Klasyfikacja wyników meczów NBA - czy drużyna gospodarzy wygra?

Jędrzej Kocięcki
Informatyka Stosowana, Politechnika Wrocławska
272711@student.pwr.edu.pl

Maj 2024

Spis treści

1 Wstęp

NBA jest największą i najbardziej prestiżową ligą koszykówki na świecie, generującą wielomilionowe wyświetlenia i miliardowe zyski w skali roku. Wraz z ciągle rosnącą popularnością ligi, NBA decydowało się na gromadzenie, przechowywanie i udostępnianie coraz to nowszych danych. Obecnie ciężko jest wymyślić statystykę, której nie dałoby się bezpośrednio pozyskać lub obliczyć na podstawie innych, już opublikowanych.

Tak ogromna ilość danych stanowi idealną przestrzeń dla osób związanych z ich analizą. W ostatnich latach zespoły sportowe coraz częściej i poważniej stawiają na zaawansowane modele predykcyjne w celu zdobycia przewagi nad konkurencją. Również w branży bukmacherskiej korzysta się z technik analizy danych, modele predykcyjne są kluczowe w przewidywaniu wyników meczów, zapewnieniu uczciwości procesu obstawiania, a z perspektywy bukmachera - oczywiście zarobku.

Niniejsza praca ma na celu ukazanie, jak nowoczesne techniki i modele predykcyjne mogą być wykorzystywane w sporcie do podejmowania bardziej świadomych decyzji i opracowywania lepszych strategii zarówno w kontekście samych gier jak i zakładów bukmacherskich. Początkowo pozyskane będą dane, które intuicyjnie mogłyby mieć wpływ na wynik danej rozgrywki. Następnie, przez zastosowanie metod uczenia maszynowego, postaram się odpowiedzieć na pytanie - czy drużyna gospodarzy wygra mecz?

2 Dane

2.1 Dane - wstęp

Mając jasno zdefiniowaną istotę problemu, należy określić, które dane i w jaki sposób trzeba pozyskać. Naturalnym wyborem były *box score*'y, które w naszym przypadku są ustrukturyzowanym podsumowaniem danej rozgrywki. Są one dostępne na oficjalnej stronie [2] *nba.com* jak i na wielu innych fanowskich portalach. Ja w celu pozyskania danych zdecydowałem skorzystać z pythonowego modułu [1] *nba_api*. Początkowo pozyskane zostają wszystkie drużyny kiedykolwiek grające w NBA, a następnie wszystkie powiązane z nimi *box score*'y. W związku z tym, że rozwijająca się liga nieustannie decyduje się gromadzić nowe, bardziej zaawansowane dane, zdecydowałem się przefiltrować zbiór, pozostawiając jedynie statystyki odnoszące się do rozgrywek po roku 2003 w celu zachowania spójności danych.

Przefiltrowany zbiór danych obejmuje 61303 rekordy, gdzie każdy z nich zawiera następujące informacje:

- **SEASON_ID**: ID sezonu, w którym odbyła się dana rozgrywka
- **TEAM_ID**: ID drużyny, do której odnoszą się poniższe statystyki
- **GAME_ID**: ID danej rozgrywki.
- **WL**: Informacja czy drużyna wygrała czy przegrała mecz - W, L
- **PTS**: Punkty zdobyte przez drużynę.
- **FGM**: Trafione rzuty z gry (z wyjątkiem rzutów osobistych).
- **FGA**: Oddane rzuty z gry.

- **FG3M**: Trafione rzuty 3 punktowe.
- **FG3A**: Oddane rzuty 3 punktowe.
- **FTM**: Trafione rzuty osobiste.
- **FTA**: Oddane rzuty osobiste.
- **OREB**: Zbiórki w ofensywie.
- **DREB**: Zbiórki w defensywie.
- **AST**: Asysty.
- **STL**: Przechwyty.
- **BLK**: Bloki.
- **TOV**: Straty.
- **PF**: Faule.
- **PLUS_MINUS**: Rożnica punktów zdobytych przez daną drużynę i drużynę przeciwną.

2.2 Wstępne przetwarzanie danych

Obecny zbiór danych musi zostać jeszcze odpowiednio przearanżowany przed przystąpieniem do dalszych kroków. Przede wszystkim nie mamy informacji, która drużyna jest drużyną gospodarzy. Dodatkowo, każde **GAME_ID** jest zduplikowane, jedna rozgrywka jest przechowywana w postaci dwóch rekordów zawierających perspektywę danej drużyny. W celu rozwiązania tego problemu musimy ponownie odwołać się do api, aby dla danego **GAME_ID** uzyskać **TEAM_ID** drużyny gospodarzy. Następnie rekordy z identycznym **GAME_ID** są scalane. Statystyki gospodarzy są zapisywane z przedrostkiem *H*, a gości *A*.

Ze zbioru należało usunąć rekordy z wybrakowanymi danymi, których liczba w kontekście całego zbioru była nieznacząca. Końcowo, zbiór danych po przefiltrowaniu i scaleniu zduplikowanych rekordów obejmuje 30809 rozgrywek.

Kolejnym problemem, jest fakt, iż każdy rekord przechowuje dane dotyczące wyniku rozgrywki. Informacje te w momencie predykcji rezultatu danego meczu oczywiście nie są jeszcze dostępne. Zaproponowanym rozwiązaniem jest zastąpienie danych podsumowujących rozgrywkę, danymi średnimi z 10 ostatnich meczy rozegranych przez konkretną drużynę.

Następnym krokiem przygotowującym dane jest wprowadzenie rankingu ELO [3] [4]. Ranking ten jest stosowany w celu określenia mocy danej drużyny. Każda drużyna otrzymuje 1500 punktów rankingowych przy pierwszym jej wystąpieniu w zbiorze danych. Następnie po każdej rozgrywce rankingi obu zespołów są aktualizowane, zgodnie z poniższymi wzorami:

Prawdopodobieństwo zwycięstwa

$$E_p = \frac{1}{1 + 10^{\frac{\text{oelo}-\text{telo}}{400}}} \tag{1}$$

gdzie:

E_p - Prawdopodobieństwo zwycięstwa drużyny.

telo - Obecny ranking ELO drużyny.

oelo - Obecny ranking ELO przeciwników.

Współczynnik k

$$k = 20 \times \frac{(MOV_{\text{win}} + 3)^{0.8}}{7.5 + 0.0006 \times \text{elo\_diff}} \tag{2}$$

gdzie:

k - Współczynnik k.

MOV_{win} - Margines zwycięstw, obliczany jako średnia różnica punktów zdobytych przez daną drużynę, a przeciwników w ostatnich 10 rozgrywkach.

elo_diff - Różnica rankingów ELO drużyn.

Aktualizacja rankingu

$$E_{i+1} = k \times (W - E_p + E_i)$$

gdzie:

E_{i+1} - Zaktualizowany ranking ELO.

k - Współczynnik k. (równanie 2).

W - Wynik meczu (1 dla zwycięstwa, 0 dla porażki).

E_p - Prawdopodobieństwo zwycięstwa. (równanie 1)

E_i - Obecny ranking ELO.

Nowy sezon

$$R_n = (R_o \times 0.75) + (0.25 * 1505)$$

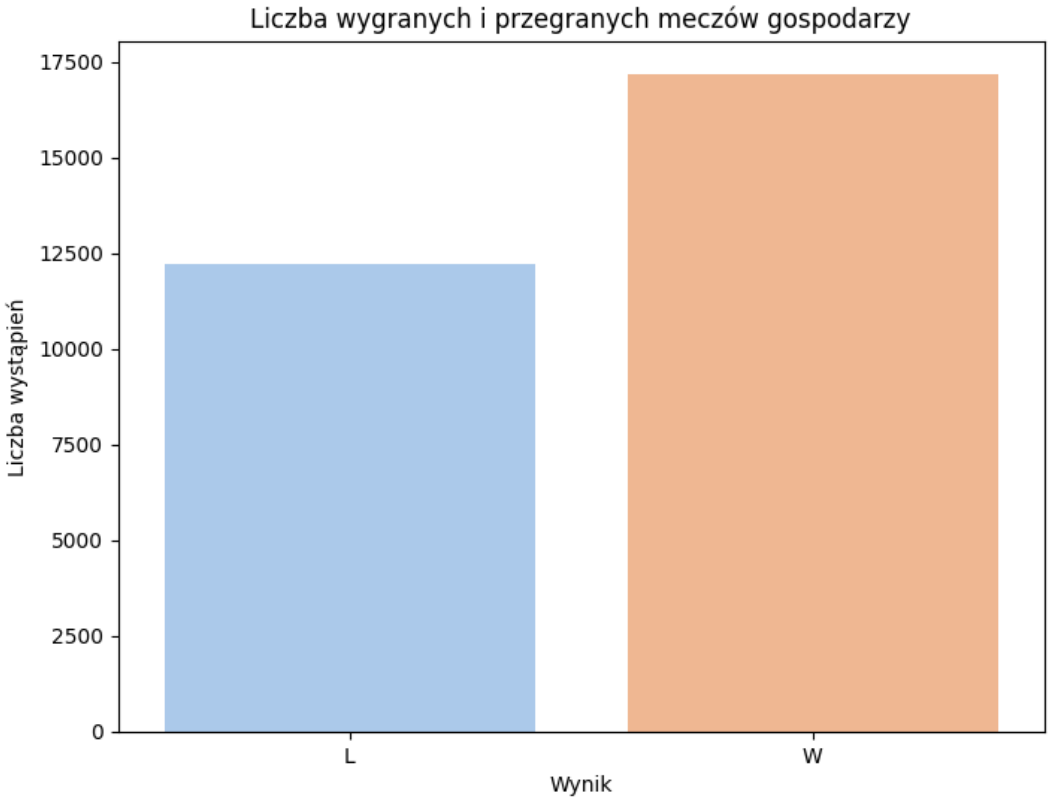
gdzie:

R_n - Ranking, z którym drużyna rozpoczyna następny sezon.

R_o - Ranking, z którym drużyna skończyła poprzedni sezon.

3 Wstępna analiza danych (Exploatory Data Analysis)

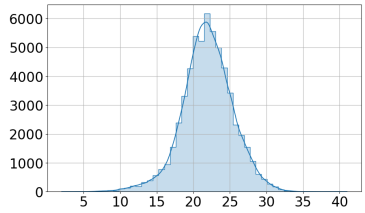
3.1 Wygrane i przegrane drużyny gospodarzy



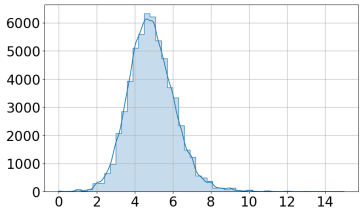
Rysunek 1: Porównanie liczby wygranych i przegranych rozgrywek przez drużyne gospodarzy

Analiza danych wykazała, że drużyny gospodarzy przegrywają około 42% meczów, co sugeruje, że zbiór danych jest względnie zrównoważony, biorąc pod uwagę stosunek liczności klas 42 do 58. Dodatkowo, ten wynik jest zgodny z intuicyjnym przekonaniem, że gra na własnym terenie może wpływać korzystnie na wynik spotkania. Jest to ważna obserwacja, ponieważ równowaga w danych jest istotna dla skuteczności modeli predykcyjnych oraz analizy statystycznej.

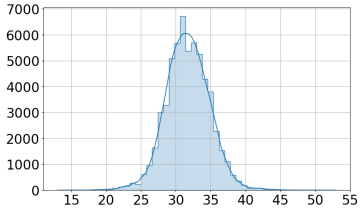
3.2 Dystrybucje poszczególnych cech



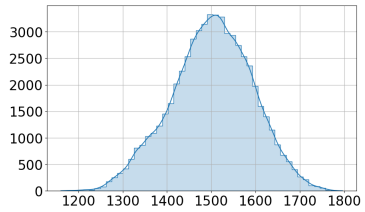
(a) AST Distribution



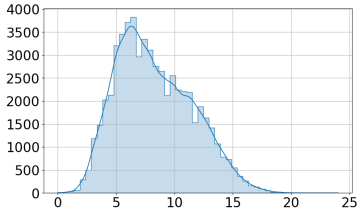
(b) BLK Distribution



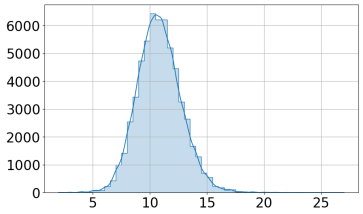
(c) DREB Distribution



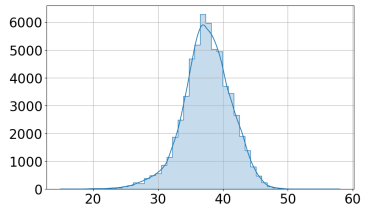
(d) ELO Distribution



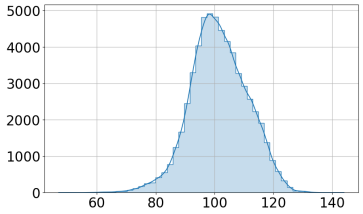
(e) FG3M Distribution



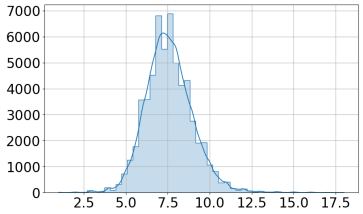
(f) OREB Distribution



(g) FGM Distribution



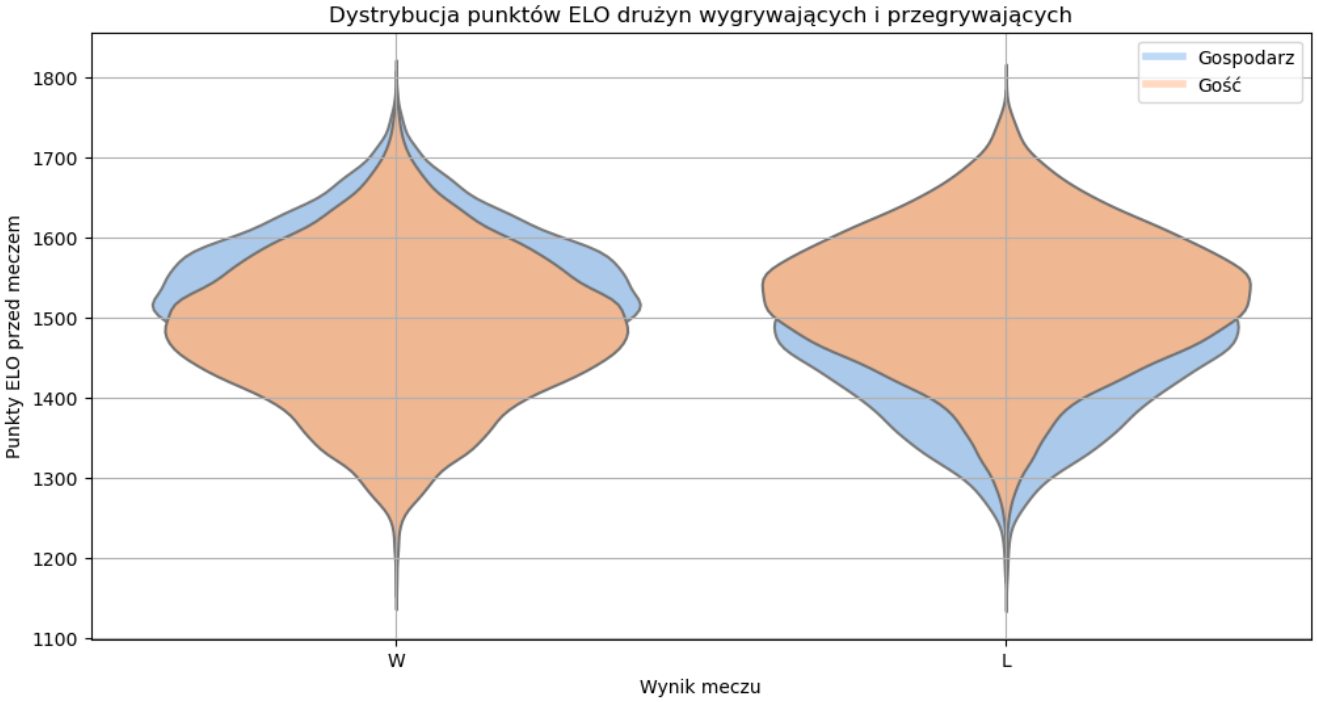
(h) PTS Distribution



(i) STL Distribution

Dane prezentują kształt histogramów zbliżony do rozkładu normalnego, co stanowi pozytywną informację. Zróżnicowanie średnich i odchyleń standardowych pomiędzy nimi jest naturalne. Istnienie rozkładu normalnego cech sprzyja efektywnej i stabilnej budowie modeli, co ma kluczowe znaczenie w analizie danych oraz predykcji z ich wykorzystaniem.

3.3 Dystrybucja poszczególnych cech uwzględniająca drużyny wygrywające i przegrywające

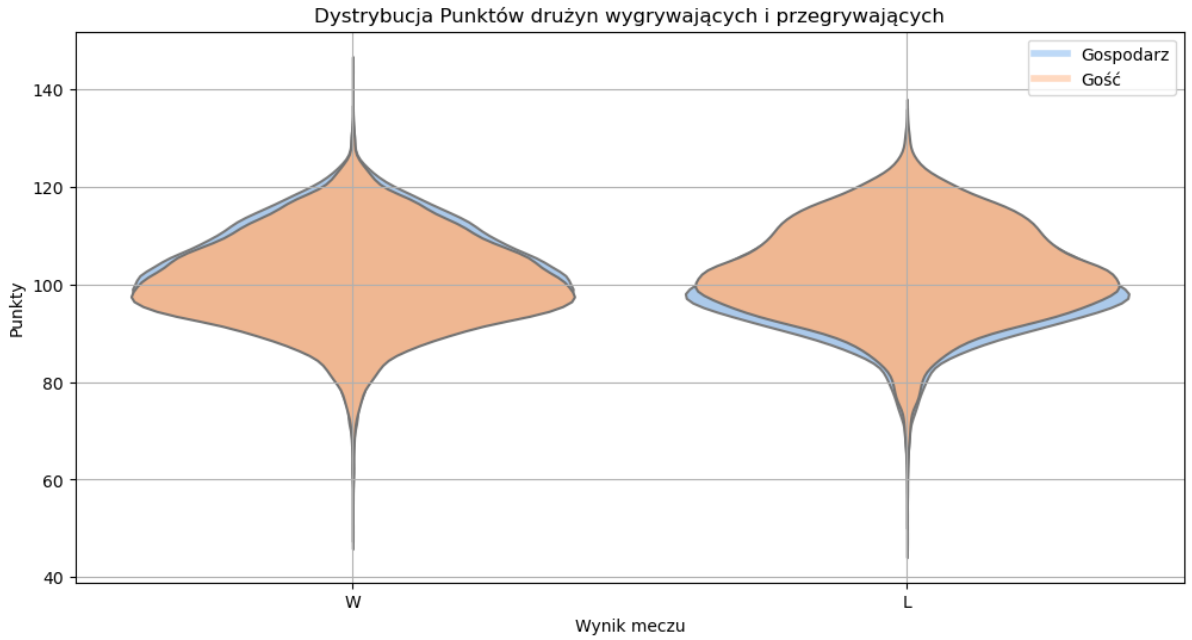


Rysunek 3: Porównanie dystrybucji punktów ELO z rozróżnieniem drużyn wygranych/przegranych i gospodarzy/gości

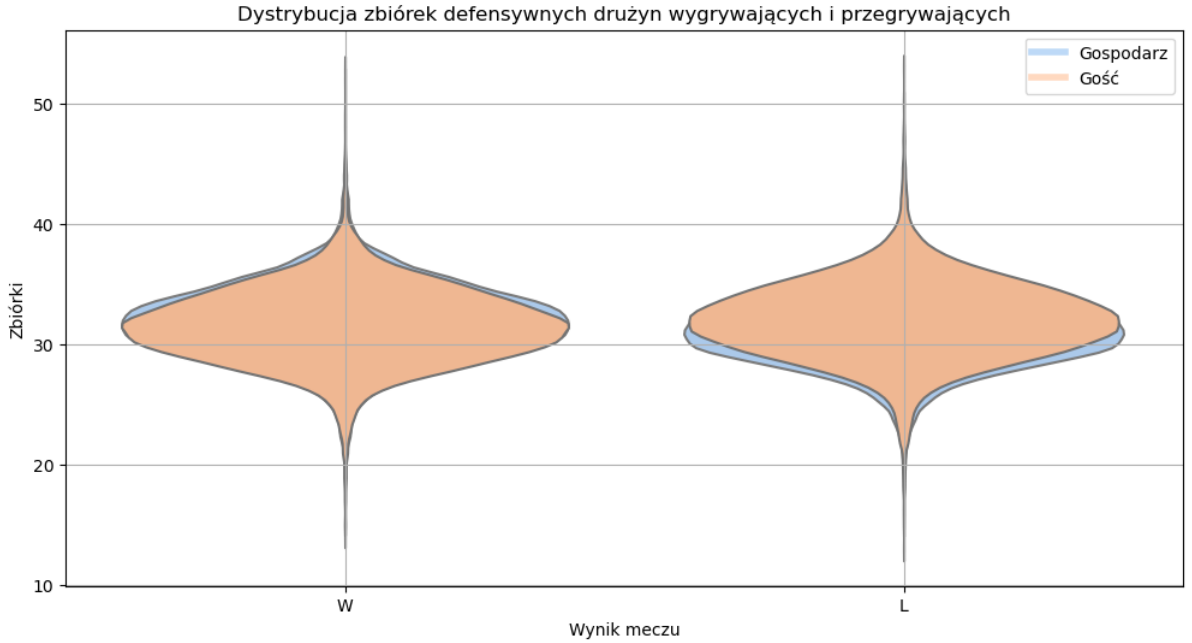
W przypadku wygranej drużyny gospodarzy, zauważalna jest tendencja, że średnia liczba punktów zdobywanych przez drużynę gości jest niższa niż średnia liczba punktów zdobywanych przez drużynę gospodarzy. Analogiczna zależność jest widoczna w przypadku wygranej drużyny gości, gdzie średnia liczba punktów zdobywanych przez gospodarzy jest niższa niż liczba punktów zdobywanych przez drużynę gości.

Podobne, choć mniej wyraźne zależności można zaobserwować w analizie zdobytych punktów oraz zbiórek defensywnych. W sytuacjach, gdy drużyna gospodarzy wygrywa, ich średnia liczba zdobytych punktów i zbiórek defensywnych jest wyższa w porównaniu do drużyny gości. Analogicznie, gdy drużyna gości odnosi zwycięstwo, ich średnie wartości w tych statystykach przewyższają wartości drużyny gospodarzy.

Te obserwacje sugerują, że kluczowe elementy takie jak zdobywanie punktów i skuteczność w defensywie, potencjalnie mogą mieć wpływ na rezultat spotkania, co jest zgodne z oczekiwaniami w kontekście analizy sportowej.



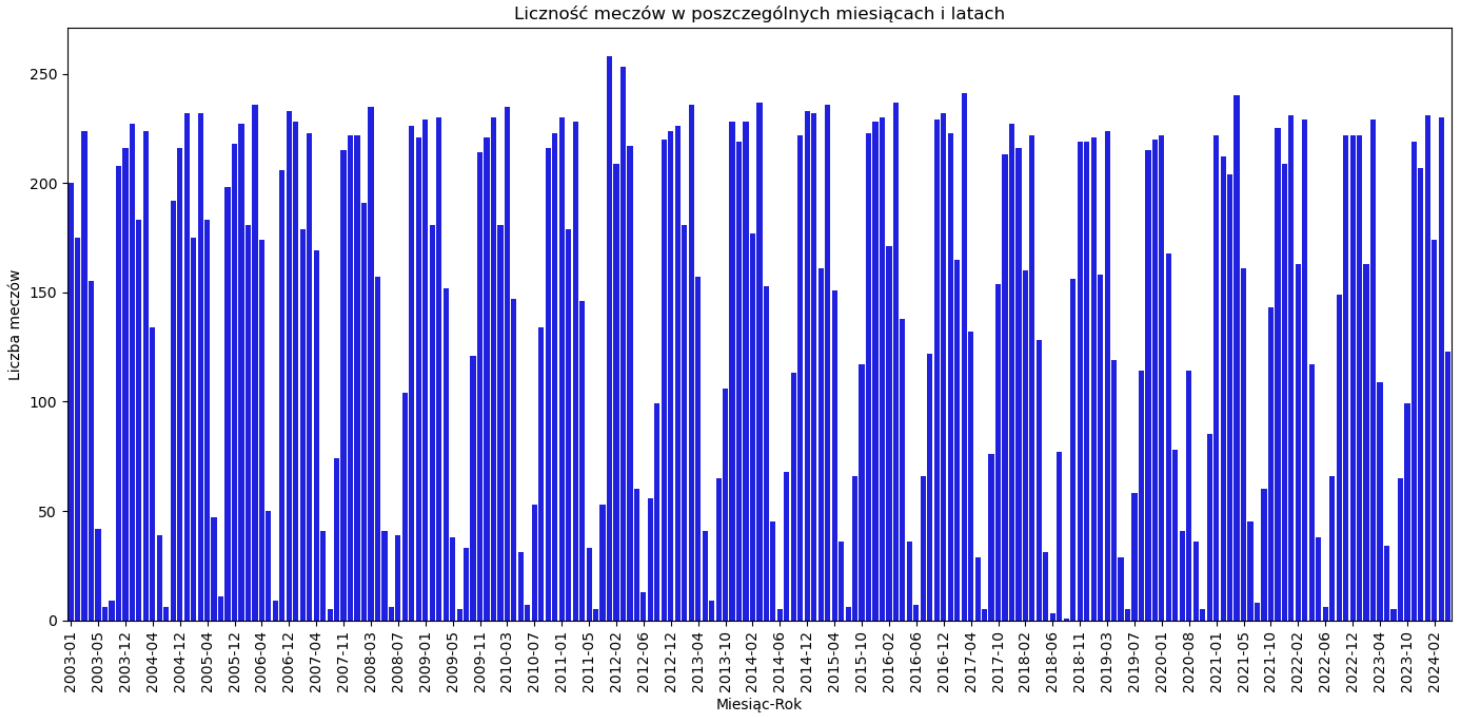
Rysunek 4: Porównanie dystrybucji zdobytych punktów z rozróżnieniem drużyn wygranych/przegranych i gospodarzy/gości



Rysunek 5: Porównanie dystrybucji zbiórek defensywnych z rozróżnieniem drużyn wygranych/przegranych i gospodarzy/gości

Obserwowane różnice w zdobytych punktach i zbiórkach defensywnych między wygrywającymi i przegrywającymi drużynami nie są na tyle znaczące, aby jednoznacznie stwierdzić, że obrona odgrywa decydującą rolę w zwycięstwach. Można więc poddać w wątpliwość stare koszykarskie powiedzenie *Atak sprzedaje bilety, obrona wygrywa mecze*.

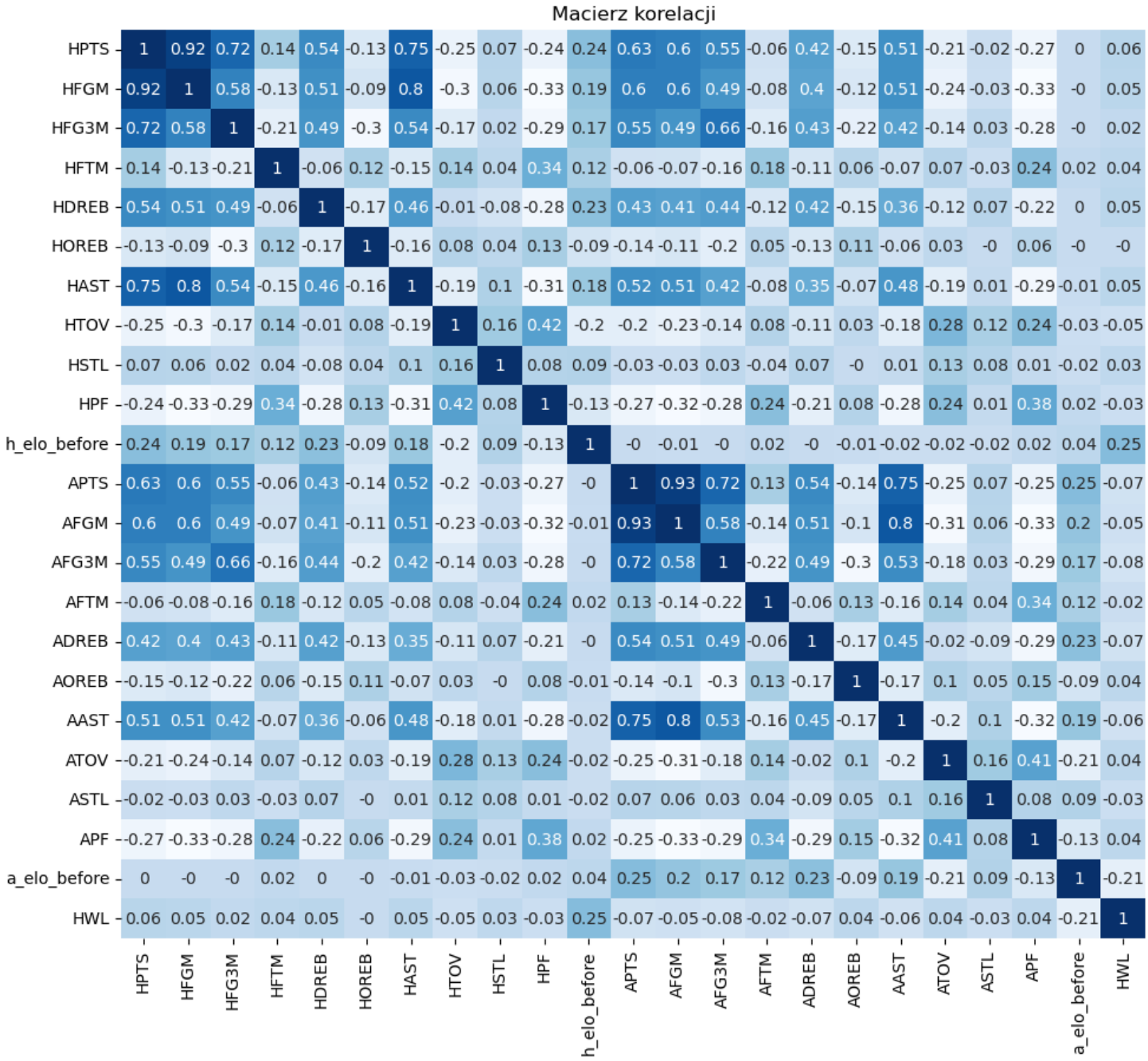
3.4 Dystrybucja meczów w zbiorze danych w zależności od danych



Rysunek 6: Wizualizacja liczności meczów w poszczególnych miesiącach i latach

W zbiorze danych zawarte są rozgrywki od roku 2003. Liczność rozgrywek w zależności od daty jest zgodna z oczekiwaniami - zauważalne są sezony, w których na początku i końcu każdego z nich występują spadki liczby rozgrywek. Największe nasilenie liczby meczów ma miejsce w trakcie sezonu, co jest typowe dla harmonogramów ligowych, gdzie najwięcej spotkań odbywa się w środkowej części sezonu regularnego.

4 Macierz korelacji



Rysunek 7: Macierz korelacji cech

Macierz korelacji pozwala nam wyciągnąć kolejne wnioski dotyczące opracowywanego zbioru danych, umożliwiając lepsze jego zrozumienie. Duże zależności występują pomiędzy asystami a liczbą zdobytych punktów i oddanych rzutów. Co ciekawe, liczba zbiórek defensywnych jest również stosunkowo mocno powiązana z tymi cechami. Interesująca wydaje się być również zależność pomiędzy stratami a faulami. Skupiając się jednak na interesującej nas cesze, którą w dalszej części będziemy chcieli predykować, możemy dostrzec, że nie jest ona zbytnio skorelowana z żadną z pozostałych

cech. Najsilniejsza korelacja, zgodnie z oczekiwaniami, występuje z punktami systemu rankingowego. Pozytywne cechy drużyny przyjezdnej (oprócz strat i fauli) są związane niską, ujemną korelacją z wynikiem rozgrywki. Należy jednak pamiętać, że ciężko utożsamić siłę drużyny z wyłącznie jedną statystyką. Można w końcu wygrywać mecze, rzucając niską liczbę punktów, robiąc wiele strat czy nie zbierając żadnej piłki. Zdecydowanie cięższym zadaniem byłaby wygrana, jeżeli każda z tych cech by zawiodła. Dlatego w celu predykcji wyników, będziemy korzystać z modeli predykcyjnych, które uwzględniają wszystkie zmienne jednocześnie. Taki holistyczny model pozwoli na bardziej dokładne prognozy, biorąc pod uwagę złożoność i wzajemne powiązania pomiędzy różnymi cechami drużyny.

5 Modele

Przed przystąpieniem do treningu modeli konieczne jest odpowiednie przygotowanie danych. Pierwszym krokiem było podzielenie zbioru danych na zmienne zależne, czyli wyniki, które chcemy przewidzieć, oraz zmienne niezależne, czyli cechy, na podstawie których będziemy dokonywać predykcji. Następnie dane zostały podzielone na dwie części: zbiór treningowy, służący do uczenia modelu, oraz zbiór testowy, służący do oceny jego skuteczności. Stosunek podziału wynosił 80:20.

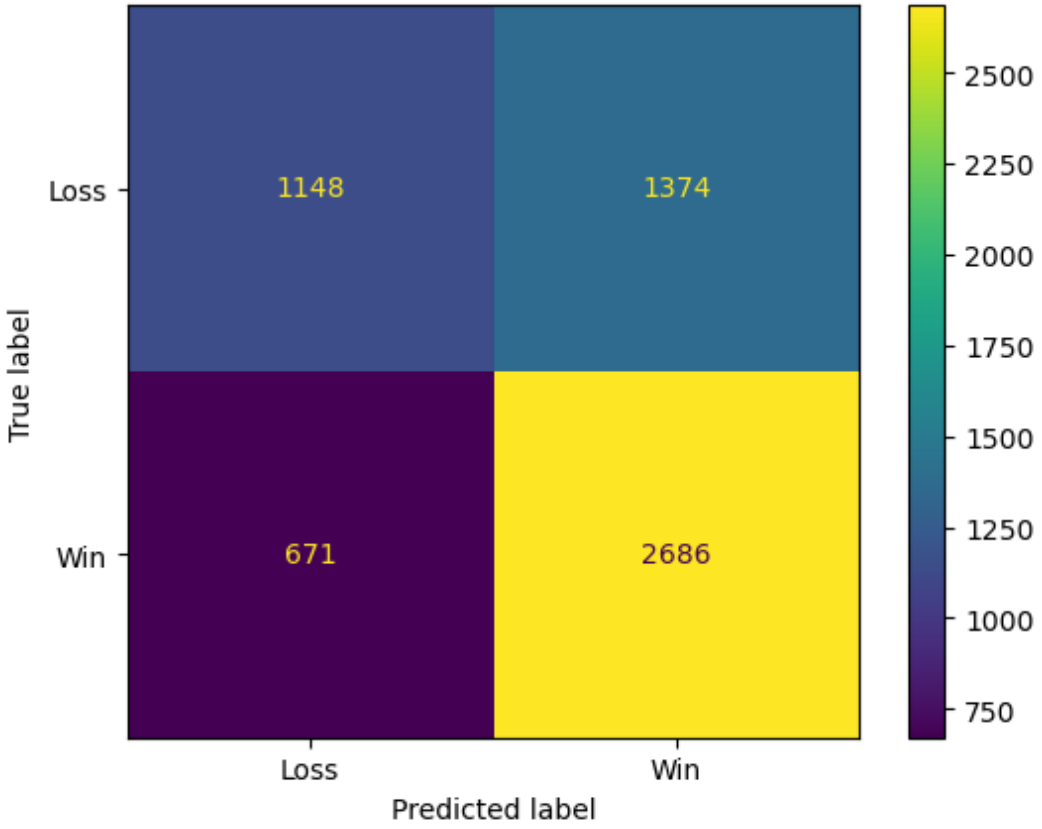
5.1 Model regresji logistycznej

W celu porównania skuteczności predykcji modelu, przeprowadziłem dwa eksperymenty. Pierwszy z nich wykorzystał domyślne wartości hiperparametrów, podczas gdy drugi opierał się na metodzie optymalizacji zwaną *Grid Search*. Polega ona na przetestowaniu wielu różnych kombinacji hiperparametrów w celu znalezienia najlepszej konfiguracji modelu. Porównując wyniki obu eksperymentów, mogłem ocenić, czy dostosowanie hiperparametrów za pomocą *Grid Search* przyniosło poprawę w skuteczności predykcji.

Wyniki pierwszego eksperymentu modelu regresji logistycznej, z domyślnymi parametrami: *C = 1.0, penalty = l2, solver = lbfgs*

Class	**Precision**	**Recall**	**F1-score**	**Support**
Loss	0.63	0.46	0.53	2459
Win	0.67	0.80	0.73	3420
Accuracy	0.65			
Macro avg	0.65	0.63	0.63	5879
Weighted avg	0.65	0.65	0.65	5879

Tabela 1: Metrics for Logistic Regression Model

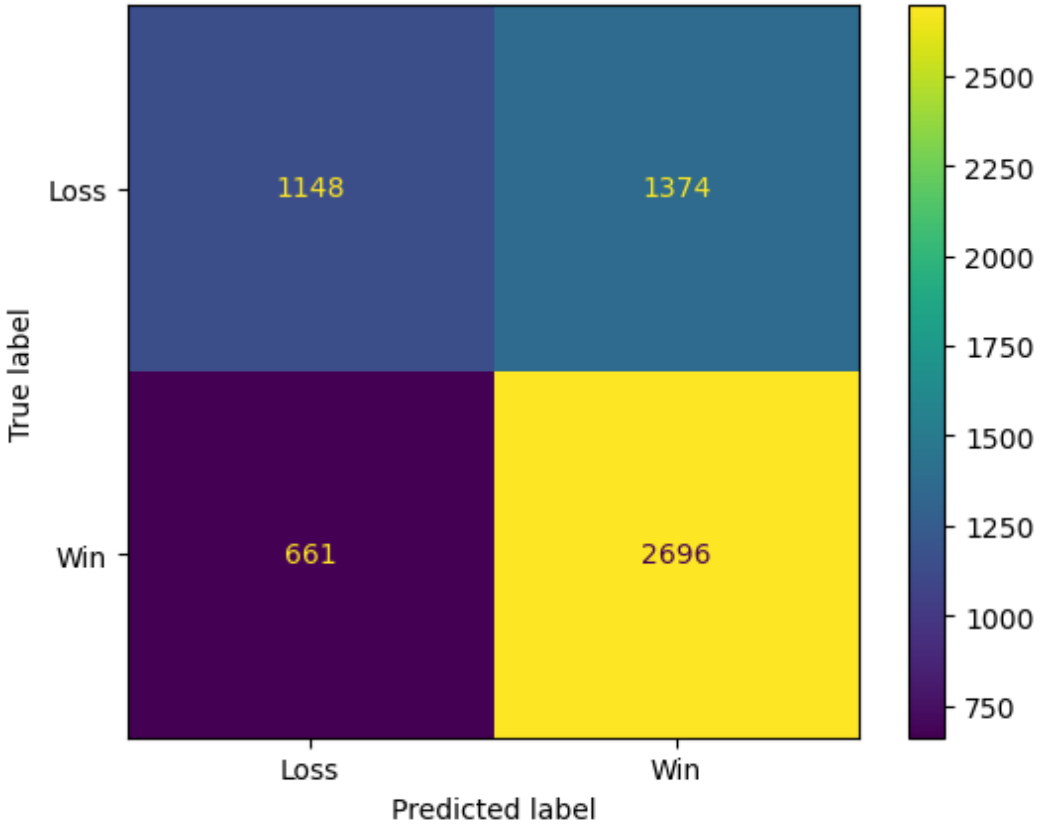


Rysunek 8: Macierz pomyłek pierwszego eksperymentu

wejściowa siatka parametrów do *Grid Search* wyglądała nastepująco:
[*C = [0.001, 0.01, 0.1, 1, 10, 100, 1000], penalty = [l1, l2, None], solver = [newton-cg, lbfgs, liblinear]*]
Po zastosowaniu optymalizacji wybrane zostały następujące hiperparametry: *C = 0.1, penalty = l2, solver = lbfgs*, dające następujące wyniki:

Class	**Precision**	**Recall**	**F1-score**	**Support**
Loss	0.63	0.46	0.53	2522
Win	0.66	0.80	0.73	3357
Accuracy	0.65			
Macro avg	0.65	0.63	0.63	5879
Weighted avg	0.65	0.65	0.65	5879

Tabela 2: Metrics for Logistic Regression Model



Rysunek 9: Macierz pomyłek drugiego eksperymentu

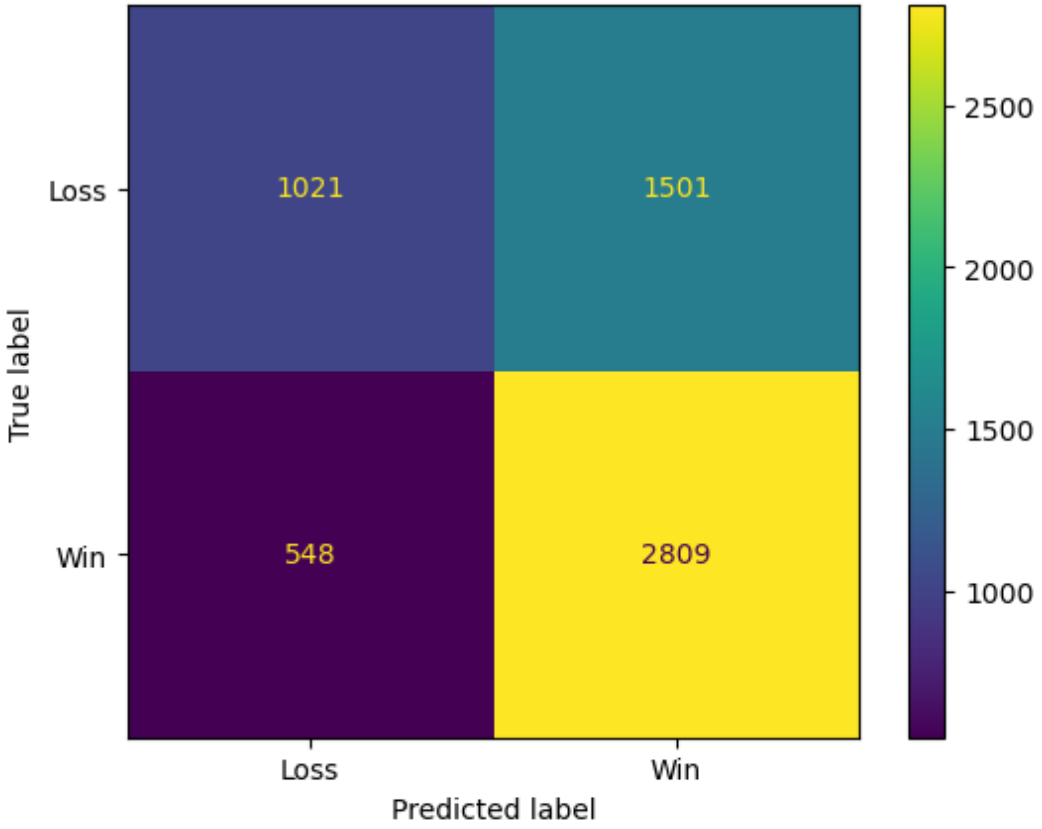
Niestety mimo przeszukiwania parametrów, proces optymalizacji nie przyniósł oczekiwanych rezultatów. Model nie uległ poprawie względem modelu z eksperymentu pierwszego. Mimo to, oba modele predykują wynik rozgrywki ze stosunkowo duża dokładnością 65%.

5.2 Klasyfikator maszyny wektorów nośnych

W dalszej części analizy predykcyjnej z użyciem klasyfikatora Maszyny Wektorów Nośnych (SVC). Podobnie jak powyżej przeprowadziłem 2 eksperymenty

Class	**Precision**	**Recall**	**F1-score**	**Support**
Loss	0.65	0.40	0.50	2522
Win	0.65	0.84	0.73	3357
Accuracy	0.65			
Macro avg	0.65	0.62	0.62	5879
Weighted avg	0.65	0.65	0.63	5879

Tabela 3: Metrics for Logistic Regression Model

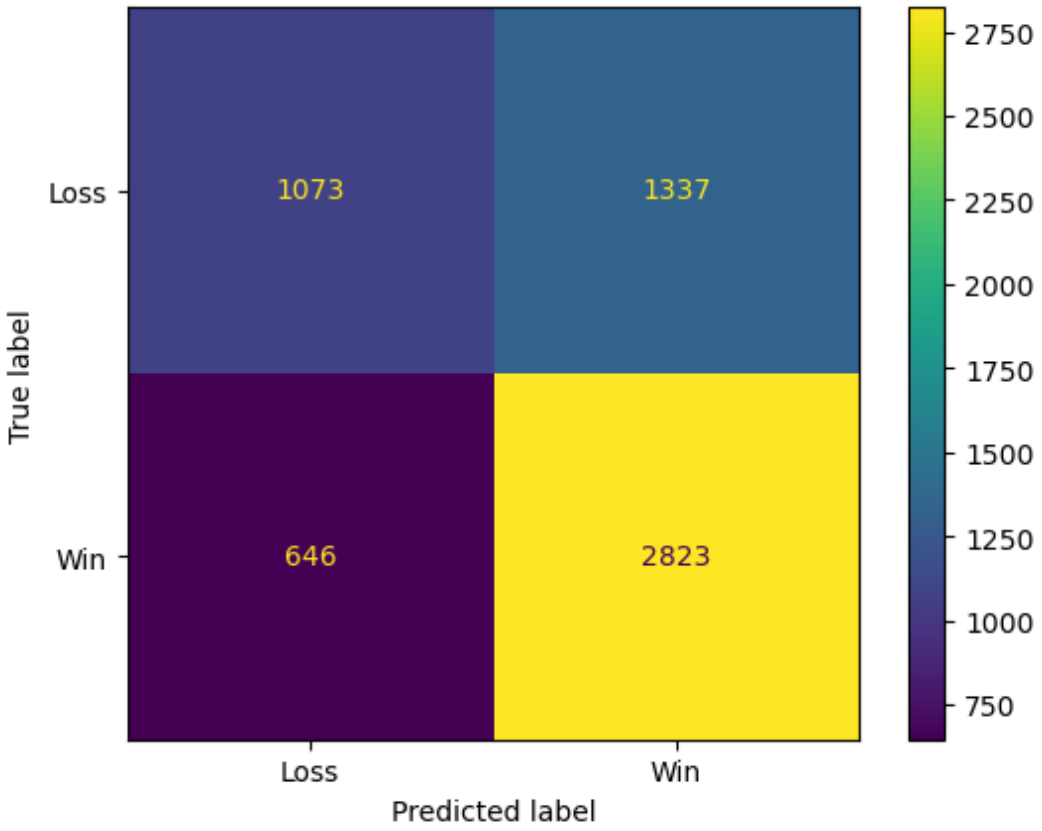


Rysunek 10: Macierz pomyłek pierwszego eksperymentu

Do optymalizacji hiperparametrów została zastosowana następująca siatka parametrów: *C = [1, 10, 100, 1000], gamma = [1, 0.1, 0.001, 0.0001], kernel = [linear, rbf]*.
Z, których najbardziej optymalną kombinacją okazało się być *C = 100, gamma = 1, kernel = linear*. Model odniósł następujący rezultat:

Class	**Precision**	**Recall**	**F1-score**	**Support**
Loss	0.61	0.45	0.52	2394
Win	0.68	0.80	0.74	3485
Accuracy	0.66			
Macro avg	0.65	0.63	0.63	5879
Weighted avg	0.65	0.66	0.63	5879

Tabela 4: Metrics for Logistic Regression Model



Rysunek 11: Macierz pomyłek drugiego eksperymentu

Optymalizacja hiperparametrów ponownie nie przyniosła większych skutków, model zwiększył swoją dokładność jedynie o 1 punkt procentowy względem modelu z pierwszego eksperymentu.

6 Wnioski

Na wynik rozgrywek sportowych wpływa ma wiele czynników zupełnie losowych i nieprzewidywalnych, takich jak dyspozycja dnia, kontuzje czy nawet czyste szczęście. W niniejszej pracy nie uwzględniono tych czynników, co może wpłynąć na precyzyjność przewidywań. Mimo wszystko, przy problemie o takiej złożoności, osiągnięcie skuteczności powyżej 50% w klasyfikacji binarnej można uznać za satysfakcjonujący wynik.

Najlepszy model, czyli klasyfikator maszyny wektorów nośnych o parametrach $C = 100$, *gamma* $= 1$, *kernel* $=$ *linear*, osiągnął skuteczność na poziomie 66%. Niestety, trudno uznać ten model za wiarygodny. Zaproponowane rozwiązanie, oparte jedynie na podstawowych danych statystycznych, mogłoby być jednym z elementów w procesie decyzyjnym dotyczącym przewidywania wygranego, ale niekoniecznie ostatecznym predyktorem.

Porównując modele klasyfikacji logistycznej i maszyny wektorów nośnych, zauważamy, że ten drugi osiąga nieznacznie lepsze wyniki. Choć lepiej przewiduje wygrane drużyny gospodarzy, to jego precyzja w przewidywaniu porażek wynosi jedynie 61%.

Warto zaznaczyć, że te modele opierały się wyłącznie na podstawowych statystykach z danych dotyczących rozgrywek, co może ograniczać ich zdolność do precyzyjnej predykcji. Istnieje potrzeba dalszego ulepszania tych modeli poprzez uwzględnienie bardziej zaawansowanych statystyk oraz czynników wpływających na wynik meczów. Najlepsze modele predykcyjne dla rozgrywek sportowych wykorzystują bardziej zaawansowane metryki i zupełnie inne podejścia. Jednym z takich podejść jest model, który całkowicie pomija zwycięstwa i porażki, a opiera się na prognozowaniu przyszłych występów każdego gracza. Tworząc taki model, można dodatkowo uwzględnić kontuzje, transfery i inne transakcje związane z graczami [5]. Wówczas siła drużyny byłaby określana na podstawie aktualnego składu wyjściowego oraz dyspozycji indywidualnych graczy, co w zaproponowanym modelu zostało pominięte. Niemniej jednak, nawet te zaawansowane modele predykcyjne operują na dokładności rzędu około 75%, w związku z czym dokładność zaproponowanego modelu można uznać za zadowalającą.

Literatura

[1] swar/nba_api. GitHub. Retrieved from `https://github.com/swar/nba_api`

[2] swar/nba_api. GitHub. Retrieved from `https://github.com/swar/nba_api`

[3] The Elo Rating System. Retrieved from `https://en.wikipedia.org/wiki/Elo_rating_system`

[4] Probabilistic Match Importance in Professional Sports. Retrieved from `https://core.ac.uk/download/pdf/154900868.pdf`

[5] FiveThirtyEight Model. Retrieved from `https://fivethirtyeight.com/methodology/how-our-nba-predictions-work/`