Dow Jones Index Case Study

Collin Real, Leonel Salazar, Seth Harris, Joaquin Ramirez

# I. Executive Summary

In this case study, we explored three types of models to predict an upcoming week’s stock price changes in Q2 for the Dow Jones Index using historical data: linear regression, gradient boosting machine, and support vector regression (SVR) models. We split our dataset into training (Q1) and testing (Q2) data to build each model and evaluate their predictive abilities. To quantify model accuracy, we calculated the Root Mean Squared Error (RMSE) for each model:

* **Linear Regression**: 1.03999924761457
* **Gradient Boosting Machine**: 1.06098845849507
* **Support Vector Regression**: 0.998923259065558

The SVR model produced the lowest RMSE among our three models, suggesting that it’s the most accurate model for predicting price changes.

The Capital Asset Pricing Model (CAPM) is a pivotal tool in finance, providing insights into the relationship between risk and return for individual assets. By analyzing the sensitivity of Dow Jones Index stocks to market fluctuations, CAPM aids in understanding systematic risk and expected returns. This report incorporates key CAPM metrics, scatterplot analyses, and portfolio strategies to paint a holistic picture of stock performance and market alignment. These findings are invaluable for constructing portfolios tailored to varying risk appetites.

CAPM calculations focus on: 1. Beta: Measuring a stock’s sensitivity to market movements. 2. Alpha: Indicating the excess return a stock provides above the market’s performance. 3. R-Squared: Quantifying how well the stock’s returns align with the market’s returns. Scatterplots of excess stock returns versus market returns were generated for visual analysis. Regression lines overlay the plots to highlight trends, providing a graphical interpretation of CAPM metrics.

# II. The Problem

Our main challenge is developing a highly accurate model for predicting future stock price movements, so investors can optimize the return on investment for their portfolios. Highly accurate models facilitate the creation of more effective investment strategies that identify stocks with greater potential of producing higher returns as well as the riskiness of buying or holding these assets. CAPM analysis provides information about the volatility and risk associated with different stocks, allowing investors to modify their portfolios and mitigate or increase risk depending on their tolerance. Predictive models create the opportunity for investors to assess their portfolios perpetually and make adjustments to their strategies based on stock performance forecasts and historical trends.

# III. Review of Related Literature

This analysis builds on previous work, including the research by Brown, Pelosi, and Dirska (2013), which highlights the application of advanced models for forecasting financial markets and assessing risk-return tradeoffs.

# IV. Methodology

The primary objective of our case study was to build a highly accurate model to predict stock prices in the upcoming weeks to optimize the return on investment of stock portfolios.

## Predictive Models

## Linear Regression

* Dependent variable is **continuous**.
* Estimates the relationship between predictors and the dependent variable by fitting a linear equation.
* Easy implementation and interpretation.
* Model assumptions:
  + **Linear relationship** between the predictors and the dependent variable.
  + Residual errors are **independent** (no autocorrelation) of each other.
  + **Multicollinearity** - independent variables are not perfectly correlated.
  + **Homoscedasticity** - error variance is **constant** across all levels of the independent variables.
  + Residuals are **normally distributed**.

## Gradient Boosting Machine

* Model assumptions:
  + **Additive model of weak learners** - the dependent variable can be predicted through an additive combination of weak independent variables.
  + **Independent observations & identically distributed data**
  + **Multicollinearity** - independent variables are not perfectly correlated.
  + **Large sample size**
  + **Homogeneity of data**
  + **Hyperparamenter tuning** - balance model complexity and accuracy.

## Support Vector Machine (SVM)

* Classification problems for both linear and non-linear outcomes.
* Identifies the hyperplane with maximum separation between the two classes (i.e., purchase vs. non-purchase) in the feature space.
* Capture non-linear relationships via multiple kernels:
  + linear
  + radial basis function (RBF)
  + polynomial
  + sigmoid
* Model flexibility and effectively identifies complex relationships for non-linear problems.
* Less interpretable and difficulty selecting the correct kernel/tuning parameters.
* Model assumptions:
  + **Linearly separable classes** - Linear SVM assumption needed to find a hyperplane separating the classes. RBF or polynomial non-linear kernels are used if the assumption is not satisfied.
  + **Independent observations**
  + **Multicollinearity** - independent variables are not perfectly correlated.
  + **Feature scaling** - accounts for distances between data points. Unscaled data might affect model results.
  + **Kernels** - assumes the correct kernel is appropriately chosen.

## Data Preprocessing & Exploration

The first step of our analysis was importing our data sources and preprocessing. We applied a Q1/Q2 training and test split which trains stock price data from Q1 to predict percent\_change\_next\_weeks\_price in Q2. The stock variable was converted to the factor data type since the information is categorical. Variables percent\_change\_price, percent\_change\_volume\_over\_last\_wk, days\_to\_next\_dividend, percent\_return\_next\_dividend, volume, high, low, open, and close were converted to the numeric data type. After this conversion, the data was scaled to avoid one independent variable from having greater influence over the others in our models. After these changes, we visualized the distribution of numeric variables using histogram plots: . Next, we calculated and visualized the correlation matrix using the correlation plot to identify and remove highly correlated variables from our model.

## Model Selection

Linear regression, gradient boosting machine, and support vector regressions models were explored to predict future stock prices. The linear regression model is used to understand the relationship between each predictor and the target variable. It also establishes a baseline to compare with more complex models. The gradient boosting machine and support vector regression (SVR) models are most advanced models that were explored to produce more accurate models with less interpretablity.

# V. Data

The case report used historical stock price data. The training data consisted of Q1 stock information and the testing data consisted of Q2. The dependent variable of the analysis was percent\_change\_next\_weeks\_price.

# VI. Findings (Results)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_