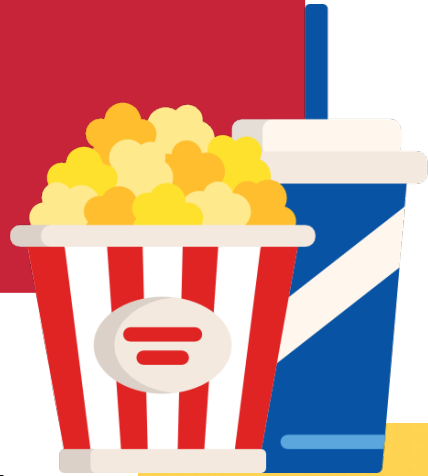
Data Science Use Case Scenario: How good is a movie?

Benazir de la Rosa
Data Science Consultant



Overview

- **How good is the movie?**
- Datasets descriptions
- Descriptive analytics
- Feature Engineering
- ML Algorithm and results
- Next Steps

How good is the movie?

- The presented use case scenario tries to predict how good is the movie base on 6 different datasets available on kaggle. See git repo

https://github.com/nonameforpirate2/Prediction-Movie-Good-or-Bad

Overview

- How good is the movie?
- **Datasets descriptions**
- Descriptive analytics
- Feature Engineering
- ML Algorithm and results
- Next Steps

Datasets Descriptions

The six datasets available are the next ones

- **genome_scores**: contains a list of tags assign to each movie with its corresponding relevance.
- **Genome_tags**: contains the list of the different tags created by users with its corresponding ID.
- **Link**: it has an homologation of the ids between movieId, imdbId and tmdbId.

Datasets Descriptions

The six datasets available are the next ones

- **Movie**: it has information about the movies, title, movieId and genre assign to the movie.
- **Rating**: it is the most important dataset. It contains the activity rating of users to movies with its corresponding timestamp. It has the user id, movie id, timestamp and the rating.
- **Tag**: it has the tagging activity of users to movies with its corresponding timestamp.

Overview

- How good is the movie?
- Datasets descriptions
- **Descriptive analytics**
- Feature Engineering
- ML Algorithm and results
- Next Steps

Descriptive Analytics

- From the available datasets, it was possible to describe the next information:
  * The most of the movies do not have a big relevance, very few movies are quite relevant.
  * People create so far 1128 different tags to assign to movies.
  * There are 27278 movies, most of them contain the "year" in the title. However, around the 19.2%

    do not have a year asign (5218).
  * There are movies from the year 1913 to the year 2013.The year with more movies is 2013.

Descriptive Analytics

- From the available datasets, it was possible to describe the next information:
  * There are 20 different genres.
  * From the rating asignation the time took place between 1995-2015.
  * Most of the movies have a 4 + rating.
  * It was found that not all the movies have a tag assign, only 71.6% (19545) have a tag.
  * Customers present a tendency to assign around 10-15 tags per movie.

Overview

- How good is the movie?
- Datasets descriptions
- Descriptive analytics
- **Feature Engineering**
- ML Algorithm and results
- Next Steps

Feature Engineering

From the available datasets, it was possible to describe the next information:

- **Relevance feature**. The feature describes the average relevance that a movie has given the tags available on it.
- **Movie Year feature**. Text mining was done on the movie titles to get the year from each movie.
- **Movie Genre feature**. Text mining was done to create label columns with each of the 20 genres.
- **Movie tag feature**. Represents the amount of tags presented in a movie.
- **User tag feature**. Represent the amount of tags that a user has use.

Feature Engineering

- **User movie feature**. Represents the amount of tags that a user use in a movie.
- **Month feature**. User behavior intention by seasonality. Represents the month of the year when the user rated the movie. Weather influence on human behavior.
- **Day feature**. User behavior intention by day. Represents the day of the month when the user rated a movie. Paycheck effect/friday and beers etc.
- **Hour feature**. User behavior intention by hour. Represents the hour in the day when the user rated a movie. Before going to sleep efect, student, house wife etc.
- **Year rating feature**. Represents the years that passed by from the moment of the rating and the movie creation.

Overview

- How good is the movie?
- Datasets descriptions
- Descriptive analytics
- Feature Engineering
- **ML Algorithm and results**
- Next Steps

ML Algorithm and Results

A machine learning algorithm was trained to predict whether the movie was good or bad, it takes into account a +4 stars rating as a good movie.

- The selected algorithm was a random forest with 50 trees, entropy criterion, depth = 5.
- Its performance was measure with a 10 k-fold cross validation and it gave as a result a 65% accuracy.

Overview

- How good is the movie?
- Datasets descriptions
- Descriptive analytics
- Feature Engineering
- ML Algorithm and results
- **Next Steps**

Next Steps

This data science model is a baseline and there is a lot to be done to improve it.

- There is plenty of room to work with NER (name entity recognition) in the part of the tagging / More feature engineering with text mining.
- More feature engineering.
- More hyper parameter tuning.
- Test more algorithms.

Thanks!

Benazir de la Rosa
Data Science Consultant