Predicting IMDb Movie Rating using Machine Learning

Ananya Jain
ananya19408@iiitd.ac.in

Manasvi Singh
manasvi19369@iiitd.ac.in

Pritish Wadhwa
pritish19440@iiitd.ac.in

Yash Bhargava
yash19289@iiitd.ac.in

1. Abstract

Is it possible to predict the rating of a movie prior to its release or production? Every year countless movies are made and released worldwide. All these movies are given ratings by viewers throughout the globe.These ratings are combined together to form the IMDb ratings. IMDb rating is the singlemost influential factor in deciding any consumer's opinion and inherently the success of a movie.

With the machine learning techniques at our disposal, we aim to predict the seemingly unpredictable IMDb rating of any movie before its theatrical release. Successfully predicting IMDb rating is beneficial for both producers (from a financial standpoint) and consumers (from an entertainment standpoint) alike.

2. Introduction

The main problem our team aim to tackle is to predict the IMDb rating of any movie prior to its release. We use a variety of features ranging from the movie overview, length of movie runtime, the country where the movie was produced, the language of the movie, to the details about the lead actors, the movie director and even the details about the movie's key poster. With this information in our arsenal, we aim to use a number of machine learning algorithms to accomplish this uphill task.

We have made sure that features which are affected by the release of movies are not taken into account. These features include properties like popularity, revenue, user ratings among others.

We have tackled this problem as a regression task. We aim to predict the ratings as close as possible to the original IMDb ratings. One thing to note is that since the IMDb ratings are influenced by the viewer reviews, these ratings might fluctuate. But over time, these ratings are more or less constant. For the task at hand, we are using the IMDb ratings as reported in the dataset without actually verifying any small change there might have been in any of the movie ratings since the publication of the dataset.

3. Literature Survey

Oghina et al. in their paper "Predicting IMDB Movie Ratings Using Social Media" [3] addressed the task of predicting IMDb movie ratings using data collected from social media services. They identified qualitative and quantitative activity indicators for a movie in social media, and extracted two sets of surface and textual features. They trained various models and upon analysing them, they found that the fraction of the number of likes and dislikes on YouTube, combined with textual features from Twitter lead to the best performing model, with strong agreement with the observed ratings and high predictive performance.

Quader et al. in their paper "A machine learning approach to predict movie box-office success" [4] proposes a decision support system for movie investment sector using machine learning techniques. The system predicts an approximate success rate of a movie based on its profitability by analyzing historical data from different sources like IMDb, Rotten Tomatoes, Box Office Mojo and Metacritic. They discovered that budget, IMDb votes and number of screens are the most important features which play a vital role while predicting a movie's box-office success.

C. izmeci et. al in their article [5] explored the IMDb ratings by exploring matrix decomposition, regression analaysis and factorization machines on social media data.

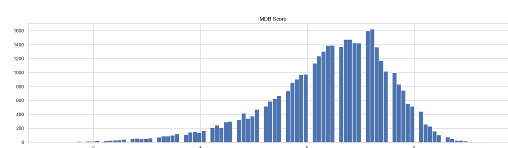

Figure 1. IMDb score distribution over the dataset

4. Dataset

We utilized open-source datasets, "The Movie Dataset" [1], and "Movie Genre from its Poster" [2] from Kaggle. We selected features like posters, synopses, cast, crew, runtime,

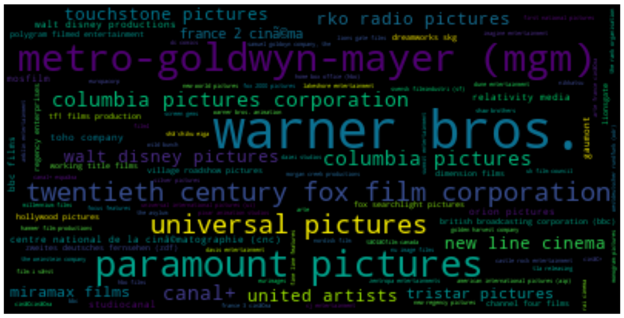

Figure 2. Word Cloud Depicting frequency of major production houses throughout the dataset

genre among others as input features and IMDb score as the prediction objective. The data set after preprocessing have the fields mentioned in table 1.

4.1. Preprocessing

The data set had variables which were discarded like Adult, describing whether a movie is adult or not, because of false and highly skewed data. Budget had about 75 percent null or zero values and Tagline also had around 75 percent null or empty values and both of these were dropped. We tried but could not find replacement for budget data. Only alternatives in front of us were illegally scraping the data or droping the column itself. We could not assign mean or median values to the empty filed as the movies on our dataset ranged over a large number of years. This might have lead to inclusion of false data in the dataset, thus we decided it better to leave out the feature itself.

Post release attributes like IMDb_id, popularity, revenue, vote_count, homepage, status, IMDb link were discarded, spoken language was also discarded since it was found redundant and non informative. We did binning on release date feature into release year. Production company and country column also required further formatting and were converted into the desired typed; after processing they were stored as pipe separated values in the intermediate preprocessed stage. We manually extracted 13 visual features which are the mean and standard deviation of red, green, blue, hue, saturation, brightness, as well as the number of

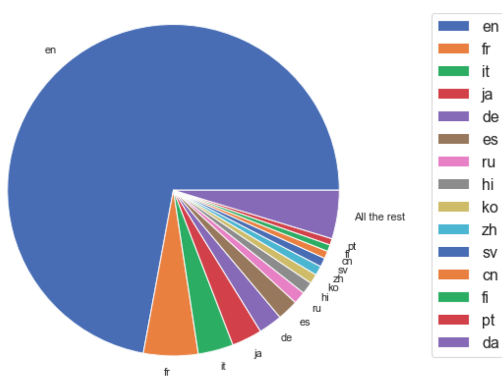

Figure 3. Language Distribution of the movies

Feature Name	Description
year	Year of release of the movie
runtime	Duration of the movie
director	The director of the movie
actors	The actor and actresses in the movie
original_language	The main language spoken
original_title	The title of the movies
overview	The summary of the movie
production_companies	The company producing the movie
production_countries	country of main production
keywords	The main themes in the movie
genre	The different genres of the movies
num_faces	Number of faces in the poster
saturation	Saturation level
hue	Hue level of the movie poster
brightness_sd	standard deviation of brightness level
saturation_sd	standard deviation of saturation level
hue_sd	standard deviation of hue level
green	green colour level in the poster
green_sd	standard deviation of green colour level
red	red colour level in the poster
red_sd	standard deviation of red colour level
blue_sd	standard deviation of blue colour level

Table 1. Features

human faces using the openCV library. For movie overview and title, we used spaCy for tokenization and nltk for stop words and only kept words that appeared in more than 75 and 10 movies respectively. For cast and crew, we extracted the director and top three actors(in the order of the credits) for each movie, and kept directors involved in at least 10 movies and actors involved in at least 20 movies in our dataset.
Our prediction attribute IMDb_score was subsequently separated from the dataset and made as the labels. After all preprocessing steps, we have a total of 21782 data points and 3561 features.

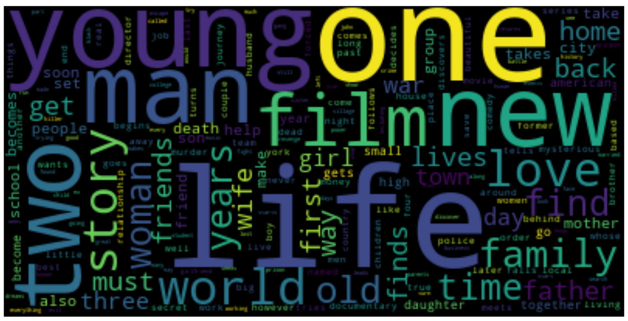

Figure 4. Word Cloud Depicting frequncy of words in the movie overview

Feature	Score
director	2259479
genres	2228322
runtime	27871
original_language	23777
green	6883
red	5974
brightness	4983
blue	4826
saturation	4117
saturation_sd	2374
hue_sd	1517
hue	981
blue_sd	364
release_date	327
num_faces	213
brightness_sd	199
green_sd	180
red_sd	171
adult	31

Figure 5. Feature Scores calculated using SelectKBest

4.2. Feature Selection

We used SkLearn's SelectKBest technique to evaluate the best features in our dataset, we realised Adult to be the feature with the least score hence dropped it. We also realised the poster data did not act as strong features due to their lower scores in the analysis but nevertheless, they were kept in the final dataset as when combined they were substantial enough.
We plotted a correlation matrix and visualized it using a heatmap. We realised our data is not very highly correlated, and dropped the features with strong positive or negative correlation. Irrespective, we found that brightness had a really high correlation with the three primary colours and thus as a consequence it was removed.
We used the Random Forest Classifier to find the features of importance and validated the output we obtained from SelectKBest technique.

4.3. Hot Encoding Data

The textual data is converted to Bag of words format and stop words are removed, analysis is done on the basis of frequency.The Bag of word format is then hot encoded.

4.4. Preparation of Training and Testing Data

We divided out dataset into training, validation and testing set with a split of 70:15:15, setting the random seed as 0 and hence shuffling the dataset to prevent any unfortunate split and patterned split.

4.5. Data Normalization

Normalization is a scaling technique where the values are centred around the mean and have a unit standard deviation. Hence the mean of the feature in consideration becomes zero and the distribution obtained has a standard deviation of one. The formula followed for the same is:

$$\hat{x_i} = \frac{x_i - \mu}{\sigma}$$

where μ is mean & σ the standard deviation of the feature.

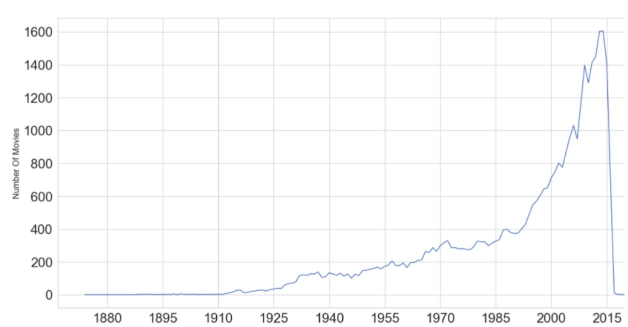

Figure 6. Number of movies vs. Release year distribution

5. Methodology

For the prediction of IMDb scores, we put to use the following models:

5.1. Regression

We approached our problem to be that of a Regression problem. We used the features extracted for our task to predict the IMDb ratings, and applied the following regression models:

5.1.1 Linear Regression

We used standard linear regression as our baseline model. To account for overfitting and making our model highly complex we used L1 and L2 regularization techniques to mitigate the same.

5.1.2 Ridge Regression

Ridge regression introduces penalty term to shrink model parameters. However ridge doesn't eliminate parameters and thus model includes all predictors which makes model interpretability difficult.

5.1.3 Lasso Regression

Lasso regularization uses shrinkage to form sparse model which helps solve multi-collinearity and aids feature selection. Since Lasso can shrink some model parameters to exactly 0 unlike ridge regression, it makes a much more interpretable model than ridge.

6. Results and analysis

We evaluate our working on this problem statement through the following:

6.1. Metrics

We have used Mean Squared Error (MSE) and R^2 metrics. MSE is used for calculating loss function. R^2, the coefficient of determination is a measurement of proportion of variance for predicted IMDb ratings.

6.2. Individual Models

After analysing with GridSearchCV for Lasso and Ridge Regression, we have set alpha as 0.01 for both Ridge and Lasso.
The results are shown in Table 2.

Model	R^2	**MSE**
Linear Regression		
Training Set	0.5951	0.4824
Testing Set	0.3464	0.7737
Lasso Regression		
Training Set	0.3154	0.8156
Testing Set	0.3424	0.7784
Ridge Regression		
Training Set	0.5951	0.4824
Testing Set	0.3464	0.7736

Table 2. Results of Models

7. Conclusion

Our expectations from the project were to predict the IMDb rating of any movie using features that are available prior to its release and analyse the discriminatory power of features to analyse how different parameters impact the rating for any movie using various machine learning models.
From the models implemented so far we observed that features like length of a movie, the director, movie genres are amongst the most important ones to make any kind of prediction. We also explored various features like the movie poster, the movie overview or summary, the language of the movie and the country of production of the movie and found out that even though they play an important part in the predictions, but when taken individually they are not as strong as the ones mentioned above.

7.1. Learning

We all learnt how to curate a dataset, perform Exploratory Data Analysis on a large dataset, how to preprocess data, handle null values, create sparse matrices and train machine learning models.

7.2. Future Work

We are yet to implement higher degree regression models, Decision Tree, Random Forest Classifier, SVM and Basic Neural Networks etc. We have to perform a comparative analysis of various machine learning models, to select the best model for our predictions. This is the fifth week since our proposal submission and we find ourselves in line with the timeline proposed having implemented the regression models.

8. Contribution

Each member played a crucial part in discussions, analysis and making this report.
Ananya and Manasvi went over with Exploratory Data Analysis, Preprocessing and the Report while Pritish and Yash dealt with curating dataset, Linear regression, Feature Selection and its analysis.

References

[1] . The Movies Dataset, 11 2017.

[2] . Movie Genre from its Poster, 05 2018.

[3] Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten Rijke. Predicting imdb movie ratings using social media. pages 503–507, 04 2012.

[4] Nahid Quader, Md. Osman Gani, Dipankar Chaki, and Md. Haider Ali. A machine learning approach to predict movie box-office success. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–7, 2017.

[5] Chuan Sun. Predict Movie Rating, 2020.