Paper 4052-2019

Exploring the Intentions of Entering Entrepreneurship for SAS® Global Forum 2019

Surabhi Arya, Prashant Gour, Dhruv Sharma, Jeroen Vanheeringen, Oklahoma State University

ABSTRACT

Entrepreneurship is fast emerging as a transformational megatrend of the 21st century given its capacity to reshape economies and industries throughout the world. This study shows that initially societal perception about an entrepreneur was influential in starting a business in the U.S. while in recent years other factors replaced this such as experience in owning a business and individual perception regarding entrepreneurship.

This study explores the determinants explaining why people likely start a business in the U.S. across the time period 2005 – 2014.

INTRODUCTION

The Global Entrepreneurship Monitor (GEM) database is a tool to provide worldwide comparable data regarding entrepreneurship. It measures differences in entrepreneurial attitudes, activity and aspirations of individuals, uncovers factors determining the nature and levels of entrepreneurial activity, and allows identifying policy implications in terms of entrepreneurship. Since its inception, GEM has conducted research in approximately one hundred countries.

The GEM database has been used by a variety of published studies throughout the years. Many of them focus on a combination of entrepreneurship worldwide related factors and issues. However, exclusively exploring the intentions of entering entrepreneurship in the U.S. and comparing the results with sets of clustered countries has often gone unremarked.

DATA

GEM gathers survey data to measure social, economic and political factors affecting entrepreneur conditions regarding 100+ economies. Survey responses are captured in different categories analogous to Likert scale having -2, -1, 0, 1 responses.

This study used the individual level GEM dataset[1] for 2005-2014 to identify significant variables for entrepreneurship in the US. In addition, it used the aggregated national level GEM dataset for 2014 to compare factors for entrepreneurship in the United States with other parts of the world. The individual level dataset has approximately one million observations representing the respondents' survey responses for each year. On the other hand, national level aggregated dataset is a harmonized individual-level data capturing results from all surveys across each country under study. The dataset has 77 observations representing each country and 277 variables representing metrics for entrepreneurship assessment (survey questions which were asked to respondents).

DATA CLEANING / PREPARATION

There were empty values in the dataset and these values are different from missing values because they can't be imputed unlike missing values. These empty values were present because the survey design was conditional and questions were asked based on the yes or no reposes in the target variable '*BSTART*'. Therefore, the transformation and data selection technique(s) were to be selected in a way that it handles missing values. The dataset had

[1] GEM datasets are available at: https://www.gemconsortium.org/data/sets?id=aps

several unexpected values which were removed using basic data step functionality in SAS. The major issue with the dataset is the distribution of variables with huge kurtosis and skewness values. If untreated, such variables would create many problems during further analysis of the dataset. As a result, it was imperative to select an appropriate variable selection method or variable reduction technique to provide unbiased analysis results.

A very important aspect of analyzing survey data is the technique used to handle the skewness in the data distribution. Generally, survey data are skewed because of the biased responses from the respondents. This bias is due to respondents' tendency to remain inclined towards extreme ends of the Likert scale which in this case are -2 and 1, or sometimes the response is neutral towards the majority of questions. This inclination of respondents impacts the statistical properties like skewness, kurtosis in the data. Therefore, to handle such skewed distribution in the data, the study used a popular data transformation technique called *double standardization technique* for the survey data.

ANALYSIS

An important aspect of the study was to adopt an appropriate variable selection technique for capturing the change in entrepreneurship dynamics in the United States from 2005 to 2014. This study utilizes several supervised variable selection techniques available through SAS Enterprise Miner. Based on a model comparison node, SAS selects the most efficient variable selection technique using the minimum misclassification criteria of applied validation data.

The study used several predictive modeling techniques (doubling as variable selection techniques) such as Decision Tree, Gradient Boosting, Random Forest, LARS, and LASSO using target variable *Bstart*, having 1 or 0 response outcome where response value 1 represented a person who is trying to start a new business with or without collaboration. After carefully assigning the input variable roles in metadata definition, the data was split into training and validation having 70-30 ratio for model assessment. Finally, based on model comparison node Gradient Boosting was the best performing model having 6% misclassification rate in the validation data.

The rationale behind using sophisticated decision tree algorithms was that it is extremely capable of handling empty values and skewed data distribution. Variable importance in the Gradient Boosting, HP Forest and HP Tree models were almost identical whereas the results for LARS and LASSO models were varied. LARS and LASSO are highly sohisticated regression techniques but the skewness and empty values in our was not handled by this technique due to very less complete cases

The models helped to find the determinants for entrepreneurship in the United States. There was an interest to understand whether they were specific to the U.S. or also applicable to other countries surveyed by GEM. For comparison, a cluster analysis was done on different GEM dataset aggregating surveyed data by country (2014).

Using cluster analysis functionality in SAS EM, the national level aggregated data was divided into five different clusters and from each cluster, a country was randomly selected to identify influencing factors in starting a business in that country using The Gradient Boosting model.

RESULTS

Variable Selection Results

The results of significant variables for starting a business in the United States are tabulated in Table 1 (appendix) for each year from 2005 to 2014. In order to avoid ambiguity in understanding significant variables definitions, actual variable names are replaced by the brief definition in Table 1 (appendix).

Based on the results, the following observations were made which captures changes in U.S. entrepreneurship dynamics in different domains like anthropology, economics, psychology, and sociology.

1. Between 2012 and 2014, economic factors such as experience in owning a business (anybusow), sociological factor like individual perception to entrepreneurship (*INDSUPyy) and anthropology factors TEAyyFEM, TEAyyMAL are highly influential factors for the entrepreneurship in the US.
2. Variable gender became an influential factor after 2011 for the U.S. as *TEAyyFEM* and *TEAyyMAL* became significant and continued to be significant until 2014. In 2012, 4.7% of the total surveyed population were female involved in *TEA. This increased to 4.9% in 2014. At the same time, 6.4% of the total surveyed population were male involved in *TEA [2] and this number increased to 7.4% in 2014.
3. Psychological factor *Fearfail* was a significant influential factor during 2008 - 2009, but no longer significant afterward. Variable '*Fearfail'* refers to survey question 'Would fear of failure would prevent you from starting a business?'
4. Variable *Knowent* is significant throughout 2005 to 2014. Variable *Knowent* signifies if the survey respondent personally knows someone who started a firm in the past two years.
5. Variables like **nbgoodc* and **nbstatus* related to societal perception about an entrepreneur were significant in the initial years of study but are no longer important.

Figure 1 represents the actual names of the survey variables which were significant for entrepreneurship in the US from 2005 to 2014. Figure 2 represents the cross-reference of relevant variables with their associated domain. For instance, *suskill* is a significant variable in several years like 2005, 2007, 2008, and 2009 (Figure 1) and it is categorized under economic factor (Figure 2).

[2] Note: *TEA - TEA (Total Early-Stage Entrepreneurial Activity), which assess the percent of working age population both about to start an entrepreneurial activity, and that have started one from a maximum of 3 years and half.

INDSUPyy : combination of skillset needed to start a business, perceiving entrepreneurship as an opportunity versus necessity.

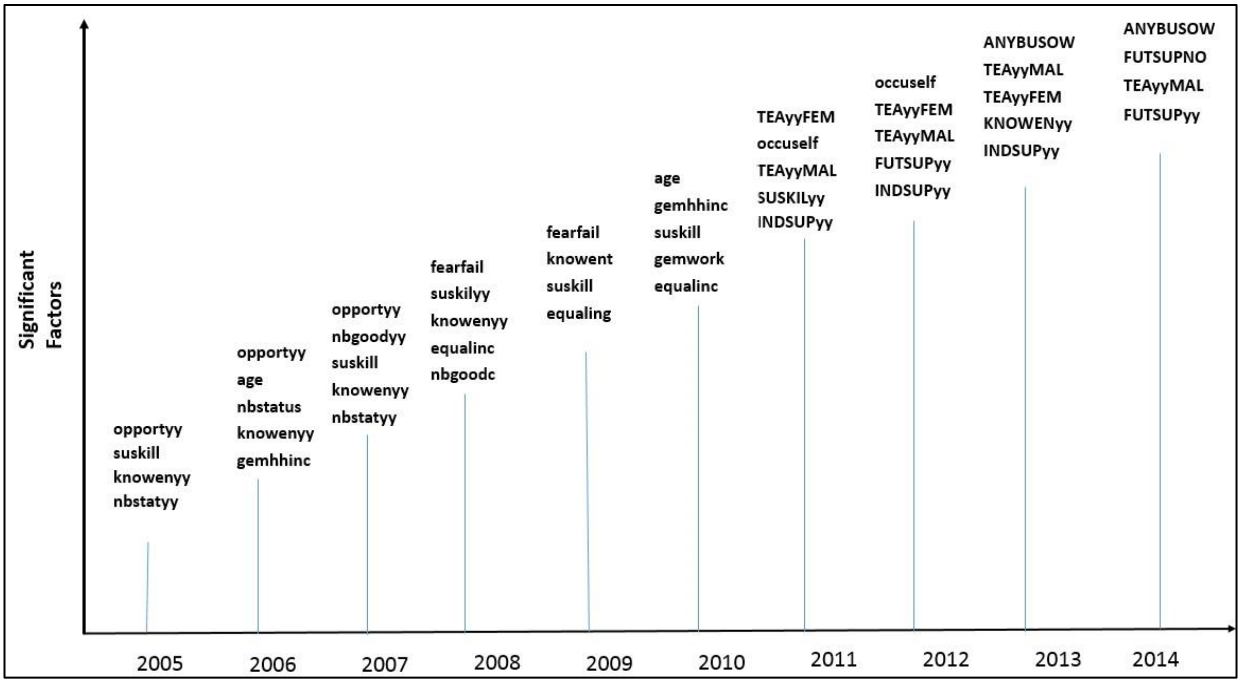

Figure 1: Significant influential factors for entrepreneurship in the US from 2005 to 2014

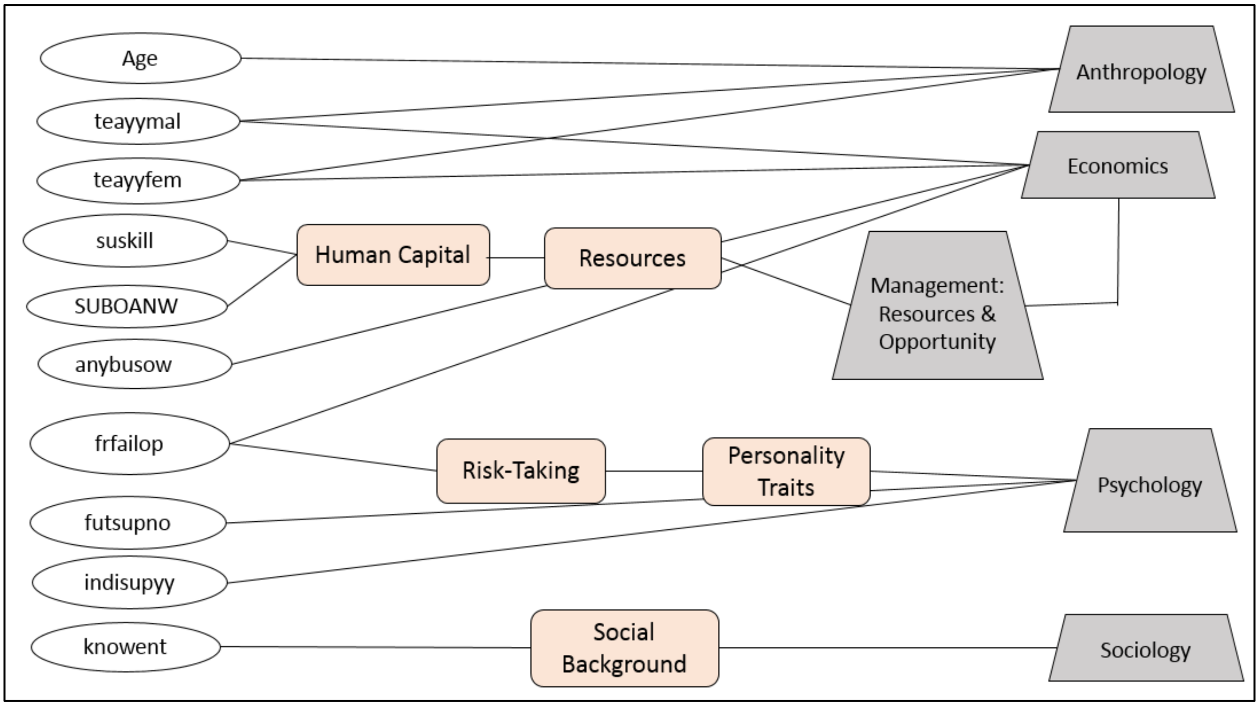

Figure 2: Significant variables and their associated categories

Clustering Results – National Level Aggregated Data

After assessing significant variables for the US, influential factors across the world were compared. Using a clustering approach, a country from each cluster was randomly selected. Table 2 (appendix) shows all significant variables that were selected by The Gradient Boosting model.

All countries were compared in all individual clusters and observed that there are no significant differences between the influential entrepreneurship factors in different parts of the world. Variables related to prior experience in business (*ANYBUSOW*), presence of necessary skillset (*SUSKILyy*), individual perception to entrepreneurship are the overlapping influential factors in majority parts of the world.

SUGGESTIONS FOR FUTURE STUDIES

The study used several variable selection techniques to find and analyze determinants for entrepreneurship in the United States by corroborating results of various applied predictive models. Further refinement of the various applied predictive models, alteration of data partition percentages, and simply more data might serve to improve the selected model's accuracy.

REFERENCES

"Eliminating Response Style Segments in Survey Data via Double Standardization Before Clustering", available at http://support.sas.com/resources/papers/proceedings11/165-2011.pdf

"SAS/STAT User's Guide The CLUSTER Procedure", available at https://support.sas.com/documentation/cdl/en/statugcluster/61777/PDF/default/statugcluster.pdf

CONTACT INFORMATION

Your comments and questions are encouraged. Contact the authors at:

Dhruv Sharma
Oklahoma State University
(405)-385-2370
dhruv.sharma@okstate.edu

Jeroen Vanheeringen
Oklahoma State University
(949)-394-7218
jvanhee@okstate.edu

Prashant Gour
Oklahoma State University
(405)-762-3697
prashant.gour@okstate.edu

Surabhi Arya
Oklahoma State University
(316)-519-5741
surabhi.arya@okstate.edu

APPENDIX

Significant Variable	Survey Question
ANYBUSOW	Any Business Owner: Nascent New Established
TEAyyMAL	Involved in TEA, male
TEAyyFEM	Involved in TEA, female
KNOWENyy	KNOWENT adapted to make it fit for national level aggregation
INDSUPyy	Individual perception to entrepreneurship Index
NEMALEyy	Nascent entrepreneur, male
occuself	Self-employed
FUTSUPyy	Expects to start-up in the next 3 years
SUSKILyy	SUSKIL adapted to make it fit for national level aggregation
NEFEMAyy	Nascent entrepreneur, female
age	What is your current age (in years)?
suskill	Do you have the knowledge, skill, and experience required to start a new business?
gemwork	GEMWORK. GEM harmonized work status
equalinc	In my country, most people would prefer that everyone had a similar standard of living.
fearfail	Would fear of failure would prevent you from starting a business?
knowent	Do you know someone personally who started a business in the past 2 years?
nbgoodc	In my country, most people consider starting a new business a desirable career choice.
opportyy	OPPORT adapted to make it fit for national level aggregation
nbgoodyy	NBGOOD adapted to make it fit for national level aggregation
nbstatyy	NBSTAT adapted to make it fit for national level aggregation
nbstatus	In my country, those successful at starting a new business have a high level of status and respect.
SUBOANW	Actively involved in start-up effort, owner, no wages yet

Table 1. Selected variables and their descriptions

China	Germany	Peru	Trinidad & Tobago	Russia
ANYBUSOW	SUSKILyy	TEAHITEC	ANYBUSOW	INDSUPyy
occuself	FUTSUPyy	FUTSUPNO	FUTSUPyy	TEAyyMAL
FUTSUPyy	TEAyyFEM	age	FUTSUPNO	SUSKILyy
KNOWENyy	INDSUPyy	GEMWORK3	NEMALEyy	age
SUSKILyy	suskill	omyint	TEAyyMAL	suskill

Table 2. Significant variables for different countries