The role of institutions in early-stage entrepreneurship: an explainable artificial intelligence approach


Published in:
Journal of Business Research

Document Version:
Publisher's PDF, also known as Version of record

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
Copyright 2024 the authors. This is an open access article published under a Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the author and source are cited.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen’s institutional repository that provides access to Queen’s research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person’s rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.

Open Access
This research has been made openly available by Queen’s academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: http://go.qub.ac.uk/oa-feedback

Download date: 21. Mar. 2024
The role of institutions in early-stage entrepreneurship: An explainable artificial intelligence approach

Byron Graham a,*, Karen Bonner b

a Queen’s Business School, Queen’s University Belfast, United Kingdom
b Economic Policy Centre, Ulster University, United Kingdom

ARTICLE INFO

Keywords:
Entrepreneurship
Explainable artificial intelligence
Institutional theory
Machine learning
Entrepreneurship policy

ABSTRACT

Although the importance of institutional conditions in fostering entrepreneurship is well established, less is known about the dominance of institutional dimensions, their predictive ability, and more complex non-linear relationships. To overcome the limitations of traditional regression approaches in addressing these gaps we apply techniques from explainable artificial intelligence to study the dominance and non-linearity of institutional dimensions in predicting country-level early-stage entrepreneurship. Eight machine learning algorithms are applied to matched data from the Global Entrepreneurship Monitor, Index of Economic Freedom, and World Bank across 573 observations from 81 countries. Findings from the most accurate random forest model reveal considerable non-linearity in the relationships between institutional dimensions and entrepreneurship, as well as heterogeneity in the importance of individual dimensions, with an overall trend towards the dominance of cultural-cognitive institutions. These findings contribute to institutional theory and highlight important areas where machine learning methods can contribute to entrepreneurship research and policy.

1. Introduction

Entrepreneurship is recognised as a key driver of economic and social development (Thurik & Wennekers, 1999), and is thus of crucial interest to policy makers who seek to shape the necessary conditions for starting a business (Baumol & Strum, 2007). Early-stage entrepreneurship is particularly important and typically the focus of policy intervention (Hart, 2003; Lundstrom & Stevenson, 2005) as it represents both the setting up and the creation of new businesses which support growth and development via increased competition, job creation and enhanced innovation capacity through the commercialisation of ideas (Audretsch et al., 2006).

Levels of entrepreneurship differ substantially across countries, and a large body of research has focused on understanding the reasons behind these differences by identifying the key determinants of entrepreneurship (Arin et al., 2015; Nikolaev et al., 2018; Urbano et al., 2019). Although various theoretical frameworks have been proposed, institutional theory has emerged as a useful theoretical framework to study the factors that foster country-level entrepreneurship (Aparicio et al., 2021; Busenitz et al., 2000; Fredström et al., 2021; Stenholm et al., 2013; Su et al., 2017; Urbano et al., 2019; Urbano & Alvarez, 2014; Valdez & Richardson, 2013). Existing research highlights the importance of formal and informal institutions in fostering entrepreneurship, and has elaborated on the roles of specific regulatory, cultural-cognitive, and normative institutions in country-level entrepreneurship (Busenitz et al., 2000; Stenholm et al., 2013; Urbano & Alvarez, 2014; Valdez & Richardson, 2013). Development of this understanding has led to increased attention around ‘next generation’ questions which include a focus on which institutions matter most for entrepreneurship and more complex heterogeneity and non-linear relationships (Audretsch et al., 2022, p. 1). To this end, recent studies have begun to explore more nuanced relationships between institutions and entrepreneurship such as asymmetry and changes over time (Mickiewicz et al., 2021), non-linearity (Audretsch et al., 2019; Chowdhury et al., 2019), heterogeneity of institutions and outcomes (Audretsch et al., 2022), as well as including a wider range of institutional dimensions (Urbano et al., 2019), and entrepreneurial outcomes such as the quality and quantity of entrepreneurship (Chowdhury et al., 2019). Despite this progress, few studies have focused on the dominance of institutions and more complex non-linearity in predicting entrepreneurship, with studies advocating for more work in these areas (Audretsch et al., 2019, 2022; Chowdhury et al., 2019).

* Corresponding author.
E-mail address: byron.graham@qub.ac.uk (B. Graham).
Machine Learning (ML) approaches are ideally suited to the study of next generation questions on the role of institutions in entrepreneurship, by providing a means of identifying more complex non-linear relationships, as well as providing measures of variable importance which can be used to study dominance (Graham & Bonner, 2022). The potential for ML to contribute novel insights has also been recognised in the wider entrepreneurship literature, with recent calls for more work adopting ML methods (Lêvesque et al., 2020; Obschonka & Audretsch, 2020), alongside similar calls in the broader business literature (von Krogh et al., 2023). ML approaches can be viewed as complementary to the more traditional regression-based techniques by helping to address some of their limitations (Molina & Garip, 2019), as well as providing additional opportunities to contribute to theory and practice (Breiman, 2001b; Delen & Zolbanin, 2018; Molina & Garip, 2019; Shrestha et al., 2021).

An emerging body of literature has begun to apply machine learning to the study of entrepreneurship across areas such as country-level opportunity prediction (Jabeur et al., 2021), individual level entrepreneurial activity (Graham & Bonner, 2022; Schade & Schuhmacher, 2023), and entrepreneurial intentions (Wei et al., 2020). Although progress has been made, further work is required to understand the potential of machine learning to complement traditional regression-based approaches in entrepreneurship research, as well as elaborating on the differences in approach and the insights that can be gained. Although a small body of past work has applied ML to the study of country-level entrepreneurship (Jabeur et al., 2021), previous studies have not applied an institutional theoretic lens alongside ML methods to study the determinants of entrepreneurship.

Although ML methods have the potential to contribute to the study of entrepreneurship and the role of institutional drivers, they have faced criticism due to a lack of interpretability (Burrell, 2016). Recent advances in explainable artificial intelligence (XAI), such as the development of permutation importance and local interpretable model-agnostic explanations (LIME) have helped to overcome issues with model interpretability (Linardatos et al., 2021; Wang et al., 2022). These advances facilitate the interpretation of models and provide explanations about how predictions are made. Despite these recent advances in XAI they have received limited attention in the entrepreneurship literature. This has resulted in important methodological gaps in the literature about how ML and XAI techniques can contribute to entrepreneurship research, and how the approach and insights differ from those gained from traditional regression-based techniques. To address these methodological gaps and respond to calls for a focus on ‘next generation’ questions around the role of institutions in entrepreneurship (Audretsch et al., 2019, 2021; Busenitz et al., 2000; Chowdhury et al., 2019; Stenholm et al., 2013; Urbano & Alvarez, 2014; Valdez & Richardson, 2013), the dominance analysis highlights heterogeneity in the importance of institutional dimensions across and within Scott’s (1995) institutional pillars. Across the three pillars, the cultural-cognitive dimensions are dominant, with the skills and opportunity recognition dimensions most important within this pillar, whilst regulatory and normative dimensions are less predictive. Past entrepreneurial behaviour and intentions are also highly predictive of future levels of Total Early-Stage Entrepreneurial Activity (TEA). These findings also contribute to the small body of entrepreneurship research that has taken a dominance analysis approach (Arin et al., 2015; Graham & Bonner, 2022). Although some studies have mentioned differences in the relative importance of the institutional predictors of entrepreneurship (Aparicio et al., 2016; Audretsch et al., 2021; Mickiewicz et al., 2021; Valdez & Richardson, 2013), dominance is usually not the focus and typically relies on comparing significance and coefficient size which has well documented limitations (Azen & Budescu, 2003).

Secondly, using the PDPs to examine the relationships between individual predictors and TEA highlights considerable non-linearity. For example, floor and ceiling effects can be observed in the predictive effect of entrepreneurial intentions and business angels. Only a small body of literature has focused on potential non-linearity in the institutional determinants of entrepreneurship (Audretsch et al., 2019; Chowdhury et al., 2019), and these studies have drawn on regression-based approaches. In contrast, our machine learning approach is able to model more complex non-linearities than traditional regression-based approaches.

Thirdly, from a predictive perspective, the results show that the more complex random forest model has the highest level of accuracy. This further suggests more complex non-linearities in the relationships which can be modelled more accurately using machine learning techniques. From a practical perspective, policy makers can consider the accuracy of the model when designing policy interventions.

This study also contributes to practice through the insights gained and from the application of ML to predict entrepreneurship levels, and the XAI techniques to understand how predictions were made. It is important for policy makers to understand the extent to which institutional factors are predictive of entrepreneurship so that policy functions as expected in achieving the desired outcomes (Nikolaev et al., 2018). The dominance analysis and visualisations of the individual variable relationships can help to inform policy decisions about where to focus targeted interventions to increase entrepreneurship. At an individual
country-level, XAI can be used to explain how predictions were made. The use of LIME provides explanations to policy makers about how predictions were made at the individual country-level, providing the insight needed to build trust and justify decisions. This element of the study helps us to bridge the academic practitioner gap (Lèvesque et al., 2020; Steffens et al., 2014), enhancing the practical relevance of the study.

This paper proceeds as follows. Section 2 reviews the relevant literature on the institutional determinants of country-level entrepreneurship as well as the literature on the analytics paradigm. This section also discusses the differences between the ML approach and traditional regression-based approaches, and highlights how ML can help to address entrepreneurship research questions particularly in areas where traditional approaches face limitations. Section 3 discusses the data, algorithms, and XAI methods that were used. This is followed by the results and discussion in sections 4 and 5, where we present and discuss the machine learning models and their interpretation. The paper then concludes in section 6 by summarising the main contributions, recommendations, and limitations.

2. Background

2.1. Institutional theory

Institutions make up the ‘rules of the game’ that guide behaviour within a society (North, 1990). North (1990) distinguishes between formal and informal institutions, with formal institutions focusing on rules and regulations, and informal institutions focusing on social norms and culture. Both types of institutions, along with their enforcement mechanisms, shape all types of individual behaviour within society by providing incentive structures and conferring legitimacy on appropriate behaviour (North, 1990, Scott, 1995). Scott (1995) further distinguishes between three institutional pillars: cultural-cognitive; normative; and regulatory, with the cultural-cognitive and normative pillars focusing on informal institutions, and the regulatory pillar focusing on formal institutions (Urbano et al., 2019). Taken together these institutional pillars provide legitimacy towards behaviours that are in line with the regulations, norms, and cultural cognitions in a society (Valdez & Richardson, 2013), thus shaping individual decision making and behaviour.

In this study, we draw on Scott’s (1995) conceptualisation of institutional pillars as a way of structuring and interpreting the individual dimensions that make up the institutional pillars. This approach allows us to consider the dominance and heterogeneity of the factors that constitute the three pillars, as well as providing a higher-level view of the dominance and heterogeneity of the overarching pillars, allowing us to examine how institutional dimensions fit together in influencing entrepreneurship. Drawing on Scott’s (1995) three pillars framework also allows us to link back to a rich body of research on the role of institutional pillars in country-level entrepreneurship (Audretsch et al., 2019; Busenitz et al., 2000; Chowdhury et al., 2019; Stenholm et al., 2013; Urbano & Alvarez, 2014; Valdez & Richardson, 2013). Although this past work has drawn on traditional regression-based methods, it provides a strong foundation when considering next generation questions around dominance and heterogeneity.

In shaping all types of social behaviour, institutions also influence entrepreneurial behaviour across the full spectrum of entrepreneurial activities spanning from opportunity recognition, through to starting and running a business (Valdez & Richardson, 2013), and ultimately business discontinuation (Beynon et al., 2021). Institutions provide the stability and incentives needed to motivate entrepreneurship (Stenholm et al., 2013). Past empirical research provides evidence to support the importance of institutions in fostering country-level entrepreneurship (Busenitz et al., 2000; Stenholm et al., 2013; Urbano & Alvarez, 2014; Valdez & Richardson, 2013).

This body of research highlights the overall importance of institutions to entrepreneurship, but also provides a more nuanced set of findings, suggesting heterogeneity in the relative importance of institutions as well as conditions under which institutions are likely to be more or less important in entrepreneurship (Audretsch et al., 2022; Busenitz et al., 2000; De Clercq et al., 2010). There have, however, also been conflicting findings, with some studies reporting evidence for the importance of all three pillars (Urbano & Alvarez, 2014) and others reporting evidence for the importance of only one or two pillars (Stenholm et al., 2013). Other studies have viewed institutions as playing a moderating relationship. For example, De Clercq et al. (2010) explores the role of Scott’s (1995) three pillars in moderating the relationship between networks and entrepreneurship, finding evidence that networks are more important when in contexts with weaker institutions. More research is required to investigate the conflicting and nuanced findings.

2.1.1. The cultural-cognitive pillar

The first pillar of Scott’s (1995) framework is the cultural-cognitive pillar, which can be defined as ‘the collective understandings of the makeup of social reality that allows for the framing of meaning within a society’ (Valdez & Richardson, 2013). This pillar builds on ideas from individual level psychology around entrepreneurial traits and cognitions (Valdez & Richardson, 2013). Entrepreneurial cognitions include shared perceptions around skills for starting a business; perceived opportunities to start a business; fear of failure; and knowing other entrepreneurs (Audretsch et al., 2021; Beynon et al., 2021; Busenitz et al., 2000; Siu & Lo, 2013; Stenholm et al., 2013; Valdez & Richardson, 2013; Xie et al., 2021). These cognitions are shared amongst groups of people, resulting in shared perceptions and attitudes (Busenitz et al., 2000; Valdez & Richardson, 2013). They then influence how individuals interpret their environment (Stenholm et al., 2013) as well as shaping aggregate group level behaviour (Valdez & Richardson, 2013). In this way, cultural-cognitive institutions also influence decisions to engage in entrepreneurship, and overall national levels of entrepreneurship (Stenholm et al., 2013; Urbano & Alvarez, 2014; Valdez & Richardson, 2013).

Past empirical work highlights the role of cultural-cognitive institutions in entrepreneurial behaviour (Stenholm et al., 2013; Urbano & Alvarez, 2014; Valdez & Richardson, 2013). Urbano and Alvarez (2014) find that skills, fear of failure, and knowing an entrepreneur are all important cognitive factors in predicting entrepreneurship. Valdez & Richardson (2013) use a composite measure, consisting of knowledge and skills and fear of failure, also finding a positive relationship with entrepreneurship. Aparicio et al. (2016) find that confidence in skills is related to entrepreneurship. Adopting a predictive analytics approach using Bayesian networks, Sohn & Lee (2013) find that country-level attitudes predict future entrepreneurial behaviour.

2.1.2. The normative pillar

The second aspect of Scott’s (1995) institutional framework is the normative pillar, which includes the country’s values and norms (Audretsch et al., 2021; Busenitz et al., 2000; Valdez & Richardson, 2013), with values focusing on what society believes is important, and norms focusing on acceptable behaviours (Valdez & Richardson, 2013). The normative pillar is closely linked to the concept of entrepreneurial culture (Busenitz et al., 2000) which can be defined as shared ‘values, beliefs, and expected behaviours’ that motivate individual entrepreneurial behaviour (Hayton & Cacciotti, 2013, p. 709). Entrepreneurial culture can be viewed as an informal institution focusing on shared norms and values which provide legitimacy and shared meaning towards entrepreneurial behaviour (Stuetzer et al., 2013). Normative institutions persist over a longer period of time compared with formal institutions which are embedded within the culture and change more frequently (Fritsch et al., 2017). Highlighting the persistence of culture, Fritsch et al. (2017) find that the entrepreneurial culture in 1925, measured as self-employment levels, influences start up activity 50 years later.

Entrepreneurial culture has been conceptualised and measured in a
variety of ways in the literature, such as drawing on Hofstede’s National cultures (Marino et al., 2010), or profiles of Big Five personality traits (Stuetz et al., 2018). However, Busenitz et al. (2000) argue that these broad conceptualisations of culture are overgeneralised, and therefore proposes a more specific measure of the entrepreneurial aspects of culture which they adopt in their institutional theoretic framework. This early work, along with more recent studies have conceptualised the normative institutional pillar as consisting of societal perceptions such as whether entrepreneurship is considered a good career choice; whether entrepreneurs have a high level of status and respect in society; and whether successful entrepreneurs are prevalent in the media (Beynon et al., 2021; Maurer et al., 2022; Stenholm et al., 2013; Turro et al., 2014). Here a distinction can also be made between the cultural-cognitive pillar and the normative pillar, with Valdez & Richardson (2013, p. 1157) stating that ‘normative elements are broader and more collective social pulses of what is acceptable; the cultural-cognitive elements are aggregates of concepts and beliefs that drive individuals’. 

In terms of empirical support for the importance of the normative pillar, past research has tended to find positive relationships between the normative pillar and entrepreneurial behaviour. For example, Aleksandrova & Verkhovskaya (2016) report positive impacts from the perception of entrepreneurship as a successful career choice and the status of entrepreneurs in society. Urbano & Alvarez (2014) further confirm that a higher degree of media attention for new businesses has a positive and statistically significant impact on entrepreneurship. Alvarez et al. (2011) report that these cultural and social norms, which they identify as informal institutions, have a stronger impact on entrepreneurial activity than formal institutions. Arabiyat et al. (2019), however, report that these cultural, or normative impacts, only hold for more basic forms of entrepreneurship rather than the more innovative type.

2.2. Entrepreneurial behaviour and intentions

Finally, we consider wider entrepreneurial activity including the extent of current and established business ownership, entrepreneurial intentions, business exit, and business angel investment. At the country-level, business ownership has been found to be predictive of nascent entrepreneurship (Sohn & Lee, 2013) while business exit has been found to be important in entrepreneurs’ growth aspirations (Auto & Pathak, 2010), and in positively predicting entrepreneurial intentions (He et al., 2020) and future entrepreneurial activity (Albiol-Sánchez, 2016). De Clercq et al. (2012) find that at the macro level new business opportunities are positively related to micro-angel investment activity. At the individual level, those who have acted as business angels have themselves been found to be more likely to become an entrepreneur (Arafat et al., 2020). Entrepreneurial intentions have also been identified as important antecedents of future entrepreneurial activity (Beynon et al., 2016; Krueger et al., 2000).

2.3. The analytics paradigm and machine learning in entrepreneurship research

Although the ML approach provides substantial opportunities to derive new insights about the role of institutions in entrepreneurship, it also differs in fundamental ways compared with the traditional statistical modelling approach. We synthesise and elaborate on the key differences between the two approaches across five dimensions encompassing the overarching paradigm, data, algorithms, model

Table 1

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Statistical Modelling</th>
<th>Machine Learning</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paradigm</td>
<td>Positivist, deductive, data modelling</td>
<td>Analytics paradigm</td>
<td>(Breiman, 2001b; Delen &amp; Zolbanin, 2018; Kitchin, 2014; Oswald &amp; Putka, 2017; Tonidandel et al., 2018; Valizade et al., 2022)</td>
</tr>
<tr>
<td>Data</td>
<td>Small, structured datasets; usually megabytes of data across tens of variables and hundreds or thousands of rows.</td>
<td>Large or small datasets which can be structured or unstructured. Large datasets can be tera- or petabytes across thousands of variables with billions of rows.</td>
<td>(Breiman, 2001b; Choudhury et al., 2021; George et al., 2014; Hair et al., 2021; Sarstedt, 2021; Tonidandel et al., 2018)</td>
</tr>
<tr>
<td>Algorithms</td>
<td>Statistical Modelling, often regression-based</td>
<td>ML algorithms (e.g. decision trees, artificial neural networks, support vector machines, K-nearest neighbours)</td>
<td>(Oswald &amp; Putka, 2017; Putka et al., 2018; Tonidandel et al., 2018)</td>
</tr>
<tr>
<td>Evaluation</td>
<td>Model fit and residual examination</td>
<td>Predictive accuracy</td>
<td>(Breiman, 2001b; Yarkoni &amp; Westfall, 2017)</td>
</tr>
<tr>
<td>Interpretation</td>
<td>Sign, effect size and statistical significance using coefficients and p-values</td>
<td>Dominance/relative importance and relationships, using XAI such as permutation importance and PDPs.</td>
<td>(Choudhury et al., 2021; Molnar, 2022)</td>
</tr>
</tbody>
</table>

Richardson (2013) do not find sufficient evidence to support a relationship between the regulatory pillar and entrepreneurship, rather finding that cultural-cognitive and normative institutions are more important determinants of entrepreneurship.
evaluation, and interpretation, as summarised in Table 1. By doing so, we develop and build on past work on the analytics paradigm, within which we situate our study (Delen & Zolbanin, 2018).

2.3.1. The analytics paradigm

Past work has highlighted differences in the overarching paradigm within which traditional statistical studies are carried out, compared with the overarching paradigm for ML studies. Contrast the two approaches, Breiman (2001b) refers to ‘two cultures’, with the traditional statistical culture focusing on ‘data models’ with functional forms specified by the researcher and the ML culture focusing on ‘algorithmic models’ which are learned from the data itself. In the wider social sciences, Chang et al. (2014) highlight the potential for a paradigm shift from the traditional approach to a computational social science approach. Focusing on the use of analytics in business and management research, Delen & Zolbanin (2018) posit the emergence of a new ‘analytics paradigm’.

These perspectives point to fundamental differences between the traditional statistical modelling approach and ML approaches. Traditionally, quantitative studies focus on statistical inference and testing hypotheses about relationships between variables derived from theory (Delen & Zolbanin, 2018), whereas ML studies focus on prediction and identifying complex patterns and relationships in the data, without the requirement for hypothesis testing (Delen & Zolbanin, 2018).

Compared with the traditional scientific method, studies carried out using ML techniques adopt different epistemological positions, where knowledge is learned from data rather than through a theory testing approach (Delen & Zolbanin, 2018; Kitchin, 2014; Tonidandel et al., 2018). Although ML approaches usually aim to maximise predictive accuracy they can also be used to develop and improve theory (Putka et al., 2018). The aim is often to identify new relationships and patterns in the data, which can contribute to theory development (Delen & Zolbanin, 2018), through ‘algorithm supported induction’ (Shrestha et al., 2021, p. 856). These new patterns and theories can also be further tested using the formal deductive hypothesis testing approach (Choudhury et al., 2021; Delen & Zolbanin, 2018; Lévesque et al., 2020; Ludwig & Mullainathan, 2023).

2.3.2. Data for machine learning

ML algorithms are effective at modelling both small and large datasets, which can be structured or unstructured (Breiman, 2001b; George et al., 2014; Hair & Sarstedt, 2021). Research in entrepreneurship, and management more generally, typically focuses on a relatively small set of variables, often combining variables to capture underlying constructs (Hair & Sarstedt, 2021). This can help to simplify models and reduce item level measurement error (Putka et al., 2018). However, combining items into a single scale potentially hides item level variance and more complex relationships that may be useful for prediction and theory (Putka et al., 2018; Woodside, 2013). In contrast to the traditional approach in which a relatively small number of variables are included in regression models, ML algorithms often perform better when more variables are included (Breiman, 2001b). Many ML methods can handle a large number of variables effectively, for example through feature selection and dimensionality reduction (Kuhn & Johnson, 2019). These approaches are particularly important when working with big data with high dimensionality (Fan et al., 2014). In addition to the size of datasets, ML approaches have also been developed to analyse and make predictions based on unstructured data, such as text, video, and images (Hair & Sarstedt, 2021), opening up additional possibilities to utilise large and novel data for entrepreneurship research.

2.3.3. Algorithms

Traditional statistical modelling studies also differ from ML studies in terms of the algorithms used to model the data, and how these models are built. Traditional statistical approaches to modelling data tend to focus on regression-based techniques (Woodsdie, 2013). In contrast, the ML approach adopts a wide range of algorithms to model the data, with the choice of algorithm depending on the specific problem and the level of accuracy achieved (Delen & Zolbanin, 2018). ML algorithms are continuously evolving in an attempt to increase predictive accuracy, which has resulted in thousands of available algorithms (Pedro Domingos, 2012).

Work on the development of improved ML algorithms has led to variations based on traditional regression algorithms, as well as the development of different approaches such as decision trees, support vector machines, K-nearest neighbours, and artificial neural networks (Kuhn & Johnson, 2019). It is also common for algorithms to be based on ensembles of models, where many ML models are combined to generate more accurate predictions (Tonidandel et al., 2018). This innovation in algorithm development, stemming largely from computer science research, has resulted in a wide range of techniques that hold promise for the entrepreneurship field. Moreover, it is common in the ML approach to apply multiple algorithms to a particular dataset to find the algorithm that results in the model with the highest predictive accuracy (Kuhn & Johnson, 2013; Tonidandel et al., 2018). This approach has been adopted in a small number of entrepreneurship studies (Jabeur et al., 2021; Schade & Schumacher, 2023).

An important advantage of ML algorithms is that they can effectively model both linear and non-linear relationships, as well as complex interactions amongst variables (Tonidandel et al., 2018). Any non-linearity and interactions between variables does not have to be specified in advance as the algorithm learns these relationships from the data. This is useful in cases where there is uncertainty about the form of the relationship between the independent and dependent variables (Putka et al., 2018). With a traditional statistical modelling approach the functional form of the model is specified in advance (Molnar et al., 2020). In contrast, with the ML approach the model that best fits the data is found by tuning the model hyperparameters using cross validation, which selects the optimal model, usually based on comparing accuracy across multiple models (Kuhn & Johnson, 2013; Molnar et al., 2020). This approach could help to reduce model uncertainty, which results from researchers uncertainty over the choice of variables to include and the functional form of the model (Arin et al., 2015; Salimans, 2012).

2.3.4. Model evaluation

The traditional statistical modelling approach focuses on evaluating the model by measuring goodness of fit on the sample data, whereas the ML approach focuses on evaluating the predictive accuracy of the model based on a holdout test dataset that is not involved in the model building process (Breiman, 2001b; Yarkoni & Westfall, 2017). This process usually involves splitting the data into a training and test set. Models are built on the training set and tested on a holdout test dataset that is not used in the model building process, thus providing a more objective measure of model performance (Kuhn & Johnson, 2013). Evaluating models based only on the sample of data used to build the model risks overfitting the model to the data, which means that it will fit the sample data well, but will not generalise well to new data and will have lower predictive accuracy (Yarkoni & Westfall, 2017). Although this approach to evaluating models focuses on predictive accuracy, it is also useful as a means of empirically evaluating the predictive accuracy and quality of a theory and in comparing the accuracy across multiple competing theories (Breiman, 2001b; Putka et al., 2018). The use of test and validation sets can also help to prevent against p-hacking by safeguarding against the overfitting that occurs during p-hacking (Choudhury et al., 2021; Yarkoni & Westfall, 2017). The metrics used to assess the accuracy of machine learning models also differ from the traditional approach, with the former focusing on measures such as root mean squared error (RMSE) and mean absolute error (MAE) for numerical outcomes, and by assessing correct and incorrect predictions, kappa, sensitivity and specificity for categorical outcomes (Kuhn & Johnson, 2013).

One limitation of current literature adopting the traditional
The importance of capturing more complex non-linearities between institutions and entrepreneurship has been recognised in recent literature, which argues that institutions can influence entrepreneurship in non-linear ways (Audretsch et al., 2022; Chowdhury et al., 2019). A small number of studies have utilised techniques to assess non-linearity in the institutional determinants of entrepreneurship. Drawing on public interest and public choice theory Audretsch et al. (2019) propose a non-linear inverted U-shaped relationship between national regulations and entrepreneurship at the city level, arguing that excessive regulation could act as a barrier to entrepreneurship, whereas effective regulation encourages entrepreneurship. They also argue that this non-linearity could exist across multiple types of regulation. Although not all of their hypotheses were supported, their findings highlight non-linear relationships between regulatory institutions and entrepreneurship. Chowdhury et al. (2019) use predictive margins to visualise non-linearity, highlighting non-linearity in some institutional factors such as entrepreneurial capital. Using regression-based approaches, Mohamadi et al. (2017) find non-linear relationships between corruption and entrepreneurship. Although these studies have identified the need to consider non-linearity, our ML approach goes beyond traditional regression-based techniques, enabling us to model and visualise more complex non-linear relationships.

The second aspect of the interpretation of ML models involves assessing the relative importance or dominance of variables. In the context of our study this centres on assessing the relative importance of the variables that make up the institutional pillars. Although few studies have directly considered dominance, some work has identified differences in statistical significance of the institutional pillars. Highlighting these nuances, Busenitz et al. (2000) propose that the three institutional pillars are important in different ways, with the normative pillar encouraging individuals to start a business, and the regulatory and cognitive pillars leading to success of the business. Valdez & Richardson (2013) find that the cultural-cognitive and normative institutional pillars are more important than the regulatory pillar. Aparicio et al. (2016) also find that informal institutions are more important in determining entrepreneurship than formal institutions. In contrast, Stenholm et al. (2013) find support for the importance of the regulatory pillar but not the cultural-cognitive or normative pillars. Mickiewicz et al. (2013) apply post estimation analysis, finding that the positive and negative changes in institutions have asymmetric effects on entrepreneurship, with deterioration of institutions having a greater effect than improvements. They also find differences in the relative importance of the rule of law and business regulation. Although adopting different approaches and different sets of institutional variables to the present study, this past work highlights the usefulness of studying the dominance of institutional factors. The contrasting findings identified in past studies highlight the need for more research to understand the most important institutional factors.

These conflicting findings over the importance of different
institutional determinants of entrepreneurship could be the result of model uncertainty, which arises due to uncertainty about the choice of variables to include in the model and the functional form of the model (Arin et al., 2015; Salimans, 2012). Although traditional regression-based studies typically ignore issues of model uncertainty (Nikolaev et al., 2018), addressing it can help to identify the most robust predictors of entrepreneurship (Arin et al., 2015). ML can help to overcome issues of model uncertainty, both around which variables to include, as well as the functional form of the model. Uncertainty over functional form is reduced by learning the model structure directly from the data rather than specifying it in advance. Uncertainty over which variables to include is reduced in algorithms which include automatic feature selection, such as tree-based approaches, which retain only those variables that improve the model fit (Kuhn & Johnson, 2013), thus reducing the need for the researcher to decide which variables to include. ML algorithms can also be used to generate measures of relative variable importance, contributing to the dominance analysis approach (Graham & Bonner, 2022). In contrast, most of the past studies assessing the importance of predictors have used regression-based approaches with a focus on measures of statistical significance and coefficient size when evaluating importance, which have well-documented limitations (Azen & Budescu, 2003).

3. Methodology

3.1. Data

The study draws on country-level time-lagged data from the GEM project from 2004 to 2019. This was matched to data from the IEF (Heritage, 2022) and GDP (World Bank, 2022). The final dataset for analysis comprises a total of 573 observations at country-year level. The use of country-year observations is a common approach to developing predictive models in the wider literature (Andres et al., 2021; Ettensperger, 2020). Data from 81 different countries are included in the final dataset. Each country-year observation consists of the independent variables measured at time t and the dependent variable measured one year later at time t + 1. Therefore, for a country to be included at least two years of data is required as well as complete observations across the variables. The GEM data was collected through the adult population survey at the individual level. For each question, respondents were asked whether or not they agreed with the statement. Weighted responses are then aggregated to country-level to give the population estimate.

Table 2 presents the descriptions of each variable. The dependent variable for the study is TEA, which is defined as the percentage of the adult population who own and manage a business less than 42 months old. Twenty-three independent variables are included in the study relating to each of the three institutional pillars as well as past entrepreneurial behaviour and intentions, and GDP. Drawing on past studies (Stenholm et al., 2013; Urbano & Alvarez, 2014; Valdez & Richardson, 2013) the cultural-cognitive pillar consists of perceived knowledge and skills to start a business, knowing another entrepreneur, fear of failure, and perceived opportunities for entrepreneurship. Previously, Stenholm et al. (2013), Urbano & Alvarez (2014), and Valdez & Richardson (2013), conceptualised the normative institutional pillar as consisting of dimensions including the perceived status of entrepreneurs, desirability of entrepreneurship as a career choice, and media coverage of entrepreneurship. To measure the regulatory pillar we draw on nine of the twelve dimensions from the IEF, which focus on areas including government size, rule of law, regulatory efficiency and open markets (Heritage, 2022). Three dimensions were not included due to high levels of missing data. The IEF has been used in past research to capture the regulatory pillar (Stenholm et al., 2013; Valdez & Richardson, 2013). In addition to the core institutional pillars, we include measures of past entrepreneurial behaviour including established business ownership, owner-managers, early-stage entrepreneurship, business angels,

| Table 2
| List of Variables. |
|-------------------|-------------------|
| Variable          | Name              | Description                                      |
| Dependent Variable: |                  | Percentage of the adult population engaged as owner and manager of a nascent or young firm (< 42 months old), at time t + 1. |
| Cultural-Cognitive Pillar: |      | Percentage of the adult population who said fear of failure would prevent them from starting a new business. |
| Media Influence: |                  | Percentage of the adult population who know someone personally who started a business in the past 2 years. |
| Regulatory Pillar: |                  | Percentage of the adult population who believe that in their country, those successful at starting a new business have a high level of status and respect. |
| Financial Freedom: |                  | Percentage of the adult population who think that in the next 6 months there will be good opportunities for starting a business in the area where they live. |
| Past Behaviour and Intentions: |      | Percentage of the adult population who believe that in their country, you will often see stories in the public media about successful new businesses. |
| Estbus: | Established business | Percentage of the adult population who are, alone or with others, expecting to start a new business, including any type of self-employment, within the next three years. |
| Owner: | Business owner | Percentage of the adult population who are, alone or with others, currently the owner of the business. |

(continued on next page)
were centred and scaled (Kuhn sensitive to the underlying measurement scale of predictors, variables features from the data, which are included in a linear regression model when generalising outside of the sample (Field et al., 2012; Woodside, 1995). The observations in the training data (Kuhn) were chosen as they provide a representative view of commonly applied methods. The analysis follows a typical ML workflow (e.g., Kuhn & Johnson, 2013). The full dataset is first split into a training set, consisting of 80% of the observations (n = 461), and a test set consisting of 20% of the observations (n = 112). To help ensure the underlying distribution of the dependent variable is maintained in the training and test sets, the splitting procedure randomly samples from within quintiles. All ML models are built using the training set and tested using the test set. Eight ML algorithms were applied: linear regression (LR); multivariate adaptive regression splines (MARS); recursive partitioning; CUBIST; random forests (RF); gradient boosted machines (GBM); K-nearest neighbours (KNN); and support vector machines (SVM). These eight algorithms were chosen as they provide a representative view of commonly applied algorithms, demonstrating regression-based approaches, decision trees, and ensembles of trees. The algorithms that were used are briefly discussed below.

3.2.1. Regression, SVM and KNN algorithms

Regression methods are commonly applied in entrepreneurship research (Coduras et al., 2016), with linear regression used to model data where the dependent variable is continuous (Field et al., 2012). The algorithm fits a hyperplane to the data, which is described by an intercept and coefficients. Linear regression has the advantage of being easily interpretable. However, it assumes a linear relationship between the independent and dependent variables, and has stringent assumptions when generalising outside of the sample (Field et al., 2012; Woodside, 2013). MARS is a regression-based technique that generates surrogate features from the data, which are included in a linear regression model (Kuhn & Johnson, 2013). This allows non-linear patterns to be modelled. The algorithm incorporates feature selection on the surrogate features so that unimportant features are not included in the model. The SVM, proposed by Corinna and Vapnik (1995), uses a different cost function to linear regression to fit the hyperplane. With the SVM the hyperplane is not found by minimising the total sum of squared errors, but is fit based on cases with higher residuals (support vectors). KNN makes predictions based on the average outcomes of neighbouring observations in the training data (Kuhn & Johnson, 2013). As KNN is sensitive to the underlying measurement scale of predictors, variables were centred and scaled (Kuhn & Johnson, 2013).

3.2.2. Tree based algorithms

Recursive partitioning (Therneau & Atkinson, 2015) is a decision tree algorithm which models the data by recursively splitting the data into increasingly homogenous groups based on the dependent variable. The algorithm begins with a root node which contains all of the data, before evaluating the improvement in node purity that would be gained by splitting on each independent variable. The algorithm splits the data based on the variable that maximises the node purity of the resulting child nodes. The procedure is repeated at each child node until a stopping criteria is reached. The model’s complexity parameter controls the size of the tree.

3.2.3. Ensembles of trees

Tree-based ensembles combine predictions from multiple decision trees, with the aim of increasing model accuracy. Three tree-based ensembles were applied: GBM, RF, and CUBIST. The GBM algorithm, developed by Friedman (2001), starts by building a single decision tree. It then builds subsequent decision trees which aim to address errors from the previous tree. The algorithm has two tuning parameters: the tree depth, and the number of trees. The RF algorithm, proposed by Breiman (2001a), builds multiple decision trees on different subsets of the training data and calculates the average of the predictions across the single trees. CUBIST adds a smoothing step to individual trees by using a linear regression model in terminal nodes (Kuhn & Johnson, 2013). To create an ensemble of trees, CUBIST uses a variation of the approaches used in GBM and RF, called committees, which implements a weighted-based approach alongside averaging predictions across trees.

3.2.4. Hyperparameter tuning and performance evaluation

For all algorithms, hyperparameters were tuned on the training data using five-fold cross validation repeated ten times. The cross-validation procedure first splits the training data into five equal parts, and then builds the model on four of the parts and tests the performance on the fifth part. This process is repeated for all combinations of the tuning parameters. The forecasting performance was assessed by applying the model to make predictions on the test data. This provides a more objective measure of performance compared to making predictions on the training data. Three performance measures are considered and reported: root mean squared error (RMSE), mean absolute error (MAE) and R-squared. The cross validation and testing approach helps to prevent model overfitting. All modelling was carried out using R and the caret package (Kuhn, 2017).

3.3. Model interpretation using XAI

A key aim of our study is to gain insight into the relationship between institutional factors and TEA by utilising XAI approaches to interpret the machine learning models. This allows us to contribute to the wider theory by identifying both the relative importance of the predictors and elucidating their non-linearity. To do so, we draw on both global and local model-agnostic XAI approaches. Global approaches focus on the relative importance of the independent variables in making predictions. Traditional approaches to calculate variable importance have focused on the size of coefficients in regression models, and the contribution of variables to improving model accuracy in tree-based approaches (Kuhn & Johnson, 2013). The former approach suffers from the limitation that it is influenced by the underlying variable measurement scale, whilst the latter approach has been found to be biased (Strobl et al., 2007). We therefore use permutation variable importance as the first global model-agnostic approach to generate variable importance scores (Fisher et al., 2019). This approach works by permuting each variable in turn and measuring the increase in model mean squared error, thus providing a measure of each variable’s contribution to the model’s predictive accuracy.

The second global model-agnostic approach that we use are PDPs. These enable a better understanding of the relationships between the
explained. The output of this approach can be visualised to show how the prediction was made for an individual observation.

4. Results

This section presents the results of the analysis, focusing on the overall model accuracy, dominance and non-linearity using global model interpretation, and individual country-level explanations using local model interpretation. Table 3 presents the descriptive statistics for the dataset. Across all country-year observations, an average of 10.96% of the working age population reported being engaged in TEA, with a minimum of 1.90% and a maximum of 74.23%. The cultural-cognitive variables of knowing an entrepreneur, opportunity perception, and fear of failure have similar averages, with just under 40% of the population responding positively to these questions. Having the knowledge, skills and experience to start a business is somewhat higher, at 47.95%. The variables making up the normative pillar range from an average of just below 60% for media attention, to 63.04% for entrepreneurship as a good career choice, to a high of just under 70% agreeing entrepreneurs receive high status. The four regulatory dimensions range from an average score of 54.65 for government spending, to 80.34 for trade freedom. Across the behavioural variables, the average business ownership is 16.56%, established business ownership is 7.46%, business discontinuity is 2.92%, and average scores of 54.65 for government spending, to 80.34 for trade freedom. This highlights the improved accuracy of the more complex random forest compared to simpler regression and single tree models.

Additional analyses were carried out to provide more insights from a theoretical perspective, rather than the purely predictive perspective. We repeat the analysis, but drop the TEA and business ownership as independent variables, as it is possible that there is some overlap due to the time periods over which the dependent variable and these independent variables are measured, even though it is very unlikely that the same individuals would be sampled at both time t and t + 1. The results of the model accuracy are presented in Table 5. The RF is again most accurate based on a RMSE of 2.725, and the RPART is the least accurate with a RMSE of 3.963. KNN also performs relatively well, with a slightly better MAE than the RF, but slightly worse RMSE and R-Squared. Across both sets of models, although the more complex models are more accurate, they are also more difficult to interpret when compared with the simpler regression and single tree-based models. This demonstrates the trade-off between model interpretability and predictive accuracy.

4.1. Model accuracy comparison

Table 4 presents the comparison of the model accuracy, which is evaluated by making predictions on the test data. The random forest model has the highest level of accuracy across all three metrics, with a RMSE of 2.392, an R-squared of 0.867 and a MAE of 1.700. Based on RMSE and R-squared, MARS has second highest accuracy with a RMSE of 2.497, R-squared of 0.864 and MAE of 1.802. Based on RMSE, KNN and CUBIST are found to be third and fourth most accurate, with the least accurate models being the SVM, RPART, and LM. This highlights the improved accuracy of the more complex random forest compared to simpler regression and single tree models.

Table 4

<table>
<thead>
<tr>
<th>Predictive accuracy for models including all variables.</th>
<th>RMSE</th>
<th>R-squared</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>2.392</td>
<td>0.867</td>
<td>1.700</td>
</tr>
<tr>
<td>MARS</td>
<td>2.497</td>
<td>0.864</td>
<td>1.802</td>
</tr>
<tr>
<td>KNN</td>
<td>2.543</td>
<td>0.847</td>
<td>1.743</td>
</tr>
<tr>
<td>CUBIST</td>
<td>2.571</td>
<td>0.855</td>
<td>1.849</td>
</tr>
<tr>
<td>GEM</td>
<td>2.588</td>
<td>0.850</td>
<td>1.901</td>
</tr>
<tr>
<td>LM</td>
<td>2.592</td>
<td>0.852</td>
<td>1.877</td>
</tr>
<tr>
<td>RPART</td>
<td>3.435</td>
<td>0.758</td>
<td>2.334</td>
</tr>
<tr>
<td>SVM</td>
<td>3.504</td>
<td>0.798</td>
<td>2.213</td>
</tr>
</tbody>
</table>

Table 3

Descriptive Statistics. *All GEM variables are measured as the percentage of the working age population who responded ‘yes’ to the statement in the GEM survey. GDP is measured in US dollars and the regulatory dimensions are measured on a scale from 0 to 100.

<table>
<thead>
<tr>
<th>Grouping Variable</th>
<th>Variable</th>
<th>Mean (%)</th>
<th>Standard dev.</th>
<th>Min (%)</th>
<th>Max (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Variable</td>
<td>TEA (t + 1)</td>
<td>10.96</td>
<td>6.80</td>
<td>1.90</td>
<td>39.91</td>
</tr>
<tr>
<td>Cultural Cognitive</td>
<td>Knowent</td>
<td>38.71</td>
<td>11.21</td>
<td>17.22</td>
<td>85.10</td>
</tr>
<tr>
<td></td>
<td>Oppert</td>
<td>39.74</td>
<td>16.18</td>
<td>2.85</td>
<td>85.54</td>
</tr>
<tr>
<td></td>
<td>Skill</td>
<td>47.95</td>
<td>13.92</td>
<td>6.55</td>
<td>87.93</td>
</tr>
<tr>
<td></td>
<td>Fruilt</td>
<td>38.70</td>
<td>9.66</td>
<td>13.15</td>
<td>72.35</td>
</tr>
<tr>
<td>Normative</td>
<td>Career</td>
<td>63.04</td>
<td>13.42</td>
<td>22.81</td>
<td>96.16</td>
</tr>
<tr>
<td></td>
<td>Status</td>
<td>69.14</td>
<td>10.34</td>
<td>31.47</td>
<td>95.20</td>
</tr>
<tr>
<td>Legislative</td>
<td>Media</td>
<td>59.42</td>
<td>13.96</td>
<td>19.22</td>
<td>88.00</td>
</tr>
<tr>
<td>Regulatory</td>
<td>PropertyRights</td>
<td>61.09</td>
<td>24.04</td>
<td>10.00</td>
<td>95.00</td>
</tr>
<tr>
<td></td>
<td>Government.Integrity</td>
<td>56.07</td>
<td>21.92</td>
<td>18.00</td>
<td>97.00</td>
</tr>
<tr>
<td></td>
<td>Tax.Burden</td>
<td>70.25</td>
<td>13.14</td>
<td>32.00</td>
<td>99.70</td>
</tr>
<tr>
<td></td>
<td>Government.Spending</td>
<td>54.65</td>
<td>23.51</td>
<td>0.00</td>
<td>95.30</td>
</tr>
<tr>
<td></td>
<td>Business.Freedom</td>
<td>74.28</td>
<td>12.97</td>
<td>37.30</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>Monetary.Freedom</td>
<td>78.04</td>
<td>7.19</td>
<td>44.00</td>
<td>94.30</td>
</tr>
<tr>
<td></td>
<td>Trade.Freedom</td>
<td>80.34</td>
<td>8.95</td>
<td>24.00</td>
<td>90.00</td>
</tr>
<tr>
<td></td>
<td>Investment.Freedom</td>
<td>64.95</td>
<td>20.43</td>
<td>0.00</td>
<td>95.00</td>
</tr>
<tr>
<td></td>
<td>Financial.Freedom</td>
<td>60.61</td>
<td>17.47</td>
<td>10.00</td>
<td>90.00</td>
</tr>
<tr>
<td>Entrepreneurial Behaviour</td>
<td>TEAprev</td>
<td>10.66</td>
<td>6.72</td>
<td>1.48</td>
<td>41.46</td>
</tr>
<tr>
<td></td>
<td>Owner</td>
<td>16.56</td>
<td>8.47</td>
<td>4.24</td>
<td>67.16</td>
</tr>
<tr>
<td></td>
<td>Estibus</td>
<td>7.46</td>
<td>4.60</td>
<td>0.42</td>
<td>37.74</td>
</tr>
<tr>
<td></td>
<td>Angel</td>
<td>4.99</td>
<td>3.62</td>
<td>0.32</td>
<td>31.28</td>
</tr>
<tr>
<td></td>
<td>Intention</td>
<td>21.11</td>
<td>14.43</td>
<td>1.09</td>
<td>72.23</td>
</tr>
<tr>
<td></td>
<td>Discont</td>
<td>2.29</td>
<td>2.35</td>
<td>0.38</td>
<td>19.88</td>
</tr>
<tr>
<td>GDP Per Capita</td>
<td>Gdp ($)</td>
<td>24754.56</td>
<td>21285.21</td>
<td>653.33</td>
<td>108307.89</td>
</tr>
</tbody>
</table>
highlighting the need for the XAI approach.

4.2. Global model interpretation: variable dominance

Variable importance scores were calculated using permutation importance to assess dominance, as discussed in the methodology. Fig. 1 shows the results of the variable importance for the RF model. The most important variable in predicting TEA is the previous year’s TEA rate. This highlights that current levels of TEA are a good indicator of the likely future levels. However, other variables also contribute towards improving the accuracy of the prediction. Of second highest importance are entrepreneurial intentions. This is followed by the cultural-cognitive and behavioural variables of current ownership levels, skills, and perceived opportunities. The regulatory dimensions of government spending and tax burden are also relatively important in sixth and seventh position. Four of the regulatory dimensions of investment freedom, property rights, government integrity, and trade freedom are found to be least important. The cultural-cognitive dimension of knowing an entrepreneur and the normative dimension of media perception are also relatively less important.

Fig. 2 shows the permutation variable importance scores for the RF model, excluding TEA and current business ownership at time t. This shows a broadly similar picture to the variable importance scores including all variables with intentions, skills, opportunities, and government spending the most important predictors. The least important predictors are knowing an entrepreneur, the perceived status of entrepreneurs, preceded by seven of the regulatory dimensions.

4.3. Global model interpretation: examining linearity using partial dependence plots

Fig. 3 shows the PDPs for the RF models for the nine most important independent variables, as identified through permutation variable importance. This shows the change in the average predicted value of TEA across different values of the independent variables. The first PDP shows the effect of the TEA rate at time t on the predicted TEA rate at time \( t+1 \). This shows a gradual increase in TEA until around 15% TEA rate, at which point there is a steeper incline. The second plot (intention) shows an increase in the predicted TEA rate as entrepreneurial

### Table 5

Predictive accuracy for models excluding TEA and current business ownership at time t.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>R-squared</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>2.725</td>
<td>0.825</td>
<td>1.952</td>
</tr>
<tr>
<td>KNN</td>
<td>2.735</td>
<td>0.822</td>
<td>1.875</td>
</tr>
<tr>
<td>GBM</td>
<td>2.802</td>
<td>0.815</td>
<td>2.036</td>
</tr>
<tr>
<td>SVM</td>
<td>2.935</td>
<td>0.794</td>
<td>2.007</td>
</tr>
<tr>
<td>CUBIST</td>
<td>2.943</td>
<td>0.815</td>
<td>2.085</td>
</tr>
<tr>
<td>LM</td>
<td>3.166</td>
<td>0.764</td>
<td>2.399</td>
</tr>
<tr>
<td>MARS</td>
<td>3.624</td>
<td>0.713</td>
<td>2.572</td>
</tr>
<tr>
<td>RPART</td>
<td>3.963</td>
<td>0.642</td>
<td>3.060</td>
</tr>
</tbody>
</table>

Fig. 1. Permutation variable importance for the RF model including all variables. The point shows the median, and the error bars show the 5% and 95% percentile across 100 iterations.
intentions increase, particularly where intentions are between around 10 % and 45 %, although this is much less prominent effect on the predicted TEA than the previous year’s TEA rate. The third PDP also shows an increase in predicted TEA as business ownership increases, particularly between 10 % and 30 % business ownership. The PDP focusing on skills, knowledge and experience shows an increase in the predicted TEA rate as skills increase, particularly above 50 %.

Fig. 4 presents the PDPs for the nine most important variables from the RF model excluding TEA and current ownership independent variables. The trends for intentions, skills, opportunities, and government spending, angel investing, discontinuity and TEA, although consistent with the dominance results the effect on predicted TEA is lower.

4.4. Local level model interpretation

The previous analyses provide both theoretical and practical insights into the findings, which are discussed further later, but here we demonstrate the application of LIME, which provides additional detail about how predictions were made. The aim of this approach is to fully explain predictions at the individual level, which will be of particular benefit to policy decision makers. We apply LIME to explain how individual country-level TEA predictions were made based on the most accurate RF model, which included TEA and business ownership at time t. Fig. 5 shows the contribution of all independent variables in predicting TEA for 20 country-year observations. Across these observations, TEA at time t is most predictive of TEA at t + 1, but there is also considerable variation in the factors that are generating the predictions at the country-year level. This is due to the underlying RF algorithm, which allows more complex and non-linear patterns to be modelled.
To further demonstrate the use of XAI techniques, we apply and visualise LIME for four country-year observations, focusing only on the five most important predictors for each observation. The results of this analysis are presented in Fig. 6. These visualisations further illustrate the heterogeneity of factors that lead to predicted levels of TEA. For example, in case 1, the predicted TEA level is 23.56%, and is based primarily on a past TEA rate of <13.39%. Intentions of <27.9%, and business ownership of <20.3%, skills of <55.9 and opportunity of <49.4% also contribute positively towards the predicted TEA rate. In contrast, in case 4 the predicted TEA rate is much lower, at 6.19%, largely due to a lower TEA rate at time t of <=5.70%. The other four variables have also contributed negatively towards the predicted TEA rate, including intentions of <=10.2%; skills of <=38.2%; government spending of <=37.1 and angel investing between 2.72% and 3.87%. These visual explanations of each predictor can be provided to policy makers to provide interpretation and insight into how each prediction was made.

5. Discussion

The findings from this study highlight the contributions that the machine learning approach can make to entrepreneurship research and practice across three key areas where traditional statistical modelling regression-based approaches face limitations: prediction, dominance analysis, and in modelling complex non-linearity. Our study complements past work which has drawn on regression-based approaches to identify the importance of institutions in entrepreneurship (Busenitz et al., 2000; Chowdhury et al., 2019; Stenholm et al., 2013; Urbano & Alvarez, 2014; Valdez & Richardson, 2013). Situating our machine learning study within this established theoretical framework helps to frame the analysis, guiding the selection and interpretation of data and allowing us to contribute to an existing body of research as well as providing a deeper understanding for practical decision making purposes (Hünermund et al., 2021).

Drawing on a machine learning methodology, eight algorithms were applied to model and predict entrepreneurship based on institutional dimensions, entrepreneurial behaviour and intentions. The more complex ‘black box’ random forest algorithm produces the most accurate predictions. This finding supports past work which has identified tree based techniques as more accurately modelling entrepreneurship (Jabeur et al., 2021; Schade & Schuhmacher, 2023). One potential explanation for this is that the relationships between institutional dimensions and entrepreneurship exhibit more complexity and non-linearities (Audretsch et al., 2019; Chowdhury et al., 2019) than can be captured using traditional regression-based models (Tonidandel et al., 2018). Although prediction is a less common aim in management research compared with theory testing, empirically testing the predictive accuracy of a model can be useful to understand how well the theory predicts future behaviour (Yarkoni & Westfall, 2017) as well as the practical utility of the model.

A commonly cited limitation of more complex ML models is that they are more difficult to interpret (Berger, 2023), which has led to recent
developments in XAI enabling insights into the relative importance of variables (dominance), the shape of relationships (non-linearity), and explanations about how predictions are made at the level of individual observations (Molnar, 2022). Utilising these techniques to draw insights from ML models also helps to overcome issues of model uncertainty faced by regression-based approaches (Nikolaev et al., 2018), by reducing the need for the researcher to make uncertain decisions about variable selection and functional form. Building on the small body of entrepreneurship literature that has taken a dominance analysis approach (Arin et al., 2015), permutation importance was used to examine the relative importance of predictor variables. The findings indicate the key importance of current levels of activity and intentions in predicting future TEA. This is consistent with wider research which indicates a ‘stickiness’ or path dependency with regards to business start-up activity, not least due to the specific history, resource endowments, physical and cultural attributes, institutions, past investments and social composition of a region which are slow to evolve (Bishop & Shilcof, 2017; Nyström, 2007). Fritsch & Wyrwich (2015) also identify the persistence of entrepreneurial activity, with its impact on rates far into the future. Intention as a predictor of future TEA is also consistent with wider findings which report a positive correlation between the two variables (Beynon et al., 2020) and show the link between intention and entrepreneurial behaviour or activity (Bogatyreva et al., 2019; Kautonen et al., 2013, 2015).

Although very few studies have adopted either a dominance analysis approach or machine learning approaches in entrepreneurship research, these findings are supported by a small body of past research. For example, focusing on predicting rural entrepreneurship, Hand et al. (2023) also find that the number of new firms in the prior year is the best predictor in both their random forest and gradient boosted models. Combined, these results point towards the importance of a general level of entrepreneurial activity in the country, and suggests the necessity of a strong entrepreneurial ecosystem to support business activity across the entrepreneurial pipeline (Autio et al., 2014).

It is particularly interesting to note, in our findings, that business discontinuity is also positively related to TEA, with a steady increase in TEA predicted for levels of discontinuity up to around 6% of the population. This highlights the importance of the dynamic process of business entry and exit and their interplay (Fok, 2009), for example, through competition or multiplier effects whereby firm entry can cause either further entry, or exit, depending on the impact on demand, and in terms of competition for the same resources (Johnson and Parker, 1996). This also corresponds with findings from the wider literature on the importance of business discontinuity for future entrepreneurial engagement and the entrepreneurial process (DeTienne, 2010; Hessels et al., 2011).

Our findings highlight important heterogeneity in the dominance of institutional dimensions across and within each pillar of Scott’s (1995) institutional theory framework. Looking at the dominance of the three pillars, the cultural cognitive pillar is most important overall, followed by the regulatory, and normative pillars. Although past work has tended to focus on statistical significance and coefficient size rather than relative importance/dominance analysis, some studies have pointed...
Fig. 5. LIME for the RF model for twenty cases.

Fig. 6. Example application of LIME for four cases.
towards differences in the importance of institutional pillars. For example, Aparicio et al. (2016) report findings that informal institutions are more important than formal institutions. Valdez & Richardson (2013) report that the regulatory pillar is relatively unimportant in new business creation, whilst Audretsch et al. (2021) report that all three pillars are important for productive entrepreneurship. Our results point to the regulatory and normative aspects being important for entrepreneurship but the role of individuals and the cognitive dimensions being more central to entrepreneurial activity.

Going beyond the dominance of the overarching pillars, there are also differences in the importance of specific dimensions within the pillars that should also be considered in institutional theory and in policy making. Within the cultural cognitive pillar, skills and opportunities are dominant, whereas fear of failure and knowing an entrepreneur are less important. This supports the findings of Sohn and Lee (2013) who apply Bayesian networks to predict TEA using the GEM data and also find entrepreneurial attitudes to be an important predictor at the national level. In particular, past research has identified the important role of perceived opportunities, and confidence in skills (Amorós et al., 2019; Aparicio et al., 2016; Boudreaux et al., 2019; Urbano & Alvarez, 2014) in entrepreneurship, linking to traditional theory such as Kirzner’s (1973, 1979) alertness to entrepreneurial opportunity and the well-studied concept of entrepreneurial self-efficacy which is found to be positively related to entrepreneurial outcomes (Glosenberg et al., 2022). Although some studies have found statistically significant relationships between fear of failure, knowing an entrepreneur and entrepreneurship (Amorós et al., 2019; Boudreaux et al., 2019; Urbano & Alvarez, 2014), we find these variables to be relatively less dominant. These findings substantiate the benefits of the use of the ML methodology which enables identification of the relative importance of variables, as opposed to traditional methods which can only report on effect size and statistical significance.

Within the normative dimension, perceived status of entrepreneurs is least important, with media portrayal and career perception slightly more important, but still outside of the most important predictors. Some past studies have found dimensions such as media portrayal (Audretsch et al., 2021) and high status (Maurer et al., 2022) to be insignificant predictors of entrepreneurship using traditional statistical tests, whilst others have found dimensions such as media portrayal to be significant (Boudreaux et al., 2022; Maurer et al., 2022; Urbano & Alvarez, 2014). The previous mixed findings may suggest that these factors are context-specific hence explaining our finding that they are relatively less important.

There is also considerable heterogeneity in the dominance of the regulatory dimensions, with government spending and the tax burden being relatively important compared with the other seven dimensions. These findings support past statistical studies that have found government spending and the tax burden to be significant determinants of entrepreneurship (Chowdhury et al., 2019) and may, for example, point to the opportunity cost of entrepreneurship, as opposed to other forms of economic activity, due to tax implications and/or the availability of employment opportunities depending on wider economic conditions. Furthermore, Dilli et al. (2018) report that the nature of institutional arrangements facilitates distinct forms of entrepreneurship and therefore the heterogeneity found in our study may relate to the generalised nature of the TEA variable which spans the range of entrepreneurial activity.

Although these findings highlight heterogeneity in the dominance of institutional dimensions, considerable non-linearity can be observed in the relationships between individual dimensions and entrepreneurship. Substantial differences in the predicted value of TEA were found across different ranges of behavioural and institutional factors, with floor and ceiling effects also observed.

These findings address calls from the wider business and management literature to consider more non-linear relationships (Arin et al., 2022; Woodside, 2013, 2014), and are supported by a small number of recent studies which have looked at non-linearity in the institutional determinants of entrepreneurship. Chowdhury et al. (2019) also report non-linear effects of entrepreneurial skills, showing that the effect of skills on their measure of productive entrepreneurship becomes negative when around 60% of the population report having entrepreneurial skills. Although our results show a levelling off at higher percentages of skills, there is a general positive trend. One explanation for this could be due to the difference in dependent variables, where a proportion of our entrepreneurs would be engaged in unproductive entrepreneurship. Audretsch et al. (2019) find U-shaped and inverted U-shaped relationships between national level regulations and city entrepreneurship. In contrast, we find generally positive effects, but with floor and ceiling relationships. Our study contributes to past work considering non-linearity (Audretsch et al., 2019; Chowdhury et al., 2019) by drawing on PDPs, which can show more nuanced and detailed non-linear patterns, and by including a different set of variables.

In addition to the theoretical contributions from examining dominance and non-linearity, a key aim of the study is in making a methodological contribution by highlighting the differences between the machine learning approach and the traditional data modelling approach, and consequently to the insights gained. The application of the machine learning method highlights that the use of traditional and machine learning approaches are not substitutes, but provide different insights. More generally, these differences in insights stem from differences across fundamental dimensions including the overarching paradigm (Delen & Zolbanin, 2018), the data used (Breiman, 2001b; Chowdhury et al., 2021), the algorithms and methods used to model the data (Putka et al., 2018; Tonidandel et al., 2018), and the evaluation and interpretation of the models generated (Breiman, 2001b; Chowdhury et al., 2021; Molnar, 2022). We agree with others who have viewed the machine learning approach and traditional regression-based approaches as complementary, and that entrepreneurship research will gain the most benefit when both approaches are used iteratively in the research process (Delen & Zolbanin, 2018; Tonidandel et al., 2018).

The analytics paradigm also places a strong emphasis on practical problem solving, and using data to derive value (Delen & Zolbanin, 2018). Addressing calls for practical relevance in the use of AI approaches in entrepreneurship research (Lèvesque et al., 2020), our study showcases the use of individual level explanations of future predictions. This allows policy makers to explain the rationale behind predicted levels of entrepreneurship, providing enhanced justification for decision making, and avoiding criticisms around the use of ‘black box’ algorithms for decision making, such as complexity and a lack of transparency.

Policy makers should be aware of differences in the relative importance of institutional factors. Policy initiatives aiming to increase entrepreneurship should focus particularly on enhancing cultural-cognitive institutions, with an emphasis on entrepreneurial skills and opportunity recognition. We also demonstrate an approach to predicting TEA, which can be implemented by policy makers, enabling evidence-based decisions to be made based on levels of forecast entrepreneurship. The XAI approach elucidates the factors that predict entrepreneurship, allowing policy makers to identify weaknesses in their entrepreneurial ecosystem and thus make appropriate interventions. Our approach helps to bridge the gap between entrepreneurship theory and practice, which is a key consideration for entrepreneurship research as a practice orientated field (Lèvesque et al., 2020).

6. Conclusion

The aim of this paper was to use XAI techniques to forecast and understand the institutional drivers of early-stage entrepreneurship, whilst also illustrating how these techniques can be used by both policy makers and researchers. We highlight the importance of cultural-cognitive institutions alongside entrepreneurial behaviour in predicting TEA, but also show that it is important to consider non-linearity in the relationship between these factors and TEA. In contrast to
expectations, normative and regulatory institutions are relatively less important determinants of entrepreneurship. Our analysis also goes further than much of the existing literature by utilizing XAI, as well as time-lagged TEA rather than a cross sectional approach. This leads to a more nuanced understanding of the relationship between institutions and TEA, which has both theoretical and practical implications. The use of XAI techniques, including permutation importance, partial dependency plots, and LIME, have opened the black box. This allows policy makers to understand how predictions are made, providing justification and confidence in the results. Researchers can also benefit from these approaches by gaining increased insight into the factors that contribute to TEA. Entrepreneurs can benefit from the practical application of explainable AI in their own businesses and technical innovations. With the increased attention around areas such as algorithmic bias (van Giffen et al., 2022), entrepreneurs and policy makers will be increasingly required to explain how predictions were made.

The study is not without limitations. Although TEA is a widely used measure of entrepreneurship, it does not capture all types of entrepreneurship, such as entrepreneurship within existing firms (Ács & Szerb, 2009). Future work could therefore consider applying these techniques to other areas of entrepreneurship research, such as predicting the formation of specific types of ventures or across specific sectors, or in predicting the future trajectory of the new businesses. Future studies could also consider incorporating additional data, which may be useful in increasing the predictive accuracy of the models. Another limitation of our study stems from the issue that some countries do not participate across all years, which reduces the size of the dataset.

The application of ML techniques in business and entrepreneurship research is an emerging area and we anticipate further developments, both on the technical side but also around further development of the analytics paradigm. Although our study aims to elucidate some of the key methodological considerations, more work is needed to explore the potential for the wide variety of machine learning approaches to contribute to entrepreneurship research. As the importance of ML grows, we believe that it can make substantial contributions to the entrepreneurship field.

CRediT authorship contribution statement

**Byron Graham:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.  
**Karen Bonner:** Writing – review & editing, Writing – original draft, Validation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**References**


