**The Art of Kickstarter Projects: An Exploratory Analysis**

**Kevin Phan | 500558667 | https://github.com/Phankd**

**Introduction**

The following project aims to explore the difficulties and hardships in the world of entrepreneurship. It is more often than not, a gray area on whether or not to embark on such a venture due to the financial challenges that stand before many individuals. The ambiguity of raising money successfully has become a challenge, very much so that we wonder if there are any factors that either hinder or accelerate the success of a Kickstarter project. Our problem lies: What are the factors that favour success over failure for any Kickstarter and what are the determinants for overfunding. To solve this problem, I will be implementing a few machine techniques after appropriate data cleansing such as logistic regression and KNN regression. Afterwards, we will compare performance measures such as recall and deploy ensemble techniques in order to create one model with the strengths from both the previous models.

**Literature review**

Project Success Prediction in Crowd-Funding

This paper implements an exploratory analysis that aimed to predict project success and developed a regression model that one can use in the event of impartial data. They utilized a dataset of 18k Kickstarters and 116K tweets regarding Kickstarters from twitter. Both a logistic and log logistic distributions was used to fit the data and survival AUC was used to rank the models of effectiveness.  They found that incorporating the failed projects with successful projects yield better results than the corresponding dataset with only successful projects. This was seen by using a 10-fold cross validation and better results across all feature set combinations. They had also seen that social network-based features from an NLP analysis of tweets contributed to prediction performance.

Project Recommendation Using Heterogenous Traits in Crowdfunding

This project had aimed to answer the simple question of “What set of features determine a project’s success?”. They separate their features into four main groups. Project based, personality based, location based and network-based traits. They used a gradient boosting tree model which is a form of ensemble method. This is good for generalization of unseen data which prevents us from worrying about heterogeneity. They have decided to not use logistic regression and SVM due to the opinion that normalization of heterogeneous data via standardization may not be suitable. AUC values were also to evaluate the model under the four main groups as well as social network, influence scores, and accumulation of backer’s categories. They had found that social and topical factors heavily influenced number of backers and that backers are also influenced by their social network. We see that a backer model was recommended in order to reach the goal.

The Dynamics of crowd funding: An Exploratory Analysis

I chose this paper because it gives a different view on understanding the nature of crowdfunding using graphs and visuals, the paper aims to prove that social connectivity, geographic location and project quantity are driving factors of successful Kickstarter projects. It provided tables of key summary statistics, correlation tables between attributes used. They also provided histograms of funding levels, as well as heat map charts for geographical distribution of successful Kickstarter. They had found that projects either succeed by small margins or fail by large ones as seen and social capital and preparedness showed higher correlations to success. Geography was also linked to the success rates.

The Determinants of Crowdfunding Success: Evidence from Technology Projects

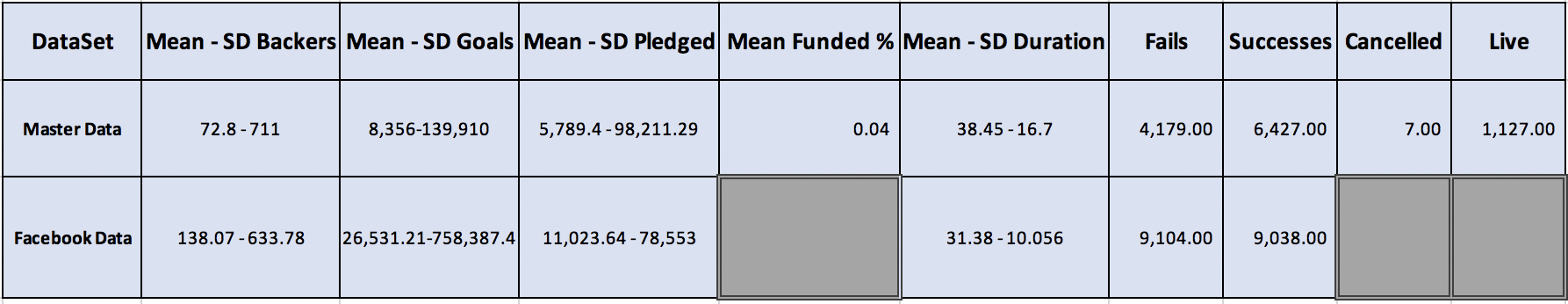
This paper aims to prove that project fund amounts, duration and dollar amount contributed a day are all correlated to a project outcome. Using a combination of datasets, the Data scientists used a Probit regression model where project outcome was regressed on few attributes including several Log transformed ones. They found that the Ln of Requested, Days and contribution frequency had strong impacts on the outcome. The same regression model was implemented on a dataset with just successful projects. This showed that successful projects had a higher mean of contributors and relatively the same attributes had similar impacts. Overall, research found that three attributes had consistent impacts across all models: the invested requested, the duration and the contribution frequency.

Crowdfunding: Determinants of Success and Funding Dynamics

This paper explores the dynamics in crowdfunding. An analysis was conducted on both the project level and the individual level. A probit regression was ran and showed that there is a negative correlation between a projects funding target and getting out of the starting phase. We are also shown pledges over funding durations and a time series of the pledges of stacked metrics (failed project, target not reached and target reached) and see that targets reached had a higher number of pledges. Overall, the paper deciphers the analysis as evidence towards the fact that only about 30% of projects collect enough pledges early to appear on track and most of them keep this momentum going until the goal is reached. The rest are a combination of failed projects and projects that attract pledges but doesn’t seem to be on track.

**My Dataset**

Background: My data set consists of an inner join of three datasets with different attributes regarding Kickstarters. The total amount of rows pre-join between the three datasets were 468,743. After joining them on Kickstarter ID, we have rows for 11,470 Kickstarter projects. We have approximately 20 different attributes (ranging from backers, pledges, location etc) being used (further dimension reduction will be performed) and it is a combination of numeric, factor and character variables. We also have one more dataset with Facebook statistics of project initiators which would help prove my social pretenses hypothesis. This particular dataset has 18,142 rows and 35 unique columns. All datasets have been retrieved from Kaggle. Elementary summary statistics of the datasets are below.



Statistics are tentative as further investigation of the data and changes may affect these numbers. Numbers may differ in the final project.

Imbalances as well as other outcomes besides successful and fails for dependent variable will be excluded. Facebook dataset did not have some of the attributes in the master dataset but a separate model will be made for this to analyze social presence on Kickstarters.

(all statistics are computed in R Studio)

**Approach**

Test KPI measures for accuracy.

Perform Data Cleansing & summary stats

Implement feature selection for models and Split

Perform variety of Machine Learning Algorithms

Perform ensemble techniques for General Model.

Create a project guideline.

Step 1: Data Cleansing and Summary Statistics

As shown earlier in the literature review, this step entails cleansing the data. Such tasks include inner joining all the datasets, checking for outliers, correcting imbalances, removing collinearity using VIF, remove NA, Null, missing or zero values and normalizing data if needed. The summary statistics will include the 5-number summary, correlation table box plots, histograms to visualize frequency and mean and standard deviation values for all numeric attributes.

Step 2: Feature Selection and Dimension Reduction

Initial data sets average around 35 attributes. We aim to reduce the number of features used by performing PCA reduction and using the first principal component as our features. We will also do information gain and see in which aspect would this differ. For the spit and models, we will use 10 cross fold validation and percentage split. We will graph the folds against its accuracy to see the optimal amount of folds and same for the percentage split. This will be done for each model respectively.

Step 3: The Model

After feature selection is done, we will perform logistic regression, log logistic regression, KNN and SVM. All this will be done on their training sets and will be used to predict test sets. Further narrowing of the best predictors for logistic regression will be done. Possible techniques could be stepwise elimination for this. Errors will be graphed. We will be using the optimal splits and cross folds that were derived from the previous step. In this step, we will also attempt to derive constraints and threshold that allow for successful projects (Censored Regression).

Step 4: KPI (Key Performance Indicators) Comparison

After the machine learning has been executed, we will compare the models using a few measures. We will first create confusion matrixes across all the models to see the true positives and true negatives of the model. Performance summaries of the model will be created in which we will compare stability, recall, F measure and AUC.

Step 5: Ensemble Techniques

After metric comparisons, ensemble techniques will be used to create one predictive model. Strengths are taken from each of the model that we performed in step 3 to create one model that will generalize everything. Certain techniques used can include bagging aggregation or strapping.

Step 6: Project Guideline

After deriving a solid predicting model, correlations and techniques to provide thresholds and constraints, a general guideline will be created with parameters that are optimal for a successful Kickstarter. This can include thing such as most popular category, optimal duration, optimal goal, amount of social presence to be on track and number of backers generally needed for goals to be reached (not restrictive to those listed).

**Literature Bibliography**

Li, Y., Rakesh, V., & Reddy, C. K. (2016, February). Project success prediction in crowdfunding environments. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 247-256). ACM.

Rakesh, V., Choo, J., & Reddy, C. K. (2015, April). Project Recommendation Using Heterogeneous Traits in Crowdfunding. In *ICWSM* (pp. 337-346).

Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of business venturing*, *29*(1), 1-16.

Cordova, A., Dolci, J., & Gianfrate, G. (2015). The determinants of crowdfunding success: evidence from technology projects. *Procedia-Social and Behavioral Sciences*, *181*, 115-124.

Crosetto, P., & Regner, T. (2014). *Crowdfunding: Determinants of success and funding dynamics* (No. 2014-035). Jena Economic Research Papers.