Analiza zależności popularności gier na platformie Twitch

Anna Majka, 266875

7 czerwca 2023 r., grupa K01-21d

Spis treści

1 Wstęp

Celem analizy danych jest ustalenie kluczowego czynnika, który wywiera największy wpływ na popularność gier na platformie Twitch. Wartości, które biorę pod uwagę, to:

- średnia liczba widzów,
- średnia liczba streamowanych godzin,
- liczba języków, w jakich gra była streamowana,
- liczba streamów z danej gry,
- liczba wystąpień gry w rankingu najpopularniejszych gier.

Przeanalizowanie tych danych pozwoli zbadać, jaki związek istnieje między tymi czynnikami a popularnością gier na Twitch. Chcę zidentyfikować najważniejszy czynnik, który może mieć największy wpływ na pozycję danej gry w rankingu popularności.

Celem projektu jest opracowanie modelu, który umożliwi estymację popularności gier na podstawie powyższych wartości. Umożliwi to lepsze zrozumienie, które czynniki są kluczowe dla osiągnięcia wysokiej popularności na platformie Twitch.

2 Pobranie zbioru danych

W celu uzyskania zbioru danych wykorzystałam oficjalne API udostępnione przez twórców platformy Twitch. Skorzystałam z odpowiednich endpointów w celu uzyskania potrzebnych informacji na temat aktualnie najpopularniejszych gier na platformie oraz wszystkich aktualnie odbywających się streamów związanych z konkretną grą.

API pozwoliło mi na dostęp do aktualnych danych, które były niezbędne do przeprowadzenia analizy. Dane dotyczące popularności gier na Twitchu zostały pobrane z oficjalnych źródeł, co gwarantuje ich wiarygodność. Pobrane dane są kluczowym źródłem informacji, które posłużą do przeprowadzenia analizy i odpowiedzi na pytanie dotyczące wpływu różnych czynników na popularność gier na tej platformie. Niestety, API nie daje możliwości pobierania danych archiwalnych, dlatego dane były pobierane ręcznie, codziennie przez niecały miesiąc, o losowych godzinach, aby uzyskać jak najbardziej wiarygodne rezultaty.

3 Przygotowanie zbioru danych

W celu przygotowania odpowiednich danych, obliczyłam odpowiednie statystyki dla każdej gry na podstawie wszystkich zebranych informacji na temat transmisji na żywo z danej gry oraz jej wystąpień w rankingu najpopularniejszych gier. W przypadku, gdy dla danej gry brakowało wszystkich informacji na temat transmisji przeprowadzonych na żywo, gra była pomijana i nie była brana pod uwagę. Uznałam to za najlepszą metodę radzenia sobie z pustymi wartościami, ponieważ na jedną transmisję na żywo składa się tak wiele czynników, że odtworzenie ich wszystkich byłoby niemożliwe. W przypadku innych nieprawidłowości związanych z danymi, sprawdzałam ręcznie, który parametr powoduje błąd i wprowadzałam odpowiednie poprawki. Po przetworzeniu wszystkich danych ostateczny zbiór składał się z informacji na temat ponad 3 tysięcy gier.

4 Wstępna analiza danych

	game name	records	id	languages	avg viewers	avg hours	average place	amount in top
0	Just Chatting	394	509658	15	8962.332487	5.845590	1.238095	21
1	VALORANT	389	516575	12	5564.159383	5.512022	3.571429	21
2	League of Legends	398	21779	14	6729.306533	5.752093	3.619048	21
3	Counter-Strike Global Offensive	397	32399	17	6396.075567	6.296331	5.666667	21
4	Grand Theft Auto V	394	32982	19	2412.241117	6.000097	5.952381	21

Rysunek 1: Fragment danych po odpowiednich obliczeniach

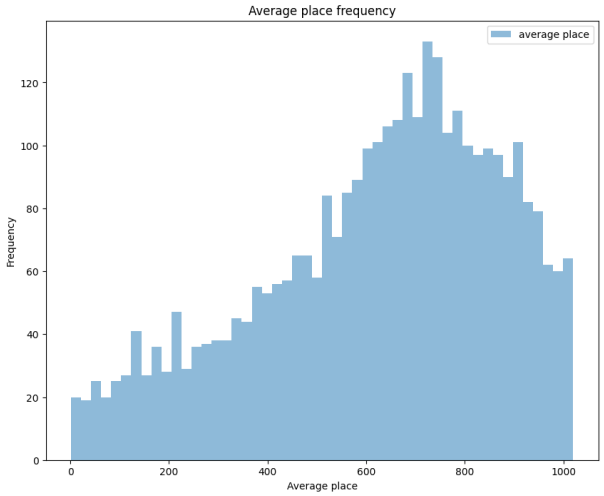
4.1 Analiza poszczególnych parametrów

4.1.1 Średnie miejsce w rankingu

Wartość minimalna	Wartość maksymalna	Średnia wartość
1,238	1020	622,712

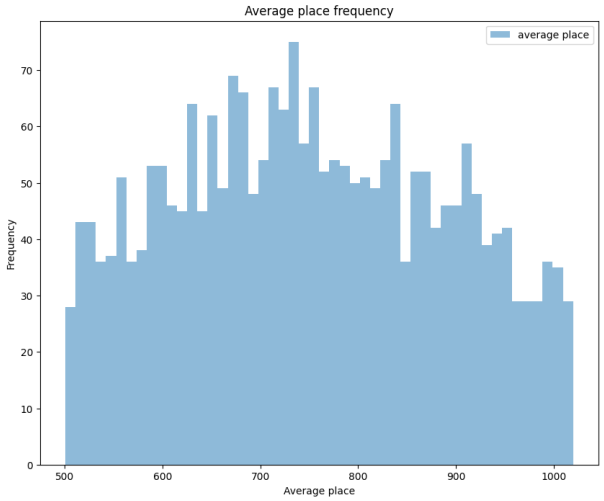
Tabela 1: Statystyki dla średniego miejsca w rankingu

Rozkład dla wszystkich danych:



Rysunek 2: Histogram średniego miejsca w rankingu

Rozkład dla najczęściej powtarzających się wartości:



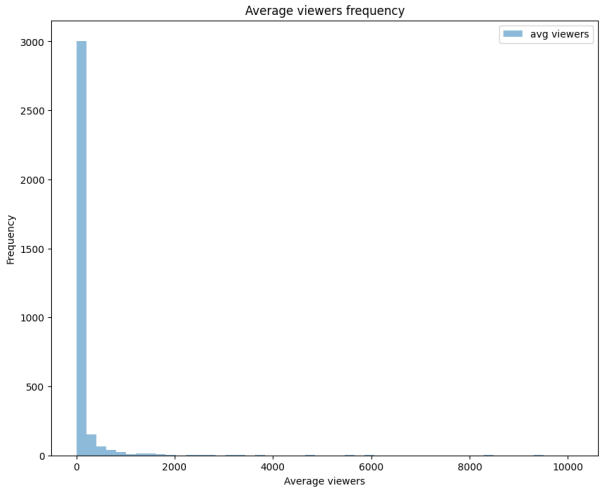
Rysunek 3: Histogram średniego miejsca w rankingu dla najczęściej pojawiających się wartości

4.1.2 Średnia liczba widzów

Wartość minimalna	Wartość maksymalna	Średnia wartość
1	10111	151,750

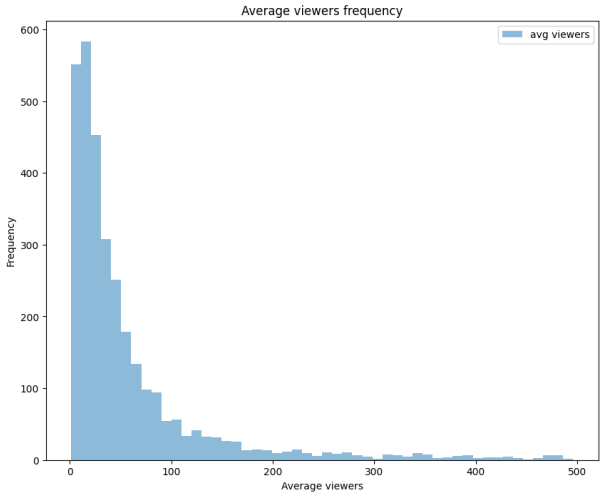
Tabela 2: Statystyki dla średniej liczby widzów

Rozkład dla wszystkich danych:



Rysunek 4: Histogram średniej liczby widzów

Rozkład dla najczęściej powtarzających się wartości:



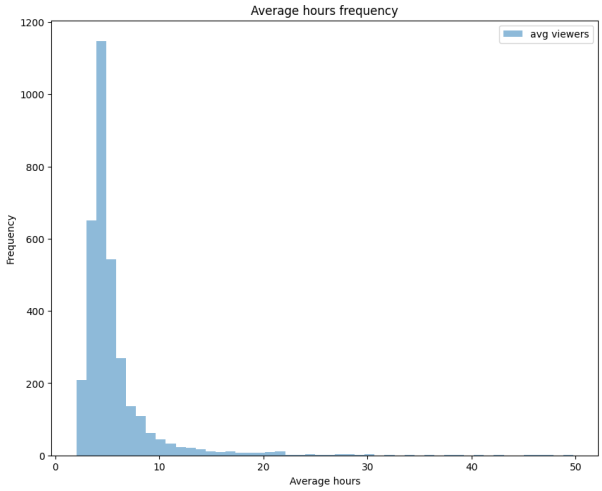
Rysunek 5: Histogram średniej liczby widzów dla najczęściej pojawiających się wartości

4.1.3 Średnia liczba streamowanych godzin

Wartość minimalna	Wartość maksymalna	Średnia wartość
2,032	49,791	5,623

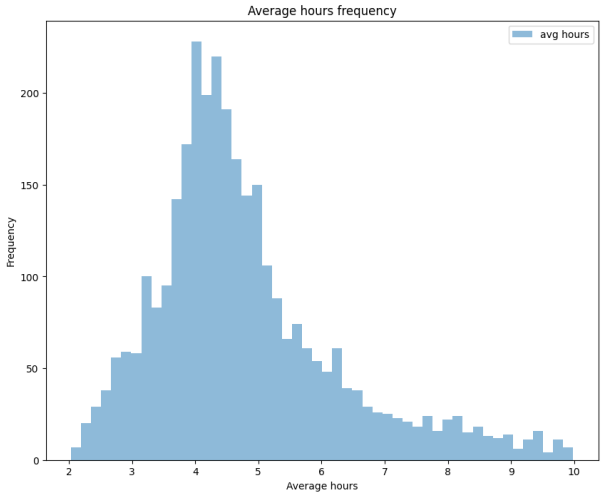
Tabela 3: Statystyki dla średniej liczby streamowanych godzin

Rozkład dla wszystkich danych:



Rysunek 6: Histogram średniej liczby streamowanych godzin

Rozkład dla najczęściej powtarzających się wartości:



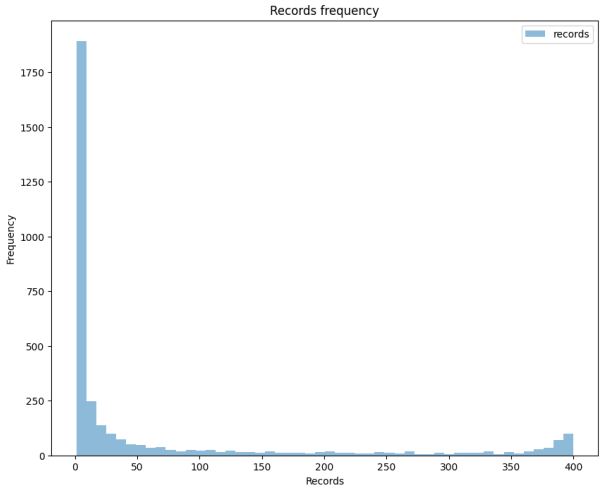
Rysunek 7: Histogram średniej liczby streamowanych godzin dla najczęściej pojawiających się wartości

4.1.4 Liczba streamów z danej gry

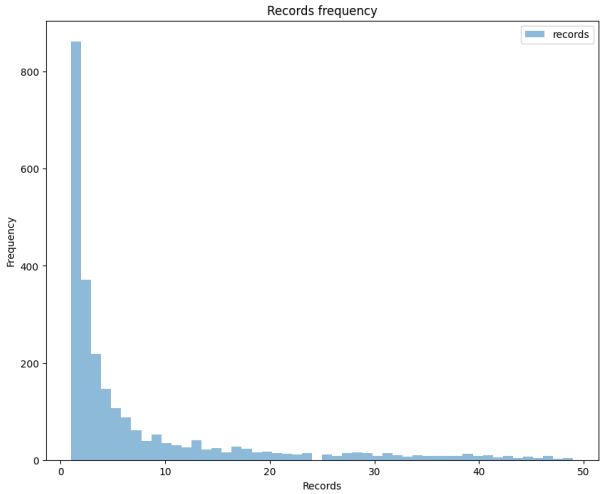
Wartość minimalna	Wartość maksymalna	Średnia wartość
1	400	65,725

Tabela 4: Statystyki dla liczby streamów z danej gry

Rozkład dla wszystkich danych:



Rozkład dla najczęściej powtarzających się wartości:



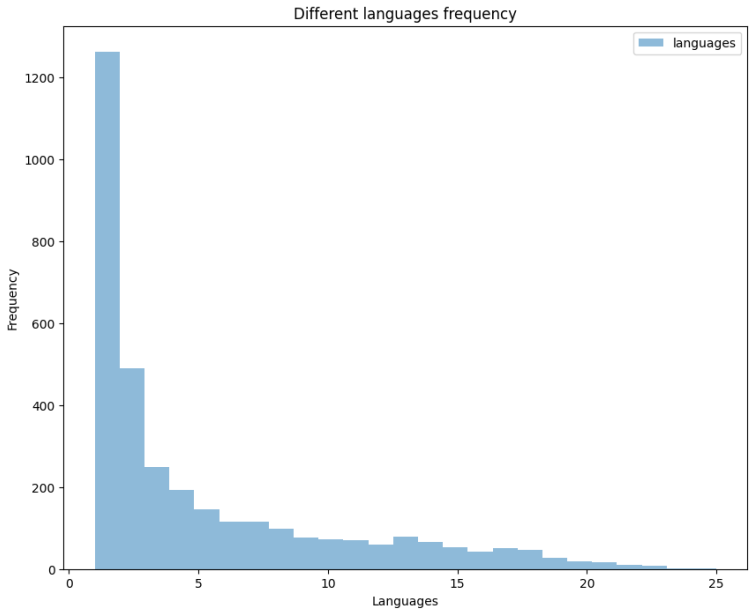
Rysunek 8: Histogram średniej liczby streamów z danej gry dla najczęściej pojawiających się wartości

4.1.5 Liczba języków, w jakich gra była streamowana

Wartość minimalna	Wartość maksymalna	Średnia wartość
1	25	4,909

Tabela 5: Statystyki dla liczby języków, w jakich gra była streamowana

Rozkład dla wszystkich danych:



Rysunek 9: Histogram średniej liczby streamów z danej gry

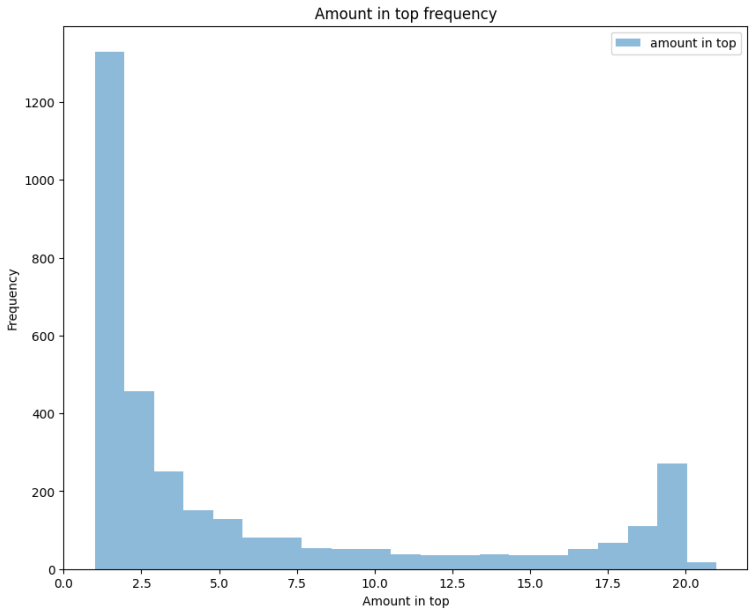
Nie ma potrzeby tworzenia dokładniejszego histogramu dla tych danych.

4.1.6 Liczba wystąpień gry w rankingu najpopularniejszych gier

Wartość minimalna	Wartość maksymalna	Średnia wartość
1	21	5,820

Tabela 6: Statystyki dla liczby wystąpień gry w rankingu

Rozkład dla wszystkich danych:

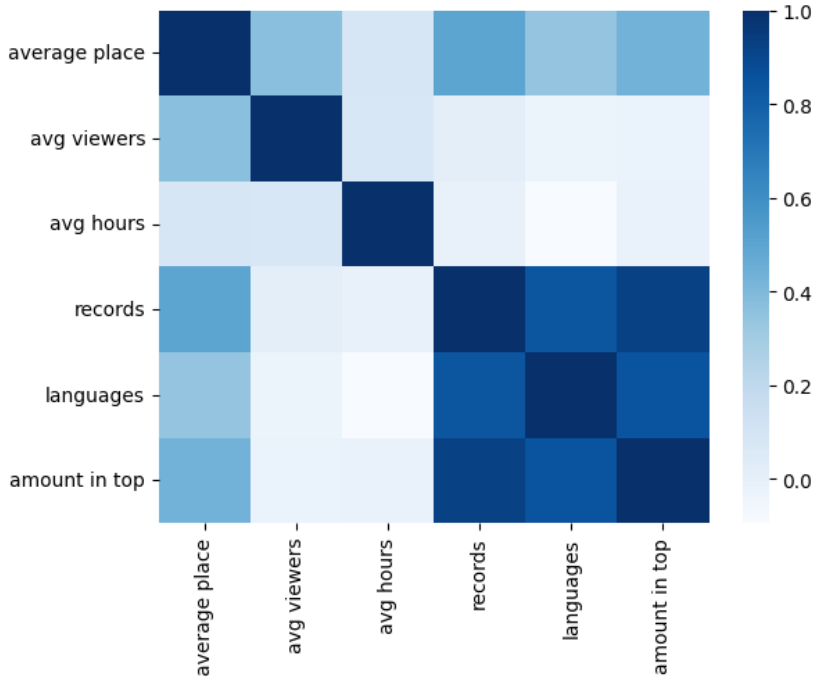


Rysunek 10: Histogram liczby wystąpień gry w rankingu

Nie ma potrzeby tworzenia dokładniejszego histogramu dla tych danych.

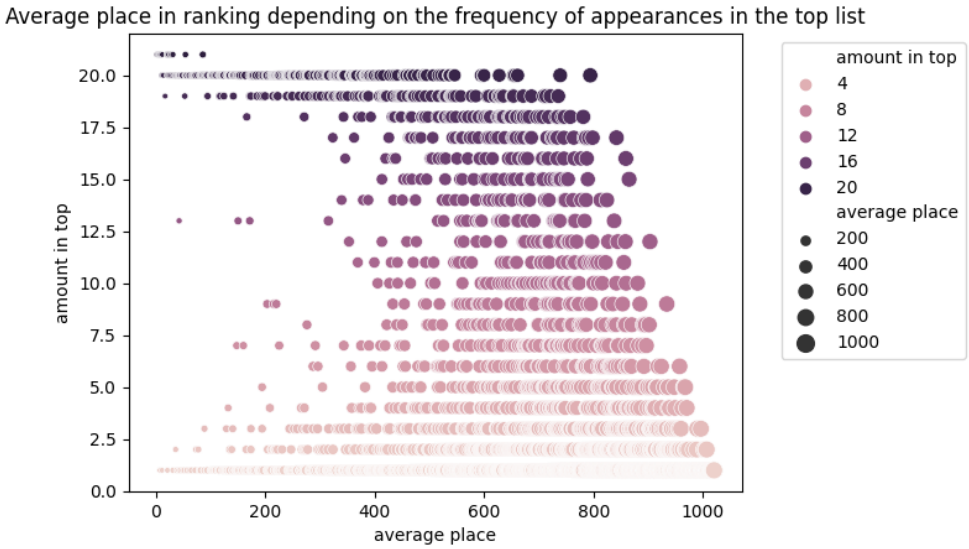
4.2 Zależności między danymi

Mapa korelacji między zmiennymi wskazuje na zwiększoną zależność między średnim miejscem w rankingu a liczbą pojawień się gry w rankingu oraz liczbą streamów z danej gry.

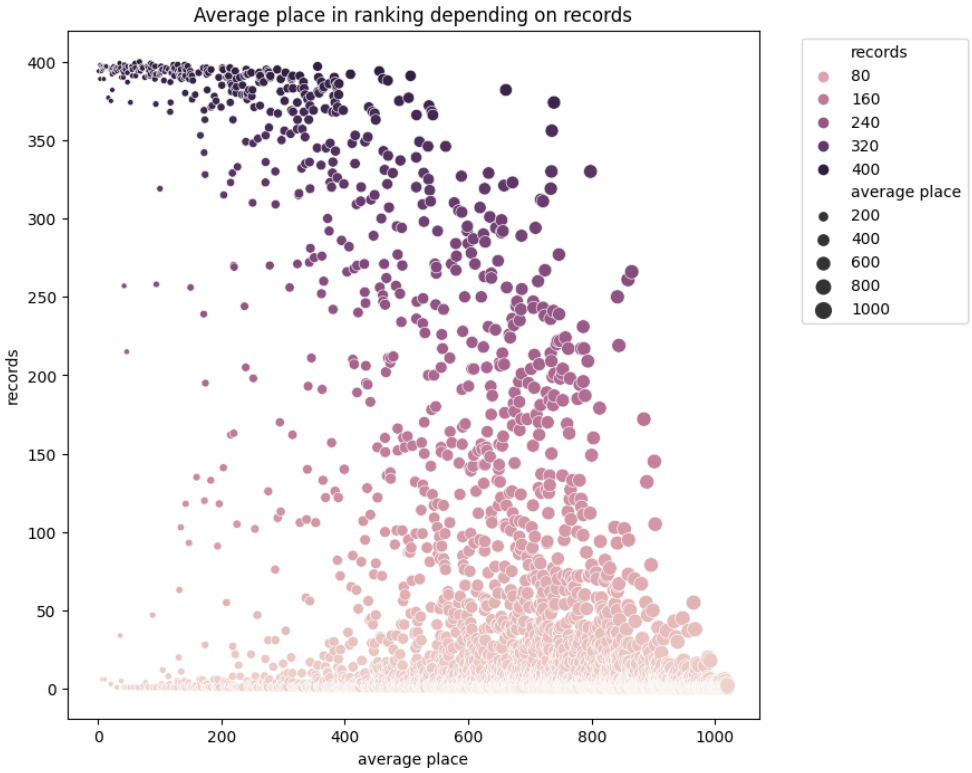


Rysunek 11: Mapa korelacji pomiędzy zmiennymi

Wizualizacja zależności pomiędzy średnim miejscem w rankingu a liczbą pojawień się gry w rankingu oraz liczbą streamów z danej gry za pomocą wykresów punktowych:



Rysunek 12: Wykres przedstawiający zależność pomiędzy średnim miejscem w rankingu, a liczbą pojawień się gry w rankingu



Rysunek 13: Wykres przedstawiający zależność pomiędzy średnim miejscem w rankingu, a liczbą streamów z danej gry

Pomimo zwiększonej korelacji pomiędzy zmiennymi trudno jest przedstawić liniowo określone zależności.

5 Działanie na modelu

5.1 Wybór odpowiedniego modelu

Uznałam, że ze względu na brak jednoznacznej liniowej zależności pomiędzy cechami opisującymi grę a jej popularnością, skorzystam z modelu Lasu Losowego (Random Forest). Działanie modelu Random Forest polega na utworzeniu wielu drzew decyzyjnych, które są kreowane na podstawie losowego podzbioru danych treningowych. Każde drzewo w lesie podejmuje niezależne decyzje na podstawie losowych podzbiorów cech. Wyniki tych drzew są następnie agregowane, aby uzyskać końcową predykcję.

Drzewa są konstruowane poprzez podział węzłów na podstawie cech, które najlepiej separują zbiór danych treningowych. Proces ten jest powtarzany rekurencyjnie, aż do osiągnięcia warunku zakończenia, na przykład maksymalnej głębokości drzewa.

Po utworzeniu wszystkich drzew w lesie, wykonuje się predykcję dla nowych danych. Każde drzewo przewiduje wynik niezależnie, a następnie wyniki są agregowane na podstawie średniej, aby otrzymać ostateczny wynik przewidywania dla całego lasu.

5.2 Podział danych

Wszystkie zebrane dane podzieliłam na zbiór treningowy oraz zbiór testowy w stosunku 80:20. Dane testowe zostały wybrane losowo. Przy każdym uruchomieniu programu korzystałam z tego samego podziału danych.

5.3 Tworzenie modelu

Dla jak najlepszego doboru parametrów modelu skorzystałam z funkcji Przeszukiwania Siatki (Grid Search).

Grid Search polega na przetestowaniu różnych kombinacji hiperparametrów (czyli takich parametrów, które nie są otrzymywane w procesie uczenia, ale wpływają na działanie modelu, jak na przykład liczba drzew w lesie losowym) zdefiniowanych przez użytkownika, aby znaleźć te, które dają najlepsze wyniki. W praktyce, dla każdej kombinacji hiperparametrów, trenuje się i ocenia model, a następnie porównuje wyniki, aby zidentyfikować optymalne ustawienia.

Ilość drzew	Maksymalna głębokość drzewa	Minimalna liczba próbek
400	10	5

Tabela 7: Najlepsze parametry uzyskane za pomocą funkcji Grid Search

5.4 Określenie jakości modelu

Do określenia jakości modelu skorzystałam z takich metryk jak błąd średniokwadratowy (MSE - Mean Squared Error) oraz współczynnik R^2.

MSE	R^2
9092,582	0,864

Tabela 8: Wyniki wybranych metryk dla modelu Random Forest

Na podstawie obliczonych metryk można stwierdzić, że model Losowego Lasu jest dobrze dopasowanym modelem do analizowanych danych, jednak błąd średniokwadratowy wskazuje na dużą rozbieżność pomiędzy przewidywanymi danymi, a rzeczywistymi.

Pomimo zwiększonej zależności pomiędzy średnim miejcem w rankingu a liczbą pojawień się gry w rankingu oraz liczbą streamów z danej gry wykazanej za pomocą mapy korelacji, według modelu czynnikiem mającym największy wpływ na popularność gry jest średnia liczba widzów.

Czynnik	Wskaźnik znaczenia
Średnia liczba widzów	0,581
Liczba wystąpień w rankingu	0,234
Liczba streamów	0,142
Średnia liczba godzin	0,03
Liczba języków	0,012

Tabela 9: Znaczenie poszczególnych czynników według modelu

5.5 Porównanie z innym modelem

Aby móc lepiej stwierdzić czy dopasowany model jest odpowiedni, porównałam jego działanie do modelu Maszyny Wektorów Nośnych (SVM - Support Vector Machine). Działanie modelu SVM polega na znalezieniu optymalnej hiperpłaszczyzny lub krzywej separacji, która w przypadku regresji minimalizuje błąd predykcji.

Analogicznie do modelu Random Forest skorzystałam z funkcji Grid Search w celu znalezienia optymalnych parametrów.

C	gamma
4,0	0,001

Tabela 10: Najlepsze parametry uzyskane za pomocą funkcji Grid Search

Zarówno dopasowanie modelu do danych jak i rozbieżność pomiędzy przewidywanymi danymi a rzeczywistymi są znacznie gorsze niż w przypadku modelu Random Forest.

MSE	R^2
48700,543	0,272

Tabela 11: Wyniki wybranych metryk dla modelu SVM

6 Wnioski

Na podstawie przeprowadzonych badań i ich wyników można stwierdzić, że pomimo dużej rozbieżności pomiędzy danymi, model dobrze dopasował się do danych i może być użyteczny nawet przy bardzo zróżnicowanych danych.

Zgodnie z przewidywaniami na podstawie mapy korelacji, popularność gier nie jest zależna od liczby streamowanych godzin oraz od liczby różnych języków, w jakich transmisje są przeprowadzane.

Model wykazał silną zależność popularności gier od średniej liczby widzów na streamie. Jednak należy wziąć pod uwagę fakt, że zebrane dane pochodzą z relatywnie krótkiego okresu czasu i dotyczą ograniczonej liczby gier, biorąc pod uwagę rozmiar platformy Twitch.

Oprócz niewielkiej grupy gier, które utrzymują stałą pozycję na szczycie, popularność większości gier jest bardzo nieregularna i może na nią wpływać wiele czynników, które nie są związane bezpośrednio z platformą Twitch, takich jak działania marketingowe producentów danej gry. Czasami nawet najbardziej absurdalne czynniki mogą wpływać na sposób, w jaki dana gra jest postrzegana przez społeczność graczy.