Raport z Analizy Danych i Modelowania, Oskar Paciorkowski, s25488

1. Podsumowanie Analizy Danych

Dane pochodzą z zestawu movie_metadata.csv i zawierają informacje o filmach, takie jak budżet, dochód, obsada, oceny i inne cechy. Celem analizy było przewidzenie oceny filmów (imdb_score) na podstawie dostępnych cech.

Wstępna analiza danych:

- Liczba rekordów przed czyszczeniem: 5043
- Liczba rekordów po czyszczeniu: 3756
- Liczba cech po czyszczeniu: 27 (w tym cechy kategoryczne i numeryczne).

Braki w danych:

- Brakujące wartości występowały w wielu kolumnach, takich jak budżet, dochód, i obsada. Usunięto rekordy z brakującymi wartościami, co zmniejszyło zbiór danych.

Rozkłady danych:

- Przeanalizowano rozkłady cech numerycznych za pomocą histogramów.
- Zidentyfikowano kilka potencjalnych wartości odstających, szczególnie w zmiennych finansowych (np. budżet i dochód).

Macierz korelacji:

- Wykres korelacji wskazał, że cechy takie jak budget i gross mają najwyższą korelację z oceną imdb_score.

2. Wyniki Profilowania Danych

Alerts

Dataset has 33 (0.9%) duplicate rows	Duplicates
actor_1_facebook_likes is highly overall correlated with actor_2_facebook_likes and 2 other fields	High correlation
actor_2_facebook_likes is highly overall correlated with actor_1_facebook_likes and 2 other fields	High correlation
actor_3_facebook_likes is highly overall correlated with actor_1_facebook_likes and 2 other fields	High correlation
budget is highly overall correlated with gross and 1 other fields	High correlation
cast_total_facebook_likes is highly overall correlated with actor_1_facebook_likes and 2 other fields	High correlation
country is highly overall correlated with language	High correlation
gross is highly overall correlated with budget and 2 other fields	High correlation
language is highly overall correlated with budget and 1 other fields	High correlation
num_critic_for_reviews is highly overall correlated with num_user_for_reviews and 2 other fields	High correlation
num_user_for_reviews is highly overall correlated with gross and 2 other fields	High correlation
num_voted_users is highly overall correlated with gross and 2 other fields	High correlation
title_year is highly overall correlated with num_critic_for_reviews	High correlation
color is highly imbalanced (79.1%)	Imbalance
language is highly imbalanced (91.7%)	Imbalance
country is highly imbalanced (74.5%)	Imbalance
content_rating is highly imbalanced (50.4%)	Imbalance
actor_1_facebook_likes is highly skewed (γ1 = 20.33840332)	Skewed
budget is highly skewed (γ1 = 44.16873671)	Skewed
director_facebook_likes has 642 (17.1%) zeros	Zeros
facenumber_in_poster has 1581 (42.1%) zeros	Zeros
movie_facebook_likes has 1742 (46.4%) zeros	Zeros

1. Korelacje Między Zmiennymi

- **Wysokie korelacje:**
    - **budget i gross**: Silna korelacja sugeruje, że filmy z większym budżetem generują wyższe przychody.
    - **num_voted_users, num_user_for_reviews, gross**: Liczba głosów użytkowników i liczba recenzji są silnie skorelowane z przychodami, co sugeruje, że popularność filmu wpływa na dochody.
    - **actor_1_facebook_likes, actor_2_facebook_likes, actor_3_facebook_likes**: Korelacje między polubieniami na Facebooku różnych aktorów wskazują na ich wzajemne powiązania w popularności.
    - **title_year i num_critic_for_reviews**: Recenzje krytyków są skorelowane z rokiem produkcji, co może wynikać z rosnącego trendu liczby recenzji w czasie.

2. Nierównomierny Rozkład Zmiennych Kategorycznych

- **Nierównowaga cech:**

- **language**: Ponad 91% wartości w tej zmiennej dotyczy jednego języka (prawdopodobnie angielskiego).
- **color**: 79% filmów jest w kolorze, co powoduje brak równowagi w tej zmiennej.
- **content_rating**: Nierównowaga wskazuje, że większość filmów ma jedną kategorię wiekową.

3. Skośność i Wartości Zerowe

- **Skośność:**
  - Zmienne takie jak **budget** i **gross** wykazują silną skośność, co może wpływać na wyniki modeli predykcyjnych. Zaleca się zastosowanie logarytmicznej transformacji, aby zmniejszyć skośność.
  - **actor_1_facebook_likes** również jest silnie skośny, co sugeruje, że niewielka liczba aktorów ma bardzo wysoką popularność.
- **Wartości zerowe:**
  - **director_facebook_likes**: 17% wartości wynosi 0.
  - **movie_facebook_likes**: Aż 46% filmów nie ma polubień na Facebooku, co wskazuje na znaczną liczbę brakujących informacji w tej zmiennej.

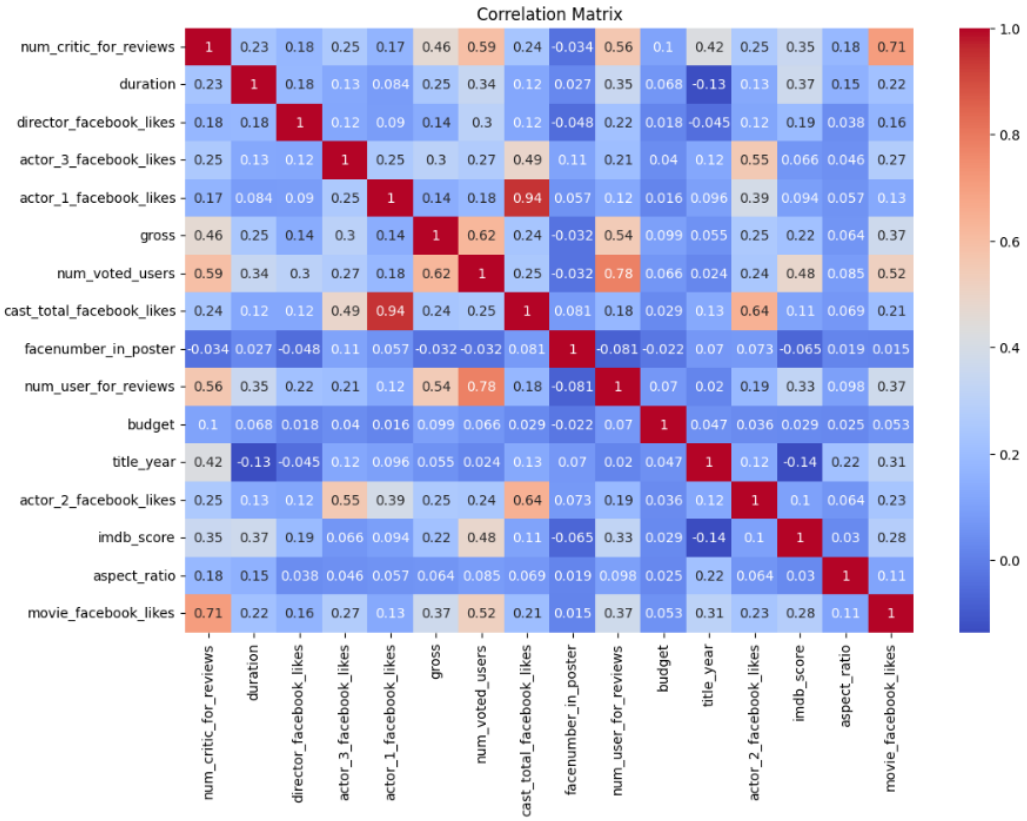
4. Rekomendacje na Podstawie Raportu

- **Usunięcie korelacji:**
  - Wysokie korelacje między zmiennymi, np. **budget i gross**, mogą być usunięte, aby uniknąć multikolinearności w modelu.
- **Transformacja cech:**
  - Zaleca się zastosowanie transformacji logarytmicznej do cech takich jak **budget** i **gross** w celu zmniejszenia ich skośności i poprawy jakości modeli.
- **Zmienne o dużym wpływie:**
  - **num_voted_users** i **duration** wydają się być kluczowymi predyktorami na podstawie analizy rozkładów oraz korelacji.

Raport zapisano jako movies_data_profiling_report.html, który szczegółowo dokumentuje wszystkie wyniki analizy i wizualizacje.

3. Analiza wykresów

Analiza Macierzy Korelacji



Silne Korelacje:

- **budget i gross (0.62):** Wyższy budżet wiąże się z wyższymi przychodami.
- **num_voted_users i gross (0.52):** Popularność filmu koreluje z większym dochodem.
- **cast_total_facebook_likes i actor_1_facebook_likes (0.94):** Gwiazdy dominują w promocji filmów.

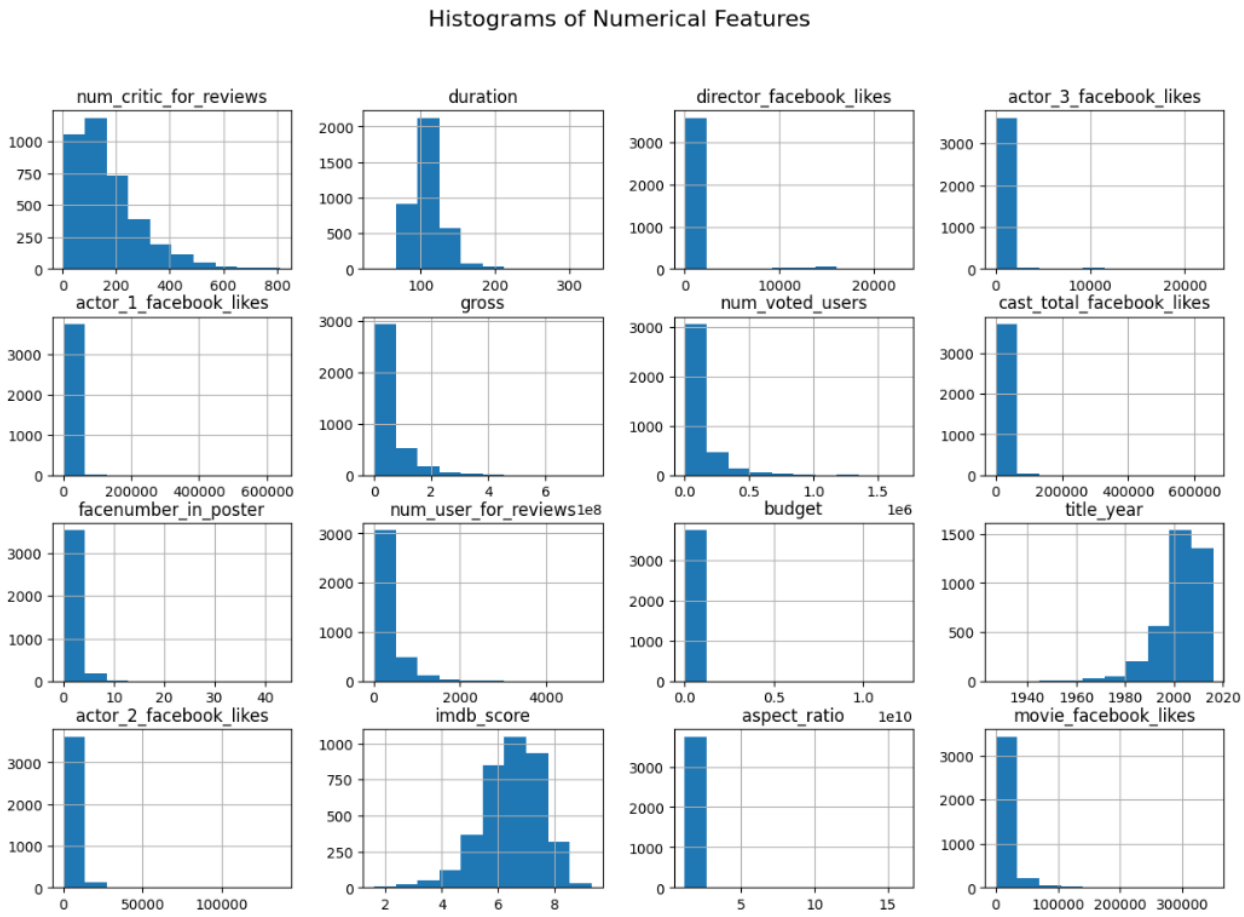
Umiarkowane Korelacje:

- **imdb_score i gross (0.37):** Wyższe oceny IMDb umiarkowanie wpływają na przychody.
- **num_user_for_reviews i num_critic_for_reviews (0.78):** Popularne filmy przyciągają użytkowników i krytyków.

Rekomendacje:

- Usunąć cechy o dużej redundancji, np. cast_total_facebook_likes.
- Skupić się na kluczowych predyktorach: budget, gross, num_voted_users, imdb_score.
- Rozważyć logarytmiczną transformację dla budget i gross.

Histogram



Cechy finansowe (budget, gross):

- Dane są silnie skośne z dużą liczbą filmów o niskim budżecie i niskich dochodach. Tylko nieliczne filmy mają ekstremalnie wysokie wartości.

Popularność (num_voted_users, num_user_for_reviews, movie_facebook_likes):

- Większość filmów ma niską liczbę głosów użytkowników i polubień na Facebooku. Niewiele filmów jest wyjątkowo popularnych.

Czas trwania (duration):

- Większość filmów trwa między 100 a 120 minut, co odpowiada standardowej długości filmu kinowego.

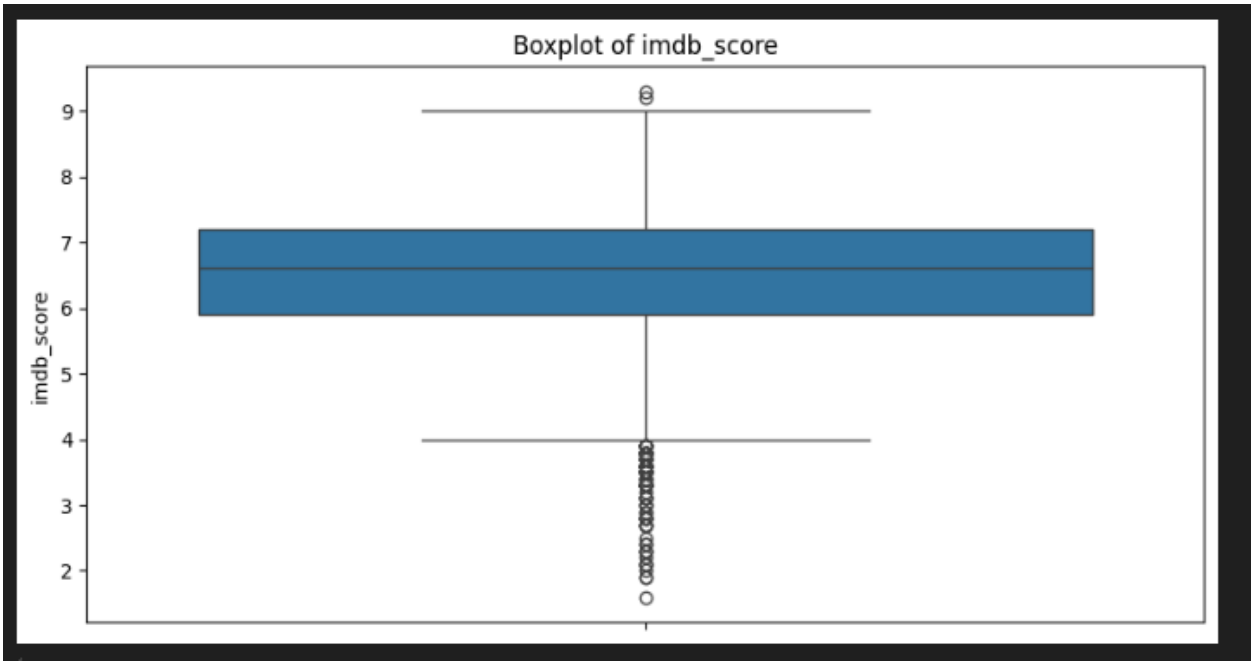
Oceny (imdb_score):

- Rozkład przypomina normalny, z najwyższą liczbą ocen w zakresie 6–8.

Polubienia aktorów (actor_1_facebook_likes, cast_total_facebook_likes):

- Wartości silnie skoncentrowane wokół zera, co sugeruje brak dużej promocji wielu aktorów.

Analiza Boxplotu *imdb_score*



Rozkład: Większość ocen mieści się w przedziale 6–8, z medianą około 7.

Odstające: Filmy z ocenami poniżej 4 i powyżej 9 są wartościami odstającymi.

Wnioski: Dane są skoncentrowane, co może utrudniać predykcję. Warto rozważyć usunięcie odstających.

4. Wyniki Modelowania

Do automatycznej analizy i doboru modeli zastosowano narzędzie **TPOTRegressor**, które wykorzystuje algorytmy genetyczne do optymalizacji pipeline'ów.

Najlepszy pipeline:

DecisionTreeRegressor z VarianceThreshold

- **Pipeline**: `DecisionTreeRegressor(VarianceThreshold(input_matrix, VarianceThreshold__threshold=0.05), DecisionTreeRegressor__max_depth=4, DecisionTreeRegressor__min_samples_leaf=13, DecisionTreeRegressor__min_samples_split=12)`
- **Wynik R^2: 0.69**
- **Opis**: Model drzewa decyzyjnego z preprocesingiem za pomocą VarianceThreshold, który usuwa cechy o niskiej wariancji. Prosty model o ograniczonej głębokości drzewa.

Pozostałe modele z najwyższymi wynikami:

1. **DecisionTreeRegressor bez dodatkowego preprocessing**
    - Wynik R^2: 0.65
    - Pipeline: Skalowanie cech + RandomForest.
    - **Opis**: Drzewo decyzyjne o maksymalnej głębokości 8. Model bez dodatkowego preprocessingu, który radzi sobie dobrze w danych o niskim poziomie nieliniowości.
2. **RidgeCV (Regresja grzbietowa)**
    - Wynik R^2: 0.62
    - O**pis**: Model regresji grzbietowej, który stosuje regularyzację w celu zmniejszenia nadmiernego dopasowania. Prostota modelu może być ograniczeniem w przypadku złożonych zależności.4. Wnioski i Dalsze Kroki

- W przyszłych iteracjach można rozważyć:
    - Usunięcie wartości odstających, które mogą wpływać na wyniki modelu.
    - Dodanie nowych cech, takich jak popularność reżysera lub gatunek filmu.

Model zapisano w pliku tpot_optimized_model.py, a logi analizy pipeline'ów w pliku tpot_log.txt.

TPOT trwał bardzo długo, ale dał mi informację na jakich modelach powinienem się skupić.

5. Plany na Przyszłość

Optymalizacja Danych:

- Usunięcie wartości odstających i transformacja cech skośnych (budget, gross) dla poprawy jakości modeli.
- Zrównoważenie zmiennych kategorycznych (language, content_rating).

Rozwój Modelu XGBoost:

- Implementacja modelu **XGBoost** dla lepszego uchwycenia nieliniowych zależności w danych.
- Przeprowadzenie optymalizacji hiperparametrów (np. learning_rate, max_depth, n_estimators) w celu poprawy wyników.
- Analiza ważności cech (feature importance) w XGBoost w celu identyfikacji kluczowych predyktorów.

Ulepszenie DecisionTreeRegressor:

- Eksperymentowanie z większą głębokością drzewa i minimalną liczbą próbek w liściu, aby zwiększyć zdolność modelu do uchwycenia złożonych zależności.
- Rozszerzenie pipeline'u o metody preprocessingu, takie jak logarytmiczna transformacja cech skośnych (budget, gross).
- Porównanie wydajności z innymi algorytmami drzewiastymi, takimi jak Random Forest lub Gradient Boosting.

Porównanie i Wdrożenie:

- Porównanie wyników modeli **XGBoost** i **DecisionTreeRegressor** na podstawie metryk takich jak R, MAE, i RMSE.
- Wdrożenie najlepszego modelu do przewidywania ocen IMDB oraz analiza jego wyników w zastosowaniu rzeczywistym.