2016

Predicting Success in Equity Crowdfunding

John (Jack) Beckwith
Wharton, UPenn

Predicting Success in Equity Crowdfunding

Abstract
Equity crowdfunding is an increasingly popular means of raising capital for early stage startups. It enables entrepreneurs to finance their companies with smaller contributions from a variety of people. This paper studies the relationship between the characteristics of a given company and its ability to raise funds on an equity crowdfunding platform. A series of statistical and machine learning models are fit to data from a U.S.-based equity crowdfunding website, including a logistic regression, a CART decision tree, a naïve Bayes classifier, and a support vector machine. This study demonstrates that a connection exists between the probability of a company's crowdfunding success and its previous funding history, Twitter presence, media buzz, size, location, and its founders' educational backgrounds. As a whole, however, the classification quality of the various models leaves something to be desired. This suggests the need for additional data inputs and more longitudinal research in the field of equity crowdfunding.

Keywords
venture capital, equity crowdfunding, machine learning

Disciplines
Business

PREDICTING SUCCESS IN EQUITY CROWDFUNDING

Jack Beckwith

Undergraduate, Wharton School of Business

Pinar Yildirim

Assistant Professor in Marketing

April 27, 2016

ABSTRACT

Equity crowdfunding is an increasingly popular means of raising capital for early stage startups. It enables entrepreneurs to finance their companies with smaller contributions from a variety of people. This paper studies the relationship between the characteristics of a given company and its ability to raise funds on an equity crowdfunding platform. A series of statistical and machine learning models are fit to data from a U.S.-based equity crowdfunding website, including a logistic regression, a CART decision tree, a naïve Bayes classifier, and a support vector machine. This study demonstrates that a connection exists between the probability of a company's crowdfunding success and its previous funding history, Twitter presence, media buzz, size, location, and its founders' educational backgrounds. As a whole, however, the classification quality of the various models leaves something to be desired. This suggests the need for additional data inputs and more longitudinal research in the field of equity crowdfunding.

Keywords: venture capital, equity crowdfunding, machine learning

INTRODUCTION

New business ventures need financial assistance to grow and succeed. Since the early 1980s, angel investors and venture capital firms have served as the primary sources of financial backing for start-up companies (Kortum & Lerner, 2000). In recent years, however, a new pipeline known as "crowdfunding" has emerged to help finance a variety of for-profit, social, and cultural projects across the world (Mollick, 2013). Crowdfunding draws inspiration from the notions of crowdsourcing (Poetz & Schreier, 2012) and micro-finance (Morduch, 1999), but refers to fundraising for new ventures conducted via the Internet. On crowdfunding websites, individuals can raise money for their ideas or projects via small contributions from a large number of people. Online crowdfunding platforms began to appear as early as 2003 with the founding of ArtistShare, a website that allows musicians to seek donations from fans to produce new records (Freedman & Nutting, 2015). Since then, the crowdfunding model has taken off. Data from the Crowdsourcing Industry Report (2015) indicates that crowdfunding websites raised a combined $34 billion in 2015 alone.

Equity crowdfunding refers to a particular model of crowdfunding, in which entrepreneurs may offer, as compensation for financial contributions, an equity stake in their company. Especially for start-ups of limited notoriety, it offers a promising way to raise money without the support of a venture capitalist. As of 2015, equity crowdfunding constituted $2.6 billion of the crowdfunding market globally (Crowdsourcing Industry Report, 2015). Its presence in the U.S. has grown substantially since April 2012, when Barack Obama signed into law the Jumpstart Our Business Startups Act (JOBS Act, 2012). The JOBS Act included Title III, a clause known as the CROWDFUND Act that directly addressed equity crowdfunding (Stemler, 2013). It lifted the ban on general solicitation and general advertising of crowdfunded equity,

allowing companies to issue equity online more liberally. It also provided for the eventual inclusion of "non-accredited" investors[1] on U.S. equity crowdfunding platforms, opening the potential of equity crowdfunding to the masses. Despite the Security and Exchange Commission's slow ratification of the CROWDFUND Act, much of the language was officially approved in October 2015 and should take affect in mid-2016.

Given the legislative changes and mounting interest surrounding equity crowdfunding, it is important to study the investment dynamics that exist on current equity crowdfunding platforms. One would like to understand how entrepreneurs "signal" their qualifications to potential investors on the basis of perhaps their educational background, their company's product description and/or video, or the amount they seek to raise. The concept of signaling theory stretches back to Spence (1973), who hypothesized that job applicants may pursue a higher education to signal their quality and reduce information asymmetry with potential employers. Subsequent research has applied signaling theory to a variety of domains, including the entrepreneur-investor relationship present in fundraising for early-stage ventures.

This study uses data collected from AngelList—a U.S.-based website that connects entrepreneurs and potential investors—to investigate those signals. AngelList offers companies with profiles on the site the ability to equity crowdfund through its platform. Companies may specify the amount of money they are seeking to crowdfund, along with the discount rate of issued equity. The data set includes a variety of information about each company, along with a binary variable indicating whether or not they received any funding.

[1] "Accredited" investors occupy a special status under financial regulation laws. The Dodd-Frank Wall Street Reform and Consumer Protection Act (2010) defines an accredited investor as including a "natural person who has individual net worth, or joint net worth with the person's spouse, that exceeds $1 million…excluding their primary residence…[or]…a natural person with income exceeding $200,000 in each of the two most recent years or joint income with a spouse exceeding $300,000 for those years." Under Regulation D of the Security Act of 1933, equity crowdfunding platforms were allowed to operate so long as companies were only able to solicit funds in exchange for equity amongst a pool of accredited investors. The JOBS Act changed that stipulation so that "non-accredited" investors would be able to invest in limited amounts.

A variety of statistical and machine learning models are employed to understand which factors about a company best predict the outcome of its equity crowdfunding campaign. The models considered—logistic regression, a CART decision tree, a naïve Bayes classifier, and a support vector machine—are all suitable to a classification-type problem. An examination of the logistic regression coefficients shows that, of the eleven features in the final model, ten are statistically significant predictors of funding success. While this doesn't confirm that these predictors are in fact true "signals" that drive investment, it is at least proves that a strong relationship exists. Still, this study finds that none of the four models built are particularly adept at correctly identifying which companies within the data set actually receive funding. Improving classification accuracy is one avenue for further research suggested in the paper's conclusion.

SURVEY OF EXISTING RESEARCH

Given that equity crowdfunding is still in its infancy, the current literature on equity crowdfunding is somewhat sparse. The majority of existing research on equity crowdfunding has focused on the theoretical implications of the JOBS Act. Argawal, Catalini, and Goldfarb (2013) point to the fact that entrepreneurs may derive a number of benefits from equity crowdfunding. First, equity crowdfunding allows entrepreneurs to access individuals with the highest willingness to pay for equity in their ventures on a truly global scale. Second, it enables entrepreneurs to bundle their equity with other rewards that consumers might covet, like early product releases. And third, equity crowdfunding can serve as a validation tool to ensure that there is substantial demand for the product, providing a particularly informative type of market research. Additionally, Macht and Weatherston (2014) hypothesize that equity crowdfunding may be attractive to entrepreneurs because it may require them to relinquish less control over

their companies than a traditional venture capital arrangement. Indeed, most crowdfunders will take on very small shareholding positions, limiting their voting rights and their ability to interfere with the entrepreneur's vision.

Other scholars have been slow to embrace equity crowdfunding as a viable alternative to angel investments. From an entrepreneur's perspective, scholars note that an equity crowdfunding campaign is not without risks. Argawal, Catalini, and Goldfarb (2013) claim that the need to disclose confidential details about an early-stage venture may provide a huge deterrent for entrepreneurs. Publicizing a new company before launch may have negative repercussions for its intellectual property rights and bargaining power with suppliers. Furthermore, entrepreneurs who opt for equity crowdfunding instead of venture capital may miss out on the industry knowledge, guidance, and connections that a VC typically provides. Finally, Valanciene and Jegeleviciute (2013) cite "administrative and accounting challenges" as a major hurdle to the adoption of equity crowdfunding by entrepreneurs. Adding a large number of shareholders will require careful bookkeeping and raise the possibility that the company becomes saddled with investors who have competing personalities and ideas.

Perhaps the most widespread criticism for equity crowdfunding has arisen out of concerns for potential investors. Individuals who help to fund equity crowdfunding ventures will likely be less experienced than the typically venture capitalist or angel investor. They will be less accustomed to reading financial documents and inferring the viability of a new start-up. Moreover, they will be unable to meet entrepreneurs in person, hindering their ability to conduct in-depth due diligence. This may increase the risks associated with investing on equity crowdfunding platforms or even make them a target for fraud (Argawal, Catalini, & Goldfarb, 2013).

Dorff (2013) points out another concern, which stems from the notion of adverse selection derived from information asymmetry. The argument goes that truly promising start-ups might continue to seek traditional angel investors for their seed funding because angels bring a wealth of experience, knowledge, and connections to the table. The start-ups that gravitate towards equity crowdfunding platforms, therefore, would be the ones that have lower potential. If that were the case, equity crowdfunding would essentially become an example of a "market for lemons" (Akerlof, 1970). At the extreme, this could lead to complete market failure. At the very least, it calls into question whether investors can expect any positive returns on investment.

Though the discussion surrounding equity crowdfunding has provided plenty of theoretical debate, few studies have attempted to test these hypotheses with data from equity crowdfunding sites. This is likely a product of the fact that such data isn't widely available yet. Because equity crowdfunding is such a new practice, even the largest crowdfunding sites have only facilitated deals for several hundred companies each. Thus, data set size in an equity crowdfunding study is somewhat constrained for the time being. Researchers are also limited by the fact that most equity crowdfunding deals worldwide have taken place within the past several years. This means that scholars will likely have to wait a few more years before longitudinal data on equity crowdfunded companies becomes a viable source of knowledge. It still remains to be seen how crowdfunded companies will fair and whether funders will make a return on their investment.

Despite the apparent data limitations in equity crowdfunding, at least one empirical paper exists analyzing an emerging equity crowdfunding platform. Ahlers et al. (2015) gathered data on 104 companies that ran an equity crowdfunding campaign between October 2006 and October 2011 on the Australian Small Scale Offerings Board (ASSOB). ASSOB is among the largest

equity crowdfunding sites outside the U.S., with AUD 125 million in funding on the platform as of April 2012. It also operates in a country in which non-accredited investors are fully able to participate in equity crowdfunding, as will be the case in the U.S. by early 2016. In that sense, Ahlers et al. (2015) were able to draw preliminary conclusions about the how equity crowdfunding might soon unfold in the U.S.

Ahlers et al. (2015) focused their work on a single motivating question: "Given different start-ups with similar observable characteristics, what leads small investors to invest in certain start-ups and not others?" Their study follows a number of others that analyze signaling theory—the signals sent by entrepreneurs that lead investors to back early-stage companies (Backes-Gellner and Werner, 2007; Certo, 2003; Hsu and Ziedonis, 2008). It is the first, however, to analyze signaling theory in an equity crowdfunding context. Using the dataset derived from ASSOB, they find that the exit strategy identified by a company has significant impact on its ability to attract investors; traditional forms of external certification—patents, grants, or government awards—did not foster more contributions; and that companies with a larger percentage of MBA graduates on their executive boards tended to garner more investors.

This study represents a notable first step towards more quantitative research in the field of equity crowdfunding. However, it begs further research for a number of reasons. Firstly, this study was conducted on a limited dataset stemming from a single Australian equity crowdfunding platform. Moreover, Ahlers et al.'s (2015) findings contradict several other prominent papers from the field of entrepreneurial signaling, including their determination that external certification doesn't influence fundraising ability. This necessarily raises the question of whether these inconsistencies are a result of the limits of the ASSOB dataset or a different status quo within the realm of equity crowdfunding.

METHODS

The Data Set

The data set is derived from AngelList, a U.S.-based website intended to connect entrepreneurs with potential investors. Companies with AngelList profiles may elect to raise funds through AngelList's equity crowdfunding platform. Prospective investors with "accredited" status may log on, survey company details, and choose to invest online. The data set includes 5,220 companies that solicited financial contributions through AngelList's equity crowdfunding platform through November 2015. Of the 5,220 companies, 2,603 have complete data for the features under consideration. Multiple imputation with chained equations, as detailed in White, Royston, Wood (2010), proved unable to effectively fill in these missing values. As a result, only those 2,603 companies with full data are used to construct the models that follow.

Each company's data contained 85 features of interest. Most of those features come directly from information listed on the start-up's AngelList profile, including details like its location, its product market, its previous mentions in the press, and its number of employees. Data about each company's funding history is collected by cross-referencing the CB Insights database for previous funding rounds and their amounts. If a Twitter handle was listed on the company's AngelList page, data about the company's Twitter activity was gathered using the Twitter API. Finally, if a company's founders provided a link to their Linkedin profiles, information on their educational background was also scraped and integrated into the data set.

For each company, there is a field indicating how much money, if any, it raised during its AngelList crowdfunding campaign. An adaptation of this field serves as the response variable in the following models. It is treated as a binary variable, coded as one if the company raised any non-zero amount of funding and zero otherwise. This field—which is subsequently referred to as

"campaign outcome"—is skewed. Of the 2,603 companies included in the data set, only 284 of them (10.9%) successfully raised money during their AngelList campaign. The skewedness of the outcome variable will function as a major consideration in the evaluation of model fit.

Empirical Models

Four different models are considered for predicting campaign outcome: a traditional logistic regression, a CART decision tree, a naïve Bayes model, and a support vector machine[2].

Logistic Regression

The logistic regression is a generalized linear model suitable for situations where a vector of features *X* is being used to predict a binary outcome. It has wide-ranging use across many disciplines of study dating back to its introduction by Cox (1958). It models the log odds of response variable *y* as a linear combination of the predictors in *X*:

$$\log \frac{P(y = 1 \mid \boldsymbol{X};\ \boldsymbol{\beta})}{P(y = 0 \mid \boldsymbol{X};\ \boldsymbol{\beta})} = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$$

The model is fit using maximum-likelihood estimation across the *n* training instances:

$$\boldsymbol{\beta} = \arg\max log \prod_{i=1}^{n} P(y_i \mid x_i;\ \boldsymbol{\beta})$$

CART Decision Tree

A classification and regression tree (CART) is a machine learning technique used for classification purposes. It recursively partitions the data set into *m* subgroups of instances with similar feature vectors. The instances in each group are then classified according to a simple prediction model, often just an assignment of the outcome that is more predominant within the group. The model is fit so as to minimize the number of classification errors in the training set. For more on the CART algorithm, see Breiman et al. (1984).

[2] The models are fit in Python using the package scikit-learn. It includes a number of useful algorithms to expedite the training and testing of machine learning models.

Naïve Bayes

The naïve Bayes classifier is another model from machine learning that is popular in the domain of text categorization, but has far-reaching applicability. It infers the conditional probability of the response variable *y* given a particular instance from the data x_i by first calculating from the data the marginal probability $P(y)$ and the conditional probabilities $P(X_j \mid y)$ and then applying Bayes' Rule:

$$P(y = k \mid X = x_i) = \frac{P(y = k) \prod_{j=1}^{d} P(X_j = x_{ij} \mid y = k)}{P(X = x_i)}$$

A new point may be classified according to the equation:

$$\hat{y} = \arg\max P(y = k) \prod_{j=1}^{d} P(X_j = x \mid y = k)$$

Support Vector Machine

Finally, a support vector machine (SVM) is considered. SVMs are among the most commonly used machine learning models for classification problems, because they are efficient, work well with few training instances, and are guaranteed to find the globally optimum decision boundary if it exists (Cortes and Vapnik, 1995). The algorithm attempts to minimize:

$$\frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i(\beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d))$$

across all training instances *n*.

Feature Selection

Only a subset of the 85 features will be included in the fitting of various models. Some features have little relationship with campaign outcome and others demonstrate significant collinearity. To assess which features should be incorporated into the subsequent models, a method of feature selection is performed. "Stability selection" is the chosen method, as outlined in Meinshausen and Bühlmann (2009). This technique of feature selection works as follows:

1) Start by creating *n* subsamples of the complete data set *Z* with replacement. Each sample should be of the same size. This study generates 1000 total subsamples and includes 75% of *Z* in each subsample.
2) For each subsample of data *i*:
    a) Fit the baseline model using all eligible features. This study uses the logistic regression model as the baseline for feature selection purposes.
    b) Determine the selection set for the subsample *i* . The selection set is defined:

$$\widehat{S_\iota} = \{\, k : \widehat{\beta_k} \neq 0 \,\}$$

    In other words, the selection set includes the subset of the *k* features that have significant coefficients in the baseline model.
3) Compute the probability that a given feature *k* appears in any of the *i* fitted models:

$$\widehat{\Pi_k} = \mathbb{P}\left(\left\{\, k \in \hat{S} \,\right\}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(\left\{\, k \in \widehat{S_\iota} \,\right\}\right)$$

The probability of selection over all subsamples *i* is an indication of the feature's explanatory power. Features with higher probabilities have significant coefficients across a larger percentage of the fitted models and should therefore be included in the final feature set.

The stability selection technique has several attractive properties. It works well on highly dimensional data, where the number of features is very large. This isn't a major concern here, given that the crowdfunding data set contains fewer than 100 features, but it is an important characteristic of stability selection in other machine learning applications. Second, it provides a bounded guarantee on the number of falsely selected variables included in the final feature set (Meinshausen and Bühlmann, 2009). And third, it offers a robust way of determining which of a number of collinear variables should be kept for modeling purposes. Stability selection assesses

the significance of the collinear variables' coefficients with various subsamples of the data and in models that include different subsets of other features. Of the collinear variables, the one that appears in the most fitted models may be chosen for the final feature set. This is the reason, for instance, that "Twitter Presence" is selected over "Previous Tweet Total," two collinear variables (see Table A in Appendix).

Using the stability selection method, the 85 features are winnowed down to the eleven features that offer the most explanatory power on "campaign outcome." These eleven features are listed in Table B of the Appendix along with a description of each one. Table C shows the average value of each of the eleven features for companies that received funding and for those that didn't. These eleven features are used to fit all models.

Parameter Optimization

The parameters for the logistic regression, CART tree, and SVM models are optimized using a stochastic search process. They were tuned to maximize the F1-score of each model across ten-folds of cross-validation. The F1-score is used because it is easily computed for each model and is a direct reflection of the model's ability to identify successful outcomes efficiently. The F1-score[3] is defined as:

$$F_1 = 2 * \frac{precision * recall}{precision + recall} = 2 * \frac{True\ Pos}{2 * True\ Pos + False\ Pos + False\ Neg}$$

By using the F1-score, the emphasis is placed on the classification of "positive" examples from the data or, in this case, companies that did receive funding. We'd like the models to identify as many of the successfully funded companies as possible, while also limiting the number of

[3] The F1-score is used frequently in machine learning as a gauge of the tradeoff between precision and recall. Precision is the ratio of correctly predicted positive cases to all cases that are predicted positive by the model. It is often described as the number of predicted positives that are "relevant." Alternatively, recall represents the "completeness" of the results, defined as the ratio of correctly predicted positives to all true positives from the data set. Ideally, both precision and recall would be one, but almost invariably that turns out impossible to achieve.

unfunded companies that the models incorrectly classify as funded. This is the trade-off captured by the F1-score.

The optimal parameters for each model may be viewed in Table D of the Appendix. For the support vector machine, a randomized grid search is used to efficiently search continuous distributions of possible parameter values, as recommended in Bergstra and Bengio (2012). For the logistic regression and CART decision tree, a complete grid search is conducted because the possible parameter values are discrete and the parameter search space is sufficiently small.

RESULTS

In this section, the models are evaluated by the quality of their fit to the AngelList data set. First, the implications of the logistic regression coefficients are reviewed in detail. Then the logistic regression's fit is compared to that of the other three models according to relevant criteria like precision, recall, and the F1-score.

Analyzing the Logistic Regression

Table E in the Appendix reports the logistic regression estimation results on the AngelList data set. Of the eleven features selected by stability selection, ten of them are statistically significant in the logistic regression at the $p < 0.05$ level. Only "Mentions in TechCrunch" fails to meet the $p < 0.05$ threshold, but just barely ($p = 0.082$). The strongest predictors—"Debt", "Previous Rounds of Funding", "Twitter Presence", and "11-50 Employees"—are highly significant ($p < 0.001$). In other words, choosing to raise debt instead of equity funding, having raised a previous round of funding, running a corporate Twitter account, and having 11-50 employees at the company (as opposed to 1-10) seem to be associated with a higher likelihood of equity crowdfunding success on AngelList, holding all else constant.

The coefficient of “Founder with MBA” is also positive and statistically significant ($p = 0.029$), seeming to echo Ahlers et al. (2015) in suggesting that holding an MBA is associated with more positive equity crowdfunding outcomes. For a sample of actual company profiles from the data, along with their associated probabilities of success according to the logistic regression, see Table F in the Appendix.

Ten of the eleven features (excluding the intercept) have positive signs, indicating that they share a positive relationship with the outcome variable. Only “S&P Close Previous Day” has a statistically significant negative coefficient (-0.0011). This predictor was intended to serve as a partial control on the country’s economic health, with the hypothesis being that users on AngelList might be more likely to invest during booming periods irrespective of company attributes. The fact that “S&P Close Previous Day” has a negative coefficient is a curious finding; it implies that, holding all else constant, a lower closing value for S&P 500 on the day prior to the start of an equity crowdfunding campaign is related to a higher probability of investment. The coefficient must be interpreted with caution, because the value is partially a reflection of the other predictors included in the model. However, Table C does demonstrate that on average unfunded companies tend to open their campaigns on the day after a higher S&P 500 close (1718.18) than funded companies (1676.87). Further research must be conducted to understand if this finding is an artifact of this particular data set, the AngelList investment setting, or a general phenomenon across equity crowdfunding platforms.

Still, the implications of the logistic regression are notable. The majority of the features selected through stability selection are positive and statistically significant. Further research will need to be undertaken to determine whether there is truly a causal relationship between any of the features and the outcome variable, as the signaling theory would dictate. But at the very least,

the coefficients of the logistic regression indicate that there is a connection between a company's location ("San Francisco"), its social media activity ("Twitter Presence"), its buzz in the press ("Previous Press Mentions"), and its founders' educational backgrounds ("Founder That Attended Top 20 School" and "Founder with MBA") and the potential for a successful equity crowdfunding effort on AngelList.

Evaluation of Model Fit

Table G shows a variety of metrics that demonstrate the classification quality of each model. The values displayed in the table are averages calculated across ten-folds of cross-validation. The metrics may be defined as follows[4]:

$$Accuracy = \frac{True\ Pos + True\ Neg}{True\ Pos + True\ Neg + False\ Pos + False\ Neg}$$

$$Precision = \frac{True\ Pos}{True\ Pos + False\ Pos} \qquad Recall = \frac{True\ Pos}{True\ Pos + False\ Neg}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

As may be seen from Table G, no model clearly dominates the others in terms of classification performance. The logistic regression does best according to overall accuracy, but given the fact that the data is highly skewed, accuracy is misleading. By simply predicting that all companies fail in their fundraising efforts, an accuracy of over 88% could be achieved. Therefore, the other three metrics—precision, recall, and the F1-score—better demonstrate the classification performance in this case.

All four models seem to perform admirably according to average precision, average recall, and the average F1-score. But upon further investigation, it becomes clear that those

[4] In machine learning, a "true positive" denotes an instance in the data set that is predicted to be of the positive class by the model (i.e. funded) and actually is. A "false positive" denotes an instance that is predicted to be positive but is actually negative. The terms "true negative" and "false negative" are defined similarly.

numbers are inflated by the fact that the models are much better at correctly identifying the more preponderant unfunded companies. When it comes to classifying companies that actually received funding, the models are markedly worse. The logistic regression again performs best in terms of average precision across the ten-folds of cross-validation (0.41). Still, this demonstrates that, of the companies that the logistic regression predicts as funded, only 41% of them actually are. Moreover, the logistic regression's recall score is the worst among the four models, illustrating the tradeoff between precision and recall. The naïve Bayes classifier seems to best at balancing this tradeoff with similar precision (0.33) and recall (0.34) scores and the highest F1-score of the four models considered (0.33).

As a secondary indication of classification performance, I present the Receiver Operating Characteristic Curve (ROC Curve) for all four models in Figure 1. The ROC Curve plots the rate of true positives to the rate of false positives for a given model. In other words, as you prod the model to predict more and more positives, how many true positive predictions do you gain for all the false positives you create? A perfect model would hit a true positive rate of 100% immediately and never waver, meaning the curve would jump from the origin to the upper left hand corner. A random model would follow the dotted line shown in Figure 1 on which the true positive rate and false positive rate increase in proportion. All four of the fitted models clearly outperform a random one, with the logistic regression and naïve Bayes doing best according to ROC. Still, they are nowhere nearly the ideal.

CONCLUSION

This study is among the first quantitative analyses conducted using data from an equity crowdfunding platform. A series of models are built to understand the way in which various

company characteristics are related to probable campaign outcomes. It contributes to the emergent literature on crowdfunding by demonstrating that there exists a relationship between a company's likelihood of crowdfunding success and its previous funding history, Twitter presence, media buzz, size, location, and its founders' educational backgrounds. These associations are made clear by the slate of positive, statistically significant coefficient values that appear in a logistic regression fitted to the data set.

This paper also contributes by showing that there remains much to learn about the investment dynamics that currently exist on an equity crowdfunding platform. Despite best efforts to select relevant features and tune model parameters, none of our four models were overly impressive in their ability to correctly classify companies according to whether or not their campaign would be successful. Precision and recall scores were markedly low, especially for the subset of companies that ultimately did receive funding on AngelList. More complex classification algorithms—artificial neural networks, Restricted Bolzmann machines, for instance—could be tried on the data set, but marginal improvements would likely result.

In reality, no model, no matter how sophisticated, can wholly supplant the need for better, more representative data. Indeed, this study was forced to remove about half of the training instances because of missing data for the educational fields ("Founder with MBA" and "Founder that Attended Top 20 School"). Moreover, there are almost certainly other factors that would demonstrate a significant relationship with the outcome variable that are unaccounted for here. None of the features included in the final feature set explicitly captures the company's product market, its video's content and quality, or the strength of its founders' professional networks (although perhaps "Previous Funding Round" may serve as a proxy). More effort should be made to collect data on companies that choose equity crowdfunding for raising capital.

More academic work in equity crowdfunding must also be done to understand the direction of causality (if any) that exists between company attributes and campaign success. The significant coefficients in the logistic regression simply demonstrate that a host of relationships *exist*; it does nothing to show that having a company Twitter or a founding team with a more impressive educational profile actually *causes* users on AngelList to invest. This requires a much more stringent set of conditions. Perhaps the most feasible way to infer causality would be through a controlled experiment in which subjects are offered a series of investment options on a fictitious, equity crowdfunding platform. Underlying characteristics in each company profile could be incrementally tweaked in an attempt to parse out drivers of investment.

Still another line of future research arises naturally from the changes taking place in equity crowdfunding as a result of the JOBS Act. The data incorporated into this study was collected in November 2015, meaning that it reflects investment decisions made only by a pool of "accredited" investors. As a result of the JOBS Act, soon AngelList and other equity crowdfunding platforms will be able to provide investment opportunities for "non-accredited" investors as well. This change has the potential to drastically alter the composition of investor pools on many equity crowdfunding platforms. Moreover, the preferences of non-accredited investors may impact the probability that a given company will receive funding. Subsequent studies must test whether the insights derived in this paper still hold on a platform where accredited and non-accredited investors are allowed to intermingle.

And, as a final avenue of future research, the longitudinal effects of using an equity crowdfunding platform need to be measured. It would be informative for both entrepreneurs and investors alike to understand how choosing an equity crowdfunding campaign over VC funding affects a company's growth rate, the likelihood of subsequent funding, or its ability to attract

high quality talent. Given how recently most of the companies in the AngelList data set launched their equity crowdfunding campaigns, such a comparison was not yet possible. After more time elapses and equity crowdfunding becomes more established in the U.S. among non-accredited investors, it is believed that such comparisons will shed much needed light on the equity crowdfunding model's ability to produce profitable companies. Only through this line of research will Dorff's (2013) hypothesis that equity crowdfunding represents a "market for lemons" be tested. Such considerations are left to future research.

APPENDIX

Feature	Selection Rate*
Previous Round of Funding	56.4%
San Francisco	51.0%
Previous Funding Round Count	50.6%
Twitter Presence	49.4%
Previous Press Mentions	48.2%
11-50 Employees	48.0%
Founder that Attended Top 20 School	45.0%
Mentions in TechCrunch	40.4%
Debt	30.0%
Days on Twitter	28.3%
Number of Founders Listed	27.8%
Percentage that Attended Top 20 School	24.4%
Founder with MBA	15.5%
Previous Funding Total	12.1%
Previous Tweet Total	10.5%
S&P Close Previous Day	8.4%
Percentage with MD	4.9%
Percentage with MBA	1.2%
Mentions in VentureBeat	0.5%
Mentions in Forbes	0.5%
Software as a Service Company	0.2%
All other features	0.0%

Table A: Results from stability selection on the AngelList data set

*Selection Rate = percentage of randomized models in which the feature was selected by the stability selection algorithm

The threshold for inclusion in the final feature set was a 5.0% selection rate. If variables were collinear (i.e. Twitter Presence and Previous Tweet Total) only one of them was included in the final feature set.

Feature	Description	Source
Previous Round of Funding	A binary variable indicating whether or not the company has previously received funding	CB Insights
Twitter Presence	A binary variable indicating whether or not the company has a Twitter profile	Twitter API
Previous Press Mentions	A count variable indicating the number of times the company has been mentioned in any press publication	AngelList
Mentions in TechCrunch	A count variable indicating the number of times the company has been mentioned in TechCrunch	AngelList
Number of Founders Listed	A count variable indicating the number of people listed as a co-founder on the company's AngelList profile	AngelList
11-50 Employees	A binary variable indicating whether or not the company's AngelList profile lists 11-50 employees as the company's size	AngelList
San Francisco	A binary variable indicating whether or not the company is based in San Francisco	AngelList
Founder with MBA	A binary variable indicating whether, among the founders listed on AngelList, there is at least one founder who holds an MBA	LinkedIn
Founder that Attended Top 20 School	A binary variable indicating whether, among the founders listed on AngelList, there is at least one founder who attended a Top 20 U.S. university for any level of schooling	LinkedIn
S&P 500 Close Previous Day	The closing number for the S&P 500 on the day prior to the launch of the company's AngelList crowdfunding campaign	Yahoo! Finance

Table B: A look at all the significant features identified through stability selection.

Feature	Campaign Outcome = 0	Campaign Outcome = 1
Previous Round of Funding	0.03	0.14
Twitter Presence	0.36	0.57
Previous Press Mentions	0.95	2.84
Mentions in TechCrunch	0.01	0.08
Number of Founders Listed	1.70	2.00
11-50 Employees	0.05	0.15
San Francisco	0.04	0.13
Debt	0.12	0.31
Founder with MBA	0.12	0.21
Founder that Attended Top 20 School	0.10	0.22
S&P 500 Close Previous Day	1718.18	1676.87

Table C: Averages of the 10 features for both classes of the response variable.

Model	Parameter Search Method	Optimal Parameters
Logistic Regression	Exhaustive Grid Search	Class Weights[5]: {0 : 0.15, 1: 0.85}
CART Decision Tree	Exhaustive Grid Search	Class Weights: {0: 0.2, 1: 0.8} # of features to consider when looking for best split: sqrt(max) Maximum depth of tree: 3 Split Criterion: entropy
Support Vector Machine	Randomized Grid Search	Class Weights: {0: 0.2, 1: 0.8} C: 175 Gamma: 0.0026

Table D: Optimal parameter values of all models considered according to grid search.

[5] The "class weights" parameter denotes the amount of weight placed upon a misclassification of each class during the training phase. In other words, if class weight = {0: 0., 1: 0.9}, a misclassification of a successfully funded company is four times as costly as a misclassification of an unfunded company. Altering the class weights is intended to steer the classifier towards making more positive predictions and to counteract the skewedness of the response variable.

Feature	Coefficient	Std Error	z	P>|z|	95% Conf. Int.	
Intercept	-1.3130	0.596	-2.202	0.028	-2.482	-0.144
Previous Round of Funding	1.1088	0.241	4.594	0.000	0.636	1.582
Twitter Presence	0.5191	0.138	3.768	0.000	0.249	0.789
Previous Press Mentions	0.0359	0.013	2.805	0.005	0.011	0.061
Mentions in TechCrunch	0.5215	0.300	1.740	0.082	-0.066	1.109
Number of Founders Listed	0.1289	0.065	1.980	0.048	0.001	0.256
11-50 Employees	0.7994	0.212	3.762	0.000	0.383	1.216
San Francisco	0.6992	0.224	3.123	0.002	0.260	1.138
Debt	0.8779	0.155	5.662	0.000	0.574	1.182
Founder with MBA	0.3847	0.176	2.180	0.029	0.039	0.731
Founder that Attended Top 20 School	0.5165	0.179	2.889	0.004	0.166	0.867
S&P Close Previous Day	-0.0011	0.000	-3.259	0.001	-0.002	-0.000
Log-Likelihood = -790.59			Pseudo-R^2 = 0.1187			

Table E: The logistic regression model. All features are significant at p = 0.05 aside from Mentions in TechCrunch.

Feature	Nimble*	Mednyma*	Zodio**
Previous Round of Funding	1	0	0
Twitter Presence	1	0	0
Previous Press Mentions	11	0	0
Mentions in TechCrunch	2	0	0
Number of Founders Listed	1	1	1
11-50 Employees	1	0	1
San Francisco	0	0	0
Debt	1	0	0
Founder with MBA	0	0	1
Founder that Attended Top 20 School	0	0	0
S&P 500 Close Previous Day	1472.05	1319.68	1265.33
$P(outcome = 1 \mid \boldsymbol{X})$	0.874	0.112	0.249
Actual Outcome	Successful	Unsuccessful	Successful

Table F: Three example company profiles along with the logistic regression predicted outcome given their attributes. *Correct Prediction **Incorrect Prediction

Model		Accuracy	Precision	Recall	F1-Score
Logistic Regression	Campaign Outcome = 0		.90	.96	.93
	Campaign Outcome = 1		.41	.19	.26
	Average	.87	.85	.88	.86
CART Decision Tree	Campaign Outcome = 0		.92	.75	.82
	Campaign Outcome = 1		.15	.40	.22
	Average	.85	.84	.71	.76
Naïve Bayes model	Campaign Outcome = 0		.92	.91	.91
	Campaign Outcome = 1		.33	.34	.33
	Average	.83	.85	.85	.85
Support Vector Machine	Campaign Outcome = 0		.91	.92	.92
	Campaign Outcome = 1		.27	.23	.24
	Average	.86	.84	.85	.84

Table G: A look at the precision, recall, and F1-scores for each model. The model parameters were tuned to maximize the F1-score when the true value of campaign outcome is one. Values reflect the average across 10-fold cross-validation on the holdout set.

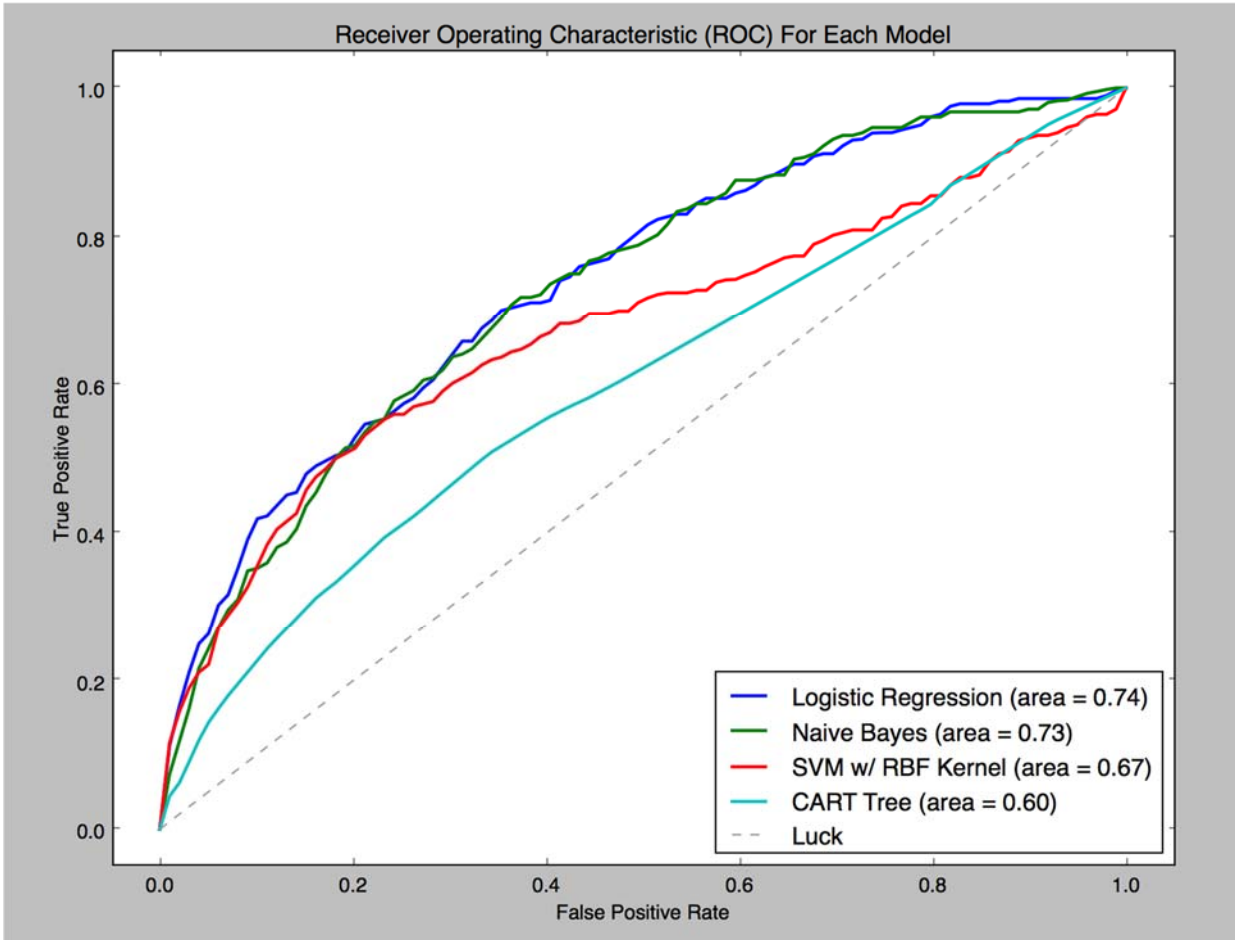

Figure 1: The ROC Curve for each model

REFERENCES

Agrawal, A. K., Catalini, C., & Goldfarb, A. (2011). *The geography of crowdfunding* (No. w16820). National bureau of economic research.

Ahlers, Gerrit KC, Douglas Cumming, Christina Günther, and Denis Schweizer. "Signaling in equity crowdfunding." *Entrepreneurship Theory and Practice* 39, no. 4 (2015): 955-980.

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *The quarterly journal of economics*, 488-500.

Backes-Gellner, U., & Werner, A. (2007). Entrepreneurial signaling via education: A success factor in innovative start-ups. *Small Business Economics*, *29*(1-2), 173-190.

Bergstra, James, and Yoshua Bengio. "Random search for hyper-parameter optimization." *The Journal of Machine Learning Research* 13, no. 1 (2012): 281-305.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and Richard A. Olshen.*Classification and regression trees*. CRC press, 1984.

Certo, S. T. (2003). Influencing initial public offering investors with prestige: Signaling with board structures. *Academy of management review*, *28*(3), 432-446.

Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of Management*, *37*(1), 39-67.

Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine learning* 20, no. 3 (1995): 273-297.

Crowdfunding Industry Report (2015). Crowdfunding Industry Report: Market Trends, Composition and Crowdfunding Platforms: Crowdsourcing, LLC.

Cox, David R. "The regression analysis of binary sequences." *Journal of the Royal Statistical Society. Series B (Methodological)* (1958): 215-242.

Dorff, M. B. (2013). Siren Call of Equity Crowdfunding, The. *J. Corp. L.*, *39*, 493.

Fink, A. (2012). Protecting the crowd and raising capital through the JOBS Act. *Available at SSRN 2046051*.

Freedman, D., & Nutting, M. (2015). *Equity Crowdfunding for Investors: A Guide to Risks, Returns, Regulations, Funding Portals, Due Diligence, and Deal Terms*. Hoboken, NJ: Wiley.

Hsu, D. H., & Ziedonis, R. H. (2008, August). PATENTS AS QUALITY SIGNALS FOR ENTREPRENEURIAL VENTURES. In *Academy of Management Proceedings* (Vol. 2008, No. 1, pp. 1-6). Academy of Management.

Jumpstart Our Business Startups Act of 2012, 15 U.S.C. §§ 78 (2012).

Kortum, S., & Lerner, J. (2000). Assessing the contribution of venture capital to innovation. *RAND journal of Economics*, 674-692.

Macht, S. A., & Weatherston, J. (2014). The Benefits of Online Crowdfunding for Fund-Seeking Business Ventures. *Strategic Change*, *23*(1-2), 1-14.

Meinshausen, Nicolai, and Peter Bühlmann. "Stability selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, no. 4 (2010): 417-473.

Morduch, J. (1999). The microfinance promise. *Journal of economic literature*, 1569-1614.

Poetz, M. K., & Schreier, M. (2012). The value of crowdsourcing: can users really compete with professionals in generating new product ideas?. *Journal of Product Innovation Management*, *29*(2), 245-256.

Security Act of 1933, 15 U.S.C. §§ 77 (2012).

Spence, M. (1973). Job market signaling. *The quarterly journal of Economics*, 355-374.

Valanciene, L., & Jegeleviciute, S. (2013). Valuation of crowdfunding: benefits and drawbacks. *Economics and Management*, *18*(1), 39-48.

White, Ian R., Patrick Royston, and Angela M. Wood. "Multiple imputation using chained equations: issues and guidance for practice." *Statistics in medicine* 30, no. 4 (2011): 377-399.