Predicting Movie Performance

A Look at Two Metrics:

Domestic Gross & Oscar Nominations

Objectives

- Predict Domestic Gross as accurately as possible
- Attempt a rough prediction of Oscar Nominations

Leveraging Available Data Sources

- Scraped Data Sources:
  - Box Office Mojo
  - The Numbers
- Other Data:
  - IMDB
- None were transformed for interpretability

- Data Tools:
  - Scikit Learn
  - Seaborn
  - Other Python Libraries

Feature Engineering

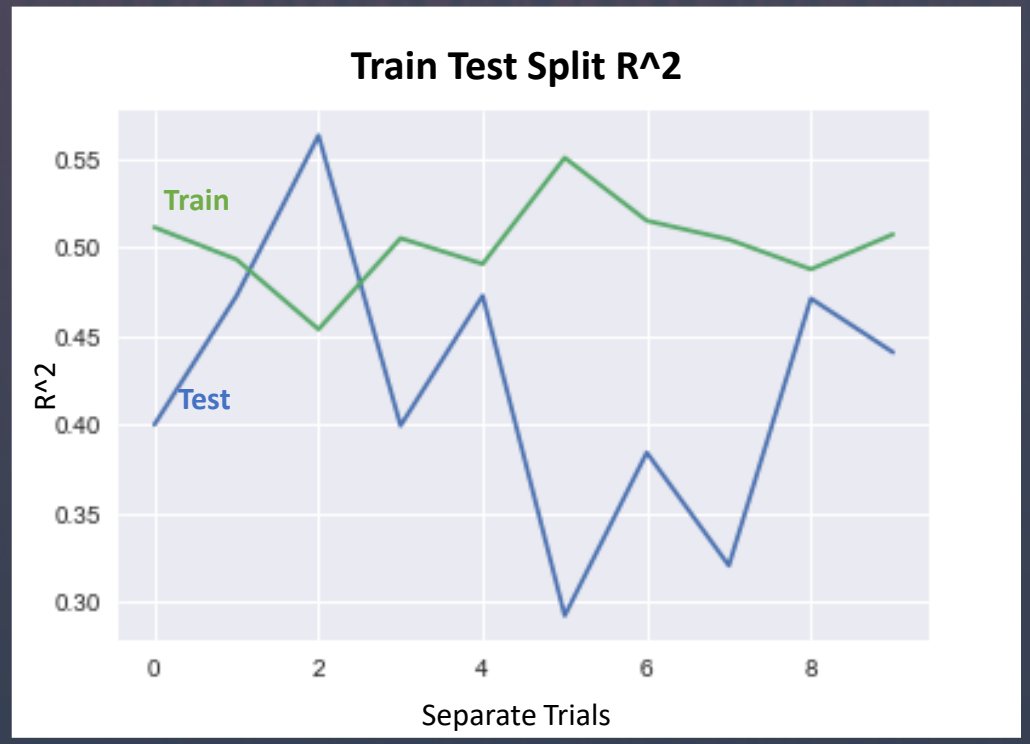
- Goal:
  - To convert Director, Actor, Writer, and Producer categorical data into useful data
- Method:
  - For each movie create cumulative metrics based on previous ROI and Oscars

	name	release_date	title	dom_gross	oscar_noms	oscars	final_budget	dom_roi
21662	Leonardo DiCaprio	1995-02-10	The Quick and the Dead	1.413773e+07	1.0	0.0	1.100000e+07	0.285248
21663	Leonardo DiCaprio	1995-04-21	The Basketball Diaries	3.277426e+07	1.0	0.0	4.300000e+07	-0.237808
21667	Leonardo DiCaprio	1997-12-19	Titanic	9.465014e+07	3.0	0.0	6.600000e+07	0.434093
21668	Leonardo DiCaprio	1998-03-13	The Man in the Iron Mask	6.954383e+08	17.0	11.0	2.660000e+08	1.614430
21669	Leonardo DiCaprio	1998-11-20	Celebrity	7.524072e+08	17.0	11.0	3.010000e+08	1.499692
21670	Leonardo DiCaprio	2000-02-11	The Beach	7.574859e+08	17.0	11.0	3.130000e+08	1.420083
21672	Leonardo DiCaprio	2002-12-20	Gangs of New York	8.370559e+08	17.0	11.0	4.130000e+08	1.026770
21673	Leonardo DiCaprio	2002-12-25	Catch Me If You Can	9.148679e+08	27.0	11.0	5.100000e+08	0.793859
21674	Leonardo DiCaprio	2006-10-06	The Departed	1.079483e+09	29.0	11.0	5.620000e+08	0.920789
21675	Leonardo DiCaprio	2006-12-08	Blood Diamond	1.211868e+09	34.0	15.0	6.520000e+08	0.858693

Predicting Domestic Gross - Baseline

- Most obvious model to beat.
- R^2 =.43

Predicting Domestic Gross - Other Features



- Added Features
  - Rating
  - IMDB Rating
  - Cast and Crew Engineered Features
- Linear Regression
- R^2 ~ .45

Predicting Domestic Gross - Lasso

	Coefs
imdb_rating^2	5.733725e+05
actor_cum_dom_roi writer_cum_oscar_noms	3.087157e+05
final_budget producer_cum_dom_roi	1.337961e-01
final_budget director_cum_dom_roi	1.304112e-01
imdb_rating final_budget	1.201354e-01
final_budget actor_cum_dom_roi	3.782103e-02
final_budget producer_cum_oscar_noms	1.444071e-03
final_budget actor_cum_oscars	-5.559616e-03
director_cum_oscars producer_cum_oscars	-4.169952e+04
writer_cum_oscars^2	-4.662014e+04
director_cum_oscar_noms writer_cum_oscars	-6.363441e+04
actor_cum_oscars writer_cum_dom_roi	-1.718337e+05
actor_cum_oscar_noms	-2.147414e+05
actor_cum_oscars producer_cum_oscars	-2.357436e+05
actor_cum_oscars	-1.498995e+06

- Used Lasso Regression to improve fit
- R^2 ~ .5

Predicting Domestic Gross - Lasso

	Coefs
imdb_rating^2	5.733725e+05
actor_cum_dom_roi writer_cum_oscar_noms	3.087157e+05
final_budget producer_cum_dom_roi	1.337961e-01
final_budget director_cum_dom_roi	1.304112e-01
imdb_rating final_budget	1.201354e-01
final_budget actor_cum_dom_roi	3.782103e-02
final_budget producer_cum_oscar_noms	1.444071e-03
final_budget actor_cum_oscars	-5.559616e-03
director_cum_oscars producer_cum_oscars	-4.169952e+04
writer_cum_oscars^2	-4.662014e+04
director_cum_oscar_noms writer_cum_oscars	-6.363441e+04
actor_cum_oscars writer_cum_dom_roi	-1.718337e+05
actor_cum_oscar_noms	-2.147414e+05
actor_cum_oscars producer_cum_oscars	-2.357436e+05
actor_cum_oscars	-1.498995e+06

- Used Lasso Regression to improve fit
- R^2 ~ .5

Predicting Domestic Gross - Lasso

	Coefs
imdb_rating^2	5.733725e+05
actor_cum_dom_roi writer_cum_oscar_noms	3.087157e+05
final_budget producer_cum_dom_roi	1.337961e-01
final_budget director_cum_dom_roi	1.304112e-01
imdb_rating final_budget	1.201354e-01
final_budget actor_cum_dom_roi	3.782103e-02
final_budget producer_cum_oscar_noms	1.444071e-03
final_budget actor_cum_oscars	-5.559616e-03
director_cum_oscars producer_cum_oscars	-4.169952e+04
writer_cum_oscars^2	-4.662014e+04
director_cum_oscar_noms writer_cum_oscars	-6.363441e+04
actor_cum_oscars writer_cum_dom_roi	-1.718337e+05
actor_cum_oscar_noms	-2.147414e+05
actor_cum_oscars producer_cum_oscars	-2.357436e+05
actor_cum_oscars	-1.498995e+06

- Used Lasso Regression to improve fit
- R^2 ~ .5

Predicting Domestic Gross - Lasso

	Coefs
imdb_rating^2	5.733725e+05
actor_cum_dom_roi writer_cum_oscar_noms	3.087157e+05
final_budget producer_cum_dom_roi	1.337961e-01
final_budget director_cum_dom_roi	1.304112e-01
imdb_rating final_budget	1.201354e-01
final_budget actor_cum_dom_roi	3.782103e-02
final_budget producer_cum_oscar_noms	1.444071e-03
final_budget actor_cum_oscars	-5.559616e-03
director_cum_oscars producer_cum_oscars	-4.169952e+04
writer_cum_oscars^2	-4.662014e+04
director_cum_oscar_noms writer_cum_oscars	-6.363441e+04
actor_cum_oscars writer_cum_dom_roi	-1.718337e+05
actor_cum_oscar_noms	-2.147414e+05
actor_cum_oscars producer_cum_oscars	-2.357436e+05
actor_cum_oscars	-1.498995e+06

- Used Lasso Regression to improve fit
- R^2 ~ .5

Predicting Oscar Nominations

- Used Lasso regression and polynomials of degree 2
- R^2 = .45
- Used qualitative features
- Most features were kept by Lasso regression

Predicting Oscar Nominations

- Used Lasso regression and polynomials of degree 2
- R^2 = .45
- Used qualitative features
- Most features were kept by Lasso regression

	title	oscar_noms	predictions
16164	Star Wars: The Last Jedi	4.0	4.326115
4752	Dunkirk	8.0	3.980012
3617	Coco	2.0	3.282210
17832	Wind River	0.0	2.477237
4874	Darkest Hour	6.0	2.421379
7120	The Greatest Showman	1.0	2.393213
797	The Big Sick	1.0	2.169370
14058	The Shape of Water	13.0	2.093809
7683	Get Out	4.0	2.041696
18287	Wonder	1.0	1.982541
13257	Phantom Thread	6.0	1.967421
7442	Guardians of the Galaxy Vol. 2	1.0	1.951976
15318	Three Billboards Outside Ebbing, Missouri	7.0	1.906537
10621	Lady Bird	5.0	1.877765
4511	Detroit	0.0	1.571309
506	Baby Driver	3.0	1.540213

Predicting Oscar Nominations

- Used Lasso regression and polynomials of degree 2
- R^2 = .45
- Used qualitative features
- Most features were kept by Lasso regression

	title	oscar_noms	predictions
16164	Star Wars: The Last Jedi	4.0	4.326115
4752	Dunkirk	8.0	3.980012
3617	Coco	2.0	3.282210
17832	Wind River	0.0	2.477237
4874	Darkest Hour	6.0	2.421379
7120	The Greatest Showman	1.0	2.393213
797	The Big Sick	1.0	2.169370
14058	The Shape of Water	13.0	2.093809
7683	Get Out	4.0	2.041696
18287	Wonder	1.0	1.982541
13257	Phantom Thread	6.0	1.967421
7442	Guardians of the Galaxy Vol. 2	1.0	1.951976
15318	Three Billboards Outside Ebbing, Missouri	7.0	1.906537
10621	Lady Bird	5.0	1.877765
4511	Detroit	0.0	1.571309
506	Baby Driver	3.0	1.540213

Issues, Assumptions, and Next Steps

- Create a proper GLM Model with Poisson Regression
- Investigate potential data leakages
- Get more consistent data from one source
- Main Issue: Bias in the Data

Thank You – Q&A

Feature Engineering - Appendix

- Goal:
  - To convert Director, Actor, Writer, and Producer categorical data into useful data
- Method:
  - For each movie create cumulative metrics based on previous ROI and Oscars

	name	release_date	title	dom_gross	oscar_noms	oscars	final_budget	dom_roi
16278	Jennifer Lawrence	2009-09-18	The Burning Plain	2.226160e+05	0.0	0.0	2.000000e+07	-0.988869
16279	Jennifer Lawrence	2010-06-11	Winter's Bone	6.754119e+06	4.0	0.0	2.200000e+07	-0.692995
16280	Jennifer Lawrence	2011-05-06	The Beaver	7.724935e+06	4.0	0.0	4.300000e+07	-0.820350
16281	Jennifer Lawrence	2011-06-03	X-Men: First Class	1.541332e+08	4.0	0.0	2.030000e+08	-0.240723
16282	Jennifer Lawrence	2011-10-28	Like Crazy	1.575286e+08	4.0	0.0	2.032500e+08	-0.224951
16284	Jennifer Lawrence	2012-03-23	The Hunger Games	5.689347e+08	4.0	0.0	2.835000e+08	1.006824
16285	Jennifer Lawrence	2012-09-21	House at the End of The Street	6.005466e+08	4.0	0.0	2.904000e+08	1.067998
16286	Jennifer Lawrence	2012-11-16	Silver Linings Playbook	7.326396e+08	12.0	1.0	3.114000e+08	1.352728
16287	Jennifer Lawrence	2013-11-22	The Hunger Games: Catching Fire	1.157308e+09	12.0	1.0	4.414000e+08	1.621902
16288	Jennifer Lawrence	2013-12-13	American Hustle	1.307425e+09	22.0	1.0	4.814000e+08	1.715882
16289	Jennifer Lawrence	2014-05-23	X-Men: Days of Future Past	1.541347e+09	23.0	1.0	6.814000e+08	1.262030
16290	Jennifer Lawrence	2014-11-21	The Hunger Games: Mockingjay - Part 1	1.878483e+09	23.0	1.0	8.064000e+08	1.329468
16293	Jennifer Lawrence	2015-11-20	The Hunger Games: Mockingjay - Part 2	2.160560e+09	23.0	1.0	9.664000e+08	1.235678
16294	Jennifer Lawrence	2015-12-25	Joy	2.217011e+09	24.0	1.0	1.026400e+09	1.159987
16299	Jennifer Lawrence	2016-05-27	X-Men: Apocalypse	2.554794e+09	27.0	1.0	1.384400e+09	0.845416