Predicting IMDb Ratings via Regression and Deep Learning

Patricio Contreras
30 April, 2021

Outline

- Business Problem
- Data and Methods
- Results
- Limitations and Conclusions
- Future Work

Business Problem

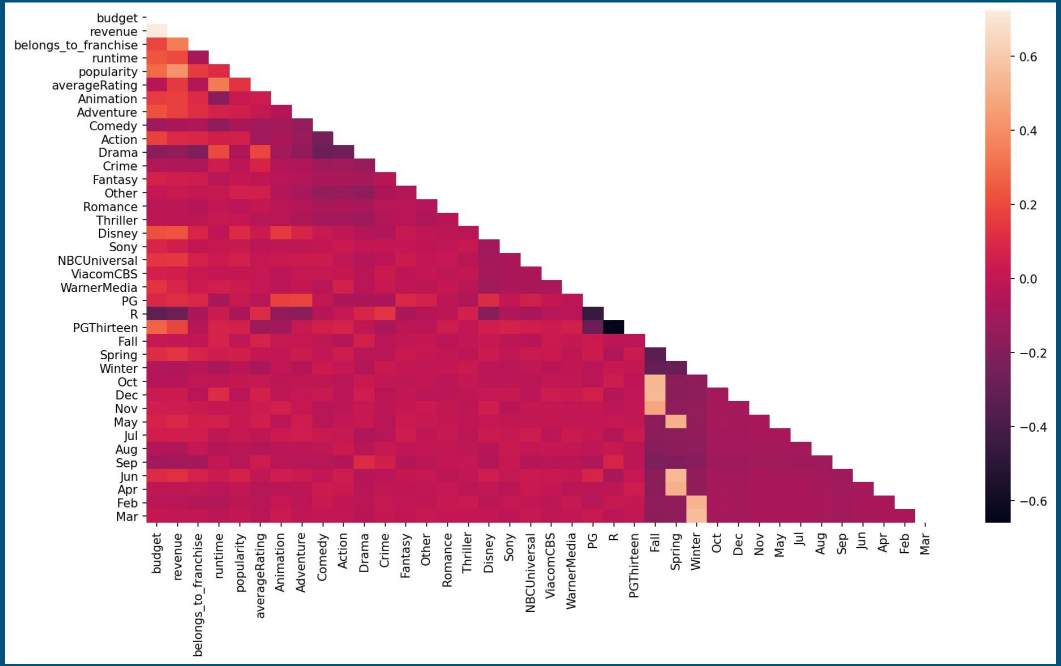
- **Predict a film’s IMDb rating given a select number of features**
- Minimise the risk of producing a “razzie”
- Determine the features that have the biggest impact on the film’s IMDb rating

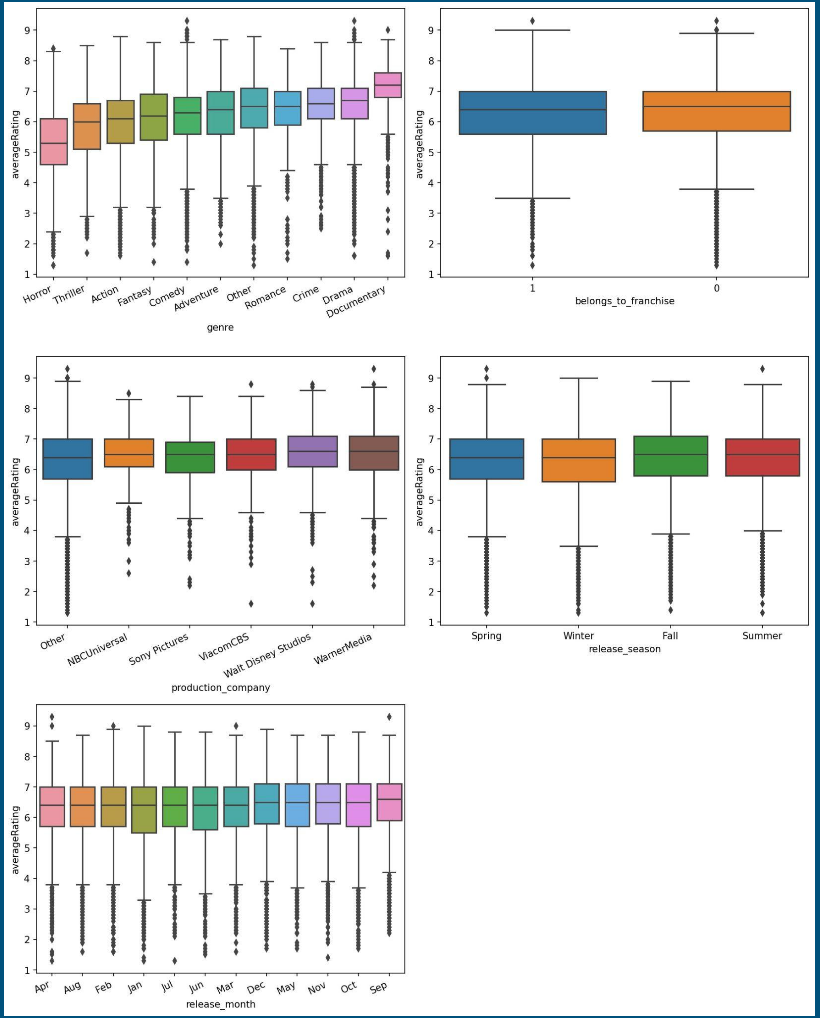
Data and Methods

- Dataset obtained from Kaggle's The Movies Dataset
- Contains information on over 40,000 movies such as genre, release date, IMDb rating, etc.
- Methods used: exploratory data analysis, regression modelling, neural networks

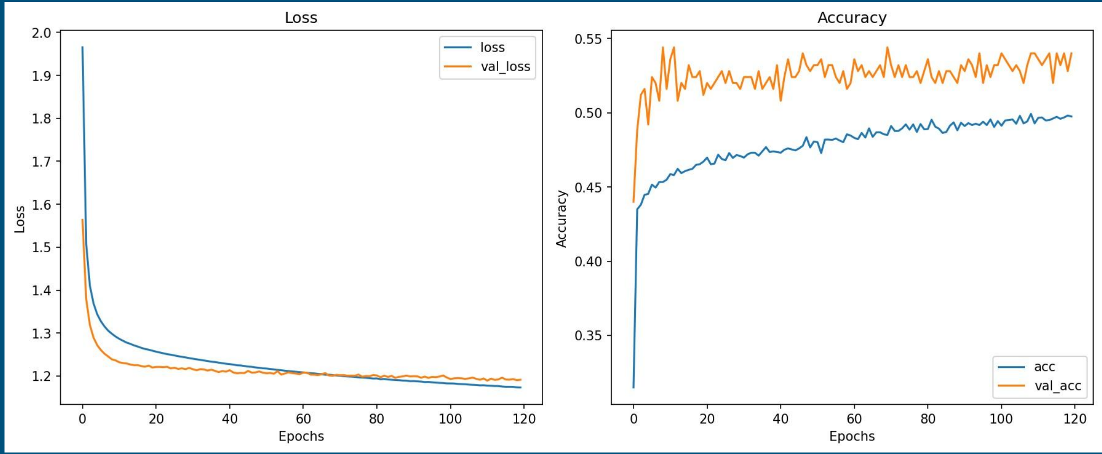
	title	genre	belongs_to_franchise	production_company	runtime	popularity	release_season	release_month	averageRating
0	Toy Story	Other	1	Walt Disney Studios	81.0	21.946943	Fall	Oct	8.3
1	Jumanji	Adventure	1	Sony Pictures	104.0	17.015539	Fall	Dec	7.0
2	Grumpier Old Men	Romance	1	Other	101.0	11.712900	Fall	Dec	6.7
3	Waiting to Exhale	Comedy	0	Walt Disney Studios	127.0	3.859495	Fall	Dec	6.0
4	Father of the Bride Part II	Comedy	1	Other	106.0	8.387519	Winter	Feb	6.1

Results


budget
revenue
belongs_to_franchise
runtime
popularity
averageRating
Animation
Adventure
Comedy
Action
Drama
Crime
Fantasy
Other
Romance
Thriller
Disney
Sony
NBCUniversal
ViacomCBS
WarnerMedia
PG
R
PGThirteen
Fall
Spring
Winter
Oct
Dec
Nov
May
Jul
Aug
Sep
Jun
Apr
Feb
Mar
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


averageRating
genre
Horror
Thriller
Action
Fantasy
Comedy
Adventure
Other
Romance
Crime
Drama
Documentary
belongs_to_franchise
1
0
production_company
Other
NBCUniversal
Sony Pictures
ViacomCBS
Walt Disney Studios
WarnerMedia
release_season
Spring
Winter
Fall
Summer
release_month
Apr
Aug
Feb
Jan
Jul
Jun
Mar
Dec
May
Nov
Oct
Sep

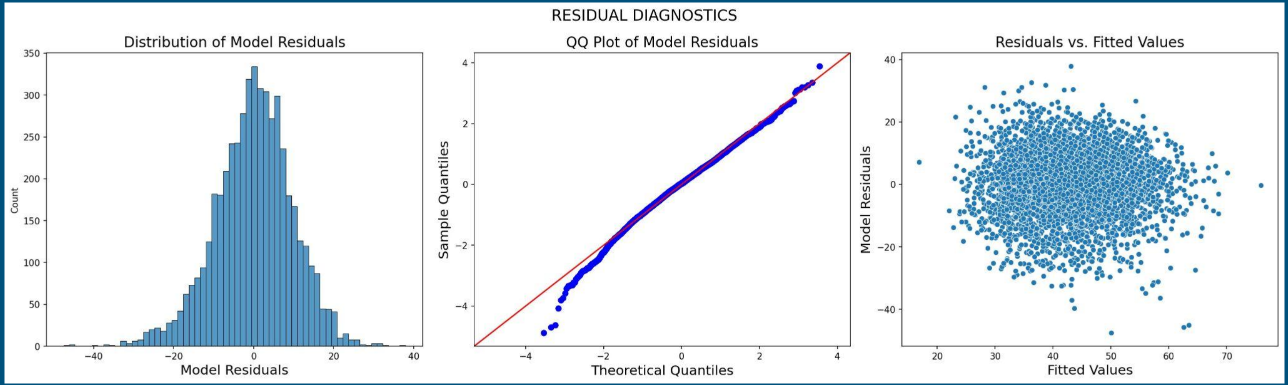
Results



Dep. Variable:	averageRating_squared	**R-squared:**	0.370
Model:	OLS	**Adj. R-squared:**	0.367
Method:	Least Squares	**F-statistic:**	139.2
Date:	Thu, 29 Apr 2021	**Prob (F-statistic):**	0.00
Time:	16:44:28	**Log-Likelihood:**	-18509.
No. Observations:	5009	**AIC:**	3.706e+04
Df Residuals:	4987	**BIC:**	3.721e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	10.4847	1.372	7.640	0.000	7.794	13.175
belongs_to_franchise	-3.0963	0.359	-8.626	0.000	-3.800	-2.393
runtime	0.2537	0.010	25.875	0.000	0.234	0.273
budget_fourth	-0.2956	0.010	-31.037	0.000	-0.314	-0.277
revenue_fifth	0.3444	0.020	17.550	0.000	0.306	0.383
sqrt_popularity	3.7822	0.184	20.596	0.000	3.422	4.142
Animation	14.1255	1.085	13.023	0.000	11.999	16.252
Adventure	8.0535	0.793	10.154	0.000	6.499	9.608
Comedy	5.4510	0.647	8.422	0.000	4.182	6.720
Action	4.8501	0.663	7.313	0.000	3.550	6.150
Drama	9.0318	0.657	13.756	0.000	7.745	10.319
Crime	9.0464	0.844	10.719	0.000	7.392	10.701
Fantasy	7.6522	1.033	7.407	0.000	5.627	9.678
Other	8.5871	0.768	11.178	0.000	7.081	10.093
Romance	7.5661	1.095	6.912	0.000	5.420	9.712
Thriller	3.7258	0.902	4.130	0.000	1.957	5.494
PG	-4.3846	0.863	-5.081	0.000	-6.076	-2.693
R	-2.5119	0.858	-2.929	0.003	-4.193	-0.830
PGThirteen	-5.4071	0.873	-6.192	0.000	-7.119	-3.695
Fall	1.0212	0.374	2.729	0.006	0.288	1.755
Spring	0.2363	0.384	0.615	0.539	-0.517	0.989
Winter	-0.7173	0.396	-1.812	0.070	-1.493	0.059

Omnibus:	117.218	**Durbin-Watson:**	1.815
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	152.383
Skew:	-0.290	**Prob(JB):**	8.14e-34
Kurtosis:	3.628	**Cond. No.**	2.01e+03



Limitations and Conclusions

- Dataset contained several missing values for features like budget and revenue
- Both models were heavily dependent on categorical data
- IMDb rating was not really affected by our categorical features
- Retrieving additional data was difficult and time-expensive

Future Work

- Natural Language Processing (NLP) could be used to identify key words and predict the IMDb rating
- Compare results between the same analysis done on TV shows
- Consider more continuous features! (opening weekend box office, marketing expenses)

Thank You!

Email: pcontreras1797@gmail.com
Github: @p-contreras
LinkedIn: linkedin.com/in/pcontreras97/