Quantifying Learning and Competition among Crowdfunding Projects: Metrics and a Predictive Model

Maryam Rahmani Moghaddam*
W.P. Carey School of Business
Arizona State University
Tempe, AZ
maryam-rahmanimoghaddam@uiowa.edu

Xiexin Liu*
Department of Data Analytics,
Dickinson College
Carlisle, PA
liuxi@dickinson.edu

Weiguo Fan
Department of Business Analytics,
University of Iowa
Iowa City, IA
weiguo-fan@uiowa.edu

Abstract

The performance of a crowdfunding project is highly situational-dependent. In this study, we quantify the interactions between crowdfunding projects in order to understand how these interactions can help predict the performance of crowdfunding campaigns. Specifically, we utilize Natural Language Processing (NLP) techniques to create a semi-automated system to label the associated product for each crowdfunding campaign. We also propose three sets of metrics to measure how crowdfunding projects learn from and compete with each other. Finally, we propose a machine learning model and demonstrate that the proposed metrics and the proposed model outperform other combinations when predicting the performance of crowdfunding projects.

Keywords: reward-based crowdfunding, interpretable machine learning, predictive analysis

1. Introduction

While entrepreneurs have a variety of sources to raise capital such as family and friends, banks, angel investors, and venture capital (VC) firms, it is significantly more difficult for newly born startups to secure external funding (Cosh et al. 2009). The reason is startups barely have tangible assets or reasonable profit margins, whereas venture capitals mostly prefer to engage and invest in mid- to late-stage larger businesses ("NVCA Yearbook - National Venture Capital Association - NVCA" 2020). Therefore, crowdfunding, as a new modern phenomenon, has become a reliable alternative source of funds for founders to raise capital from the crowd and, most probably, potential customers (Agrawal et al. 2014).

According to Kickstarter, the world's largest crowdfunding platform, as of November 22nd, 2020, *the platform had hosted over half a million campaigns and attracted a total of $5.44 billion in funds. However, 90% of pledged funds went to only 38% of the total campaigns. Although there is a large crowd of backers on crowdfunding platforms, the success rate of campaigns indicates that a large amount of available funds does not guarantee campaign success. Scholars are thus interested in studying the crowdfunding phenomenon and the factors leading to a campaign's success.

Among all the factors, a significant amount of research has been investigating the mechanism behind how campaigns interact as innovative entrepreneurs propose their novel products and enjoy worldwide crowdfunds on this platform. For example, previous overwhelmingly successful campaigns could interact with concurrent or future campaigns by bringing concurrent and lasting positive effects to both inside- and outside-category campaigns (Liu et al. 2015). Further, with hundreds of live campaigns at a time, crowdfunding campaigns would inevitably interact with each other through their competition for backers.

The task of quantifying these different interactions among crowdfunding campaigns, however, is not simple from various perspectives. First, current research on firm competition tends to examine competitors within the same industry, yet there is ambiguity regarding the definition of "industry" on crowdfunding platforms like Kickstarter. On Kickstarter, projects are divided into different categories, and each category may include projects from a variety of different industries. As an example,

* These two authors contributed equally.

URI: https://hdl.handle.net/10125/103064
978-0-9981331-6-4
(CC BY-NC-ND 4.0)

bikes and furniture are often deemed as products from different industries, however, they would both be counted into the Design category on Kickstarter. Thus, in order to analyze campaign interactions, it is necessary to build an automated system that can identify products from the same industry. Additionally, while there has been extensive research measuring how firms benefit from unexploited knowledge developed in other firms, but such conclusions cannot be drawn from crowdfunding campaigns since most of the campaigns do not possess patents.

In response to these research gaps, this study makes the following contributions. First, this study uses tools in Natural Language Processing (NLP) to build a semi-automated system to identify products within the same industry. Second, we propose that campaigns interact with each other through learning and competing, and such interactions could help us predict the performance of each crowdfunding campaign. To this end, this study focuses on quantifying crowdfunding campaign interactions and further uses machine learning methods to show that the proposed metrics help predict a campaign's performance. Third, this study also proposes a novel machine learning method to further boost the prediction performance. We show our results with a unique and large data set consisting of 27,196 Kickstarter product design campaigns, ranging from the inception of Kickstarter in April 2009 till April 2020. Based on our analysis, we found that, the herd effect is not beneficial for entrepreneurs in a crowdfunding environment since the performance of recently launched and/or concurrent campaigns carrying similar products is not a useful indicator of a campaign's success. In addition, we found that overwhelmingly successful campaigns tend to have a lasting effect on campaign performance prediction, rather than an immediate effect.

The rest of this paper is organized as follows: In the next section, we summarize the related work. Section 3 summarizes the dataset as well as the semi-automated system used to identify products from the same industry. Section 4 outlines the proposed metrics, and Section 5 outlines the proposed model and also summarizes the improvements in predicting performance given by the combination of the proposed metrics and the proposed model. Section 6 concludes this paper.

2. Literature Review

This paper focuses on how campaign interactions, i.e., campaigns learning and competing, helps predict campaign performance. From the learning perspective, we argue that entrepreneurs can *learn from other* projects, as it is widely believed that crowdfunding investors can reveal accurate information regarding the market demand (Liu and Wang 2018). Demand uncertainty is a critical issue that firms encounter mainly regarding their new products. Unless the true demand is known, these goods may be under- or over-supplied (Kennett 2008). Therefore, having knowledge of the product demand before mass production has subnational financial benefits for the creators of the product. However, gaining knowledge about the existing demand before finalizing the design and product can assist entrepreneurs in saving time and money by guiding them to design accordingly. This is made possible through crowdfunding platforms, as the historical and recent data on the amount of support gained by different types of products are publicly available.

However, current research has focused on how each entrepreneur *learns from their own* project. Mollick and Kappuswamy (2014), for example, has shown in their survey that most crowdfunding creators (59%) use a crowdfunding platform as the first step in starting their businesses, and one of the main reasons is to see if their *own* projects are in demand. Da Cruz (2018) further validates the results of Mollick and Kuppuswamy (2014) survey through empirical analysis of reward-based crowdfunding data. The authors argue that even failed projects can benefit from reward-based crowdfunding through the crowd's contributions to learn about the market demand. Strausz (2017) further shows that even marginally failed crowdfunding campaigns can use the amount of aggregated funds to alleviate the demand uncertainty concerns of VCs to raise capital. Chemla and Tinn (2020) show that through crowdfunding platforms, firms can learn about both demand and customer preferences in the early stages of developing their products. They argue that demand learning is the outcome of observing customer investments early during the campaign. Ellman and Hurkens (2019) describe backers as buyers with different valuations and study crowdfunding from the perspective of demand revelation and price discrimination which helps entrepreneurs adapting production to actual demand.

Further, we argue that the intensity of competition between campaigns that sell similar products differs from the campaigns that sell different types of products. However, current literature on campaign competition focuses on the mechanism, dynamics, efficiency, and risk, instead of differing competition from supplements and competition from complements. For example, Janku and Kucerova (2018) defined competitors as the campaigns launched in the same month or federal state while ignoring the category of

the campaign. They focused on geographical and temporal competition features, i.e., campaign's launching location, whether launched on weekends or not. Gallemore et al. (2019) investigate the impact of geography on Indiegogo campaigns and show that competition between campaigns of the same category is relatively local. Lin et al. (2018) further proposed a Dynamic Competition Model (DCM) to predict the daily collected funds of crowdfunding projects by capturing the competitiveness of projects. Nevertheless, as discussed in Lin et al. (2018), there tends to be a stronger competition among projects with similar products. While they used K-means to cluster the projects such that the ones with similar products are in the same group, the process is data-driven and thus the resulted clusters do not purely represent the specific product types.

3. The Data

The dataset contains all the Product Design campaigns from Kickstarter. The campaigns in the Product Design category, despite covering a wide range of market needs including electronics, home furniture, camping, and so forth, are mostly analogous to manufacturing goods. Further, a reward-based scheme is widely used in this category as incentives for more funding.

We started with 27,196 Kickstarter campaigns. After removing non-English and private campaigns as well as the ones we were not able to retrieve their text (no text or all images), we ended up with 26,874 campaigns. We use campaigns' text to cluster them into distinct product types such as watch, guitar, bike, etc. We exclude 2,861 canceled and 8,355 non-US-based campaigns from our regression analysis. Although we selected campaigns in the US, few of them raised funds in foreign currencies which are converted to the US dollar[1]. Subsequently, all dollar values are inflated to 2020 USD.

As we have discussed previously, crowdfunding platforms such as Kickstarter divide projects into categories. However, the definition of *category* on Kickstarter is drastically different from the definition of *industry* in any research on firm competitions. For example, the electronics, home furniture, and camping necessities are all projects in the same Product Design category but are indeed products from a wide range of industries. Further, such specific product information is not readily accessible and may only be inferred from the campaign title and/or the campaign text.

Therefore, we first leverage text analysis and clustering techniques to cluster product types based on campaigns' descriptions. Specifically, we converted each campaign's text to a set of words and built a TF-IDF matrix. A TF-IDF matrix calculates the importance of each word to a document in a corpus (Salton and McGill 1986). However, due to the size of our dataset and the uniqueness of each campaign, the resulted TF-IDF matrix is sparse and high dimensional. To avoid the curse of dimensionality, we further apply the singular value decomposition (SVD) method on our dataset so we can produce a low-rank approximation of the TF-IDF matrix (Abidin et al. 2010).

We then apply the X-means clustering method on the resulted TF-IDF matrix approximation. Although the K-means clustering method is one of the most efficient clustering techniques, it requires an a priori number of clusters. This number is difficult to determine because of the size of our dataset, and we thus use the X-means clustering instead. The X-means clustering method is the extension of the K-means clustering method in which the number of clusters is determined by the algorithm itself (Pelleg et al. 2000).

To obtain more specific product types, after we obtain clusters from the X-means clustering, we manually check each cluster. We separate the campaigns that failed to form into a meaningful cluster into a different matrix. We apply the whole procedure on this matrix until all clusters are specific enough. We ended up with 113 clusters where each one of the clusters is a distinct product type. One of these 113 clusters is 'No Type. 'No Type" contains the campaigns that did not fit a specific product type. To validate our results, we randomly selected 1000 campaigns to compare the real product type with the clustering label and obtained an 82.6% accuracy.

As a result of our clustering procedure, the products that were initially grouped under the "product design" category on Kickstarter are now labeled according to their specific type, such as watches, bicycles, etc. We categorize products based on the type of product being advertised in the campaign and therefore we do not differentiate, for example, between diving watches and luxury watches. Although diving watches and luxury watches may appear to be designed for different occasions, we contend that since they are both designed to display time, they should still be considered as one type of product.

[1] This campaign's location is Brooklyn, New York but raised funds in Japanese yen.

4. Metrics

We propose seven metrics from three different aspects in this paper. We first argue that crowdfunding campaigns interact with each other by 1) learning about *market demand, product ideas,* and *research & development* from concurrent and previous campaigns, and 2) competing on an intra- and inter-category level. We further argue that campaigns are inevitably affected by the previous overwhelmingly successful campaigns, both on an intra- and inter-category basis. When defining these metrics, the thresholds are selected based on our pilot studies to ensure that they had significant impacts on the results.

4.1. Learning metrics

We use *demand revelation* to measure the impact of previous campaigns revealing market demand. Demand revelation, also known as preference revelation in the field of economics, happens when people reveal their true preferences for public goods (McMillan 1979; Mueller 2008). Failing to know the true demand results in under- or over-supplying these goods (Kennett 2008). While the demand revelation literature in economics mainly focused on public goods to assist policy makers in providing the correct amount of public goods, in the crowdfunding (mainly a private good) context, failing to know the demand for an innovative product can also lead to project failure and result in a waste of time and money. Crowdfunding platforms are a way through which people can reveal their needs and preferences for presented products. In fact, one of the reasons that entrepreneurs use crowdfunding is to learn about the demand (Mollick and Kuppuswamy 2014).

We further use *ideation* to measure the effect of the quality ideas proposed by previous campaigns. Crowdfunding platforms facilitate ideation as they are a rich source of quality ideas. Quality of idea is one of the early determinants of product success (Goldenberg et al. 2001). While the consensus is that new ideas are encouraged and expected by the market, Calantone et al. (2006) show that the familiarity of customers with a product also plays a major role in its success. Therefore, we define a quality idea as one with a balance between newness and customer familiarity. Crowdfunding platforms, on the other hand, are full of such quality ideas as the crowds on crowdfunding platforms can identify the same signals of quality as the experts (Mollick and Robb 2016).

We define our third learning metric, *research & development sharing*, based on the knowledge spillover literature. Knowledge (R&D) spillover happens when firms benefit from the unexploited knowledge developed in other firms (Acs et al. 2009). Knowledge spillover has received a lot of attention in the entrepreneurship context (Acs et al. 2009, 2013; Jaffe 1986; Jones and Ratten 2020), and is recently discussed in crowdfunding (Johan and Taylor 2019; Martínez-Climent et al. 2020). Johan and Taylor (2019), for example, suggests that while knowledge was easily contained before internet, crowdfunding platforms provide the opportunities for campaigns to benefit from the degree of localized knowledge. Specifically, they found that art projects originate in counties with a higher proportion of creative jobs tend to be more successful in their fundraising activities.

To this end, we define *Demand revelation, Ideation* and *R&D sharing* as our learning metrics:

- *Demand revelation*. Entrepreneurs tend to gauge the demand size of a product before they decide to enter the market. Such demand information is readily available for online crowdfunding as entrepreneurs can keep track of the performance for campaigns in the same product category. We define the demand revelation by the ratio of funds obtained by the focal campaign's cluster mates[2] over the last 365 days before the launch date of the corresponding campaign.
- *Ideation*. Ideation is a proxy for the degree to which customers are familiar with the new product. This is called customer familiarity in the literature (Calantone et al. 2006). We define ideation by the cosine similarity between the focal campaign's text and its highly successful (more than 200% funded) as well as recently launched (launched within 365 days) cluster mates' text.
- *R&D sharing*. We define that a campaign has more opportunities to benefit from previous research & development if there have been more successful campaigns for this type of product. Therefore, we define the amount of received R&D knowledge sharing as the success rate of the focal campaign's cluster mates over the past 365 days before the launch date of the focal campaign.

4.2. Competition metrics

Besides learning, crowdfunding campaigns also affect each other through competition. While traditionally firms often only compete with existing firms or potential entrants in the same industry (Porter

[2] Cluster mates refer to campaigns launched in the same product type cluster.

2008), the categories defined by crowdfunding platforms can hardly be deemed as an industry as it usually contains a broad range of products, e.g., the Design category on Kickstarter would include bike, home furniture, and many other designs.

Nevertheless, the crowdfunding literature tends to define competition based on the categories provided by the platforms without further distinguishing the industry or product types. For example, Thies et al. (2016) define category competition as the number of backers a campaign gets compared to the number of existing campaigns in the same category. Similarly, Soublière and Gehman (2020) measure competition by the number of concurrent campaigns in the same category. In this research, we extend the literature by leveraging clustering techniques to get precise product type for products under the same Kickstarter category. With clustering, campaigns in the same cluster become campaigns of the same product type. We can thus consider competitions from products of the same type, as well as products in the same category but of a different type. Specifically, we define *outside-cluster competition* and *inside-cluster competition as our competition* metrics:

- *Outside-cluster competition.* We define the outside-cluster competition as the number of campaigns outside the focal campaign's cluster that end between the start and end date of the focal campaign.
- *Inside-cluster competition*. Similarly, we define the inside-cluster competition as the number of focal campaign's cluster mates whose end dates are between the start and end date of the focal campaign.

4.3. Blockbusters

The overwhelmingly successful campaigns, blockbusters, also have gained the attention of scholars. Blockbusters can exert both positive and negative effects on other campaigns (Liu et al. 2015). The positive effect of blockbusters on crowdfunding platforms is mostly due to attracting new backers as well as engaging existing ones which leads to their contribution to the fundraising activities of *similar* current and upcoming campaigns. It is also expected that there exists a negative effect for campaigns in *other* categories, since higher engagement of the backers in a specific product category could result in fewer backers in other categories (Liu et al. 2015). Moreover, the authors showed that blockbusters bring concurrent and lasting positive effects to both inside- and outside-category campaigns. We adopt a similar approach in studying blockbusters. Specifically, we define *outside-cluster blockbusters* and *inside-cluster blockbusters*:

1) *Outside-cluster blockbusters*. Researchers have found that blockbusters, i.e., the extremely successful campaigns, could direct backers' support (Liu et al. 2015). We consider the campaigns that are more than 300% funded as blockbusters. Outside-cluster blockbusters are the number of blockbusters, launched within one month before the starts date of the focal campaign, but are of different product types comparing to the focal campaign.

2) *Inside-cluster blockbusters*. In addition to outside-cluster blockbuster, we also consider the number of concurrent blockbusters that are of the same product type as the focal campaign. Similarly, we only consider the campaigns that are launched within one month before that start date of the focal campaign.

5. Predictive Analysis

We perform a predictive analysis on campaign performance based on the metrics proposed in Section 4. We define the outcome variable for each campaign as the achieved percentage above (or below) the goal. Section 5.2 outlines the three sets of basic campaign features that often used in the literature for predicting campaign performance, and we also include them as control metrics in this paper. Section 5.3 proposes a novel machine learning model to further help boost our prediction performance, and results in Section 5.4 show that the combination of the proposed metrics and the proposed model outperforms other traditional metrics and models.

5.1. Outcome variable

We define the response variable y as the achieved percentage above (or below) the goal. Specifically, we define y as:

$$y_i = \frac{Money\ Pledged_i - goal_i}{goal_i}, \qquad (5.1)$$

so that $y_i \geq 0$ if campaign i pledged more than or equal to its goal, and $y_i < 0$ if the campaign pledged less than its goal. In our dataset, 47.1% of the campaigns successfully pledged more than their goals, and less than 1% of the campaigns pledged more than 50 times of what they asked for. Figure 1 shows the distribution of the outcome variable for the campaigns that pledged less than 50 times of their goal.

5.2. Control metrics

As our goal is to predict for the performance of crowdfunding campaigns, we also include three set of basic campaign features in our predictive analysis:

rewards for backers, quality of descriptions, and usage of other visualization tools. These three sets of features are expected to interfere with backers' decisions and are typically considered in the crowdfunding literature.

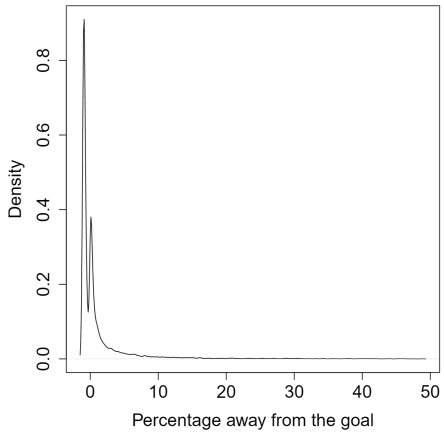

Figure 1: Distribution of the outcome variable for the campaigns that pledged less than 50 times of their goal.

Rewards for backers. Crowdfunding campaigns often offer a variety of rewards to attract investors, for example, utilitarian rewards, socioemotional rewards, and participatory rewards (Jiang et al. 2015). We thus include two rewards-related variables, namely the *length of rewards description* and the *number of reward tiers* offered.

Quality of descriptions. Further, we assess the quality of descriptions on a crowdfunding campaign page in terms of the length of different sections and the ease of readability. In particular, we include the length of the campaign, the blurb section, the story of the campaign, and the risk disclosure. The ease of readability is measured through Flesch Reading-Ease test (Flesch 1979). In this test the higher scores indicate that the text is easier to read.

Usage of visualization. Aside from text descriptions, another way to attract backers' attention is by using other visualization tools, including image, gif, and videos. We thus also measure if a campaign description includes a pitch video, the number of images used, as well as the number of animations (gif, and video) used.

5.3. The proposed model

Varying Coefficient Model (VCM) is a semi-parametric statistical model widely used in social science, business, biology, and other fields. It is a natural alternative to linear models when coefficients are expected to change across different groups. Let $X \in R^{N*P}$ be the covariate matrix with $n = 1, \dots, N$ sample units and $j = 1, \dots, P$ variables. Further, let y be the response. A varying coefficient model assumes that the impact of some variables $x_j \in X, j \in \{1, \dots, P\}$ change over different values of a group variable z. A varying coefficient model takes the following form

$$y_i = \sum_{j \in P} f(z_i) X_{ij} + \epsilon_i, \tag{5.2}$$

where for sample unit i, y_i is the response, z_i is the group information, $X_i = \{X_{i1}, X_{i2}, \dots, X_{iP}\}$ is the collection of the P variables that are believed to change over group z_i, and ϵ_i is the residual for sample unit i that can not be explained through the model. For example, Wang and Hastie (2014) showed that the impact of price of mobile computers on the number of units sold changes over different sales channels. In this case, X_{ij} is the price for mobile computer i, y is the units sold, and z is the sales channel of this computer i.

We propose a model that extends the Support Vector Regression (SVR) model by allowing its coefficients to change. Chen et al. (2014) and Lu et al. (2018) has shown the efficiency of local support vector machines in terms of classification, and the model proposed in this research further combines the strengths of SVR's high accuracy and VCM's high interpretability. Let $X \in R^{N*P}$ be the covariate matrix with $n = 1, \dots, N$ sample units and $j = 1, \dots, P$ variables. Further, let y be the response. A varying coefficient model assumes that the impact of some variables $x_j \in X, j \in \{1, \dots, P\}$ change over different values of a group variable z. To construct a hyperplane, the proposed model first assumes the following linear function:

$$\begin{aligned} y_i &= \sum_{d \in P_C} \beta_d X_{id} \\ &+ \sum_{j \in P_V} f(z_i) X_{ij,} + \beta_0, \end{aligned} \tag{5.3}$$

where P_C is the collection of variables that have constant impact of the years, P_V is the collection of variables that have varying impact of the years, and β_0 is the intercept. The decision boundary could be further defined as

$$\begin{aligned} \sum_{d \in P_C} \beta_d X_{id} + \sum_{j \in P_V} f(z_i) X_{ij} + \beta_0 \\ \leq \epsilon, \\ \sum_{d \in P_C} \beta_d X_{id} + \sum_{j \in P_V} f(z_i) X_{ij} + \beta_0 \\ \geq -\epsilon, \end{aligned} \tag{5.4}$$

where ϵ is the user-defined tolerance level. The two decision boundaries thus form the hyperplane for a Support Vector Regression model with the objective function

$$\min_{\beta_0,\beta_{P_C},\beta_{P_V}} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}|\xi_i|,$$
s.t.
$$\sum_{d\in P_C}\beta_d X_{id} + \sum_{j\in P_V} f(z_i)X_{ij} + \beta_0 \le \epsilon + \xi_i,$$
$$\sum_{d\in P_C}\beta_d X_{id} + \sum_{j\in P_V} f(z_i)X_{ij} + \beta_0 \ge -\epsilon - \xi_{i,} \quad (5.5)$$

where ξ_i represents the deviation for sample unit i from the hyperplane, and the objective function is to minimize the total deviation for all sample units.

We use the kernel technique to modify the objective function in order to determine the coefficients that change across the various groups z_i:

$$\min_{\beta_0,\beta_{P_C},\beta_{P_V}} \frac{1}{2}\sum_{d\in P_C}\|\beta_d\|^2 + \frac{1}{2}\sum_{j\in P_V}\|\beta_j\|^2 + C\sum_{i=1}^{N}K_h(z_i - z_0)|\xi_i|,$$
s.t.,
$$\sum_{d\in P_C}\beta_d X_{id} + \sum_{j\in P_V} f(z_i)X_{ij} + \beta_0 \le \epsilon + \xi_i,$$
$$\sum_{d\in P_C}\beta_d X_{id} + \sum_{j\in P_V} f(z_i)X_{ij} + \beta_0 \ge -\epsilon - \xi_i, \quad (5.6)$$

where h is the kernel bandwidth. In other words, for each z_0, where $z \in [0, T]$, we pool information from all the sample units around z_0 and weigh the i^{th} subject by a local smoothing kernel $K_h(z_i - z_0)$. Therefore, for each of the group z_0, we use gradient descent to approximate the corresponding coefficient $f(z_0)$.

5.4. Predictive performance

We use year 2010 to 2015 for training and compare the prediction performance on the consecutive years by using the aforementioned sets of metrics, including campaign features only (CF), learning metrics (LM), competition metrics (CM), and blockbusters (B). We consider a wide range of machine learning models, including XG-Boost, decision trees (DT), linear regression (LR), support vector machines (SVM) with a nonlinear kernel, as well as the proposed VCSVM model. Table 1 summarizes the Mean Absolute Error (MAE) for predicting one year in advance, three years in advance, as well as predicting five years in advance.

Model performance. Table 1 indicates that nonlinear machine learning models perform better than linear models in terms of prediction performance. Furthermore, the proposed model consistently outperforms all other machine learning models by an average of 10%. VCSVM, as opposed to benchmark models, assumes that the impact of metrics on a campaign's performance changes each year. The outperformance of the proposed model validates this assumption. We presume that this fluctuation in the effects of metrics through time is possibly due to circumstantial factors that are not included in this dataset. Moreover, given that VCSVM models this impact change as a time-dependent relationship $f(z)$, the outperformance of VCSVM may indicate that the relationship between any metric and a campaign's performance at any given time has a lasting effect on future campaigns. In other words, the impact of any metric at time t is dependent on its impact before time t, which leads to the conclusion that this impact is propagated through time.

Metric performance. In addition, Table 1 shows that using the proposed metrics enhances prediction performance in comparison with using only campaign features. Among all the proposed metrics, competition metrics help the most in predicting campaign performance. This suggests that the number of concurrent campaigns of the same product, as well as the number of concurrent campaigns of a different product but are in the same category on Kickstarter, often helps entrepreneurs predict the performance of their own product. Learning metrics contribute the least in terms of prediction of all the metrics proposed. In other words, based on our dataset, there is no strong evidence to suggest that a recently launched successful product would be expected to benefit the performance of similar products; similarly, a recently failed product would also not prevent similar products from being successful. This suggests that initiating campaigns for a product in high need does *not* necessarily guarantee its campaign's performance as 1) there are a lot of competitors out there, and 2) backers' preferences might change and there exists a highly dynamic innovative environment in the entrepreneurship world. Moreover, our results indicate that the impact of blockbusters on campaign performance is trivial in early years (e.g., a one-year prediction). However, this impact grows over time. Incorporating the number of blockbusters improves predictive performance when we are predicting years in advance (e.g., a three-year or five-year prediction). This indicates that the

relationship between the number of blockbusters and campaign performance at any given time does not have an immediate effect on future products, but the effect is likely to persist for a considerable period of time.

Table 1: Comparison of Mean Absolute Error (MAE) by using different models and different sets of metrics.

Pred range	Metrics	LR	DT	XG-B	SVM	VC-SVM
1-y	CF	3.65	3.66	3.20	3.34	**3.17**
	CF +LM	3.67	3.66	3.22	3.34	**3.17**
	CF +LM +CM	3.61	3.43	3.18	3.30	**2.95**
	CF +LM +CM +B	3.56	3.43	3.35	3.34	**2.98**
3-y	CF	4.08	3.92	4.08	3.78	**3.59**
	CF +LM	4.1	3.92	4.1	3.77	**3.59**
	CF +LM +CM	4.04	3.86	3.91	3.74	**3.61**
	CF +LM +CM +B	3.97	3.86	4.03	3.78	**3.56**
5-y	CF	4.73	4.62	4.78	4.44	**4.41**
	CF +LM	4.75	4.62	4.81	4.43	**4.41**
	CF +LM +CM	4.71	4.52	4.62	4.4	**4.32**
	CF +LM +CM +B	4.63	4.52	4.74	4.45	**4.27**

Feature importance. Further, Figure 2 shows the estimated feature importance of the proposed metrics for the years 2010 and 2015, which is the first and the last year in the training set. To ensure that the weight of each metric is comparable among different years, we take the absolute value and standardize the coefficients so that the most important feature for the year is rescaled to 1. It appears that competition metrics remain to be the most important features over the course of different years in similar fashion to our findings in Table 1. In addition, the proposed model suggests that the relationship between inside-cluster competition and campaign performance was stronger in 2010 while outside-cluster competition became the dominant factor in 2015. The reason for this is partly due to the fact that 2010 was Kickstarter's first year and there were not many campaigns with overlapping products. In our dataset, 80% of the campaigns launched in 2010 were in their own cluster.

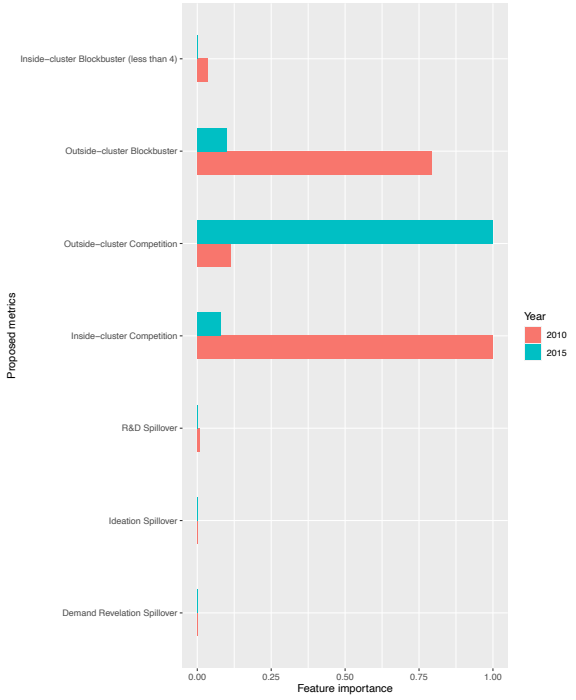

Figure 2: A comparison of the feature importance for the proposed metrics in the year of 2010 and 2015. We took the absolute value and then standardized all the coefficients to make the most importance feature 1.

6. Conclusion

Crowdfunding has gained popularity as an alternative to traditional sources of funding. It has opened up the possibility of larger funding for small businesses, but it has also created difficulties in determining the campaign features that will work to attract investors. This research leverages Natural Language Processing (NLP) techniques in order to develop a set of metrics that could aid entrepreneurs by understanding the mechanisms behind the decisions of backers. In addition, we propose a machine learning model to analyze the relationship between these metrics and campaign performance.

This paper contributes to the crowdfunding literature from the following aspects. First, we propose a new model, VCSVM, as a variation of Support Vector Machines (SVM) to allow the impact of predictors on campaign performance to change over time. Our unique Kickstarter dataset demonstrates that the proposed model outperforms other machine learning methods in terms of predicting campaign performance. This illustrates that the impact of each predictor on campaign performance may vary by year,

but due to specific circumstances that are difficult to quantify in a dataset.

Additionally, we propose three sets of metrics, including *learning, competition,* and *blockbusters*, to help entrepreneurs predict campaign performance. We have found that *learning* from the performance of recently launched products or concurrent products is not a good indicator of the performance of the same product. This means that the herd effect is not beneficial for entrepreneurs when deciding what product to launch. The present study also suggests that *competition* from both products of the same type as well as products of different types but belonging to the same Kickstarter category could help entrepreneurs predict the success of their campaigns. Lastly, by applying the VCSVM model, the results also show that the relationship between campaign performance and the number of *blockbusters*, i.e., the highly successful campaigns, can have a long-term impact on product performance prediction.

Nevertheless, we have developed our metrics and model primarily for commercial private goods on Kickstarter, a reward-based crowdfunding platform that follow an all-or-nothing methodology. These results may be subject to change if we apply the metrics and model to other types of goods, such as public goods (for example, the theater category), because consumers of private goods are more concerned with the benefit they can gain from the product than its social impact (Hong et al. 2018). Further, it would be interesting to investigate how results would differ for other types of crowdfunding, such as donation-based, lending-based, and equity-based, since the fundamental motives and drivers of these platforms differ. As an example, backers' self-perceived generosity and religiosity towards charitable giving can affect their charitable contributions (Isa et al. 2015). In addition, whether a crowdfunding platform follows a fixed funding (all-or-nothing) or flexible funding (keep what you raise) strategy can also affect the results since it influences the decision of backers to contribute to a project (Burtch et al. 2018; Strausz 2017). The answers to questions such as these are valuable for future research.

7. Reference

Abidin, T. F., Yusuf, B., and Umran, M. 2010. "Singular Value Decomposition for Dimensionality Reduction in Unsupervised Text Learning Problems," *ICETC 2010 - 2010 2nd International Conference on Education Technology and Computer* (4), IEEE, pp. 422–426.

Acs, Z. J., Audretsch, D. B., and Lehmann, E. E. 2013. "The Knowledge Spillover Theory of Entrepreneurship," *Small Business Economics* (41:4), pp. 757–774.

Acs, Z. J., Braunerhjelm, P., Audretsch, D. B., and Carlsson, B. 2009. "The Knowledge Spillover Theory of Entrepreneurship," *Small Business Economics* (32:1), pp. 15–30.

Agrawal, A., Catalini, C., and Goldfarb, A. 2014. "Some Simple Economics of Crowdfunding," *Innovation Policy and the Economy* (14:1), University of Chicago Press, pp. 63–97.

Burtch, G., Hong, Y., and Liu, D. 2018. "The Role of Provision Points in Online Crowdfunding," *Journal of Management Information Systems* (35:1), pp. 117–144.

Calantone, R. J., Chan, K., and Cui, A. S. 2006. "Decomposing Product Innovativeness and Its Effects on New Product Success," *Journal of Product Innovation Management* (23:5), pp. 408–421.

Chemla, G., and Tinn, K. 2020. "How Wise Are Crowds on Crowdfunding Platforms?," *SSRN Electronic Journal*.

Cosh, A., Cumming, D., and Hughes, A. 2009. "Outside Enterpreneurial Capital," *The Economic Journal* (119:540), Oxford Academic, pp. 1494–1533.

Da Cruz, J. V. 2018. "Beyond Financing: Crowdfunding as an Informational Mechanism," *Journal of Business Venturing* (33:3), Elsevier Inc., pp. 371–393.

Ellman, M., and Hurkens, S. 2019. "Optimal Crowdfunding Design," *Journal of Economic Theory* (184:October 2014), Elsevier Inc., p. 104939.

Flesch, R. (1979). How to write plain English. University of Canterbury. Available at http://www. mang. canterbury. ac. nz/writing_guide/writing/flesch. shtml.

Gallemore, C., Nielsen, K. R., and Jespersen, K. 2019. "The Uneven Geography of Crowdfunding Success: Spatial Capital on Indiegogo," *Environment and Planning A* (51:6), pp. 1389–1406.

Goldenberg, J., Lehmann, D. R., and Mazursky, D. 2001. "The Idea Itself and the Circumstances of Its Emergence as Predictors of New Product Success," *IEEE Engineering Management Review* (29:2), pp. 105–118.

Hong, Y., Hu, Y., and Burtch, G. 2018. "Embeddedness, Prosociality, and Social Influence: Evidence from Online Crowdfunding," MIS Quarterly: Management Information Systems (42:4), pp. 1211–1224.

Isa, N. A. M., Irpan, H. M., Bahrom, H. B., Salleh, A. B. M., & Ridzuan, A. R. B. (2015). "Characteristic affecting charitable donations behavior: Empirical evidence from Malaysia," *Procedia Economics and Finance*, 31, 563-572.

Jaffe, A. 1986. "Technological Opportunity and Spillovers of Research and Development," *American Economic Review* (79:5), pp. 984–1001.

Janku, J., and Kucerova, Z. 2018. "Successful Crowdfunding Campaigns: The Role of Project Specifics, Competition and Founders' Experience," *Finance a Uver-Czech Journal of Economics and Finance* (68:4), pp. 351–373.

Jiang, H., Wang, Z., Yang, L., Shen, J., & Hahn, J. (2021).

How rewarding are your rewards? A value-based view of crowdfunding rewards and crowdfunding performance. *Entrepreneurship Theory and Practice*, *45*(3), 562-599.

Johan, S. A., and Taylor, J. 2019. "Does Crowdfunding Democratize Success? Revisiting the Effects of Agglomeration and Localized Knowledge Spillover on Creative Projects," *SSRN Electronic Journal*, pp. 1–41.

Jones, P., and Ratten, V. 2020. "Knowledge Spillovers and Entrepreneurial Ecosystems," *Knowledge Management Research and Practice* (00:00), Taylor & Francis, pp. 1–7.

Kennett, P. ed. 2008. "Governance, Globalization and Public Policy - Google Books," *Edward Elgar Publishing*.

Kim, J., Lee, M., Cho, D., and Lee, B. 2016. "An Empirical Analysis of Semantic Network in Online Crowdfunding: Evidence from Kickstarter," *ACM International Conference Proceeding Series* (17-19-Augu).

Lin, Y., Yin, P., and Lee, W. C. 2018. "Modeling Dynamic Competition on Crowdfunding Markets," *The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018*, pp. 1815–1824.

Liu, J., Yang, L., Wang, Z., and Hahn, J. 2015. "Winner Takes All? The 'Blockbuster Effect' in Crowdfunding Platforms," *2015 International Conference on Information Systems: Exploring the Information Frontier, ICIS 2015*, pp. 1–11.

Liu, H., and Wang, Y. 2018. "The Value of Crowdfunding: An Explanation Based on Demand Uncertainty and Comparison with Venture Capital," *Emerging Markets Finance and Trade* (54:4), pp. 783–791.

Lu, X., Dong, F., Liu, X., and Chang, X. 2018. "Varying Coefficient Support Vector Machines," *Statistics & Probability Letters* (132), North-Holland, pp. 107–115.

Martínez-Climent, C., Mastrangelo, L., and Ribeiro-Soriano, D. 2020. "The Knowledge Spillover Effect of Crowdfunding," *Knowledge Management Research and Practice* (00:00), Taylor & Francis, pp. 1–11.

McMillan, J. 1979. "The Free-Rider Problem: A Survey," *Economic Record* (55:2), John Wiley & Sons, Ltd, pp. 95–107.

Mollick, E. R., and Kuppuswamy, V. 2014. "After the Campaign: Outcomes of Crowdfunding," *SSRN Electronic Journal*, pp. 1–18.

Mollick, E., and Robb, A. 2016. "Democratizing Innovation and Capital Access: The Role of Crowdfunding," *California Management Review* (58:2), pp. 72–87.

Mueller, D. C. 2008. "Public Choice: An Introduction," in *Readings in Public Choice and Constitutional Political Economy*, Springer US, pp. 31–46.

"NVCA Yearbook - National Venture Capital Association - NVCA." 2020.

Pelleg, D., Moore, A. W., and others. 2000. "X-Means: Extending k-Means with Efficient Estimation of the Number of Clusters.," in *Icml* (Vol. 1), pp. 727–734.

Porter, M. E. 2008. "The Five Competitive Forces That Shape Strategy," *Harvard Business Review* (86:1), pp. 1–17.

Salton, G., and McGill, M. J. 1986. *Introduction to Modern Information Retrieval*, USA: McGraw-Hill, Inc.

Soublière, J. F., and Gehman, J. 2020. "The Legitimacy Threshold Revisited: How Prior Successes and Failures Spill over to Other Endeavors on Kickstarter," *Academy of Management Journal* (63:2), pp. 472–502.

Strausz, R. 2017. "A Theory of Crowdfunding: A Mechanism Design Approach with Demand Uncertainty and Moral Hazard," *American Economic Review* (107:6), pp. 1430–1476.

Thies, F., Wessel, M., and Benlian, A. 2016. "Effects of Social Interaction Dynamics on Platforms," *Journal of Management Information Systems* (33:3), pp. 843–873.

Wang, J. C., and Hastie, T. 2014. "Boosted Varying-Coefficient Regression Models for Product Demand Prediction," *Journal of Computational and Graphical Statistics* (23:2), American Statistical Association, pp. 361–382.