CS 373 - FINAL REPORT

Kevin Kochpatcharin kkochpat@purdue.edu
Max O'Cull mocull@purdue.edu
Nameer Qureshi nquresh@purdue.edu
Ryan Sullivan sulli196@purdue.edu

Abstract

The two problems we are solving are the prediction of user ratings based on production tags and production experience, and the prediction of a movie's genres.

The purpose of this project is to attempt to predict a movie's success based on publicly available features from the IMDB database using linear regression and to predict if a movie should have a certain genre/tag depending on its other tags. In order to do so, we retrieve a subset of the IMDB database and filter out samples that are not as relevant to the experiment. Afterwards, we will verify our models using k-folds cross validation to determine the usefulness of our model.

This project aims to verify the usefulness of linear regression in performing multinomial classification where the predicted classes have a strict rank relative to each other. The data set we are using is a subset of the IMDb database. The intent is to train the models to predict the rating of a movie. Models will be trained on a subset of the database and verified using k-folds cross validation.

We also test the usability of decision trees in identifying missing attributes from a set of inputs. Still taking a subset of all IMDb movies as the dataset, the model will attempt to predict a movie's other genres based on given genre's. The models will also be trained on a subset of the database and verified using k-folds cross validation.

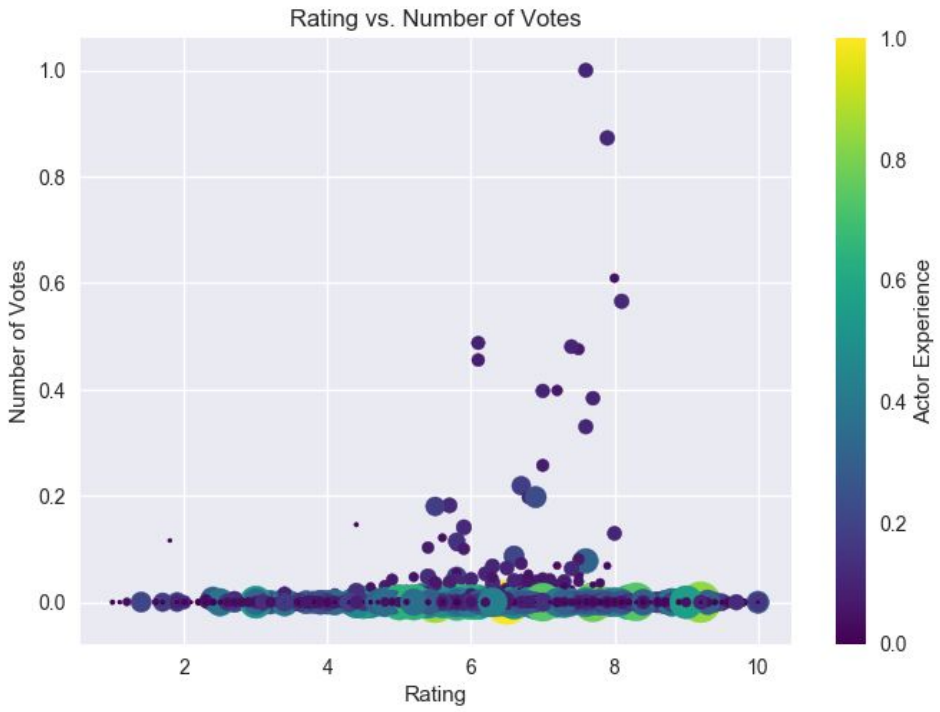
Data

Data is taken from the IMDb database. TV shows and multiple episode films are not included.

Each movie is assigned these attributes.

> { type, isAdult, startYear, runtimeMinutes, isDrama, isComedy, isShort, isDocumentary, isTalkShow, isRomance, isFamily, isNews, isAnimation, isRealityTV, isMusic, isCrime, isAction, isAdventure, isGameShow, isMystery, isSport, isFantasy, isHorror, isThriller, isSciFi, isHistory, isBiography, isMusical, isWestern, isWar, isFilmNoir, averageRating, numVotes, directorsExperience, actorsExperience }

where directorsExperience and actorsExperience is the sum of films participated in the past.



Implementation

Data from the IMDb database is loaded as a single large file. Films that are missing key attributes, such as the ratings, director, etc., are removed. The films are then labelled with a set of binary attributes, representing whether or not the film falls within a specific genre. Using the director's name and year of the movie's release, the number of films the director has released previously is also determined and assigned to the film. Then 10,000 of the films are sampled. Then the films are partitioned into several sets of samples.

The ratings model is generated using a subset of the partition samples, fitted using linear regression and then tested on the remaining samples. The percentage of correctly classified test samples is stored. This is performed against using different sets of partitions.

The genre model is generated slightly differently. The training dataset is produced from a set of the sample partitions. Then, genre attributes from the test set are masked. Model accuracy is then generated using the number of correctly reconstructed film attributes.

Preprocessing

The entire IMDb database is far too large for us to efficiently train the model on. We selected a subset of the data containing approximately 10,000 samples. We first retrieved the

entire IMDb database. Then we removed irrelevant entries. Irrelevant samples included movies that had no attributes attached to them. We did not perform any smoothing on the data.

We cleaned the data retrieved from IMDB in order to convert the labels and other data into more usable binary or numerical features. Also, in order to achieve the most relevant results, we removed all samples that are older than 50 years (about the time color television became popular). We also remove all foreign, non-English films. In order to reduce the time required to train the model, we also reduce the sample size to 10,000 randomly selected samples at this time. We then determine some of the derivative features we will be training the model with, namely the director's experience (the number of films the director has directed prior to this film), and the actor's experience (the average number of films the lead character(s) have participated in prior to the filming of this film). This information is not directly available from the database we are retrieving our data from, but can be derived by observing the entirety of the dataset. We also attach binary labels to each of the movie's genres. We do not predict the values of null values in the population all movies, we only took samples for which the data is complete.

The dataset retrieved from IMDB is stored as a CSV file.

Model and Validation

Rating prediction is performed using different masks over the training set. We experiment with the minimum information needed to be provided to the training algorithm in order to produce a model that can accurately predict a film's rating.

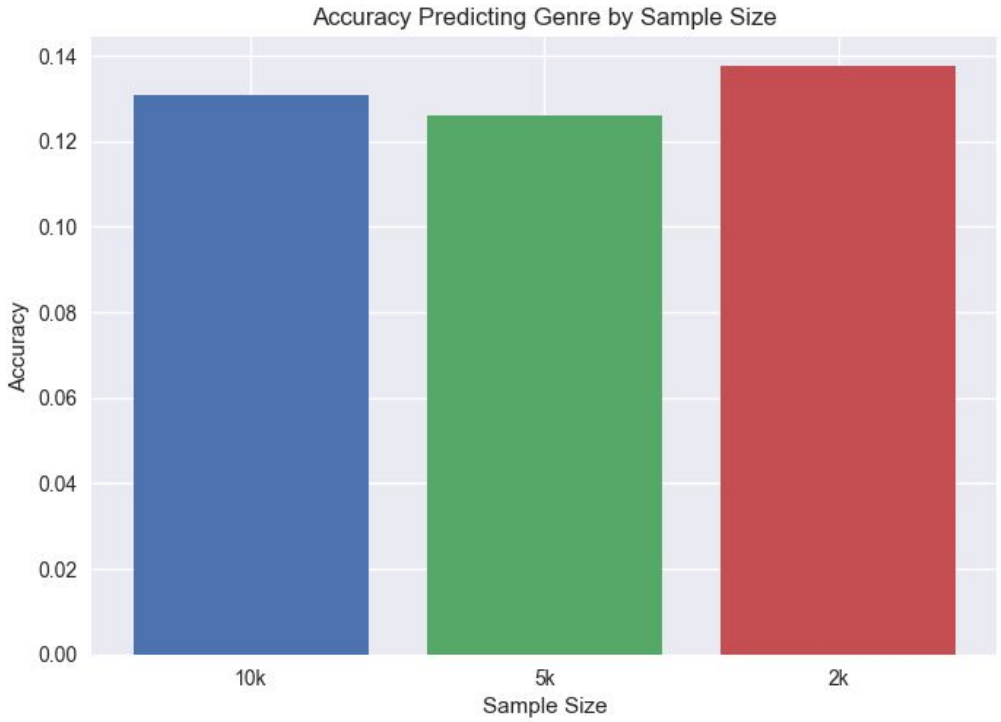
We train the genre prediction model with samples with all genres visible (27 classes), all genres visible without any other attributes, 10 genres visible, 5 genres visible, 2 genres visible, and no genres visible (but including attributes like actorsExperience).

Genre reconstruction is also performed using different masks over the training set. We experiment with the minimum information needed to be provided to the training algorithm in order to produce a model that can accurately recreate the film's original genres. We also test if we can train a model to recreate all of the the film's genre designations only with or only without knowing some of the film's genres.

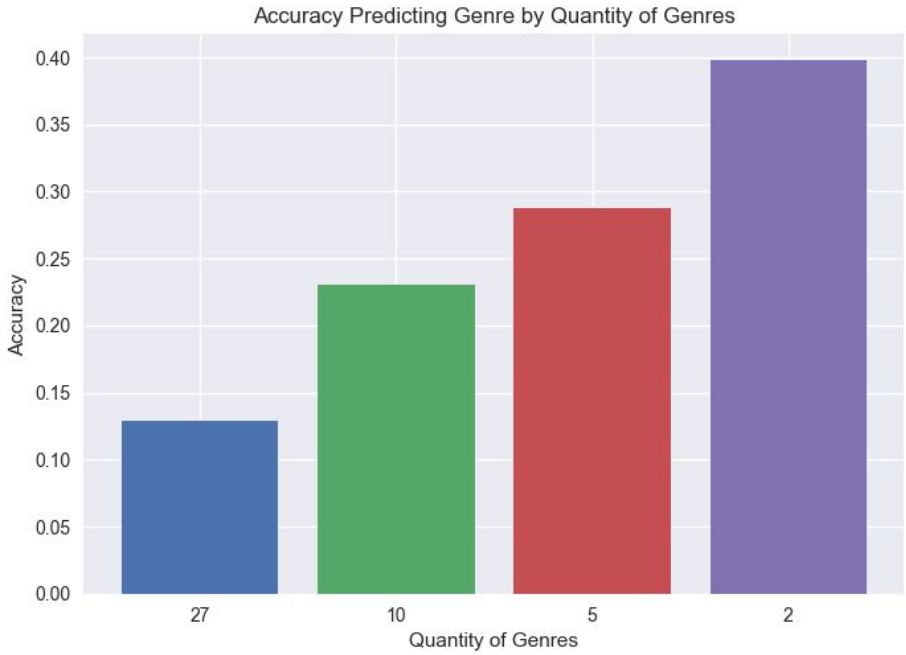
We train the genre prediction model with samples with all genres visible (27 classes), all genres visible without any other attributes, 10 genres visible, 5 genres visible, 2 genres visible, and no genres visible.

Results and Conclusion

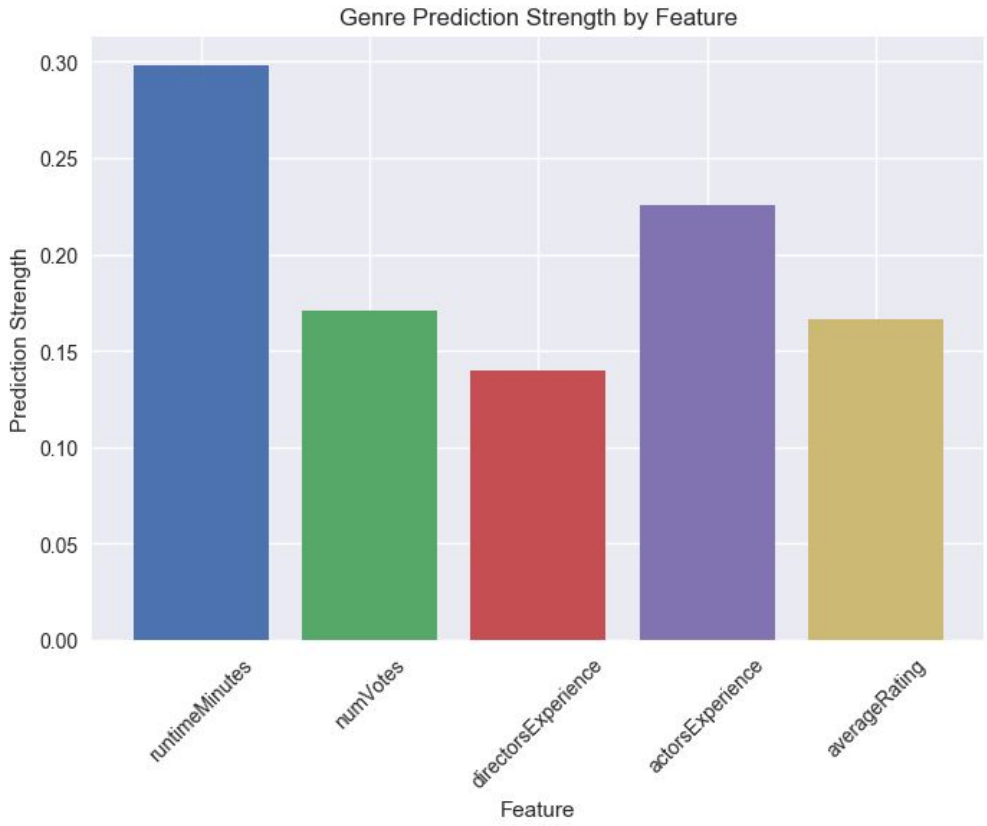
Genre Reconstruction



The number of samples in the training set of the predictor did not seem to meaningfully affect the accuracy of the model. It is possible that with sample sizes this large, there is little meaningful difference between the 10k, 5k, and 2k sample sizes. The drop in performance of the model trained on 5k samples could be attributed to randomness or possibly overfitting which could have been smoothed out in the 10k model.

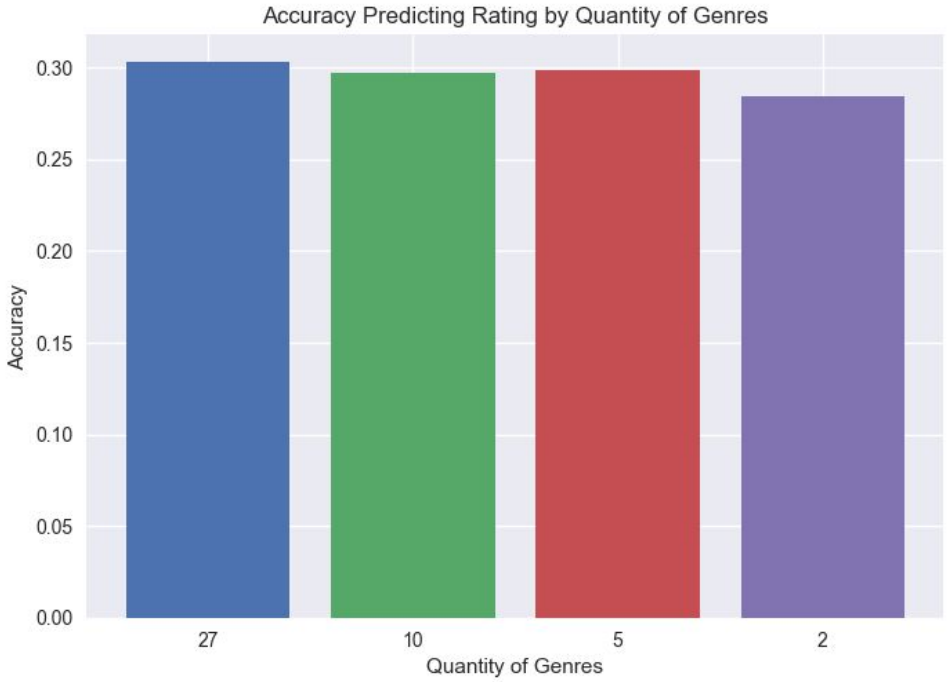


As expected, the fewer genres the model had to predict, the more accurate it became.

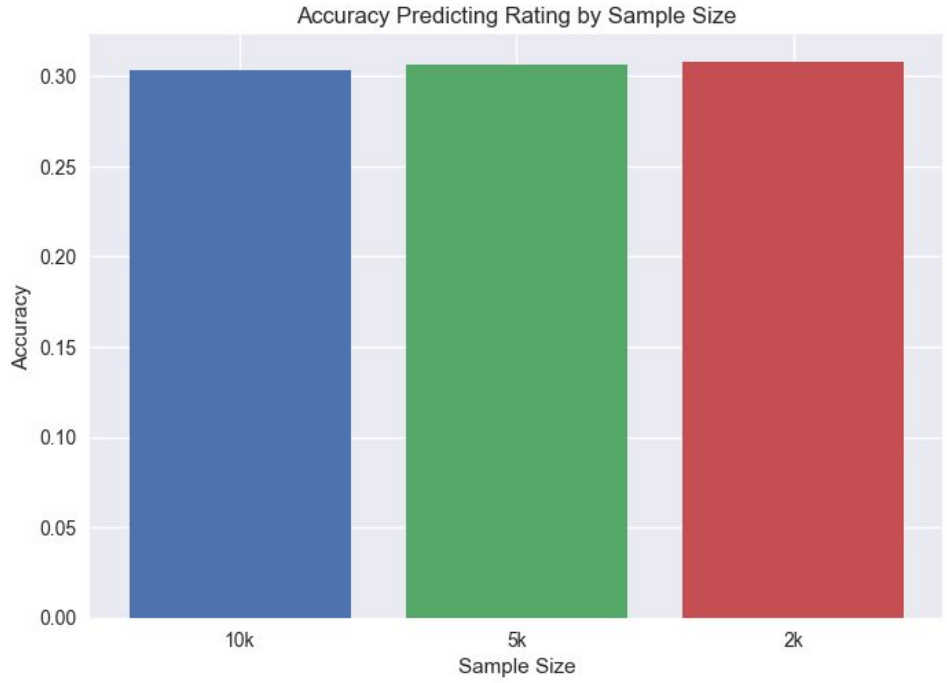


The factors which affect reconstruction of genres the most seemed to be the actors' relative experience and the length of the movie. Number of votes, average rating and the directors' experience were seemingly too generic in comparison to provide meaningful direction.

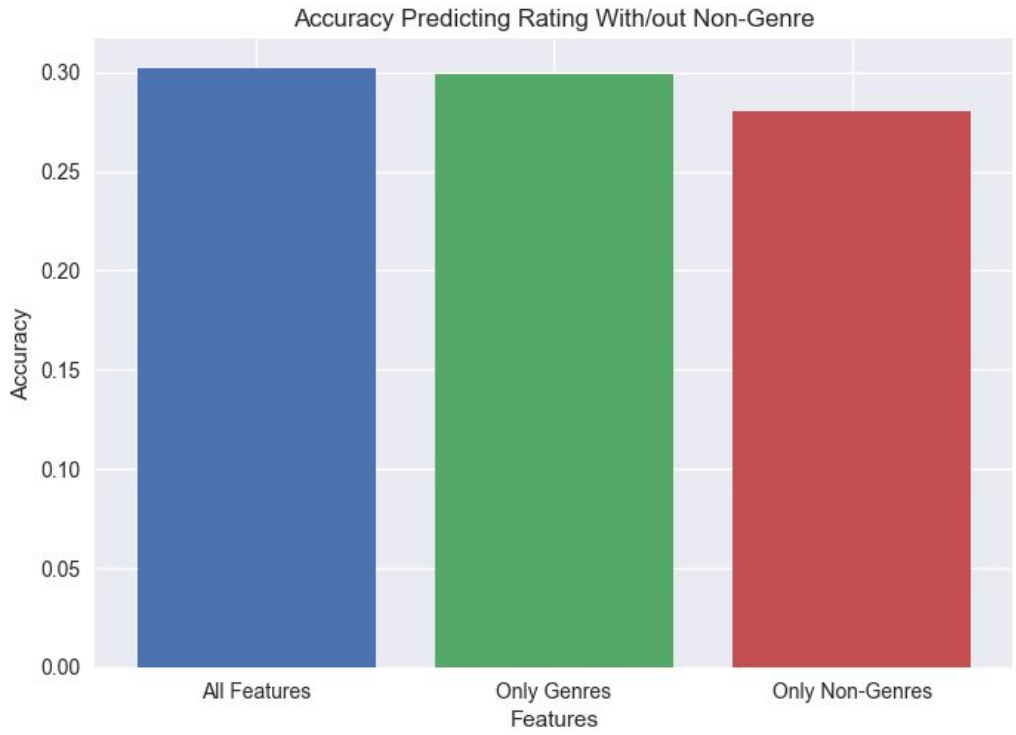
Rating Prediction



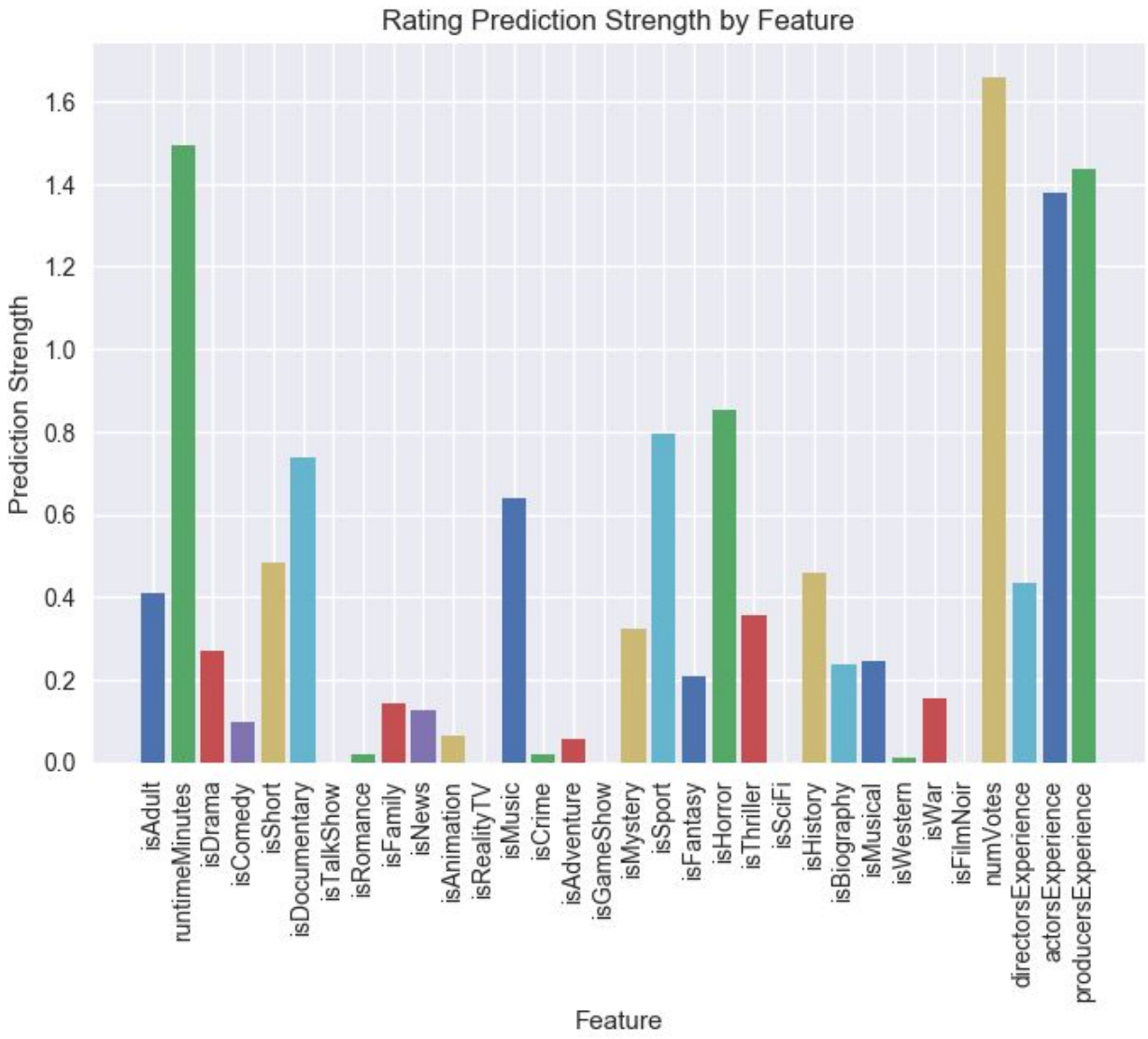
As the model was provided fewer genres to predict ratings with, it's accuracy began to dip slightly, although not incredibly significantly.



Once again, training sample size did not meaningfully affect the ratings classifier, aside from some variance.



Similar to the *Accuracy Predicting Rating by Quantity of Genres* graph, we observe that given all the features the accuracy is around 30%, however as genres are stripped away completely we reach approximately 27.5%. Leaving only genres and not experience levels, runtime, or other factors barely removed any accuracy.



The largest effects on a film's rating appeared to be the number of votes the film received, as well as the runtime of the film. The producer's experience and actors' experience also had a large effect on the film's reception. On the other hand, a film's genres, such as romance, crime, etc did not seem to affect the film's reception as significantly. Despite this, we noticed some genres, like Documentaries, Horror, and Sport movies, had more predictive power than others. Surprisingly, Director experience was a much less important feature than actor or producer experience. It is possible that the small number of well known directors did not increase the predictive strength enough on average to account for the huge number of less experienced directors

. Linear regression was unable to produce a ratings classification model that exceeded an accuracy that exceeded 0.35. This is an indication that the film's ratings may not all have been linearly separable using some of the features.