DATA SCIENCE @ THE MOVIES

BY CHRIS CRONIN - 8/20/215

PROBLEM STATEMENT AND HYPOTHESIS

Can a movie's financial success be predicted?

This project creates a linear machine-learning model that does just that!

This is an example of a supervised learning problem with a continuous response

- Haxan Films
- $600K Production Budget
- $140M Domestic Box Office

- Disney Studios
- $270M Production Budget
- $90M Domestic Box Office

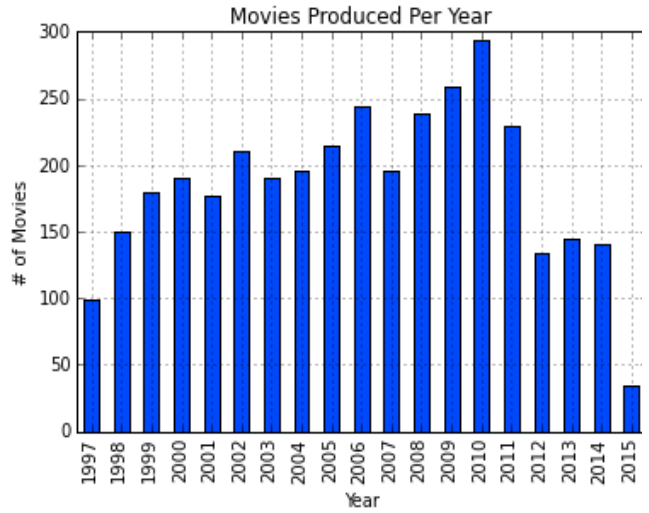
DATA SET DESCRIPTION & PREPROCESSING WORK

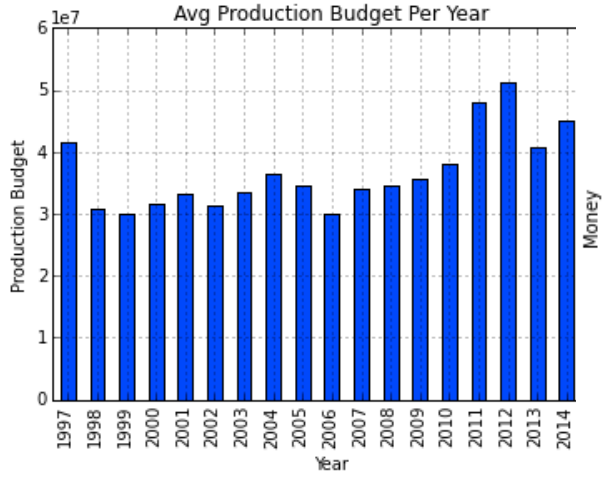
- The data for this project was obtained from OpusData, a movie database company behind the popular site *The Numbers*. The dataset included ten csv files of which seven were used for this project, all linked through an Opus Data Identification Number (ODID).
  - CSV Files: movie_summary, acting_credits, technical_credits, movie_ratings, production_companies, releases, and keywords.

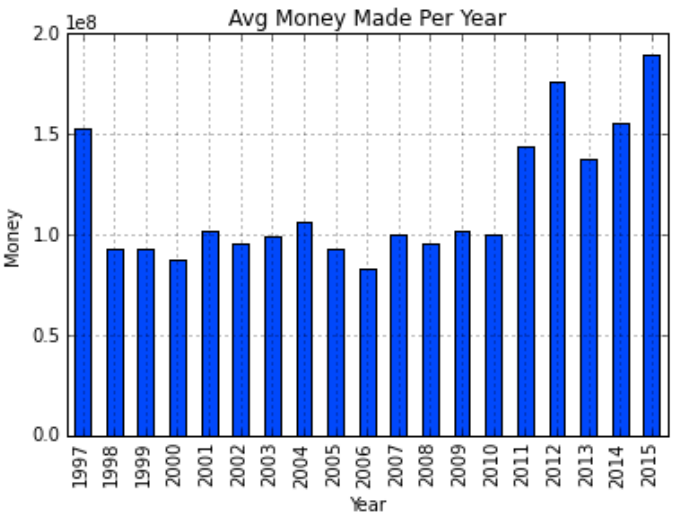
- Preprocessing took the majority of time spent on this project and a huge thanks to Ramesh, Patrick, Liam, and Sinan for all your help! It involved:
  - Creating new columns to be used as features and for analysis
  - Sorting the csv files and then merging them into one large final file with **3525 ROWS AND 901 COLUMNS!**
  - Example: The actor Leonardo Dicaprio represents one actor column out of 250, and a “1” is placed in each row of that column for a movie he is in, and a “0” for each movie he is not in.

DATA EXPLORATION & ANALYSIS

- **SEQUELS:** While sequels represented 9% of the analyzed movies, they on average made $240 million dollars more than original films.
- Dramas and comedies are the most prevalent genres, contemporary fiction the most prevalent creative type, and “R” being the most frequent rating.
- The data corroborates the “blockbuster” theory towards movie production – In order to make money, studios have to be willing to spend it!

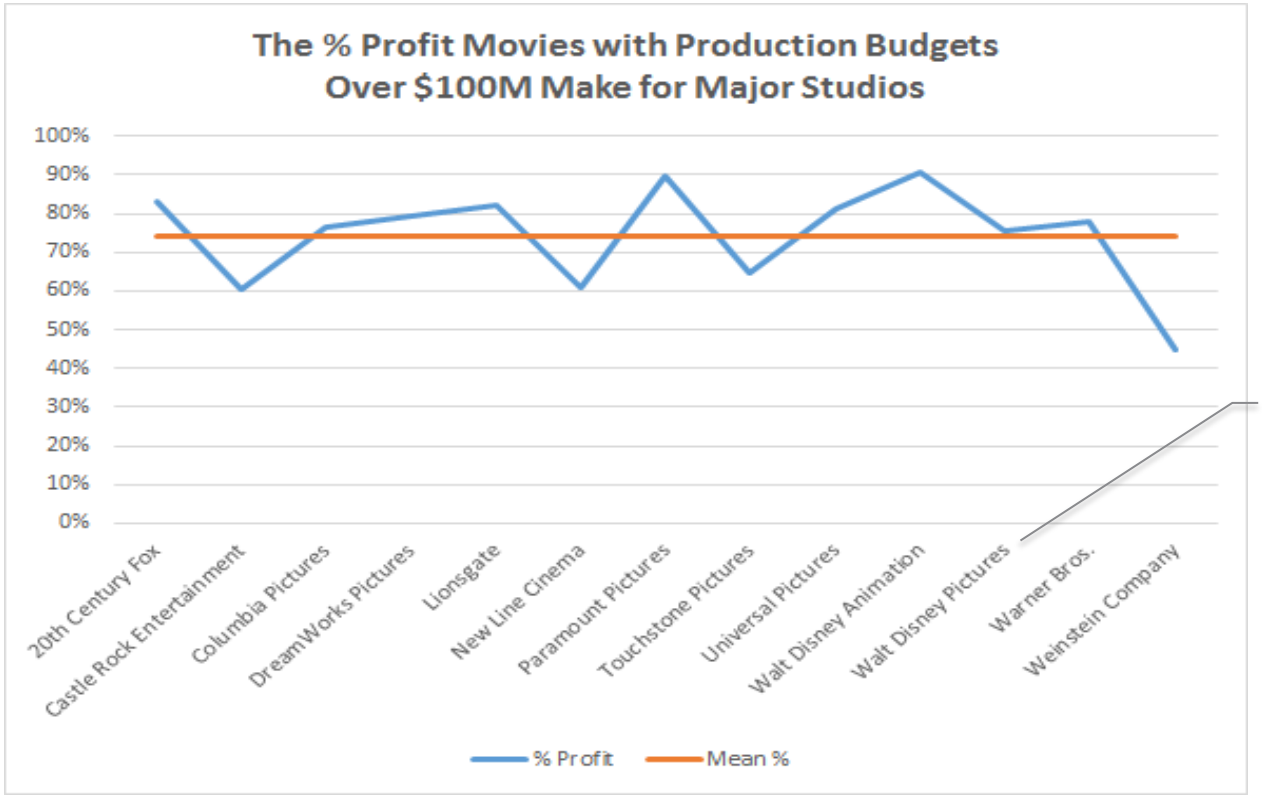






DATA EXPLORATION & ANALYSIS (CONT)

Analysis of movie production budgets and production companies revealed that for the top movie studios, on average 75% of their profits come from the movies they produced with over $100 million production budgets.



This should make Disney feel better about Lone Ranger

FEATURE SELECTION

FEATURE_COLS = 859 feature columns

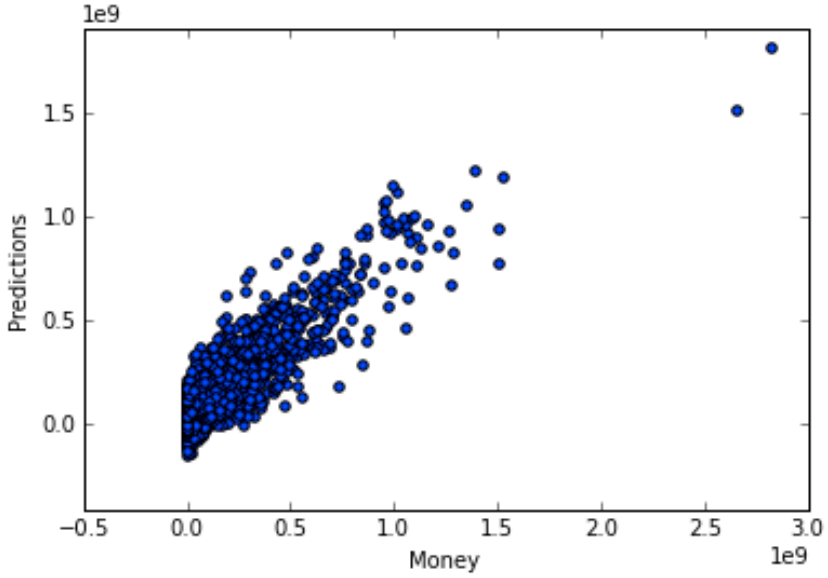
- 854 rows were from merged csv files relating to actors, directors, production companies, ratings, releases, and keywords
- 5 rows were from the original summary csv: production budget, runtimesquared, sequel, release month, and Christmas releases
    - Note: Removed features that contained metrics after a movie's release!

RELEVANT_FEATURES = 119 feature columns

- Use STATS MODEL in python to trim the FEATURE_COLS set by separating 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 - that there is a relationship between the feature and response.

REGRESSION MODELS & EVALUATION

- Model Evaluation:
  - Mainly R-squared values but also calculated intercept, Mean Average Error (MAE), and Mean Square Error (MSE).
  - Feature importance rankings
- The R-squared value represents the variance in the observed data that is explained by the model.
- **Scikit Learn:** Best model producing the highest R-squared value of 78%.



REGRESSION MODELS & EVALUATION (CONT)

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

- **Stats Model:** Used mainly to create “Relevant_Features” based on p-values
- **Decision Tree Model:** Split the data, trained on 70% of it and evaluated feature importance with a series of decisions, and then tested it on the remaining 30% it hadn’t seen.
- **Random forest model:** Ran the decision tree training test on the 70% of the data 100 different times, and averaged out the different evaluations to make predictions on the remaining 30%.

FEATURE IMPORTANCE

Decision Tree Features

production_budget 0.612
runtimesquared 0.076
keywords3_Cross-Class Romance 0.037
sequel 0.034
production_method_Live Action 0.016
release_month 0.015
person_Robert Downey, Jr. 0.013
keywords1_Boarding School 0.010
keywords2_Romance 0.008
release_pattern_Special Engagement 0.006
person_Antonio Banderas 0.006
keywords2_Family Movie 0.005
keywords2_Talking Animals 0.004
name_Sam Raimi 0.004
name_Tim Burton 0.004
name_M. Night Shyamalan 0.004
distributor_Universal 0.004
production_method_Digital Animation 0.003
source_Based on Fiction Book/Short Story 0.003
distributor_Warner Bros. 0.003
creative_type_Science Fiction 0.003
keywords3_IMAX: DMR 0.003
person_Tom Hanks 0.003
name_Andrew Adamson 0.003
production_company1_Columbia Pictures 0.002
genre_Adventure 0.002
production_company1_Touchstone Pictures 0.002
person_Dustin Hoffman 0.002
production_company1_Universal Pictures 0.002
production_company1_Paramount Pictures 0.002

Random Forest Features

production_budget 0.574
runtimesquared 0.077
sequel 0.026
keywords3_Cross-Class Romance 0.022
release_pattern_IMAX 0.016
production_method_Live Action 0.013
8release_month 0.012
distributor_Paramount Pictures 0.008
name_Andrew Adamson 0.007
production_method_Digital Animation 0.006
release_pattern_Wide 0.006
genre_Thriller/Suspense 0.004
distributor_Walt Disney 0.004
name_M. Night Shyamalan 0.004
person_Sigourney Weaver 0.004
release_pattern_Special Engagement 0.003
keywords2_Disaster 0.003
genre_Adventure 0.003
person_Robert Downey, Jr. 0.003
person_Rupert Grint 0.003
person_Kristen Bell 0.003
keywords1_Boarding School 0.003
distributor_Warner Bros. 0.003
person_Daniel Radcliffe 0.003
keywords2_Romance 0.003
distributor_Sony Pictures 0.003
person_Tom Hanks 0.003
person_John Leguizamo 0.002
person_Michelle Rodriguez 0.002
source_Based on Fiction Book/Short Story 0.002

Stats Model

'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',

FEATURE IMPORTANCE (STATS MODEL CONT)

Stats Model	Stats Model	Stats Model
'person_John Leguizamo',	'name_Chris Weitz',	'release_pattern_IMAX',
'person_Julia Roberts',	**'name_Christopher Nolan',**	'release_pattern_Limited',
'person_Kate Winslet',	'name_Jay Roach',	'release_pattern_Special Engagement',
'person_Keanu Reeves',	'name_Joseph McGinty Nichol',	'release_pattern_Wide',
'person_Keira Knightley',	'name_Louis Leterrier',	'keywords1_Animals Gone Bad',
'person_Kevin Spacey',	'name_Luc Besson',	'keywords1_Family Movie',
'person_Kirsten Dunst',	'name_M. Night Shyamalan',	'keywords1_Marvel Comics',
'person_Leonardo DiCaprio',	'name_Michael Bay',	'keywords1_Relationship Advice',
'person_Michelle Rodriguez',	'name_Raja Gosnell',	'keywords1_Secret Agent',
'person_Morgan Freeman',	'name_Rob Cohen',	'keywords1_Terminal Illness',
'person_Natalie Portman',	'name_Robert Zemeckis',	'keywords1_Visual Effects',
'person_Orlando Bloom',	'name_Roland Emmerich',	'keywords2_Animal Lead',
'person_Rachel McAdams',	'name_Sam Raimi',	'keywords2_Artists',
'person_Rachel Weisz',	'name_Steven Spielberg',	'keywords2_Disaster',
'person_Ralph Fiennes',	'creative_type_Kids Fiction',	'keywords2_Immigration',
'person_Robert Pattinson',	'production_method_Hand Animation',	'keywords2_Kidnap',
'person_Rupert Grint',	'production_method_Live Action',	'keywords2_Road Trip',
'person_Sam Rockwell',	'production_method_Rotoscoping',	'keywords2_Secret Agent',
'person_Sam Worthington',	'production_method_Stop-Motion Animation',	'keywords3_Coming of Age',
'person_Scarlett Johansson',	'production_company1_Gold Circle Films',	'keywords3_Cross-Class Romance',
'person_Steve Carell',	'production_company1_New Line Cinema',	'keywords3_End of the World',
'person_Tom Hanks',	'production_company1_Relativity Media',	'keywords3_Faulty Memory',
'person_Vin Diesel',	'production_company1_Universal Pictures',	'keywords3_Heist',
'person_Will Smith',	'production_company1_Walt Disney Animatior	'keywords3_IMAX: DMR',
'person_Woody Harrelson',	'production_company1_Warner Bros.',	'keywords3_Mistaken Identity',
'name_Andrew Adamson',	'production_company2_Participant Media',	'keywords3_Time Travel',
'name_Ang Lee',	'production_company2_Regency Enterprises',	'keywords3_Visual Effects',
'name_Bill Condon',	'distributor_Dreamworks SKG',	'sequel',
'name_Bryan Singer',	'distributor_Warner Bros.',	'production_budget']
'name_Chris Columbus',	'release_pattern_Expands Wide',	

BUSINESS APPLICATIONS & CONCLUSION

- It is possible to create a linear model to help predict a movie's financial performance!
- The models in this project are a great start to creating one that could be used by movie executives to help with "greenlight" decision making.
  - "production budget" of $80 million + "runtimesquared" of 10404 minutes + "release_pattern_IMAX" = $260 million movie!
- The challenge to this and other linear models is providing the very best features that will boost the R-SQUARED values (% of error covered by the model).