U.PORTO

FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

DO NOT TRUST YOUR GUT: ENHANCING VENTURE CAPITALISTS' INVESTMENTS USING DATA-DRIVEN DECISIONS

António Meireles Vieira

Internship report

Master in Economics

Supervised by

Professor Carlos Francisco Ferreira Alves
Professor Pedro José Ramos Moreira de Campos

Internship

LTPlabs | Eduardo Francisco Santos Silva Ribeiro

2023/2024

Acknowledgments

The conclusion of my thesis and, consequently, my Master's in Economics marks the conclusion of a significant phase in my life. This period, which I have enjoyed to the fullest, has helped me grow and develop. As one door closes and another opens, I eagerly anticipate the new challenges, adventures, and opportunities that lie ahead. As I reach the end of this journey, I would like to express my gratitude to those who have been indispensable and to whom I owe a great deal of thanks.

First and foremost, I extend my deepest gratitude to Professor Carlos Alves and Professor Pedro Campos for their availability, support, guidance, and understanding throughout this journey. Their vast knowledge and experience were invaluable to the success of this thesis.

I am also deeply thankful to LTPlabs for opening its doors and allowing me to work in an environment I never imagined I would find. These months have been a period of constant and invaluable learning. It is reassuring to see that academic research and industry can work hand in hand. Thank you to all the LTPeers for the warm welcome and all the assistance provided. A special word of appreciation to Eduardo Ribeiro for his guidance, patience, and willingness to review my work, and to Gonçalo Rodrigues for his constant support and essential help in completing this project.

This entire journey was made significantly easier thanks to the unwavering support of my family! Huge thanks for everything you have done and continue to do for me. But most importantly, thank you, Mom and Dad. Thank you for giving everything you could to provide me with the best future, for all the education and lessons you imparted, for always being proud of me, and for being an endless well of love. I hope to one day repay everything you have given me!

Of course, I cannot forget my friends. Thank you for always reminding me of the reasons why we are here. Cheers to all the shared 'tainadas' and 'patuscadas' and those yet to come. A wise person once said, "Whatever you do in this life, it is not legendary unless your friends are there to see it."

Lastly, my deepest thanks to my best friend, my partner in adventures, travel, laughter, and everyday moments. Thank you for adding more fun to my life, for bringing out the best version of me, and for dreaming with me. Thank you, Bia!

Abstract

This internship report investigates the integration of a data-driven approach into the venture capital (VC) groups' screening process. The focus of the study is on enhancing the decision-making process through a predictive model, which addresses the significant challenges faced by VCs, including the lack of automation, excess manual effort, unstructured data, and the risk of missing potential investment opportunities due to biased decisions given their reliance on heuristics. The methodology involves developing a Random Forest model capable of screening startups by evaluating more than the conventional metrics, such as keywords in descriptions. The model was trained and tested using a comprehensive dataset from Crunchbase, which provided a wide range of startup features. Results demonstrate that the model successfully captures over 80% of startups fitting the VC's desired profile, with a potential accuracy of up to 92% under optimal conditions. This signifies a substantial improvement over previous models, confirming the efficacy of incorporating advanced analytics in the VC industry. Furthermore, the model is also capable of performing well on unseen data by accurately identifying startups with the correct profile, and simulating the model's application in a real-life context. The study concludes that predictive modeling can significantly streamline the screening process, reduce biases in decision-making, and enhance the overall efficiency of capital and human allocation in venture investments.

JEL codes: C60, G11, G24, G41, L26, M13

Keywords: Behavioral Economics, Deal-Sourcing, Predictive Analytics, Startup, Venture Capital

Resumo

Este relatório de estágio investiga a integração de uma abordagem baseada em dados no processo de seleção de grupos de capital de risco (VC). O foco do estudo é aprimorar o processo de tomada de decisão por meio de um modelo preditivo, que aborda os desafios enfrentados pelos VCs, incluindo a falta de automação, o excesso de tarefas feitas manualmente, dados não estruturados e o risco de perder potenciais oportunidades de investimento devido a decisões enviesadas, dada a grande dependência em heurísticas por parte dos VCs. A metodologia utilizada envolve o desenvolvimento de um modelo de Random Forest capaz de avaliar startups através de métricas além das convencionais, como identificação de palavras-chave em descrições. O modelo foi treinado e testado usando um conjunto de dados abrangente do Crunchbase, que forneceu uma ampla gama de características de startups. Os resultados demonstram que o modelo captura com sucesso mais de 80% das startups que se ajustam ao perfil desejado pelo VC, com uma potencial precisão de até 92% em condições ótimas. Este representa uma melhoria substancial em relação a modelos anteriormente aplicados, confirmando a eficácia da incorporação de análises avançadas na indústria de VC. Além disso, o modelo apresenta bom desempenho em dados ainda não vistos ao identificar, eficazmente, startups com o perfil correto, simulando a aplicação do modelo num contexto de vida real. O estudo conclui que a modelação preditiva pode simplificar significativamente o processo de seleção, reduzir os vieses na tomada de decisão e aumentar a eficiência geral da alocação de capital e capital humano na indústria de capital de risco.

Códigos JEL: C60, G11, G24, G41, L26, M13

Palavras-chave: Análise Preditiva, Capital de Risco, Economia Comportamental, Prospeção de negócios, Startup

Table of Contents

List of Tables

List of Figures

Acronyms

AI – Artificial Intelligence

ANN – Artificial Neural Network

DT – Decision Trees

EL – Ensemble Learning

ERT – Extremely Randomized Trees

GTB – Gradient Tree Boosting

IV – Instrumental Variables

KNN – K-Nearest Neighbors

LR – Linear Regression Model

MLP – Multilayer Perceptron

NB – Naïve Bayes

RF – Random Forest

SVM – Support Vector Machines

VC – Venture Capital

XGB – XGBoost

1. Introduction

In recent decades the digital revolution has spread to all sectors of economic and social life (Ferrati & Muffatto, 2021). The rise of supercomputing power and big data has empowered and rapidly expanded artificial intelligence (AI) in the last few years (Duan et al., 2019). The dramatic changes in data storage and processing technologies provoked new opportunities to collect and leverage data, leading many managers to change decisions, relying less on intuition and more on data (Brynjolfsson & McElheran, 2016). In 2006, the British mathematician Clive Humby argued that data is the new oil. This statement not only manifests the data's importance as a fuel for innovation but also indicates that this resource is useless in its raw state as it needs to be refined and processed. Therefore, the power of differentiation lies in our capacity to transform raw data into actionable intelligence. And that is how LTPlabs distinguishes itself from the contenders in a field as competitive as consultancy.

LTPlabs is a boutique analytical-driven management consultancy firm that provides highly customized solutions to its clients, pioneering the implementation of analytical and technical models to overcome business problems. LTP is the acronym for Long-Term Partnership, which aligns with the company's mission and values. The company tries to help its clients achieve significant and sustainable performance improvements by combining analytics with business expertise. The firm pursues to meet the needs and preferences of each partner with five distinct delivery modes: Consultancy, Training & Support, Analytics-as-a-service, Assessment, and Leveraging Analytics. Founded in 2015 and showing a high-growth trajectory, LTPlabs already has over 300 successful projects across 20 countries.

Together with LTPlabs, a project was developed to improve screening early-stage investment opportunities using advanced analytical tools such as Machine Learning (ML) models. This plan originated from an internal initiative at the company, following a collaborative project with a Venture Capital (VC) group in early 2023. During this partnership, a model was developed and validated to tackle some problems that the group faced, such as the lack of formality in the Retail Technology investment vertical, the overwhelming volume of startups, the fear of missing out on key market opportunities, the lack of prioritization of analysis and the need for a more structured and consistent process among the VC group's team members. The initial project was able to solve some of the problems by automating processes and reducing manual effort, centralizing information via a database connection, and developing

the aforementioned model trained to screen potential deals by checking hard filters such as the year of foundation, number of staff members and rounds of investments that they already used and by detecting previously selected keywords on the company's descriptions that, based on historical data, are more prone to be found in startups that correspond to the VC's desired profile. This allowed for better prioritization since the model attributed a fit score to the startups and ranked them accordingly. Despite offering some advantages, this solution presented some limitations. The VC group wanted to ameliorate the fit score classification and develop a model with boosted predictive capabilities that used more features beyond the keywords approach. Not only that but VCs wanted the model to be continuously fed with new startups and updated data and a tool for improved visualization and exploitation of data by multiple team members working simultaneously.

Thus, given the challenges brought by the first solution, the primary aim of this internal project was to refine and enhance the existing model based on the insights and outcomes of the previous collaboration. The objective was to develop a robust, scalable, and generalizable model that serves as a comprehensive tool potentially marketable to various VC groups. Besides improving the model and refining the fit score by adding new variables, another goal of this project was to upgrade the user experience and data visualization by creating an integrated platform. It is also important to mention that the model does not exclude any company or make any decision by itself. Also, its results are not blindly followed without careful analysis afterwards. What the model does is that it immediately highlights those that would probably be the companies with the profile and characteristics closest to the desired ones, ensuring that the VC group focuses straight away on these, preventing them from losing a good deal while analyzing others of lesser interest.

A heavy reliance on heuristic judgments characterizes VCs' investment decision-making subject to various cognitive biases. Thus, introducing a predictive model presents a substantial opportunity for enhancement. The incorporation of a data-driven approach can significantly refine the screening process, helping VCs prioritize their opportunities more effectively and reducing the susceptibility to biases like similarity and local biases. That prioritization can streamline the evaluation process, saving valuable time for VCs, but it might also ensure a broader and more diverse consideration of potential investments. Additionally, it might ensure that decisions are made coherently and consistently and do not vary depending on the person making them. This way, employing such a model can allow

VCs to outperform their peers who continue to rely solely on traditional and less structured approaches.

The model's capability to offer a balanced analysis by mitigating the influence of subjective factors can lead to more informed and objective decisions, enhancing the performance of VC firms in a highly competitive market. The improvement of the model is, therefore, crucial for fostering confidence among VCs in its utility as a decision-making tool. Confidence in the model's accuracy and relevance is essential, as VCs must trust the tool to supplement their intuitive judgments with empirical data effectively. This report might contribute significantly to the enhancement of the model by conducting in-depth research into the specific features and criteria VCs prioritize when assessing investment opportunities. The report systematically analyzes these preferences and aims to integrate crucial insights into the model, aligning it more closely with VC's practical needs and considerations. The resulting improvements might not only improve the model's reliability and adoption but also ensure that it serves as a robust analytical framework that VCs can rely on to make more precise and successful investment decisions and that they can allocate more efficiently, analyzing more deeply the opportunities with higher potential and covering a higher stake of the market, therefore setting a new standard in the industry operations.

Therefore, it was decided to focus the study on exploring how big data and advanced analytics can empower VC firms by helping them prioritize their investment opportunities, mitigating the risk of biases in decision-making, and improving their investment decisions. In order to reach the goal of this report, the research was based on the following research question: How can the integration of a predictive model on VC's screening process improve their performance?

This report aimed to improve the previous model results, as mentioned before, so that was also kept in mind during the research. The improved model will then be compared with the previous one to check whether the objective of the study was fulfilled.

For this research, the data was collected from previous databases made available by the VC group that collaborated with LTPlabs earlier and from Crunchbase. This company provides information about a wide range of companies, most precisely startups. Quantitative methods were used to improve the model, focusing on the Random Forest Model, which will be described in more detail later in this report but also, with the complement of more theoretical

methods for the inspection of the state-of-art of the existent literature to better understand the features that could be explored and be part of the model.

This topic is extremely motivating for me due to my passion for data-driven analysis and its advantages in decision-making. Not only that, but the possibility of including financial economics, entrepreneurial finance, and behavioral economics aspects in this work is immensely pleasurable, as those are some of my main areas of interest within the Economics spectrum.

This topic can be beneficial for the host Organization since the findings of the report can be helpful in the development of this internal project by allowing different approaches to be explored, various aspects through which the model applied can be optimized, and distinct ways on how LTPlabs can use analytical-driven methods to improve the startup screening process of this or future clients. This report can serve as a step forward in the creation of a tool that could be commercialized and generalized for different VC groups. Furthermore, this study could contribute to this domain because while big data and AI have received increasing attention in various fields, little research has been done in entrepreneurship (Ferrati & Muffatto, 2021).

This report unfolds as follows: In Chapter 2, a brief literature review is made on Venture Capital, the Decision-making process and related biases, ML and predictive models applicability in VC's industry context, and Natural Language Processing techniques. Chapter 3 presents the project description and the problems faced. Chapter 4 exposes the methodology adopted in this report and the description of the data used and how and from where it was collected. Chapter 5 presents the results obtained from the implementation of the proposed methodology. Finally, Chapter 6 summarizes the report's main findings and conclusions, acknowledges the study's limitations, and provides suggestions for further work.

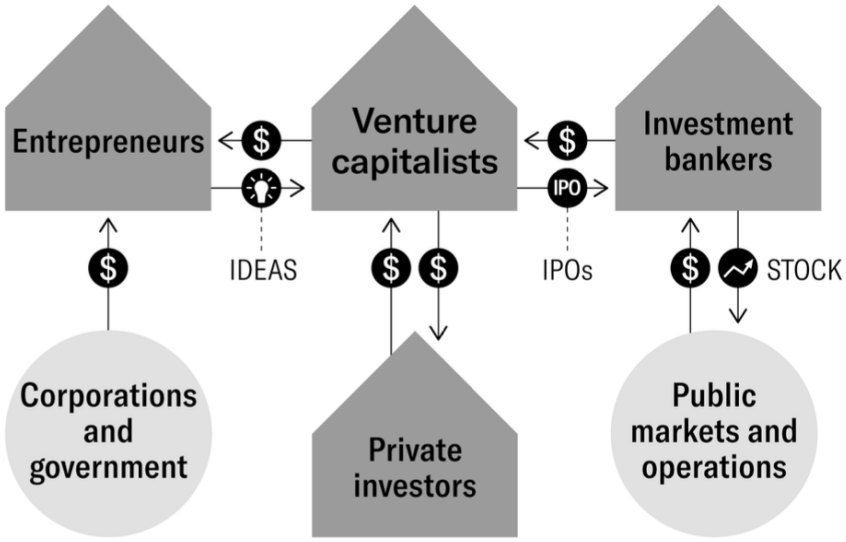
2. Literature review

2.1. Venture Capital Industry

Venture Capital (VC) is a form of private equity directed towards startups and early-stage companies. VC groups can be defined as those organizations whose mission is to finance the founding or early growth of new companies that do not yet have access to the public securities market or institutional lenders (Zacharakis & Meyer, 2000). They occupy a crucial role in the economy due to their importance in connecting entrepreneurs who have good ideas but no money with investors who have money but no ideas (Gompers et al., 2021).

Zider (1998) identifies four leading players in the VC industry: the entrepreneurs who need funding for their startups, the investors who seek high returns, the investment bankers who need companies to sell, and the venture capitalists who make money for themselves by creating a market for the other three. The players and their dynamics can be seen in Figure 2.1.

Figure 2.1 - Four main players from the VC industry by Zider (1998



Several steps in a VC's investment process can be separated between pre-investment and post-investment phases. First, there are some subtle differences between authors. Tyebjee & Bruno (1984) proposed a five-stage VC process model that included identifying potential firms, deal screening, deal evaluation and assessment of a business plan, and deal structuring, which ended with post-investment activities. Zacharakis and Meyer (2000) simplify it by

considering a three-stage investment process that starts with the VCs screening hundreds of proposals. Those who survive are then subjected to extensive due diligence, and finally, the VC and entrepreneurs negotiate the terms of the investment. Regarding the post-investment phase, Sahlman (1990) mentions eight stages, starting with the investment in the startup, followed by different periods of the company's development, culminating with the cash-out or exit. Each stage is described more completely in Table A.1.

Focusing on the pre-investment stage, screening is considered the first stage for VCs. This is characterized by gathering and identifying the best investment opportunities. VCs normally face a quandary in this phase as they question how they can screen venture proposals without prematurely eliminating high-potential investments (Zacharakis & Meyer, 2000). Much literature is dedicated to understanding how VCs make decisions and what they look for in a startup. Arguably, they rely on a mix of assessment criteria and their own experience due to the evaluated companies' lack of historical financial data. Gompers et al. (2020) state that VCs find management teams far more important than other quantitative measures for deal selection. Industry is also emphasized by Zider (1998), who argues that VCs invest in industries that are more competitively forgiving than the market. Contrarily, there is little evidence that VCs use the techniques taught at business schools, such as the discounted cash-flow model or the net present value (Gompers et al., 2021).

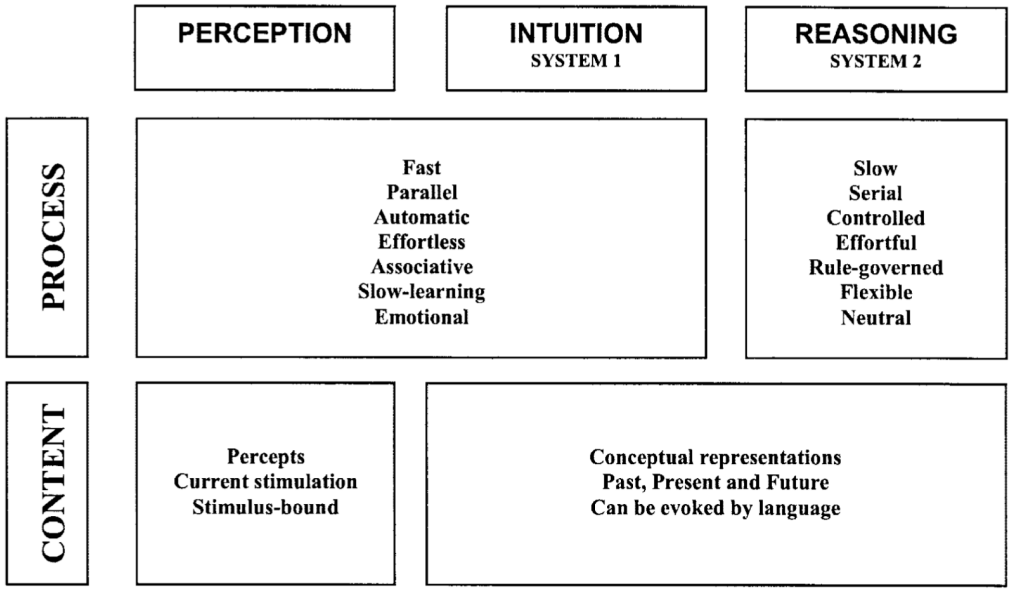
Besides that, it is also known that the selection methods of VC firms are primarily based on human judgment (Ferrati & Muffatto, 2021) and heuristics to reduce complexity and simplify decisions. Nevertheless, heuristics have been found to make decision-makers deviate from optimal decisions (Åstebro & Elhedhli, 2006). Consequently, even though VCs are experts in the new venture funding realm, they have much room for improvement (Zacharakis & Meyer, 1998). In reality, VCs lack a strong understanding of how they make decisions and are not good at self-introspecting their decision process (Monika & Sharma, 2015). According to Zacharakis and Meyer (1998), this difficulty might be due to the noise caused by information overload, as the lack of systematic understanding impedes learning and prevents them from accurately adjusting their evaluation process since they do not truly understand it.

2.2. Decision-making and Biases

The Nobel laureate in Economics, Kahneman (2003), introduced System 1 and System 2 concepts of thinking to explain cognitive processes. System 1 is fast, automatic, and intuitive,

while System 2 is slower, deliberate, and analytical. These can be seen in more detail in Figure 2.2.

Figure 2.2 - Characteristics of System 1 and 2 Thinking by Daniel Kahneman (2003)



The author argues that these systems lead to mental shortcuts, known as heuristics, often resulting in cognitive biases. These biases influence decision-making, leading individuals to rely on instinctive responses rather than engaging in deeper, more rational analysis.

Monika and Sharma (2015) refer to heuristics as methods or techniques to solve a problem despite not being certain about the optimal solution. These mental models help make decisions utilizing information in a process that can lead to biases. For these authors, biases, in the context of the VC industry, are described as significant factors that affect the VCs' mental model to evaluate ventures based on cognitive factors.

Franke et al. (2006) point out the possibility of similarity bias based on an experiment with 51 VCs respondents. They found that VCs tend to favor teams similar to themselves in the type of training and professional experience. The higher the similarity between the profile of a VC and the profile of a startup team, the more favorable the evaluation by the VC will be.

Some authors also warn about the strong evidence of local bias in VCs' decision-making, referring to the fact that VC firms invest predominantly in new ventures that are located in their home cities or states (Cumming & Dai, 2010). This is particularly odd since Chen et al. (2010) documented the geographic concentration of VC firms in three US areas,

demonstrated that these firms tend to invest in local ventures, and showed that when these VC firms outperform, that outperformance was not driven by those local investments.

Finally, much of the literature focuses on the overconfidence bias. According to Zacharakis and Shepherd (2001), overconfidence is the tendency to overestimate the likely occurrence of a set of events, with overconfident individuals making probability judgments that are more extreme than they should be given the evidence and knowledge possessed. In the case of the VC industry, overconfident VCs may overestimate the likelihood that a funded company will succeed. Koellinger et al. (2007) explain that entrepreneurs likely overestimate their control over events if entrepreneurial decisions are primarily based on perceptions, and the cognitive mechanism leads to overconfidence.

2.3. Machine Learning and Predictive Models

The screening process and evaluation of startups largely depend on the investors' personal experiences (Zhong et al., 2018), enhancing the emergence of biases, as shown previously. Since the tools that investors currently have available are not robust enough to reduce risk and help them manage uncertainty better (Arroyo et al., 2019), the literature suggests the use of analytical tools and models based on big data and AI to bridge this gap, helping them to be more effective in selecting companies with higher success probability (Corea et al., 2021).

The term "big data", which has become extremely popular in recent decades, consists of datasets that cannot be effectively processed using traditional processing applications, mainly because of its three main characteristics, known as the 3Vs: i) Volume (enormous quantities of data); ii) Velocity (the rapid rate at which data is collected, processed, and made available); iii) Variety (structured, semi-structured, and unstructured data) (Laney, 2001).

Due to big data technologies and the rise of supercomputing power, AI has been empowered and has become more popular in recent years (Duan et al., 2019). Simon (1995) defined AI as "systems that exhibited intelligence, either as pure explorations into the nature of intelligence, explorations of the theory of human intelligence, or explorations of the systems that could perform practical tasks requiring intelligence". More recently, Huang et al. (2019) included "mimicry of human intelligence" to that definition. Others mention "machines that perform tasks humans would perform" (Bolander, 2019) or "artifacts able to carry out tasks in the real world without human intervention" (Piscopo & Birattari, 2008).

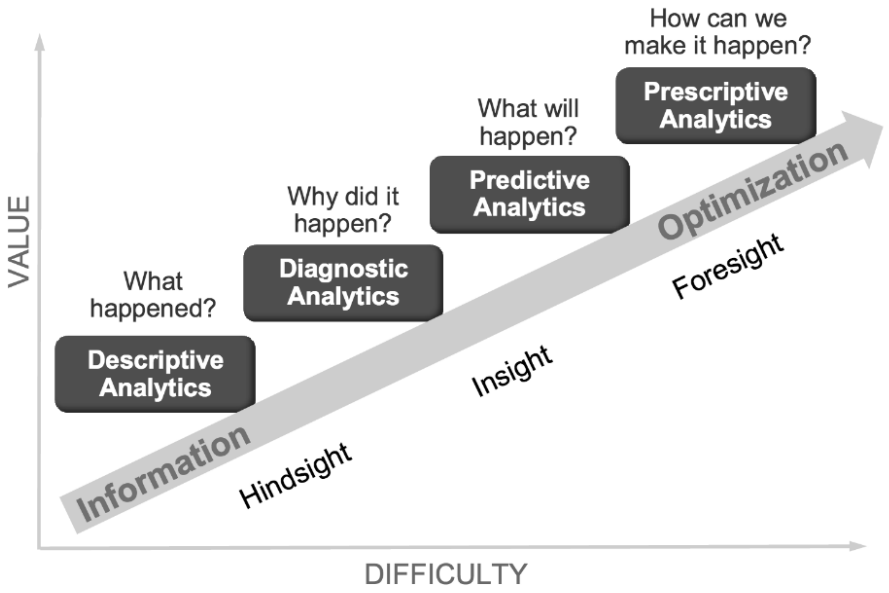
Machine Learning belongs to the broader scientific field of AI. It can be defined as the field of study giving the ability to automatically learn and improve from experience without being programmed. In the VC industry, ML and predictive models can serve numerous purposes, such as classifying companies by industry, generating investment recommendations, predicting funding events, and the exit of a company (Ferrati & Muffatto, 2021). In **Erro! A origem da referência não foi encontrada.**, a set of existing papers has been gathered. They used different types of models to predict the success of firms or merely explore what factors matter most during the screening process, allowing a better understanding of what had already been done concerning startup success prediction models.

These examples demonstrate that predictions made using machine learning models can be better than those made by humans, especially when calculating the complex interactions between different variables (Ferrati & Muffatto, 2021). However, machines also fail, especially in two kinds of situations: processing and interpreting soft information and making predictions in unknowable risk situations of extreme uncertainty (Dellermann et al., 2017). Therefore, most of the literature argues that ML cannot fully substitute humans as they remain fundamental in data collection and investment decisions (Ferrati & Muffatto, 2021). The models should instead be used as complementary, helping human VCs overcome bounded rationality constraints (Dellermann et al., 2017).

Based on this selection of papers, it is possible to verify that there are countless possibilities of variables that can be considered when analyzing startups.

Regarding the techniques, a wide variety of predictive models can be used to classify startups, as shown in **Erro! A origem da referência não foi encontrada.**. Predictive models fall within the category of Predictive Analysis, which aims to predict future events based on historical data and past behaviors. In Figure 2.3 it is possible to observe the different types of Analytics and their degree of value and difficulty.

Figure 2.3 - Types of Analytics by Gartner (2012)



Therefore, predictive models are sophisticated tools designed to forecast outcomes based on historical data. Agrawal et al. (2018) describe it as processes that take available information and use it to generate unavailable details. These models operate by recognizing patterns and relationships within the data, facilitating predictions about future events (Hastie et al., 2009). Jordan and Mitchell (2015) also elucidate that these models encompass a variety of ML techniques, each tailored to specific types of data and predictive requirements, which are central to their application across diverse domains. Besides that, James et al. (2013) highlight the practical implementation of these models, emphasizing the integration of statistical insights with computational techniques crucial for handling large datasets. Thus, predictive models provide a robust framework for making informed predictions crucial for decision-making in various contexts.

According to the presented literature in **Erro! A origem da referência não foi encontrada.**, some of the most utilized techniques in the VC context are Logistic Regression, Decision Trees, and Random Forest Models, which were applied on 32%, 26%, and 37% of the papers analyzed, respectively. They are used as predictive models for multiple reasons, and the focus will remain on these three.

Table 2.1 – Papers that used predictive models to classify startups and their characteristics

Author (year)	Data Sources	Techniques	Brief Description
Arroyo et al., 2019	Crunchbase	• ERT • DT • GTB • RF • SVM	Different trained techniques are used to predict the success of startups. Features included industry classification, funding rounds and financial performance metrics. The RF model outperformed the other models.
Corea et al., 2021	Crunchbase, LinkedIn	• GTB	Use GTB algorithm and features based on innovation potential, market viability and financial health metrics.
Dellermann et al., 2017	Previous studies	• ANN • Logit • NB • RF • SVM	Hybrid models combining human expertise and algorithms. It merges insights from VCs with predictive analytics to forecast startup success. The hybrid approach achieved superior results compared to methods relying solely on human judgement or purely algorithmic approaches.
Ewens & Marx, 2018	Crunchbase, LinkedIn, CapitalIQ, Venture Source, Companies' websites	• IV • LR	Econometric models to analyze the impact of founder turnover on startup performance. Results show that startups that replaced underperforming founders with experienced CEOs saw improved performance and higher chances of survival.
Ghassemi et al., 2020	2015 MIT Competition	• ANN • DT • EL • Logit • KNN • SVM	Various ML models predict the fate of startups based on team composition and venture properties. Use data focused on team diversity and optimism in business abstracts. Found that diverse teams are more likely to survive.

Kim et al., 2023	Crunchbase	• DT • GTB • NB • RF • SVM	Applies various techniques using data on startup characteristics, market conditions, and historical performance metrics. The study found that RF provided the best predictive accuracy.
P. Gompers & Lerner, 1998	VentureOne	• Logit • LR	Different techniques are employed to compare the success of corporate venture investments with those backed by independent venture organizations. Results are consistent with the existence of complementarities that allow corporations to select and add value to portfolio firms, effectively.
P. Gompers et al., 2010	Venture Source	• LR • Probit	Examine the persistence of performance across multiple ventures of serial entrepreneurs. Findings suggest that past success is a reliable predictor of future venture success, supporting the hypothesis that skill and experience are transferable across entrepreneurial efforts.
Retterath, 2020	Crunchbase, Pitchbook, LinkedIn, Surveys	• DT • Logit • LR • NB • RF • XGB	Comparison of human and ML models decisions. The study demonstrates that while machines can match or even exceed human performance, experienced VCs possess unique insights that are difficult to encapsulate in predictive models.
Ross et al., 2021	Crunchbase, USPTO	• MLP • RF • XGB	Creation of a proprietary ML model designed to predict startup success and exit opportunities. It includes market trends, financial health indicators and team experience as features. Model has the capacity to predict exit outcomes accurately.

Logistic regression models were initially developed by (Cox, 1958) and serve as a fundamental statistical tool for modeling binary outcomes. This technique estimates the probabilities of events by employing a logistic function (Hosmer & Lemeshow, 2000) and ensuring that the predicted probabilities are constrained between zero and one. These are valued for their simplicity, interpretability, and efficiency, making them a standard choice in many fields of study. Also known as Logit Models, they handle the interaction between variables well, providing insights into how predictor variables influence the outcome. Nevertheless, they assume a linear relationship between the log odds of the outcome and the predictor variables, which might not be appropriate (Wooldridge, 2002). Additionally, Kleinbaum and Klein (2010) state that Logit Models can be sensitive to multicollinearity, requiring careful preprocessing of the input data.

Decision tree models were developed by Quinlan (1986) and offer a graphical decision support tool by recursively portioning the data into subsets based on the value of input features, creating a tree-like model of decisions and their possible consequences. This technique involves the creation of internal nodes, branches, and leaf nodes that represent the tests on an attribute, the outcome of the test, and the class label or continuous value, respectively, making it highly intuitive and easy to interpret (Quinlan, 1986). Breiman et al. (1984) further elaborated this technique by introducing methodologies to handle both categorical and continuous data, expanding the applicability of Decision Trees in predictive modeling. Notwithstanding, they are susceptible to overfitting, especially when the tree grows too deep and becomes too specific to the training data (Rokach & Maimon, 2005).

However, Quinlan (1993) believes that pruning techniques and ensemble methods like Random Forests can mitigate the issue. Thus, this technique introduced by Breiman (2001) can be seen as a powerful extension of Decision tree models since it combines multiple decision trees to enhance predictive accuracy and control overfitting using an ensemble learning approach. This method operates by constructing many decision trees at training time and outputting the class that is the node of the classes or mean prediction of the individual trees (Breiman, 2001). Each tree in the forest is created using a different bootstrap sample from the training data, and at each node, a random subset of features is considered for splitting. The randomness from the process ensures that the individual trees are less correlated, resulting in more robust models. Random Forest Models have gained prominence for their superior performance in different contexts, particularly when dealing with large

datasets and high dimensionality, providing excellent performance even if data contains irrelevant features (Louppe, 2014). Furthermore, they also offer feature importance metrics that help in understanding the structure of the data (Biau & Scornet, 2016).

2.4. Natural Language Processing

Natural Language Processing (NLP) is a branch of AI that focuses on the interactions between humans and computers through natural language. Therefore, NLP's main objective is to enable computers to understand, interpret, and respond to human language. The applied techniques are crucial for extracting insights from unstructured data, such as startups' descriptions, in this context, and providing information from it (Manning et al., 2008).

There are many ways to perform NLP, among the most used are Sentiment Analysis, Named Entity Recognition, Summarization, Topic Modeling, Text Classification, Keyword Extraction, Lemmatization, and Stemming. These last two will be explored in more detail in the next sub-chapters.

The integration of NLP techniques in predictive models can significantly enhance the accuracy of the results by enabling the analysis of more features that contain textual data. Therefore, applying some of the techniques above allows for the efficient processing of large volumes of text, reducing complexity and focusing on relevant information that contributes to predictive accuracy (Jurafsky & Martin, 2008).

2.4.1. Stemming

Stemming is one of the most common techniques employed within NLP and works by reducing words to their root form. Stemming algorithms such as the Porter stemmer truncate words by removing inflectional endings and simplifying the analysis by reducing the variability of forms (Porter, 1980). For example, words such as "connection", "connections", "connective", "connected", and "connecting" are reduced to the root "connect".

2.4.2. Lemmatization

The lemmatization technique, on the other hand, is a more sophisticated approach since it uses vocabulary and morphological analysis to remove inflectional endings only when it can deduce the base form of a given word based on its intended meaning. This way, lemmatization ensures that the root word belongs to the language, handling exceptions and different cases more accurately. For example, the word "better" is lemmatized to "good",

and the word "ran" to "run" (Bird et al., 2009). This accuracy makes the lemmatization approach very useful for tasks where semantic consistency is critical, such as keyword extraction, which is essential to analyze a startup's descriptions and gather relevant features from it.

2.4.3. TF-IDF

Term Frequency-Inverse Document Frequency, also known as TF-IDF is an essential technique in the processing of textual data. It evaluates how important a certain word is to a document in a collection of documents. TF-IDF increases proportionally with the number of times a word appears in the document but is offset by the frequency of the word across the corpus, helping to adjust for the fact that some words appear more often (Ramos, 2003). TF-IDF can be particularly useful when combined with techniques such as stemming and lemmatization, considering that these can reduce the words to their root forms and ensure that the application of TF-IDF calculation reflects the true semantic concentration of the terms in the documents, allowing more accurate and meaningful feature extraction from textual data (Salton & McGill, 1983). The calculation formula of this method can be seen below.

$$TF\ (t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d} \tag{2.1}$$

$$IDF(t) = log\ \frac{N}{1+df} \tag{2.2}$$

$$TF - IDF(t,d) = TF(t,d)\ x\ IDF(t) \tag{2.3}$$

where *d* refers to a document, *t* refers to a term, *N* is the total number of documents and *df* is the number of documents with the term t.

3. Project Description

The main objective of this chapter is to describe the project, its stakeholders, challenges, and desired goals, as well as how it intends to help VC groups in their screening process.

The project that gave birth to this study was a project between a VC group and LTPlabs in early 2023.

The VC group in question was founded in 2015 and is part of one of the business areas of a Portuguese multinational company. Its portfolio mainly includes investments in the areas of Technology, Retail, Cybersecurity, and Software. The project was developed with the Retail Technology vertical team that looks for potential investments in this particular industry, which was one of the challenges, as it is more difficult to find the right startups' desired profile.

As explained before, VCs handle a huge number of investment opportunities at once, making the screening process highly difficult. The excess of manual efforts, the reduced number of team members on this vertical, the scope of the screening process that could let companies off the radar, the lack of a scoring system and, therefore, prioritization of the startups analyzed, and the vast amount of unstructured data registry were some of the biggest problems of the group.

During the project, it was developed and validated a model that aimed to help the VC group in the screening process. This model used startups' descriptions to classify them according to their profile and check whether they were compatible with the desired profile of the VC group. The model was successful, saving enormous time and manual effort. It is trained to screen potential deals by checking hard filters such as the year of foundation, number of workers, and rounds of investments, as the group already did, and by detecting previously selected keywords on the company's descriptions that, based on historical data, are more prone to be found in startups that correspond to the desired profile. After that, it attributes them a fit score and ranks them. The model will be explained in more detail later on.

Previously, all this work was done by hand and by different people, leading to inconsistent analysis, since the team worked independently, and each had their own files and analysis methods.

The first project was, therefore, able to tackle some of the main problems of the group, such as the lack of automation of the vertical, the excess of manual effort of the tasks, the lack of

prioritization on their analysis, and the huge amount of unstructured data that existed, solved by integration to a database to centralize the data.

However, some problems remained, and others emerged. The VCs desired a model with boosted predictive capabilities that could include other features rather than just the keywords retrieved from the startups' descriptions. Not only that, but they wanted the model to be continuously fed with updated data as they feared missing out on investment opportunities, especially in a vertical as specific as the retail tech one. Finally, despite the development of an interface already made, the group wanted to be able to visualize and exploit the data on an integrated tool where multiple team members could work simultaneously.

Thus, the objective of this new internal project was to improve the former model by adding new variables that could enhance the business fit score, empower it with predictive capabilities, and upgrade the user experience and data visualization by developing an integrated platform. Then, the new and improved model could be commercialized and used as a scalable and generalizable tool for multiple VC groups.

The automation and expansion of the number of new investment opportunities that are collected allow advantages at different levels. On the one hand, it permits a reduction in startups that remain outside the VC's radar. On the other hand, it also guarantees more data to train and consequently improve the model developed. Furthermore, it would substantially reduce the time spent on the deal-sourcing process.

Additionally, the improved business and potential fit by ameliorating the model could increase the confidence in the tool and save time to perform more important tasks for the company's growth.

4. Methodology and Data

This chapter focuses on describing the data used to develop the work, including the data sources and data transformation processes, as well as the methodologies and techniques utilized throughout the project that helped solve the problem presented to the company.

Firstly, in the Data section, the data sources consulted, the variables used, and their characteristics will be presented. Secondly, the methodology and techniques used will be introduced.

4.1. Data

Similarly to the vast majority of the consulted literature, Crunchbase was used as the primary data source of the report. Crunchbase is an American company that provides information about a wide range of companies, most precisely startups (Crunchbase Inc., 2024). Its platform has different types of information, and it is used for a vast number of reasons. In addition to providing a large amount of data about the companies, it also informs users about funding and investment information, changes in leadership positions, and corporate news.

Despite that, the starting point of the data preparation was the already existent data provided by the project client. This database consisted of an extensive list of previously analyzed startups with some data about them, including a classification attributed by the VC group that indicated whether the firm was fit or not according to the profile desired by the client. From this dataset, the names of the startups and their corresponding fit classification were extracted. After that, the list was used by uploading it to the Crunchbase website to obtain the available data about each of them. Finally, using that data from Crunchbase, a complete database of startups that could be manipulated and worked on was built.

The database contained a vast number of features for each observation, which will be briefly named and described below. The set of variables was divided into four groups:

- Startup features – that compiled information related to the startup
- Funding features – that contained information on the funding rounds of the startup
- Founders' features – that included information about the founders of the startup
- Investors' features – that comprised features about the investors (if any) of the startup

Table 4.1 shows the complete list of the features and their corresponding brief description.

Table 4.1 – List of features and brief description

Feature	Brief description
Startup features	
organization_name	Name of the startup
success	Binomial variable that indicates if the startup is fit (1) or no-fit (0) for the Retail Tech vertical
description	Brief description of the startup business activity
num_employees	Number of employees working for the company
patents_granted	Number of patents that have been granted to the company, as detected by IPqwery
trademarks_regist	Number of trademarks that have been granted to the company, as detected by IPqwery
monthly_rank_growth	Percentage change in traffic rank of site in a given country from last month, as detected by SEMrush
active_tech_count	Number of technologies currently in use by this company, as detected by BuiltWith
total_prod_active	Number of products currently in use by the company, as detected by G2 Stack
it_spend_usd	Amount of money spent on IT (in US dollars)
country_code	Country code of the startup headquarters' location
company_age	Number of years since the company foundation
has_twitter	Binomial variable that indicates if the company has a Twitter profile (1) or not (0)
has_facebook	Binomial variable that indicates if the company has a Facebook profile (1) or not (0)
has_linkedin	Binomial variable that indicates if the company has a LinkedIn profile (1) or not (0)
company_num_articles	Number of articles published on the news about the company
industry	Qualitative variable that indicates the industries in which the startup operates

Funding features	
num_fund_rounds	Number of funding rounds in which the startup took part
last_fund_amount_usd	Amount of the last funding round made (in US dollars)
total_equity_fund_amount_usd	Total amount of funds obtained through equity (in US dollars)
total_fund_amount_usd	Total amount of funds obtained (in US dollars)
months_since_last_fund	Months since the last fund
Founders' features	
num_founders	Number of founders of the startup
founders_gender	Categorical variable that takes value 0 if there are no information about founders' gender or if they prefer not to tell, 1 if founders are solely male, 1 if they are solely female and 2 if they are composed by both.
founders_num_founded_org	Number of organizations that were founded by the startup founder
founders_num_exits	Number of successful exits performed by the startup founder
founder_technical_profile	Binomial variable that indicates if the startups founders' background is of interest of the VC group
founders_num_articles	Number of articles published on the news about the startup founders
founders_school_num_alumni	Number of alumni from the university attended by the startup founders
founders_school_num_alumni_founders	Number of alumni that founded their organizations from the university attended by the startup founders
Investors' features	
num_investors	Number of investors of the startup
investor_num_investments	Number of investments made by the startup investors
investor_num_exits	Number of successful exits performed by the startups the investors invested in
investor_num_contacts	Number of Crunchbase contacts associated with the investor

The dataset contains 338 startups, of which 201 were classified as "no-fit" and, therefore, their success variable is 0, while the remaining 137 were marked as "fit" and, consequently, their success corresponds to 1. The startups in this dataset are exclusively those that participate in certain specialized events, hence the great weight of "fits". In a more comprehensive set, the percentage of "fits" would most probably be below 1%. However, this also helps in further steps to have a balanced dataset.

Additionally, a different dataset containing thousands of startups unseen by the model was used later on to simulate the model's performance in the intended context. This dataset contained the same features as the previously mentioned one, and the source was the same (Crunchbase).

4.2. Methodology

The process of building a complete database was a fundamental step towards applying the CRISP-DM methodology. This methodology stands for Cross Industry Standard Process for Data Mining, and it is a robust and widely adopted framework for developing data mining and ML projects. The methodology is used across various industries due to its flexibility and comprehensive nature. It presents several advantages: the ability to handle complex data science projects, the definition of a clear and repeatable process, and the focus on aligning technical efforts with business objectives. Due to its well-defined roadmap, CRISP-DM helps ensure that all aspects of data mining and ML projects are carefully considered and executed, leading to more accurate outcomes. Accordingly, it provides a structured approach that contains 6 phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. This iterative process begins with comprehending the business objectives, requirements, and the client's problems in this context.

The beginning of this first phase was marked by the start of the project but continued throughout it, as the process of learning and deepening business knowledge was transversal and continuous, causing the need to revisit the Business Understanding phase at various times.

It was followed by an in-depth exploration of the data to identify quality issues, patterns, and insights encompassed in the Data Understanding phase. The Data Preparation phase involves cleaning, transformation, and structuring the data to make it suitable for the

Modeling phase. These steps were mentioned and started in the previous sub-chapter by exploring the available data and extracting new ones from Crunchbase. Nevertheless, the dataset was not yet fully ready for the Modeling Phase. Therefore, additional transformations were performed to the dataset in different manners.

Some of the transformations were more profound than others, but all were crucial to ensure the appropriate performance of the models when applied later.

4.2.1. Data Understanding and Preparation Phases

Initially, all observations and features were ensured to have associated values and were checked for any missing values (NAs). In instances where NAs were found, they were addressed in a manner most fitting for the context. For the majority of the variables, assigning a value of zero was deemed appropriate, as applied to features like "founders_num_articles", "num_fund_rounds", "last_fund_amount_usd", "patents_granted", and others. Conversely, for variables such as "num_founders", assigning a zero value was illogical, so for this particular case the mode and the average of the variable were calculated to understand the most common number of founders in the dataset for each startup. The result was two, so that value was assigned in the case of NAs. In addition to these cases, there were others where a few observations contained NA values for variables such as "country_code" or "company_age". Since the number of observations in this condition was quite small and the information would be easily verified, some manual research was performed, and these values were assigned to the respective startups to ensure the highest possible data quality and reliability.

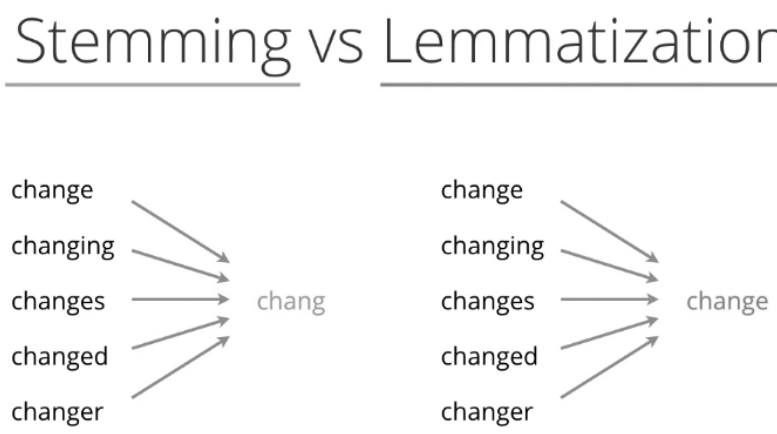
After ensuring that all observations contained the corresponding data and no missing values, more extensive data transformations were performed.

Binary variables were created based on existent features to transform categorical variables into numerical ones or enhance their usability. The "country_code" variable was used to create two dummies "is_US" and "is_EU" that takes the value one if they are from the United States or Europe, respectively, and zero otherwise. Besides that, the variable "interest_industry" was created based on the "industry" feature that takes value one if the industry that the startup operated in was of interest to the VC group based on its requests, and zero otherwise. Another transformation was applied to the "num_investors" variable to produce the "has_investors" binary variable, reflecting whether a startup has investors. This

transformation occurred due to the significant influence of the "num_investors" variable observed during the Modeling Phase.

The "description" variable that contained the description of the startup was subjected to text processing techniques such as Stemming and Lemmatization. These NLP procedures, already mentioned in the Review Literature chapter, have the function of processing unstructured data such as text and transforming it into understandable and interpretable data for computers to be used in a valuable way. Before that, the text suffered some adjustments such as the conversion of all characters to lowercase and the removal of stop words such as "a", "in", "of", "the", and "with", among dozens of others. The punctuation was also removed, and the numbers were converted to their written form. Only after these modifications took place were the stemming and lemmatization applied. These techniques convert words to their root form with the Stemming algorithm, Porter stemmer, removing inflectional endings and simplifying the analysis by reducing the variability of forms and the Lemmatization algorithm, WordNetLemmatizer, doing it in a more sophisticated way, reducing words to their root form based on its intended meaning (Bird et al., 2009). An example of a word being submitted to both procedures can be seen in Figure 4.1.

Figure 4.1 – Example of Stemming and Lemmatization techniques applied



This way, all startups had a group of words associated with them, which was crucial in calculating the TF-IDF score of each word. The TF-IDF score evaluates how important a certain word is to a document in a collection of documents and it is calculated as shown in Equations 2.1 to 2.3 in Chapter 2.

For the calculation of this score, the startups were divided into two groups based on their "fit" or "no fit" feature. Each word had, therefore, two scores, one for when it was included in the "fit" group and another when it was in the "no fit" group. Afterwards, the score obtained for being in the "no fit" group was subtracted from the score obtained for being in the "fit" group to achieve a final score for each word. This method meant that the words that appeared most frequently in startups classified as "no fit" more easily obtained negative final scores, and vice versa. Next, the formula and a few examples of final score calculations for some words are presented in Table 4.2 to explain how it is done.

$$Final\ Score(w) = Fit\ Score(w) - No\ fit\ Score(w) \tag{4.1}$$

where *w* represents each word.

Table 4.2 – Examples of Final Score calculation

Word (after text processing techniques)	Fit Score	No fit score	Final Score
deliveri	0	0,01137	-0,01137
servic	0,00787	0,02476	-0,01689
tool	0,00845	0,00130	0,00715

Table 4.3 presents a list of the 10 words with the highest and lowest scores.

Table 4.3 – Final Score ranking of words after text processing techniques

Ranking	Word (after text processing techniques)	Final Score
1	retail	0,02163
2	product	0,01908
3	build	0,01542
4	buy	0,01295
5	make	0,01217
6	payment	0,01151
7	sourc	0,01086
8	merchant	0,01063
9	growth	0,00977
10	brand	0,00947
…		
1128	order	-0,00880
1129	softwar	-0,00978
1130	enabl	-0,01019
1131	base	-0,01022
1132	social	-0,01125
1133	deliveri	-0,01137
1134	edg	-0,01199
1135	compani	-0,01247
1136	servic	-0,01689
1137	offer	-0,01964

Then, the "description_score" variable that was created represents the sum of the final scores of the words included in each startup's description, attributing a numerical classification to each startup description based on the dictionary of words developed.

Subsequent steps involved assessing the skewness and outliers of features to guarantee that the data distribution adhered to the assumptions of statistical tests and models planned for use in the following phase. This step involved visualizing the data using histograms and calculating skewness coefficients to identify any significant deviations from normality. Regarding the outliers, they were detected using z-scores, ensuring that any extreme values were appropriately handled, to enhance the robustness and accuracy of the model.

Starting with the skewness investigation, the results were obtained by applying the Fisher-Pearson coefficient of skewness methodology, which is shown below:

$$Skewness_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right) \quad (4.2)$$

where X_i are data points, $\bar{X}$ is the sample mean, σ is the sample standard deviation and n is the number of data points in the sample

According to this method, if the distribution is not symmetrical, the skewness results will be positive if skewed to the right and negative if skewed to the left. If perfectly symmetrical, skewness will be zero.

The application of the formula retrieved the skewness for each variable. From those, the ones with higher absolute values were more carefully and deeply analyzed to understand the reason behind the skewness result and if it should be performed any normalization procedure to the respective variable. The variables that justified that attentive analysis was "it_spend_usd", "last_fund_amount_usd", "total_equity_fund_amount_usd", "total_fund_amount_usd", "patents_granted", "trademark_regist" and "founders_num_articles".

In the analysis, given their considerable numerical variation, a logarithmic transformation was applied to variables related to monetary values. For instance, a funding difference of two hundred thousand dollars and four hundred thousand dollars or two million dollars versus two and a half million dollars, while significant in absolute terms, has a relatively minor impact when considering the practical scale of the companies involved. This log

transformation helps to stabilize variance and normalize the distribution, making the differences in scale more comparable and reducing the influence of outliers. However, the log transformation was not used for other variables that exhibited similar skewness such as the number of patents granted or the amount of trademark registers. The rationale was that the skewness itself provides a meaningful reflection of the underlying data distribution and should be preserved to maintain the integrity of the model's relationship with these variables. Thus, while logging transformation aids in addressing skewness in monetary variables, it was not necessary for all skewed variables, as the inherent skewness can be valuable for accurately modeling the relationships within the data.

The skewness values before and after the log transformation can be seen in detail in Table 4.4. These new variables seem more balanced, with skewness values closer to zero, thus potentially leading to better results while in the Modeling Phase.

Table 4.4 – Skewness values of a set of variables before and after log transformation

Variable	Skewness before log transformation	Skewness after log transformation
it_spend_usd	10,51	-0,09
last_fund_amount_usd	7,22	-1,13
total_equity_fund_amount_usd	7,19	-1,58
total_fund_amount_usd	7,80	0,23

The log transformations performed, consequently, gave rise to four new variables, “log_it_spend_usd”, “log_last_fund_amount_usd”, “log_total_equity_fund_amount_usd” and “log_total_fund_amount_usd”, respectively.

Regarding outliers' inspection, the z-score values and a defined threshold were used. The calculation of the z-score was performed as follows:

$$Z = \frac{x - \mu}{\sigma} \tag{4.3}$$

where Z corresponds to the z-score, x is the observed value, μ is the mean of the sample, and σ is the standard deviation of the sample.

The threshold was set to three, as is conventionally done and therefore, any z-score value greater than three or less than minus three was considered an outlier. The z-scores for the remaining variables with higher skewness values were also checked to verify if discarding certain observations was justified. Thus, the observations from the variables mentioned above like "patents_granted", "trademark_regist" and "founders_num_articles" were examined.

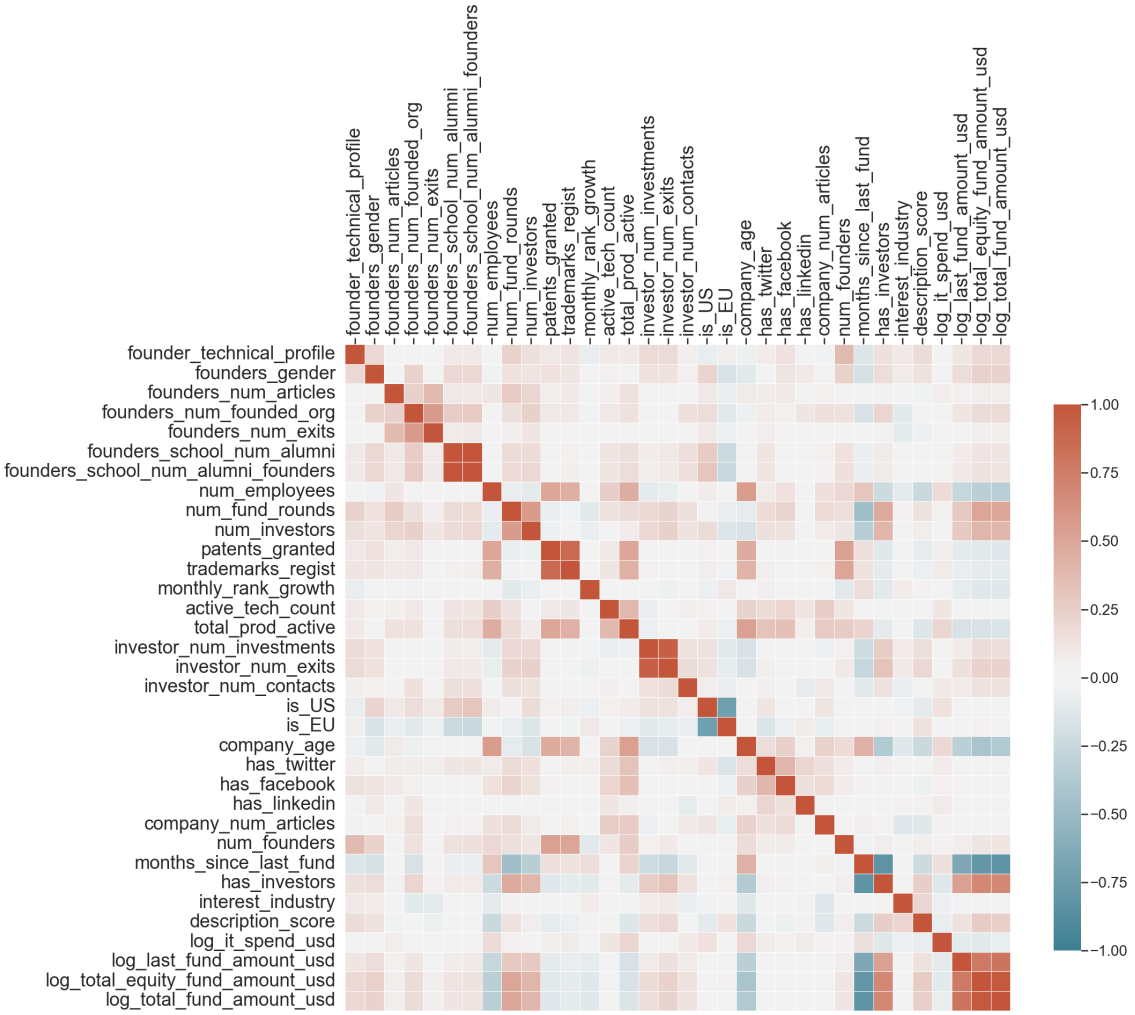
During this analysis, a modest number of outliers were identified. Upon removing them, the skewness values approached closer to zero, although not significantly. Given the limited dataset, where removing outliers would result in a loss of approximately 15% of the observations, it was decided to retain these outliers in the dataset. This decision was reinforced by a meticulous manual review of each outlier, confirming their validity and relevance within the context. Consequently, a higher degree of skewness of these variables was accepted to preserve the integrity and robustness of the model, particularly since the dataset was already constrained in size. This approach underscored a deliberate trade-off between achieving a lower skewness and maintaining a substantial number of observations to ensure a comprehensive analysis, and the second option was prioritized. In Table 4.5, the differences in skewness values on those particular variables before and after the removal of the outliers can be seen.

Table 4.5 – Skewness values of a set of variables before and after outliers' removal

Variable	Skewness before outliers' removal	Skewness after outliers' removal
founders_num_articles	9,22	5,19
patents_granted	14,29	12,73
trademarks_regist	11,78	8,57

Finally, the Data Understanding and Preparation Phase ends with a check of the correlation between all independent variables, serving not only for that but also to consult all the variables that exist after the various procedures carried out during these two phases that will be used in the Modeling Phase.

Figure 4.2 – Correlation matrix between all the variables on the dataset



Looking closely at the correlation matrix, some more pronounced correlations can be detected. Some of them were explored in more detail. There seems to exist a high positive correlation between the number of alumni from the universities attended by the founders ("founder_school_num_alumni") and the number of alumni from the universities attended by the founders that are also founders ("founder_school_num_alumni_founders") which seems plausible since it is more likely to have more students who founded their companies at universities with a greater number of alumni. The number of patents granted ("patents_granted") and trademark registers ("trademarks_regist") also present a positive

correlation given their similarity. The two binary variables "is_US" and "is_EU" that take the value one if the startups have their headquarters located in the United States or Europe, respectively, and zero otherwise, present a negative correlation, a reasonable relation considering that the majority of the startups analyzed belong to one of these two.

Regarding the number of investments from the startup investors ("investor_num_investments") and the number of exits from the startup investors ("investor_num_exits"), it appears to be a positive correlation, exhibiting the idea that more successful investors have a higher number of investments and are more prone to have a higher number of exits. Finally, exploring the funding-related variables ("log_last_fund_amount_usd", "log_total_equity_fund_amount_usd", and " log_total_fund_amount_usd) and their positive correlation among them, it is particularly expectable that a higher amount of one of these is correlated to a higher amount of the other two, or vice-versa. These three variables are also negatively correlated with the number of months since the last funding round ("months_since_last_fund"), and the possible explanation for it is that if a startup has a higher number of months that passed since its last funding round, it may mean that they have less access to funds and, therefore, they raise less money.

Despite some correlations between variables, it was decided to keep them all, in this stage, to determine in the next phase what variables might be the best for the model and take the necessary precautions regarding this problem at that time.

4.2.2. Modeling Phase

Subsequently, the Modeling Phase brought some challenges, given the need to define the variables to be used in the model, the technique to be implemented, and its assumptions and characteristics. During this phase, various algorithms are applied to the prepared data to build predictive models.

Regarding feature selection, the employed approach is based on three different techniques, combining the advanced algorithmic capabilities of the Boruta features selection method, empirical testing across various predictive models, and insights obtained from an extensive literature review.

The Boruta algorithm plays a fundamental role in the feature selection process. As an all-relevant variable selection method, Boruta works within an RF classification framework to

determine the importance of each feature(Kursa & Rudnicki, 2010). It does this by comparing the significance of the original attributes against that of randomly shuffled copies of these attributes, called shadow attributes. This method is particularly effective because it seeks to capture all features that carry significant information for prediction, ensuring a comprehensive understanding of the variables that impact VC decision-making.

To complement the Boruta analysis, it was conducted empirical testing of various predictive models, such as Random Forest, Logistic Regression, Support Vector Machine, Decision Trees, Gradient Tree Boost, and Naives-Bayes, using all the candidate variables available to evaluate the statistical significance and performance impact of each variable. This testing helped gather more information about features and confirm the robustness of the selected features by demonstrating their consistent contribution to predictive accuracy.

The process was further refined through an exhaustive literature review identifying variables traditionally emphasized in VC decision-making studies. This step ensured that the selected features were statistically significant, relevant, and validated by existing research and industry practices, giving business sense to the choices.

The integration of these three approaches allowed the adoption of a holistic approach to feature selection. This robust method leverages automated ML tools, statistical validations, and industry-specific knowledge to construct a model that can be both statistically sound and highly relevant to real-world applications. The comprehensive feature selection process is crucial for building a trustworthy model, enhancing the VC screening process, and ultimately leading to better investment decisions and performances. The selected features and their evaluation in each one of the techniques can be seen in Table 4.6.

Table 4.6 – Feature selection techniques

Variables	Selected to final model	Boruta	Empirical testing	Literature Review
founder_technical_profile				
founders_gender				X
founders_num_articles			NB	
founders_num_founded_org	X	X	SVM	X
founders_num_exits			DT; GTB; SVM	X
founders_school_num_alumni			RF; GTB; NB; SVM	
founders_school_num_alumni_founders			GTB; NB; SVM	
num_employees	X	X	NB; SVM	X
num_fund_rounds	X	X	RF; DT; GTB	X
num_investors	X	X	RF	X
patents_granted			NB	X
trademarks_regist	X	X	DT; GTB	X
monthly_rank_growth	X	X	RF; DT; GTB; SVM	
active_tech_count		X	RF; GTB	
total_prod_active			DT	
investor_num_investments		X		
investor_num_exits	X	X	RF; DT	X
investor_num_contacts				
is_US				X
is_EU				X
company_age	X	X	RF; NB	X
has_twitter			SVM	
has_facebook				

has_linkedin			SVM	
company_num_articles			DT; NB	
num_founders			DT; GTB	
months_since_last_fund	X	X	RF; DT; GTB; NB	
interest_industry				
description_score	X	X	RF; DT; GTB; NB; SVM	
log_it_spend_usd			DT	
log_last_fund_amount_usd		X	RF; DT; GTB; SVM	
log_total_equity_fund_amount_usd		X	RF; NB; SVM	
log_total_fund_amount_usd	X	X	RF; NB	X

Following the rigorous selection of features, the Random Forest Model was employed as the primary tool for this study. RF operates by constructing multiple decision trees during training and outputting the class, which is the mode of the classes or mean prediction of the individual trees. RF models are particularly effective for predictive modeling due to their robustness against overfitting, even with large datasets. This robustness arises from the model's ability to average multiple decision trees, trained on different parts of the same training set, reducing drastically the variance without substantially increasing the bias.

Regarding model training and validation, a dataset split of 80/20 was utilized. This means that 80% of the data was used to train the model, and the remaining 20% was used for test its performance. This split ratio is conventionally applied because it offers a balanced trade-off between having enough data to train the model effectively and enough data to test and validate the model's predictions. This standard split helps ensure that the model can generalize well to new data, which is critical given the predictive nature of the task.

The choice of different split ratios, such as 70/30 or 90/10, could potentially affect the model's performance and reliability. For instance, a 70/30 split, while providing more data for testing, reduces the amount available for training, which could be detrimental when working with a limited dataset, as is the case. It may impede the model's ability to learn

effectively, potentially leading to underfitting. On the other hand, a 90/10 split would provide more data for training but less for testing, which could lead to overfitting, where the model performs well on training data but poorly on unseen data. Given the scarce size of the database at hand, the 80/20 split was deemed most appropriate, balancing the need for adequate training data with sufficient testing to ensure the accuracy and generalizability of the model.

An extensive optimization process using a Grid Search method was implemented to enhance the model's predictive accuracy and generalizability, particularly for the needs of VC decision-making (Pedregosa et al., 2012).

This method finds the best possible combination of hyperparameters for a given model. Its use presents some benefits because it automates the tuning process, ensuring that the chosen model performs at its optimal level by exploring a comprehensive range of configurations, allowing considerable time savings over manual experimentation, and enhancing the model's efficacy and reliability in practical applications. The Grid Search meticulously evaluated a range of critical hyperparameters: n_estimators, which indicates the number of trees in the forest where a higher count typically boosts performance at the expense of increased computational load; max_depth, which caps the trees' growth to prevent overfitting by limiting their complexity; min_samples_split and min_samples_leaf, defining the minimum samples at a node before a split and at a leaf, respectively, being crucial for ensuring the model does not adapt too specifically to the training data; max_features, determining the number of features considered for each split to reduce overfitting risks; and class_weight, which adjusts the importance of each class in the calculations, ensuring balanced sensitivity towards majority and minority classes.

A rigorous 5-fold cross-validation accompanied the Grid Search. This method divides the dataset into five parts, training the model on four and validating on one, rotating until each part has served as the validation set. This strategy is also essential to ascertain whether the model's effectiveness is consistent across various data segments and not merely specific to a particular data portion.

According to the Grid, the optimal hyperparameters were n_estimators set at 100 to ensure a robust performance without excessive computational demands, max_depth of 10, offering a balanced approach to learning depth, max_features at "sqrt", limiting the features per split to avoid overfitting while maintaining sufficient model complexity and min_samples_leaf at

four and min_samples_split at 10, both to ensure greater model stability and prevent overfitting. The utilized string to build the model can be seen in Equation 4.4.

```
RFmodel = RandomForestClassifier(n_estimators=100, max_depth=10,
max_features='sqrt', min_samples_leaf=4, min_samples_split=10)
```
(4.4)

This finely tuned model contrasts sharply with the baseline model by demonstrating enhanced capacity to be accurate and reliable on unseen data, facilitated by cross-validation. The combination of advanced algorithmic tools and empirical validation through cross-validation underpins the model's capacity to support VCs in making more informed and data-driven decisions.

4.2.3. Evaluation and Deployment Phases

The Evaluation Phase assesses the model's performance and alignment with business goals, ensuring its validity and reliability. This phase will be explored in more detail in the next chapter, together with the Deployment Phase, which aims to integrate the model into the business processes for practical use.

Nevertheless, it is important to mention that in the preparation of these phases, it was crucial to revisit the Data Understanding and Preparation phases to perform the same data transformations and techniques on the new dataset of unseen startups and, therefore, make sure the features and their formatting corresponded to the ones applied before. This allowed for reinforced confidence in the model predictions considering that the data source and the Data Preparation phase were consistent.

5. Results and Model Evaluation

This chapter serves to show the results obtained through the application of the methodology described in Chapter 4 based on the Random Forest Model as a predictive model for sorting startups based on their fit score. This analysis aims to understand how accurate the model is in predicting startups with the desired profile by the VC and improve the results compared to the previous model. Therefore, several techniques and metrics will be displayed to explore the model's robustness and performance, such as the correlation between the explanatory variables and those with the explained variable, the significance of each variable, the ROC curve and AUC area, the comparison between actual and predicted values and consequently the construction of a confusion matrix and calculation of the Accuracy, F1, F2, and F0,5 scores. Then, the former and new models will be compared considering a stronger aspect of business sense. Finally, the model will be applied to a more extensive database of startups. According to the model, the ten highest classified companies will be analyzed to ensure the model retrieves startups with the desired profile.

The model applied, as described in Chapter 4, was a Random Forest model with its mentioned hyperparameters, including a 5-fold cross-validation and a test split with a ratio of 80/20.

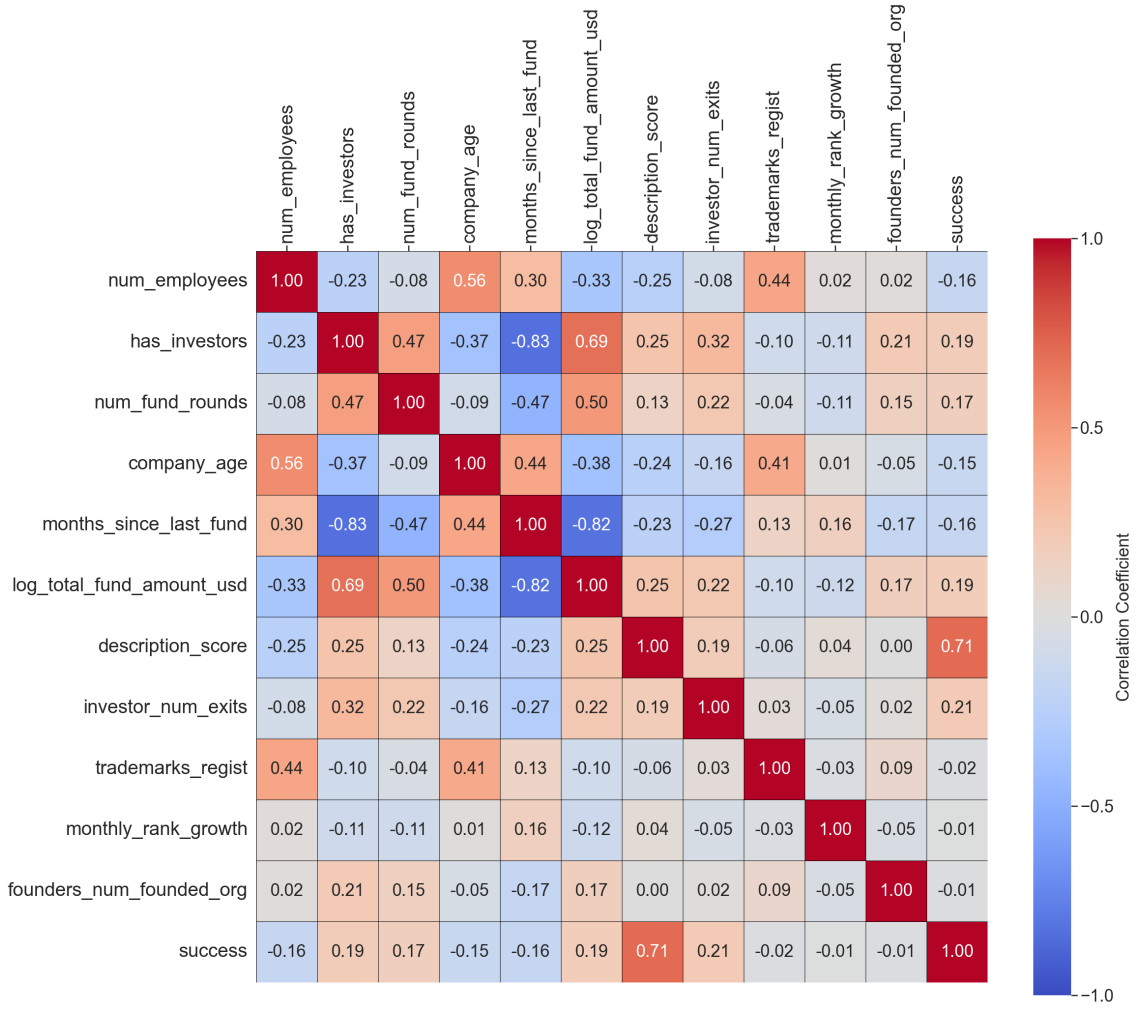
5.1. Model Evaluation

The model developed reveals several crucial insights and highlights its strengths, and areas for potential improvement.

5.1.1. Correlation matrix

Regarding the variables selected based on the techniques employed and displayed in Table 4.6, it is possible to explore each and understand their meaning considering the context of investment in VC. To help with this task, Figure 5.1 demonstrates the correlation between the explanatory variables and their correlation with the explained variable, "success".

Figure 5.1 – Correlation matrix between all selected variables



The variable "num_employees", which represents the number of employees on a given startup was selected and exhibits a negative correlation with the explained variable. This is consistent with the VC group's preferences for early-stage investment since they normally present fewer workers. Contrarily, the variable "has_investors", which retrieves the values one or zero considering the existence or not of investors, presents a positive correlation with the explained variable, demonstrating that startups that are already being backed up present are more attractive to investors, possibly due to a vetted proof of concept or initial market validation, reducing the perceived risks associated with new investments. The variable "num_fund_rounds" reflects the total number of funding rounds a startup has participated in and correlates positively with the explained variable. Similarly to the previous feature, the validation from other investors and their confidence in a given startup can make the investment more attractive.

On the other hand, it may also contradict the early-stage investment profile desired by the group if there are too many rounds. Moving to the "company_age" variable, it indicates the company's age and, as expected, has a slightly negative correlation with the explained variable, supporting the desired profile by the VC group. The "months_since_last_fund" variable, which corresponds to the number of months that passed since the last time a startup secured funding, shows a negative correlation with the explained variable, suggesting that startups who have not raised funds recently are less attractive for the VC group. This may reflect a current and active pursuit of growth and expansion strategies, which are crucial characteristics sought by investors. Interestingly, "log_total_fund_amount_usd", which represents the total amount of US dollars a given startup has raised through funding rounds, shows a positive correlation identically to other funding-related features and probably for the same reasons. The "description_score", unsurprisingly, has a higher positive correlation. Given the calculation technique utilized and described in Chapter 4, it is quite intuitive to understand that this variable is highly capable of capturing the characteristics not only of the startup but of the industry in which it operates, providing a clear picture of its activity and, therefore, making it easier to VC groups to understand if they correspond or not to its desired profile. The "investor_num_exits" is a very important variable since it shows how many successful exits (via IPO or acquisition) the startup investors are associated with. It is pretty clear that the higher this number, the more attractive the startup is since there is more confidence in investing in it. Therefore, there is a positive correlation with the explained variable. The "trademarks_regist", which represents the number of trademarks registered by the startup, suggests that while intellectual property protection is relevant, it is not a strong differentiator in the VC group's evaluation process. The same happens with the "monthly_rank_growth", which shows the monthly growth on startups' website visits, and with "founders_num_founded_org", which contains the number of organizations founded by the startup founder.

5.1.2. Feature Importance

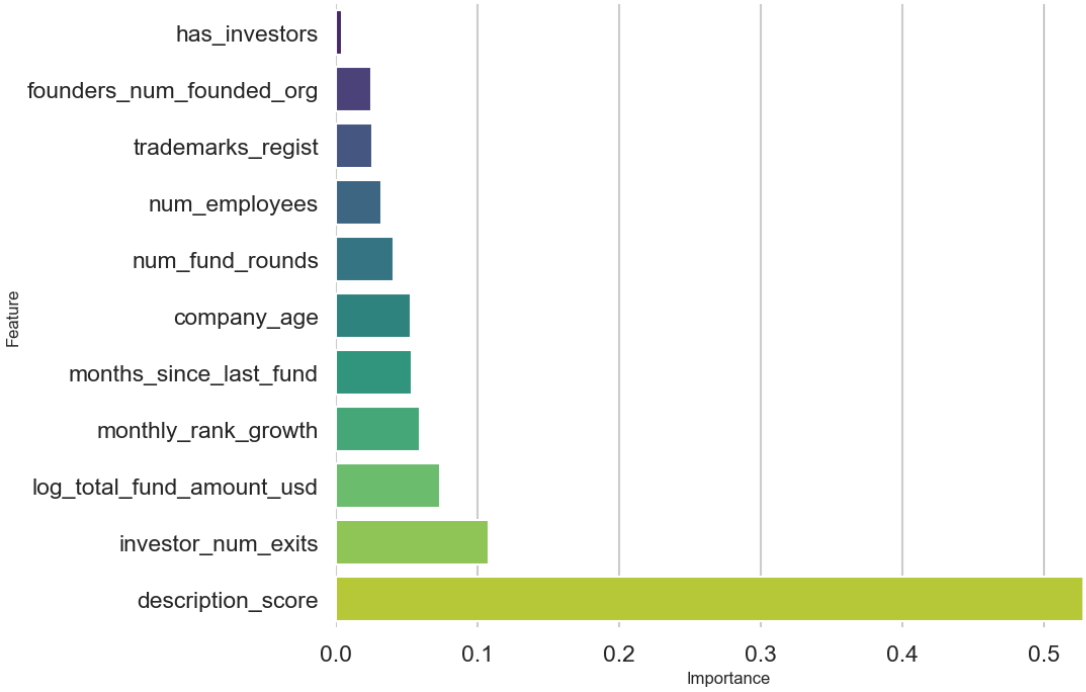
The previous analysis can be even more elucidative when complemented with a feature importance analysis, with the help of Figure 5.2., since the importance of various features in the model provides further insights into what drives the decision-making process of the VC group when evaluating startups in the screening process.

The feature importance analysis in an RF model provides a critical comprehension of which feature most significantly influences the model's predictions. This analysis evaluates each feature's contribution towards enhancing the model's predictive accuracy.

As explained before, in an RF model, each decision tree is developed by randomly selecting subsets of features and data points. Features that frequently contribute to optimal splits at key decision points, or nodes, in these trees are considered more important. This is because their use at these junctures effectively improves the node's purity, consequently impacting the tree's overall predictive decisions. Essentially, if a feature regularly aids in creating better splits that result in purer nodes, it is assigned a higher importance score. The importance of a feature is quantitatively assessed based on how much it decreases the impurity of a node, with this decrease weighted by the probability of reaching that node (derived from the probabilities of reaching preceding nodes). These importance values are then averaged across all trees within the forest to derive a comprehensive measure of each feature's effectiveness across the model. Thus, the most important features are those that not only appear frequently but also make significant contributions to the model's performance.

It is important to clarify that while high importance indicates a strong association with the response variable, it does not necessarily imply causality. Moreover, as the previous analysis did, this method does not account for correlations between features since importance might be attributed to a feature that is a substitute for others not explicitly identified by the model.

Figure 5.2 – Features' importance of RF model



Standing out as the most important part of the model, the "description_score" confirms its critical role. As explained before, there is no better feature to draw the startup profile than its description, and the TF-IDF technique applied allows us to understand the startup activity based on keywords that have positive and negative impacts on the model. Among the remaining variables, "investor_num_exits" is the second most important on the model, reflecting the value placed on a startup's backing by investors with a successful history. Other variables present very similar importance, contrary to "has_investors", which surprisingly shows the lowest significance for the model. Consequently, it was decided to remove this feature from the model, and the results will not consider this variable as part of the model. Despite the strong negative correlation between the variables "months_since_last_fund" and "log_last_fund_amount_usd", they were kept because both present substantive importance for the model, and the results would worsen without their inclusion. The updated correlation matrix and feature's importance graph and correspondent values can be seen in detail in Figure B.1, Figure B.2, and Table B.1, respectively.

5.1.3. Cross-Validation and Accuracy Scores

To further evaluate the model's performance, the cross-validation scores obtained were analyzed and can be seen in Table 5.1.

Cross-validation serves as a crucial method to assess the model's effectiveness across various subsets of data. Specifically, k-fold cross-validation, where the dataset is segmented into equal parts, offers insightful evaluations by repeatedly training and testing the model on different partitions. As explained before, this method mitigates the risk of overfitting and provides a more comprehensive assessment of performance across different data samples.

Table 5.1 – Cross Validation scores

k-folds	1	2	3	4	5
Cross-Validation Scores	0,815	0,833	0,852	0,926	0,870
Average Cross-Validation Score			0,859		

The 5-fold cross-validation applied to this model enables a robust evaluation by using different dataset segments as the test set across five iterations. The cross-validation scores appear to be consistent and stable across subsets. Despite some variations across subsets, with the lowest being 0,815 and the highest 0,926, the average of the five subsets, calculated at approximately 0,859, reinforces the model's strong predictive accuracy.

The range between subsets suggests that certain data partitions might contain certain patterns that are either very well captured by the model or represent areas where the model might still be improved. The highest score of 0,926 demonstrates the model's potential when the conditions are optimal.

The high predictive performance and the consistency presented across different scenarios and data variations are significant given the importance of these two characteristics in developing the model to be used in a real-life context since the robustness presented lends confidence in the model's utility.

Continuing to assess the model's performance, Table 5.2 shows the F1, F2, F0,5 and Accuracy scores. A threshold of 0,5 was applied to these analyses, meaning that if a startup scores 0,5 or above, it is considered as one and zero, otherwise.

Table 5.2 – F1, F2, F0,5 and Accuracy scores

Metrics	Scores
F1 Score	0,806
F2 Score	0,776
F0,5 Score	0,839
Accuracy	0,824

Accuracy is one of the most intuitive performance measures as it is simply the ratio of correctly predicted observations to the total observations. The Accuracy score stands at 82,4%, which suggests a robust performance and means that more than 82 out of 100 startups that correspond to the desired profile of the VC group are being captured by the model.

The F1 score is the weighted average of Precision and Recall. Precision is the percentage of true positive predictions out of all the positive predictions. Recall, also known as Sensitivity, is the percentage of true positive predictions out of all the positive actual cases. Therefore, this score takes both false positives and false negatives into account. It is calculated as follows:

$$F1\ Score = \ 2\ X\ \frac{Precision\ X\ Recall}{Precision + Recall} \tag{5.1}$$

The F2 score places more emphasis on Recall than Precision, meaning that this score gives more importance to capturing all the positive actual cases, even if it captures negative cases in the prediction. Contrarily, the F0,5 score places more emphasis on precision than recall, which means that the score finds it more important to make sure that the positive predictions are true positives. The calculation of both can be seen below:

$$F2\ Score = 5\ X\ (\frac{Precision\ X\ Recall}{4\ X\ Precision + Recall}) \quad (5.2)$$

$$F0{,}5\ Score = 1{,}25\ X\ (\frac{Precision\ X\ Recall}{0{,}25\ X\ Precision + Recall}) \quad (5.3)$$

The higher F0,5 score compared to the F2 Score favors the context of this project since it is more important for the VC group to make sure that all positive predictions are true positives, so as not to waste time analyzing false positive cases as they would if they did not apply the model. Therefore, even if the model does not effectively capture all the positive cases that exist, those it does capture are more precise and almost certainly positive in reality, which is very important for the investors in the screening process.

5.1.4. Confusion Matrix and Actual vs Predicted Values

The confusion matrix offers a more detailed perspective of model performance across the predictive classes. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class. The confusion matrix can be seen in Table 5.3.

Table 5.3 – Confusion Matrix

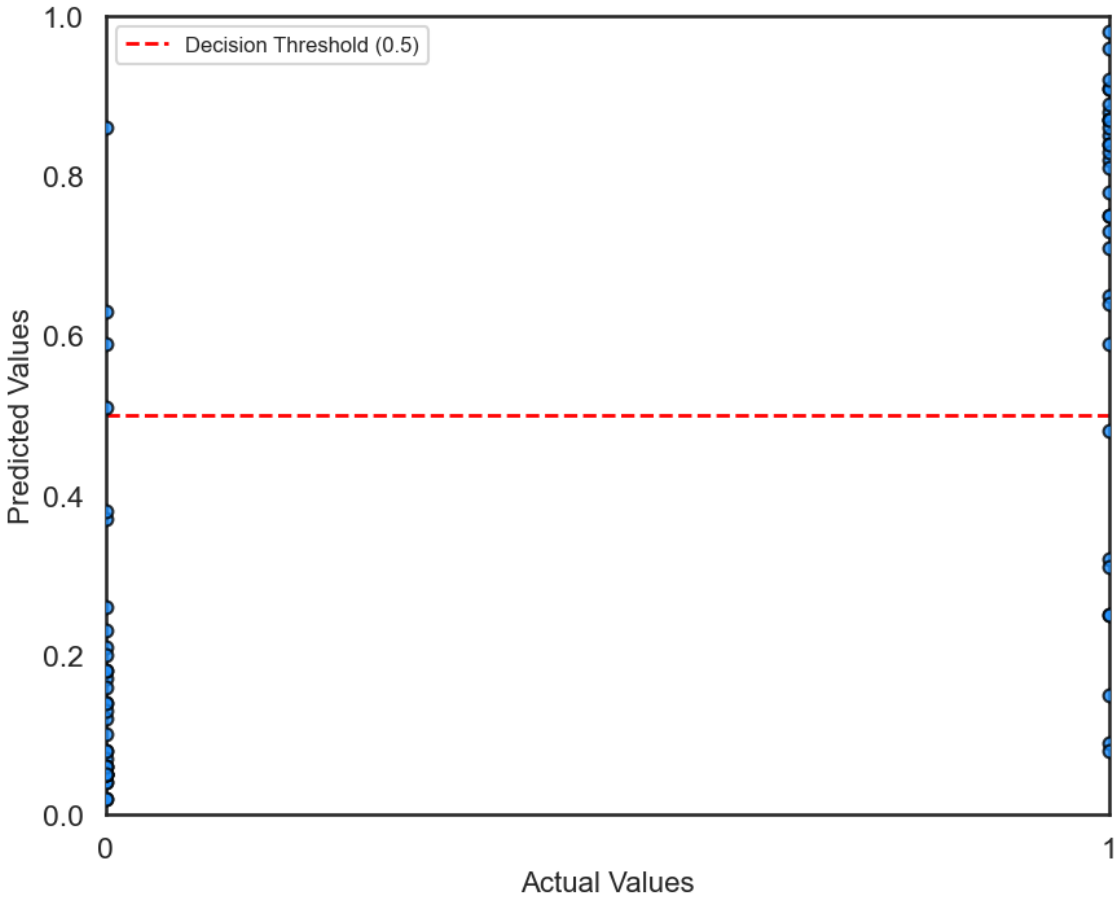
		Predicted	
		0	1
Actual	0	31	4
	1	8	25

According to the Confusion Matrix, the model predicts 31 true negative cases and 25 true positive cases, while presenting only 4 false positives and 8 false negatives, corroborating the analysis made earlier that the majority of the positive predictions it makes are true positives.

These cases are very well visible in Figure 5.3 which displays the dispersion of predicted values and their comparison with the actual values given the threshold defined at 0,5. On the left axis are the observations whose actual value is zero and the model predictions for those

observations that vary from zero to one. On the right axis are the cases where the actual value equals one and the corresponding predictions of those observations.

Figure 5.3 – Actual vs Predicted values



5.1.5. ROC Curve and AUC

The Receiver Operating Characteristic (ROC) curve is a graphical tool to assess the diagnostic ability of binary classifiers. It plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The TPR and FPR are calculated as follows:

$$TPR = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{5.4}$$

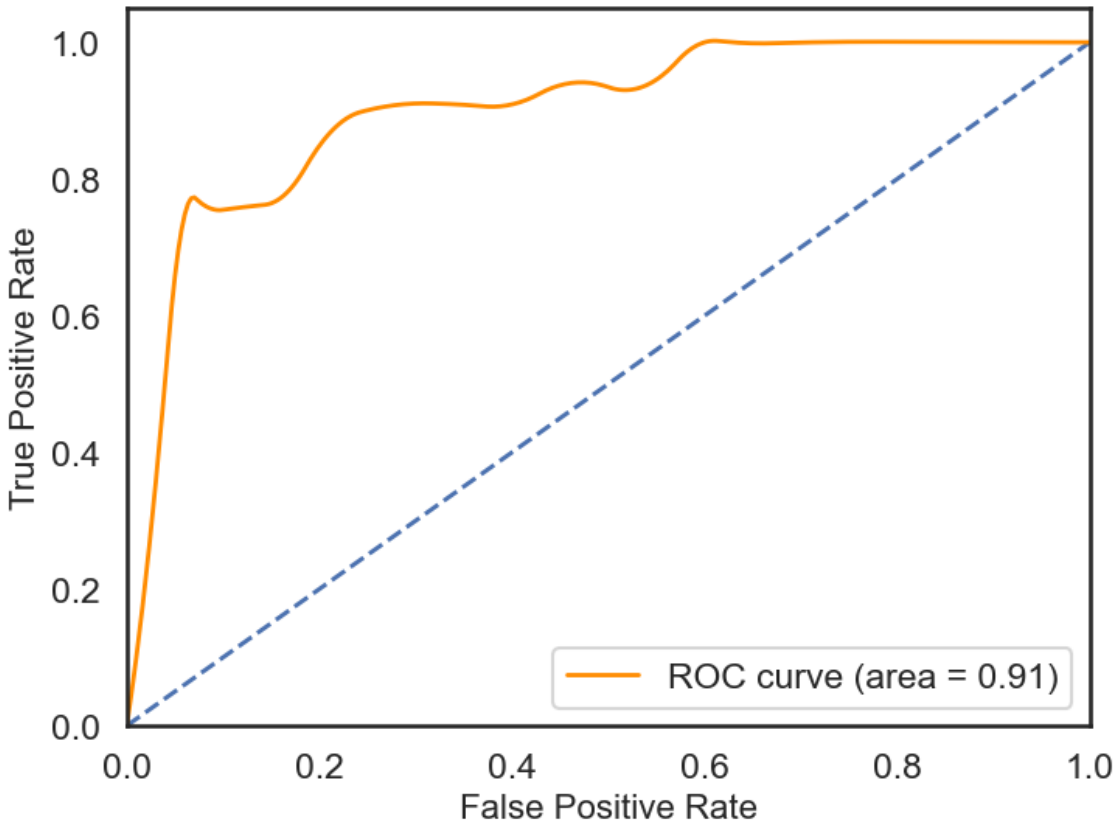
$$FPR = \frac{False\ Positives}{False\ Positives + True\ Negatives} \tag{5.5}$$

To generate the ROC curve, the model's prediction outcomes are transformed into a range of probabilities. A threshold is then varied from zero to one to calculate the TPR and FPR

at each one of those thresholds. If the model had perfect discrimination, it would have a curve passing through the left and upper axis and an Area Under the Curve (AUC) of one, although that would probably mean that overfitting problems could be the cause. Contrarily, an AUC of 0,5 suggests no discriminative ability, where predictions would be equivalent to random guessing, and the ROC curve would coincide with a diagonal line across the graph.

According to Figure 5.4, the ROC obtained in the model appears to have a high level of diagnostic ability with an AUC of 0,91. The model can distinguish between the positive and negative classes across a range of thresholds. Therefore, it can reduce the likelihood of false positives and false negatives as compared to other models with lower AUC.

Figure 5.4 – ROC Curve and AUC area



5.2. Results' comparison between models

Moving on to analysis that is more in line with the real-life context and comparing the performance of the new model developed and the model previously applied in the VC group, Table 5.4 provides the contrast of the predictive power of both models. The table shows, on a total of thirty fits, how many startups with the desired profile can the models capture on

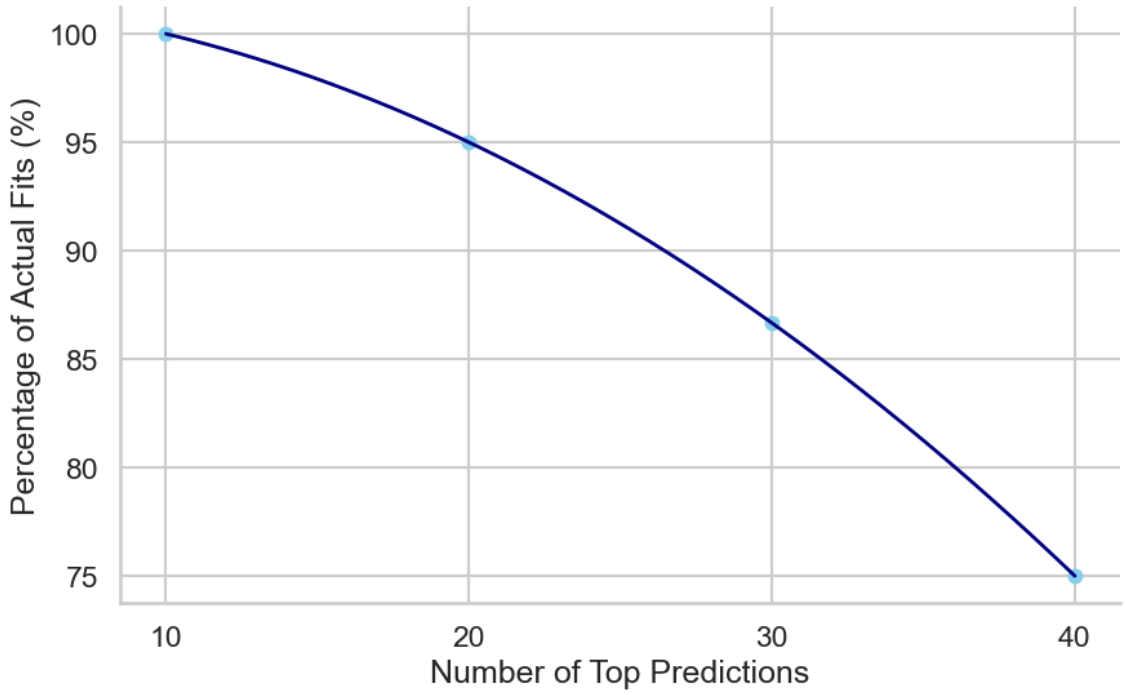
their best predictions. In other words, how many predictions (ordered by score) does the model take until it captures all the thirty real fits?

Table 5.4 – Models' comparison on fit distribution per top predicts

	Top predicts ordered by score				
	1-20	20-40	40-60	60-80	80-100
Old model	15	10	2	1	2
New model	18	12	-	-	-

The new model presents a higher predictive power than the previous model by capturing all the thirty fits in much fewer predictions. The great result of the new model illustrates its effectiveness in ranking and identifying the most suitable startups for a given VC group. As expected, capturing new ventures indicates diminishing returns on additional predictions as seen in Figure 5.5.

Figure 5.5 – Percentage of actual fits on model's top predictions



The development and evaluation of the new model underscore a significant improvement in predictive performance compared to its predecessor, fulfilling one of the project's primary goals and this particular study. This advancement is a testament to the robustness of the model's design and its effectiveness in aligning with the analytical needs of VC decision-making. Thus, the superior accuracy and reliability demonstrated by the new model suggest that it can be a valuable tool in enhancing investment decision processes by facilitating the screening process of VC groups.

Therefore, the next phase will involve deploying this model across a broader spectrum of startups. This will enable a comprehensive investigation into the model's practical applicability and the relevance of its predictive outcomes in real-world scenarios.

5.3. Model Predictions

This sub-chapter aims to apply the developed model to a wide range of unseen startups and evaluate the top-10 predictions considering the desired profile of the VC group, simulating how the model should be applied in a real-life context. Such analysis will validate the model's effectiveness and refine its predictive capabilities, ensuring that it remains a fine-tuned tool in its quest. The dataset used was the one mentioned in Chapter 4.1, corresponding to a snapshot of the Crunchbase database. Therefore, this could affect the replicability of the results.

Table 5.5 presents some features that help characterize the startups in which the VC group previously invested. The features correspond to what they were at the time of investment.

Table 5.5 - Main features of VC's previous investments in the Retail Technology vertical (at the moment of investment)

Startup	Industries	Country	Age	No. employees	Funding Rounds	Total Funding Amount (USD)
Company A	E-commerce; Marketing; SaaS	UK	4	51-100	5	4 100 000
Company B	Financial Services; FinTech; Travel	US	5	11-50	4	18 000 000
Company C	AI; E-commerce; Visual Search	SG	6	51-100	4	14 000 000
Company D	AI; Product Research; Retail Tech	US	2	1-10	0	0
Company E	AI; Retail; Retail Tech	PT	1	1-10	0	0
Company F	Cloud Computing; SaaS; Software	ES	7	51-100	3	1 323 000
Company G	E-commerce; SaaS; Software	US	2	1-10	1	2 000 000
Company H	AI; Food and Beverage; Supply Chain	US	5	51-100	4	32 800 000
Company I	Advertising	FI	4	11-50	2	550 000
Company J	AI; Image Recognition; SaaS	PT	1	1-10	0	0
Company K	Advertising; E-commerce; Software	CH	3	1-10	0	0
Company L	CRM; E-commerce; SaaS	US	3	11-50	2	25 000 000
Company M	AI; Fashion; Retail Tech	ES	4	101-250	2	2 000 000
Company N	Retail Tech; Software	PT	5	11-50	0	0

In examining the investment patterns of the Retail Technology vertical of the VC group, it becomes apparent that their strategic portfolio leans towards startups at the intersection of technology, innovation, and scalable solutions. This selection is not merely reflective of current trends in VC but suggests an alignment with companies with significant market impact and technological disruption. The vertical demonstrates a strong preference for industries such as AI, E-commerce, and Software as a Service (SaaS), for instance, investments in companies such as Company J and Company M leverage AI to revolutionize retail and E-commerce through smart solutions that enhance consumer interaction and operational efficiency. Geographically, the investments are mostly made in companies from the United States and Europe, major innovation hubs, and some emerging markets like Singapore. The predominance of the United States and Europe is also due to the way the VC group searches for investment opportunities, which is mainly through events. Companies from these geographic areas are, as a rule, more representative of startup-related events. This approach values established centers of technological advancement and emerging regions with high growth potential, helping the group expand the universe of potential opportunities and diversify its risks, taking advantage of a broader spectrum of market dynamics. Furthermore, their preference falls among younger companies with greater growth potential. The younger age of the startups, the reduced number of workers, and the smaller number of funding rounds are some of the demonstrative signs of VC's preference and are also being captured by the model.

Table 5.6 shows the top 10 predictions of the model and the correspondent fit score attributed to it.

Table 5.6 – Top-10 predictions according to the RF model

Startup	**Fit Score (attributed by the model)**
XStak Inc.	0,94
Alocalo	0,93
Inventa	0,92
Adshero	0,92
LOOX	0,90
Imagr	0,89
KwicPik	0,88
DRSSR	0,88
Retail.me	0,88
Fairmart	0,86

Additionally, Table 5.7 presents some features of these startups.

Table 5.7 - Features of the top-10 predictions

Startup	Industries	Country	Age	No. employees	Funding Rounds	Total Funding Amount (USD)
XStak Inc.	E-commerce; Marketing; SaaS	US	2	51-100	1	560 000
Alocalo	Retail Tech; Shopping; Software	DE	3	11-50	1	2 444 236
Inventa	E-commerce; Marketplace; Retail Tech	BR	3	101-250	4	80 358 865
Adshero	Advertising; E-commerce; Retail Tech	PL	3	11-50	0	0
LOOX	E-commerce; Fashion; Marketplace	BD	3	11-50	3	175 000
Imagr	AI; Retail Tech; Software	NZ	7	11-50	1	9 500 000
KwicPik	Food and Beverage; Marketplace; Retail Tech	AE	3	1-10	0	0
DRSSR	E-commerce; Retail Tech; Shopping	IT	0	1-10	0	0
Retail.me	Advertising; E-commerce; Marketing	DE	3	11-50	0	0
Fairmart	E-commerce; Internet of Things; Retail Tech	SG	3	11-50	2	1 500 000

The predictions obtained by the model reflect a continued emphasis on technological innovation with a focus on AI, E-commerce, and SaaS. Also, these startups seem to leverage advanced technologies to create novel solutions addressing specific industry challenges. The predictions appear to be consistent with previous investment options from the VC group given the shared focus on industries, the age of the companies selected, and their dimension and amount of funding rounds. Nevertheless, the startups selected by the model diverge from a geographical point of view compared with VC's previous investments. In this set, the geographic diversity is more pronounced with startups from a wider range of countries.

There are still some presences of the United States and some European countries like Poland, Italy, and Germany. Still, there are also other countries such as Bangladesh, New Zealand, Brazil, and the United Arab Emirates. The absence of location-specific variables in the RF model can explain this. This absence implies that the model does not inherently prioritize or discriminate against startups based on their geographical location, resulting in a broader array of countries represented in the output. Nevertheless, suppose this characteristic poses a concern for the VC group, particularly if there is a strategic intent to focus on specific regions. In that case, this issue can be easily rectified by implementing hard filters on the dataset before its submission to the RF model.

Next, each of the predicted startups will be carefully analyzed to understand their similarities or differences compared with the VC's desired profile to examine the predictions and understand if they were accurate qualitatively.

5.3.1. Xstak Inc.

XStak Inc. operates at the convergence of AI, e-commerce, and retail technology, particularly focusing on business intelligence and point-of-sale solutions. The firm's alignment with the VC's interest in retail technology and e-commerce platforms is evident. Notably, XStak Inc. shares a striking similarity with Company B, one of the VC's prior investments operating within retail technology and AI-driven e-commerce solutions. The fact that XStak Inc. only has had one funding round with less than 600,000 US dollars raised, and its youth makes the startup more appealing as it presents a high growth potential opportunity.

The prediction that XStak Inc. fits the VC's portfolio appears accurate, given its strong alignment with the VC's focus on technologically innovative solutions in retail and e-commerce.

5.3.2. Alocalo

Alocalo offers a localized e-commerce service that suggests nearby retailers to consumers, enhancing their shopping experience through geolocation technology. This startup's focus on local market penetration differs from any previous direct investments by the VC, which have tended towards broader applications. Despite this, Alocalo's use of technology to enhance retail experiences shares foundational principles with Company J's approach to retail innovation. Although the prediction might be seen as slightly off-target if the VC prioritizes broader market scalability over localized solutions, the potential growth of the startup given

its lack of maturity and reduced number of workers can make Alocalo worthy of VC's attention.

5.3.3. Inventa

Inventa is a B2B wholesale marketplace, connecting brands with independent buyers across the Americas. This startup from Brazil aligns well with the VC's established preference for innovative e-commerce platforms, similar to Company A's role in enhancing customer relationships through data-driven marketing. Both focus on streamlining transactions and enhancing the connectivity between sellers and their markets. Despite the higher funding rounds, the prediction aligning Inventa with the VC's interests seems accurate, reflecting the VC's inclination towards transformative e-commerce technologies.

5.3.4. Adshero

Adshero focuses on advertising technology, offering a network for targeted advertising solutions that enhance retailer engagement and sales. Its orientation towards digital marketing and retail parallels the capabilities seen in Company G, another VC investment that also exploits digital tools for retail and analytics enhancement. However, Adshero's specific focus on advertising technologies provides a narrower scope than Company G's broader analytics services, making it more akin to Company K. This specialization could be viewed positively, fitting well within the VC's interest in digital marketing innovations, suggesting a correct prediction. Once again, the startup's youth, combined with the reduced number of workers and the lack of funding rounds, make Adshero a great growth potential startup.

5.3.5. LOOX

LOOX ventures into fashion retail, providing an augmented reality platform that offers virtual try-on solutions to enhance the consumer shopping experience. While it integrates cutting-edge technology similar to that appreciated by the VC, its direct focus on consumer interaction through AR is more niche compared to the broader technological solutions previously favored. Therefore, while LOOX presents an innovative approach and despite the existence of previous investments by the VC in the fashion industry, such as Company M, its prediction as a fit might stretch the VC's typical portfolio boundaries, possibly marking it as a less precise alignment.

5.3.6. Imagr

Imagr specializes in AI-driven autonomous checkout solutions, closely mirroring the capabilities of Company E, which provides AI solutions for checkout-free retail environments. The direct overlap in technology and market application confirms Imagr's strong alignment with the VC's investment in retail technology innovations. This makes the prediction highly accurate, as Imagr exemplifies the type of transformative technology the VC aims to support in the retail sector. Despite the greater company's maturity compared to others, it is possible to verify that this factor is not exclusionary as the VC group invested in Company F when the Spanish company was the same seven years old as Imagr.

5.3.7. KwicPik

KwicPik addresses inventory management within the food and beverage sector using AI, closely paralleling Company H, which focuses on fresh produce inventory optimization. Both startups leverage AI to improve supply chain efficiencies, although KwicPik's niche in food and beverage provides a more focused approach. The prediction seems correct, given the VC's interest in AI-driven supply chain solutions (although the niche focus may slightly narrow its appeal), the lack of funding rounds, and the youth of the startup.

5.3.8. DRSSR

DRSSR integrates e-commerce with social shopping, allowing users to discover trends and styles. While it shares some characteristics with Company M, which focuses on fashion retail merchandising, DRSSR's focus on consumer social engagement through technology introduces a new dynamic. This adds a contemporary edge that might be appealing to the VC. The prediction aligning DRSSR with the VC's preferences could be seen as a correct extension into new yet related areas, and its characteristics make it a high-growth potential startup.

5.3.9. Retail.me

Retail.me operates a B2B platform for manufacturers and suppliers, enhancing retail operations through integrated data services. This startup resembles Company F in its functionality, which also focuses on optimizing product information management across retail channels. The similarity in enhancing data utilization in retail and the startup's maturity

level confirms that Retail.me's inclusion in the VC's potential investments is well-predicted, fitting within the VC's investment profile.

5.3.10. Fairmart

Fairmart enhances the visibility of retail products by making them easily searchable and purchasable globally, aligning closely with the objectives of many of the VC's previous investments, like Company H and Company L. Both focus on optimizing product accessibility and improving market reach through technology. Allied with the company's youth and a small amount of funds raised, Fairmart's global approach to retail problems echoes the VC's preference for scalable, market-wide solutions, making this prediction particularly accurate.

Thus, this analysis reveals a largely correct alignment with the VC's Retail Technology's desired profile, despite some deviations. Nevertheless, the model seems capable of finding startups that correspond to a certain established picture, and it can do it in a real-life scenario context by analyzing data that has never been seen before. This is extremely important because the model must be theoretically sound and practically effective in assisting VCs with their investment decisions in real life. By proving its utility in identifying and prioritizing promising investment opportunities, the model increases confidence among venture capitalists to rely on its insights, allowing them to delve deeper into these opportunities at a later stage. This confidence is vital for the model to be a generalizable and scalable tool that could be adopted by other VC firms, enhancing its applicability across the industry.

5.4. Startup Screening Tool

One of the other objectives of the project was the development of an integrated platform to upgrade the user experience and data visualization. Therefore, while not being fully developed, the Startup Screening Tool is a big step towards merging advanced analytical capabilities and an interactive, easily manipulated tool for VC groups. This future platform will have immeasurable value for the companies operating in this area given the ease and quickness of obtaining startups' fit classification through which they can prioritize their analyses, save a lot of time on tasks with little added value, work simultaneously with the team and make more consistent analyzes.

In Figure C.1, it is possible to see some illustrative displays of the appearance of the tool and its development, where VCs can extract data from multiple sources, apply hard filters

according to their preferences, generate results based on the developed model, and navigate through the list of startups.

6. Conclusion

Venture capital firms face a formidable challenge in the initial stage of their investments: screening hundreds, or even thousands, of potential startups. Traditionally, this process has been heavily reliant on manual analyses and heuristic approaches, which demand significant time and effort and carry a high risk of bias. Local bias or overconfidence can skew decision-making, leading to inconsistencies and potentially suboptimal investment choices. Given these challenges, there is a compelling need for methodologies that streamline the screening process while minimizing human error and bias.

This study presents a shift towards a more systematic and data-driven approach through the development and implementation of ML models. Specifically, using a Random Forest model significantly advances the VC startup screening process. The model's ability to accurately identify over 82%, and up to 92% under optimal conditions, while minimizing the false positive predictions, underscores its effectiveness. More importantly, this model surpasses the performance of the previously validated and implemented model, highlighting improvements in inaccuracy, reliability, and consistency across evaluation when applied in real-life scenarios.

The implications of integrating such a predictive model into VC's screening process are profound. VCs can expedite their decision-making processes by automating the initial analysis of potential investments. This automation ensures a more uniform application of investment criteria, thereby reducing the impact of individual biases that may arise from a diverse team of investors. Furthermore, the model aids in prioritizing investments by immediately highlighting startups that match the VCs' criteria, effectively streamlining the screening process.

This efficiency does not merely speed up analyses. It reallocates valuable resources and time towards activities that foster deeper and more strategic engagements. For example, VCs can dedicate more time to nurturing client relationships or delve deeper into market analysis to spot emerging trends and opportunities. Such tasks carry significant added value and can greatly enhance a firm's competitive edge and market positioning, improving its performance.

Integrating advanced analytics and ML models into the venture capital investment process represents a paradigm shift. This study confirms that leveraging such technologies enhances

the efficiency and accuracy of the startup screening process and plays a crucial role in modernizing the VC sector. The model's proven real-life accuracy and robustness, especially its performance with unseen data, further validate its utility and reliability. This adaptability is crucial, demonstrating the model's capacity to function effectively in dynamic, real-world environments. Such capabilities ensure that VCs can rely on the model's outputs, reinforcing their confidence in the tool as a reliable support mechanism for making informed investment decisions. As these tools continue to evolve and improve, they promise to become indispensable in the daily routines of VCs, transforming traditional, labor-intensive tasks into streamlined, strategic operations. This progression towards automation and data-driven decision-making is poised to revolutionize the landscape of VC investments, making it more dynamic, precise, and insightful than ever before.

6.1. Limitations

While the development of the RF model has demonstrated considerable improvements in the efficiency and effectiveness of the venture capital startup screening process, several limitations must be acknowledged. One significant constraint in this study was the limited availability of detailed data regarding the specific preferences or criteria that constitute a desired fit or no-fit for startups from the perspective of the VC group. This scarcity of data restricted the model's training and testing capabilities, potentially impacting the accuracy and generalizability of the results. Enhanced access to detailed, structured data on investment outcomes and preferences could improve the model's predictive power and reliability.

Another limitation was the completeness and accessibility of the data from Crunchbase. While it offers a robust dataset, the absence of key features that might influence investment decisions, such as detailed financial performance or in-depth market penetration metrics, somewhat limited the depth of analysis possible. Moreover, the lack of API access imposed additional challenges in efficiently retrieving and updating data, further complicating the research process and the model's potential to stay current with real-time changes.

6.2. Future Work

There are several ways to enhance the robustness and applicability of the predictive models used in VC startup screenings. Exploring additional features such as market conditions, competitor analyses, and more comprehensive financial projections could provide deeper insights, even though startups' financials may not always reflect their long-term viability or

innovation potential. Furthermore, incorporating qualitative data, such as management team experience or brand strength, might enrich the model's contextual understanding.

Moreover, expanding the dataset to include a broader array of startups and possibly integrating more granular historical data could allow future models to capture dynamic market trends and growth trajectories that were not feasible with the current dataset. This expansion would not only improve the model's accuracy but also its adaptability to diverse market conditions.

Additionally, experimenting with other advanced machine learning techniques and models could uncover more insights and potentially offer better performance. Techniques such as deep learning or ensemble methods might be explored to harness their capability in handling complex patterns and large datasets, especially when more comprehensive data becomes available. The application of web-scrapping techniques on other data sources would also be beneficial in obtaining more descriptions and texts about the startups and getting even more accurate fit scores.

References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence* (1st ed.). Harvard Business Review Press.

Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of Machine Learning Performance for Decision Support in Venture Capital Investments. *IEEE Access*, *7*, 124233–124243. https://doi.org/10.1109/ACCESS.2019.2938659

Åstebro, T., & Elhedhli, S. (2006). The Effectiveness of Simple Decision Heuristics: Forecasting Commercial Success for Early-Stage Ventures. *Management Science*, *52*(3), 395–409. http://www.jstor.org/stable/20110517

Bai, S., & Zhao, Y. (2021). Startup Investment Decision Support: Application of Venture Capital Scorecards Using Machine Learning Approaches. *Systems*, *9*, 55. https://doi.org/10.3390/systems9030055

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*.

Bolander, T. (2019). What do we loose when machines take the decisions? *Journal of Management and Governance*, *23*(4), 849–867. https://doi.org/10.1007/s10997-019-09493-x

Breiman, L. (2001). *Random Forests* (Vol. 45).

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification And Regression Trees*. Routledge. https://doi.org/10.1201/9781315139470

Brynjolfsson, E., & McElheran, K. (2016). The Rapid Adoption of Data-Driven Decision-Making. *American Economic Review*, *106*(5), 133–139. https://doi.org/10.1257/aer.p20161016

Catalini, C., Foster, C., & School, H. B. (2018). *Machine Intelligence vs. Human Judgement in New Venture Finance * Ramana Nanda*.

Chen, H., Gompers, P., Kovner, A., & Lerner, J. (2010). Buy local? The geography of venture capital. *Journal of Urban Economics*, *67*(1), 90–102. https://doi.org/10.1016/J.JUE.2009.09.013

Corea, F., Bertinetti, G., & Cervellati, E. M. (2021). Hacking the venture industry: An Early-stage Startups Investment framework for data-driven investors. *Machine Learning with Applications*, *5*, 100062. https://doi.org/10.1016/J.MLWA.2021.100062

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, *20*(2), 215–242. http://www.jstor.org/stable/2983890

Crunchbase Inc. (2024). *Crunchbase data: Your guide to business information*.

Cumming, D., & Dai, N. (2010). Local bias in venture capital investments. *Journal of Empirical Finance*, *17*(3). https://doi.org/10.1016/j.jempfin.2009.11.001

Dellermann, D., Lipusch, N., Ebel, P., Popp, K., & Leimeister, J. M. (2017). Finding the Unicorn: Predicting Early Stage Startup Success Through a Hybrid Intelligence Method. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3159123

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management*, *48*, 63–71. https://doi.org/10.1016/J.IJINFOMGT.2019.01.021

Eesley, C., Hsu, D., & Roberts, E. (2013). The Contingent Effects of Top Management Teams on Venture Performance: Aligning Founding Team Composition With Innovation Strategy and Commercialization Environment. *Strategic Management Journal*, *35*. https://doi.org/10.2139/ssrn.1498740

Ewens, M., & Marx, M. (2018). Founder Replacement and Startup Performance. *Review of Financial Studies*, *31*, 1532–1565. https://doi.org/10.1093/rfs/hhx130

Ferrati, F., & Muffatto, M. (2021). Entrepreneurial Finance: Emerging Approaches Using Machine Learning and Big Data. *Foundations and Trends® in Entrepreneurship*, *17*. https://doi.org/10.1561/0300000099

Franke, N., Gruber, M., Harhoff, D., & Henkel, J. (2006). What you are is what you like—similarity biases in venture capitalists' evaluations of start-up teams. *Journal of Business Venturing*, *21*(6), 802–826. https://doi.org/10.1016/J.JBUSVENT.2005.07.001

Ghassemi, M. M., Song, C., & Alhanai, T. (2020). The Automated Venture Capitalist: Data and Methods to Predict the Fate of Startup Ventures. *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:211528244

Gompers, P. A., Gornall, W., Kaplan, S. N., & Strebulaev, I. A. (2020). How do venture capitalists make decisions? *Journal of Financial Economics*, *135*(1). https://doi.org/10.1016/j.jfineco.2019.06.011

Gompers, P., Gornall, W., Kaplan, S., & Strebulaev, I. (2021, April). How Venture Capitalists Make Decisions. *Harvard Business Review*.

Gompers, P., Kovner, A., Lerner, J., & Scharfstein, D. (2010). Performance persistence in entrepreneurship. *Journal of Financial Economics*, *96*(1), 18–32. https://doi.org/10.1016/J.JFINECO.2009.11.001

Gompers, P., & Lerner, J. (1998). *The Determinants of Corporate Venture Capital Successes: Organizational Structure, Incentives, and Complementarities*. https://doi.org/10.3386/w6725

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (Vol. 2). Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Hosmer, D., & Lemeshow, S. (2000). Introduction to the Logistic Regression Model. In *Applied Logistic Regression* (pp. 1–30). John Wiley & Sons, Ltd. https://doi.org/https://doi.org/10.1002/0471722146.ch1

Huang, M.-H., Rust, R., & Maksimovic, V. (2019). The Feeling Economy: Managing in the Next Generation of Artificial Intelligence (AI). *California Management Review*, *61*(4), 43–65. https://doi.org/10.1177/0008125619863436

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Jurafsky, D., & Martin, J. (2008). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Vol. 2).

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *The American Economic Review*, *93*(5), 1449–1475. http://www.jstor.org/stable/3132137

Kim, J., Kim, H., & Geum, Y. (2023). How to succeed in the market? Predicting startup success using a machine learning approach. *Technological Forecasting and Social Change*, *193*, 122614. https://doi.org/10.1016/J.TECHFORE.2023.122614

Kleinbaum, D. G., & Klein, M. (2010). *Logistic Regression*. Springer New York. https://doi.org/10.1007/978-1-4419-1742-3

Koellinger, P., Minniti, M., & Schade, C. (2007). "I think I can, I think I can": Overconfidence and entrepreneurial behavior. *Journal of Economic Psychology*, *28*(4), 502–527. https://doi.org/10.1016/J.JOEP.2006.11.002

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, *36*(11), 1–13. https://doi.org/10.18637/jss.v036.i11

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity, and Variety*.

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*. https://doi.org/10.13140/2.1.1570.5928

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Monika, & Sharma, A. K. (2015). Venture Capitalists' Investment Decision Criteria for New Ventures: A Review. *Procedia - Social and Behavioral Sciences*, *189*, 465–470. https://doi.org/10.1016/J.SBSPRO.2015.03.195

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., & Louppe, G. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*.

Piscopo, C., & Birattari, M. (2008). The Metaphysical Character of the Criticisms Raised Against the Use of Probability for Dealing with Uncertainty in Artificial Intelligence. *Minds and Machines*, *18*, 273–288. https://doi.org/10.1007/s11023-008-9097-3

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130–137. https://doi.org/10.1108/eb046814

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/BF00116251

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Elsevier. https://doi.org/10.1016/C2009-0-27846-9

Ramos, J. (2003). *Using TF-IDF to determine word relevance in document queries*.

Retterath, A. (2020). *Human Versus Computer: Benchmarking Venture Capitalists and Machine Learning Algorithms for Investment Screening*. https://ssrn.com/abstract=3706119

Rokach, L., & Maimon, O. (2005). Data Mining and Knowledge Discovery Handbook. In *Decision Tree Induction: A Comprehensive Review* (pp. 165–192). Springer-Verlag. https://doi.org/10.1007/0-387-25465-X_9

Ross, G., Das, S., Sciro, D., & Raza, H. (2021). CapitalVX: A machine learning model for startup selection and exit prediction. *The Journal of Finance and Data Science*, *7*, 94–114. https://doi.org/https://doi.org/10.1016/j.jfds.2021.04.001

Sahlman, W. A. (1990). The structure and governance of venture-capital organizations. *Journal of Financial Economics*, *27*(2), 473–521. https://doi.org/10.1016/0304-405X(90)90065-8

Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval*. https://api.semanticscholar.org/CorpusID:43685115

Simon, H. A. (1995). Artificial intelligence: an empirical science. *Artificial Intelligence*, *77*(1), 95–127. https://doi.org/10.1016/0004-3702(95)00039-H

Tyebjee, T., & Bruno, A. (1984). A Model of Venture Capital Activity. *Management Science*, *30*, 1051–1066. https://doi.org/10.1287/mnsc.30.9.1051

Wooldridge, J. (2002). Econometric Analysis of Cross SEction and Panel Data. In *booksgooglecom* (Vol. 1).

Zacharakis, A. L., & Meyer, G. D. (1998). A lack of insight: do venture capitalists really understand their own decision process? *Journal of Business Venturing*, *13*(1), 57–76. https://doi.org/10.1016/S0883-9026(97)00004-9

Zacharakis, A. L., & Meyer, G. D. (2000). The potential of actuarial decision models: Can they improve the venture capital investment decision? *Journal of Business Venturing*, *15*(4), 323–346. https://doi.org/10.1016/S0883-9026(98)00016-0

Zacharakis, A. L., & Shepherd, D. A. (2001). The nature of information and overconfidence on venture capitalists' decision making. *Journal of Business Venturing*, *16*(4), 311–332. https://doi.org/10.1016/S0883-9026(99)00052-X

Zhong, H., Liu, C., Zhong, J., & Xiong, H. (2018). Which startup to invest in: a personalized portfolio strategy. *Annals of Operations Research*, *263*(1–2). https://doi.org/10.1007/s10479-016-2316-z

Zider, B. (1998, December). How Venture Capital Works. *Harvard Business Review*.

Annexes

A. VC industry stages

Table A.1 - 8 stages of VC investment process by Sahlman (1990)

1. *Seed investments*
   Although the term is sometimes used more broadly, the strict meaning of 'seed investment' is a small amount of capital provided to an inventor or entrepreneur to determine whether an idea deserves of further consideration and further investment. The idea may involve a technology, or it may be an idea for a new marketing approach. If it is a technology, this stage may involve building a small prototype. This stage does not involve production for sale.
2. *Startup*
   Startup investments usually go to companies that are less than one year old. The company uses the money for product development, prototype testing, and test marketing (in experimental quantities to selected customers). This stage involves further study of market-penetration potential, bringing together a management team, and refining the business plan.
3. *First stage – early development*
   Investment proceeds through the first stage only if the prototypes look good enough that further technical risk is considered minimal. Likewise, the market studies must look good enough so that management is comfortable setting up a modest manufacturing process and shipping in commercial quantities. First-stage companies are unlikely to be profitable.
4. *Second stage – expansion*
   A company in the second stage has shipped enough product to enough customers so that it has real feedback from the market. It may not know quantitatively what speed of market penetration will occur later, or what the ultimate penetration will be, but it may know the qualitative factors that will determine the speed and limits of penetration. The company is probably still unprofitable, or only marginally profitable. It probably needs more capital for equipment purchases, inventory, and receivable financing.
5. *Third stage – profitable but cash poor*
   For third-stage companies, sales growth is probably fast, and positive profit margins have taken away most of the downside investment risk. But, the rapid expansion requires more working capital than can be generated from internal cash flow. New VC capital may be used for further expansion of manufacturing facilities, expanded marketing, or product enhancements. At this stage, banks may be willing to supply some credit if it can be secured by fixed assets or receivables.
6. *Fourth stage – rapid growth toward liquidity point*
   Companies at the fourth stage of development may still need outside cash to sustain growth, but they are successful and stable enough so that the risk to outside investors is much reduced. The company may prefer to use more debt financing to limit equity dilution. Commercial bank credit can play a more important role. Although the cash-out point for VC investors is thought to be within a couple of years, the form (IPO, acquisition, or LBO) and timing of cash-out are still uncertain.
7. *Bridge stage – mezzanine investment*
   In bridge or mezzanine investment situations, the company may have some idea which form of exit is most likely, and even know the approximate timing, but it still needs more capital to sustain rapid growth in the interim. Depending on how the general stock market is doing, and how given types of high tech stocks are doing within the stock market, 'IPO windows' can open and close in very unpredictable ways. Likewise, the level of interest rates and the availability of commercial credit can influence the timing and feasibility of acquisitions or leveraged buyouts. A bridge financing may also correspond to a limited cash-out of early investors or management, or a restructuring of positions among VC investors.
8. *Liquidity stage – cash-out or exit*
   A literal interpretation of 'cash-out' would seem to imply trading the VC-held shares in a portfolio company for cash. In practice, it has come to mean the point at which the VC investors can gain liquidity for a substantial portion of their holdings in a company. The liquidity may come in the form of an initial public offering. If it does, liquidity is still restricted by the holding periods and other restrictions that are part of SEC Rule 144, or by 'stand-off' commitments made to the IPO underwriter, in which the insiders agree not to sell their shares for some period of time after the offering (for example, 90 or 180 days). If acquisition is the form of cash-out, the liquidity may be in the form of cash, shares in a publicly traded company, or short-term debt. If the acquisition is paid for in shares of a nonpublic company, such shares may be no more liquid than the shares in the original company. Likewise, if the sellers take back debt in a leveraged buyout, they may wind up in a less liquid position than before, depending on the liquidity features of the debt.

B. Model Evaluation

Figure B.1 – Updated correlation matrix between all selected variables

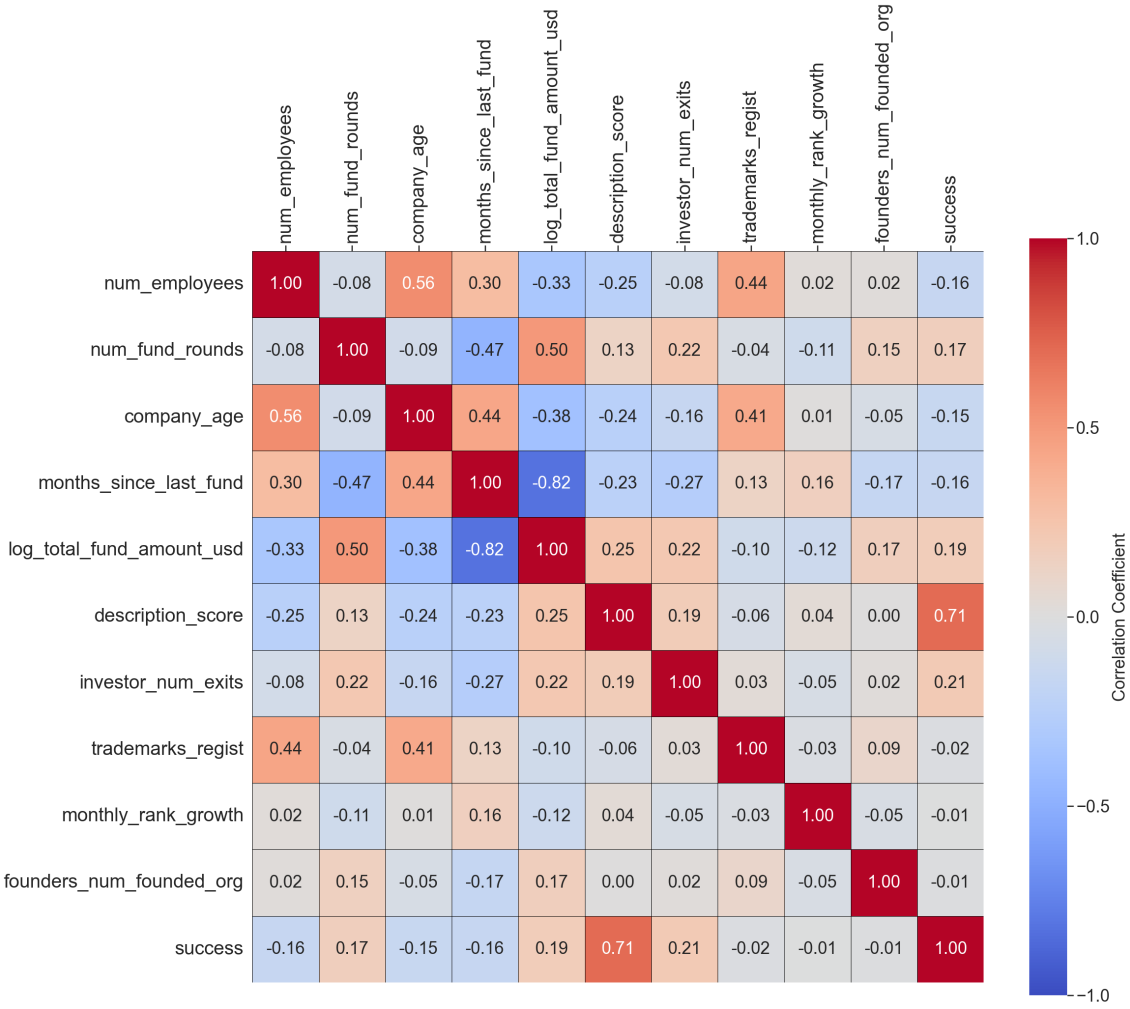


Figure B.2 - Updated features' importance of RF model

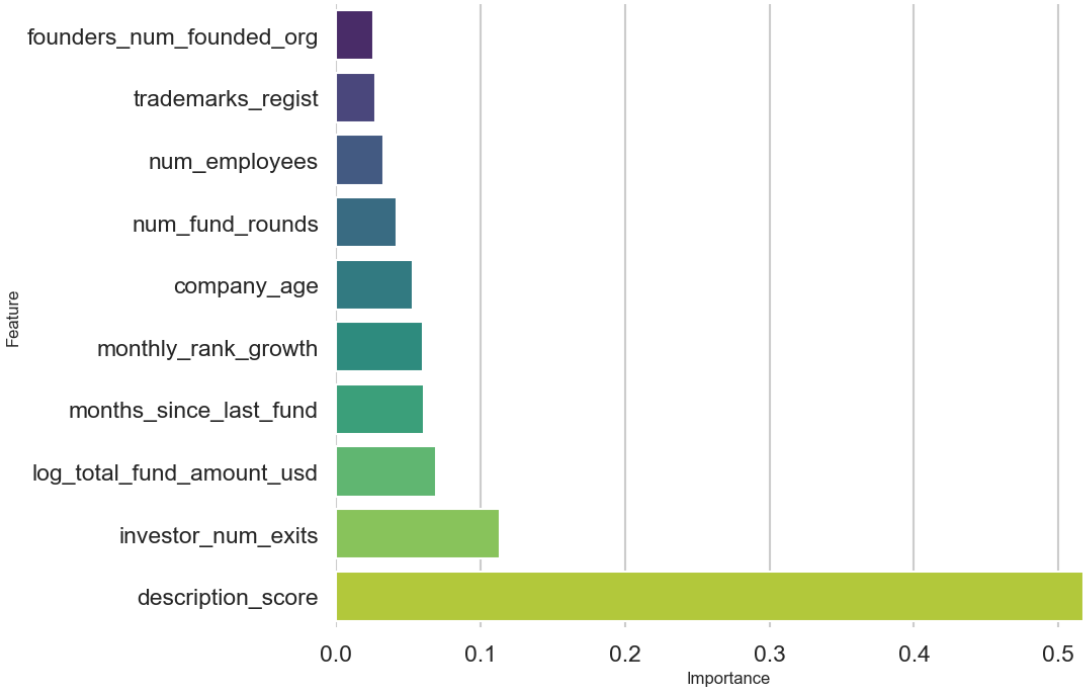


Table B.1 - Updated features' importance values

Variable	Importance
description_score	0,51698
investor_num_exits	0,11335
log_total_fund_amount_usd	0,06917
months_since_last_fund	0,06032
monthly_rank_growth	0,06015
company_age	0,05275
num_fund_rounds	0,04194
num_employees	0,03240
trademarks_regist	0,02735
founders_num_founded_org	0,02558

C. Startup Screening Tool

Figure C.1 - Startup Screening Tool platform