A Machine Learning Approach to Entrepreneurial Finance Modelling

Max Berre

Abstract Traditionally, estimating valuation relies on firm data and concrete economic indicators. So does modelling of startup investment selection and startup survivability. However, recent advancements in machine learning have given rise to customizable segmented-modelling approaches. While classical economic theory describes that firm valuations and survival rates are modelled based on revenues, growth rates, and risk, the valuation of startup often proves the exception to the rule. Meanwhile both startup investor selection and startup valuations are influenced by revenues, risks, age, and macroeconomic conditions, specific causality is traditionally a black box. Likewise, for startup survivability, which is known to be influenced by risks, revenues, age-of-firm, and access to finance, specific causality is also unclear. Because details are not disclosed, roles played by other factors (industry, business models, geography, and intellectual property) can often only be guessed at. This study is an in-depth examination outlining methods and approaches for application of segmented modelling in entrepreneurial finance, as well as ways in which they can be applied using existing data for purposes to examine selection, valuation, and survivability.

Keywords CART · Decision tree · Valuation · Startup valuation · Startup selection · Investment selection · Startup survival startup survivability · Venture capital · Entrepreneurial finance · Machine learning · Hierarchical analysis

M. Berre (✉)
Audencia Business School, Nantes, France
e-mail: mberre@audencia.com

Nyenrode Business University, Amsterdam, Netherlands

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

C. Zopounidis et al. (eds.), *Operational Research Methods in Business, Finance and Economics*, Lecture Notes in Operations Research,
https://doi.org/10.1007/978-3-031-31241-0_2

1 Introduction

Why are startup in Massachusetts-based startup substantially more likely to survive into their fourth year than those in New Hampshire? Why do startup in London attract higher valuations than those in Paris, Berlin, or Milan, even when they are based in similarly sized economies, share the same industries and many of the same investors? Why do healthcare-industry startup work their way through the VC-selection deal funnel more efficiently than IT-industry startup?

Whereas classical economic theory describes that firm valuations and survival rates are modelled based on revenues, growth rates, and risk, valuation of startup often proves the exception to the rule with startup differing significantly in terms of information environment, time structure of transactions, and linkages between investors and investees (Bellavitis et al. 2017). Given their specific characteristics, startup are notorious for being difficult to value, while their investment selection is difficult to predict, and their survivability can vary dramatically across industry (Damodaran, 2009) and across geography (Gonzalez, 2017). These difficulties are driven by opacity of both startup and investors, as well as short histories, and complex intangible assets held by startup (Damodaran, 2009), as well as disruption potential (Damodaran, 2019).

Overall, this is intended how to guide outlining approaches and methods for application of segmented hierarchical modelling in entrepreneurial finance, as well as how they can be applied using existing data and regression models.

The rest of this study proceeds as follows: the subsequent section describes segmented models, while also describing where they have appeared in both practitioner-focused grey literature, as well as in peer-review literature. After this, Sect. 3 describes how segmented models can be made hierarchical, as well as describing how they can be used for microtargeting-based approaches. Lastly, the discussion and conclusion section outlines why segmented, hierarchical, and microtargeting approaches are used by industry practitioners, by describing their added-value vis-à-vis more traditional approaches.

2 Why Segmented Models: What Do We Aggregate?

A relatively widespread theoretical approach used typically for both startup selection and startup valuation is that of the scorecard-based approach. Given that Fama (1970) describes factors as information-subsets which have the potential to drive price-signals, and which can range from historical-values to disclosures and privileged-information, their incorporation may be highly relevant. The primary advantage of the scorecard approach is the ability to incorporate qualitative, geographic, sectoral, or categorical determinants in several ways. In entrepreneurial finance, these factors might range from non-financial and deal characteristics prevalent in given sectoral or municipal ecosystems, the role of national-level or market-condition determinants, to the role that business models or ownership structures and legal form may play in survivorship or investment selection. This approach can be used to estimate either valuation or selection ranking and is capable of establishing insights even as detailed related economic and financial information is missing, scare, or unevenly available.

Segmented models are modular and relatively straightforward model approaches based on summation of market conditions, key characteristics, and deal conditions developed primarily by industry practitioners. One critical advantage of this sort of model approach is that valuation, selection, or survival probability can be modelled, captured, and understood while including specific categorical information, which can be both general, highly specific, and/or be organized as joint, combined, or hierarchical segmentation.

Mechanically speaking, reliance on optimal arrangement of multiple decision factors is central to model functionality. In parallel, multiple-criteria decision analysis (MCDA), which explicitly evaluates conflicting-criteria in decision-making is described Zopounidis et al. (2015) as being used for portfolio and investment evaluation and selection, is usually implemented in terms of fundamental factors. While valuation factors do not necessarily conflict, valuation impacts of trade-offs and fault lines may constitute important model elements which need to be taken into consideration.

In a similar vein, Zopounidis and Doumpos (2002), who examine and model investment selection, describe that choices of investment project are strategic decisions made by human agents (e.g. financial managers or venture capitalists) and not by the model; the decision makers become more and more deeply involved in the decision-making process. Citing this, Dhochak and Doliya (2019) outline apply fuzzy analytic hierarchy process to startup valuation, claiming that fuzzy AHP is a well-suited methodology to evaluate startup valuation due to close resemblance to cognitive human decision-making approaches.

Ellis et al. (2001) meanwhile make use of a segmented multi-criteria modelling approach in order to model supplier success in meeting customer expectations in the high-technology marine equipment industry. To demonstrate divergent views between shipbuilders and shipowners in both the European and Japanese markets. Figure 1 outlines ranked differences in supplier success factors, finding that Japan's shipbuilder market places higher importance on prices than do European shipbuilder markets, while Japan's shipowner market places lower importance on maintenance than do European shipowner markets.

2.1 *Practitioners: Segmented Models in Markets*

In industry, scorecard approaches are typically used by business angels. Scorecard valuation approaches which have emerged from industry prominently include Payne (2011) and Berkus (2016). Perhaps the most well-known segmented startup valuation model is the scorecard model, outlined by Payne (2011). Outlined in Table 1,

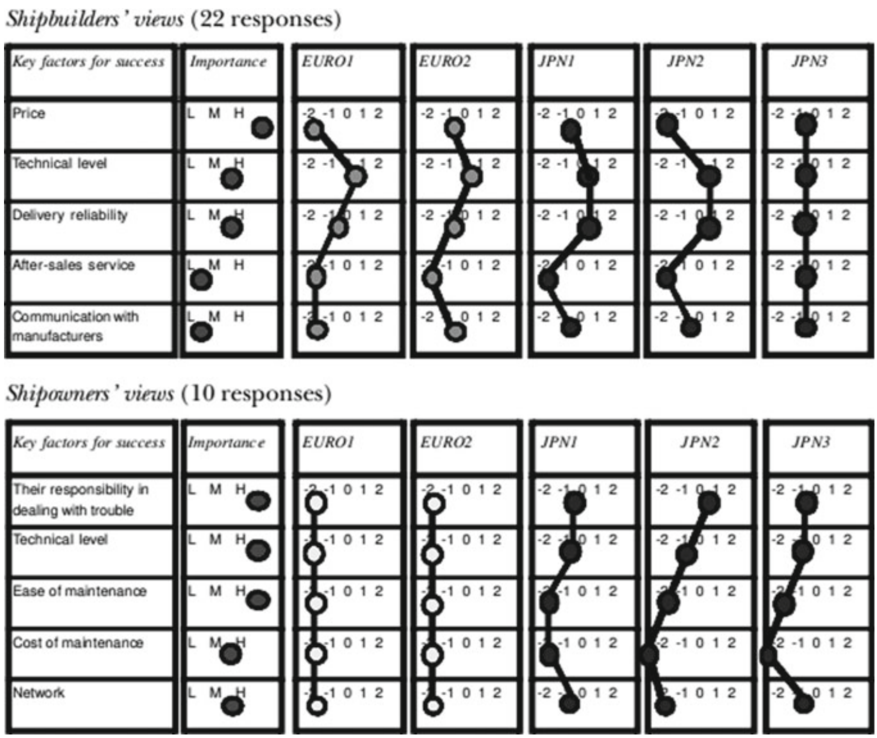


Fig. 1 Differentiated segmental positioning modelling supplier success in marine high-tech equipment. *Source* Ellis et al. (2001)

Payne's scorecard model segments valuation into management team, target market, competitive environment, and further funding need.

Focusing specifically on valuation, a well-known alternative to the scorecard model is the Berkus model (Ernst & Young, 2020). Outlined in Fig. 2, the Berkus model segments valuation into component risks.

2.2 *Segmented Models in Peer Review Literature*

Meanwhile, in published economic literature, this same concept emerges as summation-based valuation models. Prominent examples include models published by Hand (2005), Miloud and Cabrol (2011), and Sievers et al. (2013). For example, Eq. 1 describes Hand (2005)'s startup-valuation model, which is driven by deterministic valuation factors segmented into financial-statement data such as assets, Net Income, and cash flows, on one hand, and operational and industry-related data on the other.

Table 1 Abbreviated Payne scorecard model

Weighting		Impact on startup selection and valuation
0–30%	Impact	*Strength of the entrepreneur and the management team*
	+	Many years of business experience
	++	Experience in this business sector
	+++	Experience as a CEO
	++	Experience as a CFO, COO, or CTO
	+	Experience as a product manager
	–	Experience in sales or technology
	––	No business experience
0–25%	Impact	*Size of the opportunity*
		Size of the target market (total sales)
	––	< $50 million
	+	$100 million
	++	> $100 million impact
		Potential for revenues of target company in five years
	––	< $20 million
	++	$20–$50 million to > $100 million (will require significant additional funding)
0–15%	Impact	*Strength of products and intellectual property*
	–––	Not well defined, still looking for prototypes
	0	Well defined, prototype looks interesting
	++	Good feedback from potential customers
	+++	Customer orders or early sales
0–10%	Impact	*Competitive environment*
		Strength of competitors in this marketplace
	––	Dominated by a single large player
	–	Dominated by several players
	++	Fractured, many small players
		Strength of competitive products
	––	Competitive products are substantial
	++	Competitive products are weak
0–10%	Impact	*Marketing/sales/partners*
		Impact sales channels, sales, and marketing partners
	–––	Have not discussed sales channels
	++	Key beta testers identified and contacted
	+++	Channels secure, customers placed trial orders
	––	No partners identified

(continued)

Table 1 (continued)

Weighting		Impact on startup selection and valuation
	++	Key partners in place
0–5%	Impact	*Need for additional rounds of funding*
	+++	None
	0	Additional angel round
	−−	Need venture capital

Source Ernst and Young (2020)

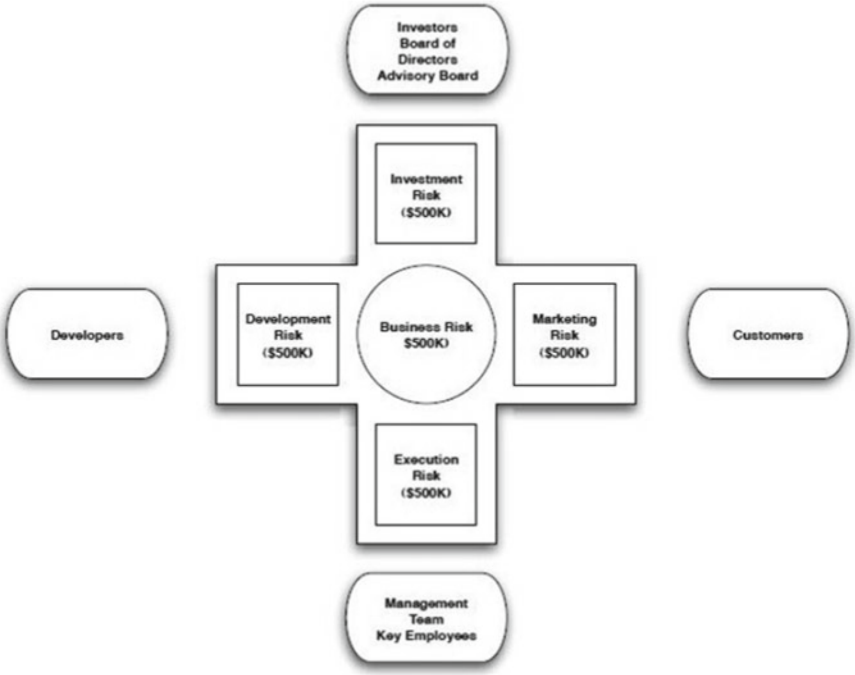


Fig. 2 Berkus model for startup valuation. *Source* Berkus (2016)

Equation 1 Hand (2005) Summation-based segmented valuation model

$$\text{Hand (2005)}\ \text{Ln(Pre-Money Valuation)} = \sum \theta_{\text{b}}\text{Ln(Financial Statement Data}_{bik}) + \sum \Upsilon_{\text{c}}\text{Ln(NonFinancial Statement information}_{cik}) + \varepsilon_{ik} \tag{1}$$

Meanwhile, Eq. 2, another prominent segmented startup-valuation model outlines the Sievers et al. (2013) summation-based valuation model, describing valuation on the basis of summation of financial, and non-financial firm attributes, as well as deal characteristics and relevant valuation coefficients. Essentially, whereas Hand (2005) segments valuation factors into accounting and non-accounting data, Sievers et al. (2013) segment valuation factors into financial factors such as revenues, risks, or capital invested, and non-financial factors including operational and industry-level data, and deal characteristics such as investor syndication, and investment-deal clauses such as redemption, tag-along, and ratchet clauses in the investment deal.

Equation 2 Sievers et al. (2013) Summation-based segmented valuation model

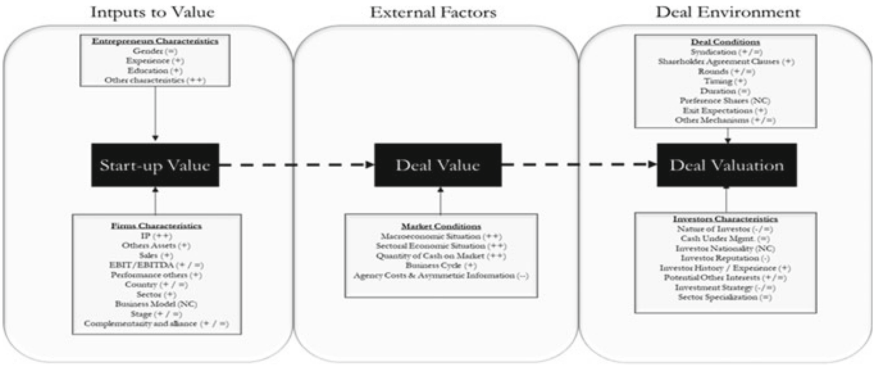
$$\begin{aligned}\log(\text{Valuation}_{it}) = \sum \Phi\text{Non-financial}_{it} + \sum \Delta\text{Financial}_{it} \\ + \sum \Psi\text{Deal Characteristics}_{it}\end{aligned} \tag{2}$$

Sievers' model estimates valuation by summation of the established segments, as is the case with the Berkus and Payne models. Overlooked, however, are interactions and hierarchies among valuation determinants.

Mechanically speaking, an alternate functional form to express segmented valuation can be elaborated via the staged valuation approach. Examples of this approach include the Startup Valuation Meta-Model developed by Berre and Le Pendeven (2022) outlined in Eq. 3. In addition to valuation factors themselves, this approach accounts for phases, interactions, and hierarchies among valuation factors. Formally:

Equation 3 Berre–Le Pendeven (2022) Valuation meta-model for startup

$$\text{Pre-Money Valuation} = f\left(\left(\left(\sum \text{Start-Up Value}\right)\sum \text{Deal Value}\right)\sum \text{Deal Valuation}\right) \tag{3}$$



3 From Segmentation to Hierarchical Microtargeting Models

Recently, the emergence and development of supervised machine learning techniques has led to increasing methodological sophistication of scorecard approaches, as predictive techniques incorporating categorical, qualitative, geo-spatial, and ordinal data have become increasingly widespread.

Mechanically speaking, microtargeting by means of datamining is described in detail by Murray and Scime (2010), as the process of inductively analysing data to find actionable patterns, fault lines, and relationships within the data, on the basis of trends drawn from both numerical and descriptive characteristics, such as average family age, family composition, and geographic area, via construction of decision trees, an analytical technique which is both explanatory and predictive, and which is used for both variable predictions, as well as to provide specific insights concerning structure, segmentation, and interrelationships among data.

This approach grants insight into how specifically any outcome variable's value is dependent on the model's deterministic factors, with each identifiable fault line constituting segments of individuals. Fundamentally, microtargeting by means of data mining can allow scorecard-based modelling approaches to incorporate qualitative and categorical data hierarchically, to a potentially extreme degree of detail.

Functionally speaking, a hierarchically structured model resulting from a micro-targeting approach can be expressed via a staged model approach. For instance, Fig. 3 displays the architectural form that the Berre–Le Pendeven Startup Valuation Meta-Model described in Eq. 2 would adopt, expressed as a hierarchical decision tree.

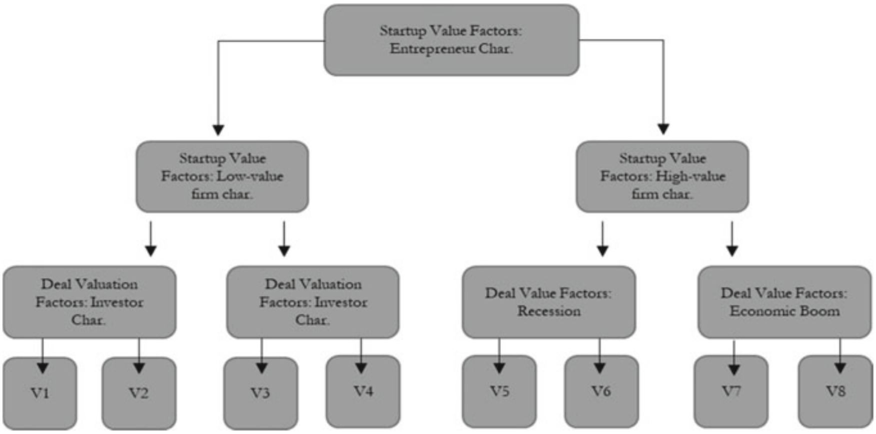

Startup Value Factors: Entrepreneur Char.
Startup Value Factors: Low-value firm char.
Startup Value Factors: High-value firm char.
Deal Valuation Factors: Investor Char.
Deal Valuation Factors: Investor Char.
Deal Value Factors: Recession
Deal Value Factors: Economic Boom
V1
V2
V3
V4
V5
V6
V7
V8

Fig. 3 Decision tree based on the Berre–Le Pendeven meta-model

3.1 Hierarchical Approaches: Regression Trees and Random Forests

Functionally speaking, CART-based microtargeting using regression tree and Random Forest approaches, the latter of which agglomerate large numbers of regression trees, can reorganize determinant impacts causing several key insights to emerge, which might otherwise be missed by regression-model approaches, or by more rudimentary estimation models. Firstly, key fault lines are expressed as threshold values along which branches diverge. Secondly, qualitative or categorical determinant factors such as geographical, sectoral, and business-model data, which has the potential to be information dense are taken into account. Thirdly, CART trees demonstrate areas and subsections of the data where given valuation determinants might be more or less influential, granting very precise insight into how valuation emerges.

For empirical-model estimation purposes, the informational content of descriptive and categorical characteristics such as geography, industry, legal form, or business model are often overlooked, despite the general possibility that these characteristics might bring-to-bear explanatory power equivalent to multiple associated numerical variables. Meanwhile, use of fixed effects to incorporate descriptive categorical characteristics suffers losses in explanatory power as the number of descriptive characteristics increases Wooldridge (2010), whereas microtargeting approaches improve their accuracy as the number and density of these characteristics increases.

Consequently, a key advantage of this approach is that startup valuation, selection, or survival probability can be microtargeted by including ever smaller and more specific categorical information, or combinations thereof.

3.2 Functional Form of Segmented Models

According to Krzywinski and Altman (2017), a CART approach does not develop or express a prediction equation. Instead, this approach partitions data along predictor axes into subsets with homogeneous values of the dependent variable. This in mind, machine learning algorithm reliance on optimal arrangement of decision factors is central to model functionality.

In tree-based approaches, this process is represented by decision trees, which can be used to make predictions from new observations. Several functional-form options exist mathematically, which are used operationally by practitioners in markets or in research settings. Furthermore, the combination and/or selective use of these can be a useful way to investigate and model causal relationships in detail. Within entrepreneurship studies, this can be used to investor selection, startup valuation, and startup survivability.

Log Transformation

Because log transformation renders summation and multiplication interchangeable, the use of variable log transformation can dramatically simplify regression models and mathematical relationships for the purposes of empirical specification (Benoit, 2011), while also lending themselves to model flexibility. Given the product property of logarithms, it is possible to express the model in its entirety as a summation model for intermediate-stage purposes, given the interchangeability of logarithm multiplication and summation (Miller et al., 2010). Specifically, this means that intermediate-stage model functional forms can reoriented in terms of variable order and in terms of interaction effects.

Moreover, log transformation "flattens" empirical relationships, by restraining the effect outliers have on variable medians and means. Because regression trees and partitioning methods in general are sensitive to outlier influence from dependent-variable outliers (Khan et al., 2013), flattening of outliers has potential to substantially increase explanatory power to regression-tree models, as log transformation reduces estimation problems associated with percentage changes from baseline (Keene, 1995), while maximizing data-scale-flattening (Ribeiro-Oliveira et al., 2018). Variables showing skewed distribution can also be made symmetric using log transformation (Keene, 1995).

Conversely, since log transformation also impacts multiplicative models (Benoit, 2011), the particular architectural shape of functions being modelled becomes unclear, as multiplication, summation, ratios, and other functional-form elements might also become unclear.

To reach a viable final outlook, one would need to see the model's log-transformed expression alongside the original expression, whose functional form captures in detail both variable order and possible interaction terms. In order to establish a tree, however, both variable order and relative variable importance need to be established. Overall, interaction terms between and among regression variables can shed some light on how specifically the model's explanatory variables interact with each other. This may indicate within-tree variable position, granting a more holistic and complete view on relationship and model causality structures.

Regression-Model Equations

Functionally speaking, regression-model equations, which consist of a summation of key variables, modified by factor coefficients, alongside constants and error terms. Fundamentally, this layout structure lends itself to near-direct transposition of segmented valuation approaches, was well as the approximation of most classically established firm valuation models, ranging from discountedcash flow valuation (DCF) approaches, to multiples-valuation approaches.

Because regression-model equations are generally expressed as summation functions, with each of the model's terms consisting of a variable and a coefficient, valuations can essentially be expressed as a summation of variables, coefficients, constants, and error terms. For instance, a discounted-revenue-based valuation, incorporating similar information to a discounted cash flow valuation (DCF), could approximate a free cash flow to equity (FCFE) approach by regressing valuation on present and historic Net Income figures in order to capture both free cash flow and its growth rate, as well as risk factors which drive the discount rate, which can be expressed as combinations of the risk-free rate and the applicable equity risk premia described in the CAPM model. This is modelled in Eq. 4.

Equation 4 Valuation regression model simulating free cash flow to equity (FCFE)

$$\begin{aligned}\text{Valuation}_{it} &= \alpha_i + \beta_1(\text{Net Income}_{it}) + \beta_2(\text{Net Income}_{it-n}) \\ &\quad + \beta_3(\text{Risk-Free rate}_t) + \beta_4(\text{Risk-Premium}_{it}) + u_{it}\end{aligned} \tag{4}$$

Meanwhile, a multiples-valuation approach, whose widespread popularity flows from its simplicity and ease of communication, as well as its ability to communicate the market's current mood (Damodaran, 2002), might seek to estimate valuation from as few as one valuation factor drawn from either a firm's balance sheet, income statement, or statement of cash flows. This however may come at the cost of sample selection, as developing a sample of relative firms and assets against which to compare valuation, can lead to standardization (or assumption of standardization) of variables outside of the valuation model. According to Damodaran (2002), the most widespread multiples-valuation model is the price/sales ratio, which describes valuation as a function of sales revenue, as outlined in Eq. 5

Equation 5 Price-to-sales ratio

$$\text{Price-to-Sales Ratio} = \frac{(\text{Firm's Total Market Share} - \text{Price})}{\text{Sales Revenue}} \tag{5}$$

Equation 6 demonstrates this ratio as an ordinary-least-squares (OLS) regression model, given by the parameter sales revenue, while β estimates price-to-sales ratio, whereas outside factors ranging from quantitative valuation determinants such as borrowing costs, R&D, CAPEX, or total assets (or asset subsets such as IP assets), to qualitative valuation factors such as those driven by industry or economic geography are sample selected to be constant, or assumed to be constant.

Equation 6 Price-to-sales ratio as an OLS regression model

$$\text{Valuation}_i = \alpha_\text{c} + \beta_\text{c}(\text{Sales Revenue}_i) + u_i \tag{6}$$

Apart from the use of regression-model functional form to express classical models, the OLS regression-model's functional form can also be used for summation-based segmented models, such as those outlined in Eqs. 1 and 2. In fact, this is even the case for models using hierarchical approaches, such as Mahmoud et al. (2022) express random-forest regressions using regression-model equations, simulating the summation-based segmented functional form of an OLS model.

Decision-Tree Model Functional Forms

Overall, substantial flexibility exists concerning the various functional forms that decision-tree models could conceivably adopt considering contexts in which they could be deployed, factors enumerated by the model, and both their relative and hierarchical explanatory power. While Krzywinski and Altman (2017) describe that a CART approach does not develop a prediction equation, CART regression-tree results can be used to modify or extend segmented models. Fundamentally, regression-tree model outputs make possible two practically viable segmentation model approaches.

Mahmoud et al. (2022), for example, express random-forest regressions using regression-model equations, simulating the functional form of an ordinary-least-squares model. This modelling approach has the advantage of capturing the causal relationship's overall directionality, which can be tested empirically, without specifically precluding existence of complex model functional forms.

Comparing Model Goodness-of-Fit

In principle, accuracy of regression-tree models can be compared to those of equivalently constructed regression models on the basis of their respective goodness-of-fit indicators. Whereas linear regressions are typically evaluated on the basis of R^2, Sandeep (2014) and Firmin (2021) outline that regression trees should be evaluated on the basis of $1 - R^2$ root-mean-squared-error.

Weighted Summation Segmentation

First, a rudimentary "back-of-the-envelope" approach to segmentation can be considered to be a modification of the Payne scorecard model, which includes model weighting to its segmentation approach. In order to obtain regression-tree model weights from the CART approach, it suffices to examine the model's variable-importance scores. While CART variable-importance outputs can aggregate to a maximum of 100%, as is the case in Table 3, aggregate variable-importance model outputs might also aggregate to less than 100%. While for CART models whose aggregate variable importance adds to 100%, it would suffice to assign variable-importance figures as weighting coefficients. For instances in which observed variable-importance outputs aggregate to less than 100%, however, factor-importance proportionality would need to be calculated as a first step, as outlined in Eq. 7:

Equation 7 CART variable-importance proportionality

$$\text{Factor-Coefficient}_i = \sigma_i(X)_i = \frac{\text{Variable Importance}_i}{\sum_n^i \text{Variable Importance}_i} \tag{7}$$

Fundamentally, this approach is useful as a generally applicable model approach, yielding a Payne-like scorecard model, which can be applied in a general fashion to entrepreneurial and startup markets at-large. For example, a Payne-style scorecard model, involving model weights, which could be constructed on the basis of firm characteristics and market characteristics, can take the form outlined in Eq. 8, combining the FCFE valuation factors with Payne model factors outlined in Table 1:

Equation 8 Weighted summation segmentation regression-tree valuation model simulating the FCFE valuation approach

$$\begin{aligned}\text{Valuation}_i &= \sigma_1\beta_1(\text{Net Income}_i) + \sigma_2\beta_2(\text{Risk-Free rate}_i)\\ &\quad + \sigma_3\beta_3(\text{Risk-Prem.}_i) + \sigma_4\beta_4\left(\text{Size of Opportunity}_i\right)\\ &\quad + \sigma_5\beta_5(\text{Competitive Env.}_i) + \sigma_6\beta_6(\text{IP}_i) \qquad (8)\end{aligned}$$

where

$$\sum_{i=1}^{n}\sigma_i = 1 \quad \text{but we observe} \quad \sum_{i=1}^{n}\widehat{\sigma_i} \leq 1$$

Here, σ refers to the weighting coefficient n of startup i, driven by variable importance (for example, the scale of Net Income's impact on startup i's valuation), while β refers to the impact coefficient n of startup i (for example, risk premium is a valuation determinant known be a constituent factor of the DCF-model discount rate (Damodaran, 2009), and as such, can be expected to have a negative β-coefficient).

Mechanically, this approach is viable for either continuous numerical variables, such as those drawn from a firm's financial statements (i.e. Net Income, fixed assets, etc.), as well as market data (i.e. business cycle and macroeconomic indicators), or for categorical and binary variables such as entrepreneur characteristics or intellectual property. Additionally, because CART regressions partition data along the predictor axes into dichotomous subsets, categorical variables (i.e. classifications such as sectoral-industry classifications and business-model classifications, as well as economic-geography variables such as counties, cities, inclusion in regional clusters) which are treated as binary variables.

Hierarchical Ordinal Segmentation

A second modelling approach can be referred to as the hierarchical ordinal segmentation approach. Given that the data are partitioned along predictor axes into subsets with homogeneous values of the dependent variable, a more complex hierarchical approach is also plausible. The basis of this approach would begin with adoption of terminal-node average values as ω-coefficients. These can subsequently be multiplied by a regression-tree's branch conditions and branch thresholds, as follows:

$$\omega_i(X)_j\left(\begin{cases} = 1 \text{ if } X \text{ is true} \\ = 0 \text{ if } X \text{ is false}\end{cases}\right)$$

Or

$$\omega(X)_j\left(\begin{cases} = 1 \text{ if } X \text{ is above the threshold} \\ = 0 \text{ if } X \text{ is below the threshold}\end{cases}\right)$$

Subsequently, the regression-tree model can be elaborated for any specific startup in accordance with the position it occupies in the regression tree. Equation 9 describes this functional form.

Equation 9 Valuation regression-tree model using hierarchical ordinal segmentation

$$\text{Valuation}_i = \omega_i \left(\prod_{i1}^{in} \text{Branch Threshold}_i \right)_i + \cdots + \omega_n \left(\prod_{n1}^{nn} \text{Branch Threshold}_n \right)_n \tag{9}$$

As a specific example building on Eq. 9, establishing a specific valuation model, Eq. 10 applies the hierarchical ordinal segmentation approach to the combined FCFE-market conditions valuation model outlined in Eq. 8 and ranking the nodes in hierarchical order following their order in Eq. 8. Note that this causes their order stated in the equation to change somewhat to reflect the conditionality relationship.

Equation 10 Valuation regression tree using hierarchical ordinal segmentation model approach

$$\begin{aligned}\text{Valuation}_i = {} & \omega_i \left(\prod_{i}^{I} \text{Net Income}_i \right)_i + \omega_j \left(\prod_{j}^{J} \text{Risk-Free rate}_j \right)_j \\ & + \omega_k \left(\prod_{k}^{K} \text{Risk-Premium}_k \right)_k + \omega_l \left(\prod_{l}^{L} \text{Size of Opportunity}_l \right)_l \\ & + \omega_m \left(\prod_{m}^{M} \text{Competitive Env.}_m \right)_m + \omega_n \left(\prod_{n}^{N} \text{IP}_n \right)_n \end{aligned} \tag{10}$$

Fundamentally, a key difference between this approach and the weighted-summation approach is that this approach is specific to the individual startup's position within the decision tree. Essentially, this means that the segmentation's functional form differs from that of weighted-summation approach, since a startup's placement on the regression tree may indicate functional form featuring either the repetition or omission of some of the regression-model's valuation determinants.

Another essential difference between the approaches is that while the weighted-summation approach can grant a holistic view of σ-weights across the dataset as a whole, the ordinal-model approach can directly provide a valuation estimate by placing the firm along regression-tree's terminal nodes (i.e. the regression-tree's leaf nodes).

Two-Tiered Approach

Given that inclusion of categorical variables has the potential to unearth valuable informational insights of both qualitative and quantitative nature and has the potential to be as information dense as the joint inclusion of multiple numerical variables, their use for research purposes remains a very valuable tool (Neter et al. 1990; Wooldridge, 2010). This is in particular the case with fixed-effects regressions, given that they can meaningfully incorporate categorical indicators such as geographical or industry-level designations. In the face of multiple information-dense categorical variables, however, this approach is subject to a hard limit in that explanatory power of joint fixed effects can be limited as the number of categorical variables grows.

What this means therefore is that either OLS or fixed-effects regressions can be deployed in order to capture the general causal overview among the model determinants and in order to detect information density and explanatory power of applicable categorical labels. In order to elaborate on any OLS or fixed-effects findings, CART (or possibly other cluster-driven approaches) can be utilized with the aim of enrichment or corroboration of findings.

Taking this into consideration, combined approaches are possible, with the potential to outperform single-method analysis in two important ways. Firstly, this approach can outperform a stand-alone OLS-based summation approach because the two-tiered approach can grant insights on the role, hierarchy, and relative-position of the model's near-significant explanatory factors (i.e. near-significant factors often have regions or subsets of the data, for which they are significant). Secondly, two-tiered approaches can provide detailed insight vis-à-vis scale and sign of factor impacts (i.e. β-coefficients), thereby improving upon stand-alone CART-based weighted summations.

4 Modelling Investment Selection and Startup Survivability

A Segmented Approach to Selection

Aside from predicting and modelling valuation, machine learning-driven segmented models also have viable applicability for modelling startup selection and startup survivability, both of which are parallel entrepreneurial finance topics which have historically encountered modelling difficulties. Similar to startup valuation, irregularity and non-transparency of data constitute considerable obstacles to model accuracy (Damodaran, 2009).

Startup selection, while a very nearby parallel entrepreneurial finance topic, which shares many of the same prominent authors, faces the additional difficulty of qualitative and intangible factor determinants and decision criteria assuming a more widespread and prominent role among business angel and venture capital investors responsible for startup-selection decisions. A segmented-model approach to startup-selection decisions faced by venture capitalists and business angels is presented by Siskos and Zopounidis (1987), which includes both decision weights and ordinal rank, as outlined in Table 2 (Siskos & Zopounidis, 1987).

Table 2 Weighting of marginal utilities for VC investment decision

Rank	Criteria	1st analysis weight	2nd analysis weight
1	Information security	0.044	0.095
2	Market trends	0.000	0.005
3	Market niche/position	0.164	0.162
4	Conjuncture sensibility	0.009	0.085
5	Result trends	0.347	0.167
6	Expected dividend rate	0.031	0.107
7	Quality of management	0.031	0.247
8	Research and development level	0.000	0.000
9	Accessibility to financial markets	0.373	0.132

Siskos and Zopounidis (1987). http://linkinghub.elsevier.com/retrieve/pii/0377221787900403

Fundamentally, while Siskos and Zopounidis (1987) use ordinal regression analysis to reach Table 2's findings, these results provide sufficient detail for the construction of a decision-tree model equation, using a weighted-summation functional form.

Citing Siskos and Zopounidis (1987)'s ordinal regression analysis approach, Dhochak and Doliya (2019) expresses ordinal selection data via fuzzy analytic hierarchy process (Fuzzy AHP), outlined in Fig. 4, consisting of decision criteria and decision sub-criteria. Essentially, the model's hierarchy represents cognitive organization, dividing criteria into various types of firm-level internal-resources, industry-level resources, and network effects.

Building on Table 2's multi-criteria decision factors, Fig. 5 proposes a hierarchical decision tree for selection ranking based on Siskos and Zopounidis' top five decision criteria, prioritizing the dominant decision criteria, according to their analysis weights, with accessibility to financial market and result trend constituting the top decision-tree branches, while market niche/position appears in multiple lower branches, due mainly to its heavy analysis-weight score in both first and second analysis. Ultimately however, HCA, CART or Random Forest results would provide the specific functional form for the final decision tree. In contrast to the Dhochak and Doliya fuzzy AHP approach, Fig. 5's decision-tree approach arranges hierarchy according to likely explanatory power, rather than cognitive factor * organization levels.

This approach may be especially useful, in particular to analyse choice models, which incorporate or suggest qualitative data points, such as the entrepreneur personality traits outlined in Murnieks et al. (2016), or the entrepreneur and investor traits described in Andreoli (2022).

A Segmented Approach to Survivability

Startup survival, on the other hand, would adopting a functional form requiring a binary dependent variable and can draw on parallels from conditional probability

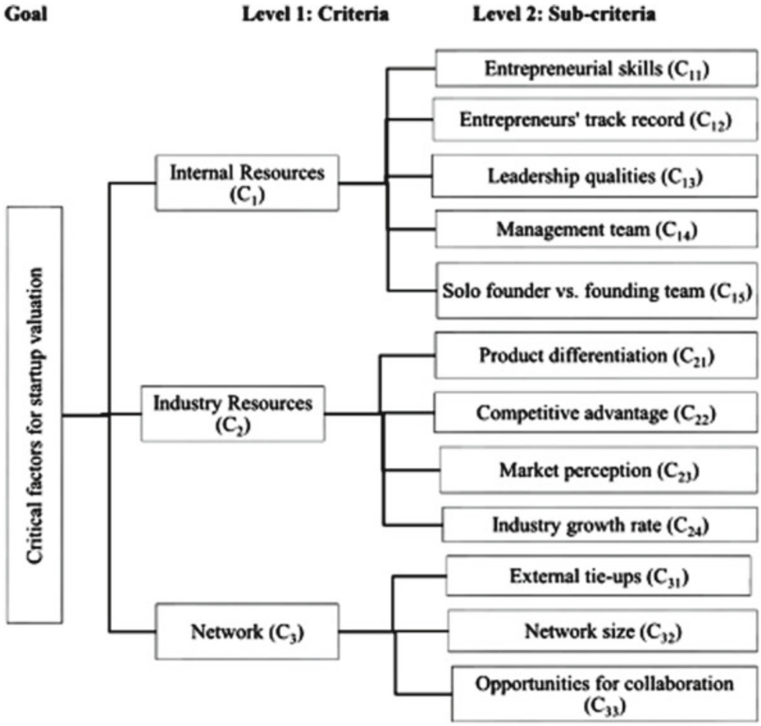

Goal
Level 1: Criteria
Level 2: Sub-criteria
Critical factors for startup valuation
Internal Resources (C$_1$)
Industry Resources (C$_2$)
Network (C$_3$)
Entrepreneurial skills (C$_{11}$)
Entrepreneurs' track record (C$_{12}$)
Leadership qualities (C$_{13}$)
Management team (C$_{14}$)
Solo founder vs. founding team (C$_{15}$)
Product differentiation (C$_{21}$)
Competitive advantage (C$_{22}$)
Market perception (C$_{23}$)
Industry growth rate (C$_{24}$)
External tie-ups (C$_{31}$)
Network size (C$_{32}$)
Opportunities for collaboration (C$_{33}$)

Fig. 4 Dhochak and Doliya startup-selection hierarchical decision model using fuzzy AHP

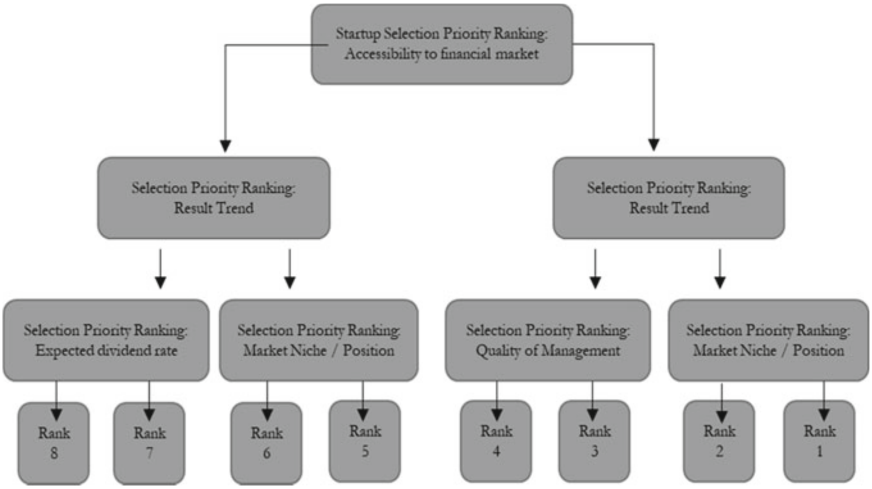

Startup Selection Priority Ranking: Accessibility to financial market
Selection Priority Ranking: Result Trend
Selection Priority Ranking: Result Trend
Selection Priority Ranking: Expected dividend rate
Selection Priority Ranking: Market Niche / Position
Selection Priority Ranking: Quality of Management
Selection Priority Ranking: Market Niche / Position
Rank 8
Rank 7
Rank 6
Rank 5
Rank 4
Rank 3
Rank 2
Rank 1

Fig. 5 Startup-selection ranking decision tree based on Siskos and Zopounidis decision weights

modelling, used in stress testing. Rebonato (2010), for example uses progression of conditional probabilities to model bank defaults and bank stress testing. A segmented machine learning approach would involve using a hierarchical tree-based algorithm (e.g. CART, ACH, or Random Forest). Alternatively, the tree-based algorithm's dependent variable can be expressed as a categorical variable instead of a binary variable, in order to capture varying degrees of financial distress, rather than simply bankruptcy as a binary term.

Specifically, determinants of startup survivability are described in the literature primarily in terms of market conditions. While Damodaran (2009) draws on the post-1998 sector-level survival likelihoods compiled by Knaup (2005) and Knaup and Piazza (2007), displayed in Table 3, in order to estimate credit-risk premium for startup valuation, purposes, it can also be used to construct sector-level startup-survival modelling.

In addition to sector-level survivability-determinants, macro-level market conditions such as GDP growth rates, prime-lending rates, and presence of business accelerators and startup accelerators are also known to play deterministic roles in modelling startup survivability (Gonzalez, 2017). In addition, Gonzalez (2017), which draws on US state-level data, describes considerable sectoral and state-level variation in one-year and four-year startup-survival likelihood. Econometrically, these findings can be represented as fixed-effects model including both state and industry-level fixed effects, as per Eq. 11.

Equation 11 Startup-survival likelihood fixed-effects model

$$\begin{aligned}\text{Survival Likelihood}_{\text{State,Industry}} = {} & \beta_1\left(\text{Real GDP Growth}_{\text{State,Industry}}\right) \\ & + \beta_2\left(\text{Prime Interest Rates}_{\text{State,Industry}}\right) \\ & + \beta_3\left(\text{Accelerators}_{\text{State,Industry}}\right) + \varepsilon_{\text{State,Industry}} \qquad (11)\end{aligned}$$

Table 3 Sector-level 7-year startup-survival likelihood

	Proportion of firms that were started in 1998 that survived through (%)						
	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Year 7
Natural resources	82.33	69.54	59.41	49.56	43.43	39.96	36.68
Construction	80.69	65.73	53.56	42.59	36.96	33.36	29.96
Manufacturing	84.19	68.67	56.98	47.41	40.88	37.03	33.91
Transportation	82.58	66.82	54.70	44.68	38.21	34.12	31.02
Information	80.75	62.85	49.49	37.70	31.24	28.29	24.78
Financial activities	84.09	69.57	58.56	49.24	43.93	40.34	36.90
Business services	82.32	66.82	55.13	44.28	38.11	34.46	31.08
Health services	85.59	72.83	63.73	55.37	50.09	46.47	43.71
Leisure	81.15	64.99	53.61	43.76	38.11	34.54	31.40
Other services	80.72	64.81	53.32	43.88	37.05	32.33	28.77
All firm	81.24	65.77	54.29	44.36	38.29	34.44	31.18

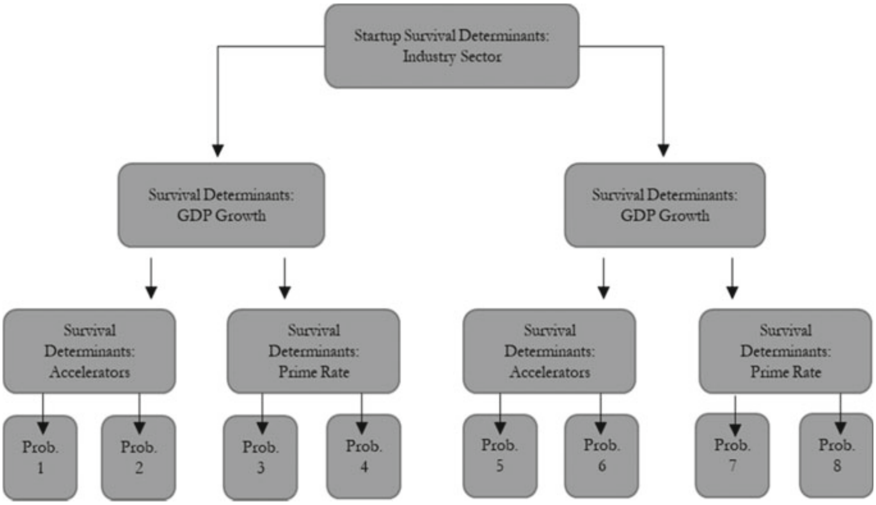

Startup Survival Determinants: Industry Sector
Survival Determinants: GDP Growth
Survival Determinants: GDP Growth
Survival Determinants: Accelerators
Survival Determinants: Prime Rate
Survival Determinants: Accelerators
Survival Determinants: Prime Rate
Prob. 1
Prob. 2
Prob. 3
Prob. 4
Prob. 5
Prob. 6
Prob. 7
Prob. 8

Fig. 6 Survival likelihood decision tree based on Damodaran (2009) and Gonzalez (2017)

Alternately, modelling startup survivability as a hierarchical decision-tree model can incorporate both the numerical determinants driving the startup-survival likelihood model outlined in Eq. 11, as well as the categorical variables which are used to construct Eq. 11's fixed effects. Hierarchically, Fig. 6 captures this relationship, adding the proposition that high-GDP growth startup survivability may by more influenced by prime-lending rates, whereas low-GDP growth startup survivability may by more influenced by presences and accessibility of startup accelerators. Because industry-level effects are described by both Damodaran (2009) and Gonzalez (2017), they are prioritized in this model, as they likely have substantial explanatory power.

5 Example of CART-Based Microtargeting Valuation Using a Single Categorical Variable

Tables 4 and 5 demonstrate the OLS and CART approaches to examine valuation-regression models drawn from Berre (2022), which include revenue, country-risk premium (capturing country-level risk-free rate), and sector-level CAPM-beta (capturing sector-level risk premium) as discounted cash flow valuation factors alongside business model.

In particular, revenues can be expected to have positive β-coefficients, while DCF-discount factor components (country-risk premium and CAPM-beta) can both be expected to have negative coefficients. Meanwhile, business model is a categorical variable, which may take the value "business-to-business" (B2B),

Table 4 OLS including DCF valuation factors and business models

OLS coefficients	Estimate	Std. error	*T*-value	*P*-value
(Intercept)	5.10E + 08	6.12E + 07	8.326	5.02E − 16***
Revenues	4.27E − 01	6.20E − 02	6.892	1.31E − 11***
Country-risk premium	− 4.19E + 09	3.18E + 09	− 1.317	0.188
Sectoral beta	− 4.61E + 08	6.02E + 07	− 7.664	6.67E − 14***
B2B&C	3.06E + 08	6.13E + 07	4.995	7.59E − 07***
B2B	9.80E + 07	6.25E + 07	1.569	0.117
B2C	6.37E + 08	6.28E + 07	10.138	< 2.00E − 16***

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Residual standard error: 546,900,000 on 644 degrees of freedom
Multiple *R*-squared: 0.2793
Adjusted *R*-squared: 0.2726
F-statistic: 41.6 on 6 and 644 DF, *p*-value: < 2.20E − 16

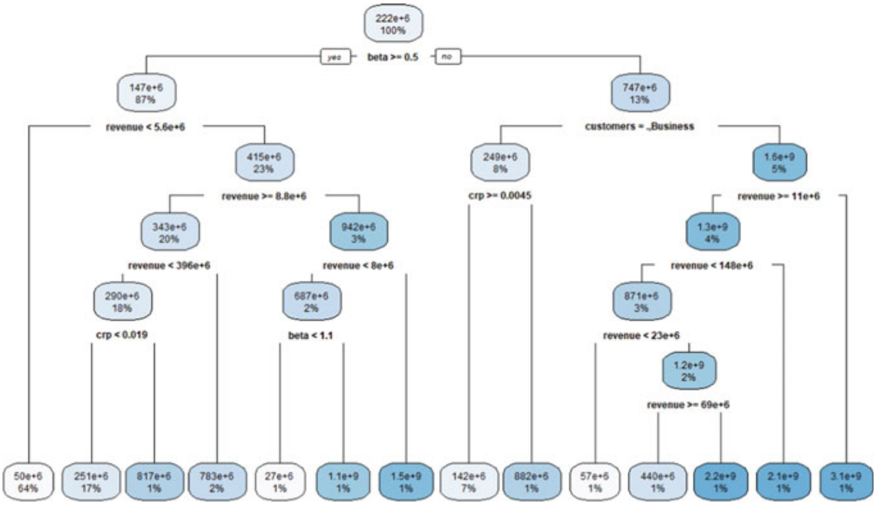
Table 5 CART including DCF valuation factors and business model

OBS: 1048				
End nodes: 15				
Complexity parameter	No. of split	RMSE	Cross-validation error	Cross-validation st. dev
0.1280	0	1.0000	1.0024	0.1634
0.0623	2	0.7441	0.8255	0.1484
0.0574	3	0.6817	0.8100	0.1479
0.0376	4	0.6243	0.7245	0.1398
0.0285	5	0.5867	0.7133	0.1397
0.0241	7	0.5296	0.7016	0.1385
0.0148	8	0.5055	0.6458	0.1366
0.0132	9	0.4906	0.6200	0.1317
0.0132	11	0.4643	0.6219	0.1318
0.0102	12	0.4512	0.6242	0.1318
0.0100	13	0.4409	0.6154	0.1318
Variable importance				
Revenue	Business model	Beta	Country-risk premium	
35	24	23	18	

"business-to-customer" (B2C), "business-to-business-and-customers" (B2B&C), or "business-to-government" (B2G).

First, Table 4 examines the relationship between startup valuations, DCF-factors, and business models, splitting business model into dummy variables, using an OLS model, finding that revenue's valuation impact is DCF consistent, while the discount factor appears to be driven by sector-level CAPM-beta. Lastly, the valuation impact of B2B is outweighed by both B2C and B2B&C.

Meanwhile, Table 5 outlines a decision tree-based CART valuation which includes revenue, country-risk premium (capturing country-level risk-free rate), and sector-level CAPM-beta (capturing sector-level risk premium) as discounted cash flow valuation factors alongside business model and describes startup pre-money valuations ranging from €27 million to €3.1 billion, and are partitioned hierarchically.



Given the structure of the regression tree in Table 5, the weighted-summation approach and the hierarchical ordinal approach would lead to somewhat-different functional forms. Equation 12 demonstrates a weighted-summation functional form example of the valuation model resulting from Table 4 regression tree, taking the resulting variable-importance indicators as σ-coefficients.

Equation 12 Valuation regression-tree model using weighted-summation segmentation

$$\begin{aligned}\text{Valuation}_i = {} & 0.35\beta_1(\text{Revenue}_i) + 0.24\beta_2(\text{Business Model}_i)\\ & + 0.23\beta_3(\text{Sectoral-Risk Beta}_i) + 0.18\beta_4(\text{Country-Risk Premium}_i)\end{aligned} \tag{12}$$

Using this approach, the highest-valuation tranche would first and foremost consist of startup with substantial revenue figures. This would be followed by firms which have business models, which focus on B2C, B2B&C, or B2G, and whose revenues are discounted by low sector-level CAPM-betas and low country-risk premiums. This means that the highest-valuation EU startup are firms which combine substantial revenue figures with a B2C, B2B&C, or B2G business model, and are located in a low-volatility industry, and based in a AAA-rated home-market such as Germany, Denmark, or Switzerland (Damodaran, 2021), whereas lowest-valuation EU startup are somewhat more likely to be based in higher-risk European markets (for example in the CEE, Baltic, or Euro-Med regions), and are characterized by high-risk industry sectors, low revenues, and B2B business model. Table 6 presents the regression-tree results outlined in Table 5, as a Payne-Style valuation scorecard.

By also drawing on the OLS findings outlined in Table 2 as a source of β-coefficients, a two-tiered approach is possible. Here, Eq. 13 and Table 6 capture the revisions possible by inclusion of β-coefficients drawn from Table 4. Because business model has been re-transcribed as its constituent (statistically significant) dummy variables, B2C and B2B&C, the functional form of the valuation model includes terms and coefficients for each of these business models, but not B2G nor B2B.

Equation 13 Valuation regression-tree model using weighted-summation segmentation

$$
\begin{aligned}
\text{Valuation}_i = {} & (0.35 * 0.4273)(\text{Revenue}_i) + (0.24 * 637{,}000{,}000_{\text{B2C}})(\text{Business Model}_i) \\
& + (0.24 * 305{,}900{,}000_{\text{B2B\&C}})(\text{Business Model}_i) \\
& + 0.23\beta_3(-460{,}900{,}000_i) + 0.18\beta_4(._i)
\end{aligned}
\tag{13}
$$

Building on this revision, Table 7 constitutes a revision of the Payne-style summation scorecard, originally outlined in Table 1, featuring incorporation of β-coefficients drawn from use of a two-tiered valuation approach.

Table 6 CART-based valuation as weighted-summation segmentation results presented in Payne-style scorecard

Weighting	Sign of β coef.	Impact on startup valuation
35%	Impact	*Revenue*
	+	Valuation is positively by revenue
		Business model
24%	Impact	*Client focus of the business*
	–	Business-to-business (B2B)
	+	Business-to-customer (B2C)
	+	Business-to-business and customer (B2B&C)
	+	Business-to-government (B2G)
		Discount factor
23%	Impact	*Sector-level CAPM-beta*
	–	Valuation is negatively impacted by sectoral risk
18%	Impact	*Country-risk premium*
	–	Valuation is negatively impacted by country-risk premium
Total		
100%		

Table 7 Two-tiered revised valuation as weighted-summation results presented in Payne-style scorecard

Weighting	β coef.	Impact on startup valuation
35%	Impact	*Revenue*
	0.4273	Valuation is positively by revenue. Per EUR of revenue
		Business model
24%	Impact	*Client focus of the business*
	–	Business-to-business (B2B)—(not significant)
	637,000,000	Business-to-customer (B2C)
	305,900,000	Business-to-business and customer (B2B&C)
	–	Business-to-government (B2G)—(not significant)
		Discount factor
23%	Impact	*Sector-level CAPM-beta*
	– 460,900,000	Valuation is negatively impacted by sectoral risk. Per 1.00 of CAPM-beta
18%	Impact	*Country-risk premium*
	–	Valuation is negatively impacted by country-risk premium. But is not statistically significant within the European EU/EEA dataset. Near-significance indicates that CRP is likely to be significant in more diverse datasets
Total		
100%		

Alternatively, hierarchical ordinal segmentation, the second valuation-segmentation approach, gives rise to a substantially larger and more complex valuation-model functional form, as each of the regression-tree's branch and terminal nodes can be represented in the model. Equation 14 demonstrates an example of the second valuation-segmentation approach, outlined in Eq. 8. Because the CART results include 14 terminal nodes, as well as numerous branch nodes, the size and complexity of the entire long-form valuation equation is substantial.

Equation 14 Valuation regression-tree hierarchical ordinal segmentation model approach

$$\begin{aligned}
\text{Valuation}_i = {} & 50{,}000{,}000(\text{Sectoral-Beta} \geq 0.5) * (\text{Revenue}_i < 5{,}600{,}000) \\
& + 251{,}000{,}000(\text{Sectoral-Beta} \geq 0.5) * (\text{Revenue}_i \geq 5{,}600{,}000) \\
& * (\text{Revenue}_i \geq 8{,}800{,}000) * (\text{Revenue}_i < 369{,}000{,}000) \\
& * \left(\text{Country-Risk-Premium}_i < 0.019\right) \\
& + 817{,}000{,}000(\text{Sectoral-Beta} \geq 0.5) * (\text{Revenue}_i \geq 5{,}600{,}000) \\
& * (\text{Revenue}_i \geq 8{,}800{,}000) * (\text{Revenue}_i < 369{,}000{,}000) \\
& * \left(\text{Country-Risk-Premium}_i \leq 0.019\right)
\end{aligned}$$

$$
\begin{aligned}
&+ 783{,}000{,}000(\text{Sectoral-Beta} \geq 0.5) * (\text{Revenue}_i \geq 5{,}600{,}000)\\
&* (\text{Revenue}_i \geq 8{,}800{,}000) * (\text{Revenue}_i \geq 369{,}000{,}000)\\
&+ 27{,}000{,}000(\text{Sectoral-Beta} \geq 0.5) * (\text{Revenue}_i \geq 5{,}600{,}000)\\
&* (\text{Revenue}_i < 8{,}800{,}000) * (\text{Revenue}_i < 8{,}000{,}000)\\
&* (\text{Sectoral-Beta} < 1.1)\\
&+ 1{,}100{,}000{,}000(\text{Sectoral-Beta} \geq 0.5) * (\text{Revenue}_i \geq 5{,}600{,}000)\\
&* (\text{Revenue}_i < 8{,}800{,}000) * (\text{Revenue}_i < 8{,}000{,}000)\\
&* (\text{Sectoral-Beta} \geq 1.1)\\
&+ 1{,}500{,}000{,}000(\text{Sectoral-Beta} \geq 0.5) * (\text{Revenue}_i \geq 5{,}600{,}000)\\
&* (\text{Revenue}_i < 8{,}800{,}000) * (\text{Revenue}_i \geq 8{,}000{,}000)\\
&+ 142{,}000{,}000(\text{Sectoral-Beta} < 0.5) * (\text{Business Model}_i = \text{B2B})\\
&* \left(\text{Country-Risk-Premium}_i \geq 0.0045\right)\\
&+ 882{,}000{,}000(\text{Sectoral-Beta} < 0.5)\\
&* (\text{Business Model}_i = \text{B2B}) * \left(\text{Country-Risk-Premium}_i < 0.0045\right)\\
&+ 57{,}000{,}000(\text{Sectoral-Beta} < 0.5)\\
&* (\text{Business Model}_i = \text{B2C or B2B\&C or B2G})\\
&* (\text{Revenue}_i \geq 11{,}000{,}000) * (\text{Revenue}_i < 148{,}000{,}000)\\
&* (\text{Revenue}_i < 23{,}000{,}000)\\
&+ 440{,}000{,}000(\text{Sectoral-Beta} < 0.5)\\
&* (\text{Business Model}_i = \text{B2C or B2B\&C or B2G})\\
&* (\text{Revenue}_i \geq 11{,}000{,}000) * (\text{Revenue}_i < 148{,}000{,}000)\\
&* (\text{Revenue}_i \geq 23{,}000{,}000) * (\text{Revenue}_i \geq 69{,}000{,}000)\\
&+ 2{,}200{,}000{,}000(\text{Sectoral-Beta} < 0.5)\\
&* (\text{Business Model}_i = \text{B2C or B2B\&C or B2G})\\
&* (\text{Revenue}_i \geq 11{,}000{,}000) * (\text{Revenue}_i < 148{,}000{,}000)\\
&* (\text{Revenue}_i \geq 23{,}000{,}000) * (\text{Revenue}_i < 69{,}000{,}000)\\
&+ 2{,}100{,}000{,}000(\text{Sectoral-Beta} < 0.5)\\
&* (\text{Business Model}_i = \text{B2C or B2B\&C or B2G})\\
&* (\text{Revenue}_i \geq 11{,}000{,}000) * (\text{Revenue}_i \geq 148{,}000{,}000)\\
&+ 3{,}100{,}000{,}000(\text{Sectoral-Beta} < 0.5)\\
&* (\text{Business Model}_i = \text{B2C or B2B\&C or B2G})\\
&* (\text{Revenue}_i < 11{,}000{,}000) \qquad (14)
\end{aligned}
$$

An interesting detail about the regression tree, described in Table 4, is that several of the nodes indicate unicorn valuation. Essentially, this tree model appears to contain a recipe for unicorn valuations. Furthermore, we see that revenue drives the majority of the lower and intermediate branches, corroborating revenue's dominant variable-importance role.

Nevertheless, while the entire regression-tree valuation function outlined in Eq. 13 is sizable and cumbersome, it is not necessary to estimate the function as whole. Rather, because segments of the function where the criteria are not met are zero, it suffices to estimate the branches and terminal node where the firm actually finds itself. For example, for a startup located in the rightmost terminal node, whose sectoral beta would be larger than 0.5, and whose revenue is less than €50,000,000, Eq. 15 reduces to

Equation 15 Valuation regression-tree model reduced-form ordinal segmentation model approach

$$\text{Valuation}_i = 50{,}000{,}000(\text{Sectoral-Beta} \geq 0.5) * (\text{Revenue}_i < 5{,}600{,}000) \quad (15)$$

Although this reduced form of the model is both compact and immediately useful for practitioner purposes, substantial detail is lost in terms of other-path branches and terminal nodes, as well as their distributions and threshold values.

6 Discussion and Further Research

Overall, segmented models are historically underappreciated within empirical finance literature, with segmented models surfacing in but a small, obscure fraction of startup-valuation literature (Berre & Le Pendeven, 2022), as well as in startup-selection and startup-survivability models. In particular, opportunities to employ this approach for modelling of startup selection are particularly relevant, given the relative prominence of qualitative decision factors, as outlined in Murnieks et al. (2016) and Andreoli (2022), and as described by Wessendorf et al. (2019).

Nevertheless, appearance of segmented models in industry and practitioner-sourced grey literature (for example, Berkus, 2016; Ernst & Young, 2020; Ewing Marion Kauffman Foundation, 2007; Goldman, 2008; Payne, 2011) serves as an unmistakable indication that segmented approaches have established traction among industry practitioners ranging from business angels and VC investors to auditing and consultancy practitioners.

Why Segmented Models Work?

While these segmented estimation models might presently be under represented within economic literature (and entrepreneurial finance literature in particular), the ongoing proliferation of machine learning techniques can be expected to increase diversity, viability and popularity of segmented models within the literature, given that there are several empirical approaches drawn from both econometrics and machine learning that segmented models can be adapted to. In principle, the industry popularity and usefulness in markets of segmented estimation models can be attributed to numerous noteworthy positive qualities which characterize them.

First, segmented models are directly transposable to empirical modelling, making investigation of their validity and accuracy a relatively straightforward task. Fundamentally, this is the case because both CART and OLS models can be expressed in segmented functional form.

Second, segmented models are mathematically straightforward, making them both straightforward to understand and easy to communicate to clients, investors, and stakeholders. This quality may help explain the widespread popularity of the Berkus and Payne methods among industry practitioners and among industry sources, given that Damodaran (2002) ascribes this quality.

Third, comes their considerable flexibility. Because the segmented estimation-models' functional form are readily transposable for the purposes of empirical modelling, they are also highly adaptable. This means that they can be altered by adding or modifying the impacts of determinant factors as the need arises, for example, by adding segments to capture interaction terms or niche functional form segments. Furthermore, they can be constructed by modifying other styles of selection models, valuation models, survivability models, and stress-testing models. For example, relative-valuation models can be combined into two-factor or three-factor segmented valuation models, while both multi-decision selection models such as Siskos and Zopounidis (1987), and fuzzy analytic hierarchy process outlined in Dhochak and Doliya (2019), can serve as the basis for segmented hierarchical models.

The rise and proliferation of hierarchical empirical approaches, including not only CART-based regression trees, but also related approaches, such as the bottom-up Hierarchical Ascending Classification decision trees, and Random Forest has yielded the proliferation of increasingly accurate and flexible prediction models, which can not only be used for improved accuracy in entrepreneurial finance modelling, but also for speedy decision making, as well as the construction of increasingly flexible segmented models. This indicates that the use of such approaches in the business and market landscape can only be expected to proliferate in future.

Contributions and Further Research

Because this study focuses on the use and import or methodological approaches from industry practitioners, as well as from political science and marketing journals into entrepreneurial finance literature, this study adds to the existing body of research in several ways by both filling existing gaps in the theory, and by elaborating on already existing published empirical findings.

First, this study ties together practitioner approaches and peer-review literature trends. While practitioner-derived or industry-oriented literature such as Ewing Marion Kauffman Foundation (2007) or Ernst and Young (2020) point to segmented valuation models such as those described by Payne (2011) and Berkus (2016), this approach, seen in studies such as Hand (2005) and Sievers et al. (2013) for valuation models and Siskos and Zopounidis (1987) for selection models, is relatively rare within peer-review literature. This may be owed to the overall need for model sophistication in order interaction effects and variable hierarchies within models. This study provides an overview and synthesis of these approaches, which can be generally deployed by practitioners and experts across a wide variety of markets, while also providing context for the ongoing debate within peer-review literature.

Second, this study elaborates on already existing published research in the entrepreneurial finance field. Existing studies which use segmented approaches devote little space to exploring model functional form. Here again, the overall need for model sophistication in order interaction effects and variable hierarchies within models is apparent.

Third, this study describes use of newly emergent empirical techniques and describes how to systematically make use of them in a consistent way. While microtargeting based on hierarchical decision trees can take several forms in terms of machine learning algorithms (i.e. recursive partitioning, agglomerative hierarchical clustering, Random Forest), the modelling functional form that can be applied for startup valuation, startup selection, or startup survival intended to accompany such modelling approaches has heretofore not yet appeared in literature. This may be owed to the overall novelty of such approaches within entrepreneurial financial literature up until now.

Given that machine learning approaches are generally confronted relatively early on with questions of model selection and algorithm selection, further research using the principles outlined in this paper should consider complexity and shape of functional form as a fundamental part of model selection and algorithm selection, as a combined model outlook. Furthermore, this combined outlook can and should be taken into consideration for all applications of machine learning approaches within finance, economics, or entrepreneurship research, as well and practice thereof in the marketplace.

Implications of this research are far reaching. For markets and industry practitioners, elaboration on why and how hierarchical segmented models work for selection, valuation, and survivability estimation, as well as how they relate to emerging machine learning approaches can lead to the development of new and bespoke entrepreneurial finance models going forward, as industry practitioners may increasing adopt this style of estimation approach. Meanwhile, the emergence of investors linked to the big data and machine learning industries (ranging from CVCs to specialized consultants and experts) may someday try to automate tree-based segmented selection, valuation, and survivability approaches, in contexts where it may be appropriate to do so (for example the implementation of trading bots in a crowdfunding platform or P2P-lending platform setting). For investors, as well as for third parties, implications are also far reaching because these models can hypothetically deliver accurate estimations via microtargeting, which in its purest form is able to bypass difficult to obtain or confidential firm data, making accurate estimations of valuation, selection, and survivability substantially more widespread within startup markets.

Meanwhile, for policy-making circles, implications of proliferation of segmented models as machine learning approaches evolve, develop, and proliferate, might be a more niche and targeting understanding of startup markets, a body of knowledge which may be very useful for the purposes of SME policy, as well as in targeting key sectors, regions, asset classes, or municipalities going forward.

Fundamentally, future research may build on this study by using the principles described here for empirical studies featuring hierarchical machine learning approaches for the development of hierarchical segmented models. Since this approach is still in its emergent phases, it may be feasible to "push-the-envelope" on what is empirically possible. Doing so can be helped, for instance by development of a taxonomy of entrepreneurial finance relevant configurations, clusters, and categorical variables, so that future microtargeting research can grow beyond reliance on industry-sector, business-model, and economic-geography variables (such as cities or postal codes).

Lastly, this research can be used as a roadmap for future studies intending to use either hierarchical machine learning techniques within entrepreneurial finance, for industry practitioners interested in using machine learning techniques to establish bespoke segmented entrepreneurial finance models, or machine learning professionals interested in deploying their expertise for entrepreneurial finance (for example in a fintech setting).

Acknowledgements This article has been sponsored by the "Finance for innovation" Chair at Audencia Business School. Special thanks is accorded to Benjamin Le Pendeven, the chair director.

References

Andreoli, J. J. (2022). *Entrepreneurial finance: Towards determinants for early-stage investments*. ProefschriftMaken. ISBN: 978-90-8980-159-3.

Bellavitis, C., Filatotchev, I., Kamuriwo, D. S., & Vanacker, T. (2017). Entrepreneurial finance: New frontiers of research and practice: Editorial for the special issue embracing entrepreneurial funding innovations. *Venture Capital: An International Journal of Entrepreneurial Finance, 19*(1–2), 1–16.

Benoit, K. (2011). *Linear regression models with logarithmic transformations* [Working Paper]. Methodology Institute London School of Economics.

Berkus, D. (2016). *The Berkus method—Valuing the early-stage investment*. Berkonomics.

Berre, M. (2022). Which factors matter most? Can startup valuation be micro-targeted? In *International Conference on Small Business (ICSB) World Conference* [Working Paper]. Washington, DC.

Berre, M., & Le Pendeven, B. (2022). What do we know about start-up valuation drivers? A systematic literature review. *Venture Capital*. https://doi.org/10.1080/13691066.2022.2086502

Damodaran, A. (2002). *Investment valuation: Tools and techniques for determining the value of any asset*. In Wiley Finance Series. Wiley. ISBN: 9780471414902.

Damodaran, A. (2009). *Valuing young, start-up and growth companies: Estimation issues and valuation challenges* [SSRN Scholarly Paper ID 1418687]. Rochester, NY: Social Science Research Network.

Damodaran, A. (2019). *The disruption dilemma: Valuing the disruptors & disrupted*. NYU Stern. https://pages.stern.nyu.edu/~adamodar/pdfiles/country/Disruption2019.pdf

Damodaran, A. (2021). *Equity risk premiums (ERP): Determinants, estimation, and implications—The 2021 edition*. Available at SSRN: https://ssrn.com/abstract=3825823. https://doi.org/10.2139/ssrn.3825823

Dhochak, M., & Doliya, P. (2019). Valuation of a startup: Moving towards strategic approaches. *Journal of Multi-Criteria Decision Analysis, 27*, 39–49.

Ellis, J. H. M., Williams, D. R., & Roscorla, F. B. (2001). Managing Japan-Europe industrial buyer-supplier relationships: A conceptual and empirical study of the Japanese market for high-technology marine equipment. *Journal of the Asia Pacific Economy, 6*(2), 232–243.

Ernst & Young. (2020). *Startup funding full eGuide—The factory*. EY—The Factory. https://thefactory.works/wp-content/uploads/2020/10/Eguide_Funding_A4.pdf?mc_cid=df85bfb95a&mc_eid=f80b29de26. Accessed June 30, 2022.

Ewing Marion Kauffman Foundation. (2007). *Valuing pre-revenue companies*. Angel Capital Association. https://www.angelcapitalassociation.org/data/Documents/Resources/AngelCapitalEducation/ACEF_-_Valuing_Pre-revenue_Companies.pdf

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance, 25*, 383–417.

Firmin, S. (2021). *Understanding the outputs of the decision tree tool*. Alteryx Designer Knowledge Base. https://community.alteryx.com/t5/Alteryx-Designer-Knowledge-Base/Understanding-the-Outputs-of-the-Decision-Tree-Tool/ta-p/144773. Accessed September 01, 2022.

Goldman, M. (2008). *Valuation of startup and early-stage companies*. The Value Examiner. http://www.michaelgoldman.com/Publications/Goldman%20Valuation%20of%20Start-ups.pdf

Gonzalez, G. (2017). What factors are causal to survival of a startup? *Muma Business Review, 1*, 097–114.

Hand, J. R. M. (2005). The value relevance of financial statements in the venture capital market. *The Accounting Review, 80*(2), 613–648.

Keene, O. N. (1995). The log transformation is special. *Statistics in Medicine, 14*(8), 811–819.

Khan, I., Capozzoli, A., Corgnati, S. P., & Cerquitelli, T. (2013). Fault detection analysis of building energy consumption using data mining techniques. *Energy Procedia, 42*(2013), 557–566.

Knaup, A. E. (2005). Survival and longevity in the business employment dynamics data. *Monthly Labor Review,* 50–56.

Knaup, A. E., & Piazza, M. C. (2007). Business employment dynamics data: Survival and longevity. *Monthly Labor Review,* 3–10.

Krzywinski, M., & Altman, N. (2017). Classification and regression trees. *Nature Methods, 14*, 757–758. https://doi.org/10.1038/nmeth.4370

Mahmoud, F., Zahoor, A., Hussain, N., & Younes, B. Z. (2022). Working capital financing and firm performance: A machine learning approach. In *Financial Economics Meeting (FEM-2022)*. Paris.

Miller, J., O'Neill, M., & Hyde, N. (2010). *Intermediate algebra* (2nd ed.). McGraw-Hill Higher Education.

Murray, G. R., & Scime, A. (2010). Microtargeting and electorate segmentation: Data mining the American national election studies, *Journal of Political Marketing*, *9*(3), 143–166. https://doi.org/10.1080/15377857.2010.497732

Miloud, T., & Cabrol, M. (2011). Les facteurs strategiques inluençant l'evaluation des start-ups par les capitaux-risqueurs. *Revue Management et Avenir, 49*, 36–61.

Murnieks, C. Y., Cardon, M. S., Sudek, R. T., White, D., & Brooks, W. T. (2016). Drawn to the fire: The role of passion, tenacity and inspirational leadership in angel investing. *Journal of Business Venturing, 31*(4), 468–484.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). *Applied linear statistical models: Regression, analysis of variance, and experimental design* (3rd ed.). McGraw-Hill Inc.

Rebonato, R. (2010). *Coherent stress testing: A Bayesian approach to the analysis of financial risk*. In Wiley Finance Series.

Ribeiro-Oliveira, J. P., Garcia de Santana, D., Pereira, V. J., & Machado dos Santos, C. (2018). Data transformation: An underestimated tool by inappropriate use. *Acta Scientiarum Agronomy, 40*(1), e35300. https://doi.org/10.4025/actasciagron.v40i1.35300

Sandeep. (2014). Difference between rel error and xerror in rpart regression trees, version: 2014-07-02. https://stats.stackexchange.com/users/45985/sandeep. URL: https://stats.stackexchange.com/q/105536. Accessed September 01, 2022.

Sievers, S., Mokwa, C. F., & Keienburg, G. (2013). The relevance of financial versus non-financial information for the valuation of venture capital-backed firms. *European Accounting Review, 22*(3), 467–511.

Siskos, J., & Zopounidis, C. (1987). The evaluation criteria of the venture capital investment activity: An interactive assessment. *European Journal of Operational Research, 31*(3), 304–313. http://linkinghub.elsevier.com/retrieve/pii/0377221787900403

Wessendorf, C., Kegelmann, J., & Terzidis, O. (2019). Determinants of early-stage technology venture valuation by business angels and venture capitalists. *International Journal of Entrepreneurial Venturing, 11*(5), 489.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). The MIT Press.

Zopounidis, C., Galariotis, E., Doumpos, M, Sarri, S., & Andriosopoulos, K. (2015). Multiple criteria decision aiding for finance: An updated bibliographic survey, *European Journal of Operational Research*, *247*, 339–348.

Zopounidis, C., & Doumpos, M. (2002). Multi-criteria decision aid in financial decision making: Methodologies and literature review. *Journal of Multi-Criteria Decision Analysis, 11*(4–5), 167–186.