Analiza posiadania piłki drużyn piłkarskich w zależności od różnych czynników meczowych

Dominik Patrzek
WIT IST
266585@student.pwr.edu.pl
MSiD 17:05 TP

15 czerwca 2023

I Wstęp

Wybranym do analizy zagadnieniem jest posiadanie piłki w zależności od różnych czynników. Drużyny mające dużo posiadania grają atrakcyjny futbol, przyjemny dla oka kibica. Jakie więc kryteria należy spełniać, aby zainteresować więcej ludzi swoimi meczami? W badaniu posiadanie piłki zostanie zbadane w zestawieniu z parametrami takimi jak:

- liczba strzałów na bramkę
- liczba oczekiwanych goli
- liczba fauli drużyny przeciwnej
- poziom drużyny
- liczba rzutów rożnych
- pogoda

Ponadto stworzony zostanie model, który estymuje posiadanie piłki drużyny na podstawie wybranych czynników.

II Zbiór danych i jego przetwarzanie

A Zbiór danych

Zbiór danych został pozyskany ze źródła footystats[2]. Przedstawia on wiele statystyk dotyczących wszystkich meczów Premier League z sezonu 2018/19 zakończonego tryumfem drużyny Manchesteru City. Posiadanie piłki jest wartością procentową, reszta statystyk to liczby całkowite i rzeczywiste dodatnie. Ponadto użyta została Weather API ze źródła visualcrossing[1], zostały do niej wysyłane zapytania, aby otrzymać dane pogodowe w miejscu danego wydarzenia.

B Przetwarzanie wstępne

Ze zbioru danych zostału usunięte dane, które nie będą brały udziału w badaniu. Pod uwagę będą brane statystyki posiadania piłki itd. w kontekście drużyny gospodarza meczu.

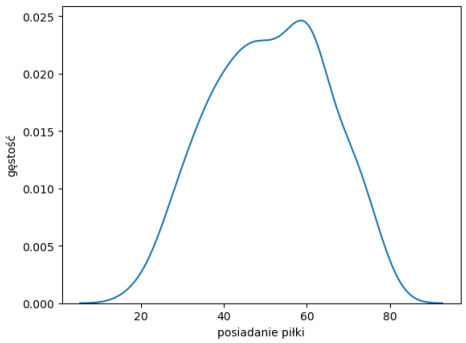
C Analiza eksploracyjna

Spójrzmy na cechy zbioru posiadania piłki

Wartość min.	18%
Wartość max.	80%
Mediana	51.0%
Odchyl. standardowe	13.84%
Najczęściej wyst. wartość	62%
Najrzadziej wyst. wartość	79%

Tabela 1: Cechy posiadania piłki

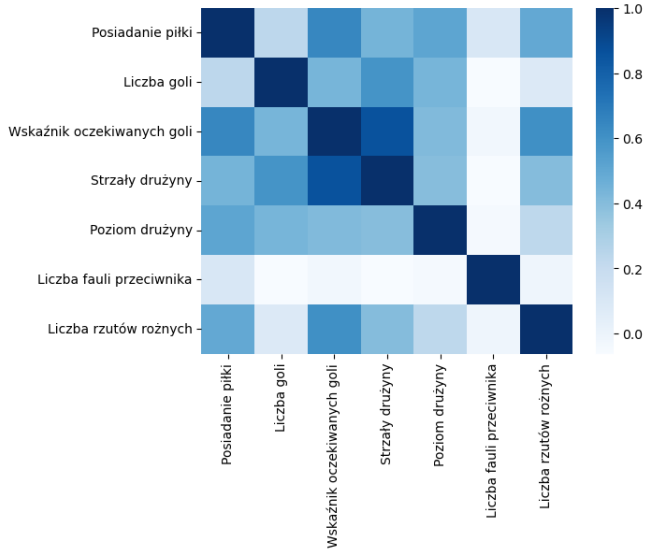
Odchylenie standardowe, które otrzymano może być uważane za umiarkowane. Oznaczające pewną zmienność, ale nie jest to ekstremalnie duże odchylenie.



Rysunek 1: Rozkład posiadania piłki

Po przetwarzaniu wstępnym zbiór zawiera dane 380 meczów scharakteryzowanych 13 cechami.

W mapie korelacji zmiennych można zauważyć wysoką zależność między posiadaniem piłki, a wskaźnikiem oczekiwanych goli i poziomem drużyny.



Rysunek 2: Mapa korelacji zmiennych

III Eksperymenty

1 Opis eksperymentów

a) Eksperyment 1 - dopasowanie modelu regresji liniowej oraz wielomianu n-tego stopnia do danych dotyczących liczby strzałów oddanych przez drużynę w meczu.

b) Eksperyment 2 - dopasowanie modelu regresji liniowej oraz wielomianu n-tego stopnia do danych dotyczących osiągniętego współczynnika goli oczekiwanych przez drużynę.

c) Eksperyment 3 - dopasowanie modelu regresji liniowej oraz wielomianu n-tego stopnia do danych dotyczących liczby fauli popełnionych przez drużynę przeciwną.

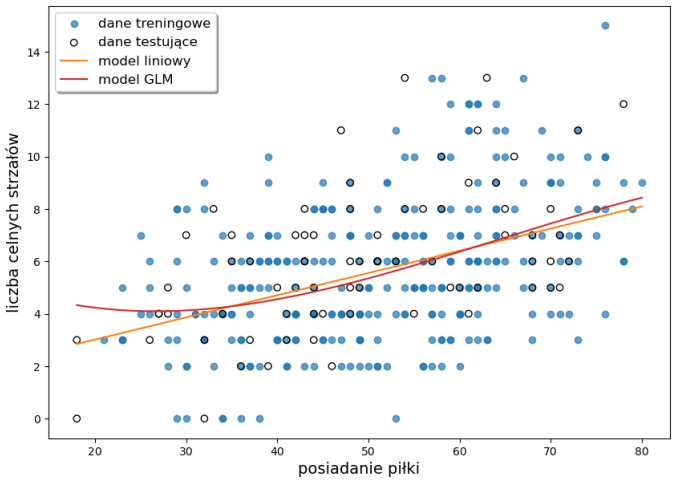
d) Eksperyment 4 - dopasowanie modelu regresji liniowej oraz wielomianu n-tego stopnia do danych dotyczących poziomu drużyny grającej mecz (mierzona średnią punktowania drużyny na mecz).

e) Eksperyment 5 - dopasowanie modelu regresji liniowej oraz wielomianu n-tego stopnia do danych dotyczących liczby rzutów rożnych wykonanych przez drużynę.

f) Eksperyment 6 - dopasowanie modelu regresji liniowej oraz wielomianu n-tego stopnia do danych dotyczących pogody w trakcie meczów Manchesteru City - najlepszej drużyny sezonu.

2 Właściwe eksperymenty

a) Eksperyment 1



Rysunek 3: Posiadanie piłki, a liczba strzałów celnych

Regresja liniowa jest postaci:

$$y = 0.08468 + 1.32104 \tag{1}$$

Parametry wielomianu stopnia 3:

a	b	c	d
-0.2102	0.005	0	6.65021

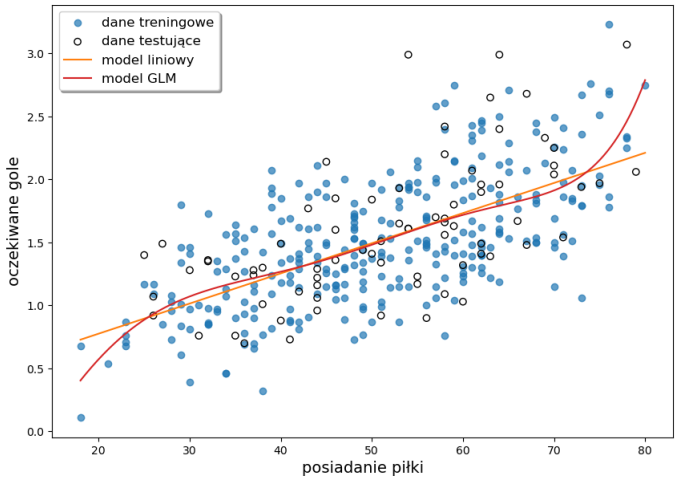
Dane zostały podzielone na zbiór treningowy i testowy w proporcji 80:20.

Błąd średniokwadratowy:

- Regresja liniowa:
  - dane treningowe: 6.06
  - dane testowe: 5.34
- Model GLM:
  - dane treningowe: 6.06
  - dane testowe: 5.61

Dane są rozrzucone po wykresie, więc błąd średniokwadratowy jest dość duży, Współczynnik regresji liniowej dodatni, tendencja jest wzrostowa.

b) Eksperyment 2



Rysunek 4: Posiadanie piłki w zależności od oczekiwanych goli

Regresja liniowa jest postaci:

$$y = 0.02377x + 0.32897 \qquad (2)$$

Parametry wielomianu stopnia 3:

a	b	c	d
0.0908	-0.0015	0	-0.56232

Dane zostały podzielone na zbiór treningowy i testowy w proporcji 80:20.

Błąd średniokwadratowy:

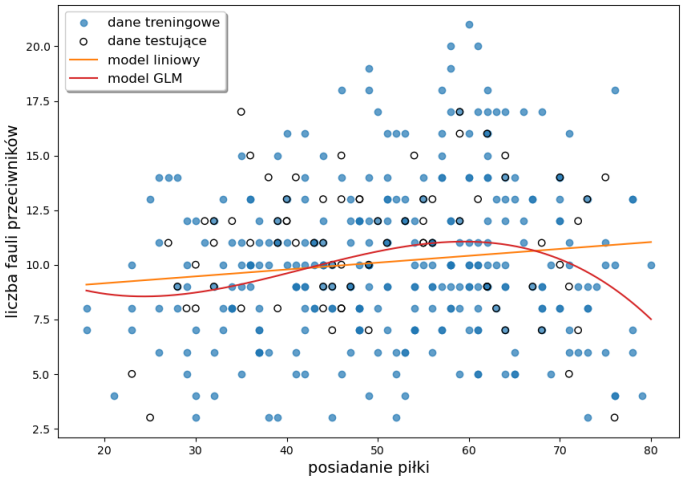
- Regresja liniowa:
  - dane treningowe: 0.16
  - dane testowe: 0.15
- Model GLM:
  - dane treningowe: 0.16
  - dane testowe: 0.149

Najlepiej dopasowanym wielomianem okazał się wielomian o stopniu 3. Dane są skupione, widać silną zależność między nimi, o czym świadczą małe wartości błędów średniokwadratowych dla danych treningowych oraz testowych. W tym kontekście należy wziąć też pod uwagę skale wartości oczekiwanych goli. Współczynnik regresji liniowej jest dodatni, tendencja danych jest wzrostowa.

c) Eksperyment 3

Regresja liniowa jest postaci:

$$y = 0.03143x + 8.52269 \qquad (3)$$



Rysunek 5: Posiadanie piłki, a faule popełnione przez przeciwnika

Parametry wielomianu stopnia 3:

a	b	c	d
-0.4992	0.0145	-0.001	13.78565

Dane zostały podzielone na zbiór treningowy i testowy w proporcji 80:20.

Błąd średniokwadratowy:

- Regresja liniowa:
  - dane treningowe: 12.7
  - dane testowe: 9.39
- Model GLM:
  - dane treningowe: 12.7
  - dane testowe: 8.85

Dane są bardzo rozrzucone, wysoki błąd średniokwadratowy. Posiadanie piłki nie ma wiele wspólnego z liczbą fauli przeciwnika.

d) Eksperyment 4

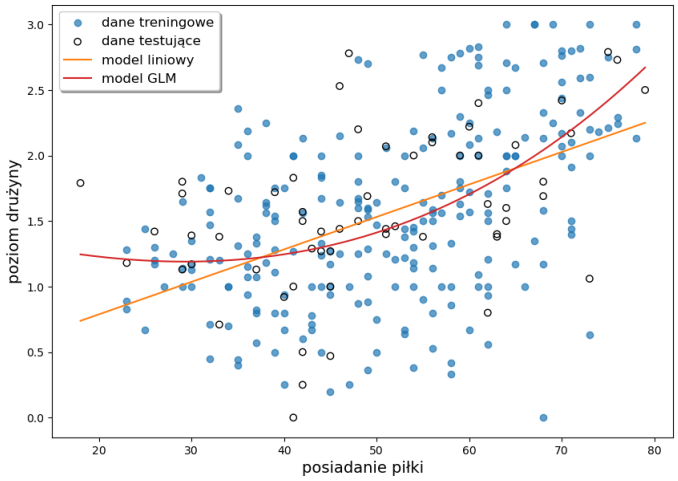
Regresja liniowa jest postaci:

$$y = 0.02477x + 0.29257 \qquad (4)$$

Parametry wielomianu stopnia 3:

a	b	c	d
-0.0198	0.0002	0	1.51629

Dane zostały podzielone na zbiór treningowy i testowy w proporcji 80:20.



Rysunek 6: Posiadanie piłki, a poziom drużyny(pkty/mecz)

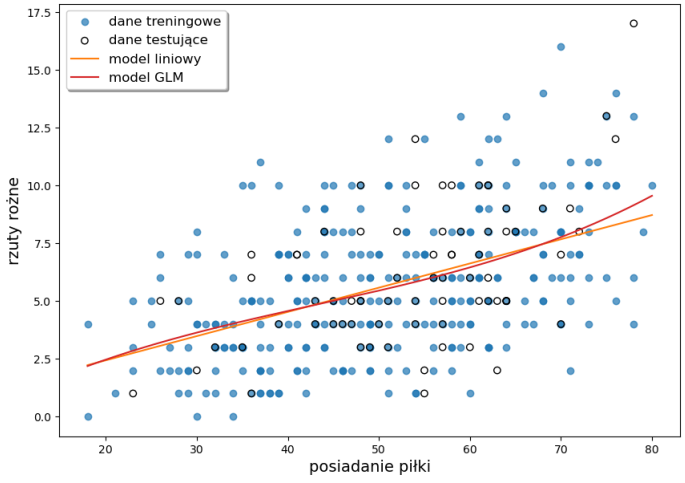
Błąd średniokwadratowy:

- Regresja liniowa:
  - dane treningowe: 0.359
  - dane testowe: 0.296
- Model GLM:
  - dane treningowe: 0.359
  - dane testowe: 0.276

Przed wykonaniem eksperymentu postanowiono odrzucić dane z pierwszych 8 kolejek, ponieważ średnia punktowa wtedy przyjmuje bardzo skrajne wartości, z każdą kolejką średnia lepiej się kształtowała.

Zauważono wysoką zależność pomiędzy poziomem drużyny a posiadaniem piłki. Drużyny, które wysoko punktują, częściej się utrzymują przy piłce i kreują sytuacje.

e) Eksperyment 5



Rysunek 7: Posiadanie piłki, a rzuty rożne

Regresja liniowa jest postaci:

$$y = 0.1047 + 0.33929 \tag{5}$$

Parametry wielomianu stopnia 3:

a	b	c	d
0.2528	-0.0039	0	-1.29023

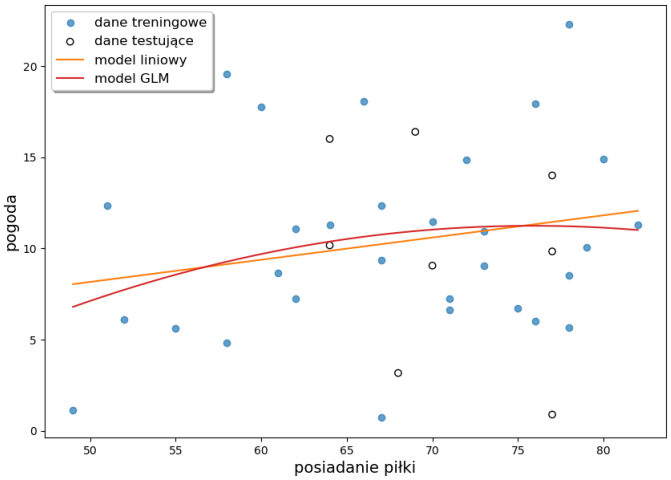
Dane zostały podzielone na zbiór treningowy i testowy w proporcji 80:20.

Błąd średniokwadratowy:

- Regresja liniowa:
  - dane treningowe: 6.97
  - dane testowe: 6.54
- Model GLM:
  - dane treningowe: 6.97
  - dane testowe: 6.3

Najlepiej dopasowanym wielomianem okazał się wielomian 3. stopnia. Dane są jednak bardzo mocno rozrzucone, Wartości błędów średniokwadratowych dla danych treningowych oraz testowych są bardzo duże. Współczynnik regresji liniowej jest dodatni, tendencja danych jest wzrostowa.

f) Eksperyment 6



Rysunek 8: Posiadanie piłki Manchesteru City, a pogoda

Regresja liniowa jest postaci:

$$y = 0.12178 + 2.07017 \tag{6}$$

Parametry wielomianu stopnia 2:

a	b	c
0.9351	-0.0062	-24.2291

Dane zostały podzielone na zbiór treningowy i testowy w proporcji 80:20.

Błąd średniokwadratowy:

- Regresja liniowa:
  - dane treningowe: 25.3
  - dane testowe: 30.9
- Model GLM:
  - dane treningowe: 25.3
  - dane testowe: 30.1

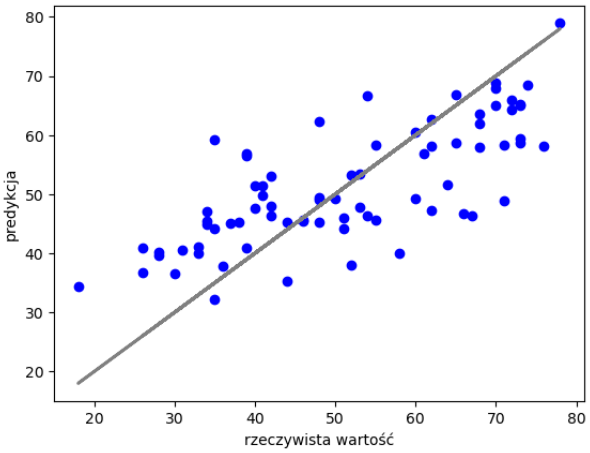
Z obserwacji wynika, że posiadanie piłki w przypadku tej drużyny nie wynika w żadnym stopniu z warunków pogodowych. Dane są mocno rozrzucone po wykresie.

IV Modelowanie

Stworzony został model predykujący posiadanie piłki. Został wytrenowany na parametrach które wykazały większą zależność z posiadaniem piłki niż pozostałe. Był to współczynnik oczekiwanych goli, poziom drużyny i liczba strzałów celnych w meczu. Najlepszym modelem okazał się ten wykorzystujący regresję liniową. Jego współczynnik determinacji wynosi

$$R^2 = 0.58 \tag{7}$$

To sugeruje, że istnieje umiarkowana zależność między tymi czynnikami a posiadaniem piłki.



Rysunek 9: Rzeczywista wartość a wartość predykowana

Wykres porównuje rzeczywistą wartość ze zbioru testowego i wartość predykowaną. Jeśli punkt na wykresie znajduje się na prostej y=x, oznacza to że model przewidział dobrze wartość.

Model jest w stanie w pewnym stopniu przewidywać, która drużyna będzie miała kontrolę nad piłką na podstawie tych zmiennych. Jednak pozostałe 42% wariantów pozostaje niewyjaśnione przez model liniowy, co sugeruje, że istnieją inne czynniki, które również wpływają na posiadanie piłki i nie zostały uwzględnione w analizie.

Dokonano jeszcze klasyfikacji zbioru na trzy klasy charakteryzujące posiadanie:

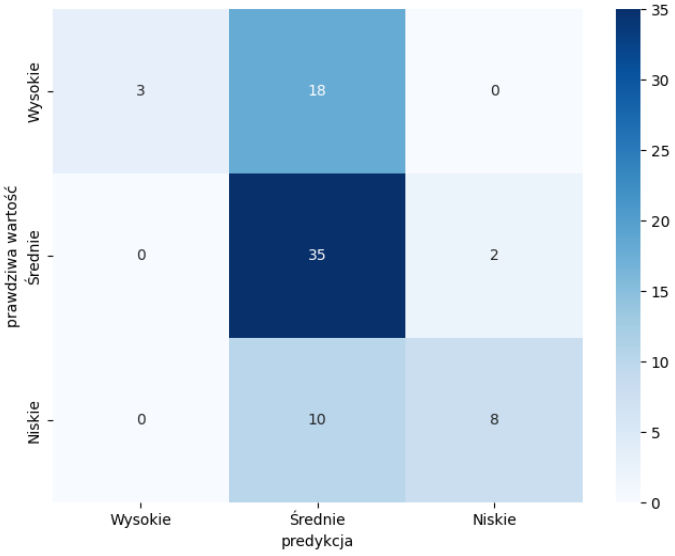
- Niskie: $\leqslant 38\%$
- Średnie: 39%-64%
- Wysokie: $\geqslant 65\%$

Wartości przedziałów zostały określone na podstawie mediany i odchylenia standardowego. Podziału dokonano używając regresji logistycznej.

Klasa	Precyzja	Czułość	Miara F1
Wysokie	1.00	0.14	0.25
Średnie	0.56	0.95	0.70
Niskie	0.80	0.44	0.57
Dokładność			0.61
Średnia	0.79	0.51	0.51
Ważona	0.74	0.61	0.55

Tabela 2: Wyniki klasyfikacji

Model regresji logistycznej osiągnął dobre wyniki precyzji dla klas "Wysokie" i "Niskie", ale słabsze wyniki dla klasy "Średnie". Czułość była za to wysoka dla klasy "Średnie". Miary F1 dla klas "Średnie" i "Niskie" są dobre.



Rysunek 10: Macierz pomyłek

Na macierzy pomyłek widać dobrze, że największa trudność zachodziła dla klasy "Średnie".
Model potrafi dobrze sklasyfikować 61% obiektów. Wartości średniej i ważonej średniej wskazują, że ogólnie model ma równowagę, ale istnieje jeszcze pole do poprawy, szczególnie dla klasy "Wysokie".

V Wnioski

Po przeprowadzeniu eksperymentów należy stwierdzić, że posiadanie piłki nie zależy od liczby fauli przeciwnika, pogody, czy liczby rzutów rożnych. Statystyki te w meczu przyjmują różne wartości, jest w nich dużo losowości.

Zauważono jednak solidną zależność między posiadaniem piłk, a wskaźnikiem oczekiwanych goli, poziomem drużyny, czy też liczbą oddanych celnych strzałów. Te statystyki mówią, że drużyna stwarza sytuacje bramkowe, przeważa, a więc posiada też piłkę. Udało się więc stworzyć model estymujący posiadanie piłki drużyny na podstawie wybranych czynników, który radzi sobie umiarkowanie.

Posiadanie piłki w meczu piłki nożnej może być wynikiem wielu czynników, również tych niemierzalnych lub ciężko mierzalnych takich jak taktyka drużyny, strategia przeciwników, zależność od wydarzeń boiskowych, kontuzji, zmęczenia, pewnej losowości.

Bibliografia

[1] Visual Crossing. Weather API. URL: https://www.visualcrossing.com.

[2] Footystats. Premier League statistics. URL: https://footystats.org.