Case 3 - Dow Jones

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

# Executive Summary

• Bottom Line: The linear regression model (LM) provided the most accurate predictions for weekly returns among Dow Jones stocks, enabling targeted investment recommendations based on risk-return analysis.

• Invest in KRFT and HPQ: Both stocks showed favorable risk-return profiles and are suitable for portfolio stability.

• Consider KO for Short-Selling: KO’s negative return, paired with a near-zero beta, presents a strong short-selling opportunity for risk-tolerant investors.

• Predictive models were tested to forecast weekly returns, with LM showing the lowest Root Mean Squared Error (RMSE).

• Capital Asset Pricing Model (CAPM) used beta values to evaluate risk, benchmarking against the S&P 500.

# 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

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

## LM

Since our problem involved regression, we opted for a linear model, which is well-suited for understanding straightforward relationships between features and the target variable. Linear models offer clear interpretability, making it easier to understand the impact of each feature on the outcome.

To improve model accuracy and reduce complexity, we focused on feature selection by identifying and removing highly collinear variables. Collinearity can inflate coefficient estimates and lead to unstable results, so we examined variance inflation factors (VIF) and removed predictors with high VIF values to ensure a more stable model.

## Regression Decision Trees

A Regression Decision tree was used to predict the percent change in next week price using all the predictor variables except ‘quarter’, ‘date’, ‘next weeks open’, and ‘next weeks close’. We use built the tree using Q1 data and tested the model on Q2 data. The table below shows the RMSE provided for each stock to determine how well our tree predicted the stocks percent change of the upcoming week. With this decision tree model, we can support our decision with what stock we believe will have the best return as it allows us to be more confident in our stock decisions. From the output below, we see the top 5 stocks with the smallest RMSE are, KO, WMT, TRV, T, and IBM.

## SVR

Due to the regression nature of this problem, we needed to use an SVM variant that handles regression tasks. SVMs are flexible because they offer different kernel types to fit different data relationships. For this project, we chose the Radial Basis Function (RBF) kernel, which is well-suited for non-linear relationships and works well with various data distributions. During model building, we tested multiple combinations of cost and gamma to find the best-performing model. We also centered and scaled the data, since SVMs are sensitive to variable scales, which helped improve the model’s performance.

## Capital Asset Pricing Model

We used the Capital Asset Pricing Model (CAPM) to understand the risk of each stock by analyzing its beta, which measures 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 relies on historical data, from this case study and S&P 500 data from Yahoo Finance (where we extracted the data), which may not fully capture future volatility. Additionally, CAPM only accounts for market risk, overlooking other factors such as company-specific risks and economic influences that 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 omited any NA, we ran the linear model to get the beta for each stock.

Our beta calculations (see table in appendix) reveal a range of values for various stocks relative to the S&P 500. Stocks like **BAC (0.9536)** and **HPQ (0.8437)** have higher positive betas, indicating they closely follow market trends. In contrast, **PG (-0.0008)** and **XOM (-0.0038)** have near-zero betas, suggesting minimal correlation with the market, which can provide portfolio stability. Notably, **INTC (-0.9063)** shows a strongly negative beta, indicating an inverse relationship to the market and potentially serving as a hedge during downturns. This range of risk and correlation profiles will be used alongside predicted returns to recommend investment decisions that align with expected returns and our client's risk tolerance.

# Data

Our dataset is fairly intact, 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. While the data was intact we needed to clean the data using lubridate package to convert a text-based date and removed '$' symbols from numerical variables to ensure data accuracy. During preprocessing we needed to scale the data at points to meet assumptions of the models we used.

# 

# Results

ADD RMSE TABLE FOR EACH MODEL HERE AND AN EXPLANATION ON WHICH IS BETTER, AND OUR DECISION AND HOW WE USED THE MODEL?

|  |  |  |  |
| --- | --- | --- | --- |
| **Stock** | **LM RMSE** | **SVR RMSE** | **DT RMSE** |
| AA | 126671324 | 126671324 | 20.05021 |
| AXP | 32653797 | 32653800 | 17.755682 |
| BA | 22107483 | 22107487 | 9.233646 |
| BAC | 637938743 | 637938739 | 11.861849 |
| CAT | 37163364 | 37163365 | 22.258886 |
| CSCO | 382385019 | 382385017 | 17.437448 |
| CVX | 36898553 | 36898556 | 12.606989 |
| DD | 26760786 | 26760786 | 12.09519 |
| DIS | 46218461 | 46218460 | 15.726054 |
| GE | 231874886 | 231874886 | 13.213855 |
| HD | 51032386 | 51032386 | 7.480315 |
| HPQ | 102925210 | 102925207 | 19.219627 |
| IBM | 23676034 | 23676055 | 4.836956 |
| INTC | 330452414 | 330452413 | 17.140916 |
| JNJ | 58363850 | 58363852 | 8.691014 |
| JPM | 146424585 | 146424583 | 9.13623 |
| KO | 35913555 | 35913557 | 3.162122 |
| KRFT | 45366153 | 45366156 | 7.121616 |
| MCD | 24965634 | 24965638 | 6.878623 |
| MMM | 15020575 | 15020577 | 8.189247 |
| MRK | 65631688 | 65631691 | 9.17664 |
| MSFT | 306000553 | 306000552 | 10.32122 |
| PFE | 207245099 | 207245100 | 8.225146 |
| PG | 46194506 | 46194511 | 5.726039 |
| T | 149761089 | 149761092 | 3.818744 |
| TRV | 15458994 | 15458999 | 3.59698 |
| UTX | 18958828 | 18958836 | 11.641442 |
| VZ | 65737755 | 65737756 | 5.210104 |
| WMT | 51055448 | 51055449 | 3.560991 |
| XOM | 88060818 | 88060821 | 12.730046 |
| **Average RMSE** | **114297253** | **114297255** | **10.53013038** |

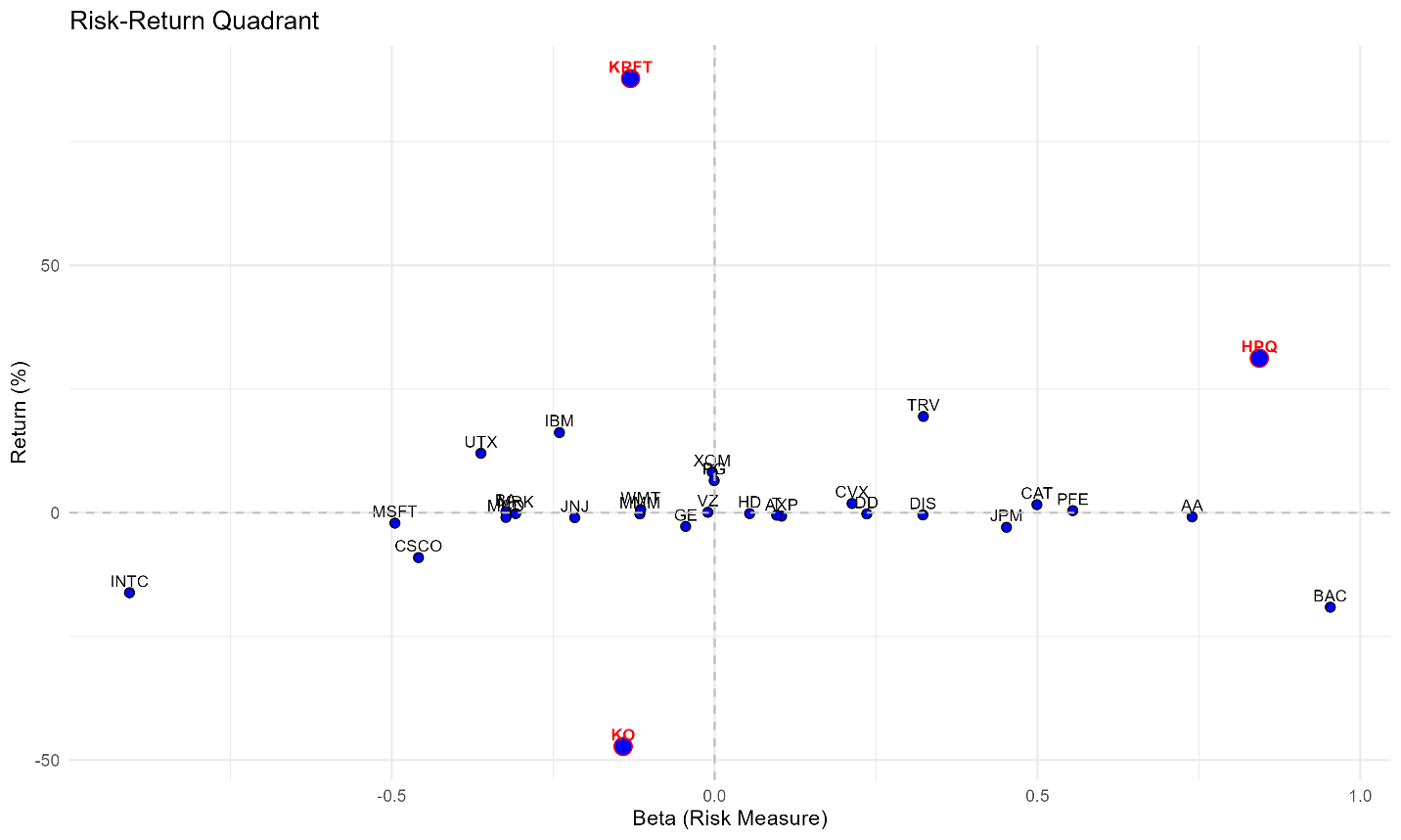
# Conclusions

The objective 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.

# Appendix