Predykcja wartości rynkowej piłkarzy

Jędrzej Sarna

12 lutego 2024

1 Wprowadzenie

W dzisiejszych czasach piłka nożna ewoluowała, stając się dyscypliną, gdzie decyzje nie są już podejmowane tylko na podstawie intuicji czy talentu. Główną rolę zaczęły odgrywać dane. Liczby, statystyki i zaawansowane analizy są teraz nieodłącznym elementem zarówno planowania strategii meczowej, jak i decyzji na rynku transferowym. Wiele klubów piłkarskich, takich jak na przykład Brighton & Hove Albion FC, zaadaptowało strategie opierające się na danych, wykorzystując zaawansowane algorytmy do rekrutacji nowych zawodników [1]. Świeżym przykładem w polskiej piłce jest przypadek Wisły Kraków, gdzie prezes Jarosław Królewski większość decyzji sportowych wokół klubu opiera na zaawansowanych analizach danych [2]. Dzięki temu podejściu, nawet kluby z mniejszymi budżetami są w stanie konkurować z największymi, oferując nowoczesny i efektywny futbol.

2 Sformułowanie problemu

W obliczu rosnącej roli danych w piłce nożnej, zdecydowałem, że chciałbym samodzielnie zgłębić świat analizy piłkarskiej. Moim celem jest zrozumienie i wykorzystanie dostępnych statystyk piłkarzy, aby na ich podstawie dokonać własnych analiz oraz stworzyć model predykcyjny pozwalający przewidywać wartość rynkową piłkarzy na podstawie ich statystyk meczowych.

2.1 Zbieranie danych

Dane na których będę opierał moją analize pochodzą ze strony [3]. Zbiór danych składa się z informacji o piłkarzach pozyskanych z dwóch wiarygodnych stron internetowych - transfermarkt.co.uk, udostępniający ogólne informację o zawodnikach oraz fbref.com, posiadający szczegółowe statystyki meczowe zawodników z poprzednich sezonów. Dane dotyczą piłkarzy, którzy grali w jednej z pięciu najlepszych lig europejskich - angielskiej Premier League, włoskiej Serie A, hiszpańskiej La Liga, niemieckiej Bundeslidze czy francuskiej Ligue 1 w sezonach od 2017/2018 do 2020/2021.

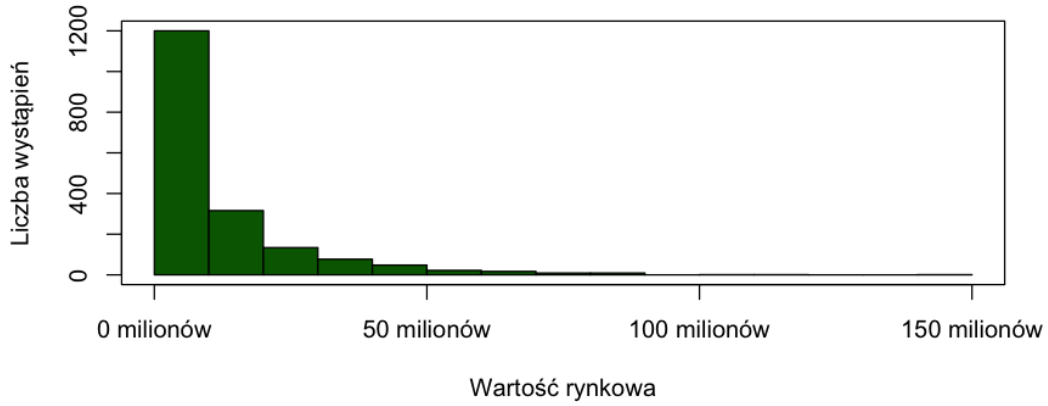
3 Eksploarcja i przygotowanie danych

3.1 Wstępna eksploracja danych

Po wgraniu danych do środowiska R można zauważyć, że jest ponad 2000 rekordów oraz co istotne, aż 548 kolumn. Zdecydowałem się na ten zbioru danych, ponieważ obecnie praca z obszernymi danymi stała się standardem, nawet w sporcie, zatem jest to okazja do podszkolenia się w tym aspekcie.

Po wstępnym przejrzeniu danych można zauważyć, że informacje można podzielić na dwie główne kategorie. Pierwsza obejmuje ogólne informacje o zawodnikach , takie jak nazwa obecnego klubu oraz ligi, pozycja na boisku, imię oraz wiek, pozostała długość kontraktu oraz zmienna wartość rynkowa (w walucie Euro), będąca naszą zmienną objaśnianą. Druga część danych skupia się na statystykach meczowych z poprzednich sezonów, zawierając szczegółowe informacje dotyczące występów każdego zawodnika w danym sezonie.

Przyjrzymy się teraz bardziej szczegółowo wartościom rynkowym piłkarzy dostępnych w naszym zbiorze danych.



Rysunek 1: Histogram wartości rynkowych w zbiorze danych

Przedstawiony histogram dobrze ilustruje rzeczywistość rynku piłkarskiego. W świecie futbolu, faktycznie tylko niewielka część zawodników osiąga bardzo wysokie wartości rynkowe. To zazwyczaj gracze wybitni, posiadający wyjątkowe umiejętności, znaczące osiągnięcia lub ogromny potencjał, co sprawia, że są oni szczególnie cenni dla swoich klubów i pożądani na rynku transferowym. 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 Wstępne przygotowanie danych

W zbiorze danych znajdują się również informacje o bramkarzach. Jednakże, należy zauważyć istotny aspekt dotyczący specyfiki danych zebranych dla tych zawodników. Mianowicie, żadna z gromadzonych metryk nie oddaje zakresu aktywności bramkarza na boisku, w związku z tym, konieczne jest wykluczenie bramkarzy ze zbioru danych. Kontynuując analizę, zmienna określająca pozostałą długość kontraktu, ma nietypową wartość 'fail'. Nie byłem w stanie zidentyfikować powodu takiego rodzaju wpisu, zatem biorąc pod uwagę małą liczność obserwacji z nim związanych, usunąłem je ze zbioru danych.

Można również dostrzec, że nie wszyscy piłkarze ze zbioru grali profesjonalnie przez ostatnie 4 lata lub ich występy miały miejsce w ligach spoza "Top 5". Z tego powodu nie posiadają oni wszystkich zebranych statystyk, zatem, aby zapewnić sprawiedliwość i spójność analizy, podjąłem decyzję o usunięciu tych zawodników z zestawienia. Dostrzegłem również, że w zbiorze jest 14 zduplikowanych zawodników, zatem usunąłem duplikaty ze zbioru.

3.3 Szczegółowa analiza statystyk meczowych

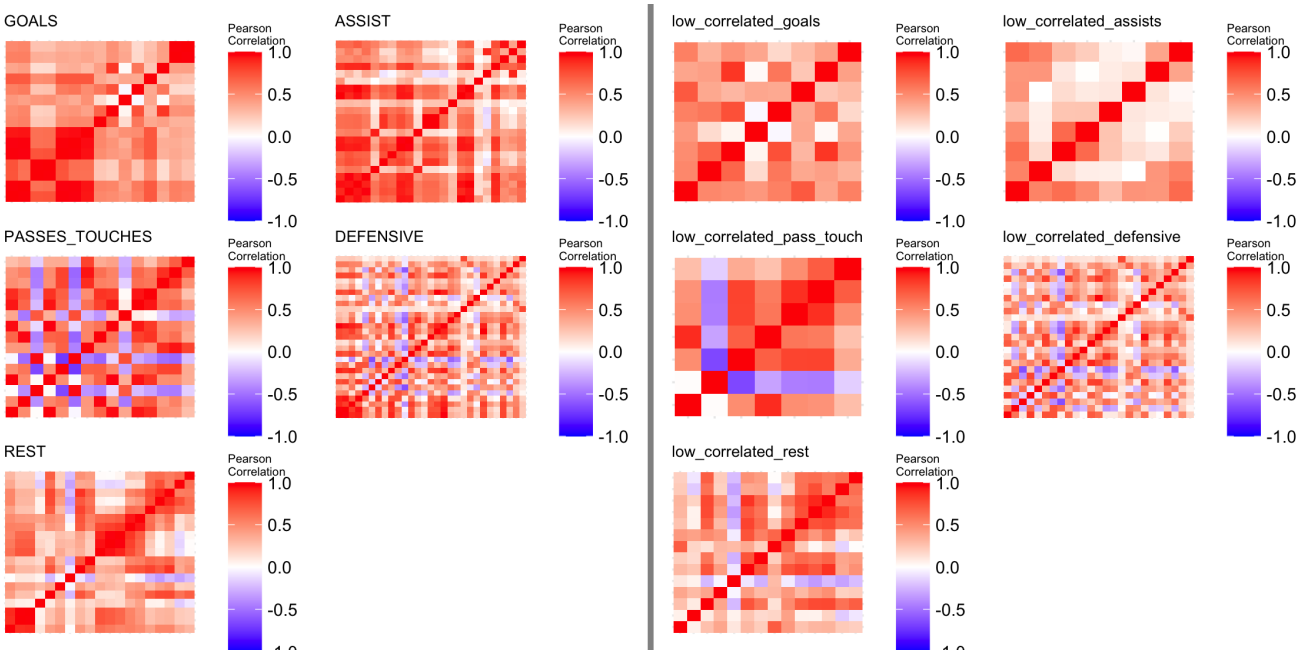
Zgromadzone statystyki charakteryzują się wysokim poziomem szczegółowości, oferując dane dla każdego sezonu z osobna. Mimo to, podjąłem decyzję o przyjęciu pewnego uproszczenia, polegającego na połączeniu danych z różnych sezonów w jedną spójną całość. To podejście ma swoje istotne zalety, pozwala na uzyskanie szerszej perspektywy na rolę i wydajność zawodników w ciągu ostatnich czterech lat. Taki przekrój danych ułatwia też analizę, zmniejszając liczbę dostępnych metryk (kolumn dotyczących statystyk) z 536 na 134.

Przystępując do procesu łączenia danych, zadbałem o staranne i właściwe przetworzenie zmiennych. W przypadku danych liczbowych, takich jak liczba strzelonych goli, dokonałem ich sumowania, aby uzyskać całkowitą liczbę dla danego zawodnika przez ostatnie cztery lata. Natomiast zmienne wyrażające wskaźniki procentowe – jak na przykład procent celnych strzałów – zostały uśrednione. Podczas dalszej analizy zauważyłem, że zmienne procentowe w zbiorze są wynikiem stosunku dwóch innych kolumn danych. Wiedząc to, postanowiłem podjąć kolejny krok w kierunku uproszczenia zestawu danych. Zdecydowałem się usunąć kolumny, które służyły do wygenerowania zmiennych procentowych, zachowując tylko te zawierające procentowe stosunki. Ta decyzja była wsparta moją wiedzą z zakresu statystyk piłkarskich i ma na celu stworzenie bardziej przejrzystego zbioru danych. Tak samo postąpiłem z kolumnami przedstawiającymi metryki na 90 minut (np. liczba celnych strzałów na 90 minut), które również powstawały na podstawie innych kolumn w zbiorze.

W tym momencie w zbiorze pozostało 99 metryk. Dalej jednak można było dostrzec wiele statystyk bardzo do siebie zbliżonych (np. kontakty z piłką oraz kontakty z piłką w grze).Podjąłem decyzję o podziale metryk na pięć kategorii, wykorzystując do tego moją wiedzę i doświadczenie w analizie statystyk piłkarskich. Taki podział pozwala na bardziej szczegółową analizę zmiennych wewnątrz grup. Kategorie te obejmują:

- **Gole**: Ta grupa skupia się na metrykach związanych ze zdobywaniem bramek, takich jak liczba strzelonych goli, skuteczność strzałów, metryka xG i podobne.
- **Asysty**: W tej kategorii analizowane są dane dotyczące asyst, kluczowych podań prowadzących do goli oraz innych statystyk, które odzwierciedlają wkład zawodnika w kreowanie okazji bramkowych.
- **Podania oraz kontakty z piłką**: Zawiera informacje o liczbie i skuteczności podań, częstotliwości kontaktów z piłką, a także o jakości i precyzji rozgrywania piłki.
- **Akcje defensywne**: Ta grupa koncentruje się na aspektach defensywnych i dyscyplinarnych, w tym takich metrykach jak odbiory, bloki, przechwyty, czy skuteczność pojedynków defensywnych.
- **Pozostałe**: W tej kategorii znajdują się wszystkie inne istotne zmienne, które nie mieszczą się w powyższych kategoriach, ale mogą mieć znaczenie dla oceny ogólnej wartości zawodnika

W procesie identyfikacji statystyk opisujących podobne zjawiska, polegałem w dużej mierze na własnej intuicji oraz doświadczeniu z zakresu piłki nożnej. Zrozumiałe nazwy zmiennych w zbiorze danych pomogły w wyodrębnieniu unikatowych metryk. Jednak, aby analiza nie opierała się jedynie na intuicji, postanowiłem podeprzeć się narzędziem statystycznym jakim jest współczynnik korelacji Pearsona. Zdecydowałem się na tą metodę ze względu na jej względną prostotę, przydatną szczególnie przy analizie większej liczby zmiennych, oraz fakt, że 'podobne' kolumny w tym zbiorze danych będą ze sobą w dużej mierze liniowo zależne (ze względu na naturę ich powstawania), zatem jest to odpowiednia miara.



Rysunek 2: Macierze korelacji dla grup metryk przed usunięciem 'podobnych' kolumn (lewa strona) oraz po usunięciu (prawa strona)

Jak widać na rysunku powyżej, wiele wysoko skorelowanych kolumn zostało usuniętych ze zbioru danych. Rzeczywiście były to kolumny o identycznym lub zbliżonym znaczeniu. Po tym zabiegu w zbiorze danych pozostały 47 różne metryki, które zostają już finalnym zbiorem statystyk meczowych oceniających występy piłkarzy.

3.4 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 w naszym zbiorze danych jest kolumna "narodowość", która zawiera aż 75 różnych wartości. Biorąc pod uwagę tę różnorodność, zdecydowałem, że bardziej efektywnym i przejrzystym podejściem będzie grupowanie narodowości zawodników według kontynentów.

Przy analizie wartości rynkowej zawodników, kluczowym aspektem wydają się być również kluby, w którym występują. Nazwa klubu sama w sobie nie daje dużo informacji, dlatego podjąłem decyzję o zastosowaniu bardziej zaawansowanego podejścia. Postanowiłem podzielić kluby na trzy wyraźnie zdefiniowane kategorie, bazując na ich wynikach w ciągu sezonów od 2017/18 do 2020/21. W tym celu stworzyłem tabelę skumulowanych punktów z czterech sezonów dla każdej z lig.

PREMIER LEAGUE		SERIE A		LA LIGA		LIGUE 1		BUNDESLIGA	
Team	Total Points	Team	Total Points	Team	Total Points	Team	Total Points	Team	Total Points
Manchester City	365	Juventus	339	FC Barcelona	341	Paris Saint-Germain	334	Bayern Monachi	322
Liverpool	340	Inter Mediolan	314	Real Madryt	315	Olympique Lyon	266	Borussia Dortmu	264
Manchester United	287	Napoli	309	Atlético Madryt	311	Olympique Marsylia	254	RB Lipsk	250
Chelsea	275	Atalanta	285	Sevilla FC	264	Lille OSC	245	Bayer Leverkuse	228
Tottenham Hotspur	269	Milan	277	Valencia CF	230	AS Monaco	234	Borussia Mönche	216
Arsenal	250	Lazio	277	Villarreal CF	223	Stade Rennes	218	Eintracht Frankfu	208
Leicester City	227	Roma	275	Real Sociedad	217	Montpellier HSC	204	TSG Hoffenheim	201
Everton	211	Sampdoria	201	Real Betis	212	OGC Nice	203	VfL Wolfsburg	198
West Ham United	198	Sassuolo	199	Getafe CF	206	AS Saint-Étienne	197	SC Freiburg	165
Burnley	187	Torino	194	Athletic Bilbao	193	Girondins Bordeaux	178	Hertha BSC	162
Crystal Palace	180	Fiorentina	187	Levante UD	180	FC Nantes	177	1. FSV Mainz 05	155
Newcastle United	178	Bologna	171	Celta Vigo	180	Angers SCO	170	FC Augsburg	145
Southampton	170	Udinese	168	Deportivo Alavés	174	RC Strasbourg	167	FC Schalke 04	135
Wolverhampton Wande	161	Cagliari	162	SD Eibar	170	Stade de Reims	138	Werder Brema	126
Brighton & Hove Albion	158	Genoa	160	RCD Espanyol	127	Dijon FCO	133	VfB Stuttgart	96
Watford	91	Verona	119	CD Leganés	124	Nîmes Olympique	115	Union Berlin	91
Aston Villa	90	Parma	110	Real Valladolid	114	FC Metz	107	1. FC Köln	69
Bournemouth	89	SPAL	100	Granada CF	102	Toulouse FC	88	Fortuna Düsseldc	44
Leeds United	59	Crotone	58	CA Osasuna	96	Amiens SC	83	Hannover 96	39
West Bromwich Albion	57	Chievo Verona	57	Girona FC	88	Brest	75	Arminia Bielefeld	35
Sheffield United	54	Benevento	54	SD Huesca	67	EA Guingamp	74	Werder Brema	31
Fulham	54	Spezia	39	Cádiz CF	44	SM Caen	71	Hamburger SV	31
Huddersfield Town	37	Empoli	38	Elche CF	36	RC Lens	57	Fortuna Düsseldc	30
Watford	34	Lecce	35	RCD Mallorca	33	Lorient	42	VfB Stuttgart	28
Bournemouth	34	Frosinone	25	Rayo Vallecano	32	Troyes	33	1. FC Köln	22
Cardiff City	34	Brescia	25	Deportivo La Coruña	29	Amiens S.C.	23	Hannover 96	21
Stoke City	33			UD Las Palmas	22			SC Paderborn 07	20
Swansea City	33			Málaga CF	20			1. FC Nürnberg	19
Sheffield United	23							FC Schalke 04	16
Norwich City	21								
Huddersfield Town	16								

Rysunek 3: Tabele lig Top 5 na podstawie sezonów 2017/18 do 2020/21

Kategorie podziału to:

- **Kluby Top 6:** Te kluby są uznawane za potęgi i najczęściej bywają w czołówce tabeli (zielone wpisy w tabeli)
- **Kluby środka tabeli:** To kluby, które zazwyczaj znajdują się w stabilnym środku (żółte wpisy w tabeli)
- **Kluby dolnej części tabeli:** Te zespoły często walczą o utrzymanie (pomarańczowe wpisy w tabeli)

Na koniec usunąłem imiona oraz nazwiska piłkarzy ze zbioru danych. Dzięki temu, analiza staje się bardziej obiektywna, skupiając się na mierzalnych osiągnięciach. Ostatecznie zbiór danych składa się z 842 różnych piłkarzy z 54 metrykami (kolumnami) dla każdego z nich oraz nie posiada żadnych brakujących wpisów.

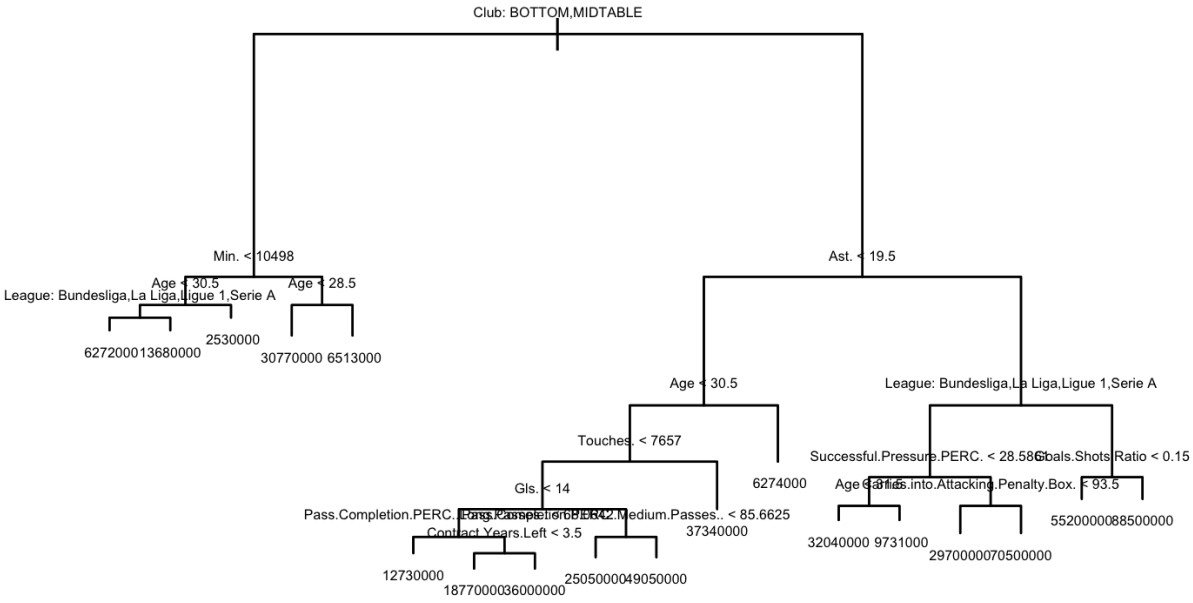
4 Budowa modelu

Dalszą analizę wartości rynkowych piłkarzy 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 wyceny zawodników będą łatwe w zrozumieniu

- w związku z cały czas stosunkowo dużą liczbą zmiennych objaśniających, zdolność modelu do wyboru zmiennych 'niosących najwięcej informacji' z dostępnego zestawu będzie przydatna
- możliwość poprawy 'jakości' predykcji modelu za pomocą metod grupowania przedstawionych podczas zajęć

Pierwszy zbudowany model jest drzewem regresyjnym opartym na wszystkich dostępnych cechach. W celu oceny 'jakości' modelu podzieliłem zbiór danych na zbiór treningowy oraz testowy w stosunku 80/20 i dokonałem głębszej analizy wyników modelu.



Rysunek 4: Wstępne drzewo regresyjne zbudowane na zbiorze treningowym

Analizując strukturę drzewa, można zauważyć, że decyzja o podziale klubów na trzy kategorie okazała się być trafionym posunięciem. Model uznał zmienną za bardzo istotną w kontekście oceny wartości rynkowej zawodników, co wydaje się być również zgodne z intuicją. Również takie podstawowe informacje jak wiek zaowdnika, liga czy ogólne statystyki jak liczba bramek, asyst, minut czy kontaktów z piłką są istotne dla modelu. Jest to z pewnością zgodne z zasadami rynku transferowego w piłce nożnej, gdzie 'suche' statystyki takie jak gole czy asysty często odgrywają kluczową rolę przy wycenie zawodnika.

5 Ocena modelu

Przechodząc do oceny wstępnego modelu, porównajmy wartości prognozowane przez model na zbiorze testowym z rzeczywistymi wartościami rynkowymi zawodników z tego zbioru.

Zbiór	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Oryginalne wartości w zbiorze testowym	360,000	3,600,000	9,000,000	16,640,947	25,200,000	90,000,000
Predykcja wartości dla zbioru testowego	2,530,000	6,272,143	8,170,000	15,650,619	21,718,421	101,700,000

Tabela 1: Porównanie predykcji z oryginalnymi wartościami rynkowymi w zbiorze testowym

Na podstawie wykresu można już dostrzec, że rozkłady predykcji oraz rzeczywistych wartości różnią się. Potwierdza to też średni błąd bezwzględny wynoszący około 8.6 mln, co nie jest zadawalającym wynikiem, nawet uwzględniając szeroki zakres wartości w zbiorze testowym (od 360 tyś do 90 mln). Szczególnie przy niższych wartościach rynkowych piłkarzy pomyłka o prawie 9 milionów Euro byłaby sporym błędem.

W tym momencie powstaje problem o którym była mowa na początku referatu, mianowicie w przygotowanym zbiorze danych również nie ma wystarczającej liczby obserwacji dla pewnych przedziałów wartości rynkowych piłkarzy. Utwierdza to też fakt, że losując różne zbiory treningowe oraz testowe, predykcje modelu znacznie odbiegają od siebie, gdyż rozkłady wartości rynkowych w zbiorach (treningowych oraz testowych) bardzo się od siebie różnią. Model jest zatem także 'niestabilny' i ma tendencję do "overfittingu". Dodatkowo, rozkłady wartości przewidywanych w liściach są z bardzo szerokich zakresów, co jest kolejnym powodem podważania skuteczności modelu.

6 Dopracowanie modelu

Model wymaga zatem 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. Poniższa tabela w dosyć zwięzły sposób opisuje zbudowany model na zbiorze treningowym.

Zmienne użyte w modelu			
Używane przy warunkach modelu.	Częstość użycia w %	Używane przy liniowych modelach w liściach	Częstość użycia w % (>70%)
Age	98	Age	100
Club	79	Gls.	100
League	63	Touches.	98
Min.	57	Min.	95
Touches.	32	Tackles.in.Attacking.3rd.	89
Gls.	24	Touches.in.Attacking.Penalty.Box.	85
Pass.Completion.PERC..Short.Passes..	19	Total.Loose.Balls.Recovered.	76
		Touches.in.Attacking.Penalty.Box.	74

Tabela 2: Lista zmiennych najczęściej używanych przy warunkach modelu oraz liniowych modelach w liściach

Utworzone drzewo składa się z 13 reguł do których używane jest 7 zmiennych wymienionych w tabeli. Zauważalne jest, że niektóre zmienne, takie jak "Age","Gls.", "Touches."czy "Min."są kluczowe zarówno dla warunków, jak i dla liniowych modeli, co ponownie wskazuje na ich znaczącą rolę.

Zbiór	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Oryginalne wartości w zbiorze testowym	360,000	3,600,000	9,000,000	16,640,947	25,200,000	90,000,000
Predykcja wartości dla zbioru testowego	0	3,959,309	9,058,428	14,529,742	20,781,668	75,840,104

Tabela 3: Porównanie predykcji z oryginalnymi wartościami rynkowymi w zbiorze testowym

Jak widać model przewiduje wartości 0 dla zawodnika, co może martwić. Można też zauważyć, że model przewiduje ogólnie niższe wartości niż pierwotny model. Błąd bezwzględny wynosi około 6.5 mln, zatem w tych aspektach widać poprawę względem wstępnego modelu. Model także stał się bardziej 'stabilny'. Losując inne zbiory treningowe oraz testowe, model uzyskuje podobne wartości błędu bezwzględnego jak i cały czas 'stosunkowo' podobne rozkłady przewidywanych wartości z oryginalnymi. Biorąc jednak pod uwagę martwiącą przewidywaną wartość 0 (która powtarzała się przy wielu różnych zbiorach treningowych oraz testowych), oraz wciąż względnie duży błąd bezwzględny, rozważmy kolejne metody poprawy modelu.

6.2 Bagging, Lasy losowe oraz Boosting

Metody grupowania ...

Po zbudowaniu modelu, ważnym aspektem jest spojrzenie na najbardziej kluczowe zmienne w modelu. Można to przeanalizować za pomocą metryki $'\%IncMSE'$, która wskazuje jak bardzo wzras ta błąd modelu (średni kwadrat błędu), gdy dana zmienna jest wyłączona oraz metryki 'IncNodePurity' która mówi, jak bardzo zmienna zmniejsza wariancję w regresji. Zmienne o jedncyh z największych wartościach w obu z tych metryk to Age 7.665844e+13 2.389021e+16 2 Club 5.732063e+13 3.713208e+16 3 Gls. 2.564758e+13 1.647120e+16 4 Min. 1.792274e+13 8.846936e+15 5 Touches. 1.590154e+13 8.835604e+15 6 Ast.

7 Zmiana zmienneh zależnej

8 Podsumowanie