Literature Review

Team: The Miner League

SRN: PES2201800688

Name: Shreya Venugopal

PAPER1:

Ericson,Jeffrey & Grodman,Jesse , 2013, A Predictor for Movie Success, CS229, Stanford University, http://cs229.stanford.edu/proj2013/EricsonGrodman-APredictorForMovieSuccess.pdf

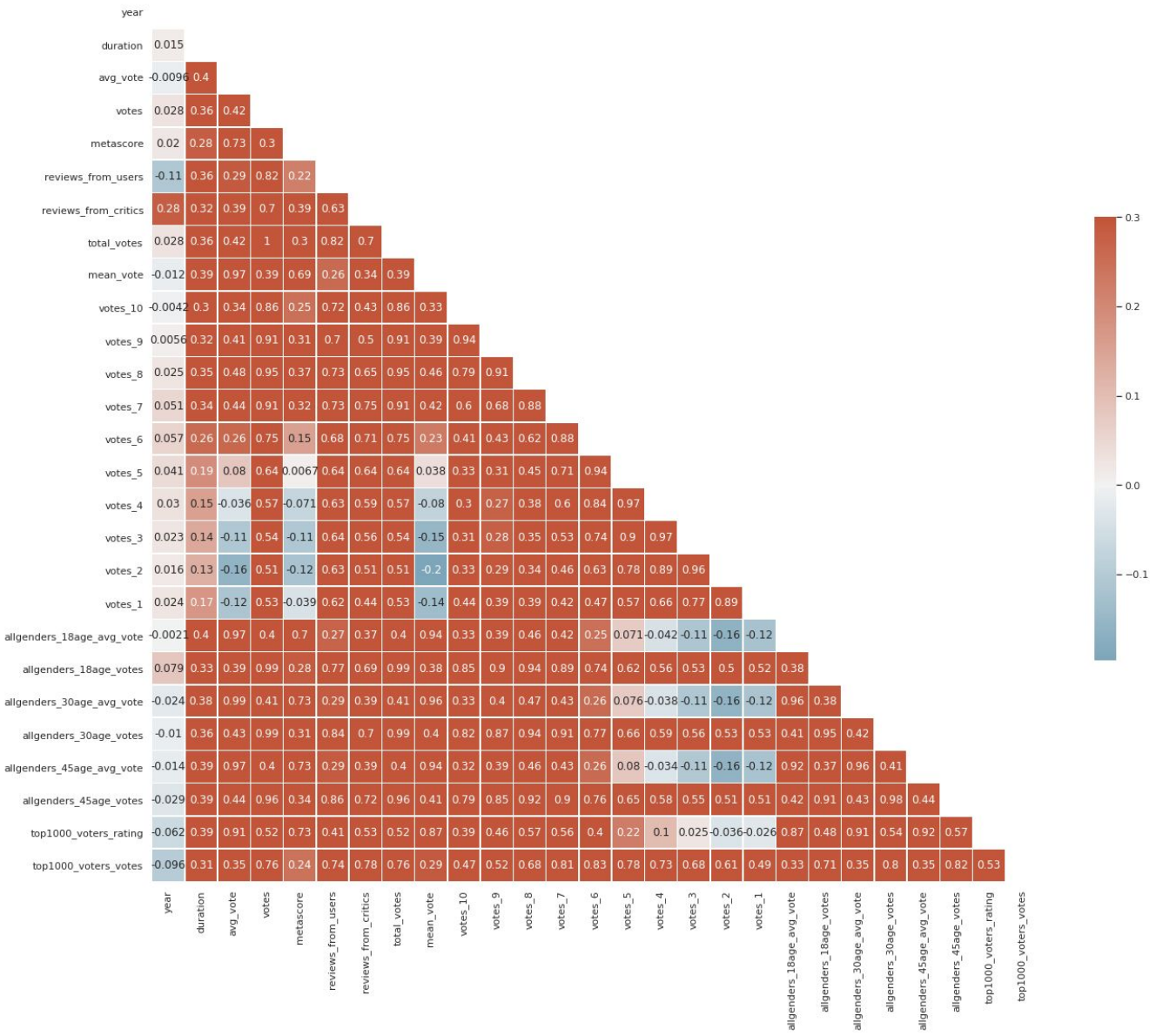
This paper is about the basic concepts of how movies become successful in the first place through their introduction. They predict five different measures of success, based solely on the attributes known about a movie before its release, such as the movie title, plot, budget, characters, and the gross rates. Python and Javascript are used in order to extract, parse and clean the data. Other features such as the time period during which a movie is released (such as on festivals or school summer holidays) are important factors to determine the success rate. Movies with more than 100 votes trend towards a higher rating with more votes. However, for those with fewer than 100 votes, there is little structure to the data. Based on this, all movies with fewer than 100 votes have been removed from the dataset. After cleaning the dataset, a model for the same was obtained.

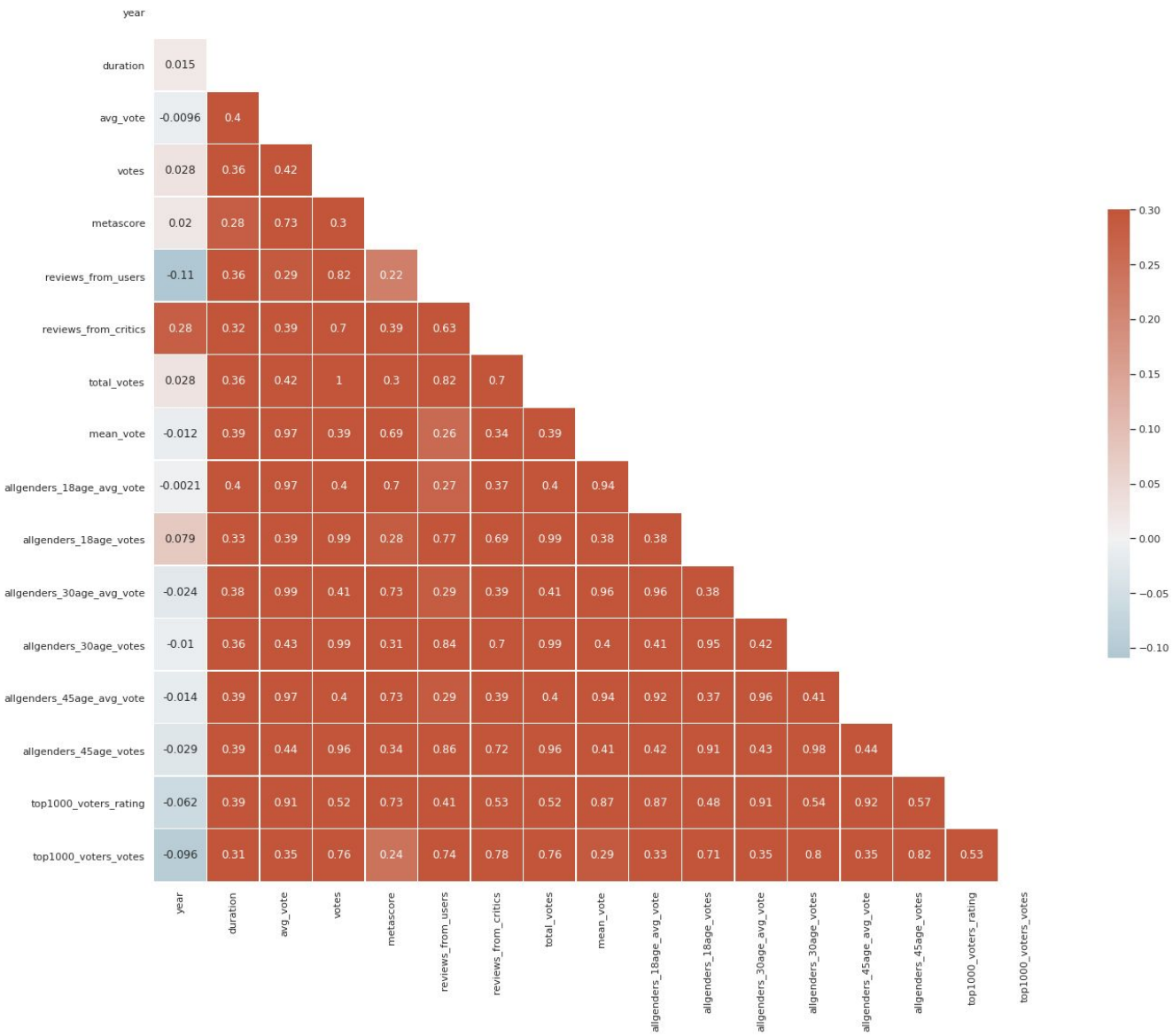
In this case, the first step used is a linear regression model with added weights based on the strength that an attribute has for determining the regression model. For example, in this case similar successful movies have been weighed heavily compared to the others to tell them apart. Training data was 70% and testing 30%, with the tau value of 0.01(weight).

The second technique applied is the SVM (Support Vector Machine), which allowed to put in all 176 of their input vectors to reduce the space used with the help of the forward search paradigm. This selects a subset of features which are best used to make accurate predictions. The following queries were considered based on it:(medians considered)
1. If a movie has a Rotten Tomatoes critics' score above 60 (on a scale to100; Rotten Tomatoes labels a movie either "fresh" with a 60+ score or "rotten" with a score below 60).
2. If a movie has a Rotten Tomatoes audience score above 64 (the median value, also on a scale to 100).
3. If a movie grossed over $7.49M in the United States (the median)
4. If a movie grossed over $500K in the United States its opening weekend (the median) 5. If

a movie has an IMDB score of 6.5 or above (the median, on a scale to10). This is then run on a linear SVM kernel, along with an L1 regularization (Ridge regression), which allows to get a classification rate well above 0.5 for each output feature. The filter feature selection helps in the retrieval of a list of features ranked by score (which was the test success rate from cross validation), which allows to perform hold-out cross-validation. On applying the above and a few final
finishes to the model, a 70% success rate for each of the rating outputs and 88% success rate for each of the monetary gross outputs was obtained.

Using a group of features was an improvement over using any one feature category, even though some single feature categories were indeed highly correlated with outputs. Majority of the popular movies had the terms “life” or “story” in them.





For the waste bodies