

Anime recommendation prediction

Jakub Kaliński s24636, Szymon Kaliński s24637

Spis treści

| | |
|---|----|
| Opis problemu | 3 |
| Dane | 4 |
| Analizy: | 11 |
| Sposób rozwiązania problemu | 14 |
| Dyskusja wyników i ewaluacja modelu | 16 |
| Podsumowanie | 18 |

Opis problemu

W branży anime istnieje wiele wyzwań związanych z przewidywaniem popularności nowych serii, identyfikacją czynników wpływających na oceny użytkowników oraz optymalizacją rekomendacji dla widzów. Problemy te wpływają na decyzje dotyczące produkcji, marketingu oraz dystrybucji. Przykłady pytań, na które można odpowiedzieć dzięki analizie danych o anime to:

- Jakiego gatunku anime są najpopularniejsze?
- Jakie cechy (np. długość serii, gatunek, typ) mają największy wpływ na oceny użytkowników?
- Jakiego trendu można zaobserwować w preferencjach widzów na przestrzeni lat?

Model analizy popularności i oceny anime może być szczególnie użyteczny dla:

- **Producentów anime:** Pomaga w podejmowaniu decyzji dotyczących wyboru gatunków, długości serii oraz studiów produkcyjnych.
- **Działów marketingu:** Mogą lepiej targetować kampanie reklamowe na podstawie preferencji widzów.
- **Analityków danych:** Dostarcza narzędzia do identyfikacji trendów i wzorców w ocenach anime, co pozwala na optymalizację strategii biznesowej.

- **Platform streamingowych:** Pomaga w rekomendowaniu treści użytkownikom na podstawie ich wcześniejszych ocen i preferencji.

Problem analizy popularności i oceny anime jest interesujący ze względu na jego praktyczne zastosowanie oraz potencjał do generowania wymiernych korzyści biznesowych. Analiza dużych zbiorów danych w kontekście branży anime pozwala na lepsze zrozumienie preferencji widzów, co jest kluczowe w obecnym konkurencyjnym środowisku medialnym. Ponadto, zastosowanie technik analizy danych i uczenia maszynowego do rzeczywistych problemów biznesowych jest fascynujące z punktu widzenia rozwoju technologii i innowacji w zarządzaniu.

Dane

Dane użyte w projekcie pochodzą ze strony internetowej kaggle.com. Umieściliśmy je w pliku anime.csv, który zawiera szczegółowe informacje na temat anime ze strony myanimelist.net. Dane te obejmują takie zmienne jak:

- anime_id - unikalny identyfikator anime w serwisie myanimelist.net.
- name - pełna nazwa anime.
- genre - lista gatunków tego anime, rozdzielona przecinkami.
- type - rodzaj anime, np. film, TV, OVA itp.

- episodes - liczba odcinków w tej serii. (1, jeśli jest to film).
- rating - średnia ocena dla tego anime w skali od 1 do 10.

Dane są wiarygodne, ponieważ pochodzą bezpośrednio z bazy danych strony myanimelist.net, co gwarantuje ich autentyczność i dokładność. Strona ta jest jedną z najpopularniejszych stron do recenzowania anime, co daje nam również ogromną liczbę danych. Niestety w zmiennej episodes występują brakujące dane, ze względu na nieznaną liczbę odcinków. Jest to spowodowane tym, że niektóre serie nie są jeszcze skończone, a regularnie emitowane. Jest to jednak niewielka liczba przy tak dużym zbiorze danych. Dodatkowo baza danych została przez nas rozszeżona o zmienne typu boolean, aby można było zliczyć ilość wystąpień każdego gatunku. W zmiennej genre, niektóre serie mają wprowadzone więcej niż jeden gatunek, co powoduje, że ocena, który gatunek jest najpopularniejszy staje się trudne do oceny. Dzięki stworzeniu dodatkowych kolumn typu boolean dla każdego gatunku, bardzo ułatwia tą diagnozę. Dlatego ostatecznie lista wszystkich naszych kolumn w pliku anime.csv wygląda następująco:

- anime_id (bigint)
- Name (string)
- Genre (string)
- Yuri (boolean)

- Yaoi (boolean)
- Vampire (boolean)
- Thriller (boolean)
- Supernatural (boolean)
- Super Power (boolean)
- Sports (boolean)
- Space (boolean)
- Slice of Life (boolean)
- Shounen Ai (boolean)
- Shounen (boolean)
- Shoujo Ai (boolean)
- Shoujo (boolean)
- Seinen (boolean)
- Sci-Fi (boolean)
- School (boolean)
- Samurai (boolean)
- Romance (boolean)
- Psychological (boolean)
- Police (boolean)
- Parody (boolean)
- Mystery (boolean)
- Music (boolean)
- Military (boolean)
- Mecha (boolean)

- Martial Arts (boolean)
- Magic (boolean)
- Kids (boolean)
- Josei (boolean)
- Horror (boolean)
- Historical (boolean)
- Hentai (boolean)
- Harem (boolean)
- Game (boolean)
- Fantasy (boolean)
- Ecchi (boolean)
- Drama (boolean)
- Demons (boolean)
- Dementia (boolean)
- Comedy (boolean)
- Cars (boolean)
- Adventure (boolean)
- Action (boolean)
- Type (string)
- Episodes (string)
- Rating (double)
- Members (bigint)

Szczegółowa ilość wystąpień wszystkich gatunków:

- **Comedy:** true 38% - 4645, false 62% - 7649
- **Action:** true 23% - 2845, false 77% - 9449
- **Adventure:** true 19% - 2348, false 81% - 9946
- **Fantasy:** true 19% - 2309, false 81% - 9985
- **Sci-Fi:** true 17% - 2070, false 83% - 10224
- **Drama:** true 16% - 2016, false 84% - 10278
- **Shounen:** true 14% - 1776, false 86% - 10518
- **Kids:** true 13% - 1609, false 87% - 10685
- **Romance:** true 12% - 1464, false 88% - 10830
- **Slice of Life:** true 10% - 1220, false 90% - 11074
- **School:** true 10% - 1220, false 90% - 11074
- **Hentai:** true 9% - 1141, false 91% - 11153
- **Supernatural:** true 8% - 1037, false 92% - 11257
- **Mecha:** true 8% - 944, false 92% - 11350
- **Music:** true 7% - 860, false 93% - 11434
- **Historical:** true 7% - 806, false 93% - 11488
- **Magic:** true 6% - 778, false 94% - 11516
- **Shoujo:** true 5% - 652, false 95% - 11642
- **Ecchi:** true 5% - 637, false 95% - 11657
- **Seinen:** true 4% - 547, false 96% - 11747
- **Sports:** true 4% - 543, false 96% - 11751
- **Super Power:** true 4% - 465, false 96% - 11829
- **Mystery:** true 4% - 495, false 96% - 11799
- **Space:** true 3% - 381, false 97% - 11913
- **Parody:** true 3% - 408, false 97% - 11886
- **Military:** true 3% - 426, false 97% - 11868

- **Harem:** true 3% - 317, false 97% - 11977
- **Horror:** true 3% - 369, false 97% - 11925
- **Psychological:** true 2% - 229, false 98% - 12065
- **Martial Arts:** true 2% - 265, false 98% - 12029
- **Demons:** true 2% - 294, false 98% - 12000
- **Dementia:** true 2% - 240, false 98% - 12054
- **Police:** true 2% - 197, false 98% - 12097
- **Game:** true 1% - 181, false 99% - 12113
- **Samurai:** true 1% - 148, false 99% - 12146
- **Thriller:** true 1% - 87, false 99% - 12207
- **Vampire:** true 1% - 102, false 99% - 12192
- **Shounen Ai:** true 1% - 65, false 99% - 12229
- **Cars:** true 1% - 72, false 99% - 12222
- **Shoujo Ai:** true 0% - 55, false 100% - 12239
- **Josei:** true 0% - 54, false 100% - 12240
- **Yaoi:** true 0% - 39, false 100% - 12255
- **Yuri:** true 0% - 42, false 100% - 12252

Gatunki zostały wpisane w kolejności od największej liczby występowania do najmniejszej. Możemy zauważyć, że nawet najczęściej występujący gatunek pojawia się tylko w 38% przypadków. Pokazuje to jak bardzo zróżnicowany jest rynek anime.

Szczegółowa statystyka type(rodzaj anime):

| Count | % | Cum.% |
|-------|---|----------------|
| • TV | - | 3787 30.8 30.8 |

| | | | | |
|------------|---|------|------|-------|
| • OVA | - | 3311 | 26.9 | 57.7 |
| • Movie | - | 2348 | 19.1 | 76.8 |
| • Special | - | 1676 | 13.6 | 90.5 |
| • ONA | - | 659 | 5.4 | 95.8 |
| • Music | - | 488 | 4.0 | 99.8 |
| • No value | - | 25 | 0.2 | 100.0 |

Szczegółowa statystyka episodes(ilości odcinków) :

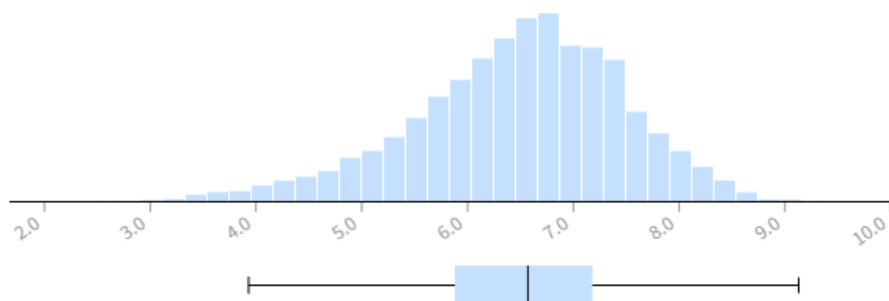
| | | |
|--------------------------|---|--------|
| • Min | - | 1 |
| • Max | - | 1818 |
| • Średnia | - | 12.383 |
| • Mediana | - | 2 |
| • Odchylenie standardowe | - | 46.865 |

Szczegółowa statystyka rating(oceny) :

| | | |
|--------------------------|---|--------|
| • Min | - | 1.6700 |
| • Max | - | 10 |
| • Średnia | - | 6.4739 |
| • Mediana | - | 6.5700 |
| • Odchylenie standardowe | - | 1.0267 |

SUMMARY

| | | |
|---------|--------|---------|
| Valid | 12,294 | 100.0 % |
| Hapax | 65 | 0.5 % |
| Invalid | 0 | 0.0 % |
| Empty | 230 | 1.9 % |



Szczegółowa statystyka members(ilości ogladających) :

| | | |
|-------|---|---|
| • Min | - | 5 |
|-------|---|---|

- Max - 1013917
- Średnia - 18071
- Mediana - 1550
- Odchylenie standardowe - 54821

Analizy:

Popularność Gatunków

| Gatunek | Średnia ocena | Suma oglądających (zaokrąglone) |
|---------------|---------------|---------------------------------|
| Yuri | 6.089 | 187k |
| Yaoi | 6.434 | 543k |
| Vampire | 6.47 | 6.4M |
| Thriller | 6.467 | 10.1M |
| Supernatural | 6.426 | 57M |
| Super Power | 6.457 | 22.8M |
| Sports | 6.46 | 7.4M |
| Space | 6.467 | 5M |
| Slice of Life | 6.445 | 32M |
| Shounen Ai | 6.471 | 1.4M |
| Shounen | 6.376 | 55.5M |
| Shoujo Ai | 6.472 | 1.9M |
| Shoujo | 6.451 | 15.3M |
| Seinen | 6.451 | 23.2M |
| Sci-Fi | 6.428 | 44.9M |
| School | 6.42 | 55M |
| Samurai | 6.469 | 3.5M |
| Romance | 6.403 | 65.7M |
| Psychological | 6.464 | 17.2M |
| Police | 6.463 | 4.9M |
| Parody | 6.473 | 8M |
| Mystery | 6.442 | 26.9M |
| Music | 5.923 | 6M |
| Military | 6.455 | 12.2M |

| | | |
|--------------|-------|-------|
| Mecha | 6.46 | 15M |
| Martial Arts | 6.466 | 7.1M |
| Magic | 6.452 | 21.2M |
| Kids | 6.113 | 3M |
| Josei | 6.47 | 2.2M |
| Horror | 6.427 | 12.6M |
| Historical | 6.455 | 11.4M |
| Hentai | 6.181 | 3M |
| Harem | 6.461 | 19.9M |
| Game | 6.469 | 6.8M |
| Fantasy | 6.428 | 57.3M |
| Ecchi | 6.463 | 25M |
| Drama | 6.375 | 58.3M |
| Demons | 6.469 | 9.5M |
| Dementia | 5.012 | 1.8M |
| Comedy | 6.383 | 107M |
| Cars | 6.469 | 514k |
| Adventure | 6.41 | 45.5M |
| Action | 6.381 | 95.5M |

k - tysiąc

M - milion

Jak te dane mogą pomóc rozwiązać problem?

Lepsze zarządzanie produkcją:

Dzięki analizie danych możemy dokładnie zidentyfikować, które gatunki anime są najpopularniejsze i jakie cechy mają wpływ na wysokie oceny użytkowników. Informacje te pozwalają na optymalne zarządzanie produkcją, co zapobiega produkcji mało popularnych serii. Na przykład, z danych wynika, że gatunek "Comedy" jest najczęściej oglądany, co sugeruje konieczność dodawania humorystycznych postaci, dialogów czy scen do tworzonej serii.

Optymalizacja rekomendacji:

Analiza ocen użytkowników pokazuje, które cechy anime są najczęściej oceniane pozytywnie. Informacje te mogą pomóc w decyzjach dotyczących rekomendacji określonych serii dla różnych użytkowników.

Poprawa satysfakcji widzów:

Dane dotyczące ocen użytkowników pozwalają na monitorowanie satysfakcji widzów. Możemy zidentyfikować, które cechy anime są wysoko oceniane, a które wymagają poprawy.

Wsparcie w podejmowaniu decyzji strategicznych:

Dane o anime mogą wspierać producentów w podejmowaniu kluczowych decyzji strategicznych, takich jak wprowadzanie nowych serii, ustalanie budżetów oraz planowanie kampanii marketingowych. Dokładne dane na temat popularności i ocen użytkowników umożliwiają podejmowanie decyzji opartych na twardych danych, co zmniejsza ryzyko i zwiększa szanse na sukces.

Sposób rozwiązania problemu

Do ewaluacji ratingu anime wybraliśmy model Random Forest.

Osiągnął on najlepsze wyniki spośród wszystkich modeli, ponieważ jego algorytm oparty na drzewach decyzyjnych radzi sobie dobrze z dużą ilością cech, które pojawiły się w naszym projekcie.

Etapy realizacji projektu

Zebranie danych: W naszym projekcie skorzystaliśmy z danych udostępnionych na stronie internetowej [kaggle.com](https://www.kaggle.com). Interesujący nas zbiór danych zawierał Nazwę, Gatunki, Typ, Ilość odcinków, Rating i Liczbę oceniających co uznaliśmy za wystarczające do stworzenia modelu.

Analiza danych: Kolejnym etapem było przeanalizowanie naszych danych. W trakcie analizy zrobiliśmy prototyp modelu, aby określić, które kategorie najbardziej wpływają na jego jakość. Na tym etapie zorientowaliśmy się też, że nasze dane będziemy musieli zmodyfikować, ponieważ sposób zapisu w nich Gatunków, nie pozwalał dostatecznie dobrze przewidywać wyniki. Chcieliśmy też sprawdzić jaki realnie wpływ na Rating mają właśnie gatunki.

Przygotowanie danych: Dane z pliku anime.csv zmodyfikowaliśmy. Sporządziliśmy listę wszystkich Gatunków, które pojawiały się w naszych danych i stworzyliśmy dla nich kategorie typu boolean, żeby otrzymywać jasną informację do których gatunków każde anime należy.

Trenowanie modeli: Po przygotowaniu danych wróciliśmy do testowania różnych modeli. Sprawdzaliśmy też jak na ewaluację wpływa wyłączenie różnych kategorii.

Wybór modelu: Z testów wynikało, że najlepiej radził sobie Random Forest, więc jego ostatecznie wybraliśmy do ewaluacji.

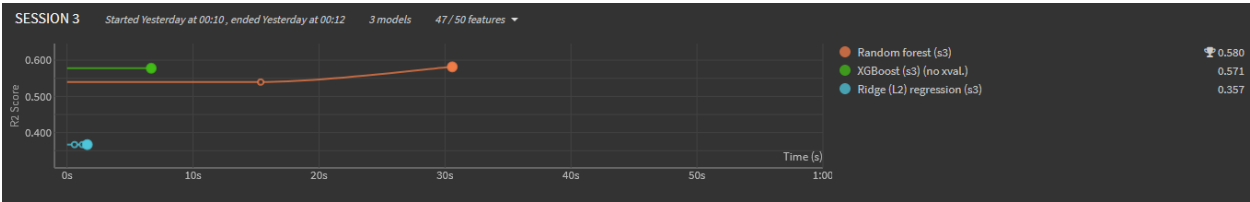
Ewaluacja modelu: Przeprowadziliśmy ostateczną ewaluację modelu na danych, aby upewnić się co do jego skuteczności, która okazała się dla nas satysfakcjonująca.

Miary ewaluacji modelu

Do oceny jakości modelu użyliśmy R^2 Score. R-squared określa, jak dobrze model wyjaśnia zmienność danych. Wyższa wartość R^2 wskazuje na lepsze dopasowanie modelu do danych. Dzięki zastosowaniu tej miary mogliśmy ocenić, jak skuteczny jest model Random Forest w przewidywaniu ratingów.

Dyskusja wyników i ewaluacja modelu

Wyniki modelowania






| | | | | |
|----------------------------|-----|------------|--|-----|
| Random forest (s3) | | 🏆 0.580 | ✓ Done 1 day ago (2024-06-21 00:12:25) | ☆ ⋮ |
| Most important features | | | | |
| Trees | 100 | members | | |
| Depth | 17 | type | | |
| Min samples | 10 | episodes | | |
| Hyperparameter search size | 2 | Drama | | |
| | | Shounen | | |
| | | Hentai | | |
| | | Train set | 9813 rows | |
| | | Test set | 2481 rows | |
| | | Train time | 2 minutes and 43 seconds | |

| | | | |
|--------------------------|----------------------------|---------|---|
| <input type="checkbox"/> | SESSION 3 | | |
| <input type="checkbox"/> | Random forest (s3) | 🏆 0.580 | ☆ |
| <input type="checkbox"/> | Ridge (L2) regression (s3) | 0.357 | ☆ |
| <input type="checkbox"/> | XGBoost (s3) | 0.571 | ☆ |

| rating | members | prediction |
|---------|---------|--------------------|
| double | bigint | float |
| Decimal | Integer | Decimal |
| | | |
| 9.37 | 200630 | 8.379058270016452 |
| 9.26 | 793665 | 8.388004599953147 |
| 9.25 | 114262 | 7.9702634816967315 |
| 9.17 | 673572 | 8.301258935583752 |
| 9.16 | 151266 | 8.08842579005209 |
| 9.15 | 93351 | 7.998923132432457 |
| 9.13 | 425855 | 8.353410433039727 |
| 9.11 | 80679 | 8.029251347392778 |
| 9.1 | 72534 | 7.851071427919368 |
| 9.11 | 81109 | 7.8792708041096375 |
| 9.06 | 456749 | 8.37017506868406 |
| 9.05 | 102733 | 8.048453916543304 |
| 9.04 | 336376 | 8.297652635304507 |
| 8.98 | 572888 | 8.372723218480608 |
| 8.93 | 179342 | 8.11917360949042 |
| 8.93 | 466254 | 8.484245078185428 |
| 8.92 | 416397 | 8.396114266310562 |
| 8.88 | 75894 | 7.974145108617267 |
| 8.84 | 226193 | 8.34192332245518 |
| 8.83 | 715151 | 8.289237973729625 |
| 8.83 | 157670 | 8.1877665220732 |
| 8.83 | 129307 | 8.098085959918365 |
| 8.82 | 486824 | 8.375953917339432 |
| 8.82 | 552458 | 8.118939395164437 |

Ewaluacja modelu

Według wyników ewaluacji wynikło, że najbardziej na rating wpływa ilość oceniających. Testowaliśmy też włączanie i wyłączanie różnych kategorii, ale nie otrzymaliśmy lepszych wyników niż 0.58. W pewnym momencie sprawdziliśmy, jak zmienia się wyniki po wyłączeniu Members i jakość znacznie się pogorszyła. Przeprowadziliśmy też przykładową predykcję wyników.

| | | |
|---|---|---|
| <input type="checkbox"/> SESSION 4 | | |
| <input type="checkbox"/>  Random forest (s4) |  0.352 |  |
| <input type="checkbox"/>  Ridge (L2) regression (s4) | 0.313 |  |
| <input type="checkbox"/>  XGBoost (s4) | 0.351 |  |

Podsumowanie

Celem projektu było przewidzenie ratingu seriali anime na podstawie danych w pliku anime.csv ze strony internetowej kaggle.com. Chcieliśmy zidentyfikować cechy najbardziej wpływających na oceny użytkowników.

Dzięki zastosowaniu algorytmu Random Forest i wysokich wyników ewaluacji udało się osiągnąć zadowalające wyniki predykcji. Proces realizacji projektu obejmował analizę, przygotowanie i modyfikację danych, trenowanie różnych modeli oraz ewaluację najlepszego z nich.

Z naszego modelu mogliby skorzystać producenci anime, działy marketingu oraz platformy streamingowe.

Producenci mogliby lepiej dostosować swoje decyzje dotyczące produkcji nowych serii, a działy marketingu efektywniej prowadzić kampanie reklamowe, a platformy streamingowe trafniej rekomendować treści oglądającym.

Pomimo satysfakcjonujących nas wyników, jesteśmy świadomi ograniczeń modelu Random Forest. Największym z nich jest wpływ ilości oceniających (members) na rating. Fakt, że model tak bardzo polega na niej może prowadzić do zaburzenia wyników w niektórych przypadkach. Ponadto, specyfika danych oraz ich ograniczona reprezentatywność mogą wpływać na ogólną skuteczność modelu w różnych kontekstach.