Моделирование и анализ данных
2023. Том 13. № 2. С. 77–84
DOI: https://doi.org/10.17759/mda.2023130204
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)


Modelling and Data Analysis
2023. Vol. 13, no. 2, pp. 77–84
DOI: https://doi.org/10.17759/mda.2023130204
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)


УДК 519.246.8; 37.018.432

# Прогнозирование рейтинга нового фильма по его метаданным

***Сологуб Г.Б.****
Московский авиационный институт
(национальный исследовательский университет) (МАИ)
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-5657-4826
e-mail: glebsologub@ya.ru

***Сазон Н.С.*****
Московский авиационный институт
(национальный исследовательский университет) (МАИ)
г. Москва, Российская Федерация
ORCID: https://orcid.org/0000-0002-9816-4585
e-mail: nikitaS1598@gmail.com

В статье описан подход к прогнозированию рейтинга нового фильма на основе данных, известных до его выхода, с использованием моделей классического машинного обучения. Подход включает в себя тестирование различных моделей с соответствующей предобработкой данных и подбором оптимальных гиперпараметров, а также выбор наилучшего алгоритма с точки зрения выбранного функционала качества.

***Ключевые слова:*** машинное обучение, пользовательская оценка фильма.

**Для цитаты:**
*Сологуб Г.Б., Сазон Н.С.* Прогнозирование рейтинга нового фильма по его метаданным // Моделирование и анализ данных. 2023. Том 13. № 2. С. 77–84. DOI: https://doi.org/10.17759/mda.2023130204

***Сологуб Глеб Борисович,** кандидат физико-математических наук, доцент кафедры математической кибернетики института «Компьютерные науки и прикладная математика», Московский авиационный институт (национальный исследовательский университет) (МАИ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000–0002–5657–4826, e-mail: glebsologub@ya.ru

****Сазон Никита Сергеевич,** студент магистратуры института «Компьютерные науки и прикладная математика», Московский авиационный институт (национальный исследовательский университет) (МАИ), г. Москва, Российская Федерация, ORCID: https://orcid.org/0000-0002-9816-4585, e-mail: nikitaS1598@gmail.com

# 1. ВВЕДЕНИЕ

В быстро развивающемся мире технологий построение прогнозов на основе данных с использованием методов машинного обучения приобретает всё большую актуальность. Современных мощностей хватает для того, чтобы эффективно обрабатывать большие объёмы данных, автоматизировать процессы и делегировать машинам задачи, которые ещё не так давно могли быть решены только человеком.

В данной работе рассматривается задача предсказания рейтинга нового фильма по данным (жанр, длительность и т.д.), известным ещё до его выхода. Такая оценка может быть полезной для различных групп людей и организаций. Например, для инвесторов, которые хотят вложить деньги в производство фильма; для студий и кинокомпаний, разрабатывающих фильмы; для режиссёров, сценаристов и актёров, присматривающих проекты для участия. Не говоря уже о миллионах людей, не связанных с киноиндустрией, но покупающих билеты в кинотеатры, основываясь лишь на собственных ожиданиях, так как реальный рейтинг ещё не успел сформироваться. Иными словами, это мощный фактор, который может иметь вес при принятии очень разнообразных решений.

Задача рассматривается в большом количестве публикаций, многие из которых вышли относительно недавно. Например, в [1] описано исследование, в котором предлагается строить прогноз с помощью линейных и метрических методов машинного обучения, а также классической полносвязной нейросети. В [2] помимо основной задачи прогнозирования кассовых сборов фильмов рассматривается задача прогнозирования пользовательской оценки, которая решается методами Random forest, gradient boosting и k nearest neighbors. Таким образом, можно сделать вывод, что настоящее время задача является достаточно актуальной.

# 2. ПОСТАНОВКА ЗАДАЧИ

В качестве источника данных была выбрана база данных IMDB [3]. Выбор обусловлен тем, что это крупнейшая в мире БД о кинематографе, и рейтинги, представленные в ней, пользуются авторитетом во всём мире. На её основе был сформирован файл со следующими характеристиками, описывающими кинокартину:

- titleType – тип/формат (например, фильм, короткометражка, сериал, видео и т. д.);
- isAdult – 0: не для взрослых; 1: для взрослых;
- startYear – год выпуска;
- runtimeMinutes – время в минутах;
- genres – список жанров, поставленных в соответствие фильму;
- directors – id режиссёра;
- (director) primaryProfession – список профессий режисёра;
- (director) knownForTitles – список фильмов, за которые режиссёр известен;
- writer – id сценариста;
- average rating – средневзвешенный рейтинг.

Также в датасете присутствует столбец tconst, содержащий уникальные идентификаторы фильмов.

| | tconst | averageRating | titleType | isAdult | startYear | runtimeMinutes | genres | directors |
|---|---|---|---|---|---|---|---|---|
| **8** | tt0000009 | 5.3 | movie | 0 | 1894 | 45 | Romance | nm0085156 |
| **34** | tt0000036 | 4.4 | short | 0 | 1896 | 0 | Drama,Short | nm0005690 |
| **74** | tt0000076 | 4.5 | short | 0 | 1896 | 1 | Drama,Short | nm0005690 |
| **89** | tt0000091 | 6.7 | short | 0 | 1896 | 3 | Horror,Short | nm0617588 |
| **106** | tt0000108 | 4.4 | short | 0 | 1896 | 1 | Drama,Short | nm0005690 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **1247779** | tt9916544 | 6.9 | short | 0 | 2019 | 12 | Drama,Short | nm3219235 |
| **1247781** | tt9916578 | 7.4 | tvEpisode | 0 | 2019 | 44 | Adventure,Biography,Comedy | nm0373673 |
| **1247782** | tt9916580 | 8.5 | tvEpisode | 0 | 2012 | 10 | Adventure,Animation,Comedy | nm0996406 |
| **1247784** | tt9916682 | 6.4 | tvEpisode | 0 | 2012 | 10 | Adventure,Animation,Comedy | nm0996406 |
| **1247785** | tt9916690 | 7.4 | tvEpisode | 0 | 2012 | 10 | Adventure,Animation,Comedy | nm0996406 |

760564 rows × 11 columns

*Рис. 1. Сведённый датасет (часть 1)*

| | writers | (director) primaryProfession | (director) knownForTitles |
|---|---|---|---|
| **8** | nm0085156 | director,writer,cinematographer | tt0000009 |
| **34** | nm0410331 | cinematographer,director,producer | tt0308254,tt0219560,tt1428455,tt1496763 |
| **74** | nm0410331 | cinematographer,director,producer | tt0308254,tt0219560,tt1428455,tt1496763 |
| **89** | nm0617588 | director,actor,producer | tt0002113,tt0215737,tt0223267,tt0000091 |
| **106** | nm0410331 | cinematographer,director,producer | tt0308254,tt0219560,tt1428455,tt1496763 |
| **...** | ... | ... | ... |
| **1247779** | nm3219235 | director,producer,writer | tt1473818,tt9916544,tt1332123,tt1830903 |
| **1247781** | nm1485603,nm1485604,nm1866876,nm0909144 | director,writer,producer | tt1618470,tt10986410,tt17501750,tt4051832 |
| **1247782** | nm1482639,nm2586970 | director,animation_department,art_department | tt0286490,tt0090315,tt0082509,tt2560206 |
| **1247784** | nm1482639,nm2586970 | director,animation_department,art_department | tt0286490,tt0090315,tt0082509,tt2560206 |
| **1247785** | nm1482639,nm2586970 | director,animation_department,art_department | tt0286490,tt0090315,tt0082509,tt2560206 |

760564 rows × 11 columns

*Рис. 2. Сведённый датасет (часть 2)*

Так как данные регулярно обновляются, зафиксируем дату обращения – 15.11.2022.

Итак, чтобы построить прогноз, нужно решить задачу регрессии, состоящую в построении алгоритма, отображающего множество объектов, описываемых признаками, во множество target-меток. На значения признаков ограничения не накладываются, а значение target-метки может быть любым вещественным числом.

В нашем случае target-меткой будем считать переменную average rating, принимающую значения в отрезке от 0 до 10, а признаками – 9 оставшихся показателей, описывающих фильм. Однако следует отметить, что это число будет меняться в ходе преобразований.

Для нахождения оптимального решения будем строить разные алгоритмы путём обучения стандартных моделей на основе собранных данных, а затем выберем лучший из них.

В качестве функционала качества будем использовать метрику MAE – средний модуль отклонения ответа алгоритма от истинного значения

$$MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i| , \quad (1)$$

где – истинное значение target-метки; – значение, полученное от алгоритма; $n$ – число объектов, по которым получен прогноз.

Выбор обусловлен классом решаемой задачи, а также отличной интерпретацией – значение MAE для данной задачи в точности равно баллу, на который в среднем ошибается алгоритм.

Таким образом, лучшим будем считать тот алгоритм, для которого значение MAE на тестовой выборке (части исходного датасета, отведённой для получения прогноза и оценки качества) будет наименьшим.

Для дальнейшей работы будем использовать язык Python, а также вычислительные ресурсы Colab от компании Google [4].

## 3. ПРЕДОБРАБОТКА ДАННЫХ

Для того чтобы строить прогноз, нужно привести все данные к числовому формату, с которым может работать компьютер.

Признаки “titleType” (тип/формат) и director (id режиссёра) являются категориальными. Признак “titleType” имеет 10 уникальных значений, поэтому для представления в числовом формате применим стандартную технику бинаризации, т.е. для каждого уникального значения сделаем отдельный бинарный столбец, в котором будет стоять 1, если данный тип соответствует фильму и 0, если не соответствует. В результате получим 10 новых признаков.

Признак “directors” же содержит в себе 157557 уникальных значений. Создавать бинарные признаки будет слишком затратным с точки зрения памяти действием, поэтому воспользуемся популярным методом и закодируем каждое значение частотой, с которой оно встречается в датасете.

Значениями признаков “genres”, “(director) PrimaryProfession” и “(director) knownForTitles” являются массивы строк. Метод кодирования частотой здесь не подходит, т.к. в этом случае одно и то же значение получат только фильмы, полностью совпадающие по спискам жанров, то есть удовлетворяющие очень сильному условию схожести. В других же случаях схожесть будет игнорироваться. Бинаризация же применима, но реализуется немного сложнее. Следует отметить, что признак “(director) knownForTitles” содержит большое число уникальных значений, поэтому для него придётся ограничиться, например, 500 самыми распространёнными фильмами с целью экономии ресурсов. Проведём соответствующие преобразования, а также

удалим признак tconst, который не понадобится нам в дальнейшем. В результате получим новый датасет, состоящий из 358 столбцов.

| | averageRating | isAdult | startYear | runtimeMinutes | directors | writers | Romance | Drama | Short | Horror | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **8** | 5.3 | 0 | 1894 | 45 | 1 | 1 | 1 | 0 | 0 | 0 | ... |
| **34** | 4.4 | 0 | 1896 | 0 | 14 | 10 | 0 | 1 | 1 | 0 | ... |
| **74** | 4.5 | 0 | 1896 | 1 | 14 | 10 | 0 | 1 | 1 | 0 | ... |
| **89** | 6.7 | 0 | 1896 | 3 | 57 | 30 | 0 | 0 | 1 | 1 | ... |
| **106** | 4.4 | 0 | 1896 | 1 | 14 | 10 | 0 | 1 | 1 | 0 | ... |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **1247779** | 6.9 | 0 | 2019 | 12 | 4 | 1 | 0 | 1 | 1 | 0 | ... |
| **1247781** | 7.4 | 0 | 2019 | 44 | 111 | 17 | 0 | 0 | 0 | 0 | ... |
| **1247782** | 8.5 | 0 | 2012 | 10 | 28 | 30 | 0 | 0 | 0 | 0 | ... |
| **1247784** | 6.4 | 0 | 2012 | 10 | 28 | 30 | 0 | 0 | 0 | 0 | ... |
| **1247785** | 7.4 | 0 | 2012 | 10 | 28 | 30 | 0 | 0 | 0 | 0 | ... |

760564 rows × 358 columns

*Рис. 3. Преобразованный датасет*

Другие признаки, а также target-метка являются являются числовыми и не требуют никаких дополнительных преобразований. Следует отметить, что при обучении определённых моделей могут понадобиться дополнительные преобразования датасета.

## 4. ОБУЧЕНИЕ МОДЕЛЕЙ

Разобьём наш датасет на обучающую и тестовую выборки таким образом, чтобы в тестовую выборку попали самые поздние фильмы (так так подобная ситуация лучше всего моделирует реальный мир, в котором мы будем получать новые объекты из будущего). Для этого нужно отсортировать все объекты по признаку startYear и разделить датасет. После этого обучим модели Linear Regression, kNN, Random Forest, Gradient Boosting. Конечно, речь идёт о версиях, адаптированных под задачу регрессии. Для первых четырёх моделей есть хорошие реализации в библиотеке scikit learn, а для последней будем использовать LGBMRegressor из библиотеки LightGBM и CatBoostRegressor из библиотеки CatBoost, так как они являются более мощными и почти всегда лучше работают на практике. Получим следующие значения MAE на тестовой выборке:

- Linear Regression – 0.958;
- kNN – 1.050;
- RandomForestRegressor – 0.943;
- LGBMRegressor – 0.947;
- CatboostRegressor – 0.946.

Как видим, результаты не сильно отличаются, но формально Random Forest оказался чуть более точным. Следует отметить, что при плохо подобранных гиперпараметрах показатели могут стать хуже, в данном случае значение 0.943 достигнуто моделью RandomForest при следующей конфигурации:

- max_depth = 61;
- n_estimators = 1750;
- min_samples_split = 3;
- min_samples_leaf = 6;
- max_features=’sqrt’.

Другие гиперпараметры были взяты по умолчанию.

Для демонстрации работы модели возьмём случайный объект и сделаем на нём предсказание:

```
[68] rf.predict(X_test.iloc[[62701]])

     array([7.38036304])
```

*Рис. 4. Получение предсказания*

Выведем реальное значение target-метки:

```
y_test.iloc[[62701]]

363154    6.8
Name: averageRating, dtype: float64
```

*Рис. 5. Получение реального значения*

Видим, что отклонение составило около 0.5 балла. Посмотрим на описание объекта в первоначальном датасете (т.е. на уровне сведённых сырых данных):

```
tconst                                                  tt1669321
averageRating                                                 6.8
titleType                                                 tvEpisode
primaryTitle                                              The Pilot
originalTitle                                             The Pilot
isAdult                                                           0
startYear                                                      2010
endYear                                                          \N
runtimeMinutes                                                   29
genres                                                      Comedy
directors                                                 nm0002433
writers                                                   nm0255910
(director) birthYear                                             \N
(director) deathYear                                             \N
(director) primaryProfession          director,producer,miscellaneous
(director) knownForTitles       tt1558182,tt4254242,tt0169190,tt1755893
(writer) birthYear                                               \N
(writer) deathYear                                               \N
(writer) primaryProfession              producer,writer,miscellaneous
(writer) knownForTitles       tt15201944,tt0218141,tt1558182,tt7599942
```

*Рис. 6. Описание объекта*

Видим, что в данном случае речь идёт об одной из серий некоторого комедийного сериала, выпущенной в 2010 году.

## 5. ЗАКЛЮЧЕНИЕ

Таким образом, можно сказать, что методы машинного обучения могут решать поставленную задачу с достаточно высокой точностью в терминах MAE, что открывает большие возможности для прикладного применения. Конечно, требования к метрикам зависят от конкретной задачи, и вполне возможно, что в реальной ситуации полученных значений будет недостаточно. Однако было показано, что подход является достаточно перспективным, также при необходимости его можно дорабатывать с целью повышения точности.

***Литература***

1. *Баев М.А.* Предсказание оценки фильма на IMDB. Материалы 60-й Международной научной студенческой конференции. Новосибирск, 2022 – с.281.
2. *Кирилина Н.А., Горбанёва Е.Н.* Применение алгоритмов машинного обучения randomforest, gradientboosting, kneighbors для прогнозирования кассовых сборов кинофильмов. Программная инженерия: современные тенденции развития и применения (ПИ-2019) Курск, 11–12 марта 2019 года – с. 25–28.
3. IMDB Datasets [https://datasets.imdbws.com/].
4. Google Colab [Электронный ресурс]. – Режим доступа: https://colab.research.google.com/ (дата обращения 01.09.22.)

# Forecasting the Rating of a New Movie Based on its Metadata

***Gleb B. Sologub****
Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia
ORCID: https://orcid.org/0000–0002–5657–4826
e-mail: glebsologub@ya.ru

***Nikita S. Sazon*****
Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia
ORCID: https://orcid.org/0000-0002-9816-4585
e-mail: nikitaS1598@gmail.com

The article describes an approach to predicting the rating of a new film based on data known prior to its release, using classic machine learning models. The approach includes testing various models with appropriate data preprocessing and selection of optimal hyperparameters, as well as choosing the best algorithm in terms of the selected quality functionality.

***Keywords:*** machine learning, film rating prediction.

**For citation:**
Sologub G.B., Sazon N.S. Forecasting the Rating of a New Movie Based on its Metadata. *Modelirovanie i analiz dannykh = Modelling and Data Analysis,* 2023. Vol. 13, no. 2, pp. 77–84. DOI: 10.17759/mda.2023130204 (In Russ., *abstr*. in Engl.).

***References***

1. Baev M.A. Predicting movie ratings on IMDB. Materials of the 60th International Student Scientific Conference. Novosibirsk, 2022 – p. 281.
2. Kirilina N.A., Gorbanyova E.N. Application of machine learning algorithms randomforest, gradientboosting, kneighbors for predicting box office revenues of movies. Software Engineering: Modern Trends in Development and Application (PI-2019) Kursk, March 11–12, 2019 – p. 25–28.
3. IMDB Datasets [https://datasets.imdbws.com/].
4. Google Colab [Electronic resource]. – Access mode: https://colab.research.google.com/ (accessed on September 01, 2022).

***Gleb B. Sologub,** Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Mathematical Cybernetics, Institute of Information Technologies and Applied Mathematics, Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia, ORCID: https:// orcid.org/0000–0002–5657–4826, e-mail: glebsologub@ya.ru

****Nikita S. Sazon,** Master's Student at the Institute of Information Technologies and Applied Mathematics, Moscow Aviation Institute (National Research University) (MAI), Moscow, Russia, ORCID: https://orcid.org/0000-0002-9816-4585, e-mail: nikitaS1598@gmail.com

Получена 17.03.2023 Received 17.03.2023
Принята в печать 17.04.2023 Accepted 17.04.2023