Machine Learning Engineer Nanodegree

Capstone Project

Carlos R. Perez

**I. Definition**

(approx. 1-2 pages)

**Project Overview**

In this section, look to provide a high-level overview of the project in layman’s terms. Questions to ask yourself when writing this section:

Has an overview of the project been provided, such as the problem domain, project origin, and related datasets or input data?

Has enough background information been given so that an uninformed reader would understand the problem domain and following problem statement?

Investors and traders use a myriad of approaches to profit off the buying and selling of stocks and derivatives in public markets. These approaches include: (1) fundamental analysis, which relies upon the examination of fundamental business factors such as financial statements, (2) technical analysis, which focuses upon price trends and momentum, and (3) quantitative analysis, which may uses a combination of both. Traders constantly attempt to discover new strategies that exploit previously unknown patterns in an attempt to predict how a stock’s price will move.

Traders generally rely on information available to the public to develop their strategies. This information can include historical stock prices, financial numbers, quarterly statements, and news. However, public information seems to provide less value every day for the average trader, as sophisticated algorithms and high-frequency traders can react to new information in fractions of a second, moving prices faster than most can even digest the news.

An approach of recent interest is to attempt to infer information from trades by company insiders. People that run a company can reasonably be expected to best know how their company will perform in the future. In the U.S, insiders (i.e., directors, officers, or owners for 10% of equity) must report any personal trades of their company’s stock with the Securities and Exchange Commission (SEC) on a Form 4 filing. While insider-trading laws restrict how insiders can trade their company’s stock, recent research has concluded that insider-trading patterns can provide clues as to future corporate events that may trigger stock price fluctuations [1]. Also, company insiders may have better information not only about their own company, but also about companies in the same industry, particularly those in the same supply chain [2].

In this project, I attempted to use data extracted from SEC forms (Form 4) to predict significant stock price increases or drops. The input dataset describes how company insiders bought and sold stock over a period of time, and how the stock price fluctuated during the same period. I then trained machine learning models to predict stock prices changes based on new insider trades input.

**Problem Statement**

In this section, you will want to clearly define the problem that you are trying to solve, including the strategy (outline of tasks) you will use to achieve the desired solution. You should also thoroughly discuss what the intended solution will be for this problem. Questions to ask yourself when writing this section:

Is the problem statement clearly defined? Will the reader understand what you are expecting to solve?

Have you thoroughly discussed how you will attempt to solve the problem?

Is an anticipated solution clearly defined? Will the reader understand what results you are looking for?

While corporate insider trades may provide clues as to how insiders think their company stock will fare, it is challenging to weed out relevant trading events from noisy trading data. For example, insiders may sell stock on a regular basis simply out of a need for cash and not with any particular strategy in mind. Therefore, a machine learning approach may be useful in analyzing insider trading data to learn what types of trading activities may predict a significant corporate event and an associated sharp stock price change. In short, the problem to be solved is how to use Form 4 data to predict significant stock price increases or drops.

To solve this problem, I used a supervised learner that uses inputs from SEC forms (Form 4) and trained the model to classify stocks into class labels comprising ranges of stock price returns At a high level, the goal of this classifier was to accurately predict stock price changes given new Form 4 data.

Datasets were constructed solely from publicly available data, in particular, Form 4 data from SEC’s Edgar database and historical stock prices from Tiingo:

* EDGAR Database: <https://www.sec.gov/edgar/searchedgar/webusers.htm>
* Free Tiingo API: <https://api.tiingo.com/>

Data was pulled from these sources and parsed using Python SOAP libraries. I then constructed datasets for a supervised learner, with data extracted from Form 4 as input features and stock price percentage change as predicted output. To simplify the input and reduce the dimensionality, certain pre-processing was done. Below are the input features and output:

Input features

* Insider **Net Buy Count** of company stock per month for 12 months (12-tuple of continuous variables from -1 to 1, can be discretized in 0.05 intervals)
* Insider **Net Buy Volume** of company stock per month for 12 months (12-tuple of continuous variables, can be discretized in 0.05 intervals)
* Company Industry **Sector** (Healthcare, Industrial Goods, Services, Technology, Utilities, Basic Materials, Conglomerates, Consumer Goods, Financial)
* **Market Cap** (Small, Medium, Large)

Output feature

* **Stock return over S&P500** for the 6 months following the inputted 12 months, discretized into 3 categories, for example: (-∞,-10%], (-10%,10%), [10%,∞).

Each datapoint will thus correspond to a single company and a range of time.

The solution trains various supervised learning classifiers to classify stocks into one of the three stock return categories. I used Support Vector Machines, Naïve Bayes, and ensemble learners, using grid searches to obtain optimal parameters, as discussed below.

**Metrics**

In this section, you will need to clearly define the metrics or calculations you will use to measure performance of a model or result in your project. These calculations and metrics should be justified based on the characteristics of the problem and problem domain. Questions to ask yourself when writing this section:

Are the metrics you’ve chosen to measure the performance of your models clearly discussed and defined?

Have you provided reasonable justification for the metrics chosen based on the problem and solution?

The model was evaluated on its ability to correctly classify stocks. The data was divided into training and testing sets and cross-validation was performed. The performance of the model was then scored using various evaluation metrics: accuracy, F1-score, precision, and recall.

F1 score was the most useful metric, since it combines both precision and recall, which are better metrics for this scenario where the output is not uniformly distributed. Precision of a class will measure the number of stocks that were correctly classified into that class over the total number of stocks classified into that class. Recall of a class measures the number of stocks that were correctly classified into that class over the number of stocks that should have been classified into that class.

**II. Analysis**

(approx. 2-4 pages)

**Data Exploration**

In this section, you will be expected to analyze the data you are using for the problem. This data can either be in the form of a dataset (or datasets), input data (or input files), or even an environment. The type of data should be thoroughly described and, if possible, have basic statistics and information presented (such as discussion of input features or defining characteristics about the input or environment). Any abnormalities or interesting qualities about the data that may need to be addressed have been identified (such as features that need to be transformed or the possibility of outliers). Questions to ask yourself when writing this section:

If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader? CHECK

If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed?

Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)

*Dataset Description*

The dataset consists of data from Jan-2010 through Dec-2016. Each datapoint is associated with a single stock and its price return from a first date to six (6) months thereafter (e.g., AAPL, January 1, 2011 through June 30, 2011). The target variable is the stock price return relative to the S&P500 index for that 6-month period, computed as:

Adjusted prices were used for stock prices, and adjusted prices of SPY fund (trust ETF that tracks the S&P500). In this manner, the feature data is being used to predict the future return of the stock (i.e., the following 6 months).

The feature set includes the monthly Net Buy Count (NBC1-12) and Net Buy Volume (NBV1-12) for each of the twelve months prior to the 6-month period of stock price return data (e.g., January 1, 2010 through December 31, 2010 for the above example). Net Buy Count reflects ratio of the number of stock purchases vs. stock sales in a month:

Net Buy Count reflects ratio of the volume of stock purchased vs. stock sold in a month:

The feature set also includes the Industry Sector associated with the stock (out of 10 categories) and the market capitalization category (small, medium, or large). The target column is the return for the stock from the first day of the 6-month period to the last day of the period, computed as (Price Last Day – Price First Day) / Price First Day. Below is a sample of 5 datapoints:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| NBC1 | NBC2 | NBC3 | ... | NBV11 | NBV12 | SECTOR | MKT  CAP | RETURN |
| 1.000000 | 0.066667 | -0.428571 | ... | -0.190212 | -0.183784 | Information Technology | small | -0.038247887838 |
| 0.066667 | 0.000000 | -0.217391 | ... | -0.183784 | -0.139860 | Information Technology | small | 0.0707736976746 |
| 0.000000 | 0.000000 | -0.578947 | ... | -0.139860 | -0.007563 | Information Technology | small | -0.0917734937395 |
| 0.000000 | 0.000000 | -0.428571 | ... | -0.007563 | 0.006516 | Information Technology | small | 0.126127808822 |
| 0.000000 | 0.000000 | 0.555556 | ... | 0.006516 | -0.072261 | Information Technology | small | -0.131451077943 |

*Dataset Statistics*

In order to balance the target categories, I calculated some simple statistics for the return prices, shown below.

|  |  |
| --- | --- |
| Min return | -0.867926289208 |
| Max return | 6.66432449583 |
| Mean return | 0.0138842243279 |
| Median return | 0.00513197463559 |
| Std Dev of return | 0.197127084811 |
| First quartile | -0.0916451313919 |
| Third quartile | 0.105930831069 |
| First third | -0.0581371249263 |
| Second third | 0.0647762819449 |

In order to create balanced classes, I used the first third and second third as the cutoff point for the 3 categories. Therefore, based on the observed statistics, the target classes were defined as follows:

* Class 1 (“below”): Under -0.05813 return
* Class 2 (“middle”): Between -0.05813 and 0.06477
* Class 3 (“above”): Above 0.06477

Histogram analysis for outliers.

**Exploratory Visualization**

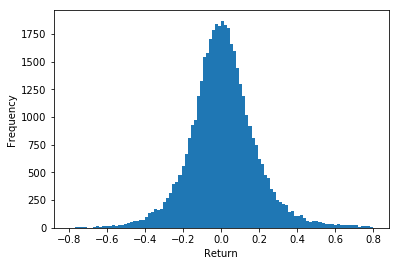
In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

Have you visualized a relevant characteristic or feature about the dataset or input data?

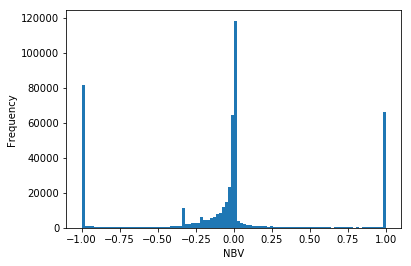
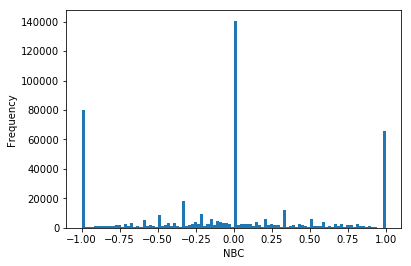
Is the visualization thoroughly analyzed and discussed?

If a plot is provided, are the axes, title, and datum clearly defined?

The following histogram shows the frequency of relative price returns over the entire dataset. Visual inspection shows that the returns are normally distributed around the mean.

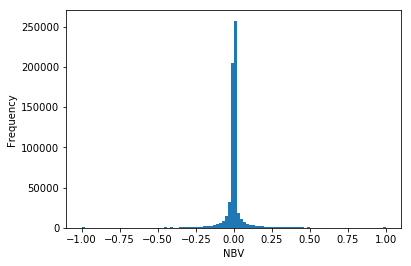
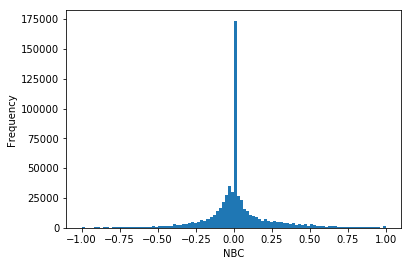


The histograms below show the frequency of values for Net Buy Count and Net Buy Volumes over all months in the dataset. The graph reveals that for a large majority of months the NBC and NBV were 0.0, indicated that for a large number of months there were no insider purchases or sales of stock. The charts also show that a disproportionate number of months have NBC and NBV values of 1.0 and -1.0, indicating that for a large number of months insiders either exclusively bought (1.0) or exclusively sold (-1.0) stocks.



This indicates a problem with the dataset where the metrics don’t yield as much information as they could. For example, a month where 1 stock was bought and a month were 1,000 stocks were bought may both have a NBV value of 1.0. As such, I decided to re-normalize the NBC and NBV values as a fraction of the maximum NBC and NBV for a given company:

*Max(Numpurchases)* represents number of stock purchases for the month with largest number of stock purchases for that company in the entire dataset. Likewise, *Max(Numsales)* represents number of stock sales for the month with largest number of stock sales for that company in the entire dataset. *Max(Volpurchases)* and *Max(Volsales)* represents the same quantities for the number of stocks purchased and sold, respectively. As such, *NBCnorm* and *NBVnorm* will also be between -1.0 and 1.0. The graph below shows the values are better distributed than with the previous normalization:



**Algorithms and Techniques**

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

Are the algorithms you will use, including any default variables/parameters in the project clearly defined?

Are the techniques to be used thoroughly discussed and justified?

Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?

I trained two classifiers to predict return categories: Naïve Bayes and an Ensemble of weak decision tree classifiers. Both classifiers are adequate to solve the problem here. While I tried Support Vector Machines as well, SVM’s running time was too long to perform adequate grid search and cross validation, and performance was similar to the other models with the few parameters I tested.

* Naïve Bayes: NB assumes each pair of features is independent of each other, which is not the case here. However, NB often still works even when this assumption is not true. NB is also very efficient, which is advantageous for a large dataset like the one used here.
* Ensemble Methods: Ensemble methods work well with a wide array of data. They can represent very complex hypotheses while being less prone than other models to over-fitting, and are more efficient than SVMs.

In order to determine the best parameters for each model, I used grid search cross-validation.

Benchmark

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

Has some result or value been provided that acts as a benchmark for measuring performance?

Is it clear how this result or value was obtained (whether by data or by hypothesis)?

Benchmark stuff

III. Methodology

(approx. 3-5 pages)

Data Preprocessing

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?

Based on the Data Exploration section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?

If no preprocessing is needed, has it been made clear why?

Preprocessing functions

Implementation

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

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

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

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

\*\*\*\*\*\*\*

Refinement

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

Has an initial solution been found and clearly reported?

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

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

\*\*\*\*\*\*\*

IV. Results

(approx. 2-3 pages)

Model Evaluation and Validation

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model’s solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

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

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

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

Can results found from the model be trusted?

\*\*\*\*\*\*\*\*\*

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?

\*\*\*

V. Conclusion

(approx. 1-2 pages)

Free-Form Visualization

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

Have you visualized a relevant or important quality about the problem, dataset, input data, or results?

Is the visualization thoroughly analyzed and discussed?

If a plot is provided, are the axes, title, and datum clearly defined?

\*\*\*\*\*\*\*\*\*

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?

Were there any interesting aspects of the project?

Were there any difficult aspects of the project?

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?

\*\*\*\*\*\*\*\*\*

Improvement

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

Are there further improvements that could be made on the algorithms or techniques you used in this project?

Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?

If you used your final solution as the new benchmark, do you think an even better solution exists?

\*\*\*\*\*\*\*\*\*

Before submitting, ask yourself. . .

Does the project report you’ve written follow a well-organized structure similar to that of the project template?

Is each section (particularly Analysis and Methodology) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?

Would the intended audience of your project be able to understand your analysis, methods, and results?

Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?

Are all the resources used for this project correctly cited and referenced?

Is the code that implements your solution easily readable and properly commented?

Does the code execute without error and produce results similar to those reported?