**Kickstarter Campaign - From an idea to fully funded reality**

Nicole Fimmen

December 2017

**1. Introduction - Why Kickstarter?**

Kickstarer's mission is to help bring creative projects to life. Since launching in 2009, 14 million people have backed a project, $3.4 billion has been pledged, and 134,135 projects have been successfully funded. Projects have many forms including food and publishing to music, fashion and technology.

Kickstarter is all or nothing. If a project doesn't get every dollar, the project will not be successful. Kickstarter released their data on Kaggle with a challenge, to predict if a project will get successfully funded or not.

**2. Kickstarter Dataset**

To predict whether a project will get funded or not, I am going to look at the Kaggle data set. The data contains the product description, goal amount, various project timelines and whether a project gets funded or not. In total there are 108,129 project from May 2009 - May 2015. For this project, I will focus on projects that use US currency.

Variable Description

- project_id: unique id of project
- name: name of the project
- desc: description of project
- goal: the goal (amount) required for the project
- keywords: keywords which describe project
- disable communication: whether the project authors has disabled communication option with people donating to the project
- country: country of project author
- currency: currency in which goal (amount) is required
- deadline: till this date the goal must be achieved (in unix time format)
- state_changed_at: at this time the project status changed. Status could be successful, failed, suspended, cancelled etc. (in unix time format)
- created_at: at this time the project was posted on the website(in unix time format)
- launched_at: at this time the project went live on the website(in unix time format)
- backers_count: number of people who backed the project
- final_status: whether the project got successfully funded (target variable – 1,0)

**3. Data Wrangling**

First I downloaded the data, train.csv, from Kaggle and uploaded the file into RStudio. Next I subsetted the data to only use projects with the US currency, which is still 85% of the total projects.

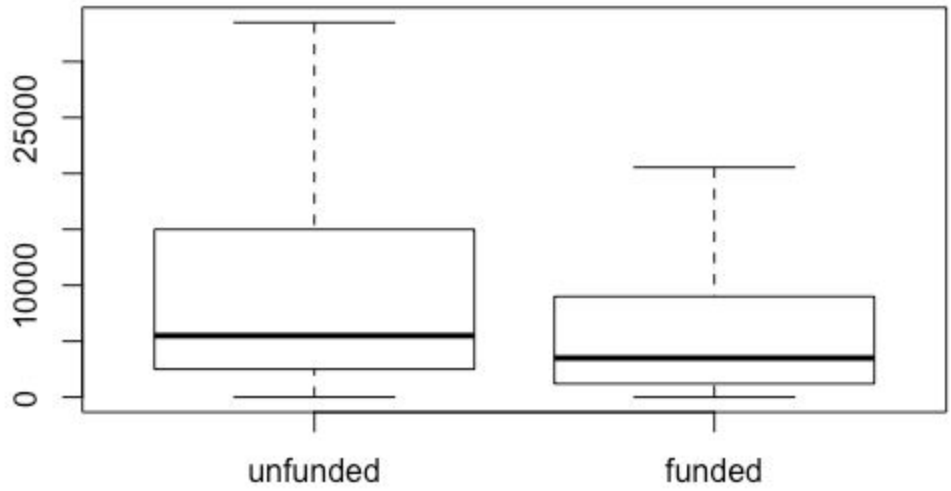
I analyzed over 92,000 projects. Of the projects, 33% were funded.

Next I took the time variables that were in unix and extracted the date and time. These variables included deadline, state_changed_at, created_at and launched_at. From these I created new variables to get a better understanding of how long the projects lasted and whether the start or end time had an impact. I also grouped the launched hour and deadline hour by morning, afternoon, night and late night.

I then turned the numeric final status with 0s and 1s to a factor with unfunded and funded.

*Limitations*
I do not have the categories of the projects. That would have been helpful to further understand the different projects by category. Also knowing exposure on the website could be helpful. What factors contribute to a project being at the top of the page versus lower on the page requiring “load more” clicks?

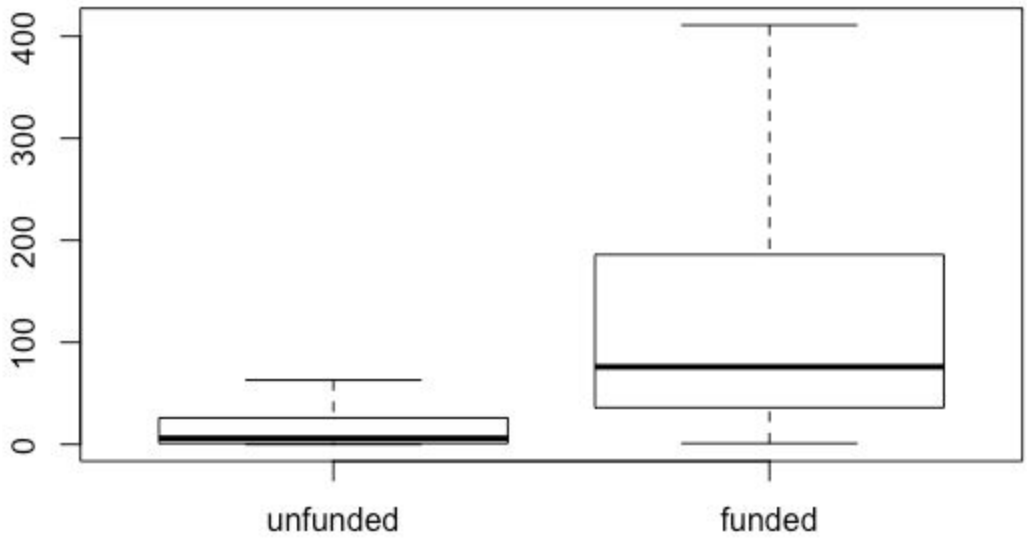
**4. Exploring the Data**
Goal Amount
Projects that got funded overall asked for less money and had a smaller range. The interquartile range was 7,800 for funded versus 12,500 for unfunded.




Disable Communication
For whatever reason the project owner can disable their communication. This obviously sounds like an action that would hinder a project from getting funded. So this option was rarely used and less than 1% of projects opted to do this.

Backers

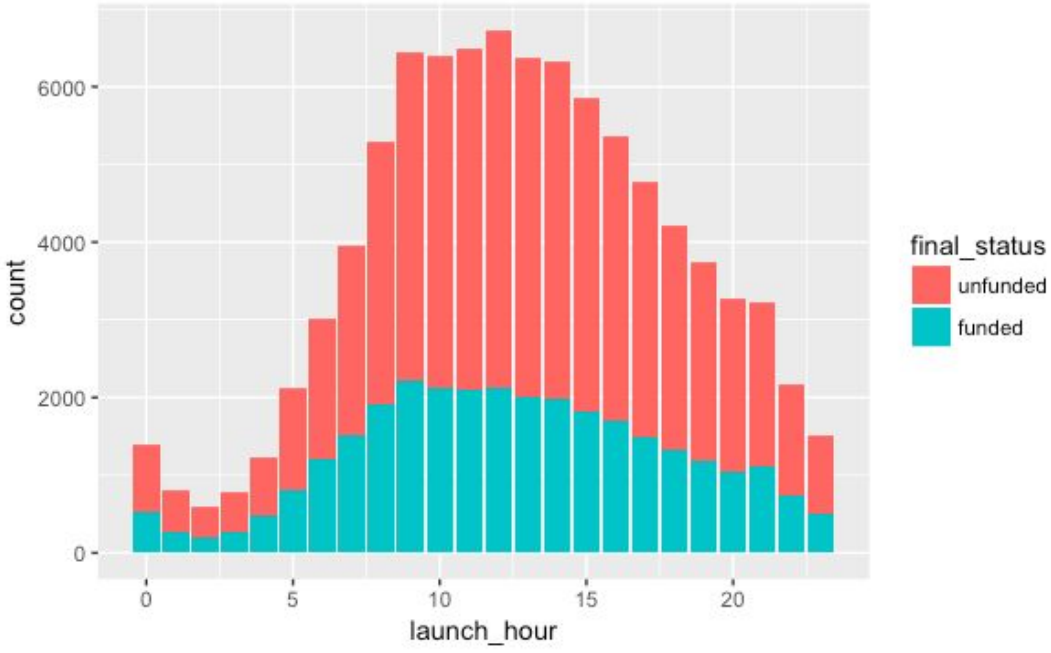
No surprises here. Funded projects have significantly more backers. The median for funded projects is 76, while unfunded backers is only 6. Additionally the interquartile range for funded is 150 versus 25 for unfunded.




**Time Variables**

Launched

For the launch, I broke out the times by hour, day and month. By hour, unfunded and funded have a similar pattern. Launch hour popularity begins to climb in the morning and peaks at 9am. Project launches stay fairly steady until 3pm where they start to decline.



The hours are also grouped into 4 sections:

- Morning: 5-10am
- Afternoon: 11am-4pm
- Night: 5-10pm
- Late night: 11pm-4am

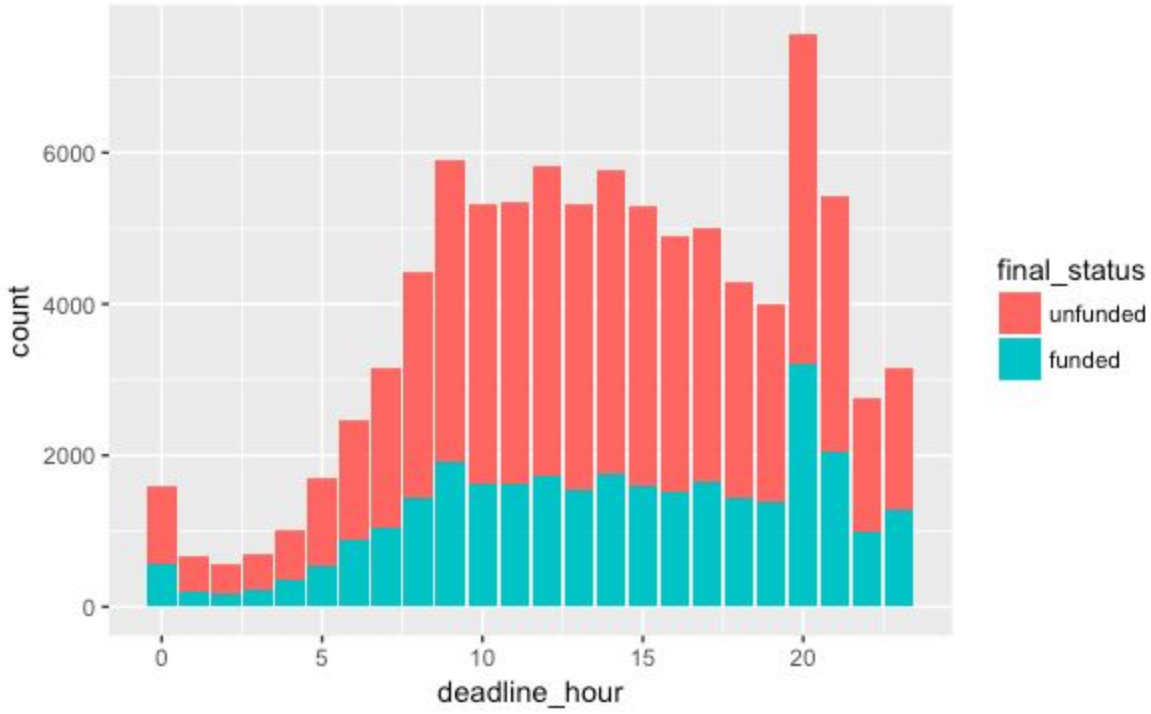
The afternoon grouping has the most launched projects.

By day, Saturday and Sunday are the least likely days to launch a project. Monday and Tuesday are the most popular days to launch a project. Wednesday, Thursday and Friday are all fairly even.

For the month, March and April are the most popular months to launch a project and December is the least popular.

**Deadline**

For the deadline variables, I broke them out the same as with launched. The deadline hour is fairly steady from early morning to the evening. Eight p.m. is the most common time for a deadline. No surprise the late night hours aren't as popular of times for a deadline.

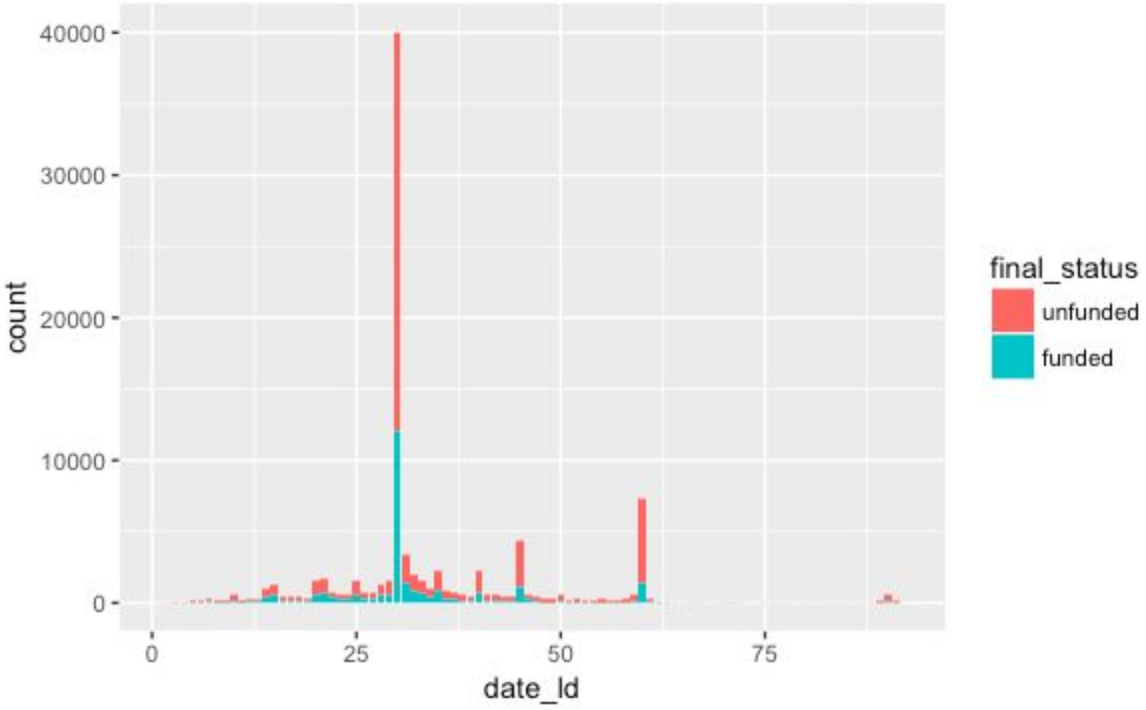



The deadline hours are grouped like the launched hours. The night section has the highest number of deadline times for funded. Unfunded has the most deadlines in the afternoon.

For deadlines by day, Tuesday has the least amount of project deadlines. Thursday - Saturday has the highest amount of project deadlines.

January and February are the least popular months to have a project deadlines. March - May are the peak months with the most project deadlines.

**Launch to Deadline**
Thirty days is the most popular time length for a project. Additionally there is a bump at 60 days. Sixty days became the max amount of days a project could run, so that probably contributes to the bump.




Overall observations from these variables show that funded projects typically have a shorter length, ask for less money and have significantly more backers.

Next I explored the description and keywords variables. I wanted to see if more or less words would have an impact. There was not a significant difference between the amount of words and funding status. I also extracted the words “I” and “we” from the description. I wanted to see if using we versus I would have a better chance at getting funded. Again nothing significant.

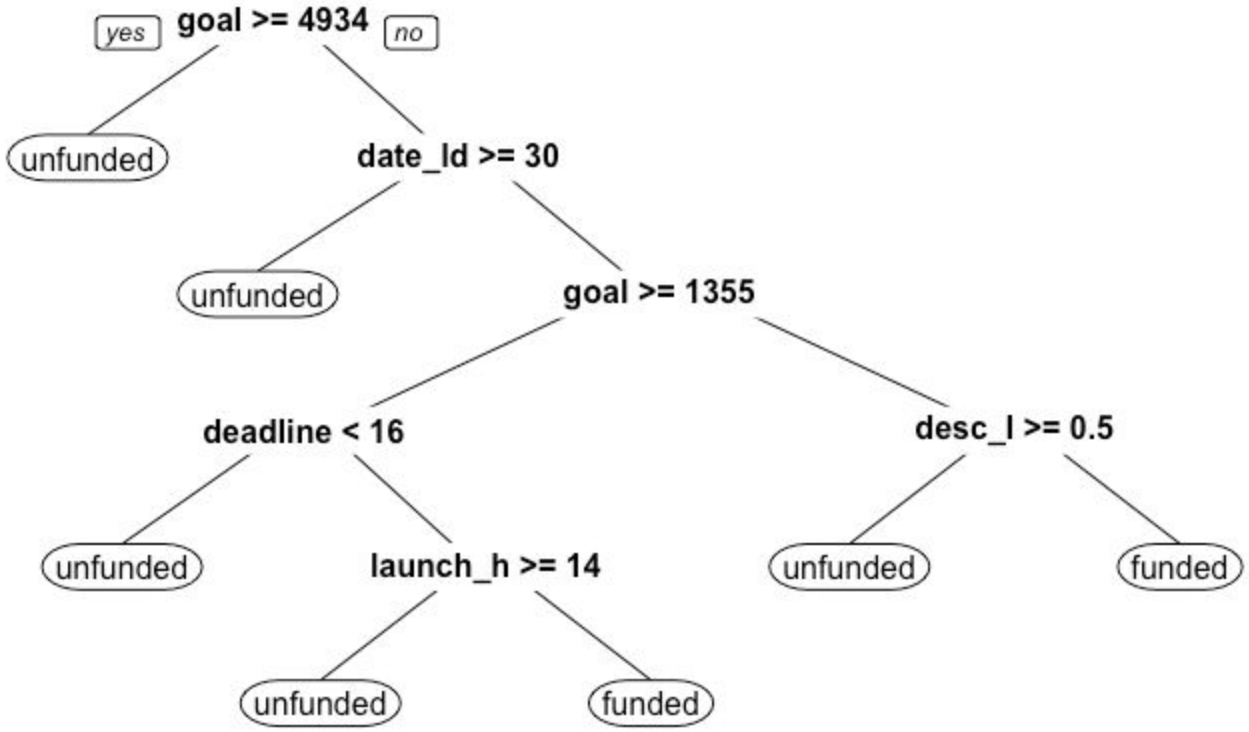
I also extracted the word “help” from the description. I was thinking potentially this active word would be more convincing in helping projects get funded. However this was not the case.

**5. Predictions**
For the next step I divided the data into a train and test set with 75% to the train and 25% for the test. I used the function nrow to get the total number of rows and multiplied that number by 75%. With this I created a random variable and divided that into the test and train data sets.

**Approach**
For this analysis I used CART, Random Forest and logistic regression.

CART

First I used the tree method CART. I applied rpart to all the variables. I used all the variables and set the complexity parameter to 0.001.  Next I picked the tree size that minimizes the misclassification rate. Then pruned the tree with the best complexity parameter. This model is used on the test data. I used the confusion matrix to see how the model performed on the test data. The accuracy was 67.7%.




Random Forest

Next for Random Forest I only used the variables that the CART tree had decided to split on. Those variables included goal, date_ld, deadline hour, launch hour and desc_l. Since the CART model did ok, I knew Random Forest wouldn't perform any better. The accuracy on the test data was 67.7%.

Logistic Regression

I also tried logistic regression to see how this model would perform. I used similar variables including goal, date_ld, deadline hour, launch hour and desc_l.  I tried a few logistic regressions with various variables and this was the best one. The accuracy on the test data was 67.6%.

6. Results

After working through the variables and trying different models, I have learned that further work is needed. For each model I used the confusion matrix with the model and test data. Here is the accuracy of each:

- CART: 0.677
- Random Forest: 0.677
- Logistic regression: 0.676

The models had almost the same accuracy. Although the models are not accurate enough to use on future projects, I do know a few important variables that are important for a project to get funded.

The variables goal amount, date_ld (days between launched and deadline), deadline hour, launch hour and the use of the word “I” had the most impact. The most obvious influencer in whether a project gets funded or not is backers count. The average backer count for funded projects was 76, as opposed to 6 for unfunded. Also asking for too much money had a negative impact on funding. The range of funding asked for for funded was $3,500 vs $5,500 for unfunded. Projects that had a 30 day window were also more likely to get funded.

7. Future Work

What I learned is that time related variables did not have as much of an impact as I had expected. The key to predicting the projects most likely is in the description. Further work is needed to analyze the project description. I would further explore specific keywords and try sentiment analysis. It is possible the tone has an effect.

**Appendix**

R code