National College of Ireland

Predicting Key Success Factors for Selecting Crowdfunding Campaigns/Projects using Machine Learning Techniques: A Case Study of Reward Based Crowdfunding Platform

MSc Research Project
FinTech

Chizoba Ezegbu
Student ID: X19109661

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Chizoba Chineze Ezegbu

Student ID: X19109661

Programme: MSc. FinTech **Year:** 2020

Module: MSc. Research Project

Supervisor: Noel Cosgrave

Submission Due Date: 15th August 2020

Project Title: Predicting Key Success Factors for Selecting Crowdfunding Campaigns/Projects using Machine Learning Techniques: A Case Study of Reward Based Crowdfunding Platform

Word Count: 6523 **Page Count**...26.....

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: 15th August 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Key Success Factors for Selecting Crowdfunding Campaigns/Projects using Machine Learning Techniques: A Case Study of Reward Based Crowdfunding Platform

Chizoba Ezegbu
x19109661

Abstract

Reward based crowdfunding platforms provide an avenue for project creators to pitch their project and source for funds in order to finance these projects. Therefore, determining the factors which could influence the successful selection and funding is considered crucial. This field of research has not only been of keen interest to the academia but also the world of finance and technology as well. In recent times, crowdfunding has gained popularity especially with the bottlenecks experienced by entrepreneurs in accessing finance from traditional banks. This popularity has enabled crowdfunding platforms to translate entrepreneurs' creativity and innovation into reality. This study focuses on identifying the success rate and key factors that are instrumental to creators receiving funding from backers for their projects. Machine learning techniques – Random Forest, Decision Tree, Support Vector Machine (SVM) and Naïve Bayes are adopted in predicting the attributes of Kickstarter campaigns which could make projects/campaigns stand out to be selected by backers for funding. The findings of this study reveal that Random Forest has the highest accuracy and using statistical analysis (chi-square), factors such as staff pick, duration, backers count and goal have been identified as the major drivers for Kickstarter project selection as well as funding. In addition, sentiment analysis is done using the Bing - Lexicon Based Approach is conducted revealing positive and negative words that could influence the decision of backers to invest specific Kickstarter projects.

1. Introduction

Crowdfunding refers to the sourcing of financial resources for the funding of entrepreneurs' projects through a collection of online investors. In recent times, the crowdfunding market has gained popularity and has been attributed as a major contributor in the financial services space, thus, accounting for several billions of dollars (Short et al., 2017). It is described as "*the next big thing*" in the Finance space as it focuses on providing seed capital for entrepreneurs and is instrumental in the creation of jobs. It is revealed that crowdfunding has

within the first-five years of its creation; experienced an explosive growth of over $5.1 billion worldwide (Li, Rakesh and Reddy, 2016). As provided in Short et al. (2017), the crowdfunding market has huge potentials of growth as the World Bank has estimated it to account for $300 billion in cumulative transactions by 2025.This booming financial innovative space is described as a concept in which entrepreneurs or start-ups in need of funds for their projects are matched with individuals or organizations that are willingly to invest their money in their projects. This intermediation is achieved through "*technological innovations*" (Hasnan, 2019). It is further revealed in Bouncken (2015) that the idea of crowdfunding gained popularity through the funding of projects in the creative sector of the economy before diversifying to other areas.

According to Forbes and Schaefer (2017), there are different models of crowdfunding markets, these include, Equity Based Crowdfunding, Peer-to-Peer Lending/Debt Based Crowdfunding, Reward Based Crowdfunding and Donation Based Crowdfunding. Short et al. (2017) further explains that these crowdfunding markets vary in line with the nature of investment and the expectations of the prospective investor. This research study will primarily focus on the Reward Based crowdfunding market; however, it will focus on predicting the key success factors for selecting Kickstarter crowdfunding campaigns using machine learning techniques – Naïve Bayes, Random Forest, Decision Tree and Support Vector Machines (SVM). It will also explore sentiment analysis (Lexicon) using Bing in determining positive and negative words that could influence successful funding of projects.

1.1 Motivation

This researcher seeks to examine the crowdfunding space with emphasis on the operational dynamics of the Reward Based crowdfunding market. This study intends to provide insights on the key factors which can influence the successful selection of crowdfunding projects or campaigns using machine learning models. As earlier highlighted, this study will center specifically on the Reward Based crowdfunding market and one of its most popular crowdfunding platforms – Kickstarter. This researcher is motivated by several factors, these include:

- The quest to add to the existing field of study/knowledge as presently limited academic studies has been conducted in this field.

- Drive to improve the present state of accessibility of entrepreneurial loans especially as bottlenecks are experienced with accessing these loans from traditional/commercial banks. Some of these bottlenecks include rigid regulations, high rates of interest, inadequate provision of collateral(s) for the loans, prolonged loan approval process and the exclusion of some categories of startup or entrepreneurial businesses.
- To show the fusion between technology and finance – reveal the accuracy of how machine learning models can determine or predict the expected outcomes of reward based crowdfunding projects; this can be considered as one of the guiding principles for both potential entrepreneurs and investors.

1.2 Objective

To examine and predict the key factors for the successful selection of Kickstarter crowdfunding campaigns or projects using machine learning models.

1.3 Research Question

What factors determine the successful selection of Kickstarter crowdfunding campaigns/projects?

1.4 Contribution

This piece of research work seeks to provide informed analysis of key factors that are considered instrumental in the Kickstarter project selection process. This researcher aims to investigate the specific factors which could influence investors' or backers' decisions to financially invest in entrepreneurs' projects. Historical research has shown that factors such as setting of realistic funding targets as well as the means of communication adopted (the narrative and videos) are considered major factors that have influenced the selection of Kickstarter projects for funding. However, this research will identify or predict other crucial factors that may contribute to the successful selection of Kickstarter projects using machine learning models. Furthermore, this study would serve as a reliable source of historical data which could be useful to Governments, Crowdfunding firms and the FinTech domain in general.

1.5 Limitations of the Study

- Time constraint which has restricted the scope of this research.

- Past projects of other Reward Based crowdfunding platforms will not be considered; also, not all Kickstarter past projects will be analyzed in this study, only a select number will.
- A limited number of variables of the dataset will be considered and analyzed in this study.

2 Related Works

2.1 The Concept of Crowdfunding

Academic research in the crowdfunding space is evolving and dates to 2010. Research studies in this field increased in 2015 and this is attributed to a significant change in the legislative environment (Cai, 2018). Crowdfunding is however, still considered to be in a "*young state of scientific research*" (Bouncken, Komorek and Kraus, 2015). According to Forbes and Schaefer (2017) crowdfunding is a means by which business owners' projects are funded by several online investors. The paper further reveals that it gained popularity over past years and is considered as one of the most profitable means through which startups or entrepreneurs source funds for executing their projects.

Ruhaab and Yisha (2019) note that most research work conducted on crowdfunding focuses on economic and behavioral studies; it will therefore be a step in the right direction if further studies on the application of data science using predictive models can be conducted. The paper further emphasizes on the limited research which incorporates predictive analytics in determining successful crowdfunding campaigns. It is thus, considered as a significant aspect of academic research as it is one of the factors accountable for exponential growth in the financial services industry. The study adopts Logistic Regression and Random Forest for the analysis of the Kickstarter projects. The paper highlights the relevance of appropriate data and states it is one of the important factors; the study however, did not have high accuracy levels and this is attributed to the limitations experienced in the number of features available in the data analyzed.

2.2 Crowdfunding Financing Models

As earlier explained, crowdfunding is the soliciting of funds to finance entrepreneurs' novel projects through a community of online investors, who in turn receive either rewards or equity depending on the crowdfunding model (Roma, Gal-Or and Chen, 2018). Pierrakis and

Collins (2013) reveal that crowdfunding consists of four distinct models; these are explained in below:

Model of Crowdfunding	Type of Contribution	Type of Return	Motivation of the Investor
Equity Based	Investment	Return on Investment	Intrinsic motivation and desire to earn returns
Reward Based	Donations	Tangible and Intangible rewards	Intrinsic motivation and desire for rewards
Donation Based	Donations	Intangible benefits	Intrinsic motivation
Debt Based or Peer-to-Peer Lending	Loan	Loan is repaid with interest in most cases.	Intrinsic motivation and desire to earn interest on loans

Table 1: Models of Crowdfunding *(Source: Pierrakis and Collins, 2013)*

2.3 Participants in Crowdfunding Markets

According to Sharma and Lertnuwat (2016), there are three parties involved in crowdfunding markets and these are explained as follows:

The Creator - This is also known as the initiator or issuer. This is the person /entrepreneur with the business idea or project seeking funding.

The Funder – The Funder is also referred to as an investor. The funder does the selection of projects which are considered viable and investment worthy. This is done online via the platforms, the innovative campaigns that are perceived as profitable are financially supported by these investors/funders with the aim of either receiving rewards, interest or as an act of charity.

The Platform – This is the intersection point of the funder and the creator. It is an interface that unites the fund seeker and the provider of the funds. This is where the projects/campaigns of the entrepreneurs or creators are hosted.

Figure 1 shows the relationship that exists between the three parties involved in crowdfunding:

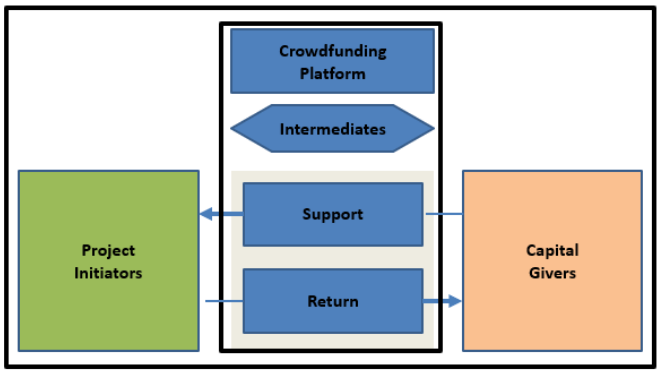


Figure 1 - The Crowdfunding Principle *(Source: Gierczak et al., 2015)*

2.4 Reward Based Crowdfunding Market and its Platforms

Roma, Gal-Or and Chen (2018) describe the reward-based crowdfunding market as a process in which investors fund entrepreneurs' projects in return for rewards which is normally in form of the finished version of the product. Figure 2 depicts this process of reward-based crowdfunding:

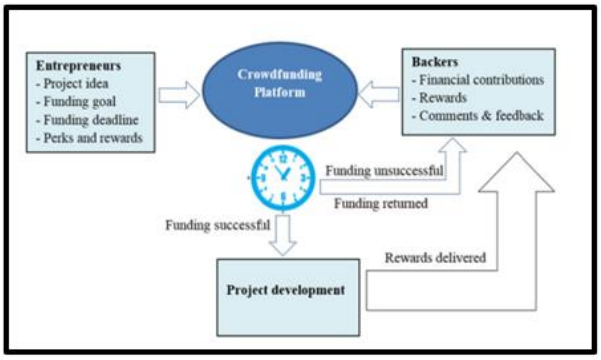


Figure 2: Process of Reward-Based Crowdfunding *(Source: Li, Jarvenpaa and Pattan, 2016)*

Cox and Nguyen (2017) further reveals that this crowdfunding market is popularly known to be a mechanism in which benefactors/entrepreneurs gain access to seed funds for their novel projects/business ideas thereby providing an alternative to traditional banks with respect to sourcing for funds. The paper contributes to this emerging field of study by showing comprehensive empirical evidence on the Reward-based crowdfunding market; it focuses primarily on the success of business-related Kickstarter campaigns.

The findings of the research paper suggest the uneven distribution of reward-based crowdfunding as only a limited number of campaigns are successful. It further shows through

a series of multiple regressions that crowdfunding campaigns or projects that are business inclined perform least compared to other categories especially in those in the creative space – such as the music and dance projects. In terms of the organization of this paper, it clearly depicts how crowdfunding can assist entrepreneurs, the theory behind this, the scope of data as well as a detailed analysis of same. The study also reviews and analyzes Kickstarter projects, but however, experiences limitations in the crowdfunding platforms' extent of coverage which includes the time frame of projects and number of platforms analyzed.

2.5 Kickstarter Crowdfunding Platform

Kickstarter is a popular crowdfunding platform which unites creators of projects with the backers that invest in projects perceived to be successful. In 2016; this crowdfunding platform was the 524th most frequently visited site worldwide (Tran et al.2016). Belleflamme, Omrani and Peitz (2015) provide that this platform launched in 2009 and few years after its creation, over 75,000 projects were funded via the platform with nearly $1.5 billion pledged funds. It further reveals that some of the projects are categorized into "*art, comics, dance, design, fashion, food, games, music, photography, publishing, technology, and theatre*". The creators on this platform offer investors in return for funds or monetary contributions tangible rewards, the investors or funders have diverse interests; however, this tilts more towards the creative projects' category.

Furthermore, Liang, Hu and Jiang (2020) suggest that the achievement of the funding goal of Kickstarter projects determine the success of the project and it adopts the "*all or nothing mechanism*" which means that if the creator of the project does not achieve the goal of the campaign/project, then, refund of funds would be made the funders/investors; thereby, leaving the creator without any funds which were previously raised. This paper investigates the success of Kickstarter projects in line with sustainable development of growing businesses.

It conducts analysis through the perspective of "*information communication and asymmetry*" theory and identifies three categories of information description and quantity – "*word count, picture count and video count*", *information attitude* in which comments is the measure and *information quality* – "*readability and update*". This study examines how these categories affect the success of crowdfunding projects through the adoption of binary logistic regression using data sourced from Kickstarter repository. The results of this study conducted shows that in relation to these three categories of information description, "*word count, picture count,*

video count, comment and update" positively impact on the success of Kickstarter projects while "*readability and comment"* negatively affect its success. Thus, these findings reveal the importance of information description in seeking funding by entrepreneurs in order to execute their projects.

2.6 Key Success Factors for Selecting Crowdfunding Campaigns

Past studies on crowdfunding shows that different variables or factors affect the successful selection of projects for funding by backers. Mollick (2014) examines crowdfunding and the determinants of successful projects. The findings of this paper through analyzing Kickstarter projects show that personal social networking, the geography of the creator as well as the quality of the project using updates and pitches via video makes a campaign stand out and be successful. Also, the research conducted by Xu et al. (2014) reveal that content and updates play major roles in the selection process of reward based crowdfunding market through the analysis of Kickstarter projects. Joenssen, Michaelis, and Müllerleile (2014) share the same opinion as Mollick (2014) and Xu et al. (2014) that communication through updates remains important in determining successful selection of Kickstarter projects by funders/investors. Joenssen, Michaelis, and Müllerleile (2014) however, adopt logistic regression mode land focus primarily on the technology category for analysis. This paper is limited by focusing its analysis on only one project category –technology, if other categories were assessed, it may yield different results.

Another school of thought explains the extent to which social media impacts the success of Kickstarter projects/campaigns. Lu et al. (2014) contributes to this field of study by suggesting that with the evolving nature of social media, campaigns promoted via social media platforms influence its success. The research study analyzes the connectivity between funds raised via crowdfunding platforms online and the corresponding promotions on social media platforms. Rakesh, Chandan and Reddy (2015) further emphasize this by revealing how promotions via Twitter impact on Kickstarter project success; this, however, is dependent on the promoters' social media connections and these assist in providing recommendations of backers for projects. Tran et al. (2016) states that predicting success factors of Kickstarter projects is considered as a crucial research problem area, hence, the need to conduct research that depicts the prediction of success or failure rate. The paper provides analysis for predicting key factors of successful Kickstarter crowdfunding projects using Naive Bayes, Random Forest and AdaboostM1.

It also reveals the extent to which these projects rely on adequate preparation and experience of the creator. The results show that there is a downward trend in terms of success rate for the Kickstarter projects that do not properly prepare and have adequate experience. Furthermore, the paper clearly presents information that smartly organizes and creates projects thereby increasing success rate as well as attracting funders. Koch and Siering (2015) reveal from the study's findings that the sample data of Kickstarter projects show variables such as "*the project description, related images, videos history of previously backed projects*" impact the projects' funding success.

It is, however, interesting that the study notes that the project initiators creation of past projects is of no importance in relation to the projects' success. This study adopts the Logistic Regression machine learning model in arriving at its findings. It is also depicted in Fernandez-Blanco (2020) that the challenge facing the prediction of success or failure rates of Kickstarter projects not only lies with the creator but platforms as well. In the study, it explains that this sparked the research in this field to investigate the behaviour of these crowdfunding platforms as it is perceived that their benefits are instrumental to the success of projects. The findings of this research show that the promotion of crowdfunding campaigns from the start/inception is instrumental to achieving a successful project and be selected by backers.

Also, Kuppuswamy and Bayus (2013) investigate the dynamics of Kickstarter platform. The purpose of the study is to identify the factors that motivate/influence backers to invest in these crowdfunding projects; the study describes the effect of social networking on Kickstarter projects. The study reveals how leveraging on promotional activity via social media platforms such Twitter can significantly impact on predicting the number of investors and funds to be raised for Kickstarter projects. This is similar to the research work conducted in Gerber, Hui, and Kuo (2012). The researchers explain that most potential investors are not inclined to donate funds to projects that have already gotten support from numerous backers.

In addition, the research findings show that the campaign deadline approaches, the project's updates increase as the creators seize it as an opportunity to appeal to backers for funding. The authors emphasize that updates of projects are considered key as it could increase funding especially at the final stages of the campaign. The study was done using Linear Regression Model – Ordinary Least Square (OLS). Another research study, Gierczak (2016) provides a greater insight on the rise of crowdfunding as an alternative source of finance. The

study discusses its major attributes, highlighting the recent development in the market space, various classification methods and its application area.

Furthermore, Dikaputra, Sulung and Kot (2019) point that the key success factors that attract potential backers are categorized into two - Organizational and Marketing factors. The paper highlights setting realistic goals, size of projects/business team, project duration as organizational factors while marketing factors are the narrative type, description through long texts and external websites or comments. The authors emphasize that both organizational and marketing factors are crucial considerations that determine the success of crowdfunding projects or campaigns. Furthermore, the paper reveals that the investors may be more attracted in funding "*all-or-nothing projects*" (reward-based crowdfunding) than "*keep-it-all project*". The paper identified its limitations which include its focus on only reward based crowdfunding market and explains that its conclusions drawn may; however, differ if other crowdfunding markets are considered. Also, the scope of the study is limited to five South Eastern Asian countries.

Furthermore, Josefy et al. contributes to the studies from the cultural perspective by showing evidence that culture is in fact one of the determinants of success in online funding communities. The result of this study indicates projects can stand out to be selected and be successful with the "*support and goodwill*" of stakeholders such as the community. This is with regard to the extent to which the entrepreneurs are aligned with community as well as their culture. The analysis is conducted using descriptive statistics and correlations and it analyzes both GoFundMe and Kickstarter projects. Research studies conducted by Thanh, Anh and Thu (2018) analyze Vietnamese crowdfunding campaigns using logistic regression model and notes that the goal, number of pictures and number of backers significantly impact on the success process. In addition, some research work adopted text mining/sentiment analysis in conducting success prediction. According to Liu (2012), this technique examines people's opinions and are considered key in influencing behaviour. Lexicon based sentiment analysis approach is employed by RM and Halkarnikar (2018) which classifies comments of a post into positive, negative and neutral, highlighting liked and disliked comments.

In summary, most of these discussed literature/studies analyzed their respective data using machine learning techniques such as Logistic Regression, Naïve Bayes, Random Forest, Multiple Linear Regression and AdaboostM1. Thus, from the literature review, the results from the analysis conducted points to updates, project goal, comments as the most important

success factors while less significant factors are the length of funding period, having a project website, quantity of pledge levels made available and the extent of competition on the date of launch.

It is worth noting that Li, Rekesh and Reddy (2016) in the study of predicting factors that determine successful crowdfunding campaigns are opined that adopting the classification model may not be very effective. The authors adopt/introduce the survival analysis based approach in solving this prediction problem where the day of success is perceived as the "*time to reach an event*" and the failed projects regarded as "*censored since the day of success is not known*". The study is also part of the school of thought that points to social network-based features as key attributes that positively influence reward based crowdfunding project selection.

This research study will, however, further investigate this and explore using machine learning models, if other factors could also influence the successful selection of Kickstarter projects.

3 Research Methodology

This research will follow the CRISP-DM (Cross Industry Standard Process for Data Mining) framework. This researcher will adopt the following steps:

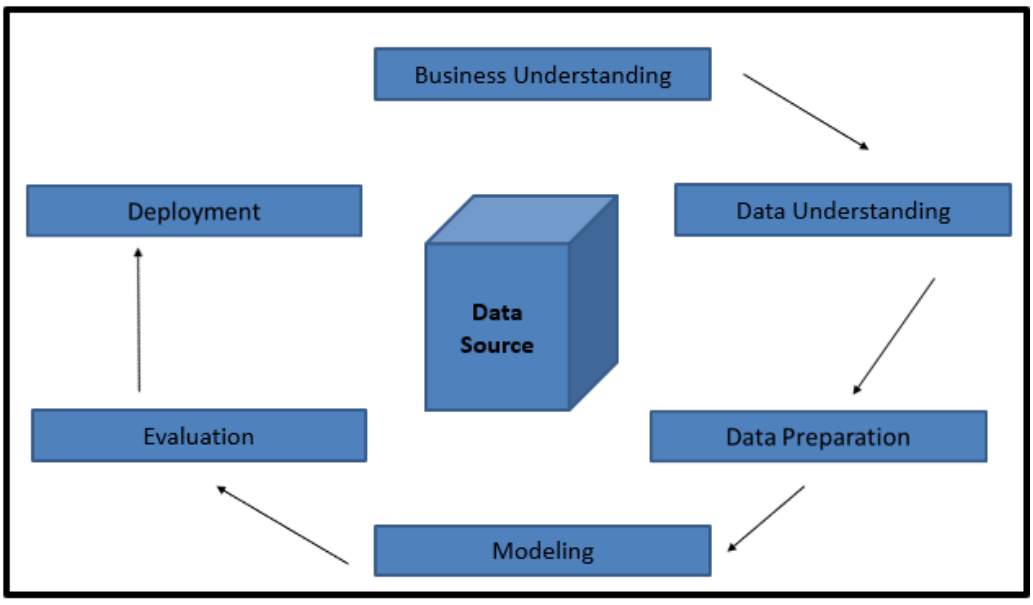


Figure 3 – Steps for Research Methodology

3.1 Business Understanding

This has been extensively explained in the previous sections of this study. It centers on sourcing of funds from the Reward –Based crowdfunding market, with a focus on Kickstarter crowdfunding platform. It explains how entrepreneurs or start-ups source funds for executing their respective projects from an online community of investors via the Kickstarter crowdfunding platform. The leadership and vast popularity attributed to Kickstarter crowdfunding platform is one of the key reasons for selecting it as the source of data to be analyzed in this study (Fernandez-Blanco, 2020). However, on this platform, not all projects or campaigns are successfully funded or meet their funding target. This study is a step in the right direction as it would predict those factors that are considered key which will make campaigns or projects stand out to be selected for funding by investors.

3.2 Data Understanding

This empirical study utilizes publicly available past Kickstarter crowdfunding projects data compiled in 2018, which covers the period 2009 to 2018. As at date, as provided by on the Kickstarter website, since its launch in 2009, the platform has $5,058,434,507 pledged to Kickstarter projects with 183,456 successfully funded projects and 18,062,963 backers[1]. The Kickstarter platform has 7 campaign categories which cut across different segments in the creative industry. These categories include Anthologies, Art, Comics, Crafts, Dance, Design, Fashion and Film. These various Kickstarter project categories have individual pages hosted on the crowdfunding platform's website. This page shows vital information on funding and the process to adopt as well. It displays Kickstarter projects that have been enrolled on the platform and shows the progress these projects or campaigns are making.

The dataset that will be analyzed comprises details of past Kickstarter projects; these are sourced from Webrobots website[2] which hosts relevant information pertaining to this study. The Webrobots site is selected because it is a web crawler that warehouses past Kickstarter projects. Also, collecting data via web makes it possible for gathering reasonably large datasets which have a high-level of data authenticity; this is because the site is actively being used and the data is generated is exclusively for web-based transmission (Chang et al. 2006).

[1] https://www.kickstarter.com/
[2] https://webrobots.io/kickstarter-datasets/

This database is regularly updated, and the data is uploaded once every monthly in CSV and JSON file format. The selected dataset consists of 65,508 past Kickstarter projects/data with 37 variables/features; these cut across the 8 campaign categories. The raw data contain a zipped folder of past projects from 2009 to 2018. This researcher will, however, explore some of these 37 features and carry out a feature selection using 'Boruta' method, in order to derive the best suited features for performing this supervised learning analysis. Also, this dataset reflects factors that determine the successful selection of Kickstarter crowdfunding campaigns/projects.

3.2.1 Classification of Variables

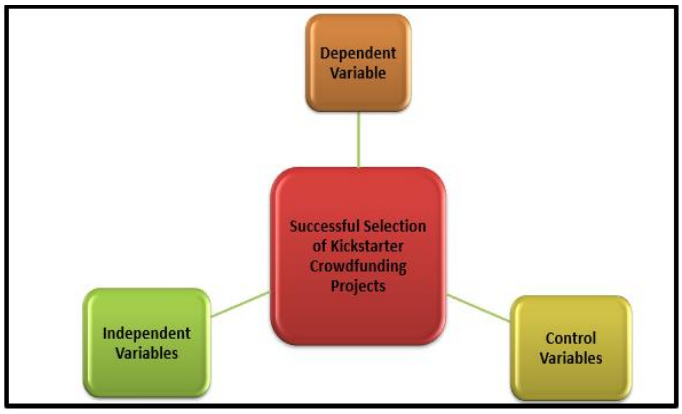


Figure 4 - Classification of Variables

Dependent Variable -The dependent variable for this study would show the state of the Kickstarter campaigns/projects. The project state explains the outcome of the campaign –if the campaign was successful or not. This dataset has five states of the campaigns - live, canceled, suspended, failed and successful. Kickstarter projects that meet 100% or more of its campaign's funding goal (successful and live states) are categorized as 'successful' while those that ae canceled, suspended or failed are categorized as 'failed'. If the creators' projects are selected for funding and meet 100% or more of the campaign goal, such projects keep all the funds for the execution of the project. For this analysis, 55.6% constitute of successful projects while 44.4% are failed projects.

This is illustrated in Figure 5:

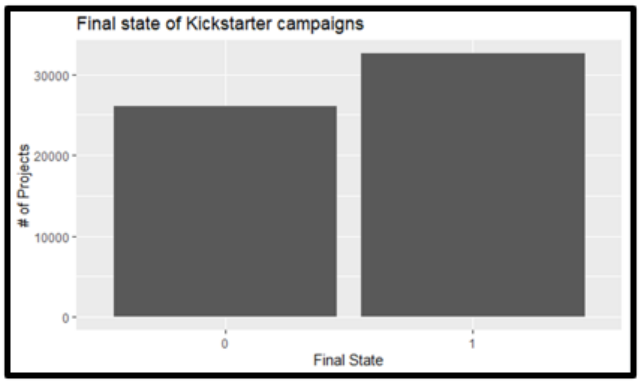


Figure 5 - Dependent Variable (Levels)

Independent Variables – In line with the research work by Bilau & Jorge Pires (2018), the following are considered as the independent variables of the study – Goal, Backers Count, Spotlight and Country.

Control Variables – In line with previous studies conducted by Koch and Siering (2015), the duration of the project as well as its category will represent the control variables of this research. It can therefore be argued that the duration of projects directly impacts the pledged amount for projects; as the longer the duration of a project, the higher the possibility of it having more funds due to greater awareness. Thus, duration and category will be included as control variables in this study.

These variables are summarized in Table 2:

S/N	Variable	Description	Type of Variable
1	State	The outcome of the project	Dependent
2	Goal	The amount of funding set to achieve/raise for the project	Independent
3	Backers Count	Number of investors supporting the project	
4	Spotlight	Updates on the project	
5	Staff Pick	Projects recommended by Kickstarter staff for funding	
6	Country	This is the location of the creator/project initiator	
7	Duration	Period of funding	Control
8	Category	Project area	

Table 2: Kickstarter Projects Variables

Descriptive Statistics

No. Unique	%missing	min	first quartile	Median	third quartile	max	mean	sd
6525	0	-4.293839	-0.6515952	0.0170916	0.6800685	3.577588	0.01172408	0.9806898

Table 3 - Distribution Data

This dataset has a standard deviation of 0.99 approximately 1 and a mean of zero.

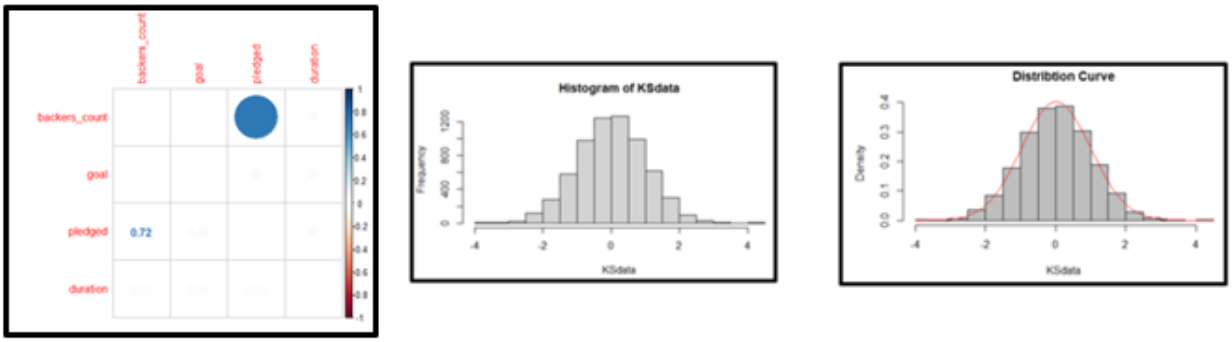


Figure 6 – Kickstarter Statistical Data

The Kickstarter projects under review have a normal distribution as shown in Figure 6 and has 2 correlating variables – "backers_count" and "pledged"; then, "pledged" is removed.

3.3 Data Preparation

This consists of the data pre-processing activities which will be conducted on the raw Kickstarter dataset. It will be pre-processed in line with the requirements of the selected machine learning techniques for this study. This researcher will import the raw data to R and this raw data which was collated in 2018 contains 65,508 projects/observations and 37 variables before any cleaning is done on R. The raw data is converted from CSV file to excel format. This data would need to be redesigned by cleaning/transforming.

This author will convert the following variables in timestamp format to date format (year, month and date) – 'created_at', 'launched_at, deadline and state_changed_at. Missing, duplicate and irrelevant data will also be identified in the dataset. A new column will also be created to show the duration of the projects/campaigns. Some variables will also need to be converted to 'character' for this analysis. As this study is geared towards solving classification problems, the categorical variables in this dataset will be converted to numerical variables, since classification models work best on numeric data.

In addition, as part of the transformation of the data, some of the text data contained in this dataset would also have to be cleaned. These text data include some URLs, symbols and special characters. These variables are cleaned as they are considered insignificant for this analysis. The types or categories of projects will be identified by this researcher and grouped accordingly as well. As the data is being prepared for modeling, it is essential to thoroughly examine the data to ensure that any components that require fixing or cleaning is done.

As part of the data organizing and transformation process, this researcher will remove any irrelevant columns/variables in the dataset as well as restore any wrongly encoded factors. Furthermore, the proportions of the five states of the dependent variables will have to be classified accordingly into two – successful and failed. This dataset will then be normalized and checked if it is balanced in order to avoid any form of bias.

Text Mining (Sentiment Analysis) – After this Kickstarter data has been cleaned and transformed, the data is ready to be processed further. A new data frame will be created for sentiment analysis using the Bing - Lexicon Based Approach on the "Blur" variable of the dataset. This variable is a comment made by the creator about the project/campaign in which funding is sought. Further explanation of these activities will be provided in the implementation section of this study.

3.4 Modeling

A number of past academic research work have adopted classification methods in predicting key factors that influence the successful selection of crowdfunding projects. In order to achieve the objective of this study, this author will consider adopting four relevant machine learning algorithms for this prediction. It is also pertinent that this researcher further investigates and examines if other/additional factors (apart from those already highlighted in the literature review) could be instrumental in determining successful Kickstarter crowdfunding projects. This study intends utilizing the following machine learning techniques for analysis:

Naïve Bayes – According to Liaw and Wiener (2002), it is a popular machine learning model used in solving classification problems. Kaviani and Dhotre (2017) reveal that it performs well on small or large datasets and is also known for its "*minimum storage and fast training*". In addition, the results are easy to comprehend and interpret. This algorithm can also be

utilized by "*unskilled users in classifier technology*". It is fast in predicting as well, this is in comparison to other techniques.

Random Forest – Donges (2019) notes that this flexible method builds an assemblage of trees. It is also very efficient in conducting predictions and it is known to eradicate issues of over fitting due to the large "*number of trees in the forest*". In addition, with the correct/appropriate randomness, this technique is most times considered to be "*accurate classifiers as well as regressors*". This model is perceived to produce accurate results. It selects the best features, thereby leading to good results. Breiman (2001) also reveals that it is proficient in tackling classification problems.

Decision Tree – According to Gepp et. al (2012), decision trees are also known as Classification Trees. This machine learning technique deals with the allocation of data to "*pre-defined groups*" It involves building of trees with the use of splitting rules. It is considered to be a relatively easy algorithm to interpret, handle missing data seamlessly, has simplified splitting rules with easy to understand graphical representations. This model is perceived to be reliable and efficient in conducting predicting analysis (Dhanpal and Gayathiri, 2012). This model is further explained in Bhargav et. al (2013) as a decision support system which adopts decisions based on tree graphs. It is also popular for handling a number of input data such as numeric and text. The model has the ability to process datasets with missing values as well and it is one of the high performing algorithms used in solving classification problems. Its flexibility allows the implementation in various data mining platforms.

Support Vector Machine (SVM) – This model was developed by Vapnik. It is revealed that this machine learning technique has the ability to provide higher classification accuracy when compared to other classification models. Over the years, it has become a popular model for solving both classification and regression problems in diverse fields/areas. This algorithm provides accurate prediction and makes use of kernel functions as well. It also trains large datasets relatively fast (Apostolidis-Afentoulis and Lioufi, 2015).

3.5 Evaluation of models

This research study will evaluate the models applied in this study - Random Forest, Naïve Bayes, Decision Tree and SVM which will determine the level of predictive accuracy in

classifying Kickstarter crowd markets projects into successful and not successful. The following will be used in evaluating the accuracy of models:

I. Confusion Matrix II. K-fold Cross Validation

3.6 Deployment

There will not be any deployment for this research study.

4 Design Specifications

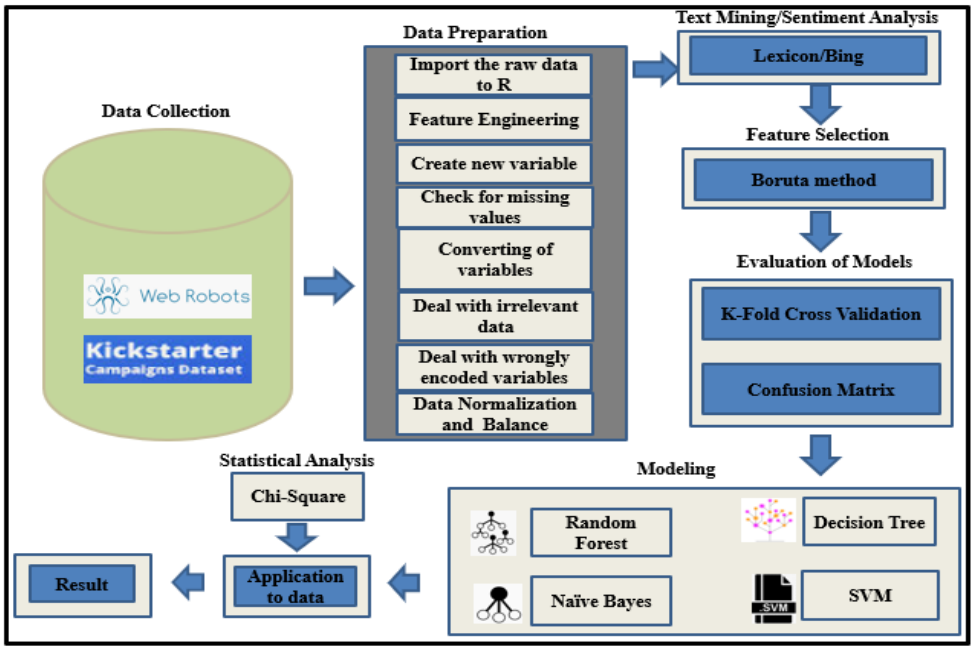


Figure 8 - Architectural Design

Figure 8 depicts the process flow diagram for this research study. Data was sourced from the Kickstarter past projects repository and this will be followed by conducting the highlighted preprocessing steps. Furthermore, text mining using Sentiment Analysis (Bing) and Word Cloud will be done. Feature selection technique (Boruta algorithm) is used to identify the most important attributes whilst others will be removed. Statistical analysis is also done using chi-square. Then, data will be split into train and test. K-fold cross validation and confusion matrix will be used in describing the performance of the test data. Finally, this processed dataset will be fed to the four models selected for this analysis, thus, providing results that will show their respective accuracy levels.

5 Implementation

This section explains the implementation of the selected models for this prediction. It describes the data preparation process undertaken before modeling including the feature selection process using Boruta algorithm. The implementation process is carried out using R programming language on a Windows machine. R has been selected for this implementation because of its ease of use, ample data preprocessing packages and readability of codes. This study is conducted using past Kickstarter projects for the period 2009 - 2018 as per the design in Figure 8.

Data Preparation – The dates on the dataset were in time stamp format and these were converted to date format and a new variable "duration" created. This variable (duration) is calculated by subtracting "deadline" from "launched_at" and these covered a period between 30 to 90 days. Also, variables were converted to factors and characters as well. Encoding is done for categorical variables (state, staff_pick, currency_trailing_code, disable_communication, and is_starrable, "state", "currency" and "country"). The dataset is also checked for duplicates and these are removed, thus, reducing the number of observations/projects from 65,908 to 58,601. Variable correlation or relationship is tested for as well with backers_count and pledged having a high correlation. Furthermore, missing data is checked for and addressed accordingly. These missing data were found in columns irrelevant to the analysis and were addressed subsequently by their removal. The irrelevant columns removed consisted of variables with text, URLs, JSON format and empty columns. This dataset is also checked for balance and it was however a balanced dataset, thus, so there was no need to balance the data.

Feature Engineering – The dates were in timestamp format; these were converted to date format.

Bing Lexicon Based Approach – This categorizes the comments on the 'Blur' variable into either positive or negative words. Some examples of positive words which attract funding include love, like, work, beautiful, support, fun, best, great and modern while negative words are limited, hard, dark, issue, lost, evil, funny, struggle and zombie. Tokenization of the comments in the 'Blur' variable is employed - this involves considering every word as separate and a token. Also, after this stop words such as "how", "and" or "but "are

filtered/removed from the tokenized comments as these mere English words that do not have any predictive insights (Lai, Lo and Hwang, 2017).

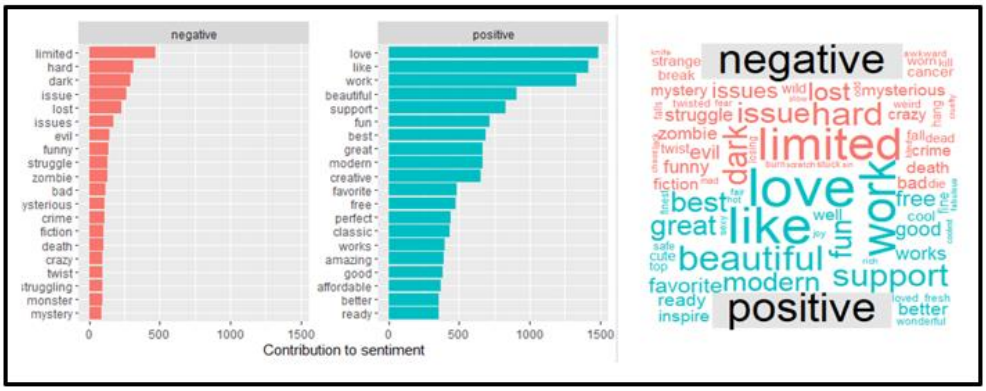


Figure 9 - Bing - Lexicon Based Approach

Feature Selection - Boruta Algorithm is adopted for the selection of important features in this analysis. After 22 iterations, 11 attributes were confirmed important while 3 (currency_trailing_code, current_currency, id,) were confirmed unimportant and removed.

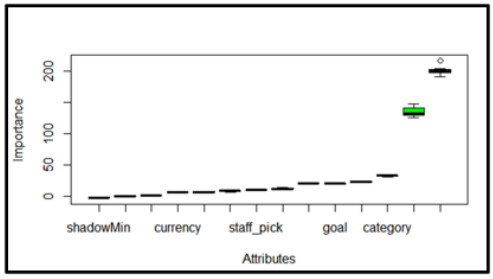


Figure 10 - Feature Selection using Boruta Algorithm

Statistical Analysis – Chi-square was used in determining the factors that are key or influential to the successful selection of Kickstarter projects. The results reveal that "duration", "staff pick", "goal" and "backer's count" are of high significance to attaining success as these factors showed significant relationship with the dependent variable (state) and had p-values < 0.05.

Models Implemented – The models implemented in this research are Random Forest, Naïve Bayes, Decision Tree and Support Vector Machines (SVM). The dependent variable is "state" which has "successful" and "failed" values. The data is split into two – training and test in a 70:30 ratio. The evaluation of models during training is done using 10-fold cross validation and confusion matrix. The evaluation metrics selected for this thesis are Accuracy, Kappa, Sensitivity and AUC. These metrics were calculated for all models and are presented

in tables for comparison. From this analysis, the models performed well achieving accuracy of over 80%. Section 6 depicts this and further provides a detailed evaluation/comparison of the results.

6 Evaluation

6.1 Results

The main focus of this research is to evaluate the performance of select models and assess their suitability for solving this problem. The evaluation metrics are selected because of their suitability in solving classification problems. The metrics are derived from the confusion matrix values.

6.1.1 Performance Matrix

Performance Matrix

Model	Accuracy	Kappa	Sensitivity	AUC
Random Forest	0.9134	0.8245	0.9027	0.9122
Decision Tree	0.9032	0.8037	0.8824	0.9011
SVM	0.851	0.7031	0.9128	0.8572
Naïve Bayes	0.8353	0.6727	0.9105	0.8428

Confusion Matrix: RF

Model RF	Actual Successful	Actual Failed
Predicted Successful	7041	764
Predicted Failed	759	9016

Confusion Matrix: Decision Tree

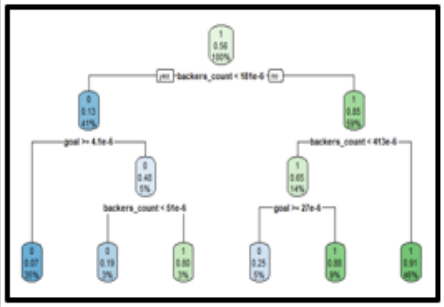
Model Decision Tree	Actual Successful	Actual Failed
Predicted Successful	6883	784
Predicted Failed	917	8996

Confusion Matrix: SVM

Model SVM	Actual Successful	Actual Failed
Predicted Successful	7120	1939
Predicted Failed	680	7841

Confusion Matrix: NB

Model Naïve Bayes	Actual Successful	Actual Failed
Predicted Successful	7102	2198
Predicted Failed	698	7582



Decision Tree

Table 4: Summary of Results

s/n	variable	x-squared	df	p-value
1	Duration	2198.4	91	< 2.2e-16
2	Backers_count	34120	1758	< 2.2e-16
3	Staff_pick	4034.4	1	< 2.2e-16
4	Goal	4672.9	2361	< 2.2e-16

Table 5: Chi-Square Results

6.2 Discussion

Four models are applied in this study in predicting success of kickstarter projects and these results are shown in section 6.1.1. The models used are Random Forest, Decision Tree, Support Vector Machine (with Linear Kernel) and Naïve Bayes. Data normalization and encoding of variables had positive impacts on the performance. Chi-square was employed answering the research question on factors that determine the successful selection of Kickstarter projects. The result identified the key factors duration, staff pick, goal and backer's count. This is an important finding as it adds to the already existing key factors found by past researchers. The classification accuracy result obtained from the models (as shown in Table 4) revealed that Random Forest is the best model which shows that it is good enough to conduct this prediction, closely followed by Decision Tree. The accuracy obtained in SVM ranked it third out of the four models under study with Naïve Bayes had the lowest accuracy. Also, the models' misclassification errors are relatively low with all being less than a quarter of their respective accuracy results.

7 Conclusion and Future Work

In summary, the stated objective of this thesis has been met and the research question answered as well. This study was able to build four classification models as shown in the results. From these models which all had relatively high accuracy rates, it can be deduced that Random Forest performed best and can be chosen as best fit based on its accuracy. It is however inferred from the results obtained that each of the models had high accuracy rate, this demonstrates that classification methods are highly effective in predicting. This study can be useful to reward-based crowdfunding markets' participants by providing a guide to seeking and investing funds via the Kickstarter platform, highlighting the success rates and key factors. Further studies can be conducted showing a comparative analysis of the success rate of Kickstarter projects prior to and post Covid-19 pandemic.

References

Apostolidis-Afentoulis, V. and Lioufi, K., 2015. SVM Classification with Linear and RBF Kernels.

Belleflamme, P., Omrani, N. and Peitz, M., 2015. The Economics of Crowdfunding Platforms. *SSRN Electronic Journal*.

Bhargava, N., Sharma, G., Bhargava, R. and Mathuria, M., 2013. Decision Tree Analysis on J48 Algorithm for Data Mining. 3(6).

Bilau, J. and Pires, J., 2018. What Drives The Funding Success Of Reward-Based Crowdfunding Campaigns?. 12(2).

Bouncken, R., Komorek, M. and Kraus, S., 2015. Crowdfunding: The Current State Of Research. *International Business & Economics Research Journal (IBER)*, 14(3), p.407.

Breiman, L., 2001. *Machine Learning*, 45(1), pp.5-32.

Cai, C., 2018. Disruption of financial intermediation by FinTech: a review on crowdfunding and blockchain. *Accounting & Finance*, 58(4), pp.965-992.

Chang, C., Kayed, M., Girgis, M. and Shaalan, K., 2006. A Survey of Web Information Extraction Systems. *IEEE Transactions on Knowledge and Data Engineering*, 18(10), pp.1411-1428.

Cox, J. and Nguyen, T., 2018. Does the crowd mean business? An analysis of rewards-based crowdfunding as a source of finance for start-ups and small businesses. *Journal of Small Business and Enterprise Development*, 25(1), pp.147-162.

Dhanpal, R. and Gayathiri, P., 2012. Credit Card Fraud Detection using Decision Tree for Tracing Email and IP. 9(5).

Dikaputra, Sulung and Kot, 2019. Analysis of Success Factors of Reward-Based Crowdfunding Campaigns Using Multi-Theory Approach in ASEAN-5 Countries. *Social Sciences*, 8(10), p.293.

Donges, N., 2019. A Complete Guide to the Random Forest Algorithm.

Fernandez-Blanco, A., Villanueva-Balsera, J., Rodriguez-Montequin, V. and Moran-Palacios, H., 2020. Key Factors for Project Crowdfunding Success: An Empirical Study. *Sustainability*, 12(2), p.599.

Forbes, H. and Schaefer, D., 2017. Guidelines for Successful Crowdfunding. *Procedia CIRP*, 60, pp.398-403.

Gepp, A., Wilson, J., Kumari, K. and Bhattacharya, S., 2012. A Comparative Analysis of Decision Trees Vis-`a-vis Other Computational Data Mining Techniques in Automotive Insurance Fraud Detection.

Gerber, E., Hui, J. and Kuo, P., 2012. Crowdfunding: Why People Are Motivated to Post and Fund Projects on Crowdfunding Platforms.

Gierczak, M., Bretschneider, U., Haas, P., Blohm, I. and Leimeister, J., 2016. Crowdfunding: Outlining the New Era of Fundraising. *FGF Studies in Small Business and Entrepreneurship*, pp.7-23.

Hasnan, B., 2019. A framework for Crowdfunding platforms to match services between funders and fundraisers. *Journal of Industrial Distribution & Business*, 10(4), pp.25-31.

Joenssen, D., Michaelis, A. and Müllerleile, T., 2014. A Link to New Product Preannouncement: Success Factors in Crowdfunding. *SSRN Electronic Journal*.

Kaviani, P. and Dhotre, S., 2017. Short Survey on Naïve Bayes Algorithm. *International Journal of Advance Engineering and Research Development*, 4(11), pp.607-611.

Kickstarter. 2020. [online] Available at: <https://www.kickstarter.com/> [Accessed 11 August 2020].

Koch, J. and Siering, M., 2015. Crowdfunding Success Factors: The Characteristics of Successfully Funded Projects on Crowdfunding Platforms.

Kuppuswamy, V. and Bayus, B., 2013. *Crowdfunding Creative Ideas: The Dynamics Of Project Backers In Kickstarter*.

Lai, C., Lo, ,. and Hwang, S., 2017. Incorporating Comment Text into Success Prediction of Crowdfunding Campaigns. p.156.

Liaw, A. and Wiener, M., 2002. Classification and Regression by RandomForest. 2(3).

Li, Y., Rakesh, V. and Reddy, C., 2016. Project Success Prediction in Crowdfunding Environments. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16*.

Li, Z., Jarvenpaa, S. and Pattan, N., 2016. Cinetics: Fueling Entrepreneurial Innovations through Crowdfunding. *Journal of Information Technology Teaching Cases*, 6(2), pp.75-83.

Liang, X., Hu, X. and Jiang, J., 2020. Research on the Effects of Information Description on Crowdfunding Success within a Sustainable Economy—The Perspective of Information Communication. *Sustainability*, 12(2), p.650.

Liu, B., 2012. Sentiment Analysis and Opinion Mining.

Lu, C., Xie, S., Kong, X. and Yu, P., 2014. Inferring the impacts of social media on crowdfunding. In: *In Proceedings of The 7th ACM International Conference on Web Search and Data Mining*. pp.573–582.

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

Pierrakis, Y. and Collins, L., 2013. Crowdfunding: A New Innovative Model of Providing Funding to Projects and Businesses. *SSRN Electronic Journal*.

Rakesh, V., Choo, J. and Reddy, C., 2015. Project Recommendation Using Heterogeneous Traits in Crowdfunding.

RM, M. and Halkarnikar, P., 2018. Text Analytics of Web Posts' Comments Using Sentiment Analysis.

Roma, P., Gal-Or, E. and Chen, R., 2018. Reward-Based Crowdfunding Campaigns: Informational Value and Access to Venture Capital. *Information Systems Research*, 29(3), pp.679-697.

Ruhaab, M. and Yisha, W., 2019. Dare to Venture: Data Science Perspective on Crowdfunding. *SMU Data Science Review*, 2(1), pp.1-16.

Sharma, S. and Lertnuwat, N., 2016. The Financial Crowdfunding with Diverse Business Models. *Journal of Asian and African Social Science and Humanities*, 2(2), pp.74-89.

Short, J., Ketchen, D., McKenny, A., Allison, T. and Ireland, R., 2017. Research on Crowdfunding: Reviewing the (Very Recent) Past and Celebrating the Present. *Entrepreneurship Theory and Practice*, 41(2), pp.149-160.

Thanh Tu, T., Anh, D. and Ha Thu, T., 2018. Exploring Factors Influencing the Success of Crowdfunding Campaigns of Startups in Vietnam. *Accounting and Finance Research*, 7(2), p.19.

Tran, T., Dontham, M., Chung, J. and Lee, K., 2016. How to Succeed in Crowdfunding: a Long-Term Study in Kickstarter.

Web Scraping Service. 2020. *Kickstarter Datasets*. [online] Available at: <https://webrobots.io/kickstarter-datasets/> [Accessed 11 August 2020].

Xu, A., Yang, X., Rao, H., Fu, W., Huang, S. and Bailey, B., 2014. Show me the money!. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14,*.