Machine Learning Assignmnet

Shilpi Mukherjeee

April 2019

1 Introduction

The assignment asked us to scrape one of the websites listed in the homework document. I started with coinmarketcap.com but because of complexity of the scrapping process in the coinmarketcap.com, I chose to move forward with boardgamegeek.com.

2 Scraping Mechanism

I created the geekparser.py script to perform scraping on the website. I could scrape all the variables except the price data, because upon scraping it gave the entire tag and even on trying to read the data through a loop, I was unable to do it. I scraped all the 1066 pages in the website for all the fields except price.. However, following this procedure, the program only scraped the first 170 pages because there was no ranking from the 171th page. This was leading in loss of some potential data. Therefore, I scraped all the 1066 pages without the rank variable. The name of the dataset is "boardgamegeekdat.csv" .Since the data was scraped in chronological order, I then added the rank field to my dataset by appending a rank column.

3 Cleaning Data

As already mentioned I appended the rank column to the dataset obtained from the scraping. I removed all the rows with missing observations and that reduced my number of observations to 22,926 from approximately 170,000 observations.

4 Machine Learning - Supervised and Unsupervised

4.1 Correlation

I find out the correlations between Geek Rating, Average Rating, Rank and the Number of Voters, as these are the only variables that I have. The difference between Geek Rating and Average Rating is that BGG Rating is artificially provided by the company to ensure that games with relatively few votes do not climb to the top of the charts. The algorithm is kept a secret so that it is not manipulated to make the games to reach the top. These matrices

Rank	0.394857
AverageRating	1.000000
GeekRating	0.457632
NumberOfVoters	0.117267

Figure 1: Correlation Matrix of Average Rating

Rank	-0.041510
AverageRating	0.457632
GeekRating	1.000000
NumberOfVoters	0.595516

Figure 2: Correlation Matrix of Geek Rating

Rank	1.000000
AverageRating	0.394857
GeekRating	-0.041510
NumberOfVoters	-0.124156

Figure 3: Correlation Matrix of Rank

show us that number of Rank is negatively correlated with number of voters, which means the greater the votes the the better is the ranking of the game. But the result is opposite for Average Rating. Also, the correlation between Number of Voters and Average Rating is only 0.117267 whereas that between Geek Rating and Number of Voters is 0.595516. I therefore choose Geek Rating as our target variable because it seems to be a better metric of the rating which is the only dependent variable in our dataset.

4.2 Supervised Machine Learning

To perform supervised machine learning, I perform a linear regression and a Random Forest Regression. In the linear regression I use a convenience function to train the algorithm, and then test the data. The training set is 70 percent and the test set is 30 percent. I perform the training with random state equals to 1 so that the results can be replicated if so desired. The Mean Square Error of the test data is 0.0648 whereas the Mean Square Error of the Random Forest Regressor is 0.000702.

The graph of the residual error on test data looks like the following, which is pretty close to Normal with mean 0.

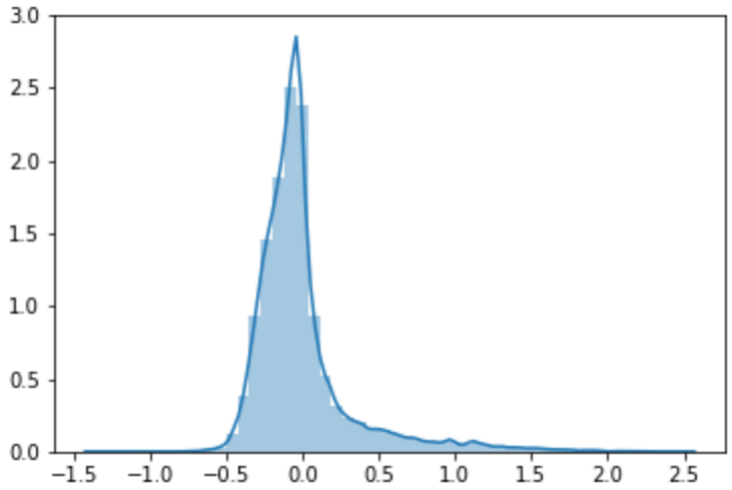

Figure 4: Plot of Residual Errors

4.3 Unsupervised Machine Learning

I then perform a k-means Clustering with five parameters using Principal Component Analysis. In this case, instead of doing a normal k-means clustering I tried to find the cluster of the correlation of Geek Rating with the relation between the other variables.The result of the clustering looks as follows.

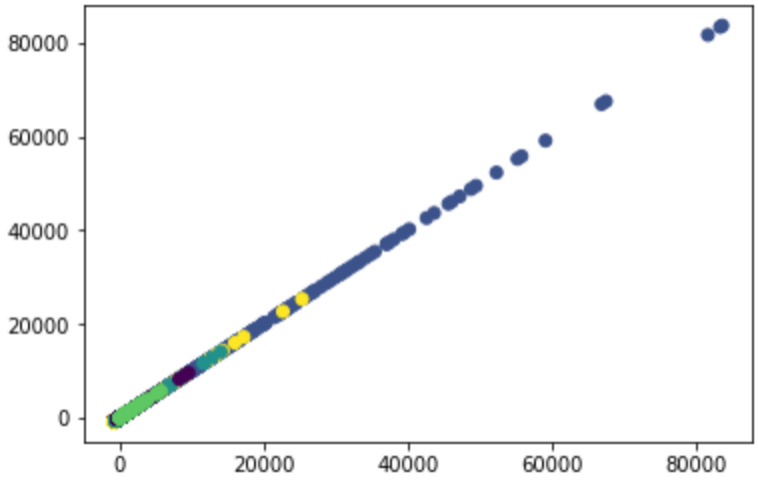

Figure 5: k-means clustering

5 Conclusion

Although I tried to perform some form of machine learning with this dataset, I don't think that any of the results obtained from the exercise were particularly illuminating, because of the absence of the price data. However, I did find that the data could we well trained using supervised machine learning techniques.