Machine Learning – Movie Success Prediction

Gabriel Solomon Holland

NORTHWEST
MISSOURI STATE UNIVERSITY

The goal of this was to see if I can predict how profitable a movie would be based on the score it received and the budget, essentially, can you buy a profitable movie?

Used a spreadsheet of 6772 movies off Kaggle. Cleaned the data by removing non-numeric data such as the writer, star actor, director, etc. Using clustering, linear regression, pipeline, and ensemble learning I tried to predict the profit of a movie using the budget and score.

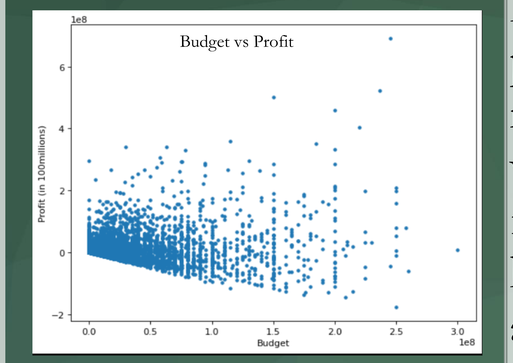
- Initial Graphs show little correlation between budget and profit, score and profit, and genre and profit. (See to the right).
- Linear Regression only scored a 51% accuracy with predicting profit based on budget
  - Had a bias of 5.7 million.
  - Overall not bad, considering how ambiguous the correlations were, 50% isn't good but for what we were using, not bad.
- Clustering could only manage to split it into 2 clusters no matter what I did, clustering couldn't handle this dataset and got me nowhere.
  - Clustering was not only unable to properly cluster the data, it was unable to make any predictions
- A Tree classifier worked best. It was able to predict profit with 65% accuracy using score and budget yet when feeding score and runtime into the same classifier it completely broke it.
  - Probably our best bet, 65% is ultimately not great and is in no way a reliable predictor but considering how impossible my project was, I think 65% is amazing.
- An SVC classifier worked well on score and runtime with a shocking 80% accuracy, yet I believe this is due to an error because score and budget completely broke the SVC, winding up with a .2% accuracy with 950% precision.
  - SVC managed to, with 80% accuracy, predict the profit of a movie based on the score and the runtime. That is so weirdly out of place that I don't think this is a reliable predictor. The fact that runtime has such a monumental effect on the classification just doesn't seem possible. Most movies fall within the same 1.5-2 hour runtime mark and I don't think it's possible to, within 30 minutes, help determine how profitable a movie will be, especially when disregarding the original theory of this project.
- Pipeline learning gave us an embarrassing 18% accuracy, and the larger the training set the lower the accuracy. Not remotely usable
- Ensemble and Neural Nets refused to run in a reasonable time
  - running at 1% training set the Ensemble and NN were unable to run. I think both of these would get a 70% or above accuracy because ensemble would implement our 65% tree classifier and more, which surely would give us >70%. Neural Nets seem to solve every problem so I really wanted to see what it could do

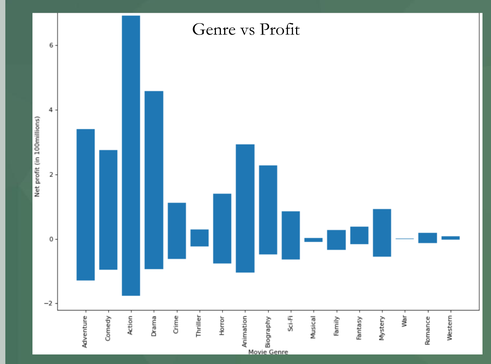
Raw Data:

	budget	company	country	director	genre	gross	name	rating	released	runtime	score	star	votes	writer	year	Profit
0	8000000	Columbia Pictures Corporation	USA	Rob Reiner	Adventure	52287414	Stand by Me	R	8/22/1986	89	8.1	Wil Wheaton	299174	Stephen King	1986	44287414
1	6000000	Paramount Pictures	USA	John Hughes	Comedy	70136369	Ferris Bueller's Day Off	PG-13	6/11/1986	103	7.8	Matthew Broderick	264740	John Hughes	1986	64136369
2	15000000	Paramount Pictures	USA	Tony Scott	Action	179800601	Top Gun	PG	5/16/1986	110	6.9	Tom Cruise	236909	Jim Cash	1986	164800601
3	18500000	Twentieth Century Fox Film Corporation	USA	James Cameron	Action	85160248	Aliens	R	7/18/1986	137	8.4	Sigourney Weaver	540152	James Cameron	1986	66660248
4	9000000	Walt Disney Pictures	USA	Randal Kleiser	Adventure	18564613	Flight of the Navigator	PG	8/1/1986	90	6.9	Joey Cramer	36636	Mark H. Baker	1986	9564613
5	6000000	Hemdale	UK	Oliver Stone	Drama	138530565	Platoon	R	2/6/1987	120	8.1	Charlie Sheen	317585	Oliver Stone	1986	132530565
6	25000000	Henson Associates (UA)	UK	Jim Henson	Adventure	12729917	Labyrinth	PG	6/27/1986	101	7.4	David Bowie	102879	Dennis Lee	1986	-12270083
7	6000000	De Laurentiis Entertainment Group (DEG)	USA	David Lynch	Drama	8551228	Blue Velvet	R	10/23/1986	120	7.8	Isabella Rossellini	146768	David Lynch	1986	2551228

Cleaned Data Description

	budget	gross	runtime	score	votes	Profit
count	6.772000e+03	6.772000e+03	6772.000000	6772.000000	6.772000e+03	6.772000e+03
mean	2.470267e+07	3.366642e+07	106.604696	6.376167	7.162959e+04	8.963749e+06
std	3.711686e+07	5.835288e+07	18.055645	1.003988	1.308859e+05	4.121820e+07
min	0.000000e+00	3.090000e+02	50.000000	1.500000	2.700000e+01	-1.769219e+08
25%	0.000000e+00	1.533799e+06	95.000000	5.800000	7.679750e+03	-5.305438e+06
50%	1.100000e+07	1.225487e+07	102.000000	6.400000	2.631700e+04	9.265295e+05
75%	3.200000e+07	4.012423e+07	115.000000	7.100000	7.635175e+04	1.411235e+07
max	3.000000e+08	9.366622e+08	366.000000	9.300000	1.861666e+06	6.916622e+08





Realistically you cannot predict how profitable a movie will be. Movies follow very few trends on the financials of a movie versus how popular it will be. There are movies that make millions and movies that lose millions with similar budgets and scores. A tree classifier was able to get a 65% accuracy with predictions, which is impressive and probably the closest we'll get to an answer. 65% isn't good enough to make full predictions on but for a general range of movies it gives us an idea of where it will probably land. There are too many outliers in the data to make a good prediction.

Kaggle.com for datasets, Jupyter Lab for IDE, Sklearn from Python Libraries for machine learning tools.

Full Code Here:

https://github.com/GabrielSolomonHolland/GSH-Machine-Learning-Project