Machine Learning I Final Project

Jinshun Su
Maeshal Hijazi
Bruce Wang

Introduction

- We, group 1, chose movies' dataset to predict a movie's **rating** based on its different features.
- In here we present our output based on different trials we did.
- Some of the movie's features we used for the ML model are the movie's Runtime in minutes, number of votes, metascore, and revenue in MILLIONS.

Dataset

- Our movies' dataset has 1000 columns and 12 rows
- It has movies from from the year 2006 to 2016
- It includes other features not mentioned before as Title, Genre, Description, Director, actors, etc.

ML Algorithm

- We used the MLPRegressor algorithm from the sklearn package.
- We used 25 percent of the data for testing and the remaining for training.
- For judging our model, we used Mean Squared Error (MSE) from the sklearn package.
- We did scatter plots for each feature with the output to see visually which feature affect the output more.
- All features had a p-value of less than 0.05 when z-tested with the output.

Results

P-value:

Table 1: P-value for each input feature

	Runtime	Votes	Metascore
Rating	0	5.42e-178	0

Study cases:
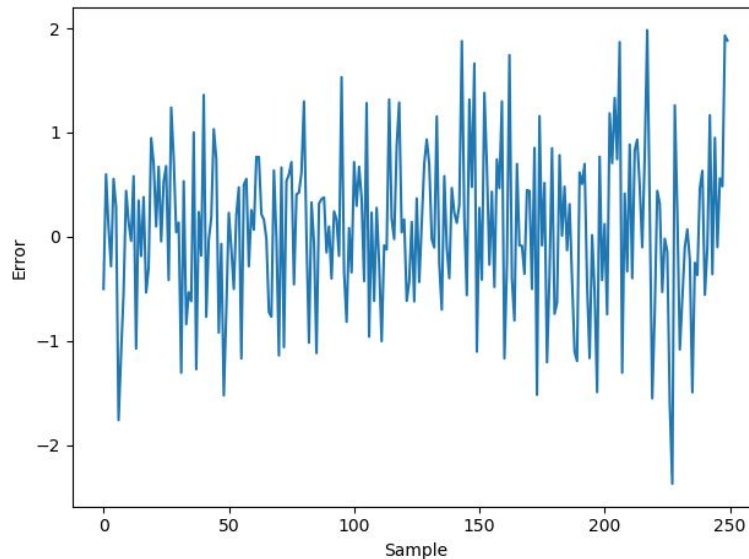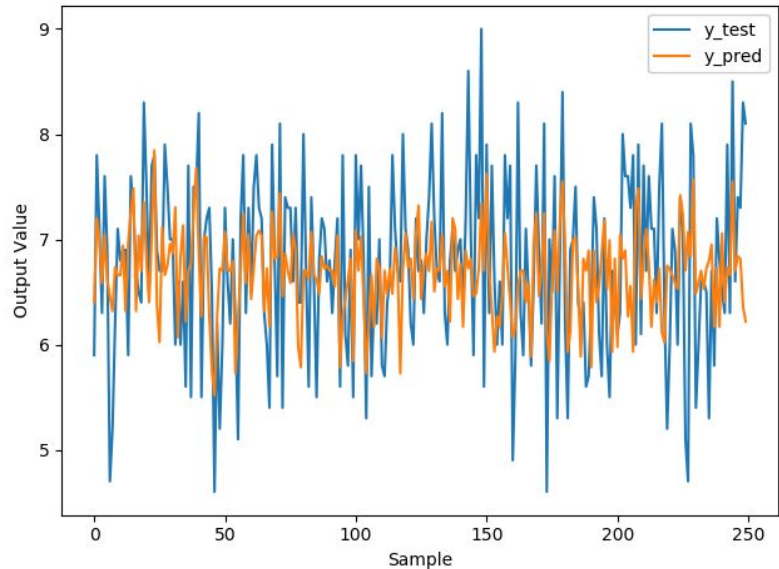
Case I: Input features: Runtime + Votes

Case II: Input features: Runtime + Metascore

Case III: Input features: Votes + Metascore

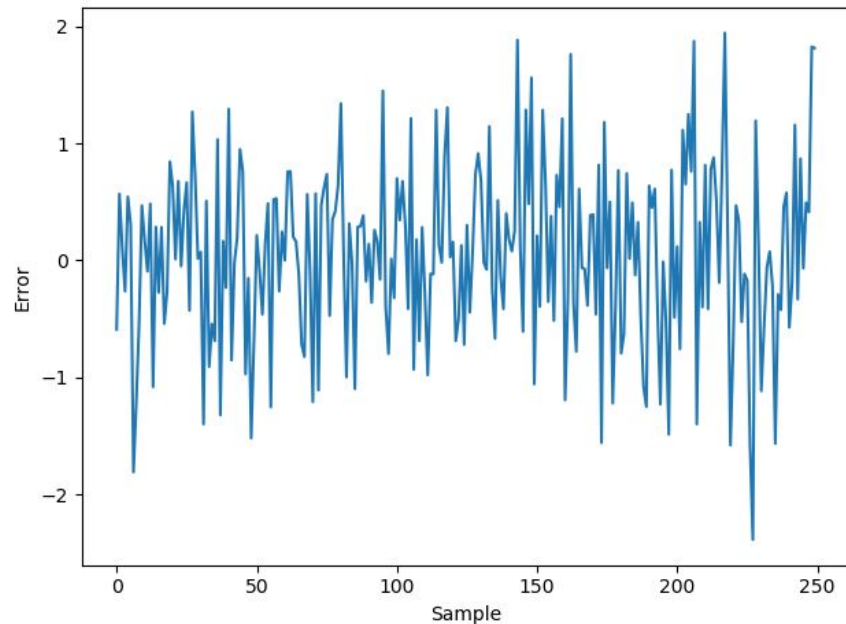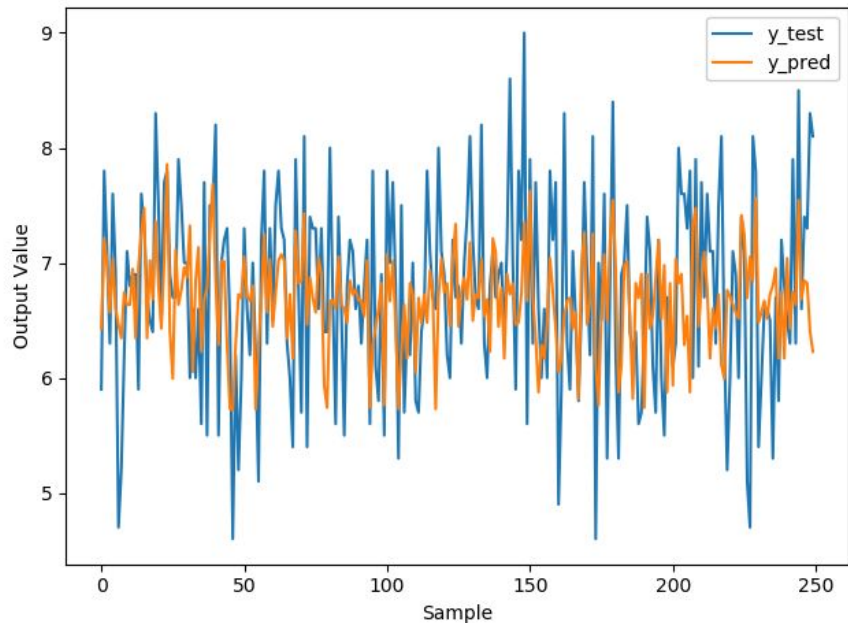Case IV: Input features: Runtime + Votes + Metascore
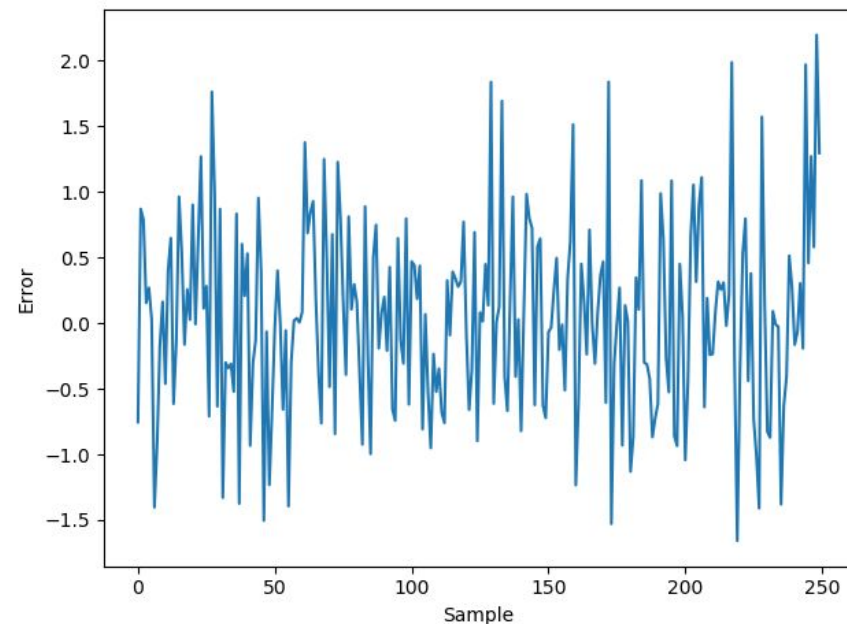
Case I: Runtime + Votes
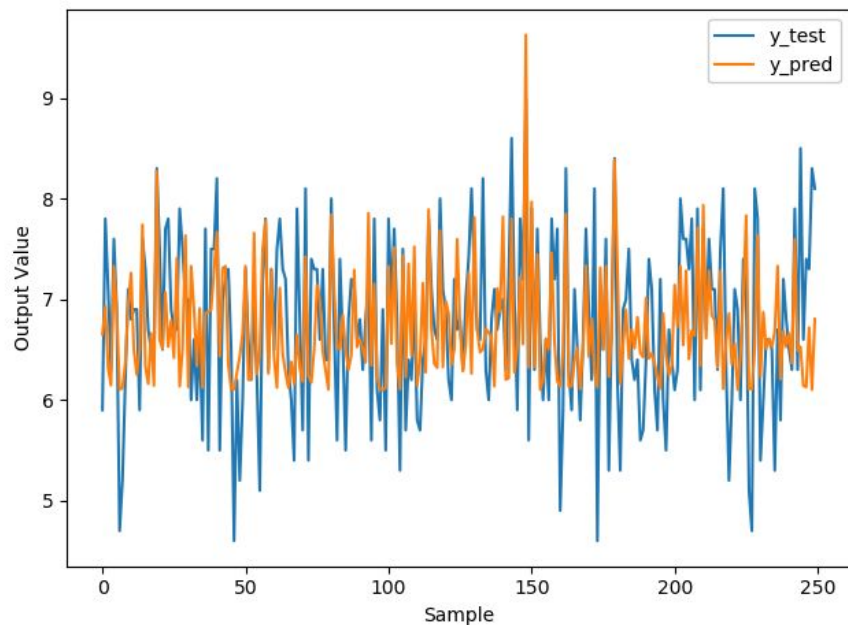
Mean squared error: 0.586

Case II: Runtime + Metascore

Mean squared error: 0.589
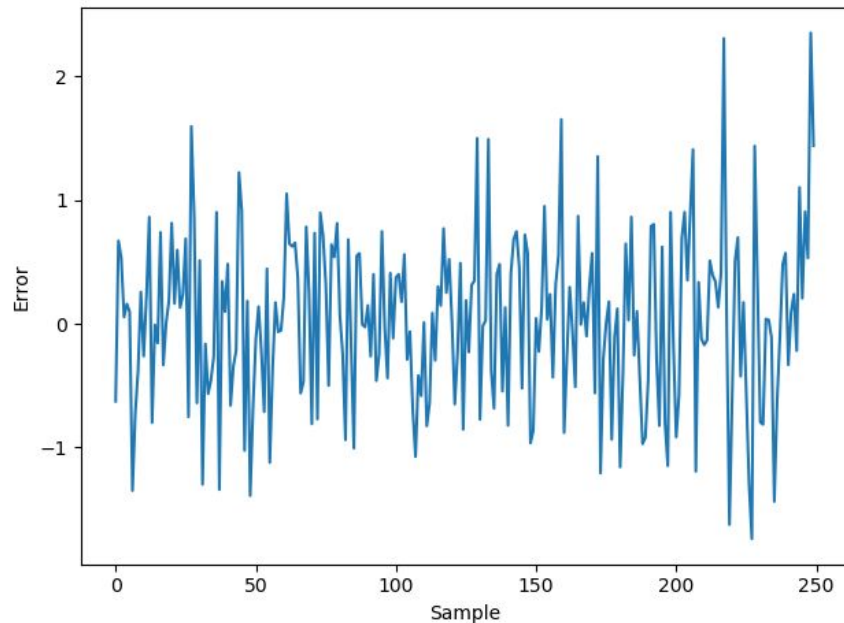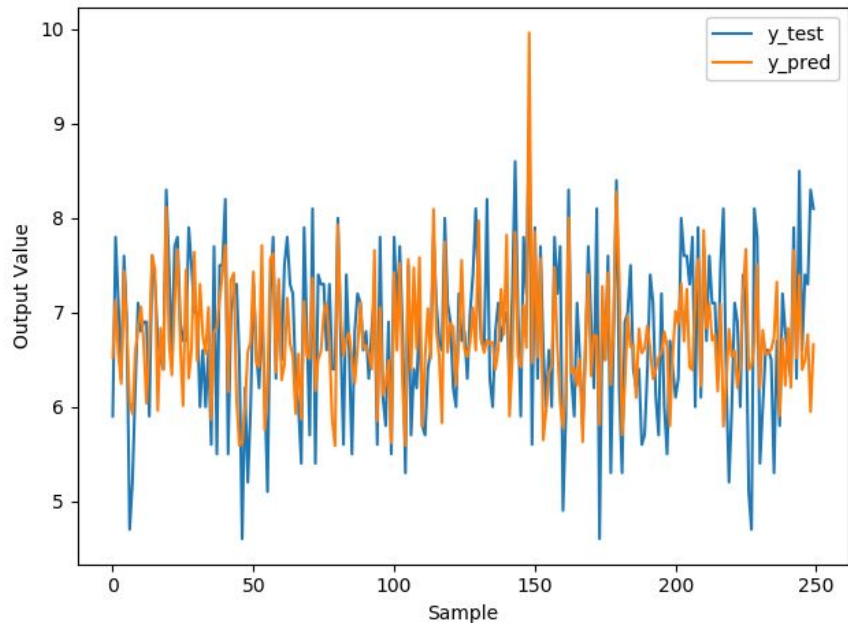
Case III: Votes + Metascore

Mean squared error: 0.518

Case IV: Runtime + Votes + Metascore

Mean squared error: 0.469

Conclusion

The model was relatively satisfactory when all properties (runtime, metascore and votes) are considered as factors to contribute to the rating of a movie.