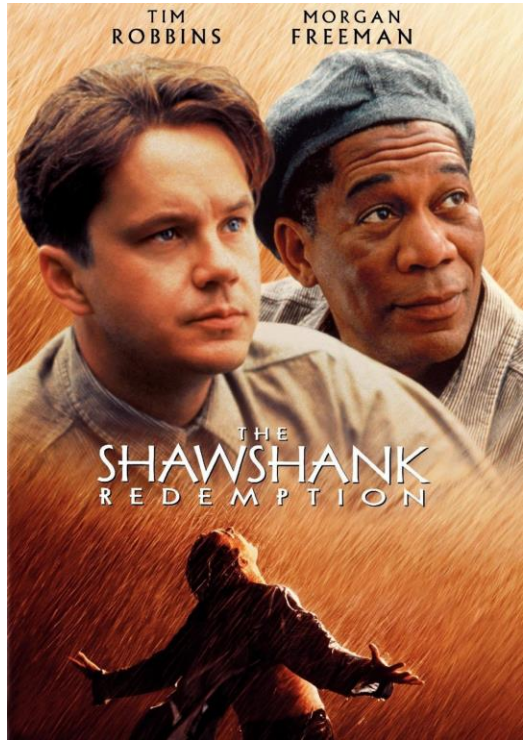
Project Luther

Jonathan Toro

Can we use analytics to identify movies that will be critically acclaimed by the audience?



The Tools used

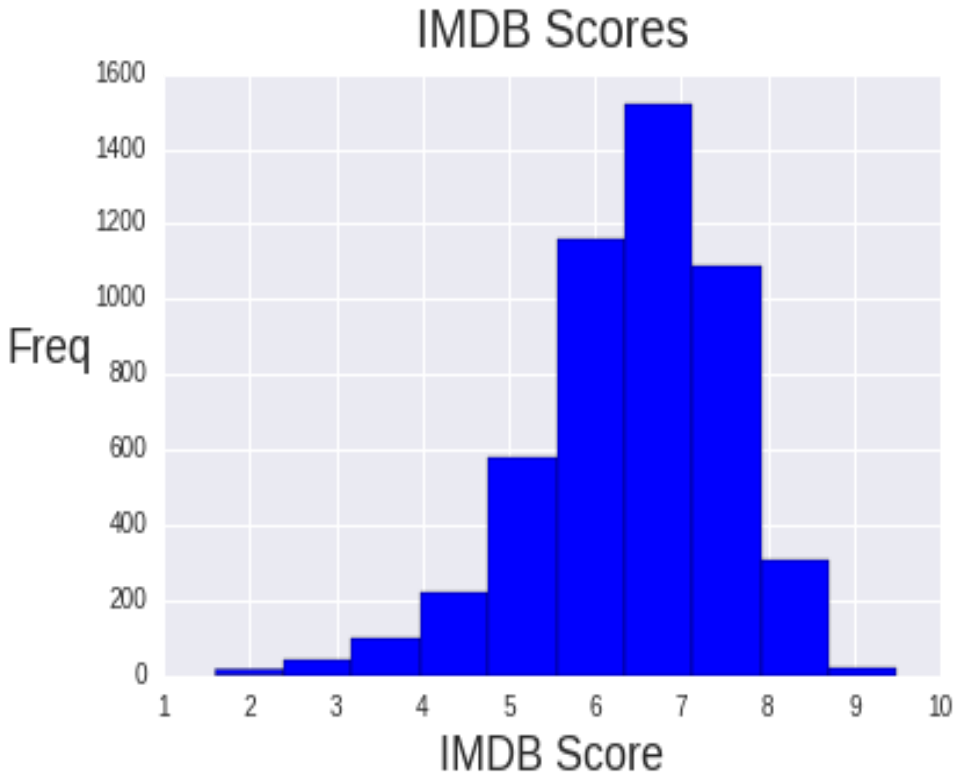
- Python
- Pandas
- Numpy
- Sklearn
- Matplotlib
- Beautiful Soup

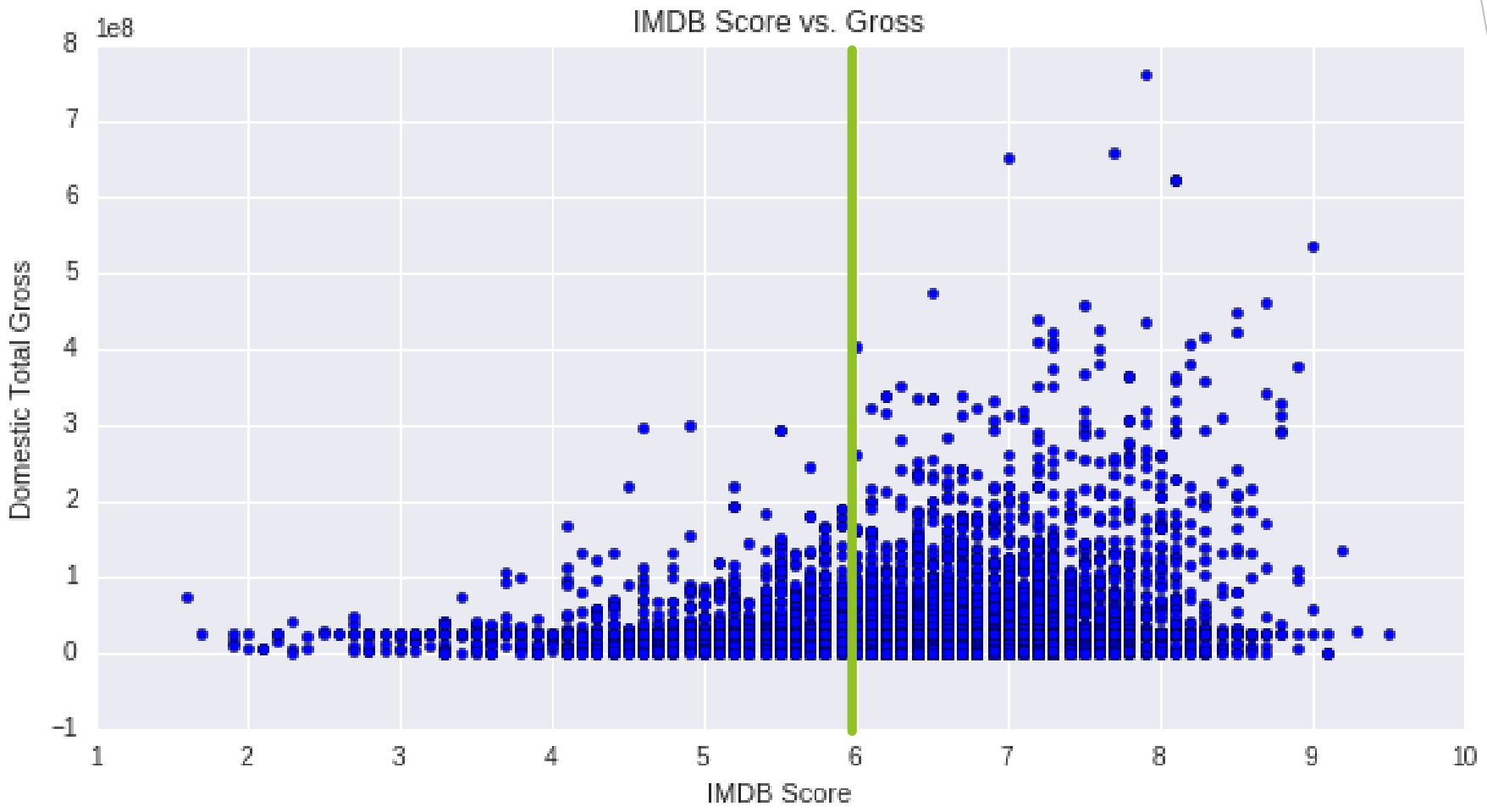
Where is the data coming from?

- Boxofficemojo.com
  - 7 variables with approximately 17000 movies
- IMDB database
  - 30 variables with approximately 5000 movies
  - Variables included: Color, director's name, actor's name, duration, social media statistics, gross, budget, genres, country, rating, IMDB score, year of release

Distribution of scores

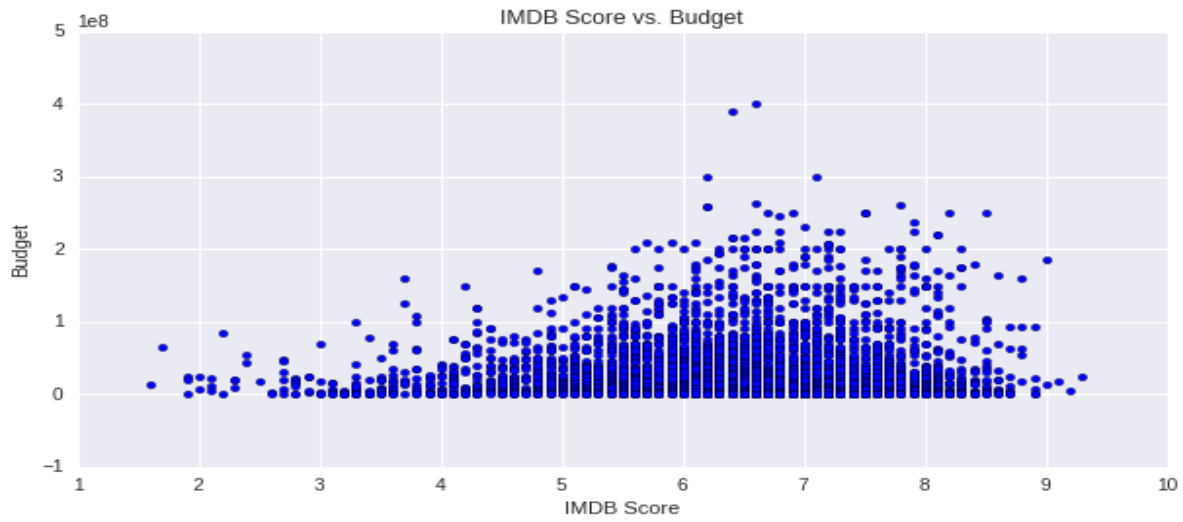
- Top 250 movies are above the score of 8
- Memorable movies are between 7 and 8
- Mediocre movies are below seven

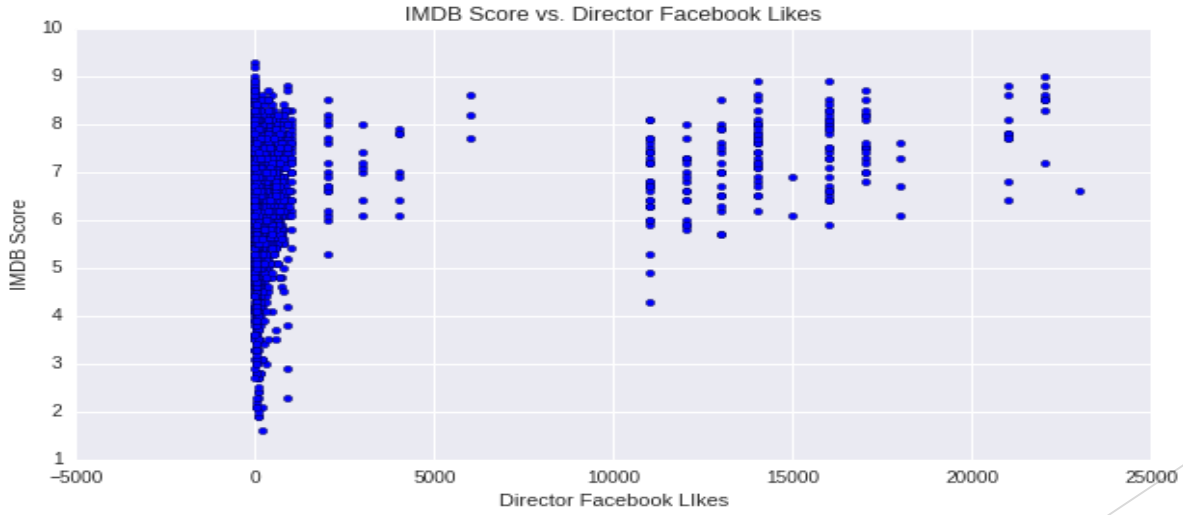




Movies below a rating of 6 won't gross more than 300 million dollars

Exploratory Data Analysis





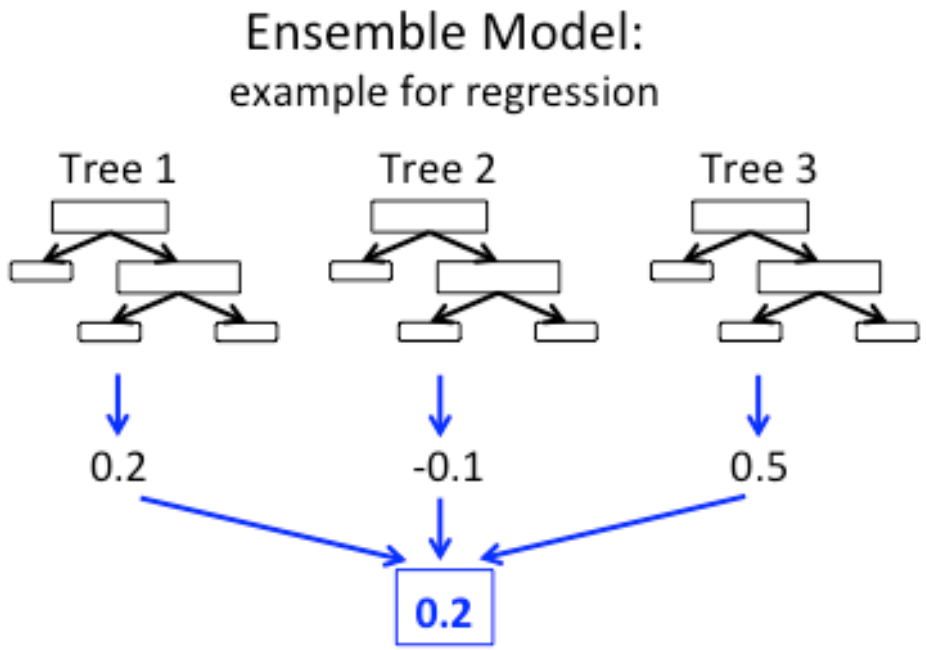
Using Linear Regression

- Used least squares model
- R^2 = 27.7%
- Feature Engineering didn't make a big difference

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
const	5.5829	0.162	34.412	0.000	5.265 5.901
director_facebook_likes	4.064e-05	5.04e-06	8.060	0.000	3.08e-05 5.05e-05
actor_1_facebook_likes	9.32e-06	1.76e-06	5.307	0.000	5.88e-06 1.28e-05
actor_2_facebook_likes	9.933e-06	4.63e-06	2.147	0.032	8.61e-07 1.9e-05
duration	0.0124	0.001	14.541	0.000	0.011 0.014
Action	-0.1894	0.046	-4.127	0.000	-0.279 -0.099
Adventure	0.1774	0.049	3.635	0.000	0.082 0.273
Comedy	-0.1979	0.043	-4.607	0.000	-0.282 -0.114
Horror	-0.3206	0.062	-5.198	0.000	-0.442 -0.200
Drama	0.3233	0.041	7.840	0.000	0.242 0.404
Documentary	0.4732	0.156	3.042	0.002	0.168 0.778
Thriller	-0.1411	0.043	-3.278	0.001	-0.226 -0.057
R	-0.5116	0.129	-3.979	0.000	-0.764 -0.259
PG-13	-0.8963	0.129	-6.924	0.000	-1.150 -0.642
PG	-0.7573	0.134	-5.660	0.000	-1.020 -0.495
G	-0.4417	0.168	-2.631	0.009	-0.771 -0.113

Experimenting with other methods

- RandomForest, CART, ridge models, Adaptive Boosting, lasso model, extratrees, gradient boosted regression
- Best R^2 value = 37%
- The most important variables are the amount of director facebook likes, duration, budget, and movie facebook likes



Conclusion

- Take social media influence into account
- Genre, MPAA ratings, and actors do not affect IMDB ratings
- It is difficult to accurately predict ratings