Predicting Kickstarter Campaign Success

ECON 573 Research Paper

Andrew Harris, Calvin Ryan, Harvey Duperier

UNC Chapel Hill
November 2022

Contents

1 Introduction

Every day, people around the world take on new challenges. This could be quitting a job to start a new business, turning a hobby into lifestyle, or transforming a passion into an obsession. People are looking to turn their dreams into entrepreneurial ventures, but they lack sufficient funding and financing. One way to finance these ventures is through Kickstarter, which is an online crowdfunding platform that helps people bring their ideas to life.

Since its launch in 2009, Kickstarter has become a popular crowdfunding option for entrepreneurs across the world. The platform has generated over $6 billion for more than 200,000 successful projects that won the support of millions of backers. Kickstarter campaigns span a variety of creative categories including art, journalism, games, and technology. Kickstarter uses an "all-or-nothing" funding approach, meaning creators set a funding goal for their campaign and receive funding if and only if they reach that goal. If a project is successful, Kickstarter collects a 5% fee of the proceeds, which is the basis of its business model. Because neither Kickstarter nor creators benefit from unsuccessful campaigns, both parties have an interest in maximizing the likelihood projects will succeed.

This paper considers whether machine learning techniques can be used to accurately predict Kickstarter campaign success and identify common characteristics of successful campaigns. This paper may be used by creators and crowdfunding platforms to identify ways to increase the likelihood of campaign success, thus increasing crowdfunding revenues and profits. Using data from over 20,000 Kickstarter campaigns between 2009 and 2017, this paper considers several machine learning techniques from ECON 573, including logistic regression, Lasso regression, Ridge regression, classification trees, bagging, random forest, boosting, SVMs, LDA, QDA, and KNN classification. Of these techniques, tree-based methods were the best tools for predicting campaign success, and pruned classification trees offered accurate predictions and interpretable results.

2 Literature Review

There happens to be a decent amount of economic literature related to predicting the success of Kickstarter campaigns. A 2016 paper, titled “Crowdfunding success factors: The characteristics of successfully funded projects on crowdfunding platforms”, looked into project-specific and factor-specific aspects that influence crowdfunding success (Koch & Siering, 2016). In addition to having a similar research topic, this paper is highly relevant to our research as it uses logistic regression as part of its methodology, and we do the same. One recent research paper published in 2019, titled “Early prediction of the outcome of Kickstarter campaigns: is the success due to virality?”, investigated factors that may predict Kickstarter campaign success including virality (Golosovsky & Solomon, 2019). This paper found that the a priori intrinsic property of the community of backers of a campaign was more influential on campaign success than the communication influence among the backers. The paper also provides advice to project founders on how to have more successful campaigns. This advice includes providing updates on project information during the funding period, increasing activity on crowdfunding platforms, and providing informative texts, pictures, and videos to backers. A 2017 research paper, titled “Guidelines for Successful Crowdfunding”, took a different approach to crowdfunding success research by examining success from the perspective of the project creator and backer; it also considered how the size of crowdfunding platforms may affect success (Forbes & Schaefer, 2017). The paper used literary research and interviews to develop four fundamental questions that are important for potential creators:

- “What crowdfunding platform should I use?”
- “What should my funding goal be?”
- “What should my reward options be?”
- “How should I construct my video?”

Though this was more of a qualitative research paper, it acts as another example of research that tries to crack the code to crowdfunding success and it uniquely provides guidelines as a tool to help predict success.

Each of the pieces of literature mentioned above demonstrate that predicting crowdfunding campaign success is an evolving research topic. Not only is this type of research beneficial for project creators, but it is beneficial for the platform itself, which collects fees from successful campaigns. It is our goal in our research to contribute to the existing literature on crowdfunding success and provide value to crowdfunding platforms and creators. Our research approach is different from existing research approaches on Kickstarter success. The dataset we have selected for this research was not used in most of the comparative literature we have seen, meaning our research may provide novel insights. Additionally, we make use of 12 classification models to predict the likelihood of campaign success and understand the common qualities of successful campaigns. Overall, this research can be considered an extension of the three papers referenced by Koch & Siering, Golosovsky & Solomon, and Forbes & Schaefer. This research expands upon the regression methods used in Koch & Siering, investigates some different factors for campaign success from Golosovsky & Solomon, and it provides more specific suggestions with deeper quantitative analysis than Forbes & Schaefer, which can be used as a tool to help predict project success.

3 Data

3.1 Data Selection

The data used for this project was collected by Rachel Downs and Muhammad Ghuari from the University of Texas at Austin. The data was scraped using webrobots.io in February 2017 and is available for download at kaggle.com. The dataset contains information on 20,632 Kickstarter campaigns launched between May 2009 and February 2017. The raw data included the following variables that were used in this study:

- Status: Factor with 4 levels: successful, failed, live, canceled, and suspended.
- Goal: Numeric campaign funding goal in native currency.
- Category: Factor with 25 levels describing project category.
- Currency: Native currency of the campaign.
- Static_usd_rate: Date-specific USD conversion rate.
- Pledged: Numeric sum of all contributions in native currency.
- Backers_count: Total number of backers for each campaign.
- Country: Factor with 21 levels describing campaign country of origin.
- Name_len and Name_len_clean: Numeric character length used in the name of the campaign.
- Blurb_len and Blurb_len_clean: Numeric character length used in the short description of the project (i.e., the “blurb”).
- Deadline/Created/Launched Weekday: Factor with 7 levels describing the weekday of campaign creation/launch/end.
- Deadline/Created/Launched Month: Numeric value describing the month of campaign creation/launch/end.
- Deadline/Created/Launched Year: Numeric year from 2009 to 2017.

- Create_to_launch_days: Number of days between pre-launch project creation and campaign launch.
- Launch_to_deadline_days: Number of days between campaign launch and campaign end.

3.2 Data Cleaning and Transformation

Currency values in denominations other than USD were converted using the static conversion rate provided in the original data. Then, campaign goals and pledges were adjusted to February 2017 dollars using monthly CPI figures from the U.S. Bureau of Labor Statistics. This was done to ensure goals and pledge amounts would be comparable across different countries and time periods.

Campaigns with State values other than "successful" or "failed" were excluded (3198 cases or 15.5% of the original data). Additionally, campaigns from Luxembourg (2) and projects in the Comedy category (1) were excluded since these occurrences were extremely rare and caused cross-validation errors. Missing values in the Category variable were replaced with an "Other" category. Year and month variables were converted to factors to account for seasonal effects and differences in funding patterns for different years. After cleaning, there were no missing values in the final data set.

A new logical "Success" variable described the outcome of the campaign. A new "Average Pledge" variable was created by dividing the CPI-adjusted pledge amount by the total number of campaign supporters. Many Kickstarter campaigns encourage consumer support by offering discounted products, early access, or merchandise to consumers, and this variable gives a rough estimate of the perceived value of these benefits. This variable may also serve as a weak descriptor of campaign quality, which is essential to understanding consumer support of crowdfunding. After the above data cleaning and transformations, the following variables were used in the our efforts to model and predict the success of Kickstarter campaigns:

- Success: Logical, with 1 indicating a successful campaign.
- Goal_adj: The goal amount in USD and adjusted for inflation.
- Country: Factor with 20 levels.
- Category: Factor with 23 levels.
- Name_len and Name_len_clean
- Blurb_len and Blurb_len_clean
- Deadline/Created/Launched Weekday: Factor with 7 levels.
- Deadline/Created/Launched Month: Factor with 12 levels.
- Deadline/Created/Launched Month: Factor with 9 levels.
- Create_to_launch_days: Numeric.
- Launch_to_deadline_days: Numeric.
- Avg_pledge: Numeric variable described above.

4 Methods

After cleaning, the data was randomly assigned to training and validation sets. Approximately 80% of observations were assigned to the training group and were used to develop the models seen below, which predict whether a Kickstarter campaign will be successful. Model performance was measured by the misclassification rate on the remaining 20% of observations. The models implemented below include logistic regression, Lasso regression, Ridge regression, classification trees, bagging, random forest, boosting, SVMs, LDA, QDA, and KNN classification.

5 Results

5.1 Logistic Regression

A logistic regression model was constructed to predict the likelihood a Kickstarter campaign would be successful, based on the variables listed above. Some brief observations:

- Funding goal (goal_adj) was negatively associated with project success and statistically significant, suggesting small funding goals make success more likely.
- Average pledge amount (avg_pledge) was positively associated with success, indicating projects that offer a high value to backers are more likely to be successful.
- Several country factors were significant and most were negative relative to the US base category, suggesting projects outside the US are less likely to succeed.
- Most of the category factor coefficients were significant, suggesting category plays an important role in the likelihood of a project gaining support from backers.
- Name length was significant and positively associated with success, indicating longer project names may help generate backer support.
- Most of the weekday coefficients were insignificant, except for the deadline being on a Tuesday and campaign being launched between Monday and Wednesday. According to this model, these timing choices may increase the likelihood of success.
- The number of days between campaign launch and the deadline is negatively associated with campaign success, suggesting longer campaigns may not have an advantage in reaching funding goals.
- None of the year or month variables appear to have a significant effect on the likelihood of campaign success, according to the logistic regression.

In addition to the above results, this model had a misclassification rate of 25.07% on the test set, which serves as a baseline for comparison for the other techniques below.

5.2 Lasso Regression

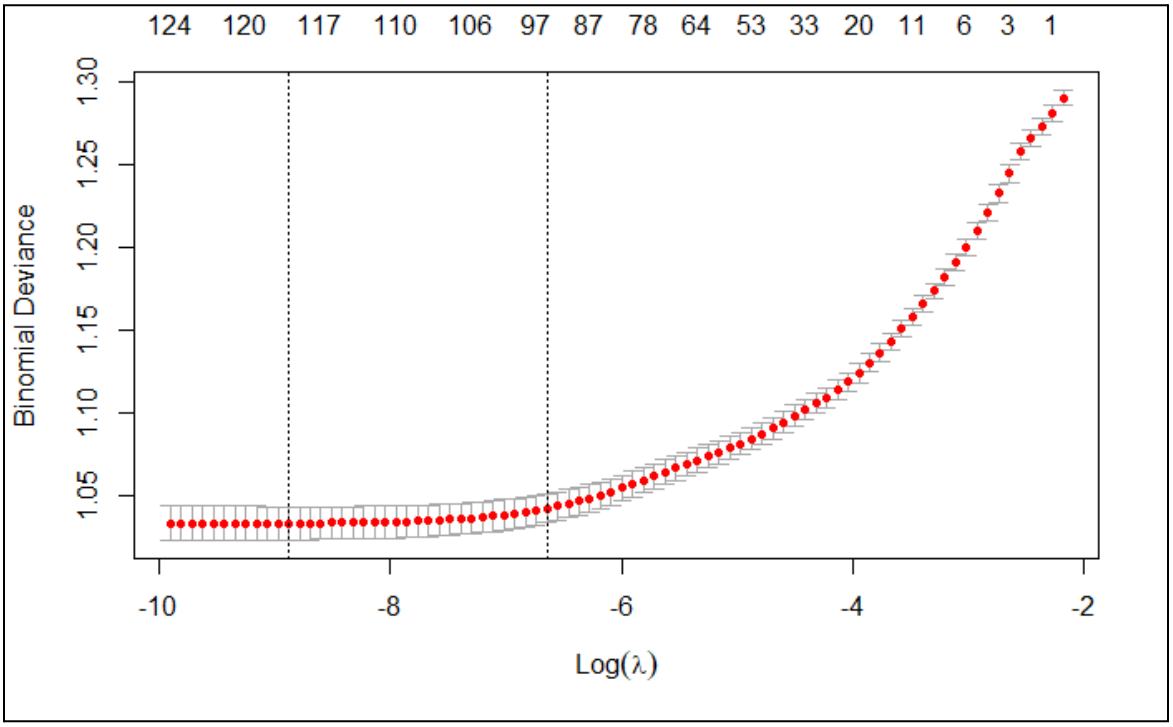

Figure 1 - Lasso Regression λ Optimization

The Lasso regression uses a tuning parameter (λ) to shrink regression coefficients, ultimately reaching 0 if λ is large enough. This allows for variable selection and should decrease overfitting commonly observed in logistic regression. Using 10-fold cross validation on the training set, the model identified the value of λ that minimized deviance and the λ within 1 SE of the minimum. The minimum and 1SE λ models eliminated 5 and 26 (of 125) regression coefficients respectively, though most exclusions were associated with the month and year variables that were insignificant in the logistic regression. When used to predict the outcomes of the validation set, the minimum λ Lasso model had a misclassification rate of 24.99%, a slight improvement compared to the logistic regression.

5.3 Ridge Regression

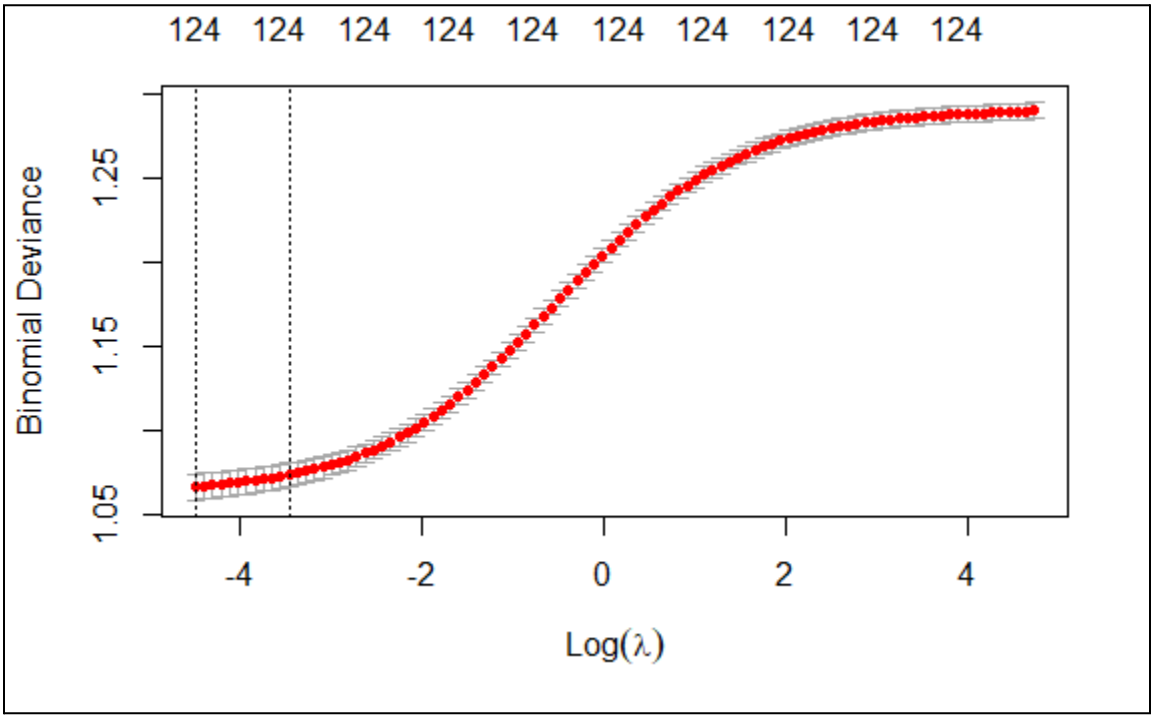

Figure 2 - Ridge Regression λ Optimization

The Ridge regression is constructed similarly to the Lasso, but its tuning parameter λ is calculated differently and shrinks model coefficients toward (but not equal to) 0. Once again, 10-fold cross validation was used to identify the λ values that minimized deviance and the λ 1 SE above the minimum. Both performed poorly on the validation set, and the minimum deviance λ Ridge model had a misclassification rate of 29.09%, which is approximately 4 points higher than the logistic regression. Given this result, the Lasso is preferable to the Ridge regression when predicting the likelihood a Kickstarter campaign will be successful.

5.4 Classification Tree

This paper implemented classification trees to predict the success and failure of Kickstarter campaigns. Classification trees are constructed using binary recursive partitioning, which iteratively divides the data to classify results into distinct groups. The initial classification tree was constructed using a complexity parameter (CP) of 0, which allows the tree to continue making partitions so long as it improves fit. This initial tree had 426 splits and a misclassification rate of 23.67% on the validation set. Using 10-fold cross validation on the initial classification tree, CV error was minimized at a CP of approximately 0.0031, which was used to prune the original tree.

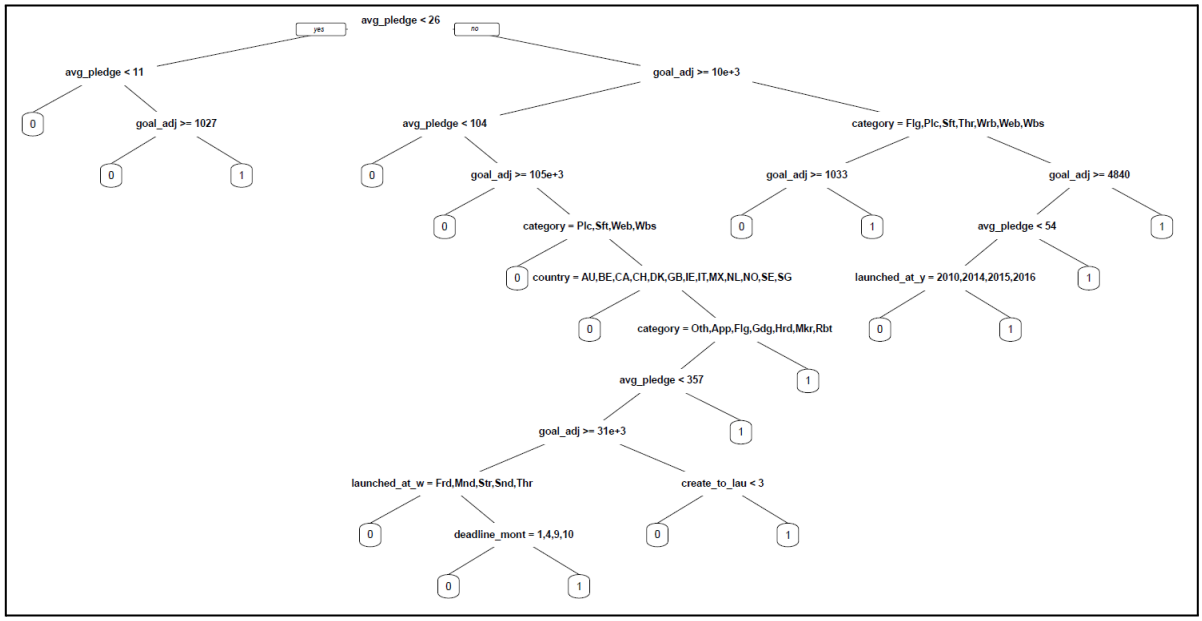

Figure 3 - Pruned Decision Tree

The pruned classification tree (shown above) had 19 splits and 20 terminal nodes, making it much simpler than the original tree. Splits included the variables for average pledge, goal, category, country, weekday of launch, deadline year, launch month, and days from project creation to launch. The misclassification rate of the pruned classification tree on the validation set was 20.80%, which is an improvement compared to the unpruned tree and the logistic regression.

5.5 Bagging

Bootstrap aggregation, also known as 'bagging', resamples the training data and trains classification trees on each of these new samples. These trees are aggregated to construct a complete model. Using this approach, 100 bootstrap replications were sampled from the training data and used to construct decision trees with a CP of 0, which were aggregated into a single model. The out-of-bag misclassification estimate from this process was 20.57%, but testing the bagged model on the validation set produced an observed misclassification rate of 18.88%, which is an improvement on the pruned classification tree and logistic regression.

However, despite producing accurate results, interpretation of bagged models is difficult. The variable importance of the model (Figure 4), illustrates which variables, if excluded, decrease accuracy the most. Project category, average pledge, and funding goal appear to play an essential role in predicting whether Kickstarter campaigns will be successful, but this model reveals little more about the impact of these variables.

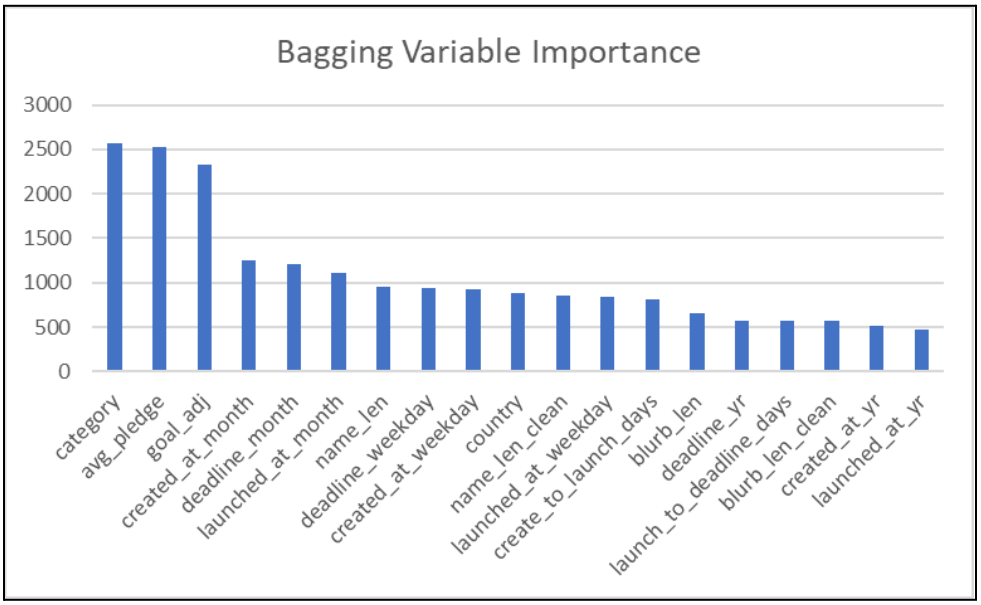

Figure 4 - Bagging Variable Importance

5.6 Random Forest

Random forest is similar to bagging in that it resamples the training data before training separate decision trees and aggregating their results to construct a single model. However, random forest also samples a subset of predictors before constructing each tree, which should produce uncorrelated trees and improve the accuracy of the aggregated model. When applied to the Kickstarter training data, 1000 trees with 4 features (mtry = 4) were sampled to produce the complete model. The out-of-bag misclassification estimate for this random forest model was 19.83%, but testing on the validation set produced an observed misclassification rate of 18.65%, which is similar to bagging and better than the logistic regression. As with bagging, random forest faces interpretability issues, and project category, average pledge, and funding goal were important features for explaining the likelihood of project success.

5.7 Boosting

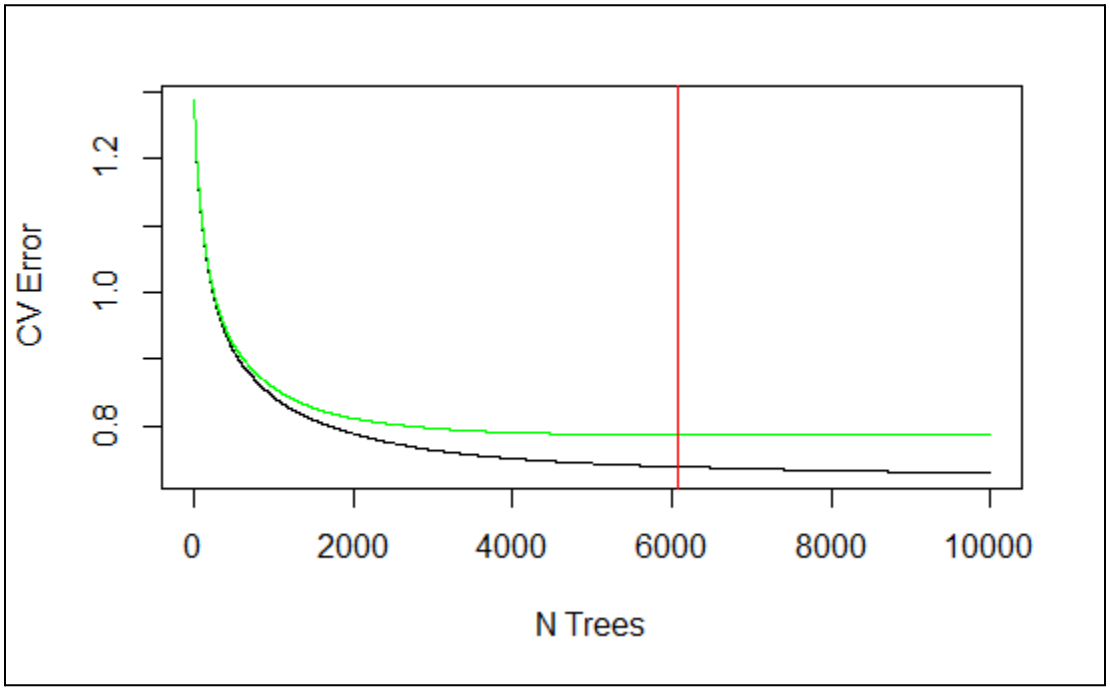

Figure 5 - Boosting Error (Out-of-Sample Test Error in Green with Minimum in Red)

Boosting is similar to bagging in that it resamples training observations to construct unique decision trees, which are ultimately aggregated into a single, complete model. However, unlike bagging, which constructs trees in parallel, boosting constructs trees sequentially by weighting poorly predicted observations in the construction of the next tree. Using 10-fold cross validation, out of sample misclassification was minimized at 6082 trees, which was used to construct the final boosted model. When tested on the validation set, the boosted model had a misclassification rate of 18.59%, which is similar to the random forest and bagging approaches. Based on these results, it appears tree-based models do a relatively good job predicting the likelihood of success for a Kickstarter campaign.

5.8 SVMs

Support vector machines (SVMs) use hyperplanes to separate data into different classes and maximize the margin between groups. To classify Kickstarter campaigns as successes and failures, two SVMs were constructed using two different kernels. Kernels define how SVMs manipulate the training data to make classifications. The first approach utilized a linear SVM kernel and used 8281 support vectors. This approach had a misclassification rate of 26.85% on the validation set. The second approach used a radial kernel, which allows for more flexible classification boundaries compared to the linear kernel. Using 8470 support vectors, the radial SVM had a test misclassification rate of 25.30%. Based on these results, SVMs were not an improvement on other methods in this analysis, including logistic regression, though it is possible another kernel or calibration may improve these models.

5.9 LDA

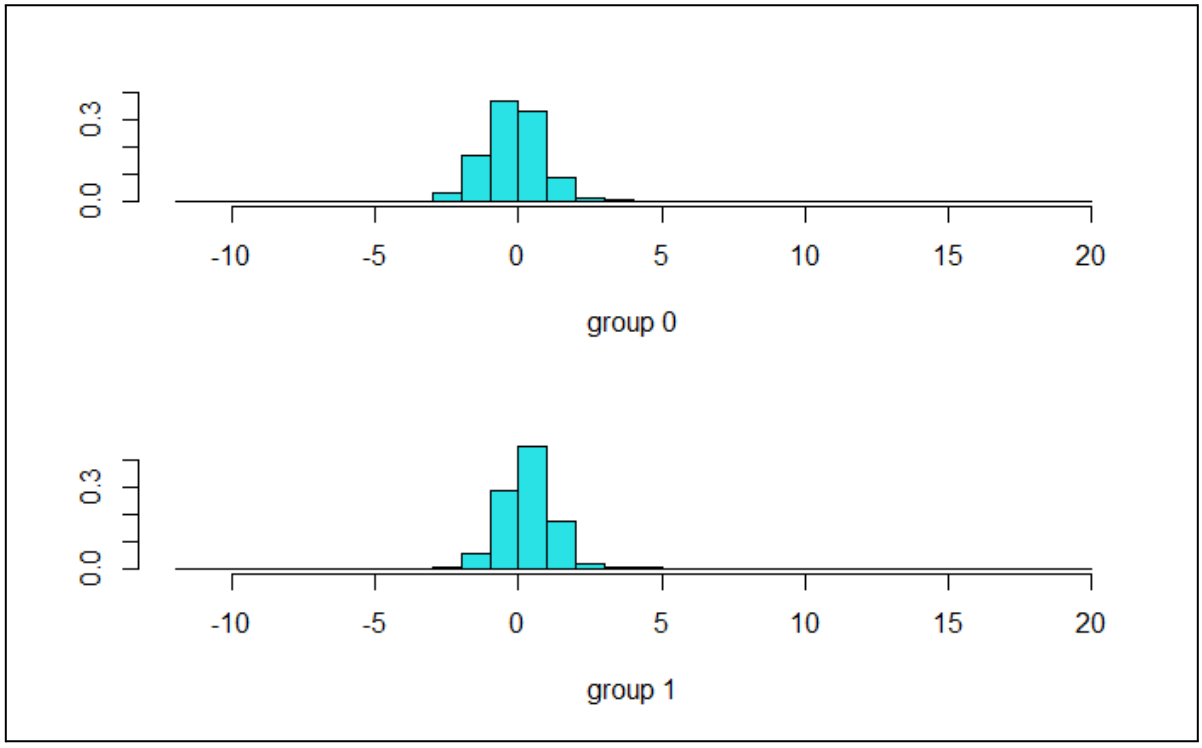

Figure 6 - LDA Histograms for Groups 1 (Success) and 0 (Failure)

Linear discriminant analysis is a dimension reduction technique that produces a linear combination of input variables that maximizes the separability between multiple classes of observations. When applied to the Kickstarter training data, factor variables are excluded by the assumptions of LDA, and the LDA distributions for successes (1) and failures (0) are seen above. Using these distributions, the model can estimate whether a campaign is likely a success or failure. The misclassification rate for LDA was 33.02% on the validation set, which is worse than all methods above, including the logistic regression. Considering the importance of factor variables in other models in this paper, the necessary exclusion of these variables in LDA may explain the relatively poor performance of this approach.

5.10 QDA

QDA is conducted similarly to LDA, but it allows for different covariance matrices between classes. As a result, QDA allows for much more flexible relationships between predictors, and it is often preferable when there are nonlinear boundaries between classes. When applied to the Kickstarter data set, the misclassification rate for QDA was 60.44%, which is much worse than LDA and more than twice as high as the misclassification rate for the logistic regression. Given these results, QDA is by far the worst performing approach used in this paper. Both LDA and QDA were relatively poor approaches for predicting the likelihood of success for a campaign, and some of this may be due to the omission of categorical predictors, which were important components in other models.

5.11 KNN Classification

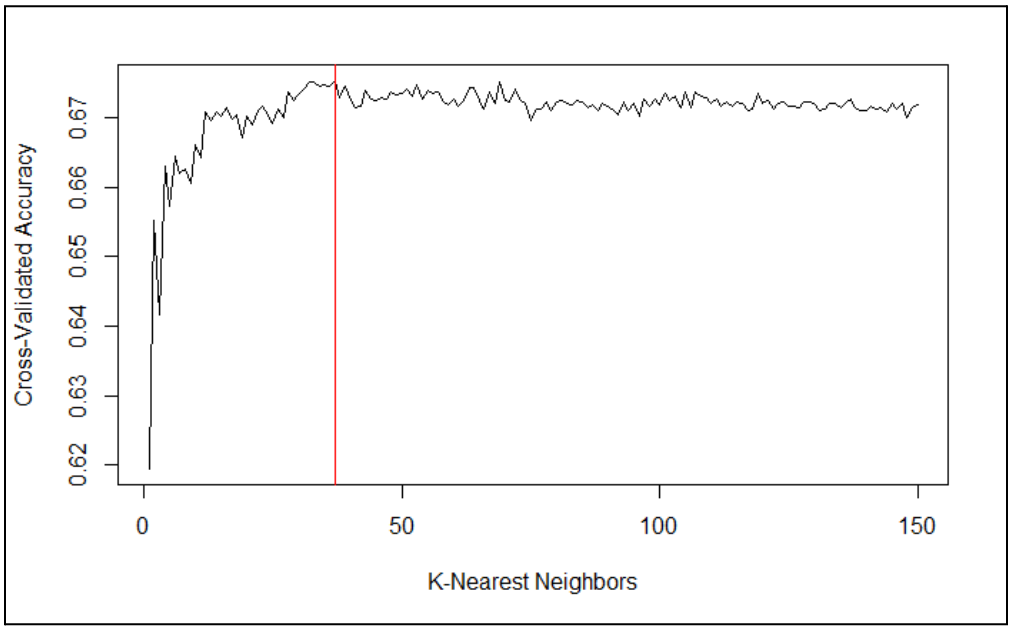

Figure 7 - KNN 10-Fold CV Accuracy (Maximum in Red)

K-nearest neighbors (KNN) classification is a relatively simple approach to predicting group membership, where an observation is classified by the closest K points using euclidean distance and majority vote. To begin, numeric predictors in the training and validation set were scaled using sample mean and standard deviation, while categorical variables were omitted by necessity. Using 10-fold cross validation on the training set, KNN accuracy peaked when K was 37. Using the closest 37 points, the validation set was classified using the training data, achieving a misclassification rate of 31.64%. While not the worst performing approach implemented in this paper, KNN significantly underperformed most other techniques, including the logistic regression. When considered with the LDA and QDA above, KNN performance suggests omission of categorical variables are detrimental to the prediction of Kickstarter campaign success.

6 Discussion

Classification Method	Misclassification Rate	Misclassification Rate (%)
Boosting	0.185886	18.59
Random Forest	0.18646	18.65
Bagging	0.188755	18.88
Decision Tree w/ Pruning	0.207975	20.80
Lasso Regression	0.249857	24.99
Logistic Regression	0.250717	25.07
SVM w/ Radial Kernel	0.253012	25.30
SVM w/ Linear Kernel	0.268503	26.85
Ridge Regression	0.290878	29.09
KNN	0.316408	31.64
LDA	0.330465	33.05
QDA	0.604418	60.44

The misclassification rate of campaign successes varied significantly depending on which classification method was used. The logistic regression served as our baseline due to its technical simplicity and interpretable coefficients, and this approach achieved a misclassification rate of 25.07%. The Lasso regression performed similarly with a misclassification rate of 24.99%, while omitting variables that did not significantly improve the model. However, the Ridge regression, another shrinkage method similar to the Lasso, performed worse than the logistic model with a misclassification rate of 29.09%.

Compared to the logistic, Lasso, and Ridge regressions, tree-based models clearly outperformed in terms of accuracy. The boosting, random forest, and bagging approaches all performed similarly with misclassification rates between 18.59% and 18.88%, while the pruned

decision tree had a misclassification rate of 20.80%. All these approaches were a significant improvement on the logistic and Lasso regressions in terms of accuracy. However, boosting, random forest, and bagging are far less interpretable than the regression approaches; whereas, the pruned decision tree is highly interpretable and still a significant improvement on previous approaches. Of the tree-based methods, the pruned decision tree is perhaps the most helpful for Kickstarter and its creators because projects can be quickly evaluated as likely successes or failures.

The remaining approaches, SVMs, KNN, LDA, and QDA, offered little additional predictive value relative to the logistic regression and tree-based methods. The SVM (Radial Kernel) and SVM (Linear Kernel) had misclassification rates of 25.30% and 26.85% respectively, which are slightly worse than the logistic regression. SVMs are also difficult to interpret due to the high dimensionality of the Kickstarter campaign data, which is a disadvantage compared to the regression approaches and the pruned decision tree. Assumptions of the LDA and QDA model necessitated factor variables were omitted from model construction, and these methods achieved misclassification rates of 33.05% and 60.44% respectively, while being difficult to interpret. KNN is a simpler alternative to these approaches but using K = 37 produced a relatively high misclassification rate of 31.64% on the validation set. Considered together, SVMs, KNN, LDA, and QDA are relatively poor classification methods for predicting Kickstarter campaign success, especially when compared to the logistic regression and tree-based methods.

Considering the findings above, random forest, bagging, and boosting had the greatest accuracy compared to all other methods, including the logistic regression baseline. However, the pruned decision tree (Figure 3) provided similarly strong predictive results with the added benefit of being highly interpretable. Therefore, we concluded that the pruned decision tree is the preferred model for predicting and understanding Kickstarter campaign success. Based on this model and other

findings throughout this paper, we formulated the following advice for increasing the likelihood of Kickstarter campaign success:

- Small funding goals increase the likelihood of campaign success, meaning creators receive more funding and Kickstarter collects more fees. Unrealistic goals decrease the likelihood of success and may discourage backers. This does not benefit creators or the platform.
- Higher average pledge amounts are associated with a higher likelihood of success. Kickstarter campaign creators should try to maximize donations per backer by providing quality incentives to support their campaign.
- The category of a Kickstarter campaign is extremely important and related to the likelihood of success. Popular categories on Kickstarter, such as apps, websites, and gadgets, are negatively associated with success. These markets may be oversaturated with campaigns and individual campaigns may be difficult to find. Creators should look for ways to stand out in these categories or consider other funding options.
- Most international funding campaigns on Kickstarter are less likely to succeed than US-based campaigns. To mitigate this effect, campaigns based outside the US, creators should explore options to make their campaigns accessible to American donors.
- Campaigns with little to no support after a significant amount of time are less likely to succeed, while possibly cluttering the market and distracting from more viable projects. Creators in this situation should adjust their campaign characteristics to be more viable by lowering goals, changing incentives, or changing categories where appropriate. Kickstarter should moderate its marketplace to decrease the number of less viable projects to increase the likelihood of success for all projects.

Future changes in the Kickstarter marketplace may mean the characteristics associated with campaign success will eventually change, meaning ongoing research is necessary. Regardless,

Kickstarter and other crowdfunding platforms will play an essential role in the world's entrepreneurial landscape for years to come, and the application of machine learning techniques will provide valuable insights on maximizing the likelihood of campaign success.

7 Works Cited

About Kickstarter. Kickstarter. Retrieved November 13, 2022, from https://www.kickstarter.com/about

Bring your creative project to life. Kickstarter. Retrieved November 13, 2022, from https://www.kickstarter.com/learn

Consumer Price Index for All Urban Consumers: All Items in U.S. City Average. U.S. Bureau of Labor Statistics. Retrieved November 12, 2022, from http://research.stlouisfed.org/fred2/data/CPIAUCSL.txt

Forbes, H., & Schaefer, D. (2017, May 9). *Guidelines for successful crowdfunding*. Procedia CIRP. Retrieved November 13, 2022, from https://www.sciencedirect.com/science/article/pii/S2212827117301178?ref=pdf_download&fr=RR-2&rr=769af1befe20adcf

Kindler, A., Golosovsky, M. & Solomon, S. Early prediction of the outcome of Kickstarter campaigns: is the success due to virality?. *Palgrave Commun* **5**, 49 (2019). https://doi.org/10.1057/s41599-019-0261-6

Koch, J.-A., & Siering, M. (2016, July 15). *Crowdfunding success factors: The characteristics of successfully funded projects on crowdfunding platforms*. Retrieved November 13, 2022, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2808424

Why kickstarter campaigns fail? Kaggle. Retrieved November 13, 2022, from

https://www.kaggle.com/datasets/thedevastator/most-kickstarter-campaigns-fail-here-s-why?resource=download&select=kickstarter_data_with_features.csv