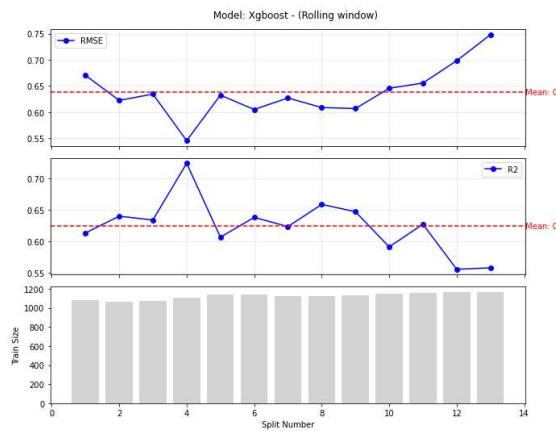
Kernel Ridge



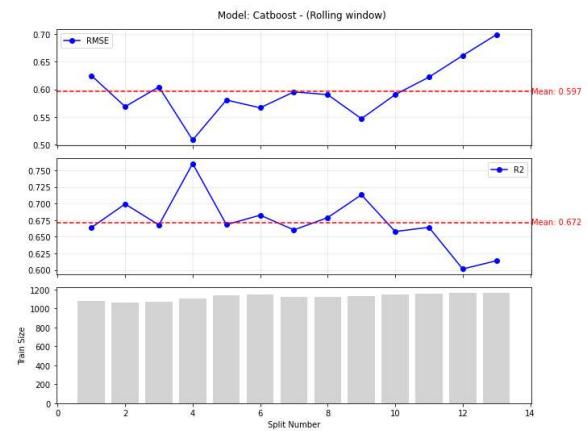
LightGBM



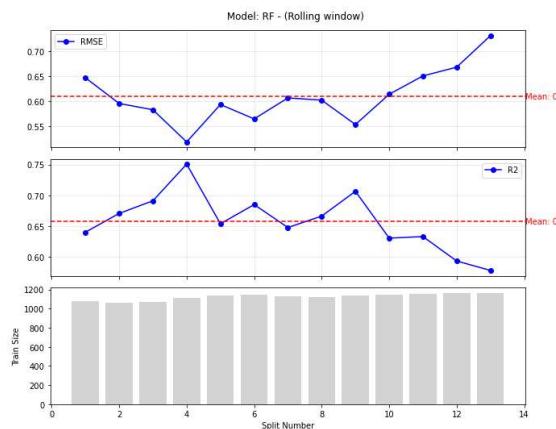
Xgboost



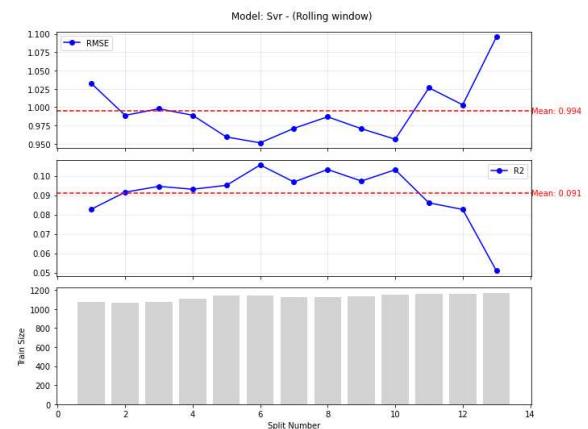
Catboost



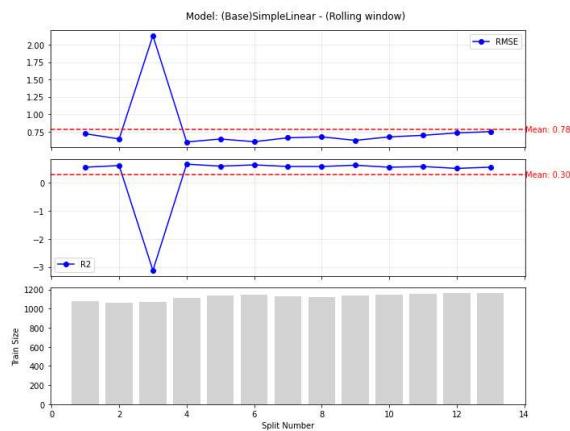
Random Forest



Svr



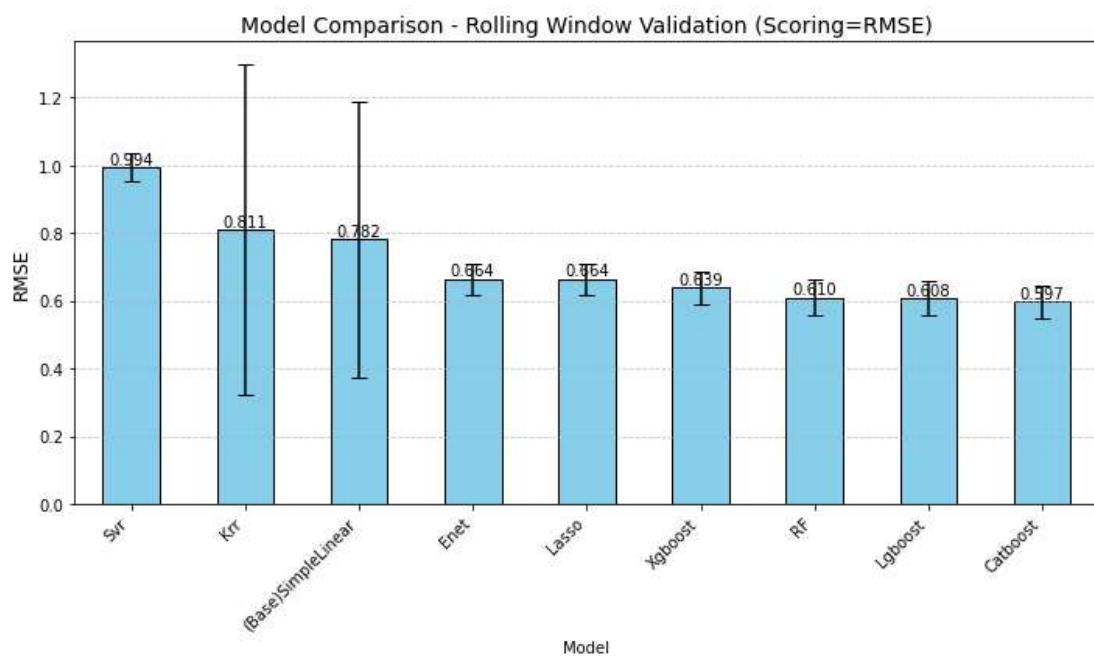
Απλή Γραμμική παλινδρόμηση



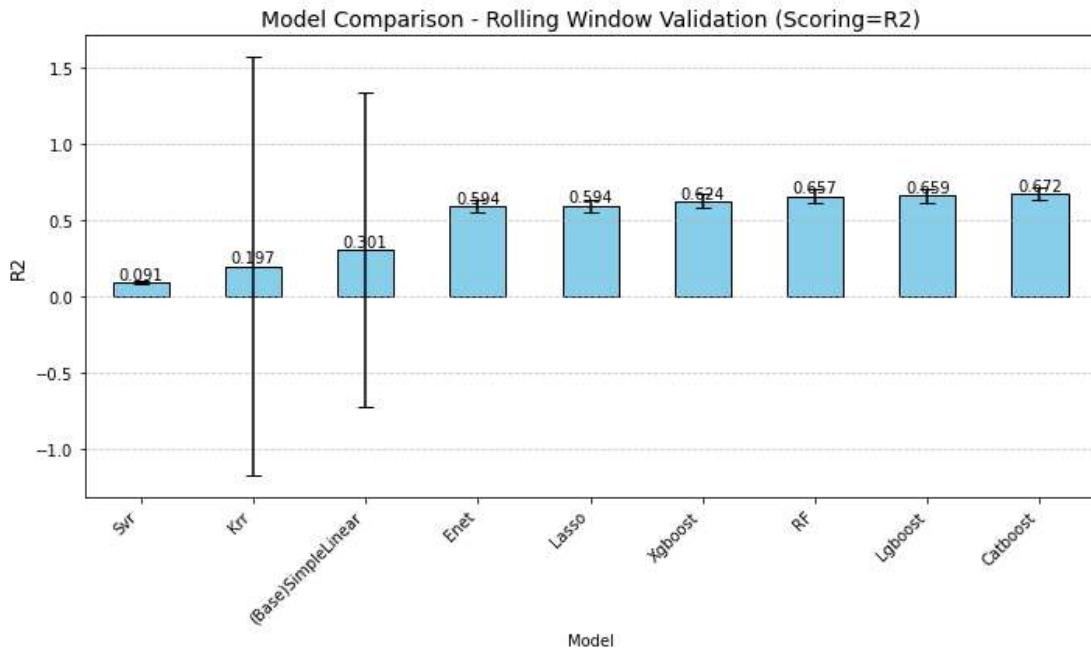
Εικόνα 4-1 έως 4-9 Επικύρωση με την μέθοδο Κυλιόμενου Παραθύρου: Σφάλμα RMSE, σκορ R2 και μέγεθος του σετ εκπαίδευσης

Συγκεντρωτικά αποτελέσματα

Τα διαγράμματα αποτυπώνουν τις μέσες τιμές (mean values) των μετρικών και την απόκλιση τους (deviation).



Εικόνα 4-10 Επικύρωση με την μέθοδο Κυλιόμενου Παραθύρου: Συγκεντρωτικά αποτελέσματα RMSE



Εικόνα 4-11 Επικύρωση με την μέθοδο Κυλιόμενου Παραθύρου: Συγκεντρωτικά αποτελέσματα R2

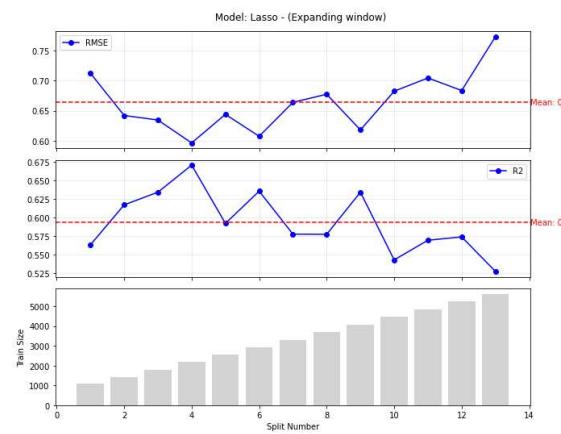
Παρατηρήσεις

Φαίνεται πως τα μοντέλα που βασίζονται στην δομή των Δέντρων απόφασης αποδίδουν καλυτέρα. Πιο συγκεκριμένα παρατηρείται πως το μοντέλο παλινδρόμησης Catboost είναι αυτό που αποδίδει καλύτερα (έχει το μεγαλύτερο σκορ R2 και το μικρότερο σφάλμα RMSE).

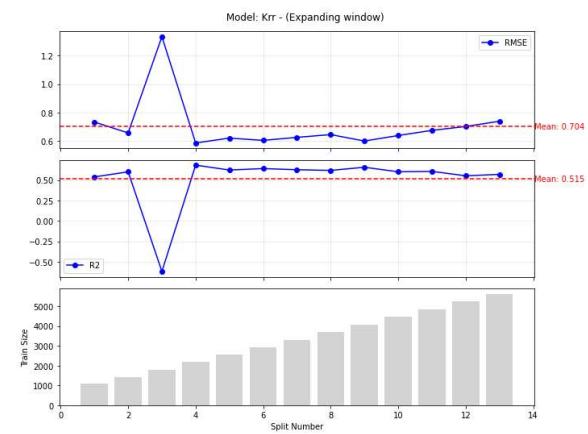
4.2 Επεκτεινόμενο παράθυρο - Expanding Window

Ακολουθούν διαγράμματα των μετρικών RMSE (root mean squared error – ρίζα μέσου τετραγωνικού σφάλματος) και R2 (Coefficient of Determination) για το κάθε μοντέλο τα οποία αποτυπώνουν την απόδοση του μοντέλου κατά την διαδικασία επικύρωσης με την μέθοδο Επεκτεινόμενου παραθύρου.

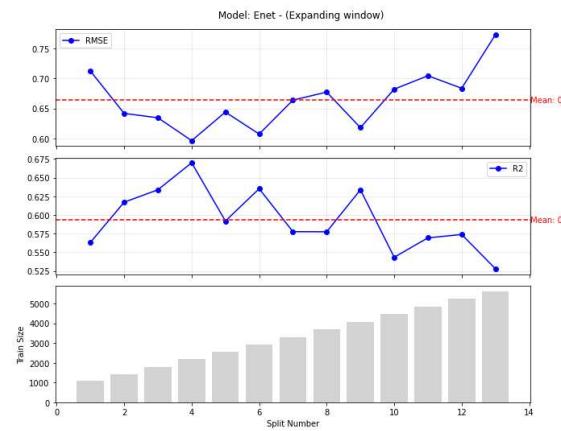
Lasso



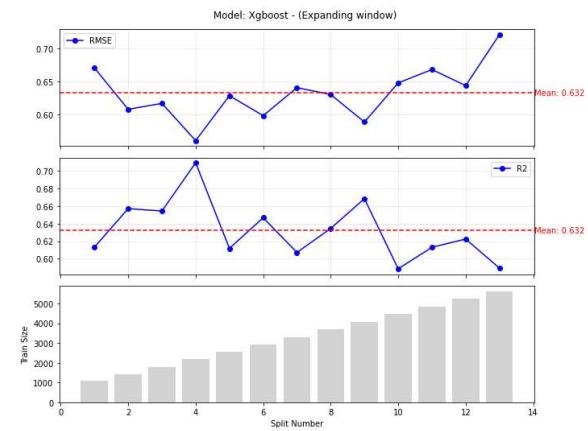
Kernel Ridge



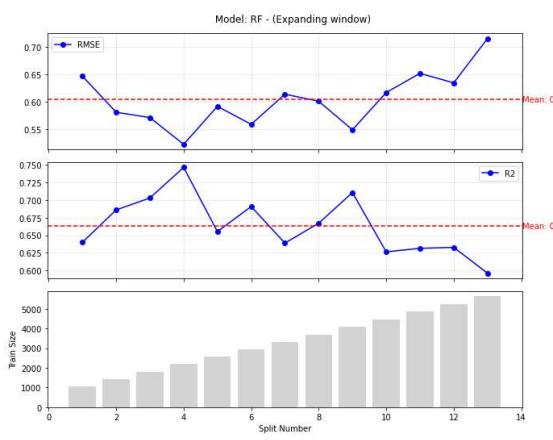
Enet



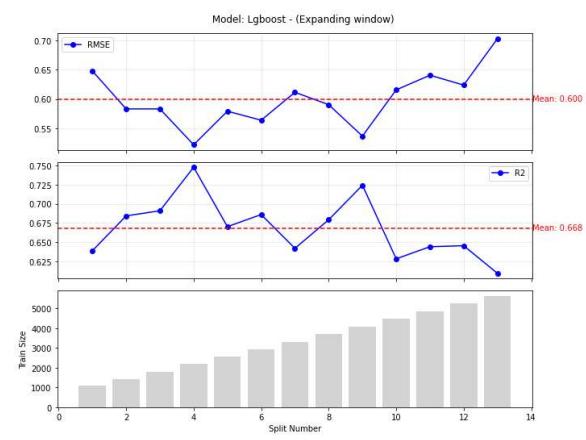
Xgboost



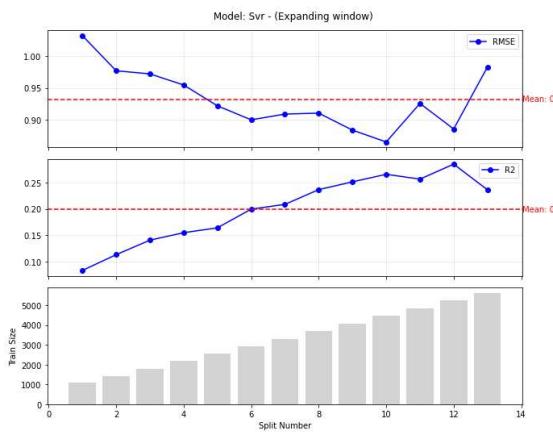
Random Forest



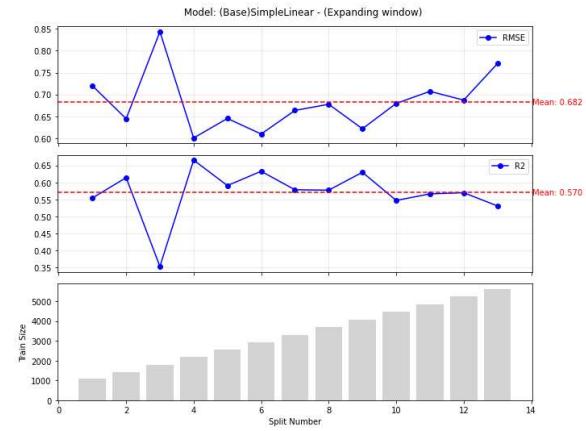
LightGBM



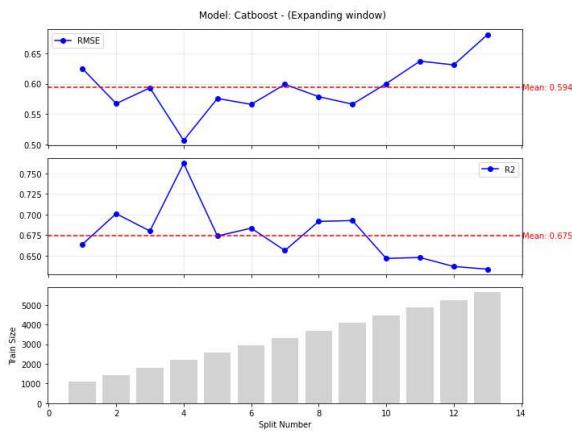
SVR



Απλή Γραμμική παλινδρόμηση



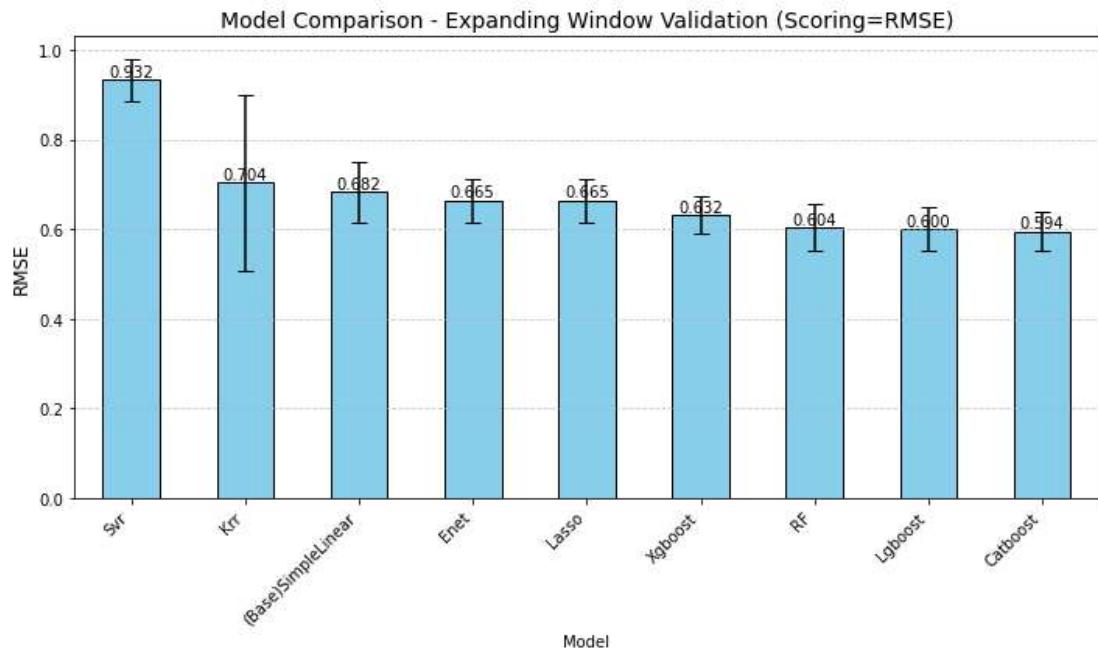
Catboost



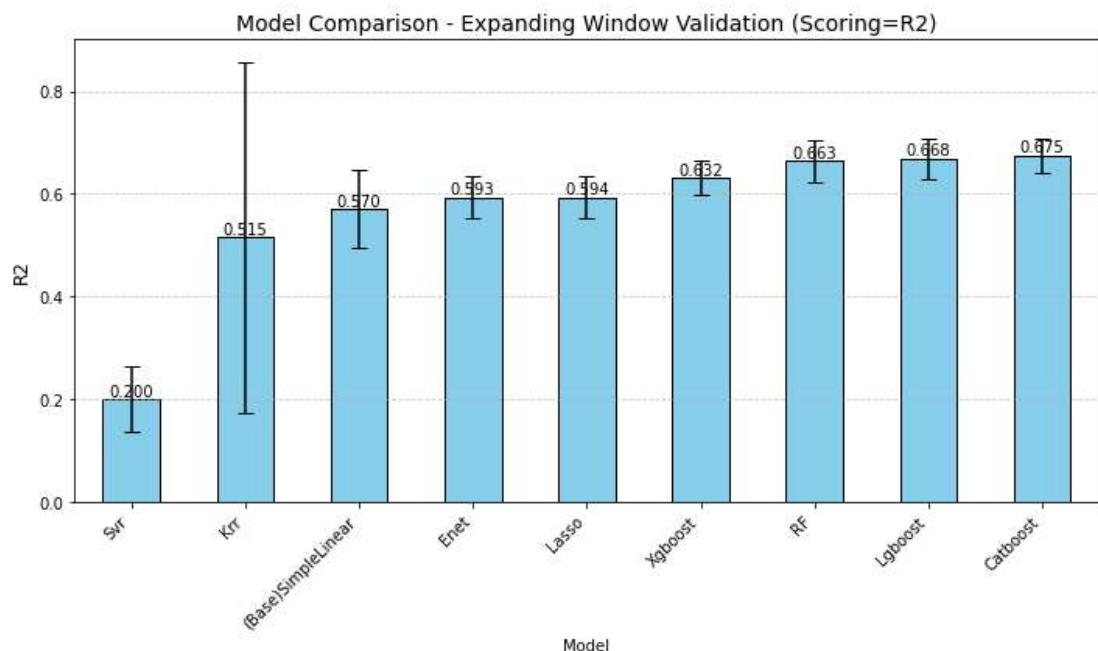
Εικόνα 4-12 έως 4-20 Επικύρωση με την μέθοδο Επεκτεινόμενου Παραθύρου: Σφάλμα RMSE, σκορ R2 και μέγεθος του σετ εκπαίδευσης

Συγκεντρωτικά αποτελέσματα

Τα διαγράμματα αποτυπώνουν τις μέσες τιμές (mean values) των μετρικών και την απόκλιση τους (deviation).



Εικόνα 4-21 Επικύρωση με την μέθοδο Επεκτεινόμενου Παραθύρου: Συγκεντρωτικά αποτελέσματα RMSE



Εικόνα 4-22 Επικύρωση με την μέθοδο Επεκτεινόμενου Παραθύρου: Συγκεντρωτικά αποτελέσματα R2

Παρατηρήσεις

Όμοια με την μέθοδο του κυλιόμενου παραθύρου παρατηρείται ότι τα μοντέλα που βασίζονται στην δομή των Δέντρων αποδίδουν καλυτέρα. Πάλι το μοντέλο παλινδρόμησης Catboost είναι αυτό που αποδίδει καλύτερα. Επίσης τα μοντέλα που απέδωσαν καλά προηγουμένως φαίνεται να αποδίδουν το ίδιο ικανοποιητικά και στην μέθοδο επεκτεινόμενου παραθύρου.

4.3 Τεχνική αναζήτησης πλέγματος - Grid search

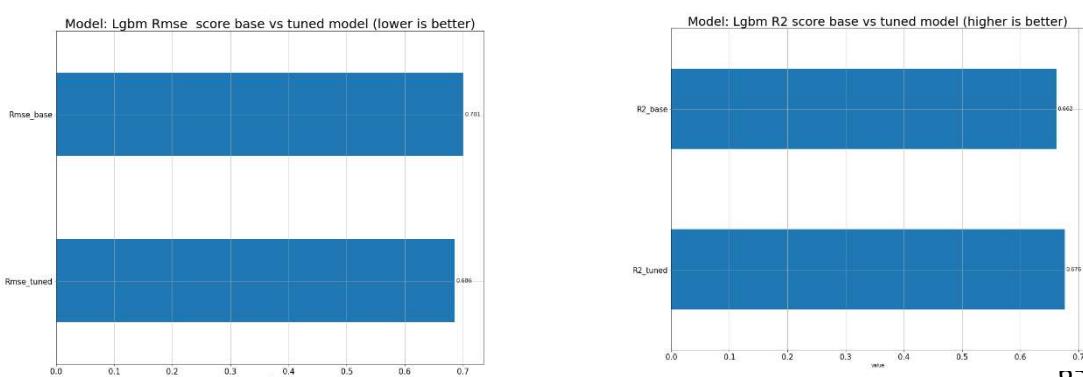
Τα μοντέλα παλινδρόμησης τα οποία επιλέχθηκαν να βελτιστοποιηθούν με την μέθοδο αναζήτησης πλέγματος (grid search) είναι τα :

- CatBoost
- LightGBM
- XgBoost

Για την επιλογή των βέλτιστων δυνατών υπερ-παραμέτρων τα μοντέλα εκπαιδεύονται (Training) και επικυρώνονται (Validation) στο σετ εκπαίδευσης (Test set) το οποίο διασπάται σε διαχωριζόμενα χρονικά διαστήματα (time splits). Τα τελικά μοντέλα με τις βέλτιστες παραμέτρους συγκρίνονται με τα αρχικά (βασικά) μοντέλα μέσω ελέγχου στο σετ δοκιμής.

LightGBM

Ακολουθούν τα αποτελέσματα της αναζήτησης πλέγματος και σύγκριση του αρχικού με το τελικό βελτιστοποιημένο μοντέλο.

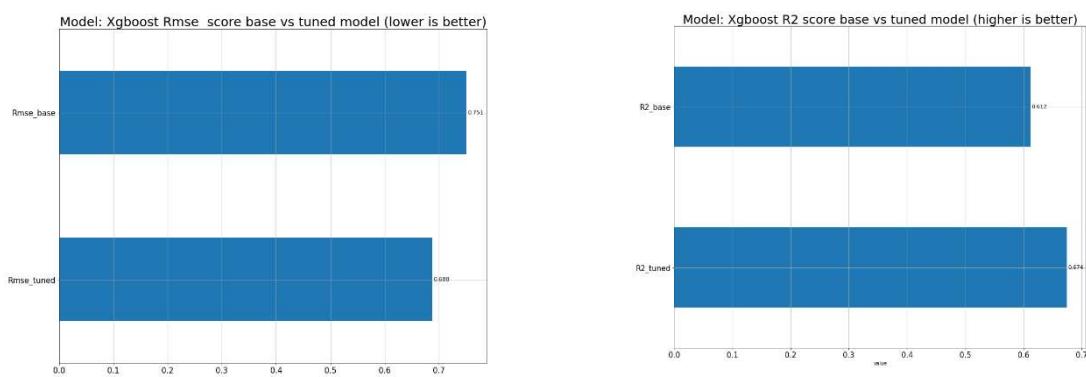


Εικόνα 4-23 και 4-24 Σύγκριση μεταξύ βασικού και βελτιστοποιημένου μοντέλου LightGBM

Παρατηρείται ότι το βελτιστοποιημένο μοντέλο LightGBM regressor αποδίδει καλύτερα από το βασικό. Παρουσιάζει μικρότερο σφάλμα (RMSE) και μεγαλύτερο R2.

XgBoost

Ακολουθούν τα αποτελέσματα της αναζήτησης πλέγματος και σύγκριση του αρχικού με το τελικό βελτιστοποιημένο μοντέλο.

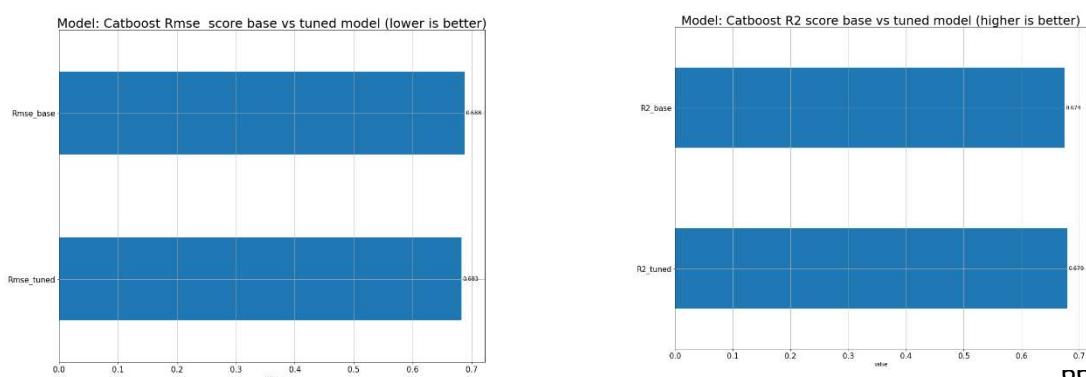


Εικόνα 4-25 και 4-26 Σύγκριση μεταξύ βασικού και βελτιστοποιημένου μοντέλου XgBoost

Παρατηρείται ότι όμοια με το LGBM μοντέλο το βελτιστοποιημένο μοντέλο XgBoost regressor αποδίδει καλύτερα από το βασικό. Παρουσιάζει μικρότερο σφάλμα (RMSE) και μεγαλύτερο R2. Η βελτιστοποίηση σε αυτή την περίπτωση είναι πιο έντονα εμφανής.

CatBoost

Ακολουθούν τα αποτελέσματα της αναζήτησης πλέγματος και σύγκριση του αρχικού με το τελικό βελτιστοποιημένο μοντέλο.



Εικόνα 4-27 και 4-28 Σύγκριση μεταξύ βασικού και βελτιστοποιημένου μοντέλου CatBoost

Παρατηρείται ότι όμοια με τις προηγούμενες περιπτώσεις το βελτιστοποιημένο μοντέλο CatBoost regressor αποδίδει καλύτερα από το βασικό. Παρουσιάζει μικρότερο σφάλμα (RMSE) και μεγαλύτερο R2. Η βελτιστοποίηση σε αυτή την περίπτωση είναι λιγότερο έντονη αλλά το μοντέλο CatBoost εξακολουθεί να έχει την καλύτερη επίδοση σε σύγκριση με τα άλλα μοντέλα. Η διαφορά αυτή όμως έχει μειωθεί αισθητά μετά την βελτιστοποίηση μέσω της αναζήτησης πλέγματος.

Ακολουθεί ο συγκεντρωτικός πίνακας των αποτελεσμάτων αναζήτησης πλέγματος (Grid search).

Μοντέλο	Base		Optimized	
	- Μη Βελτιστοποιημένο		- Βελτιστοποιημένο	
	RMSE	R2	RMSE	R2
LightGBM	0.701	0.662	0.686	0.676
XgBoost	0.751	0.612	0.688	0.674
CatBoost	0.688	0.674	0.683	0.679

Εικόνα 4-29 Συγκεντρωτικός πίνακας αποτελεσμάτων αναζήτησης πλέγματος (grid search)

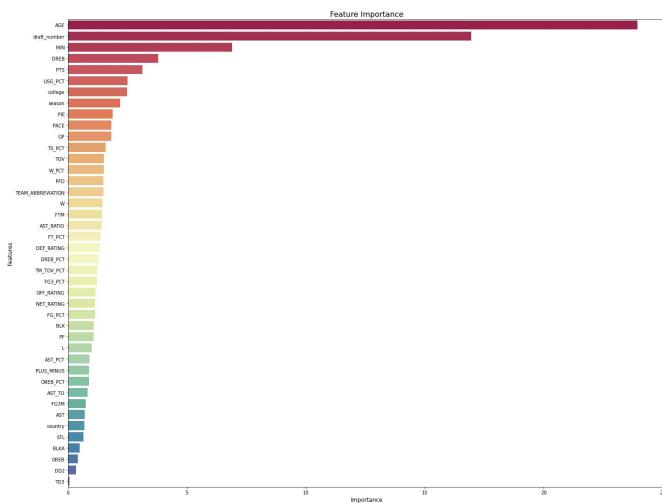
Τελικά το μοντέλο παλινδρόμησης (regression) CatBoost είναι αυτό που εμφανίζει την μεγαλύτερη απόδοση.

4.4 Σημαντικά χαρακτηριστικά - Feature importance

Ενσωματωμένη μέθοδος

Το μοντέλο παλινδρόμησης CatBoost παρέχει ενσωματωμένες μεθόδους για την αξιολόγηση της σημασίας των χαρακτηριστικών (feature importance) σε προκλήσεις παλινδρόμησης. Η κατανόηση αυτών των μετρικών σημασίας μπορεί να συντελέσει στην καλή ερμήνευση του μοντέλου. Για κάθε χαρακτηριστικό, το CatBoost ανακατεύει τις τιμές του και μετρά τον αντίκτυπο στις προβλέψεις. Μετρά πόσο αλλάζει η πρόβλεψη κατά μέσο όρο όταν αλλάζει η τιμή του χαρακτηριστικού.

Θεωρούνται πιο σημαντικά τα χαρακτηριστικά που προκαλούν μεγαλύτερες αλλαγές πρόβλεψης όταν ανακατεύονται.



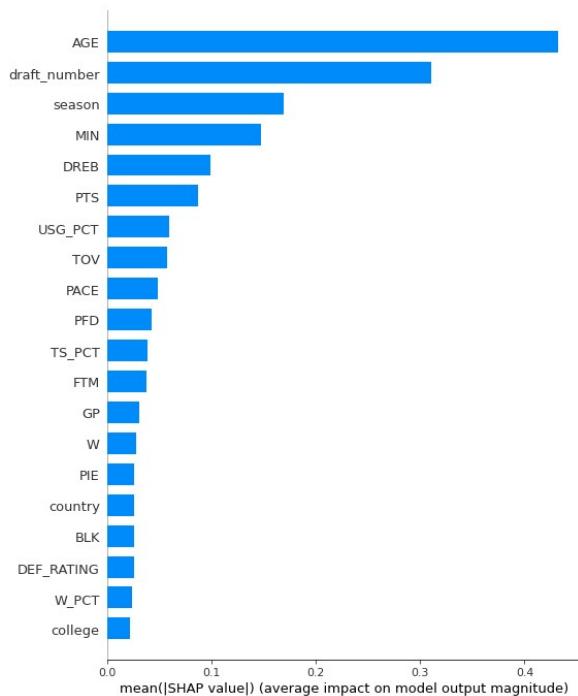
Εικόνα 4-30 Κατάταξη σημαντικών χαρακτηριστικών

Παρατηρήσεις: Παρατηρείται ότι η ηλικία, ο αριθμός επιλογής του αθλητή στην λοταρία του ντραφτ (draft number), ο μέσος όρος των λεπτών συμμετοχής του, ο μέσος όρος των πόντων και τα αμυντικά ριμπάουντ (rebound) είναι καθοριστικοί παράγοντες στην πρόβλεψη του μισθού.

Τιμές SHAP (SHAP values)

Οι τιμές SHAP παρέχουν μια ενοποιημένη προσέγγιση για την επεξήγηση των αποτελεσμάτων οποιουδήποτε μοντέλου μηχανικής εκμάθησης, συμπεριλαμβανομένων του μοντέλου παλινδρόμησης CatBoost. Βασίζονται σε έννοιες της θεωρίας παιγνίων.

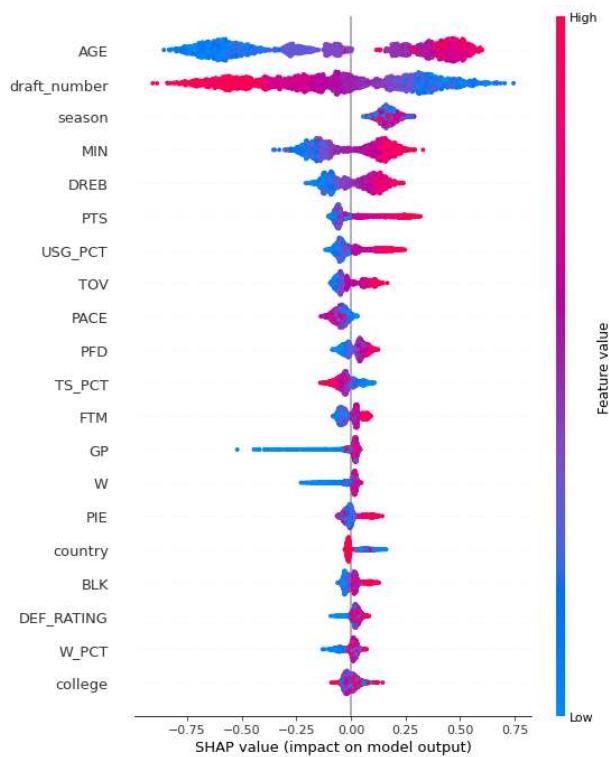
Οι μέσες απόλυτες τιμές SHAP (mean [shap values]) δείχνουν την συνολική επίδραση ενός χαρακτηριστικού στην πρόβλεψη.



Εικόνα 4-31 Μέσες απόλυτες τιμές SHAP – Σημαντικά χαρακτηριστικά

Παρατηρήσεις: Ομοίως παρατηρείται ότι οι μέσες απόλυτες τιμές SHAP αναγνωρίζουν την ηλικία, το αριθμό επιλογής στο ντραφτ (draft), τον χρόνο συμμετοχής, τα αμυντικά ριμπάουντ (rebound) και τον μέσο όρο των πόντων ως σημαντικά χαρακτηριστικά. Συμπληρωματικά διατυπώνουν ότι και η σεζόν (season) είναι σημαντική στην πρόβλεψη του μισθού.

Οι υψηλές/χαμηλές τιμές του κάθε χαρακτηριστικού που ωθούν τις προβλέψεις προς τα πάνω ή προς τα κάτω ερμηνεύουν την κατευθυντικότητα (directionality) του.



Εικόνα 4-32 Τιμές SHAP – Αντίκτυπο της τιμής των χαρακτηριστικών στην έξοδο του μοντέλου

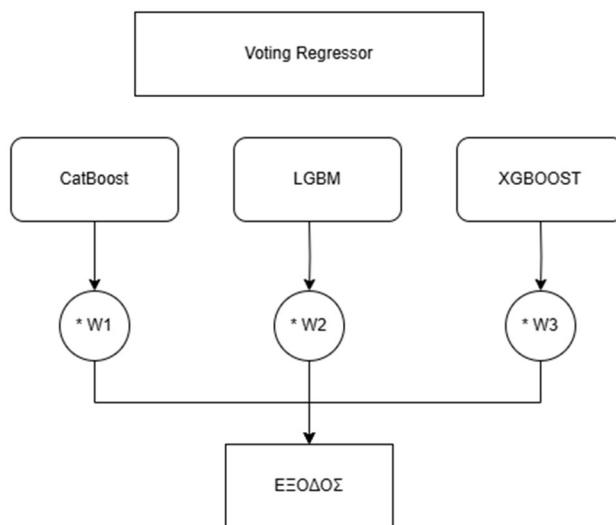
Παρατηρήσεις: Με την χρήση των τιμών SHAP μπορούν να βγουν συμπεράσματα όχι μόνο για την σημασία ενός χαρακτηριστικού αλλά και για την επίδραση του στην τελική τιμή πρόβλεψης. Παρατηρείται ότι η ηλικία όχι μόνο είναι σημαντικό χαρακτηριστικό που βρίσκεται στην κορυφή του διαγράμματος αλλά και ότι μπορεί να επηρεάσει είτε θετικά είτε αρνητικά την τιμή πρόβλεψης. Πιο συγκεκριμένα μια υψηλή ηλικία οδηγεί σε υψηλότερη πρόβλεψη μισθού ενώ μια χαμηλή οδηγεί σε χαμηλότερη. Αυτό είναι προφανές διότι η υψηλότερη ηλικία στους παίκτες του NBA συνήθως σημαίνει μεγαλύτερη εμπειρία και αποδεδειγμένη απόδοση, οδηγώντας σε υψηλότερους μισθούς, ενώ οι νεότεροι παίκτες έχουν μικρότερες αμοιβές λόγω αβεβαιότητας και περιορισμών των συμβολαίων τους (rookie contracts). Αντιθέτως, ένα υψηλός αριθμός επιλογής στο ντραφτ (draft) οδηγεί σε χαμηλότερη πρόβλεψη μισθού ενώ ένας χαμηλός οδηγεί σε υψηλότερη. Οι

παίκτες που επιλέγονται νωρίς στο ντραφτ συνήθως λαμβάνουν μεγαλύτερους μισθούς λόγω υψηλότερου δυναμικού και ικανότητάς, ενώ οι παίκτες που επιλέγονται αργότερα ξεκινούν με μικρότερα συμβόλαια, πρέπει να αναδειχθούν για να κερδίσουν υψηλές αμοιβές και συχνά καταλήγουν ως «role players» που έχουν δευτερεύοντα ρόλο στις ομάδες τους.

4.5 Μέθοδοι συνόλου – Ensembles

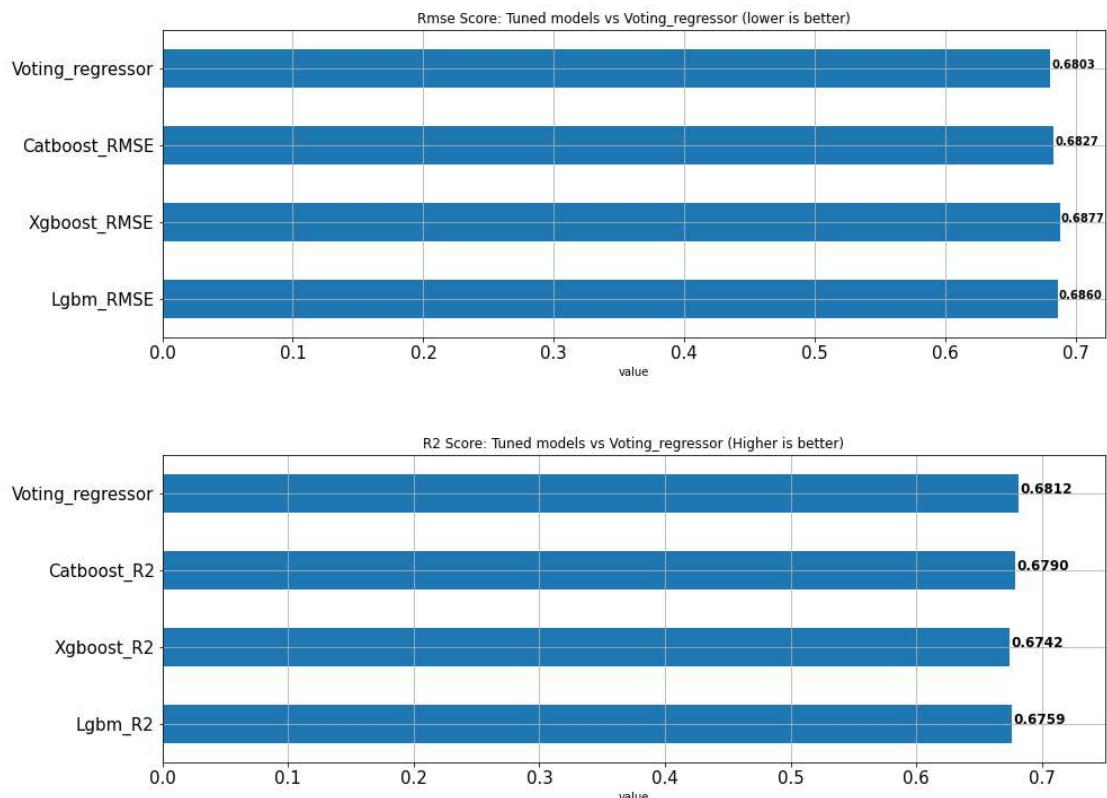
Παλινδρόμηση «ψήφου» (Voting Regressor)

Τα ίδια δεδομένα εκπαίδευσης τροφοδοτούνται και στα τρία μοντέλα παλινδρόμησης (CatBoost, XGBoost, LightGBM). Ο Voting Regressor συνδυάζει τις προβλέψεις χρησιμοποιώντας τα βάρη, με το μοντέλο CatBoost να έχει την μεγαλύτερη επιρροή αφού προηγουμένως φάνηκε να έχει την καλύτερη απόδοση.



Εικόνα 4-33 Αρχιτεκτονική μοντέλου παλινδρόμησης «ψήφου» (Voting regressor)

Ακολουθεί η αξιολόγηση του μοντέλου παλινδρόμησης «ψήφου» και η σύγκριση με τα προηγουμένως βελτιστοποιημένα μοντέλα.

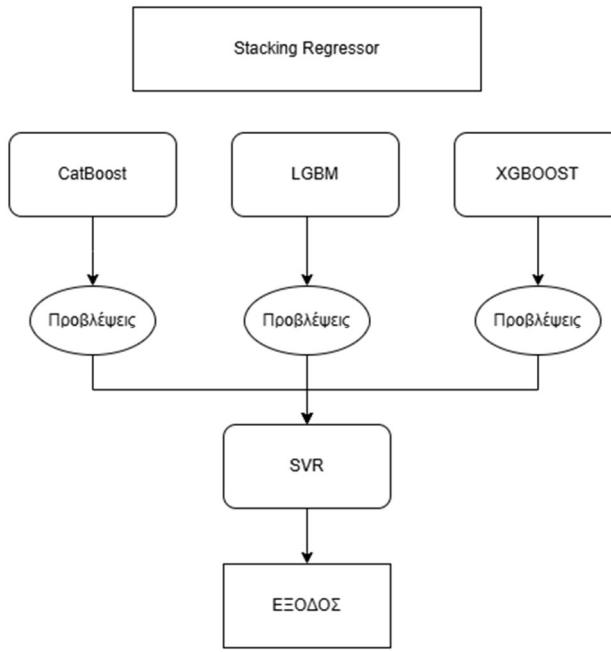


Εικόνα 4-34 και 4-35 Σύγκριση απόδοσης μεταξύ των βελτιστοποιημένων μοντέλων και των μεθόδων συνόλου

Παρατηρήσεις: Παρατηρείται ότι η μέθοδος συνόλου (ensemble) που συνδυάζει τις προβλέψεις των τριών μοντέλων έχει παρόμοια απόδοση με το ισχυρό μοντέλο παλινδρόμησης CatBoost και πιο συγκεκριμένα εμφανίζει μικρή βελτίωση.

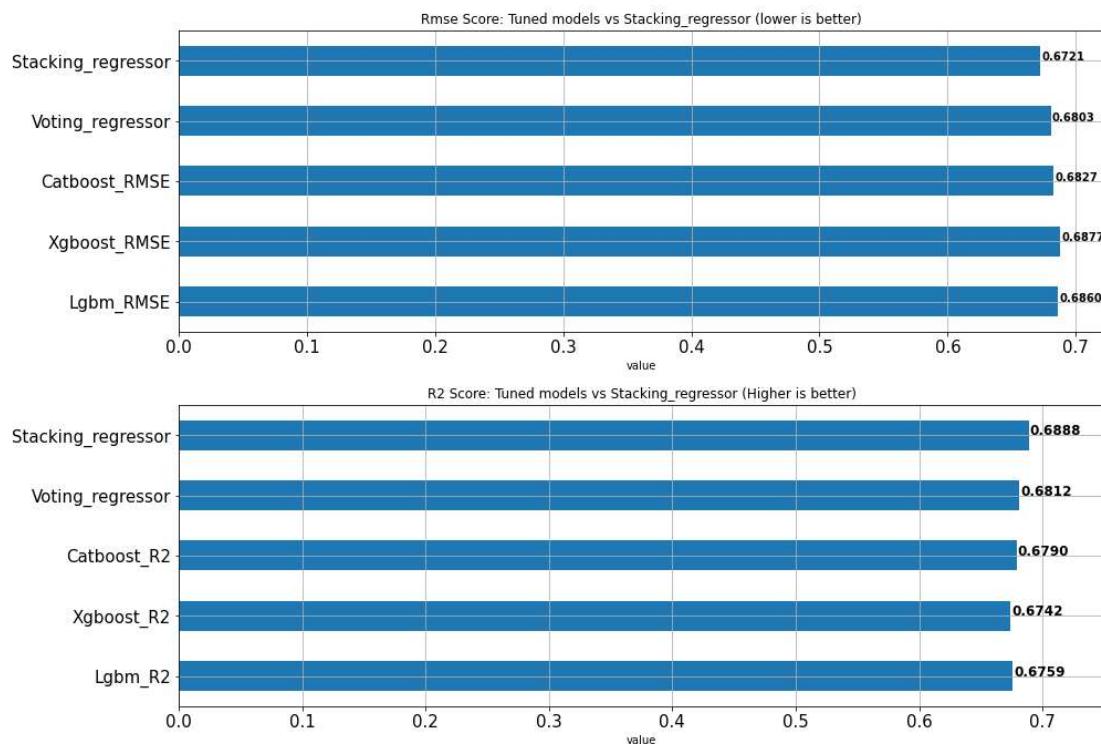
Παλινδρόμηση «στοιβας» (Stacking Regressor).

Κάθε μοντέλο δημιουργεί προβλέψεις. Αυτές οι προβλέψεις γίνονται μεταχαρακτηριστικά (meta-features). Οι προβλέψεις στοιβάζονται για να σχηματίσουν ένα νέο σύνολο δεδομένων. Το SVR εκπαιδεύεται στα μεταχαρακτηριστικά (meta-features) και μαθαίνει πώς να συνδυάζει τις προβλέψεις των βασικών μοντέλων.



Εικόνα 4-36 Αρχιτεκτονική μοντέλου παλινδρόμησης «Στοίβας»

Ακολουθεί η αξιολόγηση του μοντέλου παλινδρόμησης «στοίβας» (stacking regressor) και η σύγκριση με τα προηγουμένως βελτιστοποιημένα μοντέλα καθώς και με το προηγούμενο μοντέλο παλινδρόμησης «ψήφου» (voting regressor).

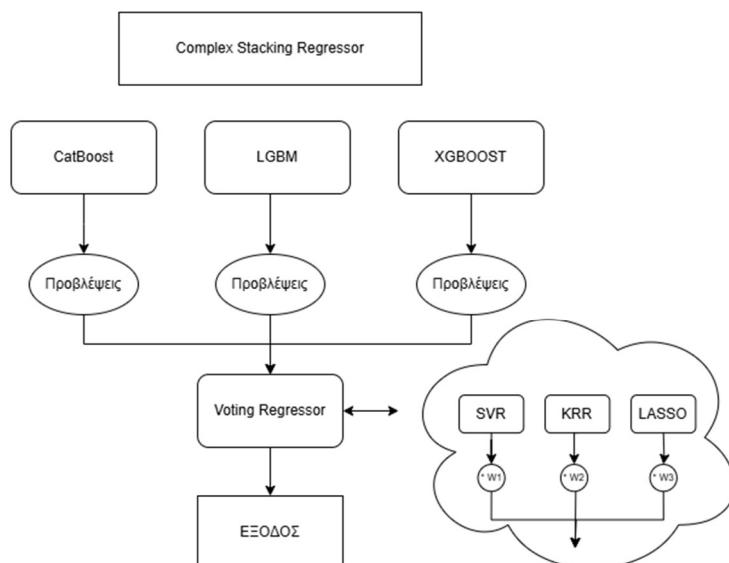


Εικόνα 4-37 και 4-38 Σύγκριση απόδοσης μεταξύ των βελτιστοποιημένων μοντέλων και των μεθόδων συνόλου

Παρατηρήσεις: Παρατηρείται ότι το μοντέλο παλινδρόμησης «στοίβας» που χρησιμοποιεί τις προβλέψεις των τριών μοντέλων για την εκπαίδευση ενός τρίτου μοντέλου SVR έχει μικρότερο σφάλμα πρόβλεψης (RMSE) και εμφανίζει μεγαλύτερο R2 συγκριτικά με τα βασικά μοντέλα και το μοντέλο παλινδρόμησης «ψήφου».

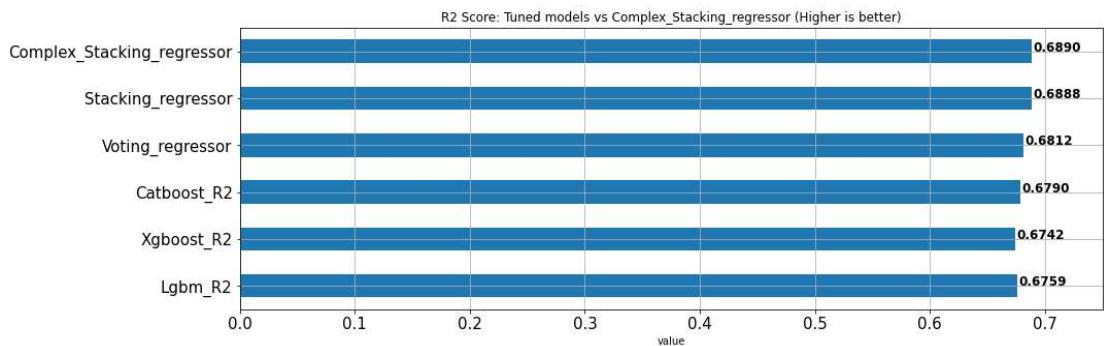
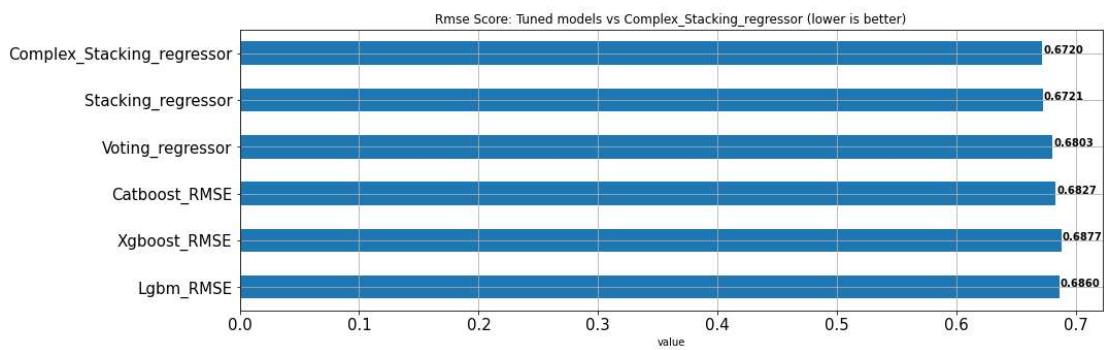
Σύνθετο μοντέλο παλινδρόμησης «στοίβας» - «ψήφου»

Συνδυάζοντας τις μεθόδους της παλινδρόμησης «στοίβας» (stacking regressor) και της παλινδρόμησης «ψήφου» (voting regressor) μπορούν δομηθούν πιο σύνθετες διατάξεις.



Εικόνα 4-39 Αρχιτεκτονική σύνθετου μοντέλου «Στοίβας» - «Ψήφου»

Ακολουθεί η αξιολόγηση του σύνθετου μοντέλου παλινδρόμησης (complex stacking regressor) και η σύγκριση με τα προηγουμένως βελτιστοποιημένα μοντέλα καθώς και με τις προηγούμενες μεθόδους.



Εικόνα 4-40 και 4-41 Σύγκριση απόδοσης μεταξύ των βελτιστοποιημένων μοντέλων (*optimized*) και των μεθόδων συνόλου (*ensembles*)

Παρατηρήσεις: Παρατηρείται ότι αυτή η σύνθετη μέθοδος συνόλου εμφανίζει παρόμοιο σφάλμα πρόβλεψης και τιμή R2 (πολύ μικρή βελτίωση) με το μοντέλο παλινδρόμησης «στοίβας». Τελικά, οι μέθοδοι συνόλου εμφάνισαν μικρότερα σφάλματα από τα βασικά μοντέλα και πιο συγκεκριμένα οι μέθοδοι «στοίβας» (απλή και σύνθετη – Stacking regressor και Complex Stacking regressor) είχαν την μεγαλύτερη απόδοση.

Ακολουθεί συγκεντρωτικός πίνακας με τις τιμές των αποτελεσμάτων.

Μοντέλα	RMSE	R2
Complex_Stacking_reg	0.6720	0.6890
Stacking_reg	0.6721	0.6888
Voting_reg	0.6803	0.6812
Catboost (optimized)	0.6827	0.6790
XgBoost (optimized)	0.6877	0.6742
Lgbm (optimized)	0.6860	0.6759

Εικόνα 4-42 Συγκεντρωτικός πίνακας απόδοσης βελτιστοποιημένων (*optimized*) μοντέλων και μεθόδων συνόλου (*ensembles*)

Το σύνθετο μοντέλο συνόλου “Complex_Stacking_reg” που συνδυάζει τις μεθόδους «ψήφου» (voting regressor) και «στοίβας» (stacking regressor) εμφανίζει την καλύτερη απόδοση.

4.6 Ακραίες τιμές - Outliers

Η ανάλυση των περιπτώσεων ακραίων τιμών (outliers) μπορεί να αποβεί χρήσιμη όταν αξιολογείται η πρακτικότητα του μοντέλου. Εξετάζοντας τις ακραίες τιμές (outliers) μπορούν να αποκαλυφθούν αδυναμίες στην ικανότητα του μοντέλου να αξιολογεί συγκεκριμένες περιπτώσεις παικτών. Επιπλέον μπορούν να αποκαλυφθούν κρυμμένα μοτίβα τα οποία αντιπροσωπεύουν ειδικές περιπτώσεις και ακολουθούν διαφορετικούς κανόνες. Στην προκειμένη περίπτωση αυτή η επίγνωση μπορεί να συντελέσει στη βελτίωση των επιχειρηματικών αποφάσεων των ομάδων και των μάνατζερ και προσδιορίζει παίκτες που ενδέχεται να υπερεκτιμώνται/υποτιμώνται σύμφωνα με τα αναλυτικά δεδομένα.

Ακολουθεί ένας πίνακας με τους παίκτες των οποίων ο μισθός προβλέπεται πολύ χαμηλότερος από την πραγματικότητα κατά τη σεζόν 2022-2023.

Πρόβλεψη (log)	Πραγματική τιμή (log)	Πρόβλεψη (εκατο- μμύρια)	Πραγματική τιμή (εκατο- μμύρια)	Όνομα παίκτη	Σ λογή	Διαφορά (ειδικήτων)
16,285	17,702	11,821	48,752	john wall	22	-36,9
16,801	17,697	19,790	48,478	russell westbrook	22	-28,6
16,096	17,458	9,779	38,198	trae young	22	-28,4
16,048	17,413	9,327	36,501	ben simmons	22	-27,1
16,955	17,717	23,095	49,497	stephen curry	22	-26,4
16,383	17,482	13,036	39,112	khris middleton	22	-26,0
16,444	17,487	13,858	39,306	rudy gobert	22	-25,4
16,490	17,473	14,500	38,750	tobias harris	22	-24,2

Εικόνα 4-43 Πίνακας ακραίων τιμών (αρνητική διαφορά)

Σύμφωνα με το άρθρο «The most overpaid players in the NBA in 2022-23» («Οι πιο υπερπληρωμένοι παίκτες στο NBA τη σεζόν 2022-23») του Hoopshype ο John wall είναι ο πιο υπερεκτιμημένος παίκτης κατά την διάρκεια αυτής της σεζόν, ο οποίος προσπάθησε να ανακάμψει από πολλαπλούς σοβαρούς τραυματισμούς στο πόδι αλλά δεν κατάφερε να

φτάσει στο επίπεδο που δικαιολογεί τον υψηλό του μισθό. Ο Russell Westbrook βρίσκεται στη τρίτη θέση αυτής της λίστας με υπέρογκο συμβόλαιο και απόδοση που δεν αντιστοιχεί σε αυτή ενός ακριβοπληρωμένου Σούπερ σταρ. Ο Ben Simmons ο οποίος αντιπροσωπεύει μια ιδιαίτερα ενδιαφέρουσα περίπτωση εξαίρεσης στο NBA, πληρώθηκε ανάλογα με το δυνητικό του δυναμικό και την μελλοντική ικανότητα να γίνει ένας από τους καλύτερους παίκτες του πρωταθλήματος. Δυστυχώς οι τραυματισμοί, η ψυχική του κατάσταση και η δραστικά μειωμένη του απόδοση δεν δικαιολογούν το συμβόλαιο του. Τα τελευταία χρόνια φάνηκε πως παρά τα τρομερά αθλητικά του προσόντα η αδυναμία του στο σουτ και η απροθυμία του να βελτιωθεί τον κατέστησε ως μια ριψοκίνδυνη επένδυση που τελικά δεν απέδωσε ποτέ. Ο Chris Middleton, ο Tobias Harris και ο Rudy Gobert αποτελούν και αυτοί για διαφορετικούς λόγους μέλη της λίστας στις θέσεις πέντε, έντεκα και δώδεκα αντίστοιχα.

Βέβαια δεν ανήκουν όλοι οι παίκτες του πίνακα σε αυτή την λίστα. Οι Stephen Curry και Trae Young είναι ακριβοπληρωμένοι αλλά δεν θεωρούνται υπερεκτιμημένοι διότι αποτελούν franchise players (αθλητές που είναι οι καλύτεροι παίκτες στην ομάδα τους) των οποίων η πραγματική αξία συχνά υπερβαίνει τα στατιστικά στοιχεία, γεγονός που καθιστά δύσκολη την αξιολόγησή τους με τα παραδοσιακά μοντέλα. Στο παρκέ ανυψώνουν τους συμπαίκτες τους και αποδίδουν σε στιγμές υψηλής πίεσης στα πλεί οφ (play off). Εκτός παρκέ η αξία τους μπορεί να φανεί στην πώληση φανελών, στην πώληση εισιτήριων, σε θεατές που προσελκύουν καθώς και σε χορηγίες και επαφές με τα μέσα ενημέρωσης.

Ακολουθεί ο πίνακας με τους παίκτες των οποίων ο μισθός προβλέπεται υψηλότερος κατά τη σεζόν 2022-2023.

Πρόβλεψη (log)	Πραγματική τιμή (log)	Πρόβλεψη (εκατο- μύρια)	Πραγματική τιμή (εκατο- μύρια)	Όνομα παίκτη	ΣΥΛΛΟΓΗ (εκατομμύρια)	ΔΙΑΦΟΡΑ
16,053	14,601	9,372	2,193	desmond bane	22	7,179
16,574	15,618	15,779	6,063	cameron johnson	22	9,717
16,239	13,845	11,290	1,030	kris dunn	22	10,260

Εικόνα 4-44 Πίνακας ακραίων τιμών (θετική διαφορά)

Παρατηρώντας τους παίκτες τους οποίους το μοντέλο θεωρεί πως η απόδοση τους δικαιολογεί υψηλότερη αμοιβή, βλέπουμε τα ονόματα των Desmond Bane, Cameron Johnson και Kris Dunn. Αυτοί οι τρεις παίκτες, οι οποίοι δεν αποτελούν σούπερ σταρς, λαμβάνουν συχνά τον τίτλο του underrated (υποτιμημένου) από τους παράγοντες και τους φαν του πρωταθλήματος, γεγονός το οποίο επιβεβαιώνεται και από τις προβλέψεις του μοντέλου.

Συμπεραίνοντας, στις ακραίες τιμές (outliers) των προβλέψεων μπορούν να εντοπιστούν υπερεκτιμημένοι παίκτες που αμείβονται δυσανάλογα με τις ατομικές τους αποδόσεις, αποθαρρύνοντας με αυτό τον τρόπο τα συμβόλαια υψηλού ρίσκου. Αντιθέτως, μπορούν να εντοπιστούν υποτιμημένοι μισθολογικά παίκτες οι οποίοι μπορούν να αποτελέσουν ικανούς υποστηρικτικούς πυλώνες της ομάδας. Παράλληλα επισημαίνεται η αδυναμία του μοντέλου να αξιολογήσει franchise παίκτες των οποίων η αξία υπερβαίνει τα κλασσικά στατιστικά.

5. Συμπεράσματα

Στην παρούσα εργασία, αναπτύχθηκαν μοντέλα παλινδρόμησης για την πρόβλεψη των μισθών των παικτών του NBA χρησιμοποιώντας εποχιακά στατιστικά στοιχεία ατομικής απόδοσης. Το μοντέλο παλινδρόμησης (regression model) CatBoost αναδείχθηκε ως το καλύτερο βασικό (base) μοντέλο, επιδεικνύοντας ισχυρή ακρίβεια πρόβλεψης πριν και μετά την βελτιστοποίηση μέσω αναζήτησης πλέγματος, ενώ οι μέθοδοι συνόλου (μέθοδος «στοιβας – stacking regressor» και «ψήφου – voting regressor») βελτίωσαν περαιτέρω την απόδοση. Τα χαρακτηριστικά με τη μεγαλύτερη επιρροή περιλάμβαναν την σεζόν (εμπειρία), τον αριθμό του ντραφτ (draft), την ηλικία, τους πόντους ανά παιχνίδι, τα λεπτά συμμετοχής και τα αμυντικά ριμπάουντ (DREB), αναδεικνύοντας την ισχυρή συσχέτισή τους με τον μισθό. Αντίθετα, χαρακτηριστικά όπως τα τριπλά νταμπλ (TD3), τα διπλά νταμπλ (DD2), τα επιθετικά ριμπάουντ (OREB), η χώρα καταγωγής, τα κλεψίματα (STL) και οι μπλοκαρισμένες προσπάθειες σουτ (BLKA) είχαν ελάχιστη επίδραση, γεγονός που υποδηλώνει ότι είναι λιγότερο κρίσιμα για τον καθορισμό του μισθού.

Αυτή η εργασία παρουσίασε αρκετές προκλήσεις, όπως ο θόρυβος των δεδομένων, η μεταβλητότητα των συμβολαίων και οι μη γραμμικές σχέσεις μεταξύ απόδοσης και μισθού. Επιπλέον, εξωτερικοί παράγοντες όπως η ζήτηση στην αγορά, η φήμη των παικτών και η αξία των franchise players (που είναι δύσκολο να ποσοτικοποιηθεί) μπορεί να επηρεάσουν τις αποδοχές πέρα από τα καθαρά στατιστικά στοιχεία.

Συνοψίζοντας, η μηχανική μάθηση (Machine Learning) παρέχει πολύτιμες πληροφορίες για τις τάσεις των μισθών στο NBA, αλλά παράλληλα η τέλεια πρόβλεψη παραμένει άπιαστη λόγω της πολύπλοκης, πολύπλευρης φύσης της αποτίμησης των παικτών.

6. Μελλοντικές Επεκτάσεις

Για την περαιτέρω βελτίωση της ακρίβειας και της αξιοπιστίας των μοντέλων πρόβλεψης του μισθού στο NBA, η μελλοντική έρευνα θα μπορούσε να ενσωματώσει μετρικές που βασίζονται στην αγορά και αποτυπώνουν καλύτερα την εμπορική επιρροή και την άυλη αξία ενός παίκτη. Η δημοτικότητα των παικτών μπορεί να μετρηθεί μέσω μετρικών, όπως οι ακόλουθοι στα μέσα κοινωνικής δικτύωσης (Twitter, Instagram, TikTok), τα οποία υποδεικνύουν τη δέσμευση των οπαδών και την εμπορική αξία του παίκτη. Ο υψηλός αριθμός οπαδών μεταφράζεται συχνά σε μεγάλες συμφωνίες και ευκαιρίες χορηγίας. Η κατάταξη των πωλήσεων της φανέλας του παίκτη (από τα δεδομένα των μαγαζιών του NBA) μπορεί να χρησιμεύσει ως μετρική της εμπορευσιμότητας, καθώς οι παίκτες με τις υψηλότερες πωλήσεις συνήθως αυξάνουν τα έσοδα και δικαιολογούν υψηλότερους μισθούς. Επιπλέον, οι τάσεις (search trends) της Google και ο όγκος αναζητήσεων των ονομάτων των παικτών αντικατοπτρίζουν το ενδιαφέρον του κοινού σε πραγματικό χρόνο, βοηθώντας στον εντοπισμό των αστέρων που ξεχωρίζουν.

Ωστόσο, αυτές οι μετρήσεις έχουν περιορισμούς. Για παράδειγμα οι αριθμοί στα μέσα κοινωνικής δικτύωσης μπορεί να διογκωθούν από πλασματικά προφίλ (bots), δεν είναι εύκολη η πρόσβαση σε ιστορικά δεδομένα ακολούθων, τα δεδομένα πωλήσεων φανέλας μπορεί να είναι δυσπρόσιτα και ελλιπή και οι τάσεις αναζήτησης μπορεί να αυξηθούν προσωρινά χωρίς μακροπρόθεσμο οικονομικό αντίκτυπο. Παρά τις προκλήσεις αυτές, ο συνδυασμός αυτών των δεικτών μπορεί να παρέχει πολύτιμες πληροφορίες για την εμπορική ελκυστικότητα ενός παίκτη πέρα από την καθαρή απόδοση στο γήπεδο.

7. Παράρτημα: Βιβλιοθήκες

```
import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
import seaborn as sns  
import cpi  
from scipy.stats import skew  
from sklearn.preprocessing import PowerTransformer  
from sklearn.linear_model import ElasticNet,  
from sklearn.linear_model import Lasso  
from sklearn.ensemble import RandomForestRegressor  
from sklearn.kernel_ridge import KernelRidge  
from sklearn.preprocessing import RobustScaler  
from sklearn.metrics import mean_squared_error  
from sklearn.metrics import mean_absolute_percentage_error  
from sklearn.metrics import mean_absolute_error  
import xgboost as xgb  
import lightgbm as lgb  
from catboost import CatBoostRegressor  
from sklearn.metrics import r2_score  
from sklearn.linear_model import LinearRegression  
from sklearn.svm import SVR  
import shap  
import webbrowser  
from sklearn.ensemble import StackingRegressor  
from sklearn.ensemble import VotingRegressor
```

```
import lime  
import lime.lime_tabular  
from sklearn.model_selection import GridSearchCV,TimeSeriesSplit  
from sklearn.model_selection import TimeSeriesSplit
```

8. Βιβλιογραφικές Αναφορές

- [1] A. L. Samuel, “Some Studies in Machine Learning Using the Game of Checkers,” IBM Journal of Research and Development, vol. 3, no. 3, pp. 210–229, Jul. 1959, doi: 10.1147/rd.33.0210.
- [2] T. Jo, Machine learning foundations: Supervised, unsupervised, and advanced learning. Springer International Publishing, 2021, doi: 10.1007/978-3-030-65900-4.
- [3] R. Lyons Jr., E. N. Jackson Jr., and A. Livingston, “Determinants of NBA Player Salaries,” The Sport Journal, May 2015, doi: 10.17682/sportjournal/2015.019.
- [4] Y. Zhao, “Model Prediction of Factors Influencing NBA Players’ Salaries Based on Multiple Linear Regression,” Proceedings of the 2022 2nd International Conference on Economic Development and Business Culture (ICEDBC 2022), pp. 1439–1445, 2022, doi: 10.2991/978-94-6463-036-7_213.
- [5] Ioanna Papadaki and Michail Tsagris, “Are NBA Players’ Salaries in Accordance with Their Performance on Court?,” Contributions to economics, pp. 405–428, Jan. 2022, doi: 10.1007/978-3-030-85254-2_25
- [6] M. Cao, “Predicting NBA Player’s Salary Based on Statistics from the Game Using Linear Regression,” pp. 476–480, Jan. 2024, doi: 10.5220/0012823900004547.
- [7] J. Zhang, “National Basketball Association Salary Prediction: A Data-Driven Linear Regression Analysis,” *Highlights in Business, Economics and Management*, vol. 24, pp. 1059–1064, Jan. 2024, doi: 10.54097/c966f539.
- [8] A. Xiong, “Analysis of NBA player salary based on multiple linear regression model,” *Theoretical and Natural Science*, vol. 51, no. 1, pp. 206–213, Nov. 2024, doi: 10.54254/2753-8818/51/2024CH0205.
- [9] X. Feng, Y. Wang, and T. Xiong, “NBA Player Salary Analysis based on Multivariate Regression Analysis,” vol. 49, pp. 157–166, May 2023, doi: 10.54097/hset.v49i.8498.
- [10] Y. Wang, “NBA Player Salary Projections Based on Gradient Boost in 2022-23 Season,” *Transactions on Computer Science and Intelligent Systems Research*, vol. 5, pp. 236–241, Aug. 2024, doi: 10.62051/ejy6xc14.
- [11] “Examining the Impact of Social Media Following on Player Salary in the National Basketball Association: A Multivariate Statistical Analysis,” *Journal of Applied Business and Economics*, vol. 25, no. 1, Feb. 2023, doi: 10.33423/jabe.v25i1.5995.
- [12] Kaggle, “Kaggle: Your Machine Learning and Data Science Community,” Kaggle. [Online]. Available: <https://www.kaggle.com>.
- [13] HoopsHype, “NBA Player Salaries,” HoopsHype, 2025. [Online]. Available: <https://hoopshype.com/salaries/>.

- [14] ESPN, "NBA Player Salaries - National Basketball Association," ESPN, 2025. [Online]. Available: <https://www.espn.com/nba/salaries>.
- [15] A. C. Atkinson, M. Riani, and A. Corbellini, "The Box–Cox Transformation: Review and Extensions," *Statistical Science*, vol. 36, no. 2, pp. 239–255, 2021, doi: 10.1214/20-STS778.
- [16] I.-K. . Yeo, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 4, pp. 954–959, Dec. 2000, doi: 10.1093/biomet/87.4.954.
- [17] V. Cerqueira, L. Torgo, and I. Mozetič, "Evaluating time series forecasting models: an empirical study on performance estimation methods," *Machine Learning*, vol. 109, Oct. 2020, doi: 10.1007/s10994-020-05910-7.
- [18] A. Jadon, A. Patil, and S. Jadon, "A Comprehensive Survey of Regression Based Loss Functions for Time Series Forecasting," *arXiv.org*, Nov. 05, 2022, doi: 10.48550/arxiv.2211.02989
- [19] A. Di Bucchianico, "Coefficient of Determination (R2)," *Encyclopedia of Statistics in Quality and Reliability*, Mar. 2008, doi: 10.1002/9780470061572.eqr173.
- [20] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [21] Scikit-learn. "Support Vector Regression (SVR) documentation" [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>
- [22] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, 2011, doi: 10.5555/1953048.2078195.
- [23] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/a:1010933404324.
- [24] Scikit-learn, "Random Forest Regressor documentation" [Online]. Available: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>.
- [25] XGBoost Documentation. [Online]. Available: <https://xgboost.readthedocs.io/>.
- [26] LightGBM Documentation. [Online]. Available: <https://lightgbm.readthedocs.io/en/stable/>.
- [27] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., Lightgbm: a highly efficient gradient boosting decision tree, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [28] CatBoost Documentation. [Online]. Available: <https://catboost.ai/docs/en/>.
- [29] L. Prokhorenkova et al., "CatBoost: unbiased boosting with categorical features," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

- [30] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992, doi: 10.1016/S0893-6080(05)80023-1.