DETERMINING KICKSTARTER PROJECT SUCCESS

FINAL REPORT

Tim Dufala
tdufala3

Bui Thi Thu Giang
gbui8

Rashmi Raju
rraju6

Abhishek Surya
asurya6

Fernanda Tello
mtello3

PROBLEM DEFINITION

We aim to predict Kickstarter[15] campaign success or failure and visually display the factors that drive success for the project.

Success is measured in terms of achieving the funding goal within targeted time frame. The prediction will be based on various parameters such as number of backers, geographic location and several more described below.

MOTIVATION

Campaign managers care about understanding the factors that would make their campaign successful, and backers care that the projects they back will meet funding. Investors also are moving toward crowdfunded products because of it is convenient, relatively less risky than other investments, provides real-time feedback, and a great marketing tool[14].

This application can help project creators improve their campaign by suggesting changes. To measure our model success we will use the standard evaluation metrics like: accuracy, cross validation, confusion matrix, and scoring to compare the different learning models.

LITERATURE SURVEY

Related Work

Based on our research, we came across several investigations that provide ideas to measure the success of crowdfunding campaigns and relate to the goal of our project.

"Project Recommendation Using Heterogeneous Traits in Crowdfunding"[7] uses attributes such as social network impact, backer profile/demographics and geo-location traits to determine success of individual campaigns achieving their funding goal. The analysis provides insight on how these attributes affect campaign success. One shortcoming is that the authors only target the influence of Twitter which may skew an accurate measure of holistic social media impact on crowdsourcing platform success.

"Issues in Crowdfunding: theoretical and empirical investigation on Kickstarter"[8] uses both empirical and theoretical models to explore reward-based projects with respect to public-good issues and advertising. Although a limitation of this study is it only applies to the platform in question, it still helps us focus on a key characteristic crucial to raising funds.

"Predicting the Success of Kickstarter Campaigns"[6] applies a linear regression model to narrow down some key factors leading to Kickstarter campaign success. An important limitation pointed out is that if all-time data was used, it may have yielded more accurate results.

"Individual crowdfunding practices"[18] investigates unique factors

which affect crowdfunding success, although this research is limited to equity crowdfunding.

Ideas and Methods

From "Delivery Rates on Kickstarter"[1], we acknowledge that we may couple the factors that lead to project success with factors that cause them to fail such as failure to deliver rewards to key stakeholders.

"Crowdsourcing Application in Marketing Activities"[9] shows that companies would prefer to know the success of deploying their products while planning its public reveal so therefore, the level of advertising promotions a campaign supports could be an additional characteristic we consider for kickstarter product success.

Besides researching specific variables, methodology is explained in more detail in "Predicting Kickstarter Campaign Success"[4] and "Predicting the success of Kickstarter campaigns"[3]. These studies comparatively evaluate and draw conclusions of successful vs failed crowdsourced projects based on different machine-learning-based models built with certain parameters. Limitation of these studies is they do not compare and contrast multiple machine-learning-based models with chosen datasets but they do give us insight into the power of ensemble learning.

This is further exemplified in "Launch Hard or Go Home!"[5] where a combined social and money-based predictor proves to have the highest prediction accuracy. The approach is useful because it achieves a highly accurate prediction or success/failure based on relation among multiple weighted characteristics.

Another approach using ensemble learning is an improved version of Random Forest model which assigns greater weights to the decisions made by high-accuracy Decision Trees in the ensemble[11].

PROPOSED METHOD

Intuition

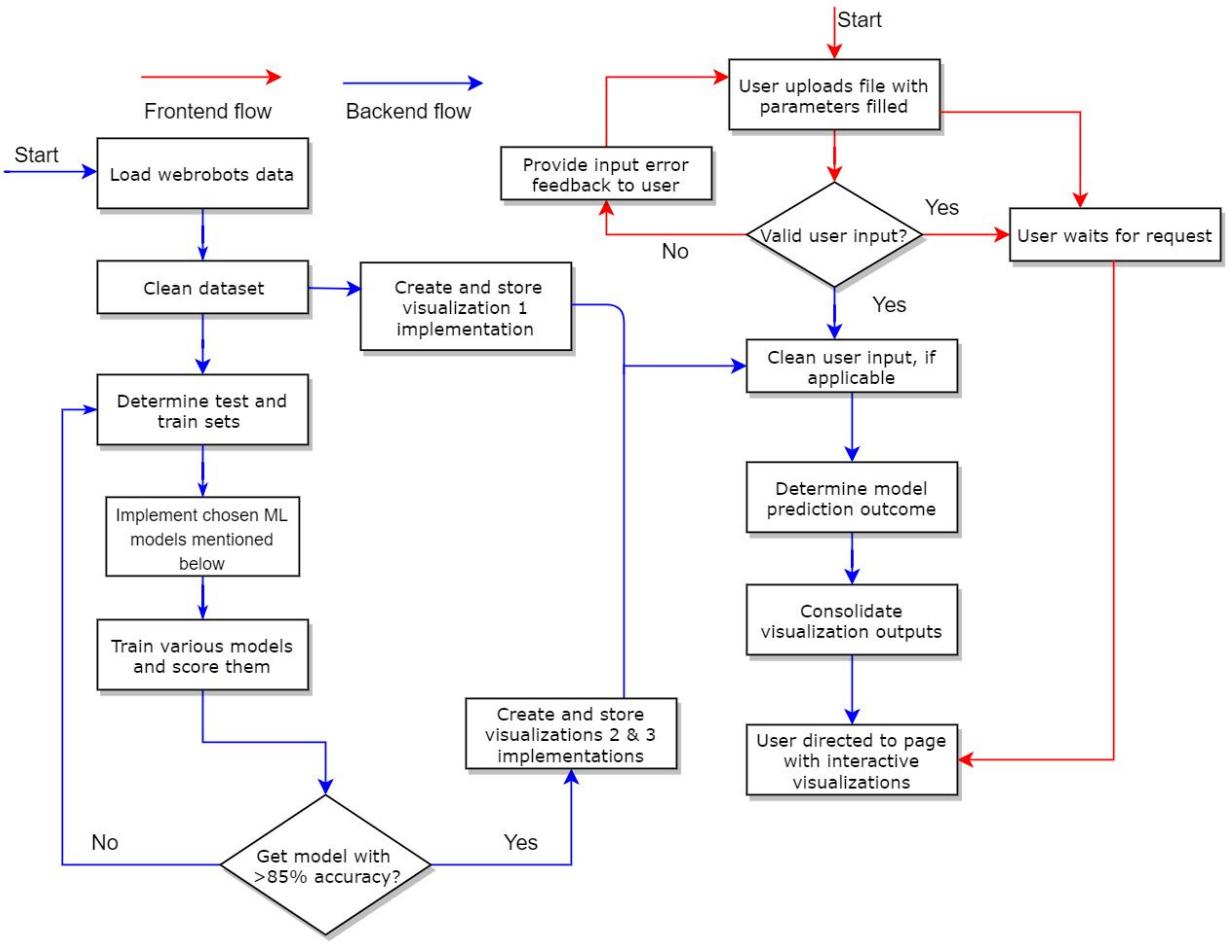
Our approach adds visualizations and suggests ways the Kickstarter project can be improved. This will be successful because it will help the project creators focus on the weak variables to improve that. We also think that using a large dataset[16] will aid in creating a more robust model.

We're aiming for at least an 85% predictive rate on randomly-sampled test data in our final model. Our innovations include:

1. Predictive model to provide overall feedback to users on how to improve their project campaign after evaluation of all identified features
2. Give users ability to train model further by allowing them to upload test-data in bulk.

Application Flow

The diagram below displays overview of frontend and backend flows of our application.



Figure#1 - Application Flow Chart

Application Architecture

The application is run on an Express / Node.js web server. User input is passed to the server via asynchronous requests (AJAX). The server will process the user data and pass it to a Python sub-process to perform analysis using pre-trained models. The server will use this to provide a response to the client. The client-side code will use this data to generate any user-input-based visualizations.

Data Preparation

The Dataset used for machine learning prediction model and statistical visualization has been downloaded (18 October 2018) from Web Robots website. Web Robots scrapes Kickstarter data by running once-a-month crawl. Raw dataset (~ 1 GB) was cleaned up by removing nearly empty columns and the rows which contains null value. Size of dataset after data cleanup is (~ 23 MB) having in-total 186,337 project records.

The feature engineering is done by generating new features like pledged amount per backer, launched year, duration time, etc. from the existing features.

For the feature selection we only choose the features which we think is useful for the model and drop the others like some json columns (after we flat it out and extract the meaningful information).

User Interface Design

The UI is part of the web-based application. It was developed using HTML5/Javascript/CSS.

In this regard, a use-case is described below:

1. User downloads input file template
2. User uploads their file with all parameters filled
3. User receives feedback if errors in input exist. Otherwise, waits for results
4. Resulting visualizations are displayed according to user's input and model prediction

Users are able to upload input file for prediction model as per instructions displayed. Statistical visualization based on dataset which was used to train model is displayed once page finishes loading along with general project introduction and description. After running the predictor, results summary is displayed along with supporting visualizations.

User Interface Implementation

User interface of the application is designed in HTML and it includes javascript elements on D3.js and jquery.

When the web page opens, it loads an interactive statistical visualization of the clean kickstarter dataset along with introduction text and guide on how to use the application.

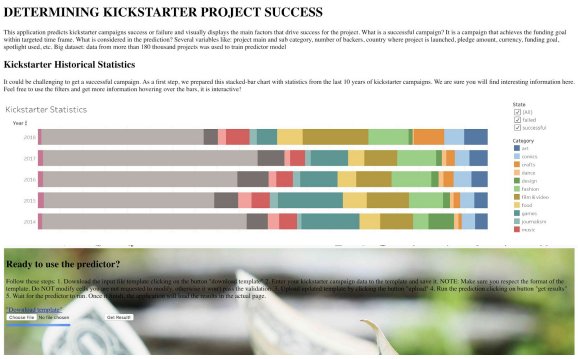


Figure - Screenshots of UI

User can download csv file template, fill it in with required feature data that includes providing following info - backers_count, converted_pledged_amount, country, current_currency, disable_communication, fx_rate, goal, is_starrable, pledged, spotlight, staff_pick, duration, launched_quarter, launched_month, launched_year, launched_week, goal_reached, pledge_per_backer, main_category, cat_name.

User can then upload the csv file which acts as input for the prediction model. Once model processes the input data, it returns final 'success/failure' verdict along with identified most influential features and their suggested values. Users can see the overall result as an interactive visualization. Additional visualization is also displayed to illustrate relative importance of influential feature sets to highlight those features which can be tweaked to maximize probability of project success (visualizations are explained in detail in next section).

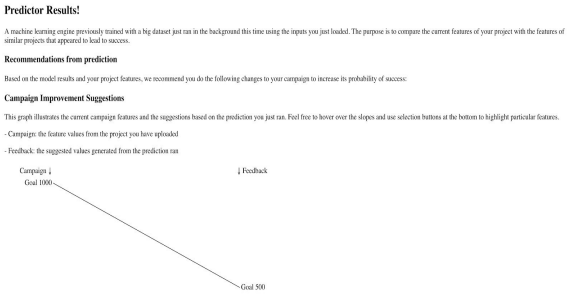


Figure - Screenshots of web page with result from the model

In the end, project credits and important project documents and link to download project source code on gatech github is provided for users who would like to work and improve setup.

Visualizations Design

One of the most important features of our project is the UX enhancement by the use of visualizations.

A variety of tools are being used to create the visualizations: D3, python seaborn(static), matplotlib, bokeh, plotly.py and tableau.

UI contains the following visualizations:

1. **Horizontal stacked bar chart:* using only our cleaned dataset, statistics are displayed in a stacked bar chart.

   Colors indicate the different campaign categories.
   Sections in stacked-bars indicate percentage of actual category pledged from the year total.
   Years labels indicate year when campaigns were launched.

   Interaction is possible in different ways:
   A. Hovering over the stacked-bars will display a tooltip with key statistics from the selected bar section
   B. Clicking on a section of the stacked-bars will highlight this section and shadow the rest of the visualization
   C. Three filters are available to handle dataset:
   - State: to choose between Successful/Failed/All campaigns
   - Category: display only desired categories

   *Note that in order to load this chart, you will need to have an account in Tableau Online. [19]

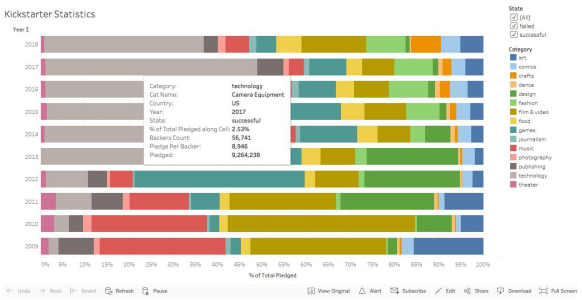


Figure - Stacked bar chart

2. *'Slopegraph'[20]:* using both, the user's campaign input and the outputs from our previously trained model, suggestions to improve the user's campaign are visualized with a slopegraph. This graph shows the trend from the current feature values in the user's input and the suggested feature values from the prediction model. Interaction in this visualization is very useful to highlight each feature either by hovering over it or by clicking on its corresponding button.

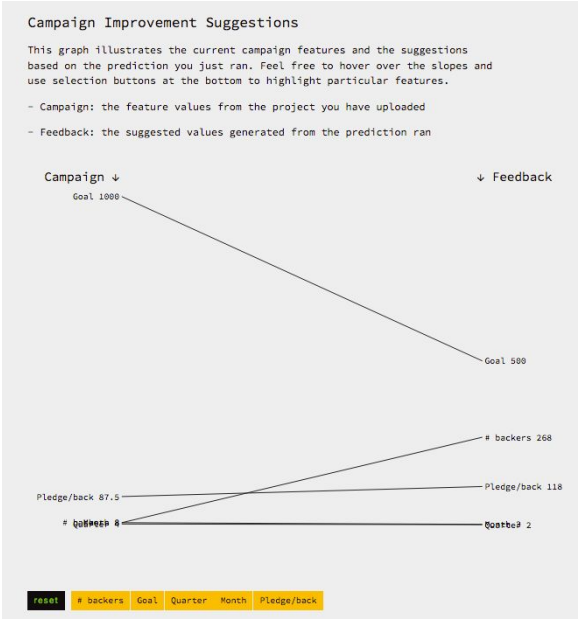


Figure - Slopegraph with predictor suggestions

3. *'Horizontal bars chart':* in order to compare the relative importance of the campaign features against the success/failure prediction, we are generating this bars chart.
   The horizontal axis shows the features level of importance while the vertical axis sorts them from top to bottom starting from the most important features. These features are in descending order of impact on projected campaign performance.

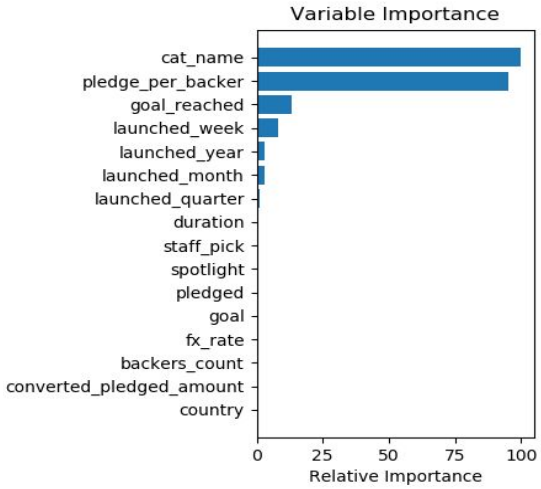


Figure - Horizontal bars chart

Visualizations Implementation

It was important for our application to use intuitive visualizations and interactive where needed.

In order to embed visualizations in HTML we are using JavaScript elements like D3.js and the Tableau API.

On the other hand, we are generating a visualization from python backend making use of matplotlib.

Model Building & Considerations

We are trying to apply various model in order to achieve high accuracy score. We may also combine them to get a better result. Here are our considerations.

Random Forests is an ensemble method which operates by constructing a multitude of decision trees and merge them together to get more accurate and stable prediction. As stated in its name, this model is added with a randomness while building the trees, it searches for the best features in a random subset of features to split the node instead of using directly the most important features in order to prevent overfitting and create a better model. This algorithm is best suited for prediction of campaign success/failure verdict and visualization 3 to display most influential attributes of user's campaign. We want to choose the attributes which achieve the highest information gain after splitting the data as the most influential.

Logistic regression is a linear model for regression rather than classification. It predicts the probability of the outcome variables based on the predictor variables. The output is a probability and we can choose a threshold value, if the probability is greater than this threshold then the event is marked as happened otherwise it is not. This model would be best suited to support visualization 2 because we would use our webrobots historical data[16] to pre-determine the thresholds of individual features that best link to campaign success. Although logistic regression would report user's values that differ from thresholds, we ended up using KMeans Clustering for threshold value determination because we want to consider targeted records based on buckets and not the entire dataset for thresholds. This use is explained in more detail under section below.

Gradient Boosting Classifier is a model which is created by "boosting" many weak predictive models into a strong one. The use of this model depends on whether we achieve at least 85% prediction accuracy from other models to drive our

visualizations. Although gradient boosting optimizes loss function, we must be careful in using this model as well since it may overfit data quickly. Random sampling and placing certain on trees would help prevent overfitting.

Model Predictions

Model predictions drive the input for visualizations. There are two integral parts to the model building and prediction process which coordinate to implement server-side modeling.

The first part encodes our cleaned Kickstarter dataset and user input by translating categorical values into integers. This is crucial for model training since models will not interpret string data
and for user input to have same encoding as our main dataset.

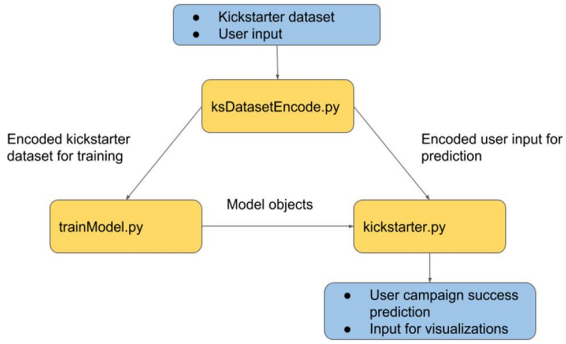
The second part is the actual model training engine. The models are built using ScikitLearn libraries and hyperparameter tuning capabilities.

For predicting the success of Kickstarter campaign, we test accuracy with three different models as mentioned in our "Model Building" section. We end up using Random Forest because it had better prediction than Logistic Regression and GBM ends up overfitting the training data. Random forest achieves ~98% accuracy. We set the n_estimators parameter to be 20 because that is the highest we could go without causing the model to overfit. To validate these three models, we scored the accuracy on 40% data being the test set against 60% being training set. We also further validated Random forest by cross validating across 10 folds which achieved approximately 98% accuracy with similar mean_fit_times. This information can be summarized in the image below.

Logistic Regression	Random Forest Classifier	Gradient Boosting Machine
• L2 ridge regression: avoids overfit and considers all features • LIBLINEAR: classifies large data quickly • Test Accuracy: 96% • Result: Not used for prediction. Random forest provides higher accuracy	• 20 estimators: largest amount to avoid overfitting • Cross Validation 10 folds Test Accuracy: 98% • Result: used to predict success of campaign	• Tried variations of depth and # estimators • Test Accuracy:100% • Result: not used due to risk of overfitting

In addition to predicting whether campaign would succeed/fail, we built a KMeans Clustering model to create suggestions for 'backers_count','goal','pledge_per_backer' features and use 50 buckets. Kmeans will be used to predict which bucket 'backers_count',
'goal', and 'pledge_per_backer' fit into and then that value is compared to the average values of that bucket to provide suggestions. We use filtering for all the other features of visualization 2. Basically, we consider the successful projects and from those, choose the most common label for that feature leading to successful project.

The interaction amongst model implementation files in backend is shown below.



EVALUATIONS

Test Plan

We will perform three types of testing mentioned below.

Usability testing: User-centered testing where we test the application as user has minimal knowledge about it.
Test that user is able to...

1. Load csv file
2. Receive feedback on unsuccessful load
3. Know if user must wait for model run to finish
4. Interact with visualizations 1 & 2 as mentioned above

Unit testing: Test each module (UI, individual visualization component) after implemented.
Test that UI is able to…

1. Accept and validate user's input
2. Output visualizations on one page and altogether
3. Allow user to interact with visuals

Regression testing: Test as we enhance already implemented parts. Use to test each model (random forest, logistic regression, GBM) after initial model build.
Test model by…

1. Use cross validation to check prediction accuracy of logistic regression
2. Output confusion matrix
3. Measure performance of classifier with prediction score

Test Observations
Usability testing: We generated sample CSV files to run through our application to test that user is able to perform the actions as expected above.
Unit testing: Before we integrated application, server, and client side parts, we performed unit testing to ensure user input is accepted, model predictions are valid, and visualizations are successfully constructed.
Regression testing: Models would get validated as they were tested with different hyperparameters such as number of estimators, depth and different types of linearity in logistic regression model. We also tried different training and test set size. Confusion matrix was used for Logistic regression model and cross validation for Random forest.

CONCLUSION
Our product makes use of the best of both worlds: Visualization and Data Analytics to provide an enhanced UX. It takes advantage of the dataset statistics with improved visualization driven by the campaign success/failure prediction using the Random Forest with prediction of ~98% (greater than 85% we initially aimed) and KMeans Clustering. Furthermore, our diverse set of visualizations allows the user to understand their campaign success in various dimensions. Our application was further validated with usability, unit, and regression testing combined to cover frontend, backend, and model capabilities. Future improvements to consider include: provide user with ability to navigate separate pages for visualizations, expand suggestion criteria to provide feedback on how user many increase their chances of Kickstarter campaign success, and train and build models with additional hyperparameter tuning for a larger feature list dataset in order to improve prediction accuracy and accept more diverse user input.

ACKNOWLEDGEMENTS
All members have contributed a similar amount of effort as shown in image below.

Member	Tasks Completed
Abhishek	• User Interface development • Testing • Final poster

Fernanda	• Visualization Design and development • Testing • Final report
Giang	• Data cleaning and Model design • Testing • Final Poster
Rashmi	• Model training and building • Testing • Final report
Tim	• Node.js application architecture • README.txt • Final report

REFERENCES

[1] Mollick, Ethan R., *Delivery Rates on Kickstarter* (December 4, 2015). Available at SSRN: https://ssrn.com/abstract=2699251 or http://dx.doi.org/10.2139/ssrn.2699251.

[2] Marom, Dan and Robb, Alicia and Sade, Orly, *Gender Dynamics in Crowdfunding (Kickstarter): Evidence on Entrepreneurs, Investors, Deals and Taste-Based Discrimination* (February 23, 2016). Available at SSRN: https://ssrn.com/abstract=2442954 or http://dx.doi.org/10.2139/ssrn.2442954

[3] Lamidi, Abedola, *Predicting the success of Kickstarter campaigns* (September 19, 2017). https://towardsdatascience.com/predicting-the-success-of-kickstarter-campaigns-3f4a976419b9

[4] McMahon, Brian, *Predicting Kickstarter Campaign Success* (February 21, 2018). https://medium.com/@cipher813/predicting-kickstarter-campaign-success-a9cf1f81e09

[5] Etter, Vincent and Grossglauser, Matthias and Thiran, Patrick, *Launch Hard or Go Home!: Predicting the Success of Kickstarter Campaigns* (October 7, 2013). https://dl.acm.org/citation.cfm?doid=2512938.2512957 or http://vincent.etter.io/publications/etter2013cosn.pdf

[6] Zhou, Haochen, *Predicting the Success of Kickstarter Campaigns* (December 5, 2017). https://www.stat.berkeley.edu/~aldous/157/Old_Projects/Haochen_Zhou.pdf

[7] Rakesh, Vineeth and Choo, Jaegul and Reddy, Chandan K., *Project Recommendation Using Heterogeneous Traits in Crowdfunding* (May 2015). https://www.researchgate.net/profile/Chandan_Reddy6/publication/280553953_Project_Recommendation_using_Heterogeneous_Traits_in_Crowdfunding/links/55b8b89108aec0e5f43ac16c/Project-Recommendation-using-Heterogeneous-Traits-in-Crowdfunding.pdf

[8] Qiu, Calvin, *Issues in Crowdfunding: Theoretical and Empirical Investigation on Kickstarter* (October 27, 2013). Available at SSRN: https://ssrn.com/abstract=2345872 or http://dx.doi.org/10.2139/ssrn.2345872

[9] Gatautis, Rimantas and Vitkauskaite, Elena, *Crowdsourcing application in marketing activities* Procedia - Social and Behavioral Sciences 110 (2014) 1243 - 1250 https://www.researchgate.net/profile/Elena_Vitkauskaite2/publication/270847840_Crowdsourcing_Application_in_Marketing_Activities/links/55e5795708aec74dbe732a7d/Crowdsourcing-Application-in-Marketing-Activities.pdf

[10] Evers, Mart, *Main drivers of crowdfunding success: a conceptual framework and empirical analysis* (September 2012). https://www.cultuurmarketing.nl/wp-content/uploads/2013/01/Mart-Evers_Master-thesis

_Main-drivers-of-crowdfunding-success_RSM_Marketing-Management_9-2012_PDF.pdf
[11] Ahmad, Fahad S. and Tyagi, Devank and Kaur, Simran, *Predicting crowdfunding success with optimally weighted random forests* (December 18, 2017). https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8286110
[12] Chung, Jinwook, *Long-term Study of Crowdfunding Platform: Predicting Project Success and Fundraising Amount* (August 2015). https://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=5486&context=etd
[13] Mitra, Tanushree and Gilbert, Eric, *The Language that Gets People to Give: Phrases that Predict Success on Kickstarter* CSCW '14 Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (February 15 - 19, 2014) 49 - 61 https://comp.social.gatech.edu/papers/mitra.cscw14.kickstarter.pdf
[14] Hendricks, Drew, *5 Reasons Why Crowdfunding Is The Next Big Investing Trend* (August 27, 2014). https://www.forbes.com/sites/drewhendricks/2014/08/27/5-reasons-why-crowdfunding-is-the-next-big-investing-trend/#1a2e9f7e6c0a
[15] Kickstarter https://www.kickstarter.com/
[16] Webrobots Kickstarter Datasets (September 13, 2018). https://webrobots.io/kickstarter-datasets/\
[17] Crowdfunding.io http://crowdfunding.io/
[18] Paul Belleflamme, Thomas Lambert & Armin Schwienbacher, *Individual crowdfunding practices (2013)*, Venture Capital, 15:4, 313-333, D OI: 10.1080/13691066.2013.785151 https://doi.org/10.1080/13691066.2013.785151
[19] Tableau Online Free Trial Request, https://www.tableau.com/products/online/request-trial
[20] Sundar, Singh, *d3 Reusable slopegraph*, http://bl.ocks.org/eesur/a4679ee453aa9357977c