Statistic Methods Final Assignment

Ekaterina Zhiganova, ECUtbildning, Data Scientics-2022, Solna

The dataset I choose names NSE\_BANKING\_SECTOR.csv.

This dataset contains an information about 5 years of Banking Stocks from National Stock Exchange if India (NSE).

It insists of 15 variables and includes 41231 objects.

Variables are: DATE (daily datetime of stocks), SYMBOL (the particular stock symbol according NSE), SERIES(different series in which a particular stock could be traded), PREV CLOSE (previous day closing price), OPEN (day’s opening price), HIGH (day’s highest price), LOW (day’s lowest price), LAST (day’s last price), CLOSE (day’s closing price), VWAP (average price the stock security has traded throughout the day), VOLUME (volume of the stocks), TURNOVER (turnover per day), TRADES (amount of trades), DELIVERABLE VOLUME (deliverable volume of the stocks), %DELIVERBLE (percentage of delivers).

First three variables as DATE, SYMBOL and SERIES have a non-numeric characteristic, so I decided to remote them from the assignment. I used the subset() function for it.



After that the dataset includes 12 variables both numerical (9) and integer (3).

I research the dataset to clear understand what kind of information in.

PREV CLOSE - is a security's closing price on the preceding time of the one being referenced. It almost always refers to the prior day's final price of a security when the market officially closes for the day.

OPEN - is the price at which a security first trades when an exchange opens for the day. An opening price is not identical to the previous day's closing price. There are several day-trading strategies based on the opening price of a market or security.

HIGH - is the highest price at which the security has traded during the current trading day.

LOW - - is the lowest price at which the security has traded during the current trading day.

LAST - is the most recently reported trading price for a market or security.

CLOSE - **is the weighted average price**of the last 30 mins of trading.

VWAP (volume-weighted average price) - the average price a security has traded at throughout the day, based on both volume and price.

VOLUME - the total number of shares traded in a specified time frame. This would include every share that is bought and sold during the period of the time in review.

TURNOVER - the number of times the stock in a business has 'turned over’ or been replaced.

TRADES - Equity and equity-linked products available for trading in the cash market include stocks, IDRs, ETFs and units of closed-ended mutual fund schemes, as well as a segment devoted to the growth of the SME's listed on EMERGE.

DELIVERABLE VOLUME – the total volume that were marked for delivery on a certain date.

%DELIVERBLE – is percentage the deliverable volume of the total volume on a certain date.

Two variables I interested the most are CLOSE and HIGH which are numerical.

I want to calculate the dependence of one variable on another and understand how it can be used to predict the next day's highest price.

The question I will answer: Is where a connection between day’s closed price and the next day highest price and if yes, how strong it is and can it be used in the continuous forecasting.

Firstly, I need to find out about a correlation coefficient.

n=41231

Sum(x) = 12225754

Sum(y) = 12039992

Sum(y2) = 11966596601

Sum(x\*y) = 12122421445

Sum(x2) = 12282222043

r = (41231\*12122421445– 12225754\*12039992

sqrt((41231\*12282222043– 122257542)\*((41231\*11966596601) - 120399922) = 0.9998892

To confirm the correctness of calculation I choose a cor() function for it. Here is r= 0.9998.

I create a plot to look at the data in visual. I create a scatter plot with explanatory variable – close, and response variable – high. I see a positive correlation between them.

After that I need to create an absolute line. Intercept and slope should be found.

x is the vector of close price; y is the vector of high price.

Mean(x) = 292.0131

Mean(y) = 296.5185

Sum(x) = 12039992

Sum(y) = 12225754

Sum(xy) = 12122421445

Sum(x2) = 11966596601

b= (41231\*12122421445– 12039992\*12225754)/(41231\*11966596601– 120399922) = 1.01202

a = mean(y) – b\*mean(x) = 296.5185 - 1.01202\* 292.0131= 0.9954025

y= 0.9954025 + 1.01202\*x

I see a strong correlation around an absolute line. Most of the data placed near it.

After that I calculate a coefficient of determination:

r2 = 0.99982= 0.9996 (almost the 1)

This means that the model is quite accurate.

I also will know is its normal distribution or not.

I use pnorm() function for it and with mean = -0.812 and sd = 0.9998.

pnorm(1) - pnorm(-1) = 0.6826895 (about 68.3% of the data is place in a mean +/- 1 standard deviation)

pnorm(2) - pnorm(-2) = 0.9544997 (about 95.4% of the data is place in a mean +/- 2 standard deviations)

pnorm(3) - pnorm(-3) = 0.9973002 (about 99.7% of the data is place in a mean +/-3 standard deviations)

It’s a normal distribution.

Answer for the question is: there is a strong positive connection between day’s closed price and the next day highest price.

Second part of the answer as “can it be used in the continuous forecasting”:

We can use the 68% range of the normal distribution and find out that with 68% probability our highest price will be between -1.8118 and 0.1878 of the median.

We can also use quantiles:

For the close price is

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 25% | 30% | 35% | 40% | 45% | 50% | 55% | 60% | 65% | 70% | 75% |
| 37.100 | 47.850 | 59.575 | 72.100 | 84.750 | 101.850 | 124.900 | 154.700 | 185.225 | 244.950 | 305.675 |

For the high price is

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 25% | 30% | 35% | 40% | 45% | 50% | 55% | 60% | 65% | 70% | 75% |
| 37.975 | 48.900 | 60.850 | 73.750 | 86.600 | 103.800 | 127.300 | 157.400 | 188.700 | 249.950 | 311.400 |

Difference between these two variables is 0.875 in the lower quartile, 1.995 in the middle and 5.725 in the upper quartile.

It can be used to continuous forecasting the highest price depended of the closed price.