Prediction of Kickstarter Projects

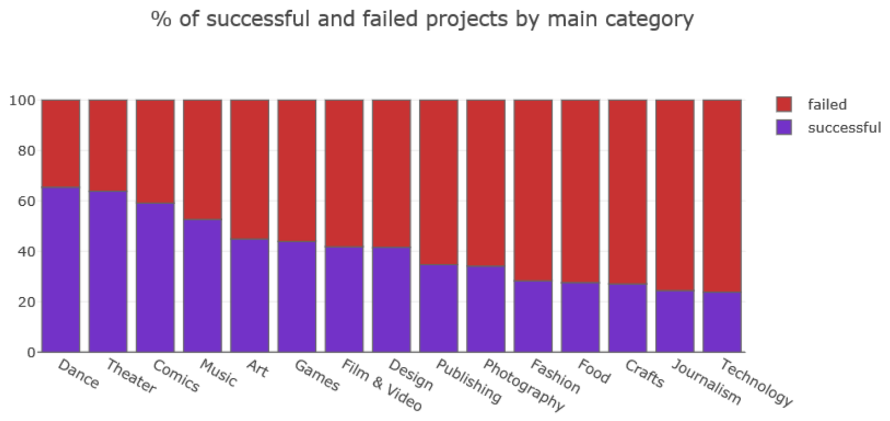
Kickstarter is the world's largest funding platform for creative projects. People post their project ideas on the website with the goal amount of money to finance their projects and attract backers to pledge. If the pledged amount of money meets the goal amount, the project is deemed as successful; otherwise, failed. This project will attempt to set up a model to predict the outcome of a potential Kickstarter project given a set of features. Why is this important? The most funded Kickstarter campaign over time was on the "Coolest Cooler," a multi-function cooler, which raised $13,285,226 and 62,642 backers with a goal amount of just $50,000. If a model can predict whether a potential project will be successful or not, it can definitely help those project owners invest into the right industry, with the right goal amount in the right country.

The data used for this project was pulled from Kaggle.com which originally collected data from Kickstarter website. The data spans from 2009 to 2018 and has 378302 rows (projects) of 15 variables (project features). The following are the variable names and their definitions:

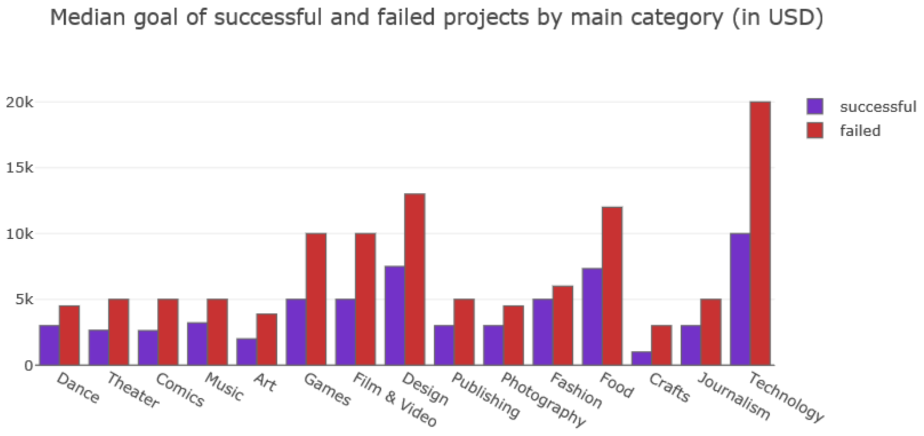
Variable	Definition	Variable	Definition
ID	Id of a project	Pledged	Pledged amount of a project
Name	Name of a project	State	State of a project
Category	Category of a project	Backers	Number of backers for a project
Main_category	Main Category of a project	Country	Country a project was in
Currency	Currency a project was pledged in	Usd_pledged	Pledged converted to usd dollars (kickstarter)
Deadline	Deadline of a project	Usd_pledged_real	Pledged converted to usd dollars (3rd party)
Goal	Goal amount of a project	Usd_goal_real	Goal converted to usd dollars (3rd party)
Launched	Launched date of a project		

There were 4 missing names and 3797 missing usd_pledged, but usd_pledged_real is basically the same variable, and names were irrelevant in the model, so nothing had to be done for those missing values fortunately.

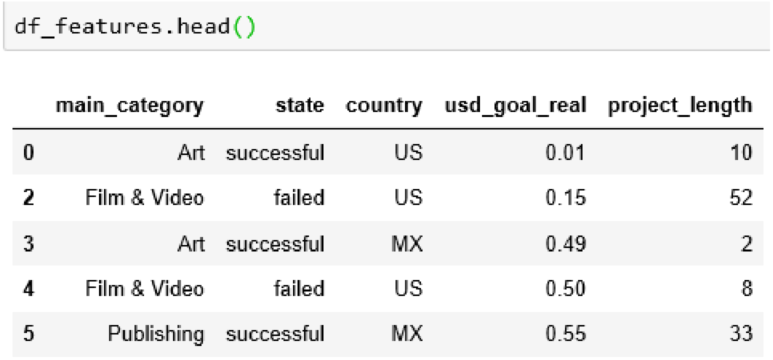
From the exploratory data analysis, there was a key finding that there were certain categories of a project that can potentially affect the outcome; art projects in dance and theater were more likely to be successful than technology projects.



This trend can be further supported by the next chart. Since the technology has a higher goal amount than dance and theater, it is more likely to fail because higher goal means it is harder to meet the goal successfully.



For data processing, first extra states other than “successful” or “failed” were left out because it wouldn’t make sense to predict any other states like “cancelled.” Then another feature called “project_length” was created calculating the number of days between the launched and deadline dates. The project length would be more correlated to the state of a project rather than start or end date. Also unnecessary features were dropped to speed the modeling code. Features used for the model were: main_category, state, country, usd_goal_real, and project_length.

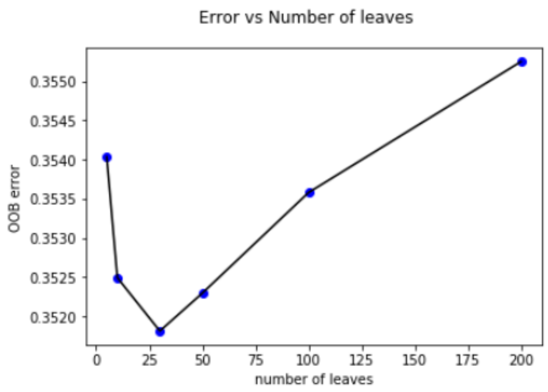


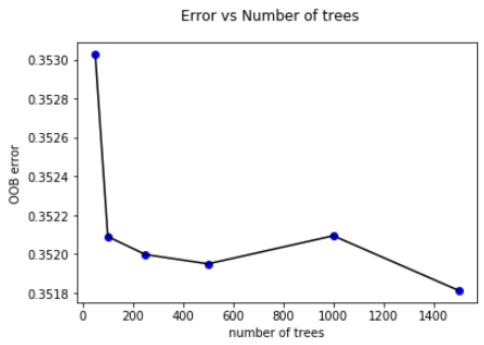
```
df_features.head()
```

	main_category	state	country	usd_goal_real	project_length
0	Art	successful	US	0.01	10
2	Film & Video	failed	US	0.15	52
3	Art	successful	MX	0.49	2
4	Film & Video	failed	US	0.50	8
5	Publishing	successful	MX	0.55	33

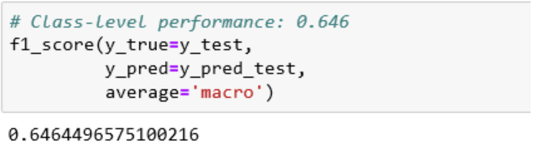
Backers and usd_pledged were dropped because they are obviously too highly correlated to the state and also a project owner wouldn't know how much pledged or how many backers one has in the planning stage. Category would be a useful feature, but it was too granular and too many categories. For categorial variables like main_category or country, dummy variables were created indicating each category or country. Then the data was split into train and test sets, with test set (25% of data) randomly chosen. Train data is used to train the model, which would later be validated with the test data.

For modeling, this project chose a random forest algorithm, because the data is quite enormous and random forest is relatively time-efficient while being accurate. The important parameters for the algorithm were number of minimum leaves and number of trees and these were chosen to be 30 and 1500 that would lead to the lowest prediction errors shown as below:



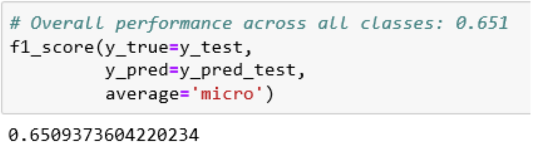


The result of the model is as follows:



```
# Class-level performance: 0.646
f1_score(y_true=y_test,
         y_pred=y_pred_test,
         average='macro')
```

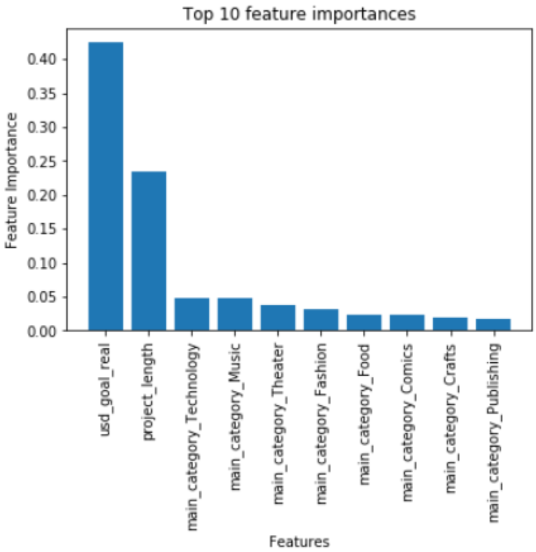
```
0.6464496575100216
```



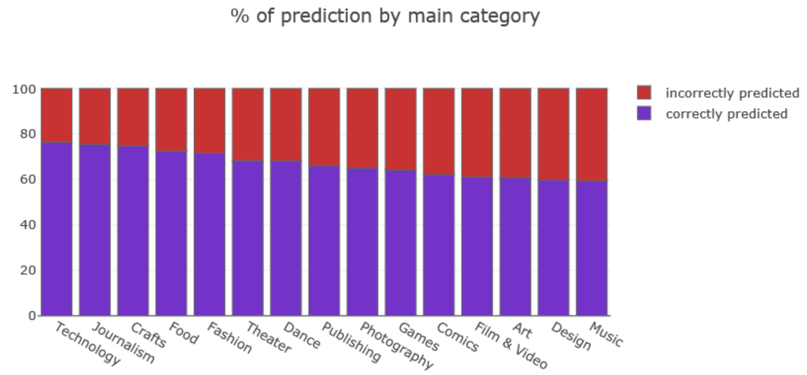
```
# Overall performance across all classes: 0.651
f1_score(y_true=y_test,
         y_pred=y_pred_test,
         average='micro')
```

```
0.6509373604220234
```

The state-level accuracy of the model was 64.6% and overall accuracy across all states was 65.1%. The below picture also indicates that the most important features in the model were the goal amount and then the project length as well as the category of a project.



Also, there were certain categories that were easier to predict the outcome/state. As you can see below, technology and journalism had higher prediction rates than art and design projects. The hypothesis here is that since technology and journalism had noticeably higher goal amounts, it would be relatively easier to predict that they will fail.



The potential project owners will input the features they want their projects to have in the following format: {'main_category': 'Comics', 'country': 'US', 'usd_goal_real': 1000, 'project_length': 55} meaning the project will be in the Comics field located in the United States with the goal amount of $1,000 USD dollars with the project length of 55 days. This will return the output of 0 or 1, 0 meaning failed, 1 meaning successful project. This will help the project owners to change their features of the project, or go ahead and stick with the plans for a successful kickstarter project.

The challenges of the project were there were not enough features from data. Most features were not useful for predicting the outcome of a project, or they were features that would not be in project owner's control because they are not something a project owner can choose before release. Still, 65% of accuracy was not bad given these challenges.

The resulting recommendations and future research topics would be to get other features that are more in project owner's control such as the amount of money a project owner invested in the planning stage. A project owner usually has some costs before release in prototype design as well as advertising. The project can also utilize another classifier algorithm other than random forest to see whether there are improvements in prediction accuracy. Lastly, this project can be implemented on other crowdfunding platforms like Indiegogo or GoFundMe.