COURSERA
Applied Data Science

Capstone Project

Movie earnings prediction

1. Background

Cinematography is a money generating business.
- The producer and film studio do not only make art, they also want to make money.
- The actors who salary is percentage of the profit.
- The investors who put money in the production.

They all want to ensure their investment will be profitable.

What will be the movie profit?

Film Studio

Actors

Producer

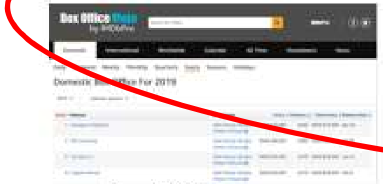
Investor

2. Data Source / Approach

Which information are available before the movie premiere to predict the earnings?
And how to get them?

Scrap popular portals to get available information.

Box Office Mojo

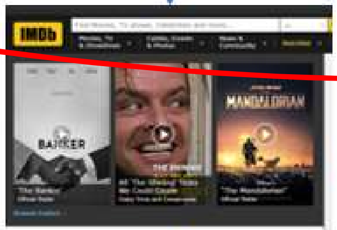


Beautiful Soup

Titles

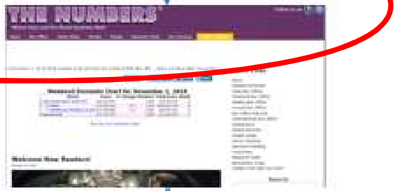
Top profitable movies per year

IMDB



requests

the-numbers.com

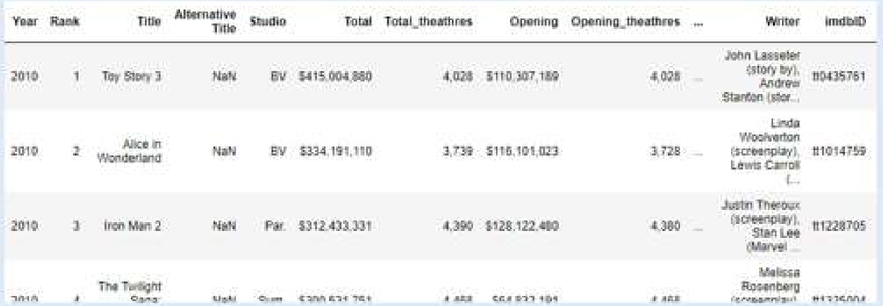


requests

- Director
- Producer
- Cast (Actors)
- Genre
- Runtime
- Language
- Country
- Rating
- Premiere date

- Budget

Top movies + details = dataset



Year	Rank	Title	Alternative Title	Studio	Total	Total_theathres	Opening	Opening_theathres	...	Writer	imdbID
2010	1	Toy Story 3	NaN	BV	$415,004,880	4,028	$110,307,189	4,028	...	John Lasseter (story by), Andrew Stanton (stor...	tt0435761
2010	2	Alice in Wonderland	NaN	BV	$334,191,110	3,739	$116,101,023	3,728	...	Linda Woolverton (screenplay), Lewis Carroll (...	tt1014759
2010	3	Iron Man 2	NaN	Par.	$312,433,331	4,390	$128,122,480	4,380	...	Justin Theroux (screenplay), Stan Lee (Marvel ...	tt1228705
2010	4	The Twilight Saga:	NaN	Sum.	$300,531,751	4,468	$64,832,191	4,468		Melissa Rosenberg (screenplay)	tt1325004

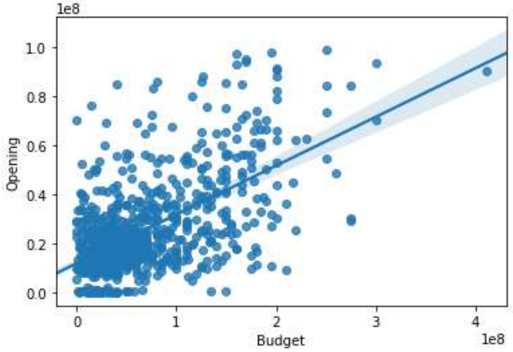
Create a dataset of most successful titles and their details in order as input for the prediction model

4. Exploratory data analysis

Available features

Numerical

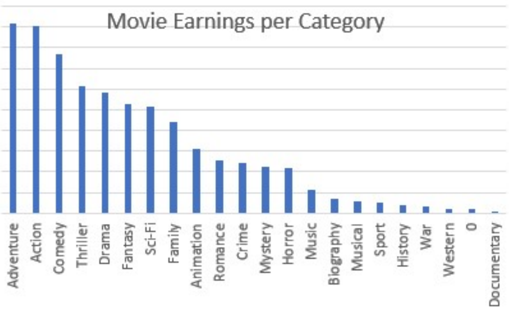
Example: Runtime, Budget



No strong relations between movie budget and its earning

Categorical

Example: Actors, Director, Plot, Genre



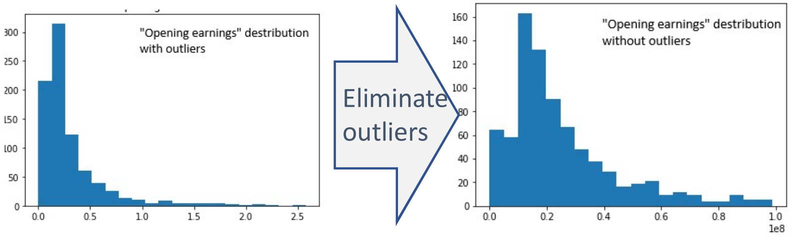
Problem: how to evaluate importance of:

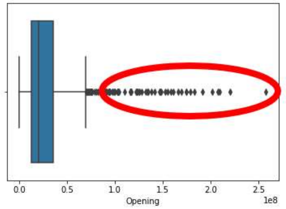
- Actors – 1778 enries
- Director – 539 entries
- Writer – 1478 enries
- Plot – 9814 words

These features need to be one hot encoded which creates high dimensionality sparse matrix

Target

Opening earnings





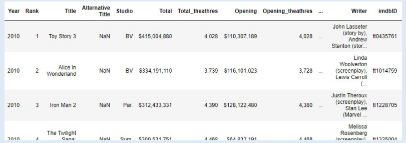
Use opening weekend earnings instead of total earning, as the result can be evaluated quicker.

5. Methodology

Preparation

Feature engineering

Dataset



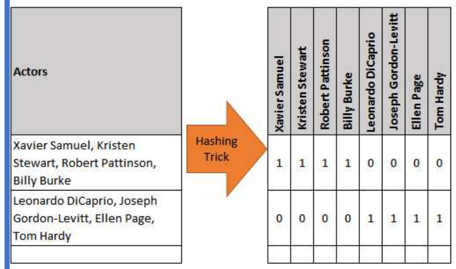
Year	Rank	Title	Alternative Title	Studio	Total	Total_theathres	Opening	Opening_theathres	...	Writer	imdbID
2010	1	Toy Story 3	NaN	BV	$415,004,880	4,028	$110,307,189	4,028	...	John Lasseter (story by), Andrew Stanton (stor...	tt0435761
2010	2	Alice in Wonderland	NaN	BV	$334,191,110	3,739	$116,101,023	3,728	...	Linda Woolverton (screenplay), Lewis Carroll (...	tt1014759
2010	3	Iron Man 2	NaN	Par.	$312,433,331	4,390	$128,122,480	4,380	...	Justin Theroux (screenplay), Stan Lee (Marvel ...	tt1228705
2010	4	The Twilight Saga:	NaN	Sum.	$300,531,751	4,468	$64,832,191	4,468	...	Melissa Rosenberg (screenplay)	tt1325004

Numerical

- Clean
- Convert to numeric type

Categorical

- Clean
- Hashing trick

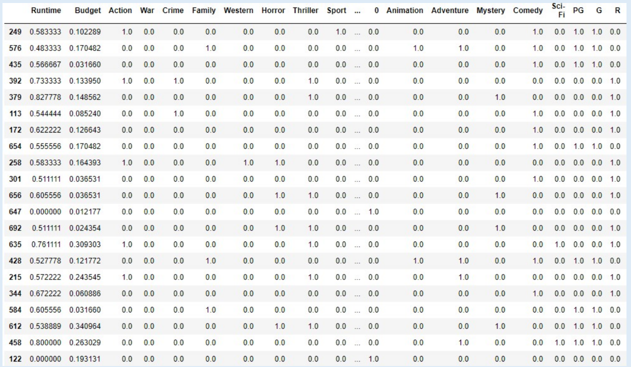


Actors
Xavier Samuel, Kristen Stewart, Robert Pattinson, Billy Burke
Leonardo DiCaprio, Joseph Gordon-Levitt, Ellen Page, Tom Hardy

Xavier Samuel	Kristen Stewart	Robert Pattinson	Billy Burke	Leonardo DiCaprio	Joseph Gordon-Levitt	Ellen Page	Tom Hardy
1	1	1	1	0	0	0	0
0	0	0	0	1	1	1	1

Combine in one big data input matrix

Dataset



	Runtime	Budget	Action	War	Crime	Family	Western	Horror	Thriller	Sport	...	0	Animation	Adventure	Mystery	Comedy	Sci-Fi	PG	G	R
249	0.583333	0.102289	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0
576	0.483333	0.170482	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	1.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0
435	0.566667	0.031660	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0
392	0.733333	0.133950	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
379	0.827778	0.148562	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
113	0.544444	0.085240	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
172	0.622222	0.126643	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
654	0.555556	0.170482	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0
258	0.583333	0.164393	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
301	0.511111	0.036531	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
656	0.605556	0.036531	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
647	0.000000	0.012177	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
692	0.511111	0.024354	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	...	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0
635	0.761111	0.309303	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
428	0.527778	0.121772	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	1.0	1.0	0.0	1.0	0.0	1.0	1.0	0.0
215	0.572222	0.243545	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	...	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
344	0.672222	0.060886	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0
584	0.605556	0.031660	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
612	0.538889	0.340964	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	...	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0
458	0.800000	0.263029	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0
122	0.000000	0.193131	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Dimension: 802 rows x 6443 columns

Sparse matrix

Min Max Scaler

Train Dataset

Split & apply MinMaxScaler

Test Dataset

6. Modelling

Which machine learning algorithm performs best, if the input is high dimensional sparse matrix?

Use different models and evaluate the best one based on R squared coefficient.

Model	R2 train data	R2 test data	Comment
Linear regression	0.820	-1.840	Linear model is not appropriate for the dataset
Decision Tree	0.238	0.185	The result is very poor
Neural Network	0.997	0.217	Overfitting
Gradient Boosting	0.671	0.435	Overfitting, however train/test result is closer
Support Vector Machine	1.0	0.0	SVM does not provide any significant result

Best performing models are Gradient Boosting and Neural Networks, however the result is still not satisfactory for model deployment. Possibly the result could be better with parameter tuning.

The reason may be the sparse data. Example:
Only 567 actors (out of 1778 total) played in more than 1 movie. This does not allow the model to use the training data, because test data contains new, different information.

7. Conclusion

The model does not perform well enough to allow deployment in business.
Reasons are:

- No linearly correlated features
- Sparse data: train data is not sufficient for modeling as test data contain new information.

Movie	Actual earnings $	Predicted earnings $	Delta
1	24’830’443	15’734’224	-36.6%
2	21’052’227	27’278’085	29.6%
3	10’609’795	24’447’134	130.4%

Not consistent enough.

Recommendation:
- create bigger training data set
- find a way to evaluate importance of each feature
- create hyper feature to better describe the data
- test other, more sophisticated algorithms
- fine tune model parameters