Kickstarter Campaigns: Analysis of Success Factors

Abilash Nair
Department of Computer Science
City University
London, GB
abilash.nair.2@city.ac.uk

***Abstract*—The coursework aims to look closely at crowdfunding analytics using publicly scraped data from Kickstarter. The analysis revolves around three main questions which seek to understand factors defining a successful campaign, the effect of money being raised on the backers as well as the campaign, and whether a predictive model with good accuracy can be built. The analysis suggests that certain main and sub-genres are strong defining characteristics of successful campaigns. Staff picked campaigns are also useful in providing visibility to a project and improving its odds. Linear regression was used to analyze the relationship between money being raised and other variables. The models proved to be inconclusive and weak with low R2 scores and high MSE. Logistic regression was used to build the predictive model using categorical variables. An accuracy of 78% and good ROC characteristics shows promising predictive ability.**

I. Introduction

Kickstarter is a website that provides a medium for crowdfunded campaigns and has successfully connected the public with many independent and creative project creators. Since its inception, one of the most difficult tasks for the public is to figure out which campaigns are feasible and which creators are credible. The money that the public pledges to the campaign creator suggests trust and a willingness to help small and independent individuals realise their vision. However, there have been several controversies surrounding a few campaigns that were fraudulent, dubious, and infeasible. Such problems arise due to the informational asymmetry that exists between the public and campaigns. The data analysis will look at a variety of important success factors that define a campaign. From location to campaign genres, the analysis will try to determine the influence of such factors on a project. Another important question is how certain factors such as the amount of money the campaign creator is trying to raise affects the public's perceptions. The analysis will also attempt to build a predictive model that can effectively capture the influences of a few variables to better predict the success outcome. A good model will help point the public to feasible and successful projects and will provide better informational symmetry between the backers and the creator.

II. Data and Research Questions

A. Data Source

The data has been sourced from Web Robots[1], a professional web scraping service that collects the data from the Kickstarter website monthly. Each monthly dataset consists of multiple csv file zipped together and in total has approximately 200,000 rows and 41 columns. Due to computational limitations for this coursework, datasets from January 2020 to October 2020 is used, although datasets from 2009 are available. Each dataset is randomly sampled, and 30 percent of each dataset is collected which is then combined into a larger dataset for analysis. The dataset requires processing as there are inconsistencies present. Fortunately, the data set is labelled and contains the state of each campaign. The dataset captures several categorical variables that will be extremely useful for building a predictive model.

B. Research Questions

The overall emphasis of the data analysis is to determine the influence of several variables on a campaign's success. What makes a campaign successful? The questions can be split into three broad parts.

1. What defines a successful campaign? Which features are important?
2. What is the influence of the campaign's monetary goal amount on backers? How does it affect the performance of a campaign?
3. Is it possible to predict a successful campaign using the metrics in the scraped data?

C. Analysis Assumptions and Limitations

When each individual dataset is randomly sampled and shrunk, the analysis assumes that the collective dataset retains the relationships between variables. Computational limitations will certainly affect the quality of the analysis. Approximately only 30 percent of the data collected from January to October is reflected in this analysis. A better sampling methodology or better computational resources can improve this analysis. However, such factors have not been investigated or made use of for this coursework.

III. Investigation and Analysis

A. Data Preparation

The data needed to be extracted from zip files and each zip file contained around 60 comma separated value (csv) files totalling to around 200,000 data points. This process was automated using python and the entire dataset was randomly sampled to be 30 percent of its original size. The sampled data was then merged into a bigger dataset. Since the data was scraped monthly, certain campaigns were included in the dataset numerous times. The duplicates were dropped, keeping the last known value.

The data needed cleaning as many values were missing, some columns were redundant for analysis, and certain rows were misclassified. Many values that were missing belonged to categorical variables. An unfortunate effect of missing data in categorical variables is that data values cannot be imputed easily. It was reasonable to omit the rows with null values as there are few compared to the size of the dataset. When the data was scraped, a few columns contained no data and were discarded.

Certain data points were misclassified. This brings focus on a very important point in this analysis. Success is measured as meeting or exceeding the goal amount. A few data points met the goal amount but was classified as a failure. Such misclassified data was omitted due to ambiguity. Since the cause for misclassification is unknown, including the points or imputing values will be inappropriate.

Outliers had to be evaluated and the goal amount was chosen as an appropriate variable for outlier detection. The goal amount was log transformed to obtain a normal distribution and outliers that deviated 3 standard deviations away from the median was discarded as shown in Fig 1. and Fig 2.

B. Data Derivation

Many data features had to be transformed or extracted from the existing columns due to either formatting issues or masking issues. The goal amount, a critical variable, was provided in different currencies. To standardise all, the conversion rate to US dollars (USD) was multiplied with the goal amount to get a standardised variable. To measure the performance of the campaign, a new column ‘raised_cap’ was required to capture how the campaign exceeded or receded the goal amount. The ‘raised_percent’ column captured the percentage difference.

Many columns were improperly formatted. Converting true or false into binary equivalents was needed for some columns. Time data included in the dataset was captured as UNIX timestamps, requiring conversion into more readable time and date formats. From the converted time columns, the launch duration and creation time of each campaign was derived. Genre and location data were mixed with other string data requiring regex data extraction techniques. The location, main genre, and sub-genre columns were extracted using this technique.

C. Model Construction

The dataset contained numerical and categorical data which prompted the use of different modelling techniques to best fit the type of data.

Linear regression was chosen to analyse the relationships between numerical variables, particularly the goal amount, number of backers, raised capital, and launch duration. Linear regression works best for numerical data that exhibits linearity and normality. During data exploration, many numerical columns exhibited extreme skew, especially to the left, or bimodal distributions. The numerical variables were log transformed to obtain normally distributed data. The dataset was not split into test and training data sets as the objective was not to predict but to understand the relationship. A scatter plot captured the regression line over the data points for analysis.

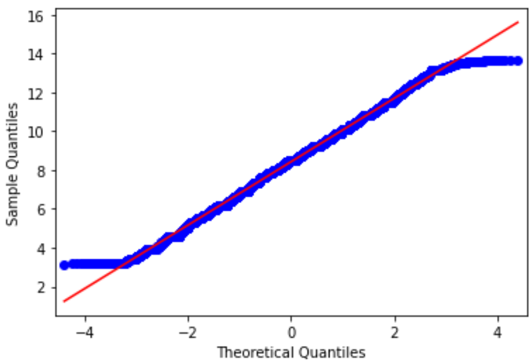

Fig 1. QQ Plot of Normalized Goal Outliers Removed

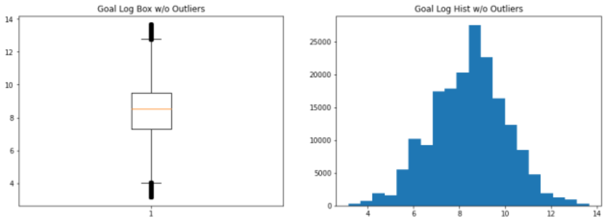

Fig 2. Normalized Goal Plots Without Outliers

Logistic regression technique was used to build the predictive model using categorical data as it works best for binary distributions. Since the labels were already known, clustering techniques were not required to label the data. The model accepted the location, main genre, sub-genre, and staff pick data to predict the success of a campaign. The idea of only accepting these variables in the model stems from the fact that these variables are readily available to a backer when the project is live. One-hot encoding technique was used to obtain dummy variables that transformed the categorical columns into binary distributions. As a result, over 140 columns were generated, increasing the complexity of the dataset. The data was randomly split into test and training data for evaluating its effectiveness. The test subset contained 33 percent of the original dataset. Iterations of the model had to be increased to 1000 due to convergence issues.

D. Model Validation

To validate the models, quantitative statistical metrics will be used to evaluate its effectiveness along with plots displaying distributions and other metrics. For the linear regression model, the R2 metric will be used to evaluate whether enough variance is captured by the model. The R2 value should be at least 0.4 and ideally above 0.5. Apart from R2, the residuals will be calculated and plotted as a boxplot and histogram to evaluate its properties. The histogram should ideally be normally distributed around a median of zero. A QQ plot is plotted to check the normality of the residual. The standard deviation is calculated to evaluate the tightness of the model. Also, the mean square error (MSE) is calculated for the model so that the accuracy of the model can be ascertained.

For the logistic regression model, the precision, recall, and f1 values are computed for both the success and failure states. The accuracy of the model is also computed which should be at least above 50 percent and ideally as high as possible. The precision metric for the success and failure states will give an estimate of how well the model predicts each state. Next, a confusion matrix is plotted to visually capture the false positive and true negative observations. The coefficients of the model are then calculated to ascertain the influence each variable has in the model. Finally, the Receiver Operating Characteristic (ROC) score is calculated, and the ROC curve is plotted along with a random classifier line.

IV. FINDINGS

A. Important features of a successful campaign

First, the proportion of successful projects by country is plotted. Hong Kong, Japan, Singapore, and Great Britain have high success rates. Sub-genres are ranked according to the proportion of most successful campaigns within each sub-genre as shown in Fig 4. A similar analysis for main genres is given in Fig 5.

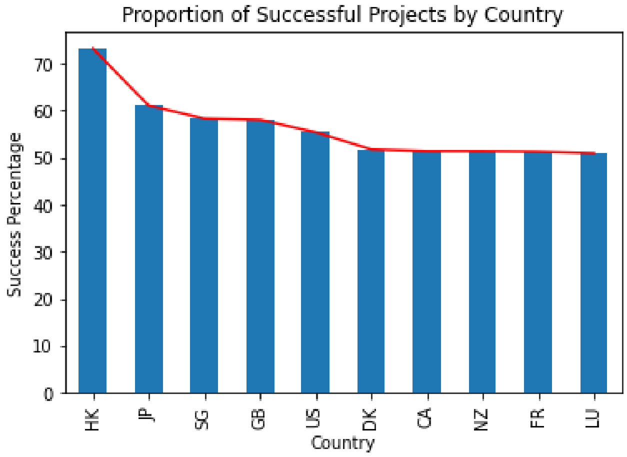

Fig 3. Success percentage by country

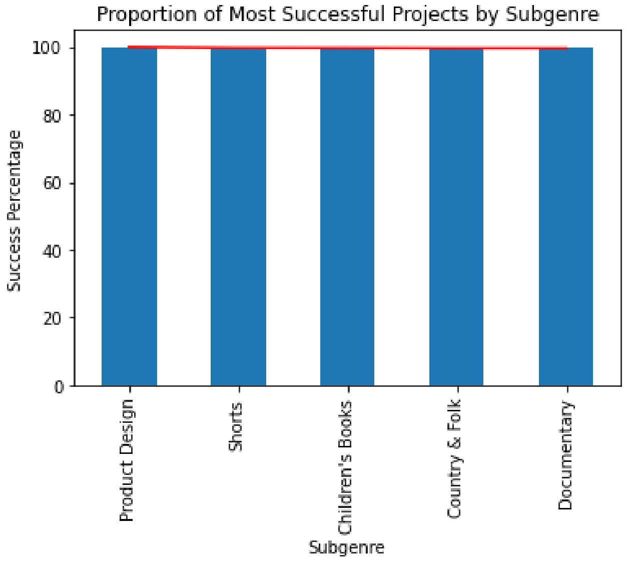

Fig 4. Top successful sub-genres

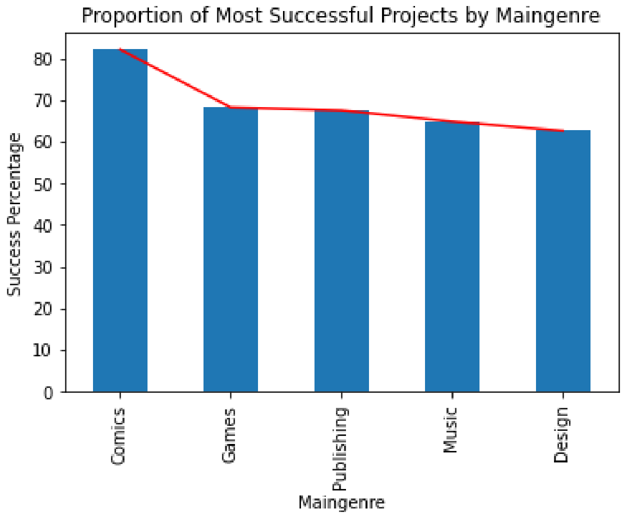

Fig 5. Best performing main genres

Staff pick is another categorical variable that shows interesting results. The success rate of staff picked campaigns is about 20 percent average. Another category that indicates how well a campaign has performed is measured by the raised capital. Fig 7. and Fig 8. is plotted by main and sub-genres.

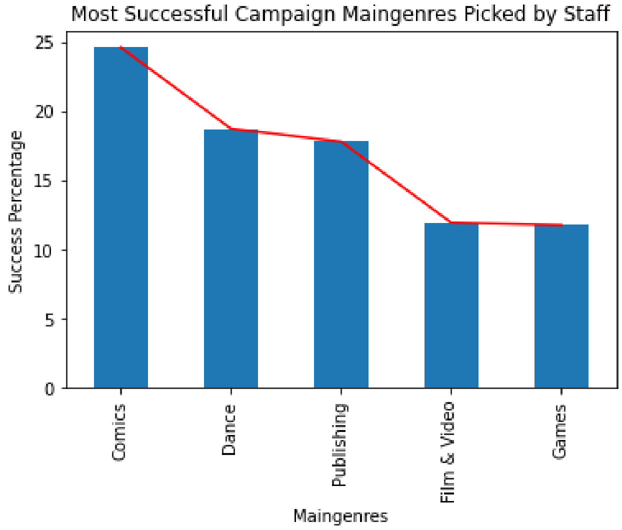

Fig 6. Best staff picks by main genre

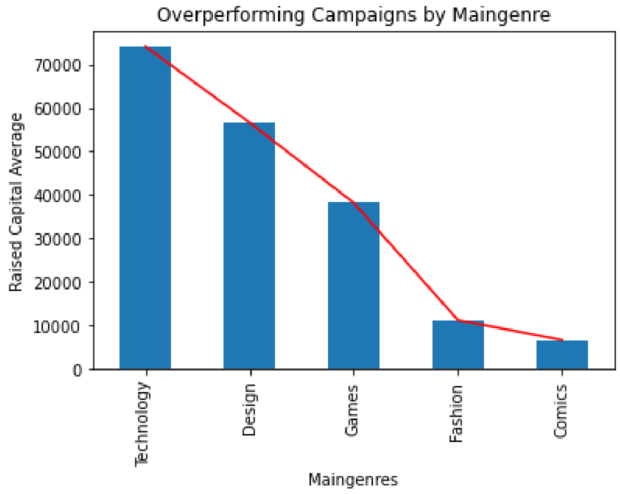

Fig 7. Overperformance by main genre

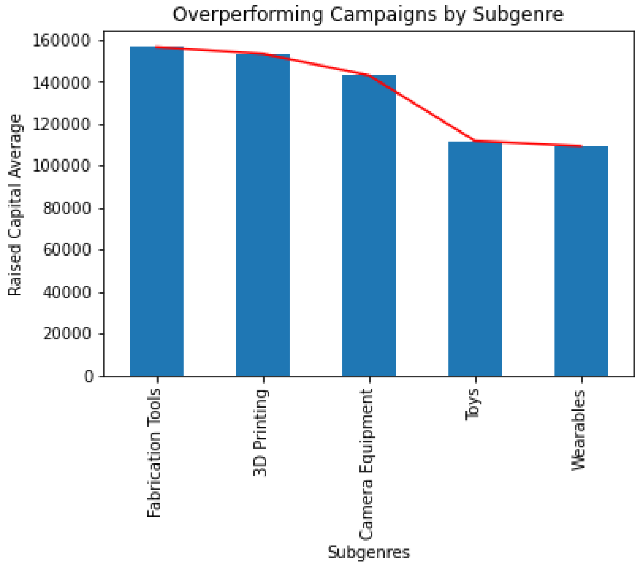

Fig 8. Overperformance by sub-genre

B. Correlation and Regression Findings

A correlation plot gives a heatmap of the correlations between variables as shown in Fig 9. The regression analysis could only be carried out for two variables and is given in Fig 10. And Fig 11. The classification is given in Table III. The confusion matrix and the ROC is plotted in Fig 12. and Fig 13.

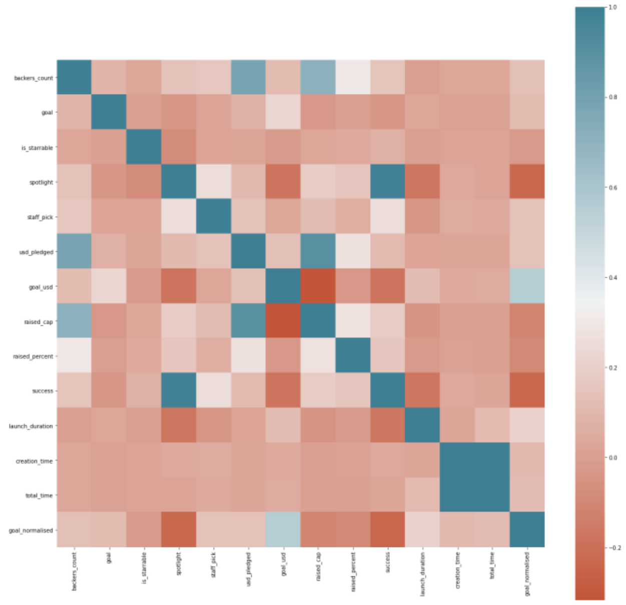

Fig 9. Correlation Heatmap

TABLE I. R2

	Backers	Raised Capital	Pledged	Launch Time
Goal	0.475	0.215	0.03	0.06

TABLE II. MSE

	Backers	Raised Capital
Goal	5.278	1.545

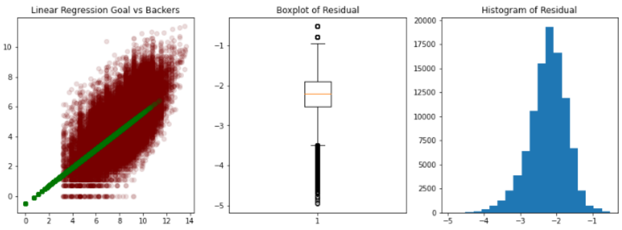

Fig 10. Goal vs Backers

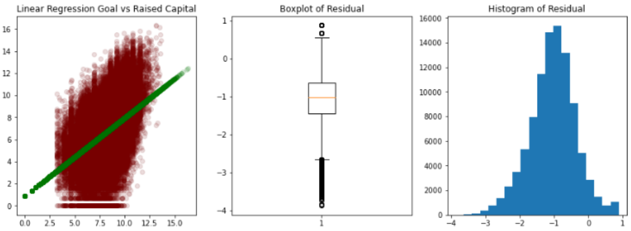

Fig 11. Goal vs Raised Capital

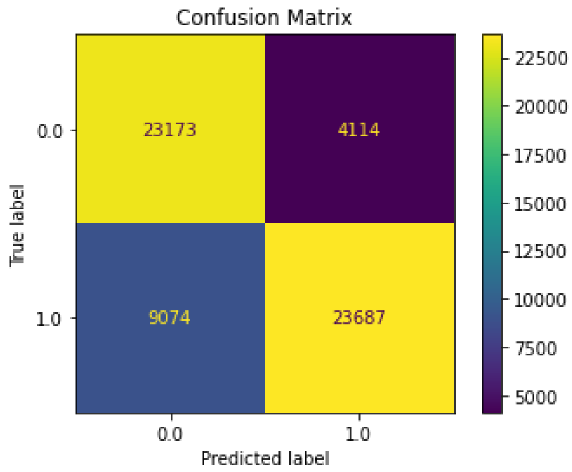

Fig 12. Confusion Matrix

TABLE III. CLASSIFICATION

	Precision	Recall	F1 Score
0.0	0.72	0.85	0.78
1.0	0.85	0.72	0.78
Accuracy			0.78

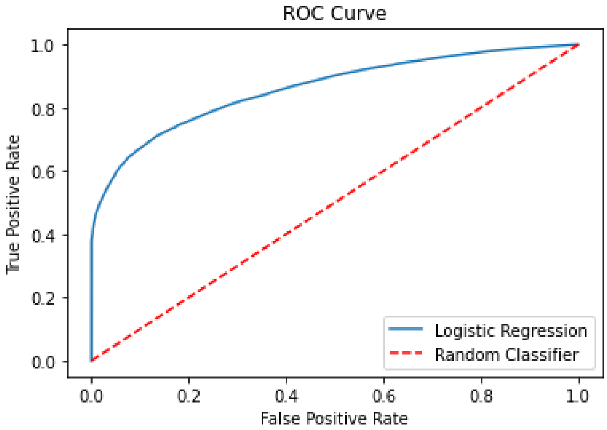

Fig 13. ROC Curve for Regression

V. REFLECTIONS AND FURTHER WORK

With regards to the first question, the most important features seem to be main and sub-genres. Exploratory analysis and logistic regression have shown the influence that some categories of genres have on success. Staff picks also seem to influence success. The coefficient for staff pick is 2.23 which is a pretty good indicator of its influence. Location seems to play a role, but its effect is diminished. Over performance of a campaign, given by raised capital, pointed to certain main and sub-genres not adequately noticed. For instance, fabrication tools and 3D printing sub-genres tended to over perform.

Linear regression proved to be quite the failure and gave inconclusive models to the second question. Correlation heatmaps revealed that not many variables had interesting correlations. The R2 coefficient was seriously low across the chosen variables. High MSE proved that models had limited accuracy. The limitation proved to be the distribution of the variables themselves. Perhaps better transformation techniques could have yielded better results. Linear regression might be ill suited for this analysis as the variables are extremely skewed, are non-normal, and have low correlation. Low confidence in the models means that the second question could not be adequately answered. The python notebook lists some cautious findings.

Logistic regression proved to be quite successful in predicting successful variables. An accuracy score of 78% is decent but the model was limited by complexity. Variables such as city location and others could not be included as it would have made the model too complex. The iterations had to be increased for the existing model due to non-convergence. The ROC curve validates the model and indicates it has good predictive behaviour. Perhaps ML techniques such as random forests could be implemented to compare its effectiveness.

VI. WORD COUNT

TABLE V. WORD COUNTS

Section	**Word Count**
Abstract	142
Introduction	216
Data and Research Questions	296
Investigation and Analysis	993
Findings	272
Reflections and Further Work	294

REFERENCES

[1] Web Scraping Service. 2020. Kickstarter Datasets. [online] Available at: <https://webrobots.io/kickstarter-datasets/> [Accessed 20 December 2020].

[2] Seaborn.pydata.org. 2020. Building Structured Multi-Plot Grids — Seaborn 0.11.0 Documentation. [online] Available at: <https://seaborn.pydata.org/tutorial/axis_grids.html> [Accessed 20 December 2020].

[3] Pandas.pydata.org. 2020. Pandas Documentation — Pandas 1.1.5 Documentation. [online] Available at: <https://pandas.pydata.org/pandas-docs/stable/index.html> [Accessed 20 December 2020].

[4] Medium. 2020. Logistic Regression: A Simplified Approach Using Python. [online] Available at: <https://towardsdatascience.com/logistic-regression-a-simplified-approach-using-python-c4bc81a87c31> [Accessed 20 December 2020].

[5] Medium. 2020. Building A Logistic Regression In Python, Step By Step. [online] Available at: <https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8> [Accessed 20 December 2020].

[6] En.wikipedia.org. 2020. Receiver Operating Characteristic. [online] Available at: <https://en.wikipedia.org/wiki/Receiver_operating_characteristic> [Accessed 20 December 2020].