Analiza danych szachowych: predykcja rankingu Elo szachistów przy użyciu modelu SVR

Sara Studzińska

Czerwiec 2023

Spis treści

1 Problem badawczy

Celem niniejszego raportu jest zbadanie możliwości przewidywania rankingu Elo szachistów na podstawie dostępnych danych, takich jak wiek, tytuł szachowy, liczba wygranych gier oraz liczba zremisowanych partii.

Problem badawczy, który stawiam sobie w tym raporcie, można zdefiniować następująco: Czy istnieje wystarczający związek między informacjami na temat szachistów, a ich rankingiem Elo? Czy na podstawie tych danych można stworzyć model, który będzie w stanie przewidywać ranking Elo szachistów z odpowiednią dokładnością?

W celu odpowiedzi na powyższe pytania przeprowadzę wstępną analizę dostępnych danych oraz stworzę model predykcyjny. Opracowane rozwiązanie pozwoli ocenić, czy istnieje możliwość przewidywania rankingu Elo na podstawie podanych zmiennych oraz jakie czynniki mają największy wpływ na ten ranking.

W dalszej części raportu przedstawię szczegółowy opis użytych danych, metodyki przeprowadzonej analizy oraz wyniki eksperymentów. Ten raport ma na celu nie tylko osiągnięcie dobrych wyników predykcyjnych, ale także zrozumienie wpływu poszczególnych zmiennych na ranking Elo oraz ocenę jakości opracowanego modelu.

2 Zebranie danych

Do pozyskania danych o szachistach skorzystałam z oficjalnej strony Federacji Szachowej (FIDE). W tym celu zastosowałam technikę web scrapingu, która umożliwiła mi pobranie informacji dotyczących poszczególnych szachistów bezpośrednio ze strony.

W celu skupienia się na szachistach utytułowanych i aktualnie aktywnych, opracowałam skrypt, który przeglądał stronę FIDE i wyodrębniał informacje tylko o tych szachistach, którzy posiadali tytuły szachowe oraz byli nadal aktywni w rozgrywkach. Dzięki temu ograniczeniu, moje badanie skupiło się na grupie szachistów, którzy osiągają znaczące wyniki i mają aktualne osiągnięcia w dziedzinie szachów.

Aby pozyskać dodatkowe informacje o szachistach, takie jak liczba rozegranych partii, skorzystałam z techniki wysyłania zapytań POST do indywidualnych profili szachistów na stronie FIDE. Poprzez te zapytania, byłam w stanie uzyskać szczegółowe dane dotyczące każdego szachisty, które miały znaczenie dla analizy i modelowania problemu badawczego.

Dzięki przeprowadzeniu tych operacji na stronie FIDE, udało mi się zebrać dane około 11 tysięcy szachistów. Otrzymane informacje obejmują imię i nazwisko szachisty, rok urodzenia, federację, płeć, tytuł szachowy, tytuł trenerski, liczbę wygranych partii, liczbę zremisowanych partii oraz aktualny ranking elo w szachach klasycznych.

3 Przygotowanie danych

W celu przygotowania danych do analizy, przeprowadziłam proces wyczyszczenia danych oraz wykonania niezbędnych operacji preprocessingu. Działania te obejmowały:

1. Usunięcie danych dotyczących tytułu trenerskiego - Liczba szachistów posiadających tytuł trenerski była niewielka, co powodowało, że nie był on reprezentatywny dla analizy.
2. Obliczenie aktualnego wieku - Na podstawie roku urodzenia szachistów, obliczyłam ich aktualny wiek w latach. Ta informacja może mieć znaczenie dla analizy, ponieważ wiek może wpływać na wyniki i ranking Elo szachistów.
3. Wybór szachistów z co najmniej 200 rozegranymi partiam - W celu zachowania odpowiedniej jakości danych, zdecydowałam się uwzględnić tylko szachistów, którzy mieli rozegrane co najmniej 200 partii. To kryterium pozwoliło na ograniczenie analizy do bardziej doświadczonych szachistów, których wyniki są bardziej reprezentatywne.
4. Obliczenie wskaźnika wygranych białymi/czarnymi - Na podstawie danych dotyczących całkowitej liczby partii, liczby wygranych i liczby remisów, osobno dla partii granych białymi i czarnymi, obliczyłam wskaźniki wygranych białymi oraz czarnymi. Te wskaźniki mogą być istotne, ponieważ wskazują na umiejętność szachisty w grze zarówno białymi, jak i czarnymi pionkami.
5. Obliczenie wskaźnika remisów białymi/czarnymi: Podobnie jak w przypadku wskaźnika wygranych, obliczyłam również wskaźniki remisów białymi/czarnymi na podstawie danych dotyczących liczby remisów i całkowitej liczby partii, rozdzielonych na partie białymi i czarnymi. Ta informacja może być przydatna, aby zrozumieć styl gry i strategie szachistów w różnych sytuacjach.

4 Wstępna analiza danych

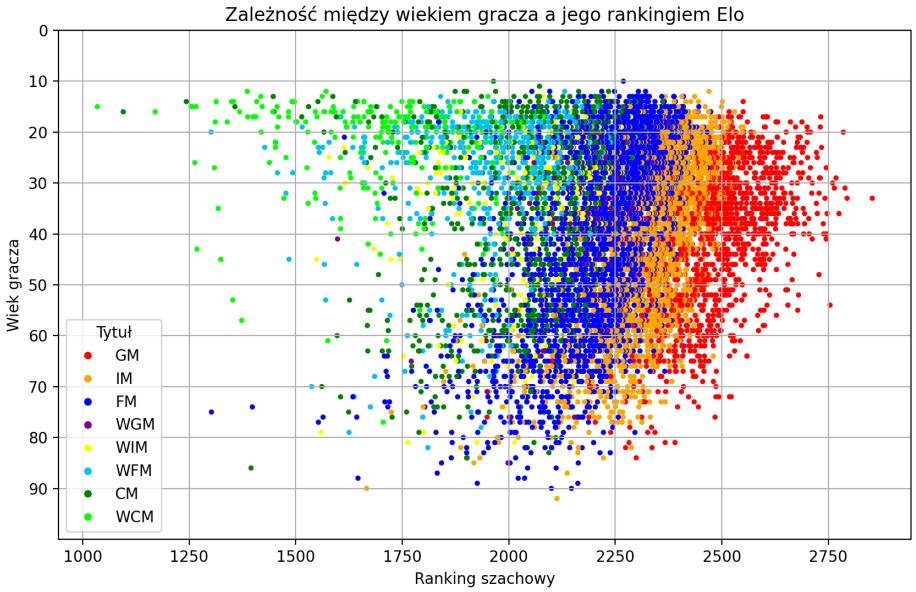
Na podstawie zebranych i przetworzonych wcześniej danych utworzyłam wykresy, które przedstawiają związki między różnymi cechami a rankingiem Elo.

4.1 Wiek a ranking Elo

Młodsi szachiści poniżej 20 roku życia często wykazują znaczące zdolności i potencjał w dziedzinie szachów. Jednak ich ranking Elo może być niższy w porównaniu do starszych graczy ze względu na ich względnie krótkie doświadczenie i ciągły rozwój umiejętności. Wielu młodych szachistów jeszcze się uczy i zdobywa doświadczenie w rozgrywkach na wyższym poziomie, dlatego ranking Elo może nie odzwierciedlać jeszcze ich pełnego potencjału.

Szachiści w średnim wieku, zazwyczaj między 20 a 40 rokiem życia, często osiągają wyższe rankingi Elo. Wynika to z ich większego doświadczenia, które gromadzą w trakcie wielu lat gry i udziału w turniejach. W tym okresie życia szachiści mają już solidne fundamenty szachowe i zdolności taktyczne, co przekłada się na osiąganie lepszych wyników.

Najstarsi szachiści, którzy przekroczyli 40 rok życia, mają bogate doświadczenie i wiedzę szachową. Ich ranking Elo może być stosunkowo stabilny, a rozwój w tej dziedzinie może być już mniej dynamiczny. Mimo że doświadczenie jest ważnym czynnikiem wpływającym na ranking, starsi szachiści mogą mieć ograniczenia fizyczne, które wpływają na ich wyniki i utrzymanie wyższego rankingu.



Rysunek 1: Wykres przedstawiający zależnośc między wiekiem gracza a jego rankingiem Elo, z uwzględnieniem tytułu gracza

Wnioski z tej analizy sugerują, że wiek, tytuł szachowy i ranking Elo są ze sobą powiązane. Wiek szachistów ma znaczący wpływ na ranking Elo. Młodsi gracze mogą mieć niższy ranking ze względu na mniej doświadczenia, ale mają potencjał do osiągania znaczących wyników. Szachiści w średnim wieku mają tendencję do osiągania wyższych rankingów dzięki swojemu doświadczeniu, podczas gdy starsi szachiści, choć posiadają duże doświadczenie, mogą utrzymywać się na pewnym poziomie rankingowym, a nawet z niego spaść.

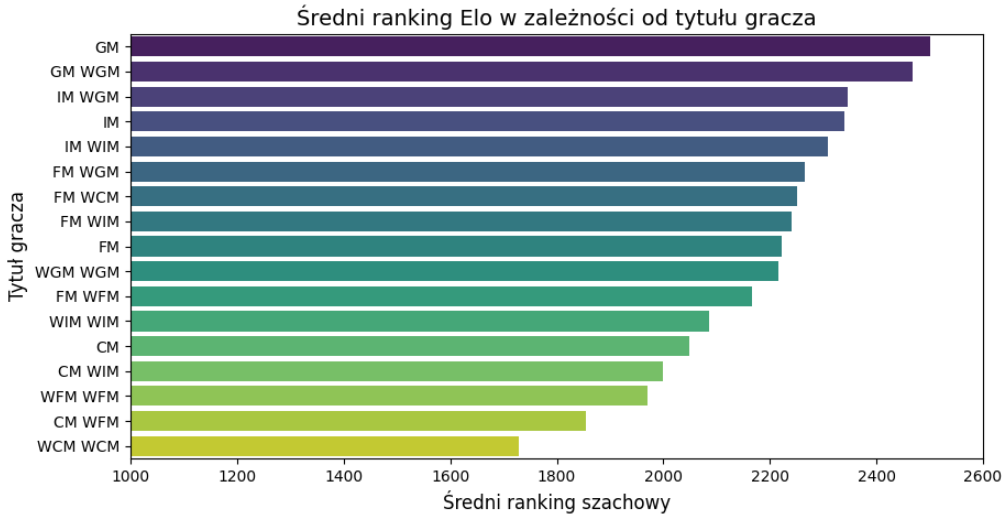
Z analizy wynika również, że posiadanie tytułu szachowego jest związane z określonymi przedziałami w rankingu Elo. Szachiści posiadający tytuł arcymi-

strza (GM) często osiągają wyższe rankingi Elo niż szachiści z tytułem mistrza międzynarodowego (IM). Jest to zgodne z oczekiwaniami, ponieważ tytuł arcymistrza jest najwyższym tytułem szachowym i wymaga osiągnięcia bardzo wysokiego poziomu umiejętności.

Analiza ta potwierdza, że tytuły szachowe są wskaźnikiem wysokiego poziomu umiejętności i wyników szachistów. Szachiści z wyższymi tytułami mają tendencję do osiągania wyższych rankingów Elo.

4.2 Tytuł gracza a średni ranking Elo

W analizie związku między tytułem szachowym a średnim rankingiem Elo gracza uwzględniłam zarówno tytuły uniwersalne (GM, IM, FM, CM) jak i tytuły kobiece. W przypadku kobiet, istnieje możliwość posiadania dwóch tytułów - tytułu uniwersalnego oraz tytułu kobiecego. Ze względu na to, że tytuły mogą informować o płci, dalsza analiza danych oraz ostateczny model nie będą uwzględniać płci, lecz jedynie tytuły.



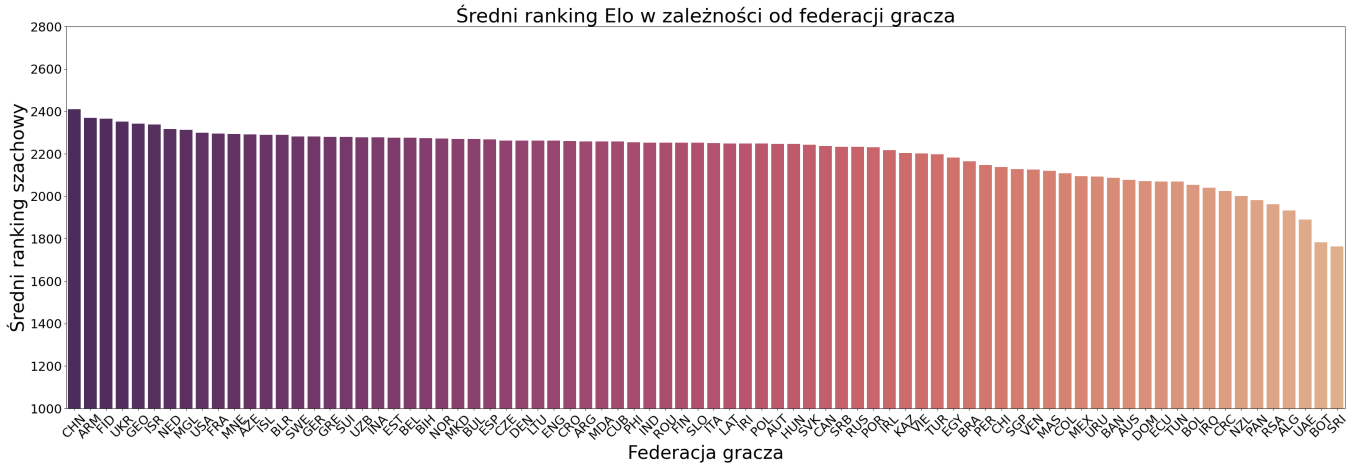
Rysunek 2: Wykres przedstawiający średni ranking Elo dla gracza o określonym tytule

Analiza wykresu przedstawiającego średnie rankingi Elo dla poszczególnych tytułów pozwoliła zaobserwować, że tytuły mają znaczący wpływ na ranking Elo gracza. Najwyższy tytuł, jakim jest arcymistrz (GM), jest powiązany z najwyższym średnim rankingiem Elo, wynoszącym około 2500. Tytuły takie jak mistrz międzynarodowy (IM), mistrz FIDE (FM) i kandydat na mistrza (CM) również odzwierciedlają poziom gry gracza, przy czym średnie rankingi Elo dla tych tytułów są niższe niż dla tytułu GM.

Wnioskiem z tej analizy jest to, że tytuły szachowe dobrze odzwierciedlają poziom gry gracza i są istotnym czynnikiem w przewidywaniu rankingu Elo. W modelu do predykcji Elo warto więc uwzględnić tytuły szachistów jako istotny parametr, który może wpływać na wyniki predykcyjne.

4.3 Federacja gracza a średni ranking Elo

Do przeprowadzenia analizy związku między federacją, z której pochodzi szachista, a jego rankingiem Elo wykorzystałam dane średnich rankingów Elo dla graczy z różnych federacji, uwzględniając jedynie te federacje, w których występowało co najmniej 20 graczy.



Rysunek 3: Wykres przedstawiający średni ranking Elo dla gracza pochodzącego z danej federacji

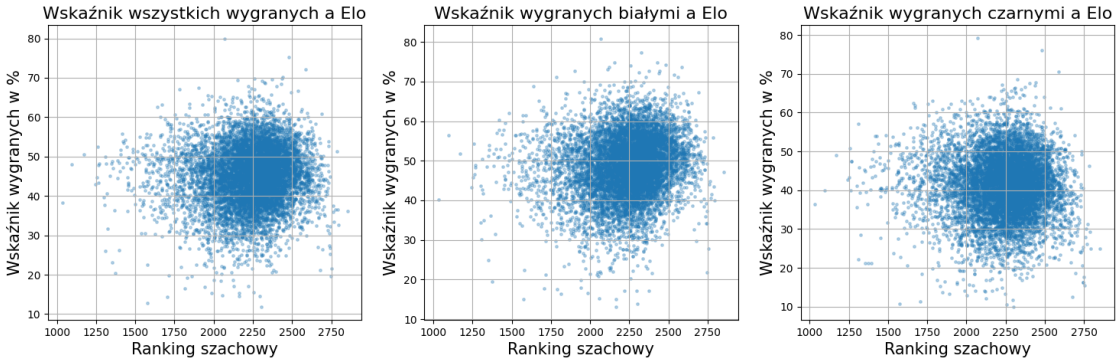
Analiza wykresu przedstawiającego średnie rankingi Elo dla poszczególnych federacji pozwoliła zauważyć pewne tendencje. W niektórych federacjach, takich jak Chiny czy Armenia, średnie rankingi graczy są relatywnie wysokie, osiągają około 2400 Elo. Sugeruje to, że gracze pochodzący z tych federacji wykazują się wysoką siłą szachową.

Z drugiej strony, istnieją federacje o słabszych średnich rankingach Elo, takie jak Botswana czy Sri Lanka, gdzie średnie Elo graczy wynosi około 1800. Może to wynikać z gorszej jakości treningów szachowych.

Jednak dla większości federacji można zauważyć, że różnice w średnich rankingach Elo są niewielkie. Większość federacji posiada graczy, których średnie Elo wynosi między 2000 a 2300.

Różnice między poszczególnymi federacjami w kontekście rankingów Elo są zazwyczaj małe i nie ma silnych dowodów na istnienie istotnego związku między federacją a rankingiem Elo. Wobec tego uwzględnienie federacji jako istotnego czynnika w modelu przewidującym ranking szachowy *nie jest konieczne*.

4.4 Wskaźnik wygranych a ranking Elo



Rysunek 4: Wykresy pokazujące zależność między wskaźnikiem wygranych partii danego gracza a jego rankingiem Elo

Analiza wykresów rozproszonych przedstawiających zależność wskaźnika wygranych partii od rankingu Elo pozwoliła na wyciągnięcie kilku istotnych wniosków. W analizie uwzględniono wskaźnik wygranych dla wszystkich partii, partii granych białymi pionkami oraz partii granych czarnymi pionkami.

Obserwując wykresy, można zauważyć, że różnice między wskaźnikiem wygranych partii dla wszystkich partii, partii białymi oraz partii czarnymi są zbliżone. Największe stężenie punktów znajduje się w okolicach 50% wygranych, co sugeruje, że większość szachistów osiąga równowagę między zwycięstwami a porażkami.

Warto zauważyć, że istnieje niewielka różnica między wskaźnikiem wygranych białymi a czarnymi. Mniej graczy zaczynających partię czarnymi ma wskaźnik wygranych między 50% a 60%, co jest zgodne z powszechnym przekonaniem, że grając czarnymi jest trudniej osiągnąć zwycięstwo. Jednak ogólnie wyniki są zbliżone dla obu grup.

Na podstawie analizy można wnioskować, że do modelu predykcyjnego rankingu Elo można wykorzystać wskaźnik wszystkich wygranych, który uwzględnia zarówno wygrane czarnymi, jak i białymi. Nie ma konieczności rozważania osobno wskaźników wygranych białymi i czarnymi, ponieważ różnice między nimi są niewielkie. Wykorzystanie wskaźnika wszystkich wygranych uprości model i zachowa wystarczającą informację dotyczącą zdolności szachistów do osiągania zwycięstw niezależnie od koloru pionków, co może prowadzić do bardziej efektywnych predykcji rankingu Elo.

4.5 Wskaźnik remisów a ranking Elo

Wykresy przedstawiają trzy różne scenariusze: wszystkie zremisowane partie, zremisowane partie białymi pionkami oraz zremisowane partie czarnymi pionkami.

Rysunek 5: Wykresy pokazujące zależność między wskaźnikiem zremisowanych partii danego gracza a jego rankingiem Elo

Każdy z trzech wykresów wykazuje podobny wzorzec. Najbardziej charakterystycznym spostrzeżeniem jest tendencja wzrostowa wskaźnika remisów dla graczy o wyższym rankingu Elo. Ilość zremisowanych partii u tych graczy wynosi nawet 50-60%. Dla graczy o rankingu mniejszym niż 2000 Elo, wskaźnik remisów oscyluje w zakresie od 10% do 40%. Istnieje możliwość, że wynika to z mniejszych umiejętności tych graczy i braku precyzji, które umożliwiałyby im osiąganie remisów.

U większości doświadczonych graczy, wskaźnik remisów również utrzymuje się na poziomie 10-40%, jednak w ich przypadku wiąże się to z wyższym wskaźnikiem wygranych. Niedoświadczeni gracze mają tendencję do wygrywania większej liczby gier, ale rzadziej remisują.

Na podstawie analizy można wnioskować, że wskaźnik remisów może być istotnym czynnikiem predykcyjnym dla rankingu Elo. Podobnie jak w przypadku wskaźnika wygranych, warto uwzględnić wskaźnik remisów dla wszystkich partii, ponieważ dostarcza on istotnej informacji o umiejętnościach szachistów i ich zdolności do remisowania niezależnie od koloru pionków.

5 Przygotowanie modelu

Wybranym przeze mnie modelem do predykcji rankingu Elo szachistów jest model SVR (Support Vector Regression).

Wybrałam regresję zamiast klasyfikacji do tego konkretnego problemu, ponieważ ranking Elo jest wartością ciągłą i reprezentuje dokładny poziom umiejętności szachisty. Model regresji pozwala uwzględnić zależności między zmiennymi, co jest kluczowe dla dokładnego przewidywania wartości numerycznych.

W przypadku rankingu Elo, klasyfikacja mogłaby być użyteczna do przewidzenia, do jakiego przedziału rankingu Elo należy dany szachista (np. niski, średni, wysoki). Jednak w tym przypadku interesuje nas konkretna wartość rankingu Elo, a nie tylko przynależność do określonego przedziału, dlatego regresja jest bardziej odpowiednia dla tego problemu.

5.1 Podzielenie danych

Model opiera się na czterech zmiennych, które zostały wybrane do przewidywania rankingu Elo. Są to: wiek szachisty, wskaźnik wygranych, wskaźnik remisów oraz tytuł szachowy, zapisany jako numer. Tytuł szachowy został zakodowany numerycznie, gdzie numer 1 odpowiada najwyższemu tytułowi (GM), a kolejne numery oznaczają niższe tytuły, według średniego rankingu Elo przypisanego danemu tytułowi (w kolejności takiej jak wynikało to z analizy danych).

Przed przystąpieniem do tworzenia modelu, dokonałam podziału danych na zbiór treningowy i zbiór testowy. 80% danych zostało użyte do treningu modelu, a pozostałe 20% zostało zarezerwowane do ewaluacji jego wydajności. Zbiór testowy składał się z około 1800 próbek.

Następnie, dla poprawy spójności i wydajności modelu SVR, przeprowadziłam skalowanie danych. Skalowanie jest szczególnie istotne dla tego modelu, ponieważ różnice w zakresach i jednostkach poszczególnych parametrów mogą wpływać na wyniki predykcji.

5.2 Wytrenowanie modelu

W procesie trenowania modelu SVR użyłam jądra rbf ze względu na jego zdolność do modelowania złożonych zależności nieliniowych, co jest istotne w przypadku analizy danych szachowych, gdzie zależności mogą być skomplikowane. Pozwala to modelowi lepiej dopasować się do zróżnicowanych charakterystyk danych szachistów.

Do przeprowadzenia treningu modelu SVR przetestowałem różne wartości parametrów gamma i C. Wybrałam zakres wartości, który obejmuje zarówno małe, jak i duże wartości, aby sprawdzić, jak różne parametry wpływają na jakość modelu i predykcje rankingu Elo szachistów.

W przypadku parametru gamma, rozważałam wartości od 0.01 do 10. Wybór takiego zakresu wynikał z faktu, że gamma kontroluje wpływ pojedynczego przykładu danych na granice decyzyjne. Wartości bliskie 0.01 oznaczają, że granice decyzyjne będą bardziej płynne i ogólne, podczas gdy wartości bliskie 10

skupiają uwagę modelu na pojedynczych punktach danych. Przez testowanie różnych wartości gamma, chciałam sprawdzić, jak dokładność predykcji rankingu Elo zmienia się w zależności od skali wpływu pojedynczych przykładów.

W przypadku parametru C, brałam pod uwagę wartości między 0.1 a 100. C kontroluje kompromis między dopasowaniem a regularyzacją w modelu SVR. Wybierając mniejszą wartość C, można ograniczyć dopasowanie modelu do danych treningowych, co może prowadzić do lepszej generalizacji. Z kolei większe wartości C pozwalają na bardziej dokładne dopasowanie do danych treningowych, ale mogą prowadzić do nadmiernego dopasowania. Przez testowanie różnych wartości C, chciałam znaleźć optymalną wartość, która zapewniłaby odpowiednie dopasowanie do danych treningowych, jednocześnie zapewniając dobre wyniki predykcji dla rankingu Elo szachistów na nowych danych.

5.3 Wyniki modelu i ocena jakości

Do oceny modelu SVR w przewidywaniu rankingu Elo wykorzystałam trzy metryki: Mean Squared Error (MSE), Mean Absolute Error (MAE) oraz R-squared (R2).

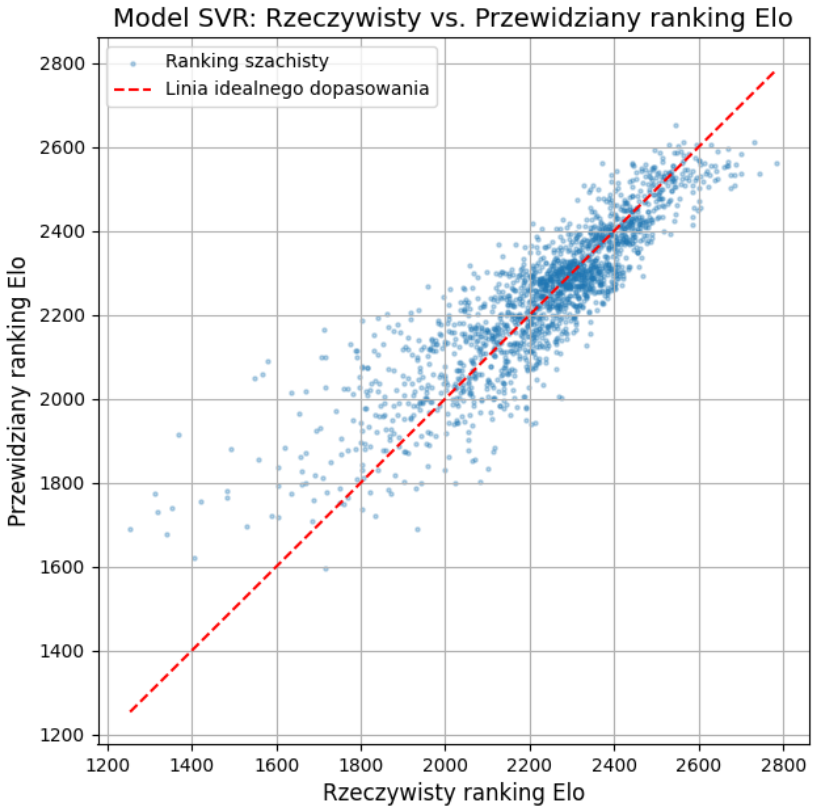
Na podstawie analizy i oceny wydajności modelu SVR dla różnych wartości gamma i C, ostatecznie wybrałam optymalne wartości gamma=0.15 oraz C=100, które zapewniają najlepszą predykcję rankingu ELO szachistów na podstawie wieku, wskaźnika wygranych, wskaźnika remisów oraz tytułu szachowego.

Dla podanych parametrów wyniki były następujące:

- Średni błąd kwadratowy (MSE): 10397.17
- Średni błąd bezwzględny (MAE): 74.60
- Współczynnik determinacji (R2): 0.76

Zaokrąglając wyniki, można stwierdzić, że średni błąd bezwzględny wynosi 75 punktów Elo. Jest to bardzo dobra predykcja, ponieważ wartości są bardzo zbliżone do rzeczywistego rankingu Elo. Średni błąd bezwzględny informuje o przeciętnej różnicy między przewidywanym rankingiem a rzeczywistym rankingiem szachisty. Im niższy jest ten błąd, tym bardziej precyzyjne są prognozy modelu.

W przypadku otrzymanego wyniku współczynnika determinacji równego 0.76, oznacza to, że około 76% zmienności w danych dotyczących rankingu Elo szachistów może być wyjaśnione przez opracowany model. Innymi słowy, model SVR jest w stanie dobrze odzwierciedlić i przewidzieć około 76% zróżnicowania wyników rankingu Elo w oparciu o dostępne informacje, takie jak wskaźnik wygranych, wskaźnik remisów, wiek i tytuł szachowy.



Rysunek 6: Wykres zestawiający rzeczywisty ranking Elo szachisty z rankingiem przewidywanym przez model SVR.

Na wykresie czerwona przerywana linia reprezentuje idealne dopasowanie, gdzie przewidywania modelu są dokładnie zgodne z rzeczywistością. Punkty znajdujące się powyżej czerwonej linii oznaczają rankingi, które model przewidział jako wyższe niż w rzeczywistości. Natomiast punkty poniżej czerwonej linii oznaczają rankingi, które model przewidział jako niższe niż w rzeczywistości.

6 Wnioski

Na podstawie analizy można wyciągnąć wniosek, że dane o szachistach uwzględniające wiek, tytuł szachowy, ilość wygranych i ilość remisów, są wystarczające, aby z dobrym przybliżeniem przewidzieć ranking szachisty. Do przewidywania rankingu odpowiednio sprawdza się model SVR o parametrach gamma=0.15 oraz C=100. Średni błąd bezwględny rankingu przewidywanego przez model wynosi 75 punktów Elo.

Model SVR osiągnął rezultaty bliskie rzeczywistemu rankingowi Elo. Jednak istnieje pewna rozbieżność w przewidywaniach modelu, szczególnie dla niektórych szachistów, których ranking został przeszacowany lub zaniżony. Model lepiej sprawdza się przy przewidywaniu szachistów o wyższym rankingu.

Największy wpływ na wyniki modelu ma tytuł szachisty, ponieważ jest on bezpośrednio związany z rankingiem Elo.