SUCCESS PREDICTION OF KICKSTARTER PROJECT

Capstone Project Milestone Report

Abstract

This document contains the initial data story for the project

Pundir, Saurabh

saurabhpundir.data@gmail.com

Date: 05/02/2018

Table of Contents

1. Introduction

1.1. About:

It is no doubt that in current time humans are connected to each other more than ever before. This has opened the new opportunity for anyone with an idea to present it to the world. The term "Startup" was coined in 1976 but become the hottest buzzword in last few years.

In the era of information and social media, it is sometimes easy and more worth to reach millions of people ready to give a dollar then one VC to invest million dollars. This is the basic concept of **crowdfunding** which again is not a new term but got muscles in its meaning in times of Internet.

Platforms like Kickstarter provide an opportunity to person with an idea to present it to everyone who has access to the internet. Also provides an opportunity for a person with belief in someone's idea to contribute, obviously with the expectation of expected rewards on investments.

1.1..1. Formal Definition - Crowdfunding:

Crowdfunding is the practice of funding a project or venture by raising monetary contributions from a large number of people in smaller amounts. Crowdfunding is a form of crowdsourcing and of alternative finance.

The company or a person comes up with a project proposal or idea and shares its details and defines the amount they are looking to raise.

1.2. Problem:

Crowdfunding has given the opportunity to various now well know products like pebble watch, Oculus or innovative project like Baubax. As of April-2018 Kickstarter (crowdfunding platform) has raised **three billion** (3,647,430,067) US dollars from **fourteen million** people (14,547,043) for **three hundred thousand** projects using crowdfunding.

But does all three hundred thousand projects were worth investing? Do all fourteen million people have the expected rewards on investment? Were all three billion dollars given to projects worth?

In the investment world, it is obvious that there is no simple answer to above questions. But there is an opportunity to provide all clients and audience involved an assistance to take better decision driven by the power of data.

The problem in this project to be addressed is that ***"Is it possible to estimate the probability of success for the recently launched or still running project?"***

1.3. Clients and Audiences

In problem section, we mentioned terms clients and audience. It is important to understand the definition of clients and audience to explore the above problem from their perspective.

There are 3 types of entities involved in any crowdfunding project.

1. **Project Owner**: Individual or company who creates a project to acquire funding from crowdsourcing
2. **Project backer**: Individual showing interest in funding the project.
3. **Crowdfunding facilitator (company hosting website)**: Company owning the platform.

All the entities mentioned above are clients and audiences because all of them are interested in one common thing **"Success"** of the project.

1.4. What does success mean

Before starting addressing these questions it is important to define ***what is a success?***

In crowdfunding ecosystem, success simply means that the project got the asked or more amount in the desired time frame. This amount is called **goal amount.**

It has a significant impact on clients mentioned above. If the project is a success.

1. **Project owner:** They will get money for the project.
2. **Project backer:** They will get an expected reward on their investment.
3. **Crowdfunding facilitator (company hosting website):** Company will get commission or fees. Also, based on prediction company can provide support by highlighting project.

The data have label called **Status**, which have value Failed, Successful, Canceled and Live. The status successful is considered as success for this project.

1.5. Scope:

There can be various crowdfunding sources, however, this Project will be limited to projects funded using a website called **Kickstarte**r.

This project is about analyzing data available for crowdfunding platform and find out the best algorithm to predict the probability of success of newly launched crowdfunding project. The project should be launched in last twenty-four hours because the project which has passed more time might have received some amount which is not a variable in this model.

1.6. Question Client(s) really care about

All the above clients and audience have a concerning question...

"What is the estimated probability of this recently launched project of being successful?"

2. Data Wrangling

The data wrangling section involved getting data and extracting useful columns or variable from the files for further processing. The individual section explains the approach applied for the cleaning and wrangling data to extract data for all further purposes.

2.1. Getting File

Data is available as CSV file on webrobots.com. The data is available based on a monthly basis as a zip file. Each month file size is 140 MB.

2.2. File Content:

Each months zip file contains nearly 40 CSV files of approximately 19 MB size each. Each CSV file contains around 4000 project information. There are 32 distinct fields for each project. The fields detail mentioned in data dictionary section.

2.3. Choosing Processing Interval:

The approach is to start with latest month file available (Jan-2018) and keep adding previous months data. Reading and consolidation of multiple month data showed that there are not many new projects added progressively in each month files. The latest file contains most projects from previous months.

2.4. 3 Step Approach

Data wrangling performed using three stages process.

1. Extracting data for each month.
2. Data Cleaning and consolidating into a single file.
3. Storing data for further processing.

2.5. Extracting Data:

There are multiple files in each month's folder. For load sharing purposes, each month folder processed separately. Reading files inside the folder is performed using "*os.path"* module. Each data extracted from individual file gets appended to a data frame for a month.

For an initial understanding of data the information on rows, columns, and data type of the data frame are explored. Each month has 170K distinct projects.

The code related to this part is in Capstone_DataWrangling_I_ReadFiles.ipynb file.

2.6. Cleaning Data:

After reading monthly data file into a data frame. The fields mentioned above were extracted further by looping each data frame row. The JSON data is further extracted using JSON normalize. The date fields are converted using date time conversion from Unix time to UTC date ("%Y-%m-%d %H:%M:%S") format.

The new temporary data frame is built in process for above extracted and converted data. This temporary data frame is merged using join and project id as a common key into main data frame after all subfields value is extracted.

The list of subfields extracted from JSON columns

- **creator**
  - name
  - is_registered
  - user id
- **location**
  - city
  - state

    - type (district, town, city)
- **Category**
    - Project category type

Missing Data 1(Location):

Location information is not available for all rows in the data frame. So, this was handled comparing location with the null & empty value before extracting. Every data frame around 2K records will not have any location value.

Missing Data 1(user info):

The field name is_registered under user info is not available from and before May 2017. This field is not relevant and hence filled with empty if not available.

Code related to this part is in Capstone_DataWrangling_I_ReadFiles.ipynb file.

2.7. Storing Data:

Challenge:

The challenge in above process of cleaning and extracting data for each month in the data frame is the processing time. Processing each month files take 8-10 hours on medium configuration machine (laptop) available. Processing 6 months data took almost three days to complete.

After cleaning data and creating final data frame for the monthly file. The file is stored on the hard disk to avoid re-running the time-consuming process. The pickle object sterilization is utilized to store each month data frame. After all monthly files are available, as individual files, a single pickle file is created with all data appended.

This file is unpickled and processed in further steps. The process of unpickling and making the file available is under 5 minutes.

This approach provided the benefit of processing the data once and then utilizing it further at a faster pace.

Further Work:

The Hadoop and another related map -reduce technology can be utilized to make this process better. But currently, this was not the scope of the project.

2.8. Merging Datasets:

The pickle files created for each month mentioned in above process are utilized to read and create a final data frame. The two data frames are created from this data.

All unique projects:

Considering the latest months data frame (July) as the base data frame, the records from all previous months are appended to create a data frame considering **unique** projects. At the end of the process, we have unique **172892** projects available. This data frame will be utilized further stages of the project.

The code related to this part is in Capstone_DatWrangling_II_Consolidate_Files file.

2.9. Data Dictionary:

a) Kickstarter.com:
Data: https://webrobots.io/kickstarter-datasets/
The dataset is available for March 2016 for every month. This data is collected from web crawler created by the company. The latest dataset contains following fields.

1. Id: the unique id
2. Photo: Info for all photo associated with the project.
3. Name: name of the project
4. Blurb: intro & detail
5. Goal: Amount needed to be raised
6. Pledged: Actual amount raised
7. State: Current state of the project (canceled, failed, successful)
8. Slug: a brief description
9. Disable_communication: communication allowed to the creator
10. country: country of campaign origin
11. currency: the currency of campaign origin
12. currency_symbol: currency symbol of campaign origin
13. currency_trailing_code: TRUE, if conversion needs to happen in a user currency
14. deadline: UNIX timestamp for a project deadline
15. state_changed_at: Unix timestamp
16. created_at: UNIX timestamp for the project created
17. launched_at: UNIX timestamp for the project started
18. staff_pick: TRUE, if staff picked
19. backers_count: Total user backed the project
20. static_usd_rate: USD conversion rate from original currency
21. Usd_pledged: pledge amount in USD after conversion:
22. creator: details like username for the creator
23. location: location of the project
24. category: the category of the project
25. profile:
26. spotlight: feature spotlight available or not
27. URLs: Url info for a project
28. source_url: seems like URL for the category

3. Exploratory Data Analysis

The section involved exploring the various relationship between columns or independent variable in the context of the state of the project.

3.1. Data wrangling:

The data needs to be further modified to get some variables in format to provide it to data visualization library like Matplotlib and Seaborn.

The information extracted in this step

- Converted Staff Pick to category type
- Extracting the no of days between project creation and project ended
- The extracted ratio between pledged and goal amounts
- The extracted ratio between the pledged amount and no of people backing project.
- Camel casing the state column so it can be used in the axis label

The code related to this part is in Capstone_DataWrangling_III_BfrDataStory file.

3.2. Exploring data relationship

The various columns are visualized to get the data representation of their relationship with the successful and fail state.

The code related to this part is in Capstone_DataStory file.

3.2..1. What is project count as per project status? What is successful and failed ratio?

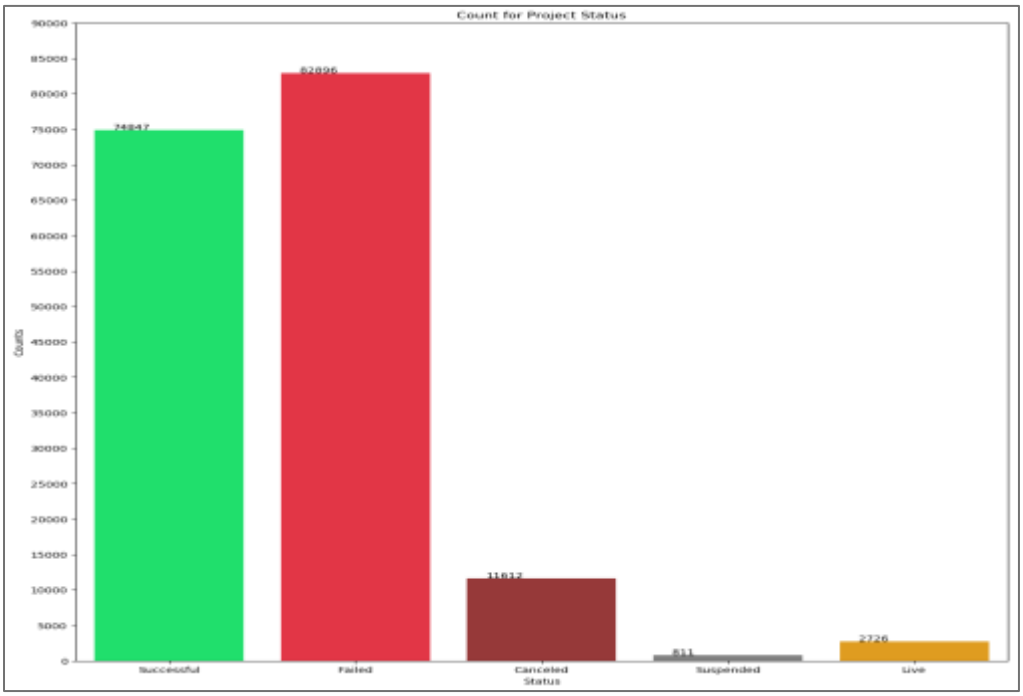

Figure 1: Project count as per project status

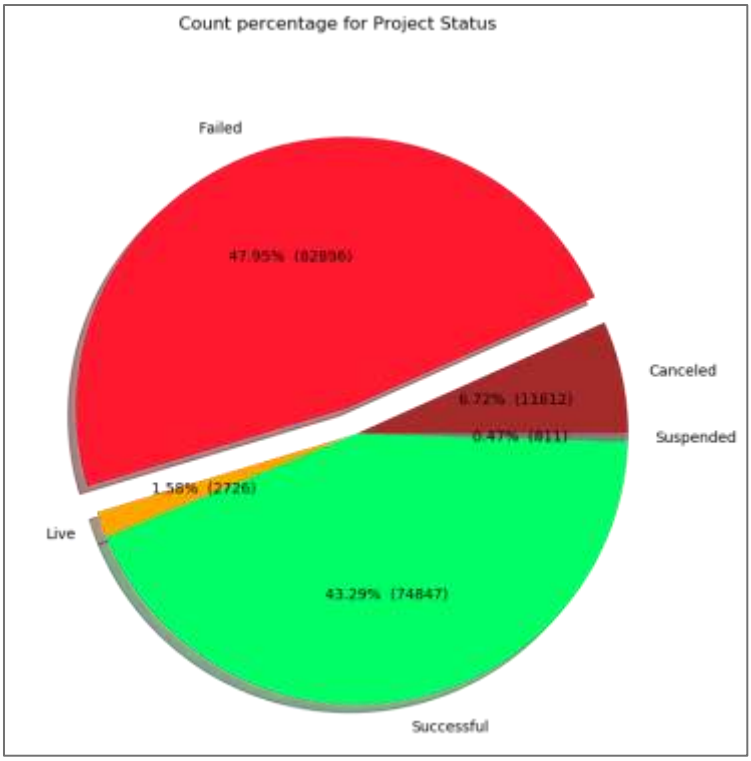

Figure 2: Project status ratio

The above displays that count of the project with each status available. The most of projects are either Successful or Failed. The other statuses are very small as compared and also not be important for the project. So they won't be mentioned much going forward.

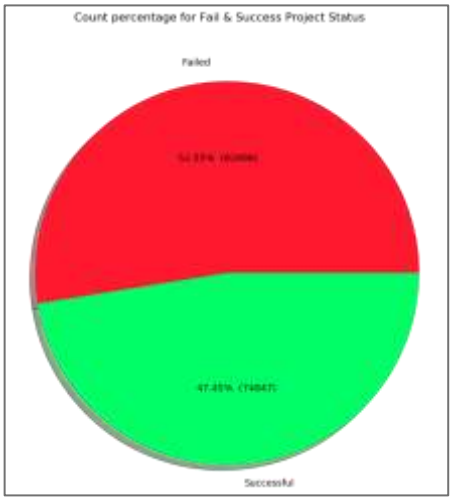

Figure 3: Ration of Failed and Successful project

In above diagram, it is very obvious that there is a more failed then successful project but still, the ratio is very close and we have data for both the states in almost equal proportion

3.2..2. Is there relationship with the year, the month of a launch date for successful or failed?

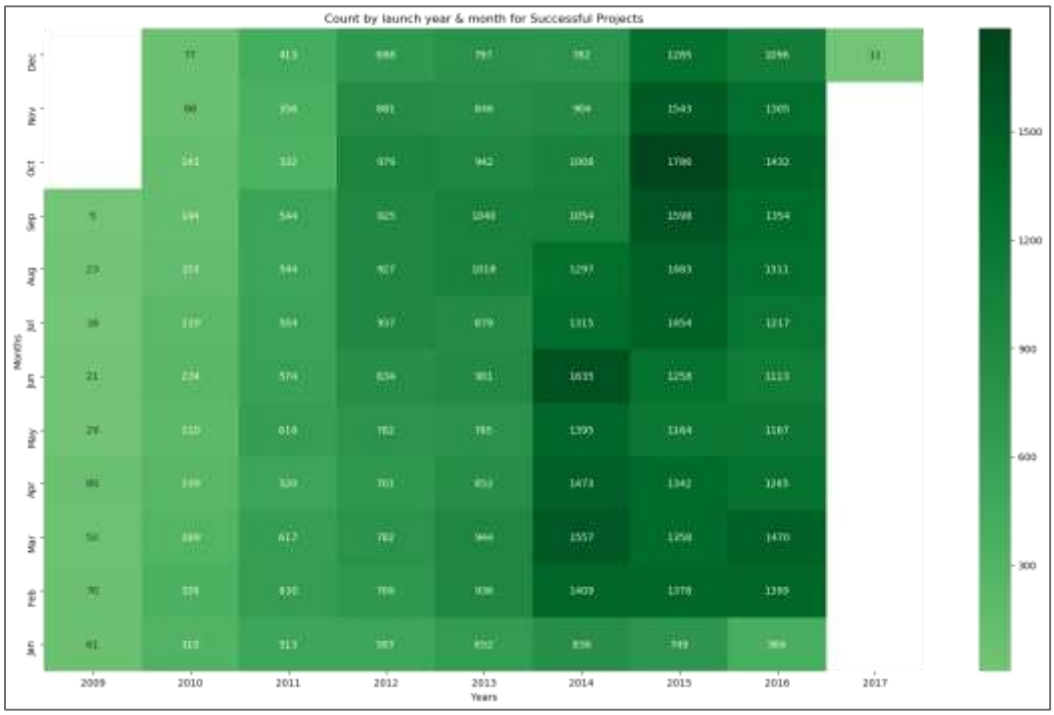

Figure 4: Heat map for year and month for successful

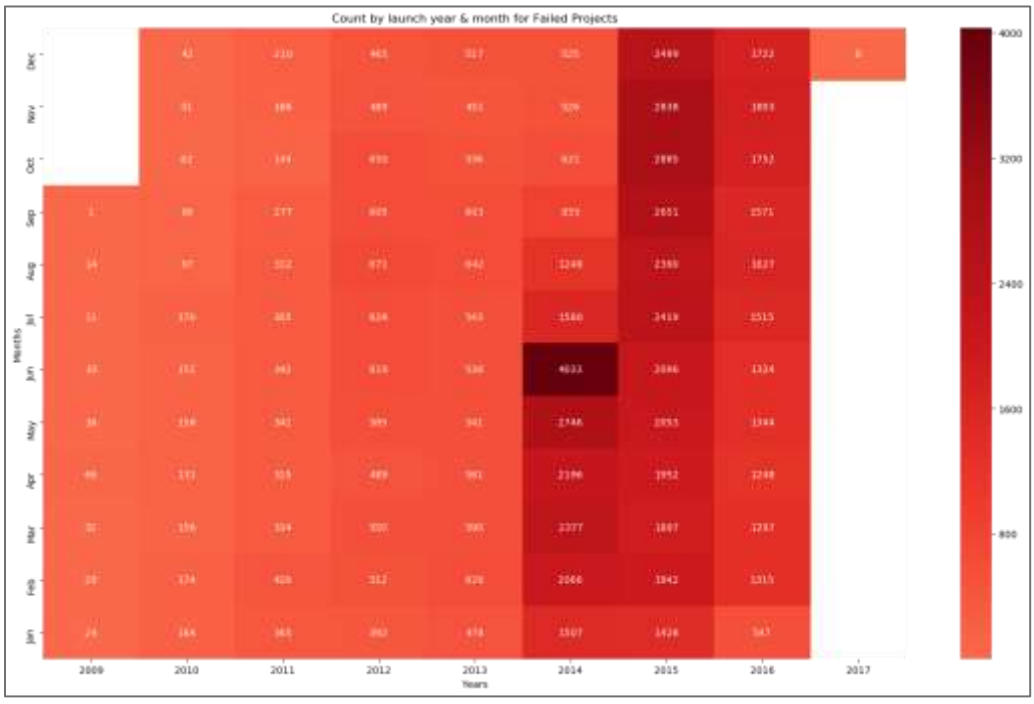

Figure 5: Heat map for year and month for failed

Figure 6: Month-wise successful and fail project

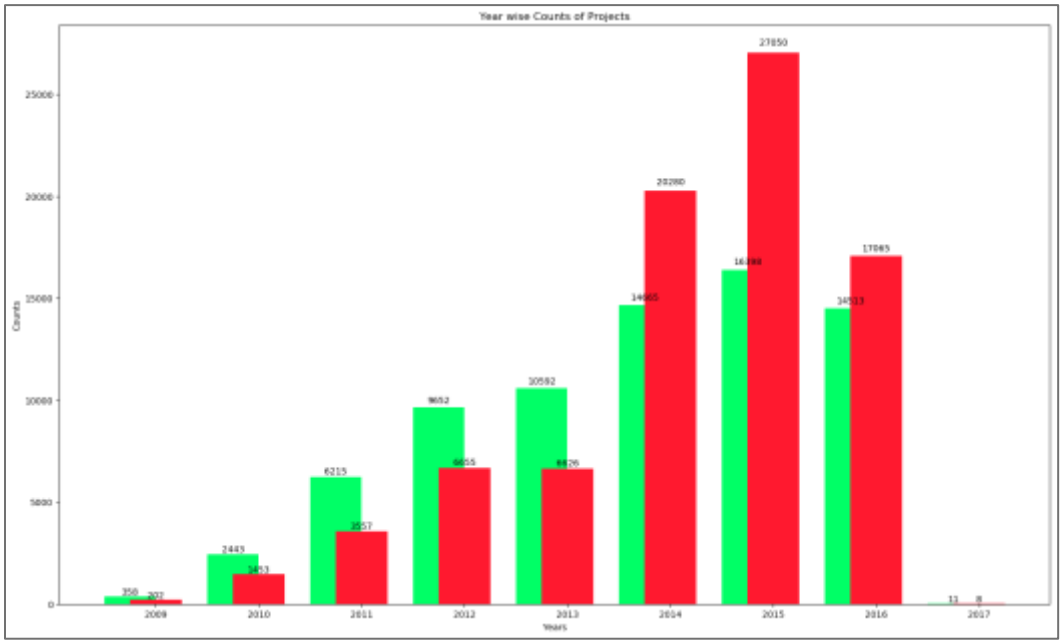

Figure 7: Year-wise successful and failed project

The above diagram shows a number of successful and failed projects in a relationship with year and month. The influence of year and month on the state is explored and following observation can be concluded.

- The projects are launched evenly all months around with few months have spiked in numbers
- Every month more projects are failed but still, the counts are close

- Kickstarter was launched in 2009. There was steady growth in a number of projects started till 2015. This is not very sure as data is only available until 2017
- Not sufficient data available from 2017 and onward
- There is an indication that Kickstarter is getting less popular and interest is on the decline
- There is an indication that Kickstarter more project failed as year's progress. This may be due to popularity is on the decline or the data is not sufficient. It should not be concluded.

3.2..3. Is there relationship with goal amount to the success or failure of the project?

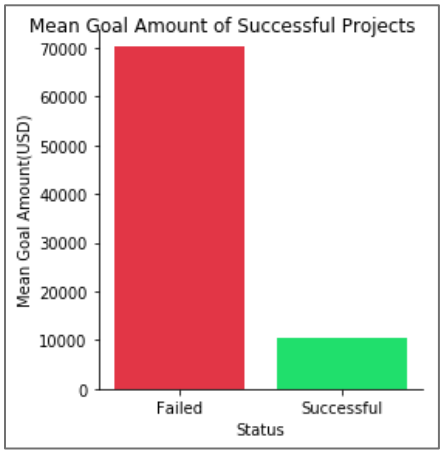

Figure 8: Mean goal amount of Successful and failed

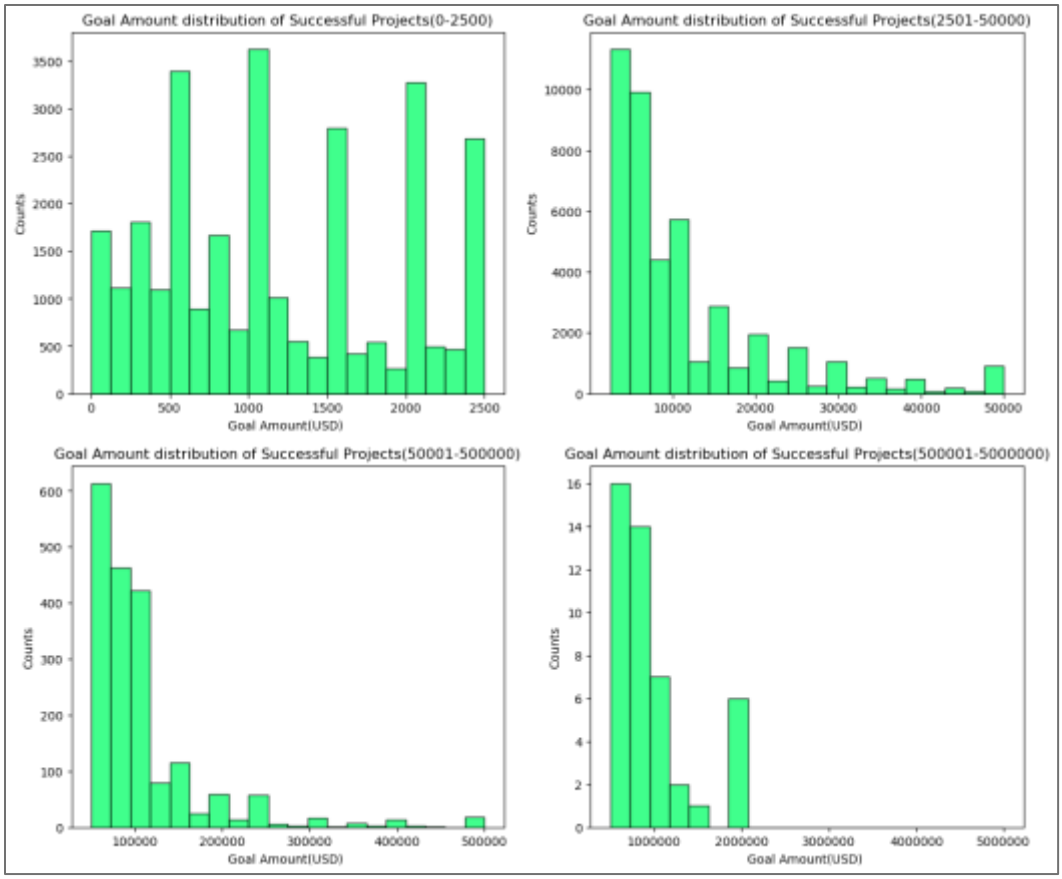

Figure 9: Goal amount distribution of Successful

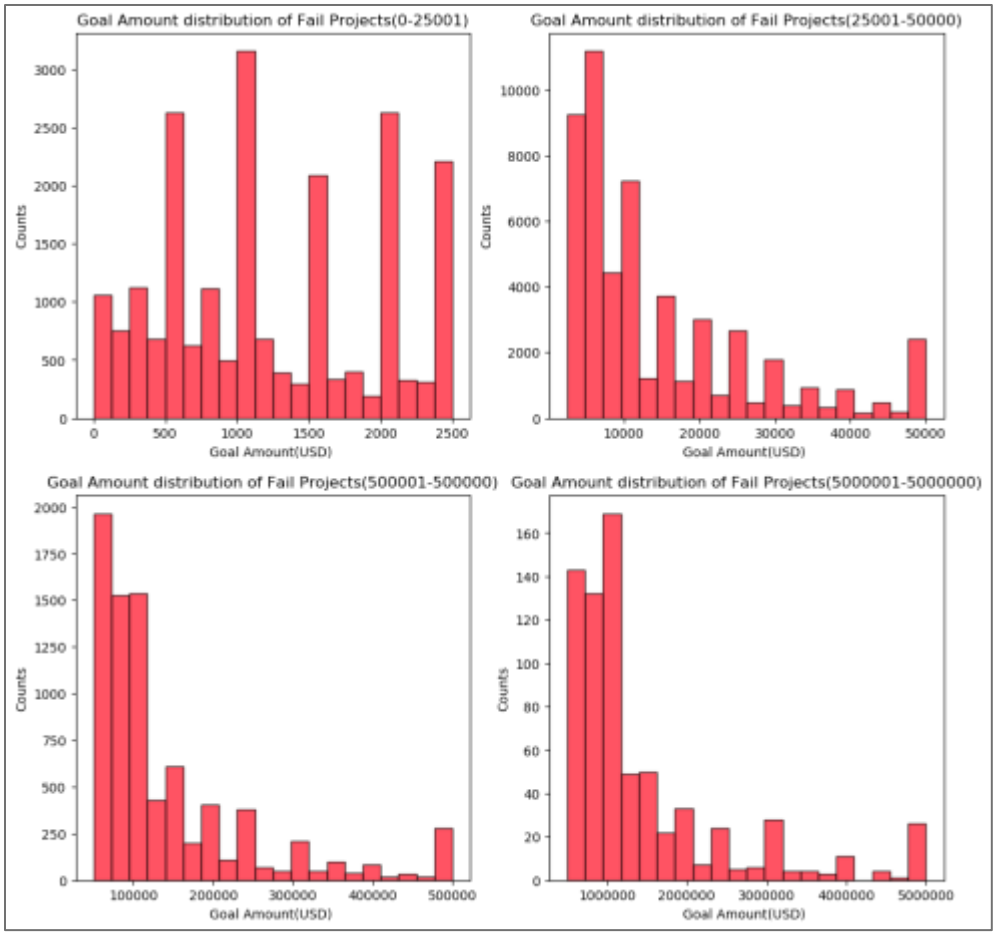

Figure 10: Goal amount distribution of failed

The above diagram shows a number of successful and failed projects in a relationship with goal amount. The goal amount is important variable and it is necessary to understand the pattern based on goal amount.

- The average goal amount of failed project is seven times higher. It indicates project with higher amount asked fails more. No surprise but still interesting.
- Most successful projects have goal amount in the range of 2500-5000. Most failed projects have goal amount in the range of 5000-7000
- The interesting observation is that from amount 0 - 4000 USD the difference between a number of the fail and successful project is very similar. It is in higher amounts the difference display impact.

3.2..4. Is there relationship with pledge amount in the success or failure of the project?

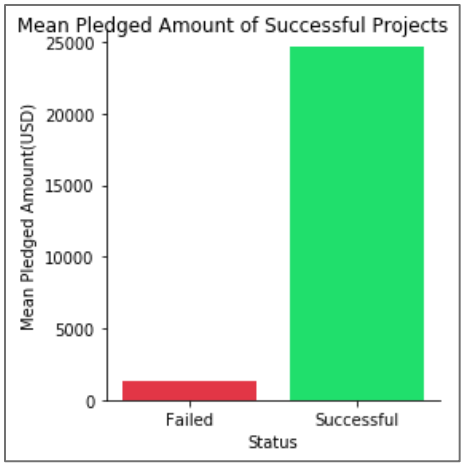

Figure 11: Mean pledged amount of Successful and failed

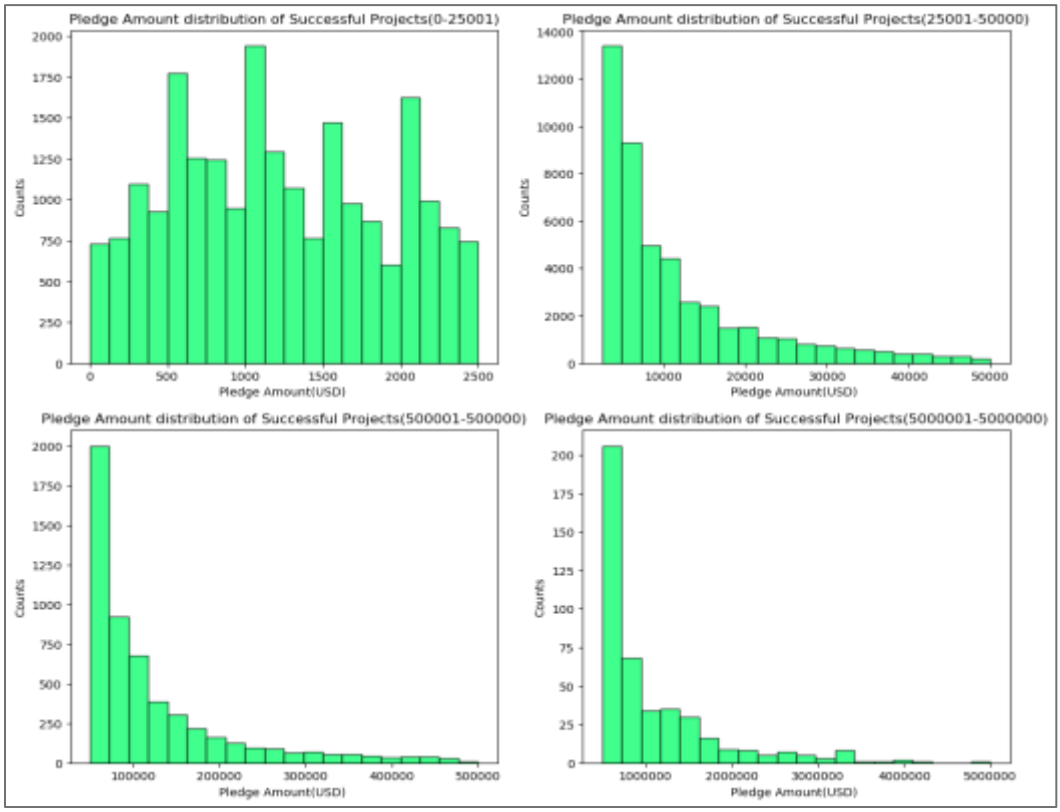

Figure 12: Pledged amount distribution of successful

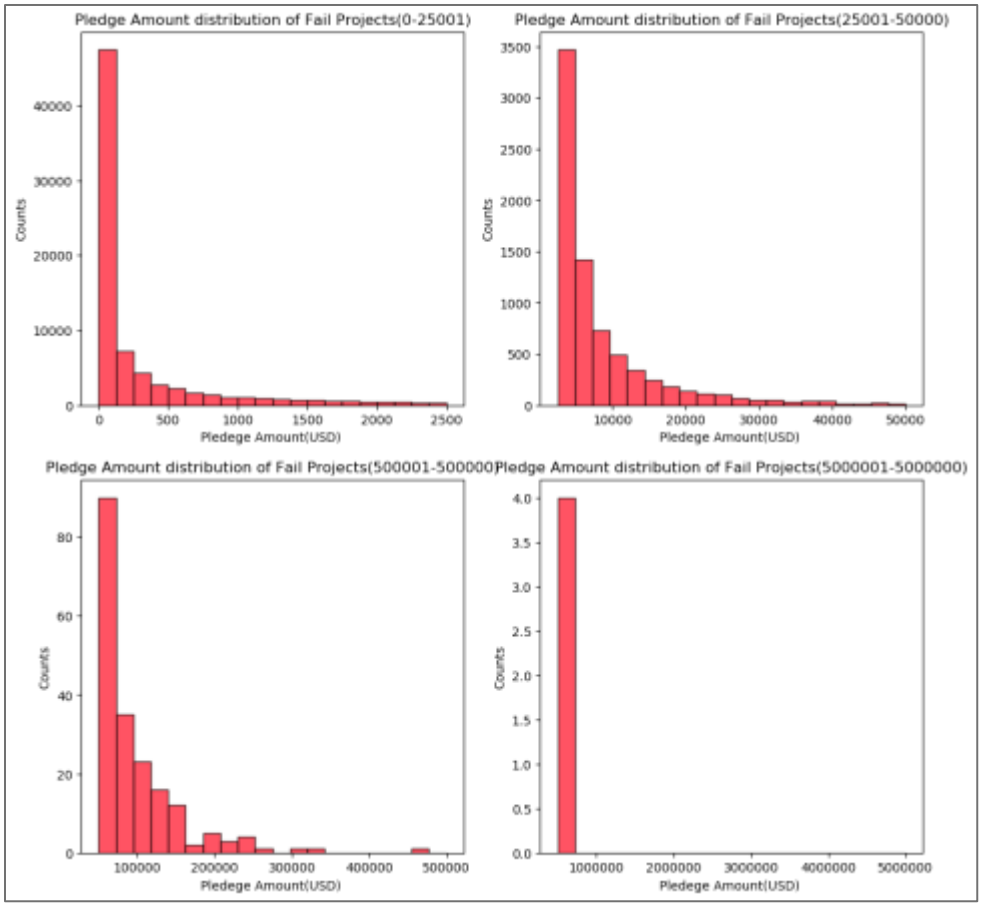

Figure 13: Pledged amount distribution of failed

The above diagram shows a number of successful and failed projects in a relationship with the pledged amount. The pledged amount is variable not included for modeling but it is necessary to understand the pattern based on pledged amount.

- The average goal amount of successful project is twenty-five times higher. It indicates project with higher amount asked fails more. No surprise but still interesting.
- The pledge amount and no of failed also have a strong negative exponential relationship.

3.2..5. Is there relationship with a pledge to goal ration in the success or failure of the project?

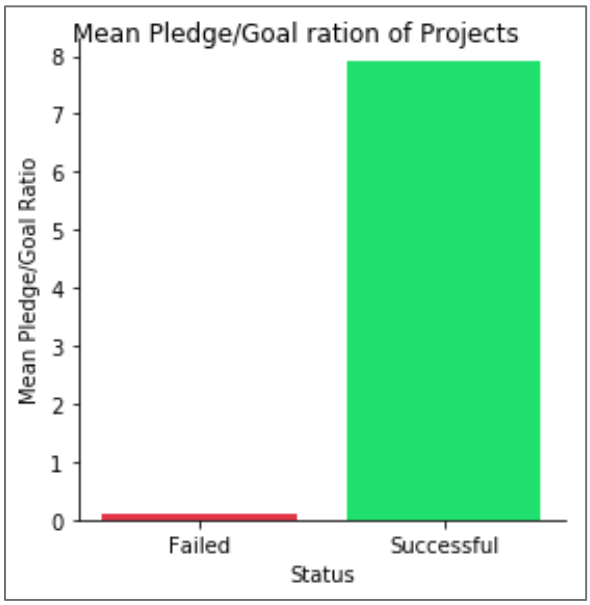

Figure 14: Mean of Pledge/Goal amount ratio

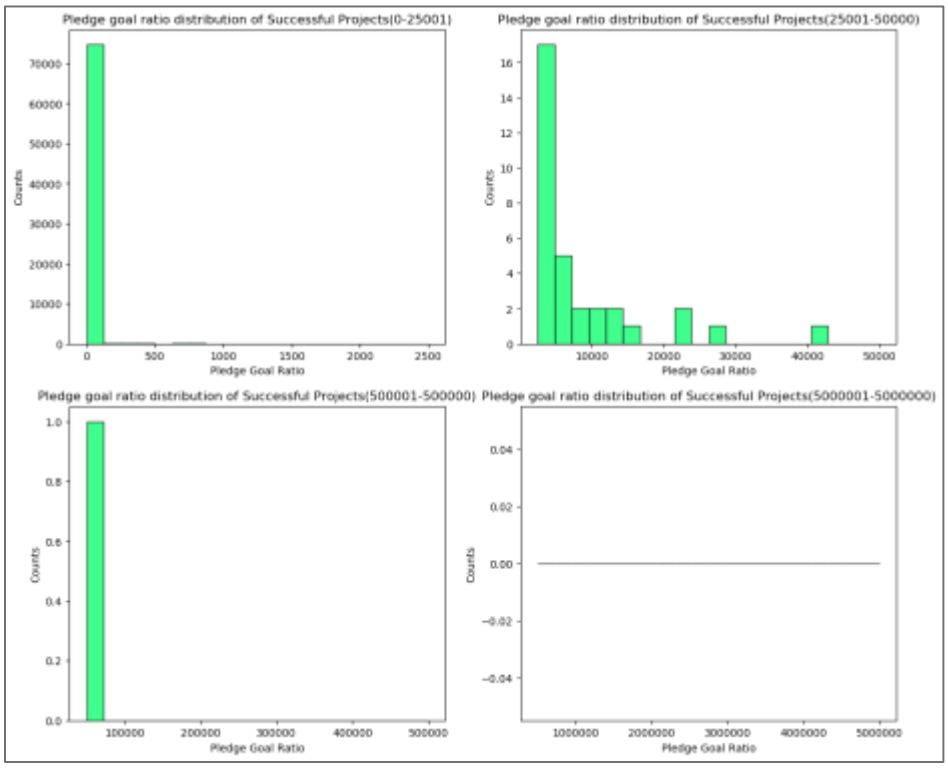

Figure 15: Pledge amount to goal amount ratio for successful

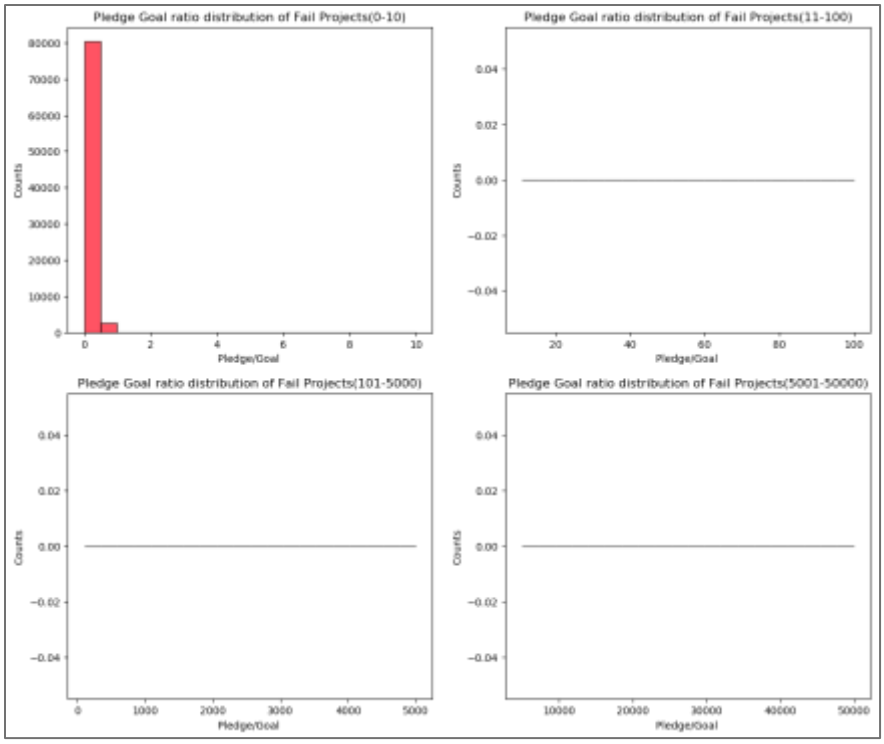

Figure 16: Pledge amount to goal amount ratio for failed

The above diagram shows the ratio of successful and failed projects in a relationship with the pledged amount by goal amount. This gives an indication of ratio the projects get than expected.

- The most successful project just achieves their goal amounts. Very rare projects get funding higher than goal amount.
- Almost all failed project miss their goal amount by more than 50%. Very rare they fail close to their targeted goal amount.

3.2..6. Is there relationship with no of backers in the success or failure of the project?

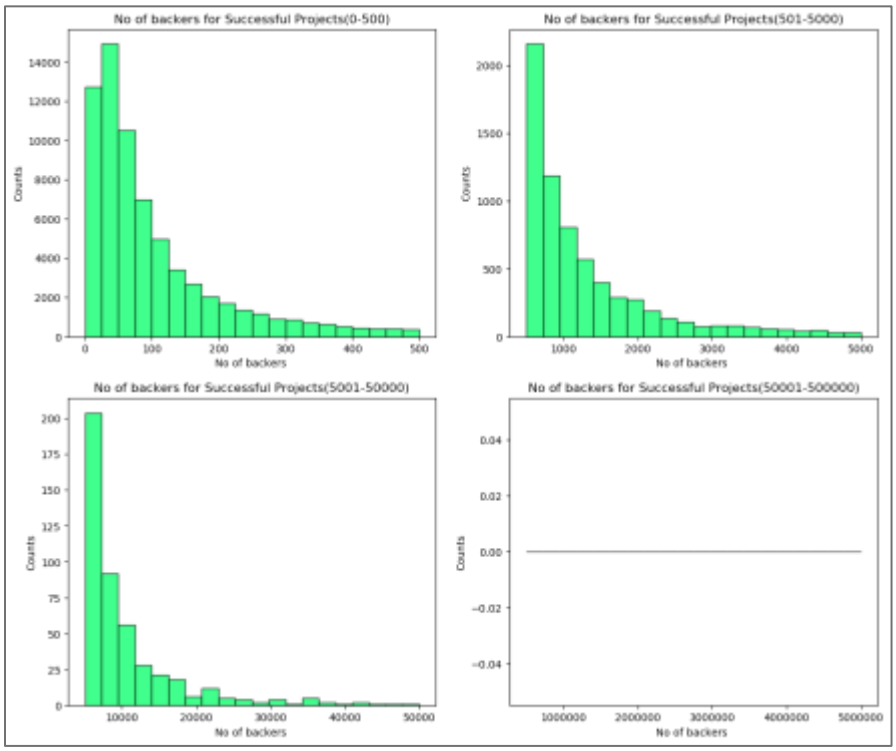

Figure 17: No of people backing the successful project

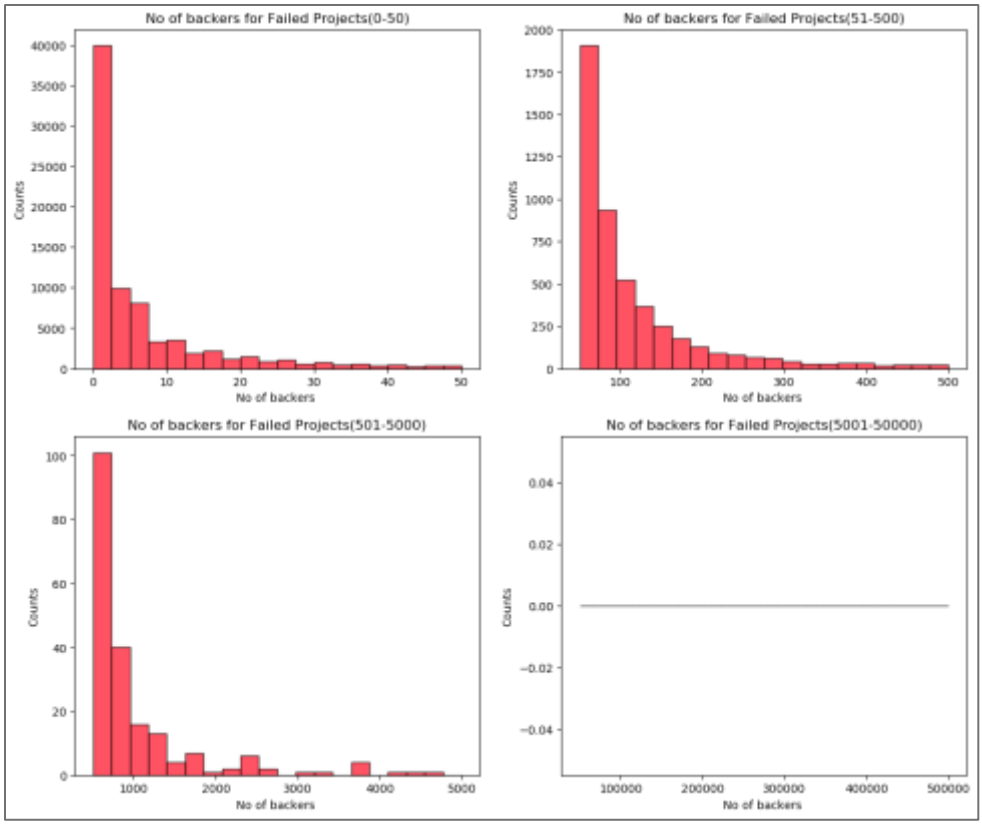

Figure 18: No of people backing the failed project

The above diagram shows the ratio of successful and failed projects in a relationship with no of people backing the project.

- Most successful projects have 25-50 backers.
- Most successful projects have 0-25 backers.
- As a number of backer declines the count of successful and failed projects decline exponentially.

3.2..7. Does staff pick mark more successful projects?

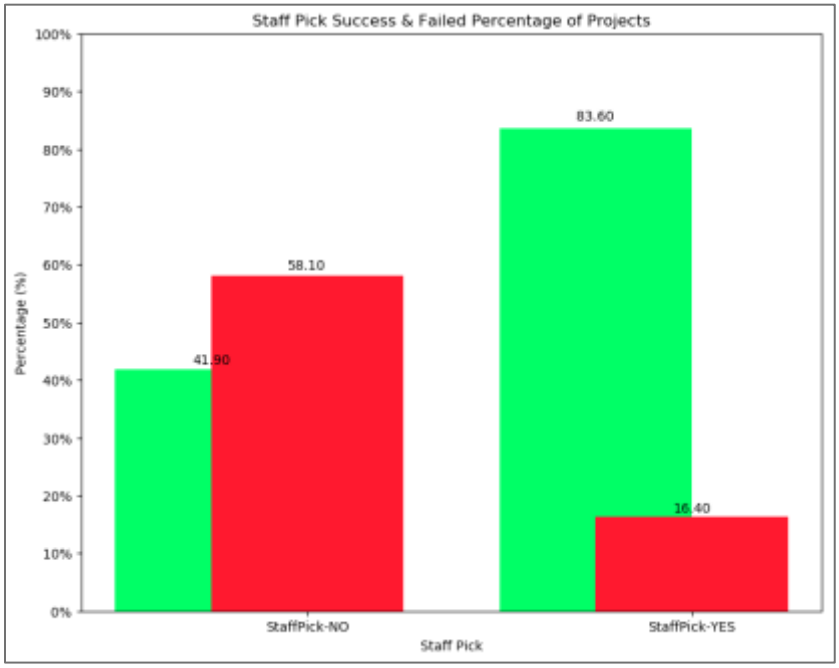

Figure 18: Staff pick for successful and failed project

The above diagram shows staff pick yes or no with successful and failed projects project.

- When the project is not staff pick the ratio of success and failed project is close
- When the project is a staff pick the ratio of success is five times higher than a failed project

3.2..8. Is there relationship with a pledge to no of backers ration in the success or failure of the project?

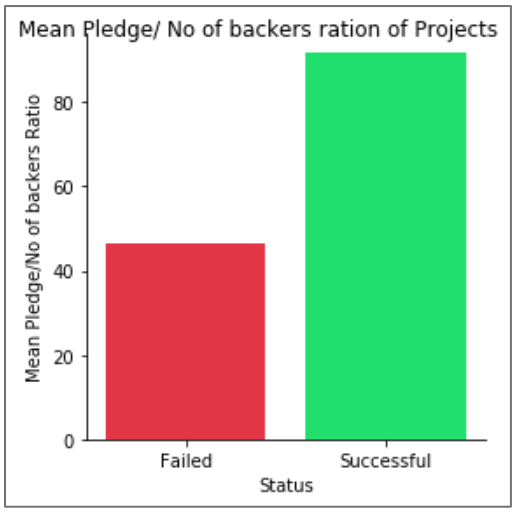

Figure 19: Pledge amount to no of the backer ratio

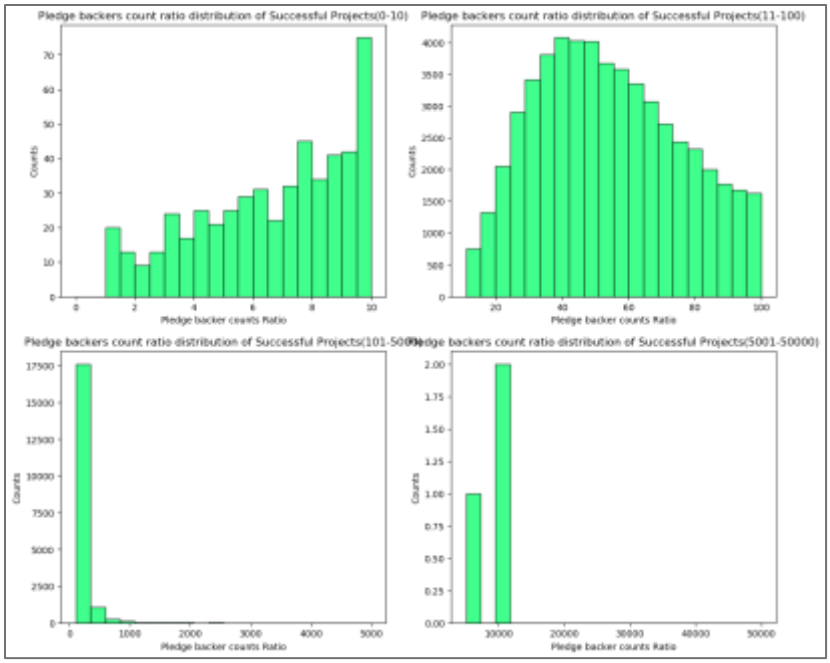

Figure 20: Pledge amount to no of the backer ratio for successful

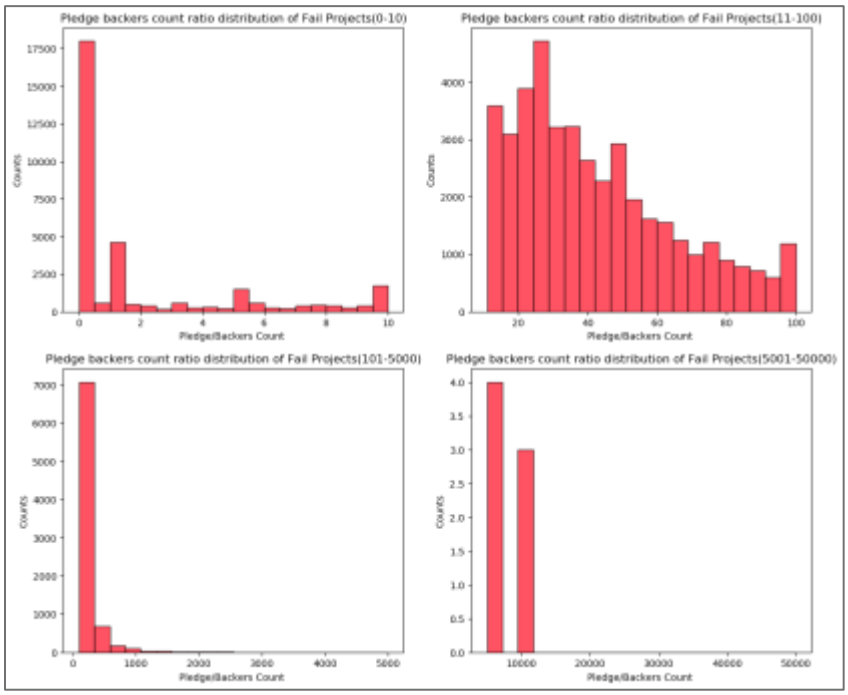

Figure 21: Pledge amount to no of the backer ratio for failed

The above diagram shows ration of pledge amount by no of backers for successful and failed projects.

- The average failed project get less than one USD per person
- The average successful project gets 35-55 USD per person

3.2..9. Which country is having most and least successful or failed project?

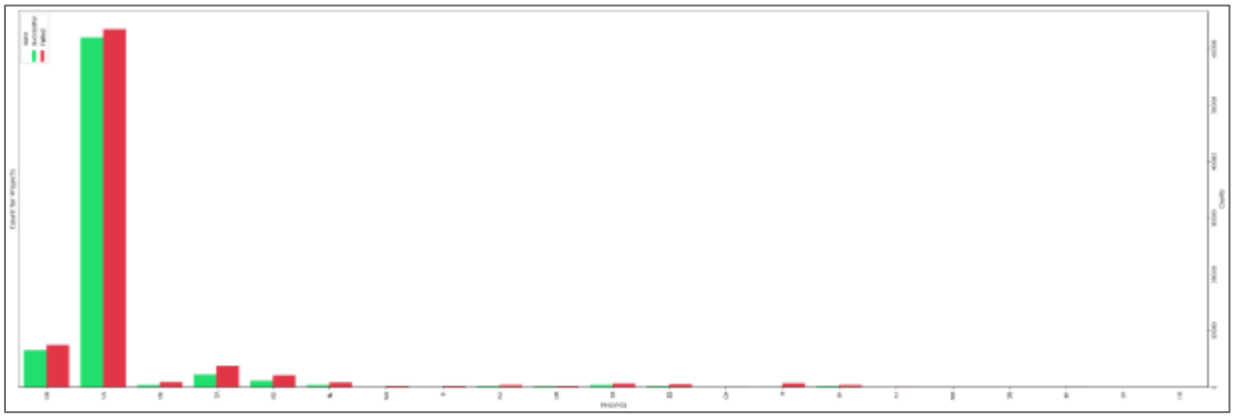

Figure 22: Country-wise distribution of projects

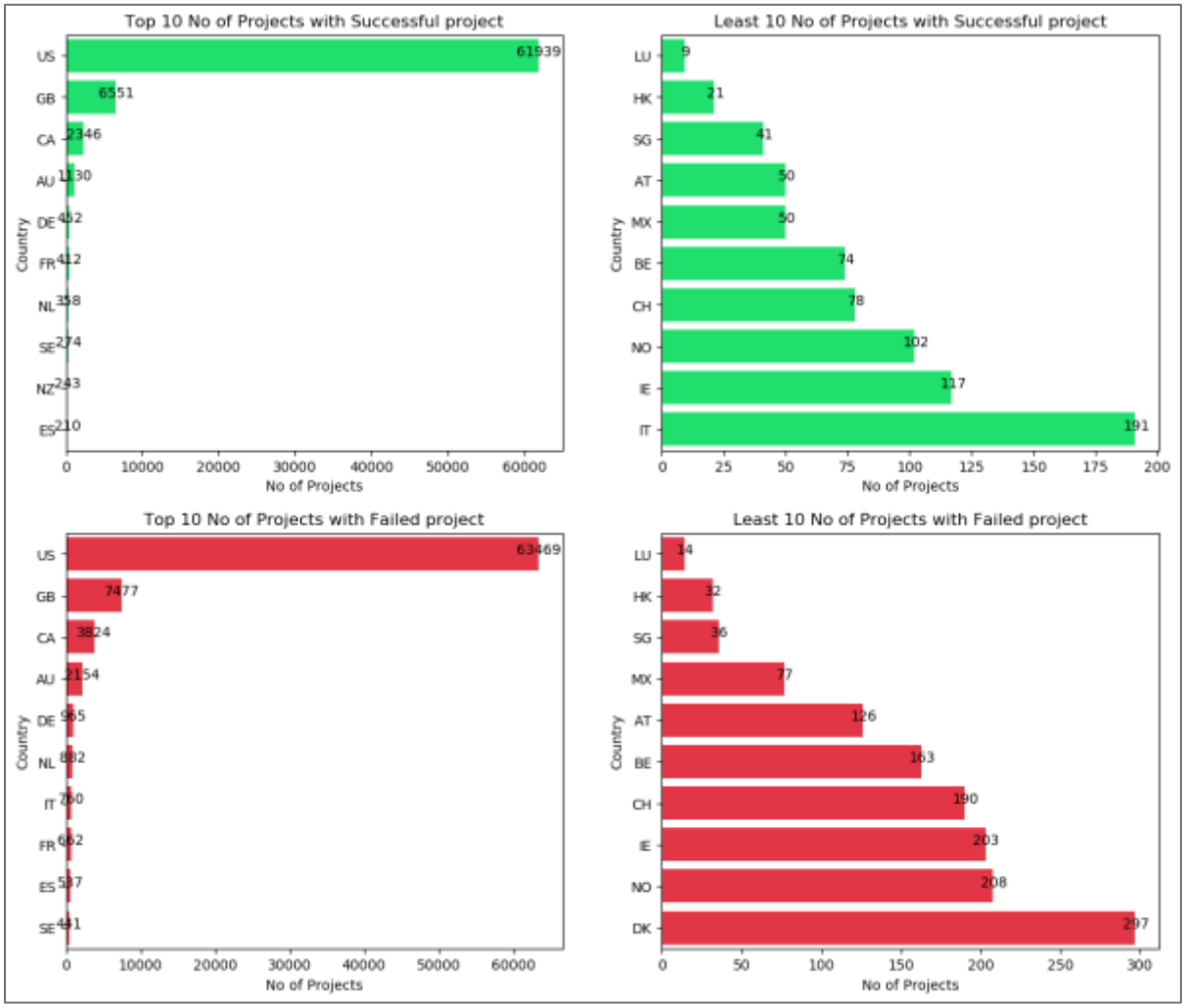

Figure 23: Country-wise Top 5 successful and fail

The above diagram shows the country-wise distribution of successful and failed projects.

- The most project belongs to the United States in both Success and Failed Status.
- Most countries in Top 10 and bottom 10 are same. It means these are most and least participating countries
- The ratio of successful and failed project for a particular country is approximately similar

3.2..10. What categories have most and least successful or failed project?

Figure 24: Category wise successful and fail

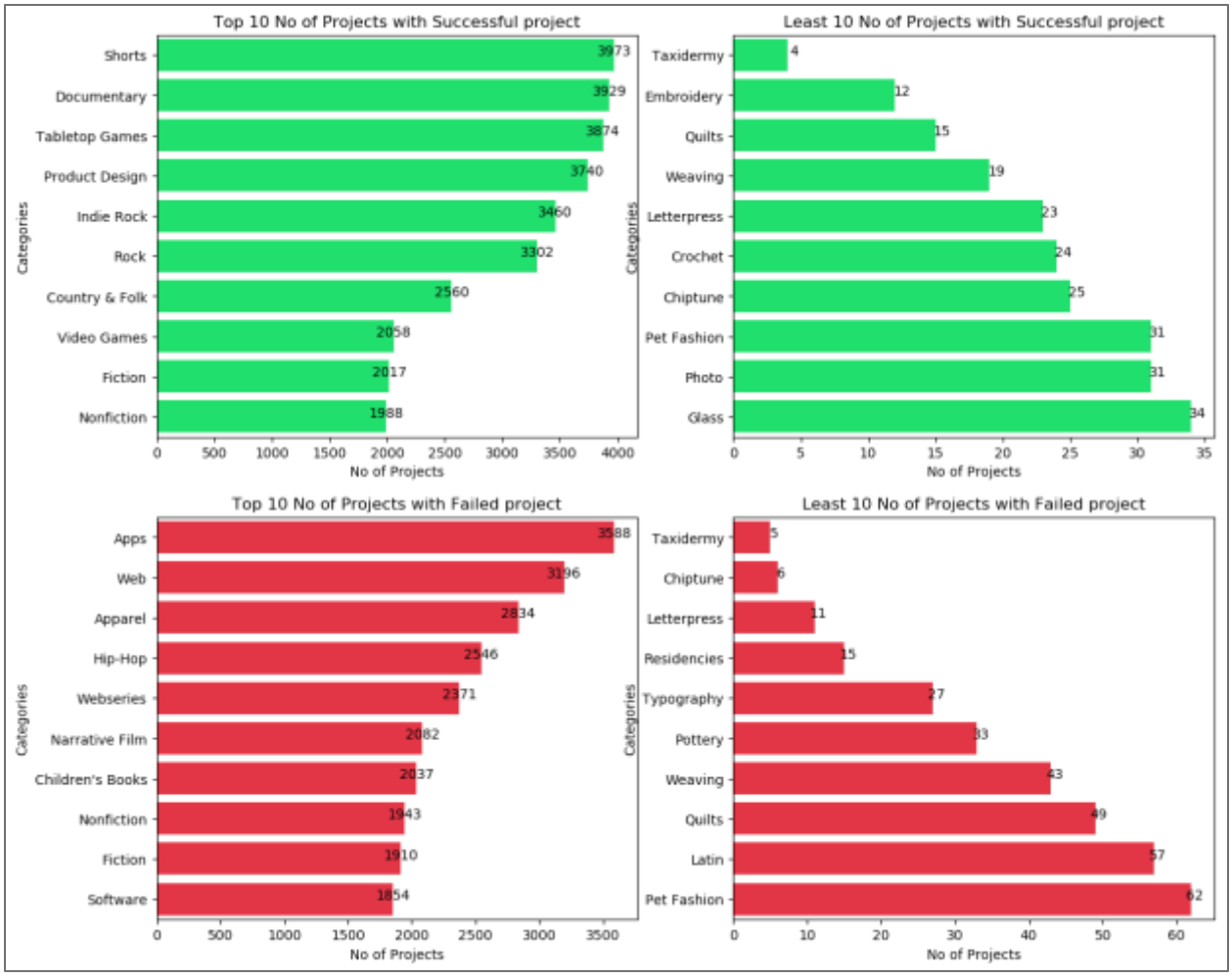

Figure 25: Top 5 categories for successful and failed

The above diagram shows the category-wise distribution of successful and failed projects.

- There are distinct 150 + categories for projects.
- Some category like Apps and Webs have high failing ratio.
- Some category like product design and shorts have high success ratio.
- No Top 10 or bottom 10 categories are common, which means categories have the impact on the state.

3.2..11. What location type is having most and least successful or failed project?

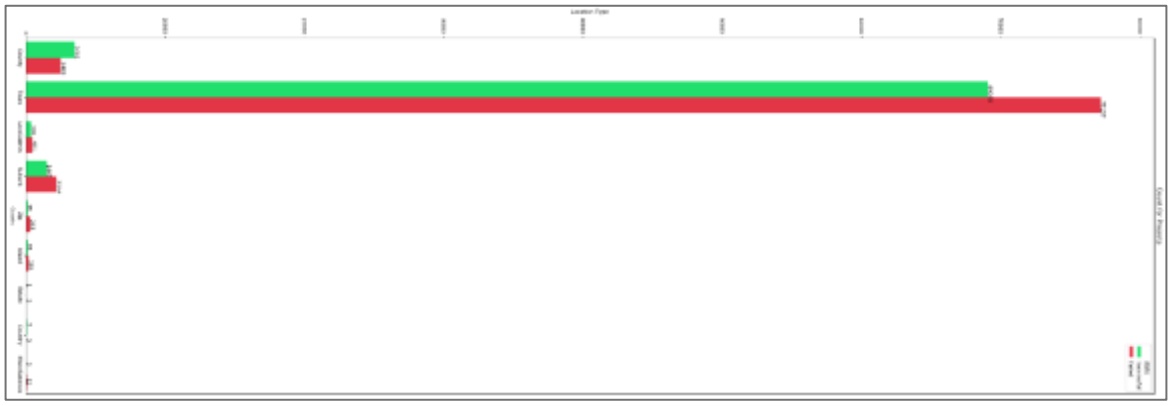

Figure 26: Location wise successful and fail

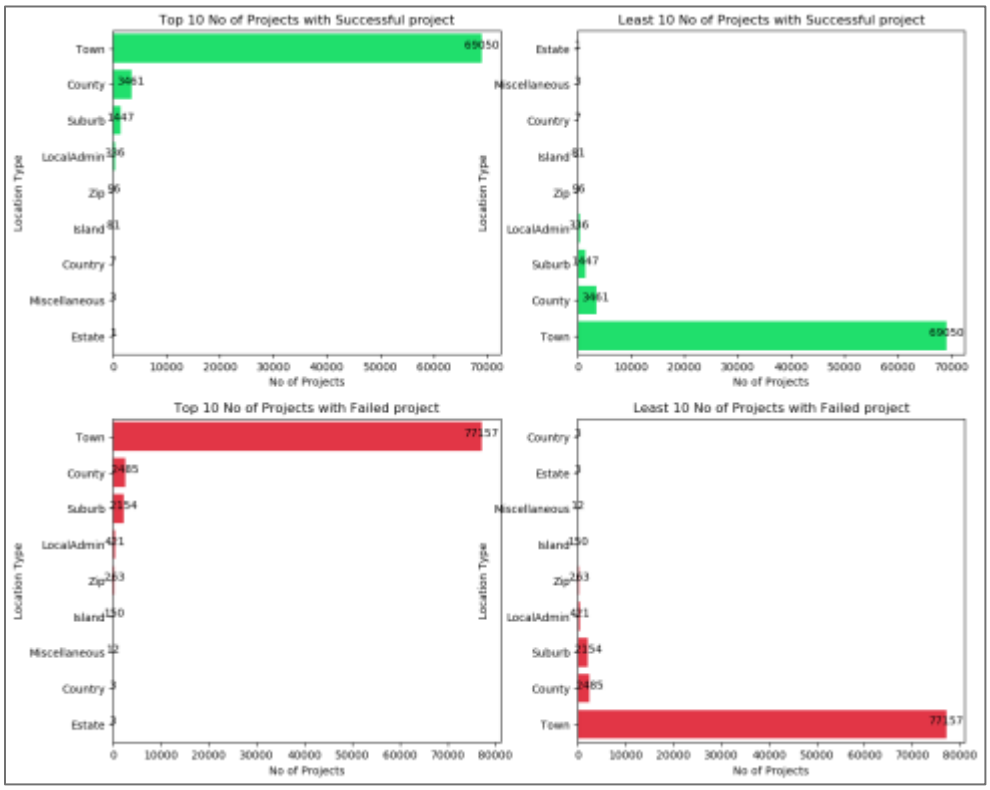

Figure 27: Top 5 location type for successful and failed

The above diagram shows location wise distribution of successful and failed projects.

- The most project belongs location type Town for both Success and Failed Status.
- The ratio of successful and failed project for a particular location is approximately similar.

3.2..12.

4. Inferential Statistic

The inferential statistic section involved exploring columns or variable from the data to perform statistical analysis. This section applies some inferential statistical concept on the data.

4.1. Exploration of a number of backers

The column or variable no of backers for the project is a very interesting column and can be explored with inferential statistic tools and can be further understood.

4.1..1. Successful project exploration

The describe from pandas data fame provides some information

```
count    74847.000000
mean       291.705680
std       1844.760094
min          1.000000
25%         34.000000
50%         72.000000
75%        173.000000
max     219382.000000
```

4.1..2. Fail project exploration

The describe from pandas data fame provides some information

```
count    82896.000000
mean        15.552765
std         82.981983
min          0.000000
25%          1.000000
50%          3.000000
75%         10.000000
max       6550.000000
```

4.1..3. Hypothesis Testing

There is the difference between the mean of backer_counts between successful and failed project. We can further analyze this hypothesis. Answer: The hypothesis is as follows

Null Hypothesis: There is no difference in no of backer in successful or failed project. Which means for no of backers for successful - means for no of backers for failed equals Zero.

Alternate Hypothesis: There is a significant difference in no of backer in successful or failed project. Which means for no of backers for successful - means for no of backers for failed not equals Zero.

Calculate Z score and p score for the null hypothesis

Calculate Z stat using ztest method in a weightstat module with significance level 0.005
The calculated values are as follow for two-sided & larger
t-statistic: 40.9166979163
p-value: 0.0

Calculate T score and p value to test the same hypothesis
Calculate T score using the ttest_ind method in stats module with significance level 0.005
The calculated values are as follow
t-statistic: 40.9166979163
p-value: 0.0

The above p value is less than our significance value and hence there is enough evidence to reject Null hypotheses and

The code related to this part is in Capstone_Inferential_Statistic file.

5. Baseline Analysis

The baseline analysis is performed by applying logistic regression machine learning model. The model obtained is considered as a baseline for further analysis

5.1. Converting columns

5.1..1. To dummies

The following category type columns are converted into dummy column using the get_dummies method in panda's module

1. Category
2. Country
3. Location type

5.1..2. To number

The column staff_pick which has value Yes or No is converted to 0 and 1

5.1..3.

5.2. Column selection

There are few columns which are filtered out from part of the dependent variable or y in the algorithm. The idea is to include as many as column possible but remove any dummy to a variable or not required column

1. Goal amount
2. Project id

The 177 columns used for the model including the one converted into dummies.

5.3. Running logistic regression

The logistic regression algorithm is used to create model [0.01, 0.1, 1, 10, 100]

5.3..1. Setting hyperparameter

The [0.01, 0.1, 1, 10, 100] are used for the hyperparameter to find the best estimator using gridsearch from the model selection.

5.3..2. Splitting train and test size

The train and test split are 80-20. The train_test_split is used to get the two sets.

5.3..3. Fitting data

The training data obtain by above split and the model return by grid search is used to fit the data and obtain the model.

5.3..4. Finding accuracy

The accuracy score is calculated on training and test data using score method of the model

```
The accuracy of training data: 0.82
The accuracy of test data:     0.83
```

5.3..5. Creating Classification report

The classification report provides the information about various parameter like precision, recall f1-score and support. These figures tell about model performance based on how good the real fit is between test and train data based on true expected vs predicted.

```
[Training Classification Report:]
yTrain length= 126194
y_predict_training length= 126194
             precision    recall  f1-score   support

          0       0.78      0.92      0.85     66280
          1       0.89      0.71      0.79     59914

avg / total       0.83      0.82      0.82    126194

[Test Classification Report:]
ytestlr length= 31549
y_predict_test length= 31549
             precision    recall  f1-score   support

          0       0.78      0.92      0.85     16616
          1       0.89      0.72      0.80     14933

avg / total       0.84      0.83      0.82     31549
```

5.3..6. Creating confusion matrix

The confusion matrix is another representation of model performance by displaying in a column the expected vs actual in matrix form

	Predicted False	Predicted True
Actual False	15336	1280
Actual True	4236	10697

5.3..7. Finding feature importance

The regression model is used to find the important feature using standardize parameter and fitting model with Xtrain /standard deviation(Xtrain,0) and ytrain . The coefficient of this model provides with high influencing independent variable for both success and failure. The top five values with absolute values provide the important features in both successful and fail

Top 5 feature with towards success (1): positive values

1. backers_count
2. usd_pledged
3. Shorts
4. Documentary
5. Tabletop Games

Top 5 feature with towards fail (0): negative values

1. Restaurants
2. Software
3. Web
4. deadline_days
5. Apps

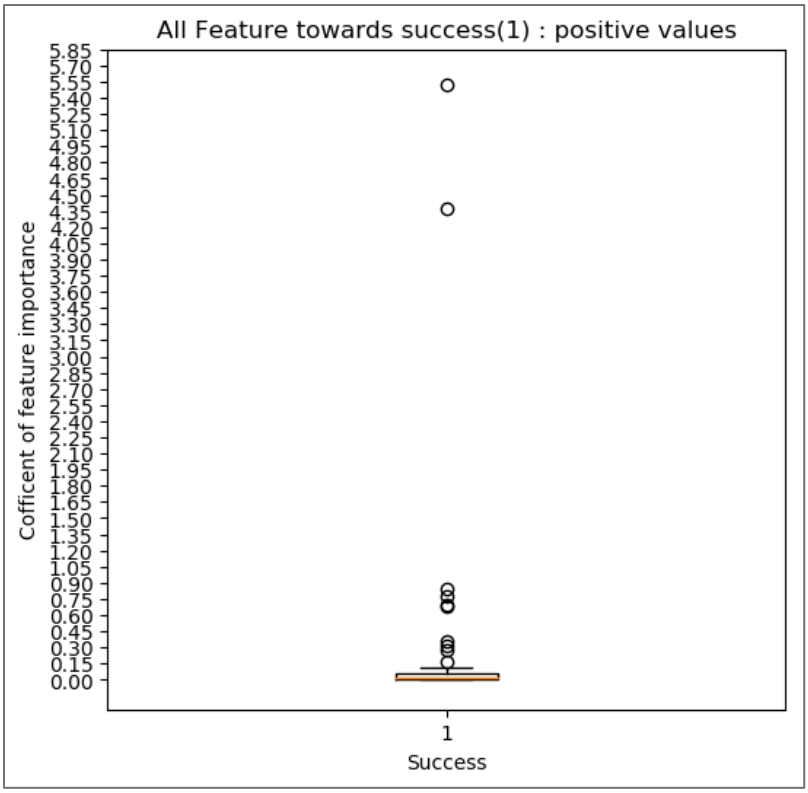

Figure 28: All Success feature spread

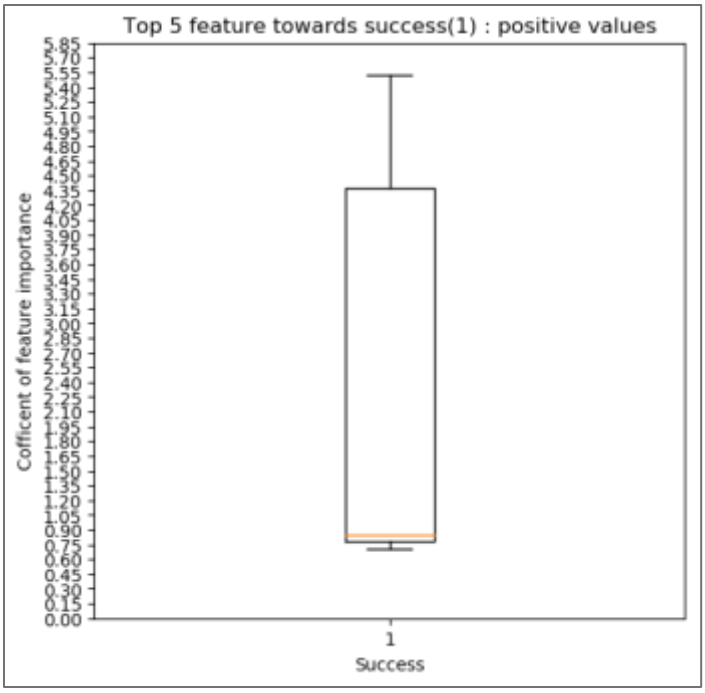

Figure 29: Top 5 Success feature spread

Top feature with towards success (1): positive values

- The spread of top positive feature is large due to top 2 value(5.55 & 4.50) being relatively very high numbers
- Most value resides in lower range providing nearly same feature importance coefficient

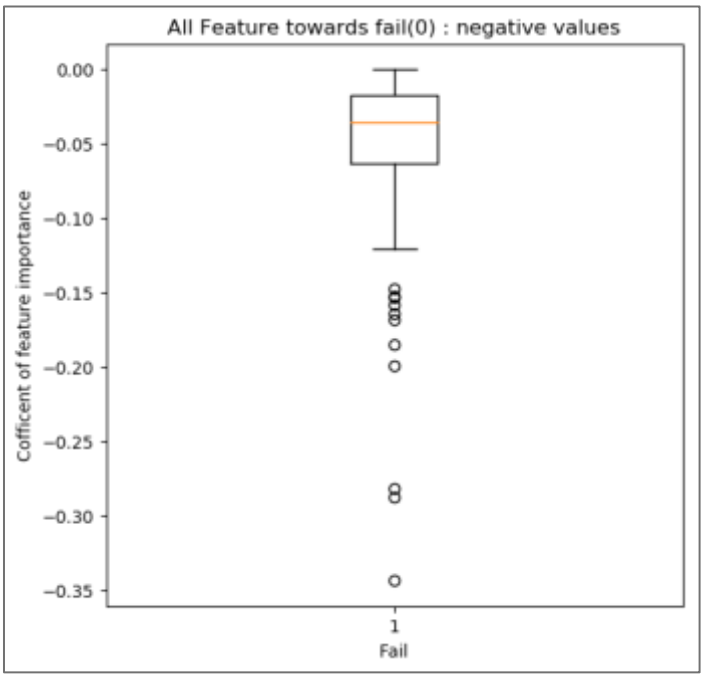

Figure 30: All fail feature spread

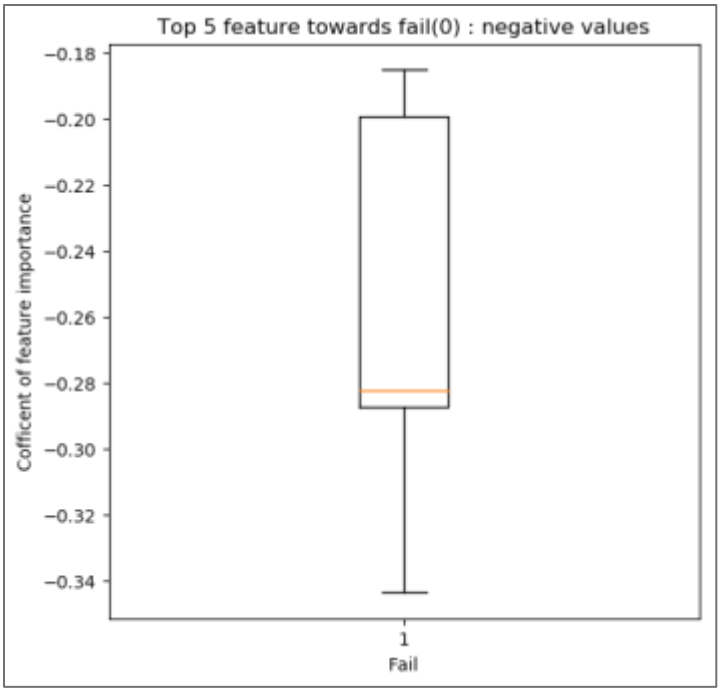

Figure 31: Top 5 fail feature spread

Top feature with towards fail (0): negative values

- The spread of top negative feature is large due to top 2 value (0.35 & 0.30) being relatively very high numbers.
- Most value resides in higher (absolute number) range providing, unlike success coefficient.
- The absolute value of fail is comparatively very small to success top most coefficient.

5.4. Multiple Run to verify

The above model is run only with one set of test and train data. To find that model was not over fit for the particular training set. The model is run with 1000 times with 1000 value of random state with 80-20 split in the model to get different set every time. The training and test accuracy are calculated for each run to verify the difference.

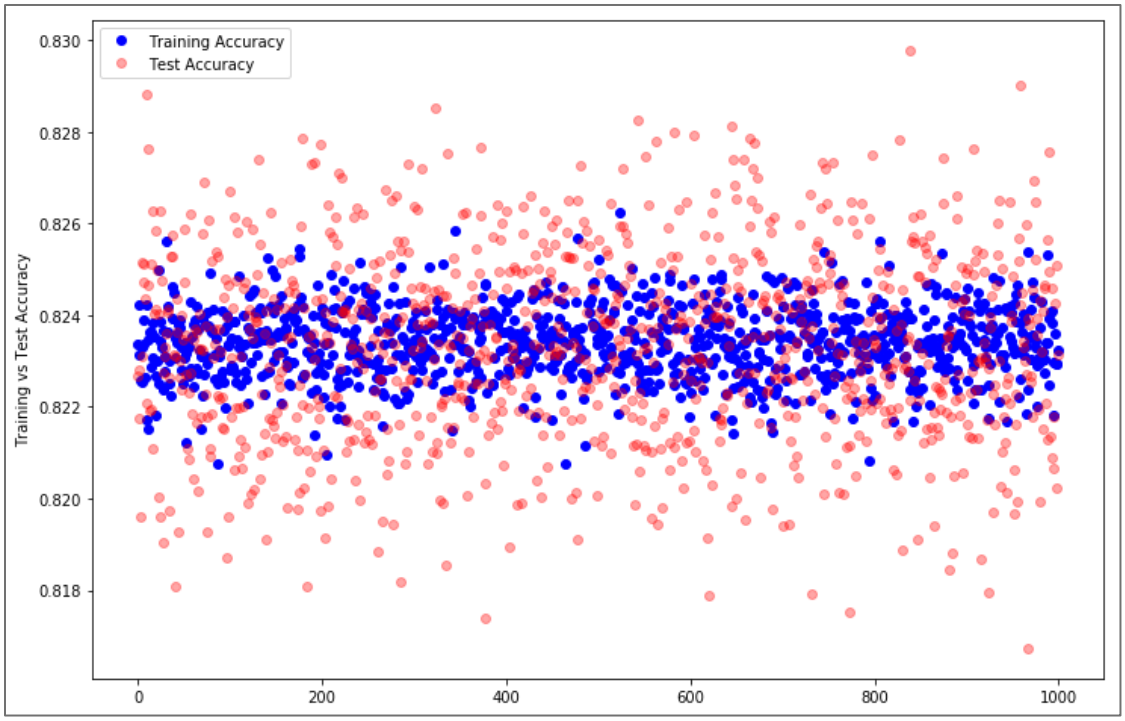

Figure 32: Training and test data values for 1000 run

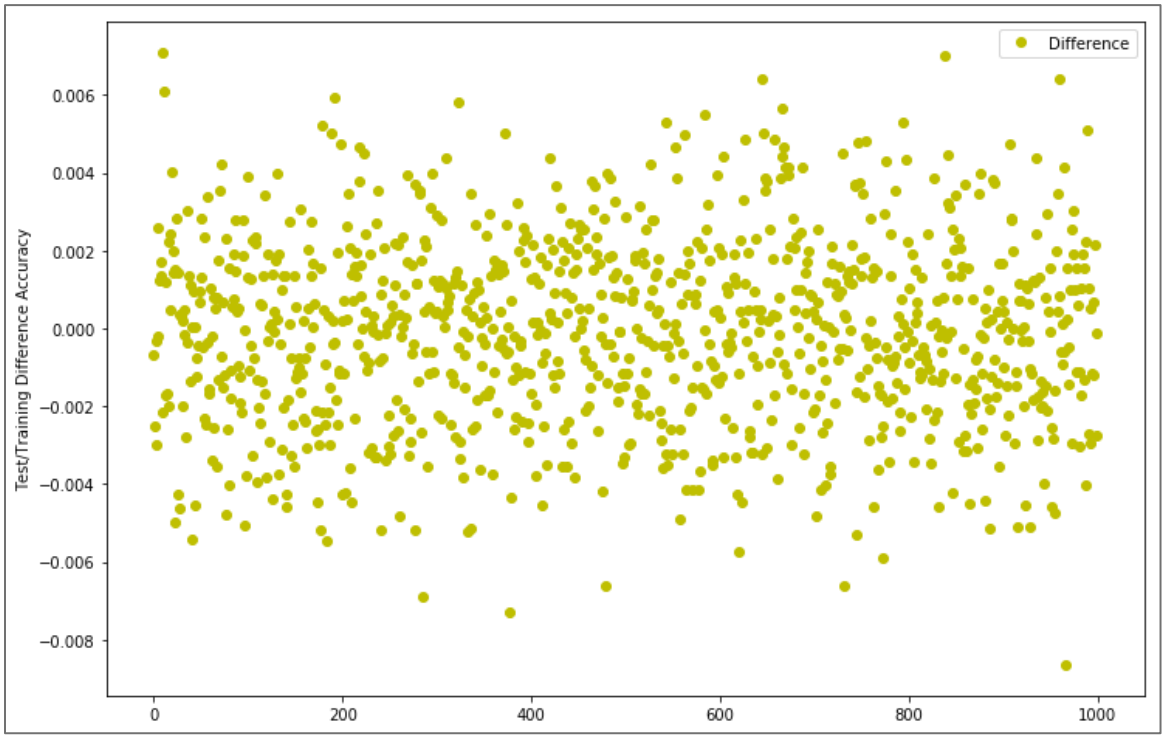

Figure 33: Difference of values for 1000 run

Training VS Test Accuracy comparison

- The distinct training accuracies are closer to each other as compared to test. This is obvious as a model is trained on that data.
- The maximum difference in test and training accuracy is 0.007.
- The above graphs and process are to get an idea of the difference in accuracy and is not final metrics. The classification report and confusion matrix above provides more insight.

6. Future Work

The project has tried to cover many aspects of data science student and covered various aspects. But there is still scope to improve the methodology or explore more in the project. As for endnotes, this will be marked as future work in reference to this project

6.1. Scope:

There can be various crowdfunding sources, however, this Project will be limited to projects funded using a website called **Kickstarte**r. There is a possibility to include the other sources.

6.2. Launch Period:

This work on finding the launch period the model is valid to predict more accurately. The work can be done in understanding that until how long after the project is launched the model is still valid to predict with accuracy.

6.3. Technology:

Currently, the files are processed using code running in a Juypter notebook on the local computer. The process took a long time to be processed on the local computer especially during wrangling and extracting data. The further work can be done to leverage cloud technologies like AWS, Azure, and Cloudera etc. to speed up the processing

There can be work done to make it batch process for getting new data available every month and update the model.

There is the scope of creating a web application to make it available to use the model and also more visualization