Applying the Random Forest Model to Forecast the Market Reaction of Start-up Firms: Case Study of GISA Equity Crowdfunding Platform in Taiwan

CHENG-SHIAN LIN
CTBC Business School, Tainan, Taiwan.
No.600, Sec. 3, Taijiang Blvd., Annan District, Tainan, TAIWAN

CHUN-YUEH LIN,
CTBC Business School, Tainan, Taiwan.
No.600, Sec. 3, Taijiang Blvd., Annan District, Tainan, TAIWAN

SAM REYNOLDS,
Taipei-based Technology Analyst Formerly with International Data Corporation,
Taipei, TAIWAN

Abstract: - In 2015, Taiwan introduced an exchange platform for equity crowdfunding called the Go Incubation Board for Startup and Acceleration (GISA) which is supervised by the OTC Taipei Exchange organization. Equity crowdfunding provides another channel for startups to access capital and allows for a new mechanism for start-up firms to establish their reputation with investors. However, the risks to investors from equity crowdfunding are high. The high-risk nature of equity crowdfunding has the potential to act as a contagion, and further erode confidence in the startup capital market by retail investors -- and this lingers over the GISA platform in Taiwan. Therefore, this study applies the of Random Forest (RF) algorithm to evaluate the market reaction for start-up firms on the GISA in Taiwan. The RF algorithm is proposed to be integrated into an AI model to forecast the market reaction to start-up firms as they get listed on the GISA equity crowdfunding platform. The results not only fulfill the gap of detecting market reaction in equity crowdfunding, but the proposed RF model can replace the traditional statistics analytical technique to evaluate the market reaction. In proposed model applied AI algorithms to predict the market reaction on Taiwan GISA platform which can provide a useful ensemble tool for start-up firms and entrepreneurs to evaluate the degree of market reaction more efficiently before listing on the Taiwan GISA platform.

Key-words: - Financial Technology (FinTech); equity crowdfunding; machine learning; GISA platform; random forest model.

Received: October 27, 2019. Revised: April 16, 2020. Accepted: April 23, 2020. Published: April 27, 2020.

1. Introduction

With the rise of Financial Technology (FinTech), crowdfunding has become more popular. The crowdfunding market has been growing fast in this era. Crowdfunding platforms include several different forms: Donation-based crowdfunding, Rewards-based crowdfunding, Debt-based crowdfunding, and Equity-based crowdfunding [6, 29, 42]. The first model is donation-based which can collect charitable funding in support of causes and projects. The

second model is rewards-based that means investors receive non-monetary rewards in exchange for their contribution. The third model is debt-based which provides a credit contract regarding the benefits policy between funders and fundraisers. The last model is equity-based that offers an equity stake in the target company [6, 29, 42].

Title III of the Jumpstart Our Business Startups (JOBS) act introduced equity crowdfunding to the United States [23]. Rossi and Walthoff-Borm [46, 59] proposed the crowdfunding has emerged as a new financing tool alongside more traditional means of financing new ventures. Therefore, the crowdfunding market has been growing fast in this era. Agrawal [4] indicated the EC has e allowed the matching of demand and supply of early-stage finance across a wider geographical area EC is also one of the approaches in external financing activity for start-up firms and EC is the important financing alternative for early-stage financing in small medium enterprise [12, 19, 59].

In 2015, equity crowdfunding was introduced to Taiwan. The best-known equity-crowdfunding platform in Taiwan is Go Incubation Board for Startup and Acceleration (GISA) and which is implemented and supervised by the OTC Taipei Exchange organization. As an OTC Exchange, the Taipei Exchange is entirely focused on helping small firms grow and providing the means to get the capital to help them his the next chapter of their development. The Taipei Exchange [53]. The Taipei Exchange [53] reports that they have about 420,000 enterprises (61%) with paid-in capital between NT$1,000,000 and NT$10,000,000 and 120,000 enterprises (18%) with paid-in capital between NT$10,000,000 and NT$50,000,000 on December 31, 2017. The advantages of start-up firms to apply for a listing on GISA fall into have four categoriesnitial Public Offering (IPO) is not necessary, free counseling, raising capital with fewer barriers and increasing the operating scale and publicity [53]. The procedures of registration on GISA platform include six steps that are applying for GISA (APF), reviewing innovation, creative opinion and comprehensive examination (RICCE), providing integrative counseling (PIC), examining the counseling results and companies' qualification (CRCQ), capital raising on GISA before registration (CRGBR) and registering on GISA (ROG). Moreover, they can get exposure to the public and retail investors with a listing on the platform. (See Figure 1).

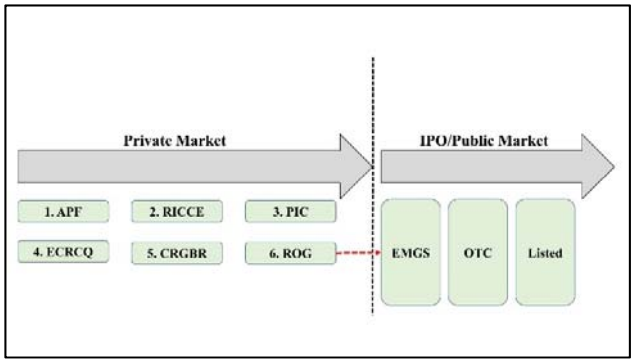

Figure 1. The processes of GISA registration
Source: Taipei Exchange (2019)
APF: Applying for GISA registration
RICCE: Reviewing innovation, creative opinion and comprehensive examination
PIC: Providing integrative counseling
ECRCQ: Examining the counseling results and companies' qualification
CRGBR: Capital raising on GISA before registration
ROG: Registering on GISA
EMGS: Emerging stocks companies
OTC: Over-the-counter companies
Listed: Listed companies

While start-up firms can efficiently access capital through GISA, the metric that is important to track is the "equity subscription rate (ESR)", that represents the market reaction of start-up firms' know-how or idea. It must achieve the 100% when they want to list on the GISA. Otherwise, if the ESR does not achieve 100%, it means the market reaction is bad then start-up firms must withdraw the registration process. For example, if start-up firms would like to make their financing by GISA equity crowdfunding platform, they have to pass the all of review procedures on GISA platform (See Figure 1 and Figure 2). In particularly, it include a threshold (ESR>100%) between step 5 and step 6 which means that have a market reaction testing in the platform. Therefore, they can list on the GISA platform and make their financing if start-up firms achieve this limitation. Otherwise, the GISA review process would reject their registration. In addition, the difficulty is higher from GISA to IPO of start-up firms when the market reaction is low in pre-listing on the GISA platform. This risk not only extends the financing cycle but also that would be enhanced the threat in urgency of working capital from start-up firms (see Figure 2). Therefore, the market reaction of start-up firms' know-how or idea when they want to list on GISA is very important. Past evidence on the market reaction are concentrate on the stock market and open market [50, 13, 27] and there is few evidence on assessing the market reaction in pre-registering on equity crowdfunding platforms. Hence, this study focuses on the evaluation of market reaction in pre-register equity crowdfunding (such as GISA) for start-up firms at Taiwan.

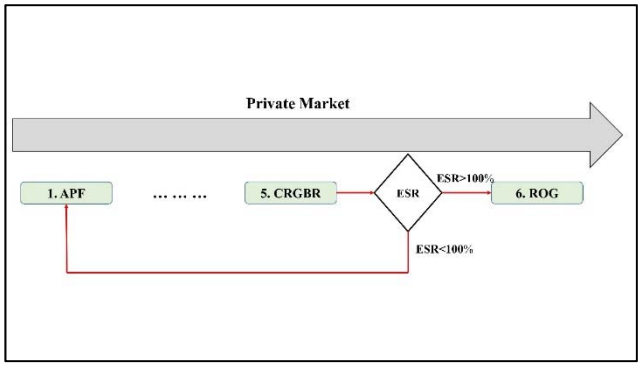

Figure 2. The process of ESR
ESR: Equity subscription rate

The rapid development of artificial intelligence (AI) and computing, many industries have implemented AI models for improving the efficiency of solving problems [24, 37, 1, 36, 17]. Thus, AI has become a multidisciplinary and interdisciplinary of natural sciences and social sciences consisting of diversified disciplines [14; 32, 56]. Nevertheless, previous studies in evaluation of market reaction are concentrated on the statistics technique [9, 52, 62, 57, 11, 18, 58]. Dangeti [16] proposed that the statistics analytical methods require assuming the shape of the model curve priori to performing model fitting on the data, whereas AI models do not need to assume the underlying shape, as machine learning algorithms can learn complex patterns automatically based on the provided data. Hand [22] indicated that statistics play a significant role in AI. Nevertheless, larger datasets and secondary data are more commonly used in AI model, as opposed to statistics. Moreover, AI techniques do not usually require particular assumptions (e.g. independence of variables) regarding the dataset, which often limit the use of parametric statistical tests [7]. Previous studies have implemented the AI model to replace the mathematical model of traditional statistics techniques for analyzing the topics in economics and finance. The results have the high reference

value in the practical application. [61, 47, 26, 20, 21].

Based on the above, the equity crowdfunding in Taiwan has to meet a requirement for start-up firms to list on GISA that is "equity subscription rate (ESR)", it means that "market reaction". If the market reaction of start-up firms or plans achieved the requirement of GISA platform which can fundraising by the GISA. Past literatures on market reaction issue were focused on the stock market and open market [50, 13, 27]. Moreover, these studies developed the analysis model by the traditional statistics techniques. The barrier of traditional statistics technique is assuming underlying shape for evaluating the data.

Therefore, this study adopts the ensemble learning AI algorithm of RF model to evaluate the market reaction for start-up firms pre-listing on GISA in Taiwan as the platform that can predict the degree of market reaction in start-up firms. The proposed AI model to forecast the market reaction of start-up firms before listing on the Taiwan GISA equity crowdfunding platform. The results not only fulfill the gap of detecting market reaction in equity crowdfunding but the proposed RF model which can replace the traditional statistics technique to evaluate the market reaction then obtain more available forecasting efficiency. From the commercial side, the proposed model applied AI algorithms to predict the market reaction in Taiwan GISA platform that can provide a useful ensemble tool for start-up firms and entrepreneurs to evaluate the degree of market reaction more efficiency before listing on the Taiwan GISA platform.

2. Random forest model and evaluation indicator

2.1 Random forest model

Breiman [3, 8, 31] proposed the random forest (RF) algorithm and it belongs to the ensemble-learning algorithm in the machine learning area based on the decision tree model. The decision tree is to obtain a structure for predicting the target variable and the application is suitable for many fields [44, 45, 48, 54, 55]. Prasad [41] considered that the principal of ensemble learning is to construct and integrate multiple base learners to achieve better generalization capabilities. By randomly extracting variables and sample data, the algorithm generates many classification trees, and then aggregates the results of classification trees, which is the RF model [25, 33, 38, 43]. Pan & Zhou [38] indicates the RF model improves prediction accuracy without significantly increasing the amount of computation workload with the neural network, support vector machine, decision tree and Adaboost model, and is not sensitive to multivariate collinearity. Furthermore, the RF model is robust for modeling of missing data and unbalanced data in the machine learning methodologies field. The RF model adopts the classification and regression tree as the base learner that is one of the algorithms of decision tree [39, 40, 63]. Acharjee [2] proposes the procedures of RF as follows: each decision tree is constructed from a bootstrap sample of the calibration dataset, containing about two thirds of the sample. Elements not included are referred to as out-of-bag (OOB) data. At each node the un-pruned decision tree is grown at each sample, one third of the predictor variables are randomly selected and the best split is chosen according to the lowest Gini index [10]. At each bootstrap

iteration, the response value for OOB data is predicted and averaged over all trees. The splitting process is repeated in each tree until a predefined stop condition is reached [15, 63], and then applies the average method or voting method to combine the prediction results of multiple decision trees to determine the RF prediction result [60]. Therefore, currently, RF is regarded as one of the best framework for this purpose [30]. Lu [30] also indicates the RF model can handle high dimensional data of many features, does not have to select the features and is adaptable for the database and it can not only deal with discrete data but also deal with continuous data without standardization. The concept of RF is shown in Figure 3.

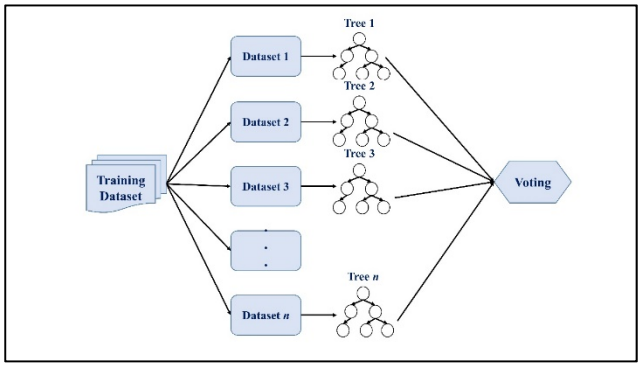

Figure 3. The concept of RF

According to Pan & Zhou [38] the RF is performed by the growth of decision trees related to the random vector θ. They assume that the training set is extracted independently from the distribution of random vectors *Y* and *X*. Let $\delta(x)$ represents the regression result of single decision tree, and then the predicted value of random forest $\beta(X,\theta_i), i = 1,2,\dots,n$ is acquired by the averaging regression results of *n* decision trees.

$$\beta(x) = \frac{1}{n}\sum_{i=1}^{n}\delta_i(X) \quad (1)$$

Where $\beta(x)$ represents the result of combined regression models. The RF is to acquire different sample sets by the method of bootstrap resampling. Hence, the RF adopts bootstrap sampling to extract *n* samples from the original training data, develops decision tree models for *n* samples, and obtains *n* classification results and votes on each sample to evaluate its final classification based on the *n* classification results.

$$\beta(x) = \arg\max_{Y}\sum_{i=1}^{n} I(\delta_i(X) = Y) \quad (2)$$

Where $\beta(x)$ represents the result of combined classification models. $\delta_i(X)$ represents the result of single decision tree. *Y* denotes the output factor. $I(\cdot)$ is a linear function.

With the number of decision tree classifications increase, the generalization error of decision trees in all forests converges to eq. (3):

$$\gamma = P_{xy}(P_\theta(\beta(X,\theta) = Y) - maxP_\theta(\beta(X,\theta) = J) < 0) \quad (3)$$

$$\gamma^* = P_{xy}(f(X,Y) < 0)$$

$$f(X,Y) = av_n I(\delta_i(X) = Y) - av_k I(\delta_i(X) = j)$$

Where *n* is the number of decision trees in the RF model. The generalization error γ will reduce when the decision tree increases.

The RF model is a kind of ensemble learning method that is developed by growing a certain number of decision trees [10, 34, 35]. The bootstrap sampling is to randomly select the samples from the original database with replacement to acquire new datasets of the same size. Selected data samples are called in-bag data, which are applied to train in decision tree model. The unselected data are called OOB data and they are not involved in the training of the RF model. Consequently, the OOB data can be applied as the validation data to predict the generalization error and thus to quantify the training accuracy [11]. It can be evaluated that the prediction result

through majority voting of decision tree classifiers [8]. Suppose the original sample size is *N*. The sample feature dimension is *M*, and the number of decision trees in the random forest model is *p*. The specific modeling steps are as follows [28, 11]:

(1) Constructing *p* decision trees from the original data through bootstrap method.

(2) Randomly selection of the *m* features in the *M* dimension as training for different decision trees, and $m < M$.

(3) Grow a decision tree for each training subset by the CART algorithm without pruning.

(4) Adopts the trained decision trees to predict the OOB validation samples. The final prediction results of an OOB validation sample is determined by the majority votes of the predication from all the decision trees that are grown without using the sample.

(5) The OOB error is estimated by the percentage of the wrongly predicted OOB data, while the training accuracy of the RF model is the percentage of the correctly predicted OOB data samples.

2.2 Model evaluation indicator

The performance evaluates such as accuracy, precision, recall, F1 score depend on the four evaluates True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) (See Table 1). TP ratio is evaluated by the correctly market reactions category to the total number of market reactions category in the dataset. TN ratios are the results that are correctly recognized. FP and FN ratio are the results that wrongly classified to be a part of wrong class. Based on the above evaluates, the precision, recall and F1 score are computed.

Precision-recall is employed here to measure the performance of classification, more specifically, the performance of classification of each class and generalization of our classifier. Furthermore, the synthesized F1 score is able to measure the performance of classification by taking precision and recall into consideration. The calculation of evaluates are as follows:

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

$$Precision = \frac{TP}{(TP+FP)} \quad (5)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (6)$$

$$F1\ score = \frac{2*Precision*Recall}{Precision+Recall} \quad (7)$$

Table 1. Confusion matrix

	Low (Actual)	High (Actual)
Low (Predicted)	TP	FP
High (Predicted)	FN	TN

3. Case Study

3.1 WEKA

WEKA is a data mining software developed by the University of Waikato in New Zealand that constructs data mining algorithms using JAVA language. Russell & Markov [49] indicated the WEKA system provides a rich set of powerful machine learning algorithms for data mining tasks, along with a comprehensive set of tools for data pre-processing, statistics and visualization, all available through an easy to use graphical user interface. In addition, Sewaiwar & Verma [51] proposed that WEKA implements algorithms for data pre-processing, classification, regression, clustering and association rules and it is not only affording a toolbox of learning algorithms, but

also a framework inside which researchers could implement new algorithms without having to be concerned with supporting infrastructure for data manipulation and scheme evaluation.

Therefore, the application of RF model is implemented by WEKA software in this study for detecting the market reaction of start-up firms in equity crowdfunding on the Taiwan GISA platform. The interface of the WEKA GUI and interface are shown in Figure 4 and Figure 5. The flowchart of building RF model is shown in Figure 6.

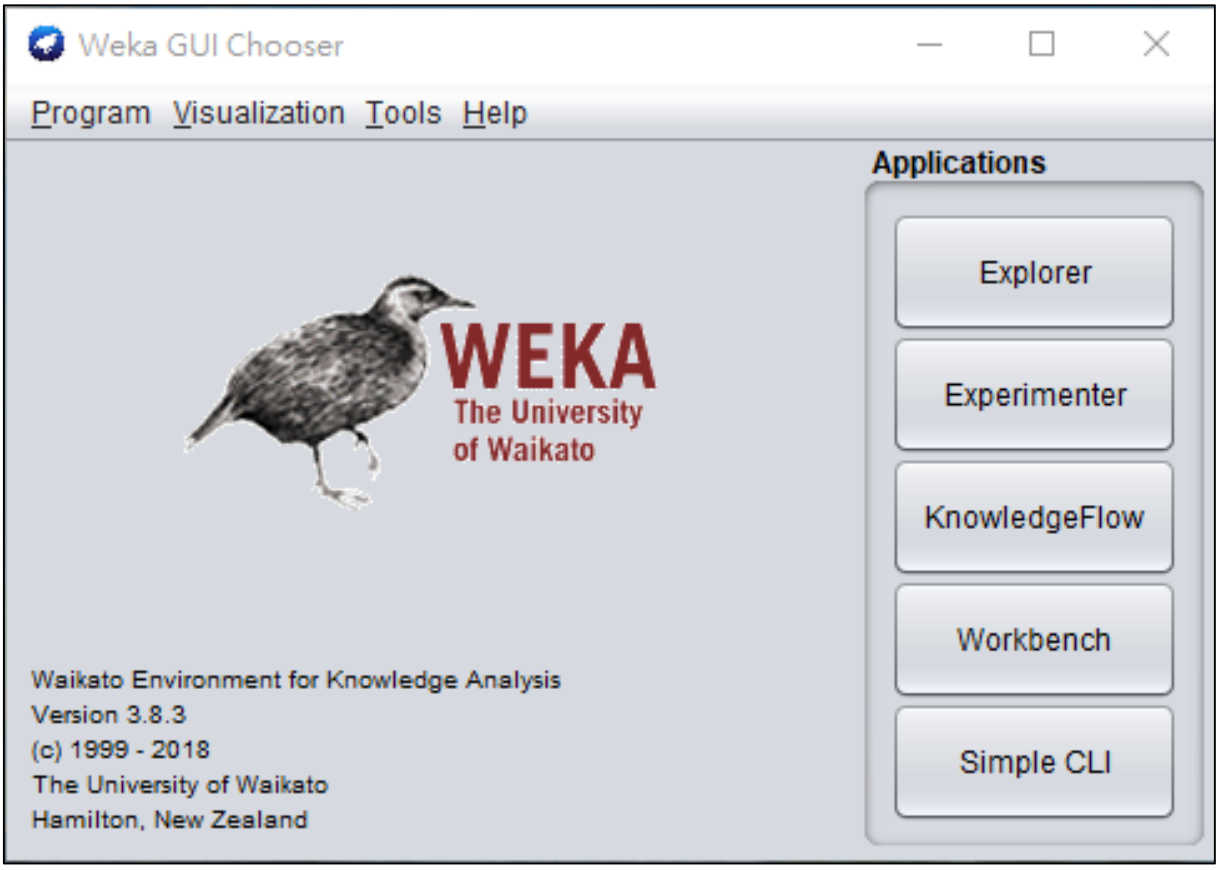

Figure 4. WEKA GUI

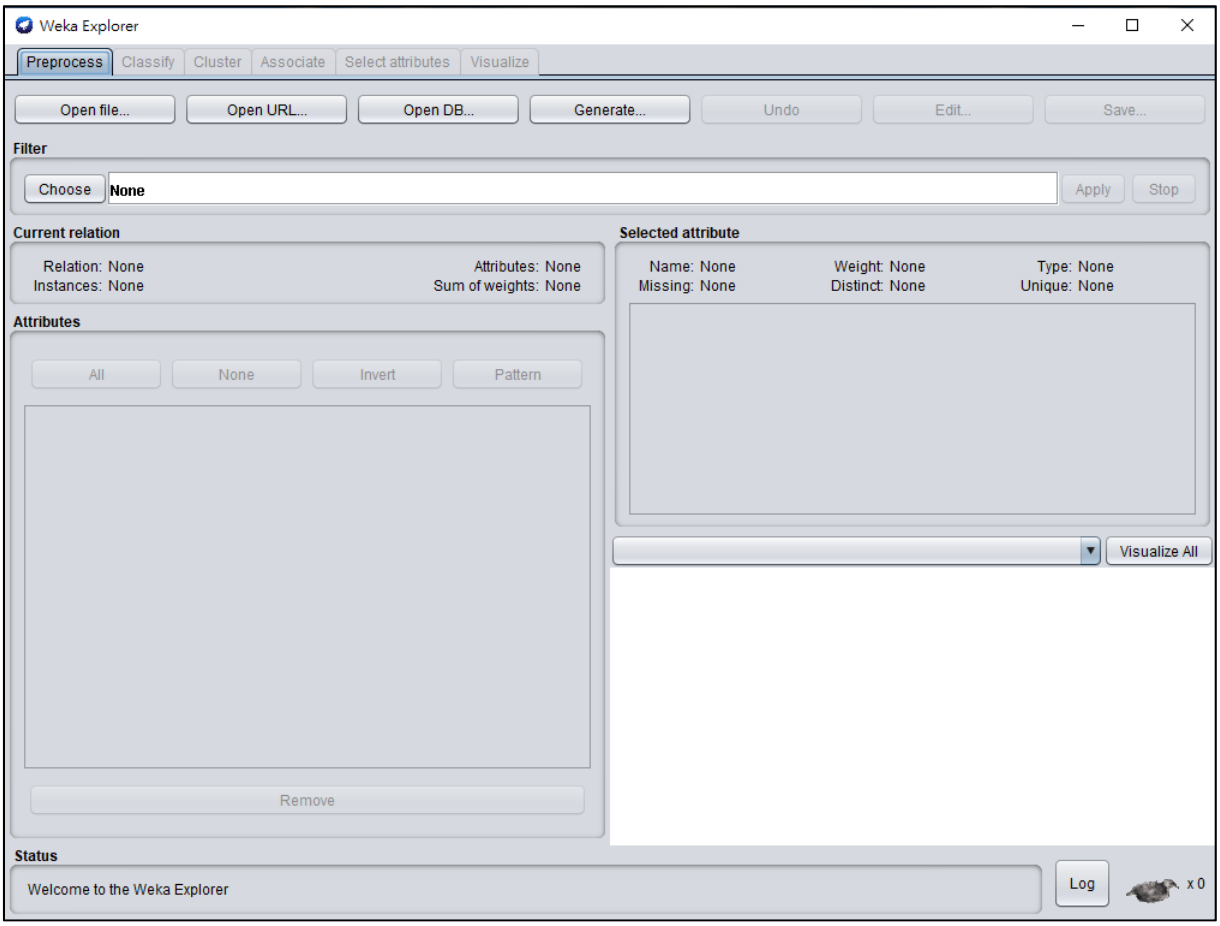

Figure 5. WEKA interface

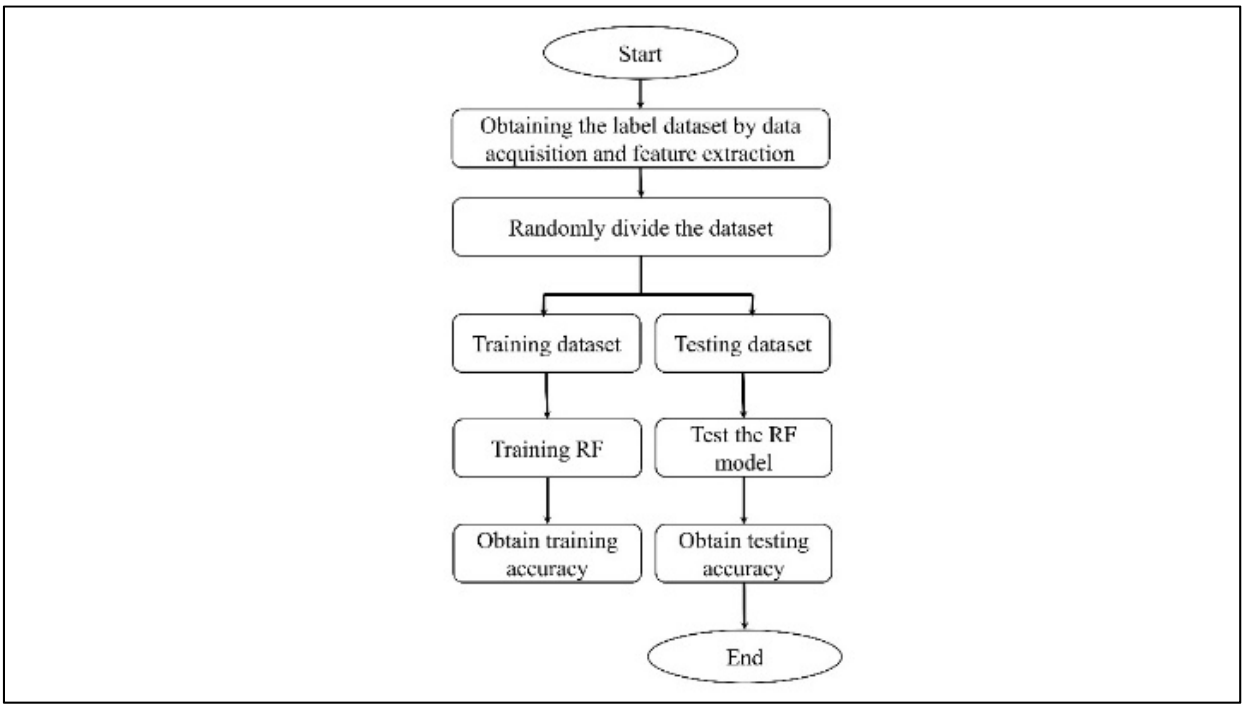

Figure 6. Flowchart of building RF model

3.2 Data Sources and Data Preprocessing

In this study, start-up firms are analyzed in the research period from January 1st, 2016 to December 31st, 2018. The major data source is the listing data from the GISA, Taipei Exchange. The testing data is from GISA by random sampling from January 1st, 2019 to June 31st, 2019 in 20 data samples.

Because the characteristics of equity crowdfunding are opaque. Start-up firms have to provide the details of the profit and loss account (P&L) Statement and balance sheet (BS) when they will list on the GISA platform in This study implemented the all factors from the profit and loss account and balance sheet. As known to all, the evaluation of the start-up firms' market reaction, which consists of the evaluation of the financial structure and operation structure that include the P&L statement and BS. The data of start-up firms that enroll in the evaluation will be preprocessed. With such purpose, the format of the data from the GISA must be unified. After data cleaning, data conversion, data integration, and data filtering, type and value of the data can be converted. During this process, it is necessary to reduce the unrelated information and noise, so that the factors which affect the market reaction of start-up firms on the GISA platform

Accordingly, discretizing the original dataset of the features and target parameter in this study is shown on Table 2 which include 126 start-up firms on the GISA platform. The descriptive statistics data is shown in Table 3 and the discretized data of each feature and target parameter is shown in Table 4.

Table 2. The original data type of features and target parameter

	Features	Type
Features	Location	Nominal
	Industry	Nominal
	Current Assets	Numeric
	Non-current Assets	Numeric
	Total Assets	Numeric
	Current Liabilities	Numeric
	Non-Current Liabilities	Numeric
	Total Debts	Numeric
	Total Capital	Numeric
	Additional Paid in Capital	Numeric
	Retained Earnings	Numeric
	Other Equity Interest	Numeric
	Total Equity	Numeric
	Net Sales	Numeric
	Operating Costs	Numeric

	Gross Profit	Numeric
	Operating Expenses	Numeric
	Operating Income	Numeric
	Total Non-Operating Income	Numeric
	Pre-Tax Income	Numeric
	Income Tax Expense	Numeric
	Profit (loss)	Numeric
	Other Comprehensive Income	Numeric
	Total Comprehensive Income	Numeric
Target	Equity Subscription Rate	Numeric

Table 3. The descriptive statistics of features and target parameter (Unit: Thousand)

Features	Max	Min	Mean	Variation	Standard Deviation
Location	Null	Null	Null	Null	Null
Industry	Null	Null	Null	Null	Null
Current Assets	10,620,706.000	319.000	134,660.960	883,579,163,227.149	939,988.917
Non-current Assets	24,775,243.000	0.000	244,972.310	4,822,271,537,385.450	2,195,967.108
Total Assets	35,395,949.000	347.000	379,633.063	9,827,490,130,756.890	3,134,882.794
Current Liabilities	8,179,878.000	16.000	97,875.660	528,837,922,538.374	727,212.433
Non-Current Liabilities	324,900.000	0.000	17,674.313	1,979,607,028.658	44,492.775
Total Debts	8,504,778.000	16.000	114,801.746	568,712,669,091.412	754,130.406
Total Capital	50,000,000.000	3,233.000	466,270.095	19,637,796,067,994.400	4,431,455.299
Additional Paid in Capital	12,000,000.000	-1,912.000	104,640.508	1,132,395,543,865.440	1,064,140.754
Retained Earnings	114,466.000	-333,473.000	-26,672.611	2,498,091,999.904	49,980.916
Other Equity Interest	8,550.000	-35,108,829.000	-279,405.175	9,704,775,953,492.830	3,115,248.939
Total Equity	26,891,171.000	-24,026.000	264,831.413	5,678,119,748,374.480	2,382,880.557
Net Sales	7,735,006.000	0.000	124,537.651	470,697,233,615.894	686,073.781
Operating Costs	14,252,529.000	0.000	156,977.214	1,593,510,943,879.720	1,262,343.433
Gross Profit	241,039.000	-6,517,523.000	-32,432.833	337,471,304,941.440	580,922.805
Operating Expenses	13,393,005.000	898.000	131,526.492	1,407,936,035,384.120	1,186,564.805
Operating Income	49,729.000	-19,910,528.000	-163,951.270	3,119,895,821,018.180	1,766,322.683
Total Non-Operating Income	899,339.000	-16,752.000	7,643.183	6,386,357,277.990	79,914.687
Pre-Tax Income	60,475.000	-19,011,189.000	-157,793.952	2,867,034,409,831.020	1,693,231.942
Income Tax Expense	11,644.000	-6,893.000	142.262	2,693,966.796	1,641.331
Profit (loss)	48,831.000	-19,004,296.000	-156,445.159	2,842,415,577,219.020	1,685,946.493
Other comprehensive income	14,415.000	-712.000	114.079	1,650,256.692	1,284.623
Total comprehensive income	48,882.000	-19,004,296.000	-156,372.913	2,842,439,127,397.510	1,685,953.477
Equity Subscription Rate	550.000%	97.960%	174.360%	102.795%	101.388%

Table 4 The discretized data of each feature and target parameter

Features	Discretized type	Classification Method
Location	Taipei City, Tainan City, Taichung City, New Taipei City, Pingtung County, Miaoli County, Kaohsiung City, Hsinchu County, Hsinchu City, Chiayi County, Changhua County, Taoyuan City	Null
Industry	Information Technology; Cultural Innovation; Biotechnology; Agriculture, Forestry, Fishery and Animal Husbandry; Social Enterprise; E-commerce; Others	Null
Current Assets	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Non-current Assets	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Total Assets	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Current Liabilities	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Non-Current Liabilities	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Total Debts	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Total Capital	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Additional Paid in Capital	High, Low	Classified by mean
Retained Earnings	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Other Equity Interest	Positive, Null, Negative	Classified by the original data
Total Equity	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Net Sales	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Operating Costs	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Gross Profit	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Operating Expenses	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Operating Income	Positive, Negative	Classified by the original data
Total Non-Operating Income	High, Medium, Low, Extremely Low	Classified by interquartile range and quartile deviation
Pre-Tax Income	Positive, Negative	Classified by the original data
Income Tax Expense	Positive, Null, Negative	Classified by the original data
Profit (loss)	Positive, Negative	Classified by the original data
Other Comprehensive Income	Positive, Null, Negative	Classified by the original data
Total Comprehensive Income	Positive, Negative	Classified by the original data
Market Reaction (Equity Subscription Rate)	High, Low	Classified by the distance of max and min in mean of equity subscription rate

3.3 Random forest and evaluation

This study implements the WEKA software for evaluating the market reaction of start-up firms on GISA platform in Taiwan by RF model. The features of data sources are 24. We obtained the most important features by correlation attributes evaluation method in WEKA software, which include Non-current assets, Total assets,

Non-current liabilities, Total capital, Additional paid in capital, Retained earnings, Profits, Total comprehensive income. The final RF training model for evaluating the market reaction on start-up firms is shown in Figure. 6, with number of iterations are 280 and number of features are (log_2 (predictors) + 1). Then the results of correctly ratio in training datasets is 67.5% and incorrectly ratio is 32.5% which is shown in Figure 7. Another, the testing data is from GISA by random sampling on 20 datasets (See Table 5), the results of correctly ratio in testing datasets is 65% and incorrectly ratio is 35% (See Figure. 8).

Table 5. The random sampling data for testing

Start-ups ID	Non-current Assets	Total Assets	Non-Current Liabilities	Total Capital	Additional Paid in Capital	Retained Earnings	Profit (loss)	Total comprehensive income	Equity Subscription Rate(Market reaction)
1	Low	Medium	High	Medium	Low	Extremely Low	Negative	Positive	High
2	Extremely Low	Low	Medium	Extremely Low	Low	Low	Negative	Negative	High
3	Medium	Medium	High	High	High	Low	Negative	Positive	Low
4	High	High	High	Medium	High	Extremely Low	Negative	Positive	Low
5	Extremely Low	Extremely Low	Medium	Extremely Low	Low	Low	Negative	Negative	High
6	Low	Extremely Low	Medium	Medium	High	High	Negative	Negative	Low
7	Medium	High	Extremely Low	High	Low	Extremely Low	Positive	Positive	Low
8	Low	Medium	Low	High	Low	Low	Negative	Negative	High
9	Low	High	Medium	High	High	Extremely Low	Negative	Negative	Low
10	Low	Extremely Low	Medium	Extremely Low	Low	Low	Negative	Negative	Low
11	Medium	Medium	Extremely Low	Medium	High	Medium	Positive	Positive	High
12	High	High	Low	High	High	Extremely Low	Negative	Negative	High
13	Medium	Medium	Extremely Low	High	High	High	Negative	Positive	Low
14	Low	Low	Low	Medium	Low	High	Negative	Negative	Low
15	High	Medium	Medium	Medium	High	Low	Positive	Positive	High
16	Medium	Low	Medium	Extremely Low	Low	High	Positive	Positive	High
17	Extremely Low	Extremely Low	Low	Low	Low	Medium	Negative	Negative	Low
18	Low	Low	Extremely Low	Extremely Low	Low	Medium	Negative	Negative	High
19	Extremely Low	Extremely Low	Low	Low	Low	Extremely Low	Negative	Negative	High
20	Medium	High	Medium	Extremely Low	High	High	Positive	Negative	Low

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          85               67.4603 %
Incorrectly Classified Instances        41               32.5397 %
Kappa statistic                          0.3485
Mean absolute error                      0.4055
Root mean squared error                  0.4694
Relative absolute error                 81.0836 %
Root relative squared error             93.8488 %
Total Number of Instances              126
```

Figure 7. The RF results of training data

```
=== Summary ===

Correctly Classified Instances          13               65      %
Incorrectly Classified Instances         7               35      %
Kappa statistic                          0.3
Mean absolute error                      0.4061
Root mean squared error                  0.4812
Total Number of Instances               20
```

Figure 8. The RF results of testing data

This study applied the performance evaluation methods including classification accuracy, precision and recall based on the confusion matrix (see Table 1), and 10-fold cross validation to evaluate the proposed method. Finally, we also used the F1-score and ROC curve to measure the performance of the RF model. Table 6 to Table 7 are the evaluation indicators on training datasets and testing datasets. The accuracy, precision, recall and F1 score in testing datasets that provide 0.650, 0.500, 0.625 and 0.556 respectively. It represents that this prediction model have 65% for detecting the market reaction of start-up firms in GISA equity crowdfunding platform in Taiwan. For example, a new startups' entrepreneur who can implement the forecast model to evaluate the market reaction in equity crowdfunding market when he (she) would like to make their funding by register GISA platform at Taiwan. Finally, the ROC curve is shown in Figure 9 and Figure 10.

Table 6. Confusion matrix of training datasets

	Low (Actual)	High (Actual)
Low (Predicted)	TP (45)	FP(19)
High (Predicted)	FN(22)	TN(40)

Table 7. Confusion matrix of testing datasets

	Low (Actual)	High (Actual)
Low (Predicted)	TP (5)	FP(5)
High (Predicted)	FN(3)	TN(8)

$$Training\ datasets\ accuracy = \frac{TP+TN}{TP+FP+FN+TN} = \frac{85}{126} = 0.674$$

$$Training\ datasets\ precision = \frac{TP}{(TP+FP)} = \frac{45}{64} = 0.703$$

$$Training\ datasets\ recall = \frac{TP}{(TP+FN)} = \frac{45}{67} = 0.672$$

$$Training\ datasets\ F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{0.945}{1.375} = 0.687$$

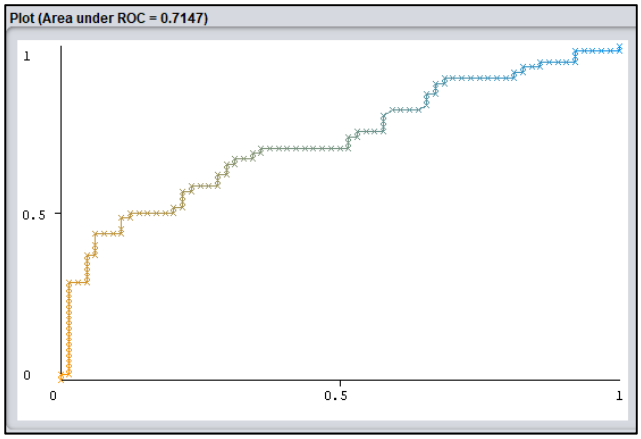

Figure 9. The ROC curve of training datasets

$$Testing\ datasets\ accuracy = \frac{TP+TN}{TP+FP+FN+TN} = \frac{13}{20} = 0.650$$

$$Testing\ datasets\ precision = \frac{TP}{(TP+FP)} = \frac{5}{10} = 0.500$$

$$Testing\ datasets\ recall = \frac{TP}{(TP+FN)} = \frac{5}{8} = 0.625$$

$$Testing\ datasets\ F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{0.625}{1.125} = 0.556$$

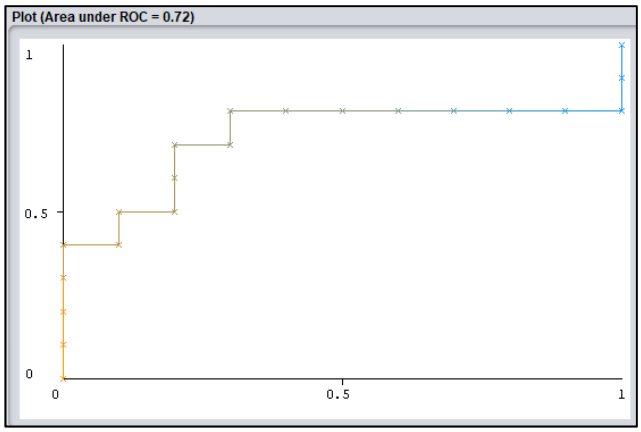

Figure 10. The ROC curve of testing datasets

Based on the above, the proposed model has 67.4% efficiency to forecast the market reaction

in pre-listing stages on GISA. The level of F1 score has 68.7% means that the robustness of the proposed RF model is suitable for predict the start-up firms market reaction. Additionally, we applied the random sampling for 20 samples from January 1st, 2019 to June 31st, 2019 to test the proposed RF model. As the results of testing data, the accuracy and F1 score are 65% and 55.6%, it represents the proposed RF model which can assist the start-up firms to predict the market reaction when they want to list on the GISA platform (see Figure 11). We can understand by the Figure 11 that indicates the startup 3, 6, 7, 10, 12, 14, 19 are failure. For example, the actual value of market reaction in startup 3 is low, but the predicted result is high; the actual value of market reaction in startup 12 is high, but the predicted result is low. Even though the theoretical results have some failure on this detecting model. In overall, the performance of accuracy have 65% for predicting the real data in market reaction on GISA platform to startup firms.

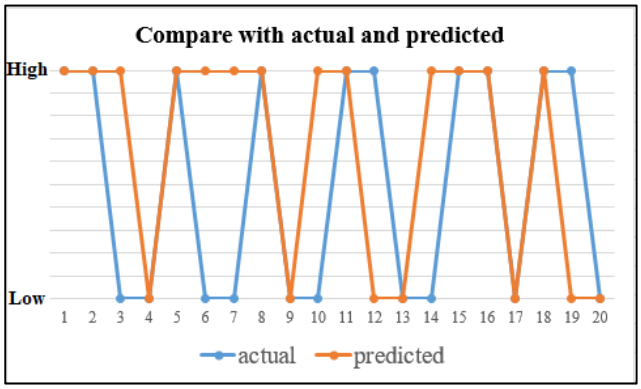

Figure 11. The comparison of results in testing datasets

4. Conclusion

The equity crowdfunding platform in Taiwan is GISA on the Taipei Exchange. If start-up firms want to list on the GIA platform, they must test the market reaction prior to listing when the market reaction is bad then start-up firms have to withdraw the registration process. This risk not only extends the financing cycle but also that would enhance the threat in urgency of working capital from start-up firms. Thus, this study proposes a machine-learning model to help the entrepreneurs for forecasting the market reaction on GISA platform in Taiwan.

This study provides empirical evidence about the market reaction for start-up firms pre-listing on the GISA in Taiwan equity crowdfunding platform by the RF model. The evaluation performance of this model which the accuracy is 67.4% and the F1 score is 68.7% then the testing dataset have 65% and 55.6%. In view of the GISA in Taiwan is an equity crowdfunding platform, the datasets of GISA is not transparent and the founding time is from 2015. Even though the performance is not very high, but it also can obtain the initial analysis for evaluating the market reaction before listing on the GISA platform to reduce the risks based on this limitation situation. Due to the founding of GISA platform in Taiwan in 2015, the data samples, start-up firms' behaviors, operation strategies and details are not transparency. Hence, the further study that can increase the new datasets and new features then reconstruct the forecasting model for retesting and improving the evaluation performance.

Consequently, this study applied the AI algorithm of ensemble RF model for predicting the market reaction of start-up firms listing on the GISA equity crowdfunding platform, which able to obtain the probabilities of market reaction of start-up enterprise and forecast the degree of market reaction. The proposed AI model to forecast the market reaction of start-up firms before joining the Taiwan GISA equity

crowdfunding platform. The results not only fulfill the gap of detecting market reaction in equity crowdfunding but the proposed RF model which can replace the traditional statistics technique to evaluate the market reaction then obtain the evaluation rules and more available forecasting efficiency., the proposed model applied AI algorithms to predict the market reaction in Taiwan GISA platform that can provide a useful ensemble tool for start-up firms and entrepreneurs to evaluate the degree of market reaction more efficiency before listing on the Taiwan GISA platform. Finally, this study proposes two recommendations for future study. The first is future research that can apply the proposed model to detect the probabilities of the private market (GISA) to the IPO market. The second is that may concentrate on implementing other machine learning algorithms such as SVM, ANN, Logistic regression and so on.

Reference:

[1] Alaka, H. A., Oyedele, L. O., Owolabi, H. A., Kumar, V., Ajayi, S. O., Akinade, O. O., & Bilal, M. (2018). Systematic review of bankruptcy prediction models: Towards a framework for tool selection. *Expert Systems with Applications*, *94*, 164-184.

[2] Acharjee, A., Kloosterman, B., de Vos, R. C., Werij, J. S., Bachem, C. W., Visser, R. G., & Maliepaard, C. (2011). Data integration and network reconstruction with~ omics data using Random Forest regression in potato. *Analytica chimica acta*, *705*(1-2), 56-63.

[3] Abellan, J., & Masegosa, A. R. (2010). An ensemble method using credal decision trees. *European Journal of Operational Research*, *205*(1), 218-226.

[4] Agrawal, A., Catalini, C., & Goldfarb, A. (2011). The Geography of Crowdfunding (= National Bureau of Economic Research Working Paper Series Nr. 16820). *Cambridge, MA*.

[5] Breiman, L., J. H. Friedman, R.A. Olsen, & C. J. Stone. (1984). *Classification and Regression Trees*. CA:Wadsworth.

[6] Bagheri, A., Chitsazan, H., & Ebrahimi, A. (2019). Crowdfunding motivations: A focus on donors' perspectives. *Technological Forecasting and Social Change*, *146*, 218-232.

[7] Bevilacqua, M., Ciarapica, F. E., & Giacchetta, G. (2008). Industrial and occupational ergonomics in the petrochemical process industry: A regression trees approach. *Accident Analysis & Prevention*, *40*(4), 1468-1479.

[8] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[9] Chen, L. H., Jiang, G. J., & Zhu, K. X. (2018a). Total attention: The effect of macroeconomic news on market reaction to earnings news. *Journal of Banking & Finance*, *97*, 142-156.

[10] Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783-2792.

[11] Chen, Z., Han, F., Wu, L., Yu, J., Cheng, S., Lin, P., & Chen, H. (2018b). Random forest based intelligent fault diagnosis for PV arrays using array voltage and string currents. *Energy conversion and management*, *178*, 250-264.

[12] Cumming, D. J., & Vismara, S. (2017). De-segmenting research in entrepreneurial finance. *Venture Capital*, *19*(1-2), 17-27.

[13] Drousia, A., Episcopos, A., & Leledakis, G. N. (2019). Market reaction to actual daily share

repurchases in Greece. *The Quarterly Review of Economics and Finance*.

[14] Došilović, F. K., Brčić, M., & Hlupić, N. (2018, May). Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210-0215). IEEE.

[15] Deng, X., Liu, Z., Zhan, Y., Ni, K., Zhang, Y., Ma, W., ... & Rogers, K. M. (2020). Predictive geographical authentication of green tea with protected designation of origin using a random forest model. *Food Control*, *107*, 106807.

[16] Dangeti, P. (2017). *Statistics for machine learning*. Packt Publishing Ltd.

[17] Eugen, B.T. (2019). Studies regarding tourism development perspectives in the existing economical and environmental context. WSEAS Transactions on Environment and Development, 15, 197-203.

[18] Fiordelisi, F., Minnucci, F., Previati, D., & Ricci, O. (2019). Bail-in regulation and stock market reaction. *Economics Letters*.

[19] Fenwick, M., McCahery, J. A., & Vermeulen, E. P. (2017). Fintech and the financing of entrepreneurs: From crowdfunding to marketplace lending. Page. 15.

[20] González-Carrasco, I., Jiménez-Márquez, J. L., López-Cuadrado, J. L., & Ruiz-Mezcua, B. (2019). Automatic detection of relationships between banking operations using machine learning. *Information Sciences*, *485*, 319-346.

[21] Güler, K., & Tepecik, A. (2019). Exchange Rates' Change by Using Economic Data with Artificial Intelligence and Forecasting the Crisis. *Procedia Computer Science*, *158*, 316-326.

[22] Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining (adaptive computation and machine learning)*. MIT Press.

[23] Ivanov, V., & Knyazeva, A. (2017). US securities-based crowdfunding under Title III of the JOBS Act. *DERA White paper*. [Accessed July 15, 2019].

[24] Kareem, S. A., Pozos-Parra, P., & Wilson, N. (2017). An application of belief merging for the diagnosis of oral cancer. *Applied Soft Computing*, *61*, 1105-1112.

[25] Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, 119-127.

[26] Khayamim, A., Mirzazadeh, A., & Naderi, B. (2018). Portfolio rebalancing with respect to market psychology in a fuzzy environment: a case study in Tehran Stock Exchange. *Applied Soft Computing*, *64*, 244-259.

[27] Linder, E., & Marbuah, G. (2019). The cost of transparency: Stock market reactions to introduction of the Extractive Sector Transparency Measures Act in Canada. *Resources Policy*, *63*, 101463.

[28] Liu, D., & Sun, K. (2019). Random forest solar power forecast based on classification optimization. *Energy*, *187*, 115940.

[29] Lu, Y., Chang, R., & Lim, S. (2018). Crowdfunding for solar photovoltaics development: A review and forecast. *Renewable and Sustainable Energy Reviews*, *93*, 439-450.

[30] Lu, S., Li, Q., Bai, L., & Wang, R. (2019). Performance predictions of ground source heat pump system based on random forest and back propagation neural network models. *Energy Conversion and Management*, *197*, 111864.

[31] Breiman, L. (1999). Random forests. *UC Berkeley TR567*.

[32] Mata, J., De Miguel, I., Duran, R. J., Merayo, N., Singh, S. K., Jukan, A., & Chamania, M. (2018). Artificial intelligence (AI) methods in optical networks: A comprehensive survey. *Optical Switching and Networking*, *28*, 43-57.

[33] Mienye, I. D., Sun, Y. & Wang, Z. (2019) Prediction performance of improved decision tree-based algorithms: a review. *Procedia Manufacturing*, *35*, 698-703.

[34] Mitchell, T. M. (1997). Machine learning. Singapore: McGraw-Hill.

[35] Michael, J. A., & Gordon, S. L. (1997). Data mining technique: For marketing, sales and customer support. *New York: John Wiley&Sons INC*, *445*.

[36] Nurhayati, A., Aisyah, I., & Supriatna, A. K. (2019). The relevance of socioeconomic dimensions in management and governance of sea ranching. *WSEAS Transactions on Environment and Development*, 15, 78-88.

[37] Portugal, I., Alencar, P., & Cowan, D. (2018). The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, *97*, 205-227.

[38] Pan, S., & Zhou, S. (2019). Evaluation Research of Credit Risk on P2P Lending based on Random Forest and Visual Graph Model. *Journal of Visual Communication and Image Representation*, 102680.

[39] Polat, K., & Güneş, S. (2009). A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Systems with Applications*, 36(2), 1587-1592.

[40] Peng, H., Zhang, X., & Huang, L. (2017). An energy efficient approach for C4. 5 algorithm using OpenCL design flow. In *2017 International Conference on Field Programmable Technology (ICFPT)* (pp. 144-151). IEEE.

[41] Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, *9*(2), 181-199.

[42] Petruzzelli, A. M., Natalicchio, A., Panniello, U., & Roma, P. (2019). Understanding the crowdfunding phenomenon and its implications for sustainability. *Technological Forecasting and Social Change*, *141*, 138-148.

[43] Pang, H., Lin, A., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., & Zhao, H. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics*, *22*(16), 2028-2036.

[44] Quinlan, J.R. (1986). Introduction of Decision Tree. Machine Learning, 1, 81-106.

[45] Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning.* CA: Morgan Kaufmann.

[46] Rossi, M. (2014). The new ways to raise capital: an exploratory study of crowdfunding. *International Journal of Financial Research*, *5*(2), 8-18.

[47] Ryman-Tubb, N. F., Krause, P., & Garn, W. (2018). How Artificial Intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. *Engineering Applications of Artificial Intelligence*, *76*, 130-157.

[48] Rokach, L., & Maimon, O. Z. (2008). *Data mining with decision trees: theory and applications* (Vol. 69). World scientific.

[49] Russell, I., & Markov, Z. (2017, March). An introduction to the Weka data mining system. In *Proceedings of the 2017 ACM SIGCSE*

Technical Symposium on Computer Science Education (pp. 742-742). ACM.

[50] Shahzad, K., Rubbaniy, G., Lensvelt, M. A. P. E., & Bhatti, T. (2019). UK's stock market reaction to Brexit process: A tale of two halves. *Economic Modelling*, *80*, 275-283.

[51] Sewaiwar, P., & Verma, K. K. (2015). Comparative study of various decision tree classification algorithm using WEKA. *International Journal of Emerging Research in Management &Technology*, *4*, 2278-9359.

[52] Sorokina, N., & Thornton Jr, J. H. (2016). Reactions of equity markets to recent financial reforms. *Journal of Economics and Business*, *87*, 50-69.

[53] Taipei Exchange. (2019). https://www.tpex.org.tw/web/regular_emerging/creative_emerging/Creative_emerging.php?l=en-us. [Accessed 11.10.2019]

[54] Ture, M., Tokatli, F., & Kurt, I. (2009). Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4. 5 and ID3) in determining recurrence-free survival of breast cancer patients. *Expert Systems with Applications*, *36*(2), 2017-2026.

[55] T'sou, B. K., Lai, T. B., Chan, S. W., Gao, W., & Zhan, X. (2000). Enhancement of a Chinese discourse marker tagger with C4. 5. In *Second Chinese Language Processing Workshop* (pp. 38-45).

[56] Tan, K. H., & Lim, B. P. (2018). The artificial intelligence renaissance: deep learning and the road to human-Level machine intelligence. *APSIPA Transactions on Signal and Information Processing*, *7*.

[57] Wood, L. C., Wang, J. X., Olesen, K., & Reiners, T. (2017). The effect of slack, diversification, and time to recall on stock market reaction to toy recalls. *International Journal of Production Economics*, *193*, 244-258.

[58] Wang, H., & Boatwright, A. L. (2019). Political uncertainty and financial market reactions: A new test. *International Economics*.

[59] Walthoff-Borm, X., Schwienbacher, A., & Vanacker, T. (2018). Equity crowdfunding: First resort or last resort?. *Journal of Business Venturing*, *33*(4), 513-533.

[60] Yoo, C., Han, D., Im, J., & Bechtel, B. (2019). Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *157*, 155-170.

[61] Ye, X., Dong, L. A., & Ma, D. (2018). Loan evaluation in P2P lending based on Random Forest optimized by genetic algorithm with profit score. *Electronic Commerce Research and Applications*, *32*, 23-36.

[62] Zhang, B., Lai, K. H., Wang, B., & Wang, Z. (2017). Shareholder value effects of corporate carbon trading: Empirical evidence from market reaction towards Clean Development Mechanism in China. *Energy Policy*, *110*, 410-421.

[63] Zhang, S., Tan, Z., Liu, J., Xu, Z., & Du, Z. (2020). Determination of the food dye indigotine in cream by near-infrared spectroscopy technology combined with random forest model. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *227*, 117551.