



**ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**

**ΠΜΣ ΔΙΟΙΚΗΣΗ, ΑΝΑΛΥΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ
ΣΥΣΤΗΜΑΤΑ ΕΠΙΧΕΙΡΗΣΕΩΝ
(M.Sc. in Business Administration, Analytics and Information Systems)**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**NBA ANALYTICS : ΠΡΟΒΛΕΨΗ ΜΕ ΒΟΗΘΕΙΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ
ΜΑΝΗΣ ΔΗΜΟΣΘΕΝΗΣ**

**ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ
ΛΑΖΑΡΟΥ ΒΑΣΙΛΕΙΟΣ**

ΑΘΗΝΑ

2022



**ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ
ΤΜΗΜΑ ΟΙΚΟΝΟΜΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**

**ΠΜΣ ΔΙΟΙΚΗΣΗ, ΑΝΑΛΥΤΙΚΗ ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΑ
ΣΥΣΤΗΜΑΤΑ ΕΠΙΧΕΙΡΗΣΕΩΝ
(M.Sc. in Business Administration, Analytics and Information Systems)**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**NBA ANALYTICS : ΠΡΟΒΛΕΨΗ ΤΗΣ ΑΠΟΔΟΣΗΣ ΕΝΟΣ ΑΘΛΗΤΗ ΜΕ
ΒΟΗΘΕΙΑ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ
ΜΑΝΗΣ ΔΗΜΟΣΘΕΝΗΣ**

**ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ
ΛΑΖΑΡΟΥ ΒΑΣΙΛΕΙΟΣ**

ΑΘΗΝΑ

2022

Copyright © Μανής Δημοσθένης, 2022

Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Οικονομικών Επιστημών (ΠΜΣ Διοίκηση, Αναλυτική και Πληροφοριακά Συστήματα Επιχειρήσεων) του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών δεν υποδηλώνει απαραίτητως και αποδοχή των απόψεων του συγγραφέα εκ μέρους του Τμήματος.

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΛΗΨΗ	6
ΕΙΚΟΝΕΣ	7
ΠΙΝΑΚΕΣ	8
ΕΙΣΑΓΩΓΗ.....	9
1.NBA Analytics	9
1.1 SWOT Ανάλυση.....	10
1.1.1 Πλεονεκτήματα.....	10
1.1.2 Μειονεκτήματα.....	15
1.1.3 Ευκαιρίες.....	16
1.1.4 Απειλές.....	17
2. Μηχανική Μάθηση.....	19
2.1 Είδη μηχανικής Μάθησης	20
2.2 Αλγόριθμοι Μηχανικής Μάθησης.....	21
3. Εφαρμογή Μηχανικής Μάθησης στο NBA.....	23
3.1 Μεθοδολογία	23
3.2 Δεδομένα	25
3.2.1 Συλλογή και Επεξεργασία Δεδομένων.....	25
3.2.2 Προετοιμασία και καθαρισμός δεδομένων.....	29
3.3 Διερευνητική Ανάλυση.....	31
3.4 Εφαρμογή Μεθόδων.....	39
3.4.1 Πρώτη Εφαρμογή Γραμμικής Παλινδρόμησης.....	39
3.4.2 Επιλογή Συντελεστών.....	44
3.4.3 Εφαρμογή Δέντρων Αποφάσεων.....	55
3.4.4 Εφαρμογή Gradient Boosting Trees.....	63
4. Συμπέρασμα.....	65
ΒΙΒΛΙΟΓΡΑΦΙΑ.....	68

ΠΕΡΙΛΗΨΗ

Η διπλωματική εργασία ασχολείται με την ανάλυση δεδομένων και εφαρμογή μεθόδων μηχανικής μάθησης για την απόκτηση χρήσιμων πληροφοριών για τις ομάδες μπάσκετ του NBA. Αρχικά παρουσιάστηκαν, μέσω μίας ανάλυσης SWOT, τα άμεσα πλεονεκτήματα (strengths), τα άμεσα μειονεκτήματα (weaknesses), οι ευκαιρίες που ενδέχεται να παρουσιαστούν καθώς και οι απειλές που ενδέχεται να παρουσιαστούν. Με αυτό τον τρόπο παρουσιάζεται η πολύπλευρη χρησιμότητα της ανάλυσης δεδομένων και της μηχανικής μάθησης στις ομάδες του NBA. Στη συνέχεια γίνεται μια συνοπτική εισαγωγή στην μηχανική μάθηση, στα είδη αυτής, καθώς και σε κάποιους βασικούς αλγορίθμους. Έπειτα καθορίζεται ο στόχος του πρακτικού μέρους της εργασίας και η μεθοδολογία για να επιτευχθεί. Το πρακτικό κομμάτι αποτελείται από την προσπάθεια εύρεσης ενός μοντέλου πρόβλεψης της απόδοσης ενός παίκτη στη καριέρα του μέσω της μεταβλητής EFF καριέρας. Το πρακτικό κομμάτι ξεκινάει με την προετοιμασία των δεδομένων και την περιγραφή των μεταβλητών. Μετά από μία περιγραφική ανάλυση του συνόλου δεδομένων, ξεκινάει η εφαρμογή των αλγορίθμων μηχανικής μάθησης στο σύνολο δεδομένων. Αρχικά γίνεται η εφαρμογή γραμμικής παλινδρόμησης και στη συνέχεια των αλγορίθμων βασισμένων σε δέντρα αποφάσεων. Κατά την διάρκεια της εφαρμογής γίνεται επιλογή συντελεστών έτσι ώστε να αντιμετωπιστούν τα προβλήματα που δημιουργούνται. Τέλος, γίνεται σύγκριση των αποτελεσμάτων από την εφαρμογή των μοντέλων και εξάγονται τα συμπεράσματα που χρειαζόμαστε.

Λέξεις Κλειδιά: NBA Analytics, Μηχανική Μάθηση, Γραμμική Παλινδρόμηση, Δέντρα Αποφάσεων

SUMMARY

NBA ANALYTICS: Predicting Player Performance via Machine Learning

In this paper the importance of data analysis and machine learning usage from NBA teams is presented. In the first section, a SWOT analysis of NBA teams that adopted data analysis and machine learning techniques is shown. Strengths and weaknesses will include the direct advantages and disadvantages from analytics adoption and opportunities and threats will include future positive and negative effects that may occur. After SWOT analysis an introduction is made, about machine learning types and some basic algorithms. Next, the thesis goal and thesis methodology are defined, which is predicting a player career performance from stats of its first rookie years. The performance measured with EFF career stat. First, we prepared the data and made an explanatory analysis on them. Then we apply machine learning algorithms in our dataset. Specifically linear regression, decision tree and gradient tree boosting are used. Subsequently adjustments in algorithms have been made, based on problems that occurred. Finally results and conclusions are presented.

Key Words: NBA Analytics, Machine Learning, Linear Regression, Decision Trees

ΕΙΚΟΝΕΣ

Εικόνα 1 : SPORT VU Camera Tracking.....	11
Εικόνα 2 : Field Goal Percentage by Attempted Shot Location.....	12
Εικόνα 3 : :Points Per Shot.....	12
Εικόνα 4 : Ετήσιος αριθμός κυβερνο-επιθέσεων	19
Εικόνα 5: Machine Learning Flowchart.....	20
Εικόνα 6: Decision Tree.....	26
Εικόνα 7: Ιστόγραμμα μεταβλητής EFF.....	34
Εικόνα 8: Ιστόγραμμα Μεταβλητής AGE.....	36
Εικόνα 9: Πλήθος της Μεταβλητής Tm.....	37
Εικόνα 10: Correlation Heatmap.....	39
Εικόνα 11: Πραγματικές - Προβλεφθείσες Τιμές(Linear Regression	43
Εικόνα 12: Ιστόγραμμα Καταλοίπων.....	44
Εικόνα 13: Ιδανικός αριθμός μεταβλητών.....	47
Εικόνα 14: Πραγματικές - Προβλεφθείσες Τιμές.....	49
Εικόνα 15: Πραγματικές - Προβλεφθείσες Τιμές.....	51
Εικόνα 16: Αρχικό Δέντρο.....	57
Εικόνα 17: Πραγματικές-Προβλεφθείσες Τιμές.....	58
Εικόνα 18 : Depth-alpha.....	59
Εικόνα 19: MSE-alpha.....	60
Εικόνα 20: Pruned Tree.....	61
Εικόνα 21: Feature Importance Decision Tree.....	62
Εικόνα 22: Gradient Tree Boosting Algorithm.....	63
Εικόνα 23: Predicted-True Values(Gradient Boosting).....	64
Εικόνα 24: Κατάλοιπα Gradient Boosting.....	65
Εικόνα 25: Feature Importnce Gradient Boosting.....	66
Εικόνα 26 : Error Comparison.....	67

ΠΙΝΑΚΕΣ

Πίνακας 1:Σύνολο δεδομένων1.....	27
Πίνακας 2:Επεξήγηση Δεδομένων	30
Πίνακας 3:Μέτρα Θέσης και Διασποράς	32
Πίνακας 4:Μέσο EFF ανά ομάδα.....	38
Πίνακας 5: Δημιουργία dummy μεταβλητών αντί της μεταβλητής Pos.....	40
Πίνακας 6: Διαφορά πραγματικών έναντι προβλεφθέντων τιμών.....	42
Πίνακας 7: Αρχικές Τιμές VIF.....	54
Πίνακας 8: Συντελεστές Παλινδρόμησης	55
Πίνακας 9: Τυποποιημένοι Συντελεστές Παλινδρόμησης.....	56
Πίνακας 10: Training and Test Error.....	67
Πίνακας 11: EFF Luka Doncic.....	68

ΕΙΣΑΓΩΓΗ

Οι επιχειρήσεις για να καταφέρουν να επιβιώσουν τον ανταγωνισμό χρειάζεται να πάρουν σωστές αποφάσεις. Η κρισιμότητα των αποφάσεων αυτών δείχνει το πόσο πολύτιμες είναι οι πληροφορίες και τα δεδομένα για τις επιχειρήσεις. Στις κορυφαίες ομάδες μπάσκετ, όπου τα κέρδη είναι τεράστια και κάθε απόφαση της διοίκησης ή του τεχνικού επιτελείου είναι ζωτικής σημασίας, είναι αναγκαίο να εκμεταλλευτούν τα δεδομένα για να αποκτήσουν πλεονέκτημα εναντίον του ανταγωνισμού. Οι κορυφαίες ομάδες καλαθοσφαίρισης, κυρίως οι ομάδες που αγωνίζονται στο NBA, τα τελευταία χρόνια προσπαθούν, εφαρμόζοντας τεχνικές data mining και business intelligence να αναλύσουν μεγάλες ποσότητες δεδομένων που θα τις βοηθήσουν στη λήψη αποφάσεων, που θα τις οδηγήσουν στις επιτυχίες. Μέσω εφαρμογής μεθόδων μηχανικής μάθησης στα δεδομένα τους, οι ομάδες προσπαθούν να πάρουν αποφάσεις που θα τις εξασφαλίσουν την επιτυχία.

Στην εργασία θα εξεταστούν οι τρόποι με τους οποίους οι ομάδες παίρνουν τα δεδομένα τους και τα αναλύουν καθώς και τα πλεονεκτήματα μειονεκτήματα από την εφαρμογή αυτή. Στη συνέχεια θα γίνει μια προσπάθεια να εφαρμοστεί στην πράξη, μέσω της εύρεσης μοντέλου μηχανικής μάθησης που θα προβλέπει την απόδοση ενός παίκτη σε όλη την καριέρα του, από τα στατιστικά των δύο πρώτων σεζόν που αγωνίστηκε στον NBA. Η ποσοτικοποίηση της απόδοσης θα γίνει βάση της μεταβλητής EFF.

Με αυτό τον τρόπο οι ομάδες θα μπορούν να αποφασίσουν νωρίς αν θα επενδύσουν σε ένα αθλητή για να γίνει Franchise Player.

1 NBA Analytics

Ένα από τα πιο δημοφιλή αθλήματα στις μέρες μας είναι το άθλημα της καλαθοσφαίρισης. Το συγκεκριμένο άθλημα είναι εξαιρετικά δημοφιλές με αποτέλεσμα οι κορυφαίες ομάδες να δημιουργούν τεράστια έσοδα. Το κορυφαίο πρωτάθλημα στον κόσμο είναι το NBA (National Basketball Association), ένα πρωτάθλημα 30 ομάδων που αγωνίζονται μεταξύ τους κάθε χρόνο με σκοπό την κατάκτηση του πρωταθλήματος. Τα έσοδα του NBA, 8.3 δισεκατομμύρια δολάρια τη σεζόν 2018-2019 (cnbc.com), αλλά και των ομάδων που αγωνίζονται σε αυτό, για παράδειγμα οι Golden State Warriors είναι 474 εκατομμύρια δολάρια έσοδα το 2019-2020 (statista.com) δείχνουν ότι δημοφιλία της καλαθοσφαίρισης μετουσιώνετε σε τεράστια κέρδη για τις ομάδες.

Η καλαθοσφαίριση σαν άθλημα παράγει πάρα πολλά δεδομένα, από τα βασικά στατιστικά που παράγουν οι αθλητές κατά την διάρκεια του αγώνα όπως οι πόντοι τα ριμπάουντ, οι τελικές πάσες και τα λάθη, μέχρι πιο προχωρημένες μετρικές. Με τον όρο NBA Analytics εννοούμε τις μεθόδους και τις τεχνικές με τις οποίες οι ομάδες καλαθοσφαίρισης του NBA αποκτούν αυτά τα δεδομένα και τα επεξεργάζονται για να πάρουν σωστές αποφάσεις. Για να καταλάβουμε τον λόγο που οι ομάδες διαθέτουν ένα μεγάλο μέρος του προϋπολογισμού τους για την δημιουργία και την λειτουργία τμήματος Analytics θα πρέπει να δούμε τα οφέλη που αποκομίζουν σε σύγκριση με τις δυσκολίες ή τις απειλές που προκύπτουν από αυτά.

1.1 SWOT Analysis

Οι αθλητικές ομάδες διαθέτουν τον προϋπολογισμό τους για την απόκτηση ικανών αθλητών, στελέχωση προπονητικού και ιατρικού επιτελείου έτσι ώστε να επιτύχουν τους στόχους τους. Τα τελευταία χρόνια με την άνοδο της επιστήμης της πληροφορικής και την ανάλυση δεδομένων από τις επιχειρήσεις, όλο και περισσότερες ομάδες επενδύουν ένα μέρος του προϋπολογισμού για την δημιουργία τμημάτων analytics.

Για να κριθεί το αν οι ομάδες χρειάζεται να επενδύσουν πάνω σε αυτό, πρέπει να κατανοηθούν τα οφέλη που θα αποκομίσουν αλλά και οι αδυναμίες που θα παρουσιαστούν. Για να γίνει αυτό θα χρησιμοποιηθεί η ανάλυση SWOT (Strength, Weakness, Opportunities, Threats) σε μία ομάδα που αποφασίζει να χρησιμοποιήσει την επιστήμη της ανάλυσης δεδομένων για να πάρει αποφάσεις.

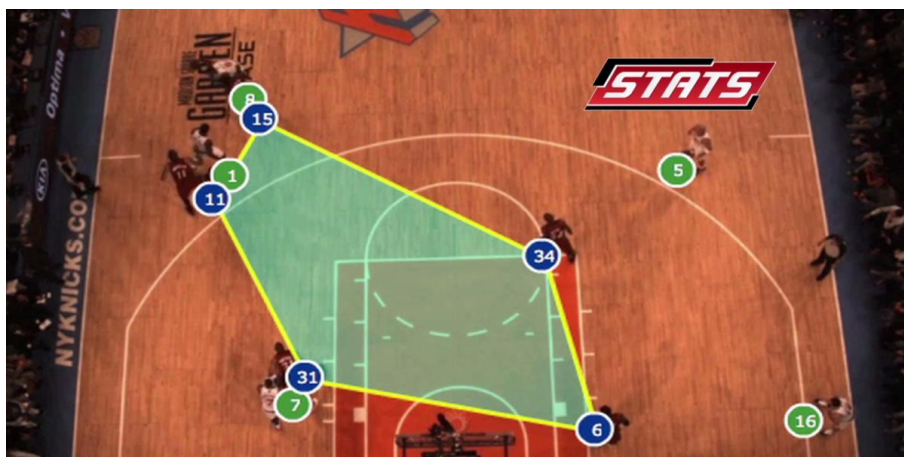
1.1.1 Πλεονεκτήματα

Το πρώτο μέρος της SWOT ανάλυσης είναι τα πλεονεκτήματα που αποκομίζει η ομάδα, δηλαδή οι τρόποι με το οποίους χρησιμοποιεί μεθόδους ανάλυσης δεδομένων για κερδίσει συγκριτικό πλεονέκτημα έναντι του ανταγωνισμού της.

Εύρεση Αποτελεσματικής Στρατηγικής

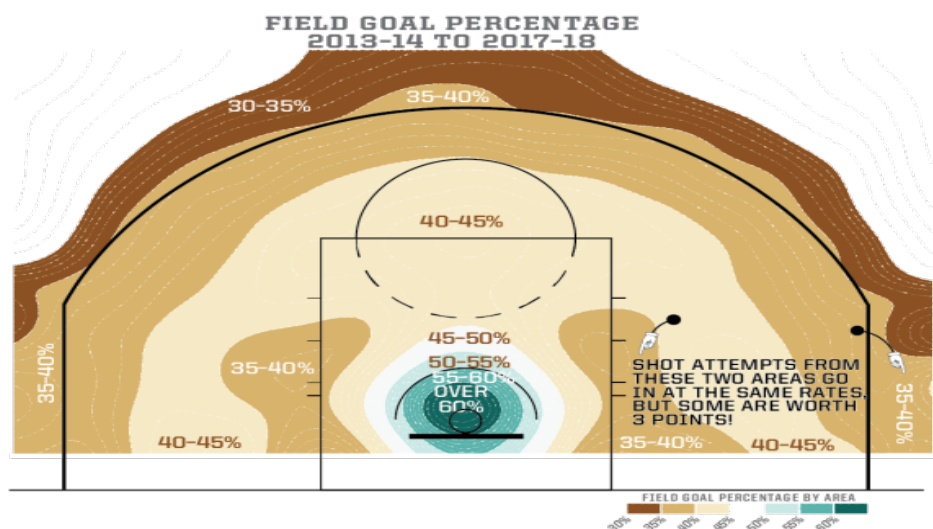
Για τις αθλητικές ομάδες, το προϊόν είναι το ίδιο το άθλημα. Η επιτυχία τους σε αυτό αποφέρει έσοδα, μέσω εισιτηρίων, διαφημιστικών συμφωνιών, πώληση εμπορευμάτων (φανέλες κ.α.), καθώς και έσοδα από έπαθλα. Οπότε η εύρεση της κατάλληλης στρατηγικής για την επίτευξη της νίκης δίνει σημαντικό πλεονέκτημα στην ομάδα.

Ως προς την εξόρυξη των δεδομένων που αναλύουν οι ομάδες για την εύρεση των αποτελεσματικών τακτικών, πέρα από τα απλά στατιστικά των αθλητών (πόντους, ριμπάουντ κ.α) οι ομάδες εγκαθιστούν στα γήπεδα κάμερες για συλλογή πιο εξιδεικευμένων δεδομένων από τους αγώνες. Αυτές οι κάμερες παρακολουθούν τις συντεταγμένες X, Y και Z των παικτών, των διαιτητών και της μπάλας 25 φορές το δευτερόλεπτο και συλλέγουν συνολικά περισσότερα από 1 εκατομμύριο δεδομένα σε ένα μέσο παιχνίδι NBA. Είναι δηλαδή ένα σύστημα παρακολούθησης των ενεργειών του αθλητή μέσα στο γήπεδο. Στη συνέχεια το πρόγραμμα χρησιμοποιεί συγκεκριμένους αλγόριθμους για να τον εντοπισμό σημαντικών γεγονότων στο παιχνίδι και να προχωρήσει στην ανάλυση αυτών. Το πλήθος των δεδομένων που δημιουργούνται δίνουν στην ομάδα πρόσβαση σε πιο προηγμένες μετρικές όπως το ποσοστό σουτ με βάση το χρόνο που ο αθλητής ελέγχει την μπάλα, τον χρόνο κατοχής, τις δευτερεύουσες ασίστ, τις ασίστ ελεύθερων βολών, τα ριμπάουντ με βάση αμφισβητούμενα ριμπάουντ, τη δραστηριότητα ανάλογα την τοποθεσία στο γήπεδο, την ταχύτητα των αθλητών, την απόσταση που διένυσαν κ.α.(
<https://www.sportsvideo.org/>)

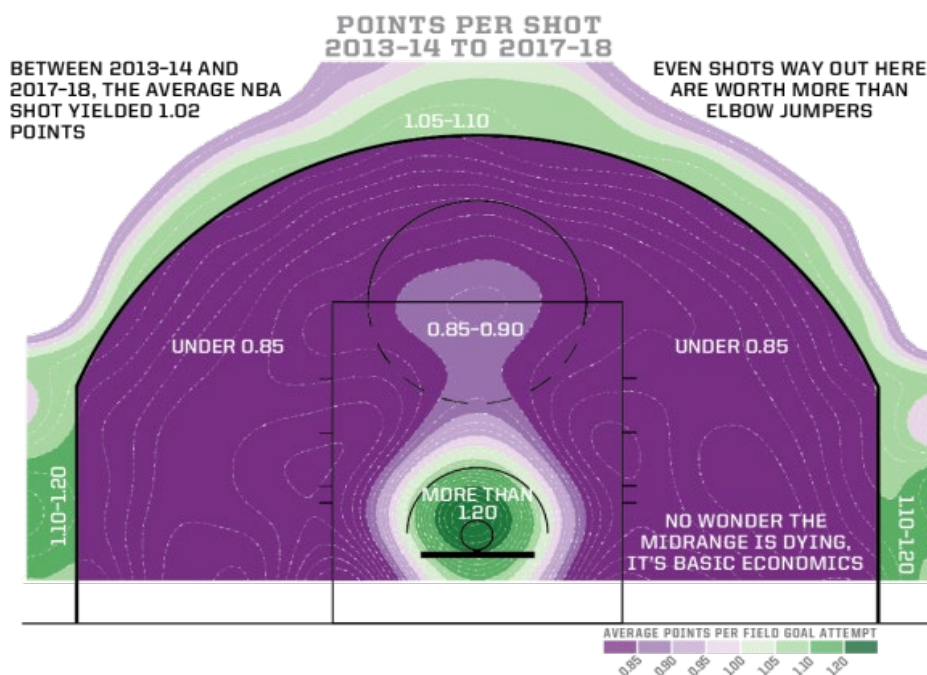


Εικόνα 1:SPORT VU Camera Tracking (Πηγή : sporttechie.com)

Οι ομάδες συλλέγουν αυτά τα δεδομένα από τις κάμερες για να δημιουργήσουν ψηφιακές αναπαραστάσεις όπως είναι οι χαρτογραφήσεις του σουτ. Σε μία απεικόνιση του ποσοστού ευστοχίας όλων των σουτ που πραγματοποιήθηκαν τις σεζόν 2013-2014 έως 2017-2018 (Kirk Goldsberry, 2019) παρατηρήθηκε μεγάλη εξάρτηση τις ευστοχίας των σουτ ανάλογα με την τοποθεσία στο γήπεδο. Για παράδειγμα την μεγαλύτερη ευστοχία είχαν τα σουτ κάτω από το καλάθι ενώ παρατηρήθηκε ότι σε ορισμένα σημεία, τα σουτ δύο πόντων είχαν το ίδιο ποσοστό ευστοχίας με τα σουτ τριών πόντων (Εικόνα 2) . Σε μια περαιτέρω ανάλυση απεικονίσθηκαν οι πόντοι ανά σουτ βάση της τοποθεσίας στο γήπεδο και παρατηρήθηκε ότι οι περισσότεροι πόντοι ανά σουτ στα σουτ τριών πόντων και στα σουτ κάτω από το καλάθι (Εικόνα 3). Μια τέτοια ανάλυση δίνει εύκολα συμπεράσματα ως προς την αποτελεσματικότητα συγκεκριμένων τακτικών , όπως για παράδειγμα το να επικεντρώνεται μία ομάδα σε επιθέσεις με σουτ τριών πόντων ή την μείωση των επιθέσεων που οδηγούν σε σουτ μέσης απόστασης.



Εικόνα 2:Field Goal Percentage by Attempted Shot Location (Πηγή : fivethirtyeight.com)



Εικόνα 3: Points Per Shot (Πηγή : fivethirtyeight.com)

Οι ομάδες μπορούν να χρησιμοποιήσουν τα δεδομένα που δημιουργούνται από τις κάμερες με ακόμα περισσότερους τρόπους, όπως η εύρεση, η ανάλυση και η κατανόηση των μοτίβων των παικτών με την βοήθεια deep learning τεχνικών (Nistala, Guttag 2018). Συγκεκριμένα πραγματοποιήθηκε μια αναπαράσταση των κινήσεων το παικτών σε κάθε επιθετική κατοχή, μέσω απεικόνισης της τροχιάς που πραγματοποιεί ο αθλητής σε εικόνας, η απεικόνιση αυτή χρησιμοποιεί τα δεδομένα κίνησης σε άξονες x και y ως χρονοσειρές για να καταγράψει την κίνηση στην εικόνα. Τρεις εκατομμύρια εικόνες κατασκευάστηκαν από τρεις σεζόν παρακολούθησης παικτών του NBA. Η ανάγκη κωδικοποίησης για κάθε εικόνα έτσι ώστε να αποτυπώνει τη σημασιολογία του τι συνέβη κατά τη διάρκεια της κατοχής, οδήγησε στη χρησιμοποίηση συνελκτικού νευρωνικού δίκτυο. Η μάθηση στο νευρωνικό δίκτυο κωδικοποίησης χαμηλών διαστάσεων για κάθε εικόνα τροχιάς του επιτρέπει να κρατάει τις πιο σημαντικές λεπτομέρειες της κίνησης.

Τα δεδομένα που δημιουργούνται μπορούν να αξιοποιηθούν από την ομάδα με διάφορους τρόπους, όπως για παράδειγμα η δημιουργία μιας σχεσιακής βάσης δεδομένων με όλες τις κατοχές. Μέσω αυτής της βάσης η ομάδα μπορεί να χρησιμοποιεί ερωτήματα (queries) όπως για παράδειγμα να βρει όλες τις κατοχές-επιθέσεις από μία σεζόν στις οποίες οι Μιλγουόκι Μπακς πραγματοποιούν μια επίθεση με pick and roll από την αριστερή πτέρυγα. Αν στη σχεσιακή βάση δεδομένων προστεθούν και άλλα δεδομένα όπως τα επιτυχημένα καλάθια, τα ριμπάουντ τα λάθη ή οι ασιστ κ.α, η ομάδα μπορεί να χρησιμοποιήσει ερωτήματα, όπως για παράδειγμα να βρει όλες τις κατοχές στις οποίες ένας αθλητής, για παράδειγμα ο ΛεΜπρόν Τζέιμς κάνει διείσδυση (drive) στο καλάθι από τα δεξιά και σκοράρει όταν απέναντι του στην άμυνα είναι ο Ντρέιμοντ Γκριν.

Επι προσθέτως τα δεδομένα αυτά, μπορούν να χρησιμοποιηθούν για την συσταδοποίηση (clustering) των κινήσεων των παιχτών, μέσω της δημιουργίας cluster profiles, στη συγκεκριμένη έρευνα παρατηρήθηκε ότι οι παίκτες όπως Stephen Curry και LeBron James έχουν παρόμοια profiles σε αντίθεση με παίκτες όπως Andre Drummond, κάτι που είναι πολύ λογικό καθώς Curry και James χειρίζονται αρκετά την μπάλα και δημιουργούν επιθέσεις, είτε για των εαυτό τους, είτε για τους συμπαίκτες τους, σε αντίθεση με τον Andre Drummond που είναι center με έφεση στο παιχνίδι κάτω από την ρακέτα. Σε περαιτέρω ανάλυση τα clusters profiles μπορούν να χρησιμοποιηθούν για την εύρεση μοτίβων στις κινήσεις των αθλητών, για παράδειγμα οι κινήσεις του Paul George ,μετά των τραυματισμό του το καλοκαίρι του 2014 ,στο το δεξί μισό του γηπέδου αυξήθηκε από 16,5% σε 46,6%, ενώ οι κινήσεις στο αριστερό μισό του γηπέδου μειώθηκαν από 54,1% σε 31,0%, ενώ ο Kevin Love μετά μετακίνηση του στους Cleveland Cavaliers , είδε τις κινήσεις του από την αριστερή πλευρά του low-post έως την κορυφή του κλειδιού (ο χώρος που βρίσκεται ακριβώς μπροστά από τη γραμμή των τριών πόντων και μόλις περνάει τη γραμμή των ελεύθερων βολών), να μειώνονται από 17,9% σε 9,4% και οι κινήσεις του αριστερά από το ζωγραφιστό να μειώνονται από 20,9% σε 13,3%, κάτι που δείχνει τον διαφορετικό ρόλο που είχε ο Love στους Minnesota Timberwolves σε αντίθεση με τον ρόλο του στους Cleveland Cavaliers που ήταν τρίτος σε ιεραρχία μετά από LeBron James και Kyrie Irving.

Το ESPN το 2015 (<https://businessoverbroadway.com/>) κατέταξε τις Αμερικάνες ομάδες καλαθοσφαίρισης, μπιέζμπολ , αμερικάνικου ποδοσφαίρου και χόκει σχετικά με τη χρήση ανάλυσης δεδομένων στη λειτουργία τους . Οι ειδικοί και οι ερευνητές του ESPN αξιολόγησαν ομάδες από το NFL, NBA, NHL και MLB με βάση τη πόσο προσωπικό απασχολεί στα analytics,την αποδοχή των analytics από στελέχη και προπονητές, το πόσο έχει επενδύσει σε απόκτηση και ανάλυση βιομετρικών δεδομένων και τέλος στον βαθμό στον οποίο η συνολική επιχειρησιακή προσέγγιση βασίζεται στα analytics. Μετά την αξιολόγηση οι ομάδες κατατάχθηκαν σε πέντε κατηγορίες : Non Believers (μηδενική εφαρμογή Analytics) ,Skeptics (πολύ μικρή εφαρμογή) , One Foot In (μέτρια εφαρμογή analytics στην λειτουργία της ομάδας , Believers (σημαντική εφαρμογή), All-in (πρωτοπόροι στην εφαρμογή analytics, βασίζουν τις βραχυχρόνιες και μακροχρόνιες αποφάσεις τους σε συμπεράσματα που πάρθηκαν με την χρήση machine learning και data analysis). Στη συνέχεια μετρήθηκε το ποσοστό νικών της προηγούμενης σεζόν (2014) και βρέθηκε ότι οι ομάδες που ανήκουν στην χαμηλότερη κατηγορία υιοθέτησης μεθόδων αναλυτικής έχουν το χαμηλότερο ποσοστό, ενώ για το NBA συγκεκριμένα οι ομάδες που ανήκουν στην κατηγορία All-in είχαν το υψηλότερο ποσοστό νικών. Τέλος παρατηρήθηκε ότι οι περισσότερες ομάδες ανήκαν στις κατηγορίες Believers και One Foot In στην κατηγορία Skeptics ανήκουν οι λιγότερες ομάδες.

Ατομική Ανάπτυξη Αθλητών και Πρόληψη Τραυματισμών

Οι αθλητές αποτελούν το σημαντικότερο κομμάτι ενός αθλητικού οργανισμού, είναι ο κύριος παράγοντας προστιθέμενης αξίας στις ομάδες. Είτε από την επίτευξη των στόχων μέσω της αγωνιστικής απόδοσης των αθλητών, είτε μέσω εμπορικών δραστηριοτήτων όπως η πώληση εμφανίσεων των αθλητών, είναι εμφανές ότι η επιτυχημένη λειτουργία της ομάδας βασίζεται σε αυτούς. Είναι σαφές οπότε για τις ομάδες να βρίσκουν μεθόδους για να διατηρούν τους αθλητές σε καλή φυσική κατάσταση. Ο σημαντικότερος κίνδυνος για τις ομάδες είναι οι τραυματισμοί των αθλητών. Ενδεικτικό είναι ότι οι ομάδες του NBA ξοδύσανε την σεζόν 2014-2015 358 εκατομμύρια δολάρια (fastcompany.com) σε αμοιβές τραυματισμένων αθλητών καθώς και σε αμοιβές αθλητών που αποκτήθηκαν για να

αντικαταστήσουν τραυματισμένους μπασκετμπολίστες. Φαίνεται και από τις οικονομικές επιπτώσεις, η ανάγκη για εύρεση τρόπων για την βελτίωση της φυσικής κατάστασης και πρόληψής των τραυματισμών των αθλητών.

Οι ομάδες με την βοήθεια ειδικών συσκευών με αισθητήρες (wearables) αποκτούν πρόσβαση σε πληθώρα δεδομένων αναφορικά με την φυσική κατάσταση των αθλητών. Ποιο συγκεκριμένα οι Toronto Raptors χρησιμοποιούν στις προπονήσεις ένα μηχανήμα μεγέθους κινητού τηλεφώνου που ονομάζεται OptimEye. Το γυροσκόπιο και το επιταχυνσιόμετρο του OptimEye παρέχουν δεδομένα σχετικά με τον τρόπο με τον οποίο οι παίκτες κινούνται και αποδίδουν κατά την διάρκεια της προπόνησης, καθώς και δεδομένα για την φυσική τους κατάσταση αλλά και το επίπεδο κόυρασης των παιχτών. Ενδεικτικά αναφέρεται ότι οι Toronto Raptors δύο σεζόν πριν την εφαρμογή του OptimEye στις προπονήσεις, ήταν μέσα στις ομάδες με τους περισσότερους τραυματισμούς ενώ δύο σεζόν μετά, το 2014 βρίσκονταν στις ομάδες με τους λιγότερους τραυματισμούς, επίσης κατά την διάρκεια της σεζόν 2018-2019 οι Raptors αποφάσισαν να μην χρησιμοποιήσουν τον καλύτερο παίκτη τους, τον Kawī Leonard σε 22 αγώνες κατά την διάρκεια της κανονικής περιόδου. Η διαχείριση φόρτου αυτή ήταν εναρμονισμένη από την ανάλυση των δεδομένων που παράγονται από την χρήση wearables. Στο τέλος της σεζόν ο Leonard αναδείχθηκε MVP των τελικών και οι Raptors πρωταθλητές (brainstation.io).

Τα δεδομένα που αποκομίζουν οι ομάδες από τα wearables αλλά και τα πιο απλά δεδομένα που έχουν οι ομάδες στη διάθεσή τους, όπως τα πόσα παιχνίδια παίζει ένα παίκτης ή το Usage Rate, το ποσοστό δηλαδή των επιθέσεων στο οποίο συμμετείχε ένας παίκτης ενώ βρισκόταν στο παρκέ, με την προϋπόθεση ότι το παιχνίδι τελειώνει με ένα από τα τρία αποτελέσματα: επίτευξη καλαθιού, προσπάθεια ελεύθερων βολών ή λάθος(bleacherreport.com), χρησιμοποιούνται σε μοντέλα μηχανικής μάθησης για πρόβλεψη τραυματισμών. Με εφαρμογή τέτοιων μοντέλων έχει βρεθεί ότι όσων αφορά τους τραυματισμούς που συμβαίνουν εντός του γηπέδου, ο πιο σημαντικός παράγοντας είναι η απόδοση των παικτών στους πιο πρόσφατους αγώνες. Όταν ένας αθλητής παίζει πιο ενεργά και παίρνει περισσότερες προσπάθειες, κυρίως από σουτ τριών πόντων έχει περισσότερες πιθανότητες να τραυματιστεί. Παρατηρήθηκε επίσης ότι οι παράγοντες, βάρος και ύψος όσο και τα συνεχόμενα παιχνίδια δεν είναι συσχετισμένες με την πιθανότητα για τραυματισμό(Wangwei Wu., 2020).

Αποτελεσματική Αναζήτηση Αθλητών (Scouting)

Το πιο σημαντικό κομμάτι για τις ομάδες είναι η στελέχωσή τους από ικανούς αθλητές. Είτε μέσω πληρωμής υψηλών συμβολαίων σε ήδη καταξιωμένους μπασκετμπολίστες, είτε απόκτησης νεαρών παιχτών, οι ομάδες θέλουν να βρουν τους αθλητές που θα τους βοηθήσουν να κερδίσουν. Η εξόρυξη δεδομένων για πληροφορίες παικτών που αγωνίζονται σε κολλέγια και λύκεια στην Αμερική, ή ακόμα και στην Ευρώπη δίνει την δυνατότητα στην ομάδα να διαλέξει αθλητές λιγότερο γνωστούς, που είναι ιδιαίτερα φθινοί, ή να ανακαλύψουν ικανούς μπασκετμπολίστες σε χαμηλές θέσεις του Draft (Draft: Ετήσια επιλογή των δικαιωμάτων νεαρών αθλητών από τα κολλέγια ή από την Ευρώπη για τις ομάδες του NBA). Ένα παράδειγμα είναι η επιλογή του Pascal Siakam από τους Toronto Raptors, μίας ομάδας που σε συνεργασία με την IBM έχει δημιουργήσει ένα data-driven επιτελείο αποφάσεων. Ο Pascal Siakam επιλέχθηκε το 2016 στη θέση 27 του draft και το 2019 αναδείχθηκε πιο βελτιωμένος παίκτης του NBA(IBM.com). Πάνω στην εύρεση παιχτών που ενδέχεται να βελτιωθούν σημαντικά στην καριέρα τους, ιδιαίτερο ενδιαφέρον έχει η προσπάθεια να προβλεφθεί μέσω μηχανικής μάθησης το αν ένας

παίχτης είναι rising star, (Mahmood, Daud 2020) με βάση την επιρροή του στους συμπαίκτες του, η αποτελεσματικότητα της πρόβλεψης φαίνεται στο ότι η σύγκριση των 20 κορυφαίων, βάση της κατάταξής τους, rising stars, με την κατάταξη τους για τις επόμενες 6 σεζόν δείχνει ότι οι περισσότεροι από αυτούς είναι στους κορυφαίους 100 του πρωταθλήματος.

Ο τρόπος που μία ομάδα μπορεί να χρησιμοποιήσει machine learning μεθόδους για αποτελεσματικό scouting φαίνεται από την προσπάθεια που έγινε (Gerasimos Plegas, 2022) να προβλέψει το μοντέλο μηχανική μάθησης, τον κατάλληλο shooting guard για την ομάδα των Milwaukee Bucks. Αρχικά βρέθηκαν ,με την βοήθεια μεθόδων clustering αλλά και με την εμπειρική γνώση στο άθλημα , τα πιο σημαντικά χαρακτηριστικά που είναι σημαντικά σχετικά με την απόδοση ενός shooting guard, και στη συνέχεια έγινε προσπάθεια πρόβλεψης της απόδοσης των τριών υποψηφίων για την θέση του shooting guard (Jrue Holiday , Bogdan Bogdanovic και Danny Green) και αξιολόγηση της απόδοσης με βάση τα σημαντικά χαρακτηριστικά που βρέθηκαν προηγούμενος. Η αξιολόγηση έδειξε τον Jrue Holiday ως την ιδανική επιλογή, και στην πραγματικότητα όντως οι Milwaukee Bucks επέλεξαν τον Jrue Holiday ως το βασικό τους shooting guard.

1.1.2 Μειονεκτήματα-Αδυναμίες

Εκτός από το τι θα κερδίσει ο ομάδα, είναι σημαντικό να κατανοήσουμε το τι μειονεκτήματα θα προκύψουν από την δημιουργία τμήματος ανάλυσης δεδομένων σε μια ομάδα μπάσκετ, δηλαδή τις αδυναμίες που θα δημιουργηθούν.

Κόστος

Παραπάνω αναλύθηκαν οι τρόποι με τους οποίους μια επαγγελματική ομάδα μπάσκετ του NBA μπορεί να εκμεταλλευτεί τα δεδομένα προς όφελός της, αλλά δεν αναλύθηκαν τα μειονεκτήματα από την εφαρμογή μεθόδων ανάλυσης δεδομένων που ίσως αποτρέπουν τις ομάδες από το να δημιουργήσουν τμήματα analytics. Η ομάδες μπάσκετ είναι επιχειρήσεις με έσοδα αλλά και έξοδα (μισθοί υπαλλήλων και αθλητών, λειτουργικά έξοδα κ.α. Η χρησιμοποίηση μεθόδων επιχειρησιακής ευφυΐας καθώς και εργαλείων data mining είναι μια επένδυση υψηλού κόστους.

Για παράδειγμα αν μία ομάδα NBA αποφασίσει να δημιουργήσει τμήματα analytics, θα χρειαστεί να επενδύσει στην απόκτηση του υλικού και λογισμικού (hardware και software) που θα χρειαστεί. Από υπολογιστές και wearables μέχρι BI software και data mining εργαλεία, η επένδυση σε εξοπλισμό χρειάζεται πόρους. Επίσης η χρησιμοποίηση BI και data mining εργαλείων προϋποθέτει επαρκώς εκπαιδευμένο προσωπικό για την δημιουργία ολόκληρων τμημάτων που θα χρειαστεί να δουλέψουν πάνω σε αυτές τις τεχνολογίες, και θα επικοινωνεί τα αποτελέσματα με τα επιτελείο της ομάδας. Είτε η ομάδα αποφασίσει να προσλάβει εξειδικευμένο προσωπικό είτε να το εκπαιδεύσει, είτε και τα δύο, το κόστος είναι μεγάλο. Το κόστος θα αυξηθεί παραπάνω καθώς η ομάδα θα πρέπει να προστατευθεί και από απειλές ασφάλειας καθώς και απειλές λόγω κακής ποιότητας δεδομένων. Οι ιδιοκτήτες θα πρέπει να κατανοήσουν το πιθανά μελλοντικά οφέλη που θα επιφέρει η αύξηση των δαπανών τις ομάδας για την δημιουργία τμήματος analytics, καθώς και αν τα διαφυγόντα κέρδη από την μη χρησιμοποίηση ανάλυσης δεδομένων υπερκαλύπτουν τα κόστη για την δημιουργία τμημάτων analytics.

(<https://www.mrc-productivity.com/>, <https://associationanalytics.com/blog/how-much-association-business-intelligence-project-cost/>)\

1.1.3 Ευκαιρίες

Εκτός από τα πλεονεκτήματα που θα αποκομίσει η ομάδα, θα παρουσιαστούν κάποιες ευκαιρίες που θα μπορούσε η ομάδα να εκμεταλλευτεί έτσι ώστε να αποκομίσει πρόσθετα οφέλη.

Εμπορικά Κέρδη

Οι επιχειρήσεις χρησιμοποιούν ανάλυση δεδομένων για να βοηθήσουν τις εμπορικές τους δραστηριότητες, έτσι και μια αθλητική ομάδα η οποία αποκτάει έσοδα και από την πώληση εμπορευμάτων (merchandise) αλλά και εισιτηρίων.

Μία μέθοδο για να βοηθήσει η ομάδα τις εμπορικές της δραστηριότητες είναι ο καθορισμός δυναμικής τιμολόγησης των εισιτηρίων, μία μέθοδο που χρησιμοποιείται κυρίως σε ομάδες baseball (MBL). Στην δυναμική τιμολόγηση αλγόριθμοι, επεξεργάζονται διάφορες κατηγορίες δεδομένων (π.χ αν η ομάδα νίκησε τον προηγούμενο αγώνα, ή τι καιρό έχει εκείνη την μέρα) και προσαρμόζει την τιμή των εισιτηρίων ανάλογα. Από έρευνες σε βάθος εικοσαετίας σε ομάδες baseball φάνηκε ότι η χρησιμοποίηση δυναμικής τιμολόγησης έδειξε αύξηση 4,2 % και 9.5 % σε κέρδη και αξία ομάδας αντίστοιχα (Courty, Davey, 2019) ενώ σε έρευνα του 2020 ανάμεσα σε managers Αμερικάνικων αθλητικών ομάδων , το 70% χρησιμοποιούν ανάλυση δεδομένων για δυναμική τιμολόγηση, αλλά το 50% δεν έχει αυτοματοποιήσει την δυναμική τιμολόγηση(Sutton, Urban, 2020) . Μέσω της δυναμικής τιμολόγησης η ομάδες μπορούν να ακολουθήσουν το παράδειγμα των ομάδων του MBL και να αποκομίσουν περισσότερα έσοδα , μέσω μεγαλύτερης προσέλευσης στα γήπεδα.

Πώληση Δεδομένων

Τα δεδομένα στον κόσμο των επιχειρήσεων είναι πολύτιμα. Πολλές επιχειρήσεις αναζητούν και είναι διατεθειμένες να επενδύσουν μεγάλα ποσά για να αποκομίσουν τα δεδομένα που χρειάζονται. Όταν μια επιχείρηση δεν μπορεί να αποκτήσει τα δεδομένα από πηγές στο εσωτερικό της συνήθως τα αποκτάει αγοράζοντάς τα από εξωτερικές πηγές (forbes.com) Οι μέθοδοι επιχειρησιακής ευφυΐας και εξόρυξης δεδομένων όπως αναλύθηκε και παραπάνω παράγει ένα τεράστιο αριθμό δεδομένων που μπορεί να χρησιμοποιήσει. Οι ομάδες θα μπορούσαν να εκμεταλλευτούν τα δεδομένα που διαθέτουν και να αποκομίσουν κέρδος. Το NBA έχει υπογράψει συμφωνία με τις εταιρείες Sportsradar και Second Spectrum αξίας 250 εκατομμυρίων δολαρίων (forbes.com). Οι εταιρείες αυτές διοχετεύουν τα δεδομένα σε στοιχηματικές εταιρείες εκτός Αμερικής, καθώς η νομοθεσία περί αθλητικού στοιχηματισμού στην Αμερική δεν τον επιτρέπει εκτός συγκεκριμένων πολιτειών (wikipedia.org). Με την άνοδο του αθλητικού στοιχηματισμού παγκοσμίως μια ενδεχόμενη αλλαγή στο νομοθετικό πλαίσιο της Αμερικής θα απογειώνει την ζήτηση σε αθλητικά δεδομένα κάτι που οι ομάδες θα μπορούσε να εκμεταλλευτεί.

1.1.4 Απειλές

Αφού παρατέθηκαν τα άμεσα οφέλη, οι ευκαιρίες που ενδέχεται να προκύψουν αλλά και τα αρνητικά που επιφέρει η υιοθέτηση μεθόδων analytics από τις ομάδες μπάσκετ, πρέπει να κατανοηθούν και οι απειλές, δηλαδή κάποια δυνητικά αρνητικά που ενδέχεται να ζημιώσουν την ομάδα.

Σκεπτικισμός

Μία απειλή που πρέπει να σκεφτούν οι ομάδες αναφορικά με την εκτεταμένη χρήση analytics από τις ομάδες είναι το κατά πόσο θα αλλάξει η φύση του αθλήματος και τι συνέπειες ενδέχεται να έχει αυτό ως προς την απήχηση στον κόσμο.

Όπως αναλύθηκε στα παραπάνω κεφάλαια, η ανάλυση δεδομένων έδειξε ότι τα σουτ τριών πόντων φαίνεται να είναι πιο αποτελεσματικός τρόπος επίθεσης σε σχέση με τα σουτ μέσης απόστασης, κάτι που έχει οδηγήσει τις ομάδες να προσανατολίζονται σε αυτή την κατεύθυνση, επενδύοντας σε αθλητές με καλό σουτ αλλά και εναρμονίζοντας τις τακτικές του σε αυτό τον τρόπο παιχνιδιού. Ο τρόπος παιχνιδιού και η επιρροή ως προς την απήχηση στον κόσμο ενδέχεται να επηρεάζει και τα έσοδα των ομάδων όπως φαίνεται και σε έρευνα (Jake Harrison, 2019) καθώς βρέθηκε ότι το συνολικό ποσοστό προσπάθειας τριών πόντων βρέθηκε να έχει αρνητικό αποτέλεσμα επί των εσόδων των ομάδων. Συγκεκριμένα σε μια προσπάθεια ποσοτικοποίησης βρέθηκε ότι μία ενδεχόμενη αύξηση του ποσοστού προσπάθειας τριών πόντων κατά 1 % θα οδηγούσε σε μείωση των εσόδων της τάξης του 0,1 %. Δηλαδή σε απώλεια περίπου 252.968 δολαρίων σε μία μέση ομάδα του NBA την σεζόν 2017-2018.

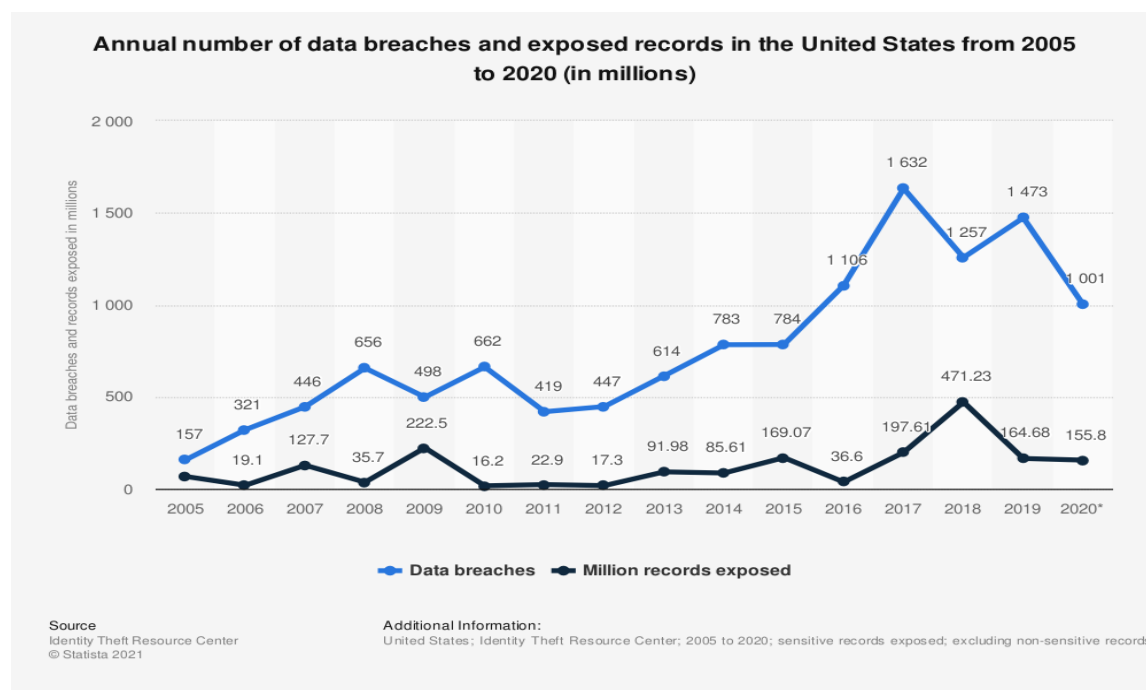
Η επιρροή που έχει η εξόρυξη και ανάλυση δεδομένων στους general managers των ομάδων, όσον αφορά τις αποφάσεις τους για την λειτουργία μίας ομάδας, έχει δημιουργήσει έντονη κριτική από εν ενεργεία αλλά και από βετεράνους αθλητές που θεωρούν ότι οι data-driven αποφάσεις στον τομέα της δημιουργίας ομάδας ή του τρόπου παιχνιδιού δεν είναι αποτελεσματικές αν δεν συνδυάζονται με τεχνογνωσία πάνω στο άθλημα. Ένα παράδειγμα είναι ο Charles Barkley, βετεράνος καταξιωμένος μπασκετμπολίστας ο οποίος έχει εκφράσει ανοιχτά την θέση του και την κριτική του κυρίως προς τον general manager των Houston Rockets Daryl Morey και την ευρεία χρήση analytics για την λήψη αποφάσεων (<https://www.si.com/>). Πολλές κριτικές υπάρχουν και για την μη-χρησιμοποίηση βασικών παικτών για λόγους ξεκούρασης, όπως αναλύθηκε και στα παραπάνω κεφάλαια λόγω των δεδομένων που αποκτούν οι ομάδες από τα wearables, και όχι λόγω κάποιου τραυματισμού. Οι οπαδοί πληρώνουν εισιτήρια για να δουν τους παίκτες αστέρες τις ομάδας και η μη-χρησιμοποίηση ενδέχεται να επηρεάσει τα έσοδα από τα εισιτήρια. Η μη χρησιμοποίηση παικτών με υψηλή δημοτικότητα ίσως επιφέρει και πρόστιμα στις ομάδες από το ίδιο το NBA ([cbssports.com](https://www.cbssports.com))

Κακή Ποιότητα Δεδομένων

Η πρόοδος της τεχνολογίας, παρέχει σε όλες τις επιχειρήσεις διάθεση περισσότερων πληροφοριών από ποτέ, αλλά αυτό είναι ταυτόχρονα και προβληματικό. Ένα πλεόνασμα δεδομένων σημαίνει ότι πολλά από αυτά που αναλύουν τα εργαλεία επιχειρησιακής ευφυΐας είναι άσχετα ή δεν βοηθούν, με αποτέλεσμα να μην διευκολύνονται οι διαδικασίες που αναφέρθηκαν στα πλεονεκτήματα. Η IBM υπολόγισε ότι το κόστος από τα χαμηλής ποιότητας δεδομένα στην Αμερική το 2016 ανήλθε στα 3.1 τρισεκατομμύρια (<https://hbr.org/>). Ο λόγος για τον οποίο τα κακά δεδομένα κοστίζουν τόσο πολύ είναι ότι οι υπεύθυνοι λήψης αποφάσεων, οι διευθυντές, αναλυτές, οι επιστήμονες δεδομένων και άλλοι τα συναντούν στην καθημερινή τους εργασία, αυτό είναι χρονοβόρο και ακριβό. Ένα άλλο πρόβλημα που προκύπτει από την πληθώρα δεδομένων και ανεβάζει την πολυπλοκότητα στο καθαρισμό και στην προετοιμασία δεδομένων είναι ότι προέρχονται από διαφορετικούς τύπους δεδομένων και διαφορετικά datasheets και οι ομάδες χρειάζεται να τα προετοιμάσουν και να υπολογίσουν στις αποφάσεις τους και τα πιθανά κόστη από δεδομένα κακής ποιότητας

Προβλήματα Ασφάλειας

Ένα από τα προβλήματα που δημιουργούνται με την χρήση οποιουδήποτε σύστημα επιχειρησιακής ευφυίας είναι ο κίνδυνος διαρροών. Κατά την χρησιμοποίηση εφαρμογών ΒΙ για τον χειρισμό δεδομένων, ένα σφάλμα ή μία κακόβουλη προσπάθεια(hacking) θα μπορούσε να το εκθέσει, βλάπτοντας την επιχείρηση, τους πελάτες ή τους υπαλλήλους σας. Ο αριθμός των data breaches αυξάνεται ολοένα και περισσότερο μαζί με την ραγδαία αύξηση της χρήσης δεδομένων από τις επιχειρήσεις. Συγκεκριμένα από το 2005 έως το 2017 ο αριθμός των data breaches αυξήθηκε κατά 900% ενώ τα δεδομένα που διέρρευσαν αυξήθηκε κατά 471% .(statista.com) Στο NBA Canada , το 2019 υπήρχε τεράστια διαρροή δεδομένων οπαδών με συνέπεια η προσωπικές πληροφορίες αρκετών ανθρώπων για παράδειγμα τα ονόματα τους και οι ηλεκτρονικές διευθύνσεις τους στα χέρια hacker(narcity.com). Οι ομάδες πρέπει να υπολογίσουν και αυτούς τους δυνητικούς κινδύνους καθώς κατέχουν στην κατοχή τους πολύ σημαντικά και πολύτιμα δεδομένα.



Εικόνα 4:Ετήσιος αριθμός κυβερνο-επιθέσεων και αρχείων που διέρρευσαν στις ΗΠΑ από το 2005 έως το 2020 (σε εκατομμύρια)(Πηγή : Statista.com)

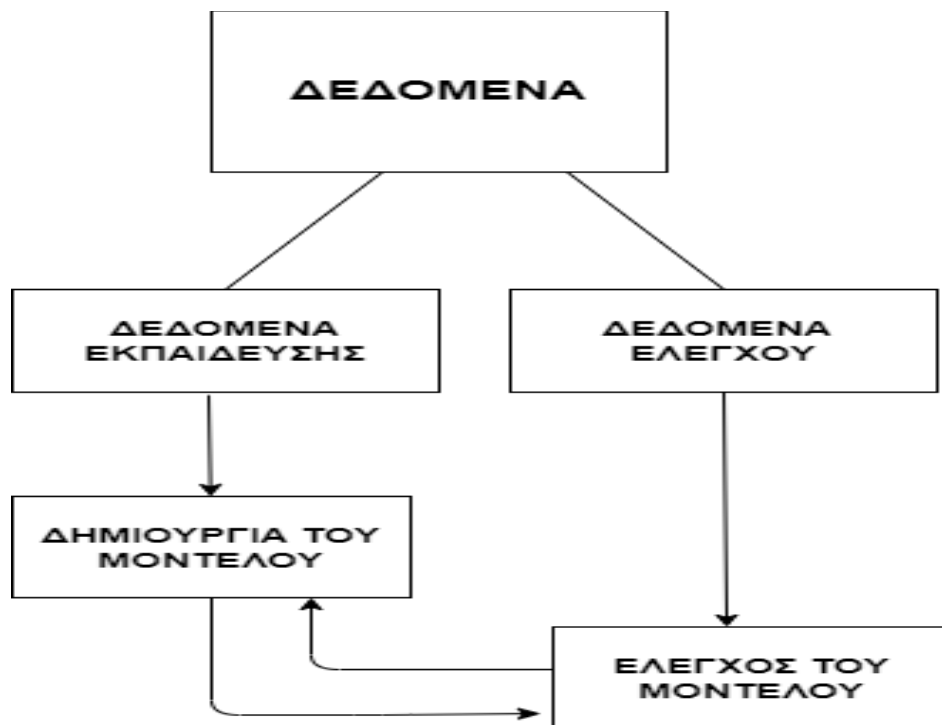
2 Machine Learning

Όπως αναφέραμε στα παραπάνω κεφάλαια, οι επιχειρήσεις και στη συγκεκριμένη περίπτωση οι ομάδες προσπαθούν να αξιοποιήσουν τα δεδομένα και να κάνουν προβλέψεις, να ανακαλύψουν συσχετίσεις και να βοηθηθούν στην διαδικασία λήψης αποφάσεων, οι ομάδες το πετυχαίνουν αυτό με την βοήθεια της μηχανικής μάθησης (machine learning).

Η μηχανική μάθηση είναι η μελέτη αλγορίθμων υπολογιστών που βελτιώνονται αυτόματα μέσα από την εμπειρία και την χρήση δεδομένων. (Tom Mitchell , 1997). Ο όρος μηχανική μάθηση χρησιμοποιήθηκε το 1959 από τον Arthur Samuel , έναν πρωτεργάτη στον τομέα των παιχνιδιών σε υπολογιστές και στην τεχνητή νοημοσύνη, ο οποίος μελέτησε το πως ένας υπολογιστής μπορεί να προγραμματιστεί έτσι ώστε να θα μάθει να παίζει, καλύτερα από τον χρήστη, το παιχνίδι ντάμες. Κατά τον Arthur Samuel, η μηχανική μάθηση είναι ο κλάδος ο οποίος δίνει στους υπολογιστές την δυνατότητα να μαθαίνουν χωρίς να έχουν προγραμματιστεί γι' αυτό, δηλαδή χωρίς να έχουν οριστεί κανόνες γραμμένοι από τους ανθρώπους. Η μηχανική μάθηση θεωρείται ως κομμάτι της τεχνητής νοημοσύνης , καθώς ως επιστημονική προσπάθεια, η μηχανική μάθηση προήλθε από την έρευνα και την μελέτη της τεχνητής νοημοσύνης. Τα τελευταία χρόνια η έμφαση της μελέτης της μηχανικής μάθησης μετατοπίστηκε από τις συμβολικές προσεγγίσεις που είχε υιοθετήσει από την τεχνητή νοημοσύνη, και προσανατολίστηκε στις μεθόδους και τα μοντέλα που είναι επηρεασμένα από την επιστήμη της στατιστικής και της θεωρίας πιθανοτήτων. (Langley P, 2011)

Η λογική της μηχανικής μάθησης είναι ότι ο υπολογιστής μαθαίνει από τα ιστορικά δεδομένα έτσι ώστε να παράξει πρόβλεψη για το μέλλον. Η μηχανική μάθηση είναι συνδεδεμένη με την επιστήμη της στατιστικής καθώς αφορά την δημιουργία προβλέψεων μέσω υπολογισμών. Η μηχανική μάθηση χρησιμοποιείται για να βοηθήσει στην ανάλυση πολύ μεγάλων ποσοτήτων δεδομένων, κάτι που τα τελευταία χρόνια με την ραγδαία αύξηση στην εκμετάλλευση των δεδομένων, κάνει την χρησιμοποίηση μεθόδων μηχανικής μάθησης, σημαντικό εργαλείο για τις επιχειρήσεις.

Η μηχανική μάθηση πετυχαίνει δύο στόχους, ο ένας είναι να ταξινομεί δεδομένα με βάση τα μοντέλα που έχουν αναπτυχθεί, ο άλλος σκοπός είναι να κάνει προβλέψεις για μελλοντικά αποτελέσματα με βάση αυτά τα μοντέλα. Ως προς τον τρόπο λειτουργίας, οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν δεδομένα εκπαίδευσης (training data) ως είσοδο με βάση τα οποία ο υπολογιστής μαθαίνει και δημιουργεί τα κατάλληλα μοντέλα και στη συνέχεια εφαρμόζει τα μοντέλα στα δεδομένα ελέγχου (test data) για να αξιολογηθεί το αποτέλεσμα, και ανάλογα ο χρήστης να το χρησιμοποιήσει ή να προσαρμόσει την μέθοδό του ανάλογα. Οι αλγόριθμοι μηχανικής μάθησης αποδίδουν καλύτερα όσο ο όγκος των δεδομένων που παρέχεται σε αυτούς μεγαλώνει.



Εικόνα 5: Machine Learning Flowchart

Η μηχανική μάθηση χρησιμοποιείται για να βοηθήσει τους ανθρώπους και τις επιχειρήσεις να αποκομίσουν αξία από τα δεδομένα. Η αναγνώριση εικόνων είναι ένα τέτοιο παράδειγμα. Ο αλγόριθμος αναγνωρίζει εικόνες βάση της σύστασής τους από πίξελ. Η πρακτική αυτή μπορεί να βοηθήσει ακόμα και σε τομείς όπως η ιατρική (αναγνώριση ανωμαλιών, μέσα από ακτινογραφίες). Όπως αναφέρθηκε παραπάνω οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται για την δημιουργία προβλέψεων, κάτι που έχει πρακτική αξία σε όλους τους τομείς των επιχειρήσεων, όπως οι επενδύσεις, το μάρκετινγκ, η ανάπτυξη ενός καινούριου προϊόντος. Το μόνο που χρειάζεται είναι δεδομένα.

2.1 Είδη Μηχανικής Μάθησης

Επιτηρούμενη (Supervised)

Επιτηρούμενη μηχανική μάθηση είναι η υποκατηγορία της μηχανικής μάθησης στην οποία τα δεδομένα που χρησιμοποιούνται για την εκπαίδευση του μοντέλου, έχουν ετικέτα, δηλαδή υπάρχει και είναι γνωστή η μεταβλητή στόχου. Στην επιτηρούμενη μηχανική μάθηση κάθε εγγραφή είναι ένα ζεύγος αποτελούμενο από την είσοδο (ανεξάρτητες μεταβλητές) και την είσοδο (εξαρτημένες μεταβλητές). Καθώς τα δεδομένα έχουν ετικέτα το μοντέλο μαθαίνει από τα ιστορικά δεδομένα και προσαρμόζεται αναλόγως μέχρι να ελαχιστοποιηθεί το σφάλμα.

Η επιτηρούμενη μηχανική μάθηση χρησιμοποιείται για ταξινόμηση και παλινδρόμηση.

Ταξινόμηση (Classification) : Στην ταξινόμηση ο αλγόριθμος αντιστοιχεί τα δεδομένα σε συγκεκριμένες κατηγορίες βάσει κοινών χαρακτηριστικών που έχουν.

Παλινδρόμηση (Regression) : Η παλινδρόμηση χρησιμοποιείται για την εύρεση σχέσης μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών και χρησιμοποιείται για την πραγματοποίηση προβλέψεων.

Μη – Επιτηρούμενη (Unsupervised)

Η μη – επιτηρούμενη μηχανική μάθηση προσπαθεί να βρει μοτίβα σε δεδομένα χωρίς ετικέτα, δηλαδή δεν υπάρχει εξαρτημένη μεταβλητή, και βάση αυτά τα μοτίβα να αναλύσει και να συσταδοποιήσει τα δεδομένα.

Η μη-επιτηρούμενη μηχανική μάθηση χρησιμοποιείται για συσταδοποίηση , συσχέτιση και μείωση διαστάσεων

Συσταδοποίηση (Clustering) : Οι μέθοδοι συσταδοποίησης προσπαθούν να ομαδοποιήσουν τα δεδομένα χωρίς ετικέτα με βάση τις ομοιότητες ή τις διαφορές τους , βρίσκοντας δηλαδή μοτίβα ή πληροφορίες στα δεδομένα.

Κανόνες Συσχέτισης (Association Rule) : Οι κανόνες συσχέτισης προσπαθούν να βρουν συσχέτιση μεταξύ των μεταβλητών ενός συνόλου δεδομένων.

Μείωση Διαστάσεων (Dimensionality Reduction) : Η μείωση διαστάσεων είναι μια τεχνική που χρησιμοποιείται όταν ο αριθμός των διαστάσεων σε ένα σύνολο δεδομένων είναι πολύ υψηλός, καθώς ο πολύ υψηλός αριθμός διαστάσεων σε ένα σύνολο δεδομένων μπορεί να επηρεάσει την απόδοση των αλγορίθμων (πχ overfitting) και η ανάλυση των δεδομένων είναι συνήθως υπολογιστικά δυσεπίλυτη.

Ενισχυμένη (Reinforcement learning)

Στην ενισχυμένη μηχανική μάθηση ο υπολογιστής προσπαθεί να βρει την καλύτερη δυνατή διαδρομή που πρέπει να ακολουθήσει σε μια συγκεκριμένη κατάσταση με βάση τον ορισμό κατάλληλων μέτρων για τη μεγιστοποίηση της ανταμοιβής στη συγκεκριμένη κατάσταση. Λόγω της απουσίας δεδομένων εκπαίδευσης, ο υπολογιστής περιορίζεται στον να μαθαίνει από την εμπειρία του όπου εκπαιδεύεται συνεχώς με μέθοδο δοκιμής και σφάλματος.

2.2 Αλγόριθμοι Μηχανικής Μάθησης

Στο κεφάλαιο αυτό θα αναφερθούν οι σημαντικότεροι αλγόριθμοι μηχανικής μάθησης

Γραμμική Παλινδρόμηση (Linear Regression) : Η γραμμική παλινδρόμηση είναι η μέθοδος που βρίσκει την γραμμική σχέση μεταξύ της εξαρτημένης μεταβλητής και της ανεξάρτητης (απλή γραμμική παλινδρόμηση) ή των ανεξάρτητων (πολλαπλή γραμμική παλινδρόμηση) μεταβλητών, προσπαθώντας να

δημιουργήσει μια ευθεία γραμμή που εμφανίζει αυτή την σχέση χρησιμοποιώντας την μέθοδο των ελαχίστων τετραγώνων. Χρησιμοποιείται για την δημιουργία προβλέψεων.

Naive Baye: Ο αλγόριθμος Naive Bayes είναι ένας εποπτευόμενος αλγόριθμος ομαδοποίησης, και βασίζεται στο θεώρημα Bayes, ακολουθώντας την αρχή ότι η παρουσία ενός χαρακτηριστικού δεν επηρεάζει την παρουσία ενός άλλου στην πιθανότητα ενός δεδομένου αποτελέσματος και κάθε προγνωστικός παράγοντας έχει ίση επίδραση σε αυτό το αποτέλεσμα. Υπάρχουν τρία είδη Naïve Bayes classifiers ανάλογα με τα δεδομένα που χρησιμοποιούνται. Όταν οι μεταβλητές είναι δυαδικού χαρακτήρα (NAI ή OXI) χρησιμοποιούμε Bernoulli Naïve Bayes ενώ κατάλληλος για ταξινόμηση σε διακριτές μεταβλητές είναι ο Multinomial Naive Bayes αλγόριθμος και τέλος σε περίπτωση μεταβλητών με συνεχείς τιμές, κατάλληλος είναι ο Gaussian Naive Bayes αλγόριθμος.

K-Πλησιέστερος Γείτονας (KNN) : Ο αλγόριθμος KNN χρησιμοποιείται τόσο για κατηγοριοποίηση όσο και για παλινδρόμηση, χρησιμοποιώντας την απόσταση (συνήθως την ευκλείδεια απόσταση) μεταξύ των δεδομένων και υιοθετώντας την παραδοχή ότι τα παρόμοια δεδομένα βρίσκονται κοντά το ένα στο άλλο. Σημαντικό στον KNN αλγόριθμο είναι ο προσδιορισμός του K. Γενικά μια μικρή τιμή του K θα αυξήσει την επίδραση του θορύβου καθώς και οδηγήσει σε πρόβλημα υπερπροσαρμογής (overfitting) ενώ μια μεγάλη τιμή του K μπορεί να οδηγήσει σε υποπροσαρμογή (underfitting) και τον καθιστά υπολογιστικά ακριβό.

Λογιστική Παλινδρόμηση (Logistic Regression) : Είναι μια μέθοδος μηχανικής μάθησης η οποία χρησιμοποιείται για να λύσει προβλήματα δυαδικής ταξινόμησης, δηλαδή όταν η εξαρτημένη μεταβλητή είναι κατηγορική. Η λογιστική παλινδρόμηση προσπαθεί να προβλέψει την πιθανότητα να συμβεί ένα από τα δύο ενδεχόμενα, με την προσαρμογή δεδομένων σε μια συνάρτηση logit.

K-Means Clustering : Είναι ένας τύπος αλγορίθμου χωρίς επίβλεψη που εφαρμόζεται σε προβλήματα συσταδοποίησης (clustering). Ο αλγόριθμος τοποθετεί τα δεδομένα παρατηρήσεις σε k συμπλέγματα, επιλέγοντας ένας K αριθμό αρχικών σημείων για κάθε συστάδα (κεντροειδή) . Κάθε σημείο δεδομένων σχηματίζει ένα σύμπλεγμα με βάση την απόσταση από τα κεντροειδή. Στη συνέχεια ο αλγόριθμος βρίσκει τα νέα κεντροειδή με βάση τα δεδομένα σε κάθε συστάδα και η διαδικασία επαναλαμβάνετε μέχρι τα κεντροειδή να μην αλλάξουν. Ο καθορισμός του K γίνεται συνήθως με την μέθοδο του αγκώνα (elbow method) και την μέθοδο της σιλουέτας (Silhouette method)

Δέντρα Αποφάσεων (Decision Trees) : Τα δέντρα αποφάσεων είναι ένας αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται είτε για κατηγοριοποίηση είτε για παλινδρόμηση. Τα δέντρα αποφάσεων προσπαθούν μέσω μια δομή που μοιάζει με κλαδιά ενός δέντρου στην οποία κάθε εσωτερικός κόμβος κάθε εσωτερικός κόμβος αντιπροσωπεύει μια δοκιμή σε ένα χαρακτηριστικό των δεδομένων, και ανάλογα το αποτέλεσμα της δοκιμής αυτής δημιουργούνται τα κλαδιά και τέλος στον τερματικό κόμβο αποτυπώνεται το αποτέλεσμα του αλγορίθμου. Πάνω στα δέντρα αποφάσεων βασίζονται και οι αλγόριθμοι **random forest** και **gradient boosting**.

Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines): Ο αλγόριθμος Support Vector Machine είναι αλγόριθμος εποπτευόμενης μηχανικής εκμάθησης που χρησιμοποιείται κυρίως για ταξινόμηση (classification). Στον Support Vector κάθε στοιχείο δεδομένων αναπαρίσταται ως ένα σημείο σε χώρο n-διαστάσεων . Ο αριθμός n καθορίζεται από τον αριθμό των μεταβλητών και η τιμή κάθε μεταβλητής είναι η τιμή μιας συγκεκριμένης συντεταγμένης. Στη συνέχεια ο αλγόριθμος βρίσκει η

την γραμμή που χωρίζει τα δεδομένα σε δύο διαφορετικά ταξινομημένες ομάδες και τα ταξινομεί ανάλογα με το πού βρίσκονται τα δεδομένα δοκιμής δηλαδή σε μια πλευρά της γραμμής.

3. Εφαρμογή Μηχανικής Μάθησης στο NBA

Στο προηγούμενα κεφάλαια είδαμε τρόπους που οι ομάδες χρησιμοποιούν ανάλυση δεδομένων και μηχανική μάθηση για να βοηθήσουν στη λήψη συμπερασμάτων, καθώς και αναφέρθηκαν περιληπτικά ορισμένες πληροφορίες για την μηχανική μάθηση. Στο κεφάλαιο αυτό θα γίνει η προσπάθεια εφαρμογής στην πράξη όσων αναλύθηκαν παραπάνω, μέσω της εφαρμογής μεθόδων μηχανικής μάθησης σε δεδομένα που σχετίζονται με ομάδες του NBA.

3.1 Μεθοδολογία

Οι ομάδες όπως αναφέρθηκε και παραπάνω, κάθε χρόνο επιλέγουν αθλητές μέσω του draft με σκοπό να ενισχύσουν την ομάδα. Οι αθλητές αυτοί προέρχονται είτε από αμερικάνικα κολλέγια είτε από πρωταθλήματα σε όλον τον κόσμο. Τα συμβόλαια που υπογράφουν οι rookies είναι συνήθως διάρκειας δύο χρόνων με επιλογή ανανέωσης από τις ομάδες για τον τρίτο και τέταρτο χρόνο. (cbabreakdown.com) Συμπεραίνετε οπότε ότι μετά από ένα διάστημα δύο χρόνων τα επιτελεία πρέπει να πάρουν τις αποφάσεις για το αν αξίζει να επενδύσουν στον αθλητή και να προσχωρήσουν στην ανανέωση του συμβολαίου.

Πέρα από την αγωνιστική αξιολόγηση, οι ομάδες μπορούν να χρησιμοποιήσουν ιστορικά δεδομένα για να βοηθήσουν την απόφαση αυτή. Στην εργασία θα γίνει προσπάθεια να προβλεφθεί η εξέλιξη του αθλητή βάσει των στατιστικών του, τις δύο πρώτες rookie seasons. Για να γίνει αυτό θα χρησιμοποιήσουμε την μεταβλητή EFF (wikipedia.org) και θα γίνει προσπάθεια να βρεθεί ένα μοντέλο πρόβλεψης της μεταβλητής αυτής. Η μεταβλητή EFF (Individual Player Efficiency) είναι μια μεταβλητή η οποία μετράει την αποτελεσματικότητα ενός αθλητή του NBA και υπολογίζεται από τον τύπο :

$$EFF = (PTS + REB + ASS + STL + BLK - MISSED FG - MISSED FT - TO) / GP$$

Θα χρησιμοποιηθεί δηλαδή οι συνολικοί πόντοι καριέρας, τα συνολικά ριμπάουντ, οι συνολικές τελικές πάσες, τα συνολικά κλεψίματα, οι συνολικές τάπες και από τα άθροισμά τους θα αφαιρεθούν τα χαμένα σουτ εντός πεδιάς, οι χαμένες βολές καθώς και τα λάθη. Τέλος το αποτέλεσμα που θα προκύψει θα διαιρεθεί με τα συνολικά παιχνίδια που έχει αγωνιστεί ο παίκτης στο NBA. Ενδεικτικά αναφέρεται ότι αθλητές όπως Michael Jordan , LeBron James, Magic Johnson, Kareem Abdul Jabar και Larry Bird έχουν EFF περίπου 30 μονάδες, ενώ οι DeMar DeRozan και Andre Iguodala κυμαίνονται περίπου στις 15 μονάδες, ενώ οι Αντώνης Φώτσης και Κώστας Παπανικολάου έχουν περίπου 5 μονάδες. Τέλος ο Giannis Antetokounmpo έχει 19 μονάδες σαν μέσο όρο καριέρας, κάτι που αναμένεται να αυξηθεί όσο αγωνίζεται περισσότερα χρόνια. Φαίνεται ότι η μεταβλητή EFF μπορεί να χρησιμοποιηθεί ικανοποιητικά σαν μεταβλητή αξιολόγησης ενός αθλητή.

Για να δημιουργήσουμε το μοντέλο πρόβλεψης της μεταβλητής EFF θα χρησιμοποιήσουμε μοντέλα μηχανικής μάθησης.

Γραμμική Παλινδρόμηση

Για να γίνει προσπάθεια εύρεσης της γραμμικής σχέσης μεταξύ των στατιστικών ενός αθλητή στα δύο πρώτα χρόνια της καριέρας του και στο συνολικό career EFF θα χρησιμοποιηθεί η πολλαπλή γραμμική παλινδρόμηση.

Όπως αναφέραμε και παραπάνω η γραμμική παλινδρόμηση είναι μια τεχνική επιτηρούμενης μηχανικής μάθησης που χρησιμοποιείται για την πρόβλεψη της μια εξαρτημένης μεταβλητής με την προσπάθεια εύρεσης γραμμική σχέσης με μία ή περισσότερες ανεξάρτητες μεταβλητές. Στη συγκεκριμένη περίπτωση θα χρησιμοποιηθεί η πολλαπλή παλινδρόμηση. Από την χρήση της πολλαπλής παλινδρόμησης μπορούμε να εξάγουμε συμπεράσματα όπως το να καταλάβουμε την επίδραση που έχουν οι ανεξάρτητες μεταβλητές σε μια εξαρτημένη μεταβλητή, όπως επίσης και το να δούμε πόσο θα αλλάξει η εξαρτημένη μεταβλητή σε μια ενδεχόμενη αλλαγή μιας ανεξάρτητης μεταβλητής, εφόσον όλες οι υπόλοιπες παραμείνουν σταθερές. Τέλος η πολλαπλή γραμμική παλινδρόμηση χρησιμοποιείται για την δημιουργία προβλέψεων.

Το μοντέλο πολλαπλής γραμμικής παλινδρόμησης για να λειτουργήσει σωστά προϋποθέτει ότι η σχέση μεταξύ ανεξάρτητων μεταβλητών και εξαρτημένων μεταβλητών είναι γραμμική. Αυτή η σχέση μοντελοποιείται με τη μορφή :

$$y_i = b_0 + b_1x_{i1} + b_2x_{i2} + \dots + b_nx_{in} + \varepsilon.$$

Το y_i είναι η εξαρτημένη μεταβλητή, τα x_{i1}, x_{i2}, x_{in} είναι οι ανεξάρτητες μεταβλητές, το b_0 είναι το σημείο τομής (intercept), τα b_1, b_2, b_n είναι οι συντελεστές κάθε επεξηγηματικής μεταβλητής και τέλος το ε είναι το σφάλμα του μοντέλου.

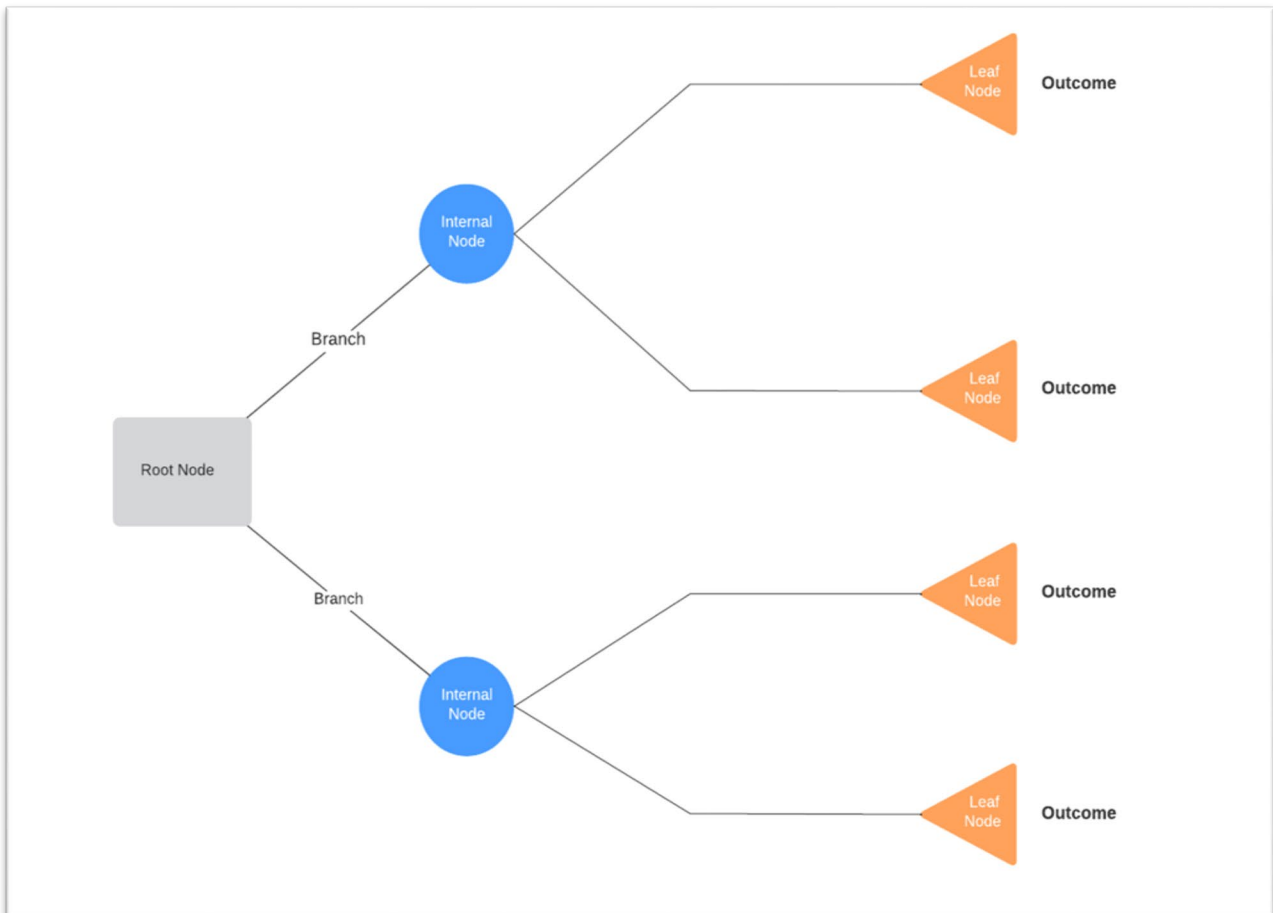
Για την εύρεση του μοντέλου πολλαπλής γραμμικής παλινδρόμησης πραγματοποιείται προσδιορισμός των συντελεστών παλινδρόμησης έτσι ώστε το σφάλμα να ελαχιστοποιείται. Για να ελαχιστοποιηθεί το σφάλμα ο αλγόριθμος βρίσκει τους συντελεστές που πετυχαίνουν την μέγιστη ελαχιστοποίηση του αθροίσματος των τετραγώνων των διαφορών μεταξύ των πραγματικών παρατηρήσεων της εξαρτημένης μεταβλητής και εκείνων που προβλέπονται από το γραμμικό μοντέλο.

Δέντρα Αποφάσεων

Σε συνέχεια της προσπάθειας εύρεσης του μοντέλου πρόβλεψης της μεταβλητής EFF, θα πραγματοποιηθεί και η εφαρμογή της μεθόδου των δέντρων αποφάσεων, για την εύρεση του μοντέλου πρόβλεψης.

Τα δέντρα αποφάσεων (Nagesh Singh Chauhan, 2020) (J.R. Quinlan, 1987) (decision trees) είναι μία μέθοδο επιτηρούμενης (supervised) μηχανικής μάθησης, που χρησιμοποιείται είτε για παλινδρόμηση (decision tree regressor) είτε για ταξινόμηση (decision tree classifier). Τα δέντρα αποφάσεων χρησιμοποιούνται για να γίνει η πρόβλεψη της τιμής μιας μεταβλητής στόχου με βάση πολλές μεταβλητές εισόδου. Στα δέντρα αποφάσεων το σύνολο των δεδομένων χωρίζεται σε υποσύνολα με βάση ένα κριτήριο μίας μεταβλητής εισόδου, στη συνέχεια πραγματοποιείται διαχωρισμός των υποσυνόλων βάση κάποιου άλλου χαρακτηριστικού σε άλλα υποσύνολα κοκ. Ο διαχωρισμός σταματάει όταν τα δεδομένα στα τελευταία υποσύνολα που έχουν δημιουργηθεί έχουν όλα τις ίδιες τιμές μεταβλητής στόχου ή όταν ο διαχωρισμός δεν προσθέτει πλέον αξία στις προβλέψεις. Κάθε σημείο του δέντρου στο οποίο γίνεται αυτός ο διαχωρισμός του

υποσυνόλου, ονομάζεται κόμβος. Ο πρώτος κόμβος ονομάζεται κόμβος ρίζας. Κάθε κόμβος δημιουργεί κλαδιά που καταλήγουν σε πιθανά αποτελέσματα. Κάθε ένα από αυτά τα αποτελέσματα οδηγεί σε πρόσθετους κόμβους. Οι τελευταίοι κόμβοι που καταλήγουν στο αποτέλεσμα (class) , ονομάζονται τερματικοί κόμβοι ή φύλλα (leaf nodes). Ο τρόπος με τον οποίο τα δέντρα αποφάσεων χωρίζουν τα δεδομένα, εξαρτάται από των αλγόριθμο που χρησιμοποιείται. Οι πιο συνηθισμένοι είναι : ID3 , C4.5, CART, CHAID ,MARS.



Εικόνα 6: Desision Tree

3.2 Δεδομένα

3.2.1 Συλλογή και επεξεργασία Δεδομένων

Όπως αναφέραμε σκοπός της εργασίας είναι να παρουσιαστεί η εφαρμογή μεθόδων μηχανικής μάθησης στο NBA για την πραγματοποίηση πρόβλεψης της καριέρας ενός παίκτη από τα στατιστικά των δύο πρώτων σεζόν. . Τα δεδομένα προέρχονται από το keggel και είναι ένα σύνολο δεδομένων όλων των στατιστικών για κάθε παίκτη από τη σεζόν 1950 μέχρι την σεζόν 2018. Η πηγή των δεδομένων είναι το site www.basketball-reference.com

Τα δεδομένα που αντλήθηκαν αποτελούνται από ένα σύνολο δεδομένων 24.691 γραμμών και 53 στηλών, και αποτελούν τα στατιστικά για κάθε παίκτη για κάθε σεζόν από το 1950 μέχρι το 2018. Για να μπορέσουμε να αναλύσουμε τα δεδομένα όπως θέλουμε πρέπει να γίνει επεξεργασία για να έρθουν στην μορφή που θέλουμε.

Για να μπορέσουμε να δημιουργήσουμε το μοντέλο πρόβλεψης θα πρέπει πρώτα να υπολογίσουμε την εξαρτημένη μεταβλητή EFF. Για να γίνει αυτό θα χρησιμοποιήσουμε τα αρχικά δεδομένα και θα από αυτά θα υπολογίσουμε τα all-career stats. Οπότε από το αρχικό σύνολο δεδομένων θα δημιουργήσουμε ένα σύνολο δεδομένων 3919 γραμμών και 10 στηλών που θα περιέχει τα στατιστικά καριέρας που χρειαζόμαστε για τον υπολογισμό της EFF. Το συγκεκριμένο σύνολο δεδομένων δημιουργήθηκε αθροίζοντας τα στατιστικά κάθε σεζόν για κάθε παίκτη (GP, PTS, REB, ASS, STL, BLK, FG, FGA, FT, FTA, TO) Αφού μαζέψαμε τα συνολικά στατιστικά καριέρας για κάθε παίκτη χρησιμοποιούμε τον τύπο της EFF και υπολογίζουμε την μεταβλητή EFF. Από το συγκεκριμένο σύνολο δεδομένων (σύνολο δεδομένων1) κρατάμε μόνο την μεταβλητή Player και EFF.

Player	EFF
Michael Jordan	29.19
Magic Johnson	29.01
Hakeem Olajuwon	27.17
Kevin Durant	26.63

Πίνακας 1: Σύνολο δεδομένων 1

Στη συνέχεια θα χρησιμοποιήσουμε τα δεδομένα των δύο πρώτων σεζόν για να βρούμε το μοντέλο πρόβλεψης. Από το αρχικό σύνολο δεδομένων θα χρησιμοποιήσουμε τους μέσους όρους των δύο πρώτων σεζόν. Αν ένας παίκτης έχει αγωνιστεί σε μία σεζόν θα χρησιμοποιηθούν τα στατιστικά της πρώτης σεζόν. Από το αρχικό σύνολο δεδομένων φτιάχνουμε ένα σύνολο 3919 γραμμών και 49 στηλών. (σύνολο δεδομένων 2). Συμπληρωματικά θα κρατήσουμε από το αρχικό σύνολο δεδομένων την ομάδα στην οποία αγωνίστηκε πρώτα ο παίκτης, την ηλικία στην οποία αγωνίστηκε πρώτη φορά καθώς και την θέση του. (σύνολο δεδομένων 3).

Τέλος ενώνουμε τα τρία σύνολο δεδομένων με κοινή μεταβλητή την μεταβλητή Player, και δημιουργούμε το τελικό σύνολο δεδομένων 3919 σειρών και 53 γραμμών. Στο παρακάτω πίνακα βλέπουμε την επεξήγηση των μεταβλητών του συνόλου δεδομένων. Η μεταβλητή EFF δείχνει το συνολικό career EFF ενώ οι υπόλοιπες μετρικές είναι υπολογισμένες σε μέσους όρους των δύο πρώτων σεζόν. Αυτό έγινε καθώς οι μετρικές είναι και σε αριθμούς (PTS, REB, ASS) αλλά και σε ποσοστά (3P%). Για παράδειγμα ένας παίκτης με 115 PTS σημαίνει ότι κατά μέσο όρο τις δύο πρώτες σεζόν πέτυχε 115 πόντους ανά σεζόν, ενώ ένας παίκτης με 40% 3P% σημαίνει ότι μέσο όρο ανά σεζόν σκόραρε σουτ τριών πόντων με ποσοστό 40%. Στον παρακάτω πίνακα παρουσιάζονται συνοπτικές επεξηγήσεις των μεταβλητών του συνόλου δεδομένων,

Μεταβλητή	Περιγραφή
Player	Το όνομα του παίκτη
Pos	Η θέση του παίκτη την πρώτη σεζόν στο NBA. Για λόγους απλοποίησης χρησιμοποιήθηκε μόνο μία θέση και όχι συνδυαστικές θέσεις(Για παράδειγμα G-F)
Age	Η ηλικία που αγωνίστηκε ο παίκτης πρώτη φορά στο NBA
G	Τα παιχνίδια που αγωνίστηκε ο παίκτης ανά χρόνο κατά μέσο όρο τα δύο πρώτα χρόνια
MP	Τα λεπτά που αγωνίστηκε ο παίκτης ανά χρόνο κατά μέσο όρο
PER	Player Efficiency Rating . Μία εναλλακτική σύνθετη μετρική της αποδοτικότητας του παίκτη.
TS%	True Shooting Percentage (μία μετρική που υπολογίζει την αποτελεσματικότητα με την οποία ένας παίκτης σουτάρει την μπάλα
3PAr	Ποσοστό προσπαθειών τριών πόντων προς συνολικές προσπάθειες εντός πεδιάς
Ftr	Αριθμός FT ανά FG
ORB%	Το ποσοστό των διαθέσιμων επιθετικών ριμπάουντ που άρπαξε ένας παίκτης ενώ ήταν στο παρκέ.
DRB%	Το ποσοστό των διαθέσιμων αμυντικών ριμπάουντ που άρπαξε ένας παίκτης ενώ ήταν στο παρκέ.
TPB %	Το ποσοστό των διαθέσιμων συνολικών ριμπάουντ που άρπαξε ένας παίκτης ενώ ήταν στο παρκέ.
AST%	Εκτίμηση του ποσοστού των τελικών πασών που έκανε παίκτης ενώ ήταν στο παρκέ.
STL%	Εκτίμηση του ποσοστού των κατοχών του αντιπάλου που τελειώνουν με κλέψιμο από τον παίκτη ενώ ήταν στο παρκέ
BLK%	Το ποσοστό των προσπαθειών για σουτ δύο πόντων του αντιπάλου που αποκρούστηκαν με τάπα από τον παίκτη ενώ ήταν στο παρκέ
TOV%	Λάθη που πραγματοποιήθηκαν ανά 100 κατοχές
USG%	Το ποσοστό των κατοχών που πραγματοποίησε ένας παίκτης ενώ ήταν στο παρκέ.
OWS	Είναι μία σύνθετη μετρική που δηλώνει την επιθετική συνεισφορά ενός παίκτη στις νίκες τις ομάδας

DWS	Είναι μία σύνθετη μετρική που δηλώνει την αμυντική συνεισφορά ένας παίκτη ή μιας ομάδας
WS	Είναι μία σύνθετη μετρική που δηλώνει την συνεισφορά ενός παίκτη στις νίκες της ομάδας
WS/48	Η συνεισφορά στις νίκες της ομάδας ανά 48 λεπτά.
OBPM	Μία μετρική που βασίζεται στο το σκορ που εκτιμά την επιθετική συνεισφορά ενός μπασκετμπολίστα στην ομάδα όταν αυτός ο παίκτης βρίσκεται στο γήπεδο.
DBPM	Μία μετρική που βασίζεται στο το σκορ που εκτιμά την αμυντική συνεισφορά ενός μπασκετμπολίστα στην ομάδα όταν αυτός ο παίκτης βρίσκεται στο γήπεδο.
BPM	Μία μετρική που βασίζεται στο το σκορ και εκτιμά την συνεισφορά ενός μπασκετμπολίστα στην ομάδα όταν αυτός ο παίκτης βρίσκεται στο γήπεδο.
VORP	Η VORP μετατρέπει το ποσοστό BPM σε μια εκτίμηση της συνολικής συνεισφοράς κάθε παίκτη στην ομάδα, μετρημένη σε σχέση με το τι προσφέρει στην ομάδα έναντι των συμπαίκτών του.
FG	Τα καλάθια εντός πεδιάς που πέτυχε.
FGA	Τα σουτ εντός πεδιάς που πραγματοποίησε.
FG%	Το ποσοστό ευστοχίας σε σουτ εντός πεδιάς.
3P	Τα εύστοχα σουτ τριών πόντων.
3PA	Τα συνολικά σουτ τριών πόντων που επιχείρησε.
3P%	Το ποσοστό επιτυχημένων σουτ τριών πόντων.
2P	Τα εύστοχα σουτ δύο πόντων.
2PA	Τα συνολικά σουτ δύο πόντων που επιχείρησε.
2P%	Το ποσοστό επιτυχημένων σουτ δύο πόντων.
eFG%	Effective Field Goal Percentage: $(2P + 1.5 * 3P) / FGA$
FT	Οι ελεύθερες βολές που σκόραρε ο παίκτης
FTA	Οι συνολικές ελεύθερες βολές που επιχείρησε ο παίκτης
FT%	Το ποσοστό των επιτυχημένων ελεύθερων βολών.

ORB	Τα επιθετικά ριμπάουντ που μάζεψε ένας παίκτης.
DRB	Τα αμυντικά ριμπάουντ που μάζεψε ένας παίκτης.
TRB	Τα συνολικά ριμπάουντ που μάζεψε ένας παίκτης.
ASS	Ο αριθμός των τελικών πασών που έδωσε ένας παίκτης.
STL	Τα κλεψίματα που πέτυχε ένας παίκτης
BLK	Ο αριθμός των block (τάπες) που πέτυχε ένας παίκτης
TOV	Ο αριθμός των λαθών στα οποία υπέπεσε ένας παίκτης.
PF	Τα φάουλ στα οποία υπέπεσε ένας παίκτης.
PTS	Οι πόντοι που πέτυχε ένας παίκτης.
EFF	Η μεταβλητή με την οποία θα μετρήσουμε την ποιότητα του αθλητή.

Πίνακας 2: Επεξήγηση Δεδομένων

3.2.2 Προετοιμασία και Καθαρισμός Δεδομένων

Όπως αναφέραμε και στα προηγούμενα κεφάλαια, ένας κίνδυνος για τις ομάδες που μπορεί να προκύψει είναι η κακή ποιότητα δεδομένων που δυσκολεύει την περαιτέρω ανάλυση. Συμπεραίνετε λοιπόν η ανάγκη να προετοιμάσουμε τα δεδομένα μας καθώς και αναζητήσουμε τυχόν προβλήματα με την ποιότητα των δεδομένων έτσι ώστε να τα φέρουμε στην κατάλληλη μορφή για να προχωρήσει η ανάλυση.

Επιλογή Χρονολογίας

Για να βρούμε το μοντέλο πρόβλεψης της μεταβλητής EFF θα χρησιμοποιήσουμε τα δεδομένα των στατιστικών από τις δύο πρώτες season του παίκτη. Στο αρχικό σύνολο δεδομένων πάρα πολλά στατιστικά έχουν null τιμές από το 1950 έως το 1980 καθώς δεν υπήρχαν πολλά στατιστικά αναφορικά με τα σουτ τριών πόντων, καθώς πρώτη φορά η γραμμή τριών πόντων εφαρμόστηκε την σεζόν 1979-1980 (wikipedia.org) Επίσης θα διαγραφούν οι παίκτες που αγωνίστηκαν στο NBA για πρώτη φορά μετά το 2010 καθώς βρίσκονται στα πρώτα στάδια της καριέρας οπότε η χρησιμοποίηση της στατιστικής συνολικής EFF για μια καριέρα που βρίσκεται ακόμα στο αρχικό της στάδιο δεν είναι ενδεδειγμένη.

Διπλότυπες Εγγραφές

Αναφορικά με τον καθαρισμό δεδομένων αρχικά έγινα έλεγχος για διπλότυπες εγγραφές που όμως δεν υπήρχαν καθώς στον μετασχηματισμό των δεδομένων για να δημιουργηθούν τα στατιστικά σε μέσους όρους των δύο πρώτων σεζόν διαγράφηκαν όλες οι διπλές εγγραφές.

Null Τιμές

Στον έλεγχο για null εγγραφές δεν βρέθηκε καμία κενή εγγραφή σε μία ολόκληρη σειρά του σύνολο δεδομένων (κάτι που υπήρχε στην αρχική μορφή των σύνολο δεδομένων) καθώς η ένωση των δύο σύνολο δεδομένων έγινε με βάση την στήλη Player. Παρουσιάστηκαν όμως πολλές κενές τιμές κυρίως στις στήλες 3P% και FT% καθώς και σε άλλες στήλες με ποσοστιαίες μετρικές . Οι κενές τιμές μετά από έλεγχο των δεδομένων φάνηκαν ότι είναι μηδενικά ποσοστά δηλαδή από ορισμένες εγγραφές που είχαν στην στήλη 3P και 3PA τιμή μηδέν, είχαν στη στήλη 3P% κενή τιμή. Το ίδιο συνέβαινε και σε εγγραφές στην στήλη FT%. Οι κενές τιμές αυτές αντικαταστάθηκαν από την τιμή 0.0.

Έλεγχος για ύποπτες τιμές

Για την ολοκλήρωση της προετοιμασίας και του καθαρισμού των δεδομένων πραγματοποιήθηκε έλεγχος για ύποπτες τιμές. Συγκεκριμένα, κατά τον υπολογισμό της μεταβλητής EFF από όλα τα δεδομένα (σύνολο δεδομένων EFF) , έγινε η προσπάθεια να βρεθούν ύποπτες τιμές του EFF. Η μεταβλητή EFF καθότι μέτρηση αποδοτικότητας μπορεί σε ακραίες συνθήκες να δώσει λάθος αποτελέσματα. Για παράδειγμα αν υπολογίζαμε την EFF σαν μέσο όρο της ετήσιας EFF θα παίρναμε πολύ υψηλές τιμές σε παίκτες που αγωνιστήκαν σε μία χρόνια πολύ λίγα παιχνίδια αλλά πετύχαν καλά στατιστικά. Για αποφευχθεί αυτό το πρόβλημα χρησιμοποιήθηκε η EFF στα συνολικά στατιστικά καριέρας. Αλλά επειδή και έτσι η αποδοτικότητα μπορεί να δώσει μη αξιόλογα αποτελέσματα θα ψάξουμε για ψηλές τιμές EFF σε αθλητές που δεν αγωνιστήκαν πολύ στο NBA. Με αυτό τον τρόπο, τα αποτελέσματα θα είναι πιο σωστά για να αξιολογήσουμε την καριέρα του παίκτη. Συγκεκριμένα πραγματοποιήθηκε αναζήτηση για τιμές της EFF πάνω από 8 μονάδες σε αθλητές που αγωνίστηκαν σε λιγότερο από 150 παιχνίδια NBA στην καριέρα τους. Ως αποτέλεσμα βρέθηκαν 77 αθλητές, η πλειονότητα των οποίων έχει αγωνιστεί στο NBA μετά το 2010, οπότε θα διαγραφόντουσαν έτσι και αλλιώς από την επεξεργασία δεδομένων που αναφέρθηκε παραπάνω στο τελικό σύνολο δεδομένων. Οι εγγραφές αυτές αφαιρούνται από το σύνολο δεδομένων 1.

Παράδειγμα ύποπτης εγγραφής: Joel Embiid: G: 31 EFF: 20

Στη συνέχεια αναζητήθηκαν αθλητές με μέτριο EFF αλλά με πολύ λίγα παιχνίδια δηλαδή EFF πάνω από 4 μονάδες και παιχνίδια λιγότερα από 50. Τα αποτελέσματα ήταν 110 εγγραφές και αφαιρέθηκαν από το σύνολο δεδομένων 1.

Παράδειγμα ύποπτης εγγραφής: Αντώνης Φώτσης: G: 28 EFF : 4

Κακή Ποιότητα Δεδομένων

Τέλος κατά την διάρκεια της ανάλυσης παρατηρήθηκαν εγγραφές στη στήλη Players, οι οποίες είχαν δίπλα στο όνομα του παίχτη έναν αστερίσκο (*) καθώς και κενά δίπλα σε κάποια ονόματα παικτών, κάτι που εμπόδιζε την σωστή ένωση των δύο αρχικών σύνολο δεδομένων. Οπότε αφαιρέθηκαν όλοι οι αστερίσκοι αλλά και τα κενά , οπότε για να είμαστε σίγουροι ότι θα περιέχονται όλοι οι αθλητές στο τελικό σύνολο δεδομένων, οι εγγραφές στο column players των αρχικών σύνολο δεδομένων θα είναι χωρίς κενό (π.χ MichaelJordan)

Το τελικό σύνολο δεδομένων προς ανάλυση περιέχει 1969 εγγραφές.

3.3 Διερευνητική Ανάλυση

Στο κεφάλαιο αυτό θα διερευνήσουμε τα χαρακτηριστικά των δεδομένων, καθώς και τις όποις πληροφορίες μπορούμε να πάρουμε για αυτά μέσω της διερευνητικής ανάλυσης.

Με την βοήθεια της περιγραφικής στατιστικής θα δούμε τα μέτρα θέσης και διασποράς για τα δεδομένα. Στους παρακάτω πίνακες παρουσιάζονται η μέση τιμή για κάθε μεταβλητή, η τυπική απόκλιση, η ελάχιστη και η μέγιστη τιμή, καθώς και το ενδοτεταρτημοριακό εύρος.

	Age	G	MP	PER	TS%	3PAr	FTr	ORB%	DRB%
count	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00
mean	22.72	45.01	860.78	10.69	0.48	0.11	0.35	7.05	12.92
std	1.79	25.47	801.41	6.64	0.11	0.15	0.26	5.24	6.22
min	18.00	1.00	0.00	-48.60	0.00	0.00	0.00	0.00	0.00
25%	22.00	23.00	168.00	8.35	0.44	0.00	0.21	3.30	8.30
50%	23.00	46.50	618.00	11.30	0.49	0.04	0.31	6.50	12.00
75%	23.00	68.00	1381.00	13.95	0.53	0.19	0.42	9.85	17.20
max	31.00	82.50	3255.00	76.30	1.14	1.00	3.65	60.00	57.40
IQR	1.00	45.00	1213.00	5.60	0.09	0.19	0.21	6.55	8.90

Πίνακας 3.1 : Μέτρα Θέσης και Διασποράς

	TRB%	AST%	STL%	BLK%	TOV%	USG%	OWS	DWS	WS
count	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00
mean	9.97	11.70	1.71	1.42	16.14	19.46	0.53	0.82	1.35
std	4.93	8.78	1.17	1.80	7.33	5.31	1.22	0.93	1.95
min	0.00	0.00	0.00	0.00	0.00	0.00	-2.15	-0.15	-1.20
25%	6.10	5.55	1.10	0.35	12.50	16.25	-0.15	0.15	0.00
50%	9.40	9.20	1.60	0.90	15.15	19.10	0.05	0.50	0.50
75%	13.25	15.70	2.15	1.90	18.85	22.10	0.90	1.20	2.15
max	56.00	57.05	13.40	39.75	100.00	56.60	8.65	7.40	16.05
IQR	7.15	10.15	1.05	1.55	6.35	5.85	1.05	1.05	2.15

Πίνακας 3.2 : Μέτρα Θέσης και Διασποράς

	WS/48	OBPM	DBPM	BPM	VORP	FG	FGA	FG%	3P
count	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00
mean	0.03	-3.14	-1.06	-4.19	0.16	133.58	292.47	0.42	11.06
std	0.12	4.08	2.36	5.23	0.84	146.20	309.17	0.11	23.26
min	-1.29	-46.40	-17.50	-53.60	-1.80	0.00	0.00	0.00	0.00
25%	0.00	-4.70	-2.05	-6.10	-0.25	21.00	51.00	0.38	0.00
50%	0.05	-2.65	-0.95	-3.60	-0.10	77.50	178.50	0.43	1.00
75%	0.09	-0.95	0.25	-1.30	0.20	204.00	446.00	0.47	9.00
max	1.44	28.40	24.15	20.70	7.25	843.00	1588.00	1.00	158.50
IQR	0.09	3.75	2.30	4.80	0.45	183.00	395.00	0.09	9.00

Πίνακας 3.3 : Μέτρα Θέσης και Διασποράς

	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%
count	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00
mean	33.71	0.17	122.52	258.76	0.44	0.44	67.55	93.64	0.66
std	64.07	0.18	137.80	280.69	0.11	0.11	80.03	107.88	0.20
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	1.00	0.00	18.00	44.00	0.40	0.40	10.00	16.00	0.60
50%	5.00	0.16	70.00	158.00	0.45	0.45	37.50	54.50	0.70
75%	33.00	0.31	181.50	386.00	0.49	0.49	98.00	137.50	0.77
max	437.50	1.00	843.00	1445.50	1.00	1.00	602.50	807.00	1.00
IQR	32.00	0.31	163.50	342.00	0.09	0.09	88.00	121.50	0.17

Πίνακας 3.3 : Μέτρα Θέσης και Διασποράς

	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	EFF
count	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00	1969.00
mean	50.51	101.47	151.98	76.51	30.01	19.20	58.53	91.78	345.78	6.89
std	58.31	113.95	169.35	107.02	32.97	32.23	60.52	77.84	378.04	5.18
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-2.00

25%	8.00	17.50	27.00	9.00	5.00	2.00	11.00	23.00	55.00	2.84
50%	29.00	62.00	92.50	36.00	18.00	7.50	38.50	73.00	204.50	5.69
75%	71.00	146.00	217.00	99.50	44.50	22.00	85.50	146.00	521.50	9.78
max	386.50	734.00	1097.00	743.50	214.00	349.50	312.50	347.50	2135.00	29.77
IQR	63.00	128.50	190.00	90.50	39.50	20.00	74.50	123.00	466.50	6.94

Πίνακας 3.4 : Μέτρα Θέσης και Διασποράς

Για την μεταβλητή **EFF** βλέπουμε ότι η μέση τιμή είναι 6.9 με τυπική απόκλιση 5.2

Η μέγιστη τιμή είναι 29.8 και ανήκει στον Larry Bird, δηλαδή έναν παίκτη που κυριάρχησε στο NBA με πληθωρική παρουσία.

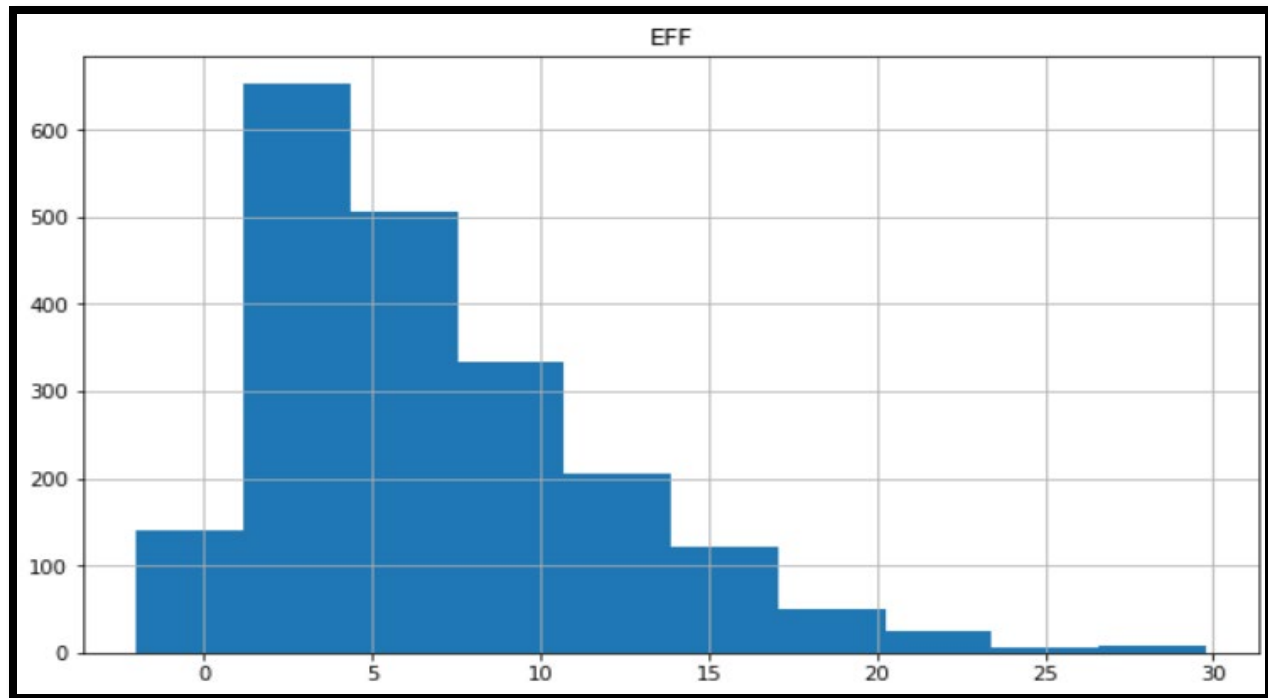
Η ελάχιστη τιμή είναι -0.2 και ανήκει στον Le Mark Baker και στον Mike Champion.

Στη μέση τιμή της EFF περίπου βλέπουμε παίκτες όπως ο Carlos Arroyo και Anthony Avent δηλαδή αθλητές που πραγματοποίησαν κάποιες αξιοπρεπείς σεζόν στο NBA.

Η τιμή του διαμέσου είναι 5.7 και ανήκει στον DeSagana Dior με μέτρια παρουσία στο NBA. Η επικρατέστερη τιμή της μεταβλητής EFF είναι 2 και εμφανίζεται 24 φορές δηλαδή στο 1% των παρατηρήσεων.

Καταλαβαίνουμε ότι οι πλειονότητα των παικτών έχει μέτρια παρουσία στο NBA κάτι που βλέπουμε και στο ιστόγραμμα για την μεταβλητή EFF. Ενώ οι αθλητές που ξεχωρίζουν είναι οι λιγότεροι.

Συγκεκριμένα πάνω από την μέση τιμή βρίσκονται 808 παρατηρήσεις, ενώ κάτω από την μέση τιμή 1159 παρατηρήσεις



Εικόνα 7: Ιστόγραμμα μεταβλητής EFF.

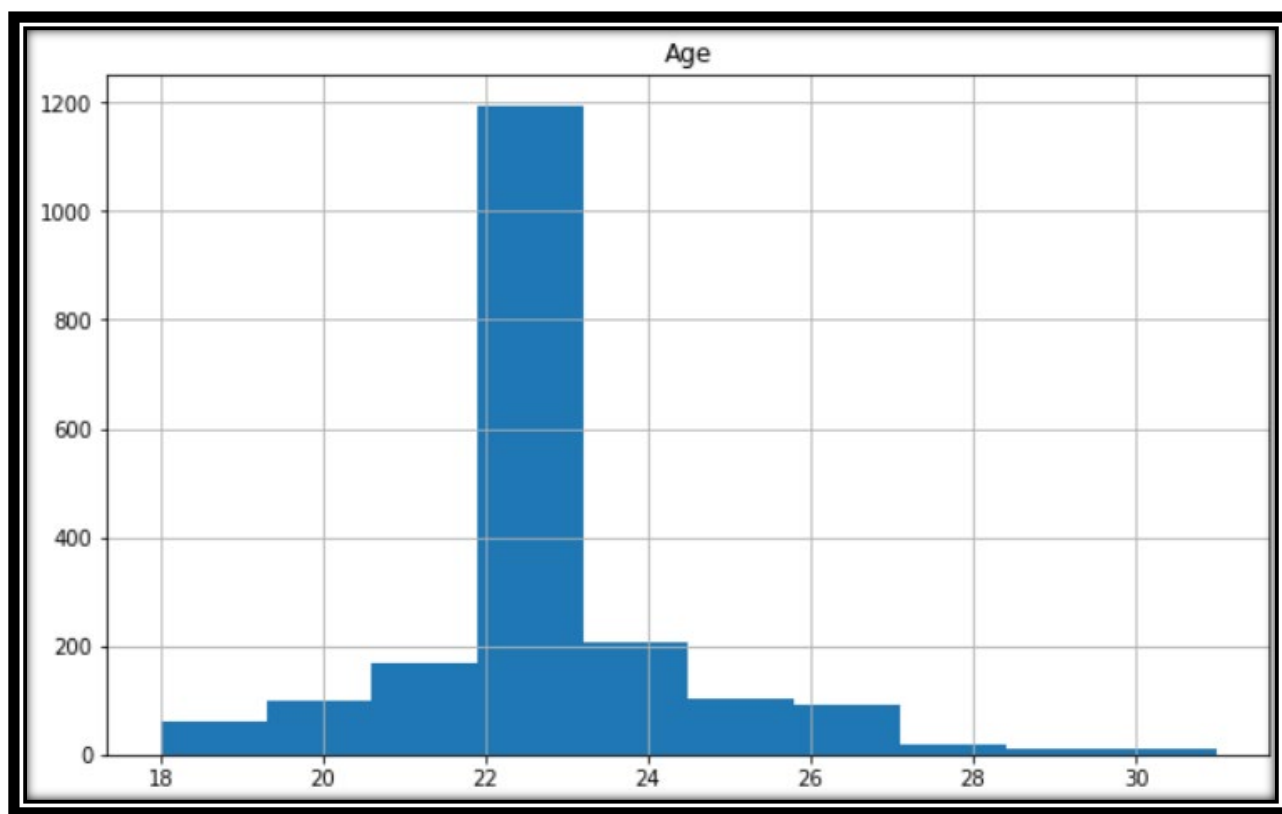
Στο ιστόγραμμα βλέπουμε ότι η πλειονότητα των παρατηρήσεων είναι γύρω από την τιμή 5. Συγκεκριμένα πάνω από 9.7 δηλαδή με αρκετά καλό career efficiency βρίσκονται 512 παίκτες.

Τέλος, παίκτες με career efficiency πάνω από 20 είναι μόνο 40.

Από την λοξότητα (skewness) = 1.08 > 0 βλέπουμε ότι το η πλειονότητα των παρατηρήσεων βρίσκεται δεξιά της κορυφής, κάτι που βλέπουμε και από το σχήμα. Αυτό το καταλαβαίνουμε και από το ότι mean=6.89 > median = 5.69 > mode=2. Ενώ η κύρτωση είναι 1.17.

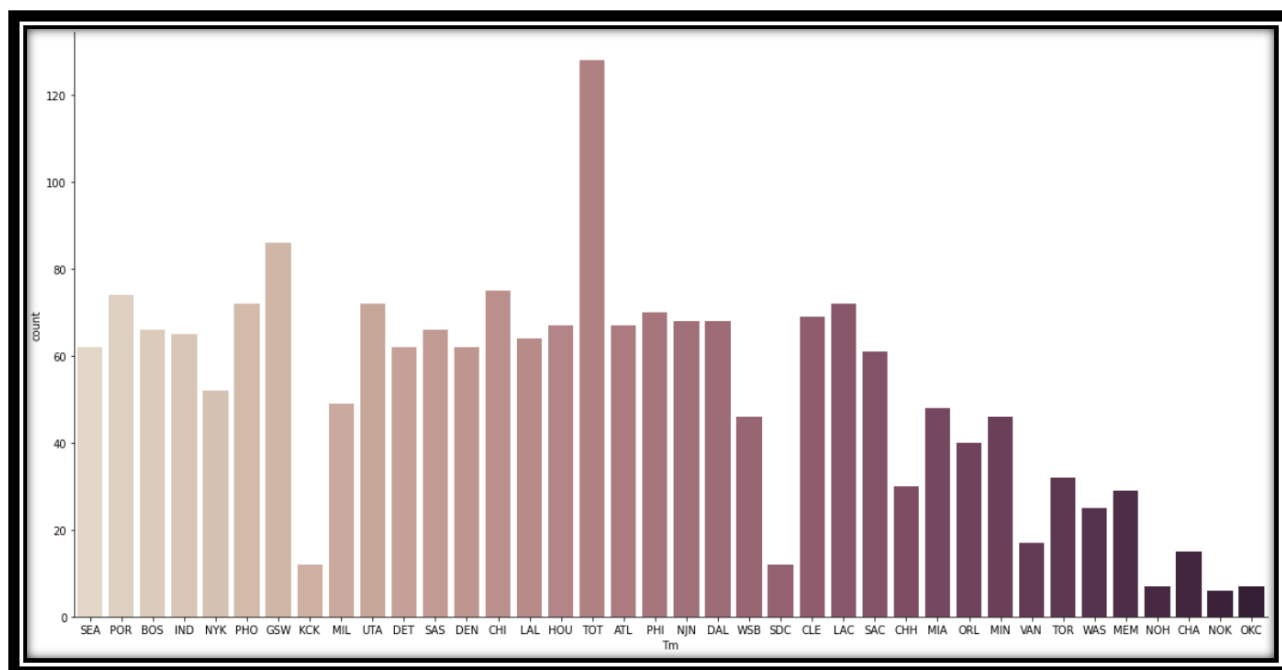
Για την μεταβλητή Age δηλαδή την ηλικία που αγωνίστηκε ένας παίκτης πρώτη φορά στο NBA βλέπουμε ότι η μέση τιμή είναι 22 με τυπική απόκλιση 1.79 κάτι που σημαίνει ότι δεν υπάρχει πολύ μεταβλητότητα στις παρατηρήσεις. Η μέγιστη τιμή είναι 31 και ανήκει στον Arvydas Sabonis, που παρότι έγινε draft σε μικρότερη ηλικία αγωνίστηκε για πρώτη φορά στο NBA σε ηλικία 31 χρονών και μάλιστα αγωνίστηκε σε 520 παιχνίδια στο NBA (Wikipedia.gr) και στον Antoine Rigaudeau με πολύ μικρή συμμετοχή στην μοναδική χρονιά που αγωνίστηκε στο NBA(Wikipedia.gr). Αντιθέτως την μικρότερη ηλικία 18 χρονών παρουσιάζουν 12 εγγραφές συμπεριλαμβανομένου του Kobe Bryant και του Tracy McGrady.

Η παρατηρήσεις της μεταβλητής Age παρουσιάζουν μετρήσεις λοξότητας = 0.84 και κύρτωσης 2 = 2.49 όπως παρατηρείτε και στο ιστόγραμμα όπου η κατανομή φαίνεται λεπτόκυρτη με ουρά προς τα δεξιά. Στο ιστόγραμμα παρατηρούμε και την μικρή μεταβλητότητα των παρατηρήσεων. Η επικρατέστερη τιμή των παρατηρήσεων είναι η τιμή 22 που εμφανίζεται σε 618 παίκτες.



Εικόνα 8: Ιστόγραμμα Μεταβλητής AGE

Όσον αφορά τις επικρατέστερες τιμές στις μεταβλητές **Tm** (Ομάδα που αγωνίστηκε ένας παίκτης για πρώτη φορά) είναι οι Toronto Raptors. Στο διάγραμμα βλέπουμε τις ομάδες και το σύνολο των παικτών που αγωνίστηκαν πρώτη φορά στο NBA σε αυτές. Οι ομάδες με λιγότερους είναι ομάδες όπως οι Oklahoma City Thunders οι οποίοι είχαν ένα ή περισσότερα ονόματα στο παρελθόν(π.χ Seattle Supersonics)



Εικόνα 9: Πλήθος της Μεταβλητής Tm

Ενδιαφέρον παρουσιάζει και το μέσο EFF ανά ομάδα δηλαδή ο μέσος όρος του συνολικού career EFF όλων των παικτών ανά την ομάδα που ξεκίνησαν την καριέρα τους στο NBA.

Ομάδα	Μέσο EFF		Ομάδα	Μέσο EFF
ATL	6.0		NJN	7.5
BOS	6.9		NOH	8.8
CHA	7.0		NOK	7.5
CHH	6.4		NYK	7.1
CHI	7.9		OKC	10.9
CLE	7.3		ORL	7.6
DAL	7.1		PHI	7.2
DEN	7.1		PHO	6.5
DET	6.8		POR	7.1
GSW	7.3		SAC	6.7
HOU	7.1		SAS	6.7
IND	7.1		SDC	6.9
KCK	7.3		SEA	7.0
LAC	6.5		TOR	7.3
LAL	6.6		UTA	6.6
MEM	6.9		VAN	6.8

MIA	7.0		WAS	6.4
MIL	7.4		WSB	6.9
MIN	6.9			

Πίνακας 4: Μέσο EFF ανά ομάδα

Από τον πίνακα βλέπουμε ότι οι OKC (Oklahoma City Thunder) έχουν τους rookies με το μεγαλύτερο ταλέντο, κάτι που είναι λογικό καθώς από την δημιουργία τους το 2008 έχουν κάνει draft παίκτες όπως James Harden και Russel Westbrook. Σε υπολογισμό του μέσου όρου της συγκεκριμένης ομάδας και σαν Seattle Supersonics (SEA) και σαν Oklahoma City Thunder παρατηρούμε ότι το συγκεκριμένη ομάδα έχει μέσο EFF 8.95. Χαμηλότερες τιμές του EFF ανά ομάδα βλέπουμε στο Phoenix Suns και Los Angeles Clippers με 6.5.

Επικρατέστερη τιμή στην μεταβλητή **Pos** (Θέση του παίκτη) είναι η θέση Shooting Guard. Αναλυτικά ανά θέση έχουμε τις εξής παρατηρήσεις:

Point Guard : 372

Shooting Guards : 420

Small Forwards : 397

Power Forwards: 409

Centers: 371

Επίσης οι τιμές της μέσης EFF ανά θέση είναι οι εξής:

Point Guards = 7.28 μέσο EFF καριέρας.

Shooting Guards = 6.23 μέσο EFF καριέρας.

Small Forwards = 6.69 μέσο EFF καριέρας.

Power Forwards = 7.49 μέσο EFF καριέρας.

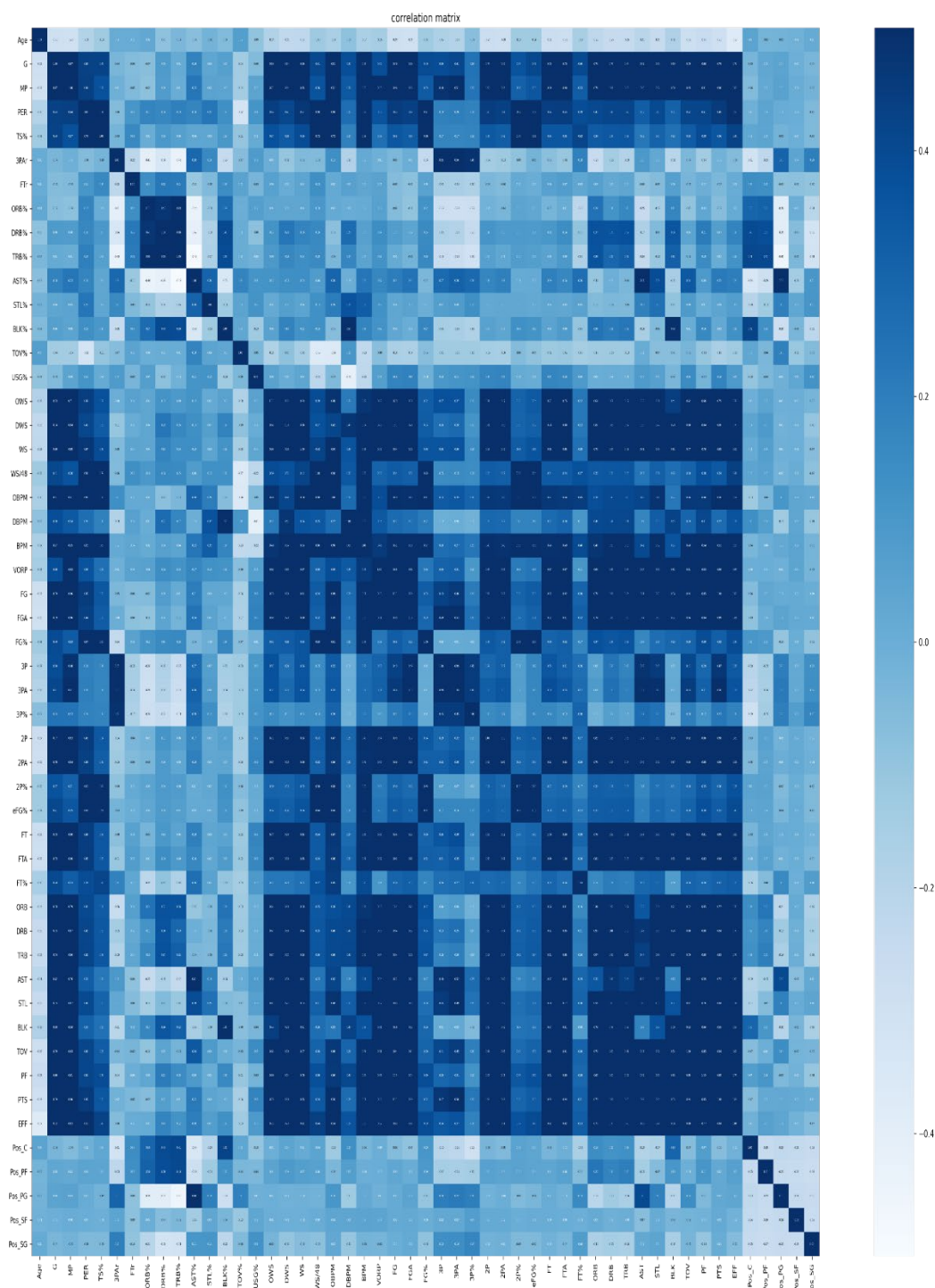
Centers = 6.75 μέσο EFF καριέρας.

Στην παρακάτω εικόνα βλέπουμε και τις συσχετίσεις ανάμεσα στις μεταβλητές τους συνόλου δεδομένων. Όσο πιο σκούρο είναι το χρώμα τόσο μεγαλύτερη η συσχέτιση.

Η μεταβλητή EFF έχει μεγάλη συσχέτιση με την μεταβλητή MP (τα λεπτά στα οποία αγωνίστηκε ο παίκτης). Με συντελεστή συσχέτισης 0.83. Αρκετά υψηλός είναι και συντελεστής συσχέτισης ανάμεσα στην EFF και στη WS(Συνεισφορά του αθλητή στις νίκες τις ομάδας) με συντελεστή 0.81. Εξίσου μεγάλοι είναι και οι συντελεστές συσχέτισης της EFF με τις μεταβλητές FG (0.83) και PTS(0.81). Πολύ υψηλή συσχέτιση έχουν οι μεταβλητές PTS και TOV (0.94) , όπως και η μεταβλητή PTS με την μεταβλητή MP (0.96, σχεδόν απόλυτη σχέση.) Όσων αφορά την μεταβλητή MP παρατηρείται υψηλή (>50) συσχέτιση με 26 μεταβλητές. Κάτι τέτοιο είναι εύκολα ερμηνεύσιμο καθώς ένας παίκτης για να καταγράψει πολλά στατιστικά πρέπει να αγωνιστεί και πολλά λεπτά.

Αναφορικά με τις αρνητικές συσχετίσεις, ενώ δεν παρατηρούνται αρνητικές συσχετίσεις ανάμεσα στις μεταβλητές AST και TRB, παρατηρούνται αρνητικές συσχετίσεις ανάμεσα στις μεταβλητές AST% και

TRB% (-0.49). Η μεταβλητή η οποία φαίνεται από το σχήμα να έχει την μικρότερη συσχέτιση με τις άλλες μεταβλητές είναι η μεταβλητή Age καθώς δεν έχει μεγάλο συντελεστή συσχέτισης με καμία άλλη μεταβλητή.



Εικόνα 10: Correlation Heatmap

3.4 Εφαρμογή Μεθόδων

Στο κεφάλαιο αυτό θα πραγματοποιηθεί η προσπάθεια εύρεσης του μοντέλου πρόβλεψης της μεταβλητής EFF από τα δεδομένα των δύο αρχικών σεζόν του παίκτη.

3.4.1 Πρώτη Εφαρμογή Γραμμικής Παλινδρόμησης

Για την εφαρμογή της μεθόδου της γραμμικής παλινδρόμησης θα χρησιμοποιήσουμε όλες τις μεταβλητές των συνόλου δεδομένων μας εκτός από την μεταβλητή Tm καθώς είναι κατηγορική μεταβλητή με πολλές διαφορετικές τιμές και δεν θα βοηθούσε το μοντέλο μας. Αντίθετα η μεταβλητή Pos θα μετασχηματιστεί σε τέσσερις διαφορετικές μεταβλητές : Pos_PG, Pos_SG, Pos_SF, Pos_PF, Pos_C. Έτσι ο κάθε παίκτης ανάλογα σε πια θέση αγωνίζεται θα έχει τιμή 1 και στις υπόλοιπες 0.

Στον παρακάτω πίνακα βλέπουμε τις πέντε μεταβλητές και στήλες που δημιουργούνται αντί της μεταβλητής και της στήλης Pos.

Player	Pos_PG	Pos_SG	Pos_SF	Pos_PF	Pos_Center
Allen Iverson	1	0	0	0	0
Michael Jordan	0	1	0	0	0
Lebron James	0	0	1	0	0
Larry Bird	0	0	0	1	0
Shaquille O'Neal	0	0	0	0	1

Πίνακας 5 : Δημιουργία dummy μεταβλητών αντί της μεταβλητής Pos.

Αφού δημιουργήσαμε τις τέσσερις αυτές μεταβλητές χρησιμοποιούμε ως εξαρτημένη μεταβλητή την μεταβλητή EFF και ως ανεξάρτητες τις μεταβλητές: 'Age', 'G', 'MP', 'PER', 'TS%', '3PAr', 'FTr', 'ORB%', 'DRB%', 'TRB%', 'AST%', 'STL%', 'BLK%', 'TOV%', 'USG%', 'OWS', 'DWS', 'WS', 'WS/48', 'OBPM', 'DBPM', 'BPM', 'VORP', 'FG', 'FGA', 'FG%', '3P', '3PA', '3P%', '2P', '2PA', '2P%', 'eFG%', 'FT', 'FTA', 'FT%', 'ORB', 'DRB', 'TRB', 'AST', 'STL', 'BLK', 'TOV', 'PF', 'PTS', 'Pos_C', 'Pos_PF', 'Pos_PG', 'Pos_SF', 'Pos_SG'.

Τα δεδομένα χωρίζονται σε δεδομένα εκπαίδευσης και δεδομένα επαλήθευσης με αναλογία 80% δεδομένα εκπαίδευσης και 20% δεδομένα επαλήθευσης. Το μοντέλο πρόβλεψης που δημιουργείται είναι το εξής:

$$\begin{aligned} EFF = & 9.79 - 0.39 * Age + 0.00005736 * G - 0.002 * MP + 0.0541 * PER + 0.717 * TS\% - 1.64 * 3Par \\ & - 0.05 * FTr - 0.149 * ORB\% - 0.09 * DRB\% + 0.27 * TRB\% - 0.0032 * AST\% - 0.1547 * STL\% + \\ & 0.03 * BLK\% + 0.001 * TOV\% + 0.02 * USG\% + 1.87 * OWS + 1.995 * DWS - 6.65 * WS/48 + 2.7 * \\ & OBPM + 2.61 * DBPM - 2.45 * BPM - 0.6 * VORP - 0.001 * FG + 0.003 * FGA - 2.46 * FG\% - 0.02 * \\ & 3P + 0.01 * 3PA + 0.86 * 3P\% + 0.2 * 2P - 0.009 * 2PA + 1.68 * 2P\% + 0.62 * eFG\% + 0.02 * FT - \\ & 0.011 * FTA + 0.9 * FT\% + 0.0012 * ORB + 0.003 * DRB + 0.0044 * TRB + 0.062 * AST + 0.02 * STL \\ & + 0.012 * BLK - 0.0018 * TOV + 0.0005 * PF - 0.0003 * PTS + 1.95 * Pos_C + 2 * Pos_PF + 2.2 * \\ & Pos_PG + 1.7 * Pos_SF + 1.9 * Pos_SG \end{aligned}$$

Για να υπολογίσουμε την στατιστική σημαντικότητα του μοντέλου θα κοιτάξουμε την τιμή F. Η τιμή F χρησιμοποιείται ως για να δούμε αν ισχύει η μηδενική υπόθεση, δηλαδή ότι όλοι οι συντελεστές παλινδρόμησης είναι ίσοι με μηδέν, δηλαδή το μοντέλο δεν έχει δυνατότητα πρόβλεψης. Η διαδικασία αυτή γίνεται μέσω μίας δοκιμής του μοντέλου μόνο με τον συντελεστή β και με μηδενικές μεταβλητές πρόβλεψης και αποφασίζει εάν οι πρόσθετοι συντελεστές βελτίωσαν τη λειτουργία του μοντέλου. Το μοντέλο που δημιουργήθηκε έχει F-Statistic 144 το οποίο σε επίπεδο σημαντικότητας $df = 45$ είναι αρκετό για να απορρίψουμε την Null υπόθεση, δηλαδή ότι η μεταβλητή EFF δεν έχει στατιστική σχέση με τις υπόλοιπες μεταβλητές.

Για να δούμε πόσο ποσοστό της μεταβολής του EFF καθορίζεται από τις υπόλοιπες μεταβλητές θα αξιολογήσουμε το R τετράγωνο (R^2). Το R^2 τετράγωνο είναι ένα στατιστικό μέτρο που δείχνει το ποσοστό της διακύμανσης για μια εξαρτημένη μεταβλητή που εξηγείται από μια ανεξάρτητη μεταβλητή ή μεταβλητές σε ένα μοντέλο παλινδρόμησης. Βρίσκουμε $R^2 = 0.809$ που σημαίνει ότι το 80.9 % της μεταβολής της εξαρτημένης μεταβλητής εξηγείται από τις ανεξάρτητες μεταβλητές. Το μοντέλο έχει ικανοποιητικό R^2 .

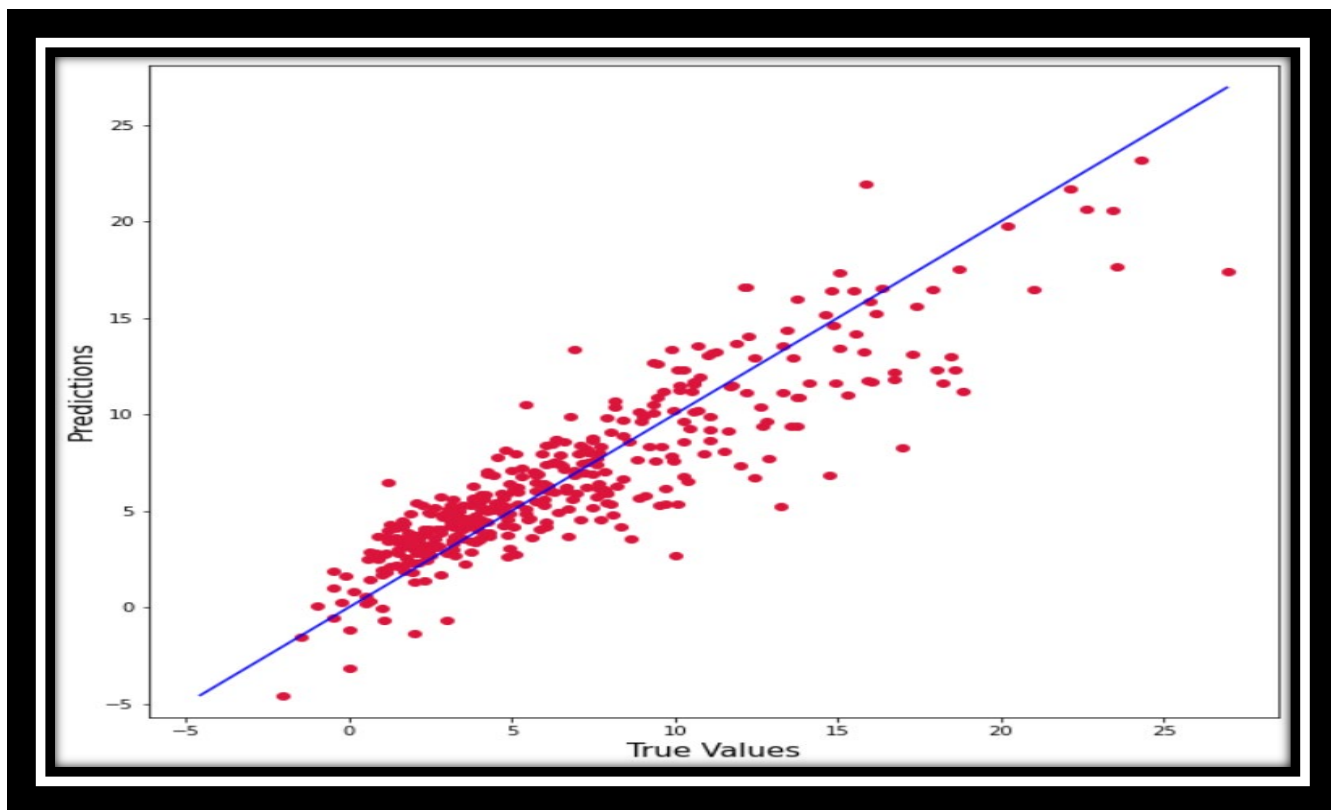
Στη συνέχεια το μοντέλο πραγματοποιεί προβλέψεις στα test data. Στον παρακάτω πίνακα βλέπουμε τυχαίες 20 προβλέψεις, την πραγματική τιμή καθώς και την διαφορά τους.

Actual	Predicted	Residual
10.55	10.13	0.41
1.05	-0.64	1.69
4.31	3.65	0.65
1.45	4.08	-2.63
1.88	4.88	-3
6.59	8.55	-1.96
10.38	6.51	3.86
3.33	4.81	-1.48
2.71	3.10	-0.39
14.82	16.39	-1.57

9.64	11.21	-1.57
1.14	2.78	-1.64
3.10	4.39	-1.29
10.16	11.48	-1.32
11.11	13.16	-2.05
0.50	0.54	-0.04
3.57	4.58	-1.01
6.27	7.52	-1.25
4.25	5.01	-0.760
9.47	10.86	-1.39

Πίνακας 6 : Διαφορά πραγματικών έναντι προβλεφθέντων τιμών.

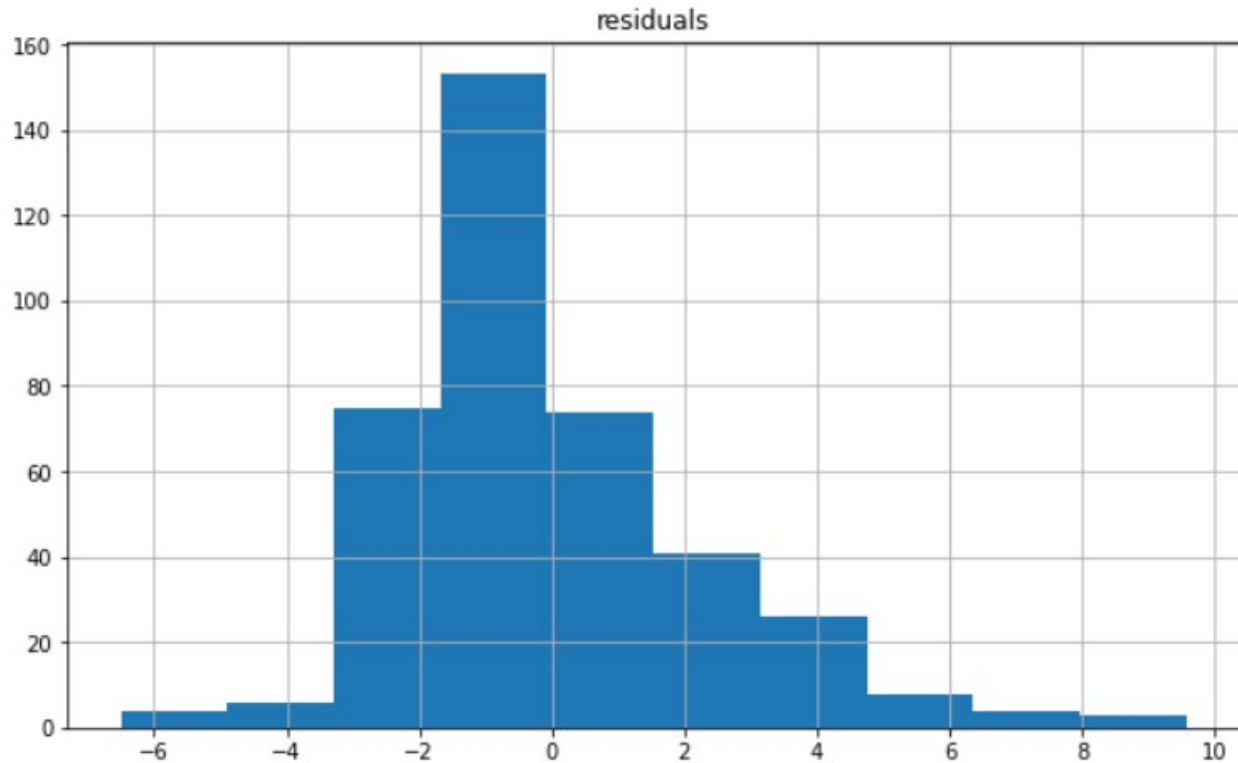
Στο επόμενο διάγραμμα απεικονίζονται οι τιμές της EFF που πρόβλεψε το μοντέλο μας στα test δεδομένα σε σχέση με τα τις πραγματικές τιμές.



Εικόνα 11: Πραγματικές - Προβλεφθείσες Τιμές(Linear Regression)

Από το διάγραμμα βλέπουμε ότι υπάρχει σχέση μεταξύ πραγματικών και προβλεφθέντων τιμών της EFF.

Αξίζει επίσης να παρατηρήσουμε την κατανομή των καταλοίπων δηλαδή της διαφοράς μεταξύ πραγματικής και προβλεφθείσας τιμής.



Εικόνα 12: Ιστόγραμμα καταλοίπων

Στο ιστόγραμμα βλέπουμε ότι τα κατάλοιπα ακολουθούν μία κατανομή η οποία είναι κοντά σε κανονική κατανομή.

Εφόσον είδαμε ότι το μοντέλο μας είναι στατιστικά σημαντικό και εξηγεί ένα καλό ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής, πρέπει να αξιολογήσουμε το προβλέψεις που παράγει. Για να αξιολογήσουμε τις προβλέψεις του μοντέλου θα χρησιμοποιήσουμε το μέσο απόλυτο λάθος (mean absolute error), το μέσο τετραγωνισμένο σφάλμα (mean squared error) καθώς και την ρίζα του μέσου τετραγωνισμένου σφάλματος

Τύπος mean absolute error :

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

Τύπος mean squared error:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - x_i)^2$$

Τύπος root mean squared error:

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

Όπου y_i η προβλεπόμενη τιμή, x_i η πραγματική τιμή και n είναι το πλήθος ο συνολικός αριθμός δεδομένων.

Καθώς στο `rmse` η απόσταση τις προβλεπόμενης τιμής από την πραγματική έχει τετραγωνιστεί αυτό σημαίνει ότι τα μεγάλα σφάλματα έχουν μεγαλύτερο βάρος και επιβαρύνουν περισσότερο την μετρική

Μία άλλη μετρική που θα μπορούσε να χρησιμοποιηθεί θα ήταν η MAPE (mean absolute percentage error) η οποία θα έδειχνε το ποσοστό του λάθους όμως σε περιπτώσεις που η x_i δηλαδή η πραγματική τιμή της εξαρτημένης μεταβλητής παίρνει τιμή 0 η MAPE δεν βγάζει σωστά αποτελέσματα.

Στο συγκεκριμένο μοντέλο το MAE είναι 1.74. Το MAE μας δείχνει ότι κατά μέσο όρο η πρόβλεψη στην μεταβλητή EFF θα απέχει 1.74 από την πραγματική τιμή. Η συγκεκριμένη τιμή δείχνει ότι μπορεί το μοντέλο να μην καταφέρνει να προβλέψει σε απόλυτη ακρίβεια την μεταβλητή EFF, αλλά δίνει στις ομάδες μία αρκετά καλή προσέγγιση. Η μετρική Mean Squared Error είναι 5.32 και η μετρική Root Mean Squared Error είναι 2.3

Αναφορικά με τους συντελεστές του μοντέλου βλέπουμε κάποια παράδοξα. Για παράδειγμα ενώ οι συντελεστές OBPM και DBPM είναι μεγάλοι (2.7 και 2.61) που σημαίνει ότι για παράδειγμα μία μεταβολή της μετρικής OBPM κατά μία μονάδα θα οδηγήσει σε αύξηση της EFF κατά 2.7 μονάδες, ενώ αντίστοιχα η μετρική BPM έχει αρνητικό συντελεστή -2.45. Οι τρεις μετρικές αυτές δείχνουν την συνεισφορά του παίκτη στην ομάδα ανάλογα με την αλλαγή του σκορ όσο βρίσκεται στο παρκέ. Η μετρική OBPM δείχνει την επιθετική συνεισφορά, η DBPM την αμυντική και η BPM την συνολική. Αν ανατρέξουμε και στο χάρτη συσχετίσεων (Εικόνα 9) θα δούμε ότι η μετρική OBPM έχει συντελεστή συσχέτισης με την μετρική BPM 0.9.

Όταν σε ένα σύνολο δεδομένων υπάρχουν πολλές μεταβλητές που συσχετίζονται μεταξύ τους παρατηρούνται προβλήματα πολυσυγγραμμικότητας (multicollinearity). Συγκεκριμένα η πολυσυγγραμμικότητα φαίνεται ότι έχει επηρεάσει του συντελεστές του μοντέλου καθώς και την στατιστική σημαντικότητα αυτών, καθώς παρότι το μοντέλο σαν σύνολο είναι στατιστικά σημαντικό, αρκετοί συντελεστές έχουν αυξημένα p-values, κάτι που τους καθιστά μη στατιστικά σημαντικούς. Κατά συνέπεια είναι δύσκολο να ερμηνεύσουμε ποια στατιστικά ενός αθλητή είναι πραγματικά σημαντικά για την εξέλιξη της καριέρας του.

Ένα ακόμη πρόβλημα που προκαλεί στα μοντέλα η πολυσυγγραμμικότητα, όσων αφορά τις προβλέψεις, είναι η υπερπροσαρμογή (overfit) του μοντέλου στα δεδομένα εκπαίδευσης. Αυτό σημαίνει ότι το μοντέλο αποδίδει πολύ καλύτερα στα δεδομένα εκπαίδευσης, σε σχέση με τα δεδομένα δοκιμής. Για να ελέγξουμε αν υπάρχει overfit θα συγκρίνουμε το MAE των δεδομένων δοκιμής με τον MAE στα δεδομένα εκπαίδευσης. Συγκεκριμένα στα δεδομένα εκπαίδευσης ο MAE είναι 1.69 δηλαδή είναι κοντά στον MAE που βρήκαμε (1.74). Το μοντέλο ως προς την ικανότητα πρόβλεψης δεν έχει επηρεαστεί από overfit.

3.4.2 Επιλογή Μεταβλητών

Η επιλογή των μεταβλητών σε προβλήματα μηχανικής μάθησης, αποτελεί ένα ευρέως χρησιμοποιούμενο εργαλείο για την βελτίωση της απόδοσης των μοντέλων, αλλά και για την βελτίωση της ερμηνευτικότητας τους. (Jianyu Miao, Lingfeng Niu, 2019)

Sequential Feature Selection

Αφού δημιουργήθηκε το πρώτο μοντέλο πρόβλεψης με την βοήθεια της πολλαπλής γραμμικής παλινδρόμησης, θα γίνει η προσπάθεια για βελτίωση του μοντέλου μέσω της επιλογής των κατάλληλων ανεξάρτητων μεταβλητών.

Ένας τρόπος να επιλέξουμε με αυτόματο τρόπο ορισμένους συντελεστές για το μοντέλο, είναι οι μέθοδοι Wrapper, μέσω των μεθόδων αυτών οι αλγόριθμοι επιλέγουν τους κατάλληλους συντελεστές μέσω μίας διαδικασίας η οποία ελέγχει το αν μία ανεξάρτητη μεταβλητή είναι κατάλληλη για πρόσθεση ή αφαίρεση από το σύνολο των επεξηγηματικών μεταβλητών με βάση κάποιο προκαθορισμένο κριτήριο, το οποίο συνήθως είναι η στατιστική σημαντικότητα μιας μεταβλητής.

Forward Selection: Το μοντέλο ξεκινάει χωρίς μεταβλητές και στη συνέχεια γίνεται η προσθήκη της μεταβλητής με βάση ένα συγκεκριμένο κριτήριο, της οποίας η συμπερίληψη δίνει την πιο στατιστικά σημαντική βελτίωση του μοντέλου. Η διαδικασία επαναλαμβάνεται έως ότου καμία προσθήκη δεν βελτιώσει το μοντέλο σε στατιστικά σημαντικό βαθμό.

Backward Elimination : Το μοντέλο ξεκινάει με όλες τις μεταβλητές, διαγράφοντας μία μεταβλητή τη φορά καθώς εξελίσσεται το μοντέλο με βάση το αν η διαγραφή αυτής της μεταβλητής επιφέρει την πιο στατιστικά ασήμαντη επιδείνωση της προσαρμογής του μοντέλου.

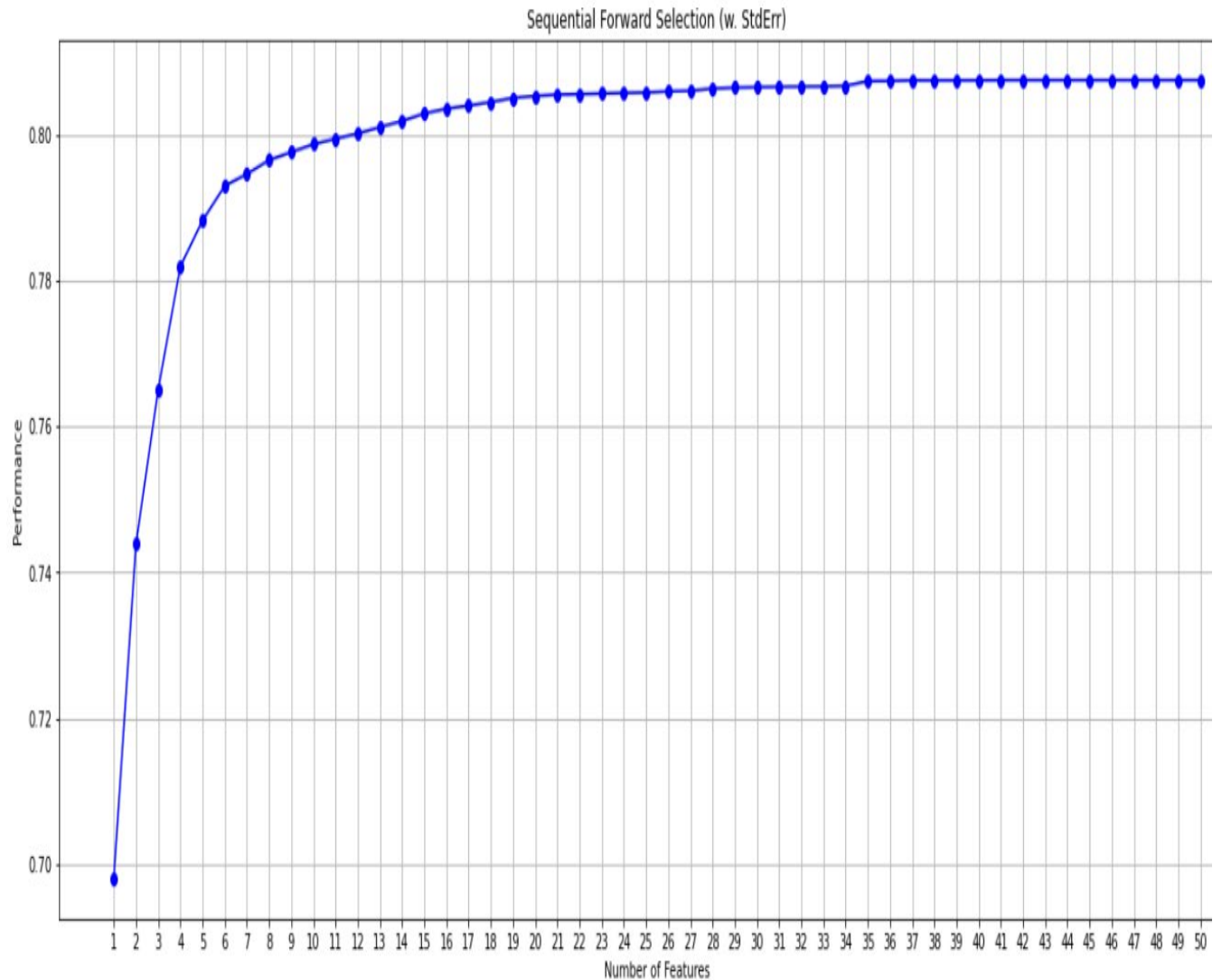
Bidirectional elimination : Συνδυασμός των δύο παραπάνω.

Διαδοχικοί αλγόριθμοι επιλογής μεταβλητών (Sequential Feature Selection Algorithms). Είναι κομμάτι των Wrapper μεθόδων και πραγματοποιούν μείωση των ανεξάρτητων μεταβλητών με σκοπό την βελτίωση της απόδοσης του μοντέλου

(Wikipedia.org, statisticshowto.com, analyticsindiamag.com)

Στη συγκεκριμένη περίπτωση χρησιμοποιήθηκε η μέθοδος Forward Selection.

Για να πραγματοποιηθεί ο Sequential Feature Selection αλγόριθμος αρχικά πρέπει να δούμε πόσος είναι ο αριθμός των ανεξάρτητων μεταβλητών που κάνει το μοντέλο να αποδίδει καλύτερα.



Εικόνα 13: Ιδανικός αριθμός μεταβλητών

Σύμφωνα με και με το διάγραμμα ο ιδανικός αριθμός μεταβλητών είναι 35.

Οι μεταβλητές που επιλέχθηκαν είναι οι :

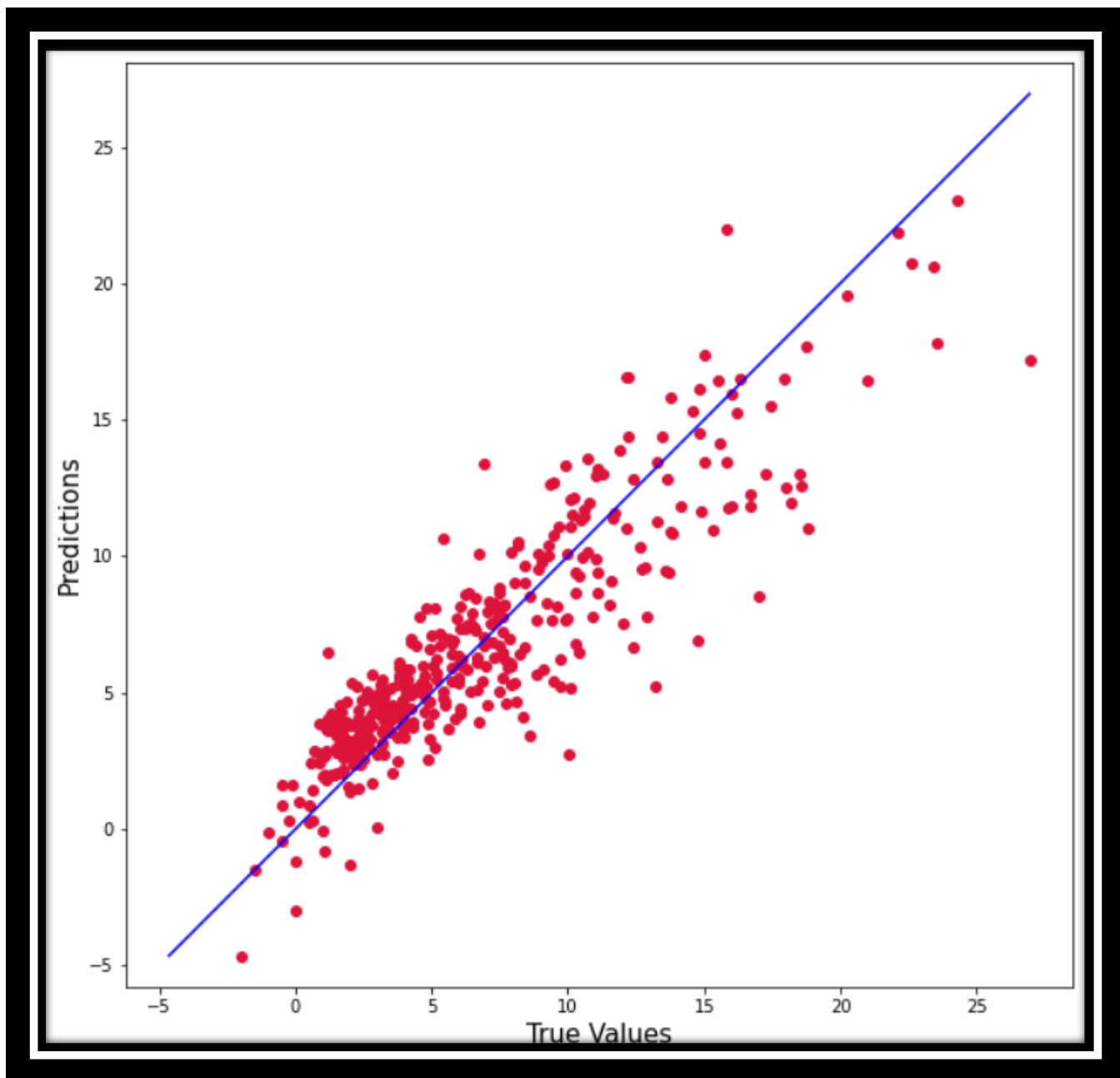
'Age','MP','PER','TS%','3PAr','FTr','DRB%','STL%','BLK%','USG%','WS','WS/48','OBPM','DBPM','BPM','VORP','FG','FGA','FG%','3P%','2PA','2P%','FT','FTA','FT%','ORB','DRB','AST','STL','BLK','TOV','PTS','Pos_C','Pos_PG','Pos_SG'

Το μοντέλο γραμμικής παλινδρόμησης θα τρέξει με μόνο αυτές τις μεταβλητές :

const	11.2708
Age	-0.3898
MP	-0.0025
PER	0.0396
TS%	0.8783
3PAr	-1.4621
FTr	-0.0423
DRB%	0.0490
STL%	-0.1696
BLK%	0.0294
USG%	0.0274
WS	0.8419
WS/48	-5.4375
OBPM	2.6557
DBPM	2.5876
BPM	-2.4125
VORP	-0.6620
FG	0.1044
FGA	0.0168
FG%	-1.6062
3P%	0.8603
2PA	-0.0219
2P%	1.5345
FT	0.0663
FTA	-0.0113
FT%	0.9279
ORB	0.0053
DRB	0.0084
AST	0.0059
STL	0.0287
BLK	0.0128
TOV	-0.0003
PTS	-0.0434
Pos_C	0.0385
Pos_PG	0.4501
Pos_SG	0.1554

Το μοντέλο έχει $R^2 = 0.809$ και $F\text{-Statistic} = 185$. Το R^2 σημαίνει ότι το 80.9 % της μεταβολής της εξαρτημένης μεταβλητής εξηγείται από τις ανεξάρτητες μεταβλητές δηλαδή ίδιο με το αρχικό μοντέλο. Το F Stat είναι ελαφρώς αυξημένο που σημαίνει ότι το μοντέλο παραμένει στατιστικά σημαντικό παρότι αρκετές μεταβλητές εμφανίζονται στατιστικά μη σημαντικές ($p\text{-value} > 0.05$)

Οι προβλεπόμενες τιμές σε σχέση με τις πραγματικές τιμές απεικονίζονται στο παρακάτω διάγραμμα.



Εικόνα 14 : Πραγματικές-Προβλεφθείσες Τιμές

Ως προς την αξιολόγηση των προβλέψεων στα test δεδομένα , το MAE παρουσίασε μία μικρή μείωση σε 1.72 από 1.74 ,το MSE είναι 5.23 από 5.32 και το RMSE έπεσε σε 2.28 από 2.30. Το μοντέλο βελτιώθηκε ελαφρώς ως προς την δυνατότητα προβλέψεων.

Επιλογή μεταβλητών με βάση τον συντελεστή συσχέτισης με την EFF.

Στη συνέχεια θα γίνει επιλογή των μεταβλητών ένα μοντέλο με βάση τις μεταβλητές που έχουν μεγάλη συσχέτιση με την μεταβλητή EFF . Σαν κριτήριο θα είναι ο συντελεστής συσχέτισης. Ο συντελεστής συσχέτισης είναι ένα μέτρο που ποσοτικοποιεί την ισχύ της σχέσης μεταξύ δύο μεταβλητών.

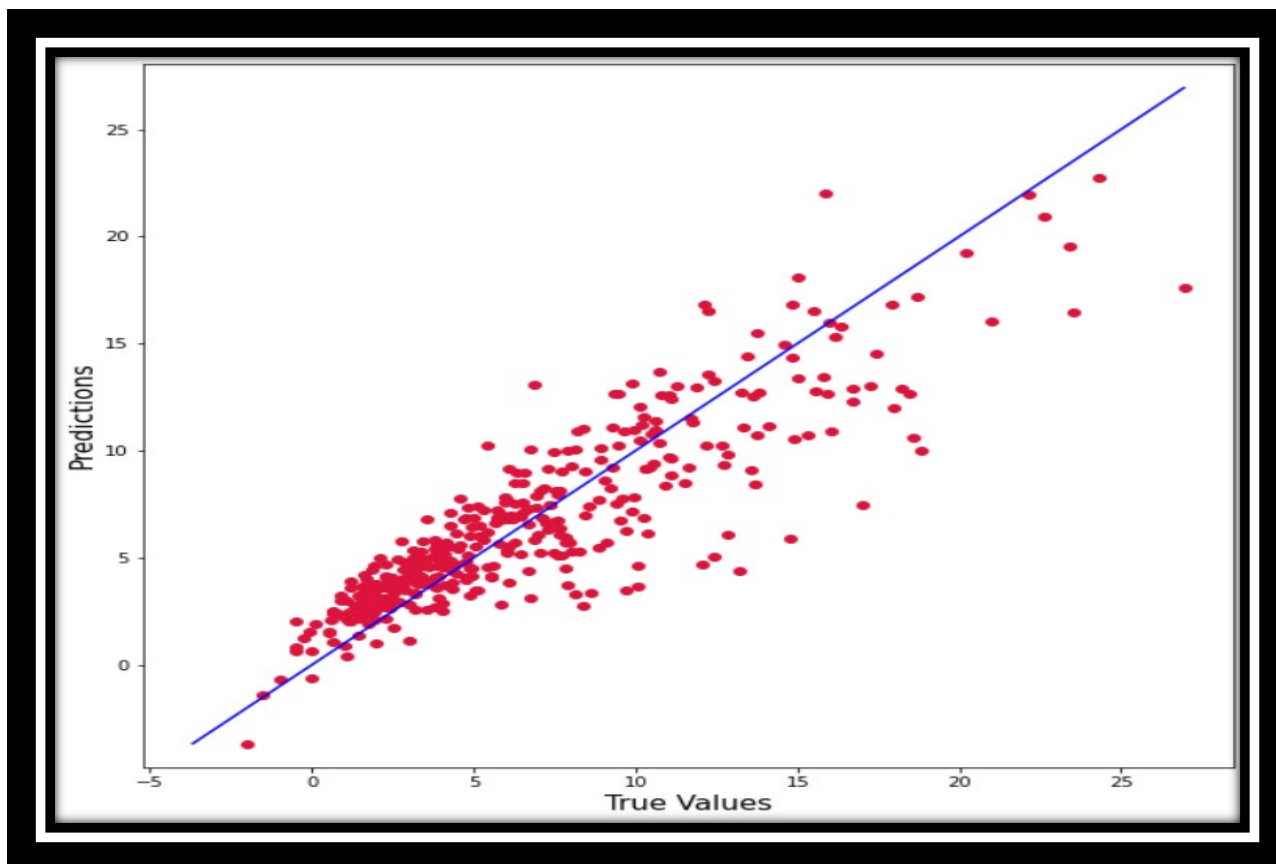
Στη συγκεκριμένη περίπτωση θα γίνει επιλογή με βάση το αν ο συντελεστής συσχέτισης είναι μεγαλύτερος του 0.5

Οι μεταβλητές που επιλέχτηκαν με βάση αυτό το κριτήριο :

G , MP, PER , OWS , DWS, WS, OBPM, BPM, VORP, FG, FGA, 2P, 2PA, FT, FTA, ORB, DRB, TRB, AST, STL, BLK, TOV, PF, PTS

Το μοντέλο πρόβλεψης μετά την επιλογή συντελεστών έχει R^2 0,786 , ελαφρώς χαμηλότερο από το προηγούμενο μοντέλο, και F 258 , ακόμα μεγαλύτερο δηλαδή , οπότε το μοντέλο παραμένει στατιστικά σημαντικό. Παρόλα αυτά αρκετοί συντελεστές παραμένουν στατιστικά ασήμαντοι ($p\text{-values} > 0.05$) . Η μειωμένη στατιστική σημαντικότητα είναι αποτέλεσμα της πολυσυγγραμμικότητας

Ως προς την αποτελεσματικότητα ως μοντέλο πρόβλεψης, το μοντέλο αποδίδει χειρότερα στα δεδομένα ελέγχου, καθώς το MAE αυξήθηκε σε 1.76 , το MSE σε 5.72 και το RMSE σε 2.39



Εικόνα 15: Πραγματικές-Προβλεφθείσες Τιμές

Αξίζει να σημειωθεί ότι με την επιλογή συντελεστών βάση του συντελεστή συσχέτισης, το μοντέλο αποδίδει καλύτερα στα δεδομένα δοκιμής από ότι στα δεδομένα εκπαίδευσης, καθώς οι δείκτες MAE, MSE, και RMSE με τιμές 1.77, 5.81 και 2.41 στα δεδομένα εκπαίδευσης, είναι μεγαλύτερες από τους αντίστοιχους δείκτες στα δεδομένα δοκιμής.

Επιλογή Μεταβλητών για Μείωση Πολυσυγγραμμικότητας

Όπως αναφέρθηκε παραπάνω τα μοντέλα γραμμικής παλινδρόμησης που χρησιμοποιήθηκαν, παρότι πραγματοποίησαν αποδεκτά αποτελέσματα ως προς της πρόβλεψη, παρουσίασαν προβλήματα λόγω της αυξημένης συσχέτισης ανάμεσα στις ανεξάρτητες μεταβλητές. Το πρόβλημα της πολυσυγγραμμικότητας οδήγησε σε μοντέλα τα οποία είναι δύσκολο να ερμηνευθούν. Η ύπαρξη πολυσυγγραμμικότητας σε ένα σύνολο δεδομένων οδηγεί δεν μας επιτρέπει να αξιολογήσουμε σωστά του συντελεστές των ανεξάρτητων μεταβλητών καθώς μία στατιστικά σημαντική μεταβλητή μπορεί να γίνει ασήμαντη εξαιτίας της γραμμικής σχέσης με άλλες ανεξάρτητες μεταβλητές. Η ύπαρξη δύο μεταβλητών οι οποίοι επηρεάζουν με ίδιο τρόπο την εξαρτημένη μεταβλητή οδηγούν λόγω πλεονασμού, ειδικά σε γραμμικά μοντέλα όπως το συγκεκριμένο που εξετάζουμε, σε αναξιόπιστους συντελεστές. (blog.clairvoyantsoft.com)

Συγκεκριμένα παρατηρούμε ότι για παράδειγμα, οι μεταβλητές OWS και WS έχουν συντελεστή συσχέτισης 0.93 αλλά οι συντελεστές τους στο αρχικό μοντέλο παλινδρόμησης έχουν διαφορετικά

πρόσιμα. Επίσης παρότι το μοντέλο είναι στατιστικά σημαντικό σαν σύνολο, οι περισσότερες μεταβλητές, οι 32 από τις 45 στο αρχικό μοντέλο έχουν $p\text{-value} > 0.05$. Το ίδιο συμβαίνει και στα υπόλοιπα δύο μοντέλα μετά την επιλογή μεταβλητών που έγινε παραπάνω.

Κρίνεται σκόπιμο να γίνει προσπάθεια εξάλειψης της πολυσυγγραμμικότητας για να μπορέσουμε να εξάγουμε ορθά συμπεράσματα. Για να καταφέρουμε να υπολογίσουμε τους συντελεστές οι οποίοι επιβαρύνουν το μοντέλο θα χρησιμοποιήσουμε την μετρική VIF (variance inflation factor). Ο VIF (Gareth James κ.α, 2013) είναι ο λόγος της διακύμανσης του ενός συντελεστή β του κατά την προσαρμογή του πλήρους μοντέλου διαιρεμένος με τη διακύμανση του ίδιου συντελεστή β εάν στο μοντέλο υπήρχε μόνο αυτός.

$$VIF(\beta) = \frac{1}{1 - R^2_{x_j|x_{-j}}}$$

Το $R^2_{x_j|x_{-j}}$ είναι το R^2 από μια παλινδρόμηση του X_j σε όλους τους άλλους προγνωστικούς παράγοντες. Αν το $R^2_{x_j|x_{-j}}$ είναι κοντά στο ένα, τότε υπάρχει συγγραμμικότητα, και έτσι το VIF θα είναι μεγάλο. Μία τιμή VIF που υπερβαίνει το 5 ή το 10 υποδηλώνει ότι η μεταβλητή αυτή έχει γραμμική σχέση με άλλες μεταβλητές, και επηρεάζει το μοντέλο.

Στη συγκεκριμένη περίπτωση θα μελετήσουμε τις τιμές VIF στο μοντέλο που δημιουργήθηκε μετά την εφαρμογή του Forward Sequential Feature Selection, καθώς ήταν και ήταν το μοντέλο γραμμικής παλινδρόμησης που απέδιδε καλύτερα, αλλά και έχει λιγότερες ανεξάρτητες μεταβλητές από το αρχικό μοντέλο.

Συγκεκριμένα οι τιμές VIF στο μοντέλο είναι οι εξής :

Μεταβλητή	VIF
Age	1.2
MP	60.8
PER	28.1
TS%	10.5
3PAr	3.9
FTTr	1.4
DRB%	3.3
STL%	3.1
BLK%	4
USG%	4
WS	34.8
WS/48	21.4
OBPM	11089.5
DBPM	3694.5
BPM	18183.4

VORP	9.4
FG	16306.7
FGA	2314.5
FG%	24
3P%	1.7
2PA	2138.1
2P%	13.3
FT	1290
FTA	119.8
FT%	2.1
ORB	15.1
DRB	19.7
AST	17.2
STL	11.8
BLK	4.5
TOV	44.2
PTS	25117.1
Pos_C	1.6
Pos_PG	2.5
Pos_SG	1.7

Πίνακας 7 : Αρχικές Τιμές VIF.

Από τις μετρήσεις τις VIF φαίνεται το πρόβλημα της πολυσυγγραμμικότητας καθώς τιμές VIF κάτω από 5 έχουν μόνο 10 μεταβλητές.

Για να πετύχουμε την εξάλειψη της πολυσυγγραμμικότητας, θα αφαιρούμε τις μεταβλητές με την υψηλότερη τιμή VIF και μετά θα ξανά υπολογίσουμε τις τιμές VIF. Η διαδικασία αυτή θα επαναληφθεί μέχρι να κρατήσουμε μόνο μεταβλητές με τιμές κάτω από 5. Αφού πραγματοποιήθηκε αυτή η διαδικασία το μοντέλο έχει τις μεταβλητές : 'Age', 'PER', '3PAr', 'FTr', 'DRB%', 'STL%', 'BLK%', 'USG%', 'DBPM', 'VORP', '3P%', '2P%', 'FT', 'FT%', 'AST', 'BLK', 'Pos_C', 'Pos_PG', 'Pos_SG'

Οι τιμές VIF είναι οι εξής :

Μεταβλητή	VIF
Age	1.1
PER	2.6
3PAr	1.7
FTr	1.2
DRB%	1.8
STL%	2.2
BLK%	2.5
USG%	1.9
DBPM	3.9

VORP	2.3
3P%	1.6
2P%	1.8
FT	4
FT%	1.5
AST	3.1
BLK	2.6
Pos_C	1.5
Pos_PG	2.2
Pos_SG	1.6

Πίνακας 7 : Τελικές Τιμές VIF.

Το μοντέλο έχει R2 0.78 με MAE 1.8, MSE 5.77 και RMSE 2.4 .

Ερμηνεία Μεταβλητών

Εφόσον το μοντέλο έχει καθαρίσει από την πολυσυγγραμμικότητα μπορούμε να διώξουμε τις μεταβλητές με $p\text{-value} > 0.05$. Τέτοιες μεταβλητές είναι οι : Pos_C ($p\text{-value} = 0.87$), Pos_PG ($p\text{-value} = 0.49$), Pos_SG($p\text{-value} = 0.76$), 3PAr ($p\text{-value} = 0.887$), USG%($p\text{-value} = 0.07$), έτσι ώστε να εξηγήσουμε τις στατιστικά σημαντικές μεταβλητές.

Αφού αφαιρέσαμε και αυτές τις μεταβλητές το μοντέλο παρουσιάζει μια πολύ μικρή βελτίωση ως προς την απόδοση στις προβλέψεις. Συγκεκριμένα το μοντέλο έχει ίδιο R2 και οι δείκτες MAE, MSE και RMSE έχουν τιμές 1.79, 5.74 και 2.39 αντίστοιχα.

Παρακάτω βλέπουμε και τους συντελεστές των μεταβλητών

Predictor	Coefficient
const	10.2236
Age	-0.4013
PER	0.0919
FTr	-0.6858
DRB%	0.053
STL%	-0.1728
BLK%	-0.1054
DBPM	0.1904
VORP	0.8704
3P%	1.0636
2P%	2.2608
FT	0.0245

FT%	1.5812
AST	0.008
BLK	0.021

Πίνακας 8 : Συντελεστές Παλινδρόμησης

Από τον πίνακα βλέπουμε ότι ο μεγαλύτερος συντελεστής είναι ο συντελεστής του 2P%. Η τιμή 2.26 σημαίνει ότι μία αύξηση της μεταβλητής 2P% κατά μία μονάδα θα αυξήσει την μεταβλητή EFF κατά 2.26 μονάδες. Αυτό όμως δεν κάνει αυτόματα την μεταβλητή 2P% ισχυρή για την επίδραση στην EFF καθώς η μεταβλητή 2P% μετριέται σε ποσοστό τις εκατό(%) και η μέση τιμή της είναι 0.43 με τυπική απόκλιση 0.11.

Εφόσον οι συντελεστές παλινδρόμησης $b_1, b_2 \dots b_n$ μετριοούνται με διαφορετικές μονάδες μέτρησης (για παράδειγμα εδώ 2P% μετριέται σε ποσοστό επί της εκατό) , η άμεση σύγκριση είναι δύσκολη καθώς ένας μικρός συντελεστής μπορεί στην πραγματικότητα να είναι πιο σημαντικός από έναν μεγαλύτερο. Για να αντιμετωπιστεί αυτό το πρόβλημα χρησιμοποιούμε τους τυποποιημένους συντελεστές παλινδρόμησης (standardized regression coefficients) (Andrew Siegel, 2017)

Οι τυποποιημένοι συντελεστές παλινδρόμησης υπολογίζονται με δύο τρόπους :

a) Με εφαρμογή τις γραμμικής παλινδρόμησης σε τυποποιημένα δεδομένα, δηλαδή στα αρχικά δεδομένα να μετασχηματιστούν σε δεδομένα ώστε να έχουν μέσο = 0 και τυπική απόκλιση = 1
Η κάθε τιμή μετασχηματίζεται με τον τύπο:

$$x_s = \frac{x - mean}{sd}$$

Όπου

X_s : η μετασχηματισμένη τιμή της μεταβλητής X_i

X : η αρχική τιμή της μεταβλητής X_i

Mean , sd : Η μέση τιμή και η τυπική απόκλιση της μεταβλητής X_i

Για παράδειγμα, η τιμή 24 της μεταβλητής Age θα μετασχηματιστεί σε 0.73

b) Με υπολογισμό των συντελεστών παλινδρόμησης στα αρχικά δεδομένα και υπολογισμό των νέων συντελεστών με τον τύπο:

$$BETA = b \frac{sd X_i}{s d Y}$$

BETA: Ο τυποποιημένος συντελεστής παλινδρόμησης

b : Ο αρχικός συντελεστής παλινδρόμησης

$sd X_i$: Η τυπική απόκλιση των τιμών τις μεταβλητής X_i

$sd Y$: Η τυπική απόκλιση των τιμών τις μεταβλητής Y .

Οι τυποποιημένοι συντελεστές παλινδρόμησης:

Predictor	Coefficient
const	0.0016
Age	-0.1385
PER	0.1177
FTr	-0.0342
DRB%	0.0636
STL%	-0.039
BLK%	-0.0366
DBPM	0.0867
VORP	0.1418
3P%	0.0364
2P%	0.0494
FT	0.3778
FT%	0.0603
AST	0.1646
BLK	0.1304

Πίνακας 9 :Τυποποιημένοι Συντελεστές Παλινδρόμησης

Από τους τυποποιημένους συντελεστές βλέπουμε ότι την μεγαλύτερη τιμή την έχει η μεταβλητή FT με τιμή 0.37. Μια τιμή βήτα 0.37 υποδεικνύει ότι μια αλλαγή μιας τυπικής απόκλισης στην ανεξάρτητη μεταβλητή FT έχει ως αποτέλεσμα μια αύξηση των τυπικών αποκλίσεων κατά 0.37 τυπικών αποκλίσεων στην εξαρτημένη μεταβλητή EFF.

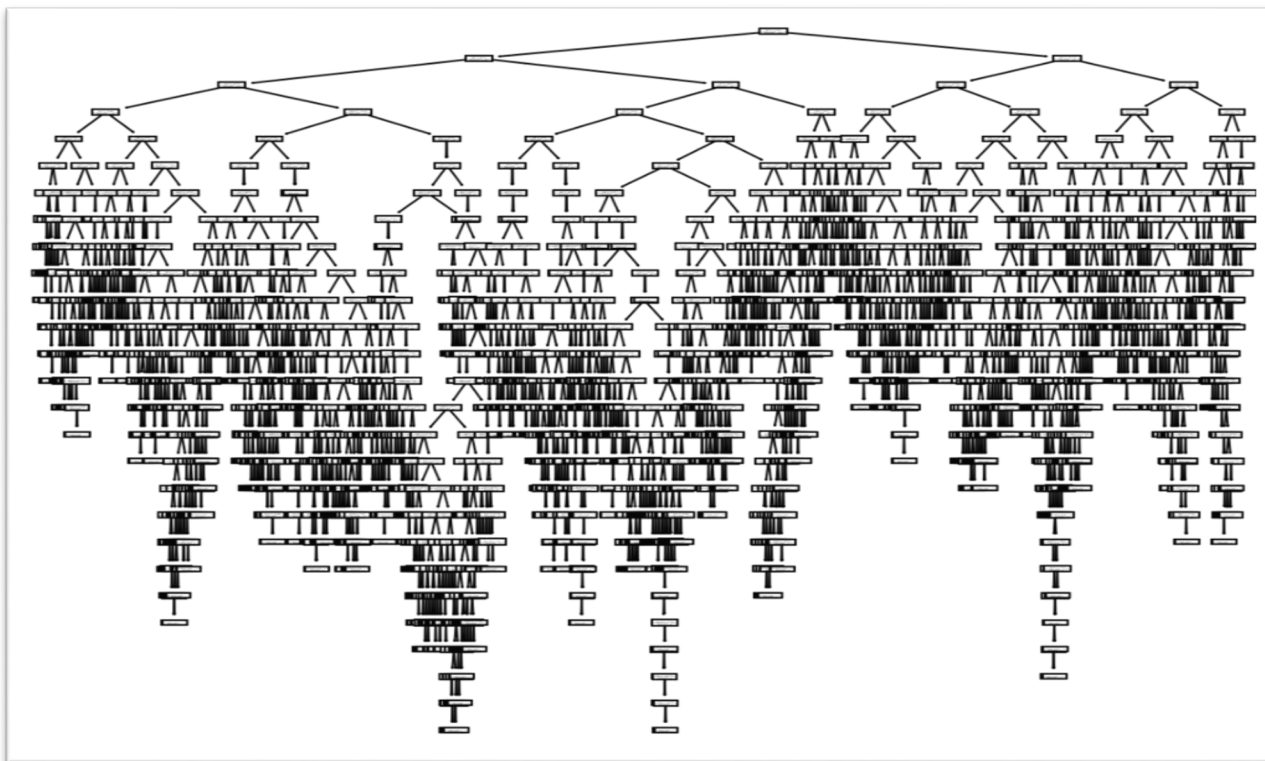
Όπως είδαμε , η μεταβλητή FT δηλαδή το πόσες ελεύθερες βολές πέτυχε ένα παίκτης τις δύο πρώτες σεζόν του στο NBA έχει την μεγαλύτερη επίδραση στον συνολικό career EFF, ενώ την μεγαλύτερη αρνητική επίδραση έχει η μεταβλητή AGE με beta -0.13 που σημαίνει ότι όσο πιο μεγάλος αγωνίζεται ένας παίκτης για πρώτη φορά στο NBA τόσο μικρότερη είναι η εξαρτημένη μεταβλητή EFF.

3.4.3 Εφαρμογή Δέντρων Αποφάσεων

Στο κεφάλαιο αυτό θα πραγματοποιηθεί εφαρμογή μεθόδου Δέντρων Αποφάσεων στα δεδομένα.

Όπως και με την γραμμική παλινδρόμηση, έτσι και στα δέντρα αποφάσεων, θα χρησιμοποιήσουμε τα δεδομένα εκπαίδευσης για να δημιουργήσουμε το μοντέλο πρόβλεψης και στη συνέχεια θα το αξιολογήσουμε με βάση την απόδοσή του στα δεδομένα εκπαίδευσης.

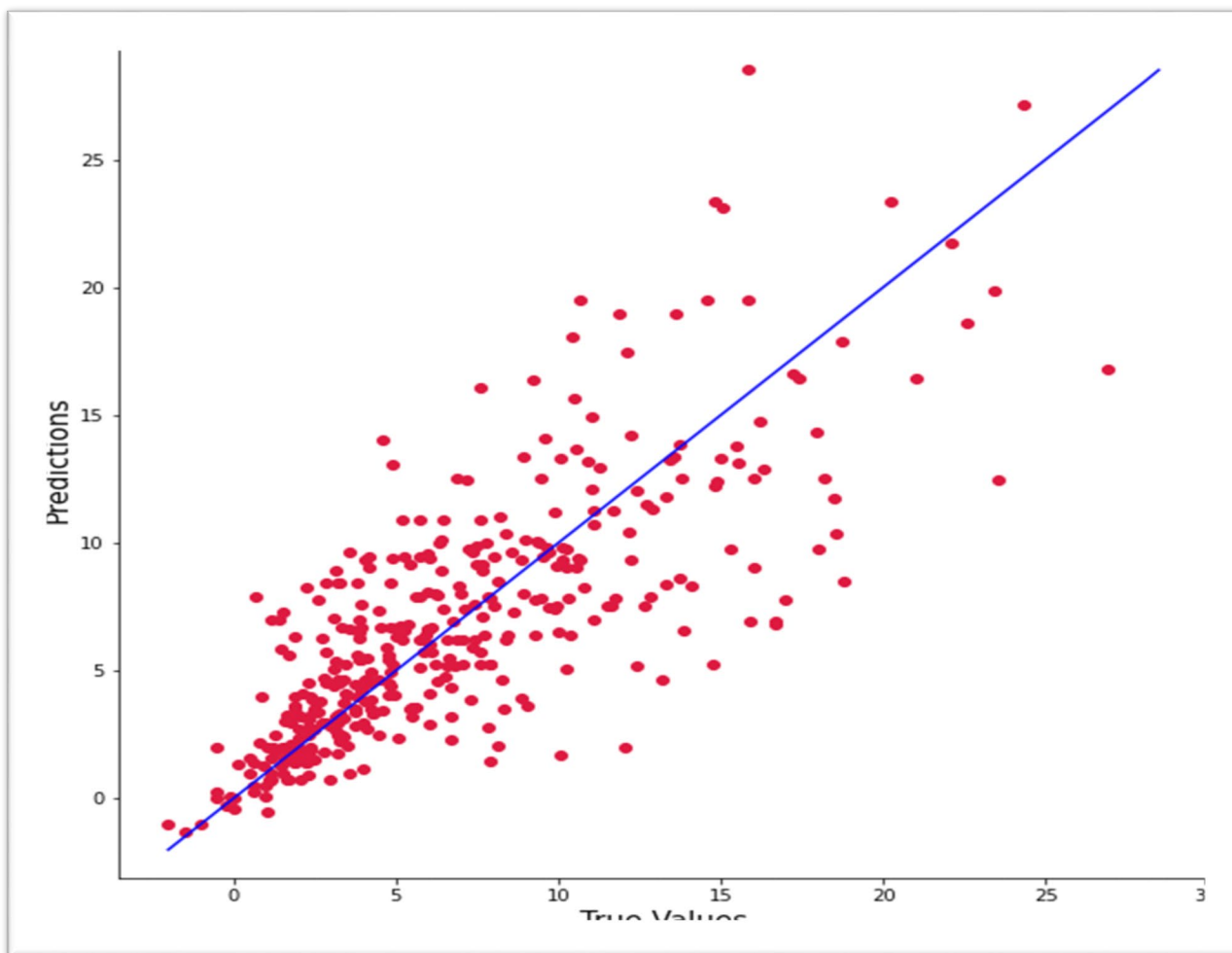
Το δέντρο που δημιουργήθηκε έχει 3069 κόμβους, το βάθος του δέντρου , δηλαδή το μήκος της μεγαλύτερης διαδρομής από μια ρίζα σε ένα φύλλο είναι 26 και μια προσπάθεια απεικόνισής του, φαίνεται παρακάτω.



Εικόνα 16: Αρχικό Δέντρο

Ένα πρόβλημα με τα μεγάλα σε μέγεθος δέντρα αποφάσεων, είναι η υπερπροσαρμογή (overfitting) στα δεδομένα εκπαίδευσης.

Πράγματι, το δέντρο έχει προσαρμοστεί τέλεια στα δεδομένα εκπαίδευσης, με μηδενικά σφάλματα και $R^2=1$. Αντίθετα σε δοκιμή στα δεδομένα ελέγχου, οι μετρικές σφάλματος είναι πάρα πολύ αυξημένες, $MAE= 2.34, MSE=11.15, RMSE=3.34$, κάτι που φαίνεται στο διάγραμμα σύγκρισης των προβλεφθέντων τιμών έναντι στις κανονικές.



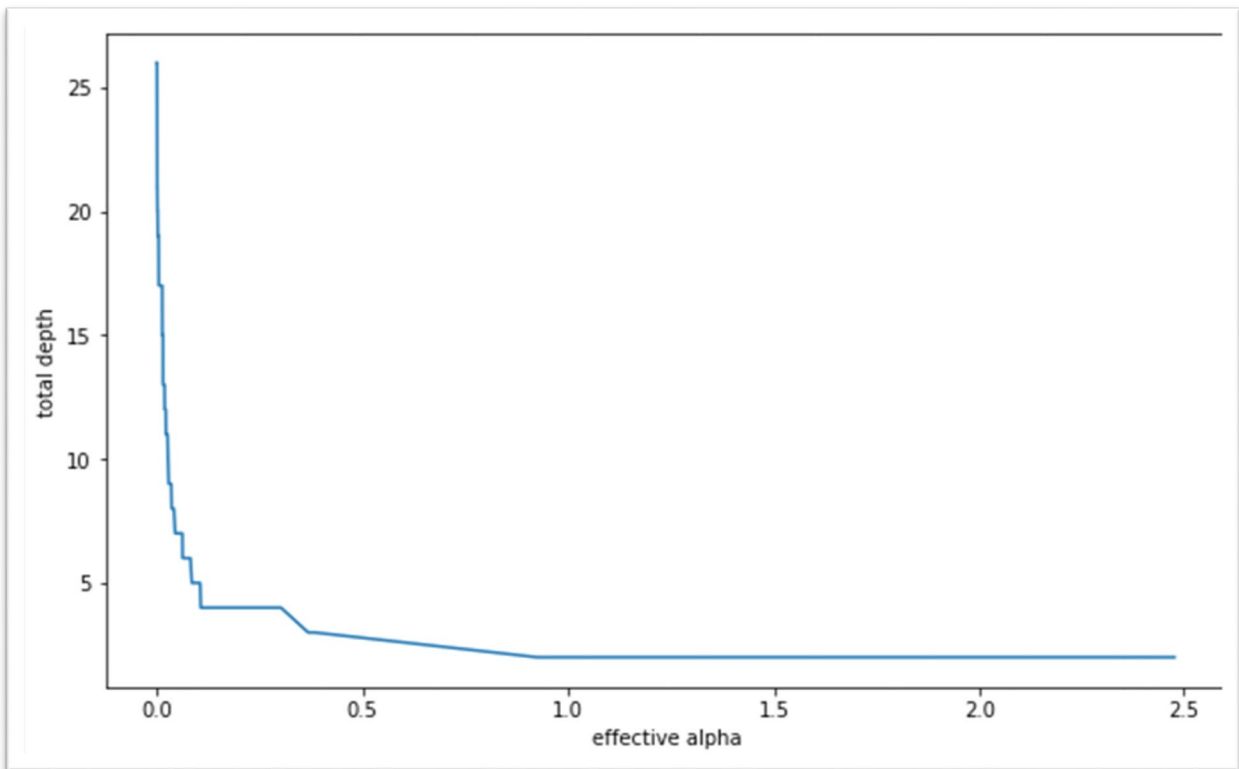
Εικόνα 17: Πραγματικές-Προβλεθείσες Τιμές

Για να μειωθεί το overfitting θα πρέπει να μειωθεί το μέγεθος του δέντρου. Το μέγεθος του δέντρου μειώνεται με μία διαδικασία που ονομάζεται pruning. (Gareth M. James, 2013) . Καθώς το μοντέλο μας ενδιαφέρει να αποδίδει καλά σε καινούρια δεδομένα θα χρησιμοποιήσουμε Cost Complexity Pruning για να βελτιώσουμε τα σφάλματα στα δεδομένα ελέγχου.(Sarthak Arora, 2020). Ο αλγόριθμος παραμετροποιείται από μία μη-αρνητική παράμετρο που ονομάζεται α (παράμετρος πολυπλοκότητας). Το α καθορίζει το μέγεθος των δέντρων που προέρχονται από το αρχικό δέντρο. Για κάθε τιμή του α υπάρχει ένα δέντρο T που είναι κομμάτι του αρχικού δέντρου T_0 για το οποίο η σχέση:

$$\sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

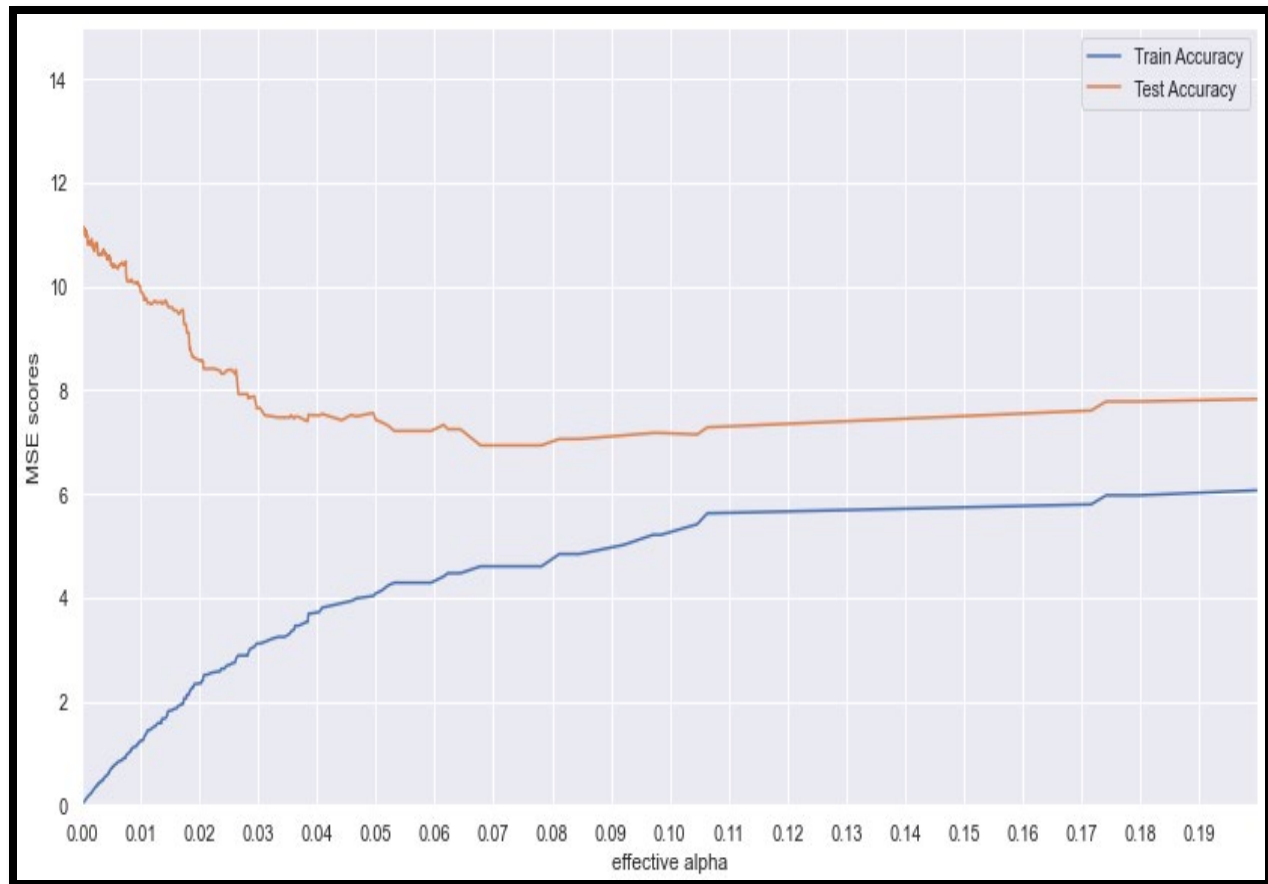
Όπου $|T|$ είναι ο αριθμός των τερματικών κόμβους του δέντρου T , R_m είναι το υποσύνολο του χώρου πρόβλεψης του τερματικού κόμβου που βρίσκεται στη θέση m , $\hat{y}_{Rm}$ είναι η προβλεπόμενη τιμή που συσχετίζεται με το υποσύνολο R_m . Σε τιμή του $\alpha=0$ το δέντρο είναι το αρχικό δέντρο αλλά όσο καθώς το α αυξάνεται, αυξάνεται και η ποινή για την πολυπλοκότητα οπότε τα δέντρα που επιλέγονται για ελαχιστοποιούν την παραπάνω σχέση είναι όλο και μικρότερα, και όσο οι τιμές του α προσεγγίζουν το δέντρο μένει μόνο με το αρχικό κόμβο. (Gareth M. James, 2013)

Συγκεκριμένα στο δικά μας δεδομένα το βάθος του δέντρου ανάλογα με την τιμή του α φαίνεται στο παρακάτω διάγραμμα.



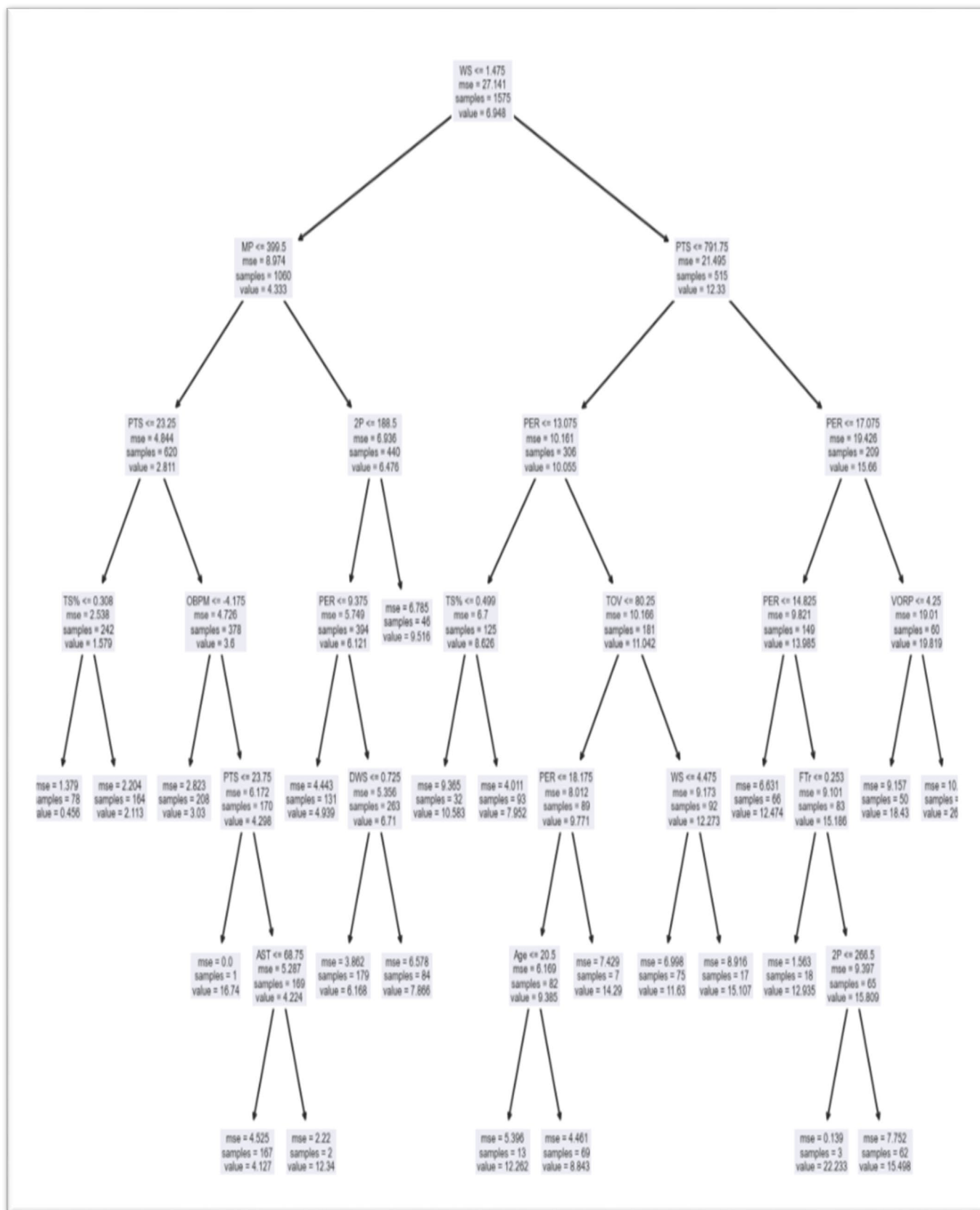
Εικόνα 18:Depth-alpha

Και οι τιμές των σφαλμάτων στα δεδομένα εκπαίδευσης και ελέγχου φαίνονται στο παρακάτω διάγραμμα.



Εικόνα 19:MSE-alpha

Οπότε βλέπουμε ότι με $\alpha=0.07$ παίρνουμε το μικρότερο σφάλμα. Δηλαδή ένα δέντρο με βάθος = 6 και αριθμό κόμβων = 45. Μία απεικόνιση του δέντρου φαίνεται παρακάτω.



Εικόνα 20: Δέντρο μετά από CCP

Οι τιμές των σφαλμάτων στα δεδομένα ελέγχου είναι φανερά βελτιωμένες συγκριτικά με τα τιμές των σφαλμάτων στο αρχικό δέντρο :

Mean Absolute Error: 1.8677175262644574

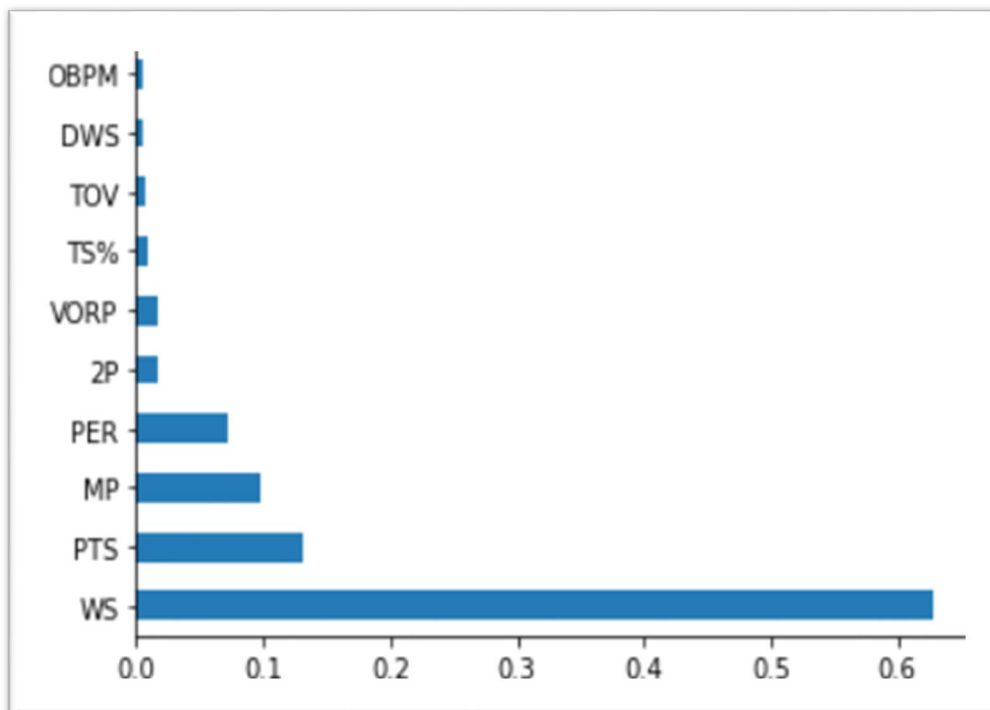
Mean Squared Error: 6.929778735567729

Root Mean Squared Error: 2.6324472901784244

Το δέντρο αποδίδει καλύτερα μετά την ολοκλήρωση της διαδικασίας pruning.

Ενώ όσον αφορά σημαντικότητα των μεταβλητών του δέντρου. Η σημαντικότητα των μεταβλητών στα δέντρα αποφάσεων ορίζεται ως η μείωση την impurity ενός κόμβου σταθμισμένη με την πιθανότητα ενός δείγματος να φτάσει σε αυτόν τον κόμβο. Υπάρχουν διάφορες τεχνικές μέτρησης του impurity ενός κόμβου, όπως το gini index (classification) ή η μείωση της διακύμανσης (regression). (Stacey Ronaghan, 2018) (machinelearninginterview.com)

Στο δέντρο μετά την διαδικασία pruning οι 10 σημαντικότερες μεταβλητές φαίνονται παρακάτω.



Εικόνα 21: Feature Importance Decision Tree

Οπότε για την πρόβλεψη της μεταβλητής EFF μέσω δέντρων αποφάσεων βλέπουμε ότι η μεταβλητή WS δηλαδή η συνεισφορά στις νίκες της ομάδας κατά τα πρώτα δύο χρόνια παίζει μεγάλη σημασία για την

πρόβλεψη της EFF ενός αθλητή, ενώ σημαντικοί είναι και οι πόνοι που πέτυχε, τα λεπτά που αγωνίστηκε και η αποδοτικότητα του.

3.4.4 Εφαρμογή Gradient Tree Boosting

Ένας ακόμα τρόπος να προβλέψουμε την μεταβλητή EFF, είναι ο αλγόριθμος Gradient Boosting. Ο φιλοσοφία πίσω από τους boosting αλγορίθμους είναι ο συνδυασμός πολλών αδύναμων μοντέλων έτσι ώστε το κάθε καινούριο δέντρο να διορθώνει τα σφάλματα των προηγούμενων, με σκοπό την δημιουργία ενός καλύτερου μοντέλου πρόβλεψης. Στον αλγόριθμο Gradient Tree Boosting η διαδικασία ξεκινάει με την δημιουργία του αρχικού μοντέλου, που είναι στην ουσία ένα δέντρο με ένα τερματικό κόμβο. Στη συνέχεια ο αλγόριθμος υπολογίζει τα ψευδο-υπολείματα και ταιριάζει ένα δέντρο σε αυτά, σχηματίζοντας τις τερματικές περιοχές. Στην συνέχεια το καινούριο δέντρο συμπληρώνει το παλιό και η διαδικασία επαναλαμβάνεται μέχρι το τέλος των επαναλήψεων. (Hastie et al, 2009)(Jerome Friedman 1999)

1. Initialize $f_0(x) = \arg \min_{\gamma} \sum_{i=1}^N L(y_i, \gamma)$.

2. For $m = 1$ to M :

(a) For $i = 1, 2, \dots, N$ compute
$$r_{im} = - \left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)} \right]_{f=f_{m-1}}.$$

(b) Fit a regression tree to the targets r_{im} giving terminal regions $R_{jm}, j = 1, 2, \dots, J_m$.

(c) For $j = 1, 2, \dots, J_m$ compute
$$\gamma_{jm} = \arg \min_{\gamma} \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma).$$

(d) Update $f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$.

3. Output $\hat{f}(x) = f_M(x)$.

Εικόνα 22: Gradient Tree Boosting Algorithm (Hastie et al, 2009)

Στο σύνολο δεδομένων μας ο αλγόριθμος μας δίνει τα καλύτερα αποτελέσματα ως προς τα σφάλματα καθώς οι τιμές των σφαλμάτων στα δεδομένα ελέγχου είναι οι εξής:

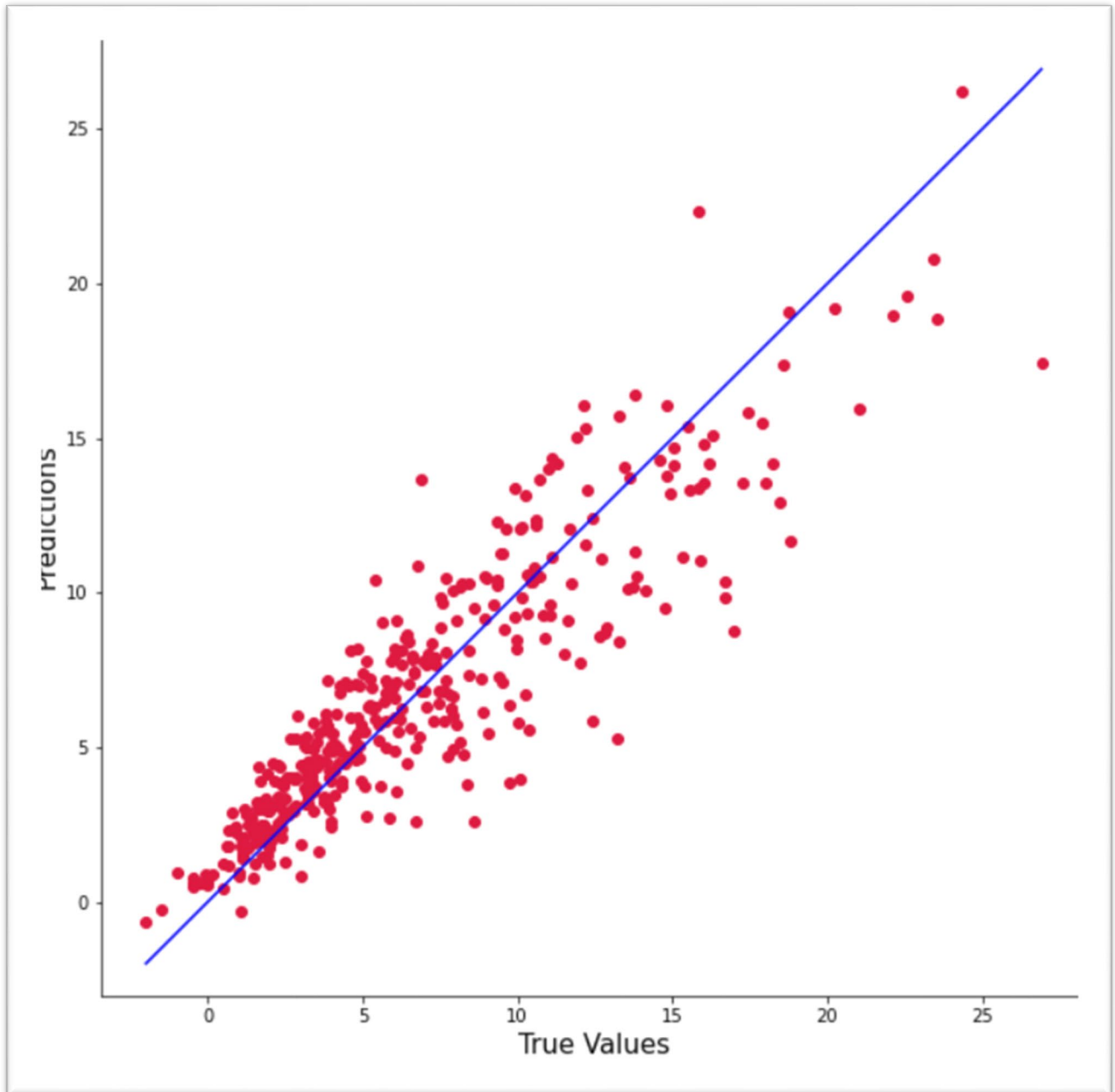
Mean Absolute Error: 1.64

62

Mean Squared Error: 4.9

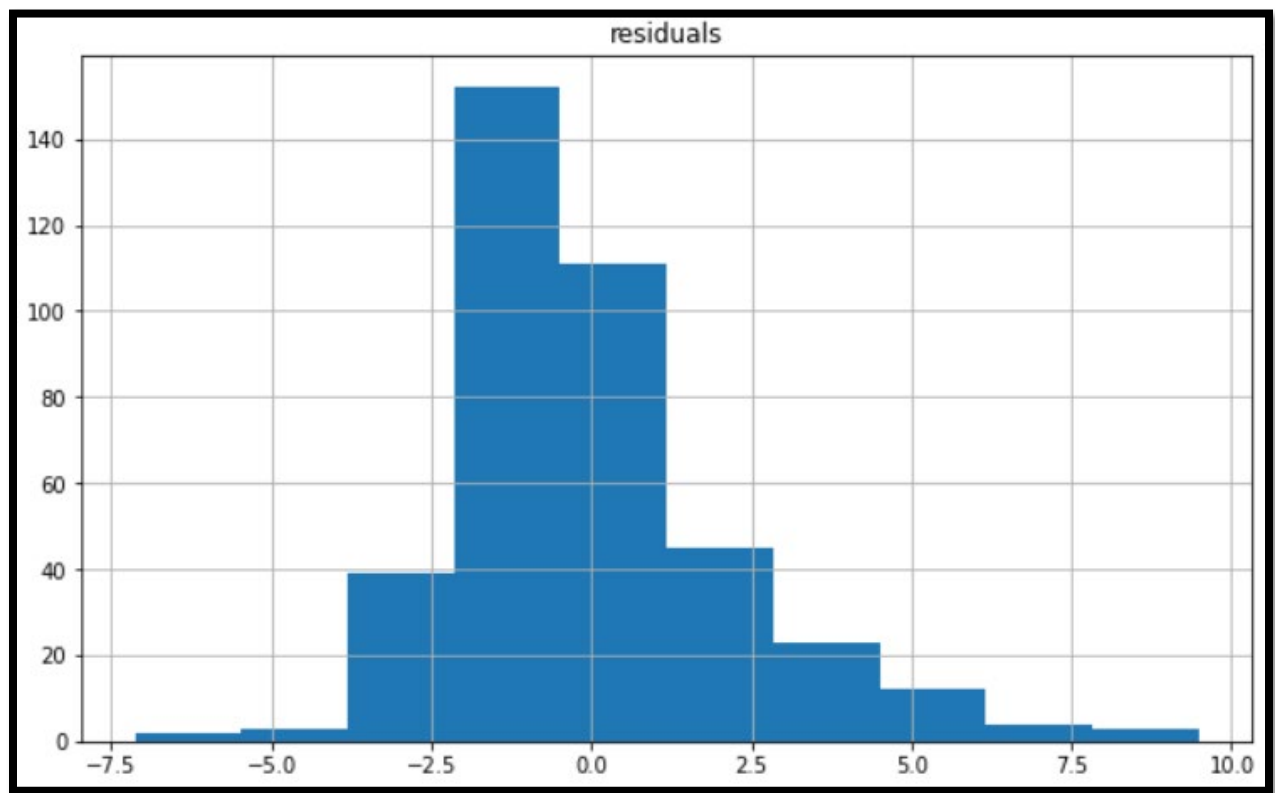
Root Mean Squared Error: 2.22

Το διάγραμμα των πραγματικών τιμών έναντι των προβλεθέντων όπως φαίνεται παρακάτω.



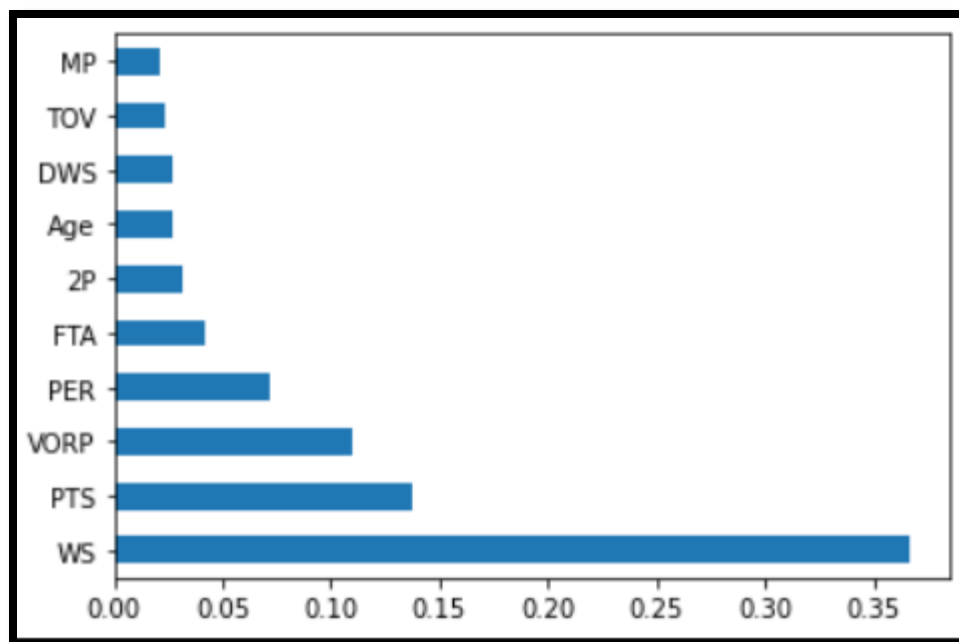
Εικόνα 23: Predicted-True Values(Gradient Boosting)

Ενώ και η κατανομή των καταλοίπων προσεγγίζει την κανονική.



Εικόνα 24: Κατάλοιπα Gradient Boosting

Όσον αφορά την σημαντικότητα των μεταβλητών, στην περίπτωση του gradient tree boosting, υπολογίζεται για κάθε μεμονωμένο δέντρο απόφασης και στη συνέχεια υπολογίζεται από τον μέσο όρο σε όλα τα δέντρα του μοντέλου.



Εικόνα 25: Feature Importance Gradient Boosting

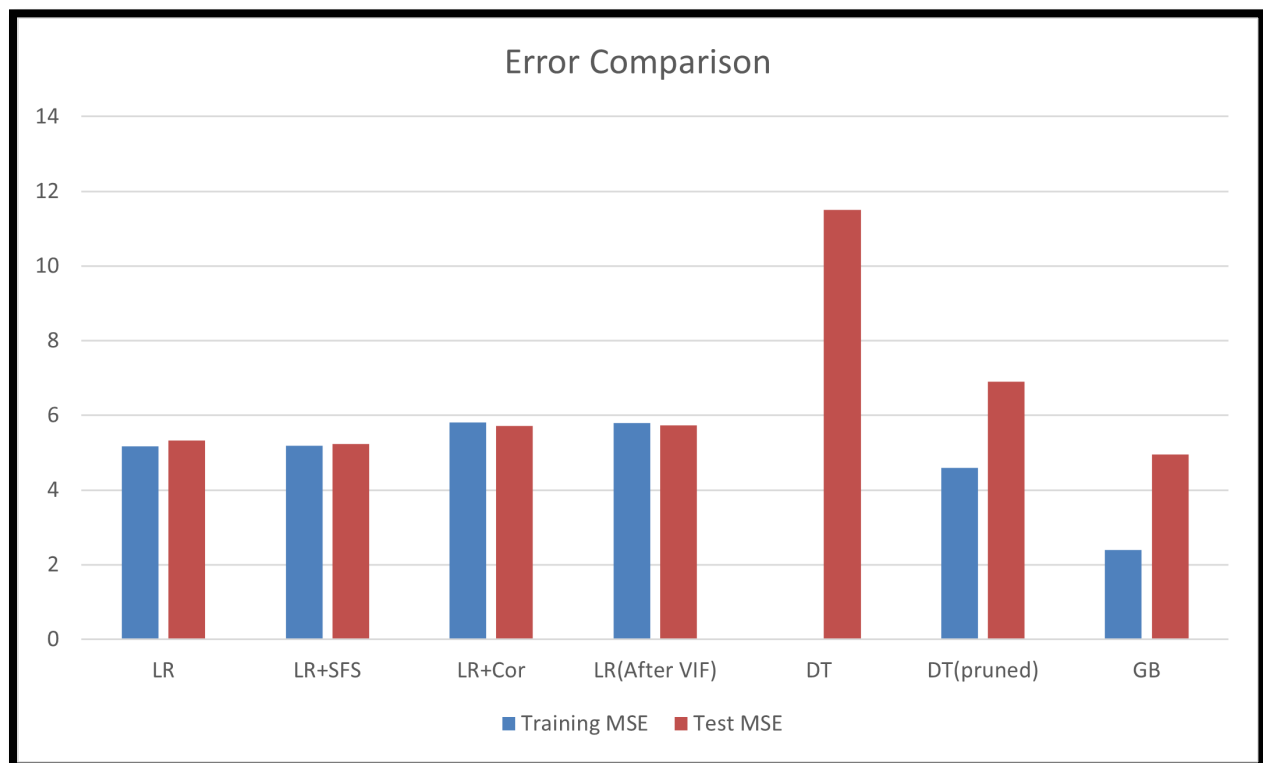
Η μεταβλητή WS παραμένει η πιο σημαντική όπως και η PTS αλλά βλέπουμε ότι τρίτη πιο σημαντική είναι η μεταβλητή VORP, δηλαδή η αξία ενός παίκτη σε σχέση με τις εναλλακτικές επιλογές (value over replacement player).

4. Συμπεράσματα

Στην εργασία αναφέρθηκαν παραδείγματα για το πως οι ομάδες του NBA, εκμεταλλεύονται τα δεδομένα για να αποκτήσουν χρήσιμα συμπεράσματα, μέσω εφαρμογής μεθόδων ανάλυσης δεδομένων και μηχανικής μάθησης. Από την υπάρχουσα βιβλιογραφία παρουσιάστηκαν ήδη αρκετοί τρόποι που μία ομάδα θα αποκομίσει οφέλη από εφαρμογή μεθόδων μηχανικής μάθησης, καθώς και προβλήματα που προκύπτουν από την ένταξη της μηχανικής μάθησης στο NBA. Από την SWOT ανάλυση παρατηρήθηκε ότι τα πλεονεκτήματα είναι πολύ περισσότερα. Στη συνέχεια παρουσιάστηκε ένα παράδειγμα του πως μία ομάδα μπορεί να αποκτήσει χρήσιμα συμπεράσματα μέσω της εφαρμογής μηχανικής μάθησης σε δεδομένα, συγκεκριμένα έγινε η προσπάθεια πρόβλεψης της απόδοσης ενός παίκτη καθ'ολη την διάρκεια της καριέρας του, μέσω πρόβλεψης της μεταβλητής EFF. Τα μοντέλα που χρησιμοποιήθηκαν για να γίνει η πρόβλεψη αυτή, είναι η πολλαπλή γραμμική παλινδρόμηση, τα δέντρα αποφάσεων και τα gradient boosted trees. Στη συνέχεια πραγματοποιήθηκαν εφαρμογές των μοντέλων με κάποιες παραλλαγές, είτε επιλογή συντελεστών για την γραμμική παλινδρόμηση, είτε pruning για τα δέντρα αποφάσεων. Το μέσω τετραγωνισμένο σφάλμα για τους αλγορίθμους φαίνεται στον παρακάτω πίνακα και.

	Linear Regression	Linear Regression +Sequential Forward Selection	LR+ Correlation Feature Selection	LR(After VIF)	Decision Tree	DT(pruned)	Gradient Boosting
Training MSE	5.17	5.19	5.81	5.8	0	4.59	2.4
Test MSE	5.32	5.23	5.72	5.74	11.5	6.9	4.95

Πίνακας 10: Training and Test Errors



Εικόνα 26 : Error Comparion

Ο αλγόριθμος που αποδίδει καλύτερα, τόσο στα δεδομένα εκπαίδευσης όσο και ελέγχου, είναι ο gradient boosting trees με τιμές MSE 2.4 και 4.95 σε training και test δεδομένα αντίστοιχα. Όσον αφορά τις τιμές των σφαλμάτων στους αλγορίθμους γραμμικής παλινδρόμησης, το μοντέλο στην αρχική του μορφή έδωσε τα καλύτερα αποτελέσματα όσον αφορά τις προβλέψεις σε test δεδομένα. Παρόλα αυτά, για να μπορέσουμε να ερμηνεύσουμε την σημαντικότητα των μεταβλητών έπρεπε να γίνει εξάλειψη της πολυσυγγραμμικότητας που παρουσιάστηκε στο σύνολο δεδομένων, καθώς πολλές ανεξάρτητες μεταβλητές είχαν συσχέτιση μεταξύ τους. Η επιλογή μεταβλητών με σκοπό την μείωση της

πολυσυγγραμμικότητας δεν βοήθησε στην βελτίωση της απόδοσης. Από την γραμμική παλινδρόμηση είδαμε ότι η μεταβλητή FT δηλαδή το πόσες ελεύθερες βολές πέτυχε ένα παίκτης τις δύο πρώτες σεζόν του στο NBA έχει την μεγαλύτερη επίδραση στον συνολικό career EFF, ενώ την μεγαλύτερη αρνητική επίδραση έχει η μεταβλητή AGE.

Αναφορικά με τα μοντέλα δέντρων που χρησιμοποιήθηκαν, το απλό decision tree, παρουσίασε σημαντικό πρόβλημα overfitting καθώς προσαρμόστηκε τέλεια στα δεδομένα εκπαίδευσης αλλά αποδίδοντας κατά πολύ χειρότερα από τα υπόλοιπα μοντέλα στα δεδομένα ελέγχου. Το πρόβλημα αυτό λύθηκε μέσω της διαδικασίας cost complexity pruning αλλά τα test errors παρέμειναν υψηλότερα σε σχέση με τα υπόλοιπα μοντέλα. Με την δημιουργία ενός μοντέλου πολλαπλών δέντρων μέσω του αλγορίθμου Gradient Boosting, δημιουργήσαμε το μοντέλο που απέδιδε καλύτερα στα δεδομένα ελέγχου. Σύμφωνα με τους αλγόριθμους των δέντρων οι μεταβλητές που έχουν την μεγαλύτερη σημαντικότητα προς την πρόβλεψη είναι οι πόντοι που πέτυχε ένας αθλητής, τα δύο πρώτα χρόνια της καριέρας του καθώς και η συνεισφορά του στις νίκες.

Καθώς ο στόχος ήταν η ανάπτυξη ενός μοντέλου πρόβλεψης της μεταβλητής EFF, το μοντέλο της γραμμικής παλινδρόμησης καθώς και το μοντέλο Gradient Boosting μας δίνουν ικανοποιητικά αποτελέσματα για κάνει μία ομάδα μία εκτίμηση για την απόδοση του παίκτη στην καριέρα του. Η ύπαρξη και άλλων παραγόντων ως προς την ανάπτυξη ενός αθλητή, κάνει την πρόβλεψη ακόμα πιο δύσκολη. Παρόλα αυτά τα μοντέλα προσφέρουν μία ικανοποιητική εκτίμηση.

Τέλος, θα δούμε πως θα διαμορφωθεί η μεταβλητή EFF καριέρας για τον αθλητή Luka Doncic σύμφωνα με τα μοντέλα.

LR	LR+SFS	LR+Cor	LR(After VIF)	DT	DT(pruned)	GB
25.5	25.5	25.36	25.75	18.1	26.7	26.42

Πίνακας 11: EFF Luka Doncic

Το μοντέλο GB προβλέπει ότι η EFF καριέρας του Luka Doncic, βάση των στατιστικών του τις δύο πρώτες χρονιές θα είναι 26.4, δηλαδή σε σύγκριση με του παίκτες στο υπάρχων dataset θα είναι μεταξύ Shaquille O'Neal (EFF= 25.6) και Chris Paul (26.6)

Βιβλιογραφία

Adam Fromal (January 26 2012) Understanding the NBA: Explaining Advanced Offensive Stats and Metrics. bleacherreport.com. <https://bleacherreport.com/articles/1039116-understanding-the-nba-explaining-advanced-offensive-stats-and-metrics>

Akhil Nistala, John Guttag (2018) Using Deep Learning to Understand Patterns of Player Movement in the NBA. Retrieved from <https://dspace.mit.edu/handle/1721.1/119728>

Andrew F. Siegel (2016), Practical Business Statistics (Seventh Edition), Chapter 12 - Multiple Regression: Predicting One Variable From Several Others, Pages 355-418, ISBN 9780128042502, <https://doi.org/10.1016/B978-0-12-804250-2.00012-2>.

Arvydas Sabonis wikipedia.org. https://en.wikipedia.org/wiki/Arvydas_Sabonis

Bob Hayes (2015) Can Analytics Improve your Game? businessoverbroadway.com <https://businessoverbroadway.com/2015/03/22/can-analytics-improve-your-game/>

Brian Kamenetzky (2014) The Next Big Thing In Sports Data: Predicting (And Avoiding) Injuries. fastcompany.com. <https://www.fastcompany.com/3034655/the-next-big-thing-in-sports-data-predicting-and-avoiding-injuries>

Contract Types. cbabreakdown.com. <https://cbabreakdown.com/contract-types>

Darren Heitner (2016) The NBA's Six Year, \$250 Million Data Deal. forbes.com.
<https://www.forbes.com/sites/darrenheitner/2016/09/22/the-nbas-six-year-250-million-data-deal/?sh=1fdbaa7c481d>

Efficiency (basketball). wikipedia.org. [https://en.wikipedia.org/wiki/Efficiency_\(basketball\)](https://en.wikipedia.org/wiki/Efficiency_(basketball))

Gambling in the United States. wikipedia.com.
https://en.wikipedia.org/wiki/Gambling_in_the_United_States

Gareth M. James, Daniela Witten, Trevor Hastie, Robert Tibshiran (2013), An Introduction to Statistical Learning: With Applications in R, Chapter 8 : Tree-Based Methods
Page 303-335
DOI 10.1007/978-1-4614-7138-7 1

Gerasimos Plegas (2021) Can a Data Scientist Replace an NBA Scout? ML App Development for Best Transfer Suggestion. towardsdatascience.com <https://towardsdatascience.com/can-a-data-scientist-replace-a-nba-scout-ml-app-development-for-best-transfer-suggestion-f07066c2773>

Hastie, T., Tibshirani, R., Friedman, J. (2009). Boosting and Additive Trees. In: The Elements of Statistical Learning. Springer Series in Statistics. Springer. Chapter 10. pp 337–387
https://doi.org/10.1007/978-0-387-84858-7_10

How Data Played a Role in the Toronto Raptors' NBA Championship (2019). brainstation.io.
<https://brainstation.io/blog/how-data-played-a-role-in-the-toronto-raptors-nba-championship>

Jabari Young (2021), NBA projects \$10 billion in revenue as audiences return after Covid, but TV viewership is a big question. cnbc.com . <https://www.cnbc.com/2021/10/18/nba-2021-2022-season-10-billion-revenue-tv-viewership-rebound.html>

Jason Dachman (July 26, 2012). STATS' SportVU Continues To Make NBA Inroads; Broadcast TV Market Up Next. .sportsvideo.org. <https://www.sportsvideo.org/2012/07/26/stats-sportvu-continues-to-make-nba-inroads-broadcast-tv-market-up-next/>

J.R. Quinlan, (1987) Simplifying decision trees, International Journal of Man-Machine Studies,

Volume 27, Issue 3, Pages 221-234,
ISSN 0020-7373, [https://doi.org/10.1016/S0020-7373\(87\)80053-6](https://doi.org/10.1016/S0020-7373(87)80053-6).

Jerome Friedman (1999) . Stochastic Gradient Boosting

Jianyu Miao, Lingfeng Niu, (2019), A Survey on Feature Selection, Procedia Computer Science, Volume 91, 2016, Pages 919-926, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2016.07.111>.

Kirk Goldsberry (May, 2, 2019). How Mapping Shots In The NBA Changed It Forever.
fivethirtyeight.com. <https://fivethirtyeight.com/features/how-mapping-shots-in-the-nba-changed-it-forever/>

Learning Feature Importance from Decision Trees and Random Forests (2021) Learning Feature Importance from Decision Trees and Random Forests. machinelearninginterview.com.
<https://machinelearninginterview.com/topics/machine-learning/learning-feature-importance-from-decision-trees-and-random-forests/>

National Basketball Association teams ranked by revenue 2020/21 season. Statista.
<https://www.statista.com/statistics/193704/revenue-of-national-basketball-association-teams-in-2010/>

Pascal Courty, Luke Davey (2019) The Impact of Variable Pricing, Dynamic Pricing, and Sponsored Secondary Markets in Major League Baseball. Journal of Sports Economics. 21. 152700251986736. DOI: 10.1177/1527002519867367.

Sutton, William & Urban, Timothy & Troilo, Michael & Bouchet, Adrien & Mondello, Michael. (2020). Business analytics, revenue management, and sport: evidence from the field. International Journal of Revenue Management. 11. 1. DOI: 10.1504/IJRM.2020.10032774.

Meta S. Brown (2016). When and Where To Buy Consumer Data (And 12 Companies Who Sell It) forbes.com. <https://www.forbes.com/sites/metabrown/2015/09/30/when-and-where-to-buy-consumer-data-and-12-companies-who-sell-it/?sh=c9de57132851>

Let's Solve Overfitting! Quick Guide to Cost Complexity Pruning of Decision Trees (2020), analyticsvidhya.com. <https://www.analyticsvidhya.com/blog/2020/10/cost-complexity-pruning-decision-trees/#:~:text=It%20reduces%20the%20size%20of,of%20Pruning%20of%20Decision%20Trees.>

Nagesh Singh Chauhan (2022) Decision Tree Algorithm, Explained. kdnuggets.com.
<https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html>

Reinforcement learning. geeksforgeeks.org <https://www.geeksforgeeks.org/what-is-reinforcement-learning/>

Supervised Learning. ibm.com. <https://www.ibm.com/cloud/learn/supervised-learning>

Stephanie Glen (2019) Decision Tree vs Random Forest vs Gradient Boosting Machines: Explained Simply. datasciencecentral.com. <https://www.datasciencecentral.com/decision-tree-vs-random-forest-vs-boosted-trees-explained/>

Stacey Ronaghan (2018). The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark. towardsdatascience.com.
<https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3#:~:text=Feature%20importance%20is%20calculated%20as,the%20more%20important%20the%20feature.>

Stepwise regression. wikipedia.org

https://en.wikipedia.org/wiki/Stepwise_regression#:~:text=In%20statistics%2C%20stepwise%20regression%20is,based%20on%20some%20prespecified%20criterion

Stepwise Regression. statisticshowto.com. <https://www.statisticshowto.com/stepwise-regression/>

Toronto Raptors Use Data-Driven Command Center on Path to NBA Finals (2019). ibm.com.

<https://www.ibm.com/blogs/think/2019/06/toronto-raptors-use-data-driven-command-center-on-path-to-nba-finals/>

Unsupervised Learning. ibm.com. <https://www.ibm.com/cloud/learn/unsupervised-learning>

Wangwei Wu (2010) Injury Analysis Based on Machine Learning in NBA Data. Journal of Data Analysis and Information Processing. 08. 295-308. DOI: 10.4236/jdaip.2020.84017.