NBA HallOfFame Prediction

SAS Model Builder

Bartłomiej Wójcik

Politechnika Warszawska, Wydział Matematyki i Nauk Informacyjnych

29.12.2024

Spis treści

1. Opis problemu

Celem niniejszego projektu jest opracowanie modelu predykcyjnego, który umożliwia przewidywanie, czy zawodnik NBA ma szansę na dostanie się do Hall of Fame. Wyróżnienie to jest zarezerwowane dla jedynie najbardziej wybitnych zawodników, trenerów, którzy mieli duży wpływ na rozwój koszykówki oraz osiągneli w lidze NBA bardzo wiele. Decyzja o przyjęciu do Hall of Fame opiera się na analizie różnych czynników, takich jak osiągnięcia sportowe, liczba zdobytych tytułów, statystyki indywidualne czy długość kariery.

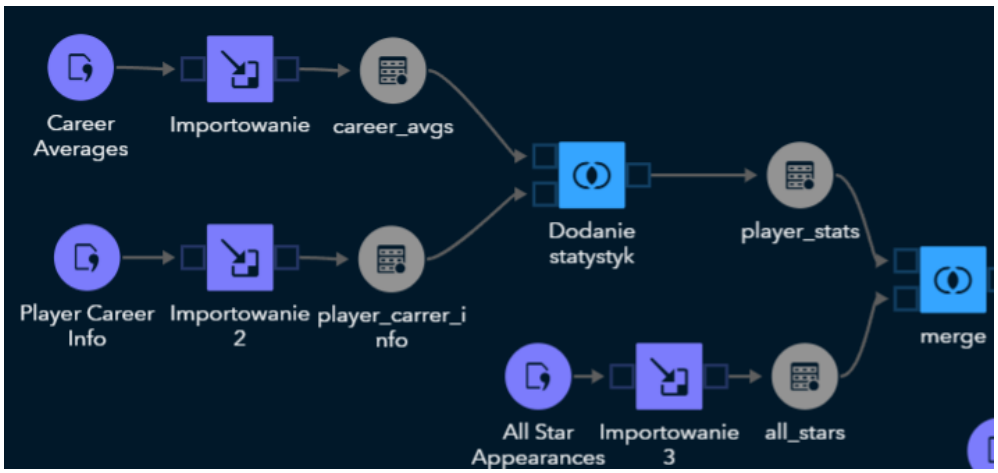
W ramach projektu, przy użyciu danych o zawodnikach NBA, postaram sie stworzyć modele, które zarówno przyporządkowują graczy do HallOfFame lub nie, a także takie, które obliczają ich procentowe szanse na dostanie się do tego prestiżowego grona. Nasza zmienna celu 'hof' jest zmienną binarną przyjmującą wartości TRUE lub FALSE.

2. Przygotowanie zbioru danych

Dane, które posłużyły mi do stworzenia docelowego zbioru danych, pochodzą z kaggle.com, są przechowywane w postaci plików CSV oraz są stale updatowane. Zmienne, których potrzebowałem znajdowały się w wielu plikach, co wymagało wstępnego ich przygotowania do załadowania do środowiska SASowego, a następnie połączeniu ich po ID zawodnika. Dla łatwiejszego dostępu do danych, pliki CSV, z których korzystałem umieściłem w repozytorium (można tam również znaleźć linki do źródeł danych.
Link do repozytorium z danymi

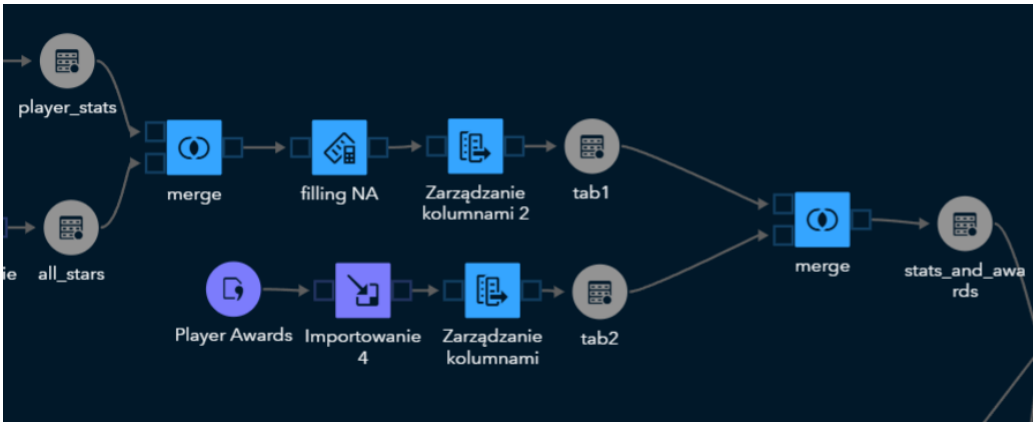
Poniżej opisane są kolejne kroki przygotowania danych i ich wstępnego processingu.

1.Informacje ogólne i statystyki



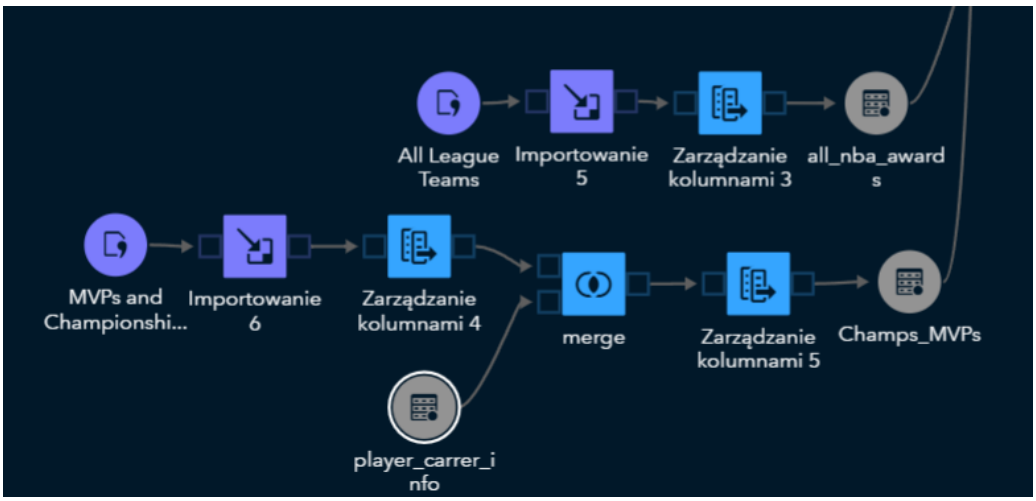
Rys. 1. Wczytanie i dołączenie statystyk do informacji ogólnych o zawodnikach (tabela player_stats).

2.Dołączenie informacji o nagrodach indywidualnych.



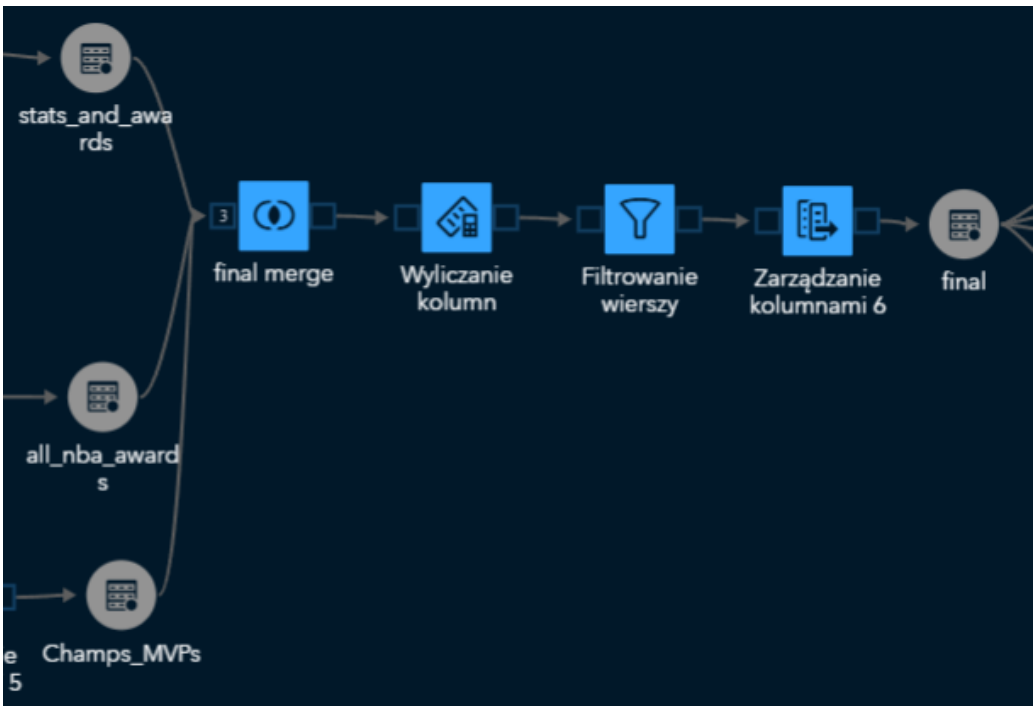
Rys. 2. Dołączenie nowo wczytanej tabeli, wybranie interesujących nas kolumn oraz zastąpienie braków danych zerami w przypadku braku nagrody danego typu (rezultat: tabela stats_and_awards).

3.Przygotowanie informacji o zdobytych mistrzostwach i ilości wystąpie w drużynie sezonu.



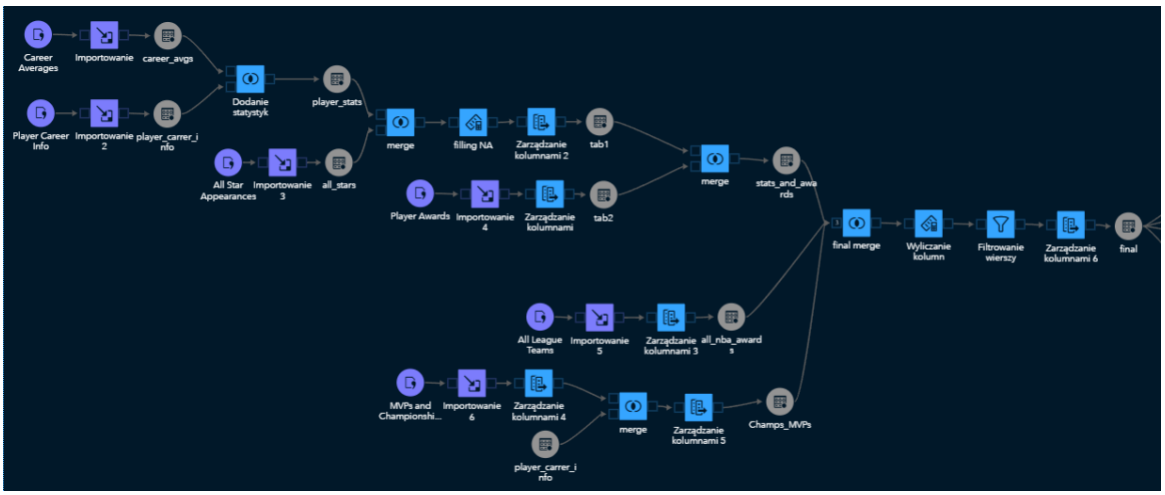
Rys. 3. Wczytanie i wstępne filtrowanie danych (rezultat: tabele all_nba_awards i Championships).

4.Połączenie zbiorów danych z punktu 3 i 4 w finalną ramkę.



Rys. 4. Połączenie tabel stats_and_awards, all_nba_awards i Championships. Wybranie interesujących nas kolumn, które są ważne w ocenie naszej zmiennej celu, flitrowanie wierszy(zawodnicy aby być brani pod uwagę muszą mieć zakończoną kariere minimum 3 lata)

5.Cały przepływ



Rys. 5. Widok całościowy, na przygotowanie danych.

2.1. Opis zmiennych

Poniżej zawarłem ogólny opis zmiennych używanych w następnych krokach projektu:

player Imię i nazwisko zawodnika.

num_seasons Liczba sezonów rozegranych w NBA, pokazująca doświadczenie zawodnika w lidze.

ppg Średnia liczba punktów zdobywanych przez zawodnika na mecz.

apg Średnia liczba asyst, które zawodnik wykonuje na mecz.

rpg Średnia liczba zbiórek, które zawodnik zbiera na mecz.

bpg Średnia liczba bloków, które zawodnik wykonuje na mecz.

MVPs Liczba nagród dla najlepszego zawodnika sezonu regularnego.

Finals_MVP Nagroda dla najlepszego zawodnika finałów NBA.

Championships Liczba zdobytych mistrzostw NBA.

AllStarCount Liczba występów zawodnika w meczu gwiazd NBA.

ROTY Nagroda dla najlepszego debiutanta roku.

hof Wskazuje, czy zawodnik został wprowadzony do Galerii Sław NBA **(zmienna celu)**.

NBA1st Liczba razy, gdy zawodnik znalazł się w pierwszym zespole NBA.

NBA2nd Liczba razy, gdy zawodnik znalazł się w drugim zespole NBA.

NBA3rd Liczba razy, gdy zawodnik znalazł się w trzecim zespole NBA.

DEF1st Liczba razy, gdy zawodnik znalazł się w pierwszym zespole obrony NBA.

DEF2nd Liczba razy, gdy zawodnik znalazł się w drugim zespole obrony NBA.

CPOY Nagroda dla najlepszego obrońcy w sezonie.

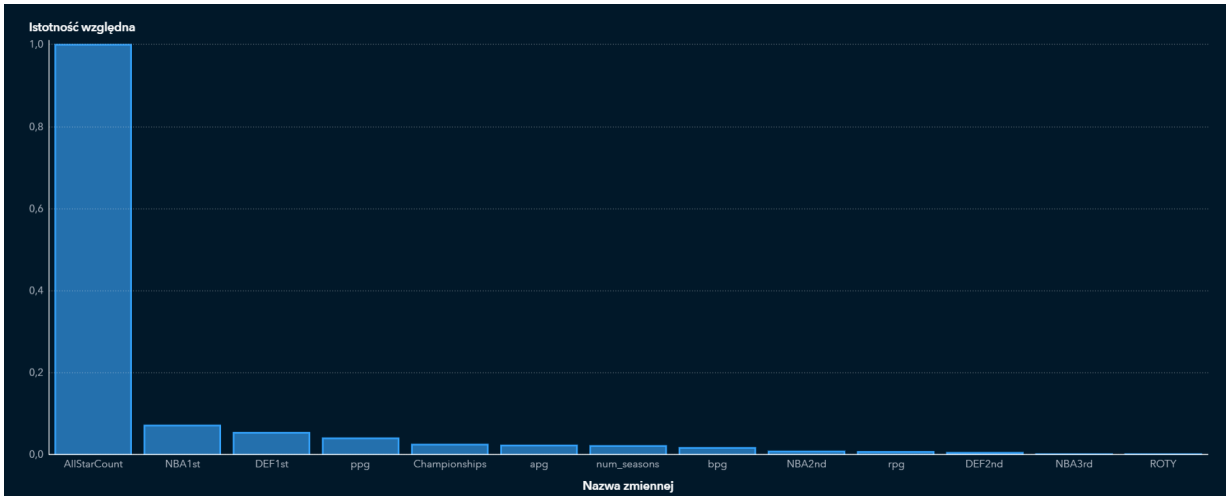
DPOTY Nagroda najlepszego obrońcy roku.

3. Eksploracyjna analiza danych

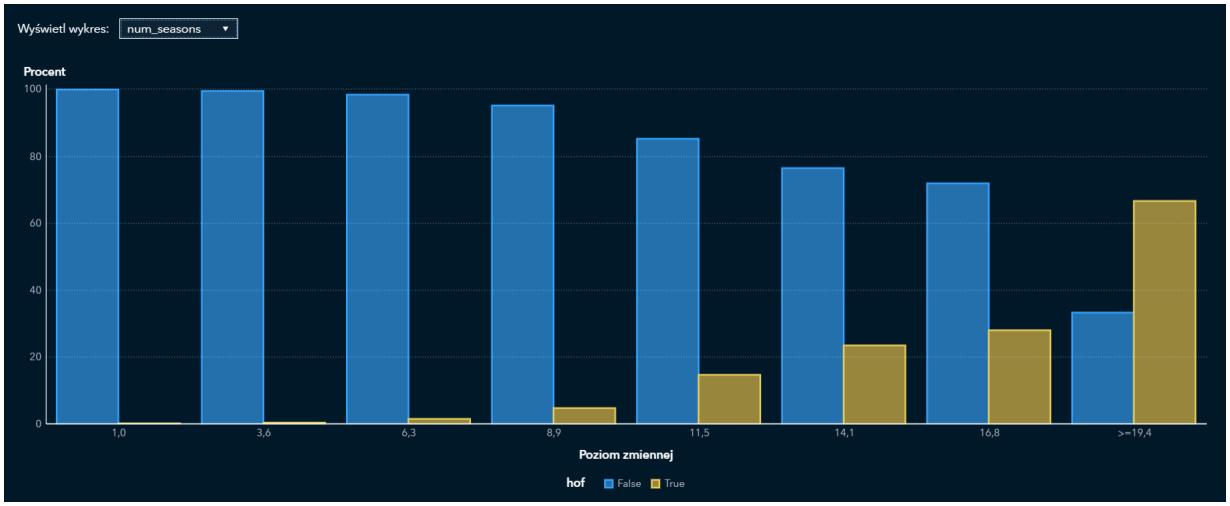
3.1. Analiza zmiennej celu i korelacji ze statystykami

Przy 3930 obserwacjach jedynie około 3% graczy należy do Galerii Sław. Poniżej przedstawiam wyniki analizy i ciekawe korelacje zmiennej celu w zależności od różnych statystyk.

Rozkład zmiennej celu w zależności od liczby rozegranych sezonów pokazuje, że realna szansa na dostanie się do Hall of Fame pojawia się od około 10 rozegranych sezonów (**Rys.7**). Wykres istotności zmiennych wskazuje, które statystyki najbardziej wpływają na naszą zmienną celu.

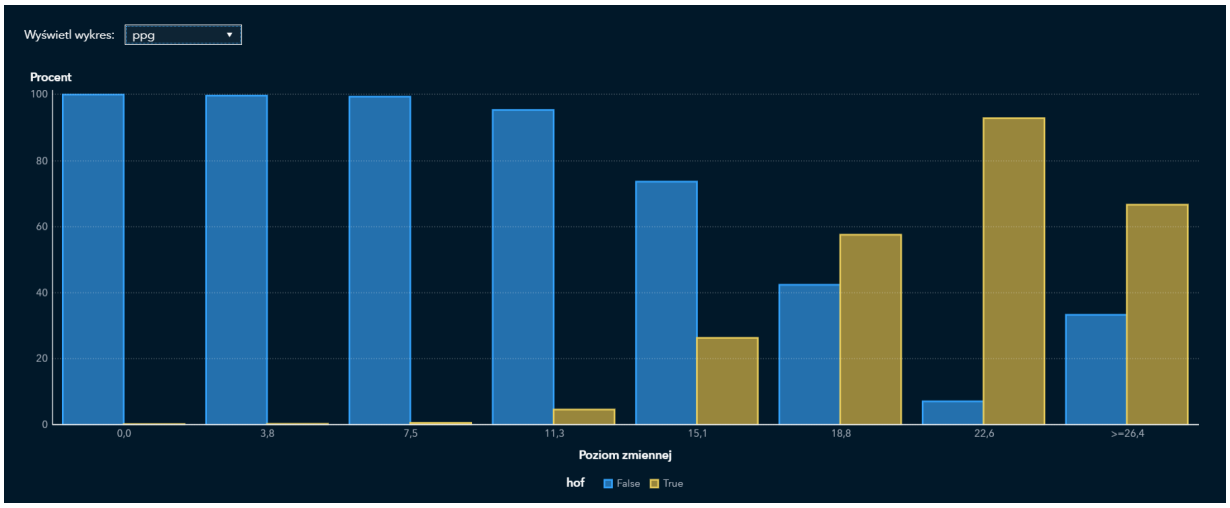


Rys. 6. Wykres istotności zmiennych.

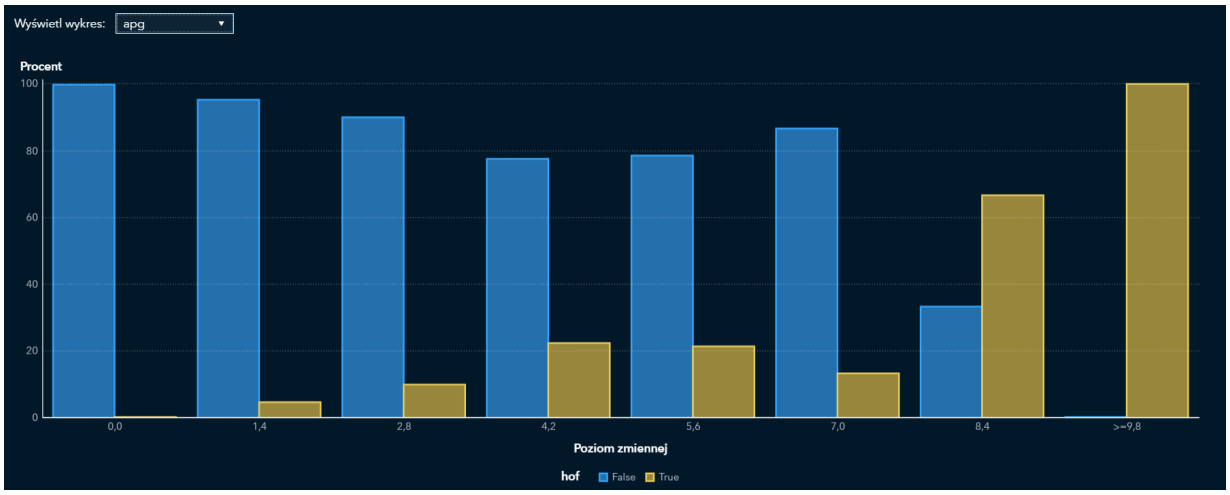


Rys. 7. Wykres liczby sezonów.

Wysokie średnie zdobycze punktowe (**Rys.8**) oraz średnia liczba asyst (**Rys.9**) znacząco zwiększają szanse na nominację do Hall of Fame. Analiza tych rozkładów potwierdza, że ofensywne statystyki są jednym z najważniejszych kryteriów.

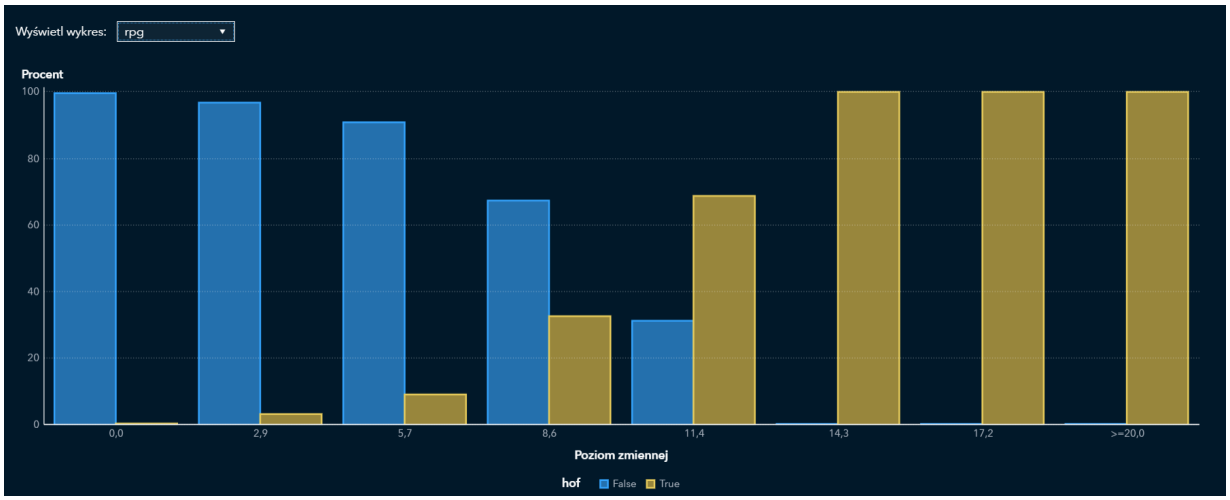


Rys. 8. Rozkład graczy w Hall of Fame w zależności od średnich zdobyczy punktowych.

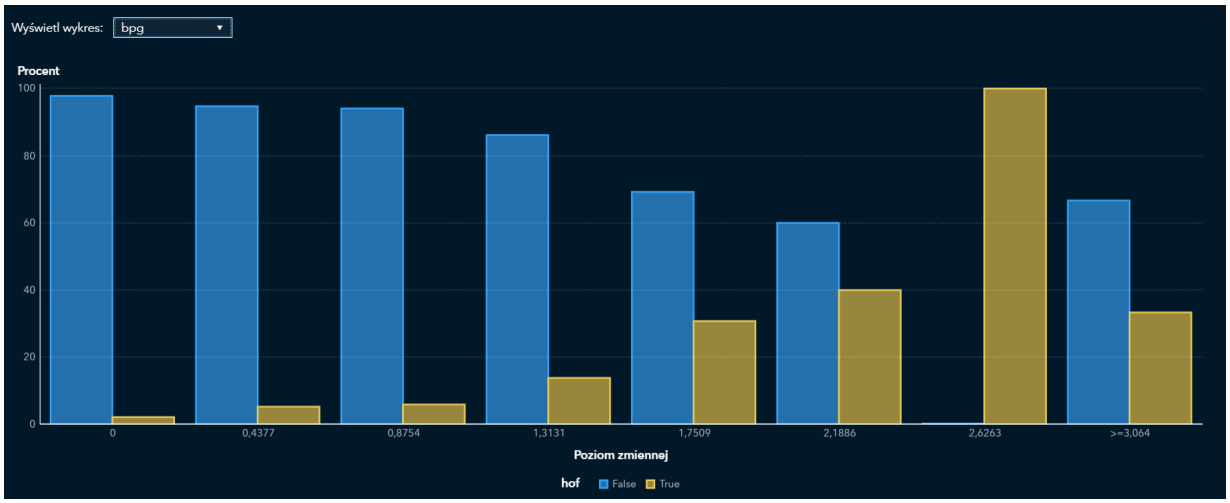


Rys. 9. Rozkład graczy w Hall of Fame w zależności od średniej liczby asyst.

Średnia liczba zbiórek (**Rys.10**) oraz bloków (**Rys.11**) również różnicuje graczy Hall of Fame, choć wyniki zbiórek powyżej 14 na mecz dotyczą głównie zawodników z dawnych lat, co może nie być reprezentatywne dla obecnych standardów.

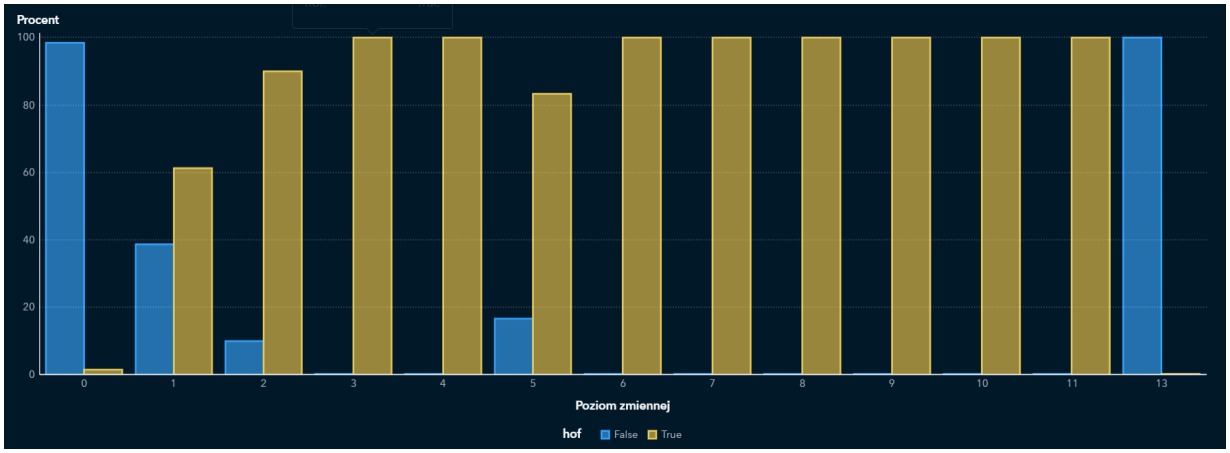


Rys. 10. Rozkład graczy w Hall of Fame w zależności od średniej liczby zbiórek.

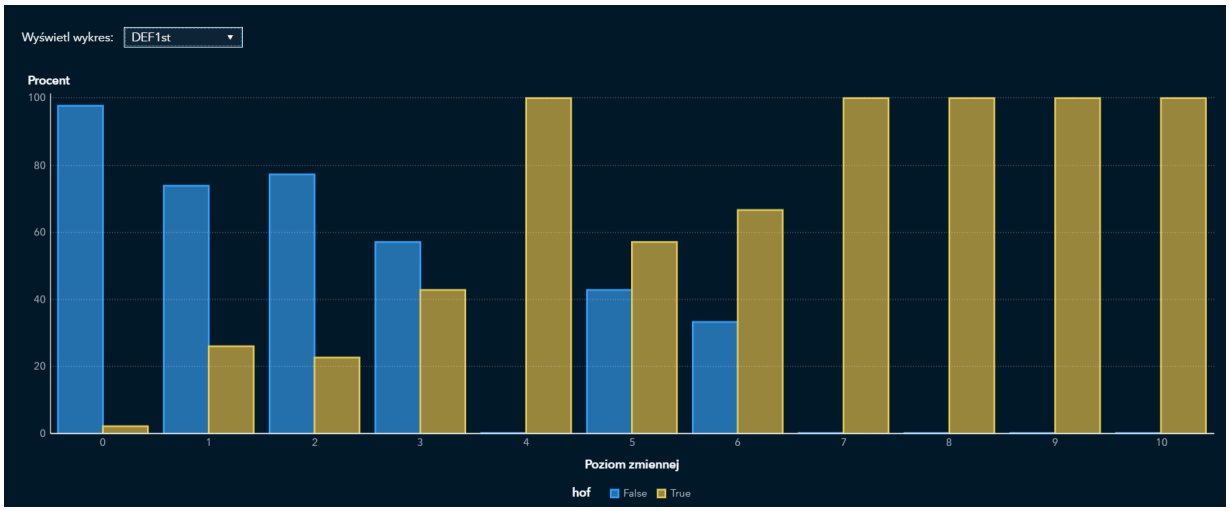


Rys. 11. Rozkład graczy w Hall of Fame w zależności od średniej liczby bloków.

Liczba nominacji do drużyn sezonu (**Rys.12**) jest unikatowym wyróżnieniem przyznawanym jedynie 15 zawodnikom każdego sezonu, co znacząco wpływa na wartość zmiennej celu. Z kolei liczba nominacji do defensywnych drużyn sezonu (**Rys.13**) również ma wpływ, choć statystyki defensywne są mniej cenione niż ofensywne.

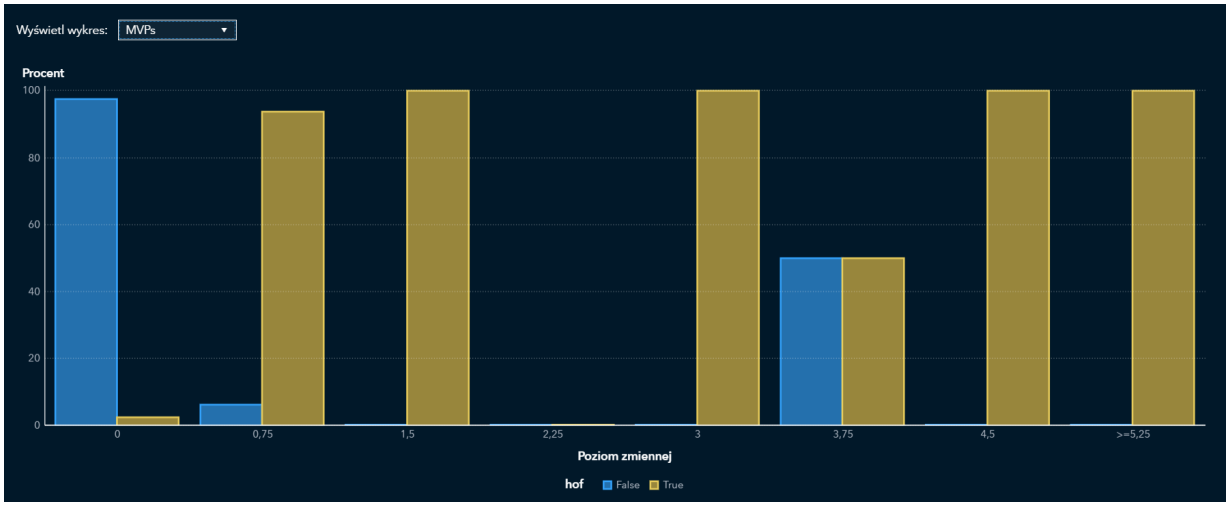


Rys. 12. Rozkład graczy w Hall of Fame w zależności od liczby nominacji do drużyn sezonu.

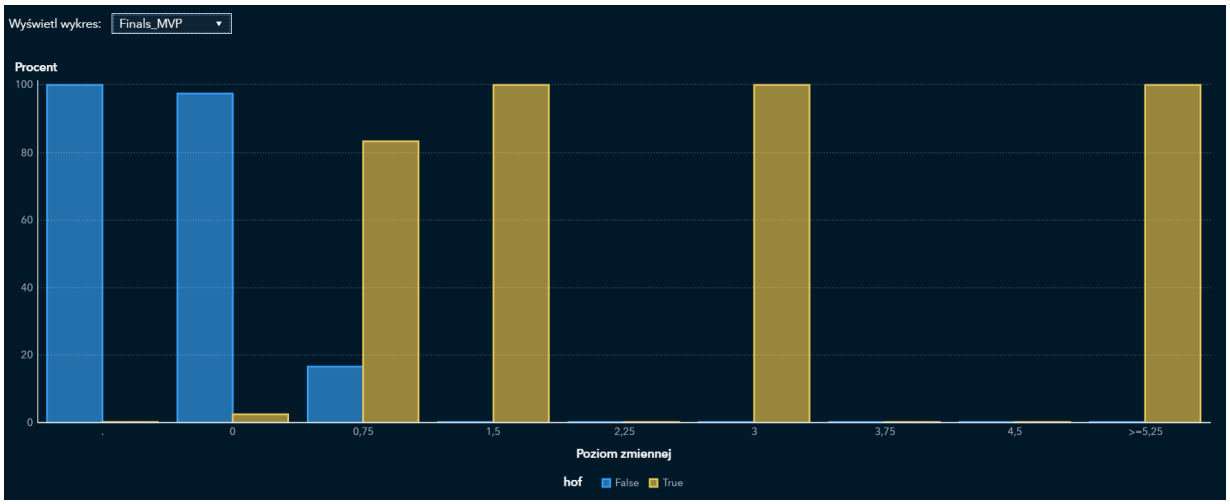


Rys. 13. Rozkład graczy w Hall of Fame w zależności od liczby nominacji do defensywnych drużyn sezonu.

Nagrody MVP sezonu regularnego (**Rys.14**) oraz MVP finałów (**Rys.15**) to prestiżowe wyróżnienia, które znacznie zwiększają szansę na znalezienie się w Hall of Fame.

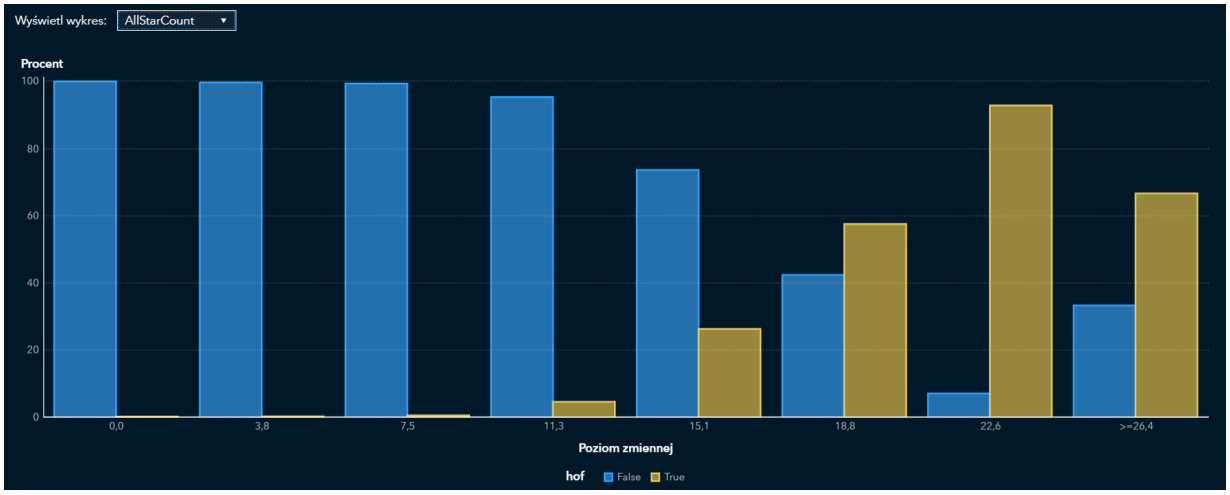


Rys. 14. Rozkład graczy w Hall of Fame w zależności od liczby nagród MVP sezonu regularnego.

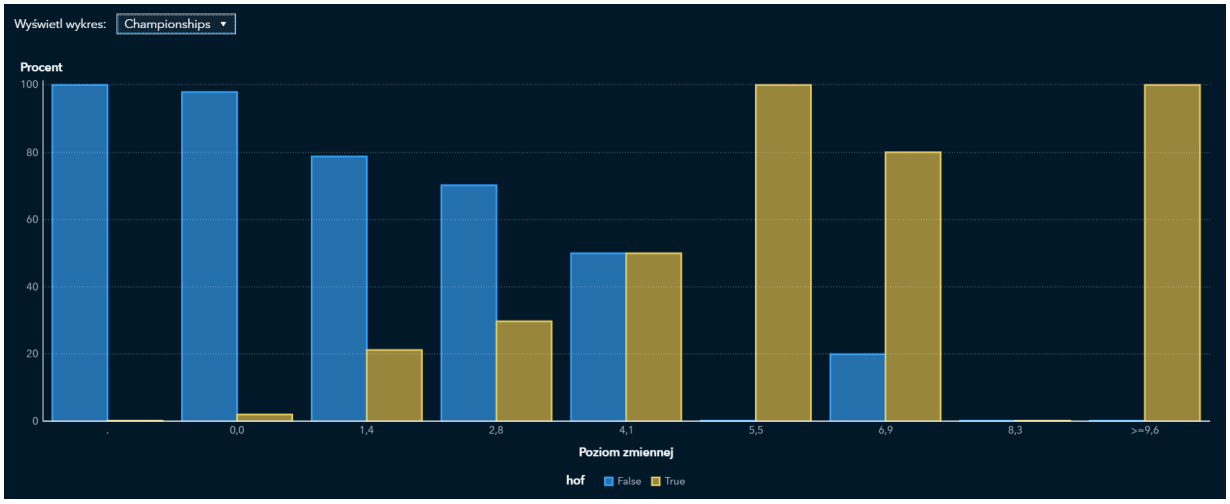


Rys. 15. Rozkład graczy w Hall of Fame w zależności od liczby nagród MVP finałów.

Liczba nominacji do meczu gwiazd NBA (**Rys.16**) oraz zdobytych mistrzostw NBA (**Rys.17**) również są kluczowymi kryteriami oceny graczy. Duża liczba występów w meczach gwiazd świadczy o długiej i wybitnej karierze, a zdobyte mistrzostwa są uważane za odnośnik sukcesu i wielkości gracza.



Rys. 16. Rozkład graczy w Hall of Fame w zależności od liczby nominacji do meczu gwiazd NBA.



Rys. 17. Rozkład graczy w Hall of Fame w zależności od liczby zdobytych mistrzostw NBA.

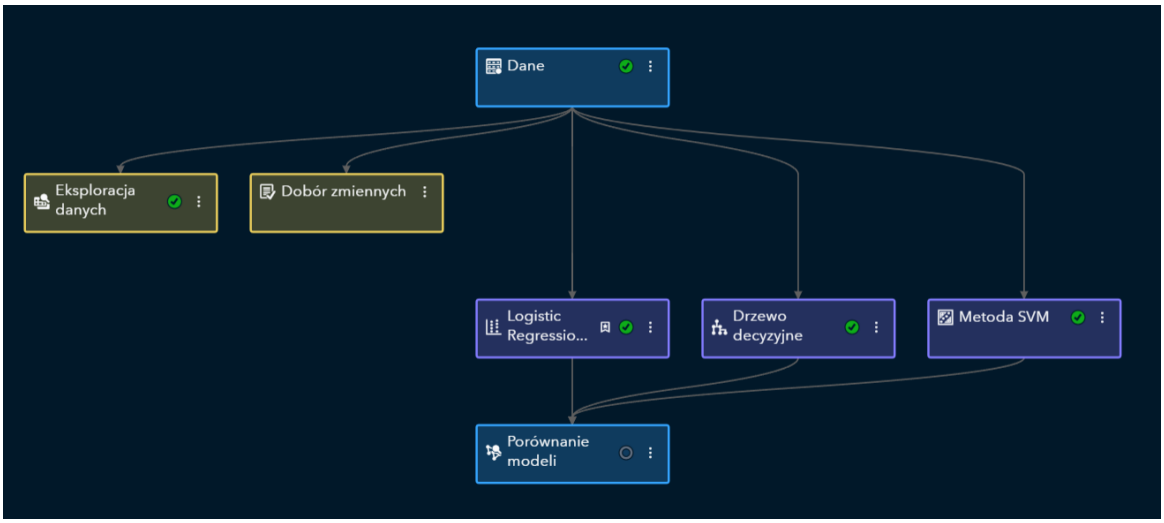
3.2. Podsumowanie eksploracji danych.

Zdecydowanie widać, że statystyki indywidualne i osiągniecia są bardzo ważnym kryterium decydującym o wartości naszej zmiennej celu. Również można zauważyć, że bardzo często statysyki i nagrody ofensywne zdają się mieć większy wpływ na naszą zmienną.
Nie powinna również dziwić bardzo duża istotność zmiennej 'AllStarCount', albowiem duża liczba wystąpień w meczu Gwiazd, mówi nam że zawodnik grał na bardzo wysokim poziomie oraz był bardzo popularny przez wiele lat swojej kariery, co jest bardzo mocno brane pod uwage przy wcieleniu do Galerii Sław NBA.
Aby się tam dostać trzeba odznaczyć się zarówno wybitnymi osiągnięciami indywidualnymi, jak i mieć realny wpływ na rozwój koszykówki i jej rewolucjonizacje (ciężko znaleźć na liście HallOfFame osobę, której fani koszykówki nie znają lub nie wiedzą z czego zasłynął lub co osiągnął).

4. Modelowanie

4.1. Pipeline

Do analizy oraz tworzenia modeli posłużył mi poniższy pipeline. Do modelowania wybrałem te metody ponieważ regresja logistyczna pozwala na procentowe okreslenie prawdopodobieństwa natomiast drzewo decyzyjne jest proste w interpretacji, co pozwala na porównanie rozumowania modelu, z faktycznym sposobem przydzielania do HallOfFame.



Rys. 18. Pipeline stworzony w SAS Model Builder.

4.2. Porównanie modeli

Porównanie modeli

Model ...	Nazwa	Nazwa ...	KS (Youden)	Dokładność	Przeciętny błąd ...
★	Logistic Regression_1	Regresja logistyczna	0,9895	0,9873	0,0077
	Drzewo decyzyjne	Drzewo decyzyjne	0,8773	0,9669	0,0235
	Metoda SVM	Metoda SVM	0,9738	0,9822	0,0287

Rys. 19. Porównanie stworzonych modeli.

Jako, że we wszystkich trzech modelach uzyskaliśmy stosunkowo dużą dokładność, to w dalszej części skupie się na bardziej delikatnej analizie logsitycznej regresji i drzewa decyzyjnego, albowiem są one łatwe w interpretacji oraz dobrze dopasowują się do rozwiązywanego przez nas problemu.

4.3. Model regresji

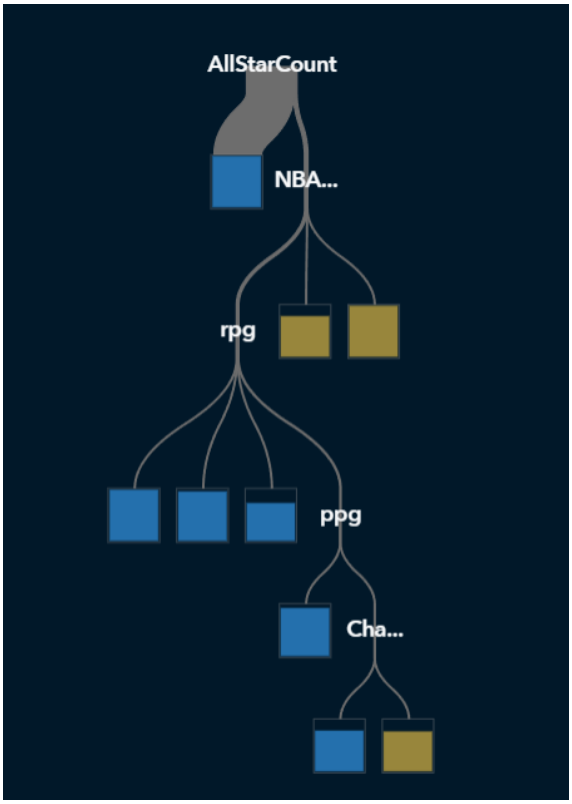
Funkcją łączącą dla zmiennej binarnej celu wybrałem **logit**. Do selekcji zmiennych w modelu wybrałem metodę **wsteczną** (*backward selection*), przy czym kryterium wyboru efektów i zatrzymania procesu selekcji zmiennych zostało oparte na *skorygowanym kryterium informacyjnym Akaike* (**AICC**). Wybrane podejście umożliwia równoważenie złożoności modelu i jego jakości dopasowania, minimalizując overfitting.

Oceny parametrów					
Parametr	Stopnie swobody	Ocena	Błąd standardowy	Chi-kwadrat	Pr. > chi-kw.
Intercept	1	-11,406345	1,338472	72,6230	<,0001
AllStarCount	1	0,838733	0,115615	52,6281	<,0001
apg	1	0,276307	0,162316	2,8978	0,0887
Championships	1	1,003249	0,203754	24,2442	<,0001
ppg	1	0,246106	0,066093	13,8653	0,0002
rpg	1	0,234798	0,092216	6,4830	0,0109

Rys. 20. Istotne zmienne i wartości ich współczynników (link do opisu zmiennych).

4.4. Model drzewa decyzyjnego

W finalnym modelu drzewa decyzyjnego zdecydowałem się na Entorpy jako kryterium klasyfikacji zmiennej celu, ponieważ dawało najlepsze rezulataty.
Minimalny rozmiar liścia oraz maksymalna głębia zostały ustalone na 10, a maksymalna liczba odgałęzień wynosiła 5.
Za metodę przycinania drzewa wybrałem 'cost-complexity', z parametrem automatycznym, co jest uniwersalną metodą dobrze sprawdzającą się w wielu modelach.



Rys. 21. Widok drzewa z podziałami. Kolorem niebieskim oznaczone jest przyporządkowanie FALSE w zmiennej hof, a złotym TRUE.

5. Podsumowanie

Przedstawione powyżej modele dobrze radzą sobie z klasyfikacją zawodników do HallOfFame. Dzięki tak obranym parametrom model równie dobrze zadziała dla przyszłych danych , kiedy zawodnicy obecnie grający, skończą już kariere czy będą mieli za sobą większą liczbe sezonów.
Analiza modeli pozwala również lepiej zrozumieć, czym kierują się komitety Basketball HallOfFame nominując zawodników do Galerii Sław.

5.1. Niedoskonałości modelu i co można, a co ciężko poprawić?

W rzeczywistości Mistrzostwa oraz nagrody MVP mają widocznie większe znaczenie niż w moich modelach, natomiast wpływa to jedynie na błędne zaklasyfikowanie pojedynczych zawodników.
Może to być spowodowane duży przedziałem danych aż od 1950 roku. Przez te 75 lat koszykówka drastycznie sie zmieniła, a co za tym również statystyki zawodników, a więc możnaby rozważyć model uczący się na jedynie zawodnikach od np. lat 90tych aby lepiej wpasowywał sie on w realia dzisiejszej koszykówki.

Jest natomiast jeden aspekt, który ciężko zawrzeć w modelu, ponieważ nie opisują go żadne dane. Mowa o popularności i wpływie na rozwój koszykówki, albowiem są gracze o przeciętnych statystykach indywidualnych, którzy jednak zrewolucjonizowali grę na zawsze czy też poprzez charyzmatyczność albo styl gry udało im się dotrzeć do serc fanów.
Analizując modele, zauważyć mogłem zależność, że z tego typu graczami modele miały najwiekszy problem, i błędnie nie przyporządkowywały ich do Galerii Sław.

5.2. Możliwe zastosowanie biznesowe

Z modeli można również skorzystać dla graczy ciągle grających w NBA, co pozwala określić ich szanse na dostanie sie do HallOfFame, czyli tak naprawde zostać ponadczasowym ambasadorem koszykówki.
Można również 'przesymulować' kariere zawodnika (np. zakładając że gracz utrzyma dyspozycje i jeszcze przez 7 sezonów bedzie grał w okolicach obecnych statystyk). Zastosowanie modelu pozwoli wtedy lepiej określić jego szanse oraz potencjał.

Takie rozwiązanie może przydać się przeróżnym firmom czy sponsorom szukającym nowych twarzy dla swoich marek. Mogłyby na podsawie ewaluacji modelu analizować „przyszłościowość" graczy, aby nie marnować pieniędzy na zawodników którzy mają bardzo niewielkie czy zerowe szanse na zostanie jedną z głównych postaci ligi NBA. Ma to szczególne znaczenie w przypadku podpisywania długoterminowych kontraktów, a firmy nie mogą zwlekać, gdyż szybko mogą stracić zawodnika na rzecz konkurencji.

Podpisanie zawodnika z dużą szansą na dostanie się do HallOfFame wiąże się z dużymi zyskami i ekspozycją marki, nawet po zakończeniu przez niego kariery.