

Case 3 - Dow Jones

Josh Gardner, Alex Martinez, Cameron Playle, and Guillermo Gallardo

Executive Summary

Brief introduction of problem. Summarizes key findings. Summarizes insights behind key findings.

Problem

Our problem is to use historical weekly return data for 30 stocks in the Dow Jones Index to predict which stock will produce the greatest rate of return in the following week. From an application standpoint, this involves analyzing stock price trends and using past performance data to inform future investment decisions. The goal is to build predictive models using this historical data to maximize future returns.

To achieve this, we will conduct separate analyses for each stock, calculating the average predictive accuracy across all stocks. This approach will allow us to identify the model that provides the highest rate of return predictions, guiding us in selecting the most effective forecasting method.

We will utilize the variables in our dataset to build several predictive models to forecast future stock returns. After constructing these models, we will evaluate their performance to determine which one offers the most accurate predictions.

We will assess the risk of each stock using the S&P 500 as our benchmark. By calculating the beta for each stock, we can gain insights into the level of risk associated with each investment. Using these risk assessments alongside stock return predictions, we seek to identify the best investment recommendations based on both returns and risk. This analysis can also help guide investors in making decisions aligned with their individual risk tolerance.

Lit. Review

Discusses and cites existing works in the theoretical and application realm. Article: <http://www.warse.org/IJATCSE/static/pdf/file/ijatcse135952020.pdf>

This article explores several advanced machine learning models for stock price forecasting—approaches we haven't yet used in our project but that offer valuable insights into different paths toward achieving similar goals. In our MSDA program, we're currently learning about CNNs and how they work to detect cyberbullying by processing images, a completely different application than stock forecasting, but it's interesting to see the versatility of these models. The authors review models like Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), each with unique strengths in time series analysis. LSTM, for instance, overcomes traditional limitations in recurrent neural networks by retaining long-term dependencies, making it effective for capturing stock trends. An extension, Bidirectional LSTM (BLSTM), improves accuracy further by processing data in both forward and backward directions, helpful for understanding nuanced market shifts (Zonathan et al., 2020)

In this study, CNN-based models were adapted for stock data by transforming one-dimensional time series into two-dimensional representations, allowing CNN to detect complex patterns. The CNNPred

model, for example, demonstrated high accuracy on major indices like the S&P 500, illustrating how CNN can be effectively applied to stock data.

The findings on hybrid CNN-LSTM models are particularly intriguing. By combining CNN's feature extraction with LSTM's temporal modeling, the CNN-LSTM model achieved the lowest RMSE score, showing strength in forecasting stock prices under volatile conditions. While our project hasn't yet explored such hybrid models, this study shows there are multiple, innovative ways we could leverage these models as we refine our forecasting objectives

Methods

Decision Trees

SVR

LM

Capital Asset Pricing Model

We applied the Capital Asset Pricing Model (CAPM) CORRECT THIS TO UNDERSTAND RISK BASED ON THIS BETA to estimate the expected return of each stock based on its beta, which indicates its volatility relative to the market. CAPM assumes that markets are efficient and that there is a direct, linear relationship between risk and expected return. However, the model has certain limitations. It depends on historical data, using information from this case study and S&P 500 data from Yahoo Finance, which may not accurately reflect future volatility. Additionally, CAPM considers only market risk, overlooking other factors such as company specific risks and economic influences, which can also impact returns.

To calculate the results in the below image, we filtered the data by stock so we could apply the `Delt()` function to calculate the returns using the closing price for each specific week. Once we calculated the weekly returns and omitted any NA, we ran the linear model with

ADD STUFF ABOUT STOCKS WITH BETA OF 1.00 INDICATED ITS PRICE IS CORRELATED WITH THE MARKET AND HIGHLIGHTS WHICH STOCKS. MENTIONED THAT THIS STOCKS HAVE SYSTEMIC RISK BUT THE BETA CALCULATION CANT DETECT ANY UNSYSTEMIC RISK

Data

Our dataset is fairly clean overall, with only 60 total NA values across two columns: 30 in *percent_change_volume_over_last_wk* and 30 in *previous_weeks_volume*. These NAs occur because they correspond to the first week of data, where there is no previous week to calculate the volume change so we decided to omit the NAs. The dataset contains 750 observations and 16 variables, which we will split by quarters: quarter one will be used for training, and quarter two for testing.

For our variables, there is a mix of variable types. Variables like *volume*, *percent_change_price*, *percent_change_volume_over_last_wk*, and *days_to_next_dividend* are numerical. However, some variables, such as *open*, *high*, *low*, *close*, *next_weeks_open*, and *next_weeks_close*, are stored as characters due to the presence of dollar signs. These will need to be transformed into numeric values for accurate analysis.

Results

ADD TABLE BY STOCK HE MENTIONED DURING CLASS

Conclusions

The objective of this study was to predict stock returns and compare risk to maximize future returns. We evaluated multiple predictive models to determine the best fit, ultimately selecting the linear regression (lm) model. This choice was based on the Root Mean Squared Error (RMSE) scores, where the lm model performed marginally better than alternatives, showing a single-digit improvement on a 10-digit scale. Risk and Reward Analysis Using CAPM: We applied the Capital Asset Pricing Model (CAPM) to assess the risk (Beta) associated with each stock. Using the model's predictions for the next week's closing price, we calculated the expected revenue. By plotting revenue against Beta, we visualized the risk-reward balance each stock offered. Key Insights and Recommendations: The risk-reward plot highlighted three stocks with favorable profiles: KRFT, HPQ, and KO. Additionally, IBM and TRV showed potential but were rated as secondary recommendations. For a stable portfolio, we recommend including KRFT as a primary option. If additional funds are available, TRV and IBM would also be suitable additions. Special Recommendation – KO: While KO stood out, it did so due to its notable negative return. Despite a near-zero Beta, we do not recommend including KO in a portfolio for long positions. However, its low Beta and performance profile make KO a viable candidate for short selling. Conclusion: By combining predictive modeling and risk assessment, we identified key stocks to consider for portfolio inclusion and developed insights into both long and short positions based on risk and return trade-offs.

