Aleksandra Derda

e-mail: 177427@student.ue.wroc.pl
ORCID: 0009-0003-6022-2238

Uniwersytet Ekonomiczny we Wrocławiu

Diagnoza nierówności płacowych wśród sportowców z wykorzystaniem metod uczenia maszynowego

DOI: 10.15611/2024.76.5.02
JEL: C5, J3, J7

Cytuj jako: Derda, A. (2024). Diagnoza nierówności płacowej wśród sportowców z wykorzystaniem metod uczenia maszynowego. W: A. Stanimir (red.), *Współczesne problemy społeczno-ekonomiczne w ujęciu analitycznym* (s. 25-41). Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu.

Streszczenie: Głównym celem pracy jest udowodnienie – z wykorzystaniem metody XGBoost – występowania luki płacowej w sporcie na podstawie wypłat koszykarzy i koszykarek z lig NBA i WNBA. Sprawdzano również prawdziwość stwierdzenia, że na wysokość wypłaty wpływają cechy fizyczne, popularność oraz skuteczność na boisku w obronie i ataku zawodnika lub zawodniczki. Utworzono jeden model, którego wyniki uznano za zadowalające. Porównano dla tego modelu wpływ poszczególnych zmiennych z wykorzystaniem trzech różnych metod. Uzyskano spójne wyniki potwierdzające, że poza skutecznością w obronie wszystkie z zakładanych czynników wpływających na wysokość wypłaty okazały się istotne. Potwierdzono również występowanie dyskryminacji płacowej ze względu na płeć.

Słowa kluczowe: luka płacowa, sport, uczenie maszynowe, XGBoost

1. Wstęp

Sport od wieków jest integralną częścią społeczeństwa i kultury każdego narodu (Waśkowski, 2011). Niepodważalna jest jego rola w formowaniu społeczeństwa, kształtowaniu tożsamości narodowej czy promowaniu ludzkich wartości. Uprawiany rekreacyjnie pozytywnie wpływa szczególnie na zdrowie i relacje społeczne jednostki, natomiast zawodowo motywuje do osiągania doskonałości.

Sport profesjonalny to także obszar, w którym pasja, talent czy zaangażowanie są nagradzane publicznym uznaniem i rozgłosem. Jest to widowisko, które gromadzi miliony fanów z całego świata, co przyciąga również największe firmy. Są one w stanie wyłożyć ogromne pieniądze na inwestycje w sport, by być zauważonym przez fanów wybranych dyscyplin. Sektor sportu uznaje się za globalny (Gratton, 1998), dlatego nie powinno dziwić, że branży tej towarzyszą zarówno wielkie

emocje, jak i wielkie pieniądze. Niestety, problemy, z którymi zmagać się muszą czołowi działacze i sportowcy również są znacznie trudniejsze i mają poważniejsze konsekwencje niż te obserwowane w życiu codziennym. Dotyczy to także nierówności społecznych.

Sport zawodowy jest jedną z branż najmocniej dotkniętych dyskryminacją płciową, gdyż w sporcie można zauważyć niemal każdy rodzaj dyskryminacji (Thakur, 2022). O różnych objawach dyskryminacji płciowej można przeczytać w pracy Tomaszewskiej (2004). W sporcie dyskryminacja ze względu na płeć jest znacznie szerszym i niejednokrotnie poważniejszym pojęciem. Dyskryminacja płciowa nie tylko wpływa na doświadczenie zawodowe sportowców na każdym etapie kariery zawodowej, na ich status materialny, ale także kształtuje oblicze całej branży sportowej. Jednym ze skutków występowania dyskryminacji płciowej jest również badana luka płacowa.

Celem tego badania jest udowodnienie występowania luki płacowej w sporcie na przykładzie amerykańskich profesjonalnych lig koszykarskich NBA i WNBA. Dodatkowym celem było przygotowanie modelu dobrze prognozującego wysokość wypłaty zawodnika. W badaniu poszukiwano odpowiedzi na pytania:

- Czy w amerykańskich ligach koszykarskich występuje dyskryminacja płacowa ze względu na płeć?
- Czy płeć ma największy wpływ na predykcję wysokości wypłaty?
- Czy czynniki, jakie mają wpływ na wysokość wypłaty, to cechy fizyczne zawodnika, jego popularność oraz skuteczność na boisku w obronie i ataku?

2. Luka płacowa w sporcie

Pojęcie „luka płacowa" jest bezpośrednim tłumaczeniem angielskiego zwrotu *gender pay gap*. Istnieje wiele definicji opisujących ten termin, jednak w większości z nich pojawia się wspólny element, który opisuje ją jako miarę nierówności w płacach kobiet i mężczyzn. Istnieją też inne definicje, jak np. Parlamentu Europejskiego (2023), zgodnie z którą luka płacowa ze względu na płeć to „różnica pomiędzy średnimi stawkami godzinowymi brutto, które otrzymują kobiety i mężczyźni" lub definicja GUS (2017, s. 6), według której luka płacowa to „najczęściej stosowany wskaźnik do porównań wynagrodzeń między poszczególnymi grupami osób".

Chociaż nie występuje osobna definicja dla luki płacowej w sporcie, jest to całkowicie inne zjawisko niż luka obserwowana w sektorze biznesowym. Zjawisko to jest zauważalne w każdym aspekcie rywalizacji sportowej, szczególnie na jej najwyższym szczeblu. Różnice płacowe w sporcie bywają nieporównywalnie większe niż znane z ogólnych raportów badających lukę płacową w biznesie (Adelphi University, 2023; Parlament Europejski, 2023). Znacznie różnią się również przyczyny występowania luki płacowej między mężczyznami i kobietami o podobnych umiejętnościach.

Sytuację kobiet i mężczyzn w wybranych sportach przedstawiono w opracowaniu Adelphi University (2023). W tabeli 1 zaprezentowano średnie zarobki dla kobiet i mężczyzn w czterech dyscyplinach. Dla każdego z tych sportów analizowano zarob-

ki w dwóch ligach rozgrywek na najwyższym poziomie, po jednej dla kobiet i mężczyzn. Nazwy lig podano w nawiasach po nazwie dyscypliny, najpierw liga męska, potem liga kobieca. Wszystkie kwoty podano w dolarach. Dodatkowo przeliczono iloraz wypłat kobiet i mężczyzn, by pokazać jaką część średnich zarobków mężczyzn zarabiają średnio kobiety.

Tabela 1. Średnie zarobki kobiet i mężczyzn w wybranych dyscyplinach sportowych

Dyscyplina	Średnie roczne zarobki mężczyzn [w dolarach]	Średnie roczne zarobki kobiet [w dolarach]	Iloraz zarobków kobiet przez zarobki mężczyzn
Koszykówka (NBA i WNBA)	10 776 383	113 295	1,05%
Piłka Nożna (MLS i NWSL)	471 279	54 000	11,46%
Tenis (100 najlepszych w ATP i WTA)	1 589 024	1 039 141	65,39%
Golf (PGA i LPGA)	1 042 917	346 360	33,21%

Źródło: opracowanie własne na podstawie (Adelphi University, 2023).

Jak wynika z tab. 1, we wszystkich analizowanych dyscyplinach mężczyźni zarabiają więcej niż kobiety i – niestety – różnica ta jest zazwyczaj znacząca. Najlepszą sytuację dla kobiet można zaobserwować w tenisie ziemnym, gdzie zarabiają one średnio około 65% średnich zarobków mężczyzn. Najgorzej natomiast stosunek ten wygląda w koszykówce, gdzie kobiety otrzymują średnio tylko 1% przeciętnych wypłat mężczyzn. Różnica ta jest ogromna i pokazuje, z jak poważną dyskryminacją płacową muszą się zmagać kobiety.

3. Wprowadzenie do tematu badania i opis wykorzystanych metod

Badanie zdecydowano się przeprowadzić na podstawie danych z amerykańskich lig koszykarskich. Dyscyplina ta jest dobrym obiektem badawczym, gdyż podobnie jak w przypadku innych opisywanych wcześniej dyscyplin, w koszykówkę również początkowo mogli grać jedynie mężczyźni. National Basketball Association (NBA), czyli amerykańsko-kanadyjska męska liga koszykarska, powstała już w 1946 roku (NBA, b.d.a), podczas gdy Women's National Basketball Association (WNBA) założono dopiero w 1996 roku (WNBA, b.d.a). Te dane wskazują, że dyskryminacja ze względu na płeć jest obecna w tym sporcie od wielu lat.

Dodatkowo popularność oraz prestiż tego sportu w Stanach Zjednoczonych przyciągają nie tylko najlepszych zawodników na świecie, ale również sponsorów (KAF, 2022), którzy są w stanie przekazać duże kwoty, by uzyskać rozgłos wśród fanów tej dyscypliny, szczególnie męskich rozgrywek. Niestety, jak pokazują przedstawione we wcześniejszej części pracy badania (tab. 1), pomimo tak dużego zainteresowania, koszykówka w Stanach Zjednoczonych wciąż jest mocno dotknięta problemem luki płacowej.

3.1. Opis wykorzystanych metod

3.1.1. Analiza głównych składowych

Analiza głównych składowych to jeden z algorytmów stosowanych do redukcji wymiarów, który polega na rzutowaniu danych na przestrzeni o mniejszej liczbie wymiarów przy jednoczesnym zachowaniu struktury danych w jak największym stopniu (Mamczur, 2019). Metoda ta polega na „transformacji zmiennych pierwotnych w zbiór nowych nieskorelowanych zmiennych zwanych głównymi składowymi" (Sztemberg-Lewandowska, 2017), gdzie każda z nich tworzona jest przez kombinację liniową zmiennych pierwotnych w taki sposób, by maksymalizować zmienność niewyjaśnioną przez poprzednią składową.

Dla każdej składowej można wyznaczyć wartość własną, która informuje, jaką część całkowitej zmienności wyjaśnia dana składowa. Najwięcej wariancji przedstawia zawsze pierwsza składowa, a każda następna tłumaczy niewyjaśnioną jeszcze część zmienności. Zsumowane wartości dla wszystkich składowych przedstawiają całkowitą wariancję (PQStat, b.d.).

Analizę głównych składowych stosuje się przede wszystkim w celu redukcji liczby zmiennych, co pozwala zarówno na wykrycie ewentualnych prawidłowości, jak i wizualizację początkowo wielowymiarowych zbiorów danych.

3.1.2. Algorytm uczenia maszynowego

Uczenie maszynowe (*machine learning*) odnosi się do szerokiego zestawu aplikacji AI, w których komputery budują modele na podstawie wzorców rozpoznanych w zbiorach danych i wykorzystują te modele do generowania hipotez na temat świata. Takie modele mają niezliczone zastosowania w programowaniu do rozwiązywania problemów zarówno klasyfikacyjnych jak i regresyjnych (Russell i Norvig, 2020).

Rys. 1. Uproszczony przepływ implementacji modelu uczenia maszynowego

Źródło: opracowanie własne na podstawie (Sundberg i Holmström, 2024).

Stosowanie metod uczenia maszynowego wymaga ścisłego trzymania się narzuconego schematu, którego uproszczony przepływ przedstawiono na rys. 1. Widać na nim, że by osiągnąć zadowalające rozwiązanie, najczęściej trzeba przejść kilka iteracji przedstawionego procesu. Modelowanie powinno składać się z czterech etapów: zbierania i przygotowania danych, trenowania modelu, ewaluacji uzyskanego modelu oraz wdrożenia modelu do użytkowania.

Przygotowanie danych

Na etapie zbierania danych niezbędne jest wskazanie, co jest problemem, który ma być rozwiązany za pomocą modelowania. Do uzyskania dobrego jakościowo rozwiązania powinno się też prawidłowo wskazać zmienne endo- i egzogeniczne.

Część metod uczenia maszynowego wymaga również normalizacji danych. Wykorzystywana w tym badaniu metoda XGBoost oparta na drzewach decyzyjnych jest nieczuła na różnice wielkości między zmiennymi, dlatego przeprowadzanie normalizacji danych jest niepotrzebne (Filho, 2022).

Podział na zbiór uczący i testowy

Zbiór uczący wykorzystywany jest do trenowania modelu – to na podstawie tych danych wykrywane są zależności (Blog Statystyczny, 2020). Zbiór ten jest zazwyczaj najbardziej liczny.

Dane ze zbioru testowego są, jak wskazuje na to nazwa, wykorzystywane do testowania modelu. Obserwacje z tego zbioru służą do symulacji działania modelu na nowych niewykorzystywanych wcześniej danych.

Trenowanie modelu – XGBoost

W badaniu wykorzystano metodę wzmacniania gradientowego – model XGBoost. Model ten jest techniką uczenia zespołowego (*ensemble learning*), która polega na łączeniu przewidywań grup predyktorów, czyli wspólnym analizowaniu wyników wielu mniejszych modeli. Takie podejście pozwala na znaczne poprawienie wyników predykcji, istotnie zmniejsza się również wariancja głównego modelu (Data Science Team, 2020). XGBoost jest modelem sekwencyjnym, czyli modele bazowe są tworzone po kolei, a każdy następny zależy od wyników uzyskanych dla poprzedniego.

Wszystkie modele wzmacniania gradientowego, należące do uczenia zespołowego, opierają się na trzech krokach (Analytics Vidhya, 2024):

1) utworzenie modelu F0 do przewidywania wartości zmiennej objaśnianej; wyznaczenie reszt modelu;

2) utworzenie nowego modelu h1, który dopasowywany jest do reszt uzyskanych dla modelu F0;

3) łączenie wniosków uzyskanych dla F0 i h1, by stworzyć model F1 – poprawiona wersja F0.

Kroki te powtarza się do uzyskania minimalnej wartości wybranego wskaźnika oceny jakości modelu (takim wskaźnikiem może być MSE) lub po przejściu wskazanej liczby iteracji.

Choć XGBoost to jeden z algorytmów opartych na wzmacnianiu gradientu, znacznie różni się od standardowych modeli z tej grupy. XGBoost jest lepszy od zwykłych algorytmów wzmacniania gradientu, gdyż zamiast trenować najlepszy model na danych (jak w przypadku tradycyjnych metod), trenuje się tysiące modeli na różnych podzbiorach zbioru uczącego, a następnie głosuje się na model o najlepszej wydajności.

GridSearchCV

GridSearchCV to proces dostrajania hiperparametrów w celu określenia optymalnych wartości dla danego modelu (Great Learning Team, 2024a). Metoda ta opiera się na zautomatyzowanym testowaniu modeli ze wszystkich zaproponowanych możliwych wartości hiperparametrów. W ten sposób wyznacza się hiperparametry dla wskazanego modelu, przy których model wykazuje najlepsze wartości wybranego wskaźnika oceniającego jakość. W tej metodzie wykorzystywana jest tak zwana walidacja krzyżowa, o której przeczytać można w Great Learning Team (2024b).

Ocena jakości

W tym etapie modelowania wykorzystuje się np.: MAE, MSE, RMSE, a także współczynnik determinacji.

Skrótu RMSE (*Root Mean Squared Error*) używa się do określenia pierwiastka średniokwadratowego błędu (Hodson, 2022 oraz Chai i Draxler, 2014):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

gdzie: n – liczba obserwacji, y_i – rzeczywista wartość zmiennej objaśnianej dla *i*-tego obiektu, $\hat{y}_i$ – przewidywana wartość zmiennej objaśnianej dla *i*-tego obiektu.

Wartość tego wskaźnika jest przedstawiana w tej samej jednostce co zmienna objaśniana, co znacznie ułatwia interpretację. Wskaźnik ten interpretuje się jako średnią oczekiwaną różnicę między wartością przewidywaną a rzeczywistą.

MAE (*Mean Absolute Error*) to wskaźnik, który oblicza się ze wzoru (Hodson, 2022; Willmott i Matsuura, 2005):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$

gdzie: n – liczba obserwacji, y_i – rzeczywista wartość zmiennej objaśnianej dla *i*-tego obiektu, $\widehat{y_i}$ – przewidywana wartość zmiennej objaśnianej dla *i*-tego obiektu.

Współczynnik jest obliczany przez wyznaczanie wartości bezwzględnych błędów predykcji, co zapobiega wzajemnemu znoszeniu się błędów z dodatnim i ujemnym znakiem. Wartość tego wskaźnika to średnia wszystkich bezwzględnych błędów predykcji. Wskaźnik ten też jest przedstawiany w tej samej jednostce co zmienna objaśniana. Wartość tego wskaźnika interpretuje się jako średnią błędów predykcji.

Współczynnik determinacji R^2 wskazuje, jaka część zmienności zmiennej objaśnianej jest wyjaśniana przez model (Walesiak, 1996). Jest on obliczany zgodnie ze wzorem (Ruijie i in., 2023):

$$R^2 = 1 - \frac{\sum_i (\widehat{y_i} - y_i)^2}{\sum_i (\overline{y_i} - y_i)^2},$$

gdzie: y_i – rzeczywista wartość zmiennej objaśnianej dla *i*-tego obiektu, $\widehat{y_i}$ – przewidywana wartość zmiennej objaśnianej dla *i*-tego obiektu.

Współczynnik ten przyjmuje wartości z przedziału [0,1]. Im wartość bliższa jedności, tym lepsze dopasowanie funkcji regresji do rzeczywistych danych. Im wyższa wartość tego współczynnika, tym większa część zmienności zmiennej objaśnianej jest wyjaśniana przez model.

Istnieje jeszcze wiele innych, mniej popularnych metod ewaluacji modelu, o których przeczytać można w Hodson (2022).

W badaniu wykorzystano do oceny jakości: MAE, MSE, RMSE, a także współczynnik determinacji. RMSE (*Root Mean Squared Error*) to pierwiastek średniokwadratowy błędu (Chai i Draxler, 2014; Hodson, 2022). Wskaźnik ten interpretuje się jako średnią oczekiwaną różnicę między wartością przewidywaną a rzeczywistą.

MAE (*Mean Absolute Error*) to wskaźnik obliczany przez wyznaczanie wartości bezwzględnych błędów predykcji, co zapobiega wzajemnemu znoszeniu się błędów z dodatnim i ujemnym znakiem (Hodson, 2022; Willmott i Matsuura, 2005). Wartość tego wskaźnika interpretuje się jako średnią błędów predykcji.

Ocena wpływu i znaczenia cech

Ocenę wpływu i znaczenia cech przeprowadzono z wykorzystaniem trzech metod: metody wbudowanej w bibliotekę *sci-kit learn*, metody permutacyjnej oraz metody SHAP.

Metoda wbudowana w bibliotekę *sci-kit learn*

Metoda wbudowana w bibliotekę *sci-kit learn* dla metod opartych na drzewach decyzyjnych bazuje na zmianach wartości wybranego wskaźnika nierównomierności,

na przykład miary Giniego. Za pomocą tej metody można określić, które ze zmiennych najmocniej wpływają na zmniejszenie wartości wskazanego wskaźnika.

Wadami tej metody są: zawyżanie ważności zmiennych liczbowych i przeprowadzanie obliczeń na danych treningowych. Przez to zmienne, które nie wpływają na przewidywanie zmiennej objaśnianej, mogą zostać wskazane jako ważne (Sci-kit learn, b.d.).

Metoda permutacyjna

W odróżnieniu od metody wbudowanej metoda permutacyjna może być stosowana dla większości modeli uczenia maszynowego. W metodzie tej bada się, jak zmienia się wartość miary błędu (takiej jak MAE czy R^2) po spermutowaniu wartości jednej ze zmiennych. Tę czynność wykonuje się dla wszystkich zmiennych z osobna, a następnie ranguje się je po wpływie, jaki miały na zmianę wartości wskazanej miary błędu (Jensen, 2022).

SHAP

W tej metodzie siła wpływu poszczególnych zmiennych jest prezentowana przez wyznaczenie wartości Shapleya (1953). Za pomocą wartości SHAP można wyjaśniać wpływ poszczególnych zmiennych dla dowolnego modelu uczenia maszynowego. Przy obliczaniu tych wartości wykorzystuje się podejście oparte na teorii gier, które mierzy wkład każdego gracza w końcowy wynik. W podejściu stosowanym w uczeniu maszynowym każdej funkcji przypisuje się wartość ważności reprezentującą jej udział w wynikach modelu (All Awan, 2023). Na podstawie wartości SHAP można wyznaczyć interpretowalność zarówno lokalną, jak i globalną.

4. Wyniki badań

4.1. Zebranie i przygotowanie danych

Badaniu poddano koszykarzy grających w lidze NBA w sezonie 2022-2023 oraz koszykarki występujące w WNBA w sezonie 2023. Wybrano te sezony, by zachować aktualność wyników uzyskanych w badaniu, gdyż na moment przeprowadzania analizy są to ostatnie zakończone rozgrywki.

Do badania wybrano 3 rodzaje zmiennych: opisujące podstawowe charakterystyki zawodników, ich popularność oraz skuteczność ofensywną i defensywną. Są to zmienne:

Plec – płeć zawodnika,
Wiek – wiek gracza w badanym sezonie,
Wyplata – wypłata bazowa zawodnika na badany sezon w przeliczeniu na jeden mecz w podstawowej części sezonu; jest to zmienna objaśniana,

OBS – liczba obserwujących na Instagramie na dzień 4 kwietnia 2024,
Min – średnia liczba rozegranych minut na mecz,
PTS – średnia liczba punktów zdobyta na mecz,
FGM – średnia liczba trafionych rzutów z gry na mecz,
FG% – średnia skuteczność prób zdobycia punktów z gry na mecz,
3PM – średnia liczba trafionych rzutów za trzy punkty na mecz,
3P% – średnia skuteczność prób rzutów za trzy punkty na mecz,
FTM – średnia liczba trafionych rzutów wolnych na mecz,
FT% – średnia skuteczność prób rzutów wolnych na mecz,
OREB – średnia liczba zbiórek ofensywnych na mecz,
DREB – średnia liczba zbiórek defensywnych na mecz,
AST – średnia liczba asyst na mecz,
TOV – średnia liczba strat na mecz,
STL – średnia liczba przechwytów na mecz,
BLK – średnia liczba bloków na mecz,
PF – średnia liczba popełnionych fauli na mecz,
DD2 – średnia liczba tak zwanych *double doubles* na mecz,
TD3 – średnia liczba tak zwanych *triple doubles* na mecz,
GPR – stosunek liczby zagranych meczy do liczby wszystkich meczy w sezonie,
GWR – stosunek liczby wygranych meczy do liczby wszystkich zagranych meczy.

Powyższe zmienne zaczerpnięto z następujących portali: Spotrac (b.d.a; b.d.b) popularbasketbal-lers.com (2024); Instagram (2024); NBA (b.d.b) oraz WNBA (b.d.b).

Niezbędne było zapewnienie porównywalności wartości zmiennych dla kobiet i mężczyzn, gdyż większość ze zmiennych była zależna od liczby meczów w sezonie. Podstawowe rozgrywki sezonu NBA składały się z 82 spotkań, natomiast w WNBA było to jedynie 40 spotkań, dlatego zdecydowano się przedstawić każdą ze zmiennych w przeliczeniu na 1 mecz.

Wszystkie obliczenia oraz część wizualizacji wykonano w języku Python. Pozostałe wizualizacje wykonano w Excelu.

4.2. Dobór zmiennych

Doboru zmiennych zdecydowano się dokonać na podstawie korelacji Pearsona oraz współczynnika zmienności. Postanowiono, że zmienne objaśniające nie powinny być ze sobą bardziej skorelowane niż 0,8.

Wystąpiły zmienne mocno ze sobą skorelowane, z tego powodu usunięto ze zbioru potencjalnych zmiennych objaśniających zmienne PTS, FTM oraz powiązaną z nią FT%.

Zdecydowano, że do badania mogą zostać wybrane jedynie zmienne, których zmienność jest wyższa niż 10%. Wszystkie zaproponowane zmienne miały wartość współczynnika wyższą niż założone 10%, dlatego wszystkie ze zmiennych brano pod uwagę w kolejnych etapach analizy.

4.3. Wyniki XGBoost

Modelowanie przeprowadzono z wykorzystaniem metody XGBoost, gdyż jest to model oparty na drzewach decyzyjnych, który jest nieczuły na różnicę wielkości zmiennych. Na metodach opartych na drzewach możliwe jest również przeprowadzanie analizy wpływu zmiennych na predykcję, co jest niezbędne do sprawdzenia zakładanych w tym badaniu hipotez. Zaletą tej metody jest również dobre dopasowywanie się do danych. W wyborze modelu oceniano również łatwość implementacji. XGBoost jest dostępny w wielu popularnych, dobrze udokumentowanych narzędziach.

Do utworzenia modelu wykorzystano bibliotekę XGBoost dostępną w Pythonie. Chociaż rozwiązanie to umożliwia implementację modelu z domyślnymi wartościami hiperparametrów, zdecydowano się dokonać strojenia hiperparametrów. Dokonano tego z wykorzystaniem metody przeszukiwania siatki (z angielskiego *Grid Search*). Jako minimalizowaną miarę oceny jakości modelu wybrano średni błąd kwadratowy MSE. Strojeniu poddano parametry:

- colsample_bytree – proporcja kolumn używanych do uczenia każdego drzewa,
- learning_rate – szybkość uczenia,
- max_depth – maksymalna głębokość pojedynczego drzewa,
- n_estimators – liczba drzew,
- subsample – proporcja próbek używanych do uczenia każdego drzewa.

Strojenie wykazało, że najlepsze wyniki można uzyskać dla modelu o hiperparametrach:

- colsample_bytree = 0.9,
- learning_rate = 0.2,
- max_depth = 3,
- n_estimators = 100,
- subsample = 1.0

Wartości powyżej zostały wskazane jako optymalne, dlatego wytrenowano model właśnie z tymi hiperparametrami.

4.4. Ocena jakości

Dla wytrenowanego modelu uzyskano błąd MSE w wysokości 4640102639,85, natomiast RMSE będące pierwiastkiem tej wartości wyniosło 68118,3. Oznacza to, że średnio przewidywana wartość wypłaty różni się od rzeczywistej o 68118,3 dolarów. Wartość ta jest równa niemal 74% wartości średniej zmiennej objaśnianej. Przedstawione powyżej wskaźniki mają stosunkowo wysokie wartości, co wskazuje, że model ten raczej słabo przewiduje wartość wypłaty dla poszczególnych zawodników i zawodniczek.

MAE dla tego modelu wyniosło 43394,6. Oznacza to, że średnia błędów predykcji jest równa 43394,6 dolarów. Wartość ta, choć przedstawia podobną charakterystykę do RMSE, znacznie się różni od wartości RMSE. Ta duża różnica sugeruje, że w zbiorze danych występują obserwacje, które znacząco różnią się między sobą.

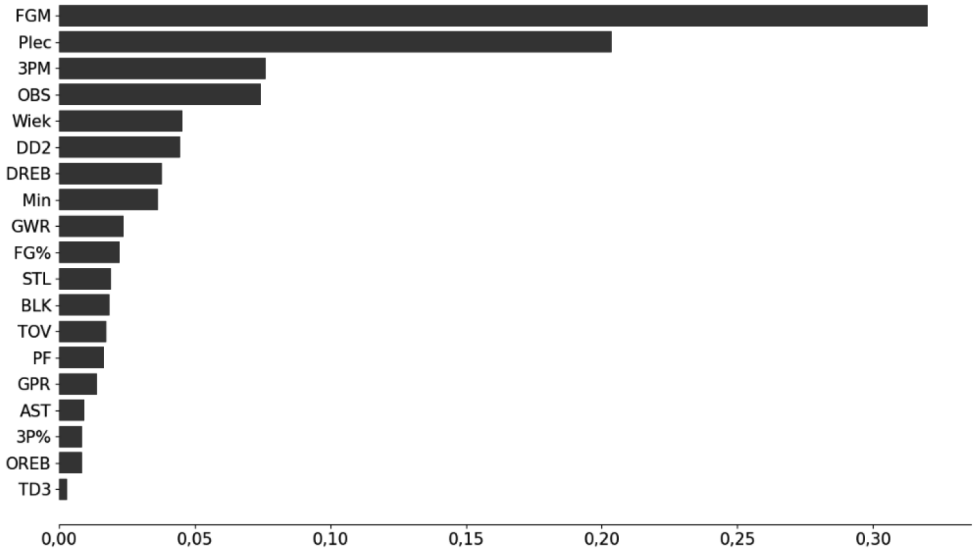
Współczynnik determinacji wyniósł 0,72, co wskazuje, że 72% wariancji jest wyjaśniana przez model. Biorąc pod uwagę, że zjawisko to jest niełatwe do przewidywania, a wysokość wypłaty często jest zależna od innych, nieuwzględnionych w tym badaniu czynników, można uznać, że zmienność tej zmiennej jest dosyć dobrze reprezentowana przez model. Trzeba pamiętać, że w rzeczywistości wysokość wypłaty jest ustalana najczęściej na podstawie osiągnięć zawodników z poprzednich lat, które nie były badane w tej analizie. Kontrakty są również ustalane przed rozpoczęciem sezonu i obowiązują kilka lat, przez co wysokość wypłaty może nie odzwierciedlać obecnej formy zawodnika. Wysokość wypłaty zapisanej w umowie między zespołem a zawodnikiem niejednokrotnie zależy też od umiejętności negocjacyjnych czy sympatii międzyludzkiej. Wszystkie te czynniki sprawiają, że zmienną Wyplata trudno jest przewidzieć za pomocą metod statystycznych, więc uzyskane dla tego modelu wyniki można uznać za zadowalające.

Wyjaśnialność globalna

By wykryć globalny wpływ poszczególnych zmiennych na predykcję modelu, zdecydowano się porównać 3 różne sposoby wyznaczania znaczenia cech: sposób wbudowany w bibliotekę *sci-kit learn* oparty na zmianach wskaźnika nierównomierności, metodę permutacyjną oraz metodę opartą na współczynniku Shapleya.

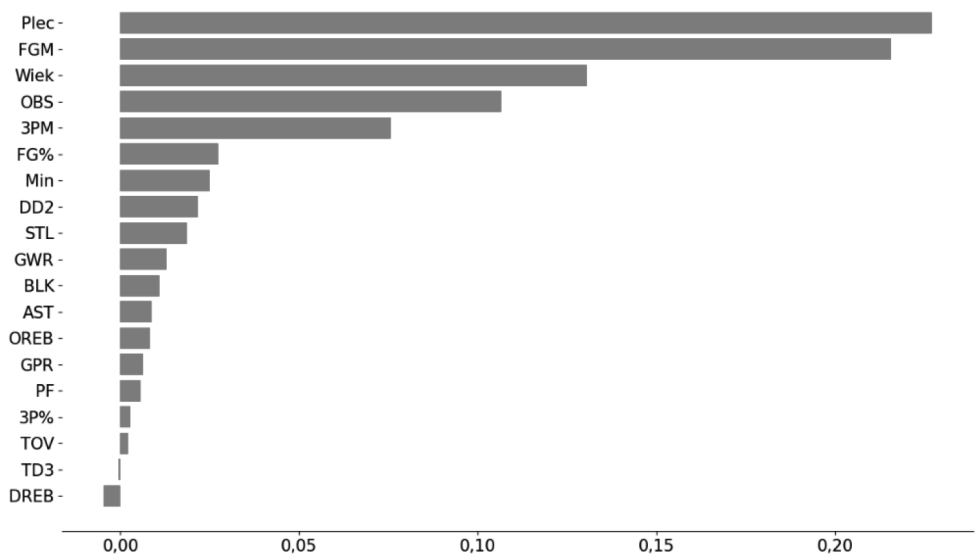
Analizując powyższe wykresy można zauważyć, że pierwsze 5 cech we wszystkich metodach jest takie same, różnią się jedynie kolejnością (rys. 2-4). Do tych cech należą: Plec, FGM (średnia liczba zdobytych punktów na mecz), Wiek, 3PM (średnia liczba udanych rzutów za 3 punkty na mecz) oraz OBS (liczba obserwujących na Instagramie). Dwie spośród tych cech są wskazywane jako ważniejsze od pozostałych: Plec i FGM. Pozostałe wybrane zmienne nie wpływają znacząco na wysokość wypłaty.

Na wykresie (rys. 5) przedstawiono wpływ poszczególnych zmiennych na wartości zmiennej objaśnianej dla każdej z obserwacji. Cechy są przedstawione osobno wierszami. Dla łatwiejszego odbioru analizy zmienne posortowano od najbardziej wpływowej do najmniej znaczącej. Kolorami zaznaczono wartości dla każdej zmiennej indywidualnie, na niebiesko zaznaczono niskie wartości każdej ze zmiennych, na różowo wysokie. Im obserwacja położona jest bliżej lewej strony wykresu, tym niższa wartość Shapleya, czyli tym bardziej wpływa ona na obniżenie wypłaty, im bardziej na prawo – tym wyższa wypłata. Jak widać na wykresie, najbardziej wpływową zmienną jest ponownie płeć. Wysokim wartościom tej zmiennej (w tym przypadku jest to wartość 1 odpowiadająca kobietom) odpowiadają niskie wartości wypłaty, co oznacza, że kobiety otrzymują niższe wynagrodzenie niż mężczyźni. Kolejną wpływową zmienną jest FGM. Wysokie wartości tej zmiennej odpowiadają wysokim i bardzo wysokim wypłatom, co sugeruje, że żeby otrzymywać wysoką wypłatę, trzeba też zdobywać dużą liczbę punktów. Wiersz dla zmiennej OBS pokazuje, że duża liczba obserwujących na Instagramie przekłada się na wyższe wynagrodzenie. Można jednak zauważyć, że niskie wartości dla tej zmiennej również



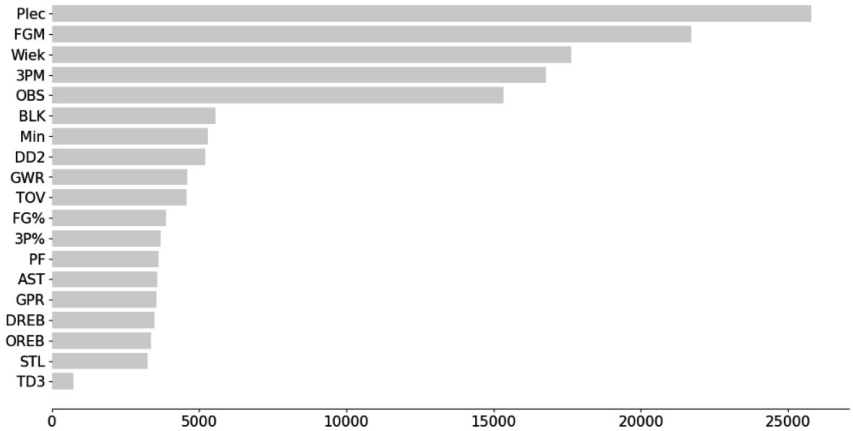
Rys. 2. Wpływ zmiennych na predykcję modelu – metoda wbudowana w bibliotekę *sci-kit learn*

Źródło: opracowanie własne.



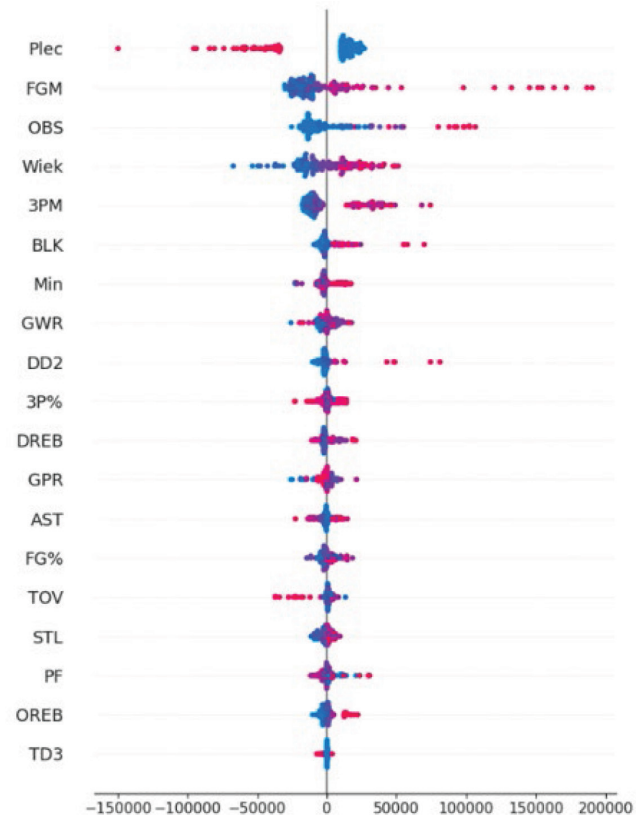
Rys. 3. Wpływ zmiennych na predykcję modelu – metoda permutacyjna

Źródło: opracowanie własne.



Rys. 4. Wpływ zmiennych na predykcję modelu – współczynnik Shapleya

Źródło: opracowanie własne.



Rys. 5. Wpływ zmiennych na wartość predykcji w podziale na obserwacje

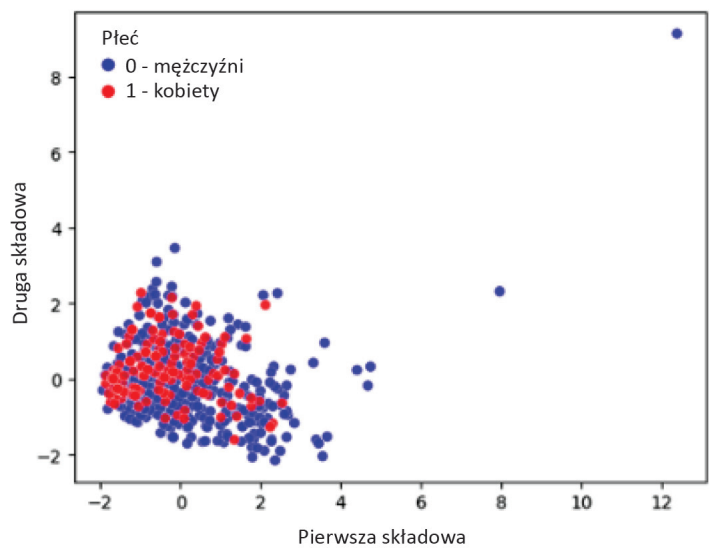
Źródło: opracowanie własne.

mogą oznaczać wynagrodzenie ponad średnią. Analizując wiersz dla wieku, zauważono, że występuje wyraźna tendencja, że im starszy zawodnik, tym wyższą wypłatę otrzymuje. Wiersz dla zmiennej 3PM pokazuje, że większość obserwacji dla tej zmiennej ma niskie wartości, które odpowiadają niskim wypłatom dla tych zawodników. Sportowcy, którzy podczas meczu średnio trafiają więcej za trzy punkty, otrzymują zazwyczaj wyższe wynagrodzenie. Pozostałe zmienne zostały wskazane przez poprzednie analizy jako mniej istotne, można jednak zauważyć tendencję, że zawodnicy, którzy przyjmują dla tych zmiennych wartości uznawane za korzystne i są one znacząco różne od średniej, najczęściej również otrzymują wyższą wypłatę.

Podsumowując wyniki analizy wpływu poszczególnych zmiennych, można zauważyć, że wszystkie 3 badane charakterystyki mają wpływ na wypłatę zawodnika: jego cechy personalne (zmienna Plec i Wiek), popularność (OBS) oraz wyniki uzyskiwane na boisku w kategorii statystyk ofensywnych (FGM i 3PM).

4.5. Szczegółowa analiza zależności

Zdecydowano się zobrazować obserwacje w przestrzeni dwuwymiarowej z uwzględnieniem jedynie zmiennych, które miały największy wpływ na predykcję modelu. W tym celu wykorzystano metodę analizy głównych składowych. Porównano, jak w grupach kobiet i mężczyzn ułożone są punkty ze względu na zmienne Wiek, OBS, FGM i 3PM. Takie przedstawienie pozwala określić, czy wartości tych zmiennych są znacząco różne między kobietami i mężczyznami.



Rys. 6. Wizualizacja dwóch pierwszych składowych głównych

Źródło: opracowanie własne.

Na rysunku 6 przedstawiono dwie pierwsze składowe dla najbardziej wpływowych zmiennych. Wyjaśniają one 70% zmienności zmiennej objaśnianej. Na czerwono zaznaczono obserwacje kobiet, a na niebiesko mężczyzn. Można zauważyć, że występuje kilka obserwacji odstających, z czego dwie z nich są wartościami ekstremalnymi. Większość tych obserwacji dotyczy mężczyzn, a tylko jedna kobiet. Pomijając wartości odstające, należy stwierdzić, że niewidoczna jest różnica między grupami kobiet i mężczyzn. Pokazuje to, że przedstawiciele obu płci nie różnią się znacząco między sobą ze względu na analizowane zmienne. Pozwala to na wysunięcie wniosku, że tak ogromna różnica w zarobkach między kobietami i mężczyznami nie wynika z ich skuteczności na boisku czy popularności w mediach społecznościowych, a właśnie jedynie z innej płci.

5. Zakończenie

Celem zaprezentowanego badania było udowodnienie, że w koszykówce występuje dyskryminacja. W tym celu stworzono model regresyjny, który w dobry sposób przewidywałby na podstawie dostarczonych danych, jakiej wysokości wypłatę powinien otrzymywać zawodnik z amerykańskich lig koszykarskich. Na podstawie wpływu zmiennych na predykcję otrzymanego modelu udowodniono, że to, jakiej płci jest badana osoba, szczególnie mocno wpływa na wysokość wypłaty. Jak wykazały szczegółowe badania, fakt, że badana obserwacja jest charakterystyką kobiety, silnie wpływa na obniżenie przewidywanej wartości. Zauważono również, że kobiety, które prezentowały wartości poszczególnych zmiennych na podobnym poziomie co mężczyźni, otrzymywały nieporównywalnie niższe wynagrodzenie. Oznacza to, że kobiety nie otrzymują odpowiedniego wynagrodzenia za wykonywaną pracę.

W badaniu wykazano, że ze względu na analizowane zmienne występuje dyskryminacja płacowa w amerykańskich ligach koszykarskich. Należy jednak pamiętać, że choć wyniki badania są jednoznaczne, mogą wystąpić czynniki, które nie zostały uwzględnione w analizie, a które mocno wpływają na to zróżnicowanie.

Wszystkie trzy badane aspekty, odnośnie do których podejrzewano, że mogą mieć wpływ na wysokość wypłaty, czyli charakterystyki zawodników, ich popularność oraz skuteczność na boisku, miały istotny wpływ na przewidywaną wartość płacy. Nieistotne okazały się jedynie zmienne, które reprezentowały skuteczność zawodników w obronie. Może to wynikać z faktu, że gracze defensywni są uznawani za mniej efektownych.

Literatura

Adelphi University. (2023). *Male vs Female Professional Sports Salary Comparison*. Pobrano 13 marca 2024 z https://online.adelphi.edu/articles/male-female-sports-salary/

All Awan, A. (2023). *An Introduction to SHAP Values and Machine Learning Interpretability*. Datacamp. Pobrano 5 maja 2024 z https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability

Analytics Vidhya. (2024). *Introduction to XGBoost Algorithm in Machine Learning*. Pobrano 5 maja 2024 z https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/

Blog Statystyczny. (2020). *Grupa ucząca, walidacyjna i testowa*. Pobrano 22 kwietnia 2024 z https://statystyczny.pl/grupa-uczaca-walidacyjna-i-testowa/

Chai, T. i Draxler, R. R. (2014). Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments against Avoiding RMSE in the Literature. *Geosci. Model Dev. 7*(3), 1247-1250, https://doi.org/10.5194/gmd-7-1247-2014

Data Science Team. (2020). *XGBoost*. Pobrano 3 maja 2024 z https://datascience.eu/pl/programowanie-komputerowe/xgboost/

Filho, M. (2022). *Does XGBoost Need Feature Scaling Or Normalization?* ? Forecastegy. Pobrano 29 kwietnia 2024 z https://forecastegy.com/posts/does-xgboost-need-feature-scaling-or-normalization/

Gratton, C. (1998). The Economic Importance of Modern Sport. *Culture, Sport, Society, 1*(1), 101-117. https://doi.org/10.1080/14610989808721803

Great Learning Team. (2024a). *Hyperparameter Tuning with GridSearchCV*. Pobrano 3 maja 2024 z https://www.mygreatlearning.com/blog/gridsearchcv/

Great Learning Team. (2024b). *What is Cross Validation in Machine Learning? Types of Cross Validation.* Pobrano 3 maja 2024 z https://www.mygreatlearning.com/blog/cross-validation/

GUS. (2017). *Wskaźniki jakości pracy*. Pobrano 2 lutego 2024 z https://stat.gov.pl/files/gfx/portalinformacyjny/pl/defaultaktualnosci/5821/10/2/1/wskazniki_jakosci_pracy_21_03_2018_pl.pdf

Hodson, T. O. (2022). Root-mean-square Error (RMSE) or Mean Absolute Error (MAE): When to Use Them or Not. Geoscientific Model Development, *15*(14),5481-5487. DOI: https://doi.org/10.5194/gmd-15-5481-2022

Instagram. (2024, 4 kwietnia). *Followers*. Pobrano 4 kwietnia 2024 z https://www.instagram.com/

Jensen, T. (2022). *Feature Importance for Any Model using Permutation*. Medium. Pobrano 3 maja 2024 z https://medium.com/@T_Jen/feature-importance-for-any-model-using-permutation-7997b7287aa

KAF. (2022). *The World's Top Basketball Leagues*. Medium. Pobrano 4 kwietnia 2024 z https://medium.com/krause-house-dao/the-worlds-top-basketball-leagues-3b634c5e426c

Mamczur, M. (2019). *Na czym polega analiza składowych głównych (PCA)?* Pobrano 21 kwietnia 2024 z https://miroslawmamczur.pl/na-czym-polega-analiza-skladowych-glownych-pca/

NBA. (b.d.a). *NBA Advanced Stats.* Pobrano 4 kwietnia 2024 z https://www.nba.com/stats/players/traditional?PerMode=PerGame&sort=PTS&dir=-1&Season=2022-23

NBA. (b.d.b). *NBA History*. Pobrano 4 kwietnia 2024 z https://www.nba.com/history

Parlament Europejski. (2023, 5 kwietnia). *Luka płacowa między kobietami a mężczyznami: definicja i przyczyny*. Pobrano 4 lutego 2024 z https://www.europarl.europa.eu/news/pl/headlines/society/20200109STO69925/luka-placowa-miedzy-kobietami-a-mezczyznami-definicja-i-przyczyny

Popularbasketballers.com. (2024, 4 kwietnia). *NBA Players Ranked by Instagram Followers*. Pobrano 4 kwietnia 2024 z https://www.popularbasketballers.com/

PQStat. (b.d.). *Analiza składowych głównych*. Pobrano 21 kwietnia 2024 z https://pqstat.pl/?mod_f=test_pca

Ruijie, S., Xiangjie, L., Yanxia, W., Shiyin, Z., Xingcan, C. i Xuejing, L. (2023). Machine Learning Regression Algorithms to Predict Short Term Efficacy after Anti VEGF Treatment in Diabetic Macular Edema Based on Real World Data. *Scientific Reports, 13*(18746). DOI: https://doi.org/10.1038/s41598-023-46021-2

Russell, S. i Norvig, P. (2020). *Artificial Intelligence: A Modern Approach (4th Edition).* Pearson.

Sci-kit learn. (b.d.). *Permutation Importance vs Random Forest Feature Importance (MDI).* Pobrano 3 maja 2024 z https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html#sphx-glr-auto-examples-inspection-plot-permutation-importance-py

Shapley, L. S. (1953). *A Value for n-Person Games. Contributions to the Theory of Games II* (s. 307-317). Princeton.

Spotrac. (b.d.a). *NBA Player Earnings*. Pobrano 2 kwietnia 2024 z https://www.spotrac.com/nba/rankings/2022-23/base/

Spotrac. (b.d.b). *WNBA Salary Rankings*. Pobrano 2 kwietnia 2024 z https://www.spotrac.com/wnba/rankings/2023/base/

Sundberg, L. i Holmström, J. (2024). Teaching Tip. Using No-Code AI to Teach Machine Learning In Higher Education. *Journal of Information Systems Education, 35*(1), 56-66. https://doi.org/10.62273/CYPL2902

Sztemberg-Lewandowska, M. (2017). Analiza niezależnych głównych składowych. *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu, 468*, 222-229. https://doi.org/10.15611/pn.2017.468.23

Thakur, A. (2022). Gender Discrimination in Sports. *International Journal of Law Management & Humanities, 5*(3), 210-219. https://doij.org/10.10000/IJLMH.113056

Tomaszewska, J. (2004). Dyskryminacja ze względu na płeć. *Dyskryminacja ze względu na płeć i jej przeciwdziałanie*. Pobrano 15 marca 2024 z https://streswpracy.pl/wp-content/uploads/2022/bzp/DYSKryminacja_ze_wzgledu_na_plec_podr.pdf

Walesiak, M. (1996). *Metody analizy danych marketingowych.* Państwowe Wydawnictwo Naukowe.

Waśkowski, Z. (2011). Integracyjna rola sportu we współczesnym świecie. *Ekonomiczne Problemy Usług*, (78), 23-32.

Willmott, C. J. i Matsuura, K. (2005). Advantages of the Mean Absolute Error (MAE) Over the root Mean Square Error (RMSE) in Assessing Average Model Performance. *Clim. Res.,* (30), 79-82, https://doi.org/10.3354/cr030079

WNBA. (b.d.a). *History*. Pobrano 4 kwietnia 2024 z https://www.wnba.com/history

WNBA. (b.d.b). *WNBA Stats*. Pobrano 4 kwietnia 2024 z https://stats.wnba.com/players/traditional/?sort=PTS&dir=-1&Season=2023&SeasonType=Regular%20Season

Diagnosis of Pay Inequality among Athletes Using Machine Learning Methods

Abstract: The main goal of the work is to prove the existence of a wage gap in sports using the XGBoost method based on the salaries of basketball players from the NBA and WNBA leagues. The truthfulness of the statement that the amount of pay is influenced by the physical characteristics, popularity, and effectiveness on the pitch in defense and attack of a player was also checked. One model was created, the results of which were considered satisfactory. The influence of individual variables was compared for this model using three different methods. Obtained results confirmed that, apart from defensive effectiveness, all of the assumed factors influencing the amount of pay are significant. The occurrence of gender-based pay discrimination was also confirmed.

Keywords: gender pay gap, sport, machine learning, XGBoost