Predicting Startup Crowdfunding Success through Longitudinal Social Engagement Analysis

Qizhen Zhang, Tengyuan Ye, Meryem Essaidi, Shivani Agarwal, Vincent Liu and Boon Thau Loo
University of Pennsylvania
{qizhen,tengyy,essaidim,ashivani,liuv,boonloo}@seas.upenn.edu

ABSTRACT

A key ingredient to a startup's success is its ability to raise funding at an early stage. Crowdfunding has emerged as an exciting new mechanism for connecting startups with potentially thousands of investors. Nonetheless, little is known about its effectiveness, nor the strategies that entrepreneurs should adopt in order to maximize their rate of success. In this paper, we perform a longitudinal data collection and analysis of AngelList - a popular crowdfunding social platform for connecting investors and entrepreneurs. Over a 7-10 month period, we track companies that are actively fund-raising on AngelList, and record their level of social engagement on AngelList, Twitter, and Facebook. Through a series of measures on social engagement (e.g. number of tweets, posts, new followers), our analysis shows that active engagement on social media is highly correlated to crowdfunding success. In some cases, the engagement level is an order of magnitude higher for successful companies. We further apply a range of machine learning techniques (e.g. decision tree, SVM, KNN, etc) to predict the ability of a company to successfully raise funding based on its social engagement and other metrics. Since fund-raising is a rare event, we explore various techniques to deal with class imbalance issues. We observe that some metrics (e.g. AngelList followers and Facebook posts) are more significant than other metrics in predicting fund-raising success. Furthermore, despite the class imbalance, we are able to predict crowdfunding success with 84% accuracy.

1 INTRODUCTION

In recent years, *crowdfunding* has emerged as a financing mechanism that has gained wide-spread popularity. In crowdfunding, a startup uses a portal such as AngelList [1], Fundable [4], or EquityNet [3] to launch a fund-raising campaign. The first generation crowdfunding platforms such as Kickstarter [6] are used to raise small amounts of funding for pet projects. In startup crowdfunding, the motivation for fund-raising is different. Investors may pledge amounts of funding as little as $1000 for equity, or invest much larger amounts as a group (otherwise known as "syndicates"). Crowdfunding companies (predominantly technology startups) then leverage social media to raise awareness among potential backers. Hence, they release a massive amount of online material concerning opinions on their industry and the background of their team. "Buyers" correspond to accredited investors that can choose to make small risky investments in growing companies.

Since this funding mechanism is a relatively new phenomenon, it is unclear whether crowdfunding is as effective for entrepreneurs to raise funding as more traditional approaches, which rely on word-of-mouth introductions and face-time with professional investors. On the one hand, the barrier to success is lower. Indeed, investors in crowdfunded companies often perform less due diligence (compared to traditional investors), due to the small amounts of capital they invest and their general lack of expertise. Companies seeking to raise capital via crowdfunding mechanisms also benefit from the ability to reach a large number of potential investors quickly, particularly via social networks that are integrated with or inherently part of these crowdfunding platforms. On the other hand, investors have a large number of companies to choose from. Given the lack of actual face-time, a company may have a hard time convincing investors to invest in it.

In this paper, we seek to answer the following questions. First, how effective is crowdfunding as a tool for fund-raising? Second, what factors result in successful fund-raising? Does a startup's ability to disseminate information about itself through social media help attract investors? Can we use a startup's social media activity levels to predict its likelihood of fund-raising success? Successful fund-raising is often time a pre-requisite to a startup's survival and subsequent success, and being able to answer these questions will help to shed some light on factors impacting a startup's success.

True causality is, of course, notoriously difficult to determine in real-world situations. Instead, what we measure is correlation, and we do so by carrying out a 10 month longitudinal study of companies actively raising funding on AngelList - a popular crowdfunding website - and their corresponding third party social networks Facebook and Twitter. We track the level of social engagement of companies currently fund-raising, and relate these rates to their ability to raise funding over time. In particular, this paper makes the following key contributions:

Longitudinal data collection. We have carried out a systematic data collection process over a 7-10 month period across different social platforms. On a daily basis, using public APIs, we have gathered data from AngelList, Twitter (twitter.com), and Facebook (facebook.com). AngelList is a U.S. website for startups, angel investors, and job seekers interested in working for startups. We tracked companies that have actively started fund-raising campaigns to determine their success rates, and correlate these rates to their level of social engagement on Facebook and Twitter.

Correlation of social engagement and fund-raising. Our first set of studies aims to quantify the relationship between the level of social engagement and fund-raising success. We observe that

CIKM'17 , November 6–10, 2017, Singapore, Singapore
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4918-5/17/11…$15.00
https://doi.org/10.1145/3132847.3132908

very few companies succeed in fund-raising on AngelList. The ones that do succeed however, are significantly more active on social networks to market themselves. For instance, successful companies are 5 times more likely to participate in social networks such as Facebook and Twitter than unsuccessful ones. Moreover, they are more likely to engage their users on these social networks. To measure this effect, we first define a series of social engagement metrics based on the number of posts, tweets, followers, and trace their changes over the duration of fund-raising. We observe that successful companies are significantly more active on social media than unsuccessful ones, and in some cases, the difference in engagement level is an order of magnitude larger. In fact, we observe that the impact spans across social networks, e.g. engagement on other social networks (Facebook and Twitter) is correlated to company's fund-raising performance on AngelList. In addition, we observe that the description text of a company is also helpful to differentiate successful and unsuccessful companies in crowdfunding.

Predicting fund-raising success from social engagement metrics. We apply a range of machine learning techniques (e.g. decision tree, SVM, k-nearest-neighbor, etc.) to predict the ability of a company to successfully raise funding based on its social engagement and other metrics. Since fund-raising is a rare event, we explore various techniques (over-sampling, under-sampling, and cost-sensitive learning) to deal with class imbalance issues. We also propose a greedy feature selection algorithm that aims to select only the most influential features for building the classification model. We observe that some social engagement metrics (e.g. AngelList followers and Facebook posts) are more significant than other metrics in predicting fund-raising success. Moreover, company description text length is also a good predictor of fund-raising success. Using these three metrics, we are able to achieve 84% accuracy (using the A-mean [24] measure) in predicting the fund-raising success, suggesting that using social engagement metrics can be an effective predictor of a startup's fund-raising success.

2 BACKGROUND

We first provide a background introduction to crowdfunding and the data sources that we have used. In the context of startups, crowdfunding is the practice of funding a venture by raising monetary contributions from a large number of people, often performed via Internet websites nowadays. The crowdfunding website that we focus on in this paper is AngelList, as it is widely used, and provides a public API to collect data. According to its online description [2], AngelList is a US website for startups, angel investors, and job-seekers looking to work at startups. The site started as an online introduction board for tech startups that needed seed funding, and evolved into one that allows startups to raise funding from accredited angel investors. AngelList is now one of the most popular crowdfunding websites in the world.

AngelList allows anyone to register and log in as an independent user. In AngelList, one can serve the role of startup founder, investor, or employee. The website allows companies to publicize their profiles, launch fund-raising campaigns, advertise jobs, and provide links to their social media websites (Twitter, Facebook).

A startup's AngelList profile page contains many features such as its overview, activity, followers, founders, team, and funding. This profile page includes several relevant links, such as the homepages of all the involved people (founders, investors, and employees), the startup's official website, as well as its Twitter, Facebook, and LinkedIn accounts. In this way, AngelList is similar to social media websites like Twitter and Linkedin, and forms a huge social networking graph in the startups' world.

AngelList allows companies to start public fund-raising campaigns. In these campaigns, companies can advertise a fund-raising goal and a duration. Progress on amount raised is displayed during their pre-set fund-raising period.

3 DATA COLLECTION

Using AngelList's default API, we perform a crawl that allows us to collect a snapshot of information on 744,036 startups. In addition to doing the initial crawl, we further use the AngelList API to track all the companies that are currently actively fund-raising. All in all, we track 4001 companies that are actively fund-raising over a 7 month period. For each company, we track the amount of funding raised over time, together with all other information on the company, including its stated valuation, target amount raised, number of followers, etc.

The AngelList dataset includes links to startups' available Facebook and Twitter URLs. Among the 4001 companies that we track, we further gather information on a daily basis of the Facebook and Twitter activities of these companies.

Facebook. Within the companies on AngelList that are actively fund-raising, we invoke the Facebook API [5] on a daily basis to get new updates on the number of likes and additional posts. Note that not all actively fund-raising companies on AngelList own Facebook accounts that are publicly accessible. In total, we tracked 388 Facebook accounts.

Twitter. Finally, we use the tweepy python library [7] to call the Twitter RESTful API methods to extract data of the Twitter platform. We track the number of new tweets posted on a daily basis, for all companies that are fund-raising on AngelList. In total, we tracked 1530 companies.

The initial AngelList snapshot was obtained just prior to the start of our longitudinal data collection. Our longitudinal data collection period is from 21 Dec 2015 to 21 July 2016 for AngelList, and from 21 Dec 2015 to 7 Oct 2016 for Facebook and Twitter. We are able to obtain 3 additional months of data from Facebook and Twitter which we are not able to on AngelList, due to recent AngelList API changes. Nevertheless, all analysis in the paper is done by matching changes from these social networks in the same time period exclusively.

4 CORRELATION ANALYSIS

We first examine whether social engagement and fund-raising success are correlated, followed by studying the impact of other startup attributes. We limit our analysis to data obtained from 21 Dec 2015 to 21 July 2016, over a 7 month period. We omit the final three months of Facebook/Twitter data (from 22 July 2016 until 7 Oct 2016) from our analysis, in order to correlate in time the actual social engagement and fund-raising period.

In our dataset, there are 4001 companies actively fund-raising on AngelList. This represents 0.53% of companies on AngelList, showing that there is a sizable number of companies who are on AngelList for other purposes (e.g. for hiring, publicity, or connecting to future investors). We observe that the startups that are actively fund-raising are spread out geographically across the world (in Africa, Asia, Australia, Europe, and North/South America), covering 926 different cities, of which the top 5 cities are New York City (6.22%), San Francisco (4.32%), Los Angeles (4.22%), London (2.75%) and Chicago (1.75%).

During this period, only 23 out of the 4001 companies attempting to raise funding are successful, which represents a success rate of 0.57%. Although the spread of cities is large, successful startups are predominantly based in the US. Indeed 20 out of the 23 successful startups are based in the US. These successful companies raise $193,665 on average (and $87,404 median) during the 7 month period (though the average total amount raised is higher at $660,038, since these companies may have already accumulated some funding raised prior to the beginning of our data collection). The most successful company raised $2M while the least successful company raised $5K, showing that there is a wide range in fund-raising success. To raise this amount, companies require 1-3 *funding events*, where each event can either correspond to individual investors investing small amounts individually or a group of investors investing larger amounts as a syndicate. While this percentage is not high, the success rate is fairly typical given the low success rate of startups. In this section, we divide up companies into two categories (successful and not successful in fund-raising) and measure their social engagement levels.

Social network participation and correlation with fund-raising success. We first analyze whether there is correlation between the presence/absence on a social network and fund-raising success. We observe strong correlation. Among successful companies, 47.8% are on Facebook and 69.6% are on Twitter. In contrast, among unsuccessful companies, only 9.5% and 38% are on Facebook and Twitter respectively. This shows that on average, companies that are successful at fund-raising are 5X and 1.8X more likely to be on Facebook and Twitter respectively, compared to companies that are not successful. Moreover, 47.8% of successful companies are on *both* Facebook and Twitter, while the rate is only 8.5% for unsuccessful companies, which represents a 5.6X difference in participation level.

Social engagement level and correlation with fund-raising success. Presence on social media is an important parameter, but level of engagement in social activity itself is even more so. Our next analysis focuses on AngelList companies that are actively fund-raising, and are also on either Facebook and/or Twitter. To measure the level of social engagements of these companies, we use the following social engagement metrics:

- **AngelList Followers (AFollowers).** The number of followers of a company on AngelList. This is usually a good proxy metric on the level of engagement by a company on AngelList, which comes in the form of having a complete profile with product information and regular news feeds.
- **Facebook Likes (FBLikes).** The number of Likes of a company on Facebook. Note that this is not Likes for a given post, but rather for the company itself. This is also a good proxy metric for the overall engagement level of a company with its potential customers/investors.
- **Facebook Posts (FBPosts).** The number of posts by the company on Facebook.
- **Twitter Tweets (Tweets).** The number of tweets by the company on Twitter.
- **Twitter Followers (TFollowers).** The number of followers for the company on Twitter.

For each of the above metrics, we explore two different measures: (1) *Delta* is the difference between the metric at the end and beginning of data collection, and (2) *Average* is the average of each of the above metric over the data collection period. The *Delta* metric captures the change from beginning to end, and hence shows the social engagement activity level during our analysis period. The *Average*, on the other hand, also factors in the initial values in social engagement, for example, a company that starts off with a high initial number of Facebook posts will have a high average, even if the subsequent number of posts are small during our measurement period. The latter metric is more useful as a way to measure a company's critical mass for its audience.

	Successful		**Unsuccessful**	
	Mean	**Median**	**Mean**	**Median**
Delta_AFollowers	5.28	3.00	0.19	0.00
Delta_FBLikes	1,432.45	106.00	505.03	4.00
Delta_FBPosts	196.91	31.00	60.62	3.00
Delta_Tweets	1,522.00	55.00	134.66	0.00
Delta_TFollowers	232.00	55.50	81.18	0.00

Table 1: Delta social engagement for companies successful and unsuccessful at fund-raising (summary table).

	Successful		**Unsuccessful**	
	Mean	**Median**	**Mean**	**Median**
Average_AFollowers	152.45	34.56	15.42	6.00
Average_FBLikes	3,928.84	1,354.37	6,316.06	701.23
Average_FBPosts	795.39	490.65	497.60	214.06
Average_Tweets	10,660.90	1,319.13	1,639.17	295.72
Average_TFollowers	2,982.32	900.14	2,730.54	257.50

Table 2: Average social engagement for companies successful and unsuccessful at fund-raising (summary table).

4.1 Social Engagement Analysis

Table 1 shows a summary of different social engagement metrics using the *Delta* measure, averaged across all companies in two categories: companies successful in raising funds on AngelList and those that do not. We observe that there is a significant difference in the level of social engagements of companies that are successful in fund-raising. For example, successful companies send out on average 1522 tweets during 7 months. However, unsuccessful companies send out on average only 135 tweets, representing a 11.3X increase in the number of tweets among successful companies. In fact, more than half of the unsuccessful companies do not engage in social media at all (as shown by the zero median values). This is in stark contrast to successful companies that are significantly more active. Based on the Delta measures, the metrics that show the most significant difference is AngelList Followers (27.2X), followed by Twitter tweets (11.3X), and Facebook posts (3.2X).

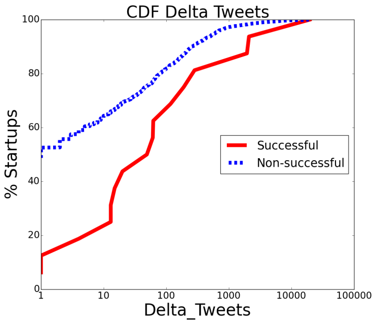

Figure 1: CDF of Tweets (X-axis in log-scale) using the Delta measure.

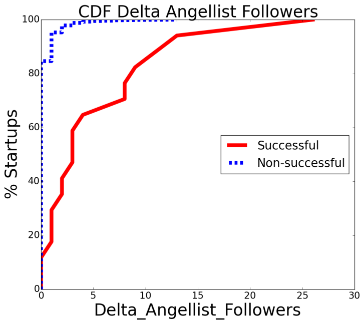

Figure 2: CDF of AFollowers using the Delta measure.

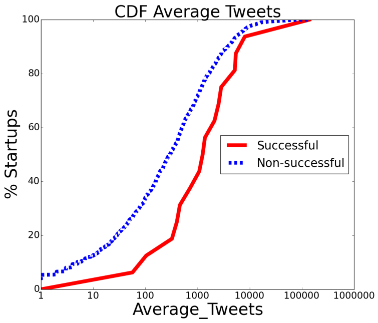

Figure 3: CDF of Tweets (X-axis in log-scale) using the Average measure.

The absolute numbers in the social engagement also matter, as shown by social engagement metrics based on the *Average* measure in Table 2. For example, successful companies have on average 10,660 tweets during this period, while unsuccessful companies have 1,639 tweets. In aggregate, successful companies have 9.9X more AngelList followers, 6.5X more tweets, and 1.6X more Facebook posts than unsuccessful companies. Interestingly, for both Delta and Average measures, we observe that the impact of social engagement spans *across* social networks. For example, being active in Twitter and Facebook have a positive impact on AngelList fund-raising, even though these are different social platforms.

Interestingly, for both measures, we observe that the increase in engagement level for successful companies is stronger on Twitter, possibly reflecting the preference of companies to use short tweets to publicize their companies, rather than lengthy Facebook posts.

We next drill down on our analysis results to look at the distribution of engagement levels among companies. Figures 1-3 show three representative CDFs of social engagement metrics for successful (blue) and unsuccessful (red) companies. We make the following observations. First, successful companies are, in aggregate, much more active in social engagement. Based on the delta measures, for companies with social media engagement, very few (if any) are able to raise funding successfully without positive increases in social engagement. The presence of a long tail distribution - as shown by the use of log-scales for tweets - shows that a small number of companies make a significantly large number of tweets as compared to other companies. Second, we further observe that while social activity is a strong indicator, it is not an absolute determiner of success. Indeed, there are many companies that we observe that are active in social media but fail in fund-raising. The other CDFs exhibit similar trends and we omit due to space constraints.

4.2 Other Attributes of Interests

We have shown that social engagement metrics are helpful to distinguish the startups that can succeed in crowdfunding and the unsuccessful ones. However, our dataset also contains other information obtained from AngelList, for example, the size of a company, and the presence/absence of marketing material (e.g. a blog, company description information, etc.) Table 3 shows a summary of all attributes that we have collected for the startups that we tracked in our dataset over the same period.

Name	Description	Pearson Correlation Coefficient
AFollowers	The difference between AFollowers at the end and the beginning of data collection	0.452
FBLikes	The difference between FBLikes at the end and the beginning of data collection	0.0065
FBPosts	The difference between FBPosts at the end and the beginning of data collection	0.057
TFollowers	The difference between TFollowers at the end and the beginning of data collection	0.132
Tweets	The difference between Tweets at the end and the beginning of data collection	0.0215
BlogURL	Binary value indicating whether a company has a blog URL in their profile	0.0088
FBURL	Binary value indicating whether a company has a Facebook URL in the profile	0.0084
LinkedInURL	Binary value indicating whether a company has a LinkedIn URL in the profile	0.0448
Video	Binary value indicating whether a company has a demonstration video	0.0793
SmallSize	Company size (<10)	0.0433
MediumSize	Company size (≥10 and <50)	-0.0194
LargeSize	Company size (≥50)	-0.0125
DescLength	The length of the description text in the profile	0.0356

Table 3: Attributes of interests in analysis.

The first and the second columns in the table describe the attributes. There are thirteen attributes associated to every startup in our dataset. *AFollowers*, *FBLikes*, *FBPosts*, *TFollowers* and *Tweets* are social engagement metrics introduced earlier. *BlogURL*, *FBURL*, *LinkedInURL* and *Video* are related to profile completeness on AngelList. *SmallSize, MediumSize* and *LargeSize* reflects the size of the company. For every company, only one of those three attributes can be true, and the others are false. *DescLength* is the length of description text that a company posted on AngelList. To have a sense of how those attributes are related to the crowdfunding success, we compute the Pearson correlation coefficient for each of the attributes with the successful/unsuccessful label vector. As background, the Pearson correlation coefficient [27] is a value between -1 and 1 that measures the strength of correlation between two variables. 1 and -1 indicate total linear correlations, and 0 means no correlation. The attribute that has the highest coefficient is AFollowers. With the exception of FBLikes, all other social engagement attributes have strong correlation to crowdfunding success. These results are shown in the third column in Table 3.

We restrict our Pearson correlation analysis to startups that have valid values for all attributes. After data cleaning to filter out startups without valid attributes, we are left with 11 successful and 260 unsuccessful companies. For consistency, these 271 startups are also the ones in our prediction analysis in the next section.

Among the profile completeness attributes, Video and LinkedInURL have higher coefficient than other attributes. The coefficients of Small, Medium, and Large imply that the size of company is related to crowdfunding success in the test. Interestingly, MediumSize and LargeSize have negative coefficients, which means that they have negative correlation with crowdfunding success. This reflects the likelihood tendencies of investors on AngelList to bias towards early-stage companies of a smaller size. The last attribute highlights the importance of having a company description, where its coefficient is even higher than Tweets.

Note that Pearson correlation coefficient can only test whether there is linear correlation between each individual attribute and crowdfunding success. In Section 5, we use machine learning techniques to predict crowdfunding success based on a subset of the attributes above. As we will show later, some of the above attributes are good predictors for classifying startups to be successful or not in crowdfunding.

5 PREDICTION

Our previous section establishes the correlation between various metrics (social engagement and others) of startups and crowdfunding success. Our next goal aims to explore whether these metrics can in turn be used to *predict* crowdfunding success. To this end, we adopt supervised learning to classify companies to successful and unsuccessful ones. As we will later show, traditional techniques do not directly apply and need to be modified to take into account the unique (and rare) nature of crowdfunding success.

This section is organized as follows. We first define the problem of predicting crowdfunding success, and we show the characteristics of our dataset, which brings challenges to the prediction problem. We then introduce the techniques that we adopt to handle the challenges, and show how we analyze the importance of each startup attribute in predicting crowdfunding success. Finally, we present the experimental results to illustrate the effectiveness of the attributes we explored and the techniques we adopted in startup crowdfunding success prediction.

5.1 Problem Definition

Given a training set of N startups, each is labeled as either *successful* or *unsuccessful* and has features F set to all attributes listed in Table 3. Our goal is to classify a new example, as *successful* or *unsuccessful*. Since we can easily label a company to be successful or not in crowdfunding (successful if it raised money during the period that we collected the data, and unsuccessful if not), we use supervised learning algorithms in the prediction.

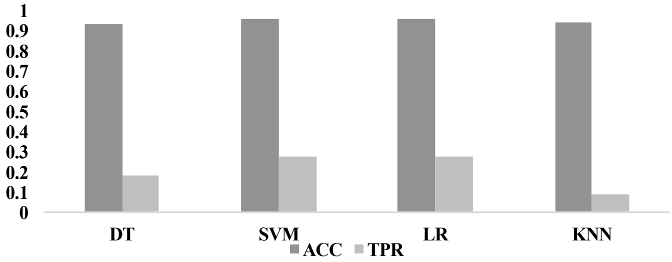

Figure 4: Accuracy and TPR of directly applying standard algorithms to learn from the data.

We explore the following standard classification algorithms on our dataset: decision tree, SVM, logistic regression and k-nearest neighbors (KNN). Each algorithm has unique characteristics and thus determines its own set of features for the best prediction. We run s-fold cross-validation for evaluating prediction performance, where s is the number of successful companies. The data is partitioned into s disjoint folds, and since there are very few successful examples, we let each fold have exactly one successful and multiple unsuccessful examples, and each unsuccessful example is placed in only one fold. One fold is withheld for testing every time (a variant of leave-one-out (LOO) cross validation), and a model is learnt from the remaining data. The results of s folds are averaged as the overall performance, using accuracy metrics that we will introduce in Section 5.2.3.

5.2 Challenges of Learning

The simple definition of this prediction problem does not make it simple to solve. We observe that the numbers of examples in two classes are seriously imbalanced. This is expected, since only a small fraction of startups (11 out of the 271 companies) are successful at fund-raising. The class imbalance requires modifications to standard machine learning algorithms.

5.2.1 *Imbalanced Classes.*

In the imbalanced classes problem, the number of successful examples (*minority*) is orders of magnitude smaller than the number of unsuccessful examples (*majority*). If we directly apply standard machine learning algorithms, the majority class can be severely over-numbered and over-represented. Intuitively, even if all successful examples are classified to be unsuccessful, a machine learning algorithm can still get low error rate. We highlight this problem in Figure 4, where we directly apply the four classification algorithms to learn from our data. ACC and TPR in Figure 4 represent accuracy and true positive rate respectively (as described in Section 5.2.3), and DT, SVM, LR and KNN represent decision tree, SVM, logistic regression and k-nearest neighbors respectively. Note that in KNN, we use cross validation on training data to determine the best k, and in this experiment, $k = 1$ is the output of cross validation. In Figure 4, all four classifiers have more than 90% accuracy, but none of them has TPR higher than 30%. This means the successful examples are rarely predicted correctly in all algorithms. Consequently, solutions are required to address the imbalance, and new metrics are needed to evaluate the performance of predictions.

5.2.2 *Learning from Imbalanced Data.*

The approach that we take is one of learning from imbalanced data [10, 11, 13, 15, 16, 18–20, 22, 24, 29]. Prior work in this area shows that standard classifiers perform better on balanced data than imbalanced data. We adopt three different methods to deal with the imbalance in our case: *over-sampling*, *under-sampling*, and *cost-sensitive learning*. Our focus is on showing that by taking care of the imbalance problem, the attributes that we extracted can be used to accurately predict startup crowdfunding success. While the results gained from our methods are promising, other more advanced imbalanced-class learning techniques can be applied to achieve even better performance as future work. We briefly summarize the three methods used:

- **Synthetic Over-Sampling.** The intuition of synthetic over-sampling is to synthesize more minority examples (*successful* examples) to make the data balanced. Specifically, let S_{suc} be the set of successful examples and S_{uns} be the set of unsuccessful examples, for every example $x_i \in S_{suc}$, it randomly selects another example $x_j \in S_{suc}$, and synthesizes a new example x_{syn} between x_i and x_j by $x_{syn} = x_i + (x_j - x_i) \times \alpha$ where α is a random number between 0 and 1. This is to generate the new example at a random position along the line from x_i to x_j. Let $S'_{suc} = S_{suc} \cup S_{syn}$ where S_{syn} is the set of synthetic successful examples, and let $|S_{uns}|/|S'_{suc}| = \beta$ where β is balancing factor. Therefore, for every example $x_i \in S_{suc}$, we need to synthesize $\lceil |S_{uns}|/(\beta \times |S_{suc}|) - 1 \rceil$ new successful examples. Note that this method is similar to the SMOTE algorithm described in [16]. The difference is that SMOTE finds the k-nearest neighbors for each $x_i \in S_{suc}$, and synthesizes the new example based on x_i and one of it k-nearest neighbors (randomly selected), while our method uses all neighbors of x_i in S_{suc} because there is no noise in our data and each successful startup is representative for generating new examples. In addition, we remove Tomek links [11, 20, 28] for cleaning the synthetic data. Tomek links are overlapping between classes [20], and thus removing such links in $S'_{suc} \cup S_{uns}$ can make the classes well-defined. Specifically, we design an incremental algorithm to find and remove Tomek links, and the processing terminates when no more links can be found. Therefore, the output of synthetic over-sampling is $S'_{suc} \cup S_{uns}$ after data cleaning.
- **Random Under-Sampling.** Random under-sampling removes part of the original majority examples (*unsuccessful* examples) to make the ratio between unsuccessful examples and successful examples balanced. Specifically, we randomly select $S'_{uns} \subset S_{uns}$, and let $|S'_{uns}|/|S_{suc}| = \beta$, where β is balancing factor, which controls the ratio between two classes after sampling. Therefore, the output dataset of under-sampling is $S_{suc} \cup S'_{uns}$. As analyzed in [20], the problem of random under-sampling is that S'_{uns} is only a small part of S_{uns}, hence it may lose important concepts in S_{uns}. To mitigate this problem, we repeat under-sampling experiments for C times, and at every repetition, we randomly choose S'_{uns}, and average the results of C times as overall performance. In this way, it makes the classes balanced and meantime S'_{uns} covers S_{uns} as much as possible.
- **Cost-Sensitive Learning.** In contrast to over-sampling and under-sampling, which address the imbalance problem by making the distribution of successful and unsuccessful examples balanced and are transparent to classifiers, cost-sensitive learning directly modifies the classifiers to let them be aware of the imbalance. Specifically, cost-sensitive learning sets different costs for misclassifications of successful examples and unsuccessful examples. The cost matrix C is defined as

$$C = \begin{bmatrix} 0 & c_1 \\ c_2 & 0 \end{bmatrix}$$

where c_1 is the cost of misclassifying successful examples to be unsuccessful, and c_2 is the cost of misclassifying unsuccessful examples to be successful. Since there is no cost for correct classification, $C[0, 0]$ and $C[1, 1]$ are both zero. The higher the cost is, the higher the penalty will be given to the classification. To deal with the imbalance, we let $c_1 > c_2$, specifically $c_1/c_2 = |S_{uns}|/|S_{suc}|$. After setting the cost matrix in classifiers, their objective is to minimize the total cost of classification. We modify all four classifiers for cost-sensitive learning.

5.2.3 Metrics. We adopt different metrics to effectively measure prediction performance. Specifically, there are four metrics being used in this paper: accuracy, true positive rate (TPR), true negative rate (TNR) and A-mean (AM). Two of these metrics (accuracy and TPR) are first introduced in Section 5.2.1, where we also show the limitations of relying on only one metric in imbalanced. We now describe all the four metrics. First of all, let TP be the number of true positives (successful examples) in the test, TN be the number of true negatives (unsuccessful examples), FP be the number of false positives and FN be the number of false negatives. The four metrics are defined as follows.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$
$$TPR = \frac{TP}{TP + FN},$$
$$TNR = \frac{TN}{TN + FP},$$
$$AM = \frac{TPR + TNR}{2}.$$

Basically, the accuracy metric measures how many examples are classified correctly in the test, the TPR metric measures how many successful examples are classified correctly, and the TNR metric measures the same thing for unsuccessful examples. The AM metric quantifies the trade-off between TPR and TNR, and has been shown in prior work [24] to be a good measure for classification under class imbalance. A classifier that has high TPR (TNR) on our dataset implies that it is effective for representing successful (unsuccessful) startups. All in all, it is difficult to use single metric to effectively assess the prediction results of classifiers, and the accuracy metric alone provides little information. Therefore, we mainly use the TPR, TNR and AM metrics to evaluate prediction performance.

5.3 Feature Selection

We select a subset of features listed in Section 4.2 in order to construct the models used in our classifiers. Intuitively, we aim to limit the features to only the most important ones that can impact the effectiveness of our classifiers. We select features through a greedy algorithm, which is described in Algorithm 1. In the algorithm, A is initialized as all features, while B is initialized as empty set. In every round, the algorithm tests whether there exists a feature in A can improve the prediction result (the algorithm is greedy on AM) based on B through cross validation ($CV(.)$ in line 9 is the process of running cross validation on the given features). If there exist one or more such features, it moves the feature with the best

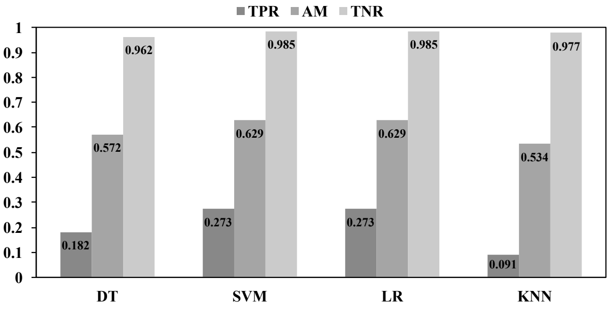

Figure 5: Results of directly applying four classifiers with feature selection.

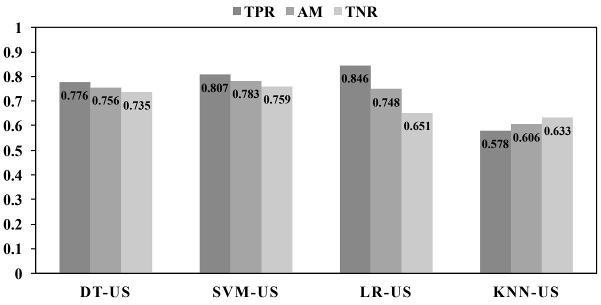

Figure 7: Results of under-sampling with feature selection.

Algorithm 1 Greedy Feature Selection

```
$A \leftarrow \{f_1, f_2, \ldots, f_k\}$
$B \leftarrow \{\}$
$bestAM \leftarrow 0$
while $|A| > 0$ do
    $maxAM \leftarrow 0$
    $maxFeature \leftarrow null$
    for each feature $f_i \in A$ do
        $TmpFeatures \leftarrow B + f_i$
        $AM \leftarrow CV(TmpFeatures)$
        if $AM > maxAM$ then
            $maxAM \leftarrow AM$
            $maxFeature \leftarrow f_i$
        end if
    end for
    if $maxAM > bestAM$ then
        $bestAM \leftarrow maxAM$
        $B \leftarrow B + maxFeature$
        $A \leftarrow A - maxFeature$
    else
        break while
    end if
end while
return $B$
```

cross validation result from A to B. This process continues until no feature in A can improve AM or A becomes empty. The final B is the result of feature selection.

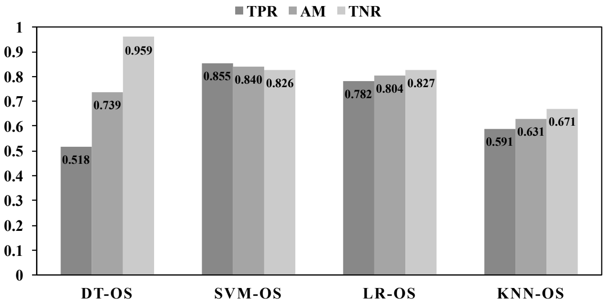

Figure 6: Results of over-sampling with feature selection.

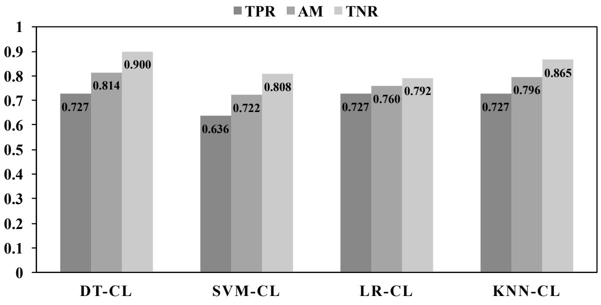

Figure 8: Results of cost-sensitive learning with feature selection.

To enable the feature selection, before we train a classifier, we run Algorithm 1 on training data and use the output features to train model and test on the reserved examples. We analyze the importance of each feature in predicting crowdfunding success based on this feature selection. Every time we do feature selection, we have output $B = \{f'_1, f'_2, \ldots, f'_l\}$, and we count the appearances of each feature in B, and we use the counted value of each feature to analyze its importance. Detailed analysis is present in Section 5.4.

5.4 Prediction Results

We present the results of predicting startup crowdfunding success using the features and the techniques introduced above. First, we compare the overall prediction performance of different classifiers on our dataset, and we show the effectiveness of the solutions to imbalanced classes problem. Then we show our analysis of startup attributes based on feature selection.

Note that in our experiments, we run cross validation on training data to determine the best set of parameters and use those parameters on training data to train classifiers, for example, the number of neighbors in KNN, the depth of decision trees and the best set of features for all classification algorithms.

5.4.1 Classification Analysis. We briefly introduced the problem of directly applying standard classifiers on our dataset in Section 5.2.1. Here we show the problem in detail. In Figure 5, we show TPR, TNR and AM of directly applying decision tree, SVM, logistic regression and KNN to predict crowdfunding success. The classifications of four classifiers are all biased to unsuccessful examples, which leads to extremely high TNR, but extremely low TPR. All TNRs in Figure 5 are higher than 95%, and none of TPRs is higher than 30%, so the AM of the classifications can only be around 50%-60%, not much better than a coin flip. SVM has the same performance with logistic regression, and they are slightly higher than decision tree and KNN. However, *none of them can achieve satisfactory prediction performance when imbalance is present.* In contrast, Figure 6, 7 and 8 show promising prediction results where imbalance is mitigated by over-sampling, under-sampling and cost-sensitive learning.

In Figure 6, all four classifiers significantly improve their prediction performance by using over-sampling. In particular, *SVM plus over-sampling is able to achieve AM as high as 84%*, which is much better than coin flipping and naive application of standard classifiers. The TPR of decision tree increases from 18% to 52%, and TNR is still very high (96%), which together lead to 74% AM. SVM performs the best in this group, with both good TPR (86%) and TNR (83%). Logistic regression achieves 80% AM which is also good, and significantly better than its performance in Figure 5 (around 60%). Like SVM, TPR and TNR of logistic regression are evenly distributed. KNN performs the worst in terms of AM, but its AM still increases by 10% compared to Figure 5.

Figure 7 shows the results of under-sampling, and compared to over-sampling, the ability of under-sampling for dealing with the imbalance problem in our dataset is worse because most classifiers perform not as good as in Figure 6. However, they are all better than Figure 5, which shows that under-sampling works for handling imbalanced data. Specifically, under-sampling is helpful to improve TPR for both decision tree and logistic regression, but not for SVM and KNN, when compared with over-sampling. Decision tree achieves 76% AM and this is better than decision tree plus over-sampling. In addition, decision tree has much better TPR (78%) than both directly applying it to our dataset (with around 60% improvement) and over-sampling (with more than 25% improvement). SVM is still the best classifier in the group with 78% AM, and it is slightly higher than decision tree in every metric. Logistic regression has lower AM than SVM because its TNR is significantly lower, which means it classifies many unsuccessful examples to be successful. Still, KNN is the worst model in under-sampling, so sampling techniques do not work well for KNN.

Cost-sensitive learning provides an alternative to deal with imbalance, and works well for decision tree and KNN as shown in Figure 8. *With cost-sensitive learning, decision tree achieves 81% AM*, which is the best performance in this group and is the second best result between groups, only lower than SVM in over-sampling. Except decision tree, none of other classifiers can achieve AM higher than 80% with cost-sensitive learning, but all increase AM by more than 10% compared to Figure 5. The AMs of decision tree and SVM are both lower than theirs in over-sampling and under-sampling, but the AM of KNN plus cost-sensitive learning is higher than KNN plus both sampling techniques. Logistic regression gets 76% AM, which is the first time better than SVM.

The main takeaway from this analysis is that direct application of standard classifiers can not achieve good predictions on our dataset, but with sampling techniques or cost-sensitive learning, some classifiers can achieve prediction performance higher than 80% in terms of AM. This shows that the startup attributes we present in this paper are good predictors for crowdfunding success.

5.4.2 Feature Analysis. We now analyze how the attributes of interests affect predicting crowdfunding success in details. We first show that feature selection can effectively improve prediction performance in our experiments. Then we use the output of feature selection to analyze the importance of features to the predictions.

In Figure 6, 7 and 8 we show the prediction performance of four classifiers where imbalance is dealt with sampling techniques and cost-sensitive learning. In those experiments, feature selection is enabled, which means before we train a model on training data, we run a cross validation on training data to select features by Algorithm 1. For comparison, we show the predictions with feature selection disabled in Figure 9, 10 and 11. Since there is no big difference between feature selection enabled or not in directly applying standard classifiers (Figure 5), we omit the corresponding figure of Figure 5 with all features. Except KNN with over-sampling and decision tree with cost-sensitive learning (2% more AM with all features in both cases), most prediction results, in terms of AM, become worse when feature selection is disabled. Specifically, in Figure 9 the AM of decision tree with over-sampling drops from 74% to 69%. SVM and logistic regression with over-sampling lose more than 10% and 5% AM respectively when feature selection is disabled. Comparing with Figure 7, Figure 10 shows that all four classifiers have lower AM without feature selection. In particular, the AM of SVM with under-sampling drops more than 7%. Figure 11 shows that feature selection is also beneficial to cost-sensitive learning. The figure shows that the AM of SVM drops from 72% to 70%, and the AM of logistic regression drops from 76% to 74%. KNN has the largest AM decrease. With feature selection, the AM of KNN is almost 80%, but it drops to 66% when feature selection is disabled.

The performance improvement by using feature selection implies that not all attributes in Table 3 are helpful for predicting startup crowdfunding success. The presence of some features when training a classifier can introduce overfitting problem, which causes prediction performance to decrease. This problem can be solved by feature selection. As we showed in above, only KNN with over-sampling and decision tree with cost-sensitive learning have better AMs when Algorithm 1 is disabled. However, even in those two cases, the features selected by Algorithm 1 do not introduce significant AM decrease (2%). Thus, the results show that Algorithm 1 is able to select a set of features for better prediction. Next, we analyze the importance of features in the predictions, i.e. what features are useful for predicting crowdfunding success, and what features may cause over-fitting.

We use *frequency* to measure the importance of each feature in predicting crowdfunding success. Frequency is defined as follows.

$$frequency = \frac{\text{\# of being selected by Algorithm 1}}{\text{\# of running Algorithm 1}},$$

where the numerator represents the times of a feature being selected feature, and the denominator represents the times of running Algorithm 1. The reason for running feature selection for multiple times is that as we mentioned in Section 5.1, we use s-fold cross validation to get the overall performance of a classifier, so in every fold, we need to select features and train model on training data. Also, there is a parameter C in both over-sampling and under-sampling to control the number of rounds. The value of frequency is between 0 and 1. 0 means that feature is never selected for prediction by

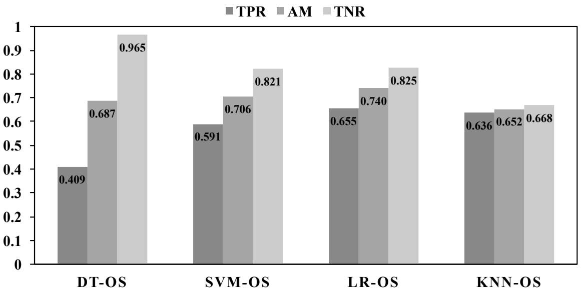

Figure 9: Results of over-sampling with all features.

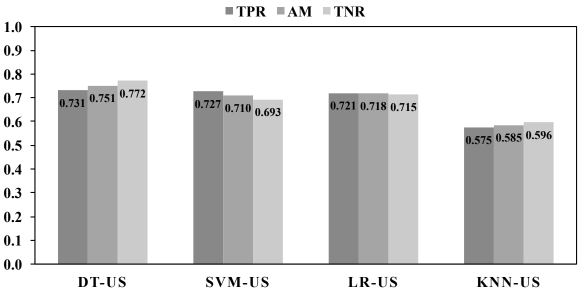

Figure 10: Results of under-sampling with all features.

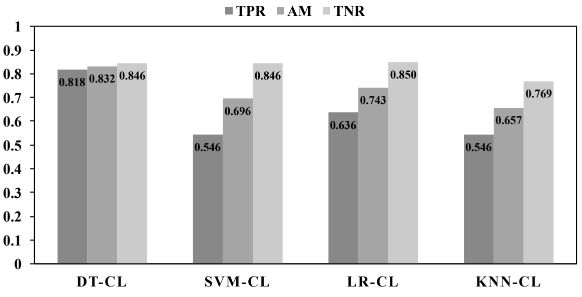

Figure 11: Results of cost-sensitive learning with all features.

Algorithm 1, while 1 means that feature is selected every time in feature selection. Therefore, 1 means "good" feature and 0 means "bad" feature if the prediction itself is a good prediction.

We select top two predictions with feature selection enabled in terms of AM for analysis, which are SVM with over-sampling (AM = 84%) and decision tree with cost-sensitive learning (AM = 81%). The results are shown in Figure 12 and Figure 13. In Figure 12, it shows that *both AFollowers and DescLength have a frequency of 1, which indicates that those two features are the most predictive in this model and are selected every time.* The feature that has the second largest frequency (0.86) is FBPosts, which is selected by this classifier in most time. The frequency of MediumSize is significantly higher than the remaining features, and SmallSize has frequency higher than 0.5, thus the size of company is also predictive in this prediction. TFollowers is also selected in some cases, but FBLikes and Tweets have very low frequencies, and they are actually among the least predictive features in this case.

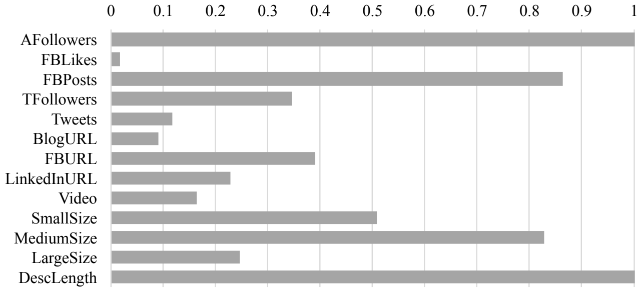

Figure 12: Frequency in SVM with over-sampling.

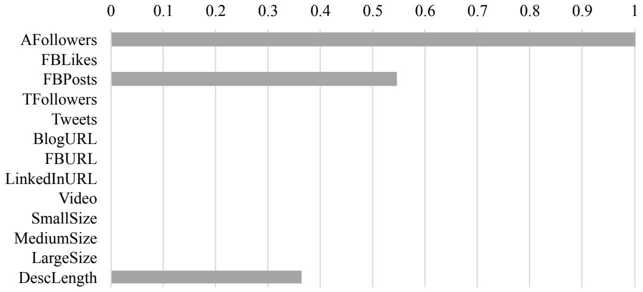

Figure 13: Frequency in DT with cost-sensitive learning.

Figure 13 shows the frequency of each feature in decision tree with cost-sensitive learning. In this prediction, only AFollowers, FBPosts and DescLength are ever selected. AFollowers is selected every time, so it is the most predictive feature in this case. FBPosts has frequency higher than 0.5 and DescLength has frequency as 0.36. The frequencies of all the other features are 0, indicating they are useless in this prediction.

The same observation between Figure 12 and Figure 13 is that AFollowers, FBPosts and DescLength are the three most predictive features, which implies that social engagement (specifically, the number of AngelList followers and the number of Facebook posts) and the description text that startups write for themselves are good predictors for classifying startups to be successful or unsuccessful in crowdfunding.

6 RELATED WORK

Prior studies on crowdfunding have explored investor recommendations [9] based on *Kickstarter* [6], a crowdfunding site for creative projects. [30] applies machine learning and text mining techniques on news article to predict the likelihood a company can be acquired. Our work is significantly different, as we focus on the AngelList platform, which is not only more recent, but also more highly focused on crowdfunding for startup investments, which has different dynamics than crowdfunding for specific projects. [25] does an exploratory study to identify factors for crowdfunding success, but provides only basic analysis on macro-level statistics. Our work is significantly more comprehensive as we integrate data from a range of data sources. Other analysis works on high-tech startups and crowdfunding research [8, 12, 14, 21, 23, 26] do not collect or use time-series data to perform longitudinal analysis, nor do they consider the impact across different social platforms. [17] performs a preliminary measurement study of crowdfunding based on a single snapshot of AngelList, but the study has a number of limitations.

First, the study captures only a snapshot on AngelList and social networks, but lacks the longitudinal aspects of our study to reason about changes over time. Second, the study uses Crunchbase as a basis for fund-raising information. Crunchbase includes information from non-crowdfunding sources, and is thus often not an accurate reflection of the progress of companies fund-raising on crowdfunding. Our study, on the other hand, tracks the actual fund-raising progress of companies on AngelList, and hence represents the ground truth on actual crowdfunding success.

The biggest challenge of learning from our dataset is imbalance between classes. We adopt three methods to deal with this problem. In fact, there is a number of works studying imbalanced learning, and those techniques can be applied to our problem to get satisfactory prediction results. *Sampling techniques.* [16] shows the imbalance problem in real world data sets, and proposes Synthetic Minority Over-sampling Technique (SMOTE) and under-sampling technique to deal with imbalance in classification. There have been a number of techniques developed based on SMOTE. In [13], authors propose safe level when doing sampling in SMOTE, and synthesize minority examples around larger safe level, and show higher accuracy than SMOTE. [10] presents Majority Weighted Minority Oversampling TEchnique (MWMOTE) to solve imbalanced learning. MWMOTE first computes a weight for each minority example, and then synthesizes new examples based on the weights. *Cost-sensitive learning techniques.* [18] reviews the problem of using different costs for misclassifications in different classes, and suggests several methods for better cost-sensitive learning based on a cost matrix. [15] explores the imbalance problem in multilayer perceptron (MLP) neural networks, and proposes a cost-sensitive learning based algorithm, called CSMLP, to improve two-class predictions in MLPs, which is based on a joint objective function and uses a cost parameter to differentiate misclassifications in different classes. In addition, [29] proposes multiset feature learning (MFL) for highly imbalanced learning. [20, 22] summarize and analyze the techniques for imbalanced learning.

7 CONCLUSION

In this paper, we have performed a longitudinal data collection and analysis of social media data, to improve our understanding of the new crowdfunding phenomenon among high-technology startups. Through a combination of correlation analysis and machine learning techniques, we have shown the strong relationships between social engagement and startup fund-raising success, and also identified the startup attributes that are most predictive.

To the best of our knowledge, this is the first longitudinal data collection of this nature. We have also shown that machine learning techniques in classification given imbalanced datasets is effective in making predictions on successful fund-raising. This has allowed us to shed light into the effectiveness of social engagements on fund-raising success. Within our current dataset, we also plan to further analyze our data by, for example, examining the contents of posts and tweets; as well as analyzing profiles of founder and investors to explore the existence of other factors involved that impact a company's crowdfunding success.

ACKNOWLEDGMENTS. We thank Miao Cheng, Sudarshan Muralidhar, Anand Sriramulu and Stuart Wagner for their earlier efforts in data collection. This work is funded in part by NSF CNS-1513679 and NSF CNS-1218066.

REFERENCES

[1] Angellist. https://angel.co/.
[2] Angellist wikipedia. https://en.wikipedia.org/wiki/AngelList.
[3] Equitynet. https://www.equitynet.com/.
[4] Fundable. https://www.fundable.com/.
[5] Graph api. https://developers.facebook.com/docs/graph-api.
[6] Kickstarter. https://www.kickstarter.com/.
[7] Tweepy. http://www.tweepy.org/.
[8] Agrawal, A. K., Catalini, C., and Goldfarb, A. Some simple economics of crowdfunding. Working Paper 19133, National Bureau of Economic Research, June 2013.
[9] An, J., Quercia, D., and Crowcroft, J. Recommending investors for crowdfunding projects. In *WWW* (2014), ACM, pp. 261–270.
[10] Barua, S., Islam, M. M., Yao, X., and Murase, K. Mwmote-majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Trans. Knowl. Data Eng. 26*, 2 (2014), 405–425.
[11] Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations 6*, 1 (2004), 20–29.
[12] Belleflamme, P., Lambert, T., and Schwienbacher, A. Crowdfunding: Tapping the right crowd. *Journal of Business Venturing 29*, 5 (2014), 585–609.
[13] Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap, C. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Advances in Knowledge Discovery and Data Mining, 13th Pacific-Asia Conference, PAKDD 2009, Bangkok, Thailand, April 27-30, 2009, Proceedings* (2009), pp. 475–482.
[14] Burtch, G., Ghose, A., and Wattal, S. An empirical examination of the antecedents and consequences of contribution patterns in crowd-funded markets. *Information Systems Research 24*, 3 (2013), 499–519.
[15] Castro, C. L., and de Pádua Braga, A. Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. *IEEE Trans. Neural Netw. Learning Syst. 24*, 6 (2013), 888–899.
[16] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR) 16* (2002), 321–357.
[17] Cheng, M., Sriramulu, A., Muralidhar, S., Loo, B. T., Huang, L., and Loh, P.-L. Collection, exploration and analysis of crowdfunding social networks. In *Proceedings of the Third International Workshop on Exploratory Search in Databases and the Web* (2016).
[18] Elkan, C. The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001* (2001), pp. 973–978.
[19] Estabrooks, A., Jo, T., and Japkowicz, N. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence 20*, 1 (2004), 18–36.
[20] He, H., and Garcia, E. A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng. 21*, 9 (2009), 1263–1284.
[21] Juan B. Roure, M. A. M. Linking prefunding factors and high-technology venture success: An exploratory study. *Journal of business venturing 1*, 3 (1986), 295–306.
[22] Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Progress in AI 5*, 4 (2016), 221–232.
[23] Kuppuswamy, V., and Bayus, B. L. Crowdfunding creative ideas: The dynamics of project backers in kickstarter. *UNC Kenan-Flagler Research Paper*, 2013-15 (2015).
[24] Menon, A. K., Narasimhan, H., Agarwal, S., and Chawla, S. On the statistical consistency of algorithms for binary classification under class imbalance. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013* (2013), pp. 603–611.
[25] Mollick, E. The dynamics of crowdfunding: An exploratory study. *Journal of business venturing 29*, 1 (2014), 1–16.
[26] Schwienbacher, A., and Larralde, B. Crowdfunding of small entrepreneurial ventures. *Handbook of entrepreneurial finance, Oxford University Press* (2010).
[27] Stigler, S. M. Francis galton's account of the invention of correlation. *Statistical Science 4*, 2 (1989), 73–79.
[28] Tomek, I. Two modifications of cnn. *IEEE Transactions on Systems, Man and Cybernetics SMC-6*, 11 (1976), 769–772.
[29] Wu, F., Jing, X., Shan, S., Zuo, W., and Yang, J. Multiset feature learning for highly imbalanced data classification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.* (2017), pp. 1583–1589.
[30] Xiang, G., Zheng, Z., Wen, M., Hong, J. I., Rosé, C. P., and Liu, C. A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on techcrunch. In *ICWSM* (2012).