CAPSTONE PROJECT

PREDICTING OF IMDB RATINGS USING SUPERVISED MACHINE LEARNING

HALAK DESAI

APRIL 2021

WHY THIS PROJECT?

- I like watching movies as most people do. I get recommendations on my subscription service all the time.

- I observed that whenever I get a recommendation, I would immediately search for its ratings meaning how other people have liked the movie.

- This got me thinking about what drives these ratings and why people like certain movie more than others and to take it a step further, **What if I can Predict a Rating!**

- If I can predict a rating, then model could be helpful to investors, casting crew, producer, writers, etc. when creating a movie!

FEATURE ENGINEERING

Considered Movies from year 2010-2020

Popular Actor (acted in more than 15 movies – top 15 by weighted average rating)

Popular Director (directed more than 5 movies – top 15 by weighted average rating)

Popular Genre (more than 4500 movies in a genre– top 4)

Duration divided into 3 buckets (0 to 60 mins, more than 60 mins and less than 90 mins, and more than 90 mins)

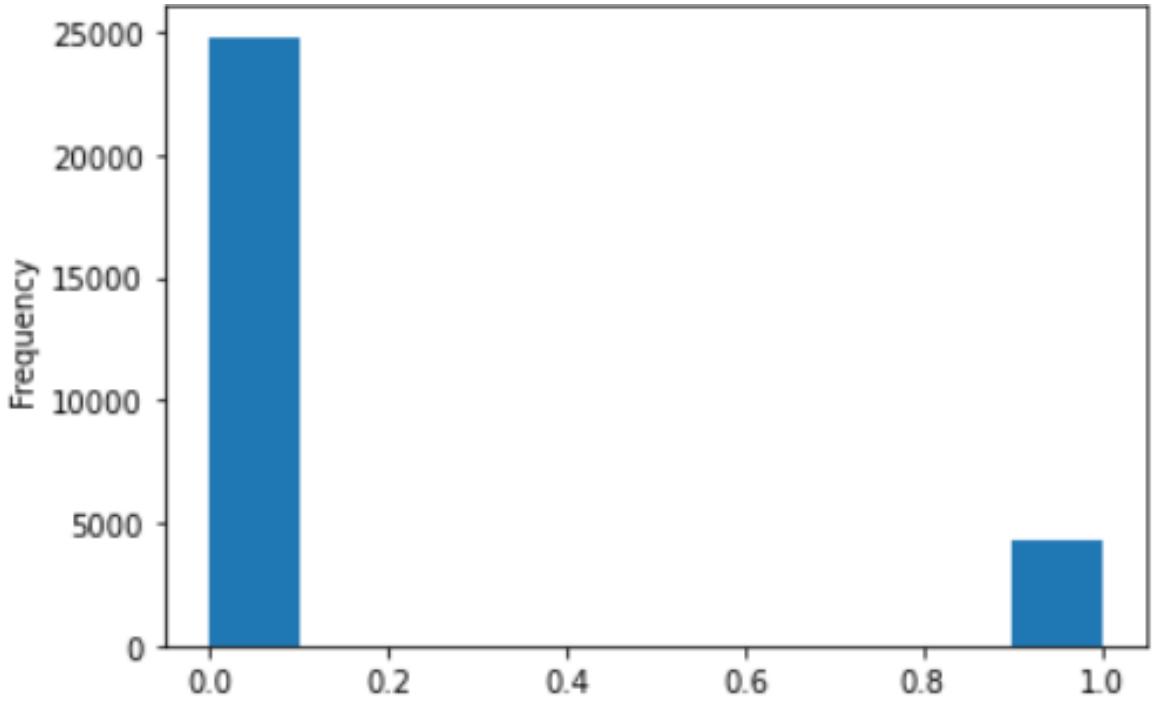
TARGET VALUE AND CLASS IMBALANCE PROBLEM

Class 0

Ratings Less than 7 = 85%

Class 1

Ratings Equal to or More than 7 = 15%



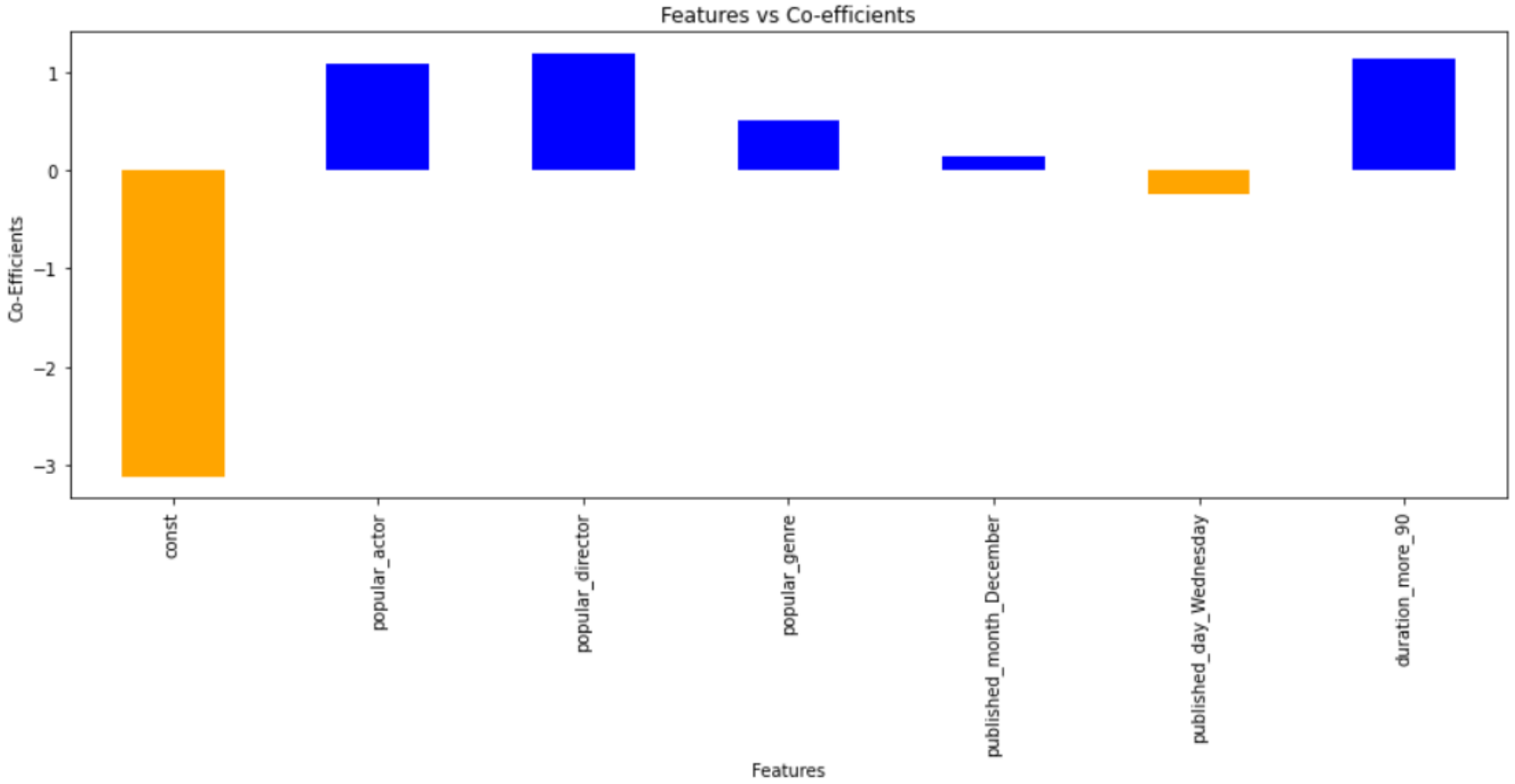
Class 0 and 1

Positive Indicators

- Casting a Popular Actor
- Popular Director
- Popular Genre
- Releasing movie in month of December
- Duration of movie more than 90 mins

Negative Indicators

- Releasing a movie on Wednesday



STATISTICAL ANALYSIS WITH LOGISTIC REGRESSION

statsmodels

Features	Imbalance Train Set	Imbalance Test Set	Oversampled – Train Set	Imbalance Test Set	Undersampled – Train Set	Imbalance Test Set
Score	85%	86%	61%	46%	61%	46%
Best Model	Decision Tree		Logistic Regression		Logistic Regression	
Precision – Class 0	Rating Less Than 7 ↑		Rating Less Than 7 – ↑		Rating Less Than 7 – ↑	
Recall- Class 0	Rating Less Than 7 – ↑		Rating Less Than 7 – ↓		Rating Less Than 7 – ↓	
Precision – Class 1	Rating More Than 7 – ↑		Rating More Than 7 – ↓		Rating More Than 7 – ↓	
Recall – Class 1	Rating More Than 7 – ↓		Rating More Than 7 – ↑		Rating More Than 7 – ↑	
Decision Tree, Logistics Regression and SVM Used in Pipeline and GridSearchCV						

MODEL RESULTS

NEXT STEPS...

OPTIMIZE MODEL TO IMPROVE PRECISION SCORE TO MAKE MY MODEL MORE RELIABLE.

- TRY ENSEMBLE MODELS

- FIND DATASET FOR REVENUE AND BUDGET FOR MOVIES

- FIND IF RATING OR BUDGET AFFECTS REVENUE AND OR MODEL

CONTACT INFORMATION:

Email: halakmajmudar@gmail.com

LinkedIn: linkedin.com/in/HalakDesai

Github: HD208