The Formula for a Blockbuster

Can movie producers predict their next hit?

Alex Chung
Shan He
Nathaniel Schub

School of Information, UC Berkeley

Problem:

Movie producers have trouble predicting the success of their investments.

King Arthur made $39.1 million and cost $175 million!

Solution:

Mine public data on movies to see if certain variables relate to box-office success.

Language	Actors	Region	Genre	Producer	Run time	Release	...

Early signals from our data:

It pays to **be an American producer.**

It pays to **make an adventure film.**

It pays to **make a mega-franchise film like Harry Potter.**

It pays to **release mid-year.**

It pays to **keep your runtime under 200 minutes.**

These findings deserve further analysis.

Acquisition and organization of information

Data Acquisition

TMDb API

TMDb is a community built movie and TV database

https://www.themoviedb.org/documentation/api

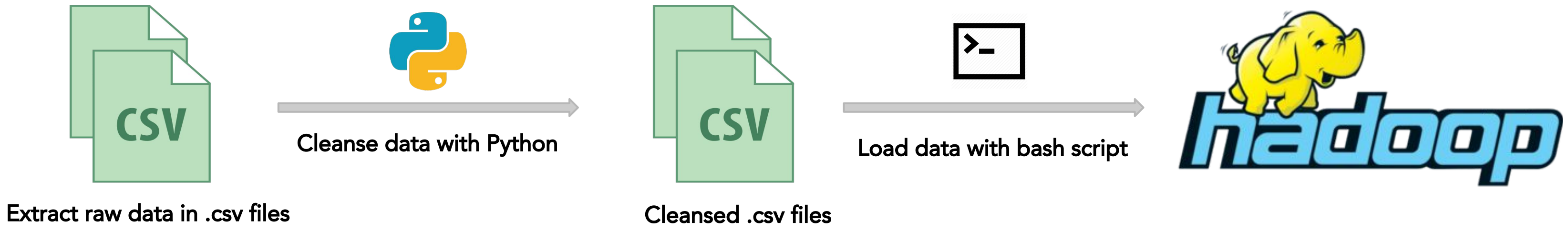
MovieLens | GroupLens

Rating datasets from the MovieLens web site, which feature a MovieLens 20M dataset
(20M ratings on 27,000 movies by 138,000 users)

https://grouplens.org/datasets/movielens/

Acquisition and organization of information

Data ETL



Architecture and Implementation Details

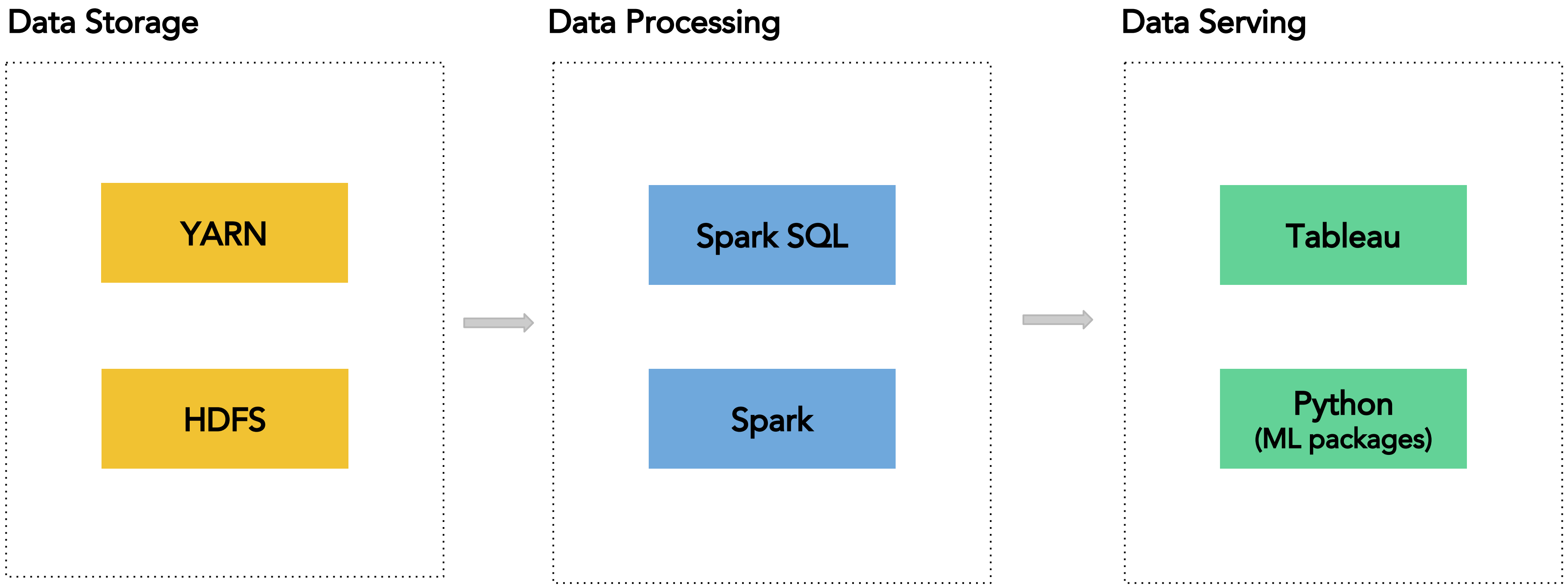


Tableau:
Key movie revenue analysis

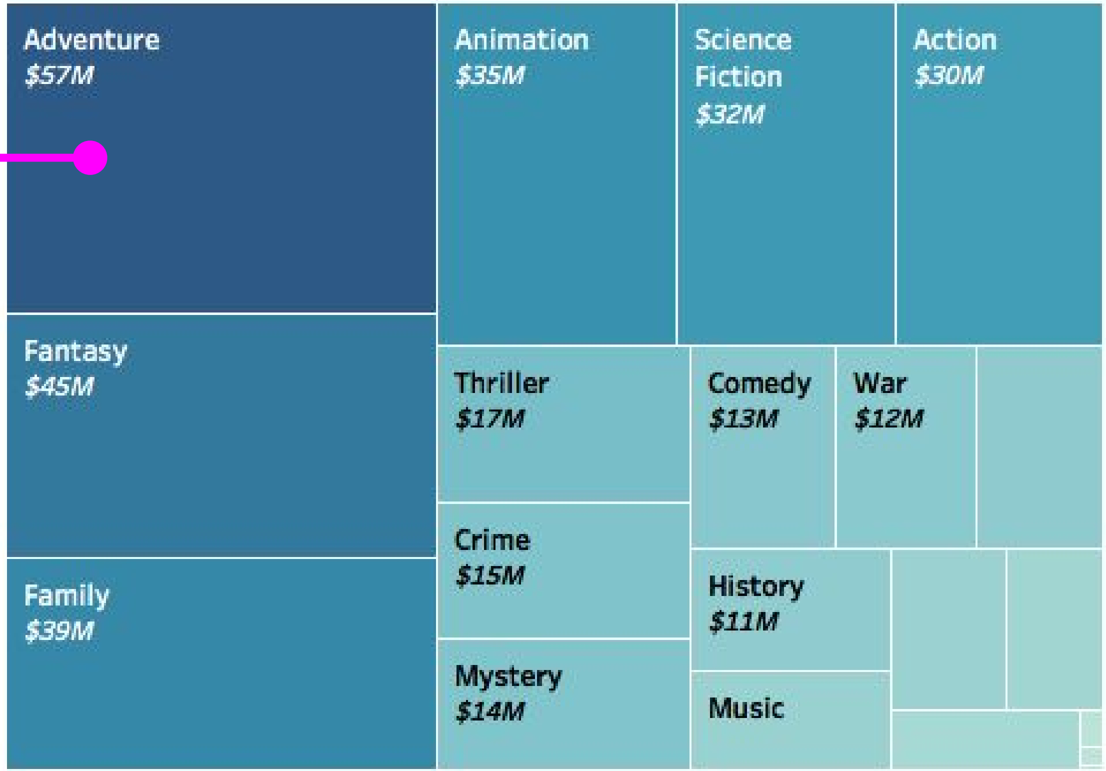
From the data collected, we looked into the relationship between Revenue and key variables of interest. And we have some interesting findings:

Findings: Total Revenue by Country

In our dataset, US movie producers generated the highest gross revenue.

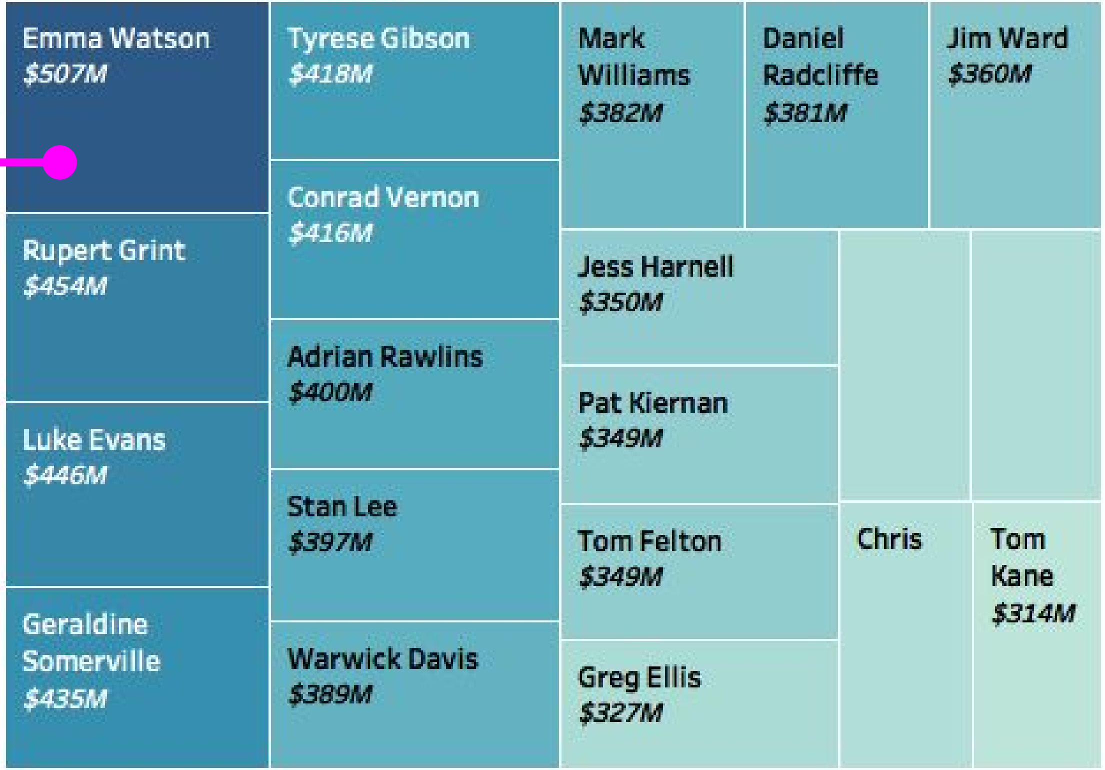
Findings: Average Revenue by Genre

Genre “Adventure” associates with the highest average revenue



Findings: Top 20 actors* by average revenue

Emma Watson leads all actors in highest average (associated) movie revenue



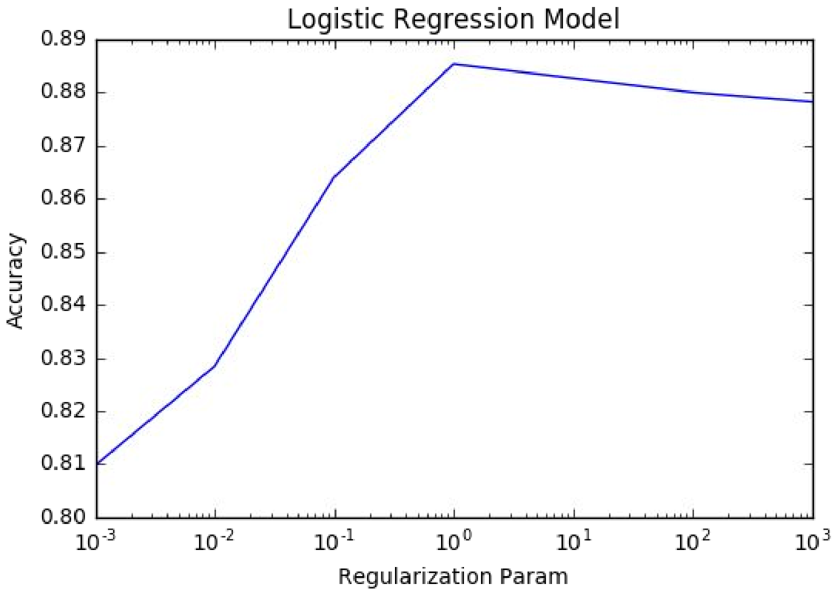
* actors casted in less than 10 movies are excluded from the list

Machine Learning Modeling

We chose a Logistic Regression model to help producers predict whether their next movie will be a success, based on:

- Title
- "Adult" movie or not
- Franchise or not
- Budget, adjusted for inflation
- Genre (action, romance, etc)
- Release Month

Accuracy for L2 Logistic Regression Model was >88%



Machine Learning Model

Logistic Model coefficients indicate which factors have the most effect on whether a movie is a blockbuster*

Increasing log odds: franchise, budget, animation, wedding, dragon… December, June

Decreasing log odds: war, next, wild, death, heaven, January, April

*defined as movies with at least 100M in income (revenue - budget) in our model. Definitions may vary from other sources

```
stack[stack[:,1].argsort()]

 [u'Thriller', u'0.111225257869'],
 [u'07', u'0.183723083853'],
 [u'four', u'0.186070645337'],
 [u'movie', u'0.19862401239'],
 [u'Comedy', u'0.265058721344'],
 [u'king', u'0.287361474103'],
 [u'love', u'0.319195235147'],
 [u'06', u'0.374026574929'],
 [u'Family', u'0.401259603749'],
 [u'12', u'0.417430445917'],
 [u'kill', u'0.586455979785'],
 [u'Romance', u'0.612692589653'],
 [u'with', u'0.675149767802'],
 [u'dragon', u'0.805016905578'],
 [u'wedding', u'0.90339667866'],
 [u'Animation', u'1.11412856526'],
 [u'budget_adj', u'1.33789462611e-08'
 [u'Franchise', u'1.54591126133']],
dtype='<U32')
```

```
stack[stack[:,1].argsort()]

[u'09', u'-0.325605602493'],
[u'death', u'-0.346711506818'],
[u'03', u'-0.362726006107'],
[u'Action', u'-0.367460655877'],
[u'Horror', u'-0.371909753977'],
[u'space', u'-0.382798084625'],
[u'04', u'-0.413441809567'],
[u'boys', u'-0.420444318114'],
[u'heaven', u'-0.438848597272'],
[u'meet', u'-0.512960480947'],
[u'legend', u'-0.588710115692'],
[u'blue', u'-0.619316291634'],
[u'01', u'-0.650842688166'],
[u'life', u'-0.654408525991'],
[u'dark', u'-1.07468133022'],
[u'wild', u'-1.4051654451'],
[u'next', u'-1.55507959666'],
[u'War', u'-1.80026895421'],
```

Machine Learning Model Testing

When we enter **Jumanji** into the logistic regression model, it gives Jumanji only a 40% chance of being a blockbuster.

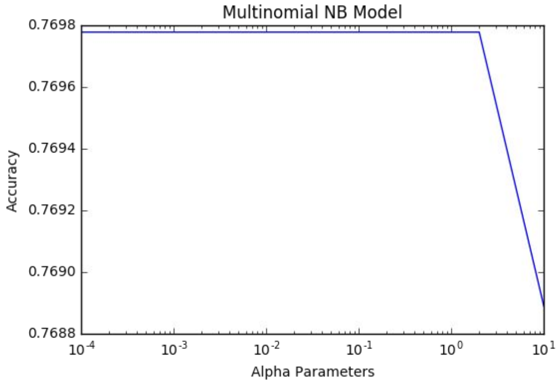
```
   title  is_blockbuster  adult  Franchise    budget_adj       genre  \
jumanji                1  FALSE          0  1.044501e+08  Adventure

runtime release_month
  104.0            12
```

Model is still a bit too conservative-only 231/1125 test movies were blockbusters, but and the model predicted that there would be 182

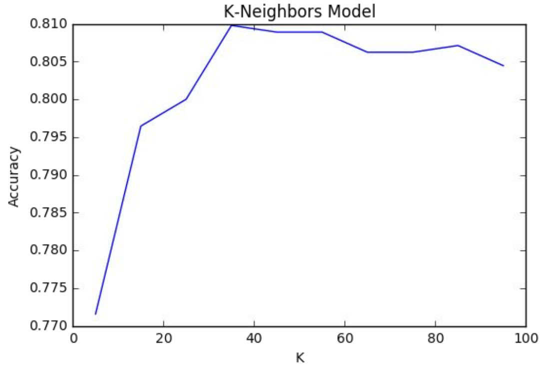
```
 [ 0.59593585  0.40406415]
```

Machine Learning Model Comparison

Multinomial Naive Bayes model was too aggressive, predicting that almost 300/1125 movies would be blockbusters



K-Nearest Neighbors Model was too conservative, predicting 139/1125 movies would be blockbusters



Scaling and Limitation

Scaling Strategies

1. Expanding dataset to global, non-English markets
2. Incorporating other sources of review data like Rotten Tomatoes
3. Incorporating other sources of revenue like video-on-demand sales
4. Analyzing live conversations on Twitter and other social media

Limitations

1. Lack of unobserved variables (e.g. creativity of plot, human values embodied in movies, acting skills) that can be highly relevant to movie revenue
2. Community-generated data can be infrequent and inconsistent
3. Lack of reporting on non-ticket (e.g. merchandise) revenue data

Q/A