IMDB MOVIE ANALYSIS

ACME CORPORATION

By: Rebecca Redmond, Wendy Merchant, George Tavlarios, John Trotto

BUSINESS RELEVANCE

- US Film industry generates $35.3 billion per year (statista)
- With the ongoing rise of streaming the industry looks to grow even more in the future
- Business questions
  - How does movie genre relate to revenue?
  - Do top directors and actors tend to make higher earning films?
  - Which rating had a stronger relationship with movie's revenue?
    - Votes, Rating, and Metascore
- By providing an accurate assessment of what factors determine a film's revenue, studios and streaming services will look to Acme as consultants when drafting up new movie ideas

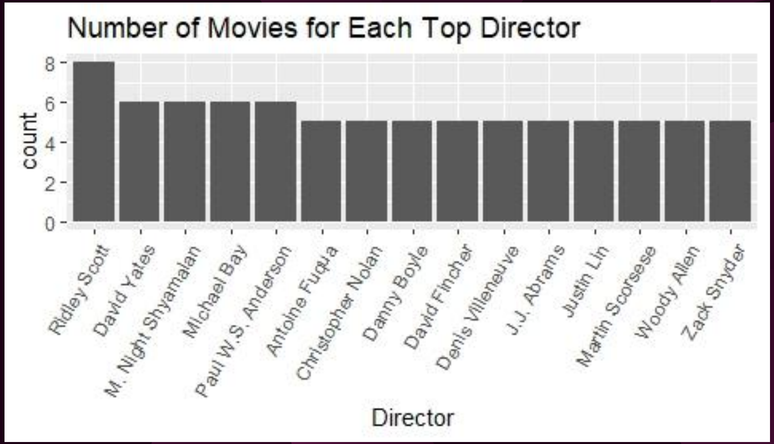
DATASET

- Data: IMDB movie data set
- Source: https://www.kaggle.com/PromptCloudHQ/imdb-data
- Units of Analysis: Data set includes the 1,000 most popular ranked movies by IMDB from 2006-2016
  - 838 movies included in analysis due to some missing values for revenue and metascore
- Key Variables: Revenue, Genre, Metascore, Ratings, Votes, Director (top director indicated by binary variable), Runtime
  - Top director includes the 15 directors with the most movies in dataset

INITIAL ANALYSIS - DIRECTOR

- Hypothesis test conducted showed a significant difference between mean revenue of top and non-top directors (p-value = .0004)
- Mean top directors = 139.13 million
  Mean non-top = 77.20 million

Many of these directors are known for popular and long series (e.g. David Yates: Harry Potter, Justin Lin: Fast & Furious). This dataset does not account for franchises, but it is worth noting from a business perspective that revenue may not necessarily be driven the directors themselves, but rather their specific projects.



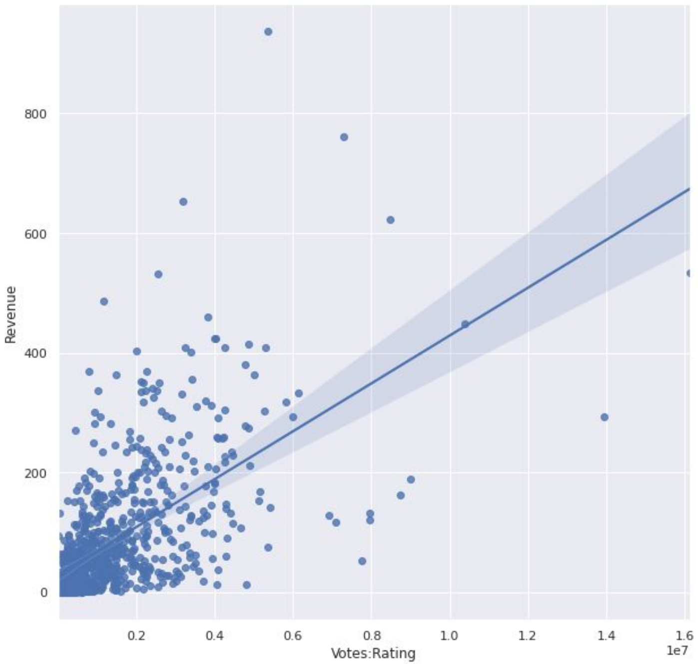
ONE-WAY ANOVA SHOWS DIFFERENCE IN GENRE

- To determine if there is a statistically significant difference in revenue across various genres, a One-Way ANOVA was conducted
  - H_0 = Average revenue is equal
  - H_A = Average revenue is not equal
- Since $p < 0.001$, there is statistical evidence to believe average revenues across genre is not consistent

	Df	Sum^2	Mean^2	F value	p-value
Genre	12	1757510	146459	16.71	**<0.001**
Residuals	859	7528477	8764		

REVENUE CORRELATES WITH POPULARITY

- Correlation test shows a positive relationship between votes, rating, and revenue
  - Votes and rating, which are determined by users, show that popularity of the film among fans and IMDB users after its release corresponded highly with a movie's high revenue
  - As votes increased, so does revenue
  - A slight decrease in votes results in an increase in revenue



Predicted Revenue =
16.07 + 0.0015(**Votes**) - 3.38(**Rating**) - 0.000142(**Votes*Rating**)

COMPARISON OF DATA MINING TOOLS

Model	Stepwise Linear Regression	Tree	KNN (k=17)	Neural Network	5th Model
Variables Included in Model	ZEV_Revenue_Director Votes ZEV_Votes_Director ZEV_Revenue_Actors ZEV_Votes_Actors ZEV_Rating_Director ZEV_Revenue_Genre ZEV_Metascore_Genre Metascore ZEV_Votes_Genre Runtime	Votes, Drama, Animation, Runtime, Adventure, Metascore	Genre, Runtime, Votes, Top Director	Year, Runtime, Rating, Votes, Metascore	Votes ZEV_Rating_Genre ZEV_Votes_Genre ZEV_Votes_Actors ZEV_Metascore_Genre Runtime
Train MSE	2229.46	4407.25	5284.9	5726.34	4845.26
Test MSE	2262.55	4718.12	6418.7	6339.71	6247.79

KNN & NEURAL NETWORK MODELS ARE NOT TOO INSIGHTFUL

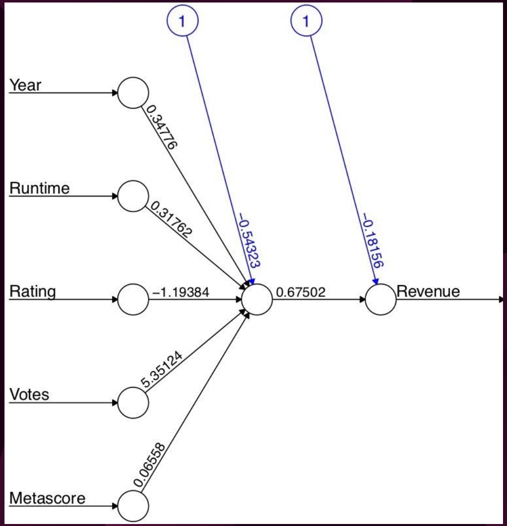
- Both had higher MSE’s when compared to Tree & Stepwise
- Both have outputs that are difficult to interpret to make any real-world conclusions
- The main finding we took from both of these models is that Votes has a great impact on revenue

KNN

Model that included Votes as the only rating metric (ie. ratings and metascore variables were left out) had the lowest train and test MSE out of all models tested

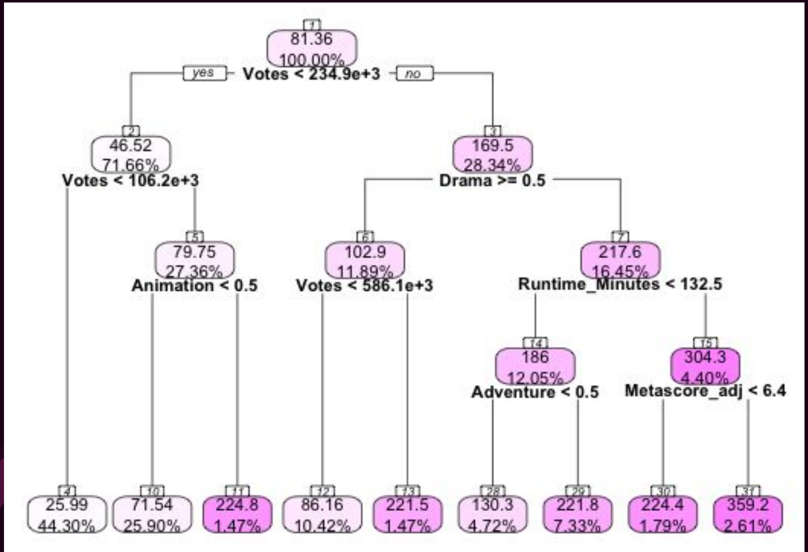
Neural Network

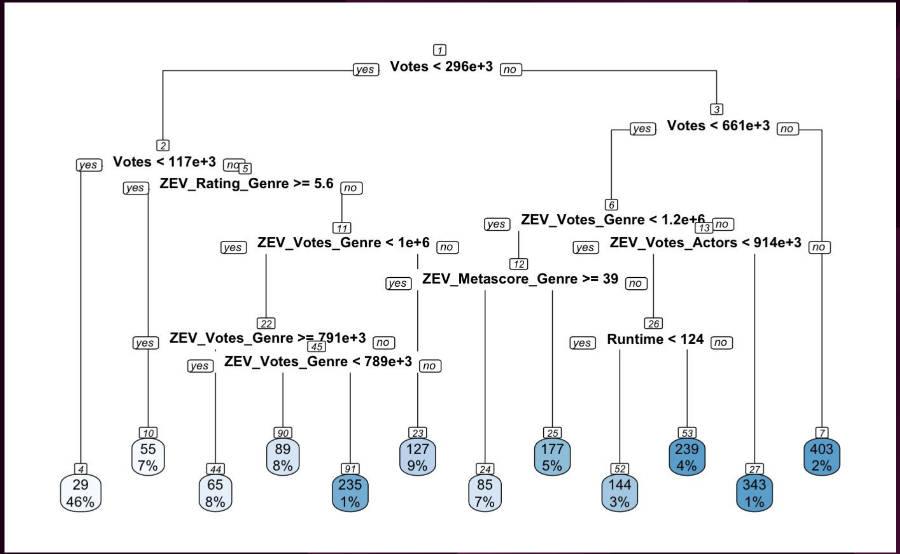
Votes was the heaviest weighted variable in model (Votes coefficient = 5.35 with next influential coefficient being ratings = -1.19)



TREE MODEL PROVIDES INSIGHT INTO REVENUE

- Votes is an important variable when deciding revenue
  - More votes are earned with a movie’s rising popularity among fans, after the film’s release
- Genre is also important, included on both trees; in the first tree, adventure and animation are the two used
- A large portion of the data (44.3%; 46%) is in node 4 on each tree, corresponding to revenue of ~26 million



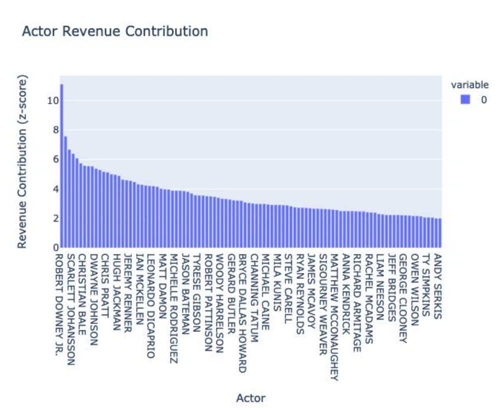


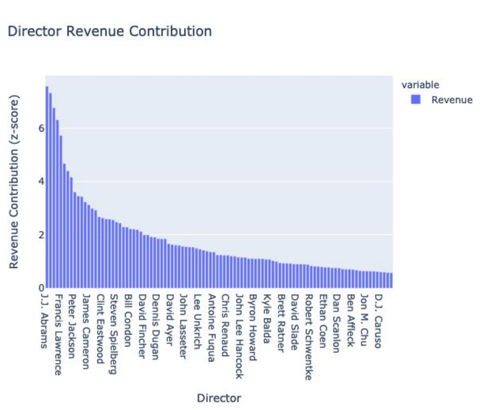
MODEL WE THOUGHT WAS BEST

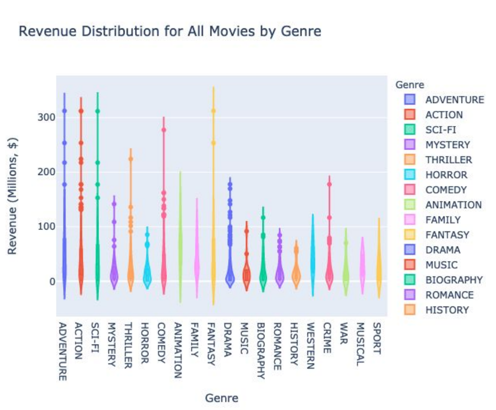
Stepwise Linear Regression
with Expected Values

- Reinforces the intuition that Revenue is more dependent on…
- … when considering the past performance of categorical factors.

… ***Who** it casts…* … ***Who** directs it…* … *and its* ***Theme(s)***…







RECOMMENDATIONS & MOVING FORWARD

- Multiple models show that Votes is a good predictor of revenue
  - Since votes are not acquired until after a movie is released, surveys about possible movies in development could be given to prospective movie goers to gauge interest and help predict votes before movies are produced
  - Further research could also be done to figure out the factors that determine why movies get a certain amount of votes
- Data analysis shows its worth hiring top talent to maximize revenue
  - Dataset included 644 different directors but top 15 created over 10% of films in the dataset
    - Dataset is top 1,000 most popular movies meaning about 2% of directors created 10% of these films
  - Top directors brought in about 80% more revenue than other directors on average
- Would be better to know the budget for further analysis
  - Anyone can spend a large sum of money to hire the top directors and actors to create a movie that will bring in a large amount of revenue, so it would be better to understand how much return on their investment they make by incorporating the budget