Available Online: 3 April 2025 DOI: 10.54254/2977-3903/2025.21937

Research on the state prediction model of crowdfunding projects based on machine learning algorithms

Han Yuantianci

Shanghai University Information Management and Information Systems, Shanghai, China

yuantianci_han@163.com

Abstract. Crowdfunding is a concept that emerged due to difficulties in raising funds for community business projects, social activities, small and micro enterprises, and startups. However, the success or failure of crowdfunding projects is often full of uncertainty. Therefore, predicting whether a crowdfunding project can succeed has become a question worthy of in-depth research. This article takes the crowdfunding website platform Kickstarter as an example and uses machine learning methods to construct an effective crowdfunding project status prediction model. This study aims to compare the performance of different machine learning algorithms in predicting crowdfunding results, while identifying the main factors that affect crowdfunding results and their relative importance. The dataset used in this study includes data from all projects published on Kickstarter between January 2017 and January 2018. This study used six classic classification algorithms to predict the status of crowdfunding activities, and calculated the accuracy of each classification model. The results showed that the accuracy of all six models was close to 1, indicating that they could effectively predict the success or failure of crowdfunding projects. However, the Gaussian Naive Bayes model had slightly lower accuracy than the other five models. Furthermore, the research results indicate that in successful crowdfunding projects, factors such as crowdfunding goals, funds raised, and the number of supporters are more crucial influencing indicators than others.

Keywords: crowdfunding platform, classification prediction, logistic regression, decision tree

1. Introduction

The origin of crowdfunding is from a foreign word, which means mass fundraising or crowd financing, representing an emerging financing model in today’s business world [1]. With the rapid advancement of technology, the Internet, functioning as a bridge between capital donors and recipients, facilitates the swift spread of crowdfunding activities and thereby mobilizes substantial funds. Lin et al. put forward that there are four major crowdfunding models currently in existence: donation - based, reward - based, equity - based, and debt - based crowdfunding [2]. The crowdfunding platform selected in this study is Kickstarter, which mainly hosts projects with an artistic inclination. Project initiators are usually individuals or small teams, with funding amounts hovering around a few thousand dollars. Despite Kickstarter being a well - developed and rule - based online crowdfunding platform, the success rate of crowdfunding projects on it is still not high. This study aims to explore and answer two research questions: the performance of different machine learning models in predicting the success of projects on Kickstarter, and the identification of key factors affecting crowdfunding results and their relative importance. Kickstarter adopts the reward - based crowdfunding model. If the crowdfunding is successful, the initiator will offer non-monetary and non-equity rewards to backers based on their contribution level. If it fails, the donated money won’t be refunded to backers [3]. Therefore, having an accurate prediction model for crowdfunding projects to assess their likelihood of success on Kickstarter and identify important factors related to success is of great significance.

In this paper, multiple research methods are applied. First, the literature induction method is used to sum up commonly-used models and indicators for crowdfunding project prediction. Then, mathematical statistics is employed to analyze the dataset with Python, so as to grasp data characteristics and variable relationships. Finally, six prediction models are built based on different machine learning algorithms, with their accuracy calculated. Meanwhile, K-fold cross-validation is conducted to comprehensively evaluate the prediction effect of each model.

2. Literature review

2.1. Research on crowdfunding projects prediction

After retrieving and reading domestic literature on crowdfunding project prediction, this paper finds that Wei et al. [4] used Kickstarter data and applied the random neural network algorithm to predict crowdfunding project success, achieving high accuracy. Xu [5] studied the online charity platform DonorsChoose.org, categorizing prediction models into single - classification and ensemble learning models. For single - classification models, he used decision - tree and logistic regression algorithms; for ensemble learning models, random forest, GBDT, XGBoost, and LightGBM were applied, with confusion matrices for evaluation. It can be seen that machine learning and ensemble learning models are widely used in crowdfunding project prediction with good effects, providing ideas and basis for this study's model construction.

2.2. Research on crowdfunding project influencing factors

After retrieving and reading domestic literature on crowdfunding project influencing factors, this paper finds that Yang [6] constructed variables from project characteristics and entrepreneurial team characteristics, discovering that text and image information extracted from variables, besides numerical information, can also convey effective information. Shu et al. [7] divided variables into project characteristics, geographical location characteristics, and project delivery time for research. It can be observed that current research on crowdfunding project influencing factors is limited, with most lacking quantitative analysis of these factors. To fill this gap, this study will determine the weight of each influencing factor through quantitative analysis and create radar charts to visually display the weight distribution of these factors.

3. Methodology

3.1. Data source

The dataset utilized in this study was obtained from Kaggle. After meticulous organization and screening, it encompassed crowdfunding project data launched on Kickstarter from January 2017 to January 2018. The original dataset included all variables shown in Table 1 below:

Table 1. Variable

Variable Name	Data Type	Description
ID	int64	Project number
Name	object	Project name
Category	object	Project category
Subcategory	object	Project sub - category
Country	object	Project origin
Launched	object	Project launch time
Deadline	object	Project deadline
Goal	int64	Crowdfunding target amount
Pledged	int64	Funds raised
Backers	int64	Number of supporters
State	object	Project crowdfunding status

3.2. Data preprocessing

In this study, the screening for missing values was initially carried out. Upon examination, it was determined that all columns had zero missing values, rendering further processing unnecessary. Then, it handled outliers using the quartile method. Specifically, it imported the mstats module from the SciPy library to determine the upper and lower bounds of normal data, and proceeded to delete data points exceeding these bounds.

To ensure the effective processing of the dataset by machine learning algorithms, non - numerical data in the dataset was converted to numerical data. For this purpose, the factorize function was utilized for label encoding, assigning an integer code starting from 0 to each unique label in every column of the dataset. Moreover, given the study's focus on predicting crowdfunding project success, data not in a successful or failed state was removed.

To gain a deeper insight into the influencing factors of crowdfunding projects, this research extracted key temporal features from the raw dataset, including the project launch time, launch month, and duration from launch to deadline.

3.3. Data statistical analysis

3.3.1. Descriptive statistics

This section uses .describe to output the descriptive statistics of preprocessed feature variables, including mean, quantiles, maxima, minima, etc., for each numerical variable. The results indicated that the crowdfunding target amount and funds raised have large standard deviations, indicating wide variation. In contrast, other features like launch time and duration have smaller standard deviations, suggesting more concentration. See Table 2 below for details.

Table 2. Descriptive statistics of feature variables

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
Category	31733	6.39	3.94	0	3.00	7.00	10.00	14.00
Country	31733	3.02	4.14	0	1.00	1.00	4.00	21.00
Goal	31733	6933.19	8548.81	1	1057.00	3500.00	10000.00	40747.00
Pledged	31733	1111.03	1839.74	0	23.00	265.00	1360.00	10585.00
Backers	31733	16.45	20.69	0	1.00	6.00	25.00	83.00
Launch Month	31733	6.35	3.42	1	3.00	6.00	9.00	12.00
Launch Hour	31733	12.98	7.45	0	6.00	15.00	19.00	23.00
Days Duration	31733	31.18	12.14	0	29.00	29.00	33.00	60.00

3.3.2. Correlation analysis

To initially explore the relationships between crowdfunding status and each variable, this section visualizes correlations through a heatmap of the correlation matrix, as shown in Figure 1. The results reveal a weak negative correlation between crowdfunding status and the target amount, a moderate positive correlation with funds raised, and a strong positive correlation with the number of supporters.

Additionally, the correlation coefficient between the number of supporters and funds raised is 0.74, suggesting potential multicollinearity. Therefore, this section calculates the Variance Inflation Factor (VIF) for each column of variables. The results show that the VIF values for all variables are less than 10, indicating no severe multicollinearity in the dataset and allowing for subsequent analysis.

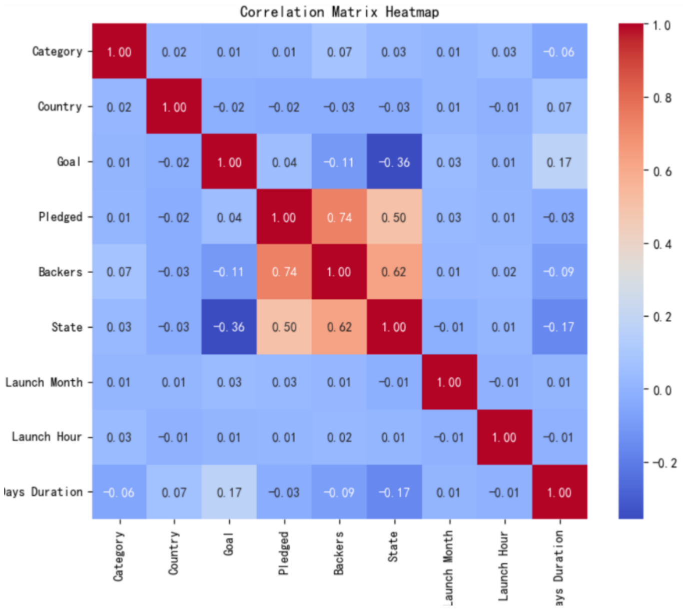

Figure 1. Heat Map4 experiment

4. Experiment

4.1. Construction of classification models

After preprocessing and statistical analysis, the total data volume is 31,733, with a success - to - failure ratio of about 3:7. The data is split into training (70%, 22,213 instances) and testing sets (30%, 9,520 instances) with $random\_state = 42$. The accuracy of each of the six models from different machine learning algorithms is calculated.

4.1.1. Logistic regression

The Logistic Regression algorithm is implemented via the LogisticRegression method in Python's sklearn.linear_model library [8]. After training on the training set data, the training and test set accuracies of the logistic regression prediction model are 1 and 0.9997, respectively.

4.1.2. Decision tree

The Decision Tree algorithm is implemented via the DecisionTreeClassifier method in Python's sklearn.tree library. After training on the training set data, the training and test set accuracies of the decision - tree prediction model are 1 and 0.9959, respectively.

4.1.3. K - nearest neighbors algorithm

The KNN algorithm is implemented via the KNeighborsClassifier method in Python's sklearn.neighbors library. After training on the training set data, the training and test set accuracies of the KNN prediction model are 0.9988 and 0.9987, respectively.

4.1.4. Gaussian naive bayes

The Gaussian Naive Bayes model is implemented via the GaussianNB method in Python's sklearn.naive_bayes library. After training on the training set data, the training and test set accuracies of the Gaussian Naive Bayes prediction model are 0.8873 and 0.891, respectively.

4.1.5. Adaptive boosting

Adaptive Boosting, an ensemble learning algorithm, is implemented via the AdaBoostClassifier method in Python's sklearn.ensemble library. After training on the training set data, the training and test set accuracies of the Adaptive Boosting prediction model are 0.9911 and 0.9910, respectively.

4.1.6. GBDT

GBDT, another ensemble learning algorithm, is implemented via the GradientBoostingClassifier method in Python's sklearn.ensemble library. After training on the training set data, the training and test set accuracies of the GBDT prediction model are 0.9954 and 0.9942, respectively.

4.2. Results

4.2.1. Accuracy of classification models

This section compares the prediction accuracies of the six models, and the results are presented in Figure 2 below. The figure indicates that the training and test set accuracies of all six models are very close to 1, signifying their good predictive performance for crowdfunding project success. Further observation reveals that the Gaussian Naive Bayes model has an accuracy of around 0.88, slightly lower than the other five models.

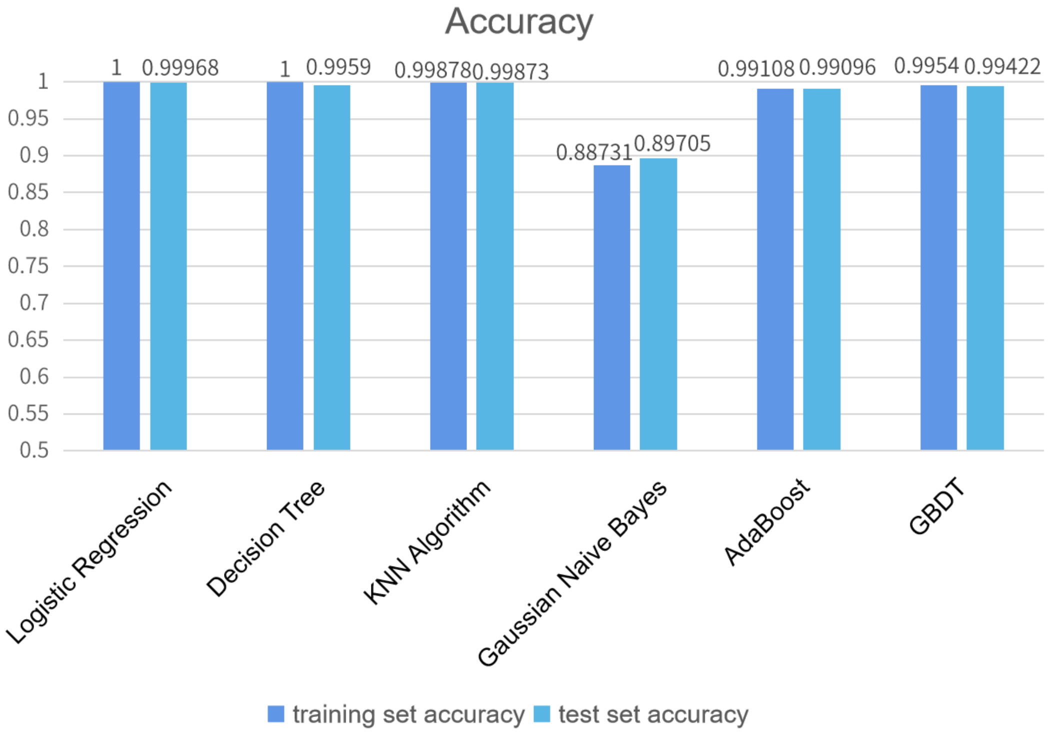

Figure 2. Accuracy comparison

4.2.2. Feature importance analysis

This section analyzes feature importance to identify key factors affecting crowdfunding outcomes. The Logistic Regression, Decision Tree, and GBDT models directly offer feature importance values. Therefore, we initially extract these values from each model, and the results are presented in Table 3 below.

Table 3. Feature importance

Feature	Logistic Regression Importance	Decision Tree Importance	GBDT Importance
Category	-0.143391	0.000430	0.000471
Country	-0.017760	0.000328	0.000013
Goal	-0.918842	0.352735	0.331958
Pledged	0.950653	0.267925	0.262234
Backers	0.090978	0.377267	0.405249
Launch Month	0.140375	0.000000	0.000000
Launch Hour	0.070039	0.000360	0.000000
Days Duration	0.014445	0.000955	0.000075

For easier comparison, the feature importance is normalized to the 0-1 interval. After normalization, the most important feature in each model has a weight of 1. The radar charts (Figure 3) for feature importance of the three models are shown below. The high-weight variables are concentrated in the upper-left corner, namely, number of supporters, funds raised, and crowdfunding target. This suggests these three factors are more crucial than others for a successful project launch.

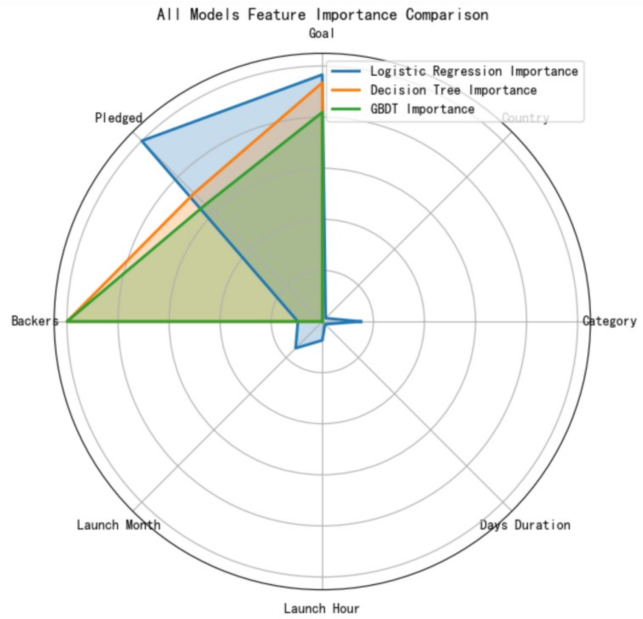

Figure 3. Radar chart

4.2.3. Robustness test

To ensure the models' prediction results are not due to specific data choices, model specifications, or random factors, this section conducts robustness tests on the six models. The scores from K-fold cross-validation are shown in the figure below. Figure 4 shows that the test scores are consistent with the previous experimental results, indicating the experimental conclusions are reliable, credible, and robust.

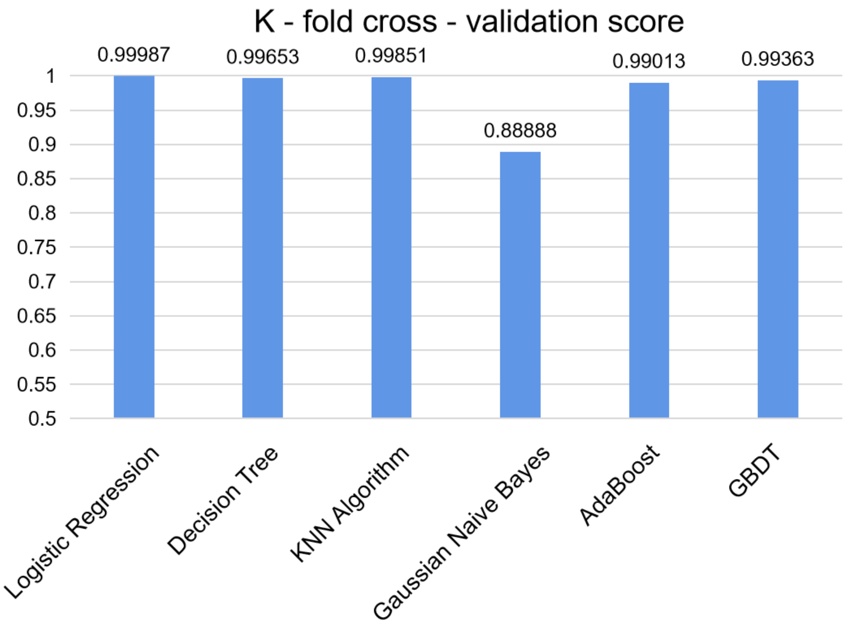

Figure 4. Cross-validation score

5. Conclusions

This paper draws the following significant conclusions. By comparing predictive models constructed using different machine - learning algorithms, it is discovered that the accuracies of the Logistic Regression, Decision Tree, K-Nearest Neighbor, Adaptive

Boosting, and GBDT models are approximately 0.99, whereas the accuracy of the Gaussian Naive Bayes model is slightly lower, at around 0.88. Overall, all six models effectively predict crowdfunding project success. This study uses K-fold cross-validation to test model robustness. The results demonstrate that the cross-validation scores are consistent with the previous experimental findings, validating the stability of the experimental conclusions. Also, feature importance analysis finds that crowdfunding target, funds raised, and number of supporters are key factors affecting project success, more important than other variables.

Due to time and data quality constraints, this research has the following limitations. First, it should have explored crowdfunding project statuses more deeply. This study focuses on success or failure, deleting other statuses during data preprocessing. However, projects can also be canceled or suspended. Future research should consider more project status scenarios. Second, multi - dimensional selection of Y. In this study, Y is categorized based on whether the crowdfunding target is met. However, the variable Y can also be defined in terms of product sales, consumer satisfaction, and reviews. Future research should explore which model has the best predictive performance for different Y selections and how feature importance changes when predicting different Y.

References

[1] Tang, Z., & Xue, X. (2024). Research on user experience evaluation of public welfare crowdfunding platform based on perceived affordance theory. *Journal of Library and Information Science, 9*(2), 34-41, 50.

[2] Lin, F., & Wen, X. (2023). Cultural and creative crowdfunding platform: types, potential and limitations. *Industrial Innovation Research,* (15), 117-120.

[3] Xu, L., & Yang, Z. (2023). Research on publishing integration ecology from the perspective of user innovation. *Publishing Wide Angle,* (12), 62-67.

[4] Wei, J., & Zhou, Z.M. (2024). Crowdfunding performance prediction model based on investor behavior analysis. *Application Research of Computers, 41*(8), 2448–2454.

[5] Xu, H. (2021). *Research on the prediction model of online public welfare crowdfunding platform projects based on machine learning algorithm*. [Unpublished doctoral dissertation]. Xidian University.

[6] Yang, J. (2023). Can artificial intelligence improve the investment efficiency of equity crowdfunding? — An empirical study based on the perspective of investor information perception. *Times Finance,* (12), 57–60.

[7] Shu, W., & Wang, M. (2021). Analysis of influencing factors of Internet crowdfunding results. *Finance and Economics,* (18), 42–43.

[8] Xue, B. (2022). *Research on influencing factors and prediction of online medical crowdfunding project financing results.* [Unpublished doctoral dissertation]. Jiangxi University of Finance and Economics.