Machine Learning Assignment

Cory Simpkins, Clemson University

April 23, 2019

1 Scraping and Parsing the Data

The site I chose to scrape for this project was *boardgamegeek.com* and I gathered data from their ranked board game list. Both the scraping and parsing sections provided unique difficulties beyond what was covered in class that I had to overcome.

1.1 Scraping

When attempting to scrape the website as we did in class with *urllibrequest*, but since the prices on the page updated when the page was loaded, the html I got from this approach did not include any prices. To solve this issue, I used the Selenium package to control a browser, slept the program to allow the prices to load, then saved the html.

1.2 Parsing

The price data presented another challenge in the parsing process. First, only some games had price data at all. Additionally, there are different prices available for different games, like list price, new Amazon, used Amazon, etc. The solution I decided on was to take an average of all available prices to create an *avg_price* variable. This may not have been the ideal solution, however I think it is reasonable for this assignment.

2 Machine Learning

Since some games had price data and others do not, this makes for a natural supervised learning approach. Using the linear regression model in *sklearn*, I used the data of the games with price information to train the machine, then used the model to predict average price for the non-priced games.

The data set gathered from the website includes game rank, average rating, geek rating, number of voters, and the year it was released. Figure 1 shows the relationship between the test observations' average ratings and average price. Figure 2 then plots the predicted prices and the average rating. Unfortunately, the R^2 of the model is only 0.1078, so this model's prediction ability is quite low. Potential improvements to this model could be made if there were more data for each board game. Each game has its own page, which has additional data such as game type, number of players, playing time, and suggested age group. Taking the time to scrape these specific pages and obtaining this data could give this model more prediction power.

Figure 1: Average Rating and Average Price

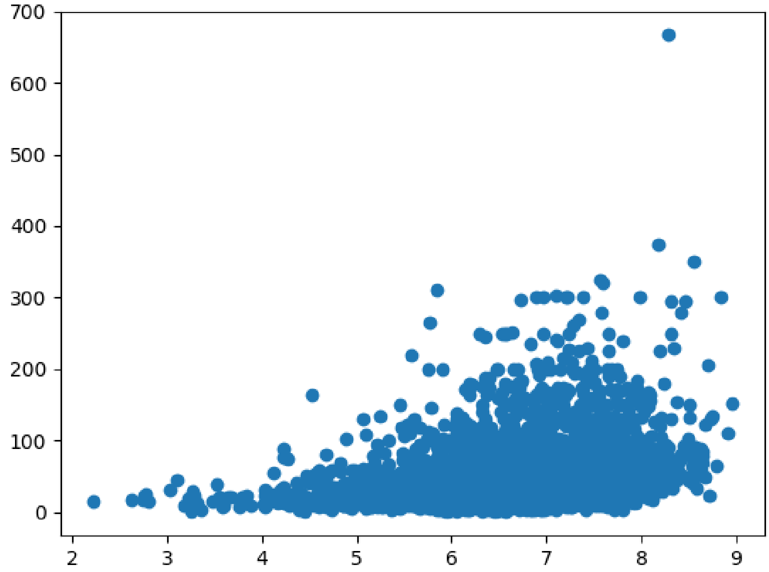


Figure 2: Predicted Average Rating and Average Price

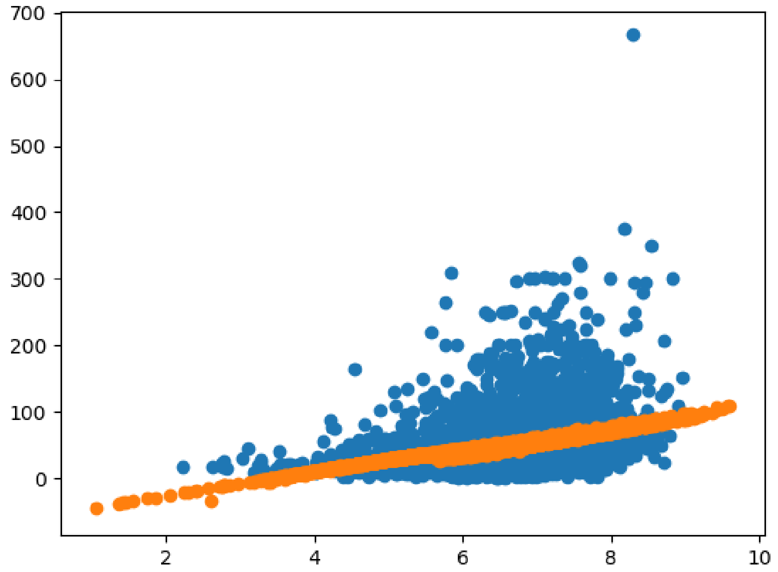
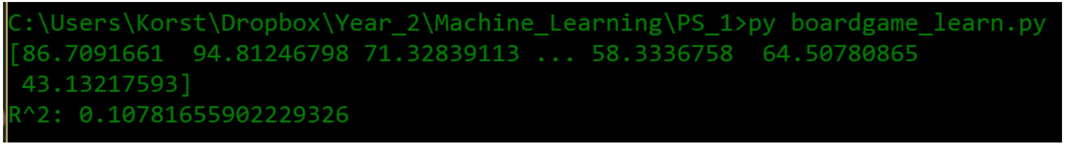


Figure 3: R^2



```
C:\Users\Korst\Dropbox\Year_2\Machine_Learning\PS_1>py boardgame_learn.py
[86.7091661  94.81246798 71.32839113 ... 58.3336758  64.50780865
 43.13217593]
R^2: 0.10781655902229326
```