Success in Democratized Financing: An Analysis of Indiegogo Campaigns Through Machine Learning and Causal Inference Perspectives

Quentin McTeer

5/02/2021

Contents

1 Abstract

This paper first utilizes a prediction dataset to employ econometric and machine learning models to calculate the likelihood that a given Indiegogo crowdfunding campaign reached its funding goal. Second, this analysis utilizes a subset of the prediction data with only campaigns that have reached their funding goal to employ a inverse probability weighted propensity score matching model to calculate the causal effect of a fully-funded Indiegogo crowdfunding campaign going inDemand on the campaign's funding percentage. The results from part one show that the random forest tree ensemble method does the best job of predicting whether a given campaign has been succesful, in other words, whether or not it has reached its funding goal. The results from part two show that, when confounding variables have been accounted for, that the causal effect of the inDemand program on funding percentage is approximately 544.73%, significant at the 5% level.

2 Introduction

2.1 Crowdfunding Background

For early-stage ventures, crowdfunding has increasingly become the gateway for overcoming financial bottlenecks. Through crowdfunding, many small investors can contribute to a proposed new product before the it hits the market. Contributions can range from a few dollars to substantial investments in high-technology tools. These contributions usually come in exchange for early access to the products, equity in the company, or some other benefit provided by the project owners seeking the capital. The market for this sort of democratized finance approach reached $13.9 billion globally in 2019, a number that will only continue to increase as it reaches more developing countries. The crowdfunding process takes place predominantly on online platforms, the most popular of which include Kickstarter and Indiegogo for early-stage business ventures. Other platforms such as GoFundMe serve as a financing option for fundraisers and other charitable campaigns. These sites allow creators to showcase their product to potential backers, and then raise capital over a set period of time typically determined by the crowdfunding platform. Backers, or funders, are the individuals who invest in the project. Crowdfunding projects can range greatly in both goal and magnitude, from small artistic projects to entrepreneurs seeking hundreds of thousands of dollars in seed capital as an alternative to traditional venture capital investment (Mollick (2014). One unique aspect of crowdfunding platforms is their rules surrounding funding goals. On Kickstarter, a campaign only receives the capital raised from its backers if the project reaches it's funding goal (which is set by the creator prior to launch on the platform). Indiegogo, however, allows for both a fixed approach (like Kickstarter) or a flexible approach where campaigns who did not meet funding goals can still receive funds raised from backers, but only after paying significantly higher fees to the platform compared to the fixed approach. Thus, most Kickstarter campaign pledges tend to cluster around 0% or 100% of the goal, whereas Indiegogo campaign pledges exhibit a slightly more even distribution across projects (Gallemore, Nielsen, and Jespersen (2019).

2.2 Research Questions

The primary question that many recent papers on crowdfunding answer relate to the likelihood that a particular project will be completely funded. This can be a crucial insight as entrepreneurs raise capital in early stages of corporate development. Understanding this likelihood can help entrepreneurs optimize their strategy and approach to fundraising so that they can be successful in their efforts to fund their ventures (Markas and Wang (2019). Most of the research on crowdfunding campaign success surrounds the prediction of Kickstarter project results (Mollick (2014), Kaminski and Hopp (2019), Markas and Wang (2019), Parhankangas and Renko (2017). The limited research that has been done on Indiegogo campaigns uses data that, as of this writing, is over six years old and focuses predominantly on the role of geography and social spaces in predicting crowdfunding success (Gallemore, Nielsen, and Jespersen (2019). Virtually none of the literature in the crowdfunding space deals with questions of causal inference or counterfactual estimation. The one exception is Chabot 2020 which utilizes both machine learning (for prediction) and econometric (for causation) techniques (Chabot (2020). The paper employs an instrumental variable to establish causation, taking advantage of a Kickstarter policy change

surrounding the maximum duration that a campaign could run for. For prediction, Chabot (2020) compared the AUC of linear and logistic regression (econometric techniques) against supervised machine learning models in their ability to predict if a Kickstarter campaign would be successful. This paper attempts to contribute to the literature on crowdfunding success. To accomplish this, the paper utilizes Indiegogo data provided by the web scraping service company Web Robots given that it is so understudied in the relevant literature. Utilizing econometric techniques, the analysis will estimate the monetary impact of a campaign's participation in Indiegogo's inDemand program. The inDemand program allows fully funded projects to continue raising capital after their campaign deadlines ends (usually 30-45 days after launch). Projects can opt into the inDemand program at any point before or after the campaign launches. This papers asks the question: What is the true causal effect of a project's participation in the inDemand program on the funding it receives relative to its goal (funded percentage). Additionally, this paper seeks to add to the ongoing discussion that sits at the intersection of machine learning and causal inference by building on the work of Chabot 2020 which found that some more sophisticated machine learning methods have higher predictive power than econometric tools when it comes to crowdfunding campaign success. To this end, the report compares LPM, logistic regression, lasso logistic regression, decision tree, and random forest (tree ensemble method) to benchmark econometric methods of prediction against those used in the machine learning world in their ability to forecast the success of an Indiegogo campaign.

3 Exploratory Data Analysis

3.1 Prediction Data Set

As mentioned above, this analysis utilizes Indiegogo data pulled from the web scraping service company Web Robots. The dataset represents 20,631 historic campaigns that are unevenly distributed between 2010 and 2020. Thus, this is a pooled cross-section spanning over one decade. Crowdfunding campaigns have notoriously low success rates as documented in previous literature. This is especially true for Indiegogo, where most research has pinpointed an average fully funded campaign rate between 8.5-11%, significantly lower than Kickstarter's 38% success rate (Gallemore, Nielsen, and Jespersen (2019). This is most likely due to the fact that Indiegogo campaigns can still receive some funds in the event that they do not reach their goal, reducing the incentive for the campaign to hit its target. Fortunately, the random sample of scraped campaigns provided for this paper by Web Robots has an approximate average success rate of 10% across all observations as shown in Figure 1 below, thereby making it representative of the true population mean.

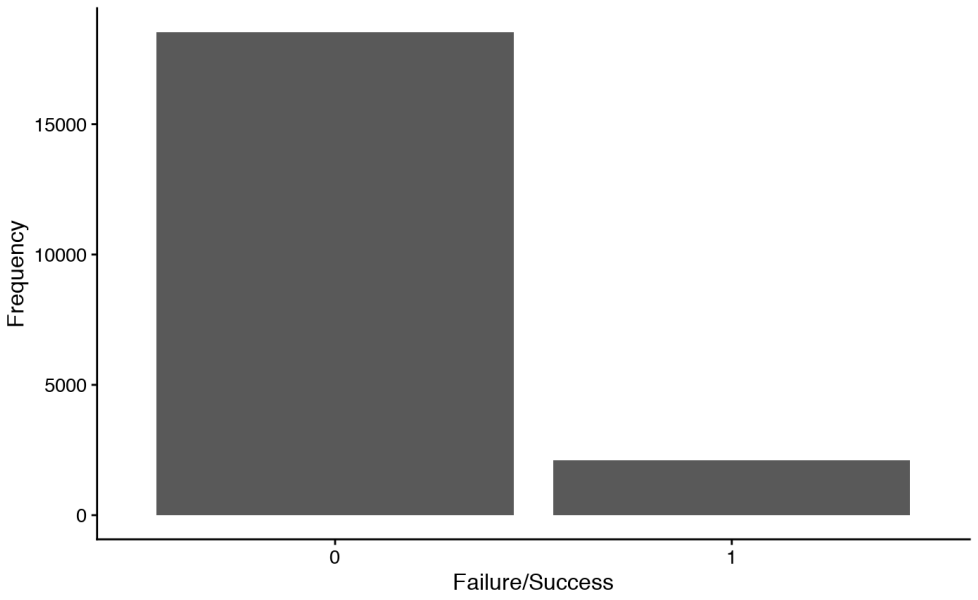

Figure 1: 10% of Indiegogo Campaigns Reach Their Funding Goal

Indiegogo's online platform breaks up projects into a variety of categories. These unique categories do not experience the same levels of funding or general success. As shown in Figure 2 below, of the twenty-eight categories listed on Indiegogo's website, Transportation, Audio, Productivity, Home, and Fashion/Wearables bring in the largest amount of capital from campaign backers.

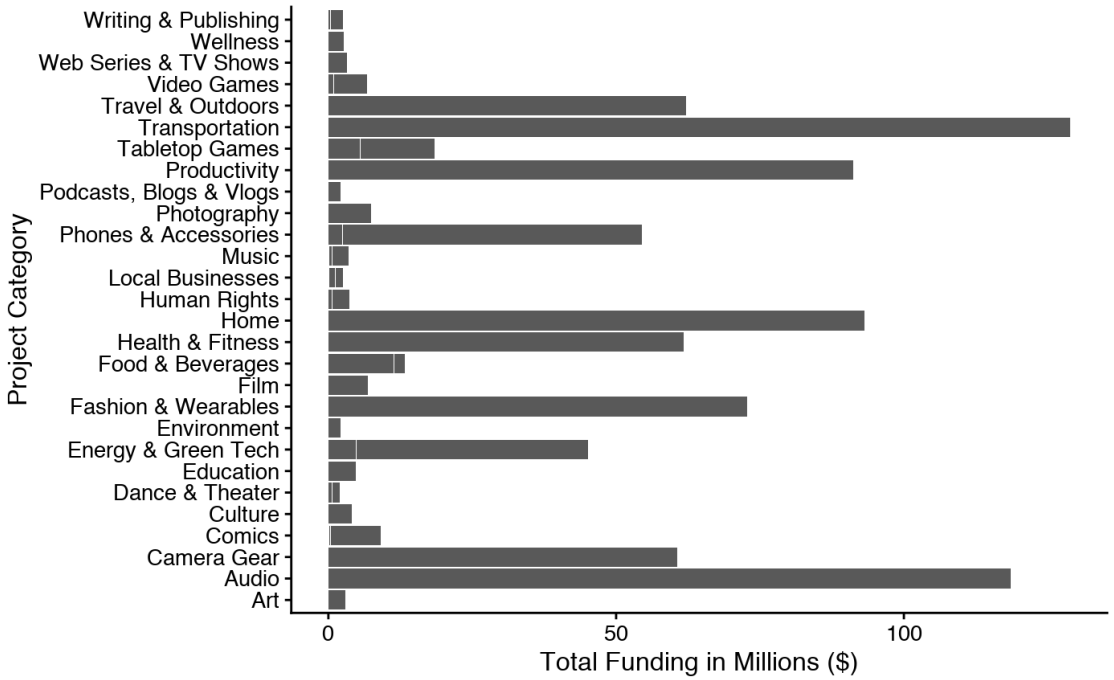

Figure 2: Total Funding By Project Category

Additionally, there are two time-specific trends to discuss. First, there is a discrepancy in how much funding projects receive by month of product launch. As shown in Figure 3 below, campaigns that launch in March or the summer months of May, June, and July experience significantly higher levels of overall funding than those launched in other months. This could be true for two reasons. Either because more campaigns are launched over the spring and summer months, resulting in higher funding, or because backers are more active and willing to finance campaigns during this time.

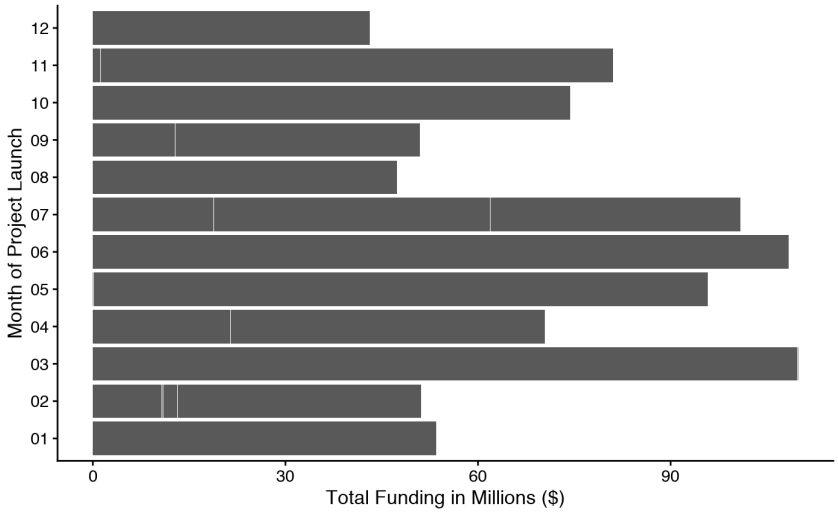

Figure 3: Total Funding By Project Launch Month

The second time-trend of interest in this paper is the success of campaigns across time. As Figure 4 shows below, there are significant time fixed-effects that impact the success of an Indiegogo project that will need to be considered in the empirical model. More recent projects are significantly more likely to succeed than older projects. This could either be due to sampling error within the data set, or it could be due to the value that experience plays in the ability of creators to launch a successful campaign. Previous literature has suggested that serial creators (those who create many campaigns) experience more success than those who only create one project (Chabot (2020). Either way, time fixed effects will need to be included in the empirical model.

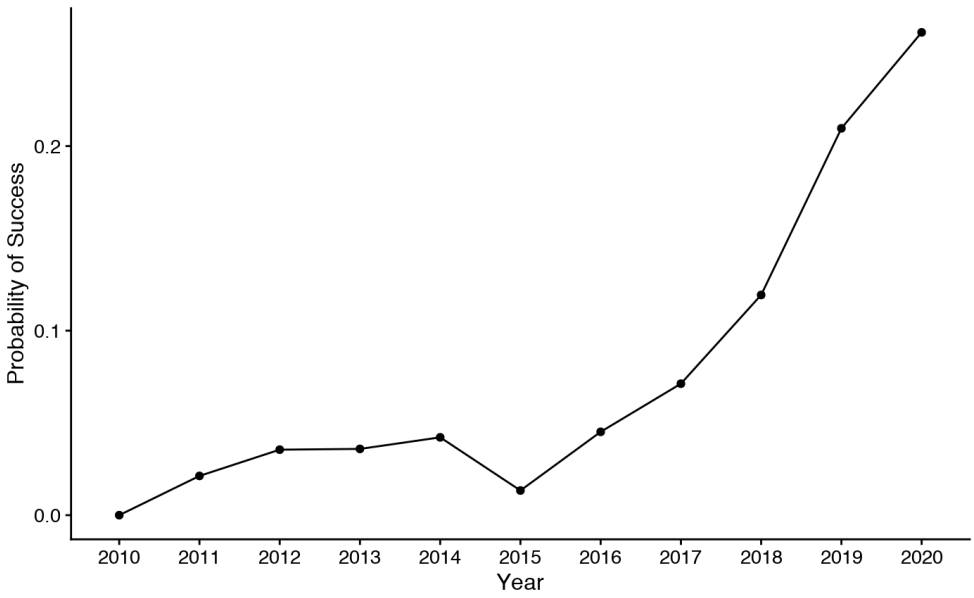

Figure 4: Probability of Project Success by Year

Projects can be started in many countries across the globe. Figure 5 below shows the distribution of projects internationally using currency of the funding goal as a proxy. The relative frequency shows that an overwhelming majority (78%) of campaigns are based in the US. Roughly 16% are based in Europe, 4% in Canada, and 1% in both Australia and Hong Kong.

currency	n.obs	relative.frequency
AUD	298	1%
CAD	922	4%
CHF	21	0%
DKK	14	0%
EUR	1625	8%
GBP	1479	7%
HKD	185	1%
NOK	1	0%
SEK	10	0%
SGD	40	0%
USD	16036	78%

Figure 5: Distribution of projects by currency

Figure 5 highlights some important summary statistics from the original dataset. State is an indicator variable taking a value of 1 if the campaign successfully met its funding goal and 0 if not. This will be the dependent variable in the prediction models. Average amount raised across campaigns is 42,895 US dollars with a large standard deviation of 312,051.

The average goal amount is approximately 26,700 US dollars a standard deviation of 48,880. Notice that min and max are 100.27 and 500,000 respectively. This is because campaigns with a project goal of less than 100 US dollars and more than 500,000 US dollars were filtered out so that only serious campaigns were being considered in the analysis. This methodology is documented in the previous literature surrounding crowdfunding (Mollick (2014).

Statistic	State	Amount Raised (USD)	Goal Amount (USD)	Time Period
mean	0.102	42,985	26,699.95	10,316
s.d.	0.303	312,051	48,879.57	5,956
min	0	1	100.27	1
max	1	17,595,741	500,000	20,631

Figure 6: Summary Statistics for Relevant Variables

3.2 Causal Inference Dataset

This paper utilizes a subset of the original Indiegogo dataset for its analysis of the monetary impact of Indiegogo's inDemand program. This subset will include 2,108 historic campaigns, representing all successful campaigns (those with a funding percentage at or over 100%). As mentioned previously, campaigns that meet their funding goal have the option to go inDemand after their campaign deadline has been reached, an opportunity that is not afforded to unsuccesful campaigns. By going inDemand, campaigns have the ability to continue raising capital from backers without setting new project goals or a deadline date. They also have the opportunity to interact and grow with their community of backers by building upon their project page post-campaign. Additionally, Indiegogo features inDemand campaigns on a subsection of their platform titled "inDemand Superstars" as well as on other pages of their website, allowing inDemand campaigns increased exposure to the 15 million users who visit Indiegogo every month. As shown in Figure 7 below, roughly 50% of successful Indiegogo campaigns choose to go inDemand. There is also a significant difference in how much funding inDemand campaigns make versus those those that choose to stop fundraising on the platform after their campaign deadline ends (2802% vs. 1646%). This funding percentage variable will be the dependent variable of this analysis as the paper seeks to determine the effect of going inDemand on total campaign funding. However, projects that choose to go inDemand and those that don't are not created the same. The figure below also shows that inDemand campaigns are less likely to be US-based, more likely to have a lower goal amount, much less represented in the transportation and audio project categories (which are the two categories that receive the most funding across all campaigns), much more likely to be represented in the home project category, and more likely to be represented in more recent time periods.

Characteristic	inDemand (TRUE)	inDemand (FALSE)	T-Stat
number of obs.	1062	1046	-
funded percent (mean)	2802	1646	-
us (mean)	0.7363	0.8394	5.834***
goal (mean)	22,631	28,609	3.71***
raised (mean)	372,489	364,441	-0.203
transport (mean)	0.037	0.098	5.677***
audio (mean)	0.037	0.2218	13.136***
home (mean)	0.1394	0.033	-8.825***
tperiod (mean)	17,117	13,056	-22.526***

Figure 7: inDemand Summary Statistics: Comparison of Means

4 Empirical Model

4.1 Prediction Models

4.1.1 Prediction Model Training

This analysis uses the entire dataset to train the linear probability model and logistic regression model, since these simple models have, in general, low variance (little chance of overfitting). Additionally, it is very common in econometric modelling to train on the whole dataset, so in order to effectively compare the method to various machine learning approaches, the whole dataset will be used. For the machine learning methods, the original dataset is split between the training data (77%) and the testing data (33%). This technique is known as the 'hold-out' method and is a form of cross-validation in order to compare the success of multiple models on new data. The split was chosen in order to create a balance in the number of successful campaigns between the two data sets, which will help improve the results of the models. It is especially useful when modeling with an outcome variable that is a 'rare occurrence,' which is the case with the successful campaign outcome variable.

4.1.2 Least Squares: Linear Probability Model & Logistic Regression

$$\text{State}_i = \beta_0 + \beta_1 \text{Goal}_i + \beta_2 \text{Tperiod}_i + \sum_{i=1}^{n} \beta_3 \text{Region}_i + \sum_{i=1}^{n} \beta_4 Month_i + \sum_{i=1}^{n} \beta_5 Category_i + \epsilon$$

The first models considered in the analysis are the econometric methods of ordinary least squares (OLS) regression and logistic regression. As shown in the above model, the binary dependent variable is State, an indicator variable taking a value of 1 if the campaign met its fundraising goal (success) and zero if not (failure). Because the dependent variable is binary, the OLS regression takes the form of a linear probability model, where the coefficients on β_1 through β_5 are interpreted as probability increases or decreases (depending on the sign). The goal of OLS regression is to minimize the sum of the squared residuals by finding a line of best fit that minimizes error between the line and the observed data. In logistic regression, the goal is to fit an s-curved line based on the maximum likelihood that the model parameters fit a particular distribution. The coefficients on logistic regression variables are interpreted as a log(odds) ratio. However, the

results can be exponentiated so that the coefficients can be intrepeted using the odds ratio. Both models will use the same independent variables to predict the chances of success for a given campaign, the fundraising goal of the campaign (Goal), the time period at the launch of the campaign (Tperiod), and fixed effects including countries and regions (Region), month of the year (Month), and category of the project type (Category). Tperiod was added to control for time-fixed effects, as the exploratory data analysis in the previous section showed more recent campaigns are more likely to succeed. Month, category, and region fixed effects are included because, as shown in the exploratory data analysis, these factors seem to receive disproportionate levels of backer funding in each of their respective levels, making them potentially strong predictors of campaign success.

4.1.3 LASSO and Ridge Logistic Regression

Both LASSO and Ridge regression are forms of regularization, a technique used to find the optimal balance between bias (error in prediction with training data) and variance (error in prediction with test data). Below is the optimization problem for LASSO (least absolute shrinkage and selection operator):

$$\min_\beta \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda \sum_{i=1}^{n} |\beta_j|$$

The goal of LASSO is to minimize the sum of squared residuals while also applying a penalty on the regression coefficient depending upon its explanatory power in the model in order to decrease bias and put more weight on variables with high predictive power. The penalty also serves to reduce the models dependence on the training data set by decreasing sensitivity in prediction of the outcome variable to changes in the feature/explanatory variable. Ridge regression applies a similiar penalty, except instead of the term $\lambda \sum_{i=1}^{n} |\beta_j|$ the penalty operator is instead $\lambda \sum_{i=1}^{n} \beta_j^2$. Because of this change in the penalty formulation, LASSO will more drastically reduce the number of coefficients because it will quickly force them to zero if they are irrelevant variables for predicting the outcome. Ultimately, the LASSO technique works better in models with many unidentified irrelevant variables and Ridge regression works better in models where most variables contribute to prediction of the outcome. However, both serve the same purpose, reduce overfitting and variance so that the model will better predict new campaign success. The penalty parameter is tuned using a bootstrap resampling technique to find the penalty parameter that maximized AUC.

4.1.4 Decision Tree

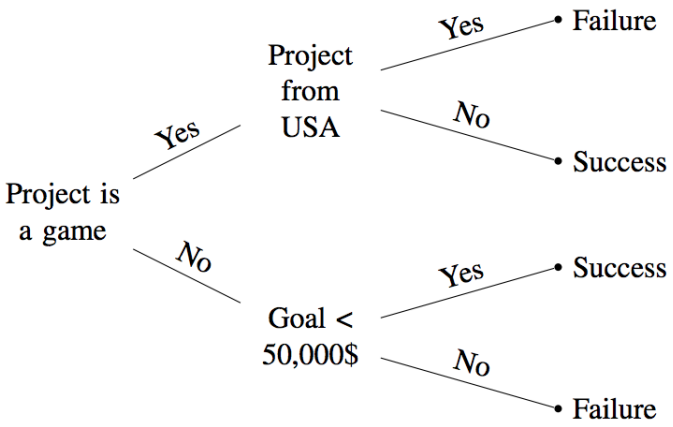

Figure 8: Example of Basic Decision Tree

Recursive partitioning (otherwise known as a decision tree classifier) is a non-parametric machine learning tool that is at the heart of other ensemble techniques like Random Forests. According to Safavian and Landgrebe (1991), the design can be divided into three steps. First, determining the optimal tree structure, then the feature selection rule at each internal node and finally, the decision rule to use at each internal node. Each of the aforementioned variables represents a different step in the tree design. When it comes to the choice of decision rule strategy to split, the most popular criteria is the Gini impurity index shown below.

$$\text{Gini} = 1 - \sum_{i=1}^{n} p^2(c_i)$$

In the equation above, $p(c_i)$ is the percentage of a class $p(c_i)$ in a given node. If the node is "pure," then one of the $p(c_i) = 1$ and the Gini impurity index is null. The decision tree will choose the split that minimizes the sum of the Gini impurity index. The decision tree package 'rpart' in R will be used to collect results for this analysis, and has three tuning parameters: cost complexity, tree depth, and minimum number of data points in a node that are required for the node to be split further. To effectively tune the model, the paper employed a bootstrap resampling technique to identify the ideal parameter values that maximizes AUC.

4.1.5 Random Forest (Tree Ensemble Method)

A random forest is an ensemble method based in the aforementioned decision tree. Ensemble methods use a single base learning algorithm to produce homogeneous base learners, learners of the same type, leading to homogeneous ensembles. In random forests, each tree in the ensemble is built from a sample drawn with replacement from the training data set. This form of sampling, otherwise known as bootstrapping, is a statistical technique for estimating quantities about a population by averaging estimates from multiple small sample series. In addition, instead of using all the features, a random subset of features is selected, further randomizing the tree. This random subset is selected at each level of the tree: the root

node, the intermediate nodes, and the leaf nodes. This process of bootstrapping and random feature selection is repeated n number of times depending on the number of trees selected. As a result, the bias of the forest increases slightly, but due to the averaging of less correlated trees, its variance decreases, resulting in an overall better model. This paper uses the 'ranger' package in R to create a random forest and will use 1000 decision trees. Aggregating trees and bootstrapping is a strategy for optimizing the bias/variance trade off called 'bagging.' There are a few hyper parameters that must be defined in the random forest model: number of random variables to consider in each tree and minimum number of data points in a node that are required for the node to be split further. These hyper parameters were tuned to find their optimal values in order to maximize accuracy and AUC using k-fold cross validation. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. In this case, for 10-fold cross validation, the dataset is split into 10 groups, and the random forest is trained and tested 10 separate times so each group gets a chance to be the test set.

4.2 Causal Model

This paper estimates the effect of a successful campaign's choice to go inDemand on the amount of funding the project receives relative to its goal (funded percentage). Because the inDemand program is optional, there could be significant selection bias in terms of the types of campaigns that choose to go inDemand versus those that do not. Projects that saw significant success during their initial one month duration might be more likely to opt into the program than those who narrowly met their funding goal. Additionally, campaigns from the US may be more likely to opt into inDemand than those from other regions. These differences make the treatment group (campaigns that choose inDemand) systematically different than those in the control group (campaigns that do not choose inDemand). In order to deal with this confoundedness issue, inverse probability weighted propensity scores will be used to add weights to a simple OLS regression of the inDemand indicator variable (1 if chose to go inDemand, 0 if not) on the funded percentage of the campaign in order to eliminate the indicator variable's correlation with the error term. This weighted propensity score technique is shown below:

$$Y_i = (\beta_0 + \beta_1 D_i + \epsilon) \times w_i$$

$$w_i = \begin{cases} 1 & \text{if } D_i = 1 \\ \frac{\hat{\Pr}(D_i|\mathbf{X}_i)}{1-\hat{\Pr}(D_i|\mathbf{X}_i)} & \text{if } D_i = 0 \end{cases}$$

$$\log[\frac{\Pr(D_i = 1)}{1 - \Pr(D_i = 1)}] = \gamma_0 + \gamma_j \mathbf{X}_i + \xi$$

The above model is broken up into three parts. First, the outcome model shows the OLS regression of inDemand status, D, on the percent of campaign goal funded, Y with error term ϵ and weights w. The second model shows that w is

calculated using the predicted probability of treatment given the covariates X. Figure 7 showed the covariates of the model, characteristics in which the average for treatment and control groups varied significantly. Those covariates include: whether or not the project is based in the US, the funding goal set by the creator, the amount raised by the campaign, a few category types (transport, home, audio, travel and outdoors, productivity, fashion/wearables), and the time period of the project. By matching with weighted propensity scores, these covariates will no longer confound the model. The final equation details the utilization of logistic regression to predict the probability that a campaign went inDemand given a certain covariate value. Once again, logistic regression uses maximum likelihood estimation to fit the most likely distribution of the data in order to calculate the parameters of the model (in this case, γ). The coefficients returned will be in the form of predicted probabilities that will become the weights used in the outcome model.

5 Results

5.1 Prediction Model Results

In classification problems (binary dependent variable) the ROC (Receiving Operating Characteristics Curve) AUC (Area Under the Curve) metric and the accuracy metric are effective measurements for comparing various models. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting 0s as 0s and 1s as 1s. The ROC curve itself has the true positive prediction rate (sensitivity) on the y axis and the false positive rate (1-specificity) on the x-axis. An excellent model has an AUC near 1 which means it has a good measure of separability. A poor model has an AUC near 0 which means it has the worst measure of separability. Additionally, accuracy is simply the proportion of the data that are predicted correctly. These two metrics are calculated using the 'yardstick' package in R, a part of the tidymodels universe of packages. Unfortunately, accuracy measurements could not be calculated for the logistic or linear regression since they were trained on all the data. In Figure 9 below, we see that the ROC-AUC value is highest for the Random Forest, with a 93.8% value as well as the highest accuracy of 92.7%. The LASSO and RIDGE regression performed about the same, while the econometric models performed slightly worse than the machine learning methods in categorizing campaign outcomes as success or failure.

Model	Accuracy	ROC-AUC
Linear Regression	-	0.890
Logistic Regression	-	0.895
Lasso Logistic Regression	0.925	0.924
Decision Tree: Gini	0.919	0.901
Ridge Logistic Regression	0.925	0.927
Random Forest: Gini	0.926	0.938

Figure 9: Comparing Econometric and Machine Learning Prediction Outcomes

Unlike the machine learning methods used here, the linear regression and logistic regression have interpretable coefficients. Figure 11 in the Appendix shows the coefficients on the campaign goal, time period, and individual campaign categories. For the exponentiated logistic regression model, the coefficients can be interpreted as probabilities from odds ratios by dividing the coefficient by one plus the coefficient. For example. if the project is in the audio category the probability of campaign success is 79.46% (3.87/4.87) larger than if the project is in the art category (the omitted category), significant at the 1% level. Further, the output from the logistic regression model shows that the categories of camera gear, audio, transportation, and home have the highest odds ratios and therefore are the most likely to succeed. The results for all four of these categories are significant at the 1% level.

Figure 12 in the Appendex reports the results for the linear probability model. The results on each coefficient are interpreted as increases/decreases in the probability that an Indiegogo campaign was successful. For example, if the campaign is in the audio category it is 31.4% more likely to succeed than if it is an art campaign (the omitted category), significant at the one

percent level. The campaigns most likely to succeed by category are audio, camera gear, tabletop games, productivity. Coefficients on all of these categories are statistically significant at the 1% level.

Both audio and camera gear show up as the most successful category types for both the logistic regression and LPM, however they start to differ on ranking the success of subsequent categories. This is most likely due to the difference in how the two models fit the data. Due to the binary nature of the dependent variable, it is likely that the results from the logistic regression are more accurate. However, since both regression types are only given a priori information about the characteristics of the campaign and also only know the goal, time period, category, month, and geographic location, both models likely suffer from omitted variable bias, which, in the best case scenario, make these results less than robust. However, despite this bias in coefficient values, the results still provide some indication of the relationship between these campaign characteristics and project success even if they do not provide the true causal effect, an important feature of econometric models not present in most machine learning approaches.

5.2 Causal Model Results

The first column of Figure 10 below shows a naive model of the impact of a campaign going inDemand on percent of funding goal met without any controls. The results of this model show that if the campaign went inDemand it increased its funded percentage by 1156.31%, ceteris paribus. However, ceteris is not paribus in this case and, as explained in the methodology section, inverse probability weighted (IPW) propensity scores were needed in order to homogenize the inDemandTRUE and inDemandFALSE campaigns across a variety of characteristics that could impact funding percentage. The results from the OLS regression with IPW propensity scores are reported in the second column. They show that, when confounding variables are accounted for, a campaign's choice to go inDemand increases its funded percentage by 544.73%, ceteris paribus. As indicated by the figure, these propensity score weighted regression results are statistically significant at the 5% level.

	Simple OLS	OLS Weighted Propensity Scores
(Intercept)	1646.144***	2080.245***
	(190.660)	(187.418)
inDemandTRUE	1156.308***	544.728**
	(268.616)	(275.110)
Num.Obs.	2108	2108
R2	0.009	0.002
R2 Adj.	0.008	0.001
AIC	42778.7	43244.3
BIC	42795.6	43261.3
Log.Lik.	-21386.330	-21619.147
F	18.530	3.921

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Figure 10: A naive OLS model vs. OLS with weighted propensity scores

5.3 Robustness Check

Because many machine learning techniques were employed in this paper, a robustness check was not included for the prediction side of the results. For the causal side, there are two methods utilized for checking the robustness of the IPW propensity score strategy. First, Eicker–Huber–White standard errors are applied to the model in the event that the data is heteroskedastic. The result of this check is a nearly identical effect of going inDemand on funded percentage (544.728%) with an identical p-value (0.04728), implying that the results are still significant at the 5% level. Second, month of campaign launch indicator variables are included in the weights calculation to test if the time of year of launch is confounding the relationship between going inDemand and funded percentage. The results of this check are an effect of 474.21% on the funded percentage of a campaign. However, the p-value is now 0.078 and the results are now only significant at the 10% level.

6 Contribution

This paper makes a few key contributions. First, the results of the econometric prediction models could actually be used by backers on Indiegogo, Kickstarter and other platforms. These models (LPM and logistic regression) highlight that projects within the audio and camera gear categories are more likely to succeed than other categories. Second, this paper contributes to the field of econometrics by measuring the predictive power of linear and logistic regression. It does this by building on the work of Chabot 2020, through the use of Indiegogo campaigns, to find that more complex machine learning algorithms (in this case, the random forest tree ensemble method) are better able to predict if a crowdfunding campaign will meet its goal than the econometric models. Second, this paper contributes to the crowdfunding literature through its causal model results highlighting the effect that going inDemand has on a campaign's funded percentage. These results could be used as both a marketing tool for Indiegogo as they try to convince successful campaigns to continue to raise capital on their site, or as important information for campaigns that have reached their funding goal and are concerned about the true added value of going inDemand. Thus, the results of the prediction and causal models provide important information for all parties in the crowdfunding space: backers, campaign creators, and the platforms themselves. Finally, the combination of both prediction and causal strategies related to Indiegogo campaign funding highlights the tremendous value of a multi-faceted approach when it comes to analyzing industry data. Marrying the machine learning world of data science and the causal inference world of economics can provide unique insights on issues in business. This can be done when used in tandem (wielding both machine learning and econometric models for prediction), or to tackle the issue through different perspectives (predicting campaign success while also measuring the causal effect of campaign's choice to go inDemand).

7 Limitations

Indiegogo campaigns only reach their funding goal 10% of the time. Because positive classification is a rare occurence within the data, classification with machine learning and econometric models will both struggle. This is the reason why even the best model (random forest) only achieved 92.7% accuracy. The other 7.3% of classifications were successful campaigns misclassified as unsuccessful. This problem is due to overfitting, the Web Robots campaign data used to train the models contained so many unsuccessful campaigns that the prediction models classified nearly every campaign as unsuccessful. Additional a priori variables would be useful for future attempts to predict campaign success, perhaps including the duration of the campaign and the number of campaigns previously started by the creator. The results are also limited because of the structural difference between Indiegogo and Kickstarter, which allocate raised campaign funds differently. As previously mentioned, unlike Kickstarter, Indiegogo will still pay out raised capital to campaigns that do not meet their goal so long as they pay a large fee. This policy works to disincentive campaigns who are near their goal towards the end of their campaign to push for full funding. This creates a campaign that has all the characteristics of success (and will probably be classified as a success by a machine learning or econometric model) but still fails to reach its goal. Finally, there are limitations to the causal model. Although the initial results of the IPW propensity score approach

showed promising results significant at the 5% level, multiple robustness checks proved that these results may only be significant at the 10%, making it more likely that the results are spurious. Additionally, because Indiegogo is the only crowdfunding platform that runs the inDemand program, there is very little external validity to apply the causal model results to other crowdfunding platforms. This makes the impact of the paper results extremely local, potentially relating only to Indiegogo itself.

8 Appendix

term	estimate	std.error	statistic	p.value
goal_usd	0.000	0.000	-4.846	0.000
tperiod	0.000	0.000	30.550	0.000
as.factor(category)Audio	3.866	0.233	16.576	0.000
as.factor(category)Camera Gear	4.204	0.174	24.201	0.000
as.factor(category)Comics	1.773	0.152	11.664	0.000
as.factor(category)Culture	0.282	0.155	1.818	0.069
as.factor(category)Dance & Theater	-1.773	0.275	-6.457	0.000
as.factor(category)Education	0.755	0.419	1.804	0.071
as.factor(category)Energy & Green Tech	2.726	0.167	16.303	0.000
as.factor(category)Environment	-1.871	0.335	-5.583	0.000
as.factor(category)Fashion & Wearables	2.488	0.115	21.585	0.000
as.factor(category)Film	-0.574	0.200	-2.869	0.004
as.factor(category)Food & Beverages	0.997	0.256	3.897	0.000
as.factor(category)Health & Fitness	2.157	0.161	13.368	0.000
as.factor(category)Home	2.893	0.135	21.362	0.000
as.factor(category)Human Rights	1.367	0.264	5.167	0.000
as.factor(category)Local Businesses	-0.973	0.432	-2.253	0.024
as.factor(category)Music	-0.354	0.158	-2.234	0.025
as.factor(category)Phones & Accessories	2.800	0.144	19.477	0.000
as.factor(category)Photography	1.413	0.251	5.638	0.000
as.factor(category)Podcasts, Blogs & Vlogs	-0.092	0.224	-0.409	0.682
as.factor(category)Productivity	3.130	0.103	30.398	0.000
as.factor(category)Tabletop Games	2.892	0.139	20.776	0.000
as.factor(category)Transportation	3.231	0.149	21.724	0.000
as.factor(category)Travel & Outdoors	2.962	0.169	17.484	0.000
as.factor(category)Video Games	0.522	0.237	2.198	0.028
as.factor(category)Web Series & TV Shows	-1.528	0.501	-3.050	0.002
as.factor(category)Wellness	-0.118	0.315	-0.375	0.708
as.factor(category)Writing & Publishing	0.103	0.214	0.484	0.629

Figure 11: Logistic Regression: Odds Ratio Coefficients

term	estimate	std.error	statistic	p.value
goal_usd	0.000	0.000	-4.210	0.001
tperiod	0.000	0.000	22.729	0.000
as.factor(category)Audio	0.314	0.023	13.696	0.000
as.factor(category)Camera Gear	0.293	0.019	15.451	0.000
as.factor(category)Comics	0.073	0.015	4.817	0.001
as.factor(category)Culture	-0.052	0.011	-4.569	0.001
as.factor(category)Dance & Theater	-0.016	0.003	-5.238	0.000
as.factor(category)Education	0.044	0.007	6.675	0.000
as.factor(category)Energy & Green Tech	0.152	0.024	6.269	0.000
as.factor(category)Environment	0.017	0.006	2.820	0.017
as.factor(category)Fashion & Wearables	0.124	0.015	8.380	0.000
as.factor(category)Film	-0.092	0.009	-10.286	0.000
as.factor(category)Food & Beverages	-0.021	0.011	-1.959	0.076
as.factor(category)Health & Fitness	0.093	0.009	10.501	0.000
as.factor(category)Home	0.186	0.021	8.914	0.000
as.factor(category)Human Rights	0.055	0.005	10.049	0.000
as.factor(category)Local Businesses	0.028	0.005	5.841	0.000
as.factor(category)Music	-0.050	0.006	-8.320	0.000
as.factor(category)Phones & Accessories	0.127	0.012	10.920	0.000
as.factor(category)Photography	0.052	0.007	7.375	0.000
as.factor(category)Podcasts, Blogs & Vlogs	-0.075	0.011	-6.983	0.000
as.factor(category)Productivity	0.198	0.018	10.975	0.000
as.factor(category)Tabletop Games	0.200	0.017	12.019	0.000
as.factor(category)Transportation	0.188	0.017	11.115	0.000
as.factor(category)Travel & Outdoors	0.185	0.014	13.346	0.000
as.factor(category)Video Games	0.019	0.008	2.450	0.032
as.factor(category)Web Series & TV Shows	-0.081	0.006	-13.029	0.000
as.factor(category)Wellness	-0.041	0.007	-5.882	0.000
as.factor(category)Writing & Publishing	-0.018	0.009	-2.026	0.068

Citations

Chabot, Victor. 2020. "Benchmarking Econometric and Machine Learning Models to Predict Crowdfunding Campaigns Outcomes on Kickstarter." PhD thesis, HEC Montreal School of Business. https://biblos.hec.ca/biblio/memoires/chabot_victor_m2020.pdf.

Gallemore, Caleb, Kristian Nielsen, and Kristjan Jespersen. 2019. "The Uneven Geography of Crowdfunding Success: Spatial Capital on Indiegogo." *Sage Publications*, Economy and Space, 1–18.

Kaminski, Jermaine, and Christian Hopp. 2019. "Predicting Outcomes in Crowdfunding Campaigns with Textual, Visual, and Linguistic Signals." *Small Business Economics* 55 (2020): 627–49.

Markas, Ruhaab, and Yisha Wang. 2019. "Dare to Venture: Data Science Perspective on Crowdfunding." *SMU Data Science Review* 2 (1): 19.

Mollick, Ethan. 2014. "The Dynamics of Crowdfunding: An Exploratory Study." *Journal of Business Venturing* 29 (1): 1–16.

Parhankangas, A., and M Renko. 2017. "Linguistic Style and Crowdfunding Success Among Social and Commercial Entrepreneurs." *Journal of Business Venturing* 32 (2): 215–36.