KickPredict: Predicting Kickstarter Success

Kevin Chen, Brock Jones, Isaac Kim, Brooklyn Schlamp
Dept. of Computing and Mathematical Sciences
California Institute of Technology
1200 E California Blvd
Pasadena, CA 91126
{kchen2, jonesb, ikim2, bschlamp} @caltech.edu

ABSTRACT

The purpose of this project was to develop a system to predict whether a Kickstarter project will be successful prior to its completion. To do this, we trained a support vector machine on a large amount of project data. This data included properties from the Kickstarter projects themselves, as well as information from external social media sources, such as Youtube and Twitter. On our testing set, our final classifier model was able to successfully predict a project's final outcome with approximately 90% accuracy given the first 40% of the project's data over time; with only initial project features (at "day zero" we were able to predict with 67% accuracy. From our analysis of the data explored, we determined that the most important features in predicting success came from the project properties, not from external media sources. To apply our model to out-of-sample projects, we developed an Android application and a Chrome extension that display a prediction percentage and relevant statistics for any Kickstarter project. do

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Machine Learning, Crowdsourced Funding, Kickstarter, Twitter

1. INTRODUCTION

1.1 Background

With the rise of the internet and the connectivity that it brings, new forms of fundraising have emerged. *Kickstarter* is a popular example of this; it is a website that provides a centralized hub to help people crowdsource funding for their personal projects. Specifically, it enables "creators" to present their project ideas by putting up individual project pages; anyone can then become a "backer" of a project by pledging money. Since its founding in 2009, Kickstarter has helped fund more than 40,000 projects (out of 100,000 created).

Each Kickstarter project consists of a target amount of funding required and a fixed time period for gathering the given amount. If the target amount is reached within the specified time, the project is considered a success, and the owner receives all of the pledged funds. Otherwise, the project is unsuccessful and the owner does not receive *any* of the pledged amount. According to the Kickstarter founders, this "protects everyone involved" by making pledging low-risk and ensuring that project creators do not have to work with insufficient funding. For our purposes, this gives a clear cut definition of success or failure; this allows us to apply a standard binary classification algorithm.

As described, projects have individual pages on Kickstarter; each project page allows a creator to communicate to their potential backers by writing a detailed description of the expected project product. The creator can also specify the different "tiers" of pledges and their corresponding rewards. In general, higher pledges provide increasing incentives. For example, an project might reward a backer with a tshirt or sticker for a $10 pledge, but promise the full product for a higher pledge.

In addition to details about the project, Kickstarter pages also show information about the creators themselves. This includes statistics such as how many projects they have previously created, how many other projects they have backed, if they are connected on facebook, etc.

Finally, in order to increase the chances of their project getting funded, creators often organize campaigns across various social media platforms. Many projects pages include videos (hosted on Youtube or internally on Kickstarter) that explain the project in more detail. All Kickstarters also have a url that can be retweeted and shared on Twitter, as well as liked and shared on Facebook.

1.2 Motivation and Goals

Kickstarter is an extremely popular service, and the number of projects proposed and funded daily is steadily increasing. As of writing, Kickstarter has overseen the transaction of over $569 million pledged funds. Clearly, crowd-sourced funding is an extremely effective and valuable approach to fundraising.

Although Kickstarter does much to limit the risk of failure (by using an "all-or-none" model), a project that does not succeed is detrimental to both the creator and its backers. Although creators are not held accountable for failed projects, they must invest time and effort to promote their ideas. This often involves creating prototypes of the project product, which be a nontrivial out-of-pocket expense for the creator. Even backers lose for unsuccessful projects; while they don't actually lose any pledged money, they waste time following and sharing a potential project.

Thus, predicting whether a project will succeed or not with a better than average accuracy can be extremely useful for both creators and backers. It can potentially save both creators and backers from wasting money on project ideas that most likely won't be funded; in addition, it can motivate and inspire creators/backers to work on projects with the most potential for success.

Since the line between successful and unsuccesful for Kickstarter projects is clearly defined, we decided to work with support vector machines to predict the success of future projects and to analyze the data. Support vector machines try to classify a set of n-dimensional points into different categories. Points with a known accuate classification are used to train the svm, which can then be used to predict the classification for future points. In our case this means we should be able to predict whether a project will be successful or not based on some set of features.

While an accurate boolean prediction would no doubt be useful for many people, our primary goal is to understand what specific factors lead to such a prediction. We are interested in which features or properties of a project are the most important in contributing to its success; in particular, we hope to look at the role of social media in a project's success or failure. This information is arguably much more valueble than a simple boolean prediction value for project creators. Since Kickstarter projects are a very personal commitment, it is expected that creators are unlikely to abandon their ideas if they know it has a small percentage of success. Knowing which features correlate with success can suggest ways for creators to potentially boost the potential of a project. For example, perhaps active usage of Youtube or Twitter to promote projects may correlate to success.

2. IMPLEMENTATION

The implementation of our project was broken up into three major steps. First, we attempted to gather as much data as we could on projects. Next, we used a standard binary classifier algorithm to identify which features were the most important and to generate prediction model. Finally, we developed an Android application and Chrome extension to apply our model to new Kickstarter projects.

2.1 Data Collection

As described above, we explored a large number of different features from Kickstarter projects and their related data in order to determine which were most important to a project's success. Specifically, we looked at the following fields:

- The amount pledged over time

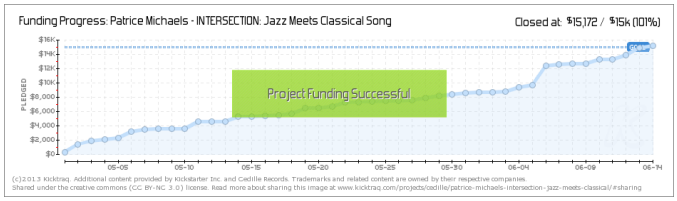

Figure 1: Kicktraq plot of pledge data over time.

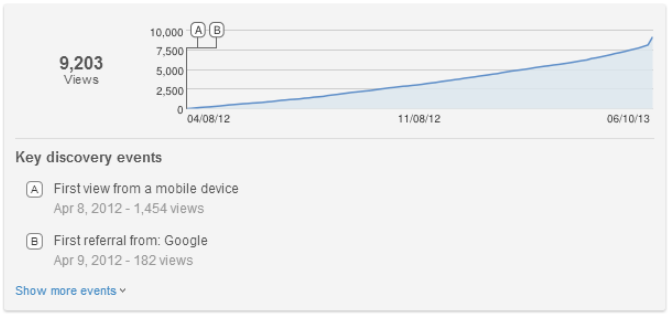

Figure 2: Youtube video views over time.

- The number of projects backed by the creator
- The total number of projects created by the creator
- Whether or not the creator is connected on Facebook
- The goal pledge amount
- The length of the project
- The number of images present in the project page
- The number of characters in the project description
- The number of pledge tiers
- The minimum and maximum pledge tiers
- Whether or not the project page has a video
- Whether or not the project page has a Youtube video
- If a Youtube video is present, the view count of the video over time
- The number of times the project Twitter link was tweeted

To get many of the project properties (such as goal pledge amount and length of project), we parsed the Kickstarter project page directly. Unfortunately, while the page contains the current pledge amount, it was not possible to get the historic pledge data directly from Kickstarter. Initially, we created a Python script to continuously scrape pages to get pledge data over time (approximately one every 2 hours). However, since we would not be able to train on the data until the projects were completed, this approach was not ideal, especially because of time constraints. Luckily, we discovered Kicktraq, a website that collects Kickstarter statistics daily. While Kicktraq unfortunately did not provide the pledge data in plain text, they did provide image graphs of the data, as shown in **Figure 1**. Although it was slightly cumbersome, we were able to get the data from the images based on measuring pixel distances. Although not 100% accurate, the estimated data was assumed to be good enough for our purposes (we believe that small fractional discrepencies in pledge data are not deciding factors in prediction).

In addition to this Kickstarter data, we used data from Youtube and Twitter. In order to gather Youtube data, we first parsed through the Kickstarter html page for each project. We looked for any urls containing *www.youtube.com* which did not contain the */user* tag in order to see if a project used Youtube videos in its description. Once we had a list of youtube urls and their associated kickstarter project, we went to each youtube url and downloaded the statistics over time from that page, as shown in **Figure 2**. This involved sending a POST request to *http://www.youtube.com/insight* with the video ID and session token from the main video page; the response consisted of Google Charts data that we were able to parse. Some of our links were broken and some did not allow statistics to be publicly viewed. Generally, we were able to get videos with statistics for approximately 10% of our initial projects.

We based our Twitter data on how many times the link to each project's Kickstarter page was tweeted. We recorded data from 20,000 completed projects and monitored the number of tweets 1000 projects received each day for a 30 day period. We then took the results from the 30 day study and grouped the projects by the final number of tweets each url received. We then determined the rate of change for each of those groups and applied it to already completed projects in an attempt to get around the fact that no historical data was unavailable from Twitter. In order to gather data we used the Get URL method of the Twitter v1.0 API and the Get Search/tweets method of the Twitter v1.1 API for the projects that were monitored for thirty days.

2.2 Classification

We collected the data described above for 20,000 completed Kickstarter projects in order to train Support Vector Machines. Specifically, 1000 were set aside for testing, and the rest were used as a training set.

Both our data mining code and SVM code were written in Python. For the SVM, we used a machine learning library for Python called scikit learn [1]. Scikit learn provides an easy implementation for various types of support vector machines. We used an SVM with a radial basis function kernel, which allowed for more precise classifications.

To accommodate temporal data (such as the pledge amount over time), we binned the data into a histogram and then used subsections of the histogram into different SVMs. For example, we binned pledge data into 20 bins over the length of the project duration (so the 1st bin included the first 5% of the pledge data). Then we made 20 SVMs, with the i-th SVM looking at the first i bins.

To find the best set of features to use, we created an SVM for every possible subset on the set of all fields we explored. As described, we used 1000 projects as our testing set in order to compare the accuracies of the different SVMs. By calculating the percentage of test projects each SVM was able to predict correctly, we were able to determine which features were the most useful for prediction.

3. RESULTS

From the analysis of all subsets of features, we discovered that most features actually contributed very little or not

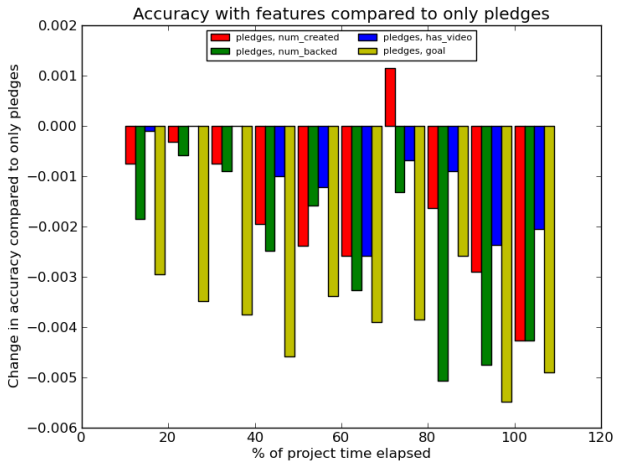

Figure 3: As is evident in the graph, pledge data with any other type of data reduces accuracy (with the exception of one fluke at 70% of time progressed).

at all to the accuracy of the SVM, as shown in **Figure 1**. The most important feature was the pledge data over time. Although some features did end up improving accuracy for certain bins, the amounts they contributed were less than a tenth of a percent; thus, we decided to only use the pledge data for our final SVM models.

Although pledge data was clearly the major contributing factor for predicting project success, we were also interested in prediction based only on day 0 starting conditions (that is, without any temporal data such as the pledge data). In order to do this, we tested all subsets of non-temporal features. We found that the most important ones were:

- The number of projects backed by the creator
- The total number of projects created by the creator
- Whether or not the project page has a video
- The total goal required for the project to receive funding

This combination of fields gave us an SVM that was 67% accurate on the training set.

To apply our model to out-of-sample projects, we ported our testing code to run on Heroku, a web application platform. The web application provides a JSON API for predicting a given project's success; it will dynamically fetch the project's data and use the appropriate SVM (based on the project's progress) to estimate the percentage of success.

To visualize the prediction, we also developed an Android application and a Chrome extension as seen in **Figure 3**. The android application allows searching for Kickstarter projects and uses the web API to display a simple prediction percentage. The Chrome extension integrates directly with the Kickstarter website; it displays an info box with relevant project data and prediction results on any project webpage.

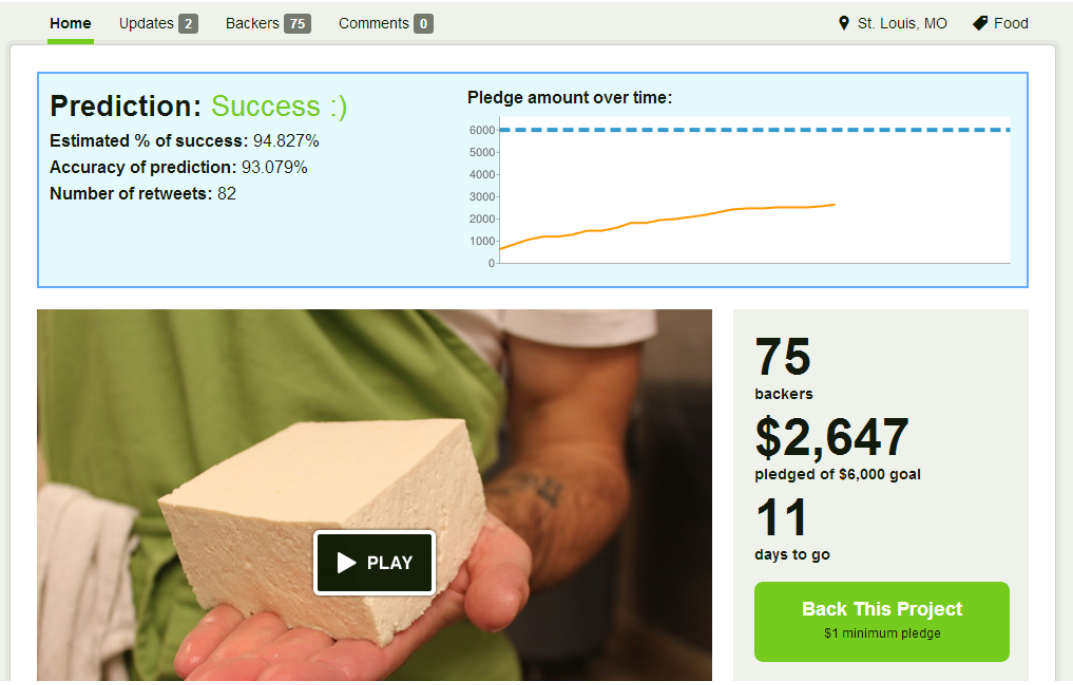

Figure 7: The prediction box is injected into Kickstarter project pages via a Chrome extension.

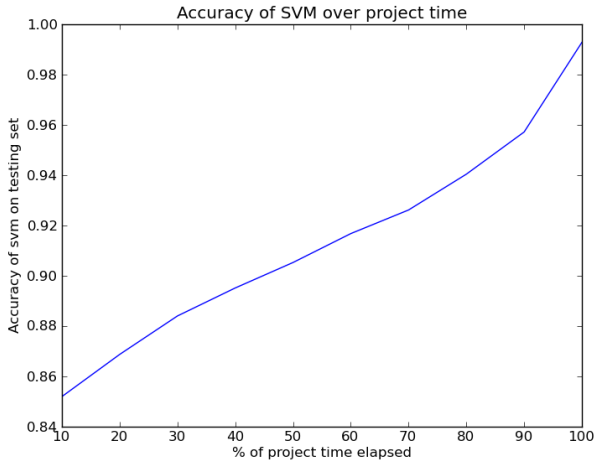

Figure 4: Accuracy of SVM on the training set vs. % project time elapsed.

4. DISCUSSION

Initially, we expected external data sources such as Youtube and Twitter to provide much more impact on predicting a project's success. We thought that projects that were advertised better through such media would receive more attention and therefore be more likely to succeed. We came to these conclusions by looking at some elementary statistics on the data **Figure 5**. These plots revealed that more successful projects have a video, and more unsuccessful projects don't have videos. Likewise it also told us that the length of the project would not play too much impact as both successful and unsuccessful videos had roughly the same distribution of Project lengths. Kickstarter itself also did some analysis on completed projects, suggesting that the project duration might make some impact on how successful a project will be [2].

However, this ended up not being the case. Youtube data did not help much with our prediction rates. Looking at about 20000 projects only supplied about 5000 videos, of which only 1000 had visible statistics. This indicates that maybe the quality of video is more important than whether the video exists. We then ran analysis using the statistics on SVM and found that the view count only made a marginal difference and did not really improve our accuracy at all.

Similarly, that data from Twitter had a small but ultimately insignificant impact on the quality of the SVM. The number of Tweets a project received was not strongly tied to its success or failure. If a project link was tweeted three or less times, it was more likely to have failed but beyond that, there was little significance to the data from Twitter.

Thus, our results lead us to conclude that the day-to-day pledge data incorporates most, if not all, of the impact of all

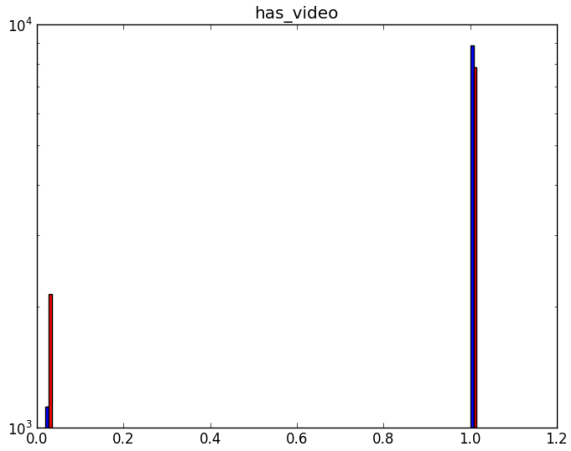

Figure 5: The distributions of whether a project has a video or not (1 corresponds to having a video, 0 to no video). Blue represents successful projects, red represents unsuccessful projects.

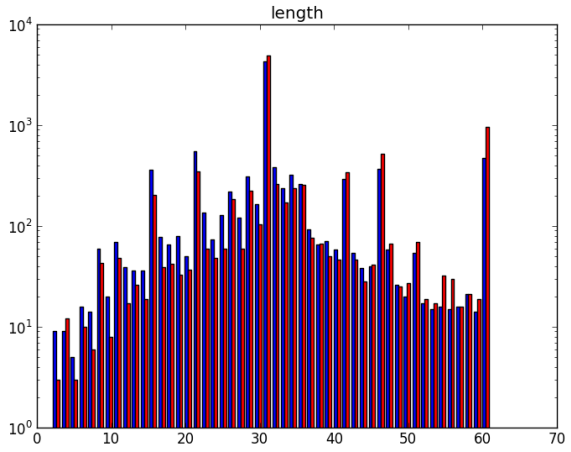

Figure 6: The distributions of project length. Blue represents successful projects, red represents unsuccessful projects.

other fields; that is, the pledge data is tightly correlated with any other predictors of success. For example, if a project is well advertised through Youtube, then people will pledge towards the project. Therefore, if Youtube does make an impact on the likelihood that a project will succeed, then so will pledge data. Another simple example is the field of whether or not a project has a video compared with whether or not a project has a Youtube video. Clearly, the field concerned with videos in general will incorporate the impact of the field concerned with just Youtube videos. We suspect this is the case in general for pledge data.

When looking at the fields relevant at day zero, we find that whether or not a project has a video does play an impact. This analysis is in line with what the basic histograms tell us about the distribution of successful and unsuccessful videos. We can interpret the importance of whether or not a project has a video as being related to how well advertised the project is at day zero.

We can interpret the day zero fields that are concerned with the project creator as being related to the reliability or reputation of the creator. Since the Kickstarter service has no way of ensuring that the project creators actually use their funds in the way advertised, backers must be able to trust the creator enough to follow through on his word. Thus, the number of projects backed by the creator and the number of projects created by the creator are likely related to the creator's trustworthiness. A video may also play a similar role by facilitating a connection between the backers and the creator, perhaps making him appear more trustworthy.

Finally, the project goal would be important at day zero because a project with a highly ambitious or undeserved goal in the eyes of backers will not be likely to succeed, while a project with a very low goal will meet it fairly quickly. To highlight this fact, consider the project "Penny Arcade Podcast, Downloadable Content: The Return." This project has a goal of a mere ten dollars. It is undoubtable that this project will receive it's funding. In fact, this project probably did not need to go through Kickstarter at all. Rather, it is using the power of crowd sourcing to market and gauge the interest of it's intended audience.

5. FURTHER WORK

Fully investigating every aspect that could affect the success or failure of a Kickstarter project wasn't feasible in the time period of the study. More social media sites (e.g. Facebook, Reddit) could potentially be investigated with additional time. Determining whether a project is shared on one of those sites affects success could provide valuable insight to future project creators about where to focus their online campaigning.

Specifically, future work could look at data from Facebook. Facebook has recently implemented a hashtag feature. Perhaps examining whether a project's page is receiving attention or whether it's hashtag is being posted will point to it's popularity among social networks. However, when looking at these external media for information, it may be worthwhile just to look at the impact of these outside sources, i.e. without pledge data itself. This may provide some insight about how well a project is advertised. In addition, this may

shed more light on the relationship between outside sources and pledge data.

A study could also attempt to determine whether the text in a projects various sections (Body, Risks and Challenges, FAQ) can be related to the success of the project. Our attempts didn't pan out as we started too late and lacked a working knowledge of text analysis. A study could also attempt to determine sentiment in the comments on the project's page and relate that to success.

Another possibility is in improving the metric used for our Twitter analysis. In our work, we only followed 1000 projects via twitter and performed basic linear regression on the projects to come up with a distribution. This may be expanded both in scale and in implementation. For example, [3] provides a method of trend analysis that may be useful in determining if the number of times a project has been tweeted will trend. Another simpler approach would be just to use a more sophisticated parametrized curve to fit the data instead of linear regression.

6. CONCLUSIONS

With this research we can accurately predict the success of a Kickstarter project. Perhaps the most interesting results of our research is that external media sources, Youtube and Twitter, did not have an impact on a Kickstarter project's success. While this is fairly accurate now, we hope to improve upon it in the future by using text analysis or checking other external media sites.

7. ACKNOWLEDGMENTS

We would like to thank Prof. K. Mani Chandy for his support and guidance as our mentor during CS145. We would also like to thank Professor Adam Wierman and Steven Low for helping to run the course.

8. REFERENCES

[1] Scikit learn reference. http://scikit-learn.org/stable/modules/classes.html.

[2] F. Benenson and Y. Strickler. Trends in pricing and duration. http://www.kickstarter.com/blog/trends-in-pricing-and-duration.

[3] S. Nikolov. Trend or no trend: A novel nonparametric method for classifying time series. http://web.mit.edu/snikolov/Public/trend.pdf.