

Przygotowanie danych do uczenia modelu

Dane z lat 2000-2023 zostały pobrane ze strony internetowej www.basketball-reference.com. Posiadały one po 28 statystyk dla każdego gracza w jednym sezonie. Finalnie dane te zostały uproszczone i wybranych zostało 10 głównych cech, które byłybrane pod uwagę podczas uczenia modelu.

Zbiór danych uwzględniający najnowszy sezon 23/24 został wydzielony z głównego zbioru, ponieważ stanowił on bazę w momencie predykcji. Dzięki czemu model nie uczył się na danych, na których dokonywał w późniejszej fazie predykcji.

Trenowanie modelu

- wczytuje dane z pliku all_stats.csv, w którym umieszczone są następujące statystyki graczy: GS, MIN, FGM, FGA, FTM, FTA, AST, STL, TOV, PTS, Player, Position
- usuwane są kolumny:
 - 'Player' - aby model nie sugerował się nazwiskiem gracza przy jego wybieraniu do drużyn,
 - 'Year' - ponieważ rok, w którym gracz został wybrany do all-nba nie ma znaczenia,
 - 'TOV' - nie została wykazana wyraźna zależność między ilością strat a wyborem do drużyny,
- mapuje wartości dotyczące pozycji gracza oraz przynależności do drużyny all-nba,
- skaluje cechy za pomocą StandardScaler,
- trenuje model regresji logistycznej (LogisticRegression),
- model jest zapisywany do pliku za pomocą joblib jako model_all_nba.joblib

Wybór modelu

Model LogisticRegression został wybrany, ponieważ w fazie testów modele Random Forest oraz SVM wykazywały znacznie gorsze rezultaty.

Model został skonfigurowany z parametrem max_iter=1000, co zwiększa maksymalną liczbę iteracji w celu zapewnienia konwergencji, oraz z parametrem class_weight='balanced', co oznacza, że model automatycznie dostosowuje wagę klas w celu zbalansowania danych (zastosowano takie rozwiązanie ponieważ klasa graczy, którzy nie zakwalifikowali się do jakiejkolwiek drużyny all-nba była wielokrotnie większa niż graczy zakwalifikowanych np. do First-all-nba-team).

Predykcja

- wczytuje dane testowe tylko z najnowszego sezonu,
- wytrenowany model grupuje zawodników według przewidywanych zespołów All-NBA i Rookie All-NBA,
- po wygenerowaniu zespołów zapisuje wyniki do pliku JSON.