

Machine Learning for Assessing StartUp Investment Attractiveness

Nataliia Dziubanovska¹, Vadym Maslii¹, Oleksandr Shekhanin¹ and Serhii Protsyk¹

¹ Western Ukrainian National University, 11 Lvivska Str., Ternopil, 46009, Ukraine

Abstract

The primary objective of this study is to utilize machine learning (ML) methods for analyzing and predicting the success of startups, particularly focusing on the IT sector. This article explores the application of ML methods for assessing the investment attractiveness of startups based on data from the crowdfunding platform Kickstarter. An analysis of the dataset covering various aspects of projects, such as funding volume, raised funds, project category, number of backers, and its status, is conducted. From all the provided data, a dataset focused on startups from the Technology category is formed, reflecting the importance of the study in the context of innovation and technological development. The application of classification methods, including decision trees, allows predicting the success or failure of a startup with high accuracy based on input data. The findings of this study can be valuable for investors, entrepreneurs, and researchers interested in risks and opportunities in the field of investing in innovative projects.

Keywords

classification, decision tree, investment, Kickstarter, machine learning, startup.

1. Introduction

The beginning of the 21st century is characterized by the emergence of a fundamentally new phenomenon in the economies of countries – startups. Every year, a significant number of new innovative companies appear, which significantly change the global economy and, thanks to their inventions, simultaneously ensure dynamic development and competitiveness of the economy of individual countries. Maximizing profit, achieved through the implementation of innovative technological solutions, and attracting venture foreign capital to finance the most attractive startups, has a positive impact on GDP growth, which ultimately affects the income level of the entire society.

In Ukraine, despite the full-scale war with Russia, the main drivers of the domestic economy have become IT companies and technological startups. As experts note, “according to the results of nine months of 2022, the industry showed growth of 13%, and the market volume amounted to almost \$5.5 billion.” Ukrainian technological startups are actively working on various innovative projects in such areas as artificial intelligence, blockchain, fintech, e-commerce, and many others. These companies are mostly aimed at the global market and have ambitions to become leaders in their fields.


The Ukrainian startup ecosystem began to form about 10 years ago and continues to be in the stage of formation. The main source of funding is bootstrapping (own funds), however, in conditions of war, economic difficulties, and other adverse factors, both the profits of startups and the personal savings of founders have significantly decreased. This forces domestic developers to enter international markets in search of investors and consumers [1].

The First International Workshop of Young Scientists on Artificial Intelligence for Sustainable Development, May 10-11, 2024, Ternopil, Ukraine

✉ n.dziubanovska@wunu.edu.ua (N. Dziubanovska); v.maslii@wunu.edu.ua (V. Maslii); shekhanin2022@gmail.com (O. Shekhanin); serhiy12308@gmail.com (S. Protsyk)

ORCID 0000-0002-8441-5216 (N. Dziubanovska); 0000-0002-9672-9669 (V. Maslii); 0009-0005-3561-4242 (O. Shekhanin); 0009-0009-8395-8946 (S. Protsyk)

© 2024 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

At the same time, the growing number of business initiatives provides investors with wide opportunities for investment, however, the process of assessing the potential profitability of a startup can be complex and risky. In the world of modern technologies, startups play a key role in shaping the economic landscape. They represent a source of innovation and stimulate competitiveness. However, considering such investment opportunities, financial participants face significant instability and uncertainty associated with the risks accompanying the startup development process. Therefore, the use of ML methods for analyzing and predicting the success of startups becomes an important tool for investors, helping them make informed decisions regarding capital placement.

This research seeks to address the pressing need for robust tools that can assist investors in making informed decisions regarding capital allocation in the dynamic and uncertain environment of startup investment.

2. Literature review

In the domestic scientific literature, considerable attention is devoted to the development of startups and the forms of their financing. In particular, in the work of O. Kurchenko [2], the results of a survey of Ukrainian startups that were in the top 100 in 2014 are presented. It was found that 70% of companies used their own funds as startup capital, so it was logically assumed that this category of startups requires the implementation of a state support program. A. Dub and M. Khlopetska [3] focus on the main sources of startup financing in conditions of instability and distinguish such types of main investors in startups: business angels, business accelerators (business incubators), and venture funds. K. Raputa [4] examines the importance of venture financing as a source of investment for startups; determines the stages at which financing takes place and the criteria for making investment decisions. L. Boltianska, L. Andreeva, and O. Lysak [5] note that in conditions of economic instability, the issue of finding sources of financing for Ukrainian startups is relevant. The authors characterize the most common financing models: personal investments, F&F, venture capital, business angels, business incubators, grants, bank loans, and crowdfunding. M. Dyba and O. Hernego [6] point out that the role of crowdfunding technologies is increasing at both the global and national levels. N. Versal and Y. Dudnyk [7], based on the use of statistical-economic methods, identified trends in the development of crowdfunding in various regions of the world and clarified the factors influencing this development. A. Schwienbacher and V. Larralde [8] identify crowdfunding as a competition, most often open on the Internet, aimed at obtaining financial resources to achieve specific goals. According to this investment scheme, investors, through specialized internet platforms or crowdfunding platforms, finance new or existing startups. T. Esen, M. S. Dahl, and O. Sorenson [9], based on the evaluation of linear probability models (LPM) built on data characterizing startup founder teams, identified which attribute most accurately predicts financing. B. Yin and J. Luo [10] analyzed datasets of real startup profiles submitted to the first stage of a startup accelerator in Southeast Asia; four criteria of value were identified, which were subsequently used to build regression models predicting screening and selection outcomes; the researchers concluded that better understanding of decision criteria could improve the decision-making process regarding startup investments.

Given that significant volumes of information characterizing projects are expressed not only in attribute features but also in quantitative characteristics, there is a need for the application of ML – a subfield of artificial intelligence that allows identifying investment-attractive objects with a high probability of success. This applied aspect of selecting startups for financing is insufficiently researched.

Thus, the purpose of the article is to investigate the possibilities of applying ML algorithms to select investment-attractive objects in the process of crowdfunding financing.

3. Methodology

When exploring the possibility of using ML methods for investment decision-making, the most intuitive algorithm for understanding and interpretation – decision tree – was applied. This method is used for classification and regression based on input feature values. Decision trees can work with different types of data, including categorical and numerical, making them flexible for use in various situations.

The main task was to divide startups into two categories: successful and unsuccessful, in order to analyze their performance and attractiveness to investors. The result of this study was the creation of a trained decision tree model capable of classifying new startups based on their characteristics. This provides investors with the opportunity to make informed decisions about investing in new projects, taking into account the likelihood of their success or failure.

The task was implemented using the DecisionTreeClassifier classifier from the Scikit-learn library in Python. It is used to build a classification model based on data using the decision tree algorithm and is responsible for creating and training the model based on input data and responses. The feature selection is based on the calculation of the Gini coefficient. When training the model, a random number generator is applied, which determines the randomness of various aspects of the process and ensures greater variability and reliability of the model results.

The trained decision tree model allows for the use of objective criteria for classifying startups, ensuring the rationale behind the decisions made. This enables risk reduction and increases the effectiveness of the investment portfolio by directing capital towards directions with the highest potential for success.

For training the decision tree model, data from Kickstarter Projects [11] were used. This dataset contains information about various startups seeking funding through the crowdfunding platform Kickstarter. It includes important characteristics such as country, category, funding goal, pledged amount, number of backers, project status, and others. This approach to analyzing startups serves as an additional tool for investors seeking opportunities for smart capital allocation and increasing its profitability. Utilizing data from Kickstarter Projects allows for considering diverse factors and assessing the potential success of each specific startup.

The Kickstarter Projects dataset contains information about 374,853 startups from various categories ('Fashion', 'Film & Video', 'Art', 'Technology', 'Journalism', 'Publishing', 'Theatre', 'Music', 'Photography', 'Games', 'Design', 'Food', 'Crafts', 'Comics', 'Dance'), different countries ('United States', 'United Kingdom', 'Canada', 'Australia', 'New Zealand', 'Netherlands', 'Sweden', 'Denmark', 'Norway', 'Ireland', 'Germany', 'France', 'Spain', 'Belgium', 'Italy', 'Switzerland', 'Austria', 'Luxembourg', 'Singapore', 'Hong Kong', 'Mexico', 'Japan'), and various states ('Failed', 'Successful', 'Cancelled', 'Suspended', 'Live'), among others (Figure 1).

	ID	Name	Category	Subcategory	Country	Launched	Deadline	Goal	Pledged	Backers	State
0	1860890148	Grace Jones Does Not Give A F\$#% T-Shirt (limi...	Fashion	Fashion	United States	2009-04-21 21:02:48	2009-05-31	1000	625	30	Failed
1	709707365	CRYSTAL ANTLERS UNTITLED MOVIE	Film & Video	Shorts	United States	2009-04-23 00:07:53	2009-07-20	80000	22	3	Failed
2	1703704063	drawing for dollars	Art	Illustration	United States	2009-04-24 21:52:03	2009-05-03	20	35	3	Successful
3	727286	Offline Wikipedia iPhone app	Technology	Software	United States	2009-04-25 17:36:21	2009-07-14	99	145	25	Successful
4	1622952265	Pantshirts	Fashion	Fashion	United States	2009-04-27 14:10:39	2009-05-26	1900	387	10	Failed
...
374848	1486845240	Americas Got Talent - Serious MAK	Music	Hip-Hop	United States	2018-01-02 14:13:09	2018-01-16	500	0	0	Live
374849	974738310	EVO Planner: The World's First Personalized Fl...	Design	Product Design	United States	2018-01-02 14:15:38	2018-02-09	15000	269	8	Live
374850	2106246194	Help save La Gattara, Arizona's first Cat Cafe!	Food	Food	United States	2018-01-02 14:17:46	2018-01-16	10000	165	3	Live
374851	1830173355	Digital Dagger Coin	Art	Art	United States	2018-01-02 14:38:17	2018-02-01	650	7	1	Live
374852	1339173863	Spirits of the Forest	Games	Tabletop Games	Spain	2018-01-02 15:02:31	2018-01-26	24274	4483	82	Live

374853 rows × 11 columns

Figure 1: Dataset Kickstarter Projects

Rapid technological advancement in the modern world dictates the pace of life and business standards. These technologies play a key role in all aspects of our lives, from communications and media to business and science. This has led to the narrowing of the existing database to startups in the Technology category (32, 562). These startups are often characterized by innovative approaches and products that open up new opportunities for the market. Due to their innovativeness, technology startups have great potential to attract investments and interest from clients. Another important advantage is the ability to scale the business, creating a sustainable and profitable model. Technological solutions allow for efficient interaction with customers and partners on a global scale, promoting rapid business growth. Successful technology startups typically have a team with relevant expertise and experience, facilitating the effective implementation of innovative solutions and addressing complex tasks.

To build the ML model, we selected startups with the status 'Failed' and 'Successful' (27, 046). This is due to their significant impact on the final outcome. 'Failed' indicates project failure, which is important for studying the reasons for failure and avoiding similar mistakes in the future. On the other hand, 'Successful' signifies successful completion of the startup, providing valuable information for studying successful strategies and practices. A detailed study of these two statuses will provide a complete picture of the dynamics of success and failure in the startup sphere. The statuses 'Cancelled', 'Suspended', and 'Live' also have their importance in understanding the dynamics of startups, but in this case, they can be considered less significant for analysis compared to 'Failed' and 'Successful'. 'Cancelled' indicates that the startup was cancelled before achieving the goal, which may be the result of strategic or financial obstacles. 'Suspended' means temporary suspension of the project, which may be caused by external factors such as legal or financial problems. 'Live' indicates that the startup is still in the active phase of fundraising or project implementation, and this information may be less representative for analysis since the project's outcome is not yet known. Thus, focusing on 'Failed' and 'Successful' will allow us to concentrate on key aspects of startup success and failure. Visualizing the number of 'Failed' and 'Successful' startups in the form of a histogram will highlight the difference in size between the two categories (Figure 2).

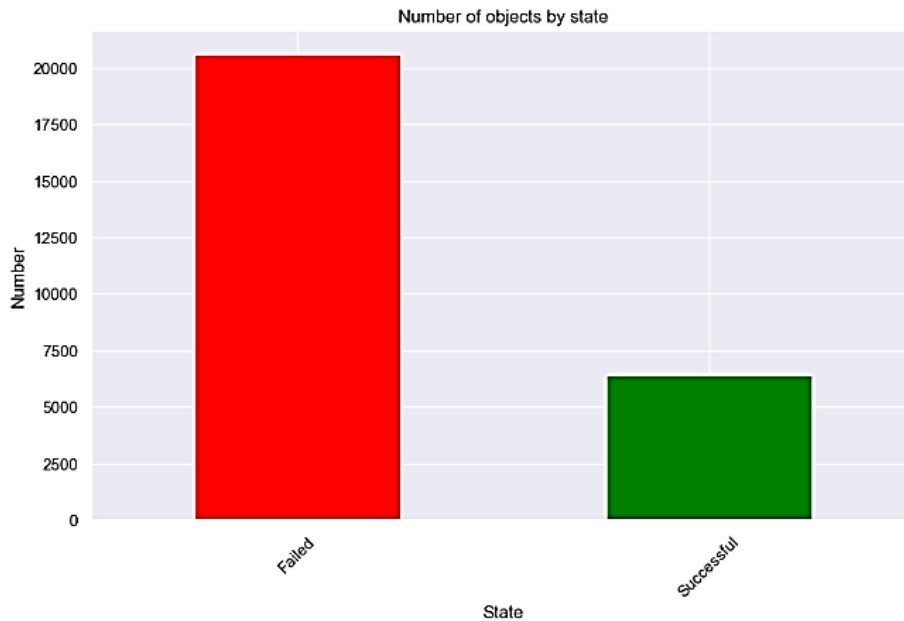


Figure 2: Number of objects by State

Considering the predominant number of 'Failed' startups, investigating the reasons for their failure is particularly relevant. Using the decision tree method allows obtaining valuable insights into the factors influencing the success of startups and can be applied as an additional tool in the decision-making process in business and investments.

4. Results

Based on the newly formed dataset of startups in the 'Technology' category with the status 'Failed' and 'Successful', we construct a decision tree. The main features selected are 'Goal', 'Pledged', and 'Backers'. Additionally, to obtain more informative insights from the 27,046 startups, we select those with 'Pledged' > 0 and 'Backers' > 0. Thus, the dataset is reduced to 21,636 entries.

Reviewing a segment of the decision tree (Figure 3) can help better understand the process of node formation and decision-making in the ML model.

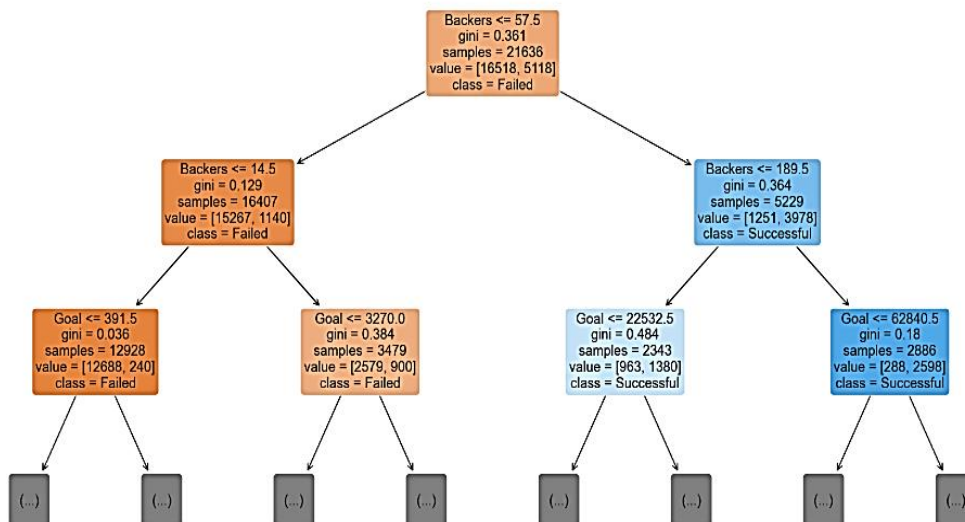


Figure 3: Segment of the Decision Tree

While analyzing the segment, important features used for data splitting can be identified, along with considering the criteria for selecting the optimal split. Investigating the structure of the tree segment can also highlight the significance of each node in decision-making and influence the final classification outcome. This approach allows for a deeper understanding of the internal mechanism of the decision tree and its interaction with the input data.

To assess the model's ability to distinguish between the 'Failed' and 'Successful' classes, we construct the Receiver Operating Characteristic (ROC) curve (Figure 4) and compute the Area Under the Curve (AUC).

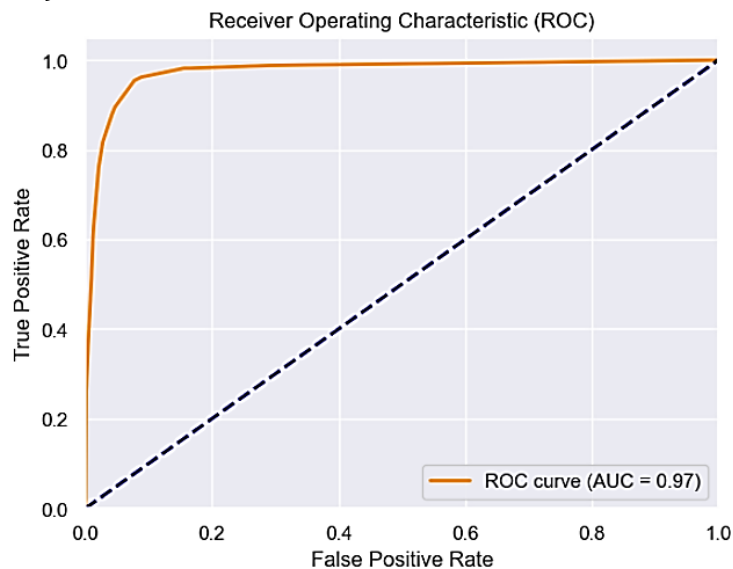


Figure 4: Receiver Operating Characteristic (ROC) of the Decision Tree

The value of the Area Under the Curve (AUC) of the ROC curve is 0.97. This high value, close to 1, indicates the model's excellent ability to distinguish between the 'Failed' and 'Successful' classes. Such a high AUC value confirms the model's high accuracy in prediction.

Now that we have a trained ML model, we can use it to predict the class of new objects by inputting their characteristics. This allows us to effectively classify new data and use the model in practical situations for decision-making.

For example, based on our dataset and key characteristics, providing the values for five new startups: 'Goal': [245000, 750000, 200000, 120000, 250000], 'Pledged': [300000, 11000, 300000, 7000, 3000], 'Backers': [850, 20, 100, 3000, 150], we obtain the predicted class values for their belonging to the respective class. In our case – Predicted Classes: ['Successful' 'Failed' 'Successful' 'Failed' 'Failed']. Thus, using this tool for investment decision-making allows for efficiently assessing the potential success of new projects and developing optimal investment strategies.

Furthermore, having a ready ML model for classifying startups into 'Failed' and 'Successful' allows us to set additional constraints for selected characteristics or add new criteria for assessing their investment attractiveness. For instance, if we are interested in startups with a 'Goal' > 100000, we add this constraint to form a new database (2089) and build a new decision tree (Figure 5). Based on this, we can also test new projects and determine their likely belonging to one of the two classes. Moreover, by continuously refining and updating our ML model with new data, we can enhance its predictive accuracy and adapt it to evolving market trends. Additionally, leveraging advanced analytics tools and techniques can provide deeper insights into the underlying factors driving startup success or failure, empowering investors to make more informed decisions. This iterative process of analysis and refinement fosters a dynamic approach to investment strategy, optimizing returns and mitigating risks in the ever-changing landscape of startup ventures.

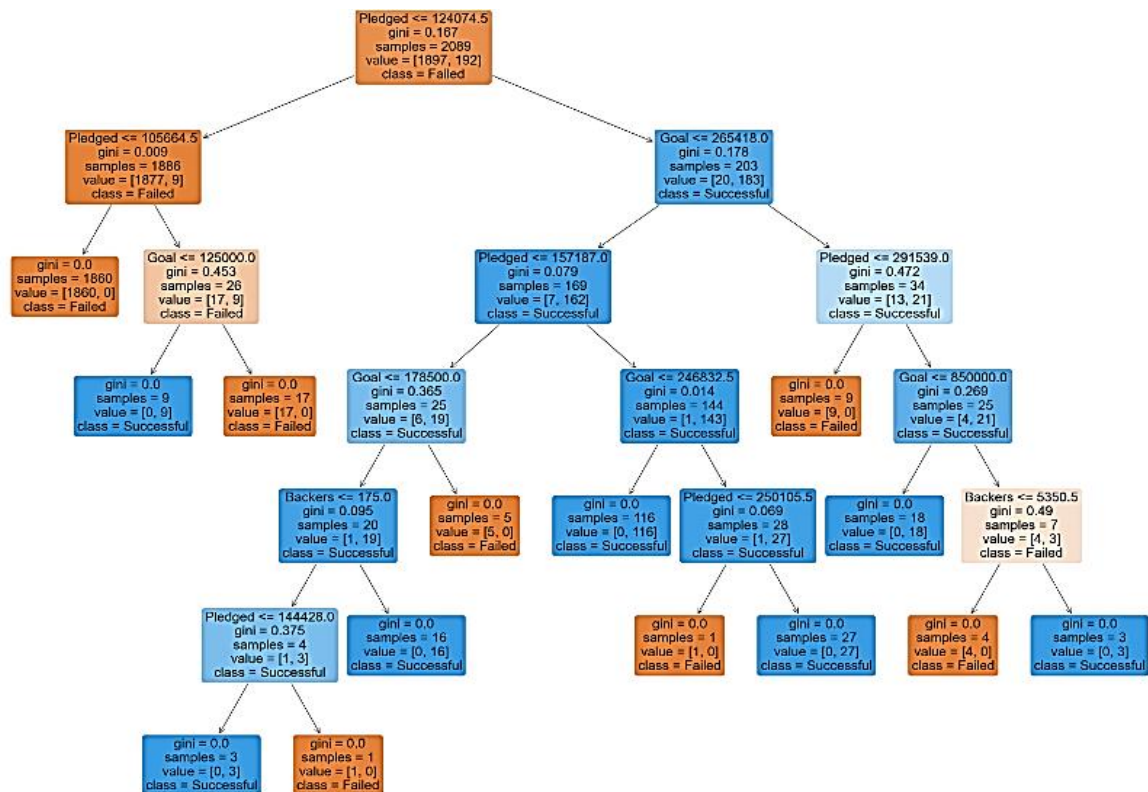


Figure 5: Decision Tree for Startups with Goal Greater than 100000

Thus, the ability to establish additional conditions and constraints enhances the accuracy of the model and ensures a more objective analysis of investment opportunities. This approach helps make informed decisions regarding the allocation of investment resources and reduces the risk of financial losses.

ML, as noted by Catalyst Fund experts, has great potential to become a powerful and effective decision-making tool for investors, but this subfield of artificial intelligence is still far from perfect or suitable as a black box. ML helps investors check their biases, identified using more reliable data, and improve operational efficiency. Thus, valuable human time can be spent on an important part of due diligence, where intuition and prudence matter. Whether ML will become so super intelligent as to become the sole decision-making tool for investors in the early stages of startup funding, only time will tell [12].

Conclusions

Therefore, considering startups as an investment object, especially early-stage startups, investors strive to conduct as thorough due diligence as possible. To make it fairer and with minimal time costs, it is advisable to apply ML. By combining quantitative characteristics of startups from various platforms, ML algorithms allow for unbiased evaluation of investment ideas and prospects.

The model obtained in the study, built based on decision tree, can be useful for various groups of interested parties. For example, investors and financial analysts can use this data to make investment decisions, assessing the potential success of startups and risks. Entrepreneurs can analyze successful and unsuccessful startups to understand the factors influencing their success and improve their own projects. Researchers can use this data to analyze trends in the startup world and technology development. Government structures and regulatory bodies can use this data to evaluate the effectiveness of policies and programs supporting startups. Students and researchers can use this data for academic research and education. Also, this decision tree model

can be used both as a ready-made database and as an example for creating their own models based on past implemented projects, methodologies, and criteria that meet specific user needs.

References

- [1] L. Shkil, Funding of Ukrainian startups during the war: where to look for investments, 2022. URL: <https://ain.ua/2022/11/16/finansuvannya-ukrayinskyh-startapiv-pid-chas-vijny/>
- [2] O. Kurchenko, Formation and development of startups in Ukraine: problems and solutions. Ukrainian society. №2 (57) (2016): 80-87.
- [3] A. Dub, M. Khlopetska, Sources of funding for startups and opportunities to attract them in Ukraine. Socio-economic problems of the modern period of Ukraine. Volume 1 (2016): 87-92.
- [4] K. Raputa, Directions for improving the mechanism of venture financing of startups in Ukraine. Investments: practice and experience. №2 (2021): 56-63.
- [5] L. Boltianska, L. Andreieva, O.Lysak, Idea selection and startup financing. Collection of scientific works of the Dmytro Motorny Tavri State Agro-Technological University (economic sciences). №1 (43) (2021): 5-12.
- [6] M. Dyba, O. Herneho, Global trends and development potential of the crowdfunding market in Ukraine. Ukraine economy. № 2 (699). (2020): 66—79. URL: <https://doi.org/10.15407/economyukr.2020.02.066>.
- [7] N. Versal, Ya. Dudnyk, Crowdfunding as an alternative FINTECH ecosystem in the financial market. Bulletin of Taras Shevchenko Kyiv National University. №4 (217) (2021): 26-37.
- [8] Schwienbacher A., Larralde B. Crowdfunding of small entrepreneurial ventures. SSRN Electronic Journal. 2010, URL: http://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID1699183_code301672.pdf
- [9] Tekin Esen, Michael S. Dahl, Olav Sorenson, Jockeys, horses or teams? The selection of startups by venture capitalists. Journal of Business Venturing Insights. Volume 19 (2023). URL: <https://doi.org/10.016/j.bvi.2023.e00383>.
- [10] Bangqi Yin, Jianxi Luo, How Do Accelerators Select Startups? Shifting Decision Criteria Across Stages. IEEE Transactions on Engineering Management. Volume 65 Issue 4 (2018): 1-16. URL: <https://doi.org/10.1109/TEM.2018.2791501>.
- [11] Kickstarter projects. Kaggle. URL: <https://www.kaggle.com/datasets/ulrikthgyepedersen/kickstarter-projects>
- [12] Kelly Nguyen, How Can Investors Use Machine Learning to Pick the Right Startups? 2017. URL: <https://bfaglobal.com/catalyst-fund/insights/how-can-investors-use-machine-learning-to-pick-the-right-startups/>