Machine Learning Project

IMDb Movie Rating Prediction

Animesh Sharma (140905526)

Mekapotula Sandeep Reddy (140905602)

Abstract

The aim of this project is to create an IMDb movie rating prediction system which utilizes the technique of machine learning to learn from a dataset that contains information on various parameters about movies released in the past, such as budget of the movie, social media presence of the actor and director, total box office earnings etc. Using the information present on these parameters in the dataset, and comparing these values with those of the movie who's rating is to be predicted, it is possible to predict an approximate IMDb rating of any movie. To get a good idea of the accuracy of prediction with multiple algorithms, a host of classification and regression methods have been used, some of which are Random Forest Classifier, Logistic Regression and Adaptive Boosting Regression.

Introduction

Machine learning is the subfield of computer science that gives "computers the ability to learn without being explicitly programmed." Evolved from the study of pattern recognition and computational learning theory in artificial intelligence, machine learning explores the study and construction of algorithms that can learn from and make predictions on data. The predicted data can then be compared with a test set containing the actual data to determine the accuracy and efficiency of the machine learning model used. Machine Learning finds applications in a plethora of environments, from email filtering, to Optical Character Recognition and Computer Vision. In the context of this project, we use Machine Learning to classify the given dataset based on the available parameters and then predict the IMDb rating of a movie from a test set based on this classification.

Survey

Movie ratings are influenced by many factors, so accurate prediction of new movie ratings can be challenging. The use of Machine Learning for classification or regression problems in areas like the one in this project is a fairly new trend, in part due to the limited computation power available in the past and due to the inadequacy of the existing algorithms used for the actual prediction. Nevertheless, attempts have been made by people to solve similar problems in the past. Some of these involve the use of image recognition of the posters of movies to better identify the top billed cast in a movie using computer vision, others use natural language processing on plot keywords and user reviews to classify movies as good or bad. Still others have used search tends to determine the relative popularity of movies over a period of time and use this data to predict the rating of a movie. The large number of possible valid ratings makes this problem both difficult and challenging.

Scope & Objectives

The scope of this project covers, but is not limited to prediction of IMDb movie ratings of any particular movie based on a set of features. The techniques used in the prediction model – feature selection, regression and cross validation – can be used for any similar problem to get desired results. In this particular implementation, the objective was to maximize the accuracy of the predicted values based on the obtained dataset by fine tuning and tweaking the parameters of the various algorithms available and come up with a rating that is close to the actual rating of a movie with as small a margin of error as possible.

Gaps

The fact that the valid IMDb rating of a movie can be anywhere from 0.0 to 10.0, which makes it a total of 101 possible values – makes this problem both difficult and challenging. If a person is asked to make a random guess about the IMDb rating of a particular movie, he/she would be correct about 1/101 or approximately 0.99% of the time, which is pretty low. Any technique which can increase this accuracy is certainly desired, and the implementation presented here can satisfy this goal. Accurate prediction, even with advanced machine learning techniques, is difficult however due to the large number of valid outcomes. However, if we take a small margin of error on either side of the predicted value, we can greatly improve the accuracy of the model. Due to inconsistencies with the dataset (a lot of values being null, have to be estimated) and some amount of overfitting, perfect accuracy can never be achieved.

Methodologies

- The first step in the process involves obtaining the dataset. This was done using web-scraping tools in python and filtering the results obtained. This involved crawling the social media pages of the actors and the director along with the IMDb page of a movie to scrape the required information.
- Once the dataset was obtained, the next step was data preprocessing. This involved dropping some unnecessary features from the dataset and using a label encoder to assign numeric values to strings so that these features can be used effectively for the classification.
- The next step involved feature selection and normalization of the values to identify the important features that have a significant impact on the IMDb rating of a movie.
- Once feature selection was complete, we used the Random Forest Classifier to predict the IMDb ratings of movies in a test set using the rest of the dataset to learn.
- Once a certain accuracy was achieved, we factored in a margin of error into the accuracy

calculations. This is justified due to the large number of valid ratings and helped boost accuracy by a large factor.

- In the next step, we cross – validated the dataset to determine the maximum accuracy that could be achieved from it. We then did cross - validation using a number of different algorithms to determine if Random Forest is indeed the best choice of algorithm or if some other algorithm is better suited to the task.
- Finally, we plotted graphs between various parameters that represent the impact of particular features on the predicted IMDb rating of the movie. A correlation analysis plot reveals the features that have a significant impact on the final rating and others that have a limited impact.

Results and Analysis

Using the methodology stated above, the following results were obtained :

Dataset :

Number of Dataset Entries – 4301

Number of Features Present - 28

Number of Features Used (After Feature Selection) – 15

Accuracy :

Random Guess (Out of 101 possible IMDb ratings) – 0.99%

Exact Accuracy (Random Forest Classifier, after algorithmic tuning) – 9.34%

Final Accuracy (Margin of Error : + or - 0.3) – 53.43%

Cross Validation (Margin of Error : + or - 0.3) :

Logistic Regression – 43.95%

Linear Discriminant Analysis – 43.63%

K-Nearest Neighbors – 42.26%

Gaussian Naïve Bayes – 15.07%

Decision Tree Classifier – 42.23%

Support Vector Machine – 48.09%

Random Forest Classifier – 53.61%

Gradient Descent – 35.12%

Adaptive Boosting Regression – 40.98%

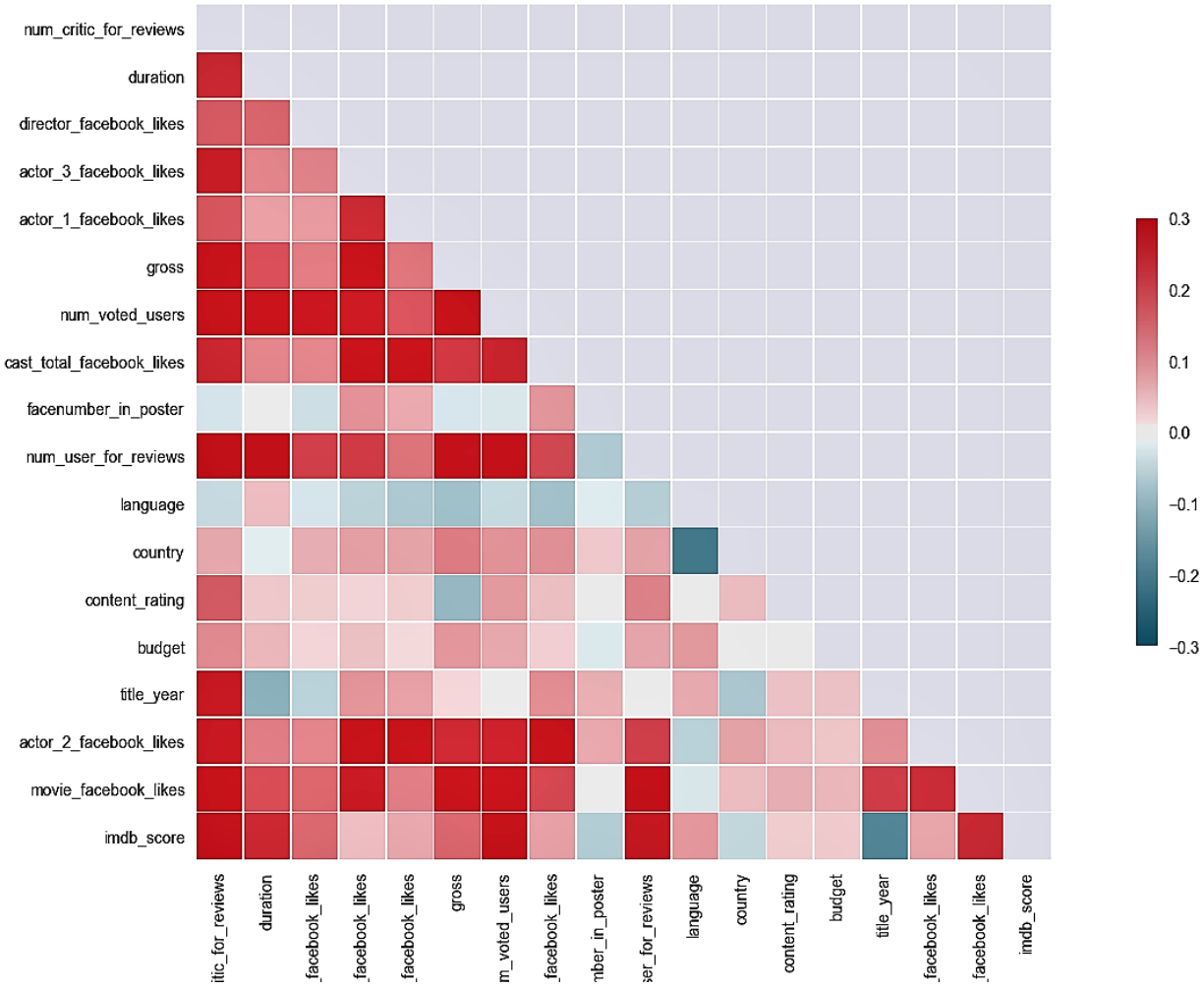
Bagging Classifier – 52.63%

Conclusion :

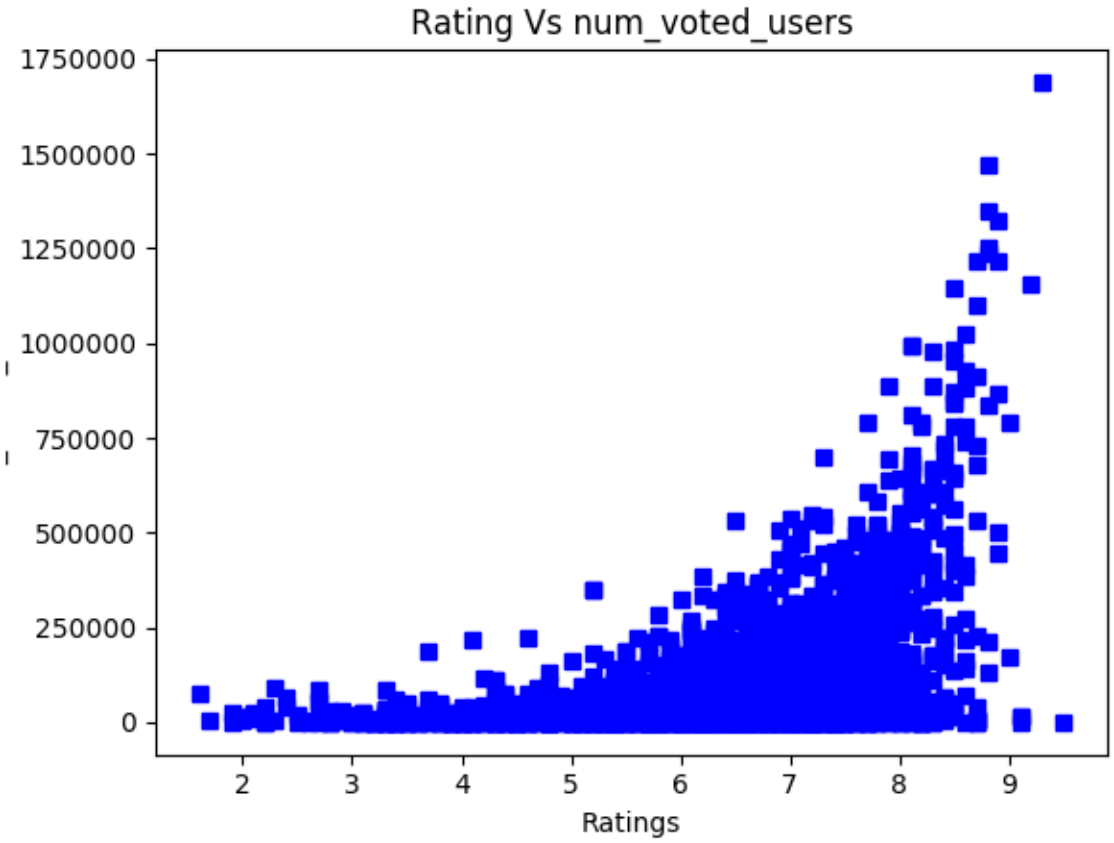
Random Forest Classifier, after some algorithmic tuning, achieves a maximum accuracy of 53.43% with a margin of error of + or – 0.3 in the predicted movie ratings. For exact predictions, it achieves an accuracy of 9.34%, which is a large improvement over that of a random guess.

Graphs

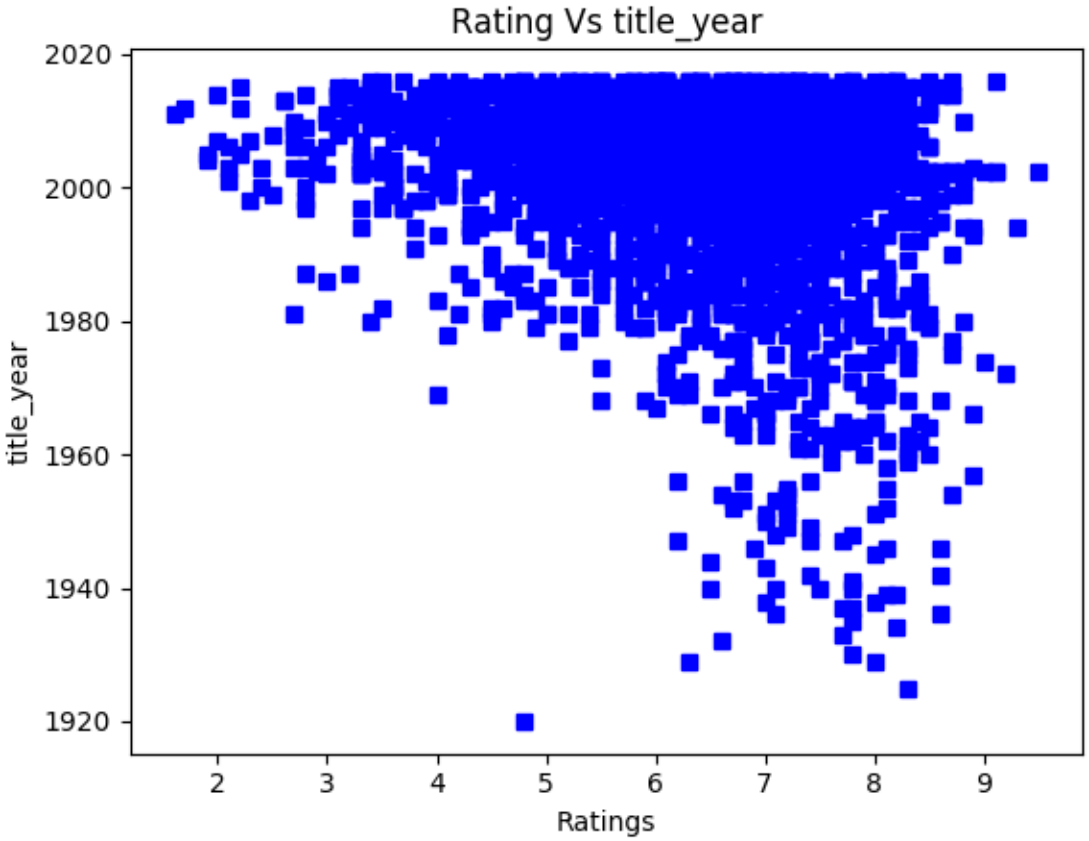
Correlation Analysis :



Number of user votes vs IMDb Rating:



Year of Release vs IMDb Rating :



References

- Inputs from Mr. Muralikrishna SN and Dr. Srikanth Prabhu, Faculty members, Deptartment of CSE, Manipal Institute of Technology
- scikit-learn.org/stable/documentation
- acm.org/citation.cfm?id=2260703
- Stack Overflow