Jay Kwon | Metis Data Science and Engineering (flex program) | Module 2 - Regression (October, 2021)

**Predicting 3M Corporation’s Weekly Stock Return using Linear Regression**

**Abstract**

A linear regression model was fitted to predict the weekly returns for 3M corporation’s stock. The goals of this project were to 1) try to create a model that most accurately predicts weekly returns and 2) gain insights on how to predict stock returns, including which features are most relevant. None of the features examined had a high correlation to returns. Average volume had the strongest correlation with returns with a correlation of -0.18. The final model had the following features with significant p-values < 0.01 : average volume, -1 week historical return, -2 to -5 week geometric average return. Unfortunately, a 5-fold cross validation resulted in a negative average R-squared value of -0.055, meaning a simple average of the weekly returns (target) would be more accurate in predicting weekly returns than our linear regression model. The wide ranged in R-squared values calculated in the cross validation suggests that more of the same type of data is necessary. In future studies, new features need to be explored such as company specific financial statement data (e.g. historic earnings per share).

**Design**

Weekly stock return is a continuous, numeric metric, making it an appropriate target for linear regression. The insights gained from this project would be beneficial to anyone interested in stock returns including investors (e.g. individuals, wealth management, hedge funds), financial news agencies, and ratings agencies. Although the model itself is not accurate enough to be used, the process of building and evaluating the model yields valuable insights about predicting stock returns, including which features might be relevant and future directions to improve the model.

**Data**

Historical price and volume traded of all S&P500 stocks were obtained from the Yahoo Finance website via web scraping. 3M’s data had 13,076 rows representing daily values from which weekly return and average volume were calculated and filtered to a relatively stable bull market from late 2010 to early 2020 (avoiding the 2008 financial crisis and the 2020 Covid-19 pandemic). Various historical returns were derived as features. Other historical data including inflation, unemployment, Fed Funds rate, and US treasury rate were obtained from federal reserve websites, formatted as weekly rates, and merged with 3M’s data to result in a final dataframe with 522 rows, 8 feature columns, and 1 target column to be used for linear regression.

**Tools**

Selenium and BeautifulSoup for web scraping

NumPy and Pandas for data manipulation

Matplotlib and Seaborn to create visualizations

StatsModels and scikit-learn for linear regression and cross validation

**Communication**

- 5 minute presentation with Slides; GitHub repo: https://github.com/jaykwon2/3M\_weekly\_return\_LR