# Learning to Identify Winning Stocks

Avinash Balachandran, Durgesh Saraph and Erjie Ang December 11, 2013

## 1 Introduction

Value investing is an approach to investment decisions that relies on buying stocks that trade at a market value below their intrinsic value and selling stocks that trade at a market value above their intrinsic value. Benjamin Graham and David Dodd were early proponents of the paradigm, and it was made famous by Warren Buffett. The intrinsic value of a company's stock is the value determined by estimating the expected future cash flows of a stock and discounting them to the present. This value is represented as the book value on financial statements. It is distinct from the market value of the stock, which is determined by the company's stock price. The market value of a stock can deviate from the intrinsic value due to reasons unrelated to the company's fundamental operations, such as market sentiment.

One major risk involved in value investing is the existence of value traps in the market. A value trap is a stock that trades at a discount to its intrinsic value due to significant strategic or operational weaknesses in the company. For example, a company trading at a relatively low market value may be unable to grow its revenues, compete effectively in its target market, or operate efficiently. It is easy for investors to confuse value traps for genuinely promising investments, because like value stocks, a value trap also trades at a market value that is below its intrinsic value, i.e. they have a high book-to-market (BM) ratio.

To mitigate the risk of value traps, investors can use accounting-based fundamental analysis to select stocks from financially strong companies. Piotroski [2000] finds that the mean annual investment return on a portfolio of companies with high BM ratio can be increased by at least 7.5% using a simple nine-point score of financial strength calculated based on accounting fundamentals. The choice of portfolio mix is determined by using Piotroski's F-score, which assigns a company one point for nine criteria related to profitability, leverage/liquidity, and operating efficiency that can be derived from the balance sheet.

In this project, we propose the use of more advanced machine learning techniques to develop a predictive model for choosing stocks that improves on the widely used F-score strategy. Our key response variable of interest is the return on a portfolio of value stocks, which is the same as in Piotroski [2000]. In our model, we will use continuous versions of the nine variables from the F-score as predictors to account for more robust relationships between the predictors and portfolio returns. In addition to the F-score variables, we used the levels and changes in financial statement variables related to profitability, leverage, and liquidity that we believed to be predictive of future returns. Specifically, we include additional measures of profitability such as gross margin, sales turnover, cash balance, liabilities, holdings and sale of property and investment and accounts receivable. Our machine learning methods will also be calibrated using historical financial data as well as returns, rather than only using current and previous year data like in the F-score strategy. We find that both the use of historical data for calibration and the inclusion of additional variables substantially increase the portfolio returns from our trading strategy.

The rest of this report is arranged as follows. Section 2 will describe the financial data sources used. Sections 3 and 4 will describe the F-score strategy and how it was used to construct the porfolios and annual returns that will serve as our benchmark. Section 5 describes how we construct training and test sets for calibrating and testing our machine learning methods. We conclude and discuss further extensions of this work in Section 7.

## 2 Data and Variable Selection

Our data is obtained from the Compustat and CRSP financial databases that are available on Bloomberg machines found on campus. Compustat contains all the quarterly financial reports of publicly traded firms in the US. The CRSP data set contains the daily closing price of all publicly listed firms. Our data set comprises of the quarterly financial reports and daily closing prices of all publicly listed firms between the years of 1977 to 2010 inclusive.

We obtain quarterly financial statement data for publicly traded firms from Compustat and data on stock returns and trading volume from the CRSP database. Our data sample comprises of the quarterly financial reports and daily closing prices of all publicly listed firms between the years of 1977 to 2010 inclusive.

We run Principal Component Analysis (PCA) on the levels and changes of all variables in our sample. We find that the first two PCs explain only 45% and the first ten PCs explain 70% of the variation, and we decide not to use PCA in our analysis.

## 3 F-Score Strategy

Using the data set described in Section 2, we can recreate the F-score strategy as follows. Using only the last quarter in a firm's financial year in our data set. The Fscore assigns a company one point for each of the following nine criteria related to profitability, leverage/liquidity, and operating efficiency: positive return on assets (ROA) in the current year, positive operating cash flow (CFO) in the current year, higher return on assets (ROA) in the current year compared to the previous year, CFO being greater than ROA, lower ratio of long-term debt to total assets in the current year than the prior year, higher current ratio this year than the previous year, no new share issuance in the previous year, higher gross margin this year than the previous year, no new share issuance over the year, higher gross margin this year than the previous year, and higher asset turnover compared to the previous year.

# 4 Creating a Portfolio

In order to test the efficacy of the machine learning approach as compared to the F-score strategy, a metric is needed to evaluate each approach. The most intuitive metric to use is yearly portfolio return. In essence, an optimal trading portfolio is created for each method per year, and the portfolio return for that year is used as the evaluation metric.

Selecting the stock price that is used to build this portfolio is challenging as all firms do not release their quarterly reports at the same time for the same quarter. Therefore, the stock price used to calculate quantities like BM must take into account the quarterly report release dates. For example, we cannot build our portfolio using the stock price on 20 December 2009 while calculating an F-score based on a quarterly report released on 1 Jan 2010 as investors will not have all the information required on 20 December 2009 to build their portfolio. In order to overcome this, we look specifically at the stock price of each company on April 30 of the year following the year when its F-score was calcu-

lated. For example, if we calculated an F-Score for IBM in December 1987, we will use the stock price of IBM in April 30, 1988. The reason we used this date is because by that time, previous year financial reports of all firms will be available to investors, and they can trade using information from these reports.

Since we only consider long positions in stocks, we begin building our portfolio by ranking firms by BM and only considering firms in the top quartile of this category. Within this high BM group, we pick the companies with the highest F-score to build our portfolio. Hence, we rank these firms by F-score and take the firms in the highest decile to form our portfolio. Based on the prices of these firms in the current year and the subsequent year, we calculate the return of the portfolio assuming equal investments in each company in the portfolio. This return is used as the metric to evaluate each method. The benchmark return used to compare the efficacy of the machine learning approach is the return obtained using the basic F-score strategy as outlined above.

## 5 Model Construction and Testing

To improve on the F-score strategy, which uses a limited set of binary variables, we will use prediction models that use real-valued variables as our predictors. In the first test, the prediction models are built using only real-valued versions of the nine F-score variables. In the second test, the predictors are expanded to include additional fundamental variables as presented in the previous section. The three prediction models used are linear regression, logistic regression and support vector machines (SVM).

Since we have cross-sectional and time-series panel data, we use a rolling window version of cross-validation to test our predictions. First, we choose the length of the training window, for example two years. This means that the prediction model is trained on two years of financial data. The trained model is then used to predict the the share price returns of firms in the year immediately following the training window. For example, if we use data from 1988 and 1989 for training, data from 1990 are used for prediction. In the next cross-validation step, we move our training window one year into the future to include 1989 and 1990. The trained model then uses 1991 data to generate predictions, so on and so forth. In total, we generated predictions for years ranging from 1982 to 2011 by moving our training and testing time periods across time. At the same time, the length of the training windows used are varied between one and four years.

In linear regression, the response variable is the return of

the stock over the following year. Our portfolio is formed by including only firms from the top decile of predicted returns. For logistic regression and SVM, all the response variables in our training sets are re-coded into a binary variable we call class. A value of 1 is assigned to class for a firm if the actual returns of that firm are in the top decile of the actual returns of all firms. Otherwise, class is coded as 0. Logistic regression and SVM models are trained using this re-coded variable as the response. The predictions output from these models are the probabilities of whether each firm belongs to the group of high-return firms. The portfolio for the test year is then formed by including only firms with predicted probabilities that are in the top decile amongst all firms available in that year.

#### 6 Results

As discussed in Section 5, we varied our portfolio construction method by varying the model used (linear regression, logistic regression, SVM), the length of training window used (1 to 4 years) and the number of variables used. Table 1 summarizes the mean and standard deviation of the predicted returns of the portfolio formed from 1982-2010. In general, we see that the portfolio returns in all our techniques have higher means than the F-score strategy (baseline). However, only the SVM techniques have standard deviations close to the F-score strategy. This means that only the SVM technique allows us to generate less volatile returns over the period of interest. Another trend is that by including more variables, the standard deviation in returns can be significantly reduced. Figure 1 shows the annual returns for all the methods using only F-score variables and a window of 1 year. Figure 2 shows the annual returns for all the methods using all variables and a window of 1 year. In the next few sections, we discuss the results generated by each machine learning model individually.

#### 6.1 Linear Regression

In general, linear regression resulted in mean returns that were higher than the F-score strategy (baseline). However, the standard deviation in these returns is much higher than the F-score strategy (baseline). This implies that while linear regression results in higher average returns, they are riskier strategies. We also investigated the change due to window length variation and data set size.

Effect of different window length: Figure 3 shows the annual returns for linear regression using only F-score variables with different window lengths ranging from 1 to 4 years. We see that using a longer window results in a worse prediction. This implies that the variables used in the F-score calculation have an expiration. Therefore, using vari-

ables more than 1 year in the past results in the addition of variables that are no longer valid to predict a company's return. This points to the fact that there might exist an "expiration date" of the usefulness of the 9 F-score variables for predicting returns. However, from Figure 4, we see the addition of more variables seem to counteract this expiration. this could mean that the additional variables could contain signals from earlier years that are not captured by the 9 F-score variables.

Effect of using more variables: Figure 5 shows the annual returns for the three different machine learning techniques used and for the two cases of using only F-score variables and using all the variables. We see that in the case of linear regression, using only F-score variables results in higher annual returns (also seen in Table 1). However, the standard deviation is higher indicating that using more

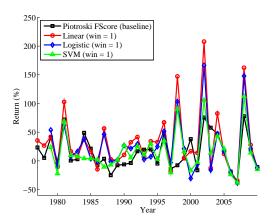


Figure 1: Annual Returns for all Techniques (F-score Variables)

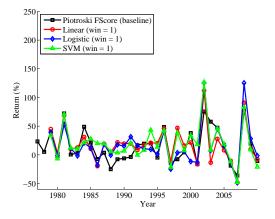


Figure 2: Annual Returns for all Techniques (All Variables)

| All Strategies   |            | F-score Variables |                  | All Variables |       |
|------------------|------------|-------------------|------------------|---------------|-------|
| Technique        | Window     | Mean              | $\mathbf{StDev}$ | Mean          | StDev |
| Linear           | 1          | 31.87             | 55.91            | 16.44         | 32.05 |
| Linear           | 2          | 22.83             | 54.23            | 17.71         | 38.36 |
| Linear           | 3          | 17.01             | 45.79            | 20.60         | 38.90 |
| Linear           | 4          | 22.61             | 50.51            | 24.88         | 54.63 |
| Logistic         | 1          | 24.10             | 48.13            | 15.60         | 37.48 |
| Logistic         | 2          | 21.59             | 49.88            | 17.29         | 32.21 |
| Logistic         | 3          | 23.26             | 54.03            | 21.11         | 39.22 |
| Logistic         | 4          | 24.43             | 54.79            | 25.09         | 48.36 |
| SVM              | 1          | 15.98             | 36.55            | 19.25         | 32.92 |
| SVM              | 2          | 17.95             | 30.50            | 15.86         | 30.48 |
| SVM              | 3          | 21.39             | 36.46            | 15.24         | 31.26 |
| SVM              | 4          | 23.52             | 40.26            | 21.76         | 42.23 |
| F-score strategy | (Baseline) | 12.51             | 29.50            | 12.51         | 29.50 |

Table 1: Mean Annual Portfolio Returns (%) and StDev for all strategies used

variables in the prediction results in a less risky portfolio building strategy.

#### 6.2 Logistic Classification

Logistic classification generally resulted in mean returns that were higher than the F-score strategy (baseline). However, similar to linear regression, the standard deviation in these returns are much higher than the F-score strategy (baseline).

Effect of different window length: For the cases where we only used the F-score variables, changing the window length does not have any clear trend on the returns (see Table 1). The variation in mean returns is under 2%. However, for the cases of using all variables, increasing the window length results in an increase of the mean return. Either way, the standard deviation based on the chosen technique is higher than the F-score strategy (baseline).

Effect of using more variables: In Figure 5, we see that in the case of logistic classification, using only F-score variables results in higher annual returns (also seen in Table 1). However, the standard deviation is higher indicating that using more variables in the prediction results in a less risky portfolio building strategy.

#### 6.3 Support Vector Machines

In general, SVM resulted in mean returns that were higher than the F-score strategy (baseline). However, unlike the other methods used, the standard deviations in return for the smaller window sizes were comparable to that of the F-score strategy (baseline).

Effect of different window length: Changing the window length for the case of using just the F-score variables

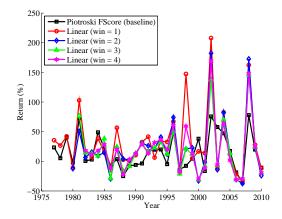


Figure 3: Annual Returns for all Windows (F-score Variables)

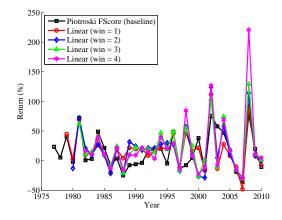


Figure 4: Annual Returns for all Windows (All Variables)

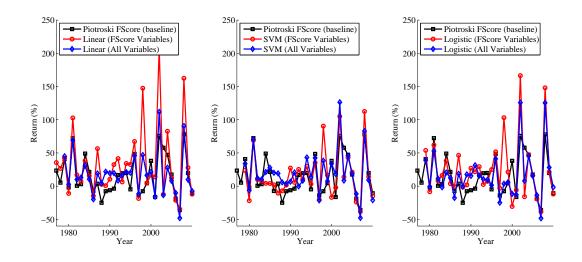


Figure 5: Annual Returns (F-score Variables vs All Variables)

and using all variables resulted in changes in the mean returns and the standard deviations. Increasing the window size resulted in higher mean returns. However, no clear trend is discernible between window size and standard deviation of returns.

Effect of using more variables: From Figure 5, we see that in the case of SVM, using either F-score variables or all variables results in comparable returns and standard deviations (also seen in Table 1). Therefore, the change in variable sets did not result in a big change in performance. One important thing to note is that using a window of 2 years, we were able to have a standard deviation comparable to the F-score strategy (baseline) while still having a higher mean return. This implies that this particular case actually beat the F-score strategy (baseline) mean return while having comparable risk.

## 7 Conclusion

Our results suggest that as expected, a strategy that gives higher mean returns also results in higher standard deviation in returns (or risk). Also, we note that including more variables in the learning algorithm results in lower standard deviation in the returns and hence results in a lower risk investment strategy. All our strategies have higher mean returns than the F-score strategy (baseline). However, most of them also have higher standard deviation. Two of our strategies, namely using an SVM with a window length of two years, result in better mean returns than the F-score strategy (baseline) while also having comparable standard deviations. Therefore, these two methods beat the F-score strategy (baseline). However, it is also important to note

that despite its relatively uncomplicated methodology, the F-score strategy (baseline) does very well. This is because it builds a strategy by analyzing the fundamental financial strengths of companies. Future work should focus on how to incorporate fundamental financial understanding of companies into the machine learning techniques proposed.

### 8 References

Piotroski, J.D., 2000. Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers. *Journal of Accounting Research* 38, pp1-41.