DOW JONES
CASE STUDY

31-OCT-2021

KRISCHELLE JOYNER | KATHY KEEVAN
DAYANIRA MENDOZA | CHARLES REYES

Executive Summary

In the following case study, “Dow Jones”, we considered four different modeling methods to understand the forecasting of weekly stock returns, while comparing the stock risk with the market. The four different methods were linear regression, decision tree, support vector machine, and capitol asset pricing model.

After going through the results and comparing each model, we can see that using linear regression is our best model that will be later explained in the report. As for decision trees model, there are no assumptions about the distribution of the data that are not influenced by outliers. The disadvantages to this can be low prediction accuracy rates. Support vector machines had a disadvantage because choosing the appropriate kernel function can become complex.

We then determine the risk level of the stocks by performing a capital asset pricing model. The beta of each stock represented how the risk was in comparison to the market rate. A beta greater than one indicated that the stock was riskier than the market, while lower than one indicated that the stock was less risky than the market.

Background

Dow Jones & Company, which was founded in 1882 by Charles Dow, Edward Jones, and Charles Bergstresser, has introduced and owns the Dow Jones Industrial Average, an index that tracks 30 major, publicly-held companies.

In its early days, the DJIA began with 12 companies mostly active in the industrial sector. As time went on, it expanded to include 30 companies, including those in railroads, cotton, gas, sugar, tobacco, and oil. The performance of industrial companies is often interpreted as a measure of the economy as a whole, making the DJIA a key metric. Although the economy's health is now tied to many other sectors, the DJIA is still revered as a key indicator. We will use DJIA data to develop a forecasting model of future stock returns.

Purpose of Study

Investors are interested in knowing how well the Dow Jones Index (DJIA) will perform in the following week. We can make future predictions by using the present Yahoo Finance data for DJIA. Because we cannot use data from any given period for future periods, we can use data from the present for future months.

The analysis will primarily focus on:

What will the future stock returns be for the Dow Jones Index?

The Dow Jones data presents several indicators of a stock, such as close or volume, on a weekly basis and can be used to predict future stock returns.

Which companies have the best performing stocks?

If you know which companies are performing the best in an index, you could explore other indexes that also include those companies if you're not getting your desired returns.

Review of Related Literature

Predict Stock Market Behavior: Role of Machine Learning Algorithms
Gurav and Sidnal (2018)

Gurav and Sidnal explored seventeen types of models and noted their recommended application for stock forecasting, and the pros and cons of the models. The algorithms that were most appropriate for either risk assessment or stock price were:

Model type	Applications	Advantages	Disadvantages
Linear regression	sales estimating, risk assessment	ease of tuning; speed	
Logistic regression	revenue prediction	ease of tuning; reduces model noise	potential for overfitting
K-means clustering	estimate the direction of price movement	tighter clusters; speed	
Support vector machine	stock market prediction	best classification performance; no overfitting	
Random forest	generally good at prediction	speed; no overfitting	large number of trees complicate analysis

Regression trees	stock market prediction	adaptive	

One challenge the authors noted was that a prediction of +1% or -1% can mean some probability of a wide range of movement of around +/- 20%, and so perhaps a better approach might be predicting the probability distribution of the stock price rather than a single prediction.[1]

A Bayesian regularized artificial neural network for stock market forecasting
Ticknor (2013)

This paper described a Bayesian neural network to predict the closing stock price the next day. Neural networks were not one of the recommended algorithms described in Gurav et al. Yet Ticknor reports a > 98% fit for future stock prices for the two stocks studied, Microsoft and Goldman Sachs. The Bayes model is a back propagation neural network where the weights between layers are adjusted during learning to reduce the error function (MSE for example). Ticknor's model had three layers, with the model adjusting the number of neurons in the hidden layer. The hidden layer neurons were adjusted between two and ten, with the optimized model having five neurons. The train-test split was 80-20. Stock data was normalized to fall within a range of -1 to 1 to guard against bias towards larger values. The training was limited to 1,000 epochs. The train and test values of the predicted Microsoft stock price are plotted against the actual price below.[2]

A hybrid ARIMA and support vector machines model in stock price forecasting
Pai and Lin (2005)

This study was interesting in that it sought to augment the ARIMA time series forecasting model with support vector machines (SVMs) to address nonlinearity in the data. The authors note that "Forecasting stock prices has been regarded as one of the most challenging applications of modern time series forecasting." In recent years, as neural network models became more feasible and matured, researchers have used various machine learning methods to improve stock price forecasting, and have found that a combination of different forecasting models can improve results compared to individual models. In particular, they found that using a hybrid model with ARIMA to capture the linear element of stock price movement and SVM to capture the non-linear element yielded better results than ARIMA or SVM alone.[3]

Methodology and Assumptions

To find the best model, we will be training our models using four different techniques.

1. Linear Regression
2. Decision Tree
3. Support Vector Machine (SVM)
4. Capital Asset Pricing Model (CAPM)

Linear Regression

Linear regressions should be the first model type that is tried. If the model performs well, it is simple to interpret and explain. Our review of research on stock price models revealed that linear regressions are suitable for this purpose. We will run a linear regression for each stock.

Decision Tree

Decision trees are also easy to understand; they are pretty good at prediction and run efficiently. The two types of decision trees are classification and regression; because the predicted values are continuous, we will run regression trees.

Support Vector Machine (SVM)

SVMs are a great model for stock prediction, and don't tend to overfit the data. We will run a SVM for each stock. However, any gains in accuracy will cost us in interpretability because these models are difficult to break down and explain in comparison to Linear Regression and Decision Tree.

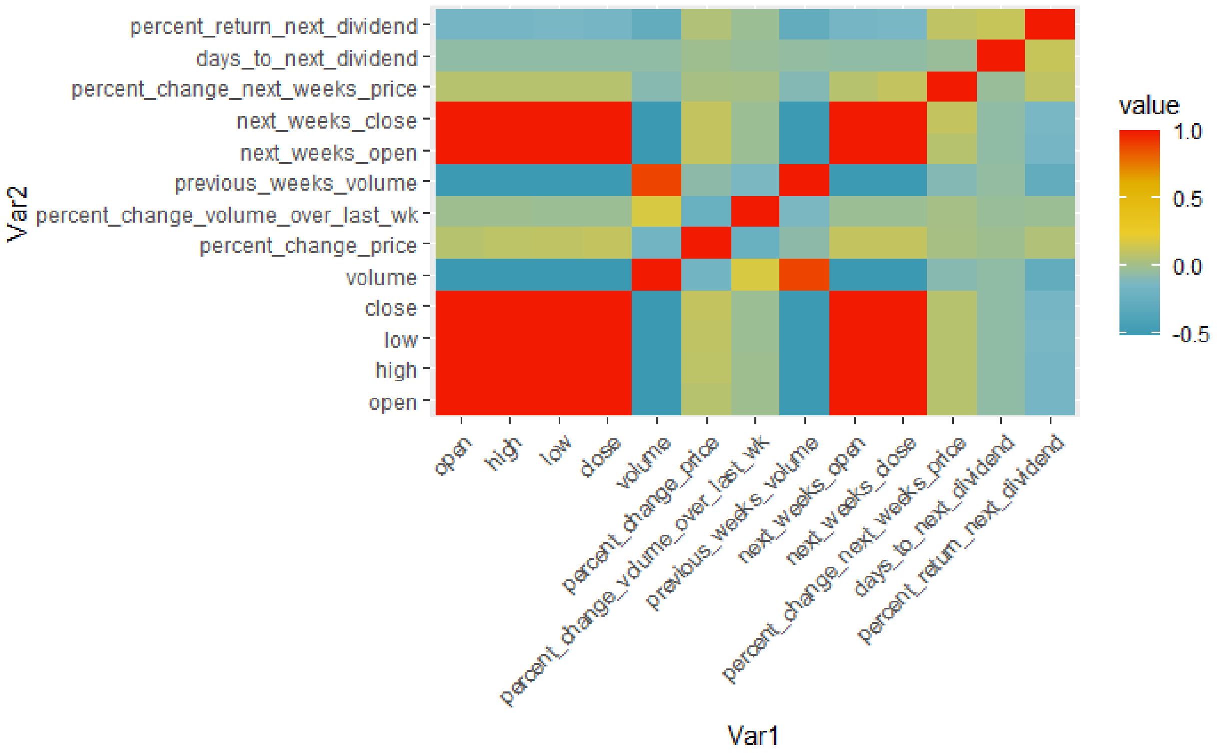
Capital Asset Pricing Model (CAPM)

We will also calculate each stock's performance against the Dow Jones Index for the same time period, using the Capital Asset Pricing Model.

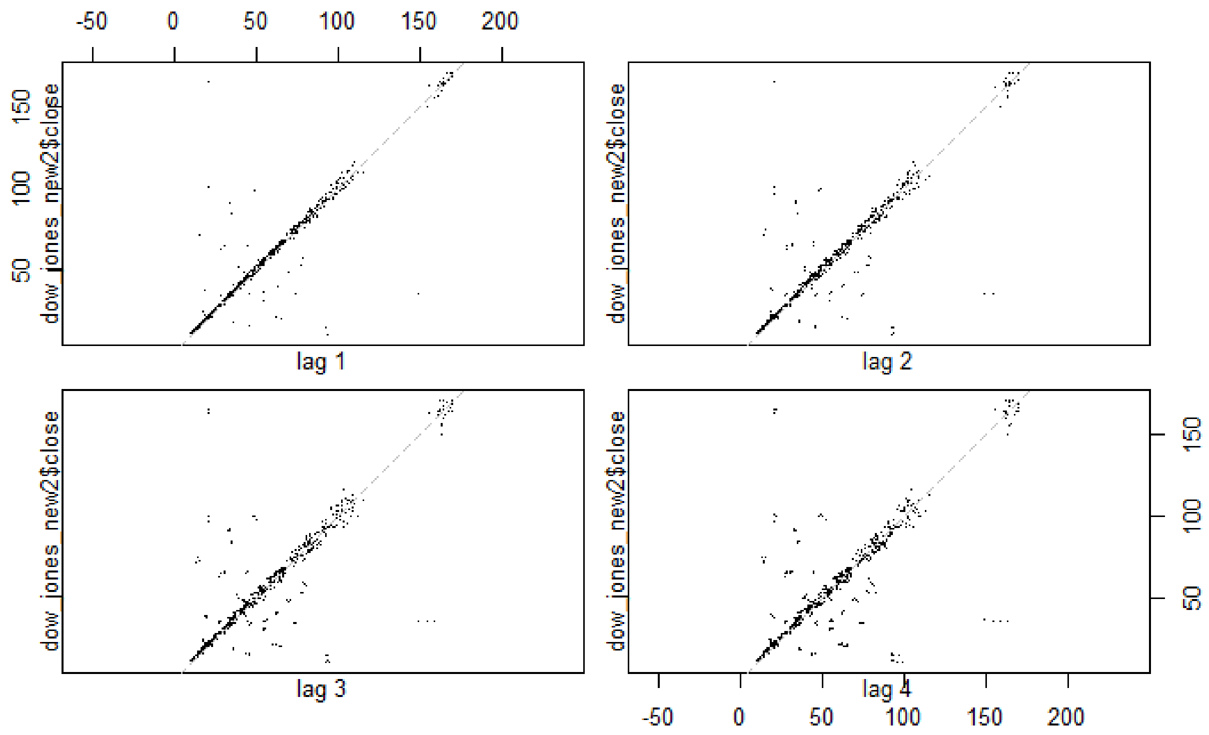
Data

The data set used for our modeling and predictions was sourced from Yahoo Finance. The data set contains data for 30 stocks from the Dow Jones Index for the periods of January 2011 to June 2011. The total data set has 750 observations and 16 variables (see appendix for variable descriptions). The data set was fairly clean and only had missing values in the "percent_change_volume_over_last_week" and "previous_weeks_volume" variables for the first observation of each stock because the previous week's data was not available. Looking at the data types for each variable, we did have to convert the formats for portions of the data that was imported as character data types. For instance, currency values were imported as character data types because they contained dollar symbols; they were converted to numeric to allow for better analysis. Date values were also not recognized properly. Although not required, stock data type was converted from character to factor.

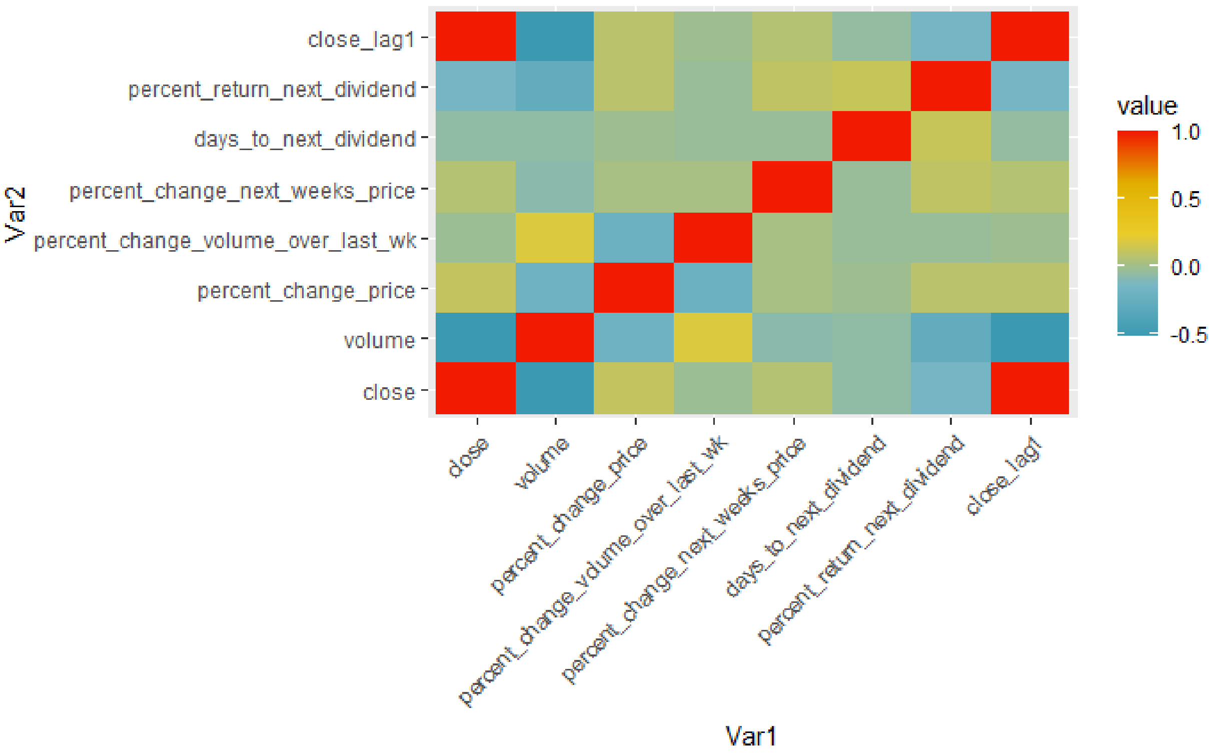
For our modeling, we also created a heat map matrix to help us identify any correlation issues among our variables.



First, we removed the “next_weeks_open”, “next_weeks_close”, “previous_weeks_volume”, and “percent_change_next_weeks_price” variables because they contained future values which typically aren’t known at the time of each prediction and shouldn’t be used to train our models. Next, we also identified high correlation with the “open”, “high”, “low”, and “close” variables (see heat map below), but we initially only removed the “high”, “low” and “open” variables because the “close” variable is part of our primary interest. Lastly, we also looked at creating a lag variable for “close” for our analysis. Based on the chart below, we have the highest correlation at lag 1.



Once we removed the variables mentioned above and created our lag variable for close, we can see most of the correlation issues we identified earlier have been solved.



The data set was split into two, one for training and one for testing. The train data set contains the Q1 2020 (January - March) data which has 360 observations, or 12 weeks of data, which we will use to train our models. The test data set contains the Q2 2020 (April - June) data which has 390 observations, or 13 weeks of data, which we will use to test the accuracy of our models. To help us evaluate the stock risk, we used the Dow Jones Industrial Average for the same period as the stocks from the Dow Jones Index (January 2011 - June 2011).

Findings

Accuracy

Accuracy proved to be a good indicator of stock returns. Using our final linear regression model, we compared the accuracy of our prediction for the *percent_change_next_weeks_price* variable with the actual stock value and found that Microsoft would be the best stock. The MSFT stock had an accuracy of 93.2%, and is therefore predicted to produce the greatest rate of return in the following week among all the stocks in the Dow Jones index.

Risk

A stock's risk is evaluated by comparing its price fluctuations with those of the market. If the fluctuations are correlated with the market, the beta value will be 1.00. When beta is greater than 1.00, the stock is more volatile than the market and will add risk to a portfolio. Likewise, if beta is less than 1.00, the stock is less volatile and therefore less risky. The stocks evaluated in this project ranged from a low of 0.8 (WMT) to a high of 1.27 (INTC). (Some stocks yielded a negative beta value; negative betas are possible but unlikely, and are generally uncommon investments like gold.)

Recommendations

Advantages and Disadvantages

Linear regression proved to be a good model as it was easy to implement. Comparing the stocks for the greatest rate of return for the following week was very easy to interpret between stocks, and made the MSFT stock easily standout as the best option.

Although decision trees do not require a lot of effort for data preprocessing and they can be easy to interpret, we found our decision tree model to result in several trees with only two terminal nodes. The accuracy was between 50% and 70% for most trees, which wasn't as good as our linear regression model.

While the Support Vector Machine model produced high accuracy percentages, we had difficulty predicting which stocks would yield the greatest rate of return because the model was difficult to interpret. One disadvantage of using SVM is choosing an appropriate Kernel function (to handle the non-linear data), which can be complex.

Conclusion

In conclusion, we recommend using a full linear regression model to determine which stock is well worth investing. The linear regression model achieved a high accuracy and we were able to explain our results easily, which makes it easier for an investor to make decisions. The simplest model is usually the best model.

Bibliography

Gurav, U., Sidnal, N., 2018. Predict Stock Market Behavior: Role of Machine Learning Algorithms, in: Bhalla, S., Bhateja, V., Chandavale, A.A., Hiwale, A.S., Satapathy, S.C. (Eds.), Intelligent Computing and Information and Communication, Advances in Intelligent Systems and Computing. Springer, Singapore, pp. 383–394. https://doi.org/10.1007/978-981-10-7245-1_38

Pai, P.-F., Lin, C.-S., 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. Omega 33, 497–505. https://doi.org/10.1016/j.omega.2004.07.024

Ticknor, J.L., 2013. A Bayesian regularized artificial neural network for stock market forecasting. Expert Systems with Applications 40, 5501–5506. https://doi.org/10.1016/j.eswa.2013.04.013

Appendix

Attribute Information:

- **quarter**: the yearly quarter (1 = Jan-Mar; 2 = Apr=Jun).
- **stock**: the stock symbol.
- **date**: the last business day of the work (this is typically a Friday)
- **open**: the price of the stock at the beginning of the week
- **high**: the highest price of the stock during the week
- **low**: the lowest price of the stock during the week
- **close**: the price of the stock at the end of the week
- **volume**: the number of shares of stock that traded hands in the week
- **percent_change_price**: the percentage change in price throughout the week
- **percent_change_volume_over_last_week**: the percentage change in the number of shares of stock that traded hands for this week compared to the previous week
- **previous_weeks_volume**: the number of shares of stock that traded hands in the previous week
- **next_weeks_open**: the opening price of the stock in the following week
- **next_weeks_close**: the closing price of the stock in the following week
- **percent_change_next_weeks_price**: the percentage change in price of the stock in the following week
- **days_to_next_dividend**: the number of days until the next dividend
- **percent_return_next_dividend**: the percentage of return on the next dividend

Endnotes

[1] (Gurav and Sidnal, 2018)
[2] (Ticknor, 2013)
[3] (Pai and Lin, 2005)