Kickstarter Project Success Prediction

Sahil Chutani (sc4617)
Harsh Mehta (hsm2148)
Umesh Bodhwani (ub2140)
Anunay Sanganal (avs2160)

Introduction

Kickstarter is the world's largest crowdfunding platform whose mission is to "bring creative projects to life". Projects across a plethora of categories like food, music, technology etc. are launched on Kickstarter with the aim of getting backed. It has 13 categories along with 36 subcategories for projects. People who list their projects on the website are called creators. People who fund the creator's projects are called backers. Creators publish their project on the platform, conditional on Kickstarter's approval, and set a goal and deadline for their project. Backers pledge money for the projects they find interesting. A project is considered successfully funded if it reaches (or exceeds) its goal by the given deadline. With such immense, varied data at our disposal, we decided to analyze this data and build models that could predict the success of the project getting funded.

Assuming that the funds for a project are available through crowdfunding platforms like Kickstarter, we tried to explore the probability of a project getting successfully funded. We built various predictive models and compared their performance. After that, we applied our models to live projects to determine their odds of getting funded before the deadline. Finally, we understood the inner workings of Kickstarter and the massive usage of this platform by creators and backers alike.

Data Exploration, Manipulation and Feature Engineering

Kickstarter was launched in 2009 and has been accumulating data related to projects, creators and backers since then. The amount of data available to us was huge. But, we decided to limit the data to be analyzed to just the past 3 months (September, October and November). We scraped the data off Kickstarter's website and collected approximately 200000 data points (projects) along with 46 attributes.

To efficiently parse textual data related to the projects like the description and the title of the project, we wrote a custom parser instead of using the JSON parser. We discarded the data points for which 80% of the columns had null values. We also dropped duplicate data to avoid overfitting later on during model building. Moreover, we applied one-hot encoding on the categorical variables to create dummy variables.

We have summarized some of the important features in the following table:

Goal	The amount requested by creators
Deadline	Estimated time by when the pledged amount must be equal to or greater than the goal for the project's success
Blurb	A short description of the project
State	Current status of the project
Category	13 categories
Location	Location of the creator
Launch day	Day of project launch
Description	Length of project description
Staff pick	Whether project is recommended by Kickstarter (0 or 1)

Table 1 Important features and their description

Through detailed data visualization of all the attributes and observing distribution plots, we were able to do feature engineering.

Through this plot, we could clearly observe that USA has the highest number of projects launched on Kickstarter and leads the other countries like Great Britain, Canada, Australia etc. by a significant margin.

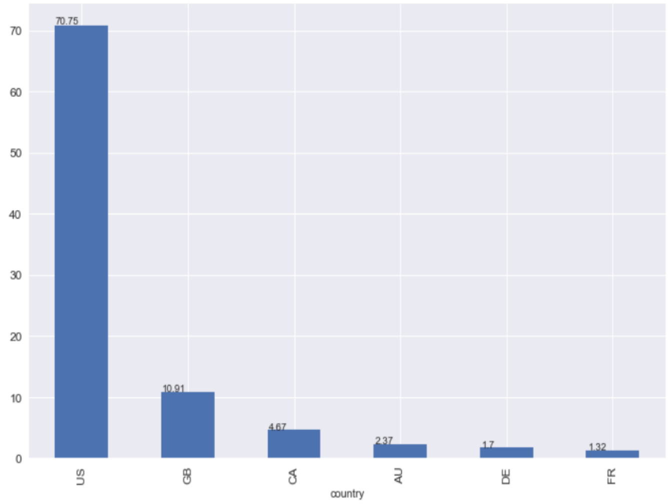

Figure 1 Projects by country

So, we decided to run our analyses on only the projects that have originated in USA. We decided to visualize how projects are spread across the country. As expected, California has maximum projects, with New York following closely. Washington, Texas, Illinois and Florida aren't far behind.

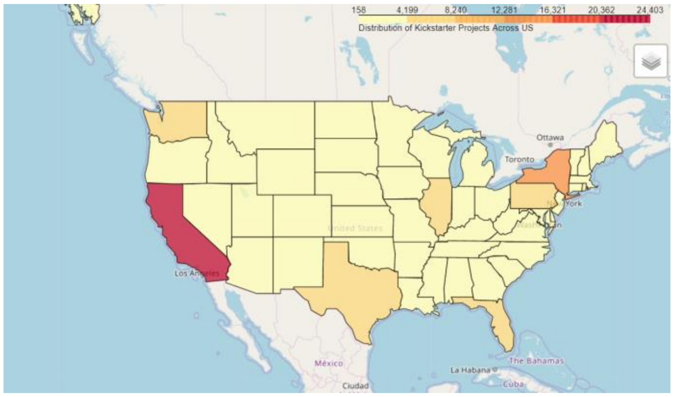

Figure 2 Project spread across USA

Bivariate analysis of percentage of projects by category indicated the top categories like music, film & food and media that see a major number of projects getting launched. Contrary to our assumptions, technology does not top this list. Interestingly, projects that are recommended by Kickstarter (captured in the feature "staff_pick"), have a very high probability of getting funded.

These were the techniques that we used to do the feature engineering for the models that we will eventually build.

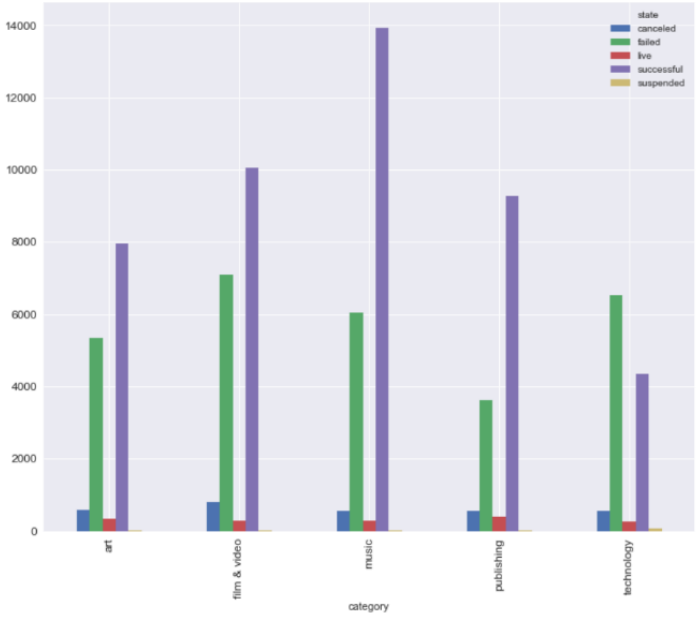

Figure 3 Project state by category

Distribution plots for goal and description length gave us an intuition as to why these features would be important predictors for the success of the project's funding goal.

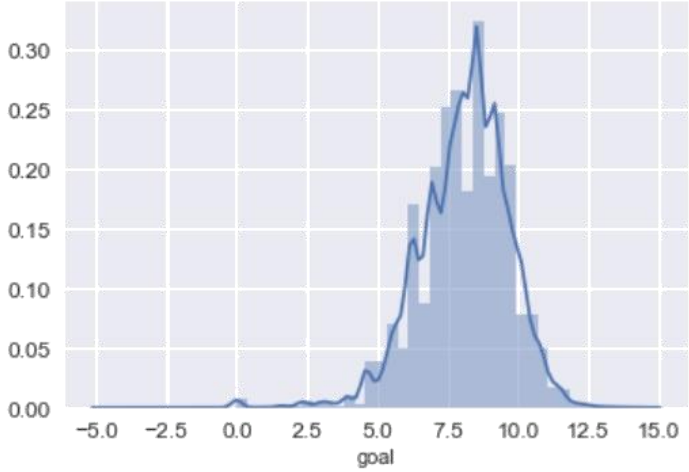

Figure 4 Distribution plot for goal

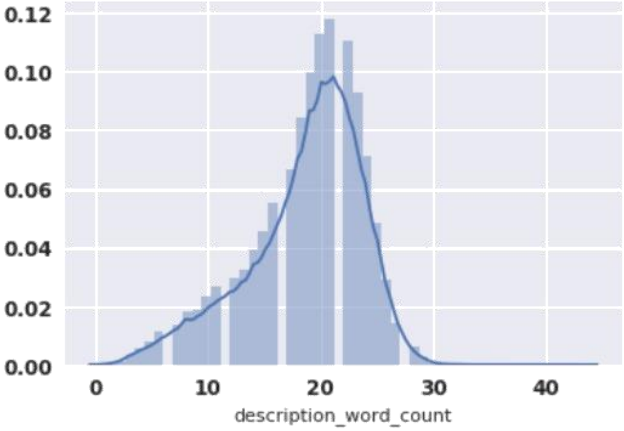

Figure 5 Distribution plot for description length

Predictive Modeling and Model Comparison

We laid more focus on the projects which originated in the US. Firstly, we used a logistic model to analyze the effect of different attributes of a project on its chances of getting successfully funded. We then predicted the probabilities for each project being successful. For training the model, we filtered the data and considered only successful and failed projects. Then we split this data into training set for training the model, and testing set for testing the model. As we can observe, all the variables came out to be statistically significant with their p-values being less than 0.05, indicating that our feature engineering was indeed accurate. The area under the ROC curve for the training set was 0.76 and that for the testing set was 0.72. The trained model can then be used to predict the outcome of the projects which are live.

Variables	Coefficients	p-Value
C(category)[T.fashion]	0.2333	0.00
C(category)[T.film & video]	0.1549	0.00
C(category)[T.food]	-1.0565	0.00
C(category)[T.games]	0.2966	0.00
C(category)[T.music]	0.5496	0.00
C(category)[T.technology]	-0.4749	0.00
C(location_type)[T.Suburb]	-0.6235	0.00
C(location_type)[T.Town]	-0.3883	0.00
staff_pick	2.5528	0.00
goal	-1.79E-05	0.00
Days_to_Deadline	-0.0232	0.00
description_length	-0.0012	0.00
Weekend	-0.1719	0.00

Figure 6 Betas obtained from Logistic Regression

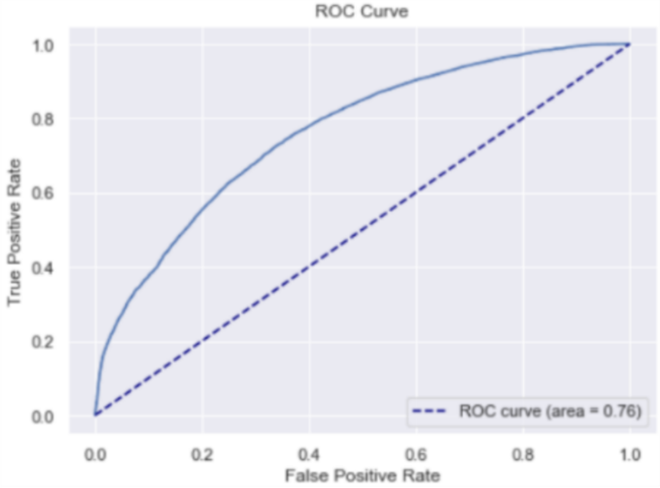

Figure 7 ROC Curve Logistic Regression

We also applied Decision Tree, Random Forest, Bagging Classifier and Feed Forward Neural Network models to our dataset. We wanted to compare the performance of different models. The best accuracy was obtained by Logistic Regression model.

Random forests gave the important features that would predict the success of a project getting funded. Looking at the top 7 features, they seemed to be as expected intuitively.

We also learnt that neural networks would need even more relevant features to make accurate predictions.

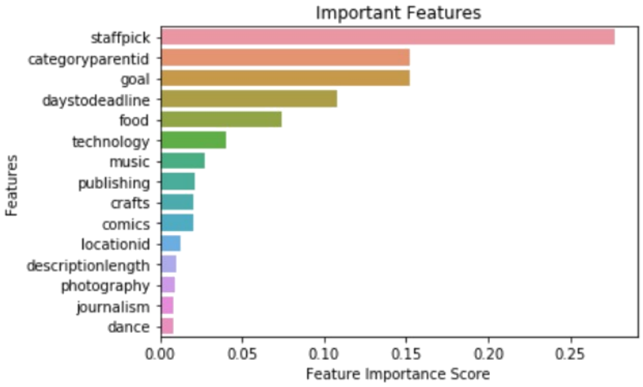

Figure 7 Important features through Random Forest

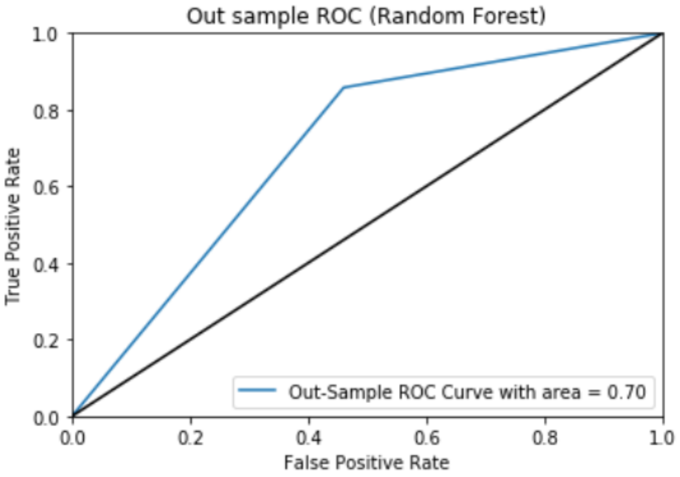

Figure 8 ROC Curve Random Forest

Model	Train Accuracy	Test Accuracy
Logistic Regression	76% (ROC)	76% (ROC)
Decision Tree	68.20%	67.87%
Random Forest	72.35%	70.01%
Neural Network	55%	46%

Figure 9 Model comparison

Key Insights

Since our stakeholders include creator of the project and the backers, we have divided insights from our models into two categories:

- For the creators:
  - Goal – The goal of the project should lie in an optimal range of $10,000 - $40,000 to ensure the project's success.
  - Deadline – 30-day deadline gives the project enough time and demonstrates confidence, but varies from project to project.
  - Description length – The optimum length of project description is 140 characters.
  - Category – If a project category is overlapping, the creator is recommended to list the project in the category which has more chances of success. For instance, if a music project backed by technology can be listed under the music category which has had historically higher odds of success.
  - Launch day: A project launched on weekday has more slightly higher chances of success.

- For the backers:
  - New backers getting started on Kickstarter should fund projects that are staff picked, which will be a safe bet for them.
  - They should observe the creator's profile over time. A creator with a good history of backed projects is more likely to be successful.
  - Location can be a big factor that impacts success of a project. California and New York has a vibrant start-up community, increasing the avenues for marketing and publicity of projects and hence raising chances of them getting funded.

Once we observed these insights from our model, we wanted to validate these claims. So, we took a project that had already been marked as failed on Kickstarter. We marked the details that could have caused the project to fail. We tweaked features like the description length and content, deadline and launched it on a weekday. We ran our model on this specific project again and observed a 10% increase in the success probability. This experiment validated our claims as to which features, that are in the hands of the creator, are actually making a difference to the project success.

Limitations

- One major limitation we faced was the lack of additional valuable features that would increase the accuracy of our models. Data related to the projects on Twitter and Facebook was almost impossible to scrape for more than a week. These attributes could have had a significant impact on our predictive model.
- Also, we did not have updated data for the projects each day (time-series data). In fact, the data was only updated after the project's deadline was exceeded. So, if a project was funded even before it reached its deadline, we did not have access to that data. Such data could prove instrumental in running a timing model that could predict the number of days a project takes to get funded.

One way to overcome these limitations is to cultivate alternate data sources that could provide more granular data for the projects. This would eventually lead to better modelling of data.

Conclusion & Future Scope

Through this project, we were able to uncover the secrets behind success of Kickstarter projects. US still leads all other countries in terms of innovation and start-up culture. Creators have good opportunities and resources at their disposal if they take advantage of crowdfunding platforms like Kickstarter. Our models can be taken into account by an amateur creator who is trying to work on his idea. After all, subjective qualities like the passion and drive shown by the creator as well as the efforts he puts into the project's publicity are going to matter significantly. Backers also need to make data-driven investment decisions to get maximum returns from Kickstarter projects. Although, backers are treading uncharted waters with most of the newer projects. That said, Kickstarter is a great platform that can serves its purpose efficiently.

The future scope for this project would majorly include collection of data from additional sources to input into our model.

- Incorporate Twitter and Facebook data about the projects to determine how marketing on social media impacts the success of a project.
- Improve the accuracy of models by taking into account more features
- Tune the hyperparameters of neural networks to improve its performance.
- Analyzing the kind of rewards that backers receive or are looking for to model the motivation behind backing a project.
- Incorporate a custom Marketing Score corresponding to each project to observe its impact on the odds of success.