Analiza wpływu danych seriali i filmów na ich popularność

Mikołaj Jastrzębski
266858@student.pwr.edu.pl
K01-21c

10 czerwca 2023

Spis treści

1 Pobocznie analizowany problem

1.1 Opis pobocznego problemu

Pierwotną ideą projektu, była analiza danych giełdowych oraz utworzenie modelu, który na podstawie historii cen jak i danych dotyczących insider tradingu byłby w stanie dokonywać prognozy przyszłych cen. Projekt zakończył się przedwcześnie po zescrapowaniu danych, analizie ich, wstępnym procesowaniu oraz wizualizacji, co zostanie opisane w kolejnych krokach sekcji.

1.2 Zgromadzenie i wstępne przetworzenie danych pobocznego problemu

1.2.1 Zgromadzenie danych

Do analizy problemu pozyskano dane scrapując je ze stron:

1. Dane dotyczące historii cen akcji ze strony Yahoo Finance
2. Dane dotyczące insider trade'ów pozyskałem ze strony Finviz (Financial Visualizations)

Z pierwszej strony internetowej pozyskano dane dotyczące trzech ostanich lat historii cen akcji 'Nvidia', 'AMD' oraz indeksu 'Vanguard Growth Index' (VUG). Dla każdej z tych pozycji zescrapowano 750 rekordów dotyczących każdego dnia giełdy w okresie ostatnich trzech lat. Dane są pozyskiwane dzięki API. Początkowo, określane są ramy czassowe, dla którego dane będą pobierane oraz interwał, z którego będą pobierane. Dane będą pobierane od 1 marca 2020 roku do 1 marca 2023 roku dla każdego dnia kiedy giełda była otwarta. Następnie, tworzony jest URL, używając symbolu akcji (ticker) oraz wcześniej zdefiniowanych wartości czasowych. W tym przypadku, korzystamy z usługi Yahoo Finance, a konkretnie z ich API, aby pobrać dane w formacie CSV. Kod pobiera dane z podanego URL, czyli pliku CSV, i zwraca je jako obiekt DataFrame za pomocą biblioteki pandas. Przetwarzamy dane oraz nadawane są im odpowiednie nazwy kolumnom. W rezultacie, funkcja zwraca obiekt DataFrame zawierający dane dotyczące cen akcji w określonym czasie. Wartości danych:

- Data
- Cena otwarcia
- Cena zamknięcia
- Najwyższa cena odnotowana danego dnia
- Najniższa cena odnotowana danego dnia
- Cena zamknięcia po regulacji
- Liczba zakupionuch bądź sprzedanych akcji

Z drugiej strony insternetowej zescrapowano dane dotyczące ostatniego roku insider trade'ów w firmach 'Nvidia' oraz 'AMD'. Kod do pozyskiwania danych pobiera dane dotyczące insider tradingu dla każdego indeksu. Na początku tworzy jest URL, używając symbolu akcji, a następnie pobierana jest zawartość strony internetowej za pomocą biblioteki urllib i modułu Request. Następnie strona internetowa jest analizowana przy użyciu biblioteki BeautifulSoup, co umożliwia ekstrakcję danych z kodu HTML. Funkcja przetwarza pobrane dane, korzystając z biblioteki pandas, aby odczytać tabele HTML ze strony. Funkcja przeprowadza odpowiednie operacje przetwarzania danych, takie jak pominięcie pierwszego wiersza (nagłówka), nadanie nazw kolumnom, ustawienie daty jako indeksu. Na koniec, przetworzone dane są zwracane jako obiekt DataFrame. Dane te zawierają wartości:

- Data
- Osoba dokonująca transakcji
- Posada w firmie

- Rodzaj transakcji
- Cena za jedną akcje
- Liczba akcji
- Wartość akcji
- Sumaryczna wartość akcji danej osoby po transakcji

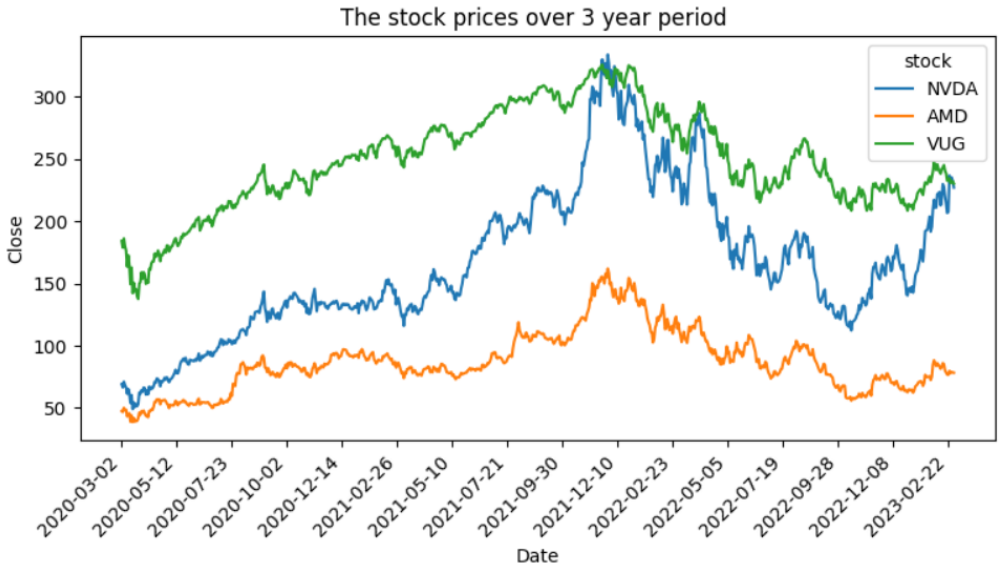
1.2.2 Wstępne przetworzenie danych

Pozyskane dane były stosunkowo kompletne, lecz zawierały kilka wad takich jak różnica w formacie daty między zbiorami danych, bądź brakujące wartości spowodowane dodawaniem kolumn.

1. Pierwszym krokiem dotyczącym poprawy jakości danych było zmienienie formatu dat na jednolity. Dane ze strony FinViz nie posiadały roku, lecz po krótkiej analizie klarownym stało się, iż są to jedynie dane z ostatnich 12 misięcy. Dla danych od stycznia do maja dodano rok 2023, a dla danych z czerwca do górnia 2022.
2. Brakującą oraz znaczącą kolumną do pracy na danych, była różnica procentowa wzrostu w porównaniu z dniem poprzednim dla każdej z akcji. Ze względu na jej brak, została dodana na podstawie obliczenia różnicy ceny zamknięcia akcji aktualnej z dniem poprzednim.
3. Konsekwencją stworzenia takiej kolumny było dodanie wartości "NaN"dla pierwszych dni każdej z akcji, ponieważ niemożliwym jest określenie procentu zmiany w porównaniu z dniem poprzednim, którego nie ma (w pozyskanych danych). Brakujące wartości zostały uzupełnione zerami, ze względu na zerową zmianę ceny w stosunku do nieegzystującego dnia poprzedniego.

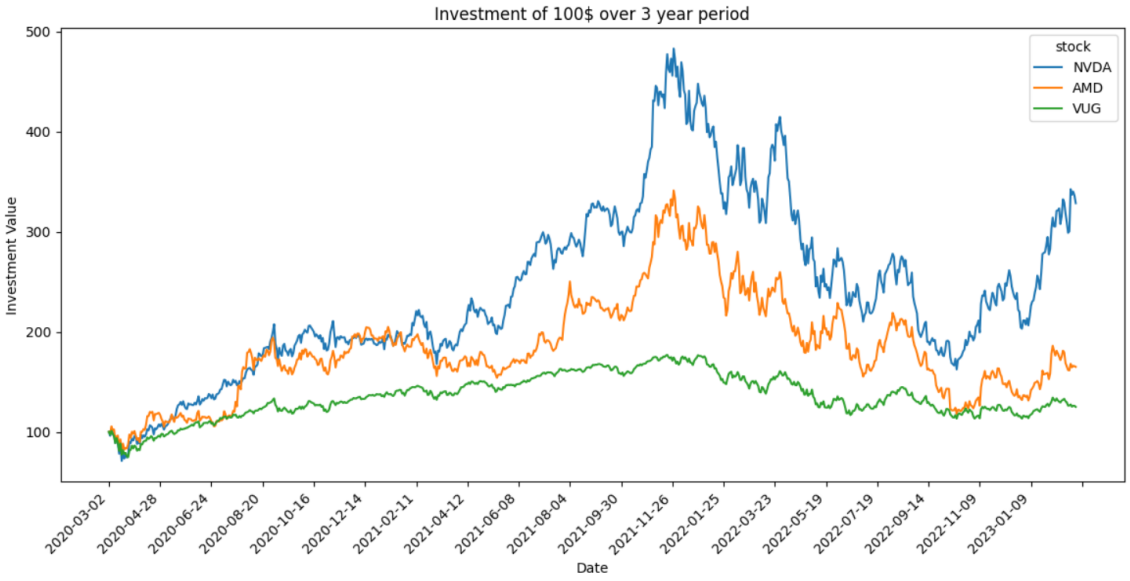
1.3 Wstępna analiza danych

Ze względu na porzucenie projektu, po dokonaniu kluczowych obserwaji, zostaną zawarte jedynie kluczowe wykresy, których wnioski zostaną określone w kolejnej sekcji.



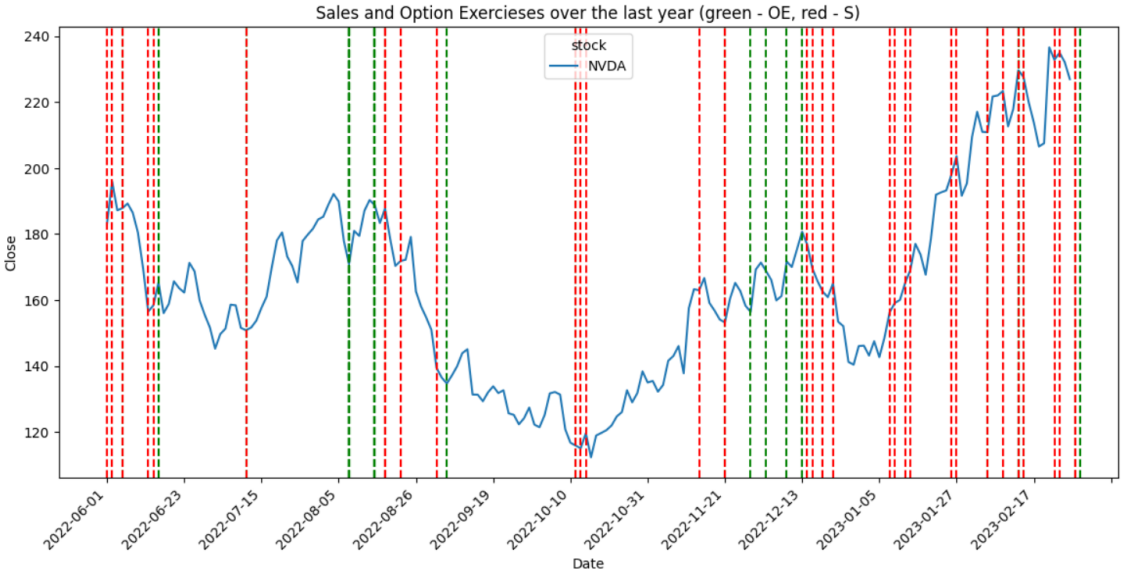
Rysunek 1: Wykres cen akcji na przestrzeni 3 lat

Rysunek 1 pozwala nam zagłębić się w szczegóły i ogólną analizę danych dotyczących cen akcji. Już na pierwszy rzut oka łatwo zaobserwować, iż ceny akcji Nvidia się stanowczo wahają, wręcz przeciwnie do cen AMD, a szczególnie indeksu VUG.



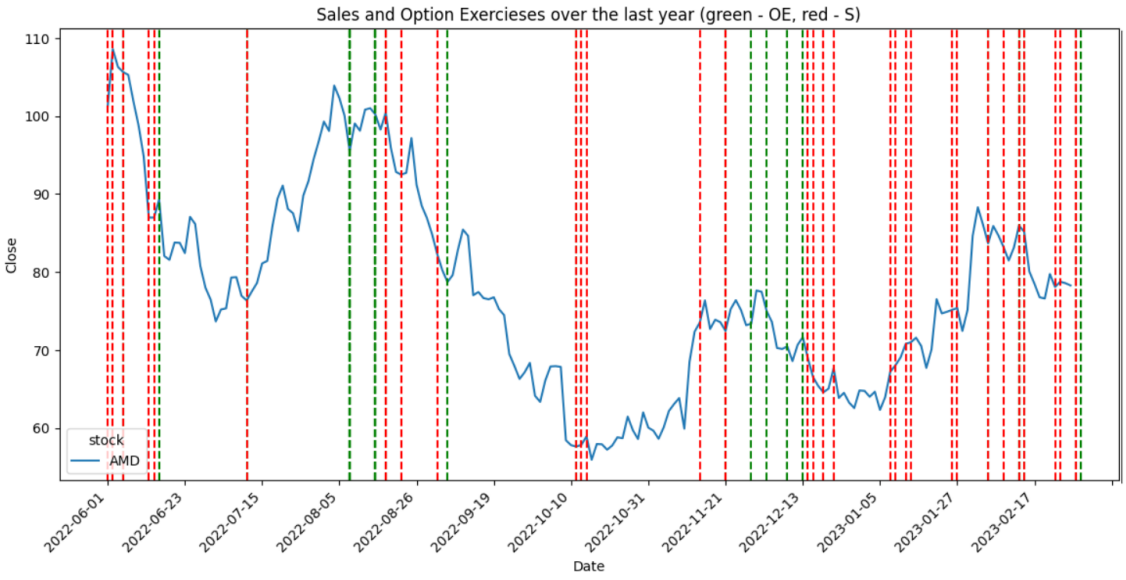
Rysunek 2: Wykres wartości zainwestowania 100 dolarów w daną akcję na przestrzeni 3 lat

Na wykresie 2 widać, iż Nvidia jest wyraźnie najsilniejszą akcją. Jest również bardzo wrażliwa na wahania, dlatego możemy uznać ją za najbardziej ryzykowną inwestycję w zbiorze danych. Dość podobna zależność występuje w przypadku AMD, lecz na wykresie można zauważyć znaczne mniejsze fluktuacje. Nvidia wydaje się być silnie powiązana z AMD. Różnica jest w wartości, ale jest to całkowicie normalne, ponieważ Nvidia jest bardziej podatna na wahania, dlatego, mniej więcej za każdym razem, gdy wartość akcji AMD przesuwa się o 1 procent na osi Y, Nvidia ma tendencję do przesuwania się 1,5 raza bardziej.
Najważniejszym i najbardziej niefortunnym aspektem jest fakt, że dane nie wskazują żadnego konkretnego wzorca. W tym momencie staje się zauważalne, że stworzenie prawidłowego modelu byłoby prawie niemożliwe.



Rysunek 3: Wykres akcji Nvidia, gdzie czerwony kolor oznacza sprzedaż, a zielony zakup

Finalnie na wykresach 3 oraz 4 widać, iż niestety sprzedaż jak i zakup akcji przez pracowników firmy ma znikome znaczenie. Jest to finalny powód dla którego tworzenie sensownego modelu mijałoby się z celem.



Rysunek 4: Wykres akcji AMD, gdzie czerwony kolor oznacza sprzedaż, a zielony zakup

1.4 Powód zakończenia prac nad problemem giełdowym

Dane dotyczące historii cen akcji oraz dane o transakcjach insiderów, które zostały zescrapowane ręcznie, niestety nie pozwalają na spełnienie celu projektu. Dane nie wskazują żadanego schematu bądź charakterystycznych trendów, ceny są niemalże losowe, a insider trade'y są w nieznacznym stopniu powiązane z cenami. Tworzenie modelu predykcyjnego na tym zbiorze danych byłoby zbliżone do wróżenia z kryształowej kuli, stąd projekt zostaje porzucony. W kolejnej części opisany został główny projekt w którym pominięto etap pozyskiwania danych poprzez scrapowanie, gdyż zostało to dokonane w tej części.

2 Wstęp do głównego problemu

Celem głównego projektu, którego finalnie dotyczy ta praca, jest zbadanie wpływu danych seriali oraz filmów na ich popularność. Dane, które wybrałem do analizy to:

- Rok wydania produkcji
- Ograniczenia wiekowe
- Gatunek
- Kraj produkcji
- Liczba sezonów
- Liczba głosów z forum filmowego Imdb
- czy produkcja jest filmem, czy serialen

Stworzony w projekcie model ma na celu prognozowanie popularności produkcji na podstawie wyżej wymienionych danych. Taka analiza może być przydatna dla fanów kinematografii szukających nowej, potencjalnie popularnej, produkcji do obejrzenia lub dla studio zainteresowanego wyprodukowaniem nowego hitu na skalę globalną.

3 Zgromadzenie i przetworzenie danych głównego problemu

3.1 Zgromadzenie danych

Dane pozyskane do głównego problemu pochodzą z Kaggle. Są to dane dotyczące Seriali oraz Filmów z platformy Netflix oraz zawierają 5806 rekordów w formcie '.csv'. Dane zawierają takie kolumny jak:

- id - identyfikator filmu
- title - nazwa produkcji
- show type - informacja czy produkcja jest serialem czy filmem
- description - opis produkcji
- release year - rok wydania produkcji
- age certification - ograniczenie wiekowe dla produkcji
- runtime - długość w minutach danej produkcji
- genres - gatunek produkcji
- production countries - kraj produkcji
- seasons - liczba sezonów
- imdb id - identyfikator imdb
- imdb score - ocena imdb
- imdb votes - liczba głosów imdb
- tmdb popularity - indeks popularności tmdb
- tmdb score - ocena tmdb

Ten zestaw danych został utworzony w celu zestawienia wszystkich produkcji dostępnych w serwisie Netflix. Dane zostały pozyskane w lipcu 2022 r. i obejmują dane dostępne w Stanach Zjednoczonych.

3.2 Przetworzenie danych

W celu ułatwienia analizy danych oraz zrozumienia ich niektóre kolumny zostały przekonwertowane w ich klarowniejsze wersje.

1. Kolumna 'Date' została zastąpiona bardziej konwencjonalną, 'Decade', symbolizującą dekadę wydania danego filmu. Generalizowanie okresu wydania produkcji jest powszechnie używane w rankingach, jak i recenzjach oraz ułatwia skonstuowanie sensownego modelu.
2. Rekordy w kolumnie 'Genres' okazyjnie zawierały informacje o większej liczby gatunków dla jednego filmu, bądź serialu. Z tego powodu kolumna została skrócona, aby zawierać jedynie informacje o najlepiej definiującym gatunku dla danej produkcji (pierwszym wspomnianym).
3. Ten sam problem dotyczył kolumny 'Production country', która zawierała informacje o kilku państwach, w których dana produkcja została nakręcona. W celu klaryfikacji dane te zostały skondensowane do wartości atomowych.

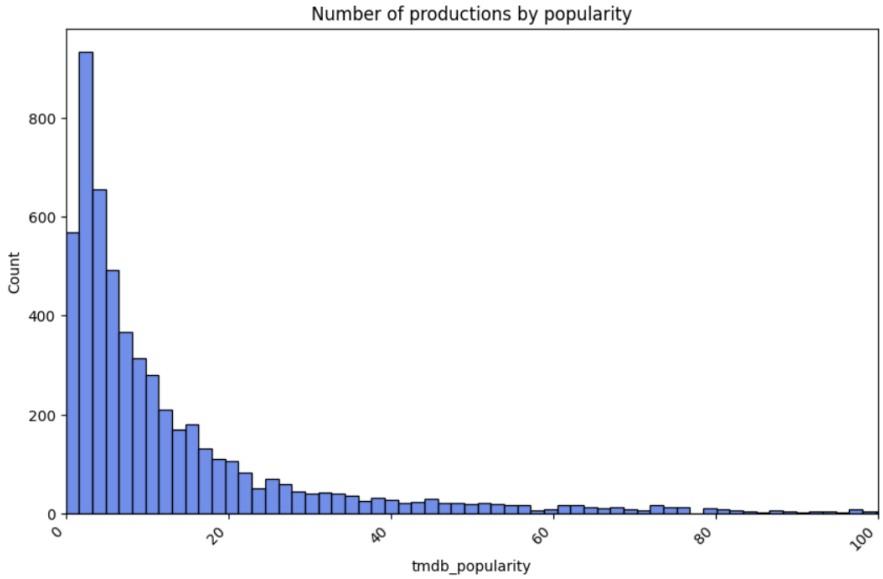
W kwestii uzupełniania danych pominięte oraz niebrane do analizy zostały rekordy niezawierające 'imdb sore', 'tmdb score' oraz 'tmdb popularity', ponieważ sztuczne uzupełnianie tych danych mogłoby negatywnie wpłynąć na wiarygodność modelu.
Uzupełnione zostały dane:

1. 'age certification' wartościami "N/A"ze względu na ogromną trudność prognozowania ograniczeń wiekowych dla danych filów i seriali.

2. 'seasons' wartościami 0, jako iż najczęściej dane nie były dostępne ze względu na fakt, iż filmy nie posiadają jakiejkolwiek liczby sezonów, posiadają one części.

3. 'imdb votes' wartością średnią, ponieważ brakowało ich niewiele, a wartość średnia w tym przypadku jest najrozsądniejszym rozwiązaniem.

3.3 Wstępna analiza danych



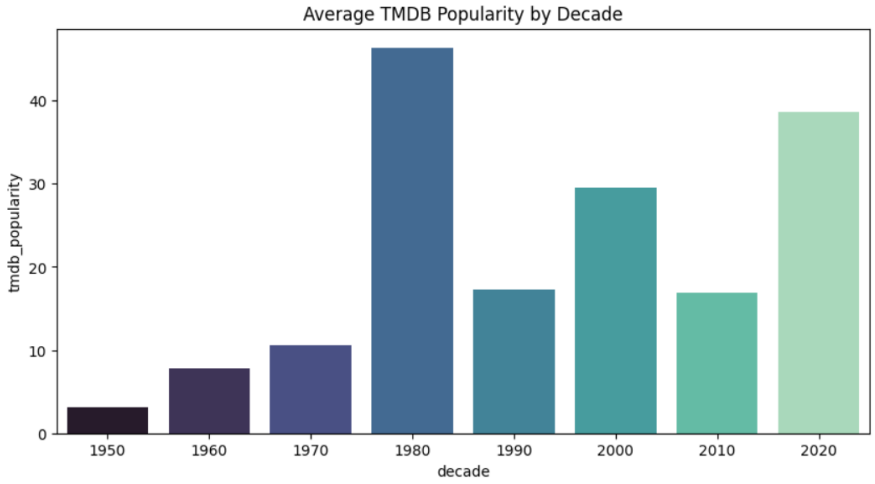
Rysunek 5: Histogram popularności produkcji z ograniczniem osi x do 100

rozmiar	wartość min	wartość max	wartość średnia
5806	0.009	1823.374	22.525

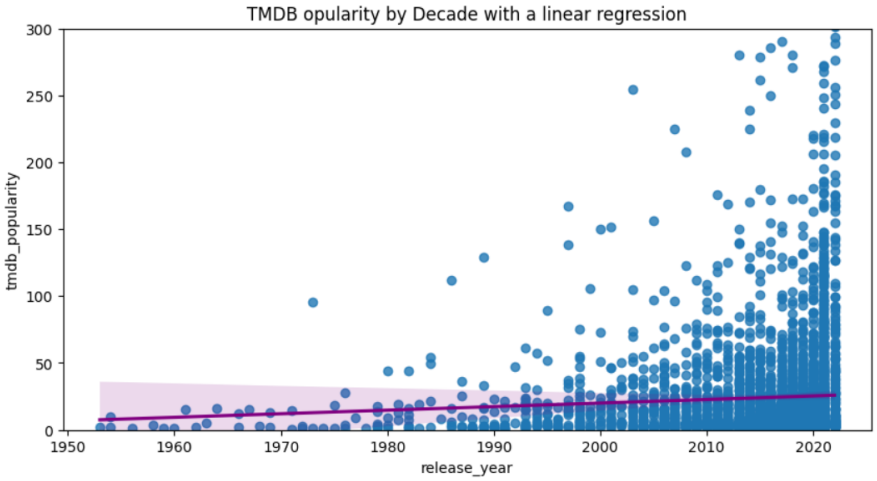
Tabela 1: Dane dotyczące rozkładu danych popularności

Dane definitywnie nie są zbliżone do rozkładu normalnego. Są skupione w obszarze 0-10 popularności, stąd mimo bardzo rozległych wartości minimalnej (0.009) oraz maksymalnej 91823.373) średnia popularność dla filmu na Netflix wynosi 22.525 dla 5806 rekordów. Może to być podstawa do dużej rozbieżności w prognozowaniu modelu.

3.3.1 Wpływ dekady wydania na popularność i ocenę

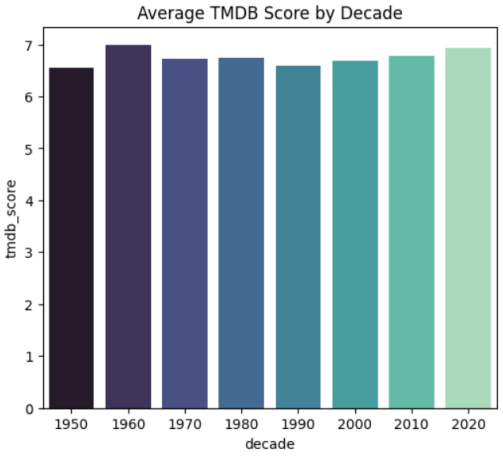


Rysunek 6: Wykres popularności do dekady wydania

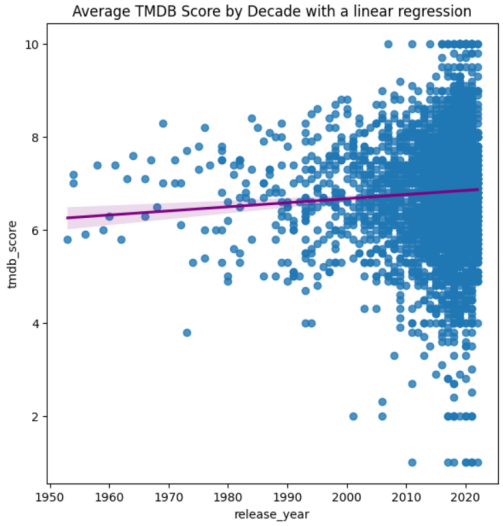


Rysunek 7: Wykres z regresją liniową popularności do roku wydania

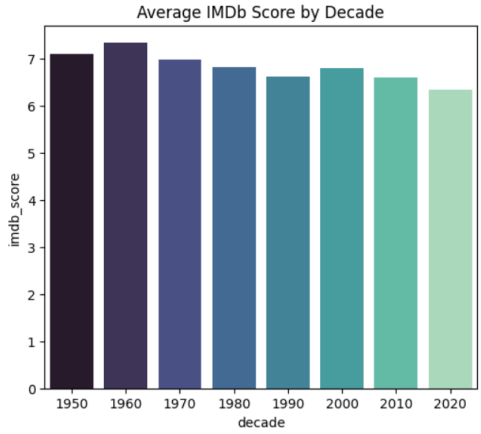
Popularność filmów wykazuje niewielką tendencję wzrostową. Sugeruje to, że w miarę upływu czasu produkcje (zazwyczaj) będą otrzymywać wyższe wyniki popularności. W latach 80. powstały produkcje o najwyższej średniej popularności TMDB. Ze względów takich jak nowe technologie, lepsza jakość oraz ogólnodostępność produkcji nowsze filmy i seriale zdobywają wyższą popularność. Zaskakującą maksymalną wartość dla lat 80 można wytłumaczyć istnieniem kilku fenomenalnych produkcji w tamtym czasie, do których fani posiadają sentyment. Filmami z lat 80 o najwyższej popularności są między innymi: 'Top Gun', 'Wheel of Fortune', 'Seinfeld'.
Dodatkowo, do celów analitycznych na poniższych wykresach także zbadano zależność między oceną produkcji, a dekadą.



(a) Wykres oceny TMDB do dekady wydania



(b) Regresja liniowa oceny TMDB do roku wydania



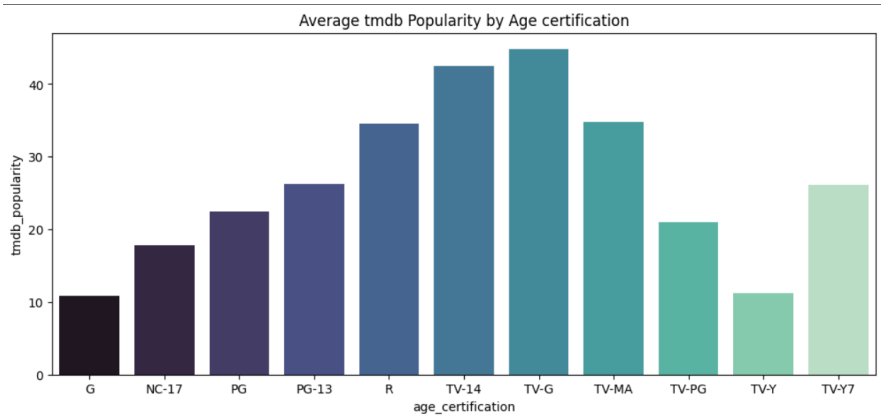
(c) Wykres oceny IMDB do dekady wydania

(d) Regresja liniowa oceny IMDB do roku wydania

Rysunek 8: Wykresy zależności oceny od dekady wydania

Co ciekawe, wnioski z powyższych wykresów wykluczają się, gdyż ocena tmdb sugeruje, iż z czasem wynik produkcji rośnie, a ocena imdb sureguje, iż z czasem wynik produkcji maleje.

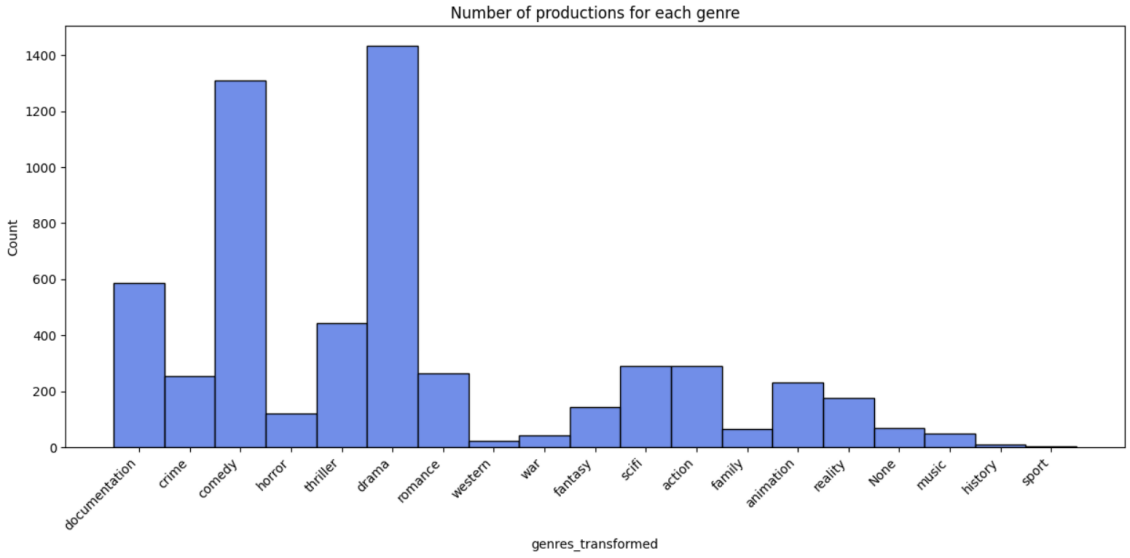
3.3.2 Wpływ ograniczenia wielkowego na popularność



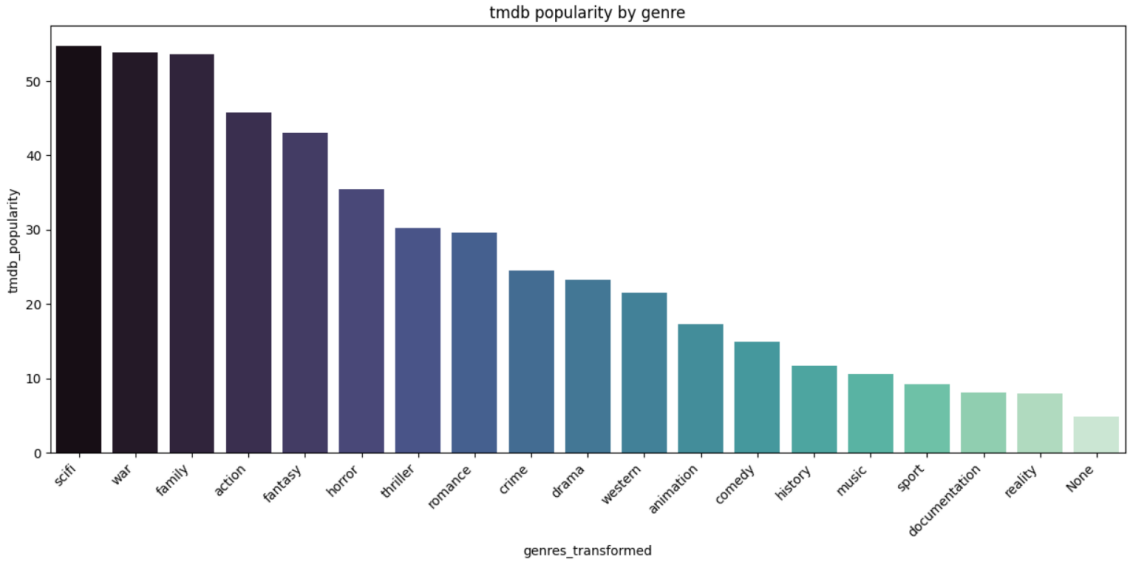
Rysunek 9: Wykres popularności do ograniczenia wiekowego

Największą popularność możemy zaobserwować przy ograniczeniach 'TV-G', 'TV-14', 'R' ORAZ 'TV-MA'. Dla produkcji 'TV-MA' jest to spowodowane tym, iż to ograniczenie umożliwia producentowi większą swobodę twórczą, dzięki czemu może szerzej objąć tematykę filmu. Ten sam powód tłumaczy występienie ograniczenia 'R' wysoko w rankingu. Dla ograniczenia 'TV-14' powodem może być ogólnodostępność dla większości grup wiekowych przez brak silnych restrykcji. Finalnie, najbardziej popularne 'TV-G' występuje na pierwszym miejscu, ponieważ jest to produkcja bez ograniczeń. Oznacza to, że może jej doświadczyć każdy widz, niezależnie od wieku

3.3.3 Wpływ gatunku na popularność



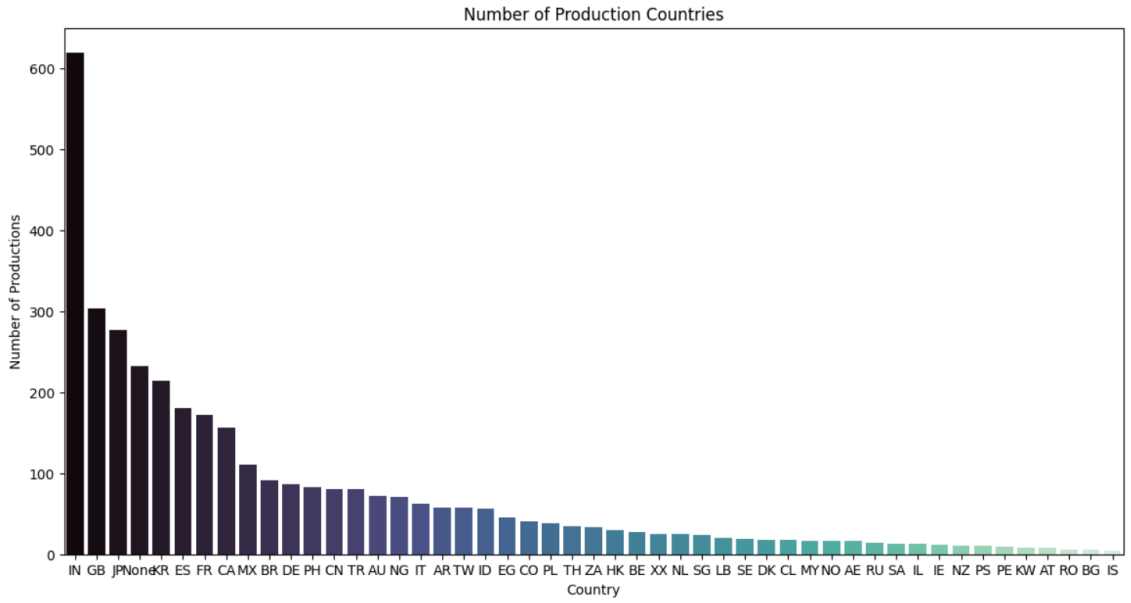
Rysunek 10: Wykres ilości produkcji danego gatunku



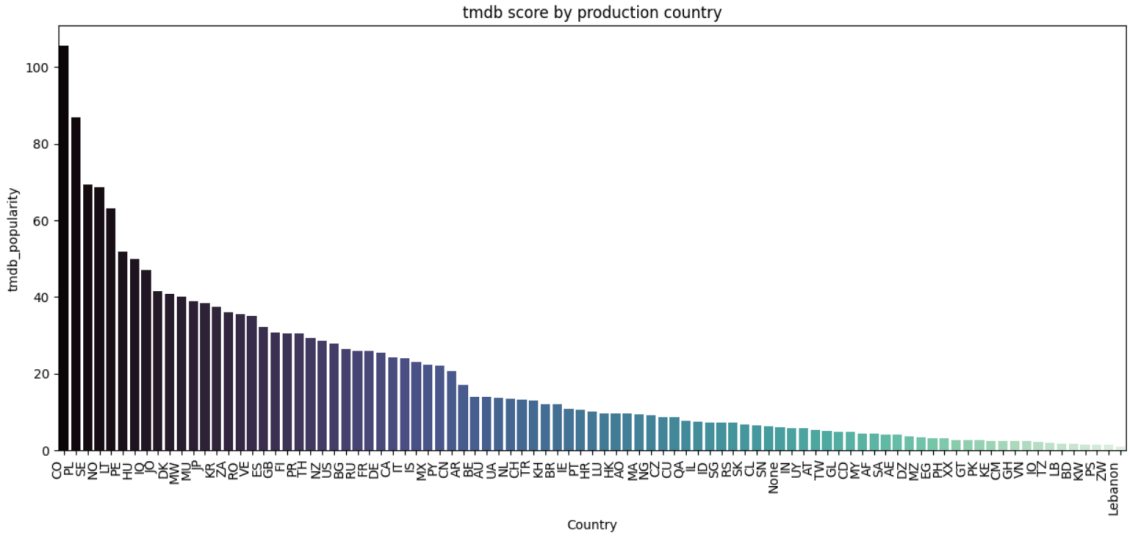
Rysunek 11: Wykres popularności do danego gatunku

Niestety, rozkład liczba różnych gatunków jest daleki od równego, przez co ciężej o wiarygodne wyniki popularności. Mimo to, okazały się być one interesujące i stosunkowo logiczne. Filmy Science Fiction zajęły pierwsze miejsce zapewne ze względu na ich niewyobrażalne efekty specjalne oraz nieszablonowość w kwestiach historii. Kolejne miejsce należy zadziwiająco do filmów dotyczących wojen, jest to najprawdopodobniej spowodowane bardzo specyficzną tematyką przez co takich produkcji jest bardzo mało, więc dane dotyczące popularności mogą się wahać.

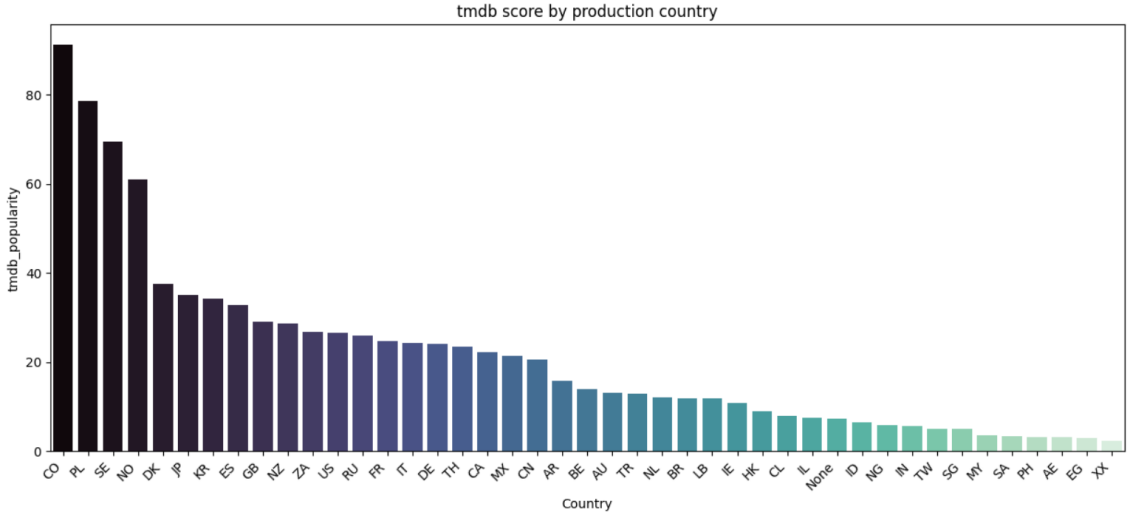
3.3.4 Wpływ kraju produkcji na popularność



Rysunek 12: Wykres liczby filmów dla danego kraju



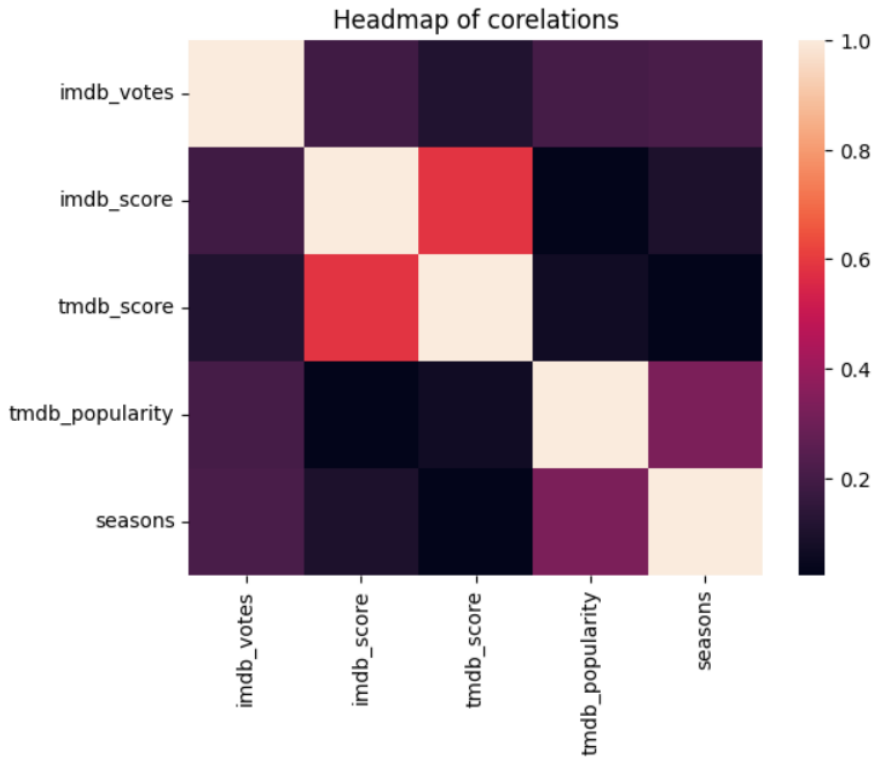
Rysunek 13: Wykres popularności do kraju produkcji



Rysunek 14: Wykres popularności do kraju produkcji, który posiada ich ponad 10

W kwestii rozkładu danych Indie, Wielka Brytania, Japonia, Korea okazały się być potęgami kinematografi, stanowiąc znaczną większość wydanych produkcji.
W kwestii popularności, dla Państw które wydały więcej niż 10 filmów lub seriali najwyższym rozpoznaniem cieszą się filmy i seriale z Kolumbii, Polski oraz Szwecji. Kolumbijskie produkcje cieszące się ogromną popularnością to 'Narcos', 'El Chapo' oraz 'La Reina del Flow'. Dla Polski, narodowymi perłami w dziedzinie kinematografii, które podbiły serca audiencji są: 'Bogdan Boner: The Exorcist', 'Hold Tight' oraz 'The Mire'97'. Dla Szwecji produkcjami tymi są 'The Bridge', 'Love and anarchy', 'Störst av Allt'.

3.3.5 Korelacja między liczbą głosów imdb, oceną imdb, oceną tmdb, liczbą sezonów i popularnością



Rysunek 15: Macież korelacji między parametrami

Powyższy rysunek ukazuje, iż między ocenami, a popularnością występuje bliska zeru korelacja. Liczba sezonów oraz liczba glosów są lepszym wskaźnikami. Występuje stosunkowo wysoka korelacja między sezonami, a popularnością. Jest to spowodowane tym, iż większość ludzi woli seriale z większą liczbą sezonów od tych z mniejszą oraz tym, że kolejne sezony są wypuszczane ze względu na dużą popularność poprzednich produkcji.

4 Model

4.1 Sposób oceniania modeli

Model został oceniony na podstawie trzech metryk:

- Średni bezwzględny błąd procentowy (MAPE) oblicza średnią wartość bezwzględną procentowych różnic między przewidywanymi, a rzeczywistymi wartościami. Niższa wartość MAPE oznacza większą dokładność modelu, a jego minimalna wartość to 0 procent. Im mniejsza różnica procentowa, tym dokładniejszy model.
- Średni błąd bezwzględny (MAE) oblicza średnią wartość bezwzględną różnic między przewidywanymi, a rzeczywistymi wartościami. Niższa wartość MAE wskazuje na większą dokładność modelu. Minimalna możliwa wartość to 0, oznaczająca idealne dopasowanie przewidywań do wartości rzeczywistych.
- Błąd średniokwadratowy to miara obliczająca średnią kwadratów różnic między przewidywanymi, a rzeczywistymi wartościami. Im niższa wartość błędu średniokwadratowego, tym większa dokładność modelu. Minimalna wartość to 0, oznaczająca idealne dopasowanie przewidywań do wartości rzeczywistych.

4.2 Rodzaje modeli

- Random Forest Regressor jest algorytmem uczenia maszynowego wykorzystującym zespół drzew decyzyjnych do przewidywania wartości numerycznych. Każde drzewo w lesie losowego lasu jest trenowane na losowej próbce danych z powtórzeniami, a następnie agreguje przewidywania wszystkich drzew, aby uzyskać końcowe przewidywanie. Ten model jest używany do regresji, czyli przewidywania ciągłych wartości na podstawie cech danych wejściowych.

- Linear Regression to algorytm regresji liniowej, który modeluje zależność liniową między zmienną niezależną a zmienną zależną. Model ten zakłada, że istnieje liniowa relacja pomiędzy zmiennymi, co oznacza, że zmiana jednej zmiennej jest proporcjonalna do zmiany drugiej. Linear Regression szuka linii najlepszego dopasowania do danych, minimalizując błąd kwadratowy między przewidywanymi, a rzeczywistymi wartościami. Przewidywania są oparte na wagach przypisanych do zmiennych wejściowych.

- Epsilon-Support Vector Regression (SVR) jest algorytmem uczenia maszynowego, który wykonuje regresję opartą na wektorach nośnych (support vectors). Ten model ma za zadanie znaleźć optymalną funkcję regresji, która minimalizuje błąd predykcji. SVR może radzić sobie zarówno z liniowymi, jak i nieliniowymi relacjami między zmiennymi. Algorytm poszukuje optymalnej funkcji regresji, która jak najlepiej dopasowuje się do danych wejściowych, minimalizując błąd predykcji i maksymalizując margines błędu epsilon.

4.3 Przygotowanie modeli

Dane dzielone są na testowe i treningowe w stosunku 1:3, co pozostawia 1264 danych testowych oraz 3791 danych treningowych.
Na podstawie manulanych testów różnych parametrów modeli jak i różnych kolumn przekazywanych do treningu modelu najlepsze wyniki osiągnięto dzięki użyciu danych:

- Dekada
- Ograniczenie wiekowe
- Gatunek
- Kraj produckji
- Liczba głosów na forum imdb
- Liczba sezonów
- Czy produkcja jest filmem czy serialem

W modelu Random Forst Regressor testowany były takie parametry jak:

- 'max features' - maksymalna liczba cech (atrybutów) branych pod uwagę przy tworzeniu każdego drzewa.
- 'n estimators' - liczba drzew decyzyjnych (estymatorów) budowanych w lesie losowym

W modelu SVR przetestowano takie parametry jak:

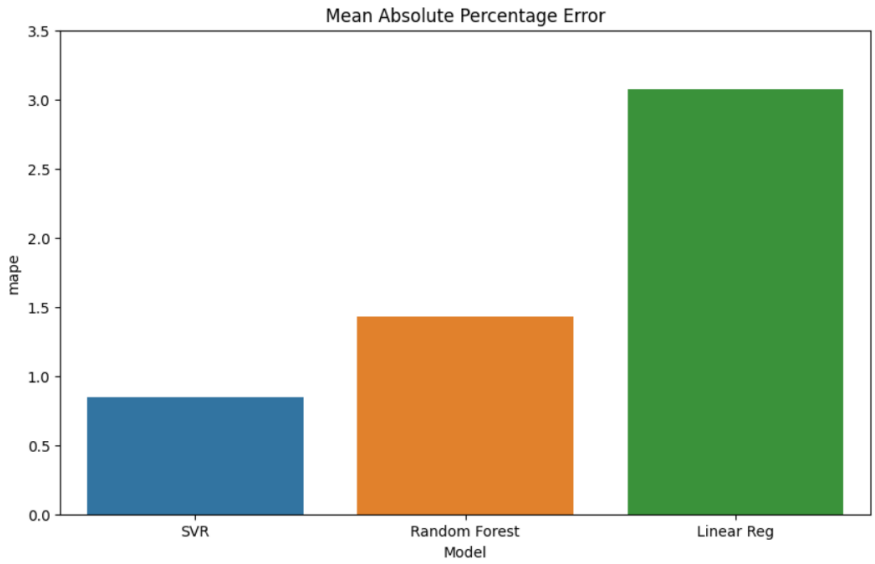
- 'kernel' - określa rodzaj funkcji jądra, która jest używana do transformacji danych wejściowych.
- 'C' - określa jak bardzo błędy są karane w procesie optymalizacji modelu.

Po głębokich analizach wszystkich stworzonych modeli do finalnego porównaniu wzięto po jednym z najlepiej sprawujących się modeli z każdego rodzaju.
Dla 'Linear Regresion' był to model bezparametrowy. Dla 'Random Forest Regressor' był to model posiadający 300 estymatorów oraz 'log2' jako parametr 'max features'. Dla 'SVM' był to model posiadający kernel 'RBF' oraz C równe 2.0.

4.4 Określenie jakości modeli

Model	MAPE	MAE	MSE
Random Forest	1.193	14.929	1691.756
Linear Reg	3.072	20.982	1535.351
SVR	0.849	14.606	1681.436

Tabela 2: Wartości oceny modeli w zależności od przyjętej metryki



Rysunek 16: MAPE poszczególnych modeli

Random Forest:

1. Model Random Forest osiągnął wartość MAPE na poziomie 1.193, co oznacza, że przewidywania różnią się średnio o 1.193 procent od wartości rzeczywistych.
2. Wartość MAE wynosi 14.929, co oznacza, że średnie bezwzględne odchylenie między przewidywanymi a rzeczywistymi wartościami wynosi 14.929.
3. Wartość MSE wynosi 1691.756, co wskazuje na średnią kwadratową różnicę między przewidywanymi a rzeczywistymi wartościami wynoszącą 1691.756.

Linear Regression:

1. Model Linear Regression osiągnął wartość MAPE na poziomie 3.072, co oznacza, że przewidywania różnią się średnio o 3.072 procent od wartości rzeczywistych.
2. Wartość MAE wynosi 20.982, co oznacza, że średnie bezwzględne odchylenie między przewidywanymi a rzeczywistymi wartościami wynosi 20.982.
3. Wartość MSE wynosi 1535.351, co wskazuje na średnią kwadratową różnicę między przewidywanymi a rzeczywistymi wartościami wynoszącą 1535.351.

SVR Support Vector Regression:

1. Model SVR osiągnął wartość MAPE na poziomie 0.849, co oznacza, że przewidywania różnią się średnio o 0.849 procent od wartości rzeczywistych.

2. Wartość MAE wynosi 14.606, co oznacza, że średnie bezwzględne odchylenie między przewidywanymi a rzeczywistymi wartościami wynosi 14.606.

3. Wartość MSE wynosi 1681.436, co wskazuje na średnią kwadratową różnicę między przewidywanymi a rzeczywistymi wartościami wynoszącą 1681.436.

5 Podsumowanie

5.1 Wnioski ogólne

1. Zgodnie z przewidywaniami znaleziona została lekko zależność między niektórymi danymi związanymi z produkcją a jej popularnością.

2. Na podstawie powyższych wyników najlepszym modelem okazał się być SVM patrząc kryterium MAE oraz MAPE.

3. Jeśli wziąć pod uwagę metrykę MSE najlepszy okazał się być model Liniowej Regresji.

4. W przypadku wszystkich modeli, wartości MSE jest stosunkowo wysoka, co sugeruje, że modele mają tendencję do popełniania błędów w przewidywaniu popularności produkcji. Powodem ku temu może być brak rozkładu normalnego wśród danych popularności oraz fakt, iż dane posiadają ekstremalne wartości, które znacząco odbiegają od reszty.

5. Zważywszy na to, iż dane nie zawierają jedynie stosunkowo popularnych produkcji oraz z faktu, iż popularność w zbiorze nie jest oceniana sensowną metryką od 0 do 100, tylko od 0 do 1823 (wartość najwyższa, lecz zapewne może być przekroczona) model można uznać za korzystny.

5.2 Potencjalne usprawnienia

Najważniejszym aspektem, który byłby w stanie ulepszyć model jest zmiana metryki popularności z bardzo rozbieżnej do skondensowanej od 0 do 100. Kolejnym ulepszeniem byłby lepszy rozkład danych w każdej kategorii, a w szczególności w popularności. Najbardziej oczywistą obserwacją w kwestiach ulepszeń jest zwiększenie liczby produkcji w bazie danych oraz uzupełnienie ich wartości brakujących. Finalnym usprawnieniem byłoby zastosowanie innych modeli z lepszymi parametrami, lecz na ten moment wykracza to poza umiejętności twórcy pracy.