Raport zaliczeniowy
Analiza cen zawodników angielskiej Premier League w zależności od ich statystyk

Maksym Selishchev
Informatyka Stosowana
Politechnika Wrocławska
MSiD Lab 9:15

May 19, 2024

1 Opis problemu

Problemem wybranym do badań jest przewidywanie cen zawodników angielskiej Premier League w zależności od różnych czynników. Analiza zbioru może wykazać przybliżoną cenę zawodnika na podstawie jego statystyk, drużyny, wieku itd. Celem projektu jest zbadanie zależności od dostępnych parametrów:

- Liczba zagranych minut w sezonie
- Drużyna
- Wiek
- Liczba bramek i asyst
- Pozycja
- Kraj pochodzenia

oraz stworzenie modelu estymującego cenę na podstawie wybranych danych.

2 Zbiór danych i jego przetwarzanie

1) Zbiór danych

W pracy zostały wykonane 4 zbiory danych:

1. Zbiór 1 "Dane zawodników i ich cena w konkretnym terminie" pozyskany z Kaggle. Zawiera dane o id zawodnika oraz jego cenę za ostatnie 25 lat.
2. Zbiór 2 "Dane osobiste zawodników" posiada id, imię, nazwisko, ligę, kraj pochodzenia, miasto urodzeina i wiele innych informacji zawodnika. Faktycznie zbiór jest powiązany ze zbiorem 1 po id. Też pozyskany z Kaggle.
3. Zbiór 3 "Statystyka zawodników angielskiej Premier Ligue w sezonie 2018/2019". Zawiera dane statystyczne graczy takie jak liczba minut zagranych w sezonie, strzałów, bramek, asyst i wiele innych. Pozyskany z footystats.
4. Zbiór 4 tabela wynników drużyn w angielskiej Premier Ligue w sezonie 2018/2019 zawiera posortowane według liczby punktów drużyny, liczbę punktów oraz cenę całej drużyny. Pozyskany z Eurosport.com

2) Przetwarzanie danych do analizy

1. W zbiorze 1 zostały tylko kolumny reprezentujące id zawodnika oraz jego cenę w terminach '2019-04-01'-'2019-07-01'.

Table 1: Cena zawodników

Id	Cena
321	80000000
204	12000000

2. W zbiorze 2 zostały tylko kolumny reprezentujące id zawodnika oraz jego imię i nazwisko. Zbiór został połączony ze zbiorem 1 po id zawodnika

Table 2: Dane zawodników

Id	Cena	Imię i nazwisko
321	80000000	Cristiano Ronaldo
204	12000000	Lionel Messi

3. W zbiorze 3 zostały tylko kolumny reprezentujące podstawowe statystyki zawodnika w sezonie i też został połączony z dwoma zbiorami wyżej.

Table 3: Statystyka zawodników

Id	Cena	Imię i nazwisko	Liczba zagranych minut	Liczba bramek	Klub
321	80000000	Cristiano Ronaldo	1234	34	Manchester United
204	12000000	Lionel Messi	5124	40	Chelsea
Pozycja		Kraj pochodzenia	Liczba asyst	Wiek	
Napastnik		Portugalia	3	38	
Napastnik		Argentyna	9	35	

4. Do zbioru 4 ręcznie była dodana cena klubu i zostały tylko nazwy, punkty i cena

Table 4: Tabela klubów

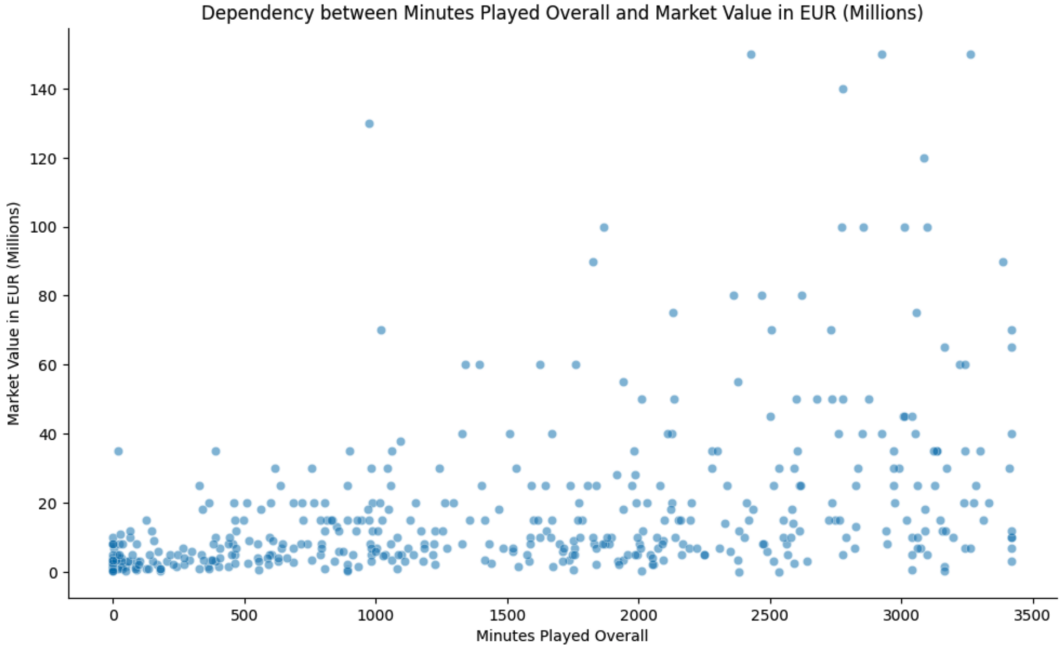
Klub	Punkty	Cena
Machester United	32	10000000
Manchester City	54	20000000

Zbiór 4 też został połaczony z 3 zbiorami wyżej po połączeniu okazało się, że tylko 354/447 zawodników mają przypisane punkty i ceny klubów, dla wypełnienia pustych komórek była wykorzystana losowa liczba między średnią z kolumny +- 30 procent.

3 Eksperymenty

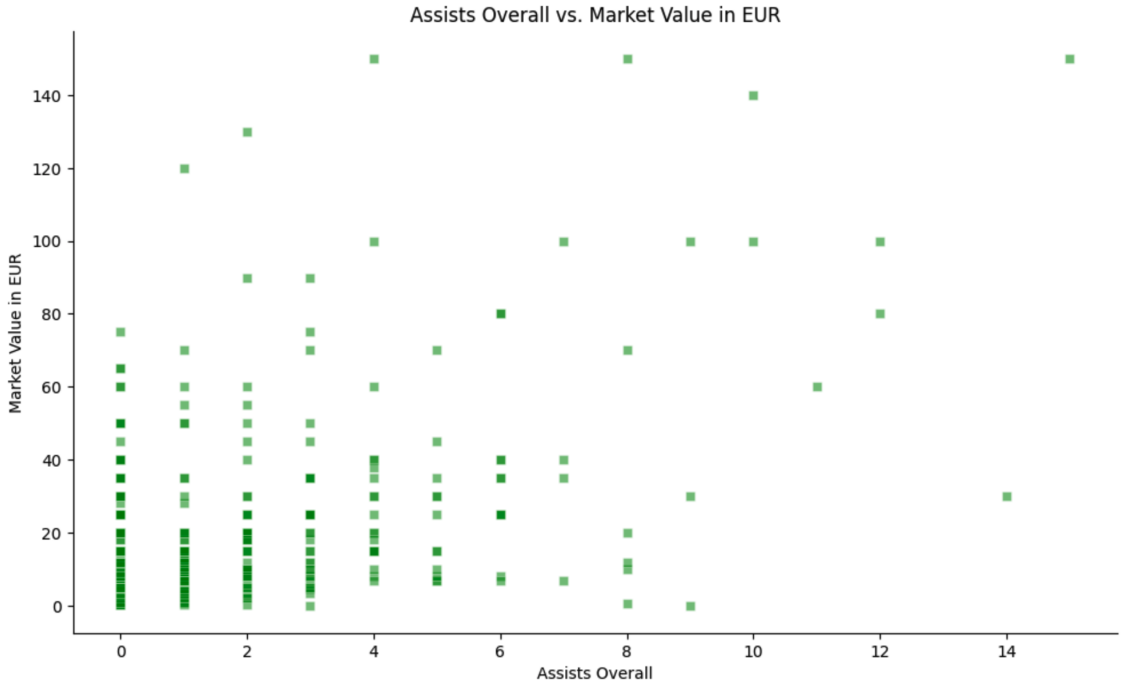
Analiza zależności

Zależność ceny od minut zagranych w sezonie :

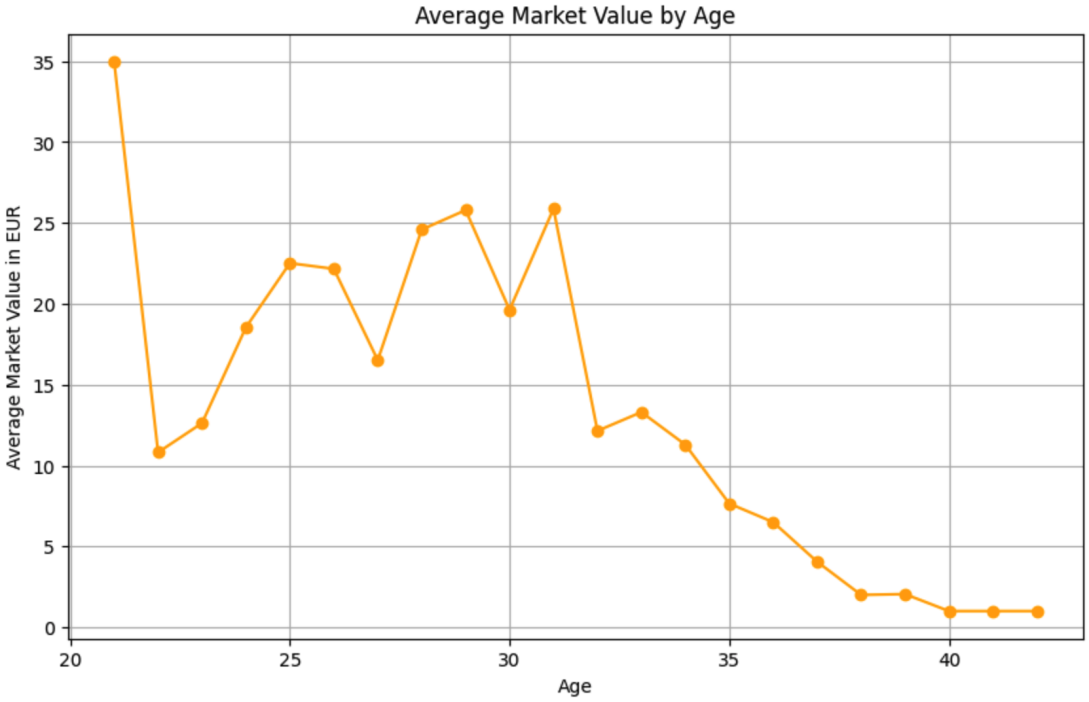


Zależność ceny od strzelonych bramek :

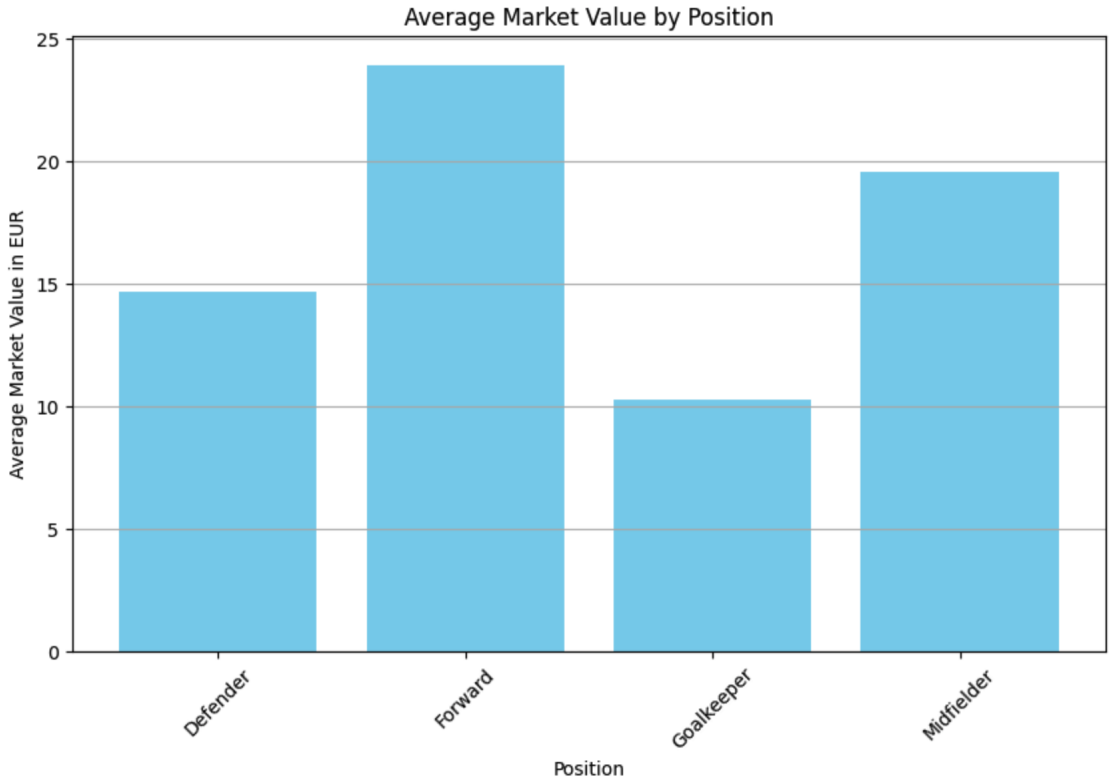
Zależność ceny od asyst :



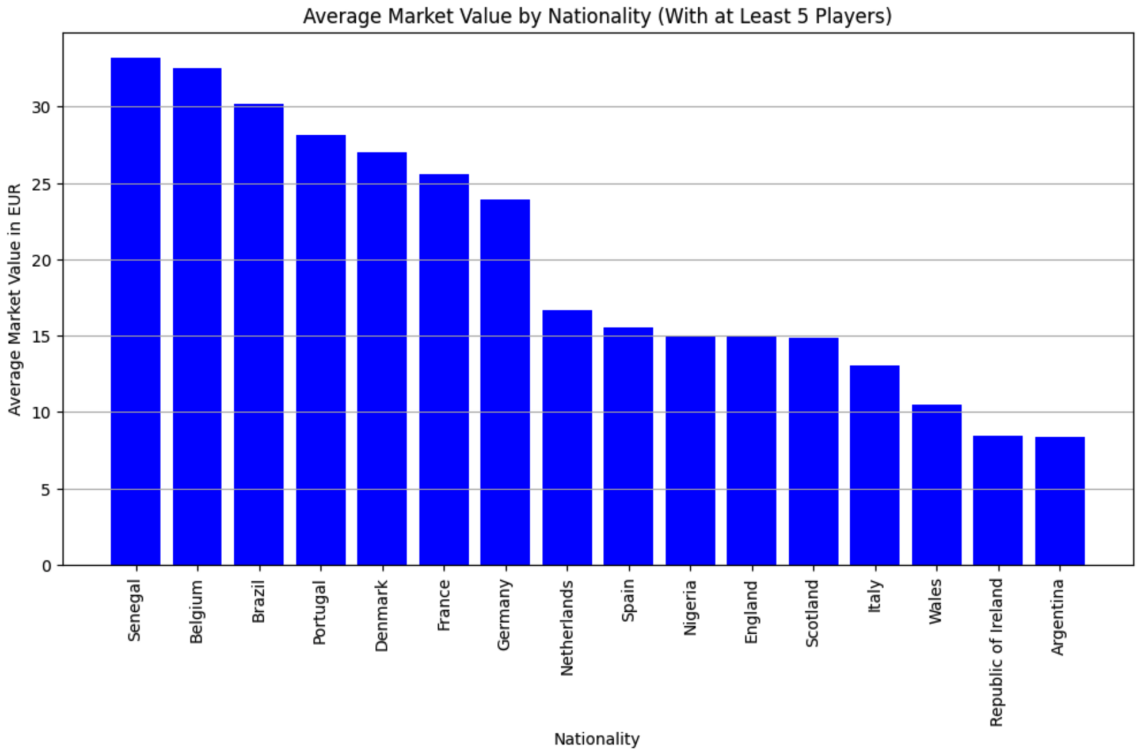
Zależność ceny od wieku :



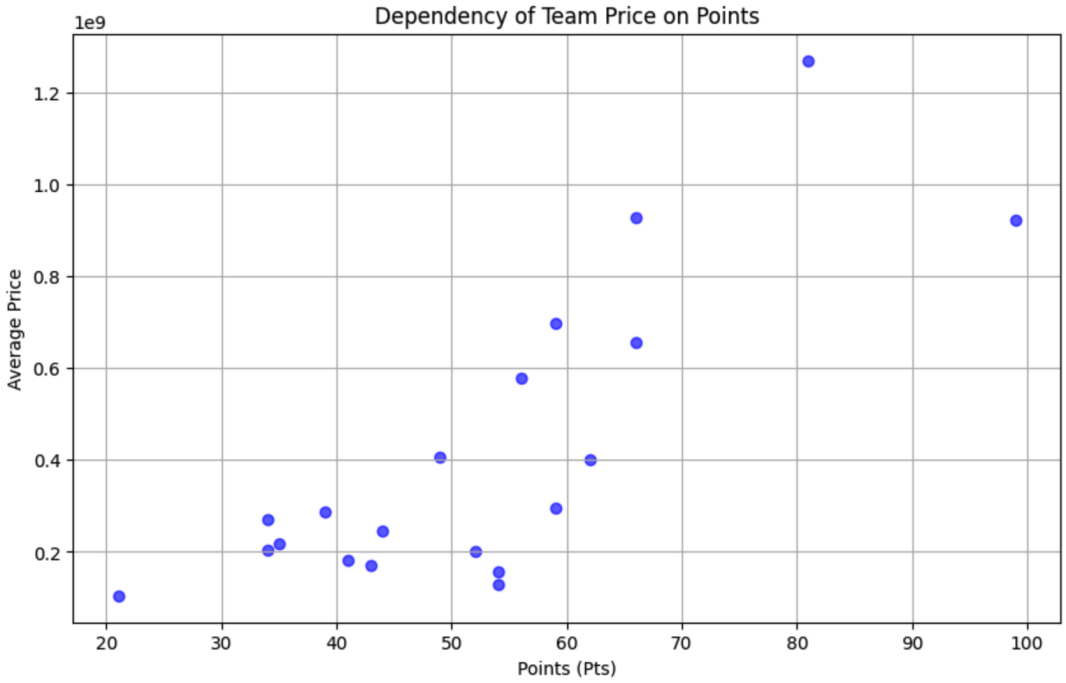
Zależność ceny od pozycji :



Zależność ceny od kraju pochodzenia :



Zależność średniej ceny zawodnika klubu od liczby punktów w lidze :



Wyszukiwanie zależnośći

1) Zależność ceny od minut zagranych w sezonie: Na podstawie danych wykresu, wydaje się, że cena zawodnika rośnie wraz z ilością minut spędzonych na boisku. Możemy przypuszczać, że kluby skupiają się na zawodnikach, którzy mają dużą liczbę minut gry.

2) Zależność ceny od strzelonych bramek: Istnieje widoczna tendencja wzrostowa między ceną zawodnika a liczbą bramek, co sugeruje, że zawodnicy, którzy zdobywają więcej goli, są bardziej pożądani na rynku.

3) Zależność ceny od asyst: Wydaje się, że liczba asyst również wpływa na cenę zawodnika, choć nie tak wyraźnie jak liczba bramek. Zawodnicy, którzy potrafią doskonale współpracować z innymi, mogą być wyceniani wyżej.

4) Zależność ceny od wieku: Istnieje pewna zależność między wiekiem a ceną, gdzie młodsi zawodnicy mogą być wyceniani wyżej ze względu na swoją perspektywę rozwoju, podczas gdy starsi zawodnicy mogą być mniej atrakcyjni ze względu na potencjalny spadek wydajności.

5) Zależność ceny od pozycji: Ceny zawodników mogą się różnić w zależności od ich pozycji na boisku. Na przykład, napastnicy mogą być wyceniani wyżej niż obrońcy ze względu na ich umiejętność zdobywania bramek.

6) Zależność ceny od kraju pochodzenia: Cena zawodnika może być również uzależniona od jego kraju pochodzenia, gdzie zawodnicy z bardziej konkurencyjnych lig lub popularnych krajów mogą być wyceniani wyżej.

7) Zależność średniej ceny zawodnika klubu od liczby punktów w lidze: Możemy zauważyć, że kluby z większą liczbą punktów w lidze mogą mieć droższych zawodnikówśś

4 Tworzenie modeli

1) Przygotowywanie danych

Dla używania zależności kraju pochodzenia i pozycji od cena, należało zamienic tekst (String) na liczbę odpowiadające każdemu kraju/pozycji.

2) Podział danych

Dane zostały podzielone na treningowe oraz testowe za pomocą funkcji w bibliotece sklearn.

3) Dane do przeanalizowania

Po przeprowadzeniu różnych prób, najlepsze wyniki uzyskano z modelami uwzględniającymi następujące cechy:

- Wiek
- Pozycja
- Liczba zagranych minut
- Kraj pochodzenia
- Liczba bramek
- Liczba asyst
- Liczba punktów drużyny
- Cena drużyny

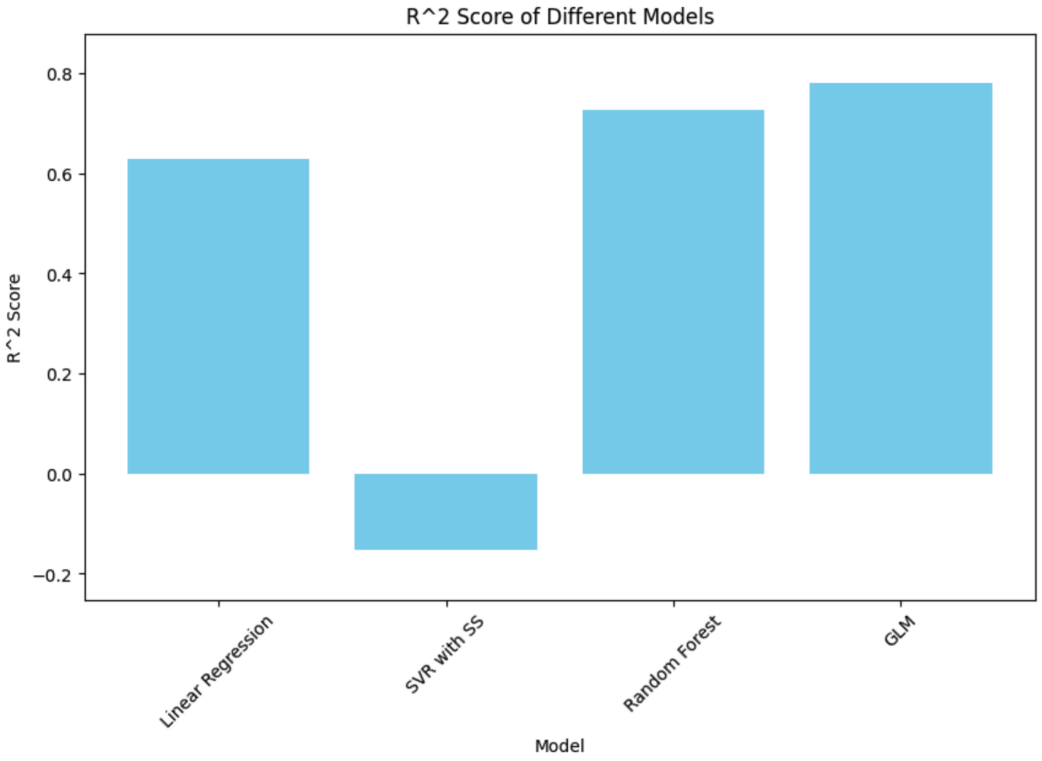
4) Tworzenia modeli przewidywających

Na podstawie zbiorów naszych danych były trenowane 4 modele :

- Regresja Liniowa
- Maszyna wektorów nośnych (SVM with Standart Scaler)
- Metoda Lasu Losowego (Random Forest)
- Uogólniony model liniowy (GLM)

Po przetrenowaniu i przewydywaniu mamy następną sytuację:

```
[0.6273014363936666, -0.15204071479532888, 0.725328341528098, 0.7790915217538648]
```



Współczynnik determinacji R^2 jest miarą, która wskazuje, jak dobrze model regresyjny wyjaśnia zmienność danych. Jego wartość wynosi od 0 do 1:

- $R^2 = 1$: Model idealnie wyjaśnia zmienność danych.
- $R^2 = 0$: Model nie wyjaśnia żadnej zmienności danych.
- **Im bliżej 1, tym lepsze dopasowanie modelu do danych.**
- **Im bliżej 0, tym gorsze dopasowanie modelu do danych.**

Jak można zauważyć, model maszyny wektorów nośnych (SVM) nie jest efektywny w naszym przypadku.

Dla analizy skuteczności modeli można dodać Średni błąd bezwzględny (MAE), Średni błąd kwadratowy (MSE) oraz Pierwiastek średniokwadratowy błędu (RMSE).

Po dodaniu tych atrybutów do naszego R^2 i usunięciu z analizy **maszyny wektorów nośnych**, mamy:

```
Linear Regression Scores:
Mean Absolute Error: 11364533.393165076
Mean Squared Error: 294612873553005.2
Root Mean Squared Error: 17164290.651029106
R2 score: 0.6273014363936666
```

```
Random Forest Scores:
Mean Absolute Error: 9015324.813988095
Mean Squared Error: 217124009824333.94
Root Mean Squared Error: 14735128.429176787
R2 score: 0.725328341528098
```

```
GLM Scores:
Mean Absolute Error: 8323452.593141168
Mean Squared Error: 174624986312154.0
Root Mean Squared Error: 13214574.768495353
R2 score: 0.7790915217538648
```

Widać, że najefektywniejszym modelem przewidywającym jest **uogólniony model liniowy (GLM)**

5 Wnioski

Analiza i przewidywanie cen zawodników na podstawie różnych zbiorów danych wykazała, że model **GLM** jest najbardziej efektywny w tym zadaniu. Zostało to potwierdzone porównaniem wyników MAE, MSE, RMSE oraz R^2 z innymi modelami. Pobieranie danych z różnych źródeł i ich powiązanie między sobą dało możliwość zaprojektowania bardziej dokładnego modelu.

6 Źródła

https://www.kaggle.com/datasets/davidcariboo/player-scores/data
https://www.kaggle.com/code/davidcoxon/football-transfer-market-eda-basic-modelling
https://www.eurosport.com/football/premier-league/2019-2020/standings.shtml
https://footystats.org/download-stats-csv