How Likely is the Investment of a Venture Capital Firm in a Latin American Country?

Machine Learning Models Based on Experience, Distances and Network Features

Marcelo Guzella*, Felipe Buchbinder

March 15, 2023

Abstract

The ability to accurately estimate whether a venture capital firm will or will not invest in a market benefits investors, companies and investment promotion agencies. In this analysis, we applied machine learning techniques to predict whether a VC firm will invest in a specific country in Latin America. Predictors included firm characteristics and past behavior, country macroeconomic situation, network-level variables, and institutional and geographic distances between the target market and the country where the firm is headquartered. The database encompasses more than 10 thousand funding rounds from 2002 to 2020. The gradient boosting algorithm presented the best predicting performance, in terms of the area under the curve that relates true positive with false positive rates. A classification using this technique has a precision 24 percentage points higher than the one of naively selecting firms that invested in the recent past, which translates into savings or gains in capital allocation or investment promotion processes.
Keywords: venture capital, machine learning, national distances, syndicated investments

1 Introduction

The venture capital market plays an important role in funding small and medium-sized enterprises (SMEs), ultimately contributing to economic development and innovation (Florida & Kenney, 1988). Venture capital and private equity (in this document called VC, in a broader scope) operations are also essential financing sources to start-ups and distressed companies (Jelic et al., 2005; Wright Robbie, 1998). Despite some downturns, the VC industry has shown relevant historic growth. The number of deals in the world increased from less than 12,000 in 2010 to more than 30,000 in 2021, while global deal

*Corresponding author: marcelo.guzella@fgv.br

value went from less than 50 billion dollars to more than 460 billion dollars in the same period (Nolting et al., 2021).

However, the VC investment process involves many challenges and a high level of uncertainty. Success depends on finding good investment opportunities, assessing those opportunities, closing the deal, monitoring the portfolio companies and exiting the investment (Mingo et al., 2018). Several studies focus on the second half of this process (from closing the deal to exiting the investment) to analyse performance, using indicators such as multiples, returns and distributions to paid-in capital (Kanuri & Hanby, 2020; Minardi et al., 2015; Phalippou, 2020). Indeed, portfolio companies deal with substantial financial, operations and market risks that lead to uncertain results and relatively high mortality rates (Castellanos, 2023). Even for companies that survived or succeed, limited liquidity, for instance, increases spreads and makes exits often challenging. Survival can be a tough task also on the side of VC firms, since it depends on securing resources, managing growth and dealing with downturns in a volatile and competitive industry (Rider & Swaminathan, 2012). VCs that invest abroad should also develop and manage resources and capabilities to face cultural and social differences, as well as geographic and institutional distances, not to mention exchange rate risk (Ruhnka & Young, 1991).

Our work focuses on the first half of the venture capital investment process: finding and assessing investment opportunities and closing the deal. In order to succeed in those stages, the VC firm should be capable of accessing industry-specific and country-specific information that flows through the network in which it is embedded. Besides the challenge of developing the necessary skills to access and analyse this opportunities, much of the uncertainty in those stages is basically related to data and environment. Many of the target companies are early-stage, SMEs and start-ups, so data about them can be scarce or not reliable. With respect to destinations that are far or culturally different or that have weak regulatory settings, data can be also harder to gather and analyse. Moreover, negotiating, handling contracts and closing the deal can be hard tasks when there are significant differences in language and norms and when there are weaker legal systems, regulatory quality or contract enforcement (Kaplan & Strömberg, 2009).

Naturally, all those uncertainty sources that affect investment likelihood propagate to VC fund investors and other stakeholders. Due to financial or strategic reasons, they may put money in a VC fund with the expectation that it will invest in a particular region. However, the VC firm may be (unpredictably) incapable or uninterested to invest in this region. When the VC firm does not make investments in the markets where the investor expected, losses should be accounted for in terms of diversification, asset allocation or the achievement of strategic objectives. This is particularly crucial when it comes to VC and cross-border investments due to the obstacles that investors would have, in terms of resources and capabilities, to perform the transactions on their own and not depending on VC firms.

In addition, investment promotion agencies or companies seeking funding would benefit from predicting the VC firms that most likely will invest in the region or country where they are. Assuming that contacts, incentives and partnerships will not marginally affect conversion in the same way to more probable and less probable firms, they can optimize results by prioritizing those efforts to a specific niche.

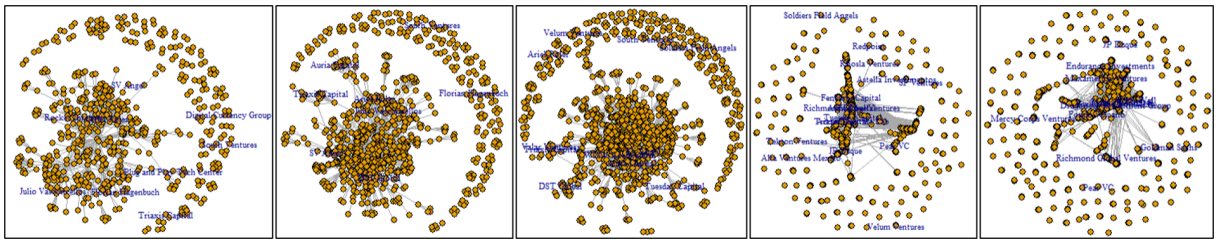

Figure 1: Graphs representing the syndicated investments in LatAm from 2016 to 2020
Note: Network corresponding to 2016 on the LHS and network corresponding to 2020 on the RHS. Each node is a VC firm and each connection is at least one coinvestment made over the 5 previous years. Only firms with at least one coinvestment over the moving window are plotted.

Last but not least, it pays VC firms to more precisely assess their own likelihood to invest in a geographic market. Finding results that are contrary to their strategic and financial objectives (or of the funds they manage, or even of their limited partners), they will be able to acquire or develop the necessary resources to improve these odds. They can, for instance, work to adapt their network position. Another alternative is to adjust their financial or strategic goals due to those odds.

Considering those benefits, we applied machine learning (ML) techniques to predict whether a VC firm will or will not invest in each of the eight Latin American countries with the most developed VC markets. Our sample involved more than 10 thousand deals closed in the region from 2002 to 2020, by around 1.4 thousand VC firms from 67 countries. We selected Latin America because of its striking growth: VC investments grew by 30% per year from 2005 to 2011 in the region. (Silva et al., 2022; Stein & Wagner, 2018).

Two of our predictors result from syndicated investment networks among VC firms. Syndicated investing is as a striking feature of VC (Lerner, 1994). Syndicates are a form of joint investment of two or more VC firms in a company, aiming at portfolio diversification, better selection, value addition and certification (Lerner, 1994). On the other hand, it implies giving up part of the return, possible informational asymmetries and conflict of interest between the leader and other syndicate members, as well as possible difficulties in coordinating the decision-making process (Wright & Lockett, 2003). Syndicated investments result in networks of interactions with attributes that have also been extensively studied in the social network literature, in order to understand their effects on performance (Bellavitis et al., 2017). Figure 1 shows graphs that represent the syndicated investments in Latin America from 2016 to 2020. Each node is a VC firm that did at least one syndicated investment. The edges link firms that co-invested at least once in the previous five years.

The cohesiveness of the network around a node is one of these attributes, measuring whether the VCs that are part of it are closely related or whether the network is more dispersed and characterized by structural holes. Cohesion creates more resource spillovers, visibility and legitimacy, especially benefiting firms with lower status and maturity, but it can also make it harder to generate unique deals and creates pressure to join partners, sometimes having to support them. A more dispersed network, on the other hand, allows for more flexibility and access to less redundant information (Bellavitis et al., 2017).

Centrality is also an important attribute, measuring how well connected a firm is to the network parts with greater connectivity. Firms with greater centrality usually have greater influence and status, due to their advantageous position in the network, which translates into better access to information, opportunities and references. On the other hand, unique and valuable knowledge can spread more easily and potentially benefit free riders in this network, compared to non-syndicated investments (Bellavitis et al., 2017).

Another aspect of great relevance in the venture capital industry is the distance between the firm and target companies (or investees), particularly the institutional and geographic distances. Firms that invest overseas commonly have to deal with institutional differences in terms of regulatory processes and rules that define contract enforcement and property rights. As pointed out by Mingo et al. (2018), firms headquartered in markets with strong institutions may overestimate the reliability of accounting information to assess opportunities and struggle when are dealing in markets with weaker institutions, with less contract enforcement and legal system robustness. Geographical distances also significantly influence investment decisions, because they determine information asymmetries, risk of adverse selection, and transaction, monitoring and communication costs.

Mingo et al. (2018) argued that the firm's centrality influences the relationship between distances and the chance to invest, while Bellavitis et al. (2017) pointed out that this centrality (and also the firm's maturity) mitigates the positive effects of network cohesion in the investment performance. Both surveys uncover relevant findings to the literature, contributing to the decision-making process of VC firms. However, Mingo et al. (2018) did not consider the network cohesiveness in the analysis of the relationship between centrality, distances and chance to invest. The effect that centrality induces in the relationship between distances and invest probability must be different for firms in cohesive and dispersed networks, due to the information flow, resources and flexibility. Likewise, the relationship between cohesion, centrality and performance might depend on how closely (geographically and institutionally) the firm is to the investees, or target companies, considering their impact on the assessment of opportunities, information asymmetries and monitoring costs.

This paper is organized as follows: in the next section we describe the research problem and how this analysis contributes to the fields of venture capital and machine learning. We then describe data, methodology, selected techniques and variables. After that we describe outputs and analyse the performance of the models. We then conclude and suggest possible improvements and future developments.

2 Research Problem and Contributions

It is commonly unclear in which markets the VC firm will invest. When the VC fund prospectus or bylaws do not restrict the activities to specific markets, one might wonder whether the process from selecting opportunities to closing the deal will be in line with his/her preferences. Even when such rules are clear and do not have flexibilities, it might be important to estimate those odds because they also depend on the ability of the VC firm to invest in a particular market and features of this market.

This work addresses the limited ability to predict the geographic operation of VC firms. Limited Partners, for instance, may apply our method to more accurately anticipate whether the capacity and interest of the General Partner, in terms of investing in a particular market, will converge to their beliefs and goals. Investors can use the same procedure to allocate resources in VC funds with a geographic coverage prospect in line with their expectations. We estimate this prospect using network cohesion, network centrality (status), VC firm maturity, geographic and institutional proximity between the VC firm and the target market, past experience of the VC firm and the macroeconomic situation of the target market.

One can argue more intuitively about how investment odds are expected to be affected by VC firm maturity, status and proximity. With respect to cohesiveness, a firm that is embedded in a disperse network (one with structural holes) is more capable of finding and benefiting from unique opportunities. The cohesion of the network and the institutional proximity between the firm and the investee are complementary attributes. Hence, the former contributes to being called for the best deals, while the latter is more useful to entering, monitoring and exiting the investments.

Being part of a cohesive network is advantageous for low-centrality VCs, but it may limit the performance of high-centrality VCs. Furthermore, it is known that, in general, centrality hinders the gains that a VC can obtain by operating within its region. Moreover, it is argued that institutional proximity complements the advantages of centrality.

Compared to Mingo et al. (2018), this analysis contributes to the literature by adding network cohesion to the scope and applying machine learning methods. The work can be a reference for further research in the VC and other industries, in the literature on networks and also on distances.

Predicting the odds of investing is useful not only to investors, but also for VC firms themselves. Hence, the results of this work contribute to investment strategies and policies of those firms, particularly in defining network partners and operating regions that are compatible with their internal resources, maximizing investment returns and the quality of the deals. Investment promotion agencies from target countries also benefit from this work. They can more accurately estimate which firms are more likely to invest and use this information to optimize their tactics and operations.

Our work adds to the literature about how machine learning techniques can support VC investing. Previous studies focused on predicting the performance of target companies or anticipating liquidity events. Halabi and Lussier (2014) and Lussier and Halabi (2010) applied success and failure models using logistic regression and extensions of it in different geographic markets. Nahata (2008) also implemented similar models to study how VC firm reputation leads to successful exits. Other studies used log-logistic hazard models (Holmes et al., 2010), linear regression (Hoenen et al., 2012), expert systems (Ragothaman et al., 2003) and ensemble classifiers (Wei et al., 2009). Arroyo et al. (2019) analysed the accuracy of several machine learning methods in a dataset of over 120,000 early-stage companies to predict closure or subsequent funding rounds. Other studies tested prediction models for mergers and acquisitions (Gugler & Konrad, 2002; Meador et al., 1996) and for corporate venture capital (Xu et al., 2017). Those studies essentially used variables related to education, experience, size, previous innovation and financial performance

and macroeconomic aspects. In general, the techniques are applied to predict portfolio performance, but not the strategic decision of investing in a specific emerging market.

Previous paper in the field of strategic management have discussed determinants of this decision (Hoskisson et al., 2000; Peng et al., 2008). More specific analyses focused on how the decision is affected by syndicate networks (Hochberg et al., 2007), geographic proximity (Chakrabarti & Mitchell, 2013) and institutional proximity (Li et al., 2014) between home and destination markets. Importantly, Mingo et al. (2018) analysed how the interplay of those dimensions determines investment likelihood. Hence, our work bridges the gap between extant literature on ML techniques applied to VC and extant literature about determinants of the decision of joining a VC funding round in emerging markets.

3 Data and Methodology

We gathered most of the data necessary for this analysis at Crunchbase, which provides information about deals in the VC industry. Data about funding rounds present the firms participating in each round, the target company, as well as sector and region-specific information. Funding rounds with announced dates from 2022 to 2020 are considered. Moreover, we obtained macroeconomic data from the World Bank's World Development Indicators (WDI), institutional data from the Worldwide Governance Indicators (WGI) and geographic distances from the website of Professor Kristian Skrede Gleditsch.

Figure 2 provides an overview of statistics of global funding rounds, and of funding rounds that took place in Latin America. Patterns are similar. The charts show the frequency distribution of the money raised and the number of investors per round. Most of the rounds have only one investor. The series of the amount of rounds per year clearly indicates an increase after 2012.

The United States are by far the country where more funding rounds were registered. In Latin America, Brazil has the leadership. Rounds are more frequent in software, engineering & transportation, and financial & professional services. Although several round stages are registered as unknown or missing, many of them are classified as seed investments.

The variables are qualitatively described in Table 1. We also mention previous studies that also used them. The classification problem that we address has the target binary variable *Invested* that is equal to one if the firm has participated in at least one funding round in a particular market during that year. Centrality and cohesion are features of the syndicated network that the firm is embedded. They measure, respectively, the status of the firm and how disperse the network around it is. Age measures the VC firm maturity. Geographic distances measure how close home and destination markets are. Institutional distance measures how those markets differ in terms of rule of law and regulatory quality. The macroeconomic situation of the destination market is measured by its foreign direct investments, GDP, GDP per capita, imports and market capitalization of listed domestic companies.

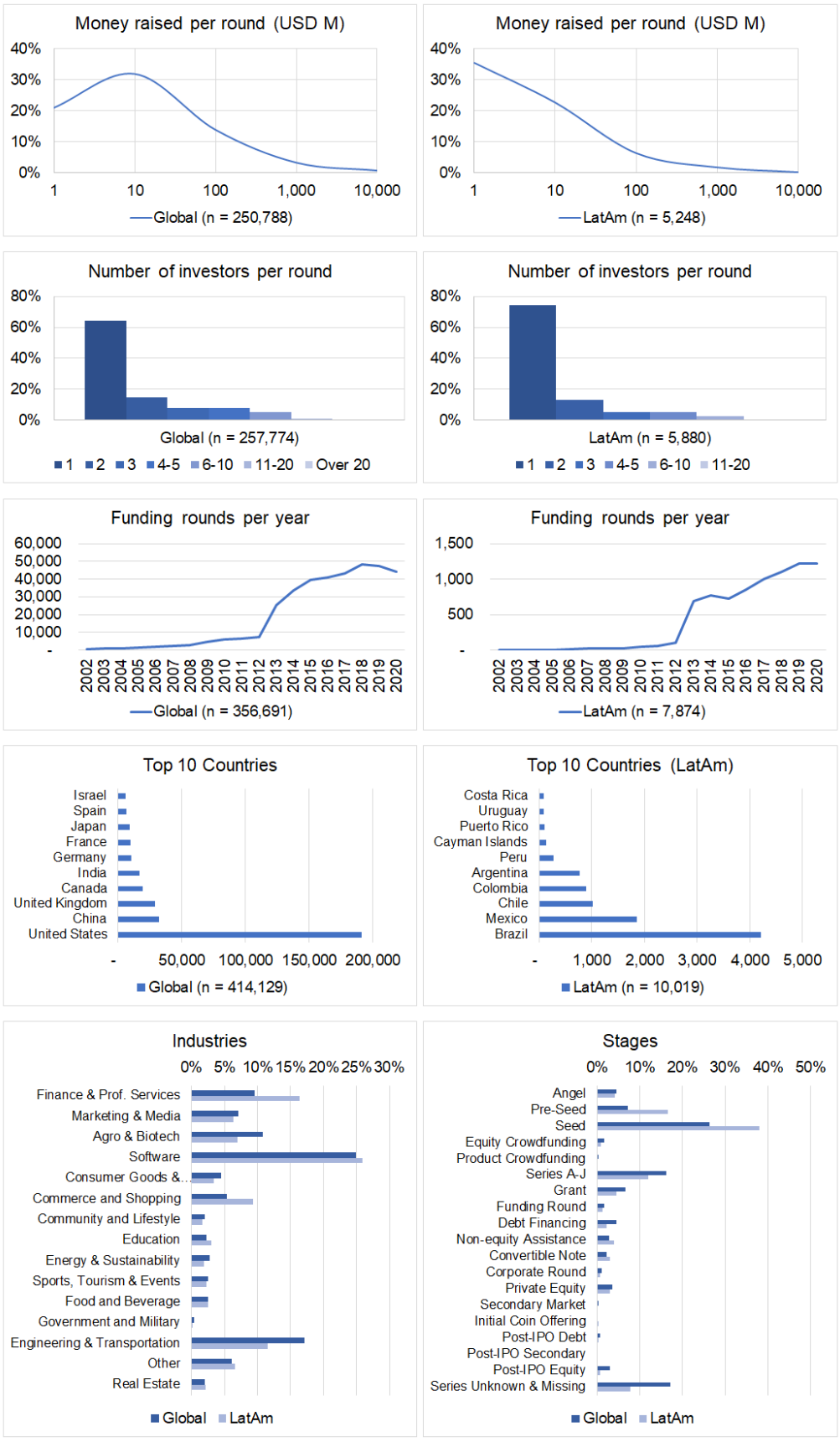

Figure 2: Charts describing the funding rounds dataset
Note: VC funding rounds cover the period from 2002 to 2020.

Table 1: Description of the variables

Variable	Description
Response	
Invested	Equal to one if the VC firm made at least one investment transaction in a particular market during a year, and zero otherwise (Mingo et al., 2018).
Syndicate network	
Centrality (status)	Eigenvector centrality score, calculated through a reciprocal process in which the centrality of each VC firm is proportional to the sum of the centralities of those firms to whom it is connected (Sorenson & Stuart, 2001).
Cohesion (vs. struct. holes)	Network constraint measure suggested by Burt (1992) and used by Bellavitis et al. (2017).
VC firm maturity	
ln(VC Age)	Log of the difference between the firm foundation year and the current year of each observation. It is commonly used as a measumre (or part of a measure) of maturity or reputation of the VC firm (Bellavitis et al., 2017; Petkova et al., 2014; Sorenson & Stuart, 2001).
National distances	
ln(Geographic)	Log of the great-circle distance (in km + 1) between the capital cities of two particular markets. (Alvarez-Garrido & Guler, 2018; Dai et al., 2012; Jääskeläinen & Maula, 2014; Mingo et al., 2018).
Institutional	Mahalanobis distance between institutional dimensions of the market of origin and of destination. As Mingo et al. (2018), the dimensions considered were regulatory quality and the rule of law.
Macroeconomic	
FDI as a % of GDP	Foreign direct investments (net inflows) as a percentage of GDP.
ln(GDP)	Log of the gross domestic product of the country, in current USD.
GDP per capita	Gross domestic product, in current USD, divided by the population.
Imports as a % of GDP	Total imports of goods and services, as a percentage of GDP.
Mkt cap as a % of GDP	Capitalization of the listed domestic companies in the country, as a percentage of GDP.
VC past behavior	
Invested in country (last 3Y)	Equal to one if the firm has invested in the country in the last 3 years.
No previous inv. in region	Equal to one if the firm has made no previous investments in the region during all previous years of the data set.
Invested in region (last 3Y)	Equal to one if the firm has invested in the region in the last 3 years.
Invested in region (last 4-5Y)	Equal to one if the firm has invested in the region between 4 and 5 years ago.
Inv. in impact in country (last 3Y)	Equal to one if the firm has made impact investments in the country in the last 3 years.

Note: VC, GDP, FDI, Mkt Cap and Y stand for venture capital, gross domestic product, foreign direct investment, market capitalization and year, respectively.

The eigenvector centrality score x_v of vertex v can be defined as: $x_v = \frac{1}{\lambda}\sum_{t\in M(v)} x_t$, where $M(v)$ is the set of neighbors of v, and λ is the greatest eigenvalue of the adjacency matrix.

Overall cohesion in each node is calculated as the sum of the cohesiveness (c_{ij}) of all relationships: $c_{ij} = (p_{ij} + \sum_{q\neq i\neq j} p_{iq}p_{qi})^2$, with the focal factor i having two contacts q and j. Hence, p_{ij} reflects the interaction between i and j while p_{qi} defines whether q and j are related.

Using the funding rounds database, we built the sample with 112,769 firm-country-year observations. It results from 8 countries, around 1,400 firms and 19 years, but it accounts for the fact that the average firm age is around 9 years, half of the sample time interval. Table 2 presents the descriptive statistics of the variables. Investments occurred only in 3% of the firm-year-country combinations, which represent around 3 thousand observations, with VC firms from 67 distinct countries. Those statistics clearly indicate imbalanced data.

Average cohesion is 0.12 in the full sample and 0.18 when investments occur, suggesting a slightly positive relationship between cohesiveness and investment likelihood. However, we should take into account that less active companies have low or null cohesiveness. After removing those cases, average cohesiveness becomes 0.31 in the full sample and 0.29 when investments occur.

Average measures suggest that centrality scores are higher when investments occur (0.02 vs. 0.07). VC foundation years range from 1903 to 2019. As expected, the past behavior of the VC firm seems to determine investment likelihood. Table shows that in 61% of the combinations the firm has never performed an investment over the region. This figure reduces to 35% when investments occur, and in 48% of the cases it occurred over the last three years.

Table 2: Descriptive statistics of the firm-year-country dataset

	LatAm full sample (112,769 obs)				**Only Invested = 1 (3,135 obs.)**			
	Mean	**SD**	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**
Response								
Invested	0.03	0.16	-	1.00	1.00	-	1.00	1.00
Syndicate network								
Centrality (status)	0.02	0.08	-	1.00	0.07	0.17	-	1.00
Cohesion (vs. struct. holes)	0.12	0.28	-	1.13	0.18	0.30	-	1.13
VC firm maturity								
Foundation Year	1997	21.27	1903	2019	2004	17.67	1903	2019
ln(VC Age)	2.20	1.15	-	4.76	1.92	1.11	-	4.76
National distances								
ln(Geographic)	8.41	1.55	-	9.89	6.44	3.83	-	9.87
Institutional	1.81	0.89	-	4.40	1.47	1.08	-	3.62
Macroeconomic								
FDI as a % of GDP	4.31	2.66	−3.99	16.23	3.34	1.31	−3.99	11.74
ln(GDP)	26.25	1.31	23.29	28.59	27.47	0.90	23.57	28.59
GDP per capita	9,430.77	3,501.96	2,021.24	15,888.14	9,719.65	2,605.60	3,349.81	15,888.14
Imports as a % of GDP	29.01	15.32	11.25	88.61	22.81	10.40	11.25	78.23
Mkt cap as a % of GDP	39.57	28.76	3.01	156.40	43.43	20.55	3.01	156.40
VC past behavior								
Invested in country (L3Y)	0.04	0.20	-	1.00	0.48	0.50	-	1.00
No previous inv. in region	0.61	0.49	-	1.00	0.35	0.48	-	1.00
Invested in region (L3Y)	0.28	0.45	-	1.00	0.62	0.48	-	1.00
Invested in region (L4-5Y)	0.05	0.23	-	1.00	0.02	0.13	-	1.00

Note: VC, GDP, FDI, Mkt Cap, L3Y and L4-5Y stand for venture capital, gross domestic product, foreign direct investment, market capitalization, last three years and last four to five years, respectively.

Table 3 presents the correlation among the variables. Most of the correlations are negligible. Cohesion and centrality correlate by 14%. As expected, institutional and geographic distance are positively correlated (0.31). Imports correlate with foreign direct investments by 62% but correlate with GDP by −62%. Previous investments in a market correlate

with geographic proximity (23%) but not so much with institutional proximity (8%).

Table 3: Correlation among predictors

	(01)	(02)	(03)	(04)	(05)	(06)	(07)	(08)	(09)	(10)	(11)	(12)	(13)
(01) Network Centrality	1												
(02) Network Cohesion	0.14	1											
(03) ln(VC Age)	−0.02	−0.01	1										
(04) ln(Geographic dist.)	−0.01	−0.02	0.03	1									
(05) Institutional dist.	0.00	−0.04	0.04	0.31	1								
(06) FDI as a % of GDP	−0.03	−0.09	0.00	0.02	−0.19	1							
(07) ln(GDP)	0.01	0.04	−0.02	−0.11	0.16	−0.48	1						
(08) GDP per capita	0.02	0.10	−0.06	0.02	−0.31	0.13	0.08	1					
(09) Imports as a % of GDP	−0.01	−0.04	0.01	0.01	−0.06	0.62	−0.62	0.12	1				
(10) Mkt cap as a % of GDP	−0.01	−0.03	0.01	0.02	−0.14	0.19	0.23	0.08	−0.10	1			
(11) Invested in country (L3Y)	0.16	0.15	−0.02	−0.23	−0.08	−0.08	0.19	0.01	−0.08	0.02	1		
(12) No previous inv. in region	−0.24	−0.54	−0.04	0.02	0.05	0.17	−0.08	−0.18	0.07	0.05	−0.26	1	
(13) Invested in region (L3Y)	0.26	0.53	−0.06	−0.04	−0.06	−0.13	0.07	0.16	−0.06-	−0.05	0.34	−0.78	1
(14) Invested in region (L4-5Y)	0.04	0.21	0.06	0.00	−0.02	−0.07	0.02	0.04	−0.02	−0.01	−0.05	−0.30	−0.15

Note: Pearson's correlation coefficient of the research covariates. VC, GDP, FDI, Mkt Cap, L3Y and L4-5Y stand for venture capital, gross domestic product, foreign direct investment, market capitalization, last three years and last four to five years, respectively.

4 Selection and Setup of the Models

We trained and tested 11 techniques that are commonly adopted for classification problems: Logistic Regression (Logit), Logistic Lasso (LogLasso), Logistic Ridge (LogRidge), Naive Bayes (NB), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Gradient Boosting (Boost) and XGBoost (XGB). Some models are simpler and more easily interpreted, such as Logit and DT, while others (particularly ensemble models such as XGB, RF and Boost) are more sophisticated and may be capable of identifying non-linear patterns that may exist in the case, while potentially reducing bias. We developed and treated the database in R Studio and applied the techniques in Jupyter Notebook running Python code.

The equation 1 describes the model considered in the application of one of the techniques, the logit model (first order multiple logistic regression).

$$logit(Invested) = \alpha + \beta Maturity + \sum_{i=1}^{2} \rho_i NetworkVariable + \sum_{i=1}^{2} \theta_i DistanceVariable_i + \\ + \sum_{i=1}^{5} \lambda_i MacroeconomicVariable_i + \sum_{i=1}^{4} \gamma_i PastBehaviorVariable_i \tag{1}$$

We trained the models using data from 2002 to 2017 and tested them using data from 2018 to 2020. Some of the hyperparameters were defined through validation and optimization processes. For instance, we used cross-validation to verify whether higher orders of the centrality variable should be considered in the model equation and it has shown no significant improvement in error rates. Moreover, we obtained the number of neighbours that resulted in the best accuracy for the KNN model in each training. The parameters

selection for the Logistic Lasso was made by checking the bast mean absolute error. The Random Forest was developed with six variables sampled at each split. Boosting was performed using 5,000 trees, with a depth of four. XGBoost was tuned with a maximum depth of two, a learning parameter of one and two rounds. For the other hyperparameters and models, we considered the default values of the implementation packages.

5 Outputs and Performance Assessment

Table 4 presents performance metrics for each technique. It shows accuracy, precision, recall, F1 score and AUC. The default threshold is 50%. The techniques are ordered by AUC. Gradient boosting presented the higher AUC (85%), compared to 50% of a dummy model that randomly classifies observations according to previous base rates.

From the point of view of an investor that wants to invest in a particular market or industry, this is a problem in which a type I error (predict that the firm will invest but then it does not) is less acceptable than a type II error (predict that a firm will not invest but then it does). Hence, precision is more important than recall, which favors using techniques such as Gradient Boosting and Logit over LDA, QDA and Naive Bayes, even though the latter ones presented higher F1 scores than the former ones.

Table 4: Performance metrics for each model

Technique	Accuracy	Precision	Recall	F1	AUC
Gradient Boosting	0.952	0.535	0.210	0.301	0.853
Logit	0.951	0.524	0.139	0.220	0.843
LogLasso	0.952	0.526	0.140	0.221	0.843
LogRidge	0.952	0.525	0.139	0.220	0.843
LDA	0.919	0.308	0.523	0.388	0.840
QDA	0.907	0.289	0.610	0.392	0.839
XGBoost	0.951	0.511	0.228	0.315	0.829
Random Forest	0.951	0.505	0.201	0.287	0.812
Naive Bayes	0.863	0.208	0.636	0.313	0.797
KNN	0.949	0.468	0.228	0.306	0.720
Decision Tree	0.927	0.270	0.284	0.277	0.635
Dummy	0.951	-	-	-	0.500

Note: Performance indicators of selected machine learning techniques to predict whether a VC firm will invest in a Latin American country. Accuracy is the percentage of correctly predicted cases. Prediction is the ratio of true positive cases to all cases predicted as positive. Recall is the ratio of true positive cases to all actually positive cases. The F1 score is the harmonic mean of the precision and recall. AUC is the area under the ROC (receiver operating characteristic) curve, which relates true positive with false positive rates. The sample period is from 2002 to 2020, with the last three years as the test set.

Figure 3 presents the ROC curves for each model. Areas under the curves are shown beside the name of each technique. The curves confirm the good performance of logit and ensemble techniques when compared to naive bayes, decision tree and KNN. The dotted line represents the dummy technique.

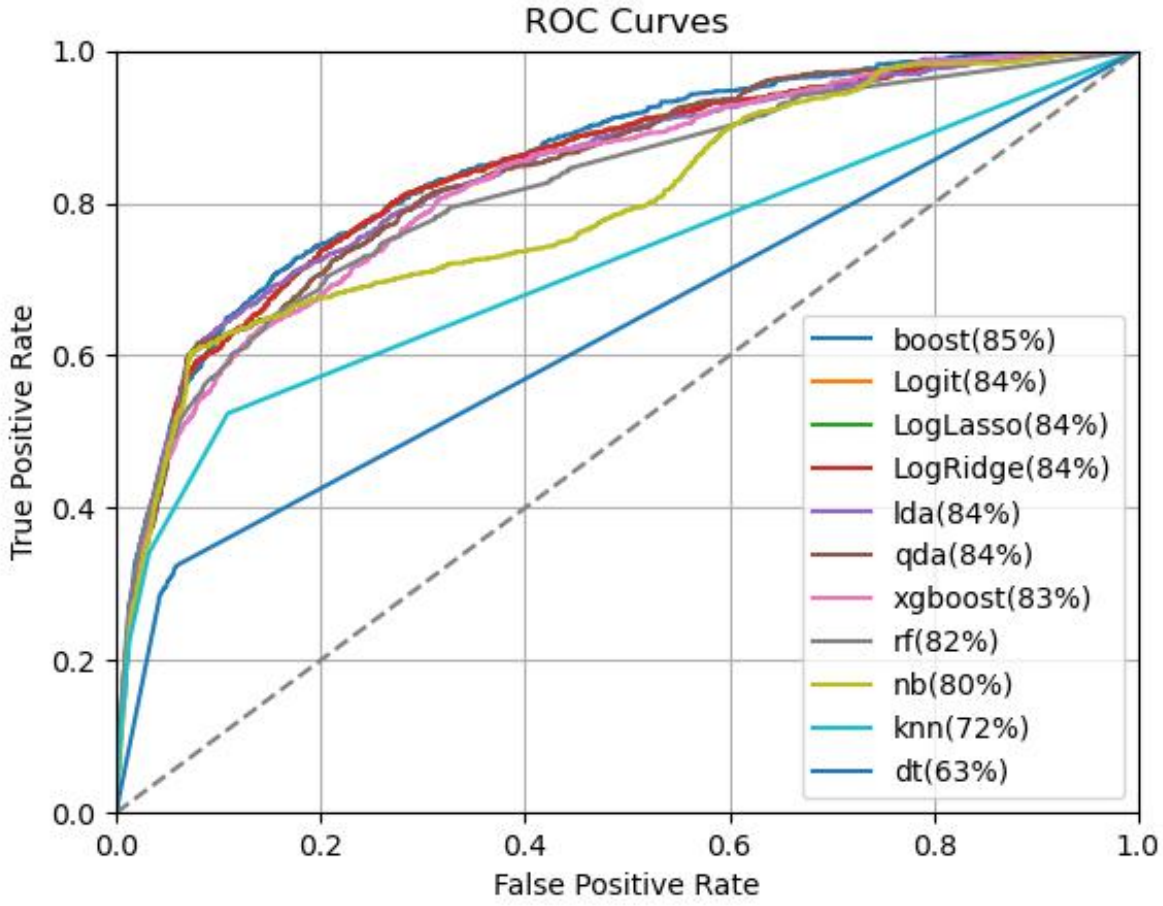

Figure 3: Plot of the resulting ROC curve for each technique

6 Final Considerations

In this study, 11 machine learning models were trained to predict whether a venture capital firm will invest in each Latin American country during a year. The predictors were characteristics of the firm and the destination market, as well as variables associated with the past investment activities of the firm, the distances between the firm headquarters and the destination market, and the network structure in which the VC firm is embedded. We used a sample of funding rounds from 2002 to 2020, training the models with the first 16 years and testing them with the last 3 years.

Despite the challenging task of dealing with very imbalanced datasets, ensemble and logit models presented reasonable performance in terms of precision and areas under the ROC curve. The Gradient Boosting technique, particularly, presented a precision of 53% and an AUC of 85%. The gain in performance is significant when we compare with simply selecting those firms that invested in that country in the last three years, which would present a precision of only 29%. The application of ML techniques can avoid the opportunity cost of allocating resources in a fund managed by a VC firm that will not invest in the markets or industries that are expected by the investor. Assembling data covering operations of 38 international banks from 1995 to 2004, Garcia-Herrero and Vazquez (2013) evidenced that international diversification gains are large and unexploited, with risk-minimizing allocations resulting in average return on assets at least 30% higher.

Moreover, it can result in savings in marketing or operation costs from investment promotion agencies responsible for fostering the development of the local VC industry. As Charlton and Davis (2007) state, investment promotion became an active area of public policy, offering services and incentives to attract investment by foreign firms. Based on a study of 58 investment promotion agencies, Morisset and Andrews-Johnson (2004) showed that, on average, agencies from emerging markets pro-actively contact 1,395 investors per year. Assuming that conversion rates of contacting more inclined investors will be higher than the ones of contacting less inclined ones, implementing an appropriate ML technique, with a precision 24 percentage points higher, would result in around 46% in savings. Morisset and Andrews-Johnson (2004) also estimated that a 10% increase in promotion expenses generates a 2.5% increase in foreign investments inflow (or 7.5% for a promotion budget between USD 2 million and USD 11 million). Assuming the same marginal gain for efficiency increase (same costs but higher output), implementing a suitable ML technique would contribute to a significant increase in FDI.

Future studies can adopt other techniques, such as neural networks, and analyse their performance. Particularly, graph neural networks might be a prominent option in this case, since syndicate network features have a notable role. Similar analyses might be performed for other regions or aggregating VC transactions all over the world. Another venue for future studies is a better understanding of causal mechanisms and the role of interaction terms to determine investment likelihood and performance.

References

Alvarez-Garrido, E., & Guler, I. (2018). Status in a strange land? Context-dependent value of status in cross-border venture capital. *Strategic Management Journal*, *39*(7), 1887–1911 (cit. on p. 8).

Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of machine learning performance for decision support in venture capital investments. *Ieee Access*, *7*, 124233–124243 (cit. on p. 5).

Bellavitis, C., Filatotchev, I., & Souitaris, V. (2017). The impact of investment networks on venture capital firm performance: A contingency framework. *British Journal of Management*, *28*(1), 102–119 (cit. on pp. 3, 4, 8).

Burt, R. S. (1992). Structural holes. In *Structural holes*. Harvard university press. (Cit. on p. 8).

Castellanos, D. (2023). Startups' mortality rate seen rising as venture capital flows tighten. *Bloomberg Línea*. www.bloomberglinea.com/english/startups-mortality-rate-seen-rising-as-venture-capital-flows-tighten (cit. on p. 2)

Chakrabarti, A., & Mitchell, W. (2013). The persistent effect of geographic distance in acquisition target selection. *Organization Science*, *24*(6), 1805–1826 (cit. on p. 6).

Charlton, A., & Davis, N. (2007). Does investment promotion work? *The BE Journal of Economic Analysis & Policy*, *7*(1) (cit. on p. 13).

Dai, N., Jo, H., & Kassicieh, S. (2012). Cross-border venture capital investments in Asia: Selection and exit performance. *Journal of Business Venturing*, *27*(6), 666–684 (cit. on p. 8).

Florida, R. L., & Kenney, M. (1988). Venture capital, high technology and regional development. *Regional Studies*, *22*(1), 33–48 (cit. on p. 1).

Garcia-Herrero, A., & Vazquez, F. (2013). International diversification gains and home bias in banking. *Journal of Banking & Finance*, *37*(7), 2560–2571 (cit. on p. 13).

Gugler, K., & Konrad, K. A. (2002). Merger target selection and financial structure. *University of Vienna and Wissenschaftszentrum Berlin (WZB)* (cit. on p. 5).

Halabi, C. E., & Lussier, R. N. (2014). A model for predicting small firm performance: Increasing the probability of entrepreneurial success in chile. *Journal of Small Business and Enterprise Development*, *21*(1), 4–25 (cit. on p. 5).

Hochberg, Y. V., Ljungqvist, A., & Lu, Y. (2007). Whom you know matters: Venture capital networks and investment performance. *The Journal of Finance*, *62*(1), 251–301 (cit. on p. 6).

Hoenen, S., Kolympiris, C., Schoenmakers, W., & Kalaitzandonakes, N. (2012). Do patents increase venture capital investments between rounds of financing (cit. on p. 5).

Holmes, P., Hunt, A., & Stone, I. (2010). An analysis of new firm survival using a hazard function. *Applied Economics*, *42*(2), 185–195 (cit. on p. 5).

Hoskisson, R. E., Eden, L., Lau, C. M., & Wright, M. (2000). Strategy in emerging economies. *Academy of management journal*, *43*(3), 249–267 (cit. on p. 6).

Jääskeläinen, M., & Maula, M. (2014). Do networks of financial intermediaries help reduce local bias? Evidence from cross-border venture capital exits. *Journal of Business Venturing*, *29*(5), 704–721 (cit. on p. 8).

Jelic, R., Saadouni, B., & Wright, M. (2005). Performance of private to public MBOs: The role of venture capital. *Journal of Business Finance & Accounting*, *32*(3-4), 643–682 (cit. on p. 1).

Kanuri, S., & Hanby, M. (2020). Private equity (PE) performance in the United States. *Journal of Applied Business and Economics*, *22*(1), 36–45 (cit. on p. 2).

Kaplan, S. N., & Strömberg, P. (2009). Leveraged buyouts and private equity. *Journal of economic perspectives*, *23*(1), 121–146 (cit. on p. 2).

Lerner, J. (1994). The syndication of venture capital investments. *Financial management*, 16–27 (cit. on p. 3).

Li, Y., Vertinsky, I. B., & Li, J. (2014). National distances, international experience, and venture capital investment performance. *Journal of Business Venturing*, *29*(4), 471–489 (cit. on p. 6).

Lussier, R. N., & Halabi, C. E. (2010). A three-country comparison of the business success versus failure prediction model. *Journal of Small Business Management*, *48*(3), 360–377 (cit. on p. 5).

Meador, A. L., Church, P. H., & Rayburn, L. G. (1996). Development of prediction models for horizontal and vertical mergers. *Journal of financial and strategic decisions*, *9*(1), 11–23 (cit. on p. 5).

Minardi, A. M. A. F., Bassani, R., Kanitz, R., Moreira Neto, J. C., & Pechlyie, K. (2015). Private equity and venture capital investments in Brazilian companies in the last 30 years. *Available at SSRN 2600355* (cit. on p. 2).

Mingo, S., Morales, F., & Dau, L. A. (2018). The interplay of national distances and regional networks: Private equity investments in emerging markets. *Journal of International Business Studies*, *49*(3), 371–386 (cit. on pp. 2, 4–6, 8).

Morisset, J., & Andrews-Johnson, K. (2004). *The effectiveness of promotion agencies at attracting foreign direct investment* (Vol. 16). World Bank Publications. (Cit. on p. 13).

Nahata, R. (2008). Venture capital reputation and investment performance. *Journal of financial economics*, *90*(2), 127–151 (cit. on p. 5).

Nolting, C., Višić, I., & Singh, S. (2021). Venture capital trends. *Deutsche Wealth*. www.deutschewealth.com/en/insights/investing-insights/asset-class-insights/venture-capital-investing-closer-look/venture-capital-trends.html (cit. on p. 2)

Peng, M. W., Wang, D. Y., & Jiang, Y. (2008). An institution-based view of international business strategy: A focus on emerging economies. *Journal of international business studies*, *39*, 920–936 (cit. on p. 6).

Petkova, A. P., Wadhwa, A., Yao, X., & Jain, S. (2014). Reputation and decision making under ambiguity: A study of US venture capital firms' investments in the emerging clean energy sector. *Academy of Management Journal*, *57*(2), 422–448 (cit. on p. 8).

Phalippou, L. (2020). An inconvenient fact: Private equity returns and the billionaire factory. *The Journal of Investing*, *30*(1), 11–39 (cit. on p. 2).

Ragothaman, S., Naik, B., & Ramakrishnan, K. (2003). Predicting corporate acquisitions: An application of uncertain reasoning using rule induction. *Information Systems Frontiers*, *5*, 401–412 (cit. on p. 5).

Rider, C. I., & Swaminathan, A. (2012). They just fade away: Mortality in the US venture capital industry. *Industrial and Corporate Change*, *21*(1), 151–185 (cit. on p. 2).

Ruhnka, J. C., & Young, J. E. (1991). Some hypotheses about risk in venture capital investing. *Journal of business venturing*, *6*(2), 115–133 (cit. on p. 2).

Silva, S., Badi, M., & Thomas, B. (2022). The surge of venture capital in Latin America. *The Lauder Global Business Insight Report*. lauder.wharton.upenn.edu/wp-content/uploads/2022/02/Venture-Capital-in-Latin-America-GBIR2022.pdf (cit. on p. 3)

Sorenson, O., & Stuart, T. E. (2001). Syndication networks and the spatial distribution of venture capital investments. *American journal of sociology*, *106*(6), 1546–1588 (cit. on p. 8).

Stein, E., & Wagner, R. A. (2018). The development of venture capital in Latin America: A comparative perspective. *Available at SSRN 3268085* (cit. on p. 3).

Wei, C.-P., Jiang, Y.-S., & Yang, C.-S. (2009). Patent analysis for supporting merger and acquisition prediction: A data mining approach. *Designing E-Business Systems. Markets, Services, and Networks: 7th Workshop on E-Business, WEB 2008, Paris, France, December 13, 2008, Revised Selected Papers 7*, 187–200 (cit. on p. 5).

Wright, M., & Lockett, A. (2003). The structure and management of alliances: Syndication in the venture capital industry. *Journal of management studies*, *40*(8), 2073–2102 (cit. on p. 3).

Wright Robbie, K., Mike. (1998). Venture capital and private equity: A review and synthesis. *Journal of Business Finance & Accounting*, *25*(5-6), 521–570 (cit. on p. 1).

Xu, R., Chen, H., & Zhao, L. J. (2017). Predicting corporate venture capital investment (cit. on p. 5).