Predicting Future IMDb Movie Scores

Essie (Yuyang) Liu, Zoe (Hao) Zhang, Nick Ohadi

Rishi Malhotra, Violet (Xuanqi) Liu

Agenda

1. Dataset Introduction
2. Research Questions
3. Methodology
4. Findings & Recs.

IMDb 5000 Movie Dataset

Total Dataset

- 28 variables
- 5043 movies
- 4906 posters
- Across 100 years and 66 countries

Selected Dataset

- 67 variables
- 1494 movies
- 1494 posters
- Across 100 years and only USA

Dataset Introduction

Data Cleaning

- Non-zero budget
- None missing
- Fixed for inflation (past CPI, 2016 as a reference)

- Budget -> Adjusted Budget
- Gross -> Adjusted Gross
- Adjusted to 2016

- Adjusted 19XX $ = 19XX $ * (2016-CPI / 19XX-CPI)

Research Questions

Can we predict a movie's IMDb rating using its quantitative attributes?

Which genres are likely to correlate with great movies?

Will the # of human faces in a movie poster correlate with the movie rating?

Methodology

- Multiple Linear Regression
- Decision Tree
- Random Forest

Disney

Forward-Stepwise Variable Selection

Full List of Variables Considered:

Group 1 - Orange
Group 2 - G1 + Blue
Group 3 - G2 + Green

Variables list:

Duration
Director_facebook_likes
Adj_gross
Cast_total_facebook_likes
Facenumber_in_poster
Adj_budg
Title_year
Action
Adventure
Animation
Comedy
Crime
Family
Fantasy
Thriller
Sci_Fi
Drama
Mystery
Romance
Biography
History
Music
War
Western
Horror
Sport
Documentary
Film_Noir
Approved
G
M
NC_17
Not_Rated
PG
PG_13
Passed
R
Unrated
rating_x

Selected Variables

Model 1:
Duration
Director_facebook_likes
Adj_gross
Cast_total_facebook_likes
facenumber_in_poster

Model 2:
Duration
Director_facebook_likes
Cast_total_facebook_likes
Title_year
Adj_gross
Animation
Drama
Horror
Comedy
Fantasy
History

Model 3:
Duration
Director_facebook_likes
Adj_gross
Cast_total_facebook_likes
Horror
Comedy
Fantasy
Music
G
Not_Rated
Approved
Passed
NC_17

Multiple Regression Models

Model 1:
R-squared: 22.64%
Error: 11.8%
Straightforward to interpret
All variables significant (5 total variables)

Model 2:
R-squared: 29.85%
Error: 10.8%
Relatively straightforward to interpret
Eight significant variables (10 total variables; 2 are marginally significant)

Model 3:
R-squared: 31.27%
Error: 10.9%
Slightly difficult to interpret
Twelve significant variables (15 total variables)

Disney

Methodology

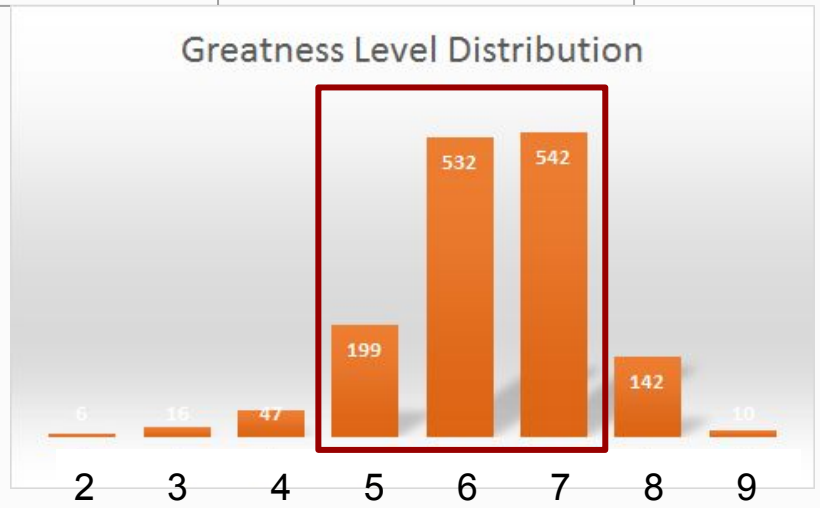
- Multiple Regression
- Decision Tree
- Random Forest

Decision Tree -- Model

Dependent Variables:

Movie Greatness
10 classes (rounded the imdb-score)

Movie	IMDB-Score	Greatness Level
1	2.3	2
2	4.9	5



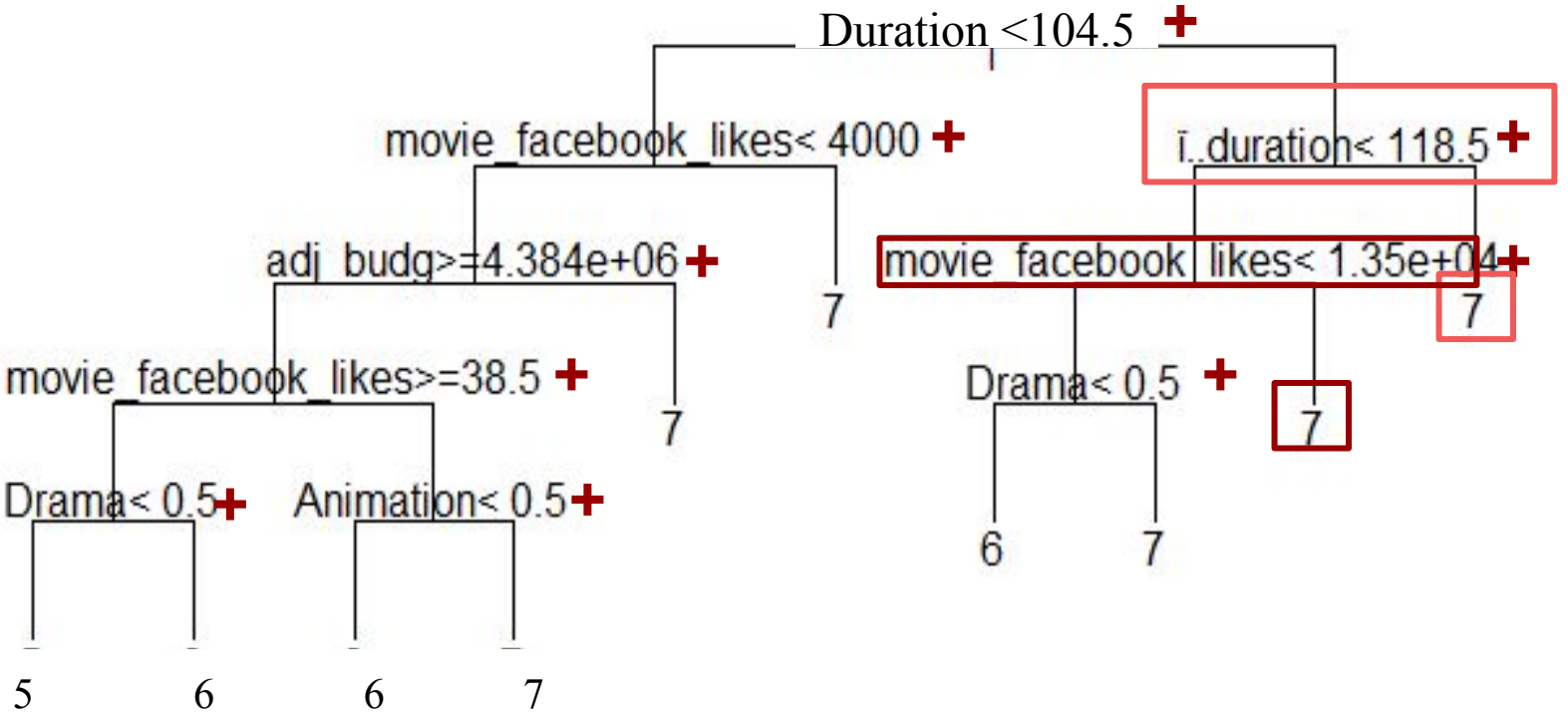
Predictor:

28 variables
(e.g. adjusted budget, genre, # human faces, movie facebook likes…)

Calibration:

- First ⅔ observations

Decision Tree -- outcome



- **Positive Features:** Duration; Movie_facebook_likes; adj_budget; Animation; Drama

Disney

Decision Tree-- Performance

Error

Error =12.25%
(sum(abs(actual-predict)/actual))/#validation

Confusion Matrix for Training

	2	3	4	5	6	7	8	9
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	1	6	12	52	32	4	0	0
6	3	5	16	67	178	79	5	0
7	1	0	4	23	110	239	55	4
8	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0

Confusion Matrix for Testing

	2	3	4	5	6	7	8	9
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	1	1	1	12	13	3	3	0
6	0	1	6	20	72	39	6	0
7	0	3	8	25	127	178	73	6
8	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0

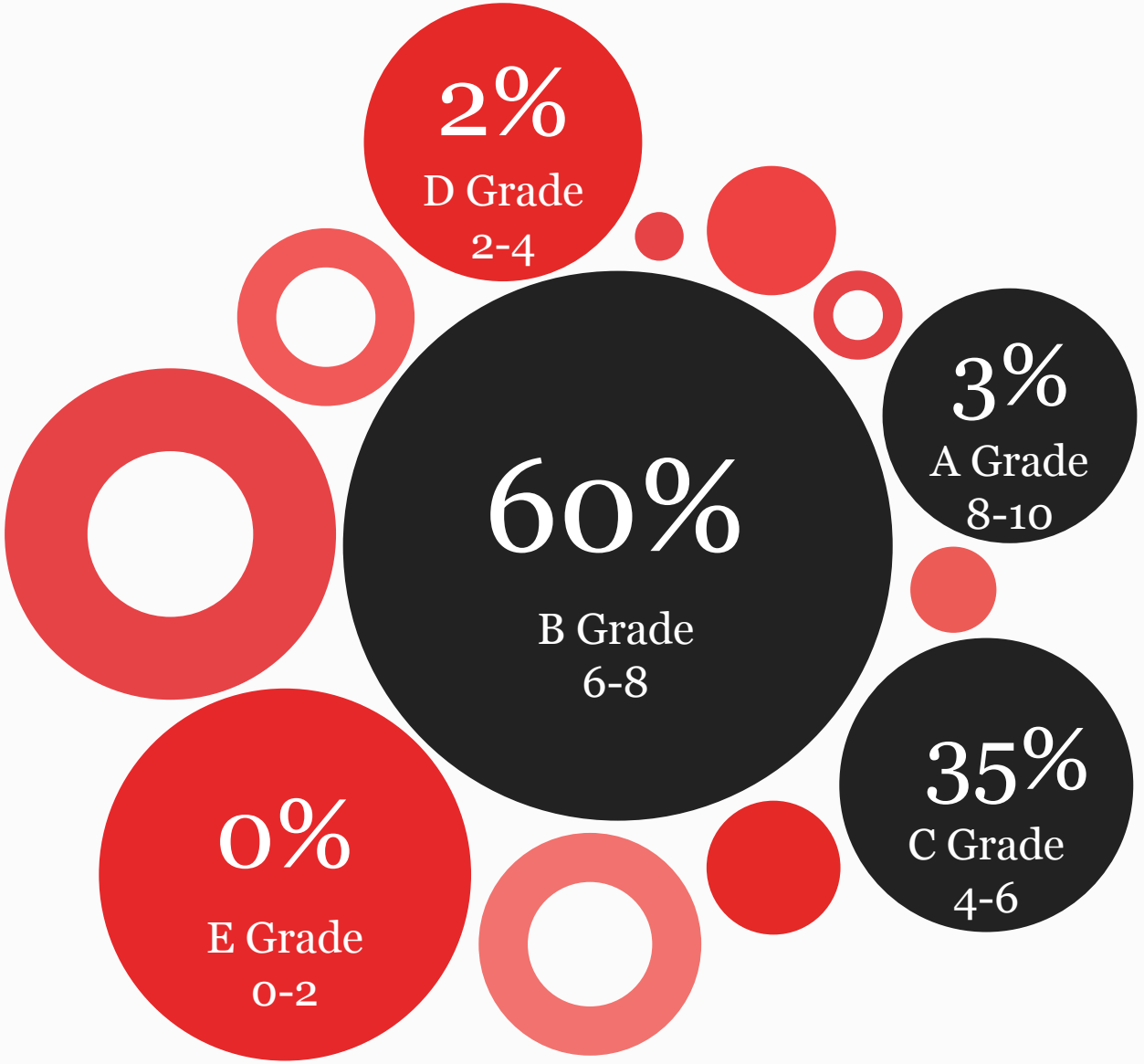
- Most common error range: 1 class
- Tends to be under estimated

Disney

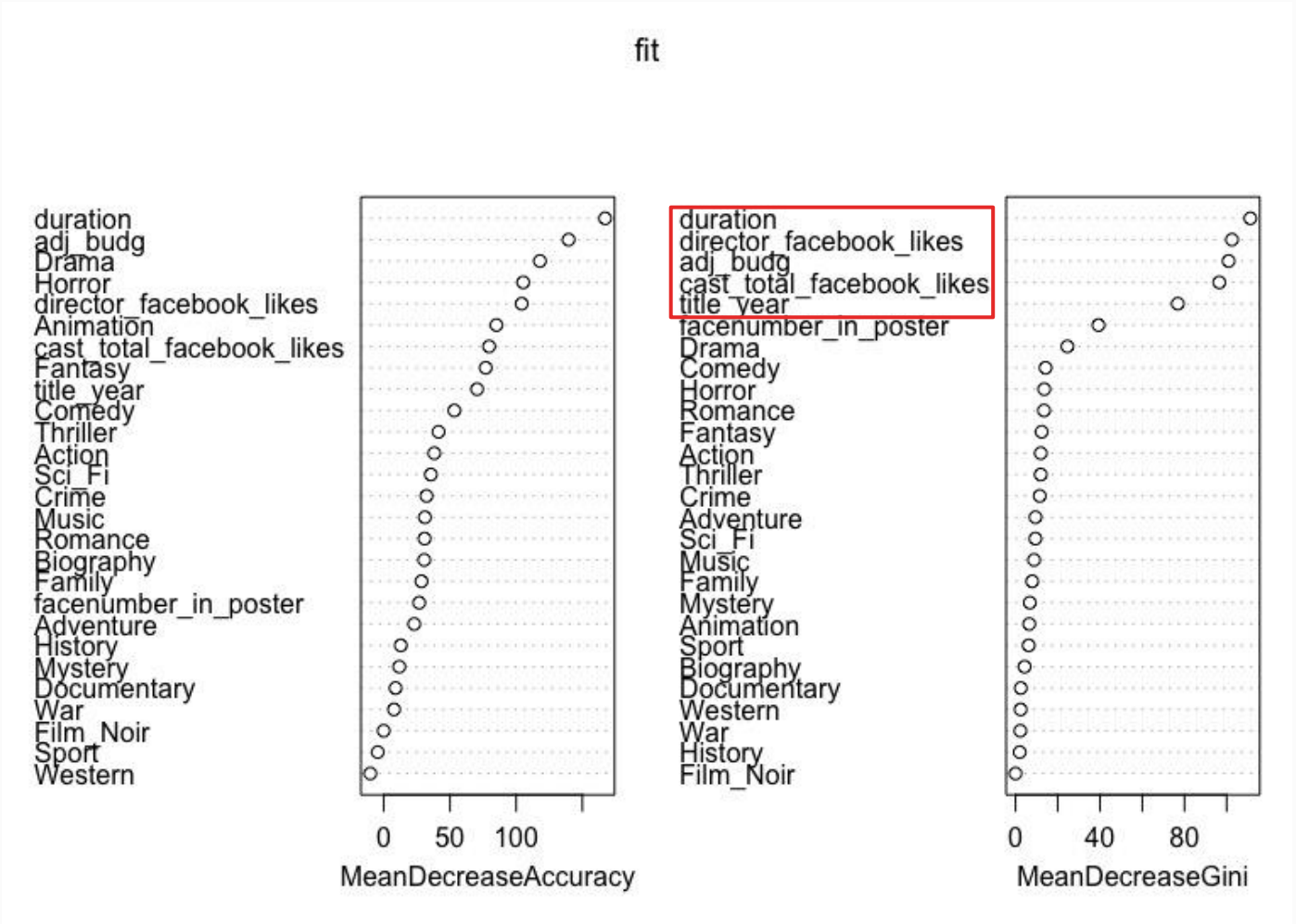
Methodology

- Multiple Regression
- Decision Tree
- Random Forest

IMDb Score Categories



Random Forest



Disney

Random Forest

Findings & Recommendations

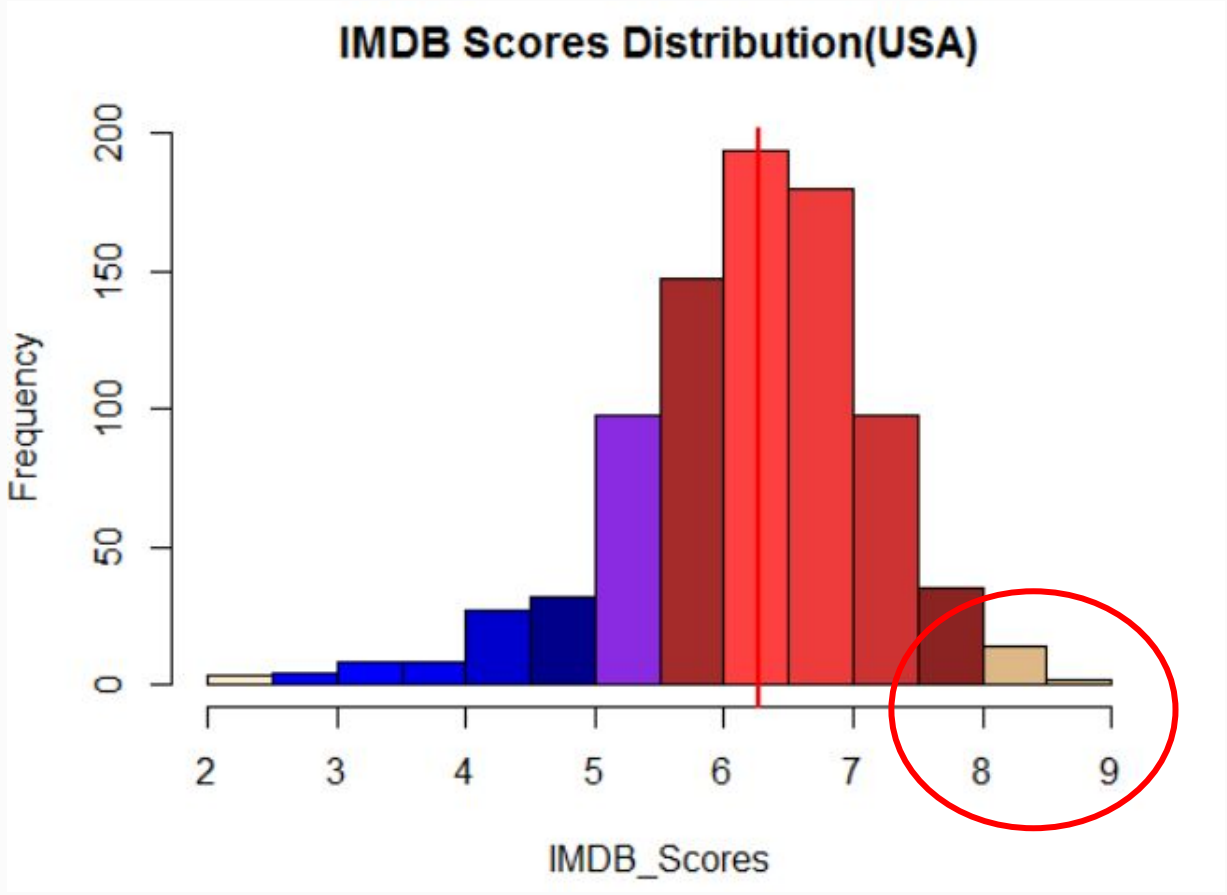
Q1: Can we predict a movie's IMDb rating using its quantitative attributes?

★ Regression and Decision Tree give insight into what traits of a movie increase/decrease ratings

★ **Increase:** duration, director's Facebook likes, gross revenue, cast total Facebook likes, budget

★ **Decrease:** number of faces in poster, G rating

Disney



Distribution of IMDb Scores(USA)

Managerial Perspective: What aspects of a motion picture can the Studio control to achieve high ratings?

- ★ The studio can control the following attributes:
  - Duration, director, cast, content rating, budget

- ★ **Recommendation:** Increase film length, pick directors with high # of Facebook likes, pick a cast with a high total # of Facebook likes, non-G rating, higher-budget films

Q2: Which genres are likely to correlate with great movies?

Genres selected by the Forward-Stepwise Regression phase, and are statistically significant (taken from model 2):

- Animation, Drama, Horror, Comedy (only Fantasy is statistically insignificant)
- Model 3 would include Fantasy (marginal significance), and Music (not significant)

Increase Ratings: Animation, Drama

Decrease Ratings: Horror, Comedy

**Interpretation of genre effects from Model 3 reflect Model 2*

**Decision Tree also reflects the effects of Animation and Drama*

Managerial Perspective: What genres should the Studio focus on that earn the best ratings in the future?

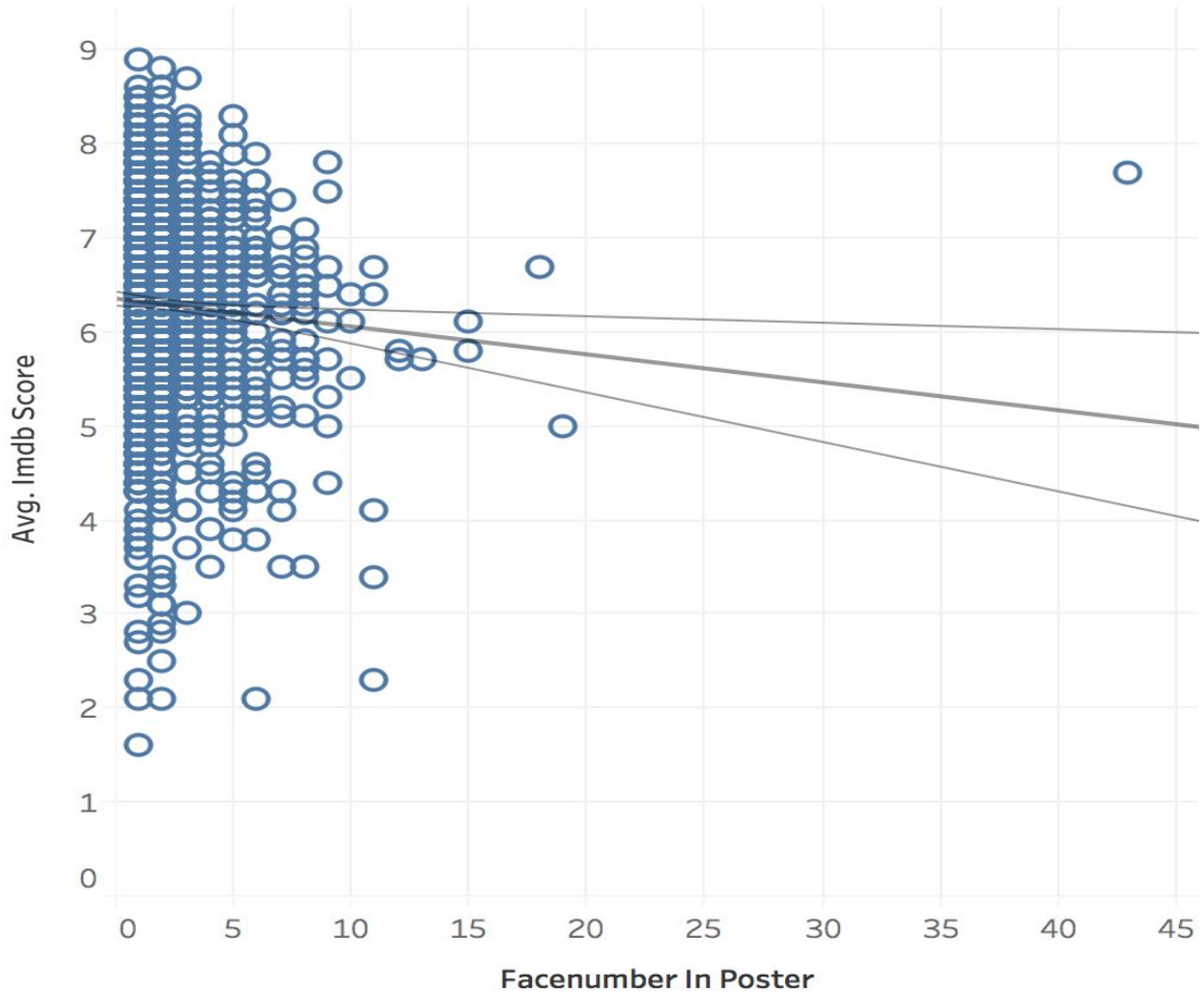
★ Animation & Drama tend to be most attractive in terms of positive ratings. Horror & Comedy should be avoided

★ Focus on producing motion pictures in these genres
  - Cross-pollinating and promoting movies with this genre are likely to boost ratings as well, due to association

Q3: Will the # of human faces in a movie poster correlate with movie rating?

- ★ Only Model 1 includes # of faces in advertising posters, conclusions based on that
- ★ With a coefficient of -.0240, we would conclude that fewer faces in a movie poster lead to higher IMDb ratings
- ★ Borderline significance : p-value is .0502
- ★ A “sweetspot” # of faces likely exists, but is not reflected by the model’s coefficient (which implies 0 faces is best)

IMDB Score Vs. Facenumber in Poster



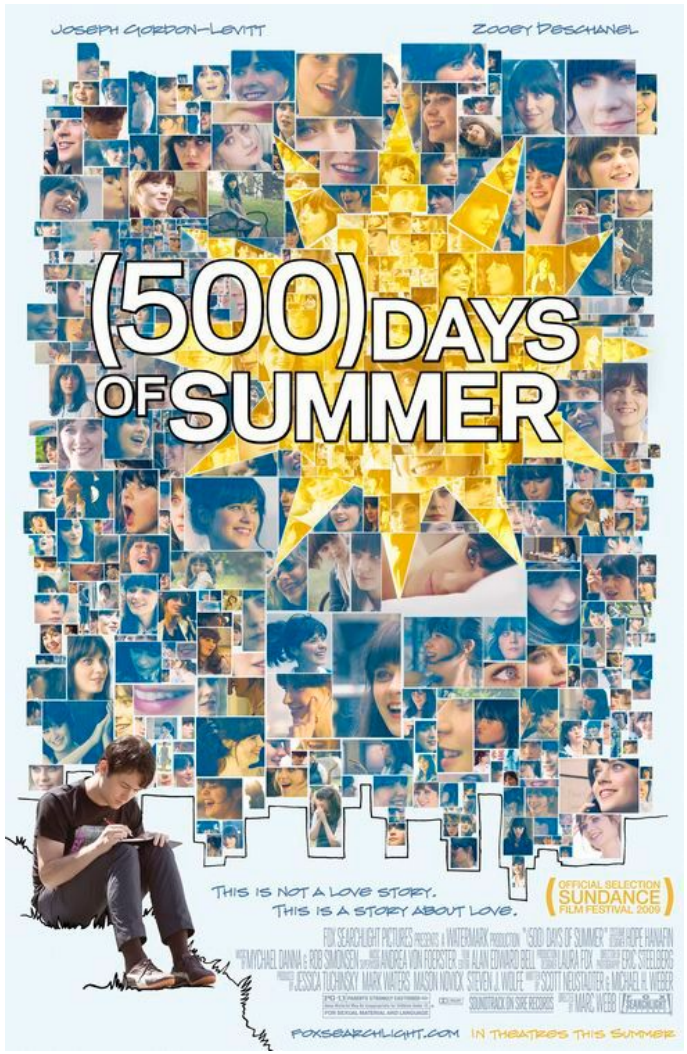
Sum of Facenumber In Poster vs. average of Imdb Score. Details are shown for Movie Title.

Managerial Perspective: How can we direct our creative advertising team to promote a movie?

★ Since there is a direct correlation, we can have internal creative teams and vendors create more effective marketing materials (i.e. posters or billboard creation)
  - Design posters with a theoretical sweet spot in mind
  - Create additional marketing materials (handouts, postcards, etc.) to experiment with # actors/actresses faces

★ We advise more research to be done on the design of posters, as # of faces doesn't capture other elements such as color, shapes, face size, composition, general aesthetic appeal, etc.

Two Movies with 7.7 Rating




BREAKING
NEWS
DataBusters

Thank You