National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science
Artificial Intelligence and Learning Systems Laboratory

Machine Learning in Fantasy Premier League

Diploma Thesis

of

SPIROS VALOUXIS

Supervisor: Stefanos Kollias
Professor, NTUA

Athens, June 2023

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science
Artificial Intelligence and Learning Systems Laboratory

Machine Learning in Fantasy Premier League

Diploma Thesis

of

SPIROS VALOUXIS

Supervisor: Stefanos Kollias
Professor, NTUA

Approved by the examination committee on June 21, 2023.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Stefanos Kollias	Athanasios Voulodimos	Giorgos Stamou
Professor, NTUA	Assistant Professor, NTUA	Professor, NTUA

Athens, June 2023

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science
Artificial Intelligence and Learning Systems Laboratory

(Signature)

.....
Spiros Valouxis

Electrical & Computer Engineer Graduate, NTUA

in memory of my mother, Evelyn

Περίληψη

Ο κόσμος των αθλημάτων fantasy έχει γνωρίσει μια έκρηξη στη δημοτικότητα, με το Fantasy Premier League (FPL) να ξεχωρίζει ως ένας από τους πρωτοπόρους στον τομέα αυτό. Αυτή η διπλωματική εργασία επικεντρώνεται στην εφαρμογή τεχνικών μηχανικής μάθησης στο Fantasy Premier League (FPL) για την ενίσχυση της απόδοσης των ομάδων και τη λήψη αποφάσεων. Οι ερευνητικοί στόχοι ήταν τρεις: να αναπτυχθούν μοντέλα μηχανικής μάθησης για την πρόβλεψη της Αναμενόμενης Αξίας (Expected Value, EV), να μπορούν να προταθούν βέλτιστες κινήσεις (μεταγραφές) με βάση τις προβλεπόμενες τιμές της αξίας των παικτών και να συνεισφέρει στον τομέα της ανάλυσης του FPL προωθώντας την εφαρμογή τεχνικών μηχανικής μάθησης.

Η αυξανόμενη δημοτικότητα του παιχνιδιού (πάνω από 11,4 εκατομμύρια ομάδες στη σεζόν 2022/23) έχει αυξήσει τη ζήτηση για στρατηγικές που βασίζονται στα δεδομένα με στόχο την απόκτηση ανταγωνιστικού πλεονεκτήματος. Παρότι προηγούμενες προσπάθειες της κοινότητας του FPL έχουν εξερευνήσει παρόμοιες έννοιες, η μελέτη μας παρουσιάζει το πρώτο ανοιχτού κώδικα, πλήρες μοντέλο FPL που χρησιμοποιεί τεχνικές μηχανικής μάθησης.

Τα αποτελέσματα της έρευνάς μας δείχνουν την ανωτερότητα του μοντέλου μας σε σχέση με τα περισσότερα άλλα υπάρχοντα μοντέλα. Μέσω αυστηρής αξιολόγησης χρησιμοποιώντας διάφορες μετρικές (MAE, RMSE, R^2 σκορ), τα μοντέλα μας παρουσίασαν συνεχώς καλύτερη απόδοση από σχεδόν όλους τους ανταγωνιστές στην πρόβλεψη της Αναμενόμενης Αξίας. Επιπλέον, ένα σημαντικό επίτευγμα ήταν η πρώτη θέση που κατέκτησε η ομάδα που διαχειρίστηκε το μοντέλο μας στη GW24 λίγκα, ξεπερνώντας πάνω από 25.000 ομάδες μέχρι το τέλος της σεζόν.

Ενσωματώνοντας με επιτυχία αλγόριθμους μηχανικής μάθησης και βελτιστοποίησης, η εργασία μας παρέχει στους παίκτες του FPL συγκεκριμένες συστάσεις για την επιλογή ομάδας, τις μεταγραφές και τη στρατηγική λήψη αποφάσεων. Αυτή η έρευνα συνεισφέρει σημαντικά στον τομέα της ανάλυσης του FPL, προσφέροντας μια καινοτόμο λύση που βελτιώνει την απόδοση των ομάδων και παρέχει πολύτιμες πληροφορίες για τους παίκτες που επιθυμούν να βελτιώσουν τις επιδόσεις τους στο FPL.

Λέξεις Κλειδιά

Fantasy Premier League, Μηχανική Μάθηση, Αναμενόμενη Αξία, Βελτιστοποίηση, FPL Analytics

Abstract

The world of fantasy sports has witnessed a surge in popularity, with Fantasy Premier League (FPL) standing out as a leader in the field. This diploma thesis focuses on the application of machine learning techniques in Fantasy Premier League (FPL) to enhance team performance and decision-making. The research objectives were threefold: to develop machine learning models for predicting Expected Value (EV), to generate optimal moves based on the predicted EV values, and to contribute to the field of FPL analytics by advancing the application of machine learning techniques.

The growing popularity of FPL (over 11.4 million teams in the 2022/23 season) has increased the demand for data-driven strategies across the community to gain a competitive edge. While previous efforts by the FPL community have explored similar concepts, our study presents the first open-source, full-scale FPL model utilizing machine learning techniques.

The results of our research demonstrate the superiority of our model compared to most of the other existing models. Through rigorous evaluation using various metrics (MAE, RMSE, R^2 score), our models consistently outperformed almost all competitors in predicting EV. Furthermore, a noteworthy achievement was the top-ranking performance of the team managed by our model in the GW24 league, surpassing over 25,000 competing teams by the end of the season.

By successfully integrating machine learning and optimization algorithms, our project provides FPL managers with actionable recommendations for team selection, transfers, and strategic decision-making. This research significantly contributes to the field of FPL Analytics by offering an innovative solution that enhances team performance and provides valuable insights for managers seeking to improve their FPL results.

Keywords

Fantasy Premier League, Machine Learning, Expected Value, Optimization, FPL Analytics

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Professor Stefanos Kollias, for his guidance and mentorship throughout the process of my diploma thesis. I would also like to extend my sincere appreciation to Paraskevi Tzouveli for her valuable contributions, assistance, and continuous support throughout this journey. Her expertise and guidance have been immensely valuable in refining the research methodology and achieving the desired outcomes.

I would like to express my deepest gratitude to my father, Apostolos, and my sister, Eva, for their unwavering support, encouragement, and belief in my abilities. Their constant presence and words of encouragement have been a source of inspiration and motivation.

I would like to extend my gratitude to all my friends, who supported, motivated, and believed in me along the way. Their friendship and support have been invaluable throughout this journey. Special thanks to Thanos, without whom I probably wouldn't finish my degree.

I would also like to express my appreciation to Sertalp, FPL Kiwi, Chris Musson, Owen, Fantasy Football Trout, and all the other members of the FPL community who have generously shared their knowledge, insights, and experiences. Their contributions and discussions have played a significant role in shaping my understanding of the subject matter and achieving my goals.

Lastly, I would like to acknowledge and thank all the individuals who have contributed in various ways to the successful completion of this thesis. Your support, encouragement, and contributions have been greatly appreciated.

Thank you all for being a part of this incredible journey and for making this achievement possible!

Athens, June 2023

Spiros Valouxis

Contents

List of Figures

List of Tables

Εκτεταμένη Ελληνική Περίληψη

Στο κεφάλαιο αυτό παρουσιάζεται μία εκτεταμένη περίληψη της εργασίας αυτής στα ελληνικά.

0.1 Εισαγωγή

Το Fantasy Premier League (FPL) είναι ένα δημοφιλές διαδικτυακό παιχνίδι που επιτρέπει στους συμμετέχοντες να δημιουργήσουν εικονικές ομάδες αποτελούμενες από πραγματικούς ποδοσφαιριστές της Premier League. Παρέχει στους λάτρεις του ποδοσφαίρου μια διαδραστική πλατφόρμα για να αλληλεπιδράσουν με την Premier League και να επιδείξουν τις "προπονητικές" τους δεξιότητες. Το παιχνίδι αναθέτει πόντους στους παίκτες με βάση την απόδοσή τους σε πραγματικούς αγώνες της Premier League, όπως τα γκολ που σκοράρουν, οι ασίστ που κάνουν, τις ανέπαφες εστίες (clean sheets) που κρατάνε και άλλες στατιστικές συνεισφορές. Το ελκυστικό σημείο του παιχνιδιού είναι η δυνατότητά του να δημιουργήσει μια εικονική εμπειρία διαχείρισης ομάδας, όπου οι συμμετέχοντες πρέπει να επιλέξουν προσεκτικά την ομάδα τους, να κάνουν αλλαγές, να επιλέξουν αρχηγούς και να αποφασίσουν σχηματισμό και τακτική. Οι παίκτες ανταγωνίζονται μεταξύ τους, προσπαθώντας να συγκεντρώσουν το μεγαλύτερο αριθμό πόντων και να επιτύχουν υψηλές κατατάξεις εντός των πρωταθλημάτων με τους φίλους τους, ή παγκοσμίως.

Το παιχνίδι έχει γνωρίσει εκθετική ανάπτυξη τα τελευταία χρόνια, προσελκύοντας εκατομμύρια συμμετέχοντες σε παγκόσμιο επίπεδο (πάνω από 11 εκατομμύρια ομάδες για τη σεζόν 2022/23). Με την αυξανόμενη δημοφιλία του παιχνιδιού, παρατηρείται μια αύξηση στη χρήση της ανάλυσης δεδομένων και τεχνικών μηχανικής μάθησης για την απόκτηση ανταγωνιστικού πλεονεκτήματος. Οι παίκτες του FPL αναζητούν τρόπους για να αξιοποιήσουν στρατηγικές που βασίζονται σε δεδομένα και προβλεπτικά μοντέλα, προκειμένου να βελτιστοποιήσουν τις επιλογές τους στην ομάδα, να βελτιώσουν τη λήψη αποφάσεων και τελικά να ενισχύσουν τη συνολική τους επίδοση στο παιχνίδι. Αυτό έχει οδηγήσει στην εξερεύνηση διαφόρων αλγορίθμων μηχανικής μάθησης, στατιστικών μοντέλων και τεχνικών βελτιστοποίησης για την πρόβλεψη της απόδοσης των παικτών, την εκτίμηση της αξίας τους και τη δημιουργία βέλτιστων στρατηγικών. Με βάση τη σημασία και τον αυξανόμενο ενδιαφέρον για την εφαρμογή της μηχανικής μάθησης στο FPL, αυτή η διπλωματική εργασία στοχεύει να συμβάλει στον τομέα αναπτύσσοντας προβλεπτικά μοντέλα για την αναμενόμενη αξία των παικτών και να προτείνει βέλτιστες κινήσεις βασιζόμενη σε αυτές τις προβλέψεις.

Το κίνητρο πίσω από αυτή τη διπλωματική εργασία απορρέει από τη δυνατότητα της μηχανικής μάθησης να βελτιώσει σημαντικά την απόδοση και τη λήψη αποφάσεων στο FPL. Παραδοσιακά, οι παίκτες βασίζονται σε υποκειμενικές αξιολογήσεις, ενστικτώδεις αποφάσεις

και περιορισμένα δεδομένα για να πάρουν αποφάσεις. Ωστόσο, υπάρχει μια αυξανόμενη αναγνώριση της ανάγκης για περισσότερες αποφάσεις βασισμένες σε δεδομένα και αντικειμενικές προσεγγίσεις στο FPL. Το παιχνίδι ανταμείβει τη συνέπεια και το μακροπρόθεσμο σχεδιασμό, καθιστώντας κρίσιμη τη χρήση δεδομένων για την κατάκτηση ανταγωνιστικού πλεονεκτήματος. Οι τεχνικές μηχανικής μάθησης προσφέρουν ισχυρά εργαλεία για την ανάλυση μεγάλου όγκου δεδομένων. Εκπαιδεύοντας μοντέλα μηχανικής μάθησης σε ιστορικά δεδομένα του FPL, μπορούν να ανακαλυφθούν μοτίβα, συσχετίσεις και τάσεις μεταξύ των στατιστικών των παικτών, των ιστορικών δεδομένων και άλλων σχετικών παραγόντων που μπορούν να επηρεάσουν την απόδοση των παικτών.

Οι κύριοι στόχοι αυτής της διπλωματικής εργασίας είναι η ανάπτυξη και η αξιολόγηση μοντέλων μηχανικής μάθησης για το FPL. Ειδικότερα, το ενδιαφέρον επικεντρώνεται στην πρόβλεψη της Αναμενόμενης Αξίας για τους παίκτες και τη δημιουργία βέλτιστων κινήσεων με βάση αυτές τις προβλέψεις. Με την επίτευξη αυτών των στόχων, η διπλωματική εργασία αποσκοπεί να συμβάλει στον τομέα της ανάλυσης του FPL και να προωθήσει τις εφαρμογές της μηχανικής μάθησης στο πλαίσιο του FPL.

0.2 Περιγραφή Προβλήματος και Στόχοι

0.2.1 Κανόνες

Το Fantasy Premier League βασίζεται στην απόδοση των παικτών στην Αγγλική Premier League. Οι παίκτες έχουν τη δυνατότητα να σχηματίσουν τη δική τους ομάδα με έναν προϋπολογισμό των £100 εκατομμυρίων. Ο στόχος είναι να επιλεγεί μια ομάδα 15 παικτών (συμπεριλαμβανομένων 2 τερματοφυλάκων, 5 αμυντικών, 5 μέσων και 3 επιθετικών) που θα σκοράρει τους περισσότερους πόντους με βάση την απόδοσή τους στους αγώνες της Premier League. Οι παίκτες πρέπει να επιλέγουν μια αρχική ενδεκάδα παικτών από την ομάδα τους με τους 15 πριν από κάθε αγωνιστική εβδομάδα, συμπεριλαμβανομένου ενός αρχηγού και ενός αντι-αρχηγού. Ο αρχηγός κερδίζει διπλούς πόντους, και ο αντι-αρχηγός μπορεί να πάρει τη θέση του αρχηγού αν αυτός δεν αγωνιστεί [Σχήμα 1]. Οι πόντοι απονέμονται για διάφορες ενέργειες στο γήπεδο, όπως το σκοράρισμα γκολ, η παροχή ασίστ, η διατήρηση ανέπαφης εστίας, οι αποκρούσεις και υπάρχουν και πόντοι μπόνους. Ο αριθμός των πόντων που απονέμονται για κάθε ενέργεια ποικίλει ανάλογα με τη θέση του παίκτη [12].

- Σε έναν αγώνα, οι τρεις καλύτεροι παίκτες καθορίζονται σύμφωνα με το Σύστημα Βαθμολόγησης Bonus του FPL και τους απονέμεται μπόνους 1, 2 και 3 πόντων αντίστοιχα. Οι Βαθμοί Bonus υπολογίζονται βάσει 32 στατιστικών του αγώνα, όπου τα γκολ που σκοράρονται, οι ασίστ και οι ανέπαφες εστίες είναι οι παράγοντες που έχουν το μεγαλύτερο βάρος.

- Για να λάβει ένας τερματοφύλακας ή αμυντικός πόντους για ανέπαφη εστία, πρέπει να παίξει τουλάχιστον 60 λεπτά, εξαιρουμένου του χρόνου καθυστερήσεων.

- Κάθε αγωνιστική εβδομάδα ο κάθε παίκτης παίρνει από το παιχνίδι 1 δωρεάν μεταγραφή. Κάθε επιπλέον μεταγραφή που πραγματοποιεί ο παίκτης κοστίζει 4 πόντους.

	Τερματοφύλακες	Αμυντικοί	Μέσοι	Επιθετικοί
Για συμμετοχή μέχρι 60 λεπτά	1	1	1	1
Για συμμετοχή 60 λεπτά και πάνω	2	2	2	2
Για κάθε γκολ	6	6	5	4
Για κάθε ασίστ	3	3	3	3
Για ανέπαφη εστία	4	4	1	-
Για κάθε 3 αποκρούσεις	1	-	-	-
Για κάθε πέναλτι που αποκρούεται	5	-	-	-
Για κάθε χαμένο πέναλτι	-2	-2	-2	-2
Για κάθε 2 γκολ που δέχεται η ομάδα	-1	-1	-	-
Για κάθε κίτρινη κάρτα	-1	-1	-1	-1
Για κάθε κόκκινη κάρτα	-3	-3	-3	-3
Για κάθε αυτογκόλ	-2	-2	-2	-2

Table 1. *Σύστημα βαθμολόγησης παικτών στο FPL*

- Το παιχνίδι δίνει στους παίκτες 4 ειδικά chips: 1.Wildcard: Δωρεάν απεριόριστες μεταγραφές για μία αγωνιστική. Μπορεί να χρησιμοποιηθεί 2 φορές τη σεζόν. 2.Free-Hit: Δωρεάν απεριόριστες μεταγραφές για μία αγωνιστική με την ομάδα σου να επιστρέφει ως ήταν την επόμενη αγωνιστική. Μπορεί να χρησιμοποιηθεί 1 φορά τη σεζόν. 3.Bench Boost: Κερδίζεις πόντους και από τους 15 παίκτες της ομάδας για μία αγωνιστική. Μπορεί να χρησιμοποιηθεί 1 φορά τη σεζόν. 4.Triple Captain: Τριπλασιάζει τους πόντους του αρχηγού για μία αγωνιστική. Μπορεί να χρησιμοποιηθεί 1 φορά τη σεζόν.

- Αν ένα γκολ σκοράρεται από φάουλ ή πέναλτι, ο παίκτης που κέρδισε το φάουλ/πέναλτι απονέμεται μια ασίστ. Επίσης, αν ένα γκολ σκοράρεται από rebound μετά από απόκρουση του αντίπαλου τερματοφύλακα, απονέμεται μια ασίστ στον παίκτη που πραγματοποίησε την αρχική προσπάθεια.

0.2.2 Πρόβλημα

Στο Fantasy Premier League, κάθε παίκτης προσπαθεί να μεγιστοποιήσει τον συνολικό αριθμό των πόντων του κατά τη διάρκεια ολόκληρης της σεζόν. Όπως φαίνεται σε αυτό το κεφάλαιο, πολλές αποφάσεις πρέπει να ληφθούν κάθε αγωνιστική εβδομάδα. Οι αποφάσεις αυτές περιλαμβάνουν ποιους παίκτες να προστεθούν στην επιλεγμένη ομάδα και ποιους παίκτες να επιλέξουν για την αρχική ενδεκάδα. Επιπλέον, πρέπει να επιλεγεί ένας αρχηγός και ένας αντι-αρχηγός. Επιπλέον, πρέπει να οριστεί μια προτεραιότητα αντικατάστασης για τους παίκτες που δεν επιλέγονται για την αρχική ενδεκάδα. Όπως έχει επισημανθεί, ένας παίκτης έχει τη δυνατότητα να κάνει μεταγραφές κατά τη διάρκεια της σεζόν. Επομένως, πρέπει επίσης να αποφασίσει εάν θα πραγματοποιήσει μια μεταγραφή και, κατά συνέπεια, ποιους παίκτες θα ανταλλάξει για κάθε αγωνιστική εβδομάδα. Έτσι, οι αποφάσεις λαμβάνονται με γνώμονα αρκετές αγωνιστικές στο μέλλον. Λαμβάνοντας υπόψη το γεγονός ότι υπάρχουν πάνω από 500 παίκτες στην Premier League κάθε σεζόν, αυτό δεν είναι εύκολο. Η αξία ενός παίκτη μπορεί επίσης να διαφέρει για διαφορετικούς παίκτες FPL, καθώς βασίζεται σε πολλούς παράγοντες (μερικές φορές και μη μετρήσιμους). Αλλά πώς μπορεί αυτό να γίνει

με αντικειμενικό τρόπο για κάθε παίκτη; Ακόμα και αν μπορούσαμε να προσδιορίσουμε την αξία κάθε παίκτη, ποια είναι η βέλτιστη στρατηγική μεταγραφών για να μεγιστοποιήσουμε τους μελλοντικούς πόντους της ομάδας μας; Πότε είναι η ιδανική στιγμή για να χρησιμοποιήσουμε τα chips; Αυτά είναι ερωτήματα που κάθε παίκτης του FPL σκέφτεται κάθε αγωνιστική εβδομάδα της σεζόν.

0.2.3 Στόχοι

Όπως αναφέρθηκε στην εισαγωγή, αυτή η εργασία έχει τους ακόλουθους 3 στόχους:

- Ο πρώτος και κύριος στόχος είναι η ανάπτυξη μοντέλων μηχανικής μάθησης που μπορούν να προβλέπουν την Αναμενόμενη Αξία για τους παίκτες της Premier League στο πλαίσιο FPL. Η αναμενόμενη αξία αντιπροσωπεύει τον αριθμό πόντων που ένας παίκτης αναμένεται να σκοράρει σε έναν συγκεκριμένο αγώνα, βασιζόμενος σε ιστορικά δεδομένα, χαρακτηριστικά του παίκτη, δύναμη αντιπάλου και άλλους σχετικούς παράγοντες. Οι ακριβείς προβλέψεις της Αναμενόμενης Αξίας είναι ζωτικής σημασίας για τους παίκτες του FPL, καθώς αποτελούν τη βάση για αποτελεσματικές στρατηγικές επιλογής παικτών και σύνθεσης ομάδας. Με την ανάπτυξη μοντέλων που μπορούν να εκτιμούν την Αναμενόμενη Αξία με υψηλή ακρίβεια, μπορούν να ληφθούν πιο ενημερωμένες αποφάσεις σχετικά με τις αποκτήσεις παικτών, τις επιλογές αρχηγείας και την συνολική στρατηγική της ομάδας.

- Ο δεύτερος στόχος είναι η πρόταση βέλτιστων κινήσεων για τους παίκτες του FPL βασισμένες στις προβλεπόμενες τιμές της Αναμενόμενης Αξίας. Οι βέλτιστες κινήσεις μπορεί να περιλαμβάνουν συστάσεις για μεταγραφές, επιλογές αρχηγείας ή προσαρμογές στη σύνθεση της ομάδας που μεγιστοποιούν τους αναμενόμενους πόντους της ομάδας.

- Ο τελικός στόχος είναι η εργασία να συμβάλει στον τομέα της ανάλυσης του FPL και να προωθήσει την εφαρμογή τεχνικών μηχανικής μάθησης στο FPL. Αυτή η διπλωματική εργασία στοχεύει να παρέχει πρακτικές εισηγήσεις, μεθοδολογίες και εργαλεία που μπορούν να χρησιμοποιηθούν από τους παίκτες του FPL για να βελτιώσουν τις διαδικασίες λήψης αποφάσεων τους. Μέσω της απόδειξης της αποτελεσματικότητας και της χρησιμότητας των μοντέλων μηχανικής μάθησης στο FPL, αυτή η έρευνα επιδιώκει να προωθήσει μια προσέγγιση που βασίζεται στα δεδομένα και να συντελέσει στην συνολική κατανόηση του παιχνιδιού.

0.3 Τα Analytics στα Αθλήματα και στο Ποδόσφαιρο

Η χρήση των στατιστικών στον αθλητισμό υπάρχει εδώ και πάνω από έναν αιώνα, αλλά ήταν τις τελευταίες δεκαετίες που πραγματικά άρχισε να συνειδητοποιείται η αξία τους και να γίνονται ολοένα και πιο δημοφιλή. Το baseball ήταν αδιαμφισβήτητα το πρώτο άθλημα που εισήγαγε την έννοια των στατιστικών στα αθλήματα από τα τέλη του 19ου αιώνα, όταν τα πρώτα στατιστικά καταγράφησαν σε αγώνες. Τις δεκαετίες 1970 και 1980, εισήχθησαν

Figure 1. *Ομάδα FPL*

και έγιναν διάσημα τα Sabermetrics, δηλαδή η πρακτική χρήσης μεθόδων στατιστικής για την ανάλυση παικτών και ομάδων baseball. Στις αρχές του 2000 ο Billy Beane και οι Oakland Athletics εξέπληξαν τον κόσμο με την επιτυχία τους σε σχέση με το μικρό τους μπάτζετ. Ο Billy Beane που ήταν ο GM της ομάδας κλήθηκε να ανταγωνιστεί ομάδες με πολλαπλάσια μπατζετ και χρησιμοποιώντας μεθόδους στατιστικής ανάλυσης κατάφερε να εντοπίσει υποτιμημένους παίκτες και να "χτίσει" μία εντυπωσιακά ανταγωνιστική ομάδα με τη γνωστή από τότε μέθοδο του Moneyball. Το βιβλίο και η ταινία που κυκλοφόρησαν για την ιστορία, έκαναν τον όρο πολύ δημοφιλή και ταυτόσημο με τη χρήση δεδομένων για την εύρεση υποτιμημένων παικτών και τη συγκρότηση ανταγωνιστικών ομάδων σε όλα τα αθλήματα και ενέπνευσαν μια ολόκληρη γενιά αναλυτών δεδομένων. Σήμερα, τα analytics αποτελούν σημαντικό παράγοντα όλων των αθλημάτων, αφού η συλλογή δεδομένων αλλά και η ανάπτυξη νέων τεχνολογικών μεθόδων συντελούν στην ανάπτυξή τους. Τα analytics σήμερα καθορίζουν πότε μία ομάδα στο αμερικάνικο ποδόσφαιρο θα παίξει κανονικά στο 4ο down, ποιά σουτ και από ποιές θέσεις θα πάρει μία ομάδα στο μπάσκετ για να έχει την πιο αποτελεσματική επίθεση ή τη στρατηγική στημένων φάσεων μίας ομάδας ποδοσφαίρου.

0.3.1 Αναμενόμενα Γκολ (xG)

Στο επίκεντρο της επανάστασης των δεδομένων στο ποδόσφαιρο βρίσκονται τα αναμενόμενα γκολ. Παρότι δεν είναι το μόνο, είναι αναμφίβολα το πιο διαδεδομένο metric στο ποδόσφαιρο και το πιο χρήσιμο για αυτήν τη διπλωματική εργασία.

Τα Αναμενόμενα Γκολ (ή xG) μετρούν την ποιότητα μιας ευκαιρίας υπολογίζοντας την πιθανότητα να μετατραπεί σε γκολ από μια συγκεκριμένη θέση στο γήπεδο κατά τη διάρκεια μιας συγκεκριμένης φάσης του παιχνιδιού. Αυτή η τιμή βασίζεται σε αρκετούς παράγοντες

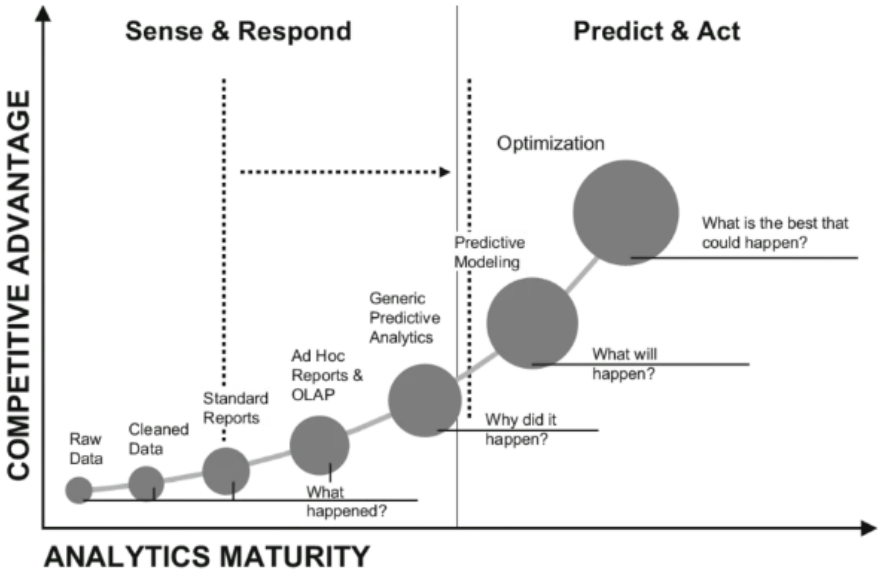


Figure 2. *Επίπεδα των Analytics [1]*

πριν από την εκτέλεση του σουτ. Το xG μετριέται σε ένα εύρος από το μηδέν έως το ένα, όπου το μηδέν αντιπροσωπεύει μια ευκαιρία που είναι αδύνατο να μετατραπεί σε γκολ και το ένα αντιπροσωπεύει μια ευκαιρία που θα αναμέναμε ένας παίκτης να σκοράρει κάθε φορά. Γνωρίζουμε ότι μια ευκαιρία από τη μέση γραμμή δεν είναι τόσο πιθανό να καταλήξει σε γκολ όσο μια ευκαιρία από το εσωτερικό της περιοχής. Με το xG, μπορούμε πραγματικά να μετρήσουμε πόσο πιθανό είναι ένας παίκτης να σκοράρει από κάθε μια από αυτές τις καταστάσεις. Για παράδειγμα, υποθέστε ότι η ευκαιρία από το εσωτερικό της περιοχής με ένα συγκεκριμένο σύνολο χαρακτηριστικών πριν το σουτ αξίζει 0,1 xG. Αυτό σημαίνει ότι ένας μέσος παίκτης θα αναμένεται να σκοράρει ένα γκολ από κάθε δέκα προσπάθειες σε αυτήν την κατάσταση. Ο όρος μπορεί να είναι νέος, αλλά αυτές οι φράσεις έχουν χρησιμοποιηθεί από φίλους του ποδοσφαίρου για πολλά χρόνια πριν εισαχθεί το xG: ¨το σκοράρει εννέα φορές στις δέκα¨ ή ¨έπρεπε να είχε κάνει χατ-τρικ¨ [8].

Κατά τη διάρκεια του παιχνιδιού, μπορούμε ενστικτωδώς να προσδιορίσουμε ποιες ευκαιρίες είναι πιο πιθανό να μετατραπούν σε γκολ. Πόσο κοντά ήταν ο παίκτης στο τέρμα; Ποιά ήταν η γωνία του σουτ; Ήταν τετ-α-τετ; Ήταν μια κεφαλιά; Το πρόβλημα είναι ότι κατά μέσο όρο υπάρχουν 25 σουτ ανά αγώνα για τα οποία πρέπει να το αξιολογήσουμε αυτό. Όλα αυτά τα σουτ μπορεί να προέρχονται από μοναδικές καταστάσεις. Το πλεονέκτημα ενός μοντέλου αναμενόμενων γκολ είναι ότι μπορούμε να πάρουμε τις μεταβλητές που αναφέρθηκαν παραπάνω - και άλλες - και να ποσοτικοποιήσουμε πώς καθεμία επηρεάζει την πιθανότητα ενός γκολ. Με αυτόν τον τρόπο, μπορούμε να αξιολογήσουμε την ποιότητα των ευκαιριών για όλα τα 9.398 σουτ που πραγματοποιήθηκαν στην Premier League τη σεζόν 2019-20 σε λίγα δευτερόλεπτα. Ας αναλύσουμε για παράδειγμα το μοντέλο xG της Stats Perform. Είναι κατασκευασμένο με χρήση ενός μοντέλου λογιστικής παλινδρόμησης που είναι τροφοδοτούμενο από εκατοντάδες χιλιάδες σουτ από τα ιστορικά δεδομένα της Opta και περιλαμβάνει αρκετές μεταβλητές που επηρεάζουν την πιθανότητα σκοραρίσματος ενός γκολ, μερικές από τις πιο σημαντικές από αυτές είναι οι εξής:

- Απόσταση από το τέρμα
- Γωνία προς το τέρμα
- Τετ-α-τετ
- Μερος σώματος που χρησιμοποιείται (π.χ. κεφάλι, πόδι)
- Τύπος πάσας που προηγήθηκε (π.χ. κάθετη πάσα, σέντρα, πάσα προς τα πίσω κτλ.)
- Κατάσταση παιχνιδιού (π.χ. κανονική ροή, αντεπίθεση, απευθείας φάουλ, κόρνερ, πλάγιο κτλ.)

Ορισμένες καταστάσεις είναι ιδιαίτερες και για αυτό μοντελοποιούνται ανεξάρτητα. Τα πέναλτι έχουν μια σταθερή τιμή που αντιστοιχεί στο συνολικό ποσοστό μετατροπής τους (0,79 xG), ενώ τα απευθείας φάουλ έχουν το δικό τους μοντέλο. Επίσης, οι ευκαιρίες με κεφαλιά αξιολογούνται διαφορετικά για στημένες φάσεις και κανονική ροή παιχνιδιού.

0.4 Θεωρία Μεθόδων Regression

Η μέθοδος regression (παλινδρόμηση) αναφέρεται ειδικά στην εκτίμηση μιας συνεχούς εξαρτημένης μεταβλητής από μια λίστα μεταβλητών εισόδου ή χαρακτηριστικών. Η παλινδρόμηση είναι μια τεχνική εκπαίδευσης με επίβλεψη που βοηθά στον εντοπισμό της συσχέτισης μεταξύ μεταβλητών και μας επιτρέπει να προβλέψουμε τη συνεχή εξαρτημένη μεταβλητή με βάση μία ή περισσότερες προβλέπουσες μεταβλητές [9].

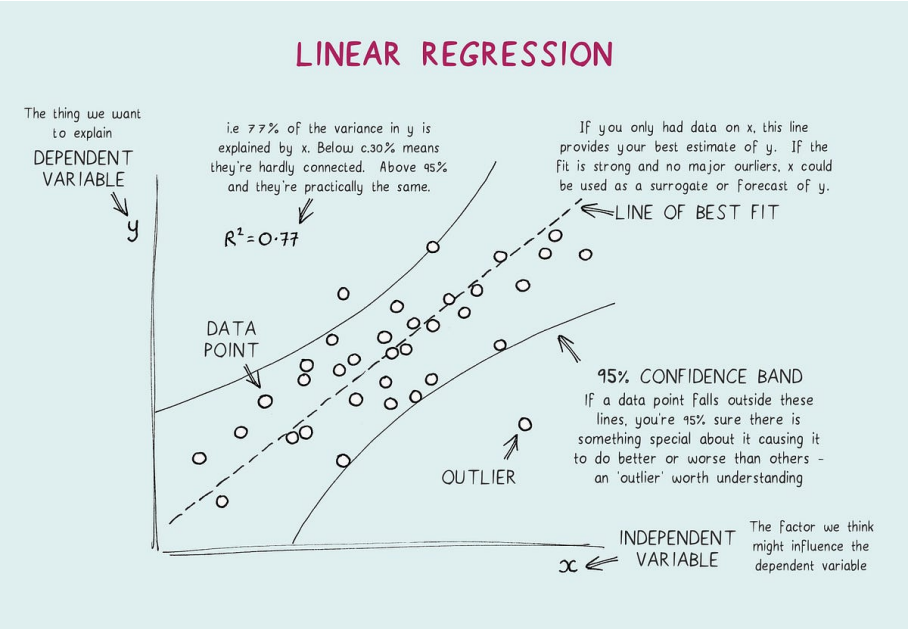


Figure 3. *Regression (Παλινδρόμηση) [2]*

0.4.1 Random Forest Regression

Η μέθοδος Random Forest Regression είναι ένας αλγόριθμος μηχανικής μάθησης με επίβλεψη που χρησιμοποιεί ensemble μέθοδο για την παλινδρόμηση. Είναι μία τεχνική bagging (bootstrap aggregating). Τα δέντρα στον αλγόριθμο λειτουργούν παράλληλα, πράγμα που σημαίνει ότι δεν υπάρχει αλληλεπίδραση μεταξύ τους κατά τη διάρκεια της κατασκευής τους [Σχήμα 4] [13].

Η μέθοδος bootstrap aggregating, γνωστή και ως bagging (από το bootstrap aggregating), είναι ένα μετα-αλγόριθμος στη μηχανική μάθηση που έχει σχεδιαστεί για να βελτιώσει τη σταθερότητα και την ακρίβεια των αλγορίθμων μηχανικής μάθησης που χρησιμοποιούνται στη στατιστική ταξινόμηση και παλινδρόμηση. Επίσης, μειώνει τη διακύμανση και βοηθά στην αποφυγή της υπερ-εκπαίδευσης [14].

Δεδομένου ενός κανονικού συνόλου εκπαίδευσης D μεγέθους n, η μέθοδος bagging παράγει m νέα σύνολα εκπαίδευσης D_i, καθένα με μέγεθος n', εφαρμόζοντας δειγματοληψία με αντικατάσταση από το D. Με τη δειγματοληψία αυτή, ορισμένες παρατηρήσεις μπορεί να επαναλαμβάνονται σε κάθε D_i. Εάν $n' = n$, τότε για μεγάλο n το σύνολο D_i αναμένεται να έχει το κλάσμα $(1 - 1/e)$ (περίπου 63.2%) των μοναδικών παραδειγμάτων του D, με τα υπόλοιπα να είναι αντίγραφα. Αυτού του είδους δειγματοληψία ονομάζεται δειγματοληψία bootstrap. Η δειγματοληψία με αντικατάσταση εξασφαλίζει την ανεξαρτησία κάθε bootstrap από τα άλλα, καθώς δεν εξαρτάται από προηγούμενα επιλεγμένα δείγματα. Έπειτα, γίνεται η εκπαίδευση m μοντέλων χρησιμοποιώντας τα παραπάνω m bootstrap σύνολα και συνδυάζονται με τον υπολογισμό του μέσου όρου των εξόδων (για παλινδρόμηση) ή της ψήφου (για ταξινόμηση). Επιπλέον, εκτός από το ότι κάθε δέντρο εξετάζει μόνο ένα bootstrapped σύνολο δειγμάτων, μόνο ένα μικρό αλλά σταθερό πλήθος μοναδικών χαρακτηριστικών λαμβάνεται υπόψη κατά την κατάταξή τους ως προβλέποντες μεταβλητές. Αυτό σημαίνει ότι κάθε δέντρο γνωρίζει μόνο για τα δεδομένα που αφορούν ένα μικρό σταθερό αριθμό χαρακτηριστικών και ένα μεταβλητό αριθμό δειγμάτων που είναι μικρότερος ή ίσος του αρχικού συνόλου δεδομένων. Ως εκ τούτου, τα δέντρα έχουν μεγαλύτερη πιθανότητα να παρέχουν πιο ποικίλες απαντήσεις, προερχόμενες από ποικίλες γνώσεις. Αυτό οδηγεί σε πλεονεκτήματα για ένα τυχαίο δάσος (random forest) σε σύγκριση με ένα μόνο δέντρο απόφασης που δημιουργείται χωρίς τυχαιότητα [14].

0.4.2 XGBoost Regression

Το XGBoost (eXtreme Gradient Boosting) είναι μια βιβλιοθήκη λογισμικού ανοιχτού κώδικα που παρέχει ένα πλαίσιο εκπαίδευσης με τη χρήση αλγορίθμων gradient boosting [15]. Στα Boosted Trees χρησιμοποιούνται σύνολα δέντρων, όπως και στα Random Forests. Η διαφορά ανάμεσα στις δύο μεθόδους βρίσκεται στη φάση της εκπαίδευσης. Ενώ τα Random Forests χρησιμοποιούν την τεχνική του bagging που αναλύθηκε παραπάνω, το XGBoost και όλα τα μοντέλα με boosted trees χρησιμοποιούν την τεχνική του boosting [16].

Η τεχνικη boosting είναι ένας ensemble μετα-αλγόριθμος για τη μείωση κυρίως της προκατάληψης (bias) και επίσης της διακύμανσης στην επιβλεπόμενη μάθηση, και μια οικογένεια αλγορίθμων μηχανικής μάθησης που μετατρέπουν αδύναμους μαθητές σε ισχυρούς [17].

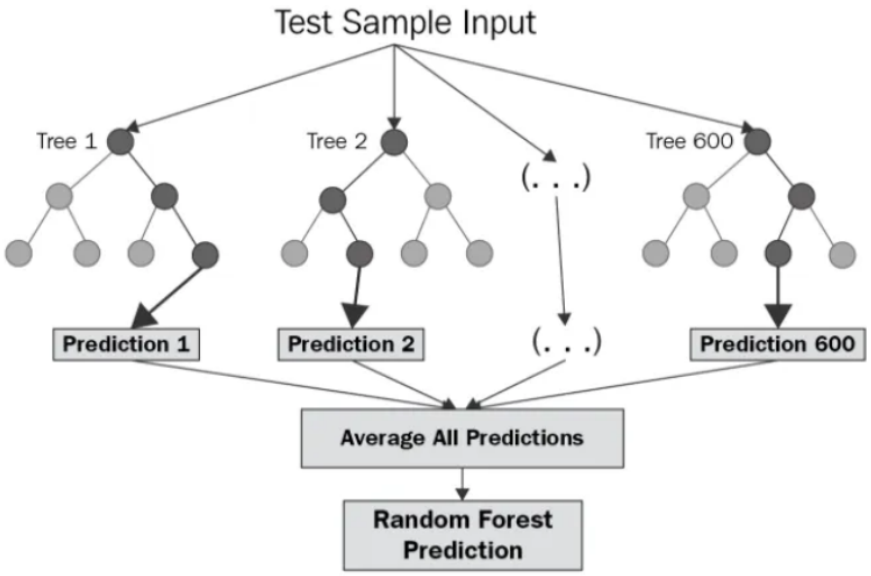


Figure 4. *Random Forest Regression [3]*

Οι περισσότεροι αλγόριθμοι boosting αποτελούνται από την επαναληπτική εκμάθηση αδύναμων ταξινομητών ως προς μια κατανομή και τον συνδυασμό τους σε έναν τελικό ισχυρό ταξινομητή. Όταν προστίθενται, οι ταξινομητές αξιολογούνται βάσει ενός συντελεστή με βάρη που σχετίζονται με την επιτυχία των αδύναμων μαθητών. Μετά την προσθήκη ενός αδύναμου μαθητή, τα βάρη των δεδομένων επαναρυθμίζονται. Τα εισερχόμενα δεδομένα που προβλέπονται λάθος αποκτούν υψηλότερο βάρος και τα παραδείγματα που προβλέφθηκαν σωστά χάνουν βάρος. Έτσι, οι μελλοντικοί αδύναμοι μαθητές επικεντρώνονται στα παραδείγματα που οι προηγούμενοι αδύναμοι μαθητές προέβλεψαν με λανθασμένο τρόπο [17]. Με αυτόν τον τρόπο, οι αδύναμοι μαθητές επικεντρώνονται όλο και περισσότερο στα παραδείγματα που οι προηγούμενοι αδύναμοι μαθητές προβλέπουν λανθασμένα, επιτρέποντας έτσι τη δημιουργία ισχυρών μαθητών.

Συνολικά, ο αλγόριθμος boosting επιτρέπει τη συνεχή βελτίωση της απόδοσης του μοντέλου, μετατρέποντας αδύναμους μαθητές σε ισχυρούς και αξιοποιώντας την πληροφορία από τα λάθη των προηγούμενων μαθητών για να βελτιώσει τις προβλέψεις του. Το XGBoost είναι ένα παράδειγμα αλγορίθμου boosting που χρησιμοποιείται ευρέως για την επίλυση προβλημάτων παλινδρόμησης και ταξινόμησης σε διάφορους τομείς της μηχανικής μάθησης.

0.4.3 Ridge Regression

Η Ridge regression είναι μια μέθοδος εκτίμησης των συντελεστών πολλαπλών γραμμικών μοντέλων σε περιπτώσεις όπου οι ανεξάρτητες μεταβλητές είναι υψηλά συσχετισμένες. Έχει χρησιμοποιηθεί σε πολλούς τομείς, όπως η οικονομετρία, η χημεία και η μηχανική. Είναι ιδιαίτερα χρήσιμη για την αντιμετώπιση του προβλήματος της πολλαπλής συσχέτισης (multicollinearity) στη γραμμική παλινδρόμηση, το οποίο συμβαίνει συχνά σε μοντέλα με μεγάλο αριθμό παραμέτρων. Η μέθοδος Ridge Regression αναπτύχθηκε ως μια πιθανή λύση στην

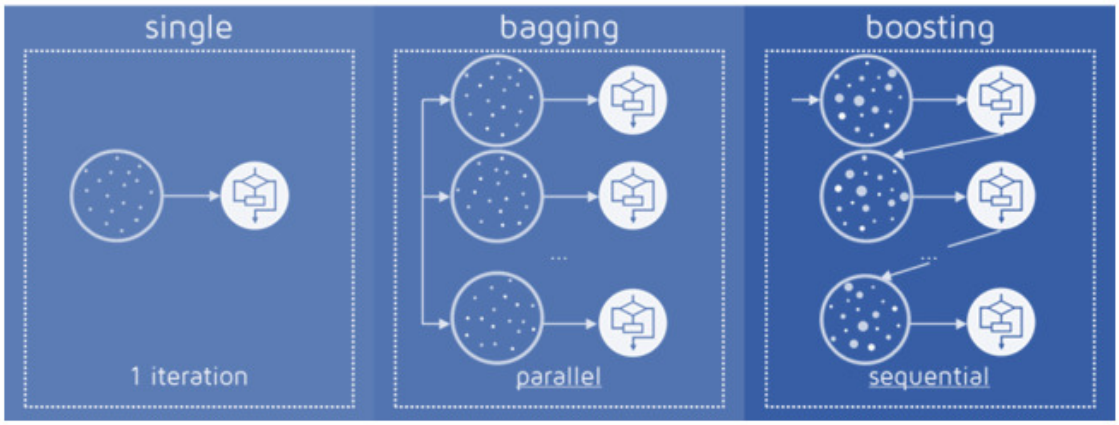


Figure 5. *Bagging vs Boosting [4]*

ανακρίβεια των εκτιμητών των ελαχίστων τετραγώνων όταν τα γραμμικά μοντέλα παλινδρόμησης έχουν μερικές πολλαπλά συσχετισμένες ανεξάρτητες μεταβλητές - δημιουργώντας έναν εκτιμητή Ridge Regression (RR). Αυτό παρέχει μια πιο ακριβή εκτίμηση των παραμέτρων του ridge, καθώς οι διακυμάνσεις του και οι τετραγωνικές μέσες εκτιμήσεις του συνήθως είναι μικρότερες από τις εκτιμήσεις ελαχίστων τετραγώνων που έχουν προηγουμένως προκύψει [18]. Ίσως ο πιο απλός αλγόριθμος που μπορεί να γίνει kernelized είναι ο Ridge Regression. Αυτό οδηγεί στον Kernel Ridge Regression [19].

0.5 Σχετικές Εργασίες

Η γενική τάση που παρατηρείται στις προηγούμενες εργασίες είναι η χρήση κυρίως ιστορικών στατιστικών δεδομένων σε συνδυασμό με αλγορίθμους Naive-Bayes, Random Forests, Boosting Machines και ensemble μεθόδους για την πρόβλεψη των μελλοντικών επιδόσεων των παικτών.

Χρησιμοποιώντας έναν αλγόριθμο Gaussian Naive Bayes, ο Thapliya κατάφερε να προβλέψει τις μελλοντικές επιδόσεις των παικτών με αναφερόμενη ακρίβεια 86% [20]. Κατέταξε τα δεδομένα σε δύο κατηγορίες: τους παίκτες που σημείωσαν 8 ή περισσότερους πόντους και αυτούς με λιγότερους από 8.

Χρησιμοποιώντας παρόμοια σύνολα εκπαίδευσης, ο Raghunandh διαπίστωσε ότι η χρήση ensemble μεθόδων όπως οι Gradient Boosting Machines (GBM) μπορεί να βοηθήσει στην πρόβλεψη πιθανών επιλογών για αρχηγεία [21].

Στο άρθρο τους, οι Bonello, Beel, Lawless και Debattista επιδίωξαν να εξετάσουν εάν η συνδυαστική χρήση δεδομένων από διάφορες πηγές θα βοηθούσε στην αύξηση της ακρίβειας και της συνολικής απόδοσης της πρόβλεψης. Παρά την υποσχόμενη αυτή ιδέα, αναφέρεται ότι αντιμετώπισαν δυσκολίες στην ενσωμάτωση δεδομένων κειμένου από τα κοινωνικά δίκτυα κατά την υλοποίηση των μοντέλων επεξεργασίας φυσικής γλώσσας (NLP). Παρόλα αυτά, κατάφεραν να επιτύχουν μια πολύ καλή κατάταξη της τάξης των 30.000 από τους 6,5 εκατομμύρια παίκτες και διαπίστωσαν ότι οι GBMs είναι οι πιο αποτελεσματικές σε αυτήν την περίπτωση, ειδικά λόγω των μη ισορροπημένων δεδομένων [22].

Ο Bonomo et Al. ανέπτυξαν ένα μαθηματικό μοντέλο βελτιστοποίησης χρησιμοποιώντας

γραμμικό προγραμματισμό για να προβλέψει ιδανικές ενδεκάδες σε κάθε αγωνιστική στην Αργεντίνικη Ποδοσφαιρική Λίγκα. Χρησιμοποίηθηκαν ιστορικά δεδομένα σε συνδυασμό με πληροφορίες από τις συνεντεύξεις των προπονητών πριν από τους αγώνες [23].

Οι Pokharel, Timalsina, Panday και Acharya χρησιμοποίησαν XGBoost regression και επικεντρώθηκαν στο ROI για να προβλέψουν την επίδοση των παικτών, εξετάζοντας επίσης την επίδραση πρόσθετων δεδομένων από αγώνες κυπέλλου που πραγματοποιούνταν μεσοβδόμαδα. Κατάφεραν ένα μέσο RMSE σκορ 2,048 για όλους τους παίκτες [24].

Σε πιο εξελιγμένα μοντέλα, οι Matthews, Ramchurn και Chalkiadakis παρουσίασαν ένα καινοτόμο προγνωστικό μοντέλο για το fantasy football, το οποίο αποτελείται από belief-state MDP μοντέλα σε συνδυασμό με αλγόριθμους Bayesian Q-Learning για την εκπαίδευση μοντέλων με τα τελευταία πέντε χρόνια δεδομένων ποδοσφαίρου [25]. Το πιο επιτυχημένο μοντέλο τους χρησιμοποίησε ένα εξελιγμένο μοντέλο Bayesian Q-Learning για να αντιμετωπίσει την αβεβαιότητα, τοποθετώντας το στους κορυφαίους 500 παίκτες από τους 2,5 εκατομμύριους. Το αρχικό μοντέλο ήταν αφελές, λειτουργούσε μυωπικά, λαμβάνοντας υπόψη μόνο τον επόμενο αγώνα. Παρά τους περιορισμούς αυτούς, το μοντέλο αυτό κατάφερε να επιτύχει ένα πολύ σεβαστό επίπεδο κατάταξης στη θέση 113,921. Με την επέκταση αυτού του μοντέλου για να εξετάσει επίσης δεδομένα από την προηγούμενη σεζόν, το μοντέλο βελτιώνει την κατάταξη στη θέση 60,633.

Τα τελευταία χρόνια, η κοινότητα του FPL έχει δημιουργήσει πολλά αξιέπαινα έργα στον τομέα. Ένα από αυτά είναι αδιαμφισβήτητα το μοντέλο του FPL Kiwi [26]. Χρησιμοποιεί αναμενόμενα δεδομένα (αναμενόμενα γκολ, αναμενόμενες ασίστ κλπ) από τις προηγούμενες 5 σεζόν και διαχωρίζει το πρόβλημα σε μικρότερα, με κάθε μικρότερο πρόβλημα να αντιστοιχεί στην πρόβλεψη μιας ξεχωριστής κατηγορίας που οδηγεί σε πόντους. Χρησιμοποιεί spreadsheets για την υλοποίηση των λεπτομερών υπολογισμών του, έχει το δικό του μοντέλο για τον υπολογισμό της ισχύος των ομάδων και ανεβάζει τις προβλέψεις του για να βοηθήσει την κοινότητα [27].

Πολλές ιστοσελίδες σχετικές με το παιχνίδι έχουν επίσης αναπτύξει πρόσφατα τα δικά τους μοντέλα. Το FPL Review [28], το Fantasy Football Scout [29], το Fantasy Football Hub [30], το Fantasy Football Fix [31] είναι όλα παραδείγματα της πρόσφατης έκρηξης δεδομένων στο παιχνίδι του FPL.

0.6 Δεδομένα

Το συνολικό διάγραμμα ροής των δεδομένων της εργασίας φαίνεται στο Σχήμα 7

0.6.1 Σημαντική Απόφαση

Η πρόβλεψη της αξίας ενός παίκτη για τις επόμενες αγωνιστικές στο FPL είναι ένα πρόβλημα για το οποίο δεν υπάρχει έτοιμο σύνολο δεδομένων για να χρησιμοποιήσουμε. Ωστόσο, η εμπειρία μας από το άθλημα και το παιχνίδι μας οδηγεί σε μετρικές που μπορούν να χρησιμοποιηθούν για να προβλέψουν αυτήν την αξία. Έτσι, έπρεπε να δημιουργήσω τα δικά μου σύνολα δεδομένων και να πειραματιστώ με διάφορα σύνολα χαρακτηριστικών προκειμένου να επιτύχω τα καλύτερα δυνατά αποτελέσματα. Μετά από προσωπικά πειράματα, συζητήσεις

με μέλη της κοινότητας του FPL και μελέτη παλαιότερων σχετικών εργασιών, αποφάσισα ότι είχε νόημα να χωρίσω το πρόβλημα της πρόβλεψης της αξίας κάθε παίκτη σε μικρότερα υ-ποπροβλήματα [Σχήμα 6]. Για κάθε ενέργεια/γεγονός στον αγώνα που παράγει πόντους για έναν παίκτη, υπάρχει ένα διαφορετικό υποπρόβλημα που πρέπει να επιλυθεί. Τα κυριότερα που απαιτούν δικά τους μοντέλα είναι τα ακόλουθα:

- Μη-πέναλτι Γκολ
- Ασίστ
- Γκολ Ομάδας
- Πέναλτι
- Αποκρούσεις
- Μπόνους Πόντοι

Τα μη-πέναλτι γκολ και τα πέναλτι, που προσθέτουν αξία μόνο στους εκτελεστές πέναλτι κάθε ομάδας, αποτελούν το στοιχείο των γκολ για την τελική αξία του παίκτη. Αυτό ισχύει για όλους τους παίκτες (με τους τερματοφύλακες να είναι οι λιγότερο πιθανοί να σκοράρουν). Τα γκολ της ομάδας, για παράδειγμα, χρησιμοποιούνται για τον υπολογισμό της πιθανότητας για ανέπαφες εστίες (χρησιμοποιώντας την διαδικασία του Poisson [32]), που δίνουν πόντους σε συγκεκριμένους παίκτες ανάλογα με τη θέση τους. Οι επεμβάσεις χρησιμοποιούνται για τον υπολογισμό των πόντων αποκρούσεων μόνο για τους τερματοφύλακες, ενώ οι ασίστ και οι πόντοι μπόνους είναι για κάθε παίκτη.

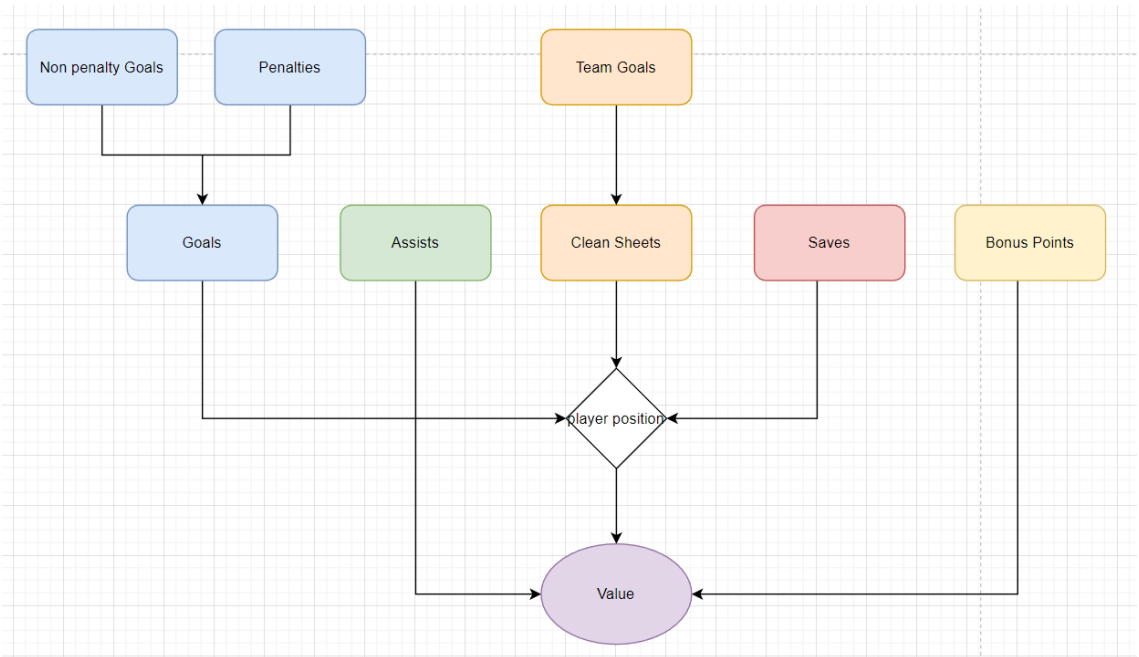


Figure 6. *Τα Στοιχεία της Τελικής Αξίας κάθε Παίκτη*

0.6.2 Πηγές Δεδομένων

Παρόλο που δεν ήταν διαθέσιμο ένα συγκεκριμένο σύνολο δεδομένων για να επιτύχω τον στόχο μου, τα ακατέργαστα δεδομένα δεν ήταν δύσκολο να βρεθούν. Εδώ, θα παρουσιάσω τις πηγές από τις οποίες συλλέχθηκαν τα ακατέργαστα δεδομένα.

- **Το API του Fantasy Premier League:** Αυτό είναι το API του επίσημου παιχνιδιού. Παρέχει εξαιρετικά χρήσιμα δεδομένα για την εργασία, όπως δεδομένα αγώνων, δεδομένα παικτών από προηγούμενες αγωνιστικές, ακόμα και στατιστικές πληροφορίες για τις προηγούμενες σεζόν των παικτών. Εκτός αυτού, μπορούμε να κάνουμε αιτήσεις στο επίσημο API για να λάβουμε δεδομένα για έναν συγκεκριμένο παίκτη του FPL ή πληροφορίες σχετικά με τις λίγκες στις οποίες συμμετέχει [33].

- **Understat:** Το Understat είναι ένας ιστότοπος που παρέχει λεπτομερή στατιστικά για τα xG (αναμενόμενα γκολ) στις κορυφαίες ευρωπαϊκές λίγκες. Ο στόχος τους ήταν να δημιουργήσουν την πιο ακριβή μέθοδο για την αξιολόγηση της ποιότητας των σουτ. Για το σκοπό αυτό, εκπαίδευσαν αλγόριθμους πρόβλεψης με νευρωνικά δίκτυα χρησιμοποιώντας ένα μεγάλο σύνολο δεδομένων (πάνω από 100.000 σουτ, πάνω από 10 παράμετροι για κάθε ένα [34]). Χρησιμοποίησα το Understat για γρήγορα και ενημερωμένα δεδομένα xG για την Premier League. Επιπλέον, το Understat παρείχε δεδομένα όπως μη-πέναλτι αναμενόμενα γκολ, σουτ και πάσες-κλειδιά ανά αγώνα για κάθε παίκτη, τα οποία βρήκα ιδιαίτερα χρήσιμα για τον στόχο μου. Χρησιμοποιώ το πακέτο understat [35] για να έχω πρόσβαση στα δεδομένα.

- **Το GitHub Repository του Vaastav:** Το API του Fantasy Premier League παρέχει δεδομένα μόνο για την τρέχουσα σεζόν. Αυτό θα καθιστούσε αδύνατη τη συλλογή δεδομένων από προηγούμενες σεζόν για την εκπαίδευση των μοντέλων μας. Το repository του Vaastav λύνει αυτό το πρόβλημα παρέχοντας δεδομένα από προηγούμενες σεζόν από το FPL API και από το Understat [36]. Αυτό το repository ήταν εξαιρετικά σημαντικό για την εργασία, καθώς αποτέλεσε τη βάση για όλα τα δημιουργηθέντα σύνολα δεδομένων.

- **Ο Χάρτης των IDs του Chris Musson:** Το API του Fantasy Premier League δεν έχει το ίδιο αναγνωριστικό για τον ίδιο παίκτη με το API του Understat, ούτε και το ίδιο όνομα σε πολλές περιπτώσεις. Σε μια νέα σεζόν του FPL, το αναγνωριστικό ενός παίκτη δεν παραμένει απαραίτητα το ίδιο, προκαλώντας προφανή προβλήματα αντιστοίχισης όταν προσπαθούμε να συνδυάσουμε το FPL API με το Understat για προηγούμενες σεζόν. Το repository του Chris Musson λύνει το πρόβλημα της αντιστοίχισης των αναγνωριστικών, καθώς έχει αντιστοιχίσει τα αναγνωριστικά του FPL με τα αναγνωριστικά του Understat για κάθε παίκτη για αρκετές προηγούμενες σεζόν [37].

- **FiveThirtyEight:** Επίσης, χρησιμοποιήθηκαν οι βαθμολογίες SPI (Soccer Power Index) και οι μετρικές των προβλεπόμενων γκολ από το FiveThirtyEight. Αυτές οι μετρικές χρησιμοποιήθηκαν σε διάφορα μοντέλα για να παρέχουν πληροφορία για τη δύναμη μιας ομάδας σε ένα συγκεκριμένο χρονικό σημείο σε σχέση με τον αντίπαλο. Το σύνολο δεδομένων Club Soccer Predictions του FiveThirtyEight χρησιμοποιήθηκε

για να ενσωματώσει αυτές τις μετρικές [38]. Στην αρχή μιας σεζόν, το SPI μιας ομάδας επηρεάζεται από το SPI με το οποίο ολοκλήρωσε την προηγούμενη σεζόν και την αξία της ομάδας στην αγορά. Κατά τη διάρκεια μιας σεζόν, το SPI μιας ομάδας αλλάζει με βάση τις επιδόσεις της ομάδας, και πιο συγκεκριμένα βασίζεται στα προσαρμοσμένα γκολ, στα αναμενόμενα γκολ με βάση τα σουτ και στα αναμενόμενα γκολ όχι με βάση τα σουτ μετά από κάθε παιχνίδι [39]. Οι βαθμολογίες SPI βοηθούν στον υπολογισμό των προβλεπόμενων γκολ για κάθε παιχνίδι και χρησιμοποιώντας τη διαδικασία Poisson [32], υπολογίζουν την πιθανότητα μιας ομάδας να νικήσει.

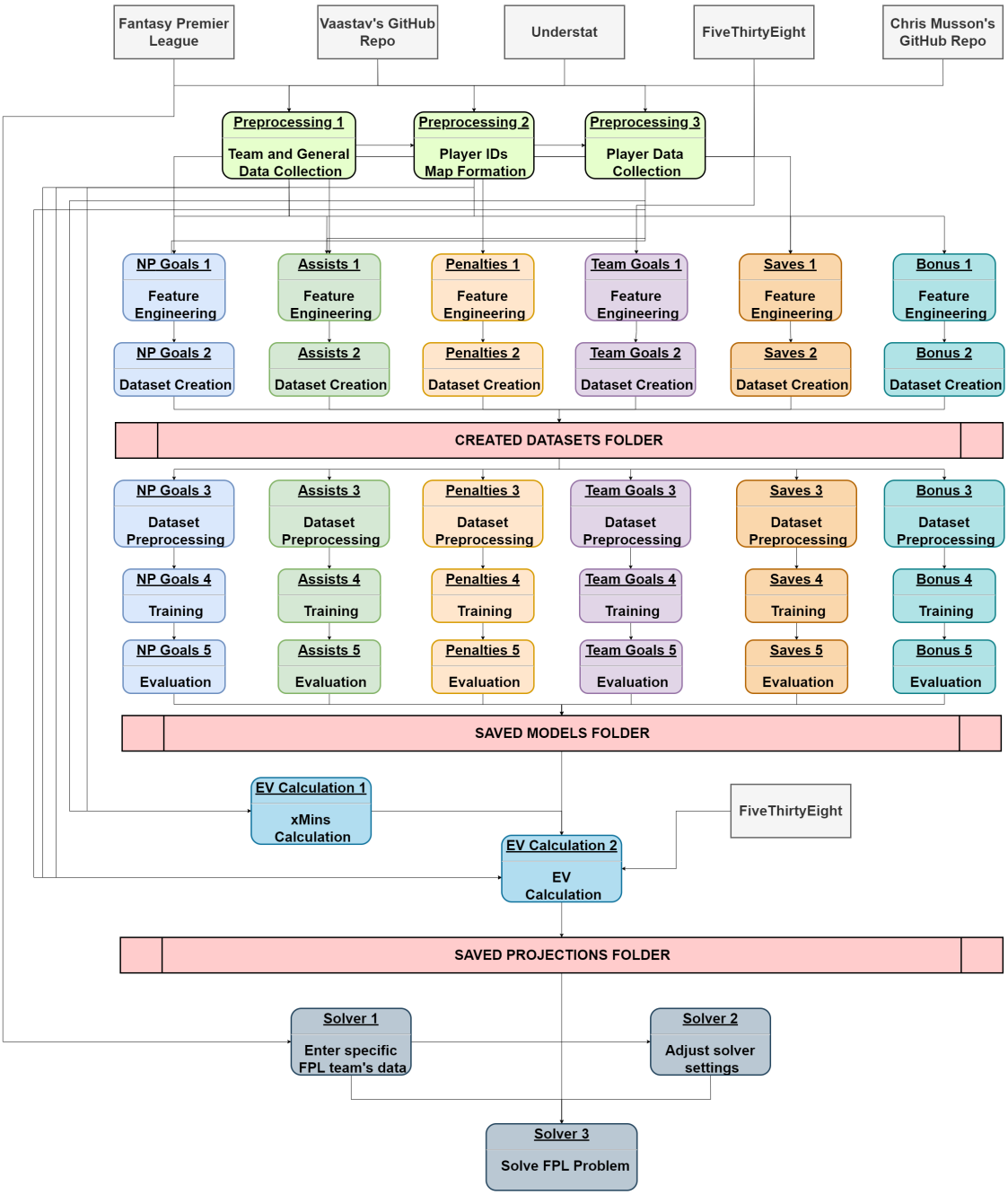


Figure 7. *Διάγραμμα Ροής Δεδομένων της Εργασίας*

Τα δημιουργηθέντα σύνολα δεδομένων αναλύονται παρακάτω στην εργασία και μπορούν

να βρεθούν στο GitHub μου [40].

0.7 Εκπαίδευση και Αξιολόγηση των Μοντέλων

Σε αυτήν την ενότητα, θα εξερευνήσουμε τις τεχνικές που χρησιμοποιήθηκαν και τις αποφάσεις που πάρθηκαν κατά τη φάση της εκπαίδευσης των μοντέλων. Θα αναλύσουμε τις μετρικές με τις οποίες αξιολογήθηκαν τα μοντέλα και θα παρατηρήσουμε αυτές τις αξιολογήσεις.

0.7.1 Μοντέλα και Μετρικές Αξιολόγησης

Επειδή τα δεδομένα για το συγκεκριμένο πρόβλημα τείνουν να είναι πολύ θορυβώδη, χρειαζόμαστε πολύ ανθεκτικά/κανονικοποιημένα μοντέλα που αποφεύγουν το overfitting. Τα μοντέλα που επιλέχθηκαν για εκπαίδευση είναι τα εξής:

- **Random Forest Regressor**
- **Kernel Ridge Regressor**
- **XGBoost Regressor**

Αυτά τα μοντέλα επέδειξαν τις καλύτερες επιδόσεις σύμφωνα με παρόμοιες εργασίες που εξετάσαμε κατά τη φάση της έρευνας στη βιβλιογραφία (βλ. Κεφάλαιο 0.5: Σχετικές Εργασίες). Οι μετρικές με τις οποίες θα αξιολογηθούν τα μοντέλα μας είναι οι εξής:

- **Mean Absolute Error (MAE):** Είναι ο μέσος όρος των απόλυτων διαφορών ανάμεσα στην πραγματική τιμή και την προβλεπόμενη τιμή του μοντέλου.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|$$

  όπου, N = συνολικός αριθμός δεδομένων, y_i = πραγματική τιμή, y_i' = προβλεπόμενη τιμή.

- **Root Mean Square Error (RMSE):** Είναι η ρίζα του μέσου όρου των τετραγωνικών διαφορών ανάμεσα στην πραγματική τιμή και την προβλεπόμενη τιμή του μοντέλου. Όσο χαμηλότερη είναι η τιμή, τόσο καλύτερο είναι το μοντέλο παλινδρόμησης [41].

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2}$$

  όπου, N = συνολικός αριθμός δεδομένων, y_i = πραγματική τιμή, y_i' = προβλεπόμενη τιμή.

- **R^2 σκορ:** Το R^2 εξηγεί σε ποιο βαθμό η διακύμανση μιας μεταβλητής εξηγεί τη διακύμανση της δεύτερης μεταβλητής. Με άλλα λόγια, μετρά το ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής που εξηγείται από τις ανεξάρτητες μεταβλητές.

$$R^2 = 1 - \frac{SSE}{SST}$$

όπου SSE είναι το άθροισμα των τετραγώνων των διαφορών ανάμεσα στην πραγματική τιμή και την προβλεπόμενη τιμή.

$$SSE = \sum_{i=1}^{N}(y_i - y_i')^2$$

και SST είναι το συνολικό άθροισμα των τετραγώνων των διαφορών ανάμεσα στην πραγματική τιμή και το μέσο της πραγματικής τιμής.

$$SST = \sum_{i=1}^{N}(y_i - \overline{y})^2$$

Εδώ, y_i είναι η παρατηρούμενη τιμή στόχου, y_i' είναι η προβλεπόμενη τιμή, $\overline{y}$ είναι η μέση τιμή, και το N αναπαριστά το συνολικό αριθμό παρατηρήσεων.

0.7.2 Εκπαίδευση

Το σύνολο εκπαίδευσης περιλαμβάνει δεδομένα από τις τρεις προηγούμενες σεζόν (2019/20 - 2021/22), ενώ το σύνολο ελέγχου περιλαμβάνει δεδομένα από την τρέχουσα σεζόν (2022/23) μέχρι την 25η αγωνιστική εβδομάδα. Αυτή είναι μια κατανομή 80%-20% για τα δεδομένα μας. Χρησιμοποιώντας την τρέχουσα σεζόν ως σύνολο ελέγχου, μας επιτρέπει να αξιολογήσουμε όλα τα μοντέλα μας στην ίδια σεζόν, ανεξάρτητα από το αν επικεντρώνονται στην απόδοση των παικτών (όπως το μοντέλο για τα γκολ χωρίς πέναλτι) ή στην απόδοση της ομάδας (όπως το μοντέλο για τα γκολ της ομάδας).

Για τη βελτιστοποίηση των υπερπαραμέτρων των μοντέλων, χρησιμοποιήθηκαν οι μέθοδοι Grid Search Cross Validation και Randomized Search Cross Validation (με n_iter=10) με τα ακόλουθα εύρη τιμών:

- n_estimators: 200 - 1200
- max_depth: 2 - 7
- learning_rate: 0.006 - 0.016
- alpha: 0.5 - 1.5

Έτσι βρέθηκαν οι βέλτιστες παράμετροι για κάθε μοντέλο.

0.7.3 Αξιολόγηση

Τα αποτελέσματα όσον αφορά το MAE των μοντέλων μας παρουσιάζονται στον παρακάτω πίνακα.

Παρατηρούμε ότι τα μοντέλα Random Forest καταλαμβάνουν τις τρεις πρώτες θέσεις ανάμεσα στα έξι μοντέλα, ενώ τα μοντέλα XGBoost κατακτούν την πρώτη θέση στα 2 από τα 3 υπόλοιπα μοντέλα, και το Kernel Ridge είναι πρώτο στο μοντέλο των Γκολ Ομάδας. Ωστόσο, οι διαφορές μεταξύ των μοντέλων Random Forest και XGBoost είναι ελάχιστες και το κυρίαρχο μοντέλο μπορεί να αλλάξει όταν δίνουμε μεγαλύτερη έμφαση στις ακραίες περιπτώσεις.

	Random Forest	XGBoost	Kernel Ridge
Μη-πέναλτι Γκολ	0.09378	**0.09366**	0.11581
Ασίστ	**0.10092**	0.10484	0.11247
Πέναλτι	0.23667	**0.20918**	0.21616
Γκολ Ομάδας	0.94811	0.94030	**0.93470**
Αποκρούσεις	**1.00484**	1.02365	1.17936
Μπόνους Πόντοι	**0.24714**	0.25096	0.28834

Table 2. *Αξιολόγηση MAE*

Τα αποτελέσματα όσον αφορα το RMSE των μοντέλων μας παρουσιάζονται στον παρακάτω πίνακα.

	Random Forest	XGBoost	Kernel Ridge
Μη-πέναλτι Γκολ	0.24316	**0.23885**	0.24850
Ασίστ	0.23798	**0.23596**	0.24017
Πέναλτι	0.31277	**0.30599**	0.31374
Γκολ Ομάδας	1.19468	1.19524	**1.18542**
Αποκρούσεις	1.54129	**1.53651**	1.59737
Μπόνους Πόντοι	0.55274	**0.55206**	0.56428

Table 3. *Αξιολόγηση RMSE*

Παρατηρούμε ότι τα μοντέλα XGBoost κυριαρχούν εκτός από το μοντέλο Γκολ Ομάδας, όπου το Kernel Ridge εξακολουθεί να βρίσκεται στην κορυφή. Είναι ενδιαφέρον ότι τα μοντέλα XGBoost είναι καλύτερα από τα μοντέλα Random Forest σε κάθε κατηγορία, κρίνοντας από το RMSE, ενώ τα μοντέλα Random Forest φαίνονταν καλύτερα κατά την αξιολόγηση με βάση το MAE. Αυτή η αλλαγή υποδηλώνει ότι τα μοντέλα XGBoost τείνουν να είναι πιο ακριβή σχετικά με τις περιπτώσεις με υψηλή διακύμανση, αφού το σκορ RMSE τιμωρεί περισσότερο τα μεγάλα σφάλματα. Ωστόσο, είναι σημαντικό να σημειωθεί ότι τα μοντέλα Random Forest παρουσιάζουν μόνο ελαφρώς χειρότερη απόδοση και εξακολουθούν να είναι αρκετά καλές λύσεις.

Τα αποτελέσματα του σκορ R^2 των μοντέλων μας παρουσιάζονται στον παρακάτω πίνακα.

	Random Forest	XGBoost	Kernel Ridge
Μη πέναλτι Γκολ	0.11814	**0.14911**	0.07900
Ασίστ	0.07706	**0.08808**	0.05996
Πέναλτι	-0.01321	**0.03020**	-0.01951
Γκολ Ομάδας	0.14130	0.14049	**0.15455**
Αποκρούσεις	0.46550	**0.46881**	0.42590
Μπόνους Πόντοι	0.12817	**0.13029**	0.09138

Table 4. *Αξιολόγηση R^2 σκορ*

Εδώ, παρατηρούμε παρόμοια αποτελέσματα με αυτά που αναφέρθηκαν παραπάνω στον πίνακα RMSE. Τα μοντέλα XGBoost κυριαρχούν σε όλες τις κατηγορίες, εκτός από τα Γκολ Ομάδας όπου το Kernel Ridge είναι πρώτο.

0.8 Τελικά Στάδια και Αποτελέσματα

0.8.1 Υπολογισμός Αναμενόμενης Αξίας

Αφού έχουμε επιλέξει τα μοντέλα με την καλύτερη απόδοση για κάθε ενέργεια στο γήπεδο που οδηγεί σε πόντους, έχει θεμελιωθεί η βάση της προβλεπτικής διαδικασίας. Το μόνο που έχουμε να κάνουμε τώρα είναι να συλλέξουμε κάθε τρέχον δεδομένο για κάθε παίκτη που απαιτείται από κάθε μοντέλο. Στη συνέχεια, τροφοδοτούμε αυτά τα δεδομένα στην συνάρτηση που υπολογίζει τους αναμενόμενους πόντους για έναν παίκτη για μια συγκεκριμένη αγωνιστική εβδομάδα και ακολουθούμε αυτήν την επαναληπτική διαδικασία για κάθε παίκτη και κάθε αγωνιστική εβδομάδα μέχρι τον επιλεγμένο ορίζοντα (την αγωνιστική εβδομάδα μέχρι την οποία θα γίνουν οι υπολογισμοί των αναμενόμενων πόντων).

Ένα πραγματικά σημαντικό μέρος της διαδικασίας υπολογισμού των αναμενόμενων πόντων (EV) είναι αναμφίβολα ο υπολογισμός των xMins (αναμενόμενα λεπτά). Αυτή η διαδικασία αναφέρεται στον υπολογισμό των λεπτών για κάθε παίκτη για κάθε μία από τις επόμενες αγωνιστικές εβδομάδες που θα γίνει ο υπολογισμός των αναμενόμενων πόντων. Τα λεπτά που παίζει ένας παίκτης έχουν κρίσιμο ρόλο στο να προβλέψεις την αξία του για τις επόμενες αγωνιστικές εβδομάδες, ένα γεγονός που είναι αρκετά ευνόητο. Όσο περισσότερο χρόνο βρίσκεται ένας παίκτης στον αγωνιστικό χώρο, τόσο καλύτερες είναι οι πιθανότητες να σκοράρει πόντους.

0.8.2 Αποτελέσματα

Σε αυτήν την ενότητα, θα συγκρίνουμε τα τελικά αποτελέσματα της εργασίας με άλλα παρόμοια έργα που προβλέπουν την αναμενόμενη αξία για παίκτες της Premier League και αναφέρθηκαν στο κεφάλαιο 0.5 με τις σχετικές εργασίες. Θα γίνει η σύγκριση βάσει δεδομένων πρόβλεψης για την επερχόμενη αγωνιστική εβδομάδα, και θα χρησιμοποιηθούν οι μετρικές MAE, RMSE και R^2 σκορ για την αξιολόγηση. Τα μοντέλα συγκρίνονται σε 6900 κοινά δείγματα από τη σεζόν 2022/23, αγωνιστικές εβδομάδες: 26-38 (πάνω από το 1/3 της σεζόν). Το μοντέλο Kiwi είχε πολλές απουσιάζουσες τιμές, επομένως συγκρίθηκε ξεχωριστά με το δικό μας μοντέλο (Αγωνιστικές εβδομάδες: 26-33 / Σεζόν: 2022-23 / N=2317) και το δικό μας μοντέλο είχε επίδοση σημαντικά καλύτερη σε κάθε μετρική. Τα αποτελέσματα των μοντέλων παρουσιάζονται στον παρακάτω πίνακα.

	MAE	RMSE	R^2
Μοντέλο Εργασίας	**1.2898**	2.3004	0.3334
FPL Review	1.3005	**2.2713**	**0.3502**
Mikkel's Model	1.3175	2.3278	0.3175
Fantasy Football Fix	1.3678	2.3606	0.2981
Fantasy Football Scout	1.3279	2.3083	0.3306
Fantasy Football Hub	1.4150	2.3798	0.2867

Table 5. *Αποτελέσματα Μοντέλων (Αγωνιστικές εβδομάδες: 26-38 / Σεζόν: 2022-23 / N=6900)*

Παρατηρούμε ότι το δικό μας μοντέλο παρουσιάζει εξαιρετική απόδοση σε σύγκριση με

τα κορυφαία μοντέλα του πεδίου. Είναι σημαντικό να σημειωθεί ότι όλα αυτά τα μοντέλα (εκτός από το μοντέλο Kiwi) αποτελούν προϊόντα για τα οποία πρέπει να πληρώσεις και θεωρούνται οι καλύτερες διαθέσιμες λύσεις στην κοινότητα του FPL.

Το δικό μας μοντέλο είναι το καλύτερο όσον αφορά το Μέσο Απόλυτο Σφάλμα και δεύτερο μόνο στις άλλες δύο μετρικές, πίσω μόνο από το FPL Review. Επομένως, μπορούμε να πούμε με ασφάλεια ότι αποτελεί μία από τις καλύτερες λύσεις που έχουν κατασκευαστεί για το πρόβλημα του FPL.

Στο Σχήμα 8 παρουσιάζονται τα καλύτερα μοντέλα. Όσο πιο πάνω δεξιά τόσο το καλύτερο για το μοντέλο και όσο μεγαλύτερο το μέγεθος της φούσκας τόσο μεγαλύτερο το R^2 σκορ του μοντέλου.

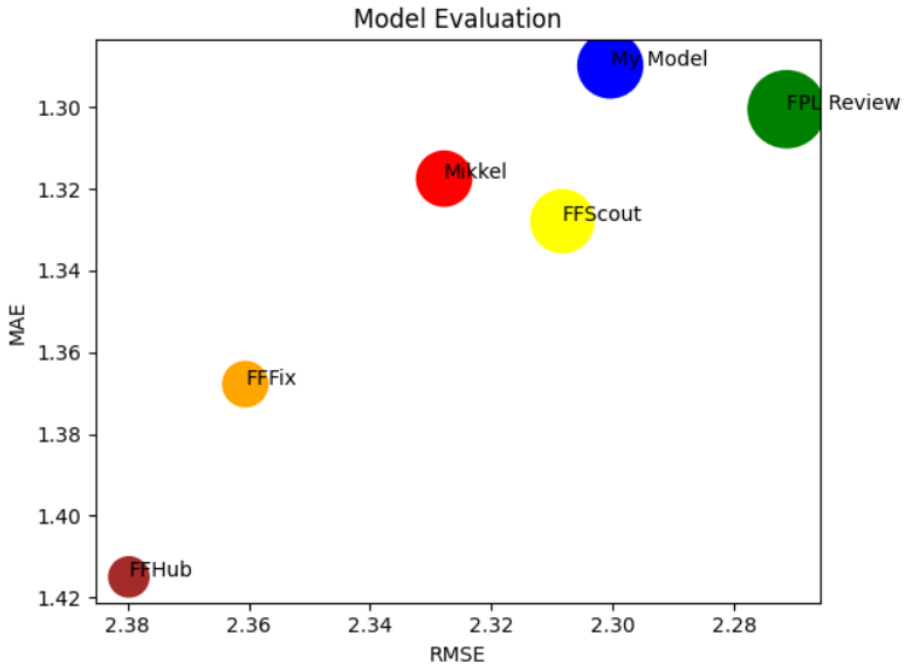


Figure 8. *Διάγραμμα Αξιολόγησης Μοντέλων*

Ένας ακόμα πολύ θετικός δείκτης της ικανότητας του δικού μας μοντέλου είναι το γεγονός ότι η ομάδα που διαχειριζόταν από τη GW24 κατέλαβε την πρώτη θέση ανάμεσα σε όλες τις ομάδες (πάνω από 25,000 ομάδες) που ξεκίνησαν από τον ίδιο γύρο [Σχήμα 9].

Figure 9. *Λίγκα Ομάδων που ξεκίνησαν τη GW24*

0.8.3 Βελτιστοποίηση στο FPL και Solvers

Η βελτιστοποίηση στο FPL αναφέρεται στη διαδικασία επιλογής του βέλτιστου συνδυασμού παικτών μέσα στους δεδομένους περιορισμούς του παιχνιδιού, με σκοπό τη μεγιστοποίηση του συνολικού αναμενόμενου αριθμού πόντων που κερδίζει η ομάδα σε έναν συγκεκριμένο ορίζοντα αγωνιστικών.

Το πρόβλημα βελτιστοποίησης του FPL είναι ένα πρόβλημα γραμμικού προγραμματισμού με μικτές ακέραιες μεταβλητές (MILP), που σημαίνει ότι ορισμένες από τις μεταβλητές περιορίζονται να είναι ακέραιοι αριθμοί, ενώ άλλες μεταβλητές μπορούν να είναι μη ακέραιες. Ο τρόπος επίλυσης αυτών των προβλημάτων είναι μέσω της μεθόδου Branch and Cut. Το Branch and Cut περιλαμβάνει την εκτέλεση αλγορίθμου Branch and Bound και τη χρήση cutting planes για την σύσφιξη των γραμμικών προγραμματισμών [42].

Σε αυτήν την εργασία, έχω ενσωματώσει τον solver του Sertalp [43]. Χρησιμοποιεί τη βιβλιοθήκη sasoptpy για να μοντελοποιήσει το πρόβλημα του FPL και να εκφράσει όλους τους διάφορους περιορισμούς και μεταβλητές. Ο solver επιτρέπει στον χρήστη να εισαγάγει τις ρυθμίσεις που επιθυμεί, όπως την απόσβεση (decay), την αξία των χρημάτων στην τράπεζα, την αξία μίας δωρεάν μεταγραφής κλπ, καθώς και τα δεδομένα της ομάδας του για να δημιουργήσει μονοπάτια μεταγραφών όπως αυτό που φαίνεται στο [Σχήμα 10].

Συνοψίζοντας, οι solvers είναι εργαλεία που χρησιμοποιούνται για να μοντελοποιήσουν και να επιλύσουν το πρόβλημα του FPL, βασιζόμενοι σε ένα σύνολο χρήσιμων τιμών αναμενόμενης αξίας (EV), όπως αυτές που παράχθηκαν σε αυτήν τη διπλωματική εργασία. Ο solver είναι ο "τυφλός εργάτης" που μεταφράζει τις χρήσιμες τιμές αναμενόμενης αξίας των παικτών σε εφαρμόσιμα σχέδια και προτάσεις.

	iter	buy	sell	score
0	0	-	-	128.436011
1	1	Salah	Fernandes	127.947591

Figure 10. *Πίνακας Προτεινόμενων Μεταγραφών από το Solver*

0.9 Επίλογος

Σε αυτή τη διπλωματική εργασία, εξερευνήθηκε η εφαρμογή τεχνικών επιστήμης των δεδομένων και μηχανικής μάθησης στο πλαίσιο του Fantasy Premier League (FPL). Οι στόχοι της εργασίας ήταν οι εξής:

- Ανάπτυξη μοντέλων μηχανικής μάθησης για την πρόβλεψη της Αναμενόμενης Αξίας των παικτών
- Πρόταση βέλτιστων κινήσεων βασισμένων στις προβλεπόμενες τιμές της Αναμενόμενης Αξίας
- Συνεισφορά στον τομέα της ανάλυσης του FPL και προώθηση της εφαρμογής τεχνικών μηχανικής μάθησης στο παιχνίδι

Κατά τη διάρκεια της εργασίας, έχουμε συζητήσει την εξέλιξη των analytics στον αθλητισμό, καθώς και τις πιο σχετικές μετρικές των analytics στο ποδόσφαιρο, τις θεωρητικές βάσεις των διαφόρων τεχνικών παλινδρόμησης που χρησιμοποιούνται, έχουμε εξερευνήσει σχετική βιβλιογραφία στον τομέα και έχουμε κατασκευάσει διάφορα σύνολα δεδομένων, αναγνωρίζοντας τους κύριους παράγοντες που επηρεάζουν την απόδοση των παικτών στο FPL.

Επιπλέον, έχουμε εξετάσει διάφορους αλγορίθμους και τεχνικές μηχανικής μάθησης, επιλέγοντας τους πλέον κατάλληλους για τη συγκεκριμένη εργασία μας.

Τα αποτελέσματα που προέκυψαν από τις πειραματικές μας μελέτες απέδειξαν τη δυνατότητα της μηχανικής μάθησης να βελτιώσει την απόδοση στο FPL. Χρησιμοποιώντας ιστορικά δεδομένα παικτών και ομάδων και ενσωματώνοντας σχετικά χαρακτηριστικά όπως στατιστικά παικτών, την ισχύ ομάδας και φόρμα ομάδας/παικτών, τα μοντέλα μας μπόρεσαν να παράξουν ικανοποιητικές προβλέψεις για τους μελλοντικούς πόντους των παικτών. Επιπλέον, ενσωματώσαμε με επιτυχία τον αλγόριθμο βελτιστοποίησης του Sertalp, ο οποίος δημιουργεί εφαρμόσιμα σχέδια για τους παίκτες του FPL μέσα στο πλαίσιο των κανόνων του παιχνιδιού. Τα ευρήματά μας υπογραμμίζουν τη σημασία της επιλογής χαρακτηριστικών και της αξιολόγησης μοντέλων στο πλαίσιο του FPL. Παρατηρήσαμε ότι ορισμένα χαρακτηριστικά, όπως η προηγούμενη απόδοση των παικτών σε διάφορα χρονικά παράθυρα, η ισχύς της ομάδας και η τρέχουσα απόδοση της ομάδας και των παικτών, είχαν υψηλή επίδραση στην πρόβλεψη της μελλοντικής απόδοσης. Επιπλέον, οι τεχνικές αξιολόγησης των μοντέλων, συμπεριλαμβάνοντας τη μέθοδο cross validation και των μετρικών απόδοσης όπως το RMSE, το R^2 σκορ και το MAE, αποδείχθηκαν κρίσιμα για την αξιολόγηση της αποτελεσματικότητας των μοντέλων μας και τη σύγκριση της απόδοσής τους.

Παρά τα ελπιδοφόρα αποτελέσματα που επιτεύχθηκαν σε αυτήν τη μελέτη, είναι σημαντικό να αναγνωριστούν ορισμένοι περιορισμοί και πιθανοί τομείς βελτίωσης. Πρώτον, ένας πιθανός περιορισμός της μελέτης αυτής έγκειται στο γεγονός ότι χρησιμοποιήθηκαν δεδομένα από τη σεζόν 2019-20 έως τη σεζόν 2022-23. Οι επιπτώσεις του COVID-19 επηρέασαν σημαντικά την Premier League, με τα γήπεδα να είναι άδεια λόγω περιορισμών COVID. Επομένως, καθώς ένα σημαντικό μέρος των δεδομένων συλλέχθηκε κατά τη διάρκεια αυτής της περιόδου, η έκταση της επίδρασης του να είναι ένας παίκτης ή μία ομάδα εντός ή εκτός έδρας μπορεί να μην έχει ερμηνευθεί σωστά στα δεδομένα. Δεύτερον, ένας τομέας για βελτίωση θα ήταν αναμφίβολα μια ξεχωριστή μελέτη για την πρόβλεψη των xMins (αναμενόμενα λεπτά των παικτών). Όπως αναφέρθηκε νωρίτερα, η σημασία των xMins είναι εξαιρετική όταν προβλέπουμε τους μελλοντικούς αναμενόμενους πόντους ενός παίκτη, καθώς η πιο σημαντική προϋπόθεση για να σκοράρει πόντους είναι να βρίσκεται στο γήπεδο. Ένας άλλος πιθανός τομέας έρευνας θα μπορούσε να είναι η δημιουργία και η αξιολόγηση ειδικών μοντέλων που προβλέπουν την αναμενόμενη αξία για υψηλών επιδόσεων παίκτες ξεχωριστά. Κάτι που αξίζει να ληφθεί υπόψη είναι τα ensemble μοντέλα, δηλαδή τη συνδυασμένη χρήση δεδομένων από πολλά ανεξάρτητα μοντέλα σε ένα μοντέλο, παρόμοια με την ιδέα της "σοφίας του πλήθους". Η δυνατότητα αποφυγής λαθών μεμονωμένων μοντέλων μπορεί να είναι χρήσιμη, αν και μερικές φορές μπορεί να γίνεται σε βάρος ειδικών "εμπνεύσεων" από αυτά. Επιπλέον, η ενσωμάτωση πιο προηγμένων τεχνικών μηχανικής μάθησης, όπως μοντέλα βαθιάς μάθησης (deep learning), μπορεί να ενισχύσει περαιτέρω την προβλεπτική δύναμη των μοντέλων μας. Τέλος, η δημιουργία heuristics για τη μείωση του χρόνου εκτέλεσης του solver είναι κάτι που θα μπορούσε να βελτιώσει την εμπειρία για έναν χρήστη, ειδικά εάν η ταχύτητα είναι πιο σημαντική από την ακρίβεια. Οι μέθοδοι με heuristics δεν είναι πάντοτε επιθυμητές, αφού δεν υπάρχει τρόπος να γνωρίζουμε πόσο μακριά βρισκόμαστε από την βέλτιστη λύση, αλλά ένας συνδυασμός παραγωγής μιας γρήγορης λύσης με heuristic και την εκτέλεση βελτιστοποίησης στο παρασκήνιο για την απόδειξη της βέλτιστης λύσης μπορεί

να αποτελέσει τον καλύτερο συνδυασμό.

Συνοψίζοντας, η εφαρμογή τεχνικών μηχανικής μάθησης στο FPL έχει αποδειχθεί αποτελεσματική στην πρόβλεψη της απόδοσης των παικτών και την βέλτιστη επιλογή ομάδας. Η δυνατότητα αξιοποίησης ιστορικών δεδομένων και εξαγωγής πολύτιμων πληροφοριών από μεγάλες ποσότητες πληροφοριών παρέχει στους παίκτες ένα σημαντικό ανταγωνιστικό πλεονέκτημα. Τα αποτελέσματα αυτής της εργασίας συνεισφέρουν στον αυξανόμενο χώρο της ανάλυσης αθλητικών δεδομένων και δημιουργούν ένα θεμέλιο για μελλοντικές έρευνες στον τομέα της επιστήμης των δεδομένων στο FPL. Η ενσωμάτωση αλγορίθμων μηχανικής μάθησης έχει τη δυνατότητα να αλλάξει τον τρόπο που οι παίκτες προσεγγίζουν την επιλογή ομάδας, τις μεταγραφές και τη συνολική στρατηγική του παιχνιδιού. Καθώς το FPL εξελίσσεται και η διαθεσιμότητα δεδομένων αυξάνεται, ο ρόλος της μηχανικής μάθησης στο παιχνίδι αναμένεται να γίνει ακόμη πιο σημαντικός. Με την παροχή προηγμένων προγνωστικών μοντέλων και αλγορίθμων βελτιστοποίησης, μπορούμε να ενισχύσουμε τη συνολική εμπειρία του παιχνιδιού και να παρέχουμε βελτιωμένες επιδόσεις. Αυτή η εργασία αποτελεί ένα πρώτο βήμα προς την απελευθέρωση του πλήρους δυναμικού της μηχανικής μάθησης στο Fantasy Premier League και παροτρύνει για περαιτέρω έρευνα και καινοτομία σε αυτόν τον συναρπαστικό τομέα.

Ο κώδικας της εργασίας μπορεί να βρεθεί στο GitHub μου [40].

Chapter 1

Introduction

Fantasy Premier League (FPL) is a popular online game that allows participants to create virtual teams of real-life Premier League players. It provides football enthusiasts with an interactive platform to engage with the Premier League and showcase their managerial skills. The game assigns points to players based on their performance in actual Premier League matches, such as goals scored, assists made, clean sheets, and other statistical contributions. The appeal of FPL lies in its ability to provide a virtual managerial experience, where participants must carefully select their team, make transfers, choose captains, and decide on formation and tactics. FPL managers compete against each other, striving to accumulate the highest number of points and achieve the top rankings within their leagues or globally.

The game has witnessed exponential growth in recent years, attracting millions of participants worldwide (over 11 million teams for the 2022/23 season). Its popularity can be attributed to several factors. Firstly, FPL allows fans to deepen their involvement with the Premier League by assuming the role of managers and having a stake in the performances of their chosen players. It enhances the overall football experience by fostering a sense of ownership and engagement. Moreover, FPL provides a competitive environment where managers can test their skills and knowledge of the game. It requires a keen understanding of player form, fixture schedules, team dynamics, and other factors that can influence player performances. FPL managers must make strategic decisions regarding player selection, formation optimization, captaincy choices, and transfers to maximize their team's potential and gain an edge over their competitors.

With the increasing popularity of the game, there has been a surge in the use of data analytics and machine learning techniques to gain a competitive advantage. FPL managers are seeking ways to leverage data-driven strategies and predictive models to optimize their team selection, improve decision-making, and ultimately enhance their overall performance in the game. This has led to the exploration of various machine learning algorithms, statistical models, and optimization techniques to predict player performance, estimate player values, and generate optimal strategies. In light of the significance and growing interest in applying machine learning to FPL, this thesis aims to contribute to the field by developing predictive models for players' Expected Value (EV) and generating optimal moves based on these predictions. By harnessing the power of machine learning, FPL managers can make informed decisions, maximize their team's

potential, and improve their rankings within the competitive FPL landscape.

The motivation behind this thesis stems from the potential of machine learning to significantly enhance FPL performance and decision-making. FPL managers face various challenges in navigating the complexities of the game and optimizing their team's performance. These challenges include the selection of the most promising players, forming an effective team composition, making timely transfers, and choosing the right captaincy strategy. Traditionally, FPL managers have relied on subjective assessments, intuition, and limited data to make decisions. However, in a game where consistency and long-term planning can make a substantial difference, there is a growing recognition of the need for more data-driven and objective approaches. This realization has fueled the exploration of machine learning techniques as powerful tools for analyzing vast amounts of data and generating valuable insights. Machine learning offers the potential to unlock patterns, relationships, and trends within player statistics, historical data, and other relevant factors that can influence player performance. By training machine learning models on historical FPL data, managers can gain a deeper understanding of how player attributes, form, fixtures, and other variables correlate with future performance. This predictive capability can enable managers to make informed decisions backed by data-driven insights.

The primary objectives of this thesis are to develop and evaluate machine learning models for FPL. Specifically, the focus is on predicting Expected Value (EV) for players and generating optimal moves based on these predictions. By achieving these objectives, the thesis aims to contribute to the field of FPL analytics and advance machine learning applications in the context of FPL.

The thesis is organized into 10 chapters. Chapter 1 provides an introduction to the topic, including the background, motivation, objectives, and an overview of the thesis structure. Chapter 2 explains the game on a deeper level. Chapters 3 and 4 focus on the evolution of sports and soccer analytics respectively, while exploring famous applications of analytics or useful metrics, such as expected Goals (xG), that play a key role in the thesis. Chapter 5 delves into the theory of the regression models we will train, and Chapter 6 presents the related work. Then, Chapter 7 is about the data flow in the project, and Chapter 8 explores the training and evaluation of our models. Chapter 9 is about the EV calculation process, the solvers that are used, and final results presentation. Finally, Chapter 10 concludes the thesis, summarizing the key findings and their implications.

The development of machine learning models for EV prediction and generating optimal moves directly benefits FPL managers. By utilizing these models, managers gain access to valuable insights and recommendations to enhance their decision-making processes. The accurate prediction of EV enables managers to identify undervalued players, exploit favorable fixtures, and optimize team composition strategies. The generated optimal moves provide managers with actionable recommendations, leading to improved team performance and higher rankings in the competitive FPL landscape.

This research extends the application of machine learning techniques to the domain of sports analytics, specifically within the context of FPL. By demonstrating the efficacy of machine learning models in predicting player performance and generating optimal moves, this research showcases the potential of machine learning to enhance decision-making

and strategy formulation in sports-related games. The methodologies, techniques, and insights derived from this research can be applied to other sports fantasy games and even real-world sports analytics, contributing to the broader field of sports data analysis.

In summary, this introduction chapter has provided an overview of the thesis, outlining the background, motivation, objectives, thesis structure, and significance of applying machine learning in FPL. The subsequent chapters will delve deeper into the research, exploring relevant literature, methodologies, predictive models for EV, generating optimal moves, and concluding with the key findings and their implications.

The code for this project can be found on my GitHub [40].

Chapter 2

Problem Description and Objectives

2.1 The Game

In this chapter, we will provide an overview of the basic elements of the game.

2.1.1 Rules and Scoring

Fantasy Premier League is based on the performance of players in the English Premier League. Players are allowed to build their own team with a budget of £100 million. The aim is to select a team of 15 players (including 2 goalkeepers, 5 defenders, 5 midfielders, and 3 forwards) that will score the most points based on their performance in Premier League matches. Managers must select a starting lineup of 11 players from their squad of 15 before each gameweek, including a captain and a vice-captain. The captain earns double points, and the vice-captain can take over if the captain does not play [Figure 2.3]. Points are awarded for various on-pitch actions, including scoring goals, providing assists, keeping clean sheets, making saves, and earning bonus points. The number of points awarded for each action varies depending on the position of the player [12].

	Goalkeepers	Defenders	Midfielders	Forwards
For playing up to 60 minutes	1	1	1	1
For playing 60 minutes or more	2	2	2	2
For each goal scored	6	6	5	4
For each assist	3	3	3	3
For keeping a clean sheet	4	4	1	-
For every 3 saves made	1	-	-	-
For each penalty saved	5	-	-	-
For each penalty missed	-2	-2	-2	-2
For every 2 goals conceded	-1	-1	-	-
For each yellow card	-1	-1	-1	-1
For each red card	-3	-3	-3	-3
For each own goal	-2	-2	-2	-2

Table 2.1. *Point System of FPL*

- In a match, the three best players are decided according to the FPL Bonus Points System and awarded a bonus of 1, 2 and 3 points. Bonus Points are calculated

according to 32 match statistics, where goals scored, assists and clean sheets are the factors that are heaviest weighted.

- For a goalkeeper or defender to receive points for a clean sheet, he has to play at least 60 minutes, excluding stoppage time.
- If a goal is scored on a direct free kick or a penalty, the player who got the free kick/penalty is awarded an assist. Also, if a goal is scored from a "rebound" after a shot saved by the opposing goalkeeper, an assist is awarded to the player that made the initial shot that was saved.

2.1.2 Transfers

Fantasy Premier League players are allowed to make a certain number of free transfers per gameweek. The number of free transfers available per gameweek depends on the player's overall strategy, and can range from 1 to 2 per gameweek. Additional transfers lead to a 4 point penalty each. When a player transfers a player out of their team, they will earn or lose money depending on the transfer fee, as well as the current value of the player. This can affect the overall value of the team, which can in turn affect future transfers. [12]

Figure 2.1. *FPL Transfer*

2.1.3 Chips

In addition to transfers, Fantasy Premier League also includes the following "chips" that players can use to gain an advantage during the season [12].

- **Wildcard.** The Wildcard allows the manager to replace his entire selected squad for free. As for the Wildcard squad selection, the same rules apply as in the regular fantasy team composition. Hence, one can only select a maximum of three players from each team, and the formation criterion must be upheld. When playing a Wildcard, a manager's budget is set to the sell price of his selected squad in that particular gameweek. Further, when playing a Wildcard, any saved transfers will be lost. The chip can be used twice a season, once in the first and once in the second half of the season.

- **Bench Boost.** The Bench Boost allows a manager to receive points for all the 15 players in the selected squad. The chip can only be used once a season.
- **Free Hit.** The Free Hit allows a manager to replace the entire selected squad for one gameweek. However, for the next gameweek, the selected squad is reversed back to the squad from the previous gameweek. The Free Hit can only be used once a season. As for the Wildcard, the same transfer- and budget rules apply for the Free Hit. The chip can only be used once a season.
- **Triple Captain.** The Triple Captain triples the points of the captain for a gameweek. If the captain does not play, the points of the vice-captain are tripled. If neither the captain nor the vise-captain play, no players are awarded triple points. The chip can only be used once a season.

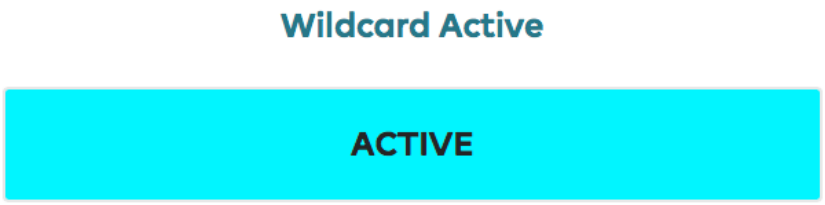


Figure 2.2. *Wildcard*

2.1.4 Leagues and Rankings

Fantasy Premier League includes several different types of leagues that players can join, including private leagues, public leagues, and head-to-head leagues. Each league has its own rules and scoring system, and players can compete against each other for prizes and bragging rights. Players can also view their overall ranking in the league, which is based on their total points earned throughout the season. [12]

2.2 The Problem

In Fantasy Premier League each manager attempts to maximize his or her total number of points over an entire season. As seen in this chapter, a number of decisions must be made each gameweek. The decisions include which players to add to the selected squad and which players to pick for the starting line-up. In addition, a captain and vice-captain must be chosen. Further, one has to set a substitution priority for the players that are not selected for the starting line-up. As pointed out, a manager is allowed to perform transfers during the season. Therefore, a manager must also decide whether to make a transfer or not, and consequently which players to be transferred in and out for each gameweek. Hence, decisions are made in a multi-period manner. Considering the fact that there are over 500 players in the Premier League each season, this isn't an easy

task. A player's value can also be different for different players, as it is based on a lot of factors (sometimes even not quantifiable). But how can that be done in an unbiased way for every player? Even if we could determine each player's value, what is the optimal transfer strategy to maximize our team's future points? What is the ideal time to play the chips? These are questions that every FPL manager is thinking about every gameweek of the season.

2.3 The Goals

As mentioned in the introduction, this thesis has the following 3 goals:

- The first and main goal is to develop machine learning models capable of predicting the Expected Value (EV) for FPL players. EV represents the expected number of points a player is likely to score in a given match based on historical data, player attributes, fixtures, and other relevant factors. Accurate EV predictions are vital for FPL managers as they form the foundation of effective player selection and team composition strategies. By developing models that can estimate EV with high accuracy, managers can make more informed decisions about player acquisitions, captaincy choices, and overall team strategy.
- The second goal is to generate optimal moves for FPL managers based on the predicted EV values. Optimal moves may include transfer recommendations, captaincy choices, or formation adjustments that maximize the team's potential points. By incorporating EV predictions into the decision-making process, managers can optimize their team's performance by identifying undervalued players, exploiting favorable fixtures, and capitalizing on emerging trends. The objective is to develop algorithms and strategies that utilize EV predictions to generate actionable insights and recommendations, enabling managers to make data-driven moves with a higher probability of success.
- The final goal is to contribute to the field of FPL analytics and advance the application of machine learning techniques in FPL. This thesis aims to provide practical insights, methodologies, and tools that can be utilized by FPL managers to enhance their decision-making processes. By demonstrating the effectiveness and utility of machine learning models in FPL, this research seeks to foster a data-driven approach to FPL management and contribute to the overall understanding of the game. The objective is to empower FPL managers with the knowledge and resources needed to improve their team's performance and achieve higher rankings.

Figure 2.3. *FPL Team*

Chapter 3

The Evolution of Sports Analytics

Sports analytics, the use of statistical analysis and data visualization techniques to extract insights and knowledge from sports data, has become an increasingly important field in recent years. The use of statistics in sports has been around for over a century, but it wasn't until the past few decades that it really began to gain traction. In this chapter, we will explore the evolution of sports analytics, from its early beginnings to the present day.

3.1 Early Beginnings

The earliest form of sports analytics can be traced back to the late 19th century, when baseball statistics were first recorded and analyzed. Baseball analyst Henry Chadwick is often credited as the father of baseball statistics, as he created the first box score in 1859. Over the next few decades, statistics continued to be recorded and analyzed, but the use of statistics was limited to basic measures. Chadwick was elected to the Hall of Fame in 1938 [44].

3.2 The Rise of Sabermetrics

The use of statistics in sports really took off in the 1970s and 1980s, with the advent of sabermetrics. Sabermetrics is the practice of using advanced statistical analysis to evaluate baseball players and strategies [45].

The term was coined by Bill James, a baseball writer and statistician who began publishing a series of books on baseball statistics in the 1980s. James' work popularized the use of advanced statistical analysis in baseball, and led to the development of new metrics that helped further the understanding of the game [46].

3.3 Moneyball

The Moneyball story began with Billy Beane, the general manager of the Oakland Athletics baseball team in the early 2000s. The Athletics were a small-market team with limited financial resources, and Beane was faced with the challenge of building a competitive team on a tight budget.

To address this challenge, Beane turned to statistical analysis as a way to identify undervalued players who could help the team compete with larger-budget rivals. He assembled a team of data analysts and began using advanced statistical models to evaluate players, focusing on measures such as on-base percentage and slugging percentage rather than more traditional measures such as batting average and stolen bases.

Through this process, Beane and his team were able to identify talented players who were undervalued by traditional scouting methods. They used this data to build a team of players who were not necessarily stars, but who had the potential to outperform their expectations.

The result was a team that was highly successful, making the playoffs in multiple seasons despite having one of the lowest payrolls in the league. The success of the Athletics led to a wider adoption of statistical analysis in baseball and other sports, and the concept of Moneyball became synonymous with the use of data to identify undervalued players and build competitive teams on a limited budget [47].

Michael Lewis's book "Moneyball: The Art of Winning an Unfair Game" chronicled the story of Beane and the Athletics, and was later adapted into a movie starring Brad Pitt as Beane. The book and movie helped to popularize the concept of sports analytics and highlight its importance in building successful teams [47].

The Moneyball story is often cited as an example of the power of data analysis in sports. By using statistical models to identify undervalued players and build competitive teams, Beane and his team were able to outperform their expectations and compete with larger-budget rivals. The success of Moneyball has inspired a new generation of data analysts and revolutionized the way many sports teams approach player recruitment and strategic decision-making.

Figure 3.1. *Oakland Athletics Logo*

3.4 Sports Analytics Today

Today, the use of analytics has spread to many other sports, including basketball, football, soccer, and hockey. Teams and analysts use a variety of data sources, including player tracking data, performance data, and scouting data, to gain insights into player performance and identify areas for improvement. The use of machine learning and artificial intelligence has also become increasingly common, allowing teams to make more sophisticated predictions and develop more effective strategies [48].

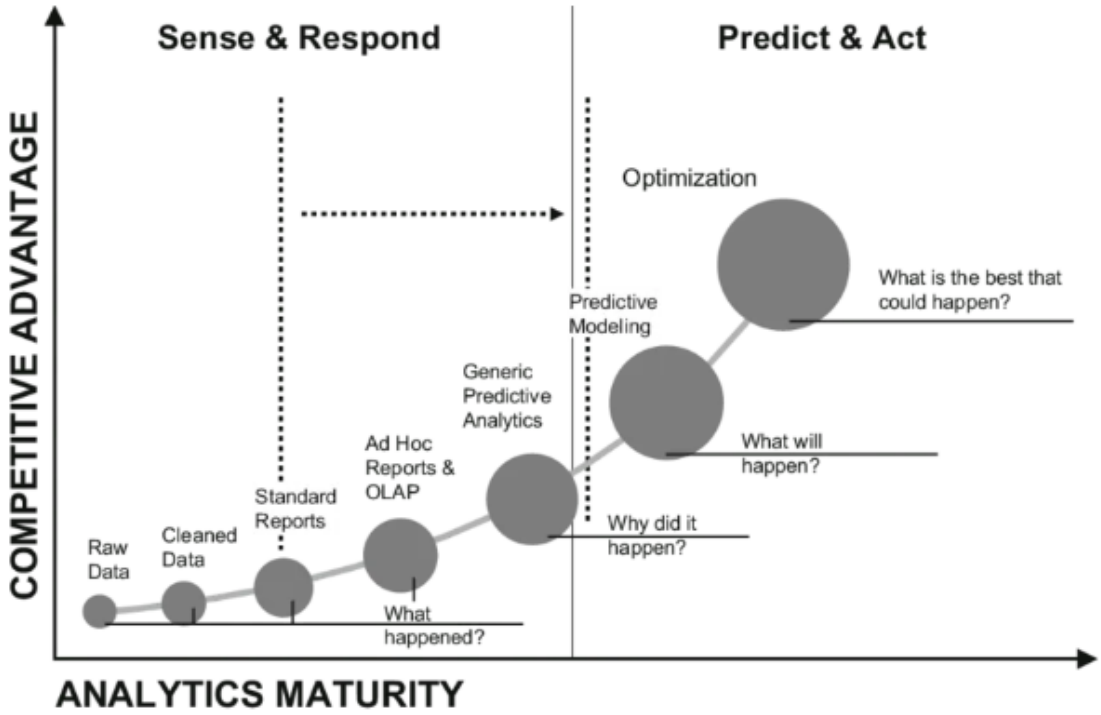


Figure 3.2. *Analytics Stages [1]*

3.4.1 Famous Applications Today

Statistics and data in sports have become commonplace. The average fan can visit websites, pull data from databases in almost every sport or even view analysis done by professionals on websites such as FiveThirtyEight.com [49]. We can view Mike Trout's on-base percentage against right-handed pitchers in a home game. We can see Stephen Curry's three-point percentage in the final moments of a playoff game. We can even find a New York Times Bot that makes recommendations on when to go for it on 4th down [Figure 3.3]

Daryl Morey of the Houston Rockets spearheaded the analytics movement in basketball and completely evolved the way the game is played. As shown in the chart below, the Rockets were way ahead of the league when it came to understanding the efficiency and value of a 3 pt. shot [Figure 3.4]. Whether it's 'Moneyball' or 'Moreyball', a lot of these concepts seem obvious now that we look back a few years. However, at the time, they were considered different and others were years off from adoption [5].

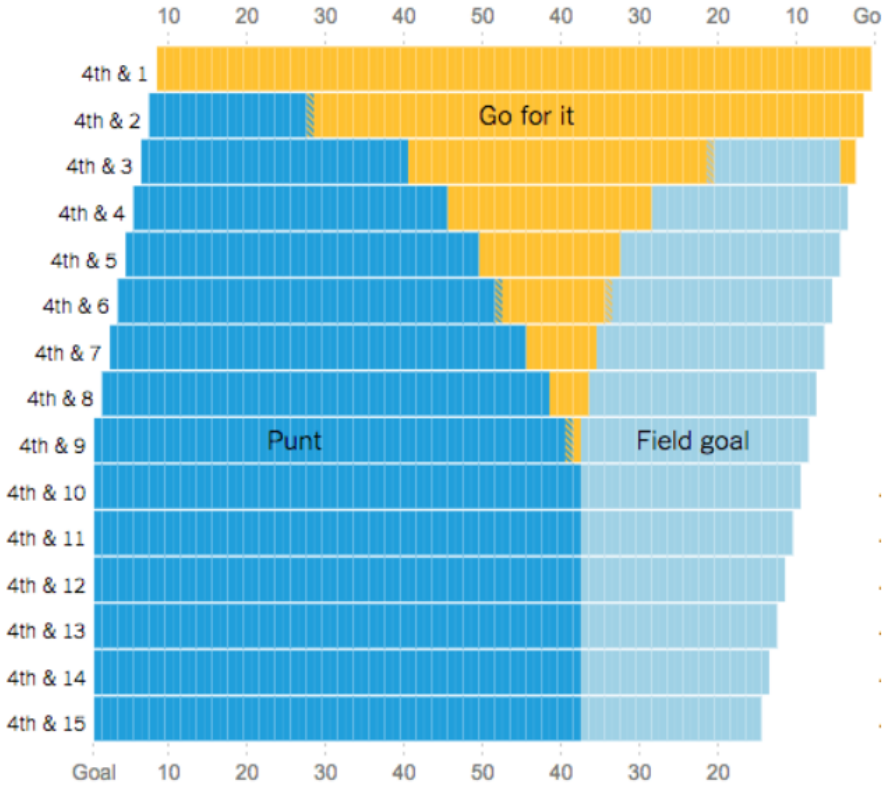


Figure 3.3. *NYT 4th Down Bot Recommender [5]*

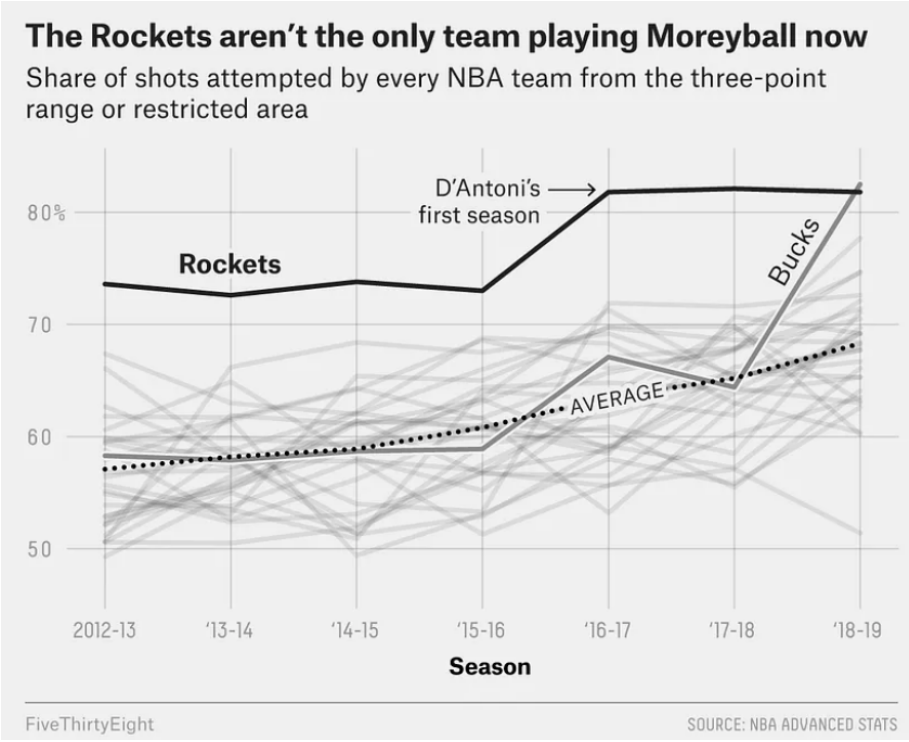


Figure 3.4. *Moreyball [5]*

In the past decades, teams across all of the NBA have widely adopted the use of analytics to better develop more efficient game styles. The midrange shots that we all

grew up watching in Michael Jordan and Kobe Bryant are now considered “bad shots” unless taken by those that are of top efficiency levels there [Figure 3.5]. In baseball, teams have begun to employ the shift in baseball. In football, teams have decided to go for it on fourth down conversions with far more frequency [5].

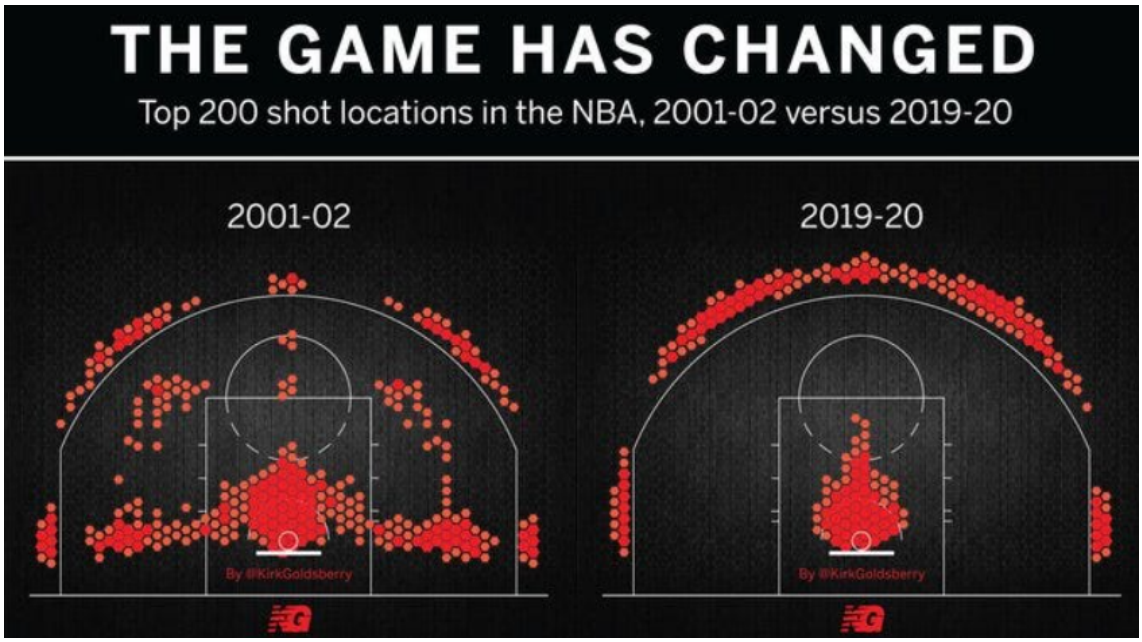


Figure 3.5. *Shot Locations Evolution [5]*

Chapter 4

Soccer Analytics

As we saw, the use of analytics — using data and statistics to better understand something — is growing across most sports. This is especially true in soccer, where the most successful teams are also frequently the most dedicated to analytics. Barcelona, Liverpool and Bayern Munich have all been public about their willingness to use data to inform decision-making. In the international game, so have U.S. Soccer and the German Football Association. More and more, winning teams' decision makers have an analytical edge over the competition [50].

Analytics are tied closely with the understanding of the game. Much of analytics is really about putting concepts that soccer people already understand into terms that can be described with data. Most people know intuitively what a counterattack is, for example. A goal of analytics, though, is to answer a question like "How valuable is a team's ability to prevent counterattacks?" To do that, we need to define and quantify a "counterattack" first. From that definition, an analyst can determine that the team's defensive ability there prevented up to five or six goals in a season. The goal of this quantification is to give an analyst a way to describe at scale what happens on the field [50].

4.1 Possession Value

A lot of useful concepts have been developed in the field throughout the years. Possession value is one of them. Possession value calculates the probability of a goal being scored at any point in a possession. Just like with shots, a team on the ball right outside the box is much more dangerous than a team knocking it around in its own half — this is why straight percentage of possession stats can be misleading. Possession value models that danger. Much of analytics focuses on determining the actions that lead to higher and lower value possessions [50].

Possession value allows us to value every action on the pitch and see how effective players have been for their team. For example, imagine that Kevin De Bruyne has the ball near the halfway line where the possession value is 1% (PV start = 0.01). If he carries the ball down the line and then plays a pass to Raheem Sterling in the box, where the possession value is now 13% (PV end = 0.13), Kevin De Bruyne has increased the likelihood of his team scoring by 12%. This is what we call possession value added (or PV+) and rewards Kevin De Bruyne for his contribution irrespective of what Sterling

chooses to do with the ball next. Manchester City are now 12% more likely to score within the next 10 seconds (PV+ = 0.12). In this way, it allows us to credit players who may previously have been undervalued by more traditional metrics such as goals and assists. We can look at all of the contributions of an individual player and evaluate whether their positive actions outweigh their negative actions [6].

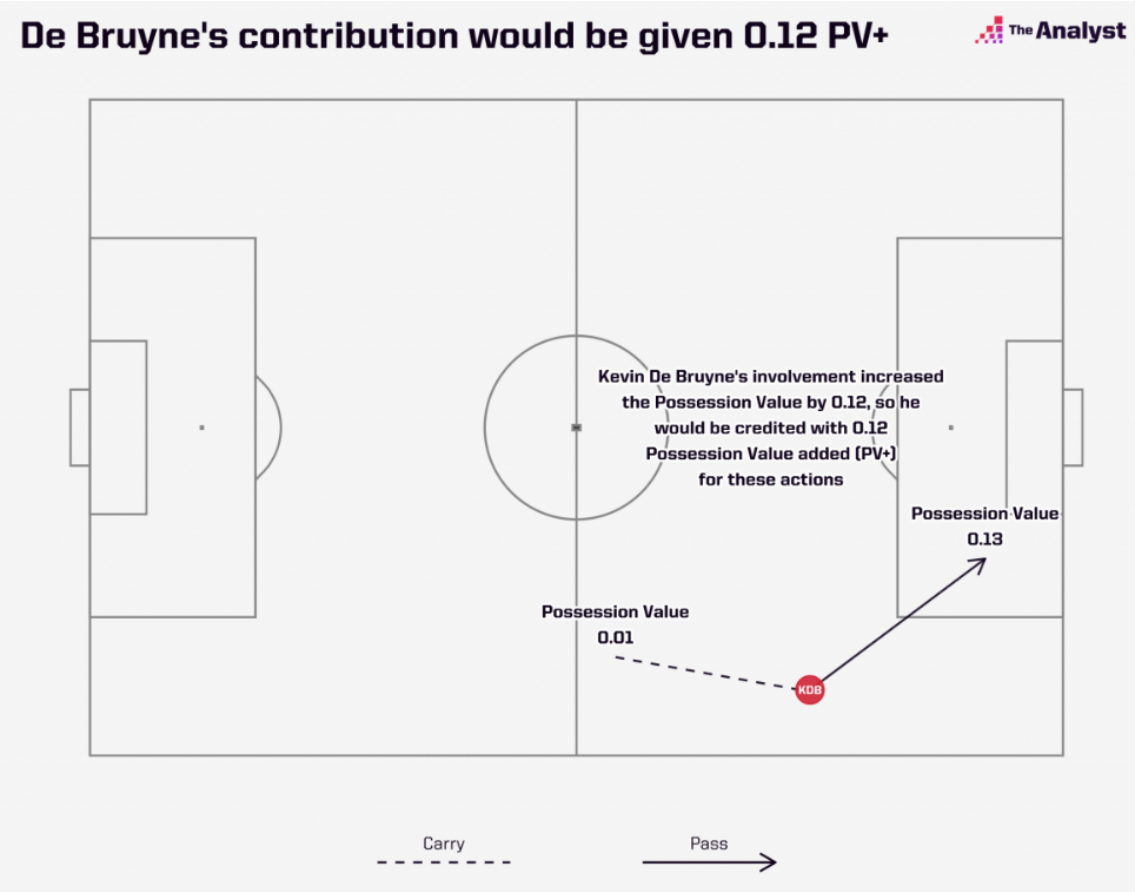


Figure 4.1. *De Bruyne's added value example [6]*

4.2 Sequences

Sequences are defined as passages of play which belong to one team and are ended by defensive actions, stoppages in play or a shot. A sequence starts with a player making a controlled action on the ball. This includes passes but not defensive events such as tackles and interceptions, unless these events are followed by a controlled action such as a pass or dribble [7].

By stringing events together, we are able to value the contributions of the players who don't necessarily score or assist goals but are still integral in the build-up. While secondary assists (the pass before an assist) provide some of this context, we can now go even further back with sequences and see who is starting these moves or who is most frequently involved before these final actions. Here is the passage of play leading to Arsenal's Martin Ødegaard's first Premier League goal in the North London derby against

Tottenham.

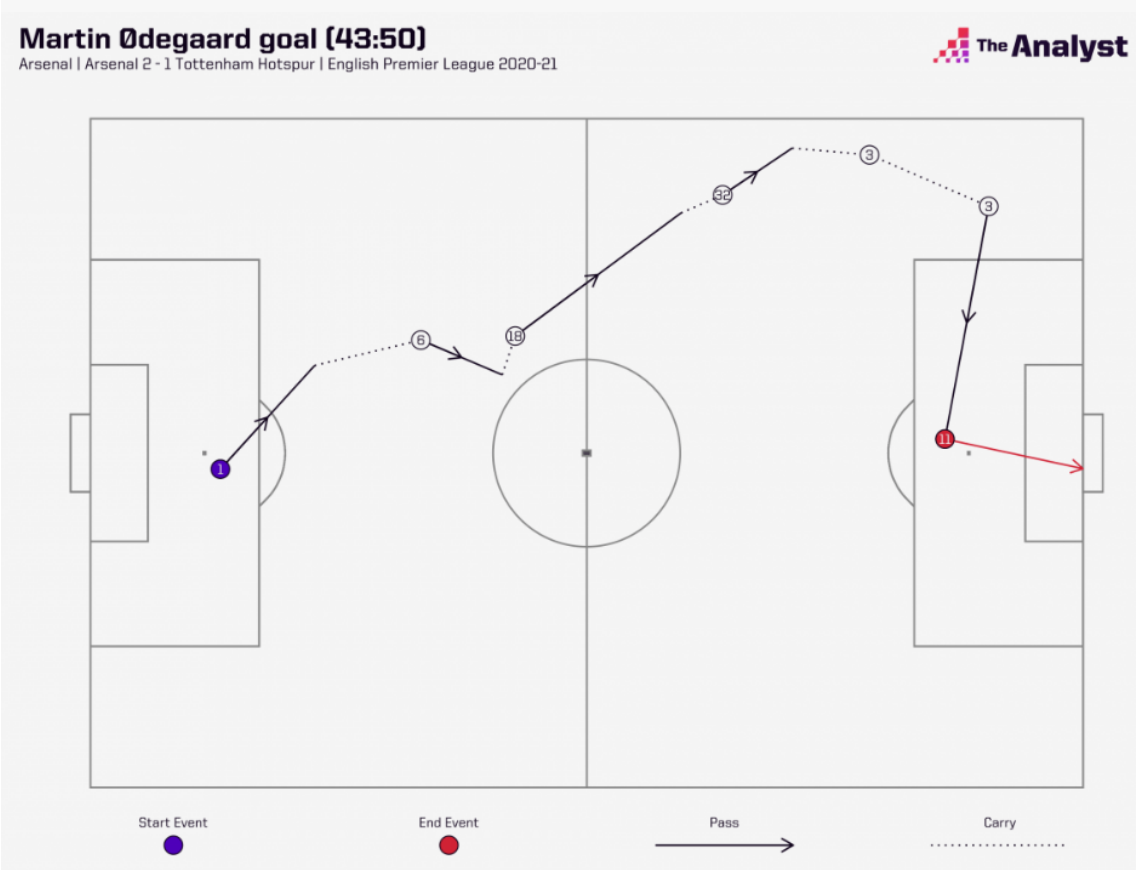


Figure 4.2. *Martin Ødegaard's first goal sequence [7]*

Each team plays differently, and their style is typically hard to quantify. With stylistic metrics derived from the sequence framework, we can now compare how teams typically move the ball. The graphic below shows how Premier Leagues 2020-21 teams currently compare in terms of their number of passes per sequence and direct speed, where direct speed is the average speed of ball movement towards the opponent goal line during a sequence [Figure 4.3].

Unsurprisingly Manchester City are out on their own. They play almost three more passes per sequence than West Bromwich Albion at the other end of the scale and are patient with progressing the ball up the pitch. Comparing teams like this gives a high level overview of how their style of plays contrast, with similar teams clustering according to their directness and passing tendencies [7].

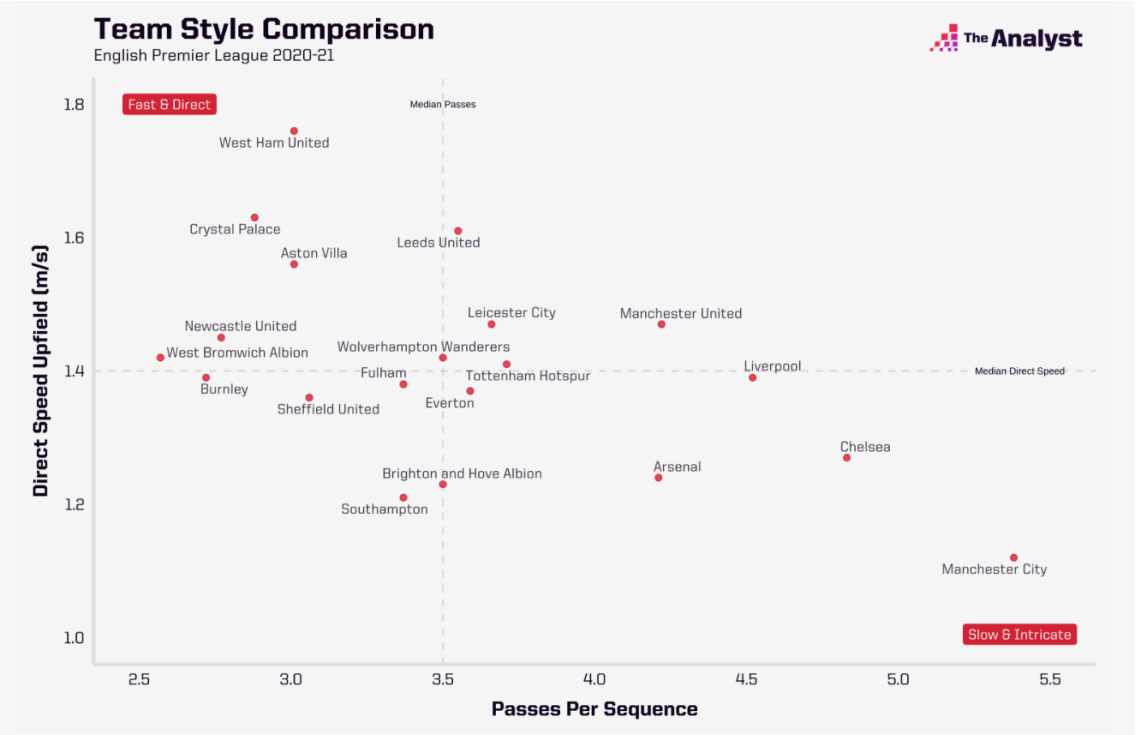


Figure 4.3. *Teams possession style [7]*

4.3 Expected Goals (xG)

At the forefront of the data boom in soccer is xG. It is without doubt one of the most widespread and insightful metrics in the soccer analytics sphere.

4.3.1 What are Expected Goals (xG)?

Expected goals (or xG) measures the quality of a chance by calculating the likelihood that it will be scored from a particular position on the pitch during a particular phase of play. This value is based on several factors from before the shot was taken. xG is measured on a scale between zero and one, where zero represents a chance that is impossible to score and one represents a chance that a player would be expected to score every single time.

We know that a chance from the halfway line isn't as likely to result in a goal as a chance from inside the box. With xG, we can actually quantify how likely a player is to score from each of these situations. For example, suppose the chance from inside the box with a given set of pre-shot characteristics was worth 0.1 xG. This means that an average player would be expected to score one goal from every ten shots in this situation or 10% of the time.

The terminology may be new, but these phrases have been used by football fans and commentators for years before xG was introduced – "he scores that nine times out of ten" or "he should've had a hat-trick" [8].

4.3.2 Common Misconceptions

The main criticisms of expected goals (xG) often appear in scenarios where the metric isn't actually being applied correctly. The most common of which is at the game level. The team that has the higher xG in a match doesn't necessarily imply that they should've won the game. xG is only measuring chance quality and not the expected outcome of the game. Exactly as the old saying suggests, goals do change games and the score line influences how teams play. If a team takes an early lead, they don't necessarily 'need' to generate more chances and we often expect to see the opposition generate more goal scoring opportunities for the remainder of the game in pursuit of a comeback.

Another misconception is in the literal interpretation of the metric name. We do not "expect" goals to occur exactly as the likelihood predicts. We also understand that fractions of goals cannot be scored. The name "expected goals" is derived from the mathematical concept of "expected value" and it is a measure of the likelihood of an outcome occurring. The expected value of a fair coin toss is 50% likely to land on heads and 50% likely to land on tails (the expected heads or the expected tails is 0.5). We do not expect exactly half of our tosses to land on each outcome, but rather that over a larger number of coin tosses, it is likely to regress to this balance. The same applies to expected goals. Variance from the expected value is inevitable and this is valuable information that we can analyze in football.

A player or team who has been overperforming their xG, does not then have to underperform to regress to expectation. This is a concept known as the Gambler's Fallacy. While we would expect them to regress back to scoring in line with their expectation with their future shots, they have already 'banked' this overperformance and so we will still expect them to overperform by this amount in the season aggregates. In the same way, if a coin toss landed on heads ten times in a row, future coin tosses are still equally likely to land on heads as they are tails, but the ten times that the coin landed on heads have already happened [8].

4.3.3 How do we calculate Expected Goals (xG)?

While watching a game, we can intuitively tell which chances were more or less likely to be scored. How close was the shooter to goal? Were they shooting from a good angle? Was it one-on-one? Was it a header?

The difficulty is that there are an average of 25 shots per game that we need to work this out for, all potentially from unique situations. The advantage of an expected goals model is that we can now take the variables above – and others – and quantify how each of these affects the likelihood of a goal being scored. With this, it allows us to value the quality of the chances for all 9,398 shots taken in the Premier League 2019-20 season in a matter of seconds.

Let's analyze the Stats Perform's xG model for example. It is built using a logistic regression model that is powered by hundreds of thousands of shots from our historical Opta data and incorporates a number of variables that affect the likelihood of a goal being scored, some of the most important of which are listed below:

- Distance to the goal
- Angle to the goal
- One-on-one
- Big chance
- Body part (e.g., header or foot)
- Type of assist (e.g., through ball, cross, pull-back etc)
- Pattern of play (e.g., open play, fast break, direct free kick, corner kick, throw-in etc)

We recognise that some situations are particularly unique and so these are modeled independently. Penalties are given a constant value corresponding to their overall conversion rate (0.79 xG); direct free kicks have their own model; and headed chances are valued differently for set-pieces and open play [8].

4.3.4 How can we use Expected Goals (xG)?

Let's compare two players from their 2019-20 seasons, Manchester City's Gabriel Jesus in the Premier League and AC Milan's Hakan Calhanoglu in Serie A. Both players took exactly 100 shots last season (excluding penalties) but scored 14 and 8 goals respectively. So, what was the difference between their shots?

By quantifying the quality of the 100 chances for each player, xG adds additional context to their shots that goes beyond the traditional metrics such as shots on target or average shot distance. We can now measure the quality of chances that each player had [Figure 4.4].

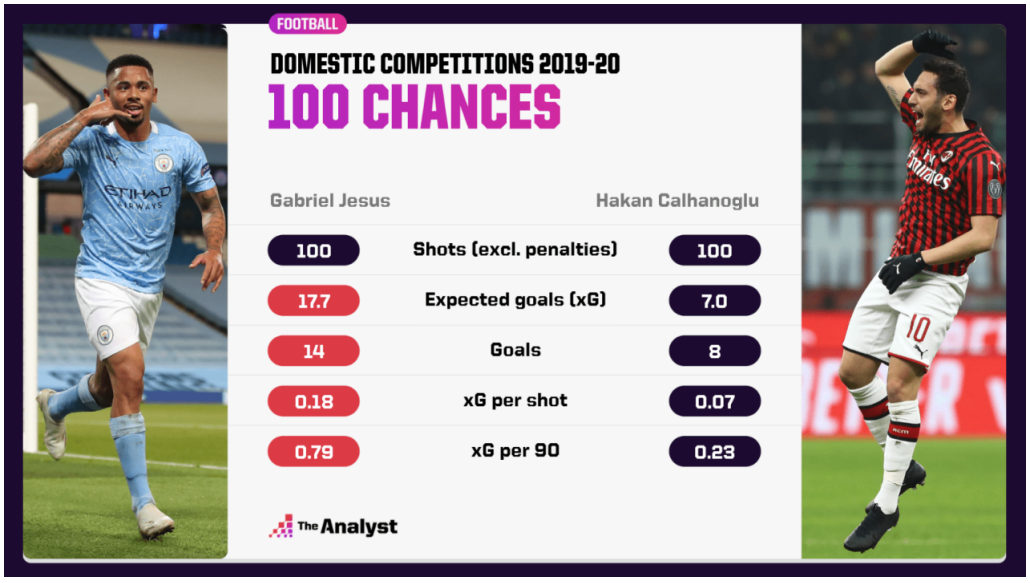


Figure 4.4. *xG Comparison [8]*

From the chances that Gabriel Jesus had we would expect the average player to score nearly 18 goals (17.7 xG). On the other hand, from Hakan Calhanoglu's chances, we would expect the average player to score only 7 goals (7.0 xG). We can immediately understand why their goal scoring output was so different. Despite Jesus overperforming and Calhanoglu underperforming slightly according to their expected goals output, their 100 chances were very different in quality and their output reflected that [8].

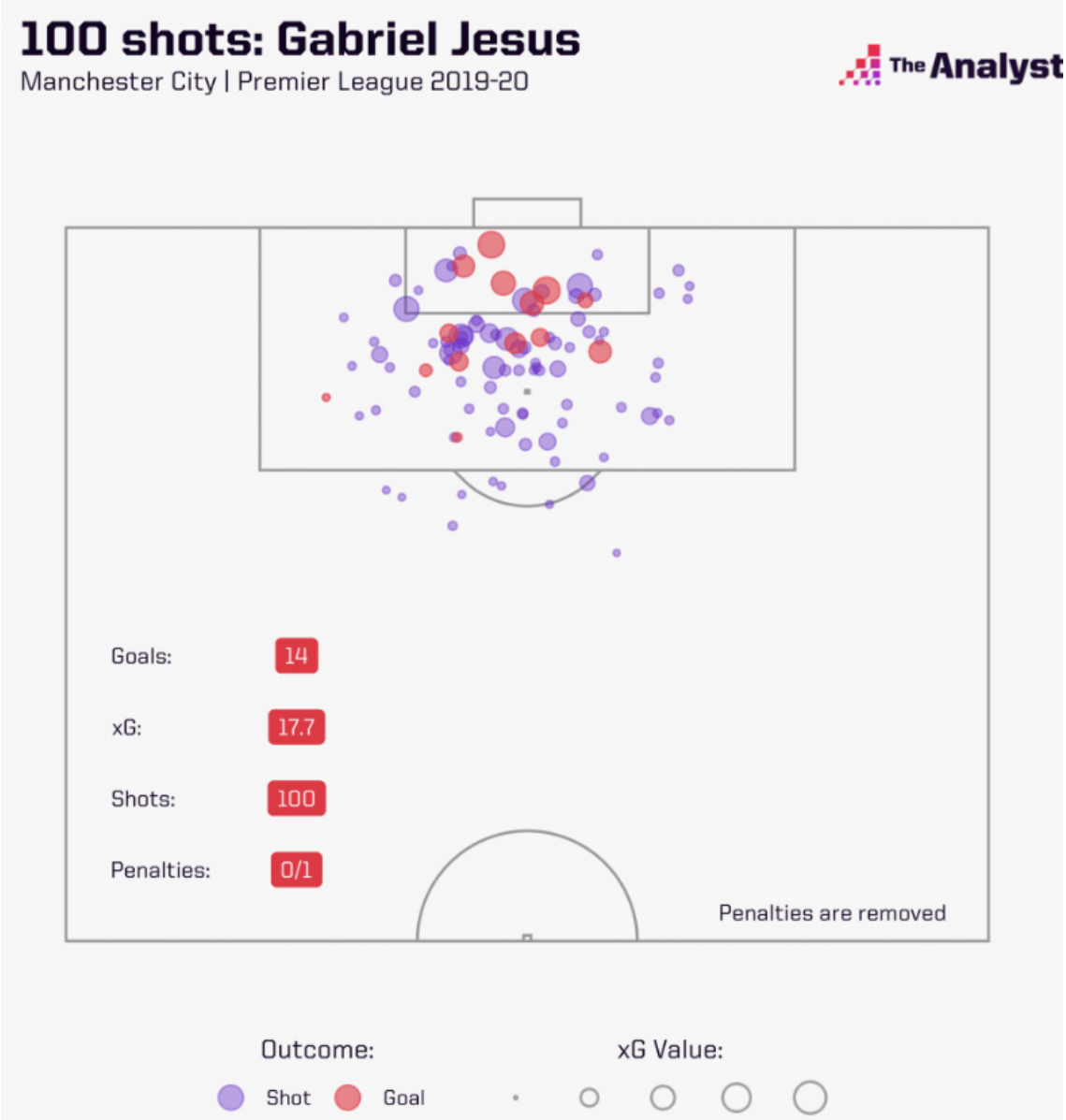


Figure 4.5. *Jesus Shot Profile [8]*

We can compare the shot profiles of the two players by looking at their expected goals per shot (or xG per shot) which values the average quality of a player's scoring chances. Gabriel Jesus' xG per shot was 0.18, meaning that he would be expected to score approximately one goal for every five shots he took. The speculative nature of Calhanoglu's shots resulted in a much lower xG per shot (0.07) that is evident in his shot map above, where the increasing size of the dot indicates an increasing xG value (and

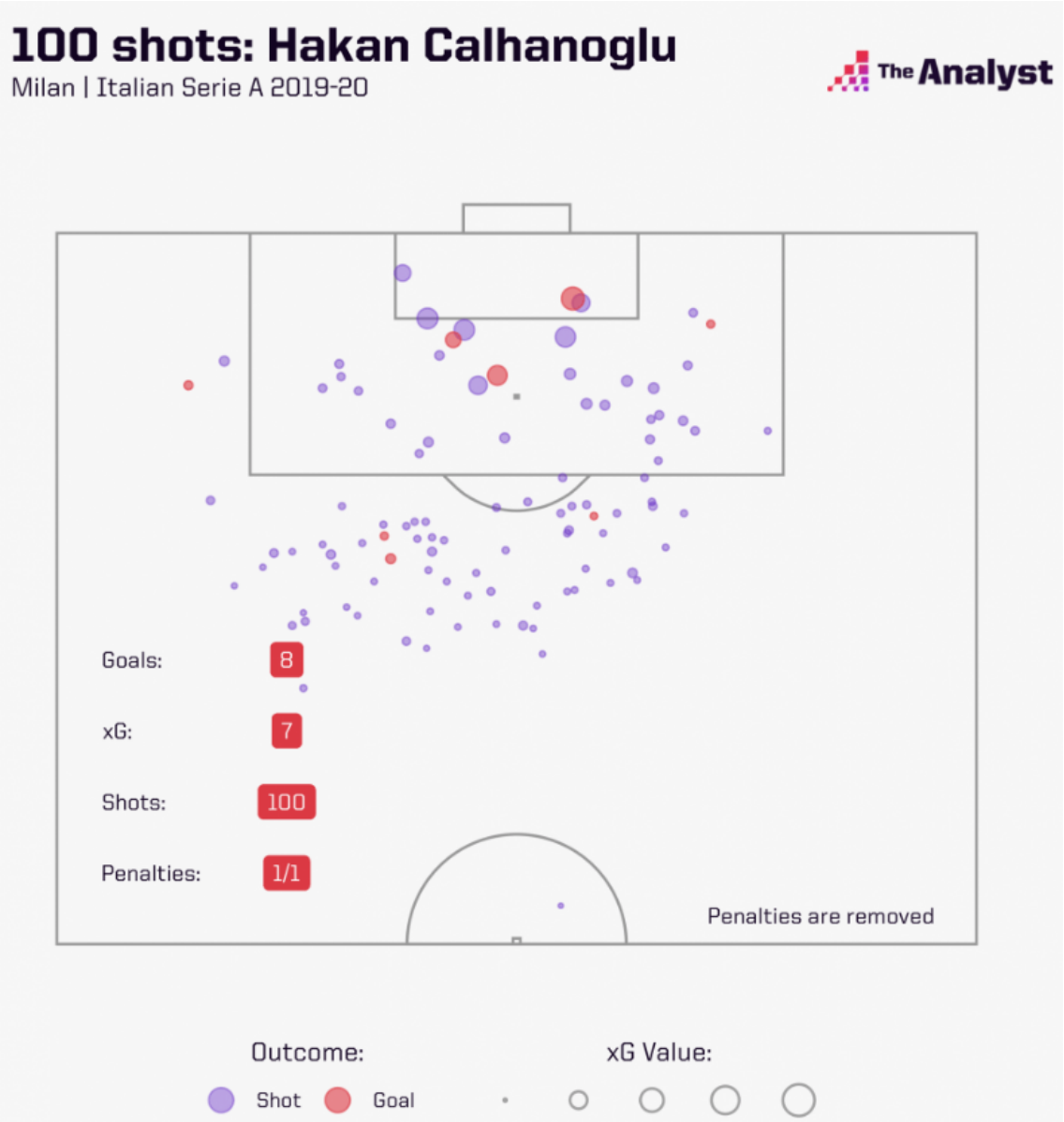


Figure 4.6. *Calhanoglu Shot Profile [8]*

hence a higher likelihood of scoring).

We've focused on an individual player example here, but the expected goals metric can also be applied to teams or games in a similar manner. Of course, we can see here that a player or team may score more or less often than their xG value suggests but this is exactly the variance we can now analyze. Is a player scoring less than he should be? Who is getting chances from high xG situations? [8]

4.3.5 What have xG models taught us?

Many of the things we have learned from xG models are, to a degree, intuitive. Even the most casual supporter could tell you that a shot from inside the six-yard box has a greater chance of producing a goal than a shot from 30 meters out. But what an Expected Goals model provides is a statistical framework to systematically evaluate the

value of each and every shot, and determine how much more likely one is to result in a goal than the other [51].

- **Central Shots Are Best**: Shots from the central part of the penalty area are more valuable than those from tighter angles.
- **Feet Over Head**: From the same distance, foot shots are more likely to result in goals than headed shots.
- **Crosses Are Hard**: In general, crosses are more difficult to convert than ground passes, through-balls and shots after dribbles.
- **Shot Quality is Key**: Given large enough sample sizes, it is possible to identify certain players who stand out for their finishing ability, but the large majority of players are close to average. In general terms, what differentiates good forwards isn't so much finishing chances at an above-average rate but generating shots from valuable locations.

Chapter 5

Regression Analysis

Regression specifically refers to the estimation of a continuous dependent variable or response from a list of input variables, or features.

Regression is a supervised learning technique which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables. In Regression, we plot a graph between the variables which best fits the given data points, using this plot, the machine learning model can make predictions about the data. In simple words, "Regression shows a line or curve that passes through all the data points on target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum." The distance between data points and line tells whether a model has captured a strong relationship or not [9].

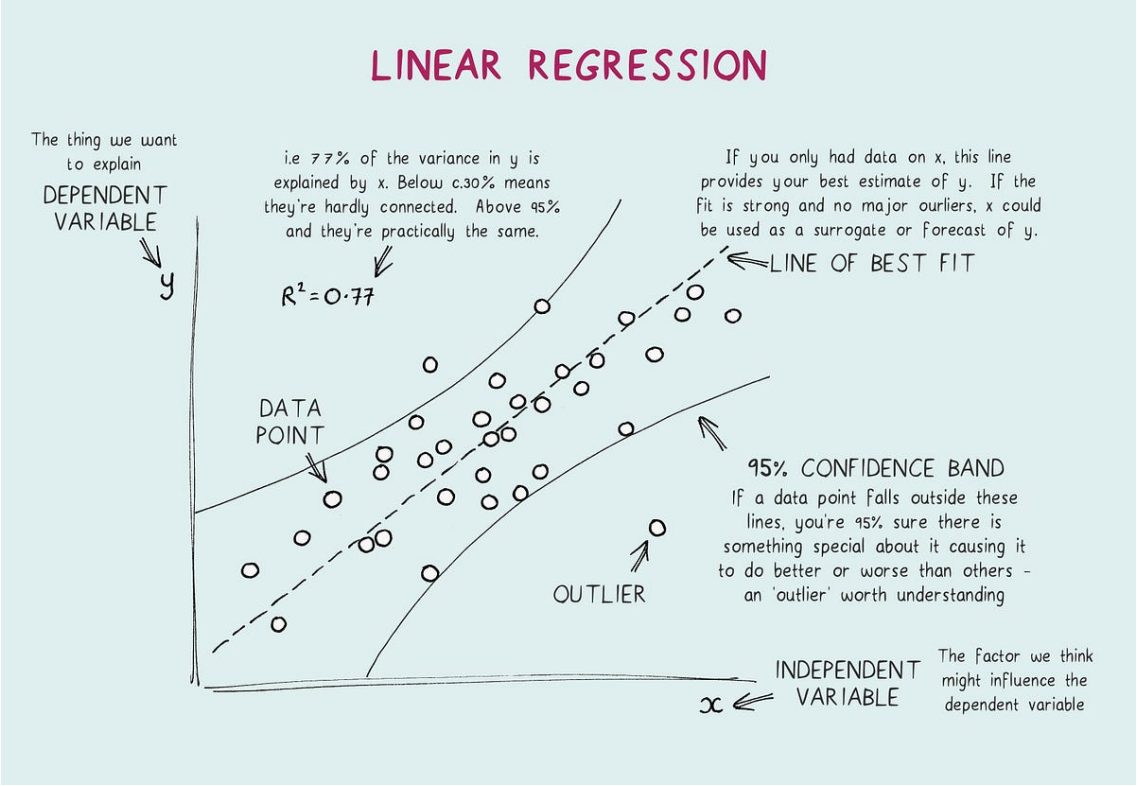


Figure 5.1. *Linear Regression [2]*

In the photo below [Figure 5.2] the main types of regression can be seen. We will then focus on three of those.

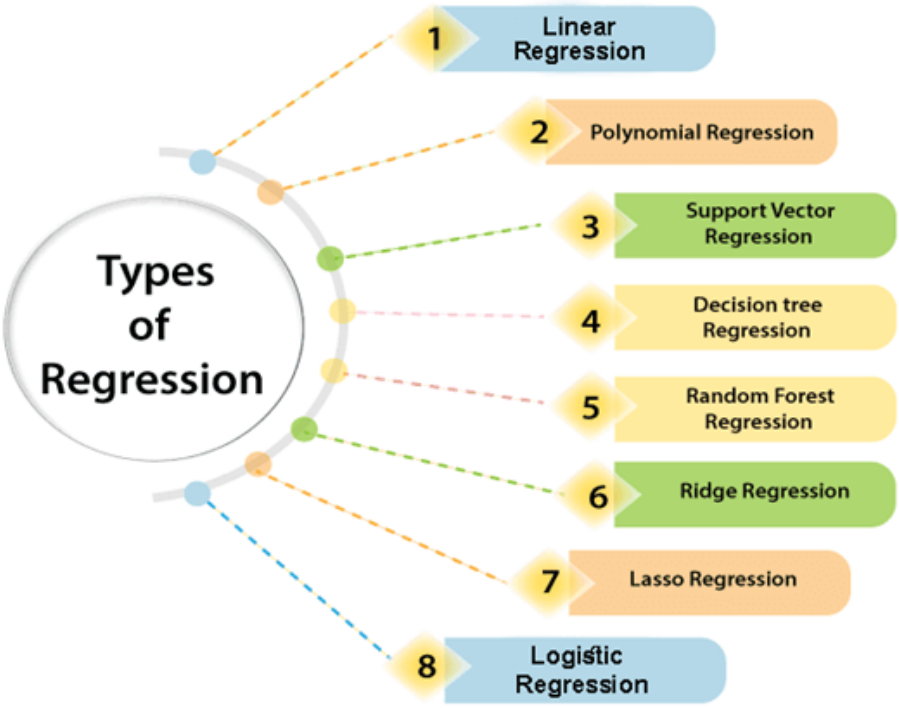


Figure 5.2. *Regression Types [9]*

5.1 Random Forest Regression

5.1.1 What is Random Forest Regression?

Random forest regression is a supervised learning algorithm that uses an ensemble learning method for regression. Random forest is a bagging (bootstrap aggregating) technique. The trees in random forests run in parallel, meaning there is no interaction between these trees while building the trees [Figure 5.3] [13].

5.1.2 Bootstrap Aggregating

Bootstrap aggregating, also called bagging (from bootstrap aggregating), is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting [14].

Given a standard training set D of size n, bagging generates m new training sets D_i, each of size n', by sampling from D uniformly and with replacement. By sampling with replacement, some observations may be repeated in each D_i. If $n' = n$, then for large n the set D_i is expected to have the fraction $(1 - 1/e)$ (about 63.2%) of the unique examples of D, the rest being duplicates. This kind of sample is known as a bootstrap sample. Sampling

with replacement ensures each bootstrap is independent from its peers, as it does not depend on previous chosen samples when sampling. Then, m models are fitted using the above m bootstrap samples and combined by averaging the output (for regression) or voting (for classification). In addition to each tree only examining a bootstrapped set of samples, only a small but consistent number of unique features are considered when ranking them as regressors. This means that each tree only knows about the data pertaining to a small constant number of features, and a variable number of samples that is less than or equal to that of the original dataset. Consequently, the trees are more likely to return a wider array of answers, derived from more diverse knowledge. This results in a random forest, which possesses numerous benefits over a single decision tree generated without randomness [14].

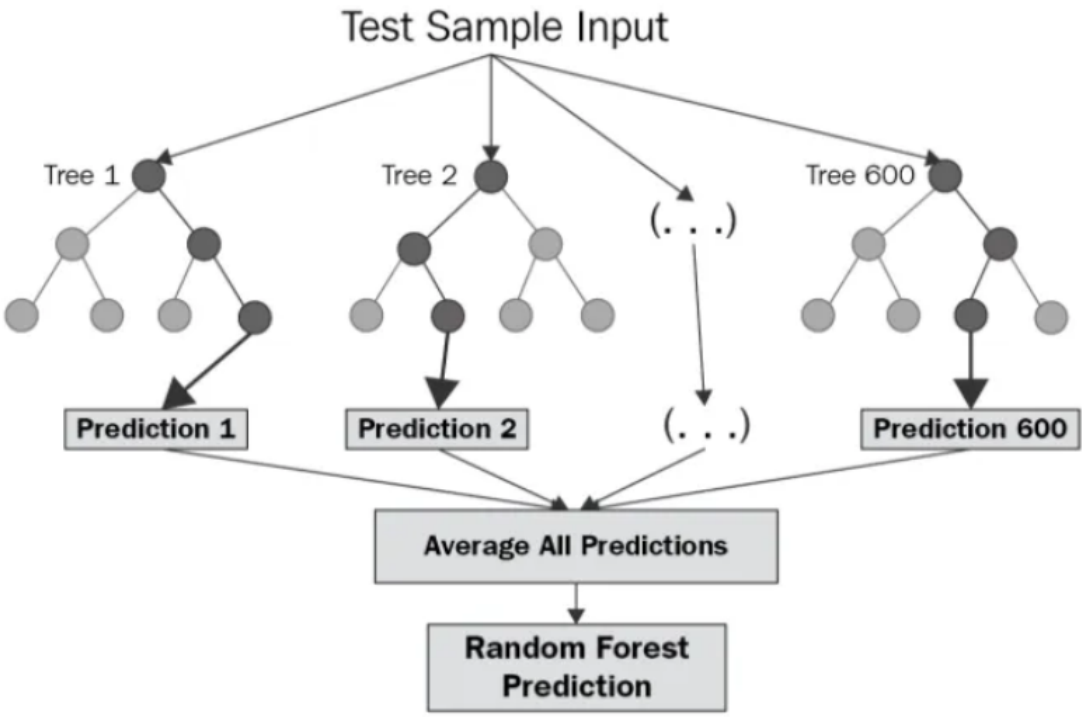


Figure 5.3. *Random Forest Regression [3]*

5.2 XGBoost Regression

XGBoost (eXtreme Gradient Boosting) is an open-source software library which provides a regularizing gradient boosting framework [15].

Tree ensembles are used in Boosted Trees like in Random Forests. The difference between the two methods can be found in the training phase. While Random Forests use the bagging technique that was analyzed above, XGBoost and all the boosted tree models use the boosting technique [16].

5.2.1 Boosting

Boosting is an ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms that convert weak learners to strong ones [17].

While boosting is not algorithmically constrained, most boosting algorithms consist of iteratively learning weak classifiers with respect to a distribution and adding them to a final strong classifier. When they are added, they are weighted in a way that is related to the weak learners' success. After a weak learner is added, the data weights are readjusted, known as "re-weighting". Wrongly predicted input data gain a higher weight and examples that were predicted correctly lose weight. Thus, future weak learners focus more on the examples that previous weak learners predicted in an incorrect way [17].

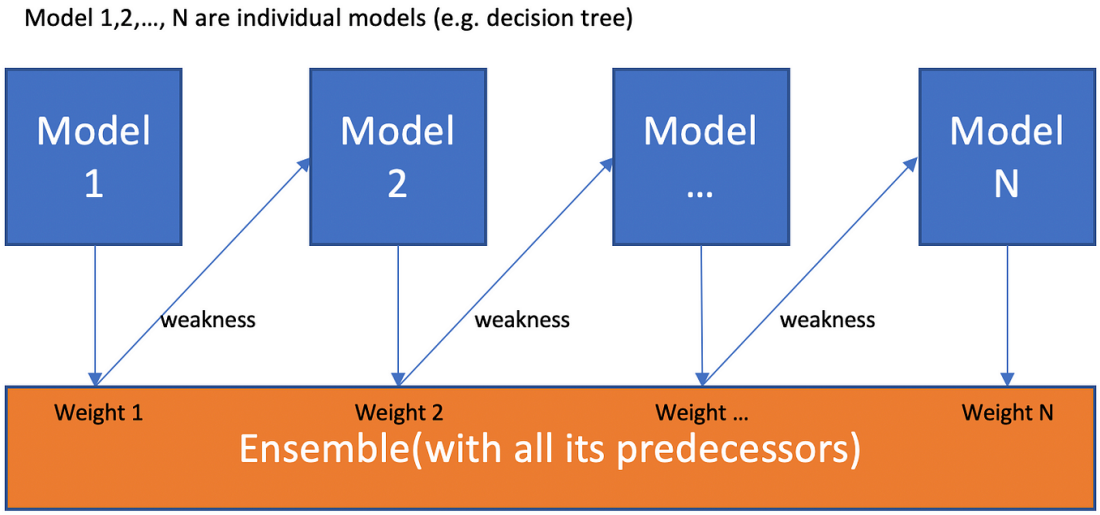


Figure 5.4. *Boosting [10]*

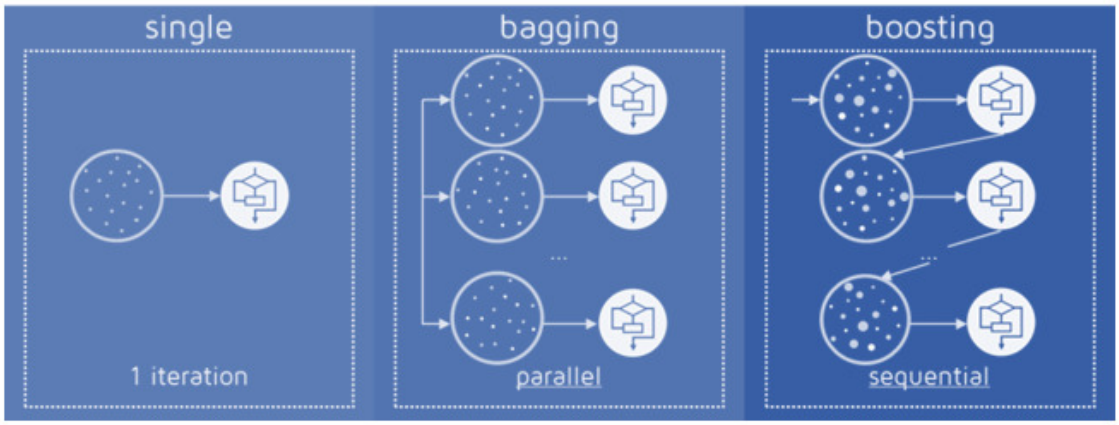


Figure 5.5. *Technique Comparison [4]*

5.3 Ridge Regression

Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where the independent variables are highly correlated. It has been used in many fields including econometrics, chemistry, and engineering. Also known as Tikhonov regularization, named for Andrey Tikhonov, it is a method of regularization of ill-posed problems. It is particularly useful to mitigate the problem of multicollinearity in linear regression, which commonly occurs in models with large numbers of parameters. Ridge regression was developed as a possible solution to the imprecision of least square estimators when linear regression models have some multicollinear (highly correlated) independent variables—by creating a ridge regression estimator (RR). This provides a more precise ridge parameters estimate, as its variance and mean square estimator are often smaller than the least square estimators previously derived [18].

Possibly the most elementary algorithm that can be kernelized is ridge regression. That leads to Kernel Ridge Regression [19].

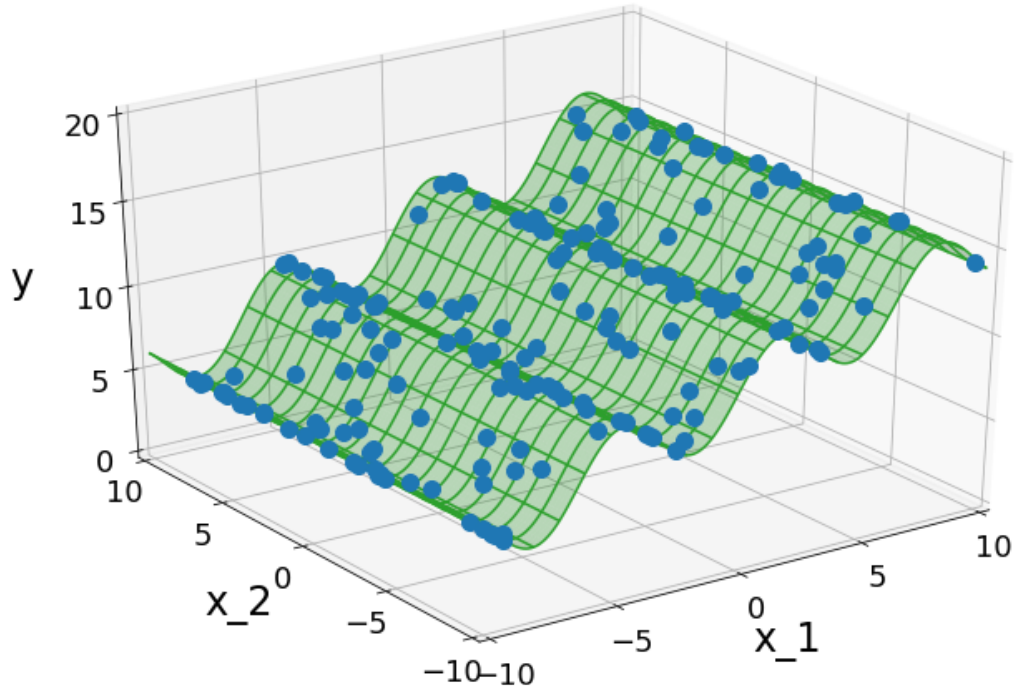


Figure 5.6. *Technique Comparison [11]*

Chapter 6

Related Work

The general inclination observed in previous work has been to mainly use historical statistical data in combination with Naive-Bayes, Random Forests, Boosting Machines and other ensemble methods to predict future player performances.

Using a Gaussian Naive Bayes algorithm, Thapliya was able to predict future player performances with a reported accuracy of 86% [20]. He classified the data labels into two classes: the ones that have scored 8 or more points, and the ones with less than 8.

Using similar training data, Raghunandh found that using ensemble methods such as Gradient Boosting Machines (GBM) could help, in his work trying to suggest potential captaincy picks [21].

In their paper, Bonello, Beel, Lawless and Debattista aimed to investigate whether combining data from different data sources would help in increasing the accuracy and overall performance of predictions. While this idea is promising, they reportedly struggled to incorporate textual data from social media etc., as they encountered difficulties during the implementation of their NLP models. However, they both managed to attain a very good rank of 30,000 out of 6.5 million players and found that GBMs are the most effective in this case, particularly due to the unbalanced data [22].

Bonomo et Al. developed a mathematical optimization model using integer linear programming to predict ideal line ups every gameweek in Argentinian Football League. They used historical data combined with information from the managers' press conference before matches were played. They used their model on posteriori stats to determine factors that could possibly help in building predictive models [23].

Pokharel, Timalsina, Panday and Acharya used XGBoost regression and focused on ROI to predict players' performance, while also examining the effect of additional data from midweek cup games. They managed a mean RMSE score of 2.048 for all players [24].

In more sophisticated models, Matthews, Ramchurn and Chalkiadakis presented an innovative fantasy football predictor that consisted of belief-state MDP models combined with Bayesian Q—Learning algorithms to train models on the past five years of football data [25]. Their most successful model used a state of the art Bayesian Q-learning model to handle the uncertainty, placing the machine within the top 500 players out of the 2.5million participants (0.01%). The original model was naive; acting myopically only considering the single next gameweek. Even with all these restrictions, this model still

achieved a very respectable rank of 113,921. By extending this model to also look back on data from the previous season, the model leaves out all the players who did not appear but still improved performance slightly, upping the rank to 60,633.

Over the past years the FPL community has produced many laudable projects in the field. One of them is undoubtedly FPL Kiwi's model [26]. He uses expected data (expected Goals, expected Assists etc.) from the past 5 seasons and breaks down the problem into smaller ones, with each smaller problem being the prediction of a separate event that leads to points. He uses spreadsheets to implement his detailed calculations, has his own model of calculating team strengths and uploads his projections to help the FPL community [27].

A lot of fantasy websites have come up with their own models recently too. FPL Review [28], Fantasy Football Scout [29], Fantasy Football Hub [30], Fantasy Football Fix [31] are all examples of the recent data boom in fantasy football.

6.1 AILS LAB - NTUA

Machine learning, including deep neural network architectures have been implemented and used in various applications by members of the NTUA Artificial Intelligence and Learning Systems Laboratory. In particular supervised CNN and CNN-RNN techniques have been applied for object categorization, in the medical diagnosis of neurodegenerative diseases, such as Parkinson's disease [52, 53, 54, 55, 56] or Covid-19 [57, 58, 59, 60, 61], involving 2-D or 3-D medical images. Emphasis has been placed on transparency and adaptation of models [62, 63, 64], but also on development of more complex architectures, i.e., Bayesian, Capsules and ones involving Uncertainty estimation [65, 66, 67, 68]. Deep self-supervised neural architectures, as well as encoder-decoder architectures have been applied to nuclear reactor fault detection [69, 70, 71], agricultural production forecasting [72, 73], recognition and synthesis of emotions [74, 75, 76, **?**, 77], while others are applied to image analysis and human-computer interaction problems [78, 79, 80, 81, 82, 83, 84, 85].

Chapter 7

Data

In this chapter, we will explore all the data related problems and processes I encountered on the way to the final results. I will explain how the data was gathered and manipulated to extract all those insights necessary to tackle our problem. The data flow diagram below [Figure 7.1] gives an overview of the process.

7.1 Important Decision

Predicting a player's value for the next gameweeks in FPL is a problem with no dataset ready to use. However, our experience from the sport and the game leads us to metrics that can be used to predict this value. Therefore, I had to create my own datasets and experiment with different feature sets in order to achieve the best possible results. After personal experimentation, conversations with members of the FPL community and studying older related work, I decided that it made sense to break down the problem of predicting each player's value into smaller subproblems [Figure 7.2]. For every action/event in the match that a player produces points, there's a different subproblem to be solved. The main ones that require their own models are the following:

- Non penalty Goals
- Assists
- Team Goals
- Penalties
- Saves
- Bonus Points

Non penalty Goals and penalties, that only add value to each team's penalty takers, constitute the Goals component of the player's final value. This translates to all players (with the goalkeepers the least likely to score). Team Goals, for example, are used to compute the probability of clean sheets (using the Poisson process [32]), that give points to specific players depending on their position. Saves are used to calculate save points for goalkeepers only, while assists and bonus points are for every player.

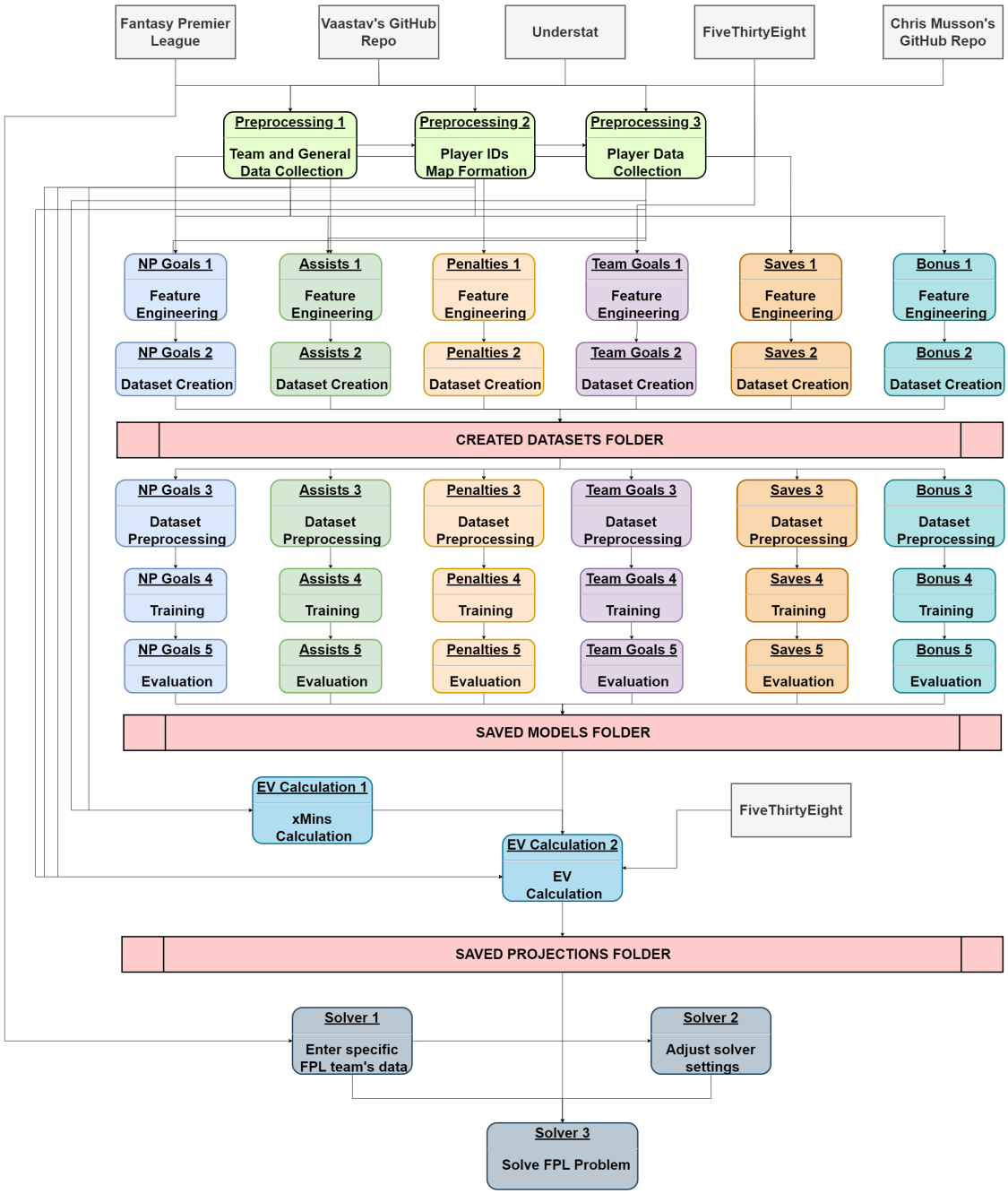


Figure 7.1. *The Data Flow Diagram of the Project*

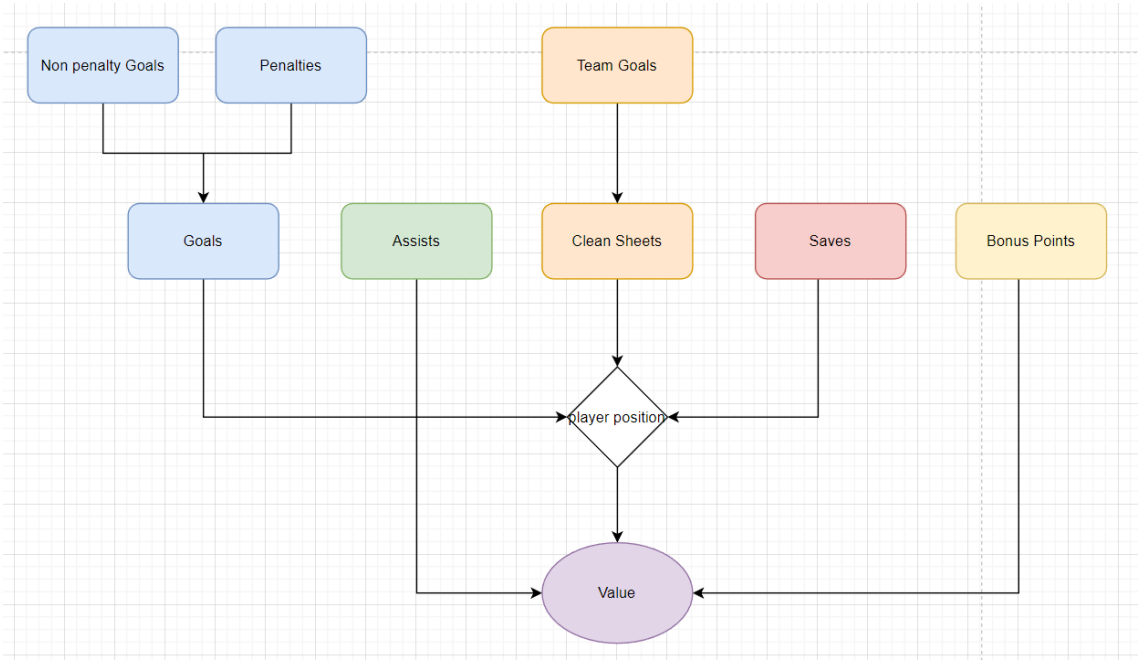


Figure 7.2. *Player's Value Components*

7.2 Data Sources

Although no specific dataset was available to achieve my goal, raw data was not difficult to find. Here, I will present the sources that the raw data were collected from.

7.2.1 Fantasy Premier League API

This is the API of the official game. It provides extremely useful data for the project such as fixture data, player data of the past gameweeks, and even statistical info on players' past seasons. Other than that, you can make requests to the official API to get a specific player's team data, or info about the leagues he/she is in [33].

7.2.2 Understat

Understat is a site that provides detailed xG statistics for the top European leagues. Their goal was to create the most precise method for shot quality evaluation. For this case, they trained neural network prediction algorithms with the large dataset (>100,000 shots, over 10 parameters for each [34]). I used Understat for quickly updated player and team xG data for the Premier League. Furthermore, understat provided data such as non-penalty goals, shots and key passes for a player per match, that I found particularly useful for my goal. I use the understat package [35] to access the data.

7.2.3 Vaastav's GitHub Repository

The Fantasy Premier League API provides data for the current season only. This would make it impossible to gather past season data to train our models with. Vaastav's repo solves that problem by providing past season data from the FPL API and from understat

[36]. This repository was extremely valuable for the project, as it is the basis of all the created datasets.

7.2.4 Chris Musson's ID Map

The Fantasy Premier League API doesn't have the same id for the same player as the Understat API, or even the same name in a lot of cases. In a new FPL season a player's id doesn't necessarily remain the same causing obvious mapping problems when trying to combine the FPL API with understat for past seasons. Chris Musson's repository solves the id mapping problem, as he has mapped the FPL ids with the Understat ids for every player for several past seasons [37].

7.2.5 FiveThirtyEight

FiveThirtyEight's SPI (Soccer Power Index) ratings and projected goals scored metrics were used as well. These metrics were used in different models to provide insight on a team's strength at a specific point in time relative to the opponent. FiveThirtyEight's Club Soccer Predictions dataset was used to incorporate those metrics [38]. At the start of a season a team's SPI is influenced by the SPI they finished last season with and the market value of the team. During a season, a team's SPI changes based on the team's performances, and more specifically based on the team's adjusted goals, shot based expected goals and non-shot expected goals after each game [39]. SPI ratings help calculate projected goals for each game and using the Poisson Process [32] calculate the probability of a team winning.

7.3 Non-Penalty Goals Dataset

From the experience of the domain and conversations with members of the FPL community, I decided to include three main feature categories: player ability, team strength, and opponent team strength. The final dataset is 51554 rows $\times$ 14 columns.

7.3.1 Features

The features are presented below, and we assume that $action_i$ is the action or event performed by the player or team i games back:

- **npg_ratel100:** The non-penalty goal rate of a player per 90 minutes in his last 100 matches (or less if data isn't available). This metric gives useful information about the player's actual ability to score over a big sample, and can even hopefully provide insights about a player's finishing ability when combined with npxGp90l100.

$$npg\_ratel100 = \frac{90 * \sum_{i=1}^{100} non\_penaltyGoals_i}{\sum_{i=1}^{100} minutes\_played_i}$$

- **npxGp90l100:** The non-penalty expected goals rate per 90 minutes of a player in his last 100 matches (or less if data isn't available). This metric gives an estimate

of the player's chance quality per 90 minutes over a big sample. Of course this can contain matches from different teams for a player, or even different positions and tactical systems, but the average can be a pretty good indicator of future chance quality.

$$npxGp90l100 = \frac{90 * \sum_{i=1}^{100} non\_penalty\_expectedGoals_i}{\sum_{i=1}^{100} minutes\_played_i}$$

- **sh_ratel100:** The shot rate per 90 minutes of a player in his last 100 matches (or less if data isn't available). This metric gives us information about the amount of a player's shots over a big sample of matches, as well as qualities such as willingness to shoot etc.

$$sh\_ratel100 = \frac{90 * \sum_{i=1}^{100} shots_i}{\sum_{i=1}^{100} minutes\_played_i}$$

- **npxGp90:** The non-penalty expected goals rate per 90 minutes of a player in the current season. The *current_GW* is the current gameweek of the season. As mentioned above, the use of this metric is to quantify the quality of the chances a player gets. Here, we get this info for the current season. Therefore, it can both be easily influenced by variance if we have only a few gameweeks of the current season, and contain useful information, such as a team's uptick in attacking performances due to a tactical change this season.

$$npxGp90 = \frac{90 * \sum_{i=1}^{current\_GW-1} non\_penalty\_expectedGoals_i}{\sum_{i=1}^{current\_GW-1} minutes\_played_i}$$

- **npg_rate:** The non-penalty goal rate of a player per 90 minutes in the current season. The *current_GW* is the current gameweek of the season. This metric gives useful information about the player's actual ability to score over a smaller sample, and specifically the current season. It tends to contain information about a player's scoring form.

$$npg\_rate = \frac{90 * \sum_{i=1}^{current\_GW-1} non\_penaltyGoals_i}{\sum_{i=1}^{current\_GW-1} minutes\_played_i}$$

- **shp90:** The shot rate per 90 minutes of a player in the current season. The *current_GW* is the current gameweek of the season. shp90 indicates the willingness of a player to shoot and is limited to the current season.

$$shp90 = \frac{90 * \sum_{i=1}^{current\_GW-1} shots_i}{\sum_{i=1}^{current\_GW-1} minutes\_played_i}$$

- **npxGp90(L4):** The non-penalty expected goals rate per 90 minutes of a player in his last 4 matches. This metric is easily influenced by variance, as it is calculated by the last 4 performances. However, it can contain really useful signals, for example if a player changes position and dramatically improves his chance quality.

$$npxGp90(L4) = \frac{90 * \sum_{i=1}^{4} non\_penalty\_expectedGoals_i}{\sum_{i=1}^{4} minutes\_played_i}$$

- **teamnpxGp90:** The average non-penalty expected goals the player's team produces in the current season. The *current_GW* is the current gameweek of the season. This is a team's metric and it is used to take into account the team's attacking strength this season.

$$teamnpxGp90 = \frac{\sum_{i=1}^{current\_GW-1} team's\_non\_penalty\_expectedGoals_i}{current\_GW - 1}$$

- **spi_team:** The Soccer Power Index rating of the player's team as calculated by FiveThirtyEight [39]. This is another team's metric and gives information about the team's general strength.

- **opp_npxGAp90:** The average non-penalty expected goals the player's opponent team concedes in the current season. The *current_GW* is the current gameweek of the season. This is a metric about the opponent team's strength and it is used to take into account the opponent team's defensive strength this season.

$$opp\_npxGAp90 = \frac{\sum_{i=1}^{current\_GW-1} opponent\_team's\_non\_penalty\_expectedGoals\_conceded_i}{current\_GW - 1}$$

- **spi_opp_team:** The Soccer Power Index rating of the player's opponent team as calculated by FiveThirtyEight [39]. This is another metric about the opponent team's overall strength.

- **minutes:** The minutes the player played in that specific match. This feature may be self-explanatory, but its importance cannot be overstated. The amount a player is involved in a game, or not, directly and greatly impacts his expected value.

- **was_home:** True if the player's team was playing in their home arena, and False if they were playing on the road. A metric for our model to take into account the possible advantage of playing with your home crowd.

- **npg:** The non-penalty Goals the player scored in that specific match. This is the label, and the number we are trying to predict.

7.3.2 Dataset Exploration

As mentioned earlier, goals in soccer, and more specifically non-penalty goals, are generally considered rare events. The average goals in a soccer match are between 2 and 3 goals. However, over 30 players can feature during a match. Furthermore, there are more players that weren't selected to play any minutes. Therefore it is easy to understand that our dataset is dominated by samples, where the player hasn't scored. This is demonstrated in the below graph [Figure 7.3].

Another thing that was discovered after exploring the dataset, and the problem in general, is that variance is dominant in the non-penalty goal prediction problem. Although we have some very good metrics to approach the problem, it is by nature a high-variance problem. We see in the scatterplot below that, in spite of the fact that most of the goals are

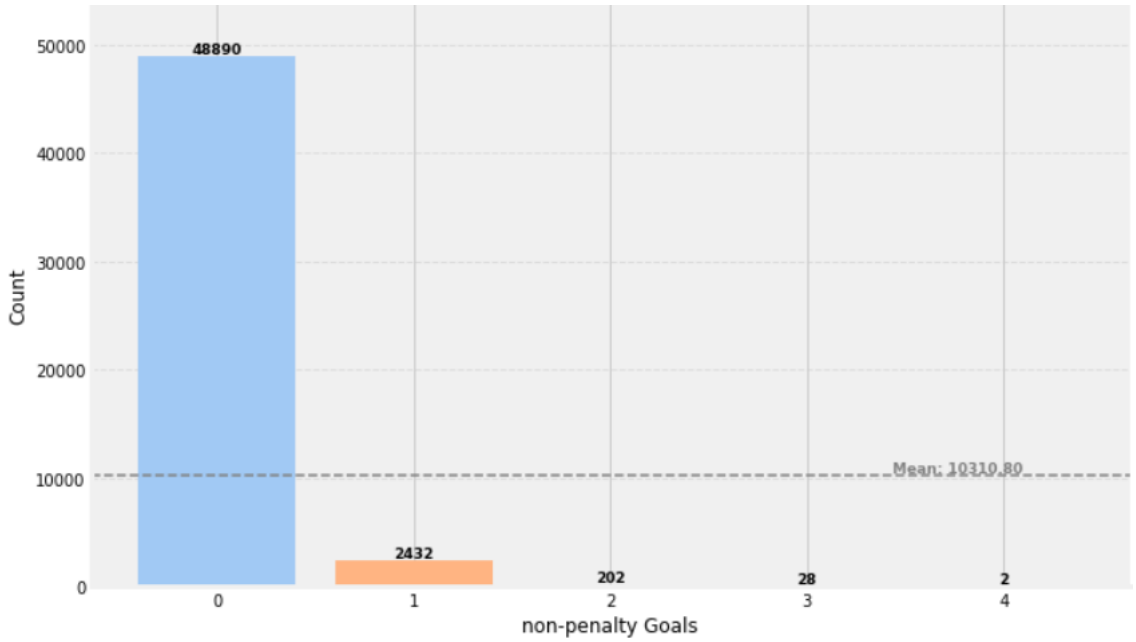


Figure 7.3. *Non-penalty Goals Dataset Barplot*

scored against teams with a 60-80 rating, which is to be expected as this range contains the most teams, goals are pretty random [Figure 7.4].

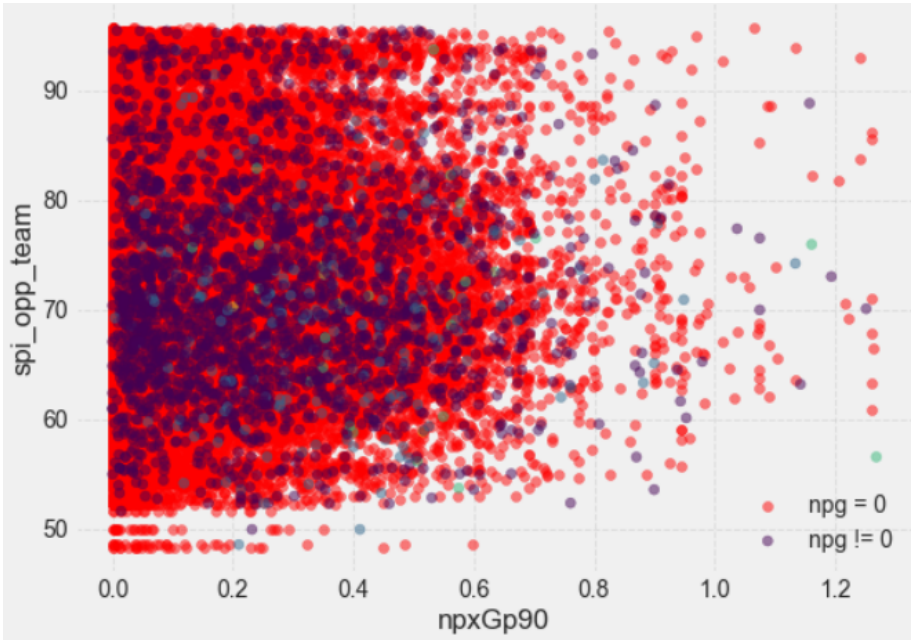


Figure 7.4. *Scatterplot of npxGp90 vs spi_opp_team*

However, not all hope is lost. We can see an obvious correlation between the npxGp90 and npg_rate metrics, especially when the sample size gets bigger [Figure 7.5].

The above correlation can be seen really well when we examine what percentage of players scored a non-penalty goal depending on their npxGp90 bracket [Figure 7.6]. We take into account players that played the greatest part of a match (>75 minutes).

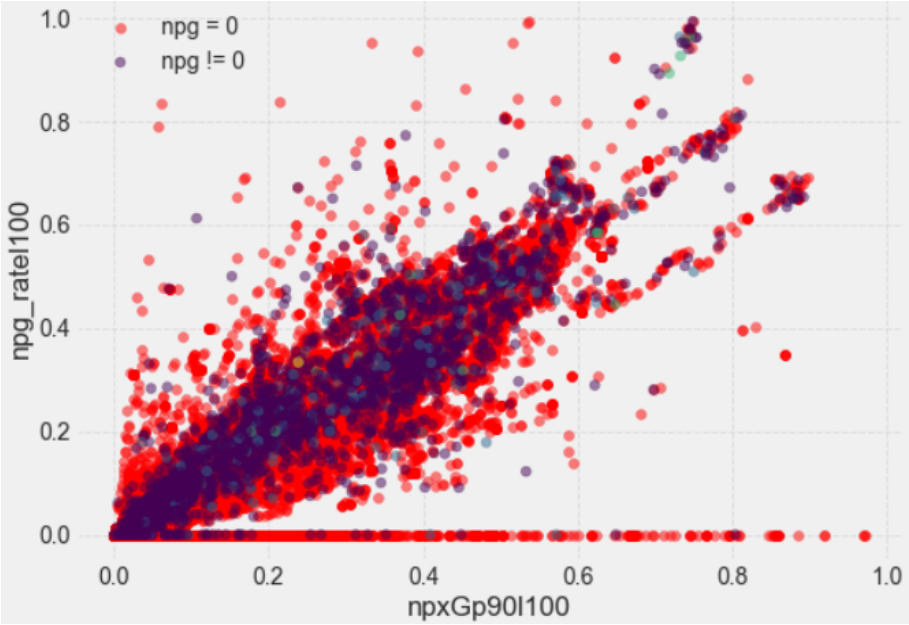


Figure 7.5. *Correlation between npxGp90 and npg_rate over a big sample*

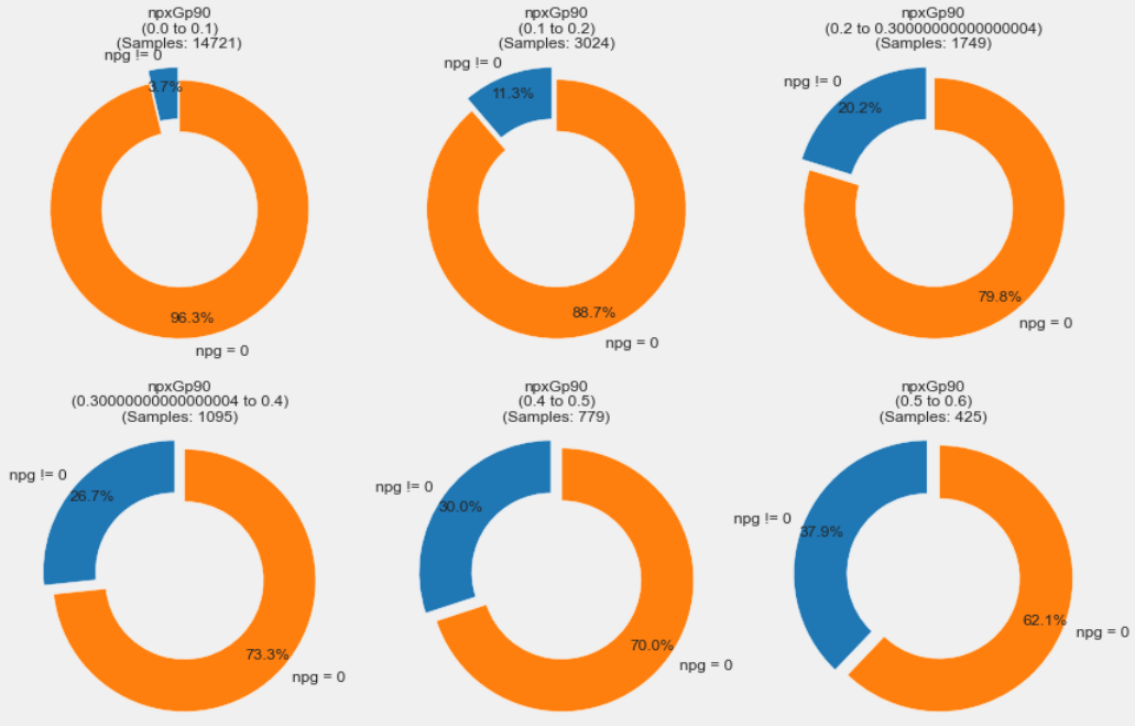


Figure 7.6. *Percentage of players that scored a non-penalty goal depending on their npxGp90 metric*

7.4 Assists Dataset

Just like with the non-penalty goals dataset, I decided to include three main feature categories: player ability, team strength, and opponent team strength. The final dataset is 51554 rows × 15 columns. The assists dataset is really similar to the non-penalty goals dataset we explored above, but with key differences.

7.4.1 Features

The features are presented below, and we again assume that $action_i$ is the action or event performed by the player or team i games back:

- **assist_ratel100:** The assist rate of a player per 90 minutes in his last 100 matches (or less if data isn't available). This metric gives useful information about the player's actual ability to create goals over a big sample.

$$assist\_ratel100 = \frac{90 * \sum_{i=1}^{100} Assists_i}{\sum_{i=1}^{100} minutes\_played_i}$$

- **xAp90l100:** The expected assists rate per 90 minutes of a player in his last 100 matches (or less if data isn't available). This metric gives an estimate of the player's chance creation per 90 minutes over a big sample. Of course this can contain matches from different teams for a player, or even different positions and tactical systems, but the average can be a pretty good indicator of future chance creation by the player.

$$xAp90l100 = \frac{90 * \sum_{i=1}^{100} expectedAssists_i}{\sum_{i=1}^{100} minutes\_played_i}$$

- **kp_ratel100:** The key passes rate per 90 minutes of a player in his last 100 matches (or less if data isn't available). This metric gives us information about the amount of a player's key passes that often lead to big chances or assists, over a big sample of matches, as well as qualities such as willingness to create etc.

$$kp\_ratel100 = \frac{90 * \sum_{i=1}^{100} key\_passes_i}{\sum_{i=1}^{100} minutes\_played_i}$$

- **xAp90:** The expected assists rate per 90 minutes of a player in the current season. The *current_GW* is the current gameweek of the season. As mentioned above, the use of this metric is to quantify the quality of the chances a player creates. Here, we get this info for the current season. Therefore, it can both be easily influenced by variance if we have only a few gameweeks of the current season, and contain useful information, such as a team's uptick in attacking performances due to a tactical change this season.

$$xAp90 = \frac{90 * \sum_{i=1}^{current\_GW-1} expectedAssists_i}{\sum_{i=1}^{current\_GW-1} minutes\_played_i}$$

- **assist_rate:** The assist rate of a player per 90 minutes in the current season. The *current_GW* is the current gameweek of the season. This metric gives useful information about the player's actual ability to create over a smaller sample, and specifically the current season. It tends to contain information about a player's

chance creation form.

$$assist\_rate = \frac{90 * \sum_{i=1}^{current\_GW-1} Assists_i}{\sum_{i=1}^{current\_GW-1} minutes\_played_i}$$

- **kpp90:** The key passes rate per 90 minutes of a player in the current season. The *current_GW* is the current gameweek of the season. kpp90 indicates the willingness of a player to create chances and is limited to the current season.

$$kpp90 = \frac{90 * \sum_{i=1}^{current\_GW-1} key\_passes_i}{\sum_{i=1}^{current\_GW-1} minutes\_played_i}$$

- **xAp90(L4):** The expected assists rate per 90 minutes of a player in his last 4 matches. This metric is easily influenced by variance, as it is calculated by the last 4 performances. However, it can contain really useful signals, for example if a player changes position and dramatically improves his chance creation.

$$xAp90(L4) = \frac{90 * \sum_{i=1}^{4} expectedAssists_i}{\sum_{i=1}^{4} minutes\_played_i}$$

- **teamnpxGp90:** The average non-penalty expected goals the player's team produces in the current season. The *current_GW* is the current gameweek of the season. This is a team's metric and it is used to take into account the team's attacking strength this season.

$$teamnpxGp90 = \frac{\sum_{i=1}^{current\_GW-1} team's\_non\_penalty\_expectedGoals_i}{current\_GW - 1}$$

- **spi_team:** The Soccer Power Index rating of the player's team as calculated by FiveThirtyEight [39]. This is another team's metric and gives information about the team's general strength.

- **opp_npxGp90:** The average non-penalty expected goals the player's opponent team produces in the current season. The *current_GW* is the current gameweek of the season. This is a metric about the opponent team's strength and it is used to take into account the opponent team's attacking strength this season.

$$opp\_npxGp90 = \frac{\sum_{i=1}^{current\_GW-1} opponent\_team's\_non\_penalty\_expectedGoals_i}{current\_GW - 1}$$

- **opp_npxGAp90:** The average non-penalty expected goals the player's opponent team concedes in the current season. The *current_GW* is the current gameweek of the season. This is a metric about the opponent team's strength and it is used to take into account the opponent team's defensive strength this season.

$$opp\_npxGAp90 = \frac{\sum_{i=1}^{current\_GW-1} opponent\_team's\_non\_penalty\_expectedGoals\_conceded_i}{current\_GW - 1}$$

- **spi_opp_team:** The Soccer Power Index rating of the player's opponent team as calculated by FiveThirtyEight [39]. This is another metric about the opponent team's overall strength.
- **minutes:** The minutes the player played in that specific match. This feature may be self-explanatory, but its importance cannot be overstated. The amount a player is involved in a game, or not, directly and greatly impacts his expected value.
- **was_home:** True if the player's team was playing in their home arena, and False if they were playing on the road. A metric for our model to take into account the possible advantage of playing with your home crowd.
- **assists:** The assists the player had in that specific match. This is the label, and the number we are trying to predict.

7.4.2 Dataset Exploration

An assist is the pass that comes before the goal. So, in order to be an assist, it needs to be a goal. Therefore, it is easy to understand that our dataset is dominated by samples, where the player hasn't assisted (just like with the previous dataset). This is demonstrated in the below graph [Figure 7.7].

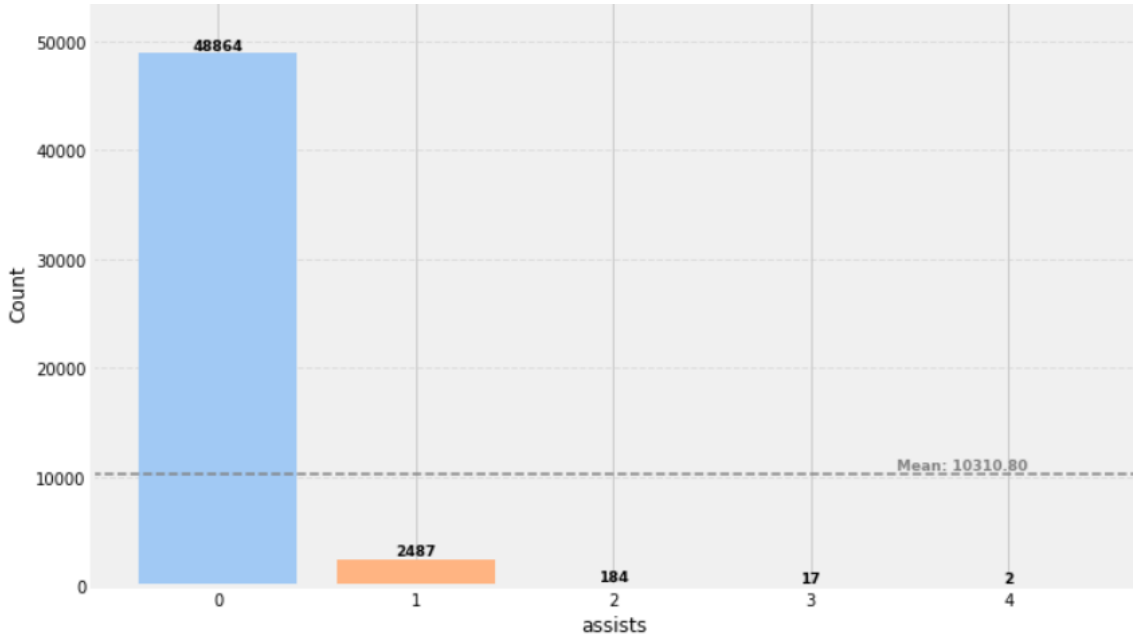


Figure 7.7. *Assists Dataset Barplot*

Assists contain even more variance than non-penalty goals. This is easily intuitive, because for an assist to happen the player who provides the potential assist depends on the player who receives it to score. This extra step in comparison to the non-penalty goal process makes it more random. Furthermore, the game rewards assists for some semi-random events that lead to goals, such as winning a penalty or shooting and someone scoring from the rebound. Although we have some very good metrics to approach the

problem, it is by nature an extremely high-variance problem. We see in the scatterplot below that, in spite of the fact that most of the assists are against teams with a 60-80 rating, which is to be expected as this range contains the most teams, they are pretty random [Figure 7.8].

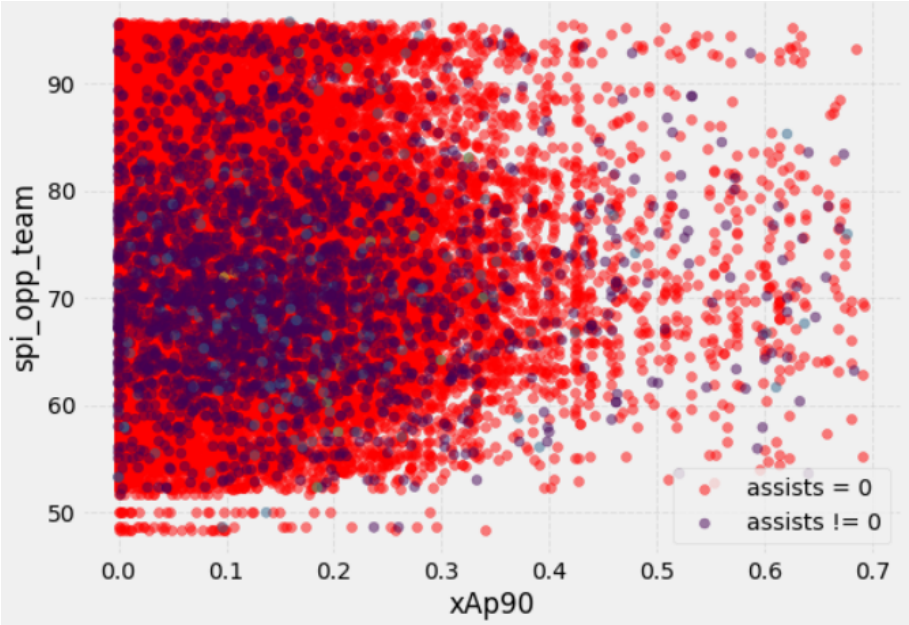


Figure 7.8. *Scatterplot of xAp90 vs spi_opp_team*

Just like we observed earlier, we can see an obvious correlation between the xAp90 and assist_rate metrics, especially when the sample size gets bigger [Figure 7.9].

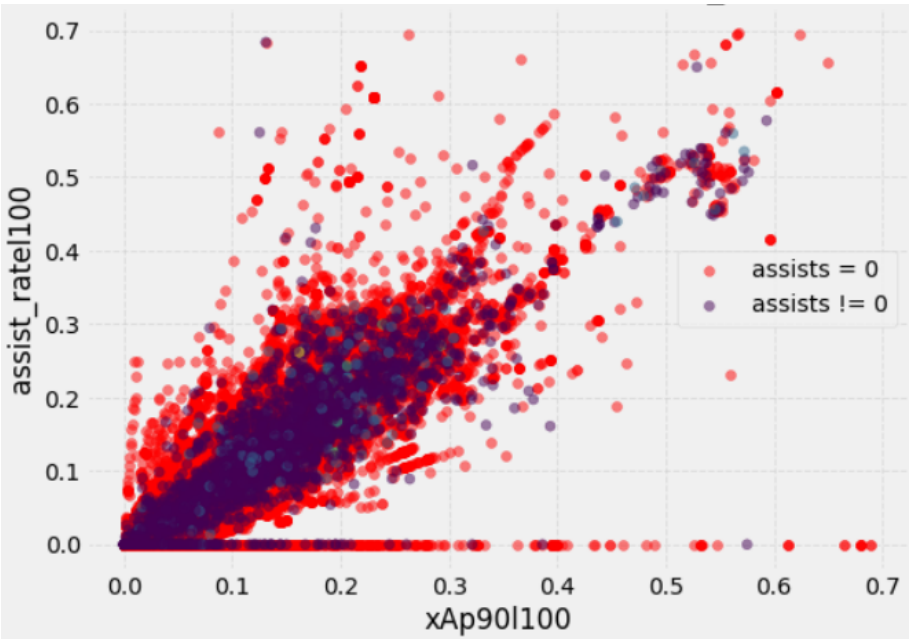


Figure 7.9. *Correlation between xAp90 and assist_rate over a big sample*

The above correlation can be seen really well when we examine what percentage of players that assisted depending on their xAp90 bracket [Figure 7.10]. We take into ac-

count players that played the greatest part of a match (>75 minutes).

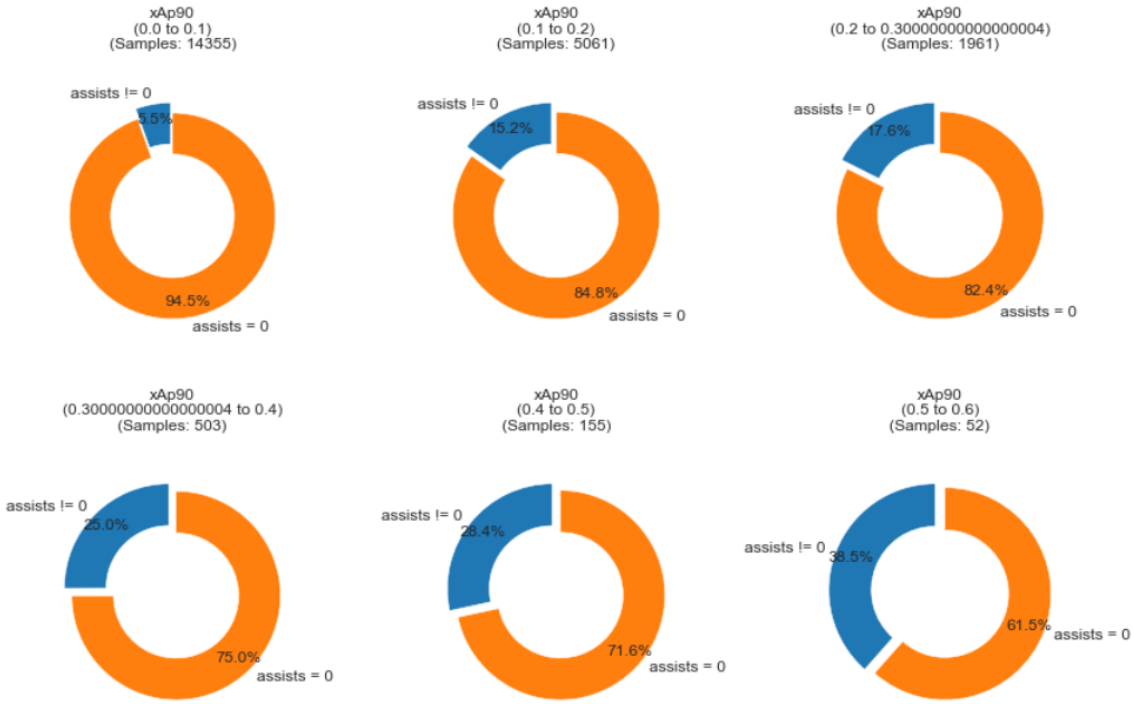


Figure 7.10. *Percentage of players that assisted a goal depending on their xAp90 metric*

7.5 Penalties Dataset

The penalties dataset is created to predict the expected penalties a team will have in the next matches. Penalties are considered almost random events in soccer matches, so I tried to create a dataset that will correspond the best way possible to a difficult task: predicting penalties. It is noteworthy that the information our eventual model will provide, will be used to get the expected penalties of each team. I have implemented a penalty hierarchy for each team (the penalty takers and their order for each team) that will translate this information into expected points for each team's penalty takers. The dataset is 2658 rows $\times$ 8 columns.

7.5.1 Features

The features are presented below, and we again assume that $action_i$ is the action or event performed by the team i games back:

- **team_npxG:** The average non-penalty expected goals the team produces in the current season. The *current_GW* is the current gameweek of the season.

$$team\_npxG = \frac{\sum_{i=1}^{current\_GW-1} team's\_non\_penalty\_expectedGoals_i}{current\_GW - 1}$$

- **team_npxGp90(L4):** The average non-penalty expected goals the team produces taking into account the last 4 matches.

$$team\_npxGp90(L4) = \frac{\sum_{i=1}^{4} team's\_non\_penalty\_expectedGoals_i}{4}$$

- **pen_rate:** The average penalty rate the team has over the course of the current season.

$$pen\_rate = \frac{\sum_{i=1}^{current\_GW-1} team's\_penalties_i}{current\_GW - 1}$$

- **oppteam_npxGA:** The average non-penalty expected goals the opponent team concedes in the current season. The *current_GW* is the current gameweek of the season. This is a metric about the opponent team's strength and it is used to take into account the opponent team's defensive strength this season.

$$oppteam\_npxGA = \frac{\sum_{i=1}^{current\_GW-1} opponent\_team's\_non\_penalty\_expectedGoals\_conceded_i}{current\_GW - 1}$$

- **oppteam_npxGAp90(L4):** The average non-penalty expected goals the opponent team concedes during the last 4 matches. The *current_GW* is the current gameweek of the season.

$$oppteam\_npxGAp90(L4) = \frac{\sum_{i=1}^{4} opponent\_team's\_non\_penalty\_expectedGoals\_conceded_i}{4}$$

- **proj_goals:** The projected goals the team will score against the opponent team as calculated by FiveThirtyEight [39]. This is another metric about the overall strength differential between the teams.

- **was_home:** True if the team was playing in their home arena, and False if they were playing on the road. A metric for our model to take into account the possible advantage of playing with your home crowd.

- **team_pens:** The penalties the team won in that specific match. This is the label, and the number we are trying to predict.

7.5.2 Dataset Exploration

As mentioned above, penalties are rare, high variance events [Figure 7.11]. However, it is only logical that the more goals a team is projected to score the more penalties it is expected to get [Figure 7.11].

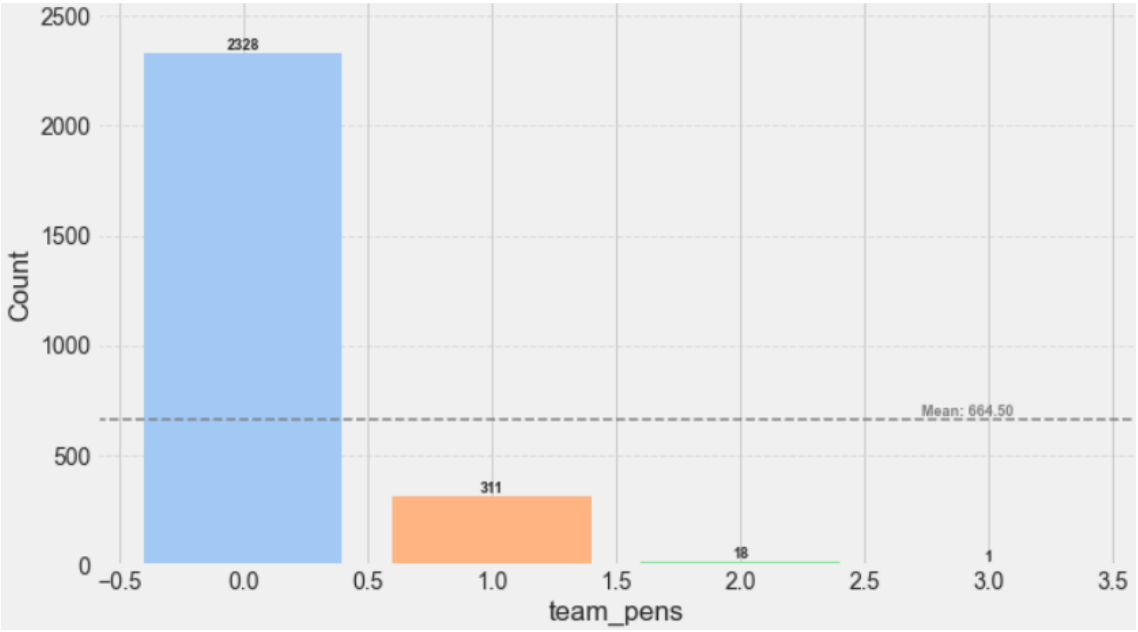


Figure 7.11. *Penalties Dataset Barplot*

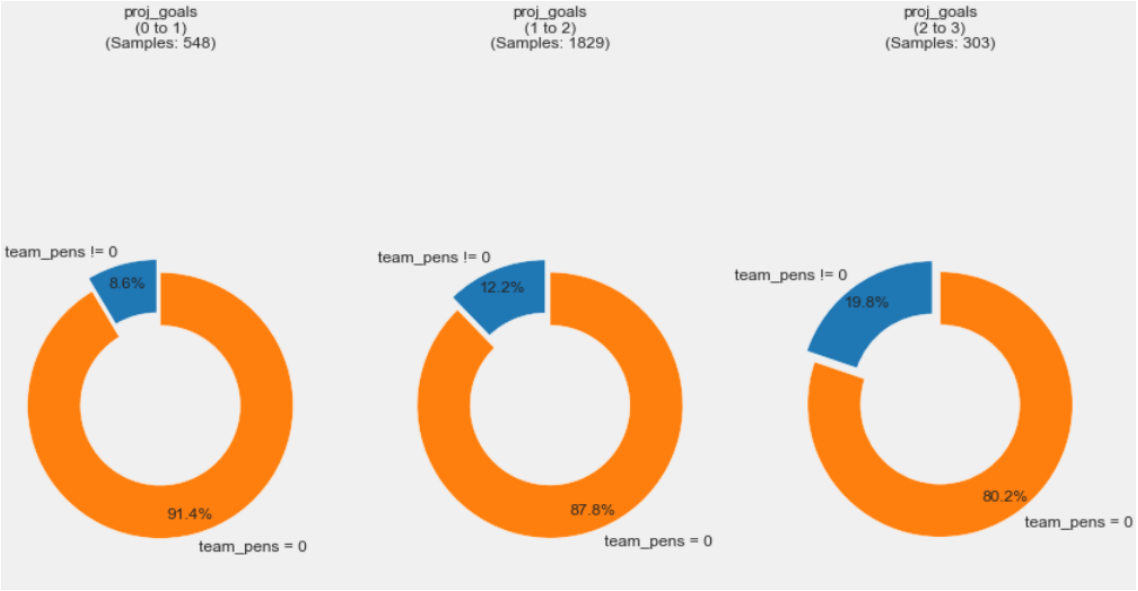


Figure 7.12. *Percentage of teams that won a penalty depending on the goals they were projected to score*

7.6 Team Goals

Although I created my own dataset and model about team goals (2260 rows × 6 columns), I decided to use the data of projected_goals provided by FiveThirtyEight [39]. In both cases we get the projected goals a team is going to score. But, we don't care about that number. What we are interested in at the end of the day is the expected clean sheet odds for each team. In order to translate the projected goals into clean sheet odds, we use the Poisson Process.

7.6.1 Poisson Process

In probability, statistics and related fields, a Poisson point process is a type of random mathematical object that consists of points randomly located on a mathematical space with the essential feature that the points occur independently of one another. The Poisson point process is often defined on the real line, where it can be considered as a stochastic process. In this setting, it is used, for example, in queueing theory to model random events, such as the arrival of customers at a store, phone calls at an exchange or occurrence of earthquakes, distributed in time [32].

A Poisson point process is characterized via the Poisson distribution. The Poisson distribution is the probability distribution of a random variable N (called a Poisson random variable) such that the probability that N equals n is given by:

$$Pr(N = n) = \frac{e^{-m}m^n}{n!}$$

where n! denotes factorial and the parameter m determines the shape of the distribution. (In fact, m equals the expected value of N) By definition, a Poisson point process has the property that the number of points in a bounded region of the process's underlying space is a Poisson-distributed random variable. [32]

7.6.2 Poisson Process Example

Let's say we want to calculate the expected clean sheet of Man City that plays against Chelsea. Let's also assume that Chelsea is projected to score 1 goal against Man City. To calculate the clean sheet chance, or the expected clean sheet, we assume that the goals Chelsea is going to score is a random variable X with expected value m=1, and that it follows the Poisson distribution. Therefore the clean sheet odds for Man City are:

$$Pr(X = 0) = \frac{e^{-1} * 1^0}{0!} = \frac{1}{e} = 0.367879$$

This is how the expected values for the clean sheets are being calculated each gameweek.

7.7 Saves Dataset

The goal of the saves dataset is to provide data for our models to predict the expected saves a goalkeeper is going to make in the next match. In order to achieve this, we take into account the goalkeeper's previous saves profile, as well as data about the strength of the two teams. The dataset is 4082 rows $\times$ 9 columns.

7.7.1 Features

The features are presented below, and we again assume that $action_i$ is the action or event performed by the player or team i games back:

- **savesp90:** The average save rate per 90 minutes of the goalkeeper in the current season. The *current_GW* is the current gameweek of the season.

$$savesp90 = \frac{90 * \sum_{i=1}^{current\_GW-1} saves_i}{\sum_{i=1}^{current\_GW-1} minutes\_played_i}$$

- **npxGAp90:** The average non-penalty expected goals the goalkeeper's team concedes in the current season.

$$npxGAp90 = \frac{\sum_{i=1}^{current\_GW-1} team's\_non\_penalty\_expectedGoals\_conceded_i}{current\_GW - 1}$$

- **npxGAp90(L4):** The average non-penalty expected goals the goalkeeper's team concedes taking into account the last 4 matches.

$$npxGAp90(L4) = \frac{\sum_{i=1}^{4} team's\_non\_penalty\_expectedGoals\_conceded_i}{4}$$

- **opp_npxGp90:** The average non-penalty expected goals the opponent team produces in the current season. The *current_GW* is the current gameweek of the season. This is a metric about the opponent team's strength and it is used to take into account the opponent team's attacking strength this season.

$$opp\_npxGp90 = \frac{\sum_{i=1}^{current\_GW-1} opponent\_team's\_non\_penalty\_expectedGoals_i}{current\_GW - 1}$$

- **opp_npxGp90(L4):** The average non-penalty expected goals the opponent team produces during the last 4 matches. The *current_GW* is the current gameweek of the season.

$$opp\_npxGp90(L4) = \frac{\sum_{i=1}^{4} opponent\_team's\_non\_penalty\_expectedGoals_i}{4}$$

- **opp_proj_goals:** The goals the opponent team is projected to score against the goalkeeper's team as calculated by FiveThirtyEight [39]. This is another metric about the overall strength differential between the teams.

- **was_home:** True if the goalkeeper's team was playing in their home arena, and False if they were playing on the road. A metric for our model to take into account the possible advantage of playing with your home crowd.

- **minutes:** The minutes the goalkeeper played in that specific match.

- **saves:** The saves the goalkeeper made in that specific match. This is the label, and the number we are trying to predict.

7.7.2 Dataset Exploration

A goalkeeper makes an average of about 2.5 saves per game as we see in [Figure 7.13]. The game rewards goalkeepers 1 point for 3 saves they make. So, the Poisson Process that was explained before is being used here as well to calculate the expected save points for the goalkeepers.

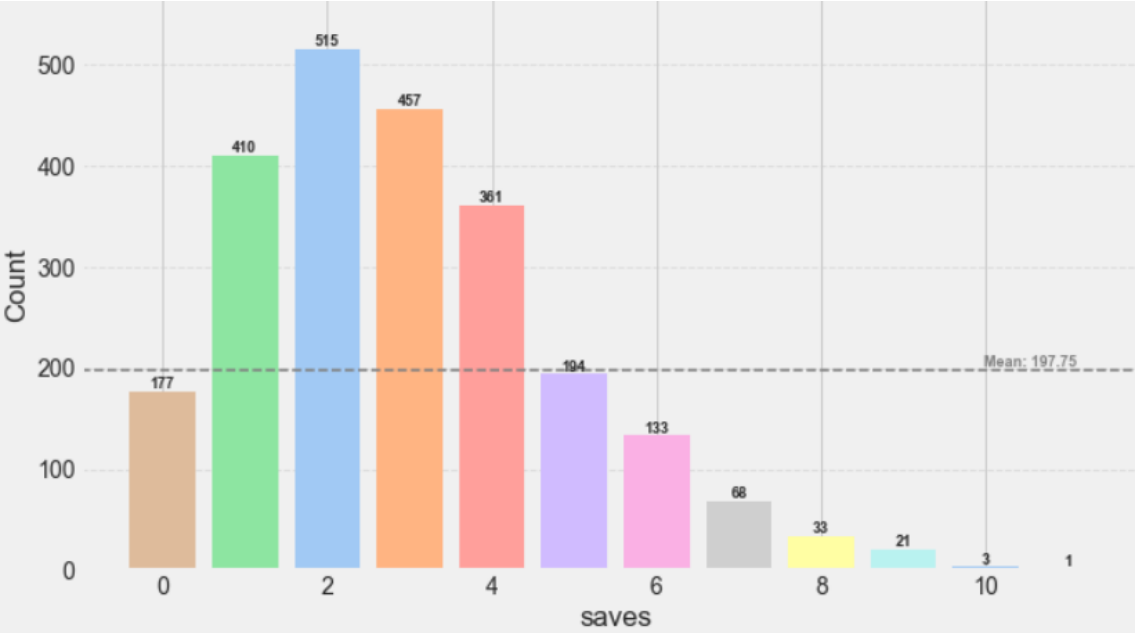


Figure 7.13. *Saves Dataset Barplot*

7.8 Bonus Points Dataset

After each match, FPL awards 3 players (sometimes more when there are ties) with 1, 2, and 3 bonus points for their performances [Figure 7.14]. The Bonus Points Dataset is created with the goal of predicting the expected bonus points for every player in the next match. Individual player statistics and attributes as well as team and opponent specific data are being collected. The dataset is 47543 rows $\times$ 10 columns.

7.8.1 Features

The features are presented below, and we again assume that $action_i$ is the action or event performed by the player or team i games back:

- **bonusp90:** The average bonus points rate per 90 minutes of the player in the current season. The *current_GW* is the current gameweek of the season.

$$bonusp90 = \frac{90 * \sum_{i=1}^{current\_GW-1} bonus\_points_i}{\sum_{i=1}^{current\_GW-1} minutes\_played_i}$$

- **position:** The player's position in the game. There are goalkeepers, defenders, midfielders and forwards. Each position class is rewarded bonus points a little

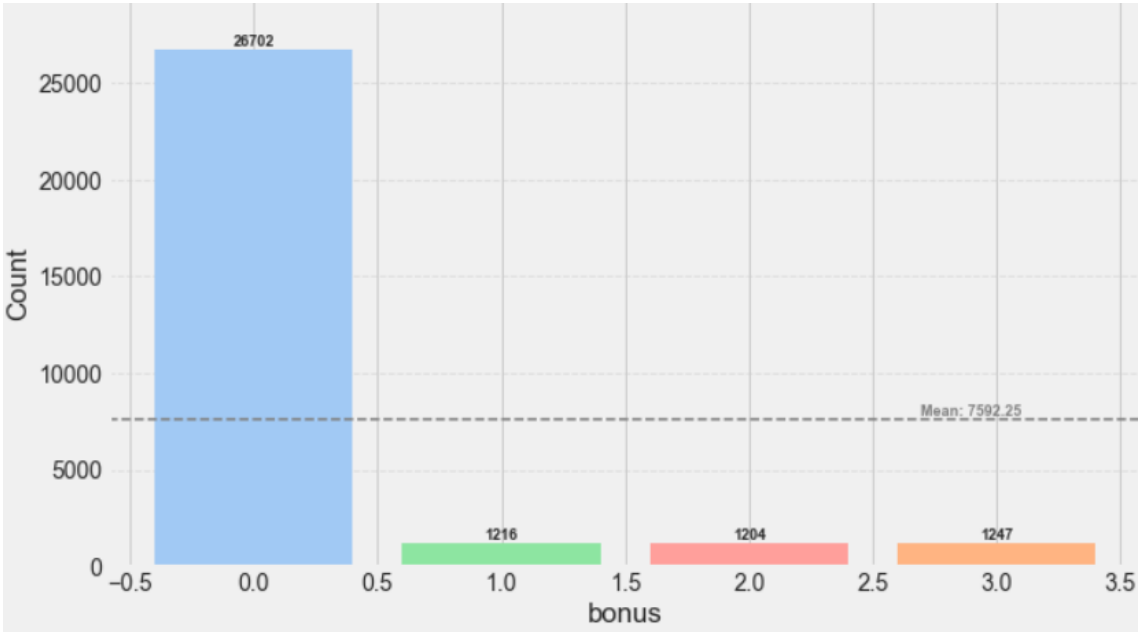


Figure 7.14. *Bonus Points Dataset Barplot*

bit differently, depending on the player's actions during the match. Therefore, the position is an important feature to consider.

- **npxGp90:** The average non-penalty expected goals the player has produced in the current season.

$$npxGp90 = \frac{\sum_{i=1}^{current\_GW-1} non\_penalty\_expectedGoals_i}{current\_GW - 1}$$

- **xAp90:** The average expected assists the player has created in the current season.

$$xAp90 = \frac{\sum_{i=1}^{current\_GW-1} expectedAssists_i}{current\_GW - 1}$$

- **npxGAp90:** The average non-penalty expected goals the player's team concedes in the current season.

$$npxGAp90 = \frac{\sum_{i=1}^{current\_GW-1} team's\_non\_penalty\_expectedGoals\_conceded_i}{current\_GW - 1}$$

- **opp_npxGp90:** The average non-penalty expected goals the opponent team produces in the current season. The *current_GW* is the current gameweek of the season.

$$opp\_npxGp90 = \frac{\sum_{i=1}^{current\_GW-1} opponent\_team's\_non\_penalty\_expectedGoals_i}{current\_GW - 1}$$

- **opp_npxGAp90:** The average non-penalty expected goals the opponent team con-

cedes in the current season.

$$opp\_npxGAp90 = \frac{\sum_{i=1}^{current\_GW-1} opponent\_team's\_non\_penalty\_expectedGoals\_conceded_i}{current\_GW - 1}$$

- **was_home:** True if the player's team was playing in their home arena, and False if they were playing on the road. A metric for our model to take into account the possible advantage of playing with your home crowd.
- **minutes:** The minutes the player played in that specific match.
- **bonus:** The bonus the player was awarded in that specific match. This is the label, and the number we are trying to predict.

The created datasets analyzed above, can be found on my GitHub [40].

Chapter 8

Training and Evaluation

In this chapter we will explore the techniques used and the decisions made during the training phase, we will examine the metrics with which the models were evaluated, and we will observe those evaluations.

8.1 Models and Evaluation Metrics

Since the data for the specific problem tend to be very noisy, we need very robust/regularized models, which avoid overfitting. The models that were selected for training are the following:

- **Random Forest Regressor**
- **Kernel Ridge Regressor**
- **XGBoost Regressor**

Those models performed the best according to similar projects we examined during the literature research phase (see Chapter 6: Related Work).

The metrics with which our models will be evaluated are the following:

- **Mean Absolute Error (MAE):** It is the average of the absolute differences between the actual value and the model's predicted value.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|$$

  where, N = total number of data points, y_i = actual value, y_i' = predicted value.

  Here, a big error doesn't overpower a lot of small errors and thus the output provides us with a relatively unbiased understanding of how the model is performing. Hence, it fails to punish the bigger error terms [41].

- **Root Mean Square Error (RMSE):** It is the average root-squared difference between the real value and the predicted value. By taking a square root of MSE, we get the Root Mean Squared Error. The lower the value, the better the regression model [41].

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2}$$

where, N = total number of data points, y_i = actual value, y'_i = predicted value.

- **R^2 score:** R-squared explains to what extent the variance of one variable explains the variance of the second variable. In other words, it measures the proportion of variance of the dependent variable explained by the independent variables.

$$R^2 = 1 - \frac{SSE}{SST}$$

where SSE is the sum of the square of the difference between the actual value and the predicted value

$$SSE = \sum_{i=1}^{N} (y_i - y'_i)^2$$

and SST is the total sum of the square of the difference between the actual value and the mean of the actual value.

$$SST = \sum_{i=1}^{N} (y_i - \overline{y})^2$$

Here, y_i is the observed target value, y'_i is the predicted value, $\overline{y}$ is the mean value, and N represents the total number of observations.

R squared is a popular metric for identifying model accuracy. It tells how close the data points are to the fitted line generated by a regression algorithm. A larger R squared value indicates a better fit.

8.2 Train-Test Split

The training set contains data from the past 3 seasons (2019/20 - 2021/22) and the test set contains data from the current season (2022/23) up to GW25. This is a 80%-20% split with our data.

Using the current season as a test set, enables us to evaluate all our models on the same season regardless of whether they are focused on player performance (such as the non-penalty goals model) or team performance (such as the team goals model).

8.3 Hyperparameter Tuning

For the Random Forest Regressors the most important hyperparameters that were tuned are:

- n_estimators: The number of trees in the forest. [13]
- max_depth: Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.

For the XGBoost Regressors the most important hyperparameters that were tuned are:

- n_estimators: The number of trees.
- max_depth: Maximum depth of a tree. Increasing this value will make the model more complex and more likely to overfit.
- learning_rate: A technique to slow down the learning in the gradient boosting model is to apply a weighting factor for the corrections by new trees when added to the model. This weighting is called the shrinkage factor or the learning rate, depending on the literature or the tool.

For the Kernel Ridge Regressors the most important hyperparameter that was tuned is:

- alpha: Regularization strength. Regularization improves the conditioning of the problem and reduces the variance of the estimates. [19]

In order to fine tune the hyperparameters of the models, Grid Search Cross Validation and Randomized Search Cross Validation (n_iter=10) were used with the following ranges:

- n_estimators: 200 - 1200
- max_depth: 2 - 7
- learning_rate: 0.006 - 0.016
- alpha: 0.5 - 1.5

This is how the best parameters for each model were found.

8.4 Evaluation

8.4.1 Mean Absolute Error Table

The MAE results of our models are being presented on the following table.

	Random Forest	XGBoost	Kernel Ridge
Non-penalty Goals	0.09378	**0.09366**	0.11581
Assists	**0.10092**	0.10484	0.11247
Penalties	0.23667	**0.20918**	0.21616
Team Goals	0.94811	0.94030	**0.93470**
Saves	**1.00484**	1.02365	1.17936
Bonus	**0.24714**	0.25096	0.28834

Table 8.1. *MAE Evaluation*

We observe that the Random Forest models top 3 of the 6 models, with XGBoost models topping 2 out of the 3 remaining models, and Kernel Ridge best at the Team Goals model. However, the differences between the Random Forest and XGBoost models are mainly minimal and the dominant model can change when we give more weight to the extreme cases.

8.4.2 Root Mean Squared Error Table

The RMSE results of our models are being presented on the following table.

	Random Forest	XGBoost	Kernel Ridge
Non-penalty Goals	0.24316	**0.23885**	0.24850
Assists	0.23798	**0.23596**	0.24017
Penalties	0.31277	**0.30599**	0.31374
Team Goals	1.19468	1.19524	**1.18542**
Saves	1.54129	**1.53651**	1.59737
Bonus	0.55274	**0.55206**	0.56428

Table 8.2. *RMSE Evaluation*

Now, we observe that the XGBoost models dominate apart from the Team Goals model, where Kernel Ridge is still on the top. What is surprising is the fact that XGBoost models are better than Random Forest models on every category judging by RMSE, while the Random Forest models seemed better evaluating by MAE. This change implies that XGBoost models tend to be more accurate about the high variance cases, since the RMSE score penalizes large errors more by taking into account the square of the errors. It is noteworthy, however, that the Random Forest models perform only slightly worse and are still pretty good solutions.

8.4.3 R^2 Score Table

The R^2 score results of our models are being presented on the following table.

	Random Forest	XGBoost	Kernel Ridge
Non-penalty Goals	0.11814	**0.14911**	0.07900
Assists	0.07706	**0.08808**	0.05996
Penalties	-0.01321	**0.03020**	-0.01951
Team Goals	0.14130	0.14049	**0.15455**
Saves	0.46550	**0.46881**	0.42590
Bonus	0.12817	**0.13029**	0.09138

Table 8.3. *R^2 score Evaluation*

Here, we observe similar results with those mentioned above on the RMSE table. The XGBoost models dominate all categories, except for Team Goals with Kernel Ridge being the clear winner in this one.

Chapter 9

Final Stages and Results

9.1 EV Calculation

After we have selected the best performing models for every on-pitch action that leads to points, the basis of the predictive process has been set. All we have to do now is collect every current data point for every player that each model requires. We then feed those data points into the function that calculates the expected points for a player for a specific gameweek and we follow this iterative process for every player and every gameweek up until the selected horizon (the gameweek up until the EV calculations will be performed).

9.1.1 xMins Calculation

One really important part of the EV Calculation process is undoubtedly the xMins (expected minutes) calculation. This process refers to the minutes prediction of each player for each one of the next gameweeks that the EV Calculation will be performed. Minutes play a crucial role in predicting a player's value for the next gameweeks, a fact that is pretty intuitive. The more time a player is on the pitch, the better the chances of scoring points and reaching the thresholds for appearance points (1 appearance point for 0-60 minutes, 2 for >60 minutes). In the figure below [Figure 9.1], we can see the non-penalty Goals model feature importances for instance, and unsurprisingly the most important feature for predicting non-penalty goals, even more important than non-penalty goal rates and npxGp90 for the player, is minutes.

In order to calculate the xMins for a player, we take into account the player's past minutes in the team, the player's form, his injury status, and important news from the press conferences the teams' managers provide before the matches. His recent performances, and the team's program congestion are also taken into account. This is not an easy task, because each manager selects his team with different criteria for each game, players can get injured without it being general knowledge, and judging a player's performance is sometimes subjective. Moreover, the process is hard to fully automate, as the managers' plans are being influenced by thousands of factors, many of which we don't even have access to as outside observers, and the managers often give limited information at press conferences for strategic reasons. Finally, the xMins calculation process aims to predict minutes for gameweeks in the future (not only the upcoming one), where the possibility of an injury, or a tactical change in the formation always looms.

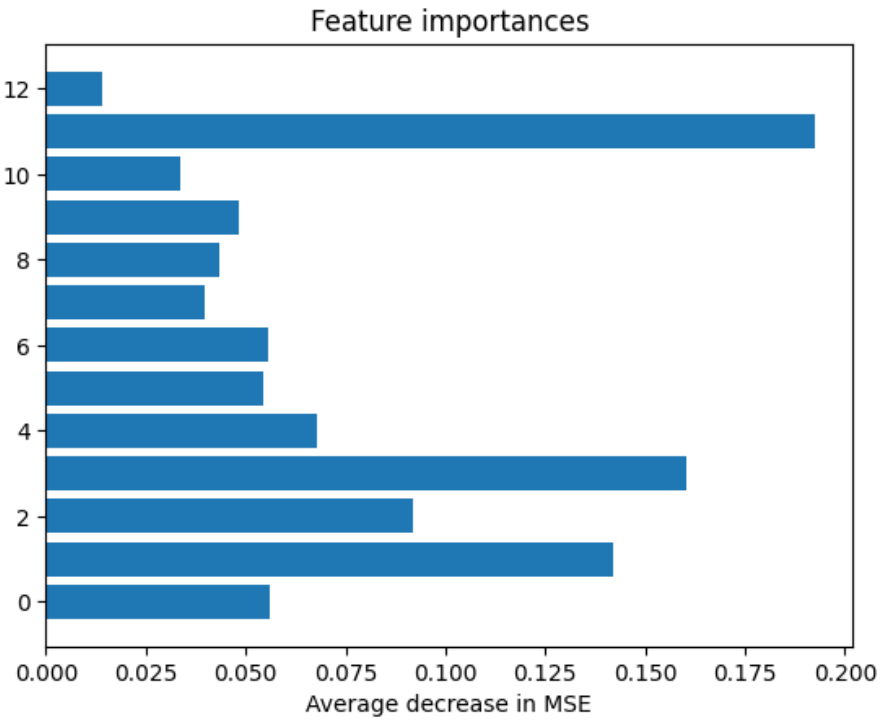


Figure 9.1. *Non-penalty Goals model Feature Importances*

9.1.2 xPoints Calculation

After all the current data points about a player, his team, and the opponent team have been collected, we use them to calculate the player's xPoints (Expected Points) for one specific gameweek.

First of all, the expected appearance points and the chance of playing at least 60 minutes (this is important because when a player plays for at least 60 minutes he is eligible for extra appearance points and clean sheet points) are being calculated depending on the player's xMins for the match. Then, the expected penalty goals of the player are being computed. The vast majority of players have 0 expected penalty goals, because they simply aren't on penalty duties for their team. Every team has about 3 players who are responsible for penalty duties. Usually, these players are in order: 1st choice penalty taker, 2nd choice penalty taker, 3rd choice penalty taker. To give the correct expected value to these players, as far as penalties as concerned, a penalty hierarchy has been implemented for every team, and the penalty takers get the expected points according to the chance of their team winning a penalty, being produced by the penalties model, the place of each player in his team's hierarchy, and the xMins of the players above him in the hierarchy (if any).

After the appearance points and penalty goals, the saved models, we explored in the previous chapter, predict the non-penalty goals, assists, saves and bonus of the player for the gameweek at hand, as well as the opponent team's projected goals. This information is not always translatable to expected value. For instance, the opponent team's projected goals are being used to assess the player's team's chance of a clean sheet, and the chance

of 2, 4 and 6 goals being conceded (the game deducts 1 point for every 2 goals conceded from defenders and goalkeepers). The expected saves are also being translated into the chance of having 3, 6, or 9 saves, as the game awards goalkeepers 1 point for every 3 saves they make. These "translations" are possible with the Poisson Process [32] we examined in an earlier chapter [see Chapter 7]. The assists and bonus models are being used directly, while we add the expected penalty goals and non-penalty goals of a player to compute the player's total goals for the match.

Finally, we add up all the expected actions/events multiplied by the points each action/event provides, according to the player's position [12]. The expected yellow cards deduct points from a player, according to his xMins. We assume a player's expected yellow cards is the average of yellow cards he has received throughout his career in the Premier League. Last but not least, we perform some basic checks (for example if a player has 0 xMins he will score 0 points) and we have the EV of the player for that specific match.

9.2 Results

In this section, we will compare the final results of the project with other similar projects that predict EV for Premier League players that were mentioned in the Related Work chapter. We will compare based on prediction data for the upcoming gameweek, and we will use MAE (Mean Absolute Error), RMSE (Root Mean Squared Error) and R^2 score as the evaluation metrics.

The projects we're comparing with are the following:

- FPL Review Premium Model [28]
- FPL Kiwi Model [27]
- Mikkel's Model [86]
- Fantasy Football Scout Model [29]
- Fantasy Football Hub Model [30]
- Fantasy Football Fix Model [31]

The models are being compared on 6900 common samples from the 2022/23 season, gameweeks: 26-38 (>1/3 of the season). The Kiwi model had a lot of missing values, so it was compared separately with our model (GWs: 26-33 / season: 2022-23 / N=2317) and our model significantly outperformed it in every metric. The results of the models are being presented on the following table.

We observe that our model performs really well compared with the top models of the field. It is noteworthy that all of them (except for Kiwi's model) are products you have to pay for and are considered the best solutions available in the FPL community.

Our model is the best as far as Mean Absolute Error is concerned and second only to FPL Review across the other two metrics. Therefore, we can safely say it is one of the best solutions built for the FPL problem.

	MAE	RMSE	R^2
My Model	**1.2898**	2.3004	0.3334
FPL Review	1.3005	**2.2713**	**0.3502**
Mikkel's Model	1.3175	2.3278	0.3175
Fantasy Football Fix	1.3678	2.3606	0.2981
Fantasy Football Scout	1.3279	2.3083	0.3306
Fantasy Football Hub	1.4150	2.3798	0.2867

Table 9.1. *Model Results (GWs: 26-38 / season: 2022-23 / N=6900)*

In [Figure 9.2] the better models are top right and the bigger the size of the bubble, the bigger the R-squared score of the model.

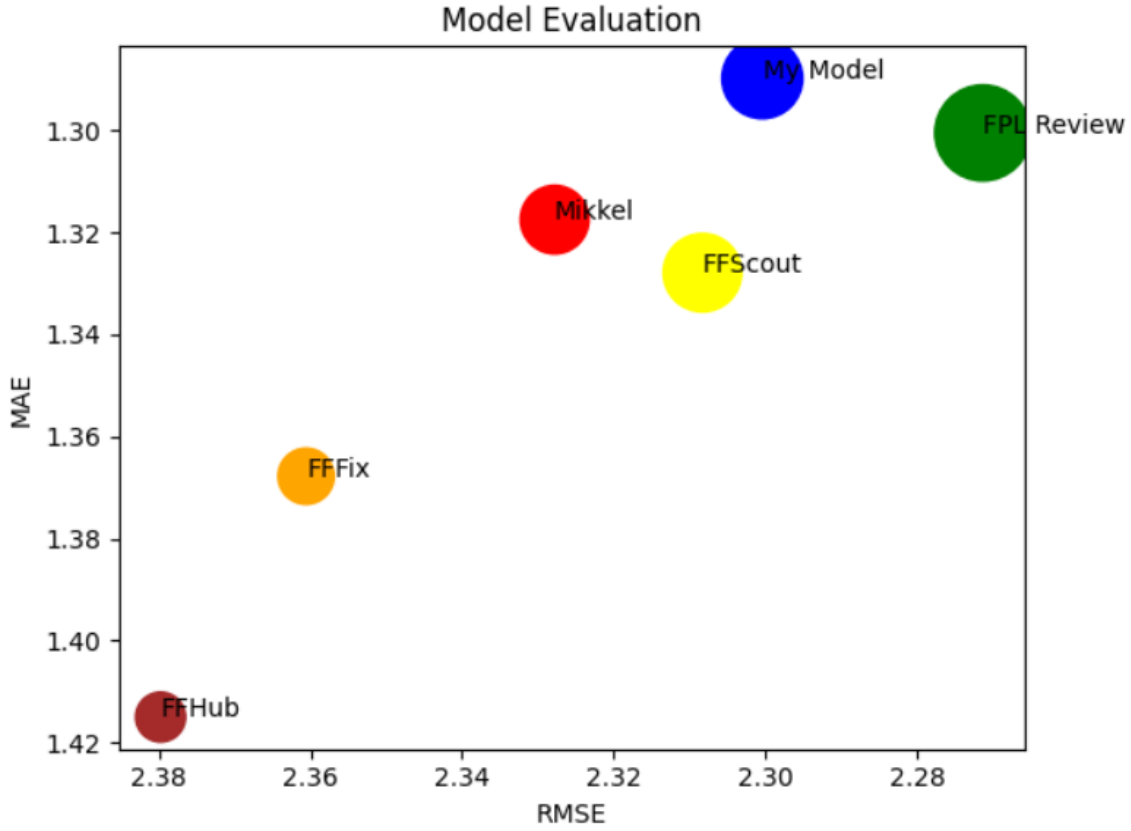
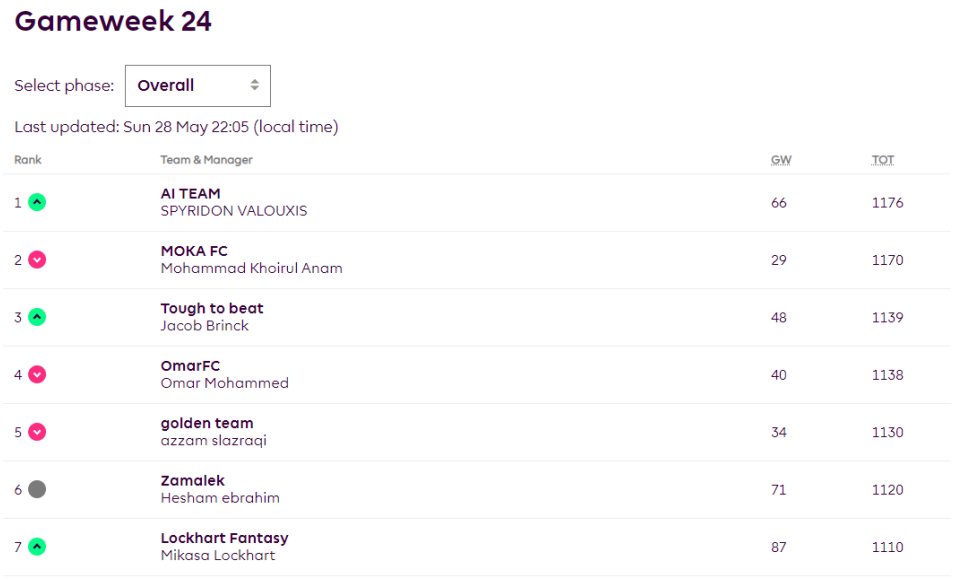


Figure 9.2. *Models Evaluation Bubbleplot*

Another really positive indicator of our model's abilities is the fact that the team it managed starting GW24, ended up 1st across all the teams (>25,000 teams) that started the same gameweek [Figure 9.3].

Figure 9.3. *Gameweek 24 League*



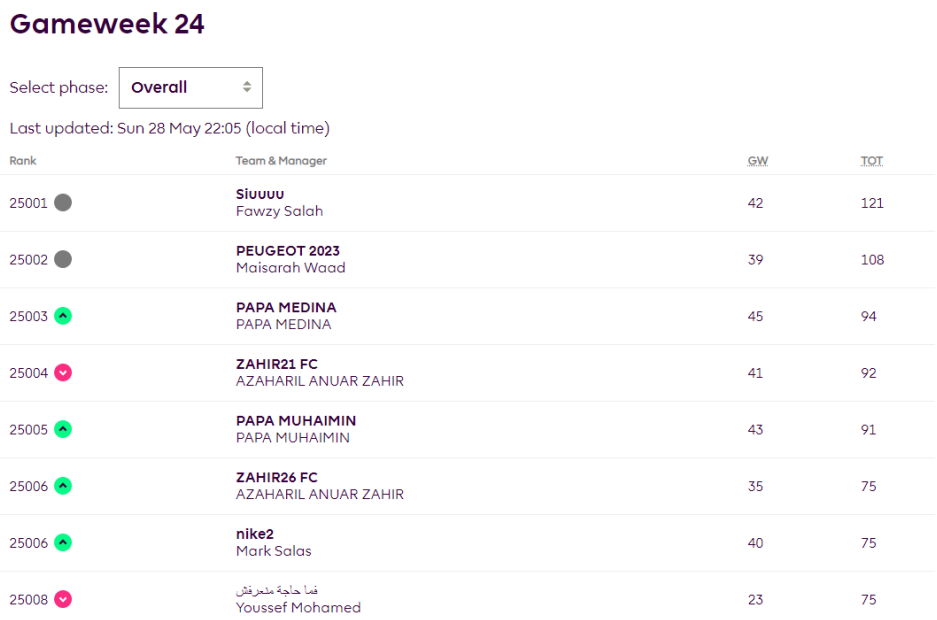
Gameweek 24

Select phase: Overall

Last updated: Sun 28 May 22:05 (local time)

Rank	Team & Manager	GW	TOT
1	**AI TEAM** SPYRIDON VALOUXIS	66	1176
2	**MOKA FC** Mohammad Khoirul Anam	29	1170
3	**Tough to beat** Jacob Brinck	48	1139
4	**OmarFC** Omar Mohammed	40	1138
5	**golden team** azzam slazraqi	34	1130
6	**Zamalek** Hesham ebrahim	71	1120
7	**Lockhart Fantasy** Mikasa Lockhart	87	1110

Figure 9.4. *Gameweek 24 League - AI Team*



Gameweek 24

Select phase: Overall

Last updated: Sun 28 May 22:05 (local time)

Rank	Team & Manager	GW	TOT
25001	**Siuuuu** Fawzy Salah	42	121
25002	**PEUGEOT 2023** Maisarah Waad	39	108
25003	**PAPA MEDINA** PAPA MEDINA	45	94
25004	**ZAHIR21 FC** AZAHARIL ANUAR ZAHIR	41	92
25005	**PAPA MUHAIMIN** PAPA MUHAIMIN	43	91
25006	**ZAHIR26 FC** AZAHARIL ANUAR ZAHIR	35	75
25006	**nike2** Mark Salas	40	75
25008	فما حاجة متعرفش Youssef Mohamed	23	75

Figure 9.5. *Gameweek 24 League - 25,000 Teams*

9.3 FPL Optimization and Solvers

Optimization is a fundamental concept used in various domains to achieve the best possible outcome. It involves finding the optimal solution from a set of possible alternatives, given certain constraints and objectives. The goal is to maximize or minimize a specific criterion, such as efficiency, cost, profit, or performance, depending on the context. FPL optimization refers to the process of selecting the best combination of players within the given game constraints to maximize the total expected points earned by the fantasy team.

Mathematical modeling is a powerful approach used to represent and solve the FPL

optimization problem. It involves formulating the problem in mathematical terms, defining variables, constraints, and an objective function. For FPL, mathematical modeling involves binary variables to represent player selection (1 if selected, 0 if not) etc., continuous variables to capture money left in the bank, budget constraints as linear inequalities, and the objective function as a function that maximizes the total points over a specific horizon. By formulating the problem mathematically, users can leverage optimization algorithms and solvers to efficiently explore the enormous solution space, considering various constraints and objectives. Mathematical models allow for systematic analysis, comparison of different strategies, and identification of the best team composition within the given constraints.

The FPL Optimization problem is a mixed-integer linear programming (MILP) problem, meaning that some of the variables are constrained to be integers, while other variables are allowed to be non-integers. Branch and cut is a method of combinatorial optimization for solving those problems and that's what the solver in the project uses. Branch and cut involves running a branch and bound algorithm and using cutting planes to tighten the linear programming relaxations. Note that if cuts are only used to tighten the initial LP relaxation, the algorithm is called cut and branch [42].

In this project, I have integrated Sertalp's solver [43]. It uses the sasoptpy library to model the FPL problem and express all the different constraints and variables. The solver allows the user to input his/her desired solver settings such as decay, the value of in the bank funds, the value of a free transfer, chips played etc. and his/her team data to create transfer paths like the one in [Figure 9.6].

In summary, solvers are tools used to model and solve the FPL problem, based on a set of useful EV values, like the ones we produced in this diploma thesis. The solver is the "blind worker" that translates the useful players' expected values into actionable plans. However, a significant drawback of FPL solvers is the fact that they are pretty computationally intensive, and thus slow, to run especially if we have big horizons. This is not a surprise as the search space is huge and the underlying algorithms try to find and prove the optimal solution, a process that can be slow sometimes.

	iter	buy	sell	score
0	0	-	-	128.436011
1	1	Salah	Fernandes	127.947591

Figure 9.6. *Solver Transfer Path Suggestion Table*

Chapter 10

Conclusion

In this study, the application of data science and machine learning techniques in the context of Fantasy Premier League (FPL) has been explored. The objectives of the thesis were as follows:

- **Develop Machine Learning Models for Predicting Expected Value (EV):** The first and main goal of the thesis was to develop a system based on machine learning models capable of predicting the Expected Value (EV) for FPL players. By developing accurate and reliable EV prediction models, the aim was to assist FPL managers in making informed decisions regarding player selection and transfers. To achieve this objective, the thesis explored various machine learning algorithms and regression techniques. Historical player performance data, along with relevant features such as expected player and team data and other statistics, were used to train the models. The performance of different machine learning models was evaluated, and the most effective approach for each separate task was selected and all of them were eventually combined for the EV prediction system.

- **Generate Optimal Moves for FPL Managers based on Predicted EV Values:** The second goal of the thesis was to generate optimal moves for FPL managers based on the predicted EV values. Once the EV values for players were estimated using the machine learning models, the thesis aimed to provide recommendations and strategies for FPL managers on player transfers, captaincy choices, and other decisions to maximize their team's overall performance. This objective involved the integration of the EV prediction models with optimization techniques. Optimization solvers were employed to identify the best combinations of player transfers and strategic moves that would lead to optimal team performance. By leveraging the predicted EV values, FPL managers can make informed decisions and optimize their team composition within the constraints of the game.

- **Contribute to the Field of FPL Analytics and Advance the Application of Machine Learning Techniques in FPL:** The final goal of the thesis was to contribute to the field of FPL analytics and advance the application of machine learning techniques in FPL. By developing accurate EV prediction models and providing optimization strategies, the thesis aimed to enhance the understanding of player performance dynamics, strategic decision-making, and team optimization in the context

of FPL. The research aimed to expand the existing knowledge base and contribute to the growing field of FPL analytics. By demonstrating the effectiveness of machine learning techniques, the thesis sought to promote the adoption of data-driven decision-making and optimization methods among FPL managers and enthusiasts.

Throughout the thesis, we have discussed the evolution of sports analytics as well as the most relevant soccer analytics metrics, the theoretical foundations of the different regression techniques used, explored relevant literature in the field, and built different datasets, identifying key factors affecting player performance in FPL. Additionally, we have examined various machine learning algorithms and techniques, selecting the most appropriate ones for our specific task.

The results obtained from our experiments and analyses have demonstrated the potential of machine learning in improving FPL performance. By leveraging historical player and team data and incorporating relevant features such as player statistics, team strength, and team/player form, our models were able to generate satisfactory predictions of players' future points. Furthermore, we successfully integrated Sertalp's optimization algorithm that creates actionable plans for FPL managers within the game constraints. Our findings highlight the importance of feature selection and model evaluation in the context of FPL. We observed that certain player attributes, such as past player performance over different time windows, team strengths, and the current season's player and team performance were highly influential in predicting future performance. Additionally, model evaluation techniques, including cross-validation and performance metrics such as mean squared error, R^2 score and mean absolute error, proved crucial in assessing the effectiveness of our models and comparing their performance.

Despite the promising results achieved in this study, it is important to acknowledge certain limitations and potential areas for improvement. Firstly, a potential limitation of this study lies in the fact that data from the 2019-20 season up until the 2022-23 season were used. The effects of COVID-19 significantly affected the Premier League, with arenas being empty due to COVID restrictions. Therefore, since an important part of the data was collected during that time period, the extent of the influence of being at home or away could be potentially misinterpreted. Secondly, an area for improvement would undoubtedly be a separate study for xMins (players' expected minutes) prediction. As mentioned earlier, the importance of xMins cannot be overstated when predicting a player's future expected points, since the most important prerequisite of scoring points is being on the pitch. Another potential area of research could be creating and evaluating specific targeted models for predicting EV for high-performers separately. Something that is worthy of consideration is the concept of Ensemble models. In simple terms this is combining data from several independent models into one model, similar to the kind of value we find in the "wisdom of the crowd" idea. The ability to dampen out mistakes in individual models may hold value, though equally sometimes it can potentially come at the cost of special insights too. Additionally, incorporating more advanced machine learning techniques, such as deep learning models, could further enhance the predictive power of our models. Last but not least, creating heuristics for shortening the solver

running time is another thing that would potentially better an FPL manager's experience, especially if speed is more important than accuracy. Heuristics are not always desirable since you have no way of knowing how far you are from the optimal solution, but a combination of both producing a fast solution with a heuristic and running optimization in the background to prove, may be the best of both worlds.

In summary, the application of machine learning techniques in FPL has proven to be effective in predicting player performance and optimizing team selection. The ability to leverage historical data and extract valuable insights from vast amounts of information provides FPL managers with a competitive advantage. The findings of this thesis contribute to the growing body of knowledge in the field of sports analytics and provide a foundation for further research and development in the area of data science in FPL. Ultimately, the integration of machine learning algorithms in FPL has the potential to revolutionize the way managers approach team selection, transfer decisions, and overall strategy. As the game continues to evolve and the availability of data increases, the role of machine learning in FPL is likely to become even more significant. By empowering FPL managers with advanced predictive models and optimization algorithms, we can enhance the overall experience of the game and provide a pathway to improved performance. This thesis serves as a stepping stone towards unlocking the full potential of machine learning in Fantasy Premier League and encourages further exploration and innovation in this exciting field.

The code for this project can be found on my GitHub [40].

Bibliography

[1] *The Evolution of Sports Analytics*. https://www.sportstechbiz.com/p/the-evolution-of-sports-analytics. Access Date: 05-10-2023.

[2] *Linear Regression*. https://towardsdatascience.com/linear-regression-explained-1b36f97b7572. Access Date: 05-17-2023.

[3] *Random Forest Regression*. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84. Access Date: 05-17-2023.

[4] *Bagging vs Boosting*. https://www.researchgate.net/figure/Comparison-of-three-methods-single-bagging-and-boosting-7_fig4_338362989/actions#reference. Access Date: 05-17-2023.

[5] David Shin. *Evolution of Analytical Data in Sports*. 2020. Access Date: 05-10-2023.

[6] *Possesion Value*. https://theanalyst.com/eu/2021/03/what-is-possession-value/. Access Date: 05-15-2023.

[7] *Sequences*. https://theanalyst.com/eu/2021/03/possessions-and-sequences-in-football/. Access Date: 05-15-2023.

[8] *Expected Goals*. https://theanalyst.com/eu/2021/07/what-are-expected-goals-xg/. Access Date: 05-15-2023.

[9] *Regression Analysis in Machine Learning*. https://www.javatpoint.com/regression-analysis-in-machine-learning. Access Date: 05-17-2023.

[10] *Boosting Algorithms Explained*. https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30. Access Date: 05-17-2023.

[11] *Kernel Ridge Regression*. https://www.cs.cmu.edu/~aarti/Class/10315_Fall20/hws/prog3.html. Access Date: 05-17-2023.

[12] *FPL Rules*. https://fantasy.premierleague.com/help/rules. Access Date: 05-10-2023.

[13] *Random Forest Regression*. https://builtin.com/data-science/random-forest-python. Access Date: 05-17-2023.

[14] *Bootstrap Aggregating*. https://en.wikipedia.org/wiki/Bootstrap_aggregating. Access Date: 05-17-2023.

[15] *XGBoost Github*. https://github.com/dmlc/xgboost. Access Date: 05-17-2023.

[16] *Boosted Trees*. https://xgboost.readthedocs.io/en/latest/tutorials/model.html. Access Date: 05-17-2023.

[17] *Boosting*. https://en.wikipedia.org/wiki/Boosting_(machine_learning). Access Date: 05-17-2023.

[18] *Ridge Regression*. https://en.wikipedia.org/wiki/Ridge_regression. Access Date: 05-17-2023.

[19] *Kernel Ridge Regression*. https://web2.qatar.cmu.edu/~gdicaro/10315-Fall19/additional/welling-notes-on-kernel-ridge.pdf. Access Date: 05-17-2023.

[20] *Using Machine Learning to Predict high- performing Players in Fantasy Premier League*. https://medium.com/@277roshan/machine-learning-to-predict-high-performing-players-in-fantasy-premier-league-3c0de546b251. Access Date: 05-18-2023.

[21] *Building an FPL Captain Classifier*. https://medium.com/datacomics/building-an-fpl-captain-classifier-cf4ee343ebcc. Access Date: 05-18-2023.

[22] Nicholas Bonello, Joeran Beel, Seamus Lawless και Jeremy Debattista. *Multi-stream Data Analytics for Enhanced Performance Prediction in Fantasy Football*, 2019.

[23] F. Bonomo, G. Durán και J. Marenco. *Mathematical programming as a tool for virtual soccer coaches: a case study of a fantasy sport game*. *International Transactions in Operational Research*, 21(3):399–414, 2014.

[24] *Fantasy Premier League - Performance Prediction*. http://conference.ioe.edu.np/publications/ioegc12/IOEGC-12-247-12361.pdf. Access Date: 05-18-2023.

[25] Tim Matthews, Sarvapali Ramchurn και Georgios Chalkiadakis. *Competing with Humans at Fantasy Football: Team Formation in Large Partially-Observable Domains*. 26:1394–1400, 2021.

[26] *FPL Kiwi Process Explained*. https://twitter.com/theFPLkiwi/status/1314641452899676163. Access Date: 05-18-2023.

[27] *FPL Kiwi Player Predictions*. https://github.com/theFPLkiwi/theFPLkiwi. Access Date: 05-18-2023.

[28] *FPL Review*. https://fplreview.com/. Access Date: 05-18-2023.

[29] *Fantasy Football Scout Rate My Team*. https://www.fantasyfootballscout.co.uk/category/rate-my-team/. Access Date: 05-18-2023.

[30] *Fantasy Football Hub*. https://www.fantasyfootballhub.co.uk/my-team/pick. Access Date: 05-18-2023.

[31] *Fantasy Football Fix*. https://www.fantasyfootballfix.com/. Access Date: 05-18-2023.

[32] *Poisson Process*. https://en.wikipedia.org/wiki/Poisson_point_process. Access Date: 05-22-2023.

[33] *FPL API Endpoints*. https://www.reddit.com/r/FantasyPL/comments/f8t3bw/cheatsheet_of_all_current_fpl_endpoints/. Access Date: 05-22-2023.

[34] *Understat*. https://understat.com/league/EPL. Access Date: 05-22-2023.

[35] *Understat Package*. https://understat.readthedocs.io/en/latest/index.html. Access Date: 05-22-2023.

[36] Vaastav Anand. *FPL Historical Dataset*. Retrieved August 2022 from https://github.com/vaastav/Fantasy-Premier-League/, 2022.

[37] *Chris Musson's ID Map*. https://github.com/ChrisMusson/FPL-ID-Map. Access Date: 05-22-2023.

[38] *FiveThirtyEight Club Soccer Predictions*. https://projects.fivethirtyeight.com/soccer-predictions/. Access Date: 05-22-2023.

[39] *How Our Club Soccer Predictions Work*. https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/. Access Date: 05-22-2023.

[40] *Project Code*. https://github.com/spiros26/FPL-Project. Access Date: 06-14-2023.

[41] *Evaluation Metrics - Regression*. https://medium.com/analytics-vidhya/evaluation-metrics-for-regression-models-c91c65d73af. Access Date: 05-31-2023.

[42] *Branch and Cut Algorithm*. https://en.wikipedia.org/wiki/Branch_and_cut. Access Date: 06-12-2023.

[43] *Sertalp's Solver*. https://github.com/sertalpbilal/FPL-Optimization-Tools. Access Date: 06-05-2023.

[44] *Henry Chadwick*. https://baseballhall.org/hall-of-famers/chadwick-henry. Access Date: 05-10-2023.

[45] *Sabermetrics*. https://onlinegrad.syracuse.edu/blog/sabermetrics-baseball-analytics-the-science-of-winning/. Access Date: 05-10-2023.

[46] *Bill James*. https://en.wikipedia.org/wiki/Bill_James. Access Date: 05-10-2023.

[47] Michael Lewis. *Moneyball: The Art of Winning an Unfair Game*. W. W. Norton Company, 2003.

[48] *Sports Analytics*. https://en.wikipedia.org/wiki/Sports_analytics. Access Date: 05-10-2023.

[49] *FiveThirtyEight*. https://fivethirtyeight.com/. Access Date: 05-10-2023.

[50] *MLS Analytics*. https://www.mlssoccer.com/news/soccer-analytics-101. Access Date: 05-15-2023.

[51] *Expected Goals Explained*. https://statsbomb.com/soccer-metrics/expected-goals-xg-explained/. Access Date: 05-15-2023.

[52] Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos Kollias και Georgios Tagaris. *Deep neural architectures for prediction in healthcare*. *Complex & Intelligent Systems*, 4(2):119–131, 2018.

[53] Athanasios Tagaris, Dimitrios Kollias και Andreas Stafylopatis. *Assessment of Parkinson's disease based on deep neural networks*. *International Conference on Engineering Applications of Neural Networks*, σελίδες 391–403. Springer, 2017.

[54] Athanasios Tagaris, Dimitrios Kollias, Andreas Stafylopatis, Georgios Tagaris και Stefanos Kollias. *Machine learning for neurodegenerative disorder diagnosis—survey of practices and launch of benchmark dataset*. *International Journal on Artificial Intelligence Tools*, 27(03):1850011, 2018.

[55] Ilianna Kollia, Andreas Georgios Stafylopatis και Stefanos Kollias. *Predicting Parkinson's disease using latent information extracted from deep neural networks*. *2019 International Joint Conference on Neural Networks (IJCNN)*, σελίδες 1–8. IEEE, 2019.

[56] James Wingate, Ilianna Kollia, Luc Bidaut και Stefanos Kollias. *Unified deep learning approach for prediction of Parkinson's disease*. *IET Image Processing*, 14(10):1980–1989, 2020.

[57] Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian και Stefanos Kollias. *Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis*. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, σελίδες 537–544, 2021.

[58] Dimitrios Kollias, Anastasios Arsenos και Stefanos Kollias. *Ai-mia: Covid-19 detection & severity analysis through medical imaging*. *arXiv preprint arXiv:2206.04732*, 2022.

[59] Anastasios Arsenos, Dimitrios Kollias και Stefanos Kollias. *A Large Imaging Database and Novel Deep Neural Architecture for Covid-19 Diagnosis*. *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, σελίδες 1–5. IEEE, 2022.

[60] Dimitrios Kollias, Anastasios Arsenos και Stefanos Kollias. *AI-MIA: Covid-19 detection and severity analysis through medical imaging*. *Computer Vision-ECCV 2022 Workshops: Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VII*, σελίδες 677–690. Springer, 2023.

[61] Dimitrios Kollias, Anastasios Arsenos και Stefanos Kollias. *A deep neural architecture for harmonizing 3-D input data analysis and decision making in medical imaging*. *Neurocomputing*, 542:126244, 2023.

[62] Dimitrios Kollias, Miao Yu, Athanasios Tagaris, Georgios Leontidis, Andreas Stafylopatis και Stefanos Kollias. *Adaptation and contextualization of deep neural network*

models. *2017 IEEE symposium series on computational intelligence (SSCI)*, σελίδες 1–8. IEEE.

[63] D Kollias, N Bouas, Y Vlaxos, V Brillakis, M Seferis, I Kollia, L Sukissian, J Wingate και S Kollias. *Deep Transparent Prediction through Latent Representation Analysis*. *arXiv preprint arXiv:2009.07044*, 2020.

[64] Dimitris Kollias, Y Vlaxos, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate και S Kollias. *Transparent adaptation in deep medical image diagnosis*. *International Workshop on the Foundations of Trustworthy AI Integrating Learning, Optimization and Reasoning*, σελίδες 251–267. Springer, 2020.

[65] Fabio De Sousa Ribeiro, Francesco Calivá, Mark Swainson, Kjartan Gudmundsson, Georgios Leontidis και Stefanos Kollias. *Deep bayesian self-training*. *Neural Computing and Applications*, 32(9):4275–4291, 2020.

[66] Fabio De Sousa Ribeiro, Georgios Leontidis και Stefanos Kollias. *Capsule routing via variational bayes*. *Proceedings of the AAAI Conference on Artificial Intelligence*, τόμος 34, σελίδες 3749–3756, 2020.

[67] Fabio De Sousa Ribeiro, Georgios Leontidis και Stefanos Kollias. *Introducing routing uncertainty in capsule networks*. *Advances in Neural Information Processing Systems*, 33:6490–6502, 2020.

[68] Nikolaos Simou και Stefanos Kollias. *Fire: A fuzzy reasoning engine for impecise knowledge*. Citeseer.

[69] Francesco Caliva, Fabio Sousa De Ribeiro, Antonios Mylonakis, Christophe Demazi'ere, Paolo Vinai, Georgios Leontidis και Stefanos Kollias. *A deep learning approach to anomaly detection in nuclear reactors*. *2018 International joint conference on neural networks (IJCNN)*, σελίδες 1–8. IEEE, 2018.

[70] Stefanos Kollias, Miao Yu, James Wingate, Aiden Durrant, Georgios Leontidis, Georgios Alexandridis, Andreas Stafylopatis, Antonios Mylonakis, Paolo Vinai και Christophe Demaziere. *Machine learning for analysis of real nuclear plant data in the frequency domain*. *Annals of Nuclear Energy*, 177:109293, 2022.

[71] Stefanos Kollias, Miao Yu, James Wingate, Aiden Durrant, Georgios Leontidis, Georgios Alexandridis, Andreas Stafylopatis, Antonios Mylonakis, Paolo Vinai και Christophe Demaziere. *Machine learning for analysis of real nuclear plant data in the frequency domain*. *Annals of Nuclear Energy*, 177:109293, 2022.

[72] Bashar Alhnaity, Stefanos Kollias, Georgios Leontidis, Shouyong Jiang, Bert Schamp και Simon Pearson. *An autoencoder wavelet based deep neural network with attention mechanism for multi-step prediction of plant growth*. *Information Sciences*, 560:35–50, 2021.

[73] Bashar Alhnaity, Simon Pearson, Georgios Leontidis και Stefanos Kollias. *Using deep learning to predict plant growth and yield in greenhouse environments*. *International Symposium on Advanced Technologies and Management for Innovative Greenhouses: GreenSys2019 1296*, σελίδες 425–432, 2019.

[74] Andreas Psaroudakis και Dimitrios Kollias. *MixAugment & Mixup: Augmentation Methods for Facial Expression Recognition*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 2367–2375, 2022.

[75] Dimitrios Kollias. *Abaw: Learning from synthetic data & multi-task learning challenges*. *European Conference on Computer Vision*, σελίδες 157–172. Springer, 2023.

[76] Dimitrios Kollias και Stefanos Zafeiriou. *Training deep neural networks with different datasets in-the-wild: The emotion recognition paradigm*. σελίδες 1–8. IEEE, 2018.

[77] Dimitrios Kollias. *Multi-Label Compound Expression Recognition: C-EXPR Database & Network*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, σελίδες 5589–5598, 2023.

[78] *Interactive content-based retrieval in video databases using fuzzy classification and relevance feedback*. *Proceedings IEEE International Conference on Multimedia Computing and Systems*, τόμος 2, σελίδες 954–958. IEEE, 1999.

[79] *Image indexing and retrieval using expressive fuzzy description logics.*

[80] Yannis Avrithis, Yiannis Xirouhakis και Stefanos Kollias. *Affine-invariant curve normalization for object shape representation, classification, and retrieval*. *Machine Vision and Applications*, 13:80–94, 2001.

[81] Deema Abdal Hafeth, Stefanos Kollias και Mubeen Ghafoor. *Semantic Representations with Attention Networks for Boosting Image Captioning*. *IEEE Access*, 2023.

[82] *Bottom-up spatiotemporal visual attention model for video analysis*. *IET Image Processing*, 1(2):237–248, 2007.

[83] *A snake model for object tracking in natural sequences*. *Signal processing: image communication*, 19(3):219–238, 2004.

[84] Manolis Wallace, Ilias Maglogiannis, Kostas Karpouzis, George Kormentzas και Stefanos Kollias. *Intelligent one-stop-shop travel recommendations using an adaptive neural network and clustering of history*. *Information Technology & Tourism*, 6(3):181–193, 2003.

[85] *Optimal filter banks for signal reconstruction from noisy subband components*. *IEEE transactions on signal processing*, 44(2):212–224, 1996.

[86] *Mikkel's Twitter*. https://twitter.com/MikkelTokvam. Access Date: 06-05-2023.

List of Abbreviations

FPL	Fantasy Premier League
EV	Expected Value
PV	Possession Value
xG	Expected Goals
xGA	Expected Goals Against
npxG	Non-penalty Expected Goals
xA	Expected Assists
npg	non-penalty goals
p90	per 90 minutes
xMins	Expected Minutes
xPoints	Expected Points
GW	Gameweek
API	Application Programming Interface
SPI	Soccer Power Index
XGBoost	Extreme Gradient Boosting
RR	Ridge Regression
GBM	Gradient Boosting Machines
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
GKP	Goalkeeper
DEF	Defender
MID	Midfielder
FWD	Forward
GM	General Manager
ROI	Return On Investment
MDP	Markov Decision Process
NLP	Natural Language Processing
FF	Fantasy Football
NTUA	National Technical University of Athens
AILS LAB	Artificial Intelligence and Learning Systems Laboratory
LP	Linear Programming
MILP	Mixed Integer Linear Programming