# Prediction of bull bear stock market based on sporting event results

**Yu Yao Hsu** yh832@nyu.edu

**Tina Yuan** thy258@nyu.edu

## Abstract

Sporting events evoke a variety of emotions in spectators. These emotions span from positive (such as joy or contentment) to negative (such as despair or anger). Research has shown that emotions can affect investors' decision: a positive mood predicts an increase in stock prices while an unpleasant mood predicts a decrease in stock prices (Cohen-Charash et al., 2013). Research has been conducted in Europe where the authors found a correlation between international soccer results and the country's stock market (Boidoa and Fasanob, 2007). This project aims to determine if there is a relationship between NBA game results and US stock prices as measured by the S&P 500 index. The study uses Twitter posts on game days and NBA game attendance as the input data. After processing the data, we built a neural network and achieved a maximum accuracy of around 60%. Although the results suggest that there may not be a strong causal relationship between NBA results and the stock market, further studies can improve on the methodology to investigate the relationship between sporting events and the stock market in the US.

**Keywords:** Predictive Analytics, NBA, Stock Market

## 1. Introduction

Emotions have severe effects on individuals' decision making (Lerner et al., 2015). Research has shown that moods can affect investors' decision: a positive mood predicts an increase in stock prices while an unpleasant mood predicts a decrease in stock prices (Cohen-Charash et al., 2013). To better measure people's moods, researchers have studied how Twitter posts and comments can be indicative of the public's mood and thus improve the accuracy of predicting stock market fluctuations (Bollen et al., 2011). As the NBA is a national league within the States, data is collected from social media platforms like Twitter to assess the public's support for individual teams and people's moods after each game. This project aims to determine if there is a relationship between NBA game results and US stock prices as measured by the S&P 500 index.

## 2. Data and Business Understanding

### 2.1 Business Understanding

Sporting events influence people's moods. A study has shown that losses in soccer matches have an economically and statistically significant negative effect on the losing country's stock market (Edmans et al., 2007). For example, when being eliminated in a major international soccer tournament, the national stock market index on the following day would be lower than average. Another study on the relationship between football and mood in the Italian stock market found that the average price/return ratio is higher when the Italian team won a soccer game (Boidoa and Fasanob, 2007). These research indicate that sporting events can alter people's mood and decision making, which suggest that these impacts can be an indicator of the fluctuations in the stock market. We would like to investigate if sporting events such as NBA basketball games in the United States have similar impacts on the US stock market.

### 2.2 Data Understanding

We collected two types of data: structured and unstructured data. For structured data, we collected "Total Games" and "Total Attendance" from the website: basketball-reference. We believe that the more games that happened within a certain time frame, the larger the population of people whose moods could have been influenced by game play results. Thus, we include total game and total attendance as our training features. For the unstructured data we used the Python library snscrape to fetch Twitter posts that contain the name of the teams that played on game day and posts that has the keyword 'NBA'. We collected a maximum of 5000 tweets for each NBA game. We then performed sentiment analysis on the tweets to see if they express positive, negative, or neutral emotions towards the NBA games played on that day. In addition to sentiment scores calculated from the tweets, we also collected the number of a tweet's like, comment, retweet, and reply to see how much support this tweet gained from the crowd. Lastly, we used Python library beautifulsoup4 to extract the S&P500's daily closing index from Yahoo Finance. We chose the S&P500 index as the index to represent the US economy because it is a weighted measurement of the 500 largest companies in the States. For all of the data we collected, we collected data from the year 2006 until 2020.

### 2.3 Data Preparation

To prepare the data for model building, we followed the procedures below on the tweets that we collected. First, we removed duplicate records by identifying duplicate tweet id's and duplicate tweet contents. Since we included the keyword 'NBA' in all the games searched, we collected duplicate tweets within the same day. Second, we removed any URLs in the record. URLs can be links to articles, pictures, or videos about NBA games but are not useful in performing sentiment analysis. Third, we removed invalid records. Some tweets we collected are invalid (e.g. wrong information in columns). We removed these kind of data so that the final input vector does not include missing values. Lastly, after all these steps, we used the Python nltk library to perform sentiment analysis, which produced output values of the tweet's positivity, negativity, and neutrality.

After we processed the tweets, we built the input matrix for the neural network. Since the S&P500 index only has values from Monday through Friday, while NBA games happened all days of the week, we used all the information learned from the current week to predict if the S&P500 index would go up or down the following Monday or the first market open day of the following week. Our total features are as following:

1. Counts of tweets(Pos, Neg, Neu): 3 columns
2. Percentages of the tweets(Pos, Neg, Neu): 3 columns
3. Counts for reply, retweet, like and quote(Pos, Neg, Neu): 12 columns
4. Number of games: 1 column
5. Total attendance: 1 column
6. Label:Next week first close index - this week first close index. If it is $> 0$ then it is assigned value 1 and 0 otherwise. 1 column.

Initially we created four different input matrices like the following:

1. (Pos, Neg, Neu) tweet count, (Pos, Neg, Neu)tweet reply, retweet, like, quote, total game count, total attendance, label
2. (Pos, Neg, Neu) tweet count percentage, (Pos, Neg, Neu)tweet reply, retweet, like, quote, total game count, total attendance, label
3. (Pos, Neg, Neu) tweet count, total game count, total attendance, label
4. (Pos, Neg, Neu) tweet count percentage, total game count, total attendance, label

The first one includes all the original features we collected. The second one we switched the three tweet count features from counts to percentages; we believe that by normalizing the values of these three columns to the same range can help the model learn better. The third and forth matrices are the same as the first and second ones, except for the fact that we omitted the tweet reply, retweet, like, and quote features.

With these 4 different input matrices, we then used the correlation matrix to keep columns that have similarity $< 0.5$. This step is performed so that the model doesn't use multiple similar features for prediction. In the end we have 8 different input matrices to test with.

## 3. Modeling and Results

### 3.1 Modeling

For the 8 different models, we went through the following procedures to get predictions. We first trained and validated the model using the cross validation with 10-fold on neural network. Next, we tested the model and optimized on the learning rate and momentum. Finally, we output the performance using the confusion matrix. See Figure 1, 2, and 3 for the model pipeline constructed in RapidMiner.

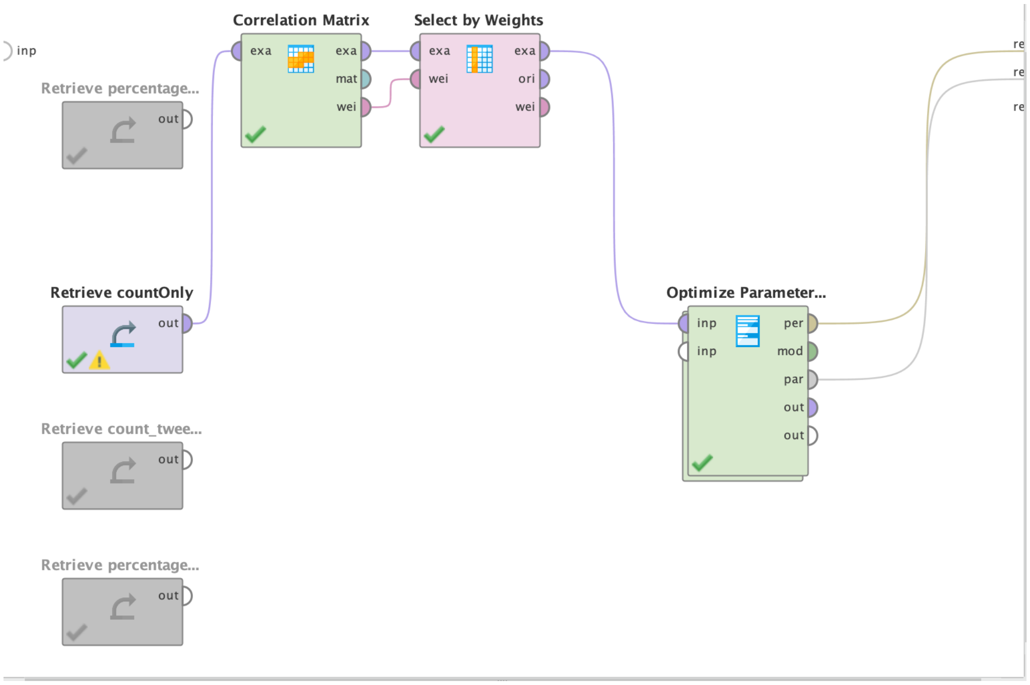


Figure 1: Four runs of initial four input matrix will go through 'Optimize Parameter' directly, while the other four runs will go through correlation matrix selection first before running the 'Optimize Parameter'

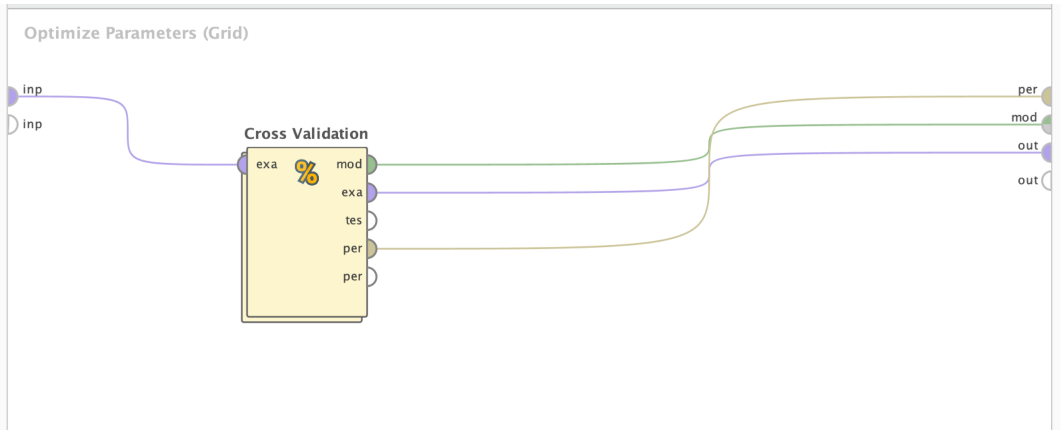


Figure 2: Optimize Parameter will run the model with 'Cross Validation'

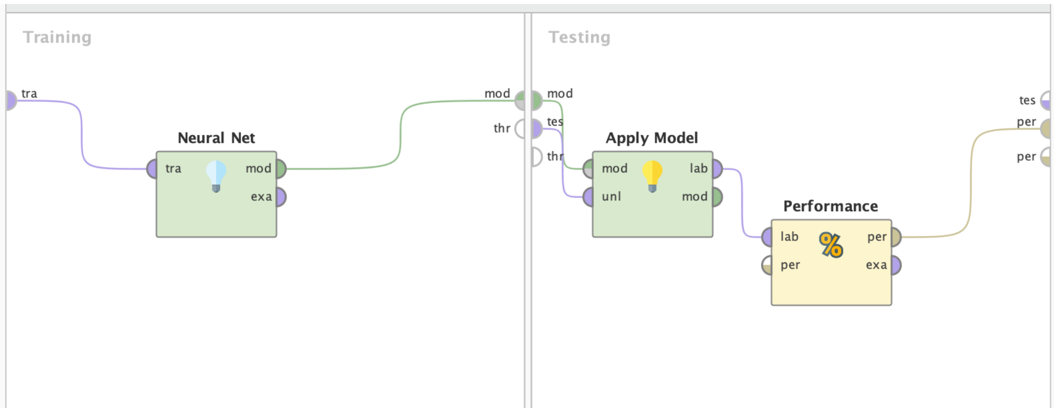
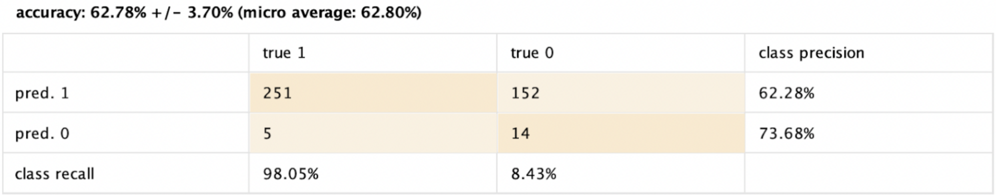


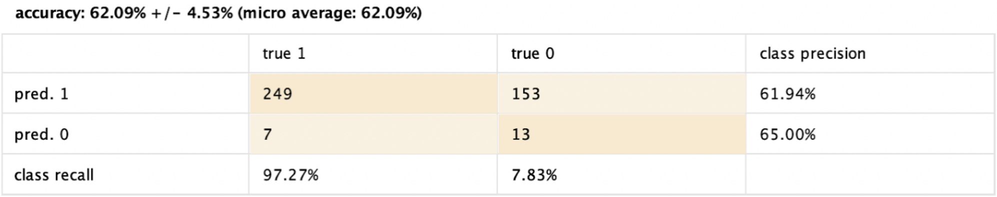
Figure 3: 'Cross Validation' will be performed with neural net and tested the model while optimizing on learning rate and momentum



**accuracy: 62.78% +/- 3.70% (micro average: 62.80%)**

| | true 1 | true 0 | class precision |
|---|---|---|---|
| pred. 1 | 251 | 152 | 62.28% |
| pred. 0 | 5 | 14 | 73.68% |
| class recall | 98.05% | 8.43% | |

Figure 4: Confusion matrix for model built with Number of Games, Total Attendance, and Label columns



**accuracy: 62.09% +/- 4.53% (micro average: 62.09%)**

| | true 1 | true 0 | class precision |
|---|---|---|---|
| pred. 1 | 249 | 153 | 61.94% |
| pred. 0 | 7 | 13 | 65.00% |
| class recall | 97.27% | 7.83% | |

Figure 5: Confusion matrix for model built with (Pos, Neg, Neu)tweet count percentage, Number of Games, Total Attendance, and Label columns

### 3.2 Results

We achieved the following accuracy for the 8 models: 60.9, 60.91, 60.9, 60.9, 61.37, 61.61, 62.78, 62.09. All models are around 60% accurate. The two that achieved $> 62\%$ accuracy are the input matrices with the following columns:

1. 62.78%: Number of Games, Total Attendance, Label
2. 62.09%: (Pos, Neg, Neu)tweet count percentage, Number of Games, Total Attendance, Label

From the two models that produced the best results, we learned that tweet count percentage is a better feature than tweet count. This confirms our hypothesis that normalizing tweet counts to percentages produce better models. In addition, we learned that the number of NBA games and game attendance might be better indicators of the stock market than Twitter moods about NBA games.

See figures 4 and 5 for the confusion matrices of the two models that produced the best results. The accuracy value suggests that we were able to correctly guess around 62.09% / 62.78% of the bull or bear markets. The precision value indicates the number of correctly predicted cases that actually turned out to be positive. Out of all the bull markets we guessed, we predicted 61.94% / 62.28% correct. Out of all the bear markets we guessed, we predicted 65% / 73.68% correct. Finally, the recall value tells us the number of the actual positive cases we were able to predict correctly with our model. Out of all the real bull markets, we predicted 97.27% / 98.05% correctly. Out of all the real bear markets, we predicted 7.83% / 8.43% correctly.

## 4. Evaluation and Discussion

The goal of this study is to predict both bull and bear markets correctly, thus the higher the model accuracy the better. Our model accuracy of around 60% indicates that when in the future our model learns from a week of NBA data, we can only be 60% certain that the first S&P500 index of the following week would be going up or down. The recall value in the confusion matrices (figures 4 and 5) communicates that the model is really good at predicting bull markets, while really bad at predicting bear markets. The model is highly biased towards predicting bull markets as the model predicted only 20 bear markets out of the 422 total cases.

Based on the experiment that we performed with NBA data, it is possible that basketball games are not be able to represent the majority mood of the investors. Unlike soccer in Europe, where it might be the biggest sport event that people care about, people in the US watch not only basketball(NBA), but also baseball(MLB), football(NFL), and ice hockey(NHL). If we want to use sporting event results to predict the stock market, we might need to take all professional sport leagues into consideration. In addition, investing firms usually have the largest impact on the stock market. We collected tweets from all kinds of people: individual investors, people that don't invest in the stock market, etc. If we want to learn from twitter mood and achieve high accuracy, we should try to collect tweets from investing firms and influential individuals, since they are the ones that affect the stock market the most. Additionally, the data set may be too small to train good models: the input matrix only has 422 rows of data. Since NBA games do not happen the entire year, there are stretches of time where there is no input data. This severely decreases the amount of data our model can be trained on. Besides, most of the tweets collected were assigned a neutral score by the nltk sentiment analysis tool. Our original hypothesis states that game results may affect people's emotions and thus be reflected on social media platforms like Twitter. However, the Twitter data collected does not show a strong indication of people's moods fluctuating due to NBA games. Moreover, our training data that helps our model learn about bear markets might not have significant traits. For example, there might be two rows of data that have really similar attributes but one with the label of bull market and the other one with the label of bear market. If most of our bear training data are similar to bull training data, then the model will not be able to learn how to differentiate bear cases. Since our data has around 60 % bull cases and 40 % bear cases, we assume that our model will classify most of these confusing cases as a bull market. Our current model is trained and then reads in one week of NBA data to predict if the following week's index will go up or down. The data is a type of time series data where each time step is one week. Model accuracy may be improved by using a recurrent neural network so that data the model uses several weeks of past data to make predictions.

## References

Claudio Boidoa and Antonio Fasanob. Football and mood in italian stock exchange. *Review of Financial Studies*, 14:1–27, 2007.

Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

Yochi Cohen-Charash, Charles A. Scherbaum, John D. Kammeyer-Mueller, and Barry M. Staw. Mood and the market: can press reports of investors' mood predict stock prices? *PloS one*, 8.8:e72031, 2013.

Alex Edmans, Diego Garcia, , and Øyvind Norli. Sports sentiment and stock returns. *The Journal of finance*, 62(4):1967–1998, 2007.

Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam4. Emotion and decision making. *Annual Review of Psychology*, 66:799–823, 2015.