Case 3 - Dow Jones

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

# START REPORT BELOW

# Executive Summary

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

# Problem

*Clear description of the problem, from an application and theoretical point of view. Outlines the report.*

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

*Discusses types of variables, sample size, and sampling techniques (if any). Discusses the model(s) and its assumptions and limitations.*

## Decision Trees

## SVR

## LM

## Capital Assest Pricing Model

We applied the Capital Asset Pricing Model (CAPM) CORRECT THIS TO UNDERSTAND RISK BASED ON THIS BETAto 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 teh Delt() function to calculate the returns using the closing price for each specific week. Once we calculated the weekly returns and omited 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

MAYBE CREATE A PLOT TO SHOW Y MONTHLY SECURITY EXCESS RETURN AND MONTHLY MARKET EXCESS RETURN

|  |
| --- |
|  |

# Data

*Discusses how data was handled, i.e. cleaned and preprocessed. Discusses distributions, correlations, etc.*

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 are is mix of variables 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

*Presents and discusses the results from model(s). Discusses relationships between covariates and response, if possible, and provides deep insights behind relationships in the context of the application.*

ADD TABLE BY STOCK HE MENTIONED DURING CLASS

# Conclusions

*Concludes with a summary of the aim and results. Discusses alternative methods that can be used.*

Based on the results (ADD TABLE VIEW BY STOCK WITH RESULTS?) we see that the model that outperforms the others is ADD STUFF. ADD STUFF.