Kickstarter Final Project

Problem:

Determine whether a project will be successful or not on Kickstarter. Understand which factors affect project success

Data Cleaning Steps:

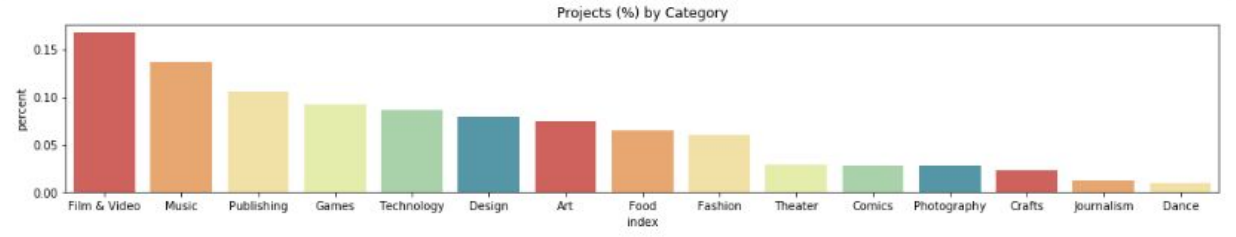
General Data Cleaning: In order to do this project, there are several data cleaning steps that were done. First, variables such as launched_date and deadline dates were converted into months. New variables were created. First,we calculated the difference between when the project was launched and when the project expected to end (launched_deadline_days_diff) and we also calculated the amount of requested money they need to raise by day (goal_per_day).

Missing Country: Also, around 3K values did not have a country and instead the value ‘N,0’ was used. First, the value was replaced by NA and then all values were converted to numbers using label encoding in order to feed the data into the KNN algorithm. This algorithm ‘predicted’ the 3K values.

Machine Learning Preprocessing: In order to feed the data into the Machine Learning Model, the following steps were done. First, the state variable that contained information about whether the model is failed or successful,into binary variables with 1 indicating project was success and 0 indicating project failed. Furthermore, we extracted the tag of speech from each word and converted all unique variables into new columns using one hot encoding. All other categorical variables (main_category,category, country) were into new columns using one hot encoding. Additional numeric columns (goal_per_day, usd_real_goal) were converted centered using Standard Scaler.

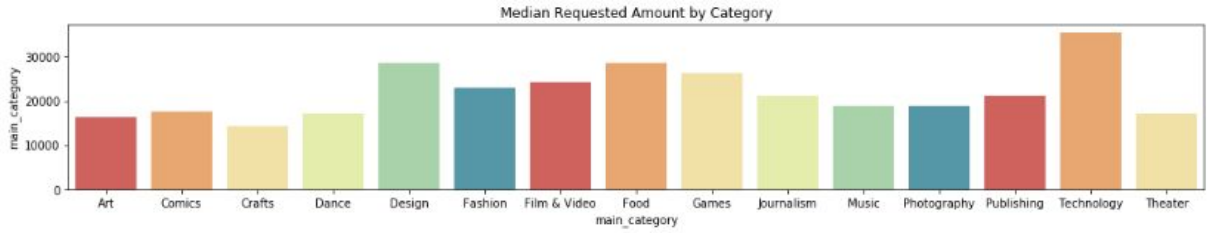
Exploratory Analysis:

1.1 Analysis Category



"Diagram 1: 17% of projects are from Film&Video while 15% of projects are from Music!

The chart above indicates that the Film & Video and Music Category are the most common categories. In contrast, very few projects are about Dance.



"Diagram 2: Technology category seems to have highest median goal requested per project!

But if we look at the data based on the amount of requested; then we see projects in the Technology category tend to request more money compared to other projects. In contrast, Music and Crafts projects tend to request the least amount of money compared to other categories.

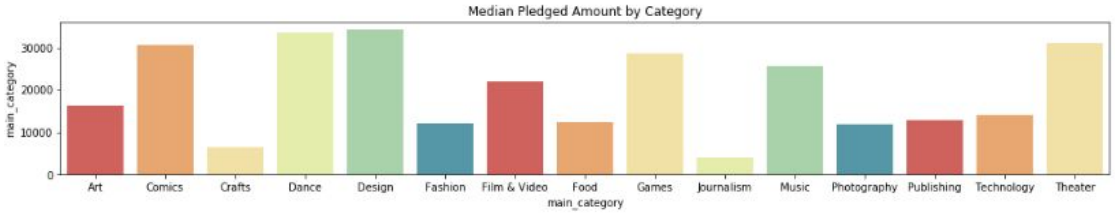


Diagram 3: Dance, Design, Comics, and Theatre have highest median USD pledged.

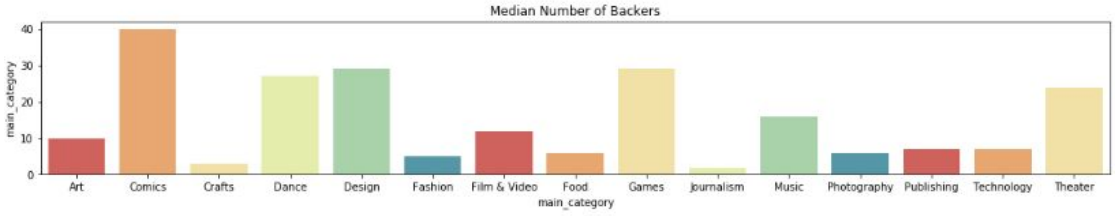
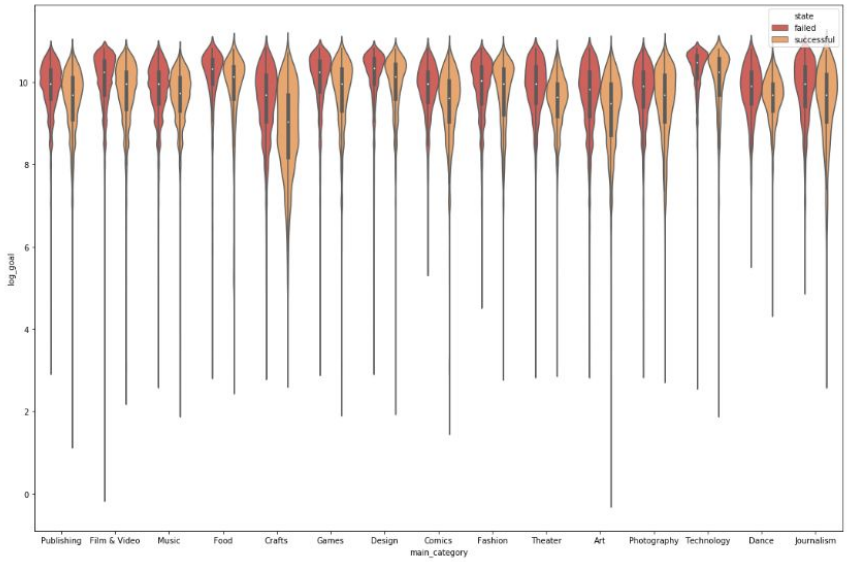


Diagram 4: Dance, Design, Comics, Games, and Theatre have highest median number of backers...

Now, if we look at the median amount pledged and median number of backers, another story emerges. Despite that Technology projects tend to request a lot of many that doesn't necessarily translate into more backers or more money pledged. In this case, for Comics,

although there are fewer comics projects that tend to request less money, on average, they tend to have more backers and more money pledged.

1.2 Analysis of Category by State



Based on the violin plot above, it seems that projects that fail tend to have higher distribution of values or at least some people tend to overestimate how much they can ask for a particular project. Based on this, it seems, Comics successful projects tend to have lower mean amounts requested compared to failed projects. On top of that, successful art projects tend to have a very wide range of possible values - in fact,they tend to have the lowest min starting value compared to other projects.

2.1 Analysis of Country

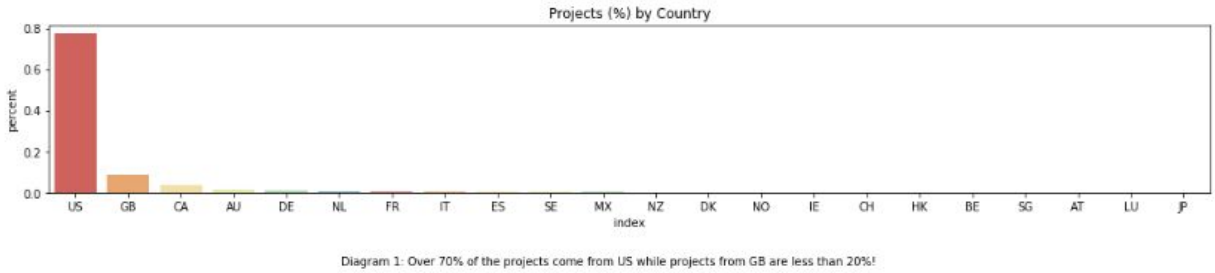


Diagram 1: Over 70% of the projects come from US while projects from GB are less than 20%!

This data is very US centric as most of the projects are from the US. Also, countries such as CA and GB tend as well are represented.

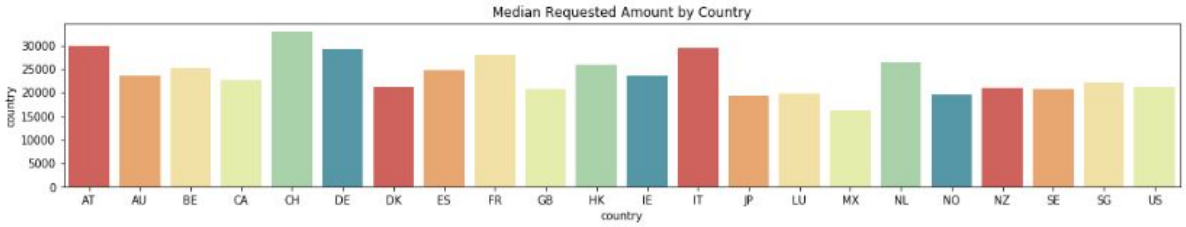


Diagram 2: It seems median requested amount is highest in CH,DE,and LU.

Despite that the projects mostly come from the US, the median requested amount is more equally distributed. In fact, the data indicates that the highest median requested amount is from CH and IT. In contrast, MX projects tend to request the least amount of money.

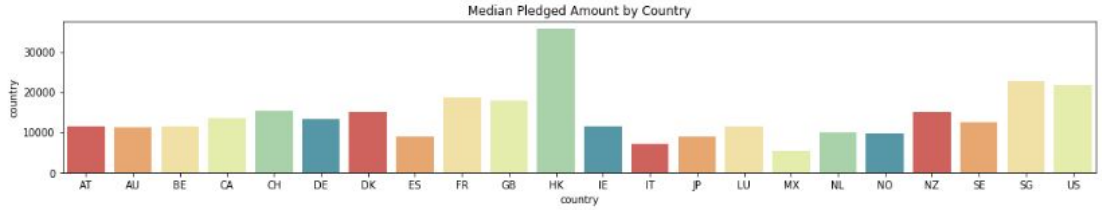


Diagram 3: Median pledged amount is highest in Hong Kong.

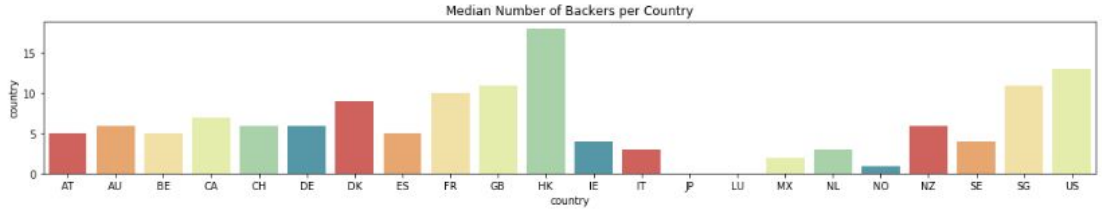
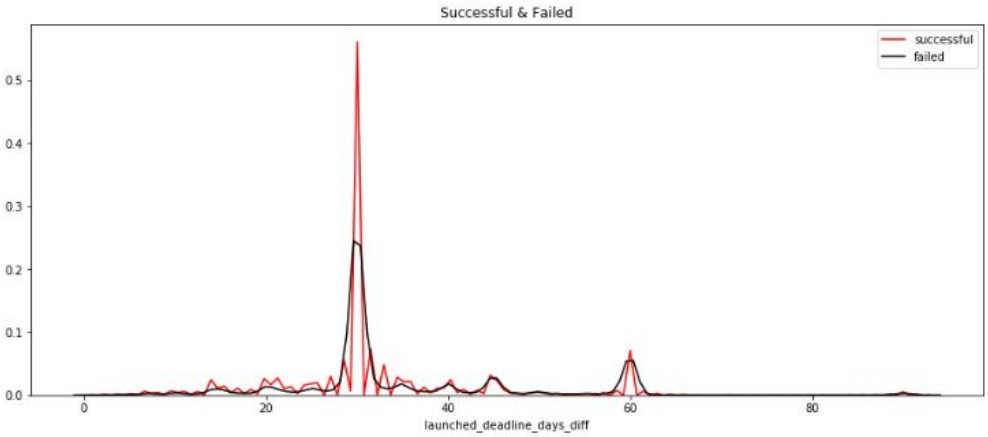


Diagram 4: HK tend to have most backers.

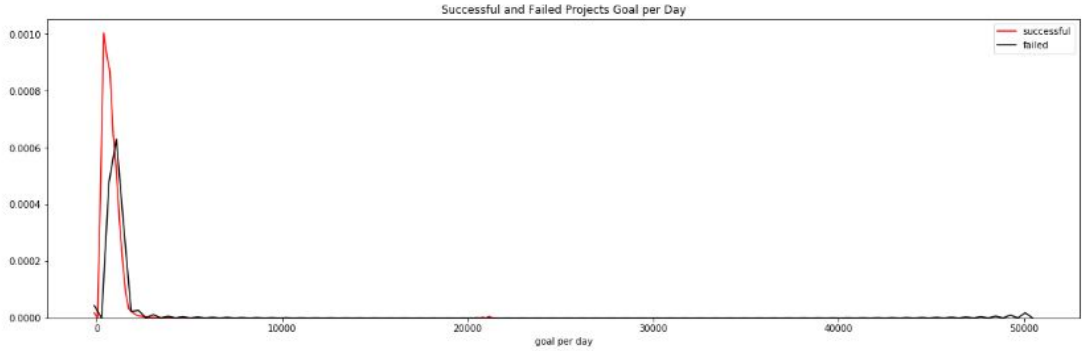
If we look at the median amount pledged and number of backers per country, there are several observations. First, again, requesting a lot of money doesn't translate automatically to having backers or a high median pledged amount. In fact, we see that although projects from IT tend to request a lot of money they barely have any backers or have low median amounts pledged.

3.1 Analysis of Project Length



This chart shows us the distribution of project length (difference between deadline date and launch date) for successful and failed projects. As we see the project length of 30 days is the most common project length. As we see successful projects are more likely to pick 30 days than failed projects perhaps indicating that people whose projects fail are more unsure as to what timeline they have to pick.

4.1 Analysis of Goal per Day



This chart indicates the distribution of how much the project creator needs to raise per day in order to be successful. The distribution for successful and failed projects are roughly similar; however, failed projects tend to have a long tail indicating there are several projects where project creators need to raise too much money per day indicating that they are requesting too much money for too little time.

Key Insights & Implications

Based on the analysis, there are several key insights:

- Amount requested and amount pledged varies by country and by category

- However, just because countries/categories have a high average requested amount that doesn't mean that this will attract a lot of pledges. Thus, there is discrepancy between people creating new projects and having those projects actually generate interest from Kickstarter
- Furthermore, it looks like there is discrepancy between the amount requested between projects that fail and projects that succeed. As we saw in the charts above, across most categories and countries, projects that fail tend to on average request more money compared to projects that succeed.
- Over 50% of successful projects take 30 days between the time projects are announced and the end date. For failed projects, around 25% of projects take 30 days to complete, perhaps indicating that for failed projects people are less certain about what timeline they should put for the project.

Machine Learning

Pre-model Variable Selection

Generally speaking, median amount pledged and requested might be good predictors based on previous analysis. However, median amount pledged can lead to data leakage as projects with high median amount pledged are likely to be successful. It would have been more useful for the business if we knew the median amount pledged in the first week of the project and see how that is related to whether the project is successful or not. Additional variables we can explore are country, main_category, and project length as they tend to differ across median amounts pledged and requested and can help determine if the project is successful. Also, we will include the project title and see if we can draw additional insights.

Machine Learning Process

1. **Preprocessing:** numerical variables that will be standardized, categorical variables that will be on hot encoded, and text which will be vectorized
2. **Basic Model:** Apply several models (Ada boost, Logistic Regression, SVC, XGB,SGD) and select the one that with highest accuracy
3. **Hyper Parametrization:** Figure out if there are any gains in performance when using hyper parametrization
4. **Stacking:** See if there are any meaningful performance gains against basic model if we deploy stacking algorithm

Results

The SGD model performed best with 69.6% accuracy on the training set and similar accuracy for the test set. Hyper parameterized model was deployed with optimized values of 1_ratio,penalty, and alpha. Final best model had alpha: 0.001, penalty: l2, and l1.ratio: 0.1. The model's performance improved only slightly by 0.1 %. Thus, the baseline SGD model was kept. We later implemented a stacking model with Adaboost, XGB, and SVC models. However, the model was only 67.5% accurate so the baseline model was kept.

Diagram 1: Baseline Model Comparison

Model	**Accuracy**
SGD	69.6%
Logistic Regression	68.3%
AdaBoost	67%
Random Forest	59.6%
SVC	48.9

Precision and Recall

Diagram 2: Precision and Recall

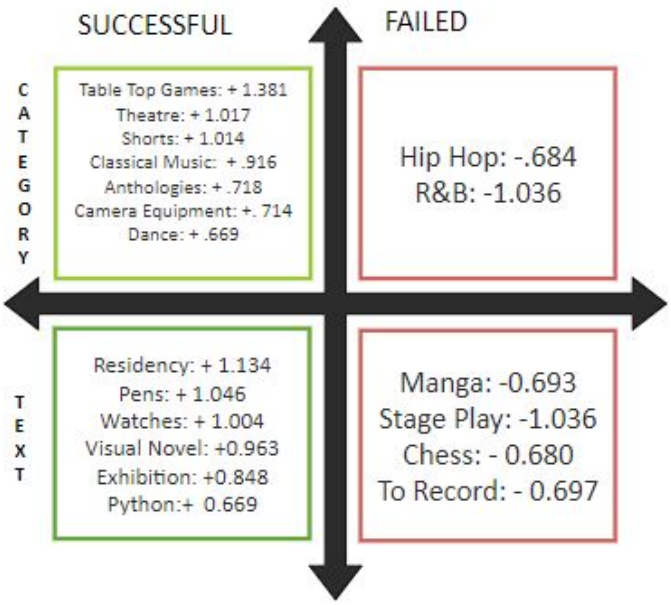
```
[[33171  6405]
 [13874 12885]]
              precision    recall  f1-score   support

      failed       0.71      0.84      0.77     39576
  successful       0.67      0.48      0.56     26759

    accuracy                           0.69     66335
   macro avg       0.69      0.66      0.66     66335
weighted avg       0.69      0.69      0.68     66335
```

This diagram indicates that the recall for success is low indicating we are more likely to predict success compared to actual successful projects.

Variable Importance

Based on the output below, it can be concluded that category and text (title) information have the biggest impact whether a project is successful or not.

Diagram 3: Variable Importance for Successful and Failed Projects



CATEGORY: It shows projects related to Tabletop Games, Theatre, Classical Music, or Dance have higher chance of being successful compared to Project related to Hip Hop, and R&B

TEXT: If the project text had residency, pens, watch, visual novel, or python in its text name, it had a higher chance of succeeding compared to projects that had manga or chess in the title.

Key Business Insights

- Categories and the title of the project are important predictors to determine if project is success or not
    - Make data easily accessible to project creators to understand what chances they have of success
    - Utilize information from text analysis help project creators to write project titles that engage their readers
    - Provide information to project creators to understand mean amount requested