Movie Endurance

Objective: To build a model that predicts the current popularity of a movie as a function of variables known at its time of release and also some other parameters of the movie that are generally available. Also, we will try to reason the functionality of the models, perform sniff tests on the model.

Introduction and the Dataset Considered: **Popularity follows the Power Law**. The short answer is network effects and the positive feedback loops. These concepts are discussed in the book "Networks, Crowds, and Markets: Reasoning about a Highly Connected World". They explain the reason for Popularity following the Power Law by presenting that **Popularity is a Network phenomenon**. If we think of it, it makes sense because a movie that is already popular is watched/followed by more people and who in turn will recommend it to more people, thus leading to an exponential increase in popularity.

We have used IMDB dataset available as a downloadable dataset from IMDB. But it lacked some crucial information which was required for our modelling. We considered around 160,000 movies after performing basic cleaning (like considering only US region movies). We scraped the IMDB website for these parameters using **Scrapy**. The parameters we extracted by scraping are Movie Budget, Opening Weekend Gross, Gross USA, WorldWide Gross, Metascore, Popularity, Color, Sound Mix.

Data Preprocessing: Of the 160,000 movies considered and scraped from IMDB, around 80,000 movies belonged to different industries (movies with negligible screening in USA, very sparse information and the budget, gross figures in regional currencies). IMDB provides **Popularity metric, which is the popularity rank for the movies based on the number of page visits of a movie updated on top of pre-existing popularity metric on a weekly basis**. Of the 80,000 movies, 4,000 movies have the popularity rank which are scraped. The other movies for which Popularity metric isn't provided are the least popular ones.

As popularity follows the Power law, the least popular ones usually correspond to the tail of the Power Law (Yellow region in the picture). Thus, the popularity rank, same as the popularity rank of the least popular labelled entity is applied to the least popular movies for which popularity metric isn't provided.

Movie Budget, Opening Weekend Gross, Gross USA, WorldWide Gross are adjusted according to the inflation of the year, in which the movie is released. For this, **Consumer Price Index** (**CPI**) dataset CPIAUCNS provided by Federal Reserve Economic Data is used. This dataset provides CPI index from as early as 1913. The index is set to give money value in terms of 2019 money.

Genres and Sound Mix are analysed and accordingly top K genres and Sound Mix are considered based on frequencies and are one-hot encoded. By this approach, a movie belonging to multiple genres would have 1 set

for the associated genres and thus able to capture all the genre information. Same is the case for Sound Mix. Infrequent genres, sound mix types are put into Not_in_K_selected bucket.

The missing values in Movie Budget, Opening Weekend Gross, Gross USA, WorldWide Gross are imputed using Iterative Imputation. It is done using Round-Robin style linear regression, where each missing value is predicted using other feature values of the data instance. Initialization is done with Mean of the feature and then iteratively estimated using linear regression technique fitted on other features of the same instance.

Features Considered: 'AspectRatio', 'Budget', 'Color', 'Cumulative Worldwide Gross', 'Gross USA', 'Metascore', 'Opening Weekend USA', 'Oscar Nominations', 'Oscar Wins', 'Other Award Nominations', 'Other Award Wins', 'Sound Mix', 'startYear', 'runtimeMinutes', 'genres', 'averageRating', 'numVotes'. Some of these are one-hot encoded for the reasons explained previously.

Prediction Models and Sniff tests for the Models

It makes sense to develop a regression/classification model when we try to predict something tangible and can be directly evaluated, like a house price, temperature estimation or a Benign/Malignant tumor classification. Popularity is a very abstract concept and there aren't standards to measure it . Thus, it makes more sense to have a ranking system based on a scoring function. Thus, we developed two hypotheses for the ideal scoring function. We present them below.

Sniff test and Analysis: (Please refer to the correlation table figures following this section)

1. Number of Oscar Wins is more correlated with Popularity than the Number of Oscar correlations. This is inline with the general view that, a movie that won an Oscar award would have more endurance than Oscar Nominated movie. *If we observe the same table, we notice that Other Nominations (Other Award Nominations) and Other Wins (Other Award Wins) are more highly correlated with Popularity. This is probably because a movie that is nominated and won various awards across the globe, would become popular in that region of the world, thus more popularity.*
2. numVotes (Number of Votes a movie received on IMDB) is the highest correlated feature with the Popularity as expected. numVotes is a direct reflection of number of People who followed the movie in the past few years.
3. Budget and Gross associated features are highly correlated with the Popularity as expected.
4. From the correlations of the Sound Mix associated features, it can be observed that Silent, Mono, Stereo are negatively correlated and Dolby and other modern Sound Mix highly positively correlated with the Popularity. These correlations clearly mark, how the technology shift caused fall in endurance of the movies with old technology. Same is the case for Color (Color_Color,

Color_Black and White) which can be observed at the end of correlation table.

5. Even though, genre alone isn't highly correlated with the Popularity of a movie, it can be observed that Western, Documentary, War, Musical are negatively correlated with the Popularity. This implies that, what used to be money-minting genres (Western, War) have gone out of style now.
6. Aspect Ratios of recent times (2.35 : 1, 2.39 : 1) are highly correlated with Popularity. This might not be a causality factor and probably just that movies of recent times are more popular than older movies.

Metascore	0.048982
Oscar Nominations	0.152971
Oscar Wins	0.200226
Other Nominations	0.468061
Other Wins	0.354776
startYear	0.112856
runtimeMinutes	0.023146
averageRating	0.079518
numVotes	0.661108
Budget_inflation_adjusted	0.514419
Cumulative Worldwide Gross_inflation_adjusted	0.506084
Gross USA_inflation_adjusted	0.450985
Opening Weekend USA_inflation_adjusted	0.536319

Sound Mix_Silent	-0.034765
Sound Mix_Mono	-0.095858
Sound Mix_Dolby	0.322284
Sound Mix_Stereo	-0.000811
Sound Mix_DTS	0.355143
Sound Mix_Dolby Digital	0.297960
Sound Mix_SDDS	0.395683
genres_Biography	0.038024
genres_Adventure	0.089498
genres_Drama	0.019262
genres_Western	-0.027567
genres_Comedy	0.020735
genres_Action	0.080648

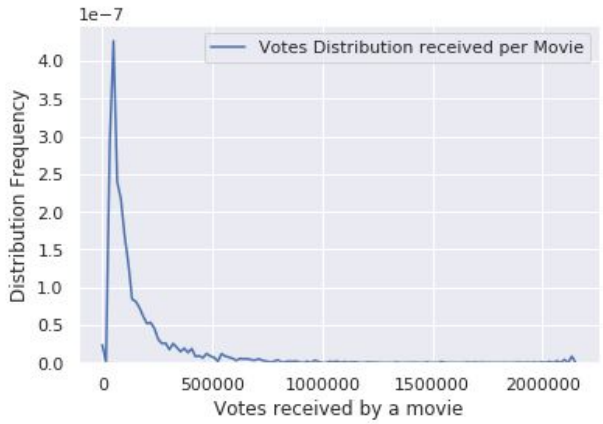
genres_Crime	0.032297
genres_Mystery	0.050941
genres_Documentary	-0.073459
genres_Fantasy	0.079692
genres_War	-0.006995
genres_Musical	-0.013664
genres_Romance	0.005550
genres_Horror	0.050294
genres_History	0.005115
genres_Family	0.021808
genres_Sci-Fi	0.075494
genres_Thriller	0.059808
genres_Music	-0.016093
genres_Adult	-0.034954
AspectRatio_1.20 : 1	-0.014972
AspectRatio_1.33 : 1	-0.038349

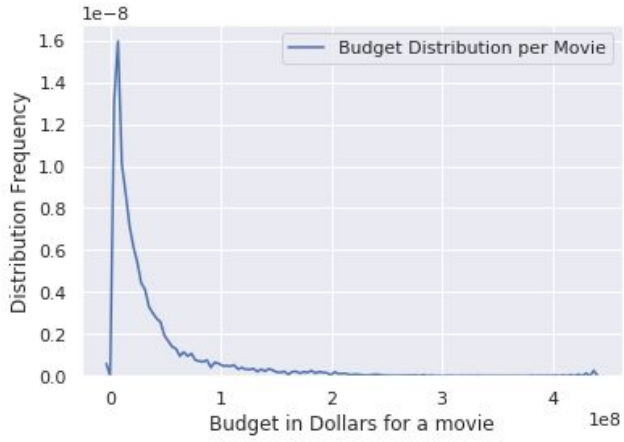
AspectRatio_1.37 : 1	-0.066320
AspectRatio_1.66 : 1	-0.011373
AspectRatio_1.78 : 1	-0.022643
AspectRatio_1.78 : 1 / (high definition)	-0.015159
AspectRatio_1.85 : 1	0.093684
AspectRatio_16:9 HD	-0.025680
AspectRatio_2.35 : 1	0.113262
AspectRatio_2.39 : 1	0.422167
AspectRatio_Not_in_selected_categories	-0.137816
Color_Black and White	-0.082268
Color_Color	0.082268

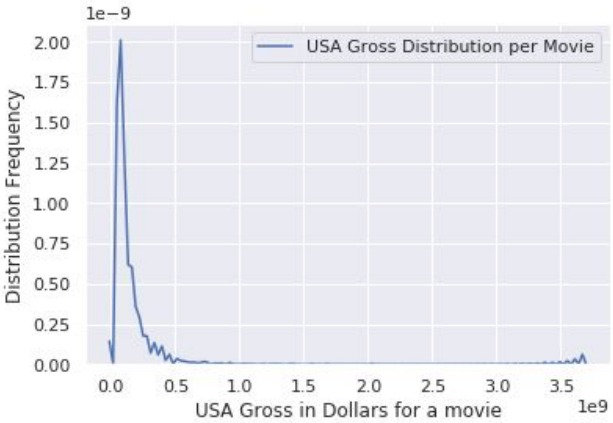
Model 1: We evaluated the correlation for all the features against the Popularity Metric.

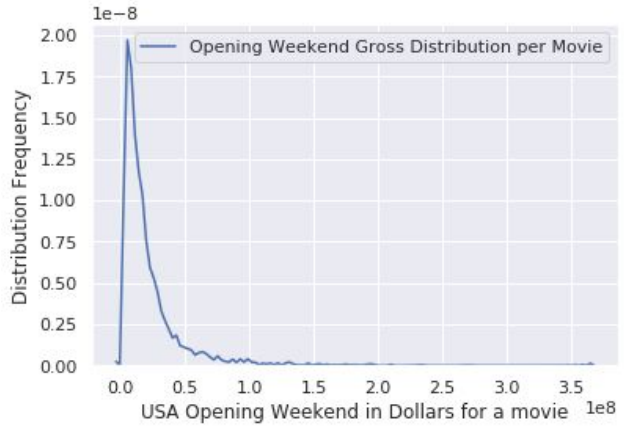
All features which had more than 0.1 correlation with the Popularity Metric are considered for the modelling. These features are Oscar Nominations, Oscar Wins, Other Award Nominations, Other Award Wins, StartYear (Year of Release), NumVotes (Votes for the movie on IMDB), Budget and Gross associated features, Sound Mix - Dolby, DTS, Dolby Digital, SDDS; Aspect Ratio - 2.35 : 1, 2.39 : 1.

We have observed that NumVotes, Budget and Gross Associated features have a Power Law Distribution.









Estimation of a Power Law governed target attribute (Popularity) by highly correlated other Power Law

governed features makes sense as there should be some features in the scoring function that scale proportional to the target variable. The Scoring function is the Summation over scaled highly correlated features multiplied by respective correlations with the Target variable (Popularity).

$$H = \sum_{i=0}^{n} \alpha_i f_i$$

α_i is the correlation of feature i, associated with the target variable. f_i is the scaled value of the highly correlated feature considered for modelling. As Power Law governed features are involved, we made sure that the scaling is done just by dividing the maximum value of the feature among the data instances considered and not fit the Gaussian Distribution. The reason for multiplying scaled feature contributions with their associated correlation values is that, a highly correlated feature should have more say in deciding the scoring function value than less correlated feature. **The correlations (α_i) evaluated on training data are fixed for testing purposes.**

The Scoring function evaluated this way, showed a correlation of 0.72 with Target metric (Popularity), which is very desirable.

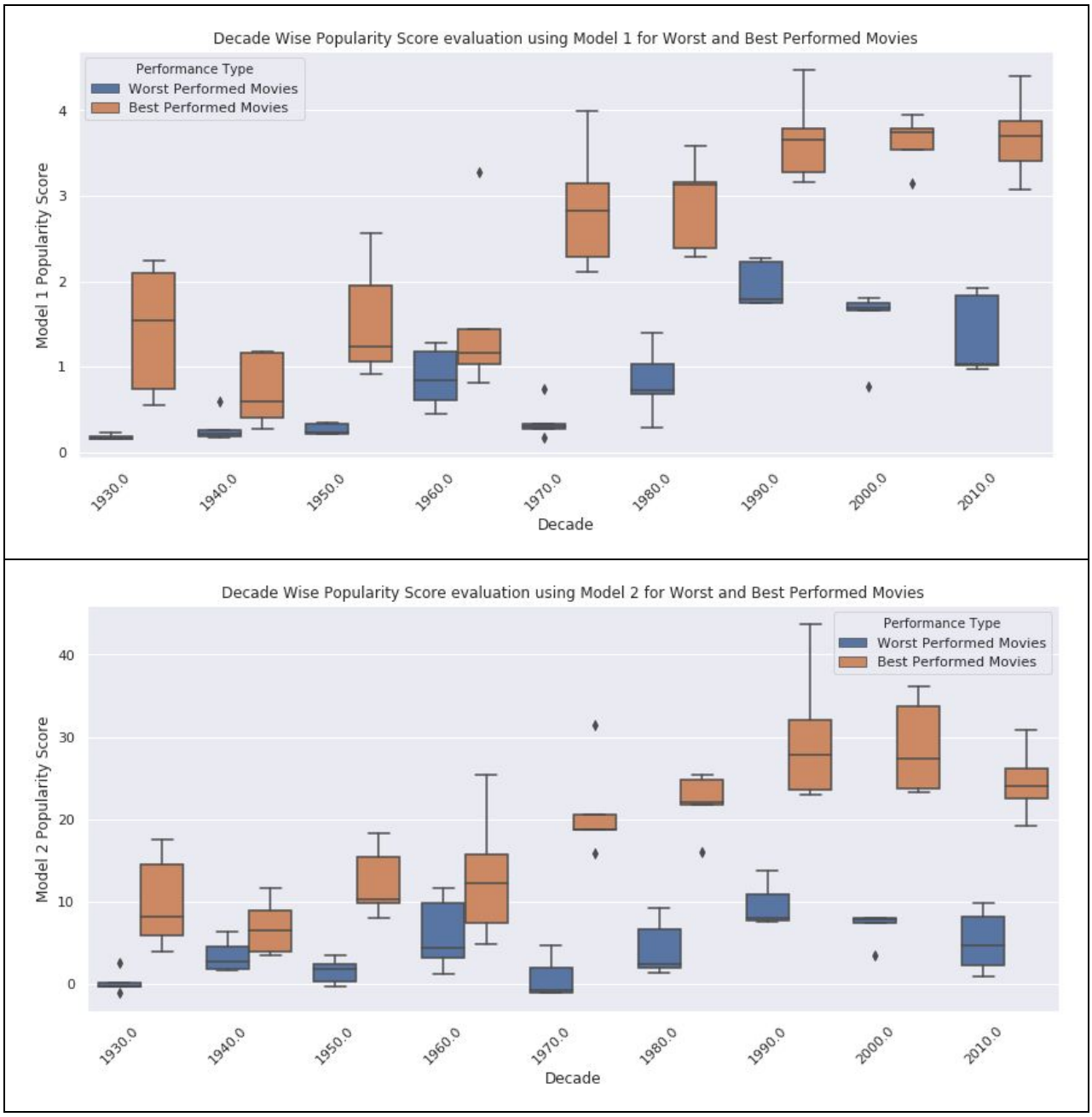
Model 2:

Logarithm of the features, which are observed to follow Power Law in Model 1, are used instead of the features themselves for the Scoring function evaluation. Everything else is the same. The logarithm of the Power Law governed features are considerably less correlated (shift from what used to be of order 0.5 to 0.2) with the Popularity metric. But the overall Scoring function has 0.79 correlation with the Popularity Metric.

Decade Wise Analysis of Movies

For decade wise analysis, we considered Performance of all movies grouped by Decade. We evaluated Performance to be Gross - Budget. We also considered Gross/Budget, but the issue with this is that a very low budget movie if received well by decent chunk audience would push its performance to a very high value. **We wanted Performance to be proportional to the impact and attention garnered thus we have gone with Gross - Budget and not Gross/Budget.** We evaluated both of our Models against 5 Best and Worst Performed movies of the respective decades.

The difference in Popularity Scores as evaluated by our models between Best and Worst Performed Movies DecadeWise is very clear for both the models.



1. The Money associated with the movies (Budget, Gross etc considered as features) and NumberOfVotes for recent movies is very high when compared to those of movies before the 1970s. As we have taken **logarithm** of these features, their say is kind of moderated by Popularity Score in Model 2 as compared to their say in Model 1. Thus, Worst Performed Movies have over the period Similar Popularity Scores in case of Model 2 which isn't the case in Model 1.
2. The Popularity Scores for Best Performed Movies from 1970 are significantly higher than the Best Performed Movies of previous decades. This can be attributed to Technology shifts captured by our Models like transition from Black and White to Color that happened from 1930s to the 1960s. This

is captured not just with technology associated features considered previously but might be in turn captured by Number of Votes, a movie received.

3. The high Popularity Scores of Worst Performed Movies in 1990 decade might be due to the reason that even though they didn't do well at the box office, they might be popular now as compared to best performed movies of older decades, which is captured by other attributes apart from money associated attributes we considered for modelling the popularity.

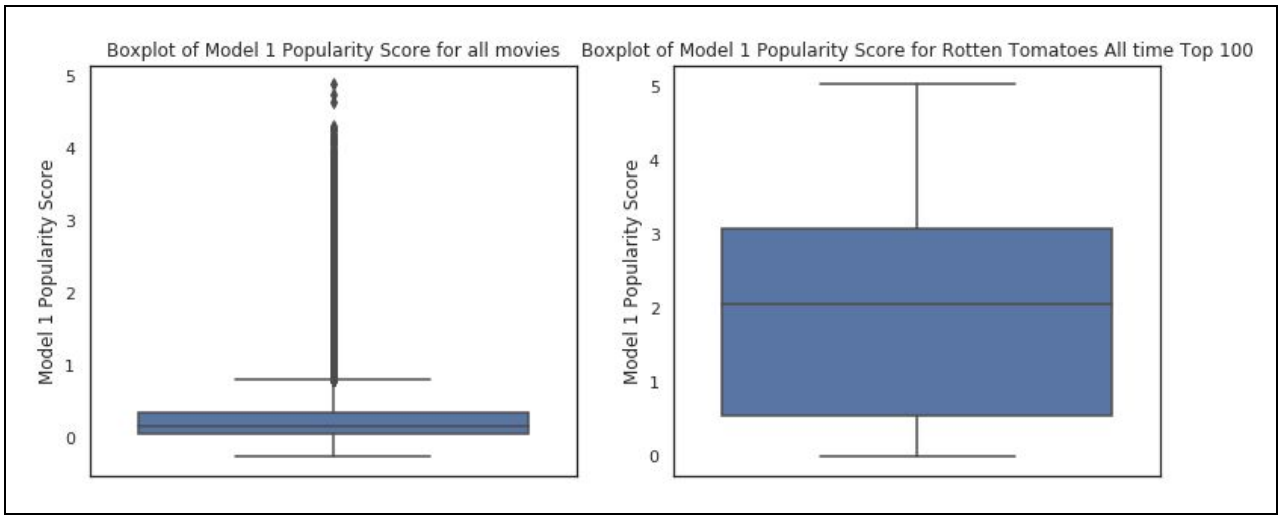
Model Validation

We have considered Rotten Tomatoes Top 100 Movies of all time for the testing of our Models. We have evaluated our model on 2 sets of movies for comparison of the distribution of scores.

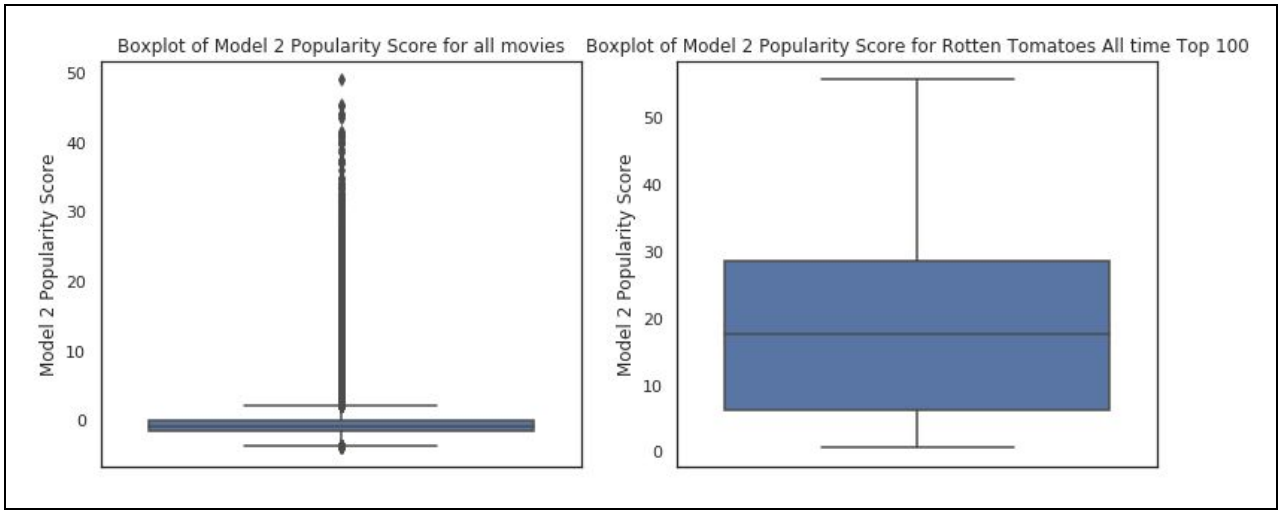
1. All movies except the Rotten Tomatoes Top 100 Movies
2. Rotten Tomatoes Top 100 Movies separately.

11 of the 100 movies are dropped in the preprocessing steps because of incomplete data for these movies on IMDB. The difference in the scale of Scoring functions of both models and the negative scores can be attributed to the difference in the way they are normalized as explained previously.

Model 1



Model 2



Summary

	Popularity Score Model 1 for All Movies Except Rotten Tomatoes Top 100 Movies	Popularity Score Model 1 Rotten Tomatoes Top 100 Movies	Popularity Score Model 2 for All Movies Except Rotten Tomatoes Top 100 Movies	Popularity Score Model 2 Rotten Tomatoes Top 100 Movies
Movie Count	85633	89	85633	89
Mean	0.31	1.95	-0.02	18.99
Std. Deviation	0.49	1.35	3.25	14.25
Min	-0.26	-0.00	-3.96	0.61
25%	0.06	0.55	-1.4	6.23
50%	0.16	2.05	-0.81	17.56
75%	0.35	3.08	0.05	28.51
Max	4.89	5.03	49.06	55.70

Both the models did very well in scoring vast segment of the popular movies well above the normal movies. Hence, we can conclude by saying that both the models did a very good job in current popularity evaluation of the movies.