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.

For this project we are going to determine what factors will predict whether a project gets funded or not. My audience is people wanting to start their own Kickstarter campaign. My project will focus on what will help boost my audiences' chance of success.

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.

Because the data was from a Kaggle competition, the data set was fairly clean. There were no odd or missing values. Most of the clean up was time related. 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. Additionally I converted all the times variables to the same time zone, Pacific Time.

For the project description and keywords I also pulled out the word count for each.

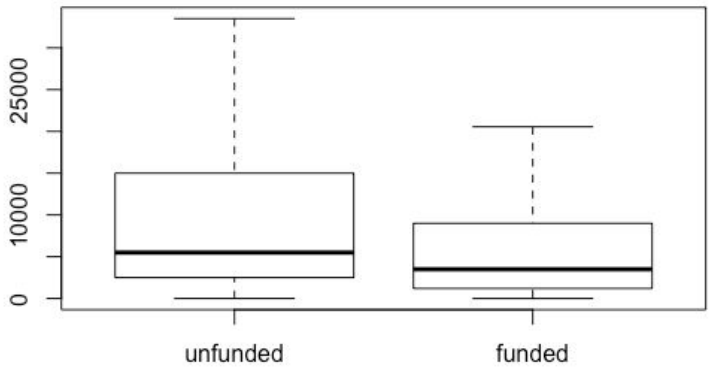
I then turned the numeric final status with 0s and 1s to a factor with unfunded and funded.

Limitations
I do not have the categories of the projects. That would have been helpful to further understand the different projects by category.

I also think having data on the projects' website exposure would have enhanced the predictability of the projects. Some projects show up on the first section in their categories while others take several clicks to find. The Kickstarter homepage rotates which category is shown first. It would be helpful to know how this is determined.

You can also click into each category within the top horizontal navigation. In each category there is a featured product, new and noteworthy, almost there and popular. Or you can click to view all and see all projects within a category. Additionally there are recommendations lower on the page that include all categories. A person can also "explore" projects and filter down by various curated lists. Or do a normal search. It would have been helpful to know what classifies a project as featured product, new and noteworthy or popular. I think including this data would have enhanced the funding status predictability.

4. Exploring the Data
Goal Amount
I assumed that projects that got funded asked for less money. Less money maybe means easier to reach the goal. Looking at the data I see the interquartile range was 7,800 for funded versus 12,500 for unfunded. Funded projects have an average goal of $3,500, while unfunded has $5,500. I used a t-test to determine if the difference in means is significant. The p-value was very low <0.0001 and rejects the null hypothesis. The null hypothesis here states there is no difference in the mean of goal amount of funded versus unfunded. Projects that got funded overall asked for less money and had a smaller range.



Disable Communication

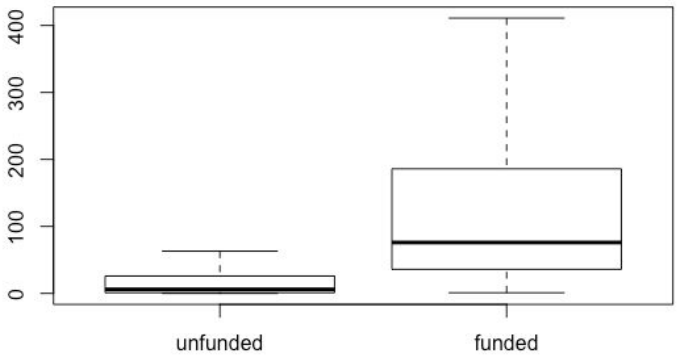
For whatever reason the project owner can disable their communication. My hypothesis is that disabling your communication will have a negative effect on funding. This obviously sounds like an action that would hinder a project from getting funded. So this option was rarely used and less than 1% of projects opted to do this. Of those projects that opted to disable their communication, none were fully funded. As this is a categorical variable, I did a chi-squared test and got a very low p-value of <0.0001. With the low value, I can conclude that the association between the variables is statistically significant. However disabling communication was rarely done. It may be an influential variable but isn't very important as less than 1% of the projects were involved in disabling communication.

Backers

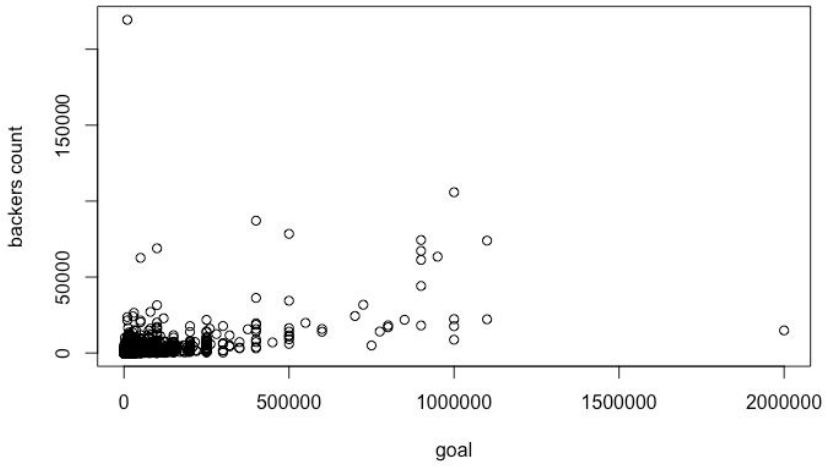
No surprises here. Funded projects have significantly more backers as more backers giving money, the more likely a project is to get funded. Although this variable was provided, it is not a factor we can use to predict whether a project gets funded or not. We do not know the amount of backers until after a project has been fully funded. I explored the variable backers anyways to see if there was anything to learn about the number of backers that could help my client.

We do have a few interesting facts on the backers. Eightyeight projects had over 10,000 backers.

However we do know that funded projects have significantly more backers than unfunded projects. 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.

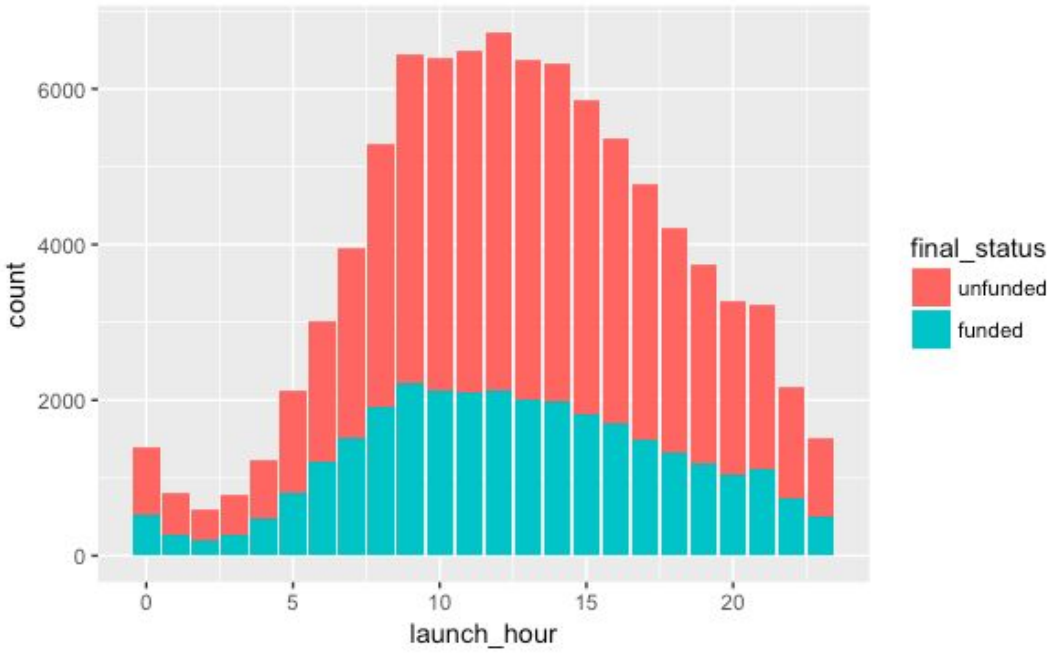


It is also hard to determine how much backers give per project. With a scatter plot for funded with backers versus goal amount, there is no clear line. Since we do not have this and only the over all backers number, using backer count is not helpful as it is a variable we only know after a project is fully funded.



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

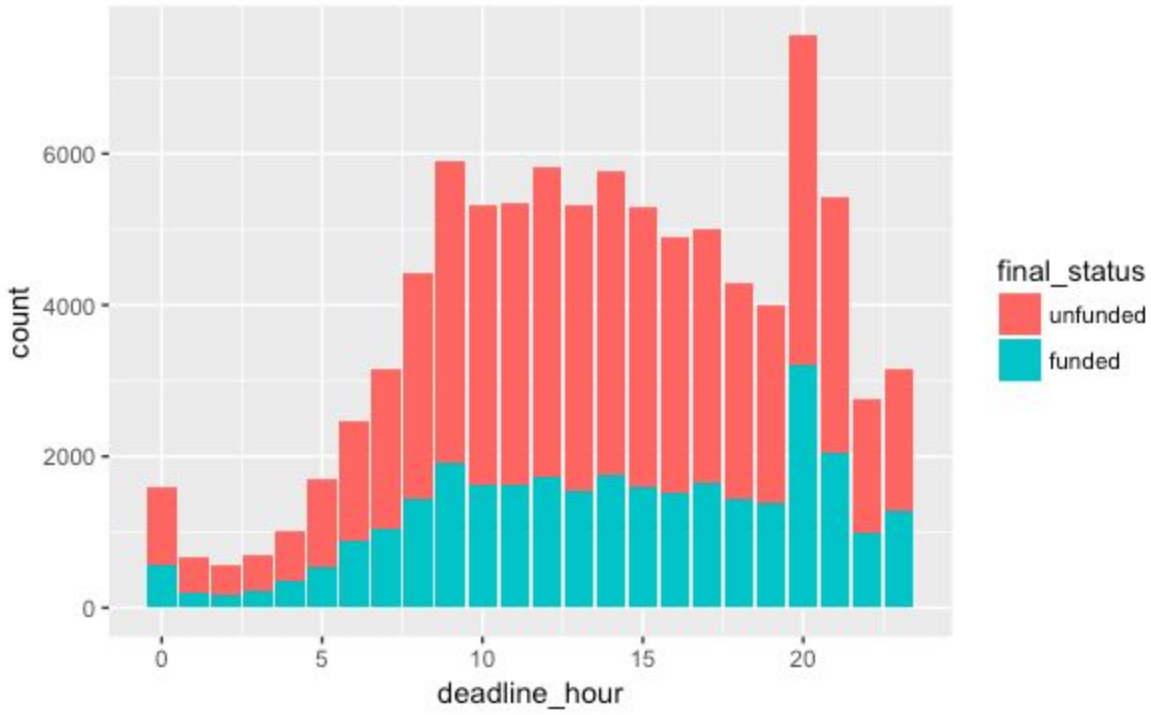
My hypothesis is that the time a project launches will impact funding status. I'm thinking if a project launches during midday, it will get more exposure. In the afternoon people are probably not commuting, not the early mornings of needing to get the initial workday started and maybe people are looking for a break. Although we do not know the intent, we do know that the afternoon grouping has the most launched projects. I did a chi-squared test (4x2) on the four time groupings and final status to see if the time groups affected the project status. With a very low p-value <0.0001, I can see that the association between the variables is significant.

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.



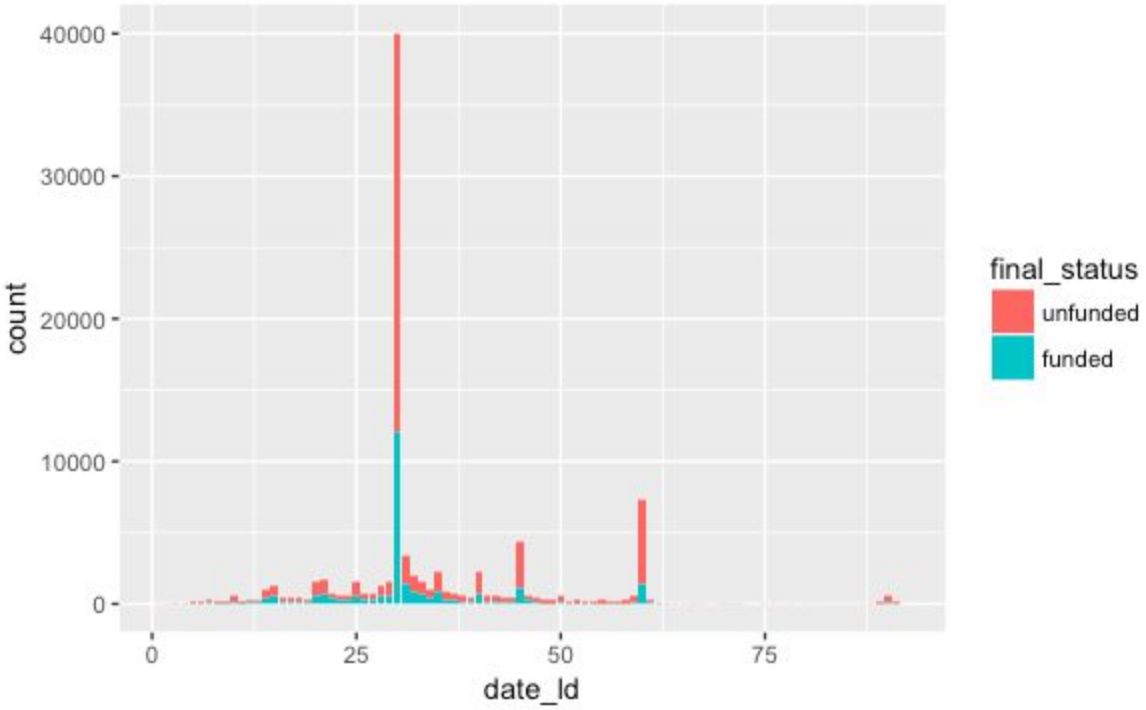
The deadline hours are grouped like the launched hours. Again I assumed deadline hour would affect funding. Projects that have a deadline after common worker hours have better funding success. The night section has the highest number of deadline times for funded. Unfunded has the most deadlines in the afternoon. Again I did the chi-squared test on the four time period and final status and had the same low p-value <0.0001, meaning the deadline time is an important variable.

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. The mean for funded is 32.5 days and the mean of unfunded is 35.3. To see if the time length mean of the projects had an impact, I did a t-test and got another low p-value <0.0001. This low value means the time from launch to deadline impacted whether a project got funded or not.



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 was thinking a shorter description, less words would increase the likelihood of a project being funded. For the description count, the p-value was 0.8601. The number was too high and so description word count had no impact. However the keyword count did have an impact. The keyword count for funded had a mean of 5.75 words and median of 6. Unfunded had a mean of 5.37 words and a median of 5. This difference in means was significant using a t-test and the p-value was <0.0001.

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. I did the chi-squared test for “I” and “we” by final status.

Twenty four percent of the projects used the word “I.” Of those projects, 30% got funded. With the chi-squared test I got a low p-value, $p < 0001$ indication the association between the variables is significant.

Around 9% of the projects used the word “we.” Of these projects, 34% got funded. For “we” the chi-squared test showed the association between the variables was not significant, $p = 0.1713$. The p-value was considered high as it was above 0.05 significant level. For the projects using “we,” I thought “we” would be more impactful, I was wrong.

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. Only 7% of the projects used the word “help” and of those 36% got funded. Using the chi-squared test on “help,” the low p-value, $p < 0001$, implies an association between the variables is significant. So “help” did have an impact on funding status.

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

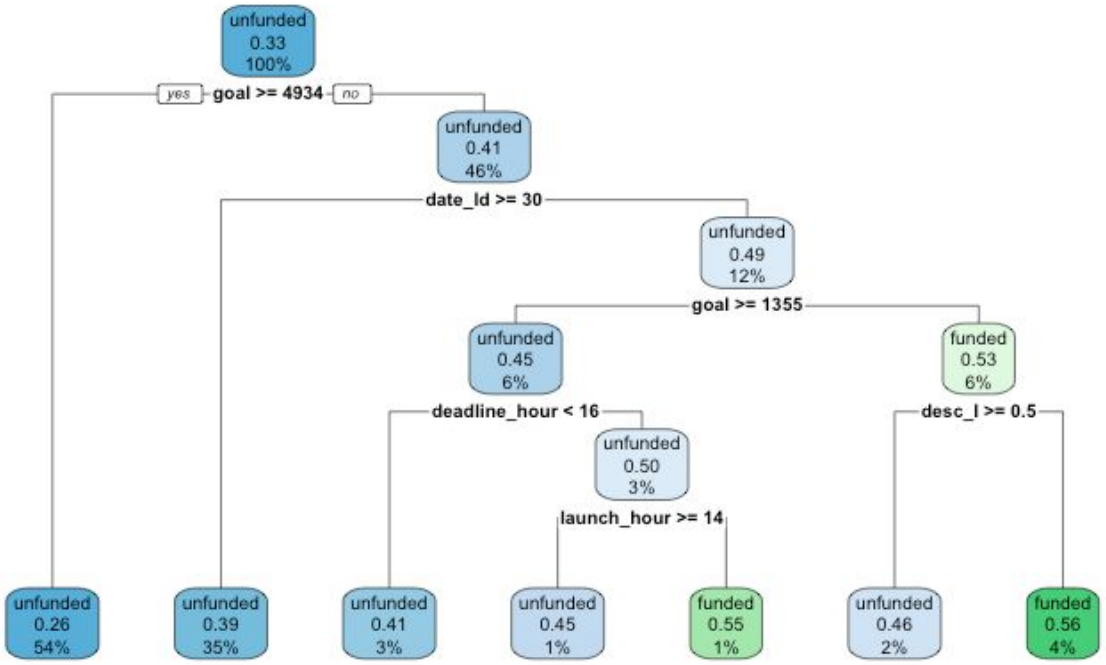
First I used the tree method CART. I applied rpart to all the variables. I used all the variables and set the complexity parameter to 0.001. Next I picked the tree size that minimizes the misclassification rate. Then pruned the tree with the best complexity parameter. After running the pruned tree, most of the splits were not over .6 or over .3, meaning that there was not much differentiation in the splits.

I used the model on the train data. With the confusion matrix, the model accuracy was 67.3% accurate.

Confusion Matrix for CART

	Predicted unfunded	Predicted funded
Actual unfunded	44,299	20,815
Actual funded	1,753	2,221

The model predicted 56% of the funded projects correctly and 68% of the unfunded projects correctly. This translates to the modeling being 56% sensitivity and 68% specific. The model is more likely to predict the unfunded projects correctly.



For the model, the first leaf splits at goal. If the goal amount is greater than or equal to 4,934 the project does not get funded. This isn't too far off as the projects that asked for the most money did not get funded. Next leaf is the date_ld (days between launched and deadline). This one definitely makes sense as several projects have 30 days as their time frame. The split then goes back to goal and onto other variables. Deadline hour over 16 (4pm) was what I observed in the data. Having the launch hour before 14 (2pm) also seemed important when I observed the data. Projects that used "I" didn't seem too significant as the tree split has it almost divided at 50/50.

Random Forest

Next for Random Forest I used as many variables as possible to see how the model would perform. I used goal, launch hour and month, deadline hour and month, project length, description and keyword count, and the use of the words "I", "we" and "help."

Using the confusion matrix, the model had an accuracy of 77%.

Confusion Matrix for Random Forest

	Predicted unfunded	Predicted funded
Actual unfunded	45,113	15,100
Actual funded	939	7,936

The model had a 89% sensitivity rate and a 75% specificity rate. So 89% of the time when the project was funded, it predict funded correctly. When a project is actually unfunded, 75% of the time the unfunded status is predicted correctly.

Logistic Regression

I also tried logistic regression to see if I could improve the model. I used similar coefficients including goal, length of project, deadline hour, launch hour and using the word I. I tried several logistic regressions with various variables and this was the best one. The coefficient p-values are extremely low meaning they are very significant.

```
Call:
glm(formula = final_status ~ goal + date_ld + deadline_hour +
    launch_hour + desc_I, family = binomial, data = ks_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5272  -0.9407  -0.8062   1.3785   6.3768

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.954e-02  3.314e-02   -0.59    0.555
goal          -9.925e-06  4.081e-07  -24.32   <2e-16 ***
date_ld       -1.428e-02  6.623e-04  -21.56   <2e-16 ***
deadline_hour  4.117e-02  1.968e-03   20.93   <2e-16 ***
launch_hour   -4.546e-02  2.117e-03  -21.48   <2e-16 ***
desc_ITRUE    -2.531e-01  1.982e-02  -12.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 87447  on 68807  degrees of freedom
Residual deviance: 84947  on 68802  degrees of freedom
AIC: 84959

Number of Fisher Scoring iterations: 8
```

On the train data, the model accuracy was 67%.

Confusion Matrix Logistic Regression:

	Predicted unfunded	Predicted funded
Actual unfunded	45,227	825
Actual funded	21,993	1,043

The model rarely predicted a funded project correctly. The sensitivity is only 5%. The specificity is 98% meaning that a majority of the time when a project is unfunded it is correctly predicts as unfunded.

6. Results

Here is the accuracy of the models on the train data.

- CART: 67%
- Random Forest: 77%
- Logistic regression: 67%

With Random Forest having the highest accuracy, I applied the model to the test data as well. Using the confusion matrix on the test data, the accuracy of the model was 68%. The model used on the test data had a 54% sensitivity and 70% specificity. The model was more likely to predict the unfunded projects correctly.

The variables that had the most impact on funding status included goal amount, project length, 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. Again we did not use this variable as this would not help my audience. People wanting to start a Kickstarter campaign will not know how many backers they will have until the project has run its course.

7. Recommendations

After researching the Kickstarter data, creating new variables from the data set and running the models, there are additional variables that could have been helpful. With the variety of projects, I think it would have been helpful to have known the categories. I’m thinking there is a major difference between the Kickstarter projects Food & Craft versus Design and Tech. When I think of tech I think of more expensive projects.

Next I would also like to know more about project placement on the website. There is the “new and noteworthy” projects. I wonder what puts a project in this section? There is a recommender at the bottom of the page. I’d like to know how this operates and if there is something I could pull from these areas. I think website exposure affects funding status.

This may be way out of scope, but I think knowing more about social media likes, share, comments and etc, would be very interesting.

Given what I know about the time variables having less of an impact than I suspected, I think there is more to the description. There is a lot that can be extracted from the description. Further work is needed to analyze the project description. I would explore additional keywords and try sentiment analysis. It is possible the tone has an effect. I think a more positive sentiment would lean towards a project being funded.

Overall this predictability of this project may reside in the description or even the keywords.

Appendix

R code