International journal of innovation in Engineering, Vol 2, No 1, (2022), 78-109

International journal of innovation in Engineering

journal homepage: www.ijie.ir

Research Paper

Forecasting Startup Return using Artificial Intelligence Methods and Econometric Models and Portfolio Optimization Using VaR and C-VaR

Milad Shahvaroughi Farahani[a1], **Amirhossein Esfahani**[b]

[a] *Department of Finance, Faculty of Finance, Khatam University, Tehran, Iran; m.shahvaroughi@khatam.ac.ir*

[b] *Department of Accounting, Eslamshahr University, Tehran, Iran; esfaaahani@gmail.com*

ARTICLE INFO

Received: 15 January 2022

Reviewed: 10 Feburary 2022

Revised: 12 February 2022

Accepted: 12 February 2022

Keywords:

Artificial Neural Network (ANN), Genetic Algorithm (GA), Econometric Models, Startup valuation, Value at Risk and Conditional Value at Risk (VaR & C-VaR).

ABSTRACT

In this paper, we have tried to study the main role of startups in economy, their characteristics, main goals and etc. The main goal of article is prediction of startup's return using artificial intelligence methods such as genetic algorithm (GA) and artificial neural network (ANN). Some global indices such as S&P500, DJAI, and economic indicators such as 10 years Treasury yield, Wilshire 5000 Total Market Full Cap Index along with some other special indicators in startups like team, idea, timing and etc. are used as input variables. GA is used as feature selection and finding the most important variables. ANN is used as an optimization model and prediction of startup's returns. We used econometric models such as regression analysis. We have estimated Value at risk (VaR) and Conditional Value at risk (C-VAR) for considered portfolios including three startups (public company) such as Dropbox, Inc. (DBX), Scout24 SE (G24.DE) and TIE.AS and optimal portfolio formation. The results show that AI based methods are more powerful in prediction of startup's return. On the other hand, VaR and C-VaR models are very beneficial approach in minimizing risk and maximizing return.

[1] **Corresponding Author**
m.shahvaroughi@khatam.ac.ir

1. Introduction

Startups are mainly young companies or teams that have new ideas (Salamzadeh, A., & Kawamorita Kesim, H. 2015). Their main characteristic is innovation. They seek to provide products or services that make the life easier for applicants and customers. So, they try to make a fundamental change in production and services. This is one of the main reasons which startups called "disruptive" (Hyrkäs, A. 2016). Most of the time, they create products or services which do not have any damage to the environment. At first, because they are not known, they may have a small income stream or lack of cash flow generation. So, they need cash. They can raise capital by parents and family or they can refer to rich investors which called angel investors or they can visit venture capitalists (Schückes, M., & Gutmann, T. 2021). Venture capital is a type of private equity financing for startups. Venture capitalists analyze startups based on different parameters such as the rate of growth, income stream or cost structure, future perspective or etc. In addition to financial analysis, other criteria such as idea, team, timing, etc. are important. From financial point of view means financial analysis, a startup may not be acceptable but they may have a good team, idea and etc. that convince investors.

Startups have some characteristics and traits which presented in Figure (1) (https://www.fundzlab.com).

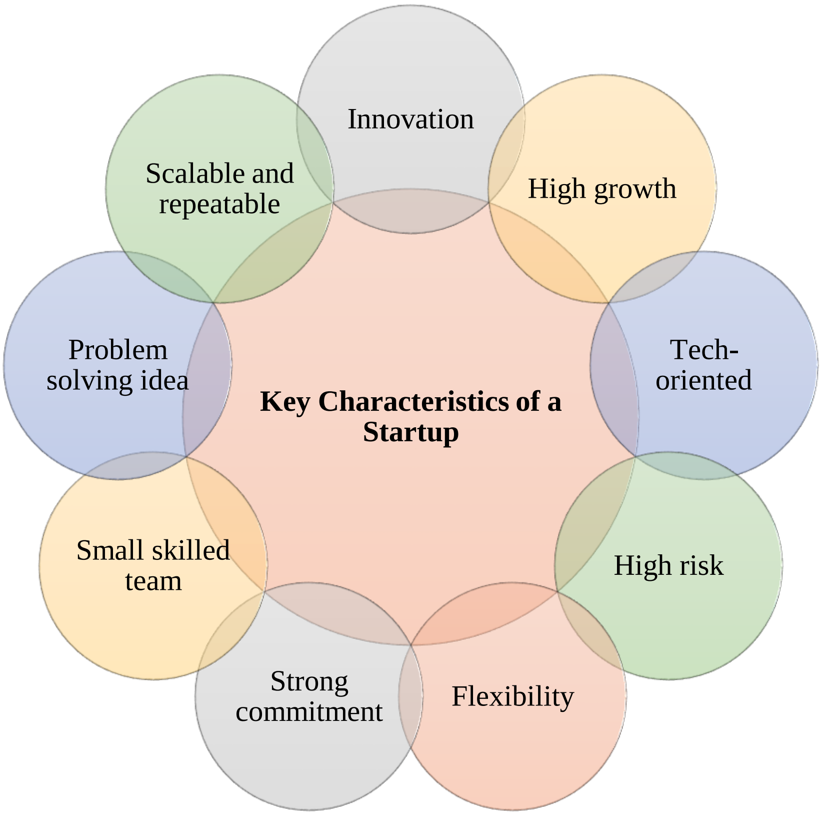


Fig. 1. The main characteristics of Startups

Team is very important in startups. Each one has a special skill and they have a deep commitment to the company. They may face a lot of problems which are risky such as financing, decreased sale and profitability. So, they should be flexible. They use technology to communicate with customers and promote products and services. Due to innovation and creativity and new idea, they have a high growth rate.

Startups play a significant role in the economy for different reasons. They maybe small but they can create jobs, increase productivity and boost the economy (Bjørnskov, C., & Foss, N. J. 2016). In startups, Growth rate is much more than other firms. There is a type of competition between startups that benefits both the companies and the customers, and the economy as a whole. Companies are always trying to decrease their prices and increase their quality in order to sell more. As a result, they can create the conditions to increase the level of quality. On the other hand, they can do different functions and take steps to create value such as building a brand, delivering excellent service or produce goods and services with special features which can lead to higher prices.

Sometimes startups may run out of cash to continue operating or develop their businesses. So, they need to be valued by investors. There are different methods for startups valuation. The main startup valuation models are Berkus model, Scorecard Valuation Method, Risk Factor Summation Method and etc. (Akkaya, M. 2020). Each model has its own limitations and deficits. Each valuation model should be compatible with the company's business models because as we mentioned earlier because each one has different assumptions.

One of the most common and notable prediction methods are artificial intelligence based methods. AI based methods include different sub-branches such as soft-computing, machine learning (ML), deep learning (DL), and etc. These methods have some characteristics that differentiate them from others. They have some features which are considerable:

I. High calculation capacity II. Speed up calculations III. Compatible with complex data structure IV. Automation of repetitive tasks V. More efficient process VI. Error reduction and etc. (Wang, M. H. 2017).

Unlike other models such as mathematical, statistical and econometric models, they do not need any pre-assumption or hypothesis. For example, in econometric models and regression analysis, you need to take a few steps such as examine stationarity, checking linearity and so on. But AI based methods just require and need data and are compatible with any type of data structure.

The rest of the paper is as follows: section2 is dedicated to literature review about different startup valuation models and their results. 3rd Section is about methodology. Section4 is findings and results and the final section is about conclusions and remarks.

2. Literature review

There are many articles about the definition of startups. Among different definitions, one of the most validated and common definitions which is adapted from Investopedia is that "A startup is a company that is in the initial stages of business". In 2019, Magalhães, R. P. C. surveyed a widespread literature about startup's definitions and research papers in his doctoral dissertation. He studies different papers with/without a startup definition per year from 2008 to 2018.

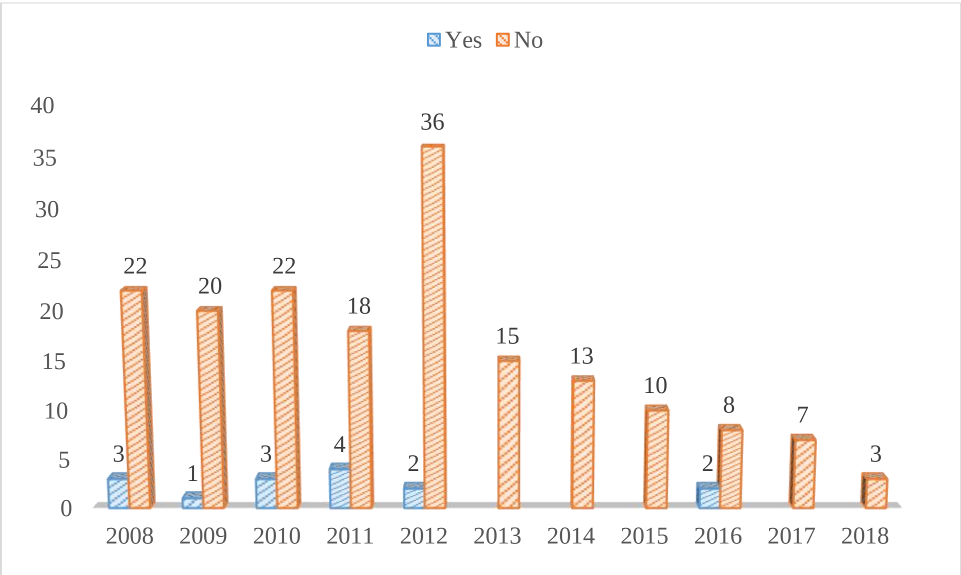


Fig. 2. Articles with/without a startup definition per year from 2008 to 2018

Out of 187 articles published in Asia and Europe, 15 articles have defined a startup that is approximately 8 percent. The results showed that these 15 papers have pointed out 4 important criteria in defining a startup. These are I. Innovation II. Lifecycle III. Investment IV. Economic impact. Although the other two criteria such as size and culture/mindset were considered, they had little effect on the definition of a startup.

Based on the mentioned features and steps, General Electric Company of Edison can be considered as one of the first startups (Edison, H. 2020). But by looking its current status, it is far from being a startup venture. Today, there are different tech startups like Facebook, Twitter, LinkedIn, Uber, Tesla, Dropbox, and Airbnb have raised billions of dollars in capital and revenue so far. Technology is the most important factor that has made this impressive growth and progress. They have some characteristics like excellent idea, timing, team, global business and etc. which differentiate them from traditional or ordinary companies or ventures.

When startups having lack of cash or liquidity, they may face to different problems such as Inability to establish a company, develop a product, marketing development and etc. One of the main ways to raise capital is to refer to an angel or venture capitalist (Hsu, D. K., et al. 2014). They analyze your financial position and financial statements, potential growth rate, team. Then, as an investment choice, decide whether to approve you or not. They may use different methods to evaluate startups. Different methods have been developed to evaluate different types of startups based on different stages of their life cycle.

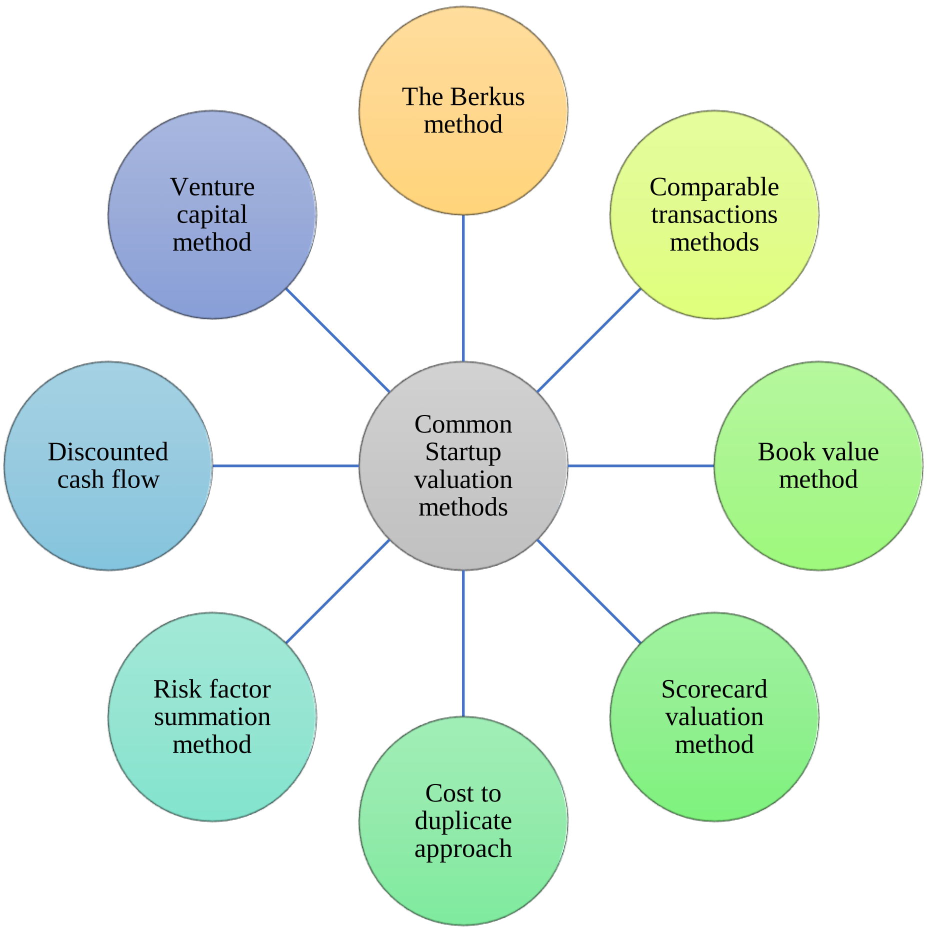


Fig. 3. Common startup valuation methods (de Oliveira, F. B., & Zotes, L. P. 2018)

In the following, recent articles on various startup valuation methods along with the results have been studied.

In a research paper, Jedlickova, M., & Kutnar, P. (2017) tried to create a fuzzy model which shows the promising results for the success prediction of hi-tech companies with a short history. They attempted to interconnect the economic and mathematical methods to produce a software for business success prediction. They concluded that their considered model is efficient in success prediction of hi-tech companies that have special qualifications such as very short history of existence and limited data to prove their overall health. Laitinen, E. K. (2019) surveyed and analyzed the impact of DCF model on startup financial success. He developed a simplified mathematical model based on revenue and expenditure and employing IRR as a measure of profitability and revenue generation as the measure of payback period. Experimental results were used as a sensitivity analysis means the relationship between DCF to the parameters of the model. The results showed that DCF model has better performance for startups that grow slowly and have a short payback period but that also exhibit a high IRR. Rahardjo, D., & Sugiarto, M. (2019) used a mixed real option method to valuation of Singapore and Indonesia startups. They believed that traditional valuation models such as Berkus comparison methods, risk factor summation and etc. cannot succeed in valuing companies that have lack of information or they are loss making companies. The results indicated that mixed real option method had better performance and had been compatible with digital startups.

Shariatpanahi, S. M., et al. (2020) believed that startup companies are better to use models which be suitable with their expectations. So, they used the real option method to value the startup companies. They used another model such as Skewness and Kurtosis Adjusted Black Scholes Model. The results showed that their models by considering the two factors means non-normal distribution of cash flows and agency costs is flexible in decision making and calculations. Montani, D., et al. (2020) reviewed different startup valuation methods to define future trend on this topic. They investigated traditional models such as the first Chicago method, Scorecard method and etc. They concluded that there is not the best and perfect model and each model has its own limitations. They considered three aspects: I. focus on future trends instead of past data. II. Considering different scenarios based on using probability and III. The unique business model of each startup and the inaccuracy of comparing competing business models. Dhochak, M., & Doliya, P. (2020) used a strategic approach to valuation of a startup. They wanted to examine the impacts of different strategic management on valuation of a new venture. So, they developed an integrative multi-criteria fuzzy decision making approach to measure the relative importance of the strategic input variables. The results validated the importance of strategic management and management theories on the valuation of new ventures. Shestakov, D. (2021) applied a testing method based on hypothesis for evaluation of startup projects. He evaluated the nature of startup risks projects based on five principal hypothesis. So, they hypothesis method could estimate risks and attractiveness of startups project easier than cash flow modeling. Hidayat, S. E., et al. (2021) tried to examine the main drivers of the company's value. They found that different factors such as financial and nonfinancial information can impact on startup valuation. Technologies like big data, clean tech and etc. can increase the value of startups and be considered as a kind of premium. Lavanchy, M., et al. (2022) tried to evaluate the most important factors in successful startup who have been successful in raising capital from Shark Tanks. They constructed a dataset to extract the unique natures. They found that entrepreneurs who offer less estimation of their companies are more successful.

3. Methodology

In this paper, AI based methods and econometric models are used as prediction models. At first, as mentioned earlier, we need to normalize data using the following equation:

$$\widetilde{S_i} = \frac{(S_i - S_{min})}{S_{max} - S_{min}} . i = 1 \dots . N \qquad (1)$$

In equation 1, numerator i is the number of data. S_i is each observation, S_{max} and $S_{min}$is the maximum and minimum observation in each indicator.

Some economic indicators and global stock indices are used as input variables as Table (1):

Table 1. Input variables

No	Name	Use
1	Open	Input
2	High	Input
3	Low	Input
4	Close	Input
5	^TNX	Input
6	CBBTCUSD	Input
7	DCOILBRENTEU	Input
8	DJIA	Input
9	DTWEXBGS	Input
10	Gold price	Input
11	NASDAQCOM	Input
12	SP500	Input
13	US dollar index	Input
14	WILL5000INDFC	Input
15	Funding	Input
16	Business model	Input
17	Idea	Input
18	Team	Input
19	Timing	Input
20	Return	Target

As you can see, there are some indicators like idea, team, timing and etc. which are related to the field of startups. They can increase predictability. According to the literature, each of the variables represents a percentage of the startup value. For example, funding, business model, idea, team, timing assign 14%, 24%, 28%, 32% and 42% of the startup value (https://www.forbes.com). We used daily data from 2019 to 2021 means three last years to prediction of three startup returns as Table (2):

Table 2. Statistical population

Startup Name (Symbol)	
1	Dropbox, Inc. (DBX)
Headquarter	San Francisco, California, US
Industry	Software-Infrastructure
Founder	Drew Houston, Arash Ferdowsi
Founded	2007, United States
Product & Services	designs and develops document management software
Valuation	9.33B$
Startup Name (Symbol)	
2	Scout24 SE (G24.DE)
Headquarter	Munich, Germany
Industry	Internet Content & Information
Founder	Beisheim Holding Schweiz AG
Founded	1998
Product & Services	a digital platform for the residential and commercial real estate sectors in Germany and internationally
Valuation	4.35B$
Startup Name (Symbol)	
3	TIE.AS
Headquarter	Netherlands
Industry	Software—Application
Founder (CEO)	Jan B. Sundelin
Founded	1987
Product & Services	It develops, distributes, and sells software solutions in the Netherlands, the United States, Germany, France, and internationally.
Valuation	35.59B$

Startups that are in pre-seed or seed stage, have shortage or lack of data. Because AI based and econometric models works with datasets, we have tried to choose startups that have some characteristics such as listed on stock exchange, headquarter in different countries, different activities etc.

3.1. Genetic Algorithm

GA is an evolutionary algorithm and it is based on the survival of superior members and Darwin's theory of evolution (Boudieb, D., et al. 2011). It is a process with different parameters such as initial population, crossover and mutation. Crossover which is called recombination, is used to generate new offspring by combining the genetic information of two parents. Mutation is an operator which used to keep genetic diversity of one generation and make it ready for next generation. Each one has an approximate value or size. For example, DeJong, K. (1975) suggested that an approximate value for crossover and mutation rate can be around the rate of 0.6 and 0.001 respectively. changing these two parameters can lead to different search space means a kind of exploration and exploitation. As you know, each algorithm begins with an initial population. Table (3) shows the GA parameters:

Table 3. GA parameters

Output Error	Output Activation Function	Input Activation Function	Mutation Rate	Crossover Rate	Number of Generation	Population size	Max Itr
MSE	Logistic	Logistic	0.1	0.9	50	20	1000
Selection parents			Mutation		Crossover		
Roulette wheel method			Binary Method		One-point method		

We used 70% of data as training and the remaining as validation and testing. We considered 0.01 as training rate which will decrease during time and repeating. chromosomes with 24 bits are used which 19 bits represents the selection or rejection of the variables and 5 other bits shows the number of neurons in hidden layer. To obtain better results, simulated annealing is used as an optimization method which can affect mutation operator. In GA, new solutions are called offspring which are the results of crossover of two parents. Figure4 shows the GA process as feature section.

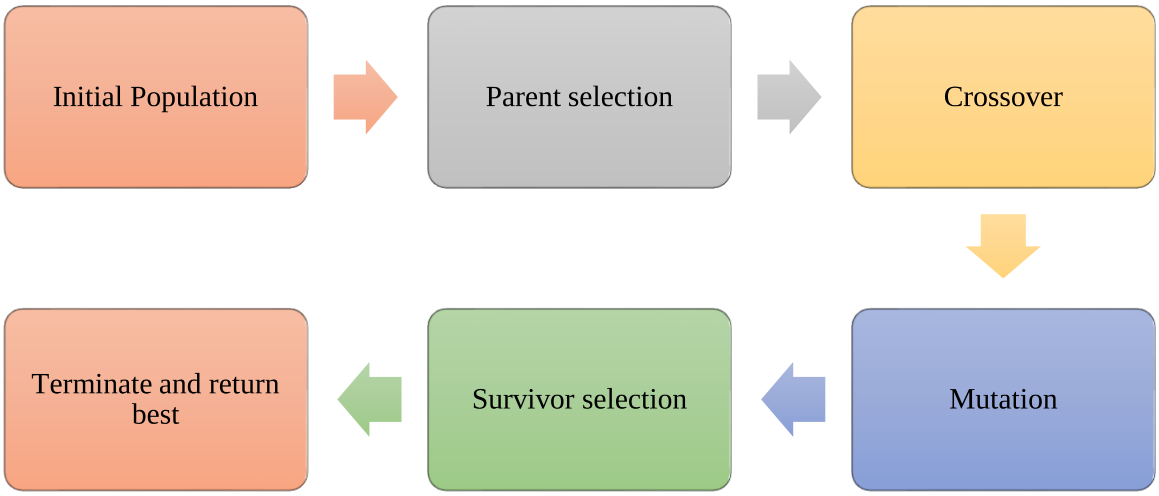


Fig 4. GA process as feature selection

New generations will be 20 best individuals and until the achievement of desired requirement, this loop and process will continue.

3.2. Artificial Neural Network

Artificial neural network is a computing system that tries to simulate the human thinking style or method (Agatonovic-Kustrin, S., & Beresford, R. 2000). The neural network can learn through data and be improved or reinforced through training. It including three layers: I. Input layer II. Hidden layer and III. Output layer. Firstly, considered variables or indicators insert in input variables. Each layer consists of two main parameters. I. weight II. Bias. In each layer, these two parameters add up together. Then they pass through an activation function which is used to recognize non-linear features. This process is done again in hidden layer but this time, weights and biases passing through a linear activation function.

There are different types of ANN such as Feedforward Neural Network (FNN), Convolutional Neural Network (CNN) and etc. In this paper, we used multi-layer perceptron (MLP). One of the most important parameters in ANN is training algorithm. In this paper, Levenberg-Marquardt (LM) is used as an optimization network. Initial training rate and the number of iteration is 0.01 and 1000 respectively. ANN parameters are as Table (4):

Table 4. parameters

Parameters	Explanations
Training	Back-propagation (BP)
Optimization algorithm	Levenberg-Marquardt (LM)
Training rate	0.01
Iterations	1000
Activation function	Tan-Sigmoid
	Pure line

Like GA, 70% of data is used to train dataset and the remaining is dedicated to validation and testing. Figure (5) shows the ANN research method to find optimal solution.

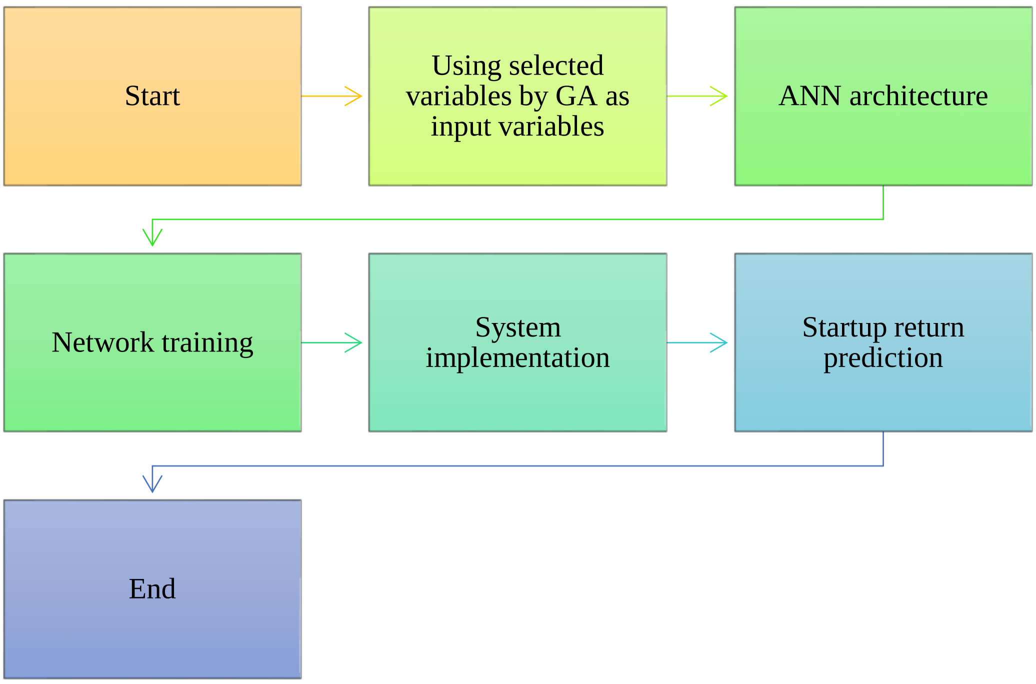


Fig 5. ANN process

3.3. Econometric Models

Economic has its own special scientific language. Econometric models are statistical models used in econometrics (Baltagi, B. H. 2011). When you want to explain the relationships between different economic indicators or variables, econometric models can be used. One of the most important concepts in econometric is regression analysis. Regression analysis is a statistical process which used for estimation (Chatterjee, S., & Hadi, A. S. 2013). In regression analysis, there are two types of variable: I. dependent variable (s) II. Independent variable (s). There are two types of regression analysis: I. multivariate regression analysis II. Univariate regression analysis.

As we mentioned earlier, every model has assumptions and limitations. So, econometric models are not exception. When you want to do regression analysis, it is necessary to take a few steps:

- Linearity: checking linearity is the first step in regression analysis and it is significance because they define the range of the method within which the results are obtained accurately and precisely. You can use Kolmogorov- Smirnov test (K-S) or Jarque-bera test for checking linearity.
- stationarity: before doing regression analysis, you should be assured that series is stationary or not. Unit root test like Augmented Dicky Fuller (ADF) test or differencing can be used to flat the trend, constant variance because The trend and seasonality will affect the value of time series at different times (Ryabko, D. 2019).

After checking these two assumption, you can do regression. A simple linear regression with one independent variable and two dependent variables is presented below:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \qquad i = 1. \dots . n \qquad (2)$$

y_i: dependent variable

β_0: intercept

β_1: x_i coefficient

x_i: independent variable

Finally, we predicted return for the next day using ARIMA model. ARIMA is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends.

Figure (6) shows the regression analysis process:

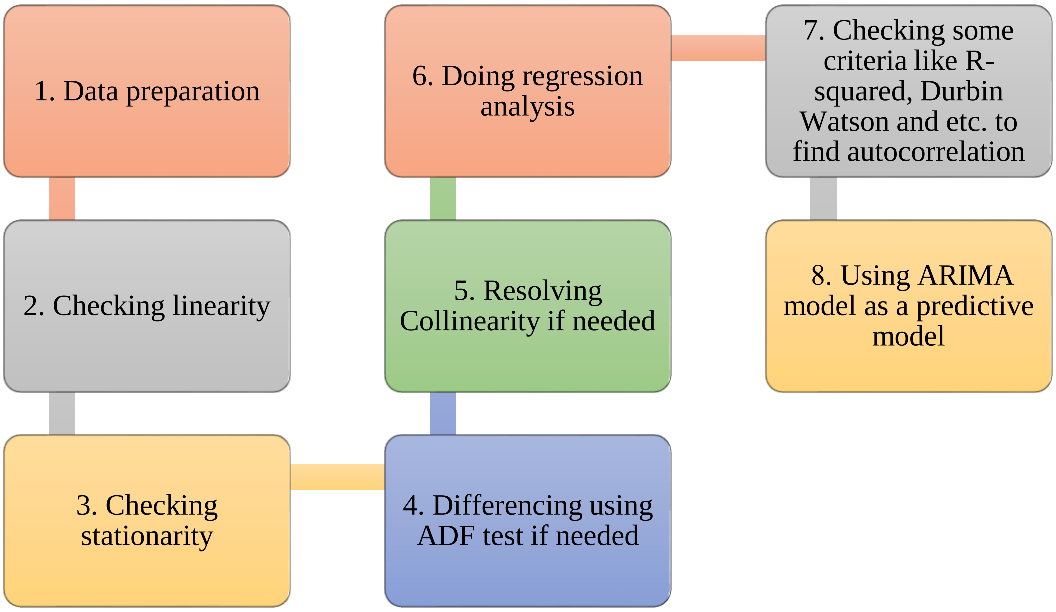


Fig 6. Regression analysis process

3.4. Portfolio Optimization

After predicting the return of companies, we will create an optimal portfolio including maximum profit and minimum risk. So, we used Value at Risk (VaR) and Conditional Value at Risk (C-VaR) to estimate the highest possibility of risk or maximum loss.

3.4.1. Value at Risk (VaR)

Because the focus of this article is about optimization of portfolio, we assume that returns follow normal distribution and correlation between risky assets are constant. In this condition, VaR is calculated as below:

$$VaR_{1-\alpha} = \mu_p - \acute{\emptyset}^{-1}(1-\alpha) \cdot \sigma_p \qquad (3)$$

Where $\mu_p$and $\sigma_p$are conditional mean and variance of portfolio respectively and $\acute{\emptyset}^{-1}$ is an inverse cumulative density function at α probability level. For portfolio optimization problem, VaR is defined as a minimum real number (γ) that does not exceed $w'r$ with α probability. This definition is expressed as:

$$VaR_\alpha(w) \ = \ Min\{\gamma: P(\gamma \leq -w'r) \leq \alpha\} \qquad (4)$$

Where r and w are returns and weights vectors of n risky assets and $\bar{R}_P = w'r$ is portfolio mean. Also, P indicates probability distribution of asset returns. Thus the portfolio optimization problem based on VaR definition can be written as:

$$Min_W \quad \gamma \qquad (5)$$

$$s \cdot t \cdot \quad w'e = 1$$

Where e is a vector of ones and budget constraint indicates that sum of assets weights equals (1).

In finance, it is assumed that distribution and return is normal with mean vector of μr and variance-covariance matrix of Ωr. By using and assuming parametric approach, the optimized portfolio is as below:

$$Min\ VaR_\alpha(w) = -w'\mu r + \acute{\emptyset}^{-1}(\alpha) \cdot \sqrt{w'\Omega rw} \qquad (6)$$

$$s \cdot t \cdot \quad 1 \cdot w'e = 1$$

$$2 \cdot \mu p \ = \ w'\mu r \ = \ \mathrm{r}$$

Where r is target gain and second constraint shows that expected mean of portfolio should be equal to r. μr and Ωr. and are achieved simply regarding equations 7 and 8 respectively.

$$\mu r = \begin{bmatrix} \mu 1 \\ \mu 2 \\ \vdots \\ \mu n \end{bmatrix}, \Omega r = \begin{bmatrix} \sigma 11 & \sigma 12 & \dots & \sigma 1n \\ \sigma 21 & \sigma 22 & \dots & \sigma 2n \\ \vdots & \vdots & \ddots & \vdots \\ \sigma n1 & \sigma n1 & \dots & \sigma nn \end{bmatrix} \quad (7)$$

Where

$$\mu_{i=E(ri).} \qquad \sigma_{ij} = E[(r_i - \mu_i)(r_j - \mu_j) \quad (8)$$

3.4.2. Conditional Value at Risk

Another measure of risk is C-VaR which have introduced by Uryasev and Rockafellar (1999). This indicator has some merits and advantages than VaR. One of them is that C-VaR can estimate risk under unfavorable economic condition. In other words, VaR measures expected loss under specified confidence or probability level in normal market state while C-VaR give helpful information about market and expected loss during unexpected economic condition. On the other hand, C-VaR provide information about left hand sie of distribution curve when expected loss exceeds VaR. C-VaR can be shown in mathematical way like as below:

$$CVaR_{1-a} = E(X.X > VaR_{1-a}) \quad (9)$$

Due to definition of VaR and assuming $f(x)$ as density function, C-VaR can represented as below:

$$CVaR_{1-a} = \frac{1}{\alpha}\int_{-\infty}^{VaR1-\alpha} x\, f(x)dx \quad (10)$$

If $f(x)$ considered as normal density function, C-VaR configured as follow:

$$CVaR_{1-a} = \mu_P - a^{-1} \cdot \varphi\, [\acute{\emptyset}^{-1}(1 - \alpha)] \cdot \sigma_P \quad (11)$$

Where φ is normal standard density and $\acute{\emptyset}$ is its cumulative distribution function. It is obvious that C-VaR is larger than VaR.

Problems can be solved and optimized by C-VaR risk measure as below:

$$CVaR_{1-a} = - w'\mu r + a^{-1}.\ \varphi\, [\acute{\emptyset}^{-1}(1 - \alpha)] \cdot \sqrt{w'\Omega rw}$$

$$s.t. \quad 1.\ w'e = 1 \quad (12)$$

$$2.\ \mu p = w'\mu r = r$$

At first, we calculated the return. After that, we have made a portfolio by a specified amount about 1milion dollar. Then this portfolio optimized with VaR and C-VaR models. Then the efficient frontier calculated.

4. Findings and results

As we mentioned earlier, we want to predict startup's return for three startups like DBX, G24.DE and TIE using daily historical data from the last three years. GA is used as feature selection and ANN used to find optimal solution. For comparability, we used econometric models like regression analysis. VaR and CVaR are two methods which are used to portfolio optimization. Finally, these two methods mean AI based methods and econometric models are compared using predictive performance metrics like precision, recall and sensitivity.

4.1. Genetic algorithm results

After getting data, we need to prepare and process them. Table (5) shows the normalized prepared data:

Table 5. data preview table

DBX										
No	Open	High	Low	Close	^TNX	CBBTCUSD	DCOILBRENTEU	DJIA	...	Return
1	-0.50804	-0.50217	-0.42956	-0.43888	0.892341	-0.96577	0.172756	-0.50024	...	0.33455
2	-0.45444	-0.50838	-0.39506	-0.50978	0.798687	-0.96102	0.151096	-0.5793	...	0.822897
...	...	...	...	...	...	...	...	...	...	...
722	0.074449	0.018001	0.136285	0.08802	-0.14836	1	0.913622	1	...	0.514308
G24.DE										
No	Open	High	Low	Close	^TNX	CBBTCUSD	DCOILBRENTEU	DJIA	...	Return
1	-0.87023	-0.85980	-0.90575	-0.86608	0.892341	-0.967665	0.172756	-0.50023	...	0.242708
2	-0.90956	-0.88431	-0.91666	-0.91593	0.798687	-0.963186	0.151096	-0.57929	...	0.564142
...	...	...	...	...	...	...	...	...	...	...
729	0.117719	0.119608	0.140873	0.121212	-0.14135	0.882458	0.650574	0.883618	...	0.467623
TIE.AS										
No	Open	High	Low	Close	^TNX	CBBTCUSD	DCOILBRENTEU	DJIA	...	Return
1	-0.96915	-0.95886	-0.96858	-0.95800	0.892341	-0.969836	0.172756	-0.51355	...	0.537009
2	-0.97943	-0.95372	-0.98429	-0.95275	0.798687	-0.965659	0.151096	-0.59050	...	0.45144
...	...	...	...	...	...	...	...	...	...	...
742	0.568123	0.568123	0.581152	0.580052	-0.23763	0.943655	0.607255	0.796214	...	0.549234

As you can see, different types of indicators are used as input variables. Table (6) shows the list of variables as input and target:

Table 6. variables table

No	Name	Use	No	Name	Use
1	Open	Input	11	NASDAQCOM	Input
2	High	Input	12	SP500	Input
3	Low	Input	13	US dollar index	Input
4	Close	Input	14	WILL5000INDFC	Input
5	^TNX	Input	15	Funding	Input
6	CBBTCUSD	Input	16	Business model	Input
7	DCOILBRENTEU	Input	17	Idea	Input
8	DJIA	Input	18	Team	Input
9	DTWEXBGS	Input	19	Timing	Input
10	Gold price	Input	20	Return	Target

20 variables are used (19 input variables and 1 target variable). There are different types of variable which can increase predictability. As we mentioned, we used 0.70% of data as training dataset and the remaining data is used to validation and testing.

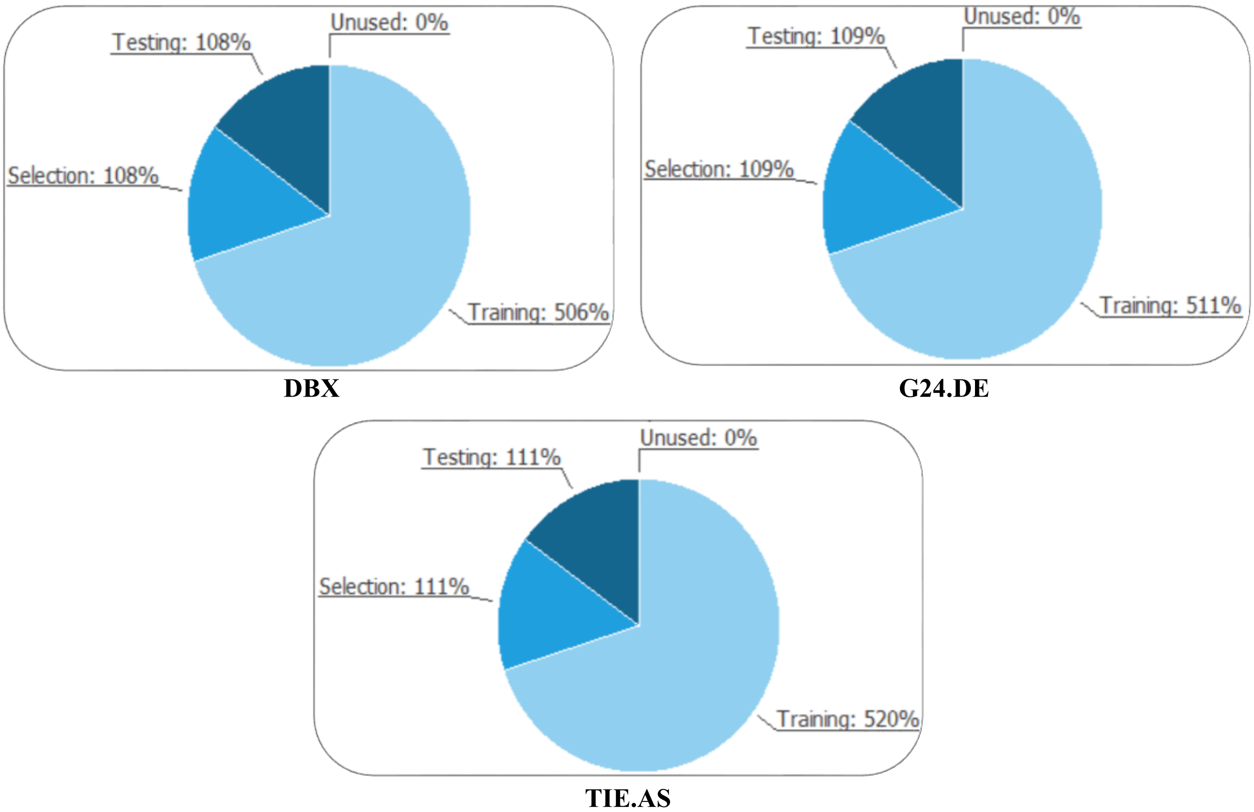


Fig 7. instances pie charts

One of the main indicators that can be useful to better understand the importance of variables is correlation. Table (7) shows the coefficient correlation between inputs and target.

Table 7. input-target correlation

DBX			G24.DE			TIE.AS		
Variables	type	Return	Variables	type	Return	Variables	type	Return
US dollar index	Linear	0.055739	US dollar index	Linear	-0.077643	^TNX	Linear	-0.055033
CBBTCUSD	Linear	-0.047404	SP500	Linear	-0.073814	DCOILBRENTEU	Linear	-0.04633
High	Linear	-0.047172	NASDAQCOM	Linear	-0.072948	US dollar index	Linear	0.043474
DCOILBRENTEU	Linear	-0.046661	WILL5000INDFC	Linear	-0.071625	Gold price	Linear	0.038137
Open	Linear	-0.044245	DJIA	Linear	-0.07064	NASDAQCOM	Linear	0.016387
Timing	Linear	-0.044023	High	Linear	-0.065849	Timing	Linear	-0.014995
Team	Linear	-0.044023	Open	Linear	-0.065164	Team	Linear	-0.014995
Idea	Linear	-0.044023	Timing	Linear	-0.063166	Idea	Linear	-0.014995
Business model	Linear	-0.044023	Team	Linear	-0.063166	Business model	Linear	-0.014995
Funding	Linear	-0.044023	Idea	Linear	-0.063166	Funding	Linear	-0.014995
Close	Linear	-0.044023	Business model	Linear	-0.063166	Close	Linear	-0.014995
Low	Linear	-0.03768	Funding	Linear	-0.063166	DTWEXBGS	Linear	0.013971
DJIA	Linear	-0.024628	Close	Linear	-0.063166	Low	Linear	-0.010277
Gold price	Linear	0.022602	Low	Linear	-0.060027	High	Linear	-0.009284
^TNX	Linear	-0.01475	DTWEXBGS	Linear	0.055138	Open	Linear	-0.008858
SP500	Linear	-0.013109	Gold price	Linear	-0.048585	WILL5000INDFC	Linear	0.006012
DTWEXBGS	Linear	-0.01257	CBBTCUSD	Linear	-0.046648	CBBTCUSD	Linear	-0.004019
WILL5000INDFC	Linear	-0.008415	DCOILBRENTEU	Linear	-0.026154	SP500	Linear	0.003305
NASDAQCOM	Linear	0.001767	^TNX	Linear	0.01507	DJIA	Linear	-0.002706

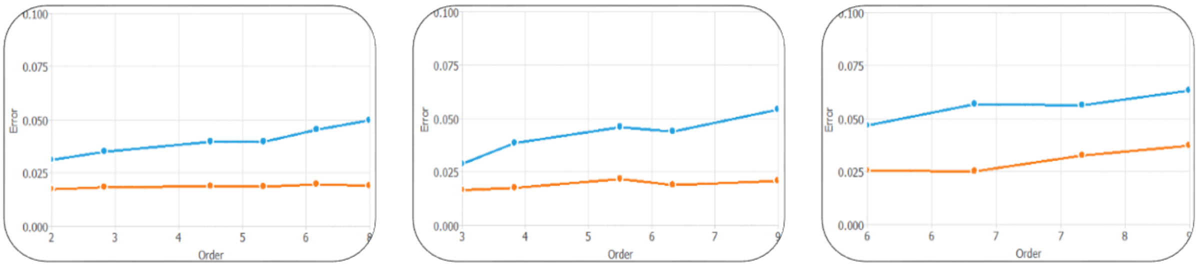
In DBX and G24.DE startups, the highest correlations are between returns and US dollar index while in TIE startup, TNX indicator has the highest correlation. Model selection is applied to find a neural network with a topology that optimize the error on new data. There are two kind of model selection: I. order selection which used to find the best architecture II. Input selection which used to find the most important variables.

The order selection algorithm chosen for this application is simulated annealing. This is a stochastic method inspired by the metallurgical industry. The parameters of order selection algorithm are as Table (8):

Table 8. Order selection algorithm parameters

Parameters	Description	Value
Minimum order	Number of minimum hidden perceptrons to be evaluated.	1
Maximum order	Number of maximum hidden perceptrons to be evaluated.	10
Cooling Rate	Temperature reduction factor for the simulated annealing.	0.5
Trials number	Number of trials for each neural network.	3
Tolerance	Tolerance for the selection error in the trainings of the algorithm.	0.01
Selection loss goal	Goal value for the selection error.	0
Minimum temperature	Minimum temperature reached in the simulated annealing algorithm.	0.001
Maximum iterations number	Maximum number of iterations to perform the algorithm.	100
Maximum time	Maximum time for the order selection algorithm.	3600
Plot training error history	Plot a graph with the training error of each iteration.	TRUE
Plot selection error history	Plot a graph with the selection error of each iteration.	TRUE

Error history for the different subsets during the SA order selection can be seen in the next charts. The blue line represents the training error and the orange line symbolizes the selection error.



DBX G24.DE TIE.AS

Fig 8. Simulated Annealing error plots

Table (9) shows the order selection results using SA optimization algorithm. They include some final states from the neural network, the error functional and the order selection algorithm.

Table 9. order selection results

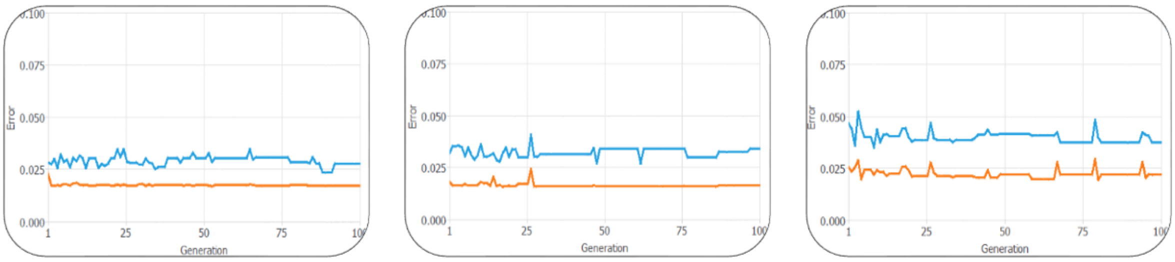
Parameters	DBX	G2.DE	TIE.AS
Optimal order	2	3	6
Optimum training error	0.0310832	0.028701	0.046772
Optimum selection error	0.0172316	0.016378	0.025404
Iterations number	5	5	5
Elapsed time	0:02	0:01	0:02

Table (9) shows that appropriate and optimal neurons in hidden layer for DBX, G.24 and TIE.AS are 2, 3 and 6 respectively. Genetic algorithm is used as feature selection. Table (10) shows the GA parameters as input selection model:

Table 10. input selection parameters

Parameters	Description	Value
Trials number	Number of trials for each neural network.	1
Tolerance	Tolerance for the selection error in the trainings of the algorithm.	0.01
Population size	Size of the population of each generation.	20
Initialization method	Initialization method used in the algorithm.	Random
Fitness assignment method	Fitness assignment method used in the algorithm.	Rank Based
Crossover method	Crossover method used in the algorithm.	Uniform
Elitism size	Number of individuals which will always be selected for recombination.	2
Crossover first point	First point used in the One Point and Two Point crossover method. If it is 0 the algorithm selects a random point for each pair of offspring.	0
Crossover second point	Second point used in the Two Point crossover method. If it is 0 the algorithm selects a random point for each pair of offspring.	0
Selective pressure	Rank-Based fitness assignment allows values for the selective pressure greater than 0.	1.5
Mutation rate	This is a parameter of the mutation operator.	0.05
Selection loss goal	Goal value for the selection error.	0
Maximum Generations number	Maximum number of generations to perform the algorithm.	100
Maximum time	Maximum time for the inputs selection algorithm.	3600
Plot training error history	Plot a graph with the optimum training error of each generation.	TRUE
Plot selection error history	Plot a graph with the optimum selection error of each generation.	TRUE
Plot generation mean history	Plot a graph with the mean of the selection error of each generation.	TRUE
Plot generation standard deviation history	Plot a graph with the standard deviation of the selection error of each generation.	FALSE

Figure (9) presents the error history during input selection using GA. the blue line represents the training error, its initial value for DBX, G24.DE and TIE are 0.0282329, 0.0318744 and 0.0467448 respectively. The final value after 100 generations are 0.0282329, 0.0353369 and 0.0522988. the orange line symbolizes the selection error, its initial values are 0.0230323, 0.0180756 and 0.0257439 and the final value after 100 generations are 0.0230323, 0.0163135 and 0.0287222.



DBX G24.DE TIE.AS

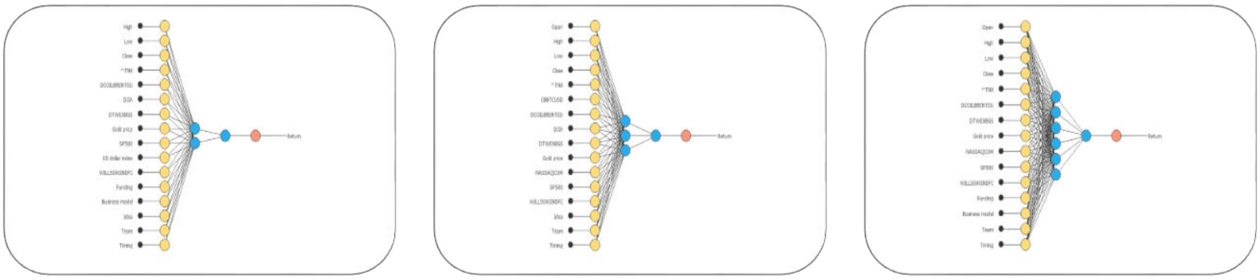
Fig 9. Genetic algorithm error plots

Table (11) presents the input selection results by the GA. It is including some optimal parameters like the number of optimal input variables, generations number and etc.

Table 11. GA input selection result

Parameters	DBX	G24.DE	TIE.AS
Optimal number of inputs	16	16	15
Optimum training error	0.028233	0.035337	0.052299
Optimum selection error	0.023032	0.016314	0.028722
Generations number	100	100	100
Elapsed time	0:02	0:04	0:15

Figure (10) represents graphical results such as the number of input variables and hidden layers. the yellow circles represent scaling neurons, the blue circles perceptron neurons and the red circles un-scaling neurons.



DBX G24.DE TIE.AS

Fig 10. Final architectures

Finally, you can see the errors based on different loss functions such as SSE, MSE and etc. in Table (12).

Table 12. error estimation based on different loss functions

DBX			
Parameters	Training	Selection	Testing
Sum squared error	130.862	29.2654	29.2952
Mean squared error	0.25862	0.270975	0.271252
Root mean squared error	0.508547	0.520553	0.520819
Normalized squared error	17.0099	15.81	16.434
Minkowski error	179.583	39.5988	39.5928
G24.DE			
Parameters	Training	Selection	Testing
Sum squared error	128.194	27.9491	26.7845
Mean squared error	0.250869	0.256414	0.245729
Root mean squared error	0.500869	0.506373	0.495711
Normalized squared error	16.5762	16.0352	15.8858
Minkowski error	177.208	38.3036	37.2006
TIE.AS			
Parameters	Training	Selection	Testing
Sum squared error	139.476	31.523	28.3731
Mean squared error	0.268224	0.283991	0.255613
Root mean squared error	0.517903	0.532908	0.505582
Normalized squared error	15.7846	14.2163	17.5963
Minkowski error	189.218	42.0131	39.0573

4.2. Artificial Neural Network results

ANN is used to find optimal solution. Input variables of ANN are those variables that found and obtained using GA. Network parameters and algorithms are as Table (13):

Table 13. Network parameters

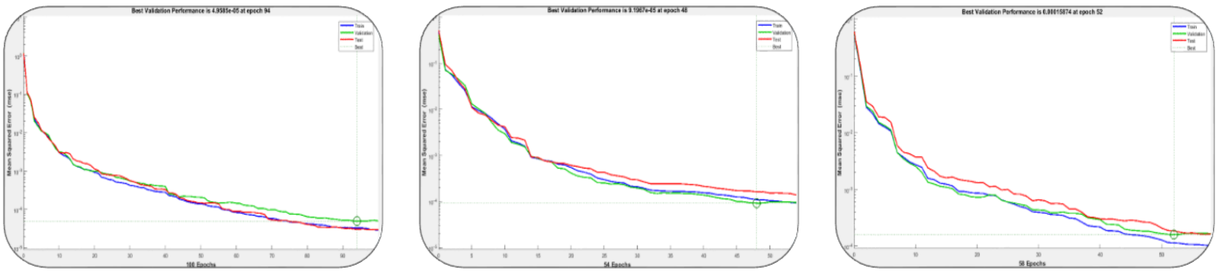
Parameters	
Data division	Random (dividerand)
Training	Scaled Conjugated Gradient (Trainscg)
Performance	Mean Square Error (MSE)
Calculations	MEX

After training network by considered parameters which mentioned in section3, the following progress and results are obtained:

Table 14. Network progress

DBX			
Parameters	Initial value	Results	Upper bound (value)
Epoch	0	100 iterations	1000
Time	-	0:00:00	-
Performance	1.22	2.91e-05	0.00
Gradient	2.79	0.000716	1.00e-06
Validation checks	0	6	6
G24.DE			
Parameters	Initial value	Results	Upper bound (value)
Epoch	0	54 iterations	1000
Time	-	0:00:00	-
Performance	0.483	9.20e-05	0.00
Gradient	1.51	0.00280	1.00e-06
Validation checks	0	6	6
TIE.AS			
Parameters	Initial value	Results	Upper bound (value)
Epoch	0	58 iterations	1000
Time	-	0:00:00	-
Performance	0.547	0.000102	0.00
Gradient	2.04	0.00178	1.00e-06
Validation checks	0	6	6

Different parameters have different tasks. For example, gradient is an index which can work as a barometer and shows training state. Figure (11) shows network performance during each epoch:



DBX **G24.DE** **TIE.AS**

Fig 11. Neural network performance during epochs

The best validation performance for DBX, GE.24 and TIE.AS are 4.9585e-05, 9.1967e-05 and 0.00015874 at epochs 94, 48 and 52 respectively. (To get more information about error histogram, please see Figure A1 in the appendix). The last step is regression. It is including three parts such as training, validation and testing. Figure (12) shows the regression (fitted data) for each datasets:

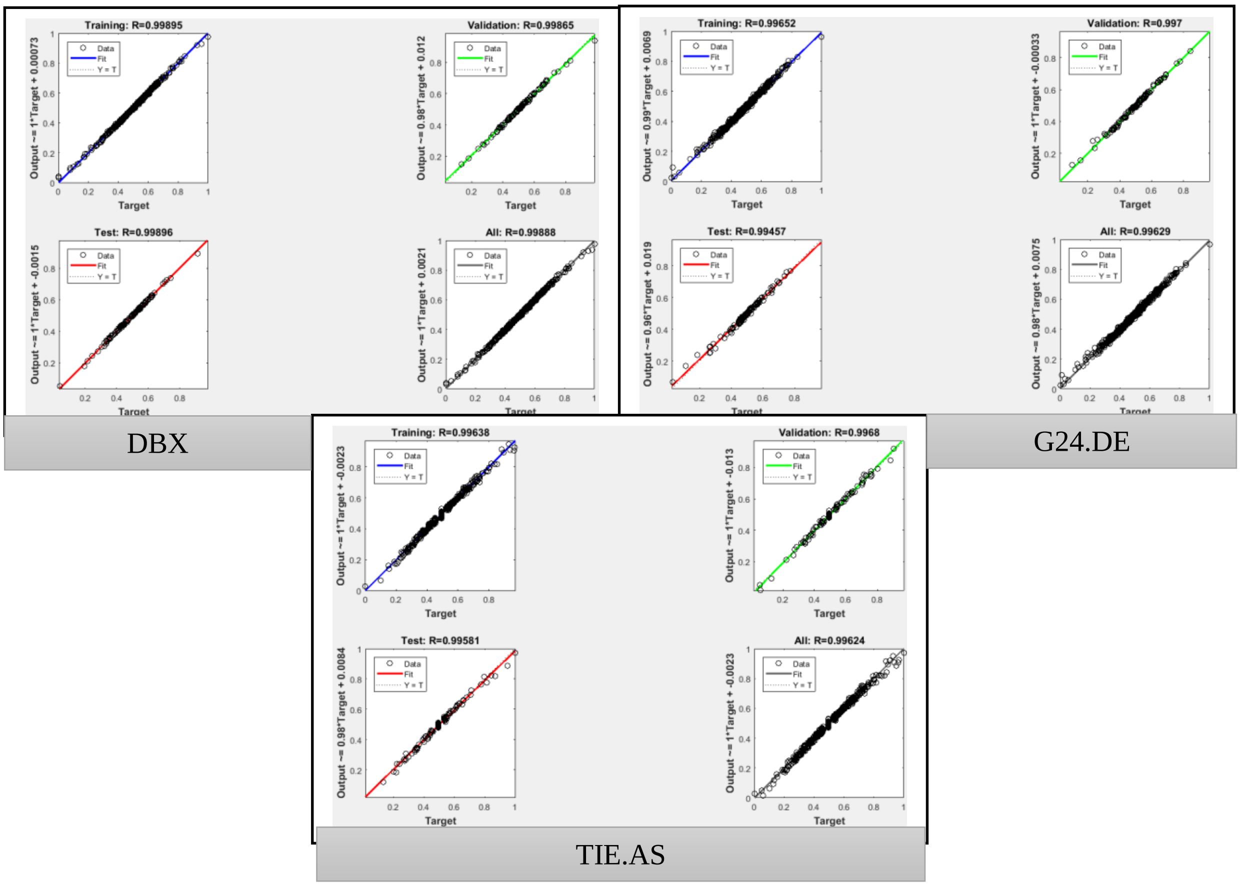


Fig 12. Regression analysis (Actual vs. predicted data)

There are some metrics which can show the goodness of fit. One of them is R-squared. As you can see, in all three startups, R-squared is more than 99% and it means the high predictability of ANN.

4.3. Econometric model results

As we mentioned earlier, econometric models have different assumptions and hypothesis like normality, linearity, stationarity and etc. in this paper, because logarithmic return is used, the series are normal and stationary. As a result, there is no need for a stationarity test. After doing regression, Figure (13) appeared for all three startups:

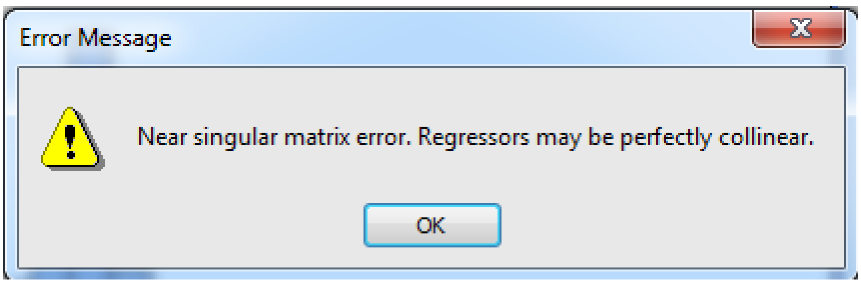


Fig 13. Error message after doing regression

Figure (13) shows that there is a collinearity between some variables. We need to solve this problem. One of the main solutions is correlation matrix. By using correlation matrix, we can identify variables that have perfect correlation to each other and then eliminate them.

As you can see, in DBX startup, there is a perfect collinearity between TEAM, HIGH, BUSINESS MODEL, CLOSE, FUNDING, LOW, IDEA, TIMING. In G24.DE company, this collinearity is between TEAM, HIGH, CLOSE, LOW, IDEA, TIMING, OPEN. In TIE.AS, there is a collinearity between TEAM, HIGH, BUSINESS MODEL, CLOSE, FUNDING, LOW, TIMING, OPEN. After eliminating determined variables, we did regression again and the following results obtained.

Table 16. DBX regression results

Dependent Variable: RETURN
Method: Least Squares (Gauss-Newton / Marquardt steps)
Date: 02/20/22 Time: 08:15
Sample: 1 756
Included observations: 755
RETURN=C(1)+C(2)*_TNX+C(3)*DCOILBRENTEU+C(4)*DJIA+C(5)
*DTWEXBGS+C(6)*GOLD_PRICE+C(7)*IDEA+C(8)*SP500+C(9)
*US_DOLLAR_INDEX+C(10)*WILL5000INDFC

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-0.028118	0.155713	-0.180576	0.8567
C(2)	0.008644	0.004215	2.050891	0.0406
C(3)	-0.000303	0.000191	-1.587952	0.1127
C(4)	1.12E-06	2.97E-06	0.376762	0.7065
C(5)	-0.000117	0.000888	-0.131635	0.8953
C(6)	-5.39E-07	2.05E-05	-0.026233	0.9791
C(7)	-0.005176	0.001500	-3.451622	0.0006
C(8)	-7.99E-05	5.91E-05	-1.351870	0.1768
C(9)	0.000699	0.000588	1.188867	0.2349
C(10)	0.001539	0.000880	1.748575	0.0808

R-squared	0.025492	Mean dependent var	0.000228
Adjusted R-squared	0.013720	S.D. dependent var	0.025957
S.E. of regression	0.025778	Akaike info criterion	-4.465426
Sum squared resid	0.495060	Schwarz criterion	-4.404145
Log likelihood	1695.698	Hannan-Quinn criter.	-4.441821
F-statistic	2.165392	Durbin-Watson stat	2.190388
Prob(F-statistic)	0.022544		

Table 15. correlation matrix

DBX

Variables	RETURN	_TNX	BUSINESS	CLOSE	DCOILBRI	DJIA	DTWEXBC	FUNDING	GOLD_PR	HIGH	IDEA	LOW	SP500	TEAM	US_DOLL	WILL5000INDFC
RETURN	1	-0.011268	-0.070273	-0.070273	-0.054555	-0.029308	-0.001329	-0.070273	0.0178917	-0.069578	-0.070273	-0.062933724	-0.020602979	-0.070273	0.0394489	-0.017381504
_TNX	-0.011268	1	-0.048267	-0.048267	0.3955912	-0.32404	0.0520471	-0.048267	-0.889667	-0.051023	-0.048267	-0.048645306	-0.441127722	-0.048267	-0.606362	-0.45405337
BUSINESS_MODEL	-0.070273	-0.048267	1	1	0.5071134	0.6974832	-0.554975	1	0.2471716	0.9966718	1	0.996766208	0.686892442	1	0.4685589	0.697637633
CLOSE	-0.070273	-0.048267	1	1	0.5071134	0.6974832	-0.554975	1	0.2471716	0.9966718	1	0.996766208	0.686892442	1	0.4685589	0.697637633
DCOILBRENTEU	-0.054555	0.3955912	0.5071134	0.5071134	1	0.6338877	-0.695456	0.5071134	-0.238811	0.5059351	0.5071134	0.514121529	0.499936421	0.5071134	-0.046768	0.49189616
DJIA	-0.029308	-0.32404	0.6974832	0.6974832	0.6338877	1	-0.718095	0.6974832	0.538915	0.6992336	0.6974832	0.702305335	0.977734615	0.6974832	0.4975904	0.974157168
DTWEXBGS	-0.001329	0.0520471	-0.554975	-0.554975	-0.695456	-0.718095	1	-0.554975	-0.169747	-0.55302	-0.554975	-0.556032802	-0.621383576	-0.554975	-0.404974	-0.636241566
FUNDING	-0.070273	-0.048267	1	1	0.5071134	0.6974832	-0.554975	1	0.2471716	0.9966718	1	0.996766208	0.686892442	1	0.4685589	0.697637633
GOLD_PRICE	0.0178917	-0.889667	0.2471716	0.2471716	-0.238811	0.538915	-0.169747	0.2471716	1	0.2503665	0.2471716	0.247437312	0.666700448	0.2471716	0.7179948	0.671879079
HIGH	-0.069578	-0.051023	0.9966718	0.9966718	0.5059351	0.6992336	-0.55302	0.9966718	0.2503665	1	0.9966718	0.99549609	0.690095958	0.9966718	0.4704537	0.701059677
IDEA	-0.070273	-0.048267	1	1	0.5071134	0.6974832	-0.554975	1	0.2471716	0.9966718	1	0.996766208	0.686892442	1	0.4685589	0.697637633
LOW	-0.062934	-0.048645	0.9967662	0.9967662	0.5141215	0.7023053	-0.556033	0.9967662	0.2474373	0.9954961	0.9967662	1	0.690468349	0.9967662	0.4657619	0.700595758
SP500	-0.020603	-0.441128	0.6868924	0.6868924	0.4999364	0.9777346	-0.621384	0.6868924	0.6667004	0.690096	0.6868924	0.690468349	1	0.6868924	0.5892559	0.998492335
TEAM	-0.070273	-0.048267	1	1	0.5071134	0.6974832	-0.554975	1	0.2471716	0.9966718	1	0.996766208	0.686892442	1	0.4685589	0.697637633
US_DOLLAR_INDEX	0.0394489	-0.606362	0.4685589	0.4685589	-0.046768	0.4975904	-0.404974	0.4685589	0.7179948	0.4704537	0.4685589	0.465761937	0.589255865	0.4685589	1	0.609784134
WILL5000INDFC	-0.017382	-0.454053	0.6976376	0.6976376	0.4918962	0.9741572	-0.636242	0.6976376	0.6718791	0.7010597	0.6976376	0.700595758	0.998492335	0.6976376	0.6097841	1

G24.DE

Variables	RETURN	_TNX	CLOSE	OILBRENT	DJIA	DTWEXBGS	OLD_PRIC	HIGH	IDEA	LOW	SP500	TEAM	WILL5000INDFC
RETURN	1	0.0182376	-0.0765	-0.046739	-0.071622	0.0588997	-0.043029	-0.077482	-0.0765	-0.074398	-0.070481	-0.076499538	-0.068331344
_TNX	0.0182376	1	-0.868411	0.3940702	-0.32226	0.0519579	-0.888796	-0.871704	-0.868411	-0.864291	-0.437178	-0.868410783	-0.4504543
CLOSE	-0.0765	-0.868411	1	-0.269787	0.4419738	-0.039378	0.892695	0.9980979	1	0.9981719	0.5465865	1	0.546514673
DCOILBRENTEU	-0.046739	0.3940702	-0.269787	1	0.6366491	-0.691308	-0.234084	-0.276473	-0.269787	-0.260795	0.5043171	-0.269786809	0.496307563
DJIA	-0.071622	-0.32226	0.4419738	0.6366491	1	-0.71057	0.5402234	0.4419141	0.4419738	0.4455528	0.9777555	0.441973795	0.974390198
DTWEXBGS	0.0588997	0.0519579	-0.039378	-0.691308	-0.71057	1	-0.168436	-0.036208	-0.039378	-0.046851	-0.611868	-0.039378157	-0.627464072
GOLD_PRICE	-0.043029	-0.888796	0.892695	-0.234084	0.5402234	-0.168436	1	0.8989189	0.892695	0.887542	0.6660337	0.892694972	0.671516338
HIGH	-0.077482	-0.871704	0.9980979	-0.276473	0.4419141	-0.036208	0.8989189	1	0.9980979	0.9969433	0.5487011	0.9980979	0.548669735
IDEA	-0.0765	-0.868411	1	-0.269787	0.4419738	-0.039378	0.892695	0.9980979	1	0.9981719	0.5465865	1	0.546514673
LOW	-0.074398	-0.864291	0.9981719	-0.260795	0.4455528	-0.046851	0.887542	0.9969433	0.9981719	1	0.5480555	0.998171862	0.547898376
SP500	-0.070481	-0.437178	0.5465865	0.5043171	0.9777555	-0.611868	0.6660337	0.5487011	0.5465865	0.5480555	1	0.546586479	0.998487172
TEAM	-0.0765	-0.868411	1	-0.269787	0.4419738	-0.039378	0.892695	0.9980979	1	0.9981719	0.5465865	1	0.546514673
WILL5000INDFC	-0.068331	-0.450454	0.5465147	0.4963076	0.9743902	-0.627464	0.6715163	0.5486697	0.5465147	0.5478984	0.9984872	0.546514673	1

TIE.AS

Variables	RETURN	BUSINESS	CLOSE	DCOILBRI	DTWEXBC	FUNDING	GOLD_PR	HIGH	LOW	SP500	TEAM	WILL5000INDFC
RETURN	1	-0.030285	-0.030285	-0.061989	0.0245075	-0.030285	0.0332601	-0.024238	-0.025761	-0.014771	-0.030285	-0.012383291
BUSINESS_MODEL	-0.030285	1	1	0.4115237	-0.604392	1	0.7032634	0.9994019	0.9992782	0.9443028	1	0.954363579
CLOSE	-0.030285	1	1	0.4115237	-0.604392	1	0.7032634	0.9994019	0.9992782	0.9443028	1	0.954363579
DCOILBRENTEU	-0.061989	0.4115237	0.4115237	1	-0.686058	0.4115237	-0.225341	0.4065179	0.4149354	0.512361	0.4115237	0.504484838
DTWEXBGS	0.0245075	-0.604392	-0.604392	-0.686058	1	-0.604392	-0.167156	-0.602453	-0.603605	-0.595769	-0.604392	-0.613375061
FUNDING	-0.030285	1	1	0.4115237	-0.604392	1	0.7032634	0.9994019	0.9992782	0.9443028	1	0.954363579
GOLD_PRICE	0.0332601	0.7032634	0.7032634	-0.225341	-0.167156	0.7032634	1	0.7059	0.7010835	0.6638752	0.7032634	0.670399987
HIGH	-0.024238	0.9994019	0.9994019	0.4065179	-0.602453	0.9994019	0.7059	1	0.9987292	0.9431499	0.9994019	0.95343642
LOW	-0.025761	0.9992782	0.9992782	0.4149354	-0.603605	0.9992782	0.7010835	0.9987292	1	0.9453297	0.9992782	0.955019453
SP500	-0.014771	0.9443028	0.9443028	0.512361	-0.595769	0.9443028	0.6638752	0.9431499	0.9453297	1	0.9443028	0.998399255
TEAM	-0.030285	1	1	0.4115237	-0.604392	1	0.7032634	0.9994019	0.9992782	0.9443028	1	0.954363579
WILL5000INDFC	-0.012383	0.9543636	0.9543636	0.5044848	-0.613375	0.9543636	0.6704	0.9534364	0.9550195	0.9983993	0.9543636	1

Table 17. G24.DE regression results

Dependent Variable: RETURN
Method: Least Squares (Gauss-Newton / Marquardt steps)
Date: 02/20/22 Time: 08:39
Sample: 1 760
Included observations: 759
RETURN=C(1)+C(2)*_TNX+C(3)*DCOILBRENTEU+C(4)*DJIA+C(5)
*DTWEXBGS+C(6)*GOLD_PRICE+C(7)*IDEA+C(8)*SP500+C(9)
*WILL5000INDFC

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-0.138724	0.074214	-1.869240	0.0620
C(2)	-0.002449	0.002526	-0.969608	0.3326
C(3)	1.15E-05	0.000112	0.103333	0.9177
C(4)	2.89E-06	1.69E-06	1.707724	0.0881
C(5)	0.001141	0.000503	2.270376	0.0235
C(6)	1.78E-05	1.27E-05	1.403734	0.1608
C(7)	-0.001771	0.000527	-3.364003	0.0008
C(8)	-5.30E-05	3.27E-05	-1.620983	0.1054
C(9)	0.000658	0.000475	1.384835	0.1665

R-squared	0.023992	Mean dependent var	0.000547
Adjusted R-squared	0.013581	S.D. dependent var	0.015407
S.E. of regression	0.015302	Akaike info criterion	-5.509825
Sum squared resid	0.175623	Schwarz criterion	-5.454900
Log likelihood	2099.979	Hannan-Quinn criter.	-5.488673
F-statistic	2.304562	Durbin-Watson stat	2.013279
Prob(F-statistic)	0.019180		

Table 18. TIE.AS regression results

Dependent Variable: RETURN
Method: Least Squares (Gauss-Newton / Marquardt steps)
Date: 02/20/22 Time: 09:02
Sample: 1 769
Included observations: 768
RETURN=C(1)+C(2)*DCOILBRENTEU+C(3)*DTWEXBGS+C(4)
*GOLD_PRICE+C(5)*SP500+C(6)*WILL5000INDFC

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	0.011341	0.069852	0.162357	0.8711
C(2)	-0.000105	0.000172	-0.608910	0.5428
C(3)	2.09E-05	0.000563	0.037183	0.9703
C(4)	1.66E-07	1.21E-05	0.013698	0.9891
C(5)	-2.29E-05	3.15E-05	-0.725079	0.4686
C(6)	0.000433	0.000573	0.754623	0.4507

R-squared	0.005321	Mean dependent var	0.001449
Adjusted R-squared	-0.001205	S.D. dependent var	0.024274
S.E. of regression	0.024289	Akaike info criterion	-4.589830
Sum squared resid	0.449534	Schwarz criterion	-4.553551
Log likelihood	1768.495	Hannan-Quinn criter.	-4.575867
F-statistic	0.815323	Durbin-Watson stat	2.335528
Prob(F-statistic)	0.538848		

Due to different probability, it is clear that values less than 0.05 are important. So, in DBX, C (2), C (7) and in G24.DE C (5), C (7) coefficients are more important than others. As it is clear, for all three startups, the rate of R-squared is very low and it means that these variables are not relevant to the return and we should find other related and more important variables.

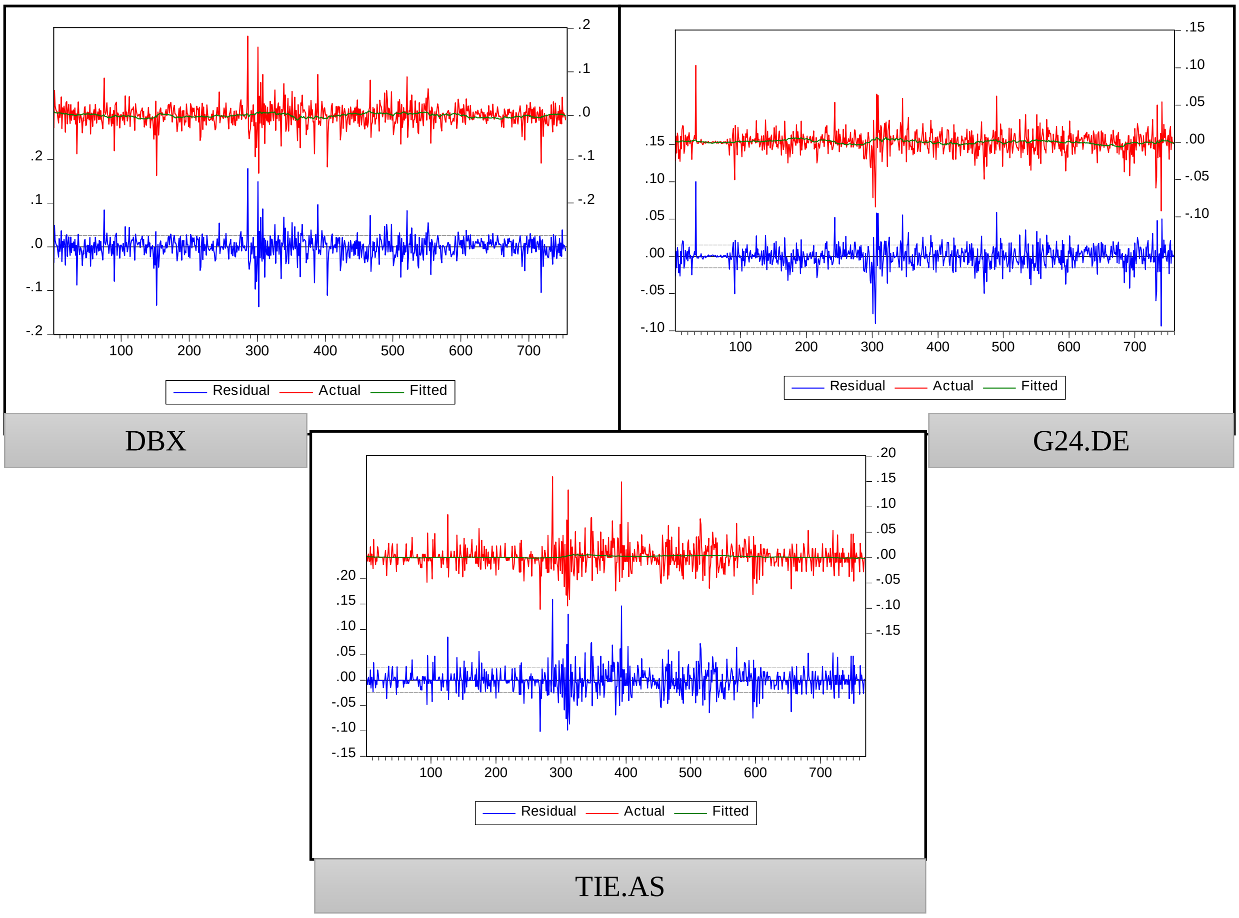


Fig 14. actual vs. Predicted

The red-line shows the actual data and the green-line shows the predicted data. The blue-line shows the residual too. As it is clear, fitted data are too far from actual data and couldn't predict volatility.

4.4. Value at Risk and Conditional Value at Risk

4.4.1. VaR estimation

The next step is the calculation of VaR with specified parameters. Here, you can see the return chart of portfolio:

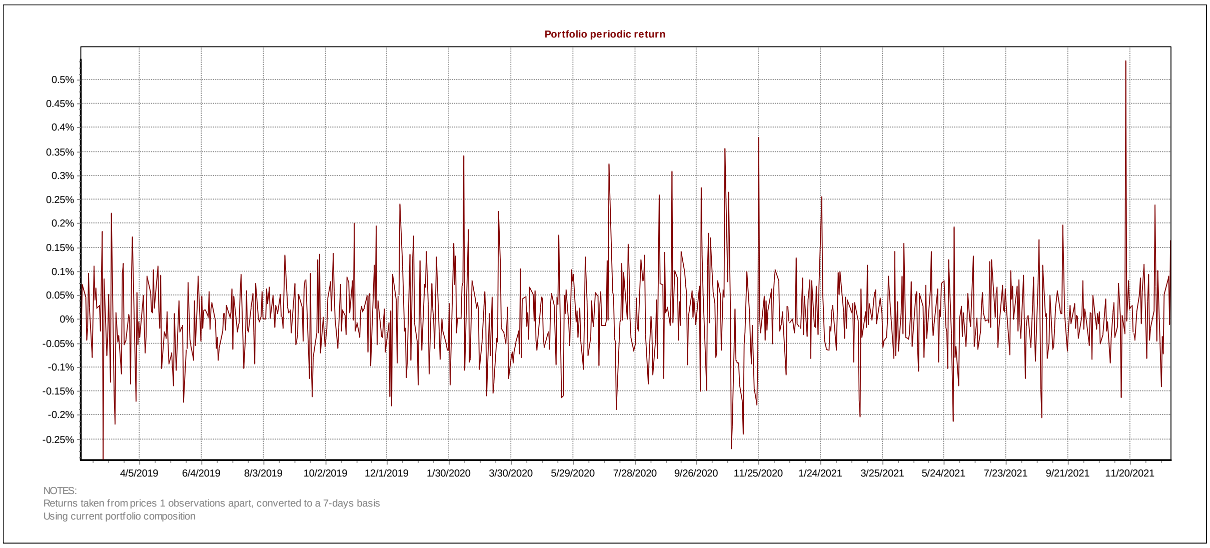


Fig 15. Periodic return

Note that, return will be essentially different according to the selected interval. For return calculation in VaR, the following formula is used:

$$Return(t) = \left[\frac{Value(t)}{Value(t - Interval)}\right]^{\frac{NumDaysVaRAnalysis}{[Date(t) - Date(t - Interval)]}}$$

Where:

- Return(t) is the return shown in the graph for the observation number t, on date Date(t)
- Value(t) is the value of the portfolio on the observation number t
- Interval is the selected observations interval
- Date(t) is the real date index from observations number t
- NumDays VaRAnalysis is the VaR horizon (in days), used as the unit of time to express the returns.

VaR parameters and VaR results can be observed in Table (19):

Table 19. VaR parameters and results

VaR parameters		
VaR horizon (in days)	7	
Significant Level	5%	
VaR results		
	Returns	Portfolio Value
Absolute	-12.825%	-13,76
Relative to mean	-13.580%	-33,665.47%
Base portfolio value to calculate VaR	107,29 (02/07/2019)	

The portfolio is not likely to lose more than 13.76 of value after 7 days following 02/07/2019 with 95% of confidence.

The results are shown in percentage and monetary units. Based on the portfolio value on the date selected. The VaR is calculated both in absolute terms (actual loss) and of the historical returns.

Figure (16) shows the portfolio periodic returns histogram:

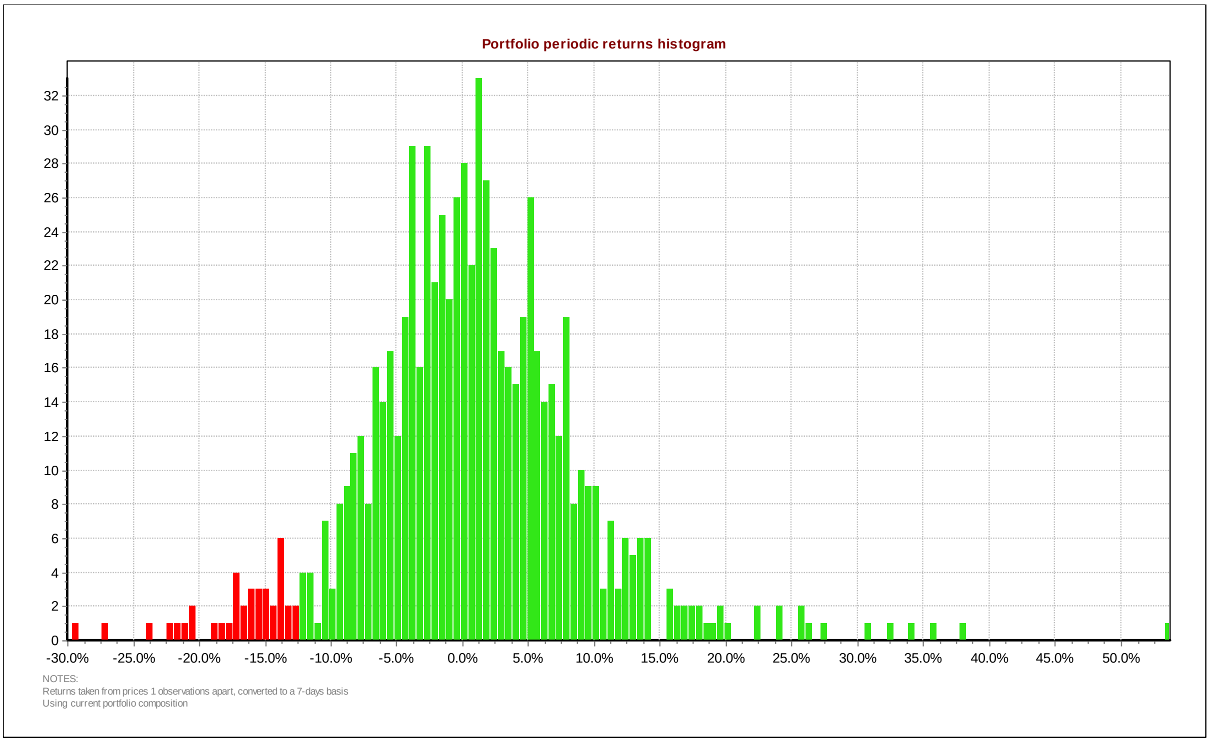


Fig 16. Portfolio return histogram with VaR significance level

This is a histogram that marks in red the losses below the specified VaR significance level (the Value at Risk limit). The results show that the portfolio is not likely to loss more than -13,76 of value after 7 days following 02/07/2019, with a 95.0% of confidence. Here, you can see a graph that shows how the VaR changes when different horizons are assumed. The table shows the VaR in percentage and in monetary units in base at portfolio value for different horizons (in days). Note that the calculated VaR is the absolute measure (actual expected loss). The base portfolio value is 107,29 (02/07/2019).

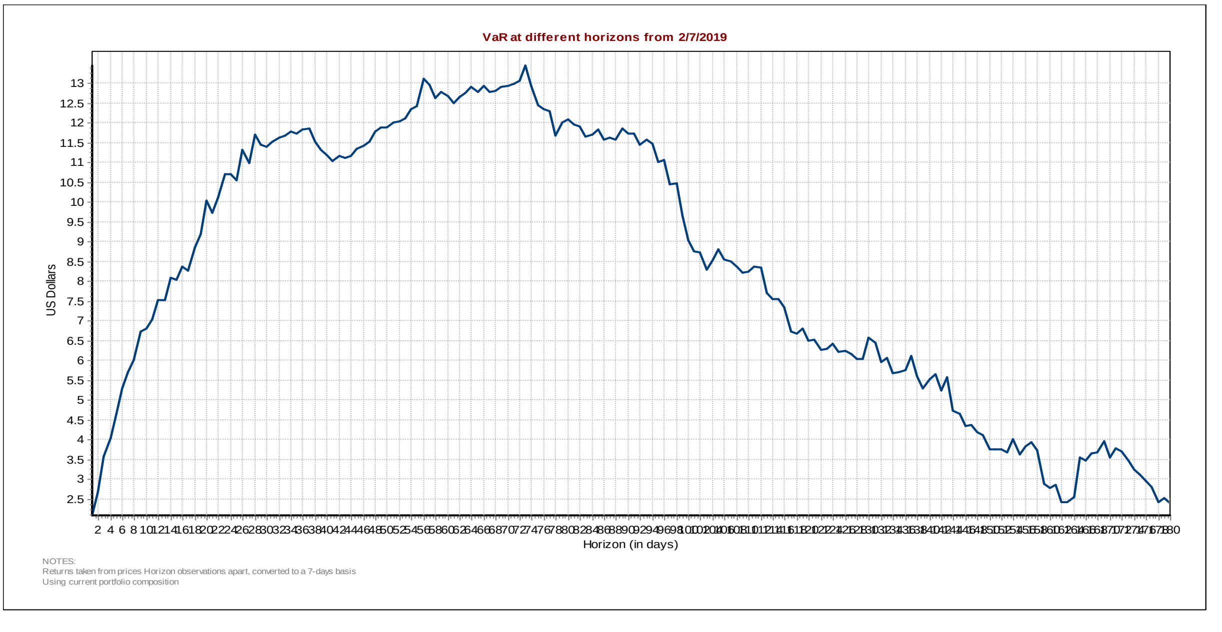


Fig 17. VaR at different horizons

Figure (18) shows a simulation of the portfolio value (on the right axis) and the VaR (in monetary units, on the left axis) along the sample.

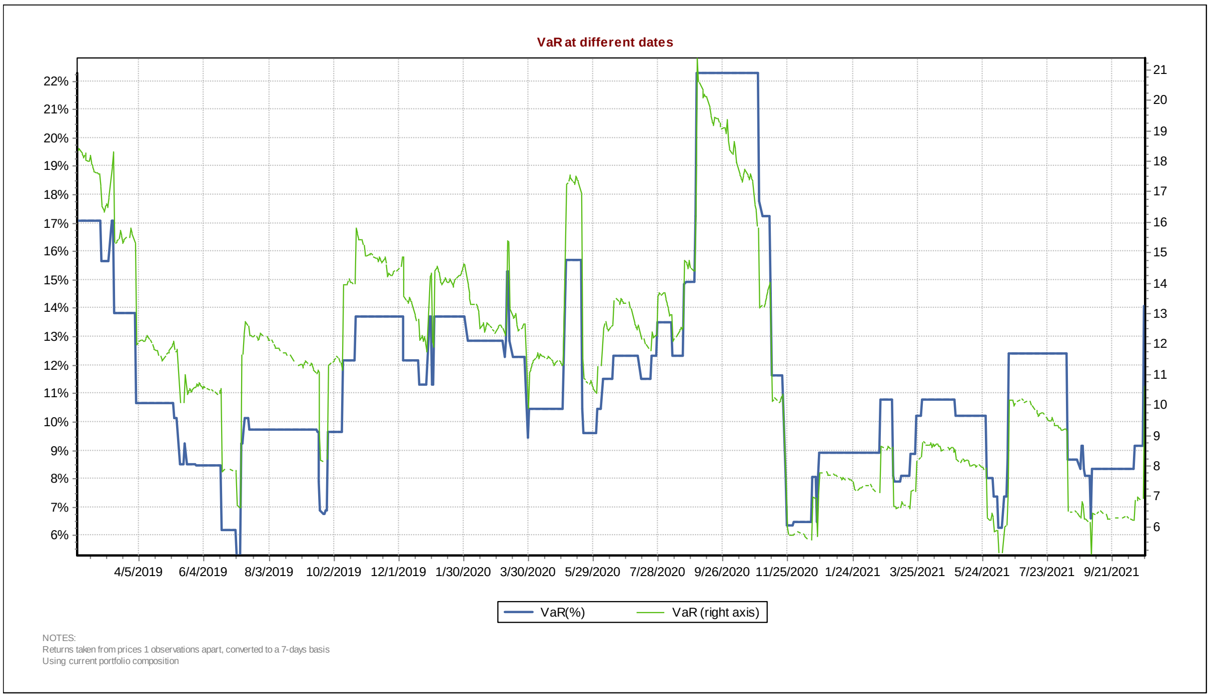


Fig 18. VaR along time

The simulation uses what is usually called "moving window" approach. This means that a fixed number of past observations will be used to calculate the VaR at all possible dates inside the sample. The window size (the number of observations used for the simulation) is a critical parameter. A large window will reduce the possible dates for simulation (because the first dates of the sample will lack enough past data to be used). The window size (obs) is 50 and the empirical alpha is 6.657%. For testing the results and confidence, back-testing is used:

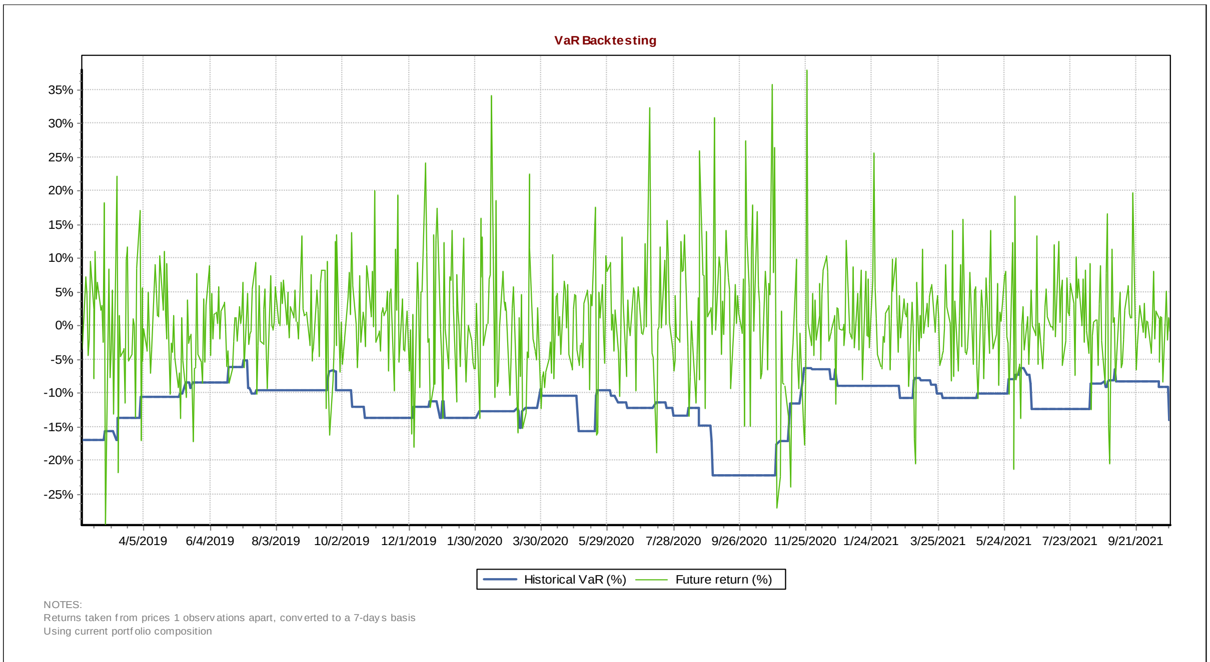


Fig 19. VaR back-testing

Figure (20) shows the Beta VaR decomposition of the VaR.

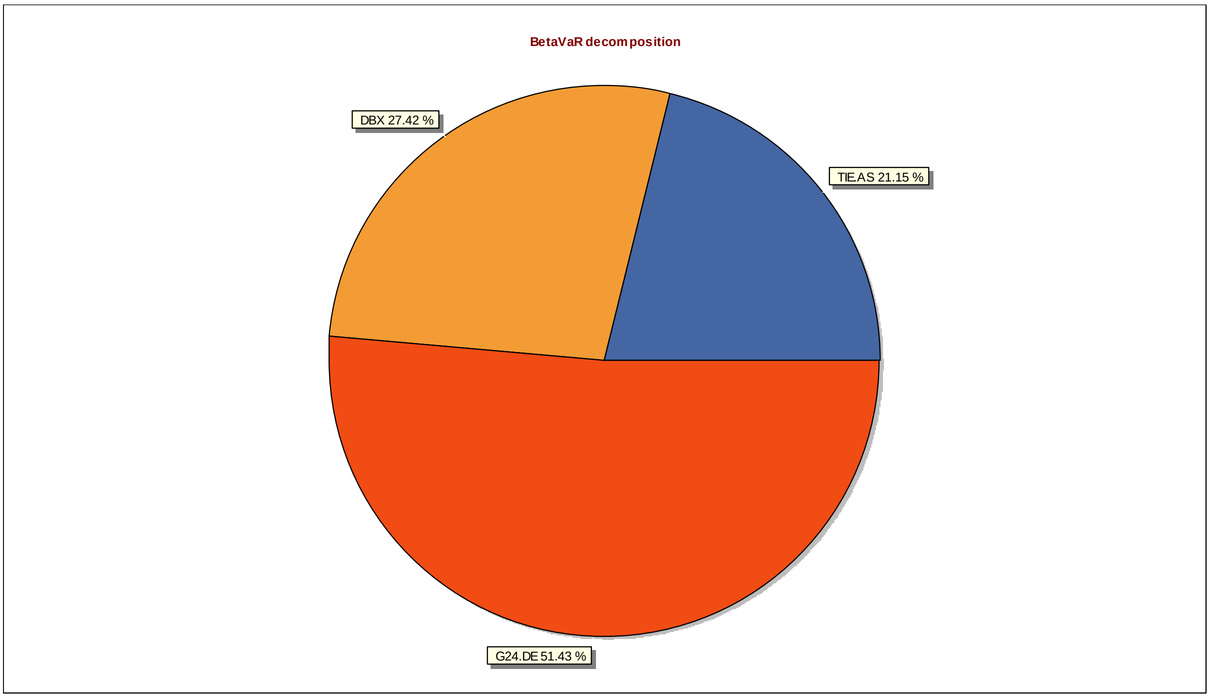


Fig 20. VaR decomposition

Table (20) shows the Beta VaR in percentage and in monetary units, based on the portfolio VaR at the specified date. The sum of all Beta VaRs is equal to the portfolio VaR.

Table 20. Beta VaR in dollars and monetary units

Asset	Position($)	BetaVaR(%)
DBX	24.63	27.424%
G24.DE	61.06	51.430%
TIE.AS	21.60	21.146%

In the following, you can see the component VaR assets that would decrease or increase risk if excluded:

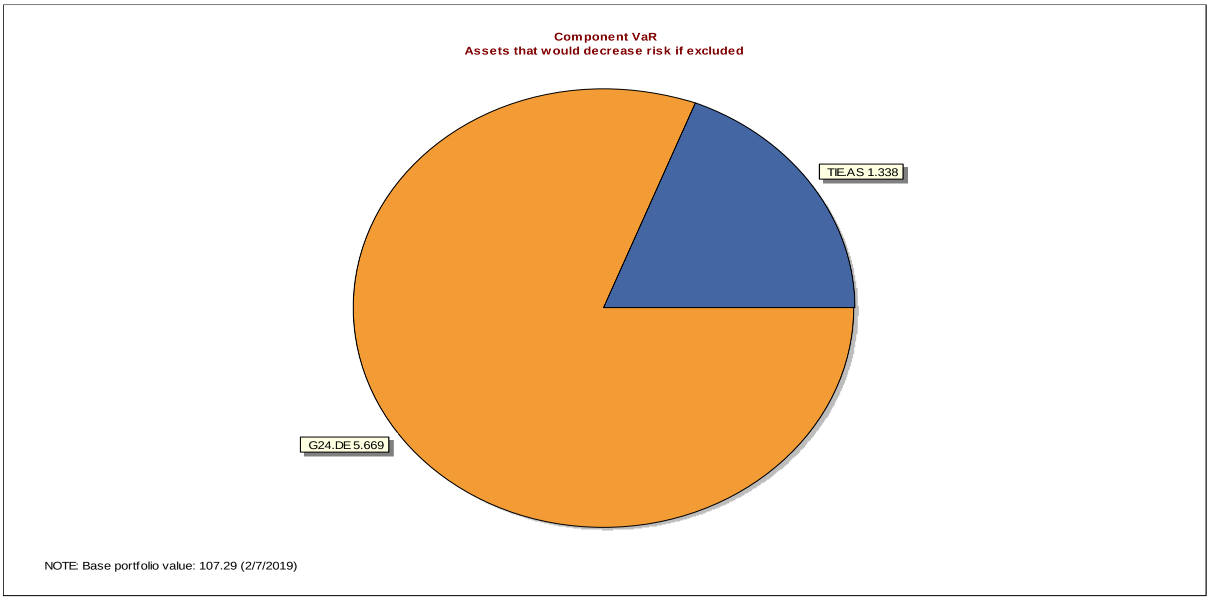


Fig 21. Assets allocation with the goal of decreasing risk

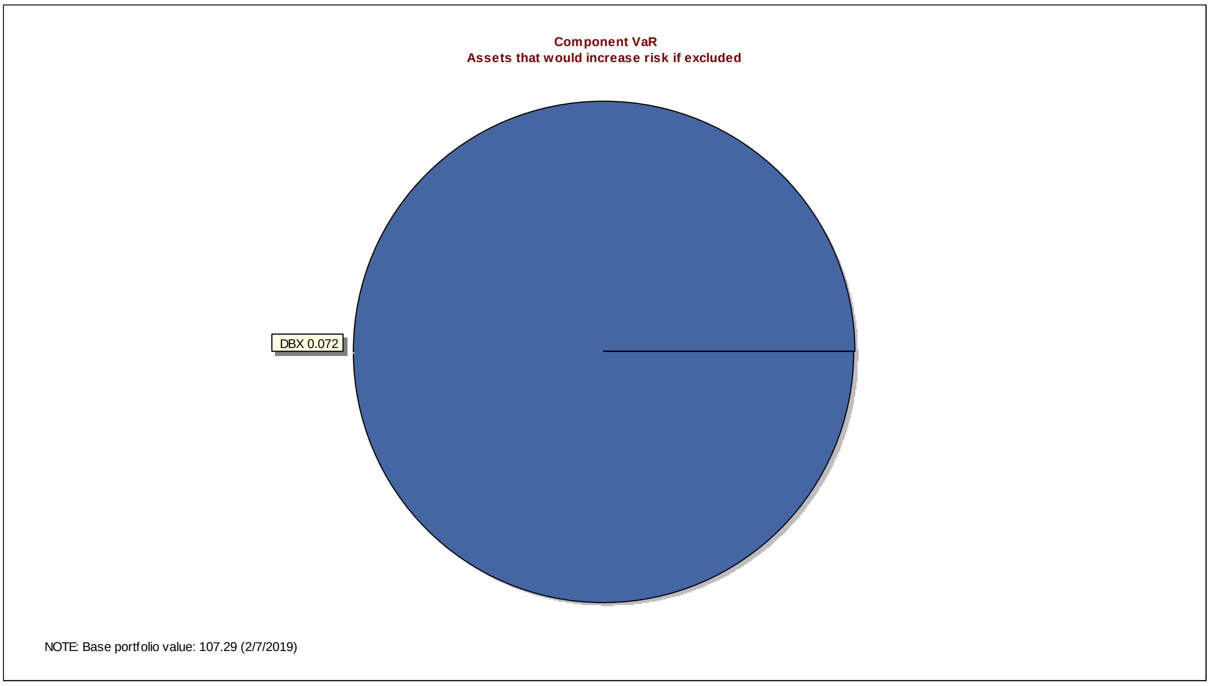


Fig 22. Assets allocation with the goal of increasing risk

Table (21), shows the results in components VaR:

Table 21. The results of components VaR

Asset	CmpVaR(%)	CmpVaR
DBX	-0.067%	-0.07
G24.DE	5.284%	5.67
TIE.AS	1.247%	1.34

4.4.2. CVaR estimation

Like VaR calculation and analysis, the related steps should be passed. VaR parameters and VaR results can be observed in Table (22):

Table 22. C-VaR parameters and results

C-VaR parameters		
CVaR horizon (in days)	7	
Significance level(%)	5(%)	
C-VaR results		
	Returns	Portfolio Value
VaR	-12.825%	-13,76
CVaR-	-17.005%	-18,24
CVaR	-17.033%	-18,27
CVaR+	-17.118%	-18,37
C-VaR relative to mean		
	Returns	Portfolio Value
VaR	-13.580%	-14,57.47
CVaR-	-17.760%	-19,05.26
CVaR	-17.788%	-19,08
CVaR+	-17.873%	-19,18
Base portfolio value to calculate VaR	107,29	

The portfolio is not likely more than 18.27 of value after 7 days following 02/07/2019 with 95% of confidence. The results are shown in percentage and monetary units. Based on the portfolio value on the date selected. The C-VaR is calculated both in absolute terms (actual loss) and of the historical returns.

Figure (23) shows the portfolio periodic returns histogram:

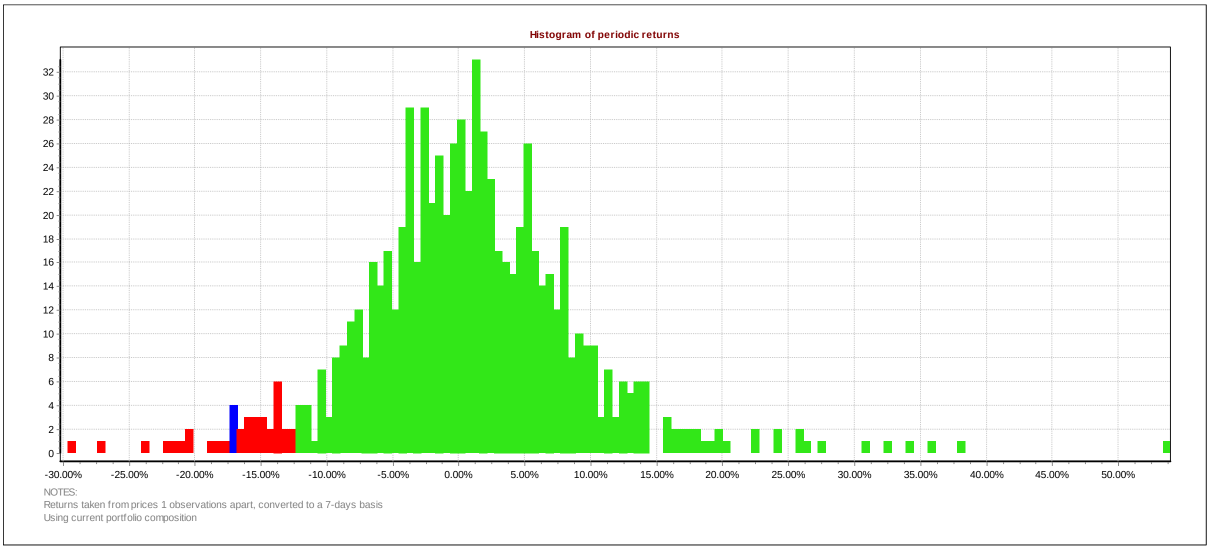


Fig 23. Return histogram

This is a histogram that marks in red the losses below the specified CVaR significance level (the Conditional Value at Risk limit). Figure (24) shows a simulation of the portfolio value (on the right axis) and the C-VaR (in monetary units, on the left axis) along the sample.

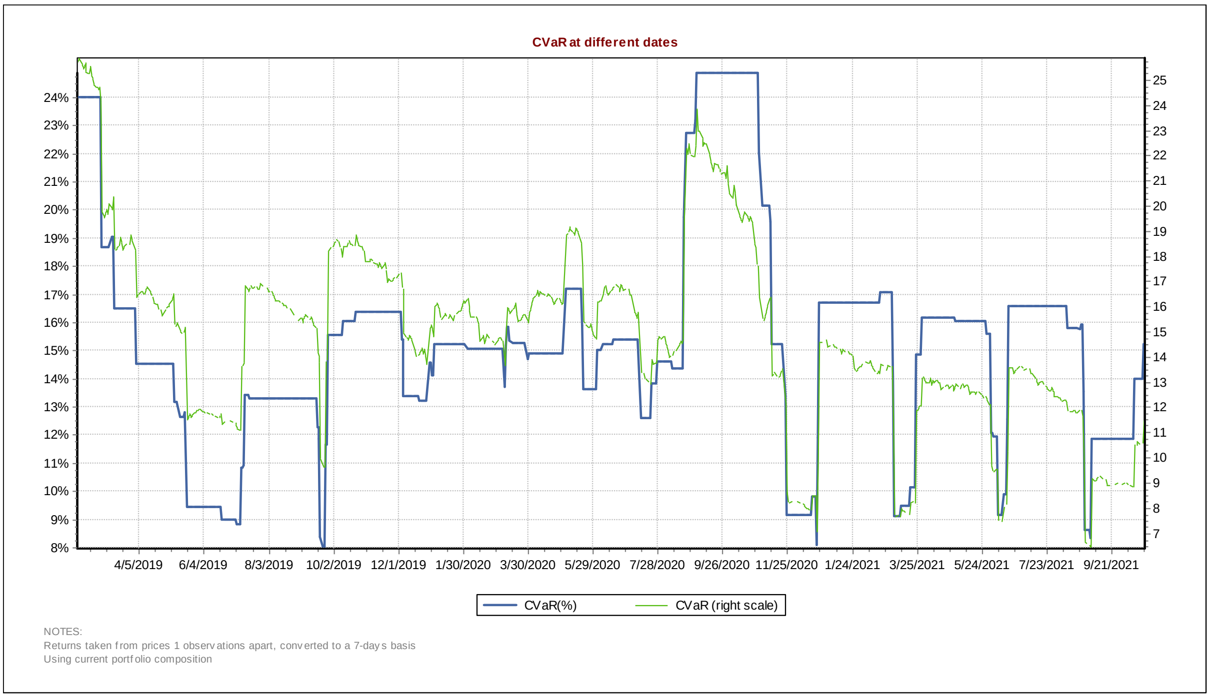


Fig 24. C-VaR at different dates

For testing the results and confidence, back-testing is used:

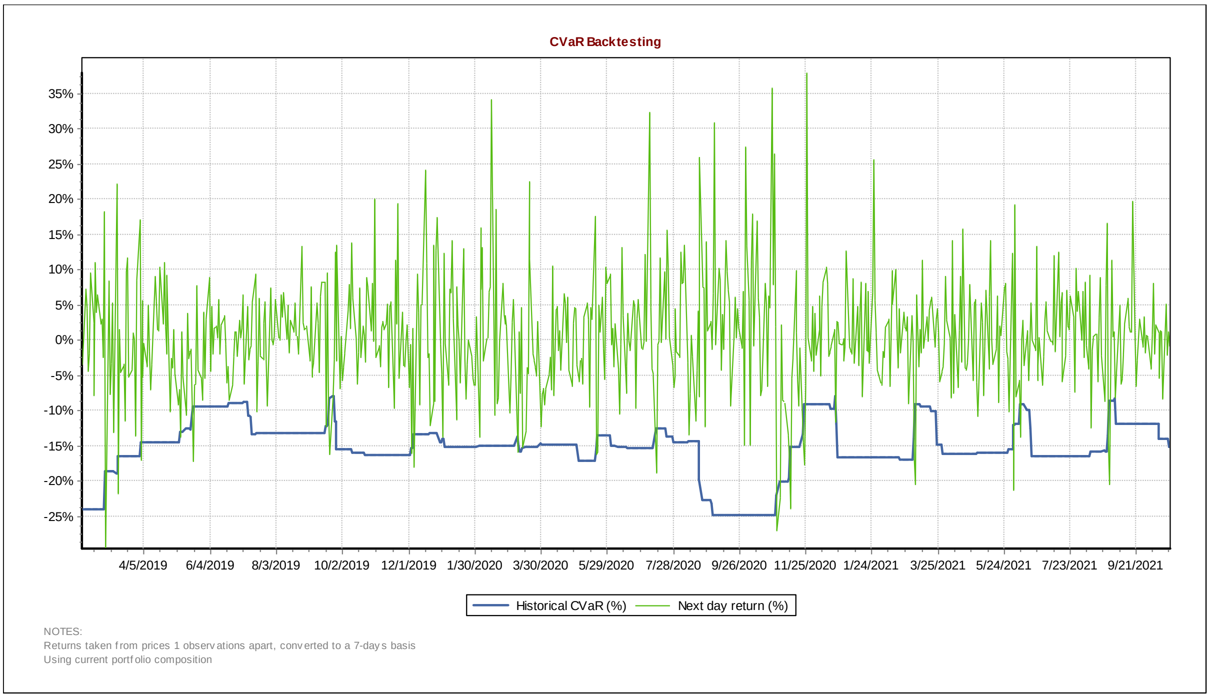


Fig 25. C-VaR back-testing

The empirical alpha is equal 3.541%. In the next steps, we have optimized the C-VaR and portfolio which has the lowest risk and highest return.

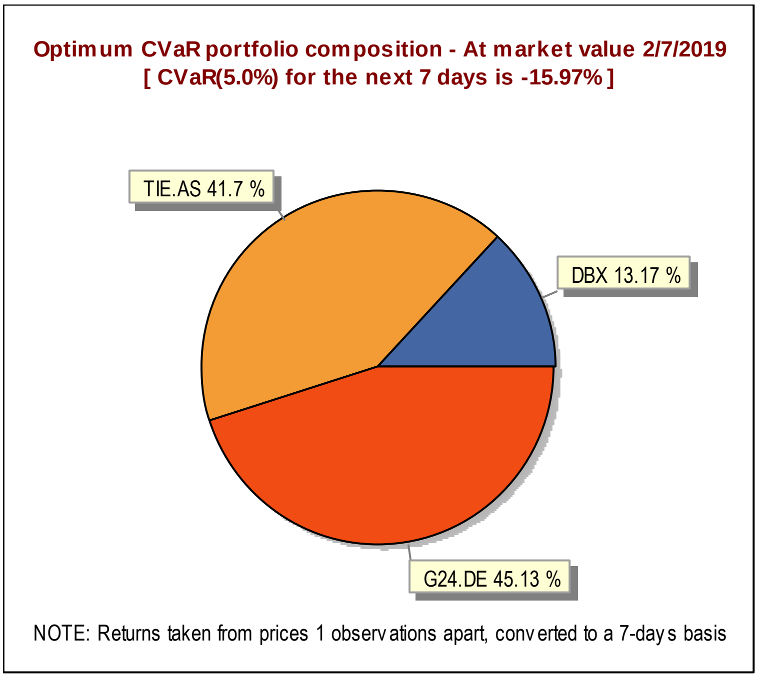


Fig 26. Optimum C-VaR portfolio composition

The method contains multiple steps:

The first step tests 100,000 random portfolios and chooses the best of them to begin the optimization process. Then, the optimizer applies an adaptive gradient-oriented algorithm that improves the precision of the result each time.

The graph shows the weight (in market value) of the most important assets considered into the optimum C-VaR portfolio.

Table (23) presents the number of titles of the current (old) portfolio and the optimum portfolio:

Table 23. number of titles of the current (old) portfolio and the optimum portfolio

Asset	Optimum	Current
DBX	0.6	1.0
G24.DE	0.8	1.0
TIE.AS	2.1	1.0

C-VaR (5.0%) for 7 days from 02/07/2019 is -15.97%.

The last step is about risk/return portfolio simulation. Figure27 shows the current portfolio using C-VaR optimized after 1,000,000 times simulations:

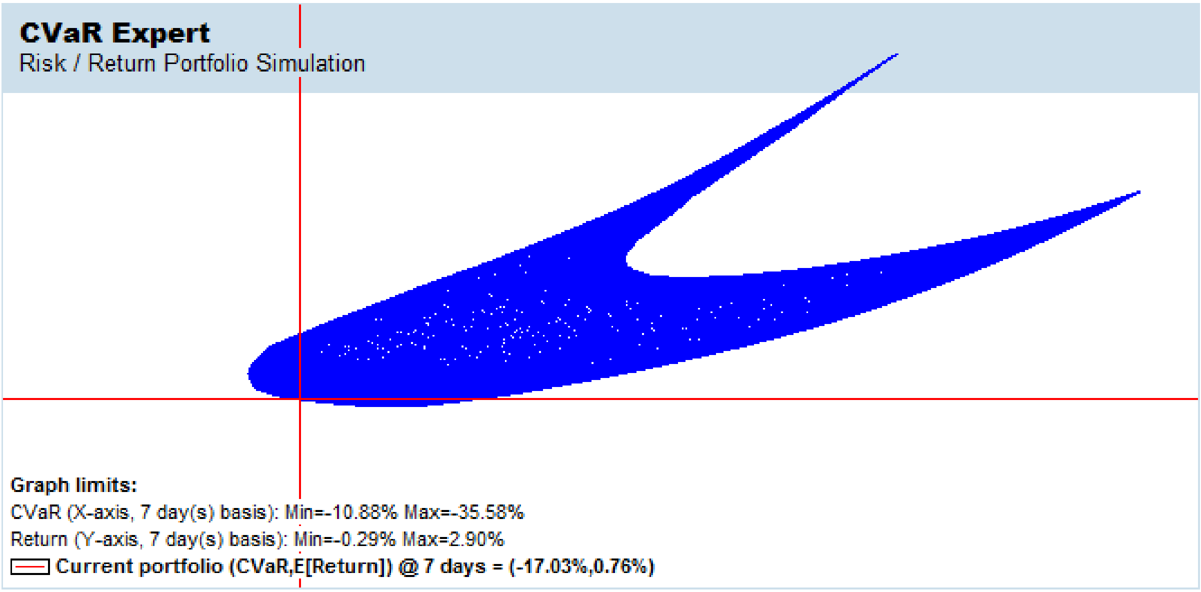


Fig 27. Risk-return portfolio simulation

As it is obvious, X-axis shows the minimum and maximum of C-VaR and Y-axis represents minimum and maximum of return respectively. The min and max of C-VaR and return is -10.88%, -35.58% and -0.29%, 2.90% respectively.

5. Conclusions and results

One of the main factors which facilitating economic growth are startups. Because of their unique characteristics and traits such as innovation, technology, knowledge and etc. they can create value. In these companies or teams, financial issue and funding for going concern and avoid the Death Valley is significant. There are numerous resources that can be helpful: I. family and friends II. Crowdfunding (i.e. people) III. Angel investors IV. Venture capitals and etc. Investors consider multiple factors such as idea, timing, team and etc. in their choices. They do financial analysis to examine whether the company has potential growth or not. One of the main tasks that they do is startup valuation. There are multiple startup valuation models such as Berkus model, DCF model, venture capital method and etc. Because startups mostly do not have information and financial statements at the beginning, we have to predict the growth rate of startups. Startups may be in different stages such as pre-seed, seed, series-A and etc. The last stage is when a startup enters the stock market.

Since, there is not any information about startups such as sale, market size, profit and etc. and most of the models works with database, so, we have tried to analyze startups that are in stock markets and passed IPO stage. In this paper, we have tried to valuate startups using artificial intelligence based model like artificial neural network (ANN) and genetic algorithm (GA) and econometric models such as regression analysis. GA used as feature selection and ANN used to find optimal solution. Finally, we make a portfolio of these three companies means Dropbox, Inc. (DBX), Scout24 SE (G24.DE) and TIE.AS and optimized it using Value at Risk (VaR) and Conditional Value at Risk (C-VaR) based on risk and return. The results showed that if you want to increase your return and risk, you would better invest in G24.DE and DBX respectively.

We found that artificial intelligence based models having high predictability based on the following characteristics:

- Speed up calculations
- Improve by training
- No assumption
- Ease of use

But econometric models have some qualifications and assumptions such as normality, linearity, stationarity and etc. which are the limitation.

As recommendations and remarks for future researches, AI based models such as ANN may face with a situation which called local minima or maxima trap. to avoid, there are solutions.

- One of them is using meta-heuristic algorithms as optimization algorithms. These algorithms can increase the capability of the network such as exploitation and exploration. So, you can increase the search space and increase your chance and speed to find the optimal solution.
- You can do a widespread literature review and finding the most important indicators in startups which constitute their value. By doing this, you can increase your model predictability with high R-squared and estimation error.

References

- Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of pharmaceutical and biomedical analysis, 22(5), 717-727.
- Akkaya, M. (2020). Startup valuation: Theories, models, and future. In Valuation Challenges and Solutions in Contemporary Businesses (pp. 137-156). IGI Global.
- Baltagi, B. H. (2011). What Is Econometrics? In Econometrics (pp. 3-12). Springer, Berlin, Heidelberg.
- Bjørnskov, C., & Foss, N. J. (2016). Institutions, entrepreneurship, and economic growth: what do we know and what do we still need to know? Academy of Management Perspectives, 30(3), 292-315.
- Boudieb, D., Mohammedi, K., Bouziane, A., & Smaili, Y. (2011). OPTIMIZATION TECHNIQUES BY DARWIN'S THEORY OF EVOLUTION. International Journal of Arts & Sciences, 4(19), 131.
- Chatterjee, S., & Hadi, A. S. (2013). Regression analysis by example. John Wiley & Sons.
- DeJong, K. Analysis of the Behavior of a Class of Genetic Adaptive. Ph.D. Thesis, University of Michigan, Ann Arbor, MI, USA, 1975.
- de Oliveira, F. B., & Zotes, L. P. (2018). Valuation methodologies for business startups: a bibliographical study and survey. Brazilian Journal of Operations & Production Management, 15(1), 96-111.

- Dhochak, M., & Doliya, P. (2020). Valuation of a startup: Moving towards strategic approaches. Journal of Multi-Criteria Decision Analysis, 27(1-2), 39-49.
- Edison, H. (2020). Lean Internal Startups: Challenges and Lessons Learned. In Fundamentals of Software Startups (pp. 251-268). Springer, Cham.
- Hidayat, S. E., Bamahriz, O., Hidayati, N., Sari, C. A., & Dewandaru, G. (2021). Value drivers of startup valuation from venture capital equity-based investing: A global analysis with a focus on technological factors. Borsa Istanbul Review.
- Hsu, D. K., Haynie, J. M., Simmons, S. A., & McKelvie, A. (2014). What matters, matters differently: a conjoint analysis of the decision policies of angel and venture capital investors. Venture Capital, 16(1), 1-25.
- Hyrkäs, A. (2016). Startup complexity: tracing the conceptual shift behind disruptive entrepreneurship. Publications of the Faculty of Social Sciences.
- Jedlickova, M., & Kutnar, P. (2017, October). Construction of a fuzzy model for the success prediction of hi-tech companies with a short history. In International Conference at Brno University of Technology, Faculty of Business and Management.
- Laitinen, E. K. (2019). Discounted Cash Flow (DCF) as a measure of startup financial success.