Chris Rutherford

Nile Kealey

Predicting Kickstarter Campaign Success

## Introduction

The primary objective of this project is to figure out what factors influence the probability of a campaign’s success in order to try and develop a model that can be used to predict whether or not a kickstarter campaign will be successful. Are there certain times of the year that make a campaign more likely to succeed? Are certain categories more successful than others? Does the length of a campaign influence how many people are likely to back the campaign?

## Plan of Action

In our dataset, we expect the most useful columns for our models will be the campaign’s goal, its main category, and the launch date and deadline. The launch date will be useful for generating new features since it contains the campaign’s launch year, month, and day.

*Is a certain time of year more likely to yield a successful campaign?*

Using the launch month of the campaigns can be used to help see how many campaigns that are launched each month actually end up being successful.

*Are certain categories more successful than others?*

*Is there an optimal length of time a campaign should be to maximize the chance of success?*

In order to prepare the data for fitting models to it, checking for missing data is a first step. The only real issue is that the **usd goal** column has about 3800 NaNs, but the **usd\_goal\_real** column is the same thing, and has no missing values, so simply dropping this column, along with **pledged** (since **usd\_pledged\_real** uses the same currency converter as **usd\_goal\_real**), will maintain consistency among the units of these columns.

## 

## Dataset Description

The dataset used for this project is a set of 378,661 Kickstarter campaigns, which comes from Kaggle (<https://www.kaggle.com/kemical/kickstarter-projects>), and consists of the following columns:

ID Likely an arbitrary number assigned by the scraper

name The name of the campaign

category The subcategory of the campaign

main\_category The main category of the campaign

currency The local currency of the campaign

deadline The day the campaign ended

goal The monetary goal in local currency

launched The day and time the campaign started

pledged How much money was pledged in local currency

state The outcome of the campaign

backers How many people pledged to the campaign

country The country the campaign started/is based in

usd pledged Amount pledged converted to USD (by kickstarter)

usd\_pledged\_real Amount pledged converted to USD (with external tool)

usd\_goal\_real Goal of campaign converted to USD (external tool)

These campaigns were scraped from the website in early January 2018, so the dataset also contains campaigns that were live at the time of scraping.

There are several kernels that have been run on this dataset, primarily dealing with performing exploratory data analysis and data transformation on the data.[[1]](#footnote-1)

## Data Processing

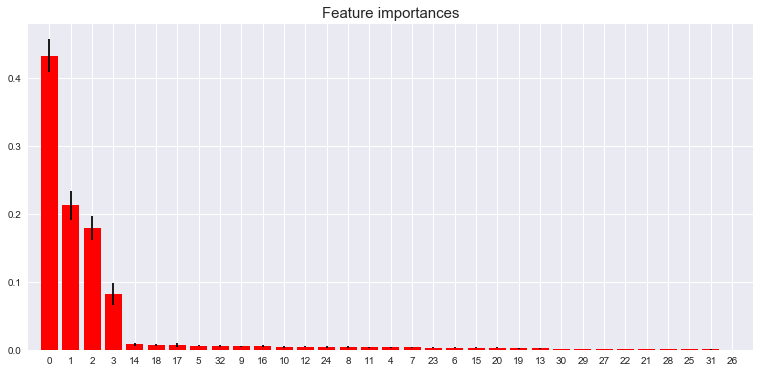
For the entirety of this project, we only dealt with this single dataset for the entirety of our analysis due to the lack of usability of other datasets, as they were either much smaller or had a limited amount of information compared to the dataset we used.

Initially, we converted the **launched** and **deadline** columns to DateTime objects in order to more easily manipulate them and extract/create additional features that can be used in our models. By converting these columns to DateTime objects, we were able to make new columns for the launch year, launch month, and the length, in days, of the campaign (by subtracting the launch date from the deadline). Adding these features provided us with more variables to use in our model in order to help us with our predictions and training the models.

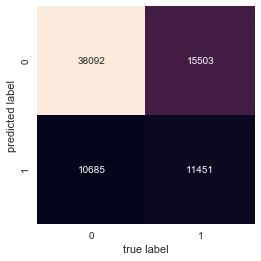
## Results

Due to the complex nature of the relationships between the variables in the dataset, the primary models being tested were random forests classifiers based on the most important features in order to predict the binary outcome of success or failure. The main hyperparameter that was tweaked was the number of estimators, but in the time allotted, identical models with different numbers of estimators yielded identical results.

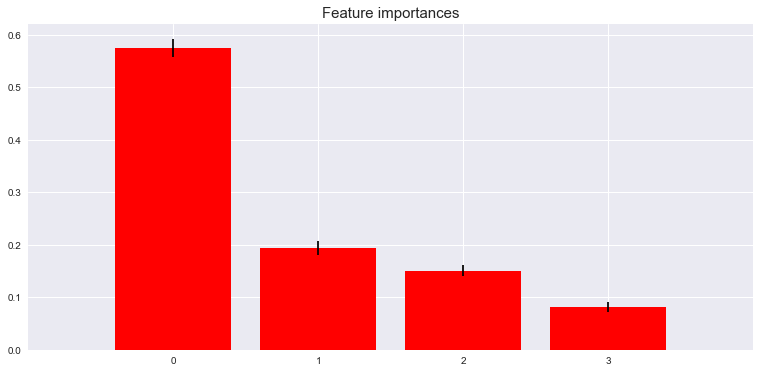
(1)



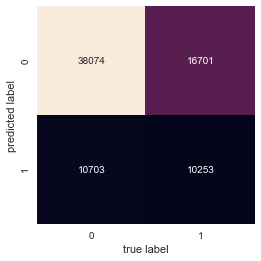
The first decision tree, using usd\_goal\_real, days, launch month, launch year, the main categories, and currencies yielded the feature importance graph in (1). The most important feature is usd\_goal\_real, followed by days, launch month, and launch year. The next 15 features were the 15 main categories, while the rest are the currencies. When predictions are made on the test set, the model yielded an average precision of 0.68, recall of 0.65, and f1 score of 0.64, along with following confusion matrix.



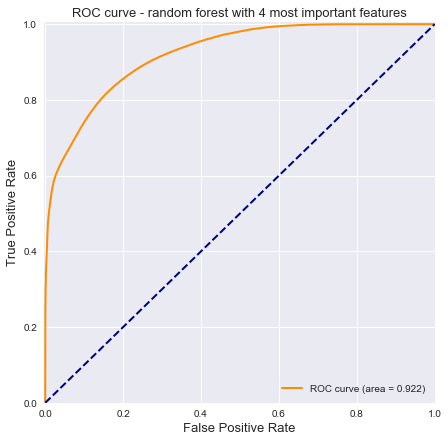
Due to the insignificance the categories and currencies have in the model, a new random forest is created with just the four most important features.



This model yields very similar results to the previous one, with an average precision of 0.67, recall of 0.64, and f1 score of 0.65.

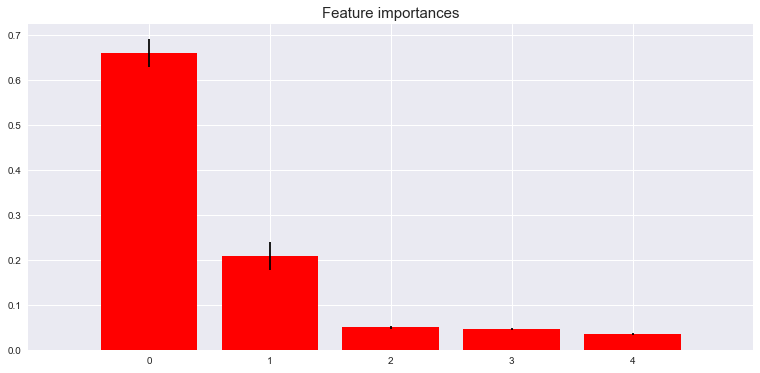


Compared to the above model’s confusion matrix, this model is slightly less able to properly label successful campaigns as successes, so it is possible that the category a campaign is in and its local currency both play some kind of role in determining success/failure. We also obtain the following ROC curve, with an AUC of approximately 0.922.

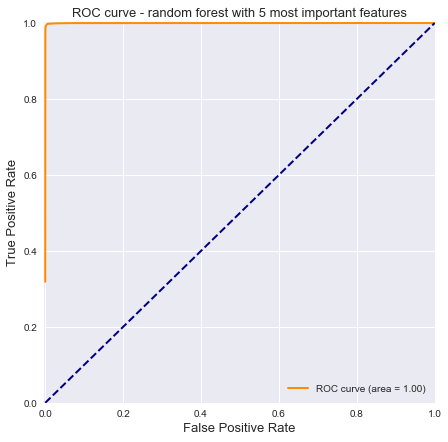


Additionally, a random forest regressor was created with these 4 features to predict the number of backers, which yielded an R2 of about 0.59 on the testing set. Due to the poorer performance of this model compared to the classifier (and time constraints), not much else testing was done with random forest regressors.

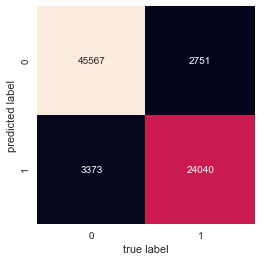
However, an important column, **backers**, can be useful for making and updating predictions as the campaign progresses. A random forest with 100 estimators was used with backers, usd\_goal\_real, days, launch month, and launch year as predictors to predict success.



Not surprisingly, backers is the most important feature used in predicting success, as the more backers a campaign had, the more likely it was to reach its goal and become successful. This model yielded an average accuracy, f1 score, and recall all of 0.92. The following ROC curve was obtained as well.

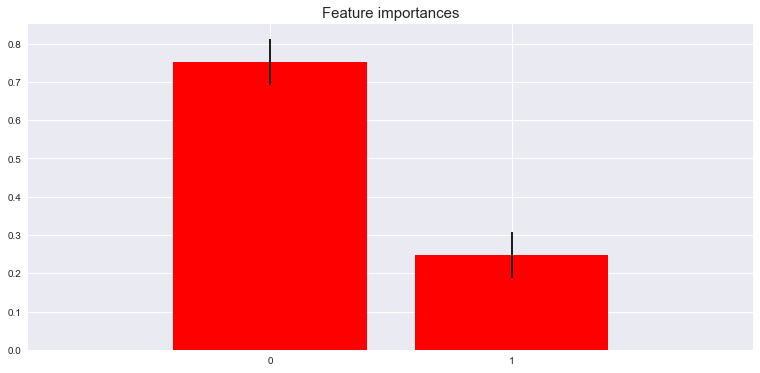


Compared to the other ROC curve, such a high AUC (nearly 1) is once again not very surprising.

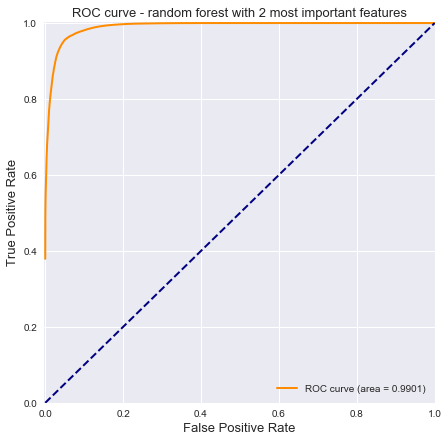


The confusion matrix shows that this model has quite high accuracy, incorrectly labeling about 6000 campaigns out of 75,731 in the testing set.

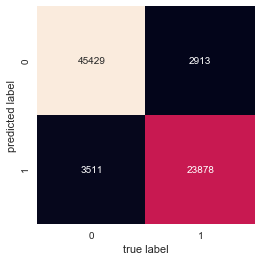
Because of how much more important backers and usd\_goal\_real are compared to the other 3 features, another random forest classifier was tested using just these 2 columns as predictors.



This model yielded an average accuracy 10-fold cross-validation of 0.91 and the following ROC curve.

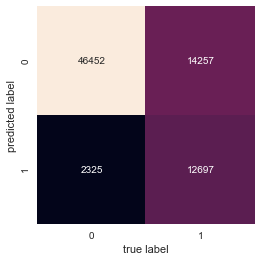


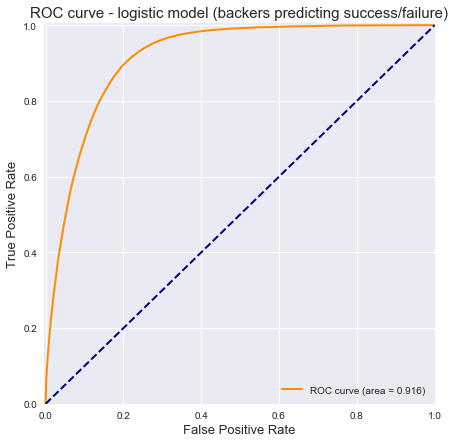
This model’s ROC AUC is still quite high at 0.99, but there is still a noticeable loss of precision when compared to the previous model. Additionally, the confusion matrix shows that more campaigns in the testing set were mislabeled.



However, the complexity of the random forests led us to attempt creating a simpler model with logistic regression. The first model used usd\_goal\_real to predict the binary outcome of success, but would only predict failure every time. The length of the campaign in days also yielded a similarly poor score, only labeling success correctly 57 times out of about 27,000.

The final model, a logistic model using backers to predict success, was unsurprisingly the most accurate, yielding an average precision of 0.86, recall of 0.78, and f1 score of 0.80.



 AUC: 0.916

Likely due to the simplicity of logistic regression compared to a random forest, this model could only predict success correctly about 47% of the time, but failure correctly 95% of the time. However, a model that uses backers as a predictor is really only useful once the campaign has already started and has been live for at least a few days, or whenever people start backing it.

## Discussion

Based on the regression tree classifiers and regressors created, the campaign’s goal is the one of the most important factors in being able to make predictions about the campaign’s likelihood for success. Additionally, the launch month is another useful predictor due to the slight variation in success rates for each month. The launch year is likely not too useful a variable to use due to the model having very little data for 2018 and none for anything in the future, so the model’s ability to predict a campaign launching in the future, even 2019 with other input variables considered, is questionable at best.

Incorporating backers as a predictor variable answers a pretty different question, and despite its close relation with success and always being 0 when a campaign launches, can be useful and more accurate for assessing a campaign’s performance as it progresses.

The complexity of the relationships observed in the dataset lean toward the random forest being the more favorable model due to its ability to capture these more complex relationships, as opposed to the logistic regression models, which are much simpler and can thus only fit to the data so well.

The main problem to solve, given more time, is to figure out which variables provide the best prediction, which variable is more suitable to predict, and which model (and its hyperparameters) would be the “best” model. Additionally, splitting the data into separate sets based on the year and fitting a model for each of these subsets, using the most important features and not using the year as an input variable.

Some additional models we would have liked to test would be more variations of the random forest (such as the number of estimators) or models that are trained and tested on campaigns that started in specific years (such as a model specifically for campaigns starting in 2016). Models that are trained and designed for specific categories would also likely yield more accurate predictions, as the variability in the most important predictor variables, such as usd\_goal\_real, can make it difficult to generalize to every single category.

1. https://www.kaggle.com/kemical/kickstarter-projects/kernels [↑](#footnote-ref-1)