Moneyball Investing: Making Stock Market Predictions with New York Yankees Game Data

Aidan Carey, Richie McNamara

PROBLEM

The problem that we explored in this project was whether a usual indicator can exist in the stock market. We chose the New York Yankees due to their location in a financial center of New York City. We examined the relationship between the New York Yankees and some of their corporate sponsors. We wanted to know if the game results of the New York Yankees could predict whether the price of some of their corporate sponsors would increase or decrease.

Unusual indicators in the stock market have been studied before. The “Super Bowl Indicator” is a well-known theory that predicts the future of the stock market based on which conference wins the NFL super bowl. It has been correct 41 out of 56 times, but there is no real correlation between the NFL and the stock market.

However, this project is more meaningful because of the real connection between the Yankees and their corporate sponsors. We are also considering game performances, which is more telling than simply observing who won the super bowl.

DATA

We are using stock data from Yahoo Finance and Yankees game results from Baseball Reference. For all stock data, we are using adjusted close price. We considered using Fangraphs for our Yankees data, but Baseball Reference had more robust statistics from each game. We did not consider other stock data sources, since Yahoo Finance is the standard.

We cleaned the data from Baseball Reference by removing insignificant features and standardizing the format to float values so that our data would work seamlessly with our decision trees. When the Yankees played doubleheaders, we omitted game 1 since there was no natural way to merge them. We altered the stock data by creating a new label for whether the next day’s stock price increases or decreases.

We only used data from the 2017, 2018, 2019, 2021, 2022 seasons. We ignored 2020 due to the COVID-19 pandemic’s shortening of the MLB season. We also only considered data from Mondays – Thursdays, since Friday’s Yankees results may not impact Monday since they could play Saturday and Sunday.

	Tomorrow's BAC Increase	Gm#	Home	Opp	Win	R	RA	Inn	W-L	Rank	GB	Time	Day Game	Attendance	cLI	Streak
2022-04-11	0.0	4	1	29	0	0	3	9.0	0	3	1.0	183	0	26211.0	1.04	-2
2022-04-12	0.0	5	1	29	1	4	0	9.0	1	2	1.0	187	0	25068.0	0.99	1
2022-04-13	1.0	6	1	29	0	4	6	9.0	0	3	1.0	190	0	30109.0	1.03	-1
2022-04-19	0.0	11	0	10	1	4	2	9.0	1	1	0.0	239	0	15498.0	0.98	1
2022-04-20	0.0	12	0	10	1	5	3	9.0	2	1	0.0	186	0	17268.0	0.99	2
2022-04-21	0.0	13	0	10	0	0	3	9.0	1	2	1.0	154	1	21529.0	1.03	-1
2022-04-26	1.0	17	1	3	1	12	8	9.0	5	2	0.5	199	0	28596.0	1.11	4
2022-04-27	0.0	18	1	3	1	5	2	9.0	6	1	-0.5	177	0	31122.0	1.18	5
2022-04-28	1.0	19	1	3	1	10	5	9.0	7	1	-0.5	231	1	29268.0	1.18	6
2022-05-02	1.0	23	0	29	1	3	2	9.0	11	1	-2.5	174	0	18577.0	1.35	10

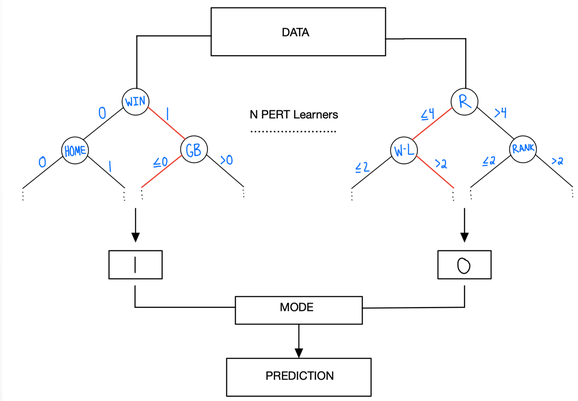
MACHINE LEARNING

The goal of our machine learning model is to predict whether the price of a specific stock will increase or decrease on the following day. We chose to use decision trees because they are explainable and popularly used in finance. We compared the performance of the CART vs. PERT algorithms to decide on which to use for our experiments. The PERT outperformed the CART algorithm often, so we decided to use the PERT algorithm for our experiments.

We elected to utilize an ensemble of PERT learners for each individual stock for our model. To do this, we needed to adjust our bootstrap and PERT learners to handle a classification problem instead.

MACHINE LEARNING (cont.)

We create a model for every stock for every year. We chose the leaf size that resulted in the best test accuracy. The following chart is an example ensemble of PERT learners.



CORPORATE SPONSORS

We researched which companies sponsored the Yankees throughout the years that we focused on. When deciding on which companies to include, we considered whether the sponsor’s relevant stock was traded often. We selected the following corporate sponsors to analyze: Bank of America (BAC), Delta Air Lines (DAL), Ford (F), Hess Corporation (HES), Nathan's Famous (NATH), PepsiCo (PEP), Mastercard Incorporated (MA), T-Mobile US (TMUS), Anheuser-Busch InBev (BUD), Sony Corporation (SONY), AT&T (T), and Volkswagen AG (VWAGY).

We narrowed down our sponsors using a performance metric. For all five years, we trained a random forest for each corporate sponsor on data from the first 65% of the season and tested on the rest of the season. We gathered the model’s accuracy and only used sponsors with over 50% accuracy for at least 3 of the 5 years. This allows the portfolio to have some predictive credibility.

	BAC	DAL	F	HES	NATH	PEP
Test Accuracy > 50%	3	3	4	4	3	1

	MA	TMUS	BUD	SONY	T	VWAGY
Test Accuracy > 50%	2	1	3	3	4	1

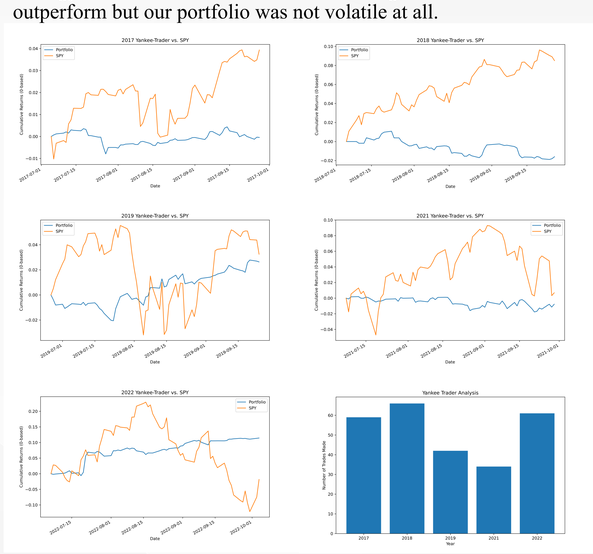
We encountered some issues with the corporate sponsors. Information for the sponsors is not as public as we'd like, which does not allow us to have a concrete metric for how much a sponsor is connected to the Yankees. We also could not find a reasonable way to separate game results from other factors that affect New York. These issues resulted in using a uniform allocation when creating a portfolio of the corporate sponsors.

YANKEE TRADING STRATEGY

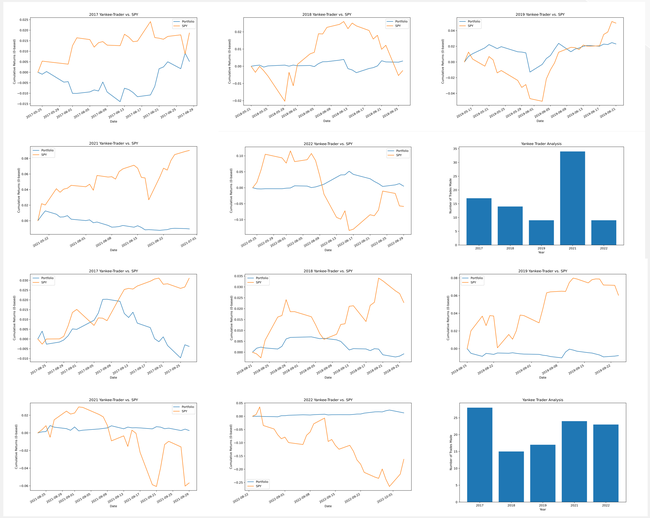
We devised a stock trading strategy using the machine learning predictions based on Yankees Data. For a given trading time range, we created a portfolio of the selected corporate sponsors. For a given year and stock, we trained models for each using data prior to the trading start date. If the model predicts an increase in price, our trader will long 1000 shares of the stock. If the model predicts a decrease in price, we will short 1000 shares. We then keep track of the portfolio over time.

EXPERIMENTS AND RESULTS

In our first experiment, we created a portfolio of the narrowed down corporate sponsors. We trained each model with the optimal leaf size over the first 60% of the games each season. Then, we tested our strategy on the last 40% of the games each season. We backtested our portfolio and compared the cumulative return of our portfolio with the SPY over that same time. We hypothesized that the SPY would outperform our portfolio consistently and that our strategy would be more volatile. In our testing, the SPY did outperform but our portfolio was not volatile at all.



In the second experiment, our goal was to see if adjusting the training and testing dates would have an impact on our returns. In our first run, we trained on the first 25% of each season and tested on the second 25%. Similarly, in the second run, we trained on third 25% and tested on the last 25%. Finally, we trained on a full season and then tested on all the next season. We hypothesized that the time of the season we tested on would not have a big impact on our results. Additionally, we hypothesized that training on a full season, then testing on the following season would not lead to great results since there is significant gap in the winter for much to change in the stock market. In our experiments, the time of year had little to no effect on the performance of our portfolio.



EXPERIMENTS AND RESULTS (cont.)

As for our experiment that trained on a full season then tested on the next season, the SPY was the better choice in 2 of the 3 years that we ran our models, which aligned with our hypothesis. The only year where our strategy outperformed the SPY was 2022 where the SPY declined in that period.

In our final experiment, we focused on just the 2022 season. We trained on the first 60% and tested on the last 40% like we did in the first experiment. However, this time we wanted to compare the performance of our trading strategy of each individual stock to a technical strategy (mean-reversion) and the Q-trader. We hypothesized that the Yankee Strategy would be outperformed by the technical strategy and the Q-trader. In our experiment, the Yankee Strategy in fact performed the best on the test data 4 out of 8 times.

CONCLUSIONS

We conclude that in general it is better to buy and hold the SPY if you are trying to make money. In the second experiment, we determined that the limited time data makes it hard to analyze how the time of the season impacts the portfolio’s performance. In addition, we determined that in most cases, the data from the previous season does not carry over to the next season. However, we did notice that our strategy is on par with the performance of the technical strategy (mean-reversion) and Q-trader for 2022, which could have just been a coincidence.

An experiment that we wanted to run but could not at this time was expanding to look at all MLB teams and their on-field uniform sponsors. Since uniform sponsors were new for the 2023 season, there is not enough data available at this moment to conduct an experiment yet. However, after this season, a project could be run to look at this relationship. Additionally, future research could be done in other sports and leagues as some have had on-field uniforms for a much longer time period. The question would just remain if there is enough game result data available as well.