CS 118 Final Project Presentation

Michelle Wang

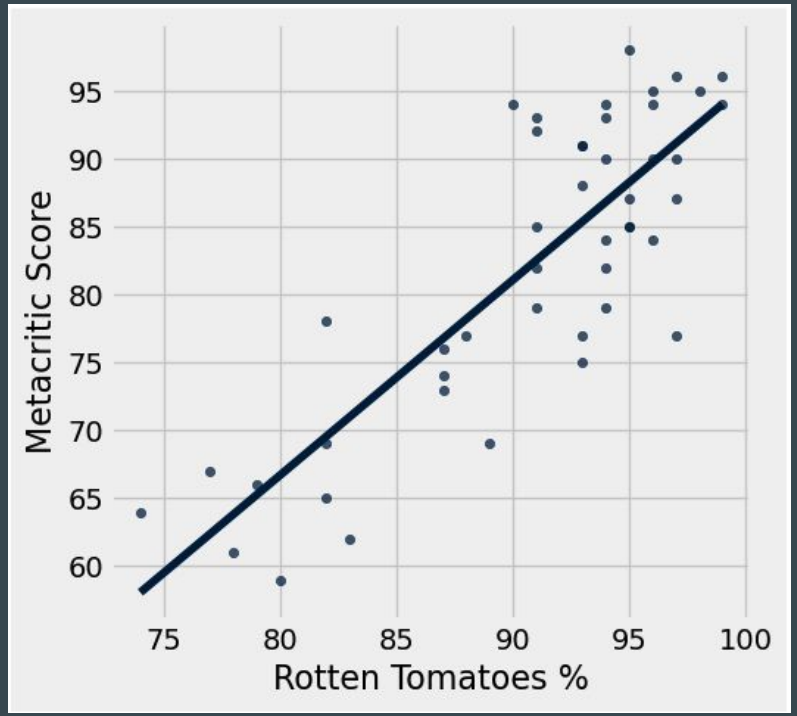
Part 1: My Dataset

- Found on Kaggle, an online user-based database of datasets, by Shayan Zulfiqar
- Data was gathered from the Top 250 Movies of IMDb as of January 2025
- 100 observations, 7 categorical variables, 6 numerical variables
- Added variable - did the movie win an Oscar or not (True/False)
- Has information on rank, ratings, Oscars, etc.

Rank	Title	Year	IMDb Rating	Metacritic Score	Box Office ($M)	Academy Award?
1	The Shawshank Redemption	1994	9.3	82	58	False
2	The Godfather	1972	9.2	100	246.1	True
3	The Dark Knight	2008	9	84	1004.9	True
4	The Godfather: Part II	1974	9	90	48.5	True
5	12 Angry Men	1957	9	96	1	False

... (95 rows omitted)

Part 2: Numerical Variables



<u>Description of My Variables</u>
Rotten Tomatoes (x): Aggregated ratings of critics and general audiences, creating a score from 0 to 100.
Metacritic Score (y): Aggregated scores of critics from 0 to 100

Standard Units

- Removes inconsistencies and errors, e.g. scale for better comparability and replication

Correlation Coefficient (r)

- -1 to 1: Measures the strength in predictability

<u>Goal</u>: Investigate the relationship between these two chosen variables and decide whether or not we can use one score to predict the other.

Part 2 Cont: Finding r

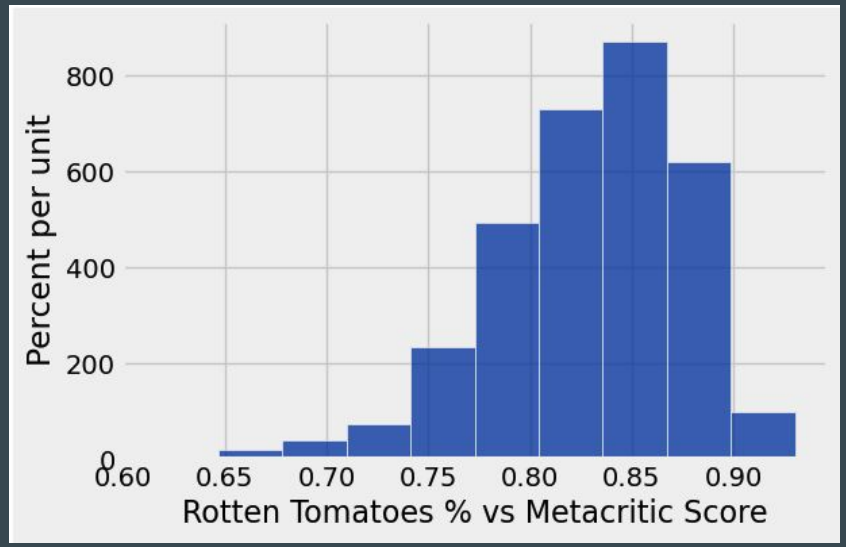
Hypothesis Testing

- We use hypothesis testing to reach our goal.
- Deciding whether or not Rotten Tomatoes and Metacritic Score are correlated

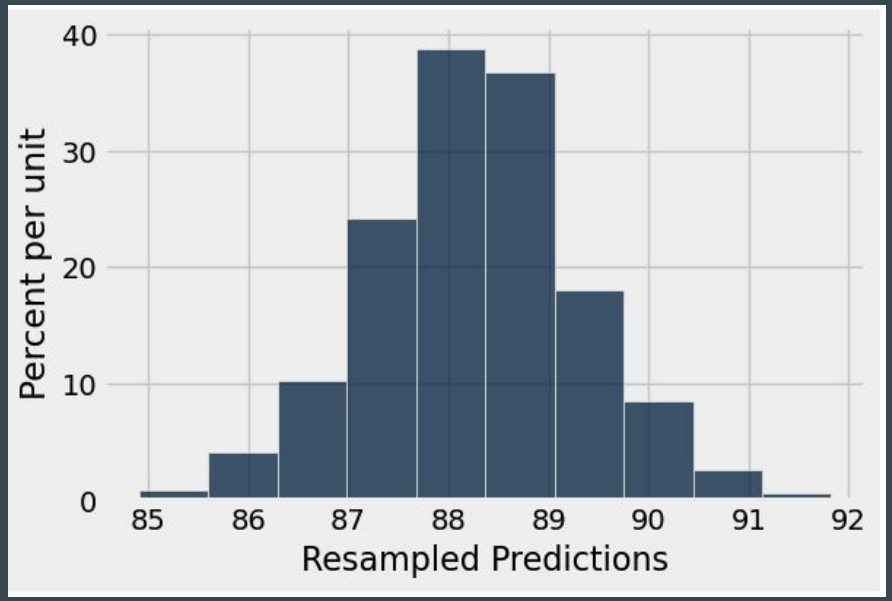
Confidence Intervals

- Range of numbers estimating the true population correlation
- 95% of the time, parameter will fall into this interval
- Use it to evaluate hypothesis

<u>95% Confidence Interval for Correlation</u>:
[0.7277712883155425, 0.9043529176298908]



Part 2 Cont: Predictions and Biases



X-Input Value Chosen for Prediction: 95

99% Confidence Interval for Y-Value:
[85.33083432304038,
91.02348903547869]

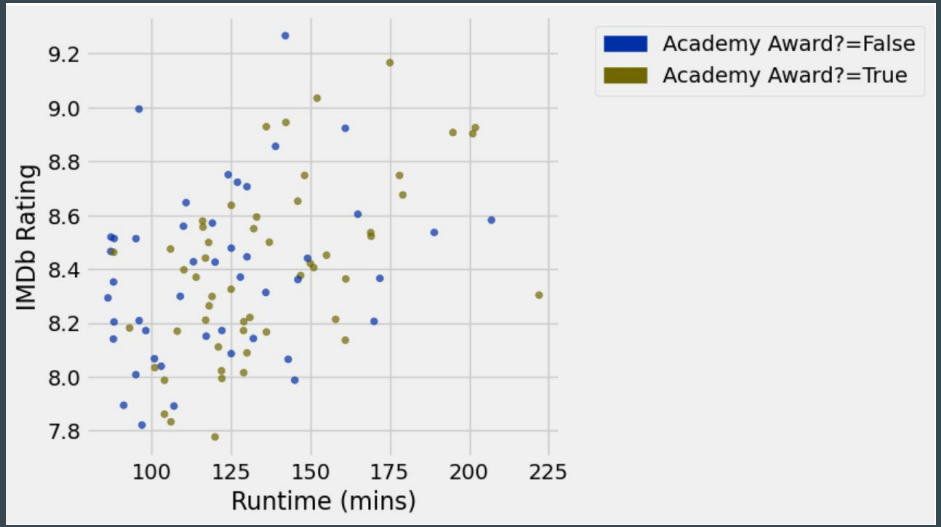
Biases and shortcomings of this data

- Doesn't tell us much about the population of viewers, because scores are aggregated
- We'd expect a positively rated movie to have positive ratings all around

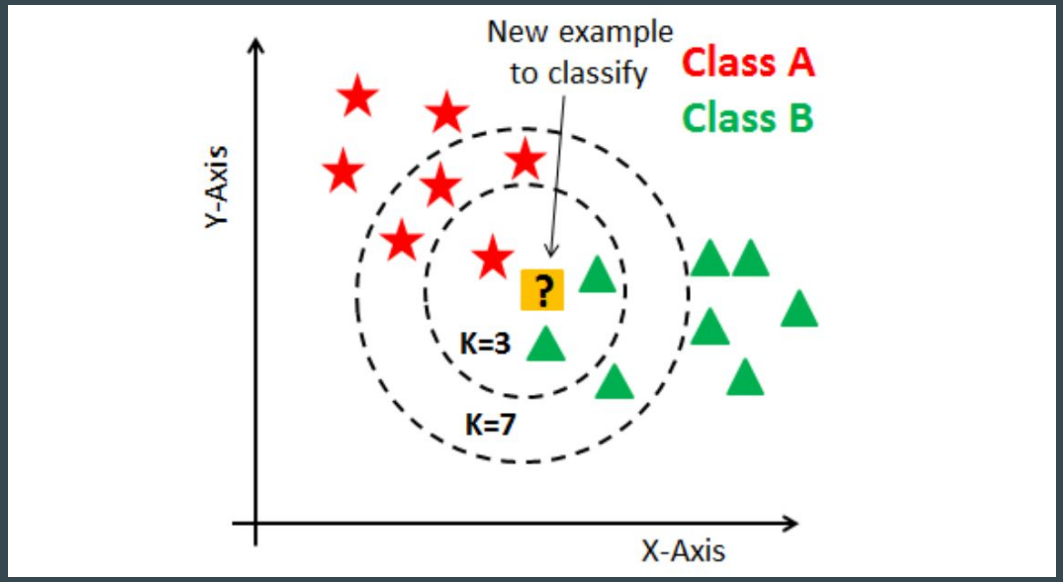
Part 3: Classifiers and Features

Can I Predict Whether or Not a New IMDb Movie Has Won an Academy Award or Not?

- Use classifiers to (attempt to) answer this question
- We use specific numerical variables to ascertain the features of award-winning movies
- Simulate the process of classifying by splitting up our dataset between testing (25) and training (75) purposes



Part 3 Cont: Process of Classification



K-Nearest Numbers (KNN)

- Choose a row from the testing set
- Select the 3 nearest training data points in terms of features (i.e. Rotten Tomatoes, box office, runtime)
- Find the majority classification (whether or not these movies won Academy Awards)
- Classify the row
- 60% Accuracy

Part 4: Improving the Classifier

- Removed one feature (Metacritic Score) that was interfering with results
- Changed from 3 to 5 k-value
- Slight improvement - 64%
- Overall classification with my data set was a failure, too many interfering biases and not enough data points

Conclusion

- Linear regression model fit for my variables of interest
- I could not effectively classify data points on whether or not the movie obtained an Academy Award or not
- My dataset was too small, had too many unusable values, wasn't randomly selected
- Oscar awards and IMDb movie rankings are inherently biased
- Going forward, I will need to be more careful in choosing reliable, worthwhile datasets in investigation.