Analysis and performance of Machine Learning for Startup Valuation

Daniel Vento Lorente

September 2020

Reus (Spain)

UNIVERSITAT ROVIRA i VIRGILI

Accounting and Finance Bachelor's Thesis

Thesis director: Prof. Xiaoni Li, PhD.

Contents

Acknowledgment

I would like to acknowledge the efforts of all the professors who are passionate about what they do, they are essential to the development of all of us, and provide much more than academic content. I would also like to thank my family, friends and partner for being with me these last few weeks of stress in which Python, Conda, Sklearn, Linux, etc., filled my day with headaches. A special thanks to Marta, my computer science teacher.

Abstract

This work seeks to carry out a review of the state of the art on the use of Artificial Intelligence techniques (Machine Learning, Deep Learning) for the valuation of companies and Startups. On the other hand, an analysis is carried out about the most common Startup valuation models and their main weaknesses, facing the two biggest problems of these companies: uncertainty and the absence of data. Finally, an experiment is proposed using the Logistic Regression and Decision Tree algorithms, to study the capacity of these techniques in the prediction of startup failure.
It is concluded that in the last 2 years there have been numerous advances in AI techniques applied to the valuation of companies, which are more effective that traditional approaches. However, further research is needed on the analysis of qualitative information (such as sentiment analysis) and its possible applications in the valuation of companies.

Resumen

Este trabajo busca realizar una revisión del estado del arte sobre el uso de técnicas de Inteligencia Artificial (Machine Learning, Deep Learning) en la valoración de empresas y Startups. Por otro lado, se realiza un análisis de los modelos de valoración de Startups más utilizadas hoy en día y principales sus debilidades afrontando los dos mayores problemas de estas empresas: la incertidumbre y la ausencia de datos. Finalmente se propone un experimento utilizando los algoritmos Regresión Logística y Decision Tree, para estudiar la capacidad de estas técnicas en la predición del fracaso de las Startups.
Se concluye que en los últimos 2 años se han realizado numerosos avances en técnicas de IA aplicadas a la valoración de empresas, que resultan tremendamente más efectivas. Sin embargo, es necesario un mayor investigación en el análisis de información cualitativa (como análisis de sentimientos) y su posible aplicación en la valoración de empresas.

Resum

Aquest treball vol dur a terme una revisió de l'estat de l'art sobre l'ús de tècniques d'Intel·ligència Artificial (Machine Learning, Deep Learning) per a la valoració d'empreses i Startups. D'altra banda, es realitza una anàlisi sobre els models de valoració de les Startups més habituals i les seves principals debilitats, davant dels dos problemes més importants d'aquestes empreses: la incertesa i l'absència de dades. Finalment, es proposa un experiment utilitzant els algoritmes de Regressió Logística i Arbre de Decisions, per estudiar la capacitat d'aquestes tècniques en la predicció del fracàs d'Startups.
Es conclou que en els darrers 2 anys hi ha hagut nombrosos avenços en tècniques d'IA aplicades a la valoració d'empreses, que són més eficaços que els models tradicionals. No obstant això, cal

investigar més sobre l'anàlisi d'informació qualitativa (com ara l'anàlisi del sentiments) i les seves possibles aplicacions en la valoració de les empreses.

Keywords

Startup valuation, Innovative Business Valuation, Machine Learning

Palabras clave

Valoración de Startups, Valoración Innovadora de Negocios, Aprendizaje Máquina

Paraules clau

Valoració d'Startups, Valoración Innovadora de Negocis, Aprenentatge Màquina

Abbreviations and terms

Abbreviations

ADASYN: Adaptive Synthetic Sampling Method for Imbalanced Data

AI: Artificial Intelligence

BA: Business Angel

DL: Deep Learning

ML: Machine Learning

NN: Neural Network

UGC: User Generated Content

VC: Venture Capitalist

Terms

Business Angel: Individual that invests his money in an early-stage startup.

Venture Capital: Firm/Fund that provides investment to startups.

1 Presentation

Despite the college degree I have studied, one of my main interests in recent years has been programming and data analysis. Of course, I have been able to acquire sufficient motivation and the necessary skills in my faculty to fulfill my ambitions and materialize this project. In fact, my goal is set on the Master's Degree Thesis, as I think this project might be more interesting if an extensive Sentiment analysis is carried out to study the influence of this technique in the outcome of startups.

I would like to cite some of the courses that have motivated me the most and have given me really interesting knowledge of high value and usefulness, as well as thank all the professors for the effort made. Without a doubt, Mathematics I and Mathematics II have given me a solid base on multivariate optimization and on how to operate with matrices; Statistics I was one of the bases that established my passion for data analysis, but Statistics II gave me a global, practical and more complex vision that I could appreciate in all its fullness, it was a turning point when I learned about probability distributions, and it helped me to understand how Bitcoin works (thanks to the exercises on the Poisson Distribution). Finally, I should name two courses that have been essential to understand the issues set out in this work: Analysis of Financial Statements (along with various Accounting courses) taught me that not everything falls on statistics and the use of algorithms, but rather the fundamental study of a company can offer detailed and very valuable information for the investor (or the auditor), and Financial Management: Investment (or the other side of the coin: Financial Management: Financing), thanks to which I started this journey.

Neural networks and other Machine Learning methods have been gaining popularity over the last few years. From insurance companies to aerospace, including entertainment, marketing, healthcare, education and retail, just to name a few, these data analysis and prediction methods are here to stay.

So, why not merge both interests to carry out a Bachelor's thesis that offers an innovative approach (Artificial Intelligence techniques) on a booming sector, such as Startups?

Nowadays it is easier than ever to create a company, we live in a world connected thanks to the internet, in which we can find a multitude of resources and tools, free and paid, to help us take the step from the idea to the business; We have a world in which governments increasingly offer more economic freedom and each time we can allocate more resources to savings, and therefore to investment, understanding this as the acquisition of assets from third parties to obtain a benefit or use of assets for the creation of a business.

The ease of starting up a business is accompanied, at the same time, by the ease of investing in this buoyant sector. Currently, small and medium investors have at their disposal hundreds, if not thousands, of investment platforms in which they can find a wide range of sectors and investment amounts. Here we must make a paragraph to explain what we mean when we talk about small and medium investors, they are those investors who are not part of investment funds, nor do they have

the amounts of data or the necessary data analysis tools to analyze companies with an acceptable level of certainty. Therefore, there is a common problematic focus in terms of investment in Startups by small and medium investors, inherent in the existence of recently created companies. We refer to the absence of information. When an investor wants to invest independently, for example, in a NASDAQ company, he can obtain a lot of reliable data from those companies that are interesting to him, be it from independent reports, the abundant information on prices, the balance sheets made publicly available by the company itself. , etc., but when someone wants to invest in a company that has been running for just a few months, or a few years, they realize that there is hardly any information, and that the most abundant information turns out to be scattered qualitative information, found mainly in platforms of investment, press releases and social networks.
So, this raises the motivation of this work: **the analysis of the existing literature on the use of modern techniques for the analysis of companies with little quantitative information (including few fundamentals) and the creation of a small experiment to validate the latest trends and ideas about this type of analysis**.

2 Introduction

2.1 Relevance

Sociologists, psychologists and marketing specialists often use the term "infoxication", or more formally "information overload", which describes the state in which being exposed to an excessive amount and diversity of information produces a deficit in the assimilation and processing of it. Today this corresponds to the fact that a large amount of information is produced, distributed in a very dispersed way, due to the internet and the increase in the processing capacity of electronic devices (that generate such amounts of information). The term was coined by Bertram Gross in 1964, in his book *The Managing of Organizations*. Despite the negative connotations that this term implies, the reality it describes has many positive aspects, be it freedom of information, quick access to large volumes of data, etc., and the one that becomes interesting in this work: abundance of qualitative information that, in other times, would turn out to be a negligible quantity.

Therefore, people whose fields of work and / or research combine data science, financial analysis, investment services, risk assessment services, etc., have a great opportunity to improve the quality of their jobs, customer satisfaction and their profit optimization. One of the key aspects of this sector is the valuation of newly created companies, the so-called Startups, which are characterized by having a high innovative component (Natalie Robehmed, 2013) and being made up of multidisciplinary teams, that follow modern techniques for business organization and product and service development. (Souza, Ghezzi, Barbosa, Nogueira & Schwengber, 2020).

Two of the most important characteristics of Startups are uncertainty (Neumann, 2019) and the absence of quantitative information, the main topics of this work. Both components damage the financing possibilities of the company, especially when they fail to raise sufficient funds from Business Angels or investment rounds (in the US only 0.96% obtain funds from Venture Capital or Business Angels (Entis, 2013)). As explained before, small and medium investors rarely have information about the financial statements, key technologies and audit reports of a startup, etc., the opposite of what happens with a company with a solid trajectory and, especially, is listed on stock markets.

Therefore, efforts should be made towards the use of modern data analysis techniques, mainly when we speak of qualitative variables, such as Machine Learning or the several types of Neural Networks.

2.2 Artificial Intelligence, Machine Learning and Neural Networks

The use of the term "Artificial Intelligence" is often used in an ambiguous way. To understand this work and, in general, any project that deals with the applications of these systems, it is essential to clarify the difference between Artificial Intelligence, Machine Learning, Neural Networks and Deep Learning.

To define these concepts, the definition given by IBM (Kavlakoglu, 2020) is chosen, one of the most influential and long-standing industries in this sector (Enderle, 2020). Thus, Artificial Intelligence

encompasses the rest of the categories, each of them being a subset of the previous one (in the order mentioned above, that is Artificial Intelligence ¿ Machine Learning ¿ Neural Networks ¿ Deep Learning). The main distinction to clarify is that classified data is always used in Machine Learning, that is, the system must be trained by showing to it what the expected output is and, based on that, the system will try to learn the pattern that the data follows: then, once trained, we can use what we have learned to make predictions with a certain level of accuracy. For example, in the present case, we can "teach" an algorithm which startups have been successful and which ones have failed, so that it tries to infer what weight is assigned to each variable (e.g., financing rounds, number of investors, country) in order to create a predictive model. This contrasts with the more complex subset of ML, Deep Learning, in which we can use a set of classified data (called Supervised Learning) or unclassified (called Unsupervised Learning), which is extremely interesting when you want to know the pattern of data that has never been analyzed before or lacks information on how the independent variables (also called "Features" influence dependent variables (also called "Targets" or "Results"); this is also especially useful in classification problems, such as image recognition or determining which factors play a key role company's performance. Finally, an Artificial Neural Network (ANN) is a set of interconnected algorithms that mimic the function of biological neurons. They imitate these in the sense that the underlying structure of all Artificial Intelligence systems is made up of a minimum of three layers of artificial neurons connected to each other, which transmit information from one to another through what is called "activation function" .

These concepts are the basis of every ML and NN technique, and are explained below.

- Datasets

  The data we are going to work with, the values of one or more independent variables with their corresponding names. They can be images, texts, numbers, etc.; e. g., a table with thousands of data about the color, size and number of sales of certain car models.

- Features

  The independent variables of the dataset. Usually for Supervised Learning, the target value is included as an additional Feature, although, we are obviously referring to the dependent variable of the rest of the factors.

	object_id	normalized_name	category_code	status_bool	status	founded_at
1	c:1001	friendfeed	web	0	acquired	2007-10-01
2	c:10014	mobclix	mobile	0	acquired	2008-03-01
4	c:100155	mtpv	cleantech	0	operating	2003-01-01
6	c:100189	locatrix communications	mobile	0	operating	2003-11-01
7	c:100228	ihirehelp	education	0	operating	2010-10-01
8	c:100238	cardiosolutions	medical	0	operating	2006-01-01
9	c:100243	deepflex	manufacturing	0	operating	2004-01-01
10	c:10026	wevod	games_video	0	operating	2006-05-04
11	c:100271	balance financial	enterprise	0	operating	2004-01-01
12	c:1003	wikinvest	web	0	operating	2006-01-01
13	c:100379	g2 web services	ecommerce	0	operating	2004-01-01
15	c:1006	youlicit	web	1	closed	2006-09-01
16	c:100607	jogabo	games_video	0	operating	2011-01-01
18	c:10075	birdpost	web	0	operating	2007-11-08
19	c:100756	ambitious minds	education	0	operating	2009-01-01
20	c:10076	popego	advertising	0	operating	2007-09-01
21	c:10082	icharts	analytics	0	operating	2008-01-01

Figure 1: Example of dataset and its features (bold).

- Algorithms

  The element that determines how the information is processed. It consists of a process, a series of instructions that are implemented through mathematical functions, which receive an input (labeled data in the case of Supervised Learning and unlabeled data in the case of Unsupervised Learning), process it and obtain an output. There are numerous types of algorithms, the choice of these depends on the type of data we are going to work with (e.g, image, text, quantitative variables), the AI technique to use (e.g, ML, DL), the structure of our neural network (e.g, Convolutional, LSTM), its processing speed, the application or not of certain techniques (e.g, Backpropagation, a widely used algorithm), etc. Most neural networks use several algorithms.

- Neurons

  The base component of any Neural Network system (Wikipedia, 2020). It consists of a mathematical function, called "Activation Function" (sometimes also called "Transfer Function") that receives one or more weighted and summed inputs, and generates an output; depending on the output value, the neuron will be activated ("fired") or not depending on the relevance of the input for the optimization of the model, usually the output is transferred to each of the neurons on the next layer (Gurney et al., 1997). There are different types of activation functions for different uses, one of the most common being the Sigmoid, which produces an output value between 0 and 1.

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

Eq. 1. The Sigmoid function that output a value between 1 and 0.

$$u_k = \sum_{n}^{i=1} w_i x_i + b \tag{2}$$

Eq. 2 represents the weighted sum of i neurons for neuron k, where w_i represents the weight of the i^{th} neuron of the previous layer, x_i represents the input of the i^{th} neuron of the previous layer and the b the bias that some models add to the function.

- Layer of neurons

  They are simply an aggregation of neurons that share the same activation function. Normally, each neuron in each layer transfers its output to all the neurons in the next layer, the input of the first layer (in the first iteration, if there are several) being external data, that is, the dataset.

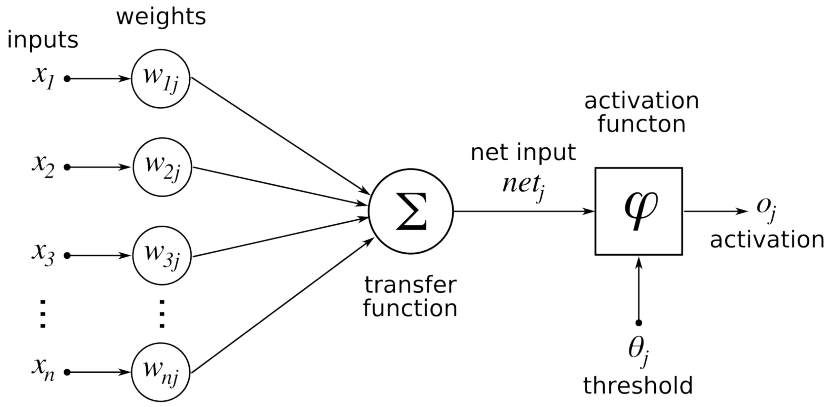

Figure 2: Artificial neuron main parts. Credit: Chrislb

- Neural Network

  It is the connected set of layers of neurons. Normally, we distinguish between single-layer networks, whose first algorithm developed, still in use today, was the "Perceptron" (Rosenblatt, 1957), and multilayer networks, for which the Perceptron is also commonly used. There are fundamentally two types of multilayer Neural Networks:

  – Simple The basis of modern neural networks. Its structure consists of a first layer that receives the inputs, the next layer called "Hidden layer", where a non-linear transformation is applied, and an output layer, which receives the outputs of the hidden layer as input.

  – Multilayer

    They are exactly the same as the previous ones, but with more than 1 hidden layer, they are common today. As a general rule, the more complex a problem, the more hidden layers are added up to some extent, as they allow for deeper learning, although this is widely disputed, and there is no clear consensus on how many hidden layers should be used (Brownlee, 2019).

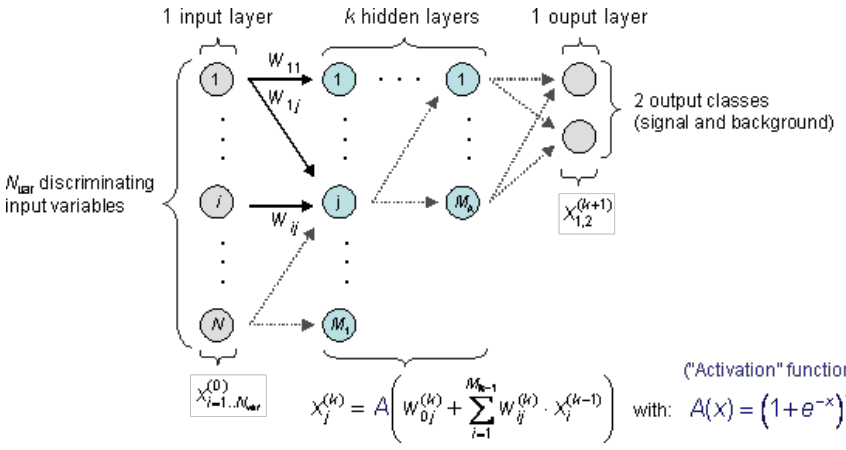

Figure 3: Example of a neural network with k layers and Sigmoid as activation function. Credit: Root CERN

2.3 Goals

This project has several objectives.

- First, review the state of the art on the use of Machine Learning and Deep Learning techniques for the valuation of companies and startups. Also, within this section, explain, broadly speaking, elementary concepts in this area, so that it also has an instructive purpose.
- Second, perform an analysis of the most used methods for the valuation of startups, analyzing their weaknesses.
- Third, carry out a small experiment that demonstrates the ease of use of Machine Learning with an acceptable level of prediction about the success or failure of startups.

2.4 Startups and investors. Some definitions

There are many valuation models for companies and it is one of the oldest areas of modern finance. However, not all companies are the same, there are, for example, listed companies and private companies, although the latter can be large, they cannot be an investment choice. In this work, the main distinction is between startups and other companies.

As stated in the introduction, "a startup in a recently created company that is characterized by having a high innovative component (Robehmed, 2013) and being made up of multidisciplinary teams that follow modern company organization and product development techniques and services (Souza et al., 2020)".

One could argue that startup valuation is a widespread branch of finance that has been around for decades, but here we find a problem. Most of the existing literature on startup valuation is fundamentally focused on individuals or companies that either have the necessary resources to carry out a broad quantitative analysis of investments, or can attract target companies (due to the large capital available to them) so that they offer internal information to receive investment. From this it is deduced that most of these investments are outside of small and medium investors, which are the motivation for this work.

What is considered a "small or medium investor"?

this is a difficult question, since it depends on several factors, for an investment round of $ 500-1,000 million, it could be ¡$ 5 million, although this type of investors rarely have access to such large rounds, unless they are IPOs . However, for a round of a few million dollars, we could define this type of investor as one who can contribute less than $ 100,000 - $ 300,000 in a single investment.

The investment rounds are usually focused on large investors, or not?

This is not true, since for many years many initiatives have been financed through Crowdfunding, being one of the most used funding sources for small initiatives2. Crowdfunding is based on financing through monetary contributions in exchange for exclusive goods and services of the company,

instead of the acquisition of capital.

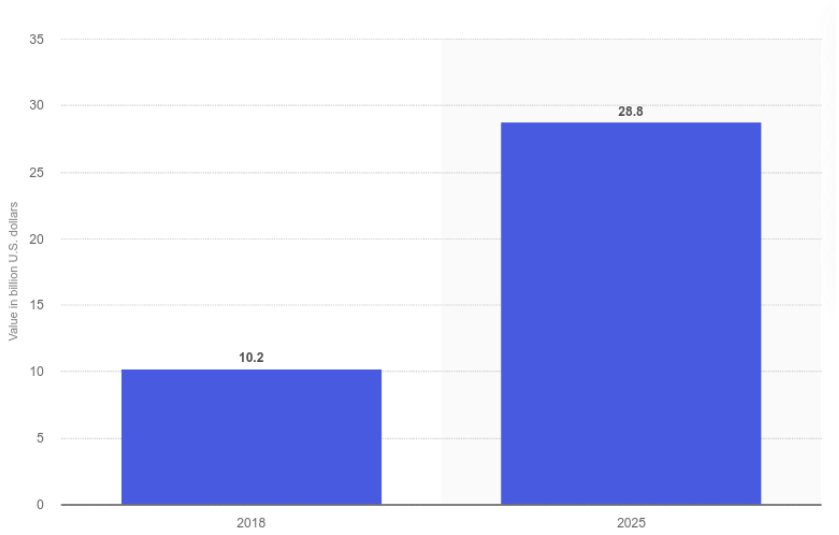

Figure 4: Value of crowdfunding in USD (billion) in 2018 and 2025 forecast. Credit: Statista

Along with this, in recent years online investment platforms for non-accredited individuals have proliferated, such as WeFunder, OneVest, Seedrs, etc. These platforms offer a fast, easy and safe way to invest small amounts (usually above $ 1,000) in startups in exchange of shares. They usually offer qualitative information, and even the possibility of contacting the founders.

So, the goal is to make it clear that the investment possibilities for this type of individuals are real and they are booming, therefore, a new approach is needed for the valuation of startups.

3 Literature review

3.1 Relevance of Artificial Intelligence in Finance

Artificial Intelligence techniques have been widely used in Finance for a long time, one of the first times they were used in this area was in the 1980s, using neural networks to predict the value of IBM shares (White, 1988). Since then, the use of AI in finance has been increasing, having experienced a significant increase in the last two decades, being the decade 2010-2020 when the use of intelligent algorithms in data analysis became widespread; a milestone that has been made possible by the increase of computers' processing power and the generation of large and diverse volumes of data. As a note before reviewing the state of the art, it must be said that most research uses Supervised Learning, given the ease of use and training of this type of AI.

Following what White showed in 1988, a multitude of research has been published in recent years. Nazemi & Tahmasbi (2013) compare Markowitz's mean-variance model, along with other alternative models, to compare various aspects of portfolio optimization with an artificial neural network. The results are that the NN is not outperformed by any of the traditional models, with special emphasis on the fact that the network can be widely scaled and can solve multitude of problems whose variables change over time. Another of the areas in which Machine Learning is used the most is High Frequency Trading (HFT), consisting of the negotiation of financial values using algorithms that work in a time frame of milliseconds or even microseconds (Haldane, 2010). One of the scenarios where HFT is most used is when arbitrage is present, a human could not take advantage of an arbitrage operation, given the short time that the differences in values are available, but an algorithm specially designed for this can.

A few years later, a new approach is tested (Aguilera-Rivera, Valenzuela-Rendón, 2016) for the optimization of multi-period portfolios to outperform traditional methods, using an evolutionary algorithm, managing to reduce the risk compared to traditional methods.

Deep Neural Networks can also be used for portfolio optimization (Lachiheb, Salah, 2018). In his paper, a DNN is used to predict, at 5-minute intervals, what the price of various stocks of a stock index will be. The predictive capacity of the network is improved by the introduction of a hierarchical design, which consists of the use of several neural networks within a neural network, so that the internal ones divide the processes to be carried out by groups of inputs (Mavrovounioti & Chang , 1992). The proposed network model offers better performance than others based on Simple Neural Networks and even DNN, achieving 73% accuracy.

In the short term but (but no as short as in HFT), NN have also been used for the prediction of stocks of the NASDAQ index with periods of 4 and 9 days (Moghaddam, Moghaddam & Esfandyari, 2015). In this case, a feed forward ANN trained using Backpropagation is used, an algorithm widely used in many AI applications. This algorithm calculates the gradient descent in the opposite direction the network is running, minimizing the cost function by adjusting the weights and biases and, therefore, reducing the error (Kostadinov, 2019) (Rumelhart, Hinton & Williams, 1986).

In the following work (Von Spreckelsen, Von Mettenheim & Breitner, 2014) a feed forward Multi-Layer Perceptron with only one hidden layer and another with the same configuration but hybrid (that is, using it together with Black's model for the pricing of options), for currency options futures (EUR / USD) pricing in 15-minute intervals. It is concluded that the network obtains acceptable results but slightly below the traditional methods.

We can find another interesting perspective on the use of AI to predict the bankruptcy of large companies (de Andrés, 2003). In this research MultiLayer Perceptron (MLP) is used to predict the bankruptcy of companies in the Spanish banking sector, which suffered a crisis in the years 1977-1985. The NN is more efficient than the classical multivariate methods, except in the type II error (classifying a solvent company as bankrupt). However, the research reveals a very interesting fact: the NN does not show which factors have been decisive in the prediction, while the classical methods, despite being less effective, are more explanatory. This is an inherent limitation of the existence of hidden layers and one of the topics that arouses the most interest among researchers (Fan, Xiong & Wang, 2020). Finally, in the same work carried out by de Andrés, an MLP is also used for the multifactorial estimation of the performance of securities portfolios, obtaining better results than its traditional counterparts (in this case, Seemingly Unrelated Regressions (SUR)).

Neural Networks have also been used to predict the credit rating of companies (Caridad, Hančlová, Woujoud & Caridad, 2019) given by rating agencies (S&P, Moody's, etc.). A positive point on this is that the paper only uses public data, making the experiment easily reproducible and showing that only with this data an acceptable model could be created. A 58.10% accuracy is only achieved if you want to predict exactly what grade the company receives, however, if a margin of error of two classes above or below is allowed, the percentage of success stands at 80.24%.

One of the most used AI techniques is clustering, a type of Unsupervised Learning, it has also been used to classify companies into different groups according to the most important factors that the algorithm decides (Machová & Vochozka, 2019). In this work, big companies in the Czech Republic are classified and then each cluster is analyzed, with the aim of creating a predictive model on the bankruptcy of such companies. One of the interesting conclusions from clustering is that larger companies benefit from higher leverage.

Finally, it is worth highlighting the use of Robo-advisors, being, perhaps, the application of AI in the financial field more focused on the main customer of financial services. Belanche, Casaló & Flavián (2019) carry out an interesting research on the adoption of these services by clients. Robo-advisors refers to financial advisory services provided by algorithms rather than human advisers; Nowadays, these platforms have proliferated (especially in the format of "applications" for mobile devices) that create risk profiles and make investment recommendations in a fully automated way, being one of the pillars of the new "Fintech" sector. The research uses the Technology Acceptance

Model (TAM) framework to understand customer reactions to the introduction of technological innovations, in this case, robo-advisors, by conducting a survey. The ease of use to investment services is the main positive reason for the adoption of these algorithms and, as a curiosity, demographics do not alter this preference.

In summary, we see that different AI techniques and models outperform traditional finance methods in most of the scenarios raised in the last 20 years, so we can conclude that these are tools that have become essential in this sector, and that they have a long way to go.

3.2 Use of Machine Learning and Neural Networks for Business Valuation

In recent years there has been a growing increase in the publication of articles on new approaches to analytics and startups. This has been motivated by the use of new analysis techniques and the analysis of new types of data that were unthinkable a few decades ago, such as the content publish on social networks by companies and their members.
Despite this, the number of publications is still very small compared to other fields, and yet both startups and investors can greatly benefit from research on the use of these techniques.
We can divide the work carried out into two types, those in which traditional data is used (with more weight on quantitative variables) and those in which a new approach is used using large datasets with qualitative variables of various kinds.

Use of new approaches for traditional datasets
One of the most concrete and quantifiable works (Kirshna, Agrawal & Choudhary, 2016), tries to predict the result of startups (11,000) based on variables such as seed funding amount, funding time, total funding, etc., doing a comparison between various ML techniques (eg, Random Forest, NaiveBayes, ADTree), using more than 30 variables. This paper's goal is not about profitability perspective by an investor, but the importance for the entrepreneur of determining which factors influence the success of a startup the most, since the data used is rarely available to investors.
With this, we are talking about a classification problem, such as Logistic Regression, with which the result of startups is predicted with a fairly good success rate.
A paragraph should be made to comment on an interesting paper (Van Gelderen, Thurik & Bosma, 2005) in which it the researches try to determinate the key failure factors in the pre-startup phase, from the perspective of human capital, following up on a 3-year period of 517 entrepreneurs who founded 310 startups in Norway, Sweden, The Netherlands and the USA. Logistic Regression is used, and it is concluded that: 1) The more capital is required before starting a business, the less likely it is to start working on it (because it is more difficult to obtain a large sum of money for a project not yet running), 2) Entrepreneurs who perceive less market risk, create their business before, regardless of whether the perception is correct or not, 3) The existence of a business plan

positively affects less ambitious people, but negatively the most ambitious people. As we can see the impact of the conclusions on the development of the company itself is minimal, since we are talking about a pre-entrepreneurship phase, while in this work an analysis of already founded startups is carried out. However, it is important to mention it, since the whole beginning of any company is based on an idea that one or more people want to bring to the market.

Random Tree and Random Forest have acquired an important relevance, and thus, an explanation on how they work must be given.

Random Forest is a combination of several Random Trees (as the name intuitively indicates), both of which are classification algorithms. The latter is an algorithm that seeks to classify the data according to the variables we gave it, prioritizing those variables that allow the classification into groups that are different from each other, starting with the largest groups and ending with small groups.

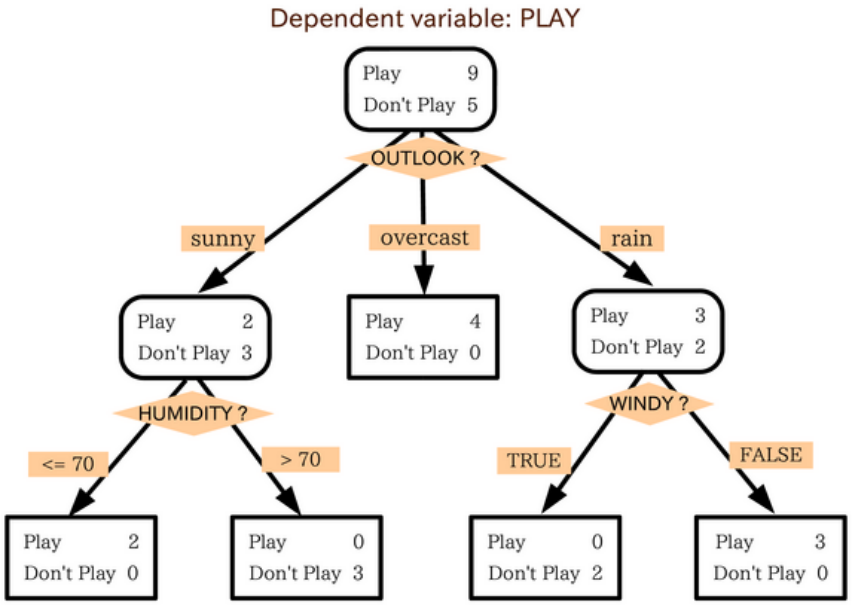

Figure 5: Decision Tree algorithm. Credit: T-Kita

Random Forest, as we said, is simply the use of several Random Trees, called "ensemble learning". In this way, it is possible to reduce the excess of variance that we find in the Trees, achieving a slight increase in bias as a negative aspect.

There's an extensive and recent research on the use of different ML techniques to predict the outcome of a startup (Ünal, 2019), using a relatively updated database and using techniques to improve the objectivity of the sample, given certain typical problems such as class imbalance.

Class imbalance is a fairly common problem in statistics and data analysis, which describes the lack of proportion between the classes under analysis (Sammut & Web, 2010). This is due to the fact that the existence of a class (or several) with a greater number of datapoints creates a bias in the model towards those majority classes, causing it to have a low accuracy on the minority class.

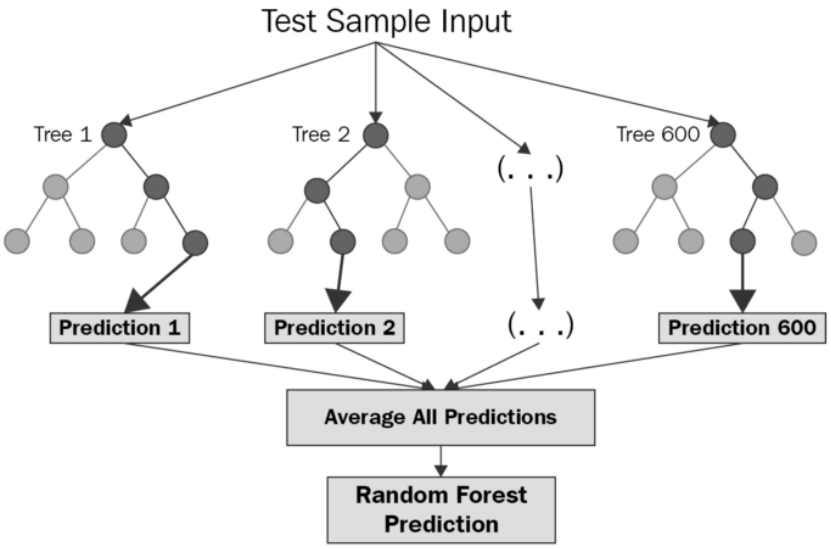

Figure 6: Random Forest algorithm. Credit: Chakure

Most of these problems are solved with one of the following techniques.

1. Get more data of the minority class. This is ideal, because new datapoints are obtained naturally that will give the model relatively objective proportionality.

2. Reduce the majority classes. This also maintains the nature of the data, although it is not always possible, as the dataset may end up being too small.

3. Generate synthetic datapoints. Since the simple duplication of data would generate linear dependence, several algorithms created in the last 20 years come into play here. The most famous being ENN, SMOTE and ADASYN (improved version of SMOTE, which will be explained in the following chapters).

According to the author of this work, first of all, a distinction must be made between the performance of a startup and the success of a business, concluding that the viable status of a startup cannot be evaluated based on the generation of profits either because most of them do not have historical data or do not generate benefits during the early stages of development, nor the analysis methods that use comparison with other companies in the sector be used can, either because the startup is disruptive, generating a market, or by the lack of data on the development of its products and / or services.
Another notable distinction is the one made regarding corporate bankruptcy and the failure of startups, that concludes the main source for this type of analysis comes from financial ratios and other data obtained from the public financial statements, which are not applicable to a startup.

A dataset composed of 44,522 startups and 19 variables is used. Here the ADASYN algorithm comes into play to solve the imbalance class (He, Bai, Garcia, & Li, 2008), oversampling the minority class. One of the conclusions is the little influence of total funding (using Simple Logistic Regression) on the success of the company, due to a combination of negative factors, the former being the trust that investors place in startups considered profitable, and being the second the cash-burning effect. Finally and with this method, they obtain an accuracy of 77%.
An insight describes that the predictive capacity of the model begins to decrease after the 3rd year. Next, other classification methods are used, the most striking being the Random Forest, with an error rate of 5.9% and the "Extreme gradient boosting", with an error rate of only 5.6%.
Finally, one of the problems this paper has is the excessive use of oversampling, which could have perhaps been reduced by changing the composition of the variables or the composition of the cleaned data. Obviously, the other problem consists in the use of variables that are sometimes not available to the small and medium investor.

Use of qualitative data

Without a doubt, it is the most interesting part because it is precisely the lack of quantitative information that produces uncertainty and leaves a large number of investors out of the possibilities offered by startups with little experience.
Several works have been carried out around how qualitative information influences startups, mainly using social networks, finding some very extensive and detailed works on such specific topics as Venture Capitalists (Hadley, 2015). In this work, two networks are generated, one formal, between managers, and the other being informal, on Twitter. This is one of the most qualitative papers on this topic, making it very interesting for its potential applications in the future. Several network theory approaches are applied. Solid conclusions on how board members and networks affect the startup success, and other insights about the success of these businesses are presented. VCs are much more relevant in both networks than non-VC members, startups owned by the former enjoy greater popularity and receive more funds.
Other interesting conclusions are the positive correlation of the number of founding members of a startup with its success.
Ling (2015) analyzed the impact of buzz in the success of startups. It did so through the valuations that VCs make of these companies, taking into account the amount of funds that are invested and their relationship with the existing User Generated Content (UGC) on social networks. Another finding was that currently VCs do not take advantage the full potential of the market excitement on new products and services, especially in the early stages of financing.
One especially interesting point comes from the dataset used. The author joined two datasets, one made up of nearly 50,000 startups and the other made up of a survey of 1,000 college students about innovative applications they use in their day-to-day. As a result, she discovered that most of the ideas mentioned corresponded to the most recent companies, which are the ones that more

funds had obtained and which more funding rounds had, so she concluded that user referral and market excitement were of key importance in the development and success of startups.

It demonstrates, as expected, a relationship between user buzz and VC investments. It also shows that top VC firms do not take advantage of the buzz as much than smaller VC firms, however, there is no evidence that the time that a company has been in the market does guarantee an increase in interaction with users.

Finally, the results of the study are proposed as a tool to increase the funds obtained in the subsequent investment rounds of startups, through the design of large user acquisition campaigns on social networks.

A paper based solely on tweets (Saura, Palos-Sanchez & Grilo, 2019), shows some interesting results on what factors positively or negatively affect the success of the company. A "Latent Dirichlet Allocation" (LDA) is used to identify the keywords of a dataset, to later identify and classify the sentiments with a "Support Vector Machine" (SVM) algorithm and finally using the quantitative analysis software Nvivo. The researchers found that important factors, such as the BAs that invest in the company, the jobs they offer, and the programming languages and frameworks they use, negatively affect the development of the company. Some of the negative aspects of the work are the relatively small size (34,000 tweets) of the sample (since for text analysis, a notable improvement is obtained the larger the sample (Kharde, Sonwane, 2016) and the topic selected for filter tweets ("#startups"), which leaves out many potential tweets.

4 Startup valuation models

4.1 Most common models

Before evaluating which AI methods could be used for the valuation of companies, it is convenient to define which are the traditional methods and why they could have a more efficient alternative.

EBT, EBIT, EBITDA

The base of the pyramid of many methods is usually formed by the EBITDA (Earnings Before Interest Taxes Depreciation and Amortization) or EBIT.

$$EBITDA = Revenues - Expenses; EBT = EBITDA - Interest - Depreciation - Amortization \tag{3}$$

Other methods are also commonly used, such as the Cash Conversion Cycle together with the Working capital. However, these two methods are rarely useful for the valuation of startups, since they depend on data that is usually not available (e.g., Average Accounts Payable, Goods Inventory Turnover), except if we have access to management reports, costs and balances.

NPV and DCF

A widely used method is the Net Present Value (NPV), which is the present value of the future cash flows offered by an investment, less the initial investment.

$$NPV = -C_0 + \sum_{T}^{t=1} \frac{C_t}{(1+r)^t} \tag{4}$$

As and implicit part of it, we have the Discounted Cash Flow (DCF).

$$DCF = \sum_{T}^{t=1} \frac{C_t}{(1+r)^t} \tag{5}$$

When the NPV = 0, obtain the Internal Rate of Return (IRR), which we can compare to the market rate (also called Opportunity Cost of Capital), what options are we discarding when we invest in a certain project, does the investment exceeds the market average rate of return?

Cost-to-duplicate

It takes into account the total cost of the company, so that its total is calculated (including RD, capital contributions, etc.) and it is assumed that an investor would not invest more than it costs to duplicate the company (Mass Challenge, 2019). The biggest disadvantage of this method is that it does not take into account the future potential of the company.

Multiples

With this method, the amount of Price-Earnings-Ratio, sales, etc., of different companies in the same sector of the target company are searched, and a multiple is calculated based on the value of the chosen variable, which can be applied to the company in which you want to invest (Investopedia, 2019).

VC-ROI

One of the most used approaches by VCs (The Business Professor, 2015).

$$ROI(ExpectedReturnonInvestment) = \frac{ExitValue}{Post - moneyValuation} \tag{6}$$

$$Post - moneyValuation = \frac{ExitValue}{ROI} \tag{7}$$

Where the exit value is usually calculated as a multiple of the sales forecast for the chosen exit moment, and the ROI is a minimum value chosen by the investor.

Berkus

Created by Venture Capitalist Dave Berkus (Berkus, 2012), it is a method that takes into account 5 categories: 1) Basic value, 2) Technology, 3) Execution, 4) Strategic relationships in its core market and 5) Production and sales.

Monetary values are assigned to these variables according to the performance of the company, to later aggregate them and obtain a value that determines the value of the company.

Risk Factor Summation

An initial monetary valuation is taken using, for example, the Berkus approach. The key risks of the company are then identified and assigned a range of positive and negative values, with which the company's valuation will increase or decrease (Payne, 2011).

4.2 The problem

The problems that startup valuation focuses on can be divided into two categories: uncertainty and lack of qualitative information.

Uncertainty

Uncertainty is inherent in any investment, even when acquiring "safe" financial instruments such as government bonds, there is always a small risk, in this case, such as a natural catastrophe, a world crisis, etc. So the relationship between risk and return must be remembered shows up, which

follow an inversely proportional relationship.

The more disruptive and young a business is, the more uncertainty there will be regarding its development, which is why startups perfectly fit this description. Going a little deeper, some of the factors that most affect this, are the absence of benefits (and even income) in the early stages of development, the registration of patents and trademarks that can generate legal conflicts and delays in development, the country risk (it can be calculated with indicators such as the Economic Freedom Index), industry competition (a startup can be innovative in the approach of a product / service in an existing market where, established companies, can see the new company as a threat, and try to mimic its approach by leveraging its own resources and the existing user base), the need for funds before generating profits, etc.
This risk is difficult to address, given the number and complexity of the factors, but it can always be minimized to some extent, depending primarily on the information available.
Minimizing this risk usually has two ways out in the case at hand.
a) Obtaining more (or higher quality) quantitative information. You can always try to do a more thorough search. For example, comparing with companies in the same sector, contacting entrepreneurs, looking for articles and other communications in other languages, etc.
b) Obtaining qualitative information through the use of AI techniques, such as sentiment analysis, especially in social networks. As mentioned in the review of the state of the art, it has been shown that this technique can be useful in the valuation of companies (Saura et al., 2019) (Ling, 2015).

Lack of quantitative information

As mentioned before, the absence of quantitative information is the main problem for startups, which separates them from other types of business. The main difference between new companies and established companies is that many of the latter publish annually (or less frequently) several reports and balance sheets, so that one can perform different assessments and calculations based on these data.
With this, we can distinguish two main sources of quantitative information.
The target startups. As mentioned in previous chapters, investment platforms in startups have proliferated, whether for capital contributions or not. It is common on these sites to find information on sales, number and type of customers, forecasts, projects, etc. This has two main problems. The first is that accessing this information requires registration in each of the platforms (and, usually, these have an exclusive contract with the startups once a financing round has started), with the consequent dispersion and loss of time . The second problem is bias, since the information that we find on these sites comes from the startup itself and, contrary to what would happen with a listed company, there is no auditing process that reviews it, nor is it public information (which could be contrasted by millions of people).

Data banks on startups. Currently there are several semi-public databases on startups, such as

Crunchbase, Zoominfo, Seedtable, etc., where we can find information regarding the number of financing rounds, total amount of financing, number of founders, etc. There are two problems with these sites as well. The first is that many only give free access to part of the information (which is obvious), with full access at a monthly subscription; and even so, being limited most of the time according to the amount of information that you want to obtain. The other problem results from outdated information or lack of information. For example, as has been observed for this work, it might take months or years until the information about the latest investment round of a startup is available in one of these databases; it gets harder in the case of a small startup, it might not be directly available.

However, these sites often offer complete but outdated information for research purposes, such as that used to carry out this work.

The solutions for implementing a model with updated information are two.

a) The payment of services that the data banks are offer, so that all information can be accessed.

b) Text and sentiment analysis techniques are used to search for information on the web about articles and press releases referring to the target startups, so that information on financing rounds, investors, sales, etc. can be collected. This will be discussed later, the possibilities of using AI for these purposes will be discussed.

5 Building a model and methodology

5.1 Software and public repository

This paper has been possible thanks to the following technologies.
Operative System: Pop_OS! (Ubuntu 20.04 LTS)
Python version: 3.8
Conda version: 4.8.4
Keras: 2.4.3
Matplotlib: 3.2.2
Numpy: 1.19.1
Pandas: 1.1.0
Sklearn: 0.23.2
Tensorflow: 2.2.0
Overleaft as a LaTeX editor.

Github user: dvento
Code available at: https://github.com/dvento/tfg

5.2 Datasets

The data has been obtained thanks to CrunchBase, due to an open database published on "Kaggle.com" website. The data includes information on nearly 500,000 statups as of 2014, most of which were founded in the beginning of the 1990s. Data such as the name of the company, a description, category, country, funding rounds (date, amount, participants, currency), milestones, offices, relationships between the founders, level of studies of the founders, etc. is provided.
This dataset has been chosen due to the large amount of data and the acceptable variety of features it offers.

5.3 Data preprocessing

One of the most important steps in data analysis is what is known as "cleaning", that is, select the time frame of the research, the selection of variables, the conversion of certain variables to various formats (such as dates), elimination of outliers, the conversion of qualitative variables, etc.
The initial sample of startups contained 462,651 values, always working on this table and adding information from other tables when necessary.
The first thing to do is to remove those Features that are not going to be used: 'entity_type', 'entity_id', 'parent_id', 'name', 'permalink', 'domain', 'homepage_url', 'twitter_username', 'logo_url ',' logo_width ',' logo_height ',' short_description ',' relationships', 'created_by', 'created_at', 'updated_at', 'region', 'first_investment_at', 'last_investment_at', 'investment_rounds',' invested_companies',

'first_milestone_at', 'last_milestone_at'.

All startups without a foundation date are eliminated, since it would make subsequent calculations ambiguous.

Startups founded before 1997 are also eliminated. This is done to cover two financial crises, the dotcom crisis, whose bubble burst in 2001, and the financial crisis of 2008.

The 'funding_rounds' table is accessed for retrieving the average of funds raised and the average number of participants per funding round are calculated, as well as the average number of months between funding rounds, and all of these variables are added to each of the startups in the main dataset.

Quantitative variables are also converted to date and numeric types, as appropriate.

After the cleaning, the sample size is 23,059 values.

We now go on to review the outliers of the most interesting variables (Funding rounds per startup, Average funds raised per startup, Average months between funding rounds, Average participants per round and Milestones per startup).

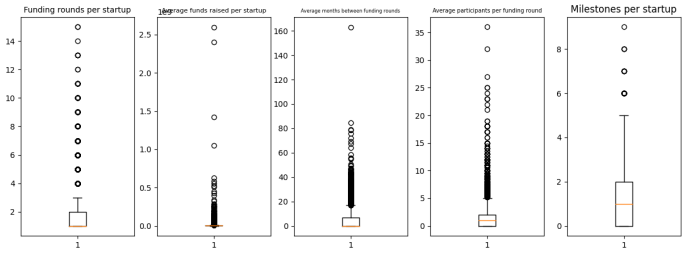

Figure 7: Main features outliers.

Once reviewed, several tests are done to determine how many to remove in order to obtain a more reliable sample. This is finally done by taking the quantiles 0.25 and 0.80, having a margin of [1, 6] above the maximum of the interquantile range as a bound according to the type of variable.

Then, after eliminating the outliers, the sample size is 20,448 values.

Finally, it is checked if there is a class imbalance in the sample and it is determined that the closed startups are the minority class, representing only 6.79% of the total sample (1,388 closed startups). This should be solved when the analysis is performed to reduce the bias of the algorithm, since with so few cases it will generalize that they are a minority and perform the classification poorly.

5.4 Features

The most interesting features that are named below.

Country code. In the sample there are 110 different countries, being the one with the most startups in the USA with 12,889 companies. The corresponding Economic Freedom Index for 2013 has been added to this variable as a new column.

Total funding (USD). With an average of $ 5,580,202, and a maximum of $ 102,148,270. It is one of the most interpretable and intuitive indicators to determine whether a startup can be profitable or not.

Average months between funding rounds. With an average of 9 months, and a maximum of 34 months.

Average funds raised per funding round (USD). With an average of $ 2,536,933.

Average number participants per funding round. An interesting fact, which could be seen as something positive more agents take part in the financing. An average of 2 participants.

5.5 What is considered to be a successful startup and what is not

Before carrying out any analysis, it is convenient to define what is a successful startup and what is not. For this small experiment, it has been considered that any company that is not closed and for which there is data (date of foundation, has received funds, etc.). The categories that meet these conditions are the following: 'operating', 'acquired' and 'ipo'.

To do this, a new column is created with a boolean on the state of the startups. If the company has closed, its value will be 1, while if the company is successful, its value will be 0.

The logic behind this reasoning is that the value of the shares of the companies that are acquired increases as a general rule many times, due to the premium offered on the shares by the company that acquires it. On the other hand, a company that goes public normally sees its valuation increased and tends to have a positive impact in the long term (Chand, 2020).

Finally, if a company is still operating today (after years of operation and financing) it is usually means that the project can be profitable.

5.6 Some insights

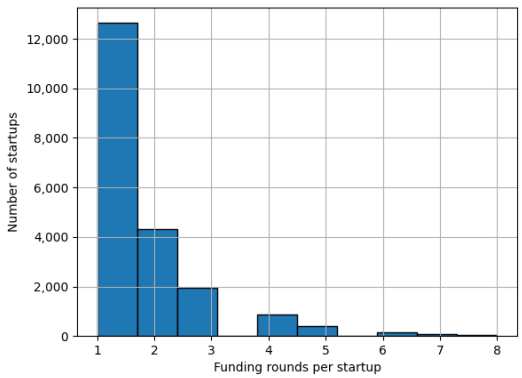

Figure 8: Funding rounds per startup.

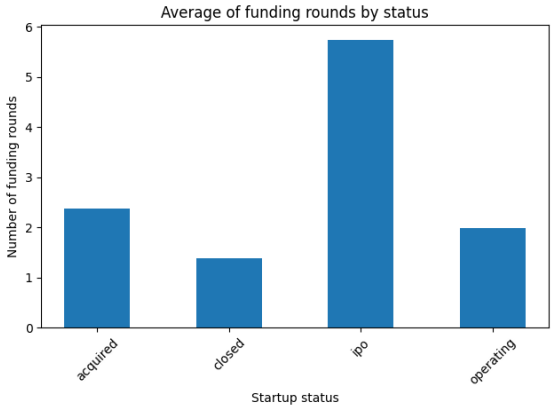

Figure 9: Average funding rounds by status.

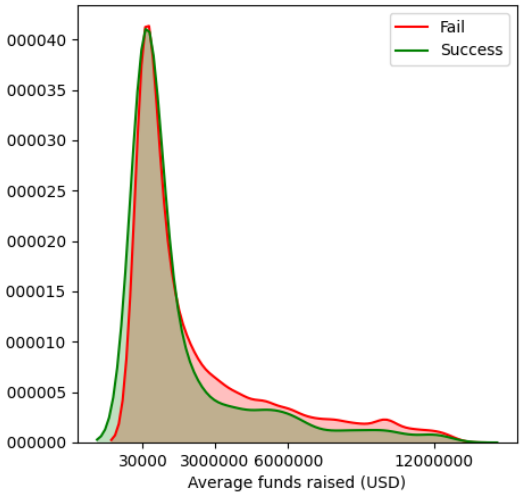

Figure 10: Average funds raised per startup (USD).

As we can see in Fig. 8, very few companies last beyond 1-3 financing rounds.

From Fig. 9, it can be concluded that those companies that have the objective of going public, make an excess effort and need a long trajectory for their exit.
The variable "Average funds raised per funding round" does not seem to be especially decisive in the success of the company, as we can see in Fig. 10.
In Fig. 11, "Average months between funding rounds" seems to have an inverse relationship with the success of startups.

5.7 Prediction using ML

Before running any model, the Class Imbalance problem must be fixed. For this, the ADASYN algorithm has been chosen for the task, which creates synthetic datapoints independent of the originals of the sample (He et al., 2008).

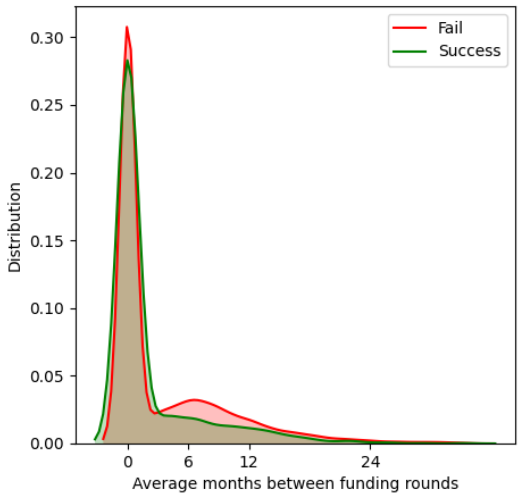

Figure 11: Average months between funding rounds.

With this method, a final sample of 34,227 startups is obtained.

In short, what ADASYN does is the following:

1. Calculate the ratio of minority / majority
2. Calculate how many datapoints needs to create in the minority class
3. Find the k-Nearest Neighbours for each minority example and find the r_i, which is the dominance of the majority class in each neighbourhood
4. Normalize the r_i so the sum of $r_i = 1$
5. Calculate the final amount of synthetic examples to generate per neighbourhood
6. Generate the necessary data for each neighbourhood

Finally, for the analysis, the following features must be converted into dummy variables: 'category_code' and 'tag_list'.

5.7.1 Logistic Regression

Logistic Regression is one of the most widely used ML methods for binary classification problems. Due to this, it has been selected to analyze startups and try to predict whether or not they will fail.

Before continuing, an explanation on precision, recall, f1score and support should be given.
TP: True Positive, a value that was correctly predicted and was positive (a successful startup, meaning a "0").

TN: True Negative, a value that was correctly predicted and was negative (a failed startup, meaning a "1").
FP: False Positive, a value that was incorrectly predicted (model predicted positive) and was negative (a failed startup, meaning a "0").
FN: False negative, a value that was incorrectly predicted (model predicted negative) and was positive (a successful startup, meaning a "1").

Accuracy. The ratio of correctly predicted values to the total of observations.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{8}$$

Precision. The ratio of true positives to the total predicted positives.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

Recall. (Sensitivity). The ratio of true positives to all observations in the actual class.

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

f1score. The weighted average of Precision and Recall.

$$Accuracy = 2 * \frac{Recall * Precision}{Recal + Precision} \tag{11}$$

Support. The number of occurrences of each class in the true values.

	precision	recall	f1-score	support
0	0.9429511793746571	0.8934511434511434	0.9175340272217775	1924.0
1	0.07657657657657657	0.14049586776859505	0.09912536443148687	121.0
accuracy	0.8488997555012225	0.8488997555012225	0.8488997555012225	0.8488997555012225
macro avg	0.5097638779756168	0.5169735056098692	0.5083296958266321	2045.0
weighted avg	0.8916889168130103	0.8488997555012225	0.8691098471740389	2045.0

Figure 12: Model 1: Average time between funding rounds, Number of funding rounds, Total funds raised (USD), Economic Freedom Index, Milestones

	precision	recall	f1-score	support
0	0.9413716814159292	0.8859968766267569	0.9128452668275676	1921.0
1	0.0759493670886076	0.14516129032258066	0.09972299168975071	124.0
accuracy	0.8410757946210269	0.8410757946210269	0.8410757946210269	0.8410757946210269
macro avg	0.5086605242522684	0.5155790834746687	0.506284129258592	2045.0
weighted avg	0.8888961963418031	0.8410757946210269	0.8635410310734897	2045.0

Figure 13: Model 2: Average time between funding rounds, Number of funding rounds, Total funds raised (USD), Economic Freedom Index, Milestones, Average participants per funding round.

As we can see, the accuracy of the model is quite acceptable, as well as the precision rate for successful startups; the most striking thing being the fact that it seems to create many false positives

	precision	recall	f1-score	support
0	0.9408866995073891	0.8939157566302652	0.9168	1923.0
1	0.06422018348623854	0.11475409836065574	0.08235294117647059	122.0
accuracy	0.8474327628361858	0.8474327628361858	0.8474327628361858	0.8474327628361858
macro avg	0.5025534414968138	0.5043349274954605	0.4995764705882353	2045.0
weighted avg	0.8885867899941469	0.8474327628361858	0.8670188062706745	2045.0

Figure 14: Model 3: Average time between funding rounds, Number of funding rounds, Total funds raised (USD), Economic Freedom Index, Milestones, Category Code

	precision	recall	f1-score	support
0	0.9384953322350357	0.8896408120770432	0.9134152859433456	1921.0
1	0.05357142857142857	0.0967741935483871	0.0689655172413793	124.0
accuracy	0.8415647921760391	0.8415647921760391	0.8415647921760391	0.8415647921760391
macro avg	0.4960333804032321 4	0.493207502812715 14	0.4911904015923625	2045.0
weighted avg	0.8848373547023769	0.8415647921760391	0.8622114857873339	2045.0

Figure 15: Model 4: Average time between funding rounds, Number of funding rounds, Total funds raised (USD), Economic Freedom Index, Milestones, Category Code, Average participants per funding round.

for startups that close. That is, of the majority of startups that, according to the model, only end up closing 5-7%, which is a negligible success rate.

This creates an opportunity cost problem, since an investor may lose the possibility of investing in a startup that ends up being a success.

This failure may be due to the use of the ADASYN oversampling algorithm and / or the similarity of startups that close with those that do not.

5.7.2 Random Tree

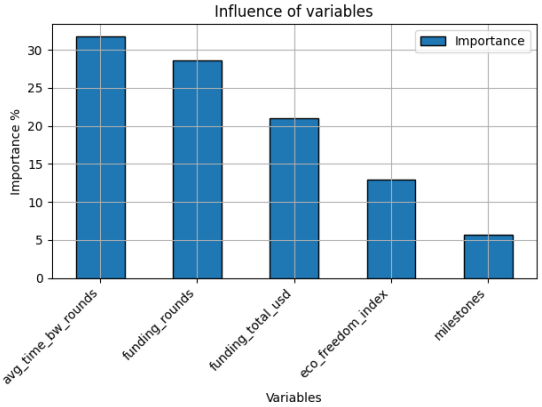

Figure 16: Model 1: Average time between funding rounds, Number of funding rounds, Total funds raised (USD), Economic Freedom Index, Milestones

We can see that the milestones of a startup do not play an essential role, much less its sector, since none of the tests carried out appear.

However, the Economic Freedom Index plays a key role in the development of startups. At the beginning in this work, it had been thought that this index would not have a great influence due to the fact that most startups are already established in countries with great financial freedom,

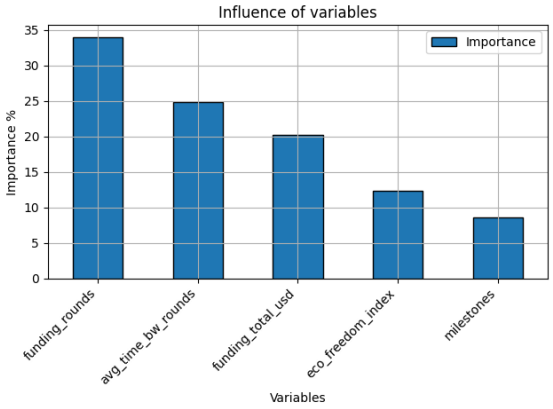

Figure 17: Model 2: Average time between funding rounds, Number of funding rounds, Total funds raised (USD), Economic Freedom Index, Milestones, Average participants per funding round.

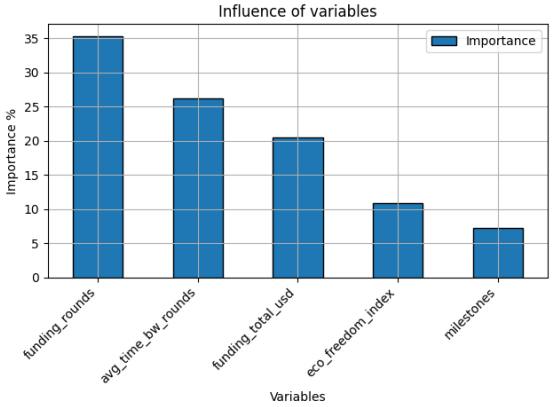

Figure 18: Model 3: Average time between funding rounds, Number of funding rounds, Total funds raised (USD), Economic Freedom Index, Milestones, Category Code

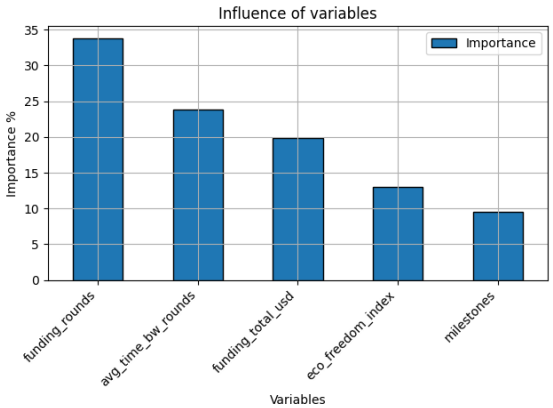

Figure 19: Model 4: Average time between funding rounds, Number of funding rounds, Total funds raised (USD), Economic Freedom Index, Milestones, Category Code, Average participants per funding round.

but it seems that there are notable differences between these countries.

6 Conclusions

The main objective of this work was to shed some light on the current state of startup valuation. The scope of large investors in the early stages of financing a business has proven to be well established, while the growing number of small investors does not find a niche in the venture investment spectrum. This is where there is a long way to go, especially if we compare the slow advance of these techniques to the growing demand for small investments and the proliferation of platforms for the small investors.

However, the literature reviewed does show that there are growing efforts (especially in recent years) to offer new approaches for valuation, not only of startups, but of companies in general. We see how innovative techniques such as text analysis, whether in digital newspapers or on social networks, play an essential role in risk assessment and in the valuation of companies. On the other hand, research seeking to improve traditional methods using less sophisticated AI techniques shows that this is already a reality.

Regarding the experiment carried out in this work, it shows that nowadays it is easy to implement sophisticated Machine Learning techniques (and others), and the get a decent quality on the results. However, a major downside for many investors that is worth noting is the data. It has been relatively difficult to get a dataset that offers quantity and quality at little or no cost, and it has been even more difficult to obtain a reliable sample, after applying various data cleaning and normalization techniques.

Finally, the AI sector has been in here for a very short period of time, a decade ago started to become widespread, including more complex techniques, such as Unsupervised Learning and the methods to determine how the User Generated Content can influence the performance of a company. Therefore, more in-depth analysis is needed on how to take advantage of these new technologies, so that investors and entrepreneurs benefit from them.

7 Bibliography

American Academy of Actuaries. (2017). "History of valuation". *ACSW Spring Meeting*, June.
Aguilar-Rivera, A. & Valenzuela-Rendón, M. (2019). "A new multi-period investment strategies method based on evolutionary algorithms". *The Natural Computing Applications Forum*, 2017. 10.1007/s00521-017-3121-6.

Belanche, D., Casaló, L. V. & Flavián, C. (2019). "Artificial Intelligence in FinTech: understanding robo-advisors adoption among customers". Emeral Insight. 10.1108.

Berkus, D. (2012, March, 25). *The Berkus Method: Valuing an Early Stage Investment.* Berkonomics. https://berkonomics.com/?p=1214
Brownlee, J. (2018, July, 27). *How to Configure the Number of Layers and Nodes in a Neural Network.* Machine Learning Mastery. https://machinelearningmastery.com/how-to-configure-the-number-of-layers-and-nodes-in-a-neural-network/

Chand, S. (2020). *10 Benefits of Issuing Initial Public Offering (IPO) for a Company.* YourArticleLibrary. https://www.yourarticlelibrary.com/business/10-benefits-of-issuing-initial-public-offering-ipo-for-a-company/24453
De Andrés, S. (2003). "Dos aplicaciones empíricas de las redes neuronales artificiales a la clasificación y la predicción financiera en el mercado español". *Revista Asturiana de Economía.* 28159348.

Enderle, R. (2020, July, 13). *The Secrets of How IBM Maintains AI Leadership.* Technewsworld. https://www.technewsworld.com/story/86752.html

Entis, L. (2013, November, 20). *Where Startup Funding Really Comes From (Infographic).* Entrepeneur. https://www.entrepreneur.com/article/230011#:~:text=According%20to%20data%20compiled%20by,funding%20from%20family%20and%20friends.

Fan, F., Xiong, J. & Wang, G. (2020). "On Interpretability of Artificial Neural Networks". arXiv:2001.02522.

Gross, Bertram, M. (1964). "The Managing Organizations: The Administrative Struggle", vol 2. pp. 856ff.

Gurney, K. (2004). "An introduction to neural networks". London: UCL Press.

Hadley, B. M. (2018). "Analyzing VC Influence on Startup Success: A people-centric network

theory approach". 10.1007/978-3-319-74295-3_1.

Haldane, A. (2010). "Patience and finance". *Speech by Mr Andrew Haldane, Executive Director, Financial Stability, Bank of England, at the Oxford China Business Forum*, September.

He, H., Bai, Y., Garcia, E. A. & Li, S. (2008). "ADASYN: Adaptative Synthetic Sampling Approach for Imbalanced Learning". *2008 International Joint Conference on Neural Networks (IJCNN 2008).* 10.1109/IJCNN.2008.4633969.

Hedayati, A., Hedayati, M. & Esfandyari, M. (2016). "Stock market index prediction using artificial neural network". Elsevier. *Journal of Economics, Finance and Administrative Science 21*, pp. 89–93.

Kavlakoglu, Eda. (2020, May, 27). *AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?*. IBM. https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks

Kharde, V. A. & Sonawane, S. S. (2016). "Sentiment Analysis of Twitter Data: A Survey of Techniques". *International Journal of Computer Applications (0975 – 8887)*, vol. 139, no. 11., pp. 5-15.
10.5120/ijca2016908625

Kopera, S., Wszendyby-Skulska, E., Cebulak, J. & Grabowski, S.(2018). "Interdisciplinarity in Tech Startups Development–Case Study of 'Unistartapp' Project". *Founder Magazine*, vol. 10, pp. 1–10.

Kostadinov, S. (2019, August, 8). *Understanding Backpropagation Algorithm.* Medium. https://towardsdatascience.com/understanding-backpropagation-algorithm-7bb3aa2f95fd

Krishna, A., Agrawal, A. & Choudhary, A. (2016). "Predicting the Outcome of Startups: Less Failure, More Success." *2016 IEEE 16th International Conference on Data Mining Workshops.* 10.1109/ICDMW.2016.103.

Kuppuswamy V., Bayus B.L. (2018) Crowdfunding Creative Ideas: The Dynamics of Project Backers. In: Cumming D., Hornuf L. (eds) The Economics of Crowdfunding. Palgrave Macmillan, Cham. 10.1007/978-3-319-66119-3$_8$

Lachiheb, O. & Salah, M. (2018). "A hierarchical Deep neural network design for stock returns prediction". Elsevier. *International Conference on Knowledge Based and Intelligent Information*

and Engineering Systems, September. 126:264–272.

Ling, C. X. & Sheng, V. S. (2010). "Class Imbalance Problem". 10.1007/978-0-387-30164-8_110

Li, W. (2015). "BUZZ: The Impact of Positive User Sentiment on Startup Company Valuations in the USA". Under the direction of Timothy Bresnahan, Honors Thesis, Stanford University.

Li, X. (2013). *Financial Management: Investment*, ch. 3. [Class slides]. Moodle URV.

Machová, V. & Vochozka, M. (2019). "Analysis of business companies based on artificial neural networks".
https://doi.org/10.1051/shsconf/20196101013.

Mavrovouniotis, M., L. & Chang, S. (1991). "Hierarchical neural networks". 10.1016/0098-1354(92)80053-C

Nazemi, A. & Tahmasbi, N. (2013). "A computational intelligence method for solving a class of portfolio optimization problems". Berlin: Springer. 10.1007/s00500-013-1186-4.

Neumann, J. (2019, November, 26). *Startups and Uncertainty.* Reactionwheel. http://reactionwheel.net/2019/11/startups-and-uncertainty.html#:~:text=Uncertainty%20can%20be%20seen%20everywhere,product%2C%20and%20in%20the%20market.&text=Startups%20that%20aim%20to%20create,Uncertainty%20becomes%20their%20moat.

Payne, B. (2011, November, 15). *Valuations 101: The Risk Factor Summation Method.* Gust blog. https://blog.gust.com/valuations-101-the-risk-factor-summation-method/

Richards, R. (2019, July, 9). *How to Value a Startup Company With No Revenue.* MassChallenge. https://masschallenge.org/article/how-to-value-a-startup-company-with-no-revenue#:~:text=Method%207%3A%20Cost%2Dto%2DDuplicate&text=In%20this%20method%2C%20you%20assess,looking%20for%20pre%2Drevenue%20investors.

Robehmed, N. (2013, December, 16). *What Is A Startup?.* Forbes. https://www.forbes.com/sites/natalierobehmed/2013/12/16/what-is-a-startup/#3bb533fc4044

Rosenblatt, F. (1957). "The Perceptron: a perceiving and recognizing automaton". Report 85-460-1. Cornell Aeronautical Laboratory.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). "Learning representations by back-

propagating errors". *Nature*, vol. 323, pp. 533-536.

Ryll, L. & Seidens, S. (2019). "Evaluating the Performance of Machine Learning Algorithms in Financial Market Forecasting: A Comprehensive Survey". arXiv: 1906.07786.

Saura, J. R., Palos-Sánchez, P. & Grilo, A. (2019). "Detecting Indicators for Startup Business Success: Sentiment Analysis Using Text Data Mining". *Sustainability*, February. vol. 11, iss. 3. 10.3390/su11030917

Smith, T. (2019, July, 14). *Multiples Approach*. Investopedia. https://www.investopedia.com/terms/m/multiplesapproach.asp

Souza, D., Ghezzi, A., Barbosa, R., Nogueira, M. & Schwengber, C. (2020). "Lean Startup, Agile Methodologies and Customer Development for business model innovation: A systematic review and research agenda". *International Journal of Entrepreneurial Behaviour & Research*, vol. 26, iss. 4, pp. 595-628. 10.1108/IJEBR-07-2019-0425.

The Business Professor. (2015, March, 10). *Venture Capital Method*. The Business Professor. https://thebusinessprofessor.com/lesson/venture-capital-method/

The Heritage Foundation. (2020). *Country rankings*. Heritage. https://www.heritage.org/index/ranking

Van Gelderen, M., Bosma, N. & Thurik, R. (2006). "Success and Risk Factors in the Pre-Startup Phase". *Small Business Economics*, May. 10.1007/s11187-004-6837-5.

Ünal, C. (2019). "Searching for a Unicorn: A Machine Learning Approach Towards Startup Success Prediction". Berlin. 10.18452/20347.

Von Spreckelsen, C., Von Mettenheim, H. & Breitner, M. H. (2014). "Real-Time Pricing and Hedging of Options on Currency Futures with Artificial Neural Networks". John Wiley & Sons. *Journal of Forecasting, J. Forecast*, 419–432. 10.1002/for.2311

White, H. (1988). "Economic prediction using neural networks: the case of IBM daily stock returns". *IEEE 1988 International Conference on Neural Networks*, July, pp. 451.10.1109 / ICNN.1988.23959.

Wikipedia. (2020, August, 2). *Artificial Neuron*. Wikipedia. https://en.wikipedia.org/wiki/Artificial_neuron

Wikipedia. (2020, August, 11). *Random Forest*. Wikipedia. https://en.wikipedia.org/wiki/Random_forest