Analyzing the Weekly S&P Stock Market Data using Logistic Regression

CMSC 6950 - Computer Based Research Tools and Applications

Term Project

6th August, 2020

Submitted by

Kwadwo Nyame Owusu-Boakye
(201990860)

Memorial University of Newfoundland
St. John's, Canada.

Abstract

1 Introduction

The S&P 500, or basically the S&P, is a stock exchange index that evaluates the stock performance of 500 huge companies recorded on stock trades in the United States. It is one of the most usually followed equity indicators, and many believe it to be probably the best depictions of the U.S. stock exchange. The normal yearly aggregate return of the index, including profits, since beginning in 1926 has been 9.8%; be that as it may, there were more than a few years where the index dropped over 30%. The index has posted yearly increases 70% of the time (Wikipedia contributors (2020)) .

The index is one of the components in calculation of the Conference Board Leading Economic Index, used to predict the course of the economy. The index is related with numerous ticker images, including: ĜSPC, INX, and $SPX, dependent on marketplace or internet site. The index value is revised each 15 seconds, or 1,559 times per business day, with price upgrade circulated by Reuters (Duggan, Wayne (2019)).

2 Materials and Methods

2.1 Logistic Regression

Let attempt to comprehend logistic regression by thinking about a logistic model with given parameters, at that point perceiving how the coefficients can be assessed from data. Think about a model with two forecasters, x_1 and x_2, and single binary (Bernoulli) response variable Y, which we represent $p = P(Y = 1)$. We believe a linear correlation sandwiched between the predictor variables and the log-odds of the outcome that $Y = 1$. This linear correlation* can be written in the ensuing mathematical structure (where ℓ is the log-odds, b is the base of the logarithm, and β_i are parameters of the model):

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{1}$$

We can retrieve the log-odds:

$$\frac{p}{1-p} = b^{\beta_0+\beta_1 x_1+\beta_2 x_2}. \tag{2}$$

By straightforward algebraic manipulation, the likelihood that $Y = 1$ is

$$p = \frac{b^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{b^{\beta_0+\beta_1 x_1+\beta_2 x_2}+1} = \frac{1}{1+b^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}. \tag{3}$$

The equation above reveals that as soon as β_i are fixed, we can effortlessly calculate either the log-odds that $Y = 1$ for a given observation, or the likelihood that $Y = 1$ for

a given observation. The main usage of a logistic model is to be provided an observation (x_1, x_2), and calculate the likelihood p that $Y = 1$. In many applications, the base b of the logarithm is generally taken to be e. Nevertheless, in some cases it can be simpler to convey results by working in base 2, or base 10 (Wikipedia contributors (2020)).

2.2 Data Collection and Description

In this project, the dataset used was downloaded or collected online from `https://www.picostat.com/dataset/r-dataset-package-islr-weekly` by executing a pandas python code.

This dataset consists of percentage returns for the stock index over 1089 weeks for 21 years, from the beginning of 1990 to the end of 2010. For each data, we have the percentage returns for each of the first previous trading weeks, Lag1 through Lag5. We have also Volume (the number of shares traded on the previous day, in billions), Today (the percentage return on the date in question) and Direction (whether the market was Up or Down on this date).

2.3 Python-Software Packages Used

The dataset downloaded was then analyzed mainly using three open source python packages. These packages are Pingouin, Statsmodels and PyCM packages.

The Pingouin package was used to determine the structure, summary, correlation, normality of variables in the dataset and also analyze the dataset with the logistic regression model (Vallat, R. (2018)). The Statsmodel was then used to determine the Confusion matrix of the logistic regression output derived. Lastly, the PyCM package was used to check the performance or accuracy or predictions of these models (Haghighi et al., (2018)).

3 Analysis and Results

4 Conclusion

References

Wikipedia contributors.(2020). S&P 500 Index. In Wikipedia, The Free Encyclopedia. Online; accessed 28-July-2020, from `https://en.wikipedia.org/w/index.php?title=S%26P_500_Index&oldid=969280492`

Duggan, Wayne (2019). (June 13, 2019). "This Day In Market History: S&P 500 Quotes Delivered Every 15 Seconds", on June 13, 2019. Benzinga.

Wikipedia contributors.(2020). Logistic regression. In Wikipedia, The Free Encyclopedia Online; accessed 28-July-2020, from https://en.wikipedia.org/w/index.php?title=Logistic_regression&oldid=967311267

Vallat, R. (2018). Pingouin: statistics in Python. Journal of Open Source Software, 3(31), 1026, https://doi.org/10.21105/joss.01026

Haghighi et al., (2018). PyCM: Multiclass confusion matrix library in Python. Journal of Open Source Software, 3(25), 729. https://doi.org/10.21105/joss.00729