Report: Kickstarter Projects – Success or Failure Prediction

Business Problem

The business problem that I seek to solve is: How can the company Kickstarter increase the success rate of its projects over the next year by spending resources on projects with the higher likelihood of success?

Data

I downloaded two Kickstarter data from Kaggle:

a. https://www.kaggle.com/kemical/kickstarter-projects
b. https://www.kaggle.com/yashkantharia/kickstarter-campaigns

Data Wrangling

I took the following steps to clean and wrangle the data.

Step 1: Exploring and Combining the dataset

I explored the two datasets to understand what columns were present in both, how many rows were there, what were the datatypes of the columns, and what were the maximum and minimum years. I wanted to use the two datasets to get as much data as possible. Looking at the two datasets, I found that after combining them, I could get information from years 2009-2019, to make the data recent and relevant. I renamed the columns that were in both datasets so that concatenating them was easier. I then combined the two datasets. The resulting dataset had above 500,000 rows.

Step 2: Cleaning the dataset

- First step was to ensure that the dataset did not have duplicated observations. Every project had a unique Kickstarted ID, which was used to find and drop duplicated observations. After dropping the duplicated IDs, I was left with 408, 278 columns. This I believe is a very good number.
- Next step was to filter for 2009-2019. I first converted the launched and deadline columns, which had date and time to datetime format. I then extracted the years out of the columns and filtered for years after or equal to 2009.
- Then, I dropped the columns that were of no analytical important. These columns were in one dataset and not in the other. Therefore, they had null values when concatenating them.
- The next step is an interesting one and has to do with the duration column in the dataset. One of the datasets had a duration column and the other did not. Duration could have contributed to the prediction of success and failure of the projects. It was necessary to calculate the duration for all the projects. I did so by subtracting the launched datetime object from the deadline datetime object if the duration column was empty in a particular row.

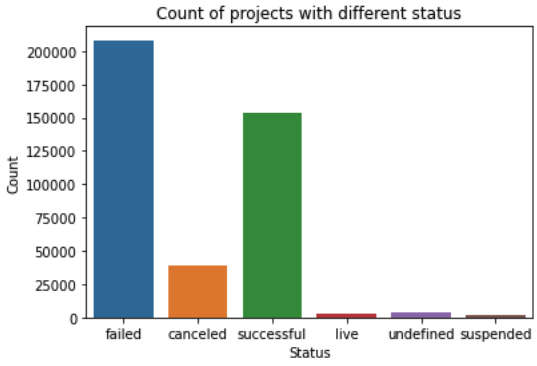
- Lastly, I checked for all the null values and made sure the columns did not have any null values.
- I then ran a simple descriptive statistic for all the numerical columns.

	id	usd_pledged	usd_goal	duration	launched_year	deadline_year
count	4.082710e+05	4.082710e+05	4.082710e+05	408271.000000	408271.000000	408271.000000
mean	1.074309e+09	9.783897e+03	4.354569e+04	33.104969	2014.521930	2014.592751
std	6.193319e+08	9.230929e+04	1.113579e+06	12.716384	2.095593	2.094641
min	5.971000e+03	0.000000e+00	1.000000e-02	0.000000	2009.000000	2009.000000
25%	5.373578e+08	3.700000e+01	2.000000e+03	29.000000	2013.000000	2013.000000
50%	1.074942e+09	7.000000e+02	5.200000e+03	29.000000	2015.000000	2015.000000
75%	1.610296e+09	4.350005e+03	1.500000e+04	36.000000	2016.000000	2016.000000
max	2.147476e+09	2.033899e+07	1.663614e+08	93.000000	2019.000000	2019.000000

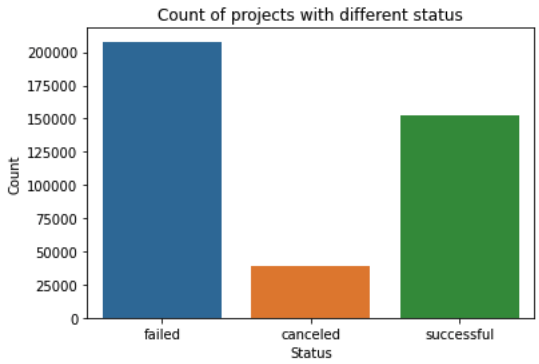
Step 3: Exploratory Data Analysis and Data Pre-Processing

This dataset required some extensive exploratory data analysis and pre-processing. I will describe the process for each variable as showing EDA and Pre-Processing step simultaneously will provide clarity.

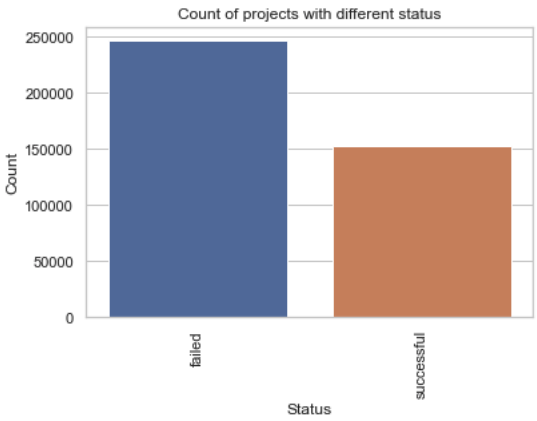
Status of the Project



We see that there are three categories 'live', 'undefined', and 'suspended', that do not give an insight into the status of the project, whether it failed or was successful. Additionally, the number of observations in these categories are also very few. So, I dropped them, leaving us with the following categories.

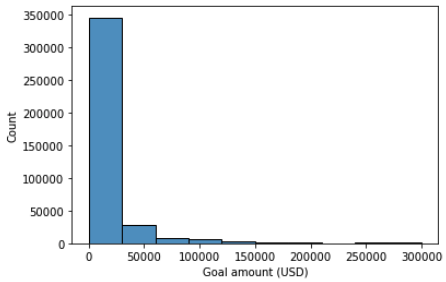


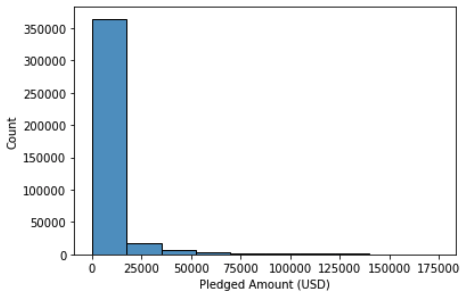
To better analyze the data, I categorized canceled projects into the failed category. Leading us to this final picture of the variable.



Goal Amount and Pledged Amount

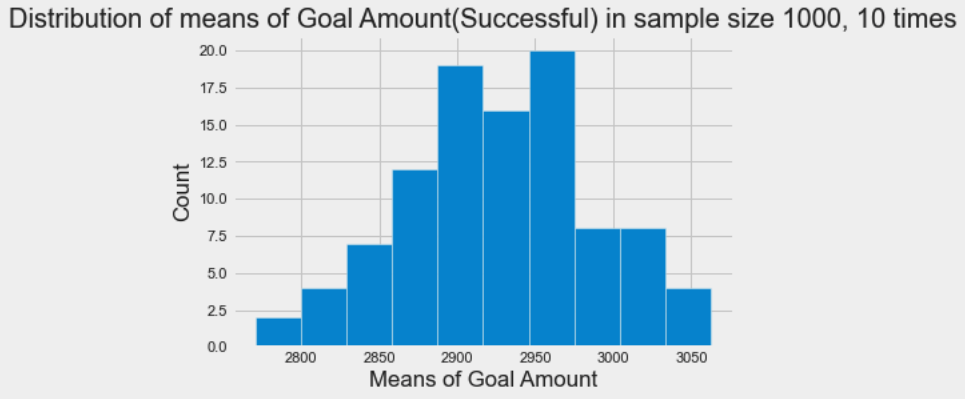
Both the Goal Amount and Pledged Amount were highly skewed towards the lower end. Even after filtering for observations less than a specific amount, the data was still skewed and not normally distributed.

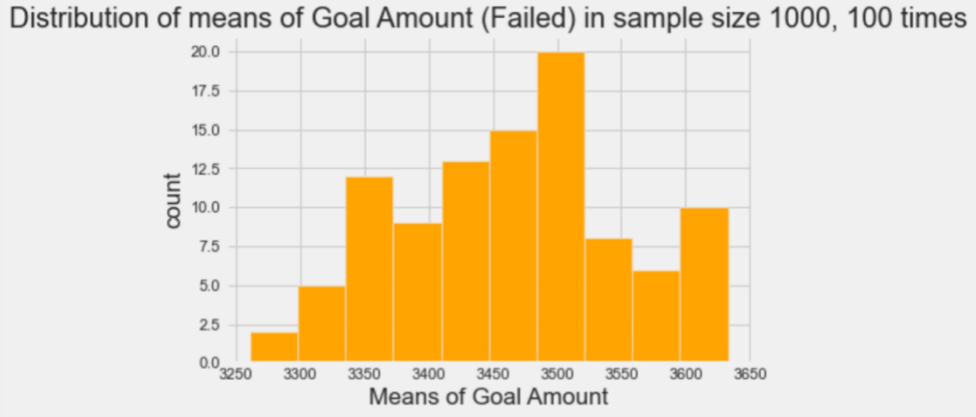




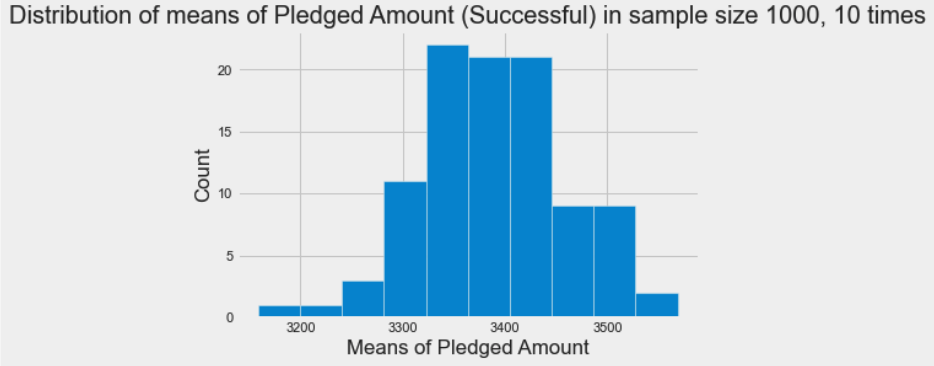
I checked for the normality of the data in these two columns using the assumption of the Central Limit Theorem: if we take a large sample of a data multiple times, calculate the mean of sample, and map the distribution of the means, the distribution should be normally distributed. I divided the data into one with the status as 'Successful' and another one with the status as 'Failed'. It was necessary to do so to run independent t-tests at the hypothesis testing phase (which will be described later). Below is the histogram of the distribution of the mean, which are normally distributed.

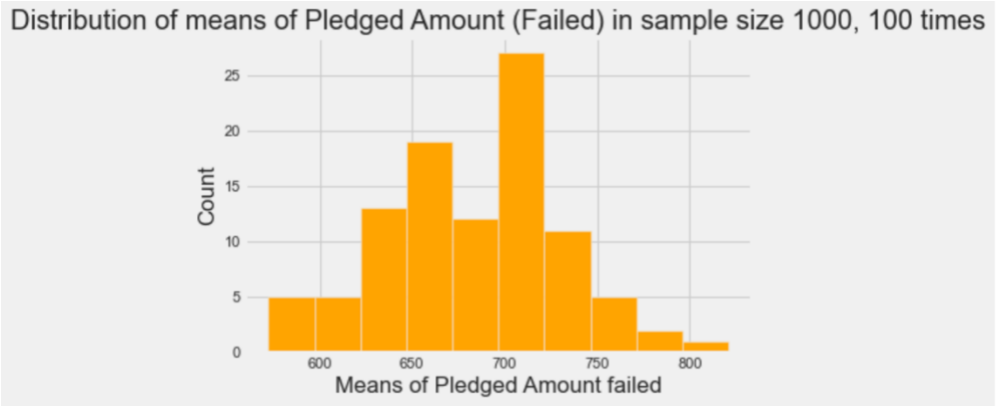
Goal Amount



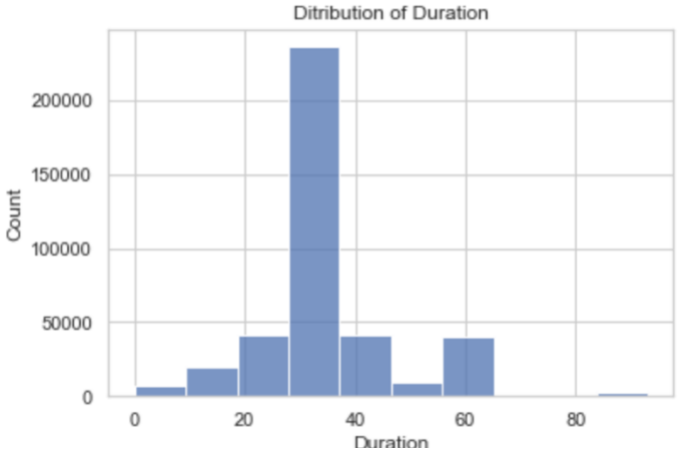


Pledged Amount



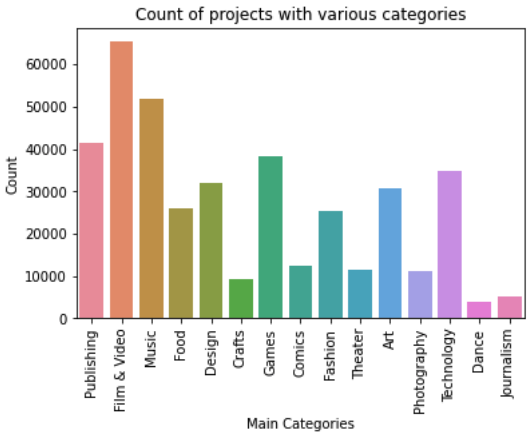


Duration



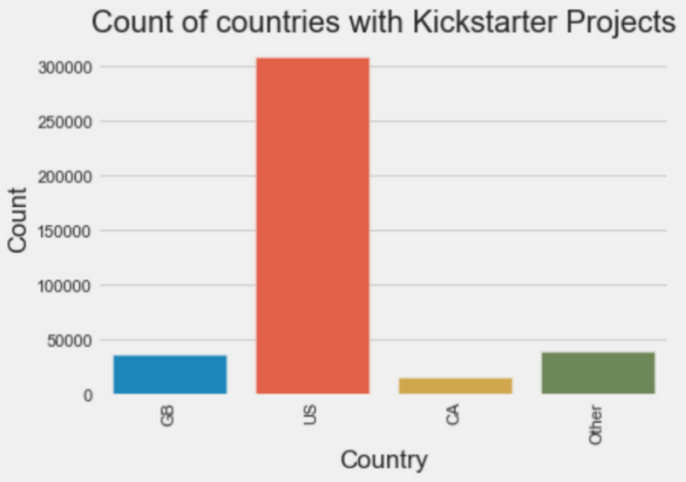
We see that the duration is normally distributed with an average of 33 days.

Main Category



In one dataset, the categories were spelled with the first letter capitalized while in another, they were spelled with all lowercase letters. I changed them to the first letter capitalized

Country



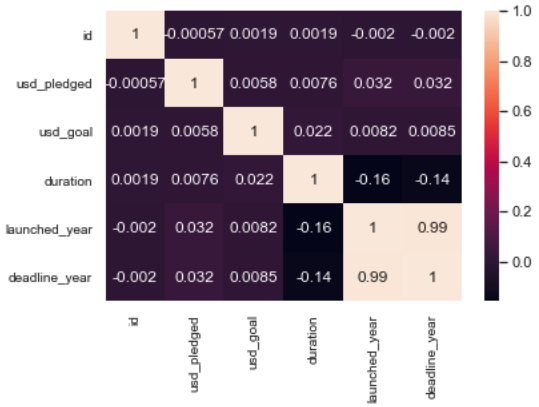
The country variable contained a value 'N,O"'. I dropped these values because there were either null values or a country that I could not recognize. I had to drop 232 rows, which is justified given that I have a large amount of data.

In the preprocessing stage, I masked all the other countries that were not Great Britain, United States, or Canada as 'Other' as other countries did not have nearly enough observations to be of analytical importance.

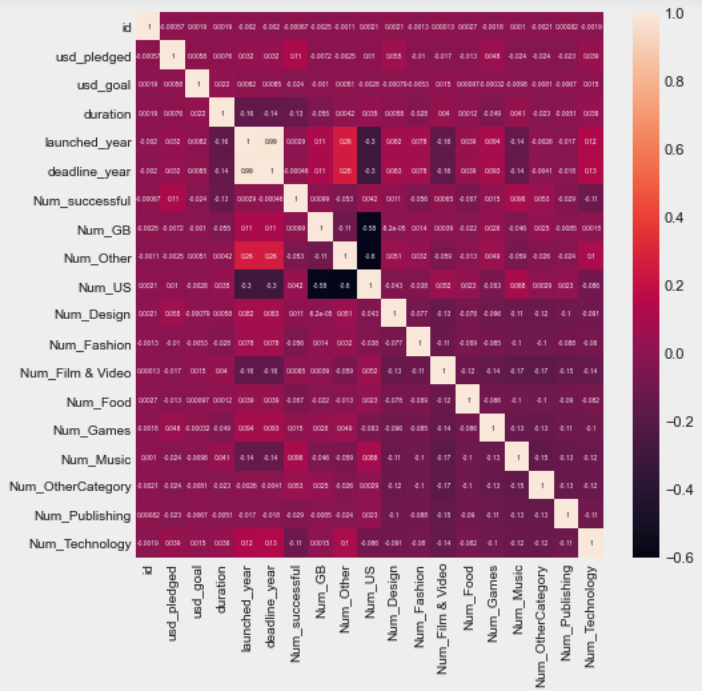
Exploring the Relationships Between Variables

The first step that I did in exploring the relationship between variables is make a heatmap that shows how all the numerical variables are correlated. The heatmap shows several interesting relationships:

Heatmap with Numerical Variables



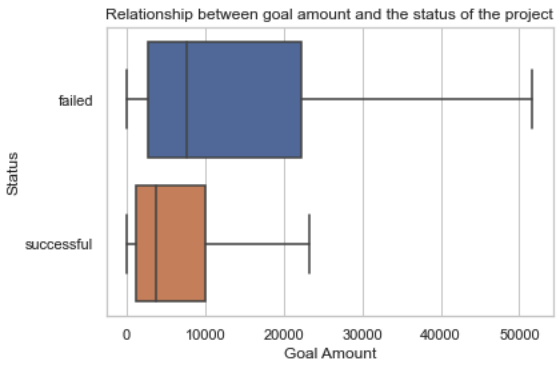
Heatmap with Dummies



a. First, the independent variables of interest are not highly correlated to one another.
b. Status (DV) maybe correlated with the pledged amount and with the country variable. The correlation might be better with the pledged amount.
c. Status does not seem to be correlated with duration.
d. The correlation between status and the goal amount appears to be very weak.

Let us dive deeper into the relationships.

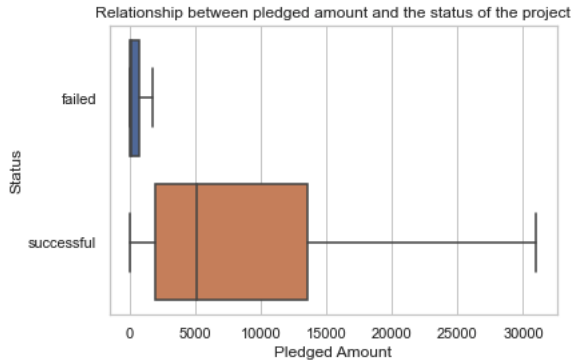
Relationship 1: Goal Amount and Status



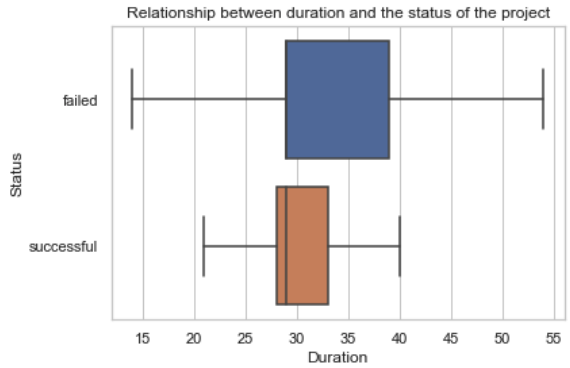
The boxplot gives a clearer picture. The picture shows that the median goal amount of successful projects is less than the median goal amount of failed projects. This indicates that with higher goal amount, there is more chances of failed projects.

Relationship 2: Pledged Amount and Status

The above boxplot shows that the median pledged amount of successful projects are a lot higher than the median pledged amount of failed or canceled projects. This indicates that success of projects is likely to be determined by higher pledged amounts.



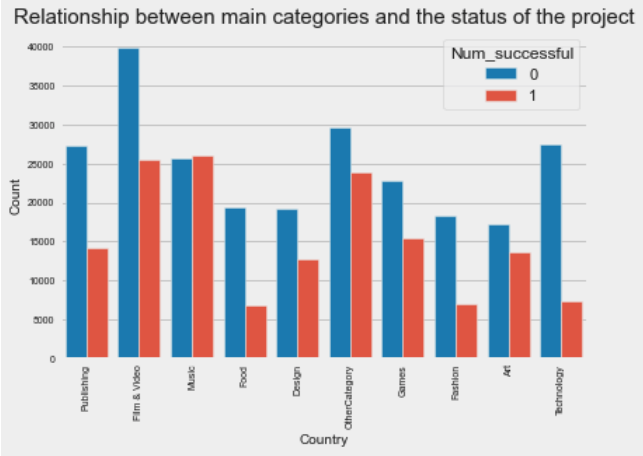
Relationship 3: Duration and Status

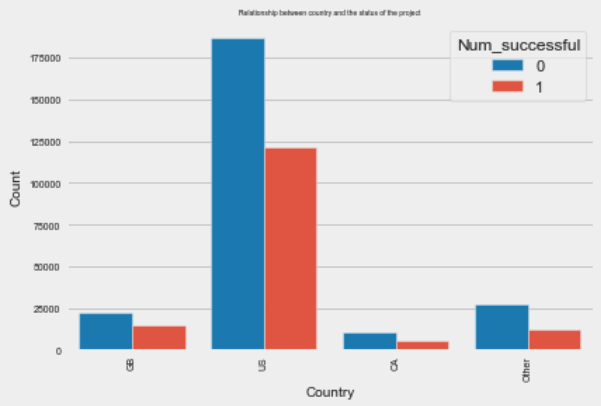


We see that for the failed status, the 25th and the 50th percentile is the same. The median of the successful projects is not different than that of the failed project, indicating that duration of the project may not be correlated with the status of the project.

Relationship 4: Main Category and Status

The above figure shows that categories may not be the best indicator of success or failure as almost every category has more failed projects than successful ones.





Relationship 5: Country and Status

It does not look like countries determine the status of the project as all countries listed have more failed projects than successful ones.

Getting Dummies

The next step was to get dummies for categorical variables: status, country, and main category.

Hypothesis Testing

Numerical Variables

I ran an independent samples t-test to explore whether there are differences in means of the Goal Amount, Pledged Amount, and Duration between successful and failed projects. We used the t statistic instead of the z statistic because we are conducting the difference of means test from a sample of Kickstarter projects.

Goal Amount	
t -statistic	p-value
52.95	0.0

The p-value is zero, which indicates that there is a significant difference in the mean of the goal amount of the failed and the successful project status. The t-statistic is good enough, however, not as high as the pledged amount.

Pledged Amount	
t -statistic	p-value
317.76	0.0

The p-value is zero, which indicates that there is a significant difference in the mean of the pledged amount of the failed and the successful project status. The t-statistic is also very high, which gives stronger evidence against the null hypothesis. This takes us back to the result we got from the heatmap. The heatmap also showed a stronger correlation with the pledged amount.

Duration	
t -statistic	p-value
0.0	1.0

The p-value is one, which is very high, and which indicates that there isn't a significant difference in the mean of the duration of the failed and the successful project status. Moreover, the t-statistic is zero which is very low to indicate any evidence against the null hypothesis.

Categorical Variables

I ran the Chi-Squared test to conduct hypothesis testing between categorical variables – status, country, and main categories.

Variables		Test Statistic	p-value
Country			
	Great Britain	35.78	0.0
	United States	163.81	0.0
	Other	997.97	0.0
Main Categories			
	Design	41.57	0.0
	Fashion	1168.80	0.0
	Film & Video	14.21	0.0
	Food	1676.03	0.0
	Games	83.01	0.0
	Music	3222.93	0.0
	Publishing	296.64	0.0
	Technology	4277.03	0.0
	Other	965.81	0.0

The p-value is 0 for all the categories across Main Category and Countries. Therefore, we can safely reject the null hypothesis.

Data Modeling

I will now detail the steps that I took to model my data.

Splitting data into train and test set

I defined my X and Y and split my data into train and test set.

Scaling the numerical variables

I scaled my variables after splitting my data into train and test set. Scaling the data before splitting would have scaled my dependent variable, which is already a categorical variable with categories 0

and 1. I also scaled only the numerical variables, which are the goal amount, the pledged amount, and duration.

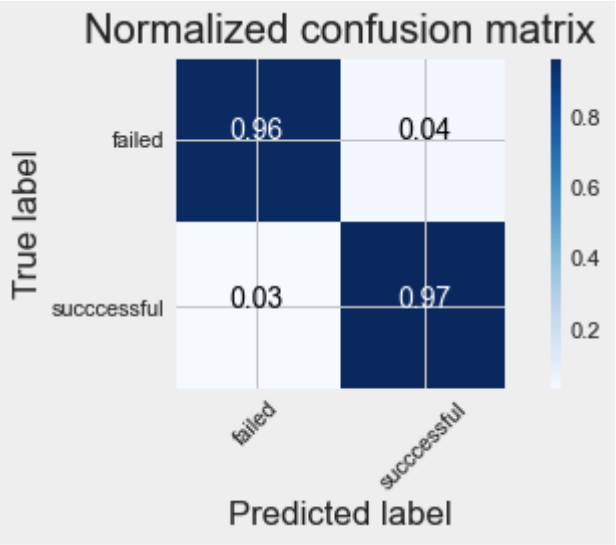
Modeling Strategies

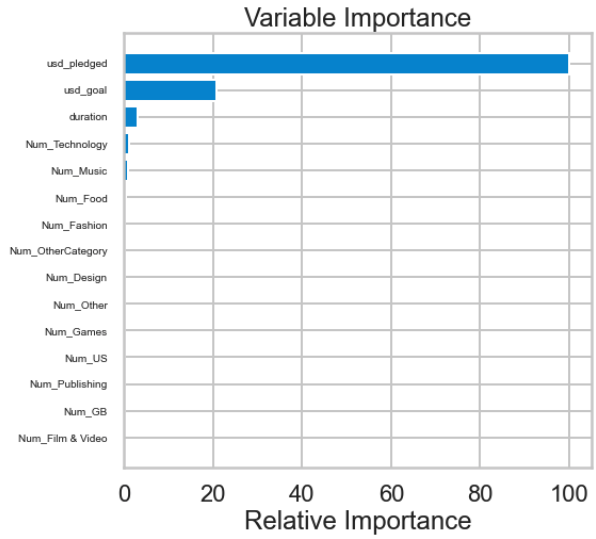
I will use Random Forest Classifier and Logistic Regression models because my dependent variable 'Status' is categorical.

Model Metric

Model	Reason	**Accuracy (before HP tuning)**	**Accuracy (after HP tuning)**	**F1 (before HP tuning)**	**F1(after HP tuning)**	**Mean CV score**
Random Forest Classifier	-Reduces overfitting in single decision trees -Quick prediction - Employs techniques to reduce variance	0.97	0.97	0.97	0.97	0.97
Logistic Regression	-Efficient training -Coefficient size and direction of association	0.90	0.98	0.90	0.98	0.90

The Random Forest Classifier model performed similarly before and after hyperparameter tuning. The classifier also shows the Pledged Amount in the most important feature in predicting the success and failure of Kickstarter Projects.





The Logistic Regression Model did significantly better after hyperparameter tuning. It also did better than the Random Forest Model after tuning the hyperparameter.

Coefficient Table (before HP tuning)

	0	Coefs
0	duration	-0.232223
1	usd_goal	-162.903573
2	usd_pledged	34.207443
3	Num_GB	0.241070
4	Num_Other	-0.079583
5	Num_US	0.120062
6	Num_Design	-0.712969
7	Num_Fashion	-0.809624
8	Num_Film & Video	-0.151221
9	Num_Food	-0.773217
10	Num_Games	-0.609592
11	Num_Music	0.168373
12	Num_OtherCategory	-0.020695
13	Num_Publishing	-0.382313
14	Num_Technology	-1.346527

Coefficient Table (after HP tuning)

	0	Coefs
0	duration	-0.121450
1	usd_goal	-1210.741467
2	usd_pledged	125.662820
3	Num_GB	0.202374
4	Num_Other	-0.083461
5	Num_US	0.173934
6	Num_Design	-0.471574
7	Num_Fashion	-0.613868
8	Num_Film & Video	0.108425
9	Num_Food	-0.589107
10	Num_Games	-0.526046
11	Num_Music	0.252544
12	Num_OtherCategory	0.046779
13	Num_Publishing	-0.139392
14	Num_Technology	-0.916331

The Coefficient Tables of the Logistic Regression Models show that the Pledged Amount is the highest positive factor in predicting the success or failure of Kickstarter Projects. The coefficient for the pledged amount increased significantly after tuning the hyperparameters. The heatmap showed that the independent variables were not highly correlated.

I therefore think that the Logistic Regression Model is the better model to predict the success and failure of Kickstarter Projects.

Kickstarter should therefore focus on increasing the pledged amount for its projects. Better advertising of the projects irrespective of country or category will be helpful.