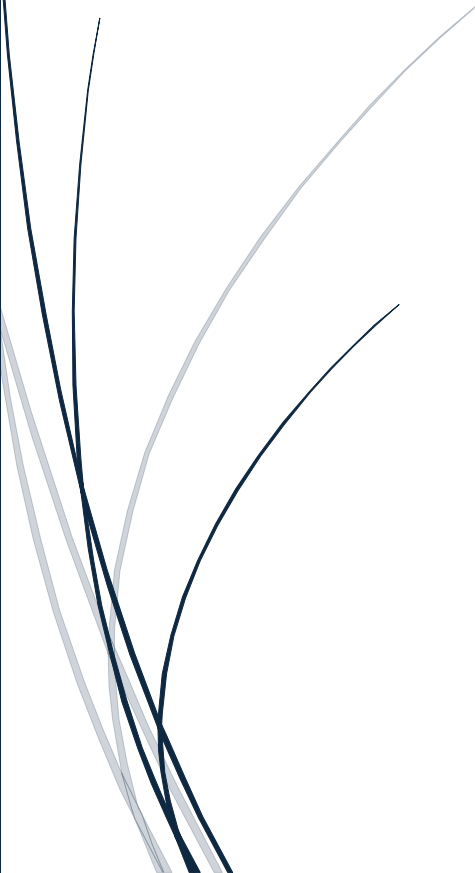


1/16/2024

Predictive Model for Stock Portfolio Value

Data Science Capstone Project
proposal



Eric S. Schneider
eschneider@bellarmine.edu

Executive Summary: In order to combat against the highly volatile nature of the stock market, I plan to create a model which will utilize past financial data, specifically ratios and earnings, in order to predict the one-year earnings or losses of a portfolio. The important ratios which will be used are return on equity, current ratio, and asset turnover ratio. The Weighted Average Cost of capital and percentage change in net income will also be used to help represent the volatility of cashflow in company value. The models which will be tested will most likely be KNN, Gradient boosting regression, and Random Forest. The validity of the model will be tested using a regression analysis in R measuring the strength of a resulting linear regression model formed from only the predicted percentage change and the actual percentage change. A Gini coefficient will also be used to show how much the percentages approximate the line which assumes predicted percentage equals actual percentage.

Project overview: Overall, I plan to create a program that will be able to predict the percentage change of a stock over the course of a full year. From this I plan to extrapolate the program's design to be able to predict what will happen to a full portfolio. In practice, the module would accept the different stocks in the portfolio, as well as their weighted capital amounts, and return the total predicted percentage change in the value of the portfolio over the course of an entire year. The predictive model would utilize numerous financial ratios related to previous years of operation meant to predict the overall profitability of a company. Additionally, the model will pull percentage differentials from previous income statements in order to gauge previous financial health through variables such as percentage change in revenue. There are multiple clear steps for what I will need to accomplish in order to ensure a fully random and fair test group, starting with creating a sample dataset to begin with.

Because of the scope I am looking for in a dataset, I am going to have to construct a dataset from scratch. However, the steps to do so are relatively easy to break up. First, I will need to construct a program which will return a list of at least five thousand stocks to use as a sample group. In this stage, only the names as they appear on the stock market will suffice. Next, I will need to construct a program which pulls the necessary data from the balance sheets and income statements of the sample stocks. Finally, I would have to clean and refine the dataset to show the desired ratios and differentials for the predictive analytics model. From there, I will be able to begin programming in the predictive models to output the predicted change in each stock, which will then be weighted by their given capital weights to output the projected earnings of the entire portfolio.

Background: There is an extremely large industry focused on the idea of using current information about a company to predict their future value. It is an incredibly important role which is entrusted to those skilled in the field of stock analysis and analytics. However, the main problem occurs with the fact that stock values fluctuate almost by the second, and it can be difficult to measure the true potential that a stock can either climb or fall to. In many cases as well, once the analysis is completed, the potential of a stock changes due to the ongoing practices of the company. However, when taken in a snapshot, much of the data provided at year end can shine light on the trajectory of the company for the coming year. Specifically, the change in revenue and net income can help predict the basic profitability of future years in terms of sheer earnings, with ratios such as return on equity being able to show the ability of the company to effectively minimize expenses. The current ratio, as well as turnover ratios can help show the ability of a company to finance short term debt. Turnover ratios such as asset turnover can be used to represent the relative risk based on the market. Finally, the weighted average cost of capital, otherwise known as WACC, can help predict the company's rate of financing cashflows, which have the potential to intensely magnify the effects on the future profitability of a company. In all, given a sample of at least three years, these ratios can help predict the future value of a company's stock.

Modeling: There are quite a few predictive models to choose from in the case of stock analysis. The most common, and simplest would be the K nearest neighbor (KNN) algorithm, which predicts the value of a new point based on its nearest neighbor points. Since the output of this model will be numeric, rather than binary, KNN will be useful as it can adequately utilize the dimensional data in order to finetune and produce an accurate model. The second model which I have the option to utilize is Gradient Boosting Regression. What is special about this model in terms of stock analysis is that it works to limit the effects of covariance in order to eliminate bias in the model. There are a number of other regression algorithms which can be implemented, such as LASSO regression and Gaussian regression, and they may prove to be useful. However, in cases where non-linear regression would be more effective, which it is possible that this situation would be, there are three other algorithms which may prove to be effective, which I would like to look into. The first is Random Forest, which utilizes multiple decision trees to create a more accurate predictive measure. Long Short-Term Memory (LSTM), which makes decisions based on preceding content, which may be useful seeing how the datapoints are based on changing financials between years. The final model is known as an Autoregressive Integrated Moving Average (ARIMA), which is specially made to predict values of highly sensitive, time-based regression models.

Most likely, the majority of these models will not be tested, but all of them are worth researching in order to find three different methods for prediction algorithms. At the moment, I believe that the most effective models would be KNN, Random Forest, and Gradient boosting regression. However, I intend to look into the other methods listed to see if they would be more effective and accurate given the desired inputs for the model.

Tools: Sci kit learn will be used extensively to implement and test the validity of each model. In addition, pandas will be used to model and clean the dataset which will be used for the predictive analysis. For evaluation purposes, I intend to run an analysis through R. Specifically, rather than qualifying each test as a binary pass or fail grade, I intend to run a linear regression analysis using the predicted percentage change, and the actual percentage change. When graphed, an accurate model will approximate a straight line with the formula $X=Y$. Additionally, the use of a Gini coefficient can help track the variability in the data. An accurate model should produce a relatively small Gini coefficient. Finally, Tableau can be used for visualization purposes of important ratios, differentials, and portfolio values. Microsoft Excel can also be used for visualization purposes in rare cases where working in tableau becomes too complex.

Conclusion: With how volatile stock prices have become in the modern financial era, it can be difficult for a person who knows very little about financial analysis to predict future prices. The prediction model resulting from this project serves to give certain insights into future stock values in order to give a prediction of the earnings of a stock portfolio. That being said, the limit of this model would be that it only shows the forward percentage for an entire year, given data that is only released once per year on financial statements. By its nature, this model will only be able to show the future stock price one year out from the release of the previous financial statements. However, given that nearly all investors believe in long term investments over short term, I do not believe that this will have a strong bearing on the validity of the model.

References:

Author links open overlay panelSupriyo Ahmed a b, a, b, c, recently, A., Alhnaity, B., Ballings, M., Bi, J., Castán-Lascorz, M. A., Chen, W., Colasanto, F., Fischer, T., Graves, A., Hsu, M.-W., Kamara, A. F., Li, J., Pérez-Chacón, R., Pradhan, T., Ronaghi, F., ... Persio, L. D. (2022, May 26). *Poly-linear regression with augmented long short term memory neural network: Predicting time series data*. Information Sciences.

<https://www.sciencedirect.com/science/article/pii/S0020025522005114#:~:text=LSTM%2C%20with%20its%20feed%20back%20connections,some%20dependent%20and%20independent%20variables.>

Gradient boosting regression. scikit. (n.d.).

https://scikitlearn.org/stable/auto_examples/ensemble/plot_gradient_boosting_regression.html

Team, O. M. E. (2023, October 13). *10 popular regression algorithms in machine learning*. Online Manipal. <https://www.onlinemanipal.com/blogs/popular-regression-algorithms-in-machine-learning>