Chris Cronin
GA Data Science
Final Paper
8/20/2015

Data Science @ The Movies

Problem Statement and Hypothesis

In Hollywood, there is no guarantee that a studio's release of a movie will be successful. Compare Disney's *The Lone Ranger*, a film that earned only $90 million in domestic gross office income off a $270 Million production budget with Haxan film's *Blair Witch Project*, a film that earned $140 million in domestic gross office income off a paltry $600 thousand production budget. Given these examples, can a movie's financial success at the box office be predicted?

This paper will seek to answer this question by creating a linear regression machine-learning model that analyzes large quantities of movie data. This model, and variations on it, will find the most important factors determining a movie's performance and by doing so predict how much return on capital movie studios can expect to make.

Data Set Description

The data for this project was obtained from OpusData, a subsidiary of Nash Information Services ('NIS'), LLC. NIS is the creator of the popular movie database website *The Numbers.* The dataset included ten csv files of which seven (described below) were used for this project, all linked through an Opus Data Identification Number (ODID).

1. Summary: Comprehensive list of movie information, the most important columns include production budget, inflation adjusted domestic box office revenue, international revenue, sequel status, production year, running time, creative type (i.e historical fiction), source (i.e. original screenplay), production method (i.e. live action), and genre (i.e. thriller/suspense).
2. Acting credits: Actors in each film with their character and acting role (i.e. lead, supporting)
3. Technical credits: Technical names in each film and their specific roles (i.e. director, screenwriter, cinematographer).
4. Ratings: Ratings for each movie (i.e. G, PG, PG-13, R)
5. Production companies: Production studios behind each movie (i.e. Warner Brothers)
6. Releases: Release date, pattern (i.e. Imax, Limited or Wide release), and movie distributor.
7. Keywords: Keywords associated with each movie (i.e. "surprise twist")

Pre-Processing Steps

The bulk of the time spent on this project was in conducting pre-processing work to create one combined csv file with the most important columns to be used as features in the linear model. This first involved reading in the seven csv files and converting them from dataframes to strings to allow merging through Python. It then encompassed the following three steps.

Step 1: Creating New Columns

Eight new columns were created to better analyze the data and they represent key features and responses in the linear model. Adding these columns involved combining columns with arithmetic operations. These new columns quantified return on investment multiples, profit, and running time for each movie. The most important new column is called "money," and is used as the response in the linear model. It represents gross income earned for a movie by adding its inflation adjusted box office income with its international box office income.

In addition, some of these new columns involved creating functions to sort and analyze key information. These new columns sorted production budgets into different ranges, created a 1 through 12 key representing which month a movie was released, and created a column that indicated whether a movie was released on Christmas or not.

Step 1: Sorting CSV Data

The seven csv files containing movie data contain far too much information to be simply added without filtering. Actors were sorted by only lead actors who appear in more than five movies and technical contributors were sorted by only directors who directed in more than five movies. Production companies were sorted by only the first two production companies for a movie and only if the company had produced more than ten movies. Similarly, distribution companies and keywords were sorted by count to get a set of the most important names.

64% of the movies had to removed from the data set leaving 3525 movies out of 9725. This was because there was no production budget value for them given by Opus Data since the company didn't have accurate figures for them. Although this was unfortunate, 3525 movies were more than enough to analyze and a future project could involve incorporating the complete movie set.

Step 3: Merging the Sorted CSV Data

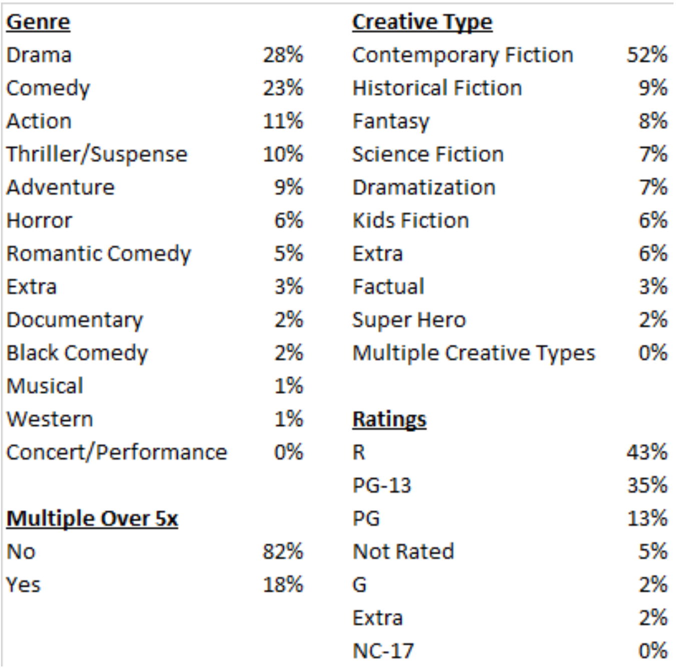
The sorted CSV data were merged into one large dataframe consisting of 3525 rows and 901 columns. Each row represents a different movie with a unique ODID number and each column represents a feature. Columns like production budget are the values given but most of the combined data involved merging movie attributes into binary values. For example, the final dataframe consists of 250 actor columns (for all actors in greater than five movies) with each column presenting either a “0” or a “1” based on whether that individual actor was in the movie or not. If you look at the row with the movie Titanic, there will be a “1” under the column “person_leonardo_dicaprio.”.

Data Exploration

Analysis of the data provided many key insights. Whether or not a movie is a sequel is a very important indicator on how well a movie will do. While sequels represented 9% of the analyzed movies, they on average made $240 million dollars more than original films. The list below of the top and bottom twenty-five profitable films reveals many sequels in the top category like *Furious 7* and *Jurassic World*. Not surprisingly, there are no sequels in the bottom category.

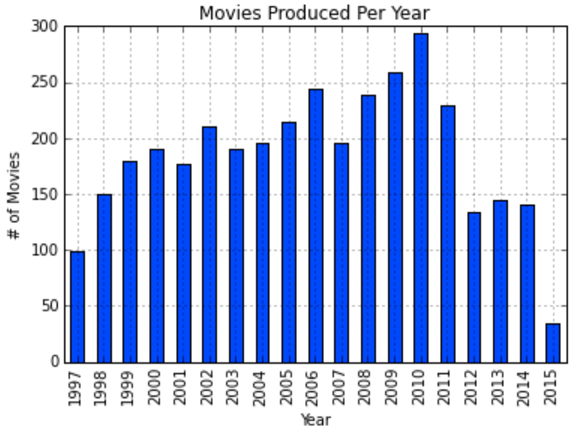
Movie	Profit	Movie	Profit
Titanic	**$2,454,100,432**	**Child 44**	**-$46,565,670**
Avatar	**$2,399,932,123**	**The Great Raid**	**-$46,690,803**
Furious 7	**$1,321,032,910**	**Beyond Borders**	**-$47,038,753**
The Avengers	**$1,307,007,774**	**Lucky You**	**-$47,440,882**
Jurassic World	**$1,298,620,315**	**Outlander**	**-$48,735,658**
Harry Potter and the Deathly Hallows: Part II	**$1,225,640,105**	**Soldier**	**-$49,777,997**
The Lord of the Rings: The Return of the King	**$1,174,721,230**	**Legends of Oz: Dorothyâ€™s Return**	**-$49,892,067**
Star Wars Ep. I: The Phantom Menace	**$1,170,821,861**	**Lucky Numbers**	**-$49,913,625**
The Avengers: Age of Ultron	**$1,143,509,368**	**Ballistic: Ecks vs. Sever**	**-$50,021,666**
Frozen	**$1,124,234,980**	**Osmosis Jones**	**-$50,493,478**
Iron Man 3	**$1,015,392,272**	**Blackhat**	**-$50,608,937**
Shrek 2	**$996,497,636**	**R.I.P.D.**	**-$50,927,536**
Harry Potter and the Sorcerer's Stone	**$986,795,917**	**Lolita**	**-$53,012,792**
The Lord of the Rings: The Two Towers	**$970,483,722**	**Eye See You**	**-$53,192,010**
Finding Nemo	**$960,994,460**	**Stealth**	**-$53,537,774**
Pirates of the Caribbean: Dead Man's Chest	**$942,682,348**	**The Alamo**	**-$61,197,147**
Transformers: Dark of the Moon	**$937,233,694**	**Fantasia 2000 (Theatrical Release)**	**-$66,285,669**
Skyfall	**$916,483,596**	**Monkeybone**	**-$67,239,351**
The Lord of the Rings: The Fellowship of the Ring	**$909,462,386**	**How Do You Know?**	**-$69,517,956**
Despicable Me 2	**$898,758,842**	**The Nutcracker in 3D**	**-$69,528,286**
Transformers: Age of Extinction	**$894,039,076**	**A Sound of Thunder**	**-$73,203,496**
The Dark Knight	**$887,603,427**	**The Spanish Prisoner**	**-$78,732,935**
Harry Potter and the Chamber of Secrets	**$882,328,061**	**The Adventures of Pluto Nash**	**-$91,151,192**
Toy Story 3	**$881,915,962**	**Town & Country**	**-$91,727,804**
Harry Potter and the Order of the Phoenix	**$845,572,695**	**Mars Needs Moms**	**-$109,937,684**

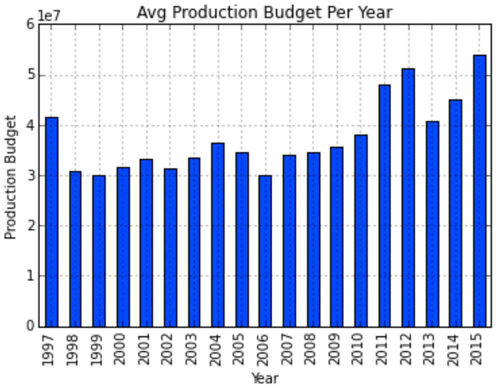
Depicted in the chart below, a counting of some of the more descriptive features reveals dramas and comedies to be the most prevalent genres, contemporary fiction the most prevalent creative type, and “R” being the most frequent rating. Additionally, roughly 20% of the movies produced more than 5x return on their production budgets.

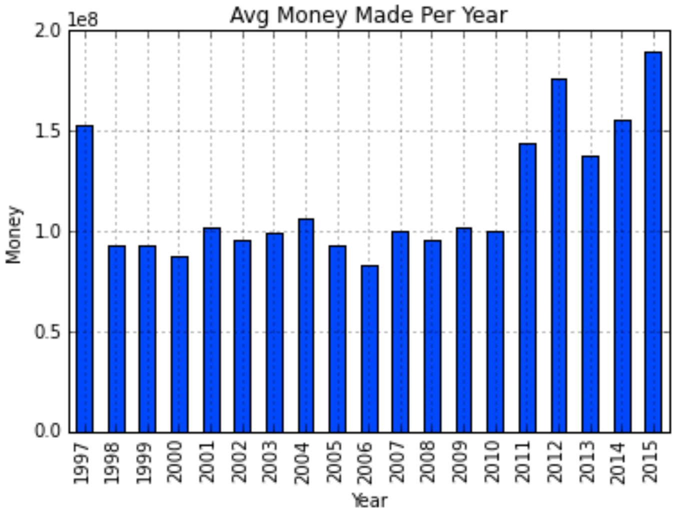


Genre		**Creative Type**	
Drama	28%	Contemporary Fiction	52%
Comedy	23%	Historical Fiction	9%
Action	11%	Fantasy	8%
Thriller/Suspense	10%	Science Fiction	7%
Adventure	9%	Dramatization	7%
Horror	6%	Kids Fiction	6%
Romantic Comedy	5%	Extra	6%
Extra	3%	Factual	3%
Documentary	2%	Super Hero	2%
Black Comedy	2%	Multiple Creative Types	0%
Musical	1%		
Western	1%	**Ratings**	
Concert/Performance	0%	R	43%
		PG-13	35%
Multiple Over 5x		PG	13%
No	82%	Not Rated	5%
Yes	18%	G	2%
		Extra	2%
		NC-17	0%

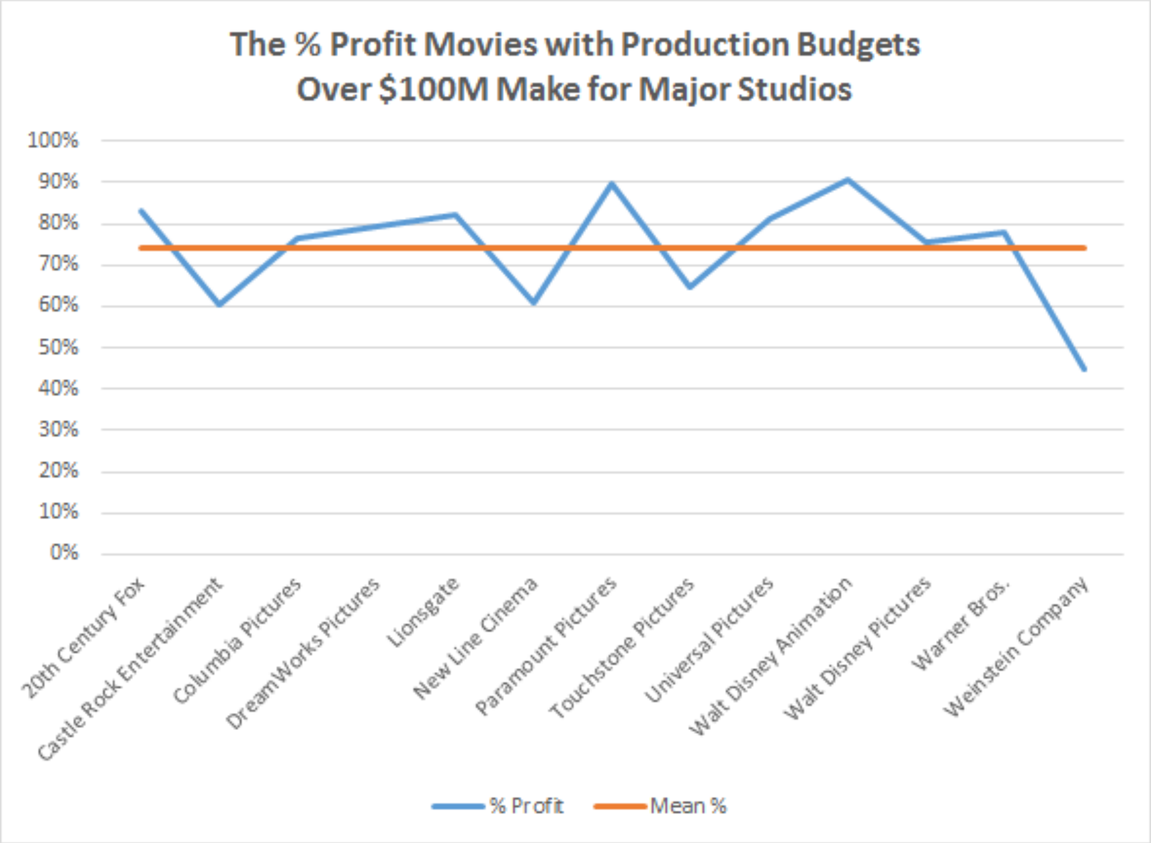
The next major insight corroborates the “blockbuster” theory towards movie production, which describes how movie studios, in order to stay competitive and profitable, have to take risks by producing movies with large production budgets. The charts below show that although the number of movies being produced each year has fallen since 2010, average production budget is on an upward trend corresponding with larger average money earned. Movie studios are putting more into less films and as of now, making more money from it.







Creating a function that sorted movie production budgets into different categories further supported the blockbuster theory depicted in the chart below. It shows that for the top movie studios, on average 75% of their profits come from the movies they produced with over $100 million production budgets. Applying this insight to Disney's *Lone Ranger* means that even though the movie was such a huge loss, Disney should still continue to make movies with large production budgets if they want to stay competitive.



Feature Selection

All of the merged and sorted/filtered columns were used as features using simple functions to represent each category. For example, "Person_Leonardo_Dicaprio" would represent one feature as well as "keyword1_surprisetwist." These represented 854 features out of 859 total.

Next, since all the csv files were merged to the main "summary" csv file, I went through the summary columns to determine the best features for the model. After running the linear model and getting suspiciously high results, it became apparent that features like 'opening weekend revenue' were too good at making predictions because they were determined after a movie's release. These and other similar features were therefore removed. New columns that were created to better organize the data were also removed and after this process, only five additional features emerged: runtimesquared[1], sequel, production budget, Christmas release, and release_month.

[1] Runtimesquare is the total time of a movie squared (which makes it usable in the model)

I titled my features “feature_cols,” which represented 859 different features to use in the model. To trim this large feature set, I used stats model to calculate the p-values for each feature, and sorted all the p-values into a new list called “relevant_features,” that showed all the features with p-values < .05.

The p-value is the probability that the relationship we are observing is occurring purely by chance. If the p-value is less than .05, that means the confidence intervals for those coefficients do not include zero, and the null hypothesis can be rejected in favor of the alternative hypothesis, which states that there is a relationship between the feature and response. “Relevant_features” trimmed the feature set to 119 features, depicted below:

```
'source_Based on Ballet',
 'source_Compilation',
 'genre_Action',
 'genre_Adventure',
 'genre_Black Comedy',
 'genre_Comedy',
 'genre_Drama',
 'genre_Horror',
 'genre_Musical',
 'genre_Romantic Comedy',
 'genre_Thriller/Suspense',
 'person_Amanda Seyfried',
 'person_Anne Hathaway',
 'person_Anthony Hopkins',
 'person_Arnold Schwarzenegger',
 'person_Ben Stiller',
 'person_Bradley Cooper',
 'person_Cameron Diaz',
 'person_Christina Ricci',
 'person_Colin Farrell',
 'person_Colin Firth',
 'person_Daniel Radcliffe',
 'person_Elizabeth Banks',
 'person_Emma Watson',
 'person_Freddie Highmore',
'person_Freida Pinto',
'person_Hugh Jackman',
'person_Jennifer Connelly',
'person_Jennifer Lawrence',
'person_Jeremy Renner',
'person_John Leguizamo',
'person_Julia Roberts',
'person_Kate Winslet',
'person_Keanu Reeves',
'person_Keira Knightley',
'person_Kevin Spacey',
'person_Kirsten Dunst',
'person_Leonardo DiCaprio',
'person_Michelle Rodriguez',
'person_Morgan Freeman',
'person_Natalie Portman',
'person_Orlando Bloom',
'person_Rachel McAdams',
'person_Rachel Weisz',
'person_Ralph Fiennes',
'person_Robert Pattinson',
'person_Rupert Grint',
'person_Sam Rockwell',
'person_Sam Worthington',
'person_Scarlett Johansson',
'person_Steve Carell',
'person_Tom Hanks',
'person_Vin Diesel',
'person_Will Smith',
'person_Woody Harrelson',
'name_Andrew Adamson',
'name_Ang Lee',
'name_Bill Condon',
'name_Bryan Singer',
'name_Chris Columbus',
'name_Chris Weitz',
'name_Christopher Nolan',
'name_Jay Roach',
'name_Joseph McGinty Nichol',
'name_Louis Leterrier',
'name_Luc Besson',
'name_M. Night Shyamalan',
'name_Michael Bay',
'name_Raja Gosnell',
'name_Rob Cohen',
'name_Robert Zemeckis',
'name_Roland Emmerich',
'name_Sam Raimi',
'name_Steven Spielberg',
'creative_type_Kids Fiction',
'production_method_Hand Animation',
'production_method_Live Action',
'production_method_Rotoscoping',
'production_method_Stop-Motion Animation',
'production_company1_Gold Circle Films',
'production_company1_New Line Cinema',
'production_company1_Relativity Media',
'production_company1_Universal Pictures',
'production_company1_Walt Disney Animation Studios',
'production_company1_Warner Bros.',
'production_company2_Participant Media',
'production_company2_Regency Enterprises',
'distributor_Dreamworks SKG',
'distributor_Warner Bros.',
'release_pattern_Expands Wide',
'release_pattern_IMAX',
'release_pattern_Limited',
'release_pattern_Special Engagement',
'release_pattern_Wide',
'keywords1_Animals Gone Bad',
'keywords1_Family Movie',
'keywords1_Marvel Comics',
'keywords1_Relationship Advice',
'keywords1_Secret Agent',
'keywords1_Terminal Illness',
'keywords1_Visual Effects',
'keywords2_Animal Lead',
'keywords2_Artists',
'keywords2_Disaster',
'keywords2_Immigration',
'keywords2_Kidnap',
'keywords2_Road Trip',
'keywords2_Secret Agent',
'keywords3_Coming of Age',
'keywords3_Cross-Class Romance',
'keywords3_End of the World',
'keywords3_Faulty Memory',
'keywords3_Heist',
'keywords3_IMAX: DMR',
'keywords3_Mistaken Identity',
'keywords3_Time Travel',
'keywords3_Visual Effects',
'sequel',
'production_budget']
```

Modeling Process and Evaluation

The modeling process involved using regression modeling because the question of predicting movie financial performance is a supervised learning problem with a continuous response. The main model used was a scikit learn model using “feature_cols” as features. The response was a column called “money” that represented inflation adjusted domestic box office revenue plus international box office revenue.

The main evaluation of the model’s success was each model’s R-squared value as well rankings of feature importance. The R-squared value represents the variance in the observed data that is explained by the linear regression model. The scikit learn model with “feature_cols” as features and “money” as its response resulted in the highest R-squared value of 78%.

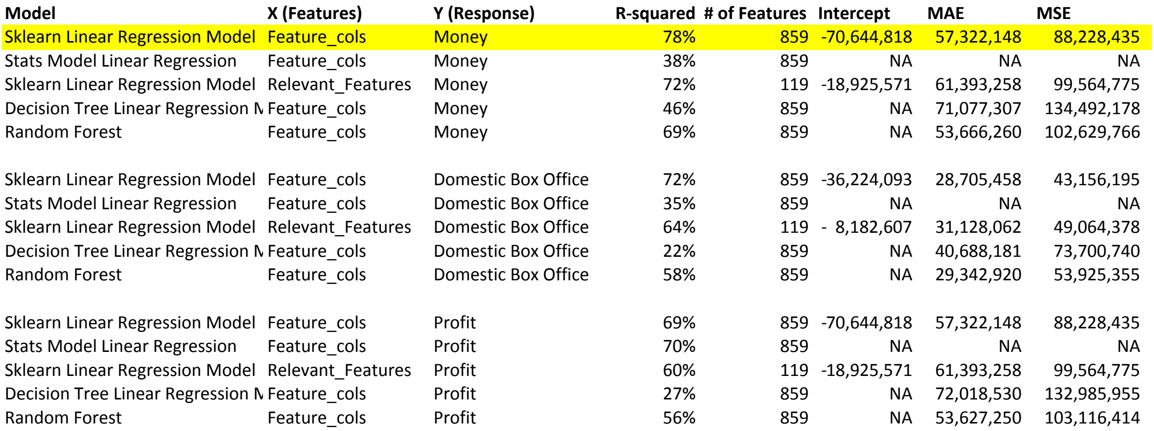
The evaluation of the importance of each “feature_cols” features was done through python’s statsmodel where “feature_cols” features were narrowed down into “relevant_features” based on p-values as previously mentioned. Although the scikit learn model was run a second time with “relevant_features” in place of “feature_cols,” it resulted in a lower R-squared value of 72%.

Regression decision tree and random forest modeling was implemented to further analyzed the data. The decision tree model split the data, trained on 70% of it (both “feature_cols[2]” and corresponding “money” responses) and evaluated feature importance with a series of decisions. It then tested these predictions on the remaining 30% it hadn’t seen. The random forest model took this a step further by running the training test on the 70% of the data one hundred different times, and averaging out the different evaluations to make predictions on the 30% dataset. Although the decisions tree and random forest models didn’t make prediction better than the scikit learn model, they ranked important features differently, and this was valuable to gaining insights about different movie influencers. Below is a list of the decision tree and random forest feature rankings:

Decision Tree Features	**Random Forest Features**
production_budget 0.612	production_budget 0.574
runtimesquared 0.076	runtimesquared 0.077
keywords3_Cross-Class Romance 0.037	sequel 0.026
sequel 0.034	keywords3_Cross-Class Romance 0.022
production_method_Live Action 0.016	release_pattern_IMAX 0.016
release_month 0.015	production_method_Live Action 0.013
person_Robert Downey, Jr. 0.013	8release_month 0.012
keywords1_Boarding School 0.010	distributor_Paramount Pictures 0.008
keywords2_Romance 0.008	name_Andrew Adamson 0.007
release_pattern_Special Engagement 0.006	production_method_Digital Animation 0.006
person_Antonio Banderas 0.006	release_pattern_Wide 0.006
keywords2_Family Movie 0.005	genre_Thriller/Suspense 0.004
keywords2_Talking Animals 0.004	distributor_Walt Disney 0.004
name_Sam Raimi 0.004	name_M. Night Shyamalan 0.004
name_Tim Burton 0.004	person_Sigourney Weaver 0.004
name_M. Night Shyamalan 0.004	release_pattern_Special Engagement 0.003
distributor_Universal 0.004	keywords2_Disaster 0.003
production_method_Digital Animation 0.003	genre_Adventure 0.003
source_Based on Fiction Book/Short Story 0.003	person_Robert Downey, Jr. 0.003
distributor_Warner Bros. 0.003	person_Rupert Grint 0.003
creative_type_Science Fiction 0.003	person_Kristen Bell 0.003
keywords3_IMAX: DMR 0.003	keywords1_Boarding School 0.003
person_Tom Hanks 0.003	distributor_Warner Bros. 0.003
name_Andrew Adamson 0.003	person_Daniel Radcliffe 0.003
production_company1_Columbia Pictures 0.002	keywords2_Romance 0.003
genre_Adventure 0.002	distributor_Sony Pictures 0.003
production_company1_Touchstone Pictures 0.002	person_Tom Hanks 0.003
person_Dustin Hoffman 0.002	person_John Leguizamo 0.002
production_company1_Universal Pictures 0.002	person_Michelle Rodriguez 0.002
production_company1_Paramount Pictures 0.002	source_Based on Fiction Book/Short Story 0.002

[2] “Feature_cols” was used instead of “Relevant_features” because it was better for the decision tree and random forest models analyze all of the feature columns.

I experimented with using these four regression models with different features and responses. For features, I used both “relevant_features” and “feature_cols.” For responses, I used “money,” “inflation adjusted domestic box office,” and “profit.[3]” I evaluated them through calculating their R-SQUARED, Mean Average Error (MAE[4]), and Mean Squared Error (MSE[5]). The best model was the original Scikit Learn model with “relevant_cols” as features and “money” as response. A depiction of these variations is below:



Model	X (Features)	Y (Response)	R-squared	# of Features	Intercept	MAE	MSE
Sklearn Linear Regression Model	Feature_cols	Money	78%	859	-70,644,818	57,322,148	88,228,435
Stats Model Linear Regression	Feature_cols	Money	38%	859	NA	NA	NA
Sklearn Linear Regression Model	Relevant_Features	Money	72%	119	-18,925,571	61,393,258	99,564,775
Decision Tree Linear Regression M	Feature_cols	Money	46%	859	NA	71,077,307	134,492,178
Random Forest	Feature_cols	Money	69%	859	NA	53,666,260	102,629,766
Sklearn Linear Regression Model	Feature_cols	Domestic Box Office	72%	859	-36,224,093	28,705,458	43,156,195
Stats Model Linear Regression	Feature_cols	Domestic Box Office	35%	859	NA	NA	NA
Sklearn Linear Regression Model	Relevant_Features	Domestic Box Office	64%	119	- 8,182,607	31,128,062	49,064,378
Decision Tree Linear Regression M	Feature_cols	Domestic Box Office	22%	859	NA	40,688,181	73,700,740
Random Forest	Feature_cols	Domestic Box Office	58%	859	NA	29,342,920	53,925,355
Sklearn Linear Regression Model	Feature_cols	Profit	69%	859	-70,644,818	57,322,148	88,228,435
Stats Model Linear Regression	Feature_cols	Profit	70%	859	NA	NA	NA
Sklearn Linear Regression Model	Relevant_Features	Profit	60%	119	-18,925,571	61,393,258	99,564,775
Decision Tree Linear Regression M	Feature_cols	Profit	27%	859	NA	72,018,530	132,985,955
Random Forest	Feature_cols	Profit	56%	859	NA	53,627,250	103,116,414

Challenges and Successes

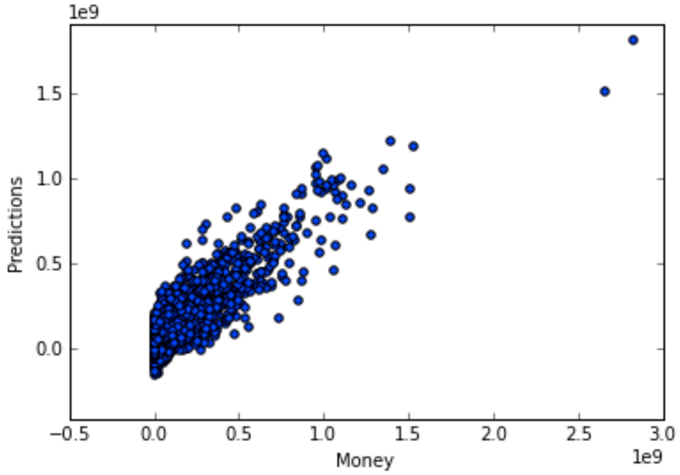
There were many challenges to creating a linear machine-learning model that would predict movie financial performance. It required extensive pre processing work combining the appropriate columns as features. The dataset was also incomplete only representing 36% of movies and having these additional movies would have produced a more accurate picture. Also, the production budgets didn’t include the cost of marketing the movies (this figure is unknown to Opus Data) and would have had a significant impact predicting responses. Additionally, it would have been insightful to include viewer and critic review information to get a sense of how much it influences a movie’s financial performance. It was also challenging creating the best features for the model, and more work can be done in this respect.

Nevertheless, the linear model and the variations on it do a solid job at making financial predictions. The Scikit Learn model has the highest R-SQUARED value of 78% and the chart below shows a correlation between the model’s “predictions” with the actual “money” amount. Therefore, with more tweaking of data features, these linear models have the potential to be used with confidence.

[3] Profit represented “money” minus “production budget”

[4] MAE is the average of the absolute errors, which is the prediction and the true value. It measures the magnitude of a set of errors, without considering their direction.

[5] MSE measures the average of the squared errors.



Business Applications

This project has the potential to be implemented by production studios in helping with "greenlight" decisions to go ahead with producing a movie. It can aid executives with deciding which types of movies they need to produce in a given year to be successful, minimizing the risk in their portfolio.

As an example of this, I created a narrow random forest model with "money" as its response and these three features ["production budget" of $80 million, "runtimesquared" of 10404 minutes, and "release_pattern_IMAX] that predicts a movie with these features will make $260 million. An executive wouldn't want to make this bet with only three features but with a more complete and evolving model, it could prove very useful in deciding whether to put resources behind making a movie.

Conclusion

This project creates linear machine-learning models to help predict a movie's financial performance, and the models in this project are a great start to creating one that could be used by movie executives. The challenge to this and other linear models is providing the very best features that will boost the R-squared values or % of error covered by the model.

Bibliography

www.opusdata.com
www.the-numbers.com