IRTG 1792 Discussion Paper 2019-022

# A Machine Learning Approach Towards Startup Success Prediction

Cemre Ünal *
Ioana Ceasu *

* Humboldt-Universität zu Berlin, Germany

This research was supported by the Deutsche Forschungsgesellschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

# A Machine Learning Approach Towards Startup Success Prediction

Discussion Papaer

by

**Cemre Ünal**

**Ioana Ceasu**

Berlin, August, 2019

## Abstract

The importance of startups for a dynamic, innovative and competitive economy has already been acknowledged in the scientific and business literature. The highly uncertain and volatile nature of the startup ecosystem makes the evaluation of startup success through analysis and interpretation of information very time consuming and computationally intensive. This prediction problem brings forward the need for a quantitative model, which should enable an objective and fact-based approach to startup success prediction.

This paper presents a series of reproducible models for startup success prediction, using machine learning methods. The data used for this purpose was received from the online investor platform, crunchbase.com. The data has been pre-processed for sampling bias and imbalance by using the oversampling approach, ADASYN. A total of six different models are implemented to predict startup success. Using goodness-of-fit measures, applicable to each model case, the best models selected are the ensemble methods, random forest and extreme gradient boosting with a test set prediction accuracy of 94.1% and 94.5% and AUC of 92.22% and 92.91% respectively. Top variables in these models are last funding to date, first funding lag and company age. The models presented in this study can be used to predict success rate for future new firms/ventures in a repeatable way.

# 1 Introduction

Startups have become an important topic in the economic policies of all developed and emerging economies around the world, not just by being a driver of economic prosperity and wealth, but also because of their major impact on innovation and technological development (Luger and Koo, 2005). Their fast growth rates, agility in deploying innovative business models and state of the art technologies, together with their fail fast and lean management approach turn them into disruptive actors in the global economy, especially since their business playground is frequently a global one. Their dynamic, and sometimes irreverent approach to business, challenges the status quo of traditional corporate business and even that of classical SMEs. Hence, the success of these new ventures is closely connected to the strategic development interests of the society (Shane, 2012). However, 90% of startups fail within the first year of their founding and less than 40% of the remaining 10% pass the 5-year milestone (Regmi et al., 2015).

In the case of startups, the stakeholders are first of all the entrepreneurs, who benefit from prediction models regarding the success or failure of their business ideas as they can make educated decisions addressing potential critical points within their business models, have the ability to pivot in a timely manner and save resources (financial, human, etc.), which are usually scarce within a startup. Other important stakeholders are startup investors (which, depending on the investment stage, can be angel investors, seed money funds, venture capital investors, etc.), who ideally benefit from such prediction models by increasing on their traditional 10% success rate with startups (Shane, 2012). Last but not least, the rest of the players have a stake in being better informed with regards to whether a startup will succeed or fail, also bearing risks related: on the one hand, the suppliers, who need to create or manage new supply chain systems and, on the other hand, the clients/customers, who might rely on the new product or service.

The environment in which startups grow and develop is very complex and risk prone, so that there are numerous intrinsic and extrinsic variables to be taken into consideration in building a prediction model. The issue becomes even more difficult for young startups, as they cannot provide any historical financial or operational data. Most of the available data is at best sparse and qualitative, and from multiple sources. In this context, it is difficult for entrepreneurs or investors to make educated and objective decisions, since humans tend to be

selective in the information they use and suffer from bias when making decisions. Some claim that intuition and gut feeling, based on decision-makers' previous experience and expertise are the best instruments in decision making when it comes to startups. Einhorn (1974), for example, states that humans are more than capable of using their intuition and making decisions based on subjective judgment of the information and are proven to recognize and use rare information pieces in various decision-making environments, where it would have been difficult to predict the outcome with an algorithm. However, there is research within the social sciences literature which challenges this argument.

Behavioural economics focuses on the bounded rationality of decision-makers and, therefore, their proneness to make errors. Bounded rationality has been linked to the limitations of humans to process vast amounts of information in a rational way (Venkatraman et al., 2009; Simon, 1955). This argument is valid, especially in today's business environment, where managers and entrepreneurs are flooded by information and data, some of which is useful and some of which is not, when considering dynamic and unstable business contexts, such as those of startups. In practice, attention and time required to collect and process information are scarce resources. Thus, decision-makers are not always paying enough attention to all the information available and, hence, cannot and will not process the underlying connections between the various pieces of information and their sources. As a consequence, decision-makers have a tendency to consider the information they value more than the information they define as unimportant. *Dual-Process Theory* proposed by Fischhoff et al. (2002), suggests two systems of thinking: (i) System 1 quickly supplies intuitive answers to judgment problems as they surface and can be described as automatic, effortless and associative, while (ii) System 2 concerns the analytical approach. When there is no experience or reference for intuitive decision-making under the so-called System 1 thinking (Fischhoff et al., 2002), poor subjective evaluations of information on hand can lead to inefficient and poor choices. Similarly, Read and Van Leeuwen (1998) claim that overestimation of the ones' own skills, lead to the use of heuristics in order to solve complex problems. Affection by transient emotions and fluctuations in attention to the different pieces of information also influence decisions and make the decisions time-variant (Luce, 1959). Therefore, bounded rationality of the decision-makers induces motivation towards a quantitative approach.

Given these limitations of decision-makers with respect to information evaluation, this paper

aims at constructing an appropriate quantitative model to predict whether a startup will succeed or fail. In the past decades, there has been extensive research on survival prediction for corporate companies, in which success drivers are strongly associated with historical financial data and KPIs. However, historical financial, sales and production data do not always exist for startups, which are an important component of success prediction of corporate companies. Startups' success is based on different dynamics, such as rapid digitization, use of innovative business models, etc. Therefore, this paper approaches the startup success prediction differently than common research for conventional company success prediction does.

This paper has 6 sections. Section 2 offers a review of the related work in the scientific literature. Section 3 presents and discusses the data, the methodology and modeling methods. Results are presented in Section 4. Sections 5 and 6 articulate the conclusion and future research possibilities, respectively.

The codes of the models and the results can be reached via respective quantlet (www.quantlet.de) links.

# 2 Literature Review

The startup definition is a controversial topic in literature. Luger and Koo (2005) emphasizes three characteristics when describing startups: New, active and independent. *New* implies the establishment of a company which did not exist before. *Active* excludes the companies which are established recently, but only exist on paper for administrative purposes such as tax avoidance. Lastly, *independent* implies that the startup is not part of an established parent company/holding. The problem of defining the "startup" concept in the dedicated research has been linked to the data measurement and collection. Hence, many researchers define startups based on the available information in their data set (Luger and Koo, 2005). In this paper, the definition of startups is also based on the available data and considers companies, which are active for less than 10 years in the industries defined by the S&P500. Due to lack of information, it is not possible to identify and exclude spin-offs and startups that are founded by larger corporations.

## 2.1 Startup Performance vs. Business Success

A lot of the research on the topic of business success focuses on corporate and SME success. In this context, the health of a firm in a competitive business environment is highly associated with its profitability and the level of financial solvency. Butler and Fitzgerald (1999) associates business success with competitive performance of the firm against its competitors. Lussier and Pfeifer (2001) considers firms as successful if they made at least industry average profits for the last 3 years. Gatev et al. (1996), on the other hand, define success as continuance of operations without owing to creditors and shareholders.

In the context of startups the definition of business success for corporate companies or SMEs does not apply. First and foremost, the majority of the early-stage startups does not generate profits and / or does not have stable, historical financial data. Therefore, the definitions of Lussier and Pfeifer (2001) or Gatev et al. (1996) do not apply. The competitive performance, on the other hand, is not always an objective metric to assess business success when it is constructed without comparative financial performance of the other players in the market, as financial KPIs can be analyzed only under an industry/peer comparison framework.

Studies focusing on what impacts startup performance frequently take an approach which examines the type of progress experienced by the new and dynamic ventures. Tavoletti (2013)

evaluates startup success by the potential of early international growth and the ability of the entrepreneur to generate valuable opportunities for its new venture. Another approach to startup success looks at the number and size of investments a startup receives (Dempwolf et al., 2014). The ability of the startup to gain traction and connect in an efficient and valuable manner to the local and global ecosystem, by proving scaling effects in a short period of time is also considered to be a measure of its performance (Ceausu et al., 2017). In a more holistic approach, Ozdemir et al. (2016) looks at startup success through a qualitative lens. They consider the global impact and contribution to the development of the entrepreneurial ecosystem as well as quantitative aspects such as revenues, users / clients and number of jobs created.

## 2.2 Corporate Bankruptcy vs. Startup Failure

There is a wide body of scientific literature dedicated to corporate and SME disruption of success or even bankruptcy. With respect to corporate / SME bankruptcy, Ooghe and De Prijcker (2008) argue that business failure is not a unique moment in time, but rather a process, with different triggers and turning points, along the life cycle of a business, i.e the disruption to success can happen in different ways. Ooghe and De Prijcker (2008) suggest three main trajectories for business failure. Firstly, there is the lack of success due to faulty management. Secondly, there is the failure of startups after a very rapid launch right at the beginning of establishing a business. The initial success is attributed to the personality traits of the management, but the company still faces failure due the neglected financial and operational duties during and after the growth phase. The third trajectory is the lack of financial sustainability due to general, immediate environments and / or corporate policies.

When it comes to startup failure, because of the more dynamic pace these ventures need to grow and develop at, the failure process window is much shorter than it is in the case of corporate companies or SMEs. Even though failure is a concept that is used frequently in the startup world (sometimes even with pride, as it is considered a source of valuable knowledge, experience and expertise mostly in North America), there are little to no scientific studies focusing on these startup specific dynamics / factors.

Ooghe and Waeyaert (2004) summarizes the factors influencing business success in five cat-

egories: (i) general environment (economics, technical advancements/aspects, foreign countries/currencies, politics etc.); (ii) immediate environment (suppliers, customers, creditors, competitors); (iii) management team characteristics (motivation, experience, skills, personality traits); (iv) corporate policy (strategy, investments, corporate governance) and (v) company characteristics (size, maturity, industry). Some other sources classify these factors under only two categories, i.e. industry specific characteristics and firm specific characteristics (Kauffman and Wang, 2001).

The research in corporate insolvency prediction has shown that data from capital markets and financial ratios (e.g cash flow/total sales, EBIT, EBITDA margins, net income etc.) based on firm's balance sheet, income and cash flow statements are proven to be useful not only in performance prediction of the established companies, but also the overall financial situation. Success prediction models in literature are designed to use financial ratios extensively, due to their standardized nature and availability for established firms. However, success prediction models for startups face some challenges. As previously stated, the majority of the early-stage startups do not generate any profits or do not have any stable financial data. This implies that the business success prediction of startups cannot be primarily based on quantitative data as for established companies. This makes the models constructed by using financial data irrelevant for startup success prediction, where this data does not exist. Even in rare cases, in which financial ratios exist for startups, these ratios by themselves may not be strong enough to build good models and other data sources are needed. Liu and Wu (2019) discusses how qualitative data can provide predictions as good as financial ratios. Also, solely using financial ratios has been heavily criticized by Doumpos and Zopounidis (2002). Dimitras et al. (1996) and Laitinen (1992) state that the financial ratios are only the symptoms but not the cause of the managerial, operating and financial problems.

## 2.3 Brief Review of Business Success Prediction Models

*"All models are wrong but some are useful.", George Box*

Business success prediction models aim to predict the status of the companies before any disruption of success happens. Ooghe and De Prijcker (2008) and du Jardin (2016) state that all firms fail in their own unique way. Directly attacking this classification problem with clustering algorithms will therefore have little use. It is important to study and analyze as

many failed firms as possible to learn and identify key factors that led to failure in the first place.

Bankruptcy prediction has been the subject of research for decades. Early studies in literature mostly rely on statistical modelling, which formalizes the relationship between variables. Statistical modelling makes predictions as accurate and consistent as possible in the context of financial decisions under extreme uncertainty (Jones and Olson, 2013). Most research has been focused on corporate bankruptcy and survival models of established companies and SMEs. The application of prediction models in this field goes back to the 1950-1960s. These models used information from financial statements such as financial ratios (Boritz and Kennedy, 1995). The early studies did not pay much attention to the ability and experience of the management team. Success prediction models traditionally used the data created by successful and unsuccessful companies from different industries. The validity of the models are assessed based on confusion matrix, i.e. Type I and Type II errors.

The research in the success prediction of early stage companies became predominant in the 1990's. Lussier (1995) implements one of the first non-financial models, which mainly used qualitative variables, in a regression model to predict new venture failure, called the $Lussier Model$. The original full model is based on 15 variables, i.e. record keeping and financial controls, capital, industry experience, management experience, planning, professional advisory, education, staffing, product/service timing, economic timing, age, partners, parent, minority business owner and marketing. There have also been many studies that show the relation between the success of a new venture and the skills and motivation of the management (Ooghe and De Prijcker, 2008).

Ooghe and De Prijcker (2008) recognizes the time dimension of success and the underlying non-financial factors. The authors emphasize the fragmented structure of the non-financial factors, which not only includes the management team, but also the relationship with different stakeholders. They come up with a framework to classify various bankruptcy cases according to the underlying reasons, as previously explained. The researchers have identified different sets of variables to be used as a proxy to predict bankruptcy of a business. du Jardin (2016), Wu (2010) and Lussier and Pfeifer (2001) state that following the multivariate discriminant analysis to differentiate between successful and failed companies, methods like logit

and probit analysis as well as linear programming have been developed and these have been frequently used. Independent of the predictive or statistical model used, the researchers have used Type I and Type II error as a basis for evaluation.

However, in the last decades, applying machine learning algorithms has become more popular, especially because many of them have proven to outperform statistical models. Although both approaches aim to learn from data, the main difference is that machine learning algorithms do not rely on rule-based programming. Cao et al. (1997) states that the continuous concern of the statistical models is the adequacy and correctness of the underlying assumptions and specifications. Haavelmo (1944) questions the validity of regression coefficients if the whole assumption of, for example, linear regression is wrong. In this framework, implementation of non-parametric models permits relaxed assumptions of the model structures.

# 3 Methodology

## 3.1 Measure of the Variables

The majority of the papers in the scientific literature present studies for which the authors have designed their own surveys and conducted interviews with the startup stakeholders in order to collect data directly from successful and failed companies. However, this approach has its limitations, such as required time, energy and the size of the used sample. Since in this paper the approach is to apply machine learning algorithms using a large amount of data to predict startup success, the data set is formed using data from the research application programming interface (API) of crunchbase.com.

## 3.2 Data Pre-Processing

The initial data set obtained from crunchbase.com had 215 729 observations with 23 variables. Data cleaning steps to obtain a complete data set are summarized in Table 1. After data cleaning, the list of variables to be used throughout this paper are summarized in Table 2. These variables provide a snapshot of the company at a given point in time.

| Action initiated | Dropped | Sample size | % |
|---|---|---|---|
| Initial observations extracted from crunchbase | | 215 729 | 100% |
| Dropped if total funding raised (USD) and # of funding rounds is missing | 95 787 | 119 942 | 55.6% |
| Only consider startups established after 2009 | 58 512 | 61 430 | 28.5% |
| Drop if the year founded and company name is missing | 8 143 | 53 287 | 24.7% |
| Drop if the domain information is missing | 1 681 | 51 606 | 23.9% |
| Drop if industry is missing | 628 | 50 978 | 23.6% |
| Drop if duplicate exists | 16 | 50 962 | 23.6% |
| Drop if region information is missing | 1 436 | 49 526 | 22.9% |
| Cleaning outliers of first funding lag, last funding lag and funding rounds | 1 224 | 48 302 | 22.3% |
| Drop if near zero of zero variance explanatory variables | 3 780 | 44 522 | 20.6% |

**Table 1:** Summary of data cleaning steps


Removing predictors has been thoroughly discussed in literature. However, Kuhn and Johnson (2013) discusses that removing variables helps reduce computing time and complexity of the models. Consider a predictor with uniform or almost uniform value, which are referred to as zero and near-zero variance predictors respectively. Such variables are not only uninformative about the characteristics of the data but also can harm the prediction accuracy. Zero and near-zero variance are calculated by dividing the unique values by the sample size and compared to a predefined threshold value. Such variables are not considered, for example, in tree-based classification models, since they do not provide varying information between classes. One approach to avoid information loss is to collect more data to abstain from zero or near-zero variance. Since collecting more information about the companies in the data set is not within the scope of this paper, the variables with zero or near zero variance are eliminated. Hence, the sectors energy, industrial, real estate and utilities as well as the continents Africa and Oceania are excluded from further analysis.

The original data defines startup status under four categories: (i) operating, (ii) acquired, (iii) IPO and (iv) closed. Chang (2004) discusses individual characteristics of each acquisition and IPO and the ambiguity in their definition. Depending on the dynamics of the deal, an acquisition can also represent failure (for example when the entrepreneur does not make any gains from the deal). There are also many unsuccessful/incomplete IPOs. However, these details of the transactions are usually not public for startups and are very resource intensive

| Variable name | Transformation | Used variables | Variable Type |
|---|---|---|---|
| Country Code | Based on country code, the respective company has been identified to avoid granularity | Continent | Categorical |
| Status | Failure: Closed<br>Success: Operating, acquired, IPO | Status | Categorical |
| Category Group List | Values with multiple industries have been split and major industry has been identified and mapped to the 11 industry classification in S&P500 | Sector | Categorical |
| Funding rounds | - | Funding rounds | Numeric |
| Total Funding (USD) | - | Total Funding (USD) | Numeric |
| Founded on | Company age has been calculated by subtracting foundation date from this year:<br>2019 - Founded on | Company Age | Numeric |
| First funding on | First funding lag is the years passed between foundation of a company and first funding received:<br>First funding on - founded on | First funding lag | Numeric |
| Last funding on | Last funding lag is the years passed between first funding and last funding received:<br>Last funding on - first funding received | Last funding lag | Numeric |
| Last funding to date | Last funding to date is the years passed since the company received the last funding to date:<br>2019 - last funding on | Last funding to date | Numeric |
| twitter_url<br>Facebook_url | A function is been created to identify the social media appearance of the firm:<br>Both: Twitter and Facebook active<br>Twitter: Only twitter<br>Facebook: Only Facebook<br>None: No social media appearance | Social | Categorical |

**Table 2:** Summary of data transformations

to obtain. Therefore, for the sake of simplicity and keeping all relevant information, startups, which are operating, acquired or issued an IPO are labelled as *successful* and startups, which are closed, are labelled as *failure*. Hence, company status (success vs. failure) is defined as the dependent variable within the framework of this paper.

There are 43 main industry categories in the raw data set. These industries are grouped under 11 industry sectors according to S&P500. These industries are: communication services; consumer discretionary; consumer staples; energy; finance; health; industrials; utilities; real estate; IT and materials. There is no company in the data set, which is doing business in the materials industry. The industry sectors energy, industrials, utilities and real estate have been removed due to near zero variance. Therefore, the total number of industries reduces to 6.

The next step is to investigate if there is an obvious difference between successful and failed companies. Figure 1 illustrates that successful and failed companies do not necessarily display different characteristics. Both types of companies have similar median values for company age, total funding (USD), number of funding rounds, first funding lag, last funding lag and last funding to date. This also supports the implementation and usage of machine learning algorithms as there is no distinct difference between the two groups, which makes the classification problem more difficult to deal with. However, characteristics of continuous variables do not differ strongly between successful and failed companies.

Table 3 gives an overview of the descriptive characteristics of the categorical variables. After feature transformation, the data reveals that 54% of the overall companies are based in the Americas. The Americas and Europe, are hosting almost 80% of the firms. The successful startups are in business mainly in consumer related industries (32%) and IT (31%) in the last 10 years. 68% of the companies have social media existence on multiple platforms. The general characteristics of the startups in the data set are in accordance with the current startup trends. Similar to continuous variables, categorical variables also do not differ strongly between two classes.

For the sake of the performance of the models built in the following sections, between-predictor correlations are needed and must be taken into consideration. No strong correlation

between variables is found. Hence, (multi-)collinearity is not being checked further.

After completing the above-mentioned data pre-processing steps, the final data set consists of 44 522 firms (20% of the initial sample size) and 19 variables. At this point, the class imbalance in the dependent variable is checked.

| | | Success | | Failure |
|---|---|---|---|---|
| Variable Name | Frequency | % | Frequency | % |
| Social | | | | |
| Both | 28 832 | 68% | 1 083 | 50% |
| Facebook | 2 577 | 6% | 69 | 3% |
| Twitter | 4 645 | 11% | 533 | 25% |
| None | 6 313 | 15% | 470 | 22% |
| Continent | | | | |
| Americas | 24 734 | 58% | 1 558 | 72% |
| Asia | 6 173 | 15% | 191 | 9% |
| Europe | 11 460 | 27% | 406 | 19% |
| Sector | | | | |
| Commercial Services | 6 855 | 16% | 507 | 24% |
| Consumer Discretionary | 7 629 | 18% | 481 | 22% |
| Consumer Staples | 5 936 | 14% | 272 | 13% |
| Finance | 3 217 | 8% | 123 | 6% |
| Health | 5 504 | 13% | 144 | 7% |
| IT | 13 226 | 31% | 628 | 29% |

**Table 3:** Descriptive statistics of categorical variables

NextUnicorn_DescriptiveStats

### 3.3 Overcoming Class Imbalance

The website crunchbase.com employs a crowd-sourcing model, in which the information is gathered through large, open and rapidly growing internet users. The interviews conducted with crunchbase team within the scope of this research revealed that the operating firms provide and update information about their enterprises. Hence, the data set obtained from crunchbase.com is subject to selection (success) bias. Success bias refers to the sampling limitation that the sample set is not representative of the true population.

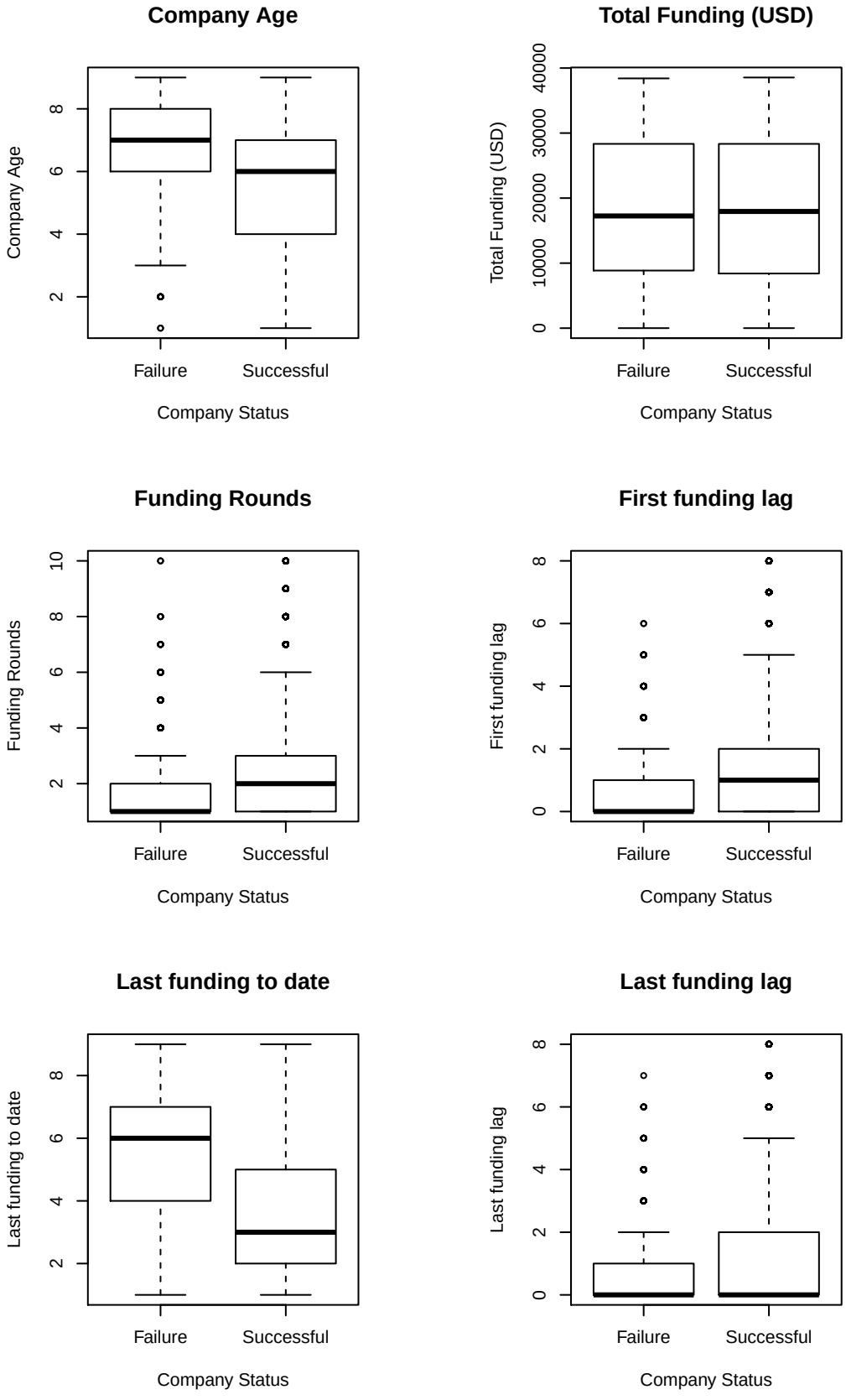



**Figure 1:** Descriptive statistics of continuous variables

The cleaned data set reveals that 95.18% of the companies are classified as successful and the remaining 4.82% are as failed/closed, indicating class imbalance. Class imbalance hinders the machine learning performance. For example, when the number of instances in one class is larger than the other, machine learning algorithms tend to label minority classes to the majority class. Although this would not have drastic effects on the accuracy, Type II error will be very high. If we determine the model performance not through accuracy but via number of false positives (FP) , the class imbalance will have a negative impact. The costs of misclassification between different classes often vary as well (Refer to Section 4).

In the scientific literature, designing smarter sampling strategies has been acknowledged as a valid approach to handling imbalanced data. However, when a new and improved sampling approach is not possible, such as in this paper, the adopted approach is to undersample the majority class or oversample the minority class (Krawczyk, 2016). In this paper, class imbalance is handled by oversampling the minority classes through synthetically creating artificial data points as described in Section 3.3.1.

### 3.3.1 Adaptive Synthetic Sampling Approach (ADASYN)

The goal of oversampling is to increase the size of minority class via synthetic observations based on the existing minority class observations to balance the size of majority and minority classes. ADASYN advances on Synthetic Minority Oversampling Technique (SMOTE) and adaptively generates minority data according to the distribution they have by adding a random value to the synthetically generated data points, in order to make them more scattered. Hence, ADASYN helps reducing the learning bias and adaptively shifts the decision boundary for the classification problem to focus more on the samples that are difficult to learn. Algorithm 1 summarizes the ADASYN process.

After completing the data pre-processing, the remaining 44 522 data points are split into training and test sets, 70% and 30% respectively. The ADASYN is adopted for training and test samples separately to prevent any dependence between two data sets.

## 3.4 Selected Models

**Logistic Regression:** The logistic regression is a specific case of linear regression where

**Algorithm 1:** Pseudocode of ADASYN based on He et al. (2008)

[H] **Input :** Training (or test) data set ($D_{k,p}$), $m_s$ and $m_l$, where $m_s \leq m_l$. $d_{th}$ is a preset threshold for the maximum tolerated degree of class imbalance ratio.

**Procedure:**

(1)Calculate the degree of class imbalance:

$$d = m_s/m_l \quad \text{where} \quad d \in (0, 1] \tag{1}$$

**if** $d < d_{th}$ **then**

(a) Calculate the number of data points, which need to be synthetically generated

$$G = (m_s - m_l) \times \beta \tag{2}$$

where $\beta \in [0, 1]$ is the parameter to satisfy $d_{th}$.

(b) **for** $x_i \in m_s$ **do**

Find KNN based on Euclidean distance in $p$-dimensional space and calculate

$$r_i = \Delta_i / K \tag{3}$$

where $\Delta_i$ is the number of examples in the K nearest neighbours of $x_i$ that belong to majority class, $r_i \in [0, 1]$

(i) Normalize $r_i$ according to

$$\hat{r_i} = r_i / \sum_{i=1}^{m_s} r_i \tag{4}$$

so that $\hat{r}_i$ is a density function

(ii) Calculate the number of data points which need to be synthetically generated for each minority example $x_i$

$$g_i = \hat{r_i} \times G \tag{5}$$

**for** *each* $x_i$ *from 1 to* $g_i$ **do**

(c) Randomly pick minority data example $x_{zi}$ from KNN of $x_i$

Generate the synthetic data example

$$s_i = x_i + (x_{zi} - x_i) \times \lambda \tag{6}$$

_where ($x_{zi}$ - $x_i$) is the difference vector and $\lambda$ is a random number: $\lambda \in [0, 1]$

**Output:** Synthetic data examples $s_i$ for i=1,...$G$

the response, $Y$, is a dichotomous variable. Logistic regression models the probability that $Y$ belongs to one of the two categories (Härdle and Simar, 2015). By design, the logistic regression function gives output between 0 and 1, which is the probability of belonging to one of dichotomous classes, $p(x_i)$.

**Recursive Partitioning Trees (Rpart):** Rpart schemes construct regression or classification models of a top level/general structure through a two-stage procedure, where the resulting models are represented as binary trees in the literature. First, the algorithm finds the best variable to best split the data into two groups. This step is then repeated for the resulting two subgroups until the subgroup size reaches a predetermined minimum size or there is no further improvement to the model to be made (Härdle, 1990).

**Conditional Reference Tree:** Algorithm for this model first tests for the hypothesis of independence between the response variable and covariates. If the hypothesis can be rejected, then the recursive steps of 1 and 2 of the general model are iterated until a stop criterion is met. The implementation uses a unified framework for conditional inference (Strasser and Weber, 1999). A split is established if the sum of the weights of two neighbouring nodes exceeds a predetermined minimum value.

**Random Forest:** Bagging is the essence of random forests. However, in the presence of one or few highly dominant predictors, each single tree would use the strongest predictor on the top level, hence trees would end up looking quite similar to each other. Random forest models are forced not to consider the whole set of available predictors. The restriction on the available predictors for each tree, therefore, prevents the model to be dominated by one (or few) very strong predictors (Breiman, 2001). Algorithm 2 summarizes the random forest formation.

**Algorithm 2:** Pseudocode of Random Forest based on Gepp et al. (2010)

**Input :** A bootstrap sample of $S$ , with F features in total and number of trees in forest is B

**function: Random Forest(S, F)**

H $\leftarrow \emptyset$

**for** $i \in 1, ...B$ **do**

$S^{(i)} \leftarrow$ *A bootstrap sample from* $S$

$h_i \leftarrow RandomTLearn(S^{(i)}, F)$

$H \leftarrow H \cup h_i$

**return** $H$

**end function**

**function: RandomTLearn(S,F)**

**At each node:**

$f \leftarrow$ *very small subset of* **F**

**Split on best feature in** $k$

**end function**

**Output: The learned tree**

**Extreme Gradient Boosting:** Gradient boosting combines weak learners in an additive manner and forms a new learner, which has maximal correlation with the negative gradient of the loss function (Friedman, 2002). In gradient boosting, the newly generated models predict the residuals (errors) of the previous models and use these predictions to form the output.

First, a subset from full training data is drawn at random and without replacement at each iteration. Then, the deviation of residuals in each iteration (partition) is derived and the best data partitioning is determined in each stage. Afterwards, the succeeding model fits the residuals from the preceding stage and builds a new model to reduce the variance of residuals. The aim here is to correct the mistakes of the first model.

Extreme gradient boosting (XGB) implements some improvements to gradient boosting (Chen and Guestrin, 2016). It penalizes trees for misclassifications, shrinks the leaf nodes and improves computing efficiency and has some other extra randomization parameters to ensure low variance. XGB reduces the space of possible feature splits based on the distribution of features across all data points in a leaf on a branch.

**Model Performance:** There are a couple of points one needs to consider while assessing model performance and concluding on the best model to implement. First of all, the performance of a learner mainly depends on the training data and the formulation of the initial hypothesis. If the training data does not provide sufficient information, it will be difficult to conclude on one single best learner. Hence, this will be another motivation for using ensemble models to benefit from multiple weak learners rather than having only one strong learner (Wang et al., 2014).

As Wang et al. (2014) and du Jardin (2016) state as well, the approach of ensemble models are reasonable. However, in practice the necessary conditions of accuracy and diversity need to be satisfied. Accuracy stands for the ability of the base learner to perform better than random guessing (generally 50%) and each base learner should have its own information about the problem, i.e. inclusion of variables/regressors.

Prior to estimating models and comparing them, it is not possible to say which modeling method will perform better in the framework of this paper, as in general, there is not a single modeling method that performs better in all research problems. In the next section, model estimations and results are being discussed.

# 4 Results and Discussion

## 4.1 Logistic Regression Implementation

Full simple logistic regression (M0) considers the remaining variables after eliminating the ones with near zero variance as explained in the earlier sections. M0 confirms the existence of the dummy trap and reveals the insignificant variables. As the second step, one level of the dummy variables and the statistically insignificant variables are excluded in the reduced logistic regression model (M1). Hence, only the coefficient estimates from M1, which are significant, are summarized in Table 4. Lasso and Ridge regularization methods are not implemented as the number of features remaining after data cleaning and feature selection steps. The training error of both M0 and M1 do not indicate over-fit as well. Determination of the best value of the regularization parameter, lambda, is outside of the scope of this paper. The most striking result is the near zero estimate of total funding (USD). This is a combined

effect of many factors, some of them are positively and some of them are negatively correlated with success rate. Positive sign is expected, since successful companies with future potential, after careful review/research from investors will get funding in favorable competitive terms. Hence, the higher the funding amount, the higher will their expectations be, that the startup has future potential. The negligible effect of total funding (USD) on success can be explained by the cash-burning of a startup. As discussed by Ooghe and De Prijcker (2008), the startups which received high investments in their rapid-growth phase often end up in bankruptcy due to poor management decisions, which includes misallocation of received funds. This result indicates high burn-rates in the failed companies.

One can argue that the number of funding rounds is a proxy of the persuasion skills of the entrepreneur towards investors when they start raising capital. Early-stage funding rounds indicate that the entrepreneur is successful in selling their idea to the investors. However, if these funding rounds are not followed by appropriate and effective managerial actions (refer to Section 2), the increasing number of funding rounds may have a negative impact on success.

| | Coefficient | Std. Error |
|---|---|---|
| (Intercept) | 3.08 | 0.01 |
| Funding rounds | -0.10 | 0.01 |
| Company age | 0.19 | 0.01 |
| Last funding to date | -0.75 | 0.01 |
| Total funding (USD) | -0.00 | 0.00 |
| Social both | 0.84 | 0.07 |
| Social Facebook | 0.85 | 0.07 |
| Social Twitter | 0.24 | 0.04 |
| Continent Americas | -0.65 | 0.03 |
| Sector Comm Serv. | -0.08 | 0.04 |
| Sector Cons. Disc. | -0.18 | 0.04 |
| Sector Cons. Stap. | -0.20 | 0.04 |
| Sector Health | 0.62 | 0.05 |

**Table 4:** Summary of reduced logistic regression (M1)

NextUnicorn_LogisticRegression

The regression coefficients of M1 change in a range of [-0.65, 0.85] for dummy variables. The

existence on both digital platforms or only on Facebook have the highest impact on business success. The negative coefficient for geographic location, continent Americas, can be explained by the intense competition and harsh business environment. As discussed before, the failure culture differs in the Americas. Hence, it is plausible to conclude that the negative coefficient confirms the *fail fast* mentality. The positive coefficient for the health sector supports the popularity of startups in the health sector in recent years. The negative coefficient of last funding to date indicates that a company is less likely to fail if their last funding was not long before 2019.

In the literature, one of the commonly used metrics, to explain the variance in the dependent variable that is explained by the independent variables for logistic regression, is McFadden's pseudo $R^2$. McFadden's $R^2$ is defined as in Equation (7), where $ln(L_M)$ is the fitted model and $ln(L_0)$ represents the null model with only the intercept as the predictor. McFadden's pseudo $R^2$ ranges between 0 and 1. If the value is closer to zero, the predictive power of the model decreases. The reduced model, M1 has a McFadden's pseudo $R^2$ of 0.26, indicating a quite weak predictive power (Hu et al., 2006).

$$McFadden's\, \tilde{R}^2 = 1 - \frac{ln(L_M)}{ln(L_0)} \tag{7}$$

Both trained models (M0 and M1) are then used to predict the failure probabilities of the startups. The status label of success is assigned if the predicted success probability is above a predetermined threshold of 50%, and failure otherwise. The confusion matrix of the test set predictions of M0 can be seen in Table 5. The prediction accuracy of M0 in the test set is 77.45%, despite the existence of the dummy trap and insignificant coefficient estimates. Although the insignificant regressors were eliminated, M1 also performed with a predictive accuracy of 77.41%, i.e. only 22.59% of the data in the test set are erroneously classified at the selected threshold level. The confusion matrix of the test set predictions of M1 can be found in Table 6. The predictive accuracy of both M0 and M1 performed better than random guessing (50%). The original empirical study on business success prediction Lussier (1995) has the predictive ability, i.e. accuracy of 70%. On the other hand, the recent extensions of Lussier's model are able reach accuracy levels of up to 85%. Despite the low McFadden's pseudo $R^2$, the reduced logistic regression model did not under-perform compared to the preceding studies.

| | Actual Failure | Actual Success |
|---|---|---|
| Predicted Failure | 4 603 (24.2%) | 2 545 (13.3%) |
| Predicted Success | 1 754 (9.2%) | 10 168 (53.3%) |

**Table 5:** Confusion matrix of the full logistic regression (M0)

NextUnicorn_LogisticRegression

| | Actual Failure | Actual Success |
|---|---|---|
| Predicted Failure | 4 354 (22.8%) | 2 521 (13.3%) |
| Predicted Success | 1 995 (10.5%) | 10 177 (53.4%) |

**Table 6:** Confusion matrix of the reduced logistic regression (M1)

NextUnicorn_LogisticRegression

The relationship between predicted probabilities of belonging to success class and continuous covariates used in M1 are illustrated in Figure 2. It can be clearly concluded that there are only few number of firms who achieve more than 6 funding rounds. Those who can achieve higher number of funding rounds attract continuous investor attention. Consistency in investor relations and financial support can be linked to higher probability of being successful. Also, although many companies receive early-stage funding, this does not necessarily promise success as explained above. The predicted probabilities are rather random for the lower end of the number of funding rounds. Also, as Section **??** and 2 elaborate, the first 1-5 years are decisive on the survival of a new venture. When the company age is considered, the predicted probability of success decreases starting from the $3^{rd}$ year. As the company age increases, it becomes difficult to make a distinct differentiation between the probability of success and failure. The lag of last funding to date reflects negative linear dependence with the predicted probabilities of success. This means ventures that received recent funding have higher odds of being successful.

On the other hand, there is no clear pattern between predicted probability of success and categorical variables. Only social both slightly exhibits positive relationship to the predicted success probability.

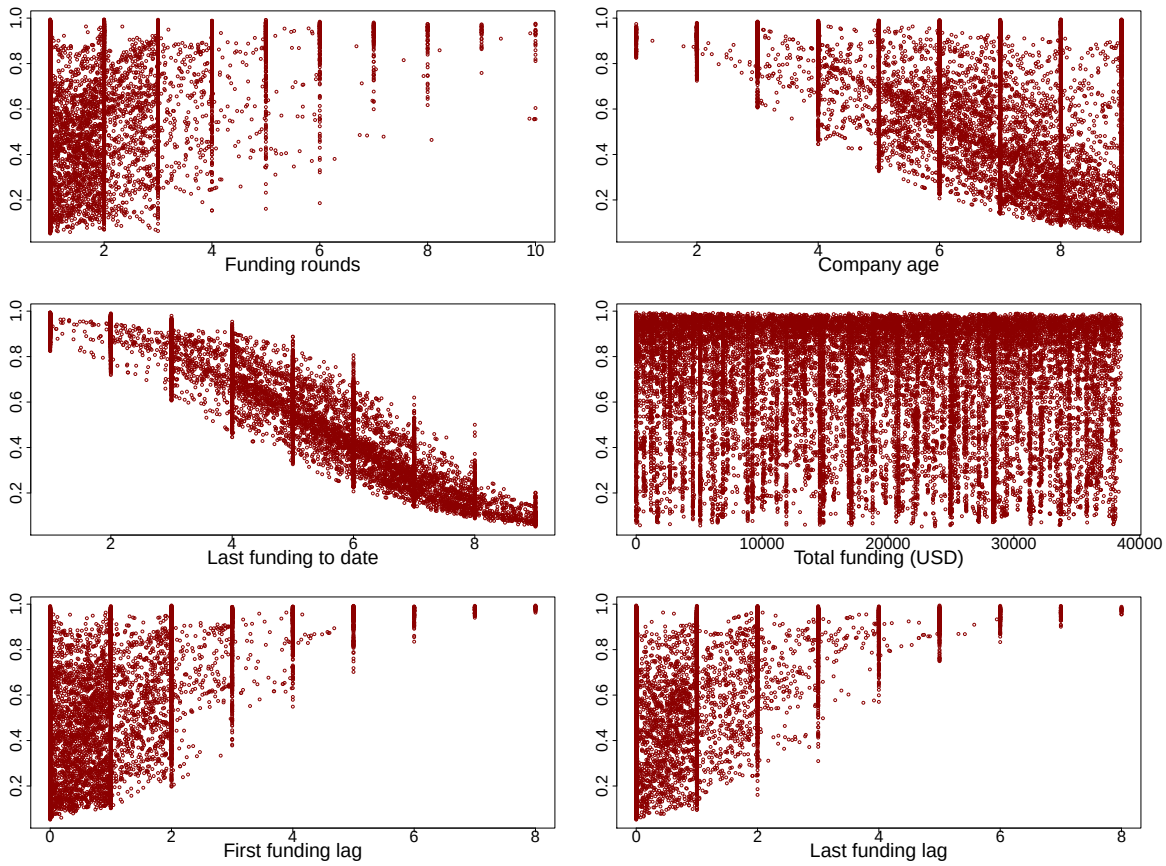



**Figure 2:** Scatter plot of success probability against regressors based on M1

NextUnicorn_Scatter

## 4.2 Recursive Partitioning & Conditional Inference Tree Implementation

There are some control parameters which affect the complexity and performance of decision trees. Two of the most important of these parameters are minimum split and minimum bucket size. Minimum split is the number of observations, which needs to exist in a node for a split to be attempted. Minimum bucket size is the minimum number of observations in a terminal node. As explained before, the startup profiles are quite unique and it is difficult to find a general fitting pattern for failed companies. Hence, the size of minimum split and minimum bucket are set to two in order to embrace the granular nature of the startup failure patterns. Furthermore, recursive partitioning tree functions in R uses a parameter called "complexity" to track and control the complexity of a tree. This measure is a combination of the ability of the tree to successfully separate the labels of the dependent variable, status, and the size of the tree. In order to determine the complexity measure, the record with

the minimum cross-validation error is identified and the complexity measure of this record is used. Hence, complexity measure is set to 0.001. The pruning attempt resulted in the same complexity measure. Therefore, pruning did not change or improve the initial construction of the recursive partitioning tree.

| | Actual Failure | Actual Success |
|---|---|---|
| Predicted Failure | 5 477 (28.7%) | 335 (1.8%) |
| Predicted Success | 880 (4.6%) | 12 378 (64.9%) |

**Table 7:** Confusion matrix of the recursive partitioning tree

NextUnicorn_RecursivePartitioning

The recursive partitioning tree performed surprisingly well with an error rate of 6.3%. Given the above explained drawbacks of stand-alone decision trees, the performance of the recursive partitioning tree can indicate overfitting. Also, fitting a single model is prone to instability after small changes in the training set (Hothorn et al., 2006).

The conditional inference tree, as explained previously, checks for the independence of the response variables and the covariates as opposed to recursive partitioning trees. The confusion matrix of the predictions from conditional inference tree is represented in Table 8. Conditional inference tree has performed with an error rate of 14.4%. With a test accuracy rate of 85.6%, conditional inference tree performs better than the literature benchmarks, mentioned earlier.

| | Actual Failure | Actual Success |
|---|---|---|
| Predicted Failure | 4 554 (23.9%) | 942 (4.9%) |
| Predicted Success | 1 803 (9.5%) | 11 771 (61.7%) |

**Table 8:** Confusion matrix of the conditional inference tree

NextUnicorn_ConditionalTree

### 4.3 Random Forest Implementation

The criticism towards stand-alone decision trees addressed the high dependence of results on the training data and alterations of the decision tree structure related to the small changes in the training data. In order to overcome these hurdles, a forest of decision trees was generated and the number of independent variables to be considered at each split is restricted.

As mentioned in the model description, the optimal number of variables for splitting at each node is the square root of the number of all available independent variables. Hence, this parameter is set to 5. The number of trees to grow is limited to 500 as the data set is quite large (Cutler et al., 2007; Strobl et al., 2007, 2008).

| | Actual Failure | Actual Success |
|---|---|---|
| Predicted Failure | 5 489 (28.7%) | 242 (1.3%) |
| Predicted Success | 868 (4.6%) | 12 471 (65.4%) |

**Table 9:** Confusion matrix of the random forest


Table 9 summarizes the confusion matrix for the predictions from the random forest model. As expected, random forest model performed well with an error rate of 5.9% and is more reliable than the recursive partitioning and conditional inference trees. Under Section 2 it has been explained that the random forest is an ensemble model and improves many shortcomings of the single decision trees. Predicting on the test set based on 500 decision trees has decreased the error rate by almost 1.5 percentage points compared to the partitioning and 3 percentage points compared to the conditional inference trees.

### 4.4 Extreme Gradient Boosting Implementation

Similar to the other models, the parameters affecting the model performance are adjusted in XGB as well. The *booster* parameter is set equal to gbtree as the model will be trained for a classification problem. As no regularization method in the logistic regression models or misclassification penalties in other decision tree based methods are implemented, *Gamma*, the loss reduction parameter to control the overfitting problem, is set to 0.

Before training the model, a 5-fold cross-validation (CV) model is implemented to identify the optimal number of iteration rounds. The maximum number of iterations for cross-validation is set to 200. The optimal number of iterations is determined by the minimum test error reached via cross-validation. For 200 rounds, a sub-sample of 5-fold is retained as the test set for validation and the remaining 4 sub-samples are used for training. If the test error of a round does not improve, i.e decrease, in 20 consecutive rounds the process is terminated and the optimal number of iteration rounds is identified. The model returned lowest test error at the $129^{th}$ iteration. The minimum test error is 0.048, indicating a CV accuracy of 95.2%. XGB performed with a predictive accuracy rate of 94.45%.

| | Actual Failure | Actual Success |
|---|---|---|
| Predicted Failure | 5 612 (29.4%) | 314 (1.7%) |
| Predicted Success | 745 (3.9%) | 12 399 (65.0%) |

**Table 10:** Confusion matrix of the extreme gradient boosting

NextUnicorn_XGBoost

### 4.5 Comparison of Models

There are plenty of options when it comes to evaluating the model performance and conclude on a metric to compare the six models, which were implemented. Below is the description of these metrics.

| | |
|---|---|
| Accuracy | : (TP + TN) / (TP + TN + FP + FN) |
| Error Rate | : 1 - Accuracy |
| Sensitivity / TPR | : (TP)/(TP+FN) |
| Specificity | : (TN)/(TN+FP) |
| FPR | : (FP)/(FP + TN) |

A Receiver Operating Curve (ROC) illustrates the performance of a classification model by plotting True Positive Rate (TPR) vs. False Positive Rate (FPR) at all classification thresholds. Area under the ROC curve (AUC) takes the integral of the ROC curve between 0 and 1 and provides an aggregate measure of performance at different threshold levels (Ling et al., 2003). Table 11 provides an overview of the various comparison metrics. Since each measure

has its benefits and drawbacks, a combined evaluation approach is adopted.

Accuracy and error rates of the models have been mentioned under the respective sections. These metrics rank the ensemble method XGB as the best performing method. This means XGB is able to label both classes, success and failure, better in comparison to the other methods. Random forest is a close second after XGB, with an accuracy rate of 94.18%. This indicates that the general classification performance of the ensemble methods dominate that of models with a more traditional approach under the accuracy metric.

| Model Name | Accuracy | Sensitivity | Specificity | Type I Error | Type II Error |
|---|---|---|---|---|---|
| Full logistic regression | 77.45% | 79.98% | 72.40% | 9.2% | 13.3% |
| Reduced logistic regression | 77.41% | 79.99% | 72.25% | 10.5% | 13.3% |
| Rpart tree | 93.63% | 97.36% | 86.16% | 4.6% | 1.8% |
| Conditional inference tree | 85.61% | 92.59% | 71.64% | 9.5% | 4.9% |
| Random forest | 94.18% | 98.10% | 86.35% | 4.6% | 1.3% |
| Extreme gradient boosting | 94.45% | 97.53% | 88.28% | 3.9% | 1.7% |

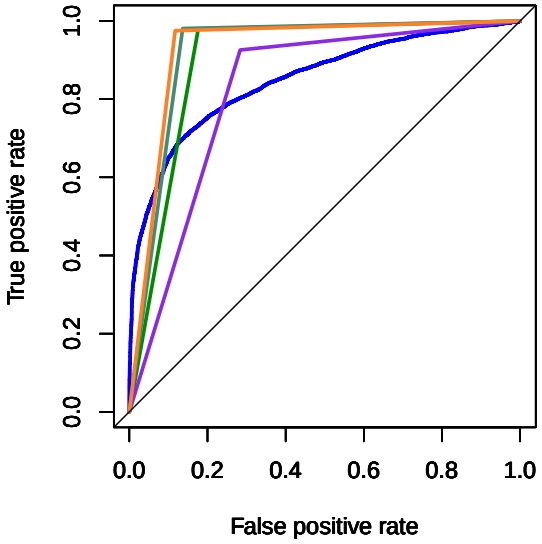
**Table 11:** Comparison of performance metrics across models

NextUnicorn_Results

One can argue that the costs of misclassifying a failed company as successful (false positive), i.e. Type II error, is costlier than misclassifying a successful company as failed (false negative), i.e. Type I error. Wang et al. (2014) and Gepp et al. (2010) state that Type II error, within startup success prediction frameworks, is more critical because Type II error bears financial losses due to investing in a startup doomed to fail, whereas Type I error creates lost opportunity costs from not investing in/dealing with a successful new business, which is also referred to as missed potential investment gains. Hence, the misclassification costs are not equal in the real world setting. With this background, one can argue that Type II error (i.e. specificity) can be used as a proxy of the real life costs of the classification errors. The model with the lowest Type II error, i.e. highest specificity, can be argued to be the best performing model. XGB labelled 88.28% of the failed companies correctly, with Type II error rate of 1.7%. Similar to the elaboration with accuracy and error rate, the second ensemble model, random forest reaches a specificity of 86.35% with the lowest Type II error rate of 1.3%. The worst performing models under this evaluation criteria are the logistic regression models.

Lastly, AUC is a scale-invariant metric. This means, AUC measures the ranking of the predictions rather than their absolute values. AUC does not require a threshold to measure the performance of the model. Although the threshold invariance of AUC metric can be advantageous in some cases, the disparity in the cost of misclassification of different classes can raise the need for a threshold to suppress this problem. Figure 3 provides a comparison of the AUC metric among models. Under AUC criteria, XGB dominates the rest of the models. Similar to the evaluation under previous metrics, the second ensemble method, random forest is the second best performing model under AUC as well.

All in all, different performance metrics point to different best-performing-models. However, it is plausible to conclude that the ensemble methods, random forest and XGB dominate the other models over all the performance metrics considered.




| Model Name | AUC |
|---|---|
| Full logistic regression | 84.89% |
| Reduced logistic regression | 84.90% |
| Rpart tree | 91.76% |
| Conditional inference tree | 82.11% |
| Random forest | 92.22% |
| Extreme gradient boosting | 92.91% |

**Figure 3:** Comparison of AUC among models

NextUnicorn_AUC

### 4.6 Discussion of Variable Importance

When the statistical models are difficult to interpret, i.e. referred in general as black-box models, variable importance constructed by the model can be useful to articulate and achieve a better understanding. The ranking of the variables according to their importance in the model construction is implemented for the top 3 best performing models: extreme gradient boosting, random forest and recursive partitioning tree.

The importance of variables in the recursive partitioning tree is calculated by adding up the improvement measures that each variable contributes as a primary or surrogate splitter. The relative importance is calculated by the sum of the goodness-of-split measure for each split plus the goodness-of-fit for all splits in which the node was a surrogate. Table 12 summarizes the normalized relative importance of the variables. The main contribution to the splits has been done by last funding to date followed by company age and social both.

| Variable | Importance Level |
|---|---|
| Last funding to date | 1.00 |
| Company age | 0.36 |
| First funding lag | 0.18 |
| Funding Rounds | 0.13 |
| Social Twitter | 0.04 |
| Social Both | 0.03 |
| Continent Americas | 0.00 |
| Social Facebook | 0.00 |
| Continent Asia | 0.00 |
| Sector Health | 0.00 |
| Sector Consumer Discretionary | 0.00 |
| Sector Consumer Staples | 0.00 |

**Table 12:** Ranked variable importance (normalized) in recursive partitioning tree

NextUnicorn_RecursivePartitioning

In the case of random forest, for each tree the prediction accuracy is measured for the out-of-bag (OOB) variables. Then, the values of the OOB variables are shuffled, while keeping all else the same. The mean decrease in accuracy represents how much the accuracy of the model decreases after shuffling the OOB variables, i.e. the respective variable is omitted. On

the other hand, mean decrease in the Gini index represents the impurity when a variable is chosen to split a node. It is calculated by the node impurity weighted by the probability of reaching that node. The higher the Gini index, the more important the feature. According to Gini index, last funding to date, company age and first funding lag are the top performing variables.

In Table 13, the importance measures are broken down by outcome class, Success (S) and Failure (F). For example, total funding (USD) is much more important for predicting failure class than predicting success. On the other hand, last funding to date is more important while predicting success than predicting failure.

| | F | S | Mean Decrease Accuracy | Mean Decrease Gini |
|---|---|---|---|---|
| Last funding to date | 42.1 | 76.6 | 89.5 | 5 840.5 |
| First funding lag | 361.9 | 70.4 | 80.9 | 2 815.0 |
| Company age | 40.5 | 62.8 | 73.3 | 2 777.2 |
| Funding Rounds | 54.6 | 38.3 | 47.7 | 1 580.0 |
| Last funding lag | 26.6 | 45.5 | 51.9 | 1 571.7 |
| Total funding (USD) | 132.4 | -9.9 | 116.9 | 1 268.8 |
| Social Both | 38.8 | 9.9 | 41.1 | 225.8 |
| Continent Americas | 49.7 | -1.32 | 46.9 | 161.4 |
| Social Twitter | 29.5 | -3.9 | 30.2 | 156.6 |
| Sector Commercial Services | 64.9 | 0.04 | 58.9 | 124.4 |
| Social None | 37.5 | 5.6 | 37.1 | 122.9 |
| Sector Consumer Discretionary | 69.1 | -8.5 | 61.8 | 113.2 |
| Sector IT | 63.1 | -3.4 | 57.1 | 109.0 |
| Sector Health | 43.4 | 13.5 | 46.9 | 97.3 |
| Sector Consumer Staples | 65.4 | -4.7 | 52.0 | 97.1 |
| Continent Europe | 33.9 | -1.1 | 34.9 | 86.1 |
| Continent Asia | 36.5 | -2.2 | 34.8 | 66.4 |
| Sector Finance | 49.2 | -9.6 | 37.7 | 56.8 |
| Social Facebook | 31.7 | 8.5 | 29.6 | 52.6 |

**Table 13:** Ranked variable importance in random forest

NextUnicorn_RandomForest

The variable importance in XGB is measured through the Gain, Cover and Frequency met-

rics. Gain represents the relative contribution of the respective variables, calculated through the contribution of each feature to each tree in the model. A higher value indicates higher importance. Cover represents the relative number of observations related to each variable. Frequency is the percentage representing the relative number of times a particular independent variable occurs in the trees of the model. The literature suggests the most relevant variable importance metric to be Gain (Chen and Guestrin, 2016).

| | Gain | Cover | Frequency |
|---|---|---|---|
| Last funding to date | 0.64 | 0.22 | 0.12 |
| First funding lag | 0.11 | 0.06 | 0.06 |
| Company age | 0.08 | 0.10 | 0.10 |
| Total funding (USD) | 0.06 | 0.42 | 0.39 |
| Funding rounds | 0.04 | 0.06 | 0.06 |
| Last funding lag | 0.02 | 0.05 | 0.04 |
| Social Both | 0.01 | 0.01 | 0.03 |
| Continent Americas | 0.01 | 0.01 | 0.02 |
| Social None | 0.00 | 0.01 | 0.02 |
| Sector Health | 0.00 | 0.02 | 0.01 |
| Sector Consumer Staples | 0.00 | 0.00 | 0.02 |
| Social Facebook | 0.00 | 0.01 | 0.01 |
| Sector IT | 0.00 | 0.00 | 0.02 |
| Sector Commercial Services | 0.00 | 0.01 | 0.02 |
| Sector Consumer Discretionary | 0.00 | 0.00 | 0.02 |
| Continent Europe | 0.00 | 0.00 | 0.01 |
| Social Twitter | 0.00 | 0.00 | 0.01 |
| Sector Finance | 0.00 | 0.01 | 0.01 |
| Continent Asia | 0.00 | 0.01 | 0.01 |

**Table 14:** Ranked variable importance in extreme gradient boosting

NextUnicorn_XGBoost

$$Gain(Y, X) = Entropy(T, X) - Entropy(X) \tag{8}$$

Importance Gain is calculated by the decrease in entropy. Using Gain measure the top performing variables are last funding to date, first funding lag and company age.

The top 3 performing models have a consensus on the most important variables, which are last funding to date, first funding lag and company age. The general ranking of the variable importance revealed that the top 3 performing models prioritized continuous variables more than categorical ones. All in all, the variable importance did not differ significantly between different models implemented as the universal function approximators choose the same variables.

# 5 Conclusion

This paper thoroughly addresses how to predict success for startups. The amount of literature work on startup success revealed the need for research in this area. Existing literature focuses on established firm success rate prediction. However, there are differences between corporate and startup success prediction, making the models in existing literature inapplicable to predicting success for startup firms.

Predicting startup success is a challenging task and the associated monetary and opportunity costs are high for making a wrong decision on which startup will be successful. Due to energy and time intensive nature of processing vast amount of information, the players of the startup ecosystem can highly benefit from a quantitative method, when it comes to making decisions in such high risk environment. Hence, this paper empirically illustrates the implementation of various machine learning algorithms to predict startup success.

The data used in the estimation is based on the information from a crowd-sourced database crunchbase.com, without allocating budget or time to interview/collect survey answers from startups. One advantage of using this data set in the paper is the sample size, which is larger compared to other research and papers in the literature. Since the majority of the firms, who provided/updated their crunchbase profiles are mostly successful firms, the used data entails a selection (success) bias. This leads to the class imbalance problem between successful (95%) and failed companies (5%). This problem is tackled by oversampling the minority class data (failed companies) by implementing ADASYN.

In total, six separate models are implemented: (i) full logistic regression; (ii) reduced logistic regression; (iii) recursive partitioning tree; (iv) conditional inference tree; (v) random forest and (vi) extreme gradient boosting. The most common method in literature, logistic

regression is implemented for comparability reasons and to construct a benchmark for the succeeding models. Logistic regression, both full and reduced models have performed better than random guessing. With McFadden's pseudo $R^2$ of 0.26 and an error rate around 22.5%, both logistic regressions performed within the predictive accuracy interval set by preceding logistic regression models in the literature. However, compared to other four implemented models, neither of the logistic regression models exhibited satisfactory predictive ability.

In order to fully use the information contained by features, two different types of decision trees have been built. Recursive partitioning trees reached AUC of 91.76% and outperformed the conditional inference trees (AUC of 82.11%) .

In order to tackle the overfitting problem of the above-mentioned decision trees, models have been extended to random forests. Random forest showed above average performance over the range of different metrics among other models implemented and provided the lowest Type II error rate (1.3%), indicating that the predictions from random forest model result in the lowest costs for misclassification of the failed companies.

Although random forest is an ensemble method itself, the research has been extended to extreme gradient boosting for its proven efficiency and performance in the recent competitions and research. Compliant with the applications in literature, XGB performed the best among other models implemented under a majority of the metrics. With an accuracy of 94.45%, a specificity of 88.28% and AUC of 92.91%, XGB slightly dominates the random forest approach. The top 3 performing models, XGB, random forest and recursive partitioning tree, ranked the same three variables as their main features, which are last funding to date, first funding lag and company age.

Predicting startup success is a challenging task and the associated monetary and opportunity costs are high. This study provides, repeatable and quantified modeling process, to predict startup success, using machine learning methods and large scale publicly available data.

# 6 Further Research

The first improvement point to address is the availability/collection of data from startups. Individual interviews and surveys with startups are not only time and resource intensive but also not reproducible. It also leads to response bias. This paper has shown that reproducible models that train on off-the shelf data with none/minimum information about the personality of the entrepreneur or the characteristics of the management team, can still reach near 95% accuracy level. However, the data used in this study lacks background information about the entrepreneur and the management team. Including these widely acknowledged variables can further improve the model performance. A future research on a common framework to conceptualize the collection of information rich data would be essential to build solid prediction models.

Also, the data used in this paper provides a snapshot at a single point in time, i.e. the time aspect of failure is being neglected. The need for panel data to better understand the triggers of failure is indisputable. Percentage/growth metrics such as the change in the number of employees or growth rate of the funding amount received and many other similar metrics generated on a longitudinal manner would help improve the prediction results.

Another improvement point is the definition of success for startups. There are some examples of startups, which filed for an IPO within the first year of their establishment. This is very uncommon in business and is not necessarily a proxy for success. Similarly, every acquisition has its own characteristics. An acquisition can represent success if the entrepreneurs benefit from the transaction or can also point to failure if the startup cannot reach financial stability. Failure on the other hand can also be more specifically defined. One can argue that a startup can be considered as failed, only after it existed long enough to officially file bankruptcy to the authorities. Such improvements to the label determination has the potential to reduce the class imbalance.

The asymmetry in terms of cost for correctly predicting startup success or failure correctly is mentioned in Section 4. Model selection for the startup success prediction also provides a research area. The minuscule difference between two ensemble methods in this paper, random forest and extreme gradient boosting, can be further investigated through the implementation of a cost function/matrix. This approach on the other hand would require intensive research

into the financial and opportunity costs of misclassification and is not trivial.

Another improvement point is to set the focus of the research on a specific industry and sub-category of these industries. The benchmark of success for firms operating in disruptive fields, such as digital and tech firms specializing on cryptocurrencies, are indeed different to the ventures, which operate in utilities or heavy machinery. Although implementing such industry specifications would have an impact on the variables defined in the data set and might result in smaller sample sizes, tailoring quantified models to the needs of different sectors can help to determine the drivers of success and predict business success with higher accuracy.

Startup success prediction is indeed in the interest of all parties involved in the startup ecosystem. In the light of the above-mentioned improvements, it might be possible that the quantitative models, such as the ones introduced in this paper, will have the predictive ability to spot the next unicorn.

## References


Boritz, J. E. and D. B. Kennedy (1995): "Effectiveness of neural network types for prediction of business failure," *Expert Systems with Applications*, 9, 503–512.

Breiman, L. (2001): "Random forests," *Machine learning*, 45, 5–32.

Butler, T. and B. Fitzgerald (1999): "Unpacking the systems development process: an empirical application of the CSF concept in a research context," *The Journal of Strategic Information Systems*, 8, 351–371.

Cao, R., M. A. Delgado, W. González-Manteiga, et al. (1997): "Nonparametric curve estimation: an overview," *Investigaciones Economicas*, 21, 209–252.

Ceausu, I., K. Marquardt, S.-J. Irmer, and E. Gotesman (2017): "Factors influencing performance within startup assistance organizations," in *Proceedings of the International Conference on Business Excellence*, De Gruyter Open, vol. 11, 264–275.

Chang, S. J. (2004): "Venture capital financing, strategic alliances, and the initial public offerings of Internet startups," *Journal of Business Venturing*, 19, 721–741.

Chen, T. and C. Guestrin (2016): "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, ACM, 785–794.

Cutler, D. R., T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler (2007): "Random forests for classification in ecology," *Ecology*, 88, 2783–2792.

Dempwolf, C. S., J. Auer, and M. DIppolito (2014): "Innovation accelerators: Defining characteristics among startup assistance organizations," *Small Business Administration*, 1–44.

Dimitras, A. I., S. H. Zanakis, and C. Zopounidis (1996): "A survey of business failures with an emphasis on prediction methods and industrial applications," *European Journal of Operational Research*, 90, 487–513.

Doumpos, M. and C. Zopounidis (2002): "Business failure prediction: a comparison of classification methods," *Operational Research*, 2, 303.

du Jardin, P. (2016): "A two-stage classification technique for bankruptcy prediction," *European Journal of Operational Research*, 254, 236–252.

Einhorn, H. J. (1974): "Expert judgment: Some necessary conditions and an example." *Journal of Applied Psychology*, 59, 562.

Fischhoff, B., D. Kahneman, P. Slovic, and A. Tversky (2002): "For those condemned to study the past: Heuristics and biases in hindsight," *Foundations of Cognitive Psychology: Core Readings*, 621–636.

Friedman, J. H. (2002): "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, 38, 367–378.

Gatev, P., S. Thomas, J.-S. Lou, M. Lim, and M. Hallett (1996): "Effects of diminished and conflicting sensory information on balance in patients with cerebellar deficits," *Movement Disorders: Official Journal of the Movement Disorder Society*, 11, 654–664.

Gepp, A., K. Kumar, and S. Bhattacharya (2010): "Business failure prediction using decision trees," *Journal of Forecasting*, 29, 536–555.

Haavelmo, T. (1944): "The probability approach in econometrics," *Econometrica: Journal of the Econometric Society*, iii–115.

Härdle, W. (1990): *Applied nonparametric regression*, vol. no.19, Cambridge University Press.

Härdle, W. and L. Simar (2015): *Applied Multivariate Statistical Analysis*, vol. 4th edition, Springer.

He, H., Y. Bai, E. A. Garcia, and S. Li (2008): "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, IEEE, 1322–1328.

Hothorn, T., K. Hornik, and A. Zeileis (2006): "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical statistics*, 15, 651–674.

Hu, B., J. Shao, and M. Palta (2006): "Pseudo-R^ 2 in logistic regression model," *Statistica Sinica*, 16, 847.

Jones, P. M. and E. Olson (2013): "The time-varying correlation between uncertainty, output, and inflation: Evidence from a DCC-GARCH model," *Economics Letters*, 118, 33–37.

Kauffman, R. J. and B. Wang (2001): "The success and failure of dotcoms: A multi-method survival analysis," in *proceedings of the 6th INFORMS Conference on Information Systems and Technology (CIST), Miami, FL, USA*, Citeseer, 3–4.

Krawczyk, B. (2016): "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, 5, 221–232.

Kuhn, M. and K. Johnson (2013): *Applied predictive modeling*, vol. 26, Springer.

Laitinen, E. K. (1992): "Prediction of failure of a newly founded firm," *Journal of Business Venturing*, 7, 323–340.

Ling, C. X., J. Huang, H. Zhang, et al. (2003): "AUC: a statistically consistent and more discriminating measure than accuracy," in *Ijcai*, vol. 3, 519–524.

Liu, J. and C. Wu (2019): "Hybridizing kernel-based fuzzy c-means with hierarchical selective neural network ensemble model for business failure prediction," *Journal of Forecasting*, 38, 92–105.

Luce, R. D. (1959): "On the possible psychophysical laws." *Psychological Review*, 66, 81.

Luger, M. I. and J. Koo (2005): "Defining and tracking business start-ups," *Small Business Economics*, 24, 17–28.

Lussier, R. N. (1995): "A nonfinancial business success versus failure prediction mo," *Journal of Small Business Management*, 33, 8.

Lussier, R. N. and S. Pfeifer (2001): "A crossnational prediction model for business success," *Journal of Small Business Management*, 39, 228–239.

Ooghe, H. and S. De Prijcker (2008): "Failure processes and causes of company bankruptcy: a typology," *Management Decision*, 46, 223–242.

Ooghe, H. and N. Waeyaert (2004): "Causes of company failure and failure paths: The rise and fall of Fardis," *European Case Study*, 1–8.

Ozdemir, S. Z., P. Moran, X. Zhong, and M. J. Bliemel (2016): "Reaching and acquiring valuable resources: The entrepreneur's use of brokerage, cohesion, and embeddedness," *Entrepreneurship Theory and Practice*, 40, 49–79.

Read, D. and B. Van Leeuwen (1998): "Predicting hunger: The effects of appetite and delay on choice," *Organizational Behavior and Human Decision Processes*, 76, 189–205.

Regmi, K., S. A. Ahmed, and M. Quinn (2015): "Data driven analysis of startup accelerators," *Universal Journal of Industrial and Business Management*, 3, 54–57.

Shane, S. (2012): "The importance of angel investing in financing the growth of entrepreneurial ventures," *The Quarterly Journal of Finance*, 2, 1250009.

Simon, H. A. (1955): "A behavioral model of rational choice," *The Quarterly Journal of Economics*, 69, 99–118.

Strasser, H. and C. Weber (1999): "On the asymptotic theory of permutation statistics," .

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008): "Conditional variable importance for random forests," *BMC Bioinformatics*, 9, 307.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007): "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, 8, 25.

Tavoletti, E. (2013): "Business incubators: effective infrastructures or waste of public money? Looking for a theoretical framework, guidelines and criteria," *Journal of the Knowledge Economy*, 4, 423–443.

Venkatraman, V., J. W. Payne, J. R. Bettman, M. F. Luce, and S. A. Huettel (2009): "Separate neural mechanisms underlie choices and strategic preferences in risky decision making," *Neuron*, 62, 593–602.

Wang, G., J. Ma, and S. Yang (2014): "An improved boosting based on feature selection for corporate bankruptcy prediction," *Expert Systems with Applications*, 41, 2353–2361.

Wu, W.-W. (2010): "Beyond business failure prediction," *Expert Systems with Applications*, 37, 2371–2376.

# IRTG 1792 Discussion Paper Series 2019

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

001 "Cooling Measures and Housing Wealth: Evidence from Singapore" by Wolfgang Karl Härdle, Rainer Schulz, Taojun Xie, January 2019.

002 "Information Arrival, News Sentiment, Volatilities and Jumps of Intraday Returns" by Ya Qian, Jun Tu, Wolfgang Karl Härdle, January 2019.

003 "Estimating low sampling frequency risk measure by high-frequency data" by Niels Wesselhöfft, Wolfgang K. Härdle, January 2019.

004 "Constrained Kelly portfolios under alpha-stable laws" by Niels Wesselhöfft, Wolfgang K. Härdle, January 2019.

005 "Usage Continuance in Software-as-a-Service" by Elias Baumann, Jana Kern, Stefan Lessmann, February 2019.

006 "Adaptive Nonparametric Community Detection" by Larisa Adamyan, Kirill Efimov, Vladimir Spokoiny, February 2019.

007 "Localizing Multivariate CAViaR" by Yegor Klochkov, Wolfgang K. Härdle, Xiu Xu, March 2019.

008 "Forex Exchange Rate Forecasting Using Deep Recurrent Neural Networks" by Alexander J. Dautel, Wolfgang K. Härdle, Stefan Lessmann, Hsin-Vonn Seow, March 2019.

009 "Dynamic Network Perspective of Cryptocurrencies" by Li Guo, Yubo Tao, Wolfgang K. Härdle, April 2019.

010 "Understanding the Role of Housing in Inequality and Social Mobility" by Yang Tang, Xinwen Ni, April 2019.

011 "The role of medical expenses in the saving decision of elderly: a life cycle model" by Xinwen Ni, April 2019.

012 "Voting for Health Insurance Policy: the U.S. versus Europe" by Xinwen Ni, April 2019.

013 "Inference of Break-Points in High-Dimensional Time Series" by Likai Chen, Weining Wang, Wei Biao Wu, May 2019.

014 "Forecasting in Blockchain-based Local Energy Markets" by Michael Kostmann, Wolfgang K. Härdle, June 2019.

015 "Media-expressed tone, Option Characteristics, and Stock Return Predictability" by Cathy Yi-Hsuan Chen, Matthias R. Fengler, Wolfgang K. Härdle, Yanchu Liu, June 2019.

016 "What makes cryptocurrencies special? Investor sentiment and return predictability during the bubble" by Cathy Yi-Hsuan Chen, Roméo Després, Li Guo, Thomas Renault, June 2019.

017 "Portmanteau Test and Simultaneous Inference for Serial Covariances" by Han Xiao, Wei Biao Wu, July 2019.

018 "Phenotypic convergence of cryptocurrencies" by Daniel Traian Pele, Niels Wesselhöfft, Wolfgang K. Härdle, Michalis Kolossiatis, Yannis Yatracos, July 2019.

019 "Modelling Systemic Risk Using Neural Network Quantile Regression" by Georg Keilbar, Weining Wang, July 2019.

**IRTG 1792, Spandauer Strasse 1, D-10178 Berlin**
**http://irtg1792.hu-berlin.de**

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

# IRTG 1792 Discussion Paper Series 2019

For a complete list of Discussion Papers published, please visit http://irtg1792.hu-berlin.de.

020 "Rise of the Machines? Intraday High-Frequency Trading Patterns of Cryptocurrencies" by Alla A. Petukhina, Raphael C. G. Reule, Wolfgang Karl Härdle, July 2019.

021 "FRM Financial Risk Meter" by Andrija Mihoci, Michael Althof, Cathy Yi-Hsuan Chen, Wolfgang Karl Härdle, July 2019.

022 "A Machine Learning Approach Towards Startup Success Prediction" by Cemre Ünal, Ioana Ceasu, September 2019.

**IRTG 1792, Spandauer Strasse 1, D-10178 Berlin**
**http://irtg1792.hu-berlin.de**

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.