A Data Science Approach for Predicting Crowdfunding Success

Ahmed Banimustafa, Sattam Almatarneh, Olla Bulkrock, Ghassan Samara and Mohammad Aljaidi

December 7, 2022

A Data Science Approach for Predicting Crowdfunding Success

Ahmed BaniMustafa, IEEE Senior Member
Data Science and Artificial Intelligence Department
Isra University, Amman, Jordan
a.banimustafa@iu.edu.jo

Sattam Almatarneh
Department of Data Science and Artificial Intelligence
Zerqa University, Zerqa, Jordan
salmatarneh@zu.edu.jo

Olla Bulkrock
Data Science Department,
Princess Sumaya University for Technology, Amman, Jordan
oll20228067@std.psut.edu.jo

Ghassan Samara
Department of Computer Science,
Zerqa University, Zerqa, Jordan
gsamara@zu.edu.jo

Mohammad Aljaidi
Department of Computer Science,
Zerqa University, Zerqa, Jordan
mjaidi@zu.edu.jo

***Abstract*— Crowdfunding is important for backing innovative projects and new startup businesses. However, success in achieving the target fundraising is a big challenge, and it depends on many complex factors. This work uses data science to predict the success of crowdfunding pledges using a historical dataset that was scrapped from the Kickstarter website. The dataset was subject to intensive data wrangling, data exploration, data engineering, and procedures. Four machine learning models were constructed in this study using four algorithms: (1) Random Forests (RF), (3) K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). The models were trained using a separate portion that makes up two-thirds of the dataset, while the remaining third was used to evaluate the models. The KNN model achieved the best performance with a classification accuracy of 97.9% and an AUC of 98.3%. The Random Forests model achieved the second-best performance with a classification accuracy of 94.9% and an AUC of 98.9%. The Precision, Recall, F1, and AUC metrics also confirmed the validity of the reported results, while the confusion matrix and the ROC curve confirmed the robustness of the constructed models.**

Keywords—Data Science, Data Mining, Machine Learning, Crowdfunding, Fundraising, Kickstarter.

I. INTRODUCTION

Crowdfunding plays a significant role in enabling ordinary people to realize their innovative ideas and supporting startup businesses. Kickstarter is ″a global crowdfunding platform that supports creative and innovative projects″. However, achieving success in a crowdfunding campaign is a complex yet risky endeavor [1, 2].

Data science can provide the tools and technologies required for forecasting and gaining insight into the success and failure of crowdfunding campaigns. These are based on performing a wide array of procedures covering web scrapping [3], data wrangling [4, 5], big data handling, feature engineering [6, 7], classical data exploration, data preparation [8], machine learning [9, 10], and model evaluation procedures [11].

This work aims at predicting the success of crowdfunding campaigns using a dataset that was scrapped from the famous Kickstarter website. The website started in 2009 and raised $4.6 billion for more than 500,000 projects, which 17 million backers funded [12, 13]. The analysis of this dataset involves intensive data processing, features engineering, and data exploration procedures which are used for preparing the dataset for modeling using several machine learning algorithms. Four machine learning algorithms were applied in this study to create four models: (1) Random Forests (RF), (2) K-Nearest Neighbor (KNN), and (3) Support Vector Machine (SVM). The models were trained using a separate portion representing 66% of the dataset. The remaining 33% were then used to evaluate the constructed model using Classification Accuracy (CA), Precision, Recall, F1, Area Under the Curve (AUC) metrics, Confusion Matrix, and ROC curve.

Section II reviews the related work, while Section III describes the dataset. Section IV describes the research methodology applied in the study, and Section V presents the results. Section VI discusses the obtained results and then draws a conclusion that also comments on future work.

II. RELATED WORK

Several related works have been reviewed in this study. These works are summarized in TABLE I.

TABLE I. A SUMMARY OF THE RELATED WORK

Work	Aims & Technique Applied	Dataset	Results Obtained
[1]	Predicting success using SVM	13,000 projects	68% CA
[14]	Predicting success using Natural language processing (NLP) based prediction	Corpus of 45,000 projects. 59 other variables	58.56% Prediction power
[15]	Predicting success using Deep Learning	378,611 projects	93.2% CA and 93.2% AUC
[16]	Predicting success using Naïve Bayes, Random Forests, and AdaboostM1	Scraped dataset of 151,608 projects. 49 features	70.7% CA, 76.3% AUC using Naïve Bayes, 83.1% CA, 90.4% using Random Forests, 84.2% CA & 91.0% AUC using AdaboostM1
[17]	Predicting success using KNN with Whale Optimization Algorithm (WOA)	21,000 projects with 24 features	64% Accuracy, F-score of 68.5%, 66.18% recall, and 71.0% precision
[18]	Predicting success using: Random Forests, CatBoost, XGBoost & AdaBoost	130,00 Projects	74%-84% CA
[19]	Predicting success with optimal weighed Random Forests	dataset 367,763 projects	94.29% CA

The analysis of the seven related works shows that most of the reported studies depended on using tabular data scraped from the Kickstarter website, except a study conducted by [14] which depended on using an NLP corpus combined with 59 other features. The size of data varies from one study to another. Most results are reported using datasets of tens of thousands of projects. The Random Forest was the most successful algorithm as it achieved a CA score of 94%, as reported in [19]; Deep Learning also achieved good results, which a CA score of 93%. However, the performance of the other reported techniques was less, as it ranged between 68% and 84% CA. The NLP-Based approach achieved the worst performance with a score of 58% only.

III. DATASET

The dataset used in this work was originally scrapped from the Kickstarter website, one of the most popular online crowdfunding websites. The original dataset comprises 300,00 records for projects described using 13 attributes. Kickstarter dataset can be scraped online using two web services: Web Robots [12] and APIFY web [3]. The scraped data is publicly available [13]. The dataset features are described in TABLE II.

TABLE II. DESCRIPTION OF DATA ATTRIBUTES

Attribute	Description	Datatype
ID	Project identification number	Nominal
Name	project name	Nominal
Main Category	The main-project type is music, food, games, design, fashion, theater, DIY, etc. (categorical values)	Categorical
Category	The sub-project type food games, design, fashion, theater, DIY, etc. (categorical values)	Categorical
Currency	Currency of the fund: USD, CAD, AUD, Euro, GBP, etc.	Categorical
Launched	Pledge start date and time.	
Deadline	Pledge end date and time.	Datetime
Goal	The targeted amount of money funded in local currency	Continuous
Pledged	Amount of money raised by the crowd for the project in local currency	Continuous
State	The current state of the project: successful, failed, canceled, suspended, live	Categorical
Country	The country of the project owner: SA, GB, AU, CA, etc.	categorical
US Pledged	Amount of money raised for the project in US dollars.	Continuous
US_goal	The targeted amount of money funded in US dollars	Continuous
Backers	The number of people who supported the projects	Continuous

IV. METHODOLOGY

The methodology applied in this study consists of seven data science phases that correspond to the phases typically found in typical data mining process models such as CRISP-DM[20] and MeKDDaM [21-23]. These cover: (A) Data Scrapping, (B) Data Wrangling; (C) Data Exploration; (D) Data Engineering; (E) Model Construction; (F) Model Evaluation; and (G) Variable Importance Ranking. Figure 1 illustrates the phases of the applied research methodology.

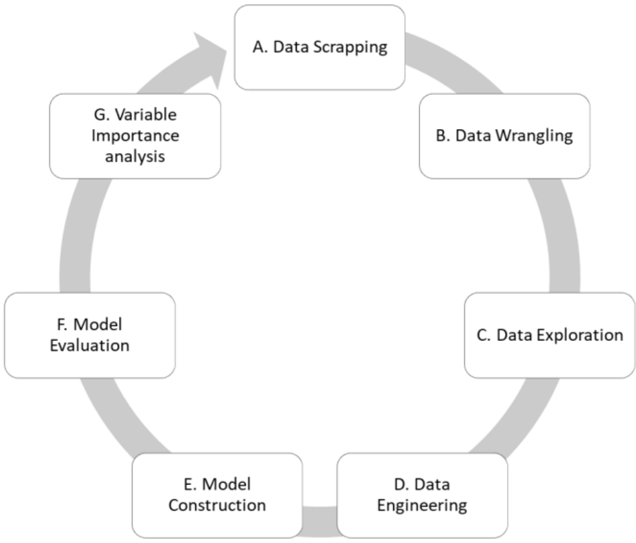

Figure 1. Applied Data Science Process

A. *Data Scrapping*

The data scrapping phase involves obtaining the dataset by scraping data from the Kickstarter website through a number of web scrapping, data capturing, and data parsing tools, APIs, and utilities [3, 12].

B. *Data Wrangling*

The data wrangling or munging involves transforming the raw data from its original web-based HTML format into a CSV tabular format to facilitate further data exploration, engineering, and analysis procedures [24, 25].

C. *Data Exploration*

Data exploration involves conducting descriptive statistics and data visualization procedures to gain insight into the data, its quality and distribution, and its trends and potential through various data visualization [21]. Issues such as missing values, outliers, and imbalanced classes are uncovered in this phase[22, 26-28].

D. *Data Engineering*

This phase covers a wide spectrum of data engineering procedures that prepare the data for modeling. These involve data sampling and transformation in addition to feature construction, transformation, and deletion [6].

E. *Model Construction*

The model construction phase involves building four prediction models using four classification algorithms which include (1) Random Forests;,(2) K-Nearest Neighbour; and (4) Support Vector Machines (SVN).

- **Random Forests:** A supervised machine learning algorithm for regression, classification, and feature ranking. This algorithm creates multiple decision trees constructed through a recursive partitioning method that splits the feature space into several regions [29-31].
- **KNN:** A nonparametric supervised learning technique introduced in the early 1950s [32] used for regression and classification. The algorithm measures the distance between the sample and its

closest K-neighbors to assign the sample membership to the most relevant classes[33, 34].

- **SVM is one of the most robust supervised learning algorithms** for solving regression and classification problems [35]. This method performs its classification tasks by mapping samples into a hyperplane to maximize the distance between the classified categories.

F. Model Evaluation

Model evaluation involves measuring the performance of constructed machine learning models using metrics such as Classification Accuracy (CA), Precision, Recall, and F1 metrics [36, 37], which are described by equations 1, 2, 3, and 4.

$$\text{Classification Accuracy (CA)} = TP+TN/ N. \quad (1)$$

$$Precision= TP/(TP+FP) \quad (2)$$

$$Recall= TP/(TP+FN) \quad (3)$$

$$F1= 2TP/2TP+FP+FN \quad (4)$$

Where TP represents the number of samples classified as belonging to the assigned class, TN represents the number of samples classified as not belonging to the assigned class. N is the total number of samples.

In addition, other metrics are also used, such as Area Under the Curve (AUC), confusion matrix [38], and Calibration Curve [39].

G. Variable Importance Analysis

Variable Importance analysis is a supplementary phase that involves analyzing the data modeling results by gaining insight into the most useful features in constructing the machine learning predictive model [31, 40]. Variable importance analysis aims at identifying factors that may influence the success of the predictive analysis and is usually performed only for the most successful models.

V. RESULTS

A. Data Scrapping Results

The dataset was scrapped from the Kickstarter website in an HTML format which was then parsed and processed into a text format. The encoding of the resulting dataset was also considered and managed to address issues encountered during the web scraping step.

B. Data Wrangling Results

The data wrangling phase involved transforming the textual data into a format that can be stored and handled in a tabular format using a CSV file. This phase was necessary for preparing the data for the exploration and also for the data modeling phases.

C. Data Exploration Results

The data exploration results found that the data has a quite acceptable distribution over successful and failed projects, shown in Figure 2. While Figure 3. shows the distribution of projects by categories. It is shown that Music, Films & Videos, and publishing are the most dominant and successful projects, while crafts, dance, and journalism are the least in terms of project numbers and success.

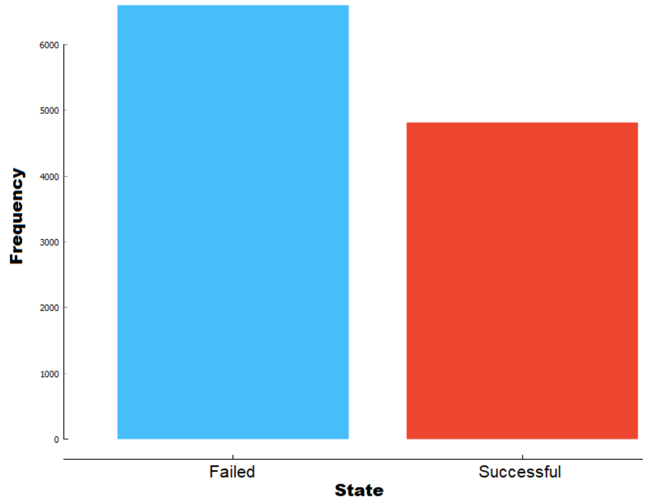

Figure 2. Distribution of dataset records over the two classes

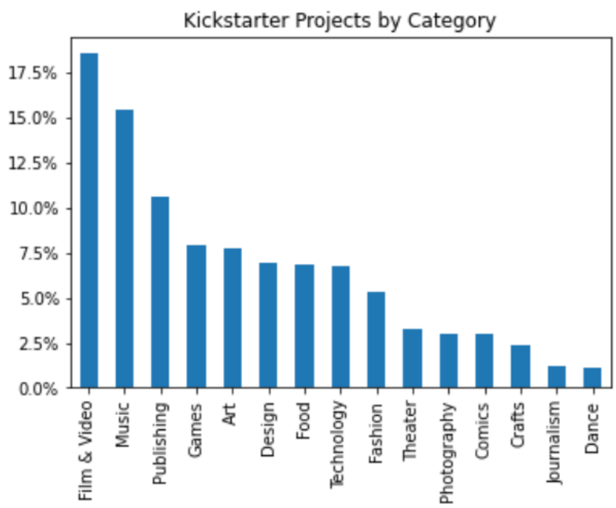

Figure 3. Distribution of projects by categories

The mosaic charts illustrated in Figure 4 uncover an interesting relationship between project categories, the pledged funding, and the project's success as the success likelihood increases in projects with more pledged money. It also shows that films, videos, and music represents the highest portion of projects and pledge for more money than other categories. Nevertheless, they enjoy a high likelihood of success. Theater, dance, and comics have the highest probability of success, representing only small portions of the pledged projects. Figure 4. explores the relationship between the goal and pledged amount of money and its influence on the project's success. The analysis results show that decreasing the goal increases the chances of the project's fundraising success.

D. Model Construction Results

Three models were constructed in this study: (1) A KNN model, A Random Forests model, and an SVM model. The models were trained using 66.6% of the dataset processed in steps A-C. On the other hand, 33.3% of the dataset was used for model evaluation.

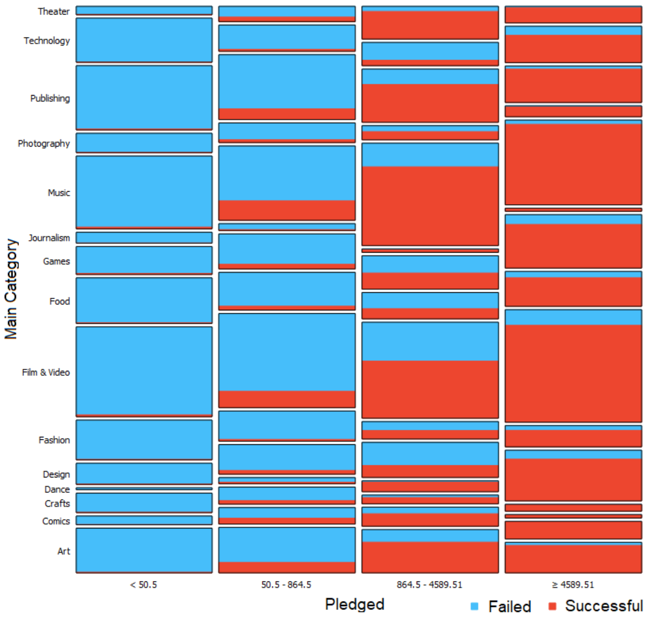

Figure 4. A mosaic chart that shows the relationship between the pledge and categories of both the successful and failed crowdfunding campaigns

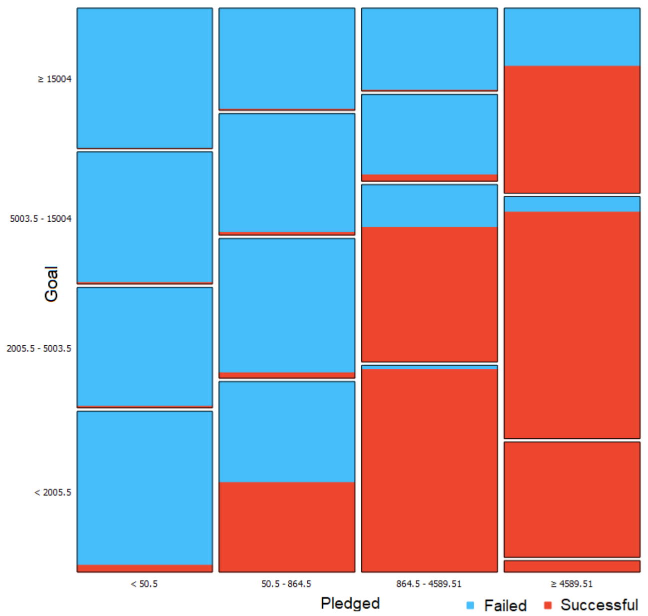

Figure 5. A mosaic chart that shows the relationship between the pledge and goal money for both the successful and failed funding campaigns

E. Model Evaluation Results

The three created models were evaluated using five performance metrics: Classification Accuracy (CA), recall, precision, F1, and Area Under the Curve (AUC). The KNN model performed best with a CA, precision, recall, and F1 score of 97.9% and an AUC score of 98.3%. The Random Forest model scored the second-best performance with 94.9% in CA, precision, recall, and f1, and 98.9% in the AUC metric. However, the SVM model failed to achieve satisfactory results, scoring between 50% and 60% in all applied performance metrics. TABLE III. Shows a comparison between the performance of the four constructed model models.

When comparing the performance of the KNN and Random Forests models constructed in this study to the models reported in the investigated work, we found that both models outperformed almost all the reported models. The confusion matrices of the KNN and Radom forests models confirm the validity of the models. TABLE IV. Shows the confusion matrix for the KNN model, while TABLE V. Shows the confusion matrix for the Random Forests model. On the other hand. The confusion matrix of the SVM model is shown in TABLE VI., which confirms its poor performance.

TABLE III. CLASSIFICATION MODELS PERFORMANCE

Model	CA	Precision	Recall	F1	AUC
KNN	97.9%	97.9%	97.9%	97.9%	98.3%
Random Forest	94.9%	95.0%	94.9%	95.0%	98.9%
SVM	52.4%	58.9%	52.4%	50.5%	50.1%

TABLE IV. SVM MODEL CONFUSION MATRIX

		Predicted		
		Success	Failure	Sum
Actual	Success	3797	7473	11270
	Failure	1773	6372	8145
	Sum	5570	13845	19415

TABLE V. KNN MODEL CONFUSION MATRIX

		Predicted		
		Success	Failure	Sum
Actual	Success	11212	58	11270
	Failure	350	7795	8145
	Sum	11562	7853	19415

TABLE VI. RANDOM FORESTS MODEL CONFUSION MATRIX

		Predicted		
		Success	Failure	Sum
Actual	Success	10686	584	11270
	Failure	397	7748	8145
	Sum	11083	8332	19415

The Calibration Curve also confirms the validity and robustness of the constructed models. The calibration curve for the created models is shown In Figure 6. The closest the curve to the logistic function curve is, the better. In comparison, the Random Forests and KNN both show excellent performance. In contrast, the performance of the SVM was poor.

F. Variable Importance Ranking Results

The Variable's importance was calculated for the two most successful models: KNN and Random Forests, which are illustrated in Figure 7 and Figure 8. While both models agree on ranking pledged as the most significant predictor, KNN ranks backers as the second most important predictor.

In contrast, Random Forest ranks the goal as the second most significant predictor. KNN ranks category, duration, and main category as important features by ranking them in fourth, fifth, and sixth place; on the other hand, Random Forests also ranks them in a different order and with fewer weights than the first three models.

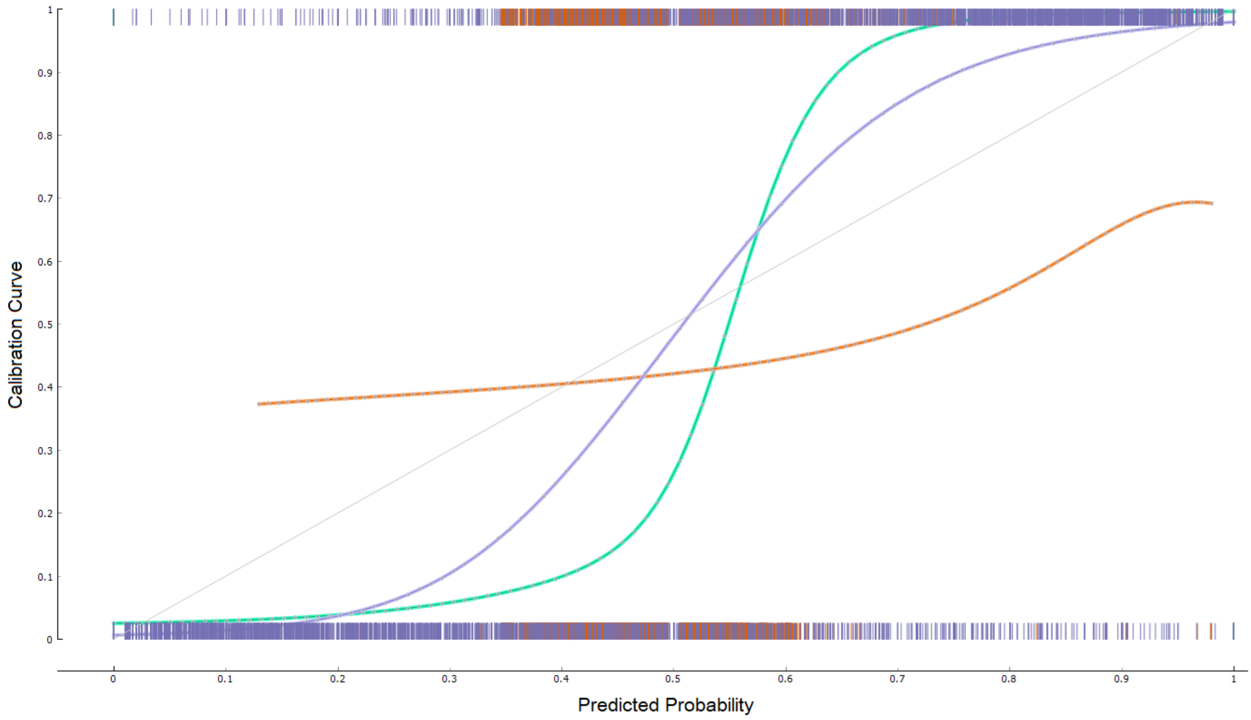

Figure 6. Calibration Curve shows the performance of the three constructed models.

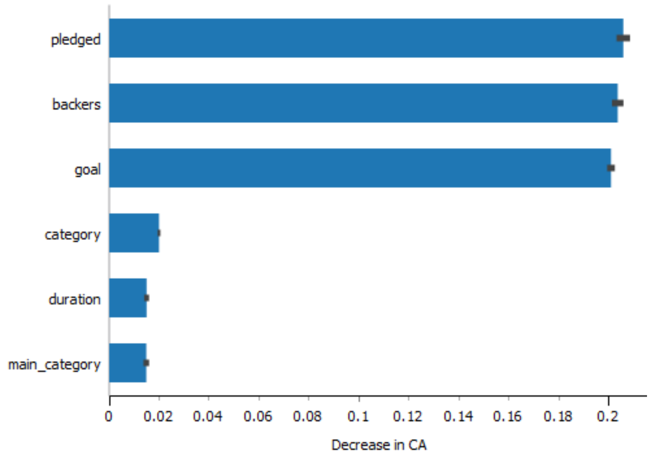

Figure 7. Variable importance ranking of the KNN model

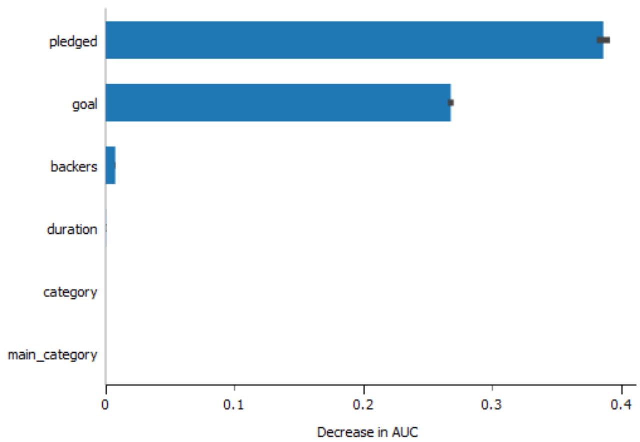

Figure 8. Variable Importance Ranking of the Random Forests Model

VI. Discussion & Conclusion

This work involved applying a data science approach for predicting the success of crowdfunding campaigns based on data sampled from a public dataset that consists of 300,000 projects. The results of this work achieved this study's aims as they successfully predicted the success of the crowdfunding campaign with excellent performance. Two of the three constructed models outperformed all the models reported in the related work of this study. The KNN model was the champion model, scoring a CA performance of 97.9% and an AUC performance of 98.3%. The Random Forest model was the second-best model, achieving a CA performance of 94.9% and an AUC performance of 98.9%. The Precision, Recall, F1, and AUC scores confirmed the validity of the two models, while the confusion matrix and calibration curves confirmed their robustness. However, the SVM model failed to score any acceptable performance.

In addition, the applied ranking algorithms were also successful in identifying the most important factors for achieving success in crowdfunding projects. The results show that the most decisive factors for success are the predetermined pledged and goal amount of money, the number of backers, and the campaign's duration. Compared to other studies reported in the related work, this work provides an additional contribution, which concerns identifying factors that influence success. The variables' importance ranking found that: pledged, goal, backers, category, and duration as the most important factors which contribute to crowdfunding success which can help project owners to influence the chance of success.

Future work that can extend this study might involve using regression and clustering techniques to predict the amount of money collected for each project. In addition, NLP can also be used to tune up the description of the project to attract more funds.

REFERENCES

[1] M. D. Greenberg, B. Pardo, K. Hariharan, and E. Gerber, "Crowdfunding support tools: predicting success & failure," in CHI'13 extended abstracts on human factors in computing systems, 2013, pp. 1815-1820.

[2] T. Tran, M. R. Dontham, J. Chung, and K. Lee, "How to succeed in crowdfunding: a long-term study in kickstarter," arXiv preprint arXiv:.06839, 2016.

[3] APIFY. (2022, 10/10/2022). APIFY Web Scraper Available: https://apify.com/misceres/kickstarter-search

[4] F. Endel and H. Piringer, "Data Wrangling: Making data useful again," IFAC-PapersOnLine, 2015.

[5] E. Mäkelä, K. Lagus, L. Lahti, and T. Säily, "Wrangling with non-standard data," Proceedings of the …, // 2020.

[6] J. Gray and P. Shenoy, "Rules of thumb in data engineering," in Proceedings of 16th International Conference on Data Engineering (Cat. No. 00CB37073), 2000, pp. 3-10: IEEE.

[7] A. Zheng and A. Casari, Feature engineering for machine learning: principles and techniques for data scientists. " O'Reilly Media, Inc.", 2018.

[8] M. G. Rossmann and C. G. Van Beek, "Data processing," Acta Crystallographica Section D: Biological Crystallography, vol. 55, no. 10, pp. 1631-1640, 1999.

[9] E. Alpaydin, Machine learning. MIT Press, 2021.

[10] A. BaniMustafa, "Predicting Software Effort Estimation Using Machine Learning Techniques," in 2018 8th International Conference on Computer Science and Information Technology (CSIT), Amman, 2018, pp. 249-256: IEEE.

[11] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Journal of Information Processing Management, vol. 45, no. 4, pp. 427-437, 2009.

[12] T. Vitulskis and P. Jonaitis. (2022, 16/10/2022). Kickstarter Datasets. Available: https://webrobots.io/kickstarter-datasets/

[13] (2022). Kickstarter Projects. Available:
https://www.kaggle.com/datasets/kemical/kickstarter-projects

[14] T. Mitra and E. Gilbert, "The language that gets people to give: Phrases that predict success on kickstarter," in Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, 2014, pp. 49-61.

[15] P.-F. Yu, F.-M. Huang, C. Yang, Y.-H. Liu, Z.-Y. Li, and C.-H. Tsai, "Prediction of crowdfunding project success with deep learning," in 2018 IEEE 15th international conference on e-business engineering (ICEBE), 2018, pp. 1-8: IEEE.

[16] S. Xiao, X. Tan, M. Dong, and J. Qi, "How to design your project in the online crowdfunding market? Evidence from Kickstarter," 2014.

[17] M. J. Ryoba, S. Qu, and Y. Zhou, "Feature subset selection for predicting the success of crowdfunding project campaigns," Electronic Markets, vol. 31, no. 3, pp. 671-684, 2021.

[18] S. Jhaveri, I. Khedkar, Y. Kantharia, and S. Jaswal, "Success prediction using random forest, catboost, xgboost and adaboost for kickstarter campaigns," in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 1170-1173: IEEE.

[19] F. S. Ahmad, D. Tyagi, and S. Kaur, "Predicting crowdfunding success with optimally weighted random forests," in 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS), 2017, pp. 770-775: IEEE.

[20] R. Wirth and J. Hipp, "CRISP-DM: Towards a standard process model for data mining," in Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 2000, vol. 1, pp. 29-39: Manchester.

[21] A. H. BaniMustafa and N. W. Hardy, "A strategy for selecting data mining techniques in metabolomics," in Plant Metabolomics, N. Hardy and R. Hall, Eds.: Springer, 2011, pp. 317-333.

[22] A. Banimustafa and N. Hardy, "A Scientific Knowledge Discovery and Data Mining Process Model for Metabolomics," IEEE Access, vol. 8, pp. 209964-210005, 2020.

[23] A. BaniMustafa, "A Knowledge Discovery and Data Mining Process Model for Metabolomics," PhD, Computer Science Dept., University of Wales, Aberystwyth Aberystwyth, 2012.

[24] F. Endel and H. J. I.-P. Piringer, "Data Wrangling: Making data useful again," vol. 48, no. 1, pp. 111-112, 2015.

[25] I. G. Terrizzano, P. M. Schwarz, M. Roth, and J. E. Colino, "Data Wrangling: The Challenging Yourney from the Wild to the Lake," in CIDR, 2015.

[26] A. BaniMustafa, "Enhancing learning from imbalanced classes via data preprocessing: A data-driven application in metabolomics data mining," ISeCure, vol. 11, no. 3, pp. 79-89, 2019.

[27] A. M. Bani Mustafa, "A knowledge discovery and data mining process model for metabolomics," Aberystwyth University, 2012.

[28] A. H. BaniMustafa and N. W. Hardy, "A Strategy for Selecting Data Mining Techniques in Metabolomics," in Plant Metabolomics: Methods and Protocols, N. W. Hardy and R. D. Hall, Eds. Totowa, NJ: Humana Press, 2012, pp. 317-333.

[29] L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.

[30] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in Ensemble machine learning: Springer, 2012, pp. 157-175.

[31] A. Liaw and M. Wiener, "Classification and Regression by randomForest," R News, vol. 2, no. 3, pp. 18-22, 2002.

[32] E. Fix and J. L. Hodges, "Discriminatory analysis. Nonparametric discrimination: Consistency properties," International Statistical Review/Revue Internationale de Statistique, vol. 57, no. 3, pp. 238-247, 1989.

[33] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN model-based approach in classification," in OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", 2003, pp. 986-996: Springer.

[34] W. J. Hwang and K. W. Wen, "Fast kNN classification algorithm based on partial distance search," Journal of Electronics letters, vol. 34, no. 21, pp. 2062-2063, 1998.

[35] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995/09/01 1995.

[36] C. X. Ling, J. Huang, and H. Zhang, "AUC: A Better Measure than Accuracy in Comparing Learning Algorithms," in Advances in Artificial Intelligence, Berlin, Heidelberg, 2003, pp. 329-341: Springer Berlin Heidelberg.

[37] J. Novakovic et al., "Evaluation of Classification Models in Machine Learning," vol. 7, pp. 39-46, 2017.

[38] R. Susmaga, "Confusion Matrix Visualization," in Intelligent Information Processing and Web Mining, Berlin, Heidelberg, 2004, pp. 107-116: Springer Berlin Heidelberg.

[39] M. Vuk and T. Curk, "ROC Curve, Lift Chart and Calibration Plot," Metodolo ški zvezki, vol. 3, no. 1, pp. 89-108, 2006.

[40] M. J. Zaki, W. Meira Jr, and W. Meira, Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press, 2014.