Individual Report
ML 1 Final Project

Professor: Amir Jafari

Date: 06/30/2020

Done by: Maeshal Hijazi

Introduction:

The goal for our project was to predict the rating of varies movies based on different features such as runtime in minutes, number of votes, metascore, and revenue in millions. The movies data had 100 rows and We, group 1, used the MLPRegressor function from the sklearn package to build our model. In order to build the optimal model, we made sure to pick the right features by picking the features that gave us the lowest mean squared error (MSE).

Description of Individual work:

In order to build an accurate model, data pre-processing had to be applied to the data. After we imported the data, we filled the empty cells with their corresponding column mean. Since all our input features were continuous, there was no need for encoding categorical features. We also did some scatter plots for every input feature vs the output to see how every feature affects the output. Then, we divided our input features and our output for some training values and some testing values. The training values were used to train the model using MLPRegressor while the testing values were used to test our trained model. In order to evaluate our work, we used mean squared error to evaluate the accuracy of our model.

Regarding my individual work, I wrote 75%-80% of the code from importing the libraries, importing the dataset, picking the needed columns, filling the empty cells with their mean values, doing scatter plots for each input vs the output, splitting the data to train set and test set, building the model, predicting the output of the test set, and finally finding the mean squared error of the prediction. I also worked on almost half of the presentation.

Results:

I have conducted many trials for which input features should we pick and conclude of using the movie's runtime and metascore to predict the rating. In Figure 1, we see the scatter plot between the runtime and the rating. Also, Figure 2 shows the scatter plot of the metascore and rating. Both scatter plots show that they are close to linear relationship and thus easier to build the model with. Therefore, I picked these two features. Furthermore, when I build the model using the MLPRegressor, I got a close relation between the the y_test and y_prediction as can be seen in Figure 3. Also, Figure 4 showed the error plot for that model. I got for this model an mean squared error of 0.475

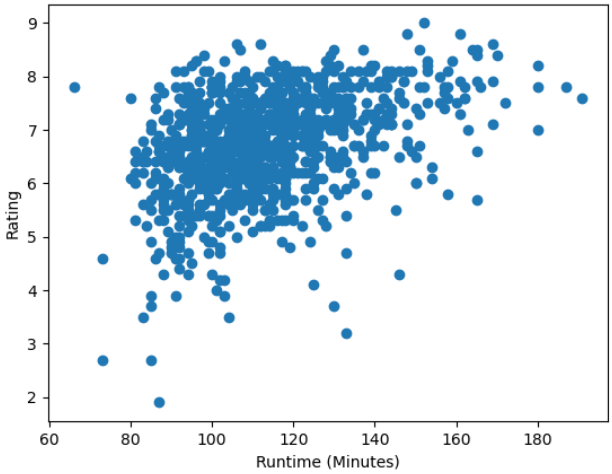

Figure 1 – Scatter plot between the Runtime and Rating

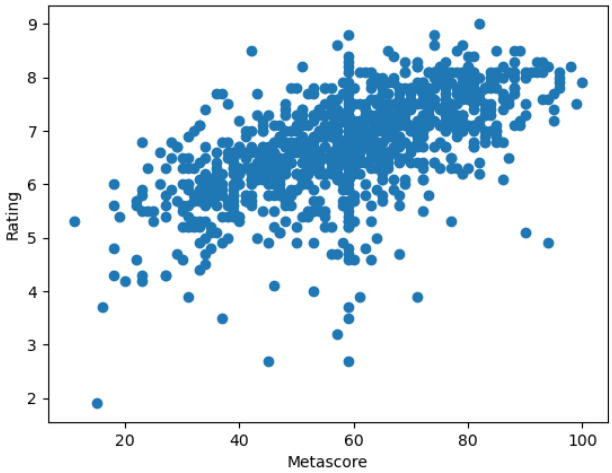

Figure 2 – Scatter plot between the Metacore and Rating

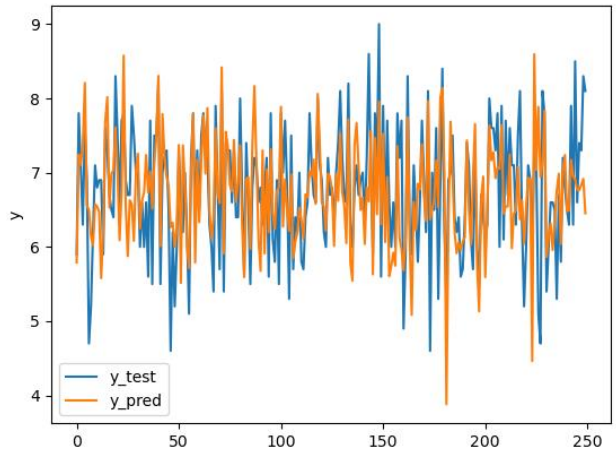

Figure 3 – y_pred vs y_test

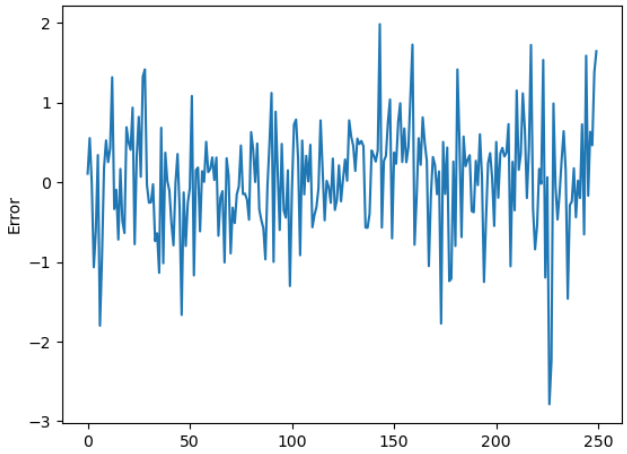

Figure 4 – Error plot

Conclusion:

The objective of this goal was to predict the rating of different types of movies base on different features. Some of the features were movie's runtime, metascore, number of voters, and many other. For the code I wrote, I used MLPRegressor with only two input features to build the ratings. From this project, I learned that the main step in order to build an accurate model is to do the best data pre-processing. Also, I used to think that regression is easier than classification but figured out that it is not. For further improvement, I would divide rating into 4 subsections. Section 1 between 0-2.5, section 2 between 2.6-5, section 3 between 5.1-7.5, and section 4 between 7.6-10. In this way, I will transform the model from regression to classification. The percentage of code I found or copied from the internet, based on the professor's equation, is 7%.