LUISS

Department
of Business and Management

Course of Digital Finance

Machine Learning application in the investment management Industry. A case study on AQA Capital.

Prof. Paolo Bonolis
SUPERVISOR

Prof. Michela Altieri
CO-SUPERVISOR

ID No. 715031 - Alberto Tittoni
CANDIDATE

Academic Year 2020/2021

Abstract

This thesis describes how to handle the transition to become an AI based Modern Investment Management Company.

It starts by defining a general Artificial Intelligence theoretical framework with a focus on Machine Learning. The second chapter aims at defining the AI technologies used in the investment management industry.

A strategic model to manage the transition to a modern investment management company is then presented. It covers all the steps to successfully introduce AI models that are coherent with the long-term strategy of the company. The model is then empirically applied to the Asset Management and Fund Management company AQA Capital to build a Machine Learning Stock Price Prediction model using Supervised Learning.

Acknowledgments

First and foremost, I would like to express my deep gratitude towards my parents who invested in my education, and Alessandro e Francesca for their support.

I would like to acknowledge and thanks Elaine Bonnici, Christian Manicaro, Gabriele Rossi, all my colleagues at AQA Capital, and my professor at Luiss University Paolo Bonolis for believing in me and in this thesis.

I would also like to thank deeply and sincerely Carlotta, Giovanni, and Lorenzo for being my travelling companions, my guides and above all my friends in this intense 5-year journey.

Finally, a special thanks to John for being an inspiration.

Contents

Table of Figures

Introduction

Investment management industry finds itself at the cusp of a major transformation. From its incubation phase in 1956 until the state of art of today, a new set of technologies has extended its reach to almost all sectors in the modern economy. This set is classified with the name of Artificial Intelligence (hereinafter AI). Under the umbrella of AI technologies, one has blossomed for its promising features that are broadly applied in the finance sector. Its name is Machine Learning (hereinafter ML).

What brought me to dig deeper in the details of AI and ML was an elegant, yet intuitive ML model called "ALI". Here is a simple illustration of the potential of AI applied to finance as it is written by Al Naqvi in its book "AI in Asset Management"[1].

"By mid-January of 2020, ALI was convinced that within the next 60 days, the US stock market would decline down to the 18,000 to 19,000 range. ALI became suspicious when a news report about some type of a viral outbreak in China caught ALI's attention. Many of us missed that little news segment as it stood too far away to make a dent in the rapidly shifting consciousness of the modern world. But not for ALI. ALI stands for Artificial Learning & Intelligence. ALI is an intelligent machine, and its story to predict the Covid-19-related crash of 2020 is as follows. It was early January, and the world was focused on turmoil in the Middle East. A war was brewing, and geopolitical tensions were rising. Fear was in the air, but as we now know, for all the wrong reasons. ALI, who neither exhibits fear nor inhibits desires, was focused on something totally different. Ignoring all that was occupying our attention, it had picked up a trigger word related to a viral outbreak in China, and it was not ready to let go. Since viral outbreak could be the trigger

[1] Naqvi, Al. 2020. Artificial Intelligence for Asset Management and Investment. Wiley.

words for ALI to identify a potentially serious risk, it was holding on to it as a dog holds on to a bone. Suddenly, the pattern-seeking mind of ALI went in hyper-stimulated mode when ALI began discovering words such as "SARS," "pandemic," "viral outbreak," "panic," and "human-to-human infection." Like hammers pounding on ALI's consciousness, these word combinations made it go in a panic mode of its own. By the third week of January of 2020, ALI began screaming for attention. Logic dictates that the above information was enough to project that in a deeply interconnected world this virus would spread throughout the globe, that it would be devastating for people, and that it would lead to a catastrophic negative financial impact. When we studied the spread of the news about coronavirus, we were able to estimate by what time markets would get infected enough to respond to the news. With historical responses for such events fed into another machine learning algorithm, we projected that the market would decline to 18,000 to 19,000 within two months. On March 23, 2020—almost 60 days from our projection—DJI declined to 18,591."

The lessons I learned from this book together with my passions for technology and my experience in the fund management company AQA Capital, gave me the motivation to write my own ML model to predict stock prices.

Creating an ML model revealed to be a challenge for two different reasons. One is purely technical, as tuning an ML algorithm requires skills in a vast number of fields such as, computer science, programming, statistic, algebra, finance etc. The other is of strategic nature. Aligning the model objectives with the ones of the company that uses them and at the same time maximise the useability of the model across the company's function was harder than expected.

For this reason, I started my own literature review on the matter and the following is what I found out. On one hand there are companies that are understanding the importance of anticipating the future by adopting new technologies. On the other there those more reluctant to change and still attached to their legacy processes but want to keep up with the rest of the industry. In both cases their willingness to progress is often exacerbated by the complexity of the transition. The problem is that for the sake of introducing new technologies at all costs the transition itself is often underestimated. Without a pre-emptive framework definition, the risk is that the ultimate result of such a transition will be an AI artifacts accumulation with a suboptimal synergies exploitation and without a cohesive attention for the strategic goal of the firm.

The main aim of this thesis is to create a broad strategic framework in which investment management companies can introduce modern AI solution to problems in a coherent and cohesive way with their broader medium long-term strategy. I will then empirically apply this framework in the company where I currently work (AQA Capital) by managing the introduction of a self-made ML stock price prediction model.

The thesis will start by giving the reader a broad understanding of AI concepts focusing on the subcategory of ML technologies. Particular attention will be given to ML categories and algorithms. After a brief introduction on the investment management industry, the analysis will then go into the more specific concept of ML application to the investment management sector. Finally, the AQA Capital case will be explored. As the company is undergoing a path to embrace and enhance the use of technology in the next year, this thesis shows a viable solution with a strategic framework to manage this transition. The framework will be applied to create and introduce the stock price prediction model that is presented in the last chapter.

Chapter 1

Theoretical framework

"The future is ours to shape. It's a race between the growing power of the technology and the growing wisdom we need to manage it."

Max Tegmark

1.1 Artificial Intelligence brief history

Genesis era (1956-1974)

It was a hot summer in Hanover, New Hampshire. The year was 1956, and ten scientists decided to spend the six-week of that summer in an extended brainstorming session about their passions: neural nets, automata theory, computer science. These workshops are a fantastic way to let new ideas flow and generate, but when you put together sharp minds in new-born fields with a contagious sense of optimism, interesting and radical things can emerge. That optimism is evident in the letter they sent to the Rockefeller Foundation to get funds for the workshop:

"We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of

these problems if a carefully selected group of scientists work on it together for a summer".[2]

This summer project will be ever remembered as the genesis of Artificial Intelligence as a field of research and the participants are often described as their founders.

The following years were a constant strive to fight scepticism over the fact that machine were able to operate certain tasks usually attributed to humans. AI researchers aim at the time was to provide proof of concept of those tasks through some ingenious and smart systems.

In 1956 the "first Artificial Intelligence program" was created by Allen Newell, Herbert A. Simon, and Cliff Shaw. It was the first program engineered to follow automated reasoning, proving that computers can also think non-numerically, a property usually associated with the mind. The program "would eventually prove 38 of the first 52 theorems in Whitehead and Russell's Principia Mathematica and find new and more elegant proofs for some." (McCorduck 2004).

In 1963 the first machine to operate Machine Learning tasks was born with the name of SVM[3]. The first chatbot of the history was born in 1965. Its creator, Joseph Weizenbaum, called it ELIZA and it was capable of simulating dialect properties of a therapist, and sometimes even tricking its users to believe it was human. This was also a proof that a machine can pass the Alan Turing test[4]. In 1966 Shakey The Robot (called like this for his peculiarity of shaking during its operations) was able to combine robotic with computer vision and NLP[5], showing that a machine could link logical reasoning with physical actions.

More advanced programs were built in the following years. The problem was that, as they were able to solve defined and narrow problems, they become terrible when the number

[2] McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, e Claude E. Shannon. 1955. «A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.» AI Magazine, 31 August.

[3] Support Vector Machine

[4] A test of a machine's ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human

[5] Natural Language Processing

and the complexity of problems increased. One reason for that the so called "combinatorial explosions" of possibilities for exhaustive search-based methods.

"To overcome the combinatorial explosion, one needs algorithms that exploit the structure in the target domain and take advantage of prior knowledge by using heuristic search, planning, and flexible abstract representations capabilities that were poorly developed in early AI systems. The performance of these early systems also suffered because of poor methods for handling uncertainty, reliance on brittle and undergrounded symbolic representation, data scarcity, and severe hardware limitations on memory capacity and processor speed. By the mid-1970s, there was a growing awareness of these problems. The realization that many AI projects could never make good on their initial promised led to the onset of the first "AI winter": a period of retrenchment, during which funding decreased and scepticism increased, and AI fell out of fashion[6]" (Bostrom 2014).

Knowledge Period (1980-1987)

The first AI winter lasted for six years until the 1980. In this year Japan was living the "post-war economic miracle". Among the other investments, Japan government, together with private companies, decided to fund a massive architecture that would serve as a platform for developing AI based programs. Soon other countries all over the world started to follow the Japanese example investing huge amounts in AI. The centre of all investment were the so called "expert systems". Those were programs that, from a knowledge base of facts, made some simple inferences in order to help decision making processes. The knowledge base was provided by expert in the relative fields. For the first time in history AI programs were starting to find solutions to practical problems like diagnosing infectious blood diseases. However, as the investment were increasing reaching 1 billion dollar in total, some problems were soon to be raised. In particular, larger expert systems were really expensive to develop, and smaller systems proved to be inefficient. The second AI winter was coming with its contribution in freezing investment and enthusiasm.

Learning Period (1993-2011)

[6] Bostrom, Nick. 2014. Superintelligence. Paths, Dangers, Strategies. Oxford University Press.

In the early 90s the winter was gradually descending. One of the hardest obstacles to the development of AI, the low processing power, was being overtaken by the developing of Moore's law. This prompted the birth of new techniques like Genetic algorithms and Neural Network. The problem with old AI is their tendency to give completely wrong outputs as the inputs were changed only by a little. Neural networks instead have the property of small degradation: "a small amount of damage to a Neural Network typically resulted in a small degradation of its performance, rather than a total crash. Even more importantly, Neural Networks could learn from experience, finding natural ways of generalizing from examples and finding hidden statistical pattern in their input" (Bostrom 2014). There was no need to have an expert to determine the number of categories in which an output should be divided or to how the features were to be weighted. The model simply learns to figure it out itself. Moreover, the invention of back propagation and the already mentioned improvement in processing power contributed to a massive focus on these techniques.

Another factor that contributed to the development of AI was the abandonment of AI systems able to find solutions to general problems for more focused and narrower AI that search for solutions to specific subsets of applications. In this way the use of AI inside businesses became more approachable fostering an increased flow of investments in the field.

Among the other invention in these years, it is worth to mention the Deep Blue system[7] that in 1997 won against the world's most established grand master Garry Kasparov at the game of Chess. In 2002 the first household appliance able to learn the environment around him was created in the form of a vacuum cleaner. In 2005 we assisted to a big leap in autonomous driving with a car designed by Stanford University which was able to win the DARPA race. In 2011 Apple launches Siri, a virtual assistant based on speech recognition with a uniquely vocal interface.

State of Art (2011-2021)

In the last decade some important steps forward have been made. Especially in those areas considered as weak points for the development of AI. The power of processors reached

[7] The system was running on a supercomputer built by IBM.

an astonishing level contributing to the functioning of some complex types of algorithms. Even though some can say that Moore's law will eventually put a brake on further development of these technologies, a new kind of computers, Quantum Computers, are in their childhood and soon will gain popularity. Another important problem, the lack of data, is now not a problem anymore. Big corporations made the collection of data one their core responsibility and activity given the discover of their immensely huge value. As the sale of the so-called Big Data to advertising company wasn't enough to gain value, they can be fed to an algorithm to extract even more value.

Important steps forward have been made also in new advanced Machine Learning techniques. AI is now able to beat humans in a wide variety of games. In 2015 AlphaGo, a computer program designed by DeepMind, was able to beat the European champion at the game of Go. This game is considered to be one of the most challenging game ever created as the possible combination of moves are 10 to the 170, which is less than the number of atoms in the known universe. In 2017 AlphaGo managed to beat the strongest Go player in the world. These achievements can be misinterpreted as useless if we think that this technology could have been used for far more serious problems. Nonetheless, a couple of moves used by AlphaGo to beat the world champion were so clever, complex, and unconventional that no humans managed to even consider in centuries. This shows how new AI-made discoveries can be used as a positive multiplier for human ingenuity. It is also worth to remember that AlphaGo made all these achievements starting from a point when it did not even know the rules of the game.

Apart from games AI is used in so many fields. Modern speech recognition programs are used in our daily life with programs like Apple's Siri or Amazon's Alexa, able to understand simple questions and operate commands. Face recognition is widely used in USA and in Europe for visa purposes among the others. Surveillance institutions can manage millions of texts, images, email to protect citizens from terrorist or malignant attacks. Military and defence industry widely invest in AI for bombs-disposing robots, drones, and other unmanned vehicles. Internet services use AI for spam detection or fraud prevention and all other kinds of defence from cyber-attacks. Finance industry is using AI driven programs to exploit arbitrage opportunities. Automated stock trading and HFT[8] now account for the majority of equity shares traded in the US market.

[8] High Frequency Trading

Given the presence and the influence of AI-driven programs in our daily life, and the impact that they can have for our future, both in term of ethical and job deployment consequences. It is worth, for the purpose of this thesis, to dig deeper into the conceptual framework of these technologies and, further on, into their application in the finance world. In this way, every allegation against or toward AI will find a broader and resistant basement to make a judgement and to evaluate its impact.

1.2 Focus on Machine Learning

The first time the term "Machine Learning" was ever used in human history in the form we know today was in 1959. At the time, a computer scientist named Arthur Samuel was working on a program based on a curious intuition. He thought it was interesting to teach a computer program to play checkers in order to develop solution to general problems. As he wrote on the IBM's Journal of Research and Development his aim was to "verify the fact that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program."[9].

A program that had the ability to learn without the need to be explicitly programmed was a simple intuition yet, as often simple intuition revealed to be, wonderful and with complex consequences. This is also the definition of ML that we have today, but what did Arthur Samuel mean with the ability to learn without the need to be explicitly programmed?

First, the program was quite simple. He could not forecast all the moves until the conclusion of the game for every move he had to make, due to limitation in computer memory. He opted for a function that measured the chance of winning for him and his opponent at every given move[10]. Then he tried to maximize his function given the fact that the opponent would try to do the same[11].

[9] Samuel, Arthur. 1959. «Some Studies in Machine Learning Using the Game of Checkers.»
[10] This algorithm is called "alpha-beta pruning".
[11] Known as minimax strategy.

What about the learning process? Arthur Samuel thought of something he referred to as "rote learning". The program could remember every move he made, as well as the value of the function during that move. While the program gained experience from playing with opponents or against itself and the function improve from professional game input, it was effectively learning.

It is important to state that ML programs do need code. The difference with traditional programs is in the input. While traditional programs need an input command, a ML program need input data. This data is fed into the machine. The programmer chooses an algorithm and select the relevant penalty function to reach his goals. The algorithm adjusts them through a process of trial and error to discover hidden patterns in the data and build a model. Next time it will get data as input the program will be able to decipher the pattern and predict future values.

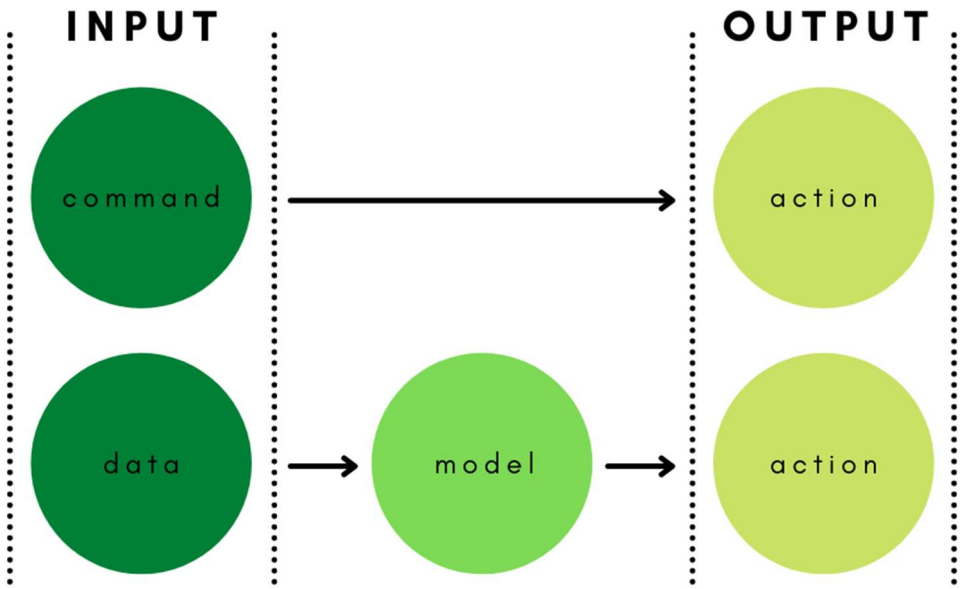


Figure 1 *- I/O process in Traditional computing and in Machine Learning.*

If we take as an example a program that analyses purchasing habits of TikTok users. The program could find a relevant relationship between TikTok users and Videomaking tools. However, the program was not explicitly written to detect this outcome. The input data and the algorithms were chosen by the programmer, but the prediction was created by the program through self-learning.

An effective analogy is comparing the process of building a data model to a guide dog. "Through specialized training, guide dogs learn how to respond in various situations. For example, the dog will learn to heel at a red light or to safely lead its master around obstacles. If the dog has been properly trained, then, eventually, the trainer will no longer be required; the guide dog will be able to apply its training in various unsupervised situations. Similarly, machine learning models can be trained to form decisions based on past experience."[12].

1.3 Machine Learning categories

There are 3 main categories of ML algorithms that are necessary to understand how hundreds of different combinations can be used to develop different programs. They are called Supervised Learning, Unsupervised Learning and Reinforcement Learning.

- **Supervised Learning**

Supervised Learning works by deciphering relationship between variables and outcomes that are given. The datasets will be consequently composed by labelled data:

$$\{(x^i, y^i)\}_{i=1}^{N}$$

x^i is called feature vector and $x^{1,2,\dots,N}$ are called features. They are values that play a role in describing the example and they contribute to describe y^i the label.

The program will be trained on labelled data, and it will try to figure out patterns in the data the describe the correct value of the label.

[12] Theobald, Oliver. 2017. Machine Learning For Absolute Beginners. Scatterplot Press.

The features could be for example the characteristics of a house (represented by x^i) that describe its price (represented by y^i). The program will analyse all the features and, given that it knows the price, can detect the degree to which these features contribute to the value of the house.

The ultimate goal of Supervise Learning models is to take a feature vector as input (x) and determine the most precise label (y). In the previous example if the model was correctly trained with a large and diversified amount of data, it can take features of a house as input and will give a prediction on the house price.

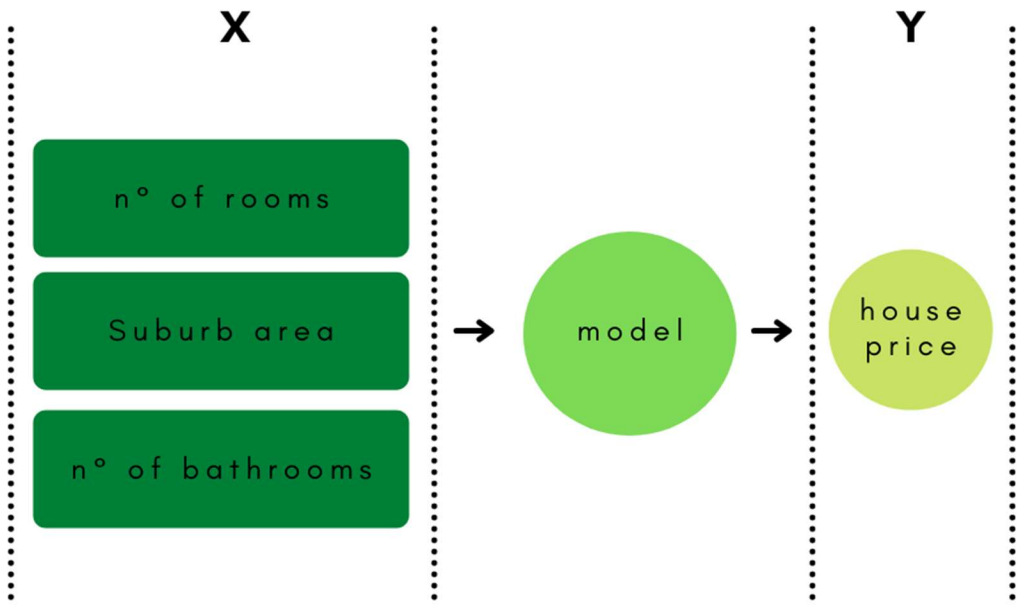


Figure 2 - I/O process in a Supervised Learning program of House price valuation.

- **Unsupervised Learning**

Unsupervised Learning works in a similar way of Supervised Learning. Nonetheless, it deals with unlabelled data:

$$\{(x^i)\}_{i=1}^N$$

x^i is still a vector of features but this time we do not have any label attached. The aim of Unsupervised Learning algorithms is to analyse data in order to detect pattern and find a viable label.

Here is an example of how useful an Unsupervised Learning algorithm called K-means clustering algorithm can operate a segmentation of the clients of a mall. Imagine you own a mall, and you have information on the clients by mean of their membership cards. From all the data you have you only need the clients' annual income and spending score to plot a graph.

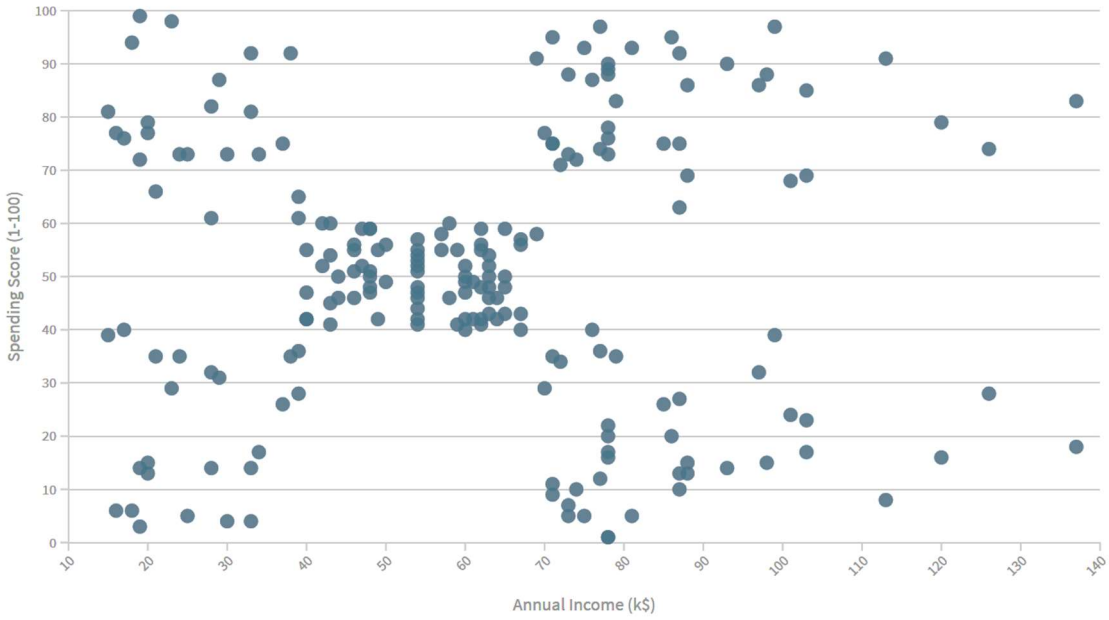


Figure 3 - Plot of mall clients' spending habits

The algorithm will discover 5 different clusters of data that will represent 5 different targets of customers without the need of specific labels that describe them:

Figure 4 - Clusters of clients highlighted [13] [14]

The important feature of Unsupervised Learning is the possibility to discover new clusters you were unaware of.

- **Reinforcement Learning**

Reinforcement Learning is the last and most complicated category of ML. In this case the machine will recognize to be present in an environment and will be able to detect the state of the environment. The machine has the possibility to take different actions that will bring to different results. These results are graded instead of being labelled. An action that takes the program closer to the goal will be graded positively and vice versa.

Reinforcement Learning can be explained through an analogy to a baby. Throughout his first years of life a baby starts with zero or few understanding of the environment around him. By taking different actions and becoming familiar with the environment, the baby will learn and will recognize the value of certain actions. Those values learned will become part of the experience of the baby and will influence its future behaviour.

The ultimate goal of the machine is to learn a policy. A policy is similar to the model we saw in previous categories. The policy takes information about the state as input and output the action that maximize the expected average reward, in other words the optimal action for that state.

The policy can be explained through the Pac-man example:

[13] Vijay Choudhary - Mall Customer Segmentation Data - https://www.kaggle.com/roshansharma/mall-customers-clustering-analysis .
[14] The blue dots are called centroid.

"A specific algorithmic example of reinforcement learning is Q-learning. In Q-learning, you start with a set environment of states, represented by the symbol 'S'. In the game Pac-Man states could be the challenges, obstacles or pathways that exist in the game. There may exist a wall to the left, a ghost to the right, and a power pill above—each representing different states. The set of possible actions to respond to these states is referred to as "A". In the case of Pac-Man, actions are limited to left, right, up, and down movements, as well as multiple combinations thereof. The third important symbol is "Q". Q is the starting value and has an initial value of "0." As Pac-Man explores the space inside the game, two main things will happen:

1) Q drops as negative things occur after a given state/action
2) Q increases as positive things occur after a given state/action

In Q-learning, the machine will learn to match the action for a given state that generates or maintains the highest level of Q. It will learn initially through the process of random movements (actions) under different conditions (states)."[15].

1.4 Machine Learning Tools

Whatever the complexity of the program there are 3 main tools used in ML whose correct use and configuration are vital for the program itself to succeed. These 3 main tools are mainly the same for a beginner and for an advanced computer scientist, yet the way, the scale, and the extent to which they are used can differ a lot. They can be described as:

1. Data.
2. Infrastructure.
3. Algorithms.

1. Data

[15] Theobald, Oliver. 2017. Machine Learning For Absolute Beginners. Scatterplot Press.

Data can be described as the samples or the cases in the domain that defines the task you want to operate or the problem you want to find a solution to.

"What we call data are observations of real-world phenomena. […] Each piece of data provides a small window into a limited aspect of reality." (Zheng e Casari 2018)[16]

In the Supervise Learning framework data is made of examples that are composed by a collection of features that have an input element and an output element that will be predicted by the model. The input data can be used in a variety of forms that include, text, images, videos, time series, etc.

The most common form of data is structured data. This is the kind of data that you can find in an excel spreadsheet or in a CSV[17] file.

In Linear Algebra we can define it as matrices. In a matrix a single row represents an example with its various features. A single column represents a single feature for each example.

It is important to state that most of the time a ML program will not work on raw data. Raw data can be defined as the data taken from your domain to solve the problem you want to solve. Obviously, you want to define the problem as a preliminary act to the collection of data.

After that, a series of action must be taken on the data samples in order to meet the requirements the program we want to write have. In other words, we must take a sample of the data that best uncovers the underlying structure of the problem we want to solve and feed it to the algorithm we are using in the most efficient and effective way.

"A feature is a numeric representation of an aspect of raw data. Features sit between data and models in the machine learning pipeline. Feature engineering is the act of extracting features from raw data and transforming them into formats that are suitable for the machine learning model." (Zheng e Casari 2018)[18].

[16] Zheng, Alice, e Amanda Casari. 2018. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media.

[17] Comma Separated Value

[18] Zheng, Alice, e Amanda Casari. 2018. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media.

These are some of the actions that can be taken on data in order to make it digestible for the algorithm.

First, the algorithms speak the maths language. They expect numbers. In your sample there can be the most disparate kind of data from numbers to text to percentages to categories etc. The act of working on data to transform it in numbers is usually taken before the definition of the model and it's called "data preparation" or "data pre-processing ". It is worth to mention a particular technique to transform text-based features into binary language called "One-hot Encoding".

It is also important to try different algorithms that better suit the data you have. There is so many algorithms to choose from that is worth trying a few of them to get the best results. However, at the same time each algorithm has specific requirements. Some algorithms, for example, require that the input variables follow a Gaussian probability distribution, so you want your data to follow the same distribution.

Finally, a useful statement to remember to better grasp the crucial role of data is "garbage in, garbage out". In other words, the performance of the program can be only as good as the data It has been used to train it.

2. Infrastructure

The term infrastructure it is usually referred to all the various platforms and tools used to carry out the ML program. Starting from the most important part of the infrastructure we have the programming language. Nowadays, especially for beginners, Python has confirmed its supremacy especially because it has simple syntax, it can be used for data collection and data piping, and it is compatible with an abundant number of libraries. Also, programming language like C and C++ are viable alternatives as well as R for more statistical uses and MATLAB for its strength in solving algebraic equations. Other languages like SQL can be useful to better interrogate and make a better use of data, especially if you are working with big numbers.

Libraries are also a vital tool in every programmer toolbox. They can be used to load and work on a dataset, clean up and perform calculations on data, and implement algorithms. There are also specialized libraries that help with data visualization or a standalone software that has other visualization techniques.

From a more advanced standpoint, it is useful to adapt more powerful infrastructure like distributed computing and cloud provider to run data processing with Graphical Processors Unit (GPU) instances.

3. Algorithms

Nowadays, there is a vast and variegate ecosystem of algorithm to choose from. There are Supervise Learning algorithms like linear regression, logistic regression, decision trees, and k-nearest neighbours and Unsupervised Learning algorithms like k-means clustering and descending dimension algorithms. There are also more advanced algorithms like Markov models, support vector machines, and Q-learning. The most interesting ones are the Neural Network algorithms that will be explained further in the next chapters.

1.5 Machine Learning algorithms

Let us take a deeper look at some of the most used algorithms. This brief analysis will start from the simplest ones like linear regression, then growing in complexity to Neural Network algorithms and Decision Trees. The class of algorithms that will be encountered are:

- Regression analysis
- Clustering
- Artificial Neural networks
- Decision trees
- Ensemble modelling

Regression analysis

Regression analysis can be described as the simplest Supervised Learning algorithm. It is used to estimate real values based on continuous values. In other words, it is used to find the best trendline to describe a dataset.

Linear regression

The simplest kind of regression analysis is the Linear Regression where the best fitting line is a straight line.

If we look at this regression from an algebraic perspective, we have some labelled examples:

$$\{(x^i, y^i)\}_{i=1}^{N}$$

N= size of the collection

x_i = Feature vector (i = 1,2,…,N)

y_i = Target vector (i = 1,2,…,N)

The model we want to build is the following:

$$f_{w,b}(x) = wx + b,$$

Where $f_{w,b}$ means that the model is parametrized by w (a vector of parameters), and b (a real number).

The goal is to find the optimal values for w and b in a manner that the model makes prediction as precise as possible. To do that we need to minimize the following expression:

$$\frac{1}{N} \sum_{i=1\ldots N} (f_{w,b}(x_i) - y_i)^2$$

This is the Objective Function, and it is called a Loss Function because it gives a penalty for a misclassification of examples i. In particular, the name of this loss function is Squared Error Loss.

To better explain this technique an experiment will be conducted by taking a cue on Oliver Theobald's book[19]. Imagine you want to predict the price of the world's most famous cryptocurrency, Bitcoin. Your idea is to use linear regression, but you lack data. You start taking note of the price of Bitcoin and the date of your observation. You start on the 27th of January and you thoroughly go on with your data collection for 1 month.

Here is what your dataset looks like:

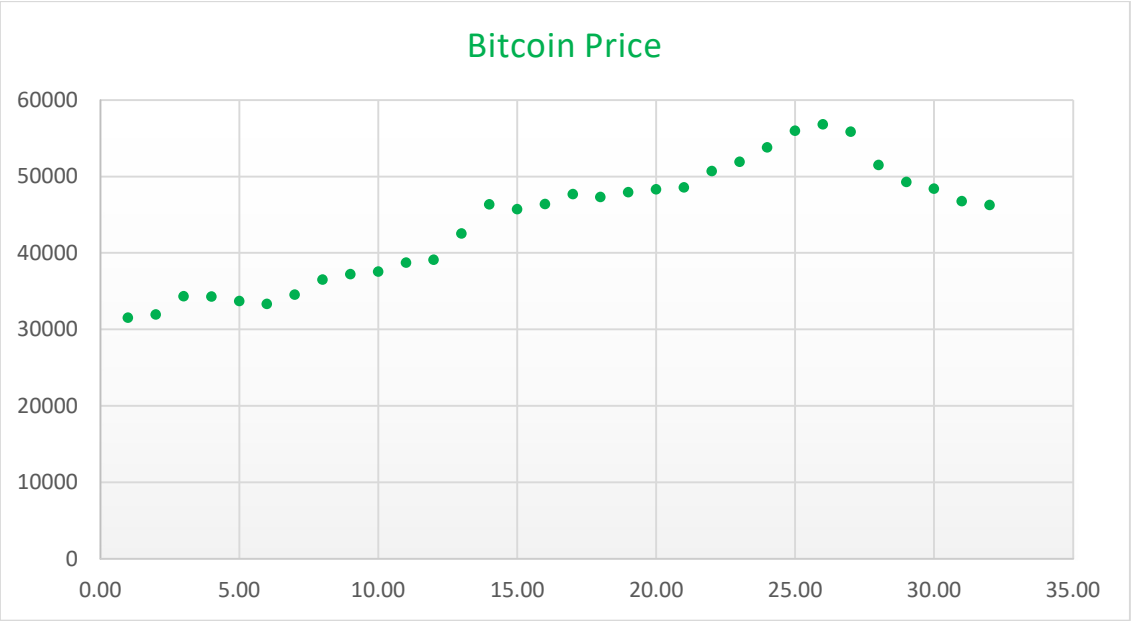
	A	B
1	**Date**	**Bitcoin Price**
2	27/01/2021	31498.28866
3	28/01/2021	31953.56909
4	29/01/2021	34317.52967
5	30/01/2021	34282.72829
6	31/01/2021	33692.61767
7	01/02/2021	33325.87703
8	02/02/2021	34521.74485

First it is necessary to convert date features into real numbers as the functions works only on real numbers. We can convert the single date into days from the first observation.

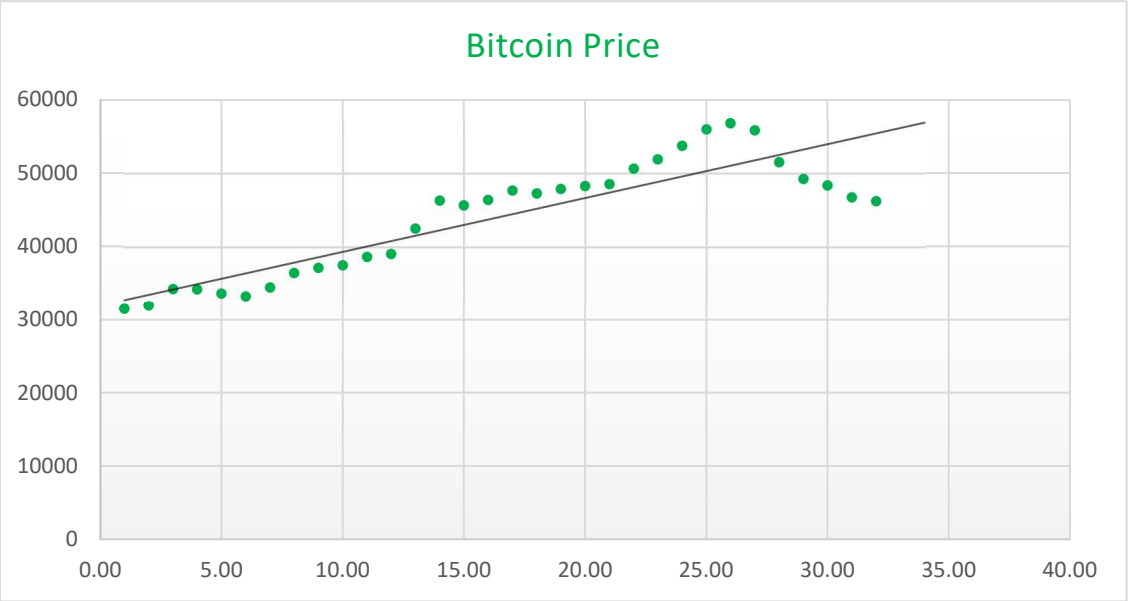
As your goal is to predict the "Bitcoin price", it will be plotted on the y axis while, "days from the first observation" will be plotted on the x axis.

Here is what the scatterplot looks like:

[19] Theobald, Oliver. 2017. Machine Learning For Absolute Beginners. Scatterplot Press.



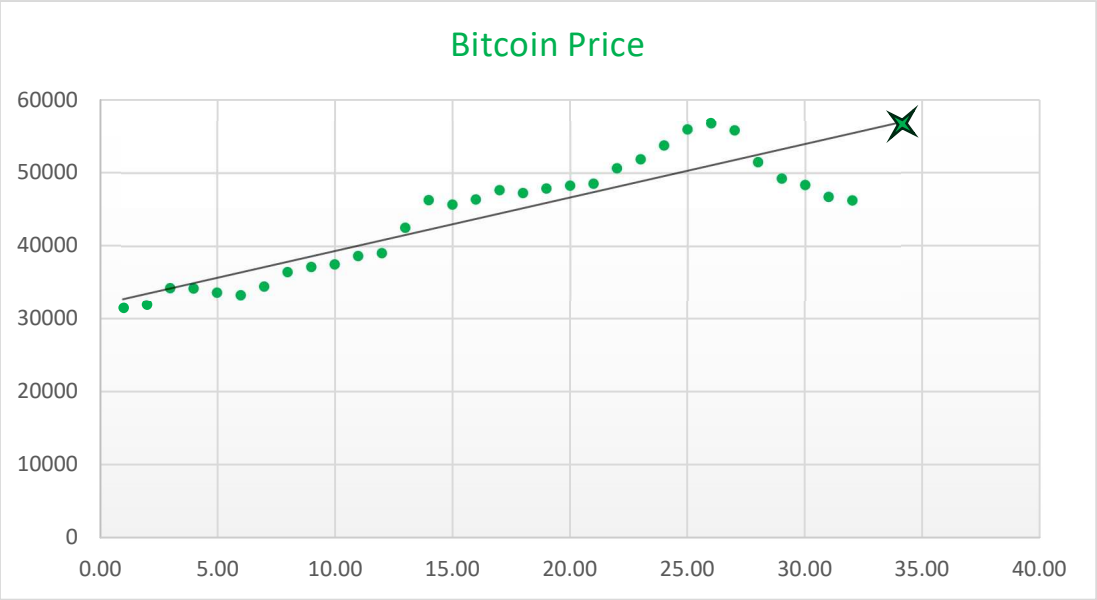
From the graph already a trend starts to appear. With linear regression it is possible to visualize the trend. The goal is to find a straight line that minimizes the aggregate distance of the data points from the line itself.



In ML the trendline is defined as Hyperplane and it could have different shapes.

While the linear regression is not fool proof when making predictions, it can be useful to understand the general future trend. If we want to know the price of bitcoin for the 28th

of February, we can see that the trendline suggest a price around USD $ 57000. However, it is more useful to know that it is predicting an upward moving of the Bitcoin price.



Logistic regression

Whereas Linear regression discerns relationship between variables through numerical predictions, the Logistic regression predicts discrete classes.

In other words, linear regression predicts continuous values like stock prices or house prices, whereas logistic regression predicts discrete classes like this person will buy an insurance or not, this email is spam or not.

There is no doubt that the prediction of discrete classes has a huge value in our society, but here is an example of how an algorithm based on a binary classification made possible for the famous supermarket Target to increase its revenues and to retain some special clients:

"Back in 2002, the discount superstore Target started looking for unusual patterns in its data. Target sells everything from milk and bananas to cuddly toys and garden furniture, and – like pretty much every other retailer since the turn of the millennium – has ways of using credit card numbers and survey responses to tie customers to everything they've ever bought in the store, enabling them to analyse what people are buying. In a story that – as US readers might know – became infamous across the country, Target realized that

a spike in a female customer's purchases of unscented body lotion would often precede her signing up to the in-store baby-shower registry. It had found a signal in the data. As women entered their second trimester and started to worry about stretch marks, their buying of moisturizer to keep their skin supple left a hint of what was to come. Scroll backwards further in time, and these same women would be popping into Target to stock up on various vitamins and supplements, like calcium and zinc. Scroll forwards in time and the data would even suggest when the baby was due – marked by the woman buying extra-big bags of cotton wool from the store.

Expectant mothers are a retailer's dream. Lock in her loyalty while she's pregnant and there's a good chance she'll continue to use your products long after the birth of her child. After all, shopping habits are quick to form when a hungry screaming baby is demanding your attention during your weekly shop. Insights like this could be hugely valuable in giving Target a head start over other brands in attracting her business. From there it was simple. Target ran an algorithm that would score its female customers on the likelihood they were pregnant. If that probability tipped past a certain threshold, the retailer would automatically send out a series of coupons to the woman in question, full of things she might find useful: nappies, lotions, baby wipes and so on.".[20]

While the ethical and legal consequences of certain algorithms will be discussed further in this thesis, let's look at the maths behind this matter.

We still need to find a target y as a liner function of the features x, but now y has only 2 possible values. We need a function with codomain (0,1) and, if the model output is closer to 0 or 1, we assign respectively a negative or positive label. This function is called standard logistic function or sigmoid function:

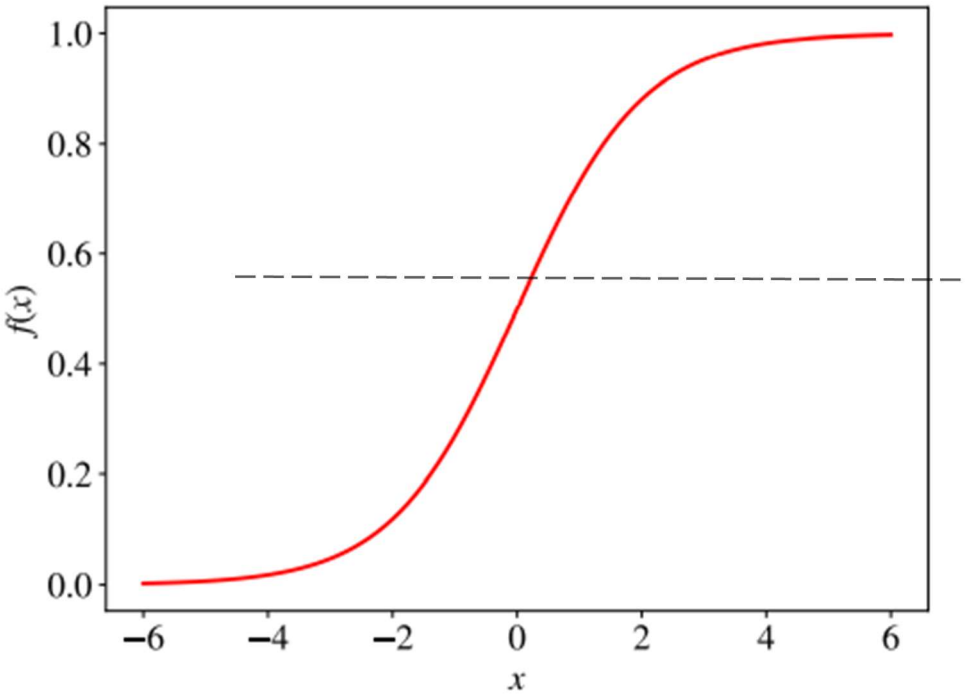
$$f(x) = \frac{1}{1 + e^{-x}}$$

And the Logistic regression model will be:

[20] Fry, Hannah. 2018. Hello world. Being human in the Age of Algorithms. W.W. Norton & Company.

$$f_{w,b}(x) = \frac{1}{1 + e^{-(wx+b)}}$$

If we consider a threshold at 0.5, we can now divide the data we have into 2 discrete classes above and below the threshold.



[21]

Figure 5 - Sigmoid function

To find a way such that the sigmoid function best fits our data we need an optimization criterion called Maximum Likelihood:

$$L_{w,b} \overset{\text{def}}{=} \prod_{i=1\ldots N} f_{w,b}(x_i)^{y_i}(1 - f_{w,b}(x_i))^{(1-y_i)}$$

[21] Theobald, Oliver. 2017. Machine Learning For Absolute Beginners. Scatterplot Press.

When y=1 we will have only $f_{w,b}(x_i)^{y_i}$,

When y=0 we will have only $(1 - f_{w,b}(x_i))^{(1-y_i)}$

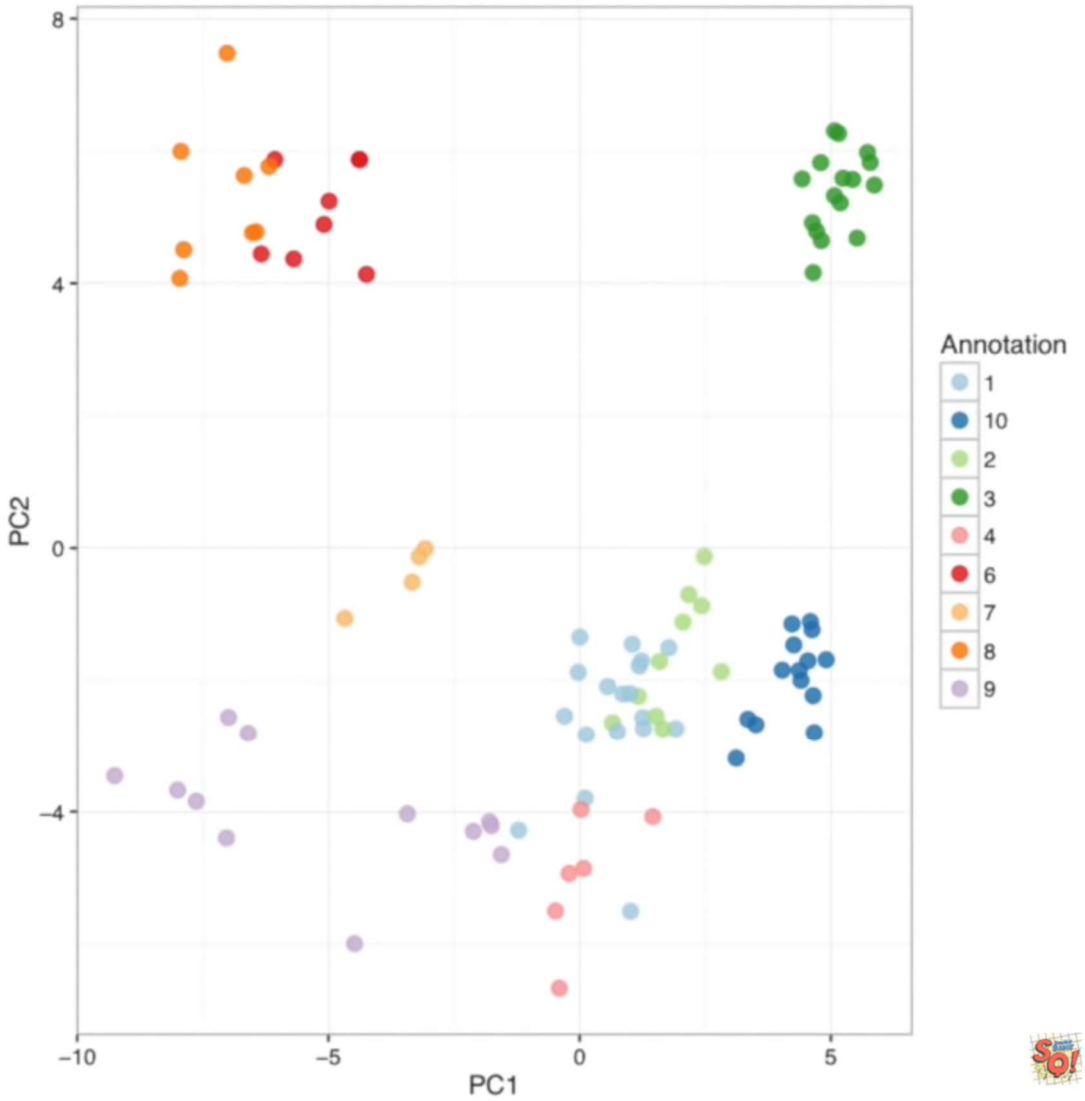
The product operator is used because we are dealing with the product of probabilities in order to get the likelihood. However, it is necessary to explain that it is preferable to maximise the Log-Likelihood.

Clustering Analysis

Another important way through which ML can extrapolate valuable information from data is Clustering analysis. These algorithms are useful to identify clusters of clients for example as it was explained in Machine Learning categories – Unsupervised Learning chapter. They can also be used to assign a new data point into existing clusters based on some information. Clustering analysis falls under the scope of Supervised and Unsupervised Learning and the 2 most popular algorithms are K-Nearest Neighbours and K-Means Clustering.

K-Nearest Neighbours

This algorithm works by assigning data points into clusters by looking at relationship with nearby data points. It all starts with a dataset with labelled examples called training data. For example, you could have different cell types from a tumour. Then the datapoints are drawn on a scatterplot like this:



22

Figure 6 - Scatterplot of different types of cell in an intestine tumour

The duty of the algorithm is to identify the clusters which a new data point belongs to. Here is a new type of cell taken from another tumour which we do not know the category yet.

[22] Starmer, Josh. 2017. «StatQuest with Josh Starmer.» Youtube. 26 june. https://www.youtube.com/watch?v=HVXime0nQeI.

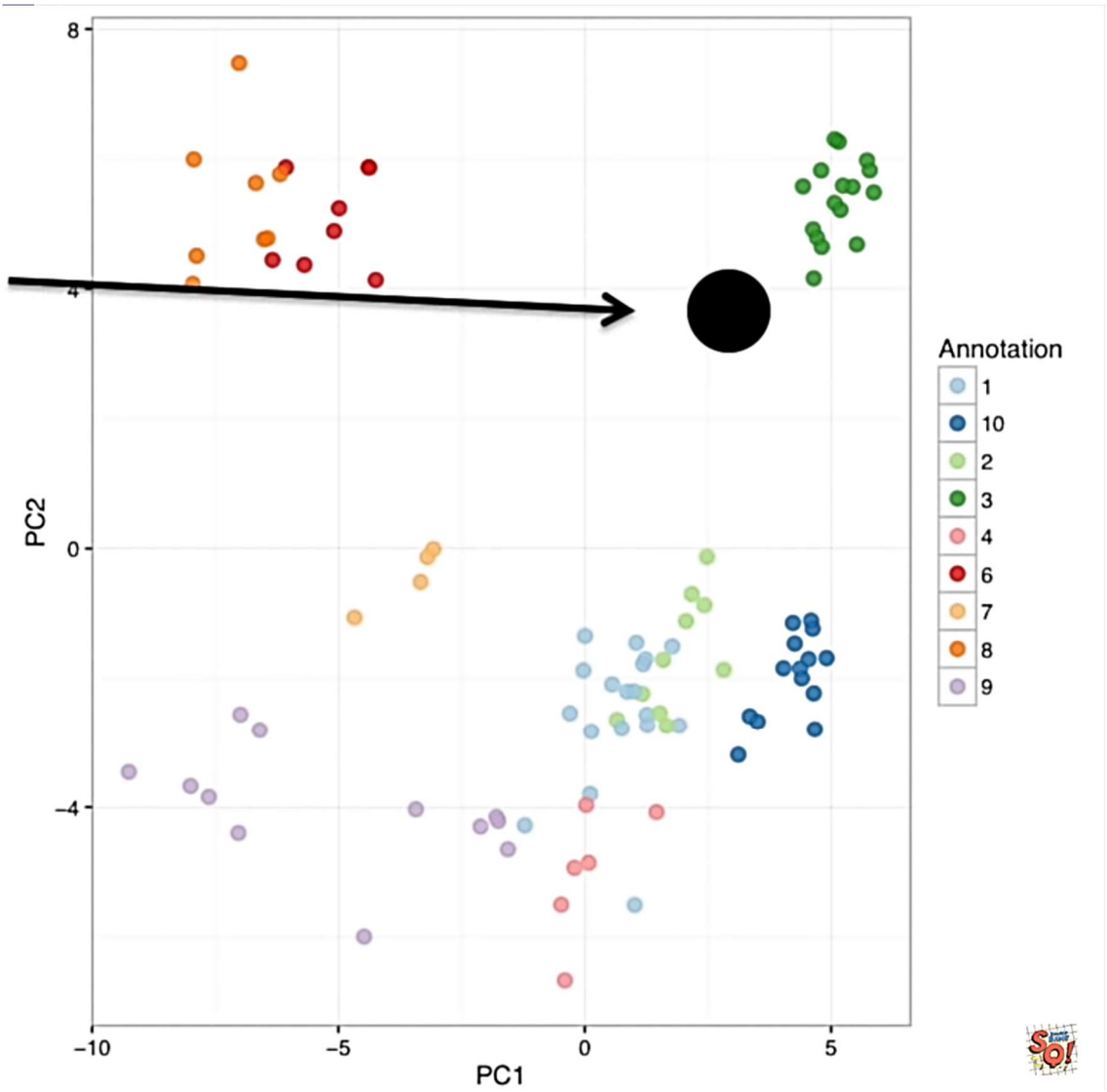


Figure 7 - New tumour cell with unknown category

The algorithm will assign this point to a category based on a parameter called k. Setting k = 1 simply means that the algorithm will assign the new point to the category of the nearest data point. In this case green:

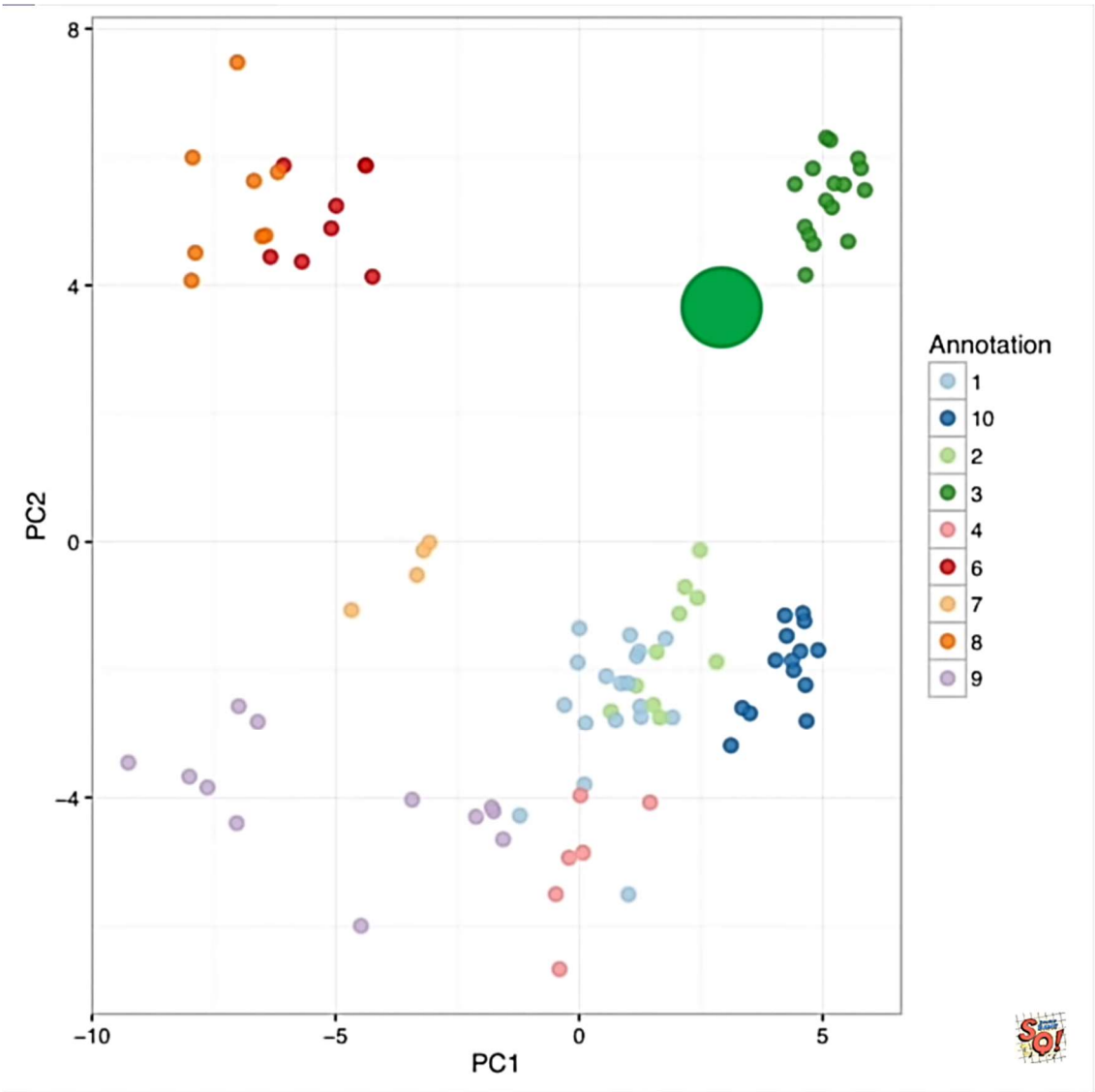


Figure 8 - Assignment of new tumour cell to the green category

Setting k = 11 means that the algorithm will assign the new point to the category of the nearest 11 data points. In this case 7 of the nearest points are red, 3 are orange and 1 is green so it will be assigned to the red category.

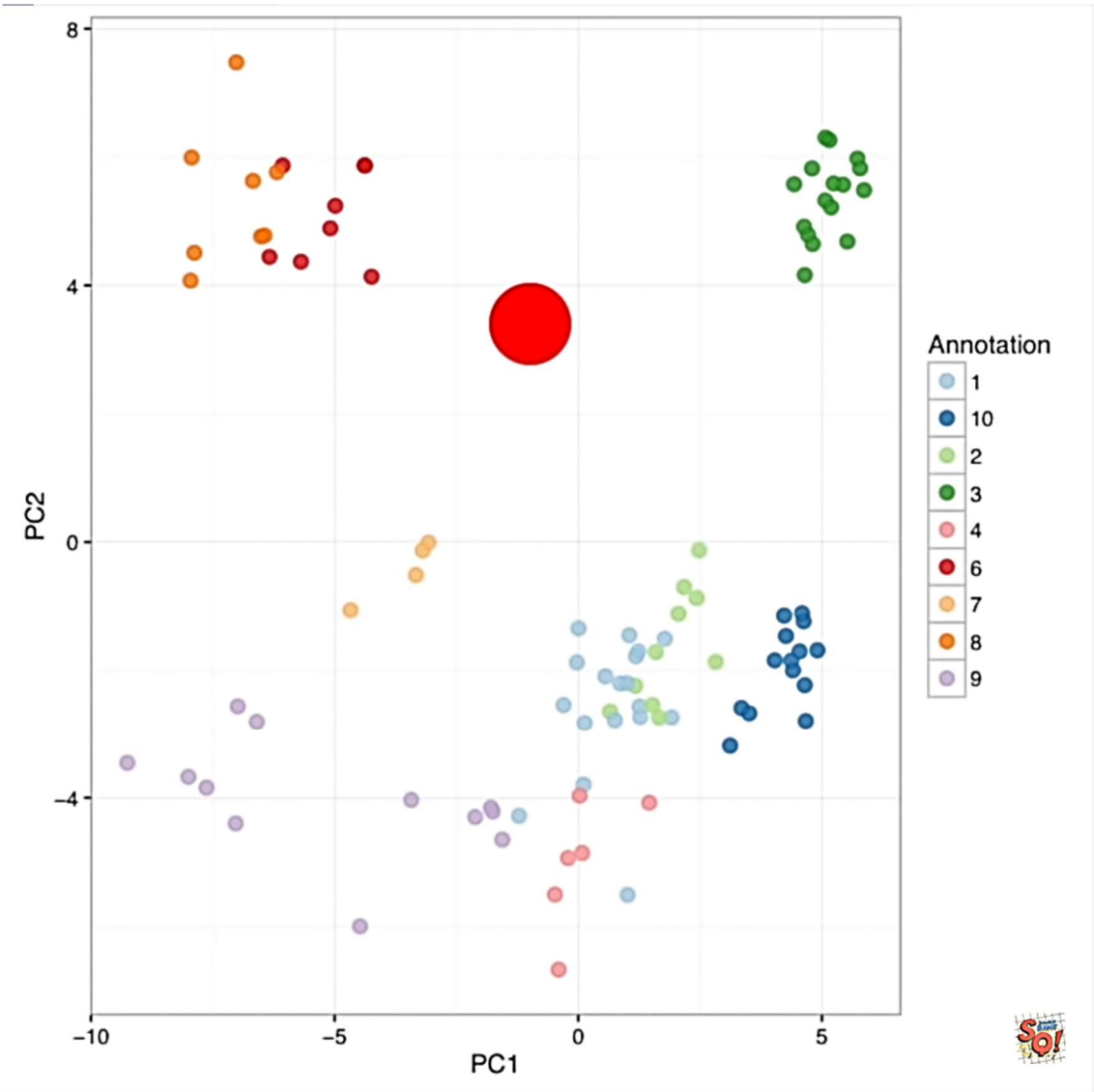


Figure 9 - Assignment of new tumour cell to the red category

The choice of K is of paramount importance. Unluckily there is no right way to choose k, so the best option is a trial-and-error process in order to find the right alternative.

K-Means Clustering

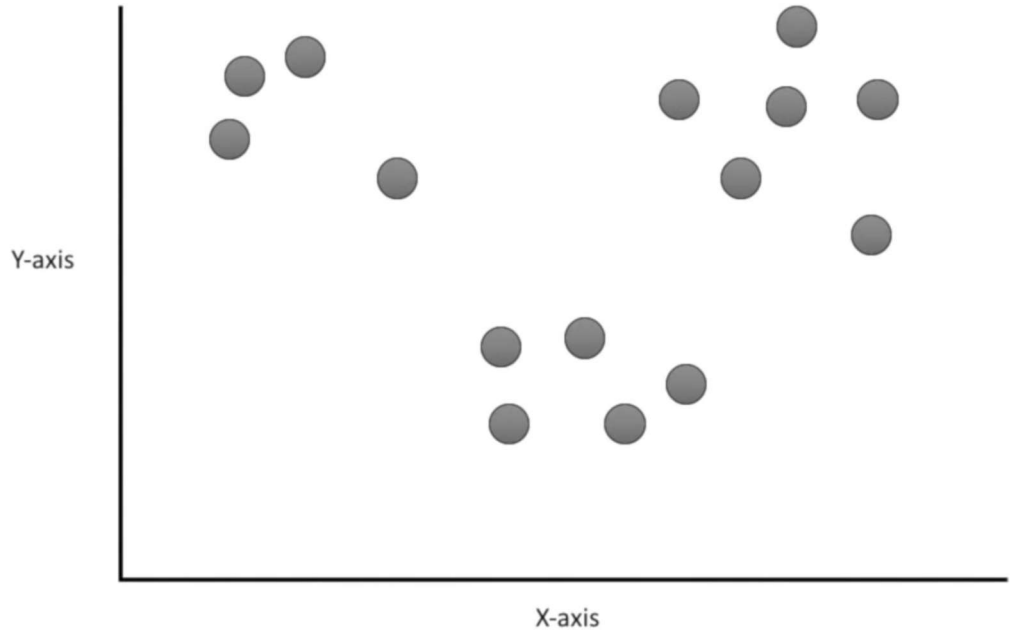
K-Means Clustering is a popular Unsupervised Learning algorithm

For instance, if you have data on different types of tumour cells, without knowing the category in advance, you want to assign each cell to the correct cluster of tumours.

First you must choose a value for k. This decision depends on the number of clusters you want to identify. If the number of clusters you need is unknown, you start with k = 1 and

then measure its total variation and iterate this process with k = 2 and so on. The total variation within each cluster is smaller when you add a cluster until the clusters equals the number of data points (total variation = 0). The right value for k will be the situation in which adding another cluster does not contribute to a high reduction in variance.

For this example, k = 3:



23

Figure 10 - Scatterplot of datapoints. 3 clusters are visible.

What the algorithm will do is take 3 random points (the blue, green and yellow dots in the figure) that will act as centroid or centre of the cluster, and then calculate the Euclidian distance from the other points with the following formula:

$$d = \sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2}$$

[23] Starmer, Josh. 2017. «StatQuest with Josh Starmer.» Youtube. 26 june. https://www.youtube.com/watch?v=4b5d3muPQmA&t=425s

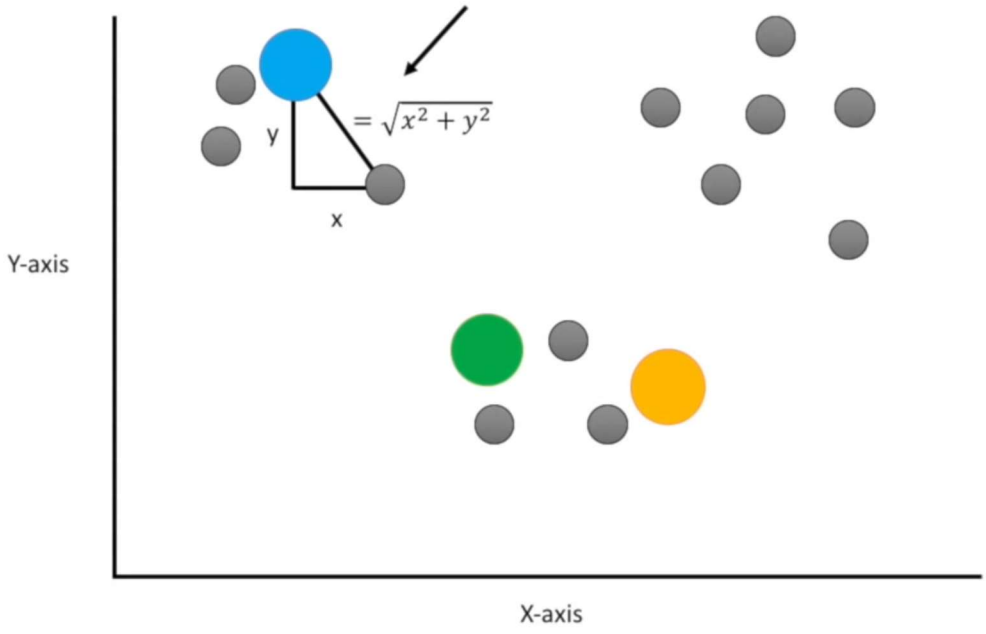


Figure 11 - The algorithm selects 3 random points, then calculate the Euclidean distance for each point

We can now assign each grey dot to the nearest cluster and calculate the centre of each cluster.

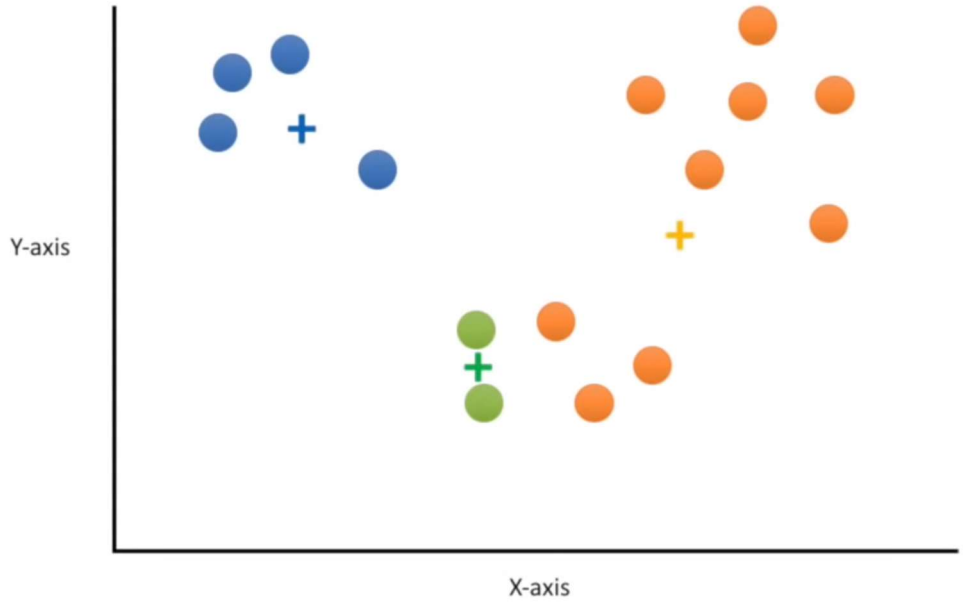


Figure 12 - The algorithm assigns each point to the nearest cluster and calculates the centre of each cluster (centroid).

The algorithm will calculate the mean value of all the x's and y's within each cluster and will plug in those values to update the coordinates of the centroid. This will result in a shift in the position of the centroid. By shifting position some points will switch cluster. The x's and y's mean value within each cluster will change and so will do the coordinates of the centroid.

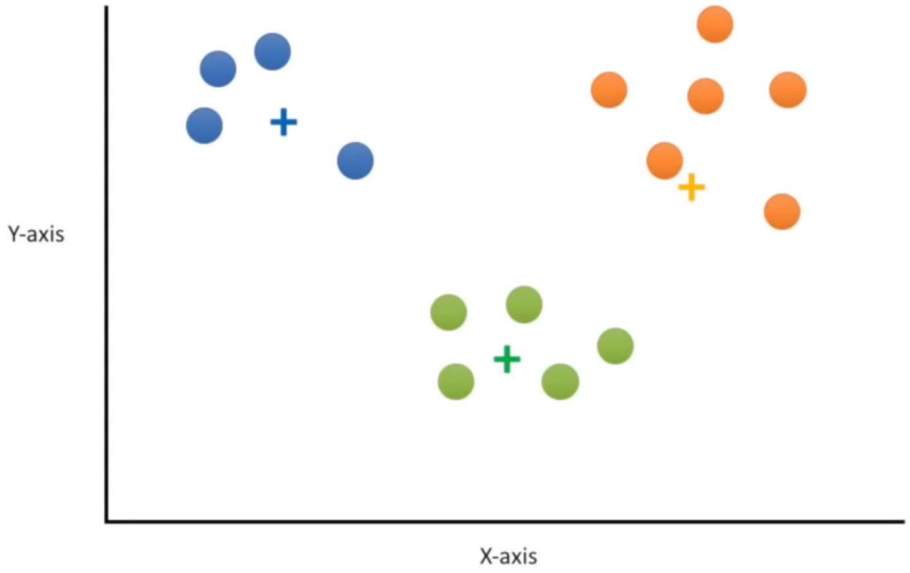


Figure 13 - The algorithm iterates the process until it finds the clusters with less variance within each cluster.

Finally, the process is iterated until no other points will switch cluster after the centroid's coordinates are updates. That is the optimal position for the centroids and the optimal size for the clusters.

Neural Networks

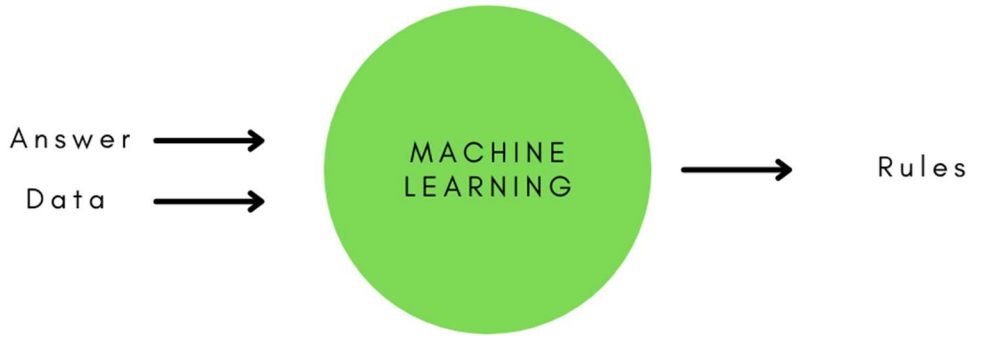
Neural Networks vs. Traditional computing

Imagine a Rock Paper Scissors Game where the computer recognizes your hand and plays with you. How would you write the code for this game? You will have to take images as input and recognize the position of the hand and fingers. You will have to tell the difference between the different figures. Practically it will require a huge amount of time to write all the code. Moreover, it will be a very difficult task.

The process of writing this program with traditional programming looks like this:

The programs will receive data (hands images), it will process the images according to the rules written on the code ("if – then" and other expressions) and it will generate answers (rock, paper, or scissor).

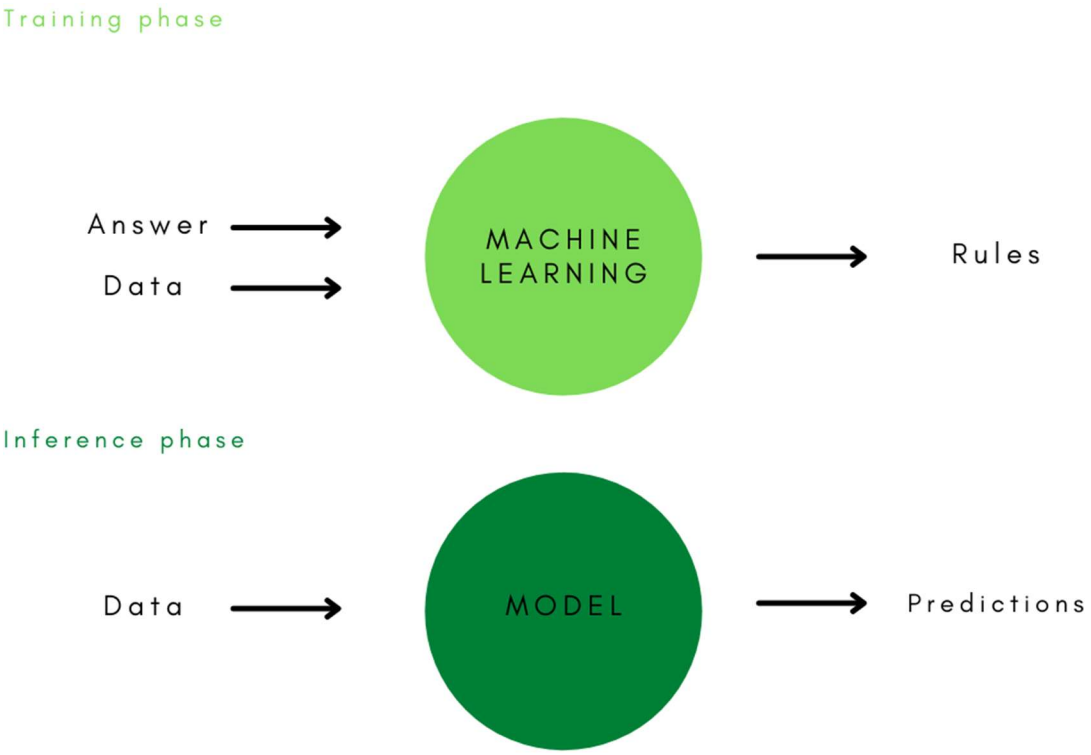
Machine Learning flip the whole process around:



The key element in this process is that writing the rules is really expensive in terms of time and complexity. The solution provided by ML is making the computer available with data (hands images) with a label attached (rock, paper or scissor) and let the program infer the rules that maps one or the other.

If the computer is fed in with the right data in terms of number, diversity of skin colour, texture, dimensions, etc. it will be able to recognize the patterns for us and we will have a ML program.

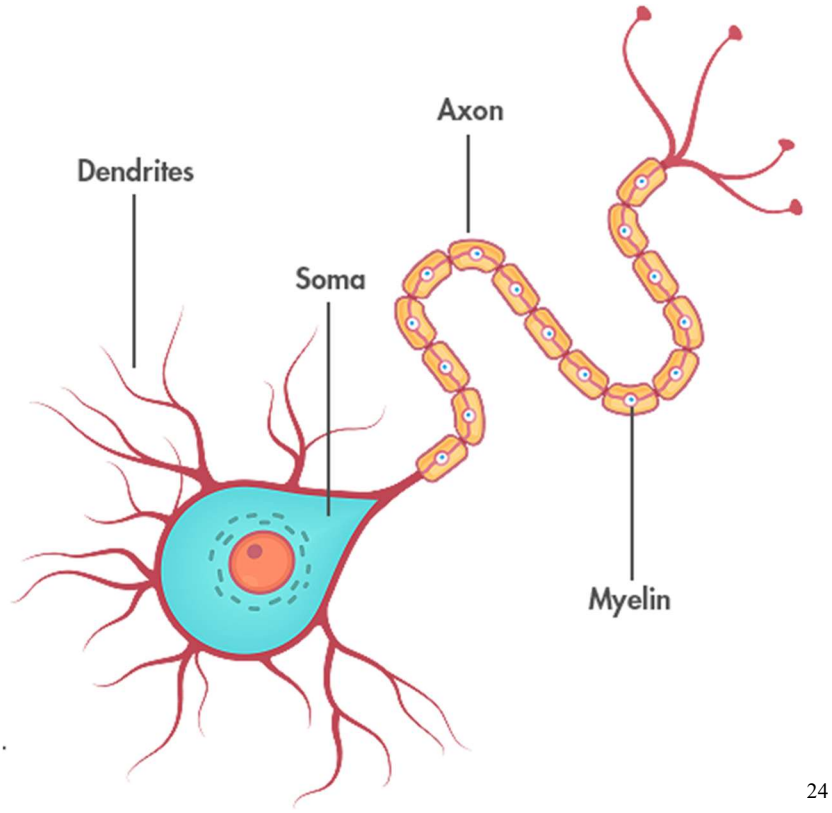
This whole process of feeding the algorithm is called training phase. The goal of the training phase is to come up with a model. The model we are dealing with is called Neural Network.



What a Neural Network does is simply getting data as input and giving predictions as output. In the Rock Paper Scissors Game example, if a closed hand is showed to a camera the computer will get the data from the hand and will generate a prediction like “this is 90% Rock, 5% Paper, 5% Scissor”.

Neural Network structure

A neural network is simply a fancy way to name models that try to emulate the functioning of the brain neurons. Obviously, the brain is still immensely more powerful and versatile of a neural network, but the intuition is clear.



[24]

Figure 14 - Brain neuron structure

A brain neuron is mainly composed by 3 parts. The Dendrites are tree-like structure responsible for receiving inputs from other neurons and bring them to the Soma. The Soma can be thought as the brain cell that receives input from the dendrites and decides whether the information is valuable enough to be passed along to other neurons. The passage occurs through the Axons that act as wires to bring information in the form of electrical signals.

[24] https://www.neuroskills.com/brain-injury/neuroplasticity/neuronal-firing/

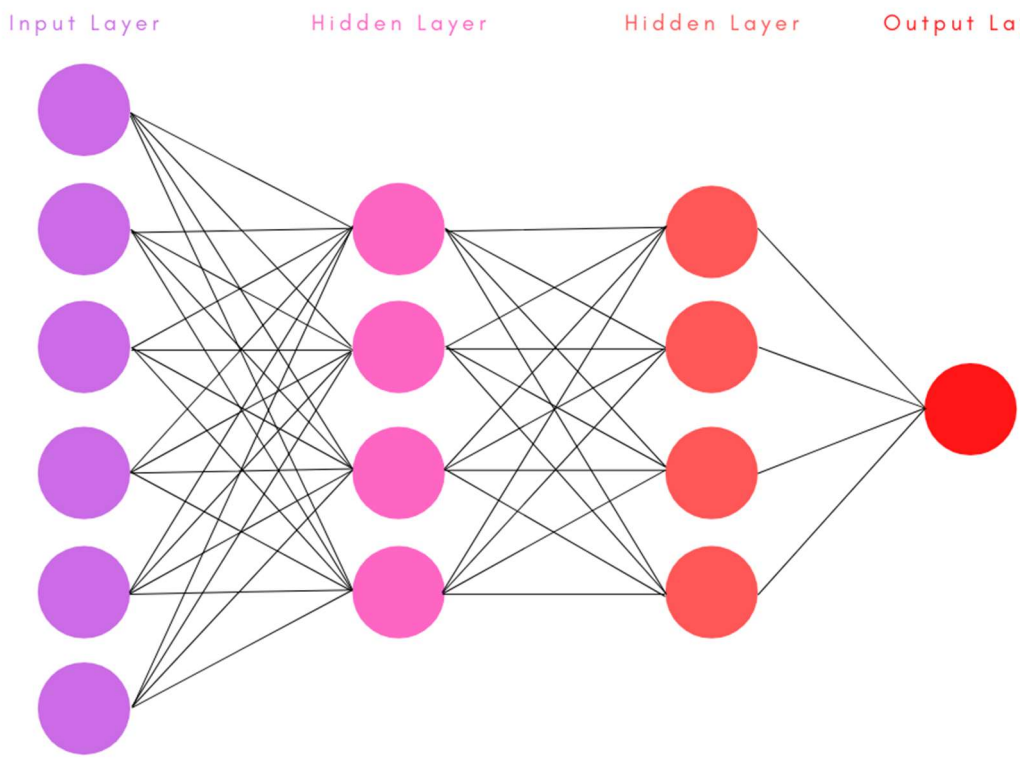


Figure 15 *- Neural Network layered structure*

Artificial NN were thought to resemble this process. The coloured dots act as neurons and are called nodes. The connections between the nodes act as axons and are called edges. The nodes store a numeric function. When they receive an input, they will fire in different patterns according to that input. The edges store a numeric weight. If the sum of the activated nodes times the weight satisfies a certain threshold, they will activate different nodes in the following layer. If the threshold is not satisfied, they will not fire the nodes in the following layer.

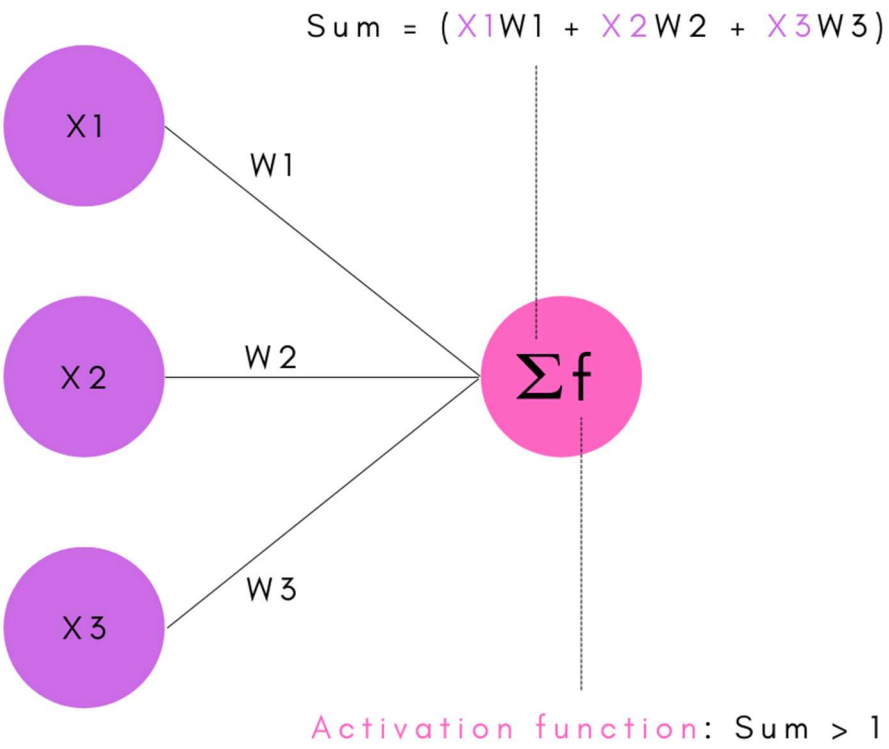


Figure 16 *- Explanation of how a node is fired. If the activation function is respected, the node will be activated.*

The first layer is called input layer and usually it receives data and stores it in a broad way. The hidden layers serve the purpose of breaking down the data between the input and output layers. Finally, the result is shown in the output layer.

Neural Network learning

The process of learning in a NN is the same state in the previous chapters for supervised learning. The Neural Network starts with random weights resulting in gross errors in the predicting the output. Nonetheless, the predicted output is compared to the actual output and the program measure the direction and the magnitude of the error (Cost value) through the so-called Cost function. The program then, will tweak and adjust all the weights accordingly to the error in order to give a better prediction for the next input. The final purpose is to reach the minimum cost value by minimising the cost function. Fair to say that this could result in an immensely difficult task as the simplest neural networks work with tens of thousands of weights. This whole process is called Back-propagation.

Neural Network "Hello, world"

In order to better grasp the concept of a Neural Network here is a simplified example that shows the code underneath a NN model[25].

Here we have two sets of numbers:

X: -1, 0, 1, 2, 3, 4

Y: -2, 1, 4, 7, 10, 13

It is easy to calculate the relationship between these lists:

$$Y = 3X + 1$$

A neural Network can be trained to solve an equivalent problem simply by feeding it with a set of X's and Y's. The NN can then infer the relationship between the two.

Looking at the code of this simple program it is easy to understand how a neural network works.

First, we import the libraries we will use:

```
import tensorflow as tf
import numpy as np
from tensorflow import keras
```

```
model = tf.keras.Sequential([keras.layers.Dense(units=1,
input_shape=[1])])
```

Hereabove we are creating a 1-layer NN with 1 neuron (units=1) and the input shape has 1 value (input_shape=[1]).

[25] https://developers.google.com/codelabs/tensorflow-1-helloworld#0

In order to compile the NN we need 2 fundamental functions: loss and optimizer.

```
model.compile(optimizer='sgd', loss='mean_squared_error')
```

The computer will try to learn by making guesses. A first guess could be Y=10X+10. The loss function will measure how badly this solution is and will give a value.

After that, the optimizer function will take the lead. Its role is to make another guess based on the previous loss function results. In other words, it will try a guess which is less bad than the previous ones like Y=5X+5.

Now the program needs to be feed with data. Data can be downloaded from library like NumPy. Here is an example of data that respects the relationship between X and Y.

```
s = np.array([-1.0, 0.0, 1.0, 2.0, 3.0, 4.0], dtype=float)
ys = np.array([-2.0, 1.0, 4.0, 7.0, 10.0, 13.0], dtype=float)
```

The last step is the training of the NN. This process relies on the function:

```
model.fit(xs, ys, epochs=500)
```

Through this function the program will run a loop in which it will consequently make a guess, measure how bad it is and using the optimizer to make another guess. It will do this for the number of epochs that will be specified (500).

In the next image it is clear that, with the first try the program is trying random guesses, but the loss is becoming ever smaller.

```
Epoch 1/500
6/6 [==============================] - 0s 16ms/step - loss: 57.4875
Epoch 2/500
6/6 [==============================] - 0s 173us/step - loss: 45.2285
Epoch 3/500
6/6 [==============================] - 0s 171us/step - loss: 35.5836
Epoch 4/500
6/6 [==============================] - 0s 195us/step - loss: 27.9955
Epoch 5/500
6/6 [==============================] - 0s 157us/step - loss: 22.0256
Epoch 6/500
6/6 [==============================] - 0s 132us/step - loss: 17.3287
Epoch 7/500
6/6 [==============================] - 0s 134us/step - loss: 13.6334
Epoch 8/500
6/6 [==============================] - 0s 129us/step - loss: 10.7262
Epoch 9/500
```

At the 50th epoch we can see that the training process made the loss extremely small

```
Epoch 45/500
6/6 [==============================] - 0s 122us/step - loss: 0.0016
Epoch 46/500
6/6 [==============================] - 0s 143us/step - loss: 0.0013
Epoch 47/500
6/6 [==============================] - 0s 136us/step - loss: 0.0010
Epoch 48/500
6/6 [==============================] - 0s 116us/step - loss: 8.2458e-04
Epoch 49/500
6/6 [==============================] - 0s 134us/step - loss: 6.6677e-04
Epoch 50/500
6/6 [==============================] - 0s 114us/step - loss: 5.4223e-04
```

The model is now ready to be used. We can use the model.predict function to verify the value of Y given a X. In this example for a X = 10 we expect a Y = 31

```
print(model.predict([10.0]))
```

The actual result will be something a little bit higher than 31 like 31.001. The reason is that NN works in term of probabilities. With only 6 data point we feed the program with, the NN will say that there is an extremely high probability that the relationship between X and Y is X = 30X + 1, but it cannot know for sure.

The more data the program will be feed with the more certain it will be with the model.

These little example covers all the fundamental tool to develop far more complex ML programs.

Chapter 2

Artificial intelligence in investment management industry

2.1 Overview

"AI will bring to the global economy as far as $15.7 trillion by 2030."

(PWC's 2019 report on AI)

From McCarthy's summer camp in the 50s, to the creation of Deep Blue by Ron Coleman and his team, to the actual state of art, AI has taken ever bigger incremental steps that already have reshaped in the past and will reshape in the future what we know as financial markets. To better grasp the magnitude of this reshaping process, a 2019 PWC's report predicted the AI will bring to the global economy as far as $15.7 trillion by 2030. In another report of McKinsey AI technologies will be able to afford 1.2% of global GDP.

As of today, AI has numerous applications in a wide range of fields. It is already dominating the trading floors of big banks, but it has also applications in lending, deposits, insurance, payments services, asset management, risk management, and the list goes on.

Technologies like neural networks and deep learning further contributed to the expansion of AI technologies use in financial industry. Just as neural networks are used in image recognition by analysing lots of inputs and breaking them down through hidden layers in order to have a reasonably correct output, the same principles are used to make predictions

about stock and bond market prices. A huge number of variables can be taken as input, like fundamentals and other companies' data, to make ever more correct predictions.

Other important progresses in NLP[26] are making chatbots more and more precise and effective, tackling that part of the financial industry closer to customers like customer services or preliminary sales.

The future adoption of this kind of technologies, however, come with some issues that will arise if preliminary and future actions will not be taken to avoid them.

AI is not perfect and has some limitations. These limitations can present themselves in the form of implementation challenges, ethical issues, and unintended consequences. Major solutions for these limitations will be further discussed in this thesis.

AI is rapidly changing the financial industry and has the potential to disrupt it. If we look at the big picture instead, it has also the potential to disrupt whole economies and change geopolitical orders. This is the reason why understanding the consequences of a mass adoption of AI is of paramount importance, especially in the finance sector.

The aim of this chapter is to briefly present investment management industry with a focus on definition of funds and their regulation. The focus will then shift on the use of Artificial Intelligence technologies and Machine learning model in the industry. Particular attention will be given to the applications in back-office operations and all the risk and challenges correlated. For every challenge a possible and ideal solution will be presented taking into consideration current state of the art models.

In the final chapter those solution will find real life applications and propositions to be applied in one of the most important Fund Management companies in Malta where I gave my personal contribution to reach business digitalization goals set by the CEO and the Founders.

2.2 Investment Management industry outlook

[26] Natural Language Processing.

Investment management is the activity that comprises the discretionary management of clients' assets. In the European Union, clients are normally classified into two distinct categories, namely, Professional Clients[27] and Retail Clients[28]. The assets under management can vary from plain vanilla financial instruments like bonds, stocks, Financial Derivative Instruments ("FDIs"), Exchange Traded Funds ("ETFs") to more exotic assets like commodities, real estate, artworks, cryptocurrencies and Non-Fungible Tokens ("NFTs").

In order to benefit from economies of scale, higher purchasing power, and diversification, certain clients prefer to invest in Collective Investment Schemes ("CIS"). CISs represent an alternative investment vehicle for those individuals who want to access the financial markets but do not have the adequate financial knowledge and/or time to attend to their own personal wealth

What is a CIS?

A Collective Investment Scheme as defined by FMSA (2000)[29] section 235(1) is *"any arrangement with respect to property of any description, including money, the purpose or effect of which is to enable persons taking part in the arrangements (whether by becoming owners of the property or any part of it or otherwise) to participate in or receive profits or income arising from the acquisition, holding, management or disposal of the property or sums paid out of such profits or income."*

Put simply, a CIS, or a fund as it is normally referred to, raises and pools capital from a number of investors, with a view to investing the capital in accordance with a defined investment policy, for the benefit of those investors. Investors acquire units or shares the collective investment scheme normally via a subscription. The shares represent a unit of proportional ownership over the assets of the fund and all the income, profits, or capital

[27] Professional Clients is defined in the EU Directive 2014/65/EU, commonly known as MiFID II as being entities that are required to be regulated or authorised to operate in the financial markets, large undertaking which satisfy certain size requirements, national and regional governments, public bodies managing public debt (excludes local authorities), central banks, international and supranational institution, and other institutional investors whose main activity is to invest in financial instruments, including entities dedicated to securitisation of assets or other financing transactions as well as other elective clients who demonstrate knowledge and experience in the financial sector.

[28] Retail Clients is defined by MiFID as well, as those clients who are not Professional Clients

[29] The Financial Services and Markets Act 2000

growth that it may create. As the value of the assets fluctuate in time so does the value of the shares to reflect this movement.

Investors have no day-to-day control over the management of the assets of the fund. CIS may be self-managed or third-party managed. A self-managed scheme is one which performs the investment management function internally through the appointment of portfolio managers internally. On the other hand, a third party managed collective investment scheme appointed a duly authorised management company to manage the assets of the fund. In view of the regulatory requirements revolving around the setting up and running of a CIS, especially in Europe, most promoters prefer to launch a third party managed CIS and appoint a regulated management company to perform the investment management function.

Investors' value in the collective investment scheme is linked to the Net Asset Value ("NAV") of the fund. The NAV is a formula that defines the value of the fund in any given moment. It is simply the value of the assets minus the value of the liabilities. The formula can be more sophisticated to reflect the needs to satisfy different jurisdictions or regulations. The NAV divided by the number of shares gives the price of a single shares of the fund, which is also the price investors pay to become shareholders. The NAV is normally calculated at the same frequency of the subscriptions and redemptions, and normally ranges from daily or once a year. An Exchange Traded Fund is a type of collective investment scheme whose price varies intra-day, since, as the name implies, the fund is traded on an exchange like a common stock.

CISs may also be set-up as multi-class or multi-funds. A multi-class fund is one which issues various classes of shares, but each share class does not constitute a separate legal entity, all share classes contribute towards the same investment objectives and strategy, but may differ in terms of, for example, investment base, distribution policy, minimum investment amounts, fees, and currency. On the other hand, a multi-fund schemes established sub-funds (also commonly referred as umbrella funds). Albeit the sub-funds within the same scheme do not have a separate legal personality, each sub-fund is a separate patrimony from the other sub-funds within the same scheme. This means that the assets of one sub-fund should not be used to make good for any shortfalls in another sub-fund. Each sub-fund can have a strategy and investment style. Moreover, each sub-fund may create more than one share class. Share classes may have different features, such as

(i) different eligible investors (ii) different fee structure (iv) different distribution policy, but they must follow the same investment objectives, policies, and restrictions.

Governance of a CIS

The founder shares of a corporate collective investment scheme are normally held by the promoters of the fund. Founder shareholders put in the initial share capital to set-up and incorporate the scheme. The founder shareholders normally retain certain rights, including the right to nominate and/or appoint the Directors of the fund. They however do not carry a right to participate in the assets of the scheme upon a winding up, other than for any remaining surplus after payment of all amounts due to creditors and investors.

The Board of Directors is responsible for the general affairs of the SICAV, of which the appointment of the service providers, including the Management Company and the Depositary Bank.

Regulated CISs are required to comply on an ongoing basis with the provisions of local laws and any applicable rules and regulations as well any other applicable regulations in any jurisdictions where it is marketed. In this regard, CIS are normally required to appoint a Compliance Officer. Moreover, the Board of the UCITS is also responsible for compliance with its obligations under the Prevention of Money Laundering and Funding of Terrorism Regulations, any rules issued thereunder, and for such reason, CISs must appoint a Money Laundering Reporting Officer.

Service Providers required to operate a third-party managed CIS

CIS require the input of several service providers in order to function properly, of which include:

- **Management Company.** In the EU, a company may only manage a CIS if it holds a specific authorisation under the UCITS Directive[30] or the AIFMD[31]. The Management Company is normally responsible for, *inter alia,* the portfolio management, risk management and valuation functions. These functions may also be delegated to third party entities, however, the Management Company has to retain certain functions internally in order not operate as a "letter-box entity".
- **Depository**. The Depositary performs the safekeeping of assets, monitoring of cashflows, and has an oversight function.
- **Fund Administrator**. The Fund Administrator is normally responsible for the preparation of the Net Asset Value ("NAV") of the shares/units of the CIS, reconciliations, fund accounting, payment of bills. The Fund Administrator normally also covers the role of the Transfer Agent, and therefore take care of the subscriptions and redemptions, keep the share register and issues contract notes.
- **Distribution Company**. A distribution company that oversees the selling of shares and marketing of the fund.

Legal Forms

The legal forms which CISs may be set up as differ from jurisdiction to jurisdiction. The below is a list of the most common forms used to set-up a CIS:

- **ICAV/SICAV.** Investment company with variable share capital. This is commonly referred to as an open-ended collective investment scheme, meaning that the fund is set-up for an indefinite time period. A ICAV is a company which enjoys separate judicial personality.
- **INVCO.** Investment company with fixed share capital, commonly applied for closed-ended funds, ie. those funds who are set-up for a definite period of time. An INVCO is a company and enjoys separate judicial personality.
- **Limited Partnership.** Commercial partnership constituted by means of a partnership deed. Limited partnerships do not possess legal personality and are

[30] Directive 2009/65/EC of the European Parliament and of the Council of 13 July 2009 on the coordination of laws, regulations and administrative provisions relating to undertakings for collective investment in transferable securities (UCITS).

[31] Directive 2011/61/EU is a legal act of the European Union on the financial regulation of hedge funds, private equity, real estate funds, and other "Alternative Investment Fund Managers" in the European Union

created by a partnership deed. There is the General Partner which assumes the responsibility for the management of the fund and the partners which are investors.

- **Unit Trust.** The two central figures in the Unit Trust are the Manager and the Trustee. The trustee and the Manager must be two separate and independent persons, and each has distinct duties. The trust manager runs the trust for profit, whilst the trustees ensure that the fund manager adheres to the investment objectives of the trust and safeguards the assets of the trust. The trustee is bound to carry out the instructions of the Manager unless such instructions are in conflict with the conditions of the Trust Deed or the license.
- **Contractual Fund. T**his type of fund does not have a separate legal personality and is set-up by form of an agreement between the management company and other parties, such as the investors (as the case for Italian contractual funds) or the Depositary (as the case for contractual funds set up in Malta.

The choice of the legal form tends to depend on the jurisdiction in which the fund is being launched, and the nature of the underlying assets. For example, the Partnership is commonly used in the UK and in Luxembourg for private equity funds. Contractual funds are common in Italy for real estate funds. SICAV/ICAVs are used in Luxembourg and Ireland for liquid and plain vanilla collective investment schemes. Whereas in Malta, the most common form is the SICAV, which works well for all types of collective investment schemes, open-ended, closed ended, real estate and private equity.

EU Directives regulating CISs

In the EU, CISs are regulated by two main Directives:

- Directive 2009/65/EC of the European Parliament and of the Council of 13 July 2009 on the coordination of laws, regulations and administrative provisions relating to undertakings for collective investment in transferable securities (UCITS).
- Directive 2011/61/EU is a legal act of the European Union on the financial regulation of hedge funds, private equity, real estate funds, and other "Alternative Investment Fund Managers" in the European Union (AIFMD).

UCITS Directive

Before the implementation of the UCITS Directive, investing in a CIS was not harmonised in the EU, and jurisdictions did not recognise the CISs of another jurisdiction, and hence marketing of funds in different EU member states was very difficult. In 1985, the EU decided to open the market by establishing a set of rules that grant a variety of fund option and additional protection for investors. UCITS funds can be established under a harmonized EU framework that give the possibility to sell a UCITS fund across the borders into other EU countries without further authorizations.

There are 2 main primary reasons for the EU launch of UCITS:

1. Create a single EU financial services entity that allow for the cross-sale of funds across EU borders.
2. Increase safeguard of investors and improve efficiency and effectiveness of regulation by creating more strictly regulated investment funds.

The UCITS Directive was designed to protect the weakest of investors, the retail investor. The protection for investors is mainly given by the respect of very strict investment restrictions and the imposition of a specific list of eligible assets. Moreover, UCITS directive gives indication on how to operate valuation on a funds.

The restrictions over the investments a fund can do are summarized in the tab below:

LIMITS FOR UCITS FUNDS	
SINGLE BODY EXPOSURE - Bonds and Equities 1. 35% Maximum exposure to: a. Member State Bonds b. Bonds Guaranteed by the Member State 2. 25%- 80% For Credit Institution Bonds with registered offices in a Member State 3. Otherwise: 5%-10%-40% rule	**BORROWING** 1. 10% Scheme may borrow up to a maximum of 10% of its assets when the scheme is set up as an investment company, a limited partnership or the UCITS is set up as a trust or common contractual fund and provided that borrowing is temporary 2. The company may not borrow for investment purposes

SINGLE BODY EXPOSURE – Collective Investment Schemes 1. 20% Exposure to one UCITS Collective Investment Scheme issuer 2. 30% aggregate exposure to non-UCITS Collective Investment Schemes	**GENERAL PROVISIONS** The UCITS Management Company may not acquire any shares carrying voting rights which would enable it to exercise significant influence over the man agent of an issuing body.
EXPOSURE- OTC Derivatives 1. 5% OTC broker exposure	**VAR APPROACH** Cannot exceed 20%
EXPOSURE- FDI (Commitment Approach) 1. Global exposure shall not exceed 100% of NAV 2. Overall risk exposure shall not exceed 200% of Total Assets	**SHORT-SELLING** A sub-fund may not carry out any uncovered sales of: 2. Transferable Securities 3. Money Market Instruments 4. Shares of Collective investment Schemes 5. FDIs
EXPOSURE- Deposits 1. 20% Deposits with a single Credit institution 2. 5% Deposits with a single broker	

Those investors can now count on more flexibility and security. UCITS funds are considered to be safer as they undergo sophisticated regulatory scrutiny, but they don't necessarily assure higher income.

UCITS fund have different regulatory and capital requirements to satisfy. Among the others:

- A tax-neutral country is usually chosen as base for the fund (Malta, Ireland, etc.)

- The country in which the fund has its headquarter dictate the law the fund will abide to. The fund is then free to market in all EU state given the provision of legal notification to the home regulator of the UCITS fund.

- The fund must provide documents for the investor to make appropriate and transparent investment decision. The documents usually take the form of an Offering Documentation (normally an Offering Memorandum/Prospectus ("OM") and Offering Supplement ("OS") in case of sub-funds) and the Key Investor Information Document ("KIID"). The contents of the Offering Documentation and the KIID is also defined by the UCITS Directive and relevant regulations.

- A pricing notification needs to be made available each time new shares are subscribed or redeemed.

AIFMD

The Alternative Investment Fund Directive creates a supervisory framework for Alternative Investment Fund Managers (AIFMs) and a marketplace at European level. The AIFMD defines an Alternative Investment Fund ("AIF") as any kind of collective investment scheme which does not fall within the definition of a UCITS fund. An AIF may be solely managed by a duly authorised AIFM or else be self-managed. A self-managed fund is considered to be an AIFM and therefore has to apply all the provisions of the AIFMD relevant for AIFMs.

Like for UCITS Directive, the AIFMD allows AIFs to be marketed across the EU, through a simple notification to the home regulator of the AIFM.

The most important requirements to undergo the above directive are listed as follow:

- The directors must nominate an independent depositary to manage the funds, usually a bank or an investment company.

- The fund must demonstrate certain level of risk management politics. The Risk Management Policy of an AIFM in relation to each AIF under management has to cover at least the following risks: Market, Liquidity, Counterparty, Credit and Operations.

- The fund must respect some transparency requirements which are specifically set out in Article 23 of the AIFMD, and include information about the investment

strategy, objectives and restrictions, leverage limits, service providers, fees, information about the pricing and risks involved.

In view that the AIFMD is a "service directive", meaning that it is addressed to regulate the AIFM, a number of EU jurisdictions have introduced types of AIFs which do not require the prior approval of the regulator or else, if they do, the regulator issues the licence in a very speedy way.

Hereunder is an explanation of the widely used AIF regimes in Malta, Luxembourg, and Ireland.

The Qualifying Investor Alternative Investment Fund ("QIAIF") (Ireland)

A QIAIF is an AIF authorized by the Central Bank of Ireland, which may be marketed freely to Qualifying Investors across the EU and the EEA states by an authorised AIFM. The QIAIF can avail of the Central Bank of Ireland's fast track, 24-hour approval process, subject to a number of conditions and confirmations submitted to the Regulator. QIAIFs are not subject to many investments, borrowing or leverage restrictions. QIAIFs are commonly set-up as ICAVs, Irish Collective Asset Management Vehicles. This is a corporate fund with its own separate legal personality which issues shares and has a Board of Directors.

The Reserved Alternative Investment Fund Regime ("RAIF") (Luxembourg)

Luxembourg has implemented the RAIF regime, which allows for a quicker time to market for the launch of the fund. A RAIF may invest in all types of assets and is subject to few investment restrictions, mainly related to diversification rules (30% of maximum allocation into one single asset). The RAIF is not subject to the Luxembourgish regulator's approval or authorization and, for this reason, the AIFM (AQA Capital) assumes full responsibility for the scheme and the fulfilment of its obligations.

The Notified Alternative Investment Fund ("NAIF") (Malta)

Malta has implemented the NAIF regime, which allows for a quicker time to market for the launch of AIFs. Once the full notification pack is submitted to the MFSA, the NAIF

may commence its business within ten [10] working days. NAIF main characteristic is that the AIFM assumes full responsibility for the fund and the fulfilment of its obligations.

2.3 The path for AI deployment

Asset management industry is at a critical juncture

In the last 50 years the Asset Management industry has undergone some radical and fundamental changes. All these changes contributed to shape the industry, bringing it to a completely different configuration. It is safe to say, then, that this process is likely to continue in the future.

From the 70s we can identify some major trends that are a direct consequence of all these reshaping forces weighting on the shoulders of the Asset Management industry:

Widening of the products offering. From investment portfolios mainly composed by stocks and bonds with little or no attention towards diversification and asset allocation, to the first Index Funds for Institutional Investors, to the state of art, we are witnessing a trend of growing complexity, broadening and evolution of the financial market. As it becomes more complex, so are the financial products offered. A meaningful care for risk management and an increasing demand for tailored and customized portfolio also bring consequences on the variety of the products asset managers are offering.

Tapering of the profit margins. With interest rates floating under the 0% line is becoming more and more difficult for Asset Management companies to build a sustainable margin. This has the obvious consequence to search for ways to cut costs whenever is possible. In this context AI solutions come very in handy as the companies can now automate certain actions that in the past required huge spending in personnel and formation, increasing the efficiency at the same time.

Change in the role of advisors and counsellors. As AI solutions take place covering some aspects of the standard role of advisors, their job is consequently changing. Repetitive and alienating tasks are being held by machines while the human workers have

more time to concentrate on more meaningful and creative projects. Even if more and more young people are taking the "do it yourself" approach to tackle their financial needs, there is still need for human to empathize, help, and nurture relationships.

Disruptive technology breakthroughs. Starting from the important phenomenon of automated stock trading in the 70s, new discoveries and tech-driven disruption the industry is now leaning towards automated investing platform, robo-advisors, and chatbot. The majority of these technologies are empowered by Artificial Intelligence and Machine Learning in particular.

These days is more and more difficult to identify what an Asset Management firm really is. Business models, infrastructure, models are all changing, providing proofs that the industry itself is evolving from a structural perspective in a self-rediscovery fashion.

"The reality is that the asset and investment management world is at the cusp of a major transformation. This transformation is not an ordinary evolution in the normal course of business. It is a revolutionary change that is creating never-seen-before opportunities and threats. It has unleashed an enormous force that is demanding new ways to respond to the challenge."[32]

To witness such a revolution can be quite overwhelming but is important to have clear in mind what are the drivers for this revolution, what direction the whole industry is taking, and to analyse what new-born and existing company can do to keep up with the extraordinary pace of these transformations. These will be, in fact, the topics that this chapter is set to explain.

AI success drivers

All the radical trends that were discussed in the previous chapter are emerging faster. They all point towards one simple conclusion. A conclusion that is similar to what is happening in more and more industries. In order to survive is vital to keep up with the new technologies. But when the technologies are so important and disruptive like AI, this issue becomes even more paramount. Investment management firms are slowly understanding this concept. This is also the first step towards an optimal outcome.

[32] Naqvi, Al. 2020. Artificial Intelligence for Asset Management and Investment. Wiley.

Acknowledging that AI is going to disrupt the whole industry and those who do not accept this concept will eventually lag behind.

"The old model of asset management is not only non-competitive, but also counterproductive. It fails to serve the best interests of clients, and it hurts the profitability of the firm. It fails to offer the quality of investment expertise that today's clients expect and deserve. It ignores important elements of human emotions, qualitative data, behaviours, narratives, and other human processes."[33].

The reason for the afore mentioned non competitiveness relies in a series of challenges that asset managers are facing. These challenges are even more exacerbated by market pressures and competitive forces from within the industry and from external factors. This will be a massive weight on operation leaders inside these firms, as their responsibilities will increase in exploiting the right opportunities and the abandoning obsolete ones. Given this watershed year (2020), they will need to rejuvenate the focus on growth, invest in innovation and reinforce products and processes.

Even if these strategic goals are shared and agreed between asset managers, research projects show that "42 percent of operations executives believe their operations and technology are not configured to adequately execute the firm's overall strategy. However, as we will see, they are not at all complacent about this situation, and are moving forward to address it."[34]

The last sentence makes the intention of reshaping the industry really clear. The rationale behind this statement can be summarized in an effective quote attribute to Abraham Lincoln himself, stating "The best way to prepare for the future is to create it".

In the next chapters I will define and analyse the challenges I just mentioned. The basement for the analysis is provided by some major studies and reports carried out by Accenture Consulting and BlackRock on the future of Asset management as well as personally conducted analysis. Moreover, I will try to highlight the best path to overcome these challenges. The results will define the use and embrace of Artificial Intelligence technologies and culture as a critical point to reach a dominant position in the industry, or at least avoid lagging behind the competition.

[33] Naqvi, Al. 2020. Artificial Intelligence for Asset Management and Investment. Wiley.

[34] Accenture Consulting; Investment Company Institute (ICI). 2019. «Reinventing Operations in Asset Management.»

Revenues

Accenture Consulting recently published a study[35] conducted on ICI[36] members. They asked to 33 asset managers representing about 15 trillion of AUM[37] what they think will be in the future the major driver for success in the asset management industry. The number one answer, new revenues opportunities, reflects the undergoing process of radical changes in the industry. When 74% of the Asset managers think that this will be the major driver for success, it means that there is a sense that the asset management industry is reaching a saturation point. Profit margins becoming more and more slender is a key hint to reach and exploit the countless opportunities of innovation in a wide sense.

[35] Accenture Consulting; Investment Company Institute (ICI). 2019. «Reinventing Operations in Asset Management.»
[36] Investment Company Institute. It is an association representing regulated funds all over the world (ETFs, Closed-end funds, UITs).
[37] Asset Under Management.

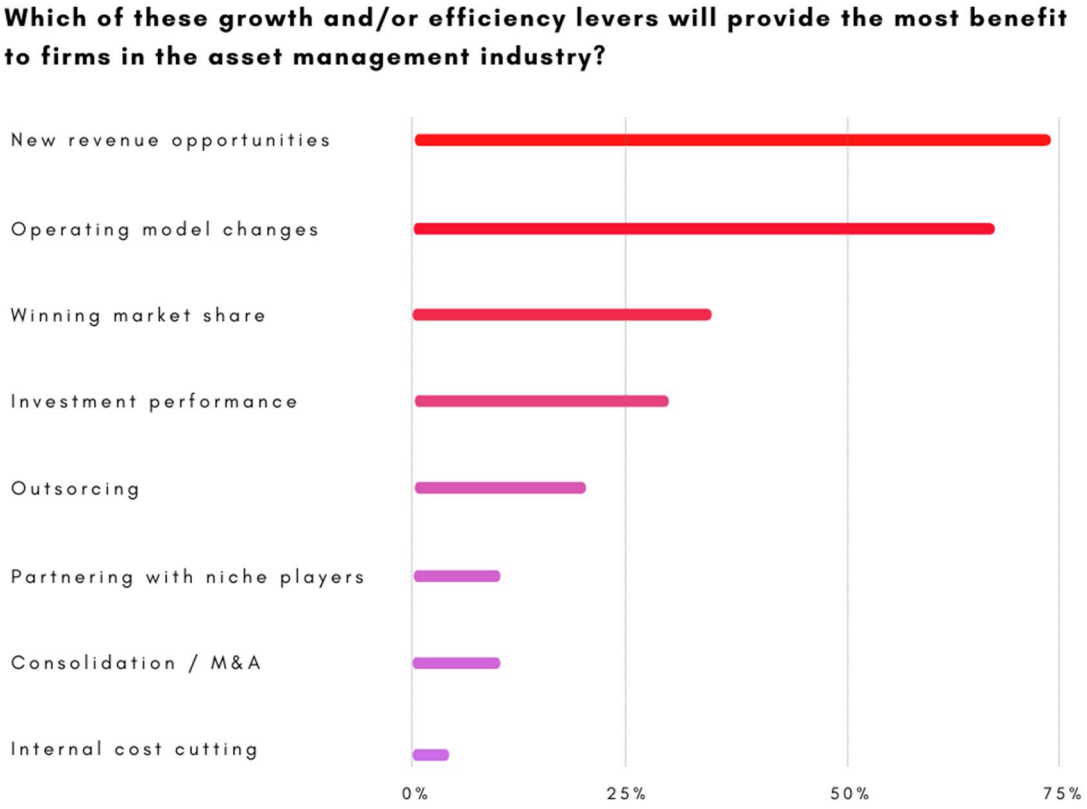


Figure 17 - Source Accenture and ICI operation study[38].

The wisest question to ask under these circumstances is: where does these new revenue opportunities are more likely to come from? With no doubt the answer falls under the spectrum of new products and distribution channels. These are more likely to generate new revenues stream. As explained in the previous chapter there are plenty of ways to exploit AI to create new products. This technology can be used to figure out future trends in asset price movements as well as efficiently uncover hidden patterns that can bring to new ways of approaching the selection of assets for a specific portfolio. Apart from the contribution that AI can give to stock picking activities, there are more and more AI based ETFs being created. As the data shows the AI sector has overperformed the market in the past 2 years. If we consider the XLK[39] benchmark which is the technology sector index

[38] Accenture. 2020. «Reinventing Operations in Asset Management.»

[39] Technology Select Sector SPDR ETF (XLK)

(as AI sector does not have its own), we have a 58% total return in a period of 12 month against a 50% of the S&P 500 as of April 30, 2021[40].

New revenues stream will not only flow from new products, the other path that asset managers are following is beyond their core competencies. Innovative and successful proprietary software (and this includes AI based software) and client platforms, can be licensed for the purpose of gaining excessive and alternative returns.

It is important to keep in mind that non-traditional competitors are knocking at the door of the asset management industry. As they were born without the burden of traditional strategies baggage, they are already comfortable in exploiting this kind of alternative revenues streams and this create a key threat for the existing firms.

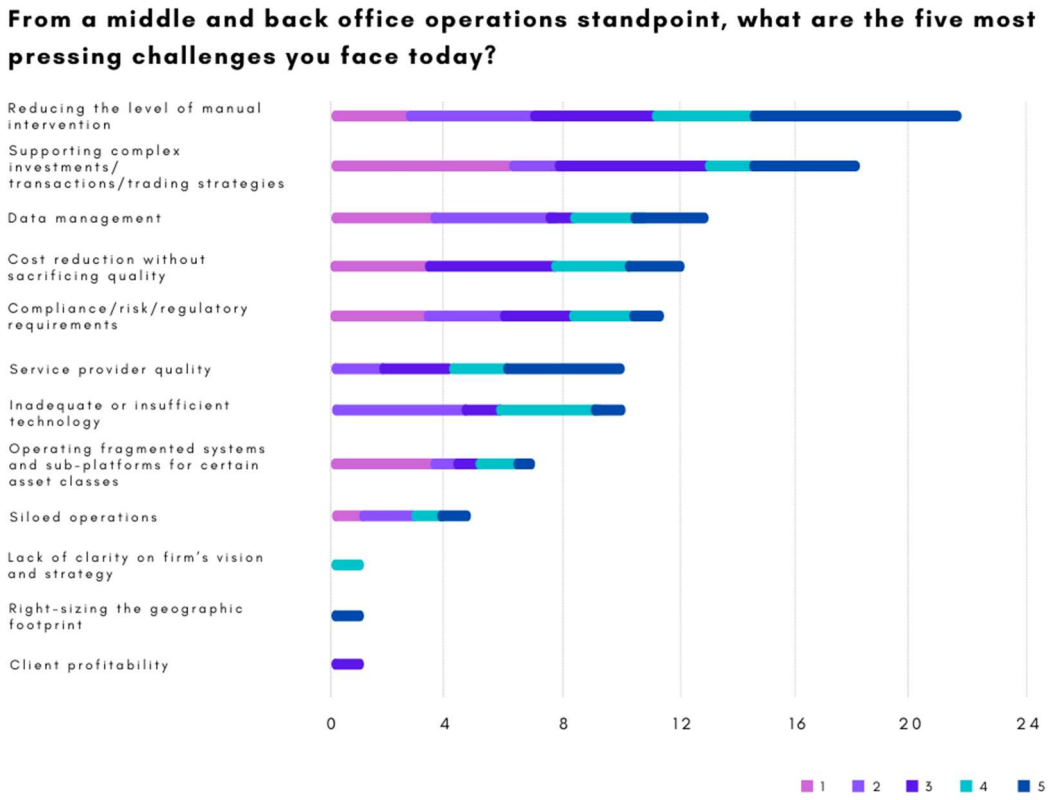
Operating model changes

Back and middle office operations play a vital role in the daily activities of an asset management company. They can be considered, in fact, the backbone of the company itself. In operations most of the daily tasks are carried out. In fund management for example the operations tasks are valuation and pricing of the fund, but also taking into account all the transactions, solving mispricing issues, or fixing portfolio management systems mistakes. Without operations the company cannot work. Given the variety and the importance of this function, it is no surprise that it contains the most dense, long, and complicated processes. That's why in a context of margin tapering, asset management companies are trying to support or change the operating model. A smooth and flexible operating model can play a decisive role in supporting the services and approaching the clients.

The major factor that will act as a catalyst in supporting the way towards new operating model is of course technology. Technology and innovation are considered to be the bridge that connects growth with efficiency. In most of the industries the capacity to discover and adopt the right technology has become vital, not only to compete but to survive. In the survey we can see that 55% of the interviewed company have a public initiative to embrace new tech tools or fintech solutions.

[40] Source: YCharts. "Financial Data." Accessed May 3, 2021.

The process of democratization of new technologies and the creation of new and new fintech companies has created the perfect background. Middle office functions such as, data and collateral management or derivative processing will have the more benefits from this background. Back office has been impacted as well, mostly in fund accounting, financial reporting, and expense management.



41

Figure 18 - Challenges for back-office operations. The number of times one of the challenges has been chosen is represented on the x axis, while the order by which they have been chosen is represented by the colour.

In the graph we can see that the industry is rotating around common themes. The most citated one is about the reduction of level of manual intervention, in other words automation. Investment management company are more and more trying to find ways to cut the unnecessary work and reduce human intervention. The first reason is purely for cost reduction purposes. Employee salary is one of the greatest expenses in financial company balance sheets and technology has the ability to decrease them notably. The

[41] Accenture Consulting; Investment Company Institute (ICI). 2019. «Reinventing Operations in Asset Management.»

other reason is that reducing human errors in back-office operations can be immensely helpful. Finally reducing the time on repetitive and alienating tasks frees up time for people to concentrate on the things that matter or that requires different skills. This has the result boost up morale and increase productivity at the same time.

Supporting complex investments/transactions/trading strategies can be quite hard for middle and back office. Usually, companies have based their technology on the ecosystem of strategies and financial instruments present at the time they were built. The rising in complexity in financial markets has reflected in the strategies portfolio managers use to manage their investments. For example, a system created to take account and evaluate bonds and equities will have a hard life trying to do the same with financial derivative instruments. That's why the majority of the companies in the industry have initiatives underway to replace or enhance the level of technology.

Another common challenge the whole industry is facing is related to data. Financial firms find themselves with tons of data, often unstructured without an actual plan to extract value from it. The challenge for the firm is represented by data quality and accessibility. High quality data doesn't refer only to the extent the information is valuable, but also the degree to which the data has been cleaned and stored in a correct way. This is fundamental to enable analytics and automation. It also creates the possibility of a more effective reporting process for consumers, clients or even directors. Uncleaned and unprocessed data is not only dangerous for the entire company by creating faulty analysis and reputational risk, but it also gives rise to opportunity cost as the world is more and more demanding towards good data management. Firms should stop assuming disruption in data management as something given and unchangeable. They should instead find ways to monetize in a systematic way the large, yet unmined data sets they have available with the constraint of the technology they have.

Possible solutions

It is clear that a process of reshaping of the middle back office is needed now more than ever. As the whole industry pushed on the development and application of new technologies the most promising ones are somehow linked together. AI and RPA[42].

[42] Robotic Process Automation.

Even if they are at different point in the adoption curve, these technologies showed to have great potential in helping investment management firms to reach their goals. AI has already deeply affected the industry. A lot of progress has been made in the use of AI in the front office, especially with the advance in NLP technologies. But that is just one of the numerous areas it is being used. Nowadays, there are funds which are being managed through AI, others have algorithms that can forecast geopolitical events and base some investment decision on the outcome.

AI in back and middle office hasn't still reached a use at scale, but larger firms have completed an AI-based reshaping process, or they have one underway. Most probably AI technologies will deliver the next wave of cost reduction throughout the industry.

RPA is considered the "getaway drug" to unlock the full potential of AI. It is able to substitute human work with a virtual workforce in order to execute redundant processes. As the research carried out by Accenture outlines:

"52 percent of respondent firms said they have incorporated RPA into their operations, with reconciliations, data management, transaction management, report generation and cash forecasting named as the top functions/tasks to which RPA has been applied. What's more, an impressive 82 percent of respondents indicate that RPA has delivered the expected results, which include increasing accuracy in some functions and freeing up employees to do more meaningful work[43]."

Finally, another viable solution is outsourcing of middle, back-office operations. This solution is faster, simpler, and effective in the short term, while it loses significance in the long run where an in-house solution can bring more to the table. It has also the advantage of being ideal for small-medium companies that do not have the necessary capabilities to build such technologies. Most of the firms in fact outsource their back-office operations while they keep the middle operations inside. "While 91 percent of respondents are satisfied that the strategic goals that led them to the outsourcing decision were achieved, 48 percent of respondents have considered bringing a function back in-house—including collateral management, derivatives processing and tax services—with service provider quality cited as the primary reason. Meanwhile, firms continue to expand

[43] Reinventing Operations in Asset Management – Accenture Conuslting, Investment Company Institute (ICI).

the scope of outsourcing using non-bank partners such as technology firms and professional services firms."[44]

The risks of AI deployment

As the use of AI in finance increase in popularity, risk and challenges associated with this process arise as well. The need for quick and meaningful action by the policy makers has become of primary importance in an ever-evolving context. The aim of tis chapter is to analyse some of these challenges and examinate some tools to mitigate risk.

Data management

Data is the centre of any AI algorithm or process. However, the quality and the appropriate choice of data are fundamental as they can introduce some degree of non-financial risk. These risks are mainly related to privacy, confidentiality, and concentration.

Quality of data is at the base of a correct functioning of AI programs, but the quality itself can be put in discussion as the size of the datasets increase. A great number of studies[45] shows how the level of truthfulness of big data is defined by a certain amount of uncertainty. The 3 most important characteristics are:

1. Exhaustivity.
2. Extensionality.
3. Veracity.

The first one refers to how wide the scope is, the second one refers to the capacity to add, remove or change fields, and the third and most important one refers to the degree to which the data is complete and can be trusted. Other prerequisite to avoid risk in the data collection is an adequate process of data revising through labelling and structuring. Ideally the ML programs need to identify signals distinguishing them from noise with the final aim to recognize patterns in the data.

[44] Reinventing Operations in Asset Management – Accenture Consulting, Investment Company Institute (ICI).
[45] IBM Big Data & Analytics Hub. 2020. The Four V's of Big Data. IBM.

An effective solution to the problem of noise in the data is through the creation of artificial dataset (also called synthetic datasets) in ML modelling[46]. They allow for the improvement of feature like scale and diversity in a dataset that lack of them. Another important feature of synthetic dataset is that they can be anonymous allowing the company to comply with consumer privacy requirements.

Privacy is in fact one of the major factors of risk for ML models as they have a high capacity of making inferences. The universe of data points that are considered sensitive can and will be expanded by the creation of more sophisticated algorithms as they can be proficient in the user identification.

Other concerns need to be raised as the data used in the financial sector as the data needs to be aggregated, transmitted (often across borders), stored, processed. The importance of a good plan for data governance is becoming more and more paramount.

Discrimination and algorithmic bias

The correct utilization of data is a recurrent problem in almost every aspect connected to AI risks. In fact, a biased dataset can lead to discrimination in financial services. It is important to bear in mind that AI has the potential to reduce discrimination and unfair treatment, but only if biases in data are reduced to the minimum.

The reason why AI algorithms can bring to increase discrimination is to be found in their capacity to compound existing bias in the data. If the ML model is trained with a dataset that contains biases it will perpetuate it. This outcome can be also not intentional, and they can derive also from good-quality data. This is the reason why companies need to implement into their data management/governance so, auditing and testing of models. These mechanisms can sense check the result of the model against some baseline data. Ideally, whoever supervise the algorithm has to have the capacity to score the model in his accuracy and fairness. This is where human intervention becomes absolutely important, both in the phase of preparation of data and in the phase of explanation od the model result as well as the phase of correction and design of the model.

[46] S&P. 2019. "Avoiding Garbage in Machine Learning." https://www.spglobal.com/en/research-insights/articles/avoiding-garbage-in-machine-learning-shell.

The explainability puzzle

Another frisk factor for AI or ML empowered models is the afore mentioned explainability puzzle (or conundrum). This can be explained as the difficulty and sometime the impossibility of describing how a model arrived at a certain result. Or in other words the difficulty in decomposing the result in its underlying drivers. The explainability conundrum represent one of the most pressing challenges in the AI community. It is also exacerbated by the fact that sometimes companies intentionally hide the mechanism behind their models for intellectual property protection purposes.

In this context a clear problem in terms of policy compliance gain importance. As it is mentioned in this OECD paper: "In the most advanced AI techniques, even if the underlying mathematical principles of such models can be explained, they still lack 'explicit declarative knowledge' (Holzinger, 2018[38]). This makes them incompatible with existing regulation that may require algorithms to be fully understood and explainable throughout their lifecycle (IOSCO, 2020[39]). Similarly, the lack of explainability is incompatible with regulations granting citizens a 'right to explanation' for decisions made by algorithms and information on the logic involved, such as the EU's General Data Protection Regulation (GDPR)13 applied in credit decisions or insurance pricing, for instance. Another example is the potential use of ML in the calculation of regulatory requirements (e.g. risk-weighted assets (RWA) for credit risk), where the existing rules require that the model be explainable or at least subject to human oversight and judgement (e.g. Basel Framework for Calculation of RWA for credit risk – Use of models 36.33).[47]"

In the absence of interpretation of AI and ML algorithms the problems are not only at individual companies' level but also at a macro-level for the market itself. It is becoming more and more difficult for companies and policymakers to forecast the effect of AI-based decision on the markets leading to increase in possibility of market shocks. It is also quite frightening that if you don't know the drivers underlying the model you are also unable to adjust it or stop it in conditions not expected by the algorithm. This comprises periods of illiquidity, aggravated flash crashes, negative prices, and far more.

[47] OECD. 2021. OECD Business and Finance Outlook 2021: AI in Business and Finance. Paris: OECD Publishing. https://doi.org/10.1787/ba682899-en.

Training and testing of AI models

The performance of a great AI model is also based on its capacity to capture tail events in the data or non-linear relationship. However, it is also one of the greatest challenges during the process of training models, as tail events are rare and scarce in most of datasets and they are also unique events with connotations hard to foresee. This challenge is undoubtably creating weak points in the financial system whenever periods of unprecedented crisis arise. One viable solution would be to train models with a lot of different datasets containing tail events such as global financial crises, natural disasters or, more recently, Covid-19 crisis. The result of this will be models with increased accuracy and optimal parameters reducing the risk of overfitting[48].

The second optimal solution would be an excellent process of validation. The validation is carried out with samples unknown to the model. As a result of training on the validation set the optimal parameter will be chosen as the ones that performed better (minimum validation error), and the model accuracy can be quantitatively determined. Again, in this case synthetic datasets represents an effective and cheap way to validate a model by creatin simulated data.

AI models performance during Covid 19 pandemic

ML models whenever idiosyncratic one-time events are about to happens are considered to be a double-edged sword. The most cutting-edge algorithms, but mostly the one created for this purpose have the potential to predict tail events by uncovering new patterns in the data, often invisible to the human eye. Other instead have proven to suffer in these circumstances. A survey conducted on UK banks showed evidence that nearly 35% of banks had negative results on ML models during the virus outbreak[49]. The reason for this poor performance has to be found in the effects of the pandemic on real economy. The unprecedented rise in unemployment and the temporary stop of companies' activities brought the models in unseen territories. Whenever a model is created to predict the behaviour of a variable and that variable shows unexpected changes, the predictive capability of the model loses effectiveness resulting in performance degradations. That's

[48] The risk of over fitting is whenever a model performs well on the training samples, but poorly on new and unknown samples.

[49] Bholat, M. Gharbawi, and O. Thew. 2020. The impact of Covid on machine learning and data science in UK banking. Bank of England, Bank of England Quarterly Bulletin Q4.

why it is important to stress the value of validation with synthetic and normal data that include extreme scenarios.

Chapter 3

AQA capital case study

"I made my first investment at age eleven. I was wasting my life until then."

Warren Buffet

3.1 AQA Capital presentation

AQA Capital is an Asset Management Company authorised by the Malta Financial Services Authority ("MFSA") as a full scope Alternative Investment Fund Manager in terms of the Directive 2011/61/EU and a UCITS manager in terms of the Directive 2009/65/EC as amended. AQA Capital holds a Category 2 investment services licence issued by the MFSA in terms of the Investment Services Act (chapter 370 of the laws of Malta). AQA Capital can act as investment manager for both AIF and UCITS Funds and promotes these funds throughout the EEA. AQA Capital is also authorized to provide discretionary portfolio and investment advisory service to retail and professional clients.

AQA operates fully authorized branch in Milan (subject to CONSOB supervision) and representative offices in Prague, and London with a team of 29 specialists spread across three offices. AQA operates as well in Luxembourg and shortly also in Ireland.

AQA is an operationally ready and economically viable solution that aims to cover the entire value chain of asset management industry, providing its clients tailored made services, including:

- **Financial engineering solutions.** Offering bespoke solutions to support the structuring of dedicated investment funds, vehicles and single financial operations

- **Fund management services.** acting as Management Company in relation to the launch and the ongoing management of Collective Investment Schemes, irrespective of their nature (UCITS or Alternative Investment Funds), domicile (i.e. Luxembourg, Malta, Italy, Ireland or any other EU or non-EU jurisdiction) and underlying assets (listed equity, private equity, bonds, FDIs, real estate, etc.).

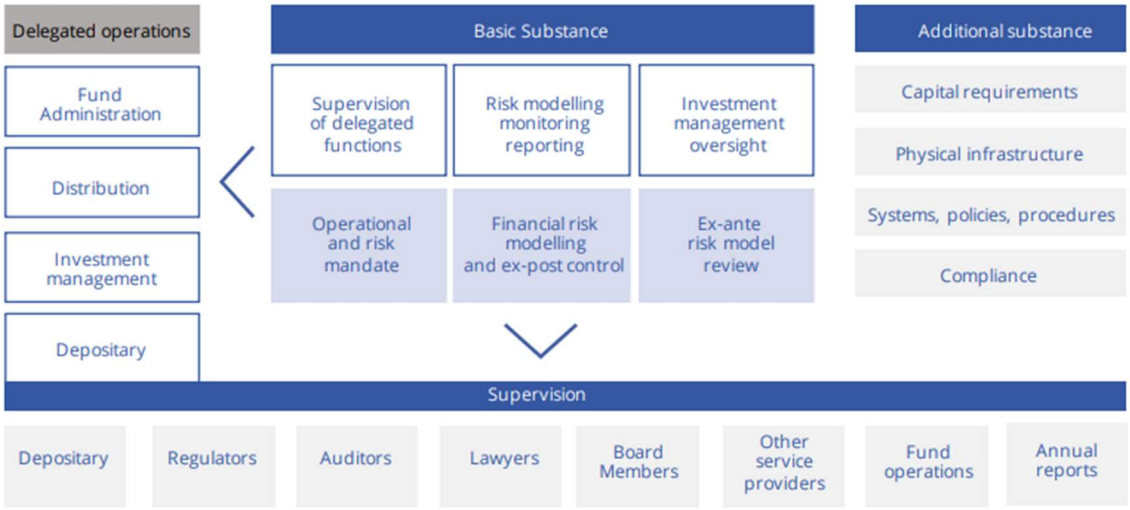


Figure 19 - AQA Capital services offer

Fund management services

A Management Company ("ManCo") has the responsibility to:

- Provide the core services of portfolio management and/or risk management.
- Supervise the delegated functions on an ongoing basis and ensure that the delegate has the necessary resources to undertake such functions adequately.
- Coordinate and liaise with all the service providers and ensure compliance with applicable laws and regulations.
- Regularly report to the Regulator.

This will apply to any EU or offshore jurisdictions, either with AQA brand or in white label for third parties, irrespectively of asset class and risk profile.

Through an extensive network, AQA provides legal and operational support for the distribution and placement of collective investment schemes both in and outside EU, in compliance with each targeted jurisdiction

AQA also operate on the entire Value Chain so that the clients can choose the segment of activities most suitable for them.

Figure 20 - AQA Capital Value Chain

AQA Capital 2.0

From the start of 2021 AQA Capital is undergoing a digitalisation process to enhance its technological footprint. One important goal of this venture is to decrease the level of human intervention in its day-to-day operations by automating repetitive and alienating tasks either the consequence of using staff work for more meaningful and complex tasks.

Another important ambition for AQA is to increase its technological reputation among stakeholders, showing its willingness to embrace innovation in an active and passionate way.

Whoever worked inside the company was solicited to suggest some ideas that could contribute to reach the goal of this digitalisation process. After a round of presentations in May 2021, the ML learning stock price prediction model that I suggested was chosen as the best initiative to carry out. The model together with the strategical framework to apply it inside the company are shown in the next chapter.

3.2 Designing a modern investment company

Goals definition

Before starting the long and complex process of designing a modern investment firm the idea is to set some ambitious yet reachable goals in order to always keep in mind what we are aiming to. The goals need to be created by considering the values and medium-long term goals of the company that will use them. AQA is a company that invested a vast number of resources into improving their most valuable assets, people. Not only people that work inside the company, but a far greater group of people that can be recognized in its stakeholders. Moreover, AQA valorise people as individuals and as part of a group. It also valorises connections and synergies that emerge the group itself.

The following goals perfectly embed these values:

- **Structural coherence**. A firm can be seen as a collection of capabilities. No single capability determines the success, but the management and correct use of all of them is the key.
- **Interdependence**. Each capability can interact with the other creating added value by giving importance to the interdependence.
- **External Interaction**. These capabilities need to process both internal and external information. However, particular importance needs to be given to external flows.
- **Performance Maximisation**. The Capabilities need to have performance as ultimate goal with the constraint of maximising the interaction of all of them instead of favouring a single one.
- **Cohesive value building**. The design of the capabilities is focused on three aspects: automation, intelligence, and data. The first one is when machine replace human work and its performance is measured through efficiency. The second one is referred to the ability of the machine to navigate in uncertainty with success.

A company that manages to reach all these goals will reach excellence in becoming a modern investment company. This will be true as all functional area of the company will be interconnected and will function as a network.

Intelligence, Automation, and Data

An Intelligence-centric competitive advantage

The ultimate goal of such a process is to reach a sustainable and effective competitive advantage. In order to do so, it must be built on some stable factors. A characteristic that the right technology can confer to the building of competitive advantage is intelligence. Intelligence can be defined as "being able to successfully perform work by displaying goal-directed behaviour in situations of uncertainty. When machines display intelligence, they perform work and resolve uncertainty in accordance with goal-directed behaviour. It can also be viewed as an attribute of an artifact by which it accomplishes work by successfully tackling uncertain situations in accordance with its goals."[50]

In order for a competitive advantage to be considered intelligent, it has to be built on models and machines that can understand the environment from an internal and external perspective. Nonetheless, they will also have to handle competitive, social, and cultural dynamics in a context of opportunities and threats. Machine learning models are perfect to fulfil these tasks as they can be viewed as learning systems that can evolve, mature, and adapt to the changes in the environment.

The above-mentioned factors declined in an investment management company environment are defined as follow:

- **Humans and machines**. Only by exploiting weaknesses and strengths of both human and machines the competitive advantage can be stable. Moreover, it is essential to define them clearly and to take advantage of synergies that are inherent inside this connection.
- **Products**. Machine learning models will enable the possibility to discover new strategies, eliminating human biases and reduce costs and risk.
- **Services**. Learning machine will be useful to help compliance with regulatory requirements and help risk managers to find signals or red flags on assets managed by the funds.

[50] Naqvi, Al. 2020. Artificial Intelligence for Asset Management and Investment. Wiley.

A firm designed on these actors is certain to build a sustainable and intelligent competitive advantage. However, intelligence is only one side of AI. The other side is represented by automation.

Automation

One of the key challenges for AQA Capital is to increase automation of repetitive processes in order to divert the potential of people to more interesting and valuable tasks. In order to address the challenge of automation is mandatory to understand and analyse its definition.

Automation can be defined as the process by which the human input is minimized. There are different types of automation. The ones that will be addressed by this model will be mostly three as defined by IBM:

"Process automation. Process automation manages business processes for uniformity and transparency. It is typically handled by dedicated software and business apps. Using process automation can increase productivity and efficiency within your business. It can also deliver new insights into business challenges and suggest solutions. Process mining and workflow automation are types of process automation.

Integration automation. It is where machines can mimic human tasks and repeat the actions once humans define the machine rules. One example is the "digital worker." In recent years, people have defined digital workers as software robots that are trained to work with humans to perform specific tasks. They have a specific set of skills, and they can be "hired" to work on teams.

Artificial intelligence (AI) automation. The most complex level of automation is artificial intelligence (AI) automation. The addition of AI means that machines can "learn" and make decisions based on past situations they have encountered and analysed. For example, in customer service, virtual assistants powered can reduce costs while empowering both customers and human agents, creating an optimal customer service experience."

Obviously, the automation requires intelligence to be really effective, as the model can both automate a repetitive task and navigate through uncertainty granting an adequate response.

The Machine Learning model that will be presented in the next chapters is a simple stock price prediction program based on Artificial intelligence automation. It will be responsible of automating the task of analysing tons of data tables and all the processes that concern this task with a simple and fast model that can learn how to do it by itself. It is a task that AQA capital is decided to address with intelligent automation as it takes a lot of resources in terms of energies, money, and time.

Data

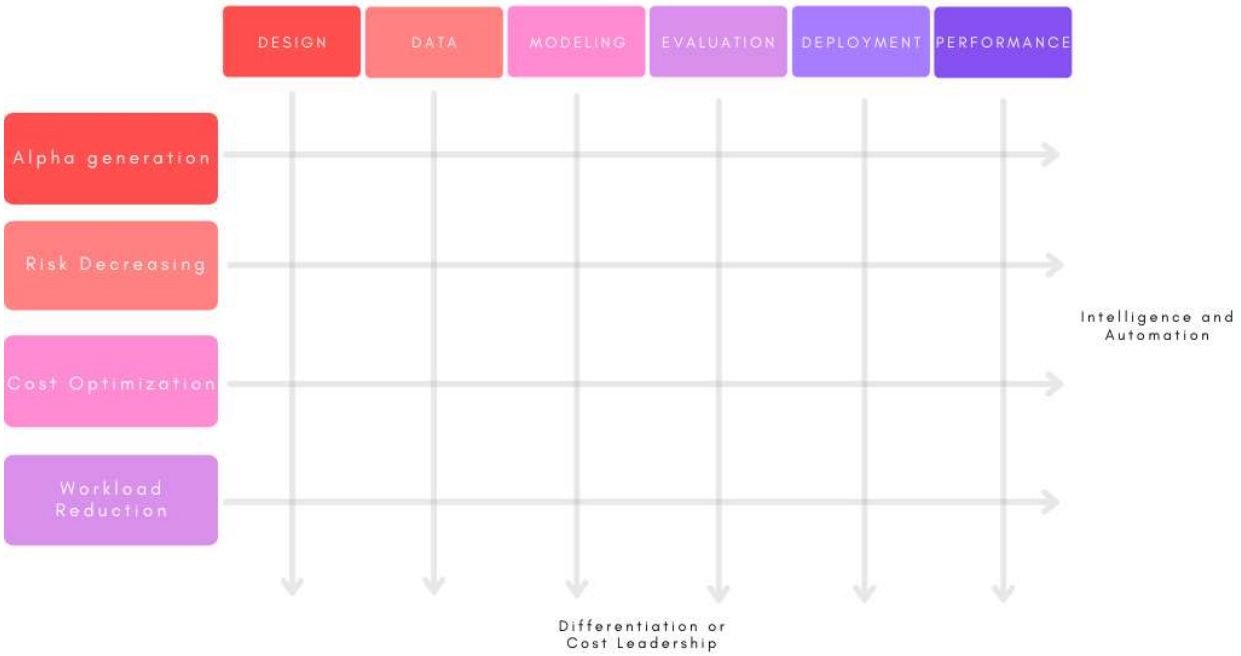
Finally, the factor on which a sustainable AI-based competitive advantage is built is data. In any respectful AI or ML model data is centric. Not only it needs a vast amount of data to function properly, but the quality is also important. Having superior quality data will ensure optimal performance for the model. Moreover, the right type of data must be chosen as the one that is closer and more useful to the goal that was set in the previous chapters. In other words, data is gathered thinking about completeness, timeliness, and relevance.

The strategic framework

After having defined the goals of the process it is necessary to build a reusable model that can be employed not only by AQA, but any other company that shares its value of integration and valorisation of human capital and that want to increase the level of technological deployment.

By taking the above value drivers as underlying assumption is it possible to create a model that will manage the strategic transition from a chaotic technology management to a modern AI approach. The assumption above considers that the companies are collection of capabilities, and those capabilities needs to be extended above the singular functional area and across the whole company.

As AQA is still at an early stage in its life and its structure still highlights a functional level division it is even more paramount to define value drivers that can be shared across the functions.



On a vertical level we have the normal value chain of a firm. The value drivers have different underlying objectives.

- **Cost optimization**. Reducing the costs across the whole company.
- **Workload reduction.** Decreasing the hours spent on repetitive tasks and freeing up time to do more important works.
- **Risk Decreasing**. Reducing human error in order to decrease the number of mistakes on a given task and consequently reducing risk.
- **Alpha Generation**. Alpha is defined as returns greater than those of the market and holding back risk at the same time.

These value drivers are defined as capabilities area and one or more capability can influence one or more functional area of the organization.

The horizontal level is based on the scientific method. In this particular case it is declined to reflect the life cycle of ML and AI models. The core processes are:

- Design
- Data
- Modelling
- Evaluation
- Deployment
- Performance

As explained in the previous chapters these areas are competency of Machine Learning and are independent of the capability areas on the vertical axis. These competencies are active by the development of Machine Learning models and solutions.

With this model every ML solution can be created, developed, deployed, nurtured, and eliminated at the end of its work.

The performance of the models is compared to human performance in terms of efficiency and effectiveness and analysed by keeping in mind the goal of the company and the external environment.

Design

Design is the first step towards the optimal AI evolution, and it is aimed to build alternative processes or replace old ones. It is defined as "the process of architecting a firm's AI plan that supports the firm's strategy[51]."

Design usually starts from the partner of a firm that recognize a need or want to fix a certain problem. Obviously not every problem has an optimal AI solution, or better not every problem has the necessity to be solved through an AI model. Usually, a problem that naturally lend itself to AI solution it's a problem that can be breakdown in multiple steps and has a high degree of repetitive processes.

These problems need to be analysed in terms of automation and intelligence. How much the problem requires an automation solution? And how much degree of intelligence is needed? By answering this simple yet challenging questions it is possible to specify the objectives for automation and intelligence and exclude some type of AI solution that are not optimal.

In the AQA capital case the problem was to increase the technological awareness of the entire company by building some AI artifacts thar reflect the company's values and that can generate a certain degree of competitive advantage. From a marketing standpoint this will also enhance and improve how the image of AQA capital is seen among stakeholders. From and operational perspective these artifacts need to automate repetitive task and free up part of the employee's time.

Stock price prediction is a process that ML algorithms are very keen to solve in an elegant and efficient way. It is also a process that requires a huge quantity of time spent in data gathering, parameters tuning and problem fixing. Nonetheless by applying an AI framework to the problem it is possible to generate much more accurate and fast prediction. On one side this will increase the gap towards competitors still attached to legacy technologies, and on the other will create an interesting gain in reputation towards clients and service providers.

While the technical aspects of the solution to such a problem will be addressed in the next chapters, let's move on along the path of the creation of AI artifacts.

Data

[51] Naqvi, Al. 2020. *Artificial Intelligence for Asset Management and Investment.* Wiley.

Usually data is the most challenging part of the life of an AI model. As previously addressed in the previous chapters it needs to be of good quality and in the right quantity to avoid both underfitting and overfitting.

What is really missing in a vast number of investment management company is a good and wealth data management program. These programs were critical in the past and with the advent of AI are even more imperative today.

"From a strategic perspective a good data management program needs to:

1. Help managing data for the entire enterprise.

2. Provide data for all the data science and AI efforts.

3. Operationalizes data collection and sensing.

4. Implement and support data pre-processing for AI.

6. Understands the current and future needs for data.

7. Performs the traditional data management functions such as data governance, data quality, metadata management, and master data management."[52]

As data availability has reached an impressive status by increasing in quantity, velocity, and granularity, it is more and more challenging to keep the pace with its advance. Most of financial companies nowadays rely heavily of financial data providers such as Refinitiv or Bloomberg that cannot be consumed systematically by a single individual in an efficient way. That is when APIs come into play. API stand for Application Programming Interface and, which is a set of protocols for integrating and building application software. It also allows the computer code to retrieve, select, and process arbitrary amount of data from any site with API enabled. This is the best way to gather data from financial data providers as it enables the user to access large amount of data in a relatively small amount of time.

Data gathering and processing in the AQA capital stock price prediction program was relatively easy. The main idea is to predict the return of a stock by using information embedded in other stocks, indices and FX currencies. Data was gathered by accessing

[52] Naqvi, Al. 2020. *Artificial Intelligence for Asset Management and Investment.* Wiley.

two important and free financial data providers such as Yahoo Finance and Fred. The composition of this data was mainly formed by weekly and daily return for seven different assets from three different asset classes.

While data quality was easy to verify, it still required cleaning process to adequate the greater number of returns for the FX rates to the lower number of weekly returns of stocks.

Modelling

Modelling is the phase where ML solutions are created. It is the phase in which data is used to extract a pattens with the finale ai to build a model. In an investment management firm AI artifacts fall under three main categories:

Enterprise strategy. In this category there are all the models used to evaluate, improve, and predict the strategic position of the company. This part heavily leans towards intelligence and less on automation.

Investment. In these category AI artifacts are created to give birth, improve, support or replace investment strategies. This is done by converting information inside data into a working algorithm. The strategy can be focused either on single asset classes or on entire portfolios.

Enterprise function. This is a residual category to include everything that is not part of Investment or Enterprise strategy. It can be related to marketing, client management, supply chain, or others.

The model used to predict stock prices is called a learning model. It uses the method of Machine Learning to make a prediction about an unknown part of reality. What prediction really means in the ML ecosystem is simply the formula that estimates an unknown value. In other words, prediction is estimating the output (Y) given the inputs (Xs). It does so by recognizing the relationship between the variables and assigning the optimal parameters to reflect this relationship. As there are many algorithms that can fulfil this task in different ways and with different feature, part of the modelling process is choosing the one that maximises performance while minimizing errors and risk.

Evaluation

Evaluation is all about making sure that the model works. the term “works” can be expressed in different connotations, but mainly it means that the algorithms are accurate, they do not overfit or underfit and they are efficient.

Accuracy means that the model can deliver accurate prediction. This is not a problem that has a static solution, as the accuracy can change over time. Solution to the accuracy problem can be adding or changing features if the one used are sub optimal, making sure that the training data distribution contains cases and conditions in which the algorithms will perform, check if the underlying distribution has changed and make an update if necessary.

Underfitting and overfitting, as explained in the previous chapter, are respectively the situation in which the data on which the algorithm is trained does not contain enough information, and the situation in which the algorithms over-learn from training data and performs badly in the test data.

In a stock price prediction situation, for example, an algorithm that can perform well on training data and poorly on test data or on the actual deployment is not of much use.

Efficiency is a feature related to the capacity of the model to work without using to much computational power. The result of this an excess of computational power used will be that the model will take a long time to run. This feature must be considered not only in the phase in which the program is running but also in the training phase and scaling phase.

The method used to evaluate the model in the stock price prediction model are splitting of dataset in training and test phase (80 – 20 split) and k-fold cross validation. The first method is used to train the model on 80% of the datasets and then test its effectiveness on the remaining 20% to check for possible overfitting and underfitting issues. The second method iterate the first method 10 times randomizing the point in which data is split. Both of them based on mean squared meter errors metrics to evaluate their effectiveness.

Deployment

In the deployment phase models are finally used inside the organization. Not every model reaches this stage, only the ones that managed to reach satisfying levels of accuracy and performance on the test set can be deployed. Usually, decisions on this stage are taken based on the aim that the model was created for. If the model was created for investment

purposes than the deployment phase is governed by the CIO[53], or the portfolio manager. If the model was created with enterprise function or enterprise strategy it can be governed by the CIO, but also by the COO[54] or CEO[55].

Performance

The evaluation of performance is based on the hypothesis that business and technology strategies are convergent, interdependent, and integrated. That is why when evaluating the performance of a single AI artifacts this must be done by considering its contribution on the overall strategy of the company. The two key questions to ask are:

Did the AI projects contribute to deployment and operationalization of the overall strategy of the company?

To which extent did the AI projects contribute?

Put simply we are trying to verify if the models deployed were successful in the first place and then find the most correct way to measure this success/failure.

We can divide the performance measurement in business performance and technological performance. Business performance can vary based on the strategic goals of the company analysed, but the overall concept is to verify if those goals were reached or not. This kind of performance is segmented in six capabilities:

Strategy performance. Here we measure factors related to the competitive advantage of a firm. These are usually, AUM variation, clients variation, brand power, brand awareness. These measures give indication on whether the business model is helping to reach a competitive advantage.

Functional performance. Here we measure the performance of each function against the goal that were set. For an investment management firm usually we have operations, risk management, compliance, and trading.

[53] The Chief Information Officer has the role to manage technology systems and handle their application inside the business.

[54] Chief Operation Officer.

[55] Chief Executive Officer.

Investment performance. Usually, the investment performance is measures with the mechanism of GIPS[56].

CSR. This capability measures the performance in terms of corporate social responsibility. It shows the impact that the company has on society.

ROI. This is an overall assessment of the company. We should expect that introducing AI models will contribute to performance with a satisfactory return on investment. Therefore, ROI is measure at corporate and at project level.

Technological performance is a more straightforward performance to measure. It is divided in four main areas.

Data assessment. This measures whether the data available for the project was of good quality and if it was effective for the purpose of the model.

Number of AI models developed. It simply measures the number of AI artifacts created inside the company considering the ones that reached the modelling stage.

Number of AI evaluated. It measures the number of artifacts that reached the evaluation phase. A great number is not always a sign of an efficient technology department as what really matters in the end is how many artifacts are actually deployed.

Number of AI deployed. This measures the number of artifacts deployed. This area also measures productivity, effectiveness, and collaboration. Productivity is how quickly can the organization create new artifacts and the percentage of automation goal reached with respected of all automation goal set. Effectiveness measure if the artifacts reached the goal. The indicators used for effectiveness are usually bias, variance, and overfitting test. Collaboration is a measure of how the AI solutions performs well in a work chain when they undergo human-machine or machine-machine interaction.

[56] Global Investment Performance Standards are used to measure investment performance by investment managers throughout the world. They are a set of voluntary and easy to compare standards.

3.3 ML Stock price prediction model using supervised learning

Problem definition

Stock prices prediction is one of the key challenges in an investment management company, and it has been for decades. However, several companies are still stuck with the same deterministic model that they have been using for years. The field has evolved immensely in the last decade with new technologies like Machine Learning. They naturally lend to stock prediction problems using historical data given their ability to learn the relationship between variables in a nondeterministic way.

In this chapter I will illustrate a model that makes use of supervised learning to predict stock prices. In supervised learning the goal is to train a model using labelled data in order to make predictions on future data. This means that in this case the data sample used has already an output signal attached. The model that will be explained make use of regression, whereas regression is used to predict a continuous variable using a number of explanatory variables.

The following model is inspired by the stock prediction model as described in the book "Machine Learning & Data Science Blueprints for Finance".[57]

The variable that will be predicted by the model is the weekly return of Microsoft shares. The explanatory variables that can be chosen to predict the stock return mainly fall under 3 categories:

Correlated assets. Every company is fully immersed in a network of factors that can have an important influence. Competitors, financial companies, clients, but also commodities, currencies, indices can have a correlation with the company we want to analyse.

Technical indicators. Indicators such moving average or exponential moving average are followed by a vast number of investors.

Fundamental analysis. Mainly composed by performance reports further declined in metrics such ROE or ROI, and news.

[57] Tatsat, Hariom, Sahil Puri, and Brad Lookabaugh. 2020. Machine Learning and Data Science Blueprints for Finance - From Building Trading Strategies to Robo-Advisors Using Python. O'Reilly.

For the scope of this thesis the model will make use of correlated assets as it is more valuable for the reader to understand the concept on a simple model as opposed to a model with a too high degree of complexity. Nonetheless, other features can be incorporated inside the model to increase its precision in predicting the desired output.

The independent variable used by this model are the following:

Stocks. IBM, Alphabet
Currencies. USD/JPY, GBP/USD
Indexes. S&P 500, Dow Jones, VIX

These are the candidates chosen for their potential to be correlated to the explanatory variable.

Data will be composed mainly by the closing price[58] of the above-mentioned assets in the past 10 years (2009-2019). Data regarding stocks will be taken from Yahoo Finance while data on currencies and indices will be taken from Fred[59].

Data gathering

The first step in designing a model is the most important one. Not only it is important to gather good quality data, but also the data that has embedded the right amount of information on the explanatory variable. A first stride in this direction is done by choosing the right independent variables and this will be checked in the next pages both in an analytical and a visual way. Secondly, given that the data used will be composed by past closing prices of assets, it is important to check the completeness of the dataset and the reliability of the data provider. Finally, it is time to export the data from Yahoo Finance and Fred through a function called Pandas DataReader:

[58] The number of stocks closing price is of course different from the one of currencies, as stocks do not register a closing price during bank holidays. For this reason, the number will be aligned for the time series analysis.

[59] Short for Federal Reserve Economic Data, FRED is an online database consisting of hundreds of thousands of economic data time series from scores of national, international, public, and private sources.

```
# Next, we define variables and we extract the data required for
our analysis using pandas datareader.

y_stk = "MSFT"

x_stk1 = "IBM"
x_stk2 = "GOOGL"

x_ccy1 = "DEXJPUS"
x_ccy2 = "DEXUSUK"

x_idx1 = "SP500"
x_idx2 = "DJIA"
x_idx3 = "VIXCLS"

start = datetime.datetime(2008, 12, 31)
end = datetime.datetime(2019, 12, 31)

stk_tickers = [y_stk, x_stk1, x_stk2]
ccy_tickers = [x_ccy1, x_ccy2]
idx_tickers = [x_idx1, x_idx2, x_idx3]
stk_data = web.DataReader(stk_tickers, 'yahoo', start, end)
ccy_data = web.DataReader(ccy_tickers, 'fred', start, end)
idx_data = web.DataReader(idx_tickers, 'fred', start, end)
```

Next, the dependent and independent variable are defined as “y” and “x”. Microsoft weekly return is the dependent variable, and it is computed using five trading days as an approximation of a week. The independent variable are the 5-days return of IBM, Google, USD/JPY, GBP/USD, S&P 500, Dow Jones, and VIX. Moreover, the historical lagged return of Microsoft (5, 15, 30, 60 days) to incorporate the time series component.

With this step we are reframing the times series data points into a supervised model framework based on regression.

```
return_period = 5

Y = np.log(stk_data.loc[:, ('Adj Close',
y_stk)]).diff(return_period).shift(-return_period)
Y.name = Y.name[-1] + '_pred'

X1 = np.log(stk_data.loc[:, ('Adj Close', (x_stk2,
x_stk1))]).diff(return_period)

X1.columns = X1.columns.droplevel()

X2 = np.log(ccy_data).diff(return_period)
X3 = np.log(idx_data).diff(return_period)


X4 = pd.concat([np.log(stk_data.loc[:, ('Adj Close',
y_stk)]).diff(i) for i in
```

```
                [return_period, return_period * 3, return_period *
6, return_period * 12]], axis=1).dropna()
X4.columns = [y_stk + '_DT', y_stk + '_3DT', y_stk + '_6DT', y_stk
+ '_12DT']

X = pd.concat([X1, X2, X3, X4], axis=1)

dataset = pd.concat([Y, X], axis=1).dropna().iloc[::return_period,
:]
Y = dataset.loc[:, Y.name]
X = dataset.loc[:, X.columns]
```

Data analysis

It is possible to visualize the dataset that has been generated through dataset.head(), which will give as output the first 5 lines of the table.

	MSFT_pred	GOOGL	IBM	DEXJPUS	DEXUSUK	SP500	DJIA	VIXCLS	MSFT_DT	MSFT_3DT	MSFT_6DT	MSFT_12DT
2010-03-31	0.021	1.741e-02	-0.002	1.630e-02	0.018	0.001	0.002	0.002	-0.012	0.011	0.024	-0.050
2010-04-08	0.031	6.522e-04	-0.005	-7.166e-03	-0.001	0.007	0.000	-0.058	0.021	0.010	0.044	-0.007
2010-04-16	0.009	-2.879e-02	0.014	-1.349e-02	0.002	-0.002	0.002	0.129	0.011	0.022	0.069	0.007
2010-04-23	-0.014	-9.424e-03	-0.005	2.309e-02	-0.002	0.021	0.017	-0.100	0.009	0.060	0.059	0.047
2010-04-30	-0.079	-3.604e-02	-0.008	6.369e-04	-0.004	-0.025	-0.018	0.283	-0.014	0.007	0.031	0.069

An effective and fast way to confirm the quality of the data is to visualize it. For example, it is paramount to check the correlation and interdependence of data, not only between dependent and independent variables, but also among independent variables. Some algorithms in fact can have a poor performance if the inputs are too much correlated. The best way to check these instances is through a correlation matrix. There are Python modules to fulfil this task in a fast and effective way:

```
# In order to get a sense of the interdependence of the data we
look at the scatter plot and the correlation matrix

correlation = dataset.corr()
pyplot.figure(figsize=(15, 15))
pyplot.title('Correlation Matrix')

sns.heatmap(correlation, vmax=1, square=True, annot=True,
   cmap=sns.diverging_palette(271.1, 12.2, s=97,
l=60,as_cmap=True))
```

This will generate as output the following correlation matrix.

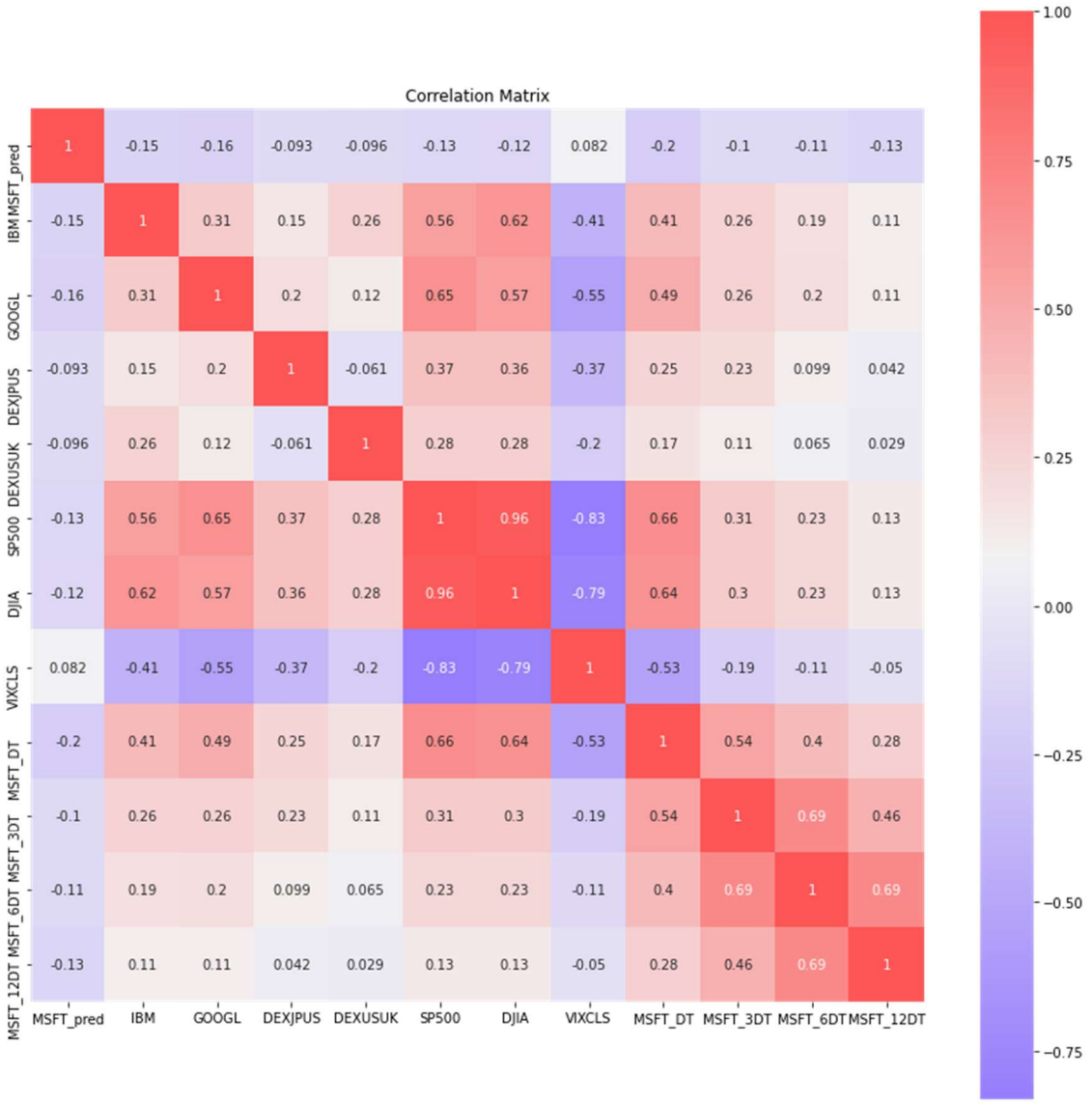


Figure 21 - Correlation matrix for the model variable. Negative values highlighted in magenta represent negative correlation, positive values highlighted in red represent positive correlation.

From this simple matrix we can gather a great number of useful information. We can see for example as the predicted variable has a good correlation with the lagged returns of Microsoft. We can also see the high correlation between Dow Jones and S&P 500 returns or the negative correlation between the volatility index and the other assets, which is quite intuitive.

It is also useful to have a look at the decomposition of the time series of Microsoft weekly return into trend and residual components.

```
# Next, we look at the seasonal decomposition of our time series

res = sm.tsa.seasonal_decompose(Y, freq=52)
fig = res.plot()
fig.set_figheight(8)
fig.set_figwidth(15)
pyplot.show()
```

The output generated is:

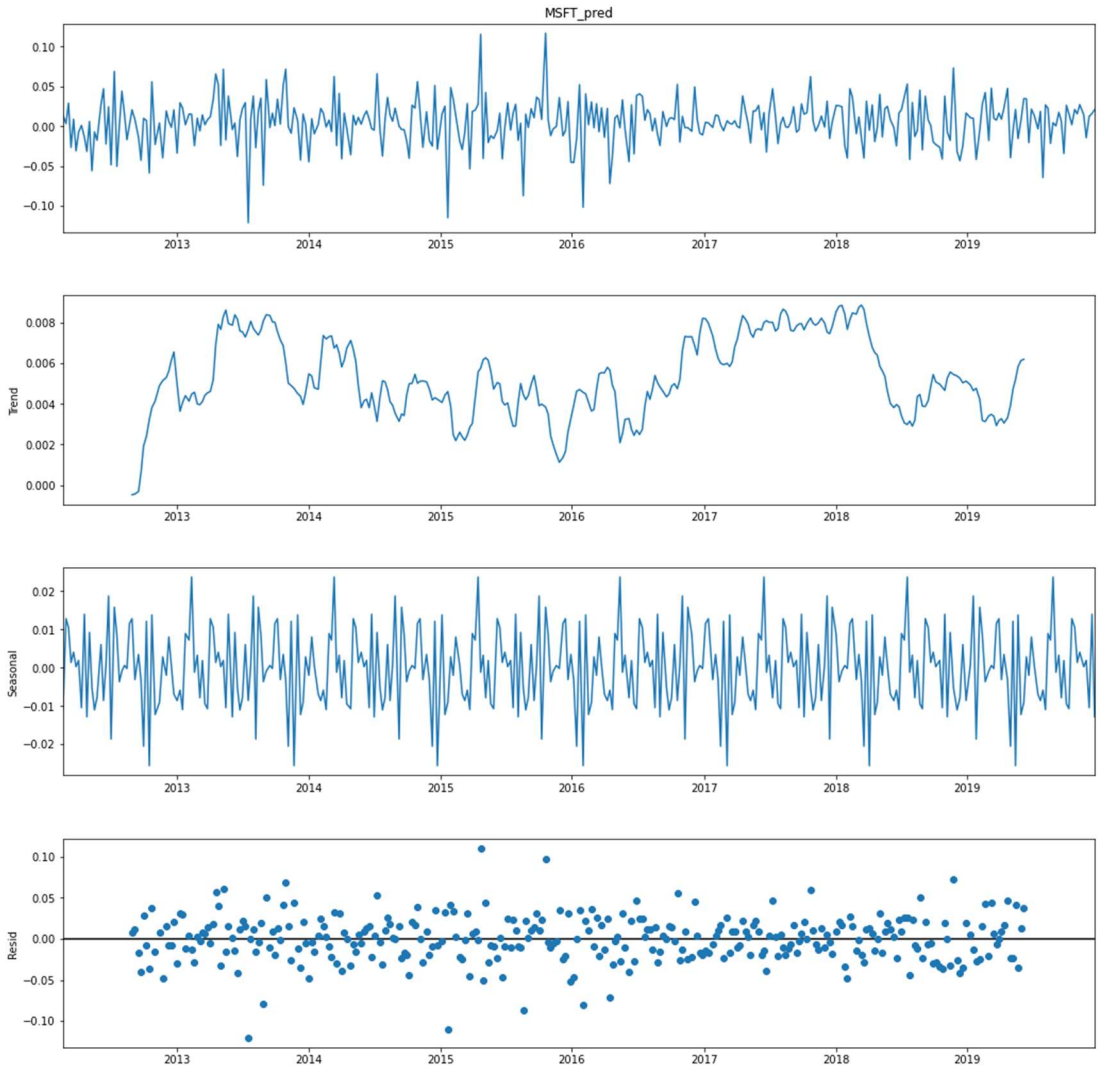


Figure 22 - decomposition of time sieries into observed data, trend, seasonality, and residuals.

In the trend graph we can observe an upward trend that may show up in the model prediction, while in the residual graph we can see a relatively small white noise throughout the whole analysis period.

Another useful tool is the SelectKBest feature from Sklearn that assigns a score to every feature in order to give a sense of their importance in predicting the dependent variable.

```
bestfeatures = SelectKBest(k=5, score_func=f_regression)
fit = bestfeatures.fit(X, Y)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X.columns)
# concat two dataframes for better visualization

featureScores = pd.concat([dfcolumns, dfscores], axis=1)
featureScores.columns = ['Specs', 'Score']
# naming the dataframe columns

featureScores.nlargest(12, 'Score').set_index('Specs')
# print 12 best features
```

As output we can see every feature with its own score. Microsoft historical 5days lagged return are intuitively the most important feature, in comparison 6days and 3days lagged returns give a relatively small contribution:

MSFT_DT	15.473
GOOGL	9.654
IBM	8.070
MSFT_12DT	6.667
SP500	6.287
DJIA	5.176
MSFT_6DT	4.337
MSFT_3DT	4.052
DEXUSUK	3.480
DEXJPUS	3.263
VIXCLS	2.49

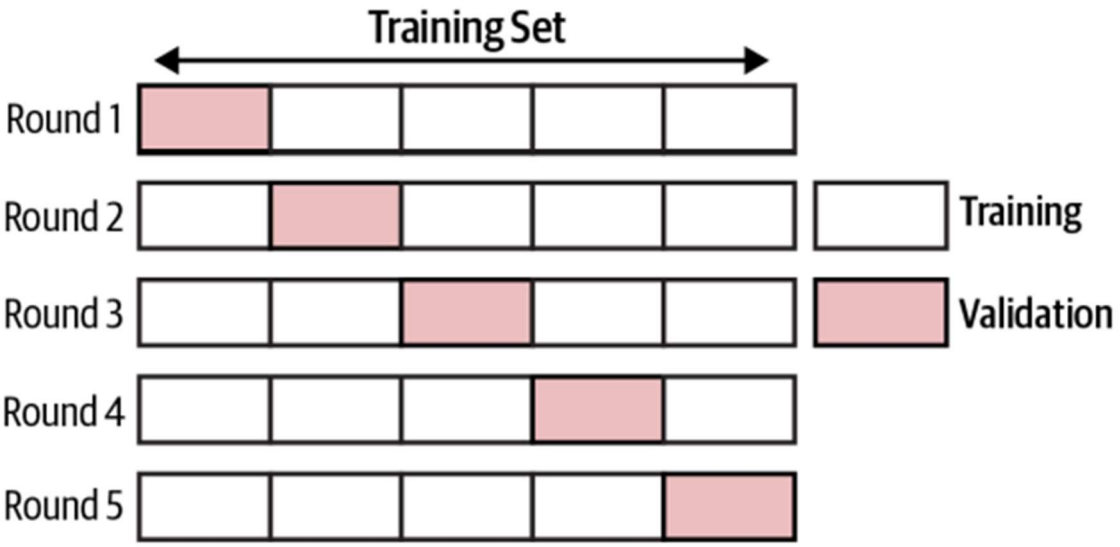
Last step would be the cleaning of data from unperforming or biased feature, or the processing of feature importance. However, the small number of feature and the relative simplicity and good quality of the data is enough to not to go on with further processing.

Model evaluation

When evaluating a ML model, the two common problems that need to be avoided are both underfitting and overfitting of data. Underfitting of data happens when the model is not complex enough and can't fully capture the underlying trends. Overfitting of data is the exact opposite. The model has overlearned from the data used to train it and can't generalize whenever is fed with new data. There are some expedients to combat overfitting, one of which could be to simply use more training data. Nonetheless, in certain circumstances it is hard to gather additional quantity of data, or it is simply too expensive in terms of computing power.

In these cases, a very helpful and ingenious tool is called cross validation. First step is to divide the data into two sets: training set and test set. The first one will be used to train the data and to discover underlying trends, the second one will be used to test if the model performs well with data on which it has not been trained with.

To further increase the effectiveness of this process k-fold cross validation comes on handy. The idea behind cross validation is to optimize the hyperparameters of the models in the training set before assessing them with the test set. This is done by splitting the training set into k equal parts and then training the model on k-1 parts and validating on the k^{th} part. The process is then repeated using different configurations. In this manner it is possible to obtain a reasonable estimate of the generalization of the error of the model.



[60]

Figure 23 - Visual explanation of K-fold validation analysis

The above figure represents an example of k-fold cross validation in which data is split in training and validation set in five different ways.

In the stock price prediction model, there are some observations to be made. First, data will be split into 70% training set and 30% test set as this split seems to perform better than other common splits. Secondly, data cannot be split randomly. Being a time series, the sequence of data points is important, so the dataset has to be split by selecting an arbitrary point in the list of observations.

This process can be done with few lines of code:

```
validation_size = 0.3

train_size = int(len(X) * (1 - validation_size))
X_train, X_test = X[0:train_size], X[train_size:len(X)]
Y_train, Y_test = Y[0:train_size], Y[train_size:len(X)]

num_folds = 10
seed = 7

scoring = 'neg_mean_squared_error'
```

[60] Tatsat, Hariom, Sahil Puri, and Brad Lookabaugh. 2020. Machine Learning and Data Science Blueprints for Finance - From Building Trading Strategies to Robo-Advisors Using Python. O'Reilly.

The k-fold cross validation is executed ten times. In this way we select the optimal hyperparameters for the model. The evaluation method for the algorithm is the negative mean squared error metric. The reason why it is negative is to allow better comparison with other models.

Model comparison

Now that a framework in which the model can run, perform, and be tested is ready, it is time to choose the most adequate model to use. This choice is based on some key factors:

- **Simplicity**. A simple model is usually more scalable and quicker and can be easily explained.
- **Training time**. Time required to train the model.
- **Nonlinearity handling**. Some models can or cannot handle nonlinearity relationship in the data.
- **Resistance to overfitting**. Degree to which a model is prone to encounter overfitting issues.
- **Size of the dataset**. Capacity of the model to handle large training datasets.
- **Number of features**. Ability to handle multiple dimensionalities in the feature space.
- **Model interpretation**. Degree to which a model is easily explainable.

The ecosystems of model to choose from in a supervised learning framework have been described in chapter two. They are listed as following:

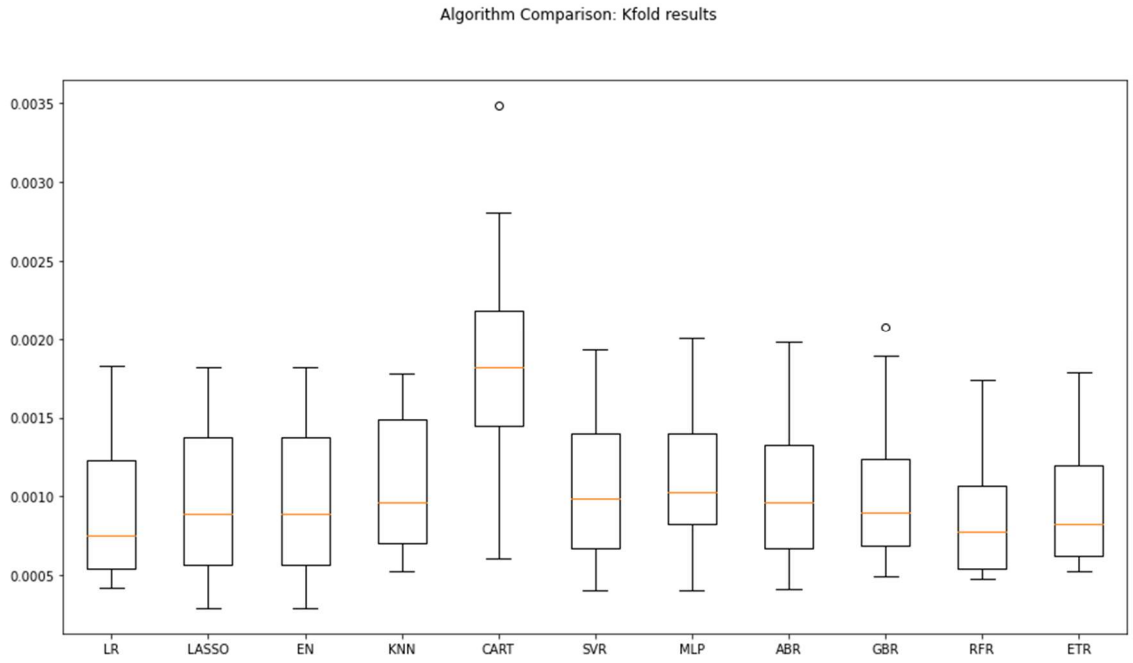
- **Regression and tree regression algorithms**
    - Linear Regression (LR)
    - Lasso (LASSO)
    - Elastic Net (EN)
    - K neighbours Regressor (KNN)
    - Decision Trees Regressor (CART)
    - Support Vector Machine (SVR)
- **Neural network algorithms**
    - MLP Regressor (MLP)

- **Ensemble models**
    - Ada Boost Regressor (ABR)
    - Gradient Boost Regressor (GBR)
    - Random Forest Regressor (RFR)
    - Extra Trees Regressor (ETR)

The aim is to run the k-fold analysis on each of the above models and then on the entire dataset split in training set and test set.

```
names = []
kfold_results = []
test_results = []
train_results = []
for name, model in models:
    names.append(name)

    ## K Fold analysis:

    kfold = KFold(n_splits=num_folds, random_state=seed)
    # converted mean square error to positive. The lower the beter
    cv_results = -1 * cross_val_score(model, X_train, Y_train,
cv=kfold, scoring=scoring)
    kfold_results.append(cv_results)

    # CHECK
    # Full Training period
    res = model.fit(X_train, Y_train)
    train_result = mean_squared_error(res.predict(X_train),
Y_train)
    train_results.append(train_result)

    # Test results
    test_result = mean_squared_error(res.predict(X_test), Y_test)
    test_results.append(test_result)

    msg = "%s: %f (%f) %f %f" % (name, cv_results.mean(),
cv_results.std(), train_result, test_result)
    print(msg)
```

The results of the k-fold analysis are represented in the below image:



From this image we can see that the models that performed better are the one that can best handle linear relationships (LR, LASSO, EN). This demonstrate that the independent and dependent variables are bonded by a strong linear relationship. This was also confirmed by the visualisation of the correlation matrix in the previous pages.

The results of the models on the entire k-fold validation and test sets are the following:

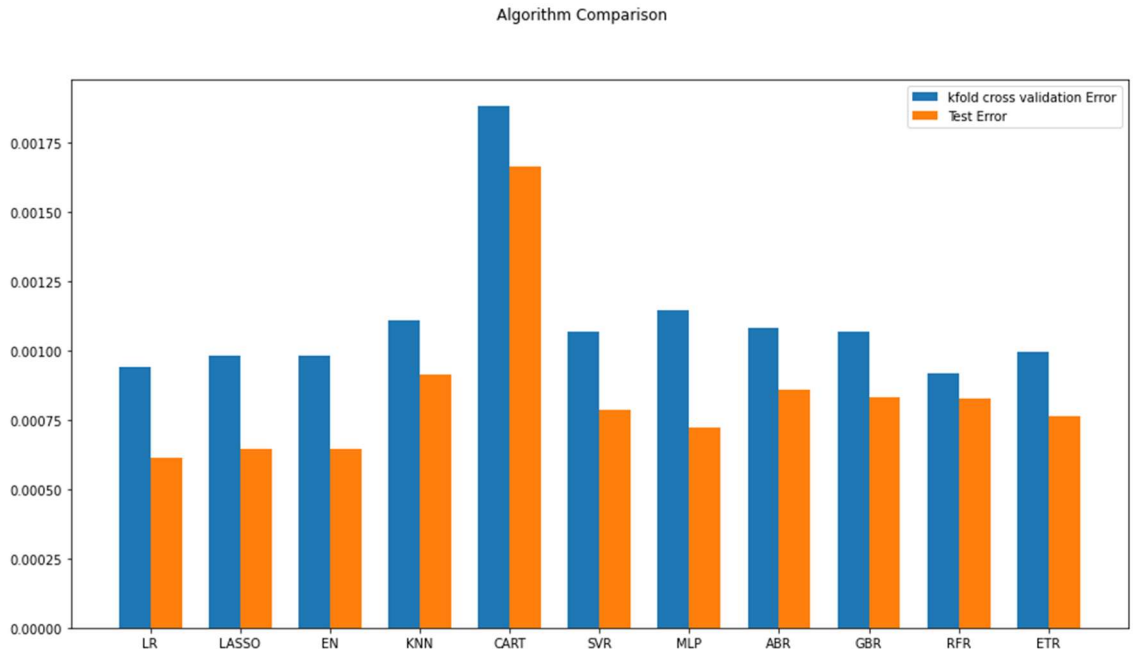


Figure 24 - comparison between errors in k-fold validation and test set in different AI models.

By having a deeper look at these results, we can finalize further conclusion on the most effective model and on the others as well. Decision Tree Regressor (CART) and Extra Trees Regressor (ETR) for example, extremely overfitted data in the validation set resulting in high errors on the test set. A similar conclusion can be drawn on Gradient Boosting Regressor (GBR) and Random Forest Regressor (RFR). A powerful tool like the artificial Neural Network resulted in high errors in both sets. This can be due to the fact that NN are very effective in explaining non-linear relationships and less effective in explaining linear relationships, or because the model was not trained enough time. The stronger performance is shown by linear models, especially the Linear Regression (LR). The original intuition showed in the correlation matrix and in the k-fold analysis is once again confirmed.

Model finalization

Finally, we can take the Linear Regression as the selected model and visualize its output. This can be done by plotting together the forecasted data and the actual data together. The assumption for the sake of simplicity is that the price at the beginning of the test set (01.07.2017) is one.

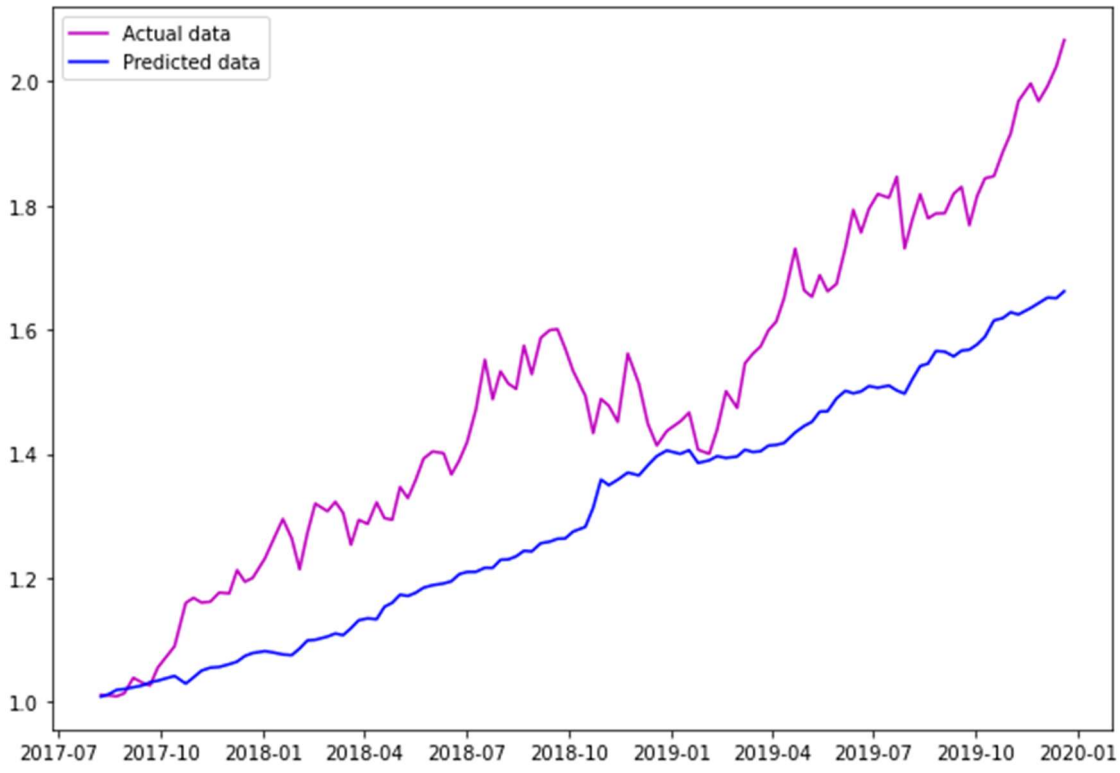


Figure 25 - Actual vs predicted data graph

By looking at the graph it is clear that the model managed to capture the trend perfectly. Volatility is intuitively lower in the predicted series than in the actual one and the alignment between with the real data persists for a few months.

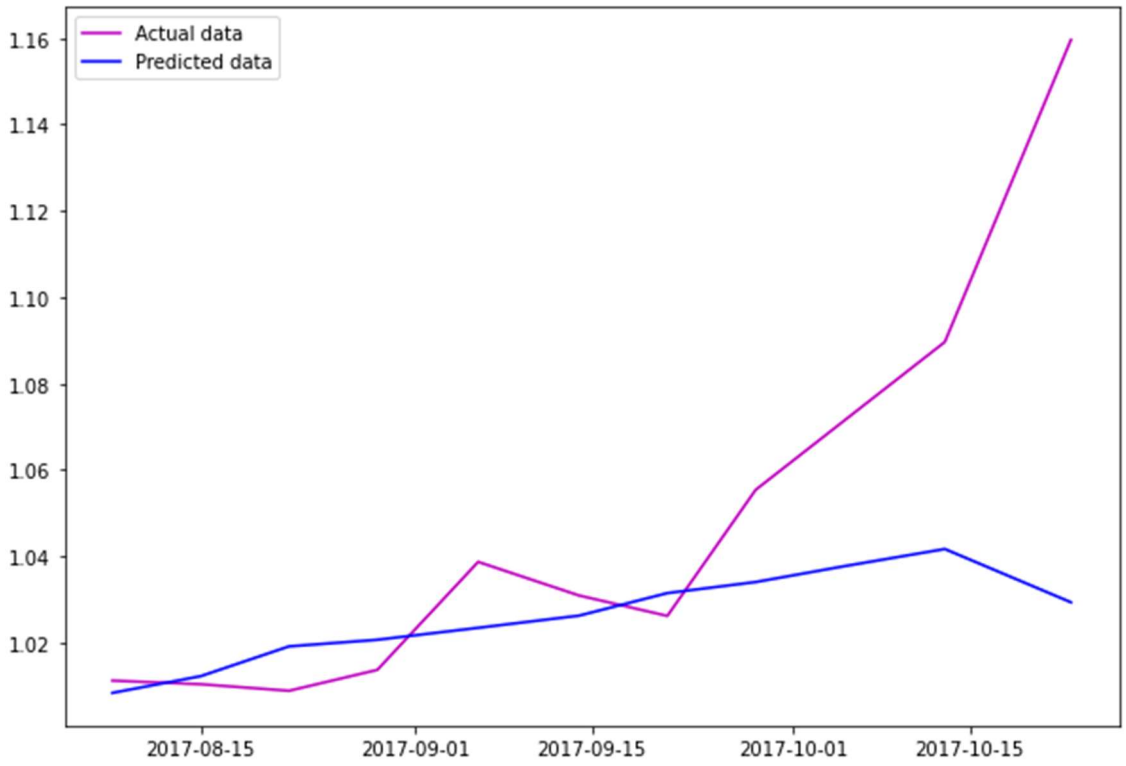


Figure 26 - First 3 months of actual vs predicted data

It is important to notice that the aim of the model is to predict the return of the next day given all the information fed before that day. This is the reason why the model performs well in the first months before moving away from the actual data after about 3 months after the start of the test set.

Model conclusions

This model shows how promising simple supervised regression-based models are to approach complex problems like stock prediction. Moreover, a key problem in finance such as overfitting and underfitting are dealt with in an effective way. Even though one of the key strengths of this model is simplicity, it can be increased in complexity by adding other variables in order to have a much more accurate prediction. These variables can vary from news, analyst reports, volumes, technical indicators, P/E ratio, and much more.

With few lines of code, a supervised-regression modelling framework has been created allowing the possibility to predict stock price using historical data. It is a quick and

effective tool to be used to generate a fast prediction that can be further analysed in terms of risk and profitability to avoid capital loss.

Conclusion

We started by defining some key theoretical concept on Artificial Intelligence and Machine Learning. We then dived into the Investment management industry discovering the massive impact these technologies are creating, concluding that is imperative for companies to start adopting AI solutions to keep pace with competition.

The literature review and the Accenture survey analysis demonstrated how Investment Managers are trying to introduce new technologies or at least enhancing legacy ones. The most important growth/efficiency levers were new revenues opportunities and operating model changes. Two processes that AI solutions naturally lend themselves to solve. This analysis also confirmed the complexity of the transition process to adopt new technologies.

The issue is that for both big and small companies is hard to find a way to systematically introduce new AI artifacts in a way that is coherent and cohesive with the strategic goals and at the same time that unlocks their full potential in terms of effectiveness and synergies. In order to solve these problems, the proposed solution is a strategic framework in which AI solutions can be created, developed, deployed, nurtured, and, if necessary, dismantled.

We demonstrate that in the empirical example of the investment management company AQA Capital and its willingness to adopt a strategic framework to manage the transition to a more technology-based company.

The strategic framework presented in this thesis starts by defining the goal of the strategy and by comparing them with the medium-long term strategic goals of the company. Keeping this in mind it is possible to define the value chain and decline it in single value drivers. Each value driver needs to be further declined following the steps in the AI/ML lifecycle.

- **Design**. There are certain problems that can be solved with excellent effectiveness through AI models, especially the ones that requires high degree of automation and/or intelligence.
- **Data**. This must be done in a bigger context of a healthy data management program. It is paramount to maximise quality and optimise quantity of data.
- **Modelling**. Through modelling the actual AI model is developed keeping in mind the overall strategic goals of the company.
- **Evaluation**. In this delicate phase the model is tested to uncover potential inaccuracies. Underfitting and overfitting are the most common issues.
- **Deployment**. Whenever the model is accurate and it demonstrates to show good performance, it can be finally deployed inside the organization.
- **Performance**. Periodical controls to the model need to be executed in order to check if the performance is adequate to the goals. This check is done both on the company's overall goals level and on a single model level.

This strategic framework was applied in the context of the AQA Capital 2.0 project to enhance technological awareness inside and outside of the company.

The model created is a machine learning stock price prediction model using supervised learning. It serves the purpose of automating the task of data gathering, model tuning, and problem fixing of the old legacy model for small day to day investment decisions. The program takes as input stocks indexes and currencies returns and predicts the weekly returns of a correlated asset. The model managed to predict with extremely high accuracy the weekly return of Microsoft stock by revealing the underlying relationship with other connected variables like other stock returns, indices, currencies, and lagged return to incorporate a time series factor. These tasks were computed by using Machine Learning technologies in the context of Supervised Learning algorithms.

To conclude we can affirm how Investment management companies are more and more willing to embrace new technology solutions to their operational and strategic needs. We showed how AI and ML are one of the most revolutionary and effecting technologies to solve these needs and they naturally lend themselves to be applied in financial contexts. The problem for investment companies that adopt AI and ML solutions is to introduce single artifacts without a cohesive and coherent approach with their medium long-term strategy. This usually bring to a so-called chaotic AI artifacts accumulation. We empirically demonstrated how the application of a strategic framework to introduce a ML

model inside AQA Capital brought to an effective synergy's exploitation across the company functions and a coherent application of the model with the company objectives.

Resume

Introduction

In the last decade some important steps forward have been made in the field of AI. The power of processors has reached an astonishing level contributing to the functioning of some complex types of algorithms. Lack of data is now not a problem anymore. Big corporations made the collection of data one their core responsibility and activity given the discovery of its immensely huge value as it can be fed to an algorithm to extract even more value. These are only few of the reasons why Investment management industry finds itself at the cusp of a major transformation.

On one hand there are companies that are understanding the importance of anticipating the future by adopting new technologies. On the other there are those more reluctant to change and still attached to their legacy processes but want to keep up with the rest of the industry. In both cases their willingness to progress is often exacerbated by the complexity if the transition. The problem is that for the sake of introducing new technologies at all costs the transition itself is often underestimated. The main aim of this thesis is to create a broad strategic framework in which investment management companies can introduce modern AI solution to problems in a coherent and cohesive way with the broader strategy of the company.

For the purpose of this thesis, it is worth to dig deeper into the conceptual framework of these technologies and, further on, to their application in the finance world. In this way whatever the background of the reader he or she will find a broader and resistant basement to make a judgement and to evaluate its impact.

Theoretical framework

Machine Learning is a subset of AI technologies. By defining it first it is easier to zoom out to grasp the broader concept of AI. The first person to discover what we define today as ML is called Samuel Arthur (1959). He come up with this idea while playing the game of checkers. “I want to verify the fact that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program.” This idea led to the definition of ML that persists today:

“A ML program has the ability to learn without the need to be explicitly programmed.”

The difference with traditional programs is in the input. While traditional programs need an input command, a ML program need input data. This data is fed into the machine. The programmer chooses an algorithm and select the relevant penalty function to reach his goals. The algorithm adjusts them through a process of trial and error to discover hidden patterns in the data and build a model. Next time it will get data as input the program will be able to decipher the pattern and predict future values.

There are three main categories of ML algorithms Supervised Learning, Unsupervised Learning and Reinforcement Learning.

Supervised Learning works by deciphering relationship between variables and outcomes that are given. The program will be trained on this labelled data, and it will try to figure out patterns in the data the describe the correct value of the label. The goal of Supervise Learning models is to take a feature vector as input (x) and determine the most precise label (y).

Unsupervised Learning works in a similar way of Supervised Learning. Nonetheless, it deals with unlabelled data. The aim of Unsupervised Learning algorithms is to analyse data to detect pattern and find a viable label.

Reinforcement Learning is the last and most complicated category of ML. In this case the machine will recognize to be present in an environment and will be able to detect the state of the environment. The machine has the possibility to take different actions that will bring to different results. These results are graded instead of being labelled. An action that takes the program closer to the goal will be graded positively and vice versa. The goal of the machine is to learn a policy. A policy takes information about the state as input and output the action that maximize the expected average reward, in other words the optimal action for that state.

The main categories of ML algorithms are Regression analysis, Clustering, Artificial Neural networks, Decision trees, Ensemble modelling.

Regression analysis can be described as the simplest Supervised Learning algorithm. It is used to estimate real values based on continuous values. In other words, it is used to find the best trendline to describe a dataset.

Clustering analysis algorithms are useful to identify clusters. They can also be used to assign a new data point into existing clusters based on some information. Clustering analysis falls under the scope of Supervised and Unsupervised Learning and the 2 most popular algorithms are K-Nearest Neighbours and K-Means Clustering.

Artificial neural network algorithms are models that try to emulate the functioning of the brain neurons. As neurons they are composed of nodes and edges. The nodes store a numeric function. When they receive an input, they will fire in different patterns according to that input. The edges store a numeric weight. If the activated nodes and its weight satisfy a certain threshold, they will activate different nodes in the following layer. If the threshold is not satisfied, they will not fire the nodes in the following layer. Finally, some nodes will fire to reach the output layer, where the output is registered. This output is then evaluated with a penalty function and the weights are adjusted and improved to have a better output in the following epoch. The process is then repeated and the NN is trained until the output reaches certain satisfaction threshold.

Artificial intelligence in investment management industry

One industry that perfectly suits the huge potential of AI technology is the investment management industry. To better grasp the magnitude of this reshaping process, a 2019 PWC's report predicted the AI will bring to the global economy as far as $15.7 trillion by 2030. In another report of McKinsey AI technologies will be able to afford 1.2% of global GDP.

As of today, AI has numerous applications in a wide range of fields. It is already dominating the trading floors of big banks, but it has also applications in lending, deposits, insurance, payments services, asset management, risk management, and the list goes on. Just as neural networks are used to analyse lots of inputs and breaking them down to have a reasonably correct output, the same principles are used to make predictions about stock

and bond market prices. A huge number of variables can be taken as input, like fundamentals and other companies' data, to make ever more correct predictions.

From the 70s we can identify some major trends that are a direct consequence of all these reshaping forces weighting on the shoulders of the Asset Management industry: widening of the products offering, tapering of the profit margins, change in the role of advisors and counsellors, disruptive technology breakthroughs.

All the aforementioned radical trends are emerging faster. They all point towards one simple conclusion. A conclusion that is similar to what is happening in more and more industries. To survive is vital to keep up with the new technologies. When the technologies are so important and disruptive like AI, this issue becomes even more paramount.

"The old model of asset management is not only non-competitive, but also counterproductive. It fails to serve the best interests of clients, and it hurts the profitability of the firm. It fails to offer the quality of investment expertise that today's clients expect and deserve. It ignores important elements of human emotions, qualitative data, behaviours, narratives, and other human processes."[61].

The reason for the non-competitiveness mentioned above relies in a series of challenges that asset managers are facing. These challenges are even more exacerbated by market pressures and competitive forces from within the industry and from external factors.

"42 percent of operations executives believe their operations and technology are not configured to adequately execute the firm's overall strategy. However, as we will see, they are not at all complacent about this situation, and are moving forward to address it."[62]

In the next paragraphs I will define and analyse the challenges I just mentioned. The basement for the analysis is provided by some major studies and reports carried out by Accenture Consulting and BlackRock on the future of Asset management.

[61] Naqvi, Al. 2020. Artificial Intelligence for Asset Management and Investment. Wiley.

[62] Accenture Consulting; Investment Company Institute (ICI). 2019. «Reinventing Operations in Asset Management.»

Accenture Consulting recently published a study[63] conducted on ICI[64] members. They asked to 33 asset managers representing about 15 trillion of AUM[65] what they think will be in the future the major driver for success in the asset management industry. The number one answer, new revenues opportunities, reflects the undergoing process of radical changes in the industry. Profit margins becoming more and more slender is a key hint to reach and exploit the countless opportunities of innovation in a wide sense.

Technology can be used to figure out future trends in asset price movements as well as efficiently uncover hidden patterns that can bring to new ways of approaching the selection of assets for a specific portfolio. Apart from the contribution that AI can give to stock picking activities, there are more and more AI based ETFs being created.

Back and middle office operations play a vital role in the daily activities of an asset management company. That's why in a context of margin tapering, asset management companies are trying to support or change the operating model. A smooth and flexible operating model can play a decisive role in supporting the services and approaching the clients.

The major factor that will act as a catalyst in supporting the way towards new operating model is of course technology. Technology and innovation are considered to be the bridge that connects growth with efficiency. In most of the industries the capacity to discover and adopt the right technology has become vital, not only to compete but to survive. In the survey we can see that 55% of the interviewed company have a public initiative to embrace new tech tools or fintech solutions.

Investment management company are more and more trying to find ways to cut the unnecessary work and reduce human intervention. The first reason is purely for cost reduction purposes. Employee salary is one of the greatest expenses in financial company balance sheets and technology can decrease them notably. The other reason is that reducing human errors in back-office operations can be immensely helpful. Finally reducing the time on repetitive and alienating tasks frees up time for people to concentrate on the things that matter or that requires different skills.

[63] Accenture Consulting; Investment Company Institute (ICI). 2019. «Reinventing Operations in Asset Management.»
[64] Investment Company Institute. It is an association representing regulated funds all over the world (ETFs, Closed-end funds, UITs).
[65] Asset Under Management.

Another common challenge the whole industry is facing is related to data. Financial firms find themselves with tons of data, often unstructured without an actual plan to extract value from it. The challenge for the firm is represented by data quality and accessibility. High quality data doesn't refer only to the extent the information is valuable, but also the degree to which the data has been cleaned and stored in a correct way. This is fundamental to enable analytics and automation. It also creates the possibility of a more effective reporting process for consumers, clients or even directors. Uncleaned and unprocessed data is not only dangerous for the entire company by creating faulty analysis and reputational risk, but it also gives rise to opportunity cost as the world is more and more demanding towards good data management

From the vast ecosystem of different technology to use to solve these problems AI are the most promising owns. AI has in fact already deeply affected the industry. A lot of progress has been made in the use of AI in the front office, especially with the advance in NLP technologies. But that is just one of the numerous areas it is being used. Nowadays, there are funds which are being managed through AI, others have algorithms that can forecast geopolitical events and base some investment decision on the outcome.

AI in back and middle office hasn't still reached a use at scale, but larger firms have completed an AI-based reshaping process, or they have one underway. Most probably AI technologies will deliver the next wave of cost reduction throughout the industry.

Finally, another viable solution is outsourcing of middle, back-office operations. This solution is faster, simpler and effective in the short term, while it loses significance in the long run where an in-house solution can bring more to the table. It has also the advantage of being ideal for small-medium companies that do not have the necessary capabilities to build such technologies.

AQA capital case study

AQA Capital is an Asset Management Company authorised by the Malta Financial Services Authority ("MFSA") as a full scope Alternative Investment Fund Manager in terms of the Directive 2011/61/EU and a UCITS manager in terms of the Directive 2009/65/EC as amended. AQA Capital holds a Category 2 investment services licence issued by the MFSA in terms of the Investment Services Act (chapter 370 of the laws of Malta). AQA Capital can act as investment manager for both AIF and UCITS Funds and

promotes these funds throughout the EEA. AQA Capital is also authorized to provide discretionary portfolio and investment advisory service to retail and professional clients.

AQA is an operationally ready and economically viable solution that aims to cover the entire value chain of asset management industry, providing its clients tailored made services, including:

- Financial engineering solutions.
- Fund management services.

From the start of 2021 AQA Capital is undergoing a digitalisation process to enhance its technological footprint. One important goal of this venture is to decrease the level of human intervention in its day-to-day operations by automating repetitive and alienating tasks with the consequence of using staff work for more meaningful and complex tasks. Another important ambition for AQA is to increase its technological reputation among stakeholders, showing its willingness to embrace innovation in an active and passionate way. Whoever worked inside the company was solicited to suggest some ideas that could contribute to reach the goal of this digitalisation process. After a round of presentations in May 2021, the ML learning stock price prediction model that I suggested was chosen as the best initiative to carry out. The model together with the strategical framework to apply it inside the company are shown in the following lines.

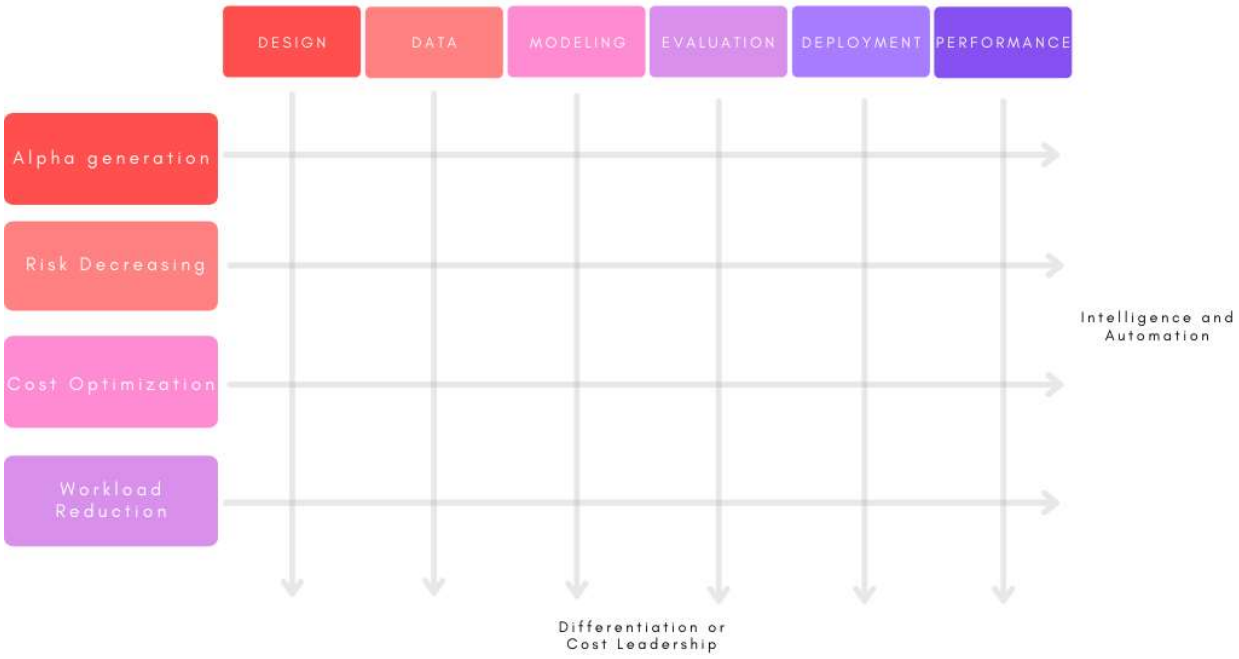
IN order to design a modern investment company, it is necessary to set some ambitious yet reachable goals and comparing them to the medium long-term ones of the company.

AQA for example is a company that invested a vast number of resources into improving their most valuable assets, people. Not only people that work inside the company, but a far greater group of people that can be recognized in its stakeholders. Moreover, AQA valorise people as individuals and as part of a group. It also valorises connections and synergies that emerge the group itself. The following goals perfectly embed these values:

- Structural coherence.
- Interdependence.
- External Interaction.
- Performance Maximisation.
- Cohesive value building.

A company that manages to reach all these goals will reach excellence in becoming a modern investment company.

By taking the above value drivers as underlying assumption is it possible to create a model that will manage the strategic transition from a chaotic technology management to a modern AI approach. The assumption above considers that the companies are collection of capabilities, and those capabilities needs to be extended above the singular functional area and across the whole company.



On a vertical level we have the normal value chain of a firm. The value drivers have different underlying objectives.

The horizontal level is based on the scientific method. In this particular case it is declined to reflect the life cycle of ML and AI models.

As explained in the previous chapters these areas are competency of Machine Learning and are independent of the capability areas on the vertical axis. These competencies are active by the development of Machine Learning models and solutions.

With this model every ML solution can be created, developed, deployed, nurtured, and eliminated at the end of its work.

The performance of the models is compared to human performance in terms of efficiency and effectiveness and analysed by keeping in mind the goal of the company and the external environment.

Design is the first step towards the optimal AI evolution, and it is aimed to build alternative processes or replace old ones. It is defined as “the process of architecting a firm’s AI plan that supports the firm’s strategy.

Data is the most challenging part of the life of an AI model. As previously addressed in the previous chapters it needs to be of good quality and in the right quantity to avoid both underfitting and overfitting. The key for an effective use of data inside a company is a good and healthy data management program.

Modelling is the phase where ML solutions are created. It is the phase in which data is used to extract a pattens with the finale ai to build a model. In an investment management firm AI artifacts fall under three main categories:

- Enterprise strategy. Related to the strategic position of the company
- Investment. Based on strategic investment objectives.
- Enterprise function. Residual class to include other types of artifacts.

Evaluation is all about making sure that the model works. the term “works” can be expressed in different connotations, but mainly it means that the algorithms are accurate, they do not overfit or underfit and they are efficient. Accuracy means that the model can deliver accurate prediction. This is not a problem that has a static solution, as the accuracy can change over time. Solution to the accuracy problem can be adding or changing features if the one used are sub optimal, making sure that the training data distribution contains cases and conditions in which the algorithms will perform, check if the underlying distribution has changed and make an update if necessary.

In the deployment phase models are finally used inside the organization. Not every model reaches this stage, only the ones that managed to reach satisfying levels of accuracy and performance on the test set can be deployed. Usually, decisions on this stage are taken based on the aim that the model was created for.

The evaluation of performance is based on the hypothesis that business and technology strategies are convergent, interdependent, and integrated. That is why when evaluating the performance of a single AI artifacts this must be done by considering its contribution on the overall strategy of the company. The two key questions to ask are:

Did the AI projects contribute to deployment and operationalization of the overall strategy of the company?

To which extent did the AI projects contribute?

Put simply we are trying to verify if the models deployed were successful in the first place and then find the most correct way to measure this success/failure.

ML Stock price prediction model using supervised learning

Stock prices prediction is one of the key challenges in an investment management company, and it has been for decades. However, several companies are still stuck with the same deterministic model that they have been using for years. The field has evolved immensely in the last decade with new technologies like Machine Learning. They naturally lend to stock prediction problems using historical data given their ability to learn the relationship between variables in a nondeterministic way.

This model makes use of supervised learning to predict stock prices. The variable that will be predicted by the model is the weekly return of Microsoft shares. For the scope of this thesis the model will make use of correlated assets as independent variables as defined below:

- **Stocks**. IBM, Alphabet
- **Currencies**. USD/JPY, GBP/USD
- **Indexes**. S&P 500, Dow Jones, VIX

Data will be composed mainly by the closing price[66] of the above-mentioned assets in the past 10 years (2009-2019). Data regarding stocks will be taken from Yahoo Finance while data on currencies and indices will be taken from Fred[67].

The first step in designing a model is the most important one. Not only it is important to gather good quality data, but also the data that has embedded the right amount of information on the explanatory variable. A first stride in this direction is done by choosing the right independent variables and this will be checked in the next pages both in an analytical and a visual way. Secondly, given that the data used will be composed by past closing prices of assets, it is important to check the completeness of the dataset and the

[66] The number of stocks closing price is of course different from the one of currencies, as stocks do not register a closing price during bank holidays. For this reason, the number will be aligned for the time series analysis.

[67] Short for Federal Reserve Economic Data, FRED is an online database consisting of hundreds of thousands of economic data time series from scores of national, international, public, and private sources.

reliability of the data provider. Finally, it is time to export the data from Yahoo Finance and Fred.

Next, the dependent and independent variable are defined as “y” and “x”. Microsoft weekly return is the dependent variable, and it is computed using five trading days as an approximation of a week. The independent variable are the 5-days return of IBM, Google, USD/JPY, GBP/USD, S&P 500, Dow Jones, and VIX. Moreover, the historical lagged return of Microsoft (5, 15, 30, 60 days) to incorporate the time series component. An effective and fast way to confirm the quality of the data is to visualize it. For example, it is paramount to check the correlation and interdependence of data, not only between dependent and independent variables, but also among independent variables. Some algorithms in fact can have a poor performance if the inputs are too much correlated. It is also useful to have a look at the decomposition of the time series of Microsoft weekly return into trend and residual components.

When evaluating a ML model, the two common problems that need to be avoided are both underfitting and overfitting of data. In these cases, a very helpful and ingenious tool is called cross validation. First step is to divide the data into two sets: training set and test set. The first one will be used to train the data and to discover underlying trends, the second one will be used to test if the model performs well with data on which it has not been trained with.

To further increase the effectiveness of this process k-fold cross validation comes on handy. The idea behind cross validation is to optimize the hyperparameters of the models in the training set before assessing them with the test set. This is done by splitting the training set into k equal parts and then training the model on k-1 parts and validating on the k^{th} part. The process is then repeated using different configurations. In this manner it is possible to obtain a reasonable estimate of the generalization of the error of the model.

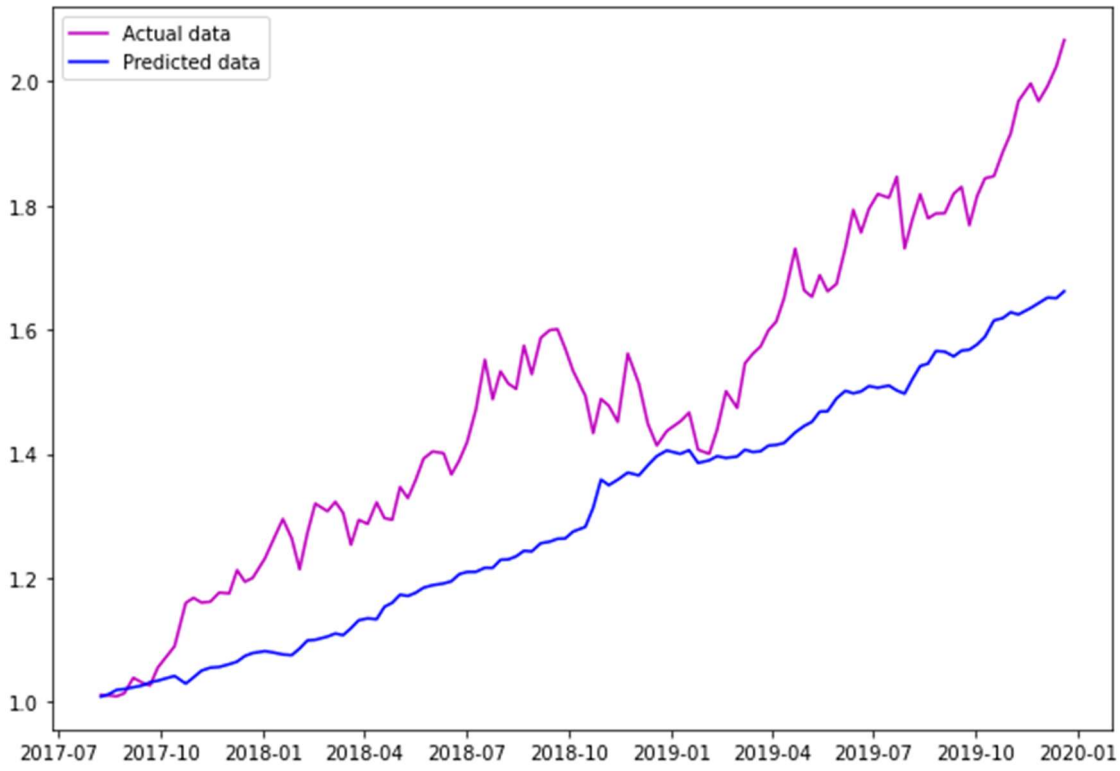
In the stock price prediction model, there are some observations to be made. First, data will be split into 70% training set and 30% test set as this split seems to perform better than other common splits. Secondly, data cannot be split randomly. Being a time series, the sequence of data points is important, so the dataset has to be split by selecting an arbitrary point in the list of observations.

Now that a framework in which the model can run, perform, and be tested is ready, it is time to choose the most adequate model to use. The ecosystems of model to choose from in a supervised learning framework is listed as following:

- **Regression and tree regression algorithms**
    - Linear Regression (LR)
    - Lasso (LASSO)
    - Elastic Net (EN)
    - K neighbours Regressor (KNN)
    - Decision Trees Regressor (CART)
    - Support Vector Machine (SVR)
- **Neural network algorithms**
    - MLP Regressor (MLP)
- **Ensemble models**
    - Ada Boost Regressor (ABR)
    - Gradient Boost Regressor (GBR)
    - Random Forest Regressor (RFR)
    - Extra Trees Regressor (ETR)

the models that performed better are the one that can best handle linear relationships (LR, LASSO, EN). This demonstrate that the independent and dependent variables are bonded by a strong linear relationship.

Finally, we can take the Linear Regression as the selected model and visualize its output. This can be done by plotting together the forecasted data and the actual data together. The assumption for the sake of simplicity is that the price at the beginning of the test set (01.07.2017) is one.



By looking at the graph it is clear that the model managed to capture the trend perfectly. Volatility is intuitively lower in the predicted series than in the actual one and the alignment between with the real data persists for a few months. It is important to notice that the aim of the model is to predict the return of the next day given all the information fed before that day. This is the reason why the model performs well in the first months before moving away from the actual data after about 3 months after the start of the test set.

This model shows how promising simple supervised regression-based models are to approach complex problems like stock prediction. Moreover, a key problem in finance such as overfitting and underfitting are dealt with in an effective way. Even though one of the key strengths of this model is simplicity, it can be increased in complexity by adding other variables in order to have a much more accurate prediction. These variables can vary from news, analyst reports, volumes, technical indicators, P/E ratio, and much more.

This model is a quick and effective tool to be used to generate a fast prediction that can be further analysed in terms of risk and profitability to avoid capital loss.

Conclusions

We started by defining some key theoretical concept on Artificial Intelligence and Machine Learning. We then dived into the Investment management industry discovering the massive impact these technologies are creating, concluding that is imperative for companies to start adopting AI solutions to keep pace with competition. The issue is that for both big and small companies is hard to find a way to systematically introduce new AI artifacts in a way that is coherent and cohesive with the strategic goals and at the same time that unlocks their full potential in terms of effectiveness and synergies. We demonstrate that in AQA Capital willingness to adopt a strategic framework to manage the transition to a more technology-based company.

The strategic framework starts by defining the goal of the strategy and by comparing them with the medium-long term strategic goals of the company. Keeping this in mind it is possible to define the value chain and decline it in single value drivers. Each value driver needs to be further declined following the steps in the AI/ML lifecycle.

- **Design**. There are certain problems that can be solved with excellent effectiveness through AI models, especially the ones that requires high degree of automation and/or intelligence.
- **Data**. This must be done in a bigger context of a healthy data management program. It is paramount to maximise quality and optimise quantity of data.
- **Modelling**. Through modelling the actual AI model is developed keeping in mind the overall strategic goals of the company.
- **Evaluation**. In this delicate phase the model is tested to uncover potential inaccuracies. Underfitting and overfitting are the most common issues.
- **Deployment**. Whenever the model is accurate and it demonstrates to show good performance, it can be finally deployed inside the organization.
- **Performance**. Periodical controls to the model need to be executed in order to check if the performance is adequate to the goals. This check is done both on the company's overall goals level and on a single model level.

This strategic framework is applied in the context of the AQA Capital 2.0 project to enhance technological awareness inside and outside of the company. The model created is a machine learning stock price prediction model using supervised learning. It serves the purpose of automating the task of data gathering, model tuning, and problem fixing

of the old legacy model for small day to day investment decisions. The model takes as input stocks indexes and currencies returns and predicts the weekly returns of a correlated asset.

Bibliography

I. Accenture Consulting; Investment Company Institute (ICI). 2019. “Reinventing Operations in Asset Management.”

II. Accenture. 2020. “Reinventing Operations in Asset Management.”

III. Bholat, M. Gharbawi, and O. Thew. 2020. *The impact of Covid on machine learning and data science in UK banking.* Bank of England, Bank of England Quarterly Bulletin Q4.

IV. Bostrom, Nick. 2014. *Superintelligence. Paths, Dangers, Strategies.* Oxford University Press.

V. Burkov, Andriy. 2019. *The hundred-page Machine Learning book.*

VI. Fry, Hannah. 2018. *Hello world. Being human in the Age of Algorithms.* W.W. Norton & Company.

VII. IBM. 2020. “ What is automation?”

VIII. IBM Big Data & Analytics Hub. 2020. *The Four V’s of Big Data.* IBM.

IX. McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. 1955. “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.” *AI Magazine*, 31 August.

X. McCorduck. 2004.

XI. Naqvi, Al. 2020. *Artificial Intelligence for Asset Management and Investment.* Wiley.

XII. OECD. 2021. *OECD Business and Finance Outlook 2021: AI in Business and Finance.* Paris: OECD Publishing. https://doi.org/10.1787/ba682899-en.

XIII. S&P. 2019. “Avoiding Garbage in Machine Learning.” https://www.spglobal.com/en/research-insights/articles/avoiding-garbage-in-machine-learning-shell.

XIV. Samuel, Arthur. 1959. “Some Studies in Machine Learning Using the Game of Checkers.” *IBM Journal of Research and Development, Vol. 3, Issue. 3,* Vol. 3, Issue. 3.

XV. Starmer, Josh. 2017. “StatQuest with Josh Starmer.” *Youtube.* 26 june. https://www.youtube.com/watch?v=HVXime0nQeI.

XVI. Tatsat, Hariom, Sahil Puri, and Brad Lookabaugh. 2020. *Machine Learning and Data Science Blueprints for Finance - From Building Trading Strategies to Robo-Advisors Using Python.* O’Reilly.

XVII. Theobald, Oliver. 2017. *Machine Learning For Absolute Beginners.* Scatterplot Press.

XVIII. Zheng, Alice, and Amanda Casari. 2018. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists.* O'Reilly Media.

Code

Neural Network “Hello world”

```
import tensorflow as tf
import numpy as np
from tensorflow import keras

model = tf.keras.Sequential([keras.layers.Dense(units=1,
input_shape=[1])])

model.compile(optimizer='sgd', loss='mean_squared_error')

xs = np.array([-1.0, 0.0, 1.0, 2.0, 3.0, 4.0], dtype=float)
ys = np.array([-2.0, 1.0, 4.0, 7.0, 10.0, 13.0], dtype=float)

model.fit(xs, ys, epochs=500)

print(model.predict([10.0]))
```

ML Stock Price Prediction Model using Supervised Learning

```
# Loading python libraries

import numpy as np
import pandas as pd
import pandas_datareader.data as web
from matplotlib import pyplot
from pandas.plotting import scatter_matrix
import seaborn as sns
from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import GridSearchCV

from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso
from sklearn.linear_model import ElasticNet
from sklearn.tree import DecisionTreeRegressor
from sklearn.neighbors import KNeighborsRegressor
from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.ensemble import ExtraTreesRegressor
from sklearn.ensemble import AdaBoostRegressor
from sklearn.neural_network import MLPRegressor

# Libraries for Deep Learning Models
from keras.models import Sequential
from keras.layers import Dense

# from keras.optimizers import SGD
from tensorflow.keras.optimizers import SGD
from keras.layers import LSTM
from keras.wrappers.scikit_learn import KerasRegressor

# Libraries for Statistical Models
import statsmodels.api as sm

# Libraries for Saving the Model
from pickle import dump
from pickle import load

# Time series Models
from statsmodels.tsa.arima_model import ARIMA
# from statsmodels.tsa.statespace.sarimax import SARIMAX

# Error Metrics
from sklearn.metrics import mean_squared_error

# Feature Selection
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2, f_regression
```

```
from pandas.plotting import scatter_matrix
from statsmodels.graphics.tsaplots import plot_acf

import warnings

warnings.filterwarnings('ignore')

import datetime



# Next, we define variables and we extract the data required for
our analysis using pandas datareader.

y_stk = "MSFT"

x_stk1 = "IBM"
x_stk2 = "GOOGL"

x_ccy1 = "DEXJPUS"
x_ccy2 = "DEXUSUK"

x_idx1 = "SP500"
x_idx2 = "DJIA"
x_idx3 = "VIXCLS"

start = datetime.datetime(2008, 12, 31)
end = datetime.datetime(2019, 12, 31)

stk_tickers = [y_stk, x_stk1, x_stk2]
ccy_tickers = [x_ccy1, x_ccy2]
idx_tickers = [x_idx1, x_idx2, x_idx3]
stk_data = web.DataReader(stk_tickers, 'yahoo', start, end)
ccy_data = web.DataReader(ccy_tickers, 'fred', start, end)
idx_data = web.DataReader(idx_tickers, 'fred', start, end)



# Next, we choose to predict the seires using weekly returns
approximated by using 5 business day period returns.

return_period = 5

Y = np.log(stk_data.loc[:, ('Adj Close',
y_stk)]).diff(return_period).shift(-return_period)
Y.name = Y.name[-1] + '_pred'

X1 = np.log(stk_data.loc[:, ('Adj Close', (x_stk2,
x_stk1))]).diff(return_period)
X1.columns = X1.columns.droplevel()
X2 = np.log(ccy_data).diff(return_period)
X3 = np.log(idx_data).diff(return_period)

X4 = pd.concat([np.log(stk_data.loc[:, ('Adj Close',
y_stk)]).diff(i) for i in
                [return_period, return_period * 3, return_period *
6, return_period * 12]], axis=1).dropna()
```

```
X4.columns = [y_stk + '_DT', y_stk + '_3DT', y_stk + '_6DT', y_stk
+ '_12DT']

X = pd.concat([X1, X2, X3, X4], axis=1)

dataset = pd.concat([Y, X], axis=1).dropna().iloc[::return_period,
:]
Y = dataset.loc[:, Y.name]
X = dataset.loc[:, X.columns]

# with the below is possible to visualise the dataset created

pd.set_option('precision', 3)
dataset.describe()

dataset.head()

dataset.tail()



# The interdependence of data is shown through the correlation
matrix…

correlation = dataset.corr()
pyplot.figure(figsize=(15, 15))
pyplot.title('Correlation Matrix')
sns.heatmap(correlation, vmax=1, square=True, annot=True,
            cmap=sns.diverging_palette(271.1, 12.2, s=97, l=60,
as_cmap=True))

# …and the seasonal decomposition of the time series

res = sm.tsa.seasonal_decompose(Y, freq=52)
fig = res.plot()
fig.set_figheight(8)
fig.set_figwidth(15)
pyplot.show()

# Sklearn's SelectKBest function is used to get a value of feature
importance.

bestfeatures = SelectKBest(k=5, score_func=f_regression)
fit = bestfeatures.fit(X, Y)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X.columns)
# concat two dataframes for better visualization
featureScores = pd.concat([dfcolumns, dfscores], axis=1)
featureScores.columns = ['Specs', 'Score']
# naming the dataframe columns
featureScores.nlargest(12, 'Score').set_index('Specs')
# print 12 best features



# Train Test Split and Evaluation Metrics
# We start by splitting our dataset in training and test sets.
```

```
validation_size = 0.3
# test 70/30


# In case the data is not dependent on the time series, then train
and test split should be done based on sequential sample, by
selecting an arbitrary split point in the ordered list of
observations and creating two new datasets.

train_size = int(len(X) * (1 - validation_size))
X_train, X_test = X[0:train_size], X[train_size:len(X)]
Y_train, Y_test = Y[0:train_size], Y[train_size:len(X)]

num_folds = 10
seed = 7

# In order to avoid confusion, and to allow comparison with other
models, we invert the final scores
scoring = 'neg_mean_squared_error'



# Regression and Tree Regression algorithms
models = []
models.append(('LR', LinearRegression()))
models.append(('LASSO', Lasso()))
models.append(('EN', ElasticNet()))
models.append(('KNN', KNeighborsRegressor()))
models.append(('CART', DecisionTreeRegressor()))
models.append(('SVR', SVR()))

# Neural Network algorithms
models.append(('MLP', MLPRegressor()))

# Boosting methods
models.append(('ABR', AdaBoostRegressor()))
models.append(('GBR', GradientBoostingRegressor()))

# Bagging methods
models.append(('RFR', RandomForestRegressor()))
models.append(('ETR', ExtraTreesRegressor()))




# Once we have selected all the models, we loop over each of them.
First, we run the K-fold analysis. Next we run the model on the
entire training and testing dataset.

names = []
kfold_results = []
test_results = []
train_results = []
for name, model in models:
    names.append(name)
```

```
    ## K Fold analysis:

    kfold = KFold(n_splits=num_folds, random_state=seed)
    # converted mean square error to positive. The lower the beter
    cv_results = -1 * cross_val_score(model, X_train, Y_train,
cv=kfold, scoring=scoring)
    kfold_results.append(cv_results)

    # CHECK
    # Full Training period
    res = model.fit(X_train, Y_train)
    train_result = mean_squared_error(res.predict(X_train),
Y_train)
    train_results.append(train_result)

    # Test results
    test_result = mean_squared_error(res.predict(X_test), Y_test)
    test_results.append(test_result)

    msg = "%s: %f (%f) %f %f" % (name, cv_results.mean(),
cv_results.std(), train_result, test_result)
    print(msg)

# K Fold results

fig = pyplot.figure()
fig.suptitle('Algorithm Comparison: Kfold results')
ax = fig.add_subplot(111)
pyplot.boxplot(kfold_results)
ax.set_xticklabels(names)
fig.set_size_inches(15, 8)
pyplot.show()

# compare algorithms
fig = pyplot.figure()
ind = np.arange(len(names))  # the x locations for the groups
width = 0.35  # the width of the bars

fig.suptitle('Algorithm Comparison')
ax = fig.add_subplot(111)
pyplot.bar(ind - width / 2, np.mean(kfold_results, axis=1),
width=width, label='kfold cross validation Error')
pyplot.bar(ind + width / 2, test_results, width=width, label='Test
Error')
fig.set_size_inches(15, 8)
pyplot.legend()
ax.set_xticks(ind)
ax.set_xticklabels(names)
pyplot.show()

print(test_results)

# Time Series Model - ARIMA Model

X_train_ARIMA = X_train.loc[:, [x_stk1, x_stk2, x_ccy1, x_idx1,
x_idx2, x_idx3]]
X_test_ARIMA = X_test.loc[:, [x_stk1, x_stk2, x_ccy1, x_idx1,
x_idx2, x_idx3]]
tr_len = len(X_train_ARIMA)
```

```
te_len = len(X_test_ARIMA)
to_len = len(X)

modelARIMA = ARIMA(endog=Y_train, exog=X_train_ARIMA, order=[1, 0,
0])
model_fit = modelARIMA.fit()

error_Training_ARIMA = mean_squared_error(Y_train,
model_fit.fittedvalues)
predicted = model_fit.predict(start=tr_len - 1, end=to_len - 1,
exog=X_test_ARIMA)[1:]
error_Test_ARIMA = mean_squared_error(Y_test, predicted)
error_Test_ARIMA

seq_len = 2  # Length of the seq for the LSTM

Y_train_LSTM, Y_test_LSTM = np.array(Y_train)[seq_len - 1:],
np.array(Y_test)
X_train_LSTM = np.zeros((X_train.shape[0] + 1 - seq_len, seq_len,
X_train.shape[1]))
X_test_LSTM = np.zeros((X_test.shape[0], seq_len, X.shape[1]))
for i in range(seq_len):
    X_train_LSTM[:, i, :] = np.array(X_train)[i:X_train.shape[0] +
i + 1 - seq_len, :]
    X_test_LSTM[:, i, :] = np.array(X)[X_train.shape[0] + i -
1:X.shape[0] + i + 1 - seq_len, :]


# Lstm Network
def create_LSTMmodel(neurons=12, learn_rate=0.01, momentum=0):
    # create model
    model = Sequential()
    model.add(LSTM(50, input_shape=(X_train_LSTM.shape[1],
X_train_LSTM.shape[2])))
    # More number of cells can be added if needed
    model.add(Dense(1))
    optimizer = SGD(lr=learn_rate, momentum=momentum)
    model.compile(loss='mse', optimizer='adam')
    return model


LSTMModel = create_LSTMmodel(12, learn_rate=0.01, momentum=0)
LSTMModel_fit = LSTMModel.fit(X_train_LSTM, Y_train_LSTM,
validation_data=(X_test_LSTM, Y_test_LSTM), epochs=330,
                              batch_size=72, verbose=0,
shuffle=False)

# Visual plot to check if the error is reducing
pyplot.plot(LSTMModel_fit.history['loss'], label='train')
pyplot.plot(LSTMModel_fit.history['val_loss'], label='test')
pyplot.legend()
pyplot.show()

error_Training_LSTM = mean_squared_error(Y_train_LSTM,
LSTMModel.predict(X_train_LSTM))
predicted = LSTMModel.predict(X_test_LSTM)
error_Test_LSTM = mean_squared_error(Y_test, predicted)

test_results.append(error_Test_ARIMA)
```

```
test_results.append(error_Test_LSTM)

train_results.append(error_Training_ARIMA)
train_results.append(error_Training_LSTM)

names.append("ARIMA")
names.append("LSTM")

# compare algorithms
fig = pyplot.figure()

ind = np.arange(len(names))  # the x locations for the groups
width = 0.35  # the width of the bars

fig.suptitle('Comparing the performance of various algorthims on
the Train and Test Dataset')
ax = fig.add_subplot(111)
pyplot.bar(ind - width / 2, train_results, width=width,
label='Train Error')
pyplot.bar(ind + width / 2, test_results, width=width, label='Test
Error')
fig.set_size_inches(15, 8)
pyplot.legend()
ax.set_xticks(ind)
ax.set_xticklabels(names)
pyplot.ylabel('Mean Square Error')
pyplot.show()


# Grid Search for ARIMA Model
# Change p,d and q and check for the best result

# evaluate an ARIMA model for a given order (p,d,q)
# Assuming that the train and Test Data is already defined before
def evaluate_arima_model(arima_order):
    # predicted = list()
    modelARIMA = ARIMA(endog=Y_train, exog=X_train_ARIMA,
order=arima_order)
    model_fit = modelARIMA.fit()
    error = mean_squared_error(Y_train, model_fit.fittedvalues)
    return error


# evaluate combinations of p, d and q values for an ARIMA model
def evaluate_models(p_values, d_values, q_values):
    best_score, best_cfg = float("inf"), None
    for p in p_values:
        for d in d_values:
            for q in q_values:
                order = (p, d, q)
                try:
                    mse = evaluate_arima_model(order)
                    if mse < best_score:
                        best_score, best_cfg = mse, order
                    print('ARIMA%s MSE=%.7f' % (order, mse))
                except:
                    continue
    print('Best ARIMA%s MSE=%.7f' % (best_cfg, best_score))
```

```
# evaluate parameters
p_values = [0, 1, 2]
d_values = range(0, 2)
q_values = range(0, 2)
warnings.filterwarnings("ignore")
evaluate_models(p_values, d_values, q_values)

# prepare model
modelARIMA_tuned = ARIMA(endog=Y_train, exog=X_train_ARIMA,
order=[2, 0, 1])
model_fit_tuned = modelARIMA_tuned.fit()

# estimate accuracy on validation set
predicted_tuned = model_fit.predict(start=tr_len - 1, end=to_len -
1, exog=X_test_ARIMA)[1:]
print(mean_squared_error(Y_test, predicted_tuned))

length = len(predicted_tuned)

five_index = length // 5

first_fifth = list[:five_index]

print(first_fifth)

first_fifth = predicted_tuned[0:11]

print(first_fifth)

Y_first_fifth = Y_test[0:11]

print(Y_first_fifth)

print(predicted_tuned)

# Comparison graph for actual vs. predicted data
predicted_tuned.index = Y_test.index
pyplot.plot(np.exp(Y_test).cumprod(), 'm', label='Actual data')  #
plotting t, a separately
pyplot.plot(np.exp(predicted_tuned).cumprod(), 'b',
label='Predicted data')
pyplot.legend()
fig.set_size_inches(18.5, 10.5, forward=True)
pyplot.show()

# Comparison graph for first 3 months of actual vs. predicted data
first_fifth.index = Y_first_fifth.index
pyplot.plot(np.exp(Y_first_fifth).cumprod(), 'm', label='Actual
data')  # plotting t, a separately
pyplot.plot(np.exp(first_fifth).cumprod(), 'b', label='Predicted
data')
pyplot.legend()
fig.set_size_inches(18.5, 10.5, forward=True)
pyplot.show()
```