UNIVERSITÀ DEGLI STUDI DI GENOVA

SCUOLA POLITECNICA

Dipartimento di Ingegneria Meccanica

Dottorato in Ingegneria Matematica e Simulazione

XXXIV Ciclo del Dottorato

Engineering Innovative Technology For The Financial Industry

Relatore:

Chiar.^{mo} Prof. Ing. Agostino Bruzzone

Allievo:

Giuliano Fabbrini

Ringraziamenti

I ringraziamenti per questo percorso vanno principalmente al Simulatation Team, in particolare al Professor Agostino Bruzzone per avermi supportato, corretto e incentivato ad iniziare questa avventura. Ringrazio anche Marina Massei e Kirill Sinelshchikov collaboratori del Simulation Team per avermi dato la loro disponibilità e il loro aiuto in molti progetti e conferenze.

Sono molto grato a Koby Shemer per avermi dato fiducia e la possibilità di collaborare con la sua azienda Apha Beta, nodo fondamentale per il mio progetto finale.

Tengo a menzionare Saverio Scopelliti, mio responsabile per l'azienda dove lavoro. In questi tre anni mi ha sempre lasciato la giusta libertà per portare avanti il mio dottorato con un ottimale bilanciante fra obbiettivi lavorativi e studio da dedicare al dottorato. Ringrazio Saverio per avermi trasmesso i giusti messaggi.

Marta. Che dire, grazie per la pazienza, per il supporto, per avermi preparato mille volte la cena quando finivo tardi di scrivere la tesi, per avere ascoltato il mio discorso ecc. potrei continuare all'inifinito.

Arianna, Luca, Fabiana vi ringrazio per aver creduto in me e soprattutto per avermi fatto arrivare qua. Il merito è vostro. La mia famiglia è lo specchio della persona che sono.

Marco Gotelli…ti menziono in fondo semplicemente perché non cambierebbe nulla nel menzionarti all'inizio o in mezzo. Questo non è il mio risultato ma il nostro risultato. E dopo? (vedremo…). Principalmente ti ringrazio per l'amicizia che abbiamo e i risultati che abbiamo ottenuto sempre e costantemente insieme in tutto il nostro percorso di studi!

INDEX

ABSTRACT

Today innovative technologies allow to develop new performing systems; therefore, also complex in terms of interactions among components and emerging. The innovative technologies are a crucial point for many fields. The filed that we have focused on it is the financial engineering. The financial engineering notion is central to this thesis; as technology advances, the financial sector is becoming increasingly connected to the engineering sector. Financial engineering is a diverse subject of research and practice in which an engineering approach and methodology are applied to the world of finance. Financial engineering functions as a connector of data from several sectors, such as Economics, Mathematics, and IT. It is the application of mathematical concepts to financial problems, as well as the use of tools (due to technological advancements) and expertise from other fields. It used arithmetic to solve current financial problems and create new and innovative financial solutions. Regular commercial banks use financial engineering, which is known as quantity analysis. I looked at two major applications of financial engineering in this paper: machine learning and blockchain. In the financial sector, these two technologies are critical for engineering solutions. We investigated one application of blockchain: the crypto mining server. Machine learning is employed in the development of a specific financial indication (in the case of this work).

1 – INTRODUCTION

The main concept of this thesis is the financial engineering, thanks to the evolution of the technology the financial sector is becoming closed to the engineering sector.

The financial engineering is multidisciplinary field of study and practice that applies an engineering approach and methodology to the world of finance. The financial engineering is like an integrator of information from different fields, like:

- Economics
- Mathematics
- Computer Science

In fact, it is the use of mathematics techniques to solve financial problem, using tools (thanks to the technology) and knowledge from different sector. It applied mathematics to address current financial issues as well as to devise new and innovative financial products.

The financial engineering is referred to as quantities analysis and it is used by regular commercial banks, investment banks, insurance agency etc.

It is a quite new filed study, the first recognized program offering degree in financial engineering is the New York University Polytechnic School of Engineering.

In the field of financial engineering is present also an association, the International Association for Quantitative Finance (IAQF) formerly the International Association of Financial Engineers (IAFE), is a non-profit professional society dedicated to fostering the fields of quantities finance and financial engineering.

The International Association for Quantitative Finance was established in 1992, the IAQF has expanded its reach to host events in San Francisco, Toronto, Boston, and London.

The financial industry is always coming up with new and innovative investment tools and products for investors and companies. Most of the products have been developed through techniques in the fields of financial engineering.

In general, the function of the financial engineering is devoted to test new tools or methods like:

- Investment analysis
- New debt offering
- New investment
- New trading strategies
- New financial models

The financial engineering is used across a board range of tasks in the financial world. Some of the areas where it is most applied are the following:

- Arbitrage trading
- Corporate Finance
- Risk management and analytics
- Technology and algorithms finance
- Behavior finance
- Pricing of options and other financial derivates
- Quotative portfolio management
- Creation of structured financial products and customized financial
- Credit risk and credit management

The financial engineering can be also declined in a various type, like the derivates trading. In this case the financial engineering uses stochastic, simulations and analytics to design and implement new financial processes to solve problems in finance.

Thanks to the effort of the financial engineering many strategies were born, like:

- Married Put is the name given to an options trading strategy where an investor, holding a long position in a stock, purchases an at-the-money put option on the same stock to protect against depreciation in the stock's price.

- Protective Collar. A collar, also known as a hedge wrapper or risk-reversal, is an options strategy implemented to protect against large losses, but it also limits large gains.

- Long Straddle is an options strategy where the trader purchases both a long call and a long put on the same underlying asset with the same expiration date and strike price.

- Short Stingless consists of one short call with a higher strike price and one short put with a lower strike. Both options have the same underlying stock and the same expiration date, but they have different strike prices.

- Butterfly Spreads. The term butterfly spread refers to an options strategy that combines bull and bear spreads with a fixed risk and capped profit. These spreads are intended as a market-neutral strategy and pay off the most if the underlying asset does not move prior to option expiration. They involve either four calls, four puts, or a combination of puts and calls with three strike prices.

Other example of financial engineering application is the stock return system identification, this specific model can identify the system with more than 96% accuracy. Furthermore, one of the indispensable approaches in the financial market, which presents a significant investment role, is forecasting the trends accurately, thanks to the application of a new algorithm, multiple adaptive forecast (MAF), based on adaptive control systems, estimation, and stochastic processes to detect the future trends of prices.[22] (Azarnejad A el all 2021)
Regarding the speculation financial engineering play a crucial role, for example, instruments such as the Credit Default Swap (CDS) were initially created in the late 90s to provide insurance against defaults on bond payments, such as municipal bonds. However, these derivative products drew the attention of investment banks and speculators who realized they could make money from the monthly premium payments associated with CDS by betting with them. In effect, the seller or issuer of a CDS, usually a bank, would receive monthly premium payments from the buyers of the swap. The value of a CDS is based on the survival of a company the swap buyers are betting on the company going bankrupt and the sellers are insuring the buyers against any negative event. If

the company remains in good financial standing, the issuing bank will keep getting paid monthly. If the company goes under, the CDS buyers will cash in on the credit event.

Reassuming the financial engineering is:

- Financial engineering draws on tools from applied mathematics, computer science, statistics, and economic theory.

- Financial engineering is the application of mathematical methods to the solution of problems in finance.

- These businesses apply the methods of financial engineering to such problems as new product development, derivative securities valuation, portfolio structuring, risk management, and scenario simulation.

- Investment banks, commercial banks, hedge funds, fintech companies, insurance companies, corporate treasuries, and regulatory agencies employ financial engineers.

- Financial Engineers, having built a very strong foundation of skills, are also able to succeed in data science, ML/AI, and developer roles.

- As the pace of financial innovation accelerates, the need for highly qualified people with specific training in financial engineering continues to grow in all market environments.

In this work I have considered two main applications of the financial engineering, machine learning and blockchain. These two technologies play a crucial role for the engineering solutions in the financial sector. We studied a specific utilization of blockchain, the crypto mining server. The machine learning is used for building a specific financial indicator (in the case of this work).

Before to start with the description of this work I will introduce another concept that it is the strategic engineering, it is not only a concept, but it is a master in University of Genova that is which inspired me to apply my mathematical and engineering studies and knowledge in finance sector. Strategic Engineering is a new approach, emerging in excellence centers around the world, where new technologies and methods face challenges and uncertainty in complex systems. Strategic engineering is the process of designing and analyzing new solutions to achieve strategic results. It is exactly the approach that has been used to define and complete the work that will be presented in this thesis.

In the specific I have implemented a specific algorithm to implement a crypto mining ring server. In the crypto mining ring server, I studied and implemented an algorithm for the Bitcoin Z crypto value and the possibility to use the crypto mining in a mining pool.

Regarding the machine learning I had a possibility to collaborate with the Alpha Beta company. Thanks to this company I have implemented an algorithmic to predict international stock returns in many markets.

2 – THE BLOCKCHAIN TECHNOLOGY

In this chapter the whole concept of blockchain will be explained. There will be a review of what blockchain is and how it works including some possible bottlenecks to overcome.

2.1 – Blockchain overview

Blockchain is a distributed database. The blockchain stores information eletronicaly in digital format. The databased is ahsred among the nodes of a computer network.

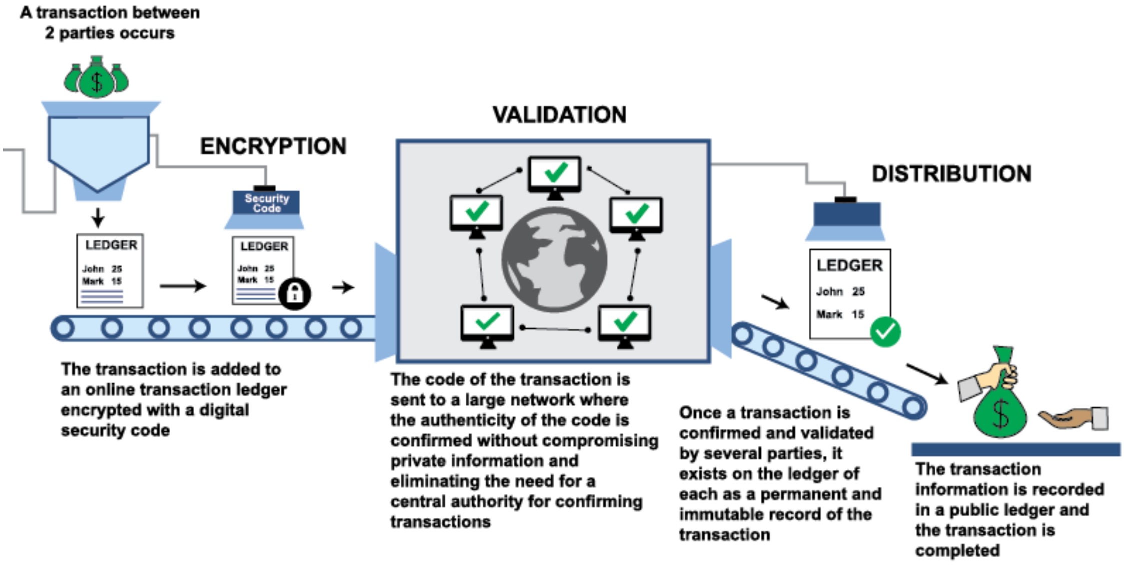

A transaction between 2 parties occurs
ENCRYPTION
Security Code
LEDGER
John 25
Mark 15
The transaction is added to an online transaction ledger encrypted with a digital security code
VALIDATION
The code of the transaction is sent to a large network where the authenticity of the code is confirmed without compromising private information and eliminating the need for a central authority for confirming transactions
DISTRIBUTION
Once a transaction is confirmed and validated by several parties, it exists on the ledger of each as a permanent and immutable record of the transaction
The transaction information is recorded in a public ledger and the transaction is completed

Figure 2.1: Example of a blockchain transaction.

In the specific for the Bitcoin tacnology, every 10 minutes, it is constantly growing by adding new blocks to the chain. Miners do it to record the most recent transactions. All blocks are in the Blockchain in a chronological order. Every node has a copy of the Blockchain that is automatically downloaded when the miner enters the Bitcoin network. All information's about all transactions ever executed is recorded in the blockchain. Blockchain is both the network and database, secure and integrate. Blockchain can build the transactions based on rules defined mathematically and enforced mechanically. The main point is that blockchain does not have one definition because of its various dimensions, including technological, operational, legal, and regulatory. One model of understanding blockchain is through comparing it to the new application layer for internet protocols because blockchain can enable both immediate and long-term economic transactions, and more complicated financial contracts. It can be a layer for transactions of different types of assets, currency, or financial contracts. Moreover, a registry and inventory system for recording, tracking, monitoring, and transacting of all assets could be managed with blockchain.

Consequently, Blockchain can be used for any form of asset, including every area of finance, economics, and money.

2.1.1 - Technical concepts

It is important to consider the technical concepts of blockchain to understand the consequence of the various architectures with respect to regulation, security, performance, and privacy. There is a variety of different technologies based on Blockchain that were developed to solve various problems. Thus, for different needs there are different available technologies.

Generally, Blockchain is a digital platform that keeps the whole history of all transactions between users across the network in a tamper. Also, Blockchain is a database for providing transactions in digital currency such as Bitcoin and Ethereum networks. All transactions that were created between users or counterparties are checked by cryptographic algorithms and then grouped into blocks that are added to Blockchain. No one can change the information in blocks because they are chained to each other. Concerning Bitcoin, every node

in the network has its own copy of Blockchain, synchronized with other nodes using a peer-to-peer protocol. This demonstrates the uselessness of a central authority and consequently leads to confidence of participants in the integrity of any single entity. Blockchain enables to process different transactions and securely reach consensus without third parties.
Fundamental technical concepts of Blockchain technology are the following:

Node: Peer or Node is a computer with the special software that maintains a Blockchain. All nodes are connected to the Blockchain network so they can receive and submit transactions.

Network: It is a result of cooperation of all nodes that run Blockchain software to communicate with each other.

Smart contracts: These are contracts converted into codes to be carried.

Submit transaction: When users submit transactions, they are sent to the nodes on the network who subsequently send them to other nodes.

Transaction Validation: All transactions are cryptographically validated by the nodes on the Blockchain network. Invalid transactions are ignored.

Block: It is a group of transactions collected by nodes into a bundle. To be valid blocks must be formed according to pre-determined set of rules: They must not exceed a maximum size in bytes, contain more than a maximum number of transactions, and must reference to the most recent valid block.
Blockchain: It Is a chain of blocks that is organized by the following system: Each new block is attached to the most recent valid block.

Consensus: It is an agreement of all nodes in the Blockchain. To enable distributed system operation, multiple processes cooperate with each other. Faults in such systems can occur anywhere, that is why they use consensus protocols.

Hash function: It is a one-way function that reflects an input of selectable size to a fixed sized output called hash. Properties of a cryptographic hash function:

1) easy to generate the hash given the input
2) infeasible to generate the original input given the hash
3) virtually impossible for two similar inputs to have the same output in a so called “collision”

2.2 - In the middle of the crisis

Seven weeks after Lehman Brothers filed for the bankruptcy, a paper of in total nine pages appeared on The Cryptography Mailing List under the pseudonym Satoshi Nakamoto (2008). In this paper Nakamoto introduces a completely new concept of a digital currency that could and still can change the entire financial sector and everything related to it. The name of this cryptocurrency is bitcoin, although we can only deduct that from the title since he does not mention it anywhere else in the paper.

The author describes the processing of transactions by financial institutions and therefore the trust in these institutions that is required. By-products are transaction costs, mediating costs and necessaries provisions for ‘accounts receivable’ on the balance sheet (as far as they are not all the same), that are increasing. An electronic payment system without the need of intermediaries seems to be attainable, but the big problem is that of double spending. That is, up until that moment as Nakamoto (2008) found a solution and claims to present a safe and revolutionary system if most of the computer power is in the hands of honest owners, reducing fraud and every kind of transaction cost to a minimum. The system beyond bitcoin is called blockchain, even though Nakamoto nowhere in his paper uses this specific word but he is talking about a 'chain of blocks'.

2.2.1 - Metrics

Blockchain is developing, and many different database technologies and distributed protocols appear. All these technologies are applicable for many different industries and as such require several specifications. The main objective of the development of such technologies is improving blockchains, solving the scalability and throughput capacity of them and ensuring their security, performance, and robustness. These areas are being covered by various types of distributed ledger technologies with varying degrees of decentralization.

The current and past state of the whole network are stored at a blockchain node. Below are qualitative and quantitative metrics that can evaluate the performance of a blockchain architecture.

1. *Submission Throughput*: It is the maximum number of transaction submissions per second possible/ permitted by each node and by the entire network.
2. *Maximum/Average Validation Throughput*: It is the parameter, that determines the maximum/average transaction processing speed of the network.
3. *Average Transaction Validation Latency*: It is the average period that is token to validate the transaction from the time of its submission. This

metric determines the period of waiting of the users for their transaction to be validated and placed in a block. Very important is that the block confirmation and notion of validation could be different in every blockchain.

4. *Latency Volatility*: It is a measure of possible variety of the transaction processing time.
5. Security: Evaluation of the security system requires a threat model, that can define the type and scope of adversaries and attacks on the system. Such threat models could be different in any Blockchain applications. For the security evaluation following analysis are required:
    - Transaction and block immutability
    - Transaction censorship resistance
    - Denial of Service resilience
    - Trust requirement of users and oracles
    - Protocol governance and node membership services
    - Transaction confidentiality and user anonymity

6. Confidentiality: It is the ability of nodes to conceal the contents of the transaction or even the identity as having participated in that transaction from other nodes.
7. Transaction fees: It is the small fee, that users must pay to the network to process transactions or execute smart contracts. These fees cover maintenance costs of the blockchain and provide the protection from frivolous of malicious computational tasks.
8. Hardware requirements:
    - Memory/storage: it is a total memory/storage that is required per node

- Processor: it is an amount of processing resources that are required to validate transactions and blocks.
- Network usage over time, including throughput and latency requirements
- Hardware requirements will change as the network scales.

9. Scalability:
    - Number of nodes: the increase of the number of nodes leads to the change of system performance
    - Number of transactions: the increase of the number of transaction submission per second leads to the change of system performance.
    - Number of users: the increase of the number of active users submitting transactions leads to the change of system performance
    - Geographic dispersion: the increase of the geographic dispersion of nodes leads to the change of system performance
10. *Validation process*: It is an important factor, that is necessary to determine the performance of the network.
11. *Complexity*: it is a measure of the development, maintenance, and operation complexity of Blockchain infrastructure.
12. *Smart-contract limitations*: The main limitations that can influence on the ability of the code deployed on the blockchain are the smart contract scripting language and the underlying consensus protocols.

2.3 - The blockchain protocol

In the article of Nakamoto, the blockchain protocol is mainly explained in the light of the new digital currency bitcoin. As will turn out later this protocol can be used not only for transferring digital money but for all kinds of applications.

Because bitcoin takes place in the digital world, the character of the currency is abstract. Unlike a dollar, there is nothing that can be touched with bitcoin. To that respect it can be compared with transferring dollars or euros at home via internet through a bank. The difference is that bitcoin is not transferred via the bank but directly peer-to-peer and therefore that it is not possible to withdraw bitcoins from a bank, they only exist in a digital form. With blockchain being the underlying layer of bitcoin, one can imagine that explaining this protocol is even a little more difficult and abstract. Others compare it with holding a slippery soap in your hand (Management Events, 2016). Below the protocol will be presented as simple as possible, but not any simpler.

2.3.1 - Requirements

For blockchain, there are a few basic requirements that must be met. These are:

- *Contributors*: A contributor can be anybody with a computer and access to the internet. With multiple contributors the network becomes distributed instead of the usual centralized or decentralized network. This is depicted in Figure 2.2. In a centralized network every kind of communication is transmitted via the central node to the one you want to reach. A modern example is the instant messaging application WhatsApp. Every message that is sent goes through the node of the company of WhatsApp and then reaches the recipient. The decentralized network works the same, but now there are a couple of nodes available through which a message can be delivered. The weaknesses of both types of networks are obvious. If the node crashes for whatever reason either the entire or a part of the system goes down.

The odd one out here is the distributed network, with which blockchain works. Any transaction or any message does not have to go through another node to reach the consignee, it is directly sent from a peer to another peer. Every contributor can be a node on his own. An intermediary is obsolete for this exact reason and a crash of one node has no consequences for the other users.

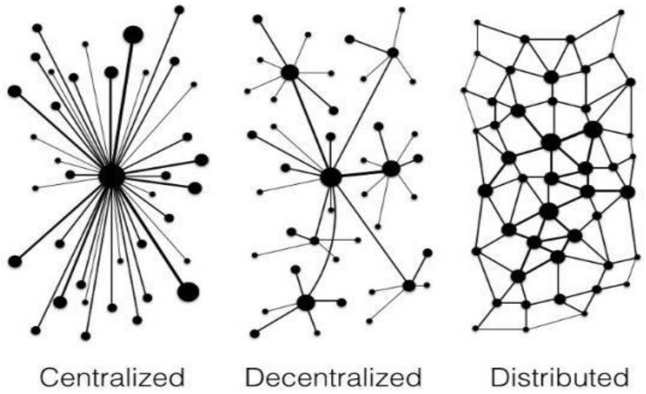


Figure 2.2: Centralized, Decentralized and Distributed networks

- *Ruleset of the system*: new transactions must be approved to be able to execute them. When person A wants to transfer 25 bitcoins to person B, this transaction is put into a block. To implement this block into the existing blockchain, it needs to be mined. The exact details of the mining process will be explained later. Once a block is mined, this is communicated to the other users who can also check whether the mining is done in the correct way and if the transaction is valid. Of course, this can only be the case if person A indeed had 25 bitcoins in his wallet.

  Because new blocks are implemented into the existing blockchain, any discrepancies will immediately become clear, also because it can be revealed who the validators are. So new blocks are always connected to old blocks and new transactions are always connected to old transactions. Once a block is mined and implemented, this cannot be changed anymore. The only way that a block can be altered is by adding a new transaction, discarding fraud with transactions that were already closed in the past.

- *System itself*: The contributors are participating in a network that is distributed. Every dot in the system can do a transaction, be a validator or a miner. In principle every node has got the same rights and obligations. If most of the miner's act on an honest basis, basically no fraud can be committed with the mining. Every node act on itself, so one rotten apple does not spoil the whole barrel. Chances are even high that the rotten apple can be pinpointed and excluded. In this way the network becomes incredibly flexible, and any additional host or facilitator is redundant.

Because of these requirements trust in a middleman or in the transaction partner is not a prerequisite anymore, as committing fraud is almost impossible. Trust is the end of the protocol of blockchain instead of the beginning. In this context a transaction can be interpreted in the broadest way possible. It is not specifically or exclusively related to transferring money.

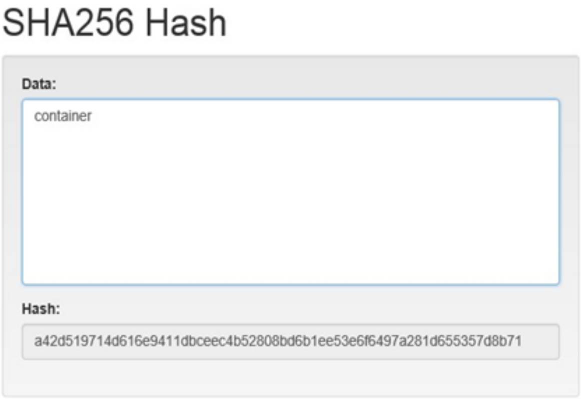
2.3.2 - Hashing

Whenever a letter is written, it can be signed to give some guarantee that this message is indeed coming from a specific person. This can be done with every letter that is sent and the same signature can be used every time. Once somebody can copy the signature, he can pretend that messages are coming from that specific person. To make things more secure different signatures can be used for every day of the week, not only does the forger now needs to imitate in total seven signatures, but it also needs to know on which day which signature is used. This story can be extended to using a different signature for every day of the year or for different messages, making it harder and harder to become a copycat.

Before the digital age coming up with a range of different signatures is a wearisome task. With the current level of technology, digital signatures can be made, and they can even be fabricated for different messages. An example of these signatures is a secure hash algorithm. This is abbreviated with SHA after which a number is put to indicate the output size in bits. A SHA-256 is a secured hash algorithm with an output of 256 bits. SHA-512 has an output that is twice as big and is much more than twice as safe. The secured hash can be compared with guessing a pin. The odds of guessing a pin of 4 numbers correct is 1 out of 10,000. When guessing a pin of eight digits, the chance that somebody is right is only 1 out of 100,000,000. At the other hand it also requires a lot more space and computational power, to which will be come back later. For bitcoin SHA-256 is used and this type of hashing will be taken as the basis for the rest of the explanation.

For hashing the text ‘container’, somebody can go to several hashing or encrypting sites.

This is not only an online signature; it is a digital fingerprint. The code is unique for this message, and not only will it appear every time that the text ‘container’ is typed in into any SHA-256 hashing system, but it also works the other way around. If the hash is put into an unhashing or decrunlashingte, the message ‘container’ appears. To every message - no matter how big or how small - another unique hash gets attached that always has the same length.



SHA256 Hash

Data:

container

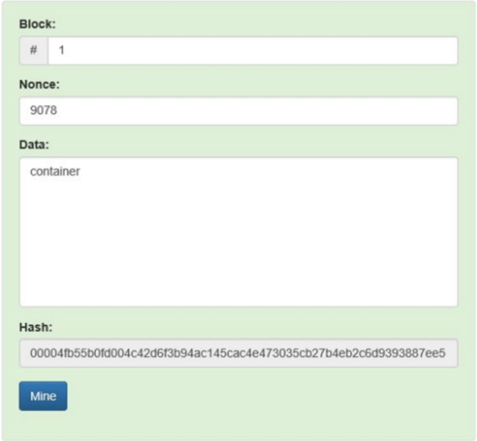
Hash:

a42d519714d616e9411dbceec4b52808bd6b1ee53e6f6497a281d655357d8b71

Figure 2.3: Hashing digital fingerprint (Anders, 2017).

2.3.3 - Blocking

Now that hashing is made clear, we can go one step further into putting the message into a block. Except for the text that can be put in and the associated hash, also several the block in the chain and a nonce appears. A nonce is another number that makes sure that the first four digits of the hash are zero. Things are illustrated in Figure 2.4 and again ‘container’ is the text. Because it is the first block, the number of the block is one and the nonce is constructed in a way that the first four digits of the hash are zero. The hash is different than in figure because also A block number and nonce is added.



Block:

# 1

Nonce:

9078

Data:

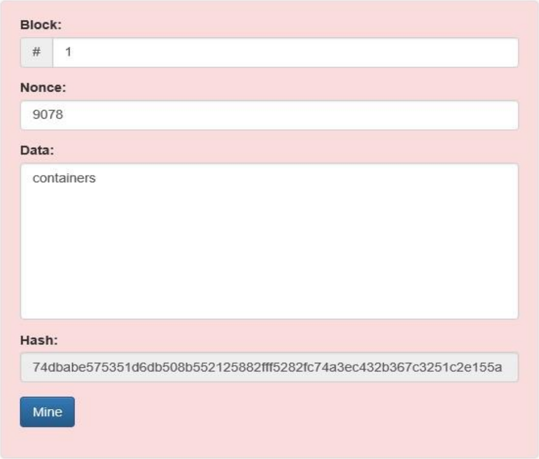
container

Hash:

00004fb55b0fd004c42d6f3b94ac145cac4e473035cb27b4eb2c6d9393887ee5

Figure 2.4: The text 'container' put in a block (Anders, 2017).

If now suddenly, the text is changed to plural, so from ‘container’ to ‘containers’, what happens can be seen in Figure 2.5. The number and the nonce did not change, however the data has indeed changed and so did the hash. On top of that the block turned red, this is an analogy of the figurative red flag. Apparently, something is not right. The hash for these data is correct but with the old nonce the first four digits are not all zero. Some of the readers might already have noticed the button that says ‘mine’. If this button is clicked the computer searches for a nonce that will make the hash starting with four zeros again through trial and error. If it has found one, it will both change the nonce and the hash and the block will turn green again. Of course, both the block number and the data itself will not be edited due to the mining. If all the content is restored to the original information, the block will also turn green again. Mining for a second time is in that case not necessary.



Block: # 1

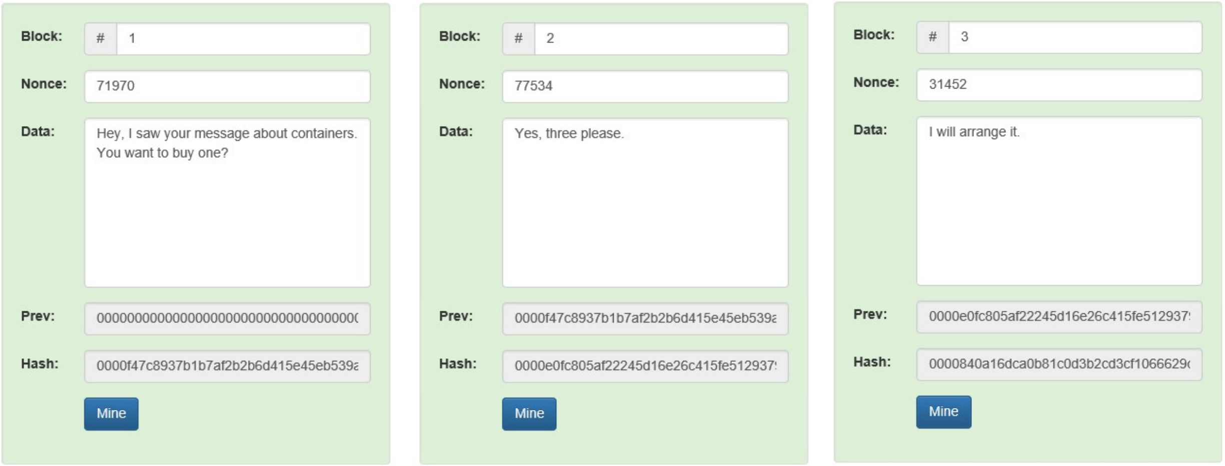
Nonce: 9078

Data: containers

Hash: 74dbabe575351d6db508b552125882fff5282fc74a3ec432b367c3251c2e155a

Figure 2.5: The text ‘container’ of the previous figure changed to plural (Anders, 2017).

Quite intuitively, a blockchain is a chain of the blocks that are mentioned in the previous paragraph. An example of such a chain can be found in Figure 2.5. Again, the same kind of blocks as in Figure 2.4 appear, but now they are linked to each other and another box is added, which is called ‘Prev:’. This stands for previous and with this is meant the hash of the previous block.



Block: # 1

Nonce: 71970

Data: Hey, I saw your message about containers. You want to buy one?

Prev: 0000000000000000000000000000000000

Hash: 0000f47c8937b1b7af2b2b6d415e45eb539

Block: # 2

Nonce: 77534

Data: Yes, three please.

Prev: 0000f47c8937b1b7af2b2b6d415e45eb539

Hash: 0000e0fc805af22245d16e26c415fe512937

Block: # 3

Nonce: 31452

Data: I will arrange it.

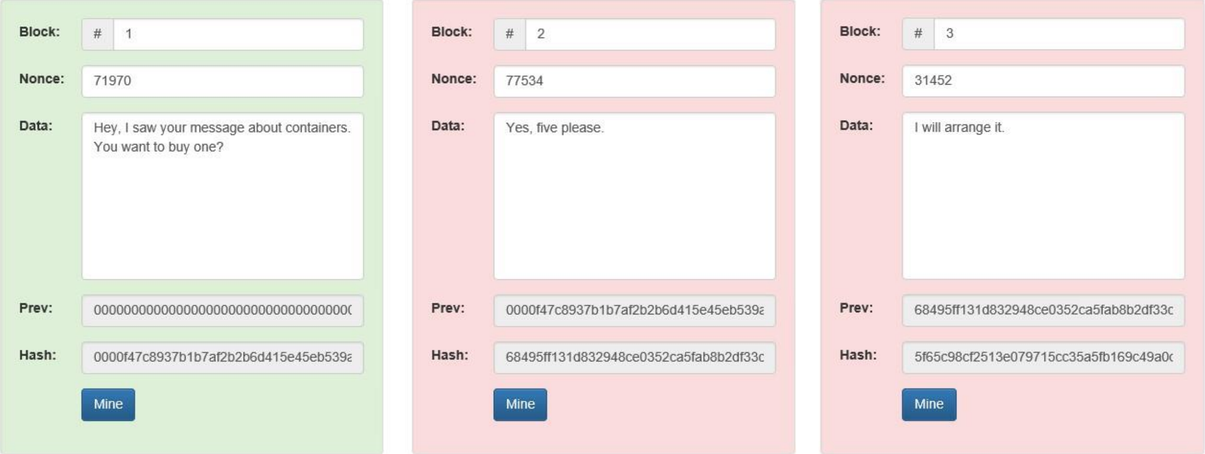
Prev: 0000e0fc805af22245d16e26c415fe512937

Hash: 0000840a16dca0b81c0d3b2cd3cf1066629

Figure 2.6: A conversation between two parties in a blockchain (Anders, 2017).

In the first block the hash of the previous is ‘000000...’ because the first block naturally does not have a previous block or hash. The previous hash of the second block starts with ‘0000f47...’ and indeed this is the hash of the first block. The same goes for the third block.

If suddenly, the content of the second block is changed, the following in Figure 2.7 happens. The current hash of the second block has changed and so did the colour. This is like Figure 2.4 and Figure 2.5, the nonce stays the same but the result is that the hash does not start with four zeros anymore. Although nothing in the content of the third block has been changed, the third block is red as well. This is because the third block is linked to the second block via the hashes, and these do not add up anymore. The chain is broken in block two. Every additional block in the blockchain will therefore turn red.



Block: # 1
Nonce: 71970
Data: Hey, I saw your message about containers. You want to buy one?
Prev: 0000000000000000000000000000000000
Hash: 0000f47c8937b1b7af2b2b6d415e45eb539a

Block: # 2
Nonce: 77534
Data: Yes, five please.
Prev: 0000f47c8937b1b7af2b2b6d415e45eb539a
Hash: 68495ff131d832948ce0352ca5fab8b2df33c

Block: # 3
Nonce: 31452
Data: I will arrange it.
Prev: 68495ff131d832948ce0352ca5fab8b2df33c
Hash: 5f65c98cf2513e079715cc35a5fb169c49a0c

Figure 2.7: Consequences of a change of the content in a blockchain (Anders, 2017).

This problem can be solved by mining all the red blocks again, starting with the first one. In this chain that does not cost too much time or effort, but it can be imagined that when block number two is broken in a chain of 200 blocks, mining everything all alone is impartible.

The content in this chain can indeed be changed and the blocks can be mined again, only two blocks need to be remined. But the fact that only two blocks need to be remined, also proves that this blockchain is accessible for one individual. Things again require more effort and become more complicated when the chain is for example accessible for five people. Imagine first the blockchain of Figure 2.6 times five, so five chains of three blocks, which indicates that there are indeed five participants in this blockchain. If a forger now wants to change the content as in Figure 2.7, it must do that in five chains and then he must remine ten blocks. The moment that either the blocks or the number of participants is of a significant amount, it becomes practically impossible to commit fraud, if most of the computer mining power acts in good faith.

2.3.4 - Mining

In real the mining is a little more difficult than explained in the previous paragraph. The idea is the same, but it is mainly related to the fact that blockchains are introduced on a much bigger scale than presented before. Therefore, the mining is not just a matter of pushing a button and finding a solution after several seconds. The nonce in combination with the hash is a cryptographic, mathematical puzzle. In case of bitcoin this puzzle is difficult, and the solution is hard to find. However, once found the correctness can be checked easily. It can be compared with cracking a safe. It is hard to find the correct numbers, but when somebody pretends to have them, putting them in and checking is easy. Adding a transaction to an existing blockchain will work as follows:

1. New transactions are broadcast to all nodes in the distributed network.
2. Each node collects new transactions into a block.
3. Each node works on finding a difficult so-called proof-of-work (mining with the help of the hash and nonce) for the block.
4. When a node finds the proof-of-work, it broadcasts the block together with the solution to all nodes.
5. Nodes accept the block only if all transactions in it are valid and if the coins were not already spent.
6. Nodes express their acceptance of the block by creating the next block in the chain and using the hash of the accepted block as the previous hash.

When the block indeed is approved, the miner who found the solution will receive 25 bitcoins as a reward, but only after 99 blocks have been added to stimulate the entire mining process. The mining takes place with the help of a lot of computational power. It is possible to be a solo miner, but often miners unite their forces in a mining pool with which they are connected either through installed software or through the cloud.

2.3.5 - Smart contracts

Almost all concrete applications of blockchain in various fields will be discussed in the next chapter. However, the idea of smart contracts will already be explained in this paragraph. With the current system it is possible to make smart contracts that operate based on blockchain. These contracts are smart because the rules of the contract can be translated into computer code, replicated, and executed across the nodes of the blockchain. In this way, the contract can become self-enforcing, monitoring external inputs from trusted sources to settle according to the stipulations of the contract (Peters & Panayi, 2016).

Once the conditions of the contract are put in, it will execute itself as soon as all these conditions get fulfilled. For example, ten barrels of light sweet crude oil will be transferred from A to B now that the dollar to euro exchange rate gets above $1.20 and when the oil price gets below $40 per barrel. When these conditions are met, the transfer will automatically take place. No intervention of any party is necessary anymore. The consequences can be revolutionary.

2.4 - Blockchain bottlenecks

As both the idea of blockchain and bitcoin became bigger and more accepted, some bottlenecks were found on the way. For most of the rising problems a (temporary) solution is found. The most prominent and important matters are discussed below.

2.4.1 - Multiple blockchains

On itself multiple blockchains are no problem at all, and they can exist next to each other. Some small problems do arise when transactions need to be done between different blockchains, for example regarding the transfer of assets. For this the pegged sidechain is invented. Pegged sidechains allow the user to transfer assets between multiple blockchains. Sidechains can easily interoperate, but they remain separate systems. As a result, malicious designs will be confined to that specific sidechain. The sidechain developed by Back e.a. has a two-way peg which means that assets can be transferred between chains and can be returned if necessary.

2.4.2 - Privacy

In a blockchain as presented in Figure 11, every piece of information is fully public. This would also be the case for personal and sensitive data that preferably should be kept private. In the contribution of Zyskind and Nathan (2015) the blockchain is combined with an off-blockchain storage solution. The users of it still do not have to have trust in any third party for protecting their data and at the same time companies do not have to be continuously worrying about securing and compartmentalizing them. Whenever a company wants to use the data, the original owner is being notified.

The Hawk-system of Kosba has the same kind of application, but then specifically focused on financial transactions to retain from the eye of the public how much money is transferred to which person. Just like with the storage solution of Zyskind and Nathan (2015), the system keeps being involved in interactions with the blockchain. Heilman, Baldimtsi and Goldberg (2016) discovered that within the financial transactions although often claimed otherwise bitcoin is not a fully anonymous currency. By using a third party -

that does not have to be trusted - who issues vouchers, this state of fully anonymous transactions can be achieved. The solution to ensure privacy can be desirable for people with both good and less good intentions.

2.4.3 - Speed, size, and security

If blockchain is really going to be introduced on a massive scale, the earlier mentioned mining process seizes a lot of time, maybe even too much. The creation of a new block with bitcoin currently takes around ten minutes. The number of transactions within one block is limited to 500. Therefore, the throughput in the bitcoin network is maximized to 7 transactions per second now. If we compare this to for example VISA, which has 2000 tops, the throughput is way too small and currently there are discussions about how to increase it. This is possible with SegWit, an extension that doubles the size of the blocks. For some of the investors and developers this is not enough; they introduced the alternative 'bitcoin cash' that can handle blocks that are 8 times the size of a current bitcoin block. Earlier, Kraft already presented a model to predict the time it will take to create a new block in the long-term future, given that the mining power either grows with a constant rate or exponentially. Taking Moore's Law into account the latter seems the most plausible but finding an estimator for the block times does not solve the underlying problem for bitcoin. The question also is whether bitcoin cash is the solution and if its bigger blocks are big and safe enough for worldwide scaling as it is only introduced a few weeks ago there still is a lot unclear about this new cryptocurrency.

So, in any case latency can become a big issue here and the same goes for the size and bandwidth. In February 2016 the size of a blockchain in the bitcoin network was over 50,000 MB. If the throughput increases, blockchain could grow with 214 Petabyte each year (which is more than 220 billion MB). This is related to the fact that everybody has got a full copy of the same data; the consequence is a lot of redundancy. Meanwhile, mining bitcoin transactions costs a lot of energy because of all the computing power that is necessary to solve the cryptographic puzzles. This is estimated at 15 million dollars/day. For these problems of speed, size and security, a few systems are in development as well and reviewed below.

2.4.4 - Cryptographic puzzle alternatives

The mining of the new blocks goes together with solving a difficult cryptographic puzzle that cost a lot of time, computer power and energy. The way of mining described with the nonce and hash is how new blocks in for example bitcoin need to be created, which is called proof-of-work. This 'proof' shows that the miner is acting in good faith and deters frivolous and malicious. The only downside of this method is that most of the miners need to act in honest faith. The moment that this is not the case, the system becomes vulnerable for a so called 51%-attack. If 51% of the miners or more can get on the same malicious page, they are able to change most of the information on the blockchain, which then automatically becomes the 'truth'. With the proof-of-work protocol and the increasing size of the blockchains, mining demands more and more computer power. For this reason, people increasingly participate in mining pools. Once that one mining pool possesses more than half of the computing power, things can get somewhat tricky. Currently, with bitcoin, the top-three mining pools in fact own more than 50% of the computer power (Gervais, 2014). With that they do have the power to do a lot of harm. On top of it, because of the increasing required computer performance and steady mining fees it is possible that in the future mining is not lucrative at all anymore. This depends on a lot of parameters such as the development of the IT-sector, Moore's Law and the growth and maturation of blockchain in general.

Especially the method of proof-of-work requires a lot of effort in every meaning of the word, but at the other hand it matches the most safety conditions compared to other proofs (Beigel, 2014). Some alternative proofs will be mentioned below.

2.4.4.1 - Proof-of-stake

In the proof-of-stake the miner does not have to solve a cryptographic puzzle. Proving ownership of a certain amount of cryptocurrency such as bitcoin is sufficient. For example, if some person owns 2 out of the 100 bitcoins that are available, in total this person is allowed to mine 2% of all the possible transactions. The advantage is that everybody is invited to mine a certain amount of the transactions compared to proof-of work because not a lot of computational power is necessary. Moreover, no energy is wasted because ownership is easily shown. Other than that, are proof-of-stake protocols less vulnerable to 51% attacks than proof-of-work protocols. Hou (2014) shows that the vulnerability to a 51%-attack of the proof-of-stake is equal to the proof-of-work protocol if the motivation of the attacker is large enough. Recently, Kiayias e.a. (2016) introduced a proof-of-stake protocol that can provide almost the same security guarantees as the proof-of-work. The authors also present possible different types of attacks and how they are undermined. So, this negative side of the proof-of-stake might have been tackled already.

Another disadvantage is that the person is required to have an amount of cryptocurrency to be able to engage in the mining activities. So, ownership of bitcoin is required to mine bitcoin transactions but next to this cryptocurrency also other cryptocurrencies are starting to be introduced such as Dash, Safecoin and Litecoin. If one needs to possess all these cryptocurrencies to mine transactions of them, the job of mining might become somewhat costly before having even properly started.

2.4.4.2 - Proof-of-activity

This version is a hybrid of both the proof-of-work and the proof-of-stake protocol. It starts with the first protocol where miners are trying to solve the cryptographic puzzle. Afterwards, the blocks that need to be mined are randomly assigned to the miners that were able to solve the puzzle, but the more coins the miner owns, the greater the chances of assignment to that specific miner.

At this point it is not completely clear what the advantages are of this proof-of activity. The authors mention that it offers good security against possible future attacks on for example bitcoin. Intuitively this makes sense since two ‘proofs’ are combined, although except for the 51%-attack there does not seem to be a lot of security issues with the proof-of-work protocol alone. Furthermore, it is understandable that the proof-of activity will not solve the problem of energy waste at all.

2.4.4.3 - Other proofs

Other type of ‘proofs’ are proof-of-burn, the proof-of-capacity and the proof-of-elapsed time. The proof-of-burn requires the burning of coins. The only way to accomplish this is by sending an amount to an address from which they are irretrievable (Coindesk, 2017). The proof-of-capacity works the same as the proof-of-burn. The difference is that now not an amount of cryptocurrency needs to be burned, but an amount of space on the hard disk.

Intel in 2017 developed the proof-of-elapsed-time. This protocol roughly works the same as the proof-of-work, but it consumes far less electricity. The only problem is that this type of proof is only made available by Intel. Again, it would be then a third party in which we must have trust.

2.4.4.4 - Verification and validation of information

In the end the different proofs are only necessary for a miner to show that he is acting in good faith. Rather than just stating that his intentions are sincere it must give evidence that indeed it is, ranging from solving puzzles to giving up cryptocurrency or hard disk space.

Of all the proofs presented here, the proof-of-work still seems to be the best for the job of verification and validation of information, but the proof-of-stake is catching up.

The proof-of-stake is the best alternative and will probably also be introduced on a big scale in the not so far future. Besides, the proof-of-elapsed-time that belongs to Intel looks promising, but the question is how this will develop over time and whether the public really wants to trust a third party again while this is not completely necessary. Of the other mentioned types of proof is not that much information available and thus so far, they do not look like they are going to improve blockchain considerably.

2.4.5 - Legal implications

As Steve Jobs said: technology either needs to be beautiful or invisible (Entrepreneur, 2016). Hopefully at this point blockchain is a little more demystified. It can still be hard to understand how it really works in terms of computer code. This does certainly not imply that both businesses and consumers should stay far away from blockchain in general. How many people really know how internet works? To use it effectively it is not necessary to know exactly how it works and the same goes for blockchain. It is not only a technical, but also a business and management topic.

Nonetheless, like the internet we might have to come up with novel legal mechanisms because of this new set of technology. No parties need to interfere in the transactions of other people anymore, but this also means that supervision and control of authorities can be a lot harder. To that respect Wright and De Filippo speak of the evolution from the Lex Mercatoria, that during medieval times emerged organically from the interactions of merchants analogous to the story in chapter one, to the Lex Informatica, which are contractual agreements between internet service providers and online operators due to the lack of appropriate (inter)national regulation, to the Lex Cryptographic, where again parties have to agree about the set of rules themselves because a higher suitable authority is missing. This authority might have serious issues with regulating blockchain in any way because of for example the application of smart contracts.

Kiviat recognizes these problems, next to his posed questions about jurisdiction, which activities to regulate and to what extent. Now of writing his contribution, the author stated that bitcoin was not only easy to use for the average consumer, but also for the less ordinary part of society considering black-market transactions, tax evasion, money laundering and terrorist financing.

Without a doubt, implementing one-size-fits-all rules for every application with blockchain as a basis is impossible. It must go through trial and error. Law will always be one step behind compared to technology, but this seems to make sense as first the technology needs to be introduced to the public. One needs to know what it is, how it works and where the growing pains are before any guidelines can be made.

Equal to the internet, we should not stay far away of blockchain for this reason. When it is used in practice slowly but steadily the gaps will appear, and the rules will crystallize during time.

2.4.6 - Longlisting problems

When looking at the enumeration of the main problems - the real practical issues will follow in the next chapter - the list looks impressive and infinite. Other questions are how bugs in the system are going to be repaired and if somebody can reverse transactions in case something goes wrong without a middleman or anybody responsible. The answer to the first question is that in bitcoin for example a few administrators exist that can do bitcoin operations like protocol updates and resolution of incidents. These administrators do not need any computer power at all. The answer to the last question is basically no, but transactions can be done in the same but exact opposite way. With blockchain, mutual transactions can be executed when and only when both parties are able to do so. The benefit is that money does not need to be transferred before or after ownership of an asset, it can happen at the same time if that is what parties wish. In case that transactions happen subsequently instead of synchronously and the other party refuses to honour his part of the deal, intuitively they can go to court to legally reverse the contract but in terms of blockchain transfer the asset back to the original owner. This will not yield problems of any significance.

In the end it probably all looks worse than it really is. The most important matter is that of privacy and security and the balance between them. But this is no different than when introducing databases for insurance companies and in the healthcare sector, for Facebook and with the introduction of public surveillance cameras.

Taking fingerprints, making pictures, and storing information about individuals will always be a precarious matter. This is rooted in saving data and it does hold for blockchain. We may need some time to get used to new ideas, doubtlessly mistakes will be made during further development but in the end, it might help with the daily activities in life and businesses.

2.5 - Different views on blockchain

After a meticulous explanation on how blockchain exactly works now some alternative views will be given. As will turn out, considering history this development also completely makes sense.

On closer examination, blockchain is nothing more than a big database (Gustier, 2016). To that extent it can be compared with Google Sheets on the internet; all relevant parties have access to these sheets, the information that is available is real-time and updates can be made by anybody that will be more or less immediately visible to everybody. New rows can be added but only under the already existing rows, to which they must be connected in some way. The rules of the games can be changed, but only from the new row further down. Additionally, every participant has a copy of the sheet, so the moment that there are two individuals with a copy saying one thing and 98 copies indicating another thing it becomes clear quite fast who is right and who is not.

Another way to look at blockchain is that of one big ledger. Where usually in the past there were multiple separate ledgers that were managed by a few administrators, now blockchain offers the opportunity of one big ledger for everybody. This so-called mutual distributed ledger can become a single source of truth in a very cheap and robust way.

Being fraudulent with this big ledger is harder as well. This activity is compared with stealing a cookie; it is easier to steal a cookie our of a jar in a room where often nobody is than when this jar is standing on the middle of a marketplace with a big amount of people passing by every minute. Considering the history that was given in the first chapter from North (1994) that we were going from informal rules in the time that everybody knew each other, to the need of formal institutions because distances between trading partners were increasing and trust could not be taken for granted anymore, now the third paradigm may have arrived. Already for a while, with the internet we try to put these formal institutions online. Platforms like Amazon and eBay were built to facilitate human economic activity (Warburg, 2016). Blockchain can then be the next step within this paradigm of online institutions. While still online we can now suffice with technology alone, without the platforms.

2.6 - Blockchain in a political setting

In the same line as the online institutions as mentioned in the previous paragraph, blockchain can also be considered in a governmental dimension. Atzori (2015) researched the blockchain applications from a political perspective and states that blockchain can be seen as a hyper-political tool with which central authorities can be dismissed as the state was often seen as single point of failure on the long term. This on its turn can lead to a stateless global society as the ultimate stage of globalization.

However, the author recognizes problems and does not think that any authority can be completely absent. The role of the State is a necessary central point of coordination in society that cannot be replaced by algorithms only, but those algorithms can help with the execution of that role. To that respect blockchain and its platforms can better be a pre-political tool instead of a hyper-political tool. By utilizing this information infrastructure governments can become much smarter and a lot less expensive.

2.7 - An (un)limited applicability

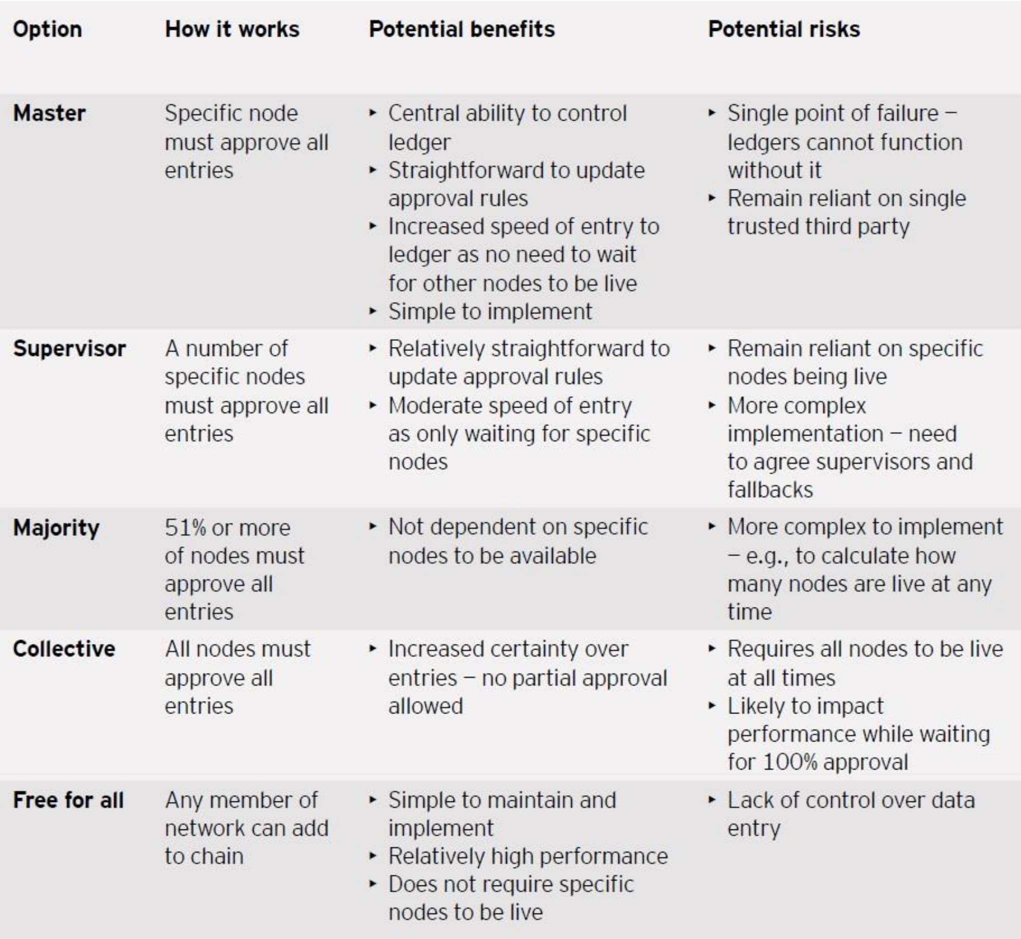
Up until today the real identity of the inventor of blockchain and bitcoin, the person behind Satoshi Nakamoto, is still unknown. In the past, Australian computer scientist and businessman Craig Wright claimed to be him backed up by cryptographic proof and other information, but later it turned out to be a scam. Earlier Dorian Satoshi Nakamoto was wrongfully identified as the inventor. The question also is whether the real Satoshi Nakamoto will ever be revealed; the pseudonym was created for a reason. Not every individual and company will appreciate a revolutionary system that makes a portion of activities superfluous.

The system can come in very handy, but the applicability should also not be overestimated. During my research I came across a video on YouTube, which was a compilation of various videos about bitcoin. In that video people are calling Wall Street a fraud, hosts are yelling that banks are financial terrorists that are here to kill us and kill themselves and a person is crying because that think that bitcoin can save children in the Middle East. Personally, I think this is going a couple of steps too far.

It is not that blockchain is beatific and going to ensure world peace. [4] Swan (2015) suggests an idea of blockchain thinking, in which the brain is acting as a decentralized autonomous organization. Blockchain thinking is proposed as a computational system of input-processing-output, which can be beneficial for both human enhancement and artificial intelligence. Naturally this paper is very forward looking and speculative. It is aimed at the discovery of explorative concepts with no immediate feasibility. Swan (2015) is speaking about mind cloning and multispecies intelligence and although blockchain can help in this development, it will not be the breakthrough that is decisive for the feasibility of those concepts.

Others are sceptic about the general usability of blockchain. Considering the already mentioned barriers of scaling and regulation, blockchain is good at removing third parties but unlikely to offer economic advantages for any commercial problem other than the one for which it was specifically engineered to solve. With bitcoin in existence for many years, no other application than digital cash have been found.

The upside is that some of the negative characteristics of blockchain can be removed by choosing for another architectural structure. The architecture can be created in such a way that a new entry must be approved by one specific person (node) in the system or by all nodes.



Option	How it works	Potential benefits	Potential risks
Master	Specific node must approve all entries	‣ Central ability to control ledger ‣ Straightforward to update approval rules ‣ Increased speed of entry to ledger as no need to wait for other nodes to be live ‣ Simple to implement	‣ Single point of failure – ledgers cannot function without it ‣ Remain reliant on single trusted third party
Supervisor	A number of specific nodes must approve all entries	‣ Relatively straightforward to update approval rules ‣ Moderate speed of entry as only waiting for specific nodes	‣ Remain reliant on specific nodes being live ‣ More complex implementation – need to agree supervisors and fallbacks
Majority	51% or more of nodes must approve all entries	‣ Not dependent on specific nodes to be available	‣ More complex to implement – e.g., to calculate how many nodes are live at any time
Collective	All nodes must approve all entries	‣ Increased certainty over entries – no partial approval allowed	‣ Requires all nodes to be live at all times ‣ Likely to impact performance while waiting for 100% approval
Free for all	Any member of network can add to chain	‣ Simple to maintain and implement ‣ Relatively high performance ‣ Does not require specific nodes to be live	‣ Lack of control over data entry

Table 2.1: Different architectural choices with a blockchain.

The architectural options all have their own strengths and weaknesses. The master option exhibits a lot of similarities with a bank that needs to approve the transactions while the option of free for all looks very accessible and easy, but here there is a risk of disorder.

2.8 Standardization of the system

As blockchain is a relatively recent invention, an innovation race to the bottom is going on regarding all kinds of aspects within this system. The race is characterized:

- *High levels of investment*: During the first years of blockchain, the spending on research and development was low and cautious. Since 2014 investments are getting higher and are not only done by start-ups, but also by large established companies and governments. Businesses are starting to break the mold.
- *Standards race*: Also, in this thesis a lot of improvements and modifications are presented. At the other hand the basic standards are not set either. The most lucrative applications still need to be picked out.
- *An effort to control the directions of the diffusion, in which the problem of a lack of standards so far is rooted*: There is a diversity of developments to be found in which every party hopes to introduce the decisive platform for businesses. Potential market leaders are on the way. Another effect now is fragmentation which threatens a uniform development of blockchain.

The fact that there is a lot of turbulence around blockchain can be a sign. A lot of parties are getting involved in different developments for several reasons and this amount only keeps increasing.

Occasionally, individual enthusiasts of blockchain are coming together in so-called hackathons. There are different kinds of hackathons; with or without specific goals and often sponsored by software companies. Other than that, there are embassies, centres, conferences, workshops, and hackerspaces in relation to blockchain. Embassies provide coordinating information flows for people that are positive about blockchain. Hackerspaces are physical places where the necessary tools are offered to explore technology and to work on projects. The hackathons and hackerspaces are not exclusively focused on blockchain, but also here the interest in this system is growing. One example of a Hackerspace is Pixel bar, which is situated in the west of Rotterdam.

2.9 - Bank versus blockchain

After having discussed blockchain extensively, it can be concluded that this new system can cause a revolution in various industries and in several ways although it must be admitted that it is still in its infancy (Warburg, 2016). So far, banks have executed their tasks quite well but there are a few problems:

- *Banks are centralized and can be hacked*: especially with the apparent rise of hackers this should be taken into serious consideration. Recently hackers are focused on shutting down the supply of electricity and so far, this seems an achievable goal, next to the various ransomware attacks that only seem the beginning of a lot more.
- *Banks exclude billions of people as not everybody has access to a bank that easily*: furthermore, if there is any access, banks take up a relatively big part of the whole. International official remittances to developing countries can take up to 20% of the total amount and the money takes quite a while to get there. Old research showed that in the USA 45% of the GDP was devoted to this transactions sector. The work of Wallis and North is considered as one of the most prominent efforts in quantifying transaction costs.

- *Banks are the only one that have access to the whole picture, and this might undermine our privacy*: at the other hand the consumers only have access to their own account.

Blockchain can make a change in these banking problems. In essence it is nothing more than a big database in which all the information is public, distributed, synchronized, and cryptographically secured. It disintermediates, does not need any trust, and can provide transparency and reliability at low costs. However, new issues arise around privacy, security, and scaling.
The possibilities should not be embellished but in general the benefits outweigh the costs. With a solid blockchain fundament a lot of practical applications - of which bitcoin is often briefly mentioned - can be further developed. What these applications look like and what they can do, will be discussed in the next chapter.

3 – BLOCKCHAIN - IMPLEMENTATION AND APPLICATIONS

3.1 – Examples of Digital Ledger Implementation

Blockchain technology was developed under a digital ledger named Bitcoin by Satoshi Nakamoto and a lot of people think that these two terms are the same. But Blockchain and Bitcoin are totally different. Bitcoin was the first application that utilized blockchain technology, and especially Bitcoin fulfilled the potential of it.

3.1.1 - Bitcoin

Bitcoin is an open source, peer-to-peer crypto-currency that was developed by Satoshi Nakamoto in 2008 and launched in 2009. The system is based on public-private key technology and the decentralized clearing of payments to allow quasi-anonymous transactions. Bitcoin is an independent currency, and it does not belong to any government or legal entity. It is not possible to exchange bitcoins for gold or other object of utility. Adherents of this system argue that Bitcoin has many features which can make it a perfect currency for main merchants and consumers.

For better understanding the functioning of Bitcoin, it has been studied a typical transaction. Both participants in a payment have a private and a public key. To confirm the ownership of a balance of bitcoin the payer needs its private key. To identify the payee, the payer should use payee’s public key, that is open to everyone in the system. For accepting the transaction, the bitcoin software requests all peers on the network to acknowledge the payment is valid. Once the transaction is verified, all other peers are informed that the balance of payer was transferred to the payee. To spend the money, the new owner should repeat this process.

In Bitcoin system there is no need in a central clearing authority. All transactions are grouped together in a block for authenticating that requires the system to solve a complicated cryptography problem. One by one all peers on the network should complete their “proof-of-work “and share it with others by adding the transaction to the blockchain – a place for recording all previous payments and transactions. The fact that all peers can observe the transaction makes it impossible to spend the same balance more than once.

The proof-of-work requires the high level of computing power that leads to high costs. But members of this network can incur such costs, because of the reward for authentication the transactions. The reward is the ability to have their own newly created bitcoin. This decentralized clearing process is called mining.

Bitcoin has several clear strengths. In comparison with fiat currency or precious metals bitcoin could not be confiscated. It also avoids capital controls and disproportionate *taxation*. The one who owns bitcoin can have access to the funds if one can connect to the Internet and keeps a copy of the private keys.

There are *no warehousing costs.* It means that there are no additional costs to storing bitcoins except the initial set up and the properly securing a wallet for Bitcoin users.

Bitcoins are *easy to transport*. Everything that is necessary for logging into the system is private keys. They can be saved in storage media (USB flash driver) or uploaded to the cloud.

The insufficiency is *fixed by an algorithm*. Bitcoin documentation provides that any changes in monetary supply of bitcoins can be made to unanimously consent of all bitcoin-holders, but the fact is, that the resulting currency could not be called Bitcoin, as it is totally different from the original design. Due to this no central authority can decide to debase it. Critics think that the decision about the changes to the money supply of Bitcoin could be done through a majority decision of people who are not monetary experts. In compare with fiat currencies, very often there is a central bank in commission with the keeping relatively stable value for the currency. Bitcoin utilizes *cryptographic security*. In contrast with precious metals that require physical security or fiat currency with the institutional security.

Bitcoin provides *automatic record keeping*. When all payments are recorded in the blockchain, and records are automatically produced.

Bitcoin is *deflationary,* it considers its fixed money supply. Further still, the loss of private keys became more widespread for bitcoin-holders and that leads to a decreasing money supply. In accordance with economical rules and laws it is well-known that deflationary currency has a harmful influence for the economy, because it increases the burden of depts. That are usually denominated in normal terms. Nevertheless, there are some Bitcoin supporters, such as Austrian School economists, who insist that a fixed monetary supply is not necessarily harmful, as deflation would be produced by technological progress.

Even though Bitcoin has several advantages over fiat currencies, it is not without weaknesses. Critics think that this situation will lead to increase the number of cryptocurrencies which will compete, whereas that will end hyperinflation and collapse. This view intends that all cryptocurrencies achieve the same level of acceptance. Supporters of Bitcoin respond that cryptocurrencies are subjects to network effects, in view of infrastructure investment, marketing, mindshare and liquidity. Today Bitcoin holds the superiority on the market, but if other currencies were to replace it in future, the network effect would conduce the leading cryptocurrency and lead to gathering the market around it.

Critics argue that Bitcoin is *volatile*, and it should not be used as a store of value. In compare with fiat currency, Bitcoin does not have any authorities, such as central banks, to assure everyone in the stability of the value. Therefore, the price of bitcoin could have the self-fulfilling dynamics where an incident could blame on itself, becoming a huge confidence crisis.

In Bitcoin system *money supply is not under control*. The amount of money can be changed through the open-source project, through miners and users agreeing to the change.

Critics argue that by holding bitcoins users could not avoid extra wastes related to inflation, because in the case of inflation the difference in prices of fiat and bitcoins will be *taxable*. Proponents of Bitcoin appeal, that it concerns most assets, and in any case, bitcoin holders will protect their money from inflationary increase in the money supply of fiat currency with hedge.

In compare with fiat currencies, cryptocurrencies do not have a *status of legal means of payment*.

Governments may want to *ban cryptocurrencies*. They could prevent illegal uses of cryptocurrencies and enforce currency control. But it is very difficult task, because of distributed structure of cryptocurrencies. However, the ban for exchanges and payment processors could be realized.

Bitcoins have *no physical backing*. Therefore, there is *no intrinsic value* to support them. Bitcoin proponents appeal that gold does not have intrinsic but monetary value too, and furthermore, some supporters of bitcoin argue that the proof-of-work performed by miners is the intrinsic value of it.

The downtrends in the price of bitcoins could be acute because Bitcoin *does not have a marginal cost of production* to stabilize the price. In contrast with commodities, such as gold, the marginal cost of production acts as a support for price levels.

There is *no deposit insurance* for users of Bitcoin in compare with banks. But supporters of Bitcoin reply that there is no need in them, if the security practices, followed by issuers and services, are verified.

Since the supply of bitcoin follows a predetermined trajectory, change in demand of it causes the fluctuations of bitcoin’s purchasing power. In March 2013 there was a fall of exchange rate of bitcoin/dollar, caused by the problems with an updating to the system. On the other hand, such low price pushed the purchasing power of bitcoins up. The speculators were buying bitcoins with the feasible ability to sell them at a higher price. On November 28, 2013, the price has risen to $1132, and after that the price had a decreasing tendency again, perhaps because the speculators were not confident in the future of bitcoin. At the end of May 2014, bitcoin was $631. The existence of demand shocks like these influence on the purchasing power of Bitcoin and makes it unpredictably variable.

Although there have been some network problems, Bitcoin has managed to gain a wide acceptance. The market of Bitcoin is developing and, today, users have many ways to obtain and spend bitcoins. One of the ways, as mentioned above, is mining. Due to it, many of the early holders of Bitcoin acquired their profit. However, not every average user can obtain bitcoins by mining because of the high-level computer technologies that are required. To complete the proof-of-work users need a network of custom-built computers. Nowadays, it has become more common for users to buy bitcoins via an online exchange. There are many exchange services that convert bitcoin to/from a large variety of currencies (USD, EUR, JPY, CAD, GBP, CHF, RUB, AUD, SEK, DKK, HKD, PLN, CNY, SGD, THB, NZD, and NOK), such as Bit Stamp and Coinbase.

Recent research is full of expectations of the possible effects of Bitcoin on the monetary policy of internationally acclaimed currencies. Due to the Quantity Theory of Money, many economists argue that the wide usage of bitcoins could lead to an increase in the velocity of fiat currencies, whereas the necessity of holding them could decrease. Moreover, the result of such an increase in the velocity could be an inflation that will force central banks to decrease the money supply and consequently implement a tightening of the monetary policy. On the other hand, there are some economists who see the future of bitcoin as a positive event for the monetary policy of fiat currencies. For example, economists of the Austrian School view the development of Bitcoin as a return to the Gold Standard.

Finally, it has also been another thought that cryptocurrencies could increase the resilience of the economy. In the case of turmoil or malfunctioning of the existing financial structures, Bitcoin, and cryptocurrencies in general could be useful, as they create an alternative payment system.

3.1.2 - Ethereum

Ethereum is an open Blockchain platform that enables building, executing, and using decentralized applications. It creates applications that automate and facilitate the direct interaction between peers across the network. The same as Bitcoin, Ethereum allows the creation of a payment system without any third-party authorities. Rapid development time, security for applications and the ability of different applications to interact efficiently with each other became very important factors in the context of developing Ethereum. That is why Ethereum began to employ a special programming language – "Turing complete". It enables to create applications that run on the Ethereum system in different programming languages. Ethereum has high level of security, and it relies on a proof-of-work mining. It utilizes Ethereum Virtual Machine "EVM", where smart contract computations are paid for using a cryptocurrency called Ether. Every node of EVM runs such computations to maintain all operations in the blockchain. Due to this process, Ethereum can work with extreme levels of fault tolerance. However, the massive use of synchronized computing across the whole network makes the processes slower.

3.1.3 - IBM Open Blockchain and Hyperledger Fabric

IBM OBC was created on the assumption that blockchain technology would be well regarded with many networks that serve and provide different goals. IBM is a part of the Hyperledger Project, a Linux foundation project. The main objective of this project is to promote blockchain technology by identifying and addressing important characteristics for a cross-industry open standard for distributed ledgers.

The system that is utilized in IBM OBC is self-maintained and does not need any other network requirements. As Ethereum, OBC uses "Turning complete". The Hyperledger Fabric permit many different uses of Blockchain; therefore, it allows the creation of distinct levels of permission. Due to the ability to encrypt the transaction, participants can conceal their identity, transaction patterns and terms of confidential contracts from third parties. The Hyperledger Fabric relies on Byzantine Fault Tolerant algorithm to secure consensus in the network, differing from Bitcoin that utilize proof of-work mining.

3.1.4 - Eris DB / Tender mint

Eris, similarly, to Ethereum, is an open source blockchain platform for building, testing, maintaining, and operating digital applications. The main difference between these two platforms is that Eris DB allows the creation of both permissioned and permission-less blockchains. This platform was meant to be deployable in many distinct environments. Eris DB supports the EVM, thus any smart contract code written for Ethereum can also execute on an Eris DB blockchain. This platform is aimed to permit easier building of digital applications for users.

Moreover, it has developed its own platform, using Tender mint's consensus protocol. The Tender mint project includes an open source BFT consensus protocol implementation for smart contracts.

3.1.5 - R3CEV

R3CEV is a technology firm, aimed to research, develop, and improve the integration of blockchain in the financial sphere and building a financial grade ledger. This firm expects to ask financial institutions and regulatory bodies to be involved with the creation of a distributed ledger standard.

3.1.6 – Blockchain & IoT

Blockchain is a revolutionary paradigm for the whole society and the Internet of Things. Probably it can be named as the enabling currency of machine economy. There were 26 billion devices and 1.9$ trillion economies in the 2020. Consequently, "Internet of Money" should manage the transactions between all these devices, and micropayments could develop into a new layer of the economy. Connections in M2M (machine-to-machine) sphere are growing faster than any others. A machine economy can provide a fast and efficient decentralized system of handling and allocating resources on a machine scale, just like money economy allows to do it on a human scale.

The visual example of M2M micropayments could be the automatic negotiation "between" connected to each other automobiles on the higher-speed highway. If they are in a hurry, micro compensating road peers on a more relaxed schedule. The next example could be drones, especially coordinating personal air delivery by them with a device-to device micropayment network. The agricultural sphere could be developed with blockchain likewise. Their sensors can use economic principles to filter out routine data and fulfil the database with the most relevant, depending on the environmental conditions.

Generally, at the most basic level, blockchain technology's decentralized model of thrustless peer-to-peer transactions means intermediary-free transactions. However, the massive shift to this system on a large-scale global basis could mean a totally different operation of humanity in the spheres that cannot yet be foreseen, but where all that system could easily loose its utility.

3.2 Blockchain Applications

Several applications regarding blockchain technology will be described below.

3.2.1 – Financial Markets

Clearing, trading, and replacing the intermediary:

The settlement of financial assets and the clearing are traditional functions of the banking industry. In the U.S., Canada and Japan, there is a 3-day settlement cycle, and in the EU, Hong Kong and South Korea, this cycle takes two days. This one-day difference can bring many risks related to liquidity and credits. That is why in the U.S. the Federal Reserve pressed all stakeholders to act on increasing end-to-end payment speed. Some argue that blockchain does not only move value, but it also integrates several components of the trading-clearing-settlement value chain in an elegant and efficient way. Therefore, the sphere of clearing and settlement trades is one of the potential applications for blockchain.

Blockchain technology can change the clearing and settlement process by means of decentralization and disintermediation. The use of blockchain could make the settlement cycle reducing time consuming. Moreover, back-office costs could be reduced by using Blockchain technology because all reporting, compliance and collateral management can be handled through it. Also, an important feature in using blockchain is that placed funds will not be allowed to release until each party is satisfied with the actions of the other. It will be useful to add to a transaction a digital signature of a third or even more parties, who play the role in authenticating performance.

However, there are critics who think that Blockchain is always going to be more expensive than a central clearer because the processing job will be done by a multiple of agents, not by one. This will define such clearing service as not cheap.

Payment systems:

Nowadays, all payments are checked and ensured by third party authorities, therefore experts in this industry predict that permissioned blockchains will take a significant part in the payments by 2020. The first bank that decided to introduce Blockchain technology for international payments was Santander UK in June 2016.

Particularly in the U.S., non-depository financial services such as blockchain payment companies have been traditionally regulated. However, there is a chance that the laws that establish licensing and compliance standards for money transmitters may be enhanced if the number of blockchain-based systems increases. Still there are several blockchain-based payment providers that may be subject to money services business (MSB) regulations issued by the Department of the Treasury's Financial Crimes Enforcement Network (FinCEN). On the other hand, the EU has a uniform legal framework for regulating electronic money.

Operational risks in financial markets:

Clearing intermediaries is applicable to a category of regulated entities called financial market infrastructure (FMI). The Federal Reserve states that FMIs include the system operator that settle or record payments, securities, derivatives, or other financial transactions. Thus, FMIs are regulated. Due to the usage of blockchain technology, there is no need for a trusted intermediary which could present operational risks. Consequently, the blockchain system will lead to the automation of trade clearing or of payment system.

Smart contracts:

Initially, the blockchain was developed to improve cryptocurrencies, but entrepreneurs are now developing a new way of using blockchain – smart contract. It is a contract between parties that is coded and uploaded to the blockchain. The smart contract does not rely on the third-party authorities. All processes in dealing with such contracts are automatically controlled. The clauses of a contract are executed after all parties have accomplished their duties. This function removes all ambiguity regarding the execution of contract conditions concerning the existence of external dependencies.

Smart contracts may make the negotiation process and performance of a contract easier and more efficient. Usually, the interface of a smart contract is clear, and it imitates the logic of contractual clauses. The main aim is to secure the contractual processes and reduce the cost related to contracting.

One of the main features of blockchain in smart contract is enabling “thrustless” transactions. This type of transaction defines as validated, monitored and bilaterally enforced transactions over a digital network. Smart contracts can incorporate multiple digital signatures for necessary approval of participants. If the conditions of a smart contract depend upon real world data, systems called “oracles” can be implemented to monitor and verify this data.

Another potential use of smart contracts is in financial transactions. There are various features of smart contracts that make them appropriate for this sphere. For example, margin could be automatically transferred upon margin calls, and if there is counterparty default, the contract could terminate itself. Due to the “custodial functions” of blockchain, recordkeeping, auditing, and custodial functions it will be possible to reduce transactional costs for the parties.

Even if the use of smart contracts may be limited, the use of them may increase the automation of the contractual processes and reduce transactional costs related to them. Smart contract is a replicated form of a real contract in the digital network.

3.2.2 – Other industry applications

Blockchain has the potential to provide not only financial applications but disruptive applications to other industries too.

Real estate industry

Blockchain can be applied to both public and private sectors of the real estate industry. All information concerning land registry records and public records of land ownership can be easily uploaded to blockchain. This opportunity will allow the relevant stakeholders and agencies to have access to the ownership data. This considerably decreases the amount of disputes and the need in the third party, consequently saving time and cost for the consumers.

Information about the private sector, such as residential rental agreements between private counterparties, can be uploaded to blockchain. Also, smart contracts can be a good way of improving the real estate industry. This will regulate the workflow of real estate agencies and save resources and time.

Smart Government:

The instantaneous and simultaneous access to a database that keeps public records is a considerable benefit for the government agencies. A good example could be identity management. However, there is much space for improvement for blockchain in this sphere, but applications that will be the result of that kind of work will propel the whole sphere to the next level. For example, the placing of passports or drivers' licenses on blockchain can enable different agencies to verify identification in real time. The Estonian government is experimenting with such management solutions based on blockchain.

Regulatory and taxation applications are also a good example of implementing blockchain technology. Many banks are working on implementing blockchain-based systems. Consequently, if regulators do the same, they could directly and automatically impose restrictions on the execution of transactions. This fact leads to the reduction of regulatory compliance and auditing costs. Financial transactions can also be taxed automatically since the ledger keeps track of transfer of ownership of assets, as each transaction is visible to the relevant tax agencies. This decreases the overhead and the need for various intermediaries in the process. Foreign Aid is another interesting application for blockchain technology. Foreign aid can be utilized in a more efficient manner using cross-border transfers to reach the targeted zone. This fact gives the possibility to avoid corruption and misuse of funds. Last, but not least, there are voting systems in Smart government. Blockchain technology can help to improve them. By using blockchain, every individual can vote on an anonymized ledger, and all results can be counted and defined without the identity of participants. Due to this, the voting environment overhead will be eliminated.

Artificial intelligence:

Artificial intelligence is a new area for integration with blockchain. Applications based on blockchain in this sphere will have far-reaching implications in the future. Nowadays, smart contracts work on a "narrow intelligence", but they can be programmed to accomplish different tasks according to pre-determined rules and conditions. The development of blockchain will lead to the sophistication of the smart contract's implementation. Integration with artificial intelligence can help nodes on the blockchain to function on their own in a semi-autonomous way.

The results of this integration could be following:

- Negotiations between nodes on the blockchain on asset price discovery.
- Discovering ownership networks of financial assets.
- Blockchain nodes cooperating to optimize household energy consumption within the broader Internet of Things model.

NFT:

Non-fungible token uses cryptocurrency to conduct the transaction using the blockchain technology.

NFT is a digital asset that include artwork, miusic, or in-game assets like avatars. A non-fungible taken is a unique that cannot be exchanged for any other NFT. The cost of NFTs varies across the board.

The main famus NFT are:

- Ozzy Osbourne's CryptoBatz
- Under Armour Stephen Curry Sneaker
- Applebee's Metaverse Mondays
- Snoop Dogg's "Da Dogg Gone Gym"
- Lindsay Lohan's Fursona
- Human One
- Everydays: The First 5000 Days. The selling price was $63.9 million.

3.3 – Limitations

The blockchain industry is still in the early stages of development, and there are many kinds of potential limitations. The classes of limitations are both internal and external including those related to technical issues with the underlying technology, ongoing industry thefts and scandals, public perception, government regulation, and the mainstream adoption of technology.

3.3.1 - Technical Challenges

Several technical challenges related to the blockchain, whether a specific one or the model in general, have been identified. The issues are in clear sight of developers, with different answers to the challenges posited, and avid discussion and coding of potential solutions. Insiders have different degrees of confidence as to whether and how these issues can be overcome to evolve into the next phases of blockchain industry development. Others are building different new and separate blockchains (like Ethereum) or technology that does not use a blockchain (like Ripple). One central challenge with the underlying Bitcoin technology is scaling up from the current maximum limit of 7 transactions per second (the VISA credit card processing network routinely handles 2,000 transactions per second and can accommodate peak volumes of 10,000 transactions per second), especially if there were to be mainstream adoption of Bitcoin. Some of the other issues include increasing the block size, addressing blockchain bloat, countering vulnerability to 51 percent mining attacks, and implementing hard forks (changes that are not backward compatible) to the code, as summarized here:

Throughput: The Bitcoin network has a potential issue with throughput in that it is processing only one transaction per second, with a theoretical current maximum of 7 tops. Core developers maintain that this limit can be raised when it becomes necessary. One way that Bitcoin could handle higher throughput is if each block were bigger, though right now that leads to other issues about size and blockchain bloat. Comparison metrics in other transaction processing networks are VISA (2,000 tps typical; 10,000 tps peak), Twitter (5,000 tps typical; 15,000 tps peak), and advertising networks (>100,000 tps typical).

Latency: Right now, each Bitcoin transaction block takes 10 minutes to process, meaning that it can take at least 10 minutes for your transaction to be confirmed. For sufficient security, you should wait more time and for larger transfer amounts it needs to be even longer, because it must outweigh the cost of a doublespend attack (in which Bitcoins are double spent in a separate transaction before the merchant can confirm their reception in what appears to be the intended transaction). Again, as the comparison metric, VISA takes seconds at most.

Size and bandwidth: The blockchain are 25 GB and grew by 14 GB in the last year. So, it already takes a long time to download (e.g., 1 day). If throughput were to increase by a factor of 2,000 to VISA standards, for example, that would be 1.42 PB/year or 3.9 GB/day. At 150,000 tops, the blockchain would grow by 214 PB/year. The Bitcoin community calls the size problem “bloat,” but that assumes that we want a small blockchain; however, to really scale to mainstream use, the blockchain would need to be big, just more efficiently accessed. This motivates centralization, because it takes resources to run the full node, and only about 7,000 servers worldwide do in fact run full Bitcoin nodes, meaning the Bitcoin daemon (the full Bitcoin node running in the background). It is being discussed whether locations running full nodes should be compensated with rewards. Although 25 GB of data is trivial in many areas of the modern “big data” era and data-intensive science with terabytes of data being the standard, this data can be compressed, whereas the blockchain cannot for security and accessibility reasons. However, perhaps this is an opportunity to innovate new kinds of compression algorithms that would make the blockchain (at much larger future scales) still usable, and storable, while retaining its integrity and accessibility. One innovation to address blockchain bloat and make the data more accessible is APIs, like those from Chain and other vendors, that facilitate automated calls to the full Bitcoin blockchain. Some of the operations are to obtain address balances and balances changes and notify user applications when new transactions or blocks are created on the network. Also, there are web-based block explorers (like https://blockchain.info/), middleware applications allowing partial queries of blockchain data, and frontend customer-facing mobile wallets with greatly streamlined blockchain data. There are some potential security issues with the Bitcoin blockchain. The most worrisome is the possibility of a 51-percent attack, in which one mining entity could grab control of the blockchain and double-spend previously transacted coins into his own account. The issue is the centralization tendency in mining where the competition to record new transaction blocks in the blockchain has meant that only a few large mining pools control most of the transaction recording. At present, the incentive is for them to be good players, and some (like Ghash.io) have stated that they would not take over the network in a 51-percent attack, but the network is insecure. Double-spending might also still be possible in other ways—for example, spoofing users to resend transactions, allowing malicious coders to double-spend coins. Another security issue is that the current cryptography standard that Bitcoin uses, Elliptic Curve Cryptography, might be crack able as early as 2015; however, financial cryptography experts have proposed potential upgrades to address this weakness.

Wasted resources: Mining draws an enormous amount of energy, all of it wasted. The earlier estimate cited was $15 million per day, and other estimates are higher. On one hand, it is the very wastefulness of mining that makes it trustable—those rational agents compete in an otherwise useless proof-of-work effort in hopes of the possibility of reward—but on the other hand, these spent resources have no benefit other than mining.

Usability: The API (Application Programming Interface) for working with Bitcoin (the full node of all code) is far less user-friendly than the current standards of other easy-to-use modern APIs, such as widely used REST APIs.

Versioning, hard forks, multiple chains: Some other technical issues have to do with the infrastructure. One issue is the proliferation of blockchains, and that with so many different blockchains in existence, it could be easy to deploy the resources to launch a 51-percent attack on smaller chains. Another issue is that when chains are split for administrative or versioning purposes, there is no easy way to merge or cross-transact on forked chains.

Another significant technical challenge and requirement is that a full ecosystem of plug-and-play solutions be developed to provide the entire value chain of service delivery. For example, linked to the blockchain there needs to be secure decentralized storage (Maid Safe, Story), messaging, transport, communications protocols, namespace and address management, network administration, and archival. Ideally, the blockchain industry would develop similarly to the cloud-computing model, for which standard infrastructure components, like cloud servers and transport systems, were defined and implemented very quickly at the beginning to allow the industry to focus on the higher level of developing value-added services instead of the core infrastructure. This is particularly important in the blockchain economy due to the sensitive and complicated cryptographic engineering aspects of decentralized networks. The industry is sorting out exactly how much computer network security, cryptography, and mathematics expertise the average blockchain start-up should have—ideally not much if they can rely on a secure infrastructure stack on which this functionality already exists. That way, the blockchain industry's development can be hastened, without every new business having to reinvent the wheel and worry about the fact that its first customer-facing wallet was not multiset (or whatever the current industry standard is, as cryptographic security standards will likely continue to iterate). Some of the partial proposed solutions to the technical issues discussed here are as follows:

Offline wallets to store most coins: Different manner of offline wallets could be used to store the bulk of consumer crypto coins, for example, paper wallets, cold storage, and bit cards.

Dark pools: There could be a more granular value chain such that big crypto-exchanges operate their own internal databases of transactions, and then periodically synchronize a summary of the transactions with the blockchain—an idea borrowed from the banking industry.

Alternative hashing algorithms: Litecoin and other cryptocurrencies use script, which is at least slightly faster than Bitcoin, and other hashing algorithms could be innovated.

Alternatives to proof of work for Byzantine consensus: There are many other consensus models proposed such as proof of stake, hybrids, and variants that have lower latency, require less computational power, waste fewer resources, and improve security for smaller chains. Consensus without mining is another area being explored, such as in Tender mint’s modified version of DLS (the solution to the Byzantine Generals’ Problem by Work, Lynch, and Stockmeyer), with bonded coins belonging to byzantine participants. Another idea for consensus without mining or proof of work is through a consensus algorithm such as Hyperledger’s, which is based on the Practical Byzantine Fault Tolerance algorithm.

Only focus on the most recent or unspent outputs: Many blockchain operations could be based on surface calculations of the most recent or unspent outputs, like how credit card transactions operate. “Thin wallets” operate this way, as opposed to querying a full Bitcoin node, and this is how Bitcoin wallets work on cellular telephones. A related proposal is Kryptonite, which has a “mini-blockchain” abbreviated data scheme.

Blockchain interoperability: To coordinate transactions between blockchains, there are several side chains projects proposed, such as those by Block stream.

Posting bond deposits: The security of proposed alternative consensus mechanisms like Tenements’ DLS protocol (which requires no proof-of-work mining) could be reinforced with structural elements such as requiring miners to post bond deposits to blockchains. This could help resolve the security issue of the “nothing at stake in short time ranges” problem, where malicious players (before having a stake) could potentially fork the blockchain and steal cryptocurrency in a double-spend attack. Bond deposits could be posted to blockchains like Tender mint does, making it costly to fork and possibly improving operability and security.

REST APIs: Essentially secure calls in real time, these could be used in specific cases to help usability. Many blockchain companies provide alternative wallet interfaces that have this kind of functionality, such as Blockchain. Info's numerous wallet APIs.

3.3.2 - Business Model Challenges

Another noted challenge, both functional and technical, is related to business models. At first traditional business models might not seem applicable to Bitcoin since the whole point of decentralized peer-to-peer models is that there are no facilitating intermediaries to take a cut/transaction fee (as in one classical business model). However, there are still many worthwhile revenue-generating products and services to provide in the new blockchain economy. Education and mainstream user-friendly tools are obvious low-hanging fruit (for example, being targeted by Coinbase, Circle Internet Financial, and Apo), as is improving the efficiency of the entire worldwide existing banking and finance infrastructure like Ripple—another almost "no brainer" project, when blockchain principles are understood. Looking ahead, reconfiguring all of business and commerce with smart contracts in the Bitcoin 2.0 era could likely be complicated and difficult to implement, with many opportunities for service providers to offer implementation services, customer education, standard setting, and other value-added facilitations. Some of the many types of business models that have developed with enterprise software and cloud computing might be applicable, too, for the Bitcoin economy—for example, the Red Hat model (fee-based services to implement open-source software), and SaaS, providing Software as a Service, including with customization. One possible job of the future could be smart contract auditor, to confirm that AI smart contracts running on the blockchain are indeed doing as instructed and determining and measuring how the smart contracts have self-rewritten to maximize the issuing agent's utility.

Another significant technical challenge and requirement is that a full ecosystem of plug-and-play solutions be developed to provide the entire value chain of service delivery. For example, linked to the blockchain there needs to be secure decentralized storage, messaging, transport, communications protocols, namespace and address management, network administration, and archival. Ideally, the blockchain industry would develop similarly to the cloud-computing model, for which standard infrastructure components (like cloud servers and transport systems) were defined and implemented very quickly at the beginning to allow the industry to focus on the higher level of developing value-added services instead of the core infrastructure. This is particularly important in the blockchain economy due to the sensitive and complicated cryptographic engineering aspects of decentralized networks. The industry is sorting out exactly how much computer network security, cryptography, and mathematics expertise the average blockchain start-up should have—ideally not much if they can rely on a secure infrastructure stack on which this functionality already exists. That way, the blockchain industry's development can be hastened, without every new business having to reinvent the wheel and worry about the fact that its first customer-facing wallet was not multiset (or whatever the current industry standard is, as cryptographic security standards will likely continue to iterate).

3.3.3 - Scandals and Public Perception

One of the biggest barriers to further Bitcoin adoption is its public perception as a venue for (and possible abettor of) the dark net's money-laundering, drug-related, and other illicit activity—for example, illegal goods online marketplaces such as Silk Road. Bitcoin and the blockchain are themselves neutral, as any technology, and are “dual use”; that is, they can be used for good or evil. Although there are possibilities for malicious use of the blockchain, the potential benefits greatly outweigh the potential downsides. Over time, public perception can change as more individuals themselves have wallets and begin to use Bitcoin. Still, it must be acknowledged that Bitcoin as a pseudonymous enabler can be used to facilitate illegal and malicious activities, and this invites in-kind “Red Queen” responses (context-specific evolutionary arms races) appropriate to the blockchain. Computer virus detection software arose in response to computer viruses; and so far, some features of the same constitutive technologies of Bitcoin (like Tor, a free and open software network) have been deployed back into detecting malicious players.

Another significant barrier to Bitcoin adoption is the ongoing theft, scandals, and scams (like so-called new altcoin “pump and dump” scams that try to bid up new altcoins to quickly profit) in the industry. The collapse of the largest Bitcoin exchange at the time, Tokyo-based MtGox, in March 2014 came to wide public attention. An explanation is still needed for the confusing irony that somehow in the blockchain, the world’s most public transparent ledger, coins can disappear and remain lost months later. The company said it had been hacked, and that the fraud was a result of a problem known as a “transaction malleability bug.” The bug allowed malicious users to double-spend, transferring Bitcoins into their accounts while making MtGox think the transfer had failed and thus repeat the transactions, in effect transferring the value twice. Analysts remain unsure if MtGox was an externally perpetrated hack or an internal embezzlement. The issue is that these kinds of thefts persist.

Blockchain industry models need to solidify and mature such that there are better safeguards in place to stabilize the industry and allow both insiders and outsiders to distinguish between good and bad players. Oversight need not come from outside; congruently decentralized vetting, confirmation, and monitoring systems within the ecosystem could be established. An analogy from citizen science is realizing that oversight functions are still important and reinforce the system by providing checks and balances. In DIYgenomics participant-organized research studies, for example, the oversight function is still fulfilled, but in some cases with a wholly new role relevant to the ecosystem—independent citizen ethicists—as opposed to traditional top-down overseers (in the form of a human-subjects research Institutional Review Board). Other self-regulating industries include movies, video games, and comic books.

There is the possibility that the entire blockchain industry could just collapse (either due to already prognosticated problems or some other factor yet unforeseen).

There is nothing to indicate that a collapse would be impossible. The blockchain economy does have a strong presence, as measured by diverse metrics such as coin market capitalizations, investment in the sector, number of start-ups and people working in the sector, lines of GitHub code committed, and the amount of "newspaper ink" devoted to the sector. Already the blockchain industry is bigger and better established than the previous run at digital currencies (virtual-world currencies like the Second Life Linden dollar). However, despite the progress to date and lofty ideals of Bitcoin, maybe it is still too early for digital currency; maybe all the right safeguards and structures are not yet in place for digital currencies to go fully mainstream (although Apple Pay, more than any other factor, may pave the way to full mainstream acceptance of digital currencies). Apple Pay could quite possibly be enough for the short term. It will be a long time before Bitcoin has the same user-friendly attributes of Apple Pay, such as latency of confirmation time.

3.3.4 - Government Regulation

How government regulation unfolds could be one of the most significant factors and risks in whether the blockchain industry will flourish into a mature financial services industry. In the United States, there could be federal- and state-level legislation; deliberations continue into a second comment period regarding a much-discussed New York Bitlicense. The New York Bitlicense could set the tone for worldwide regulation. On one hand, the Bitcoin industry is concerned about the extremely broad, wide-reaching, and extraterritorial language of the license as currently proposed. The license would encompass anyone doing anything with anyone else's Bitcoins, including basic wallet software (like the QT wallet). However, on the other hand, regulated consumer protections for Bitcoin industry participants, like KYC (know your customer) requirements for money service businesses (MSBs), could hasten the mainstream development of the industry and eradicate consumer worry of the hacking raids that seem to plague the industry.

The deliberations and early rulings of worldwide governments on Bitcoin raise some interesting questions. One issue is the potential practical impossibility of carrying out taxation with current methods. A decentralized peer-to-peer sharing economy of Airbnb 2.0 and Uber 2.0 run on local implementations of OpenBazaar with individuals paying with cryptocurrencies renders traditional taxation structures impossible. The usual tracking and chokehold points to trace the consumption of goods and services might be gone. This has implications both for taxation and for the overall measurement of economic performance such as GDP calculations, which could have the beneficial impact of drawing populaces away from being overly and possibly incorrectly focused on consumption as a wellness metric. Instead, there could be an overhaul of the taxation system to a consumption-based tax on large-ticket visible items such as hard assets (cars, houses). Chokehold points would need to be easily visible for taxation, a “tax on sight” concept. A potential shift from an income tax–based system to a consumption tax–based system could be a significant change for societies.

A second issue that blockchain technology raises about government regulation is the value proposition offered by governments and their business model. Some argue that in the modern era of big data, governments are increasingly unable to keep up with their record-keeping duties of recording and archiving information and making data easily accessible. On this view, governments could become obsolete because they cannot fund themselves the traditional way—by raising taxes. Blockchain technology could potentially help solve both challenges and could at minimum supplement and help governments do their own jobs better, eventually making classes of government-provided services redundant. Recording all a society’s records on the blockchain could obviate the need for entire classes of public service. This view starkly paints governments as becoming redundant with the democratization of government features of the blockchain.

However, just as there might be both centralized and decentralized models to coordinate our activities in the world, there could likely be roles for both traditional government and new forms of blockchain-based government. There might still be a role for traditional centralized governments, but they will need to become economically rationalized, with real value propositions that resonate with constituencies, shrink costs, and demonstrate effectiveness. There could be hybrid governments in the future, like other industries, where automation is the forcing function, and the best “worker” for the job is a human/algorithmic pairing. Perfunctory repetitive tasks are automated with blockchain registries and smart contracts, whereas government employees can move up the value chain.

3.3.5 - Privacy Challenges for Personal Records

There are many issues to be resolved before individuals would feel comfortable storing their personal records in a decentralized manner with a pointer and possibly access via the blockchain. The potential privacy nightmare is that if all your data is online and the secret key is stolen or exposed, you have little recourse. In the current cryptocurrency architecture, there are many scenarios in which this might happen, just as today with personal and corporate passwords being routinely stolen or databases hacked—with broad but shallow consequences; tens of thousands of people deal with a usually minor inconvenience. If a thorough personal record is stolen, the implications could be staggering for an individual: identity theft to the degree that you no longer have your identity at all.

3.3.6 - Overall: Decentralization Trends Likely to Persist

However, despite all the potential limitations with the still nascent blockchain economy, there is virtually no question that Bitcoin is a disruptive force and that its impact will be significant. Even if all the current infrastructure developed by the blockchain industry were to disappear (or fall out of popularity, as virtual worlds have), much of their legacy could persist. The blockchain economy has provided new larger-scale ideas about how to do things. Even if you don't buy into the future of Bitcoin as a stable, long-term cryptocurrency, or blockchain technology as it is currently conceived and developing, there is a very strong case for decentralized models. Decentralization is an idea whose time has come. The Internet is large enough and liquid enough to accommodate decentralized models in new and more pervasive ways than has been possible previously. Centralized models were a good idea at the time, an innovation and revolution in human coordination hundreds of years ago, but now we have a new cultural technology, the Internet, and techniques such as distributed public blockchain ledgers that could facilitate activity to not only include all seven billion people for the first time, but also allow larger-scale, more complicated coordination, and speed our progress toward becoming a truly advanced society. If not the blockchain industry, there would probably be something else, and in fact there probably will be other complements to the blockchain industry anyway. It is just that the blockchain industry is one of the first identifiable large-scale implementations of decentralization models, conceived and executed at a new and more complex level of human activity.

4 BLOCKCHAIN – CASE STUDY

4.1 Introduction

In this chapter will be presented a specific crypto mining ring server that has been development in collaboration with University of Genoa.

As mentioned before, mining (or extraction) is the method used by many crypto systems-currencies in general, to mine, and thus create coins.

The cryptocurrency network stores transactions within data structures called "blocks." Even a block can be added to the block-chain, that is to the huge public database containing all the transactions, it is necessary for a computer to validate it by finding a particular code, which it can only be guessed by math-based cryptographic attempts and algorithms.

Each transaction block, i.e., the set of all transactions that took place over a period that varies from blockchain to blockchain, is suitable for a single node in the network.

The task of the miner, or of the node, is to install the software necessary for the computation of hash functions on his machines; this software can calculate transaction data, to which a random or pseudo-random.

This value, together with the transaction block data, generates an alphanumeric string called “hash.” To calculate the contents of a hash string, the miner requires many tries and calculations, therefore many nonces.

In the calculation process, the hash of the previous block is also added which, together with the data of the transaction block and the nonce, generates the hash of the current block.

The characteristic that makes this calculation complex, but essential to be considered correct by the system, is because the hash must start with a fixed number of zeros.

When the string is validated, the transaction block is made valid as well.

The operation ends with the extraction of the cryptocurrency on which the miner is working and the update of the blockchain ledger, the ledger of transaction blocks.

The whole process allows the system to be extremely secure, as the transaction blocks are linked together by the sharing of the hash; this allows a hypothetical attacker to renounce the alteration of the transactions as he would inevitably change his hash as well.

Every day the network you work for creates a certain number of coin rewards and distributes them to online miners.

To increase the likelihood of receiving them, you need very powerful specify machines with greater computing power. The greater the computing power, the greater the likelihood of mining and receiving cryptocurrencies.

Over time many of these machines have been purchased, increasing the difficulty of the algorithm: this is how Pool mining was born.

More specifically, it is the difficulty distributed by the protocol for confirming a block of the blockchain. Pool mining.

In case the difficulty increases, miners react by starting to compete and increasing the computing power of the systems until the right sequence is found by a block.

The difficulty is calculated for each block, it depends on the cryptocurrency and, when other miners are added to the network, there will be an increase in the generation of new blocks: the difficulty level will be recalculated, and the production of new ones will be slowed down. blocks.

Miners who do not meet the required level of difficulty will be verifiably excluded from the network, banning them for a limited period; also, for this reason, in the various mining pools, there are various network ports where to access them. Each gate is a nailed for nodes of certain bands of computing power.

Returning to the discourse of Pool mining, this consists of a group of miners, each of which transfers a part of the resources of its hardware to solve the algorithms and earn coins. If the miners get rewards, they will be divided according to the transmission power that the members of the pool have given away for the extraction.

4.2 Mining Pool and Protocol – Stratum Protocol

The mining pool effectively reduces the gradualness of the block generation reward, decreasing it more evenly over time.

The Stratum Protocol is essentially the evolution of the "get work" protocol, created to support pooled mining. In the past, with get work, the block header was passed from the server to the client, without any transaction.

The only way to change the block was through the nonce value. The most the client could do was try all nonce values by requesting more work from the server.

The " get work "Remote Procedure Calls (RPC) method was the simplest and most original method, directly constituting a header for the miner.

Since a header contains only a single 4 Byte nonce valid for around 4 giga hash, many modern miners should run dozens or hundreds of " get works "per second. Solo-miners, or network nodes, can still use 'get work' on older versions, but most mining pools today either advise against it or do not allow it to be used.

An improved method is the “getblocktemplate” RPC, which provides the mining software with much more information, such as the information needed to build a transaction by paying for the solo-miner's pool or wallet.

Stratum is a line-based protocol that uses a transition control protocol (TCP) socket, with payload encoded as JSON-RPC messages.

JSON is a standard text-based format for representing structured data based on JavaScript object syntax.

The client simply opens the TCP socket and sends requests to the server in the form of JSON format messages. Each line received by the client is again a valid JSON-RPC fragment, containing the response.

Stratum is an easy protocol to implement and easy to debug, because both sides are talking in a readable format.

In addition, JSON is widely supported on all platforms and current miners already have JSON libraries. Compressing and decompressing the message is simple and convenient.

Stratum is a protocol also tested through penetration testing attacks. In particular, the pool and the miners communicate via the Stratum protocol, to assign jobs and to send results. Famous penetration tester Ruben Recabarren shows different types of possible attacks by creating an "opponent model, for example Eavesdrop attack Eavesdropping is a type of attack in which an attacker tries to passively capture radio signals by decoding, or trying to decode, the transmitted data.

Interfering and modifying the miners' Stratum communications Rather common attack, that of changing for example the payment address of miners or exchange pools.

Analyzing the above attacks, Recabarren worked on two passive attacks called: StrapTap, where the opponent can capture and access all information on Stratum packets, and ISP Logs, where the opponent can only access packet time.

The solution required to achieve a stronger Stratum protocol must consider:

- Security protect yourself from Stratum attacks and be resilient to attacks.

- Efficiency the encryption of all Stratum packages is inefficient and insecure.

- Adaptability of minimal modalities to the Stratum protocol.

Bedrock isa secure and efficient extension of the Stratum protocol. It seeks to prevent adversaries from deducing miners' hash rates and efficiently authenticate Stratum messages. Bedrock has 3 components, each of which addresses several Stratum vulnerabilities.

The first component authenticates and obfuscates work assignment and shares submission messages. The second component protects notices of sharing difficulties. The third component protects the pool from inference on the miner's capabilities.

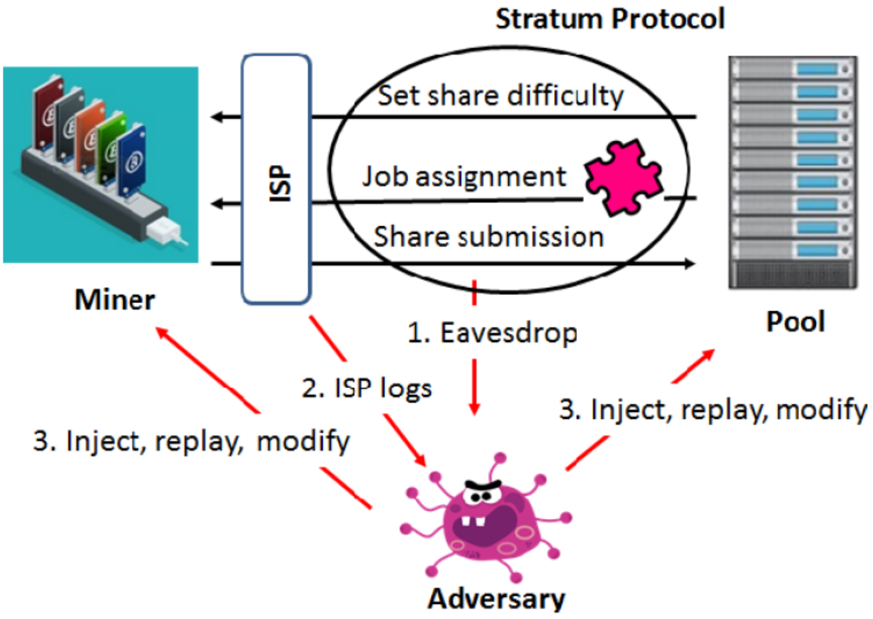


Figure 4.1: Stratum protocol with attack

4.3 BitcoinZ Cryptovalute

4.3.1- Background

BitcoinZ was officially launched by an anonymous developer on September 9 th, 2017 at Block #1 with the followiwing hash:

- BlockHash 0007844681f84249ad7829f9673ea4b6d26a139c741c5847926aff944337d908

The following timeline of events occurred to memorialize and publicly announce the launch:

- GitHub announces at Block #71 with no public pools
- Announced in Zclassic Slak in # general at Block #284 with 15KSol/s network hash
- BitcoinTalk Announcement ('ANN') on September 10th, 2017.

Within four weeks of lunch, the block height broke 15,000 and had e networks hash of 4-5 MSol/s

The founding principles are:

- One-hundred percent decentralized development
- Zclassic spirit, Zcash core, Bitcoin fundamentals
- Always immutable, hardforks only for improvements, changing history is banned
- Fair proposal system
- Everyone is equal and every coin is made by the community and for the community
- Everyone should be able to mine (ASIC resistant)
- No pre-mine, no ICO, no dev taxes

4.3.2 Tech/ Coin Supply

BitcoinZ is a bitcoin compatible cryptocurrency based on the zcash core. It utilizes the Equihash algorithm with t-addresses and zk-snarks anonymous z-addresses.

The combination of these technologies enables BitcoinZ to operate as a proven cryptocurrency with the ability to offer graphics processor unit (GPU) mining to anyone in the world with access to off the shelf graphics cards, also known as commodity hardware. Additionally, it enables the portability and compatibility with BitcoinZ and other cryptocurrency blockchains. Since BitcoinZ has adopted the zcash core, it inherits a ‘halving interval’, adjusted for BitcoinZ’s total supply. The halving interval is set for block 840,000 and is estimated to occur in approximately 4 years from the Sept 9th, 2017 launch.

The approximatre calculation can be copleted by the following steps:

- Blocks Per day = Second in A 24hr Day/ Block Time = 1,440 / 2.5 = 576
- Bloks Per Year = Blocks Per Day x Days In Years = 576 x 365 = 210,240
- Halving Internal in Years = Having Blocks Intrnal / Blocks Per Years = 840,000/ 210,240=3.99

So, in approximately 3.99 years, the block reward will go from 12,500 to 6,250, in accordance with the consensus code Halving Interval set at the 840,000th block.

4.3.3 Decentralized Techniques

In furtherance of the fairness core formula, methods to continue the decentralized nature are
strongly encouraged. Innovative methods of decentralization should always be researched to
keep the integrity and principal vision of “a coin for all”.

- 21 billion coins as total supply to enable every person on the planet to own at least one BTCZ
- Equiphasic PoW algorithm to enable mining with commodity hardware, thereby reducing the barrier to entry for mining.

- Decentralized development by volunteers with no geographical boundaries. Everyone can and is encouraged to participate and contribute to the project, to further progress BitcoinZ as a gift to the world.
- Fair start of the coin by posting in public forum, offering the opportunity to all. No pre-mine, no development fund, all coins to be mined by the community.

4.4 Crypto mining ring server

In this project a cryptomonad ring server has been developer. We had the opportunity to develop the algorithmic part and the hardware part.
Starting from the harder part in the following lines will be explained each part:

Motherboard

The model MSI z170 has been used. We decided to use this model for the material that it is been built. The cryptomining ring server is implemented in a warehouse with a lot of problems like humidity. The Titanium Choke and Dark CAP ensure that our PC runs stable under extreme conditions.

- Titanium Choke: titanium choke uses a titanium core that has better ability in thermal and higher temperature tollerence. This allows the Titanium Choke to turn at a 200-degree Celsius high temperature, have a 40% higher current capacity, a 30% improvement in power efficiency and better overclocking power stability
- Dark cap: with their aluminum core design, Dark Caps has been a staple in high-end motherboard designs and provides lower equivalent series resistance (ESR) as well as its over-10-year lifespan.

Thanks to these materials the following problems have been solved:

- High temperature protection: All key components used in MSI motherboards have all passed military testing ensuring stable operation in the harshest environments.

- Circuit protection: Carefully selected materials, multiple PCB layers and shielding result in the best circuit protection.
- ESD protection: Each I/O port is protected against the hazards of Electrostatic Discharge
- EMI protection: All MSI motherboards comply with strict American FCC regulations and reduce the impact of Electromagnetic Interference.

When the motherboard is used for mining scope the water cooling is necessary, in fact in the server crypto mining, in the following picture is shown the cooling system that we have been implemented.

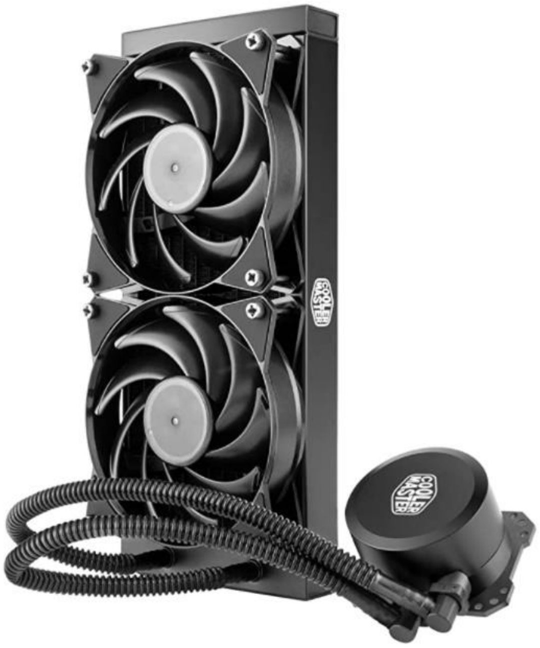
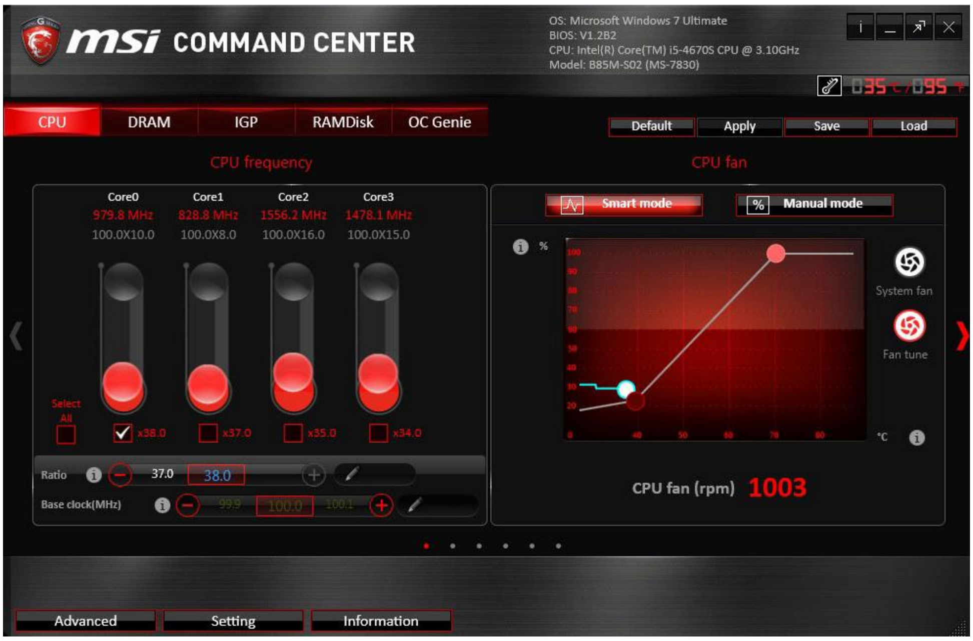


Figure 4.2: Water cooling system

Regarding the software the command center in very intuitive for the end user. It very powerful tool to push the motherboard to the max. Command center allows users to tune settings to increase system stability, maximize overclocking and adjust cooling features.



OS: Microsoft Windows 7 Ultimate
BIOS: V1.2B2
CPU: Intel(R) Core(TM) i5-4670S CPU @ 3.10GHz
Model: B85M-S02 (MS-7830)

Figure 4.3: Motherboard command centre

In the following table is presented the product specifications

CPU (MAX SUPPORT)	i7
SOCKET	1151
CHIPSET	Intel® Z170 Express
DDR4 MEMORY	3600(OC)/ 3200(OC)/ 3000(OC)/ 2800(OC)/ 2600(OC)/ 2400/ 2133 MHz
MEMORY CHANNEL	Dual
DIMM SLOTS	4
MAX MEMORY (GB)	64
PCI-E X16	3
PCI-E GEN	gen-03
SATAIII	6
PCI-E X1	4
M.2 SLOT	2 x 2280 Key M(PCIe Gen3 x4/SATA)
RAID	0/1/5/10
TPM (HEADER)	1
SATA EXPRESS	2

LAN	10/100/1000*1
USB 3.1 PORTS(FRONT)	2
USB 3.1 PORTS (REAR)	4(Gen1), 2(Gen2, Type A+C)
USB 2.0 PORTS (FRONT)	4
USB 2.0 PORTS (REAR)	2
AUDIO PORTS (REAR)	5+Optical SPDIF
HDMI	1
DVI	1
DIRECTX	DX12
FORM FACTOR	ATX
SLI	Y
3-WAY SLI	Y
CROSSFIRE	Y

Table 4.1: Motherboard command centre

CPU

The CPU implanted is the Intel Core I5-6600K. It is part of the Core i5 line-up, using the Skylake architecture with Socket 1151. Core i5-6600K has 6MB of L3 cache and operates at 3.5 GHz by default, but can boost up to 3.9 GHz, depending on the workload. Intel is making the Core i5-6600K on a 14 nm production node, the transistor count is unknown. You may freely adjust the unlocked multiplier on Core i5-6600K, which simplifies overclocking greatly, as you can easily dial in any overclocking frequency.

With a TDP of 95 W, the Core i5-6600K consumes a good deal of power, so decent cooling is needed. Intel's processor supports DDR4 memory with a dual-channel interface. The highest officially supported memory speed is 2133 MHz, but with overclocking (and the right memory modules) you can go even higher. For communication with other components in the machine, Core i5-6600K uses a PCI-Express Gen 3 connection. This processor does not have integrated graphics, you will need a separate graphics card.

Hardware virtualization is available on the Core i5-6600K, which greatly improves virtual machine performance. Programs using Advanced Vector Extensions (AVX) can run on this processor, boosting performance for

calculation-heavy applications. Besides AVX, Intel is including the newer AVX2 standard, too, but not AVX-512.
The performance are:

- Frequency: 3.5 GHz
- Turbo Clock: up to 3.9 GHz
- Base Clock: 100MHz
- Multiplier: 35.0x
- Multiplier Unlocked
- TDP: 95W

This CPU can generate more than 0.96 USD monthly income with a 1544.03 H/s hash rate on the XMR – Random algorithm

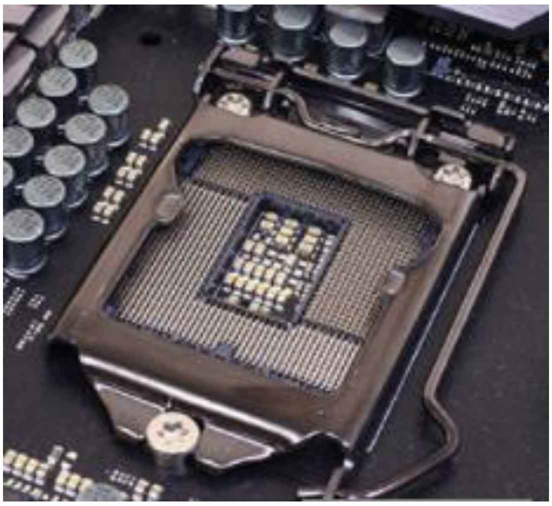


Figure 4.4: intel Core i5-6600K

RAM

8GB 2113 DDR4 with the main following features and advantages:

- Speeds start at 2133 MT/s and faster date rates are expected to be available as DD4 technology matures
- Increase bandwidth by up to 32%
- Reduce power consumption by up to 40%

- Faster burst access speeds for improved sequential data throughput
- Single ranked, x8 based, unbuffered DIMM

The double data rate fourth generation synchronous dynamic random-access memory, (DDR4 SDRAM). DDR4 can achieve higher speed and efficiency thanks to increased transfer rates and decreased voltage.

The chips can support transfer rates between 2133 MT/ and 4266 MT/s. This memory uses little power -- 1.2 Volts. DDR4 doesn't fundamentally change the way memory operates, but it features a new command signal to indicate the active command. AMD has been sampling DDR4 in its latest chip sets.. Under-the-hood technology is the engine that drives computers, and DDR4 is compact piece of muscle.

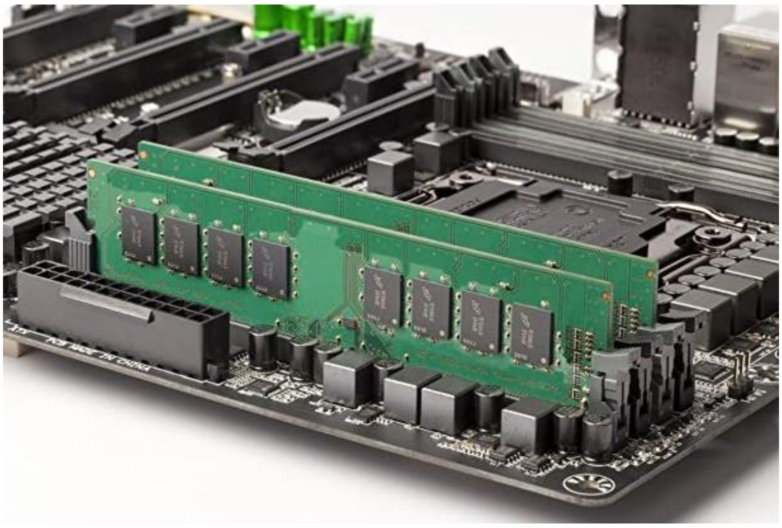


Figure 4.6: 8GB 2113 DDR4

SSD

256 GB Samsung 870 Evo. It is a good balance between performance and price. The 870 EVO also offers an endurance of up to 2,400TBW for the highest capacity model.

Performance-wise, the Samsung 870 EVO is quoted to deliver sequential read and write speeds of 560MB/s and 530MB/s, respectively, while random performance is expected to hit up to 98,000 IOPS read and 88,000 IOPS writes. In the following table is shown the specification

Interface	SATA 6Gbps
Form Factor	2.5-inch
Storage Memory	Samsung V-NAND 3-bit MLC (TLC)
Controller	Samsung MKX Controller
DRAM	4GB LPDDR4 (4TB) 2GB LPDDR4 (2TB) 1GB LPDDR4 (1TB) 512MB LPDDR4 (250/500GB)
Capacity	**4TB**, 2TB, **1TB**, 500GB, 250GB
Sequential Read/Write Speed	Up to 560/530 MB/s
Random Read/Write Speed	Random Read 98K, Write 88K IOPS
Management Software	Samsung Magician
Total Bytes Written	2,400TBW (4TB) 1,200TBW (2TB) 600TBW (1TB) 300TBW (500GB) 150TBW (250GB)

Table 4.2: Motherboard command centre

Power supply unit (PSU)

Two different PSU are implemented:

- PSU1 : Crosair RM850
- PSU2: Itek 700w

We use the following requirements to choose the best power supply unit:

- Capacity 2x higher than you need. Ideally, you’ll run at 50 - 60% capacity.
- 80 PLUS Gold or Cybenetics Gold for higher efficiency.

- Lower than 50mV ripple at +12V under full load at increased operating temperatures (>40°C).
- Quality fan (FDB or Rifle bearing; ideally, it should use ball or magnetic bearings).
- Enough 6+2 pin PCIe connectors to handle the graphics card(s) that you plan to use. You should not use any adapters, in any case! 4-pin Molex connectors are not meant for high loads.
- Enough 4-pin Molex or SATA connectors for the PCIe riser cards that you will use.
- All cables should use 18AWG wires (lower is better) maximum. For PCIe cables, 16-gauge wires are ideal.
- For 1.4kW (better still, 1kW) and stronger PSUs, a C19 coupler is required. An AC power cord with 14AWG wires should be used. For lower-capacity PSUs, an AC power cord with at least 16AWG wires is required.
- Support the essential protection features (SCP, OPP), including over-temperature protection.
- Over 12ms hold-up time (17ms+ is ideal) and an accurate power-good signal. The power-good signal must have at least a 1ms delay, dropping at least 1ms before the rails go out of spec.
- Complete EMI filtering stage (minimum components: 4x Y caps, 2x X caps, two CM chokes, an MOV), along with inrush current protection (an NTC thermistor is required, which ideally should be supported by a bypass relay).
- Impeccable build quality, including quality MOSFETs and high-quality bulk/filtering capacitors (105°C rating and a majority of filtering caps on the secondary side must have >4000h lifetime). The use of polymer caps on the secondary side is preferred.
- 15cm distance between peripheral connectors to allow PCIe risers to be installed further away

Figure 4.7: Motherboard command centre

Graphics Processing Unit (GPU)

Over the years, the mining process and its efficiency have improved with the use of better hardware. Graphics Processing Units (GPU) have been used in the mining process for years, simply because they are more efficient than their immediate counterparts.

Cryptocurrency mining was originally performed using CPUs, or Central Processing Units. However, its limited processing speed and high-power consumption led to limited output, rendering the CPU-based mining process inefficient. Enter GPU-based mining, which offered multiple benefits over the use of CPUs. A standard GPU, like a Radeon HD 5970, clocked processing speeds of executing 3,200 32-bit instructions per clock, which was 800 times more than the speed of a CPU that executed only 4 32-bit instructions per clock. It is this property of the GPU that makes them suitable and better for cryptocurrency mining, as the mining process requires higher efficiency in performing similar kinds of repetitive computations. The mining device continuously tries to decode the different hashes repeatedly with only one digit changing in each attempt. GPUs are also equipped with many Arithmetic Logic Units (ALU), which are responsible for performing mathematical computations. Courtesy of these ALUs, the GPU can perform more calculations, leading to improved output for the crypto mining process.

Each standard computer is equipped with a Central Processing Unit (CPU), which is a processing device that acts as a master of the whole computer system.

It performs the controlling functions for the whole computer based on the logic of the operating system and the software installed on the computer. Typical functions—like save this file as MS Word, print this spreadsheet, or run that video in VLC Media Player—are controlled by the CPU. A GPU is another processing device, but one that works solely for handling display functions. It is the part of a computer that is responsible for its video rendering system.

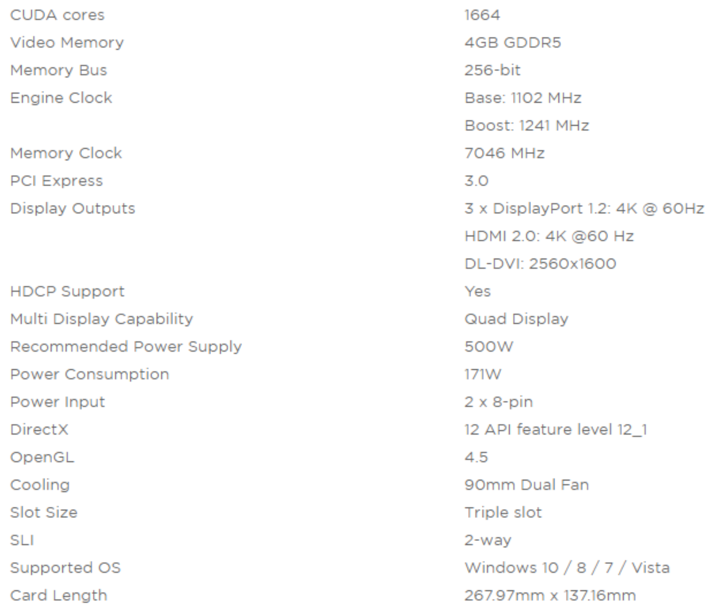
The typical function of a GPU is to perform and control the rendering of visual effects and 3D-graphics, so the CPU doesn't have to get involved in minute details of video-rendering services. It takes care of graphics-intensive tasks such as video editing, gaming display, and decoding and rendering of 3D videos and animations. To draw an analogy, the master (CPU) managing the whole organization (the computer system) has a dedicated employee (GPU) to take care of a specialized department (video-rendering functions). This setup allows the CPU to perform the high-level diversified tasks for managing the whole computer, while the GPU oversees the video functions of which it is a specialist. A CPU will perform the function to open a video file in Windows Media Player, but once the file opens, the GPU takes over the task of displaying it properly.

In our Crypto mining ring server, the following GPU have been used, the main factor that we utilized for the choice are the budget of the project and the permeance of the GPU. In total there are 4 GPU

- GPU 1: 980 pailt
- GPU 2: 970 Strix
- GPU 3 :970 Zotac Gerforce GTX
- GPU 4: 970 Zotac Gerforce GTX

Taking in example one of that (970 Zotac Gerforce GTX) The GeForce® GTX 970 is a high-performance graphics card designed for serious gaming. Powered by new NVIDIA® Maxwell™ architecture, it features advanced technologies and class-leading graphics for incredible gaming experiences. When you demand the absolute best graphic card available, look no further than the ZOTAC GeForce® GTX 970 AMP! Extreme Edition. These graphics cards feature a triple fan IceStorm enhanced cooling system and FireStorm for expert adjustments to fine-tune overclocking with extreme precision.

In the following table the main specifs are presented:



CUDA cores	1664
Video Memory	4GB GDDR5
Memory Bus	256-bit
Engine Clock	Base: 1102 MHz Boost: 1241 MHz
Memory Clock	7046 MHz
PCI Express	3.0
Display Outputs	3 x DisplayPort 1.2: 4K @ 60Hz HDMI 2.0: 4K @60 Hz DL-DVI: 2560x1600
HDCP Support	Yes
Multi Display Capability	Quad Display
Recommended Power Supply	500W
Power Consumption	171W
Power Input	2 x 8-pin
DirectX	12 API feature level 12_1
OpenGL	4.5
Cooling	90mm Dual Fan
Slot Size	Triple slot
SLI	2-way
Supported OS	Windows 10 / 8 / 7 / Vista
Card Length	267.97mm x 137.16mm

Figure 4.9: Zotac Gerforce GTX Specifications

In the following pictures are represented the model of GPU that have been used for this project.

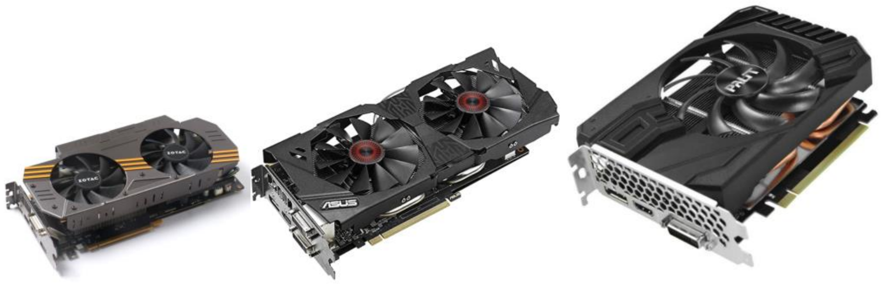


Figure 4.10: GPU

All these components have been implanted in a wooden frame; the frame was built in the laboratory. We chose the wooden material for budget reason.

4.5 Mining Pool

To use our crypto mining ring server, we decided to implement the power of our crypto mining ring server in a specific mining pool system.

A mining pool is a joint group of cryptocurrency miners who combine their computational resources over a network to strengthen the probability of finding a block or otherwise successfully mining for cryptocurrency. Idividually, participants in a mining pool contribute their processing power toward the effort of finding a block. If the pool is successful in these efforts, they receive a reward, typically in the form of the associated cryptocurrency.

Rewards are usually divided between the individuals who contributed, according to the proportion of everyone’s processing power or work relative to the whole group. In some cases, individual miners must show proof of work to receive their rewards.

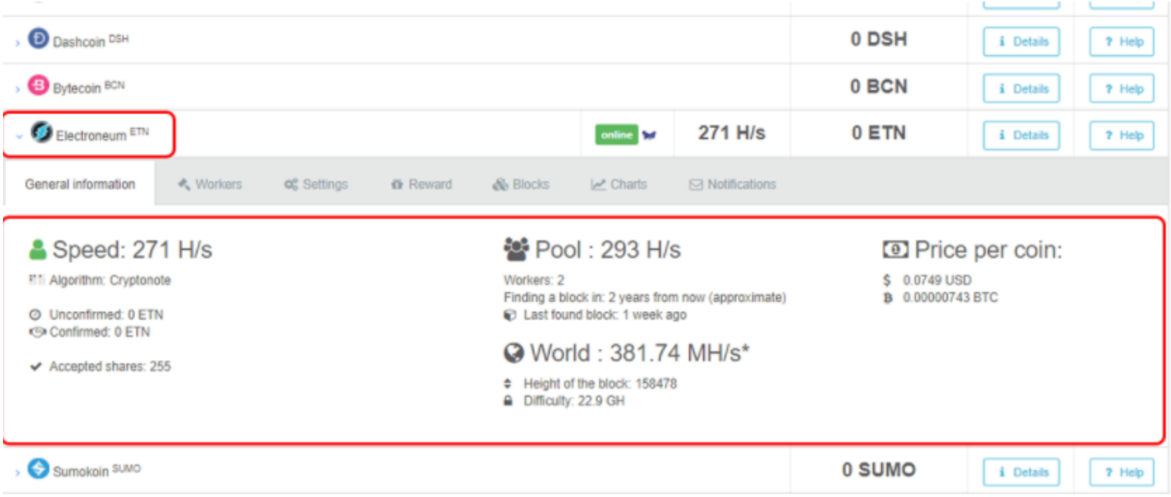
Anyone who wants to make a profit through cryptocurrency mining has the choice to either go solo with their own dedicated devices or to join a mining pool where multiple miners and their devices combine to enhance their hashing output. For example, attaching six mining devices that each offers 335 mega hashes per second (MH/s) can generate a cumulative 2 giga hashes of mining power, thereby leading to faster processing of the hash function.

Not all cryptocurrency mining pools function in the same way. There are, however, several common protocols that govern many of the most popular mining pools. Proportional mining pools are among the most common. In this type of pool, miners contributing to the pool's processing power receive shares up until the point at which the pool succeeds in finding a block. After that, miners receive rewards proportional to the number of shares they hold. Pay-per-share pools operate somewhat similarly in that each miner receives shares for their contribution. However, these pools provide instant payouts regardless of when the block is found. A miner contributing to this type of pool can exchange shares for a proportional payout at any time. Peer-to-peer mining pools, meanwhile, aim to prevent the pool structure from becoming centralized. As such, they integrate a separate blockchain related to the pool itself and designed to prevent the operators of the pool from cheating as well as the pool itself from failing due to a single central issue. While success in individual mining grants complete ownership of the reward, the odd of achieving success is very low because of high power and resource requirements. Mining is often

not a profitable venture for individuals. Many cryptocurrencies have become increasingly difficult to mine in recent years as the popularity of these digital currencies has grown and the costs associated with expensive hardware necessary to be a competitive miner as well as electricity oftentimes outweigh the potential rewards. Mining pools require less of each individual participant in terms of hardware and electricity costs and increase the chances of profitability. Whereas an individual miner might stand little chance of successfully finding a block and receiving a mining reward, teaming up with others dramatically improves the success rate. By taking part in a mining pool, individuals give up some of their autonomy in the mining process. They are typically bound by terms set by the pool itself, which may dictate how the mining process is approached. They are also required to divide up any potential rewards, meaning that the share of profit is lower for an individual participating in a pool. A small number of mining pools, such as AntPool, Poolin, and F2Pool dominate the bitcoin mining process, according to blockchain.com.1 Although many pools do try to be decentralized, these groups consolidate much of the authority to govern the bitcoin protocol. For some cryptocurrency proponents, the presence of a small number of powerful mining pools goes against the decentralized structure inherent in bitcoin and other cryptocurrencies.
In the specific the minepool that we decided to use is Minepool.ch, is a Swiss multicurrency mining pool BitcoinZ (the crypto that we decided to mine).
We started with the installation of the mining pool in the server, with the following steps:

- Download the mining pool from PandaPool wecsite
- Register the account
- Download the Cryptocurrency GUI miner section
- Open the GUI application, and here you have 2 tabs, (this miner can work in 2 modes) In a simple mode everything works automatically (lite), and in a professional you will get more settings (professional version). For this guide we are going to use the lite version, so simply choose a coin, fill the email gap with the same email you had used to create the account on PandaPool and press “Start Mining”

Int the following picture is present the main page of the mining pool:



Dashcoin DSH 0 DSH

Bytecoin BCN 0 BCN

Electroneum ETN online 271 H/s 0 ETN

Speed: 271 H/s
Algorithm: Cryptonote
Unconfirmed: 0 ETN
Confirmed: 0 ETN
Accepted shares: 255

Pool : 293 H/s
Workers: 2
Finding a block in: 2 years from now (approximate)
Last found block: 1 week ago

World : 381.74 MH/s*
Height of the block: 158478
Difficulty: 22.9 GH

Price per coin:
$ 0.0749 USD
Ƀ 0.00000743 BTC

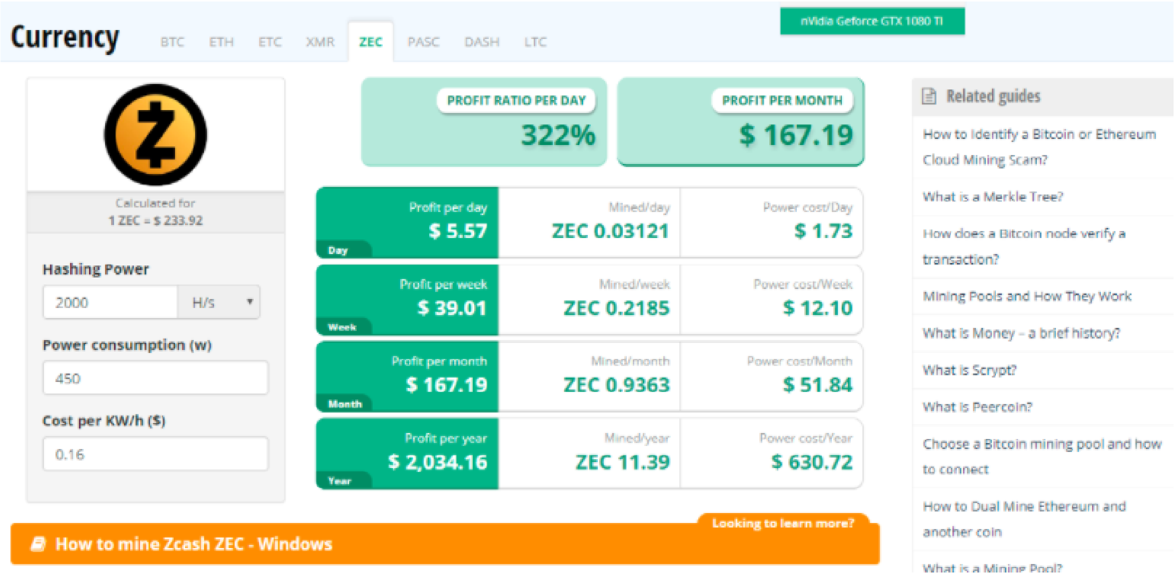
Sumokoin SUMO 0 SUMO

Figure 4.11: Mining Pool

One of the advantages of this mining pool was the possibility to use a Crypto Compare Mining calculator. With this calculator is possible to calculate the cost of the electricity and the gain from the crypto market.
In total there are their gaps:

- Hashing Power: A hash is the output of a hash function, and the Hash Rate is the speed at which a computer is completing an operation in the specific coin algorithm. A higher hash rate is better when mining as it increases your opportunity of finding the next block and receiving the reward. When you run your miner, you can see how many hashes you can do per second.
- Power consumption (w): Is how much power you consume using your machine, any component on your machine will use electricity, so you need to know how to confirm your total consumption, there are many ways to confirm your total power consumption, such as Electricity Usage Monitors (known as Kill-a-watts) and websites like outer vision which do advanced calculus.
- Cost per KW/h ($): Is how much you pay in dollars per kilowatt, the essential thing you need to know to calculate your ongoing profitability is the cost of your electricity. Check with your provider or look at your last electricity bill.

In the following picture in shown how is possible calculate the cost and the gain with the Crypto Compare Mining calculator for a specify GPU



Currency

BTC ETH ETC XMR ZEC PASC DASH LTC

nVidia Geforce GTX 1080 Ti

Calculated for
1 ZEC = $ 233.92

Hashing Power: 2000 H/s

Power consumption (w): 450

Cost per KW/h ($): 0.16

PROFIT RATIO PER DAY: 322%

PROFIT PER MONTH: $ 167.19

Day	Profit per day $ 5.57	Mined/day ZEC 0.03121	Power cost/Day $ 1.73
Week	Profit per week $ 39.01	Mined/week ZEC 0.2185	Power cost/Week $ 12.10
Month	Profit per month $ 167.19	Mined/month ZEC 0.9363	Power cost/Month $ 51.84
Year	Profit per year $ 2,034.16	Mined/year ZEC 11.39	Power cost/Year $ 630.72

Looking to learn more?

How to mine Zcash ZEC - Windows

Related guides

- How to Identify a Bitcoin or Ethereum Cloud Mining Scam?
- What is a Merkle Tree?
- How does a Bitcoin node verify a transaction?
- Mining Pools and How They Work
- What is Money – a brief history?
- What is Scrypt?
- What is Peercoin?
- Choose a Bitcoin mining pool and how to connect
- How to Dual Mine Ethereum and another coin
- What is a Mining Pool?

Figure 4.11: Dashboard of Crypto Compare Mining calculator

5 MACHINE LEARNING TECNOLOGY

Machine learning is a procees of solving a practical problem:

1) Gathering a datset
2) Algorithmically building a statistical model based on that datast

Machine learing ia a branch of computational algorithms with the basic concept to emulate human intelligence.

Machine learing is used to teach machines how to handle the data more effincnetly.

These machines made the human life easy by enabling people to meet various life needs, including travelling, industries, and computing. And Machine learning is the one among them.

Machine learning relies on different algorithms to solve data problems.

The fusion of blockchain tacnology and machine lenaring can bring several advantages. Machine learing is one of the most interesting technologies that makes machines like humans, on the other hands the primary function of blockchain is making secured transactions between participants. Both technology con helps each other.

These technologies have helped replace conventional approaches used in the education sector with highly technical and affective methods.

Thinking to a combination of technology between artificail intelligence, machine learing and blockchain we can refer to a manufacturing application.

In this caseAI can enable comprehensive remote status and performance monitoring of machines across the globe, machine learning-powered proactive maintenance, and the assignment of the most suitable technician with replacement parts, while blockchain can provide a secure way of purchasing.

5.1 Typese of learning

Supervised Learning

In supervised learning the dataset is the collection of the label examples {(xi, yi)} N i=1., each element is called a feature vector. The value is called a feature. The goal of a supervised learning algorithm is to use the dataset to produce a model that takes a feature vectors x as input and outputs information that allow deducind the label for this feature vector. For instance, the model created using the dataset of people could take as input a feature vector describing a person and output a probability that the person has cancer.

Any classification learning algorithm that builds a model implicitly or explicitly creates a decision boundary. The decision boundary can be straight, or curved, or it can have a complex form, or it can be a superposition of some geometrical figures. The form of the decision boundary determines the accuracy of the model (that is the ratio of examples whose labels are predicted correctly). The form of the decision boundary, the way it is algorithmically or mathematically computed based on the training data, dibranchiates one learning algorithm from another. In practice, there are two other essential denunciators of learning algorithms to consider: speed of model building and prediction processing time. In many practical cases, you would prefer a learning algorithm that builds a less accurate model fast. Additionally, you might prefer a less accurate model that is much quicker at making predictions.

Unsupervised Learning

In this case the concept is to create a model that takes a feature vector x as input and either transforms it into another vector or into a value that can be used to solve a practical problems.

Semi Supervised Learning

The goal of a semi- supervised learning algorithm is the as the goal of the supervised learning algorithm bur using many unlabelled examples can help the learning algorithm to find a better model.

Reinforcement Learning

In this case in possible to excite actions. The goal of the reinforcement learning algorithm is to learn a policy. Policy is a function that the takes the feature vector of a state as input ad outputs an optimal action to execute in that state. The decision making is sequential.

5.2 Fundamental Algorithms

I will describe five algorithms that there are use as building block for the most effective learning algorithms

5.2.1 Linear Regression

The idea is to build a model $f_{\mathbf{w},b}(\mathbf{x})$ as a linear combination of feature of example x:

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

Where:

- X : D-dimensional vector
- b : real number
- w: value of the parametrization

The goal is to find a aptiaml values $(\mathbf{w}_{u}, b_{u})$, to do that we have to define to most accurate (regarding prediction) model.

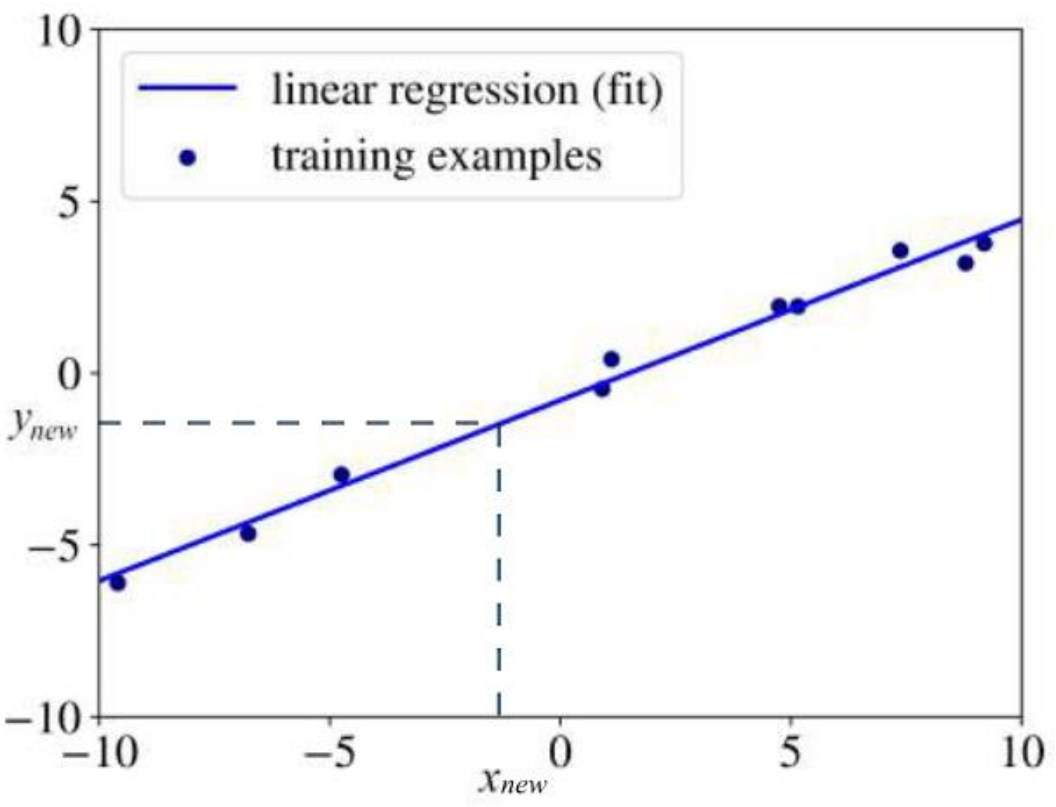


Figure 5.1: Linear Regression for one-dimensional examples

In the Figure 5.1 is shown an example of linear regression, the blue line is the line of the regression, if the blue line was from the blue point, the prediction of new would have fewer chances to be correct.

The goal is to find a solution, starting from the minimization of the following expression:

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

In mathematics, the expression we minimize or maximize is called an objective function, or, simply, an objective. The expression $(f(\mathbf{x}_i) \neq y_i)_2$ in the above objective is called the loss function. It's a measure of penalty for misclassification of example i. This choice of the loss function is called squared error loss. All model-based learning algorithms have a loss function and what we do to find the best model is we try to minimize the objective known as the cost function. In linear regression, the cost function is given by the average loss, also called the empirical risk. The average loss, or empirical risk, for a model, is the average of all penalties obtained by applying the model to the training data.

The choice of the linear from for the model is that it's simple and overfitting. Overfitting is the property of a model such that the model predicts very well labels of the examples used during training but frequently makes errors when applied to examples that weren't seen by the learning algorithm during training.

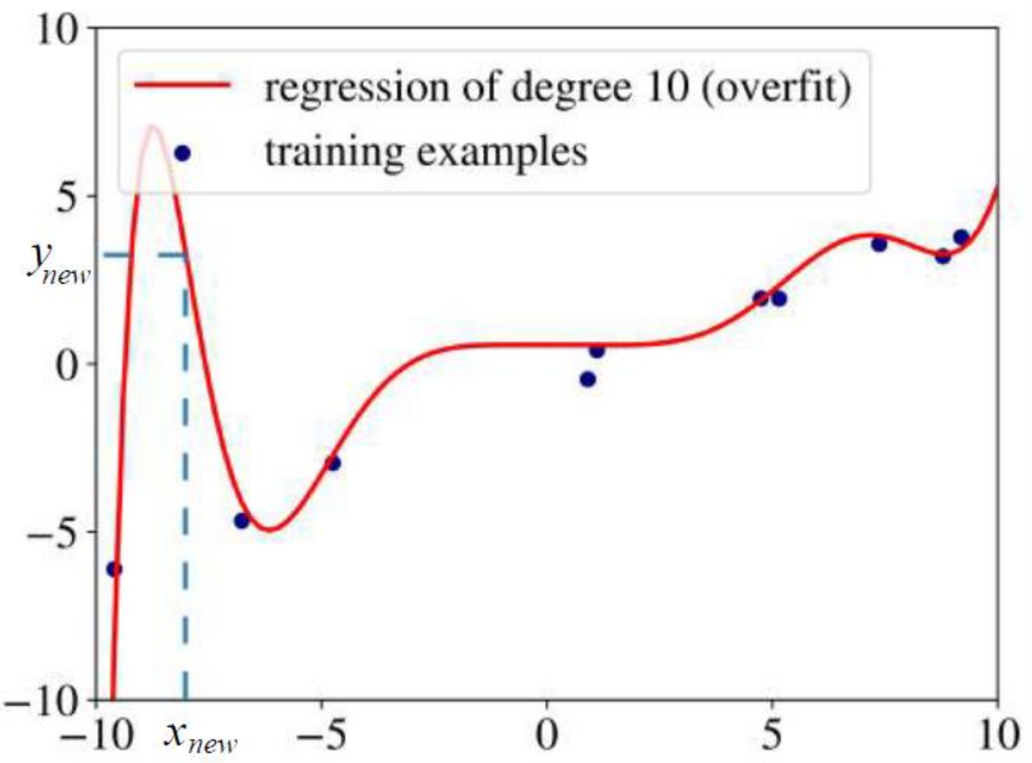


Figure 5.2: Overfitting

In the Figure above is shown an example of overfitting, the blue line in the same of the figure 5.1 but the difference is that this time this is the polynomial regression with a polynomial degree 10.

5.2.2 Logistic Regression

Logistic regression is a classification learning algorithm, not a regression. Be attention to not confuse due to the name.

Starting from the standard logistic function that represent the logistic regression. The idea is to model yi as a linear function of xi. The linear combination of features such as wxi + b is a function that spans from minus infinity to plus infinity, while yi has only two possible values.

Standard logistic function:

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

Where:

- e: base of the natural logarithm

The standard logistic function can be also written in the logarithmic form:

$$f_{\mathbf{w},b}(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{1 + e^{-(\mathbf{wx}+b)}}.$$

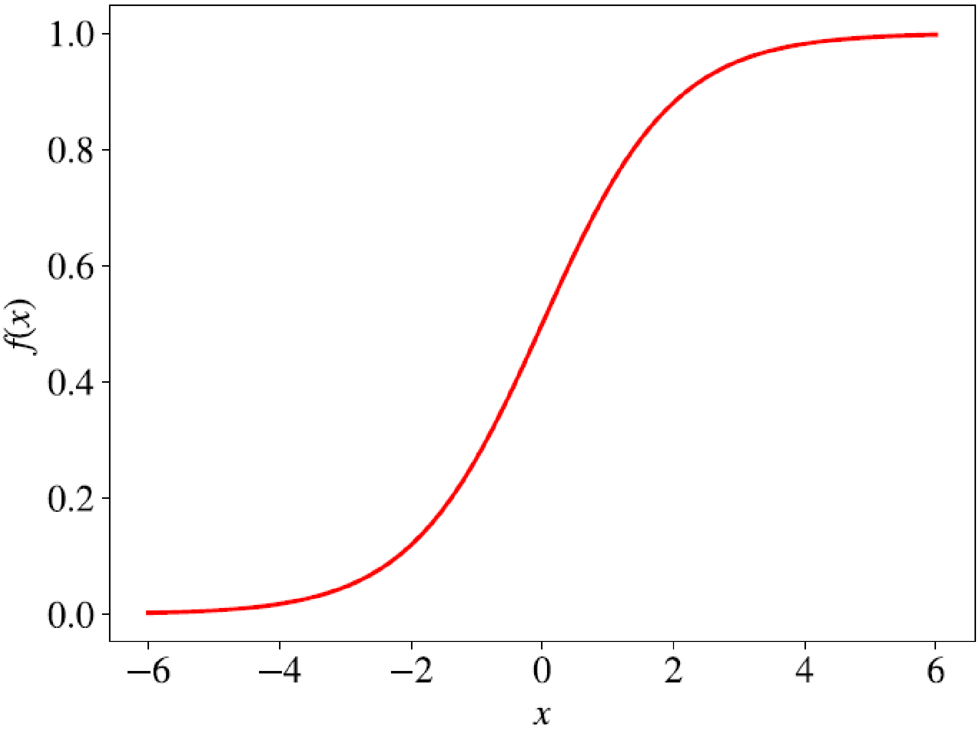


Figure 5.3: Standard logistic function

To find optimal **w**u and *b*u the empirical risck will be minimized thanks to tfe mean squared error or MSE.

Likelihhod: likelihood function defines how likely the observation (an example) is according to our model. The goal is to mi maximize this function. The optimization criterion in logistic regression is called maximum likelihood. Instead of minimizing the average loss, like in linear regression, we now maximize the likelihood of the training data according to our model:

$$L_{\mathbf{w},b} \stackrel{\text{def}}{=} \prod_{i=1\ldots N} f_{\mathbf{w},b}(\mathbf{x}_i)^{y_i}(1 - f_{\mathbf{w},b}(\mathbf{x}_i))^{(1-y_i)}.$$

Because of the exp function used in the model, in practice, it’s more convenient to maximize the log-likelihood instead of likelihood. The log-likelihood is defined like follows:

$$LogL_{\mathbf{w},b} \stackrel{\text{def}}{=} \ln(L(_{\mathbf{w},b}(\mathbf{x})) = \sum_{i=1}^{N} y_i \ln f_{\mathbf{w},b}(\mathbf{x}) + (1 - y_i)\ln(1 - f_{\mathbf{w},b}(\mathbf{x})).$$

A typical numerical optimization procedure used in such cases is gradient descent.

5.2.3 Decision Tree Learning

A decision tree is an acyclic graph that can be used to make decision, it can learned from data. The ude is to build a decision tree that would allow to predict the class given a feature vector.

The optimization criterion, in this case, is the average log-likelihood:

$$\frac{1}{N}\sum_{i=1}^{N} y_i \ln f_{ID3}(\mathbf{x}_i) + (1 - y_i)\ln(1 - f_{ID3}(\mathbf{x}_i)),$$

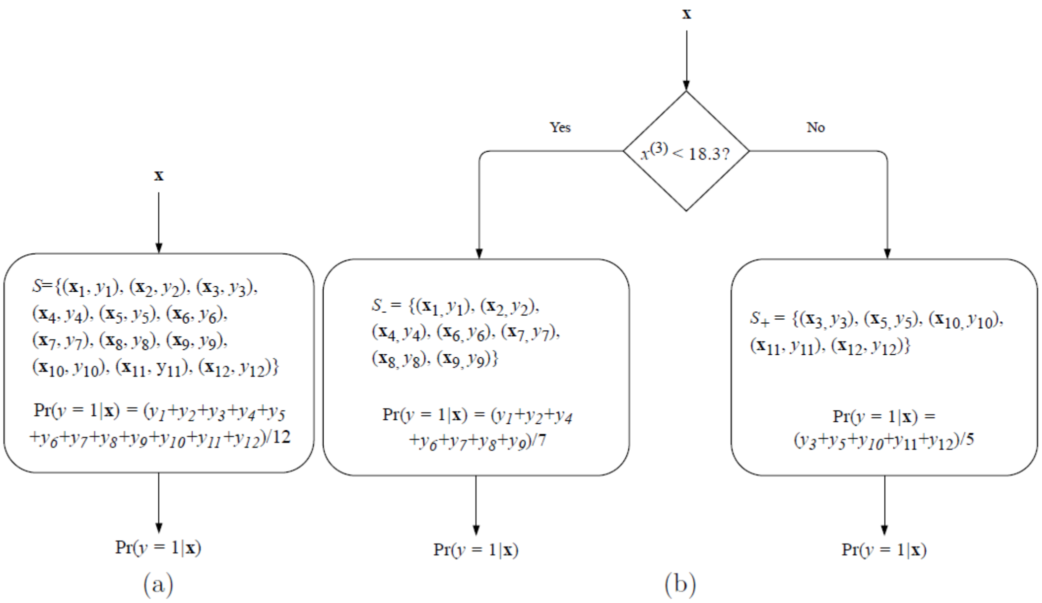
f_{ID3} is a decision tree. The ID3 algorithm optimizes it approximately by constructing a non-parametric model $f_{ID3}(\mathbf{x})$ d=ef Pr(y = 1|$\mathbf{x}$).

This algorithm works with the following steps:

start node that contains all examples: $S_{d=ef}\{(\mathbf{x}_i, y_i)\}_{Ni=1}$. Start with a constant model $f_{S\ ID3}$:

$$f_{ID3}^{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x},y)\in\mathcal{S}} y.$$

At this moment we must split with the pieces os S- and S+, in the following figure is shown a decisomn tree after one split:



x
S={(x1, y1), (x2, y2), (x3, y3), (x4, y4), (x5, y5), (x6, y6), (x7, y7), (x8, y8), (x9, y9), (x10, y10), (x11, y11), (x12, y12)}
Pr(y = 1|x) = (y1+y2+y3+y4+y5 +y6+y7+y8+y9+y10+y11+y12)/12
Pr(y = 1|x)
(a)
x
x(3) < 18.3?
Yes
No
S- = {(x1, y1), (x2, y2), (x4, y4), (x6, y6), (x7, y7), (x8, y8), (x9, y9)}
Pr(y = 1|x) = (y1+y2+y4 +y6+y7+y8+y9)/7
Pr(y = 1|x)
S+ = {(x3, y3), (x5, y5), (x10, y10), (x11, y11), (x12, y12)}
Pr(y = 1|x) = (y3+y5+y10+y11+y12)/5
Pr(y = 1|x)
(b)

Figure 5.4: Split

Entropoty is the criteria that is usude for the goodness of the split, it is a nmeasurement of uncertainty about a random variable.
The formula of entropy is:

$$H(\mathcal{S}) = -f_{ID3}^{\mathcal{S}} \ln f_{ID3}^{\mathcal{S}} - (1 - f_{ID3}^{\mathcal{S}}) \ln(1 - f_{ID3}^{\mathcal{S}}).$$

In the figure 5.4 the (a) In the beginning, the decision tree only contains the start node; it makes the same prediction for any input. (b) The decision tree after the first split; it tests whether feature 3 is less than 18.3 and, depending on the result, the prediction is made in one of the two leaf nodes.
The algorithm stops at a leaf in any of the below situation:

- All examples in the leaf node are classified correctly by the one-piece model
- We cannot find an attribute to split upon.
- The split reduces the entropy less than some ‘ (the value for which has to be found experimentally3).
- The tree reaches some maximum depth d (also must be found experimentally).

5.2.4 Support Vector Machine

Support vector machine (SVM) is studied in two different situations:

- Linearly non-separable case with the present of noiuse
- Linearly non-separable case inherent nonlinearly

The goal is to satisfy the following constraints:

$$\begin{aligned}&\mathbf{wx}_i - b \geq 1 \text{ if } y_i = +1, \text{ and}\\&\mathbf{wx}_i - b \leq -1 \text{ if } y_i = -1\end{aligned}$$

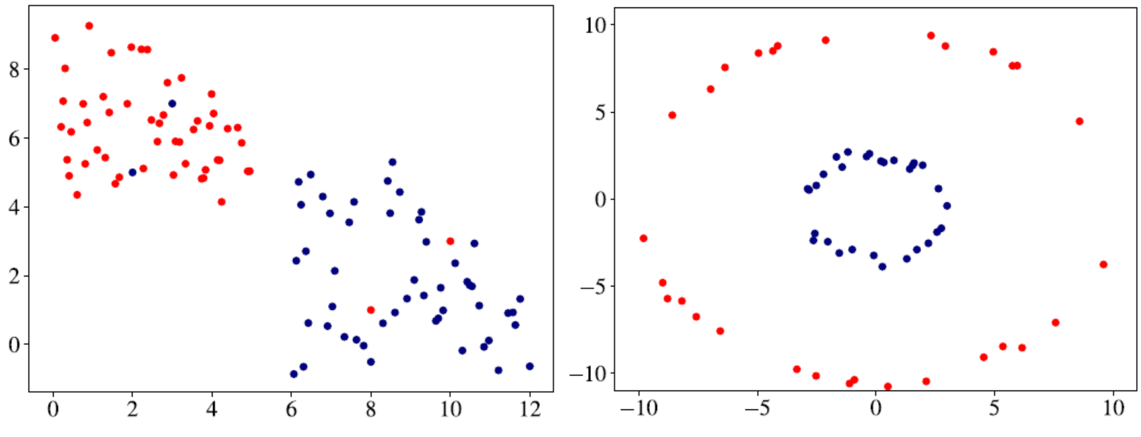


Figure 5.5: Linearly non-separable cases. Left the presence of noise, rightr the inherent nonlinearity

In this case the data are non-linearly we must introduce the hinge loss function: $max (0, 1 \neq y_i(\mathbf{wx}_i \neq b)).$

We then wish to minimize the following cost function:

$$C\|\mathbf{w}\|^2 + \frac{1}{N}\sum_{i=1}^{N} \max\left(0, 1 - y_i(\mathbf{w}\mathbf{x}_i - b)\right),$$

where the hyperparameter C determines the trade-off between increasing the size of the decision boundary and ensuring that each xi lies on the correct side of the decision boundary.

The value of C is usually chosen experimentally, just like ID3's hyperparameters 'and d. SVMs that optimize hinge loss are called soft-margin SVMs, while the original formulation is referred to as a hard-margin SVM.

In case of inherent Non -Linearity in SVMs, using a function to implicitly transform the original space into a higher dimensional space during the cost function optimization is called the kernel trick.

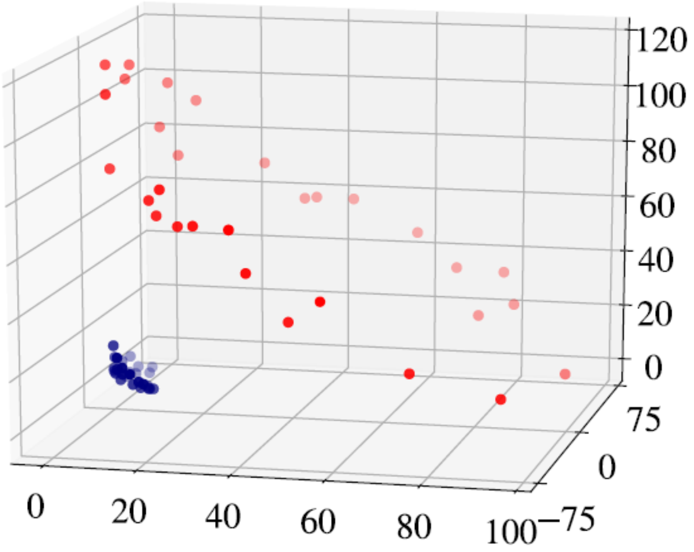


Figure 5.6: Inherent Non-Linearity

It's possible to transform a two-dimensional non-linearly separable data into a linearly separable three-dimensional data using a specific mapping.

The method traditionally to solve the optimization problems is the method of Lagrange multiplayer with the following formula:

$$\max_{\alpha_1 \ldots \alpha_N} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} y_i \alpha_i (\mathbf{x}_i \mathbf{x}_k) y_k \alpha_k \text{ subject to } \sum_{i=1}^{N} \alpha_i y_i = 0 \text{ and } \alpha_i \geq 0, i = 1, \ldots, N,$$

Where α_i are called lagrange multipliers.

5.2.5 K-Nearest Neighbors

K-Nearest Neighbors (kNN) is a non-parametric learning algorithm. Contrary to other learning algorithms that allow discarding the training data after the model is built, kNN keeps all training examples in memory. Once a new, previously unseen example x comes in, the kNN algorithm finds k training examples closest to x and returns the majority label (in case of classification) or the average label (in case of regression).

$$s(\mathbf{x}_i, \mathbf{x}_k) \stackrel{\text{def}}{=} \cos(\angle(\mathbf{x}_i, \mathbf{x}_k)) = \frac{\sum_{j=1}^{D} x_i^{(j)} x_k^{(j)}}{\sqrt{\sum_{j=1}^{D} \left(x_i^{(j)}\right)^2} \sqrt{\sum_{j=1}^{D} \left(x_k^{(j)}\right)^2}},$$

5.3 Anatomy of Learning Algorithm

In general, it possible to consider each leering algorithm composed in three parts:

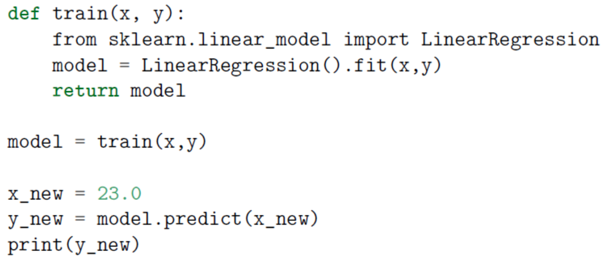
- Loss Function
- Optimization criteria based on the loss function
- Optimization routine lev earning a training data to find a solution to the optimization criteria

When the optimization criterion is differentiable there are two different optimization algorithms:

- Gradient descent is an iterative optimization algorithm for finding the minimum of a function. To find a local minimum of a function using gradient descent, one starts at some random point and takes steps proportional to the negative of the gradient (or approximate gradient) of the function at the current point.

- Gradient descent can be used to find optimal parameters for linear and logistic regression, SVM and neural networks which we consider later. For many models, such as logistic regression or SVM, the optimization criterion is convex. Convex functions have only one minimum, which is global. Optimization criteria for neural networks are not convex, but in practice even finding a local minimum suffices.

This thesis is concentrated in the application of machine leering for the financial sector, we are not going to explain the details of gradient, gradient descent, in fact in case of engineering is realy difficult that there is a possibility to implemt the formula for gradient., but using libraries, most of which are open source. A library is a collection of algorithms and supporting tools implemented with stability and efficiency in mind. The most frequently used in practice open-source machine learning library is scikit-learn. It’s written in Python and C. Here’s how you do linear regression in scikit-learn:



```
def train(x, y):
    from sklearn.linear_model import LinearRegression
    model = LinearRegression().fit(x,y)
    return model

model = train(x,y)

x_new = 23.0
y_new = model.predict(x_new)
print(y_new)
```

5.4 Basic Practice.

In this chapter will be study the promlems that be to solve to implement a machine learning in a engineering word. The main problems are:

- Feature engineering
- Overfitting
- Hyperparameter tuning

5.4.1 Feature Engineering

The firs stepo is to build a dataset. In oder to transform a row data into dataset there is a proble that is called feature engineering. Feature engineering is a labour-intensive process that demands from the data analyst a lot of creativity and, preferably, domain knowledge.

One- Hot Encoding

It is possible to transform specific categorical feature into several binary bons, only when the dataset is categorical.

For example, if there is a categorical of colours on this feature there are three possible values.

$$\begin{aligned} red &= [1,0,0] \\ yellow &= [0,1,0] \\ green &= [0,0,1] \end{aligned}$$

Binning

It is the process of converting a continuous feature into multiple binary features called bins or buckets, typically based on value range.

Normalization

Normalization is the process of converting an actual range of values which numerical feature can take, into a standard range of values, typically in the interval [≠1, 1] or [0, 1].

Normalization formula:

$$\bar{x}^{(j)} = \frac{x^{(j)} - min^{(j)}}{max^{(j)} - min^{(j)}},$$

Normally the normalization is useful for improve the speed of learning

Standardization

Standardization (or z-score normalization) is the procedure during which the feature values are rescaled so that they have the properties of a standard normal distribution with µ = 0 and ‡ = 1, where µ is the mean (the average value of the feature, averaged over all examples in the dataset) and ‡ is the standard deviation from the mean.

Standard scores (or z-scores) of features are calculated as follows:

$$\hat{x}^{(j)} = \frac{x^{(j)} - \mu^{(j)}}{\sigma^{(j)}}.$$

Data Input Techniques
It is the methodology that consist in replacing the missing value of a feature by an average value of this feature in the dataset:

$$\hat{x}^{(j)} = \frac{1}{N} x^{(j)}.$$

M is the number of examples.

5.5 Learning Algorithm Selection

How choose a machine learning algorithm?
Hear is present e several question that must be ask before to choose a machine learning algorithm:

- Explain ability
- In memory ora out of memory
- Number of feature and examples
- Categorical or numerical features
- Nonlinearity of the data
- Taring speed. More the algorithm is simple and more the train is fast
- Prediction speed, the popular way to choose one is by testing it in the validation test

5.6 Three sets

The data analyst works with three sets:

- Training set
- Validation set
- Test set

The three sets are used for setting to assess the model. In the following figure is shown a section decision diagram

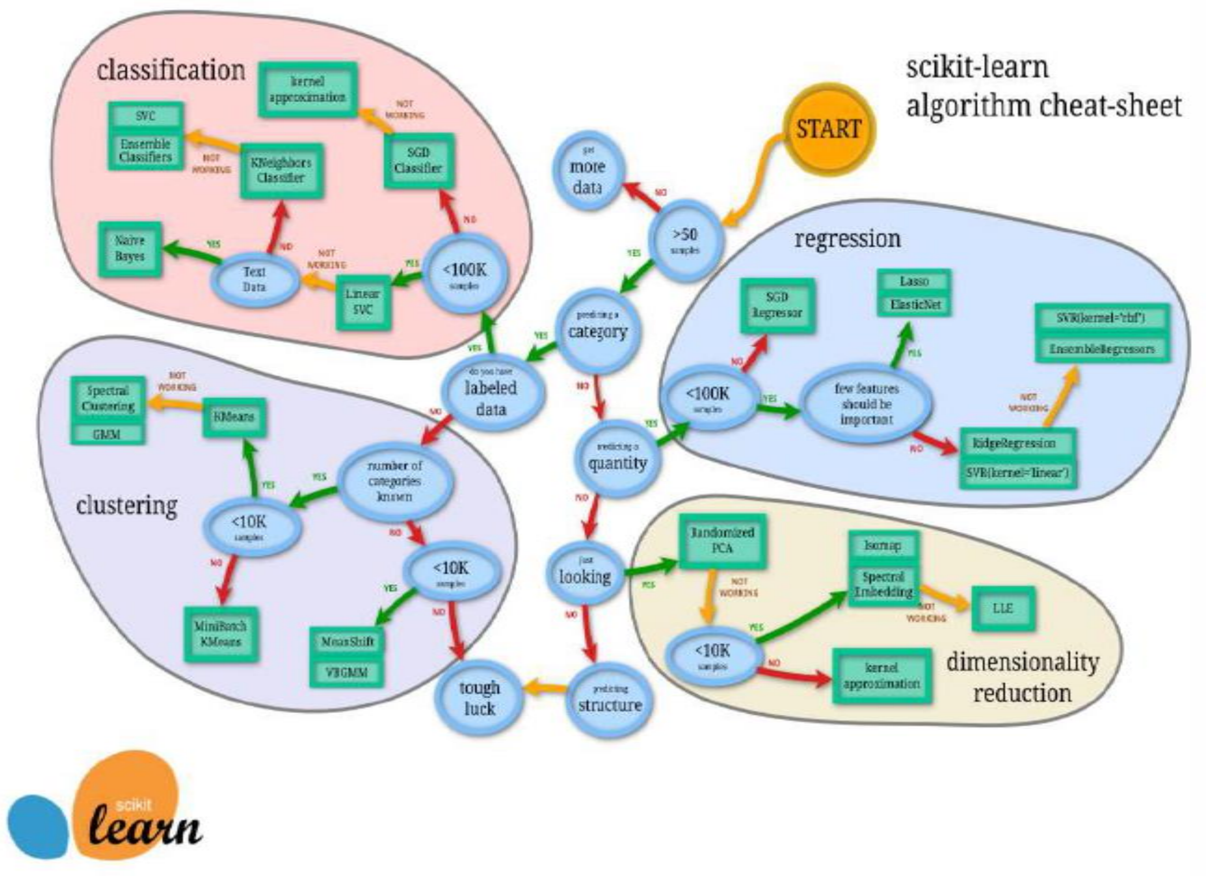


Figure 5.7: Algorithm section diagram

5.7 Underfitting and Overfitting

Underfitting is the inability of the model to predict well the labels of the data it was trained on. There most important reason are:

- The model is too simple for the data
- The features you engineered are not enough

Overfitting is a problem a model can exhibit. The model that overfits predicts very well the training data but poorly the data from at least one of the two holdouts sets. The main reason are:

- The model is to complex for the data

- Too many features but a small number of training examples

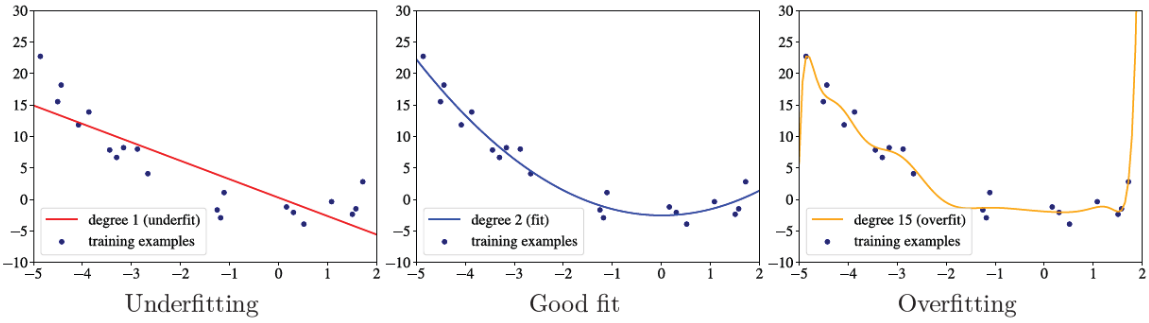


Figure 5.8: Examples of fit

In the figure below is represented different tape of fit. To avoid the overfitting, there are several solutions

- Try a simpler model
- Reduce the dimensionality of examples in the dataset
- Add more training data
- Regularize the mode, regularization is the most widely used approach to prevent overfitting.

5.8 Regularization

There are two different types of regulation, L1 and L2 regularization. The goal is created e regulation model modifying the object function adding a penalization term.

L1 and L2 regulation methods can be combined. The combination of L1 and L2 is called elastic net regularization.

Taking in example the linear regression, the L1-regularized looks like:

$$\min_{\mathbf{w},b} C|\mathbf{w}| + \frac{1}{N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2,$$

The L2-regularized objective looks like:

$$\min_{\mathbf{w},b} C\|\mathbf{w}\|^2 + \frac{1}{N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2, \text{ where } \|\mathbf{w}\|^2 \stackrel{\text{def}}{=} \sum_{j=1}^{D}(w^{(j)})^2.$$

The L1 regulation produces a sparse model, L2 usually provide better solution with the advantage of being differentiable.

5.9 Model Performance Assessment

When a machine learning algorithm is done is very useful to have some methods to define the performance of their model.
For the regression is simple to define the performance, because is dependent from the quality of the fitting.
For the classification, the most widely used and tools assess the classification model are:

- Confusion Matrix
- Precision/Recall
- Accuracy
- Cost-Sensitive Accuracy
- Area under the ROC Curve (AUC)

In the following lines we are going to explain in detail each method.

Confusion Matrix
It is used to calculate two different performance metrics:

- Precision
- Recall

Confusion Matrix is a table that summarizes how successful the classification model is at predicting examples belonging to various classes.

Precision/Recall
They are the two most frequently used metrical, precision is the rate oof correct positive predictions to the overall number of positive precisions

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

Recall is the ratio of correct positive predictions:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Accuracy

It is given by the number of correctly classified examples divided by the total number of classified examples

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

Cost- Sensitive Accuracy

We must use the equation of the accuracy, it necessary to assign a cost to bth types of mistakes.

Area under the ROC Curve (AUC)

It is a combinate between true positive rate and false positive rate, it is shown in the following formula:

$$\text{TPR} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \text{ and } \text{FPR} = \frac{\text{FP}}{(\text{FP} + \text{TN})}.$$

5.10 Neural Networks and Deep Learning

5.10.1 Neural Networks

A neural network (NN) is a function, this function is nested function.

The neural networks function for a 3-layer returns and scalar and it’s look like:

$$y = f_{NN}(\mathbf{x}) = f_3(\boldsymbol{f}_2(\boldsymbol{f}_1(\mathbf{x}))).$$

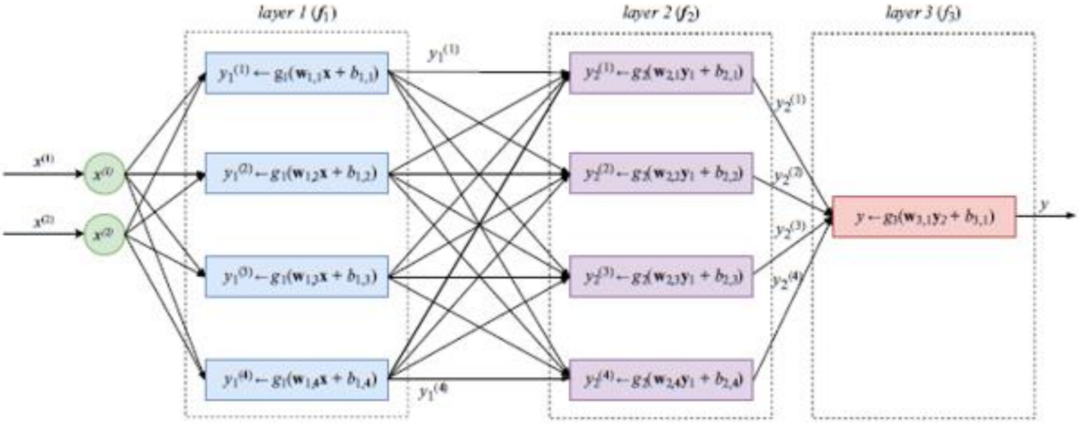
f1 and f2 are a vectors functions of the following from:

$$\boldsymbol{f}_l(\mathbf{z}) \stackrel{\text{def}}{=} \boldsymbol{g}_l(\mathbf{W}_l\mathbf{z} + \mathbf{b}_l),$$

Where:

- l is the layer index
- g ia called activation function
- W ia a matrix
- B in a vectors

In the following picture is shown how a neural network as a connected combination of units logically organized into one or more layer.



layer 1 (f_1)
layer 2 (f_2)
layer 3 (f_3)
$x^{(1)}$
$x^{(2)}$
$x^{(1)}$
$x^{(2)}$
$y_1^{(1)} \leftarrow g_1(\mathbf{w}_{1,1}\mathbf{x} + b_{1,1})$
$y_1^{(2)} \leftarrow g_1(\mathbf{w}_{1,2}\mathbf{x} + b_{1,2})$
$y_1^{(3)} \leftarrow g_1(\mathbf{w}_{1,3}\mathbf{x} + b_{1,3})$
$y_1^{(4)} \leftarrow g_1(\mathbf{w}_{1,4}\mathbf{x} + b_{1,4})$
$y_1^{(1)}$
$y_1^{(4)}$
$y_2^{(1)} \leftarrow g_2(\mathbf{w}_{2,1}\mathbf{y}_1 + b_{2,1})$
$y_2^{(2)} \leftarrow g_2(\mathbf{w}_{2,2}\mathbf{y}_1 + b_{2,2})$
$y_2^{(3)} \leftarrow g_2(\mathbf{w}_{2,3}\mathbf{y}_1 + b_{2,3})$
$y_2^{(4)} \leftarrow g_2(\mathbf{w}_{2,4}\mathbf{y}_1 + b_{2,4})$
$y_2^{(1)}$
$y_2^{(2)}$
$y_2^{(3)}$
$y_2^{(4)}$
$y \leftarrow g_3(\mathbf{w}_{3,1}\mathbf{y}_2 + b_{3,1})$
y

Figure 5.9: A multiplayer perceptron with two-dimensional input

In figure 5.9 , the activation function gl has one index: l, the index of the layer the unit belongs to.

Usually, all units of a layer use the same activation function, but it's not strictly necessary.

Each layer can have a different number of units. Each unit has its own parameters $\mathbf{w}_{l,u}$ and $b_{l,u}$, where u is the index of the unit, and l is the index of the layer. The vector $\mathbf{y}_{l\neq 1}$ in each unit is defined as $[y(1)_{l\neq 1}, y(2)_{l\neq 1}, y(3)_{l\neq 1}, y(4)_{l\neq 1}]$. The vector x in the first layer is defined as $[x(1), \ldots, x(D)]$.

5.10.2 Deep Learning

The term deep learning refers to training neural networks using modern algorithmic and mathematical toolkit independently of how deep neural network is. Many businesses problem can be solved with neural networks.

Convolution neural network

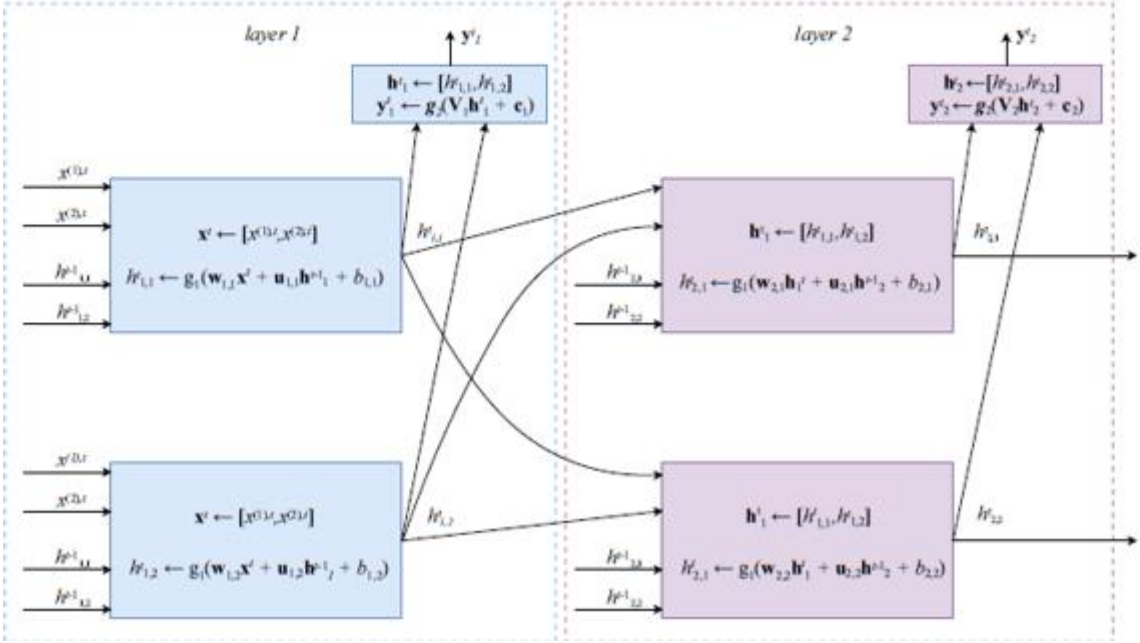
Convulution neural network (CNN), is a special kind of feed-forward neural network (FFNN)that signific ally reduces the number of parameters ina deep neural network with many units without losing too much in the quality of model. Convolutional neural network has found applications in image and text processing where they beat many previously established benchmarks.

Recurrent Neural Network

Recurrent neural networks (RNNs) are used to label, classify, or generate sequences. A sequence is a matrix, each row of which is a feature vector, and the order of rows matters.

Labeling a sequence means predicting a class to each feature vector in a sequence. Classifying a sequence means predicting a class for the entire sequence. Generating a sequence means to output another sequence (of a possibly different length) somehow relevant to the input sequence.

To have an example in figure 5.10 is shown how each training example ia a matrix in which row ia a feature.



layer 1
layer 2
$\mathbf{y}^t_1$
$\mathbf{h}^t_1 \leftarrow [h^t_{1,1}, h^t_{1,2}]$
$\mathbf{y}^t_1 \leftarrow g_2(\mathbf{V}_1\mathbf{h}^t_1 + \mathbf{c}_1)$
$\mathbf{y}^t_2$
$\mathbf{h}^t_2 \leftarrow [h^t_{2,1}, h^t_{2,2}]$
$\mathbf{y}^t_2 \leftarrow g_2(\mathbf{V}_2\mathbf{h}^t_2 + \mathbf{c}_2)$
$x^{(1),t}$
$x^{(2),t}$
$h^{t-1}_{1,1}$
$h^{t-1}_{1,2}$
$\mathbf{x}^t \leftarrow [x^{(1),t}, x^{(2),t}]$
$h^t_{1,1} \leftarrow g_1(\mathbf{w}_{1,1}\mathbf{x}^t + \mathbf{u}_{1,1}\mathbf{h}^{t-1}_1 + b_{1,1})$
$h^t_{1,1}$
$h^t_{1,2}$
$h^t_{1,2} \leftarrow g_1(\mathbf{w}_{1,2}\mathbf{x}^t + \mathbf{u}_{1,2}\mathbf{h}^{t-1}_l + b_{1,2})$
$h^{t-1}_{2,1}$
$h^{t-1}_{2,2}$
$\mathbf{h}^t_1 \leftarrow [h^t_{1,1}, h^t_{1,2}]$
$h^t_{2,1} \leftarrow g_1(\mathbf{w}_{2,1}\mathbf{h}^t_1 + \mathbf{u}_{2,1}\mathbf{h}^{t-1}_2 + b_{2,1})$
$h^t_{2,1}$
$h^t_{2,2}$
$h^t_{2,1} \leftarrow g_1(\mathbf{w}_{2,2}\mathbf{h}^t_1 + \mathbf{u}_{2,2}\mathbf{h}^{t-1}_2 + b_{2,2})$

Figure 5.10: The first two layer of an RAN. The input vectors are two-dimensional; each layer has two units

Other important extensions to RNNs include bi-directional RNNs, RNNs with attention and sequence-to-sequence RNN models. Sequence-to-sequence RNNs are frequently used to build neural machine translation models and other models for text to text transformations. A generalization of RNNs is a recursive neural network model.

5.11 Problems and Solutions

In this chapter the main function for the solution of the problem will be listed. It is not present a global explanation because the goal of this work is to analyse how machine learning is used in the financial sector.

Kernel Regression
It is no parametric method, there are no parameters to learn. The model on the data itself. In the simplest form, in kernel regression we look for a model like this:

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} w_i y_i, \text{ where } w_i = \frac{N k(\frac{x_i - x}{b})}{\sum_{k=1}^{N} k(\frac{x_k - x}{b})}.$$

Multiclass Classification
In multiclass classification, the label can be one of the C classes: y oe {1, . . . ,C}. Many machine learning algorithms are binary; SVM is an example. Some algorithms can naturally be extended to handle multiclass problems. ID3 and other decision tree learning algorithms can be simply changed like this:

$$f_{ID3}^{S} \stackrel{\text{def}}{=} \Pr(y_i = c|\mathbf{x}) = \frac{1}{|S|} \sum_{\{y \,|\, (\mathbf{x},y) \in S, y=c\}} y,$$

One – class Classification
The idea behind the one-class gaussian is that we model our data as if it came from a Gaussian distribution, more precisely multivariate normal distribution

(MND). The probability density function (pdf) for MND is given by the following equation:

$$f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^{D}|\boldsymbol{\Sigma}|}} \quad ,$$

In the figure 5.11 is shown an application of one- class classification solved the Gaussian method

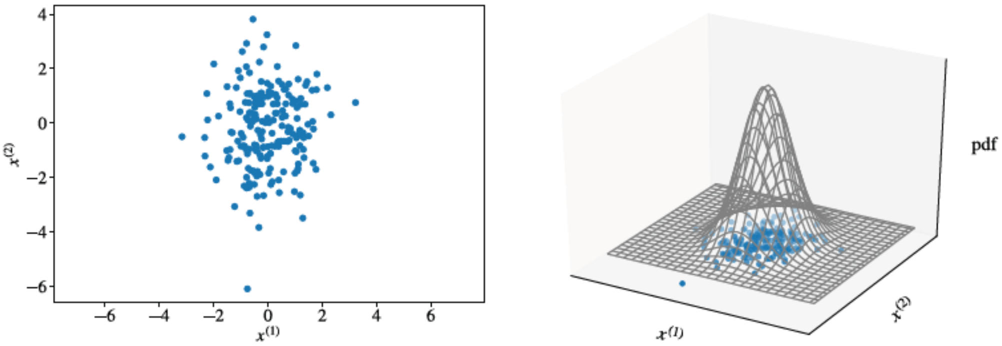


Figure 5.11: One-class classification solved using the one-class gaussian method

Multi-Label Classification

Neural networks algorithms can naturally train multi-label classification models by using the binary cross-entropy cost function. The output layer of the neural network, in this case, has one unit per label. Each unit of the output layer has the sigmoid activation function.

Accordingly, each label l is binary ($y_{i,l}$ oe {0, 1}), where $l = 1, \ldots, L$ and $i = 1, \ldots, N$. The binary cross-entropy of predicting the probability ˆyi,l that example xi has label l is defined as $\neq (y_{i,l} \ln(\hat{}y_{i,l}) + (1 \neq y_{i,l}) \ln(1 \neq \hat{}y_{i,l}))$. The minimization criterion is simply the average of all binary cross-entropy terms across all training examples and all labels of those examples. It also uses in the case of number for example.

Enseble Learning

It is an approach to boost the performance of simple learning algorithms.

Enseble learning is a learning paradigm is a learning paradigm that, instead of trying to learn one super-accrue model, focuses on training e large number of low-accuracy models and then combining the predictions given by those weak models to obtain a high-accuracy meta-model. In total there are boosting and bagging.

Learning to Label Sequences

Sequence ladling is the problem of automatically assigning a label to each element of a sequence. A label sequential training example in sequence libelling is pair of list (X,Y) where X is a list of feature vectors, one per time step, Y is a list of the same length of labels.

The model called Condition random Fields (CRF) is a very effective alternative that often well in practice for the feature vectors that have many informative features.

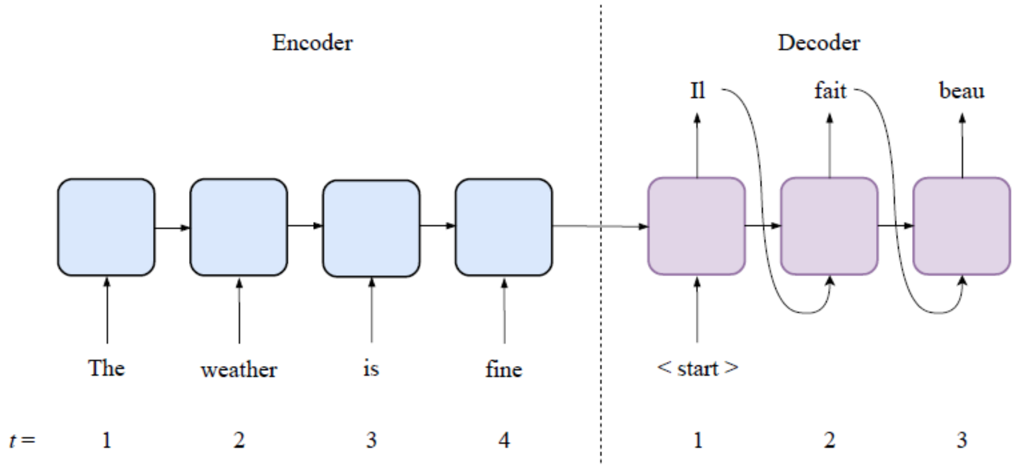


Figure 5.12: A traditional seq2seq architecture

Sequence-to-sequence Learning

Sequence-to-sequence learning (often abbreviated as seq2seq learning) is a generalization of the sequence labelling problem. In seq2seq, Xi and Yi can have different length. seq2seq models have found application in machine translation (where, for example, the input is an English sentence, and the output is the corresponding French sentence), conversational interfaces (where the input is a question typed by the user, and the output is the answer from the machine), text summarization, spelling correction, and many others.

A traditional seq2seq architecture is illustrated in fig. 4. More accurate predictions can be obtained using an architecture with attention. Attention mechanism is implemented by an additional set of parameters that combine some information from the encoder (in RNNs, this information is the list of state vectors of the last recurrent layer from all encoder time steps) and the current state of the decoder to generate the label. That allows for even better retention of long-term dependencies than provided by gated units and bidirectional RNN. A seq2seq architecture with attention is illustrated in Figure 5.12

Active Learning

Active learning is an interesting, supervised learning paradigm. It is usually applied when obtaining labelled examples is costly. That is often the case in the medical or financial domains, where the opinion of an expert may be required to annotate patients ‘’or customers’’data. The idea is that we start the learning with relatively few labelled examples, and many unlabelled ones, and then add labels only to those examples that contribute the most to the model quality.

There are two main important strategies.

- Data density and uncertainty
- Support vector-based

Semi -Supervised Learning

In semi-supervised learning (SSL) we also have labeled a small fraction of the dataset; most of the remaining examples are unlabeled. Our goal is to leverage many unlabelled examples to improve the model performance without asking an expert for additional labelled examples.

An autoencoder is a feed-forward neural network with an encoder-decoder architecture. It is trained to reconstruct its input. So, the training example is a pair (x, x). We want the output x of the model f(x) to be as like the input x as possible, like in the figure 5.13

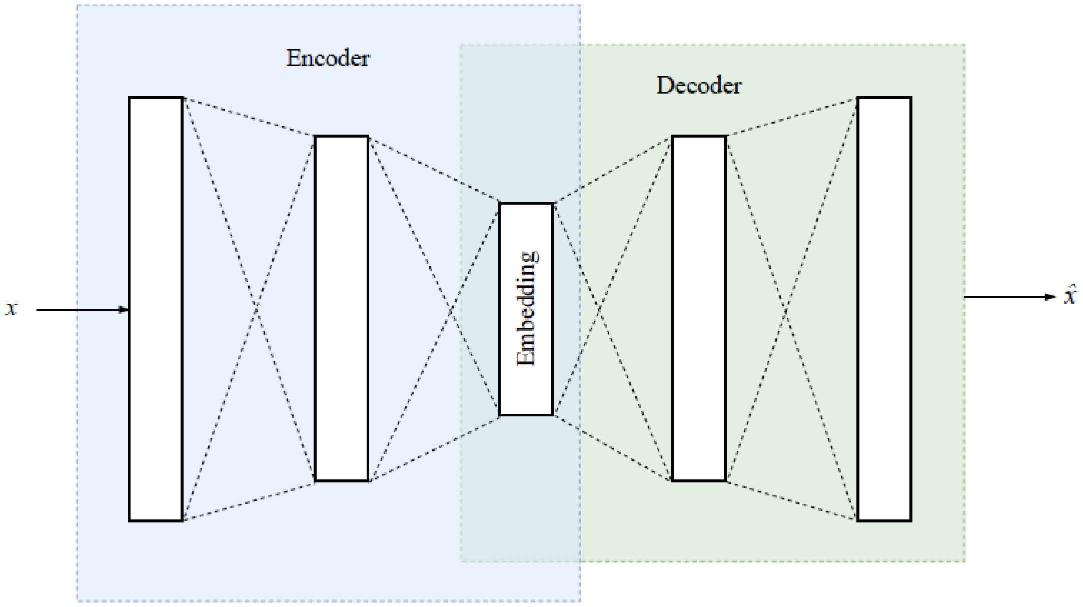


Figure 5.13: Autoencoder

One- Shot Learning

In one-shot learning, typically applied in face recognition, we want to build a model that can recognize that two photos of the same person represent that same person.

It’s a common misconception that for one-shot learning we need only one example of each entity for training. In practice, we need much more than one example of each person for the person identification model to be accurate. It’s called one-shot because of the most frequent application of such a model: face-based authentication. For example, such a model could be used to unlock your phone. If your model is good, then you only need to have one picture of you on your phone and it will recognize you, and it will recognize that someone else is not you. When we have the model, to decide whether two pictures A and Aˆ belong to the same person, we check if //f(A) ≠ f(Aˆ)// is less than some threshold · , which is another hyperparameter of the model.

Zero-Shot Learning

In zero-shot learning (ZSL) we want to train a model to assign labels to objects. The most frequent application is to learn to assign labels to images.

The trick is to use embeddings not just to represent the input x but also to represent the output y.

5.12 Unsupervised Learning

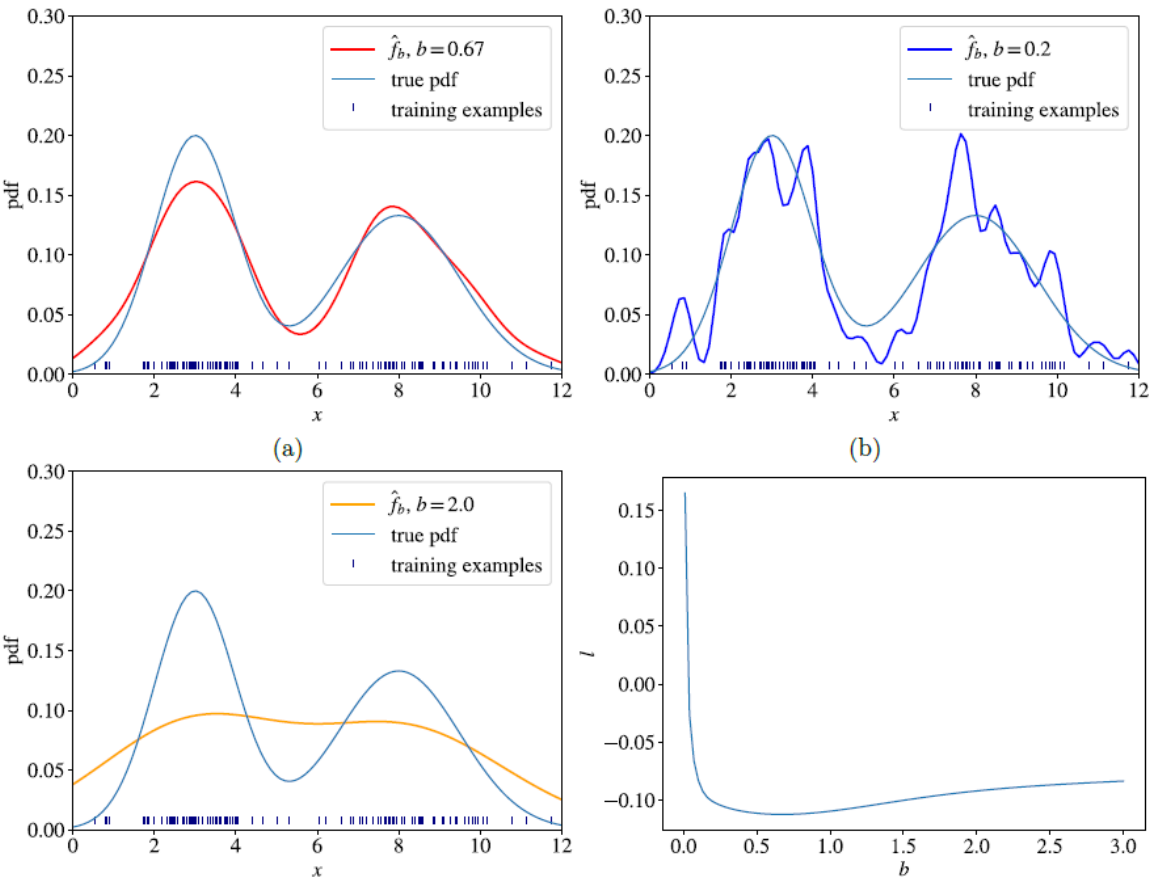
In unsupervised learing label the problems are that there are note labels, it is very problematic for many applications.

5.12.1 Density Estimation

It is a problem of modelling the probability density function of a known distribution from which the data set has been drawn.
It is possible to use a nonparametric model in kennel regression, the formula of kernel regression is:

$$\hat{f}_b(x) = \frac{1}{Nb}\sum_{i=1}^{N} k\left(\frac{x - x_i}{b}\right),$$

In the following figure is shown the kenel refression in different form:



(a) (b) (c) (d)

Figure 5.13: Kernel Density Estimation

A responsible choice of measure of this difference is the mean integrator square error (MISE):

$$\text{MISE}(b) = \mathbb{E}\left[\int_{\mathbb{R}} (\hat{f}_b(x) - f(x))^2 \, dx\right].$$

It is possible to replace the summation with the integral

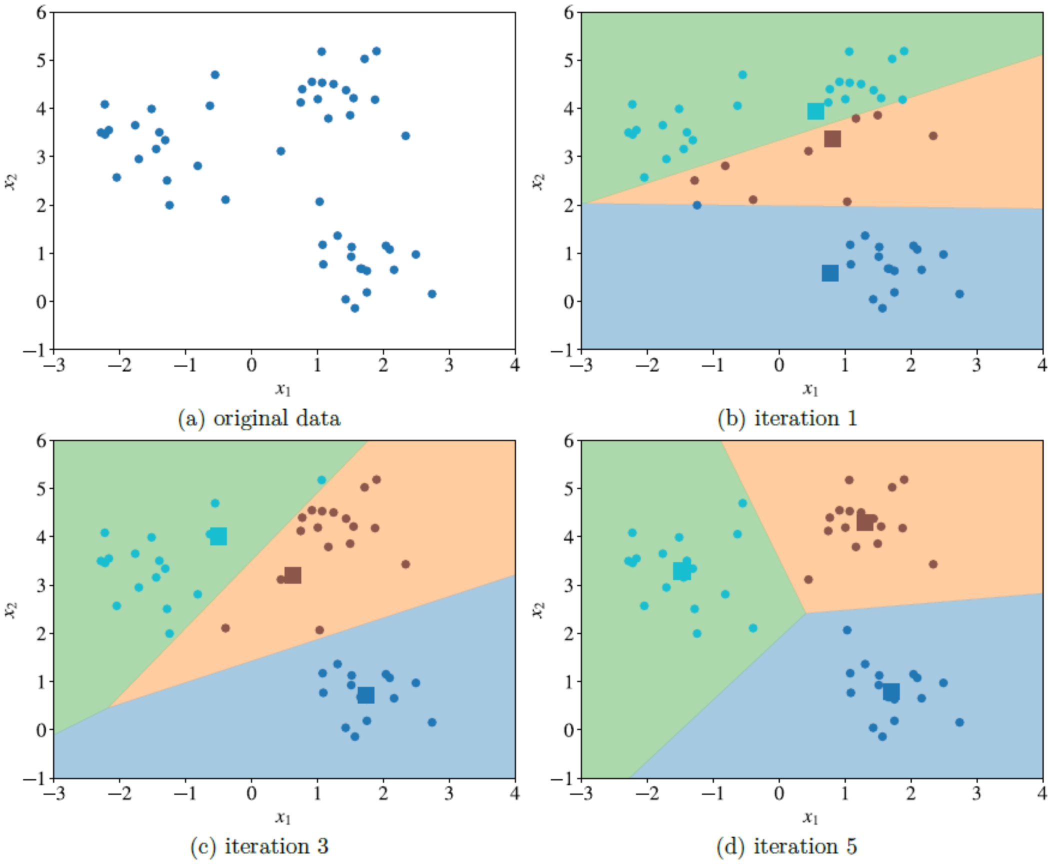
$$\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_b^2(x)dx\right] - 2\mathbb{E}\left[\int_{\mathbb{R}} \hat{f}_b(x)f(x)dx\right] + \mathbb{E}\left[\int_{\mathbb{R}} f(x)^2dx\right].$$

The second term can be second term can be approximated by ≠ 2NqNi=1 ˆ f(i)b (xi), where f(i)b is a kernel model of f computed on our training set with the example xi excluded. In order to fins the optimal value b* for b, is necessary to minimize the cost:

$$\int_{\mathbb{R}} \hat{f}_b^2(x)dx - \frac{2}{N}\sum_{i=1}^{N} \hat{f}_b^{(i)}(x_i).$$

5.12.2 Cluster

Clustering is a problem of learning to assign a label to examples by leveraging an unlabeled dataset. Because the dataset is completely unlabeled, deciding on whether the learned model is optimal is much more complicated than in supervised learning.



(a) original data

(b) iteration 1

(c) iteration 3

(d) iteration 5

Figure 5.14: Progress of the k-means algorithm for k=3

In the following lines the main important algorithms will be analized.

K-Means

The k-means clustering algorithm works as follows. First, the analyst has to choose k — the number of classes (or clusters). Then we randomly put k feature vectors, called centroids, to the feature space1. We then compute the distance from each example x to each centroid c using some metric, like the Euclidean distance. Then we assign the closest centroid to each example (like if we labeled each example with a centroid id as the label). For each centroid, we calculate the average feature vector of the examples labeled with it. These average feature vectors become the new locations of the centroids. We recompute the distance from each example to each centroid, modify the assignment and repeat the procedure until the assignments don't change after the centroid locations were recomputed. The model is the list of assignments of centroids IDs to the examples. The initial position of centroids influence the final positions, so two runs of k-means can result in two di_erent models. One run of the k-means algorithm is illustrated in figure 5.14. Di_erent background colors represent regions in which all points belong to the same cluster.

The value of k, the number of clusters, is a hyperparameter that has to be tuned by the data analyst. There are some techniques for selecting k. None of them is proven optimal. Most of them require from the analyst to make an “educated guess” by looking at some metrics or by examining cluster assignments visually. Later in this chapter, we consider one technique which allows choosing a reasonably good value for k without looking at the data and making guesses.

DBSCAN and HDBSCAN

While k-means and similar algorithms are centroid-based, DBSCAN is a density-based clustering algorithm. Instead of guessing how many clusters you need, by using DBSCAN, you define two hyperparameters: ‘and n. You start by picking an example x from your dataset at random and assign it to cluster 1. Then you count how many examples have the distance from x less than or equal to ‘. If this quantity is greater than or equal to n, then you put all these ‘-neighbors to the same cluster 1. You then examine each member of cluster 1 and find their respective ‘-neighbors.

If some member of cluster 1 has n or more neighbors, you expand cluster 1 by putting those neighbors to the cluster. You continue expanding cluster 1 until here are no more examples to put in it. In the latter case, you pick from the dataset another example not belonging to any cluster and put it to cluster 2. You continue like this until all examples either belong to some cluster or are marked as outliers.

An outlier is an example whose ‘-neighborhood contains less than n examples.

The advantage of DBSCAN is that it can build clusters that have an arbitrary shape, while kmeans and other centroid-based algorithms create clusters that have a shape of a hypersphere.

An obvious drawback of DBSCAN is that it has two hyperparameters and choosing good values for them (especially ‘) could be challenging. Furthermore, having fixed, the clustering algorithm cannot e_ectively deal with clusters of varying density.

HDBSCAN only has one important hyperparameter: n, that is the minimum number of examples to put in a cluster. This hyperparameter is relatively simple to choose by intuition. HDBSCAN has very fast implementations: it can deal with millions of examples e_ectively. Modern implementations of k-means are much faster than HDBSCAN though, but the qualities of the latter may outweigh its drawbacks for many practical tasks. It is recommended to always try HDBSCAN on your data first.

Determining the Number of Cluster

The problem is when is necessary to for D-dimensional data (D>3).

We must split the data into training and test set. We must training and test S_{tr} of size N_{tr} and S_{te} of size N_{te} respectively, you fix k, the number of clusters, and run a clustering algorithm C on sets S_{tr} and S_{te} and obtain the clustering results $C(S_{tr}, k)$ and $C(S_{te}, k)$.

Define the $N_{te} \lozenge N_{te}$ co-membership matrix $\mathbf{D}[A, S_{te}]$ as follows: $\mathbf{D}[A, S_{te}]_{(i,i0)} = 1$ if and only if examples $\mathbf{x}_i$ and $\mathbf{x}_{i0}$ from the test set belong to the same cluster according to the clustering A. Otherwise $D[A, S_{te}]_{(i,i0)} = 0$.

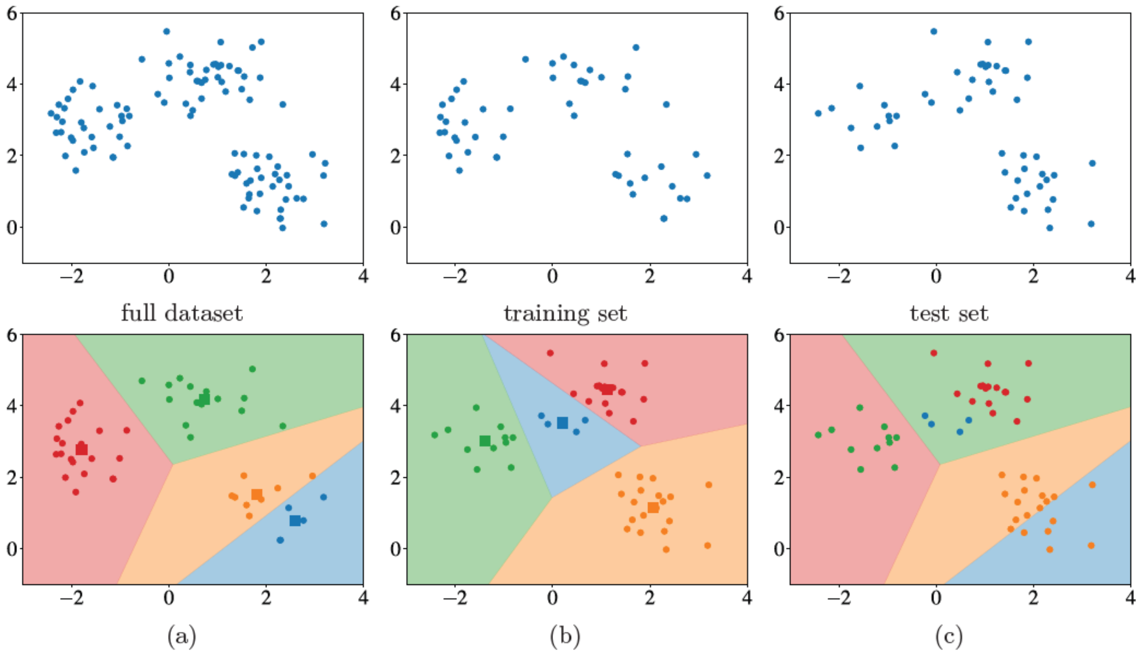


Figure 5.15: The clustering fot k=3

The prediction strength for the number of cluster k is given by:

$$\mathrm{ps}(k) \overset{\mathrm{def}}{=} \min_{j=1,\ldots,k} \frac{1}{|A_j|(|A_j|-1)} \sum_{i,i' \in A_j} \mathrm{D}[A, \mathcal{S}_{te}]^{(i,i')},$$

Where:

- A: cluster
- J: cluster from the clustering
- Aj: is the number of the examples in cluster

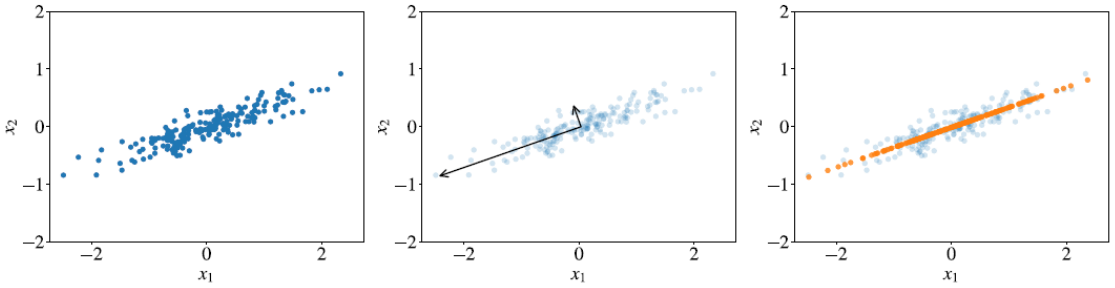
Another effective method for estimating the number of cluster is the gap statistic method. Other, less automatic methods, still used by some analysts, include the elbow method and average silhouette method.

5.12.3 Dimensionality Reduction

Dimensionally reduction remove redundant or highly correlated features, reducing the noise in the data. This contributes to the interpretability of the model.
There are two main techniques of dimensional reduction:

- Principal component analysis (PCA)
- Uniform manifold approximation ad project (UMAP)

In the figure is shown the application of the model mentioned before.



(a) (b) (c)

Figure 5.16: Application PCA and UMAP

Principal Component Analysis

Principal components are vectors that define a new coordinate system in which the first axis goes in the direction of the highest variance in the data. The second axis is orthogonal to the first one and goes in the direction of the second highest variance in the data. If our data was three-dimensional, the third axis would be orthogonal to both the first and the second axes and go in the direction of the third highest variance, and so on.

To describe each orange point, we need only one coordinate instead of two: the coordinate with respect to the first principal component.

When our data is very high-dimensional, it often happens in practice that the first two or three principal components account for most of the variation in the data, so by displaying the data on a 2D or 3D plot we can indeed see a very high-dimensional data and its properties.

Uniform manifold approximation and project

We must consider w:

$$w(\mathbf{x}_i, \mathbf{x}_j) \overset{\text{def}}{=} w_i(\mathbf{x}_i, \mathbf{x}_j) + w_j(\mathbf{x}_j, \mathbf{x}_i) - w_i(\mathbf{x}_i, \mathbf{x}_j) w_j(\mathbf{x}_j, \mathbf{x}_i).$$

The function w(xi,xj):

$$w_i(\mathbf{x}_i, \mathbf{x}_j) \overset{\text{def}}{=} \exp\left(-\frac{d(\mathbf{x}_i, \mathbf{x}_j) - \rho_i}{\sigma_i}\right),$$

Where:

- d (xi,xj) is the Euclidean distance between teo examples
- ρ_i Is the distance from x and the closest neighbor
- σ_i Is the distabce from x to its k

Now we can introduce the concmept of fuzzy set. Fuzzy set is the definition of membership strength. Thanks to intrudaction of the fuzze set it's possible to define the two fuzzy sets called fuzzy set cross-entropoty. The formula is:

$$C(w, w') = \sum_{i=1}^{N} \sum_{j=1}^{N} w(\mathbf{x}_i, \mathbf{x}_j) \ln \left(\frac{w(\mathbf{x}_i, \mathbf{x}_j)}{w'(\mathbf{x}'_i, \mathbf{x}'_j)} \right) + (1 - w(\mathbf{x}_i, \mathbf{x}_j)) \ln \left(\frac{1 - w(\mathbf{x}_i, \mathbf{x}_j)}{1 - w'(\mathbf{x}'_i, \mathbf{x}'_j)} \right)$$

In the following figure it is possible to see the result of dimensionality reduction applied to the MNIST:

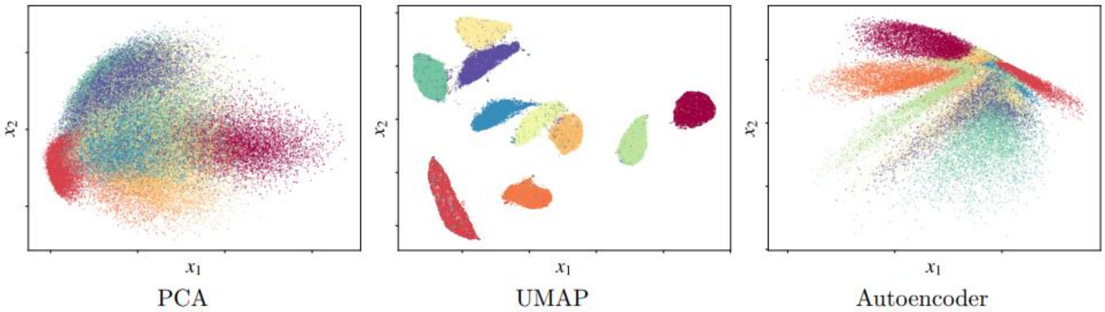


Figure 5.17 : Application of MINIST dataset.

6 MACHINE LEARNING IMPLEMENTATION AND APLLICATION

In the following chapter we are going the explore the main application of machine learning, starting from an overview of the financial sector.

6.1 Machine Learning in the financial sector

Financial services, banking, and insurance remain one of the most significant sectors that has a very high potential in reaping the benefits of machine learning and artificial intelligence with the availability of rich data, innovative algorithms, and novel methods in its various applications. While the organizations have only skimmed the surface of the rapidly evolving areas such as deep neural networks and reinforcement learning, the possibility of applying these techniques in many applications vastly remains unexplored. Organizations are leveraging the benefits of innovative applications of machine learning in applications like customer segmentation for target marketing of their newly launched products, designing optimal portfolio strategies, detection, and prevention of money laundering and other illegal activities in the financial markets, smarter and effective risk management is credit, adherence to the regulatory frameworks in finance, accounts, and other operations, and so on. However, the full capability of machine learning and artificial intelligence remains unexplored and unexploited. Leveraging such capabilities will be critical for organizations to achieve and maintain a long-term competitive edge. While one of the major reason for the slow adaption of machine learning models and methods in financial applications is that the algorithms are not well known n and there is an inevitable trust deficit in deploying them in critical and privacy-sensitive applications, the so-called “black-box” nature of such models and frameworks that makes analysis of their internal operations in producing outputs and their validations also impede faster acceptance and deployment of such models in real-world applications.

6.1.2 Main application of machine learning in finance

With the increasing availability and declining cost for complex models executing on high-power computing devices exploiting the unlimited capacity of data storage, the financial industry is geared up to exploit the benefits of machine learning to leverage a competitive business edge. While some of the use cases have already found their applications in the real world, others will need to overcome some existing business and operational challenges before they are deployed. Some of the applications are mentioned below.

Risk Modeling

The extensive domain of risk modeling and management is one of the major applications. The risk modelling credit and market is a critical application of machine learning, but a non-finance application such as operational risk management, compliance, and fraud management is also quite important. Most of the classification approaches and modeling techniques in machine learning such as binary logistic regression, multinomial logistic regression, linear and quadratic discriminant analysis, and decision trees, etc., are the foundational building blocks of applied modeling in the real world.

In data science application the risk and availability of the data play a pivotal role.

Hence, in data-rich applications such as credit risk modeling and scoring, designing mortgage schemes, the machine learning models have already made substantial inroads in comparison to scenarios such as low default credit portfolios for well-known parties that lack the availability of data. Fraud analytics remains another intensive area of Machine Learning applications in the non-financial domain.

Portfolio Management

The portfolios are designed based on the recommendations of algorithms the need to optimize various parameters with return and risk being the two most important ones.

Using the information provided by the users such as their ages of retirement, amount of investment, etc., and other associate details such as their current ages, current assets at hand, the algorithm allocate the invested amount into diverse asset classes to optimize the return and the risk associated with the portfolio.

When the initial location is made, the algorithm monitors the market environment and changes the allocation so that the portfolio remains at its optimized level always. These AI-enabled portfolio managers, known as the robo-advisors are increasingly being used in real-world portfolio design due to their superior adaptability and optimization skill to their human counterparts.

Algorithmic trading

Algorithmic trading exploits the use of algorithms to carry out stock trading in an autonomous manner with the minimal human intervention.

This methodology was benne invented in 1970, Machine Learning has pushed algorithmic trading into a new dimension where not only advanced trading strategies can be made very fast but also deep insights can be made into the stock price and overall market movements.

Machine Learning playing an increasingly important role in calibrating high-frequency, high-volume trading decisions in Realtime for critical applications.

Fraud detection and analysis

They are one of the critical applications in the finance sector. There is an increased level of security and privacy risk associated with sensitive information both from the organization and personal front due to ubiquitous availability of connectivity, high computational power in devices, an increased amount of data stored and communicated online. These issues have changed the way online fraud analysis and detection are being made. While detection in earlier days used to depend on matching a large set of complex rules, the recent approaches are largely based on the execution of learning algorithms that adapts to new security threats making the detection process more robust while being agile.

Loan and insurance underwritings

At large banks and insurance firms, the availability of historical data of consumers, financial lending/borrowing information, insurance outcomes, and default-related information in paying debts, can be leveraged in training robust machine learning models. The learned patterns and trends can be exploited by the learning algorithms for lending and underwriting risks in the future to minimize future defaults. The deployment of such models can lead to a paradigm shift in business efficiency and profit. However, at present, there is a limited utilization of such models in the industry as their deployments are largely confined within large financial institutions.

Financial chatbots

It is the definition of automation in the finance industry. Accessing the relevant data, machine learning models can yield an insightful analysis of the underlying patterns inside them that helps in making effective decisions in the future. In many cases, these models may provide recommended actions for the future so that the business decision can be made in the most efficient and optimum way. In this case also an artificial intelligent system

Risk management

Risk management examples are diverse and infinite ranging from deciding about the amount a bank should lend a custode or how to improve compliance to a process or the way risk associated with a model con be minimized

Asset price prediction

The price of an asset is affected by numerous factors driven by the market including speculative activities. In the classical approach, an asset price is determined by analyzing historical financial reports and past market performances. With rich data available, of late, machine learning-based models have started playing significant roles in predicting future asset prices in a robust and precise manner.

Derivative pricing

Machine Learning models have got rid of the requirement of such assumptions as they attempt to fit in the best possible function between the predictors and the target by minimizing the error. Accuracy and the minimal time needed for the deployment of the models in real-world use cases make machine learning the most impactful paradigm in the task of derivative pricing.

Money laundering

With the machine learning models is possible to find applications in detecting money laundering activities minimal false-positive cases.

Several new capabilities and approaches and frameworks in machine leaning and data science have become available to the modelers and engineers for all disciplines including finance professionals and research. We are going to list the main important:

- Virtual agents: The machine learning paradigm will witness the increasing deployment of agents in various tasks. These agents have the capability of performing complex data mining tasks through a large set of policy rules, defined procedures, and regulations, and provide automated responses to queries.

- Cognitive robotics: The robots in the cognitive domain have the power of automating several tasks which are currently done by humans. This automation comes with an additional level of sophistication, speed, and precision in performing the tasks.

- Text analytics: The applications of sophisticated algorithms, frameworks, and models of natural language processing in analyzing voluminous and complex financial contracts and documents help processing and decision making faster and more accurately with minimal associated risks.

- Video analytics: Advancements in the fields of computer vision, image processing, speech processing, and speech recognition together with the exponential growth in hardware capabilities have led to very promising progress in compliance, audit, model validation in various financial applications including automated generation and presentation of financial reports.

There are many models that can be used in the financial sector, the main important and influence are the following

Sparsity-aware learning
Sparsity-aware learning has evolved as an alternative model regularization approach to address several problems that are usually encountered in machine learning Considerable effort has been spent in designing schemes such frameworks in an iterative manner in solving estimation tasks of model parameters avoiding overfit. Sparsity-aware learning systems are well-suited in financial modeling applications leading to extremely robust and accurate models for various applications in finance.

Reproducing kernal Hilbert Spaces

Reproducing Kernel Hilbert Spaces (RKHS) is essentially a Hilbert space function that evaluates a continuous function in the linear space. These functions find important applications in statistical learning as every functional representation in RKHS represents minimization of an empirical function embodying the associated risk, and the representation is made as a linear combination of the data points in the training set transformed by the kernel function. Accordingly, RKHS has a very high potential in risk modeling and evaluation in finance.

Monte Carlo simulation

This method of modeling provides the modeler with a large range of possible outcomes and probabilities that they will occur for any choice of action that is taken. It is used in a diverse set of domains like finance, energy, project management, and monitoring, research and development, and insurance. It performs risk analysis by designing models of possible results by substituting a range of values – a probability distribution – for any factor that has inherent uncertainty. The ability in handling uncertainty makes this approach particularly popular in modern-day modeling in finance.

Graph theory

Multivariate financial data pose a very complex challenge in processing and visualization in addition to being difficult in modeling. Graph theory provides the modeler with a very elegant, efficient, and easily interpretable method of handling multivariate financial data.

Particle filtering

Particle filtering is a method of modeling nonlinear and non-Gaussian systems with a very high level of precision. Its ability to handle multi-modal data makes it one of the most effective and popular modeling techniques in many fields including finance. Stated in simple words, particle filtering is a technique for identifying the distribution of a population that has a minimum variance by identifying a set of random samples traversing through all the states to obtain a probability density function that best fits into the original distribution and then substituting the integral operation on the function by the mean of the sample.

Parameter learning and convex paths

While optimization methods have been proved to be very effective in training large-scale deep neural networks involving millions of parameters, the regularization of these methods has become of paramount importance for proper training of such networks.

Accordingly, intensive work has been also carried out in estimating the biases associated with the optimum value of the objective function arrived at by the algorithms. The estimation of such biases provides the modeler with an idea about the degree of inaccuracy in the models for critical applications including financial modeling.

Deep learning and reinforcement learning

The application of machine learning in finance has largely been manifested in the form of models built on deep neural network architecture and smarter algorithms for the optimization and training of such networks. Reinforcement learning-based models have made the automation of such models a reality. A vast gamut of applications, such as algo trading, capital asset pricing, stock price prediction, portfolio management can be very effectively designed and executed using deep learning and reinforcement learning frameworks.

Looking to the future of the machine learning in the financial sector there are some initial challenges like:

- Data challenges: While the availability of data in finance is quite plenty, the time series data in finance (e.g., stock prices) are quite small for data-hungry machine learning and deep learning models. Models built on limited time series data are naturally less trained and improperly designed. The result is a sub-optimal performance of the models. Another problem in finance is that financial data cannot be synthesized unlike images in the fields of computer vision and image processing. Since finance data cannot be synthesized, one must wait for financial data to be produced in the real world before using them in model training and validation. The third challenge with financial data is the high level of noise associated with high frequency trading data. Since high-frequency data in finance are invariably associated with a high level of noise, the machine learning models trained on such noisy data are intrinsically imprecise. Data evolution in finance poses the fourth challenge. Unlike data in most other fields, where the data features do not change with time, the features in financial data in harmony with financial markets evolve and change with time. This implies that financial variables will not carry the same meaning and significance over a long period, say one decade. The changes in the semantic meaning and significance of financial variables make it particularly difficult for the machine learning model to derive a consistent explanation and learning over a reasonably long period of time.

- Black-box nature of the models: Machine learning and artificial intelligence-based models are black-box in nature. In these models, while the outputs from the model are available and most of the time, they are easily interpretable, the biggest shortcoming is their lack of power of explanation of the output. In many critical applications in finance, mere outputs are not sufficient, and strong logical support for explaining the output is mandatory to instill sufficient confidence in the minds of the decision-makers. In absence of such explainable characteristics of the machine learning models, it will always remain a difficult job for the modelers to advocate the suitability of such models in critical business use cases.[19] (Jaydip Sen, Rajadeep Sen el all 2021)

- Validation challenges of the models: Due to their higher complexity and opaqueness in operation, the machine learning models pose some significant challenges to risk management and validation [29] (Fabiano B et all 1995). While the regulators demand the machine learning models to comply with the SR 11-7 and OCC 2011-12 standards of risk management, the optimum execution of the models may not be possible if all those guidelines are to be strictly adhered to. Model risk is an event that occurs when a model is designed following its intended objective, but it introduces errors while in execution yielding inaccurate results from the perspective of its design and business use case. Another manifestation of model risk can happen when a model is built and deployed inaccurately or with faults without proper knowledge about its limitations and shortcoming.

- Challenges in model testing and outcome analysis: The performance of a model and its accuracy in testing are evaluated by outcome analysis. Since the neural network model has a natural tendency to overfit or underfit the data based on the training, it is imperative on the part of the model evaluation team to address the biasvariance trade-offs in the training and validation process. Since the traditional k cross-validation procedure used in backtesting of predictive models does not work effectively for machine learning model validation, the machine learning model validators should take particular care in carrying out normalization and feature selection before model training and validation. Validation loss and accuracy should be strictly monitored and analyzed to ensure that no overfitting or underfitting is present in the model before it is sent to the production phase. Neural network models are also difficult to evaluate on their sensitivity analysis as these models lack explain ability. Since establishing a functional relationship between the explanatory variables and the target variable in a neural network model is a difficult task unlike statistical models, sensitivity analysis between the inputs and the output may involve a computationally involved challenge while the results of the analysis may be quite complex.

- Challenges with models designed by vendors: As per the requirements specified in SR 11-7 and OCC 2011-12 standards, models supplied by vendors are required to adhere to the same rigor as internally developed models. However, in many practical situations, due to proprietary restrictions testing of vendors supplied models becomes a challenge. For vendor-supplied models, banks and financial institutions will have to mostly rely on softer forms of validation. The softer form of validation may include periodic review of model performance and conceptual soundness, stringent assessment of model customization, review of the development process, and applicability of the model in the portfolio of operations of the bank.

All domains including finance, the major cause that contributes to the risk in a machine learning model is the complexity associated with the model. The machine learning algorithms are intrinsically very complex as they work on voluminous and possibly unstructured data such as texts, images, and speech.

Therefore, training of such algorithms demands sophisticated computing infrastructure and a high level of knowledge and skill on the part of the modelers. However, countering such complexities with overly sophisticated models can be very counterproductive. Instead of adopting a complex approach, the banks and financial institutions should understand the risks associated with their business and operations and manage them with simple model-validation approaches. The issues that will dominate the risk management landscape for machine learning models in the financial industry are:

- Interpretability of the models
- Model bias
- Extensive feature engineering
- Importance of model hyperparameters
- Production readiness od models
- Dynamic calibration of models
- Explainable artificial intelligent

In the days to come, the financial industry will show increasingly more reliance on machine learning and artificial intelligence-based emerging methods and models to leverage competitive advantages. While the regulatory and compliance will evolve into a more standardized framework, machine learning will continue to provide the banks and other financial institutions more opportunities to explore and exploit emerging applications, while being more efficient in delivering the existing services. While the emerging techniques discussed in the chapter will play their critical roles in mitigating future risks in models, they will also guide the authorities in designing effective regulations and compliance frameworks in risk-intensive applications like creditworthiness assessment, trade surveillance, and capital asset pricing. The model validation process will increasingly be adapted to mitigate machine learning risks, while considerable effort and time will be spent in fine-tuning the model hypermeters in handling emerging applications. However, banks will have more opportunities to deploy the models in a large gamut of applications, gaining competitive business advantages and mitigating risks in operations.

6.1.2.1 Machine Learning Techniques for Stock Perdition Example

I decided to add this sub chapter because it is a very significantly application of the machine learning in the finance sector.
In general, there are two stock prediction methodology:

- Fundamental Analysis: Performed by the Fundamental Analysts, this method is concerned more with the company rather than the actual stock. The analysts make their decisions based on the past performance of the company; the earnings forecast etc.
- Technical Analysis: Performed by the Technical Analysts, this method deals with the determination of the stock price based on the past patterns of the stock (using time-series analysis.)

Technical Analysis: Performed by the Technical Analysts, this method deals with the determination of the stock price based on the past patterns of the stock (using time-series analysis.), The EMH hypothesizes that the future stock price is completely unpredictable given the past trading history of the stock. There are 3 types of EMH's: strong, semi-strong, and weak form. In the weak EMH, any information acquired from examining the stock's history is immediately reflected in the price of the stock. The Random Walk Hypothesis claims that stock prices do not depend on past stock prices, so patterns cannot be exploited since trends to not exist. With the advent of more powerful computing infrastructure (hardware and software) trading companies now build very efficient algorithmic trading systems that can exploit the underlying pricing patterns when a huge number of data-points are made available to them. Clearly with huge datasets available on hand, Machine Learning Techniques can seriously challenge the EMH.

Regarding the indicator can be any of the following:

- Moving Average (MA)
- Exponential Moving Average (EMA)
- Rate of Change (ROC)
- Relative Strength index (RSI)

In this example the EMA is considered the primary indicator because of its ability to handle an almost infinite amount of past data, a trait that is very valuable in time series prediction.

EMA (t) = EMA (t-1) + alpha * (Price (t) - EMA (t-1))

Where: alpha = 2/ (N+1), Thus, for N=9, alpha = 0.20

In theory, the Stock Prediction Problem can be considered as evaluating a function F at time T based on the previous values of F at times t-1,t-2,t-n while assigning corresponding weight function w at each point to F.

$$F\ (t) = w1{*}F\ (t\text{-}1) + w2{*}F\ (t\text{-}2) + \ldots + w{*}F\ (t\text{-}n)$$

In the following picture is shown how the EMA models the actual Stock price:

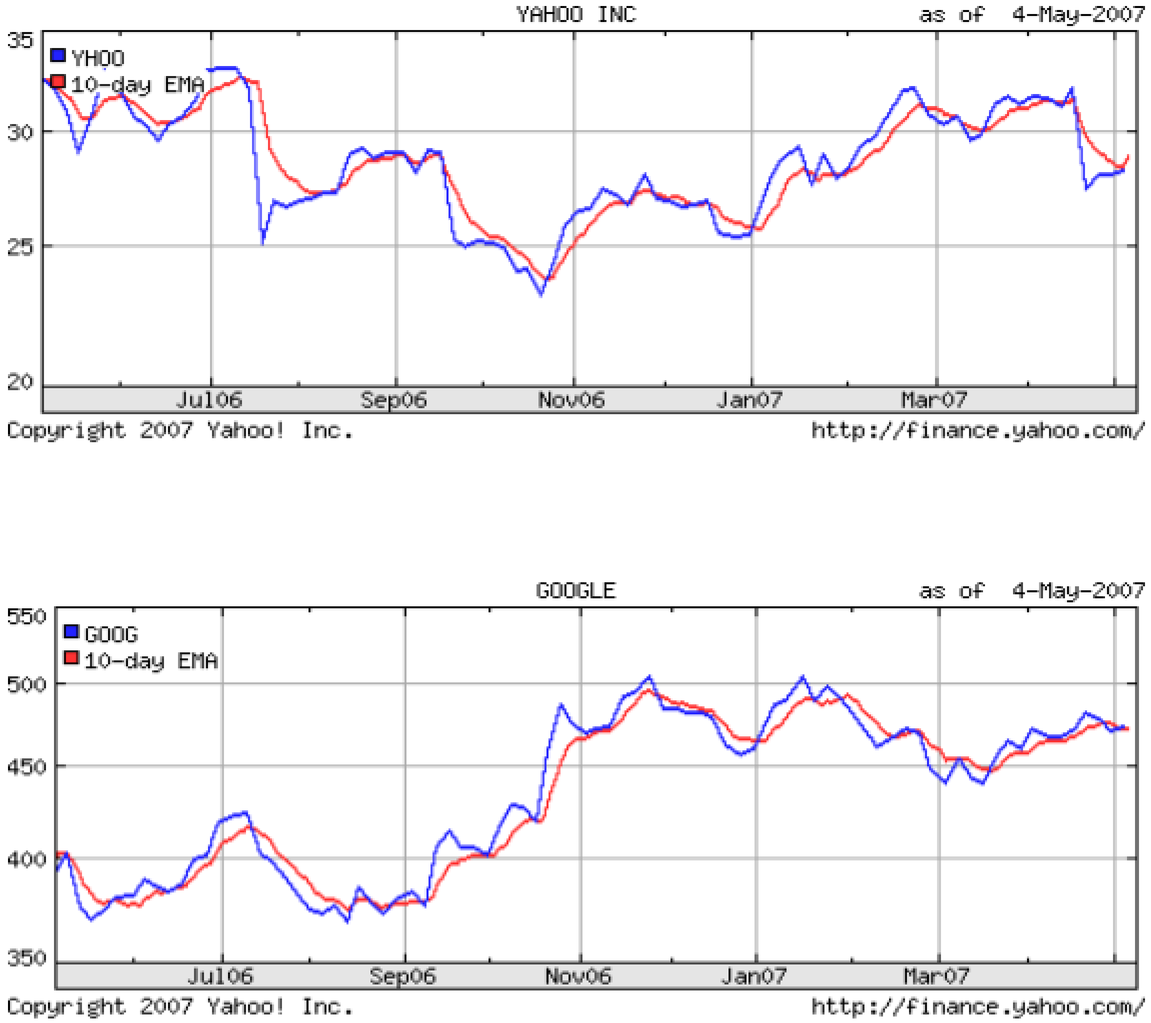


Figure 6.1: EMA application for the actual Stock Price

6.1.2.2 Application on Credit Scoring Using Machine Learning: Case of Morocco

The credit scores are useful for the banks to measure the creditworthiness of the customer using a numerical score. In the case of Marrocan banks the risk management depertaments was use to applyinhg old statistical model to decide where or not to grant credit to a certain client. Thanks to the machine learning tacnology has been found out that the problem of scoring a client can be optimized.

To optimaize the moodily the model logistic regression was used. It consists of classifying data or output variables into binary or nominal extremes by creating a logarithmic line to distinguish them. It was developed by David Cox and is one of the most frequently used statistical model in credit scoring. The logistic regression predicts the probability that the response variable belongs to a certain class, and this probability is continuous and vary between 0 and 1.

In Credit Scoring, this is the probability sought of according a Loan or not. Logistic regression consists of creating a mathematical model of a set of explanatory or predictor variables, to predict a log its transformation of the dependent variable.

The logistic regression equation is:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Where β are the regression coefficients that are to be estimated from the data during the training phase. Now is possible to calculate the probability with the inverse function: logistic.

Since the logistic regression lacks adaptability according to the points mentioned above, this study proposes the application of recent non-parametric algorithms like Random Forest, known for their accuracy and robustness, for Credit Risk Scoring.

The performance of machine learning models in the case of credit scoring has already proven in several applications of foreign banks like Bank of America and Wells Fargo. In fact, instead of a usual statistical model, we decided to compare it with the Random Forest algorithm. We chose to establish this comparison based on the same key concepts of the Random Forest model namely: Bootstrapping and Feature sampling, explained in the following. This study aims to propose the use of machine learning algorithms for the case of scoring credit risk in Moroccan banks.

Other application is the random forest algorithm. In this case the concept is to building multiples decision trees out of random samples of the data (Bootstrapping) and a random selection of variables/features (Feature Sampling). [20] (Nabila hoamdoum el all 2019)

It is a non-parametric model, which means that it possible to have no assumptions about the form or parameters of a frequency distribution.

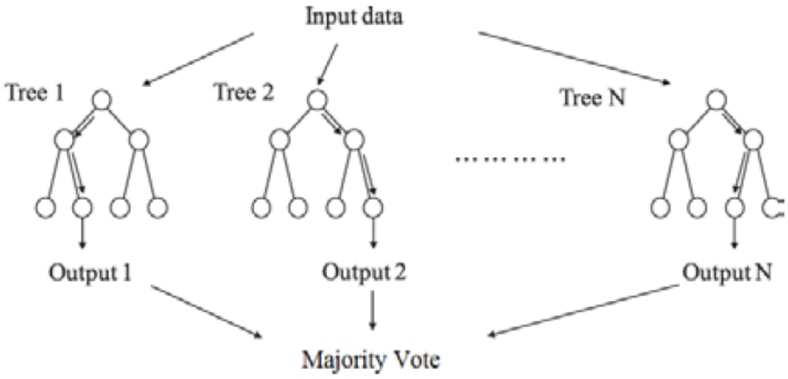


Figure 6.2: Illustration of random Forest Model

To explain the case study, we applied Random Forest and Logistic Regression algorithms on the same datasets using various features at each step. The goal is to compare the resulting accuracy for each run to prove that a nonparametric algorithm has better performance than a statistical model in Credit Scoring.

To build the Dataset the real data from Moroccan bank customers has been usded.

Datasets following specific characteristic of population. Each dataset includes more than 7 selected features randomly. The common point is that all datasets contain the target variable “Credit Risk” which is a Boolean variable returning two values: 1 if the individual is not solvent and the credit will not be granted, and 0 if he is not risky. To adapt Logistic Regression to give relevant results, we have encoded all categorical variables, using One-Hot-Encoding during the Feature Engineering step.

For the implementation phase, IBM SPSS Modeler is a software suitable for Feature Engineering and modeling applications. Both logistic regression and random forests are available in the software. The comparison of the results obtained from each algorithm was implemented on IBM SPSS Statistics which provided a wide choice of statistical test. We used a Partition to avoid over fitting by splitting data: 70% for training the model, and 30% for testing the model. In the Classification step, for every task to be performed there is a separate node, and the process consist to connect nodes to create a stream or a flow chart as can be seen in Figure 6.3.

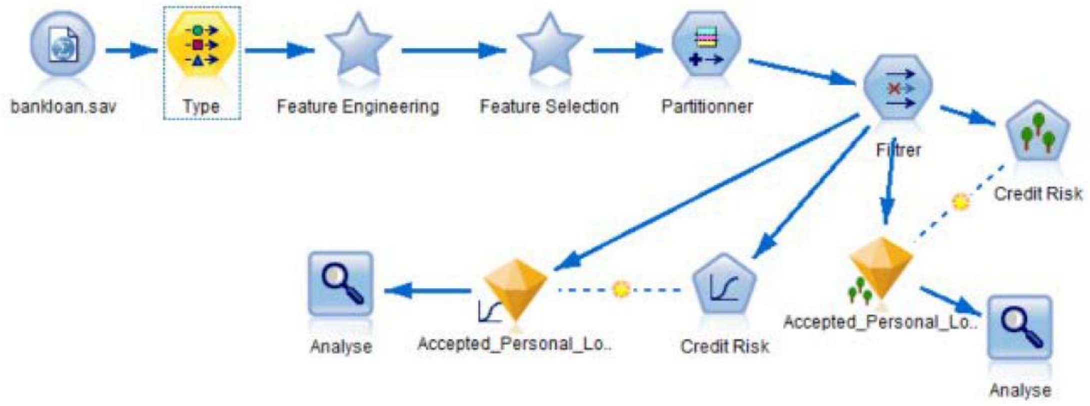


Figure 6.3 SPSS Modeler Strea, Related to tp the Steps Follwed in the study

In the following table is possible to see the caparison between two algorithms:

Datasets	Number of Obs	Number of Random Feature	Accuracy of Logistic Regression	Accuracy of Random Forest
1	3500	8	81,71%	93,86 %
2	2500	14	80,31%	88,76%
3	1900	10	77,33%	84,73%
4	3800	7	87,2%	87,51%
5	3900	9	86,91%	89,44%

The mean performance, across all datasets, of logistic regression is 82.96% and the mean performance of random forest is 88.86%. To compare the results of the two algorithms, we applied a student's t-test that justifies whether the difference in mean between the accuracy of the two algorithms is statistically significant or only due to chance.

The test statistic for equal variance is F = 0.993 with p = 0.348. With this p-value we choose to not reject the null hypothesis of equal variance, therefore, the test assumes equal variance. Then, the t statistic obtained is t = −2.542 with corresponding p-value p = 0.036. The significance level α is a measure of our confidence that the results are significant and not random. The p-value obtained is smaller than 0.05 which is the most common significance level. Thus, we reject the null hypothesis that there is no difference between the two performances.

To sum up, the difference between mean performance of Random Forest and Logistic Regression is Statistically Significant. Therefore, we may conclude that the Random Forest model performs better than the Logistic Regression model in Credit Scoring.

6.2 Machine Learning application

The following figure is a great explanation of the application of the machine learning.

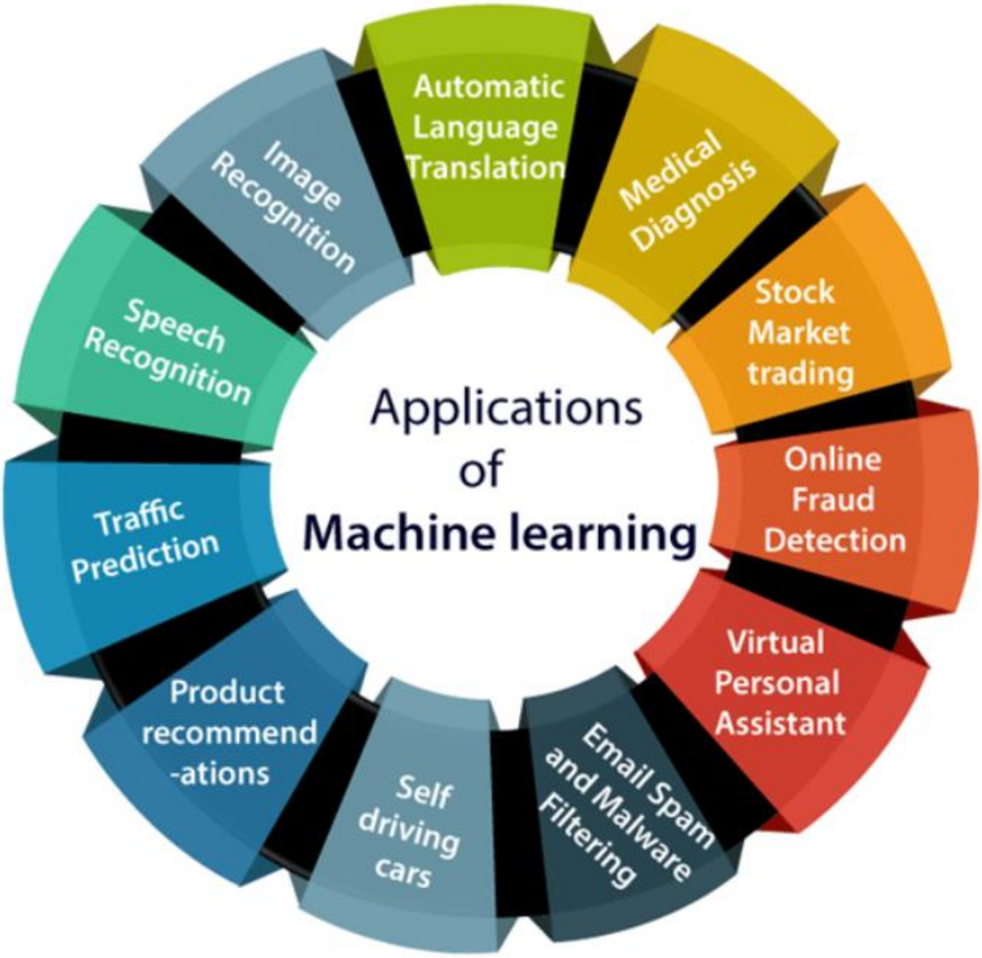

Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

Figure 6.4 Machine Learning Applications

In the following lines is shown a short explanation of the main applications of Machine Learning

Image Recognition
It is one of the most common applications of machine learning. Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a

photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's face detection and recognition algorithm.
In the details it is based on Facebook project named Deep Face.

Speech Recognition
The typical example is the voice search of Google.
Speech recognition is a process of converting voice instructions into text, and it is also known as "Speech to text", or "Computer speech recognition." At present, machine learning algorithms are widely used by various applications of speech recognition. Google assistant, Siri, Cortana, and Alexa are using speech recognition technology to follow the voice instructions

Traffic Prediction
The traffic condition is based in two frameworks:

- Real Time location
- Average time. Taken on past days at the same time

Everyone who is using Google Map is helping this app to make it better. It takes information from the user and sends back to its database to improve the performance.

Products recommendations
Companies such as Amazon, Netflix, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning.
Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest.

Self – Driving cars
The most important company in this sector is Tesla. It is using unsupervised learning method to train the car models to detect people and objects while driving.

Email Spam and Malware Filtering
Some machine learning algorithms such as Multi-Layer Perceptron, Decision tree, and Naïve Bayes classifier are used for email spam filtering and malware detection.

Virtual Personal Assistant
These virtual assistants use machine learning algorithms as an important part. These assistants record our voice instructions, send it over the server on a cloud, and decode it using ML algorithms and act accordingly.
The most important examples are:

- Google assistant
- Alexa
- Cortana
- Siri

Online Fraud Detection
For each genuine transaction, the output is converted into some hash values, and these values become the input for the next round. For each genuine transaction, there is a specific pattern which gets change for the fraud transaction hence, it detects it and makes our online transactions more secure.

Medical Diagnosis
In this sector the machine learning is used for diseases diagnoses. It can build 3D models that can predict the exact position of lessons in the brain.

Automatic Language Translation

Goodes's GNMT (Google Neural Machine Translation) provide this feature, which is a Neural Machine Learning that translates the text into our familiar language, and it called as automatic translation.
The technology behind the automatic translation is a sequence-to-sequence learning algorithm, which is used with image recognition and translates the text from one language to another language.
Regarding the industry I decided to esamited 17 industry the use the benefit of Machine Learning. I the following it is possible to see the company:

- Trading Technology

- Capital One. It uses machine learning to detect, diagnose and remediate anomalous app behavior in real time. It also uses the technology as part of its anti-money laundering tactics to adapt quickly to changes in criminals' behaviors.
- Fit Analytics. It uses machine learning to make recommendations on the best-fit styles.
- Amazon. It offers free machine learning services and products such as its Amazon Sage Maker to help developers and data scientists build, train, and deploy ML models.
- Netflix. It uses machine learning to analyze the viewing habits of its millions of customers to make predictions on which streaming video shows you may likely enjoy the most and make recommendations based on those predictions.
- Yelp. It relies on machine learning to sort through tens of millions of photos users upload to its site and then uses the technology to group them into various categories, such as, food, menus, inside the establishment or outside photos.
- Waymo. Self-driving vehicles business
- Duolingo. Data collected from your answers undergoes Duolingo's statistical model that predicts how long you will remember a certain word before needing a refresher course. Duolingo, as a result, knows when to ping you with a suggestion to retake the course.
- ASOS. Machine learning aids Asos in determining which customers are likely to continue buying its products and which customers are likely to have low CLTV, which in turn could affect Asos offering them free shipping or other promotions.
- Deserve. It uses machine learning for its creditworthiness assessment technology.
- OPTIMAL. Marketing automation platform company Optimail uses machine learning to evaluate a treasure trove of customer data to determine the best time to send them an email
- Twitter. Twitter's machine learning now ranks tweets with a relevance score based on what you engage with the most and other metrics. High ranking tweets are placed at the top of your feed, so you're more likely to see them.
- Blue river technology. The ida is to differentiate between crops and weeds, as well as achieve proper spacing between plants
- Quora. It is a social media question and answer website. The company ranks answers based on results from its machine learning, such as

thoroughness, truthfulness, and reusability, when seeking to give the best response to a question.
- Civis Analytics.
- Label Isight. It’s product metadata platform uses machine learning to give a personalized view of each food product, such as ingredients, suppliers, and supply chain history, which, in turn, aims to help you decide whether to purchase the item.
- HubSpot.It’s product metadata platform uses machine learning to give a personalized view of each food product, such as ingredients, suppliers, and supply chain history, which, in turn, aims to help you decide whether to purchase the item.

The idea of this list is to have an idea how much machine learning is implement in the system of the company. There is different way to use machine learning and the benefit is different depending on the case of each single company.

7 MACHINE LEARNING CASE STUDY

This is chapter contain the most important part of this thesis. Thanks, a great collaboration between university of genova and Alpha Beta Company I had the opportunity to analyze how a machine learning was implemented in a real company. The main business of Alphabeta is the implementation of machine learning and selling the result of the calculation.

The business in found on machine learning and the concept of the company is based on machine learning.

In this environmental there are different professionals that work together with differed capability but with the same goal.

In the next page I will introduce in the detail the company.

7.1 Alpha Beta Company

Alpha Beta is an Israeli Fintech company that specializes in developing quantitative investment strategies and technological asset management tools.

Founded in 2010 by Koby Shemer, co-founder of Analyst IMS Investment (TLV: ANLT), AlphaBeta develops innovative indices and investment strategies which are backed by decades of academic research combined with extensive knowledge and experience.

The company's activities are based on the constant evaluation of factors that drive the various markets in the world, resulting in the formation of practical strategies that can be applied to customer portfolios.

The organizational culture and intellectual curiosity keep us innovative and make us stand out. By challenging long-standing beliefs and conducting in-depth research, they transform theory into practical investment tools.

Alpha Beta collaborates with leading investment houses which execute the innovative investment strategies and asset management methods.

In addition, at the RPS FUND LP (by Alpha-Beta), they harness various strategies to achieve Absolute Return with a very low correlation (Beta) to market returns.

The team members are:

- Koby Shemer. Funder
- Shmuel Oderberg. General Menager
- Ron Shemer. Founder and RPS Hedge Fund Manager

- Idan Schatz. Research Team Leader
- Lital Shemer. CMO – RPS Hedge fund
- Vasil Anoshin. Software Engineer
- Oded Zimerman. Senior Researcher
- Oleg Dizengof. Researcher and platform manager
- Hanan Libhaber. Researcher
- Guy Lavi. Trading Manager

The main asset of the company is to use in-house technological tools and vast amount of data to constantly examine factors that drive markets around the globe. The investment strategies are based on decades of academic research which we combine with the accumulated practical experience. There are many collaborations with leading investment houses which implement the unique strategies and methods and manage the propriety hedge fund.
Lingering in the different wokres listed before is interesting to see the position. This company is exactly a perfect mix between research, training, technology, and assets manager.

AphaBeta has a proletary platform that it is based on:

- Artificial Intelligent
- Propriety strategy
- Established brand and knowhow

This tecnoly is useful for offers alternative investment solutions using various investment strategies. Their goal is to try and achieve excess returns over the traditional indices. This indicate are AlphaBeta propriety.
The company has an artificial intelingent asset pricing lab. Thanks to this lab AlphaBeta has activated several collaborations with academic partners on research and development of AI-based technologies to create innovative investment products.
In the following figure is shown a reassumed that is useful to understand how the AphaBeta platform works.

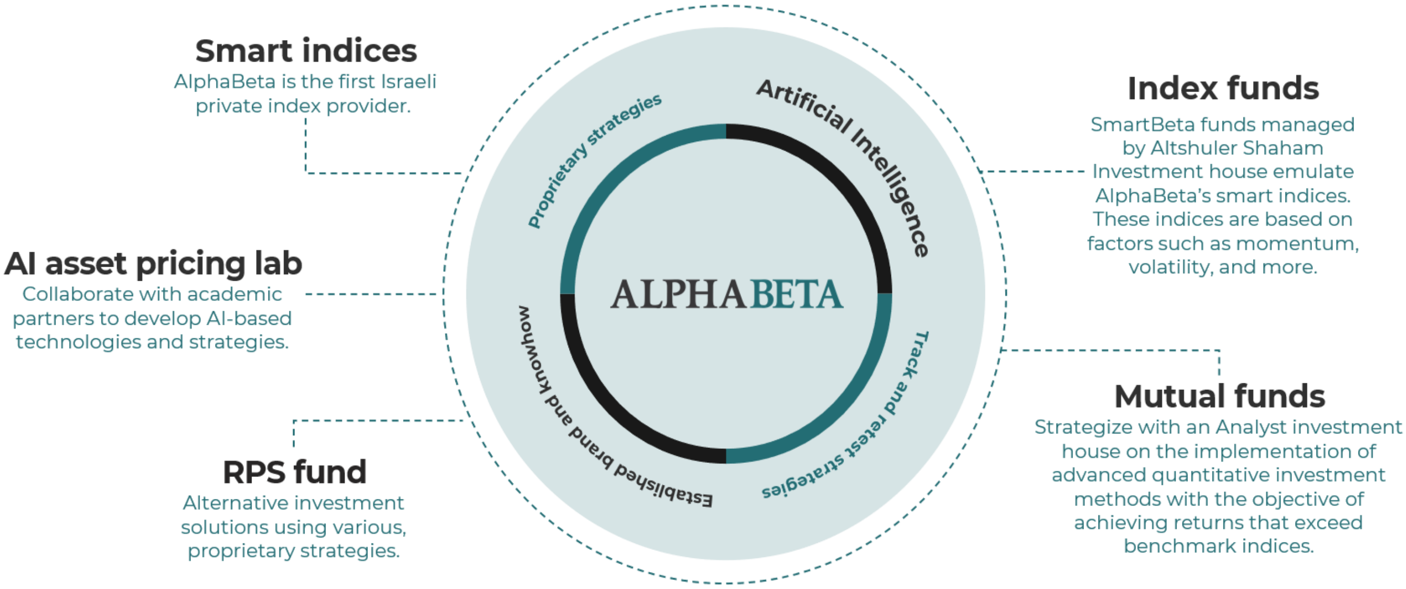

Smart indices
AlphaBeta is the first Israeli private index provider.
AI asset pricing lab
Collaborate with academic partners to develop AI-based technologies and strategies.
RPS fund
Alternative investment solutions using various, proprietary strategies.
Index funds
SmartBeta funds managed by Altshuler Shaham Investment house emulate AlphaBeta's smart indices. These indices are based on factors such as momentum, volatility, and more.
Mutual funds
Strategize with an Analyst investment house on the implementation of advanced quantitative investment methods with the objective of achieving returns that exceed benchmark indices.
Proprietary strategies
Artificial Intelligence
Track and retest strategies
Established brand and knowhow
ALPHABETA

Figure 7.1 AlphaBeta Platform

The company has large infrastructure of investment factors that provide valuable signals. Each based on academic research by well-known professors that works to create an interactive descriptive statistical and back testing engine in which an analyst can construct different portfolios considering factor combinations for asset management process.

The artificial intelligent is the base to build a portfolio construction and the prediction of the financial and fundamental ratios and values of different variables in the P&L or the balance sheet.

Finally, there is a collaborate system where researchers from different disciplines and departments from computer science, statistics, and finance will be able to publish a proposal for research. The research is the fundamental concept of the company.

I know that I already repeat this concept, but I want to do in purpose. It is a real company that really believe in a research activity. It connected the revenues of the company directly with the output of the research activity.

The academic research has been successful in developing cutting-edge models that consistently, beat classic asset management methods both in terms of higher yield and lower risk.
In the end the fundamental of the company is simple:

The public does not have access to the best asset management methods that currently available.

The success is the general public’s success

In the following pictures is easy to understand the positioned to capitalize on demand and the strategy of the company

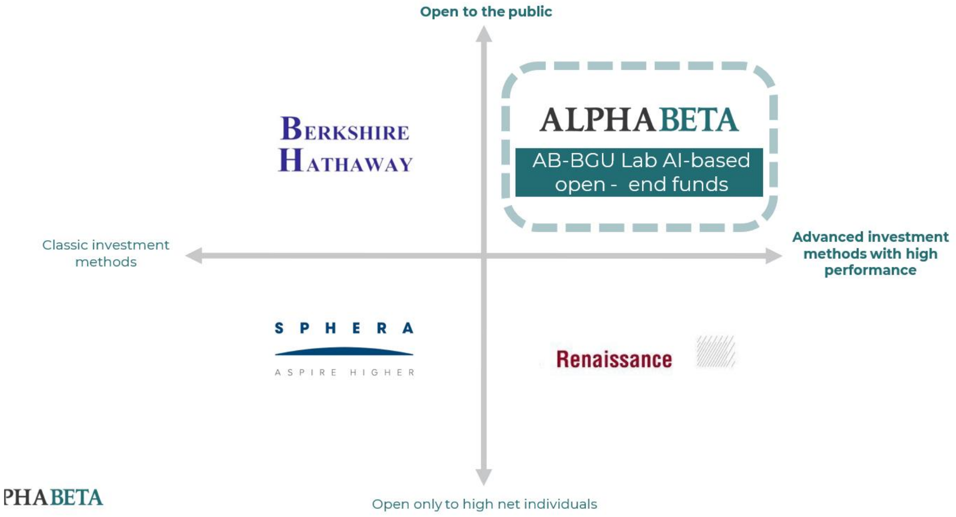

Open to the public
BERKSHIRE HATHAWAY
ALPHABETA
AB-BGU Lab AI-based open - end funds
Classic investment methods
Advanced investment methods with high performance
SPHERA
ASPIRE HIGHER
Renaissance
PHABETA
Open only to high net individuals

Figure 7.2 AlphaBeta positioned to the capital demand

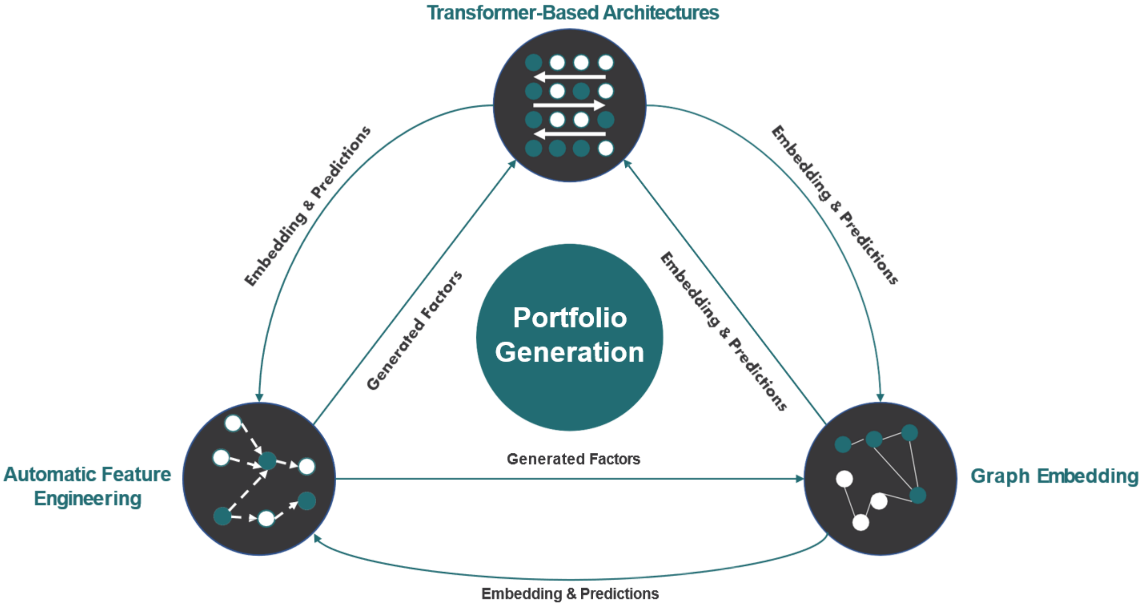


Figure 7.3 Hight level concept

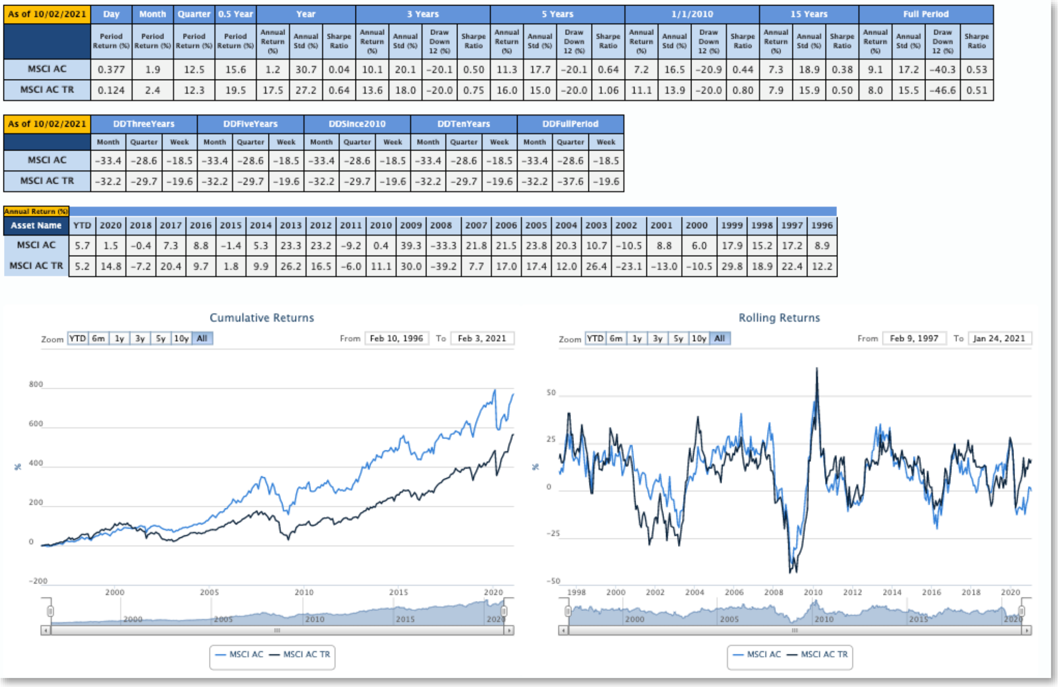
The concept works thanks the following steps

- Search. Investment managers look for advice, data, talent, and ways to surface valuable insights to enhance performance.

- Alpha Beta contribution with AI tecnology

- Discovery. Through word of mouth, clients find AlphaBeta based on the firm’s established track record and brand. The research-as-a-service further solidifies AlphaBeta as a leader in the industry, retaining clients for years.

- Research. Client performs highly quantitative research (right) to turn AlphaBeta’s bespoke knowledge into smart strategies to test and compare strategies as if they had a quant department in-house.



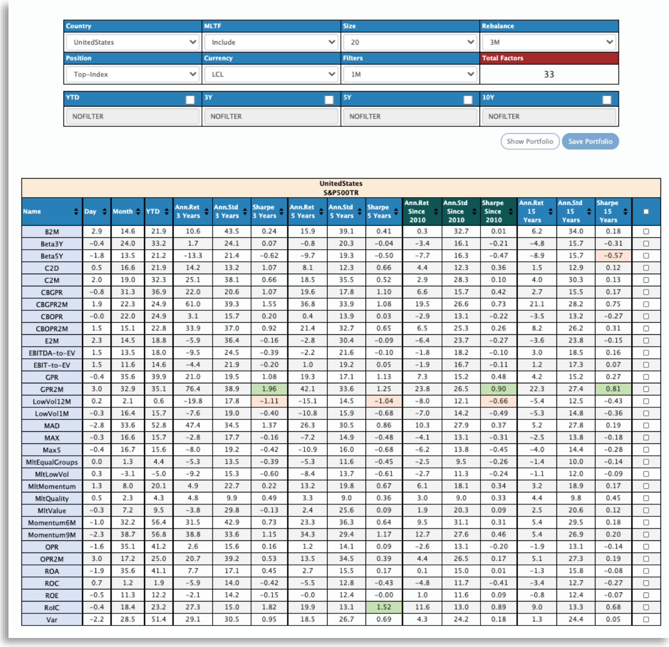
As of 10/02/2021	Day	Month	Quarter	0.5 Year	Year			3 Years				5 Years				1/1/2010				15 Years			Full Period			
	Period Return (%)	Period Return (%)	Period Return (%)	Period Return (%)	Annual Return (%)	Annual Std (%)	Sharpe Ratio	Annual Return (%)	Annual Std (%)	Draw Down 12 (%)	Sharpe Ratio	Annual Return (%)	Annual Std (%)	Draw Down 12 (%)	Sharpe Ratio	Annual Return (%)	Annual Std (%)	Draw Down 12 (%)	Sharpe Ratio	Annual Return (%)	Annual Std (%)	Sharpe Ratio	Annual Return (%)	Annual Std (%)	Draw Down 12 (%)	Sharpe Ratio
MSCI AC	0.377	1.9	12.5	15.6	1.2	30.7	0.04	10.1	20.1	-20.1	0.50	11.3	17.7	-20.1	0.64	7.2	16.5	-20.9	0.44	7.3	18.9	0.38	9.1	17.2	-40.3	0.53
MSCI AC TR	0.124	2.4	12.3	19.5	17.5	27.2	0.64	13.6	18.0	-20.0	0.75	16.0	15.0	-20.0	1.06	11.1	13.9	-20.0	0.80	7.9	15.9	0.50	8.0	15.5	-46.6	0.51

As of 10/02/2021	DDThreeYears			DDFiveYears			DDSince2010			DDTenYears			DDFullPeriod		
	Month	Quarter	Week	Month	Quarter	Week	Month	Quarter	Week	Month	Quarter	Week	Month	Quarter	Week
MSCI AC	-33.4	-28.6	-18.5	-33.4	-28.6	-18.5	-33.4	-28.6	-18.5	-33.4	-28.6	-18.5	-33.4	-28.6	-18.5
MSCI AC TR	-32.2	-29.7	-19.6	-32.2	-29.7	-19.6	-32.2	-29.7	-19.6	-32.2	-29.7	-19.6	-32.2	-37.6	-19.6

Annual Return (%)																										
Asset Name	YTD	2020	2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996
MSCI AC	5.7	1.5	-0.4	7.3	8.8	-1.4	5.3	23.3	23.2	-9.2	0.4	39.3	-33.3	21.8	21.5	23.8	20.3	10.7	-10.5	8.8	6.0	17.9	15.2	17.2	8.9
MSCI AC TR	5.2	14.8	-7.2	20.4	9.7	1.8	9.9	26.2	16.5	-6.0	11.1	30.0	-39.2	7.7	17.0	17.4	12.0	26.4	-23.1	-13.0	-10.5	29.8	18.9	22.4	12.2

Figure 7.4 Example of research activity

- Results. Client utilizes the test strategy (from above) to form a portfolio of proprietary insights to drive outsized returns and track performance across all factors



Country	MLTF	Size	Rebalance
UnitedStates	Include	20	3M
Position	**Currency**	**Filters**	**Total Factors**
Top-Index	LCL	1M	33
YTD □	3Y □	5Y □	10Y □
NOFILTER	NOFILTER	NOFILTER	NOFILTER

UnitedStates
S&P500TR

Name	Day	Month	YTD	Ann.Ret 3 Years	Ann.Std 3 Years	Sharpe 3 Years	Ann.Ret 5 Years	Ann.Std 5 Years	Sharpe 5 Years	Ann.Ret Since 2010	Ann.Std Since 2010	Sharpe Since 2010	Ann.Ret 15 Years	Ann.Std 15 Years	Sharpe 15 Years	■
B2M	2.9	14.6	21.9	10.6	43.5	0.24	15.9	39.1	0.41	0.3	32.7	0.01	6.2	34.0	0.18	□
Beta3Y	-0.4	24.0	33.2	1.7	24.1	0.07	-0.8	20.3	-0.04	-3.4	16.1	-0.21	-4.8	15.7	-0.31	□
Beta5Y	-1.8	13.5	21.2	-13.3	21.4	-0.62	-9.7	19.3	-0.50	-7.7	16.3	-0.47	-8.9	15.7	-0.57	□
C2D	0.5	16.6	21.9	14.2	13.2	1.07	8.1	12.3	0.66	4.4	12.3	0.36	1.5	12.9	0.12	□
C2M	2.0	19.0	32.3	25.1	38.1	0.66	18.5	35.5	0.52	2.9	28.3	0.10	4.0	30.3	0.13	□
CBGPR	-0.8	31.3	36.9	22.0	20.6	1.07	19.6	17.8	1.10	6.6	15.7	0.42	2.7	15.5	0.17	□
CBGPR2M	1.9	22.3	24.9	61.0	39.3	1.55	36.8	33.9	1.08	19.5	26.6	0.73	21.1	28.2	0.75	□
CBOPR	-0.0	22.0	24.9	3.1	15.7	0.20	0.4	13.9	0.03	-2.9	13.1	-0.22	-3.5	13.2	-0.27	□
CBOPR2M	1.5	15.1	22.8	33.9	37.0	0.92	21.4	32.7	0.65	6.5	25.3	0.26	8.2	26.2	0.31	□
E2M	2.3	14.5	18.8	-5.9	36.4	-0.16	-2.8	30.4	-0.09	-6.4	23.7	-0.27	-3.6	23.8	-0.15	□
EBITDA-to-EV	1.5	13.5	18.0	-9.5	24.5	-0.39	-2.2	21.6	-0.10	-1.8	18.2	-0.10	3.0	18.5	0.16	□
EBIT-to-EV	1.5	11.6	14.6	-4.4	21.9	-0.20	1.0	19.2	0.05	-1.9	16.7	-0.11	1.2	17.3	0.07	□
GPR	-0.4	35.6	39.9	21.0	19.5	1.08	19.3	17.1	1.13	7.3	15.2	0.48	4.2	15.2	0.27	□
GPR2M	3.0	32.9	35.1	76.4	38.9	1.96	42.1	33.6	1.25	23.8	26.5	0.90	22.3	27.4	0.81	□
LowVol12M	0.2	2.1	0.6	-19.8	17.8	-1.11	-15.1	14.5	-1.04	-8.0	12.1	-0.66	-5.4	12.5	-0.43	□
LowVol1M	-0.3	16.4	15.7	-7.6	19.0	-0.40	-10.8	15.9	-0.68	-7.0	14.2	-0.49	-5.3	14.8	-0.36	□
MAD	-2.8	33.6	52.8	47.4	34.5	1.37	26.3	30.5	0.86	10.3	27.9	0.37	5.2	27.8	0.19	□
MAX	-0.3	16.6	15.7	-2.8	17.7	-0.16	-7.2	14.9	-0.48	-4.1	13.1	-0.31	-2.5	13.8	-0.18	□
Max5	-0.4	16.7	15.6	-8.0	19.2	-0.42	-10.9	16.0	-0.68	-6.2	13.8	-0.45	-4.0	14.4	-0.28	□
MltEqualGroups	0.0	1.3	4.4	-5.3	13.5	-0.39	-5.3	11.6	-0.45	-2.5	9.5	-0.26	-1.4	10.0	-0.14	□
MltLowVol	0.3	-3.1	-5.0	-9.2	15.3	-0.60	-8.4	13.7	-0.61	-2.7	11.3	-0.24	-1.1	12.0	-0.09	□
MltMomentum	1.3	8.0	20.1	4.9	22.7	0.22	13.2	19.8	0.67	6.1	18.1	0.34	3.2	18.9	0.17	□
MltQuality	0.5	2.3	4.3	4.8	9.9	0.49	3.3	9.0	0.36	3.0	9.0	0.33	4.4	9.8	0.45	□
MltValue	-0.3	7.2	9.5	-3.8	29.8	-0.13	2.4	25.6	0.09	1.9	20.3	0.09	2.5	20.6	0.12	□
Momentum6M	-1.0	32.2	56.4	31.5	42.9	0.73	23.3	36.3	0.64	9.5	31.1	0.31	5.4	29.5	0.18	□
Momentum9M	-2.3	38.7	56.8	38.8	33.6	1.15	34.3	29.4	1.17	12.7	27.6	0.46	5.4	26.9	0.20	□
OPR	-1.6	35.1	41.2	2.6	15.6	0.16	1.2	14.1	0.09	-2.6	13.1	-0.20	-1.9	13.1	-0.14	□
OPR2M	3.0	17.2	25.0	20.7	39.2	0.53	13.5	34.5	0.39	4.4	26.5	0.17	5.1	27.3	0.19	□
ROA	-1.9	35.6	41.1	7.7	17.1	0.45	2.7	15.5	0.17	0.1	15.0	0.01	-1.3	15.8	-0.08	□
ROC	0.7	1.2	1.9	-5.9	14.0	-0.42	-5.5	12.8	-0.43	-4.8	11.7	-0.41	-3.4	12.7	-0.27	□
ROE	-0.5	11.3	12.2	-2.1	14.2	-0.15	-0.0	12.4	-0.00	1.0	11.6	0.09	-0.8	12.4	-0.07	□
RoIC	-0.4	18.4	23.2	27.3	15.0	1.82	19.9	13.1	1.52	11.6	13.0	0.89	9.0	13.3	0.68	□
Var	-2.2	28.5	51.4	29.1	30.5	0.95	18.5	26.7	0.69	4.3	24.2	0.18	1.3	24.4	0.05	□

Figure 7.5 How the results look like

7.2 Academic research in smart investment strategies

Thanks to Alpha Bena company we are going to present the application of machine learning methods to predict international stock returns in many markets.

In particular we follow the metodoly of Gu,Kelly, and Xiu (GKX, 2020) and use their set of protentional hyperparameter value.

We estimate the parameters and hyperparameters from the U.S. data and apply them to 31 major markets.

To the extent that international markets are not perfectly correlated with the U.S., we run

substantially more out-of-sample tests. Overall, the evidence is mixed-while neural network (NN) models work well internationally, regression trees (RTs) show signs of overfitting on some occasion.

The second part of the analysis is very useful a long- debated question in international asset pricing.

A market-specific model typically needs to be pre-specified with the knowledge of the institutional details, but machine learning can detect non-monotonic relationships and complex interactions between returns and many characteristics even without such knowledge.

In the part of examination, we examine the cross-market integration, finding a specific market-specific NN models con be further improving by adding U.S.-based variables.

The results indicate that international markets are integrated in the sense that the U.S. market is relevant for other markets.

In the GKS we made a comparations linear and non-linear models. With the following variants:

- OLS with Hunber loss function
- LASSO, that selects a subset of predictors
- RDIGE, that restricts the magnitude of the regression coefficient
- ENET. It is a connation of LASSO and RIDGE

After that we focused in in two classes of non-linear models' regression trees (RTs) and neural network (NN) models.

The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The NN models is useful for aggregate and transform amount signals into outputs.

For each test we set aside training and validation periods to train our models and select the hyperparameters, using the construct forecasts of one.month.ahead stock returns.

In total we list 36 variables including:

- Market capitalization
- Trading volume
- Past return of the industry
- Accounting information

We use the hyperparameter and parameter values estimated from the U.S. to form the predictions in all 32 markets (31 international markets plus the U.S.). Using 94 characteristics, 8 macroeconomic predictors, and 74 industry dummies (i.e., a total number of 920 (= 94 X (8 + 1) + 74) covariates).

The interpretation of GKY is that both RTs and NN models ouperforme linear models in terms of out-of- sample R^2 and long-shorts portfolio Sharpe rations.

Regarding the NN models generate large economic profits that trees and linear models in the most international markets.

However, the best tree model yields a lower out-of-sample R^2 than the best linear model in 23 of the 31 markets.

In second test of set we had to examinate an issue regarding the overfitting. We had to train and validate molded separate for each market.

GKX show that the highest equal-weighted Sharpe ratio (2.45) and value-weighed Sharpe ratio (1.35) are achieved by a NN model in the U.S. The NN is the also the most profitable model because it can generate annualized Sharpe rations that are close to or above 2 or are above 1.

The R^2 and Sharpe ratios in these market-specific NN models are usually higher than those in our U.S.-trained NN models. Generally, NN models appear to capture international return-characteristic relationships effectively.

The best tree model gives a lower value-weighted Sharpe ratio than the best linear model in 19 markets.

While we cannot provide a rule to state how much data are needed, we show evidence that the underperformance of RTs is more pronounced in markets where there are fewer stocks and a shorter time, suggesting the effectiveness of RTs heavily depends on the number of observations. We will see in the next lines the results of NN models produce higher equal-weighted and value-weighted Sharpe Rations than linear models.

NN models perform well in sorting stocks into top and bottom deciles but not as good for other deciles and in forming return predictions.

Among the 36 variables, we show that firm size, one-month return reversal, and daily return volatility are the most important predictors in the U.S., while in other large international markets some other predictors can dominate. NN models

also face lower downside risk of the maximum drawdown and the maximum one-month loss are usually lower than those of other models.

Although machine learning is powerful, this study shown that overfitting can be a problem and we should exercise caution when applying it to different markets.

Using local factors and characteristics is important, our results suggest that the relationships between returns and characteristics seem to vary across markets. The NN models provide a way to price stock returns locally with more characteristics and in a larger set of markets.

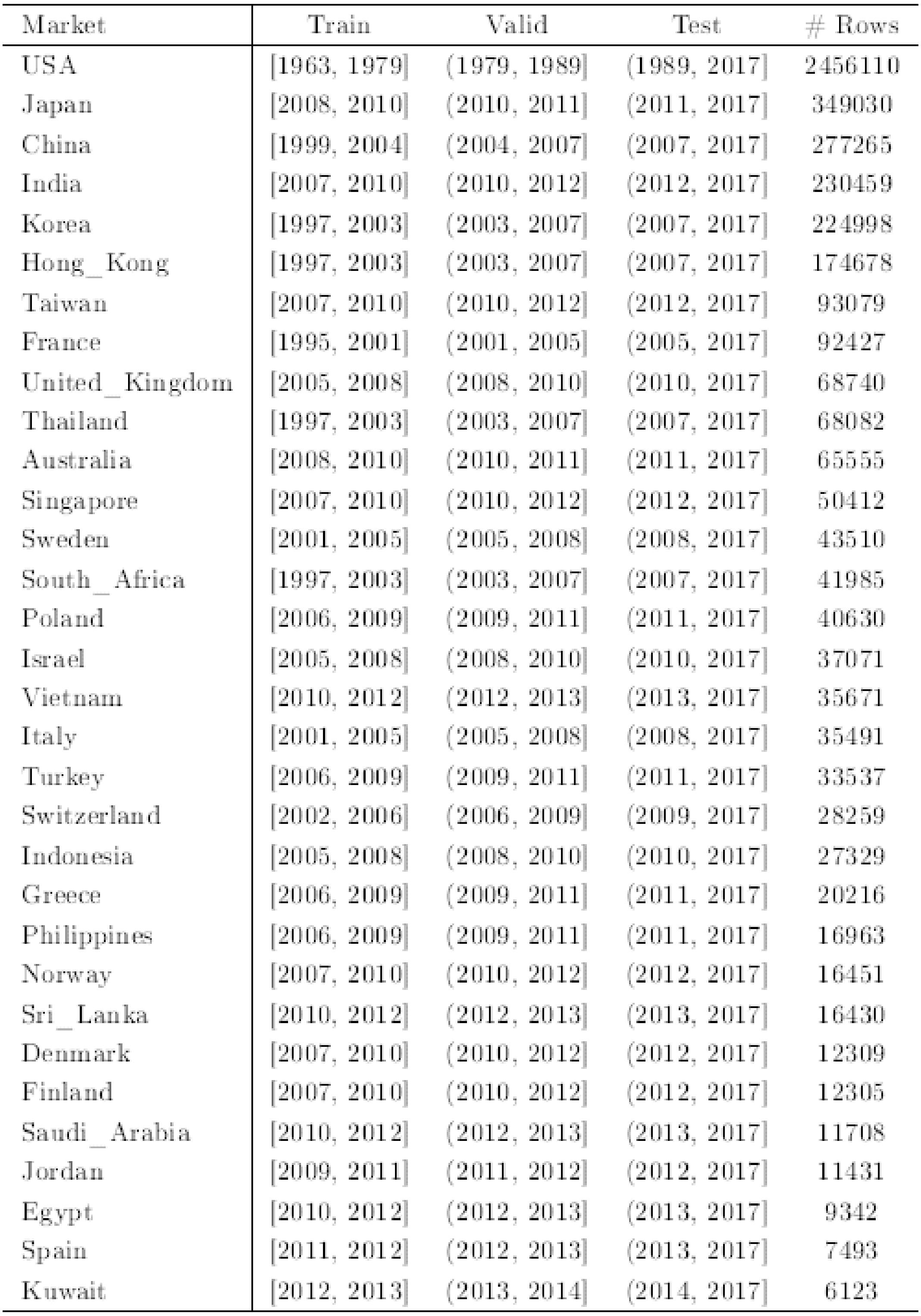
7.2.1 Methods

7.2.1.1 Data

We obtain data on stock returns, trading volume, market capitalization, and industry information from DataStream. We minorize raw returns at the top and bottom 2.5% in each exchange in each month to correct for potential data errors. Following Hou, Karolyi, and Kho (2011) and Ince and Porter (2006), all monthly returns that are above 300% and reversed within 1 month, as well as zero monthly returns, are removed (DataStream repeats the last valid data point for delisted forms). Accounting ratios (book-to-market and sales-to price) are available from Factset. For the U.S. and China, we use the data with CRSP and CSMAR, respectively, because of better coverage.

We download data for as many markets as possible and require each market to have at least 100 stocks with valid observations of return and the 36 characteristics for at least 3 years. As a result, 32 markets, including the U.S., are in the final sample. Our data range from 2.4 million stock-month observations in the U.S. to around 6,100 in Kuwait.

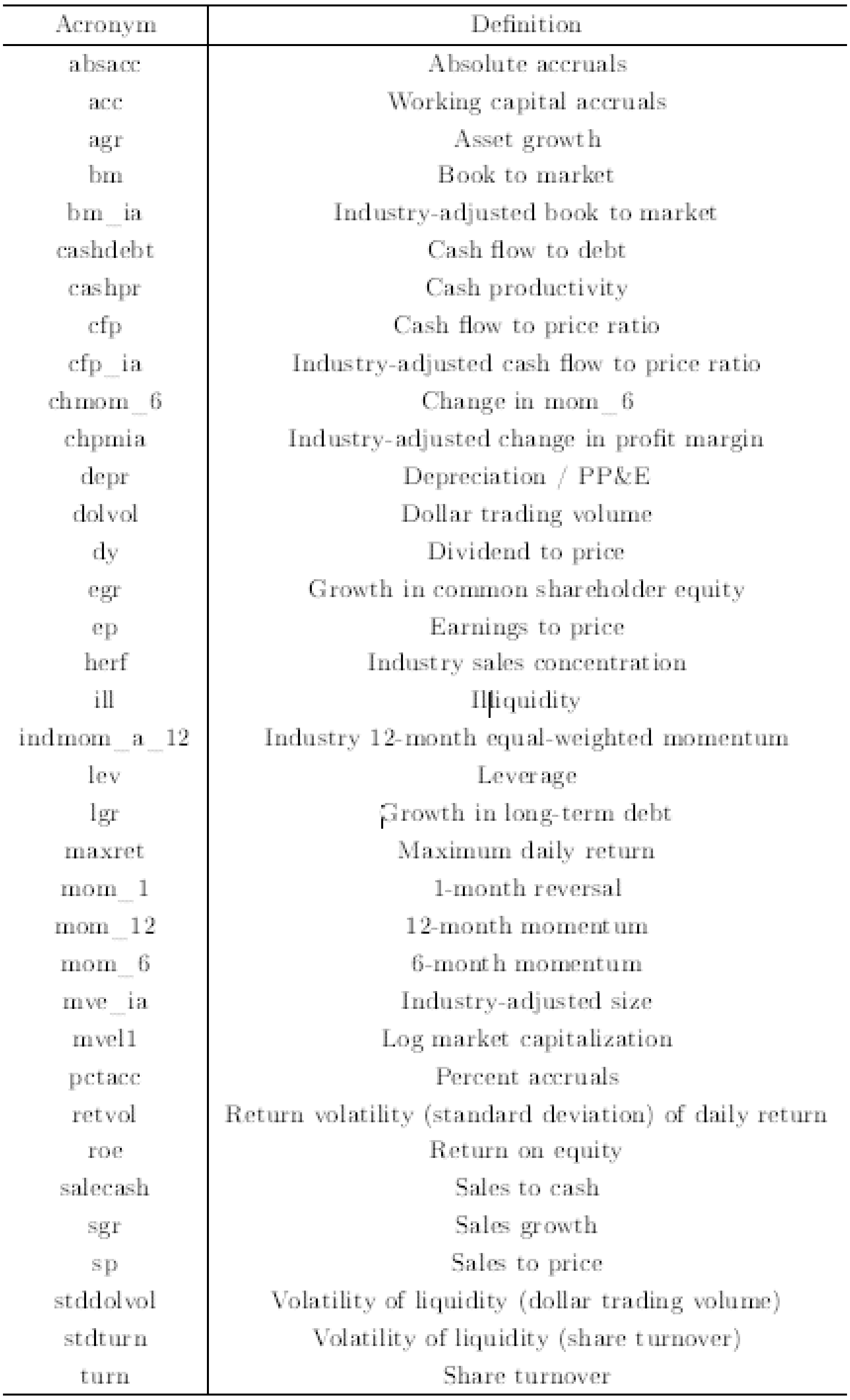
In the following table is shown the list of market:



Market	Train	Valid	Test	# Rows
USA	[1963, 1979]	(1979, 1989]	(1989, 2017]	2456110
Japan	[2008, 2010]	(2010, 2011]	(2011, 2017]	349030
China	[1999, 2004]	(2004, 2007]	(2007, 2017]	277265
India	[2007, 2010]	(2010, 2012]	(2012, 2017]	230459
Korea	[1997, 2003]	(2003, 2007]	(2007, 2017]	224998
Hong_Kong	[1997, 2003]	(2003, 2007]	(2007, 2017]	174678
Taiwan	[2007, 2010]	(2010, 2012]	(2012, 2017]	93079
France	[1995, 2001]	(2001, 2005]	(2005, 2017]	92427
United_Kingdom	[2005, 2008]	(2008, 2010]	(2010, 2017]	68740
Thailand	[1997, 2003]	(2003, 2007]	(2007, 2017]	68082
Australia	[2008, 2010]	(2010, 2011]	(2011, 2017]	65555
Singapore	[2007, 2010]	(2010, 2012]	(2012, 2017]	50412
Sweden	[2001, 2005]	(2005, 2008]	(2008, 2017]	43510
South_Africa	[1997, 2003]	(2003, 2007]	(2007, 2017]	41985
Poland	[2006, 2009]	(2009, 2011]	(2011, 2017]	40630
Israel	[2005, 2008]	(2008, 2010]	(2010, 2017]	37071
Vietnam	[2010, 2012]	(2012, 2013]	(2013, 2017]	35671
Italy	[2001, 2005]	(2005, 2008]	(2008, 2017]	35491
Turkey	[2006, 2009]	(2009, 2011]	(2011, 2017]	33537
Switzerland	[2002, 2006]	(2006, 2009]	(2009, 2017]	28259
Indonesia	[2005, 2008]	(2008, 2010]	(2010, 2017]	27329
Greece	[2006, 2009]	(2009, 2011]	(2011, 2017]	20216
Philippines	[2006, 2009]	(2009, 2011]	(2011, 2017]	16963
Norway	[2007, 2010]	(2010, 2012]	(2012, 2017]	16451
Sri_Lanka	[2010, 2012]	(2012, 2013]	(2013, 2017]	16430
Denmark	[2007, 2010]	(2010, 2012]	(2012, 2017]	12309
Finland	[2007, 2010]	(2010, 2012]	(2012, 2017]	12305
Saudi_Arabia	[2010, 2012]	(2012, 2013]	(2013, 2017]	11708
Jordan	[2009, 2011]	(2011, 2012]	(2012, 2017]	11431
Egypt	[2010, 2012]	(2012, 2013]	(2013, 2017]	9342
Spain	[2011, 2012]	(2012, 2013]	(2013, 2017]	7493
Kuwait	[2012, 2013]	(2013, 2014]	(2014, 2017]	6123

Table 7.1

We construct 36 stock characteristics. It’s possible to lists of the acronym and definition of the 36 stock characteristics used as model inputs.in the following table.



Acronym	Definition
absacc	Absolute accruals
acc	Working capital accruals
agr	Asset growth
bm	Book to market
bm_ia	Industry-adjusted book to market
cashdebt	Cash flow to debt
cashpr	Cash productivity
cfp	Cash flow to price ratio
cfp_ia	Industry-adjusted cash flow to price ratio
chmom_6	Change in mom_6
chpmia	Industry-adjusted change in profit margin
depr	Depreciation / PP&E
dolvol	Dollar trading volume
dy	Dividend to price
egr	Growth in common shareholder equity
ep	Earnings to price
herf	Industry sales concentration
ill	Illiquidity
indmom_a_12	Industry 12-month equal-weighted momentum
lev	Leverage
lgr	Growth in long-term debt
maxret	Maximum daily return
mom_1	1-month reversal
mom_12	12-month momentum
mom_6	6-month momentum
mve_ia	Industry-adjusted size
mvel1	Log market capitalization
pctacc	Percent accruals
retvol	Return volatility (standard deviation) of daily return
roe	Return on equity
salecash	Sales to cash
sgr	Sales growth
sp	Sales to price
stddolvol	Volatility of liquidity (dollar trading volume)
stdturn	Volatility of liquidity (share turnover)
turn	Share turnover

Table 7.2

Before we input all features in the model, we normalize them to zero mean and unit standard deviation by month and market.

7.2.1.2 Model Estimation, Hyperparameter Tuning, and Out-of sample Test

The model is set to predict the next month stock returns in U.S. Dollars in excesses of the risk-free rate. We made a separation the sample of the market into 34 non-overlapping parts, while maintaining their chronological order in Orde to train the model for each market.
The chronological ores are:

- Training data, which consist of the first 30% of the periods, are used to estimate the model subject to a particular set of hyperparameter values.
- Validation data, accounting for 20%, are deployed to construct forecasts and calculate objective functions based on the estimated model from the training sample. we iteratively search for the best set of hyperparameters that optimizes the objective functions.
- testing data are the remaining 50%; they are ‘’out-of-sample’’ to provide objective assessments of the models' performance after determining hyperparameters and normal parameters for the models.

We predict the returns in the next calendar year, the training data expands by one year whereas validation samples are maintained with the same size.
If we have a lock to the Table 7.1, when we predict the cross-sectional returns in 1991, the training and validation samples are [1963, 1980] and (1980, 1990], respectively.
We first train and validate the model for the U.S. market following the above-mentioned procedure. Then, we apply the U.S.-estimated models to the corresponding years of other markets.
At now we have described how we decide to ouy-of-sample test using the international data to investigate if the model overfits the U.S. data.
The second zest ph machine learning consist in to trained and validate using each market’s data with the same set of potential hyperparameter values.
If the machine learning model with its estimation and regularization techniques can truly capture the underlying data generating process, which presumably varies across markets, market-specific models should outperform the U.S.-estimated model in non-U.S. markets.

7.2.1.3 Post-estimation Evaluation

We decide to implement Sharpe rations and out-of-sample R^2 to evaluate the performance of machine learning models.
It is important understand how the Sharpe rations and out-of-sample R^2 works.

Out-of-sample R^2
We report the out-of-sample R^2 based on the following equation:

$$R^2_{oos} = 1 - \frac{\sum_{(i,t)\in\text{Test}} (r_{i,t} - \hat{r}_{i,t})^2}{\sum_{(i,t)\in\text{Test}} r^2_{i,t}}$$

Where the determinator is the sum of squared excess returns without demeaning.
Second, since the distribution of portfolio returns is sensitive to the dependence among
stock returns, a good stock-level prediction model does not necessarily produce accurate
portfolio-level forecasts. Thus, it is worth examining the model's predictability at the portfolio level. We sort stocks into deciles based on the prediction of models. Within each decile, we consider both equal weight and value weight by stocks' market capitalization, and we rebalance the portfolio every month.
In addition, we calculate a model's in-sample R^2, R$_{is}^2$ which uses the return data in the
training sample. This comparison can help identify model overfitting.

Sharpe Ratio
The key difference between R$_2$ and Sharpe ratios is that the calculation of the former involves predicted returns while the latter only uses the rank of predicted returns. This makes R^2 an arguably more rigorous measure of model performance. However, the value of R^2 is more sensitive to outliers (i.e., extreme prediction errors) and can be noisy. As shown later, R^2 and Sharpe ratios occasionally suggest different results. Sharpe ratios are more economically meaningful from the perspective of investors. We discuss both measures in our main analysis.

Relative importance of Predictors

To identify the predicts we have utilized the approach by Dimopoulos Bourrewt, and Lek (1995).

For the contribution of the j-th input variable, we calculate:

$$\mathrm{SSD}_j = \sum_k \left(\left. \frac{\partial f}{\partial x_j} \right|_{x=x^k} \right)^2$$

7.2.2 Predicting Stock Returns Using Machine Learning

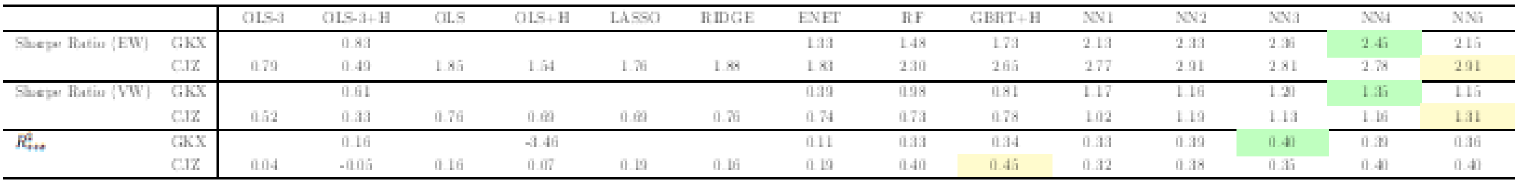
7.2.2.1 Predicting US Stock Returns

The main purpose in to compare the performance of the models with those in GKX, with the input more than 900 features.

First, we pick a smaller set of hyperparameters in NN models to save computing capacity.

Second, we normalize all variables in each month and each market to zero mean and unit standard deviation, while GKX rank-transform variables onto the range of [-1; +1].

Standardization is applied to achieve accelerated algorithm convergence rate, which is especially critical for NN models. For some machine learning models, the objective function may not work properly when inputs have various ranges. Third, we subtract the future cross-sectional average return in the corresponding market from the future return of each stock in each month.



		OLS-3	OLS-3+H	OLS	OLS+H	LASSO	RIDGE	ENET	RF	GBRT+H	NN1	NN2	NN3	NN4	NN5
Sharpe Ratio (EW)	GKX		0.83					1.33	1.48	1.73	2.13	2.33	2.36	2.45	2.15
	CJZ	0.79	0.49	1.85	1.54	1.76	1.88	1.83	2.30	2.65	2.77	2.91	2.81	2.78	2.91
Sharpe Ratio (VW)	GKX		0.61					0.39	0.98	0.81	1.17	1.16	1.20	1.35	1.15
	CJZ	0.52	0.33	0.76	0.69	0.69	0.76	0.74	0.73	0.78	1.02	1.19	1.13	1.16	1.31
R^2_{oos}	GKX		0.16		-3.46			0.11	0.33	0.34	0.33	0.39	0.40	0.39	0.36
	CJZ	0.04	-0.05	0.16	0.07	0.19	0.16	0.19	0.40	0.45	0.32	0.38	0.35	0.40	0.40

Table 7.3

In the table 7.3 we have taken in consideration three metrics for evaluation model performance.

Uut-of-sample R_2 (R_{2oos}, in percent), and equal- and value-weighted Sharpe ratios (SR) of the long-short portfolio returns. Our plain OLS model generates an R_{2oos} of 0.16%, which is higher than the R_{2oos} of the OLS with Huber loss in

GKX, -3:46%. The improvement comes from restricting OLS to a sparse parameterization, as we force the model to include only 36 covariates. Next, when we restrict to only three input features (i.e., size, value, and momentum), we obtain model performance inferior to GKX. The OLS-3+H model in GKX produces an R_{2oos} of 0.16% and equal-weighted (value-weighted) SR of 0.83 (0.61), while in our experiment, R_{2oos} is -0.05% and equal-weighted (value-weighted) SR is 0.49 (0.33). The di_erence is plausibly due to the way of standardization.

Regularizing the linear model via dimension reduction or shrinkage gives similar redactions to OLS in our setting. The R_{2oos} of LASSO and RIDGE are 0.19% and 0.16%, respectively, while the equal-weighted (value-weighted) SR is around 1.8 (0.7). In comparison with GKX, our ENET model obtains a similar R_{2oos} but a higher Sharpe ratio. Tree models, i.e., RF and GBRT+H, exhibit stronger predictive power than the linear models. For example, RF produces an R_{2oos} of 0.40% and equal-weighted (value-weighted) SR of 2.30 (0.73). Such pattern is also present in GKX.

The best performing model is arguably NN, consistent with the _endings of GKX. In GKX, neural network with 4 hidden layers (NN4) has the highest equal-weighted (value- eighted) Sharpe Ratio 2.45 (1.35). Our results are close: we obtain 2.78 (1.16) in NN4, and our equal-weighted Sharpe ratio is the highest with NN2 and NN5 (tied) and reaches 2.91, while our value-weighted Sharpe Ratio is the highest with NN5 at 1.31. The lower Sharpe ratio of value-weighted returns than equal-weighted is consistent with the evidence from GKX. This is not surprising as the literature has shown that larger stocks are subject to less limits to arbitrage, making their abnormal returns more di_cult to forecast. Also, note that in our analysis of the U.S. market, tree models perform as well as NN models, when measuring with R_{2oos}, but NN models can generate considerably higher Sharpe ratios than trees. Overall, with the 36 stock characteristics, our trees and NN models appear to have similar return predictability to models in GKX using more than 900 inputs. One may be surprised by this _nding, but it is consistent with some of the results in GKX and other studies. For example, GKX show that via dimension reduction, the ENET model selects only 20 to 40 features because the inputs and characteristics are partially redundant and fundamentally noisy signals (see Figure 3 of GKX). Furthermore, a few recent studies, such as [167] Feng, Giglio, and Xiu (2020), [168] Freyberger, Neuhierl, and Weber (2020), [169] Kozak, Nagel, and Santosh (2020), argue that a modest number of factors can explain cross-sectional U.S. stock returns. As we show later, using 36 characteristics seems to predict cross-sectional stock returns in international markets as well. Last, we analyze the relative importance of the 36 characteristics in each model; see Figure 7.5.

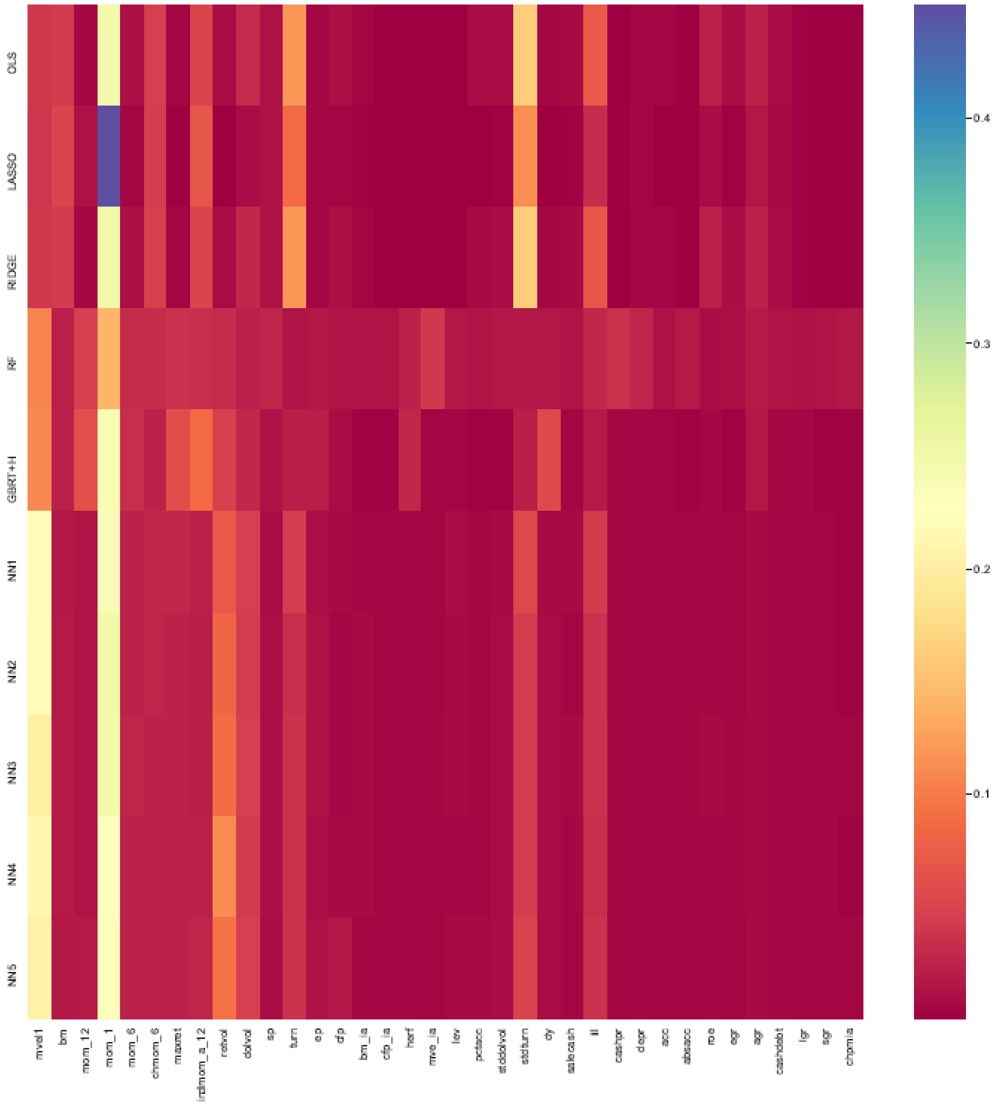


Figure 7.5 Relative Importance of the U.S. – estimated model

In NN models, the strongest predictors are log market capitalization (mvel1), reversal (mom_1), and daily return volatility (retvol). While the pattern is generally consistent with GKX, stock size appears to be more important in our models. This is possibly due to the exclusion of some measures that are shown to be useful in GKX. Firm size may in part capture the eect of these variables in our setting.

7.2.1.2 Predicting International Stock Returns with the U.S. Estimated Models

Focusing on the stringent test, we estimated a model to each of the 312 international markets. For the following year, the training data expands by a year and the validation period maintains the same size (again both using only past U.S. data).

International markets are considered out-of-sample data relative to the U.S. market and allow us to test model overfitting. Given that tuning the hyperparameter values is critical to achieve desirable model performance, the possibility of obverting is an important concern.

The Figure 7.6 shown the reports equal- and value-weighted Sharpe ratios of long-short portfolio returns, along with the Sharpe ratio of the market portfolio during the sample period. We list the markets based on the descending order of the number of observations and highlight the method that gives the highest Sharpe ratio in each market.

Starting with the equal-weighted portfolios on the left, we find two interesting observe tons. First, in every market, machine learning-based models outperform traditional models or the passive market portfolio.

Second, models considering nonlinear and complex interaction effects (i.e., RTs and NN models) outperform linear machine learning models (LASSO and RIDGE). The patterns are similar but slightly weaker for value-weighted Sharpe ratios.

The predictive power of NN models is economically sizable: using the best NN model in each market, the average equal-weighted (value-weighted) Sharpe ratio of the 31 markets is 1.94 (1.07); 19 markets have equal-weighted Sharpe ratio above 1.5 and 26 markets above one, and 15 markets have a value-weighted Sharpe ratio greater than one and 26 larger than 0.75.

However, the out-of-sample R_2 reported in Figure 7.7 shows a different picture. While NN models are still the best model in more than half (17) of the markets, OLS-3 stands out in

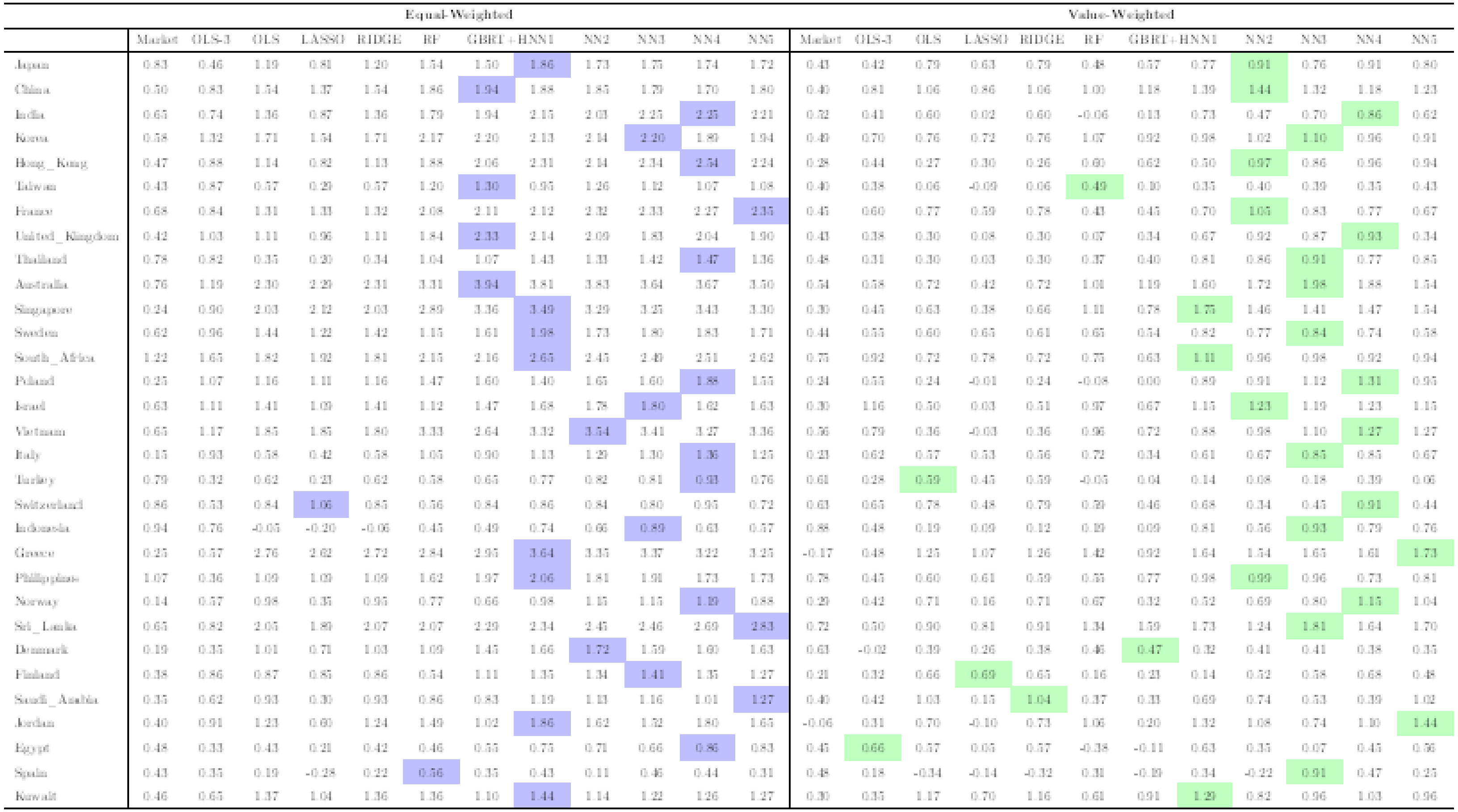
11 markets, consistent with the claim by Fama and French (2012) that OLS-3 is useful performance benchmark for international markets. Regression trees do not give the best prediction in any markets, and in many markets, they generate negative out-of-sample R^2.

We systematically compare the models' performance in Figure 7.8. We pick the best-performing model in each of the three categories (i.e., linear, trees, and NN) and calculate the difference of Sharpe ratios and out-of-sample R^2 between them. We find that on average the best performing tree model can generate an equal-weighted Sharpe ratio that is 0.41 higher than the best linear model across the 31 markets, and among them 26 (or 84%) markets have a positive difference. Comparing NN with linear models, the average difference is even higher, at 0.65, with 30 (or 97%) markets being positive. The best NN model outperforms the best tree by 0.25 in the Sharpe ratio on average, and 25 (or 81%) out of 31 markets have a positive difference. This finding in the international markets is consistent with GKX's conclusion in the U.S. market.

For value-weighted Sharpe Ratios, regression trees do not appear to signific Aly outperform linear models: only 15 markets (48%) have a positive difference, while NN models still signific ally outperform linear and tree models. For out-of-sample R2, regression trees generally have the poorest performance, and NNs give a higher R2 than linear models in 17 (or 55%) of 31 markets.

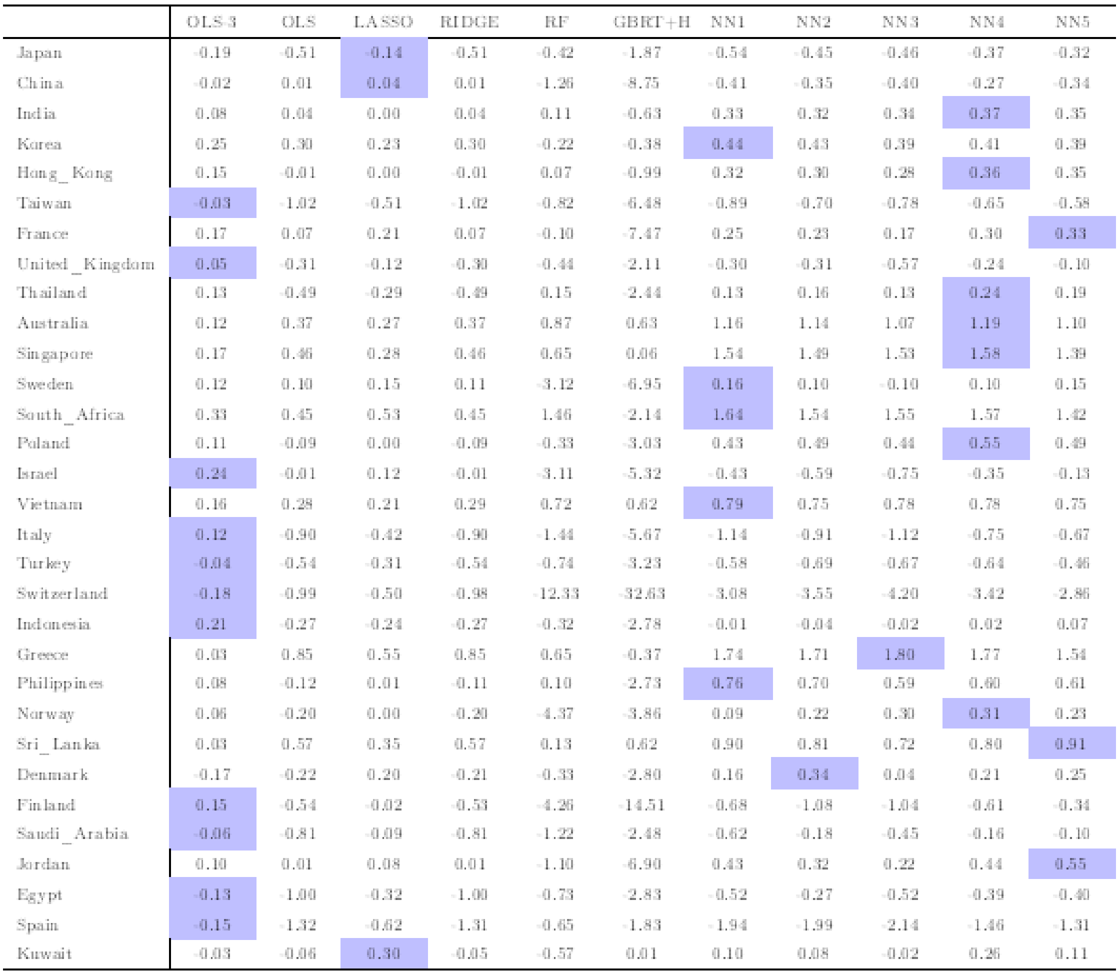
The result in the Figure 7.6 shown mixed evidence regarding the concern of overfitting.

NN models do reasonably well in capturing the common components of return-characteristic relationships in equity markets worldwide.



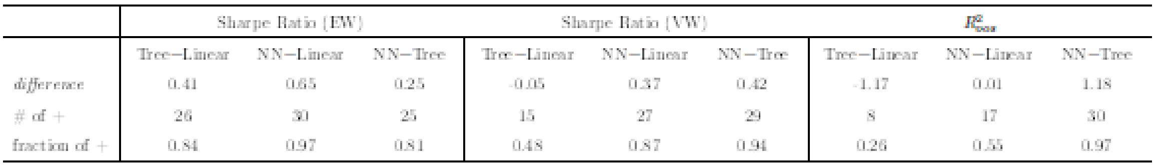
	Equal-Weighted												Value-Weighted											
	Market	OLS-3	OLS	LASSO	RIDGE	RF	GBRT+H	NN1	NN2	NN3	NN4	NN5	Market	OLS-3	OLS	LASSO	RIDGE	RF	GBRT+H	NN1	NN2	NN3	NN4	NN5
Japan	0.83	0.46	1.19	0.81	1.20	1.54	1.50	1.86	1.73	1.75	1.74	1.72	0.43	0.42	0.79	0.63	0.79	0.48	0.57	0.77	0.91	0.76	0.91	0.80
China	0.50	0.83	1.54	1.37	1.54	1.86	1.94	1.88	1.85	1.79	1.70	1.80	0.40	0.81	1.06	0.86	1.06	1.00	1.18	1.39	1.44	1.32	1.18	1.23
India	0.65	0.74	1.36	0.87	1.36	1.79	1.94	2.15	2.03	2.25	2.25	2.21	0.52	0.41	0.60	0.02	0.60	-0.06	0.13	0.73	0.47	0.70	0.86	0.62
Korea	0.58	1.32	1.71	1.54	1.71	2.17	2.20	2.13	2.14	2.20	1.89	1.94	0.49	0.70	0.76	0.72	0.76	1.07	0.92	0.98	1.02	1.10	0.96	0.91
Hong_Kong	0.47	0.88	1.14	0.82	1.13	1.88	2.06	2.31	2.14	2.34	2.54	2.24	0.28	0.44	0.27	0.30	0.26	0.60	0.62	0.50	0.97	0.86	0.96	0.94
Taiwan	0.43	0.87	0.57	0.29	0.57	1.20	1.30	0.95	1.26	1.12	1.07	1.08	0.40	0.38	0.06	-0.09	0.06	0.49	0.10	0.35	0.40	0.39	0.35	0.43
France	0.68	0.84	1.31	1.33	1.32	2.08	2.11	2.12	2.32	2.33	2.27	2.35	0.45	0.60	0.77	0.59	0.78	0.43	0.45	0.70	1.05	0.83	0.77	0.67
United_Kingdom	0.42	1.03	1.11	0.96	1.11	1.84	2.33	2.14	2.09	1.83	2.04	1.90	0.43	0.38	0.30	0.08	0.30	0.07	0.34	0.67	0.92	0.87	0.93	0.34
Thailand	0.78	0.82	0.35	0.30	0.34	1.04	1.07	1.43	1.33	1.42	1.47	1.36	0.48	0.31	0.30	0.03	0.30	0.37	0.40	0.81	0.86	0.91	0.77	0.85
Australia	0.76	1.19	2.30	2.29	2.31	3.31	3.94	3.81	3.83	3.64	3.67	3.50	0.54	0.58	0.72	0.42	0.72	1.01	1.19	1.60	1.72	1.98	1.88	1.54
Singapore	0.24	0.90	2.03	2.12	2.03	2.89	3.36	3.49	3.29	3.25	3.43	3.30	0.30	0.45	0.63	0.38	0.66	1.11	0.78	1.75	1.46	1.41	1.47	1.54
Sweden	0.62	0.96	1.44	1.22	1.42	1.15	1.61	1.98	1.73	1.80	1.83	1.71	0.44	0.55	0.60	0.65	0.61	0.65	0.54	0.82	0.77	0.84	0.74	0.58
South_Africa	1.22	1.65	1.82	1.92	1.81	2.15	2.16	2.65	2.45	2.49	2.51	2.62	0.75	0.92	0.72	0.78	0.72	0.75	0.63	1.11	0.96	0.98	0.92	0.94
Poland	0.25	1.07	1.16	1.11	1.16	1.47	1.60	1.40	1.65	1.60	1.88	1.55	0.24	0.55	0.24	-0.01	0.24	-0.08	0.00	0.89	0.91	1.12	1.31	0.95
Israel	0.63	1.11	1.41	1.09	1.41	1.12	1.47	1.68	1.78	1.80	1.62	1.63	0.30	1.16	0.50	0.03	0.51	0.97	0.67	1.15	1.23	1.19	1.23	1.15
Vietnam	0.65	1.17	1.85	1.85	1.80	3.33	2.64	3.32	3.54	3.41	3.27	3.36	0.56	0.79	0.36	-0.03	0.36	0.96	0.72	0.88	0.98	1.10	1.27	1.27
Italy	0.15	0.93	0.58	0.42	0.58	1.05	0.90	1.13	1.29	1.30	1.36	1.25	0.23	0.62	0.57	0.53	0.56	0.72	0.34	0.61	0.67	0.85	0.85	0.67
Turkey	0.79	0.32	0.62	0.23	0.62	0.58	0.65	0.77	0.82	0.81	0.93	0.76	0.61	0.28	0.59	0.45	0.59	-0.05	0.04	0.14	0.08	0.18	0.39	0.06
Switzerland	0.86	0.53	0.84	1.06	0.85	0.56	0.84	0.86	0.84	0.80	0.95	0.72	0.63	0.65	0.78	0.48	0.79	0.39	0.46	0.68	0.34	0.45	0.91	0.44
Indonesia	0.94	0.76	-0.05	-0.20	-0.06	0.45	0.49	0.74	0.66	0.89	0.63	0.57	0.88	0.48	0.19	0.09	0.12	0.19	0.09	0.81	0.56	0.93	0.79	0.76
Greece	0.25	0.57	2.76	2.62	2.72	2.84	2.95	3.64	3.35	3.37	3.22	3.25	-0.17	0.48	1.25	1.07	1.26	1.42	0.92	1.64	1.54	1.65	1.61	1.73
Philippines	1.07	0.36	1.09	1.09	1.09	1.62	1.97	2.06	1.81	1.91	1.73	1.73	0.78	0.45	0.60	0.61	0.59	0.55	0.77	0.98	0.99	0.96	0.73	0.81
Norway	0.14	0.57	0.98	0.35	0.95	0.77	0.66	0.98	1.15	1.15	1.19	0.88	0.29	0.42	0.71	0.16	0.71	0.67	0.32	0.52	0.69	0.80	1.15	1.04
Sri_Lanka	0.65	0.82	2.05	1.89	2.07	2.07	2.29	2.34	2.45	2.46	2.69	2.83	0.72	0.50	0.90	0.81	0.91	1.34	1.59	1.73	1.24	1.81	1.64	1.70
Denmark	0.19	0.35	1.01	0.71	1.03	1.09	1.45	1.66	1.72	1.59	1.60	1.63	0.63	-0.02	0.39	0.26	0.38	0.46	0.47	0.32	0.41	0.41	0.38	0.35
Finland	0.38	0.86	0.87	0.85	0.86	0.54	1.11	1.35	1.34	1.41	1.35	1.27	0.21	0.32	0.66	0.69	0.65	0.16	0.23	0.14	0.52	0.58	0.68	0.48
Saudi_Arabia	0.35	0.62	0.93	0.30	0.93	0.86	0.83	1.19	1.13	1.16	1.01	1.27	0.40	0.42	1.03	0.15	1.04	0.37	0.33	0.69	0.74	0.53	0.39	1.02
Jordan	0.40	0.91	1.23	0.60	1.24	1.49	1.02	1.86	1.62	1.52	1.80	1.65	-0.06	0.31	0.70	-0.10	0.73	1.06	0.20	1.32	1.08	0.74	1.10	1.44
Egypt	0.48	0.33	0.43	0.21	0.42	0.46	0.55	0.75	0.71	0.66	0.86	0.83	0.45	0.66	0.57	0.05	0.57	-0.38	-0.11	0.63	0.35	0.07	0.45	0.56
Spain	0.43	0.35	0.19	-0.28	0.22	0.56	0.35	0.43	0.11	0.46	0.44	0.31	0.48	0.18	-0.34	-0.14	-0.32	0.31	-0.19	0.34	-0.22	0.91	0.47	0.25
Kuwait	0.46	0.65	1.37	1.04	1.36	1.36	1.10	1.44	1.14	1.22	1.26	1.27	0.30	0.35	1.17	0.70	1.16	0.61	0.91	1.29	0.82	0.96	1.03	0.96

Figure 7.6 Equal- and value- whether Sharpe ratio



	OLS-3	OLS	LASSO	RIDGE	RF	GBRT+H	NN1	NN2	NN3	NN4	NN5
Japan	-0.19	-0.51	-0.14	-0.51	-0.42	-1.87	-0.54	-0.45	-0.46	-0.37	-0.32
China	-0.02	0.01	0.04	0.01	-1.26	-8.75	-0.41	-0.35	-0.40	-0.27	-0.34
India	0.08	0.04	0.00	0.04	0.11	-0.63	0.33	0.32	0.34	0.37	0.35
Korea	0.25	0.30	0.23	0.30	-0.22	-0.38	0.44	0.43	0.39	0.41	0.39
Hong_Kong	0.15	-0.01	0.00	-0.01	0.07	-0.99	0.32	0.30	0.28	0.36	0.35
Taiwan	-0.03	-1.02	-0.51	-1.02	-0.82	-6.48	-0.89	-0.70	-0.78	-0.65	-0.58
France	0.17	0.07	0.21	0.07	-0.10	-7.47	0.25	0.23	0.17	0.30	0.33
United_Kingdom	0.05	-0.31	-0.12	-0.30	-0.44	-2.11	-0.30	-0.31	-0.57	-0.24	-0.10
Thailand	0.13	-0.49	-0.29	-0.49	0.15	-2.44	0.13	0.16	0.13	0.24	0.19
Australia	0.12	0.37	0.27	0.37	0.87	0.63	1.16	1.14	1.07	1.19	1.10
Singapore	0.17	0.46	0.28	0.46	0.65	0.06	1.54	1.49	1.53	1.58	1.39
Sweden	0.12	0.10	0.15	0.11	-3.12	-6.95	0.16	0.10	-0.10	0.10	0.15
South_Africa	0.33	0.45	0.53	0.45	1.46	-2.14	1.64	1.54	1.55	1.57	1.42
Poland	0.11	-0.09	0.00	-0.09	-0.33	-3.03	0.43	0.49	0.44	0.55	0.49
Israel	0.24	-0.01	0.12	-0.01	-3.11	-5.32	-0.43	-0.59	-0.75	-0.35	-0.13
Vietnam	0.16	0.28	0.21	0.29	0.72	0.62	0.79	0.75	0.78	0.78	0.75
Italy	0.12	-0.90	-0.42	-0.90	-1.44	-5.67	-1.14	-0.91	-1.12	-0.75	-0.67
Turkey	-0.04	-0.54	-0.31	-0.54	-0.74	-3.23	-0.58	-0.69	-0.67	-0.64	-0.46
Switzerland	-0.18	-0.99	-0.50	-0.98	-12.33	-32.63	-3.08	-3.55	-4.20	-3.42	-2.86
Indonesia	0.21	-0.27	-0.24	-0.27	-0.32	-2.78	-0.01	-0.04	-0.02	0.02	0.07
Greece	0.03	0.85	0.55	0.85	0.65	-0.37	1.74	1.71	1.80	1.77	1.54
Philippines	0.08	-0.12	0.01	-0.11	0.10	-2.73	0.76	0.70	0.59	0.60	0.61
Norway	0.06	-0.20	0.00	-0.20	-4.37	-3.86	0.09	0.22	0.30	0.31	0.23
Sri_Lanka	0.03	0.57	0.35	0.57	0.13	0.62	0.90	0.81	0.72	0.80	0.91
Denmark	-0.17	-0.22	0.20	-0.21	-0.33	-2.80	0.16	0.34	0.04	0.21	0.25
Finland	0.15	-0.54	-0.02	-0.53	-4.26	-14.51	-0.68	-1.08	-1.04	-0.61	-0.34
Saudi_Arabia	-0.06	-0.81	-0.09	-0.81	-1.22	-2.48	-0.62	-0.18	-0.45	-0.16	-0.10
Jordan	0.10	0.01	0.08	0.01	-1.10	-6.90	0.43	0.32	0.22	0.44	0.55
Egypt	-0.13	-1.00	-0.32	-1.00	-0.73	-2.83	-0.52	-0.27	-0.52	-0.39	-0.40
Spain	-0.15	-1.32	-0.62	-1.31	-0.65	-1.83	-1.94	-1.99	-2.14	-1.46	-1.31
Kuwait	-0.03	-0.06	0.30	-0.05	-0.57	0.01	0.10	0.08	-0.02	0.26	0.11

Figure 7.7 Individual stock R_{2oos}



	Sharpe Ratio (EW)			Sharpe Ratio (VW)			R^2_{oos}		
	Tree−Linear	NN−Linear	NN−Tree	Tree−Linear	NN−Linear	NN−Tree	Tree−Linear	NN−Linear	NN−Tree
difference	0.41	0.65	0.25	-0.05	0.37	0.42	-1.17	0.01	1.18
# of +	26	30	25	15	27	29	8	17	30
fraction of +	0.84	0.97	0.81	0.48	0.87	0.94	0.26	0.55	0.97

Figure 7.8 Comparison of model performance

7.2.2.3 Predcting International Stock Returns with Market-Specific Models

When there is any market-specific component in the return-characteristic relationship due to some institutional friction or investor culture in the local market, training the model by each market should presumably achieve even stronger return predictability.

7.2.3 Result

The Figure 7.6 and the Figure 7.7 summarizes the permeance of Shape ratios and R^2_{oos}. The Figure 7.8 represent a caparison between the two models.

Like what we Nd with U.S.-estimated models, NN models exhibit the strongest return predictability in most of the markets in terms of Sharpe Ratios. For the equal-weighted (value-weighted) Sharpe ratio, NN models outperform linear and tree models by 0.60 (0.41) and 0.44 (0.61) on average or in 81% (78%) and 78% (88%) of the markets, respectively. Also like the previous test, the out-of-sample R2 is not as strong: the best-performing NN model's R_{2oos} outperforms by 0.17% and 0.03% on average or in 69% and 47% of the markets, compared to the best of linear and tree models, respectively. Economically, the best NN model achieves an equal-weighted (value-weighted) SR above 1.5 (1) in 21 (20) of the 32 markets.

Market-specific tree models do not seem to dominate linear models. In Figure 7.8, relative to linear models, the average equal-weighted Sharpe Ratio of trees is higher by 0.17, while the average value-weighted Sharpe Ratio of trees is lower by 0.21. The average $R_{2oos}$of treesis higher than that of linear models by 0.13.

Figure 7.8 shows that trees' performance are especially poor in markets where the number of observations is low: in the top half of markets with more observations, tree models' average equal- and value-weighted Sharpe Ratio is higher than that of linear models in 81% and 50% of the markets, respectively; but in the bottom half, these numbers fall to 38% and 25%. This suggests that tree models need more data to converge to a stable parameter estimation.

By comparison, despite its complex structure as well, NN models appear to be more robust. While Panel C documents a corresponding drop from the top half to the bottom half, the drop is smaller (94% and 81% vs. 69% and 75%). Linear models are also robust in estimation due to their simpler model structure, which in turn, however, limits their ability to capture complex return-characteristics relationships.

To further examine the proneness of overtting, we compare a model's in-sample and outof- sample R^2. In Figure 7.7, the in-sample R^2 (R^2_{is}) is reported in the parenthesis under the corresponding out-of-sample R^2 (R^2_{oos}). R^2_{is} is calculated using data and predictions from the training sample In the U.S., the R^2_{oos} of each model is slightly lower than itsrespective R^2_{is}. However, for markets with lower number of observations, R^2_{oos} of tree models are substantially lower than their R^2_{is}. Using Monte Carlo simulations, GKX also find that tree models are prone to overfiring and experience a larger decrease in R^2 from in-sample toout-of-sample. Our finding complements theirs by using the international equity markets as the true out-of-sample data.

While there is no clear theoretical explanation that trees are more vulnerable to overfitting, our out-of-sample tests confirm that, at least for this type of financial data, the structure, and the regularization settings of NN can fit and learn in a more robust manner and generate stronger return predictability. This is consistent with evidence in the machine learning literature that random forests can be inconsistent and that NN models with multiple layers do not overfit the training data.

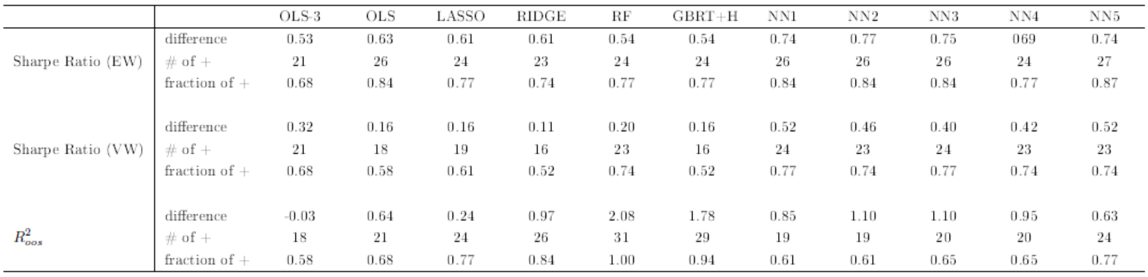
7.2.4 Return-Characteristics Relationship: Common or Market-Specific

Based on the model present in the Figure 7.6 are the top 15 markets based on the number of observations.

On the one hand, there are similarities in the return-characteristic relationship across international equity markets.

An important question follows naturally is whether market-specific models perform better than their U.S.-estimated counterparts. to make the comparison sensible, we require the U.S. model to be estimated only using the data over the same sample years that the market-specific model uses.

For example, China's data are available from 1999 to 2017, with 1999-2004 as the training period and 2005-2007 as the validating period. To compare the China-specific model with the U.S.-estimated counterpart, we train and validate with the U.S. data in 1999-2004 and 2005-2007, respectively. Then, for each machine learning method, we compare the return predictions from the U.S-estimated model with those from the market-specific model, based on Sharpe ratios and out-of-sample R2. We repeat this procedure for each of the 31 international markets in our sample and summarize the differences across all markets.



		OLS-3	OLS	LASSO	RIDGE	RF	GBRT+H	NN1	NN2	NN3	NN4	NN5
Sharpe Ratio (EW)	difference	0.53	0.63	0.61	0.61	0.54	0.54	0.74	0.77	0.75	069	0.74
	# of +	21	26	24	23	24	24	26	26	26	24	27
	fraction of +	0.68	0.84	0.77	0.74	0.77	0.77	0.84	0.84	0.84	0.77	0.87
Sharpe Ratio (VW)	difference	0.32	0.16	0.16	0.11	0.20	0.16	0.52	0.46	0.40	0.42	0.52
	# of +	21	18	19	16	23	16	24	23	24	23	23
	fraction of +	0.68	0.58	0.61	0.52	0.74	0.52	0.77	0.74	0.77	0.74	0.74
R^2_{oos}	difference	-0.03	0.64	0.24	0.97	2.08	1.78	0.85	1.10	1.10	0.95	0.63
	# of +	18	21	24	26	31	29	19	19	20	20	24
	fraction of +	0.58	0.68	0.77	0.84	1.00	0.94	0.61	0.61	0.65	0.65	0.77

Table 7.4

Table 7.4 represent the result; we find that market-specific models generally outperform their U.S. estimated counterparts. For example, market-specific models improve equal weighted Sharpe ratios by 0.53 to 0.77 and value-weighted Sharpe ratios by 0.11 to 0.52 on average across the 31 markets. If we focus on NN models, the improvement is visible: 74,87% of the markets experience an increase in Sharpe ratio, while 61,77% of the markets see an increase in R^2.

7.2.5 The dimension where machine learning improves Return Predictability

For trees, there is only limited improvement in equal-weighted Sharpe Ratios when the number of observations is scient. While NN models and trees are designed to capture nonlinearity and complex interaction effects, in our tests they do not outperform linear models in all dimensions. Here we further examine the return predictability by focusing on the difference between Sharpe Ratios and out-of-sample R^2.

While both are measures of models' performance, a high Sharpe Ratio indicates that a model does well in sorting stocks by their predicted returns, especially those in the top and bottom deciles. A high out-of-sample R^2 indicates that a model does well in predicting the actual returns generally, across all stocks. Therefore, it is possible that a model achieves a high Sharpe Ratio but low R^2, if the stocks that it puts in the extreme deciles indeed have extreme actual returns but the predictions (particularly in other deciles) are poor.

In this section, we offer more evidence that NN models perform well in sorting stocks into top and bottom deciles but not as good for other deciles and in forming return predictions.

Actual return deciles. We first verify that the stocks that NN place in the top and bottom deciles are those that produce extreme actual returns. Recall that the top (bottom) decile is the decile with the highest (lowest) predicted returns.

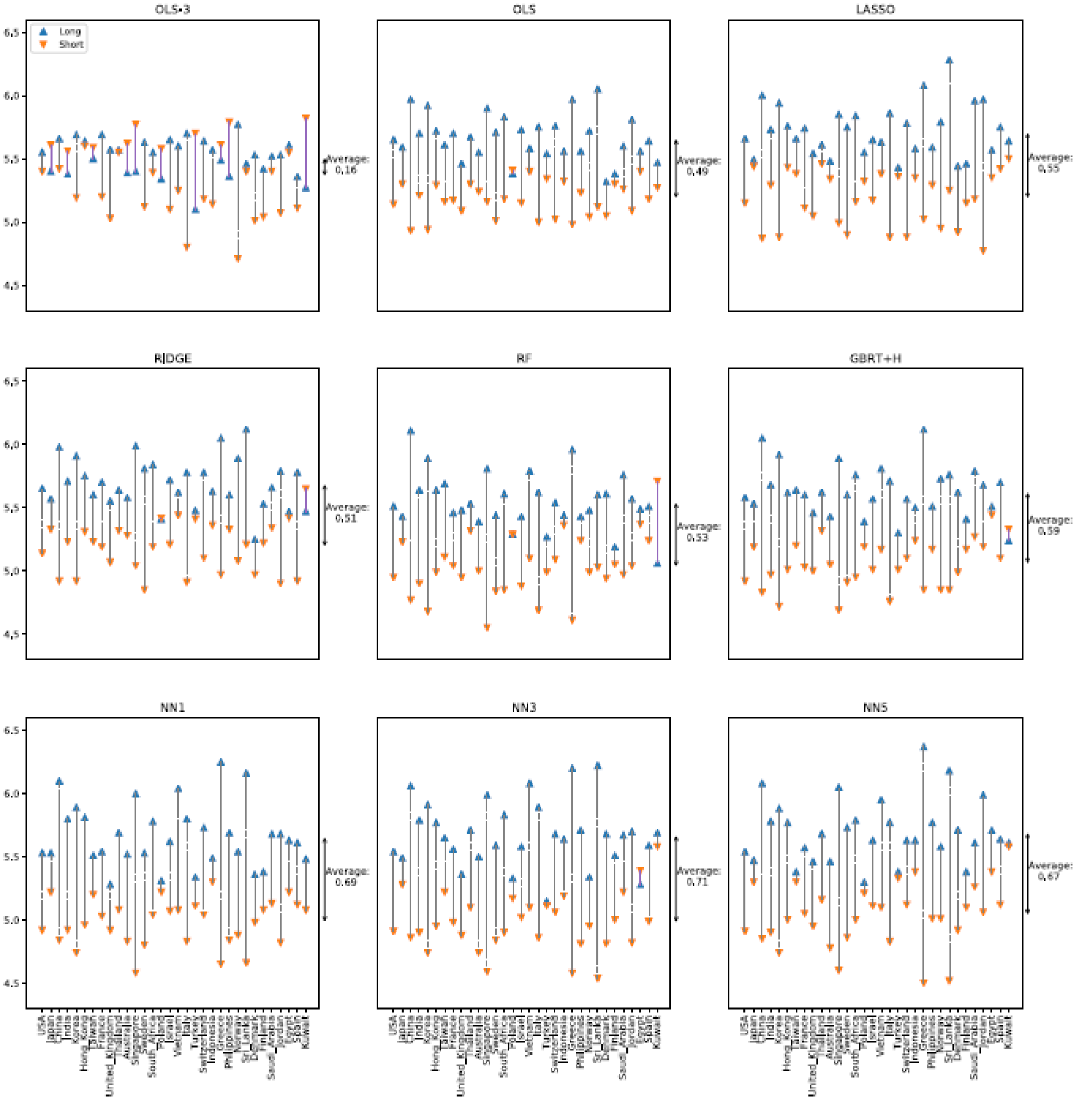


Figure 7.9 Comparison of actual return deciles of the long and the short portfolios.

In Figure 7.9, we graph then average actual return deciles of the long and the short portfolios in each market in the market-specific tests and report the average distance of each model (for brevity, NN2 and NN4 are omitted). If a model has zero predictive power, the actual return deciles would be 5.5 for both the long and the short portfolios, and the average distance would be zero. If a model has perfect predictive power, the actual return decile for the long (short) portfolio would be 10 (1), and the average distance would be 9.

LASSO, RIDGE, and RF show an average distance of 0.510.55, while GBRT+H is slightly higher at 0.59. The distance in RF and GBRT+H is wider in markets with more 6observations, consistent with the results in Figure 7.7. NN models report an average distance of around 0.7, meaning that the long portfolio is on average 0.7 decile higher than the short portfolio in terms of actual returns. This is substantially higher than OLS and OLS-3, which have an average distance of 0.49 and 0.16, respectively.

9th minus 2nd long-short portfolios. Instead of using the top and bottom deciles to construct the long-short portfolio, here we examine the returns of the portfolio that longs the 9th and shorts the 2nd decile. Table 7 reports the result of Sharpe ratios. Figure 7.6 uses the U.S.-estimated (market-specie) model, and Panel C compares the models' performance.

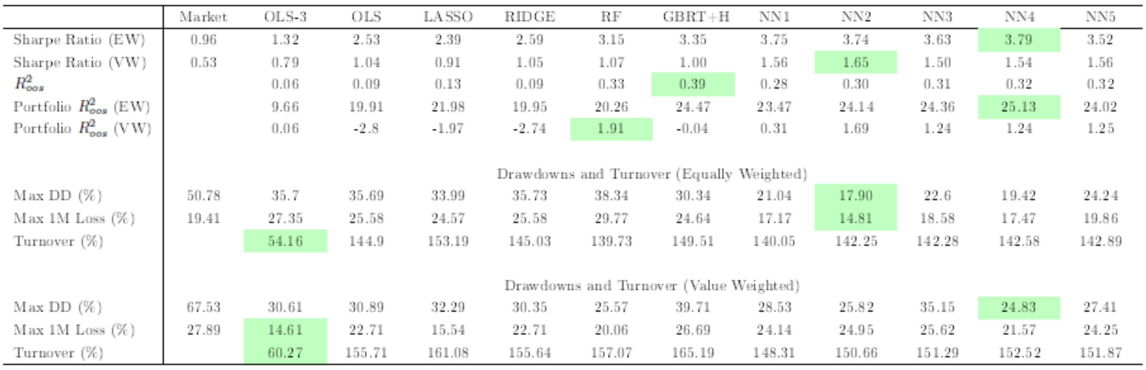
Within the U.S.-estimated models, the predictions based on NN models again appear to be the best: in terms of equal-weighted (value-weighted) Sharpe ratios, NN models outperform linear models by 0.29 (0.27) on average or in 74% (81%) of the markets. For market-specific models: NN models outperform linear models by 0.12 (0.14) on average or in 56% (59%) of the markets. The outperformance is weaker than that in Figure 7.6 and Figure 7.7, where 10th and 1st deciles are used. In the 9th and 2nd deciles, NN models are less impressive. Portfolio R^2. R^2 for individual stock returns can be noisy due to outliers. Figure 7.6 uses the portfolio R^2 instead. Although there are many ways to construct the portfolio, we focus on the decile portfolios sorted on the return prediction of the machine learning model. We consider both equal- and value-weighted portfolio returns. The out-of-sample R^2 is calculated as specified in the following Equation

$$R^2_{oos} = 1 - \frac{\sum_{(i,t)\in T_{test}} \left(r_{i,t} - \hat{r}_{i,t}\right)^2}{\sum_{(i,t)\in T_{test}} r^2_{i,t}}$$

Figure 7.6 and Figure 7.7 present the result of U.S.-estimated and market-specific models, respectively. The results are generally aligned with those using individual stock R^2. Figure 7.8 shows that NN models produce better portfolio R^2 in 52,74% of the markets, again not as dominant as they are in Sharpe Ratios.

7.2.6 Polling all stocks

At this moment we will poll all stocks our global sample to train and validate the machine learning models and predict expected returns. With more data and larger space for portfolio selection, machine learning models can be better trained and have stronger predictive power.
While the sample starts from 1963, our testing period is from July 1992 to December 2017 due to the availability of risk factors.



	Market	OLS-3	OLS	LASSO	RIDGE	RF	GBRT+H	NN1	NN2	NN3	NN4	NN5
Sharpe Ratio (EW)	0.96	1.32	2.53	2.39	2.59	3.15	3.35	3.75	3.74	3.63	3.79	3.52
Sharpe Ratio (VW)	0.53	0.79	1.04	0.91	1.05	1.07	1.00	1.56	1.65	1.50	1.54	1.56
R^2_{oos}		0.06	0.09	0.13	0.09	0.33	0.39	0.28	0.30	0.31	0.32	0.32
Portfolio R^2_{oos} (EW)		9.66	19.91	21.98	19.95	20.26	24.47	23.47	24.14	24.36	25.13	24.02
Portfolio R^2_{oos} (VW)		0.06	-2.8	-1.97	-2.74	1.91	-0.04	0.31	1.69	1.24	1.24	1.25
						Drawdowns and Turnover (Equally Weighted)						
Max DD (%)	50.78	35.7	35.69	33.99	35.73	38.34	30.34	21.04	17.90	22.6	19.42	24.24
Max 1M Loss (%)	19.41	27.35	25.58	24.57	25.58	29.77	24.64	17.17	14.81	18.58	17.47	19.86
Turnover (%)		54.16	144.9	153.19	145.03	139.73	149.51	140.05	142.25	142.28	142.58	142.89
						Drawdowns and Turnover (Value Weighted)						
Max DD (%)	67.53	30.61	30.89	32.29	30.35	25.57	39.71	28.53	25.82	35.15	24.83	27.41
Max 1M Loss (%)	27.89	14.61	22.71	15.54	22.71	20.06	26.69	24.14	24.95	25.62	21.57	24.25
Turnover (%)		60.27	155.71	161.08	155.64	157.07	165.19	148.31	150.66	151.29	152.52	151.87

Table 7.4 Performance of International Portfolios: Pooling All Stocks

The table 7.4 represent the results, NN models perform the best in most dimensions. The global equal-weighted (value-weighted) long-short portfolio based on NNs yields a Sharpe ratio of 3.79 (1.65), a large improvement from previous tables, while the Sharpe ratio of the market portfolio is 0.96 (0.53). With more data available for training, the performance of tree model also improves, with equal-weighted (value-weighted) Sharpe ratios of 3.15-3.35 (1.00-1.07). One should take the high Sharpe ratio with caution for investment purpose, as the estimates here do not consider transaction costs or other frictions, such as short-sale constraints, in the international equity markets. When R^2 measures are used, NN models and trees still outperform linear models. R$^2_{\text{oos}}$ reaches the highest of 0.39% from GBRT+H. Note that the performance of tree models is close to that of NN models and better than linear ones, consistent with the conjecture that the estimation of tree models requires a large data sample.

We next examine the risk of the machine learning based long-short 10,1 portfolios. Following GKX, we first look at the maximum drawdown (MaxDD), maximum one month loss (Max 1M Loss), and portfolio turnover rate. The maximum drawdown of a strategy is defined as:

$$\mathrm{MaxDD} = \max_{0 \leq t_1 \leq t_2 \leq T}(Y_{t_1} - Y_{t_2}),$$

Where:

- Y_{t} in the calculation long term from month zero through t

The maximum one-month loss (Max 1M Loss) is the lowest monthly return of the trading strategy. For equal-weighted portfolios, NN-based strategies have the lowest maximum drawdown and one-month loss.

For value-weighted portfolios, NN models have the lowest maximum drawdown, but OLS 3 provides the lowest one-month loss.

The portfolio average monthly turnover is calculated:

$$\mathrm{Turnover} = \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{i}\left|w_{i,t+1} - \frac{w_{i,t}(1+r_{i,t+1})}{\sum_j (1+r_{j,t+1})}\right|\right)$$

Where:

- $\mathrm{W}_{\mathrm{i,t}}$ is the weight of stock i in the portfolio at month t.

It appears that the monthly turnover rate of NN-based strategies is approximately 150%, which is about 20 to 30% higher than the number shown in GKX based on the U.S. market. Given the larger pool of stocks and the important role of price trend predictors in machine learning models, it is not surprising that the outperformance is achieved with relatively higher portfolio turnover rate. The previous results are all based on raw returns more than the U.S. risk free rate. Last, we turn to risk-adjusted returns to examine whether the machine learning models capture something beyond the commonly known factors. We adopt three international asset pricing models to calculate risk-adjusted returns: the Fama-French five -factor model augmented with a momentum factor, the 6-factor model developed by Hou, Karolyi, and Kho

(2011), and the partial-segmentation Carhart model in Karolyi and Wu (2018). 14 In the Fama- French model, we include a set of the 6 factors for developed markets and a set for emerging markets. That is, in total 12 factors are used for the risk adjustment of the global portfolio returns.

In the Table 7.4, the monthly equal-weighted (value weighted) alphas based on the best performing NN model are significantly positive, at 3.66% - 4.88% (1.86% - 2.38%) with t-statistics well above 5. The R^2 of the factor models is low, typically below 10%, particularly for NN, suggesting that the factors can only explain a small fraction of the returns of the NN-based strategies. Therefore, NN models' information ratio (IR) ranges from 1.06 to 1.18 for equal weighting and 0.38 to 0.51 for value weighting.

For most measures, NN and tree models, which consider nonlinear and complex interaction effects, significantly outperform linear models. An interesting observation here is that with a larger data sample, the performance of tree models stably improves.

7.2.7 Polling all Non-U.S. Stocks

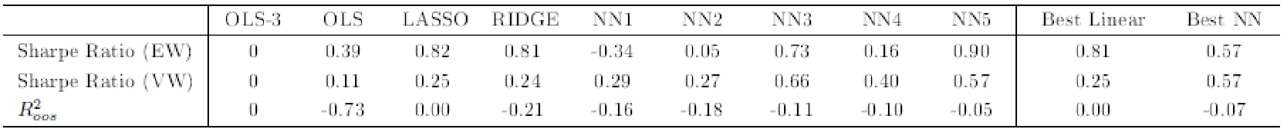
While there are multiple ways to extract information from U.S. stocks, we add new variables that are like those commonly used in the literature. We construct three types of variables:

- U.S. Factors: In each month, for each of the 36 characteristics, we sort U.S. stocks into 10 deciles in descending order and compute the value-weighted returns for each decile. Then we done a factor as the return of the top decile portfolio minus the return of the bottom decile portfolio. This is similar to the way that common risk factors are constructed, such as Fama and French (1993, 2015, 2017).

- U.S. Characteristic Gaps: In each month, we compute the characteristic gap as the divergence between the 95th percentile and the 5th percentile of a corresponding stock characteristic in the U.S. market. Cohen, Polk, and Vuolteenaho (2003) and Huang (2021) show that gaps in book-to-market and in past returns, respectively, can predict future returns.

- Local Factors: As a comparison, we compute local factors in the same way as the U.S. factors. Stocks that are in the same market as the stock in question are used.

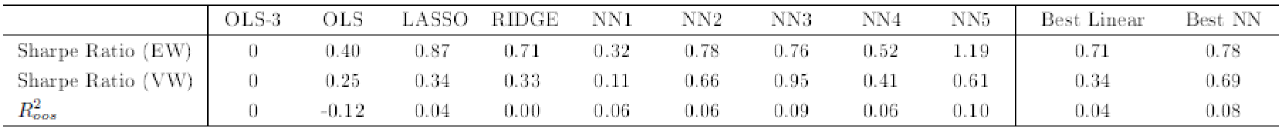
We also compute the interaction terms for each stock characteristic and its respective factor or characteristic gap. Therefore, on top of the 36 raw stock characteristics, the augmented model in this section adds 36 X 3 factors or characteristic gaps + 36 X 3 interaction terms = 216 independent variables.
For the pooled sample of non-U.S. stocks, we only study linear and NN models and drop trees because of computational constraints.



	OLS-3	OLS	LASSO	RIDGE	NN1	NN2	NN3	NN4	NN5	Best Linear	Best NN
Sharpe Ratio (EW)	0	0.39	0.82	0.81	-0.34	0.05	0.73	0.16	0.90	0.81	0.57
Sharpe Ratio (VW)	0	0.11	0.25	0.24	0.29	0.27	0.66	0.40	0.57	0.25	0.57
R^2_{oos}	0	-0.73	0.00	-0.21	-0.16	-0.18	-0.11	-0.10	-0.05	0.00	-0.07

Table 7.5 comparison of Performance between Augmented non-U.S. model and Original non-U.S. Model

Table 7.5 reports the difference in equal-weighted and value-weighted Sharpe Ratios and out-of-sample R_2 between the augmented models and the original models using only 36 stock characteristics. For both linear and NN models, the augmented model generally improves equal- and value-weighted Sharpe Ratios but not R_2.

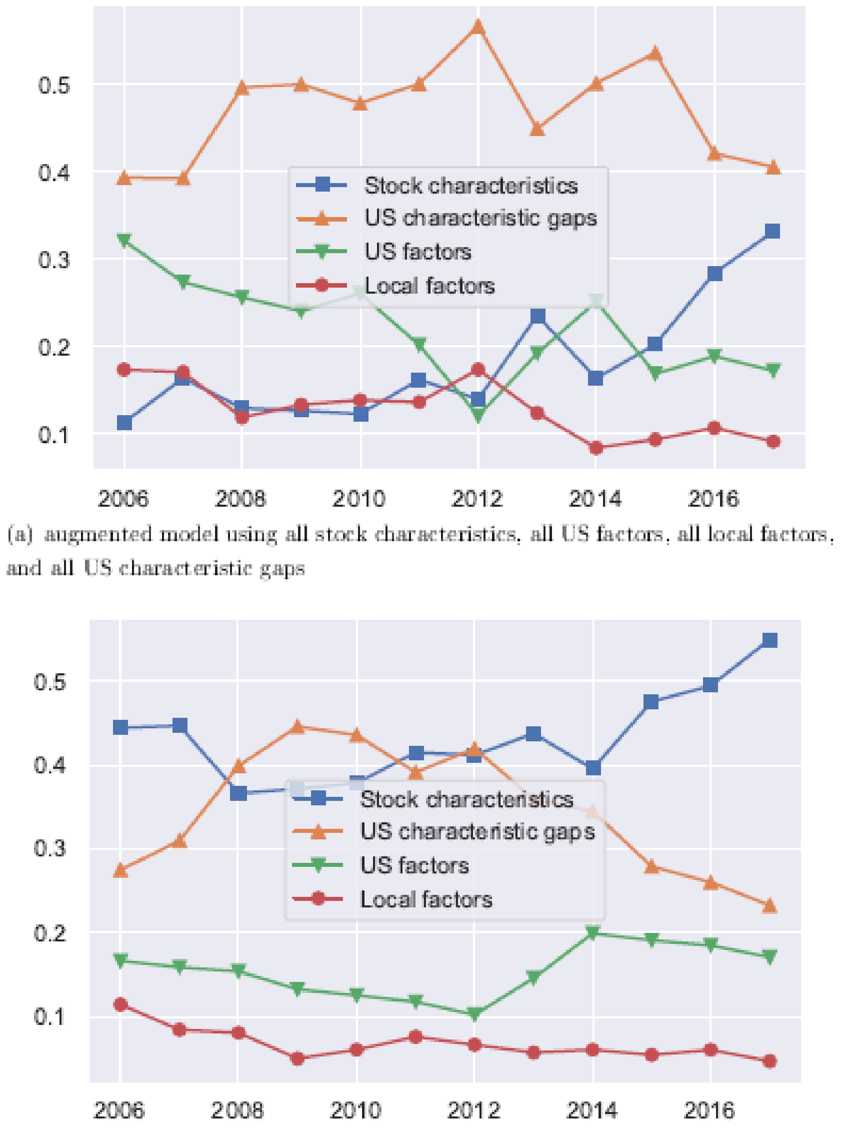


	OLS-3	OLS	LASSO	RIDGE	NN1	NN2	NN3	NN4	NN5	Best Linear	Best NN
Sharpe Ratio (EW)	0	0.40	0.87	0.71	0.32	0.78	0.76	0.52	1.19	0.71	0.78
Sharpe Ratio (VW)	0	0.25	0.34	0.33	0.11	0.66	0.95	0.41	0.61	0.34	0.69
R^2_{oos}	0	-0.12	0.04	0.00	0.06	0.06	0.09	0.06	0.10	0.04	0.08

Table 7.6 comparison of Performance between Augmented non-U.S. model and Original non-U.S. Model

In the Table 7.6 we reduce the number of additional variables by focusing on the top 10 characteristics in each market. For each market, we select the top 10 characteristics according to their variable importance on the raw market-specific models. Therefore, in each test we add 10 X 3 factors or characteristic gaps + 10 X 3 interaction terms = 60 independent variables (on top of the 36 stock characteristics). The performance of augmented linear models here is like that in Table 7.5 but augmented NN models with top 10 characteristics show even higher equal- and value-weighted Sharpe Ratios and a slightly higher R^2. Comparing with the best original NN model, the best augmented NN model's equal-(value-) Sharpe Ratio is higher by 0.78 (0.69), and the R2 is higher by 0.08.

The difference between Table 7.5 and Table 7.6 highlights that NN models do not necessarily become more powerful when having more independent variables. Even with many observations, the full augmented NN models with 36 characteristics do not generate the best results.



(a) augmented model using all stock characteristics, all US factors, all local factors, and all US characteristic gaps

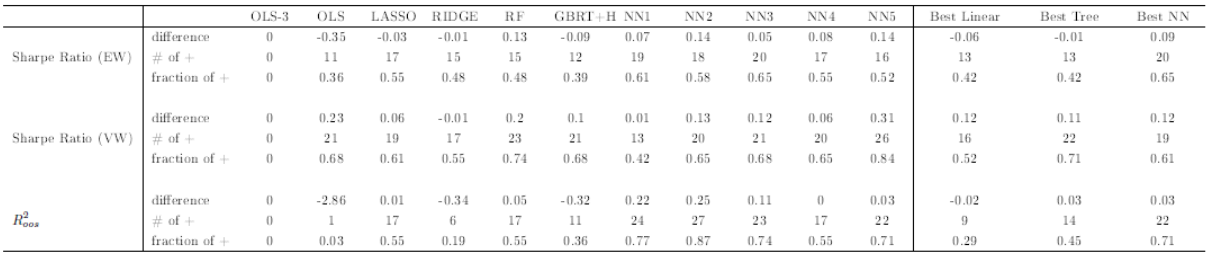
(b) augmented model using all stock characteristics, top 10 US factors, top 10 local factors, and top 10 US characteristic gaps

Figure 7.10 Group variable importance over time.

The Figure 7.10 shown the variable importance of each type of variables in the best augmented NN model. The sum of variable importance is normalized to one. U.S. characteristic gaps are even more important than stock characteristics in the full model with 36 characteristics. They are still very important, with an average value of above 0.3, in the augmented model with top 10 characteristics. Local factors have the lowest variable importance, especially in recent years.

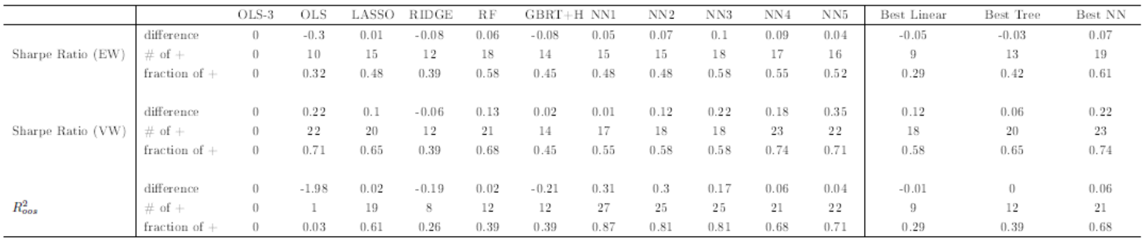
7.2.8 Market- Specific Models

Considering the market-specific models with the additional variables. Given our findings in the previous subsection, we only add U.S. characteristic gaps and U.S. factors based on the top 10 characteristics in each market (because the market-specific models contain a much lower number of observations).



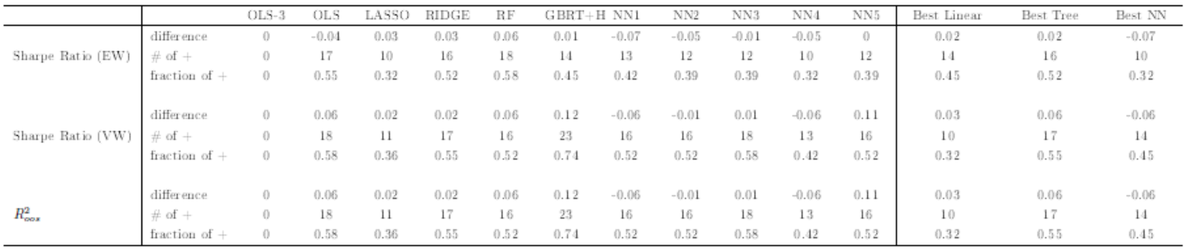
		OLS-3	OLS	LASSO	RIDGE	RF	GBRT+H	NN1	NN2	NN3	NN4	NN5	Best Linear	Best Tree	Best NN
	difference	0	-0.35	-0.03	-0.01	0.13	-0.09	0.07	0.14	0.05	0.08	0.14	-0.06	-0.01	0.09
Sharpe Ratio (EW)	# of +	0	11	17	15	15	12	19	18	20	17	16	13	13	20
	fraction of +	0	0.36	0.55	0.48	0.48	0.39	0.61	0.58	0.65	0.55	0.52	0.42	0.42	0.65
	difference	0	0.23	0.06	-0.01	0.2	0.1	0.01	0.13	0.12	0.06	0.31	0.12	0.11	0.12
Sharpe Ratio (VW)	# of +	0	21	19	17	23	21	13	20	21	20	26	16	22	19
	fraction of +	0	0.68	0.61	0.55	0.74	0.68	0.42	0.65	0.68	0.65	0.84	0.52	0.71	0.61
	difference	0	-2.86	0.01	-0.34	0.05	-0.32	0.22	0.25	0.11	0	0.03	-0.02	0.03	0.03
R^2_{oos}	# of +	0	1	17	6	17	11	24	27	23	17	22	9	14	22
	fraction of +	0	0.03	0.55	0.19	0.55	0.36	0.77	0.87	0.74	0.55	0.71	0.29	0.45	0.71

Table 7.7 Comparison of Performance between Augmented market-specific models and Original market-specific models



		OLS-3	OLS	LASSO	RIDGE	RF	GBRT+H	NN1	NN2	NN3	NN4	NN5	Best Linear	Best Tree	Best NN
	difference	0	-0.3	0.01	-0.08	0.06	-0.08	0.05	0.07	0.1	0.09	0.04	-0.05	-0.03	0.07
Sharpe Ratio (EW)	# of +	0	10	15	12	18	14	15	15	18	17	16	9	13	19
	fraction of +	0	0.32	0.48	0.39	0.58	0.45	0.48	0.48	0.58	0.55	0.52	0.29	0.42	0.61
	difference	0	0.22	0.1	-0.06	0.13	0.02	0.01	0.12	0.22	0.18	0.35	0.12	0.06	0.22
Sharpe Ratio (VW)	# of +	0	22	20	12	21	14	17	18	18	23	22	18	20	23
	fraction of +	0	0.71	0.65	0.39	0.68	0.45	0.55	0.58	0.58	0.74	0.71	0.58	0.65	0.74
	difference	0	-1.98	0.02	-0.19	0.02	-0.21	0.31	0.3	0.17	0.06	0.04	-0.01	0	0.06
R^2_{oos}	# of +	0	1	19	8	12	12	27	25	25	21	22	9	12	21
	fraction of +	0	0.03	0.61	0.26	0.39	0.39	0.87	0.81	0.81	0.68	0.71	0.29	0.39	0.68

Table 7.8 Comparison of Performance between Augmented market-specific models and Original market-specific models



		OLS-3	OLS	LASSO	RIDGE	RF	GBRT+H	NN1	NN2	NN3	NN4	NN5	Best Linear	Best Tree	Best NN
	difference	0	-0.04	0.03	0.03	0.06	0.01	-0.07	-0.05	-0.01	-0.05	0	0.02	0.02	-0.07
Sharpe Ratio (EW)	# of +	0	17	10	16	18	14	13	12	12	10	12	14	16	10
	fraction of +	0	0.55	0.32	0.52	0.58	0.45	0.42	0.39	0.39	0.32	0.39	0.45	0.52	0.32
	difference	0	0.06	0.02	0.02	0.06	0.12	-0.06	-0.01	0.01	-0.06	0.11	0.03	0.06	-0.06
Sharpe Ratio (VW)	# of +	0	18	11	17	16	23	16	16	18	13	16	10	17	14
	fraction of +	0	0.58	0.36	0.55	0.52	0.74	0.52	0.52	0.58	0.42	0.52	0.32	0.55	0.45
	difference	0	0.06	0.02	0.02	0.06	0.12	-0.06	-0.01	0.01	-0.06	0.11	0.03	0.06	-0.06
R^2_{oos}	# of +	0	18	11	17	16	23	16	16	18	13	16	10	17	14
	fraction of +	0	0.58	0.36	0.55	0.52	0.74	0.52	0.52	0.58	0.42	0.52	0.32	0.55	0.45

Table 7.9 Comparison of Performance between Augmented market-specific models and Original market-specific models

In the Table 7.7 the augmented models include both U.S. characteristic gaps and U.S. factors.

In the Table 7.8 and the Table 7.9 the augmented models include only U.S. characteristic gaps and U.S. factors, respectively. Focusing on the best NN models, Table 7.7 and Table 7.8 show similar results while table 7.9 is weaker. The best augmented NN models in Table 7.7 and Table 7.8 yield higher Sharpe Ratios by 0.07-0.22 on average, when compared with the best NN models using only 36 stock characteristics. Out-of-sample R^2 is 0.03-0.06 higher on average. Overall, information from U.S. stocks seems to be useful in producing better rankings of local stocks' predicted returns, and hence higher Sharpe Ratios. While NN models cannot explain why U.S. characteristic gaps are more important than U.S. factors, one possible reason is that the U.S. characteristic gaps contain information about the return premium of the characteristics in the global market. A wider spread may suggest that the corresponding characteristic becomes more important in the return generating process. U.S. factors may carry such information too but returns can be contaminated by noise and other variables.

On the other hand, local factors do not appear to help enhance return predictability.

8– CONCLUSIONS

Thanks to this work we have the possibility to shown two specific application of engineering solution in the financial sector.

Starting from blockchain we have implemented a real crypto mining server.

Al crypto value were utilized for cover the cost of energy and the investment for building the crypto mining server. It is a scientific prototype; the goal is to test the crypto mining server not to have some benefit.

The collaboration with the mining pool was very interesting sharing the knowloges that were very useful for the calibration of the crypto mining server. The crypto mining space is constantly changing as new technologies emerge. The professional miners who receive the best rewards are constantly studying the space and optimizing their mining strategies to improve their performance. As more and more people turn towards cryptocurrencies, there’s a vast increase in the need for experts in the field.

The crypto mining that has been built for this work is very ‘’basic’’, in the Figure 8.1 is shown one of the steps of the crypto mining construction

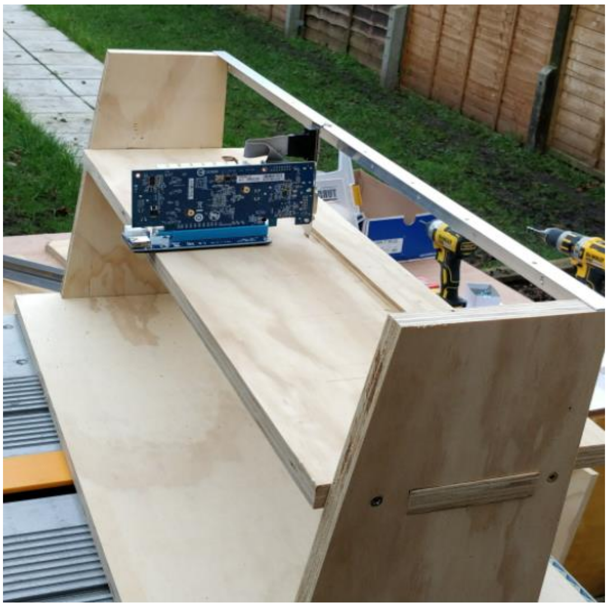


Figure 8.1 Crypto mining server construction

Now a days is possible to buy a crypto mining server already composed, we decided to build it by our self for budget reason.

The most important problem of the crypto mining server is the amount of energy consumed, infact according to data from the University of Cambridge's Bitcoin Electricity Consumption Index, the facility for the creation of Bitcoin consumes about 134 terawatt hours, a figure comparable to the consumption of a medium-sized nation. For example, Kazakistan is the second largest producer of digital currency in the world, after the United States. This is where 18% of global cropto-currency miners are located. Last year, something like 90 thousand companies chose this country after fleeing China due to the energy-related squeeze. In Kazakhstan, most of the electricity is still produced with coal and it's a problem for the environment, but now, to make up for the huge consumption of the cryptocurrency farms, they are studying to power them with alternative sources: wind, solar and even nuclear.

Other researchers have already studied different algorithms like the Green-PoW. The Green-PoW is an energy-efficient consensus algorithm that reduces the computation load to nearly 50% compared to the original Bitcoin’s PoW algorithm, without affecting the other properties of the system. The algorithm divides time into epochs, where each epoch consists of two consecutive mining rounds.

The second goals of our crypto mining server will be the study and implementation of a specific algorithms in other to reduce the consume of energy.

The second main technology that we have been studied is the machine learning. The finance sector is one of the key pillars of any nation’s economy. However, with the emergence of big data and rapid advancements in technology, the finance sector is processing significant amounts of heterogenous data. Institutions in the finance sector are increasingly using machine learning algorithms and techniques to process these heterogenous data. This exploratory review provides an in-depth look at the machine learning applications in the finance sector. The state-of-the-art machine learning applications in the finance sector were reviewed in this exploratory study.

Thanks to the collaboration with Alpha Beta company, we construct a dataset of 32 international markets and document common machine learning models performance in predicting the cross-section of stock returns. In the U.S. market, even with only 36 characteristics, the predictive power and profitability of complex machine learning models are comparable to those documented in previous studies using hundreds of variables (from Alpha Beta company). More important, training our models using U.S. data and applying them on international stocks a stringent test to address potential overfitting issues-concludes that neural network (NN) outperforms linear models, particularly in forming profitable portfolios. We achieve even stronger Sharpe ratios and out-of-sample R2 if we train the models separately for each market, so that the models can pick up market-specific return-characteristic relationships. However, there are signs that regression trees overfit the in-sample data and underperform linear models, especially in markets where there are few observations.

While the return-generating process seems to vary across markets, international markets are not totally segmented. Market-specific NN models are even more powerful when we add U.S. characteristic gaps and the interactions between stock characteristics and them respective U.S. characteristic gap as independent variables.

We conclude that NN models, previously focusing on the U.S. market, can be applied to equity markets around the world. With a reduced set of predictors, one can examine more closely the return-characteristic relationships generated by the algorithms and link them to the market-specific structure. For example, Leippold, Wang, and Zhou (2021) show that the most relevant variables when using NN models to predict Chinese stock returns are liquidity and fundamental factors, which they attribute to the short-termism of retail investors in China. Future research can provide more economic insights on other variables and other markets.

Another possible future research direction is to better explain the power of NN models using an asset pricing model.

The outperformance of NN models suggests that nonlinear and complex interactions among the predictors should not be overlooked, and the additional information carried by U.S. characteristic gaps in international markets is valuable. We adopt machine learning to help us understand return characteristic relationships and market integration, and it is interesting to see how the findings can be linked to equilibrium asset pricing.

References

[1] Abeyratne, S. A., & Monfared, R. P. (2016). Blockchain ready manufacturing supply chain using distributed ledger

[2] Aksentijević, S., Tijan, E., & Hlaća, B. (2009). Importance of organizational information security in port community systems. Information System Security, MIPRO.

[3] Erik Hofmann, Marco Rüsch (2017). Industry 4.0 and the current status as well as future prospects on logistics.

[4] Swan Melanie (2015). Blockchain: Blueprint for a new economy.

[5] Omran, Henke, Heines, Hofmann (2016). Blockchain-driven supply chain finance: Towards a conceptual framework from a buyer perspective.

[6] Biggs, Hinish, Natale, Patronick (2015). Blockchain: Revolutionizing the Global Supply Chain by Building Trust and Transparency.

[7] Cachin C. (2017). Architecture of the Hyperledger Blockchain Fabric.

[8] Korpela, Hallikas, Dahlberg (2017). Digital Supply Chain Transformation toward Blockchain Integration.

[9] IBM (2017). IBM Blockchain Platform Technical Overview.

[10] Mattila, J. (2016). The blockchain phenomenon - Tech. rep.BRIEWorkingPaper, UC Berkeley.

[11] Park, N. K., Choi, H. R., Lee, C. S., Kang, M. H., & Yang, J. W. (2005). Port management information system towards privatization.

[12] Ramberg, J. (1999). ICC Guide to Incoterms 2000. ICC.

[13] Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system.

[14] McGinnis, M. A., Boltic, S. K., & Kochunny, C. M. (1994). Trends in logistics thought: an empirical study.

[15] Froystad, P., & Holm, J. (2016). Blockchain: powering the internet of value.

[16] BircoinZ Community (2018). BitccoinZ – Community Paper A Community Gift To The World

[17] Jake Frankenfield, Amilcar Chavarria, Katharine Beer, Mining Pool, (2022).

[18] Andriy Burkov, The Hundre-Page Machine Learning Book (2019)

[19] Jaydip Sen, Rajadeep Sen, Abhishek Dutta, Machine Learing in Finance-Emerging Trends and Challenges, Department of Data Science, Praxis Busisnes School, INDIA, Independent Researcher and Financial Analyst 3School of Computing and Analytics, NSHM Knowledge Campus, Kolkata, India. (2021)

[20] Nabila hoamdoum, Impact of AI and Machine Learning on Financial Industry: Application on Moroccan Credit Risk Scoring, Journal of Advanced Research in Dynamical and Control Systems · November 2019

[21] David Choi, Wenxi Jiang, Chao Zhang, Aplha Go Everywhere : Machine Learning and International Stock Returns, 2021

[22] Azarnejad A., Khaloozadeh H, Department of Control and Systems, Faculty of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran, Stock return system identification and multiple adaptive forecast algorithm for price trend forecasting, 2021.

[22] Noureddine Lasla, Lina Alsahan, Mohamed Abdallha, Mohammed Yunis, Green-PoW: An Energy-Efficient Blockchain Proof-of-Work Consensus Algorithm, 2020

[23] J. T. Reason and J. J. Brand (1975) “Motion Sickness”, London: Academic press

[24] Gary E. Riccio and Thomas A. Stoffregen (1991) “An Ecological Theory of Motion Sickness Postural Instability”, Ecological Psychology, 3:195-240.

[25] Cruz-Neira, C., Sandin, D.J., DeFanti, T.A., Kenyon, R.V., Hart, J.C. (1992) "The CAVE: audio visual experience automatic virtual environment", ACM Commun., 35 (6), 64–72

[26] Pimentel, K., Teixeira (1992) "Virtual reality: through the New Looking Glass", TAB Books, NYC

[27] George L. Danek (1993) “Vertical Motion Simulator Familiarization Guide”, NASA Technical Memorandum, Mofflet Field, California 94035-1000.

[28] Kennedy, R.S., Lane, N.E., Berbaum, K.S., & Lilienthal, M.G(1993) “Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness, International Journal of Aviation Psychology”, 3:203-220.

[29] Fabiano B., Parentini, I., Ferraiolo, A., Pastorino (1995) “A century of accidents in the Italian industry, Safety Science”, 21: 65-74

[30] Giribone P., Bruzzone A.G. & Tenti M.(1996) "Local Area Service System (LASS): Simulation Based Power Plant Service Engineering & Management", Proc. of XIII Simulators International Conference, New Orleans LA, April 8-11

[31] Minister of transport Canada (1996) Helicopter flight training manual.

[32] Brown, L. (1999). “A Radar History of World War II”. Institute of Physics Publishing, Bristol

[33] Djellal, F., & Gallouj, F. (1999) “Services and the search for relevant innovation indicators: a review of national and international surveys”, Science and Public Policy, 26(4), pp. 218-232

[34] Bowman, D., Hodges, L. (1999) "Formalizing the design, evaluation and application of interaction techniques for immersive virtual environments", J. Visual Lang. Comput., 37–53

[35] Amico Vince, Bruzzone A.G., Guha R. (2000) "Critical Issues in Simulation", Proceedings of Summer Computer Simulation Conference, Vancouver, July

[36] Montgomery D.C. (2000) "Design and Analysis of Experiments", John Wiley & Sons, New York

[37] Bass, T. (2000) "Intrusion detection systems and multisensor data fusion" Communications of the ACM, 43(4), pp.99-105

[38] Hereld, M., Judson, I. R., & Stevens, R. L. (2000). "Introduction to building projection-based tiled display systems". IEEE Computer Graphics and Applications, 20. pp. 22-28.

[39] Nakatsu R., Tosa, N.(2000)"Active Immersion: the goal of Communications with interactive agents", Proc. of Int.Conf. on Knowledge-Based Intelligent Engine

[40] Joseph J., LaViola Jr., (2000), “A Discussion of Cybersickness in Virtual Environments”, Department of Computer Science, Brown University.

[41] Mobley, R. K. (2001) “Plant engineer's handbook”, Butterworth-Heinemann, Oxford, UK

[42] Majumdar, A., Polak, J. (2001). “Estimating capacity of Europe's airspace using a simulation model of air traffic controller workload”. Journal of the Transportation Research Board, (1744), 30-43.

[43] Richards A., J. Bellingham, M. Tillerson, and J. P. (2002) “How: Co-ordination and control of multiple UAVs”, Proc. of the AIAA Guidance, Navigation, and Control Conference, Monterey, CA, August

[44] Feddema, J.T.; Lewis, C.; Schoenwald, D.A., (2002) "Decentralized control of cooperative robotic vehicles: theory and application,

"Robotics and Automation, IEEE Transactions on, vol.18, no.5, pp.852,864, Oct

[45] Jacobson J.,Hwang Z.(2002) "Unreal tournament for immersive interactive theater", ACM Comm., 45 (1), 39–42

[46] Sherman, W. R., & Craig, A. B. (2003) “Understanding Virtual Reality—Interface, Application, and Design. Presence”, Morgan Kaufmann Publisher, SF, 12(4)

[47] Whisker, V. E., Baratta, A. J., Yerrapathruni, S., Messner, J. I., Shaw, T. S., Warren, M. E., Rotthoff E.S., Winters J.W., Clelland J.A. & Johnson, F. T. (2003). "Using immersive virtual environments to develop and visualize construction schedules for advanced nuclear power plants". In Proceedings of ICAPP Vol. 3, pp. 4-7

[48] Keating, C., Rogers, R., Unal, R., Dryer, D., Sousa-Poza, A., Safford, R., Peterson W. & Rabadi, G. (2003) “System of systems engineering”, Engineering Management Journal, 15(3), 36-45

[49] Vail D. & M. Veloso, (2003) “Dynamic multi-robot coordination”, Multi-Robot Systems: From Swarms to Intelligent Automata, Vol II, pp. 87-100.

[50] Stilwell D. J., A. S. Gadre, C. A. Sylvester and C. J. Cannell (2004) “Design elements of a small low-cost autonomous underwater vehicle for field experiments in multi-vehicle coordination”, Proc. of the IEEE/OES Autonomous Underwater Vehicles, June, pp. 1-6

[51] National Research Council. (2004) “Assessing the national streamflow information program”, National Academies Press, Washington DC, USA

[52] Stilwell D. J., A. S. Gadre, C. A. Sylvester and C. J. Cannell (2004) “Design elements of a small low-cost autonomous underwater vehicle for field experiments in multi-vehicle coordination”, Proc. of the IEEE/OES Autonomous Underwater Vehicles, June, pp. 1-6

[53] Jones, D. (2005) “Power line inspection-a UAV concept”, Proc. of the IEE Forum on Autonomous Systems, Ref. No. 11271, November

[54] Keating, C. B., Sousa-Poza, A., & Kovacic, S. (2005) “Complex system transformation: a system of systems engineering (SoSE) perspective”, Proc. of 26th ASEM National Conference, pp. 200-207

[55] Keegan, D. (2005) “The incorporation of mobile learning into mainstream education and training”, Proc. of World Conference on Mobile Learning, Cape Town, October.

[56] Shah, A. P., et al (2005, June). “Analyzing air traffic management systems using agent based modeling and simulation” In Proceedings of the 6th USA/Europe Seminar on Air Traffic Management Research and Development

[57] Bruzzone, A. G. et al. (2006) “Simulation and Optimization as Decision Support System in Relation to Life Cycle Cost of New Aircraft Carriers”,

Proceedings of Modelling Simulation and Optimization, Gaborone, Botswana

[58] Ross, S., D. Jacques, M. Pachter, and J. Raquet, (2006) “A Close Formation Flight Test for Automated Air Refueling,” Proceedings of ION GNSS-2006, Fort Worth, TX, Sep

[59] Grocholsky, B., Keller, J., Kumar, V., Pappas, G., (2006) "Cooperative air and ground surveillance", Robotics & Automation Magazine, IEEE, vol.13, no.3, Septmber, pp.16-25

[60] Jans, W., Nissen, I., Gerdes, F., Sangfelt, E., Solberg, C. E., & van Walree, P. (2006) “UUV covert acoustic communications- preliminary results of the first sea experiment”, in Techniques and technologies for unmanned autonomous underwater vehicles- a dual use view”, RTO Workshop SCI-182/RWS-016, Eckernförde, Germany

[61] Ross, S., D. Jacques, M. Pachter, and J. Raquet, (2006) “A Close Formation Flight Test for Automated Air Refueling,” Proceedings of ION GNSS-2006, Fort Worth, TX, Sep

[62] Doherty, P., & Rudol, P. (2007) "A UAV search and rescue scenario with human body detection and geolocalization", Proceedings of the Australian Conference on Artificial Intelligence, Vol. 4830, December, pp. 1-13

[63] Tanner H. G. (2007) “Switched UAV-UGV cooperation scheme for target detection”, IEEE International Conference on Robotics and Automation, Roma, Italy, April, pp. 3457-3462

[64] Tanner H.G., D.K. Christodoulakis, (2007) “Decentralized cooperative control of heterogeneous vehicle groups”, Robotics and Autonomous Systems 55,pp 811–823

[65] Tanner H. G. (2007a) “Switched UAV-UGV cooperation scheme for target detection”, IEEE International Conference on Robotics and Automation, Roma, Italy, April, pp. 3457-3462.

[66] Tanner H.G., D.K. Christodoulakis, (2007b) “Decentralized cooperative control of heterogeneous vehicle groups”, Robotics and Autonomous Systems 55,pp 811–823

[67] Tumer, K., Agogino, A. (2007, May). “Distributed agent-based air traffic flow management” In Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems (p. 255). ACM

[68] Shafer, A.J., Benjamin, M.R., Leonard, J.J., Curcio, J., (2008) "Autonomous cooperation of heterogeneous platforms for sea-based search tasks", Oceans, , September 15-18, pp. 1-10

[69] Bruzzone A.G., E. Bocca (2008) "Introducing Pooling by using Artificial Intelligence supported by Simulation", Proc.of SCSC2008, Edinbourgh, UK

[70] Jamshidi, M. (2008) "Introduction to system of systems. System of Systems Engineering", CRC Press, NY, pp. 1-43

[71] Mittal, S., Zeigler, B. P., Martín, J. L. R., Sahin, F., & Jamshidi, M. (2008) “Modeling and Simulation for Systems of Systems Engineering”, in Systems of Systems Innovation for 21th Century, Wiley & Sons, NYC

[72] Sousa-Poza A., Kovacic S., Keating C. (2008) "SoSE: An Emerging Multidiscipline", Int.Journal of Systems Engineering, Vol.1, Nos.1/2

[73] Rhodes, D. H., Valerdi, R., & Roedler, G. J. (2009) “Systems engineering leading indicators for assessing program and technical effectiveness”, Systems Engineering, 12(1), 21-35

[74] Stroeve, S. H., Blom, H. A., Bakker, G. B. (2009). “Systemic accident risk assessment in air traffic by Monte Carlo simulation”, Safety science, 47(2), 238-249.

[75] Sujit, P. B., Sousa, J., Pereira, F.L., (2009) "UAV and AUVs coordination for ocean exploration", Oceans - EUROPE, vol., no., pp.1,7, 11-14 May

[76] Tether, T. (2009) "Darpa Strategic Plan", Technical Report DARPA, May

[77] Sujit, P. B., Sousa, J., Pereira, F.L., (2009) "UAV and AUVs coordination for ocean exploration", Oceans - EUROPE, vol., no., pp.1,7, 11-14 May

[78] Tether, T. (2009) "Darpa Strategic Plan", Technical Report DARPA, May

[79] Prof. Alessandro Colonna ,(2009), “Costruzione di strade ferrovie e aeroporti”, University of Bari, Bari (Italy).

[80] Nolan, M., (2010), “Fundamentals of air traffic control”, Delmar, Boston, MA

[82] Mark McCall, Bob Mury, May 2010, Distributed Interactive Simulation, DIS PDG Chair.

[83] Jamshidi, M. (2011) “System of systems engineering”, Innovations for the twenty-first century, vol 58, John Wiley & Sons, NYC

[84] Ncube, C. (2011) “On the Engineering of Systems of Systems: key challenges for the requirements engineering community”, Proc. of IEEE Workshop on Requirements Engineering for Systems, Services and Systems-of-Systems (RESS), August, pp. 70-73

[85] Merabti, M., Kennedy, M., & Hurst, W. (2011) “Critical infrastructure protection: A 21 st century challenge”, Proc. of IEEE Int.Conf. on Communications and Information Technology, ICCIT, March, pp. 1-6

[86] Bürkle, A., Segor, F., Kollmann, M. (2011). "Towards autonomous micro uav swarms". Journal of intelligent & robotic systems, 61(1-4), pp. 339-353

[87] Cárdenas, A. A., Amin, S., Lin, Z. S., Huang, Y. L., Huang, C. Y., & Sastry, S. (2011) "Attacks against process control systems: risk assessment, detection, and response", Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security, March, pp.355-366

[88] Merabti, M., Kennedy, M., & Hurst, W. (2011) "Critical infrastructure protection: A 21 st century challenge", Proc. of IEEE Int.Conf. on Communications and Information Technology, ICCIT, March, pp. 1-6

[89] Bruzzone A.G., Fadda P, Fancello G., Massei M., Bocca E., Tremori A., Tarone F., D'Errico G. (2011a) "Logistics node simulator as an enabler for supply chain development: innovative portainer simulator as the assessment tool for human factors in port cranes", Simulation October 2011, vol. 87 no. 10, p. 857-874, ISSN: 857-874

[90] Bruzzone A., Massei M., Longo F., Madeo F., (2011b). Modeling and simulation as support for decisions makers in petrochemical logistics. Proceedings of the 2011 Summer Computer Simulation Conference, pp. 130

[91] Dai, F. (Ed.). (2012) "Virtual reality for industrial applications", Springer Science & Business Media

[92] Quero, S., Botella, C., Pérez-Ara, M. A., Navarro, M., Baños, R. M., Maciá, M. L., & Rodríguez, E. (2012). The use of Augmented Reality for safety in health: The European Project ANGELS. In ICERI2012 Proceedings (pp. 215-218). IATED

[92] Francesca Giardini, Frèdèric Amblard, (2012), "Multi-Agent-Based Simulation", Springer, Spain.

[93] Kastek, M., Dulski, R., Zyczkowski, M., Szustakowski, M., Trzaskawka, P., Ciurapinski, W., Grelowska G., Gloza I., Milewski S, Listewnik, K. (2012) "Multisensor system for the protection of critical infrastructure of seaport" In Proc. of SPIE, Vol. 8288, May

[94] Shkurti, F., Anqi Xu, Meghjani, M., Gamboa Higuera, J.C., Girdhar, Y., Giguere, P., Dey, B.B., Li, J., Kalmbach, A., Prahacs, C., Turgeon, K., Rekleitis, I., Dudek, G., (2012)"Multi-domain monitoring of marine environments using a heterogeneous robot team", Proc. of IEEE Intelligent Robots and Systems (IROS), vol., no., pp.1747,1753, October 7-12

[95] Spillane, J. P., Oyedele, L. O., & Von Meding, J. (2012) "Confined site construction: An empirical analysis of factors impacting health and safety management", Journal of Engineering, Design and Technology, 10(3), pp.397-420

[96] Bruzzone, A.G., Berni, A., Fontaine, J.G., Cignoni, A., Massei, M., Tremori, A., Dallorto, M., Ferrando, A. (2013c) "Virtual Framework for

Testing/Experiencing Potential of Collaborative Autonomous Systems”, Proc. of I/ITSEC, Orlando. FL USA

[97] Magrassi C. (2013) “Education and Training: Delivering Cost Effective Readiness for Tomorrow's Operations“,ITEC Keynote Speech, Rome, May

[98] Maravall D., J. de Lopea,b, R. Domíngueza, (2013) “Coordination of communication in robot teams by reinforcement learning”, Robotics and Autonomous Systems 61, pp.661–666

[99] Nano, G., & Derudi, M. (2013) "A critical analysis of techniques for the reconstruction of workers accidents", Chemical Engineering, 31

[100] Biocca, F., & Levy, M. R. (2013) "Communication in the Age of Virtual Reality", Routledge, London

[101] Pérez-Ara, M. A., Quero, S., Navarro, M. V., Botella, C., & Baños, R. M. (2013) “Augmented Reality for Safety at Work: Needs Analysis of the Priority Risks for Safety”, Proc. of Health Context in Spain, INTED, pp. 812

[102] Seidel, R. J., & Chatelier, P. R. (Eds.). (2013). "Virtual reality, training’s future? Perspectives on virtual reality and related emerging technologies". Springer Science & Business Media.

[103] Nystrom Robert , (2014), Decoupling Patterns, Game Programming Patterns, 5:215

[104] Pizzella, L. A. E. (2014) “Contributions to the Configuration of Fleets of Robots for Precision Agriculture”, Thesis, Universidad Complutense, Madrid, Spain, May

[105] Siebert, S., & Teizer, J. (2014) “Mobile 3D mapping for surveying earthwork projects using an Unmanned Aerial Vehicle (UAV) system”, Automation in Construction, 41, pp.1-14

[106] Valavanis, K. P., & Vachtsevanos, G. J. (2014) "Handbook of unmanned aerial vehicles", Springer Publishing Company, NYC

[107] Bruzzone A.G., Massei M., Agresta M., Poggi S., Camponeschi F., Camponesch M. (2014) “Addressing Strategic Challenges on Mega Cities through MS2G”, Proceedings of MAS, Bordeaux, France, September 12-14

[108] Clarke, R., & Moses, L. B. (2014) “The Regulation of Civilian Drones' Impacts on Public Safety” Computer Law & Security Review, 30(3), 263-285

[109] Pizzella, L. A. E. (2014) “Contributions to the Configuration of Fleets of Robots for Precision Agriculture”, Thesis, Universidad Complutense, Madrid, Spain, May

[110] Valavanis, K. P., & Vachtsevanos, G. J. (2014) "Handbook of unmanned aerial vehicles", Springer Publishing Company, NYC

[111] Siebert, S., & Teizer, J. (2014) “Mobile 3D mapping for surveying earthwork projects using an Unmanned Aerial Vehicle (UAV) system”, Automation in Construction, 41, pp.1-14

[112] Perlman, A., Sacks, R., & Barak, R. (2014). "Hazard recognition and risk perception in construction". Safety science, 64, pp. 22-31.

[112] Liu E., Theodoropoulos G.(2014)"Space-time matching algorithms for interest management in distributed virtual environments”, ACM TOMACS, vol.24

[113] Zhang, J., Hou, H. T., & Chang, K. E. (2014) “UARE: Using reality-virtually-reality (RVR) models to construct Ubiquitous AR environment for e-Learning context”, Proc. of IEEE Science and Information Conference (SAI), August, pp. 1007-1010

[114] Alzahrani, A., Callaghan, V., & Gardner, M. (2014) “Towards the Physical Instantiation of Virtual People and Components in Physical Mixed-Reality Tele-Presence Environments”, Proc. of Intelligent Environments Workshops, pp. 285-294, July

[115] Benes, F., & Kodym, O. (2014) “Application of Augmented Reality in Mining Industry”, Proc. 14th SGEM GeoConference on Informatics, Geoinformatics and Remote Sensing, Vol. 1, 1-35, June 19-25

[116] Hale, K. S., & Stanney, K. M. (2014). "Handbook of virtual environments", CRC Press, Boca Raton, FL

[117] Kenyon, A., Van Rosendale, J., Fulcomer, S., & Laidlaw, D. (2014). "The design of a retinal resolution fully immersive VR display", Virtual Reality (VR), IEEE, pp. 89-90

[118] Bruzzone, A. G., Massei, M., Tremori, A., Longo, F., Nicoletti, L., Poggi, S., Bartolucci C., Picco E. & Poggio, G. (2014b) “MS2G: simulation as a service for data mining and crowd sourcing in vulnerability Reduction” Proc. of WAMS, Istanbul, September.

[119] Bruzzone, A. G., Massei, M., Tremori, A., Poggi, S., Nicoletti, L., & Baisini, C. (2014a) “Simulation as enabling technologies for agile thinking: training and education aids for decision makers” International Journal of Simulation and Process Modelling 9, 9(1-2), 113-127

[120] Dorn A.W. (2014) “Aerial Surveillance: Eyes in the Sky” in Air Power in UN Operations: Wings for Peace (A. Walter Dorn, Ed.), Ashgate Publishing, Farnham, UK, pp. 119-134

[121] Harvey C., Stanton N.A., (2014). Safety in System-of-Systems: Ten key challenges, Safety Science, vol. 70, pp. 358-366

[122] Longo F., Chiurco A., Musmanno R., Nicoletti L., (2015). Operative and procedural cooperative training in marine ports. Journal of Computational Science, vol. 10, pp. 97-107.

[123] Kehoe, B., Patil, S., Abbeel, P., & Goldberg, K. (2015) "A survey of research on cloud robotics and automation", IEEE Transactions on automation science and engineering, 12(2), pp.398-409

[124] Kim, D. H., Kwon, S. W., Jung, S. W., Park, S., Park, J. W., & Seo, J. W. (2015) "A Study on Generation of 3D Model and Mesh Image of Excavation Work using UAV", Proceedings of the International Symposium on Automation and Robotics in Construction, Vol. 32, Vilnius, January

[125] Leão, D. T., Santos, M. B. G., Mello, M. C. A., & Morais, S. F. A. (2015) "Consideration of occupational risks in construction confined spaces in a brewery", Occupational Safety & Hygiene III, 343

[126] Merwaday, A., & Guvenc, I. (2015) "UAV assisted heterogeneous networks for public safety communications", Proc. of IEEE Wireless Communications and Networking Conference Workshops, March, pp. 329-334

[127] Apvrille, L., Roudier, Y., & Tanzi, T. J. (2015) "Autonomous drones for disasters management: Safety and security verifications", Proc. 1st IEEE URSI Atlantic, May

[128] Floreano, D., & Wood, R. J. (2015) “Science, technology and the future of small autonomous drones”, Nature, 521(7553), 460

[129] Kehoe, B., Patil, S., Abbeel, P., & Goldberg, K. (2015) "A survey of research on cloud robotics and automation", IEEE Transactions on automation science and engineering, 12(2), pp.398-409

[130] Kim, D. H., Kwon, S. W., Jung, S. W., Park, S., Park, J. W., & Seo, J. W. (2015) "A Study on Generation of 3D Model and Mesh Image of Excavation Work using UAV", Proceedings of the International Symposium on Automation and Robotics in Construction, Vol. 32, Vilnius, January

[131] Kovacevic, M. S., Gavin, K., Oslakovic, I. S., & Bacic, M. (2016). "A new methodology for assessment of railway infrastructure condition". Transportation research procedia 14, pp. 1930-1939

[132] Leão, D. T., Santos, M. B. G., Mello, M. C. A., & Morais, S. F. A. (2015) "Consideration of occupational risks in construction confined spaces in a brewery", Occupational Safety & Hygiene III, 343

[133] Merwaday, A., & Guvenc, I. (2015) "UAV assisted heterogeneous networks for public safety communications", Proc. of IEEE Wireless Communications and Networking Conference Workshops, March, pp. 329-334

[134] Bednarz, T., James, C., Widzyk-Capehart, E., Caris, C., & Alem, L. (2015). "Distributed collaborative immersive virtual reality framework for the mining industry" in Machine Vision and Mechatronics in Practice. Springer Berlin Heidelberg. pp. 39-48

[135] Chafkin M. (2015) "Why Facebook’s $2 Billion Bet on Oculus Rift might one day connect Everyone on Earth", Vanity Fair Hive, October

[136] Le, Q. T., Pedro, A. K. E. E. M., Lim, C. R., Park, H. T., Park, C. S., & Kim, H. K. (2015) “A framework for using mobile based virtual reality and augmented reality for experiential construction safety education”, International Journal of Engineering Education, 31(3), 713-725

[137] Muhanna, M. A. (2015) "Virtual reality and the CAVE: Taxonomy, interaction challenges and research directions", Journal of King Saud University-Computer and Information Sciences, 27(3), 344-361

[138] Safir, I. J., Shrewsberry, A. B., Issa, I. M., Ogan, K., Ritenour, C. W., Sullivan, J., & Issa, M. M. (2015). Impact of remote monitoring and supervision on resident training using new ACGME milestone criteria. Can J Urol, 22, 7959-7964

[139] Documentation Team (2016) GIMP- GNU Image Manipulation Program (User Manual).

[140] Bower, M., Lee, M. J., & Dalgarno, B. (2016) “Collaborative learning across physical and virtual worlds: Factors supporting and constraining learners in a blended reality environment”, British Journal of Educational Technology

[141] Bruzzone A., Longo F., Nicoletti L., Vetrano M., Bruno L., Chiurco A., Fusto C., Vignali G. (2016a). Augmented reality and mobile technologies for maintenance, security and operations in industrial facilities. 28th European Modeling and Simulation Symposium, EMSS 2016, pp. 355.

[142] Bruzzone A.G., Massei M., Maglione G., Agresta M., Franzinetti G., Padovano A. (2016b) “Virtual and Augmented Reality as Enablers for Improving the Service on Distributed Assets”, Proc. of I3M, Larnaca, Cyprus, September

[142] Bruzzone A.G., Massei M., Maglione G.L., Di Matteo R., Franzinetti G. (2016c) "Simulation of Manned & Autonomous Systems for Critical Infrastructure Protection", Proc. of DHSS, Larnaca, Cypurs, September

[143] Gonzalez, D. S., Moro, A. D., Quintero, C., & Sarmiento, W. J. (2016, August). Fear levels in virtual environments, an approach to detection and experimental user stimuli sensation”, Proc. of XXI IEEE Symposium on Signal Processing, Images and Artificial Vision (STSIVA), pp. 1-6

[144] Lindgren, R., Tscholl, M., Wang, S., & Johnson, E. (2016) "Enhancing learning and engagement through embodied interaction within a mixed reality simulation". Computers & Education, 95, 174-187

[145] Peña-Rios, A., Hagras, H., Gardner, M., & Owusu, G. (2016) “A Fuzzy Logic based system for Mixed Reality assistance of remote workforce”, Proc. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), July, pp. 408-415

[146] Scudellari M. (2016) "Google Glass Gets a Second Life in the ER", IEEE Specturm, May 25

[147] Spanu S., Bertolini M., Bottani E., Vignali G., Di Donato L., Ferraro A., Longo F., (2016). Feasibility study of an augmented reality application

to enhance the operators' safety in the usage of a fruit extractor. International Food Operations and Processing Simulation Workshop, FoodOPS 2016, pp. 70.

[148] Wagner K. (2016) "Two Years Later: Facebook's Oculus Acquisition Has Changed Virtual Reality Forever", Recode, March 24

[149] Bruzzone A.G., Maglione G.L. (2016) "Complex Systems & Engineering Approaches", Simulation Team Technical Report, Genoa

[150] Davis M., Proctor M., Shageer B., (2016). A Systems-Of-Systems Conceptual Model and Live Virtual Constructive Simulation Framework for Improved Nuclear Disaster Emergency Preparedness, Response, and Mitigation. Journal of Homeland Security and Emergency Management, vol. 13, no. 3, pp. 367-394

[151] Altawy, R., & Youssef, A. M. (2016) "Security, Privacy, and Safety Aspects of Civilian Drones: A Survey" ACM Transactions on Cyber-Physical Systems, 1(2), 7

[152] Bruzzone A.G., Longo F., Massei M., Nicoletti L., Agresta M., Di Matteo R., Maglione G.L., Murino G., Antonio Padovano A. (2016a) “Disasters and Emergency Management in Chemical and Industrial Plants: Drones simulation for education & training”, Proc. of MESAS, Rome, June 15-16

[153] Bruzzone A.G., Massei M., Longo F., Cayirci E., di Bella P., Maglione G.L., Di Matteo R. (2016b) “Simulation Models for Hybrid Warfare and

Population Simulation", Proc. of NATO Symposium on Ready for the Predictable, Prepared for the Unexpected, M&S for Collective Defence in Hybrid Environments and Hybrid Conflicts, Bucharest, Romania, October 17-21

[154] Bruzzone A.G., Massei M., Maglione G.L., Di Matteo R., Franzinetti G. (2016c) "Simulation of Manned & Autonomous Systems for Critical Infrastructure Protection", Proc. of I3M, Larnaca, Cyprus, September

[155] Pulina, G., Canalis, C., Manni, C., Casula, A., Carta, L. A., & Camarda, I. (2016) "Using a GIS technology to plan an agroforestry sustainable system in Sardinia", Journal of Agricultural Engineering, 47(s1), 23-23

[156] Sanchez-Lopez, J. L., Pestana, J., de la Puente, P., & Campoy, P. (2016) "A reliable open-source system architecture for the fast designing and prototyping of autonomous multi-uav systems: Simulation and experimentation", Journal of Intelligent & Robotic Systems, 84(1-4), pp.779-797

[157] Spanu S., M. Bertolini, E. Bottani, G. Vignali, L. Di Donato, A. Ferraro, F. Longo (2016) "Feasibility study of an Augmented Reality application to enhance the operators' safety in the usage of a fruit extractor", Proc. FoodOPS, Larnaca, Cyprus, September 26-28

[158] Steven M. La Valle, (2017), Virtual Reality, University of Illinois Cambridge University

[159] Di Donato (2017) “Intelligent Systems for Safety of Industrial Operators, the Role of Machines & Equipment Laboratories”, SISOM Workshop, Rome

[160] Palazzi, E., Caviglione, C., Reverberi, A.P., Fabiano, B. (2017) “A short-cut analytical model of hydrocarbon pool fire of different geometries, with enhanced view factor evaluation”, Process Safety and Environmental Protection, August

[161] Salvini, P. (2017) “Urban robotics: Towards responsible innovations for our cities”, Robotics and Autonomous Systems, Elsevier

[162] Gonzalez-Franco, M., Pizarro, R., Cermeron, J., Li, K., Thorn, J., Hutabarat, W. & Bermell-Garcia, P. (2017) “Immersive Mixed reality for Manufacturing Training” Frontiers, 4(3), 1

[163] Palazzi, E., Caviglione, C., Reverberi, A.P., Fabiano, B. (2017) "A short-cut analytical model of hydrocarbon pool fire of different geometries, with enhanced view factor evaluation”, Process Safety and Environmental Protection, August

[164] Tatic, D., & Tešic, B. (2017) “The application of augmented reality technologies for the improvement of occupational safety in an industrial environment”, Computers in Industry, 85, 1-1

[165] Bruzzone A.G., et al. (2017) "A STRATEGIC SERIOUS GAME ADDRESSING SYSTEM OF SYSTEMS ENGINEERING", Proc. of I3M, Barcelona, September

[166] Bruzzone A.G., Massei M., Agresta M., di Matteo R., Sinelshchikov K., Longo F., et al. 2017, "AUTONOMOUS SYSTEMS & SAFETY ISSUES: THE ROADMAP TO ENABLE NEW ADVANCES IN INDUSTRIAL APPLICATIONS", Proc. of MAS, Barcelona, September

[167] Guanhao Feng, Stefano Giglio, Dacheng Xiu 2020 ‘’TAMING THE FACTOR ZOO: A NTEST OF NEW FACTOR ‘’

[168] Joachim Freyberger, Andreas Neuhierl, Michael Weber 2020 ‘DISSECTING CHARACTERISTICS NONPARAMETRIC ‘’ May

[169] Sherhiy Kozak, Stefan Nagel, Shrihari Santosh 2020 ‘SHRINKING THE CROSS-SECTION ‘’

Web references

- it.wikipedia.org
- www.Blockchain4innovation.it
- www.internet4things.it
- bitcoin.org
- www.i-scoop.eu/industry-4-0
- www.IBM.com
- www.forbes.com
- www.bcg.com
- www.ethereum.org
- www.eventbrite.com
- www.ilpost.it
- techcrunch.com
- www.javatpoint.com/applications-of-machine-learning
- builtin.com/artificial-intelligence/machine-learning-examples-applications
- corporatefinanceinstitute.com/resources/knowledge/finance/financial-engineering/

Sorridere rafforza.