



Stock Market Prediction with Decision Trees:

A 3-Year Window

Final Report

Data Mining for Business, Fall 2020 (Professor Michael Deamer)

By:

Prannoiy Chandran

Marissa Maffa

Sebastian Carvajal

David Reyes

Abstract

The purpose of this analysis is to explore the efficacy of using decision trees to predict stock-related performance. The analysis will seek to predict whether or not stocks will outperform the market based on a miscellany of stock-related performance metrics and valuation methodologies. The time-frame for this analysis is a three year period between Dec 2016 - Dec 2019. Data from 2020 has been omitted due to the recent periods of elevated volatility associated with the pandemic. Accuracy will be measured in conjunction with other supervised models however, the decision tree model will be the focal point of our analysis given its descriptive qualities.

Introduction

Supervised models are powerful in their ability to predict outcomes. In a day and age, where being able to predict decisions based on a future outlook is of critical importance, technology serves a vital role. Supervised models are at the forefront of these developments and proving their efficacy is pivotal to their implementation. The application of these tools can be imperative to a wide-range of industries however, the investment industry is a key area of focus. In 2019 alone, the global asset management industry reached \$89 trillion dollars.

Data Description

Fundamental data is the linchpin of security analysis and is essential to the investment process regardless of style and risk level. The FactSet Fundamentals Database provides users a comprehensive global database of financial statement information. The FactSet Fundamentals collection methodology uses multiple methods to acquire filings for publicly traded companies. Sourcing 8-Ks, Business Wire, PRNewswire, Stock Exchanges, company websites, and other news sources for preliminary documents enables FactSet to capture information as soon as it is available.

FactSet Fundamentals collects 1000+ data items, including detailed financial statement content, per share data, and calculated ratios with deep history and broad global coverage. The sample universe we chose is publicly

traded companies from the U.S. (~9,000 entities). The Fundamental data we chose are annual year end valuations from 2011-2016 calendar year ends, and the Pricing data was collected on an annual basis from December 2016 to December 2019.

Data Item (Inputs)	Data Item Description	Statistic Date
Net Sales	Sum of a company's gross sales minus its returns, allowances, and discounts. It's one of the most important line items on the income statement	5Y growth, '11-16
Free Cash Flow	The cash available after accounting for cash outflows for operations and maintaining capital assets. It's a good starting point for shareholders to evaluate how likely the company will be able to pay their expected dividends or interest.	5Y growth, '11-16
Dividend Yield	Dividend income per share, divided by the price per share	5Y growth, '11-16
EBITDA	Earnings Before Interest, Taxes, Depreciation, and Amortization. EBITDA margins give investors a more accurate snapshot of a firm's operating profitability.	5Y growth, '11-16
Net Income Margin	Ratio of net income to revenues. It typically is expressed as a percentage but can also be represented in decimal form. Also called the net profit margin, it represents how much each dollar in revenue collected translates into profit. A higher profit margin is always desirable, but can vary by industry.	As of Dec 2016
Total Debt % Total Assets	Total amount of debt relative to assets owned by a company. It can reflect how financially stable a company is. The higher the ratio, the higher the degree of leverage and, consequently, the higher the risk of investing in that company.	As of Dec 2016
Earnings per share	Net profit divided by the number of common shares outstanding. A higher EPS indicates greater value because investors will pay more for a company's shares if they think the company has higher profits relative to its share price.	5Y growth, '11-16
P/E ratio	Current share price by the total EPS earnings over the past 12 months. indicates the dollar amount an investor can expect to invest in a company in order to receive one dollar of that company's earnings.	5Y growth, '11-16
Price	The current or historical price that a share of stock is trading for on the market	1Y and/or 3Yr %Chg, '16-19

Fig 1: Data items selected for model building and analysis

Problem Statement

This analysis will test the efficacy of using decision trees to predict stock-related performance. The prediction model will be based on a wide-range of market-related performance metrics and valuations. The metrics used in the model include PE ratios, Profit Margins, EPS Growth, Revenue Growth, Debt to Asset Ratios, and EBITDA Growth. The targeted time-frame for the analysis will be 2016-2019 therefore, omitting the volatile periods

associated with the pandemic. The model will be benchmarked against two other supervised models, namely, neural and Bayesian networks. This will allow us to differentiate and analyze the prediction accuracy of each model. However, decision trees will be the focal point of our analysis given their descriptive benefits.

Methodology

The dataset, compiled from FactSet's database, covers several characteristics including revenue, growth, risk and liquidity. Addressing a variety of characteristics ensures a model that can make accurate predictions across different industries and company sizes. Data attributes were selected from key ratios that are used by banks and hedge funds to analyze a company's performance and optimize investment portfolios.

The C5.0 Decision Tree algorithm was selected as the primary method of analysis. Its advantages include not being overly sensitive to missing values and its ability to address a large number of scenarios. Disadvantages to be cognizant of include the algorithm's tendency to overfit the dataset by performing overly complex analysis. To ensure that the model will perform well when applied to other datasets, the decision tree might be pruned before or after it has been trained, in addition to experimenting with different partition sizes for the dataset. In addition, the Neural Net and Bayesian Network algorithms were selected as comparison algorithms. These comparisons serve to evaluate the relative effectiveness of supervised algorithms in this context.

The primary metrics for evaluating the model will be accuracy and misclassification rates, using the external testing dataset (historical data from the S&P 500 index) to gauge generalization performance. In addition, the model's results will be compared to the results of the neural net and Bayesian network models. The relative accuracies of each model when applied to the training and testing samples will be studied to identify signs of overfit or inaccuracy.

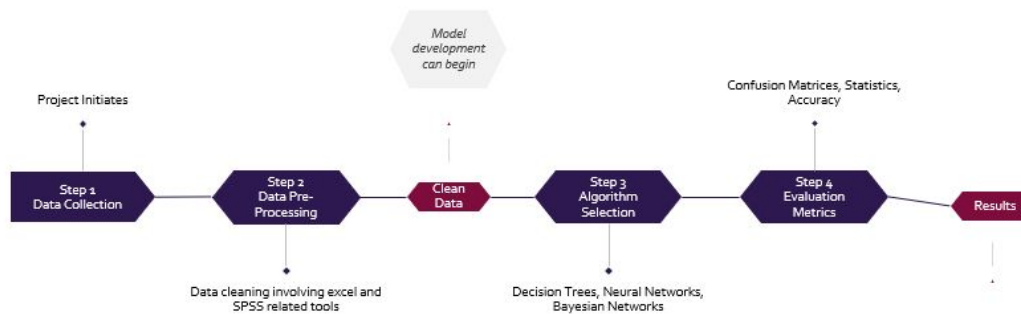


Fig 2: Stock market prediction workflow

In order to prepare the data for analysis, the data underwent several data cleaning procedures. Firstly, we had to introduce a supervised column. This binary column consisted of a Yes or No field indicating whether or not an individual company outperformed the broader market index. The price appreciation of the S&P 500 Index during the 2016-2019 targeted time-frame consisted of 43.88%. We set the outperformance benchmark to 3% over the market performance therefore, any company showing a price appreciation of 46.88% or better would meet the outperformance criteria. Following this procedure we removed all companies that contained inaccurate, incomplete, or misleading financial information. This consisted of companies in the dataset that did not report on any of the targeted metrics during the specified time frame. The specified time frame for our growth rates consisted of the 5 years preceding 2016. The objective behind using 5-yr preceding growth rates is to establish historical financial trends or inputs that may potentially influence price performance. After conducting these Excel-based data cleaning procedures, we shifted to cleaning the data in SPSS. The first procedure consisted of omitting any companies that IPO'd after 2011. We wanted to exclude these particular stocks since they would de-facto be unable to fulfill the 2011-2015 financial reporting requirement. Following this procedure, we noticed that our GICS_Sector field ([Global Industry Classification Standard](#)) contained “NA@” values for companies that did not fit the standard sector classifications. We used the Filler node and the replace function to replace “NA@” with “Other”. These SPSS adjustments finalized the data cleaning process, allowing us to transition into the model building & analysis stage.

Field	Measurement	Values	Missing	Check	Role
Symbol	Typeless			None	None
Name	Typeless			None	None
GICS_SECTOR	Nominal	"Communication Services", "Consu...		None	Input
IPO_DATE	Continuous	[19841105, 201111215]		None	None
LAST_TRADE_DATE	Flag	20201023/20200615		None	None
Outperformance(3% over 43.88...	Flag	Yes/No		None	Target
Average 5-yr Net Sales Growth:	Continuous	[-0.838621971, 1245.1306]		None	Input
Average 5-yr FCF Growth:	Continuous	[-387.7052733, 6044.892547]		None	Input
2016 Dividend?	Flag	1/0		None	Input
Average 5-yr EBITDA Growth	Continuous	[-147.5387449, 196.186296]		None	Input
Average 5-year EPS Growth	Continuous	[-1524.741288, 4.555999868E7]		None	Input
NET_INCOME_MARGIN_2016	Continuous	[-563713.572, 3825.159822]		None	Input
DEBT_TO_ASSETS_2016	Continuous	[0.0, 88575.0]		None	Input

Fig 3: Sample of inputs (with datatypes) used for model building

Results and Discussion

Decision Tree (C5.0 Classifier)

We started our analysis with a simple C5 Decision Tree model in SPSS. Decision trees are one of the most popular supervised machine learning algorithms, so we first wanted to know if the decision tree model can take in various continuous financial metric fields and still yield meaningful results on predicting if stock was a good or bad investment. Since our target field is binary (outperform or underperform) and not a continuous field, the results were easy to interpret. In the first stream, the model had a predictive accuracy of 75% and a tree depth of 8 splits. The field with the highest predictor importance was Avg 5Y EPS Growth. This wasn't a surprise because a higher EPS (earnings per share) indicates a profitable status, and typically suggests that the company is increasing its stock dividend payout over time. What was more interesting was the second split of the tree was on the field GICS Sector. This was the one nominal input field we included. It made us wonder if we were to know the GICS Sector of a company up front, could the DT be refined in a way to have a decision tree per sector? Could we split the records by GICS Sector cohorts?

We changed our DT model slightly to split the records by GICS Sector field, rather than have it as an input. This strategy mimics what many portfolio managers and financial advisors in the industry do to make investment decisions. The SPSS model produced an independent Decision Tree nugget per GICS Sector cohort. This adjustment increased our predictive accuracy for certain sectors (Energy at 92%), and reduced predictive accuracy in other sectors (Financials at 66%). This makes sense because certain industries are more volatile than others within a given time period. This also explains why Portfolio Managers and financial advisors stress the importance of diversification across sectors, asset classes and asset types to align with an individual's tolerance for risk.



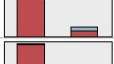



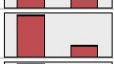

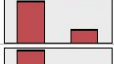





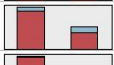



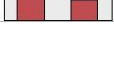

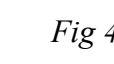

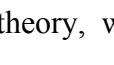

Graph	Model	GICS_SECTOR	No. Records in Split	No. Fields Used	Overall Accuracy (%)
		Energy	181	0	92.818
		Real Estate	127	2	86.614
		Communication Services	119	3	84.034
		Materials	153	1	81.699
		Other	55	0	80.0
		Consumer Staples	135	0	76.296
		Consumer Discretionary	367	0	73.842
		Information Technology	457	3	70.678
		Industrials	461	0	70.282
		Health Care	390	2	69.487
		Utilities	78	0	67.949
		Financials	116	0	66.379

Fig 4: C5 DT results by GICS Sector cohort

Lastly, to really prove our theory, we adjusted our DT model again by removing the GICS Sector field altogether from the DT model. The predictor importance changed to 5YR Net Sales Growth (.53), tree depth was reduced to 3 splits, and overall accuracy was reduced to 72%. We concluded that not knowing the GICS Sector of a company can reduce your predictability on if the stock will be a good or bad investment relative to the performance of the S&P 500. Secondly, we can conclude that DT models produce more meaningful results when including nominal fields as inputs, rather than many continuous fields as inputs.

Alternative Model #1: Neural Net

Several neural nets were generated to test various parameters and methods, with the goal of optimizing model stability and accuracy. GICS Sector was identified during data pre-processing and model building as having the potential to significantly influence output and results. The net was split on the sector field to generate predictions within each sector, yielding results with a high variance across sectors. In addition, the model was run both with and without sector as an input to examine differences in output. Ensemble methods and bootstrapping were also used to compare results. These methods produced models with high accuracy but also high variance across partitions, suggesting overfit. To avoid skewing results, their results were not used in the final evaluation.

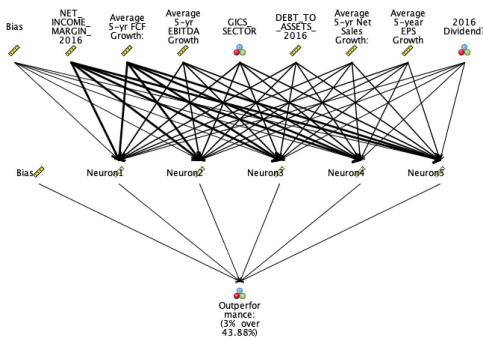


Fig 5: Final neural net

■ Results for output field Outperformance:(3% over 43.88%)

■ Comparing \$N-Outperformance:(3% over 43.88%) with Outperformance:(3% over 43.88%)

'Partition'	1_Training	2_Testing	3_Validation
Correct	1,099 72.83%	622 71.66%	190 72.52%
Wrong	410 27.17%	246 28.34%	72 27.48%
Total	1,509	868	262

■ Performance Evaluation

'Partition' = 1_Training	
No	0.0
'Partition' = 2_Testing	
No	-0.001
Yes	-0.332
'Partition' = 3_Validation	
No	0.0

Fig 6: Results of neural net (with 60:30:10 partitions)

The final partition split used was 60:30:10 for training, testing and validation respectively. This produced a model with minimal variance across partitions as well as reasonable accuracy (at a similar level to the other models). A combination of continuous, flag and nominal attributes were used as inputs, and the resulting multi-layer perceptron had 5 neurons in the hidden layer. The neural net predicted the target variable, Outperformance, with 72.8% accuracy in the training set, with similar accuracy levels in the testing (71.7%) and validation (72.5%) datasets.

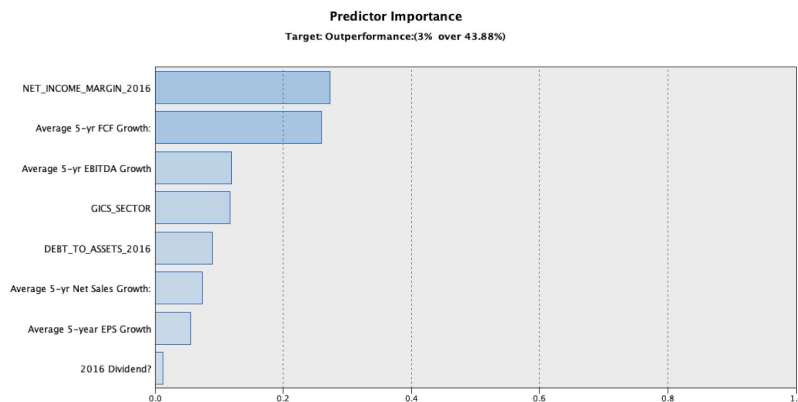


Fig 7: Predictor importance rankings for neural net

Net income margin and average 5-year FCF growth were identified as the most significant predictors by a large margin. Both ratios are used frequently in finance to assess a company's financial viability, as they offer insights into profitability as well as the company's ability to maintain operations and capital assets. When the model was run without sector as an input, the relative ranks of the predictors changed, but net income margin and FCF growth maintained their high ranks.

GICS sector was found to be a fairly important predictor when used as an input, particularly in influencing the relative importance of the other inputs. However, the sector attribute also had to be prevented from skewing

results and producing an overfitted model that would perform poorly on unseen data. Comparing models generated with and without sector as an input during the model building stage addressed this concern by ensuring that results from the final model were reliable to a large extent.

Alternative Model #2: Bayesian Network

Similarly to our neural network, we generated several Bayesian networks to test for accuracy and stability. The Bayes net was again split on the GICS Sector field to generate predictions on each sector, yielding high variance across all sectors. Also like the neural network, the model was run without sector as an input to examine differences in output.

Several combinations of the model were run, including the split on sector, without the split, with a partition, and without a partition. Accuracy and variance ranged widely across each model, producing low accuracy and high variance for one with a split and partition and high variance for another with just the split on sector. Ultimately, these models were omitted from our final evaluation.

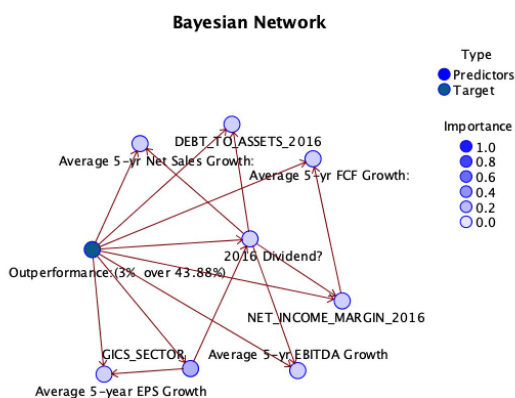


Fig 8: Final Bayes net

■ Results for output field Outperformance:(3% over 43.88%)

■ Comparing \$B-Outperformance:(3% over 43.88%) with Outperformance:(3% over 43.88%)

'Partition'	1_Training		2_Testing		3_Validation	
Correct	1,493	71.57%	195	67.01%	186	70.99%
Wrong	593	28.43%	96	32.99%	76	29.01%
Total	2,086		291		262	

■ Coincidence Matrix for \$B-Outperformance:(3% over 43.88%) (rows show actuals)

'Partition' = 1_Training		No	Yes	\$null\$
No		1,468	18	27
Yes		540	25	8

'Partition' = 2_Testing		No	Yes	\$null\$
No		191	5	13
Yes		77	4	1

'Partition' = 3_Validation		No	Yes	\$null\$
No		181	2	7
Yes		65	5	2

Fig 9: Results of Bayes net (80:10:10 partitions)

Our final bayes net was run on a partition split of 80:10:10 for training, testing, and validation, respectively. This resulted in a model with high accuracy and minimal variance, especially when compared to our decision trees and neural net. Inputs included a combination of continuous, nominal, and flag attributes, and our resulting model displays a directed graph with conditional dependencies reflected upon each input. Accuracy appears to

be highest for our training set at about 71.6%, with similar values for testing (67.0%) and validation (71.0%).

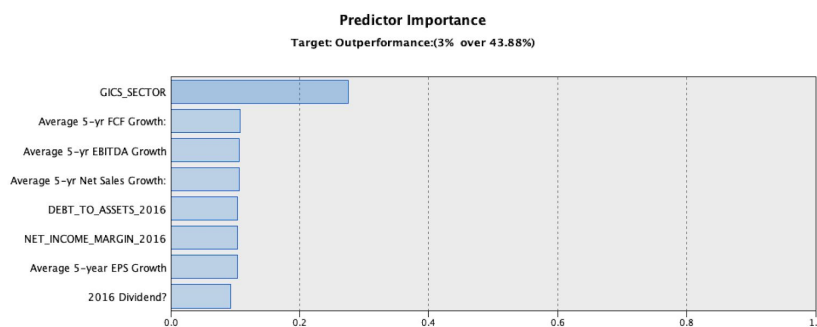


Fig 10: Predictor importance for Bayesian network

Predictor importance for the Bayes net indicated that the GICS Sector was the most important input by a large margin, again suggesting that quality of data in a particular sector can influence predictions. All other inputs were of similar importance, with the exception of 2016 Dividend, which is similar to our neural net. For the models run with a split on sector, we could predict those sectors that would perform poorly on their testing and validation sets. However, these models appeared to be overfit as variance ran high and accuracy was not consistent, leading us to omit the models.

Conclusion

The results from all three models indicate that revenue and cash flow ratios can offer powerful predictive insights within sectors. In particular, the ability to maintain operations and capital assets is closely correlated with changes in stock prices (and by extension, company performance). Net income margin, FCF growth and the debt to assets ratio were all identified as key predictors in at least 2 out of 3 models. In addition, GICS sector plays a crucial role in influencing the relative predictive importance of other inputs, and results are far less reliable when the GICS sector is unknown. This suggests that GICS Sector classification coverage quality can influence a model's overall performance. Nevertheless, industry-specific trends were captured to varying degrees by the ratios used as inputs. Split models were highly performant (more than 85% accuracy) in some sectors, such as energy, real estate and communications, but unreliable in others. This implies that the models must be deployed in carefully calibrated cohorts.

While all three models registered roughly 70% accuracy, decision trees proved to be the most informative when nominal fields are included as inputs. This confirms the hypothesis that the C5.0 classifier would be the most appropriate algorithm in this context. However, attention must be paid to the relative number of continuous inputs during deployment, and the results must be examined for signs of skew or overfit.

Regarding the assumptions made, future models could be trained using other benchmarks (especially other major indices like the Dow Jones Industrial Average or the Nasdaq Composite) to improve generalization performance. Also, sectors with low predictive accuracy could be addressed by collecting more or better quality data. Such measures are likely to improve the overall performance of the predictive models as well as the investment portfolios they are applied to.

Appendix

Group Member Tasks

Prannoiy Chandran: Neural net model and analysis, Conclusion (worked on report and presentation slides)

Marissa Maffa: Decision tree model and analysis, Data collection (worked on report and presentation slides)

Sebastian Carvajal: Data pre-processing and cleaning (worked on report and presentation slides)

David Reyes: Bayes net model and analysis (worked on report and presentation slides)

References

[pwc.com/gx/en/industries/financial-services/asset-management/publications/asset-management-2020-a-brave-new-world.html](https://www.pwc.com/gx/en/industries/financial-services/asset-management/publications/asset-management-2020-a-brave-new-world.html)

<https://www.bcg.com/en-us/publications/2020/global-asset-management-protect-adapt-innovate>