Fundamentals of Data Science Project:

Movie Gross Prediction from IMDB 5000 Movie Dataset (by Anjana Tiha)

1.1 Introduction:

I have used IMDB 5000 movie dataset to predict movie gross. Both categorical and numerical features of the database have been used to predict gross. It has been treated as regression problem.

2. Preprocessing:

Data has been separately preprocessed for different experimentation (using both numerical and textual data and treating textual features as categorical features (eg. country, actor names, director name and so on)).

2.1. Features:

Following features have been used for gross prediction:

Numerical Features: 5 numerical features have been used. 3 top actors' and director's Facebook likes, budget are the numerical features that has been used to predict gross.

Text Features: 7 textual features have been used in gross prediction. They are 3 top actors' names, director's name, country, content rating, language.

2.2. Features Preprocessing:

Data Cleaning has been cleaned before data type specific preprocessing. Rows with missing gross value has been removed. Records with missing major feature values were also removed from data.

Numerical Features Preprocessing: Numerical data has been scaled to 0 – 1 by using fitting in minmax scaler. As data had wide range of values, using 0-1 scaling was very helpful.

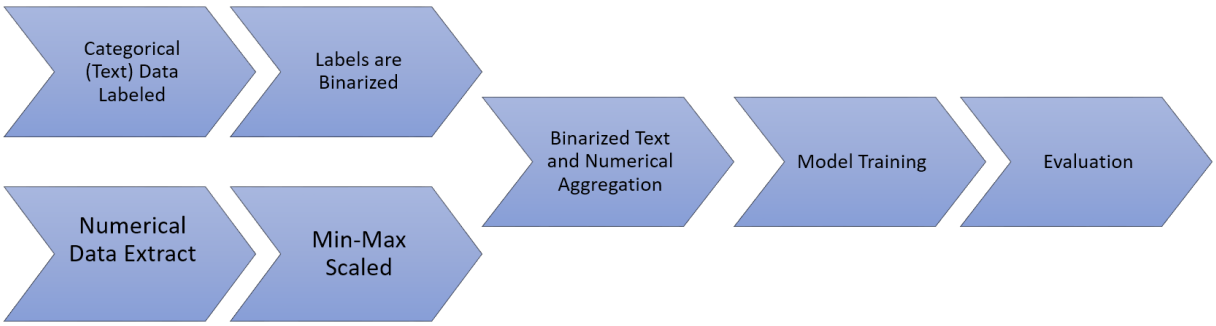


Table 1:Gross prediction process

Textual Features Preprocessing:

Textual data has been labeled for each column separately. Each column was assigned label for each distinct feature. Each textual data column has been labeled and then they have been transformed to binary form.

Motivation for using Textual Data as Categorical Data:

Motivation behind using the textual features like actor names, director names, country, content rating as categorical data is that they contain information that can be crucial for good gross prediction. Using this data in any other form could make it lose the utility of using cast and textual data. Also, transforming actor, director names and some of the textual data helps maintaining the complex relation among the feature columns. For example, a set of features like actor, director combination could attain certain range of gross for a movie. So, using the features as category could help preserve their interrelations.

4. Regression Models:

Random Forest Regression and Decision Tree Regression has been used to predict gross. Other machine learning models like linear regression, Support Vector Machine (SVR) has also been applied but were less successful.

5. Regression Evaluation:

Cross Validation has been applied to evaluate the regression model performance. Five - Fold Cross Validation has used. Cross validation with Mean Absolute Error(MAE) and Mean Squared Error(MSE) has been calculated. In addition to MAE and MSE, Median Absolute Error, Explained Var Score, Root Means Square Error has also been calculated for evaluation without cross validation (train test split function used which does random splitting).

Table: Regression Performance Evaluation for gross prediction using both numerical and categorical data

Evaluation Metrics	Random forest Regression	Decision Tree Regression
Mean Absolute Error	0.0398	0.0456
Mean Squared Error	0.0048	0.0065
R^2 Score	0.4628	0.3225

Random forest Regression (Other Scores)
Median Absolute Error : 0.0199578950989
Explained Var Score : 0.468926439852
R^2 Score : 0.449712010893

Decision Tree Regression (Other Scores)
Median Absolute Error : 0.0206232734035
Explained Var Score : 0.265120200536
R^2 Score : 0.241282066425

Performance Plot(Random):

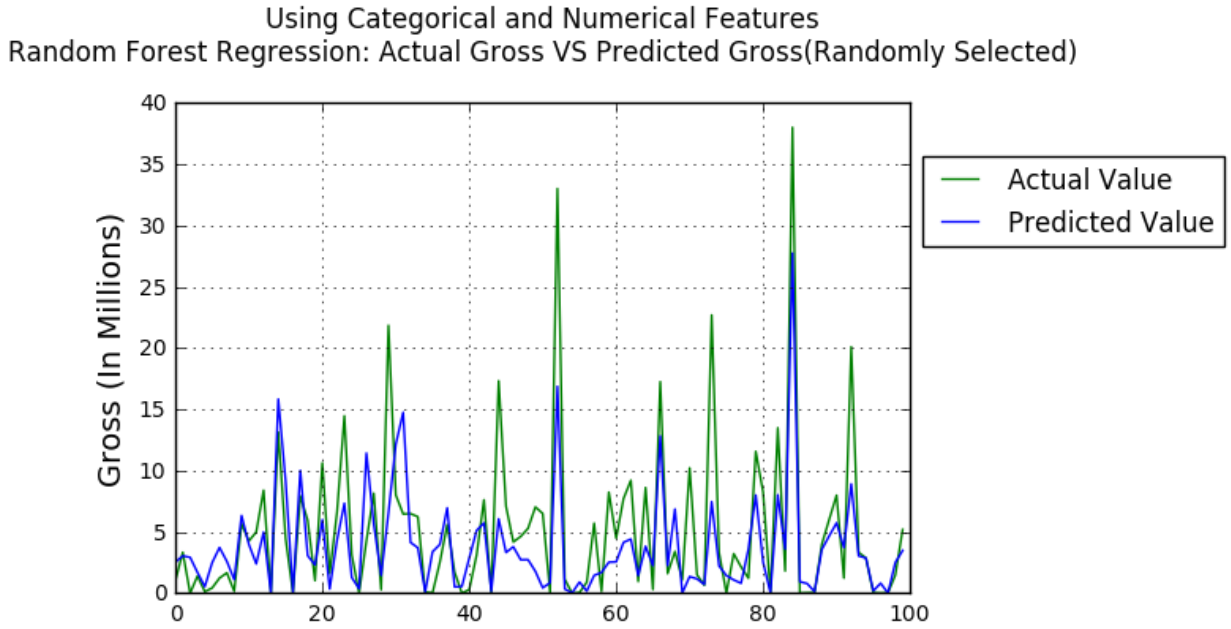


Table 2: Prediction Performance (Random 100 Data)

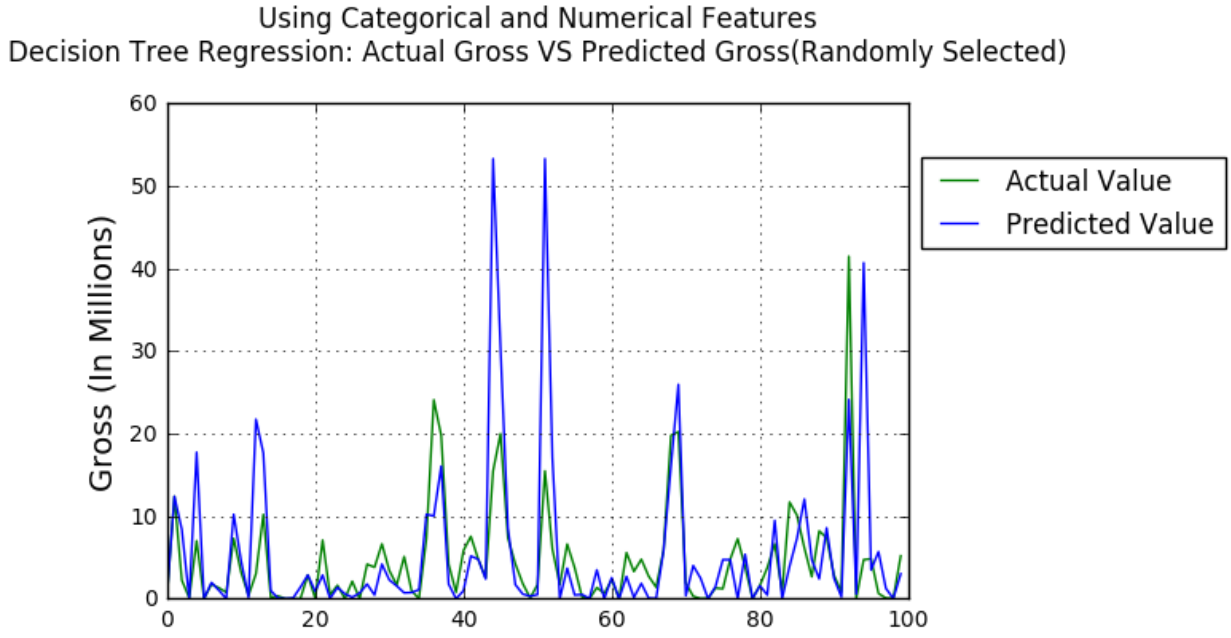


Table 3: Prediction Performance (Random 100 Data)

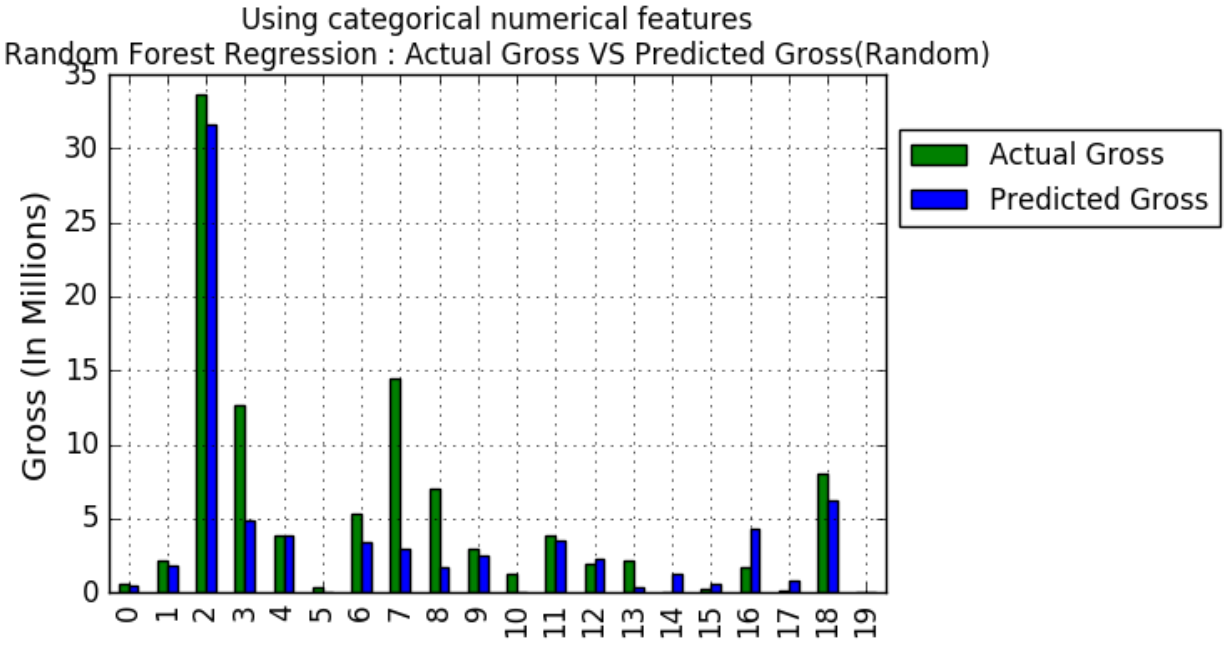


Table 4:Prediction Performance (Random Data)

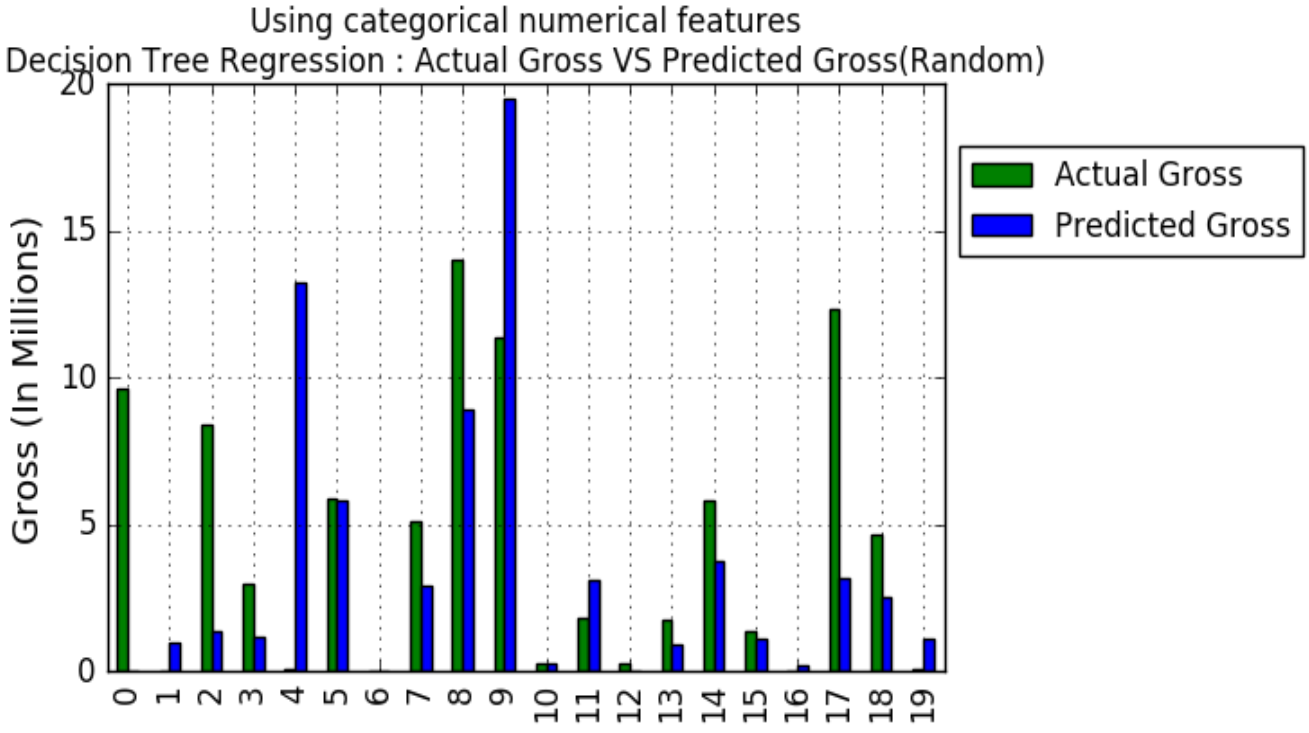
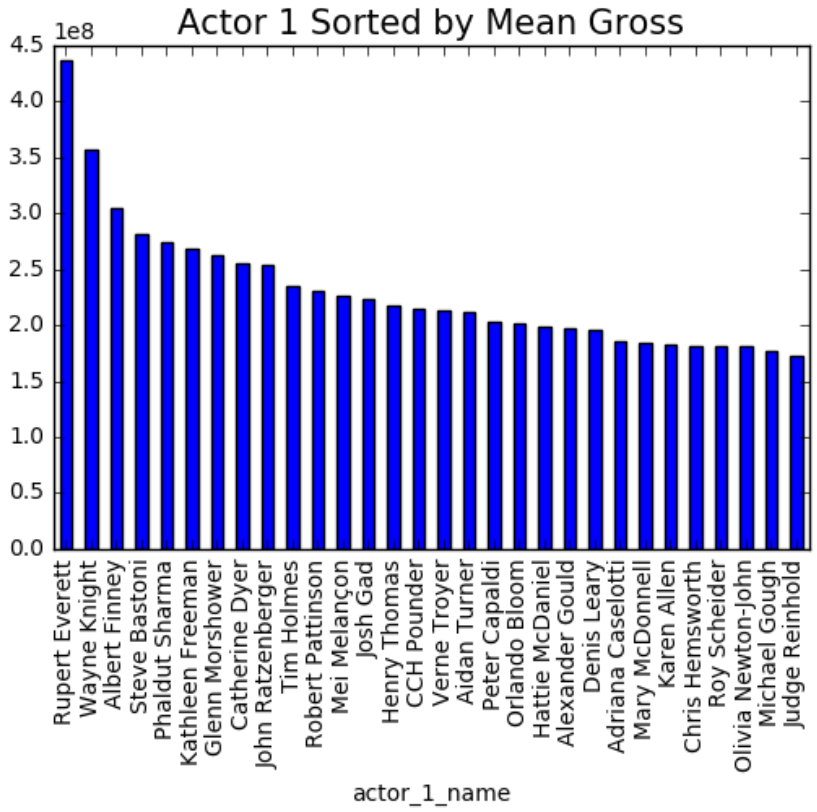
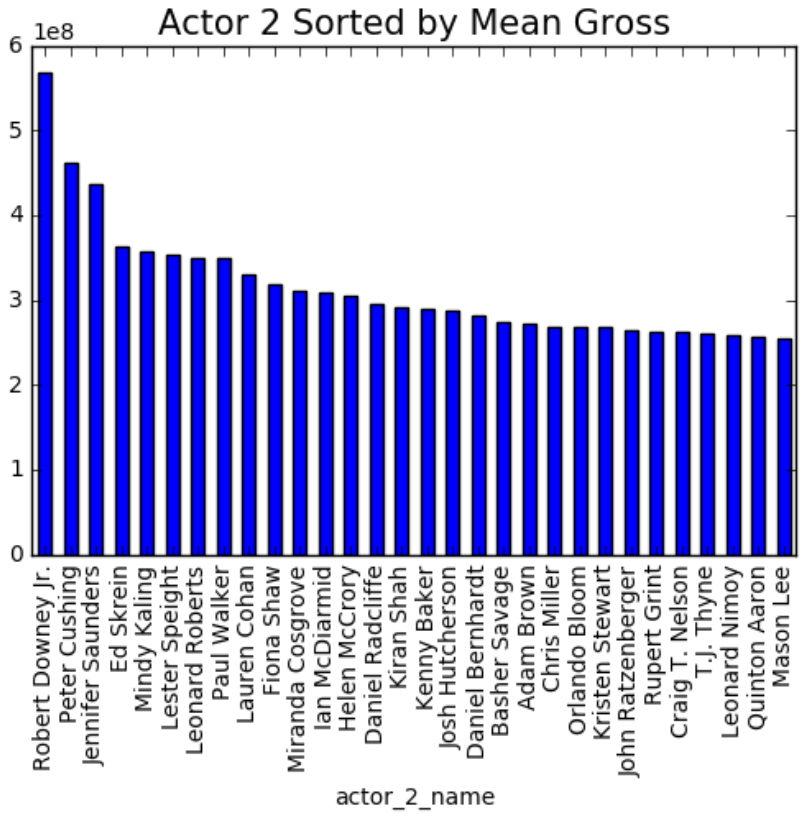


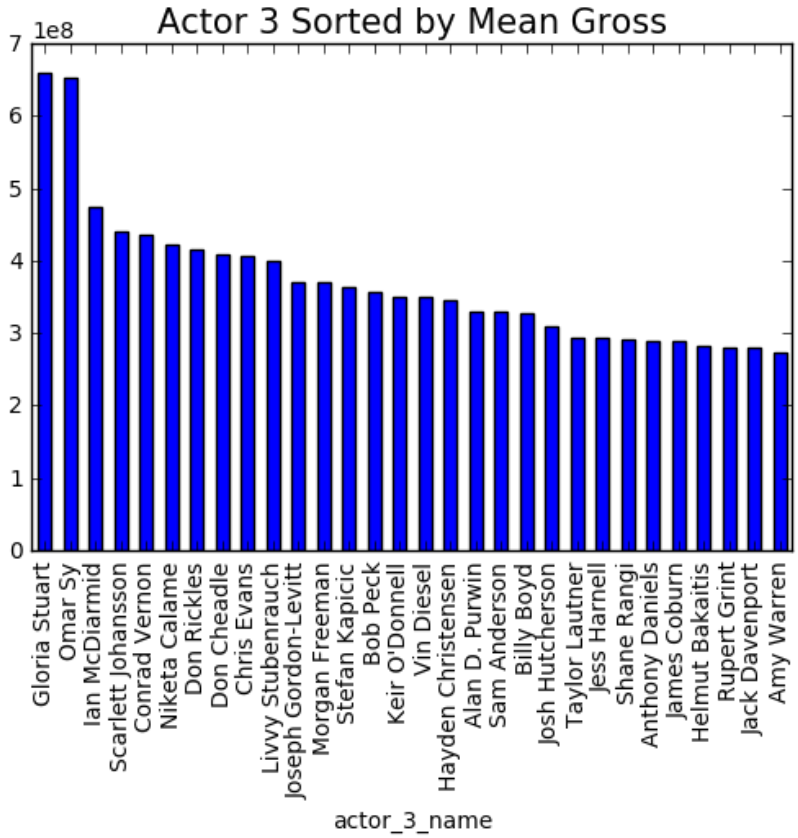
Table 5:Prediction Performance (Random Data)

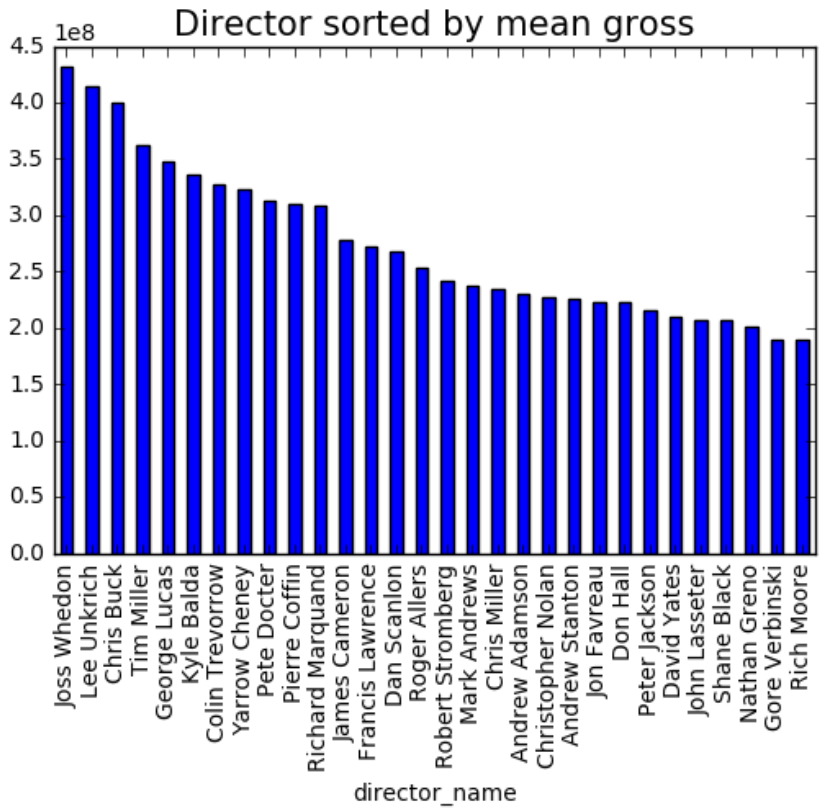
6. Data Visualizations:

Visualization have been added for all selected features in gross prediction problem. Mean gross for Top 3 Actor, director, content rating, country, and language have been plotted to show top candidates in each feature. Result data has been plotted and shown in different way. Prediction vs actual gross HAS been plotted for customized view in a range sequentially. User can select number of predicted and actual gross to be shown though line chart and bar chart.




Actor 2 Sorted by Mean Gross
1e8
6
5
4
3
2
1
0
Robert Downey Jr.
Peter Cushing
Jennifer Saunders
Ed Skrein
Mindy Kaling
Lester Speight
Leonard Roberts
Paul Walker
Lauren Cohan
Fiona Shaw
Miranda Cosgrove
Ian McDiarmid
Helen McCrory
Daniel Radcliffe
Kiran Shah
Kenny Baker
Josh Hutcherson
Daniel Bernhardt
Basher Savage
Adam Brown
Chris Miller
Orlando Bloom
Kristen Stewart
John Ratzenberger
Rupert Grint
Craig T. Nelson
T.J. Thyne
Leonard Nimoy
Quinton Aaron
Mason Lee
actor_2_name


Actor 3 Sorted by Mean Gross
1e8
7
6
5
4
3
2
1
0
Gloria Stuart
Omar Sy
Ian McDiarmid
Scarlett Johansson
Conrad Vernon
Niketa Calame
Don Rickles
Don Cheadle
Chris Evans
Livvy Stubenrauch
Joseph Gordon-Levitt
Morgan Freeman
Stefan Kapicic
Bob Peck
Keir O'Donnell
Vin Diesel
Hayden Christensen
Alan D. Purwin
Sam Anderson
Billy Boyd
Josh Hutcherson
Taylor Lautner
Jess Harnell
Shane Rangi
Anthony Daniels
James Coburn
Helmut Bakaitis
Rupert Grint
Jack Davenport
Amy Warren
actor_3_name



7. Conclusion:

Adding textual data like plot keywords, movie name, genre for gross prediction could help further improve the result. Also, using other machine learning techniques like *Deep Neural Network(DBN)* could help make result more accurate.