Forecasting Box Office Success via Optimization: A Data Driven Approach

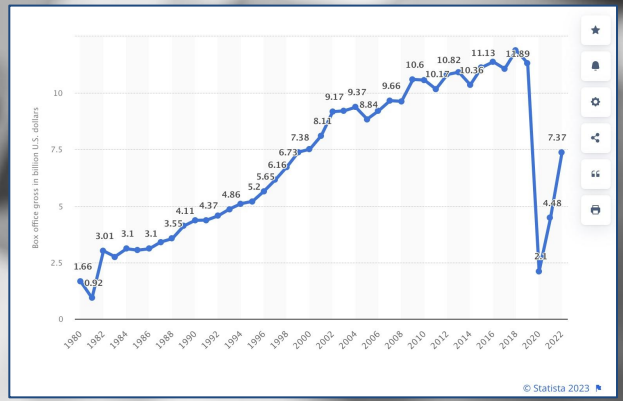
Presented by Team 17

The Film Industry

- Involves substantial financial investments such as production costs and marketing expenses as with many other industries
- Crucial for industry stakeholders to know relevant factors for a movie concept to be successful before going forward with production

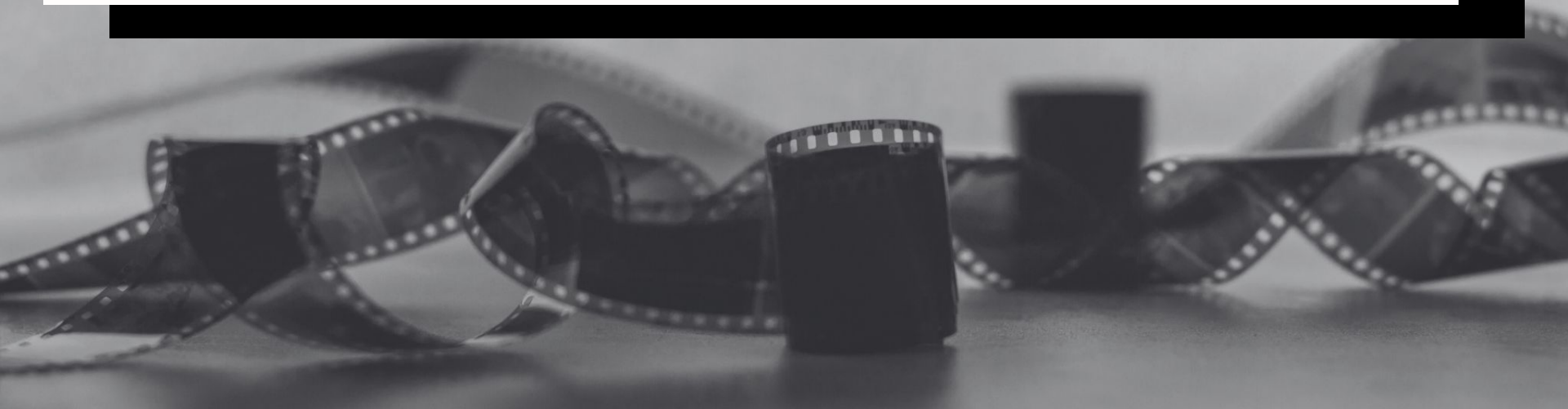
Predicting Success of Movies

- In the domestic market alone, $1 billion+ are spent on movies by corporations such a Amazon and large distributor studios
- Decisions and investments are risky due to volatility in prediction
- Success is affected by a multitude of factors from different sectors



Model Objective

- Optimize cast and directors based on calculated utilities within specific categories of movies and their types
- Predict box office revenue of given upcoming movies using optimal cast and directors



Significant Features

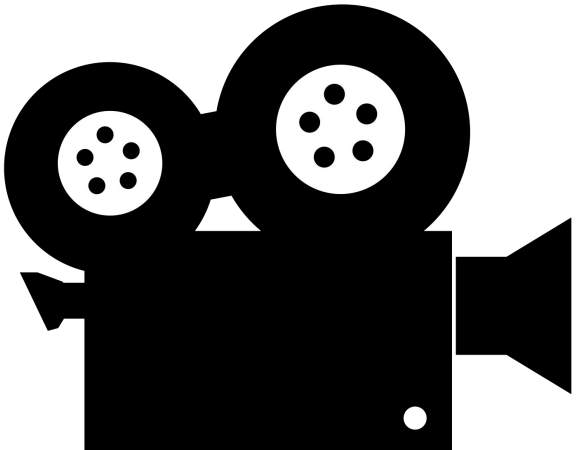
- Genres
- Creative Types
- Production Methods
- Source Materials
- MPAA Ratings
- Theatrical Distributors
- Cast Size
- Budget

- The cast and directors of a movie have a direct influence on the box office, both financially and in terms of overall popularity

- Utility values were calculated for actors and directors based on financial success and number of movies made

Data Collection

THE NUMBERS



All relevant data web scraped from "The Numbers" film database

Final movie dataset contains 2566 samples

Final actor dataset contains 3878 samples

Final director dataset contains 1285 samples

Model Datasets

Actor Utility Values and Salaries

Contains the names, utility values, and the salaries of actors

Director Utility Values and Salaries

Contains the names, utility values, and the salaries of directors

Movie Training Set for Linear Regression

Containins movies partitioned between 1995 to 2020

Movie Testing Set for Linear Regression

Containins movies partitioned between 2021 to 2023

Actor & Director Value

$$v_i = ln(a_i \cdot n_i^{\lambda}),$$

where
v_i = utility value of actor/director *i*
a_i = Average domestic box office
n_i = Number of movies participated in
λ= order of correlation for experience

Derived due to the following:

- a_i represents the bankability for all movies taken part in
- n_i represents experience and cumulative skill
- λ changes the degree to which experience is valued
- $\lambda = 2$ has been set for this model following EDA and general domain knowledge

Actor & Director Salary

$$s_{ij} = B_j \cdot \alpha_i + G_j \cdot \gamma_i,$$

where
s_{ij} = Salary of actor/director *i* for movie *j*
B_j = Production budget for movie *j*
α_i = Budget cut for actor/director *i*
G_j = Domestic gross for movie *j*
γ_j = Gross cut for actor/director *i*

Derived due to the following:

- $B_j \cdot \alpha_i$ represents the earnings received from the production budget of the movie
- $G_j \cdot \gamma_i$ represents the earnings received from the domestic gross box of the movie
- s_{ij} is then grouped by name *i* and averaged to produce the mean salary s_i

Salary Cut Distributions

Status	Value Range	Percent Cut
A-List	**≥ 22.0**	$\alpha_i = 0.10$ $\gamma_i = 0.02$
B-List	**20.5 ≤ and < 22.0**	$\alpha_i = 0.07$ $\gamma_i = 0.0$
C-List	**< 20.5**	$\alpha_i = 0.05$ $\gamma_i = 0.0$

Optimization Model

$$\textit{Maximize} \quad \sum_{i=1}^{n} v_i x_i \quad \forall i = 1,2,\dots,n$$

$$\textit{Subject to} \sum_{i=1}^{n} x_i \leq C \quad \forall i = 1,2,\dots,n$$

$$\sum_{i=1}^{n} s_i x_i \leq 0.3B \quad \forall i = 1,2,\dots,n$$

$$x_i \in \{0,1\} \quad \forall i = 1,2,\dots,n$$

- Objective function maximizes actor/director utility
- Optimal actor count cannot exceed user-defined cast size C
- Decision variable x_i determines whether actor/director is selected

Final Dataset

	Movie	genres	mpaa-ratings	theatrical-distributors	Production Budget	Domestic Gross	Release Date	creative-types	production-methods	sources	Crew Value
0	Toy Story	adventure	g-(us)	walt-disney	30000000.0	192523233.0	November, 1995	kids-fiction	digital-animation	original-screenplay	46.021657
1	Crimson Tide	action	r-(us)	walt-disney	55000000.0	91387195.0	May, 1995	contemporary-fiction	live-action	original-screenplay	46.475760
2	Judge Dredd	action	r-(us)	walt-disney	85000000.0	34687912.0	June, 1995	science-fiction	live-action	based-on-comic-or-graphic-novel	40.164674
3	The Jungle Book	adventure	pg-(us)	walt-disney	175000000.0	364001123.0	April, 2016	fantasy	animation-and-live-action	based-on-fictional-book-or-short-story	28.066436
4	The Lion King	adventure	g-(us)	walt-disney	260000000.0	543638043.0	July, 2019	kids-fiction	animation-and-live-action	remake	35.609616
...	...	...	...	...	...	...	...	...	...	...	...
2503	The New Mutants	horror	pg-13-(us)	sony-pictures-classics	67000000.0	23855569.0	August, 2020	super-hero	live-action	based-on-comic-or-graphic-novel	34.361398
2504	The Gentlemen	action	r-(us)	miramax-dimension	22000000.0	36471796.0	December, 2019	contemporary-fiction	live-action	original-screenplay	40.861833
2505	Honest Thief	thriller-or-suspense	pg-13-(us)	united-artists	30000000.0	14163574.0	October, 2020	contemporary-fiction	live-action	original-screenplay	35.646974
2506	The Last Full Measure	drama	r-(us)	walt-disney	20000000.0	2949212.0	January, 2020	dramatization	live-action	based-on-real-life-events	34.623697
2507	Words on Bathroom Walls	drama	pg-13-(us)	walt-disney	9300000.0	2542518.0	August, 2020	contemporary-fiction	live-action	based-on-fictional-book-or-short-story	35.035282

Linear Regression

```
df_train = df_encoded[df_encoded['Release Date'] < 2021]
df_test = df_encoded[df_encoded['Release Date'] > 2021]
drop_for_x = ['Movie','Domestic Gross','Release Date','Genre','MPAA Rating','Distributor','Creative Type','Production Method','Source']

X_train = df_train.drop(columns = drop_for_x)
y_train = df_train['Domestic Gross']

X_test = df_test.drop(columns = drop_for_x)
y_test = df_test['Domestic Gross']
```

```
linreg = LinearRegression()
linreg.fit(X_train, y_train)
train_score = linreg.score(X_train, y_train)
test_score = linreg.score(X_test,y_train)
betas = pd.Series(linreg.coef_, index=X_train.columns)
betas = betas.append(pd.Series({"Intercept": linreg.intercept_}))
betas.idxmax()
```

Final Results

- Crew Value has a greater influence on the box office than other factors such as production budget
- Relatively high correlation suggests reliability of the model

Production Budget	13.59102
Crew Value	16.20999

Train Score: 0.6509503408368086
Test Score: 0.6290361312172759

Future Insights

- Forecasting the success and profitability of a given conceptual movie
- Deeper exploration on relations between cast and director
- Expand the training data from domestic to international