Predykcja zarobków koszykarzy ligi NBA

Jędrzej Sarna

12 lutego 2024

1 Wprowadzenie

Współczesna koszykówka przeszła transformację, w której intuicja i talent ustępują miejsca szczegółowej analizie danych. Decyzje dotyczące strategii gry, wyboru zawodników do składu, a nawet ustalania zarobków graczy NBA są teraz oparte na precyzyjnych liczbowych analizach. Kluby NBA, wykorzystując zaawansowane technologie i algorytmy, coraz śmielej inwestują w badania danych, aby wyjść na prowadzenie w zaciętej rywalizacji sportowej. Odpowiednie manewrowanie dysponowanym budżetem przy ustalaniu zarobków zawodników umożliwia przeznaczenie większych nakładów finansowych w inne sektory klubu.

2 Sformułowanie problemu

W obliczu rosnącej roli danych w koszykówce, zdecydowałem, że chciałbym samodzielnie zgłębić świat analizy w tej dziedzinie sportu. Moim celem jest zrozumienie i wykorzystanie dostępnych statystyk, aby na ich podstawie dokonać własnych analiz oraz stworzyć model predykcyjny pozwalający przewidywać zarobki koszykarzy korzystając z zebranych statystyk meczowych oraz ogólnych informacji o zawodnikach.

2.1 Zbieranie danych

Dane na których będę opierał moją analize pochodzą ze strony (link). Zbiór danych składa się z informacji o koszykarzach pozyskanych z dwóch wiarygodnych stron internetowych - *hoopshype.com*, udostępniającej ogólne informację o zawodnikach oraz *basketball-reference.com*, posiadający szczegółowe statystyki meczowe zawodników z poprzednich sezonów. Dane dotyczą koszykarzy, którzy grali w lidze NBA w sezonie 2022/2023.

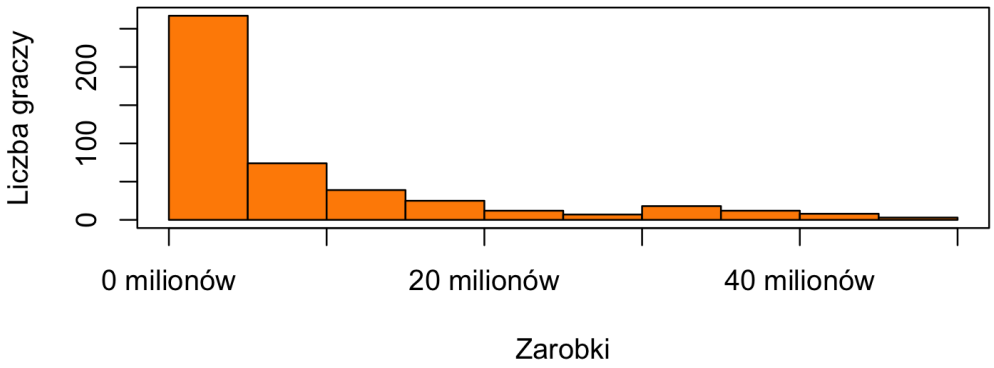
3 Eksploarcja i przygotowanie danych

3.1 Wstępna eksploracja danych

Po wgraniu danych do środowiska R można zauważyć, że zawiera on 467 wierszy, gdzie każdy odpowiada pojednczemu graczowi, oraz 32 kolumny opisujące wyniki sportowe poszczególnych zawodników.

Po wstępnym przejrzeniu danych można zauważyć, że w zbiorze danych występują dwie kolumny nie wnoszące żadnych informacji. Jest to kolumna przedstawiająca indeks obserwacji oraz kolumna odpowiadająca za 'skrótowce' imion oraz nazwisk zawodników. Przedstawione zmienne możemy zatem usunąć, doprowadzając do 30 kolumn w zbiorze danych. Kontynuując analizę, informacje w zbiorze można podzielić na dwie główne kategorie. Pierwsza (pierwsze 5 kolumn) obejmuje ogólne informacje o zawodnikach, takie jak nazwa obecnego klubu, pozycja na boisku, imię i nazwisko, wiek oraz aktualne zarobki w skali roku (w dolarach amerykańskich), będąca naszą zmienną objaśnianą. Druga część danych (pozostałe 25 kolumn) skupia się na statystykach meczowych z sezonu 2022/2023, zawierająca szczegółowe informacje dotyczące występów każdego zawodnika. Zbiór danych posiada wartości *NA* dla niektórych statystyk, rozwiążemy ten problem w dalszej części referatu.

Przyjrzymy się teraz bardziej szczegółowo zarobkom koszykarzy dostępnych w naszym zbiorze danych. Analizując pensję, można zauważyć, że istnieje dwóch zawodników, których zarobki wynoszą znaczniej mniej od reszty (5849 $). Opisani gracze, to juniorzy, którzy zagrali w jedynie jednym meczu w sezonie, zatem te obserwacje możemy usunąć, gdyż odstają od standarów ligi. Przedstawmy teraz graficznie rozkład zarobków w lidze NBA.



Rysunek 1: Histogram zarobków koszykarzy w zbiorze danych

Przedstawiony histogram dobrze ilustruje rzeczywistość świata koszykówki. W lidze NBA, chociaż ogólne zarobki są bardzo wysokie, faktycznie tylko niewielka część zawodników osiąga astronomiczne kwoty wypłat liczące w dziesiątkach milionów dolarów. To zazwyczaj gracze wybitni, posiadający wyjątkowe umiejętności, znaczące osiągnięcia lub ogromny potencjał, co sprawia, że są oni szczególnie cenni dla swoich klubów, zatem wymagają kontraktów na ogromne kwoty. Należy jednak wziąć pod uwagę fakt, że taki rozkład zmiennej objaśnianej może być problematyczny przy budowie modelu ze względu na małe ilości obserwacji przy dużych wartościach. W dalszej części referatu dokładniej przeanalizuję ten problem.

3.2 Szczegółowa analiza statystyk meczowych

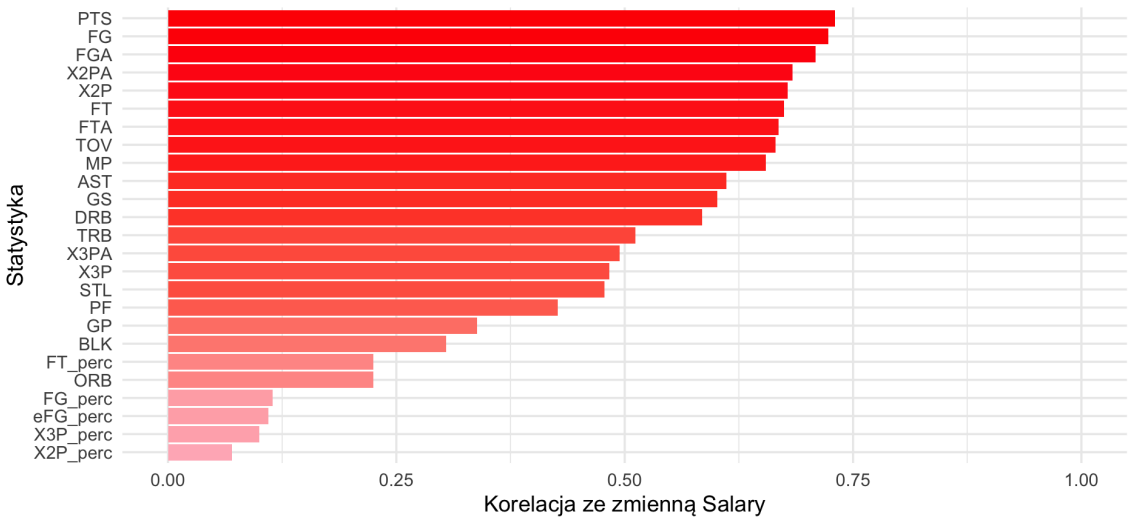
Zgromadzone statystyki charakteryzują się wysokim poziomem szczegółowości, oferując dane obejmujące różne aspekty meczu koszykówki. W poniższej tabeli zostały przedstawione metryki z jakimi mamy do czynienia.

Skrót	**Opis**	**Skrót**	**Opis**
GP	Gry Rozegrane	X3P	Rzuty za 3 Punkty (średnia na mecz)
GS	Gry rozpoczęte w pierwszym składzie	X3PA	Próby Rzutów za 3 Punkty (średnia na mecz)
MP	Minuty na Boisku (średnia na mecz)	X3P_perc	Procent Skuteczności Rzutów za 3 Punkty
FG	Rzuty z Gry (średnia na mecz)	X2P	Rzuty za 2 Punkty (średnia na mecz)
FGA	Próby Rzutów z Gry (średnia na mecz)	X2PA	Próby Rzutów za 2 Punkty (średnia na mecz)
FG_perc	Procent Skuteczności Rzutów z Gry	X2P_perc	Procent Skuteczności Rzutów za 2 Punkty
FT	Rzuty Wolne (średnia na mecz)	eFG_perc	Skuteczność Rzutów Efektywnych z Gry (*)
FTA	Próby Rzutów Wolnych (średnia na mecz)	FT_perc	Procent Skuteczności Rzutów Wolnych
ORB	Zbiórki Ofensywne (średnia na mecz)	TRB	Zbiórki Ogółem (średnia na mecz)
DRB	Zbiórki Defensywne (średnia na mecz)	AST	Asysty (średnia na mecz)
STL	Przechwyty (średnia na mecz)	BLK	Bloki (średnia na mecz)
TOV	Straty (średnia na mecz)	PF	Faule Osobiste (średnia na mecz)
PTS	Punkty (średnia na mecz)		

Tabela 1: Opisy statystyk koszykarskich ze zbioru danych

W zbiorze danych dla pewnych statystyk pojawiły się wartości *NA*, o czym już wspominałem. Te braki danych dotyczyły wskaźników skuteczności, gdzie liczba prób rzutów równała się zero, co logicznie uniemożliwiało wyliczenie skuteczności rzutów. W takich przypadkach, sensownym rozwiązaniem było zastępienie tych wartości zerem.

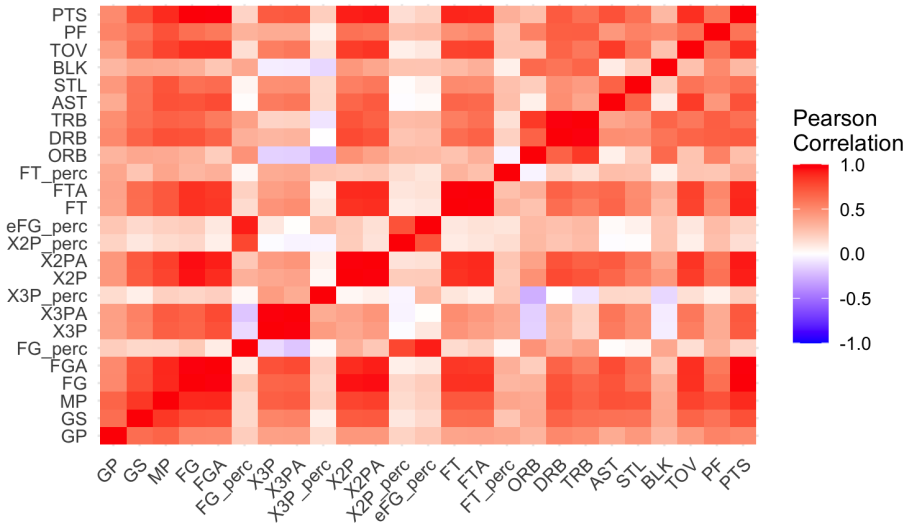
Chociaż zbiór zawiera dużą ilość statystyk, istotne jest zrozumienie, że nie wszystkie z tych danych w równym stopniu wpływają na zarobki zawodnika. W projekcie skoncentruję się na analizie i interpretacji tych metryk, aby zidentyfikować kluczowe wskaźniki, które mają największy wpływ na finansowe aspekty kariery koszykarzy. Rozpoczynając analizę zależności, pierwszym krokiem będzie przedstawienie wykresu korelacji, który zilustruje związki między zarobkami zawodników NBA, a statystykami meczowymi.



Rysunek 2: Korelacja poszczególnych statystyk meczowych z zarobkami zawodników

Biorąc pod uwagę korelacje na rysunku powyżej, moglibyśmy rozważyć skupienie się na tych najbardziej skorelowanych cechach w naszym modelu. Są to cechy takie jak "PTS", "FG", "FT", "MP". Rzeczywiście, wpływ tych zmiennych na finanse zawodników może być zgodny z intuicją, przykładowo zdobyte punkty w meczu bezpośrednio wpływają na wynik, który determinuje sukces klubu, a co za tym idzie z pewnością zarobki. Zaskakujący może być tak niski wpływ każdej z metryk dotyczących skuteczności na zarobki (5 z ostatnich 6 cech).

Należy jednak pamiętać, że jest tutaj przedstawiona jedynie liniowa zależność między zmiennymi. Dodatkowo, powinniśmy zachować ostrożność na współliniowość zmiennych, ponieważ może ona zniekształcać ich znaczenie oraz przewidywania modelu. Przykładowo, dla różnych typów rzutów, dysponujemy trzema zmiennymi charakteryzującymi każdy typ: liczbą rzutów celnych, liczbą prób rzutu oraz skutecznością. Te zmienne określają bardzo podobne zjawiska, więc możliwa jest pewna współliniowość. W celu analizy, przygotowałem macierz korelacji statystyk , która umożliwi wnikliwą analizę zależności potencjalnie problematycznych zmiennych opisanych powyżej jak i wszystkich innych relacji.



Rysunek 3: Macierz korelacji statystyk meczowych

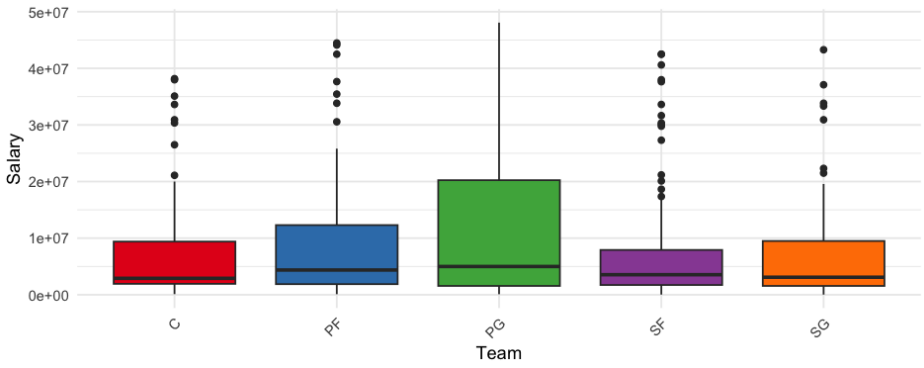
Wartości współczynników korelacji potwierdzają spekulację o współliniowości w przypadku niektórych zmiennych. Rzuty celne, odpowiadające im próby oraz skuteczność są ze sobą wysoko skorelowane, zatem postanowiłem usunąć kolumny opisujące liczby prób rzutów wszelkiego rodzaju (kolumny "FGA", "X3PA", "X2PA", "FTA"). Te informację można łatwo uzyskać za pomocą liczby celnych rzutów oraz skuteczności, a usunięcie ich upraszcza analizę i usuwa współliniowość w pewnym stopniu. Również kolumna "TRB"opisująca wszystkie zbiórki, została

usunięta, gdyż występujące w zbiorze cechy określające osobno zbiórki ofensywne oraz defensywne, dostarczają wystarczająco informacji o tym aspekcie gry.

W tym momencie pozostało 20 statystyk, które nie są już w tak dużej mierze od siebie zależne. Statystyki opisują różne aspekty gry na boisku od defensywnych zagrań po ofensywne. Dokonaliśmy już wstępnej analizy zależności statystyk meczowych od zarobków, wyróżniając pewne zmienne. Dalszą analizę wpływu poszczególnych cech na zarobki przeprowadzę przy budowie modelu.

3.3 Szczegółowa analiza ogólnych informacji o zawodnikach

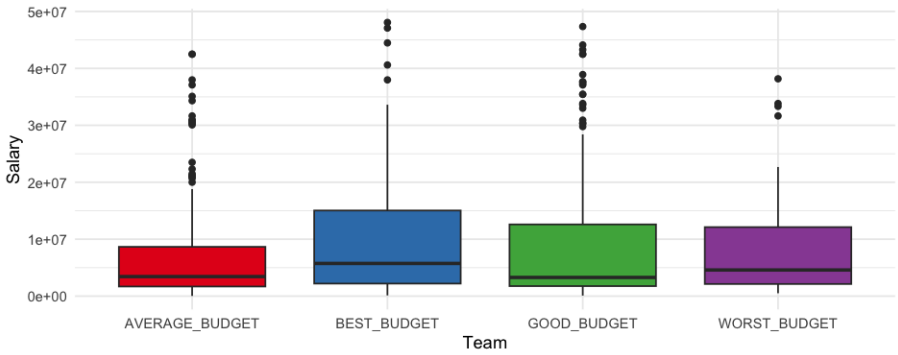
Po dokładnej analizie statystyk sportowych przyszła pora na bliższe przyjrzenie się ogólnym informacjom dotyczącym zawodników. Jednym z interesujących aspektów jest pozycja zawodnika na boisku. W zbiorze danych występuje 9 unikatowych pozycji, jednak ja pogrupuję je w 5 najpopularniejszych pozycji, czyli PG - Rozgrywający, PF - Silny Skrzydłowy, SG - Rzucający Obrońca, SF - Niski Skrzydłowy, C - Środkowy. Utworzone grupy mają stosunkowo podobne liczności. Sprawdźmy, jak przedstawiają się zarobki w każdej z tych grup.



Rysunek 4: Rozkłady zarobków zawodników dla każdej z pozycji

Widać, że rozkłady ze względu na pozycję nie różnią się znacznie. Głównie wyróżnia się pozycja PG, rzeczywiście niektórzy twierdzą, że gra na tej pozycji wymaga największych umiejętności (więcej informacji tutaj), co z pewnością wiąże się z większymi wymaganiami finansowymi. Bazując na tym, uznałem, że wyróżnię pozycję PG, a resztę połączę w jedną grupę, co również uprości analizę.

Przy analizie zarobków zawodników, ważnym aspektem wydają się być również kluby, w których występują. Sama nazwa klubu nie daje wiele informacji, a uwzględniając dodatkowo dużą ilość zespołów (30), podjąłem decyzję o zastosowaniu bardziej zaawansowanego podejścia. Postanowiłem podzielić kluby na cztery kategorie, bazując na wycenianych wartościach danego klubu (więcej informacji tutaj). Logika finansowa jest klarowna: kluby o wyższej wartości rynkowej powinny oferować wyższe wynagrodzenia. W lidze NBA w sezonie 2022/23 można wyróżnić 3 kluby z największym budżetem oraz 3 z najmniejszym. Pozostałe zespoły mają zbliżone wartości, które podzieliłem na dwie grupy. Ważnym aspektem, jest również fakt, że w zbiorze danych mamy zawodników z przypisanymi dwoma klubami (jest to związane z transferem w środku sezonu). W takich przypadkach postanowiłem skupić się tylko na pierwszym klubie zawodnika, gdyż dla tego klubu jest przypisana jego pensja w zbiorze.



Rysunek 5: Rozkłady zarobków zawodników dla każdej z grup klubów

Jak widać, zarobki zawodników nie zależą w dużej mierze od finansów klubu. W lidze NBA kluczowi gracze często rotują między zespołami, aby zapewnić pewnego rodzaju równowagę w lidze, przykładowo Rudy Gobert grający w klubie o niskim budżecie zarabia ponad 38 mln dolarów (13 wynik w zbiorze). Chociaż różnice nie są duże, widać jednak, że zawodnicy o najwyższych zarobkach grają głównie w klubach o największych lub wysokich budżetach. Postanowiłem więc połączyć utworzone grupy w bardziej ogólne dwie kategorie - kluby o małych budżetach (WORST_BUDGET oraz AVERAGE_BUDGET) oraz kluby o wysokich budżetach (BEST_BUDGET i GOOD_BUDGET).

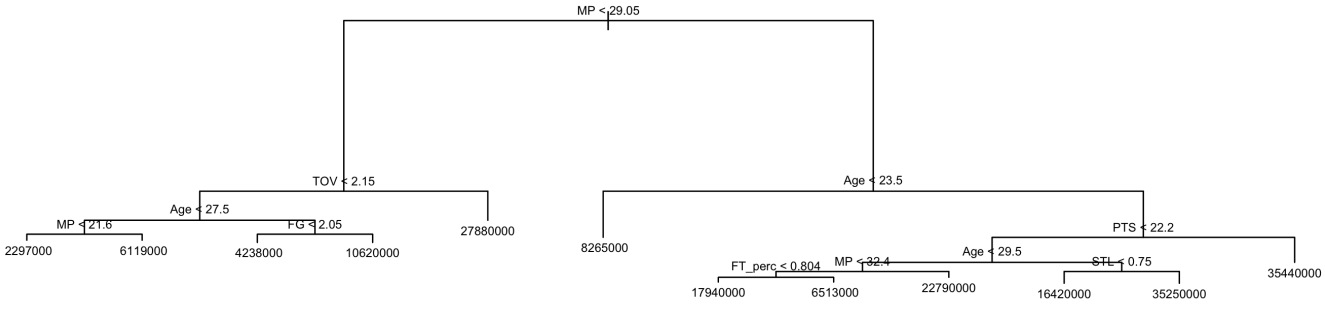
Pozostałe zmienne określające ogólne informację o zawodniku to wiek oraz imię i nazwisko. Wiek zdecydowanie może mieć wpływ na zarobki, doświadczenie w koszykówce ma duże znaczenie i często trzeba za nie "zapłacić". Z drugiej strony, zawodnicy młodzi z dużym potencjałem, również otrzymują większe pensję od władz, aby przekonać ich do pozostania w klubie. Zmienna wiek może mieć zatem duże znaczenie przy predykcji zarobków. Ze zbioru danych usunąłem jednak imiona oraz nazwiska zawodników. Dzięki temu, analiza staje się bardziej obiektywna, skupiając się na mierzalnych osiągnięciach. Ostatecznie zbiór danych składa się z 465 różnych koszykarzy z 24 metrykami (kolumnami) opisującymi każdego z nich oraz nie posiada żadnych brakujących wpisów.

4 Budowa modelu

Dalszą analizę oraz predykcję zarobków koszykarzy chciałbym oprzeć na modelu drzewiastym, a dokładniej drzewie regresyjnym. Zdecydowałem się na ten rodzaj modelu z kilku powodów, między innymi:

- model może z łatwością być używany dla zmiennych ilościowych jak i jakościowych, które występują w opracowanym zbiorze danych
- podstawowy model jest prosty w interpretacji, utworzone kryteria określające zarobki będą łatwe w zrozumieniu
- w związku ze stosunkowo dużą liczbą zmiennych objaśniających, zdolność modelu do wyboru zmiennych 'niosących najwięcej informacji' z dostępnego zbioru będzie przydatna
- istnieje możliwość poprawy 'jakości' predykcji modelu za pomocą metod grupowania przedstawionych podczas zajęć

Na początek podzieliłem zbiór danych na zbiór treningowy oraz testowy. Zdecydowałem się na proporcję 80/20 ze względu na niedużą ilość danych w zbiorze. Zadbałem również, aby zmienna objaśniana w obu zbiorach osiągała podobne wartości. Następnie przeszedłem do budowy modelu. Pierwszy powstały model jest drzewem regresyjnym opartym na wszystkich dostępnych cechach, a jego struktura jest przedstawiona poniżej.



Rysunek 6: Wstępne drzewo regresyjne zbudowane na zbiorze treningowym

Analizując strukturę drzewa, można zauważyć, że decyzje o podziałach na grupy przy zmiennych określających pozycję oraz zespół nie miały większego znaczenia. Rzeczywiście, jak było widać wcześniej w analizie tych cech, nie separowały one wartości zarobków zawodników w sposób klarowny. Idąc dalej, model uznał zmienne takie jak MP czy Age za bardzo istotne w kontekście oceny zarobków, co wydaje się być również zgodne z intuicją. Widać również, że PTS odgrywa kluczową rolę przy dużych zarobkach. Zawodnicy zdobywający więcej niż 22.2 punktów na mecz (spełniający również poprzednie warunki drzewa) według modelu powinni zarabiać ponad 35 milionów dolarów. Rzeczywiście jest to ciężkie do osiągnięcia i tylko nieliczni (najlepiej zarabiający) potrafią tego dokonać.

5 Ocena modelu

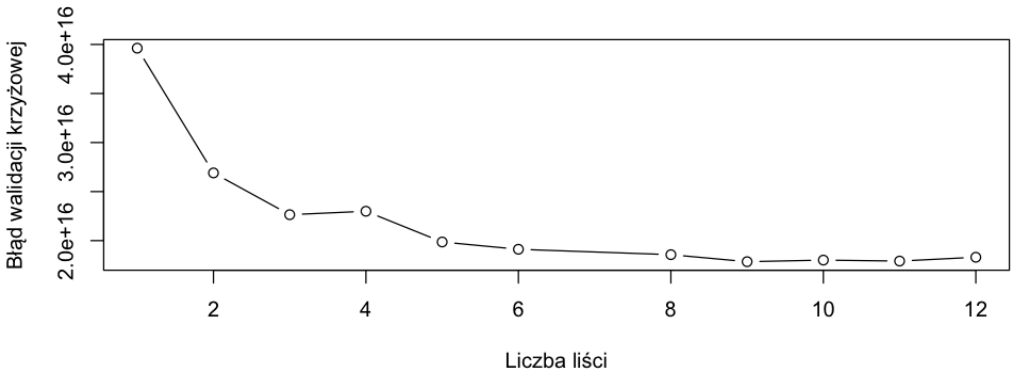
Oceniając jakość predykcji modelu, na początek posłużyłem się średnim błędem bezwzględnym, gdyż jest on prosty w interpretacji. Dla wybranego zbioru treningowego oraz testowego wartość tego błędu to ponad 4.2 miliona dolarów, co można potraktować jako duży błąd. Analizując dalej, model zdecydowanie obniża wartości dla największych zarobków zawodników (maksymalna wartość predykcji to około 35 mln $, gdzie w zbiorze testowym wynosi aż 48 mln $). Również bardzo niskie zarobki nie są otrzymywane przez drzewo. Wartości charakteryzujące rozkłady wartości przewidywanych oraz rzeczywistych są przedstawione w tabeli poniżej.

Zbiór	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Oryginalne wartości w zbiorze testowym	35,096	1,563,518	2,905,851	9,375,162	10,900,634	48,070,014
Predykcja wartości dla zbioru testowego	2,296,709	2,296,709	4,237,860	9,568,981	10,624,731	35,440,234

Tabela 2: Porównanie predykcji z oryginalnymi wartościami zarobków w zbiorze testowym

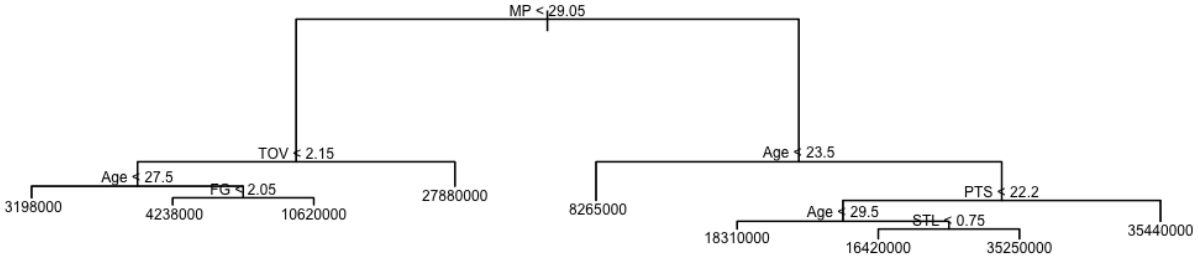
Sumując wszystkie różnice między predykcjami, a rzeczywistymi wartościami, model zawyża zarobki o 18 mln $. Największe błędy popełnia jednak dla predykcji największych wartości zarobków. Omawiane błędy są spowodowane naturą zarobków koszykarzy - mało zawodników zarabia ogromne kwoty, zatem model nie ma wystarczająco informacji o wartościach tego rzędu. Problem ten przytaczałem już na początku referatu, analizując naszą zmienną objaśnianą.

Sprawdźmy jeszcze czy model nie dopasowuje się zbytnio do danych. Przeprowadzimy walidację krzyżową i sprawdzimy, czy drzewo nie wymaga przycięcia.



Rysunek 7: Wykres średniego błędu w zależności od rozmiaru drzewa

Analizując wykres, można zauważyć, że rozmiar drzewa minimalizujący błąd walidacji krzyżowej to 9. Przytnijmy zatem drzewo i sprawdźmy jego strukturę.



Rysunek 8: Przycięte drzewo regresyjne

Po przycięciu drzewa zamiast 12 liści mamy 9, utraciliśmy zatem 3 warunki modelu. Średni błąd bezwzględny dla przyciętego drzewa wynosi około 4.5 mln $, zatem pogorszył się o 0.3 mln $. Uzyskaliśmy jednak model mający lepszą zdolność do generalizacji na nowe dane.

6 Dopracowanie modelu

Model wymaga jednak dopracowania. W tym rozdziale skupię się na różnych sposobach udoskonaleniu modelu. Rozważę różnorodne metody poznane na zajęciach mające na celu poprawę jego jakości, aby ostatecznie wybrać jeden finalny model.

6.1 Regresja w liściach

Pierwsze podejście poprawy modelu jest oparte na zastosowaniu regresji w liściach drzewa. Stosując ten sam zbiór treningowy zbudowałem omawiany model. Poniższa tabela w zwięzły sposób opisuje zmienne wykorzystywane w modelu.

Zmienne użyte w modelu			
Używane przy warunkach modelu	Częstość użycia w %	Używane przy liniowych modelach w liściach	Częstość użycia w % (>51%)
PTS	82	PTS	98
Age	69	Age	98
X2P	43	X2P	98
FT	18	FT	98
X2P_perc	14	FG	98
MP	8	X3P	98

Tabela 3: Lista zmiennych najczęściej używanych przy warunkach modelu oraz liniowych modelach w liściach

Utworzone drzewo składa się z 7 reguł do których używane jest 6 zmiennych wymienionych w tabeli. Zauważalne jest, że niektóre zmienne, takie jak PTS, Age, X2P czy FT są kluczowe zarówno dla warunków, jak i dla liniowych modeli, co ponownie wskazuje na ich znaczącą rolę.

Oceniając model, średni błąd bezwzględny wynosi około 3.6 mln $, zatem "średnio"model wypada lepiej od pierwotnego. Sumując jednak różnice między predykcjami, a rzeczywistymi wartościami, model niedoszacowuje zarobki na prawie 140 milionów! Jest to spowodowane między innymi faktem, że model przewidział zarobki równe 0 dla dwóch zawodników. Zatem zmniejszył się średni błąd, jednak jego ochylenie się zwiększyło (z 4.6 mln $ na 5 mln $). Podsumowując, model "średnio"zachowuje się lepiej, jednak ma większe odchylenie błędu oraz ciężej go interpretować. Biorąc to pod uwagę, nie jest on wystarczająco lepszą metodą od pierwotnej. Przejdźmy więc do kolejnych metod dopracowywania modelu.

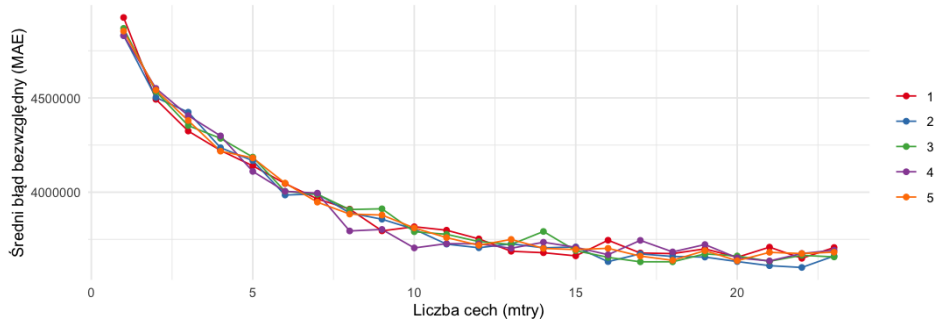
6.2 Bagging oraz Lasy Losowe

Kolejne metody mające na celu poprawę jakości modelu to bagging oraz lasy losowe. Zbudujmy odpowiednie modele.

W przypadku metodologii Bagging korzystamy z wszystkich dostępnych 23 parametrów. Uzyskany model ma średni błąd bezwzględny wynoszący około 3.7 mln $ oraz odchylenie 4.6 mln $. Poprawiliśmy zatem dokładność predykcyjną kosztem utraty interpretowalności.

Budując model oparty o Lasy Losowe korzystałem z 8 zmiennych ze zbioru danych. Te metoda wprowadza pewną losowość celem zmniejszenia korelacji pomiędzy drzewami składowymi. Otrzymany model posiada średni błąd bezwzględny na poziomie 3.9 mln $ oraz odchylenie, jak przy metodzie Bagging, 4.6 mln $.

Przy każdej próbie metoda Bagging dawała jednak lepsze wyniki niż Lasy Losowe. Średni błąd bezwzględny oraz odchylenie błędu maleją wraz z rosnącą ilością cech używanych do podziału. Poniżej jest zwizualowana zależność średniego błędu bezwgzlędnego od liczby parametrów.



Rysunek 9: Wykres średniego błędu bezwględnego dla różnej liczby cech dla 5 iteracji

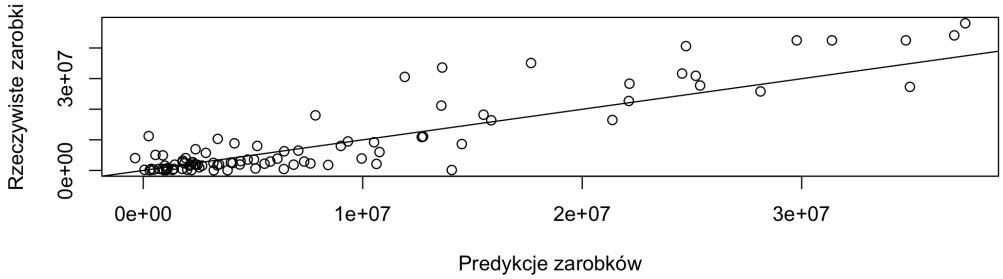
Widać zatem, że metoda Bagging przy omawianym zbiorze danych wypada lepiej niż Lasy Losowe. Może to być jednak spowodowane zbytnim dopasowaniem do danych. Minimalny błąd spośród 5 powtórzeń wyniósł blisko 3.6 mln $, zatem jest to pewna przybliżona dolna granica błędu dla tych metod. Dla obu metodologii jednak te same zmienne są najbardziej istotne. Zmienne Age, MP, TOV, PTS, FT mają najwyższą wartość *%IncMSE*, co oznacza, że wykluczenie każdej z nich pojedynczo ze zbioru danych najbardziej zwiększyłoby błąd średniokwadratowy. Jest to kolejne już potwierdzenie znaczącego wpływu tych zmiennych na zarobki.

6.3 Boosting oraz XGBoosting

W dalszej kolejności chciałbym przeanalizować metody Boosting oraz XGBoosting dla naszych danych.

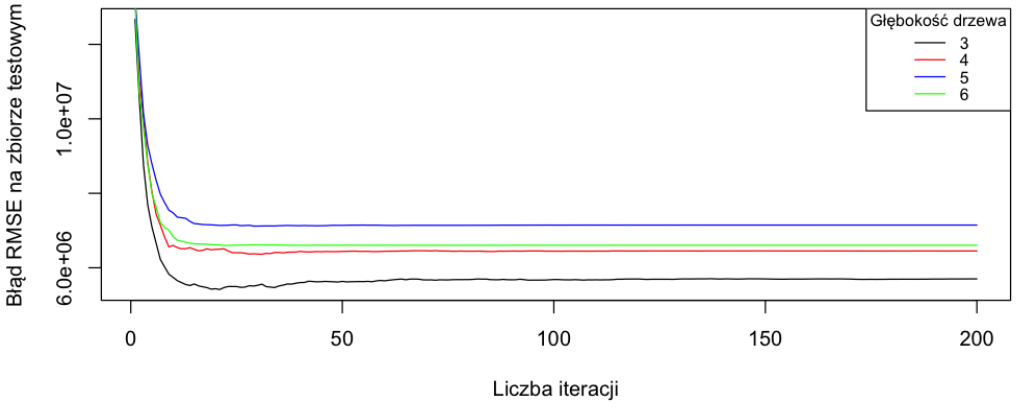
Zaczynając od metodologii Boosting, przygotowanie modelu rozpocząłem od znalezienia odpowiednich paramertów za pomocą walidacji krzyżowej. Rozważyłem siatkę parametrów metody, zatem liczby drzew B, parametru λ, liczby podziału d w drzewach oraz minimalnej liczby obserwacji w liściu. Po dokonaniu walidacji krzyżowej otrzymane parametry to B=1000, λ=0.01, d=5 oraz 20 będąca minimalną liczbą obserwacji w liściu. Następnie przeszedłem do budowy modelu z otrzymanymi parametrami.

Analizując otrzymany model, wciąż zmienne takie jak Age, TOV, FT, MP, AST, PTS odgrywają największą rolę. Mają największe wartości miary *rel.inf*, która określa, jak ważna jest każda zmienna dla prognozy modelu. Średni błąd bezwzględny modelu wynosi 3.8 mln $, a odchylenie 4.4 mln $. Widać zatem poprawę względem poprzednich modeli. Poniższy wykres przedstawia rozrzut błędów predykcji. Widać, że dla większych zarobków model popełnia większe błędy (odległość od prostej jest większa). Dodatkowo, można zauważyć, że dla większych wartości model zaniża zarobki, a dla mniejszych zawyża. Ten problem przewija się przez cały referat i jest spowodowany między innymi zbyt małą ilością danych.



Rysunek 10: Wykres porównujący predykcje z rzeczywistymi wartościami (linia prosta oznacza dokładną predykcję)

Przechodząc do algorytmu XGBoosting, analizę rozpoczynam od ustalenia parametrów metody. Zbiór danych nie jest duży, zatem liczbę iteracji ustalam na 200. W celu określenia optymalnej maksymalnej głębokości drzewa, porównuję wyniki dla różnych parametrów. Rozpatruję wartości głębokości drzewa od 3 do 6, gdyż najczęściej przyjmuję się owy parametr z tego przedziału.



Rysunek 11: Błąd RMSE w zależności od liczby iteracji dla różnych głębokości drzewa

Na podstawie wykresu widać, że model z maksymalną głębokością równą 4 daje najmniejszy błąd średniokwadratowy na zbiorze testowym, zatem wybieram ten parametr. Przejdźmy do budowy oraz oceny modelu.

Najmniejszy błąd średniokwadratowy występuje po 31 iteracjach, zatem dla tej liczby iteracji tworzymy ostateczny model. Średni błąd bezwględny modelu to 4 mln $, a odchylenie wynosi prawie 5 mln $. Porównując poszczególne predykcje, zbudowany model wypada gorzej od modelu opartego na zwykłym Boostingu w prawie każdym aspekcie, ma jedynie mniejszy minimalny błąd predykcji (najmniejszą różnicę między wartością przewidywaną oraz prawdziwą). Co ciekawe, oba modele niedoszacowują wartości na bardzo podobnym poziomie (blisko 80 mln $ łącznie), jednak model oparty o Boosting w ogólności jest lepszy i pozostaje najlepszą z rozważanych metod.

7 Podsumowanie

Predykcja zarobków koszykarzy z pewnością nie należy do prostych zadań. Podczas analizy mogliśmy jednak dostrzec zmienne, które szczególnie wpływają na podział zarobków wśród zawodników NBA. Zmienne takie jak Age, PTS, MP, FT, TOV powtarzały się jako najbardziej istotne we wszystkich opisanych modelach. Rzeczywiście może to być zgodne z intuicją, wiek z pewnością ma wpływ na zarobki, a wymienione statystyki są kluczowe w meczach koszykówki, zatem przekładają się na pensje zawodników. Jednym z interesujących wniosków jest brak większego wpływu pozycji czy klubu zawodnika na wartość jego zarobków.

Podsumowując wyniki predykcyjne modelów, osiągnięte rezultaty nie są szczególnie satysfakcjonujące. Średnie błędy bezwgzlędne na poziomie 4 mln $ są stosunkowe duże, a żadna z opisanych metod nie poprawiła jakości w sposób tak znaczący, aby uznać go za dokładnym przy predykcjach. Modele jakie chciałem jednak wyróżnić, to pierwsze przycięte drzewo regresyjne oraz model oparty na metodzie Boosting. Przycięte drzewo ma trochę większy błąd w stosunku do reszty, jednak z pewnością przez walidację krzyżową nie jest zbytnio dopasowany do danych. Jest on też bardzo łatwy w interpretacji, co jest ważną zaletą. Model oparty o Boosting z kolei ma najmniejszy średni błąd bezwzględny (3.8 mln $) oraz niskie odchylenie błędu (4.4 mln $), co w porównaniu do skali zarobków wynoszących od kilkudziesięciu tysięcy dolarów do prawie 50 milionów $ daje stosunkowo zadowalający wynik. Modele boostingowe są też mniej podatne na zbytnie dopasowanie do danych, co jest kolejną zaletą modelu. Niestety model nie jest łatwo interpretowalny, jednak biorąc pod uwagę wcześniej wymienione aspekty, wybieram ten model jako finalny i najlepiej dostosowany do przewidywania zarobków koszykarzy.

Ogólnie rzecz biorąc. powodem przeciętnych wyników jest między innymi dostępny zbiór danych oraz natura zarobków koszykarzy. Poprzez niewielką liczbę graczy w lidze NBA oraz bardzo zróżnicowane zarobki, aglorytmy nie wyciągneły istotnych statystycznie wniosków na temat niektórych pensji. Dodatkowo, badanie zależności wypłat od statystyk zawodników jest zdecydowanie dużym uproszczeniem problemu. Potraktowanie zagadnienia jako wyłączną zależność zarobków od osiągnięć sportowych, wydaje się być podejściem wyidealizowanym. Zawodnicy NBA to jednak nie tylko mistrzowie koszykówki, ale również celebryci. Dzięki temu przynoszą swoim klubom znaczące korzyści finansowe. Ich renoma, imię i nazwisko mogą znacząco wpływać na wartość marki klubu, przyciągając sponsorów, fanów i media, co z kolei może mieć bezpośredni wpływ na ich wynagrodzenia. Uwzględnienie tych aspektów w analizie mogłoby znacząco wzbogacić model i uczynić go bardziej wszechstronnym w uchwyceniu pełnego obrazu zarobków w lidze NBA.

W ramach projektu dokonałem zaawansowanej analizy postawionego problemu, jednak istnieje wiele możliwości udoskonalenia modelu, zwiększających jego precyzję i użyteczność. Segmentacja zarobków do grup o zbliżonej liczności może pozwolić algorytmom na efektywniejsze identyfikowanie wartości odstających. Dodatkowo, wzbogacenie modelu o nowe zmienne (dotyczące renomy zawodnika) z pewnością przyczyni się do głębszego zrozumienia czynników wpływających na wynagrodzenia. Kolejnym rozwiązaniem jest przeanalizowanie zarobków z kilku sezonów ligi, dostając w ten sposób wiecej informacji o różnych zawodnikach i ich zarobkach.