# Capstone Project

## Machine Learning Engineer Nanodegree

Carl Gosselin

September, 2016

## I. Definition

### Project Overview

**The problem domain**

I've chosen to apply my new-found knowledge in machine learning algorithms to the investment and trading domain. I keep hearing that many firms are using machine learning algorithms to gain an edge in the market. For this project, I'd like to:

A) understand how to apply machine learning algorithms to stock data to predict future price

B) understand the current challenges of using machine learning algorithms to help predict stock prices

Machine Learning Algorithms

For this project, I will be applying two machine learning algorithms to the stock data to predict the stock price 5 days later:

A) sklearn's Linear Regression (http://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LinearRegression.html)

B) sklearn's KNN Regression (http://scikit-learn.org/stable/modules/neighbors.html)

**Project origin**

This project originated from Udacity's course - <a href="https://www.udacity.com/course/machine-learning-for-trading--ud501?\_ga=1.193565739.763209811.1447720300"> Machine Learning for Trading </a>.

I've always been interested in understanding the stock market. I'm happy to have the opportunity to apply machine learning algorithms to stock data.

<a href="https://www.youtube.com/watch?v=j-6pr72cves" target="\_blank"> "\_I am very happy to be here!\_" </a>

**Related datasets**

CSV files captures the daily stock data for the following stocks from the last 5 years

1. SPY (SPDR S&P 500 ETF)

2. UPRO (ProShares UltraPro S&P500)

3. GOOG (Google)

4. AAPL (Apple)

5. AMZN (Amazon)

6. DIS (Disney)

7. NFLX (Netflix)

8. FB (Facebook)

9. AXY (Alterra Power)

10. VIX (Volatility Index)

11. TSLA (Tesla)

12. GWPH (GW Pharmaceuticals)

13. MSFT (Microsoft)

14. GLD (SPDR Gold Shares)

### Problem Statement

My problem statement is simple and straightforward - How can I increase my chances of picking "winning" stocks?

My current analysis involves tracking a handful of stocks on google.com/finance. When time permits, I print-off quarterly and other reports and read them in detail. I believe that another strategy that would complement my current analysis is creating predictive models on the price of the stock. I feel that my feeble brain is no match for machine learning algorithms when it comes down to crunching numbers.

### Metrics

I will be attempting to use the train/test split utility on the stock data from sklearn's cross\_validation module . A section of the data will be used to train the model and another section will be used to test the model on unseen data (out of sample data). I will use the score method to measure the difference between predicted prices and real prices in the testing (out-of-sample) dataset.

Sample lines of code:

splitting the train / test data: `X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3, random\_state=0)`

Displaying the predictive score of linear regression or knn regression on unseen data -> `print('regr.score(X\_test, y\_test): %.2f' % regr.score(X\_test, y\_test))`

## II. Analysis

### Data Exploration

For this project, I will be using data in the form of datasets. I have selected 12+ stocks that are actively traded on the stock market. I've also included data related to the volatility index (VIX). This is also known as the fear gauge. It is explained that when the VIX is high, the market moves lower as they are (supposedly) inversely related. I plan to include the VIX data with other stocks in visual comparisons.

All stock data files have the same columns:

- Date (date of the stock price)

- Open (opening price of the stock

- High (highest price of the stock for the day)

- Low (lowest price of the stock for the day)

- Close (price of the stock at close)

- Volume (the trading volume of the stock for the day)

- Adj Close (the adjusted close price)

note: the adjusted close price will be different from the "Close" price when a company chooses to split the stock, give dividends, etc...

Sample data - SPY stock:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Date** | **Open** | **High** | **Low** | **Close** | **Volume** | **Adj Close** |
| 2011-08-10 | 115.26 | 116.27 | 111.94 | 112.29 | 662607400 | 101.24 |
| 2011-08-11 | 113.26 | 116.27 | 111.94 | 112.29 | 662607400 | 105.79 |
| 2011-08-12 | 118.40 | 119.20 | 117.27 | 118.12 | 313731600 | 106.50 |
| 2011-08-15 | 119.19 | 120.73 | 119.00 | 120.62 | 258810600 | 108.75 |
| 2011-08-16 | 119.47 | 120.69 | 118.30 | 119.58 | 294095200 | 107.82 |

Statistical analysis of sample data using dataframe.describe:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Open** | **High** | **Low** | **Close** | **Volume** | **Adj Close** |
| Count | 1258 | 1258 | 1258 | 1258 | 1258 | 1258 |
| Mean | 174.59 | 175.46 | 173.68 | 174.65 | 134220000 | 166.88 |
| Std | 30.69 | 30.72 | 30.65 | 30.70 | 65142420 | 33.61 |
| Min | 108.34 | 112.58 | 107.43 | 109.93 | 373178000 | 99.63 |
| 25% | 143.43 | 144.09 | 142.62 | 143.36 | 910950200 | 132.71 |
| 50% | 183.71 | 184.27 | 182.80 | 183.83 | 1189386000 | 174.88 |
| 75% | 203.62 | 204.52 | 202.55 | 203.46 | 1589144000 | 199.05 |
| max | 218.39 | 218.75 | 217.80 | 218.17 | 6626074000 | 218.17 |

note: I don't have anything to highlight in the data. I will however be appending an extra column at the end of the dataset to capture the adjusted close price 5 days later. To do this, I will copy the existing "Adj Close" column and shift the data 5 rows up. Below is the snippet of the code to do this:

df['Adj\_Close\_5\_Days\_Later'] = df['Adj Close']

df['Adj\_Close\_5\_Days\_Later'] = df['Adj\_Close\_5\_Days\_Later'].shift(-5)

Other comments about the data:

- By default, the csv files are indexed by a number starting from "0". The code will need to explicitly identify the "Date" column as the index column.

- The data in the csv files begin with the latest trade day. Visualizing this data untouched, the graph will show a downward trend for stocks with increasing prices. The data will need to be re-organized to show a proper linear progression through time.

- It is possible that some of the stocks did not trade on a certain day. Stock with no trades on a specific day will need a "nan" (or similar) inserted into the empty cell.

- CSV files will need to be joined to combine data for comparison. In other words, csv files will need to be joined.

- When joining csv files for different stocks, column names will need to be modified to prevent duplication of column titles. To avoid processing errors, columns will be renamed to the stock ticker. E.g. "Adj Close" -> "GOOG". Updating column names will avoid overlapping of column names during csv/table joins.

- To prevent duplicating code, a utility function will need to be built to process all csv stock files in an efficient manner.

- abnormality -> dividends and stock splits. Need to use adjusted close. Historical data gets adjusted for this purpose.

- To be able to compare stocks, the stock data will need to be normalized to view the differences in performance through time.

### Exploratory Visualization

Below is a series of files that visualized the data is various ways:

(note: comments are included in the ipynb files)

1) [160\_exploratory\_visualization.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/160_exploratory_visualization.ipynb)

- Visuals in this file display normalized data in comparison to the benchmark stock (SPY)

- The visuals quickly show which stocks performed better (or worse) than the benchmark stock (SPY)

2) [170\_bollinger\_bands.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/170_bollinger_bands.ipynb)

- Visuals in this file display the famous bollinger bands TM for a selected number of stocks

- The space between the upper and lower bands indicate the amount of risk (aka standard deviation) in a stocks movement

3) [180\_daily\_returns.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/180_daily_returns.ipynb)

- Visuals in this file show the daily returns of a stock in comparison to the benchmark stock (SPY)

- The visuals also display how much a stock moves along (or against) the movement of the benchmark stock (SPY)

4) [200\_fill\_missing\_values.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/200_fill_missing_values.ipynb)

- The visuals in this file show the gaps in trading for stocks that are not traded every day

- Forward filling and back filling techniques were applied to be able to compare, as best as possible, against other stocks traded on a daily basis

5) [210\_plot\_histograms.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/210_plot_histograms.ipynb)

- Another way to visualize the data was to display the stock with histograms

- In this file, daily returns were compared to the benchmark stock, SPY, in the same graph

6) [220\_scatterplots\_beta\_alpha\_and\_correlation.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/220_scatterplots_beta_alpha_and_correlation.ipynb)

- Scatterplots were created to compare stocks to the benchmark stock (SPY)

- This file also captured the beta and alpha variables.

- The beta variable indicates how much more reactive the stock is to the market than the benchmark stock (SPY)

- the alpha variable indicates how well the stock performs with respect to the benchmark stock (SPY).

7) [280\_portfolio\_statistics\_visual.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/280_portfolio_statistics_visual.ipynb)

- The visuals in this file shows the performance of selected stocks as a portfolio

- This portfolio of stocks was then compared to the benchmark stock (SPY)

- The portfolio of stocks performed better than the SPY benchmark stock

8) [290\_portfolio\_allocation\_optimization.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/290_portfolio_allocation_optimization.ipynb)

- This visual shows the performance of a portfolio of stock after portfolio optimization

- Wow, what a difference. The optimization focuses on only two stocks in the portfolio of stocks (AMZN and AXY)

9) [300\_CAPM\_with\_optimized\_portfolio\_allocation.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/300_CAPM_with_optimized_portfolio_allocation.ipynb)

- This visual takes the data from the previous file and display the CAPM visual

- This file also captured the beta and alpha variables.

- The beta variable indicates how much more reactive the stock is to the market than the benchmark stock (SPY)

- the alpha variable indicates how well the stock performs with respect to the benchmark stock (SPY).

10) [310\_return\_vs\_risk\_scatterplot.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/310_return_vs_risk_scatterplot.ipynb)

- This visual displays a graph for the amount of risk vs return for a selected number of stocks

### Algorithms and Techniques

For this project, I will be using two supervised learning algorithms to "predict" stock prices:

1) sklearn's Linear Regression

2) sklearn's K Nearest Neighbors

### Benchmark

The benchmark for both the linear regression and knn regression algorithms will be a predictive score of 50%. If one of the algorithms return a score above 50% with unseen data, then this would indicate a better than chance indicator for predicting the future price of a stock.

## III. Methodology

### Data Preprocessing

As previously discussed, 12+ stocks have been selected for this project. The following pre-processing steps were taken:

- By default, the csv files are indexed by a number starting from "0". The code needed to explicitly identify the "Date" column as the index column.

- The data in the csv files begin with the latest trade day. The data needed to be re-organized to show a proper linear progression through time.

- Some of the stocks did not trade on a certain day(e.g. AXY). Stock with no trades on a specific day needed to a "nan" (or similar) inserted into the empty cell. Backfill and Forwardfill techniques were applied to the data for the ability to compare one stock to another.

- CSV files needed to be joined to combine data for comparison. A utility was created to join files efficiently.

- When joining csv files for different stocks, column names needed to be modified to prevent duplication of column titles. Columns were renamed to the stock ticker. E.g. "Adj Close" -> "GOOG". Updating column names avoided overlapping of column names during csv/table joins.

- To prevent duplicating code, a utility function needed to be built to process all csv stock files in an efficient manner.

You can also find additional data preprocessing comments in the ipynb files:

1) [160\_exploratory\_visualization.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/160_exploratory_visualization.ipynb)

2) [170\_bollinger\_bands.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/170_bollinger_bands.ipynb)

3) [180\_daily\_returns.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/180_daily_returns.ipynb)

4) [200\_fill\_missing\_values.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/200_fill_missing_values.ipynb)

5) [210\_plot\_histograms.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/210_plot_histograms.ipynb)

6) [220\_scatterplots\_beta\_alpha\_and\_correlation.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/220_scatterplots_beta_alpha_and_correlation.ipynb)

7) [280\_portfolio\_statistics\_visual.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/280_portfolio_statistics_visual.ipynb)

8) [290\_portfolio\_allocation\_optimization.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/290_portfolio_allocation_optimization.ipynb)

9) [300\_CAPM\_with\_optimized\_portfolio\_allocation.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/300_CAPM_with_optimized_portfolio_allocation.ipynb)

10) [310\_return\_vs\_risk\_scatterplot.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/310_return_vs_risk_scatterplot.ipynb)

### Implementation

*Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?*

I used two algorithms for predicting stock prices five days later. To do this, I copied over the adjusted price column in the csv file and "shifted" the column upwards 5 days (5 rows). I then had to manually split the data between the training group and the testing group as sklearn's train\_test\_split function did not have roll-forward cross-validation functionality. With the dataset split between training and testing, I applied two algorithms to the data; linear regression and knn regression. To my surprise, linear regression performed better than the knn regression algorithm. Linear regression received a score of 0.68 in the testing phase and knn regression received an abismal -22.54. I really thought knn regression would have performed better. Related code files:

1) [320\_supervised\_linear\_regression.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/320_supervised_linear_regression.ipynb)

2) [330\_supervised\_knn\_regression.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/330_supervised_knn_regression.ipynb)

*Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?*

yes, predicting time-series data without cheating requires chronological data to properly train and test the algorithm. using sklearn's train\_test\_split function, data in the training group had datapoints that were further into the future than data in the testing set. This means that the algorithm was able to peek into the future. With this setup, both algorithms generated a perfect score of 1 for predicting future prices. With this realization, I was forced to manually split the date with custom code.

*Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

All of my code is documented/captured in my github repository:

1) [320\_supervised\_linear\_regression.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/320_supervised_linear_regression.ipynb)

2) [330\_supervised\_knn\_regression.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/330_supervised_knn_regression.ipynb)

or

3) [Project Repository - top folder](https://github.com/carldgosselin/build-a-stock-price-indicator)

### Refinement

*Has an initial solution been found and clearly reported?*

The initial solution for knn regression had a k\_neighbors values of 10. This value produced a score of on the out-of-sample data of -22.71 (this is abismal, I know). Passing a number of k values through GridsearchCV recommended a k\_neighbors value of 25. This value produced a score on out-of-sample data of -22.54 (only marginally better). In conclusion, knn regression on this dataset is abismal.

*Is the process of improvement clearly documented, such as what techniques were used?*

I created a function call train\_knn to apply the GridsearchCV function to find the optimal values for 'n\_neighbors', 'leaf\_size', and 'weights'. GridsearchCV came back with the following best parameter: {'n\_neighbors': 25, 'weights': 'uniform', 'leaf\_size': 1}

Code:

1) [330\_supervised\_knn\_regression.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/330_supervised_knn_regression.ipynb)

*Are intermediate and final solutions clearly reported as the process is improved?*

n/a

## IV. Results

### Model Evaluation and Validation

*Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?*

I chose to apply both a linear regression and knn regression on stock data. The purpose of this exercise was to see the effectiveness of these two algorithms in predicting the stock price 5 days later.

For the linear regression model, no parameters were tweaked. I applied the linear regression algorithm from sklearn to the SPY stock data (as well as other stocks). The algorithm produced a score of 0.68 for the SPY stock. I visualized the predicted stock price vs the real-world price results in the out-of-sample (or test set):

1) [320\_supervised\_linear\_regression.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/320_supervised_linear_regression.ipynb)

For the knn regression model, I ran the k, leaves, and weights through the Gridsearchcv. As previously discussed, GridsearchCV came back with the following best parameter: {'n\_neighbors': 25, 'weights': 'uniform', 'leaf\_size': 1}:

1) [330\_supervised\_knn\_regression.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/330_supervised_knn_regression.ipynb)

*Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?*

I ran multiple stocks through the linear regression model. In conclusion, the linear regression model is superior to the knn model. However, I would personally not rely solely on the predictions of these algorithms for trading stocks. The models and the data would need to be improved drastically to be considered as a serious input for trading decisions.

*Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?*

No, unfortunately I find that both the linear regression and knn regression models are not robust models for trading. I would not rely on these predictions to make a decision on buying or selling stocks. I think that additional data will need to be added to the dataset to increase the strength of the predictions. I am ok with this as through this project, I realize that I am more of a fundamental analyst than a technical analyst. I will keep machine learning algorithms in my back pocket when researching stocks but it won't be the primary tool to drive any of my decisions. Instead, I prefer to rely on all of the visualizations in the data exploration section of this project for making decisions on trades.

Part of the reason for not using machine learning algorithms as a primary tool is that I find myself gravitating towards the Efficient Market Hypothesis where the price of shares always incorporate all relevant information in near real-time fashion. However, I realize that professional trading firms have their trades automated with extremely sophisticated machine learning tools that find market inconsistencies within milliseconds.

*Can results found from the model be trusted?*

No. They cannot. As discussed above. Although it's a tool I will continue to build and incorporate in some manner, I will rely more on fundamental analysis to make my decisions on buying and selling stocks.

### Justification

In this section, your model’s final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You

should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

*Are the final results found stronger than the benchmark result reported earlier?*

*Have you thoroughly analyzed and discussed the final solution?*

*Is the final solution significant enough to have solved the problem?*

with linear regression, I was able to beat the benchmark of 50%. In other words, the predictions produced by linear regression are better than chance. The following file displays a score of 0.68 (68%) for predicting the price of the SPY stock five days out:

1) [320\_supervised\_linear\_regression.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/320_supervised_linear_regression.ipynb)

On the other hand, the knn regression did not perform as well. For the SPY stock prediction, knn regression produced a score of -22.54:

1) [330\_supervised\_knn\_regression.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/330_supervised_knn_regression.ipynb)

## V. Conclusion

### Free-Form Visualization

Most, if not all, of the visualization produced are now invaluable to my research on stocks. I've already talked at length about all of my visualizations in one of the sections above. If I had to choose one, I would choose the first visualization that normalizes all stocks for an easy comparison on performance. This was the first "oh wow" moment. And the "oh wow" moments just kept building after each visualization.

You can view the "normalized" visuals in the following file in my github repository:

1) [160\_exploratory\_visualization.ipynb](https://github.com/carldgosselin/build-a-stock-price-indicator/blob/master/160_exploratory_visualization.ipynb)

### Reflection

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are

expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this

section:

*Have you thoroughly summarized the entire process you used for this project?*

I want to start off by saying that this was one of the most amazing educational journeys I've been on. At the beginning of this course, I never would have imagined how engaging this course would be. I had my doubts about how I would connect with other students and Udacity teachers. I must say that I felt more engaged than any of my courses that I've attended at University. I was also not enthusiastic of the forums for answering questions and helping others but, in the end, I found this medium to be a saviour. Hearing about other people's challenges, helping others and getting input from others gave me that extra push to move forward.

*Were there any interesting aspects of the project?*

Some of the most eye-opening moments in this project was the visualization during data exploration. Previously, my research on the stock market consisted of navigating to google.com/finance to review numbers and downloading financial reports of the stocks I was monitoring. Turning numbers into visualization and comparing my favorite stocks to the SPY stock gave me a whole new perspective on the companies I've been tracking.

Visualizing the normalized data for each stock and then comparing them to each other gave me a better view of which stocks where the most successful.

Basically, I found all of the visualization techniques in Tucker Balch's class really enlightening and useful: normalization, bollinger bands, daily returns, scatter plots, risks vs returns comparison... the list goes on.

*Were there any difficult aspects of the project?*

- I followed Tucker Balch's course on Machine Learning for Trading. As the course progressed, there was less and less support and guidance for the python code related to the concepts being taught. This was a challenge for me as it took days, if not weeks, to research and experiment with the code to see the concepts in action, such as the Sharpe Ratio, CAPM, etc...

In the end, however, I was able to push through and build the code that would create the desired results and visualization. I'm a better person for going through these challenges.

*Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?*

Yes, the model and solution does meet my expectations. Honestly, I'm just happy that I applied a machine learning algorithm to stock data. However, I do realize that using simple machine learning algorithms such as linear regression and knn by themselves won't make me a rich man. These may have worked in the 70s but I'm realizing that today's market is much too efficient to make money on basic machine learning algorithms. I think the main reason why simple algorithms doesn't cut it anymore is that too many people are looking at the data in the same manner. Therefore, if too many players are using the same strategies to win in a zero-sum game, that strategy becomes useless. In today's market, I would think that the market is won by the people, or companies, that collect and analyze data in unique and creative ways in addition to having the fastest connection and servers to the market.

### Improvement

The first thing I would improve in this project is adding additional data to increase the strength of the prediction. For example, I would experiment with additional data such as:

- P/E ratio

- Market capitalization

- Insider trading

- Range

- 52 week high and low

- looking at local weather where trades are made

I would also like to experiment with more creative data related to the people-side of the business. It would be great to interject data that related to how "happy" people are working in the organization or how much volunteering employees participate per year. Who knows, such stats may increase the prediction strength.

I would also like to apply Ensemble learners, boosting and bagging techniques (as discussed in Tucker Balch's Machine Learning for Trading course). From what I understand, Ensemble learning amalgamates different learning algorithms to increase the strength of a prediction. I would also like to further experiment with algorithms in the reinforcement learning, Q-Learning and Dyna space.

To answer your question about if even better solutions exists, I would imagine there are better solutions out there. I hear that most trades on the market are done by machines. I can images that these machines are using much more sophisticated algorithms than a simple linear regression or knn. Perhaps I'll get there someday.