Drużyna gwiazd w liczbach

Jędrzej Ruciński

2 stycznia 2023

Spis treści

Rysunek 1: Allstar game 1978 - Atlanta

Rysunek 2: Allstar game 2000 - Oakland

Rysunek 3: Allstar game 2016 - Toronto

1 Wstęp

Półmetek sezonu w najlepszej lidze koszykarskiej na świecie to czas, w którym wszystko zaczyna się już powoli wyjaśniać. Po około 4 miesiącach gry, w których każda drużyna pokazała się w około 50 meczach rozrzuconych po całych Stanach Zjednoczonych, wiadomo które z nich trzymają ścisłą kontrolę nad pierwszymi pozycjami w tabeli. Wiadomo kto dalej walczy o awans do rozgrywek o mistrzostwo a kto już może sobie odpuścić końcówkę sezonu.

Poza tym okres ten związany jest z niezwykłym wydarzeniem jakim jest **Allstar weekend**. Trzy dniowa przerwa od rozgrywek podczas, której najlepsi indywidualni zawodnicy biorą udział w wszelakich konkurencjach, które mają zapewnić rozrywkę kibicom na całym świecie. Zwieńczeniem całego weekendu jest Mecz gwiazd (*Allstar game*), w którym biorą udział najlepsi gracze w bieżącym sezonie.

Co prawda wynik tego meczu nie jest w żadnym stopniu istotny a często nawet zawodnicy biorący w nim udział nie grają na 100% w obawie przed kontuzją. Za to kto bierze udział w tym meczu jest bardzo dużym wyróżnieniem dla wszystkich zawodników i co roku, przynajmniej w moim odczuciu, ciekawszy od samego meczu jest proces selekcji zawodników do niego, a w szczególności pierwszych piątek dla każdej konferencji (wschodniej i zachodniej).

Rysunek 4: Pierwsze piątki z sezonu 1988.

2 Przeprowadzane badania

Projekt będzie dotyczył właśnie wyżej wspomnianych pierwszych piątek - tak zwanych *starterów*. Celem jest na podstawie danych z wielu poprzednich sezonów NBA, być w stanie w nowym sezonie przewidzieć która piątka graczy będzie zaszczycona tym wyróżnieniem. Warto też wspomnieć, że w 1999 allstar weekend nie odbył się ze względu na ogólnoligowy lockdown, więc ten sezon nie będzie uwzględniany w badaniach.

2.1 Proces wyboru zawodników

By móc dokonywać takie właśnie predykcje musimy w pełni rozumieć jak wygląda selekcja tych najlepszych koszykarzy. Aktualnie jest to niestety bardzo skomplikowany proces. Odbywa się w formie głosowania, lecz prawa głosu mają trzy grupy - eksperci, zawodnicy/trenerzy oraz kibice. Ze względu na dużą dysproporcje co do wielkości tych grup głosy są oczywiście odpowiednio ważone i następnie łączone. Tu pojawia się dość duży problem ponieważ głosy zawodników, trenerów i ekspertów są tajne, tzn nie są ujawniane szerokiej publice. Na szczęście do 2016 roku liczyły się tylko głosy kibiców, które są regularnie udostępniane. Na tych właśnie danych się skupimy, zatem nasza predykcja będzie się opierała tylko na opinii fanów z całego świata, która, jak niektórzy mogliby stwierdzić, jest najważniejsza.

2.2 Dostępne miejsca

Jak sama nazwa wskazuje do najlepszej pierwszej piątki może się dostać pięciu zawodników. Na początku meczu gwiazd znajduję na boisku po pięciu najlepszych zawodników z każdej konferencji. System ten jest często krytykowany, ponieważ bywają lata gdzie jedna z konferencji prezentuje się o wiele lepiej do drugiej co wiąże się z brakiem wyróżnienia dla zawodników, którzy na takie zasłużyli, lecz nie zmieścili się w limicie miejsc. Zatem w części predykcyjnej tego projektu nie będzie brana pod uwagę konferencja w jakiej aktualnie gra dany zawodnik.

W skład pierwszej piątki wchodzi trzech zawodników pola ataku (*frontcourt*) oraz dwóch zawodników pola obrony (*backcourt*) i tego podziału będziemy się trzymać.

Ostatnią rzeczą, którą warto wziąć pod uwagę to fakt, że głosowanie do allstar odbywa się w tak zwanym półmetku sezonu, kiedy została rozegrana lekko ponad połowa meczy. W tym projekcie będę za to patrzył na całkowity sezon i na podstawie danych ze wszystkich meczy oceniał czy w danym roku konkretny zawodnik mógłby trafić do drużyny gwiazd. Dane z części sezonu są dosyć trudno dostępne, a do tego zawodnicy bardzo rzadka po "przerwie allstarowej" schodzą z konkretnego poziomu, na którym grali.

2.3 Uproszczenia

Zatem, podsumowując wszystkie uproszczenia przyjęte w tym projekcie prezentują się następująco:

- Brane pod uwagę są tylko głosowania do 2016 roku, które zawierały tylko głosy kibiców.
- W predykcjach nie są brane pod uwagę konferencje.
- Na statystyki z danego sezonu patrzymy całościowo, tak jakby głosowanie odbywało się po jego ukończeniu.

3 Dane

Dane użyte w tym projekcie pochodzą z dwóch źródeł.

3.1 Głosy

Dla każdego z sezonów wybrani zawodnicy do allstarowych piątek umieszczeni są na stronie Basketball Reference wraz z ilością zdobytych przez nich głosów. Za pomocą skryptu w języku `python` dane z pierwszej tabelki widocznej na stronie zostały ściągnięte do obiektów pandasowych oraz złączony w

duży plik csv zawierający te dane dla lat 1978-2016. Plik: *allstar_starters.csv*. Próbka z tego pliku:

Position	Player	Votes	Year
C	Bob McAdoo	98325.0	1975
G	Anfernee Hardaway	1050461.0	1996
F	LeBron James	2516049.0	2007
F	Blake Griffin	700615.0	2015

3.1.1 Skrypt pythonowy

Pozyskanie danych do formatu csv z wyżej wspomnianej strony okazało się jednak dosyć wymagające, gdyż umieszczone są one na stronie w postaci tabelki html, do której trzeba było się w jakiś sposób dostać. Skrypt ten powstał na wzór skryptu kanału Dataquestio, na którym przeprowadzany był projekt przewidujący MVP każdego sezonu NBA i również ściągane były dane ze strony Basketball Reference.

Skrypt przechodzi po wszystkich porządanych przez nas latach i dla każdego z nich wchodzi na odpowiednią stronę, następnie zapisując całą tą stronę do pliku allstars/year.

```
for year in years:
  url = f"https://www.basketball-reference.com/allstar/NBA_{year}_voting.html"
  data = requests.get(url)

  with open(f"allstars/{year}.html", "w+") as f:
    f.write(data.text)
```

Następnie musimy znaleźć jaki szukana przez nas ramka ma html-owy id.

Rysunek 5:

Używając tych id przechodzimy po każdej stronie i zapisujemy tabelkę html do ramki pandasowej, korzystając z biblioteki `BeatifulSoup` pozwalającej na szukanie elementów języka html po ich id.

```
dfs = []

for year in years:
   with open(f"allstars/{year}.html") as f:
    page = f.read()
    soup = BeautifulSoup(page, "html.parser")
    starters_E = soup.find(id = "all-star-starters_1")
    starters_W = soup.find(id = "all-star-starters_2")
    starters__E = pd.read_html(str(starters_E))[0]
    starters__E = starters__E.iloc[1:,:]
    starters__W = pd.read_html(str(starters_W))[0]
    starters__W = starters__W.iloc[1:,:]
    starters = pd.concat([starters__E,starters__W],axis = 0)
    starters = starters.reset_index()
    starters = starters.sort_values(2, ascending = False)
    starters["Year"] = year
    del starters["index"]

    dfs.append(starters)

allstar_starters = pd.concat(dfs)
allstar_starters = allstar_starters.reset_index()
del allstar_starters["index"]
allstar_starters.rename({0: 'Position', 1: 'Player', 2: 'Votes'}, axis = 1, inplace =
    True)
```

Ramki na koniec łączymy w dużą tabelkę pandasową, w której pozostało nam usunąć index oraz nadać nowe nazwy zmiennym. Na sam koniec całość exportujemy do pliku csv.

```
allstar_starters.to_csv("allstar_starters.csv")
```

3.2 Statystyki Zawodników

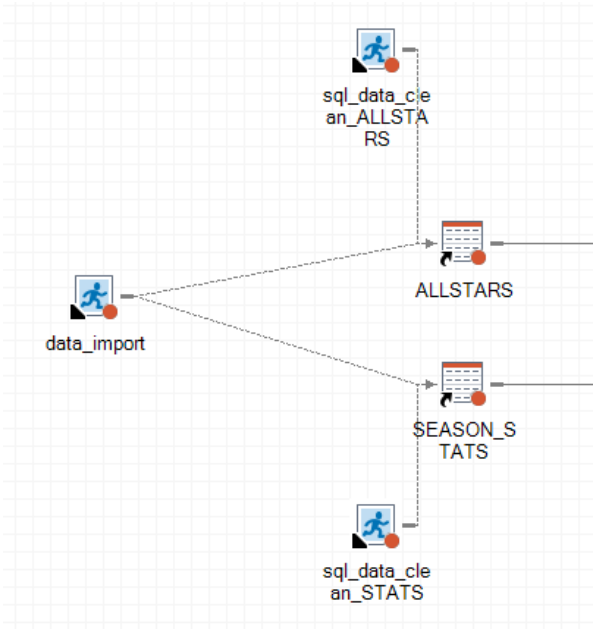
Aby porównać tych wyróżnionych zawodników do reszty potrzebujemy również statystyk każdego zawodnika w lidze dla każdego z sezonów 1978-2016. Dane te otrzymaliśmy ze strony **Kaggle**. Ramka którą użyłem została przygotowana przez użytkownika *OMRI GOLDSTEIN* i zawiera statystyki wszystkich zawodników od sezonu 1950. Dane znajdują się pod linkiem NBA Players stats since 1950. Plik: *Season_Stats_2.xlsx* Zawiera on multum informacji o każym zawodniku, który w tym okresie zagrał chociażby jeden mecz na parkietach NBA. Każda linijka jest podsumowaniem pojedynczego sezonu w wykonaniu jednego zawodnika. Najważniejsze informacje o nim to między innymi:

- **Games** - Liczba zagranych spotkań,
- **Points** - Liczba zdobytych punktów,
- **FG%** - Procent trafionych rzutów,
- **Assists** - Liczba asyst,
- **Total Rebounds** - Liczba zdobytych zbiórek

I wiele innych o wiele bardziej zaawansowanych statystyk, opisanych na stronie Kagglowej.

4 Przygotowanie danych

4.1 Import oraz pierwsze czyszczenie



Rysunek 6:

Dane z obu plików csv importujemy do projektu i za pomocą komendy `Proc SQL` dokonujemy pierwszego czyszczenia na obu tabelach (funkcje odpowiednio `sql_data_clean_` dla obu tabel).

4.1.1 sql_data_clean_ALLSTARS

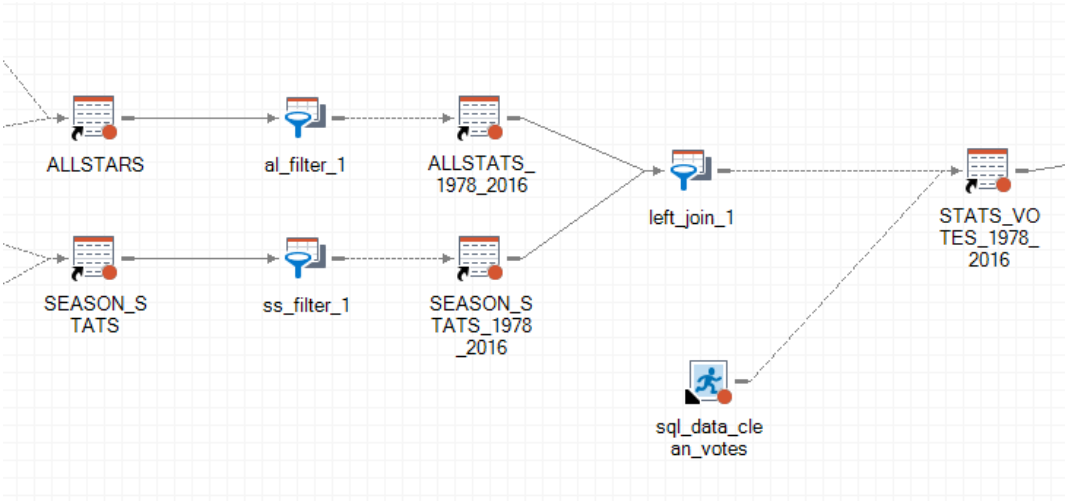
Jedyne zmiany jakie musimy zrobić w tej tabeli to uogólnienie pozycji zawodników z *G - guard, F - forward, C - Center* do pożądanych *F - frontcourt, B - backcourt*. Oznacza to po prostu zamianę pozycji $C \longrightarrow F, G \longrightarrow B$.

4.1.2 sql_data_clean_STATS

W tej tabeli wykonujemy już o wiele więcej zmian:

- W tabeli na Kaggle wszystkie statystyki liczbowe są podane sumarycznie, więc wprowadzamy tak zwane *Per Game Stats*, czyli po prostu daną statystykę podzieloną przez liczbę rozegranych meczy w tym sezonie. Robimy to dla Punktów, Asyst, Zbiórek, Przechwytów, Bloków, Strat i Fauli.
- Jeżeli w danym sezonie gracz grał w paru drużynach to w tabeli pojawia się wiele razy. Na szczęście dla wszystkich takich zawodników jest też pozycja TOT - czyli całkowite statystyki z tego sezonu. Zatem dla tych graczy usuwamy wszystkie inne pozycje prócz tych gdzie $Team = TOT$.
- Zawodnicy, którzy po ukończeniu kariery dostali się do **Basketball Hall of Fame** mają przy imieniu znaczek $*$. Jest to wyróżnienie za osiągnięcia w całej karierze czyli nie jest dla nas istotne, zatem usuwamy ten znaczek.
- Ostatecznie również sprowadzamy pozycje do *F - frontcourt, B - backcourt*.

4.2 Łączenie ramek i kolejne czyszczenie



Rysunek 7:

4.2.1 Wybór lat i łączenie tabel

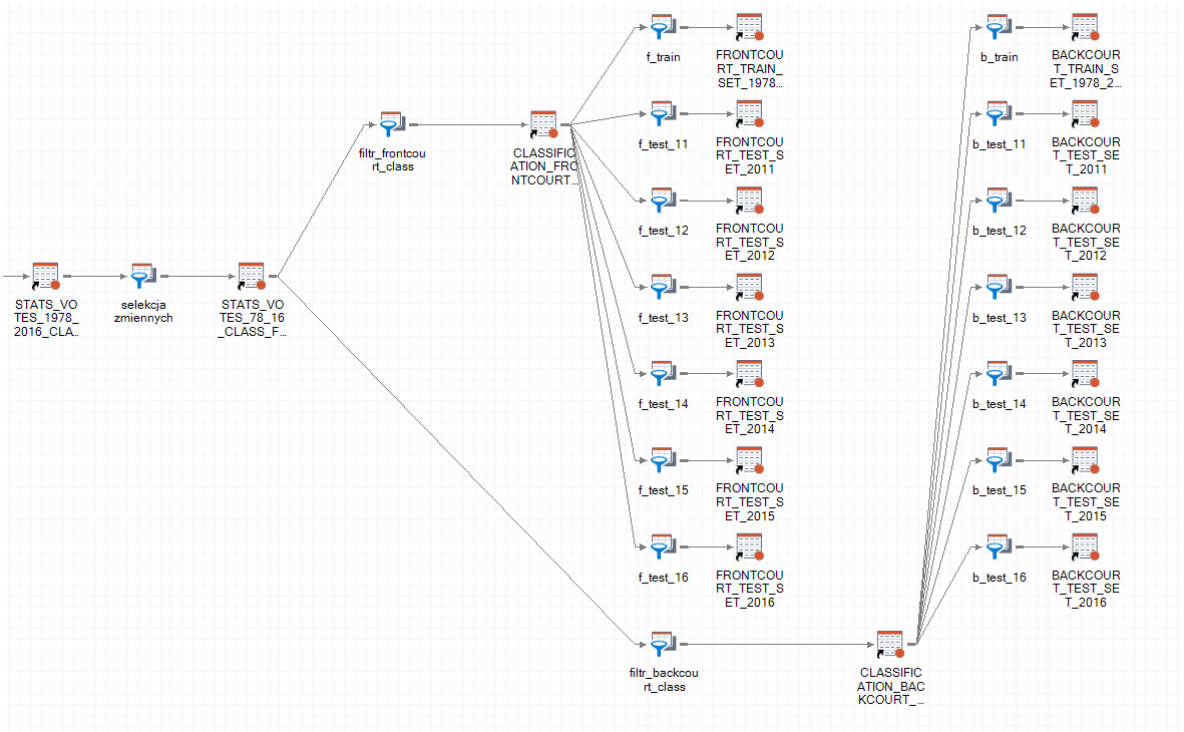
Dla obu uzyskanych ramek filtrujemy tak aby uzyskać spójny okres czasowy. Sezony 1978 - 2016 wybrałem ponieważ miały najbardziej wspólne dane co do liczenia statystyk (np. dopiero pod koniec lat 70 zaczęto liczyć przechwyty oraz wprowadzono linię rzutu za 3 punkty). Temu właśnie służą `al_filter_1` oraz `ss_filter_1`.

Następnie łączymy te tabele po kolumnach **Year** oraz **Player** gdyż dopiero razem tworzą one unikatowy klucz. W ten oto sposób otzymaliśmy ramkę *STATS_VOTES_1978_2016*, która po małym czyszczeniu będzie naszą główną bazą do tworzenia podzbiorów treningowych i uczących.

4.2.2 Końcowe zmiany SQL

Aktualnie nasza ramka zawiera statystyki danego zawodnika z każdego sezonu, w którym grał oraz liczbę głosów jaką dostał w przypadku gdy był wybrany do pierwszej piątki. Oznacza to, że zdecydowaną większość tej kolumny jest pusta gdyż tylko 10 zawodników na sezon ma tam konkretne wartości. Do tego liczby głosów fanów w ogólności znacząco rosną z roku na rok ze względu na ciągle rosnącą popularnością koszykówki na świecie. Model regresyjny przewidujący liczbę głosów nie ma więc zbyt dużego sensu i może zwrócić bardzo mylące dane. Wprowadzamy więc zmienną binarną *Starter* $\in \{0, 1\}$, przyjmującą wartość 1 gdy zawodnik został wybrany do pierwszej piątki i 0 w przeciwnym przypadku. Wszystko to dzieje się w funkcji `sql_data_clean_votes`.

5 Tworzenie zbiorów uczących i testowych



Rysunek 8:

5.1 Pierwsze podziały

W celu utworzenia zbiorów predykcyjnych pierwszym krokiem jest usunięcie zmiennej *Votes* mówiącej o liczbie głosów otrzymanej przez danego zawodnika. W przypadku tworzenia modelu klasyfikacji zmienna ta nie będzie nam potrzebna. Czyni to filter `remove_votes`.

Następnie dzielimy zbiór na graczy pola obrony (*Backcourt*) oraz graczy pola ataku (*Frontcourt*). Tak jak wspominałem w 2.1, wybory do pierwszych piątek odbywają się wewnątrz tych grup.

5.2 Zbiory treningowe i testowe

Bardzo ważnym spostrzeżeniem w całym tym projekcie jest to iż zawodnicy wybierani są do składów Allstar w obrębie jednego sezonu. Musimy więc brać to pod uwagę przy dokonywaniu podziałów na zbiory testowe i treningowe. Między innymi dlatego podział ten dzieje się już w obrębie środowiska SAS Enterprise Guide, gdyż nie możemy sobie pozwolić na typowo losowy podział jaki oferuje SAS Enterprise Miner. Mogło by to skutkować zbiorem testowym, w którym mamy, na przykład, tylko 3 zawodników z danego sezonu. Wprowadzamy więc następujący ręczny podział:

- Jako zbiór testowy w obu grupach (*Frontcourt* oraz *Backcourt*) przyjmujemy wszystkie sezony od 1978 aż do 2010 roku włącznie.
- Dla obu grup tworzymy 6 rozłącznych zbiorów testowych dla każdego z sezonów od 2011 do 2016 .

5.3 Wybór zmiennych

Nasze dane zawierają dalej dla każdego zawodnika ogromną liczbę statystyk, z których część jest od siebie w prost zależna, część nie jest dla nas zbytnio istotna, a część może nawet wprost psuć dokładność naszego modelu. Musimy więc przeprowadzić wstępną selekcje tych zmiennych.

Dokładny opis wszystkich aktualnie rozważanych zmiennych znajduję się w pliku PDF o nazwie `Charakterystyka zmiennych`.

Sprowadzamy więc tą obszerną listę do następujących zmiennych:

- **Year** - wskazuje sezon
- **Player** - zawodnik
- **Pos** - pozycja zawodnika (F,B)
- **G** - liczba rozegranych gier w danym sezonie
- **MP** - liczba rozegranych minut w danym sezonie
- **PER** - *Player Efficiency Rating*, zaawansowana statystyka mówiąca o wydajności danego zawodnika; Rozwinięty opis
- **TS%** - *True Shooting Percentage*, zaawansowana statystyka mówiąca o celności danego zawodnika; Rozwinięty opis
- **USG%** - *Usage Percentage*, procent akcji przechodzących przez zawodnika
- **OWS** - *Offensive Win Share*, zaawansowana statystyka, liczy przyczynienie się do skutecznych akcji ofensywnych
- **DWS** - *Deffensive Win Share*, zaawansowana statystyka, liczy przyczynienie się do skutecznych akcji defensywnych
- **BPM** - *Box Plus Minus*, liczba punktów zyskanych gdy zawodnik był na boisku
- **FGA** - *Field goals attempted*, oddane rzuty z gry
- **FG%** - procent trafionych rzutów z gry
- **2PA** - oddane rzuty za dwa punkty
- **2P%** - procent trafionych rzutów za dwa punkty
- **3PA** - oddane rzuty za trzy punkty
- **3P%** - procent trafionych rzutów za trzy punkty
- **FTA** - oddane rzuty wolne
- **FT%** - procent trafionych rzutów wolnych
- **PPG** - *Points Per Game*
- **APG** - *Assists Per Game*
- **RPG** - *Rebounds Per Game*
- **SPG** - *Steals Per Game*
- **BPG** - *Blocks Per Game*

- **TPG** - *Turnovers Per Game*
- **FPG** - *Fouls Per Game*
- **Starter** - zmienna wyznaczająca czy w tym sezonie zawodnik dostał się do pierwszej piątki allstar (0 - nie, 1 - tak)

Charakterystyka wyżej wymienionych zmiennych prezentuję się następująco:

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
2PA	2PA	15577	0	6108769.00	0.0	392.17	300.00	2213.00	2.82915

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
3P%	3P%	15577	0	3069.88	0.0	0.20	0.22	1.00	0.00148

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
3PA	3PA	15577	0	1036661.00	0.0	66.55	11.00	886.00	0.84768

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
APG		15577	0	30206.59	0.0	1.94	1.33	14.54	0.01513

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
BPG		15577	0	6701.81	0.0	0.43	0.25	5.56	0.00421

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
BPM	BPM	15577	0	-30640.80	-86.7	-1.97	-1.60	36.20	0.03561

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
DWS	DWS	15577	0	21066.50	-1.0	1.35	1.00	9.10	0.01012

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
FG%	FG%	15532	45	6884.81	0.0	0.44	0.45	1.00	0.00070

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
FGA	FGA	15577	0	7145430.00	0.0	458.72	368.00	2279.00	3.14214

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
FPG		15577	0	31259.56	0.0	2.01	2.00	6.00	0.00694

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
FT%	FT%	15162	415	10973.81	0.0	0.72	0.75	1.00	0.00113

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
FTA	FTA	15577	0	2232538.00	0.0	143.32	99.00	972.00	1.15300

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
G	G	15577	0	865900.00	1.0	55.59	64.00	85.00	0.20092

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
OWS	OWS	15577	0	22391.50	-3.3	1.44	0.70	15.20	0.01732

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
PER	PER	15574	3	199028.70	-90.6	12.78	12.90	129.10	0.04661

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
PPG		15577	0	132352.96	0.0	8.50	7.00	37.09	0.04887

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
RPG		15577	0	56761.21	0.0	3.64	2.97	18.66	0.02087

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
SPG		15577	0	10802.56	0.0	0.69	0.60	3.67	0.00392

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
Starter		15577	0	378.00	0.0	0.02	0.00	1.00	0.00123

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
TPG		15577	0	20563.75	0.0	1.32	1.14	5.67	0.00672

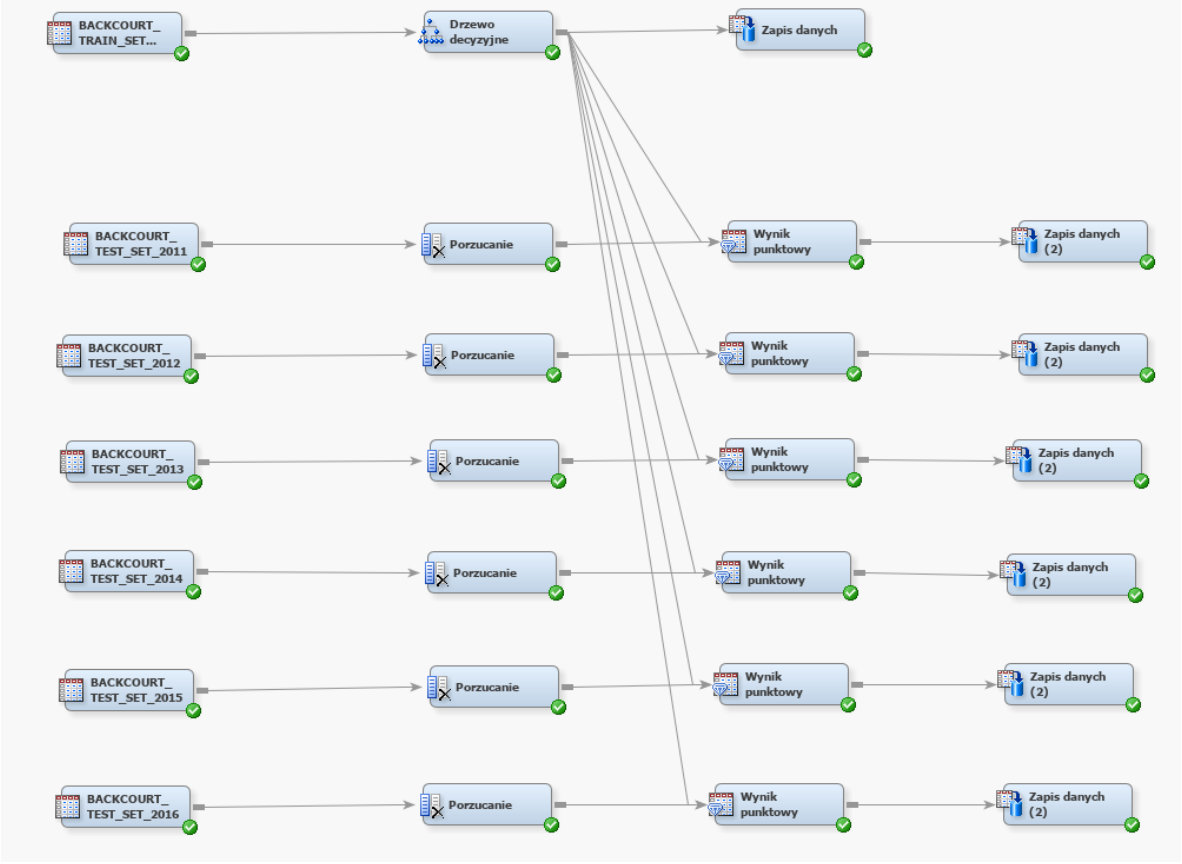
Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
TS%	TS%	15539	38	7888.02	0.0	0.51	0.52	1.14	0.00068

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
USG%	USG%	15574	3	295367.10	0.0	18.97	18.70	100.00	0.04270

Variable	Label	N	NMiss	Total	Min	Mean	Median	Max	StdMean
Year	Year	15577	0	31134531.00	1978.0	1998.75	1999.00	2016.00	0.08759

6 Drzewo Decyzyjne

Pierwszym modelem, który stworzyłem jest znane wszystkim bardzo dobrze, lecz niekoniecznie przez wszystkich lubiane **drzewko decyzyjne**. Drzewo inicjowane jest według poniższego schematu:



Rysunek 9:

Identyczne proces generowany jest dla zawodników *Frontcourt*. Dane treningowe wrzucane są do modelu drzewa o roli wyłącznie danych uczących i wynik przewidywań na tym zbiorze od razu zapisywany jest do odpowiedniej biblioteki w postaci tabeli SASowej. Następnie dla każdego zbioru testowego z osobna usuwana jest kolumna *Starter* i liczony jest wynik punktowy na podstawie tych danych oraz wygenerowanego drzewa. Te poszczególne wynik również są zapisywane w postaci Tabel.

6.1 Pierwszy run, ustawienia domyślne

Model, w którym wykorzystane został domyślne ustawienia Enterprise Minera zwrócił ciekawe wyniki, lecz nieoptymalne. Rozważmy predykcję dla obu grup pozycji, w porównaniu do rzeczywistych starterów.

6.1.1 Backcourt

Rok	Faktyczni starterzy	Przewidziani starterzy
2011	Kobe Bryant, Dwayne Wade, Chris Paul, Derrick Rose	Kobe Bryant, Dwayne Wade
2012	Kobe Bryant, Derrick Rose, Dwayne Wade, Chris Paul	Kobe Bryant
2013	Kobe Bryant, Dwayne Wade, Rajon Rondo, Chris Paul	Kobe Bryant, Dwayne Wade
2014	Kobe Bryant, Stephen Curry, Dwayne Wade, Kyrie Irving	
2015	Kobe Bryant, Stephen Curry, John Wall, Kyle Lowry	Kobe Bryant, Dwayne Wade
2016	Dwayne Wade, Stephen Curry, Kyle Lowry, Russel Westbrook	Dwayne Wade

Jak widać przewidywani zawodnicy są bardzo podobni z roku na rok, i pomimo tego iż można by dyskutować godzinami o zasługach Koby'ego oraz Dwade'a, w poszczególnych latach byli zawodnicy, którzy o wiele bardziej zasługiwali na te pozycje.

Sprawdźmy więc, które zmienne drzewo brało pod uwagę w tej predykcji:

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance
Player	Player	1	1.0000
FTA	FTA	1	0.4107
SPG		1	0.2390
FPG		1	0.2343
RPG		1	0.2091
TS_	TS%	1	0.1903
FGA	FGA	1	0.1824
FT_	FT%	1	0.1633
PPG		1	0.1513

Rysunek 10:

Całe szczęście, powyższy rysunek dużo wyjaśnia. Model jako najważniejszą zmienną w predykcji uznał imiona zawodników, co wyjaśnia tak częste przewidywanie Koby Bryanta oraz Dwayna Wade'a, którzy w zbiorze treningowym (lata 1978-2010) mogli pojawiać się najczęściej jako starterzy. Tłumaczy to również tak małą liczbę przewidywanych zawodników każdego roku, gdyż według modelu nikt inny nie zasłużył na to wyróżnienie.

6.1.2 Frontcourt

Rok	Faktyczni starterzy	Przewidziani starterzy
2011	Dwight Howard, Lebron James, Amar'e Stoudemire, Yao Ming, Kevin Durant, Carmelo Anthony	Dwight Howard, Lebron James, Tim Duncan, Kevin Love
2012	Dwight Howard, Lebron James, Carmelo Anthony, Blake Griffin, Kevin Durant, Andrew Bynum	Dwight Howard, Lebron James, Tim Duncan, Kevin Love
2013	Dwight Howard, Lebron James, Carmelo Anthony, Blake Griffin, Kevin Durant, Kevin Garnett	Anderson Varejao, Lebron James, Tim Duncan, Kevin Durant
2014	Paul George, Lebron James, Carmelo Anthony, Blake Griffin, Kevin Durant, Kevin Love	Andre Drummond, Kevin Durant, Kevin Love, Lamarcus Aldridge, Lebron James, Tim Duncan
2015	Pau Gasol, Lebron James, Carmelo Anthony, Blake Griffin, Anthony Davis, Marc Gasol	Andre Drummond, Lebron James, Tim Duncan, Deandre Jordan
2016	Paul George, Lebron James, Carmelo Anthony, Kobe Bryant, Kevin Durant, Kawhi Leonard	Andre Drummond, Lebron James

W tym przypadku możemy zauważyć, że predykcje są bardziej trafne, lecz dalej młodzi zawodnicy wydają się mieć mniejszą szanse na kwalifikacje. Przykładem jest Kevin Durant, który już w 2010 roku znajdywał się na szczycie ligi, lecz u nas pojawił się dopiero w 2013 roku, gdy części starszych zawodników już nie było.

Sprawdźmy co tym razem brało pod uwagę drzewo:

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance
PER	PER	1	1.0000
Player	Player	1	0.8194
_2PA	2PA	2	0.3656
FTA	FTA	2	0.3039
BPM	BPM	2	0.2953
RPG		2	0.2911
FGA	FGA	1	0.2882
FG_	FG%	1	0.2124
Year	Year	1	0.2116
PPG		1	0.2059
FPG		1	0.2046
TPG		1	0.1560
DWS	DWS	1	0.1373

Rysunek 11:

Nazwiska zawodników nie miały tym razem największej wagi w predykcjach, lecz wciąż wysoką. Przeważyło jednak *Player Efficiency Rating*, jedna z bardziej zaawansowanych statystyk dostępnych aktualnie na stronach NBA. Optymalnie chcieli byśmy jednak aby model nie brał w ogóle pod uwagę imion zawodników, choć jest to dosyć "realistyczny" parametr. Przy głosowaniu kibiców często większy wpływ ma, niestety, renoma i marka zawodnika niż aktualny poziom, który pokazuje na boisku.

6.2 Wskazane zmienne

Tym razem zmusiliśmy model do wykorzystania wszystkich zmiennych znajdujących się w zbiorze uczącym. Do tego zabroniliśmy korzystania ze zmiennej *Player*, czyli imion i nazwisk zawodników. Zobaczmy jakie wyniki tym razem otrzymaliśmy.

6.2.1 Backcourt

Rok	**Faktyczni starterzy**	**Przewidziani starterzy**
2011	Kobe Bryant, Dwayne Wade, Chris Paul, Derrick Rose	
2012	Kobe Bryant, Derrick Rose, Dwayne Wade, Chris Paul	Kobe Bryant, Dwayne Wade, Derrick Rose
2013	Kobe Bryant, Dwayne Wade, Rajon Rondo, Chris Paul	Kobe Bryant, James Harden
2014	Kobe Bryant, Stephen Curry, Dwayne Wade, Kyrie Irving	
2015	Kobe Bryant, Stephen Curry, John Wall, Kyle Lowry	James Harden, Jimmy Butler, Russel Westbrook
2016	Dwayne Wade, Stephen Curry, Kyle Lowry, Russel Westbrook	James Harden, Jimmy Butler, Russel Westbrook, Kemba Walker, Stephen Curry

Jak widać wyniki są o tyle lepsze, iż zaczynają się pokazywać coraz młodsi zawodnicy, którzy rzeczywiście zasługują na te pozycje. Absencja Stephena Curriego z większości lat może być zaskakująca, lecz wynika ona z faktu iż model trenowaliśmy na danych z dawnych lat, w których jego niepodważalnie najlepsza cecha (rzuty za 3) nie była aż tak wykorzystywana, przez co model nie brał jej pod uwagę.

Sprawdźmy więc, które zmienne i z jaką wagą były wykorzystywane:

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance
PPG		3	1.0000
BPM	BPM	1	0.5057
APG		1	0.4109
BPG		2	0.3123
TPG		1	0.3024
FTA	FTA	1	0.2815
_2PA	2PA	1	0.2717
USG_	USG%	1	0.2591
FG_	FG%	1	0.2109

Rysunek 12:

Zatem tak jak przepuszczałem, ze względu na roczniki, na których trenowaliśmy dane, rzuty za trzy punkty nie są kompletnie brane pod uwagę. Z drugiej strony zmienna taka jak bloki - *Blocks Per Game* ma zaskakująco dużą wagę, jak na to, że wysokie liczby w tej kategorii często nie są osiągane przez graczy na pozycjach *Backcourt*.

6.2.2 Frontcourt

Rok	Faktyczni starterzy	Przewidziani starterzy
2011	Dwight Howard, Lebron James, Amar'e Stoudemire, Yao Ming, Kevin Durant, Carmelo Anthony	Dwight Howard, Lebron James, Kevin Durant
2012	Dwight Howard, Lebron James, Carmelo Anthony, Blake Griffin, Kevin Durant, Andrew Bynum	Dwight Howard, Lebron James, Blake Griffin, Kevin Durant, Paul Milsap
2013	Dwight Howard, Lebron James, Carmelo Anthony, Blake Griffin, Kevin Durant, Kevin Garnett	Lebron James, Tim Duncan, Kevin Durant
2014	Paul George, Lebron James, Carmelo Anthony, Blake Griffin, Kevin Durant, Kevin Love	Blake Griffin, Kevin Durant, Carmelo Anthony, Kevin Love, Lebron James
2015	Pau Gasol, Lebron James, Carmelo Anthony, Blake Griffin, Anthony Davis, Marc Gasol	Andre Drummond
2016	Paul George, Lebron James, Carmelo Anthony, Kobe Bryant, Kevin Durant, Kawhi Leonard	Kevin Durant, Lebron James

Tutaj również otrzymaliśmy trochę lepsze wyniki. Pojawili się nieco młodsi zawodnicy, tacy jak Blake Griffin. Za to jedyny zawodnik wyznaczony w 2015 to Andre Drummond, co jest całkiem zaskakujące gdyż nie dostał się on w tym roku nawet do całej drużyny allstars, a co dopiero do pierwszej piątki.

Tym razem zmienne wykorzystane prezentują się następująco:

Variable Importance

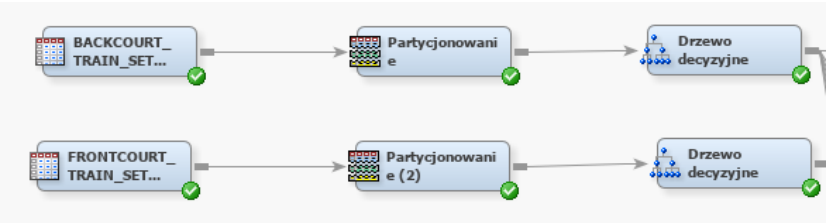
Variable Name	Label	Number of Splitting Rules	Importance
PER	PER	3	1.0000
TPG		2	0.5889
DWS	DWS	2	0.3512
FPG		2	0.2492
FTA	FTA	1	0.2464
OWS	OWS	1	0.2280
APG		1	0.2081
BPM	BPM	1	0.2062
_2PA	2PA	1	0.2021
TS_	TS%	1	0.1776

Rysunek 13:

W tym przypadku możemy zauważyć, że przeważyły zaawansowane statystyki takie jak *Player Efficiency Rating*, *Defensive & Offensive Winshare* oraz *Box Plus Minus*.

6.3 Walidacja

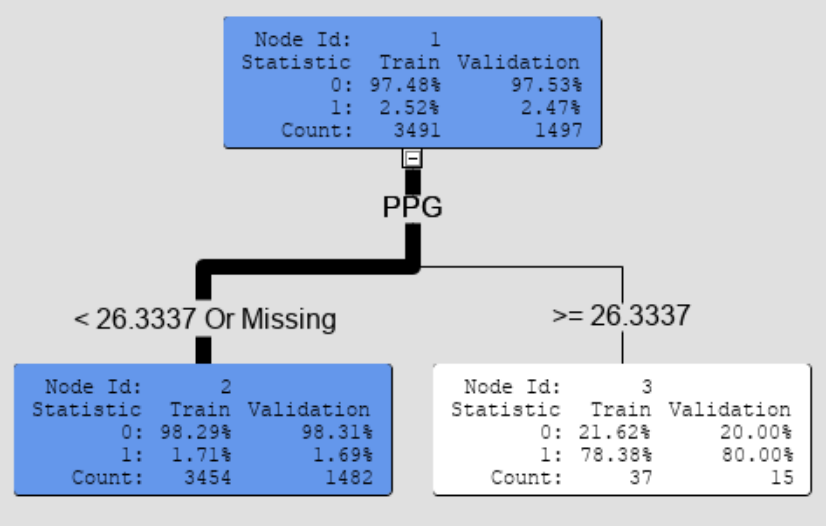
Tym razem wprowadzimy podział zbioru uczącego, wyznaczając 30% z niego jako zbiór walidacyjny. Pozwoli to minerowi na obcięcie drzewa do optymalnej głębokości oraz wybranie jedynie optymalnych zmiennych. Dalej jednak rezygnujemy z korzystania zmiennej *Player* aby przysłowiowo *nie oceniać książki po okładce*.



Rysunek 14:

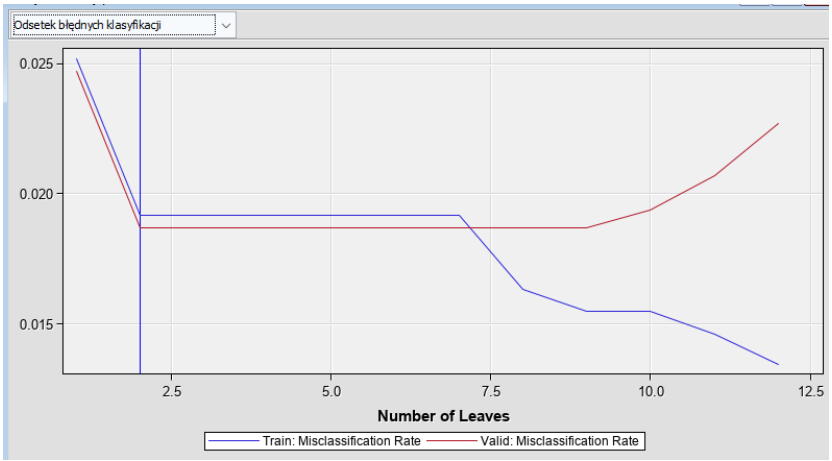
6.3.1 Backcourt

Co ciekawe przy użyciu zbioru walidacyjnego drzewo zostało drastycznie obcięte i korzysta z tylko jednej zmiennej:



Rysunek 15:

Moment tego "optymalnego" odcięcia możemy odczytać z wykresu oceny poddrzewa, na którym oś pionowa oznacza procent błędnie sklasyfikowanych pozycji:



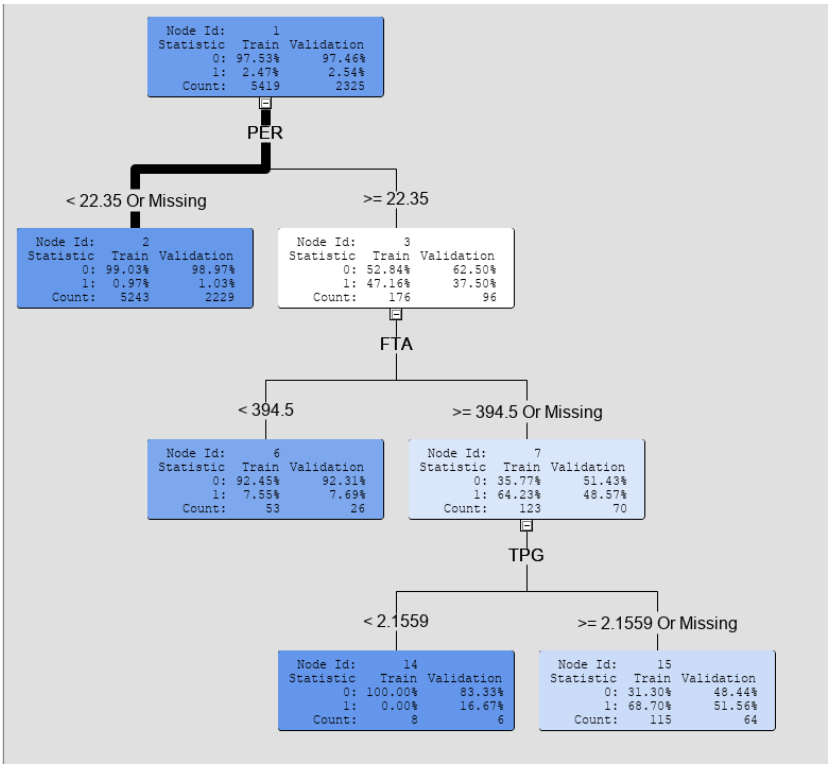
Rysunek 16:

Jak można by się spodziewać, w rzeczywistości taki podział nie jest najlepszy gdyż aktualnie niewiele zawodników osiąga średnią punktową powyżej 26, a do tego wyróżniają się w innych statystykach. W każdym razie predykcje wyglądają następująco:

Rok	Faktyczni starterzy	Przewidziani starterzy
2011	Kobe Bryant, Dwayne Wade, Chris Paul, Derrick Rose	
2012	Kobe Bryant, Derrick Rose, Dwayne Wade, Chris Paul	Kobe Bryant
2013	Kobe Bryant, Dwayne Wade, Rajon Rondo, Chris Paul	Kobe Bryant
2014	Kobe Bryant, Stephen Curry, Dwayne Wade, Kyrie Irving	
2015	Kobe Bryant, Stephen Curry, John Wall, Kyle Lowry	James Harden, Russel Westbrook
2016	Dwayne Wade, Stephen Curry, Kyle Lowry, Russel Westbrook	James Harden, Russel Westbrook

6.3.2 Frontcourt

W przypadku tych zawodników drzewko zostało trochę łagodniej potraktowane, a wykorzystane został jedynie zmienne *Player Efficiency Rating*, *Free Throw Attempts*, *Turnovers Per Game*. Drzewko prezentuje się następująco:



Rysunek 17:

Spójrzmy również na optymalny moment obcięcia drzewa. Wykres przedstawia znów zależności liczby liści od procentu błędnie sklasyfikowanych pozycji:

Otrzymujemy więc następujące predykcje:

Rysunek 18:

Rok	**Faktyczni starterzy**	**Przewidziani starterzy**
2011	Dwight Howard, Lebron James, Amar'e Stoudemire, Yao Ming, Kevin Durant, Carmelo Anthony	Amar'e Stoudemire, Dwight Howard, Lebron James, Kevin Durant
2012	Dwight Howard, Lebron James, Carmelo Anthony, Blake Griffin, Kevin Durant, Andrew Bynum	Dwight Howard, Lebron James, Blake Griffin, Kevin Durant, Kevin Love
2013	Dwight Howard, Lebron James, Carmelo Anthony, Blake Griffin, Kevin Durant, Kevin Garnett	Lebron James, Blake Griffin, Kevin Durant, Carmelo Anthony
2014	Paul George, Lebron James, Carmelo Anthony, Blake Griffin, Kevin Durant, Kevin Love	Blake Griffin, Kevin Durant, Carmelo Anthony, Kevin Love, Lebron James, DeMarcus Cousins
2015	Pau Gasol, Lebron James, Carmelo Anthony, Blake Griffin, Anthony Davis, Marc Gasol	Blake Griffin, DeMarcus Cousins, Lebron James
2016	Paul George, Lebron James, Carmelo Anthony, Kobe Bryant, Kevin Durant, Kawhi Leonard	Kevin Durant, Lebron James, DeMarcus Cousins

Jak widać walidacja na zbiorze zawodników *Froncourt*, tzn głównie centrów oraz skrzydłowych przyniosła o wiele lepsze wyniki niż na zbiorze *Backcourt*. Jest to jednak uzasadnione, znowu przez strukturę zbioru uczącego. Pozycje tzw. podkoszowe nie zmieniły się aż tak drastycznie jak role rozgrywających. W latach 2011-2016 można z przymrużeniem oka powiedzieć, że te same cechy czynią dobrego gracza podkoszowego co w latach 1978-2010. Jednak w przypadku rozgrywających jest to kompletnie inna sprawa.

7 K najbliższych sąsiadów

Drugi trenowany przeze mnie model to **K najbliższych sąsiadów**. Model ten pozwoli na określenie prawdopodobieństwa każdego z graczy na uzyskanie pozycji w pierwszej piątce drużyny allstar. Metodę do liczenia odległości pomiędzy pozycjami ustawiamy jako *Random Decision Tree*. Ograniczamy również model do 10 sąsiadów, oraz liczbę kubełków na 8.

Dane uczące znów dzielimy wyznaczając 50% z nich jako dane walidacyjne. Następnie liczymy wynik punktowy za pomocą każdego ze zbiorów testowych 2011 - 2016.

W ten sposób otrzymaliśmy następujące wyniki:

7.1 Backcourt

7.1.1 Poprawność na danych walidacyjnych

Data Role=VALIDATE Target Variable=Starter Target Label=Starter

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	98.0606	99.7944	2427	97.3136
1	0	1.9394	77.4194	48	1.9246
0	1	26.3158	0.2056	5	0.2005
1	1	73.6842	22.5806	14	0.5613

Rysunek 19:

Tak jak mogliśmy się spodziewać najliczniejszą grupą są poprawnie przewidziane wartości 0 zmiennej *Starter*. Za to ponad 77% procent starterów nie zostało poprawnie sklasyfikowanych co może być bardzo niepokojące, lecz pamiętajmy, że te dana nie zamykają się w obrębie jednego sezonu, skąd może wynikać ta dysproporcja.

7.1.2 Predyckje

Sprawdźmy więc faktyczne nazwiska, które osiągnął najwyższe prawdopodobieństwo dostania się do pierwszej piątki. Tabelki są kolejno dla lat 2011-2016.

	Player	EM_EVENTPROBABILITY
1	Kobe Bryant	0.6
2	Russell Westbrook	0.3
3	Derrick Rose	0.2
4	Monta Ellis	0.2
5	Dwyane Wade	0.2

	Player	EM_EVENTPROBABILITY
1	Marcus Thornton	0.1
2	O.J. Mayo	0.1
3	Nick Young	0.1
4	Lou Williams	0.1
5	Jason Terry	0.1

	Player	EM_EVENTPROBABILITY
1	James Harden	0.3
2	Kobe Bryant	0.3
3	DeMar DeRozan	0.2
4	Russell Westbrook	0.2
5	J.J. Barea	0.1

	Player	EM_EVENTPROBABILITY
1	James Harden	0.3
2	Arron Afflalo	0.2
3	Goran Dragic	0.2
4	DeMar DeRozan	0.2
5	Monta Ellis	0.1

	Player	EM_EVENTPROBABILITY
1	DeMar DeRozan	0.4
2	Russell Westbrook	0.3
3	James Harden	0.2
4	Giannis Antetokounmpo	0.2
5	Monta Ellis	0.1

	Player	EM_EVENTPROBABILITY
1	James Harden	0.3
2	Giannis Antetokounmpo	0.3
3	DeMar DeRozan	0.2
4	Jimmy Butler	0.2
5	Isaiah Thomas	0.1

Możemy zauważyć, iż wkrada się tu wiele nazwisk, których nie było wcześniej, lecz wiele z nich miało zaskakująco dobre sezony i osiągnęli dobre liczby w wielu statystykach. Przykładami są tacy zawodnicy jak Giannis Antetokuompo, James Harden, Russel Westbrook oraz Demar Derozan, którzy bardzo mocno skorzystali z faktu, iż algorytm KNN korzystał ze wszystkich podanych zmiennych.

Z drugiej strony rok 2012 jest bardzo zaskakujący. Nie został tam poprawnie przewidziany ani jeden zawodnik, a do tego wskazani koszykarze raczej nie mieli w tym roku prawa być nawet w konwersacji najlepszych w swojej roli.

7.2 Froncourt

7.2.1 Poprawność na danych walidacyjnych

Data Role=VALIDATE Target Variable=Starter Target Label=Starter

Target	Outcome	Target Percentage	Outcome Percentage	Frequency Count	Total Percentage
0	0	97.6672	99.7881	3768	97.2889
1	0	2.3328	92.7835	90	2.3238
0	1	53.3333	0.2119	8	0.2066
1	1	46.6667	7.2165	7	0.1807

Rysunek 20:

Znów najliczniejszą grupą są wartości {0,0}, ze względu na niezbilansowanie zbioru uczącego. Tym razem fakt iż walidacja wykonywana jest na okresie wielu lat i do tego partycja danych przeprowadzana jest losowo, miało ogromny wpływ na poprawne przewidzenie starterów, gdyż jedynie 2% z nich uzyskało poprawną wartość.

7.2.2 Predykcje

Sprawdźmy więc faktyczne nazwiska, które osiągnęły najwyższe prawdopodobieństwo dostania się do pierwszej piątki. Tabelki są kolejno dla lat 2011-2016.

	Player	EM_EVENTPROBABILITY
1	Dwight Howard	0.7
2	Kevin Durant	0.4
3	Amar'e Stoudemire	0.4
4	Blake Griffin	0.4
5	Al Jefferson	0.3
6	LaMarcus Aldridge	0.3
7	Carmelo Anthony	0.3
8	Chris Bosh	0.3
9	LeBron James	0.3

	Player	EM_EVENTPROBABILITY
1	LeBron James	0.2
2	Blake Griffin	0.2
3	Carmelo Anthony	0.1
4	Dwight Howard	0.1
5	Dirk Nowitzki	0.1
6	Kevin Garnett	0.1
7	DeMarcus Cousins	0.1
8	Kevin Love	0.1
9	Tristan Thompson	0.1

	Player	EM_EVENTPROBABILITY
1	LaMarcus Aldridge	0.3
2	Dwight Howard	0.2
3	Greg Monroe	0.2
4	Tim Duncan	0.2
5	Kevin Durant	0.2
6	DeMarcus Cousins	0.1
7	Ersan Ilyasova	0.1
8	Ryan Anderson	0.1
9	Blake Griffin	0.1

	Player	EM_EVENTPROBABILITY
1	Kevin Durant	0.5
2	Al Jefferson	0.3
3	Carmelo Anthony	0.3
4	Blake Griffin	0.3
5	LaMarcus Aldridge	0.2
6	DeMarcus Cousins	0.2
7	Anthony Davis	0.1
8	Dwight Howard	0.1
9	Gordon Hayward	0.1

	Player	EM_EVENTPROBABILITY
1	Anthony Davis	0.3
2	DeMarcus Cousins	0.2
3	Andrew Wiggins	0.2
4	LaMarcus Aldridge	0.2
5	Kevin Love	0.1
6	Nikola Vucevic	0.1
7	Gordon Hayward	0.1
8	Blake Griffin	0.1
9	Marc Gasol	0.1

	Player	EM_EVENTPROBABILITY
1	Steven Adams	0.1
2	Kevin Love	0.1
3	Kawhi Leonard	0.1
4	Marcin Gortat	0.1
5	Jahlil Okafor	0.1
6	Gordon Hayward	0.1
7	Anthony Davis	0.1
8	Paul Millsap	0.1
9	Brook Lopez	0.1

Znów widzimy wiele nowych nazwisk, często młodych zawodników rozpoczynających dopiero swoją przygodę w NBA. Znów wiele z nich jak najbardziej mogło by znaleźć się w tym elitarnym gronie starterów. Zaskakująca za to jest nieobecność w większości lat Lebrona James'a, przez wielu uważanych za niepodważalnie najlepszego zawodnika naszych czasów.

8 Podsumowanie

Możemy więć jasno stwierdzić, że drzewa decyzyjne poradziły sobie z tym zadaniem znacznie lepiej niż metoda k najbliższych sąsiadów, która nawet przy innych ustawieniach i metrykach dalej okazywała się dosyć zawodna.

Jednym jednak z niewielu plusów metody KNN jest fakt, iż mogliśmy zauważyć, że gdy wszystkie statystyki są brane pod uwagę nowi zawodnicy dostają szansę na wyróżnienie. Oczywiście jest to mało realna sytuacja, gdyż w prawdziwym NBA mało rzeczy jest aż tak ważnych jako samo zdobywanie punktów, co trafnie pokazało nam drzewo ze zbiorem walidacyjnym.

Ogólnie rzecz biorąc powstałe modele stworzyły, moim zdaniem, bardzo ciekawy i zarazem kompletnie nowy punkt spojrzenia na proces selekcji najlepszych zawodników danego sezonu. Z niecierpliwością czekam więc na tego roczny allstar weekend aby móc przetestować algorytm i porównać wyznaczonych zawodników z rzeczywistymi wyborami!

Rysunek 21:

Literatura

[1] dr hab. inż. Maciej Grzenda. Wykłady z przedmiotu *Wybrane algorytmy i systemy analizy danych*. 2022

[2] Basketball Reference https://www.basketball-reference.com/

[3] Dokumentacja SAS https://support.sas.com/en/documentation.html

[4] Dataquest https://www.youtube.com/c/Dataquestio