Predicting Investor Behavior and Investment Patterns in Equity and Lending Crowdfunding Campaigns

Rosa Porro*, Thomas Ercole, Giuseppe Pipitò, Gennaro Vessio and Corrado Loglisci

Department of Computer Science, University of Bari Aldo Moro, Bari, Italy

Abstract

Crowdfunding has become a popular way for entrepreneurs and companies to raise capital. However, increasing competition in the crowdfunding market has led to the need for predictive models to estimate the success or failure of crowdfunding campaigns. This study presents preliminary results on predicting investor behavior and investment patterns in equity and lending crowdfunding campaigns using a machine and deep learning approach. We collected a new dataset of crowdfunding campaigns in Italy and used several algorithms to model and predict investor behavior and investment patterns. The results obtained are encouraging. Our study can contribute to a better understanding of crowdfunding strategies and campaign design guidelines.

Keywords

Crowdfunding, Data science, Machine learning, Predictive modeling, Time series analysis

1. Introduction

Crowdfunding is a business model in which numerous individuals, often Internet users, invest small amounts of capital in a project. These small contributions are then pooled to support individual entrepreneurs, businesses, or nonprofit groups financially. This model operates without financial intermediaries, allowing ideas and products to be proposed directly through a platform. By harnessing "the power of the crowd", the crowdfunding model reflects collective wisdom, as the group of investors collectively determines the success of project funding. This approach has enabled many projects to reach their funding goals and thrive [1].

However, the increasing number of crowdfunding projects also leads to intense competition in the crowdfunding market because of the limited number of potential investors. Platform managers must determine whether an idea is worth designing and promoting as a campaign. Also, as different and alternative crowdfunding platforms develop, it is essential to consider the one that can encourage more investment. A very promising idea may be a failed campaign if it is hosted on a platform with investors disinterested in the issue or if it has to convey multiple points of innovation. These factors cause investors to be concerned about the results and thus eager to know the chances of successful funding before investing [2].

For this reason, technologies with predictive capabilities that can provide advanced information may be critical. These tools could estimate success or failure in advance based on predictive

SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy

*Corresponding author.

rosa.porro@uniba.it (R. Porro); gennaro.vessio@uniba.it (G. Vessio); corrado.loglisci@uniba.it (C. Loglisci)

models built once the stated duration of the campaign has expired. However, investment and investor interest may change due to the influence of campaign design and its presentation to investors. Therefore, studying the temporal variability of investor behavior and investment dynamics becomes critical to adapt campaign design over time [3].

Unfortunately, investor and investment patterns and trends are often complex, nonlinear, and characterized by wide diversity. Investor and investment data are inherently time series and, therefore, may exhibit fluctuations, sparsity, and irregularities; there may be numerous investments in a few days and no movement for a while. In addition, the fundraising phase may have different duration in different campaigns, and thus the lengths of these time series are always different. This requires data-driven algorithms that can account for the diversity of these time series and capture latent patterns while maintaining the ability to generalize [4]. It should not be overlooked that investor and investment dynamics represent the effect of a stimulus-effect mechanism attributed to the multimodal information content with which the campaigns were designed [5]. The data, such as textual description and metadata, are adapted by the creators according to the investment trend, but, unfortunately, their changes are not tracked by the platforms, as is the case with investors and investments, but kept as they were at the time of the opening of the fundraising phase, or as they are at the time of the closing. Time-varying and static data can be well handled by machine and deep learning solutions that allow us to investigate the task of investor and investment prediction by accommodating hidden relationships with multimodal information content and historical investor and investment data.

In light of this, this paper aims to explore and analyze the dynamics of investor behavior and investment patterns, particularly equity and lending crowdfunding campaigns in Italy, using a new dataset we collected. Equity crowdfunding focuses on financing companies, and investors become partners in the company, while in lending crowdfunding, investors lend money to fund a project and become creditors. Specifically, we propose to develop predictive models that can estimate the success or failure of crowdfunding campaigns by considering the temporal variability of investor behavior and investment dynamics, multimodal information content, and historical investor and investment data. We propose to use machine and deep learning algorithms to extract latent patterns and hidden relationships between campaign characteristics and investor behavior. The results of this study can contribute to the development of more effective crowdfunding strategies and guidelines for designing campaigns that attract more investment and increase the chances of funding.

The paper proceeds as follows. Section 2 discusses related work. Section 3 introduces the new equity and lending crowdfunding campaign dataset. Section 4 presents the methodology. Section 5 showcases the results and discussion. Section 6 concludes the paper and outlines future research directions.

2. Related work

Effective communication is the key to success in today's globalized world, particularly crowdfunding, where creators must effectively communicate their vision and build rapport with potential funders. Du et al. [6]'s research shed light on the influence of language on the success of crowdfunding campaigns, highlighting the importance of incorporating funders' preferred language into campaign strategies. Yang et al. [7] used natural language processing and feature selection algorithms to identify content factors in film crowdfunding campaigns. They found that specific words in the project name significantly impact project outcomes. The study by Song et al. [8] found that specific words used in project names significantly impact crowdfunding campaigns' success in the game industry. These findings provide valuable insights for creators seeking funding, enabling them to optimize communication strategies and increase the likelihood of success.

The articles by Wang et al. [9], Al-Qershi et al. [10], and Zhong et al. [11] all focus on using data analysis techniques to predict the success of crowdfunding campaigns. Each study focuses on a different type of data analysis, including video content, image, sentiment, and social media data analysis. In particular, Zhong et al. proposed the "project network" concept to predict crowdfunding campaign success by creating interconnections between projects and extracting network-based features. Also of interest is the study conducted by Shi et al. [12] on a recently developed deep learning framework based on audio analytics that can extract audio features to predict the fundraising outcomes of projects. The authors use transfer learning to train models in the proposed framework and multi-task learning to extract features. Overall, these studies demonstrate the potential of data analysis techniques in providing valuable information to improve crowdfunding campaign presentations and engage online communities, thereby increasing the likelihood of success. Although the results obtained from previous research are noteworthy, they offer only limited information and suggestions for selecting the best platform to ensure the success of a fundraising project. The long-term goal of our research is to predict the success of a project and recommend the best platform based on its characteristics.

In addition, it is worth noting that many countries have sought to address legal issues related to equity and lending crowdfunding to protect investors. While this may limit the growth of these types of crowdfunding, it may foster it by increasing the perceived safety of lenders. However, there is currently a lack of consistent regulation across countries, resulting in platforms focusing exclusively on campaigns within their nations. Due to regulatory and geographic limitations and lower investor participation compared to reward-based campaigns, the data associated with equity and lending crowdfunding are not as extensive. This study aims to expand previous research on non-reward-based platforms, particularly on Italian portals and simultaneously multiple websites. The research uses machine learning methods traditionally considered more effective than those previously employed in this field. In particular, the focus is on the time series associated with the phenomenon, which is often ignored due to the difficulty of finding data.

3. Materials

An online equity and lending crowdfunding project aggregator, Startups Wallet,[1] provided our dataset, acquired through web scraping. It is currently unpublished and its subsequent publication will be a milestone in our research, as there are no datasets involving multiple equity/lending crowdfunding portals/platforms like ours in the literature. The data were

[1] https://startupswallet.com/

collected without quality constraints, with a partial manual step of imputing missing data and checking for errors. The following main challenges characterize the dataset:

- The collection integrates data from nearly 40 Italian platforms, each with its formatting and some of which with missing information. The platforms considered represent a significant sample of all Italian platforms; the fact that the data refer only to 40 platforms out of about 80 available is irrelevant because, as pointed out in [13], many platforms have launched only a few campaigns.
- Some data related to the evolution of the campaigns are unavailable in the days following the update, as they are hidden and therefore not easily retrievable in case of errors (a fairly common situation);
- The observed failure rates are 21% for equity projects and 3% for lending projects. As a result, the classification problem is unbalanced;
- Some campaigns exhibit qualities of one type, but the platform categorizes them as the opposite type. Therefore they are considered in both cases.

In this work, we considered the following features:

- *Type* (lending: 1084 projects; equity: 907 projects) and *sector* (real estate: 781 projects; company: 1210 projects of crowdfunding). Some cumulative information is shown in Fig. 1.
- The campaign *name* and *region*;
- The crowdfunding *platform/website*;
- *Pre-commitment of money* and *pre-commitment of investors*;
- Crowdfunding campaign performance with a double time series, i.e. *money* and *investors*;
- The *minimum* and *maximum amount* to be raised;
- *Start date* and *end date* of the crowdfunding campaign;
- Multiple economic *categories*;
- *Minimum subscription* that can be invested by a funder in that project;
- *Holding time*, duration of the investment;
- *Annual ROI*, the annual return on investment;
- *Pre-money evaluation* of the company.

Given that the two phenomena analyzed, equity and lending, have only partially overlapping semantics and that *annual ROI* and *holding time* are relevant only for lending campaigns and *pre-money evaluation* only for equity campaigns, we decided to divide the dataset into two distinct groups and then conduct the analysis separately for the two groups.

As mentioned above, the time series reflecting the dynamics of the crowdfunding campaigns we are interested in are two for each crowdfunding observation/campaign: *money* and *investors*. They reflect the investor behavior and investment patterns we want to study. These time series can have distinct start and end dates. They also vary in length. Finally, they may have different frequencies, although they are never less than daily due to the collection methods. It is worth noting that these data are not necessarily monotonic because of the presence of platforms that allow withdrawals and investor activity before closing.

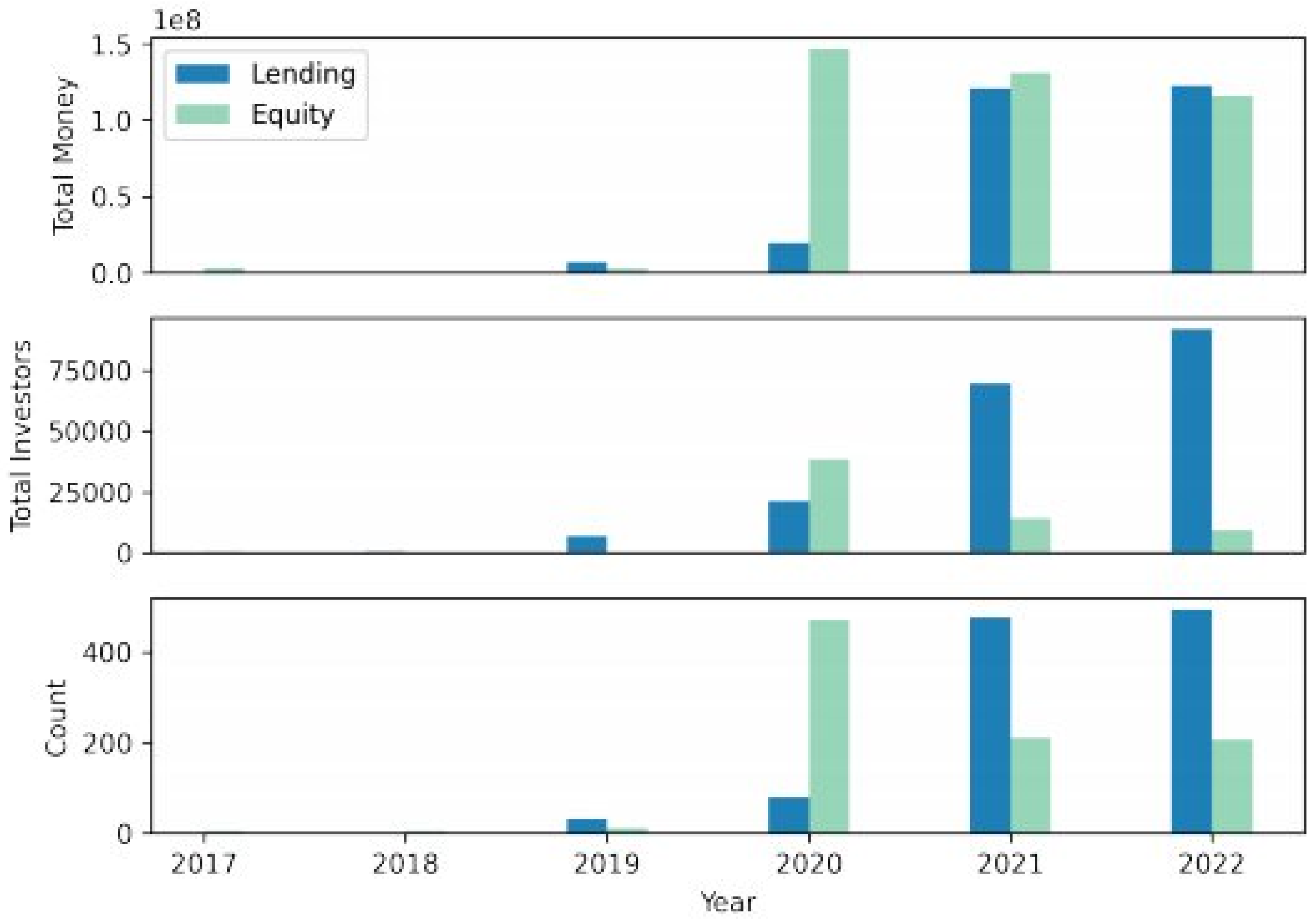

Figure 1: Cumulative amount of money, investors, and projects by year and typology.

4. Methods

In this section, we present the applied methodologies. As this was a new dataset, we describe the abundant pre-processing we applied to have clean data suitable for subsequent classification with learning algorithms in the first subsection. Time series classification is described in detail in the following subsection.

4.1. Preprocessing

Unfortunately, the dataset had many missing values, especially for the two time series. Denoted with (x_i, y_i) the pairs of values of the time series $\mathbf{x} = [x_1, \ldots, x_t, \ldots, x_n]$ and $\mathbf{y} = [y_1, \ldots, y_t, \ldots, y_n]$ associated with the i-th observation, the most frequent of these (non-mutually exclusive) cases are:

1. Missing values for both series at the same date, e.g. x_t and y_t;
2. Missing initial values for one or both series, e.g. x_0 and y_0 (or just y_0 or x_0);
3. Other miscellaneous missing value problems;
4. Complete absence of values for either series, e.g. $x_t \; \forall \; t = 1, \ldots, n$ (or $y_t \; \forall \; t = 1, \ldots, n$).

The first problem was solved by eliminating the pair (x_i, y_i) since there is no information. Nevertheless, we checked the following constraints if both first values were missing. If the time series length is 1, we imputed the values of *investors* and *money* before commitment (which are imputed to 0 if necessary). Otherwise, if the values of the pre-committed *investors* and *money* were equal to the second element of the double time series, we deleted the first pair; otherwise, we imputed them as in the case of length 1. This solves problem (2).

As for the other issues related to problem (3), since there is a high correlation between the time series (see Table 1), we employed a dual linear regression approach to estimate the missing values of one variable by exploiting the predictions of the other when possible (i.e., when at least two valid pairs of (x_i, y_i) values exist).

Table 1
Overall correlation between *money/investors* on *typology* and *sector*.

Typology and sector	Correlation
Company lending	0.427
Real estate lending	0.704
Company equity	0.462
Real estate equity	0.806

Finally, to solve problem (4), we eliminated all instances with consistently missing values of *money* due to their small number, while instances with all missing values for *investors* were retained and marked. For those observations that do not have problems related to this issue, we reported the previous values of individual instances (which exist due to the imputation of the first values made earlier). This procedure solves all the problems regarding missing values in the time series. After imputing missing values for the time series, the distributions of correlations remain almost identical.

In addition, to meet the requirement of constant length for time series data in machine learning algorithms, we selected two unique values of 10 (as the median of the scraping campaign duration) and 14 (as biweekly intervals) as fixed lengths. We interpolated the daily series and adjusted their lengths to 10 or 14. In cases where the time series was shorter than the target value, we increased the frequency by adding more points to exceed the target value, carefully preserving the previous data through appropriate frequencies. We performed uniform sampling for time series longer than the target value, including those resulting from the previous case. We maintained the starting and ending points to bring the length to the desired value. The goal was to ensure that the correct time series retained as much helpful information as possible while meeting the desired length requirements. Having obtained the *money* and *investors* series of fixed length, we calculated the *returns* for both series and lengths using the formulas:

$$r_j^x = \frac{x_j - x_{j-1}}{x_{j-1}} * 100, \quad r_j^y = \frac{y_j - y_{j-1}}{y_{j-1}} * 100.$$

In cases where platforms or campaigns did not specify a *maximum amount* to be achieved (but only a *minimum amount* was provided), we used a *k*-NN imputation method based on the *minimum amount* ($k = 5$). This approach also considers the *type* and *sector* of the campaign, allowing for accurate imputation of missing values. We applied a similar method to impute the *pre-money evaluation* in the case of equity campaigns. Comparing distributions before and after imputation, no significant differences were observed.

Finally, we calculated the *campaign duration* from the difference between the end date and the start date and extracted the year, month, quarter, and day of the week for both. Since the latter three are cyclic features, we coded them using the appropriate n-th unit roots. The number of words, characters, capital letters, and numeric characters were extracted from the *campaign title*. In addition, we created Boolean features based on the *campaign type* to check for frequent words. One-hot encoding was used to code multiple *categories* of campaigns, as well as *region* and *platform name*.

4.2. Time series classification

After considering the various business policies implemented by the different platforms to evaluate the effectiveness of a campaign, we concluded that defining the binary label of *success* at 98% of the value of the *minimum amount* would have been more appropriate. This decision was made because some campaigns narrowly missed their targets but were still considered successful by the platforms, as it would be easy to replicate the campaign and achieve the desired result. Then, before performing classification, we divided the dataset into training and test sets using *stratified* sampling based on the *success* label. Considering the imbalanced distribution of successful class instances, we employed the BorderlineSMOTE algorithm [14] and applied undersampling on the training set to increase the proportion of the minority class and obtain more robust results.

We performed sequences of classification tasks in which the time series length was n points (10 or 14, as mentioned above), after standardizing all features. First, we attempted classification using the features and the first value of the *money* and *investors* time series, i.e. (x_1, y_1). Next, we gradually added the i-th value of all time series (x_i, y_i, r_i^x, r_i^y), $i = 2, ..., n$, and trained the models accordingly. We conducted n training cycles. Each round involved training ten models using the following algorithms:

- Decision Tree: a model that partitions the data into smaller subsets based on a set of rules or criteria;
- *k*-NN: the algorithm calculates the distance between a new input and existing data points and then predicts the output value based on the values of its *k*-nearest neighbors;
- Logistic Regression: a statistical model that is used to predict a binary outcome based on predictor variables, using a logistic function to convert a linear combination of the input features into a probability value between 0 and 1;
- Support Vector Machine (SVM): a supervised learning algorithm that separates data into classes by finding the best hyperplane that maximizes the margin between the classes;
- Bootstrap Aggregating (Bagging): an ensemble learning method that combines multiple decision trees to create a more accurate model. Each tree is trained on a random subset of the data, and their results are combined to make the final prediction;
- Random Forest: an extension of Bagging that also randomly selects subsets of features used in each data sample;
- Adaptive Boosting (AdaBoost): a method similar to Bagging. However, unlike Bagging, it assigns weights to the training examples and modifies them in each iteration to concentrate on the examples that were inaccurately classified in the previous iteration;
- eXtreme Gradient Boosting (XGBoost): is an optimized implementation of Gradient Boosting that uses a more regularized model formulation to prevent overfitting and a more efficient algorithm to speed up the computation;
- Histogram Gradient Boosting: a variant of Gradient Boosting that uses histogram-based techniques to discretize the input features and speed up the computation;
- Multi-Layer Perceptron (MLP): a fully-connected artificial neural network that consists of multiple layers of interconnected nodes and uses a backpropagation algorithm to learn from the training data.

Algorithm 1 Training procedure

```
procedure TRAINING_MODELS(typology_list, Models, $\mathbf{x}$, $\mathbf{y}$, $\mathbf{r}^{\mathbf{x}}$, $\mathbf{r}^{\mathbf{y}}$, other_data)
    for typology in typology_list
        for model in Models
            for $i \in 1, \ldots, len(\mathbf{x})$
                if $i$==1 then
                    training model on $x_1$, $y_1$, other_data[typology]
                else
                    training model on $(x_j, y_j, r_j^x, r_j^y)$, $\forall j \leq i$, other_data[typology]
                evaluation on the test set
```

Algorithm 1 reports the overall training procedure. We repeated this procedure on both the equity and lending campaign data subsets, evaluating the models with the following metrics:

- Accuracy: measures the overall performance of a classification model and represents the percentage of instances classified correctly out of the total number of instances;
- Precision: measures the percentage of true positives (correctly classified positive instances) out of the total number of instances classified as positive. It is a measure of the model's ability to avoid false positives;
- Recall: measures the percentage of true positives out of the total number of true positive instances in the dataset. It is a measure of the model's ability to identify all positive instances, including those incorrectly classified as negative;
- F1-score: is a harmonic mean of precision and recall and is used as a measure of the overall performance of a classification model. It ranges from 0 to 1, with 1 representing the best possible score. It attaches equal importance to precision and recall, making it useful when classes are unbalanced.

5. Experimental results

For space reasons, we do not report all the results obtained for equity and lending by varying the models and the length of the time series, but we only show the best results. These are shown in the following tables. L10 and L14 refer to the corresponding cases with lengths of 10 and 14, respectively.

Specifically, the models that performed best on the test set were:

- For equity: AdaBoost (L10, see Table 2) and Histogram Gradient Boosting (L14, see Table 3);
- For lending: Bagging and Random Forest (L10, see Tables 4 and 5), and Histogram Gradient Boosting (L10 and L14, see Tables 6 and 7).

Summarizing:

- Lending-based models performed better than their equity counterparts. However, the overall better results were very promising. Performance gradually improved when more time data points were considered; however, the "initial" data points were already good predictors of campaign success;
- Among the models evaluated, ensemble models achieved the highest level of performance, followed by MLP. The simpler models did not perform satisfactorily for the equity case. When considering the equity models, *k*-NN showed a considerable discrepancy in scores compared to all other models. In contrast, both SVM and Decision Tree models achieved slightly lower accuracy than the average of the best-performing models.

Table 2
AdaBoost equity L10 results.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
Accuracy	0.94	0.934	0.934	0.956	0.956	0.973	0.978	0.978	0.984	0.984
Precision	0.973	0.96	0.966	0.98	0.973	0.986	0.987	0.987	0.987	0.987
Recall	0.953	0.96	0.953	0.966	0.973	0.98	0.987	0.987	0.993	0.993
F1	0.963	0.96	0.959	0.973	0.973	0.983	0.987	0.987	0.99	0.99

Table 3
Histogram Gradient Boosting equity L14 results.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}
Accuracy	0.918	0.945	0.951	0.956	0.962	0.945	0.951	0.973	0.973	0.962	0.978	0.978	0.978	0.989
Precision	0.953	0.954	0.967	0.973	0.986	0.973	0.979	0.986	0.98	0.98	0.987	0.987	0.987	0.993
Recall	0.946	0.98	0.973	0.973	0.966	0.96	0.96	0.98	0.987	0.973	0.987	0.987	0.987	0.993
F1	0.949	0.967	0.97	0.973	0.976	0.966	0.969	0.983	0.983	0.976	0.987	0.987	0.987	0.993

Table 4
Bagging lending L10 results.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
Accuracy	0.972	0.982	0.977	0.972	0.977	0.986	0.982	0.982	0.977	0.986
Precision	0.977	0.986	0.986	0.981	0.986	0.991	0.986	0.986	0.99	0.986
Recall	0.995	0.995	0.99	0.99	0.99	0.995	0.995	0.995	0.986	1.0
F1	0.986	0.991	0.988	0.986	0.988	0.993	0.991	0.991	0.988	0.993

Table 5
Random Forest lending L10 results.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
Accuracy	0.968	0.968	0.977	0.972	0.977	0.991	0.982	0.982	0.982	0.986
Precision	0.977	0.977	0.981	0.981	0.981	0.991	0.986	0.986	0.986	0.986
Recall	0.99	0.99	0.995	0.99	0.995	1.0	0.995	0.995	0.995	1.0
F1	0.983	0.983	0.988	0.986	0.988	0.995	0.991	0.991	0.991	0.993

Table 6
Histogram Gradient Boosting lending L10 results.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
Accuracy	0.972	0.977	0.977	0.982	0.982	0.982	0.982	0.977	0.977	0.991
Precision	0.977	0.977	0.981	0.99	0.986	0.986	0.986	0.986	0.986	0.991
Recall	0.995	1.0	0.995	0.99	0.995	0.995	0.995	0.99	0.99	1.0
F1	0.986	0.988	0.988	0.99	0.991	0.991	0.991	0.988	0.988	0.995

Table 7
Histogram Gradient Boosting lending L14 results.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}
Accuracy	0.972	0.982	0.986	0.986	0.982	0.986	0.991	0.982	0.982	0.977	0.986	0.986	0.986	0.995
Precision	0.977	0.981	0.986	0.986	0.981	0.986	0.991	0.986	0.986	0.986	0.991	0.986	0.986	0.995
Recall	0.995	1.0	1.0	1.0	1.0	1.0	1.0	0.995	0.995	0.99	0.995	1.0	1.0	1.0
F1	0.986	0.991	0.993	0.993	0.991	0.993	0.995	0.991	0.991	0.988	0.993	0.993	0.993	0.998

6. Conclusion and future work

The results of our study are positive. They suggest that it is possible to anticipate the financial success of an equity or lending campaign during its initial stages. The data from multiple platforms analyzed in the study reinforce the conclusions and imply potential applicability in various scenarios. In essence, this research represents a significant advance in improving the ability to predict the financial success of equity and lending campaigns. This could produce favorable results for investors, entrepreneurs, and crowdfunding platforms.

In the future, other methods could be developed to predict trends in the values of the sums raised at different stages of the campaign and estimate the final value achieved. Regression models could also be considered to predict relative success considering the maximum amount reached. This could significantly impact the planning and implementation of a crowdfunding campaign. More accurately predicting performance and the final value of the amount raised would allow promoters to manage the campaign, optimize marketing expenses better, define the timing of the various phases, and adjust communication strategies according to performance. Finally, we plan to release a pseudonymized version of our dataset to further promote research in this growing field. For example, in addition to predicting the success of a campaign, information about the characteristics of a good campaign could be derived from the data.

References

[1] L. B. Junge, I. C. Laursen, K. R. Nielsen, Choosing crowdfunding: Why do entrepreneurs choose to engage in crowdfunding?, Technovation 111 (2022) 102385.

[2] W. Wang, Zheng, Hongsheng, Wu, Y. Jim, Prediction of fundraising outcomes for crowdfunding projects based on deep learning: a multimodel comparative study, Soft Computing 24 (2020) 8323–8341.

[3] J. Wang, H. Zhang, Q. Liu, Z. Pan, H. Tao, Crowdfunding Dynamics Tracking: A Reinforcement Learning Approach, in: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, AAAI Press, 2020, pp. 6210–6218. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6087.

[4] X. Ren, L. Xu, T. Zhao, C. Zhu, J. Guo, E. Chen, Tracking and Forecasting Dynamics in Crowdfunding: A Basis-Synthesis Approach, in: 2018 IEEE International Conference on Data Mining (ICDM), 2018, pp. 1212–1217. doi:10.1109/ICDM.2018.00161.

[5] C. Cheng, F. Tan, X. Hou, Z. Wei, Success Prediction on Crowdfunding with Multimodal Deep Learning, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019, 2019, pp. 2158–2164. URL: https://doi.org/10.24963/ijcai.2019/299. doi:10.24963/ijcai.2019/299.

[6] Q. Du, J. Li, Y. Du, G. A. Wang, W. Fan, Predicting crowdfunding project success based on backers' language preferences, Journal of the Association for Information Science and Technology, 72(12) (2021) 1558–1574. doi:10.1002/asi.24530.

[7] K.-F. Yang, Y.-R. Lin, L.-S. Chen, Discovering critical factors in the content of crowdfunding projects, Algorithms 16 (2023). URL: https://www.mdpi.com/1999-4893/16/1/51. doi:10.3390/a16010051.

[8] Y. Song, R. Berger, A. Yosipof, B. R. Barnes, Mining and investigating the factors influencing crowdfunding success, Technological Forecasting and Social Change Volume 148 148 (2019) 43–58. doi:org/10.1016/j.techfore.2019.119723.

[9] Y. J. W. Wei Wang a, Lihuan Guo b, The merits of a sentiment analysis of antecedent comments for the prediction of online fundraising outcomes, Technological Forecasting and Social Change (2022). doi:https://doi.org/10.1016/j.techfore.2021.121070.

[10] O. M. Al-Qershi, J. Kwon, S. Zhao, Z. Li, Predicting crowdfunding success with visuals and speech in video ads and text ads, European Journal of Marketing (2022).

[11] C. Zhong, W. Xu, W. Du, Success prediction of crowdfunding campaigns with project network: A machine learning approach, Journal of Electronic Commerce Research 23 (2022) 99–114.

[12] W. X. Jiatong Shi, Kunlin Yang, Leveraging deep learning with audio analytics to predict the success of crowdfunding projects, The Journal of Supercomputing (2021). doi:10.1007/s11227-020-03595-2.

[13] O. E. F. . Innovation, 7° Report italiano sul CrowdInvesting, Technical Report, Politecnico di Milano, 2022. URL: https://www.osservatoriefi.it/efi/wp-content/uploads/2022/07/reportcrowd2022.pdf.

[14] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-SMOTE: A New Over-Sampling Method

in Imbalanced Data Sets Learning, in: D.-S. Huang, X.-P. Zhang, G.-B. Huang (Eds.), Advances in Intelligent Computing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 878–887.