
HOLLYWOOD

movieConnoisseur.io

Movies Dataset ML & NLP

Team -- Tyson's Big Cat

Alannah
Danni
Tianchi

Agenda

Objective

Methods

Models

Demo

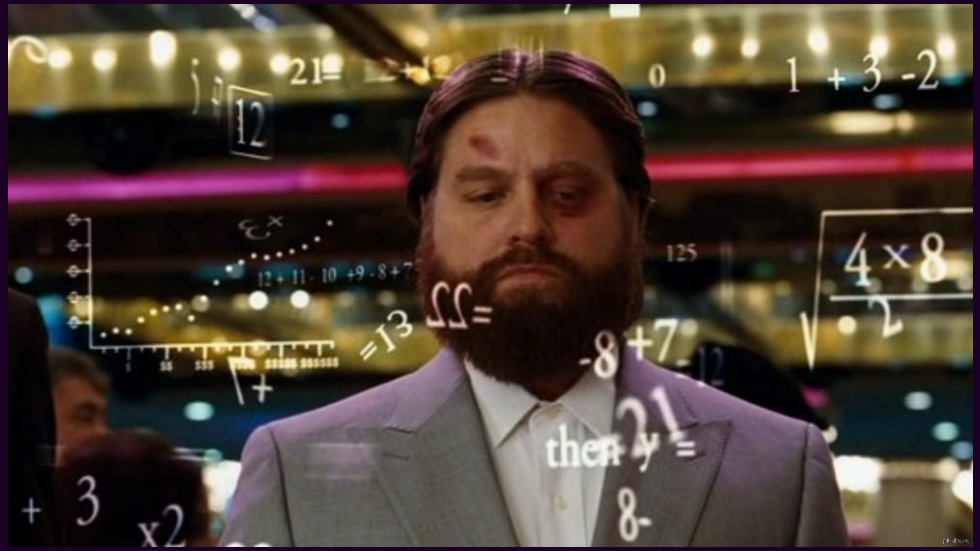
Limitations & Next Steps

Objective

- Predict movie box office
- Predict movie oscar nominations
- Predict movie MPA ratings

Methods

- **Data Collection**
  - Kaggle
  - OMDB API
- **Machine Learning**
  - Multiple Linear Regression
  - Logistic Regression
  - SVM
  - Random Forests
  - Deep Learning
- **NLP**
  - Movie plot
- **Predictor App**



Data Collection

- Our base dataset was a collection of movies data obtained from Kaggle
- 200 movies per year from 1986 - 2016
- Variables:

Budget	Company	Country	Director	Genre	Box Office	Name	Rating
Year	Released Date	Runtime	IMDB score	IMDB votes	Star	Actors	Writers
Action	Adventure	Fantasy	Sci-Fi	Crime	Drama	History	Comedy
Biography	Romance	Horror	Thriller	Adult	Film-Noir	Documentary	Musical
War	Animation	Family	Sport	Music	Mystery	Short	Western

Non genres

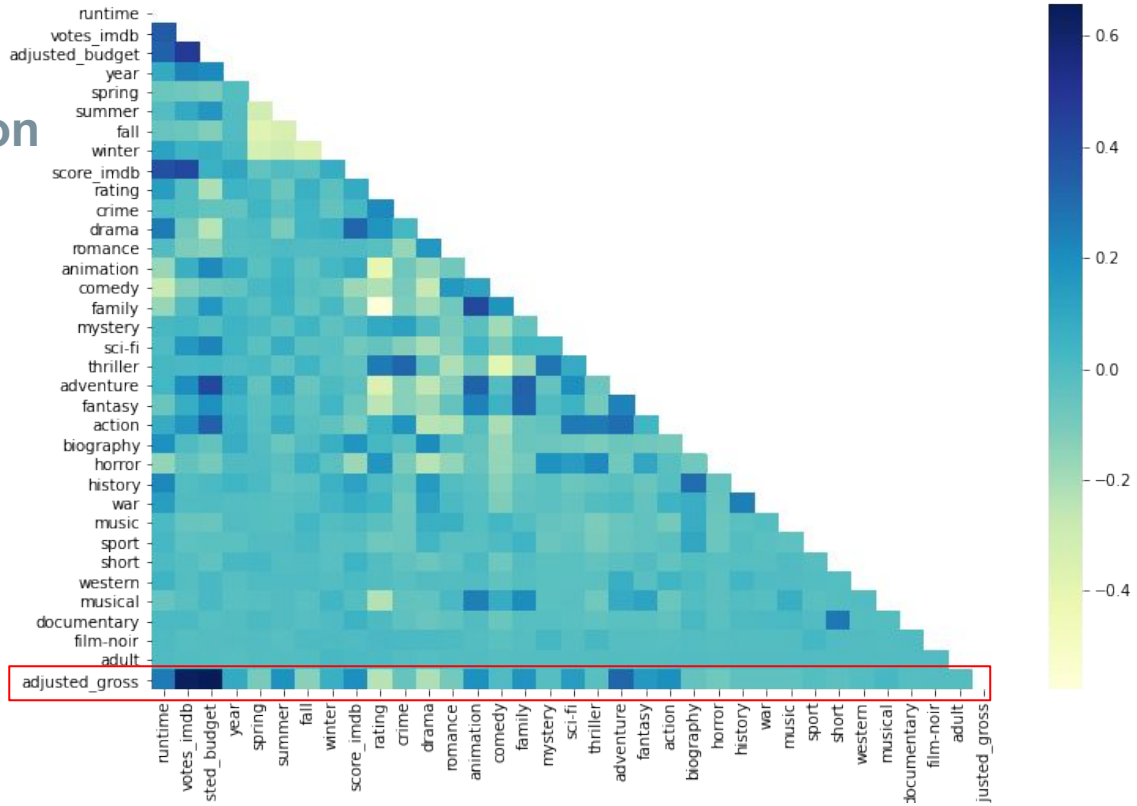
Genres

- We then used the OMDB api to retrieve additional information on each film, such as **plot, genre (multiple), awards and nominations**

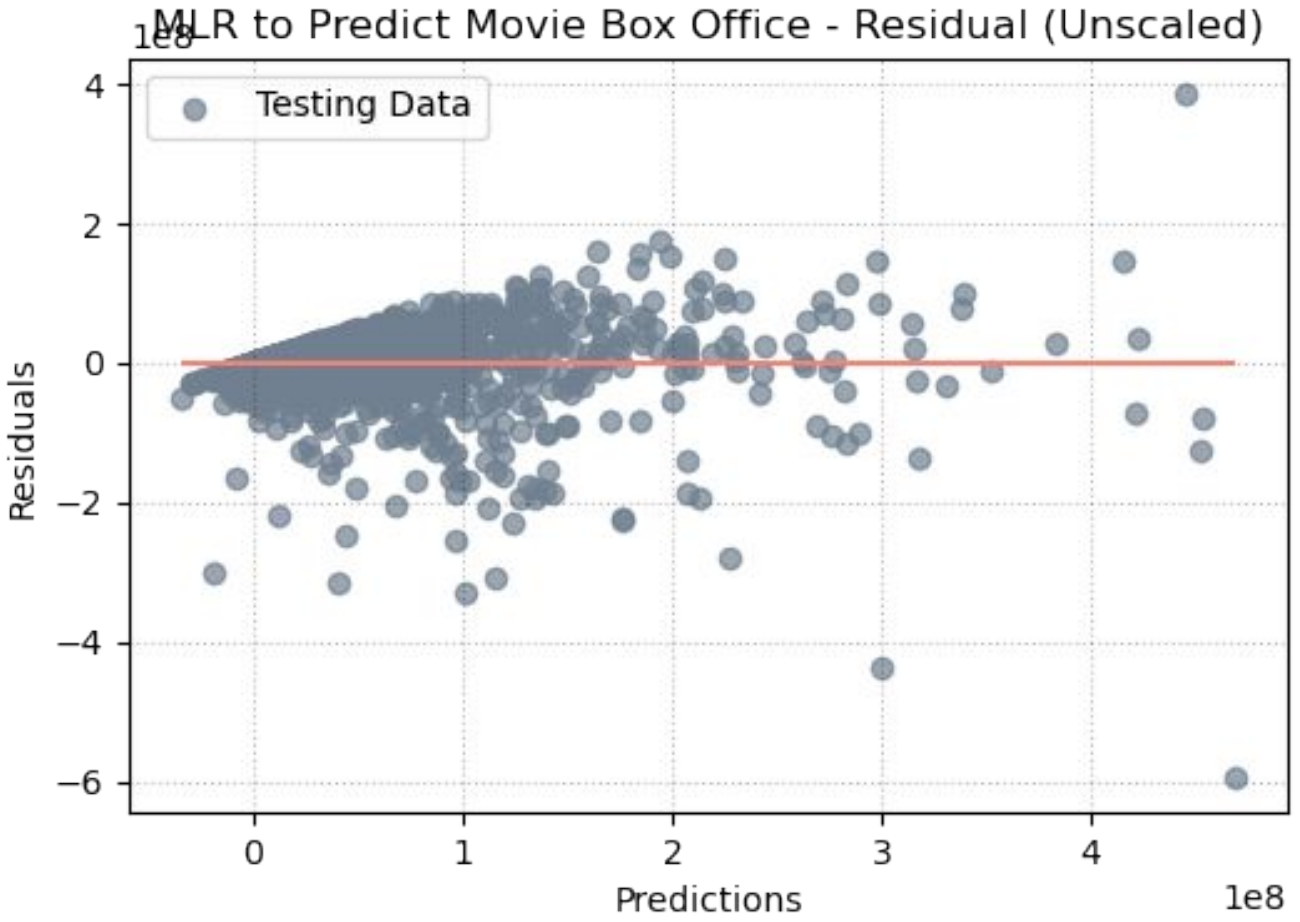
Models & Process

Box Office

- **Multiple Linear Regression**
- Features:
  - **Time**
    - year
    - month (seasonalized)
  - **Genres (24)**
    - One movie can have multiple genres
  - **General**
    - score (IMDB)
    - votes (IMDB)
    - runtime
    - budget (adjusted for inflation)



Residual


MLR to Predict Movie Box Office - Residual (Unscaled)
1e8
Testing Data
Residuals
Predictions
1e8
4
2
0
−2
−4
−6
0
1
2
3
4

Evaluation

```
Model Scores:
------------------
Training Data MSE: 2888532442004727.0
Testing Data MSE: 2888532442004727.0
Training Data Score: 0.6158
Testing Data Score: 0.5895
```

	features	coefficients	feature_type
1	votes_imdb	0.4599	General
2	adjusted_budget	0.2235	General
0	runtime	0.0219	General
8	score_imdb	0.0031	General
9	rating	-0.0068	General

	features	coefficients	feature_type
	film-noir	0.0231	Genre
	family	0.0117	Genre
	animation	0.0098	Genre

Top 3

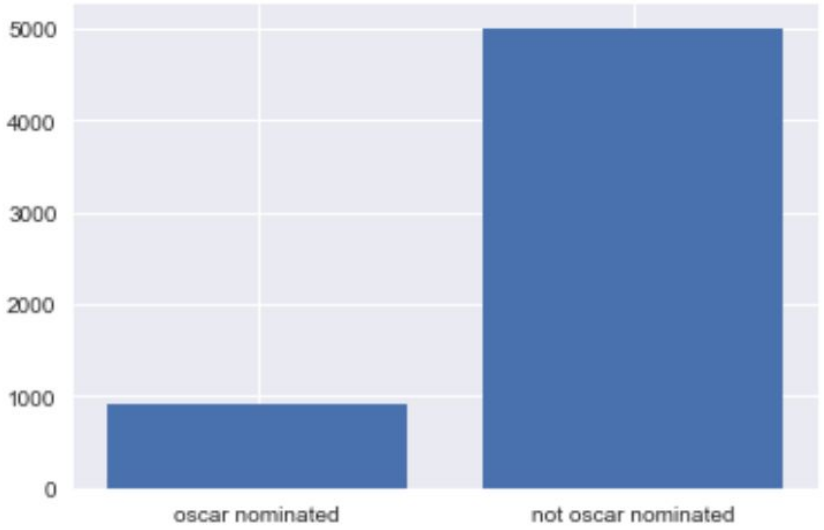
	features	coefficients	feature_type
	history	-0.0109	Genre
	musical	-0.0142	Genre
	documentary	-0.0158	Genre

Bottom 3

	features	coefficients	feature_type
0	year	-0.0008	Year
1	summer	0.0082	Season
2	winter	0.0008	Season
3	spring	-0.0016	Season
4	fall	-0.0074	Season

Oscar Nominations

- **Logistic Regression**
  - Binary classification problem
- **Features**
  - Plot
  - Plot length
  - Genres
- **NLP**
  - HashingTF Vectorizer
  - TF-IDF transformation
  - Pyspark vs sklearn



Oscar Nominations

Unweighted

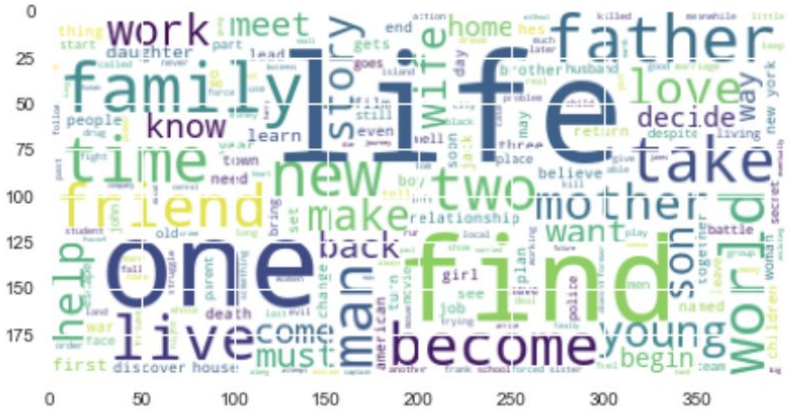
	precision	recall	f1-score	support
0	0.86	0.99	0.92	1006
1	0.46	0.06	0.11	177
accuracy			0.85	1183
macro avg	0.66	0.52	0.51	1183
weighted avg	0.80	0.85	0.80	1183

Weighted

	precision	recall	f1-score	support
0	0.90	0.81	0.86	1006
1	0.32	0.50	0.39	177
accuracy			0.77	1183
macro avg	0.61	0.66	0.62	1183
weighted avg	0.82	0.77	0.79	1183

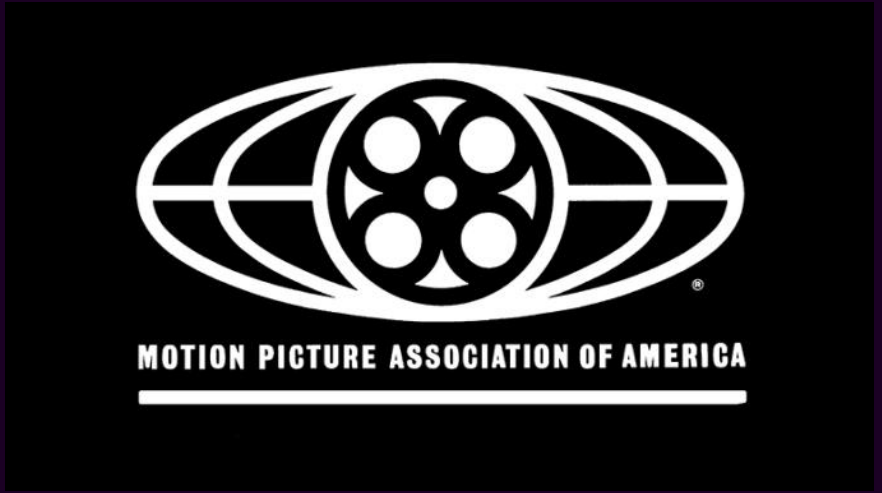
Oscar Nominations

- **Limitations and lessons:**
  - Small set of oscar winners
  - Overall plot data was similar between classes



MPA Rating

- **Features:**
  - Plot
  - Plot length
  - Genres
- **Models**
  - **Multiclass classification problem**
    - Decision Tree & Random Forests
    - Deep Learning
    - Linear SVM (for application)



Random Forests

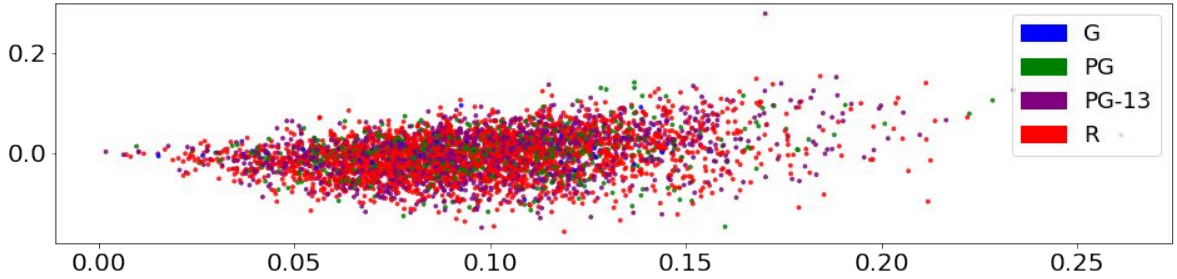
- Feature importances
- Good f1-score
- Number of trees (200 vs. 3000)
- Balance issue
- **Genres play an important part**

	precision	recall	f1-score	support
G	0.00	0.00	0.00	33
PG	0.74	0.49	0.59	219
PG-13	0.60	0.29	0.39	444
R	0.65	0.94	0.77	778
accuracy			0.66	1474
macro avg	0.50	0.43	0.44	1474
weighted avg	0.64	0.66	0.61	1474

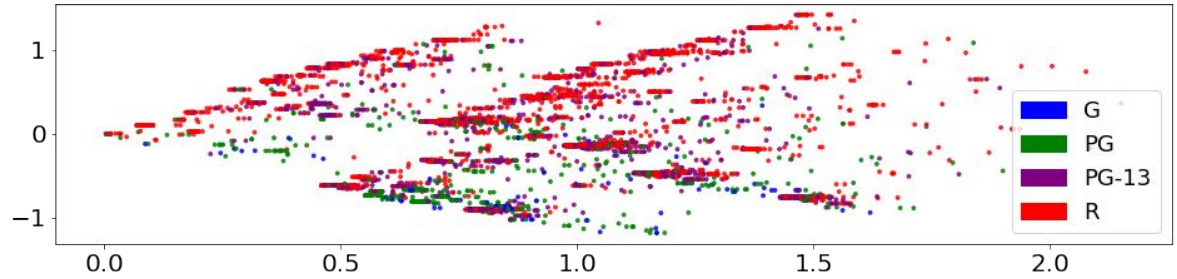
Random Forests model feature importances

```
[(0.032852177149722335, 'family'),
 (0.013346226312506912, 'animation'),
 (0.010207138268718689, 'adventure'),
 (0.008127192141042662, 'thriller'),
 (0.006962489022086924, 'length'),
 (0.006900903206820905, 'crime'),
 (0.005912805986435832, 'comedy'),
 (0.005706837066290208, 'fantasy'),
 (0.005398193847265688, 'musical'),
 (0.005079553221749293, 'horror'),
 (0.004001656943561991, 'drama'),
 (0.003198898198649997, 'action'),
 (0.0022222759262747445, 'romance'),
 (0.0011141939563632242, 'mystery'),
 (0.0010973285156303742, 'sci-fi'),
 (0.0008325564880059716, 'sport'),
 (0.0006532225028020425, 'music'),
 (0.0004819905010564446, 'biography'),
 (0.0004202569022487683, 'short'),
 (0.0003463805653817972, 'history'),
 (0.0003365315296118395 7, 'war'),
 (0.0001951077897751112 2, 'western'),
 (0.0001533602447019888 2, 'documentary'),
 (1.1635011838982446e-05, 'film-noir'),
 (4.26660331001538e-06, 'stemmed'),
 (0.0, 'adult')]
```

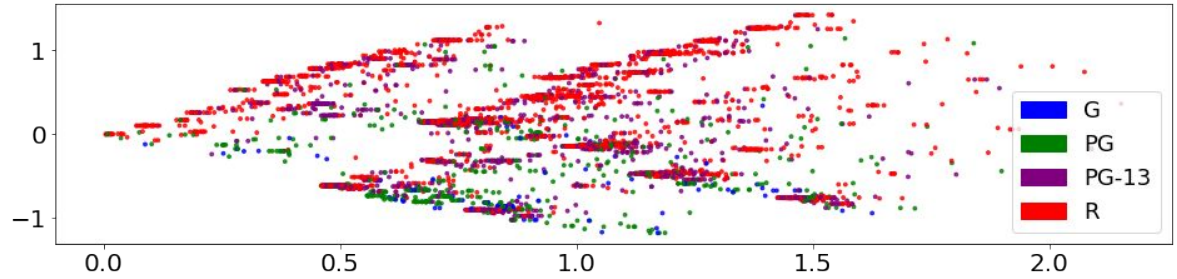
Just Plot Vectors



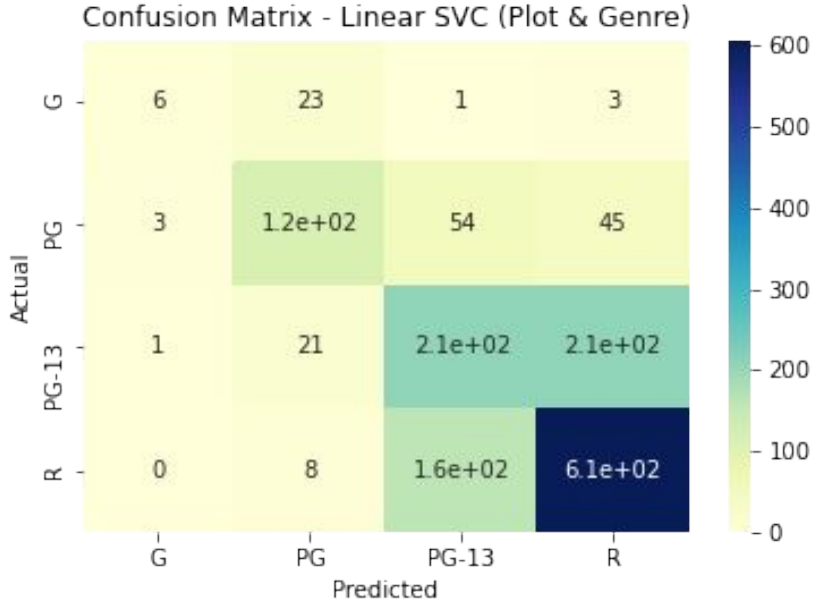
Just Genre Vectors



Combined Vectors



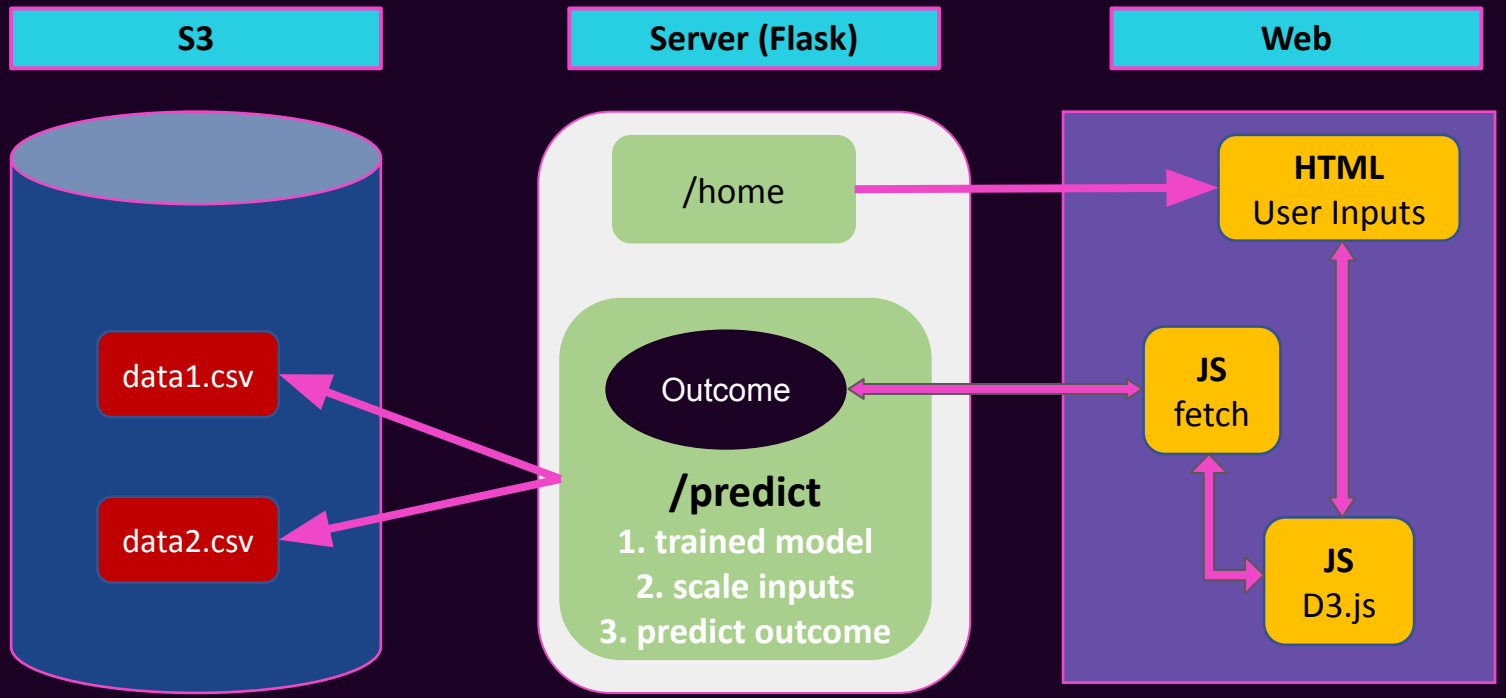
Linear SVM



	precision	recall	f1-score	support
G	0.58	0.21	0.31	33
PG	0.68	0.53	0.60	219
PG-13	0.51	0.48	0.49	444
R	0.70	0.78	0.74	778
accuracy			0.64	1474
macro avg	0.62	0.50	0.53	1474
weighted avg	0.64	0.64	0.63	1474

Application

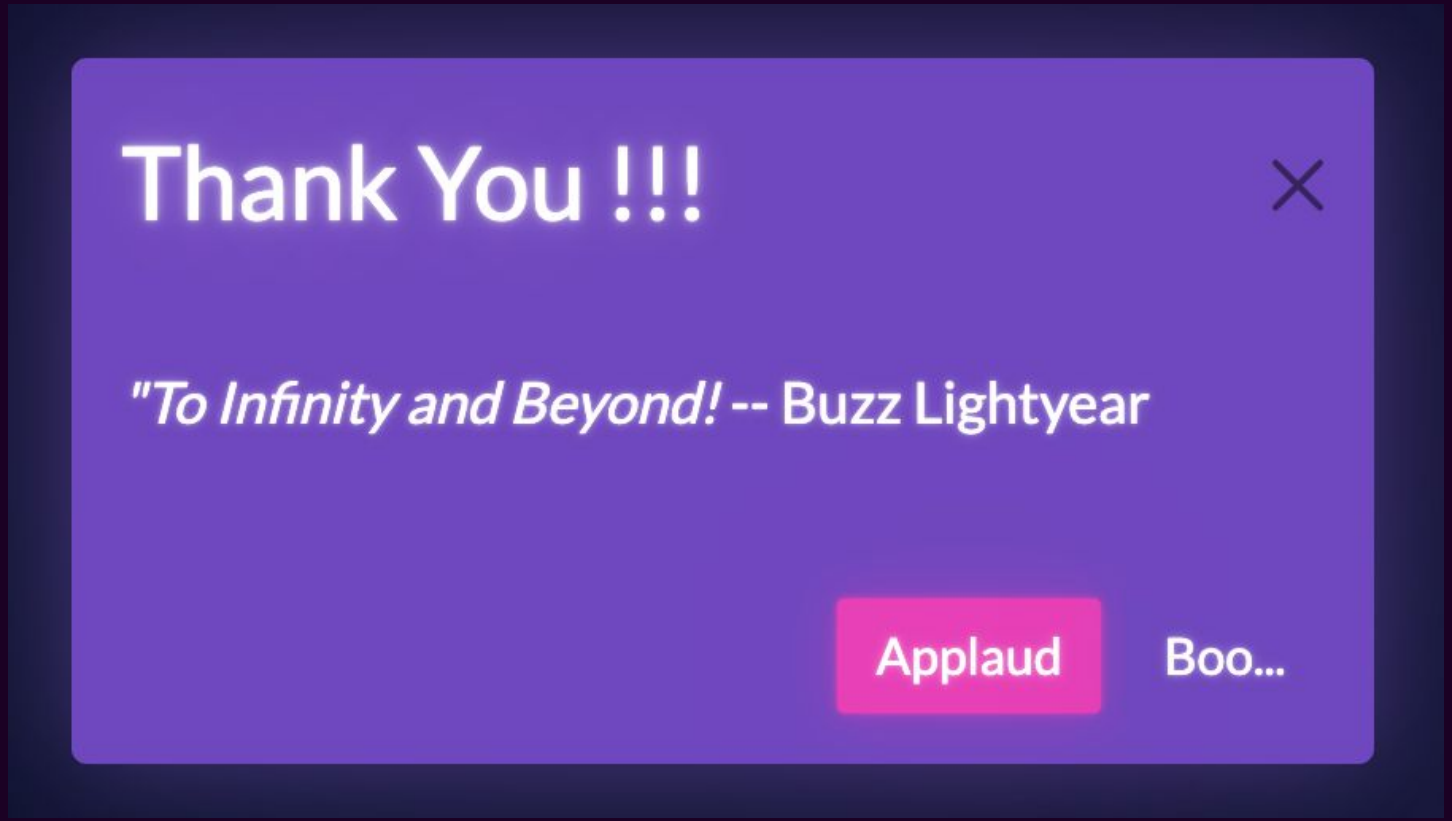
Technical Diagram



Live Demo

Limitations & Next Steps

- Data:
  - Volume
  - Unbalanced
- Features:
  - Box office: factoring in cast and production
  - Oscar model: factoring in cast and production
  - Ratings model: production company data
- Next steps:
  - Training the plot
  - Incorporate “sureness” (certainty of prediction) into the App


Thank You !!!
"To Infinity and Beyond! -- Buzz Lightyear
Applaud
Boo...

Demo

- R
  - detective finds murderer, cuts off own leg.
  - Crime, Thriller, Mystery
- PG-13
  - harry porter goes to magic school, goes through wall.
  - Drama, Fantasy
- PG
  - dragon gets trained and fights bad guys.
  - Action, Adventure, Animation
- G
  - lion loses father, comes back to be the king.
  - Animation, Family, Musical

guidelines

1. You should keep your presentation around 10 minutes and allow for 5 minutes of questions
2. When discussing your model, talk about…
   a. Why the problem you're working on lends itself to ML problem vs a traditional if/then logic
   b. How you made your feature selection (i.e., why the features you focused on were important)
   c. The challenges you had in avoiding bias or balanced/unbalanced data sets, etc…
   d. Why the model you chose makes sense for this problem (i.e., this model works well for categorization vs regression or this model works well when you have more or less data, etc...)
   e. What you'd consider adding in terms of features or alternative models
3. Notice what I avoided above...not a lot of talk on your accuracy. Obviously discuss it, but don't make it the focus.