National College of Ireland

Prediction of Crowdfunding Project Success Probability using Machine Learning

MSc Research Project
MSc in FinTech

Ashwani Teotia
Student ID: x17160715

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ashwani Teotia
Student ID:	x17160715
Programme:	MSc in FinTech
Year:	2019
Module:	MSc Research Project
Supervisor:	Noel Cosgrave
Submission Due Date:	12/08/2019
Project Title:	Prediction of Crowdfunding Project Success Probability using Machine Learning
Word Count:	5500
Page Count:	17

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 September 2019

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):	

Prediction of Crowdfunding Project Success Probability using Machine Learning

Ashwani Teotia
x17160715

Abstract

Crowd markets have established as alternative financing means where the borrowers preserve funding from the crowd in small increments. However, crowd markets mainly reward-based crowd markets are facing the problem of very low project success rate. Consequently, it's leading to low revenues for platforms and the project creators are challenged to meet fund targets. These problems are due to lack of project optimization tools for project creators. This research will contribute academically to establish a research area to provide a solution in terms of probability of success of crowd markets projects and on the other hand will act as a foundational work for practitioners to optimize project success decisioning systems. This research is evaluating the predictive accuracy of probability predicted by the machine learning models. Support vector machine (SVM), k-nearest neighbors (k-NN) and Random forest are used to predict probability. Sorting smoothing method (SSM) is used to predict the estimated actual probability. Linear regression is used to evaluate the relation between predicted probability and the estimated actual probability. Probability is the better parameter to provide project success insights facilitating borrowers for the decision making, increased revenue for the platform and increased transparency for the funders. Results of this research found random forest model have the highest predictive accuracy as (R-Squared=0.93) in predicting the probability of the success of the project and would be used to build decisioning systems.

Contents

1 Introduction

Reward-based crowd markets are the platforms mostly online used to raise the funds required by the founders of the project from a larger number of people. This form of funding is called crowdfunding which comprises of crowdsourcing and microfinancing (Wallmeroth, Wirtz and Groh; 2018). Reward-based markets have emerged as an established source of entrepreneurial financing (Li, Rakesh and Reddy; 2016). Reward-based crowdfunding has a rich history and a noticeable example is presented in the research where the creator raised the funding to translate a book offering to include names in the acknowledgment for two gold guineas (Janků, Kučerová and Dařena; 2019).

Crowd markets have been long used, nevertheless, after the year 2008 financial crisis led to a worldwide liquidity issue have given rise to FinTech innovative crowd markets platforms (Mollick; 2014). These platforms are online distributed applications acting as alternative financing channels by connecting project creators and the funders to facilitate project funding. These platforms have been widely accepted by the crowd market's participants, serving the ease of connectivity and are backed up by the government policy to promote sustained growth (Wallmeroth et al.; 2018). Research related to reward-based crowd markets are primarily divided into two categories one is to find the factors affecting the success of the project to achieve target fund and the other is to find accuracy in prediction of project funding success (Xu and Zhu; 2018).

Reward-based crowdfunding mainly works on two models, Keep-It-All or All-Or-Nothing the first model provides creator the amount raised from the user even the target is not achieved while the second model provides funds to the creators only when the target is met (Xu and Zhu; 2018).

Ren, Xu, Zhao, Zhu, Junliang and Chen (2018) have observed the need to optimize the resources at funder and creator ends in All-or-Nothing models for the project's success, moreover, it is stressed by the project parameters such as duration of the project to get timely funding.

This research used reward-based crowd market platform named Kickstarter. This platform is based on the All-Or-Nothing model. The main problem reward-based crowd markets facing is the low success rate of 30% to 40% (Yu et al.; 2018; Li et al.; 2016; Zhou et al.; 2016). The failure of the projects is due to the fact of lack of project optimization tools, insights are unavailable on unsuccessful projects and the current solutions are intuitive by predicting the output of a project in terms of label- successful or failure (Zhou et al.; 2016).

This research establishes the importance of probability in predicting the success of crowd markets projects. Probability serves as a better parameter providing insights to the unsuccessful reasons of a project. Probability of success is an important parameter to define risk management solution for the crowd market platforms and the decision-making systems for creators and the funders (Menon, Jiang, Vembu, Elkan and Ohno-Machado; 2012; Yeh and hui Lien; 2009).

1.1 Project Specification

As per the literature review conducted from the year 2008 onwards and analysis of the Kickstarter platform data, specifications of this research work is setup to propose the following research gap.

- Research Question: *What is the accuracy of the probability of success of reward-based crowd markets projects using machine learning?*
- Research Objective
  - *By using machine learning models predict optimized success probability of projects*
  - *To derive the estimated actual probability of success of a project using sorting smoothing method*
  - *To find the predictive accuracy of the probability of success of projects*
- Purpose: The purpose of this research is to find the predictive accuracy of predicted probabilities from the machine learning models where the models are used to predict the success of reward-based crowd markets projects. The prediction of the success of crowdfunding projects is a binary classification problem. The probability is an intermediate output of the models which act as an input parameter to decide on project output by applying a threshold. Research to predict the accuracy of success prediction of the markets projects is widely available however lacking in the area to predict the accuracy of success probability. Interestingly disruptive to the traditional financial systems crowd markets have established as highly used, however, there is huge scope available to optimize success prediction of the market's projects.

The approach used to carry out this research is supervised machine learning methods. Firstly, the model's performance output matrix is evaluated iteratively to get the maximum accuracy. By using probability and the found classes from tuned models, sorting smoothing method is used to calculate the actual estimated probability. Further regression is used to evaluate the accuracy of the predicted probabilities. The result of the regression will reveal the predictive accuracy of the probability of success of projects basis on the models used.

This research work is using the Kickstarter [1] platform global project data from Kaggle [2] . Scope of this research is limited to the analysis of Kickstarter projects originated in the United Kingdom.

The remainder of this paper is organized as follows. Reward-based crowd markets project success as a domain is introduced in section 1.2. Section 2 presents the body of literature assessed to complete this research. Section 3 provides details regarding methodology followed to perform research. Section 4 gives the design specifications of the methods used to implement research steps. Section 5 briefs the implementation related details of the research. Section 6 evaluates and discusses the results found by carrying out this research. Section 7 concludes this research work with the suggestions for future work.

[1] Kickstarter website: `http://www.kickstarter.com`

[2] Kaggle website: `http://www.kaggle.com`

1.2 Reward-Based Crowd Markets Success

Reward-based crowd markets is a type of crowdfunding where a project creator or a venture needs the resources such as finances and using crowd markets, they seek the finances from the funders in return of rewards. However, funding is not the only goal of the creators, they may even use the crowdfunding to know the product's market demand. The returns may include finished products, future services and special acknowledgments (Wallmeroth et al.; 2018; Mollick; 2014).

The success of a reward-based market project is to meet the targeted funding amount. Platforms are online available to a global audience and many times the funded amount is way more than asked by the creators as in case of pebble watch (Yu et al.; 2018). On the other hand, projects fail at a rate of 60% to 70%. As per the analysis of Kickstarter historical data the success rate of the platform can be observed in figure 1.

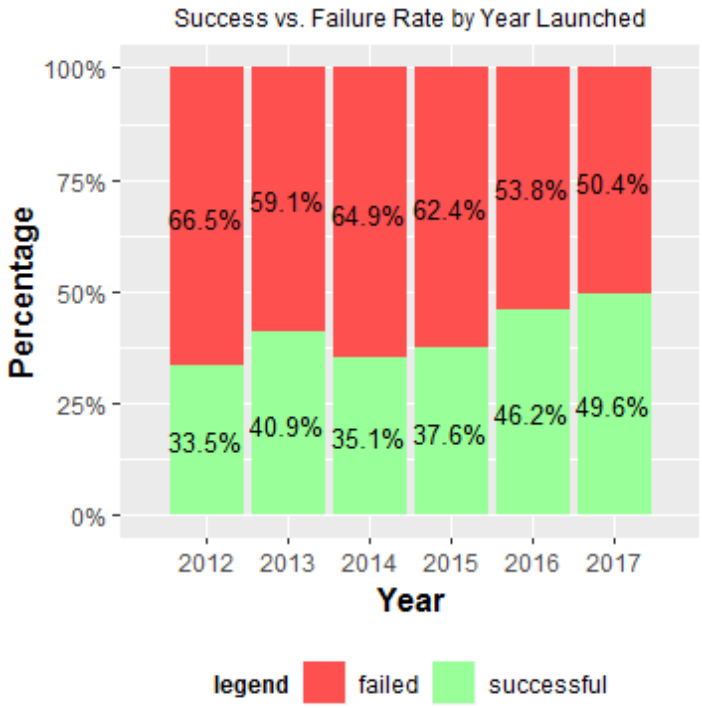

Figure 1: Kickstarter Project Success Rate (United Kingdom)

The components of an end to end machine learning-based solution for the crowd markets success is summarized in figure 2. The goal of such a solution is to enable the project creator to optimize their projects for success. The components of the solution can be used as a risk management solution for platforms while the funders can have transparency on the success of project they funded.

This research is scoped to the project success probability component in the figure 2.

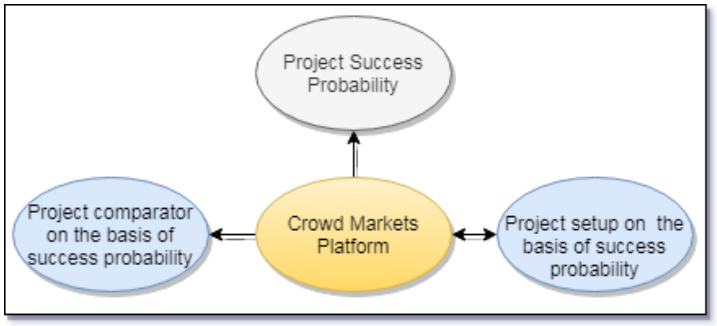

Figure 2: Project Success Components

2 Related Work

The success of reward-based crowd markets projects has been widely researched from 2008 onwards. Scope of the research mainly pertains to four categories: success prediction of the projects to meet funding target (Yu et al.; 2018), selecting factors having high impact on success of projects (Ahmad, Tyagi and Kaur; 2017), reward product supplies optimization of the reward projects with product design strategy (Wang et al.; 2018); and the dynamics of project progress with respect to time, i.e., daily funding prediction of the projects (Ren et al.; 2018). This research work corresponds to the first category and a part of the work has an overlap with finding influential factors affecting project success during data exploration.

Ren et al. (2018) research focused on tracking and forecasting per day funding amount, which is of high significance to all the three participants to derive success similar to this research work.

Ahmad et al. (2017) have designed random forest with optimally weighted classifiers to select features signifying success prediction and the model to predict success. Accuracy found was 94.289% and have confirmed period of the project and the funding target have high importance in success prediction of the project. This model is limited to apply at the inception of the project.

Chung and Lee (2015) have studied the effect of project's temporal features to predict the range of expected pledge and the performance of models to predict success. They found AdaboostM1 having accuracy around 75% for various feature combination (Kickstarter projects temporal features + twitter user profile features). Li et al. (2016) have evaluated the effect to include temporal parameters to find that it dramatically improves the performance of success predictions at the start of the projects. This work will use the duration of the project as a temporal feature.

Chung and Lee (2015), have noticed two pledge peaks in the start and end of the experiments, and an affinity to pledge towards the deadline of the project. By using machine learning models, they have observed a consistent increase in project success rate by including project's static, project temporal features and twitter user profile features.

Mollick (2014) proposed as a holistic view that the underlying project quality, personal networks and the location of project origin are the most important factors to determine the success of projects. Pan, Guo and Yan (2018) have established research by considering the maximum set of features, location as categorical, funding goals as numeric, description of the projects as textual. They have achieved an accuracy of 72.78% by using neural networks. Data was collected from Kaggle with 100K rows. They have analyzed the

wrongly predicted projects and found most of the errors are due to data, i.e. data is not labeled properly and the issues in recording textual data. Since this research is using similar dataset, data anomalies detection and the feature engineering will be applied to accurately classify the projects.

Beier and Wagner (2015) conducted an empirical study on a reward-based market(100-days.net) using 740 projects. They observed the effect of high updates and the media richness in the project presentation concludes to increase in success rate. Their model is based on off/on platform communication activities to increase project success. Dataset used in this work lacks data regarding updates to the project though the data regarding media richness feature is included using the length of the project name.

Mollick (2014) work has found that the projects fail by a large amount or succeeds by a low margin and the project success is linked with the quality of the project. Zhou et al. (2016) studied antecedents to meet the funding target. They found two features project description and the project creator experience and expertise playing an important role in the project success. Project quality is made up of a complex group of features, on the other hand, project creators are usually not experienced to judge the project quality. This observation will be used in the research to optimize the models.

Sawhney, Tran and Tuason (2017) have found that moderate success predictions can be made by using project content, linguistic features and the meta-information gathered from multiple projects. They used data from Kickstarter having 160K observations. SVM is used for classification and the accuracy found is 71%. They found the generalization of SVM improves as the data grows. The research will consider the scope to optimize SVM accuracy by providing multiple region data in addition to the United Kingdom. Though the larger training time is envisioned for SVM to use the bigger dataset.

Greenberg, Pardo, Hariharan and Gerber (2013) research have noticed the need for a feedback tool regarding project success to provide decision control to the project creators similar to the current research. They used different decision trees and SVM with different kernels to predict success. The accuracy of SVM and decision tree found is 54.43% and 60%-70% respectively.

Much of the crowd markets research work includes analysis of binary class labels output, success or fail specifically in terms of the confusion matrix. Prediction accuracy is acting as the primary factor to evaluate the effectiveness of the model in identifying the classes. Feedback tools have been recommended on top of accuracy solutions to facilitate decisioning to the project creators. Research has established probability-based solutions to have more value in terms of decisioning to value risk management solution. Feedback tools can be built on output probability to provide decision control to all participants of the crowd markets (Keramati, Yousefi and Omidvar; 2015; Yeh and hui Lien; 2009).

Menon et al. (2012) have mentioned the importance of probability in place of binary labeled output in various applications such as defining meta classifiers, non-standard learning tasks and use predictions to take actions.
Importance of accurate probability measurement to assist in accurate predictions have been observed in other domains as well, as in healthcare to model clinical decisions (Steyerberg et al.; 2010); in weather forecasting, to forecast the weather in terms of a score (Brier; 1950); in financial services to predict credit defaults (Yeh and hui Lien; 2009).

Niculescu-Mizil and Caruana (2005) examined the relationship between the predicted probability of the models and the corresponding posterior probability. They have noticed machine learning models have characteristic to pull probabilities away from the

(0,1) or pushing probabilities towards (0,1). They corrected the biased probabilities by using two calibrations methods- Platt scaling and Isotonic regression. They have found two categories of models based on goodness of predicted probability, some models have better probability prior to calibration while others post calibration. These finding will be observed in this work.

Yeh and hui Lien (2009) work have pointed out the importance of probability in risk management solutions and have compared six machine learning models by devising a method to compare model predicted probability with the estimated actual probability. They have used 25K observations of a credit card default dataset. Neural networks are found to have better-predicted probability than the other compared models. For neural networks predicted probability and estimated actual probability regression had (R2=0.9647). This methodology was further used by (Keramati, Yousefi and Omidvar; 2015) to observe that the Optimally Weighted Fuzzy K-Nearest Neighbor (OWFKNN) algorithm performs best to classify credit defaults based on probability. This research work will use this methodology and will include changes as per domain and dataset requirements.

3 Methodology

The cross-industry standard process for data mining (CRISP-DM) methodology is used is used to realize this research (Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer and Wirth; 2000). This method covers all the phases of a data analytics project from inception to deployment and involves iterations for continuous improvements. First phase of this methodology includes business value context understanding for the analysis project. Next step is data understanding to explore the data from business and analysis perspective. This phase succeeds by the data processing to include data transformation and prepare the data for modelling. Further phases include modelling, evaluation and deployment stages respectively (Kelleher, Namee and D'Arcy; 2015).

3.1 Dataset Selection

As discussed above, the focus of this research is to find predictive accuracy of the probability of success of crowd markets, mainly the reward-based crowd markets. This research has used the Kickstarter reward-based crowd markets data. Kickstarter is a well-known crowd market started in 2008. It is one of the FinTech platform representatives of reward-based crowd markets.

- Data Sourcing: Data for this research is sourced from Kaggle . This data has been used by the researcher before to conduct research (Yu et al.; 2018) and Kaggle is owned by Google LLC an established organization. Thus, the data is considered trustworthy to conduct research. Projects originated from the United Kingdom are in the scope of this research. The global dataset has 378,661 rows and 15 features. Dataset [3] has projects data created from April 2009 to January 2018. Considering deadline or the target date dataset has the range from May 2009 to March 2018 (Yu et al.; 2018). Dataset has following fields available: Id, Name, Category, Main_category, Currency, Deadline, Goal, launched, Pledged, State, Backers, Country, Usd.pledged, Usd_pledged_real, Usd_goal_real.

[3]Dataset used: `https://www.kaggle.com/kemical/kickstarter-projects/version/7`

- Data Cleaning: The data collected from Kaggle is cleaned up using R. Data had the issues in metadata interpretation, missing values and garbled data in textual fields. Cleaning of the collected data is discussed in detail in the design and implementation section 4 5.

- Data Processing: This phase has explored the data further to process/ transform the data as per findings in the data understanding phase and required by the models in the modelling phase. Two new fields- Length of the project name and Duration of the project are created to include as have been studied in research (Ren et al.; 2018) and (Ahmad et al.; 2017).

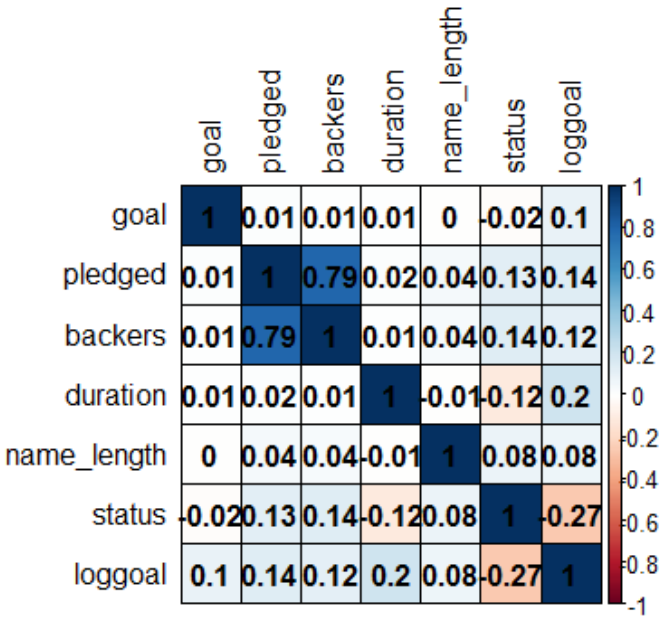

Figure 3: Features Correlation-Pearson

3.2 Modelling

This phase included planning and execution of the required models in this research. Support vector machine (SVM), k-nearest neighbors (k-NN) and Random forest are used to predict crowd markets probability. Models are tuned for the optimized accuracy. More details available in the sections 4 5.

The contribution of each model in this work is as follows:

- K-nearest neighbor (K-NN): This is an analogy-based machine learning model, which derives analogy from the k-nearest neighbor to predict the class of the required sample to validate. Analogy is based on distance from the k-nearest neighbors. This model does not produce/ derive a prediction formula in training step while uses the complete training data during validation. Thus, the advantage is fast training but disadvantage is to have increased time and space during validation (Zhang, Ye, Essaidi, Agarwal, Liu and Thau Loo; 2017).

- Support Vector Machine (SVM): As used in research by Greenberg et al. (2013), SVM finds optimal decision boundary to separate classes. It is a discriminative

classifier uses linear and non-linear planes to classify data. An advantage in using SVM is kernel trick which transforms the data to represent data into higher dimensional space without increasing feature space. SVM is a black box in terms of kernel trick and have high training time for the larger dataset.

- Random Forest: Random forest builds multiple decision trees during the training period. It chooses the decision of most of the trees as the decision of classification. Random forest has less training time while consistent on higher accuracy. These avoid overfitting due to the selection of multiple decision trees. This model can work fine even the data have missing or inconsistent observations (Ahmad et al.; 2017).

The sorting smoothing method (SSM), presented in the research by Yeh and hui Lien (2009) is used in this work. This technique is used to calculate estimated actual probability by using the class predictions made by the models. This method includes sorting the predicted output based on predicted probability and then derives estimated real probability with the help of predicted probability and the class with respect to each observation. Linear regression is used to find the extent of the predicted probability of success representing the estimated real probability of success.

Keramati et al. (2015) have used SSM to baseline their work and the other researchers have used this method in credit default domain to conduct comparative study. This method converts discrete binary classes into continuous representation to approximate the actual probabaility. Scatter plots and linear regression are used to calculate predictive accuracy of probability of success.

Predicted probability and the estimated real probability, independent and dependent variables respectively are continuous and representing similar information- the probability of success; the relation assumed between these variables is linear and is verified using linear regression.

R-squared as the output of the linear regression is used to calculate the representation of variance in estimated actual probability by the model predicted probability.

Yeh and hui Lien (2009) have used scatter plots and linear regression to observe the relation between predicted probability and the estimated real probability and further they find predictive accuracy of probability.

3.3 Evaluation

This research includes two levels of evaluation, one is to evaluate applied models k-NN, SVM and Random forest for the accuracy to classify crowd markets projects into success and failure, the other evaluate the predictive accuracy of the models predicted probability.

Models accuracy will be evaluated by using confusion matrix, while the R-squared with other coefficients of linear regression will be used to predict predictive accuracy of the probability of the models used. Scatter plots are used to visualize the relation between predicted probability and the estimated real probability. Following evaluation parameter are used to evaluate the results.

- Confusion matrix: This is the performance matrix used in the classification problems to know various performance parameters and error types. Accuracy parameter of the confusion matrix is used to determine the efficiency of the models in this work.

Accuracy is selected as the parameter to evaluate since the more accuracy in classification leads to more number of better probabilities and further have significance to establish relation between predicted and the estimated actual probability.

- R-Squared (co-efficient of determination): R-Squared value is used in this work to evaluate the representation of estimated actual probability by the predicted probability. By definition R-squared value determines the proportion of variance in the target variable that can be represented by the explanatory variable. This evaluation parameter is selected since only one independent variable is required to model.

- k-fold cross-validation: This technique is used in this research to evaluate the performance of models. K is selected as 10. 80% of the observations are used in the cross-fold validation during training. Using this technique model performances are averaged on 10 model runs to avoid over and underfitting.

4 Design Specification

High level design of this research work is presented in the figure 4.

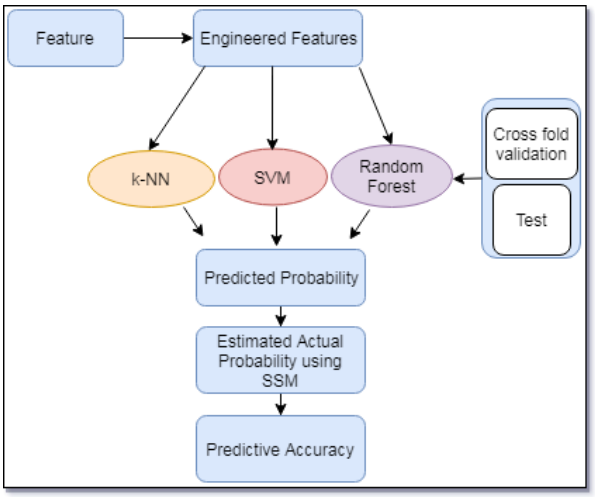

Figure 4: High Level Design

- Data Sourcing: To conduct the research to predict the accuracy of prediction probability from the machine learning models, the real trustworthy data had to be available. This is considered as the first step of the project. Data is collected from a reliable source- Kaggle corresponding to the Kickstarter platform. The projects available in the dataset are available on the Kickstarter website to review. To define the project scope as per training time of the models and complexity of the regional data only the projects in the United Kingdom are considered for the work of this research. After subsetting to consider only United kingdom projects, dataset has 33,672 rows and 15 features. Data have six statuses of the projects as live, canceled, suspended, undefined, failed and successful. There are 35.37% successful projects while the unsuccessful projects are 52.21% out of all the categories.

Table 1: Dataset Fields

Field	**Data Type**	**Description**
Project Id	number	Project Id
name	category	Project name
Category	category	Sub-Category of the project
Main_Category	Category	Project maincategory
Currency	Category	Project base currency
Deadline	Date	Project deadline
goal	Number	Project goal amount in USD
Launched	Date	Project launched date
pledged	Number	Pledged amount in project currency
State	Category	Project status out of six statuses
Backers	Number	Number of backers
Country	category	Country of Project origin
usd.pledged	Number	Amount pledged in USD (Conversion by Kickstarter)
Usd_pledge_real	Number	Amount pledged in USD (Conversion by fixerio API)
Usd_goal_real	Number	Project target amount in USD

- Data Cleaning: Data quality report (DQR) is prepared to understand the data anomalies. Data is further explored to find anomalies and to prepare data planning report (DPR) Kelleher et al. (2015). Backers and goal fields have outliers though these values are of significance to the business. Backers and goal fields are normalized in the processing. Following are the cleaning actions performed on the collected dataset.
  - Rows are removed with respect to NA values in name field.
  - Launched and deadline features are converted from string to Date format to use in data processing.
- Data Processing:
  - ID column is removed as not giving a business feature value and being independently unique.
  - Duration is created as a new field and is derived from the deadline and launched fields. This field have negative values, rows with -ve values were removed.
  - Project name length a new field is created from the name field to include in analysis.
  - Rows are filtered to contain only the rows with state of either successful or failed
  - Country column being representative of the dataset and independently unique is removed from the dataset.
  - Currency column is removed since was not having a significance value in terms of location or currency of the amount in pledge and goals. For pledge and goal currency used is USD.

– pledged and backers columns have high correlation as in figure 3 and are redundant to use together thus only backers field is included in final feature table.

Since k-NN and SVM required numerical columns for modelling Category column is one-hot encoded into 15 columns to identify a category in terms of 0s and 1s. Numerical values are scaled and centred as required by SVM. Analysis base table with the fields as in Table 2 is prepared as the result of data processing to use with the machine learning models used.

Table 2: Dataset Features

Feature	**Data Type**	**Description**
goal	Number	Project goal amount in USD
Backers	Number	Number of backers
Duration	Number	Time duration of the project (days)
Project Name Length	Number	Numeric length of project name
State	Category	Project status out of six statuses
Category	Category	Project main-category (Encoded into 15 features)

Zhang et al. (2017) have pointed out that fund raising is a rare event and to resolve the issue of data imbalance, the sampling techniques would be used. Furthermore, the authors have used three sampling techniques (over-sampling, under-sampling, and cost-sensitive learning) to resolve the data imbalance issue. Since the classes to classify in this research are not highly imbalance thus random sampling is used.

Seeding is used to initialize random number generator with a fixed state at the time of data sampling and the training of the model to maintain reproducibility of the results.

5 Implementation

R language was used to implement the research artefacts. Quad core Windows machine is used to carry out implementation work required to realize this research. Random forest and SVM were taking a long time to train in the global dataset and even on the scoped dataset. The research was started with the global dataset of Kickstarter and scoped to United Kingdom projects as required per design section 4. Parameters were tuned by using hyperparameter optimization for all the models used in the research. k-fold cross validation (with k=10) is used to measure model accuracy by avoiding under and overfitting.

Models implemented in this research are as follows:

k-nearest neighbor (k-NN), SVM and Random forest models are implemented using caret package of R. Analysis base table used for the research have six categorical and numerical features. Five features are independent features. "state" is the dependent feature in the analysis base table with the values of "successful" and "failed".

Independent feature, Category is one-hot encoded to be used in k-NN and SVM while used as categorical in random forest. Numerical columns are preprocessed to center and scale. 10-fold Cross validation is used to evaluate models during training. The data is divided into two parts to be used in training and testing in the ratio of 80:20. Validation data part is avoided due to the use of cross-fold validation.

By using hyperparameter optimization, the value of k=5 found for k-NN is 5, for Random forest mtry is found as 15 while SVM provided maximum output for gamma=0.03743058 and C=128.

Linear and non-linear kernels are used with SVM to optimize accuracy. Non-linear kernel is considered for final model run.

For Random forest, random and grid search are conducted to find optimized mtry.

By using tuned models output, analysis base table is appended with two columns- predicted probability of success and the class output of the model prediction.

Appended analysis base table is used to predict estimated actual probability by using sorting smoothing method (SSM). Linear regression model is used to estimate linear relation between predicted probability and the estimated actual probability.

6 Evaluation

6.1 Results

The results of this research work comprise of two sets- one is the intermediate performance matrix output from machine learning models to classify projects into success and failure and the other is the linear regression model output generated by comparing the models predicted probability and the estimated actual probability. Accuracy results as set one are collected from the tuned model and the confusion matrix. Second set of results are collected as output coefficients of linear regression.

Result Set 1: Performance matrix of the models

Table 3: Performance Matrix

Model	**Accuracy**	**AUC**	**kappa**
k-NN	0.8163	.8118	0.6213
SVM	0.8664	0.8521	0.718
Random Forest	0.9279	0.9266	0.851

Table 4: Confusion Matrix: k-NN

Model k-NN	**Actual Failed**	**Actual Successful**
Predicted Failed	TN=2916	FN=510
Predicted Successful	FP=572	TP=1893

Table 5: Confusion Matrix: SVM

Model: SVM	**Actual Failed**	**Actual Successful**
Predicted: Failed	TN=3242	FN=541
Predicted: Successful	FP=246	TP=1862

Table 6: Confusion Matrix: Random Forest

Model: Random Forest	**Actual Failed**	**Actual Successful**
Predicted: Failed	TN=3255	FN=192
Predicted: Successful	FP=233	TP=2211

Result Set 2: Linear Regression Output

Table 7: Linear Regression Output

Model	Intercept	Slope	R-Squared
k-NN	-0.06465	1.1624	0.8222
SVM	-0.1409	1.2030	0.8703
Random Forest	-0.028157	1.070224	0.9315

6.2 Discussion

Models used in this research to predict probability have seen many turning points as per the results presented in the result section 6.1 and explained as follows:

Models showed significant performance improvements as the data used for modelling is processed to suit the models used. Data standardization to center and scale the data, one-hot encoding to convert categorical data into numerical had significant impact on performance. One-hot encoding turned category column to 15 dummy features and has significantly improved the classification accuracy of SVM from 0.81 to 0.87 while only slight improvement for k-NN . Thus, the category feature played a significant role to increase classification accuracy and in turn better predicted probabilities. However random forest in grid search tune have the better classification accuracy than SVM and k-NN and the model have better probability prediction.

As per confusion matrix in result section, the order of classification accuracy of the models used in this work in ascending order is k-NN, SVM and random forest.

Scatter plot visualization of predictive probability and the estimated actual probability has interesting findings, scatter plots of all the models have mainly four regions- the estimated actual probability remains close to zero in the start, then rises, then remains at one and then decreases. This is according to the smoothing done by using SSM. Further this is due to high accuracy of SVM and random forest, moreover success and failure observations are balanced.

Regarding predictive accuracy of probability of success of the projects random forest have the slope close to one and the intercept close to zero thus random forest is representing the predictive accuracy of probability of success and is equal to the R-Squared value (0.9315).

Results of linear regression from result section 6.1 and the scatter plots in figures 5 6 7 presents the relationship in the predicted and actual estimated probabilities.

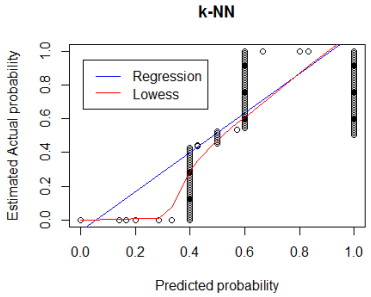

Figure 5: k-NN Scatter

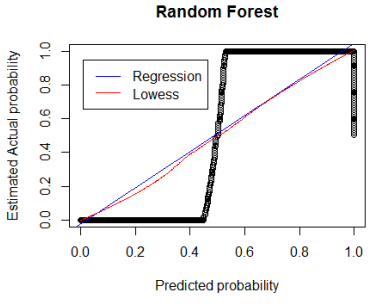

Figure 6: Random Forest Scatter

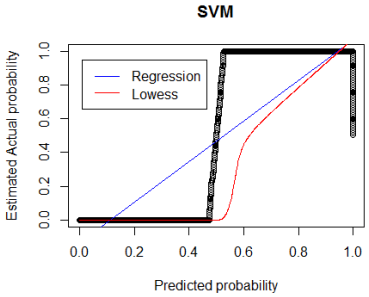

Figure 7: SVM Scatter

7 Conclusion and Future Work

In summary, this research work objective was to evaluate the predictive accuracy of the probability of project success in reward-based crowd markets. k-NN, SVM and Random forest machine learning algorithms are used for this research with sorting smoothing technique to find the estimated actual probability of project success. Linear regression is further used to test the representation of the estimated actual probability of project success by the probability predicted by the models. Accuracy improvement of the classification is found to be highly related to feature inclusion and feature processing.

Selected models were tuned to achieve higher accuracy in the performance matrix. Random forest classifies the projects with more accuracy than the SVM and k-NN. R-Squared value for the random forest is observed as 0.93. Results showed a high predictive accuracy of the probability of success using random forest and thus random forest would lead to establishing a decisioning system for the crowd markets platforms participants. As discussed earlier the projects for this research were scoped in the United Kingdom, in future, the other locations can be added and the feature set could be extended to add locations to analyze with deep learning.

This research can be considered as a stepping-stone towards building a risk management system enabling decision making to the crowd markets platform like Kickstarter.

References

Ahmad, F. S., Tyagi, D. and Kaur, S. (2017). Predicting crowdfunding success with optimally weighted random forests, *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS)*, pp. 770–775. doi: 10.1109/ICTUS.2017.8286110.

Beier, M. and Wagner, K. (2015). Crowdfunding success: A perspective from social media and e-commerce.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability.

Chapman, P., Clinton, J. M., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. R. H. and Wirth, R. (2000). Crisp-dm 1.0: Step-by-step data mining guide.

Chung, J. and Lee, K. (2015). A long-term study of a crowdfunding platform: Predicting project success and fundraising amount, *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, HT '15, ACM, New York, NY, USA, pp. 211–220.
URL: *http://doi.acm.org/10.1145/2700171.2791045*

Greenberg, M., Pardo, B., Hariharan, K. and Gerber, E. (2013). Crowdfunding support tools: Predicting success and failure, pp. 1815–1820.

Janků, J., Kučerová, Z. and Dařena, F. (2019). Behavioural insights from crowdfunding financing: Power of nudges.

Kelleher, J. D., Namee, B. M. and D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*, The MIT Press.

Keramati, A., Yousefi, N. and Omidvar, A. (2015). Default probability prediction of credit applicants using a new fuzzy knn method with optimal weights', in tavana, m. and puranam, k. (eds.), *Handbook of research on organizational transformations through big data analytics*, pp. 429–465. Hershey, PA, USA, doi: 10.1109/ICTUS.2017.8286110.

Li, Y., Rakesh, V. and Reddy, C. K. (2016). Project success prediction in crowdfunding environments, *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, ACM, New York, NY, USA, pp. 247–256.
URL: *http://doi.acm.org/10.1145/2835776.2835791*

Menon, A. K., Jiang, X., Vembu, S., Elkan, C. and Ohno-Machado, L. (2012). Predicting accurate probabilities with a ranking loss, *Proceedings of the ... International Conference on Machine Learning. International Conference on Machine Learning* **2012**: 703–710.

Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study, Vol. 29, pp. 1–16.
URL: *http://www.sciencedirect.com/science/article/pii/S088390261300058X.*

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning, *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, ACM, New York, NY, USA, pp. 625–632.
URL: *http://doi.acm.org/10.1145/1102351.1102430*

Pan, C., Guo, Y. and Yan, C. (2018). Predicting the success of crowdfunding. `https://cs230.stanford.edu/projects_spring_2018/reports/8289614.pdf`.

Ren, X., Xu, L., Zhao, T., Zhu, C., Junliang, G. and Chen, E. (2018). Tracking and forecasting dynamics in crowdfunding: A basis-synthesis approach, pp. 1212–1217.

Sawhney, K., Tran, C. and Tuason, R. (2017). Using Language to Predict Kickstarter Success, p. 9.

Steyerberg, E., J Vickers, A., Cook, N., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. and Kattan, M. (2010). Assessing the performance of prediction models a framework for traditional and novel measures, *Epidemiology (Cambridge, Mass.)* **21**: 128–38.

Wallmeroth, J., Wirtz, P. and Groh, A. (2018). Venture capital, angel financing, and crowdfunding of entrepreneurial ventures: A literature review, *Foundations and Trends® in Entrepreneurship* **14**: 1–129.

Wang, G., liu, q., Zhao, H., Liu, C., Xu, T. and Chen, E. (2018). Product supply optimization for crowdfunding campaigns, *IEEE Transactions on Big Data* **PP**: 1–1.

Xu, Y. and Zhu, N. (2018). Successful factors and prediction of crowdfunding on wechat, *American Journal of Industrial and Business Management* **08**: 946–962.

Yeh, I.-C. and hui Lien, C. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, *Expert Systems with Applications* **36**(2, Part 1): 2473 – 2480.
URL: *http://www.sciencedirect.com/science/article/pii/S0957417407006719*

Yu, P., Huang, F., Yang, C., Liu, Y., Li, Z. and Tsai, C. (2018). Prediction of crowdfunding project success with deep learning, *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*, pp. 1–8.

Zhang, Q., Ye, T., Essaidi, M., Agarwal, S., Liu, V. and Thau Loo, B. (2017). Predicting startup crowdfunding success through longitudinal social engagement analysis, pp. 1937–1946.

Zhou, J., Lu, B., Fan, W. and Wang, G. (2016). Project description and crowdfunding success: An exploratory study, *Information Systems Frontiers* **20**.