MACHINE-LEARNING APPLICATIONS IN INVESTMENT STRATEGY

Assignment 2

25579 Applied Portfolio Management - Autumn 2023

Vi Nguyen

13592629 | Chi.V.Nguyen@student.uts.edu.au

Table of Contents

1. Advantages and disadvantages of machine learning based alpha models... 2

1.1 Why are the machine learning becoming popular now? 2

1.2 Benefits of Machine Learning models 2

1.3 Risk associated with machine learning 3

2. Predictive power of decision tree 4

2.1 Creating the decision tree 5

2.2 Evaluating the model 4

3. Economic interpretation 6

4. Optimization 8

5. Regression Tree 11

References 12

1. Advantages and disadvantages of machine learning based alpha models.

1.1 Why are the machine learning becoming popular now?

Machine learning is a very powerful tool for analysing large quantitative data. Regardless, the utilisation of the machine learning approach to investing was just recently widespread globally. Before that, there were many constraints that discouraged the development and application of machine learning in the financial field, varying from technical aspects, data availability and computing power.

First, the evolution of big data in the last two decades has generated substantial amounts of data and created many new data types. Specifically, the emerging data sources include Social media, satellite technology, sensor devices, and many original data types like video, images or sound. According to Forbes, 2018 alone generated 90% of the entire data in the world (Marr, 2018). The machine learning algorithm required some massive quantity of data to train and work effectively.

Secondly, the computing power and resource have become more advance and cost-effective. The development of supercomputers allows machine learning algorithms to process faster and more efficiently. Moreover, the storage cost was a big problem two decades ago, with a significant event Y2K that impacted many industries that utilise computer-based technology. Cloud technology development allows businesses to store their data 'on the cloud' in ample virtual storage at a sufficient cost.

Lastly, the unstructured data has been growing very fast recently. According to Forbes, unstructured data contributed over 80% of the total data generated (Marr, 2019). Unstructured data was not very useful to most of the investors out there. Regardless, this indicated that there are many opportunities out there to be discovered. With the recent machine learning advancement, particularly in the pattern recognition field, investors with sufficient data analytics skills can utilise this technology to discover new patterns and develop new investing strategies. More and more investment funds worldwide have applied these statistical models to their daily operations.

1.2 Benefits of Machine Learning models

The benefits of applying machine learning models produce cost-effective solutions to analyse many data types originating from different data sources. Recently, the trend in social media like Tik Tok or Twitter generated a substantial amount of data that seem impractical to many people. Regardless, the recent evolution of machine learning methodologies like image processing or pattern recognition allowed investors to derive meaningful stories from these unstructured data.

For an investor looking for a set of companies with a pre-defined characteristic, the supervised machine learning algorithms can scan and analyse large datasets, returning companies that satisfy the requirements. Specifically, this part of supervised machine learning is called classification. Moreover, based on the current trend of the factors, the regression analysis can be applied to predict the future return of stocks based on past trends.

On the other hand, if the fund managers wish to seek new and unexplored patterns other than regular profitability measurement, they can employ the unsupervised machine learning algorithm. Unsupervised machine learning focuses on clustering, which groups the stocks with the best performance together to find any possible factors that explain the outstanding performance. The program will then recommend any potential companies having similar factors to stock that perform well.

Machine learning is very cost and time effective. A single employee on the data analysis software can replace many roles, such as data collectors, statisticians, business analysts, etc. The automation and integration capacity allows one machine learning program to run across many platforms and data sources with some modifications. The analyst can then focus on more critical investment decisions that require qualitative analysis, which the machine learning algorithms unable to deliver.

With an emerging trend of data-driven investing, the machine learning approach allows many fund managers to disregard traditional investing methods with no competitive advantages and focus only on data-driven techniques. Hence, investors can construct unique investment strategies that benefit different parties.

1.3 Risk associated with machine learning

On the other hand, the power of machine learning comes with many risks and associated problems if the data analysis cannot handle them carefully. The problems came from the nature of machine learning, some ethical issues as well as the safety of the data.

Most machine learning techniques are very technical and challenging to interpret by any data or business analyst. Regarding critical decisions like medical surgery or investing a billion-dollar portfolio, an ambiguous explanation from the machine learning algorithm can raise many problems.

The machine learning model can over-train or overfit the data. Training the data can increase the accuracy of the training model, but it would be useless for the actual data. Overtraining the data can make the algorithm less generalised, unable to work with the other data set. The data analyst's job is to find the optimal point that balances the in-sample and out-of-sample accuracy.

There is also some ethical concern surrounding machine learning. These algorithms require enormous amounts of data to learn and produce meaningful patterns. Most data providers may not be aware of how their data is collected and used. When customers sign up for a membership or subscription, the organisations would not inform the customer about how they monitor and analyse the customer behaviours.

Moreover, the security of these data raises more severe concerns. These personal data can be sensitive and potentially cause dramatic problems if not protected from cyber-attacks. The hackers can use this data to causing financial damages, like taking out mortgage loans and buying unwanted products. More seriously, they can commit criminal offences under any person's identity.

The reliance on machine learning can deliver some problems. For instance, a few financial funds nowadays rely solely on statistical machine learning algorithms and ignore the traditional value. If the data analysts have some conflict of interest, they can biasedly select the parameters and decide which stock has an outstanding performance.

2. Predictive power of decision tree

2.1 Creating the decision tree

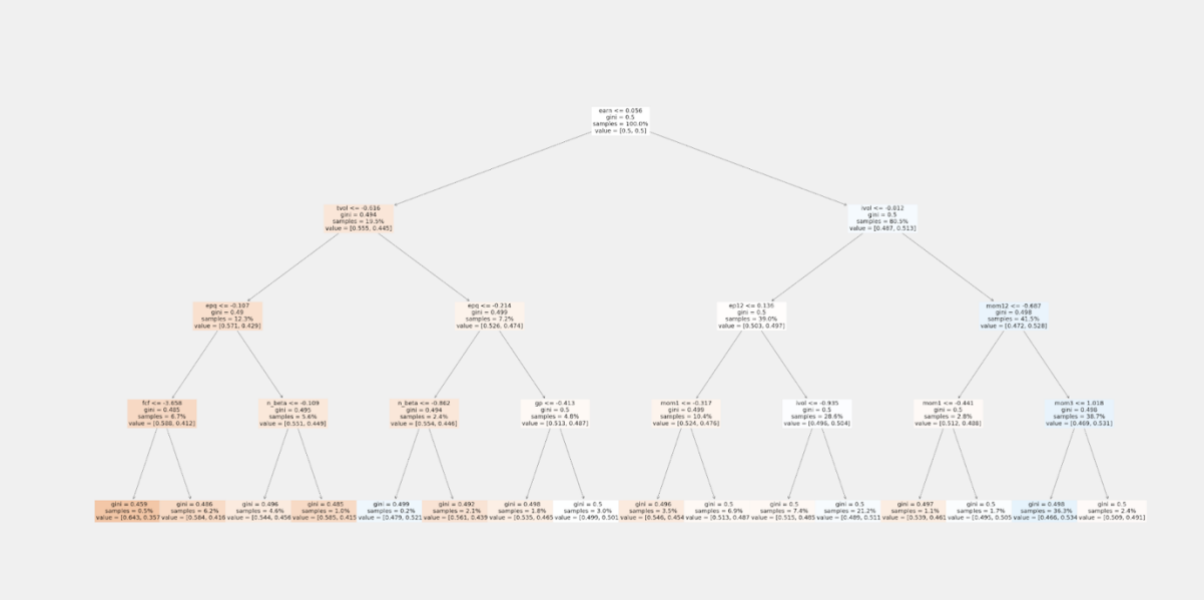


Figure 2.1: Decision tree

2.2 Evaluating the model

The predictive power of the model is being evaluated using two different ways:

1. The statistical evaluation of the model measures the accuracy rate, which is the percentage of correct guesses. There are two accuracy rates:

- **In-Sample accuracy rate** examines the training set.
- **Out-of-sample accuracy rate** examines the testing set.

2. The financial performance measure: Mean Return, Information Ratio (RR ratio) and others financial ratio.

2.2.1 Evaluating the model using statistical measures

The accuracy of the In-Sample and Out-of-Sample have no substantial difference. The similarity signals that the timeline train-test split performance is entirely satisfactory and did not overfit the data. Moreover, the cross-validation model has a similar accuracy with the in-sample accuracy and out-of-sample accuracy using the ten different splits of our sample.

Table 2.1: Statistical accuracy

In-Sample Accuracy	Out-of-Sample Accuracy	10-Folds Cross Validation Average Accuracy
0.53	0.52	0.52

Regardless, the model's performance is slightly above 50%, which is a just a bit better than a random guess. This is due to the volatile nature of financial data. Hence, having a model that beats the random guess by just a few percentages can still be ultimately beneficial for the portfolio manager.

On the other hand, the performance of the model is just slightly above 50% which is a random guess. Regardless, the financial data is relatively volatile. Hence, having a model that beats

that beat the random guess by just a few percentages still can be ultimately beneficial for the portfolio manager.

2.2.2 Evaluating the model using investment returns & performance

The statistical measure **does not reflect** the financial performance of the investment strategy. Since the portfolio manager can choose to short-sell the bad stock and overinvest in the good stock, the investment performance will differ greatly from the statistical performance.

Table 2.2: Portfolio performance diagnostics

	Predicted loser	Predicted winner	Benchmark	Active	Neutral
Mean Return	-9.49%	4.03%	0.02%	4.01%	13.52%
Standard Deviation - Tracking error	35.31%	25.21%	27.66%	5.54%	17.17%
Information Ratio (RR ratio)	-26.89%	15.98%	0.06%	72.40%	78.75%
Positive Period (%)	51.43%	65.71%	62.86%	60.00%	60.00%
Worst Month Performance	-33.09%	-24.63%	-27.14%	-4.28%	-12.80%
Best Month Performance	21.11%	13.38%	16.04%	2.99%	8.46%
Max DrawDown	-58.67%	-26.71%	-35.13%	-11.84%	-36.04%

Overall, the classification performs very well in predicting the winner and the loser, as shown in Table 2.2. The mean return of the predicted winner is 13.52% higher than the predicted loser, which is very impressive. The tracking error of the winner is also lower than the loser, signalling a safer and less volatile strategy. The information ratio (RR ratio) is the return of the average return of the strategy, adjusted the risk by dividing the tracking error. The risk-adjusted return of the predicted winner is 16%, beating the performance of the predicted loser by 78%. On the other hand, the risk-adjusted average return of the predicted winner also beat the benchmark by 72%. The tracking error of the benchmark is slightly higher than the predicted winner.

Among the three, the predicted winner has the highest percentage of positive periods, followed by the benchmark and the predicted loser. The order is reversed for the worst performance month. However, the best performance measure is quite abnormal where the predicted loser has the higher return, followers by the benchmark and the predicted winner have the lowest best month performance. The max drawdown of the predicted loser is the worst, followed by the benchmark and predicted winner.

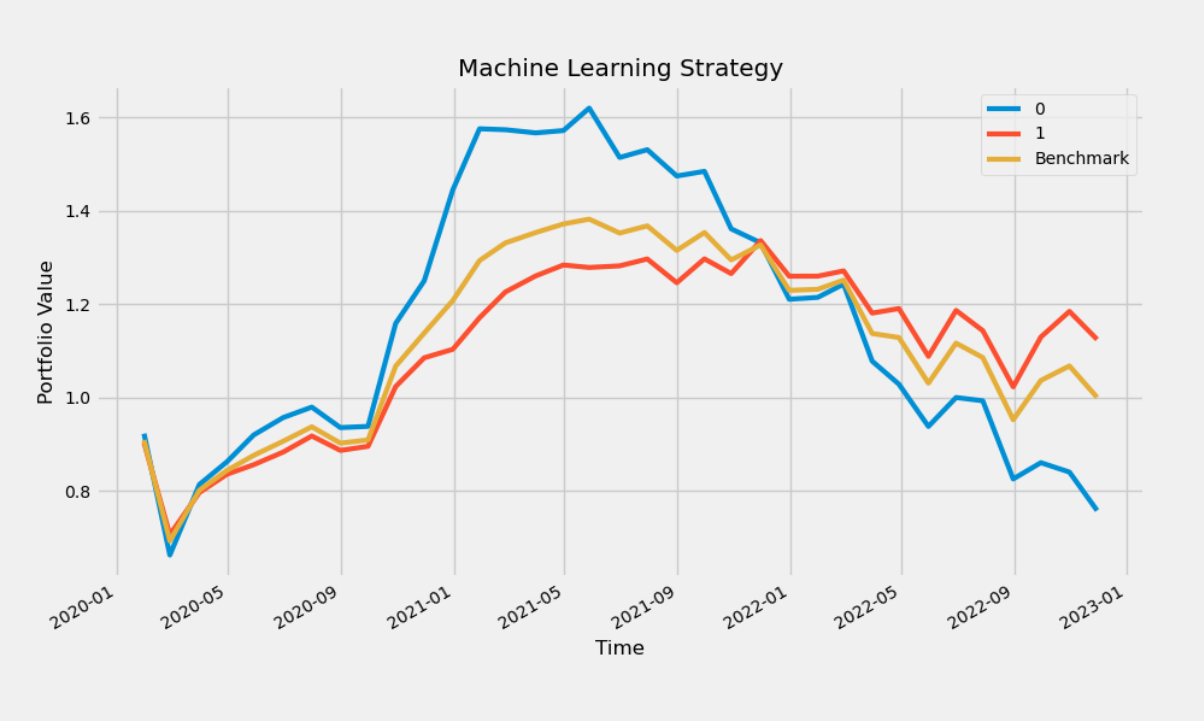


Figure 2.3: Accumulated return of the strategy overtime

Figure 2.3 Illustrated the cumulative return of the three strategies over the Covid period of the sample test set. There are no significant differences between the three strategies before October 2020. However, the predicted loser (0) outperformed the other two portfolios consistently over 1 year until the beginning of 2022. This is also the period where Benchmark outperform the predicted winner (1) by roughly 5-10%. After that, the predicted losers dropped significantly under 80% at the end of 2022. The Predicted winner had outperformed the Benchmark over one year since the start of 2022. The result indicates that this machine-learning strategy did not perform well during the pandemic.

3. Economic Interpretation

Table 3.1 Feature Importance

Profitability	54.9%	gp	1.1%
		fcf	1.2%
		earn	52.7%
Value	9.6%	mom1	2.8%
		mom3	3.1%
		mom12	3.6%
Momentum	9.3%	epq	4.9%
		ep12	4.4%

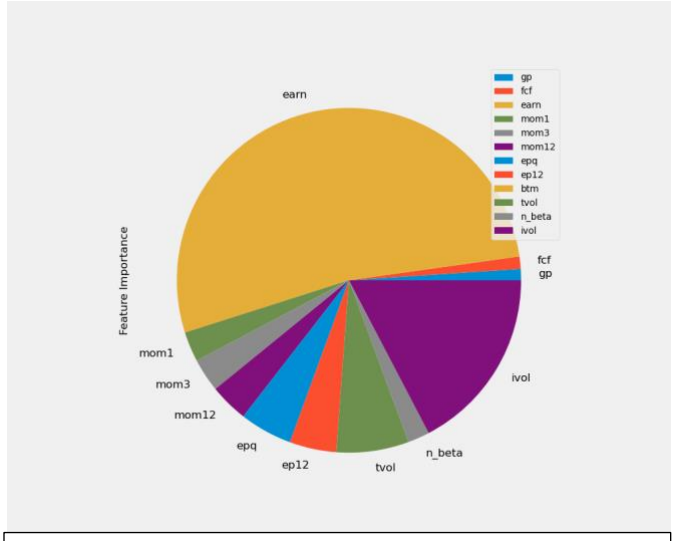
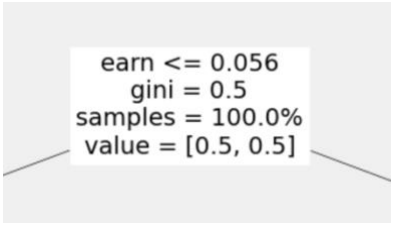


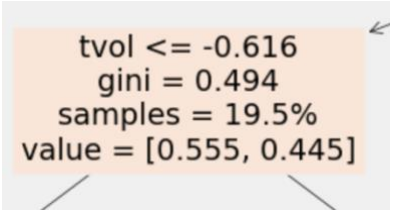
Figure 3.1: Feature importance of the factors

		btm	0.0%
		tvol	6.8%
Low-Votatility	26.2%	n_beta	2.0%
		ivol	17.3%

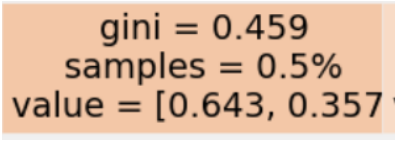
Based on Table 3.1: Feature Importance, the most important factors are that profitability accounts for 55% of gini reductions, followed by low volatility, which contributes 26.2%. The value and momentum factor contributes around 9%. Among the three profitability factors, the Earnings Profitability (Income Before Extraordinary Items / Total Assets) or 'earn' have the most significant impact. This high contribution indicates that profitability had a high correlation with the stock's future return. The increase in profitability always attracts long-term and short-term investors. The low volatility is also a good indication of safe and stable stock, which attract more long-term investors.

Back to the decision tree, we have a similar pattern. The tree will be divided into two branches for each level based on the condition of the first line. For instance, the stock is divided into two branches based on the condition of the first line: earn <= 0.056. The second one is the Gini impurity of the node. The third line is the sample percentage, illustrating how much this node contributed toward the total sample. Lastly, the first element in the square cracket value illustrates the probability of the stocks with specific factors will perform below the median next month. For instance, the left branch nodes are on the left. The probability of the stock with Earning Profitability below 5.6% will have a 55.5% chance that they will perform below the median.





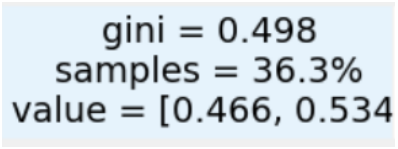
Therefore, the node with the high first element will be bad stock. The node on the left at the lowest level is the very bad node, containing 0.5% of the sample, with all the conditions:



- Earning Profitability < 5.6%
- Volatility < -61.6%
- Quarterly EPS < -10.7%
- Cash Profitability < 3.658

On the other hand, the ‘good’ stock contains the highest probability that its performance will be higher than the median for the next month. This is the node on the right of the leftmost node. The node contain 36.3% of the sample, having 53.4% chance that the stocks with these conditions would have a higher return than the median next month. The conditions are:

- Earning Profitability > 5.6%
- Volatility > -0.012
- 12-month momentum > -0.687
- 3-month momentum <= 101.8 %



In conclusion, the best stocks are stocks with high Earning Profitability, slightly high volatility, no significant downtrend in the last 12 months, and having a recent strong up-trend month. This strategy does not have a lot of difference from the traditional investment strategy focusing on the profitability of the stock first, then searching for the safety.

4. Optimization

Figure 4.1 indicates that the optimal depth of the tree for out-of-sample accuracy is five, with an accuracy rate of approximately 52.4%. The best range of tree depth is from 5 - 7, providing an accuracy above 52.3%. The accuracy started to decrease as the depth exceeded 9.

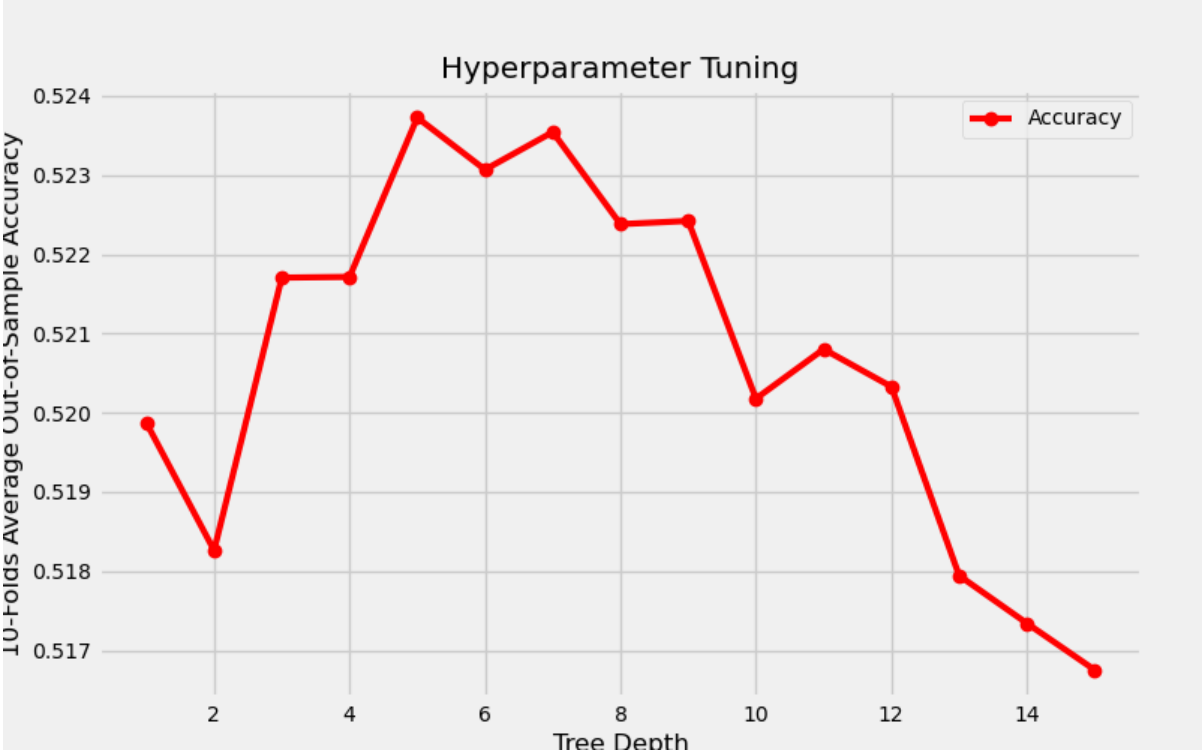


Figure 4.1: Hyperparameter Tuning for 10-Folds Average Out-of-Sample Accuracy

Regardless, this is not the optimal depth for the Information Ratio. The Information Ratio for the predicted winner was 15.98% when the depth equals 4. As the tree depth increased to 5, the Information ratio reduced to 14.62%. With higher complexity, the Information Ratio decreased.

Figure 4.2 and Table 4.1 illustrates that the optimal depth for Information Ratio is 2, yielding an Information ratio of 0.22. Regardless, this level of depth can lead to underfitting. As shown in Figure 4.3, where the tree depth is 2, there are two nodes with a sample size of approximately 40%. Hence, it would be more meaningful to go a bit deeper. As the tree depth exceeded 4, the Information Ratio increased slightly from 1-2% and decreased after 9. This minor improvement would not satisfy the extra complication the decision tree would generate going a few levels deeper.

Table 4.1: Hyperparameter Tuning for Information Ratio of Predicted Winner

Tree Depth	1	2	3	4	5	6	7
Information Ratio	15.46%	21.66%	13.65%	15.98%	14.62%	15.78%	16.70%

Tree Depth	8	9	10	11	12	13	14	15
Information Ratio	16.48%	16.56%	15.49%	14.55%	14.92%	13.36%	12.39%	14.23%

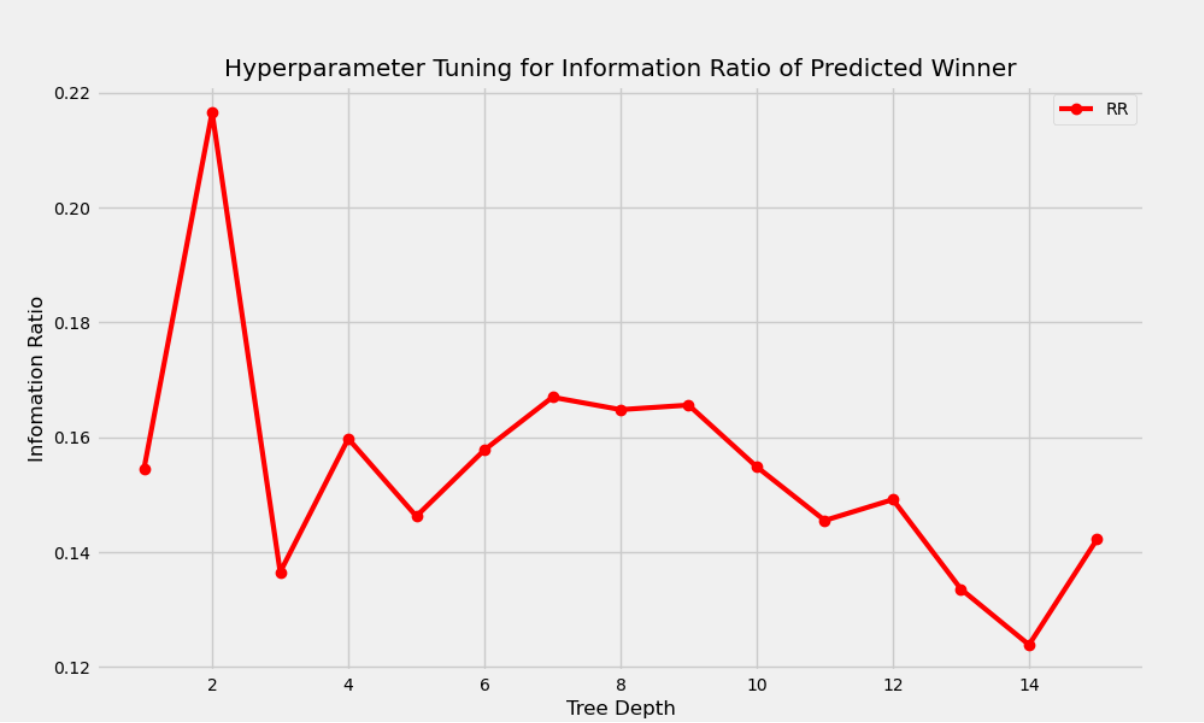


Figure 4.2: Hyperparameter Tuning for Information Ratio of Predicted Winner

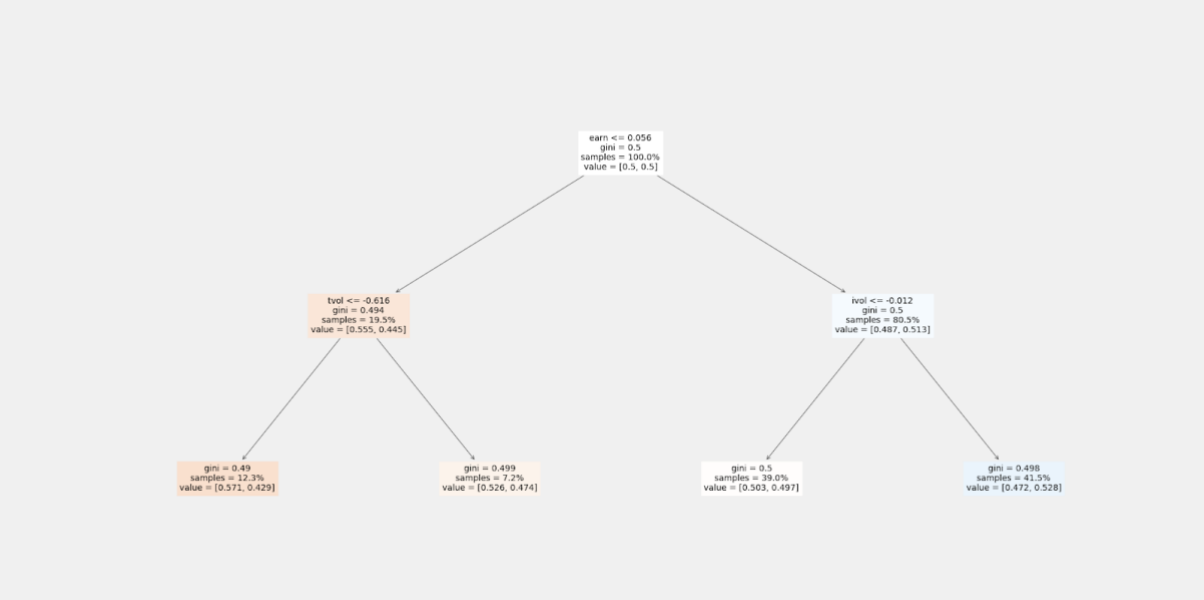


Figure 4.3: Decisitrion with depth of 2

Regarding the min sample leaf, the optimal sample leaf should be 5% where it yield the highest out-of-sample accuracy, as shown in Figure 4.4. The optimal depth is also 4 as the accuracy did not increase after 4.
In conclusion, the optimal measurements are the tree depth of 4 and the minimum sample leaf of 5%.

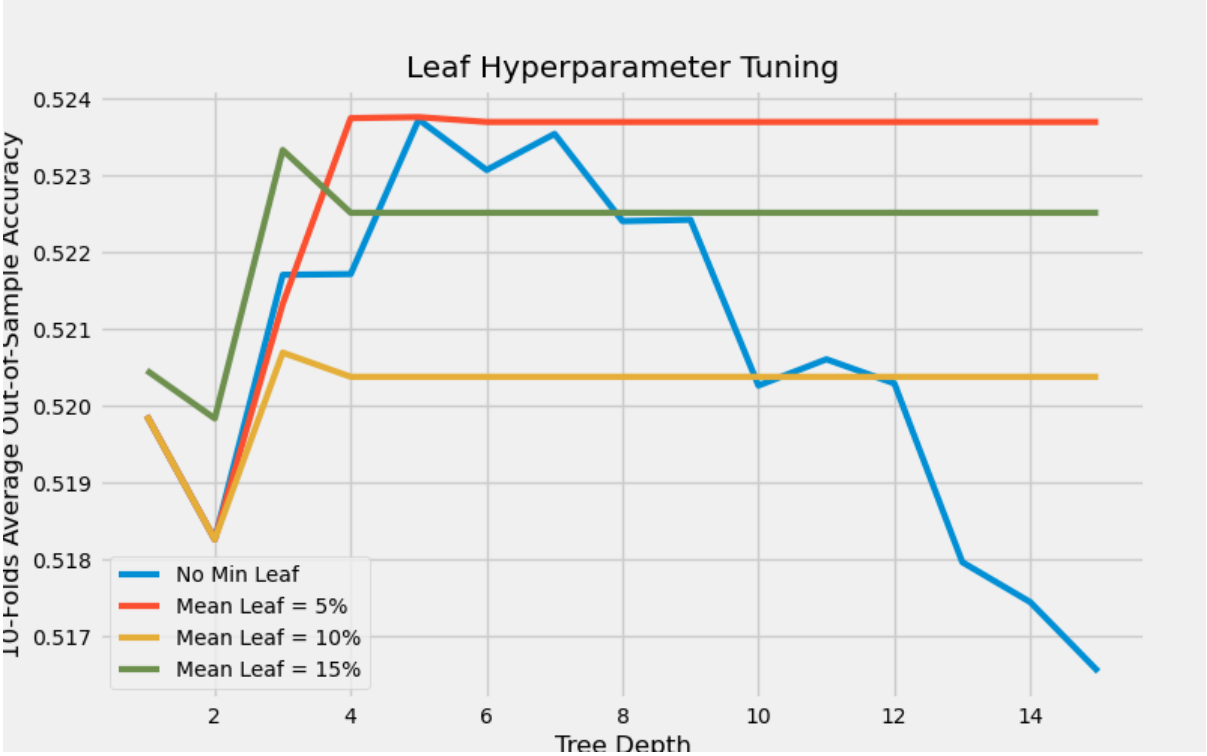


Figure 4.4: Leaf Hyperparameter Tuning

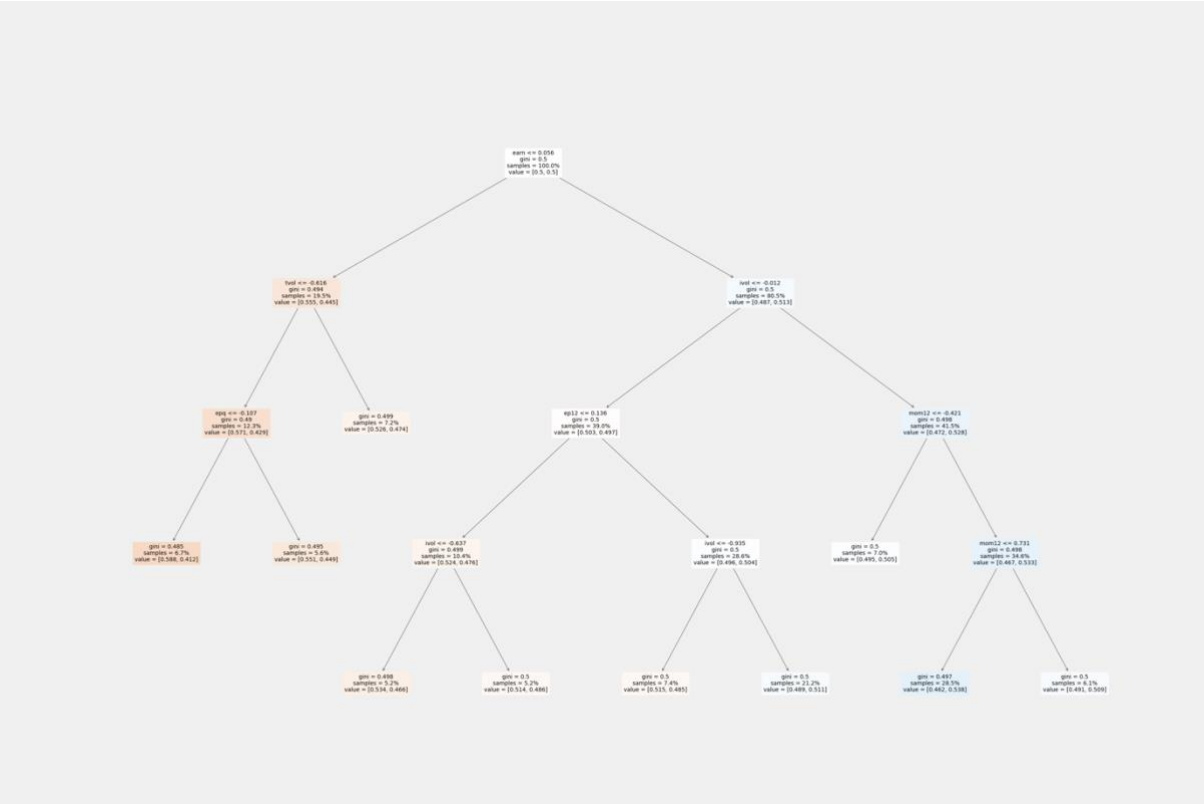


Figure 4.5: Decisitrion with depth of 4 (Optimal)

5. Regression Tree

The in-sample and out-of-sample accuracy can differ greatly, especially when analysing and predicting financial data. The reasons behind the difference initiated from the machine learning algorithm itself, like overfitting, as well as the nature of the financial market.

First, the overfitting of training data creates a significant difference between the training and testing accuracy. For instance, allocating 95% of the dataset to training can increase the in-sample accuracy. Regardless, this will greatly reduce the out-of-sample accuracy. Therefore, the overfitting result is outstanding but impractical when applied to other datasets.

Even if the dataset is spitted appropriately, and the accuracy of the predicting model is high, there is no guarantee that the model can return a high performance in the future. For instance, Figure 2.3 illustrates that there is one year that the predicted losers outperformed the predicted winner in 2021 -2022. This error is due to Black Swan events like Covid 19. The changes in macroeconomic factors can significantly impact the predicting power of any machine learning algorithm. Like any other algorithm, the decision tree predicts the stock's future return based on the past pattern. If the pattern changes suddenly, the machine learning algorithms cannot acknowledge the modification.

Moreover, many components of the equation can vary over time. The transaction costs or the investment amounts can increase that, reducing the actual return of the strategy. Factors such as liquidity can also impact the strategy's performance, where it does not allow the fund manager to sell the stock or make a short position.

Another reason for the difference could be the bias of the data analytics. In constructing the decision tree algorithm, the data analytic can decide to explore certain factors that they think fit the most. They can choose to over-train the factors that they think are most suitable.

In conclusion, the data analyst must comprehend that these machine algorithms should only work as an assistant to discover new insights and patterns. The data analyst must continue to monitor, evaluate and enhance the quality of the machine learning algorithm over time.

Reference

Marr, B. (May, 2018) *How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read.* Forbes. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/?sh=9b77d0360ba9

Marr, B. (2019, Oct 10). *What Is Unstructured Data And Why Is It So Important To Businesses? An Easy Explanation For Anyone*. Forbes. https://www.forbes.com/sites/bernardmarr/2019/10/16/what-is-unstructured-data-and-why-is-it-so-important-to-businesses-an-easy-explanation-for-anyone/?sh=45c7d57115f6