IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Assessment of machine learning performance for decision support in venture capital investments

JAVIER ARROYO[1], FRANCESCO COREA[2], GUILLERMO JIMENEZ-DIAZ[1], and JUAN A. RECIO-GARCIA[1]

[1]Department of Software Engineering and Artificial Intelligence. University Complutense of Madrid, Spain
[2]Ca' Foscari University, Venice, Italy. Four Trees Merchant Partners Inc. Madrid, Spain

Corresponding author: Juan A. Recio-Garcia (e-mail: jareciog@fdi.ucm.es).

This work was supported by Four Trees Merchant Partners Inc. and in part by the the Spanish Committee on Economy and Competitiveness (TIN2017-87330-R) by the European Union's H2020 coordination and support actions under grant agreement No 825215.

ABSTRACT The venture capital (VC) industry offers opportunities for investment in early-stage companies where uncertainty is very high. Unfortunately, the tools investors currently have available are not robust enough to reduce risk and help them managing uncertainty better. Machine learning data-driven approaches can bridge this gap, as they already do in the hedge fund industry. These approaches are now possible because data from thousands of companies over the world is available through platforms such as Crunchbase.

Previous academic efforts have focused only on predicting two classes of exits, i.e., being acquired by other company or offering shares to the public, using only one or a few subsets of explanatory variables. These events are typically related to high returns, but also higher risk, making hard for a venture fund to get repeatable and sustainable returns. On the contrary, we will try to predict more possible outcomes including a subsequent funding round or the closure of the company using a large set of signals. In this way, our approach would provide VC investors with more information to set up a portfolio with lower risk that may eventually achieve higher returns than those based on finding unicorns (i.e., companies with a valuation higher than one billion dollars).

We will analyze the performance of several machine learning methods in a dataset of over 120,000 early-stage companies in a realistic setting that tries to predict their progress in a 3-year time window. Results show that machine learning can support venture investors in their decision-making processes to find opportunities and better assessing the risk of potential investments.

INDEX TERMS CrunchBase, Decision Support Systems, Investment, Machine Learning, Risk Assessment, Venture Capital.

I. INTRODUCTION

After the last financial crisis, one of the most immediate reactions for financial institutions and regulators has been to artificially lower the interest rates. We live indeed today in a historical time where interest rates are recording the lowest levels of the last several decades. As a consequence, traditional public markets do not longer represent the solution to achieve sustainable returns for investors.

In this difficult environment, venture capital (VC) as an asset class has emerged as one of the potential poles of attraction for investors that are both looking for financial returns and innovation sprints. Many of the biggest empires have indeed been created (and funded) in the last 10-15 years and the sector itself is evolving at an incredibly high speed because of the huge interest raised by private and institutional investors.

Fast forward ten years, this investment frenzy has bought a couple of different considerations: first of all, it is getting harder and harder nowadays to get sustainable returns also in VC. The industry is therefore polarizing: many bigger funds have been quickly raised in the last two years to invest bigger tickets in faster-scaling companies and to double on winners. On the other side, a good opportunity still exists for investments in early-stage projects but the related uncertainty

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938659, IEEE Access

makes many of those deals hard to be finalized. Hence, if from one hand we have crazy valuations completely detached from any fundamental, on the other hand, we have the number of seed deals that is shrinking down.

This creates a gap and therefore an opportunity for smart investors that are willing to bet on early-stage companies with not so much validation as their more mature peers. Unfortunately, the toolbox investors currently have available is not robust enough to reduce their risk and help them managing uncertainty in a better way. The main rationale of this work is, therefore, to provide the investment community with tools that can make early-stage deals more attractive. Artificial intelligence and machine learning could be that new tool. Being a data-driven investor is a very well-known concept in the hedge fund industry, but in VC is yet not so popular and a lot of work can be done in this space. Machine learning can indeed support VCs investor by helping them spotting business opportunities, performing better portfolio management, matching co-investors and deals, obtaining intelligence on the competitorsâĂŹ landscape, identifying potential acquirers, and much more. In other words, it has the potential to make venture investors better and more informed, even in the post-investment phase where they need to help companies to grow.

However, there is another specific case we are interested in. A venture capitalist could be either a great financial investor or a great operator. In either case (and often the edges blur) VCs have to possess two other skills in addition to post-investment support abilities: i) they have to be able to generate interesting deal flow and understanding where good companies are; ii) they have to be able to identify patterns or signals of potential success in a company and pay the right price for it.

Those skills are hard to acquire and can only be developed spending years in the industry. We believe though that machine learning can speed up the learning curve for an investor. This paper is then an attempt to establish a data-driven approach that might be useful for early-stage investors to predict the future success of a company and understand the associated risks. We will show that it is possible to draw some insights from mostly qualitative data and that those insights have a positive correlation with the probability of a company to progress. In order to prove it, we test different machine learning methods that could better inform an investor on what company deserves funding. Our dataset will consist of over 120,000 early-stage companies retrieved from Crunchbase,[1] a platform that gathers business information about companies, e.g. funding sources, founders, business sector, etc.

While many studies and VC focus on predicting whether a company will be eventually acquired or go through an initial public offering (IPO), we will try to predict what will happen to the company next, including not only acquisition or IPO, but also obtaining more funds, or closure. In this way, we aim to offer VC investors the opportunity to consider not only high-risk/high-reward companies (such as potential unicorns) but also to be able to set up a portfolio with lower risk.

The rest of the work is structured as follows: Section II discusses how this work relates to previous similar studies. Section III introduces our main models, techniques and dataset composition and collection. Section IV shows the empirical results of our study, while Section V analyzes the technical challenges and the business implications of our predictions. Section VI finally summarizes our main results and discusses future research directions.

II. LITERATURE REVIEW

We can find in literature several approaches for the prediction of the success of early-stage or startup companies. A popular one is the success/failure model presented based on logistic regression [1]. It considers 15 dependent variables obtained from the review of 20 previous works. This model has been extended and validated for different markets such as the United States, Chile or Croatia [2], [3]. In the case of the U.S. market, only 4 of these variables were statistically significant (planning, professional advice, education, and staffing) possibly due to the relatively small sample size.

There may also be different ways to take into account and measure the reputation of a VC investor. In fact, using a logit framework, studies show that reputable VC firms are more likely to lead their companies to successful exits [4]. This evidence also holds when analyzing the performance of individual VC investments. In another work, logistic regression analysis is used to study the relation of growth in 200 Finnish firms with founders' motive, their background characteristics, management styles, etc [5].

A study with a different approach analyzes the survival of 181 newly established manufacturing firms in north-east England [6]. The work uses log-logistic hazard models to study the relationship between the survival time of the firm and signals either related to the firm (plant size) or the macroeconomic aspects.

Other approach models the total amount of VC funding raised with a linear regression [7] . In this case, the dataset contains information about biotechnology US companies created from 1974 to 2011. The variables that define these companies are related to the number of VC investments received, the number of patents, the citations of these patents, and other geographical information.

So far, most of the approaches reviewed are based on regression analysis, mainly logistic regression. However, fewer works have explored alternatives based on artificial intelligence.

One of the earliest works in this stream presents a rule-based expert system that predicts the acquisition of companies [8]. This system achieves a success rate of 70% although evaluation is limited to a dataset of 200 companies. Even if an expert system can be used to select successful companies, data-driven approaches based on machine learning have been more popular. For example, Wei et al. (2009) propose the use of ensemble classifiers to predict about 600 cases of mergers

[1] https://www.crunchbase.com/

and acquisitions in Japan [9]. In this work, predictors are technological variables from patent analysis and both profiles of investors and candidate target companies. The authors report a global accuracy of 88% and precision over 40% when predicting an acquisition.

Similarly, Yankov et al. (2014) compare several machine learning methods to predict the success of 142 Bulgarian startups using data from a questionnaire [10]. The authors show that decision trees are the most accurate method and use them to reveal startup success factors, e.g. the presence of competitive advantage, founders experience in a similar position, etc. A more sophisticated approach combines supervised (Support Vector Machines) and unsupervised learning (clustering) for the prediction of business models with higher growth expectations and chances of survival [11]. The work considers startups from USA and Germany and achieves an accuracy of 83.6% when trying to predict the survival of a venture, but again using a small dataset of 181 companies.

Other authors compare the performance of human experts and machine learning techniques when predicting outcomes of early-stage firms [12]. They consider 2,506 Nigerian firms that participate in a business plan competition and conclude that machine learning methods do not achieve significant improvements compared to human experts. They report a 63% success rate for machine learning and 58% for human experts that led them to conclude that human experts also have difficulty in identifying which firms will succeed.

While machine learning seems a promising venue, the works presented so far have either worked with a small dataset or with data retrieved ad-hoc. More representative samples can be obtained by platforms like Crunchbase, which gathers data from hundreds of thousands startups, even though the data retrieved is not as rich as in the case of ad-hoc datasets.

Xiang et al. (2012) try to predict acquisitions through machine learning for companies founded between 1970 and 2007 [13]. As predictors, they use different kind of firm descriptors, including information on the management team and finance sources, but also from TechCrunch news. They highlight the problems related to the sparsity of the dataset, dropping-off approximately 20,000 companies because of the lack of a complete description. This way, they use a dataset with 60,000 companies described by 22 features and they segment them according to their business sector. They enrich the dataset with the distribution of news for each company in the 5 most representative topics from each business sector using a corpus of over 38,000 news. Unfortunately, only about 5,000 companies had a presence in the corpus. Finally, they show that considering the information provided by the news improves the results. They achieve a precision between 60% to 79.8% from half of the categories and Bayesian Networks outperform both SVM and logistic regression.

Other works use a Crunchbase dataset of over 80,000 startups from five states in the US from 1985 to 2014 to predict either an M&A (merger and acquisition) or an IPO (initial public offering) [14]. The author compares the performance of logistic regression, SVM and random forests for the prediction of the success of startups using the data provided by Crunchbase. The proposed approach comprises a data acquisition and selection stage. Next, a pre-processing stage tries to avoid the sparsity problem reported in other works [13]. To address this problem, the author proposes to re-code several variables and generate synthetic variables to represent potential interesting features. Another problem faced when trying to create the predictive model was the large class imbalance between successful and non-successful companies. After pre-processing, only 16.8% of the dataset consisted of successful companies. Therefore, he employs an oversampling strategy of the minority class to fix that issue. The author reports a precision close to 92%. However, the impact of the artificial increase of the dataset caused by the oversampling (with 60% more instances) is not taken into consideration when discussing the results. Moreover, the time window precedes the Crunchbase creation, so it is possible to observe a survival bias since successful companies whose foundation precedes that of Crunchbase are over-represented.

As a result, we can conclude that the literature reveals the potential of using machine learning to exploit Crunchbase data to create a decision support system for the prediction of the success of early-stage companies. However, we consider that previous approaches focus on too long time-windows, which are unrealistic for a VC, and only on IPO or acquisitions, rather than including a larger set of potential outcomes. The following section presents our approach and differences in the use of Crunchbase dataset to help VC investors screening promising early-stage companies.

III. PREDICTION OF SUCCESS IN EARLY-STAGE COMPANIES

The main goal of this work is the development and evaluation of a data-driven approach that uses machine learning to help VC investors scouting and selecting the best companies to support. As in some previous works, our approach relies on the data provided through Crunchbase.

The main features of our approach are the following:

- **Focused on early-stage companies**: We will consider that *early stage companies* are defined as active companies with less than a precise age at the start of a simulation window and that are in an early funding stage (earlier than series C), according to the startup fundraising stages.[2]
- **Time-aware analysis**: Our approach defines a realistic time window, which represents a reasonable investment window for a VC investor.

[2]There are various types of funding rounds: Seed, Series A, B, and C, and so on. While Seed can be achieved through investments of business angels and Micro-VCs, in Series A typically institutional investors and large funds kick in. Series B usually comes into play for a startup that is already profitable and that wishes to increase its profit margin. The next round is renamed as Series C. Many companies utilize Series C funding to help boost their valuation in anticipation of an IPO. However, some companies can go on to Series D and even Series E rounds of funding as well, mainly because they are in search of a final push before an IPO.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938659, IEEE Access

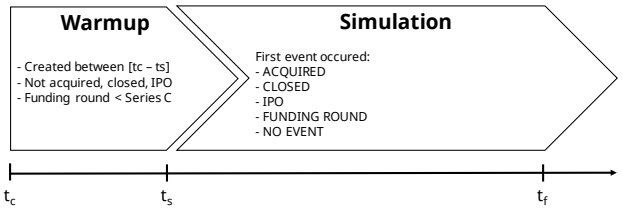

FIGURE 1. Warmup and Simulation window.

- **Multi-class prediction problem**: Our target variable represents the next event that a company will achieve in a defined simulation window. The prediction of these events is useful for the VC investor to make an investment decision because they represent the company progress. Namely: a company closes; it obtains investment through a new funding round; the company is acquired by another company; or it goes public via an IPO (allowing the company to raise capital from public markets). We also consider the case where no event is reported for a company in Crunchbase during the simulation window.

We want to emphasize that our *time-aware* approach offers a more realistic prediction than other works detailed in Section II. Instead of using all the data available from the dataset, our approach uses data from the companies that were early-stage startups before the time window under investigation. The approach aims to predict what will happen first to each company in the window.

More specifically, our approach defines the following time windows (shown in Figure 1):

- The approach defines a timestamp t_s that represents the moment where the VC investor decides to invest in a startup company.
- The Simulation window represents the temporal window where the company will evince its success. It is defined by a Start timestamp (t_s) and a Finish timestamp (t_f). The interval will be selected according to the time that VC investors consider to evaluate the success of a company.
- The Warmup window represents the period that defines what happened in a company *before* the VC investor decided to invest in that company. As our goal is startups, our approach only takes into account companies created between the start of the warmup window (t_c) and the time when the simulation begins (t_s).

In our experiment, we keep the time windows close to the current time. In this way, we want to minimize the survivorship bias that is obviously present in Crunchbase (or in any similar platform). Companies that succeeded are to some extent over-represented because the ones that failed early may have not even appeared in the database in the first place. By studying a period close in time, we also aim to learn the way startups and VCs operate nowadays, which might be different from previous best practices.

Furthermore, it also minimizes the problems of considering Crunchbase data *as is* when downloaded and not *as was* at the beginning of the simulation. There is a problem with some Crunchbase variables that are not dated and may not reflect the situation of the company at the time where the simulation started (e.g. the number of employees, the managers of the company, etc). While we try to avoid such variables, there is a more subtle bias that is present and that we will mention in Section V. All these aspects help make the results of our experiment more reliable.

Next sections will describe how the training and test sets are created according to our temporal constraints, which variables are employed by the decision models, and how these models are evaluated.

A. DATA SELECTION

Our data sample is extracted from a *Daily CSV Export* of Crunchbase from August 2018. The full dataset contains 623,232 companies, information about 799,446 company founders and 227,172 tuples about funding round events (mainly from the last 20 years).

We define a Simulation window of 3 years ($t_s = August\ 2015$ and $t_f = August\ 2018$) and a Warmup window of 4 years ($t_c = August\ 2011$). These windows mean that an early-stage company will not be older than 4 years at the time of the investment and we expect that it will raise new funds in no more than 3 years after the prediction (or the VC investment) is provided. These windows are considered adequate for a startup given the high failure rate in the early years; for example, at four years was about 44 percent in the US.[3]

According to these windows, we filtered companies using their creation date and removed the ones acquired, closed or that went public by an IPO during the Warmup window, or the companies which closed a funding round above Series C, which is not interesting for early-stage VC investments.

The final data sample consists of 120,507 companies with the company name, sector, country, age, and other additional information, and 34,180 funding round events about the selected companies.

B. TARGET VARIABLE

The ultimate goal of a VC investment is to invest in companies that will be acquired or go for an IPO. However, these companies are rare due to the natural selection inherent in the venture capital process [15]. A VC investor may also look for companies that advance toward new injections of capital and hopefully larger outcomes. This way, we have defined a multi-class target variable whose value is extracted from the events occurred during the Simulation window. Using only the first event occurred for a company, we define the following classes:

[3]https://smallbiztrends.com/2019/03/startup-statistics-small-business.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938659, IEEE Access

Class	Frequency	Ratio
CLOSED (CL)	686	0.57%
ACQUIRED (AC)	3,293	2.73%
FUNDING ROUND (FR)	21,682	17.99%
IPO (IP)	143	0.12%
NO EVENT (NE)	94,703	78.58%

TABLE 1. Class value distribution of the success measure.

- ACQUIRED (AC): The company is acquired during the simulation window.
- FUNDING ROUND (FR): The company reaches at least another round of funding in the simulation window.
- IPO (IP): The company will go for an IPO during the simulation window.
- CLOSED (CL): The company is closed during the simulation window. This class is overridden by ACQUIRED when the closed and acquired events occurred simultaneously in a short period (indicating that the company was successfully acquired and then closed by the acquirer).
- NO EVENT (NE): None of the previous events occurred during the simulation window.

In a perfectly rational environment, it can be argued that subsequent funding may be a proxy of non-success since it will indicate that a company is burning money and need a new capital injection. In the industry though, this is not the case because VCs urge companies to spend money quickly to grow faster to automatically elicit the ones that can return them the money in a reasonable time frame from the ones that are dead ends.

In the same fashion, a prior closure of a venture can be seen as money and energy saving both from the entrepreneur perspective as well as the VC side. Again, even though this makes perfect sense in a theoretical model setting, it is not what happens in practice and very rarely occurs. In fact, it is hard to disentangle the reasons for product failure and entrepreneurs tend to stick to that as long as possible. Furthermore, they prefer to use every penny at their disposal to improve the product, launch a new one, or pivot the company to win the market rather than going from a preemptive closure. There is no stigma in failing, and even VCs often prefer to try them all before considering the company as a failure. Their solution is not, in fact, withdrawing the money, but rather avoid to keep investing in companies that are not profitable.

The value distribution of the target variable in our data sample is shown in Table 1. While a priori only the CL-class represents a failure, the percentage of early-stage companies closed after three years is unreasonably low (less than 3%) given the low startup survival rate. We believe that many NE class companies may be closed indeed, but the Crunchbase database does not reflect it, as closure generally happens without any official announcement or regulatory filing. Furthermore, from a VC investor point of view, if a startup does not show any progress after 3 years (the time of our simulation window) it is not a good investment. Hence, we decided to consider that the NE class denotes a failed investment.

As a result, we consider three "success" classes (AC, FR, and IP) that roughly represent the 21% of the companies in our sample, and two "failure" ones (CL and NE) that represent the remaining 79%.

C. PREDICTOR VARIABLES

After a pre-processing stage for removing unnecessary variables, computing new synthetic variables and cleaning missing values, we compiled a set of 105 variables for each company in the data sample. Most of them were extracted from the events occurred during the Warmup window and some variables that may have changed the simulation window –like the number of employees– were omitted.

Although some of the predictors are selected based on previous research on predicting early-stage company performance –like variables concerning founders education [2] or company location [7], among others–, we included other variables available in Crunchbase.

The set of predictors can be divided into the following categories.

1) Company information

This category comprises general information about the companies, like location (we considered that *country_code* provided an acceptable granularity level and a homogeneous set of values) and the business sectors where the company operates from a list of 46 categories predefined by Crunchbase.

Additionally, we included the company age in months (*age_months*) at the beginning of the simulation (t_s) and we added variables that measure the presence of the company in social media networks. These binary variables indicate whether the company registered in Crunchbase its contact information (*has_email* and *has_phone*) or a Facebook (*has_facebook_url*), Twitter (*has_twitter_url*) or Linkedin (*has_linkedin_url*) account.

2) Funding information

The variables in this category summarize the funding events occurred in a company during the Warmup window. The availability of a temporal series of funding round events in Crunchbase allowed us to synthesize information about:

- Number of funding rounds that the company achieved before t_s (*round_count*) and the total amount raised in those rounds (*raised_amount_usd*).
- Data about the last funding round in Warmup window: funding round type (*last_round_investment_type*), amount raised (*last_round_raised_amount_usd*), company valuation after this funding round (*last_round_post_money_valuation*) and time lapsed, in months, between the beginning of the simulation (t_s) and when this funding round occurred (*last_round_timelapse_months*).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938659, IEEE Access

- Number of (unique) investors who participated in the funding rounds during the Warmup window (*investor_count*) and, specifically, in the last funding (*last_round_investor_count*).
- Qualitative information about the investors. We defined a category of renowned investors as the investment companies registered in Crunchbase and created variables for the number of unique renowned investors who participated in the funding rounds during the Warmup window (*known_investor_count*) and, specifically, in the last funding (*last_round_known_investor_count*).

3) Founders information

The last category comprises information about the people who founded a company. In addition to the number of founders (*founders_count*), we synthesized new variables that provide information about the heterogeneity of the founders according to their origin –number of different countries where the founders come from (*founders_dif_country_count*)– and gender –number of male (*founders_male_count*) and female (*founders_female_count*) founders.

Previous studies highlighted the importance of having a college education when building a new company [2], [3]. Crunchbase provides information about the education of most of the founders. However, the way this information is stored (in a free-form text) hinders the synthesis of qualitative variables about the education received by company founders. After revisiting the information contained in Crunchbase and observing that most of the education entries refer to higher education, we decided to synthesize quantitative variables about the total number of degrees obtained by company founders (*founders_degree_count_total*), as well as the maximum (*founders_degree_count_max*) and the average number of degrees (*founders_ degree_count_mean*) among them.

The sparsity problem is evident also in this category, as most of the companies do not have information about their founders or their education. In this case, we consider the absence of data as useful information and consider 0, where it corresponds. It means that the company has not updated the information about the founders in Crunchbase.

D. MODELS AND ALGORITHMS

We have considered five different machine learning classifier algorithms:[4] Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), Extremely Randomized Trees (ERT) and Gradient Tree Boosting (GTB).

Decision Trees, Random Forests, Extremely Randomized Trees and Gradient Tree Boosting are tree-based classifiers. We considered tree-based classifiers for the following reasons:

- They incorporate *feature selection*, so they are able to cope with a high number of variables of presumably very different importance for the classification problem, as it is in our case.
- They are *white-box* classifiers that can be interpreted since it is possible to measure the relevance of the features used for the classification. This is particularly useful because we are interested in understanding the classification decision and in identifying success drivers.

[4] Algorithms implemented in Scikit-learn. http://scikit-learn.org/

Classifier	Accuracy
Decision Trees	74.6
Random Forests	81.8
Extremely Randomized Trees	81.9
Gradient Tree Boosting	**82.2**
Support Vector Machines	81.7

TABLE 2. Global accuracy of the classifiers in percentage

We consider decision trees as a baseline and because it has been successfully used in the literature [10]. However, decision trees have typically a limited success compared with more sophisticated classifiers, such as the tree-based ensemble classifiers considered (RF, ERT, and GTB). Ensemble classifiers are known to be more accurate than any of its members if the classifiers in the ensemble are accurate and diverse [16].

As a complement of the tree-based classifiers, we will use a classification method based on a different paradigm: Support Vector Machines (SVM). This method has already shown good results in plenty of fields, including VC investing [13], [14]. While originally designed for binary classification they are extended for multi-class classification using a one-vs-one scheme. It is also a method that is effective in high dimensional spaces as the one we are facing. As a disadvantage, it is important to mention the lack of transparency of its results, contrary to what happens with tree-based classifiers. Another disadvantage may be its computational time when using non-linear kernels, so we will use a linear version of the method.

To evaluate the classifiers we use stratified k-fold cross-validation with $k = 5$. The validation is stratified to preserve the same amount of companies of each class in each fold. This is important because the data is imbalanced. We also use the same dataset partition (i.e., the same folds) for training all the classifiers. In this way, we eliminate the impact of different partitions when comparing their performance. Naturally, we will report the performance of the classifiers in the validation set (aggregating the k validation subsamples).

IV. RESULTS

Table 2 shows the global accuracy of our classifiers in the validation subsamples. The classifier that performed worse was unsurprisingly the decision tree with 74.6%. This classifier performs in principle worse than a naive classifier that always predicts the majority class (i.e., "no event"). Such classifier would have an accuracy of 78.6%, because this value is the frequency of the "no event" class in the dataset (see Table

1). However, decision tree predictions are more informative as we will see below when analyzing the performance for each class in the target variable. The rest of the classifiers considered performed better than the naive classifier and the GTB classifier was the best, followed closely by the other methods considered.

Global accuracy is a general performance indicator, but a VC is especially interested in those classes that reflect a successful scenario (FR, AC, and IP). If the classifier performs well in flagging companies that belong to those classes, the VC could become more confident in betting on those companies. As a result, we analyze below the classifier performance for each class considered in the target variable.

A. ANALYSIS OF THE RESULTS FOR EACH CLASS

Table 3 shows the performance of the considered machine learning algorithms for each class in our target variable. As in an information retrieval problem, the metrics provided will be precision, recall and F1 score for every class. For a given class, high precision means that an algorithm returned substantially more relevant instances than irrelevant ones, while high recall means that an algorithm returned most of the relevant instances. The F1 score is the harmonic mean of the precision and recall.

In this problem, recall is not a critical measure. VCs do not need to find all the "interesting" companies available in the world (or in the database) since they cannot invest or even consider investing in all of them. However, a minimum recall is important because the classifier should provide the VC with a sufficiently large subset of relevant companies to invest in. In this sense, we suggest going beyond the percentage value of recall and considering the number of companies actually retrieved by the classifier for the considered class.

On the other side, a VC investor is primarily interested in increasing the success rate when making the decision about which company invest in. Hence, we need to focus on precision of profitable classes (FR, AC, and IP). Thus, precision will serve us for establishing the main comparisons among classifiers, while recall will be used to nuance our analysis. F1 score is merely added as a complementary measure for comparisons, but will not inform our analysis.

Looking at the results for the "closed" companies in Table 3, no single classifier performed well. In the case of SVM, the classifier did not even activate for this class, while in the rest of the classifiers activations were false positives (except for few true positives in the case of decision trees). This could be due to the small number of companies labeled as "closed" in the dataset, but the figures in Table 3 show that this does not apply to the "acquired" companies that are even fewer (686 versus 143). Thus, we believe that the real problem is that the "no event" class includes companies that are closed, but whose closure has not been updated in Crunchbase, as we already anticipated in Section III-B. This fact makes very difficult for the classifier to really discriminate between those two classes.

Regarding the performance for the "no event" class, we do not notice relevant differences in the performance of the classifiers in terms of precision (all between 0.83 and 0.85). Remarkably, all of them outperform a naive classifier always predicting the "no event" class (precision of 0.79), and therefore an investor could discard companies flagged as "no event" using any of our classifiers. From a practical perspective, we would choose the classifier with a higher recall because it means that it will activate more times. Hence, we do not consider the decision trees but rather take into account the rest of the classifiers as equally good, being able to identify around the 95% of the "no event" companies. Bear in mind that, as above mentioned, we consider the "no event" class as not interesting for a VC investor since it probably includes companies that are not attractive (i.e., not in the startup cycle) or closed ones where closure was not reflected in Crunchbase.

The differences in terms of classifier performance increase for the "funding round" class. The decision tree has the worst precision (0.43), while the precision for the rest of the classifiers varies between 0.6 and 0.64. The highest precision is offered by SVM and GTB. Since they are very similar, we could prefer GTB because is the classifier with the highest recall (0.4 versus 0.33). However, in this case, recall is not highly relevant because the number of "funding round" in our dataset is 21,682 and even the 33% of them still represents an incredibly large number of potential investments to analyze for a single VC.

In the "acquired" class the precision decreases notably. Random Forest and Extremely Randomized Trees are the best algorithms with 0.33 and 0.31, respectively, and the same recall. While the performance is apparently low, the percentage of the "acquired" class in the population is the 2.7% and it is never over 14% in the VC funnel of US companies [15]. Considering these facts, these classifiers enhance more than 10 times the chances of finding a company that will be acquired in the sample, which would represent an outstanding performance for a VC investor. While the recall is low in both cases (0.03), its translation to absolute numbers means that each of these classifiers activated close to 300 times and got it right around 90 times. Again, the absolute numbers are satisfactory for the investment potential of many if not most VC investors.

Finally, we analyze the results in the "IPO" class, which is the less frequent event of our dataset, with only 143 companies. As in the "acquired" class, Random Forest and Extremely Randomized Trees are the best algorithms with a precision of 0.44 and 0.27, respectively. However, their success rate is 4/9 and 3/11, respectively, and the differences may be due to mere chance. Moreover, the number of activations for all the algorithms is very low for this class, which means that is a difficult class to discriminate. This may be due to similarities between the companies within "IPO" class and the other classes that represent success, but also that the "IPO" class is algorithmically neglected because of its low frequency.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938659, IEEE Access

	CL			NE			FR			AC			IP		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
DT	0.02	0.02	0.02	0.85	0.85	0.85	0.43	0.43	0.43	0.09	0.10	0.10	0.04	0.03	0.04
RF	0.00	0.00	0.00	0.85	0.94	0.89	0.60	0.44	0.51	0.33	0.03	0.05	0.44	0.03	0.05
ERT	0.00	0.00	0.00	0.85	0.94	0.89	0.61	0.43	0.51	0.31	0.03	0.05	0.27	0.02	0.04
GTB	0.00	0.00	0.00	0.85	0.95	0.90	0.64	0.40	0.49	0.17	0.003	0.01	0.07	0.01	0.02
SVM	-	-	-	0.83	0.96	0.89	0.64	0.33	0.44	0.00	0.00	0.00	-	-	-

TABLE 3. Precision, recall and F1 score for each class (columns) and each machine learning algorithm.

In summary, the main conclusions of the analysis are:

- The classifiers are of no use for predicting the "closed" companies, and of little practical use for the "IPO" ones.
- For the "no event" and "funding round" classes ensemble classifiers provide the best results, with a slight preference for the Gradient Tree Boosting.
- Random Forests and Extremely Randomized trees are the best classifiers for the "acquired" class, even if their recall is low.
- SVM was not able to generalize the features of those classes with few instances (CL, AC, and IP), and its performance in the rest of the classes was not the best one. Hence, we consider that should not be considered in the present approach.

We should proceed with caution to select the best model among the ensemble classifiers. In fact, if we look at global performance in the majority classes (NE and FR), we would choose GTB. However, if we also consider the precision of the classes that represent acquisition, we would choose RF, as it offers a good balance of precision in the FR and AC classes.

B. REINTERPRETATION OF THE CLASSIFICATION ERROR FOR VC INVESTORS

In this section, we turn our attention to classification errors. In a typical classification problem, errors are undesirable. However, in this domain, some errors can cause no harm to the portfolio of a VC investor or can even turn into pleasant surprises. For example, if a classifier predicts that a given company will be "acquired", but it turns out that the company just obtains a "funding round", the error is not vital, because it means that the company keeps growing. On the opposite case, finding that a company the classifier predicted would obtain a "funding round" is instead being "acquired" is again a successful event, because the investor exits its investment with a profit.

Following this idea, we will reinterpret the classification error, collapsing our multi-class target variable into a binary variable with two classes: "good" and "bad". The "good" class includes the classes "funding round", "acquired", and "IPO", while the "bad" one is including the "no event" and the "closed" classes. Below we show the "binarized" classification error of the algorithms considered.

In terms of global accuracy, the decision tree classifier is the worst with 0.77, while the rest of the classifiers obtain a similar result, i.e., 0.83.

In Table 4 we show the "extended" precision values for each class, that is, we consider that an error only happens if you predict one of the "good" classes and you get one of the "bad" ones or vice versa. We also include the number of "extended" true positives to better contextualize the precision value. By definition, all precision values must be (and in fact are) equal to or better than the ones shown in Table 3.

The comparisons among methods are similar to those that could be extracted from the previous table. However, for a VC investor the chances to do wrong decrease when investing in companies flagged as "good".

In Table 5 we aggregate the results of Table 4 to show the precision of two categories: "good" and "bad". This table makes possible to better compare the global performance of the classifiers in terms of good and bad investments. According to the numbers, the precision of GTB and SVM are superior to the other methods for the "good" category and slightly worse for the "bad" one.

Given the importance of the "good" signal when constructing a portfolio, we consider that GTB and SVM are the methods that better suit this domain. An investor following the "good" signals provided by the GTB or SVM would have a success rate close to 7 out of 10, which is truly remarkable. Given the better performance of GTB in terms of recall and precision of the "bad" class, we would finally recommend using GTB for the portfolio construction using binary signals.

In experiments not reported here for the sake of brevity, we trained the classifiers using the binary target variable with the "good" and "bad" classes. Results were roughly similar, so it does not seem an advantage to consider the binary target instead of the more nuanced multi-class one.

C. FEATURE IMPORTANCE

In this section, we analyze the feature importance of the multi-class approach shown in Section IV-A and not the binary one.

The problem with machine learning classifiers is that they are usually black-box decision tools. However, decision-makers need to understand why a decision is suggested or at least which factors are taken into account.

From the classifiers used in our work, SVM are black boxes, but tree-based classifiers make possible to analyze the importance of different features. From all of them, decision trees are the most transparent and understandable, since the tree can be read and understood. On the other hand,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938659, IEEE Access

	CL		NE		FR		AC		IP	
	Prec.	TP	Prec.	TP	Prec.	TP	Prec.	TP	Prec.	TP
Decision Trees	0.74	535	0.86	80638	0.47	10195	0.34	1228	0.39	53
Random Forest	0.33	1	0.86	89397	0.64	10318	0.53	155	0.56	5
Extremely Randomized Trees	0.33	1	0.86	89721	0.65	10039	0.47	128	0.55	6
Gradient Tree Boosting	0.44	24	0.85	91016	0.68	9139	0.55	35	0.41	11
Support Vector Machines	-	0	0.84	91817	0.68	7604	1	1	-	0

TABLE 4. Precision values and True Positives for each machine learning algorithm after reinterpreting the classification error.

Classifier	Bad	Good
Decision Trees	0.86	0.45
Random Forests	0.86	0.64
Extremely Randomized Trees	0.86	0.64
Gradient Tree Boosting	0.85	0.68
Support Vector Machines	0.84	0.68

TABLE 5. Precision values for each machine learning algorithm after reinterpreting the error and aggregating the original classes into "good" and "bad".

tree-based ensemble classifiers are more difficult to analyze because they use multiple trees to classify, but still, it is possible to estimate the importance of the features used for classification.

Figure 2 shows the ten most important features of the tree-based ensemble classifiers.[5] In order to estimate features relevance, we have re-trained the classifier with the whole dataset. The Figure shows that 4 of the variables in the top 5 are the same for the three classifiers, and the other appears in two of them –namely, *age_months*, *founders_count*, *has_linkedin_url*, *founders_dif_country_count* and *raised_amount_usd*. Given the frequency of the classes in the dataset and the performance of the classifiers, these variables most likely help to discriminate between the "no event" and the "funding round" classes.

For example, *age_months* represents the age of the company in months. This makes sense because for a startup the older it is the higher the probability to survive and hence the higher the probability to receive funding or to be acquired. Since we consider companies up to 4-years old, the variable is related to the maturity of an early-stage company.

The variable *has_linkedin_url* means that the company has a LinkedIn profile and that link appears in Crunchbase. Being LinkedIn the most relevant professional networking website, the presence of the company there and the appearance of the link in Crunchbase seems to have a relevant role in signaling potentially attractive companies.

In the case of *founders_count*, the variable is important and it means that some information about the founders exists in Crunchbase. As mentioned in Section III-C3, if no information on the founders is available the respective value would be zero. However, the variable *founders_dif_country_count*, which represents the number of different countries of origin of the founders, provides different information. It probably speaks of the internationalization of the company, which is interpreted as a good sign.

Finally, *raised_amount_usd*, which is the amount of money raised in US dollars, is related to the historical ability of the company to obtain funds. This likely means that capital is an edge, and having receive (or receiving) funds will increase the probability of success.

The rest of the variables are related to the completeness of the company information in Crunchbase (*has_email*, *has_phone*), to the gender and education of the founders (*founders_male_count*, *founders_degree_count_mean*), or to the funding rounds (*last_round_timelapse_months*, *round_count*). Interestingly, some business sectors looks more popular and attractive, e.g., *Health Care* or *Science and Engineering*. Some countries of origin also look to have a strong relevance. Those countries, unsurprisingly, represent strong economies, e.g., USA, China, and Sweden.

V. DISCUSSION

We have shown the potential of using machine learning algorithms to predict different levels of success — and not only IPO or acquisitions— of early-stage companies in a medium-term time window. This approach had not been explored before in the literature. The results show a global accuracy of around 82% of the best algorithm, Gradient Tree Boosting. Looking at the details, most algorithms explored obtain a precision for determining that a company will achieve the next funding round between 64% and 68%. This result means that a VC investor can construct a portfolio with less risk using such classifiers. The best method for flagging high-gain companies — those that are acquired or go public through an IPO— is Random Forest which obtains a precision of 0.33 and 0.44 for acquisitions and IPOs, respectively. The result is outstanding given that the percentage of such companies in the dataset is less than 3%. The recall is low, but it sums up to over 100 companies between the acquired ones and the ones that went public, which is more than enough for most VC firms. In rough terms, a VC firm would need to invest in 300 companies flagged as acquired or IPO by the classifier and expect only 100 to turn into high-reward investments.

These results are certainly promising for VC investors, who typically do not obtain such success rates. Classifiers with predictions of different levels of success can be incorporated in their screening process and their portfolio con-

[5]Decision trees are not shown because they obtained worse results than the ensemble classifiers.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2019.2938659, IEEE Access

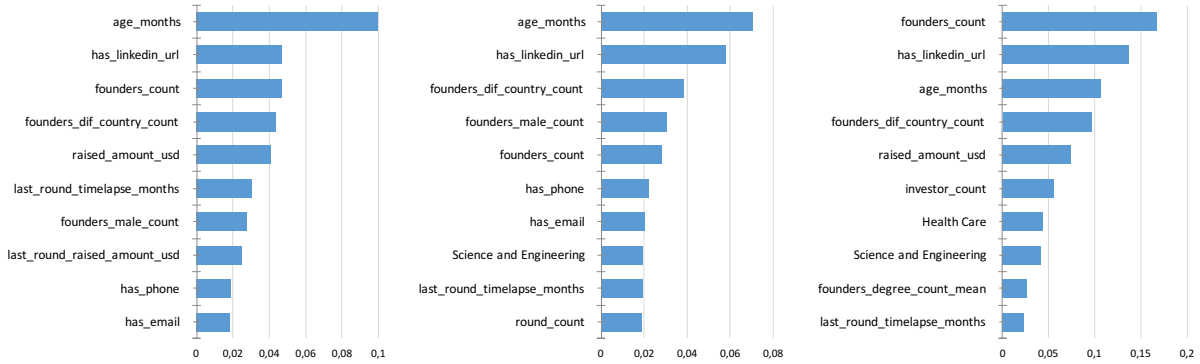

FIGURE 2. Feature importance for the tree-based ensemble classifiers (left: random forests. center: extremely randomized trees. right: gradient tree boosting)

struction. However, the results reported might not be free of potential biases that need to be considered and are therefore discussed below.

It is important to bear in mind that the experiment may have a bias since all the data was retrieved at the end of the simulation window (t_f). However, our simulation starts in August 2015 (t_s) and we ideally should be using the information available in Crunchbase at that time. For example, perhaps not all the companies we are considering were already in Crunchbase in August 2015. They could have been added later. It is also reasonable to think that this problem creates a bias towards successful companies since those are more likely to be added in Crunchbase.

A similar and more subtle bias can be present as well. It is related to the company information available in Crunchbase at the start of the simulation. It is possible that a company at that time did not have specific information, such as the profile of its founders that was very relevant for classification. This information could have been added later, especially in the case of companies that progressed and had success to some extent. The Crunchbase data dump does not determine when this information was added and this may distort the results.

Finally, it is important to remember that the tool does not increase the investor's ability to actually close the deal, but it only augments the ability to process information and assess companies in absence of more traditional financial data. However, we believe that even if these problems may affect the performance of the classifier presented, they should still produce useful advice for VC investors. The effect of the bias might be evaluated by simulating the use of the system in real-time, which takes years to be done, given the time window considered.

VI. CONCLUSIONS

This work has shown that machine learning can help in the baseline screening to early-stage investors that are looking at potential investments with no relevant quantitative data or track record. Our experiment in a realistic setting demonstrates that a multi-class machine learning classifier can help to increase the success rate of an investor.

Clearly, this tool is only one of many in a VC's toolbox, but it is very relevant when it comes to quickly skim through the thousands of opportunities an investor sees every year. Being aware of specific features that may signal a company will outperform its peers is key to reduce both risk and uncertainty when investing. It cannot and should not be the only tool used to evaluate a company, but it is definitely the first step an investor should undertake in order to move forward a conversation or not.

From the perspective of a venture capitalist, a multi-class approach such as the one proposed is much more useful than a binary one, which only gives "good" or "bad" scenarios. On one hand, our approach can help to reduce the risk of the portfolio, as we considered a class that represents moderate success for a company — to proceed to a subsequent funding round— which is easy to spot for the classifiers. On the other hand, since best performing funds are mostly derived from a very few numbers of companies that end up producing outsized results, a VC may decide to focus just on the "acquired" signal, which has a higher risk, but higher potential rewards. We also showed that some classification errors of the multi-class approach cannot be considered harmful or dramatic for the portfolio. The nuanced information from the multi-class approach can be used to set up a detailed portfolio strategy. Moreover, the classifier's information can be combined with "fundamental" analysis on the company activity, the sector where it operates, the profile of the founders and managers, etc.

Furthermore, the work presented here can be refined further and some of its limitations can be tried to be overcome. Besides those related to training the classifiers, such as hyperparameters optimization or using sampling strategies for imbalanced classes, new variables could be considered. For example, the information from the founders has proven to be relevant in our approach — even in the form of rough numerical variables with high sparsity— and its relevance is documented in the literature [5], [10], [13]. Since LinkedIn provides founders and managers data and in a structured way, more informative variables describing their professional experience or academic background could be included in our

approach, which could lead to better performance.

founding teamÂt's diversity (so various types of degrees instead of mere number of them

Additionally, specific classifiers for different countries and/or business sectors could be trained. These classifiers could outperform the ones presented and offer additional insights and signals to spot interesting companies for each country or sector.

The practitioners implications of this first work are potentially huge. We believe that a more refined tool could change the entire dynamics of the industry. The baseline assumption any fund does is that most of its investments will be a loss and will be written off the book and that the entire return will be driven by no more than 10% of the investments made. The risk-return profile is thus very unstable, which forces investors to only pursue opportunities that might turn into moonshots and reject those investments that, although good, cannot achieve at least a 10x return. However, if there would be a way to better predict a company success, an investor could easily invest in a portfolio of companies all producing a 2x returns rather than looking for a single unicorn, eventually achieving a higher return than what most funds do nowadays. This would also drive the valuations down and potentially increase innovation and entrepreneurial activity, and therefore should be widely embraced by policy-makers as well.

REFERENCES

[1] R. N. Lussier and S. Pfeifer, "A Crossnational Prediction Model for Business Success," Journal of Small Business Management, vol. 39, pp. 228–239, 2001.

[2] C. E. Halabí and R. N. Lussier, "A model for predicting small firm performance," Journal of Small Business and Enterprise Development, vol. 21, no. 1, pp. 4–25, feb 2014.

[3] R. N. Lussier and C. E. Halabi, "A three-country comparison of the business success versus failure prediction model," Journal of Small Business Management, vol. 48, pp. 360–377, 2010.

[4] R. Nahata, "Venture capital reputation and investment performance," Journal of Financial Economics, vol. 90, no. 2, pp. 127–151, 2008.

[5] H. Littunen and H. Niittykangas, "The rapid growth of young firms during various stages of entrepreneurship," Journal of Small Business and Enterprise Development, vol. 17, no. 1, pp. 8–31, 2010.

[6] P. Holmes, A. Hunt, and I. Stone, "An analysis of new firm survival using a hazard function," Applied Economics, vol. 42, no. 2, pp. 185–195, 2010.

[7] S. Hoenen, "Do Patents Increase Venture Capital Investments between Rounds of Financing," Master's thesis, Wageningen University and Research Center, the Netherlands, 2012.

[8] S. Ragothaman, B. Naik, and K. Ramakrishnan, "Predicting Corporate Acquisitions: An Application of Uncertain Reasoning Using Rule Induction," Information Systems Frontiers, vol. 5, no. 4, pp. 401–412, dec 2003.

[9] C.-P. Wei, Y.-S. Jiang, and C.-S. Yang, "Patent analysis for supporting merger and acquisition (m&a) prediction: A data mining approach," in Designing E-Business Systems. Markets, Services, and Networks, C. Weinhardt, S. Luckner, and J. Stößer, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 187–200.

[10] B. Yankov, P. Ruskov, and K. Haralampiev, "Models and Tools for Technology Start-Up Companies Success Analysis," Economic Alternatives, no. 3, pp. 15–24, 2014.

[11] M. Böhm, J. Weking, F. Fortunat, S. Müller, and I. Welpe, "The Business Model DNA: Towards an Approach for Predicting Business Model Success," in Proceedings der 13. Internationalen Tagung Wirtschaftsinformatik, 2017, pp. 1006–1020.

[12] D. J. McKenzie and D. Sansone, "Man vs. Machine in Predicting Successful Entrepreneurs: Evidence from a Business Plan Competition in Nigeria," CEPR Discussion Paper No. DP12523, Tech. Rep., 2017.

[13] G. Xiang, Z. Zheng, M. Wen, J. I. Hong, C. P. Rosé, and C. Liu, "A Supervised Approach to Predict Company Acquisition with Factual and Topic Features Using Profiles and News Articles on TechCrunch," ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, pp. 2690–2696, 2012.

[14] F. R. d. S. R. Bento, "Predicting start-up success with machine learning," Master's thesis, Universidade Nova do Lisboa, Portugal, 2018.

[15] CB Insights, "Venture Capital funnel shows odds of becoming a unicorn are about 1%," https://www.cbinsights.com/research/venture-capital-funnel-2/, 2018, accessed: 2018-01-11.

[16] L. K. Hansen and P. Solomon, "Neural network ensembles," IEEE Trans. Pattern Analysis and Machine Intelligence, no. 12, pp. 903–1002, 1990.

JAVIER ARROYO is Associate Professor at Universidad Complutense of Madrid (UCM) since 2013. He got a PhD degree in Computer Science from Universidad Pontificia Comillas (2008).

He has research experience in time series forecasting, agent-based simulation, and machine learning applied to different domains and real-life problems.

FRANCESCO COREA is Vice president at Four Trees Merchant Partners and researcher at Ca'Foscari University of Venice.

His focus is on venture capital, entrepreneurship and artificial intelligence, and has worked as an investor and startup advisor for the last few years. He holds a Ph.D. in Economics from LUISS University and he's a former fellow in Applied Math at UCLA.

GUILLERMO JIMENEZ-DIAZ is a Computer Research Scientist and Associate Professor at Universidad Complutense of Madrid where he received his Ph.D. in Computer Science in 2008.

His research is concerned to Recommender Systems and its combination with Social Network Analysis. His main domains of application are tourism and e-learning, but he is also interested in Augmented Reality technologies in Museums.

JUAN A. RECIO-GARCIA is Head of Department of Software Engineering and Artificial Intelligence at Universidad Complutense of Madrid, where he obtained a PhD in Computer Science in 2008.

His research has focused on the confluence of Software Engineering and Case-Based Reasoning (CBR), developing the COLIBRI platform to build CBR systems. He is also working in the areas of Context-aware and social Recommender Systems.