An Unostentatious Analysis of Stock Market Tax Filings and Positive Return on Investment

Agamdeep Singh Chopra, MBA Student, B.S. Physics

Stevens Institute of Technology
BIA 600
2020 Fall

Introduction

Stock market data is very chaotic for the untrained eye and can sometimes go against our intuitions and common sense. Investing big sums of money in the stock market can be very risky. But this risk can be minimized by analyzing the stock market data and making informed decisions. Today, many tools are available to the investors utilizing complex data analysis techniques and models to do such a thing. In this project, I attempted to develop a proof of concept for a similar tool to guide the investors to make better informed investments.

The primary goal of this project was to investigate if one can predict weather investing and holding a certain stock for one year will have positive cash inflow for the investor or not. This was done by analyzing a sample of income tax statements of publicly listed corporations being traded in the New York Stock Exchange.

The secondary goal of this project was to scrutinize the impact of the annual Lifestyle Satisfaction Scores on the performance of the individual stocks.

Dataset Explanation

The raw data contained 220 financial indicators (variables) and about 23000 examples ranging from the years 2014 to 2018 for all companies listed on the New York Stock Exchange. The data

was known to have errors and misreporting, for what the original data was cleaned using Rapid Miner to remove any empty rows, non-significant variables for model development, and repeated values giving about 22000 examples with 41 variable lists. This data was then standardized and further cleaned by selecting examples whose variables were within 1 standard deviation for that variable list giving about 16000 examples. This data was still very noisy, so a subset of only the data that showed some linear tendency with respect to the binary classification was extracted with about 9000 examples to develop a predictive model. This approach enabled development and training of models on a small sample size that could be quickly optimized and later scaled up to test against the entire 22000 examples set.

There were 5 key variables used in this experiment namely Asset Growth, Debt Growth, Receivables Growth, Book Value per Share Growth, and Inventories that were identified as the most significant variables in the 9000-example subset and were used to develop the predictive models. These variables were identified by running a Pearson Correlation test on the sample and got a score of (r (39) > 0.9, p < 0.01) against the binary expectation list.

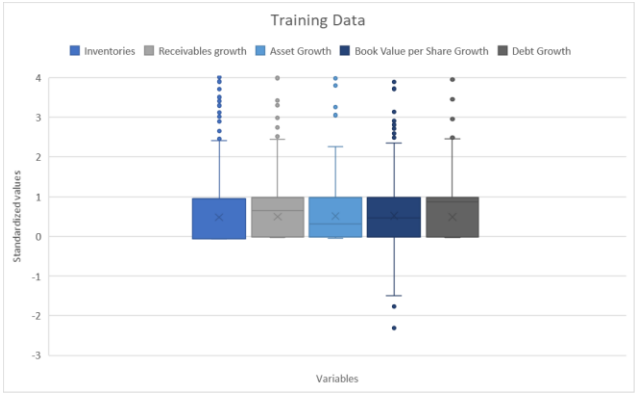

Figure 1. The graph represents the distribution of the key variables in the train set. As expected, most values fall in similar range after standardizing with most outliers in the positive region.

- **Asset Growth** is the degree at which a company's assets change in value over time. It was identified as an interval scale variable with no natural zero point. For the Training

set, (*M*= 0.51, *SD*= 1.78, *Kurtosis*= 6226.09) and for the test set, (*M*= 0.51, *SD*= 0.40, *Kurtosis*= -1.61).

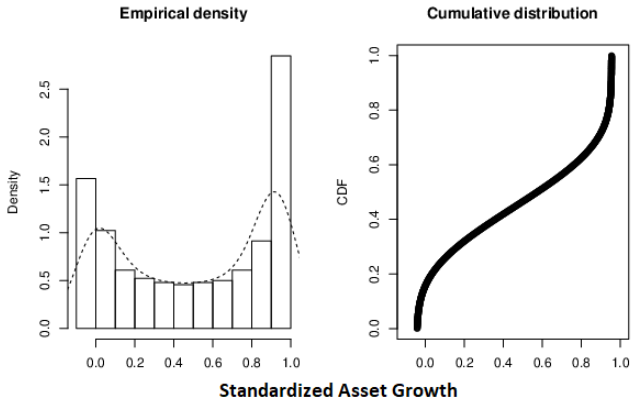

Fig 2. The graph represents the test set distribution of Asset Growth. It can be inferred from the graph that most values of Asset Growth lie between -0.2 and 1 after standardizing.

- **Debt Growth** is the amount of change in the debt financing of a company by issuing bonds, notes, and more. Higher value usually indicates expanding operations. It was identified as an interval scale with no natural zero point. For the Training set, (*M*= 0.49, *SD* = 0.53, *Kurtosis* = 16.47) and for the test set, (*M*= 0.50, *SD*= 0.50, *Kurtosis*= -1.88).
- **Receivables Growth** is the change in outstanding revenues of a company. i.e., Revenue that is not collected immediately as cash after a sale. This can be an indicator of future cash inflow. It was identified as an interval scale with no natural zero point. For the Training set, (*M* = 0.49, *SD* = 0.61, *Kurtosis* = 328.82) and for the test set, (*M*= 0.51, *SD*= 0.50, *Kurtosis*= -1.87).
- **Book Value per Share Growth (BVPSG)** is the change in the minimum value of a company's equity. Growth in the book value (assets – liabilities) per share. Undervalued stocks have a higher BVPS value. Investing in undervalued shares can lead to profit in time if there is a significant price appreciation. It was identified as an interval scale with

no natural zero point. For the Training set, (*M* = 0.52, *SD* = 1.76, *Kurtosis* = 3508.80) and for the test set, (*M*= 51, *SD*= 0.52, *Kurtosis*= -0.32)

- **Inventories** Cash value of any items or goods held by a company whose sales/liquidation will result in cash inflow. It was identified as a ratio scale with natural zero being no inventories on hand. For the Training set, (*M* = 0.48, *SD* = 0.53, *Kurtosis* = 2.29) and for the test set, (*M*= 0.50, *SD*= 0.52, *Kurtosis*= -0.40)
- **Lifestyle Satisfaction Score** An index on the scale of 1 to 10 that measures average happiness of a country based on various socioeconomic factors on an annual basis. It was identified as an ordinal scale. (*M* = 6.94, *SD* = 0.06, *Kurtosis* = 0.62)

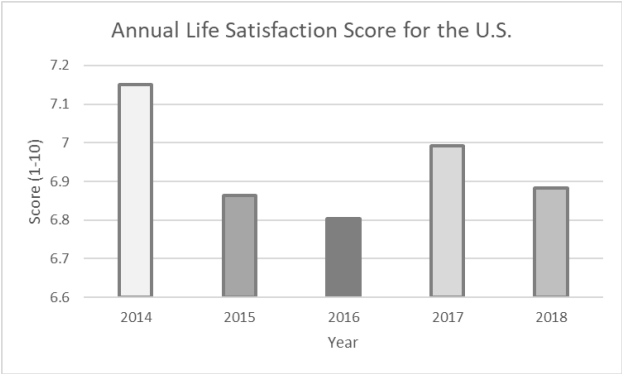

Fig 3. The graph shows the LS Score for the US between 2014 – 2018.

- **Binary Expectation Score** is the yes or no expected prediction outcome for the given example. It was identified as a Categorical variable with True = 1, and False = 0 binary values. A true value means that the stock will increase in value the next year and hence should be invested in. A False value will be the opposite.

Analytics

After conducting the Pearson Correlation test and compiling the 9000-example list “simplified set”, train and test sets were created to develop a predictive model. This was approached by

using the significant variables to develop a multiple regression model and code and optimize a Shallow and a Deep Machine Learning Model to see which of these approaches might be the best for the chaotic raw data.

I. Multiple Regression Model

Using the simplified set, multiple regression gave (99%) accuracy. It failed to produce a reliable prediction when applied to rest of the data. Book Value Per Share had a negative coefficient that goes against our expectations but according to the model it was not a significant variable to get an accurate prediction.

Variable Name	*Coefficient*	*p-value*	*Significant? (95% conf.)*	*Affect*
Inventories	0.383811574	< 0.001	Yes	Positive
Receivables growth	0.107929503	< 0.001	Yes	Positive
Asset Growth	0.003973607	2.86E-13	Yes	Positive
BVPSG	-0.000504733	0.36	No (greater than 0.05)	Negative
Debt Growth	0.452693293	< 0.001	Yes	Positive

Table 1. Multiple Regression Variable Summary

II. Shallow Neural Network Model

Using the simplified set, a shallow neural network algorithm was programed and optimized in Python. ADAM optimization, learning rate decay, input normalization, vectorization, Swish and Sigmoid functions were implemented in this model. The network was optimized to have 2 hidden layers with 10 and 5 neurons, respectively. The Shallow model gave (100%) accuracy on both training and test simplified sets. But the error increased significantly as rest of the data was included. The resulting accuracy dropped down to about (45%) for the entire dataset.

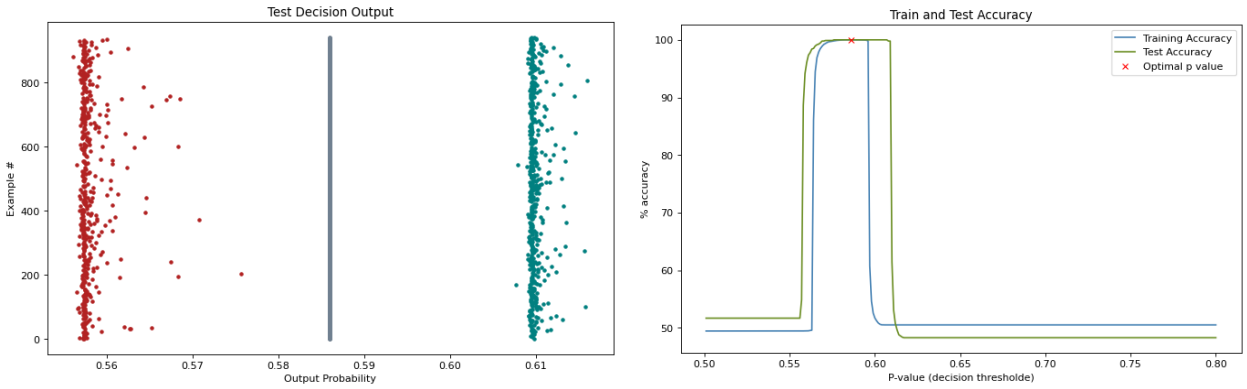

Fig 4.1. (left) The graph shows the test decision output of the model on the simplified set. Data on the left of the threshold is prediction of False and data on the right is True. Red represents the prediction was expected to be false and blue represents that it was expected to be true. The grey line represents the decision boundary/threshold.

Fig 4.2. (right) The graph helps optimize threshold value by showing accuracy as a function of the threshold.

III. Deep Neural Network Model

Using the simplified set, and modifying the shallow neural network script, a Deep learning model was developed. The model was optimized to have 10 hidden layers with 200, 200, 100, 100, 50, 150, 50, 50, 20, and 10 neurons, respectively. The model achieved an accuracy of (99%) on both the training and test simplified sets. Again, the accuracy dropped down significantly when applied to the entire data set to about (57%). But this time, the cost function had a significant negative slope even after long training at iteration 5000, indicating that if the model was let to run for a longer time, preferably days compared to overnight, the accuracy might be significantly higher. The Deep Neural Network model seems to be the best approach to create a predictive model from all 3 models that were tested on this dataset.

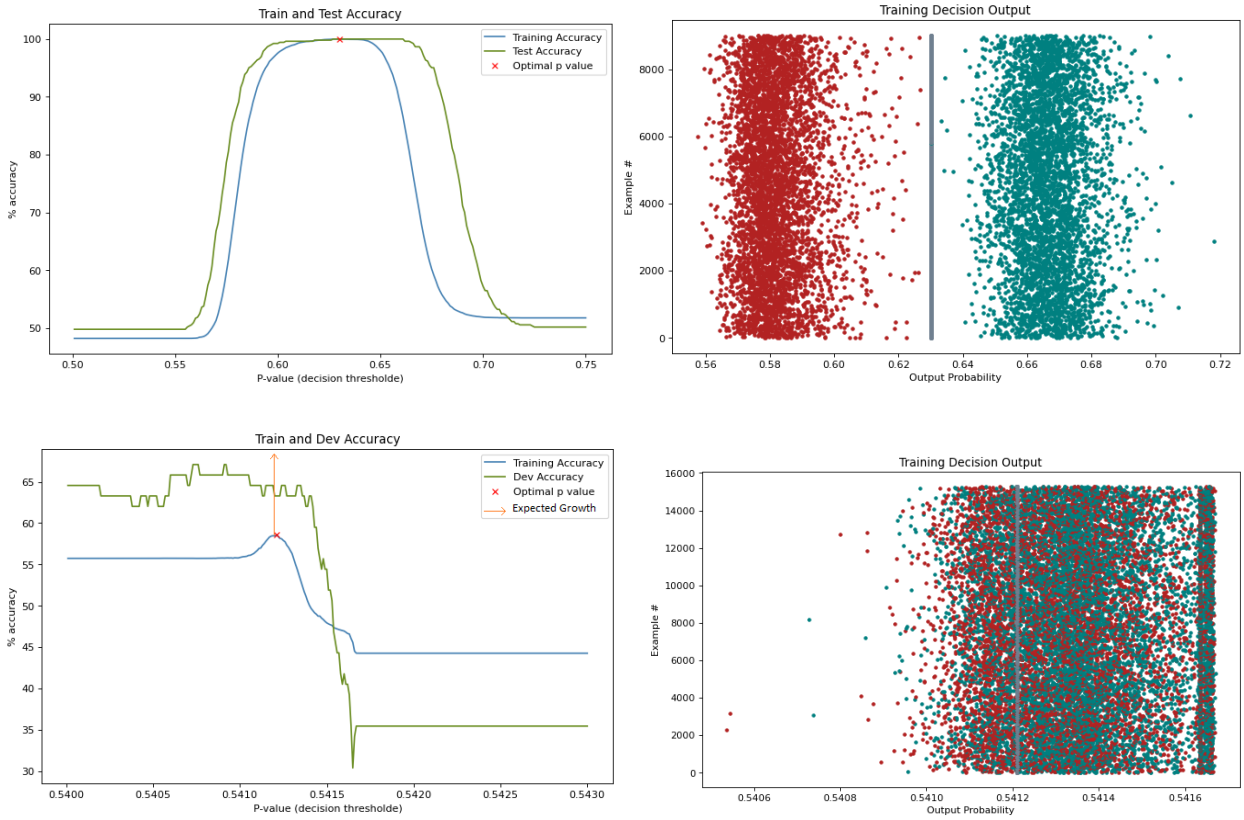

Fig 5.1. (Top Left) The graph shows accuracy as a function of decision boundary/threshold for the simplified set after 5000 iterations.

Fig 5.2. (Top Right) The graph shows the training decision output of the model on the simplified set. Data on the left of the threshold is prediction of False and data on the right is True. Red represents the prediction was expected to be false and blue represents that it was expected to be true. The grey line represents the decision boundary/threshold.

Fig 5.3. (Bottom Left) The graph shows accuracy as a function of decision boundary/threshold for the entire set after 5000 iterations. The orange line is the direction of expected growth of accuracy with more training time and/or algorithm and hyperparameter optimization.

Fig 5.4. (Bottom Right) The graph shows the training decision output of the model on the entire set. Note how the model misclassifies almost half the data. As runtime is increased, the model output starts to polarize more and more, and misclassification decreases as well.

IV. Lifestyle Satisfaction Score

Lifestyle satisfaction score had a Pearson Correlation value of (r (39) = 0.4, p < 0.01) which is statistically insignificant. None of the models appear to have any affect with the inclusion of this variable list. Hence, Lifestyle Satisfaction Score does not seem to be a good indicator for investing in stocks.

Interpretations and Recommendations

- From this analysis, it seems that Deep learning algorithm gave the best prediction for the chaotic stock market data. Longer training time and hyperparameter tuning might give higher accuracy. Better hardware and Neural Network architecture might significantly improve both the Shallow and Deep Neural Network Algorithms.
- Lifestyle Satisfaction score was not significant because it does not capture a companies' individual environment. It would be wise to use a per company metric for a better analysis in the future. For example, using a companies' employee rating from a job search website might be a more informative and accurate metric per company than Life Satisfaction score for the entire country.
- Multiple Regression and Shallow models worked better on data that showed linear tendency. This refers to variables whose change has a proportional change in the prediction outcome. The Deep Neural Network did perform good on both the chaotic and linear data, but the former models performed slightly better on the linear data with Shallow Neural Network being the most accurate and Regression being the fastest.
- Interestingly, it appears that the models perform better on companies that are in the process of rapid expansion or expecting exceptional growth. This can be inferred from the most significant variables as discussed earlier.