BI SOLUTION FOR AT

The best BI solution to solve AT's investment decision issues

OCTOBER 10, 2018

THE UNIVERSITY OF AUCKLAND

Owusu Frimpong kfri750 246378566

Table of Contents

Executive Summary

Alison Trumps Investment management company manages about 1.2billion in assets, primarily in the share markets and in investment funds. Due to the fact that it operates mainly in financial markets, there is a huge need to be able to predict market outcomes and use it to make money for their clients.

The overall problem we are trying to solve here is that analysts can't effectively predict movements in the stock market due to the high amount of complex data. This problem can be broken down into the fact that there are too many different potential data points and too much data to analyse manually. Each factor holds a different weighting which means they can have nonlinear effects on stock prices. Similar factors can have different effects on different stocks depending changes in third-party stocks which may make it difficult to predict. The results of these factors change periodically depending on changes in the macroeconomy.

The Business Intelligence solution that we have chosen is a combination of Artificial Neural networks and a genetic algorithm. This solution is the most suited to solve each of the 4 subproblems mentioned earlier. The genetic algorithm will act to optimise the parameters used to make stock predictions in the neural network. This solution was chosen due to the fact that it most closely was able to meet requirements specified in the requirement specification. The reason a combination of GA and ANN was selected as opposed to just ANN was due to the risk of the neural network overtraining itself and tracking day to day changes as opposed to trends in the medium to long-term.

Problem

The problem is being analysed from the perspective of a financial analyst. Financial analysts assess the performance of stocks, bonds, and other types of investments. This is difficult because, on a day-to-day basis, random shocks, crowd psychology, and short-lived trends influence financial markets in complex ways.

Problem: Analysts can't adequately predict movements in the stock market due to the high amount of complex data.

Requirements Specification

- The tool needs to give insight and make informed predictions on the direction and magnitude of price movements over time.
- Investments need to perform higher than the market which is tracked by trade funds such as the NZX50 or ASX200 since the market index provides a general standard of market performance.
- The system must have an embeddability high enough such that it can be used with current analytical processes.
- The system needs to be able to produce results in a reasonable time frame.
- The system needs to have high consistency and accuracy
- To be useful to AT the system needs to be able to interpret and analyse large amounts of market data and "update its view of the world" frequently and easily.
- AT does not need the system to make specific point predictions for prices on a specific date but needed it only to provide the decision maker with estimates of a share's upside and downside potential.
- The system needs to be scalable so that new scenarios can be added later.
- For the model to be practical it should also be flexible enough to accommodate new market trends, new types of data, and portfolio objectives.

Stakeholders

As the CEO of the company, Shahab would make decisions relating to the analysis of revenues, expenses and profits. If this BI solution is implemented financial analysts and portfolio managers would be able to make more informed and accurate decisions. This will have a positive effect on overall revenue made through share profits for AT. This increase in share revenue will also increase the reputation of the company, and more shareholders will want to use AT to invest their money.

Financial analysts assess the performance of stocks, bonds, and other types of investments. They then report this data to portfolio managers who make the final decision on whether to buy, sell or keep shares. The BI solution would make their jobs significantly easier as they now don't have to trawl through vast amounts of data and analyse different data points.

Portfolio managers do extensive research to make investment decisions for a fund or group of funds under their control. They may spend the day meeting with analysts, researchers, and clients checking the financial markets, keeping up on company news, and buying and selling investments as things change. As economic analysis becomes more and more accurate due to the implementation of this BI solution, decisions made by the portfolio managers become more and more accurate this leads to higher gains made by shareholders. With this BI solution, it is possible to filter out unsuitable shares for specific clients, for example, if the best possible investment is in a cattle farm but the investor is Buddhist the next best solution will have to be looked at, and this can be repeated until a suitable personalised solution is found. This overall will allow portfolio managers to do their jobs better.

Subproblems

Too many different potential data points and too much data to analyse manually

- For simplicity, I have chosen to use the example of stocks in the construction industry in New Zealand. This subproblem describes that in the situation of a stock market there are hundreds, possibly thousands of different potential data points and an almost infinite amount of possibly relevant data that could be used for stock analysis.

Each factor holds a different weighting which means they can have nonlinear effects

- The effect of an earthquake is likely to have a higher impact on stock prices than a server restart.

Similar factors can have different effects on different stocks

- This problem describes that a result may have a different impact on stock prices. For example, there are 2 construction companies in Wellington, one much larger than the other and there is an earthquake in Wellington that doesn't affect the structure of either firm. Suddenly there is a large amount of construction work that could be contracted to both firms. In this case, the larger firm is likely to have a higher percentage increase in their share price as investors are expected to flock to a more trusted stock, assuming nothing else changes. These are the types of distinctions a possible solution must address.

The effects of these factors change periodically depending on changes in the macroeconomy

- If the NZX50 is going up significantly the impact of a reduction in construction projects would have a smaller overall effect than if there was low confidence in the financial market (drop or slowdown in NZX50)

Possible Solutions

Decision trees

Classification is the task of assigning objects to one of several categories, the system then uses these categories to predict the probability of an outcome. In this case, the classifying attribute would be the different factors that could be utilised predict stock values such as changes in the market capitalisation. This algorithm is worth mentioning as it ticks a couple of the boxes asked for in the requirement specification such as that it uses simple calculations which would allow insight to be gained quickly when required and that it doesn't need to function in real time. The input attributes: (what they should be, what ranges of value should they have) would have to be programmed individually which means it would require a lot of expert time to train the system. Another downside is that there is also no continuous prediction which means every time the system is to be used a new set of data would have to be inputted into the system. Another key thing is that it can't handle complex data as well as other solutions in fact, the more data and more trees get added, the less accurate the solution becomes. This is catastrophic considering the amount of data and different data points are needed to predict stock changes. For this reason, it would be difficult for this algorithm to answer the subproblems mentioned above.

Dimension	Target Solution	Decision Trees
Accuracy	Moderate	Moderate
Explainability	Moderate	Low
Response Speed	Moderate	High
Scalability	Moderate	Low
Flexibility	High	Low
Embeddability	High	Low
Tolerance for complexity	High	Moderate
Tolerance for noise in data	High	Moderate
Independence from experts	High	Low

Naïve Bayes

This algorithm works with a similar concept to decision trees in that it uses the probability of previous events to predict events for similar situations in the future. A fundamental assumption it uses, however, is that instances are independent of each other, but in the real world and especially the stock market this is untrue. The system requires supervised learning which would require a high amount of expert time to implement. Similar factors can have different effects on different stocks and each factor holds a different weightingleading to them having nonlinear effects because of this, past probability-based algorithms such as this would struggle to make accurate decisions.

Dimension	Target Solution	Naïve Bayes
Accuracy	Moderate	Low
Explainability	Moderate	Low
Response Speed	Moderate	High
Scalability	Moderate	High
Flexibility	High	High
Embeddability	High	Moderate
Tolerance for complexity	High	Moderate
Tolerance for noise in data	High	Moderate
Independence from experts	High	Moderate

Rule-based systems

Another option we can use to these predictions is a rule-based system. They use predefined knowledge to solve problems. It uses rules specified by experts to create IF-THEN statements which in this case can be used to help analysts make decisions on their stock choices. The benefits of using an algorithm based on this are that it can provide justification for every decision it makes, another advantage is that it doesn't require the whole solutions worth of rules to be explained before it can be used, rules can be added as needed. If there is a massive amount of data machine learning algorithms can be used to create new rules. The major disadvantage of this, however, is that it cannot handle the complex data associated with the stock market, where different rules may have different effects depending on third party factors. Another flaw is that it cannot learn from its own mistakes if a rule is added that doesn't apply to every situation the system won't know this and will continue to make wrong recommendations.

Dimension	Target Solution	Rule-based systems
Accuracy	Moderate	High
Explainability	Moderate	High
Response Speed	Moderate	High
Scalability	Moderate	Low
Flexibility	High	Moderate
Embeddability	High	Moderate
Tolerance for complexity	High	Low
Tolerance for noise in data	High	Low
Independence from experts	High	Low

Case-based reasoning

Another possible solution is case-based reasoning. It is the process of finding similar cases and adjusting the solution to account for the differences between the case being solved and cases in the case-base. A record of each past attempt is stored as a case in a case-base. Cases are stored in a database with columns describing attributes and rows showing cases, this allows for manipulation using SQL. An example could be the 2009 global financial crisis, as many variables as possible from that time period broken down by date could be added and relationships can be gleaned from it. These acquired relationships can then be applied to other similar cases. Stored cases can be searched for by similarity of attributes when a case needs solving. Some downsides to this algorithm are that if attributes have subtle dependencies or there are interactions between factors it becomes more difficult to compare one case with another. In the example of the stock market where there are vast amounts of different data points with subtle dependencies and interactions, it makes it very difficult to use this and gain accurate results.

Dimension	Target Solution	Case-Based Reasoning
Accuracy	Moderate	Moderate
Explainability	Moderate	Moderate
Response Speed	Moderate	High
Scalability	Moderate	High
Flexibility	High	High
Embeddability	High	Moderate
Tolerance for complexity	High	Moderate
Tolerance for noise in data	High	Moderate
Independence from experts	High	Moderate

Fuzzy Logic

Another possible option for this is to use fuzzy logic. Fuzzy logic allows the algorithm to make judgements on the extremity of general statements such as “there was a huge earthquake in Wellington”. This allows the algorithm of choice to quantify the effect of certain factors, e.g. how much this event will affect Weta Workshops potential output. However, it may be difficult to implement this in a financial analyst’s toolkit as the reasoning used may hinder the stability or reliability of the system. Another downside is that the system cannot learn from its mistakes or adapt itself, it must be adjusted manually by an expert which means it has low independence from experts.

Dimension	**Target Solution**	**Fuzzy Logic**
Accuracy	Moderate	High
Explainability	Moderate	Moderate
Response Speed	Moderate	Moderate
Scalability	Moderate	High
Flexibility	High	Moderate
Embeddability	High	Moderate
Tolerance for complexity	High	Low
Tolerance for noise in data	High	Low
Independence from experts	High	Low

Best-fit Solutions

Artificial Neural Network (ANN)

Another choice is the Artificial Neural Network which works by finding patterns in data and using it to build models and find relationships between data points. It is an excellent option because it can work with incomplete or noisy data. It is also able to correct its own mistakes. It learns from its mistakes by bring presented a new piece of data to which it guesses a result, that result is compared with the actual result, if it was right nothing happens, however, if it was wrong the system will examine itself to determine which parameters to adjust. This process is repeated until sufficiently trained. This is very good in the example of a stock market because there are years of previous cases that can be used to teach the system. Some more pros to this system are that it can very accurately approximate complex non-linear functions while keeping the calculations relatively simple so there isn't a need for powerful hardware or processing power. This solution is the most able to solve the subproblems mentioned earlier it would easily be able to generate rules and relationships based on a vast amount of complex data even if there is missing or noisy data.

However, the model can be overtrained in which case the network learns the noise in the data, e.g. day to day movements in stock prices, rather than the underlying patterns, e.g. long-term trends of stock prices, which is what a financial analyst is mostly interested in.

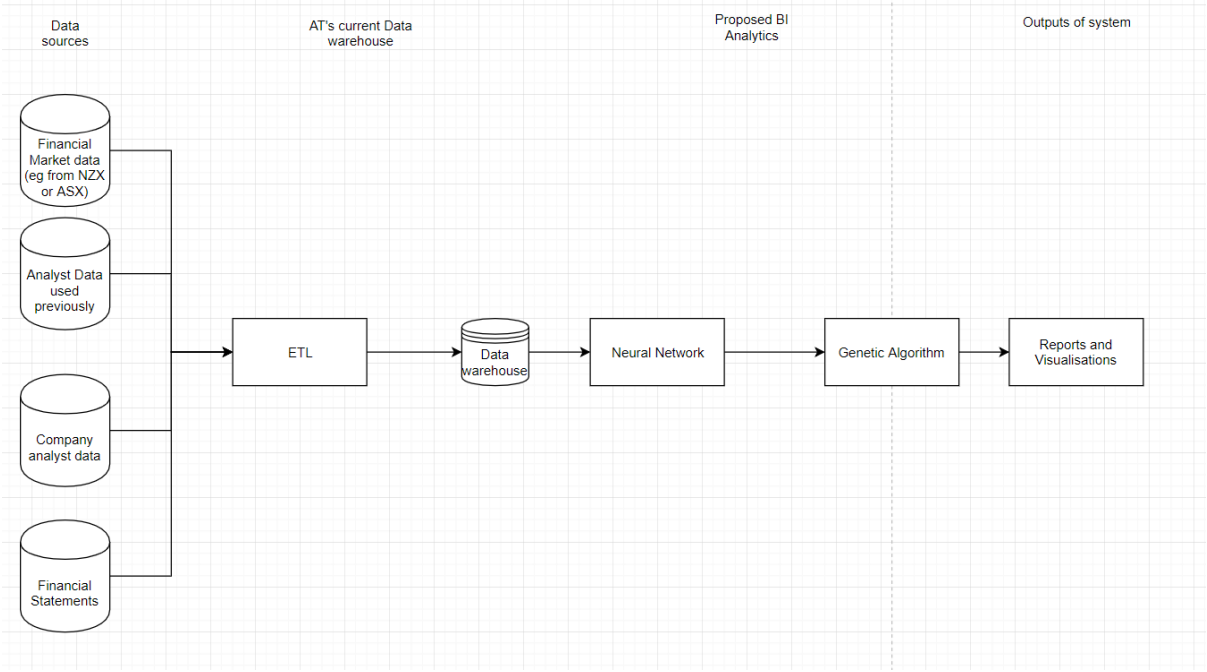
Dimension	Target Solution	Artificial Neural Network
Accuracy	Moderate	High
Explainability	Moderate	High
Response Speed	Moderate	Moderate
Scalability	Moderate	High
Flexibility	High	High
Embeddability	High	High
Tolerance for complexity	High	High
Tolerance for noise in data	High	High
Independence from experts	High	High

Genetic Algorithms

A Genetic Algorithm is a data mining method that uses optimisation; the term refers to a set of adaptive procedures applied within a computer system. It starts by making random guesses on how to solve an optimisation problem. The system then ranks the solutions and chooses the best one. Next, it removes bad answers and replaces them with new solutions made by combining bits of good solutions and records interim results, in some cases an entirely new solution will be introduced to the population and allowed to evolve.

Dimension	Target Solution	Genetic Algorithms
Accuracy	Moderate	High
Explainability	Moderate	High
Response Speed	Moderate	High
Scalability	Moderate	High
Flexibility	High	High
Embeddability	High	High
Tolerance for complexity	High	High
Tolerance for noise in data	High	Moderate
Independence from experts	High	High

Conceptual Framework Diagram



The data sources for the system are the financial markets, previous analyst knowledge and relationships, company analyst data and financial statements. Financial market data is the data gained from organisations such as NZX and ASX who show historical trends in stock prices and information about each company. Data that analysts used previously can also be used as input data. Company analyst data is data each individual company may provide shareholders about their predicted financial position and plans for the future. Financial statements for each firm describe the current financial situation of firms with assets, liabilities, costs, revenue and profits and other such information.

The ETL Extracts transforms and load these inputs and imports them into the data warehouse. From there it is sent to the Neural Network to be analysed. The genetic algorithm then uses the data and outputs to optimise the neural network solution. The final outputs of this are reports and visualisations that can be used by analysts and portfolio managers to make investment decisions.

Detailed instructions

My final choice of algorithm is the combination of an Artificial Neural Network and a Genetic algorithm. The neural network would be used to make stock decisions based on relationships it has gleaned from the data inputted. The genetic algorithm would be used to optimise the neural network solution. The genetic algorithm will choose the best parameters for the neural network to use to prevent it from overtraining itself.

An example of a gene that we would use is the name of the city that the firm works in. A chromosome is a series of genes, an example of which is a group of cities such as all the cities that a firm operates in. A decoder would then tell the system the significance of the chromosome, for example, that it represents cities the firm operates in. Overall the system can be used to find the best-combined parameter values in a trading rule which can be built into the artificial neural network. The neural network then uses these trading rules to make guesses on stock market values. The ANN can adapt to changing variables, so it will be able to handle these changes.

When the system is first introduced, the traders can define a set of parameters which can then be optimised by the genetic algorithm. These parameters are different factors that may affect stock prices such as natural disasters and improvements in a country's financial performance. These parameters are then fed to the Artificial Neural Network to make predictions on changes in the stock market. The Artificial Neural network then uses its own prediction rules stored in the hidden layer below to make predictions which are outputted through the output layer below.

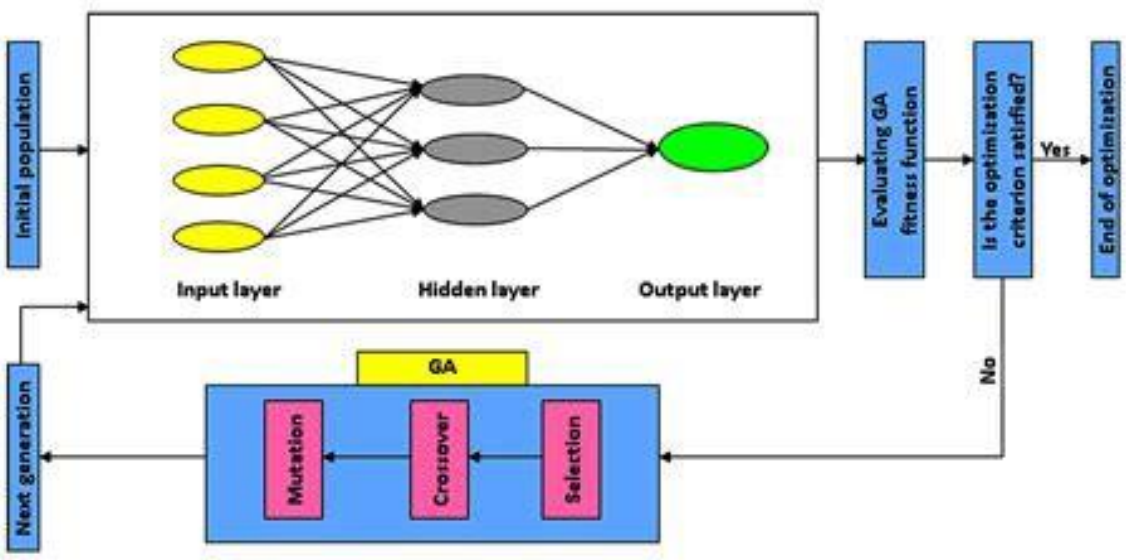


Image from: https://towardsdatascience.com/gas-and-nns-6a41f1e8146d

As shown in the diagram above the GA would evaluate whether the solution gained is similar to the one expected. If it is right, nothing happens. However, if it is wrong, the Genetic Algorithm will use Selections where it chooses rules generated that were accurate and gets rid of ones that were wrong. It also uses crossovers where it picks parts of rules that were right and combines them with other parts that were also right. It also applies Mutation where it will change parts of a rule randomly and input these into the next generation to be tested with.

Critique

Overall this system matches quite closely the requirement specification listed above covering the main problem base, a possible set of flaws that this solution may present is it may be expensive to use a complex composite system such as this. The Artificial Neural Network may also overtrain itself

if given too much information to learn from, if this happens it will stop providing overall trends in stock prices and start showing day to day movements which is known to be very volatile. However the genetic algorithm will act to prevent this.