Predicting Stock Market Trends using Logistic Regressions and Feature Engineering

Rishi Kumar Tunuguntla
Dept. of Civil Engineering
Indian Institute of Technology Bombay
Mumbai, India
rishikumar@iitb.ac.in
Analytics General Championship Report

***Abstract*—This document contains all the information about the approach, research, references and everything related to the approach I have taken to solve the problem statement of predicting the dependent variable y using various techniques to increase the accuracy of the model.**

***Index Terms*—stock markets; technical analysis; statistics; machine learning; classifiers; logistic regression; naive bayes; feature engineering**

I. Introduction

We were given a dataset containing 10,000 points of OHLC (Opening-High-Low-Close) prices of a financial institution and had to predict a dependent variable y using the independent variables x_1, x_2, x_3, x_4 (Opening-High-Low-Close respectively). The meaning or the significance of the y was also not given, we had to predict the values of the binary dependent variable y and also interpret it's significance.

II. Methodology

A. Analyzing the problem statement

In the given data set there were 4 independent variables x_1, x_2, x_3, x_4 and one dependent variable y. By observing we can say that the variable y is a binary variable. The values that y taking is 0 and 1 only, according to which I came to an assumption that the variable y might be Boolean binary variable. Where 0 might be representing some negative statement and 1 might be representing a positive statement.

B. Choosing the Machine Learning Models

Basically to predict a binary variable we use classifiers. Here as the variable y is a binary Boolean variable. I have first started experimenting with different classifier models keeping x_1, x_2, x_3, x_4 as independent variables and y as a dependent variable. First I started testing the following models and the accuracy scores I got -

1) Logistic Regression with an accuracy score of 0.52.
2) KNN with an accuracy score of 0.51.
3) SVM with an accuracy score of 0.49.
4) Decision Trees with an accuracy score of 0.48.
5) Random Forests with an accuracy score of 0.50.

C. Instability of the models

After observing the accuracy scores of all the models, we can say that some of the model's performances were not too bad. But the main problem was varying the test and train sizes severely affected their performances. For example, the accuracy with Logistic Regression Model was 0.52 when the training data is 70% of the today data, but the accuracy drastically dropped down to 27%.

Then I realized to increase the stability and accuracy of model we can do Feature Engineering.

III. Significance of y

As the variable y is Boolean binary variable taking only 0 and 1 values. Those values can indicate the performance of the market, for example 1 indicates it is a good day (if the price of the financial institution is increased after x days. And 0 indicates it is a bad day if the price of the financial institution is decreases after x days, say. Going forward with this assumption of the significance of y, we will create new features on this basis. We can say that it predicts direction of market on the basis of % return of x previous days and volume of shares traded on previous days.

IV. Feature Engineering

Now we have to create new features to increase the accuracy of the model. So 10 new technical indicators will be used. The new technical indicators that are being used are -

A. One Day Returns

ODR on a stock is used to measure the day to day performance of stocks, it is the price of stocks at today's closure compared to the price of the same stock at yesterday's closure. Positive daily return means appreciation in stock price on daily comparison.

B. Momentum

Momentum is the speed or velocity of price changes in a stock, security, or tradable instrument. Momentum shows the rate of change in price movement over a period of time to help investors determine the strength of a trend.

C. Return of Investment

Return on investment (ROI) is a performance measure used to evaluate the efficiency or profitability of an investment or compare the efficiency of a number of different investments. ROI tries to directly measure the amount of return on a particular investment, relative to the investment's cost.

ROI is calculated by subtracting the initial value of the investment from the final value of the investment (which equals the net return), then dividing this new number (the net return) by the cost of the investment, then finally, multiplying it by 100.

We will calculate ROI for 10, 20 and 30 day periods.

D. Relative strength index

The relative strength index (RSI) is a momentum indicator used in technical analysis that measures the magnitude of recent price changes to evaluate overbought or oversold conditions in the price of a stock or other asset.

Formula for Calculating RSI: RSI = 100 – [100 / (1 + (Average of Upward Price Change / Average of Downward Price Change))].

We will calculate relative strength index for 10, 14, and 30 days periods.

E. Exponential Moving Average

The exponential moving average (EMA) is a technical chart indicator that tracks the price of an investment (like a stock or commodity) over time. The EMA is a type of weighted moving average (WMA) that gives more weighting or importance to recent price data.
Formula for Calculating EMA:
$EMA = EMA = Price(t)k + EMA(y)(1k)$
where
t = today
y = yesterday
N = number of days in EMA
k = 2÷(N+1)

F. Moving Average Convergence/Divergence

It is a trading indicator used in technical analysis of stock prices, created by Gerald Appel in the late 1970s. It is designed to reveal changes in the strength, direction, momentum, and duration of a trend in a stock's price.

Formula for Calculating MACD:
MACD = Moving Average of EMA(n) - EMA(m2) for each row

G. Stochastic RSI

The stochastic RSI (StochRSI) is a technical indicator used to measure the strength and weakness of the relative strength indicator (RSI) over a set period of time.

Formula for Calculating EMA:
SRSI = (RSI$_t oday - min(RSI_p ast_n))/(max(RSI_p ast_n) - min(RSI_p ast_n)$

Calculating the Stochastic RSI for 10, 14, and 30 day periods.

H. True Range

True range is a technical analysis volatility indicator originally developed by J. Welles Wilder, Jr. for commodities. The indicator does not provide an indication of price trend, simply the degree of price volatility. The average true range is an N-period smoothed moving average of the true range values.

Formula for Calculating True Range: TR = MAX(high[today] - close[yesterday]) - MIN(low[today] - close[yesterday])

I. Williams %R oscillator

It compares a stock's closing price to the high-low range over a specific period, typically 14 days or periods. Williams %R oscillates from 0 to-100; readings from 0 to -20 are considered overbought, while readings from -80 to -100 are considered oversold.

Formula for Calculating Williams %R oscillator: %R = (Highest High - Close)/(Highest High - Lowest Low) * -100

J. Commodity Channel Index

Definition: The Commodity Channel Index (CCI) is calculated by determining the difference between the mean price of a security and the average of the means over the period chosen. This difference is compared to the average difference over the time period.

Formula for Calculating CCI: CCI = (Typical Price - 20-period SMA of TP) / (.015 x Mean Deviation) — Typical Price (TP) = (High + Low + Close)/3 — Constant = 0.015.

V. Implementing

Now after creating all the new features we will use all the features including the initial given variables x_1, x_2, x_3, x_4. Removing the high(x_1) and low(x_2) variables to increase the accuracy. Cleaning the data frame created and reomoving all the unwanted features and missing spaces, the data frame will be ready to be used. Using Logistic Regression, Gaussian Naive Bayes, Decision Trees, Random Forests and SVM again on the new data set (included of newly developed features).

VI. Results

After experimenting on all the models the best accuracy was shown by Logistic regression and Naive Bayes. Below is the Classification Report for the Logistic Regression model on the new data set with new features included

	precision	recall	f1-score	support
0	0.55	0.51	0.53	1637
1	0.55	0.59	0.57	1653
accuracy			0.55	3290
macro avg	0.55	0.55	0.55	3290
weighted avg	0.55	0.55	0.55	3290

Below is the Classification Report for the Gaussian Naive Bayes model on the new data set with new features included.

	precision	recall	f1-score	support
0	0.51	0.78	0.62	1637
1	0.54	0.25	0.34	1653
accuracy			0.52	3290
macro avg	0.52	0.52	0.48	3290
weighted avg	0.52	0.52	0.48	3290

VII. Conclusion

For predicting the value of y we used a logistic regression model, because the dependent variable y only takes two values (which indicate the increase or decrease of the stock price of the market). For increasing the accuracy we had used feature engineering to create new technical indicators which served the purpose. We had obtained an accuracy of 0.55 with the Logistic Regression model and an accuracy of 0.52 with Gaussian Naive Bayes Model.

References

[1] Huiwen Wang, Wenyang Huang and Shanshan Wang “Forecasting open-high-low-close data contained in candlestick chart”

[2] Usha Ananthakumar, Ratul Sarkar ”Application of Logistic Regression in assessing Stock Performances”

[3] Andrew D. Mann and Denise Gorse, “A New Methodology to Exploit Predictive Power in (Open, High, Low, Close) Data”

[4] Ashwini Pathak, Sakshi Pathak,“Study of Machine learning Algorithms for Stock Market Prediction”

[5] Rebwar M. Nabi, Soran Ab. M. Saeed, Habibolah Bin Harron and Hamido Fujita, “Ultimate Prediction of Stock Market Price Movement”.