Startup Success Prediction

In the Dutch Startup Ecosystem

by

Diego Camelo Martinez

in partial fulfillment of the requirements for the degree of

Master of Science
in Management of Technology

at the Delft University of Technology,
to be defended publicly on October 30, 2019 at 10:30 am.

Student number:	4735560		
Project duration:	April, 2019 – September, 2019		
Thesis committee:	Prof. dr. C. P. Cees van Beers,	Chairman	TU Delft
	Dr. ing. V. E. Scholten,	Supervisor	TU Delft
	Dennis Huisman	External Supervisor	Startup Delta

TU Delft Delft University of Technology

"The secret to successful hiring is this:."
look for the people who want to change the world."
- Marc Benioff, Salesforce CEO

Foreword

With this thesis I culminate a two year journey at TU Delft and I earn my MSc diploma in *management of technology* (MOT). Through this enriching experience, I have been able to complement my engineering perspective, by understanding the broader picture in which technological development is embedded in. I understood that technological development does not occur in isolation, it involves the participation of multiple stakeholders that collaborate to attain real social impact. Thanks to the MOT programmed I have been able to complement my strong technical background by enhancing my managerial and human capabilities.

This master thesis was not an easy journey and I want to thank to all the people that collaborated to make this master thesis possible. Thanks to my supervisors at TU Delft, Victor Scholten and Cees van Beers, and my supervisor at techleap.nl, Dennis Huisman, for their feedback and support. Special thanks to all the startup founders and interviewees that participated in this research for dedicating part of your time to respond to my questionnaire and participate in my interviews. Lastly, thanks to my fellow colleagues at *techleap* and other friends who helped me reviewing my thesis and constantly giving me support.

In this research I focused on the startup ecosystem of the Netherlands. I worked on this topic as I am passionate about entrepreneurship & innovation. I intend in the future, to apply all the acquired knowledge and experience to strengthen and enrich the innovation system of my home country Colombia. My journey at TU Delft, and my internship at *techleap.nl*, have greatly contributed towards the achievement of this personal and professional goals.

I dedicate this thesis to my family, my eternal inspiration to keep growing.

Diego Camelo Martinez
Delft, the Netherlands, October 2019

"I knew that if I failed I wouldn't regret that,
but I knew the one thing I might regret is not trying."
- Jeff Bezos, Amazon Founder and CEO

Abstract

What makes a startup successful? How to define success? Existent models on startup success prediction often left aside significant predictors which may not be available in typical business databases such as *crunchbase.com*. In this research, focused on a population of five thousand organisations from the Dutch startup ecosystem, we go beyond previous approaches and deliver predictive models that include novel and distinctive variables. To achieve this goal, we depart from an extensive selection of variables drawn from the literature review. The initial selection is discussed, refined and enriched by carrying out interviews with knowledgeable actors in the ecosystem. At the end of the study, a total of eight significant predictors are used to construct three predictive models on startup success. The first model predicts a startup having total funding of one million euros or above, the second model predicts a startup having ten or more employees, and the third model predicts a startup having an average annualized return of at least 20% in the past three years. After testing the models, accuracies of 71%, 71% and 76% respectively are obtained. The results of this research are meant to be used by the organisation techleap.nl. By enriching the data, employing more sophisticated ML models and conducting this research at different points of time, techleap.nl will be capable of monitoring and predicting the performance of the ecosystem both accurately and dynamically.

Keywords

Startup Success Prediction, Predictive Modeling, Logistic Regression.

Executive Summary

The purpose of this executive summary is to provide the reader with an overview of the entire content of this document. More importantly, to state the usefulness of this research and illustrate its implementation in a real context.

Startup and entrepreneurial ecosystems have become an essential element of innovation systems and economies worldwide. Startup ecosystems are rapidly growing, and it is vital to monitor their performance and drive their growth. To monitor startups' performance, it is important to analyse what makes a startup successful, and how to define its success. Existing models on startup success prediction often exclude significant predictors which may not be available in typical business databases such as *crunchbase.com*. In this research, we go beyond previous approaches and deliver three predictive models on startup success that include novel and distinctive success factors. To achieve this goal, we work with the database of *techleap.nl*, a non-profit organisation that exists to strengthen, connect and grow the thriving and competitive ecosystem of the Netherlands. Using these data, we build three predictive models from a sample (*seventy-three entities*) of the population (*approximately five thousand organisations*) of startups & scaleups in the Dutch startup ecosystem.

What makes a startup successful? How to define success? To solve this question, we perform three main activities: a literature review, a discussion with knowledgeable actors in the ecosystem, and an exploratory analysis of the data. All three steps are conducted sequentially to deliver a comprehensive and distinctive set of variables, which are later used to construct three predictive models on startup success. From the literature review, we obtained an initial selection of forty-three predictors and eight criteria of startup success. This selection is discussed and refined through unstructured interviews with actors in the Dutch startup ecosystem. After the discussions, a total of twenty-two variables are removed, and sixteen new variables are added. A reduced selection of thirty-seven success factors and four success criteria is obtained. Lastly, data is collected through questionnaires and filtered once more after exploring the data. A final selection of twenty-eight independent variables and three dependent variables are explored for the construction of the three predictive models on startup success. In the end, a total of eight predictors are used to predict the outcomes of three dependent variables. The three predictive models built are displayed in figure 1 and discussed in the next paragraph.

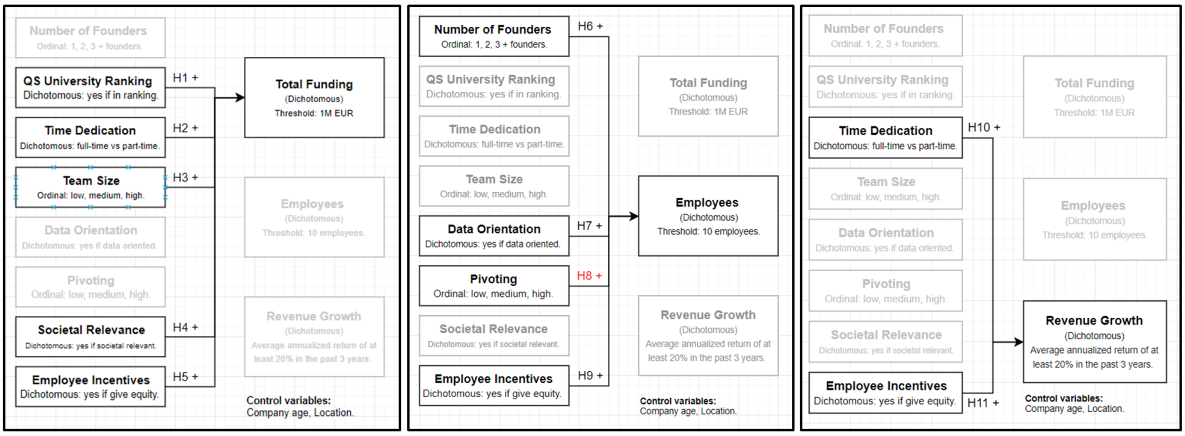

Figure 1: Predictive models on startup success.

Based on logistic regression, the three models predict the values of the three dependent variables going above a certain threshold. For the first model, five variables (*university ranking, time dedication, team* *size*, *societal relevance* and *employee incentives*) proved to be significant predictors of a startup reaching total funding above one million euros. For the second model, four variables (*number of founders*, *data orientation*, *number of pivots* and *employee incentives*) proved to be significant predictors of a startup having ten or more employees. For the third model, two variables (*time dedication* and *employee incentives*) proved to be significant predictors of a startup having an average annual growth higher than twenty per cent per annum over three years (definition of a scaleup). It is important to notice that we measured variables in different ways, and some of them were easier to evaluate than others.

All the organisations employed to build the model are headquartered in the Netherlands. These companies were founded between 2009 and 2017 and can all be classified as startups according to qualitative definitions provided in this research. Furthermore, it is important to highlight that these models evaluate the characteristics of startups & scaleups in their very early stages and intend to predict outcomes typical of the scaling stage in the startup life cycle. Typical benchmarks that occur at this point in the life cycle include having more than ten employees, reaching total funding above one million euros, achieving a series A round of investment, among others. The three delivered models are tested in a new dataset (*seventeen entities*). The characteristics of the three models are displayed in table 1.

	Characteristic	**Total Funding**	**Number of Employees**	**Revenue Growth**
Train (n=74)	*Number of Predictors*	5	4	2
	Number of Occurences	19	37	22
	McFadden R2	0.43	0.31	0.13
	AUC	90.4%	84.8%	73.2%
Test (n=17)	*Accuracy*	0.71	0.71	0.76
	Sensitivity	0.17	0.75	0.63
	Specificity	1.00	0.6	0.89

Table 1: Performance for the three predictive models on startup success.

Although all three models perform reasonably at making general predictions (as determined by the accuracy), they do not come without limitations. The model on *total funding* performs poorly at predicting positive outcomes (sensitivity of 0.17). The model on *number of employees* has fair values of accuracy, sensitivity and specificity, but some of its predictors are somewhat questionable (e.g. *number of pivots*). Lastly, the model based on *revenue growth* is superior at making predictions, its predictors are straightforward to evaluate, but with only two dichotomous predictors the model is limited by its simplicity.

How can the results of this study be suited for real-life applications? Primarily, this research serves techleap.nl's purpose of strengthening the startup ecosystem in the Netherlands. To use this thesis to attain this specific goal, we recommend taking the following actions:

1. **Longitudinal Study:**. Select a largely-enough sample of early-stage startups to monitor their performance through time. Provide them with an incentive so that they are willing to participate in the study and report their data.

2. **Enrich the data**. Explore new variables to be included in the models, do this by carrying out brainstorming sessions with stakeholders and partners. Enhance the evaluation of the success factors, come up with standardised methodologies to evaluate difficult variables such as *data orientation*. Besides collecting data directly from startups, integrate external databases that include valuable information such as financial KPIs.

3. **Success prediction**. Implement more sophisticated *machine learning* methods to enhance the predictions on startup success.

Although this research is focused on techleap.nl's organisational goals, the insights obtained and recommendations presented can be valuable for other stakeholders such as venture capitals, startups and accelerators.

Contents

LIST OF FIGURES

List of Tables

"If you are not embarrassed by the first version of your product, you've launched too late."

- Reid Hoffman, LinkedIn Co-Founder

1

Introduction

1.1. Initial Remarks

The purpose of this section is to provide the reader with information so that he or she can navigate comfortably throughout the contents of this document. In subsection 1.1.1, we provide an overview of the structure of this document. In subsection 1.1.2, key concepts are introduced, and we explain how these terms are used in the report.

1.1.1. About this Document

This document was written using LaTeX document preparation system. The design of this document is based on a TUDelft template and adapted to the contents of this research. Throughout the text the reader will encounter multiple hyperlinks (highlighted in light blue), these connect to other chapter/sections of this thesis, the glossary, the acronym lists or the references. Keywords, other concepts and acronyms are highlighted at their first appearance, a link to the glossary or the acronym list is provided for clarification of these concepts.

This document is divided into six chapters. In chapter one, Introduction, the motivation, relevance, objectives and the general planning of this research are presented. In chapter two, Startup Theory, essential concepts to understand this research are thoroughly discussed. In chapter three, Predictors and Criteria, we deliver a set of independent and dependent variables that resulted from the literature review and discussions with knowledgeable actors of the Dutch startup ecosystem. In chapter four, Predictive Model, all the process carried out to build the startup success predictive model, from data collection to model testing, is carefully explained. In chapter five, Discussion, all the sub research questions are responded by discussing the results achieved in this study. Lastly, in chapter six, Final Remarks, conclusions and recommendations for further improvements of this research are given.

1.1.2. Concepts

The purpose of this subsection is to clarify how key concepts are employed throughout the text. These concepts are further discussed in chapter 2.

- **Startup:** in this thesis, *startups* and *scaleups* are sometimes mentioned indifferently. Both startups and scaleups are subjects of study in this research. To avoid excessive repetition, these may also be referred to as *companies*, organisations, entities, subject of study, and so forth.

- **Predictor:** the independent variables of this research are primarily discussed as *predictors*. To avoid repetition, these may also be referred to as *factors* or simply *independent variables*.

- **Criteria:** the dependent variables of this research are primarily discussed as *criteria*. To avoid repetition, these may also be referred to as *measures of success*, *success metrics*, or simply *dependent variables*.

- **Success:** in this research, success is defined as a dichotomous variable. A startup is classified as successful or not successful depending on the dependent variable used.

1.2. MOTIVATION

All through this thesis, the concept of startups is frequently discussed. Most readers might be familiar with the term. Perhaps, some may even be able to provide a proper definition. Although not always the case, startups are often associated with high-tech, and high-tech is followed by an extensive collection of buzz words and hyped technologies. Why does this matter? Startups are on hype, and the definition of the concept is always subject to interpretation. What is the difference between startups and SMEs? What is the relationship between startups and entrepreneurship? When is a startup consider a scaleup? Delimitation remains fuzzy. A thorough understanding of startups is vital for the development of this project. A proper definition allows us to construct a consistent set of data and to exhaustively understand the life cycle of startups, from formation to maturity.

Although the meaning is malleable, most would agree on one thing: startups are designed for rapid growth. Startups aim to cause big bang disruption, to create radical innovation [Groenewegen and Langen, 2012], and to cause meaningful changes in the economy and society. How startups cause these meaningful changes? For starters, although startups represent a small portion of small businesses, when controlling for firm age, they make a disproportionately larger contribution to job creation and sales growth [Ayyagari et al., 2011]. "Firm startups account for only 3% of employment but almost 20 per cent of gross job creation" [Haltiwanger and Jarmin, 2010].

Startups are also drivers of innovation. Austrian economist Joseph Schumpeter stated that "innovation is the centre of economic change causing gales of creative destruction" [Schumpeter, 1942]. He defined innovations as inventions turned into useful business ideas and entrepreneurs as the actors responsible for this transformation. In Schumpeter's later publications, entrepreneurship is seen as an institutional effort; he even stated that "the country itself, or its agenda, can act as an entrepreneur" [Sledzik, 2013]. In this thesis we adopt this collectivist approach: innovations do not occur in isolation; startups need the collaboration of multiple entities to turn inventions into successful business ideas. Startups are embedded in an ecosystem that works to advance the economy through innovation, the startup ecosystem. This ecosystem is formed primarily by startups, government, academia, investors and corporations. This thesis is done conjointly with techleap.nl, a non-profit organisation that exists to strengthen, connect and grow the thriving and competitive startup ecosystem of the Netherlands.

As discussed before, startups aim to be disrupters. Undoubtedly being so is not an easy task. The success rate of startups is reduced and investing in them is accompanied by significant risk. The definition of success is relative. Success can be measured in many ways and different stakeholders are interested in different metrics. Success can occur at different levels, from simply surviving to being a unicorn. By exploring the database of dealroom.co (strategic partner of techleap.nl), with an estimate of 450 thousand global startups & scaleups, we can easily determine the success rate for different criteria. From the population of this database, less than 0.25% of the companies are unicorns, 0.45% have reached an IPO and 8% have been acquired. Startups can also be classified as successful depending on their funding stage (10% of companies with funding over one billion euros), or by the company size (40% with more than ten employees, 16% with more than fifty, and 7% with more than one hundred). Success rates can greatly vary, located in a spectrum depending on the metric used. As for this research, where data is collected from a limited sample (around one hundred companies), predicting rare outcomes such as an IPO or unicorns is nearly impossible. Success is herewith defined in the more conservative side of the spectrum, where success rates are in the range of 10% to 50%.

How to evaluate the worthiness of a startup? World-class, radical innovators, game-changers, and so forth. Hackneyed words that may sound appealing to the common, not to the experienced investors. For them, predicting startup success is a science; some rely on numbers to assess worthiness; some judge from more qualitative attributes. Most of the quantitative studies that predict startup's success rely on hard data that is available in business databases. In this study we go beyond those approaches by exploring a broader selection of variables. A total of thirty-seven variables obtained from thorough desk research and discussions with knowledgeable actors are explored as candidates of startup success. With the variables, three models were built to predict the outcomes of three dependent variables (*total funding, number of employees, revenue growth*). At the end of this document we provide techleap.nl with a series of recommendations on how to implement the insights of this study to attain their organisational goal of strengthening the startup ecosystem in the Netherlands.

1.3. RESEARCH OBJECTIVE

The purpose of this research is to build a predictive model (or models) for the success of startups in the Netherlands. At the end of this document, in section 6.2, we provide techleap.nl with a series of recommendations on how to use the insights of this study to strengthen the startup ecosystem. In the next subsection, the research questions are introduced; these questions help make the objective of this research more tangible. We respond to all sub-questions in the discussion 5 chapter and the main research question in the conclusion 6.1 section. In subsection 1.3.2, the research scope of this study is made clear, by clearly defining the subject of study, and the boundaries that frame this research (e.g. geographical boundary).

1.3.1. RESEARCH QUESTIONS

The research questions listed below are answered sequentially throughout this research. Sub questions one to five (henceforth SQs) build the foundation to answer the main research question (henceforth MQ). SQ1 is responded in section 3.1, here an initial set of independent variables is delivered. SQ2 is discussed in chapter 3.2 and a new modified selection of variables is displayed in section 3.3. SQ3 and SQ4 are responded in section 4.6, where three predictive models based on logistic regression are constructed to predict success determined by total funding, number of employees and revenue growth. SQ5 is answered in section 4.7, where the three mentioned models are tested on a different set of startups. For more information on how the research questions are related to the research design, please refer to the flow diagram in subsection 1.5.3.

- **SQ1:** What is a comprehensive selection of predictors and criteria for startup success, as mentioned in the literature?
- **SQ2:** How the selection of predictors and criteria can be improved according to knowledgeable actors of the Dutch startup ecosystem?
- **SQ3:** Which of the selected predictors proved to be significant at predicting startup success?
- **SQ4:** What is the relationship between the significant predictors and the criteria for startup success?
- **SQ5:** How does the relationship between predictors and criteria perform at predicting startup success?
- **MQ:** How startups' success can be predicted from a comprehensive selection of predictors and criteria in the Dutch startup ecosystem?

1.3.2. RESEARCH SCOPE

This research is focused on the startup of the Dutch Startup Ecosystem. To the date 3 May 2019, using techleap.nl database, after applying the definitions explained in subsection 2.2, a population of around five thousand entities is estimated. For a good representation of the population, with a 95% confidence level and 10% margin of error, the sample size must be around one hundred entities. As data is collected through questionnaires, and responses are typically low (around 10%), an approximate of one thousand startups should be reached. Startups in the sample pertain to diverse industries and regions in the Netherlands. These are founded between 2009 and 2017. For more details about this process please refer to section 4.1. Although the database is supposed to be composed only by startups (and not SMEs), we further filter the selection by including a question in the questionnaire delivered to the startups in which they confirm (or not) their startup categorisation based on a qualitative definition provided by the researcher.

1.4. RELEVANCE

In this section, we discuss the relevance of this study from a societal, scientifical and organisational perspective. First, we discuss the societal relevance from the perspective of three main actors in the Dutch Startup Ecosystem: investors, startups and the government. Second, the organisational relevance is made clear, explaining how the results of this research are aligned to the organisational goals of techleap.nl (main collaborator in this research). Lastly, we describe how this research contributes to the scientific & academic community by addressing the knowledge gap found in previous research.

1.4.1. SOCIETAL

Although the main goal of this thesis is to contribute to techleap.nl's organisational goals, different stakeholders in the ecosystem may significantly benefit from this research in different ways. In this subsection we particularly discuss the relevance of this study for startups, the government and investors.

STARTUPS

As subjects of study of this research, startups benefit from this study in various ways. Most importantly, as participants of this study, some startup founders will receive a summarized report of the results. Startup founders can learn which are the presumed key factors for startup success as mentioned by the literature and knowledgeable actors in the ecosystem (e.g. researchers, investors, startup founders, accelerators, and so forth). They can also learn which are the factors that result to be the most significant as determined by the predictive model. With this insights, startup founders will be able to observe the relations between the success factors and criteria and solve questions such as: how many founders are too many founders? Is developing a business plan critical for the success of a company? Should I launch my startup now or should I acquire more hands-on experience? This study focuses on factors at the very early stages. Results can be taken in retrospect by existent startups or by new startups to improve their decision making.

GOVERNMENT

Government is a key actor as it regulates the ecosystem at the policy level. While other actors in the ecosystem are focused on their individual benefits, the government works for the well being of the society at large. The startup ecosystem is just one item in the government's agenda. The government's role is vital for the proper development of any economy, safeguarding the interest of the citizens. The model developed in this thesis may be useful to identify gaps which may require potential actions by the policy-makers (e.g. a region showing poor performance, inclusiveness problems, sustainability, and so forth).

INVESTORS

Perhaps, the stakeholders that benefit the most from this research are investors in the Dutch startup ecosystem. From the results of the predictive models, investors can observe which variables are significant predictors of success and which are not. With this information investors can improve their decision making when evaluating possible investments. Furthermore, from the results of the qualitative research, investors can observe which other success factors are discussed by both the academia and knowledgeable actors in the ecosystem.

1.4.2. ORGANIZATIONAL RELEVANCE

In this organisational relevance, we want to discuss how techleap.nl can benefit from this research. Most importantly, this research contributes to techleap.nl's goal of strengthening the Dutch startup ecosystem. Using this study as a departing point, techleap.nl can enrich and improve the obtained models to quantitatively monitor, predict and steer the success of startups in the ecosystem. In section 6.2 we provide techleap'nl with a series of recommendations of how this objective can be achieved. Besides this primary reason, the following secondary benefits where also identified:

- **Better understanding:** from the desk research and the interviews, techleap.nl can expand their understanding of how different factors and criteria for startup success are perceived by knowledgeable actors. Through the predictive model, which determines critical factors of startup success, techleap.nl can better evaluate the potential of startups. This task is critical as techleap needs to frequently select top-performing startups for their various projects. Results can also be used to help startups in their decision making throughout their life cycle and advising the government in their role as policymakers.
- **Reporting:** publishable reports can be extracted from the different chapter of this thesis. First, from chapter 2, where we thoroughly explain the startup concept, an introductory conceptual report can be made for readers that are not familiar with the topic. This knowledge is typically tacit and acquired through experience in the startup-world. A simplified report with all relevant concepts may be valuable for the unacquainted reader. Second, insights from the desk research and conducted interviews in chapter 3 can be reported. These insights can be very valuable to understand how factors and criteria of success are noted by researchers and knowledgeable individuals. Lastly, results from the predictive model in chapter 4 can be published, so that the ecosystem at large may be able to observe and even interact with the data, to understand how critical success factors (e.g. the number of founders) influence the dependent variables (e.g. total funding above one million euros).

- **Database:** one of the goals of techleap.nl is to make the ecosystem visible through data. Currently data is exported from dealroom.com (strategic partner), but techleap.nl intends to build a more extensive and richer database in the coming years. Throughout this master thesis, we continuously work towards this goal. First, data were previously analysed through independent tables; these are now connected in the structure of a proper database. Second, the contact database was significantly enriched, collecting hundreds of emails from startup founders. Lastly, the results of the desk research, the interviews and the predictive model; help to determine which new data should be collected when constructing the new database.

1.4.3. SCIENTIFIC RELEVANCE

The scientific relevance of this research becomes evident after reading section 3.1. In this section, thorough desk research on predictive models on organisational success is conducted. Two main observations are made:

- **Predictors:** most predictive models on startup success rely on financial information, web metrics and basic demographics collected from business databases such as Crunchbase. This selection of predictors may be limited, leaving critical success factors out of scope.

- **Criteria:** different definitions of startup success are employed across different sources. Although it is clear that success metrics can vary depending on various factors (e.g. size of the sample), most studies do not evaluate or discuss the different alternatives. Success may vary from company survival to acquiring the status of unicorn. Success also varies depending on the type of variable (e.g. dichotomous, categorical, ordinal).

A model is built to represent something. In this case, the predictive model represents the relationship between startup success factors and criteria. A model should not only be accurate, but it should also be useful. When reviewing the literature, we noticed that predictive models on organisational success tend to focus more on the model than on the selection of the variables. This lack of reflection the success factors and criteria, can lead to models with high accuracies of poorly understood systems. It can be particularly noticed that most predictive models do not include a theoretical framework. Moreover, intangible factors (e.g. culture, motivation, team quality), which may be significant predictors, are rarely considered, and models are built with easy-to-use data.

This research aims to go one step further, reducing the existent knowledge gap. In chapter 3, a comprehensive set of variables is delivered after conducting desk research and carrying out interviews with knowledgeable actors in the ecosystem. By doing this, we aim to construct a distinctive model (or models) that predict startup success with variables that have not been yet included in previous studies. Although many critical factors are hardly tangible, this research may discover novel predictors that can be easily measured.

1.5. RESEARCH DESIGN

The nature of this research is primarily quantitative. At the end of this study, we deliver three predictive models based on logistic regression. One model is built to predict a startup in the sample having total funding above one million euros, the second model predicts a startup having ten or more employees, and the third model predicts a startup having an accelerated revenue growth as defined by the scaleup concept.

Nevertheless, this thesis is not exclusively quantitative. The selection of startup success predictors and criteria is delivered after conducting thorough desk research and carrying out interviews with knowledgeable actors in the Dutch startup ecosystem. In this section, we briefly present what is done in the three main activities of this thesis: desk research, interviews, and predictive modelling.

1.5.1. Qualitative

Qualitative research in this thesis consists of two main tasks: desk research, where the literature is reviewed to deliver an initial selection variables, and interviews, that are carried out to refine the previous selection.

Desk Research

In the desk research, carried out in chapter 3, literature related to the keywords of *startup success prediction* was reviewed. Out of a total of twelve consulted articles, six were read in their entirety, and four were used to compile an initial selection of success predictors. The entire set of twelve articles is skimmed to review the different success criteria employed, seven candidates for dependent variables are listed. The initial list of independent variables consists of forty-three candidate predictors grouped in the categories of *founders*, *business*, *innovation*, *actions & decisions*, *resources*, *environment* and *third-party support*. To deliver the selection of predictors, all candidate variables are coded in the four documents using atlas.ti software for qualitative analysis and a code-document table is created. Due to the limited number of codes in atlas.ti free version, variables are coded once per document, and no frequency is displayed. This initial selection of variables is discussed with knowledgeable individuals and further refined.

Interviews

The interviews process is described in section 3.2. The purpose of the interviews is to refine the selection of variables obtained from the desk research. The selection of predictors and criteria for startup success is reviewed with knowledgeable actors of the Dutch startup ecosystem. A total of thirteen persons were reached, and seven were interviewed. To avoid bias, an heterogeneous sample of interviewees is selected, including individuals with different backgrounds, experience levels and roles in the ecosystem. Interviews are conducted face-to-face, in an unstructured manner, and with no predefined questions given. All interviews were conducted in a similar fashion: interviewer and interviewee introduce themselves to each other, the interviewee is asked to provide his or her critical success factors and criteria, and the pre-selected list of variables from the desk research is discussed. After reviewing the list of variables with the interviewees, a total of thirty-seven variables and four dependent variables are selected for further analysis.

1.5.2. Quantitative

The quantitative study is divided into six sections. First, in section 4.1 Sampling, data set from techleap.nl is filtered and we obtained an estimate population size of 4.5 thousand startups & scaleups. Taking into account the size of the population, we calculated that the required sample size was at least one hundred entities. Second, in section 4.2 Data Collection, data is collected through a questionnaire delivered to 990 startup founders from 820 startups & scaleups. Third, in section 4.3 Data Preparation, data is thoroughly prepared. Data preparation includes many critical steps including *data integration* (merging data sources), *data enrichment*, *data cleaning* (removing outliers, dealing with blank spaces) and *data transformation*. Fourth, after data prepration, in section 4.4 Variable Selection 3, we deliver a third and final selection of variables. This final selection consists of twenty-eight independent variables (predictors) and three dependent variables (criteria). Fifth, in section 4.5 Logistic Regression, a brief overview of logistic regression is provided, the machine learning technique used to build the models. Sixth, in section 4.6 Model Construction, we build three models that predict startup success. One model to predict a startup having a total funding above one million euros, a second model to predict a startup having ten or more employees, and a third and last model to predict a startup achieving a revenue growth as defined by the scaleup definition. Lastly, in section 4.7 Model Testing the three proposed models are tested on new data to evaluate their performance.

Even though this research is primarily quantitative, its core lies in the combination of the results obtained from the qualitative (comprehensive selection of variables) and the quantitative research (predictive model based on logistic regression). It is not the goal of this thesis to conduct extensive qualitative research nor to employ sophisticated machine learning models to predict startup success. The goal of this thesis is to employ a comprehensive selection of variables to create a simple model on startup success, hence the use of logistic regression. Why logistic regression? Despite this technique being limited in it's predictive capabilities, logistic regression is perhaps the simplest and easiest to understand of all machine learning methods. For now, this research focuses on the essentials, exploring the critical factors and criteria of startup success, and adequately understanding the phenomena behind their relationship. For a better understanding on how this thesis was conducted, the research flow diagram is provided in the next subsection.

1.5.3. FLOW DIAGRAM

In figure 1.1 flow diagram for this research is displayed. As it can be seen in the legend, the quantitative research steps are coded in yellow whereas the qualitative research steps are coded in light blue. Tools, techniques and other details are also stated. Every box in the diagram is related to a chapter or section from the table of contents. Not all sections are represented in the flow diagram, only the critical ones. As it is was mentioned before, SQ1 is answered in section 3.1, SQ2 in section 3.2, SQ3-Q4 in subsection 4.6 and SQ5 in section 4.7. The main research question is discussed in chapter 6.

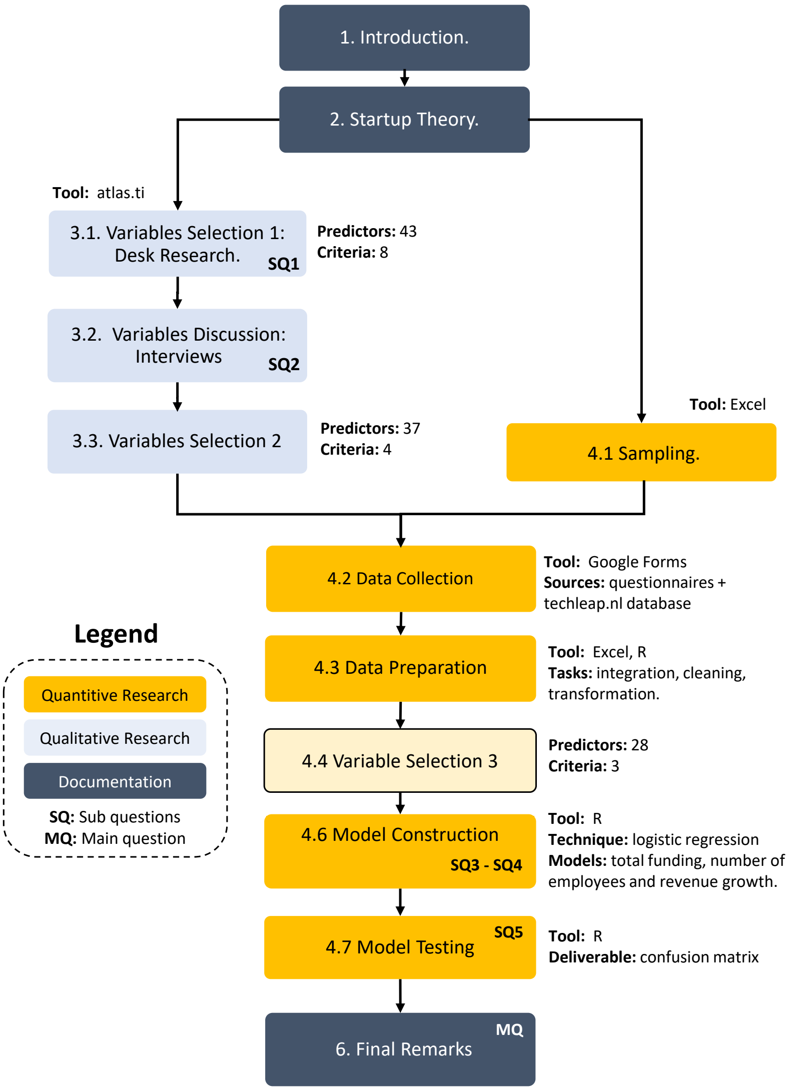

Figure 1.1: Research Flow Diagram

From all the presented tasks the ones that results the most critical are the discussions with knowledgeable actors in the Dutch startup ecosystem, data collection from startup founders through questionnaires and the data preparation. The interviews require some time because it involves reaching out to people who are most likely busy. The collection of data through questionnaires requires significant time because of two main reasons: (1) a large number of emails need to be collected, (2) achieving the desired number of responses is challenging and requires significant effort. Lastly, data needs to be carefully prepared to be included in the predictive model, although this task may be demanding it does not depend on external individuals so the risk of not completing it on time is much lower. In the next section, the general planning for this thesis is provided by means of a Gantt diagram.

1.6. SCHEDULING

Because of the combined nature of this research, including both qualitative and quantitative methods, the time scheduling of this project must be carefully planned. Using free online platform teamgantt.com we are able to organize the scheduling of our project as it is shown in figure 1.2.

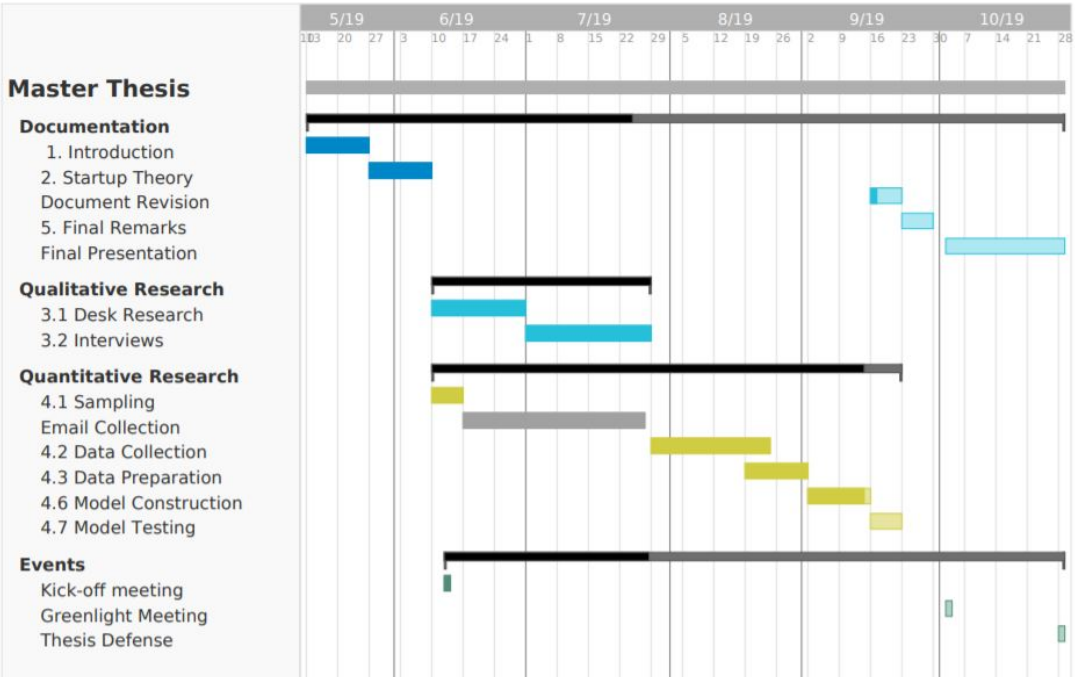

Figure 1.2: Thesis Gantt Diagram.

The schedule shown above is organized in a similar manner to the flow diagram presented in the previous section. The tasks are divided into documentation, qualitative research and quantitative modelling and the colour classification match that of the flow diagram. Every task is listed and matches the chapters and sections of this document. The whole thesis is completed within a time frame of twenty-four weeks, from proposal to defence. The document is divided into five chapters. The first two chapters form an extended version of the thesis proposal. In the second chapter, desk research and interviews are conducted to deliver a comprehensive selection of independent and dependent variables for startup success in the Dutch startup ecosystem. In chapter four, three predictive models on total funding, number of employees and revenue growth as dichotomous variables are built and tested. Lastly, in chapter five, the main research question is responded in the conclusion and recommendations for further improvements of this research are provided. The main bottlenecks in the development of this project are the administration of interviews and questionnaires to knowledgeable actors and startup founders respectively. Because of these bottlenecks, time must be managed carefully to finish this thesis before the established time frame.

The first stage of this research, which involves writing chapter 1 (Introduction) and 2 (Startup Theory) of this document is done in a time frame of four weeks. Following this, the qualitative research in chapter 3 (Startup Success: Predictors and Criteria) is carried out in seven weeks. In this chapter, the interviews demand time and effort as these tasks involve reaching out to individuals that are often busy; some may not be even willing to participate. In parallel to the qualitative research, hundreds of emails are collected to prepare for the administration of the questionnaires. Once the qualitative research is finalized, and the list of candidate variables is defined, the questionnaire is designed and administered to collect data from startup founders, this process takes place in a time frame of four weeks, a low response rate is expected. Next, as it can be seen from the diagram, the quantitative analysis (from data preparation to model testing) is carried out only until the last five weeks before the green light meeting. Lastly, once the model is tested the document will be revised in its entirety, the fifth chapter (that includes conclusions and recommendations) is written, and the presentation for the defence is prepared.

2

Startup Theory

In this section, we provide a detailed description of the key concepts of this thesis. The literature review serves as a foundation for selection of variables in chapter 3. This chapter is divided into two main sections. The first section is an overview of the *entrepreneurial theory*, and the second section is an overview of the concept of *startups*.

Search Terms: keywords stated at the bottom of the abstract have many related concepts. For example, startup ecosystems may be considered as a subset of the entrepreneurial ecosystem and of the innovation system. Some of the search terms used for this chapter include *entrepreneurial theory*, *startup theory*, *startup's life cycle*, *startup's financing*, *startup ecosystem*, *dutch startup ecosystem*. Besides these concepts, which represent broad topics, other specific terms were searched separately, definition for concepts that are related but not core to this research are provided in the glossary (e.g. *triple helix model*).

Bibliography Manager: EndNote and BibTeX. EndNote was used as the primary reference manager, BibTex was used to create the connection to the LaTeX editor. There is a total of 9 sources used for section 2.1 and 30 for section 2.2.

Sources: Information in this chapter was collected from books, journals and grey literature. Papers published in journals were found by using one of the following databases and search engines: Google Scholar, ScienceDirect, Scopus, Web of Science and TU Delft research repository. Grey literature such as reports, theses and articles were obtained using Google or TU Delft's Thesis education repository (for theses). For this section a total of 38 sources were consulted, in table 2.1 the distribution of references by type and sections of this chapter is shown in table 2.1.

Table 2.1: Literature Review Sources.

	Entrepreneurship Theory 2.1	**Startup Theory** 2.2
Journals	4	7
Books	5	2
Reports	-	13
Theses	-	7
Web Pages	-	6

2.1. Entrepreneurial Theory

n this section, we briefly discuss the concept of *entrepreneurship*. It is essential to introduce this concept as it provides the foundations to understand the more contemporary concept of *startup*. Entrepreneurship is a concept that has been extensively studied, the use of the concept can be found as early as the 18th century when Richard Castillon defined an entrepreneur as "a person that does not retreat from engaging in risky business ventures" [Sledzik, 2013]. For long time there was no research to be done regarding entrepreneurship, it was thought to be a skilled solely acquired through hand-on experience [Kuratko, 2017]. Nowadays there is much research on the field, thanks to pioneer scholars in the topic, we now celebrate the immense growth in entrepreneurship research Kuratko [2017].

Many definitions are provided for the concept of *entrepreneurship*. Although, the concept of entrepreneurship was introduced long ago before the neoclassical economic era, the Austrian economist Joseph Schumpeter, is considered by many as the earliest scholar on the topic [Bull and Willard, 1993] [Cuervo Garcia et al., 2007] [Sledzik, 2013]. Before Schumpeter, entrepreneur was simply "the organizer and manager of production or trade" [Sledzik, 2013]. Schumpeter's view on entrepreneurship and innovation can be seen throughout his life-time works [Schumpeter, 1934] [1942] [1947]. In his early works, also known as Mark I theory, entrepreneurship was defined from the perspective of the individual, the following definition is provided: "the function of entrepreneurs is to reform or revolutionize the pattern of production by exploiting an invention, or more generally, an untried technological possibility for producing a new commodity or producing an old one in a new way by opening a new source of supply materials or a new outlet for products, by reorganizing an industry and so on". An innovation is often defined as a commercialized invention, an invention alone has no economic value and the entrepreneur is the actor in charge of giving an invention this value [Lundstrom and Stevenson, 2005]. In the later works of Schumpeter [Schumpeter, 1942], also known as Mark II theory, Schumpeter stated that it was not the efforts of the individual entrepreneur but those of the large corporations through research and development that mattered to drive innovation in the economy. Both Mark I and Mark II theory are important to acknowledge for our study, for our predictive analysis we take into consideration both the characteristics of the entrepreneur and the organization as drivers of innovation.

Similarly to startups, there is confusion between the concepts of entrepreneurship and *Small & Medium Enterprises* (henceforth SMEs). As written in [Lundstrom and Stevenson, 2005], researchers continuously refer to the concepts of self-employed, small business owner/manager, and entrepreneur interchangeably. As described in the previous paragraphs, entrepreneurs drive innovation in the economy, and both, individuals and organizations can conduct entrepreneurial activities. As mentioned in [Cuervo Garcia et al., 2007], the critical factor that distinguishes between entrepreneurs and SMEs owners is innovation. We conclude that an entrepreneur seek rapid growth and strive to be innovators, we later discuss how this differs from the definition of startups.

In this section we have briefly discussed entrepreneurship from an economic perspective, research and theories on the topic go further than that. In [Bull and Willard, 1993], it is mentioned that literature on the matter can be divided into five broad categories: studies on the definition (that we have already discussed), on the psychological traits of entrepreneurs, on the success strategies, on the business development and on the environmental factors. Categories two to five are later discussed when carrying out the desk research for success predictors and criteria in section 3.1. In a similar manner, in article [Kwabena and Simpeh, 2011], entrepreneurship studies are divided into six main theories: economic, psychological, sociological, anthropological, opportunity-based and resource-based. These theories are briefly described next.

- Economical Theory: "explore the economic factors that enhance entrepreneurial behavior", from the perspectives of the classical, neoclassical and Austrian (i.e. Schumpeter) schools of economics.

- Psychological theory: explore entrepreneurship at the level of the individual: one theory states that the entrepreneur have inborn entrepreneurial qualities, another theory defends that success comes not only from the abilities of the entrepreneur but also from the external support and one last theory mentioned proposes that entrepreneurs are driven by the natural human need to succeed, accomplish, excel or achieve.

- Sociological theory: focuses on the social context, social contexts relates to entrepreneurship in four ways: through social networks (the entrepreneur builds strong connections based on trust), the life experiences (that may steer entrepreneur's decisions to do something meaningful), through ethnic identification (e.g. the obstacles individials face due to the social background they are located) and lastly, population ecology, refers to the environmental factors (e.g. political,legislation, customers, employees, competition,etc).

- Anthropological theory: says that for someone to successfully initiate a new venture, the cultural context must be considered.

- Opportunity-based theory: contrary to the Austrian School of Economics, states that entrepreneurship do not cause change but rather exploit opportunities from change (e.g. a new technology, presence of emerging markets,etc).
- Resource-based theory: states that success of new ventures is highly dependent on the access to capital by the founders, capital may be financial, social (i.e. network) or human (valuable depending on education and experience).

All the theories and definitions herewith presented provide this study with valuable insights to understand the concept of entrepreneurship that serves as foundation to later understand the concept of *startups*. As can already be inferred, both concepts significantly overlap and a clear distinction must be provided for this study.

2.2. Startup Theory

This section aims to provide the reader with a proper definition of the startup concept, understand its life cycle and the broader term of the startup ecosystem. Most of the definitions for the hereby presented concepts are not formally stated in any source. Information from both scientific and business sources is combined to provide the definitions in this section. Definitions and ideas stated by some of the most recognized and knowledgeable actors in the field are included. Although these may not be scholars or academics, these individuals are highly recognized by the worldwide startup ecosystem. These individuals are Peter Thiel (co-founder of Paypal), Ben Horowitz (co-founder of VC Andreessen Horowitz), Steve Blank (former advocate of the Lean Startup movement), Paul Graham (founder of Y Combinator), Max Marmer (founder of Startup Genome) and Eric Ries (author of the Lean Startup Book).

2.2.1. Definition

Although there are different definitions that can be assigned to the startup concept, most would agree on one thing: startups are designed for rapid growth. This matches with the definition by Paul Graham who says the only essential thing to define startups is growth, "everything else we associate with startups follows from growth" [Graham, 2012]. Unlike Graham, Peter Tiel thinks that startups, fundamentally, are about creating technological innovation. Tiel explicitly refers to vertical innovation, the development of a new technology that has not been created before unlike horizontal innovation which consists of bringing existing technologies to new places [Thiel, 2014]. Although startups are often associated with technology (like Peter Thiel), this is not always the case. In the early uses of the term (late 1970s), rapid growth was explained by the technological developments of the time and the later burst of the internet (2000s) [StartupCommons, 2019]. Nowadays, being high-tech is not a condition for being categorized as a startup. A startup can be for example an organization developing an advance manufacturing method (technology-based) as well as an e-commerce platform to buy products directly from farmers (technology-enabled). As mentioned by Graham, the only essential thing is growth.

In his book, "the Lean Startup", Eric Ries define a Startup as a "human institution designed to create a new product or service under conditions of extreme uncertainty" [Ries, 2011]. He defines a product as a source of value for the people who become customers. Steve Blank defines a startup as a "temporary organization used to search for a repeatable and scalable business model", this scalable business often have a global ambition, raise capital through angel investors and venture capital funds, and innovate to solve problems. Once a startup finds a repeatable and scalable business model, it ceases to be a startup [Blank, 2013].

According to techleap.nl, main collaborator in this research, a startup in the ecosystem is limited to: not being a service provider, not a subsidiary of a larger enterprise, having more than 1 employee and being less than 20 years old (they do this in order to include some outliers such as Ayden and Swapfiets). Startup Delta does not provide a clear definition in terms of company size by employee count. For this study, startups with solo-founders are not included, nor are startups as old as twenty years old. Boundaries in terms of size by employees and the age of the organisations must be defined.
Regarding company age, an organisation may be defined as a startup if this is younger than ten years old [M.Sc. Steigertahl et al., 2018]. Regarding company size by employees, the OECD [OECD, 2017] provides the following categorization for SMEs: micro (less than ten employees), small (ten to forty-nine) and medium enterprises (fifty to two-hundred-fifty). We limit our study to organizations with less than 250 employees.

As mentioned in section 1.1.2, we sometimes refer in this text to startups and scaleups indifferently. Although it is irrelevant for this research, we may associate scaleups (size-wise) with medium enterprises and startups with small and micro-enterprises. As defined by the OECD, a startup turns into a scaleup according to the definition provided in the glossary. A scaleup successfully turns into a sustainable company when reaching an exit either via a M&A or an IPO. For this research, exited organizations are not taken into account; the main reason to exclude them is that the differentiation between scaleups and corporations is blurry in the database.

What is the distinction between startups and SMEs? According to the previous definitions, a startup is designed for rapid growth [Graham, 2012] and is accompanied by high uncertainty [Ries, 2011]. Contrary to startups, SMEs show slow and steady growth during their life cycle. Startups are often financed by risk-seeking investors such as angels and VCs whereas SMEs are funded by more risk-averse options such as bank loans. SMEs have a 75% if likelihood of survival after two years whereas startups have a much lower 25% [Compass, 2015]. Risk is high but so do returns, your local ice cream shop will probably never scale into a billion-dollar market capitalisation organisation that creates thousands of jobs, a successful startup has higher odds to achieve this.

What is the relation between entrepreneurship and startups? Although a startup founder is also an entrepreneur, the founder of the local ice cream shop is so. Contrary to the early definitions of the entrepreneurship concepts, where innovation seems to be intrinsic to the entrepreneur, nowadays an entrepreneur is any individual who carries out entrepreneurial activities by creating new businesses.

2.2.2. THE RISE OF STARTUPS

Humanity has already passed the tipping point between the industrial and information eras. In the last 50 years, blue-chip companies have lost power, and they became vulnerable to newcomers. In the industrial era blue-chip companies were protected with low levels of competition, information obscurity, and growing consumption. The information age made these barriers obsolete. Today's technologies make it easier to start new businesses and information to be more transparent. Driven by frugality and environmental consciousness, new consumers have also shifted to a sharing or renting economy. Big companies failed to innovate and keep up with the trends, and startups are on the rise. Photography was disrupted by Instagram, book stores by Amazon, music by Apple and Spotify, hotel chains by AirBNB, Taxis by Uber, traditional HR by Linked In, newspapers by social media and retail stores by e-commerce [Compass, 2015].

Although startups may provide big returns, they are still accompanied by significant risk. If a startup has failure rates as high as 90%, why do we see startups everywhere? In his book "The Four Steps of the Epiphany" [Blank, 2013], Steve Blank gives four reasons for this phenomenon: (1) startups can now be built for thousands rather than millions, (2) new investors with smaller checks (e.g. angels, accelerators, micro VCs), (3) new management science for start-ups and (4) a faster adoption of new technologies by consumers (e.g. due to a more connected and globalized society).

Blue-chip companies are declining, and startups are booming. How does this matter? Why are startups important? First of all, startups mean a lot for economic and job growth. Most successful tech startups like Google, Facebook and Amazon drive today's economy. In the next decade, most likely, the global economy will have big players the reader has never heard of, they may not even exist yet. Startups growth at an accelerated pace and cause meaningful disruptions to the global economy. Beyond creating revenue, startups create jobs, lots of them. As a matter of fact, over the past 28 years "startups were responsible for all new net job creation in the US" [Kane, 2010]. Startups have also helped to reinvent power structures, and we are transitioning to a society where power is participatory and held by many [Heimans and Timms, 2014]. A new set of values centred on participation has been created: informal decision making, networked governance, open-source collaboration, radical transparency, do-it-ourselves culture, short-term affiliation,etc.

2.2.3. LIFE CYCLE

For the life cycle of startups there exists many frameworks, they are all very similar, and they all start with an idea and finish with a mature organisation. In figure 2.1 we can observe a typical life cycle for a startup organisation, contrary to an SME which would typically show slow and steady growth, a startup faces a period of economic losses followed by an accelerated growth that eventually settles at maturity.

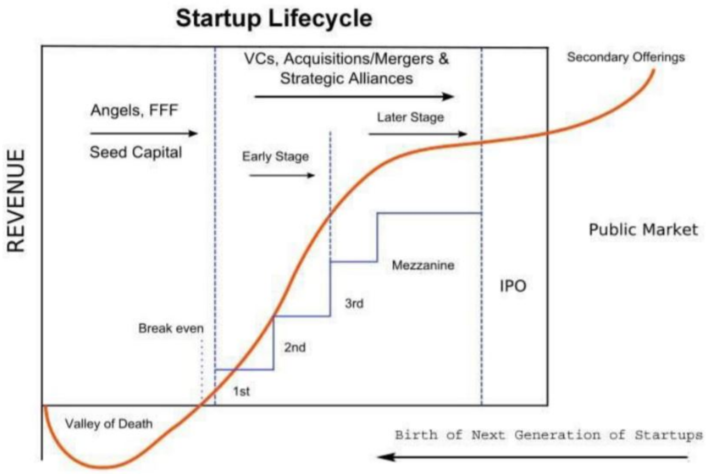

Figure 2.1: Startup typical life cycle [Valley, 2018].

Two important concepts to highlight from this figure are the *valley of death* and the *break even point*. At the beginning of the life cycle, a startup burns cash much faster than they create revenue; in the phases of exploration and earlier development of the product, a startup creates no revenue at all. Resources are consumed as research & development is conducted. Once the product is launched and commercialised successfully, startups start creating positive cash flow. Once the cumulative profit/loss reach zero once again, startups reach the *break even point* and exit the *valley of death*. Various models of startup's life cycle converge to a single common point: once a startup achieves product-market fit (henceforth PMF) a company starts being successful, it goes out of the valley of death, reaches the break-even point and starts scaling into a mature company.

Another framework is the one presented by Steve Plank in his book *Four Steps to Epiphany"* [Blank, 2013], according to this framework a startup goes through the following four stages: (1) customer discovery, (2) customer validation, (3) customer creation and (4) company building. This framework focuses on understanding customers' needs. Blank divides this four phases into two broader groups: to the first two phases he refers as "search", to the two later he refers to as "build", and to everything after that he refers to as "growth". In the search phase a startup's goal is to achieve PMF. A startup goes from search to build once it achieves customer validation and this occurs once the following three conditions are satisfied: "a sales channel that matches how the customer wants to buy and the costs of using that channel are understood", "sales are achievable by a sales force", and customer acquisition and activation are well understood. Once a company reach the build phase, it achieves a positive cash flow and starts scaling into a mature company, Blank affirms that this occurs typically when a company has 40 employees. Once a company starts scaling, it leaves the casual, informal culture into a more established organisation with culture, training, processes and procedures. We see many similarities with the framework proposed by Steve Blank and the life cycle shown in figure 2.1.

Another framework is the one propose by Startup Genome's founder Max Marmer's, this framework is more explicit and identify the key actions and milestones that take place during a startup life cycle. From this framework it is important to highlight and make a distinction between the concepts of prototype and minimum viable product (henceforth MVP) [Bichara, 2018]. The former serves to demonstrate a concept, it serves to illustrate an idea rather than talking about it. The later actually does the concept, not finished, but still functional. In Marmer's framework the following stages take place [Marmer et al., 2011]:

1. **Discovery:** this phase takes around 5-7 months, the goal of this stage is to validate if the startup is solving a meaningful problem and if there are customers interested in its solution. In this stage the founding team is formed, a prototype is created, potential customers are interviewed, and the value proposition is defined.

2. **Validation**: this stage takes around 3-5 months. The goal of this stage is that the product is acquired by initial customers, achieve PMF. Startups achieve this goal by raising pre-seed money, recruiting key hires and creating a MVP. Startups pivot if necessary or proceed to next stage.

3. **Efficiency**: this stage last 5-6 months and its main goal to refine the business model and to improve the efficiency of the customer acquisition process. Startups achieve a repeatable and scalable business model.

4. **Scale:** in this very last stage, startups try to drive growth more aggressively. In this stage positive revenues are first achieved (break-even point).

From Marmer's framework is important to highlight the concepts of MVP and prototype. These concepts will be later mentioned when selecting the success predictors and criteria.

Financing Life Cycle

The financing life cycle of a startup is the process of collecting funds from formation to maturity. In figure 2.2 the different investments that take place during a startup's life cycle can be observed, from friends & family to initial public offerings. It is essential to highlight that the concepts of incubator and accelerator are often used interchangeably as they both serve similar purposes; they both (most of the times) provide seed money and mentoring in exchange for startup's equity. The difference relies on, as the names suggest, the stage of the startups they focus on; incubators typically invest in very early-stage startups, whereas accelerators typically focus on those who have more significant traction.

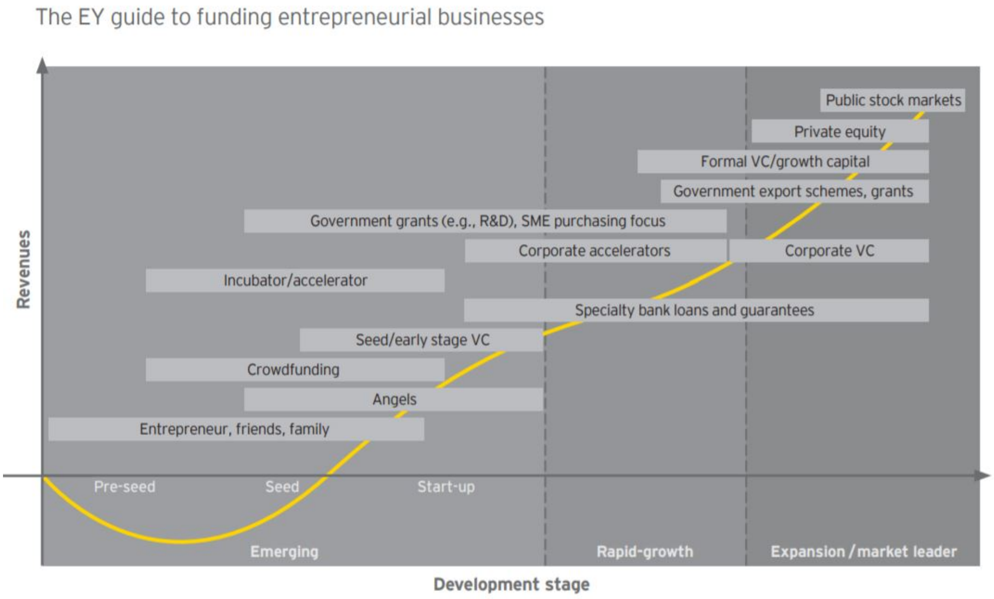

Figure 2.2: Startup typical financing life cycle [Young, 2013].

1. **Pre-seed:** in the very early stages of a startup's life cycle money is typically collected through FFFs. Once the startup has a little bit of traction, they start collecting money through crowdfunding, incubators/accelerators and angels. Accelerators (or incubators) are organisations that invest in startups in very early stages; they typically invest in startups that have some early development and will rarely invest in startups with just an idea. Startups that go through an accelerator program usually receive pre-seed money and mentorship so that they can develop a MVP, start commercialising their product and scale into a mature company. An angel investor is a wealthy individual, typically with more than 1M euros of net worth that invest in early-stage startups. Typical pre-seed money is in the order of tens of thousands of euros.

2. **Seed:** once the startup has a MVP they start collecting funds through more wealthier investors: accelerators, wealthier angels and early-stage venture capitalist (a.k.a micro VCs). A venture capitalist

(henceforth VC) is a professional investor that works for a venture capital fund; funds are raised collectively from institutions and individuals and invested in seed-stage startups. The investment at this stage is considerably larger; the median deal size for seed companies in the US is around $2.2M [Fundz, 2018].

3. **Series:** once the startup starts scaling up it raises money from a series of rounds. The first round is called Series A and is usually the first significant round of venture capital financing. The average Series A round size is $10.5M [Fundz, 2018]. As the startups scale up more investors buy in stocks of the firm through series of consecutive rounds. The median deal size for Series B rounds in the US is around $24.5M [Fundz, 2018]. The median deal size for Series C rounds in the US is $50M [Fundz, 2018]. Series C round is the first of what are typically called later-stage investments, and this can continue into Series D, E, F and even G. The size of these rounds is in the order of hundreds of millions. Investors in later-stage rounds are larger VCs, private equity firms, hedge funds and banks.

4. **Exit:** the life cycle of a startup typically finishes when an exit is achieved. An exit can occur either through merging and acquisitions (henceforth M&A) or through an initial public offering (henceforth IPO). As the name suggests, in a M&A a company can either be acquired by a larger corporation or merge with a comparable size one. An IPO occurs when the startup sells equity to the public and then becomes a stand-alone corporation. Although many startups dream with achieving unicornization (valuation over $1B) before exiting, most exits occur earlier, according to CB insights 97% of the exits were M%As and most happened before Series B [Joffe and Eversweiler, 2018].

2.2.4. Startup Ecosystem

In the previous subsection, the definition of startup was provided. What can be done to drive startups' growth? In nowadays' information era, innovation is conducted differently. Society has moved from a closed innovation framework, where big companies conducted R&D with little to no collaboration, to an open innovation framework, where boundaries of the firm are permeable and collaboration happens in a higher degree. Innovation no longer occurs in isolation. Other scholars have studied how innovation occurs outside of firms' boundaries more collaboratively. When innovation occurs from the interaction of multiple entities, we refer to this as an innovation system. The concepts of regional innovation system (RIS) and national innovation system (RIS) got popularity in the early 1990s. Other concepts that were crucial for the development of the concept of innovation systems were Porter's industrial clusters as well as Etzkowitz and Leydesdorff's triple helix model [Lundvall, 2007].

The previous concepts serve as a foundation to understand the concept of *startup ecosystems.* Innovation systems frameworks focus on innovation carried out by big companies and the public sector, whereas startup ecosystems focus on innovation driven by startups. As innovation may occur by different actors, we may say that startup ecosystems are a subset of the broader innovation system. Some may refer to the concepts of entrepreneurial ecosystem and startup ecosystems interchangeably. In strict terms, the startup ecosystem is a subset of the entrepreneurial ecosystem where only rapid growth, innovative and highly disruptive companies are included.

Startup ecosystems are formed by different organisations interacting as a system to create new startup companies and help them grow. Actors in the ecosystem may include startups themselves, academia, government, corporations, investors, support organisations (e.g. co-working spaces, incubators & accelerators, event organisers), service providers (e.g. consulting, accounting, legal) and advisory & mentorship organisations (e.g. techleap.nl). Startup ecosystems also vary according to their geographical scope. It is not rare that cities (e.g. Amsterdam) or regions (e.g. Silicon Valley) significantly outperform and undermine the broader national system. Robust NIS, typically managed by policy-makers, consists of cities or regions working conjointly in a balanced relationship of cooperation and competition (a.k.a. coopetition).

How to manage startup ecosystems? Like any system, startup ecosystems also consist of inputs, outputs, processes and feedback. Ecosystem management is driven by clearly defined goals at the output of the system. Inputs and processes are managed through the execution of policies, protocols, and practices. Feedback takes place by continuously and carefully monitoring the interactions between the inputs of the system and the outputs. Managing startup ecosystems results to be a highly complex task.

How to properly define the inputs and outputs of startup ecosystems? Startup genome is an organisation entirely dedicated to studying ecosystems worldwide. Together with global thought leaders, they define robust strategies and implement programs to drive lasting change through startup ecosystems. In most countries, despite long efforts to build startup ecosystems, not enough jobs are being created and not enough economic growth can be seen. Startup Genome attributes this failure to the models being used to assess startup ecosystems and a lack of comparable local and global data. In the past, industrial clusters were constructed from assets accumulated by large corporations over the decades. As this was the case, the government was able to control outputs by providing support and incentives to a handful of well-known and stable organisations. This strategies are outdated and no longer applicable. Porter's industry cluster framework "has been bogged down by the fundamentally different shape of startup ecosystems" [Genome, 2018]. To deal with this issue, Startup Genome has developed a new framework for the assessment of startup ecosystems.

Let us begin by identifying the characteristics of startup ecosystems. First, startup ecosystems can be defined in terms of size, according to Michael Porter's framework, innovation, productivity and rate of new entrants increase as the ecosystem grows. Second, ecosystems can be defined in terms of connectedness, both at the local level (similar to Porter's framework) and the global level (original contribution by Startup Genome). Lastly, ecosystems can be defined by their scaleup production. Among startups, the top 10% contribute to 80% of gross revenue and job creation [Genome, 2018]. Hence, scaleup production becomes the primary goal of startup ecosystems.

Startup Genome proposes two frameworks to asses startup ecosystems. The first framework is called the "Life Cycle Model" and the second the "Success Factor Model". The former model describes how ecosystems evolve through different phases. It serves as a lens through which to look at the gaps (and strengths) of the later model and understand the importance of each.

According to the first model, startup ecosystems' performance increase with size, resources and experience. What drives this growth is their global ambition and their connectedness. In the process, ecosystems go through the phases of activation, globalisation, expansion and integration. Performance is measured in terms of the production of scaleups, exits and unicorns [Genome, 2018].

The second model, the *success factor model*, shows how startup ecosystems scale to create a greater economic impact. On the one hand, economic impact increases as the ecosystems jump from local to global systems. In the early activation phase of startups, as mentioned in the life cycle model, ecosystems solely focus on the local system. The quality of ecosystems at the local level is measured by their experience and connectedness. Performance is captured by startup output and output growth index. The local system also includes the local context which entails the "collection of cultural issues, English proficiency, coding proficiency, infrastructure, size of the local and national economy, and the general laws and regulations" [Genome, 2018]. As ecosystems move towards the late activation phase, global system becomes the focus, communities develop global business models, global market reach, and boost their performance through exits. It is not rare that a handful of large exits combined may cause a sharp acceleration in the ecosystem growth by provoking a surge in net resource attraction (i.e. resource recycling).

Startup Genome measures performance according to various indicators. (1) Production of scaleups, unicorns and exits. (2) Startup valuation. (3) Ecosystem value which they define as the sum of startups' valuation and exits' total value. (4) Startup output (i.e. count of startups) (5) Growth indexes, which capture the growth of exits, startup output and funding.

The previous model also includes the following factors as success factors for startup ecosystems. (1) Founder, a combination of their mindset, ambitions, demographics, economic and educational data, startup strategy and know-how. (2) Talent, factor determined by the ability of early-stage startups to hire and attract key employees (particularly those with startup experience), the quality of the employees and the cost of acquisition. (3) Funding, as measured by the proportion of startups obtaining seed funding, attrition rate from seed to Series A, median seed amounts, median Series A amounts and the number and proportion of experienced VC Firms. (4) Startup experience. (5) Global connectedness. (6) Local connectedness, as captured by the sense of community, local relationships, collisions and density. (7) Global Market Reach. (8) The organisation, measurement of the quantity and quality of organisations, programs, events, and other activities. (9) Economic impact. As it can be inferred by now, all these factors and criteria are highly complex to measure and are being continually refined and redefined by organization such as Startup Genome. The approach herewith presented serves to explain, in a more understandable way, how exactly startup ecosystems drive economic growth in the societies they are embedded in.

Dutch Startup Ecosystem

The Netherlands has shown outstanding entrepreneurial growth in the last decade. Nevertheless, as it was discussed previously, entrepreneurial activity is not necessarily causal to the creation of highly innovative and rapid growth companies. The rapid growth on the entrepreneurial activity in the Netherlands is explained in a higher degree by an increase in self-employment rather than in the creation of new innovative firms, to this phenomenon we refer to as the *Dutch Entrepreneurship Paradox* [Stam, 2014].

Regardless of the relative stagnation in the production of rapid growth firms at the beginning of the decade, the Dutch startup ecosystem has shown significant improvements in recent years. The Netherlands is ranked on the 15th position in Startup Genome's World Rankings on *startup ecosystems* and 2nd on the Global Innovation Index (GII) [Dutta and Lanving, 2018]. According to the previous ranking, the ecosystem is remarkably strong in the sub-sectors of fintech, health/life sciences and agriculture/new foods. The later evaluates countries by their innovative performance according to several metrics in various categories. Particularly, the Netherlands leads the categories of business sophistication, knowledge & technology and creative outputs. Another framework, the Global Entrepreneurship Monitor (GEM) [Bosma and Kelley, 2018], evaluates entrepreneurial ecosystems based on enabling conditions for ecosystem's development. As it can be seen in figure 2.3, the Netherlands outperforms the mean in all the criteria. The remarkable infrastructure conditions of the ecosystem are the result of the efforts made by the many actors in the ecosystem. A detailed map of the actors in the ecosystem, designed by Halbe & Koenraads can be found in the appendix **??** [Latham, 2018].

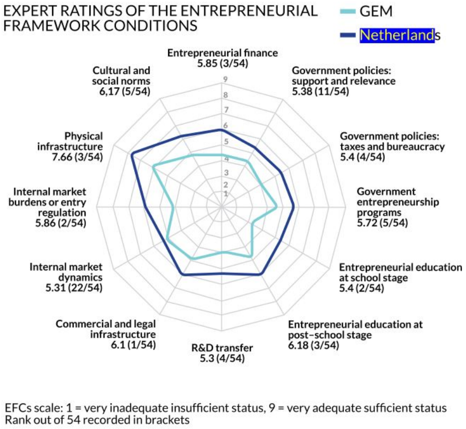

Figure 2.3: Dutch Startup Ecosystem Framework Conditions Bosma and Kelley [2018].

Lastly, to conclude with the overview of the Dutch startup ecosystem, basic information based on techleap.nl's data is provided. The *startup finder* tool from techleap.nl is an online platform/tool that serves to explore the Dutch startup ecosystem. All the information hereby shown is publicly available and can be collected from their webpage.

To the date, 6th June 2019, the startup ecosystem has an estimate of 6870 companies and 456 thousand employees. A total of 715 million euros of venture capital was invested in 2018. The regions with the most startups & scaleups are North Holland with 2470, Zuid Holland with 1415 and Utrecht with 1184. In the year 2018, the three industries that received the most substantial funding were health, fintech and enterprise software. There is a total of 1002 startups & scaleups in enterprise software, 712 in health and 585 in fintech. Other noticeable industries are energy with 510 startups and food (includes agriculture) with 318 startups. There is a total of 2044 investors, eighty-four incubators/accelerators (or similar), forty-six higher education institutions and ninety-five working spaces.

"Your most unhappy customers are your greatest source of learning."
- Bill Gates, Microsoft Founder

3
PREDICTORS AND CRITERIA

In this chapter, the selection of startup success predictors and criteria is delivered. This selection serves as the foundation for the predictive model in chapter 4. This chapter is divided into three sections: desk research 3.1, interviews 3.2 and variable selection 2 3.3. In the first section through a literature review, we deliver an initial set of independent and dependent variables. A total of forty-three independent variables and eight dependent variables are initially explored, to this set of variables we referred to as “variable selection 1”. In the second section, through a series of interviews, the initial selection of predictors and criteria of success is discussed. The selection is refined, and novel variables are included. A total of thirty-seven independent variables and four dependent variables are selected after the previous stage. This set of variables we referred to as “variable selection 2”; this set is explored for the quantitative analysis.

3.1. VARIABLES SELECTION 1: DESK RESEARCH

Which set of variables should be selected for the predictive model? Although this research intends to be distinctive from previous approaches, undoubtedly, we need to start by conducting a literature review on similar studies. To achieve this, we follow a similar process to the one carried in chapter 2:

Search Terms: as the title of this research, we focused our desk research around the keywords *startup success prediction*. By altering these keywords, we reviewed other terms such as *startup success*, *startup success factors* and *startup success criteria*.

Bibliography Manager: EndNote and BibTeX. EndNote is used as the primary reference manager, BibTex is used to create the connection to the LaTeX editor.

Sources: Information in this chapter was collected mostly from academic and conference journals. We obtained these papers by using the following databases and search engines: Google Scholar, Science Direct, Scopus and Web of Science. Aside from journals, we also included master theses, these were found through Google Scholar and TU Delft academic repository. After an initial screening, twelve sources were consulted. In table 3.1, the distribution of references by type is shown. Sources are divided into two categories: *factor specific* and *generic*. Factor specific includes papers that focus on the influence of a single factor to explain startup success. Papers classified as generic, study the relationship between multiple variables to explain startup success. For this study we focus on the later as we intend to build a predictive model based on a comprehensive selection of variables. A total of six sources are thoroughly read to determine the initial set of variables.

We selected a total of no more than six sources for the following reasons. First, we observed a high level of redundancy across sources. Second, the selected studies do extensive research to select their set of variables. Third, this study intends to include new and distinctive variables that result from the series of interviews with actors in the ecosystem; hence, the desk research is not our only focus. We divide this section into three subsections: in subsection 3.1.1 the selected articles are reviewed, in subsection 3.1.2 a pre-selected set of success predictors based on the desk research is delivered, in subsection 3.1.3 several the candidates for success criteria are also presented.

	Factor Specific	**Generic**	Σ
Journal Article	4	3	7
Conference Paper	1	2	3
Master Theses	1	1	2
Σ	6	6	12

Table 3.1: Theoretical Framework Sources.

3.1.1. LITERATURE REVIEW

In this subsection, we discuss six literature sources that explain startup success as a result of various independent variables. For every source, if available, the following elements are mentioned: title, brief description, research method, count of success predictors, categorisation of the predictors, success criteria, other variables (mediating, moderating, control), conceptual model and hypotheses.

1. FINDING THE UNICORN: PREDICTING EARLY STAGE STARTUP SUCCESS THROUGH A HYBRID INTELLIGENCE METHOD [Dellermann et al., 2018].

DESCRIPTION: Study on a *hybrid intelligence method*, combining the strengths of humans and machines to predict startup success.

RESEARCH METHOD: Quantitative, it combines machine learning algorithms together with collective intelligence automation to predict startup success.The study compares the results of the following ML algorithms: logistic regression, naive bayes, support vector machine (SVM), artificial neural network (ANN) and random forests.

THEORETICAL FRAMEWORK: A total of twenty-one variables were selected in this study, seven of which are considered *soft variables*. Soft variables are evaluated through collective intelligence (people). The variables are categorized into the following six categories: meta(similar to demographics), value proposition, market, resources and third-party support. No control, mediating or moderating variables are identified in this study. The criteria for success is the achievement of a Series A funding.

2. A MODEL OF ENTREPRENEUR SUCCESS: LINKING THEORY AND PRACTICE [Limsong et al., 2016].

DESCRIPTION: Drawing from human capital theory and theory of opportunity identification, this research aims to develop and test a theoretical model of entrepreneur success.

RESEARCH METHOD: Quantitative, data is collected through questionnaires from two-hundred successful entrepreneurs. Data is analysed using confirmatory factor analysis(CFA) to test the relationships between observed indicators and latent constructs.

THEORETICAL FRAMEWORK: twenty-three variables are grouped into six main factors, divided into internal and external. It is hypothesized that all six factors contribute positively to entrepreneurial success. Success is measured as financial and non-financial. All variables are measured by employing questionnaires. The vast majority of the variables are ordinal and evaluated using a Likert scale. The hypotheses are confirmed and it is shown that external factors cause more considerable variability in the dependent variable. No control, mediating or moderating variables are identified.

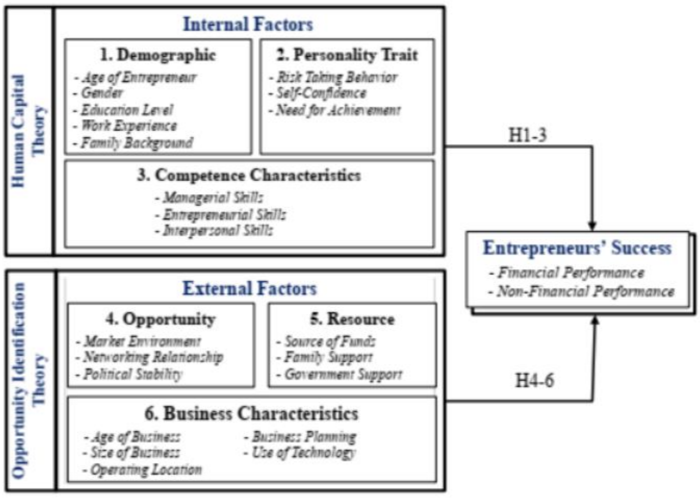

Figure 3.1: Conceptual Model of Source 2.

3. CRITICAL SUCCESS FACTORS OF THE SURVIVAL OF START-UPS WITH A RADICAL INNOVATION [Groenewegen and Langen, 2012].

DESCRIPTION: The purpose of this study is to determine which factors are most important for the success of a startup with radical innovation in the first three years. It is shown that different factors correlate in different ways with the different growth variables (turnover and employee).

RESEARCH METHOD: Quantitative, data from 125 Dutch startups is collected through questionnaires and analysed using correlation analysis.

THEORETICAL FRAMEWORK: A total of 19 independent variables are selected to explain startup success. These variables are grouped into three main categories: innovation, organisation and entrepreneur. Two dependent variables are used to measure success, employee and turnover growth in the first three years of the startup. No hypotheses are stated. No control, moderating or mediating variables are explicitly identified. Companies older than fifteen years are excluded.

4. SUCCESS AND RISK FACTORS IN THE PRE-STARTUP PHASE [Gelderen et al., 2005].

DESCRIPTION: Longitudinal study on the startup success factors. It focuses on studying nascent entrepreneurs throughout their entrepreneurial journey. Contrary to cross-sectional studies, this study collects data before the startup is even launched and keeps track of it throughout their life cycle.

RESEARCH METHOD: Quantitative, this study collects data from a sample of 517 nascent entrepreneurs (2.7% out of 49936 phones dialled). After three years, 192 entrepreneurial efforts were successful, 115 were abandoned and 210 are still trying. Data were analyzed utilizing principal component analysis(PCA) and logistic regression.

THEORETICAL FRAMEWORK: A total of 19 independent variables are selected to explain startup success. These variables are grouped into the following categories: individual demographics, human capital, motivation, process, environment financial, network, ecological and intended organisation. This study states that "the first success of a firm is its birth", nascent entrepreneurs are periodically interviewed overtime to evaluate if they have successfully started their business, abandoned or are still trying. No hypotheses are explicitly mentioned. Mediating or moderating effects are mentioned but not included in the analysis of this study.

5. EXAMINING THE CRITICAL SUCCESS FACTORS OF STARTUP IN THAILAND USING STRUCTURAL EQUATION MODEL [Nalintippayawong et al., 2018].

DESCRIPTION: This study examines the success factors from a sample of Thai startups. This study shows that supporting partner, business model, market opportunity and customer perspective result to be critical predictors.

RESEARCH METHOD: Quantitative, data is collected from a sample of 152 Thai startups. Causal relationships between the variables are studied applying the structural equation model(SEM). This model employs confirmatory factor analysis to evaluate latent constructs and multi-variable regression to determine the links between the constructs.

THEORETICAL FRAMEWORK: A total of 16 independent variables are selected to explain startup success. These variables are grouped in latent constructs that are later validated with factor analysis. These latent constructs are support partners, business model, market opportunity and customer perspective. The dependent variable, named as startup potential, is measured by combining profit and startup funding. The variables are related to each other as shown in the conceptual diagram in figure 3.2. It is hypothesized that the variables have a positive effect on each other as shown in the conceptual diagram (H1-H5). No control, moderating or mediating variables are identified.

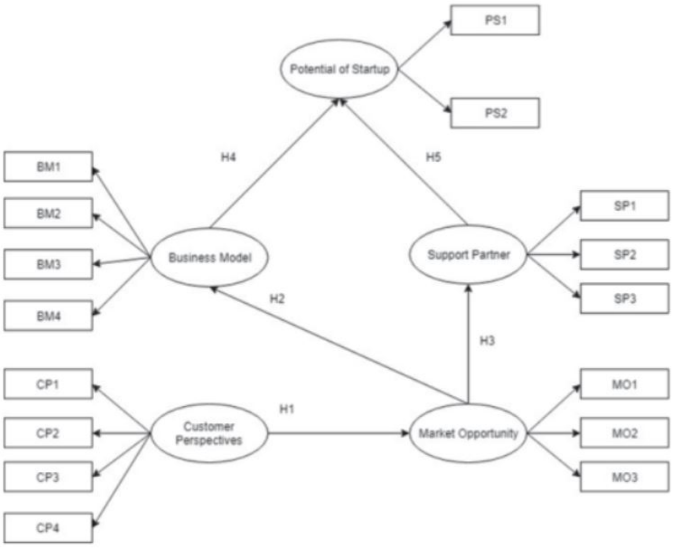

Figure 3.2: Conceptual Model of Source 5.

6. PREDICTING STARTUP SUCCESS WITH MACHINE LEARNING [Bento, 2017].

DESCRIPTION: This master thesis explores and analyses the data of 495798 startups by using the metrics provided by the website CrunchBase.com, it employs machine learning techniques to classify startups as successful or not successful.

RESEARCH METHOD: Quantitative, data is collected from a sample of 86 588 startups. ML supervised learning classification techniques like Logistic Regression, Support Vector Machines and random forests are employed to evaluate the sample. The later technique showing the best results, a 94.1% True Positive Rate and a 93.2% area under the curve (ROC) are achieved.

THEORETICAL FRAMEWORK: A total of 158 binary independent variables are used to predict the dependent variable. A startup is considered successful if it has achieved an exit (exit or M&A). These variables are not grouped into any particular categories in the model. No hypotheses are explicitly made. No control, moderating or mediating variables are identified. No conceptual model is drawn.

3.1.2. PREDICTORS SELECTION

As it can be seen from the previous subsection, there are many choices for predictors and criteria of startup success. These variables can be organised into various categories and related to one another in many different ways. In this subsection we intend to deliver the initial set of variables that are further validated and enriched in section 3.2. After carefully reading the six sources mentioned in subsection 3.1.1, we excluded sources five and six for the selection. Source five was excluded as the variables are ambiguous and proper definitions are not provided. Source six because the selection of variables is far too extensive. After reviewing all the variables from the selected four articles, the following categorization is proposed: *founders' characteristics, business characteristics, innovation characteristics, actions & decisions, resources, external environment* and *third-party support.*

As the number of variables is considerably large, the different categories have been divided into three different tables (3.2, 3.3 and 3.4). All tables are displayed in a cross-tabular format, with variables on the rows and sources in the columns. Inside the cells, an alternative name (if any) and the type of variable (e.g. categorical) are provided. This information helps the reader to understand how the variables are measured in the different sources. Ideally, to observe the recurrence of the variables, these should be coded multiple times within the same document. Taking into consideration that a total of more than forty variables (codes) are identified, the free trial of Atlas.ti (max one hundred quotes) results very limiting. Instead, as the core of this study is not qualitative, the analysis is kept simple, and the variables are coded once per document. The first table 3.1.1 displays the success factors related to the founders' characteristics.

	VARIABLES	**SOURCE 1**	**SOURCE 2**	**SOURCE 3**	**SOURCE 4**	Σ
FOUNDERS	1. Founders' Age		(numeric)		(numeric)	2
	2. Gender		(categorical)		(categorical)	2
	3. Work Experience		(numeric)	(numeric)	(numeric)	3
	4. Entrepreneurial education	(categorical)	Entrepreneurial skills (ordinal)			2
	5. Entrepreneurial experience	Previous founded ventures (binary)	Entrepreneurial skills (ordinal)	Years of exp (numeric)	Experience in firm founding (binary)	4
	6. Industry experience			Technological expertise (ordinal)	(numeric)	2
	7. Managerial Skills		(ordinal))	(numeric)	(numeric)	3
	8. Interpersonal Skills		(ordinal)			1
	9.Education level		(categorical)	Higher education (binary)	(categorical)	3
	10.Social Capital			Social network (ordinal)		1
	11. Personal Dedication				(binary)	1
	12.Risk Profile		Risk taking behavior (ordinal)	Willingness to take risks (ordinal)		2
	13.Personal Motivation		Need for achievement (ordinal)			1
	14.Self Confidence		(ordinal)			1

Table 3.2: Pre-selection of independent variables 1 out of 3.

As it can be seen from table 3.2, many success factors are related to the founders' characteristics. Entrepreneurial experience is the only factor that is mentioned in all the sources and is measured in different ways. Work experience is mentioned in three out of four articles and is evaluated in a similar manner (number of years of work experience). Some articles make a distinction between entrepreneurial education and experience (source 1), whereas other sources are more general and refer to entrepreneurial skills (source 2). Some of these variables may be correlated, such as age and experience; nevertheless, some sources employ them separately in their analysis. Source 2 provides this research with the highest amount of variables related to the founders' characteristics; it also includes some distinctive variables such as personal motivation and self-confidence. In table 3.3, we present the variables that are related to the business characteristics and to the innovation.

	VARIABLES	SOURCE 1	SOURCE 2	SOURCE 3	SOURCE 4	Σ
BUSINESS	15. Firm Age	(numeric)	(numeric)			2
	16. HQ Location		(categorical)			1
	17. Industry	(categorical)			(categorical)	2
	18. B2B vs B2C	(binary)				1
	19. Tech Level		Use of Technology (ordinal)		Techno nascent (binary)	2
	20. Ambition to grow				(binary)	1
	21. Networking				Membership to formal networks (Binary)	1
	22 Business Model	(categorical)				1
	23. Revenue Model	(categorical)				1
	24. Scalability	(not defined)				1
INNOVATION	25. Innovativeness	(not defined)		Degree of radicalness (ordinal)		2
	26. Unique advantage			(ordinal)		1
	27. Technological hype	(categorical)				1

Table 3.3: Pre-selection of ndependent variables 2 out of 3.

From table 3.3, we observe that there is a considerable amount of variables linked to the business characteristics and a few to the innovation itself. For these categories, there is no dominating variable (no fully filled rows). Source 1 dominates with a total count of eight factors proposed. Sources 2 and 3 have a count of only three and two variables respectively. A completely different situation to what is shown in table 3.1.1 where the selection of variables from sources 2 and 3 related to the founders' characteristics is much larger than that of source 1. From both tables it can be observed that source 1 has a stronger focus towards the individual whereas source 2 and 3 focus on the organisation. Source 4 presents a significant amount of variables in both founders' and business characteristics but no variables related to the innovation.

At first sight, it is difficult to predict how these variables affect the potential success of startups as these are related to one another in complex ways. As an example, although overall success is more likely to be achieved over time, the life cycle of startups varies significantly across industries, and hence firm age is not a very fair success predictor. In this section it is important to have an initial approach to the variables. In section 3.2, the selection of variables is refined, and the variables are thoroughly discussed. In table 3.4, the selection of variables related to the categories of *actions & decisions*, *resources*, *external environment* and *third-party support* is presented.

	VARIABLES	SOURCE 1	SOURCE 2	SOURCE 3	SOURCE 4	Σ
Act & Dec	28. Customer Proactiveness			(numeric)		1
	29. Proof of Value	(numeric)				1
	30. Proof of Concept	(binary)			(categorical)	2
	31. Business Planning		(ordinal)		(binary)	2
Resources	32. Number of Founders			Multiple Owners (binary)		1
	33. Team	(numeric)			(binary)	2
	34. Money to Market	Capital raised (numeric)		75K Euro seed capital (binary)	Startup-capital (categorical)	3
Environment	35. Competition	(numeric)				1
	36. Market Environment		(ordinal)		Risk of market (ordinal)	2
	37. Political Stability		(ordinal)			1
Third-Party Support	38. Mentorship			External advice and knowledge (binary)	Information and guidance (binary)	2
	39.Funding Source	Financial support (categorical)	(categorical)	(Investors capital (binary)	3rd-party money (binary)	4
	40 Webpage Analytics.	(numeric)				1
	41. Social Medial Analytics	(numeric)				1
	42. Family Support			(ordinal)		1
	43. Government Support			(ordinal)		1
	Total	18	18	16	19	71

Table 3.4: Pre-selection of Independent Variables 3 out of 3.

After the categories of *founders' characteristics* and *business characteristics*, the third-largest group of variables is third-party support. From this category, the variable *funding source* is used in all four sources. Related to this, the variable *money to market* from the resources category is employed across three sources. It can be observed that funding (as expected) is a commonly used predictor for startup success. All sources have at least one variable for all four categories presented. There are no dominant sources when observing the four categories as a whole. Source 3 has a slightly higher count of variables in the *third-party support category*, including family and governmental support as startup success predictors.

In table 3.4, the total count of variables per source across all categories is shown in the total row at the bottom. A total of seventy-one variables from all four sources are combined into forty-three startup success factors. The total count per source across all categories is comparable, ranging from sixteen variables (source 3) to nineteen (source 4). The only source that has at least one variable in all the categories is source 1. For this pre-selection of independent variables, we tried to include all variables mentioned in the literature. Nevertheless, we excluded some variables because of the following reasons: difficult to measure, ambiguous definition, redundancy, risk of significant bias, irrelevant for this study, among others. This pre-selection of independent variables is refined and enriched in section 3.2. In the next subsection 3.1.3, we present and discuss the first selection of startup success criteria (dependent variables).

3.1.3. Criteria Selection

In this subsection, various candidates for success criteria (a.k.a the dependent variable) are presented and discussed. The selection of these candidate variables was made taking into consideration the desk research and the researcher's own experience as data analyst trainee at techleap.nl.

CRITERIA	DESCRIPTION
Employees	Best measured as total full employment(TFE). TFE as a success criterion can be measured either as an accumulated value or as a year-over-year (YOY) value (absolute or percentual). This variable presents some problems. Estimates in databases such as dealroom.com are often exaggerated; this can happen because the number of employees scrapped from LinkedIn often differs from the actual TFE. Furthermore, employee count as a measure of success is industry-dependent (a nanotech company requires fewer highly specialized employees).
Financial KPI	Revenue, or any other relevant financial KPI, may be used as a measurement of startup success. Either as an accumulated or YOY value (absolute or percentual). The advantage of using financial KPIs is that they are less susceptible to differences across industries (especially if measured as a percentual change). Although financial KPIs are accurate (as required by the law) and they seem to be the right candidate for success criteria, obtaining this data can be extremely challenging.
Funding Round	Funding can be measured as a categorical variable. Startups can be positioned in different stages as explained subsection 2.2.3. Startups in early stages would be in pre-seed or seed stages whereas more developed startups (or scaleups) would be positioned in further Series A, B, later series and exits.
Profitability	As explained in subsection 2.2.3, startups go through various stages and milestones during their life cycle. A significant milestone for start-ups is to reach the break-even point (see figure 2.1). Break-even occurs once accumulated revenue equals accumulated cost. Although this is just the start for many startups, this point is often seen as the first benchmark to be reached. It is important to notice that some startups may not reach this point and still be considered as very successful startups (e.g. Uber).
Scaleup	As mentioned in subsection 2.2.1, "the only essential thing to define startups is growth". The scaleup concept, commonly used in the startup world, is defined based on sustained revenue growth. Although revenue information is challenging to obtain, startups could classify themselves as scaleups without disclosing any financial data.
Total Funding	Startup success can be measured as a continuous variable depending on the total amount of funding collected to date. This information can be easily obtained from most business databases. Total funding may be misleading. Some startups may have a significant amount of funding, but they are having considerable losses or have not even launched their product in the market yet. Total funding is a standard metric when discussing the worthiness of startups in the startup ecosystem, and this metric is related to the valuation.
Valuation	Valuation is the process that determines the economic value of a business. The valuation of a business is calculated taking into account various elements such as the accountancy, team quality, traction, reputation, etc. For startups valuation is highly speculative whereas for established business the calculation is more straightforward. Despite being speculative, startup valuation is frequently used when discussing success. A *unicorn*, considered by many as the ultimate goal for every startup, is a startup that reached a valuation of US$1 billion within the first five years of existence.

Table 3.5: Candidates for dependent variable.

In the previous table, we discussed seven potential candidates for startup success. As the objective of this research is to predict success as a dichotomous variable, if selected as criteria, numerical variables have to be transformed. For the predictive model, one or more variables can be selected for exploration. Variables can even be combined through an index or as an if-else condition. In the next section, we discuss the candidate variables with knowledgeable actors in the ecosystem and a second selection is delivered. This second selection is then used as a foundation for the quantitative analysis in chapter 4.

3.2. Variables Discussion: Interviews

The previous section delivered an initial set of success predictors and criteria. The purpose of this section is to validate and enrich the previous set and to deliver a second selection of independent (predictors) and dependent (criteria) variables for startup success. To achieve this goal, the initial set of variables was discussed with knowledgeable actors in the ecosystem. We reached a total of thirteen persons and interviewed seven of them. The interviews took place in an unstructured manner, and no predefined questions were given. Overall, the interviews were carried out in the same way: introduction by both the interviewer and the interviewee, a general discussion about key startup success factors and criteria, and a discussion about the first selection of variables obtained from the literature review. From the interviews, predictors were validated, removed or altered, and novel variables were proposed. This section is divided into two subsections. In subsection 3.2.1 we provide a review of the interviews conducted. In subsection 3.2.2, we explain the modification made to the first selection of predictors and criteria delivered from the previous section.

3.2.1. Discussion Overview

As previously mentioned, a total of seven actors were interviewed. Most interviewees were audio-recorded and all were requested to sign an informed consent. In this subsection, a brief review of all seven interviews is provided. Interviewees significantly differ from one another. They have different levels of experience, areas of expertise, roles in the ecosystem. For each interview we provide the profile of the interviewee (keeping anonymity) together with a summarised version of the interview. Within the summary, factors and criteria for startup success are examined, existent variables are evaluated, and new variables surface from the discussions. All interviews are combined into one single audio and uploaded to YouTube, subtitles were extracted, and a transcript to analyse the data was obtained.

Contrary to the previous subsection, Atlas.ti is not used to conduct qualitative analysis. As the core of this thesis is not the qualitative research, we decided to keep the analysis of this subsection as simple as possible. In order to employ Atlas.ti for the analysis of this section, a clean transcript must be provided to the software. The transcript obtained employing the YouTube technique is sufficient for a superficial analysis but presents a lot of typographic errors and text is not divided between interviewer and interviewee. Cleaning the transcript would have required an unnecessary amount of manual work. To improve the subsection of this thesis, future researchers can employ more sophisticated transcript techniques (through paid software). The interviews are analysed by identifying and highlighting startup success factors, criteria and supporting quotes along the transcript.

Interviewee 1

Profile: The first interviewee works directly with techleap.nl. He has over twenty years of international business experience. His areas of expertise include startups, high growth strategies, business development, M&A, corporate finance & law, sales & marketing, distribution channels, global trade, negotiations, operations, opening global markets, and organisational development & team building.

Interview Insights: At the beginning of the discussion, the interviewee was asked to discuss startup success. Various factors were mentioned, but the most critical factor, the one that discussed the most was **experience**. When talking about experience, the interviewee highlighted the fact that the Dutch startup ecosystem is significantly different from the one in the US (in which the interviewee has experience). The interviewee mentioned that entrepreneurs in the US are on average much older than Dutch entrepreneurs. Local entrepreneurs have more **governmental support**, and with this financial support, founders dare to take the risk and start their new venture. US entrepreneurs, on the contrary, wait some time, raise some money, gain more knowledge, build a **professional network** and gain credibility. The interviewee strongly thinks that financial support is insufficient to turn a startup into a big success. Startups, with inexperienced entrepreneurs, need strong **mentorship** so they can smoothly go through their financial life cycle. The interviewee also mentioned **team quality** as success factors, the **potential of the proposition**, and the **business model** (in the sense of B2B vs B2C). Lastly, when asked about the dependent variable, the interviewee has a preference for **revenue** and **cash-flow positive** threshold as indicators of success.

Interviewee 2

Profile: The second interviewee also works directly with techleap.nl. He has around twenty years of working experience, including entrepreneurial experience and four years of experience as managing director of an accelerator. His main areas of expertise are entrepreneurship, start-ups and business strategy.

Interview Insights: In this occasion, aside from the general discussion regarding startup success factors and criteria, the initial list of variables (as shown in tables 3.2, 3.3 and 3.4) was thoroughly discussed. One of the most important things to highlight from this interview is that the interviewee believes success factors are hardly tangible. The two main factors of success, as mentioned by the interviewee, were **organizational culture** and the **ability to adopt new insights** with the number of pivots as a suggested proxy. At the human capital level, **passion** and **team diversity** were mentioned to be key success factors. **number of founders** and **B2B vs B2C** variables were both mentioned to be worth considering. **competition** was discussed as a complex variable; having few competitors can be a signal of a non-existent market whereas too many competitors indicate an overcrowded market. The variables of **business and revenue model** were mentioned to be conflicting, overlapping with one another. Regarding the dependent variable, the interviewee has a preference for **total amount of funding**, **total amount of jobs created**, **revenue** and **social media metrics** as criteria of startup success.

Interviewee 3

Profile: The third interviewee, also works directly with techleap.nl. She has 15+ years of working experience, frequent public speaker, host and moderator at tech events. Areas of expertise include international development, business strategy, entrepreneurship and start-ups.

Interview Insights: In this interview, contrary to the previous interviews, factors and criteria of startup success were not discussed as a broad topic of conversation. Rather, the pre-selected list of variables was thoroughly discussed. The interviewee, when discussing the founders' gender variable, mentioned that it is worth looking at **diversity** in a broader sense (backgrounds, races, nationalities). Concerning the educational background in the team, the interviewee mentioned that many tech startups are conformed solely by people with a technical background. The interviewee asserts that it is important to differentiate between different kinds of **experience** (general, industry-specific, managerial, entrepreneurial). Like interviewee two, interviewee three emphasises the importance of building a strong **corporate culture**. **Intrinsic motivation**, **connectedness** and **innovativeness** were mentioned as interesting although difficult to measure. The variables of **time dedication** and **risk aversion** where also mentioned as worth analyzing.The **business planing** variable, was modified from a binary variable (originally related to the existence of a business plan) to a categorical variable. Novel variables that resulted from the discussion with the interviewee include **university ranking**, **previous income** earned and **employee incentives**. The variables of **business and revenue model** were found complicated, with not a clear categorisation and highly industry-dependent. From the **mentorship** variable, the interviewee considers important to analyse which is the role of accelerators in the ecosystem. Lastly, when discussing the dependent variables, the interviewee has a preference for firm **valuation**.

Interviewee 4

Profile: The fourth interviewee works connecting the Dutch startup ecosystem with Silicon Valley. He works developing, supporting and advising businesses. He has around twenty years of working experience, and the areas of expertise include recruiting, business development, partnerships, community building, event management, marketing strategy, client relations, product strategy, entrepreneurship, among others

Interview Insights: This interview was conducted virtually. In this situation, the pre-selected list was not discussed. Instead, the interviewee was asked to indicate which he considered as key factors and criteria of startup success. The interviewee mentioned **timing**, **team quality** and **uniqueness** as critical success factors. A startup entering the market too early may incur in unnecessary R&D expenses, users may not be ready for the innovation, and complementary products or services may not be available. When discussing uniqueness, besides the radicalness of the product, it is important to evaluate if the problem is worth solving, if someone is having that problem and willing to pay for the solution. Lastly, when discussing the dependent variable, the interviewee considered **revenue growth** and **profitability** as key indicators of success.

Interviewee 5

Profile: The fifth interviewee works for the *Erasmus Centre for Entrepreneurship* (ECE). This institution makes part of the Erasmus University Rotterdam and works to reduce the gap between university and market. To achieve this, besides providing educational programs to corporates, the centre also provides 4-weeks startup programs to coach startups in taking their first steps. As part of the ECE, the interviewee is the program manager for various education programs, including the startup programme.

Interview Insights: In this interview, the pre-selected list was presented and discussed. The interviewee mentioned the **number of pivots** to be very valuable for startup success; he argued that startups that pivoted more achieve a better PMF. When discussing the **number of founders**, the interviewee stated that although multiple founders typically achieve PMF in less time. The interviewee strongly suggests startups to **market as quick as possible**, startups that focus too much on the technology and not in the client/user have fewer chances of achieving PMF. **Competition** was also discussed and mentioned as an important variable; the interviewee states that startups that do not adequately identify and analyse competition have fewer chances of success. The variables of **culture**, **organicity** and **innovativeness** were discussed as problematic, either difficult to measure or susceptible to bias. The variables of **industry** and **business model** were discussed as difficult to categorize. For the dependent variables, the interviewee has no strong preference for any variable. The **break-even point** or **valuation** were mentioned as good alternatives.

Interviewee 6

Profile: Interviewee number six is a former entrepreneur with around ten years of experience. She has been part of the founding team of five companies. Aside from her experience as an entrepreneur, the interviewee has experience working with accelerators. She currently runs her consultancy firm and works together with Techstars mentoring early-stage startups.

Interview Insights: Interviewee six has a strong entrepreneurial experience; she strongly believes that failure is a startup's greatest mentor. The interviewee affirms that **team quality** together with **entrepreneurial experience** and **mentorship** are the most critical success factors. Team quality was described as intangible and difficult to measure. Regarding the **number founders**, the interviewee mentioned that many accelerators would hardly accept solo founders. The interviewee, discussing mentorship, emphasises the role of incubators and accelerators on startup success. Similar to previous interviewees, the **willingness to learn** and **motivation** of the founders was emphasised. The interviewee also mentioned the importance of having a significant customer base, concept related to the total addressable market(TAM). Regarding the dependent variables, the interviewee has a preference for **exits** (particularly IPO) as a measure of success.

Interviewee 7

Profile: The last interviewee is a startup founder, academic researcher and lecturer. It has more than twenty-five years of experience in the areas of strategy, program management and management consulting. The startup of the interviewee uses artificial intelligence and mathematical algorithms to reduce the failure of startups.

Interview Insights: Contrary to the previous interviews, interview seven was not recorded. Interview seven took place as an informal conversation that lasted several hours. As the topic of this research is aligned with the interviewee's research, the conversation was considerably more exhaustive than the previous interviews. The interviewee mentioned **founders' trust** as a critical factor of success. Another distinctive variable, included in the interviewee's research, is the creation of a **user persona**. The interviewee uses **revenue** and **revenue growth** in his research as the main criteria of success. The interviewee's startup, contrary to this study, employs non-supervised machine learning techniques. He categorises the variables in his research into actions & decisions, human capital, team capital and social capital.

3.2.2. Selection 1: Modifications

Insights drawn from the conducted interviews serve to refine and enrich the initial selection of success factors. In this subsection, two tables are presented. The first table 3.6 lists all the variables that have been removed from the pre-selection in subsection 3.1.2. The second table 3.7 lists all new factors that resulted from the interviews. In section 3.3 the final list of variables is delivered with detail.

EXCLUDED VARIABLES

For the following table, the numbers of the variables matches those from subsection 3.1.2.

Variable	Note
1. Founder's age	Age may be a success factor as it is related to knowledge and experience; this variable is left out as it is effects can be explained employing other variables.
2. Gender	Any difference in the success metrics, when comparing between genders, may be explained by complex societal phenomena which goes beyond the scope of this study. Results may be misleading and wrongly interpreted.
6-7. Industry and managerial experience	The effects of these variables will be studied by measuring overall working experience, to simplify the analyses and reduce the number of variables.
8. Interpersonal Skills	This variable is highly difficult to evaluate, it typically requires the measurement of multiple items.
14. Self-confidence	Like interpersonal skills, self-confidence is a problematic construct that requires the evaluation of multiple sub-factors.
15. Firm age	This variable is not suitable for success predictor. Firm age serves as a control variable. Companies with age between three to ten years are the subject of this study.
16. HQ Location	This variable is not suitable for success predictor. HQ Location serves as a control variable. Only companies HQ in the Netherlands are included in this study. A correlation between location and startup success may be wrongly interpreted.
17. Industry	The industry of a startup is part of their DNA, is related to what they do and can not be decided upon. A correlation between industry and startup success may be wrongly interpreted.
21. Networking	We believe that what matters the most is the connectedness of the individuals rather than the organisation.
22-23. Business and revenue model	Both variables are challenging to categorise, overlapping and not a determinant success predictor.
24. Scalability	Although this variables seem to be determinant when evaluating the potential of startups, this variable is hardly tangible.
26. Unique advantage	The definition of this variable, as read in the literature significantly overlaps with the variable of product innovativeness.
27. Technological hype	Not all variables in the sample can be linked to a specific technology. This variable only applies to high-tech startups actively working on breakthrough technologies.
29. Proof of value	As mentioned in source 3 Groenewegen and Langen [2012], proof of value refers to the customer base built during the early stages of the startup. As this number may significantly vary across industries, we have decided to exclude this variable and measure the effects of customer involvement with the variable of customer pro-activeness.
36. Market environment	A proper evaluation of the competitive environment requires thorough analyzes such as the one proposed in Porter's five forces model. To simplify this research, we decided to exclude this variable and measure the level of competitiveness by merely counting the number of identified competitors during the early stages.
37. Political stability	This variable was identified as hardly tangible and irrelevant.
40-41. Web-metrics	This variable seems to be a very good candidate as a success predictor. Nevertheless, the web-metrics of startups during their early stages is difficult to measure in this cross-sectional study.
42-43. Government & Family Support	Government support is mostly targeted to specific projects (e.g. health, energy). Family support was discussed to be much less influencing than the mentorship variable.

Table 3.6: Variables removed from 3.2, 3.3 and 3.4.

Added Variables

New variables are introduced in table 3.7. The final set of variables is presented in detail in section 3.3.

Variable	**Note**
Mutual trust	Variable proposed by interviewee 7. This factor can be estimated measuring the number of years the founders' have known each other for before startup launch.
Previous salary	Variable proposed by the researcher and supported by interviewee 3. Although this variable may present multicollinearity with work experience, this variable was discussed to be worth looking into as it may better represent general knowledge & experience.
University Ranking	Variable proposed by the researcher. This variable complements the variable of education level and it is measured according to the *Quarelli Simons* world university ranking.
Team diversity	Variable proposed by the researcher and supported by interviewee 3. This variable intends to analyse the correlation between diversity (in terms of gender, background, races and nationalities) and startup success.
Organicity	This variable intends to capture the intangible construct of startup culture, which was mentioned by interviewees 2, 3 and 5. Organicity refers to the organisational structure of a firm. An organic firm has a flexible working culture, operates in cross-functional teams, has a free flow of information, low formalisation and decentralised decision making.
Data orientation	Variable proposed by the researcher. Included to measure the influence of data usage in startup success.
Driving Force	Variable proposed by the researcher. This variable intends two compare the two driving forces of an innovation, market pull vs technology push.
Pivoting	This variable intend to capture the intangible factors mentioned by interviewees 2 and 6. Interviewee 2 referred to the *ability to adopt new insights*. Interviewee 6 referred to the *willingness to learn* of startup founders. As mentioned by interviewee 2, pivoting can serve as a proxy for this factor. Furthermore, interviewee 5 highlights the importance of pivots. He affirms they help startups achieve a better product-market fit.
Societal Relevance	Variable proposed by the researcher and inspired by the conversation with interviewee 4. Do startups that aim to solve relevant societal and economic problems perform better? This variable aims to respond to this question.
Time to market	Variable proposed by interviewee 2 and supported by interviewee 5. Although this variable may vary significantly across industries, it was mentioned by interviewee 5 that startups are suggested to market as soon as possible.
TAM Calculation	Variable suggested by an external agent (actor not interviewed). The purpose of this variable is to observe is startups that forecast their demand perform better.
Market Estimation	Variable suggested by the researcher. The purpose of this variable is to observe how many startups calculate their total addressable market, and if these perform better than those who do not.
Employee Incentives	Variable suggested by the researcher. The purpose of this variable is to analyse if startups that give equity to employees perform better as employees are motivated to achieve better results.
Marketing Persona	Variable included in the research of interviewee 7. This variable intends to analyse if the creation of a user/consumer persona influences startup success.

Table 3.7: New variables added to the framework.

In the first section of this chapter, desk research was conducted to deliver an initial set of variables. In the second section, we interviewed a total of seven knowledgeable actors in the Dutch startup ecosystem. Insights from the interviews were used to get a better understanding of the variables, discuss the selection, exclude and modify variables from the initial set, and add new variables. In the next section 3.3, we deliver the final selection of variables that is further explored in the qualitative analysis in chapter 4.

3.3. VARIABLES SELECTION 2

This section is divided into two subsections. In subsection 3.3.1, we deliver the final list of success predictors. In subsection 3.3.2, we select the criteria for startup success and discuss it briefly.

3.3.1. PREDICTORS SELECTION

The selection of independent variables is divided into tables 3.8 and 3.9. Taking into account the current selection of variables, these have been grouped into the categories of *human capital*, *business characteristics*, *third-party support* and *strategy & planning*. The column **ID** serves to enumerate the list of variables. The column **DS** identified the source of the data, if the value is DB this means that the value will be extracted from techleap.nl's database, if the value is Q followed by a number X, this means that the variable is measured through the questionnaire that is further discussed in chapter 4.2. The **Type** column indicates the type of variable (numeric, ordinal, categorical or binary); this information is very helpful when manipulating the data. The **Description** column provides a brief description of the variable, as most of the variables are obtained through the questionnaire, we suggest the reader consult the appendix A and have a look at the question related to the variable for a full understanding of how the variable is evaluated. The reference column indicates the source of the variable, if the value is S1-S6 this refers to the literature reviewed in section 3.1.1, if the value is I1-7 this refers to the interviewees conducted in section 3.2, if the value is *author* this means the variable was proposed by the author of this research, one variable has the value of *external*, this means the variable was proposed by an individual who was not interviewed for this thesis.

	ID	DS	Variable	Type	Description	Reference
Human Capital	1	DB	Number of Founders	Numeric		S3
	2	Q1	Mutual Trust	Numeric	Number of years the founder's have known each other for.	I7
	3	Q2	Work Experience	Numeric	Number of years worked before launching the startup.	S2,S3,S4
	4	Q9	Previous Salary	Numeric	Previous salary before launching the startup.	Author
	5	Q3	Entrepreneurial Knowledge & Experience	Ordinal	Measured with a Likert Scale.	S1,S2,S3,S4
	6	Q4	Education Level	Ordinal	Six categories from Elementary to PhD.	S2,S3,S4
	7	Q5	University Ranking	Ordinal	Based on QS world university rankings.	Author
	8	Q6	Field of Study	Categorical	Five categories based on the QS ranking's field of study	Author
	9	Q7	Social Capital	Ordinal	Likert agree/disagree scale.	S3
	10	Q8	Time Dedication	Binary	Full time vs part time.	S4
	11	Q10	Risk Profile	Ordinal	Measured with a Likert Scale.	S2,S3
	12	Q11	Motivation	Categorical	Six Categories	S2
	13	Q12	Team Size	Numeric	Number of employees before market introduction.	S1,S4
	14	Q13	Team Diversity	Ordinal	Measured with a Likert Scale.	Author

Table 3.8: Selection of Independent variables table 1 out of 2.

	ID	DS	Variable	Type	Description	Reference
Business Characteristics	15	DB	B2B vs B2C	Binary	Self explanatory	S1
	16	Q14	Organicity	Ordinal	Measured with a Likert Scale	I2,I3,I5
	17	Q15	Tech Orientation	Ordinal	Five Levels from tech-enabled to tech-driven	S2,S4
	18	Q16	Data Orientation	Ordinal	Five Levels from data resistant to data-driven	Author
	19	Q18	Driving Force	Binary	Market Pull vs Technology Push,	Author
	20	Q22	Ambition to Grow	Ordinal	Measured with a Likert Scale.	S1
	21	Q26	Pivoting	Numeric	Self explanatory	I2,I6
	22	Q19	Societal Relevance	Ordinal	Measured with a Likert Scale	Author
	23	Q17	Innovativeness	Ordinal	Measured with a Likert Scale	S1,S3
	24	Q25	Outperformance	Ordinal	Measured with a Likert Scale	S3
	25	Q23	Competition	Numeric	Number of identified competitors	S1
Support	26	Q20	Funding Source	Categorical	Five Categories (e.g. angel)	S1,S2,S3,S4
	27	Q21	Mentorship	Categorical	Five Categories (e.g. accelerator)	S3,S4
	28	Q24	Alliances	Numeric	Number of alliances (horizontal,upstream or downstream))	S4
Strategy & Planning	29	Q27	Time to Market (T2M)	Numeric	Time elapsed between launch and market introduction.	S2,S5
	30	Q28	Money to Market (M2M)	Numeric	Money consumed between launch and market introduction.	S1,S3,S4
	31	Q29	MVP	Ordinal	Four Levels describing the use of POC & Prototypes before MVP	S1,S4
	32	Q30	Communication Tool	Categorical	Describes the use of various tools to communicate the startups' ideas and planning (e.g. Business Plan, Canvas, Pitch Deck, others).	S2,S4
	33	Q31	Forecasted Demand	Binary	Forecasted the Demand? (Yes/No)	External
	34	Q32	Market Estimation	Binary	Calculated the TAM? (Yes/No)	Author
	35	Q33	Employee Incentives	Binary	Equity to Employees (Yes/No)	Author
	36	Q34	Marketing Persona	Binary	Created a Marketing Persona? (Yes/No)	I7
	37	Q35	Customer Proactiveness	Ordinal	Measured with a Likert Scale.	S3

Table 3.9: Selection of Independent variables table 2 out of 2.

3.3.2. Criteria Selection

We explore four variables as startup success criteria: *total funding, employees, profitability* and *revenue growth*. The first two variables are obtained from techleap.nl's database, whereas the latter two are collected through the questionnaire. Total funding is measured as a dichotomous variable, and a startup is considered successful if this has above one million in total funding. The employees variable is also dichotomous, and a startup is considered successful if this has ten or more employees. The variables of profitability and revenue growth are measured as a yes/no question in the questionnaire. A startup is profitable if this has reached the break-even point. A startup is said to have rapid growth if this matches the criteria of the scaleup definition.

3.4. CONCEPTUAL MODEL

This and the following section were added at the end of this research and do not appear in the general planning. The purpose of this section is to facilitate the connection between the selected variables and results obtained and summarized in tables 5.2 and 5.3 in the discussion section. As it is not practical to present here an extremely large list of hypotheses for all the considered variables, we must disclose part of the results obtained in this research. Summarised, at the end of this research we deliver a total of three predictive models based on eight success predictors. In figure 3.3 the conceptual diagram for the first model is presented.

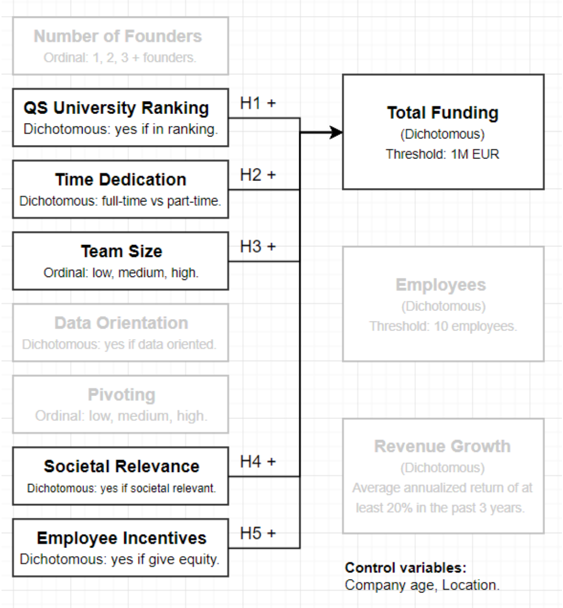
Number of Founders
Ordinal: 1, 2, 3 + founders.
QS University Ranking
Dichotomous: yes if in ranking.
H1 +
Total Funding
(Dichotomous)
Threshold: 1M EUR
Time Dedication
Dichotomous: full-time vs part-time.
H2 +
Team Size
Ordinal: low, medium, high.
H3 +
Employees
(Dichotomous)
Threshold: 10 employees.
Data Orientation
Dichotomous: yes if data oriented.
Pivoting
Ordinal: low, medium, high.
Revenue Growth
(Dichotomous)
Average annualized return of at least 20% in the past 3 years.
Societal Relevance
Dichotomous: yes if societal relevant.
H4 +
Employee Incentives
Dichotomous: yes if give equity.
H5 +
Control variables:
Company age, Location.

Figure 3.3: Conceptual Diagram for Model 1.

Although more detail will be given on how did we arrive to these results, what is important to notice here is the conjectural relationship between the different success factors and criteria. For this first model, it was hypothesized that the variables of *university ranking, time dedication, team size, societal relevance* and *employee incentives*; were all positively correlated to the criteria of total funding. In figure 3.4, the conceptual diagram for the second model is presented.

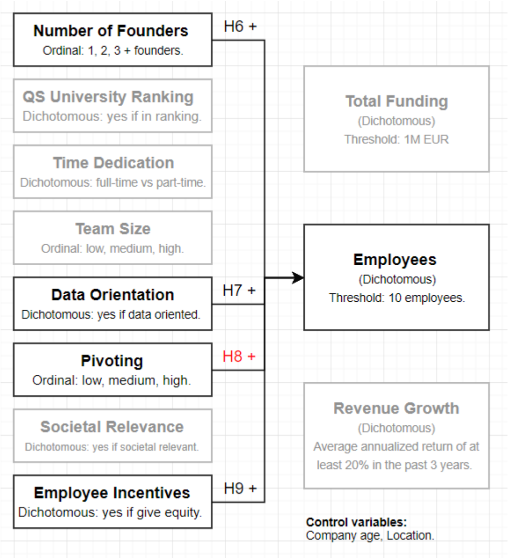
Number of Founders
Ordinal: 1, 2, 3 + founders.
H6 +
QS University Ranking
Dichotomous: yes if in ranking.
Total Funding
(Dichotomous)
Threshold: 1M EUR
Time Dedication
Dichotomous: full-time vs part-time.
Team Size
Ordinal: low, medium, high.
Employees
(Dichotomous)
Threshold: 10 employees.
Data Orientation
Dichotomous: yes if data oriented.
H7 +
Pivoting
Ordinal: low, medium, high.
H8 +
Revenue Growth
(Dichotomous)
Average annualized return of at least 20% in the past 3 years.
Societal Relevance
Dichotomous: yes if societal relevant.
Employee Incentives
Dichotomous: yes if give equity.
H9 +
Control variables:
Company age, Location.

Figure 3.4: Conceptual Diagram for Model 2.

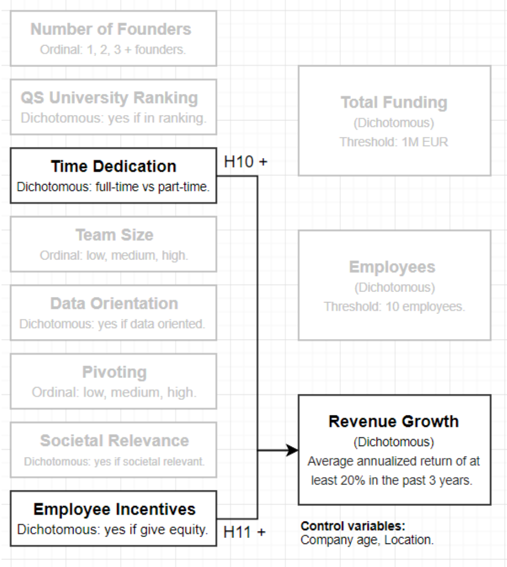
Number of Founders
Ordinal: 1, 2, 3 + founders.
QS University Ranking
Dichotomous: yes if in ranking.
Total Funding
(Dichotomous)
Threshold: 1M EUR
Time Dedication
Dichotomous: full-time vs part-time.
H10 +
Team Size
Ordinal: low, medium, high.
Employees
(Dichotomous)
Threshold: 10 employees.
Data Orientation
Dichotomous: yes if data oriented.
Pivoting
Ordinal: low, medium, high.
Societal Relevance
Dichotomous: yes if societal relevant.
Revenue Growth
(Dichotomous)
Average annualized return of at least 20% in the past 3 years.
Employee Incentives
Dichotomous: yes if give equity.
H11 +
Control variables:
Company age, Location.

Figure 3.5: Conceptual Diagram for Model 3.

For the second model, a total of four independent variables proved to be significant predictors of a startup having ten or more employees. It was hypothesized that the variables of number of founders, data orientation, pivoting and employee incentives; were all positively correlated to the criteria of total funding. As it will be explained in subsection 5.4, the pivoting variable contrary to what was discussed with the interviewee I6, shows to have a negative correlation with the dependent variable. In figure 3.5, the conceptual model for the third model is presented.

For the third model, two variables proved to be significant predictors of a startup having a revenue growth according to the scaleup definition. It was hypothesized that the two variables of time dedication and employee incentives were positively correlated to the criteria of revenue growth. In the next section we articulate the eleven hypotheses which can be observed in the three conceptual diagrams presented for the three predictive models built.

3.5. HYPOTHESES

As it can be observed in the conceptual models of the previous subsection, a total of eleven hypotheses could be articulated for the relationship between the startup success predictors and criteria. The first five hypotheses pertain to the model of total funding, hypotheses six to nine pertain to the model on the number of employees and the last two hypotheses belong to the model on revenue growth.

1. Startups whose founders studied in a top-ranked university have greater chances of having total funding above one million euros.
2. Startups with full-time dedicated founders have greater chances of having total funding above one million euros.
3. Startups with larger core teams have greater chances of having total funding above one million euros.
4. Startup focused towards societal relevance have greater chances of having total funding above one million euros.
5. Startups that give equity to their employees have greater chances of having total funding above one million euros.
6. Startups with more than one founder have greater chances of having ten or more employees.
7. Startups which are data-oriented have greater chances of having ten or more employees.
8. Startups that pivot more have greater chances of having ten or more employees.
9. Startups that give equity to their employees have greater chances of having ten or more employees.
10. Startups with full-time dedicated founder have greater chances of having an average annualised return of at least twenty per cent in the past three years.
11. Startups that give equity to their employees have greater chances of having an average annualised return of at least twenty per cent in the past three years.

Although the results of the constructed models are presented in section 4.6, these are better summarized in section 5.4. In this section, we briefly discuss the hypotheses here presented and we will discover that almost all of them are confirmed except for the hypothesis number eight which showed a negative relationship.

"Waiting for perfect is never as smart as making progress."
- Seth Godin, Author

4

Predictive Model

In this chapter, all quantitative research is presented. This chapter starts with the selection of the sample and ends with testing the predictive models on startup success. In section 4.1, the data set exported from techleap.nl's database is filtered to estimate the population size (4.5 thousand companies) and the required sample size (at least 73). In section 4.2, we explain how the data was collected from startup founders through a questionnaire, and we provide a basic description of the characteristics of the sample. After delivering the questionnaire to over 950 companies, we obtained 91 responses. These responses were divided into 74 companies to train the model and 17 companies to test it. In section 4.3, all the process of data preparation is explained, from data integration to data transformation. Once the data is prepared, the current selection of variables suffers some modifications and the new selection of variables is provided in section 4.4. A total of twenty-eight independent variables and three dependent variables are selected for the construction of the predictive models. Three predictive models are built to predict total funding (above EUR 1M), employees (above ten) and revenue growth (according to the scaleup definition). In section 4.5 an overview of logistic regression is provided, this is the technique we used to build the predictive models and is essential that the reader fully understands the principles. In section 4.6, the three predictive models are constructed and their predictive power is assessed. Lastly, in section 4.7 the three predictive models are tested using the set of 17 companies, and the confusion matrices are delivered. The results of the later two sections are summarized and discussed in sections 5.4 and 5.5 of the discussion chapter.

4.1. Sampling

The initial data set was exported from techleap.nl's database (available in) on 11 June 2019. The set exported contained 7723 rows and 92 columns. This data set was reduced implementing the definitions discussed in section 2.2. The filtering process was conducted as follows:

1. HQ's Location: 7331 startups & scaleups. A total of 391 companies with HQ's outside of the Netherlands are excluded from the data set. These companies are included in the database because they have offices and a strong presence in the Netherlands. These companies perform significantly different than companies that are headquartered in the Netherlands, and hence we exclude them from our analysis.

2. Subsidiaries: 7007 startups & scaleups. A total of 324 companies are excluded from the previous set. These companies are either owned or controlled by another larger, more mature, company or corporation.

3. Company age: 4553 startups & scaleups. A total of 2454 companies founded before 2009 are excluded from the previous set. As discussed in subsection 2.2.1 a company should no longer be labelled as a startup if it is older than ten years old.

By implementing the startup definitions established in subsection 2.2.1 we have filtered a total of 2778 companies from the database. It can be inferred that the population size of startups & scaleups in the Netherlands is around 4.5 thousand entities. For a population of this size, with a confidence level of 99% and confidence interval of 15% it was calculated that the required sample size must be above 73 entries.

4. Company size: 4517 startups & scaleups. A total of 36 companies with more than 250 employees are excluded from the previous set after excluding companies that have more than 250 employees. As it was discussed in subsection 2.2.1, a company is categorised by the OECD as a large corporation once it surpasses this threshold.

5. Company age: 3898 startups. A total of 619 startups founded in the last two years have been excluded from the previous set. It has been observed that the count of companies significantly decreases after 2017; this occurs because these companies are very young and they have not got into the radar. To this phenomenon, techleap.nl refers to as reporting lag. Furthermore, as this research intends to evaluate the success of startups, it is reasonable to exclude young companies that are still in their early stages.

6. Founders' information: as we intend to collect data directly from the startup founders, we need to know who the founders are. From the previous set, we have the founders' information from a total of 1491 startups & scaleups.

From a total of 1491 startups & scaleups, we identified a total of 2255 founders. From this set, a total of 1183 emails from 969 startups & scaleups were collected. Besides the desired sample size of 73 startups to train the predictive model, extra responses are needed to test the model. A typical train/test ratio is machine learning classification models is 80/20. Taking into account this ratio, we need at least 92 total responses (73 train/19 test), a response rate above 9.5% from the total number of startups reached. Email data was collected using *rocket reach* and *hunter* add-ins for google chrome.

4.2. DATA COLLECTION

4.2.1. QUESTIONNAIRE

Most of the data used for the predictive model was collected through the administration of questionnaires to startup founders. This questionnaire was created using google forms, and it was delivered to 969 startups & scaleups. With a response rate of 10.2%, a total of 91 data points were collected (one below the minimum required). In appendix A, the full list of questions from the questionnaire is also displayed. Reading the questionnaire is essential for the reader, as they can relate the questions to the variables presented in section 3.3. To check the questionnaire in full detail please consult the following URL: `https://forms.gle/tFGsh3atRm9UCdqPA`.

4.2.2. SAMPLE CHARACTERISTICS

For the sample characteristics, we included all 91 responses. As it was mentioned before, companies in this study were launched between 2009 and 2017. The lower limit is selected according to the startup definition provided in section 2.2.1. The upper limit excludes young startups and avoids the reporting lag. In figure 4.1 the distribution of companies by launch date can be observed.

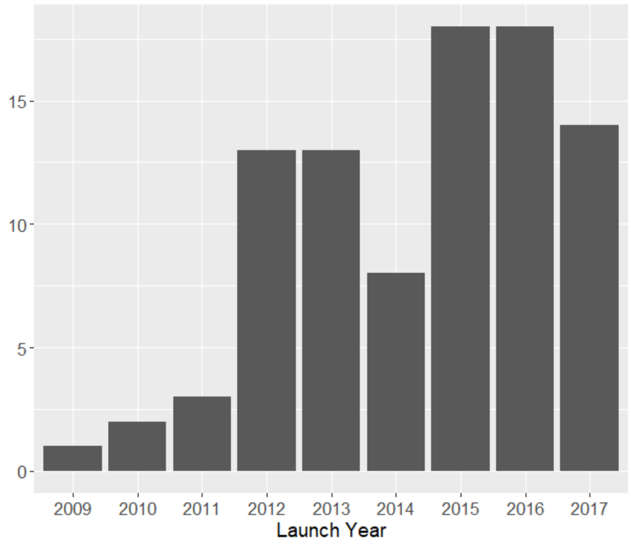

Figure 4.1: Launch Year of the Companies in the Sample.

As it can be noticed, the data is left-skewed. There tend to be more startups & scaleups in recent years. The apparent increase in companies can be attributed to the following reasons: (1) the number of startups founded have been increasing in the last decade (2) the older the startup is, the larger the chances that this is no longer active, (3) if the company is still operating, the founders may be less interested or busy to participate in this study. In table 4.1, the count of startups per industry is listed.

Industry	**Count**
Information Technology & Services	10
Health	9
Internet	7
Renewables & Environment	6
Education	6
Computer Software	6
Marketing	5
Manufacturing	5
Travel	4
Apparel & Fashion	4
Others	29
Total	**91**

Table 4.1: Count of companies by industry in the sample.

In section 3.2.2 the variable *industry* was excluded from the selection of success predictors. In our sample, we include startups from all sorts of industries as shown in table 4.1. Industries with only one or two companies are grouped into the *others* category. Some industries are more research-intensive (e.g. health) compared to other industries (e.g. internet). We expect the influence of these industry-dependencies to average out, causing a minimal influence on the proposed model. Lastly, as the scope of this research is to analyse the startups & scaleups in the Netherlands. The list of companies per city is listed in table 4.2. As it can be noticed the great majority is headquartered in Amsterdam. Other clusters can be identified around the cities of Eindhoven, The Hague-Delft-Rotterdam and Utrecht.

Industry	**Count**
Amsterdam	32
Delft	9
Utrecht	8
Eindhoven	7
The Hague	4
Rotterdam	3
Haarlem	3
Enschede	3
Wageningen	2
Naarden	2
Groningen	2
Leiden	2
Others	14
Total	**91**

Table 4.2: Count of companies by HQ location.

4.3. Data Preparation

For proper analysis and model construction, data needs to be thoroughly prepared. Data preparation is formed by several sub-processes including integration, enrichment, import, cleaning, transformation.

4.3.1. Data Integration

The primary sources of data for this research are techleap.nl's database and the data collected through the questionnaire. The data from both sources are integrated into one single table with forty-six columns and ninety-one rows. Columns one to five contain identifying and demographic data from the startups (name, ID, HQ location, industry, launch year). Columns six and seven are independent variables obtained from techleap.nl's database (Number of Founders and B2B vs B2C). Columns eight to forty-two contain the independent variables collected through the questionnaire. Columns forty-three and forty-four dependent variables from techleap.nl's database (total funding and employees). The last two columns contain the dependent variables collected from the questionnaire (Break-Even and Scaleup). A total of thirty-seven independent variables and four candidate dependent variables are included in this current selection of variables.

4.3.2. Data Enrichment

Before the data is analysed and a model can be built, we need to ensure the quality of the data, particularly that of the dependent variables. When observing the values of *total funding*, a lot of missing values are identified. Because of the reduced size of the sample, removing the rows with missing values is not the desired option. Missing values are often replaced with descriptive statistics such as the mean or the median. These replacements may not be the best alternative for replacing missing values in total funding as the dispersion in the data is considerably big. Instead, the researcher replaces these missing values with an educated estimate. We also noticed that the employee values in the database are sometimes overestimated. The processes of how the data for these two dependent variables were enriched is described next.

- **Total Funding**: In the case that the round type is known, the empty value was replaced by the median of the population (e.g. seed round EUR 0.1M, Series A EUR 1.1M). In the case that the round type is unknown, but the investor is known, the empty value is replaced by the average deal size of the investor. The data about investors and rounds can also be consulted from techleap.nl's database in https://finder.startupdelta.org.

- **Number of Employees**: The employee count is estimated from the number of LinkedIn user profiles connected to the company's LinkedIn. This method can display inflated values that do not represent real TFE. To correct the data, the researcher verified that the number of employees lied in between the employee range defined by the company on their LinkedIn profile. If the number falls out of the range, the number is replaced by the maximum.

Besides the dependent variables previously mentioned, the variable of *university ranking* was also preprocessed and enriched. In the questionnaire, respondents were requested to type the name of their educational institution. The researcher assigned a score from cero to one hundred for each educational institution according to the Quarelli Simons world university rankings. The ranking can be consulted in https://www.topuniversities.com/university-rankings.

4.3.3. Data Import into R Studio

To this point, the data is stored in an excel spreadsheet, containing 91 rows and 46 columns. Before the data can be further processed, this is imported into the RStudio environment. When the data is imported, the variables of *ID*, *HQ location* and *industry* are excluded. These variables are excluded because they do not take any role in the construction of the predictive model; the variable name is kept as the only identifying variable. Furthermore, after reviewing the responses from the questionnaire, we decided to exclude the variables of *field of study* (Q6), *innovativeness* (Q17), *driving force* (18), *out-performance*(19), *ambition to grow*(23) and *competition* (24). After importing the data in R Studio, the data frame is formed by 32 independent variables, 4 dependent variables and one identifying variable. The 91 responses are divided into a train set of 74 data points and a test set of 17 data points. All the data preparation tasks here presented are conducted only on the training set.

Next step, before the data can be further processed, is to convert the data types of the imported variables. The imported data consist of numeric and character type columns. All character type columns (except *company name*) need to be converted into categorical type data (also known as factors in R). The variables *B2B vs B2C*, *Time Dedication*, *Motivation*, *Funding Source*, *Mentorship*, *MVP*, *Communication Tool*, *Forecasted Demand*, *TAM.Calculation*, *Employee Incentives*, *Persona*, *Break.Even* and *Scaleup* are converted into categorical (factor) data type. After the transformation, the data set consist of one character column, eleven independent categorical variables, two dependent categorical variables, twenty-one independent numeric variables and two dependent numeric variables.

4.3.4. Data Cleaning

Once the data is imported and converted into the adequate data type, this needs to be cleaned before it can be transformed and further analysed. In this subsection, two tasks are carried out, the replacement of missing values and outliers. As it was mentioned before, due to the size of the sample, removing rows is not the desired option. Empty spaces and outliers need to be replaced with an adequate value. In the following two subsubsection the process of data cleaning is clearly explained.

Missing Values

To observe the sparsity of the data, we used the *Amelia* package in R studio, this tool allows us to observe the data sparsity graphically. We observed that a total of ten variables contained missing values, a 2% sparsity. The variable with the most substantial amount of empty values was *university ranking*, assuming that a university not in the QS ranking has a score of zero, we replaced all empty values for zeros. This replacement could be a problem because this ranking includes comprehensive universities but not very specialised ones (e.g. MBA schools). For all the remaining variables, empty values were replaced by the median. The mean was not used as this showed to be significantly affected by the outliers. At the end of this task, all the set was free of empty values with a data sparsity of 0%.

Outliers

The next step in the data cleaning process is to remove outliers in the data. This task is conducted only in those variables that are entirely numeric. Despite that at this point the data frame consists of twenty-three numeric variables, many of these are ordinal and measured through the use of Likert scales in the questionnaire. These variables cannot have outliers as they are bounded by the limits of the scale they are measured. A total of ten variables are not measured using a Likert scale: number of founders, trust, work experience, university ranking, salary, core team size, alliances, pivoting, T2M and M2M. The QS Score variable is not taken into account for the outliers identification as this is bounded by the scale range (0 to 100).

As a first approach to identify outliers, we draw the histograms of the ten numeric variables. When observing the histograms of the these, it was clearly observed that the variable *number of founders* had no outliers and the variables of *alliances*, *money to market* (M2M) and *pivoting* had outliers. For the remaining five variables we used box-plots to identify their outliers. As it can be seen from figure 4.2, the only variable without outliers is *work experience*, all other variables had their outliers values replaced by the median.

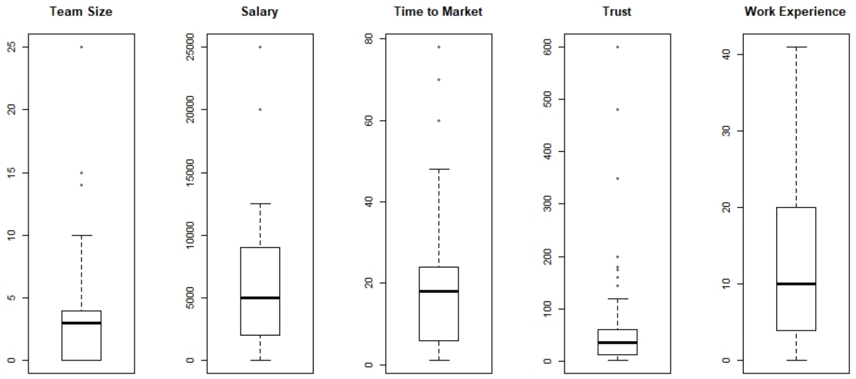

Figure 4.2: Identification of outliers from numeric variables.

4.3.5. Data Transformation

In this subsection three activities of data transformation are carried out. The purpose of this section is to group data into different levels or categories so that the data is distributed uniformly. We want the data to be distributed this way because for the logistic regression we need sufficient data in each of the level or category. In the first part of this subsection, we deal with categorical variables; three categorical variables are excluded from the data set, and one is transformed into a dichotomous variable. In the second part, we deal with the nine numeric variables; these are discretized into bins or clusters that make the data uniformly distributed. Seven numeric variables are reduced to three levels, one to four levels and one to five levels. Lastly, we deal with the ordinal variables, all seven ordinal variables are transformed into dichotomous variables.

Categorical Variables

To this point the data set contains a total of four categorical variables. From these, *motivation* has a total of six categories, the variable *funding source* ten categories, *mentorship* eight categories and *MVP* four categories. Besides these four variables, one variable was codified before being imported into RStudio. The variable *communication tool*, with originally ten categories, was reduced to a dichotomous variable with values of "business plan" or "pitch deck, canvas & others".

The variables of mentorship, funding source and motivation were eliminated from the data set. This action was taken because we did not observe any easy way of grouping the different categories. The variable MVP, with initially four levels, was coded into a dichotomous variable with the levels "no prototype or POC" and "prototype or POC". These last variable intend to capture the idea that startups that test their ideas before development may perform better. To this point, the data set contains seventy-four rows and thirty-three columns. One identifying variable, twenty-eight independent variables and four dependent variables.

Numeric Variables

The next step is the discretisation of numeric variables. To achieve this, we used the function *bin* from the *OneR* library in RStudio. This function groups the numeric variables into several bins as defined by the user. The binning is carried out through different methods; in this case the method "content" is selected. The content method groups the data into intervals of equal content. To determine the number of bins, the researcher observed the histograms of the variables and qualitatively determined which was the best way to group the data. The histogram of the variables is not displayed in this thesis as it does not add any valuable information to the predictive model and on the contrary it can confuse the reader. For a detail description of all these processes refer to the R code in appendix B.

- **Alliances:** the responses from the *alliances* variable were binned into three levels: "low" level which contain 0 recognized alliances, "medium" level contains values of 1-2 alliances and "high" with 3 or more recognized alliances.
- **Team Size:** the *time size* values have also been binned into three levels: "low" for 0 or 1 persons, "medium" for a count of 2 to 3 persons and "high" for 4 or more persons.
- **Money to Market (M2M)**: The values of M2M have been categorized into three levels: values between 0 and 32K+ are labelled as "low", between 32K and 75K as "medium" and 75K+ as "high". All values in thousands of euros.
- **Number of Founders:** this variable have been divided into three levels: 1 founder, 2 founders and 3 or more founders.
- **Pivoting**: the data in the pivoting variable was grouped into three levels: "low" level for 0 pivots, "medium" for 1-2 pivots and "high" for 3 or more pivots.
- **Previous Salary**: the *salary* variable was divided into five levels: "low" for monthly salaries below "1.5K", "average" for salaries between "1.6K and 3K", "above average" between 3.1K and 5K, "high" between 5.1K and 9K, and "very high" for 9.1K and above. The name of the levels serves only to order the categories. This means that "above average" is solely speculative.
- **Time to Market (T2M)**:the time to market variable was divided into four levels: "fast" for a time to market below 8 months, "average" between 9 and 18 months, and "slow" for 19 months or above.

- **Mutual Trust**: the trust variable was divided into three levels: "low" for founders that know each other for less than 20 months, "medium" between 21 and 36 months, and "high" above 37 months.
- **Work Experience**: the work experience variable has been divided into four levels: "low" for experience between 0 and 4 years, "medium" 5 to 10, "high" 11-19 years and "very high" above 20 years of working experience.

Ordinal Variables

In a similar way numerical variables were discretised, we collapsed the ordinal variables into dichotomous variables. To transform an ordinal variable into dichotomous, a threshold must be defined. We defined the threshold according to the median; this way, we guaranteed that both levels in the new dichotomous variable have similar amount of responses. Depending on the variable the levels are labelled differently. For a full understanding of how the current selections of variables are measured, please refer to the next subsection.

4.4. Variables Selection 3

Up to this point, due to the activities carried out in the data preparation section, the initial set of independent variables listed in table 3.9 underwent numerous transformations. In the following table, the set of variables that are selected prior to the analysis is listed once more.

4.4.1. Predictors Selection

	ID	DS	Variable	Type	Description
Human Capital	1	DB	Number of Founders	Ordinal	1 founder, 2 founders or 3+ founders.
	2	Q1	Mutual Trust	Ordinal	Time the founders have known each other for. If < 20, "low", 21-36 "medium" and > 37 "high".
	3	Q2	Work Experience	Ordinal	Labeled "low" if < 4 years, "medium" 5 to 10, "high" 11 to 19 and "very high" if > 20.
	4	Q9	Previous Salary	Ordinal	Labeled "low" if < EUR 1.5K, "average" if between 1.6K and 3K, "above average" 3.1K - 5K, "high" 5.1 - 9K and "very high" if > 9.1K.
	5	Q3	Entrepreneurial Know & Exp	Binary	High vs low.
	6	Q4	Education Level	Ordinal	Six levels from elementary to PhD.
	7	Q5	University Ranking	Binary	Categories: yes/no. A university is considered a "top university" if within the QS ranking.
	8	Q7	Social Capital	Ordinal	High vs low.
	10	Q8	Time Dedication	Binary	Full time vs part time.
	11	Q10	Risk Profile	Binary	Categories: yes/no. Yes if the founder has a risk-seeking profile.
	12	Q12	Team Size	Ordinal	Labeled as "low" if 0 or 1 persons, "medium" if 2 or 3 persons and "high" if 4 or more.
	13	Q13	Team Diversity	Binary	Categories: yes/no. Yes if the team is diverse.
Company	14	DB	B2B vs B2C	Binary	B2B vs B2C
	15	Q14	Organicity	Binary	High vs low.
	16	Q15	Tech Orientation	Binary	Categories: yes/no. Yes if tech oriented.
	17	Q16	Data Orientation	Binary	Categories: yes/no. Yes if data oriented.
	18	Q26	Pivoting	Ordinal	Labeled as "low" if no pivots, "medium" for 1-2 pivots and "high" for 3 or more pivots.
	19	Q19	Societal Relevance	Ordinal	Categories: yes/no. Yes if stated societal relevance.

	ID	DS	Variable	Type	Description
Strategy & Planning	20	Q24	Social Capital	Ordinal	Labeled "low" if no alliances, "medium" if 1-2 and "high" if 3 or more.
	21	Q27	Time to Market (T2M)	Ordinal	Labeled "fast" if T2M below 8 months, "average" if between 9 to 18 months and "slow" if above 19 months.
	22	Q28	Money to Market (M2M)	Ordinal	Labeled "low" if between 0 to 32K euros, "medium" if between 32 and 75K and "high" if 76K or above.
	23	Q29	MVP	Binary	Two categories: "prototype or poc" vs "no prototype nor poc".
	24	Q30	Communication Tool	Binary	Two categories: "full business plan" vs "pitch deck, canvas & others".
	25	Q31	Forecasted Demand	Binary	Forecasted the Demand? (Yes/No)
	26	Q32	Market Estimation	Binary	Calculated the TAM? (Yes/No)
	27	Q33	Employee Incentives	Binary	Equity to Employees (Yes/No)
	28	Q34	Marketing Persona	Binary	Created a Marketing Persona? (Yes/No)

Table 4.3: Third selection of independent variables.

As it can be seen from table 4.3, the original list of 37 variables in table 3.9 has been reduced to a list of 28 variables which will be used for the predictive modelling. As it was discussed in subsection 3.1.3, there were various possibilities of dependent variables (success predictors), in the following subsection we proceed to graphically explore this and have a better understanding on how these are related to one another.

4.4.2. CRITERIA SELECTION

For dependent variables, we have at the moment two numeric and two dichotomous variables. It is essential to take into account that the revenue growth variable is measured as a dichotomous variable according to the definition of scaleup. According to this definition, besides having a rapid revenue growth, a scaleup also has more than ten employees and cannot be older than ten years old. Taking this into consideration, we proceed to plot the two numeric variables in a scatterplot.

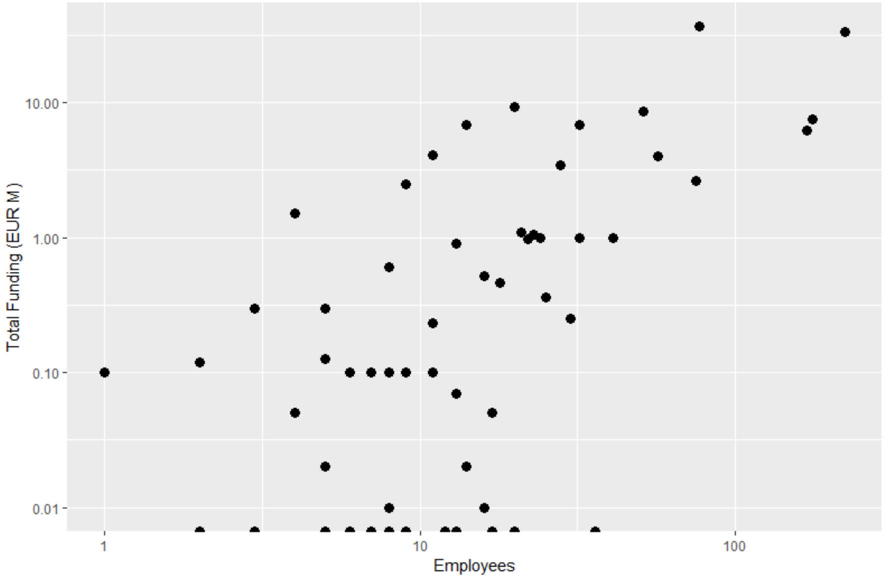

Figure 4.3: Total Funding vs Number of Employees.

From the scatterplot, where both numerical variables are plotted in a logarithmic axis, a positive correlation can be observed. To confirm this, we calculated the correlation between the variables. With a p-value of 0 a correlation of 0.67 was obtained. Next, we proceed to explore the two dichotomous variables, revenue growth and profitability. In table 4.4 the number of occurrences for both variables is shown.

	Rapid Growth (YES)	Rapid Growth (NO)	**Total**
Profitable (YES)	11	16	27
Profitable (NO)	11	36	47
Total	22	52	74

Table 4.4: Number of Occurrences for Success Criteria.

Before collecting the data, the results obtained in table 4.4 were not expected by the researcher. The fact that the data is equitably distributed across all boxes highlights that a rapidly growing business is not necessarily a profitable one. Businesses that present rapid growth typically incurred in significant investments and achieving profitability is more challenging than for a regular SME. Companies that achieve profitability but do not have an accelerated growth may share characteristics and attributes of an SME.

Despite it is out of the focus of this research, the table above and the previous discussion may be very interesting for future studies. As mentioned, it can be argued that achieving profitability with an accelerated growth may be a sign of a more conservative approach and it can be hypothesised that these companies share attributes common of a successful SME. To observe this more clearly let's plot the *profitability* dichotomous variable on top of the *total funding* vs *employees* scatterplot.

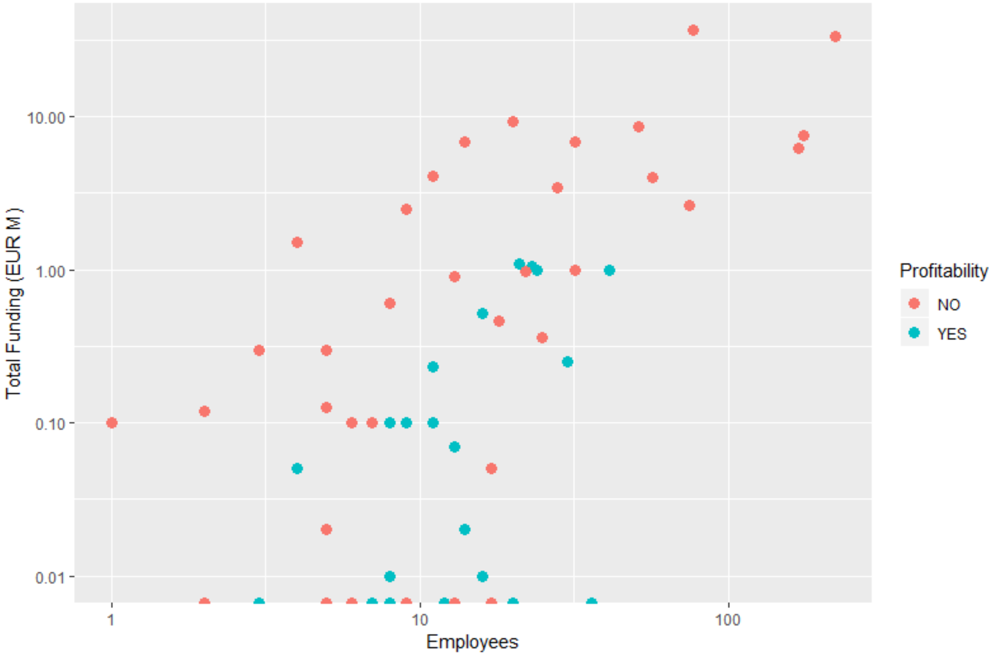

Figure 4.4: Profitability vs Total Funding vs Employees.

It can be clearly seen in figure 4.4 that the profitable startups (green) can be enclosed in the lower left quadrant, limited by the one million euros funding and fifty employees horizontal and vertical lines respectively. Next, in figure 4.5, we present the dichotomous variable *revenue growth* on top of the *total funding* vs *employees* scatterplot.

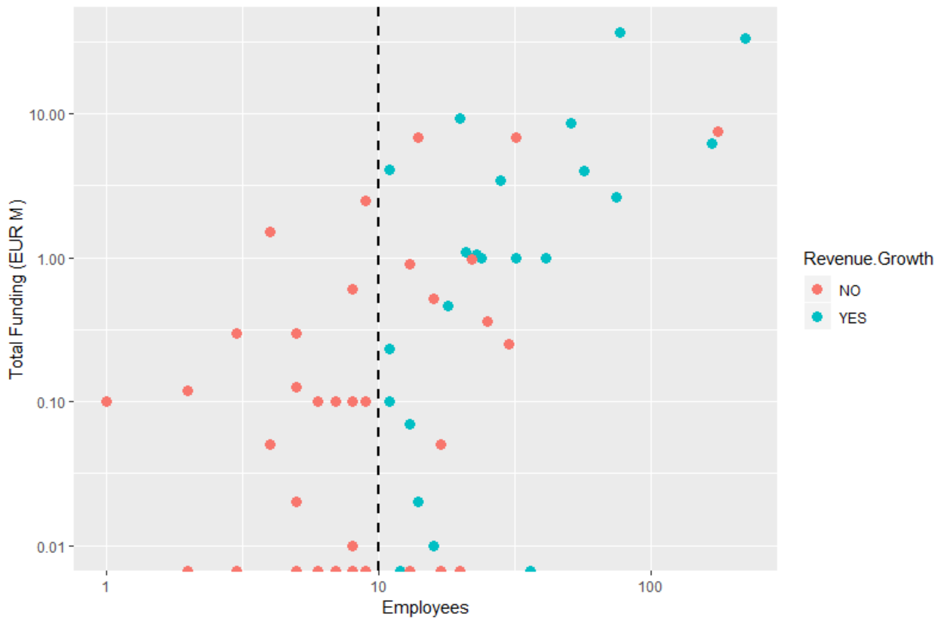

Figure 4.5: Profitability vs Total Funding vs Employees.

By definition, a scaleup has more than ten employees, and that is why all green points in figure 4.5 are located on the right side of the chart. Hence, for this plot, it is essential to observe the data only to the right of this threshold. It can be seen that most of the data points located on the upper right corner, those with a large funding and employment, have a rapid revenue growth. Furthermore, the ratio of green to red points seem to be larger towards the right side (50+ employees). To corroborate this information we proceed to draw the scatterplot of each of the numeric dependent variables against the rapid growth variable. These type of scatterplots (dichotomous on the y-axis and continuous on the x-axis) are typical of logistic regressions.

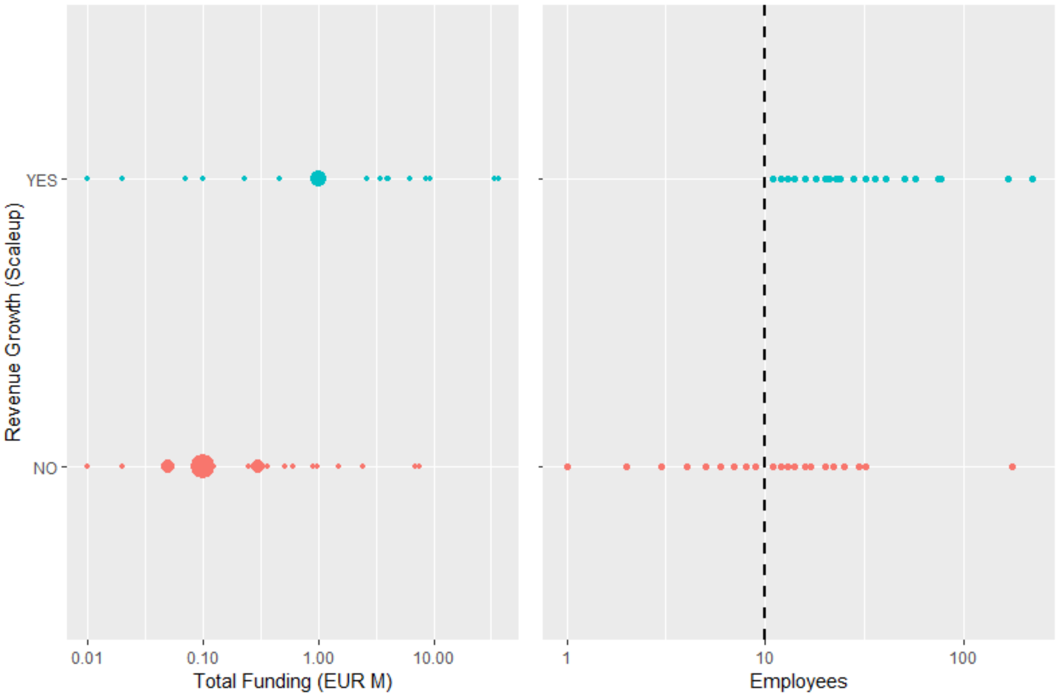

Figure 4.6: Rapid Growth vs Total Funding (left), Rapid Growth vs Employees (right).

As it can be seen from the plots in figure 4.6 there exists a relationship between both numeric variables (total funding and employees) with the dichotomous variable representing rapid growth (scaleup). On the left, although there exist rapid growth companies across all levels of funding, a slightly higher concentration of these is found towards the right side of the chart. The fact that there are green points on the left side highlights that a bootstrapped company can also achieve rapid growth. On the right, a clearer relationship between employees and rapid growth can be observed. As it was mentioned, a scaleup by definition has more than ten employees so we need to look at the scatter plot only to the right of this threshold. In the region above the threshold (dotted line), it can be observed that most of the companies above 50 employees present a rapid growth; this strengthens the observations made for figure 4.6. Lastly, we want to explore how all four candidate dependent variables are related in one single plot. To do this, we plot the data shown in table 4.4 on top of the *total funding* - *employee* scatterplot plane.

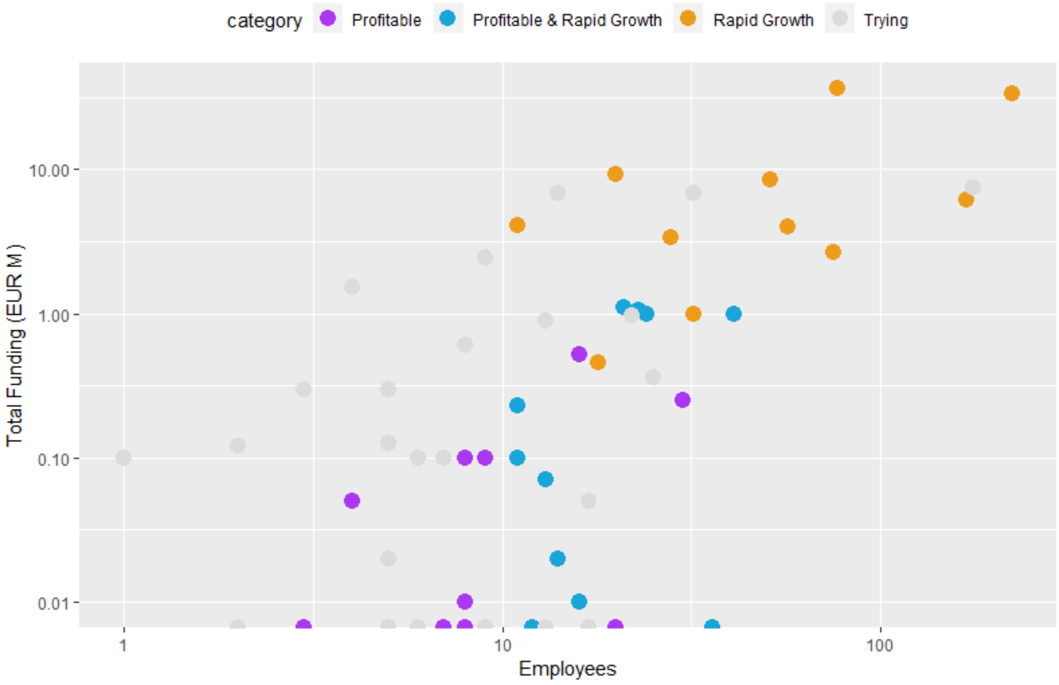

Figure 4.7: Rapid Growth and Profitability vs Total Funding and Employees.

As can be seen from this plot, all companies that have rapid growth but not profitable yet are sited on the upper right corner of the figure. This observation strongly highlights the definition by Paul Graham mentioned in section 2.2.1, "the only thing essential for startups is growth". It can also be argued that the most successful startups are those who have both rapid growth and profitability. As it can be observed in the figure, the startups that share these characteristics are located in the square delimited by the range of ten to sixty employees and one million euros funding or less. Perhaps, these companies are more conservative, and the cluster of blue points differ significantly from the cluster of orange points.

To conduct an appropriate classification analysis, which is out of the scope of this research, the data on revenue growth needs to be much more reliable. For this study, startup founders were asked to classify themselves according to the scaleup definition which is based on percentual revenue growth. Depending on the real value, we could be talking about very different companies. A percentual increase in revenue may be easier to achieve for a company having revenue in the order of hundreds of thousands of euros compared to a company having revenue above the million euros. It would be interesting for further studies to include the recurring revenue variable and analyse its relationship with the other dependent variables here presented.

4.5. Logistic Regression

In the next section, we deliver the final model of this research. As mentioned before, the goal of this thesis is to construct a predictive model of the success of startups in the Dutch startup ecosystem. As independent variables, we explore the collection of twenty-eight variables listed in table 4.3. As success predictors, we explore the variables of *total funding, number of employees* and *revenue growth*. To predict the success of startups, we carried out a *logistic regression* between the success predictors and criteria. This technique is perhaps the simplest and easiest to understand of all machine learning methods. Before proceeding to the construction of the model in section 4.6, it is necessary to have all concepts of logistic regression very clear. This section is divided into three subsections: in subsection 4.5.1 we provide the reader with important remarks to take into account when building a logistic regression, in subsection 4.5.2 we present the general equation for the logistic regression and in subsection 4.5.3 we carefully explain how the model's performance can be evaluated using different metrics.

4.5.1. General Remarks

The immediate advantage of this technique is that the model is straightforward; the relationship between the variables is modelled by an equation and can be easily interpreted. Because the model is simple and easy to understand, it is also limited in the predictions. The decision boundary must be linear, and hence the logistic regression falls short to explain complex relationships. For this research, we use this technique as an initial approach to predict startup success, for further studies other more sophisticated machine learning techniques should be evaluated. Below, we provide a list of important remarks to take into account when building a logistic regression:

- **Linear Relationship:** logistic regression does not require a linear relationship between the dependent and the independent variables. It does require a linear relationship between the independent variables and the log odds.
- **Distribution:** the variable does not need to be normally distributed.
- **Variable Types:** the dependent variable must be dichotomous. The independent variables can be numerical, ordinal, or dichotomous (categorical variables must be divided into several dichotomous variables).
- **Multicollinearity:** logistic regression requires independent variables not to be highly correlated.
- **Sample Size:** highly susceptible to over-fitting. To avoid over-fitting, large sample sizes are typically required. As a rule of thumb, for logistic regression, the number of variables in the model is defined based on the number of occurrences of the least likely outcome. Typically ten events per variable (EPV) are suggested. Depending on the number of events per dependent variable, the model may be more or less over-fitted.

4.5.2. Equation

As mentioned, the relationship between the success criteria and the predictors is modelled by an equation in logistic regression. In an ordinary linear regression, the output variable y is modeled as $y = \beta_0 + b_1 x_1 + b_k x_k$ where β_0 is the intercept and β_n are the coefficients that multiply the variables x_n. In ordinary linear regression, both dependent and independent variables need to be normally distributed. Contrary to ordinary linear regression, the general linear model is more flexible and allows the variables to not be normally distributed. In generalised linear models, the relationship between the independent variables is modelled by a specific function. For the logistic regression, this function is the *logit* function and the logit function is defined as $logit(x) = \frac{x}{1-x}$.

For logistic regression, the variables are related to the probability p of a given startup to be successful. The relationship between the variables is modelled by the logit function as follows:

$$logit(p) = log(\tfrac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

To make the relationship more clear, we must have a look at the plot of a typical logistic regression in figure 4.8. In the plot, the variable x is the independent variable, and the variable y is the dependent variable. From the plot we can infer that a subject with a low value of x is less likely to have an outcome of y equal to one. The model, represented by the fine line in blue, assigns a probability p to every possible outcome of y. As an example, in the figure, if we know that a subject has $x = 4$ we can say that it has a probability of 0.2 of y being equal to one. To the shape of the curve we refer to as a *sigmoid*, which is the inverse function of the *logit* function where $sigmoid(logit(p)) = p$.

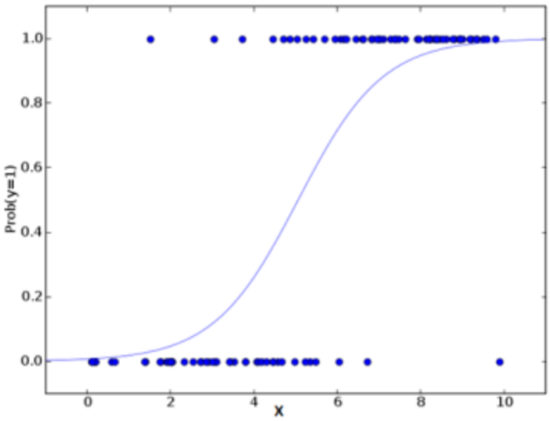

Figure 4.8: Typical plot for a logistic regression.

As explained, the previous plot and model represented by the sigmoid function gives us the probabilities of a subject to present a positive outcome. In order to make this a dichotomous variable, a threshold needs to be defined (typically 0.5). All subjects with that have a probability below the threshold are classified as negative, and all above are classified as positive. Although sometimes done that way, the selection of this threshold is not arbitrary. As an example, if we are trying to predict the outcome of a deadly disease, we may want to lower the threshold so that subjects with lower probabilities of having a positive outcome are also classified as positive. The selection of the threshold depends on the nature of the application and the ROC curve which is explained in the following section.

4.5.3. Model Evaluation

When conducting a logistic regression in R (or any other statistical tool), several values are delivered and is essential to know how these should be interpreted. In this subsection, we briefly explain relevant concepts that are necessary to understand the model proposed models.

Deviance

The deviance is a measure of the goodness of fit of the proposed model. A model is conformed by an equation in which an intercept and several coefficients are included. The *null deviance* indicates how good is the fit of the model when no predictors are included. The *residual deviance* indicates how the fit is improved when the model gets more sophisticated (when variables are added to the model). Although the absolute values are not relevant on their own, when we build logistic regression, and we include variables into the model, we should observe a reduction in the deviance.

Goodness of Fit

For assessing the quality of the model, the *Akaike Information Criterion* is typically employed. This criterion is calculated based on the deviance but it penalises the model when it gets too complicated (when many variables are included in the model). In general, when constructing a logistic regression, we want the AIC to decrease.

Although general linear models do not have an exact equivalent of the r-squared used in ordinary linear regressions, the *Mc Fadden's pseudo-R-squared* is typically used. This measure can be calculated from the deviance or from the AIC. The R-squared gives us a measurement for the overall effect size of the model.

Confusion Matrix

When a model is constructed, this can be tested, and the results can be displayed in the way of a confusion matrix. The general structure of a confusion matrix is displayed in the table below.

	PREDICTED: YES	**PREDICTED: NO**	
ACTUAL: YES	True Positive (TP)	False Negative (FN)	Sensitivity
ACTUAL: NO	False Positive (FP)	True Negative (TN)	Specificity
	Precision		**Accuracy**

Table 4.5: Structure of a confusion matrix.

This table is typically used to evaluate the results of a predictive model. In the table, some of the metrics that are used to evaluate the results, are presented. On the lower right corner, we have the *accuracy* metric, and this is calculated as the total number of correctly predicted values divided over the total number of values. In terms of the confusion matrix, the accuracy is calculated as $\frac{TP+TN}{TP+TN+FP+FN}$. For many models, accuracy is selected as the primary performance criteria. Depending on the application of the model, other metrics may be more desirable. When the population size is vastly large compared to the number of positive events (e.g. identifying a rare disease) accuracy may not be a good performance metric. In the case of identifying startup success (depending on the metric of success) the accuracy metric is a good metric of performance.

The sensitivity (also known as true positive rate or recall) tells us how good is a model at correctly identifying positives, whereas the specificity tells us how good is a model at correctly identifying negatives. The sensitivity is computed as $\frac{TP}{TP+FN}$ and the specificity as $\frac{TN}{TN+FP}$. The precision metric is used to determine which is the proportion of predicted positive outcomes that are positive. Depending on the application there is always going to exist a trade-off between the different metrics.

ROC Curve

Previously, we explained how a model could be evaluated using the confusion matrix. One issue with the confusion matrix is that the results are dependent on the threshold defined in the logistic regression. One way to evaluate the model regardless of the defined threshold is using the ROC curve. The ROC curve is obtained when the true positive rate (or sensitivity) against the false positive rate (1-specificity) are plotted for every possible threshold.

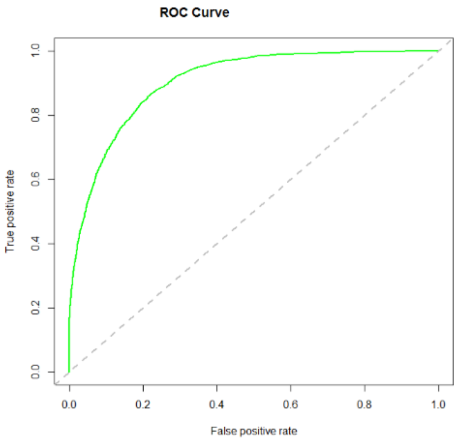

Figure 4.9: Typical ROC curve.

The ROC curve is typically used to evaluate the performance of the model. The imaginary diagonal line that goes from the bottom left corner to the upper right corner represents the points where the sensitivity equals the false positive rate. A more natural way to interpret this line is that this is the curve obtained by just doing random guessing. The ROC curve of the model is better, the further it goes from this line, towards the upper-left corner. A commonly used metric is the *area under the curve* the closest this number is to one, the higher the predictive power of the model is.

4.6. Model Construction

In subsection 4.4.2, four candidates for dependent variables (success criteria) were explored; profitability (as determined by the break-even point), was discussed as a poor metric of success. An SME can achieve profitability because the investment is not very large and hence the investment is relatively easy to recover. With this in mind, a predictive model is built to predict the outcome of three different dependent variables: *total funding, number of employees* and *rapid growth* (according to the scaleup definition). For a summarized version of the results of this section, please refer to the section 5 on the discussion chapter. As the purpose of this research is to predict a dichotomous outcome, the variables of total funding and number of employees need to be transformed into dichotomous variables. To construct the model, we carry out the following procedure:

1. A bivariate logistic regression is built for all 28 independent variables listed in 4.3 and all significant variables are selected.
2. A multivariate logistic regressions are built with the variables selected from the previous step and non-significant variables are dropped from the model.
3. The correlation between significant predictors is revised for a safety check on multicollinearity. The results of the correlation analysis are not included in this section but the detailed process can be observed in the R script in appendix B B.
4. A deviance analysis is carried out to observe the variation in the residual deviance when every variable is added one at a time. If needed, variables that do not cause big changes in the residual deviances can be dropped to avoid over-fitting.
5. The final model is built with the remaining significant variables. The McFadden R^2 is computed to assess the model fit.
6. The ROC curve is plotted and the AUC is computed to determine the predictive power of the model based on the training data set.

4.6.1. Model 1: Total Funding

To transform the *total funding* variable into a dichotomous variable, a threshold of one million euros is defined according to the average value of Series A rounds in techleap.nl's database. A startup is considered successful if it has total funding above the threshold. From the total sample of 74 data points, a total of 19 startups (26%) have total funding above this threshold. In the following table the results for the bivariate regression can be observed.

ID	Variable	β_n	$p-value$	AIC
7	University Ranking	1.21	0.052	80.02
10	Time Dedication	1.29	0.039	79.46
12	Team Size	0.77	0.022	78.66
18	Pivoting	-0.92	0.026	78.66
19	Societal Relevance	1.73	0.006	75.31
27	Employee Incentives	2.00	0.003	73.28

Table 4.6: Total funding greater than one million euros, bivariate logistic regressions results.

From the previous table, it can be observed than a total of six variables are good predictors for startup success if success is defined as having more than one million euros in total funding. According to this success definition, bootstrapped startups that prove to show good results in other metrics are excluded. The next step is to construct the multivariate regression with the six candidate variables. As the number of events is low (19), if we use six predictor variables the model is likely to be over-fitted. When building the multivariate model, one variable showed to be no longer significant. In table 4.7, the results from the multivariate regression are shown. The new model has an AIC of 59.82.

ID	Variable	β_n	$p-value$
7	University Ranking	1.39	0.1
10	Time Dedication	2.21	0.017
12	Team Size	1.09	0.027
18	Pivoting	-0.74	0.19
19	Societal Relevance	1.95	0.019
27	Employee Incentives	2.40	0.006

Table 4.7: Total funding greater than one million euros, multi-variate logistic regressions results.

Next, we can remove the pivoting variable from the model to obtain our final model with five predictors. Although the number of responses is low (19/74), we decide for the moment to leave the model over-fitted. We proceed to deliver the final set of coefficients and deviance residuals. The final model to predict a startup having more than one million euros in total funding is displayed in table 4.8.

ID	Variable	β_n	$p-value$	Deviance Residual	Δ
	Intercept	-6.36	0	84.3 (null)	
7	University Ranking	1.56	0.06	80.0	4.3
10	Time Dedication	2.27	0.01	75.1	4.9
12	Team Size	1.1	0.02	64.7	10.4
19	Societal Relevance	2.04	0.01	57.3	7.3
27	Employee Incentives	2.35	0.006	47.7	9.7

Table 4.8: Logistic Regression model to predict Total Funding above 1 Million Euros.

To evaluate the model fit and the predictive power, we compute McFadden R^2 and plot the ROC with the training data. The McFadden R^2 is 0.43 with a p-value of 0. The ROC is displayed in figure 4.10; the achieved AUC is 90.4%. This value is high, so we must suspect of over-fitting.

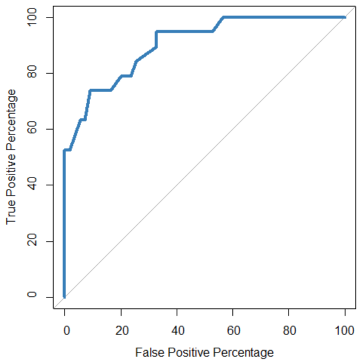

Figure 4.10: ROC for Total Funding model. AUC = 90.4%.

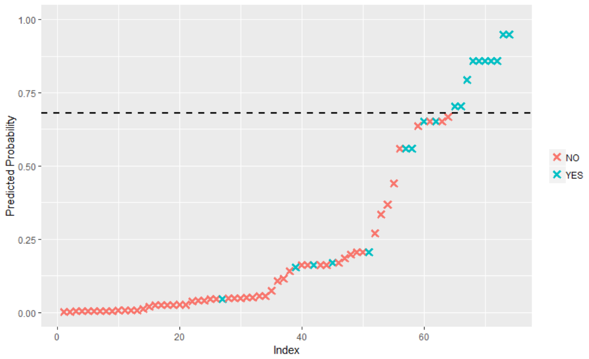

Figure 4.11: Graphical representation of model 1.

Furthermore, a graphical representation of the model is also given in figure 4.11. In this plot we can observe the predicted probabilities for the seventy-three organisations in the training data set. All data points are categorised according to their actual success outcome (green for companies with total funding above one million euros). In the plot the threshold that will be used to test the model is also represented, for an explanation on how this threshold was defined please refer to the next section on model testing.

4.6.2. Model 2: Number of Employees

In a similar manner to what was done with the *total funding* variable, a threshold of ten employees was defined to make the *employee* variable dichotomous. This threshold is defined according to the definitions discussed in section 2.2.1 and to match the *scaleup* definition. For this model, a startup is considered to be successful if this has at least ten employees after three years of its foundation. From the total sample of 74 points, a total of 37 variables (50%) have a total count of employees above the threshold. In table 4.9, the results of the bivariate regressions for each of significant predictors are displayed.

ID	Variable	β_n	$p-value$	AIC
1	Number of Founders	1.26	0.0003	90.70
10	Time Dedication	1.02	0.037	102.04
17	Data Orientation	0.90	0.063	103.03
18	Pivoting	-1.08	0.0028	96.3
21	Time to Market (T2M)	-0.63	0.035	101.91
27	Employee Incentives	1.00	0.038	102.15

Table 4.9: Total funding greater than one million euros, bi-variate logistic regressions results.

From the previous table, it can be observed that a total of six variables are good predictors of a startup having ten employees or more. It is essential to notice the effect of the different variables by looking at the AIC. Particularly the variable *number of founders* has a strong effect on the model. The interpretation of the values of the coefficients depends on the scale in which the variables are measured. As the variables we previously transformed into dichotomous or ordinal variables with a reduced number of levels the effects (as measured by the log of the odds ratio), is comparable among all variables. The next step is to build the multivariate model with the selected six variables. In table 4.10, the results of the regression are displayed. The multivariate model with six variables has an AIC of 82.277.

ID	Variable	β_n	$p-value$
1	Number of Founders	1.31	0.0015
10	Time Dedication	0.61	0.34
17	Data Orientation	1.13	0.083
18	Pivoting	-1.04	0.019
21	Time to Market (T2M)	-0.41	0.27
27	Employee Incentives	1.15	0.06

Table 4.10: Employee count above 10 employees, multivariate logistic regressions results.

Next, we proceed to remove the variables of *time dedication* and *time to market* from the model and construct a new model with four variables. As the number of positive events is relatively large (37/74), by selecting four predictor variables the model should not be over-fitted. We proceed to deliver the final set of coefficients and deviance residuals. The final model to predict a startup having more than ten employees is displayed in table 4.11.

ID	Variable	β_n	$p-value$	Deviance Residual	Δ
	Intercept	-1.55	0.16	102.6 (null)	
1	Number of Founders	1.34	0.001	86.7	15.9
17	Data Orientation	1.23	0.004	83.06	3.6
18	Pivoting	-1.14	0.007	74.5	8.6
27	Employee Incentives	1.17	0.05	70.6	3.9

Table 4.11: Logistic Regression model to predict number of employees above ten.

From the results of the model, looking at the changes in the deviance residuals, it can be noticed that the variables of *number of founders* and *pivoting* have a substantial effect on the model. To evaluate the model fit and the predictive power we compute McFadden R^2 and plot the ROC with the training data. The McFadden R^2 is 0.31 with a p-value of 0. The ROC is displayed in figure 4.12, the achieved area under the curve is 84.8%.

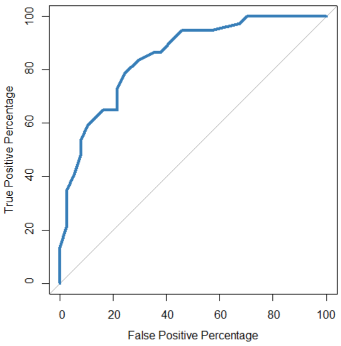

Figure 4.12: ROC for Employee Size Model. AUC = 84.8%.

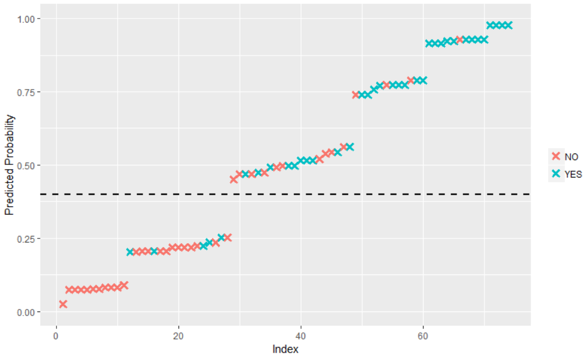

Figure 4.13: Graphical representation of model 2.

As it can be observed in figure 4.13, the model presents a less continuous distribution than the previous model depicted in figure 4.11. In the same way, the y-axis represents the calculated probabilities for each of seventy-three organisations in the training set and the colour represent the actual outcome (green for organisations with more than ten employees). The threshold used to test the model is also represented in the figure, for more details about the selection of this threshold and the results of the model testing please refer to the next section.

4.6.3. MODEL 3: RAPID GROWTH

The last model we built, predicts a startup having a rapid revenue growth as defined by the scaleup definition. We expect this model, focused on predicting rapid revenue growth, to present similarities with the previous model. From the total sample of seventy-four points, a total of twenty-two companies (30%) are classified as scaleups. In table, 4.12 the results of the bivariate regressions for each of the significant independent variables are displayed.

ID	Variable	β_n	$p-value$	AIC
10	Time Dedication	1.58	0.01	86.3
17	Data Orientation	0.98	0.08	90.7
18	Pivoting	-0.91	0.02	87.9
21	Time to Market (T2M)	-0.62	0.06	90.3
27	Employee Incentives	0.92	0.09	91.00

Table 4.12: Scaleup stage, bi-variate logistic regressions results.

The first thing that can be noticed from the previous table is that the results are similar to those of the model from total employment. As it was argued before, the set of scaleups from the total responses correspond to a subset of the startups with more than ten employees. Contrary to the total employment model, to predict startups with more than ten employees and rapid growth, the variable *number of founders* is not a good predictor. By looking at the AIC in the rightmost column, we can observe that the variables that have a larger effect in the model are *time dedication* and *pivoting*. The next step is to build the multivariate model with the selected five variables. In table 4.13, the results of the regression are displayed. The multi-variate model with five variables has an AIC of 83.749.

ID	Variable	β_n	$p-value$
10	Time Dedication	1.38	0.046
17	Data Orientation	0.77	0.22
18	Pivoting	-0.63	0.14
21	Time to Market (T2M)	-0.44	0.24
27	Employee Incentives	1.2	0.06

Table 4.13: Scaleup stage, multi-variate logistic regressions results.

The three not-significant variables from the multivariate regression are dropped. Two variables are kept as predictors of scaleup stage. The number of predictors, taking into account the number of events (22) is suitable to avoid over-fitting. We proceed to deliver the final set of coefficients and deviance residuals. The final model to predict a startup reaching the scaleup stage is displayed in table 4.14.

ID	**Variable**	β_n	$p - value$	Deviance Residual	Δ
	Intercept	-0.86	0.04	90.1 (null)	
10	Time Dedication	1.7	0.008	82.3	7.8
27	Employee Incentives	1.08	0.06	78.5	3.8

Table 4.14: Logistic Regression model to startups reaching the scaleup stage.

From the results of the model, looking at the change in the deviance residuals, we can infer that the variable of *time dedication* has the greatest influence in the model. The AIC for this model is 84.5 which is slightly larger than the one from the multivariate model with five variables. We prefer this number as we obtain greater significance and avoid over-fitting of the model. Finally, to evaluate the model fit and the predictive power we compute McFadden R^2 and plot the ROC with the training data. The McFadden R^2 is 0.13 and the p-value is 0.003. This value is considerably lower than the models to predict total funding and number of employees. The ROC is displayed in figure 4.14, the achieved area under the curve is 73.2%.

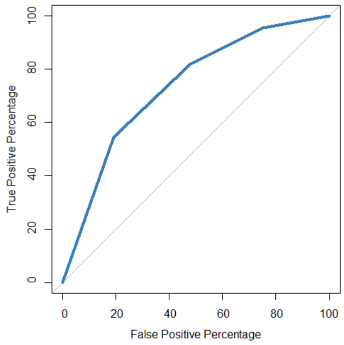

Figure 4.14: ROC for Scaleup Model. AUC = 77.9%.

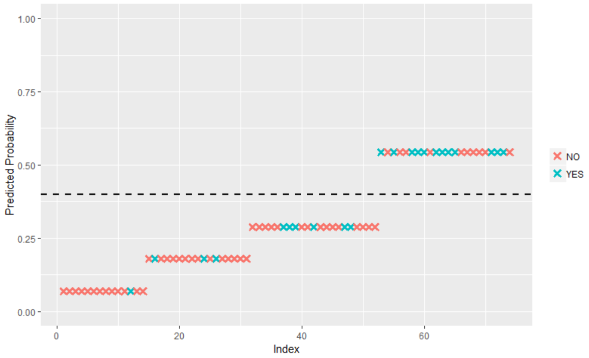

Figure 4.15: Graphical representation of model 2.

As it can be observed in figure 4.15, the model on revenue growth is the simplest of all as it is formed by two dichotomous independent variables. Because of this, only four possible outcomes are possible and the corresponding calculated probabilities can be noted in the figure. In the same way than the previous models, the y-axis represents the calculated probability and the colour represent the actual outcome (green if the company has a revenue growth as described by the scaleup definition). A threshold of 0.4 is also drawn in the figure.

4.7. MODEL TESTING

In previous section three models where constructed, one to predict total funding above one million euros, another to predict employee count above ten, and the last one to predict revenue growth according to the scaleup definition. From the three selected dependent variables, a total of eight unique variables showed to be good predictors of success. In this section, we aim to test the three proposed models on a different set of data. To collect new data, a shorter version (with only the significant variables) was delivered to a new list of startup founders. A total of seventeen responses were collected. To test the models, we make use of confusion matrices as explained in section 4.5.3. Together with the confusion matrix we deliver relevant performance metric (accuracy,sensitivity and specificity). The thresholds to test the models were selected by experimenting with the results, selecting those thresholds that maximize the accuracy without sacrificing sensitivity or specificity. For a summarized version of these results please refer to section 5.5.

4.7.1. MODEL 1:TOTAL FUNDING

As we can recall from the construction of total funding model, this model was built to predict a startup reaching a total funding above EUR one million based on the predictors of *university ranking, time dedication, team size, societal relevance* and *employee incentives.* Although this model showed the best performance in terms of McFadden R^2 (0.43) and the AUC (90.4%) we also argued that this model could be over-fitted due to the number of variables and positive events (number of successful companies in the sample). When testing the model, for a threshold of 0.7, we tested the model on the test sample and obtained the results observed in the confusion matrix of figure 4.15.

	PREDICTED: YES	**PREDICTED: NO**
ACTUAL: YES	1	5
ACTUAL: NO	0	11

Table 4.15: Confusion matrix for model 1.

For this model, an *accuracy* of 0.71, a *sensitivity* of 0.17 and a *specificity* of 1.0 were obtained. Although the model presents a fair accuracy, the sensitivity is very poor, and this means the model is not good at correctly identifying positives. This model may also have some problems due to the definition of some of the variables; these issues are further discussed in the 6.1.

4.7.2. MODEL 2:NUMBER OF EMPLOYEES

The number of employees model built to predict a startup having ten or more employees, the prediction is made based on the variables of *number of founders, data orientation, pivoting* and *employee incentives.* This model showed a slightly smaller performance compared to the total funding model; this model has a McFadden R^2 of 0.31 and an AUC of 84.8%. The advantage of this model is that the number of positive events is adequate for the number of predictors selected (avoiding over-fitting). We tested the model with a threshold of 0.4 on the test sample and obtained the results observed in figure 4.16.

	PREDICTED: YES	**PREDICTED: NO**
ACTUAL: YES	9	3
ACTUAL: NO	2	3

Table 4.16: Confusion matrix for model 2.

For this model, the calculated *accuracy* is 0.71, *sensitivity* of 0.75 and *specificity* of 0.6. This model presents good values of accuracy and sensitivity and a fair specificity.

4.7.3. MODEL 3:REVENUE GROWTH

The revenue growth model was built to predict a startup having an average annual revenue growth higher than 20% in the past three years (according to the scaleup definition). To predict a startup being a scaleup, or having a rapid growth, the predictors of *time dedication* and *employee incentives* showed to be significant. Taking into account the number of positive events and predictors, this model is most likely not over-fitted. The reduced number of predictors makes this model simple, compared to the more sophisticated models of *number of employees* (four predictors, not over-fitted) and *total funding* (five predictors, over-fitted). When building this model, the performance metrics were lower than those of the other models. The McFadden R^2 is 0.13, and the AUC for the ROC curve is 73.2%. We tested the model with a threshold of 0.4, and we obtained the results displayed in figure 4.17.

	PREDICTED: YES	**PREDICTED: NO**
ACTUAL: YES	5	3
ACTUAL: NO	1	8

Table 4.17: Confusion matrix for model 3.

For this model, the calculated accuracy is 0.76, the sensitivity is 0.63, and the specificity is 0.89. This model presents outstanding values of accuracy and specificity and a fair value of sensitivity.

5
Discussion

In the fifth chapter of this thesis, we respond to the research questions stated in subsection 1.3.1. This chapter is divided into five sections, one section for each sub research question. In the first subsection, we discuss the results from section 3.1, where an initial set of variables was delivered based on desk research. In the second subsection of this chapter, we discuss the results from sections 3.2 and 3.3, where a second selection of variables is delivered after the previous selection is discussed with knowledgeable actors in the Dutch startup ecosystem. In the third and four subsection of this chapter, we discuss the results from section 4.6. In the third subsection, we analyse the results of the bivariate regressions, which indicate which variables, from the selection of twenty-eight variables in table 4.3, are significant predictors of startup success for the three proposed models. For the fourth subsection, we discuss the results of the multivariate regressions for the three dependent variables. The multivariate logistic regressions are the models that describe the relationship between the three startup success criteria and the significant predictors. Lastly, in the fifth section, we discuss the results from section 4.7, where the three models are tested on a set of test data and evaluate using confusion matrices and associated metrics (accuracy, sensitivity and specificity).

5.1. Variable Selection 1: Desk Research

As stated in 1.3.1, the first sub-question in this thesis was: "*what is a comprehensive selection of predictors and criteria for startup success, as mentioned in the literature?*". To respond to this question, we compiled the selection of variables from four articles into one initial list of forty-three predictors and eight success criteria. The forty-three predictors were grouped into the categories of *business*, *innovation*, *actions & decisions*, *resources*, *environment* and *third-party support.* From the selection of variables, various things are worth discussing.

First, it is interesting to notice how the different variables are measured. The measurement of the variables can be critical for the construction of the predictive models. Entrepreneurial experience, for example, can be measured as a dichotomous variable ("have you had any previous ventures?"), or as a numeric variable ("how many years of entrepreneurial experience do you have?"). The previous variable is relatively tangible, easy to measure, and we would expect the variable to influence the model similarly despite the scale this is measured. Other variables are less tangible, less straightforward, and there is no consensus on how these should be measured (e.g. *innovativeness*, *scalability*). These intangible variables can be highly problematic, and an adequate evaluation of these is mandatory if these are to be included in the model. Abstract variables are often measured through Likert-scales in questionnaires, but this may not always be adequate. *Innovativeness* for example, if measured by means of a Likert-scale, can delivery very biased results. Intangible variables can also be measured with proxies. The *social capital* of a startup founder, can be measured with a Likert-scale in the questionnaire (as we did) but it can also be measured with proxies such as the number of connections on LinkedIn. Proxies can be very valuable, but they must be carefully implemented. The number of connections in LinkedIn for example, do not take into account the quality of these connections. To conclude this observation, we observed that the way variables are measured, greatly varies, and this can be very influential in the construction of any predictive model.

Second, it was worth observing which variables are recurrent in the literature. From the results of section 3.1.2, we can observe that some variables, or related variables, are included across the different sources. Entrepreneurial knowledge and experience, for example, is included in all four sources consulted. Excluding the first source, all other sources take into account several metrics of general knowledge & experience, such as, *entrepreneur age*, *work experience*, *industry experience*, *managerial skills* and *education level*. Other variables that are recurrent in the literature are *money to market (M2M)* and *funding source*. Interestingly, none of these recurrent variables (except perhaps education level) resulted in being significant predictors of startup success in our research. On the contrary, *personal dedication* (or time dedication as defined in our model), which was one of the most significant predictors in our model, was only included in fourth source. From this, we can conclude, that either we should have included more references in the review, or that current research is leaving aside variables that could be good predictors of success.

Third, it is also worth mentioning how the literature focuses on the different categories of factors. Except for source number one, all other sources strongly focus on the characteristics of the entrepreneur (particularly source two). In this research, we strongly agree with this position, and we believe that predictive models (including this one), can be further improved by introducing distinctive variables related to the characteristic of the individual. For this research, we did not focus on attacking this objective since evaluating certain behavioural characteristics can be extremely challenging (e.g. self-confidence). Besides, it is not worth making an excessive effort on variables that are not confirmed to be significant predictors; thus we decide to explore some of these variables by using a simple Likert-scale in the questionnaire.

The desk research conducted on the four predictive models on organizational success served us to answer the first sub-question of this research. The goal of the first sub-question is to give us an initial set of variables to be explored before collecting the data and building the predictive model.

5.2. Variable Selection 2: Interviews

The second subquestion of this research is "*how the selection of predictors and criteria can be improved according to knowledgeable actors of the Dutch startup ecosystem?*". To achieve this goal, we carried out a total of seven interviews, with individuals from different backgrounds and levels of experience. The purpose of the interviews was twofold: (1) to identify key success factors and criteria as perceived by the interviewees, (2) to discuss, refine and enrich the selection of variables delivered from the desk research.

First, during the interviews, interviewees were explicitly asked to give their critical success factors and criteria. In the form of "which factors you find the most critical when determining startup success?" for the predictors, and "how to measure the success of startups?" for the criteria. Overall, responses for the predictors were mostly intangible (e.g. team quality, culture, scalability) whereas the criteria responses were more measurable (e.g. revenue, profitability, growth). The consensus, amongst all interviewees, is that team's characteristics are what matters the most when discussing about success factors. Interestingly, it is often mentioned in the startup world, that a good idea is nothing without a proper team. Reciprocally, investors would even invest in a faulty-idea if this is led by a great team. Team characteristics that were mentioned by the interviewees included *experience*, *passion*, *ability to adopt new insights*, *team diversity*, *number of founders*, *motivation*, *trust between founders* amongst others.

Besides the characteristics of the people, characteristics of the business (as an entity), actions and decisions, and the innovation (the idea) were also mentioned. Regarding the business, *culture* was mentioned by one interviewee to be a critical success factor. Although culture is a very interesting variable, this is very intangible and difficult to evaluate. In this research, as attempt to capture this factor, we included the organicity variable (ID 16 in table 3.3.1). Perhaps in further research a proper methodology to evaluate corporate culture can be implemented. Regarding the actions & decision of startups during their early stages, *timing*, *time to market* and *number of pivots* were mentioned by the interviewees. Timing is not included in this research as it is difficult (if not impossible) to evaluate quantitatively. Timing refers to the adequate timing of market introduction, sometimes it is too early (e.g. consumers are not ready for the innovation), or it is too late and incumbents already dominate the market. Time to market is included in our research, this variable is defined as the time elapsed between startup launch and market introduction. During the interviews it was discussed that startups are encouraged to launch as soon as possible, this idea is confirmed as the variable showed to be a significant predictor in the models of *number of employees* and *revenue growth*. To explain why an early market launch shows to be a significant predictor of success, multiple arguments can be made. For example, we can say that a startup that launch their product faster could be decisive, efficient and passionate. Evidently, arguments can also be made against the same idea. Lastly, the variable *pivoting* was also mentioned during one of the interviews. It was discussed that startups that pivot, have greater capacities at receiving and incorporating feedback and can deliver a more developed solid ideas. Contrary to what it was expected, the pivoting variable was negatively related to success in all the three models constructed. These results can be explained with the same argument made for the *T2M* variable. Startups that pivot in excess, are not decisive and take too long to launch their innovation to the market. Because of the contradiction in this variable, this should be implemented with skepticism.

Second, the interviews were conducted to discuss the variables in the list delivered by the desk research section. The initial list consisted of forty-three independent variables and eight dependent variables. After the discussion with the key actors, a second set was delivered with thirty-seven independent variables and four dependent variables. Although the set of independent variables seems to only differ by six variables, the changes made were much larger than that. Out of the forty-three factors from the first selection, twenty factors were removed, and fourteen were added, which gives the net difference of five variables.

Although we do not intend to repeat why variables were excluded and added, it is important to observe the general reasons why these actions were taken. In the case of the excluded variables, these were removed because of five main reasons. First of all, variables that were consider too complex to evaluate were excluded, as an example of this, we have variables such as *self-confidence*, *scalability*, *unique advantage*, *political stability*, *market environment*, among others. Second, variables that were thought to be plain demographics were also excluded, this includes variables such as *founders' gender*, *HQ location* and *industry*. Third, although some variables are tangible, data can be difficult to determine because we are discussing about early-stage characteristics, one example for this case are the web-metrics, it is not practical to ask founders for information from several years ago, these variables could be included in longitudinal studies. Fourth, some variables were excluded to avoid multicollinearity, an example of this are all the variables related to experience which can be reduced into fewer variables (e.g. founders' age, industry experience, managerial experience). Fifth, variables were excluded because they didn't seem to be adequate casual factors, for example trying to predict startup success based on firm age. The relevance of discussing all the reasons why the variables were excluded is to highlight the fact that many studies do not stop to reflect on the selection of variables. For a predictive model to not only be accurate but useful, variables need to be chosen adequately. Besides excluding variables, discussion with the interviewees also served to propose new variables for the predictive models, out of the total fourteen added variables represented in table 3.7, the variables of *university ranking*, *data orientation*, *societal relevance*, *time to market (T2M)*, *pivoting*, and *employee incentives* showed to be significant predictors of startup success in section 4.6. This is very important to recognize because they are novel variables added that were not found in previous predictive models. Highlighting the fact that the core of this research is to construct a predict mode of startup success based on a comprehensive selection of variables.

5.3. Startup Success: Significant Predictors

In this research, we built three predictive models for three different definitions of startup success. The first model is built to predict a startup having total funding above one million euros, the second model is built to predict a startup having ten employees or more, and the third model is built to predict a startup having a revenue growth according to the scaleup definition. For the third sub-question, "*which of the selected predictors showed to be significant at predicting startup success?*", we can refer to tables 4.6, 4.9 and 4.12 of section 4.6. To obtain these tables, we conducted a total of eighty-four (twenty-eight variables and three models) bivariate logistic regressions and listed the factors that proved to be significant predictors. It is important to recall that the second selection of variables, resulting from the discussions with knowledgeable actors in the ecosystem, which contained a selection of thirty-seven predictors and four criteria, was further reduced to twenty-eight predictors and three criteria after data preparation in section 4.3.

From the results of the bivariate logistic regressions, we obtained that a total of nine significant predictors for startup success depending on the success criteria. In table 5.1, the list of significant predictors for each of the three models are shown.

ID	Variable	Total Funding	Number of Employees	Revenue Growth
1	*Number of Founders*		X	
7	*University Ranking*	X		
10	*Time Dedication*	X	X	X
12	*Team Size*	X		
17	*Data Orientation*		X	
18	*Pivoting*	X	X	X
19	*Societal Relevance*	X		
21	*Time to Market (T2M)*		X	X
27	*Employee Incentives*	X	X	X

Table 5.1: List of significant predictors for three success criteria.

From the list of significant predictors, a few observations can be made. Three factors are significant predictors for all three models, these are the variables of time *dedication*, *pivoting* and *employee incentives*. For the first and third variables the measurement is quite straightforward and is less susceptible to errors. The second variable, on the contrary, as it was discussed in the previous section, may be conflicting and subject to interpretation.

In general, we observe that there are predictors that are very straightforward and others that should be implemented doubtfully. The variables of *number of founders*, *time dedication* and *employee incentives* are the most straightforward and hence preferred for the predictive models. The variables of *university ranking*, *team size*, and *time to market* are relatively quantifiable but these are slightly more prone to errors. The variable of *university ranking*, as discussed before, may exclude top-tier educational institutions that are not included in the general rankings (e.g. MBA Schools). The factor *time to market* can significantly vary when not defined properly. Is T2M measured from the generation of the idea? From the launch of the startup? Or perhaps from the day startups started working on the MVP? Lines are blurry, and this makes these concepts conflicting. The variable *team size* has a similar problem; for this research, founders were requested to indicate the number of employees in their early stages. This variable can have similar problems to employee count (e.g. ten persons working a few hours a week). Lastly, we find the variables of *data orientation*, *pivoting* and *societal relevance* the most conflicting and the ones that should be implemented with the most scepticism in our models. The variables of data orientation and societal relevance can be highly biased when evaluated from the perspective of the founders themselves. Data and societal impact are hyped concepts that often are overused, and not many founders are self-reflective when responding to the questions in the questionnaire. For adequate use of these variables, proper methodologies should be designed and implemented. Lastly, the pivoting variable, besides also be prone to errors due to interpretation (what exactly can be considered a pivot?) the results contradict the hypotheses mentioned by one of the interviewees.

Besides the significant predictors, it is also important to discuss some of the variables that did not show to be good predictors of success. Surprisingly, the variables of *mutual trust* and *work experience*, did not show to be good predictors of startup success in this research. The variable of *mutual trust* was mentioned by one of the interviewees to be a critical factor of success. The interviewee carries out similar research on startup success and he claimed that this variable (measured in the same way), is one of the most critical factors in his model. The variable of *work experience*, which is often discussed to be startup success factor, both during the interviews and in the literature [Azoulay et al., 2018], did not result to be a good predictor in the models of this research. This two variables, perhaps, were the ones that the researcher found the most surprising to not show as predictors. Other variables that were also expected in a less degree include the variables of *previous salary*, *entrepreneurial knowledge & experience B2B vs B2C*, *education level*, and *MVP*.

5.4. Predictive Models

Following next, we proceed to discuss the fourth subquestion of this research. "What is the relationship between the significant predictors and the criteria for startup success?. To address this question, we have to look at the results delivered from the multivariate regressions for the three constructed models in section 4.6. The three multivariate regressions are represented in tables 4.8, 4.11 and 4.14. It is important to remember, that some variables that resulted significantly in the bivariate regressions may no longer be in the final models based on the multivariate logistic regressions. For the creation of the three predictive models, a total of eight variables were used. In table 5.2, the three models are presented. Inside the cells the values of the β coefficients are given.

ID	Variable	Total Funding	Number of Employees	Revenue Growth
	Intercept	-8.64	-1.55	-0.86
1	*Number of Founders*		1.34	
7	*University Ranking*	1.56		
10	*Time Dedication*	2.27		1.7
12	*Team Size*	1.10		
17	*Data Orientation*		1.23	
18	*Pivoting*		-1.14	
19	*Societal Relevance*	2.04		
27	*Employee Incentives*	2.35	1.17	1.08

Table 5.2: Three predictive models on startup success.

Perhaps the most interesting thing that we can highlight from the results is that the variable of *employee incentives* is a significant predictor in all three models of startup success. For the interpretation of the results, it is essential to recall the logit function as explained in section 4.5. The logit function, delivered by the multivariate logistic regression relates the predictors to the probability of an event of being classified as positive. The general form of the equation is displayed below.

$$logit(p) = log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

When discussing the advantages of the logistic regression, as compared to other machine learning techniques, we mentioned that the results of a logistic regression could be easily interpreted. The fact that the relationship between the variables is modelled by an equation, is strong evidence of this advantage. To understand how the models should be interpreted, we provide an example using the simplest model of revenue growth. We replace the values of the coefficients of the equation and the variables are shown as x_k, where k corresponds to the ID of the variable as listed in table 4.4.

$$logit(p) = -2.6 + 1.7x_{10} + 1.08x_{27}$$

Interpreting this equation is very easy, especially because the two variables that form the model are dichotomous, and there are only four possible outcomes. If a startup founder is dedicated full-time, the variable x_{10} takes a value of one, zero if otherwise. If the startup gives equity to its employees, the variable x_{27} takes a value of one, zero if otherwise. For the first scenario, where both predictors are zero (presumably wrong decisions), the probability p of a startup of being successful is given only by the -2.6 intercept. For this scenario the computed probability is 6.8%. On the other hand, when both variables are one (presumably the best scenario), the computed probability is 54.5%. Other information that can be obtained from the interpretation of the equation comes from the β coefficients which represent the *log of the odds ratio*. As an example, the 1.7 β coefficient for the *time dedication* variable in the *revenue growth* model, indicates that the odds of being successful, for a full-time dedicated founder, are 4.5 times ($e^{1.7} - 1$) higher than the odds of a part-time dedicated founder.

5.5. Models' Performance

In the last section of this chapter, we respond to subquestion five: "how does the relationship between predictors and criteria perform at predicting startup success?". To respond to this question, we discuss several metrics or indicators that resulted from the construction of the models in section 4.6 and the model testing in section 4.7. For a better understanding of the meaning of the different metrics, please refer to section 4.5.

To construct the model, we used a total of seventy-four responses. From the design of the models, the most important indicator of performance is the perhaps the AUC (area under the curve) of the ROC curve. Explained simply, an area under the curve of 50%, or a straight diagonal in the ROC curve, means that the model is as good as random guessing. The closer the AUC is to 100%, the higher the predictive power of the model. The ROC is a plot that relates the significancy and the specificity of a model for all possible thresholds of decision. A given subject is classified as positive if its computed probability (calculated from the model) is above the defined threshold. The McFadden R^2, like in the normal R^2 for an ordinary linear regression, is an indicator of the goodness of the fit of the model. Although low values are expected for the R^2, this indicator is useful to compare the different models. If the model is over-fitted, perhaps because of the number of variables is too large, considerable values of AUC and R^2 can be deceiving. To assess if the model is over-fitted, we can observe at the ratio between the number of positive occurrences and the number of variables in the model. For a model to not be over-fitted, a suggested ratio of at least 10:1 is recommended. Lastly, it is essential to highlight that the mentioned metrics are calculated from the training data, to evaluate the actual performance of the models, these must be tested in a separate set of data.

To test the models, we collected new data and seventeen new responses were obtained. In section 4.7, we tested the three predictive models on startup success and delivered the achieved confusion matrices. Confusion matrices are typically used in machine learning to evaluate predictive models and various metrics can be computed from these. For this research we focus on the indicators of *accuracy*, *sensitivity* and *specificity*. Accuracy is a recurrent term when discussing a predictive model, and this is an overall indicator of how good is a model at making predictions. Nevertheless, accuracy alone is not sufficient to evaluate a predictive model. As can be seen from one of our models, a model can have a fair accuracy but a poor sensitivity or specificity. Sensitivity is an indicator of how good the model is at predicting positive outcomes whereas specificity is an indicator of how good is the model at predicting negative outcomes. Now that all metrics have been briefly discussed, we proceed to discuss the performance achieved for the proposed models of this research. In table 5.3, the various indicators of performance are displayed for all three models.

	Characteristic	**Total Funding**	**Number of Employees**	**Revenue Growth**
Train (n=74)	*Number of Predictors*	5	4	2
	Number of Occurences	19	37	22
	McFadden R2	0.43	0.31	0.13
	AUC	90.4%	84.8%	73.2%
Test (n=17)	*Accuracy*	0.71	0.71	0.76
	Sensitivity	0.17	0.75	0.63
	Specificity	1.00	0.6	0.89

Table 5.3: Performance for the three predictive models on startup success.

All three proposed models have advantages and disadvantages. It can be said that the total funding model is the most sophisticated one as this includes five factors to predict startup success. Nevertheless, a ratio of approximately 4:1 positive events to number of predictors tells us that this model is most likely over-fitted. The high value of R^2 and AUC in the total funding model, compared to the other models, can be attributed to over-fitting. The model on number of employees also includes a fair amount of four success predictors. Contrary to the total funding model, the number of employees model has a much better ratio of positive events to number of predictors (9:1), which makes this model less prone to over-fitting. Lastly, the revenue growth model, the simplest of the three models, presents low values of R^2 but fair value for the AUC indicator.

When testing the models, we could verify that the total funding model did not perform as well as expected (presumably because of over-fitting). For a threshold of 0.71, a maximum accuracy of 0.71 was obtained for this model. Although the model is overall good at making predictions, with a sensitivity of 0.17 the model performs poorly at predicting positive outcomes. In the context of predicting startup success, it is perhaps more important to be better at predicting positive outcomes (sensitivity) than at predicting negative outcomes (specificity). For the model on number of employees, better results were achieved. With an accuracy of 0.71 and a sensitivity of 0.75, the second model is much better at predicting positive outcomes than the first model. Although the sensitivity has a lower priority, it is good to observe that this has a value of 0.6 (better than random guessing). Lastly, the less sophisticated revenue growth model shows outstanding results on the test data. With a sensitivity of 0.63 and a specificity of 0.89, this model is reasonably good at predicting positive outcomes and very good at predicting negative outcomes. The accuracy of this last model is also superior to the ones of the other two models despite only including two variables.

Which is the best model? According to the results obtained from section 4.7, both the number of employees model, and the revenue growth model, are good predictors of startup success. The advantage of the former is that this is more sophisticated (more factors included) and hence more interesting. Nevertheless, as it was discussed before, some of the variables included for the construction of this model should be implemented with skepticism (particularly the pivoting variable). Furthermore, despite being the simplest model (with only two factors), the *revenue growth model* is perhaps the best of the three proposed models. Not only because the test results are quite good but also because the variables used to construct the model are very straightforward, easy to measure and easy to implement.

"I'm convinced that about half of what separates
the successfull entrepreneurs from the non-successful ones
is pure perseverance."
- Steve Jobs, Co-Founder and CEO, Apple

6
Final Remarks

In the conclusions section, we respond to the main research question. To achieve this, we summarise the insights obtained from the discussion chapter 5 and deliver a conclusion for this thesis. Lastly, in the last section of this chapter and this thesis, we deliver a list of recommendations on how this research can be further improved.

6.1. Conclusion

The purpose of this thesis, titled "Startup Success Prediction in the Dutch Startup Ecosystem", is to answer to the main research question of "how startups' success can be predicted from a comprehensive selection of predictors and criteria in the Dutch startup ecosystem?. To achieve this goal, two general actions needed to be carried out: to come up with the comprehensive selection of variables, and to predict startup success from this selection. The final results of these two task are presented in table 4.3 and table 5.2 respectively.

The core of this research comes from the comprehensive selection of variables, not from the predictive model, which is rather simple. As it was mentioned in the introduction, on section 1.4, previous research focuses too much on the construction of the model but does not reflect on the selection of the variables. For a model to not only be accurate but also to be useful, variables should be carefully selected and measured.

To deliver a comprehensive selection of variables, we carried out the following actions. First, compiled from previous research, an initial list of forty-three independent variables and eight-dependent variables. Second, we discussed the previous selection with knowledgeable actors from the Dutch startup ecosystem. The list was refined into a new selection of thirty-seven independent variables and four dependent variables. Lastly, once the data was prepared and explored, the final list of variables presented in table 4.3 consisted of a total of twenty-eight independent variables and three dependent variables.

From the process of selection of predictors and criteria for startup success, we learned numerous lessons. We got acquainted with the meaning and interpretation of all variables in the selection. We learned that most success factors, as perceived by the knowledgeable actors, are hardly tangible and difficult to evaluate (e.g. team quality). In order to include these less tangible variables in a model, proper evaluation methodologies or proxy metrics should be carefully designed and implemented. Perhaps, it is difficult for very abstract concepts like *team quality*, but entirely possible to achieve for slightly more tangible variables such as *data orientation*, *social impact*, or even *psychological traits*. Lastly, we also observed that other more tangible variables, which are rarely included in previous research, could be significant predictors of startup success (e.g. time dedication, equity to employees, time to market). The whole process of the selection of the variables was very insightful and made this research distinctive from previous studies.

How to define success? Besides the exploration and selection of the success factors, success criteria were also reviewed from the literature and discussed with the actors from the Dutch startup ecosystem. In general, we observed that different measurements and levels of success could be used, depending on aspects such as who is interested in the research. As an example, the government might be more interested in factors such as the total employment created by startups, whereas investors might be interested in factors such as financial KPIs or company valuation. Furthermore, success is not a one-point event, it can occur at many levels, and success is experienced at many stages during a startup's life-cycle. A startup can experience success when launched, when finalising a MVP, when obtaining the first returns from sales, when achieving the first round of investment, and so forth. From the idea to unicornization, the startup journey consists of many stages where success (or failure) is experienced multiple times.

For this research, we evaluated three metrics of success as dichotomous variables: total funding, number of employees and revenue growth. The advantage of defining success as dichotomous depends on the criteria. For the total funding variable, and particularly for the number of employees variable, business databases contain meaningful errors which make the use of continuous data not suitable. Furthermore, revenue is rarely public. To determine revenue growth, we asked the startup founders to classify themselves according to the scaleup definition provided in the glossary. With the three success criteria, we constructed three predictive models on startup success: one model to predict a startup reaching total funding above one million euros, a second model to predict a startup having ten or more employees and a third model to predict a startup having an average annualized revenue growth of at least 20% in the past three years. The first two models were selected because the data was available in techleap.nl's database. The third model was selected, inspired on the definition of startup by Paul Graham "the only essential thing to define startups is growth, everything else we associate with startups follows from growth".

For the three predictive models, a total of eight significant success factors were used: *number of founders*, *university ranking*, *time dedication*, *team size*, *data orientation*, *pivoting*, *societal relevance* and *employee incentives*. For a better understanding on how these variables were evaluated please refer to table 4.3. From the three predictive models, the second model (employees) and the third model (revenue) showed the best results when testing the models in section 4.7. The number of employees model is more sophisticated (four predictors) but includes variables that should be implemented sceptically. The model on revenue growth, on the contrary, is more simple (two predictors), but the measurement of its predictors is straightforward and less prone to criticism. The variable *employee incentives*, a dichotomous variable that states if a startup gave equity to their employees, showed to be a significant predictor for all the three models. It is particularly interesting to notice how the *revenue growth* model presents such a good predictive performance despite using only two variables, *employee incentives* and *time dedication*.

Evidently, this research was not carried out without limitations. In the next section, we discuss how this research can be improved for future research.

6.2. RECOMMENDATIONS

As it was mentioned at the beginning of this thesis, this study is intended to contribute to techleap's organisational goal of strengthening the startup ecosystem of the Netherlands. This study should be used as a departing point. It must be significantly improved to create much more sophisticated prediction on startup success in the ecosystem. The way techleap.nl can improve this research is by attacking the limitations of this.

This study was primarily limited due to a lack of resources which is typical for a master thesis. First of all, this research was conducted by a single individual as a cross-sectional study. Second, as the researcher is a university student, collecting the data from startup founders through questionnaires was challenging due to the lack of credibility or lack of benefits for the respondent. Third, this research is based exclusively on techleap.nl database. Fourth, the conducted qualitative research, which includes both the literature review and the discussions with knowledgeable actors, can be done more thoroughly. Lastly, the implemented machine learning techniques are rather simple and more sophisticated models can be implemented. To improve this research, we suggest that the following actions are taken:

- **Longitudinal study:** Select a largely-enough sample of early-stage startups to monitor their performance through time. One of the greater limitations of this study was that we evaluated startups' success based on their early-stage characteristics. Some variables might be wrongly estimated by the founders (e.g. money to market), and some variables were excluded as it was cumbersome to obtain the data from past years (e.g. web metrics). To improve this research, we suggest to conduct longitudinal research by collecting information on new start-ups now, collect their data periodically and conduct a time-series analysis to predict their performance in the future.

- **Enrich the data:** to improve the quality of the data three main actions can be taken. First, explore new variables to be included in the models. Do this by carrying out brain storming sessions with stakeholders and partners, and conduct a more thorough literature review. Second, improve the number and quality of the responses. To achieve this use techleap.nl's credibility to reach a larger number of respondents and provide them with incentives so are willing to participate in the study. Lastly, the quality of the data can be improved by implementing adequate methodologies or frameworks to evaluate soft variables (e.g. societal impact).

- **Success prediction:** as mentioned, the techniques herewith implemented are rather simple. Logistic regression, the machine learning technique here employed, is simple, easy to understand, but limited in their predictive power. A considerable sample size is required, and this makes a model difficult to implement. For further research, other machine learning predicting techniques such as decision trees and neural networks should be explored. Furthermore, in longitudinal studies, it would be interesting to construct the predictions based on a periodical review of the independent and dependent variables (a time-series prediction).

A

Appendix A: Questionnaire

In this appendix the list of questions included in the questionnaire are provided for a rapid consultation. For a detailed view of the questionnaire content please consult the following link: https://forms.gle/tFGsh3atRm9UCdqPA.

A.1. Predictors

1. When you launched your startup, for how long (in months) did you know your co-founders? Leave blank if no co-founders and compute the average if various.

2. How much working experience did you have (in years) when launching your startup? Feel free to use decimals.

3. Agree or Disagree: At the very early stages of my startup, I had strong entrepreneurial knowledge and experience.

4. What is your highest level of education achieved?

5. Type in the name of the educational institution where you achieved your highest digree.

6. Select the subject that better matches your field of study.

7. Agree or disagree. At the very early stages, I was very well connected (quantity and quality) to individuals or organizations that could facilitate the development of my startup.

8. At the early stages, I was dedicated (full-time/part time) to my startup.

9. Which was the average salary (in euros) per month you earned before launching your startup?

10. Please select your risk profile (from risk averse to risk seeking).

11. Personal Motivation: What drove you (primarily) to start your own business?

12. Core team. How many employees (not counting founders) did your startup have before market introduction?

13. Agree or disagree. My core team consisted of a varied mix of people from different genders, educational backgrounds and cultures.

14. Please locate your startup according to its organicity (from highly mechanic to highly organic):

15. Please locate your startup according to its technological orientation (from tech-enabled to tech-driven).

16. Please locate your startup according to its data orientation (from data-resistant to data-driven).

17. To what degree was your product, process or business model new for the market? From incremental innovation to radical innovation.

18. Please locate your startup according to its market orientation (from technology push to market pull).

19. Agree or disagree: my product significantly outperforms existing solutions from other competitors and substitutors.

20. Societal relevance. Agree or disagree: my startup help important stakeholders and addresses major societal and economic issues.

21. Which was your main source of funding?

22. Mentorship: which of the following entities provided you with the most valuable feedback/advice?

23. Ambition to grow. Agree or disagree: at the very early stages, I envisioned my startup rapidly growing, crossing borders and reaching global markets.

24. How many direct competitors did you recognize at the very early stages of your startup?

25. How many alliances (horizontal, upstream or downstream) did you form at the very early stages of your startup?

26. How many times did you pivot before market introduction?

27. Time to market: how long (months) did it take your startup to reach market introduction?

28. Money to market: how much money (in euros) did you spend before market introduction?

29. Did you create a POC and a Prototype before your MVP? (4 categories)

30. Which of the following alternatives did you primarily use to communicate to the outside world? (4 categories)

31. Did you explicitly forecast your demand? (y/n)

32. Did you calculate your total addressable market? (y/n)

33. Did you give equity to your employees? (y/n)

34. To what degree was the customer involved in the development of your product/business? From not involved to highly involved.

A.2. CRITERIA

1. Has your startup reached the break-even point?

2. According to the provided definition, can your company (in the present time) be considered a scaleup? (y/n)

B

Appendix B: R Studio Code

```
# ———————————————————————————————————————————————————————————————————————
#—————————————————————————————— IMPORT—————————————————————————————————————
#————————————————————————————————————————————————————————————————————————

rm(list=ls()) #Clear Variables
dev.off() #Clear Plots

#Set Working Directory
setwd("C:/Users/Diego Camelo/Documents/Professional/TU Delft/MOT/Master Thesis/Data")

#Import data_train from Excel
library(XLConnect)
book <– loadWorkbook("MasterSheet3.xlsx")
data <– readWorksheet(book,sheet=1,startRow=1,endRow=92)

# #Sample Characteristics
#
# library(ggplot2) #Library to plot data_train
# data_train$Launch.Year <– as.character(data_train$Launch.Year)
# ggplot(data_train,aes(x=Launch.Year)) + geom_bar(start="count")
# + xlab("Launch Year") + ylab("")
# + theme(axis.text.x = element_text(size=13),axis.title.x=element_text(size=14),axis.text.y = element_
    text(size=13))
#
# sort(table(data_train$Industry)) #To tabulate industries
# sort(table(data_train$HQ.Location)) #To tabulate locations


#Exclude Variables: Sample Characteristics
data <– subset(data,select = –c(ID,HQ.Location,Industry,Launch.Year))

#Exclude variables: no longer interesting variables
data <– subset(data,select=–c(Field.of.Study,Innovativeness,Driving.Force,Outperformance,Ambition.to.Grow,
    Competition))

#str(data_train) #Check the data_train Structure

data_train<– data[1:74,]#Training data_train
data_test <– data[75:91,] #Test data_train (to be used at the end)

# ———————————————————————————————————————————————————————————————————————
#———————————————————————————————TYPE CONVERTION————————————————————————————————
#————————————————————————————————————————————————————————————————————————

#Categorical variables as Factors (Independent 11)
data_train$B2B.vs.B2C <– as.factor(data_train$B2B.vs.B2C) #2 Levels: B2C/B2B
data_train$Time.Dedication<– as.factor(data_train$Time.Dedication) #2 Levels: Full time vs part time
data_train$Motivation <– as.factor (data_train$Motivation)
data_train$Funding.Source<– as.factor(data_train$Funding.Source)
data_train$Mentorship <– as.factor (data_train$Mentorship)
```

```
data_train$MVP <- as.factor (data_train$MVP)
data_train$Communication.Tool <- as.factor(data_train$Communication.Tool)
data_train$Forecasted.Demand <- as.factor(data_train$Forecasted.Demand)#2 Levels: yes/no
data_train$TAM.Calculation <- as.factor(data_train$TAM.Calculation) #2 Levels: yes/no
data_train$Employee.Incentives <- as.factor(data_train$Employee.Incentives) #2 Levels: yes/no
data_train$Persona <- as.factor(data_train$Persona) #2 Levels: yes/no

#Categorical variables as Factors (Independent 2)
data_train$Revenue.Growth <- as.factor(data_train$Revenue.Growth) #DV related to rapid growth
data_train$Profitability <- as.factor (data_train$Profitability) #DV related to profitability

# str(data_train) #Check the structure of the transformed data_train


#----------------------------------------------------------------------------------------------------
#----------------------------------------Cleaning ----------------------------------------------------
#----------------------------------------------------------------------------------------------------

library(purrr) #Library to manipulate data_train
library(tidyr) #Library to manipulate data_train


#----------------------------------------Missing Values ----------------------------------------------

# #Visual Check of Missing data_train (Sparsity)
# library(Amelia)
# missmap(data_train,main = "data_train Sparsity")
#
# #Check how many rows have missing (NA) data_train
# sort(sapply(data_train,function(x) sum(is.na(x))),decreasing = TRUE)

#Replace NA Values from QS Rating to 0 score
data_train$University.Ranking[is.na(data_train$University.Ranking)] <- 0


#Replace all the rest of NA values with median data_train (to avoid outliers)
data_train$Mutual.Trust [is.na(data_train$Mutual.Trust)] <- median(data_train$Mutual.Trust,na.rm=T)
data_train$M2M[is.na(data_train$M2M)] <- median(data_train$M2M,na.rm=T)
data_train$Previous.Salary [is.na(data_train$Previous.Salary)] <- median(data_train$Previous.Salary,na.rm=T)
data_train$T2M [is.na(data_train$T2M)] <- median(data_train$T2M,na.rm=T)
data_train$Pivoting [is.na(data_train$Pivoting)] <- median(data_train$Pivoting,na.rm=T)
data_train$Alliances [is.na(data_train$Alliances)] <- median(data_train$Alliances,na.rm=T)
data_train$Customer.Proactiveness [is.na(data_train$Customer.Proactiveness)] <- median(data_train$Customer.Proactiveness,na.rm=T)
data_train$Competition [is.na(data_train$Competition)] <- median(data_train$Competition,na.rm=T)
data_train$Education.Level [is.na(data_train$Education.Level)] <- median(data_train$Education.Level,na.rm=T)


#Check new sparsity
# missmap(data_train,main = "Data Sparsity")

#---------------------------------------------------Outliers ------------------------------------------------

#Numeric variables vector (10)
num_var<- c("Number.of.Founders","Mutual.Trust","Work.Experience","Previous.Salary",
"Team.Size","Alliances", "Pivoting","T2M","M2M")

# #Histogram Plots to Spot Outliers
# data_train[,num_var] %>% gather() %>%
#   ggplot(aes(value)) +
#   facet_wrap(~ key, scales = "free") +
#   geom_histogram()
#
# #Note: Alliances,M2M & Pivoting Clearly have outliers.
#
# #Box plots for a closer look to outliers. ------------------------------
# windows(10,5)
# par(mfcol = c(1,5)) #Creates a grid of plots
```

```
#
# boxplot(data_train[,"Team.Size"], main = "Team Size") #Outliers detected
# boxplot(data_train[,"Previous.Salary"],main="Salary") #Outliers detected
# boxplot(data_train[,"T2M"],main="Time to Market") #Outliers detected
# boxplot(data_train[,"Mutual.Trust"],main="Trust") #Outliers detected
# boxplot(data_train[,"Work.Experience"],main="Work Experience") #Outliers NOT detected
#
# par(mfcol=c(1,1)) #Reset the grid size
# dev.off() #Clear Plots
# ----------------------------------------------------------------------------

#Box plots for the numeric dependent variables
# boxplot(data_train[,"Employees"]) #The presence of Outliers is expected
# boxplot(data_train[,"Total.Funding"]) #The presence of Outliers is expected


#Create Function to provide outliers indexes
outliers_pos<-function(vec){
q<-as.numeric(quantile(vec))
IQR<-q[4]-q[2]
upper_limit<-q[4]+1.5*IQR
outliers<-which(vec>upper_limit)
return (outliers)
}

#Replace Outliers
data_train$Alliances <- replace(data_train$Alliances,outliers_pos(data_train$Alliances),median(data_train$Alliances))
data_train$Team.Size <- replace(data_train$Team.Size,outliers_pos(data_train$Team.Size),median(data_train$Team.Size))
data_train$M2M <- replace(data_train$M2M,outliers_pos(data_train$M2M),median(data_train$M2M))
data_train$Pivoting <- replace(data_train$Pivoting,outliers_pos(data_train$Pivoting),median(data_train$Pivoting))
data_train$Previous.Salary <- replace(data_train$Previous.Salary,outliers_pos(data_train$Previous.Salary),median(data_train$Previous.Salary))
data_train$T2M <- replace(data_train$T2M,outliers_pos(data_train$T2M),median(data_train$T2M))
data_train$Mutual.Trust <- replace(data_train$Mutual.Trust,outliers_pos(data_train$Mutual.Trust),median(data_train$Mutual.Trust))

# #Plot Histograms once more to check outliers cleaning
# data_train[,num_var] %>% gather() %>%
#   ggplot(aes(value)) +
#   facet_wrap(~ key, scales = "free") +
#   geom_histogram()


#--------------------------------------------------------------------------------------------------
#----------------------------------------Transformation----------------------------------------
#--------------------------------------------------------------------------------------------------

#Categorical variables removed
data_train <- subset(data_train,select=-c(Motivation,Funding.Source,Mentorship))
# str(data_train)

#Categorical variables
data_train$MVP <- as.factor(ifelse(data_train$MVP=="No, I directly built the MVP.","no prototype or poc","prototype or poc"))

library(OneR) #For using the bin function

#QS Factor to Dichotomous
data_train$University.Ranking <- as.factor(ifelse(data_train$University.Ranking>"0","YES","NO")) # Threshold = 0

#Numerical variables
levels(bin(data_train$Alliances,3,method="content")) #levels: 0, 1-2, 3+
data_train$Alliances <- bin(data_train$Alliances,3,label=c("low","medium","high"),method="content")

levels(bin(data_train$Team.Size,3,method="content")) #levels: 0-1,2-3,4+
data_train$Team.Size <- bin(data_train$Team.Size,3,label=c("low","medium","high"),method="content")

```

```
levels(bin(data_train$M2M,3,method="content")) #levels: 0 -32k, 32k - 75k, 75k +
data_train$M2M <- bin(data_train$M2M,3,label=c("low","medium","high"),method="content")

levels(bin(data_train$Number.of.Founders,2,method="content")) #levels: 1, 2, 3+
data_train$Number.of.Founders <- bin(data_train$Number.of.Founders,3,label=c("1","2","3+"),method="content")

levels(bin(data_train$Pivoting,3,method="content")) #levels: 0,1-2,3+
data_train$Pivoting <- bin(data_train$Pivoting,3,label=c("low","medium","high"),method="content")

levels(bin(data_train$Previous.Salary,5,method="content")) #levels: 0-1.5k,1.6k-3k,3k-5k,5.1k-9k,9.1k+
data_train$Previous.Salary <- bin(data_train$Previous.Salary,5,label=c("low","average","above average", "high", "very high"), method="content")

levels(bin(data_train$T2M,3,method="content")) #<8, 9-18, 19+
data_train$T2M <- bin(data_train$T2M,3,label=c("very fast","average","slow"),method="content")

levels(bin(data_train$Mutual.Trust,3,method="content")) #<=20, 21-36, 37+
data_train$Mutual.Trust <- bin(data_train$Mutual.Trust,3,label=c("low","medium","high"),method="content")

levels(bin(data_train$Work.Experience,4,method="content")) #0-4,5-10,11-19,20+
data_train$Work.Experience <- bin(data_train$Work.Experience,4,label=c("low","medium","high","very high"),method="content")

#Convert transformed variables to numeric
data_train$Alliances<- as.numeric(data_train$Alliances)
data_train$Team.Size <- as.numeric(data_train$Team.Size)
data_train$M2M <- as.numeric(data_train$M2M)
data_train$Number.of.Founders <- as.numeric(data_train$Number.of.Founders)
data_train$Pivoting <- as.numeric(data_train$Pivoting)
data_train$Previous.Salary <- as.numeric (data_train$Previous.Salary)
data_train$T2M <- as.numeric (data_train$T2M)
data_train$Mutual.Trust <- as.numeric (data_train$Mutual.Trust)
data_train$Work.Experience <- as.numeric(data_train$Work.Experience)

# #Check New Histogram with Discretized Variables
# data_train[,num_var] %>% gather() %>%
#   ggplot(aes(value)) +
#   facet_wrap(~ key, scales = "free") +
#   geom_histogram()

#Likert - Ordinal Variables

# likert_var <- c("Entrepreneurial.K.E","Social.Capital",
#                 "Risk.Profile","Team.Diversity",
#                 "Organicity","Tech.Orientation",
#                 "Data.Orientation","Societal.Relevance")

# #Histogram of Ordinal Variables
# data_train[,likert_var] %>% gather() %>%
#   ggplot(aes(value)) +
#   facet_wrap(~ key, scales = "free") +
#   geom_histogram()

data_train$Team.Diversity <- as.factor(ifelse(data_train$Team.Diversity>="5","YES","NO"))
data_train$Data.Orientation <- as.factor(ifelse(data_train$Data.Orientation>="4","YES","NO"))
data_train$Entrepreneurial.K.E <- as.factor(ifelse(data_train$Entrepreneurial.K.E>="4","High","Low"))
data_train$Organicity <- as.factor(ifelse(data_train$Organicity>="6","High","Low"))
data_train$Risk.Profile <- as.factor(ifelse(data_train$Risk.Profile>="6","Risk-Taker","Risk-Averse"))
data_train$Social.Capital <- as.factor(ifelse(data_train$Social.Capital>="5","High","Low"))
data_train$Societal.Relevance <- as.factor(ifelse(data_train$Societal.Relevance>="6","YES","NO"))
data_train$Tech.Orientation <- as.factor(ifelse(data_train$Tech.Orientation>="6","YES","NO"))

#Dichotomous variables don't need to be transformed into numerical variables straight away.
#They do have to be transformed in the case we need to do correlation analysis (further).

# #Let's plot the transformed ordinal variables
# library(gridExtra)
#
# p1 <- ggplot(data_train,aes(x=Team.Diversity))+geom_bar(stat="count")
# p2 <- ggplot(data_train,aes(x=Data.Orientation))+geom_bar(stat="count")
# p3 <- ggplot(data_train,aes(x=Entrepreneurial.K.E))+geom_bar(stat="count")
# p4 <- ggplot(data_train,aes(x=Organicity))+geom_bar(stat="count")
# p5 <- ggplot(data_train,aes(x=Risk.Profile))+geom_bar(stat="count")
# p6 <- ggplot(data_train,aes(x=Social.Capital))+geom_bar(stat="count")
# p7 <- ggplot(data_train,aes(x=Societal.Relevance))+geom_bar(stat="count")
# p8 <- ggplot(data_train,aes(x=Tech.Orientation))+geom_bar(stat="count")
#
# grid.arrange(p1,p2,p3,p4,p5,p6,p7,p8,nrow=2)
# rm(p1,p2,p3,p4,p5,p6,p7,p8)


# ----------------------------------------------------------------------------------------------------------------
#--------------------------------------------Exploration---------------------------------------------------------
#----------------------------------------------------------------------------------------------------------------

# #Scatter Plot of Total Fundings vs Employees
# ggplot(data_train,aes(x=Employees,y=Total.Funding)) + geom_point(size=3) + scale_x_log10() + scale_y_log10() +ylab("Total Funding (EUR M)")

# #We can observe a positive linear correlation. Let's calculate this correlation.
# #cor.test(data_train$Employees,data_train$Total.Funding) #We obtain a high and significant correlation.
#
# #Let's now explore the new proposed success metrics
# #xtabs(~Revenue.Growth + Profitability,data_train=data_train)
# #Note: I didn't expect to have startups in the upper right and lower left corners.
#
# #Scatter plot 1: Funding vs Employees vs Profitability ----------------------------------------------------
# ggplot(data_train,aes(x=Employees,y=Total.Funding,color=Profitability)) + geom_point(size=3) +scale_x_log10() + scale_y_log10() + ylab("Total Funding (EUR M)")
#
# #Note: we observe profitable startups clustered in the lower left corner.
# #Hypothesis: More typical of SMEs, less investment, less risk.
#
# #Scatter plot 2: Funding vs Employees vs Revenue.Growth (a.k.a Scaleup) ----------------------------------------------------
#
#    ggplot(data_train,aes(x=Employees,y=Total.Funding,color=Revenue.Growth)) + geom_point(size=3) +scale_x_log10() + scale_y_log10() + ylab("Total Funding (EUR M)") + geom_vline(xintercept=10,linetype="dashed",color="black",size=1)
#    #Note: we observe that dots are gathered towards the right side of the chart.
#
#    #Hypothesis 1: The Revenue.Growths (y/n) variable is not related to total.funding.
#
#      p1 <- ggplot(data_train,aes(x=Total.Funding,y=Revenue.Growth,color=Revenue.Growth)) + geom_count() + scale_x_log10() + theme(legend.position = "none")+xlab("Total Funding (EUR M)") + ylab("Revenue Growth (Scaleup)" )
#      #Note: total funding is somehow related to the Revenue.Growth variable. Why??? Not sure.
#
#      #model <- glm(Revenue.Growth~Total.Funding,family="binomial",data_train)
#      #summary(model) #Total Funding 0.3137 P: 0.0129 *
#
#    #Hypothesis 2: The Revenue.Growths (y/n) variables is related to employee size
#      p2 <- ggplot(data_train,aes(x=Employees,y=Revenue.Growth,color=Revenue.Growth)) + geom_point() + scale_x_log10() + theme(legend.position = "none",axis.title.y=element_blank(),axis.text.y=element_blank()) + geom_vline(xintercept=10,linetype="dashed",color="black",size=1)
#      #It shows a much stronger relationship. Let's confirm by creating a logistic regression of one variable.
#
#      #model2 <- glm(Revenue.Growth~Employees,family="binomial",data_train)
#      #summary(model2) #Total Funding 0.03837 P: 0.0234 *
#
#
#    grid.arrange(p1,p2,nrow=1)
#
#
# category <- ifelse(data_train$Revenue.Growth=="YES"&data_train$Profitability=="YES","Profitable & Rapid Growth",
#                          ifelse(data_train$Revenue.Growth=="NO"&data_train$Profitability=="YES","Profitable",
#                           ifelse(data_train$Revenue.Growth=="YES"&data_train$Profitability=="NO","Rapid Growth","Trying")))
#
```

```
# data_train_temp <- cbind(data_train,category)
#
# ggplot(data_train_temp,aes(x=Employees,y=Total.Funding,color=category)) + geom_point(size=4) + scale_x_log10() + scale_y_log10() + theme(legend.position="top") +  scale_color_manual(values = c("#ac37f0","#18a5d9","#ed9c1a","#dbdbdb")) +ylab("Total Funding (EUR M)")
#
# rm(data_train_temp)

# ------------------------------------------------------------------------------------------------
#----------------------------------MODEL EXPLORATION----------------------------------------------
#-------------------------------------------------------------------------------------------------


emp_10 <- ifelse(data_train$Employees>=10,"YES","NO")
data_train <- cbind(data_train,emp_10)

fund_1 <- ifelse(data_train$Total.Funding>=1,"YES","NO")
data_train <- cbind(data_train,fund_1)


#---------------------Model 1: Employees ---------------------------------------------------------


DV <- "emp_10"
DV_index <- as.numeric(which(colnames(data_train)==DV))
subset <- data_train[,c(2,17,20,27,DV_index)]
model_emp_10 <- glm(emp_10~.,family="binomial", subset)
# summary(model_emp_10)

# #Anova Test
# anova(model_emp_10,test="Chisq")

# #Mc Fadden Pseudo R2
# ll.null <- model_emp_10$null.deviance/-2
# ll.proposed <- model_emp_10$deviance/-2
# MFR2<- (ll.null-ll.proposed)/ll.null
# p_value <- 1-pchisq(2*(ll.proposed-ll.null),df=(length(model_emp_10$coefficients)-1))
# MFR2
# p_value

# #ROC Curve
# library(pROC)
# par(pty = "s")
# roc(data_train$emp_10,model_emp_10$fitted.values,plot=TRUE,
#      legacy.axes=TRUE, percent = TRUE, col="#377eb8",lwd=4,
#      xlab="False Positive Percentage",ylab="True Positive Percentage")


#---------------------Model 2: Revenue.Growth ---------------------------------------------------------

DV <- "Revenue.Growth"
DV_index <- as.numeric(which(colnames(data_train)==DV))
subset <- data_train[,c(10,27,DV_index)]
model_rev <- glm(Revenue.Growth~.,family="binomial", subset)
summary(model_rev)


#Anova Test
anova(model_Revenue.Growth,test="Chisq")

#Mc Fadden Pseudo R2
ll.null <- model_Revenue.Growth$null.deviance/-2
ll.proposed <- model_Revenue.Growth$deviance/-2
MFR2<- (ll.null-ll.proposed)/ll.null
p_value <- 1-pchisq(2*(ll.proposed-ll.null),df=(length(model_Revenue.Growth$coefficients)-1))
MFR2
p_value

#ROC Curve
par(pty = "s")
```

```
roc(data_train$Revenue.Growth,model_Revenue.Growth$fitted.values, plot=TRUE,
legacy.axes=TRUE, percent = TRUE, col="#377eb8",lwd=4,
xlab="False Positive Percentage",ylab="True Positive Percentage")

#-------------------------Model 3: Total Funding ----------------------------------------------

DV <- "fund_1" #Significant variables: 8,10,13,20,27
DV_index <- as.numeric(which(colnames(data_train)==DV))
subset <- data_train[,c(8,10,13,18,27,DV_index)]
model_funding<- glm(fund_1~.,family="binomial", subset)
# summary(model_funding)
#
#
# #Anova Test
# anova(model_funding, test="Chisq")
#
# #Mc Fadden Pseudo R2
# ll.null <- model_funding$null.deviance/-2
# ll.proposed <- model_funding$deviance/-2
# MFR2<- (ll.null-ll.proposed)/ll.null
# p_value <- 1-pchisq(2*(ll.proposed-ll.null),df=(length(model_funding$coefficients)-1))
# MFR2
# p_value
#
# #ROC Curve
# par(pty = "s")
# roc(data_train$fund_1,model_funding$fitted.values, plot=TRUE,
#      legacy.axes=TRUE, percent = TRUE, col="#377eb8",lwd=4,
#      xlab="False Positive Percentage",ylab="True Positive Percentage")


# ---------------------------------------------------------------------------------------------------
#-------------------------------------Correlation Analysis-------------------------------------------
#---------------------------------------------------------------------------------------------------

# predictors <- c("Number.of.Founders","University.Ranking",
#                 "Time.Dedication","Team.Size",
#                 "Data.Orientation","Societal.Relevance",
#                 "Pivoting","Employee.Incentives") #Significant Predictors
#
# criteria <- c("Revenue.Growth","emp_10","fund_1") #Criteria of Success (Three models)
#
# mod_var <- c(predictors, criteria)
#
# subset <- data_train[,mod_var]
#
# subset <- as.data.frame(sapply(subset,as.numeric))
#
# #Correlation Matrix
# CM_ME <-round(cor(subset), digits=2)
# library(corrplot)
# corrplot(CM_ME,method = "square",type="lower")

#Notes: no collinearity problems.

#Receiver Operating Characteristic (ROC) Curve


# ---------------------------------------------------------------------------------------------------
#-------------------------------------DATA PREPARATION: TEST SET-------------------------------------
#---------------------------------------------------------------------------------------------------

#Only significant variables
data_test <- subset(data_test, select = c(Name,Number.of.Founders,University.Ranking,
Time.Dedication,Team.Size,Data.Orientation,
Societal.Relevance,Pivoting,T2M,Employee.Incentives,
Total.Funding,Employees,Revenue.Growth))



#Categorical variables as Factors (Independent 11)
```

```r
data_test$Time.Dedication<- as.factor(data_test$Time.Dedication) #2 Levels: Full time vs part time
data_test$Employee.Incentives <- as.factor(data_test$Employee.Incentives) #2 Levels: yes/no


#Categorical variables as Factors (Independent 2)
data_test$Revenue.Growth <- as.factor(data_test$Revenue.Growth) #DV related to rapid growth

#Replace NA Values from QS Rating to 0 score
data_test$University.Ranking[is.na(data_test$University.Ranking)] <- 0

# #Numeric variables vector (10)
# num_var_test<- c("Number.of.Founders","Team.Size","Pivoting","T2M")
#
# #Histogram Plots to Spot Outliers
# data_test[,num_var_test] %>% gather() %>%
#   ggplot(aes(value)) +
#   facet_wrap(~ key, scales = "free") +
#   geom_histogram()

#Replace Outliers
data_test$Team.Size <- replace(data_test$Team.Size,outliers_pos(data_test$Team.Size),median(data_test$Team.Size))
data_test$T2M <- replace(data_test$T2M,outliers_pos(data_test$T2M),median(data_test$T2M))

#Data transformation
data_test$Number.of.Founders <- ifelse(data_test$Number.of.Founders==1,1,ifelse(data_test$Number.of.Founders==2,2,3))
data_test$University.Ranking <- as.factor(ifelse(data_test$University.Ranking==0,"NO","YES"))
data_test$Team.Size <- ifelse(data_test$Team.Size<=1,1,ifelse(data_test$Team.Size<=3,2,3))
data_test$Data.Orientation <- as.factor(ifelse(data_test$Data.Orientation>=4,"YES","NO"))
data_test$Societal.Relevance <- as.factor(ifelse(data_test$Societal.Relevance>="6","YES","NO"))
data_test$Pivoting <- ifelse(data_test$Pivoting==0,1,ifelse(data_test$Pivoting<=2,2,3))
data_test$T2M <- ifelse(data_test$T2M<=8,1,ifelse(data_test$T2M<=18,2,3))

emp_10 <- ifelse(data_test$Employees>=10,"YES","NO")
fund_1 <- ifelse(data_test$Total.Funding>=1,"YES","NO")
data_test <- cbind(data_test,emp_10,fund_1)


#Model Testing (total funding)

Threshold <- 0.7
prediction <- predict(model_funding,data_test,type="response")
prediction <- ifelse(prediction > Threshold,"YES","NO")
misClasificError <- mean(prediction != data_test$fund_1)
print(paste('Accuracy',1-misClasificError))
table(data_test$fund_1,prediction) #Confusion Matrix

#Model Testing (employees)

Threshold <- 0.4
prediction <- predict(model_emp_10,data_test,type="response")
prediction <- ifelse(prediction > Threshold,"YES","NO")
misClasificError <- mean(prediction != data_test$emp_10)
print(paste('Accuracy',1-misClasificError))
table(data_test$emp_10,prediction) #Confusion Matrix

#Model Testing (rapid growth)

Threshold <- 0.5
prediction <- predict(model_rev,data_test,type="response")
prediction <- ifelse(prediction > Threshold,"YES","NO")
misClasificError <- mean(prediction != data_test$Revenue.Growth)
print(paste('Accuracy',1-misClasificError))
table(data_test$Revenue.Growth,prediction) #Confusion Matrix
```

Acronyms

ANN Artificial Neural Network. *Glossary:* artificial neural network, 21

CFA Confirmatory Factor Analysis. *Glossary:* confirmatory factor analysis, 21

GEM Global Entrepreneurship Monitor. 18

GII Global Innovation Index. 18

KPI Key Performance Indicator. 27

MVP Minimum Viable Product. 67

NIS National Innovation System. *Glossary:* national innovation system, 16

PCA Principal Component Analysis. *Glossary:* principal component analysis, 22

RIS Regional Innovation System. *Glossary:* regional innovation system, 16

SEM Structural Equation Model. *Glossary:* structural equation model, 23

SME Small and Medium Enterprise. 3

SVM Support Vector Machine. *Glossary:* support vector machine, 21

TAM Total Addressable Market. *Glossary:* total addressable market, 30

TFE Total Full Employment. *Glossary:* total full employment, 27

YOY Year-over-Year. 27

GLOSSARY

artificial neural network a biologically inspired, non-parametric learning algorithm that can model extremely complex non-linear functionDellermann et al. [2018]. 21

big bang disruption a dramatic new kind of innovation. Instead of entering the market as a product that is either inferior to or more expensive than those of established incumbents, a Big Bang Disruptor is both better and cheaper from the moment of creation [Review, 2013]. 3

closed innovation contrary to the term open innovation, closed innovation assumes that the best route to innovation is to have control over the firm's processes and resources [Chesbrough, 2003]. 16

confirmatory factor analysis multivariate statistical procedure that is used to test how well the measured variables represent the number of constructs [Solutions, 2013]. 21

fintech financial technology, often shortened to fintech, is the technology and innovation that aims to compete with traditional financial methods in the delivery of financial services. It is an emerging industry that uses technology to improve activities in finance [Lin, 2016].. 18

industrial cluster geographic concentrations of interconnected companies and institutions in a particular field", a typical example is Silicon Valley, a cluster where suppliers, customers, employees, governmental institutions and academia converged due to the explosive growth in the number of companies doing semiconductor research in the 1960s [Porter, 1998]. 16

innovation system the interaction and flow of information among people, enterprises and institutions to drive innovation. 16

logistic regression a very well-known linear regression algorithm used as the baseline algorithm, frequently applied for binary choice models [Dellermann et al., 2018]. 21

naive bayes bayesian parameter estimation problem based on some known prior distribution [Dellermann et al., 2018]. 21

national innovation system "the elements and relationships which interact in the production, diffusion ande use of new, and economically useful, knowledge...and are either located within or rooted inside the borders of a nation-state" Lundvall [2007]. 16

open innovation a paradigm that assumes that firms can and should use external ideas as well as internal ideas, and internal and external paths to market, as the firms look to advance their technology [Chesbrough, 2003]. 16

principal component analysis a principal component analysis describes a number of variables with a smaller number of variables, termed the principal components, that still contain as much information, exhibited in the original variables, as possible. 22

radical innovation an invention that destroys or supplants an existing business model. Unlike architectural or incremental innovation, radical innovation blows up the existing system or process and replaces it with something entirely new. Some see radical innovation and disruptive innovation as interchangeable terms.[TechTarget, 2013]. 3

random forests a popular ensemble method that minimizes variance without increasing bias by bagging and randomizing input variables [Dellermann et al., 2018]. 21

regional innovation system a system stimulating innovation capabilities of firms in a region so as to enhance the region's growth potential and regional competitiveness[Cooke et al., 1997]. 16

reporting lag concepts used to the phenomenon that explains why investment data always show a decrease in the last twenty-four months.. 39

scaleup a company who has an average annualized return of at least 20% in the past 3 years with at least 10 employees in the beginning of the period Institute [2016]. 3, 6, 13, 27, 34, 45, 55, 56, 60, 67

startup a company initiated by individual founders or entrepreneurs to search for a repeatable and scalable business model Blank [2013]. 3

startup ecosystem a system formed by startups, investors and various types of organization working together to create and scale new startup companies. 3

structural equation model statistical method frequently used to study the structural relationship between factors by means of factor analysis and linear regressions [Nalintippayawong et al., 2018]. 23

support vector machine classification algorithm based on a linear discriminant function, which uses kernels to find a hyperplane that separates the data into different classes[Dellermann et al., 2018]. 21

total addressable market revenue opportunity if the totallity of possible customers was served. 30

total full employment FTE stands for full-time equivalent (not full-time employee) and translates the total hours worked by part-time employees into the number of equivalent full-time employees. To calculate FTE, you have to know how many employees you have, and the average number of hours they work. You can then determine the equivalent number of full-time workers you employ. [Handrick, 2018]. 27

triple helix the interaction of government, industry and academia to foster social and economic development Leydesdorff [2010]. 16

unicorn company with a valuation over 1 Billion Euros. Definitions may vary across countries. In the United States a company is considered a startup if and only if this has reached the 1 Billion USD valuation within a time frame of 5 years from launch.. 3

Bibliography

Ayyagari, M., Demirguc, K., Asli, M., and Vijislav (2011). Small young firms across the world: contribution to employment, and growth. *The World Bank*.

Azoulay, P., Jones, B., Kim, D., and Miranda, J. (2018). The average age of a successful startup founder is 45. Online article.

Bento, F. R. d. S. R. (2017). *Predicting Startup Success with Machine Learning*. Thesis.

Bichara, T. (2018). Why an MVP is not a prototype. Online article.

Blank, S. G. (2013). *Fourt Steps to the Epiphany*. 2nd edition.

Bosma, N. and Kelley, D. (2018). Global entrepreneurship monitor. Report.

Bull, I. and Willard, G. E. (1993). Towards a theory of entrepreneurship. *Journal of Business Venturing*, 8(3):183–195.

Chesbrough, H. (2003). *Open Innovation: The New Imperative for Creating and Profiting fromTechnology*.

Compass (2015). The global startup ecosystem ranking 2015. Report.

Cooke, P., Gomez Uranga, M., and Etxebarria, G. (1997). Regional innovation systems: Institutional and organisational dimensions. Journal Article 4.

Cuervo Garcia, A., Ribeiro, D., and Roig, S. (2007). *Entrepreneurship : concepts, theory and perspective*. Springer, Berlin, 1st edition.

Dellermann, D., Lipusch, N., Ebel, P., Popp, K. M., and Leimeister, J. M. (2018). Finding the unicorn: Predicting early stage startup success through a hybrid intelligence method. Conference article.

Dutta, S. and Lanving, B. (2018). Global innovation index. Report, Cornell SC Johnson College of Business, INSEAD, World Intelectual Property Organization.

Fundz (2018). What is Series A funding, Series B funding and more. Online article.

Gelderen, M. v., Thurik, R., and Bosma, N. (2005). Success and risk factors in the pre-startup phase. *Small Business Economics*, 24(4):365–380.

Genome, S. (2018). Global startup ecosystem report. Report, Genome.

Graham, P. (2012). Startups equal growth. Web page.

Groenewegen, G. and Langen, F. (2012). Critical success factors of the survival of start-ups with a radical innovation. *Journal of Applied Economics and Business Research*, 2:155–171.

Haltiwanger, J. C. and Jarmin, R. S. (2010). Who creates jobs? small vs large vs young. *US Census Bureau Center for Economic Studies Paper*.

Handrick, L. (2018). What is TFE and how to calculate it. Webpage.

Heimans, J. and Timms, H. (2014). Understanding "new power". Online article.

Institute, S. (2016). The scaleup manifesto. Webpage.

Joffe, B. and Eversweiler, C. (2018). What every startup founder should know about exits. Online article.

Kane, T. (2010). The importance of startups in job creation and job destruction. Report.

Kuratko, D. F. (2017). *Entrepreneurship : theory, process, practice*. Cengage Learning, Boston, USA, 10th edition.

Kwabena, N. and Simpeh, K. (2011). Entrepreneurship theories and empirical research: A summary review of the literature. *European Journal of Business Management*, 3:1–8.

Latham, J. (2018). Dutch startup ecosystem map. Figure.

Leydesdorff, L. (2010). The knowledge-based economy and the triple helix model. *Annual Rev. Info. Sci & Technol.*, 44(1):365–417.

Limsong, S., Sambath, P., Seang, S., and Hong, S. (2016). A model of entrepreneur success: Linking theory and practice. *WEI International Academic Conference Proceeding*.

Lin, T. C. (2016). Infinite financial intermediation. Journal article.

Lundstrom, A. and Stevenson, L. (2005). *Entrepreneurship policy : theory and practice*. Springer, New York.

Lundvall, B.-k. (2007). *National Innovation Systems—Analytical Concept and Development Tool*, volume 14.

Marmer, M., Herrmann Lasse, B., Dogrultan, E., and Berman, R. (2011). A new framework for understanding why startups succeed. Report.

M.Sc. Steigertahl, L., Prof. Dr. Mauer, R., Say, J.-B., and Entrepreneurship, I. o. (2018). EU startup monitor. Report.

Nalintippayawong, S., Waiyawatpattarakul, N., and Chotipant, S. (2018). Examining the critical success factors of startup in thailand using structural equation model. pages 388–393.

OECD (2017). Enterprises by business size. Web page.

Porter, M. E. (1998). Clusters and the new economics of competition. Report.

Review, H. B. (2013). Big-bang disruption. Webpage.

Ries, E. (2011). *The Lean Startup*. Crown Business, United States, 1 edition.

Schumpeter, J. (1934). *The theory of economic development*. Harvard University Press, Cambridge, 6th edition.

Schumpeter, J. (1942). *Capitalism, Socialism, and Democracy*, volume 3.

Schumpeter, J. A. (1947). The creative response in economic history. *The Journal of Economic History*, 7(2):149–159.

Sledzik, K. (2013). Schumpeter's view on innovation and entrepreneurship. *SSN Electronic Journal*.

Solutions, S. (2013). Confirmatory factor analysis. Webpage.

Stam, E. (2014). *The Dutch Entrepreneurial Ecosystem*.

StartupCommons (2019). What is a startup? Online Article February.

TechTarget (2013). Radical innovation definition. Webpage.

Thiel, P. (2014). *Zero to One*. Crown Publishing Group, United States, 1 edition.

Valley, N. (2018). Startup life cycle. Figure.

Young, E. . (2013). The EY G20 entrepreneurship barometer 2013. Report.