MOVIE GROSS PREDICTION

Nikita Redkar
Information Technology
Rutgers University
ndr33@scarletmail.rutgers.edu

Ajinkya Rasam
Information Technology
Rutgers University
ar1341@scarletmail.rutgers.edu

Aditya Sawant
Information Technology
Rutgers University
ads238@scarletmail.rutgers.edu

Abstract- Important factors influencing movie gross are genre, directors, actors, production house that makes the movie but everybody misses the sentiment of people which play an important role in determining the success of a movie. In this report we present a model which will predict the gross income of any movie while taking into consideration attributes which represent people's views before it is released on the box office. Using this model we were able to predict gross income of movies which were released in the first week of May 2017.

I. INTRODUCTION

Movie industry is a high earning business industry. Every year the amount of revenue earned by the industry grows substantially. With the huge budget investment for movies it is becoming important for movie producers to know what would be their gross earned after the release. Hence it is important to forecast pre-release movie grosses, because investors in the movie market want to make
wise decisions [1]. The goal of our project is to predict gross of movies that would be released in the first week of May 17.

Predicting the gross of movies solely depends on parameters available before the release of movie. Our approach to the project includes Data collection, Data Analysis, Prediction and Analysis. The major task for our project was to gather data of movies from different sources and to clean and merge the data together. For data extraction we used API's from different websites like IMDB, OMDB and Number's. Merged and cleaned the data in Python. Important parameters that would affect the gross of the movies can be directors, actors, cast, the budget spent on the movie, promotions before the release but other than these there are also parameters like imdbVotes, number of users who critic the movie before the release which tell us about the sentiments of general public who see the trailer and rate the movie. Using correlation testing and Principal Component Analysis we were able to find out which parameters could help predict gross of movie accurately before its release. For prediction we used multiple regression and ridge regression models.

The contents of this paper are organized as follows. First, we will review some related work briefly. Second, we will describe the movie data sources, and give correlation analysis. We then set up different models for prediction and evaluate their performance. Finally, we conclude that our model which considers user reviews and votes predicts the gross accurately.

II. RELATED WORK

Different people use different approach for prediction. Natural language processing community uses movie reviews as a domain for sentiment analysis. To understand emotions from the review comment's they use polarity and subjectivity of the comment to classify the movie [2] [3]. But this technique has not been used to predict the gross of movie before the release.

Some advanced models divided the data into high grossing and low grossing movies based on a threshold and then performed regression tests separately [4]. We tried to use this technique but this makes our problem a classification example and would not help us to predict an amount for gross which can only be found out by regression model.

III. DATA COLLECTION

We first extracted our raw movie data from IMDB website. The data obtained included 5043 records and had 28 attributes which included directors names, actor names, number of critic reviews, duration of movie, director Facebook likes, actors Facebook likes, gross, budget, genre, movie title, number of votes users, imdb link, number of users for review, imdbVotes and score, Awards won by the cast of movie, Production house and some more parameters.

Since we needed release date of the movies in our data we extracted that data from OMDB database. OMDB provides an API key to connect to it which in return gets data from IMDB. By providing the imdb id we can extract the movie data for that id. After collecting the required data we removed the unwanted columns from the data and cleaned the Excel data which included unwanted characters. After merging the data we had 43 attributes also since genre was a categorical variable we had to convert it to binary to check individual effect of genre on gross of movie.

Since data collection is the most important part of analysis we spent most of our time cleaning the data and removing the null values in Python. As we had data from year 1916 to 2016 initially we started our analysis steps considering all the data but then realized that the past years 1916 data was prune to inflation factors and this data affected our gross model prediction and we were not getting accurate results as the movies in 1916 did not make much gross as they used to have very less budget in those days for movies as compared to the movies released in 20's. So for our analysis we considered data from year 2010-2016 of recent years with 1510 data rows.

IV. DATA ANALYSIS

Our first step in Analysis started with correlation test in Python. Correlation test helps to find out which parameters are closely related to our predictor variable which is Gross. The value of correlation coefficient varies between +1 and -1 [5]. When the value of the correlation coefficient lies around ± 1, then it is said to be a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. The positive sign + indicates a positive relationship between the variables and negative sign – indicates a negative relationship between the variables. The sign of correlation is important. We plotted correlation matrix using heat map in Python Fig 1.

Also Table 1. depicts the correlation coefficients of variables with gross. Both the heat map and coefficients shows that Gross has very strong correlation with num_critic_for_reviews, num_voted_users, num_user_for_reviews, budget, movie_facebook_likes, imdbVotes. Genre-Adventure, Action and Sci-Fi had good correlation with gross of movie compared to other genres that is these genre movies were amongst the top gross earnings. This can be seen in Figure 2.

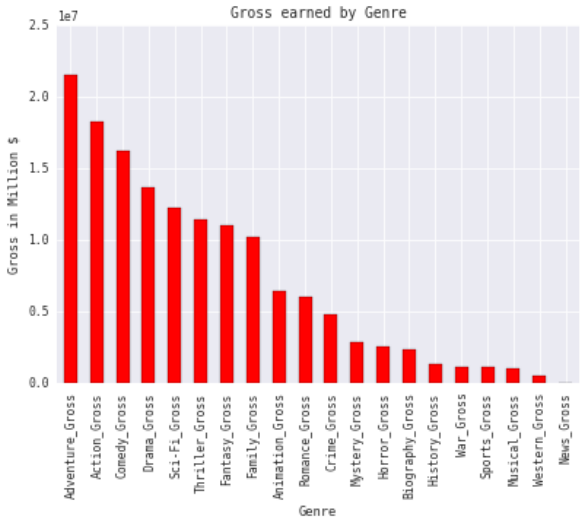

Figure 2: Top Genres and the gross earned by them in Million Dollars

Since budget and gross had a very strong correlation we plotted a scatter plot Figure 3. to see their relation. It can be seen that they have a positive correlation and as budget for a movie increases their gross earned also increases. This is true because more the money spent on a movie for cast, promotions, directors, actors more the movie is hit on Box office and they earn high gross.

To understand the growth of gross and budget in movie industry plotted the trend over the years 2010-2016 in Figure 4. It is clear that initially in 2010 when the budget of the movies was less gross earned was also less which went on

increasing in successive years later.

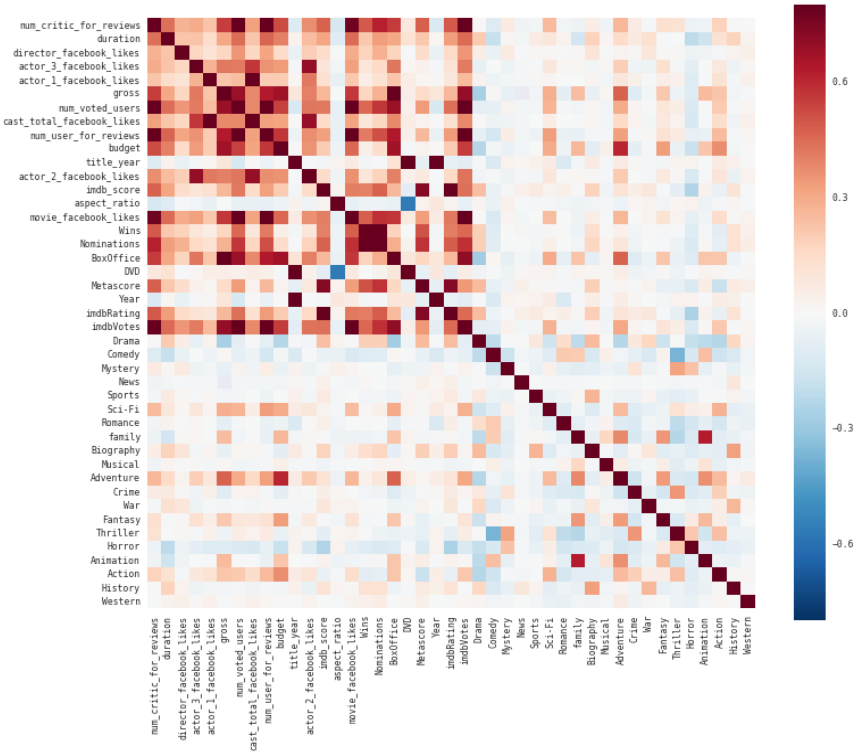

Figure 1: Heat map depicting correlation matrix of all variables. Dark red indicates strong positive correlation and dark blue indicates strong negative correlation

Variables	Correlation coefficient with gross
num_critic_for_reviews	0.553949
duration	0.309045
director_facebook_likes	0.171733
actor_3_facebook_likes	0.422479
actor_1_facebook_likes	0.221833
gross	1.000000
num_voted_users	0.689425
cast_total_facebook_likes	0.383249
num_user_for_reviews	0.634466
budget	0.676046
title_year	0.093572
actor_2_facebook_likes	0.400591
imdb_score	0.265839
aspect_ratio	-0.004301
movie_facebook_likes	0.564660
Wins	0.169877
Nominations	0.279641
BoxOffice	0.989214
DVD	0.051452
Metascore	0.155552
Year	0.094980
imdbRating	0.259830
imdbVotes	0.701152
Drama	-0.263602
Comedy	0.012159
Mystery	-0.063333
Sports	-0.030572
Sci-Fi	0.292980
Romance	-0.075229
family	0.247542
Biography	-0.058106
Musical	0.017175
Adventure	0.474752
Crime	-0.095443
War	-0.023350
Fantasy	0.216047
Thriller	-0.045395
Horror	-0.093921
Animation	0.251526
Action	0.224402
History	-0.026445
Western	-0.006566

Table 1: Correlation coefficients with gross

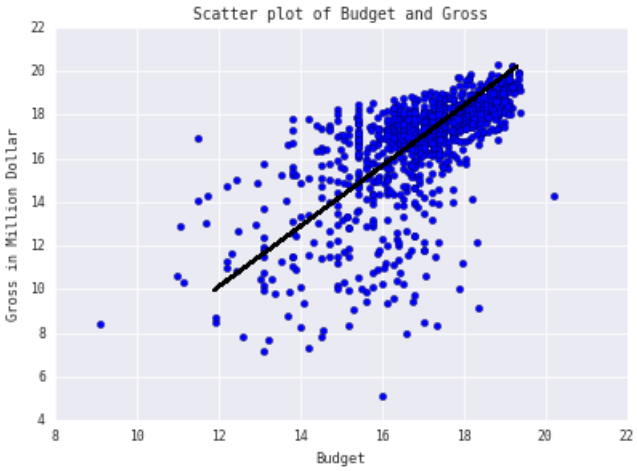

Figure 3: Scatter plot of Budget and Gross

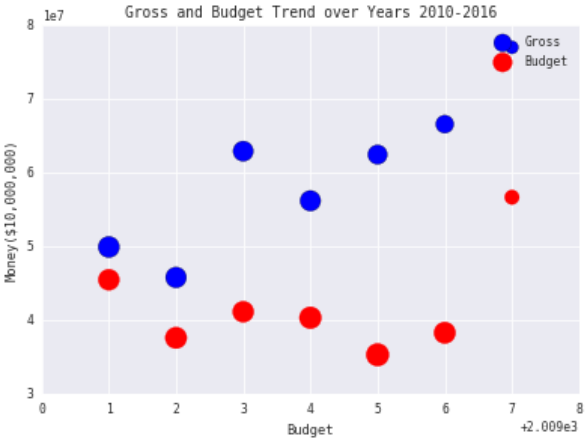

Figure 4: Gross and Budget trend over years 2010-2016

Directors, Production house also play a very significant role in determining the gross of movies. In order to understand who were the top directors who earned maximum gross profit through movies we plotted a bar graph of directors and their gross in Figure 5. From 2010-2016 these are the top directors in movie industry who earn maximum gross. Joss Whedon, Francis Lawrence were amongst the top directors with highest gross earned.

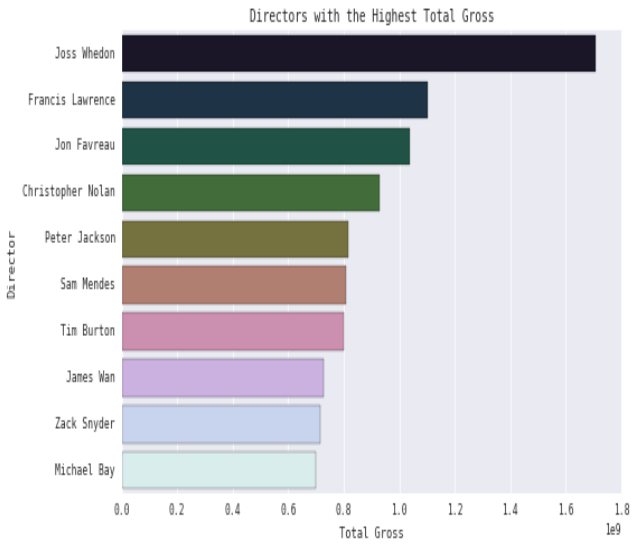

Figure 5: Directors with highest total gross

Figure 6 shows the box plot of directors and their average gross distribution. Box plot displays the distribution of data based on the minimum, first quartile, median, third quartile, and maximum. It is standardized way of displaying the distribution of data.

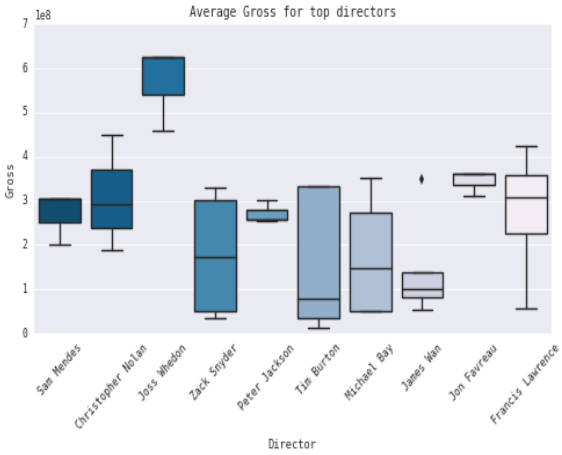

Figure 6: Box plot of Average Gross earned by top directors

Figure 7 depicts the production house that earns maximum gross. Big names like Walt Disney, Universal pictures, Warner Brothers who have very high budget earn high gross because of their movie directions and actors.

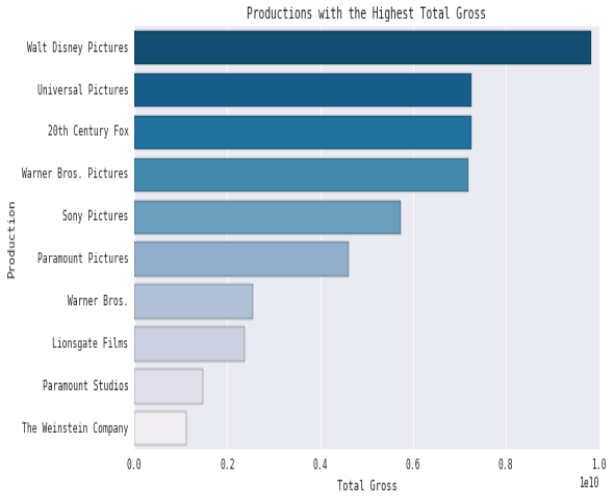

Figure 7: Productions with Highest Total gross

From all the above plots it is clear that gross is dependent on many factors. For predicting gross based on Directors and Production houses which are categorical variables we need to use feature vector technique which assigns weights to each attribute. Feature vectors are equivalent to the vectors of explanatory variables used in statistical procedures such as linear regression. Feature vectors are often combined with weights using a dot product in order to construct a linear predictor function that is used to determine a score for making a prediction [6]. But in this paper we have concentrated only on user rating factors

V. FEATURE SELECTION- Principal Component Analysis

Principal Component Analysis is exploratory or descriptive method of data analysis. It is used for reducing the original variables into orthogonal, synthesized (non-correlated) variables. It's often used to make data easy to explore and visualize. [7] The number of principal components is less than or equal to the smaller of the number of original variables or the number of observations.

We used this concept to reduce the number of attributes or features required to train our regression model. Figure 8 plot shows change in cumulative explained variance versus the number of components. We can observe that more than 85% of variance in the dataset can be explained by 3 components which is a good indication and now we can reduce the dimensionality of our data.

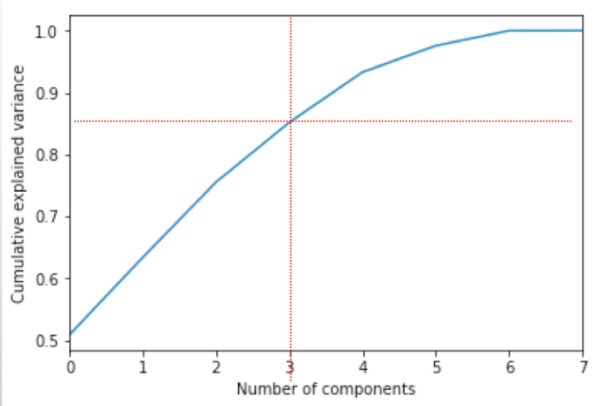

Figure 8: Number of components versus cumulative explained variance

Further, to determine feature's importance or to select the features which can be used to train our model, we utilized Sklearn's 'Feature Selection' library [8]. This library can be used for feature selection/dimensionality reduction on sample sets, either to improve estimator's accuracy scores or to boost their performance on very high-dimensional datasets. It acts like a meta-transformer on any estimator which has coef_ attribute after fitting. We used Random Forest estimator for this purpose. Figure 9 shows importance of each feature. 'IMDB_votes' being the most important feature followed by the rest.

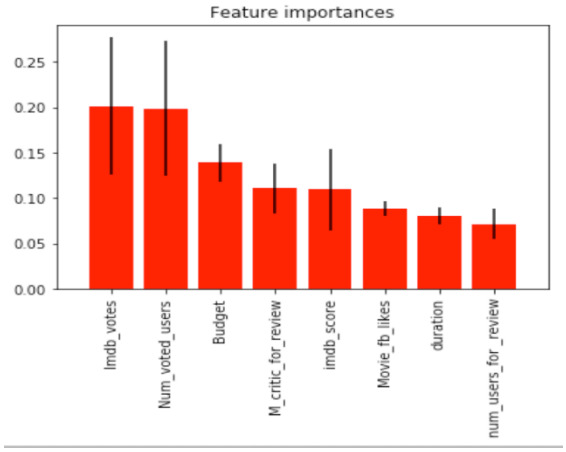

Figure 9: Important features and their importance

From Figure 9 it is clear that imdb_votes, number of voted users, budget are the three important features for gross prediction amongst the rest features so we have used these features in our multiple regression model for prediction.

VI. PREDICTION

Two important modeling methodologies used in this paper are Multiple linear regression using OLS (Ordinary Least square) and Ridge regression. Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y [9]. Formally, the model for multiple linear regression, given n observations, is given by:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots \beta_p x_{ip} + \varepsilon_i \text{ for } i = 1,2, \dots n.$$

For our model first we have used 3 features (imdb_votes, number of voted users, budget) suggested by PCA for regression and predicted Gross. Figure 10 shows the regression model summary results. The model equation to predict gross is given by OLS is:

Gross= (766.7163)* imdb_votes+ (0.6081)* budget+(-535.1927)*num_voted_users+ (-1.826e+06)

OLS Regression Results

Dep. Variable:	gross	R-squared:	0.629
Model:	OLS	Adj. R-squared:	0.628
Method:	Least Squares	F-statistic:	415.1
Date:	Fri, 05 May 2017	Prob (F-statistic):	1.57e-157
Time:	22:01:57	Log-Likelihood:	-14152.
No. Observations:	737	AIC:	2.831e+04
Df Residuals:	733	BIC:	2.833e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
const	-1.826e+06	2.73e+06	-0.669	0.503	-7.18e+06	3.53e+06
num_voted_users	-535.1927	145.267	-3.684	0.000	-820.383	-250.003
budget	0.6081	0.038	16.117	0.000	0.534	0.682
imdb_votes	766.7163	137.403	5.580	0.000	496.966	1036.466

Omnibus:	297.448	Durbin-Watson:	2.109
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5217.942
Skew:	1.347	Prob(JB):	0.00
Kurtosis:	15.754	Cond. No.	1.12e+08

Figure 10: OLS linear regression model summary for 3 features

From the above model we found that the conditional number was very high which means there was multicollinearity between the variables. Multicollinearity is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated to each other and this affects the predictor variable. To solve this problem we performed Ridge Regression on the same 3 variables.

Ridge Regression is a remedial measure taken to alleviate multicollinearity amongst regression predictor variables in a model [10]. Often predictor variables used in a regression are highly correlated. When they are, the regression coefficient of any one variable depend on which other predictor variables are included in the model, and which ones are left out. Ridge regression adds a small bias factor to the variables in order to alleviate this problem [10]. Ridge regression model gives us better correlation between the predicted variable and independent variables than OLS linear model.

Ridge regression model equation for 3 features is given by:

Gross= (195.72595505)* imdb_votes+ (0.61467986)* budget+(54.83219198) *num_voted_users+ (837004.633781)

The parameters which we considered for models comparison include Mean Squared Error (MSE), correlation factor, Explained variance score, R- squared value.

The mean squared error (MSE) for a predictor model measures the average of the squares of the errors or deviations that is, the difference between the predictor and what is predicted [11]. MSE is given as:

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2.$$

Where y_i is the true value and $\hat{y}_i$ is predicted value.

The Explained variance score computes the explained variance regression score [12]. The best possible score is 1.0, lower values are worse. Explained variance score is given by:

$$\texttt{explained\_variance}(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}}$$

Where $\hat{y}$ is estimated target output and y is the corresponding (correct) target output and Var is the variance the square of standard deviation.

R-squared is a statistical measure of how close the data are to the fitted regression line. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance [13].

Our both multiple regression and ridge regression model with 3 features gave good results for predicting gross which can be seen in Table 2. But we improvised our model by using one more feature (movie_facebook_likes) which had highest correlation with Gross and replaced number of voted users with

num_critic_for_reviews because our aim was to predict gross of movies which would release in first week of May and for this we should have all 3 features available for prediction but we found out on IMDB that number of voted users is sometimes not available. So next we again ran Multiple Regression and ridge regression on 4 features (imdb_votes, movie_facebook_likes, budget, num_critic_for_reviews). Figure 11 shows the summary of OLS regression model for 4 features and Table 2 gives summary results of both the models.

OLS Regression Results

Dep. Variable:	gross	R-squared:	0.574
Model:	OLS	Adj. R-squared:	0.571
Method:	Least Squares	F-statistic:	246.1
Date:	Fri, 05 May 2017	Prob (F-statistic):	7.44e-134
Time:	22:49:58	Log-Likelihood:	-14125.
No. Observations:	737	AIC:	2.826e+04
Df Residuals:	732	BIC:	2.828e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]	
const	4.391e+06	4.03e+06	1.090	0.276	-3.52e+06	1.23e+07
imdb_votes	237.6858	25.527	9.311	0.000	187.572	287.800
num_critic_for_reviews	-1.952e+04	2.11e+04	-0.926	0.355	-6.09e+04	2.19e+04
budget	0.6024	0.038	16.038	0.000	0.529	0.676
movie_facebook_likes	91.1498	98.826	0.922	0.357	-102.866	285.166

Omnibus:	321.338	Durbin-Watson:	2.015
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6830.028
Skew:	1.440	Prob(JB):	0.00
Kurtosis:	17.633	Cond. No.	1.62e+08

Figure 10: OLS linear regression model summary for 4 features

The model equation for predicting gross with 4 features by OLS is given by:

Gross= (237.6858)* imdb_votes+ (0.6024)* budget+(-1.952e+04)*num_critic_for_reviews+ (91.1498)*movie_facebook_likes+(4.391e+06)

There was a good improvement in MSE and variance score for Ridge regression model with 4 features which can be observed in Table 2. Ridge regression model equation for 4 features is given by:

Gross= (2.11276139e+02)* imdb_votes+ (7.22947947e-01)* budget+(-3.05122849e+04)* num_critic_for_reviews+ (1.78433198e+02)*movie_facebook_likes+(2218633.0302)

	4 Features		3 Features	
	Linear Regression	Ridge Regression	Linear Regression	Ridge Regression
Correlation	-0.083	0.729	-0.001	0.851
MSE	3344671046889592.500	2878388040639787.000	2013627976900174.250	2053895232695707.000
Explained Variance Score	0.640	0.494	0.604	0.698

Table 2: Summary Result of Linear regression and Ridge Regression

From statistical analysis it is found that Ridge regression model is a better predictor model for our project for predicting Gross.

Table 3 shows the results of test data where we have calculated predicted gross of movies and have checked the difference with the actual gross of test data. Also table 4 shows the predicted gross of recent release movies.

Model Test

Predicted Gross	Gross Actual	Difference (%)	Title
$54,552,794.66	$54,414,716.00	0.25	The Warrior's Way
$3,838,616.45	$4,074,023.00	-5.78	Sparkle
$45,931,251.65	$46,280,507.00	-0.75	The Host
$65,930,363.12	$66,359,959.00	-0.65	The Maze Runner
$24,707,375.44	$24,268,828.00	1.81	Step Up 3D
$22,864,602.47	$22,331,028.00	2.39	Zulu
$74,736,438.72	$75,573,300.00	-1.11	Unbroken
$3,808,048.32	$2,848,578.00	33.68	The Runaways

Table 3

Predictions:

Movie Title	Predicted Gross
Guardians of the Galaxy Vol. 2	$388287985
3 Generations	$8167983.45

VII. CONCLUSION

We have predicted the gross of movies using regression model and to do so we have used features such as budget, IMDB Votes, Number of critics for Review and movie Facebook likes. We selected these features based on the output of PCA and correlation. PCA selects the number of components such that the amount of variance that needs to be explained is greater than the percentage specified by components. Features selected had multicolinearity and therefore we used Ridge regression to predict the gross. Our model was tested using MSE and explained variance and MSE was huge whereas explained variance was very low. This was caused because we previously used last 100 years of data which was not inflation adjusted. To overcome this problem we fitted our model with only 10 years of data to predict gross of future movies. And then the results were as expected, with small MSE and explained variance close to 1. Therefore this data of budget, IMDB Votes, Number of critics for Review and movie Facebook likes are proven to be capable of predicting movie gross with good accuracy. For future work, we plan to take data from more sources and adjust its inflation to current date and see if that increase the model's accuracy. We are also planning to quantify actors and directors to build the feature vector to predict the gross.

VIII. REFERENCES

[1] https://www.cs.cmu.edu/~nasmith/TDF/ZhangWenbinISF2009Paper.pdf

[2] B. Pang and L. Lee, "Thumbs up? sentiment classification using machine learning techniques," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, July 2002, pp. 79–86.

[3] P. Chaovalit and L. Zhou, "Movie review mining: a comparison between supervised and unsupervised classification approaches," in Proceedings of the Hawaii International

[4] http://www3.cs.stonybrook.edu/~skiena/591/final_projects/movie_gross/

[5] http://www.statisticssolutions.com/correlation-pearson-kendall-spearman/

[6] https://en.wikipedia.org/wiki/Feature_vector

[7] https://en.wikipedia.org/wiki/Principal_component_analysis

[8] Scikit Learn – Feature selection: http://scikit-learn.org/stable/modules/feature_selection.html

[9] http://www.stat.yale.edu/Courses/1997-98/101/linmult.htm

[10] https://stats.stackexchange.com/questions/52653/what-is-ridge-regression

[11] https://en.wikipedia.org/wiki/Mean_squared_error

[12] http://scikit-learn.org/stable/modules/model_evaluation.html

[13]http://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables