Predicting Which Stocks Will Beat the Market

Junlin Liu, Yongshang Wu, Hao Wang

{junlin93, wuy, wanghao}@stanford.edu | Computer Science Department, Stanford University

Abstract

In stock markets, there are usually two types of measurements for a stock or a portfolio: absolute return and relative return. Sometimes, it is hard for investors to make absolute return in a bear market like in 2008. On the other hand, in a long period of bull market such as that from 2009 to 2017, most investors can make some absolute return. Under such circumstances, relative return becomes more important to judge whether or not an investor is successful or a portfolio manager is good.

If we can predict which stocks can beat the market using machine learning, it will be easier for us to do better than other investors! In this project, we try to predict whether or not a stock can beat the market next 1, 2 and 4 quarters, and finally we make it! The most different thing of our project from others is that we predict future returns of stocks not only on past prices, but also on fundamental data of companies.

Data

- **Stocks**: All 30 constituent stocks of Dow Jones Industrial Average Index
- **Time Range**: From 2013 Q4 to 2017 Q2 (15 quarters)
- **Content**: 1) Historical price information; 2) Fundamental data (e.g., revenue, operation income, dividend of the company)
- **Source**: FactSet(a financial information provider)

Features

Basic features

1. return of the market in recent 1, 2, 3 and 4 quarters
2. return of the stock in recent 1, 2, 3 and 4 quarters
3. variation from the highest price in recent 1, 2, 3 and 4 quarters
4. variation from the lowest price in recent 1, 2, 3 and 4 quarters
5. revenue growth of the company in recent 1, 2, 3 and 4 quarters
6. operation income growth of the company in recent 1, 2, 3 and 4 quarters
7. net income growth of the company in recent 1, 2, 3 and 4 quarters
8. by how much earnings of the company beat analysts' expectation in recent 1, 2, 3 and 4 quarters
9. dividend ratio of the company in recent 1, 2, 3 and 4 quarters
10. dividend growth of the company in recent 1, 2, 3 and 4 quarters

Advanced features

Another set of features we use is built upon the set of basic features. For each feature x, we add feature

$$f(x) = \frac{1}{x}.$$

Then for each two features x_1 and x_2, we add feature

$$g(x_1, x_2) = x_1 x_2.$$

Feature Selection

In our project, we also use feature selection to improve results. Two feature selection methods we use are:

1. Wrapper (backward search)
2. Filter (pick the most informative features by feature importance)

Results

Basic features and feature selection

In the first experiment, we fit two models on basic features. Then we apply feature selection (Wrapper) to see if the results becomes better. The table below show our result. The number outside of the parenthesis is the average accuracy while that in it is the maximum accuracy by simple parameter tuning.

	Logistic Regression		Gradient Tree Boosting	
	Before	After	Before	After
1 Quarter	0.49 (0.61)	0.51 (0.61)	0.52 (0.63)	0.55 (0.65)
2 Quarters	0.50 (0.63)	0.50 (0.67)	0.53 (0.65)	0.55 (0.71)
4 Quarters	0.56 (0.69)	0.54 (0.64)	0.55 (0.69)	0.57 (0.76)

Table 1: Result for basic features and feature selection

Performance of different models:

We tried a variety of models to see which one has the best performance. From the figures below, we see that Gradient Tree Boosting is the best model in terms of our project.

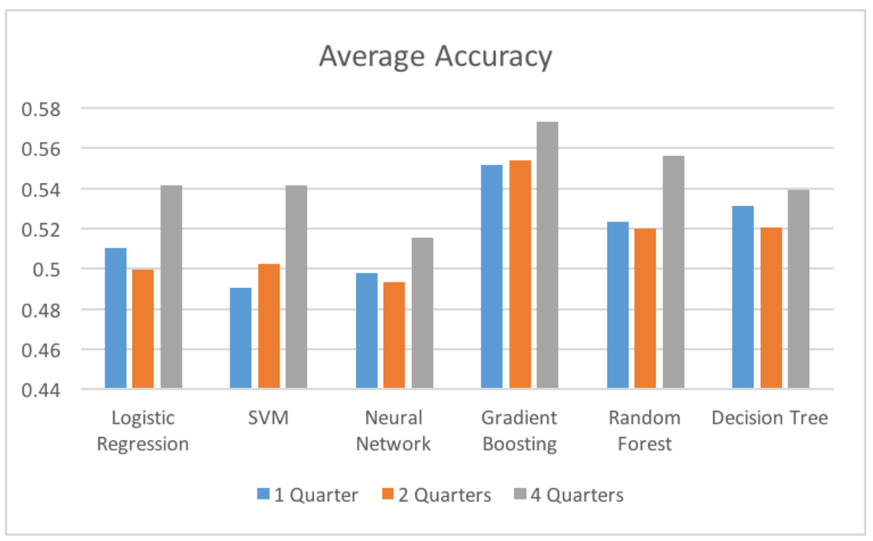

Figure 1: Average accuracy

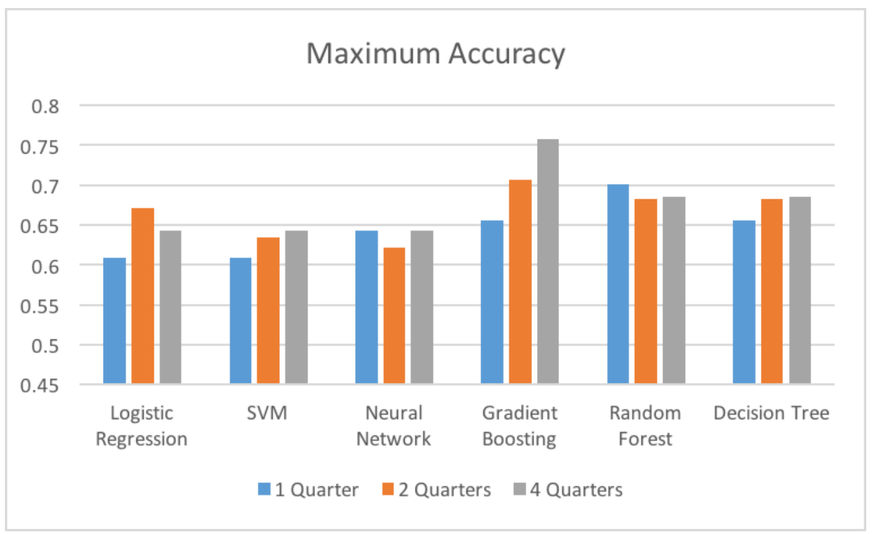

Figure 2: Maximum accuracy

Advanced features

After a chunk of experiments, we still can not get better performance. We begin to think of more advanced features. By extracting new features, we get a set of more complicated features. By fitting Gradient Tree Boosting on new features, we obtain very good results! We tried two different ways. The first is performing feature selection (Wrapper) first and then extracting advanced features. The second is extracting advanced features first and then performing features selection (Filter). It turns out that the second way leads us to goal.

	Origin	Method 1	Method 2
1 Quarter	0.52 (0.63)	0.54 (0.67)	**0.67 (0.80)**
2 Quarters	0.53 (0.65)	0.52 (0.63)	**0.70 (0.82)**
4 Quarters	0.55 (0.69)	0.55 (0.71)	**0.72 (0.86)**

Table 2: Results after adding advanced features