Breaking the Zuckerberg Myth: Successful Entrepreneurs Have 10 Years of Prior Employment

Utilizing Data Science and Machine Learning to Study Socio-Economic Patterns Among Successful Entrepreneurs

Thomas Ferry
Sutardja Center for Entrepreneurship and Technology
University of California, Berkeley
Berkeley, CA, USA
tferry@berkeley.edu

Mudit Goyal
Sutardja Center for Entrepreneurship and Technology
University of California, Berkeley
Berkeley, CA, USA
muditg@berkeley.edu

Ikhlaq Sidhu
Sutardja Center for Entrepreneurship and Technology
University of California, Berkeley
Berkeley, CA, USA

Alexander Fred-Ojala
Sutardja Center for Entrepreneurship and Technology
University of California, Berkeley
Berkeley, CA, USA

***Abstract*—With the advancement of Big Data technologies, and the creation of new, open source Machine Learning and Data Science tools, we now have new methods and approaches to obtain quantitative entrepreneurship research results. This approach benefits from a seemingly inexhaustible source of information, i.e. data. In this paper, we apply data science techniques from UC Berkeley's Data-X framework to research socio-economic characteristics and traits, related to successful entrepreneurs, and their link to early stage venture success. Such techniques include the use of web scraping scripts, data cleaning, the creation of data pipelines, and visualization using open source Python libraries (e.g. Scikit-learn, Pandas) as well as other tools commonly used in the Machine Learning development stack. This paper offers two types of results of the performed quantitative study, that is in the intersection between Machine Learning and Entrepreneurship Research (Fig 1.). First that the Machine Learning development stack offers powerful tools to produce entrepreneurship research results, and this might be a new paradigm in this field of research. Second, we can now quantitatively verify from public data sources that, contrary to popular belief, a successful entrepreneur is not a young, college drop-out that had a genius idea in their dorm room. Instead, the most predictive trait of entrepreneurial success is the work experience of the founder, i.e., the number of years the founder has been employed before starting their company. Optimally they will have been employed for 10 to 12 years.**

Keywords: data science; venture valuation; entrepreneurial success; machine learning; evaluation methods

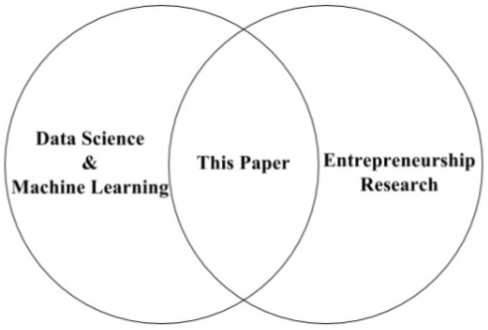

Fig. 1.

978-1-5386-1469-3/18/$31.00 ©2018 IEEE

I. Introduction

Throughout the 1990's, entrepreneurship research was focused on figuring out the motivations for, the patterns of, and the pace of industrialization for new ventures. Recently however, the focus has more been on looking into entrepreneurial activities with the aim to uncover the key patterns of activities associated with success. [3] The fundamental questions that researchers strive to answer about entrepreneurship are "how, why, and when do entrepreneurs discover and explore opportunities?" Some factors that may help to answer some of those questions might include composition, ability or exposure. Other factors such as financial strength may also play a large role, as more stability is correlated to increased propensity for risk taking. [1]

Entrepreneurship research has been highly qualitative. Can we quantify some of those factors? Even though economic factors are the most easily quantifiable, personality and social factors can also be put into numbers, with the onset of Natural Language Processing and Machine Learning. A jump in computing power coupled to a democratization of Data Sciences are making highly quantitative methods for entrepreneurship research more relevant by the day. Today Random Forests and Stability Selection, among others, can be applied to discover valuable insights in the endless stream of data the internet provides. [2]

In this paper we will explore the socio-economic traits of entrepreneurs and their link to early stage success. More specifically, we will be looking at the work and educational experience start-up founder-CEOs have at the time they start their company and how predictive those characteristics are of a Valuation Increase between the Series A round of funding and the Series B.

We state the following hypothesis: *A founder-CEO's work experience before starting the venture, as measured in units of time, is the most predictive feature of early stage valuation increase of that new venture (in comparison to e.g., where the employment took place, academic background of the founder, what type of jobs the founder has had)*

We aim to verify this hypothesis if the following is true: *Different feature selection techniques will unveil a larger predictive importance of Years of Employment than of Worked at Google, Worked at Microsoft, Worked as PM and Worked in Sales features etc., with regards to Valuation Increase* We will first start by describing the analysis methods used to obtain the results and then describe the characteristics of the results and explore their implications.

II. Methodology

A. Objective

The general purpose of this study was to explore the characteristics of entrepreneurs to find the most predictive features of early stage venture success. The definitions of what constitutes an entrepreneur and entrepreneurial success are still subject to debate in the entrepreneurship research community, but for this study the following criteria and assumptions are established:

- An entrepreneur is/was founder and CEO of a private company that reached at least a Series B round of funding.
- A Series B round of funding is a second or third round of financial funding backed by one or more Venture Capital firms.
- We measure success in terms of relative valuation increase between series A and series B rounds of financing of the founders' startups

The methods used to analyze the data are quantitative and differ from more 'traditional' methods used in social science research and its interrelated research fields. In this study a variety of feature selection techniques such as Pearson correlations, distance correlations, random forest regression, and stability selection implemented on a randomized lasso regularization; statistical learning techniques typically present at the core of data science, machine learning and artificial intelligence. The Pearson correlation [4] allowed us to determine the existence of linear relationships in our data, the distance correlation [5] allowed us to determine the existence of non-linear relationships as well as the degree of independence of variables with regards to the increase in valuation between series A and series B, the random forest regressor [6] allowed us to calculate the predictive importance of each founder characteristic with regards to the valuation increase between Series A and Series B of an early stage company (emphasizing the importance of the top three features). Finally, stability selection [7] allowed us to have a less top-skewed view of the importance of each feature.

B. Tools and Methods of Data Collection

There were three distinct parts to our process of data collection. First, the gathering of a list of startup founders-CEOs and their respective companies. Second, the gathering of Series A and Series B valuations of the said companies, and finally the gathering of the characteristics of each founder. For the list of the characteristics we chose to explore, see APPENDIX.

The first phase of our data collection was performed with the export of a filtered search from the www.crunchbase.com database. Cross-referencing a list of founders whose company had at least reached a Series B round and their corresponding startups, we turned towards www.pitchbook.com and their valuation estimates to gather the Series A and Series B valuation information for each founder. Finally, we primarily used public profiles on www.linkedin.com of the founders alongside a few web searches to gather their characteristics. The collection was performed manually by a team of five

University of California, Berkeley students and over the time period September 2017 – November 2017.

With the novel aspect of our methodology in mind, this study constitutes only the first step towards what we believe will be a larger movement of highly quantitative analyses and applications of machine learning methods in entrepreneurship research. We are also aware of the limitations of the data collection process and we were faced with unavoidable imperfections accompanying the explorations and analysis. This mainly due to constraints ranging from legal concerns regarding data ownership to limited time and resources. All in all, we collected data on 244 founders along with 16 features which were all collected from the available data repositories, i.e., it might contain irrelevant features and the sampling method might also be biased. However, regardless of the outlined limitations of the study it still brings valuable results to the academic field of entrepreneurship research – as our findings are deemed to be statistically relevant, hence our results display the characteristics of valid statistical inference.

Before we move onto the specifics of our analysis method, we formally define the *Valuation Increase* feature. As shown in equation (1), we defined the *Valuation Increase* to be the ration of the post-money valuation of the start-up at Series B to the post-money valuation at Series A.

$$\text{Valuation Increase} = \text{Series B Post-Money Valuation} / \text{Series A Post-Money Valuation} \quad (1)$$

C. Processing and Analysis

Along with a more classic observation of correlations we built a Random Forest Regressor (RFR) and performed a stability selection to observe which characteristics of the entrepreneur were most predictive of what we defined to be early entrepreneurial success: a change in valuation between Series A and Series B stages of funding. It is interesting to note that none of our continuous variables had missing values.

1) Distance Correlation

Distance correlation is a robust method of correlation estimation, that was designed to address the shortcomings of Pearson Correlation [5]. For the Pearson Correlation method, a correlation value of 0 does not imply independence (as can be observed in the classic example of x and x^2). On the other hand, a Distance correlation value of 0 does indicate independence between the variables. [8] In the case of complex systems such as ours, where the relationships between variables are most likely non-linear, Pearson Correlation is not the best choice to measure the relationship between two variables. As a result, we chose to use Distance correlation to study the relationships between the entrepreneur's characteristics and the valuation change of the founder's early stage venture.

2) Random Forest Regressiona

For the reader unfamiliar with Random Forest Regressors (RFRs), we will briefly describe the principles behind the method and explain why we believe this to be a great method to be applied to entrepreneurship research.

A random forest, as its name indicates, is an ensemble of decision trees. Decision Trees (used in our case for regression) are a type of supervised Machine Learning algorithm that iteratively split a population or sample into two or more homogenous sub-populations based on the most significant variable. To choose the best split, i.e., the most significant variable maximizing the information gain, regression trees use a Reduction in Variance algorithm. This algorithm choses the split which minimizes the most the overall 'impurity' of each subpopulation. For more information we invite the reader to look at the resources offered at the Data-X at Berkeley website https://data-x.blog.

Random Forests are therefore a powerful tool that can show us which features are most important in predicting a valuation change. They are well-suited for analyzing and representing complex non-linear systems and relationships, and they allow for a quick and elegant understanding of the relative importance of each feature without having to calculate their correlations with the target variable.

In our case, we implemented the Random Forest Regression algorithm with 1000 decision trees using the Scikit-learn Python library. [11] The following hyperparameters were used:

max_depth=None, *min_samples_split=2*, *min_samples_leaf=1*, *min_weight_fraction_leaf=0.0*, *max_features='auto'*, *max_leaf_nodes=None*, *min_impurity_decrease=0.0*, *min_impurity_split=None* *bootstrap=True*, *oob_score=False*, *n_jobs=1*, *random_state=None*, *verbose=0*, *warm_start=False*

We chose to use those parameters as they are of standard in early stage machine learning projects. Finally, we used two ways to measure 'impurity': the mean squared error and the mean absolute error and observed the results.

3) Randomized Stability Selection

Stability selection [7], is a recent feature selection method based on subsampling in combination with selection algorithms. On a high level, the idea is to apply a feature selection algorithm on different subsets of data. These subsets of data include different subsets of features. The process is repeated until the selection results can be aggregated. This aggregation can be performed for example by checking how often a feature ends up being selected as important. As a

result, we can expect predictive features to have scores close to 100%; while weaker, but still relevant features will also have non-zero scores. Finally, irrelevant features will have scores close to zero, as they never appear among the selected features. Stability selection is therefore useful for both pure feature selection, but also for data and feature interpretation. Indeed, good features will not get a low score just because they are similar and/or correlated to another feature in the data set.

In our study, we applied stability selection on a Randomized LASSO regression [9] algorithm implemented in the Python package Scikit-learn. We chose this implementation because it offers a convenient way to determine the regularization parameter automatically and is simple to use yet powerful.

4) Comparison of Results and Exploring Relationships

To cross-reference the 4 methods employed and to validate our results we employ an ensemble technique where we rank the features according to their importance scores. Every time a feature is ranked as most predictive, it scores 16 points, while each subsequent feature scores a point less than the previous one. We cannot in this case use a mean of the scores we obtain as they are all calculated differently and are on different scales.

Finally, to study the nature of the relationship between our most predictive feature and our measure of early stage success we fit Polynomial Regression models [10] of respective degrees: 1,2,3,4 and 5.

5) Standardizing Variables

The reader can notice that in the name of certain variables we appended the notation 'standardized'. This is meant to signify that the original value of the variable was modified to fit a certain scale or be grouped in a joint category. For example, "UC Berkeley" and "University of California at Berkeley" both became "University of California Berkeley".

III. RESULTS

A. Distance Correlation

Distance Correlation showed that the three characteristics / features most predictive of a valuation change are: *Years of Employment*, *Standardized Graduate Studies* and *Worked in Sales?* As shown in Fig. 2.

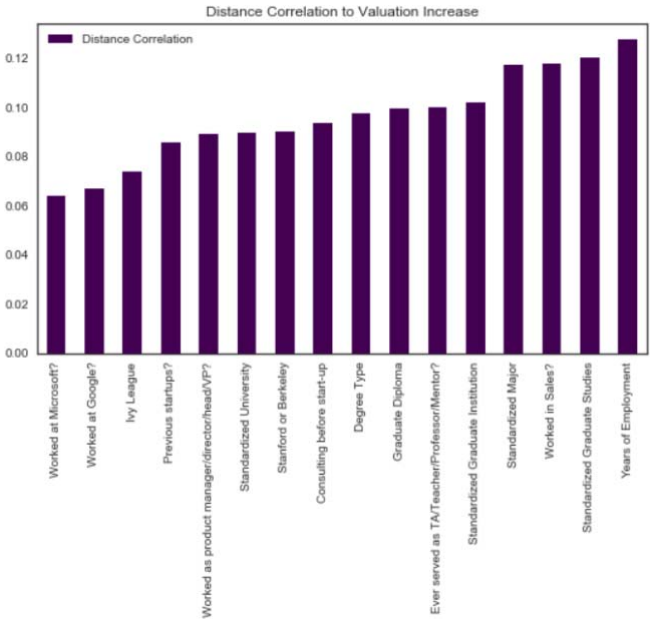

Fig. 2. Distance Correlation of Features to Valuation Increase vs Features

B. Random Forest Regression

1) Mean Squared Error

RFR with the Mean Square Error method employed as the impurity measure showed that the three characteristics / features most predictive of a valuation change are: *Standardized Major*, *Standardized University* and *Years of Employment?* As shown in Fig. 3.

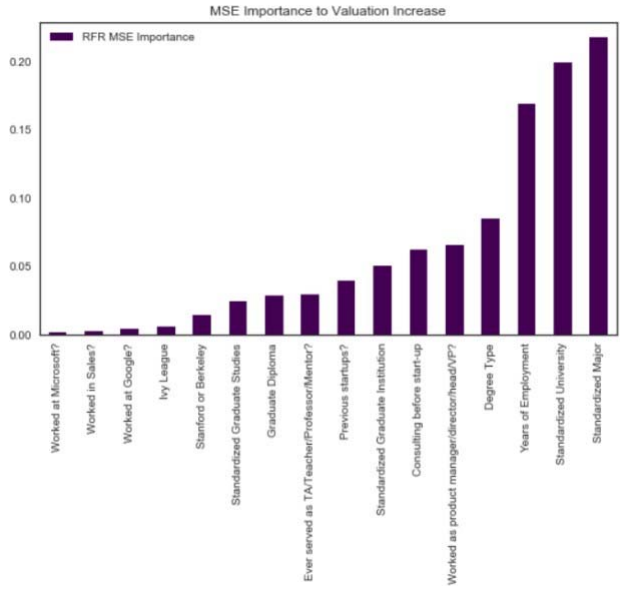

Fig. 3. Random Forest Regression Mean Squared Error Importance Score of Feature to Valuation Increase Prediction vs Features

2) Mean Absolute Error

RFR with an Mean Absolute Error method employed as the impurity measure showed that the three characteristics / features most predictive of a valuation change are: *Standardized Major*, *Standardized University* and *Years of Employment?* As shown in Fig. 4.

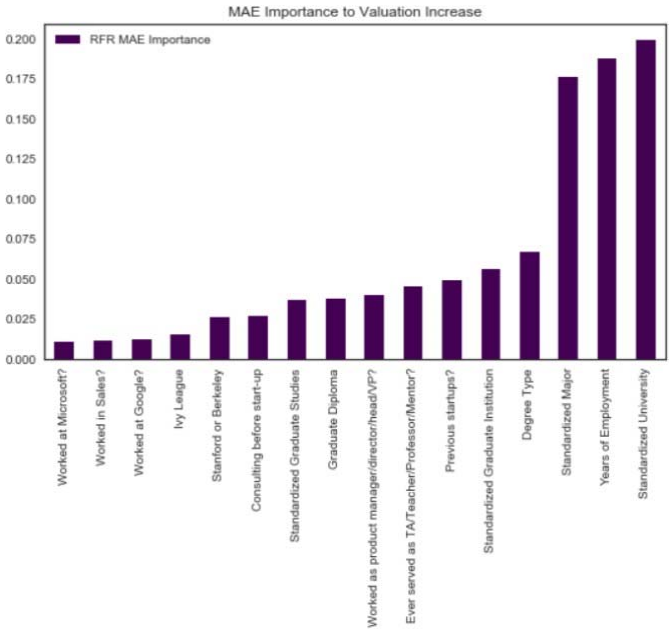

Fig. 4. Random Forest Regression Mean Absolute Error Importance Score of Feature to Valuation Increase Prediction vs Features

C. Stability Selection

Stability Selection showed that the three characteristics / features most predictive of a valuation change are: *Years of Employment*, *Worked in Sales?* and *Standardized Major* as shown in Fig. 5.

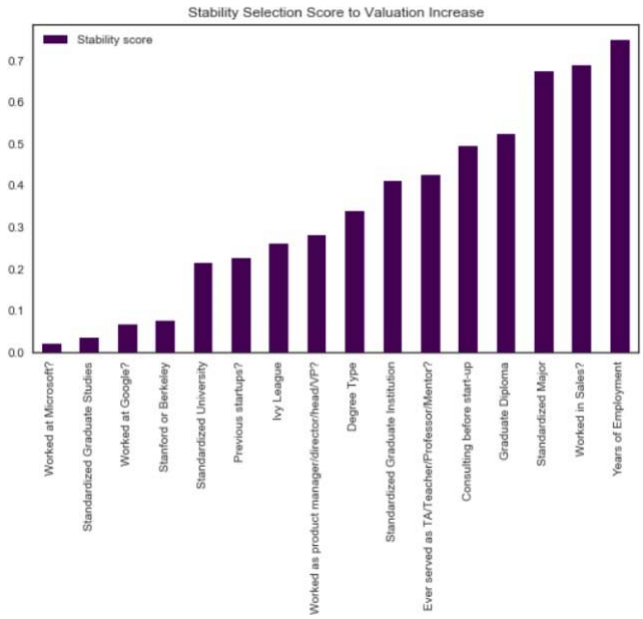

Fig. 5. Stability Selection Score for Features to Valuation Increase vs Features

D. Summary

Our ranking method showed that the three characteristics / features most predictive of a valuation change are: *Years of Employment*, *Standardized Major* and *Standardized Graduate Institution* as shown in Fig. 6 and Table I.

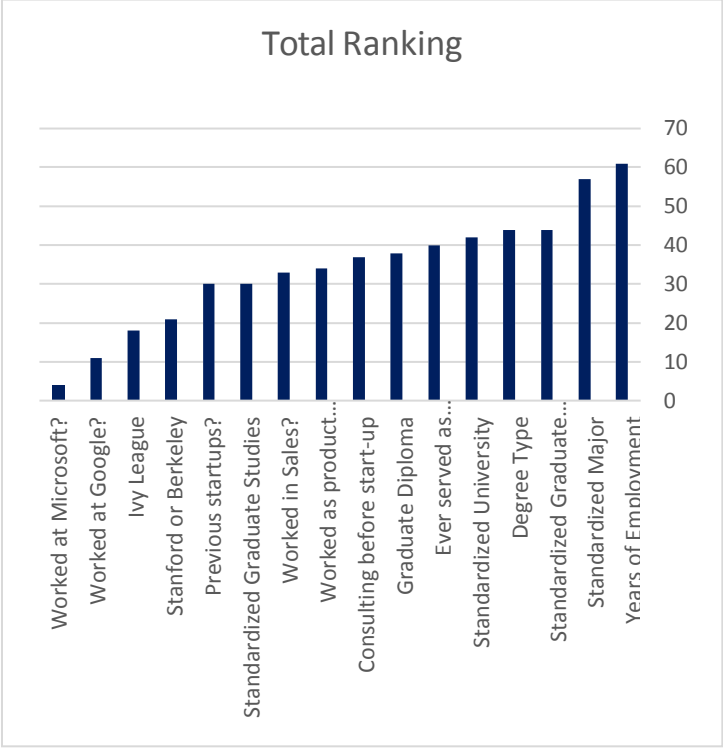

Fig. 6. Summary Sum of Rankings

TABLE I. SUMMARY RANKING TABLE

Summary Ranking Table	Scoring Method				
	Distance Correlation	*RFR MSE Importance*	*RFR MAE Importance*	*Stability score*	*Total*
Years of Employment	16	14	15	16	61
Standardized Major	13	16	14	14	57
Standardized Graduate Institution	12	10	12	10	44
Degree Type	9	13	13	9	44
Standardized University	6	15	16	5	42
Ever served as TA/Teacher/ Professor/Mentor ?	11	8	10	11	40
Graduate Diploma	10	7	8	13	38
Consulting before start-up	8	11	6	12	37
Worked as product manager/ director/head/VP?	5	12	9	8	34
Worked in Sales?	14	2	2	15	33
Standardized Graduate Studies	15	6	7	2	30
Previous startups?	4	9	11	6	30
Stanford or Berkeley	7	5	5	4	21
Ivy League	3	4	4	7	18
Worked at Google?	2	3	3	3	11
Worked at Microsoft?	1	1	1	1	4

E. Relationship Between Years of Employment and Valuation Increase

In Fig.7, from the polynomial regressions we fit to our data it seems that

1- Years of Employment is positively correlated to a Valuation Increase, and that
2- The companies lead by entrepreneurs with 10 to 12 years of prior employment display the highest valuation increases between Series A and Series B.

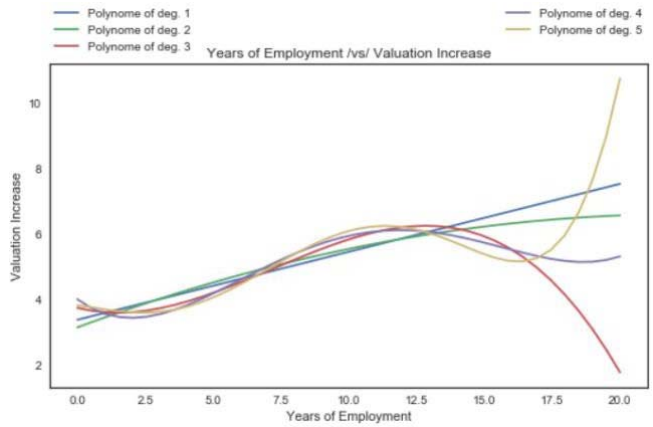

Fig. 7. Polynomial Regression models fit with target variable of Years of Employment and predictor Valuation Increase.

IV. IMPLICATIONS AND CONCLUSION

With the resounding success stories of the Mark Zuckerbergs, Steve Jobs and Bill Gates of the world, grew the idea in the western collective subconscious that successful entrepreneurs are genius college drop outs that built their company in their dorm room. Our quantitative, data-driven machine learning study shows that this is a myth. The most indicative characteristic of early stage success among the features considered is by far Years of Employment, which turns out to be positively correlated with a post-money valuation increase between the Series A and Series B rounds of investment. In America, the most successful entrepreneurs tend to have 10 to 12 years of prior work experience before they launch their successful ventures.

We do realize that our sample size of 244 entrepreneurs could be increased. However, we believe our approach to this entrepreneurship research as well as the robustness of the method can justify further research to be undertaken in this direction to validate, or dismiss, by utilizing a more comprehensive data set.

Today, in the age of Big Data, there is an exponentially growing quantity of information available to all. This allows new research methodologies that are more quantitative in nature to be applied to new fields of research. Machine Learning and Data Science can offer entrepreneurship research a new set of tools that can showcase patterns and results inferred from these vast collections of data. Only a decade ago studies like these were impossible to conduct.

We imagine that if entrepreneurship researchers either directly have data and Machine Learning tool skills, or if they work in collaboration with others who have these capabilities, then we think that new results could be obtained that would also better complement qualitative work in this area.

V. ACKNOWLEDGEMENTS

We would like to acknowledge our fellow students and friends Julian Chan, Nitin Sampath and Yuan Zho, as well as the staff and faculty of the Sutardja Center for Entrepreneurship and Technology without whom this paper would not have been possible. Thank you, Julian, Nitin and Yuan, for your tireless data collection work as well as precious inputs and support.

VI. Appendix

Name	Type	Description
Previous startups	Discrete	How many other startups did the founder launch ?
Consulting before start-up?	Binary	Did they do consulting before launching their startup?
Standardized University	Categorical	Where did they do their undergraduate studies?
Standardized Major	Categorical	What was their primary field of undergraduate studies?
Degree Type	Categorical	What degree did they obtain? BS, BA, BEng…
Standardized Graduate Institution	Categorical	Where did they do their graduate studies?
Standardized Graduate Studies	Categorical	What was their primary field of graduate studies?
Graduate Diploma	Categorical	What graduate degree did they obtain? MBA, MS, PhD…
Ever served as TA/Teacher/Professor/Mentor ?	Binary	Did they put on their linkedin profile that they served as a Teaching Assistant, Graduate Student Instructor, Professor or Mentor?
Years of Employment	Continuous	How many years were they employed before founding their startup?
Worked as product manager/director/head/VP?	Binary	Did they put in their Linkedin profile that they had the title of product manager, director of product, head of product or VP of product at some point in their career?
Worked at Google?	Binary	Did they put in their linkedin profile that they were ever employed at Google?
Worked at Microsoft?	Binary	Did they put in their linkedin profile that they were ever employed at Microsoft?
Worked in Sales?	Binary	Did they put in their linkedin profile that they ever had a sales related job?

VII. References

[1] Bell, C. & McNamara, J. 1991. High-tech ventures: The guide for entrepreneurial success. Reading, MA: Addison-Wesley Publishing Company, Inc.

[2] Networks of Venture Capital Firms in Silicon Valley: Emilio J. Castilla. 2003.

[3] Bell, C. & McNamara, J. 1991. High-tech ventures: The guide for entrepreneurial success. Reading, MA: Addison-Wesley Publishing Company, Inc.

[4] Pearson, Karl. "Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia", 1896

[5] Székely, Gábor J. et al. "Measuring and testing dependence by correlation of distances", 2007

[6] Breiman, Leo. "Random Forests", 2001

[7] Meinshausen, Nicolai. et al. "Stability Selection", 2009

[8] Coppak, S. W. "Limitations of the Pearson Product-Moment Correlation", *Clinical Science,* 1990

[9] Wang, Sijian et al. "Random Lasso", 2011

[10] Gergonne. "Design and Analysis of Polynomial Regression Experiments", 1815

[11] Scikit Learn 0.19.1, https://www.scikit-learn.org