SCIENTIFIC HORIZONS

Journal homepage: https://sciencehorizon.com.ua

Scientific Horizons, 24(11), 72-84

UDC 339.727.22:334.012.4]:005.521
DOI: 10.48077/scihor.24(11).2021.72-84

Applying Machine Learning Approach to Start-up Success Prediction

Olena Piskunova*, Larysa Ligonenko, Rostyslav Klochko, Tetyana Frolova, Tetiana Bilyk

Kyiv National Economic University named after Vadym Hetman
03057, 54/1 Peremohy Ave., Kyiv, Ukraine

Article's History:
Received: 20.10.2021
Revised: 19.11.2021
Accepted: 22.12.2021

Suggested Citation:
Piskunova, O., Ligonenko, L., Klochko, R., Frolova, T., & Bilyk, T. (2021). Applying machine learning approach to start-up success prediction. *Scientific Horizons*, 24(11), 72-84.

Abstract. Predicting the success of a new venture has always been a topical issue for both investors and researchers. Nowadays, it has become even more relevant concerning start-ups-young innovative and technology enterprises aimed at scaling their businesses. The purpose of this study is to create a model for predicting start-ups' success based on their descriptive characteristics. A model that connects such start-up features as the period from foundation to the first financing, the area of activity, type, and amount of the first financing round, business model, and applied technologies, with the start-up investment success, which refers to re-investment, has been developed using data from the Dealroom platform on statistics of start-ups activity and their description. The final sample included 123 start-ups that are founded or operate in Ukraine. Three machine learning algorithms are compared: Logistic Regression, Decision Tree, and Random Forest. Acceptable results were obtained in terms of Accuracy, Sensitivity, and F-score, despite the limited data. The best model concerning start-up success prediction is determined by a Decision Tree, with an average effectiveness of 61%, 55%, and 52%, respectively. The AUC level for the Decision Tree achieved 58%, which is lower than the Logistic Regression and Random Forest scores (65%). But the last models had done so well by better predicting start-up failures, while more practical is the ability to predict their success. All models showed an acceptable level of AUC to confirm with confidence their effectiveness. The decision support system for the investment object can be helpful for entrepreneurs, venture analysts, or politicians who can use the built models to predict the success of a start-up. This forecast, in turn, can be used to drive better investment decisions and develop relevant economic policies to improve the overall start-up ecosystem

Keywords: start-up ecosystem, start-up, innovation, decision support system, classification, data modelling, predicting the success of a start-up

*Corresponding author

INTRODUCTION

Start-ups, which according to the European Association of Start-ups are recognised as "an independent organisation, which is younger than five years and aimed at creating, improving, and expanding a scalable, innovative, technology-enabled product with high and rapid growth" [1], are the driving force of rapid and innovative change in present-day business environment. According to the World Economic Forum, the global start-up economy was estimated at 2.8 billion dollars in 2019, and its development is about three to four times faster than the growth of economies in many countries [2]. Their high growth rates and the flexibility in deploying innovative business models make them an increasingly visible element of the global economy, as creators of economic value and destroyers of existing industries, as a source of employment generation, and a place for talent development, as a way to commercialise science and research and development (R&D).

Summarising the studies available, several research groups can be distinguished, depending on the research object. The first group of researchers based their models on quantitative methods for assessing the key factors of start-up success. For example, in [3], using statistical tests, researchers try to evaluate which business model brings the highest success rates. By success, they mean the level of profitability, the level of profit growth, the start-up valuation, and the amount of investment. Another example of this type of research is the study that determines the impact of indicators such as the level of innovation, education of founders and employees, the level of investment, and the start-up size on the growth of a start-up's business performance. As the criterion for success, researchers consider the level of income per employee [4]. A similar approach was applied in [5], but its authors consider the fact of continuing to carry out start-ups activities as a criterion of success.

The second group of researchers focused on identifying the qualitative internal factors of start-up success. In particular, in [6], based on a survey of 25 experts, the authors summarise the answers to the question "What are the key factors for the success of a start-up?" Using the AHP method, eight main factors of start-up success were identified as follows: product uniqueness, product characteristics, customer demand, marketing promotion, distribution channels, after-sales service, new product development, and financial support. In [7], researchers assessed the influence of the digital activity of the founders and their affiliation with venture investors on the business performance of a start-up by constructing multiple regression. The researchers consider the asset turnover ratio to be the criterion for the success of a start-up.

The third group of researchers sets themselves the task of identifying external factors that increase the probability of start-up success, at a country level in particular. Thus, in [8], the application of the principal component analysis allowed identifying five main factors, the presence of which contributes to the development of a start-up ecosystem in the country: 1) access to human capital; 2) the quality and results of the institutions' activities; 3) focus on market conditions; 4) business environment; 5) development potential.

The methodology of start-up success prediction, which is used in research on this issue, is very diverse. However, most frequently, researchers use machine learning methods. The level of complexity of research using this technique is also very different: from simple works [9; 10], where one method is used – a logistic model, to complex (from a methodological standpoint) research using 9 algorithms, including Random Forest, Bayesian network, Decision tree, their comparison and selection of the best [11]. In most studies, 2-3 algorithms are used, followed by the selection of the best one based on the level of forecasting accuracy.

Considering the complexity of information support of this study (initial data), researchers choose different approaches to defining the success of a start-up, which leads to different accuracy of the developed models:

- in [12], the success of a start-up is defined by conducting an IPO, selling a start-up, or receiving re-financing in an amount exceeding the previous maximum. In [10], this is the fact that a start-up has a profitability of more than 20 percent, new patent applications, and participation in the innovation subsidy programme is considered a success;
- in [13] argues the expediency of changing the criterion of "obtaining re-financing" to the criterion of "reaching the round". Researchers applied 3 algorithms (logistic regression, support vector machine, gradient boosting), obtaining F-score accuracy rates of 57%, 34%, 43%, respectively;
- in [14], the random forest and gradient boosting methods were used in combination with the success definition "the start-up is still functioning" (the researchers predicted that the start-up would fail, not succeed). The maximum performance of the study was set at 94.5% in terms of accuracy, and 92.91% in terms of AUC (Area Under the Curve);
- a higher forecasting accuracy was obtained by studying the correlation between a start-up's digital activity and its chances of surviving for more than 5 years [15]. The researchers managed to predict the start-up survival probability with 91% accuracy, using random forest and gradient boosting methods.

Thus, there are two types of problems that have not been adequately addressed in previous studies so far: 1) what start-up should be considered as a success, what factors determine it; 2) what techniques, having a relatively small amount of public data on the start-up activities, are used to evaluate the prospects for its success.

That is why the actual scientific task is to further investigate the essence, manifestations, and varieties of the concept of "start-up success", to identify the relationships between indicators of start-up activity, and their

impact on the achievement of various types of success. The limited information field of the research determines the interest in improving the input data preparation methods, setting up machine learning models, and evaluating their quality.

Critically comprehending the available approaches to determining and predicting the success of a start-up, this study yielded the following conclusion: various events that meet the interests of the start-up in terms of implementing a development strategy or the interests of its founders can be considered the success of a start-up. Understanding this polyvariety allows distinguishing between different types of success as follows:

- investment success – getting additional financing;
- customer success – achieving the target volume or increase in the number of users (consumers) of a start-up product;
- market success – achieving target sales of a start-up's product or achieving target market share;
- adaptive success – continued existence for a certain period (more than 5 years), which allows identifying the transition from the status of a young new to an ordinary enterprise;
- financial success – IPO (Initial Public Offering) or selling a start-up, which allows its founders and primary investors to exit the start-up and monetise their investments. Financial success is in line with the "classic understanding" of start-up success [12].

The purpose of this study is to develop a methodology for predicting the success of a start-up using machine learning methods based on its activity data from open sources. The results of this study will considerably facilitate the decision-making process of choosing objects and areas of investment (start-up specialisation) for entrepreneurs, potential co-founders of start-ups, and venture capitalists.

To achieve the said purpose, *the following tasks* were identified:

– to prepare the available public data on the activities and funding of start-ups for the features of the application of machine learning algorithms and determine the criterion indicating the success of a start-up;

– to systematise the model building methodology for the start-up success prediction: choosing algorithms and adjusting their hyperparameters, setting up cross-validation, selecting indicators of forecasting accuracy;

– according to the developed methodology, to implement three different algorithms for predicting the investment success of a start-up;

– to conduct an in-depth comparative quality analysis of the constructed model and, on this basis, to substantiate the best methodology for predicting the investment success of a start-up in terms of accuracy.

MATERIALS AND METHODS

Data pre-processing

The study uses open data from the Ukrainian version of the Dealroom resource [16]. Dealroom.co is a leading provider of data on start-ups and technology ecosystems in Europe and all around the world. The Ukrainian version allows analysing data on start-ups that are based or operate in Ukraine with ease. Due to limited resources, information about only 566 start-ups related to Ukraine could be accessed.

The generated data set contains 98 start-up characteristics fields. A list of all fields is presented in Figure 1.

NAME	TAGS	FACEBOOK LIKES	FOUNDERS UNIVERSITES
PROFILE URL	B2B/B2C	TWITTER FOLLOWERS	FOUNDERS COMPANY EXPERIENCE
WEBSITE	REVENUE MODEL	TWITTER TWEETS	FOUNDERS FIRST DEGREE
TAGLINE	LAUNCH DATE	TWITTER FAVORITES	FOUNDERS FIRST DEGREE YEAR
ADDRESS	CLOSING DATE	SW TRAFFIC 6 MONTHS	FOUNDERS LINKEDIN
HQ REGION	INDUSTRIES	SW TRAFFIC 12 MONTHS	FOUNDERS FOUNDED COMPANIES TOTAL FUNDING
HQ COUNTRY	SUB INDUSTRIES	ANGELLIST	LISTS
HQ CITY	DELIVERY METHOD	FACEBOOK	COMPANY STATUS
LATITUDE	GROWTH STAGE	TWITTER	APP DOWNLOADS LATEST (IOS)
LONGITUDE	DEALROOM TAG	LINKEDIN	APP DOWNLOADS 6 MONTHS (IOS)
LOCATIONS	YEARLY GROWTH (SIMILARWEB)	GOOGLE PLAY LINK	APP DOWNLOADS 12 MONTHS (IOS)
FOUNDING LOCATION	ALEXA GROWTH (ALL TIME)	ITUNES LINK	APP DOWNLOADS LATEST (ANDROID)
TEAM (DEALROOM)	EMPLOYEES	EACH ROUND TYPE	APP DOWNLOADS 6 MONTHS (ANDROID)
TEAM (EDITORIAL)	EMPLOYEES (2016, 2017, 2018, 2019, 2020, 2021)	EACH ROUND AMOUNT	APP DOWNLOADS 12 MONTHS (ANDROID)
INVESTORS	EMPLOYEES IN HQ country (2016, 2017, 2018, 2019, 2020)	EACH ROUND CURRENCY	TRAFFIC COUNTRIES
EACH INVESTOR TYPES	LAST KPI DATE	EACH ROUND DATE	TRAFFIC SOURCES
LEAD INVESTORS	PROFIT (2016, 2017, 2018, 2019)	TOTAL ROUNDS NUMBER	SIMILARWEB RANK 3/6/12 MONTHS
TOTAL FUNDING (EUR M)	EBITDA (2016, 2017, 2018, 2019)	EACH ROUND INVESTORS	EMPLOYEE RANK 3/6/12 MONTHS
LAST ROUND	REVENUE (2016, 2017, 2018, 2019)	LOGO	APP RANK 3/6/12 MONTHS
LAST FUNDING	FINANCIALS CURRENCY	FOUNDERS	NUMBER OF ALUMNI EUROPEAN FOUNDERS THAT RAISED > 10M
LAST FUNDING DATE	VALUATION	FOUNDERS STATUSES	TECHNOLOGIES
FIRST FUNDING DATE	VALUATION CURRENCY	FOUNDERS GENDERS	INCOME STREAMS
SEED YEAR	VALUATION (EUR)	FOUNDERS IS SERIAL	TECH STACK DATA (BY PREDICTLEADS)
OWNERSHIPS	VALUATION DATE	FOUNDERS BACKGROUNDS	TRADE REGISTER NUMBER
SDGS	CORE SIDE VALUE		

***Figure 1**. List of all fields in the start-up data set*

Evidently, many fields are not acceptable for usage in modelling. For example, PROFILE URL (link to the start-up site), FOUNDERS (names of the founders), or LOGO (link to download the start-up logo). Accordingly, the first step will be to clean up the redundant fields from the data. The rest of the factors are presented in Figure 2.

NAME	B2B/B2C	FACEBOOK LIKES	TWITTER FOLLOWERS
REVENUE MODEL	TWITTER TWEETS	LAUNCH DATE	TWITTER FAVORITES
SW TRAFFIC 6 MONTHS	SW TRAFFIC 12 MONTHS	EACH ROUND TYPE	APP DOWNLOADS LATEST (IOS)
HQ REGION	INDUSTRIES	EACH ROUND AMOUNT	APP DOWNLOADS 6 MONTHS (IOS)
HQ COUNTRY	SUB INDUSTRIES	EACH ROUND CURRENCY	APP DOWNLOADS 12 MONTHS (IOS)
HQ CITY	DELIVERY METHOD	EACH ROUND DATE	APP DOWNLOADS LATEST (ANDROID)
TOTAL FUNDING (EUR M)	GROWTH STAGE	TOTAL ROUNDS NUMBER	APP DOWNLOADS 6 MONTHS (ANDROID)
LAST ROUND	YEARLY GROWTH (SIMILARWEB)	PROFIT (2016, 2017, 2018, 2019)	APP DOWNLOADS 12 MONTHS (ANDROID)
LAST FUNDING	ALEXA GROWTH (ALL TIME)	EBITDA (2016, 2017, 2018, 2019)	SIMILARWEB RANK 3/6/12 MONTHS
LAST FUNDING DATE	EMPLOYEES	REVENUE (2016, 2017, 2018, 2019)	APP RANK 3/6/12 MONTHS
FIRST FUNDING DATE	EMPLOYEES (2016, 2017, 2018, 2019, 2020, 2021)	VALUATION	TECHNOLOGIES
SEED YEAR	EMPLOYEES IN HQ country (2016, 2017, 2018, 2019, 2020)	VALUATION CURRENCY	INCOME STREAMS

***Figure 2**. Available fields after data clearing*

Another problem with this dataset is that it lacks a criterion for whether a start-up made an attempt to get funding. Accordingly, there is no reliable information that a start-up that has not attracted external investment is unsuccessful. It is probable that the start-up did not require financial support from investors (Fig. 3), developing based on its own financing sources.

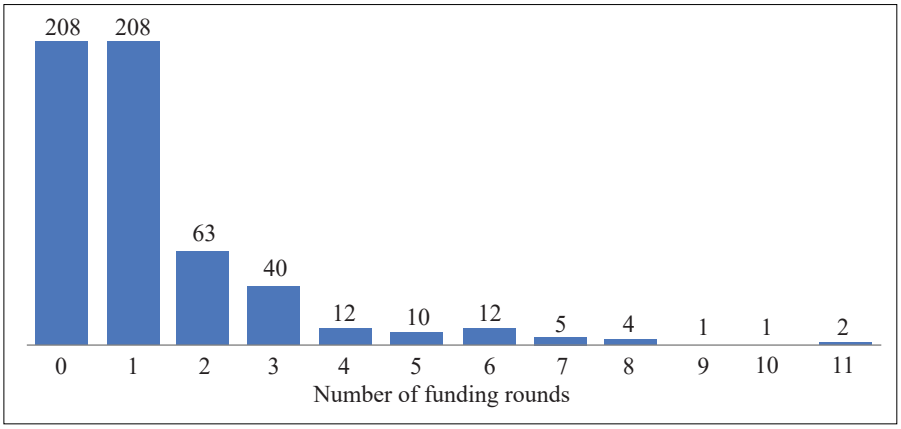

***Figure 3**. Distribution of start-ups according to the number of funding rounds*

Therefore, to avoid inaccuracies, start-ups for which information on funding was not reflected were removed from the data set. Thus, 358 start-ups were left in the data array.

The next problem is that data on Internet activity (SW TRAFFIC 12 MONTHS, APP DOWNLOADS 12 MONTHS (IOS), and TWITTER FAVORITES) are available either for the current date or for the last year. Accordingly, it is impossible to apply these factors for enterprises that had the last round before 2019 (except for the situation that it was a single round of funding). For the reliability of the conclusions, the sample was limited to start-ups that had the first funding in 2017-2019. After these manipulations, only 149 start-ups were left. The distribution by type of the last round of funding is presented in Table 1.

***Table 1**. Distribution of start-ups by type of last funding*

Type of last round	Number of start-ups
SEED	82
ANGEL	18
ACQUISITION	13
EARLY VC	10
GRANT	10
SERIES A	5
ICO	4
SERIES C	2
LATE VC	2
CONVERTIBLE	1
DEBT	1
GROWTH EQUITY	1

In Table 1, only 14 start-ups fit the classic definition of a successful start-up – ACQUISITION or GROWTH EQUITY, which is insufficient for building a model. Taking this into account, it was decided to limit the study solely to investment success, the criterion of which was recognised as obtaining repeated funding. Since part of the agreements on attracting investments in a start-up is confidential information and not all start-ups know the amount of preliminary investments raised, an increase in funding cannot be used as a success criterion (as suggested in [12]). A factor that can also affect the reliability of the obtained results is the presence of start-ups that received A+ funding or made an exit already in the first round. It is unknown what influenced the decision of investors in this situation, so the data on such start-ups were also removed from the data set.

As a result, 123 start-ups were left, which were transformed according to the binary method (1 – a successful start-up (fact of repeated investment), 0 – unsuccessful (there is no fact of repeated financing). The number and ratio of successful and unsuccessful start-ups is presented in Figure 4.

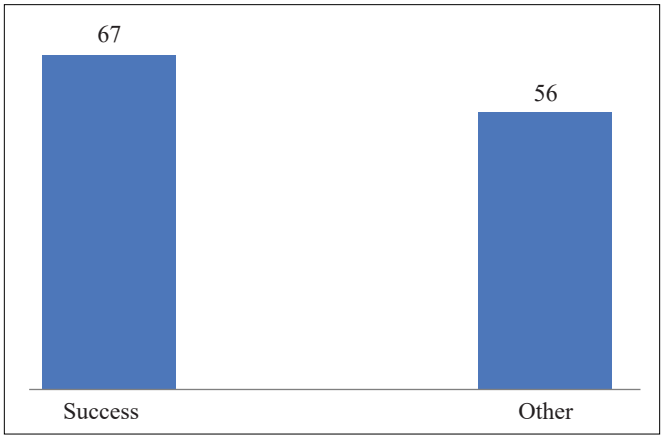

***Figure 4**. Number of successful and unsuccessful start-ups in the dataset*

The next step was to create new variables from the available ones. Thus, for example, a new variable was created by subtracting LAUNCH DATE and the year of the first investment in the EACH ROUND DATE field (Time_to_first_funding). To create First_round_type variable, the first value was taken from the EACH ROUND TYPE. From the EACH ROUND AMOUNT field, the amount of the first investment was allocated – First_round_amount.

For an integrated assessment of the volume of traffic to the site (SW_traffic), the value SW TRAFFIC 12 MONTHS was taken as a basis. To estimate the number of downloads (APP_downloads) – the sum of APP DOWNLOADS 12 MONTHS (IOS) and APP DOWNLOADS 12 MONTHS (ANDROID).

The presence of a start-up in the top ratings of Similar Web was evaluated by creating a new variable TOP_Rank_SW, where the presence in the chart of any rating SIMILAR WEB RANK 3/6/12 MONTHS or APP RANK 3/6/12 MONTHS was indicated under the "TOP" value.

The next step is to verify factors gaps. If a field value is missing for most start-ups – it is removed. If only a part is missing – the value was supplemented manually. Qualitative information was received from the start-up website and open sources, and quantitative information was filled in by the mean method to replace the missing values. For the missing values, based on the filled ones, was assigned the average for each ratio B2B_B2C, INDUSTRIES, First_round_type, INCOME STREAMS, REVENUE_MODEL, and TECHNOLOGIES. Since the algorithm for developing a machine learning model assumes dividing the data into a test and training set (25%/75%), and training set is divided into 5 equal groups for cross-verification, the quantity of each of the qualitative values must be greater than 6 (5 per each cross-verification and one per training set). Therefore, it is necessary to group the values of the model parameters according to the given threshold value.

The grouping results are as follows:

– HQ_COUNTRY – turned into a binary variable. Characterises the presence of Head office in Ukraine or another country.

– Time_to_first_funding – now characterises the time to the first investment in the context of three groups: receiving funding in the year of opening, for the next year, for over 1 year.

– First_round_type – the variable is grouped into two categories SEED_EARLY_VC and ANGEL_GRANT.

– INDUSTRIES – divided into 4 categories: Heals_Food_Education, Enterprise_Software_Manufacturing, Creative_Services, Home_Travel_Hobbies.

– TECHNOLOGIES – grouped into 4 categories: App_programs_development, New_product_development, Online_services, and VR_AR_ML_AI.

As a result, there were 14 fields that characterise a start-up. The description of fields is presented in Table 2.

***Table 2**. Characteristics of model variables*

Variable name	Description
Success	Dependent variable characterising the factor of start-up success (Yes/No)
HQ_COUNTRY	Country of the head office location
Time_to_first_funding	Time to the first investment
B2B_B2C	Client type (business or individual)
INDUSTRIES	Area of activity
YEARLY_GROWTH_SIMILARWEB	The annual increase in digital activity in the SW rating
SW_TRAFFIC	The number of visits to the site on average per year
First_round_type	Type of the first funding round
First_ROUND_AMOUNT	The first funding round amount
APP_DOWNLOADS	The average number of app downloads over the last year
TOP_Rank_SW	Presence in the top ranking by digital activity in their area of activity
INCOME_STREAMS	Income channel
REVENUE_MODEL	Revenue model
TECHNOLOGIES	What technologies/innovations have been proposed

This means that if there is an x number of factor levels, x-1 dummy variables will be created and all but the first-factor level are converted to new columns. All numerical factors have been normalised. To perform this operation, the R programming language and the Rstudio IDE were used. Tidymodels package was chosen as the main software package for tuning machine learning algorithms.

Methodology and technology of start-up success prediction

During the study, 3 classification models will be applied – decision tree, random forest, and logistic regression. A decision tree develops solutions using a tree model. The algorithm splits the sample into two or more homogeneous sets (branches) based on the most significant differentiators of the input variables. To select a differentiator (predictor), the algorithm takes into account all the features and produces a binary partition. Then it chooses that option with the least cost (i.e., high precision) and repeats recursively until the data is successfully split across all branches (or reaches the maximum depth).

A random forest is an ensemble model that builds multiple trees and classifies features based on a "vote". The object belongs to the class that has the majority of votes from all trees. The algorithm trains several decision trees on different datasets and uses the mean to improve the forecasting accuracy of the model. Logistic regression is a technique for modelling the probability of an event. Like linear regression, it helps understand the relationship between one or more variables and a target variable, except that in this case, the target variable is binary: its value is 0 or 1.

These machine learning methods were chosen proceeding from the results of the analysis of existing studies and because of their ease in development and interpretation.

The modelling procedure requires the division of our sample into training and test. The distribution takes place in the proportion of 75% – training, 25% – test. The first set was used to tune the parameters of the model. The second was to test the algorithm against previously unseen values.

The next step is to set up the cross-verification procedure. Cross-verification is a statistical technique used to evaluate the predictive capabilities of machine learning models. It is commonly used in applied learning to compare and select models for a specific predictive modelling problem.

The general methodology is as follows:

1. The dataset in the training sample is shuffled randomly.
2. The sample is divided into k groups.
3. Each group is divided into two samples for training and testing.
4. A test sample is taken from one group, and all the rest serve as a training sample. This is done k-times.
5. The model learns on the training set and evaluated on the test.
6. The results of constructing k-models are averaged.

Usually, 10-fold cross-validation is used. But due to the limited data, a 5-fold check was used. Each algorithm is cross-validated, and the best model was chosen as the basis for prediction.

In machine learning, measuring accuracy is an important task. Therefore, when it comes to classification, can be relied upon the AUC (Area Under the Curve) – ROC (Receiver Operating Characteristics) curve. It is one of the most important metrics for testing the effectiveness of any classification model. The AUC-ROC curve is a performance measurement for classification tasks at various threshold settings. ROC is the probability curve and AUC is the degree or measure of the distribution.

It describes how the model can distinguish between classes. The higher the AUC, the better the model predicts 0 as 0 and 1 as 1.

The ROC curve is plotted with TPR (True Positive Rate) to FPR (False Positive Rate), where TPR is on the Y-axis and FPR is on the X-axis.

TPR (True Positive Rate)/Recall/Sensitivity – the share of correctly distributed positive values among all positive values is determined as follows:

$$TPR = \frac{TP}{TP + FN} \quad (1)$$

where TP is the true positive classification, FN is the false negative classification.

Specificity – the proportion of correctly distributed negative values to all negative values is calculated according to the formula:

$$Spec = \frac{TN}{TN + FP} \quad (2)$$

where TN is the true negative classification, FP is the false positive classification.

FPR (False Positive Rate) – the proportion of incorrectly distributed negative values to all negative values is calculated according to the formula:

$$FPR = 1 - Spec = \frac{FP}{TN + FP} \quad (3)$$

where TN is the true negative classification, FP is the false positive classification.

An excellent model has an AUC close to 1, which means that it divides classes well. A bad model has an AUC close to 0, which means the worst allocation score. This effectively means that the algorithm matches the reverse values. It predicts 0 as 1 and 1 as 0. And when the AUC is 0.5, it means the model has no class separation ability at all.

Another classical method for evaluating the quality of a classification model is Accuracy, which is calculated according to the formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

where TN is the true negative classification, FP is the false positive classification, TP is the true positive classification, FN is the false negative classification.

The last metric in our study will be the weighted accuracy estimation (F-score), which is calculated according to the formula:

$$F = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \quad (5)$$

where FP is the false positive classification, TP is the true positive classification, FN is the false negative classification.

A feature of the machine learning algorithm is that they are based on numerous hyperparameters. These hyperparameters have their default values, which are applied if no adjustments are made. Obviously, they do not always provide the most accurate forecasts. Therefore, the optimal hyperparameters were selected for each model. And for each hyperparameter, a 5-fold cross-verification is done to find their most optimal ratio.

– Decision Tree model hyperparameters:

cost_complexity: The complexity of the model

tree_depth: The maximum depth of the tree

min_n: The minimum number of points in a node that the node will require for further separation.

– Random Forest model hyperparameters:

mtry: The number of variables (predictors) that will be randomly selected at each division when creating models.

trees: The number of trees contained in the ensemble.

min_n: The minimum number of points in a node that the node will require for further separation.

– Logistic regression model hyperparameters

penalty: The total amount of regularisation in the model.

mixture: A mixture of different types of regularisation.

The method of random search of parameters was used to select the hyperparameters. Since there are an infinite number of variants of parameter compounds, a simpler method of parameter determination is required. One option is a random search. Where the "n" is a number of randomly selected values of compounds of hyperparameters. Next, a model is studied for each compound and 5-fold cross-verification is performed. In this study, 25 random parameter compounds for each model will be calculated.

RESULTS AND DISCUSSION

As already noted, the first stage, common to all models, is the division of the initial sample into test and training samples. 93 start-ups were included in the training sample, 30 in the test sample.

Decision tree implementation

After the implementation of the Decision Tree algorithm, Accuracy was set at 0.633, Roc_Auc – 0.665. The evaluation was based on a test sample.

But one iteration of the evaluation is not enough to assert with complete confidence the effectiveness of the model. Therefore, to finalise the results, it was necessary to cross-validate the model's performance. The cross-verification results are presented in Table 3.

***Table 3**. Decision Tree cross-verification results*

Metric	Min	Median	Mean	Max
Roc_Auc	0.444	0.581	0.550	0.644
Accuracy	0.444	0.632	0.590	0.722
Sensitivity	0.333	0.625	0.572	0.778
Specificity	0.3	0.5	0.602	0.909
F-score	0.462	0.5	0.553	0.667

To improve the forecasting efficiency, the model hyperparameters tuning procedure was carried out. Table 4 shows the hyperparameters that will provide the model with the largest minimum values for Roc_Auc and Accuracy. Table 5 demonstrates the forecasting results with the new hyperparameters considered. Figure 5 demonstrates the Roc curve.

***Table 4**. Optimal hyperparameter values*

Hyperparameter	Default value	Optimal value
cost_complexity	0.01	9.45078958391999e-08
tree_depth	30	14
min_n	20	33

***Table 5**. Decision Tree cross-verification results after customisation*

Metric	Min	Median	Mean	Max
Roc_Auc	0.512	0.581	0.576	0.644
Accuracy	0.5	0.632	0.612	0.722
Sensitivity	0.125	0.625	0.547	0.875
Specificity	0.3	0.8	0.662	0.909
F-score	0.182	0.636	0.523	0.667

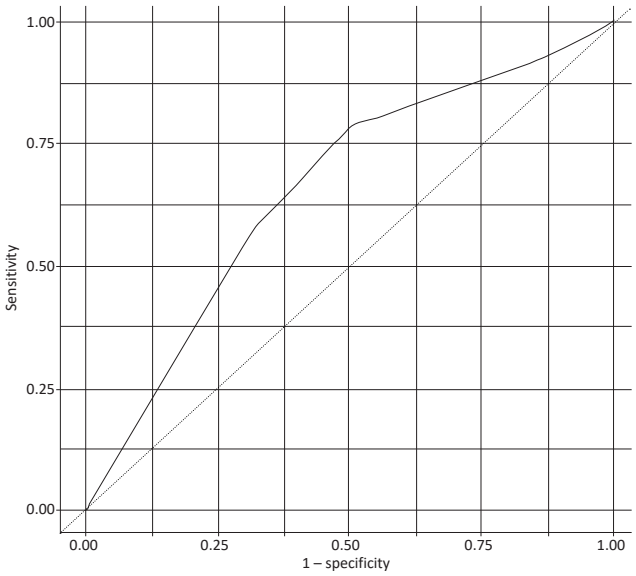

***Figure 5**. Roc curve for decision tree*

Logistic regression implementation

After the implementation of the Logistic Regression algorithm, Accuracy was set at 0.467, Roc_Auc – 0.473. The evaluation was based on a test sample. Logistic Regression cross-verification results are presented in Table 6.

***Table 6**. Logistic Regression cross-validation results*

Metric	Min	Median	Mean	Max
Roc_Auc	0.462	0.495	0.544	0.689
Accuracy	0.389	0.55	0.504	0.579
Sensitivity	0.333	0.375	0.45	0.667
Specificity	0.3	0.5	0.545	0.727
F-score	0.375	0.421	0.445	0.6

Table 7 demonstrates the hyperparameters that provide the model with the largest minimum values for Roc_Auc and Accuracy. Table 8 shows the forecasting results with the new hyperparameters considered. Figure 6 shows the Roc curve.

***Table 7**. Optimal hyperparameter values*

Hyperparameter	Default value	Optimal value
penalty	0.1	0.0621016941891562
mixture	1	1

***Table 8**. Logistic Regression cross-verification results after customisation*

Metric	Min	Median	Mean	Max
Roc_Auc	0.575	0.606	0.645	0.733
Accuracy	0.5	0.611	0.600	0.684
Sensitivity	0.125	0.5	0.45	0.625
Specificity	0.5	0.8	0.724	0.818
F-score	0.182	0.533	0.486	0.625

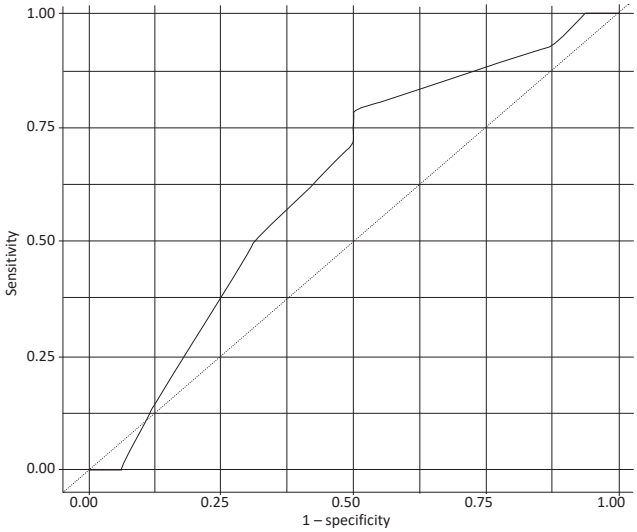

***Figure 6**. Roc curve for logistic regression*

Random forest implementation

After the implementation of the Random Forest algorithm, Accuracy was set at 0.467, Roc_Auc – 0.5. The evaluation was based on a test sample. Random Forest cross-verification results are presented in Table 9.

The hyperparameter optimisation process did not give significant results, so conclusions were based on their default values. Figure 7 shows the Roc curve.

Comparison of models

To compare the models, the average values of the characteristics of forecasting efficiency were calculated (Table 10).

***Table 9**. Random Forest cross-verification results*

Metric	Min	Median	Mean	Max
Roc_Auc	0.562	0.626	0.645	0.8
Accuracy	0.5	0.556	0.570	0.684
Sensitivity	0.111	0.5	0.358	0.556
Specificity	0.5	0.8	0.744	0.9
F-score	0.167	0.471	0.399	0.625

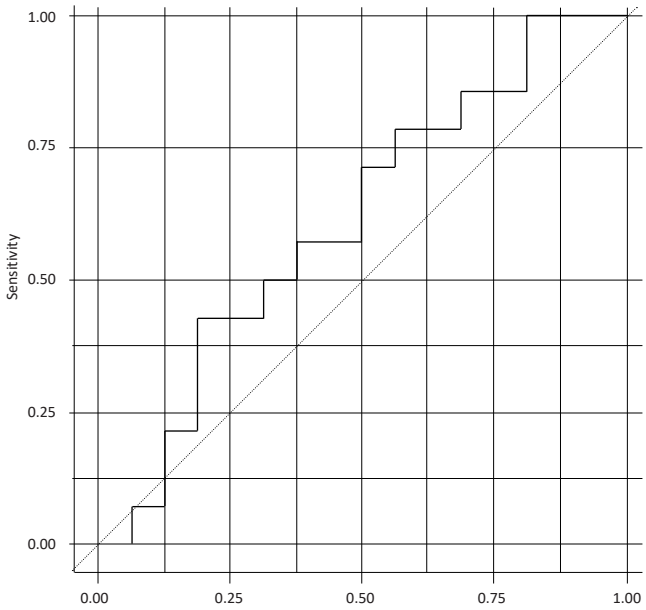

***Figure 7**. Roc curve for random forest*

***Table 10**. Comparative table of the model's effectiveness*

Metric	Decision tree	Logistic regression	Random forest
Roc_Auc	0.576	0.645	0.645
Accuracy	0.612	0.600	0.570
Sensitivity	0.547	0.45	0.358
Specificity	0.662	0.724	0.744
F-score	0.523	0.486	0.399

Discussion of machine learning modelling results

This study provides a complete guide to implementing a start-up success prediction model. In the beginning, 566 start-ups were presented in the data set, which were characterised by 98 variables. Since this is an open-access database, it contains certain inaccuracies, redundant and contradictory information (Fig. 1). All parameters that, according to the authors, do not characterise a start-up, but are only technical fields were ignored (Fig. 2).

One of the main tasks was to define a clear concept of "start-up success". After doing an exploratory analysis of the data, it was concluded that, based on the sample under study, the object of research could be a financial success – successful start-ups are those that made Acquisition, while GROWTH EQUITY or investment success are start-ups with more than one round of funding (Table 1). The first criterion was rejected due to the extremely small number of start-ups that meet it. The model was further tuned for start-ups that had at least one round of funding for the current period (Fig. 3). The final sample included 123 start-ups (Fig. 4).

Since the algorithm for developing a machine learning model assumes dividing the data set into a test and training set, and the training set into 5 equal groups for cross-verification, the quantity of each of the qualitative values must be greater than 6 (5 for each cross-check and one for the training sample). Therefore, it was necessary to group the values of the model parameters according to the given threshold value. If a field value was missing for most start-ups – it was removed. If only a part was missing – the value was manually supplemented. Qualitative information was taken from the start-up website and open sources, and quantitative information was filled in by the mean method of missing values replacement. As a result, 14 fields characterising a start-up were formed (Table 2). In the final, for better operation of machine learning algorithms, all qualitative variables, except for the dependent one, were converted to binary form.

All stages of the implementation of the start-up success prediction model were carried out, namely:

- selection of machine learning algorithms;
- the division of the sample into test and training samples;
- tuning model hyperparameters;
- selection of metrics for evaluating the model quality;
- implementation of algorithms;
- cross-verification of all forecasting accuracy metrics;
- comparative analysis of forecasting efficiency.

Thus, according to the developed methodology and the proposed algorithm, three models for predicting

the success of a start-up were implemented: Decision Tree, Logistic Regression, and a Random Forest.

During studying the results of forecasting by the Decision Tree method the calculated results according to the ROC-AUC metric were set at 66.5%. This level means that the model allows to predict the success of a start-up more likely than it would be done randomly. After cross-validation, lower model performance results were obtained. The minimum values for Roc_Auc and Accuracy are of particular concern. They are less than the threshold value for the effective model – 0.5 (Table 3). To improve the forecasting accuracy, the model hyperparameters were adjusted (Table 4). This allowed increasing the minimum values of the forecasting accuracy using the Roc_Auc and Accuracy metrics to 0.512 and 0.5, respectively (Table 5). In Figure 5 the ROC curve is above the dashed line, which means that model is more efficient than random distribution into classes.

As Table 6 demonstrates, the logistic regression model yielded unacceptable results to claim that this model is effective – it did not pass the 0.5 thresholds. The next step was to improve the prediction results by adjusting the model hyperparameters. Table 7 demonstrates the model hyperparameters that gave us the highest minimum values for Roc_Auc and Accuracy. Table 8 presents the forecasting results considering the new hyperparameters: Roc_Auc – 0.575 and Accuracy – 0.5. Figure 6 the ROC curve is above the dashed line, which means that the model presented in this study is more efficient than random distribution into classes.

Evidently, the minimum values of the Roc_Auc and Accuracy parameters after cross-verification of the random forest model were established at the level of more than 0.5, which indicates the acceptability of the model for decision-making. The average values were established at the level of 0.645 and 0.57, respectively (Table 9). The hyperparameter optimisation process did not give significant results, so conclusions were made based on their default values. In Figure 7, the ROC curve is above the dotted line, which means that model is more efficient than random distribution into classes.

To compare the models, the average values of the characteristics of forecasting efficiency were calculated (Table 10). Since, for all models, the highest accuracy rates are observed in terms of Specificity, which means that models are better at recognising start-up failure than its success. The most descriptive indicators are Sensitivity and Accuracy because they are focused on measuring the accuracy of predicting the success of start-ups. Therefore, even if the weighted accuracy indicator Roc_Auc in the logistic regression (0.645) and random forest (0.645) is higher, but due to a more accurate prediction of negative factors, the decision tree will be considered as the most effective model (Sensitivity – 0.547, Accuracy – 0.612).

As a result of the work performed, the key factors influencing various manifestations of a start-up success are defined. In addition, the study revealed a methodology for cleaning and preparing data for modelling using machine learning methods. For the first time, this paper describes an algorithm for working with missing data by manually adding information from open sources and according to the method of mean missing values replacement.

The efficiency of the method proposed can be confirmed by the effectiveness of forecasting based on generally accepted assessment metrics and performance thresholds. To make the research more practical, it relied on publicly available data. This is a considerable advantage over the studies [4], [5] due to the simplicity of implementation and the possibility of repeating the modelling processes. Achievement of high-quality results of predicting success, based on a much smaller list of start-ups and their characteristics, makes the study more universal for application in emerging markets than studies [14] and [15], which were based on start-ups from all over the world, mainly from the USA and Europe.

In this study, much more attention is paid to methods of improving forecasting efficiency than in similar studies [9; 10]. The process of tuning the model hyperparameters was implemented, conducted multiple cross-validations against five different metrics for evaluating the quality of the models. This allowed investigating the model from different angles of its efficiency as opposed to only one – the number of positively predicted observations, as in the study [11].

Upon critically evaluating the obtained results, it is advisable to pay attention to certain limitations that are inherent in them as follows:

1. The results of the practical implementation of the models are very strongly influenced by the country's economic environment. In some countries with a better investment climate, start-ups can achieve better results even with worse characteristics than in the models provided in this study. While in countries with the worst estimates of the investment climate, even promising start-ups can fail due to the lack of infrastructure to support their development.

2. The disadvantage of the proposed methodology is that the success of a start-up is highly dependent on the field of start-ups activity (specialisation). Since every day new trends are formed, new technologies appear, new inventions are commercialised, some start-ups and areas that are relevant and promising now, over time, will no longer be of interest to investors.

3. The proposed algorithms are incapable of predicting the success of a start-up in the early stages. To forecast success, one needs to wait for at least some results of activity, and only then predict. For many real-life investors, this may be too long.

CONCLUSIONS

Predicting start-up success is a challenging task that is critical for many stakeholders making decisions about their start-up investments. The decision support system for the investment object can be useful for entrepreneurs, venture analysts, or politicians who can use the built models to predict the success of a start-up, using such start-up characteristics as the economic sector and business model, basic technologies, and indicators of digital activity, time to first funding and its amount. This forecast, in turn, can be used to drive better investment decisions and develop relevant economic policies to improve the overall start-up ecosystem.

The machine learning algorithms used are based on data on the receipt of investments by these start-ups in 2017-2021. Given the limited number of start-ups, it was decided to accept the fact of re-investment (investment success) as a start-up's success. 3 machine learning models, that allow predicting the probability of success of start-ups, were built: Decision Tree, Logistic Regression, and a Random Forest. Evaluation of the accuracy of the developed models was a key task of the work. The effectiveness of the algorithms was tested according to five indicators: Accuracy, AUC, Sensitivity, Specificity, and F-score. Compared to the initial experiments, the accuracy of start-up success prediction has been successfully improved by adjusting the model hyperparameters. The built models are qualitative without signs of overfitting, which can be seen by evaluating the results of cross-verification.

The results were cross-verified using metrics such as AUC (Area Under the Curve), Sensitivity (True Positive Rate), Specificity (True Negative Rate), Accuracy, Recall, Precision, etc. The value of these metrics allows asserting that predictions made based on developed models are better than based on our judgments or randomness. The decision tree model demonstrated the highest Accuracy, Sensitivity, and F-scores for the tested algorithms, averaging 61%, 55%, and 52%, respectively. This allows recommending the use of a decision tree algorithm to predict the start-up success. The AUC score for the decision tree settled at 58%, which is lower than the logistic regression and random forest result (65%). That is, the last two models allow for better prediction of the start-up's failure which is also extremely useful proceeding from the interests of practical usage. All models have an acceptable level of AUC classification accuracy to confidently confirm their efficiency.

Further research may lie in scaling the proposed approaches to other markets or groups of regions at the same stage of economic development. It is necessary to continue the search for other algorithms that would allow for a better understanding of the correlation between different types of start-up activity and different manifestations (types) of its success.

REFERENCES

[1] VISION. (2020). Retrieved from https://europeanstart-upnetwork.eu/vision.

[2] 4 ways governments can support start-ups and save their economies. (2020). Retrieved from https://www.weforum.org/agenda/2020/06/4-ways-governments-can-support-start-ups-and-save-their-economies.

[3] Haddad, H., Weking, J., Hermes, S., & Böhm, M. (2020). Business model choice matters: How business models impact different performance measures of start-ups. *Proceedings of the 15th international conference on business information systems 2020 "Developments, opportunities and challenges of digitisation", WIRTSCHAFTSINFORMATIK 2020*. doi: 10.30844/wi_2020_h4.

[4] Aminova, M., & Marchi, E. (2021). The role of innovation on start-up failure vs. its success. *International Journal of Business Ethics and Governance*, 4(1), 41-72. doi: 10.51325/ijbeg.v4i1.60.

[5] Weking, J., Böttcher, T., Hermes, S., & Hein, A. (2019). Does business model matter for start-up success? A quantitative analysis. In *27th European Conference on Information Systems (ECIS)*. Sweden: Stockholm & Uppsala. Retrieved from https://aisel.aisnet.org/ecis2019_rip/77.

[6] Chen, Y., Tsai, C., & Liu, H. (2019). Applying the AHP model to explore key success factors for high-tech start-ups entering international markets. *International Journal of E-Adoption*, 11(1), 45-63. doi: 10.4018/ijea.2019010104.

[7] Gloor, P., Fronzetti Colladon, A., Grippa, F., Hadley, B., & Woerner, S. (2020). The impact of social media presence and board member composition on new venture success: Evidences from VC-backed U.S. start-ups. *Technological Forecasting and Social Change*, 157, article number 120098. doi: 10.1016/j.techfore.2020.120098.

[8] Skawińska, E., & Zalewski, R. (2020). Success factors of start-ups in the EU—A comparative study. *Sustainability*, 12(19), article number 8200. doi: 10.3390/su12198200.

[9] He, S., & Yu, C. (2020). Analysis of the crucial success factors for tech-based start-ups. *International Journal of Innovation in Management*, 8(1), 17-22.

[10] Kaiser, U., & Kuhn, J. (2020). The value of publicly available, textual and non-textual information for start-up performance prediction. *Journal of Business Venturing Insights*, 14, article number e00179. doi: 10.2139/ssrn.3570379.

[11] Krishna, A., Agrawal, A., & Choudhary, A. (2016). Predicting the outcome of start-ups: Less failure, more success. In *16th International Conference on Data Mining Workshops (ICDMW)* (pp. 798-805). Barcelona: IEEE.

[12] Ang, Y., Chia, A., & Saghafian, S. (2020). Using machine learning to demystify start-ups funding, post-money valuation, and success. *Systems*, 9(3), article number 55. doi: 10.2139/ssrn.3681682.
[13] Żbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management*, 58(4), article number 102555. doi: 10.1016/j.ipm.2021.102555.
[14] Ünal, C., & Ceasu, I. (2019). *A machine learning approach towards start-up success prediction. International research training group 1792 "High dimensional nonstationary time series", Humboldtuniversität Zu Berlin (IRTG 1792 Discussion Paper, No. 2019-022)*. Retrieved from https://www.econstor.eu/handle/10419/230798.
[15] Antretter, T., Blohm, I., & Grichnik, D. (2018). Predicting start-up survival from digital traces: Towards a procedure for early stage investors. In *International Conference on Information Systems (ICIS)*. San Francisco. Retrieved from https://www.alexandria.unisg.ch/publications/255532.
[16] Official website of the Dealroom Ukraine. (n.d.). Retrieved from https://ukraine.dealroom.co/.

Застосування методів машинного навчання для прогнозування успіху стартапу

Олена Валеріївна Піскунова, Лариса Олександрівна Лігоненко, Ростислав Сергійович Клочко, Тетяна Олександрівна Фролова, Тетяна Олександрівна Білик

Київський національний економічний університет імені Вадима Гетьмана
03057, просп. Перемоги, 54/1, м. Київ, Україна

Анотація. Прогнозування успіху новоствореного підприємства завжди було актуальною задачею як для інвесторів, так і для дослідників. У теперішній час вона набула ще більшою актуальності стосовно стартапів – молодих інноваційних технологічних підприємств, спрямованих на масштабування свого бізнесу. Дане дослідження спрямоване на створення моделі прогнозування успіху стартапу на основі його описових характеристик. Використовуючи дані платформи Dealroom стосовно статистики фінансування стартапів та їх опису, розроблена модель, яка пов'язує такі характеристики стартапу як: період від заснування до отримання першого фінансування, сфера діяльності, тип і сума першого раунду фінансування, бізнес-модель і застосовані технології із інвестиційним успіхом стартапу, під яким розуміється отримання повторного фінансування. У фінальну вибірку увійшло 123 стартапи, які засновані або ведуть свою діяльність в Україні. Порівняно три алгоритми машинного навчання – логістичну регресію, дерево рішень і випадковий ліс. Незважаючи на обмеженість доступних даних, отримані прийнятні результати з точки зору Accuracy, Sensitivity та F-score. Найкращою моделлю з точки зору передбачення успіху стартапу визначено – дерево рішень, із показниками середньої точності 61 %, 55 %, і 52 % відповідно. Оцінка AUC для дерева рішень встановилася на рівні 58 %, що нижче показників логістичної регресії та випадкового лісу (65 %), але останні моделі досягли таких високих результатів за рахунок кращого передбачення провалів стартапів, у той час коли ж більш практично-значущим є можливість передбачення їх успіху. Всі моделі показали прийнятний рівень точності класифікації AUC, що з впевненістю дозволяє стверджувати про їх ефективність. Система підтримки прийняття рішень стосовно об'єкту інвестування може бути корисною для підприємців, венчурних аналітиків або політиків, які можуть використати побудовані моделі для прогнозування успіху стартапу. Цей прогноз, зі свого боку, може використовуватися для прийняття більш ефективних інвестиційних рішень і розробки релевантної економічної політики, спрямованої на поліпшення загальної екосистеми стартапів

Ключові слова: стартап-екосистема, стартап, інновації, система підтримки прийняття рішень, класифікація, моделювання даних, передбачення успіху стартапу