Predicting the Success or Failure of a Kickstarter Project

Jonathan Ching, David Han, Rajat Jain, Pranav Tawar, Jay Juang
Statistical Learning at HKU

Introduction

Overview

Crowdfunding is a powerful alternative means of funding a project that has been made widespread over the last few years due to the advent of technology, namely the internet and online payment. Rather than approach a traditional lender such as a bank or a venture capitalist, an entrepreneur can seek many small loans from a large number of people. He or she would pay them back either with interest like a regular loan or with perks and rewards from the project.

Kickstarter is perhaps the most renowned and widely used crowdfunding platform in the world. It is an online site where people with a project or idea ranging from manufacturing drones to producing a documentary can call for funding from the general public. To date, there have been over 4 billion dollars pledged to various projects on Kickstarter from a total of roughly 49 million pledges[1]. Kickstarter is home to hundreds of thousands of projects, and has allowed a variety of innovative ideas to come to life.

The funding process generally follows the following procedure. First, a project creator registers for Kickstarter and customizes his profile and personal information. The creator then makes a project page and uploads a project description, photos, reward schemes, and sometimes a video advertisement. The creator sets a target funding goal and a funding timeline for the project and opens his project up to the public. (Unlike traditional loans with interest, the public is incentivized to fund the project in return for certain exclusive products.) At the end of the funding duration, if the funding goal has been reached, the project is deemed as successful and the creator is given the crowdfunded money; however, if the funding goal was not reached, the project is deemed unsuccessful and the creator receives none of the money.

From this system of "all or none" approach on Kickstarter, it is clear that it is critical to reach funding that is equal to or greater than your target goal. A project with a $1,000,000 goal that raises $800,000 will end up with an empty pocket, compared to a project with a $10,000 target that is successful in reaching its goal.

Motivation

Our motivation for conducting our research is to help creators on Kickstarter better understand the different aspects of their project in relation with success. We aim to empower them with the knowledge on how to focus on the key areas for success. In addition, by creating predictive models, creators may be able to estimate their own likelihood of success and make appropriate changes to better reach their goal.

[1] https://www.kickstarter.com/help/stats

Research Questions

Within this report, we explore two central questions: (1) What factors contribute to the success of Kickstarter projects, and (2) How much do these different factors play into the success of a project. In other words, we aim to explore the significance of different variables on the success of a Kickstarter project and identify the project features that are the most critical.

Methods

Data Source

Our data was retrieved directly from Kickstarter's internal API by crawling through available Kickstarter web pages and storing the JSON response for each web page on October 18, 2018. We collected over 200,000 projects from the last decade, though we are missing some of the older projects due to a change in Kickstarter policy that limits the number of viewable projects to a single user[2]. However, all recent projects are included in the dataset; it is only older projects dating before April 2015, that have been randomly excluded. This resulted in data that is skewed more heavily toward the trends of recent years, but seemed acceptable since the data was weighted more towards recent trends at the time of writing and should not have a large impact on predictive models for the future.

Data Cleaning and Feature Engineering

After collecting the data as a series of JSON object, we parsed the data in python and stored them as a CSV file using the pandas dataframe. The original JSON representing a project contained over 100 different key-value pairs, most of which we dropped during the JSON to CSV conversion since they did not have any applicable meaning such as links to the creator profile avatar images or the link background color for the project page.

After converting the data into a table, we derived new data from columns in the original data that seemed useful, but were not directly applicable in building a predictive model. For example, we derived a column for funding duration based on the original data on the funding start date and end date. Another example of derived data was converting the project description into two new columns of word count and character length. We also normalized all currency to USD using the USD conversion rate (in the original data from Kickstarter) for better interpretability.

One important decision we made was to drop the column pertaining to subcategories and only using the parent categories. Kickstarter projects are grouped into 15 broad parent categories such as technology, film & video, games, etc. However, these 15 categories are then further subdivided into hundreds of smaller categories such as classical music, horror movies, and web comics that were far too detailed and specific to provide meaningful insights. Since there is obvious correlation between parent category and their subcategories, we decided to utilize parent category only.

Since our data was directly scraped from Kickstarter's database, the data was very clean. Roughly 1000 projects (~0.5%) had missing location information, and we performed listwise deletion on those data points. We also performed listwise deletion on

[2] https://webrobots.io/kickstarter-datasets/

all duplicate elements, of which there were roughly 20,000 (~10%). We also removed all projects that were currently ongoing, cancelled by the creator for external reasons, or suspended by Kickstarter due to fraud because those project states were not of use to our research purpose. The projects in unusable states were roughly 10,000 rows (~5%), so our final version of the dataset consisted of approximately 174,000 rows.

Variables of Interest

After scraping, processing, and cleaning the data, we were left with 18 columns of data.

Variable	Meaning
parent_category	project category e.g. dance
launch_month	project launch month
deadline_month	project deadline month
country	country of origin
location.type	type of location e.g. city
staff_pick	whether or not it was selected as a staff favorite
spotlight	Whether or not a project was specially featured
creator_has_slug	whether or not creator's user profile has a customized url endpoint
state	successful is project reaches goal funding or failed if not

Table 1.1 Categorical Data

Variable	Meaning
funding_duration_days	number of days for which funding was open
pre_funding_duration_days	number of days between project creation and funding start
blurb_length	length of project description
blurb_word_count	word count of project description
name_length	length of project name
name_word_count	word count of project name
usd_goal	funding goal in USD
usd_pledged	amount actually funded in USD
backers_count	number of project backers

Table 1.2 Numerical Data

Half of the data columns are categorical (Table 1.1) and half of are quantitative (Table 1.2). The variables for staff_pick, spotlight, and creator_has_slug are all boolean type.

Analytical Procedure

Our goal is to decode the underlying structure and hidden distribution of factors of how a Kickstarter project becomes successful. We would like to: (1) identity whether a project at its inception has a high chance of getting successfully funded and (2) understand the distribution of successful projects across the variables. We implement our classification and models in R.

We split the 170k rows of data into Training and Testing set, by sampling 30% out of all indexes as testing dataset, and setting the rest as training dataset. The testing dataset was used to evaluate if the model overfit the data and yielded a low accuracy outside of the non-training data. We used the test set to estimate the error rate and assess the performance of each of our classifiers.

Within the scope of our project, we have attempted a total of five different approaches for predictive models. These are: Logistic Regression, Random Forest, Bagging, Naive Bayes, and Support Vector Machines (SVM). We implemented logistic regression because predicting success vs failure is a clear case where logistic regression could provide a reasonably accurate but highly interpretable model. We chose random forest and bagging because our data was relatively noisy, and these types of models tend to deal better with noisy data. We chose Naive Bayes since, like logistic regression, modelling a probability of success fits its use case. Lastly, SVM was chosen because it is currently one of the most popular training methods for obtaining a high predictive accuracy despite its lack of interpretability.

In order to test forecasting potential, we have the same setup as the above except removing variables that would only be known after the campaign has completed. Such variables include backers_count, usd_pledged, and spotlight[3]. The idea is to assess if the success of the projects could be attributed to the remaining variables that are known before the funding has begun.

Finally, to evaluate which of the models are most effective in predicting Kickstarters' success, we compute the mean squared error (MSE) of success/failed classification from all models. For binary tasks, we used the area under the ROC curve (AUC) which was computed using outputs of the classifiers (separating hyperplane distances for SVMs and outcome probabilities for RFs). We also look to see which models have the highest accuracy rate and use that as the ultimate judge of which models are the most effective. We note that both AUC and MSE are more discriminative than the accuracy metric, and are not sensitive to handle imbalance of 2-class distributions.

Result

Descriptive Statistics

Our dataset contained exactly 174,297 projects, none of which were missing any data (see data cleaning section above for more details). Of these projects, 100,076 were successful and 74,221 failed, giving us an overall success rate of 57.4%. This shows that the data was not too skewed towards either the successful projects or the failed projects.

	mean	std	min	25%	50%	75%	max
funding_duration_days	33.189711	12.417624	1	29	30	35	91
pre_funding_duration_days	41.166285	108.73645	0	2	10	34	2593
launch_month	6.2087357	3.3379701	1	3	6	9	12
deadline_month	6.6347671	3.2855949	1	4	7	9	12
blurb_length	114.34084	25.138261	1	105	125	132	150
blurb_word_count	19.290848	4.8248561	1	17	20	23	35
name_length	34.888047	15.790832	1	22	34	49	85
name_word_count	5.7180904	2.7247951	1	4	6	8	21
usd_goal	38333.44	1041018.2	1	1500	5000	12000	130994986
usd_pledged	11731.506	86146.233	0	114	1558	6216	10266846
backers_count	142.65137	971.70022	0	4	27	86	105857

Table 2.1 Numerical Data Characteristics

[3] Projects are only listed on Spotlight if the project has exceeded the goal amount within a quick time.

We had a total of 15 different parent categories. The largest category by number of projects was music with 24,781 projects, closely followed by film & video with 24,577 projects. Technology and art came third and fourth at 18,674 and 18,126 projects respectively. Dance and journalism were the two least active categories with 3066 and 4117 projects respectively.

136,798 projects (roughly 78% of the total projects) originated from the United States. The second most active country, the United Kingdom only produced 15,594 projects in comparison. As for making staff pick, only 23,429 projects, or about 13% of projects, were able to make it to the list. 50,572 creator profiles were customized to the extent of having their own personal url, corresponding to the creator_has_slug column.

Detailed information regarding the distribution of quantitative variables outlined in Table 1.2 can be found in Table 2.1. It is interesting to note that projects are launched almost uniformly throughout the year, and that the majority of projects have a funding duration of 25 to 35 days.

Exploratory Analysis

Our first approach to exploring our data for additional insights was to break it down by category (Figure 2.1). For each category of Kickstarter projects, we conducted a distribution breakdown of user and project statistics per category. According to the analysis, contrary to public perception of Kickstarter as a technology focused platform, the category with highest number of success cases is music, while the most number of failed projects are in the technology category. Dance projects had the highest rate of success, but was one of the lowest in terms of absolute success count. The average contribution per backer is highest in the technology category, with around $120 per person. Each project had an average of 260 backers. The average goal amount of technology projects is $78,000, second only to the film & video at $110,000.

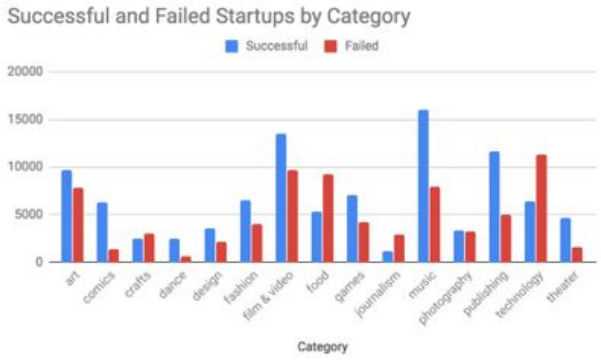

Figure 2.1 Success vs Failure by Category

For the country breakdown, the country with the highest average numbers of backers is Hong Kong (Figure 2.2). It is higher than US, with 150 backers on average. However considering the number of projects that are produced each year, US has a much higher volume than that of HK and other countries combined.

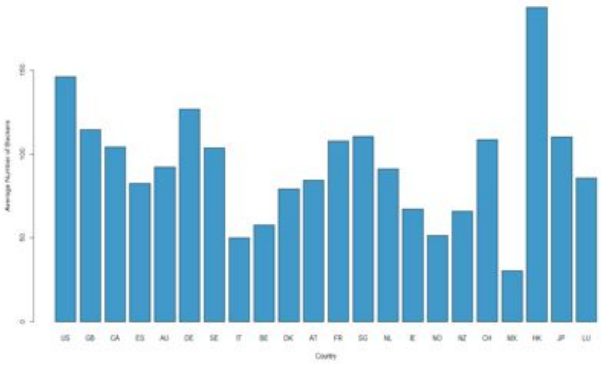

Figure 2.2 Number of Backers by Country

One important metric that we would like to understand is that how traction affects the success rate of one project. However, we lack the time series to do meaningful analysis for tractions. What we were able to determine was that there was a minimal number of projects that achieved between 50-99% of their project goal. If a project reached 50% of its funding goal before the deadline, there was a 97% likelihood that it would continue to reach its full funding goal. Properly exploring the role traction plays in crowdfunding would be one of the future directions that we would like to process in the future if we had the relevant data.

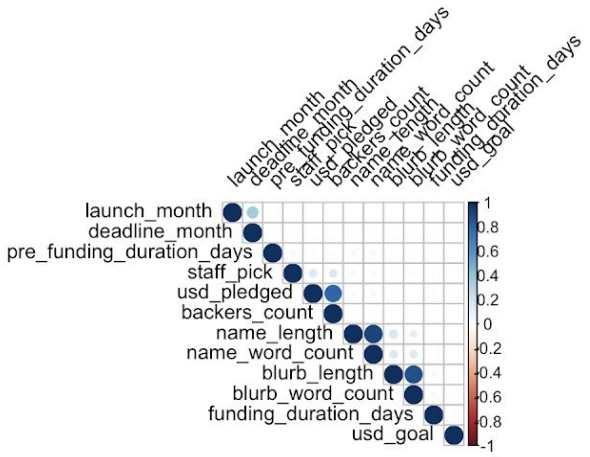

Figure 2.3 Correlation Matrix

A correlation matrix between our numerical categories can be found in Figure 2.3. We found strong correlations between backers count and usd pledged, name length and name word count, and blurb length and blurb word count as we would expect. There is little correlation elsewhere in the matrix.

Model Forecasting

Summary of models and their respective performances are illustrated in Figure 4.1. The precision of constructed models lie the range from about 60% to 70%. This is consistent with prior work regarding the volatile and unpredictable nature of Kickstarter.[4] No known model yields accuracies in the 90% range.

When implementing logistic regression, we achieved an result of 70.71% for training accuracy, and 70.51% for testing accuracy. The similarity between training and testing accuracy implies that we did not overfit our data during training. AUROC of Logistic regression is around 0.969.

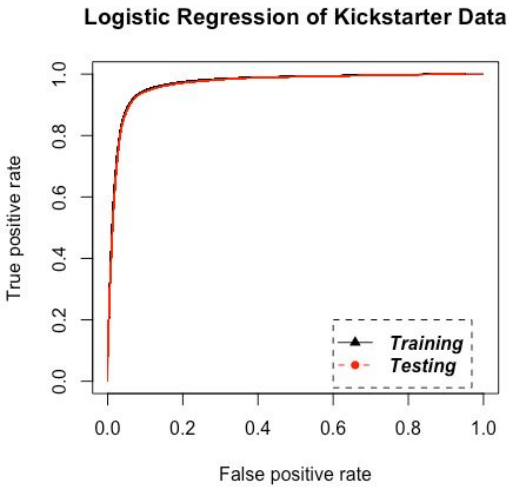

Figure 3.1 AUROC of Logistic Regression

For our model, all the p-values for the variable coefficients are well below 1% with almost every variable having a p-value of less than 0.000001. Furthermore, most of the variables chosen pass the assumption of no multicollinearity. The logistic regression model has captured the trend fairly well, and can yield good predictions for project status.

[4] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5935031/

	Failed *(predicted)*	Successful *(predicted)*
Failed *(actual)*	12732	5865
Successful *(actual)*	9557	24136

Table 3.1: Confusion Matrix for Logistic Regression model

We constructed a model using Random Forests, achieving a training accuracy of 99.98%, and a test accuracy of 74.22%, the best out of all our models.

	Failed *(predicted)*	Successful *(predicted)*
Failed *(actual)*	13560	3584
Successful *(actual)*	9897	25250

Table 3.2 Confusion matrix for Random Forest model

We chose to build the model with 3 random variables, and to grow 500 trees. We suspect that the good performance of the Random Forests model is because Random Forests can take advantage of being invariant to outliers or non-linear relationships, instead only focusing on the clustering within data (of which Kickstarter is a dataset with many separate clusters of different characteristics).

Another feature of the Random Forest method is that along with its model, it also simultaneously generates a measure of the relative importance of different variables.

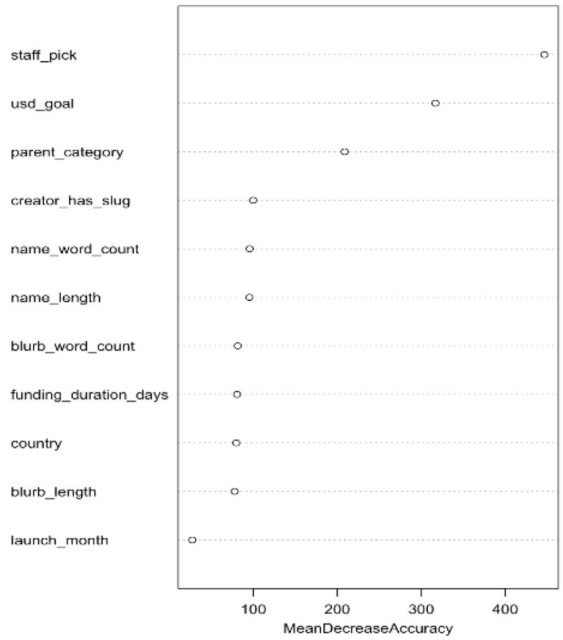

Figure 3.2 Mean Decrease in Accuracy by variable for Random Forest

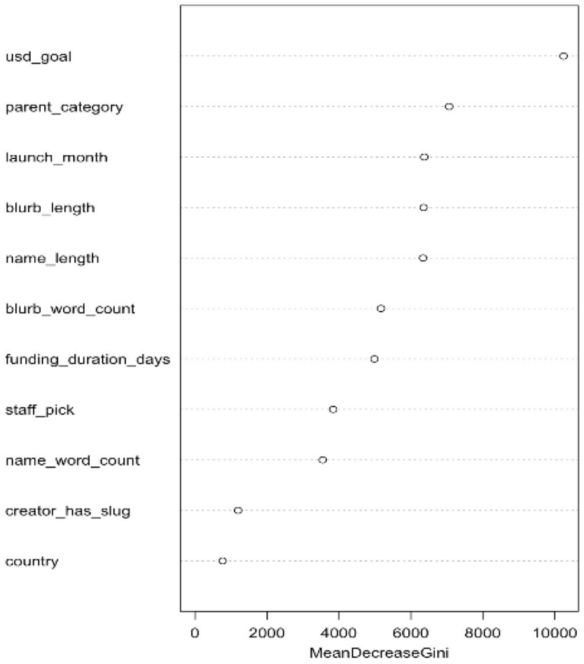

Figure 3.3 Mean Decrease in Gini by variable for Random Forest

From Figure 3.3 and 3.4, it can be observed that the three variables staff_pick, usd_goal, and parent_category are significantly correlated with success rate, while other factors. Interestingly, even though the variable launch_month shows very little decrease in mean accuracy, it actually shows as the third highest variable in terms of mean decrease in Gini.

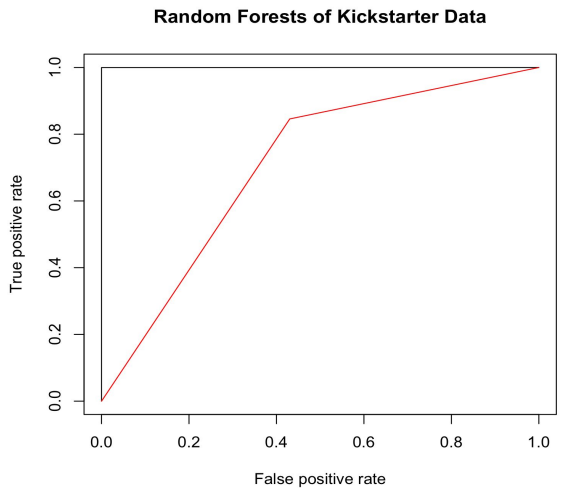

Figure 3.4 AUROC of Random Forests

We constructed a model using Bagging (with decision trees), achieving a training accuracy of 99.99%, and a test accuracy of 72.13%, almost as good as the Random Forest model. We chose to build the model by also growing 500 trees, the same number as used in Random Forest.

	Failed *(predicted)*	Successful *(predicted)*
Failed *(actual)*	13065	5308
Successful *(actual)*	9224	24693

Table 3.3 Confusion matrix for Bagging model

As expected, Bagging methods yields correlated trees, and thus producing models with higher variance than models produced by Random Forest, which explains why the Bagging model performs worse than Random Forest. However, its performance is still competitive with the Random Forest result. Bagging is also capable of generating measures of variable importance, but as the results are generally similar to the Random Forest result, it will not be reproduced.

The Naive Bayes Classifier model has the worst test prediction accuracy of 59.8%; barely higher than the trivial model always predicting successful, with 57.4%. We suspect that the model fails, because of the inherent assumption regarding the independence among variables. In fact, many of the variables, such as blurb_length and blurb_word_count, are highly correlated. Because of its assumption that all variables contribute independently to the probability of project success, the model fails to capture possible correlations between the variables.

	Failed *(predicted)*	Successful *(predicted)*
Failed *(actual)*	1469	20820
Successful *(actual)*	197	29804

Table 3.4 Confusion Matrix for Naive Bayes Classifier model

From the confusion matrix, it is plain that the the Naive Bayes Classifier has essentially learnt a trivial model (by predicting success

for up to 96.8% cases). Ultimately, the Naive Bayes Classifier did not yield strong results.

When implementing SVM on our data, we were somewhat surprised at its relatively weak performance compared to other methods, as the SVM classifier is relatively resistant to overfitting. Its test accuracy is 70.53% with 78748 vectors used.

	Successful *(predicted)*	Failed *(predicted)*
Failed *(actual)*	11742	5151
Successful *(actual)*	10547	24850

Table 3.5 Confusion matrix for SVM model

One possible improvement is to change the SVM kernel from radial to sigmoid or other types. As our data does not require a high dimensional kernel, the classifier might perform better.

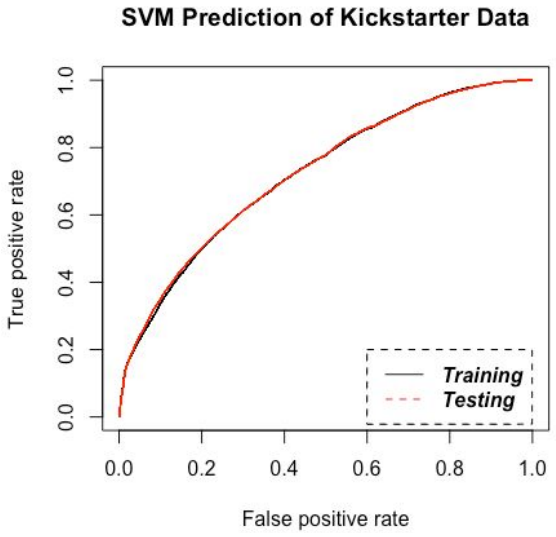

Figure 3.4 AUROC of SVM

Discussion

Summary of Findings

One important finding was identifying the fact that all of the variables we identified did play some sort of role in contributing to the success of a project. In our logistic regression model, all elements had a statistically significant p-value. More importantly, we were able to identify the three variables that played the largest role in predicting success: the project's funding goal, the project category, and whether or not the project was selected as a staff pick.

We found that smaller goal sizes were indicative of a higher chance of success. We also found that being selected as a staff pick also greatly increases chances of success due to the extra attention those projects would receive. Category also played a very important role, as certain categories far outperformed other categories despite the fact that different categories were associated with larger or smaller average goal sizes. For example, crafts and dance both had similar average goal sizes at $8,555 and $6,724 respectively. However, the vast majority dance projects were able to be successful, while more than half of the crafts projects failed. It is likely that different categories appeal to different types of backers.

As for our predictive models, simply by taking a look at the accuracy of our models, we can see that the best performance came from the random forest model, followed by the bagging model. Naive Bayes classification was by far the worst technique, and SVM's performance was quite average.

The reason for Naive Bayes' lackluster performance could be because the model assumes that each of the variables are

conditionally independent. Random forests and bagging were our best performers likely because our training data was very noisy and because our number of observations greatly exceeded the number of features. The large number of observations allowed us to maximize the effectiveness of these models. A summary of our models and their performance can be found in Table 4.1.

Model	AUROC
Logistic Regression	0.9695117
Random Forest	0.7076112
Bagging	0.703873
Naive Bayes	0.5312282
SVM	0.723588

Table 4.1 Model Summaries

Merits, Limitations and Future Directions

Our predictive models were good at looking at the basic project features to come up with a predictive model with roughly 75% accuracy. We had a large dataset spanning many years, so with regards to having a sufficient sample size, there were no concerns.

However, we hope that with further time and resources, we can extend our work by utilizing the large amount of unstructured data available in Kickstarter projects, such as the quality of promotional videos, pledge options offered, and social media interaction of project owners. For example, with regards to the project description, our team solely looked at its word count and character length. Other research has been done using sentiment analysis on the project description to achieve a 4% increase in predictive accuracy.[5] It is abundantly clear that some potential indicators of Kickstarter success requires the use of advanced techniques such as Natural Language Processing to take advantage of.

Future iterations of this research may also leverage deep learning techniques such as neural networks to better leverage unstructured data for a higher predictive accuracy. One must keep in mind, however, that this may result in lower interpretability of which variables are important.

Conclusion

Projects looking to be successful on Kickstarter should be aware of the roles that different categories play – Kickstarter may great to get funding for dance projects, but a bad place for journalism projects. Projects should also consider lowering their Kickstarter funding goals compared to their real funding goals in order to better achieve success, since funding continues until the duration ends even if the goal is reached. Lastly, projects should aim to be engaging and get onto the staff pick list, though the process by which that occurs is not currently transparent.

Ultimately, there is much work to be done regarding predicting the success of a Kickstarter project, primarily revolving about extracting additional features and incorporating them into our models. More research ought to be carried out, particularly regarding analysis of reward schemes and the quality of project advertisement videos and photographs.

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5935031/