CAPSTONE PROJECT – PREDICTING IMDb RATINGS USING SUPERVIZED MACHINE LEARNING
HALAK DESAI
BRAINSTATION - APRIL 2021

WHY THIS PROJECT?

I have always been curious to know `"WHY"` a decision was made and `"WHAT"` information supported it.

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!

WHERE IS MY DATA FROM?

- My dataset is from IMDb sourced from Kaggle.com , and have multitude features. I had 4 different csv files to work.
- After cleaning data and my final shape of dataset is about 85k rows and about 18 features.
- Each row represents information about a unique movie. It has information for the movies from 1911 to 2020(Not full year).
- I also observed that the actors, directors, and genres have list of entries per row.
- Another important information is that my dataset has more categorical values than numerical values.

FEATURE ENGINEERING

- As per insights gained during my initial analysis, I applied feature engineering to prepare my data for modeling by creating some columns to simplify large dataset.

I had about 194K unique actors, 41K unique directors and about 27 unique genres.

Hence, I decided to create 3 new columns:

Popular Actor – Movies with more than 5000 votes, Actors acted in more than 15 movies, and selected top 15 actors based on calculated mean weighted average of all movies that actors have acted.

Popular Director – Movies with more than 5000 votes, Directors directed in more than 5 movies, and selected top 15 actors based on calculated mean weighted average of all movies that actors have directed.

Popular Genre – Movies with more than 5000 votes, Genres selected that have more than 1000 movies.

MODELING

Results of my Logistic Regression by using Stats Model:

- if a popular actor casted in the movie the odds of getting a good rating are increased by a factor of 2.947.
- if a popular director is directing the movie the odds of getting a good rating are increased by a factor of 3.297.
- if a movie is made in a popular genre the odds of getting a good rating are increased by a factor of 1.672.
- if a movie is released in December the odds of getting a good rating are increased by a factor of 1.147.
- if a movie is made for duration longer than 90 mins the odds of getting a good rating are increased by a factor of 3.109.
- if a movie is released on Wednesday the odds of getting a good rating are decreased by a factor of 1.265.

CAPSTONE PROJECT – PREDICTING IMDb RATINGS USING SUPERVIZED MACHINE LEARNING

HALAK DESAI

BRAINSTATION - APRIL 2021

I decided to run Logistic Regression, SVM and Decision Tree for easy interpretability of models for prediction from plethora of Machine Learning Tools.

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	Logistic Regression		Decision Tree		Logistic Regression	
Precision – Class 0	Rating Less Than 7- 85%		Rating Less Than 7 – 93%		Rating Less Than 7 – 93%	
Recall- Class 0	Rating Less Than 7 – 100%		Rating Less Than 7 – 40%		Rating Less Than 7 – 40%	
Precision – Class 1	Rating More Than 7 – 67%		Rating More Than 7 – 19%		Rating More Than 7 – 19%	
Recall – Class 1	Rating More Than 7 – 0%		Rating More Than 7 – 83%		Rating More Than 7 – 83%	
Decision Tree, Logistics Regression and SVM Used in Pipeline and GridSearchCV						

Imbalance Train Set

For Class 1 for Good Ratings (ratings greater than 7)

Recall score of 0% means that the model is not classifying the class efficiently. It is predicting more False Negatives than True Positives.
Precision score of 67 % means that model is predicting marginally more than equal True Positives than False Positives.

So, in my case, model is misclassifying good ratings (ratings greater than 7) as poor ratings (ratings less than 7), but when it does detect a good rating, it is moderately likely that it is true good rating.

Balanced Train Set UnderSampling using RandomUnderSampling method and OverSampling Methods using SMOTE.
Both techniques gave me similar results and accuracies.

For Class 1 for Good Ratings (ratings greater than 7)

Recall score of 83% means that the model is classifying the class very efficiently. It is predicting less False Negatives than True Positives.
Precision score of 19 % This means that model is predicting very less True Positives and predicting more False Positives.

So, in my case, it does a good job classifying ratings into good and bad very well, but when it does detect we are not sure if it is indeed a good rating, it is highly likely that it is truly a poor rating.

FINAL THOUGHTS

My initial thoughts and expectations were that I would find some trends and patterns on how ratings are affected by actors casted or director' s name and popularity.

After my analysis I did find a relationship between these attributes.

It was interesting to find that most of the ratings fell between 5 and 7 on a scale of 0 to 10. Hence, I had to deal with class imbalance when predicting good vs bad ratings.

My model has proven to identify Class very well by 83% but precision is poor just about 19% increasing my False Positives when my class was balanced.

Hence, I would rephrase my question to identify good ratings and how to avoid a poor rating for a movie!
I would pay more focus on precision score, as I want my model to focus on increasing my True Positives, i.e., True Good Ratings. I would be comfortable, If I have to trade-off my Recall Score, because it would make my model conservative. The model would predict more False Negatives (True Good Ratings misclassified as Poor Ratings) then it is currently predicting, but by improving Precision Score, cost of investing in poorly rated movie would be low. Hence it will be a risk averse model.

INDUSTRY STANDARD FOR SIMILAR PROBLEM

Predicting ratings, votes, reviews are prediction which have human emotions and preferences involved. It does not have any measurements which are explainable by a scientific technique or any equipment.

I have come across accuracies as high as 87% and as low as 30%. Though I want to make a note that higher accuracies were obtained when analyzing top 1000 to top 5000 movies (different features were analyzed mostly used reviews as a feature).

I want to make a note that my model analyzes about 29000 movies. There is more noise in my dataset.

FUTURE DIRECTION

Optimize model to improve precision score to make my model more reliable and trustworthy. I would like to try some methods below:

- Try Ensemble Models
- Find dataset for Revenue and budget for movies
- Find if rating or budget affects revenue and or model.

Finally, I would also like to try to make a model which would predict ratings for each genre to make it more focused and narrow the scope to get a better result.