Capstone project report **Kiem Nguyen**
Machine Learning Engineer Nanodegree **February 2021**

Project Overview

Trading for profit is a difficult problem to solve. In an efficient market, buyers and sellers would have all the information needed to make a rational trading decision so the stock is should remain at its fair values. The reality is the financial market is not efficient in real life especially now when automated trading allows for thousands of transactions to occur within a nanosecond.

This project tries to select Alpha signal from trading using data science and machine learning techniques. The data set is a 10-year trading data of SPY from January 2021 to the most recent date February 2021. SPY is a stock market index that measures the stock performance of 500 large companies listed on stock exchanges in the US. This fund is the largest ETF in the world. The SPY movement could signal the direction of the market.

I use pandas-datareader library to extract end-of-day stock pricing data from Yahoo Finance from 2011-2020. This library provides functions to extract data from various internet sources in a pandas data frame. I received an error when trying to collect SPY data from the Quandl data source. It turned out that Quandl only has SPY data for the paid subscribers so the API key in my account doesn't work. Google also retired their stock data source. Fortunately, the Yahoo Finance data source works great so I decided to use this data source.

Problem Statement

Between 2010 and 2020 the S&P 500 has an annual average return of 13.6% in the past 10 years. This project would provide a practical approach to predict future price movements in financial markets based on past returns. The assumption is that certain patterns in financial markets repeat themselves such that past observations can be leveraged to predict future price movements. (Hilpisch, Yves. 2020. Artificial Intelligence in Finance: A Python-Based Guide.)

Metrics

For the random forest model, I used R2 scores from the DecisionTreeRegressor to evaluate the model performance. In addition, I used the scatter plot to visualize the prediction vs actual values.

For Neural Network models and ensembling models, I used the r2_score() function from sklearn.metrics to calculate the score. I also used the scatter plot to visualize the prediction vs actual values.

Data Exploration

The data has both stock's closing price (Close) and adjusted closing price (Adj Close). The closing price is the last transaction price before the market closes. The adjusted closing price factors in corporate actions, such as stock splits, dividends, and rights offerings. I decided to use the adjusted closing price for the models.

Date	High	Low	Open	Close	Volume	Adj Close
2021-02-12	392.899994	389.769989	389.850006	392.640015	50505700.0	392.640015
2021-02-16	394.170013	391.529999	393.959991	392.299988	50700800.0	392.299988
2021-02-17	392.660004	389.329987	390.420013	392.390015	52290600.0	392.390015
2021-02-18	391.519989	387.739990	389.589996	390.720001	59552200.0	390.720001
2021-02-19	392.380005	389.549988	392.070007	390.029999	83142800.0	390.029999

The SPY data from 2010-01-01 to the most recent date 2021-02-19 has 2802 rows and 6 columns

```
df.shape # get the number of rows and columns
```

```
(2802, 6)
```

```
df.describe() # generate descriptive statistics
```

	High	Low	Open	Close	Volume	Adj Close
count	2802.000000	2802.000000	2802.000000	2802.000000	2.802000e+03	2802.000000
mean	210.783105	208.517659	209.704789	209.734333	1.237649e+08	192.583213
std	69.809154	69.075299	69.468840	69.454677	7.465493e+07	75.348641
min	103.419998	101.129997	103.110001	102.199997	2.027000e+07	82.872505
25%	142.442497	141.355003	141.982498	141.982498	7.239802e+07	120.491968
50%	206.800003	204.629997	205.614998	205.614998	1.038480e+08	184.856987
75%	268.587502	265.575005	267.577507	267.287506	1.530642e+08	254.170013
max	394.170013	391.529999	393.959991	392.640015	7.178287e+08	392.640015

We can see there is no null value in the data, and the data index is the Date column:

```
df.info()
```

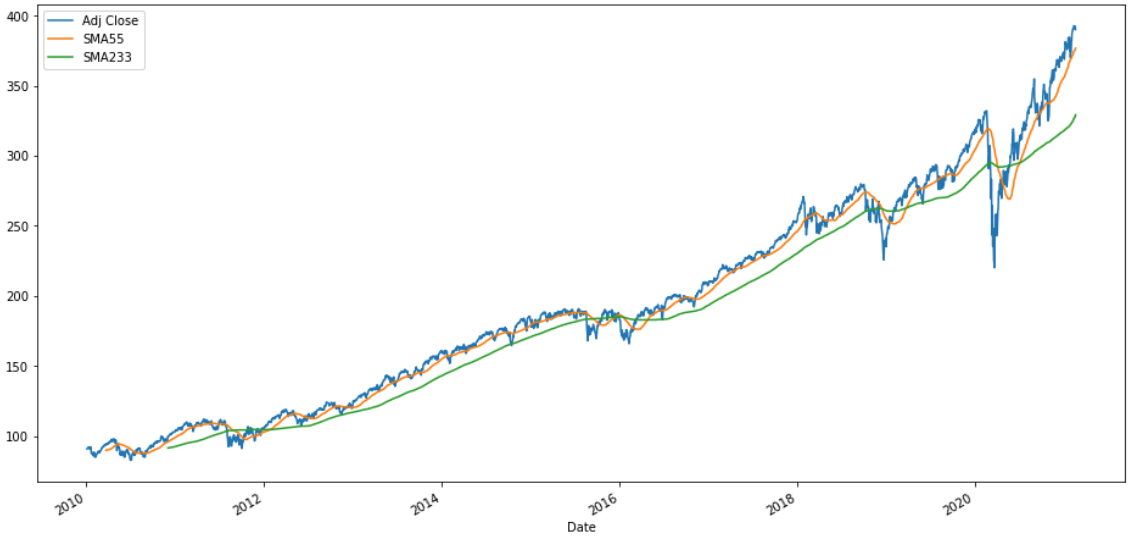
```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 2802 entries, 2010-01-04 to 2021-02-19
Data columns (total 6 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   High       2802 non-null   float64
 1   Low        2802 non-null   float64
 2   Open       2802 non-null   float64
 3   Close      2802 non-null   float64
 4   Volume     2802 non-null   float64
 5   Adj Close  2802 non-null   float64
dtypes: float64(6)
memory usage: 153.2 KB
```

```
df.index
```

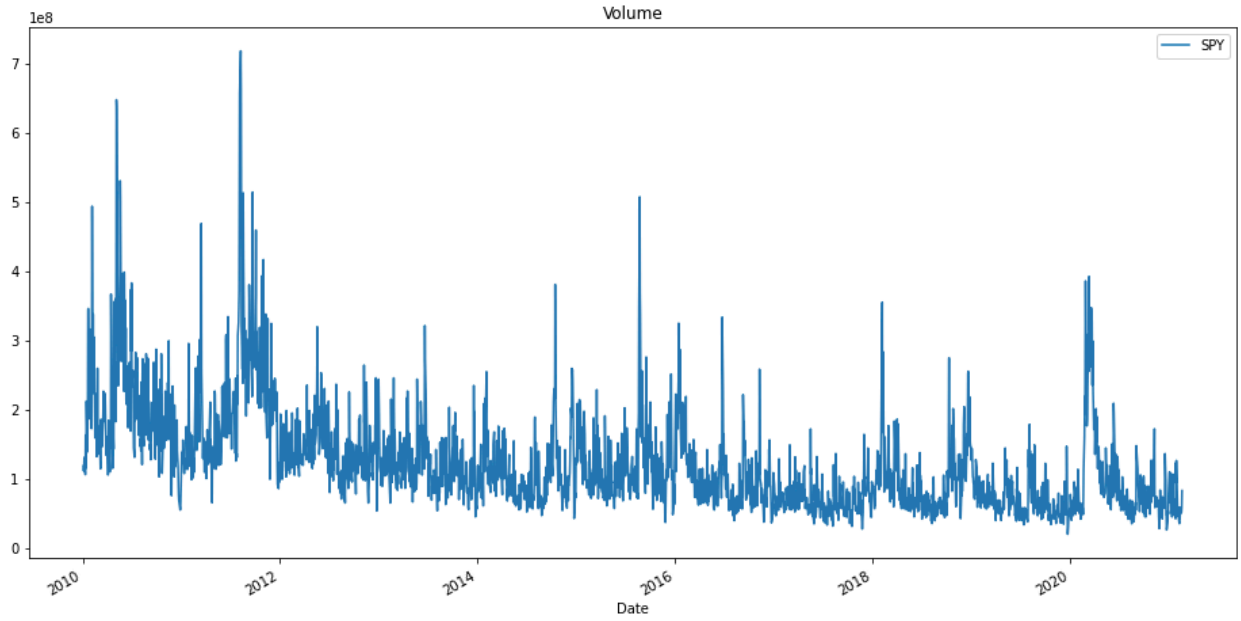
```
DatetimeIndex(['2010-01-04', '2010-01-05', '2010-01-06', '2010-01-07',
               '2010-01-08', '2010-01-11', '2010-01-12', '2010-01-13',
               '2010-01-14', '2010-01-15',
               ...
               '2021-02-05', '2021-02-08', '2021-02-09', '2021-02-10',
               '2021-02-11', '2021-02-12', '2021-02-16', '2021-02-17',
               '2021-02-18', '2021-02-19'],
              dtype='datetime64[ns]', name='Date', length=2802, freq=None)
```

Exploratory Visualization

The adjusted closing price of SPY since 2010-01-01 with 55 days and 233 days simple moving average (SMA). A simple moving average (SMA) calculates the average of a selected range of closing prices, by the number of periods in that range. It is a technical indicator that can aid in determining if an asset price will continue or if it will reverse a bull or bear trend.



The volume of SPY since 2010-01-01:



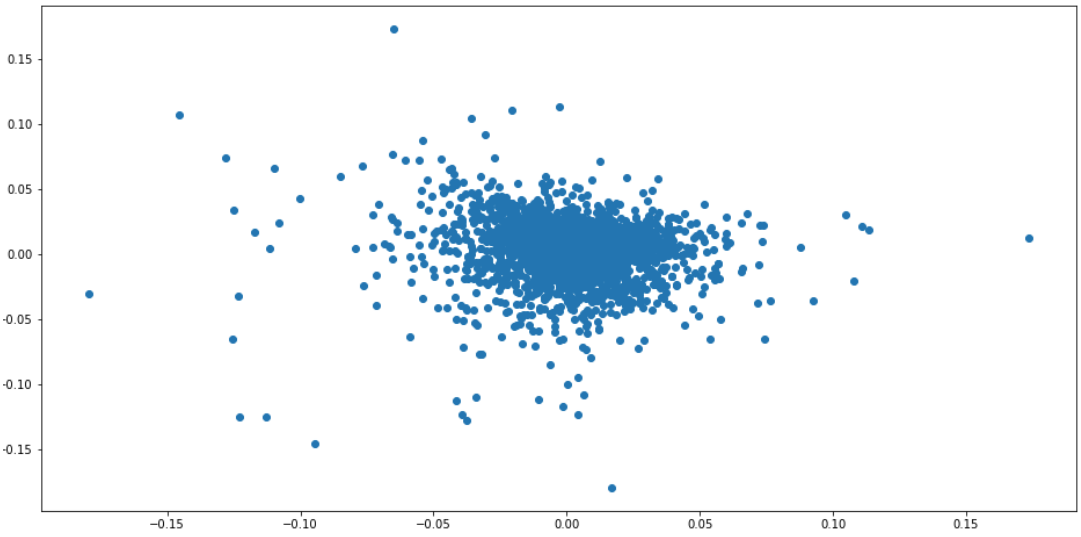
SPY has a huge volume in 2011. Let's find out the exact date of this event to see what happend that date

```
df['Volume'].idxmax()
```

```
Timestamp('2011-08-09 00:00:00')
```

The event was on Black Monday when US and global stock markets crashed.

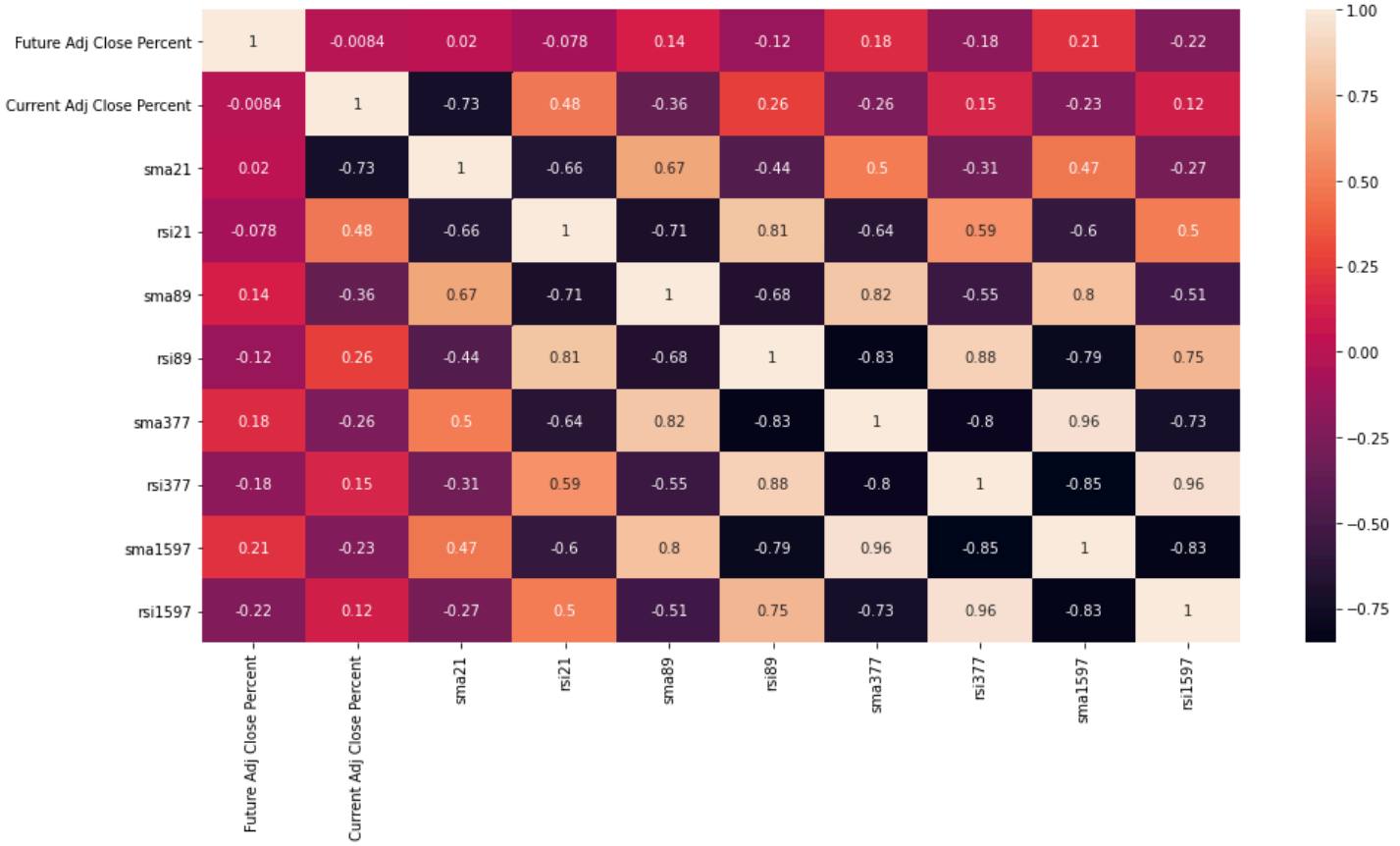
The correlation coefficient would tell us how strong the correlation of previous price changes vs future price changes. If it's highly correlated, then the stock price is a trend following. Otherwise, the stock price is mean reverting.



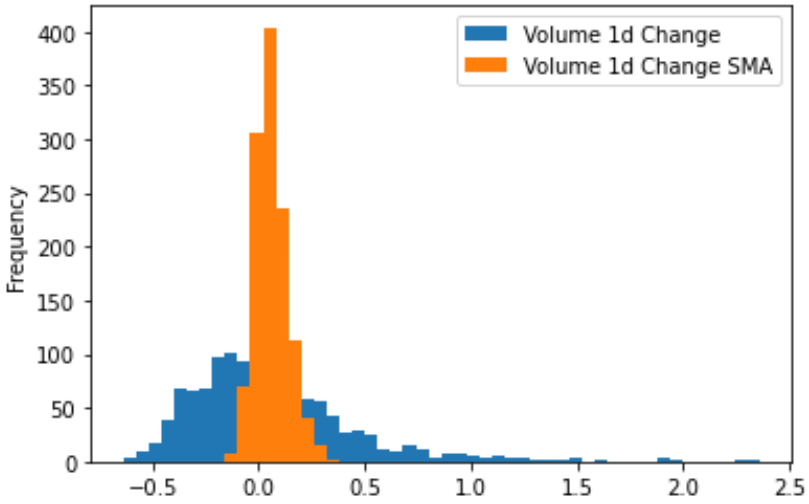
We can see the future price change is negatively correlated to the previous price change for a 5 days trading period. This tells us that a mean reversion trading would be a good trading strategy.

Future Engineering:
Before building the models, I want to add more data to make a better prediction by doing feature engineering. The new features I add in the models are:

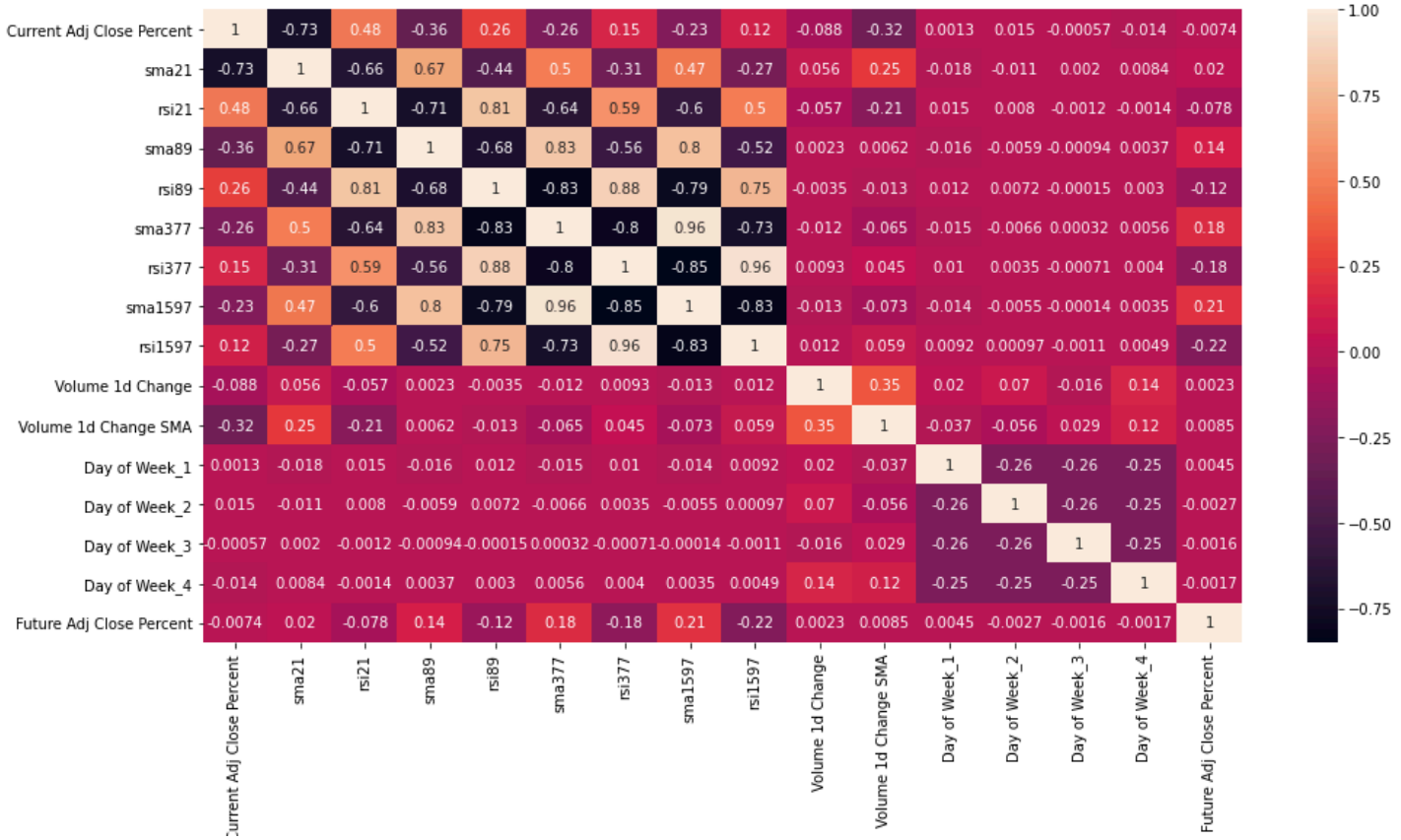
1. An RSI (Relative Strength Index) for different time period. RSI provides signals about bullish and bearish price momentum. A stock is usually considered overbought when the RSI is above 70% and oversold when it is below 30%. Below is the heatmap to see the correlation between the new features and the target (Future Adj Close Percent):



2. A simple moving average (SMA) for different time period which is one of the most common indicators.



3. Day of week feature: As for these new features, I used numbers in the Fibonacci Sequence for different time periods. First, I tried 55, 89, 144, 233 time period, but then switched to 21, 89, 377, 1597 because this give better correlations between the features and targets which could help the model.



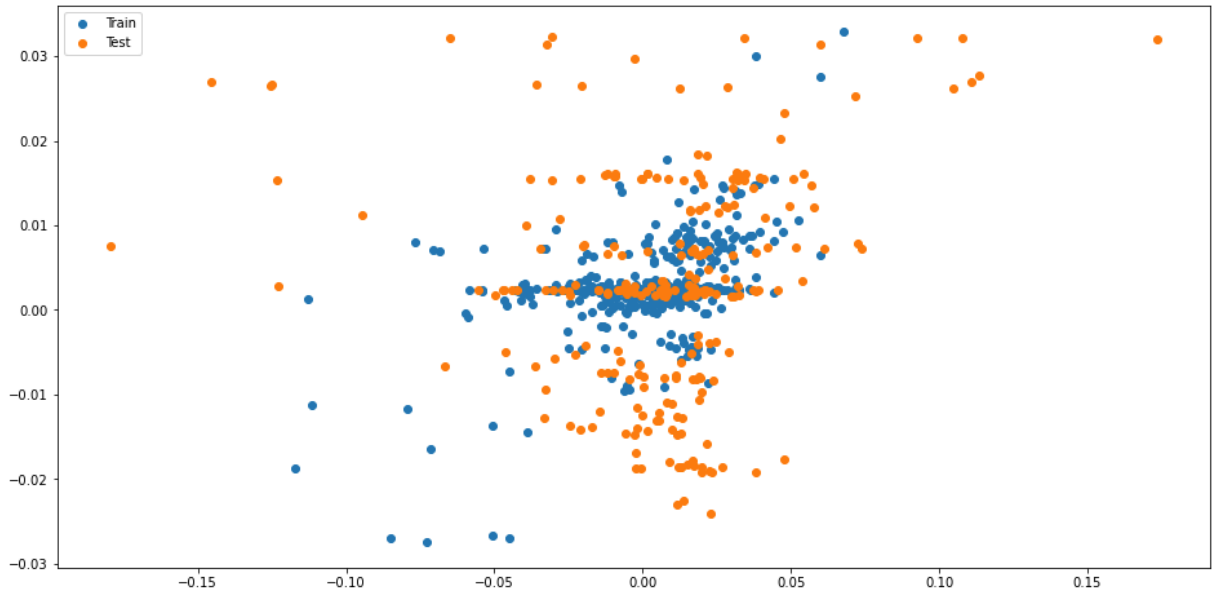
Algorithms and Techniques

Random Forest:

I use Random Forest model first to predict the future price change of SPY because it usually performs well, and it has a lot of settings to tune the model performance.

After getting a low score from a based line random forest model, I used the ParameterGrid to turn the hyperparameter to improve the test score, then I fit the model with the best hyperparameter found.

```
# Check the model performance

# fit the model with the best hyperparameter we found above
rf = RandomForestRegressor(n_estimators=500, max_depth=2, max_features=4, random_state=40)
rf.fit(X_train, y_train)

train_predict = rf.predict(X_train)
test_predict = rf.predict(X_test)

plt.figure(figsize=(16,8))
plt.scatter(y_train, train_predict, label='Train')
plt.scatter(y_test, test_predict, label='Test')
plt.legend()
plt.show()
```



Next, I use StandardScaler to scale the train and test data to build two Neural Network models with different settings using keras library. Neural nets can capture the interaction between different hyperparameters very well.

Neural Networks:

The first Neural Network model has 3 layers. The first two layers both have relu activation function and the last layer has a linear activation function. I also apply 20% Dropout after the first layer to avoid over fitting.

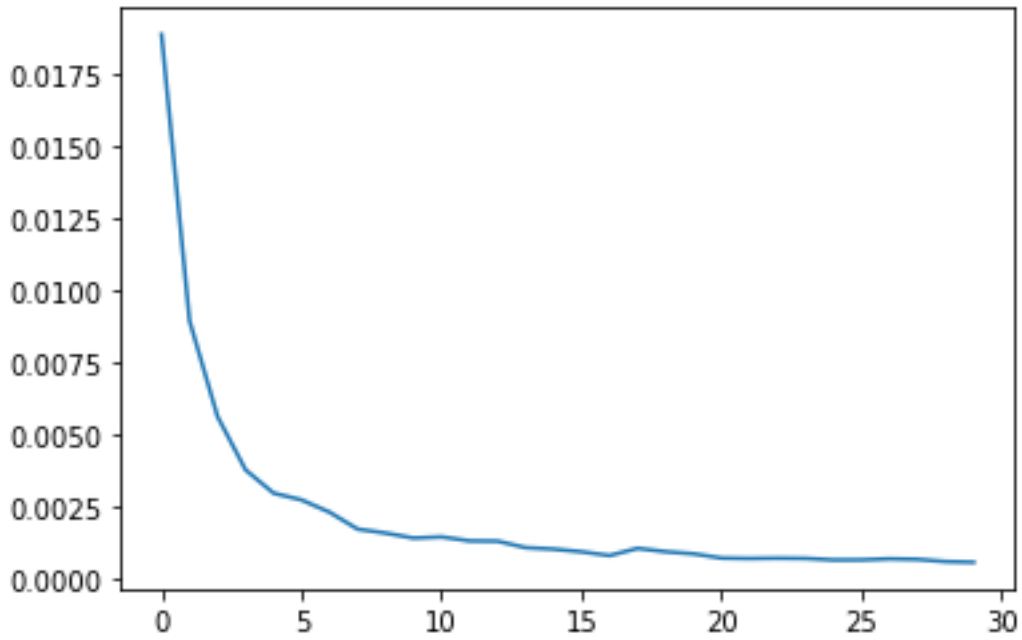
```
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import Dropout

nn = Sequential()
nn.add(Dense(200, input_dim=X_train_scale.shape[1], activation='relu'))
nn.add(Dropout(0.2)) # add dropout to prevent overfitting
nn.add(Dense(50, activation='relu'))
nn.add(Dense(1, activation='linear'))
```

Next, I fit the model with a mean squared error loss function and train with 30 epochs:

```
nn.compile(optimizer='adam', loss='mse')
history = nn.fit(X_train_scale, y_train, epochs=30)
```

```
Epoch 1/30
960/960 [==============================] - 1s 738us/step - loss: 0.0189
Epoch 2/30
960/960 [==============================] - 0s 99us/step - loss: 0.0089
Epoch 3/30
960/960 [==============================] - 0s 104us/step - loss: 0.0056
Epoch 4/30
960/960 [==============================] - 0s 106us/step - loss: 0.0038
Epoch 5/30
960/960 [==============================] - 0s 106us/step - loss: 0.0030
```

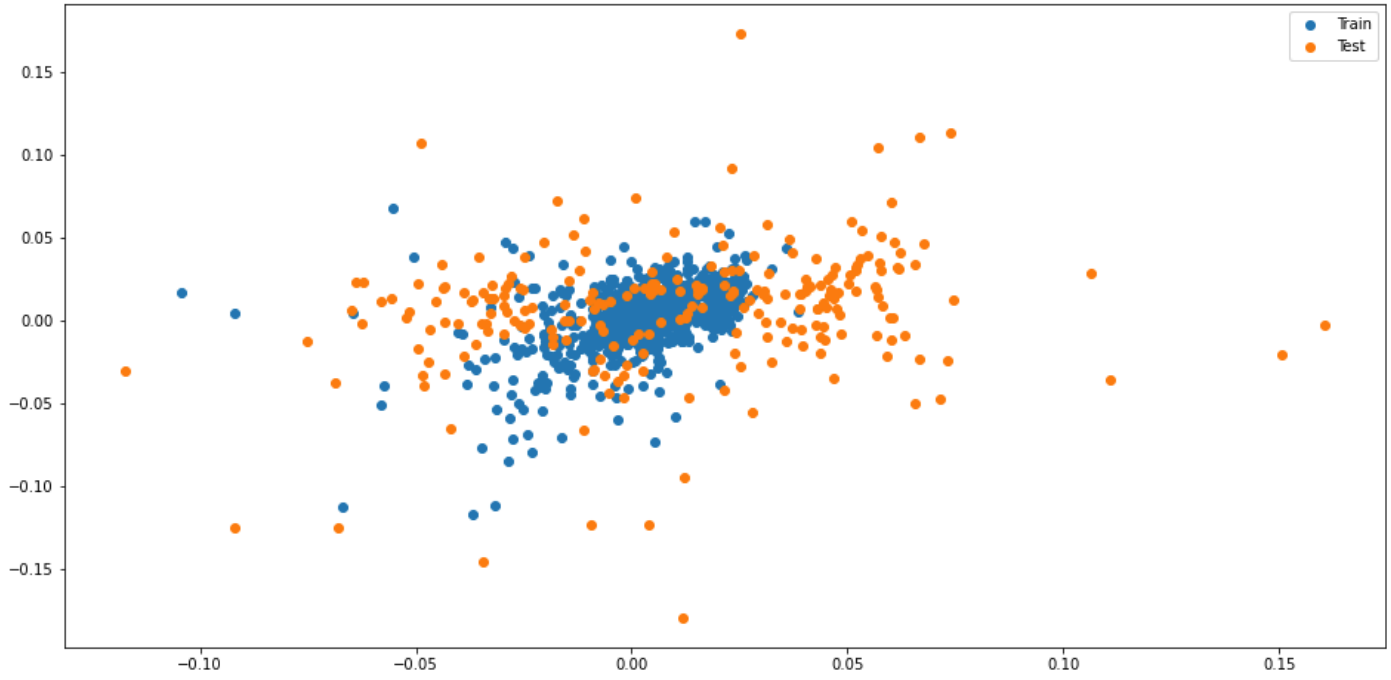
We can see the loss curve is flattened out, so the neural net was sufficiently trained:



After fitting the model, I use the r2_score function to check the model performance then plot the prediction vs actual values to see how the model preformed:

```
from sklearn.metrics import r2_score

train_predict_nn = nn.predict(X_train_scale)
test_predict_nn = nn.predict(X_test_scale)

print(r2_score(train_predict_nn, y_train))
print(r2_score(test_predict_nn, y_test))

plt.figure(figsize=(16,8))
plt.scatter(train_predict_nn, y_train, label='Train')
plt.scatter(test_predict_nn, y_test, label='Test')
plt.legend()
plt.show()
```

```
-0.6066821374793427
-0.48706220744139594
```



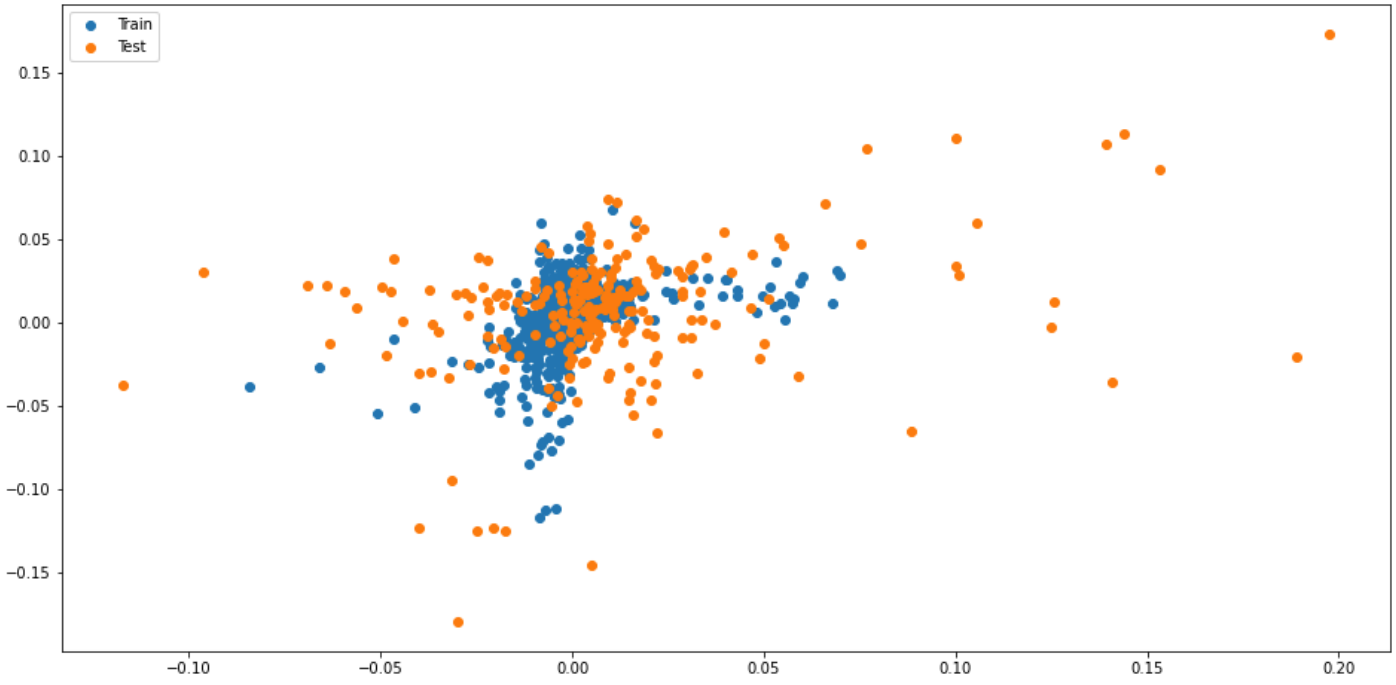
Next, I use keras.losses and tensorflow to create a custom loss function for the second Neural Network model instead of using the previous mean squared error loss function. This custom loss function would give more penalty weight for predicting the wrong stock's closing price.

```
import keras.losses
import tensorflow as tf

def custom_loss(true_val, predict_val):
    penalty = 500
    loss = tf.where(tf.less(true_val * predict_val, 0), penalty * tf.square(true_val - predict_val), tf.square(true_val

    return tf.reduce_mean(loss, axis=-1)

keras.losses.custom_loss = custom_loss
```

The second Neural Network model performs better than the first Neural Network model on the test data:

```
train_predict_nn2 = nn2.predict(X_train_scale)
test_predict_nn2 = nn2.predict(X_test_scale)

print(r2_score(train_predict_nn2, y_train))
print(r2_score(test_predict_nn2, y_test))

plt.figure(figsize=(16,8))
plt.scatter(train_predict_nn2, y_train, label='Train')
plt.scatter(test_predict_nn2, y_test, label='Test')
plt.legend()
plt.show()
```

```
-1.3679893187634766
-0.29260031074236004
```

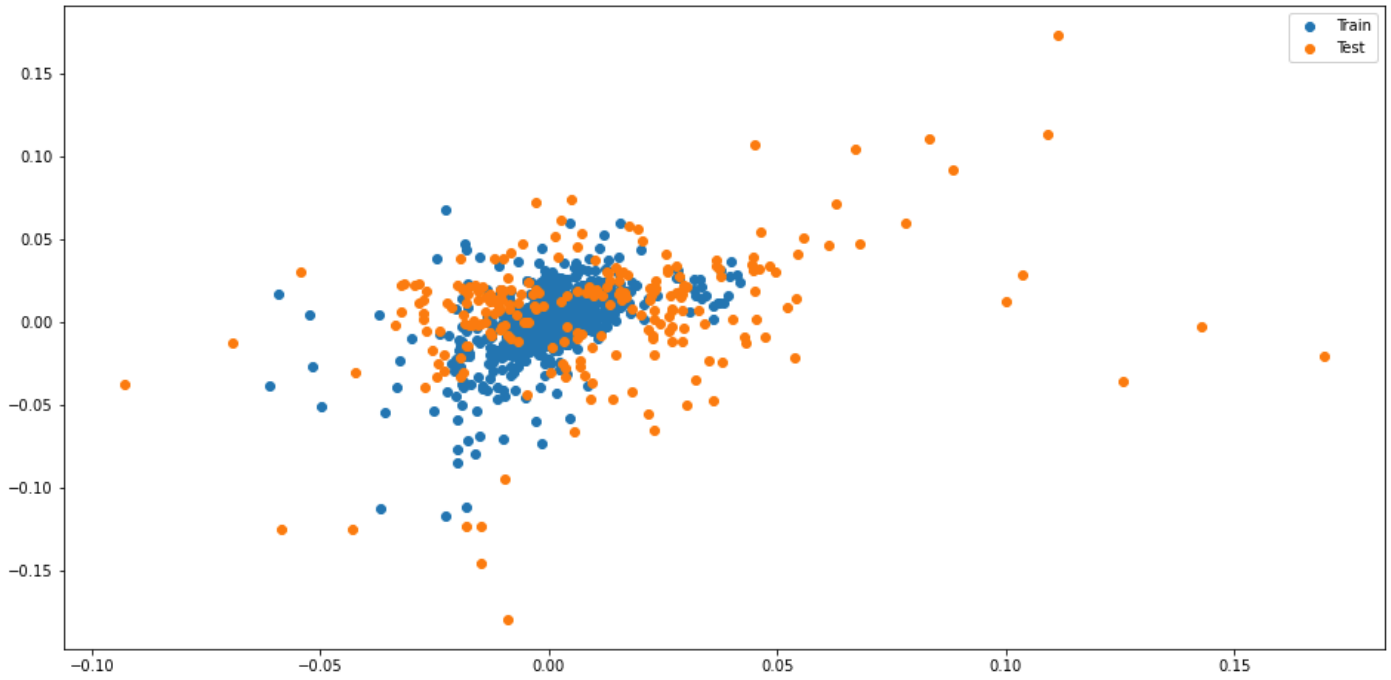


Finally, I stack both neural network models and take the average prediction scores to improve the prediction.

Ensembling Model Performance

```
from sklearn.metrics import r2_score

print(r2_score(ensembling_train_predict, y_train))
print(r2_score(ensembling_test_predict, y_test))

plt.figure(figsize=(16,8))
plt.scatter(ensembling_train_predict, y_train, label='Train')
plt.scatter(ensembling_test_predict, y_test, label='Test')
plt.legend()
plt.show()
```

```
-1.257762666961324
-0.5796675982504489
```



For the final result, we can see the r2_score values are about the average of both models. I have trained with many different parameters, but the scores still can't improve significantly. However, by doing feature engineering, hyper parameter tuning, apply drop out and stacking models, I do notice the improvements.