# Machine Learning in Stock Price Trend Forecasting

Yuqing Dai, Yuning Zhang

yuqingd@stanford.edu, zyn@stanford.edu

#### I. INTRODUCTION

Predicting the stock price trend by interpreting the seemly chaotic market data has always been an attractive topic to both investors and researchers. Among those popular methods that have been employed, Machine Learning techniques are very popular due to the capacity of identifying stock trend from massive amounts of data that capture the underlying stock price dynamics. In this project, we applied supervised learning methods to stock price trend forecasting.

According to market efficiency theory, US stock market is semi-strong efficient market, which means all public information is calculated into a stock's current share price, meaning that neither fundamental nor technical analysis can be used to achieve superior gains in a short-term (a day or a week). Indeed, our initial next-day predication has very low accuracy around 50%. However, as we tried to predict long-term stock price trend, our models achieved a high accuracy (79%). Based on our prediction result, we built a trading strategy on the stock, which significantly outran the stock performance itself.

#### II. IMPLEMETATION

## A. Data Collection

The training data used in our project were collected from Bloomberg Database. In this project, we picked 3M Stock to apply our method. The data contains daily stock information ranging from 1/9/2008 to 11/8/2013 (1471 data points). There are 16 features that we can use to apply our learning theory. In addition, we used the daily labeling as follows: label "1" if the closing price is higher than that of the previous day. Otherwise label "-1". For example, if the closing price of stock A on 11/11/2013 is higher than that on 11/10/2013, and on 11/10/2013, the PE ratio, PX volume, PX ebitda,...,S&P 500 index are  $X_1$ ,  $X_2$ ,..., $X_{15}$ , so the training data of A on 11/10/2013 is (X, Y), where  $X = (X_1, X_2,...,X_{15})^T$ , Y = (+1)

Stock	3M Co (NYSE: MMM)
Features	PE ratio, PX volume, PX ebitda, current enterprise value, 2-day

	net price change, 10-day volatility, 50-day moving average, 10-day moving average, quick ratio, alpha overridable, alpha for beta pm beta raw overridable, risk premium, IS EPS, and corresponding		
	S&P 500 index		
Data Source	Bloomberg Data Terminal		

#### B. Model Selection

#### 1. Next-Day Model

In our project, we mainly applied supervised learning theories, i.e. Logistic Regression, Gaussian Discriminant Analysis, Quadratic Discriminant Analysis, and SVM. The most important result that we should watch closely is the accuracy of prediction, which we define as follows:

# $= \frac{the \ number \ of \ days \ that \ model \ correctly \ classified \ the \ testing \ data}{total \ number \ of \ testing \ days}$

We used 70% of the data sets as training data and tested our fitted models with the remaining 30% data sets.

Model	Logistic	GDA	QDA	SVM
	Regression			
Accuracy	44.5%	46.4%	58.2%	55.2%

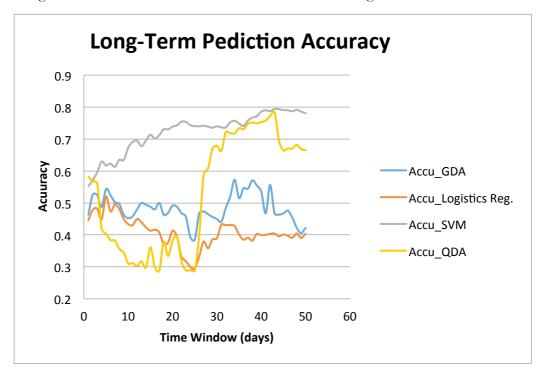
It turned out that the next-day prediction has a very low accuracy with the highest accuracy (QDA) being only 58.2%. We know that by flipping a coin we can probably get an accuracy of roughly 50% since the investing decision is binomial. Such result can be explained by the semi-strong efficient market theory, which states that all public information is calculated into a stock's current share price, meaning that neither fundamental nor technical analysis can be used to achieve superior gains.

## 2. Long-Term Model

Although our next-day prediction isn't very positive, we believe the financial data of a particular stock can still provide some insights for the stock's future movement. After all, that' why so many financial institutions/individual investors believe their work is meaningful.

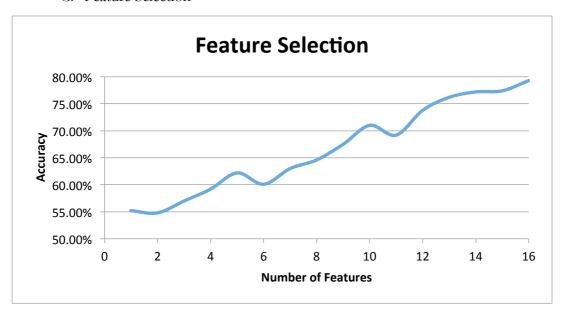
Especially, we think sometimes because of the existence of market sentiment, some information will not be reflected in the stock price immediately. Besides, in the eyes of investors, we also care about the predictions results of longer term to design our long-term investment strategy.

Here, we define our problem as predicting the sign of difference between tomorrow's stock price and that of certain days ago. Again, we used 70% of the data set as training data and tested our fitted models with the remaining 30% data sets.



From the chart, we can see that for SVM and QDA model, the accuracy increases when the time window increases. Furthermore, SVM gives the highest accuracy when the time window is 44 days (79.3%). It's also the most stable model.

## C. Feature Selection



From the chart established by backward stepwise feature selection, we can see that when we used all of the 16 features, we get our highest prediction accuracy. It makes sense because with over 1400 data points but only 16 features, there's no need to reduce

the number of features.

#### III. TRADING STRATEGY

#### A. Predictor Characteristics

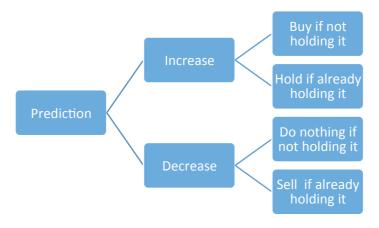
From the previous analysis, we have already determined that the best predicting model for 3M stock is SVM model. Here we use SVM as our predictor in order to develop our trading strategy.

Predictor	SVM
Kernel	Polynomial
Number of Features	16
Time Window	44 days

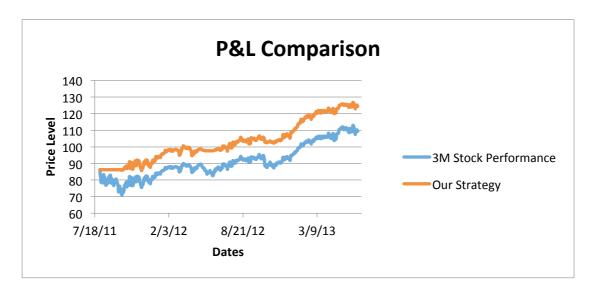
## B. Strategy Implementation

Initially, we used 990 of our 1470 data points to fit our model. Then we used our model to predict the stock price and made according investment decision on an on-time basis, meaning we will take in new information and update our predictor every trading date. Our back-testing of the strategy is over the course of December 2011 to October 2013.

On each day of the beginning 44 days, we will make a decision whether to buy the stock or not based on our prediction of whether the stock price would go up after 44 days. After the first 44 days, on each day we will make an investment decision again. It's better illustrated in the following decision tree:



Equivalently, we can interpret the strategy as if there are 44 traders. Trader i is responsible for trading his portfolio on i, 44+i,..., 44n + i,...day. Traders are independent to each other.



From the plot above, it's obvious that our strategy has outrun the performance of the stock, with an annualized return 19.3% vs. 12.5%.

## IV. CONCLUSION

In this project, we applied supervised learning techniques in predicting the stock price trend of a single stock. Our finds can be summarized into three aspects:

- 1. Various supervised learning models have been used for the prediction and we found that SVM model can provide the highest predicting accuracy (79%), as we predict the stock price trend in a long-term basis (44 days).
- 2. Our feature selection analysis indicates that when use all of the 16 features, we will get the highest accuracy. That's because the number of data points is much bigger than that of the features.
- 3. The trading strategy based on our prediction achieves very positive results by significantly outrunning the stock performance.

As for our future work, we believe we can make the following improvements:

- 1. Test our predictor on different stocks to see its robustness. Try to develop a "more general" predictor for the stock market.
- 2. Construct a portfolio of multiple stocks in order to diversify the risk. Take transaction cost into account when evaluating strategy's effectiveness