PREDICTING THE ENTREPRENEURIAL SUCCESS OF CROWD FUNDING CAMPAIGNS USING MODEL-BASED MACHINE LEARNING METHODS

Mohammed Abrar Ul Hasan[1],Neha Unnisa[2], Mohd Imroz[3]
PG Scholar, Department of CSE, Deccan College of Engineering and Technology, Hyderabad, Telangana
abrarhasan123@gmail.com
Asst Professor, Department of CSE, Deccan College of Engineering and Technology, Hyderabad, Telangana
nehaunnisa@deccancollege.ac.in
Asst Professor, Department of CSE, Deccan College of Engineering and Technology, Hyderabad, Telangana
mohdimroz@deccancollege.ac.in

ABSTRACT: -

Finding the metrics that make these campaigns remarkably effective is a typical phenomenon that piques the curiosity of corporations, investors, and entrepreneurs engaged in crowd funding activities, especially on the Kick starter website. This study aims to compare the chosen various machine learning algorithms, discover model-based machine learning approaches based on performance assessment that forecast campaign success, and analyse the significance of key predictive factors or characteristics based on statistical analysis. In order to fulfil our study goals and optimise our understanding of the dataset, feature engineering was carried out. After that, cross-validation techniques were used to compare the machine learning models, which included Logistic Regression (LR), Support Vector Machines (SVMs) in the forms of Linear Discriminate Analysis (LDA), Quadratic Discriminate Analysis (QDA), and Random Forest Analysis (bagging and boosting), with respect to their test error rates, F1 score, Accuracy, Precision, and Recall rates. Bagging and gradient boosting (the SVMs), two of the machine learning models used for predictive analysis, were found to be more reliable techniques for predicting the success of Kick starter projects based on test error rates and other classification metric scores acquired across the three cross-validation approaches. The performance of important statistical learning techniques, which direct the selection of learning techniques or models and provide a gauge for the final model's quality, has allowed us to meet the main research goals of this work. Nonetheless, Bayesian semi-parametric methods should be explored in subsequent studies. Using an unlimited number of parameters to obtain information about the underlying distributions of even more complex data is made easier by these techniques.

1. INTRODUCTION

Another way to get money for a project or concept through online donations is through crowd funding. Instead than approaching a select number of knowledgeable investors, an entrepreneur uses crowd funding to raise external finance from a vast audience (the "crowd"), in which each individual provides a very modest amount[1]. Online crowd funding websites began to appear in the mid-2000s, coinciding with the internet's phenomenal rise in popularity. Among these websites are Kickstarter, IndieGoGo, GoFundMe, and Kiva. Founded in 2009 and headquartered in Brooklyn, New York, Kickstarter has grown to become one of the most well-known websites for crowd funding. Since then, initiatives have received billions of dollars from a wide range of global investors. Within the crowd funding community, Kick starter has established itself as a venue for innovative ideas to potentially get funded. Art, comics, crafts, dance, fashion, design, film/video, food, gaming, journalism, music, photography, publishing, technology, and theatre are just a few of the genres in which these could fit. Due to the amount of money raised through Kickstarter, investors, businesses, and entrepreneurs are becoming more interested in this new means of funding. These groups frequently ask: Is it possible to pinpoint the essential elements of a successful Kickstarter campaign? The Pebble Time smart watch, which raised over 20 million US dollars (USD), the Coolest Cooler, which raised over 13 million USD, and the Exploding Kittens board game, which raised over 8 million USD, are a few noteworthy Kickstarter success stories. A Kick Starter campaign must receive 100% of the cash requested by the project founder within a predetermined window of time (between 1 and 92 days) in order for the campaign to be considered successful. If not, it is considered unsuccessful. This strategy is regarded as "all or nothing" as, in the event that a campaign is a failure, backers get their money back. In addition to protecting the backers, this business model encourages creators to set reasonable goals[1]. In reward-based or donation-based crowd funding, contributors exchange their monetary contributions for intangible prizes or other forms of remuneration. The amount of pay is directly correlated with the contributions made [2] . Online, crowd fundraising takes place on numerous platforms. Numerous websites for crowdsourcing and fundraising exist, each with unique features that cater to the campaign objectives of their clientele. Comprehending the distinct attributes of these websites is important for the triumph of crowd fundraising.

Previous research has indicated that among the various forms of crowd fundraising campaigns, potential funders find the reward-based kind to be especially appealing [3, 4]. Among these reward-based crowdsourcing websites, IndieGoGo, GoFundMe, and Kickstarter are noteworthy. One of the most well-known reward-based crowd fundraising systems in the world, Kicks Tarter, serves as the foundation for this study. It organises fundraising events for a range of artistic endeavours, including games, films, music, and technology. Kickstarter initiatives typically have an aim that is quite clear. Three different sorts of actors typically participate in the crowd financing model: creators who submit projects for funding, backers who promise financial support for the initiator's concept, and a mediator. Both sides are mobilised by the Kicks tarter platform. It is accessible to producers and supporters worldwide. In fact, Kicks Tarter has hosted more than 170,000 successfully financed projects since its founding in 2009, earning over 4.5 billion dollars from more than 16 million backers. Because Kick Starter uses a "all-or-nothing" financing model, no one is paid for a pledge towards a project until it meets its financial target, lowering the risk for all parties. Each project has a target budget limit or goal that must be reached within a predetermined amount of time; a project is deemed successful only when this objective is accomplished. In the event that a project's funding target is not met, backers will not be billed, creators will not get any of the cash committed, and they are not forced to finish projects without the necessary finances. A five percent charge is subtracted by Kicks tarter from the money raised through a campaign once a project is successfully funded. It is this indicator of a campaign's success or failure that allows academics to use algorithms for classification In order to accomplish their objectives, potential participants (creators and backers) typically want to know the likelihood that a Kickstarter campaign will succeed. This could protect them from spending time and resources on proposals that have little to no chance of getting funded and, more significantly, steer them towards initiatives with better chances of success. There are a few things that contribute to a successful crowd fundraising campaign[5], like the developer's reputation and past performance[3]. A number of factors were looked into for the success of crowd fundraising, including the campaign's content, financial incentives, the qualities of the developers and sponsors, the campaign's duration, deadlines, goal expectations, and the accuracy of the information provided[6–8]. It can be challenging to identify which variables are crucial, though. This paper sheds light on prediction model selection and optimization in crowd funding success by presenting a case study in which feature selections and contrasted statistical models are used to assess successful crowd funding forecasts. The study's specific objectives are to assess and contrast functional models that can accurately forecast Kicks tarter campaigns, determine the significance of important variables or characteristics, such as the number of backers and the amount of money pledged in USD, and evaluate various machine learning algorithms, such as Logistic Regression (LR), Support Vector Machines (SVMs), which include linear and quadratic discriminate analysis, and finally random forests (bagging and boosting).

2. SCOPE OF THE PROJECT

This study aims to analyses and compare functional models that can accurately predict Kicks tarter campaigns; evaluate the significance of important variables or features, like the number of Backers and the amount pledged in USD; and compare various machine learning algorithms, such as Decision Tree (DT) with GridSearchCV, Gradient Boosting with GridSearchCV, and Logistic Regression (LR), which includes linear and quadratic discriminate analysis.

3.EXISTING SYSTEM

- Big data has developed into a useful tool. Big data is being used in many different industries, including telecoms, e-commerce, banking and insurance, military and surveillance, oil and gas development, oil and gas networks, academics, healthcare, aerospace, and transport planning.
- But only if we can make data communicate will this vast amount of data start to become useful.
- Data analytics is the instrument that uses data to tell stories in an understandable and visual way.
- According to a well-known social listening tool called Brand Watch, as of May 2019, over 3.4 billion people used social media platforms globally.

3.1 PROPOSED SYSTEM

- The use of crowd funding by businesses, investors, and entrepreneurs to raise capital for their initiatives is growing in popularity. Specifically, Kick starter has grown in popularity as a crowd funding platform. Effective prediction models are necessary to assist campaign developers, as the success of these campaigns is not always assured.
- In order to tackle this problem, a recent study aimed to examine the efficacy of several machine learning models in forecasting campaign success and find critical predictive characteristics for successful crowd funding campaigns. Several machine learning models, such as logistic regression, linear discriminate analysis (LDA), quadratic discriminate analysis (QDA), random forest analysis (bagging and boosting), and decision trees with GridSearchCV, were applied

after feature engineering was used to extract pertinent information from the dataset. Model comparison revealed that Decision Trees with GridSearchCV performed better than the other models.

3.1.2 PROPOSED SYSTEM ADVANTAGES:

- It can manage intricate and non-linear correlations between the response and predictor variables. By methodically identifying the ideal hyper-parameters that produce the best accuracy, grid search CV improves prediction accuracy.
- This lowers the possibility of over-fitting and enhances the model's ability to generalize to new data.

4. Feature and variable selection

An attempt is made to determine potential correlations between the dataset's continuous variables. In order to do this, a correlation plot for a number of chosen variables was created (see Fig. 4). Some continuous variables have very positive associations when you examine the plot in more detail. PledgedUSD and promised, for instance, have a strong correlation. Given that the data in both variables is remarkably similar, this makes sense. The same may be stated for a number of other continuous variables, including "days_spent_making_campaign" and "days_inception_to_deadline.". It is significant to remember that a high correlation between these variables is a sign of multicollinearity and could lead to inaccurate statistical conclusions. Variance Inflation Factor (VIF), condition indices, and variance decomposition proportions are used as detection tools to find and fix multicollinearity problems. The overall impact of the regressors' dependences on each other's variance is measured for each term in the model by the variance interval function (VIF)[9]. Multicollinearity is indicated by one or more big VIFs.

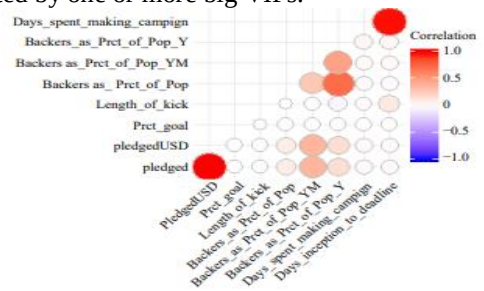


FIG 4 Correlation plot of continuous feature variables

"cursorily" using a model that is linear. The findings pertaining to the variance decomposition proportion and VIF measures for the continuous variables "length of Kick starter," "backers count," "goal," and "pledge per person" made it easier to exclude the other continuous variables. It is impractical for investigators or researchers to include every possible predictor in a model when dealing with extremely large datasets that have a multitude of potential technical predictors, like the dataset used in this Kickstarter project. This is because a significant number of these variables might not be related to the outcome being predicted. Predicting an outcome and identifying a "parsimonious" subset of variables that are connected to the outcome may be of relevance in certain situations. This implies that we can identify the key variables for study using a dimension reduction approach or procedure. In this instance, we examine the application of the Least Absolute Shrinkage and Selection Operator (LASSO), which helps choose the subset of variables that minimises prediction error to help investigators who are interested in making predictions[10]. In this case, a few less influential variables' coefficients are compelled to equal zero. The factors that are deemed most relevant or contributing are retained. We are also presented with a variable important measure via the random forest approach or the criterion known as Gini important or Mean Decrease in Impurity (MDI) that determines the importance of each feature[11]. The variables that were ranked as more contributive or significant in minimising prediction error when both methods were used were the goal, backers count or number of backers, pledge per person, length of Kick starter project, project categories, time factor, and population factor.

5.METHODOLOGIES

5.1MODULE:

- **Data Collection**
- **Dataset**
- **Data Preparation**
- **Model Selection**
- **Analyze and Prediction**
- **Accuracy on test set**
- **Saving the Trained Model**

5.2 MODULE DESCRIPTION

5.2.1.Data Collection:

Gathering data is the first significant stage in the actual creation of a machine learning model. This is a crucial phase that will have a cascading effect on the model's quality; the more and better data we collect, the more effective the model will be.Data can be gathered using a variety of methods, including physical interventions, web scraping, and more.

5.2.2. Dataset:

The dataset consists of 1383 individual data. There are 9 columns in the dataset, which are described below.

1. Category – sub category of the project/ startup
2. main_category – main category of the project/ startup
3. currency – what is the type of the currency used

4. Goal - the final goal amount
5. Pledged - in return for lending funds
6. State - These will determine whether our project/ startup is going to make a success or failure
7. Backers - A backer is an investor who either has a limited experience in startup investing or, even if he has, would rather let someone else manage the investments on his behalf or even choose the startups in which he must invest.
8. Country – what is the country code
9. usd pledged - in return for lending funds in usd currency
10. usd_pledged_real – actual in return for lending funds in usd currency
11. usd_goal_real - the final goal amount usd currency
12. dead_year – ended year
13. dead_hour - ended hour
14. dead_minute - ended minute
15. dead_day – ended day
16. dead_month - ended month
17. launch_hour - started hour
18. launch_minute - started minute
19. launch_day - started day
20. launch_month - started month
21. launch_year - started year
22. Duration – period of project/duration
23. name_char_count – number of characters in the project name
24. name_word_count - number of words in the project name
25. name_avg_word_len - number of words length in the project name.

5.2.3.Data Preparation:

Sort through data and get it ready for training. Clean up everything that could need it (get rid of duplicates, fix mistakes, handle missing numbers, normalise data, convert data types, etc.).Data can be made random to eliminate the impact of the specific order in which it was gathered and/or prepared.Use data visualisation to carry out additional exploratory research or to identify pertinent correlations between variables or class imbalances (bias alert!). Divided into sets for evaluation and training.

5.2.4 Model Selection:

We integrated the GridSearch CV algorithm with Decision Tree after achieving a 99.86% accuracy on the train set.

5.2.5 Analyze and Prediction:

We selected just two features from the actual dataset:
1 Overview - comprehensive principles of the project/startup/entrepreneurship
2 State: This shows the likelihood of success or failure for the project, startup, or entrepreneurial endeavour.

5.2.6 Accuracy on test set:

We got an accuracy of 97.28% on test set.

5.2.7 Saving the Trained Model:

The first thing to do is store your trained and tested model into a.pkl format file using a library like Pickle, after you're comfortable enough to take it into a production-ready environment.Verify that Pickle is installed in your setting. Let's import the module and then upload the model to the.pkl file next.

6. CLASSIFICATION ALGORITHMS

6.1 PROPOSED ALGORITHM

DECISION TREE (DT) WITH GRID SEARCH CV, LOGISTIC REGRESSION (LR) GRADIENT BOOSTING WITH GRID SEARCH CV

Because decision trees can manage complicated and non-linear interactions between the predictor and responder variables, their use in predicting the entrepreneurial success of crowd fundraising campaigns has received attention. With the help of grid search cross-validation (CV), which is the focus of this proposed content, we want to improve the accuracy with which crowd financing campaign success is predicted. Data preprocessing will be a part of the suggested methodology, where we will deal with any missing or inconsistent data and eliminate any redundant or irrelevant features. Next, we will divide the dataset into training and testing sets. The testing set will be used to assess the model's performance, while the training set will be used to construct and refine the decision tree model using grid search CV. The grid search CV will look over the hyper parameters in a methodical manner to find the best combination that will produce the best accuracy.

6.2 EXISTING ALGORITHM

SUPPORT VECTOR MACHINE (SVM)

For the benefit of the publications included in this study, the notion of machine learning algorithms is somewhat expanded rather than strictly defined. To increase the breadth of work, machine learning algorithms encompass not only commonly used algorithms but also a method, strategy, or process that may use an algorithm in the background to assess any type of social media data.

7. REQUIREMENTS OF PROJECT

- Operating System : Windows 7/8/10
- Platform : Spyder3
- Programming Language : Python
- Front End : HTML, CSS

8. SYSTEM ARCHITECTURE

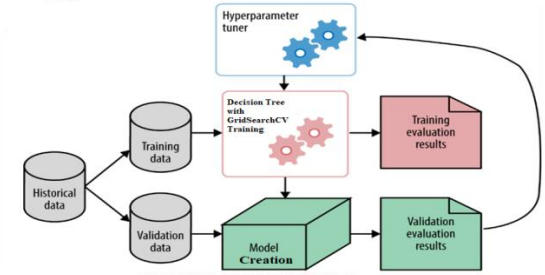
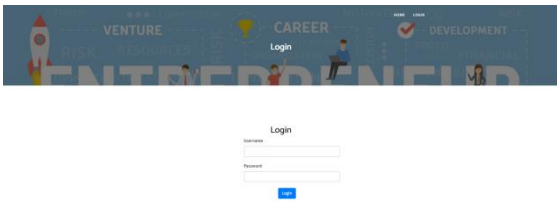


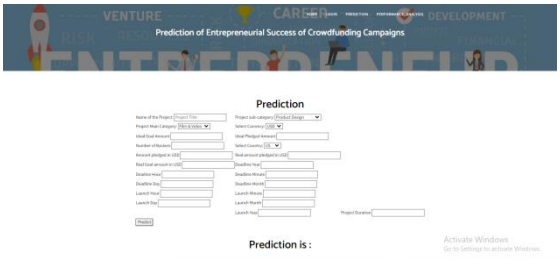
FIG 8.ARCHITECTURE OF MODEL

9. RESULT

This project implements like application using python and the server process is maintained using the SOCKET & SERVERSOCKET and the design part is played by cascading style sheet.
These are some of the snapshots

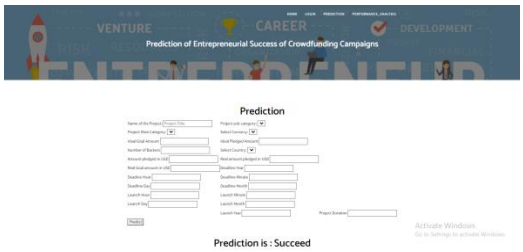


9.1 Login Page

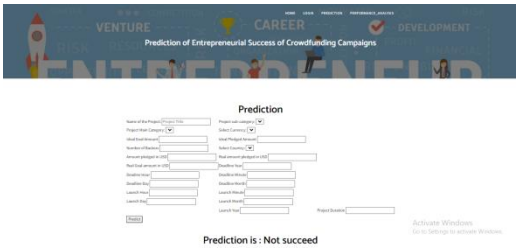


9.2 Landing Page

9.3 A test case for prediction by entering campaign details



9.4 Result if a campaign is successful



9.5 Result if a campaign is unsuccessfull

10.FUTURE ENHANCEMENTS

In order to encourage generalization, the Bayesian nonparametric approach will therefore be a more tenable strategy and deserving of future consideration. Because they enable the use of an infinite number of parameters to represent the properties of the distribution underlying the complicated data, the Bayesian nonparametric models are more reliable and applicable to a wider range of situations. Furthermore, if the identification and comprehension of the impact of specific variables taken into consideration on the success rate is of interest.

11.CONCLUSION

The primary goal of this study was to examine statistical learning techniques and related feature-engineering-based machine learning algorithms that provide the most accurate predictive models for predicting Kick starter campaign success. Kick Starter, one of the largest reward-based crowd financing platforms globally, provided the data that was used through site scraping. 61 characteristics and more than 80,000 observations were used. The focus of the study was on American-originated Kick starter initiatives, as around 96% of them were conducted in this country. To start, the most important variables were targeted via feature engineering. Goal, supporters count, time, and population were determined to be the most important and productive variables in reducing the prediction error of all the machine learning techniques we intended to employ after dimensionality reduction using GridSearchCV, the Decision Tree procedure, and multicollinearity diagnostics. We then looked at three machine learning algorithms. Three cross-validation techniques were used to assess the validity of these methods' performances, and classification metrics

including test error rates, accuracy, F1 scores, precision, and recall were monitored..According to the results, the Decision Tree with GridSearchCV methods for classification had the lowest test error rates and overall very high rates of recall, precision, and accuracy (99.86%), which suggests that these methods are superior for classifying and predicting Kick starter campaign success rates. Thus, the main research question has been resolved. It is crucial to remember that the presumptions made by some classification techniques, such the parametric LR analysis, have been demonstrated to be unrealistic and unadoptable when dealing with extremely complicated data. This is due to the fact that it starts with the premise that the sample data originates from a population with a fixed number of parameters and an identical probability distribution. For complicated datasets, the second assumption—that of the independence of observations—is not always tenable.

12. REFERENCES

[1] P. Belleflamme, T. Lambert, and A. Schwienbacher, Crowd funding: Tapping the right crowd, J. Bus.Venturing, vol.29, no. 5, pp. 585–609, 2014.

[2] V. Kuppuswamy and B. L. Bayus, Crowd funding creatynamics oive ideas: The df project backers, in The Economics of Crowd funding, D. Cumming and L. Hornuf, Eds. Cham, Germany: Springer, 2018, pp. 151–182.

[3] E. M. Gerber and J. Hui, Crowd funding: Motivations and deterrents for participation, ACM Trans. Compute.-Human Interact., vol. 20, no. 6, p. 34, 2013.

[4] J. S. Hui, M. D. Greenberg, and E. M. Gerber, Unders tanding the role of community in crowd funding work, in Proc. 17th ACM Conf. on Computer Supported Cooperative Work & Social Computing, Baltimore, MD, USA, 2014, pp. 62–74.

[5] E. Mollick, The dynamics of crowd funding: An exploratory study, J. Bus. Ventur., vol. 29, no. 1, pp. 1–16, 2014.

[6] M.J. Zhou, B. Z. Lu, W. P. Fan, and G. A. Wang, Project description and crowd funding success: An exploratory study, Inf. Syst. Front., vol. 20, no. 2, pp. 259–274, 2018.

[7] N.X. Wang, Q. X. Li, H. G. Liang, T. F. Ye, and S. L. Ge,Under standing the importance of interaction between creators and backers in crowd funding success, Electron. Commer. Res. Appl., vol. 27, pp. 106–117, 2018.

[8] K. Choy and D. Schlagwein, Crowd sourcing for a better world: On the relation between it affordances and donor motivations in charitable crowd funding, Inf. Technol. People, vol. 29, no.1,pp. 221–247, 2016.

[9] H. Yu, S. H. Jiang, and K. C. Land, Multicollinearity in hierarchical linear models, Soc. Sci. Res., vol. 53, pp. 118–136, 2015.

[10] S. L. Kukreja, J. Löfberg, and M. J. Brenner, A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification, IFAC Proc. Vol., vol. 39, no. 1, pp. 814–819, 2006. B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and [11] F. A. Hamprecht, A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data, BMC Bioinformatics, vol. 10, no. 1, p. 213, 2009.

[12] P. McCullagh and J. A. Nelder, Generalized Linear Models. 2nd ed. London, UK: Chapman & Hall/CRC, 1989.

[13] J. Franklin, The elements of statistical learning: Data mining, inference and prediction, Math. Intell., vol. 27, no. 2, pp. 83–85, 2005.

[14] M. Grandini, E. Bagli, and G. Visani, Metrics for multi-class classification: An overview, arXiv preprint arXiv: 2008.05756, 2020.

[15] N. Japkowicz and M. Shah, Evaluating learning algorithms: A classification perspective. Cambridge, UK: Cambridge University Press, 2011