Predicting Performance of Stocks in the Dow Jones Index

Jul 24 2022 Session A6

Laura Lu

Presentation Outline

1. Introduction
2. Clustering
    a. Literature Review
    b. Model Investigation
    c. Findings Discussion
3. Random Forest
    a. Literature Review
    b. Model Investigation
    c. Findings Discussion
4. Linear Regression
    a. Literature Review
    b. Model Investigation
    c. FIndings Discussion

I. Introduction

The Significance

Total market capitalization is 93.8 trillion: more and more people are investing and it’s becoming a popular topic of research

The Dow Jones Index

- Dataset taken from UCI Machine Learning Repository
- **6 month range** of samples of the 30 companies in the Dow Jones Index from Jan to Jun 2011
- **Measures 16 variables**:
  - open, high, low, close, volume, percent change price, percent change volume over last week, previous weeks volume, next week's open, next week's close, percent change next week's price, days to next dividend and percent return next dividend
- **First Quarter** is used as **test data** and **Second Quarter** is used as **train data**

quarter	stock	date	open	high	low	close	volume	percent_change_price	percent_change_volume_over_last_wk	previous_weeks_volume
1	AA	1/7/2011	$15.82	$16.72	$15.78	$16.42	239655616	3.79267	NaN	NaN
1	AA	1/14/2011	$16.71	$16.71	$15.64	$15.97	242963398	-4.42849	1.380223	239655616.0
1	AA	1/21/2011	$16.19	$16.38	$15.60	$15.79	138428495	-2.47066	-43.024959	242963398.0
1	AA	1/28/2011	$15.87	$16.63	$15.82	$16.13	151379173	1.63831	9.355500	138428495.0
1	AA	2/4/2011	$16.18	$17.39	$16.18	$17.14	154387761	5.93325	1.987452	151379173.0
...	...	...	...	...	...	...	...	...	...	..
2	XOM	5/27/2011	$80.22	$82.63	$80.07	$82.63	68230855	3.00424	-21.355713	86758820.0
2	XOM	6/3/2011	$83.28	$83.75	$80.18	$81.18	78616295	-2.52161	15.221032	68230855.0
2	XOM	6/10/2011	$80.93	$81.87	$79.72	$79.78	92380844	-1.42098	17.508519	78616295.0
2	XOM	6/17/2011	$80.00	$80.82	$78.33	$79.02	100521400	-1.22500	8.811952	92380844.0
2	XOM	6/24/2011	$78.65	$81.12	$76.78	$76.78	118679791	-2.37762	18.064204	100521400.0

My Approach

I used three different algorithms: linear regression, k means clustering and random forest to evaluate which one best fit the data.
X Values: dates
Y Values: Percent__change__next__week__price

quarter	stock	date	open	high	low	close	volume	percent_change_price	percent_change_volume_over_last_wk	previous_weeks_volume
1	AA	1/7/2011	$15.82	$16.72	$15.78	$16.42	239655616	3.79267	NaN	NaN
1	AA	1/14/2011	$16.71	$16.71	$15.64	$15.97	242963398	-4.42849	1.380223	239655616.0
1	AA	1/21/2011	$16.19	$16.38	$15.60	$15.79	138428495	-2.47066	-43.024959	242963398.0
1	AA	1/28/2011	$15.87	$16.63	$15.82	$16.13	151379173	1.63831	9.355500	138428495.0
1	AA	2/4/2011	$16.18	$17.39	$16.18	$17.14	154387761	5.93325	1.987452	151379173.0
...	...	...	...	...	...	...	...	...	...	...
2	XOM	5/27/2011	$80.22	$82.63	$80.07	$82.63	68230855	3.00424	-21.355713	86758820.0
2	XOM	6/3/2011	$83.28	$83.75	$80.18	$81.18	78616295	-2.52161	15.221032	68230855.0
2	XOM	6/10/2011	$80.93	$81.87	$79.72	$79.78	92380844	-1.42098	17.508519	78616295.0
2	XOM	6/17/2011	$80.00	$80.82	$78.33	$79.02	100521400	-1.22500	8.811952	92380844.0
2	XOM	6/24/2011	$78.65	$81.12	$76.78	$76.78	118679791	-2.37762	18.064204	100521400.0

- Linear Regression
- K Means Clustering
- Random Forest

Point of Criticism: Data is a decade old so it shouldn't be assumed that models will perform the same in present day.

II. Clustering

II. Clustering

Literature Review

Dynamic-radius species-conserving genetic algorithm for the financial forecasting of Dow Jones index stocks.

Brown, M. S., Pelosi, M. J., & Dirska, H. (2013, July)

Summary

The DSGA will form **clusters** of individuals that are within a predetermined **radius**.

- Moves the strongest member of a cluster within a certain radius to the **next generation**.
- Selected by **highest fitness score**

Results

Rate of Return: summing the weekly return and dividing by the initial investment amount

- Produced a **0.54% weekly rate of return**
- All 1 week trials **outperformed** the Dow Jones Index

Dynamic-radius species-conserving genetic algorithm for the financial forecasting of Dow Jones index stocks.

Brown, M. S., Pelosi, M. J., & Dirska, H. (2013, July)

Shortcomings and Bias

Differences in how someone would invest in one week time vs month vs year frame

- Predictive model is suitable for an investment type suitable to day traders

Wrong prediction can also generate greater losses than the Dow

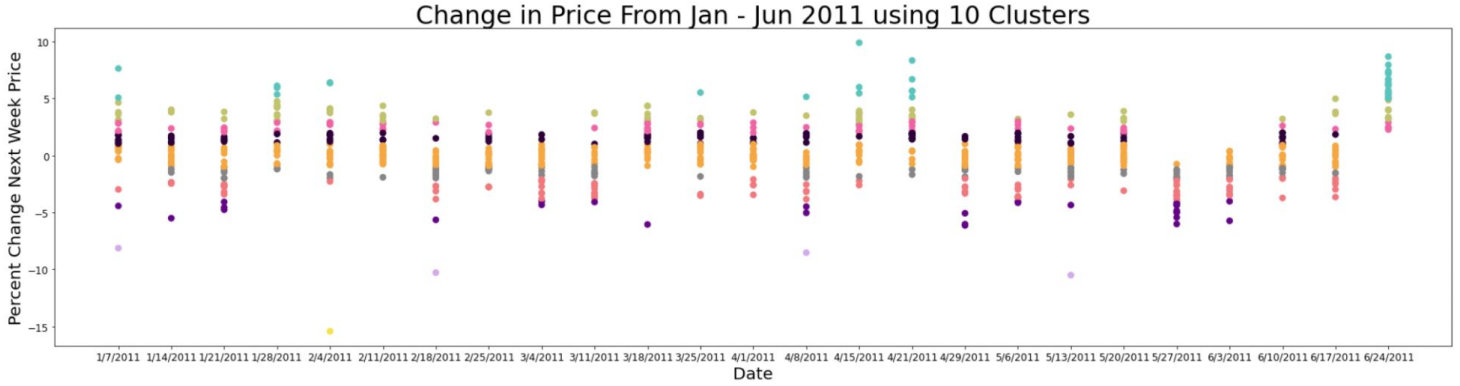
- Dow Jones Index minimizes the losses of an inaccurate prediction

II. Clustering

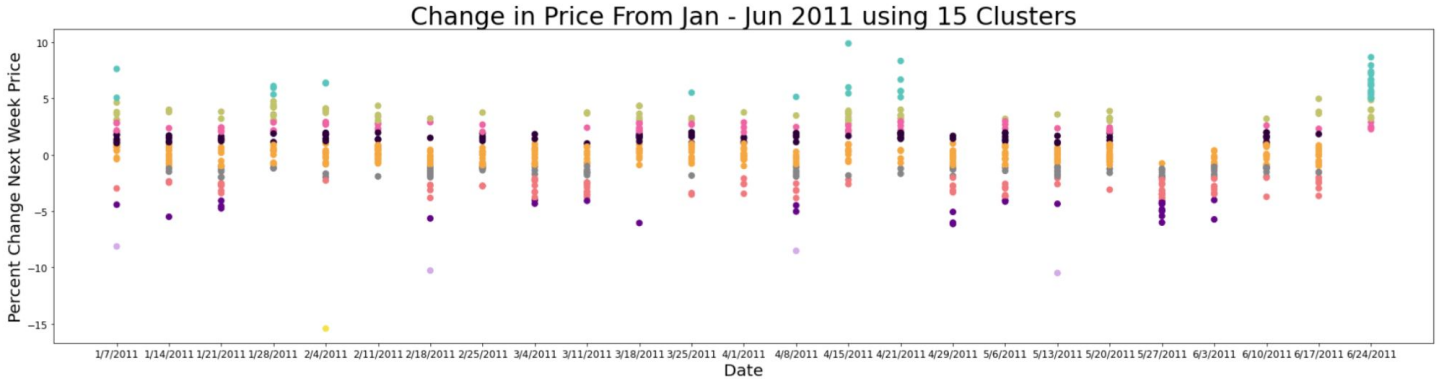
Model Investigation

Hyperparameter: Inertia

measures how well a dataset is clustered by k-means



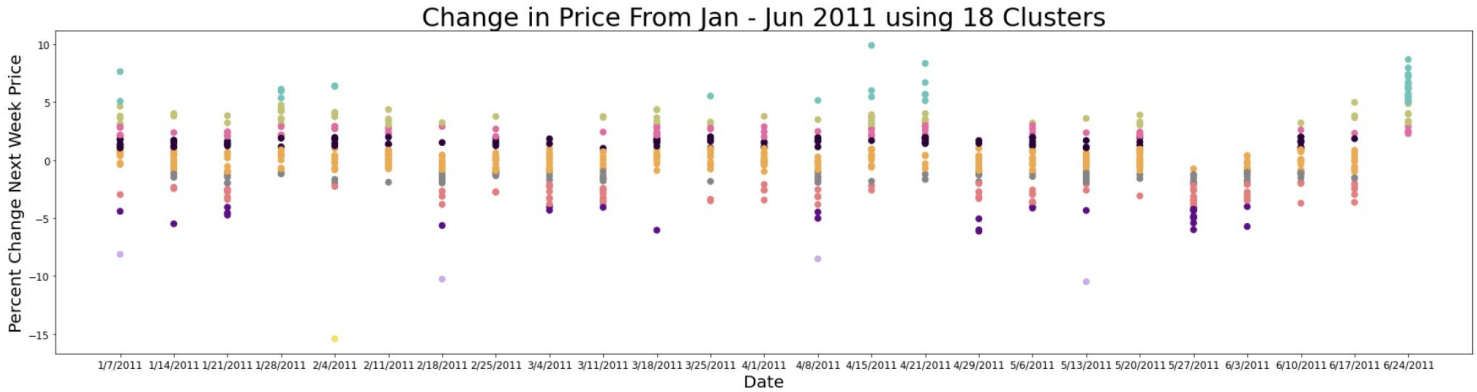
10 Clusters
Inertia: 162



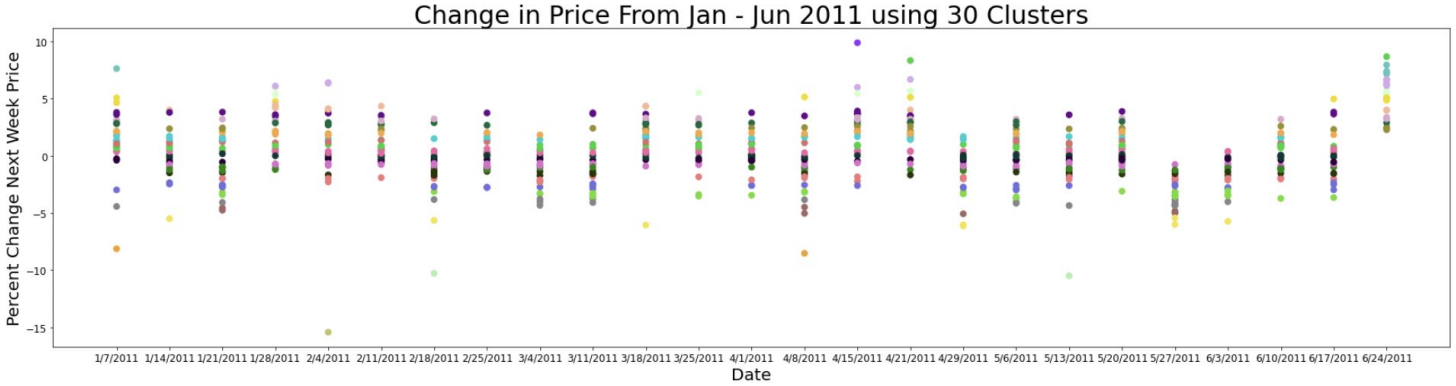
15 Clusters
Inertia: 63

Hyperparameter: Inertia

The more clusters, the lower the inertia. 30 clusters had the lowest inertia.



18 Clusters
Inertia: 44



18 Clusters
Inertia: 13

K Means Clustering Algorithm Model

```
kmeans = KMeans(n_clusters=18)

kmeans.fit(x)

•••

kmeans.labels_

•••

kmeans.predict(x)
```

Shortcomings and Bias

Assumption: Each cluster represents the average change in price from the present week to the next week

- May **not be accurate**

Difficult to **interpret results of cluster** of whether stakeholder should buy or sell

- Needs to be revised to provide stakeholder with more cluster information
- As a result, it is difficult to compare with findings in the paper

III. Random Forest

III. Random Forest

Literature Review

Predicting the direction of stock market prices using Random Forest

Khaidem, L., Saha, S., & Dey, S. R. (2016)

Summary

Split criteria was based on **Gini Coefficient** and **Shannon Entropy**

- Gini Coefficient: quality of split
- Shannon Entropy: information disorder

Results

Tested on Apple, GE and Samsung Stocks

Days	Accuracy %	Precision %
30	86	81
60	90	91
90	93	94

Predicting the direction of stock market prices using Random Forest

Khaidem, L., Saha, S., & Dey, S. R. (2016)

Bias and Shortcomings

Number of Trees was cross validated while Depth was not cross validated

- May not be an optimized model

Unclear Timeframe

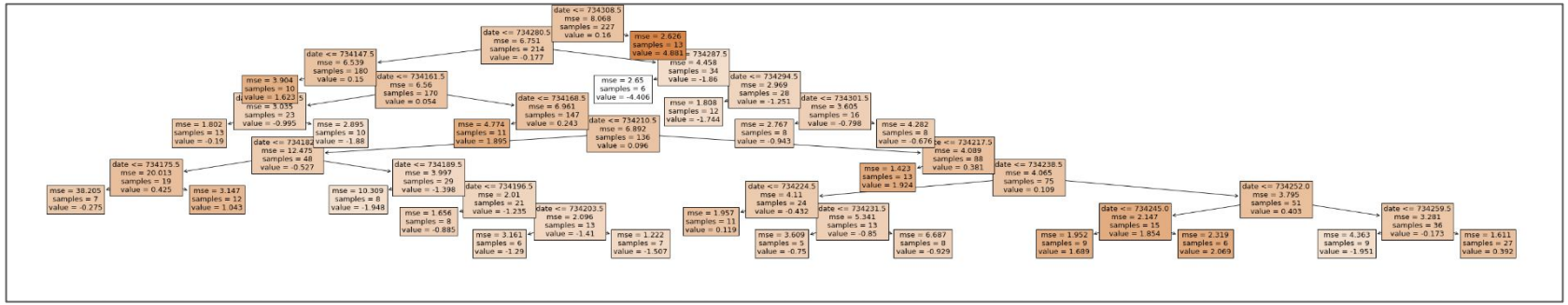
- Accuracy and precision could be attributed to a selected time frame when the economy was already steady

III. Random Forest

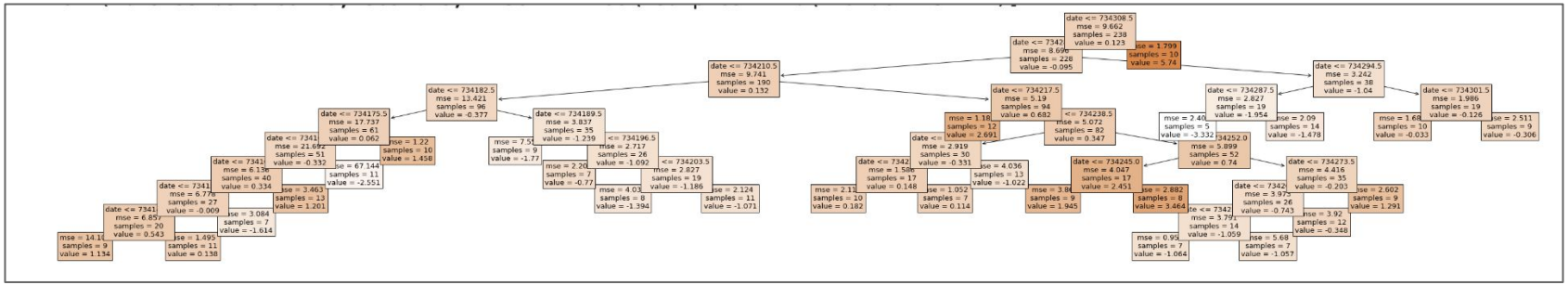
Model Investigation

Hyperparameter: Inertia

Observation: Different plots even though same first index max depth



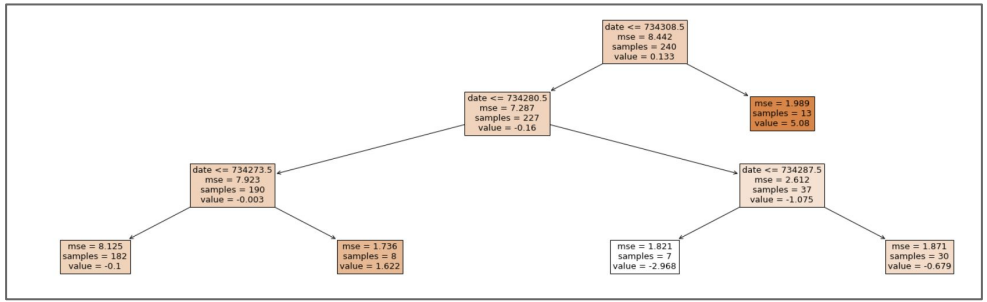
Max Depth 10 MSE: 4.47



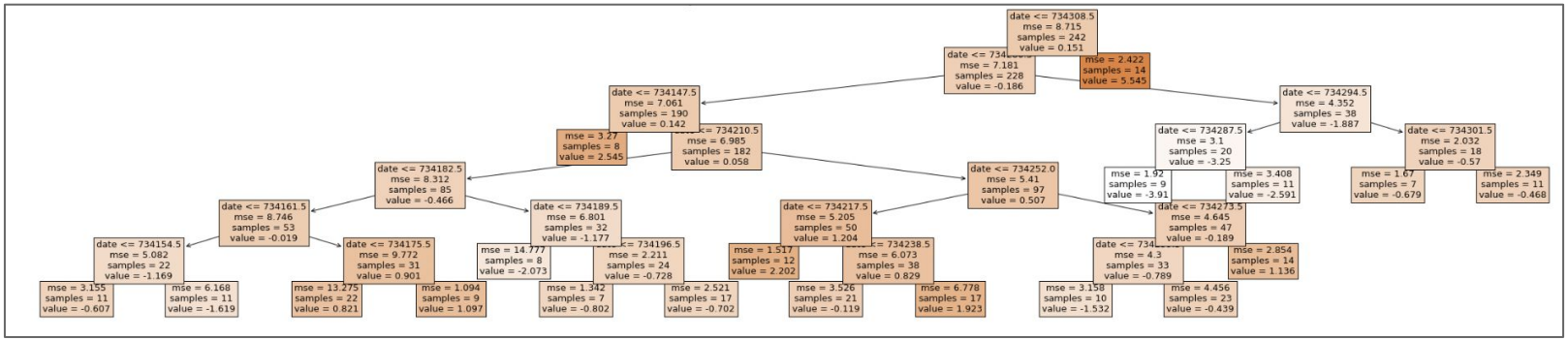
Max Depth 9 MSE: 4.47

Hyperparameter: Inertia

Max Depth of 7 produced the lowest Mean Squared Error



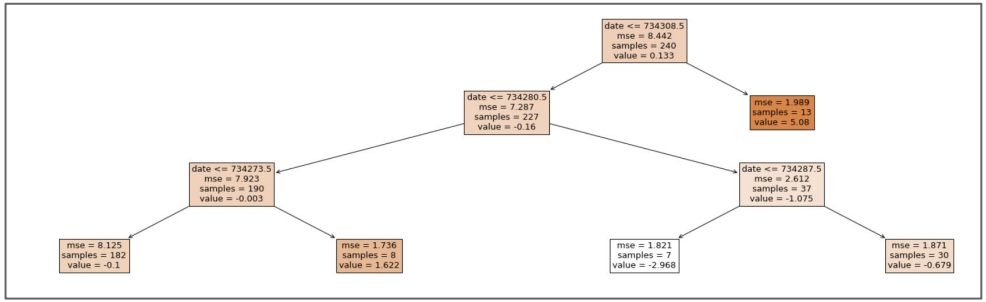
Max Depth 3 MSE: 5.02



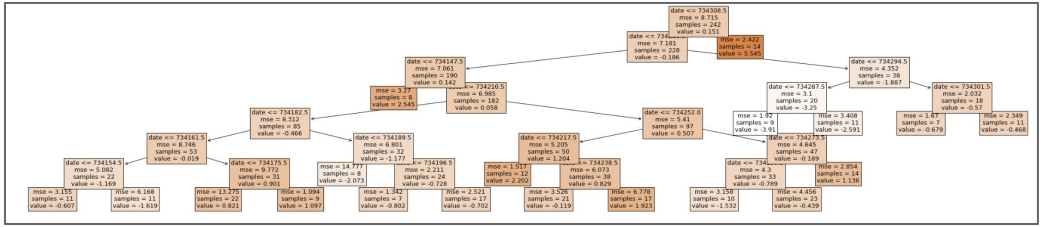
Max Depth 7 MSE: 4.46

Hyperparameter: Tree Depth

The low MSE indicates that the Random Forest Model strongly fits the data, similar to the arguments made in the paper



Max Depth 3 MSE: 5.02



Max Depth 7 MSE: 4.46

IV. Linear Regression

IV. Linear Regression

Literature Review

An Assessment of Stock Return Prediction using Machine Learning

Timmerman, N. (2021)

Summary

Measured the amount of returns using a linear regression model from 10-10-2013 to 31-5-2014 and 29-12-2017 to 31-10-2020

Results

Expected large difference between predicted and actual value for some stocks

- 65% difference for Amazon Stocks

An Assessment of Stock Return Prediction using Machine Learning

Timmerman, N. (2021)

Bias and Shortcomings

Two different time periods with different vastly economic conditions and lengths

- Underfitting data

Only tested on 11 stocks from similar industries

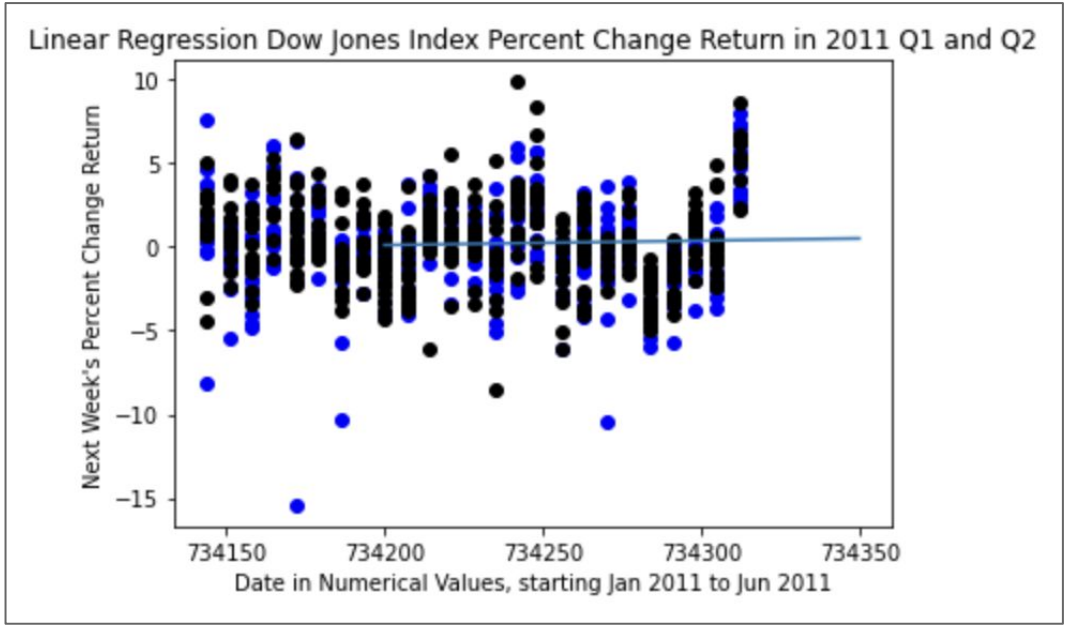
- Wider test data in order for model to be generalizable across the industry

IV. Linear Regression

Model Investigation

Linear Regression

The MSE was 6.6 while the R^2 was -0.0125. The negative R^2 indicates that the Linear Regression model did not fit the data well.

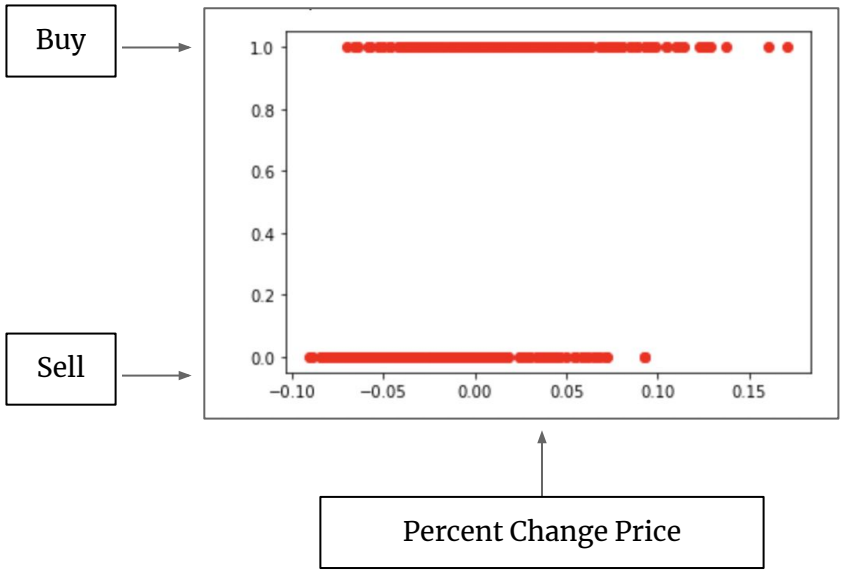


V. Next Steps of Research

Using a Logistic Regression Model

Classify stocks to be Buy or Sell depending on the percent change price

Hypothetical Visualization



VI. Conclusion

Conclusion

Thank you for listening!

K-Means Clustering model fit the data but provided vague predictions

Random Forest model strongly fit the data

Linear Regression model poorly fit the data and further investigation should be done with a Logistic Regression model