Research Bulletin 21/4 | Profiling Entrepreneurs using Machine Learning Approaches

Dr Karen Bonner, Ulster University Economic Policy Centre & Dr Byron Graham, Queen's University Management School

December 2021

Summary

Advances in analytics have created opportunities for researchers to apply machine learning techniques to address entrepreneurship research questions. This study aims to illustrate some of the opportunities available from the application of machine learning techniques to better understand entrepreneurial activity. Drawing on data from the Global Entrepreneurship Monitor (GEM), this study adopts a machine learning methodology to examine the relative importance of the determinants of entrepreneurial intentions and total early-stage entrepreneurship. The results show that across all models, perceptual variables such as having the skills to start a business, and knowing an entrepreneur are found to be relatively more important determinants, along with age. Cultural factors and other demographics are less important. Notably, the techniques highlight the complex interrelationships between factors and lack of a single set of characteristics to define entrepreneurs. Overall, total early-stage entrepreneurship can be modelled more accurately than entrepreneurial intentions, but it remains challenging to accurately predict either. The results contribute to our understanding of the determinants of entrepreneurship, as well as highlighting the application of the machine learning methodology.

Introduction

Entrepreneurship has been found to play an important role in economic growth, innovation, and social development. Given its importance it is no surprise that a substantial body of research has sought to understand the determinants of entrepreneurial activity and entrepreneurial intention[i]. Although results from the research suggest that the determinants include demographic, psychological, contextual and institutional factors there have been conflicting results on their importance, particularly in terms of the socio-cultural and demographic factors[ii], for example whether younger or older people are more likely to be entrepreneurs. Partly, these conflicting results are thought to be due to methodology, including uncertainty in the models used and the choices made by researchers around model specification[iii]. The existing methodological approaches have also been found to have limited predictive accuracy.

Given these methodological concerns and the fact that there is still debate therefore over which set of factors has greatest explanatory power in terms of determining entrepreneurship, an alternative approach using machine learning methods has been adopted here. The machine learning algorithms allow for the inclusion and evaluation of a large number of variables in the model building process, through the use of an automatic feature selection process. This reduces the model uncertainty associated with the researchers' specification of the model. This approach also generates variable importance measures which allows us to identify the most dominant variables in predicting entrepreneurial intention and early-stage entrepreneurial activity.

Machine Learning Approach

Use of a machine learning approach in entrepreneurship research has been facilitated by advances in computer science, which have created opportunities to draw on new data and techniques to answer new and existing research questions. Despite the opportunities presented by such techniques they have only been implemented to a limited extent to date in the entrepreneurship research sphere, primarily due to the lack of familiarity with the methods[iv].

Machine learning involves the application of an algorithm to learn relationships between variables in a dataset. There are two broad categories of machine learning: supervised learning, and unsupervised learning. Supervised learning involves using an algorithm to learn the relationships between input features (variables) and a target (outcome). There are many learning algorithms that can be used, such as regression, support vector machines, decision trees, random forests, gradient boosting, and artificial neural networks. In contrast to supervised learning, unsupervised learning involves the application of an algorithm to learn the patterns in the data. Common algorithms include k-means and hierarchical clustering. The key difference between supervised and unsupervised learning is that there is no target variable in the latter, rather the aim is to identify patterns of relationships in the data.

The machine learning approach we adopt is supervised learning, implemented using three common algorithms: recursive partitioning, logistic regression, and gradient boosting[v]. Use of the different algorithms allows us to identify differences between the methods and their potential utility in addressing entrepreneurship research questions. Our study also allows for identification of the most important predictors of entrepreneurship, by calculating the relative importance of the determinants, as well as identifying non-linear patterns, and the combinations of factors that lead to entrepreneurship.

To highlight the potential of machine learning techniques in entrepreneurship research our central question is thus: Which factors are most important in predicting entrepreneurial intentions and early-stage entrepreneurship?

Data Source

We draw on worldwide data from the 2017 Global Entrepreneurship Monitor (GEM) to undertake this study. Although the GEM framework has been studied extensively in entrepreneurship research, few studies have applied machine learning techniques. The GEM survey is based on an underpinning conceptual framework, which focuses on the factors that influence entrepreneurial behaviour, and in particular the influence of individual characteristics, social values, and national framework conditions[vi]. Individual level factors include demographics, as well as perceptual and motivating factors such as whether the individual knows an entrepreneur, whether they have the skills to start a business, opportunities to start a business, and fear of failure. Social values focus on whether entrepreneurship is a good career choice, media portrayal of entrepreneurs, status of entrepreneurs in society, and the ease of start-up. We draw on this framework, alongside the literature on the topic, as a guide to select variables for inclusion in the models.

Two dependent variables are considered: entrepreneurial intentions and early-stage entrepreneurial activity (TEA). Entrepreneurial intentions are measured using the survey question: 'Are you, alone or with others, expecting to start a new business, including any type of self-employment, within the next three years?'. Early-stage entrepreneurship, or TEA, is a widely used measure encompassing nascent and new business ownership. The nascent stage reflects businesses within the first three months of start-up, and new businesses as those between 3 and 42 months old. Both dependent variables are measured on binary scales where 0 = no and 1 = yes.

For the independent variables, we focus on individual level determinants from the GEM framework. This includes perceptual variables relating to self-perceptions, cultural factors, and individual demographics and experience. The perceptual variables include self-efficacy, networks, opportunities, and fear of failure. The cultural factors include perception of good opportunities for start-up; perception of entrepreneurship as a good career choice; perception of status of entrepreneurs; status of entrepreneurs in the media; ease of starting a business and social entrepreneurship.

The individual demographics include age, gender, ethnicity, occupation, household size, whether the person has acted as a business angel, and whether they have closed a business in the past two years.

Results

One of the aims of the machine learning approach is to identify the most important variables in predicting entrepreneurial intentions and early-stage entrepreneurial activity. We draw on the variable importance measures produced through the machine learning approach to identify the relative importance of the independent variables.

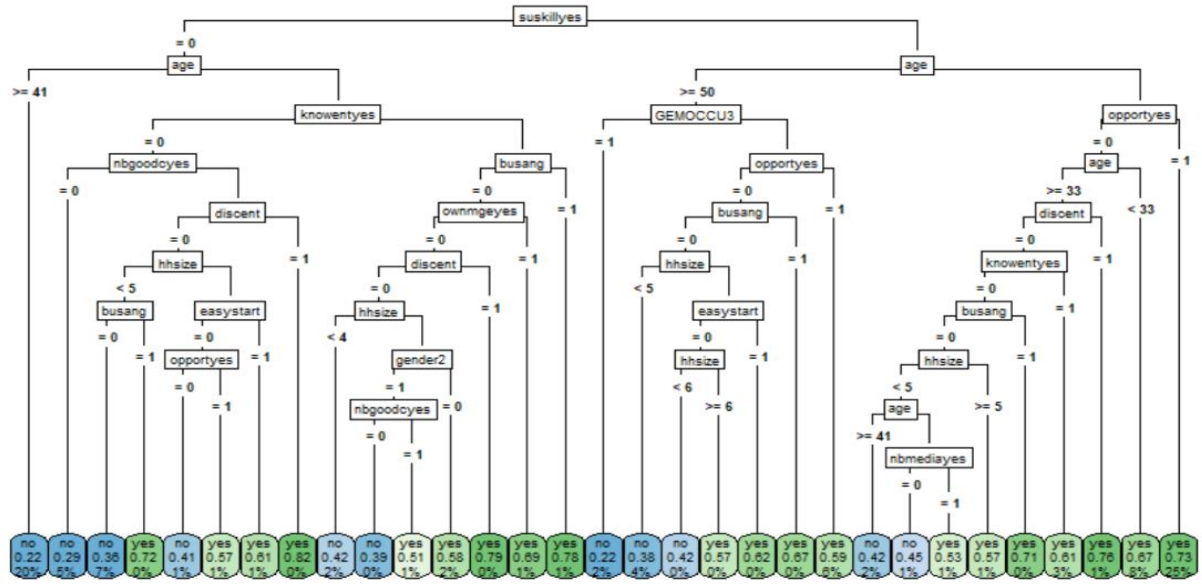
Turning first to the determinants of entrepreneurial intention, the results of the logistic regression model show that self-efficacy (having the skills to start a business) is the most important predictor, followed by age, having closed a business, and perceiving good opportunities for entrepreneurship.

In the recursive partitioning model, age and self-efficacy are also the two most important predictors, followed by a work status of retired/disabled, and perceiving good opportunities for entrepreneurship.

In the gradient boosting model, the most important predictors are an occupation of retired/disabled, having closed a business, experience as a business angel, and self-efficacy.

The structure of the final decision tree that was built using recursive partitioning is show in Figure 1.

Figure 1: Decision Tree Model Predicting Entrepreneurial Intention



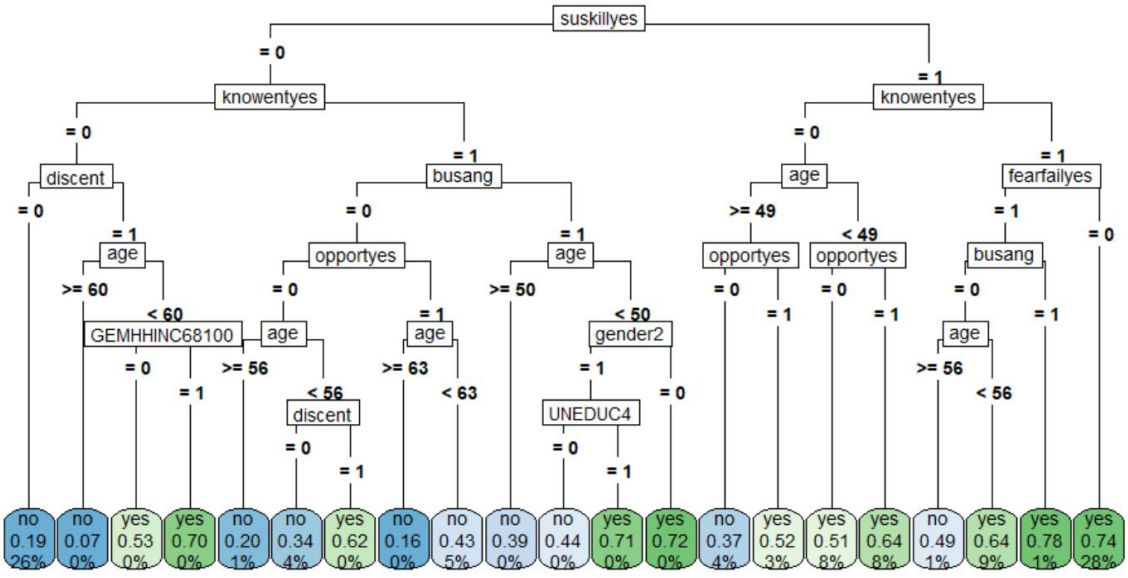
This tree structure can be interpreted alongside the above variable importance measures to add additional information about the structure of the final model. The tree shows that the variable that best separates the data into future entrepreneurs and non-entrepreneurs is self-efficacy, with individuals less likely to intend to start a business when they feel they do not have the skills to do so (i.e. suskillyes = 0). This is followed by age on both the left-hand side and right-hand side of the tree. On the left-hand side of the tree, people who do not have the skills to start a business and who are aged 41 or over are predicted not to have entrepreneurial intentions. Traversing the right-hand side of the tree shows that people with the skills to start a business, who are aged under 50 and who perceive good opportunities are predicted to have entrepreneurial intentions.

Analysing the determinants of early-stage entrepreneurial activity (TEA) using the same methodology results in the three perceptual variables of self-efficacy, knowing an entrepreneur, and perceiving good opportunities as the most important predictors in the logistic regression model. This is followed by age, and

fear of failure. The same pattern emerges in the recursive partitioning model. The gradient boosting model also exhibits a similar pattern, with self-efficacy and knowing an entrepreneur the two most important predictors. This is followed by age, perceiving good opportunities and having discontinued a business.

Figure 2 shows a pruned version of the final decision tree for predicting TEA. As with entrepreneurial intention, the tree splits first on self-efficacy, with individuals who do not have the skills to start a business being less likely to do so. The second split on both sides of the tree is on whether or not the person knows an entrepreneur, and in both cases a higher proportion of people that know an entrepreneur are engaged in early-stage entrepreneurship compared with those who do not.

Figure 2: Pruned Decision Tree Model Predicting Total Early-Stage Entrepreneurial Activity



Discussion

Across all models, the perceptual variables are consistently amongst the most predictive factors of both entrepreneurial intention and TEA. Notably, the cultural variables are not found to be important predictors of either. This is consistent with the wider literature that has found these variables to have low predictive power[vii].

Having the skills to start a business is the most important predictor of TEA across all three models and is of relatively high importance across the three models for entrepreneurial intention. This is consistent with theoretical arguments about the importance of self-efficacy[viii]. Opportunities to start a business are also important predictors of both entrepreneurial intention and of TEA, which aligns with arguments from the literature about the fundamental role of opportunities in entrepreneurship[ix].

Knowing an entrepreneur is consistently of relatively higher importance in predicting TEA compared to predicting entrepreneurial intention. This could suggest that networks of entrepreneurs are important when actually in the process of starting a business but are of less importance when one is intending to start a business. Fear of failure is also found to be of relatively lower importance in predicting entrepreneurial intention, but consistent with the literature, is amongst the top predictors of TEA.

In terms of the demographics, age is consistently of relatively high importance across the entrepreneurial intention and TEA models, although the decision tree structures in Figures 1 and 2, highlight the complex relationship between age and entrepreneurship. Gender is also found to be less important in predicting both entrepreneurial intention and TEA. One potential reason for this could be due to interrelationships between gender and other concepts such as self-efficacy[x], which account for the majority of the predictive ability.

In terms of overall model accuracy, the gradient boosting models were found to be most accurate while overall the models were stronger at predicting TEA compared to entrepreneurial intention. We suggest that this is because TEA is a more tangible and specific measure than entrepreneurial intention. The analysis does indicate, however, that it is difficult to predict both entrepreneurial intention and TEA as all models have quite a high level of predictive error. Again, this is consistent with the wider literature that entrepreneurship is a difficult phenomenon to predict, with multiple determinants[xi].

Conclusion

In this study we have implemented a machine learning methodology to examine the relative importance of predictors of entrepreneurial intention and early-stage entrepreneurship. Perceptual variables are found to be the most important predictors of entrepreneurship, particularly having the skills to start a business. Age is an important demographic predictor, but other demographic factors are relatively less important. Cultural perceptions are also found to be relatively less important. The decision tree further highlights the complex interrelationships between factors, indicating that there is no unique set of attributes that predict being an entrepreneur or a future entrepreneur, rather a combination of different attributes can result in entrepreneurship. The findings contribute to our understanding of the entrepreneurial phenomenon. From a policy perspective the models developed can help us to better understand those most likely to start a business, as well as the most important predictors. This information could enable more effective targeting of resources and improved focus of policy interventions to increase entrepreneurship.

Karen Bonner, Byron Graham

For more information please contact ka.bonner@ulster.ac.uk or byron.graham@qub.ac.uk

[i]For example, for a review, see Audretsch D.B., Erdem D.K. (2005) Factors Affecting Entrepreneurial Activity: Literature Review. In: Alvarez S.A., Agarwal R., Sorenson O. (eds) Handbook of Entrepreneurship Research. International Handbook Series on Entrepreneurship, vol 2. Springer, Boston, MA. Factors Affecting Entrepreneurial Activity: Literature Review

[ii]Liñán, F., & Fayolle, A. (2015). A systematic literature review on entrepreneurial intentions: citation, thematic analyses, and research agenda. International Entrepreneurship and Management Journal, 11(4), 907–933. A systematic literature review on entrepreneurial intentions: citation, thematic analysis, and research agenda

[iii]Arin, K. P., Huang, V. Z., Minniti, M., Nandialath, A. M., & Reich, O. F. M. (2015). Revisiting the Determinants of Entrepreneurship: A Bayesian Approach. Journal of Management, 41(2), 607–631. Revisiting the Determinants of Entrepreneurship: A Bayesian Approach

[iv]Obschonka, M., & Audretsch, D. B. (2020). Artificial intelligence and big data in entrepreneurship: a new era has begun. Small Business Economics, 55, 529–539. Artificial intelligence and big data in entrepreneurship: a new era has begun

[v]For space reasons, in the results section we only display the decision tree output of the recursive partitioning method.

[vi]GEM. (2018). Global Entrepreneurship Monitor 2018/2019 Global Report. Global Entrepreneurship Monitor. Retrieved from GEM 2017/2018 GLOBAL REPORT

[vii]Liñán, F., Urbano, D., & Guerrero, M. (2011). Regional variations in entrepreneurial cognitions: Start-up intentions of university students in Spain. Entrepreneurship and Regional Development, 23(3–4), 187–215. Regional variations in entrepreneurial cognitions: Start-up intentions of university students in Spain

[viii]Bandura, A. (1997). Self-efficacy: The exercise of control. New York, NY: Worth Publishers

[ix]Shapero, A. (1984). The entrepreneurial event. In C. A. Kent (Ed.), The environment for entrepreneurship (pp. 21–40). Lexington, KY: Lexington Books.

[x]Chowdhury, S & Endres, M. (2005). Gender Difference and the Formation of Entrepreneurial Self-Efficacy (2005). University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship

[xi]Kuckertz, A., Berger, E.S.C. & Allmendinger, M. (2015). What Drives Entrepreneurship? A Configurational Analysis of the Determinants of Entrepreneurship in Innovation-Driven Economies. Die Betriebswirtschaft/Business Administration Review, 75,4, pp. 273-288, Available at SSRN: What Drives Entrepreneurship? A Configurational Analysis of the Determinants of Entrepreneurship in Innovation-Driven Economies