**Final Project Report**

**Kickstarter Projects Data Analysis**

**Group 7**

Ryan French

Saurabh Gupta

Vedika Shenoy

**IST 707 – M002**

**Data Analytics**

Instructor:

Frank Marullo

Spring2019

Table of Contents

[1. INTRODUCTION 3](#_Toc6781646)

[2. OBJECTIVE AND BUSINESS QUESTIONS 3](#_Toc6781647)

[3. DATA DESCRIPTION 4](#_Toc6781648)

[4. DATA PREPROCESSING AND PREPARATION 5](#_Toc6781649)

[5. DATA ANALYSIS 6](#_Toc6781650)

[6. CHALLENGES 11](#_Toc6781651)

[7. CONCLUSION 11](#_Toc6781652)

# INTRODUCTION

Kickstarter is the world’s largest platform for people who want to develop creative projects through crowdfunding. Consisting of projects from a variety of categories such as Arts, Design and Tech, and Music, Kickstarter enables artists, musicians, filmmakers, designers and other creators to bring their ideas to life by providing them with the required funding. As for the patrons who “pledge” to help projects, they are frequently presented with rewards or experiences for providing their support.

Over the course of this report, we will be detailing our project which endeavored to identify key components contributing to project success, building classifiers capable of predicting project success, and predicting which projects will ultimately be funded successfully. In order to achieve this outcome, we have utilized a variety of Data Science techniques as well as the languages Python and R and the Weka platform.

# OBJECTIVE AND BUSINESS QUESTIONS

Through the use of our business questions we hoped to frame the objectives initially highlighted above and establish concrete inquiries with which to direct our work.

1. What are the factors that primarily contribute to the overall success of a Kickstarter project?
2. Can these factors be manipulated to increase the likelihood of project success?
3. Can we build predictive models that can classify projects as successful or unsuccessful?
4. If so, what is the accuracy with which these models can predict project outcomes?

# DATA DESCRIPTION

This dataset consists of Kickstarter projects from various categories along with the end result of the project (whether it was a success, failure, or if it was cancelled) and was obtained from the popular database repository Kaggle. The dataset consists of approximately 379,000 records and 13 fields.

The Kickstarter dataset can be found here: <https://www.kaggle.com/kemical/kickstarter-projects>

The attributes of the dataset are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| No | Attribute Name | Attribute Type | Definition |
| 1 | ID | Numeric | Internal Kickstarter ID |
| 2 | name | String | Name of the project |
| 3 | category | String | Sub category of project |
| 4 | main\_category | String | Main category of project |
| 5 | currency | String | Currency used to gather funding |
| 6 | deadline | Datetime | Deadline for crowdfunding |
| 7 | goal | Numeric | Fundraising goal |
| 8 | launched | Datetime | Date launched |
| 9 | pledged | Numeric | Amount pledged by backers |
| 10 | state | String | Current state of the project |
| 11 | backers | Numeric | Number of backers |
| 12 | country | String | Project country of origin |
| 13 | usd pledged | Numeric | Amount of money pledged |

# DATA PREPROCESSING AND PREPARATION

We began the process of data processing by performing some exploratory analysis to better understand our data. Once we felt comfortable with it, we then transitioned into the process of data cleaning. This proved quite an extensive operation due to the wide variety of errors contained in the data. This took numerous forms from strings being entered in fields intended to be numerical, to invalid factor levels being added to categories with specific possible inputs. We have summarized the data cleaning operations that were performed below:

* Removed empty columns
* Dropped rows with invalid main category
* Dropped rows with invalid currency
* Removed projects with a state of ‘undefined’ (they were never finished and therefore could never receive funding)
* Updated invalid ‘country’ or ‘name’ values to ‘none’
* Converted invalid dates to '1990-01-01 01:00:00’
* Split datetime columns into year and month
* Dropped old datetime columns
* Removed entries with strings in 'backers’
* Removed entries with strings in ‘usd\_pledged’
* Engineered dummy variable to indicate project success

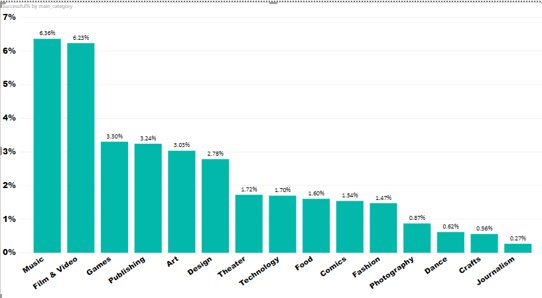
After data cleaning was complete, we were left with 374373 rows (of 378648 rows originally), 98.8% of the initial dataset.

# DATA ANALYSIS

We performed an initial descriptive analysis to get general insights from our dataset. For each of the descriptive analysis, we took the percentage of Successful Projects instead of their count so as to gain a better understanding regardless of the distribution. Our results are as follows:

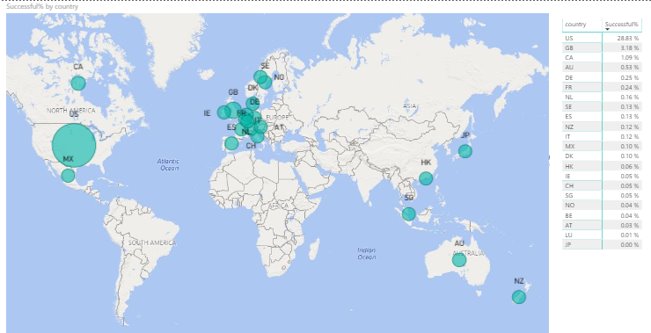
1) Percentage of Successful Projects by Main Category

The following bar plot shows the percentage of successful projects in each main category. We can see that the most successful category was Music, followed by Film & Video, Games, Publishing, and Art.



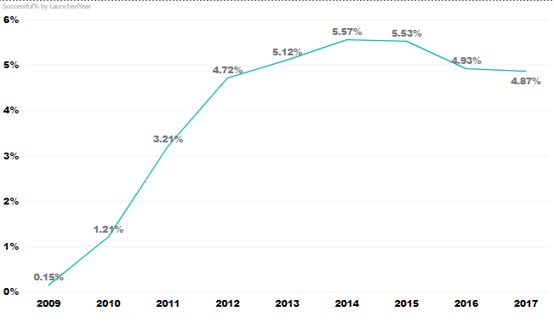
2) Percentage of Successful Projects by Country:

The following visualization shows the percentage of successful projects by the country they were launched in. The sizes of the circles show their relative percentage. A larger radius implies greater percentage of successful projects in that region. Hence, we can see that USA, followed by Great Britain and Canada, has the largest percentage of successful projects.



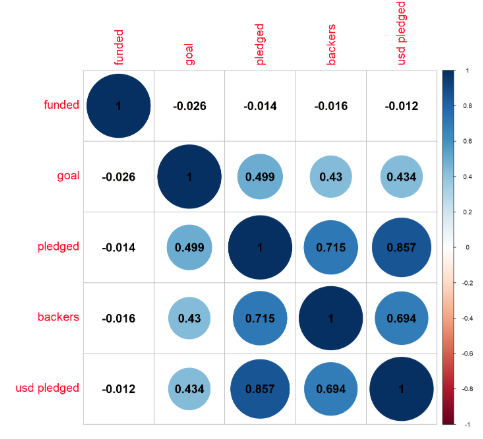
3) Percentage of Successful Projects by Launch Year:

Following graph shows the percentage of successful projects by the year in which they were launched. It is understandable from the following graph, that the initial years were the foundation years of the KickStarter website. Over the years, the percentage of successful projects has increased.



4) Correlation Matrix between the target variable (funded) and other variables:

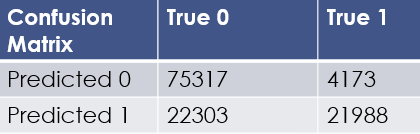
Following visualization shows the relation between the target variable Funded (yes/no) and other variables such as goal, backers, usd pledged, and pledged. Interestingly, we can see that the target variable is inversely proportional to each of these variables which makes sense too. As there is more cost involved, or more money asked for by a project, the project is more likely to not receive the necessary funding. Also, if a high amount was pledged by the crowd, a higher number of backers is observed.



For further analysis we performed logistic regression and association rule mining. We performed following prediction modelling on the important features:

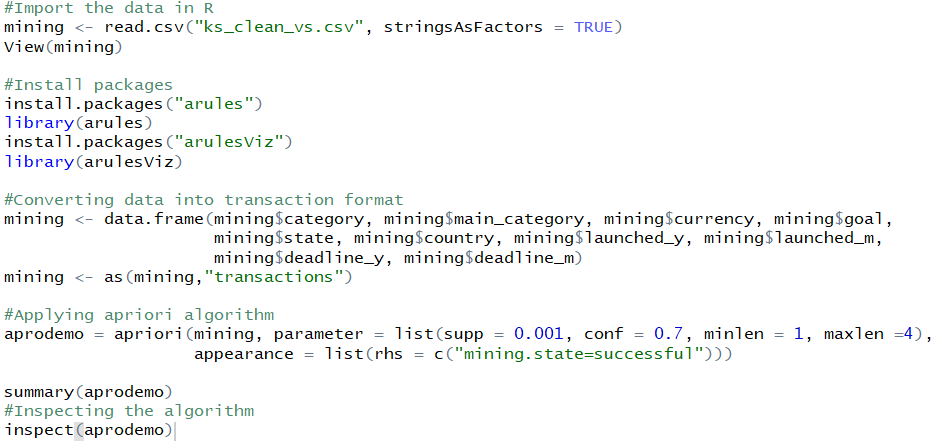
**Logistic Regression**

Utilizing a 2 to 1 split between test and training data, we generated a Logistic Regression Model which resulted in an accuracy of 78.6%, a precision of 84.0%, and a recall of 49.6%. The confusion matrix for this model is as follows:

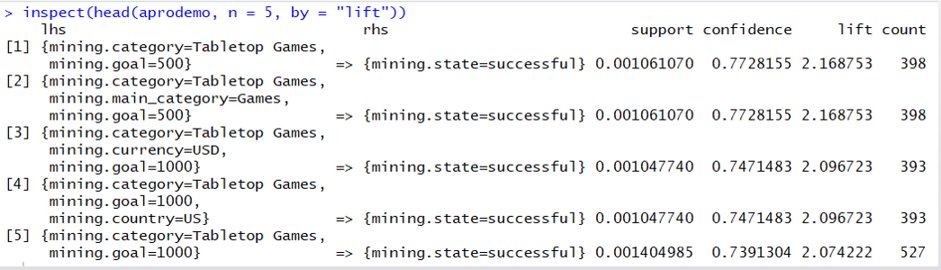


**Association Rule Mining**

We used association rule mining to find the attributes that contribute to Kickstarter project success. This was done by generating rules with a support of 0.001 and confidence of 0.7, which resulted in 90 rules.



Also, the following are the top 5 rules:

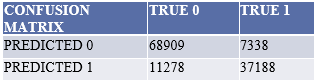


The following are the important attributes as generated by association rule mining:

|  |  |
| --- | --- |
| **Attribute​** | **Value​** |
| Main Category​ | Games, Theater, Dance​ |
| Sub Category​ | Tabletop games, Indie Rock​ |
| Goal​ | 500 – 1000 for Games, ​ 1000 – 3000 for Theater​ |
| Launch Year​ | 1970 – 2014 for Indie Rock and Dance, ​ 2015 – 2018 for Theater​ |
| Country​ | US for Theater, Games and Dance, ​ GB for Theater​ |

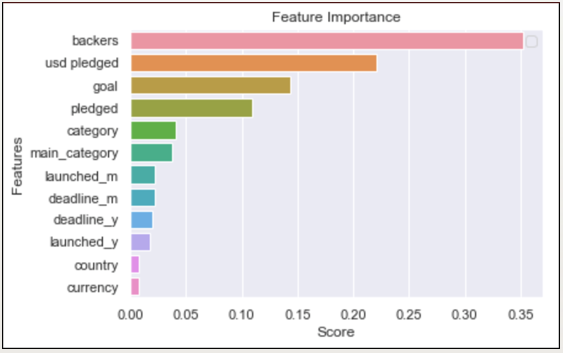
**Naïve Bayes**

Using 2 to 1 split, we passed the important variables through Naïve Bayes model for predicting whether a project received funding or not. We got the accuracy of 85.07%. Following is the confusion matrix:



**Random Forest**

Our Random Forest model was created utilizing a 2 to 1 split between test and training data, 1000 trees, and resulted in an accuracy of 93.4%, a precision of 90.7%, and a recall of 90.8% While this accuracy was already quite high, we then viewed the feature importance ranking in order to refine it further. The feature importance rankings are as follows:



As can be determined from this chart, the features ‘backers’, ‘usd pledged’, ‘goal’, and ‘pledged’ all have significant impact on the outcome of the project while the rest are rather negligible. With this information in mind, we then removed all variables but those previously mentioned. The results from this refined model were an accuracy of 96.1%​, a precision of 94.4%​, and a recall of 94.7%.

**Model Comparison**

We observed that Random Forest gives us the highest accuracy on the data set.

|  |  |
| --- | --- |
| **MODEL NAME​** | **ACCURACY​** |
| LOGISTIC REGRESSION​ | 78.6%​ |
| NAÏVE BAYES​ | 85.07%​ |
| RANDOM FOREST​ | 96.1%​ |

# CHALLENGES

1) We could not run SVM model on our dataset as the dataset was huge. We tried partitioning the data in the ratio 2:1 and even used the smaller part for SVM modelling, but the model ran for 40 minutes and crashed.

2) On using all the variables in Naïve Bayes for predicting the Funded target variable, we faced overfitting of model. So, we dropped the columns ‘launched’, deadline days, and ‘state’ in modelling.

# CONCLUSION

After performing analysis on the data using various algorithms, we came up with the following business suggestions for ensuring success in launching Kickstarter projects:

1. Setting goals between **$1000 to $3000**for **Theater** will contribute to project success​.
2. Launching projects in the categories **Dance, Games**and**Theatre**tends to result in higher success rates​.
3. In **Gaming** category, launching projects related to **Tabletop Games**will lead to higher chances of project success​.
4. Setting a goal of **$500 to $1000**for **Games** will contribute to project success​.
5. Launching projects of **Theatre** category will be successful in **US** as well as **GB** especially when launched between **May to August.**
6. Project success can be predicted with an accuracy of 96.1%.