

## **A development of a venture business feasibility evaluation model using XGBoosting algorithm**

Seung-Yeon Hwang<sup>a</sup>, Jeong-Joon Kim<sup>b</sup>

<sup>a</sup> Dept of Computer Engineering, University of Anyang, Anyang-si, Gyeonggi-do, Republic of Korea

<sup>b</sup> Corresponding Author, Dept of Software, University of Anyang, Anyang-si, Gyeonggi-do, Republic of Korea

### **Abstract**

As of 2018, Korea's youth employment rate of 15-24 years is 27.4%, which is 14.0% lower than the OECD average youth employment rate of 41.4%. The unemployment rate for youth aged 15 to 29 was 10.0%. On the other hand, in the last decade, despite the instability of youth employment in the Republic of Korea, it is preparing or thinking about venture businesses and young entrepreneurs. However, as a result of examining the reasons why the ratio of young people to entrepreneurs is low compared to the ratio of considering young venture start-ups, difficulty recovering (65.5%), the limit of business ideas (17.2%), and lack of entrepreneurship support system (17.2%). Therefore, in this study, we analyzed the business factor evaluation factors of BMO indicators using the XGBoosting algorithm. The main purpose of this study is to provide a reliable and predictable business model for each business entity, investor and government industry support team.

**Key Words :** R, Xgboost, Big Data, Business evaluation, Govern Driven venture support system, Young businessman, Statistical analysis

### **1. Introduction**

The 2008 Lehman Brothers crisis caused the Great Depression around the world. Countries around the world have made great efforts to boost economic growth and create jobs over the past decade to overcome the global economic crisis. We are overcoming the crisis by discovering and supporting new growth businesses, not by increasing productivity and increasing quantity. As a result, the real unemployment rate, excluding the natural unemployment rate, has been declining since mid-2017, and the world is preparing for a second leap in economic growth as of 2018.

On the other hand, unlike the economic trend that is changing from a low-growth phase to a growth phase in the world, as of 2018, South Korea achieved the largest youth unemployment since the 2008 global financial crisis.

The government authorities are trying to revitalize the economy through revitalization of small and medium-sized enterprises and venture businesses and support for start-ups. Despite the government's will for improvement, the youth employment situation has not changed significantly. According to the National Statistical Office, the number of unemployed people increased from 807,000 since 2013, from 937,000 in 2013 to 101,2000 in 2016, and recently increased to 102,8000 in 2017, an increase of 220,000 in four years. Accordingly, the government greatly strengthened venture investment and re-startup investment support projects in order to find new growth engines using the creative economy as a motif. It has established a small and medium-sized venture business department and is developing a business focused on management practices such

## A development of a venture business feasibility evaluation model using XGBoosting algorithm

as intellectual property rights and patents, market analysis, and marketing through a support project called k-startup.

However, due to the nature of venture start-ups, due to difficulties in judging sustainability through financial statements-based business feasibility indicators (for example, DCF evaluation models), Korea's success rate in venture business support is significantly lower than other OECD countries. In particular, the survival rate of start-ups after three years in the Death Valley period, which is called the tomb of important start-up venture companies, is 41%, ranking last among OECD.(DataOECD,2013) Also, according to an international comparison of Korean entrepreneurship activities, the proportion of 15-24 year-olds among 64 countries was 1.8%, ranking 62nd among 64 countries. (Source: Global Entrepreneurship Monitor (GEM)) As such, the overall start-up environment falls short of support, and there is no business feasibility assessment analysis of the city's proper business model, causing social losses not only for government officials but also for founders and related venture investment advisors. For this reason, finding a venture company with a relatively safe business success rate is important from the perspective of founders, government agencies, and investors. With the recent development of big data technology and the influence of government 3.0 open data, the institutional and technological foundation to solve these problems has been laid. Therefore, it became possible to analyze the factors of feasibility evaluation using public data. Therefore, if quantitative business evaluation of similar businesses in the future becomes possible based on the data so far, it will be possible to present a new milestone for founders and investors. Therefore, analyzing business feasibility evaluation factors through recent business data is more important than selecting business items.

The process of selecting the top 1000 businesses in Seoul is divided into the first and second evaluations, and the evaluation of business plans and external experts. The main contents of the plan consist of motivation, ideas, outlines, business promotion plans, marketability, profitability, and ripple effects on recruitment, and there is no content that reviews the founder's capabilities or refers to evaluation indicators that suit the characteristics of the industry. Looking at overseas cases, y-combinators or tech stars in the US select about 1% of teams in the US according to a strict selection process, and make and evaluate whether there is a real demand behind the business, such as profitability and marketability, and whether it is a sustainable business in the future. Therefore, the success rate was about 11% higher than that of Korea. In addition, through accelerator and exit investments, about 61% were acquired for more than 300 billion won.

Therefore, in this study, the predictive model methodology was constructed through the selection and analysis of start-up support projects hosted by government ministries, and the quantitative evaluation model was added as a support criterion index, and the success probability prediction model of venture and re-startup companies was applied to the existing application. It is constructed based on the xgboosting algorithm along with the BMO evaluation model index. Based on the results derived from this study, exit investments in the growth phase (stage 2) and the government's assessment of funded investments can be used as indicators of objective and universal success measures through data, rather than just financial statement-based or expert analysis.

## II. Related Technology

### 1. Bigdata

Big data refers to a large set of structured or unstructured data that exceeds the ability to collect, store, manage, and analyze data with existing database management tools, and technology to extract value from such data and analyze the results. It refers to 3V as a property of big data, meaning size, speed, and diversity. Recently, it adds value or complexity.

### 2. R[1]

The R programming language (R) is a programming language and software environment for statistical calculations and graphics. It was started by Robert Gentleman and Ross Ihaka of the University of Auckland in New Zealand and is currently being developed by the R Core team. R is an implementation of the S

programming language distributed under the GPL license, sometimes referred to as GNU S. R is widely used in statistical software development and data analysis, and is a traditional scripting tool among statisticians due to its ease of package development, and is used as a big data analysis tool along with Python. R has a variety of statistical packages ranging from statistics, machine learning, finance, bioinformatics and graphics, all of which are provided free of charge.

### 3. Hadoop[2]

Hadoop (hereafter, Hadoop) is a data analysis application platform that supports a computer clustering environment that can process large amounts of data. It is provided as an open source library by the Apache Open Source Foundation and is a base platform that uses a file system called HDFS to operate an ecosystem that spans large amounts of data storage, processing, and analysis. It was developed in versions 1.0 through 2.0 3.0 and handles data with the concept of MapReduce.

### 4. Correlation analysis

Correlation analysis is a method of analyzing the linear relationship between two variables in probability theory and statistics. The two variables can be correlated from independent relationships, and the strength of the relationship between the two variables is called correlation. In correlation analysis, the number of parental relationships  $\rho$  is used as the unit of the degree of correlation.

The correlation coefficient is a determination of the correlation, which only represents the degree of association between the two variables, but does not explain the causal relationship. The direction, degree, and mathematical model of causality can be identified through regression analysis as to whether there is a causal relationship between the two variables. The correlation coefficient measurement method uses Pearson correlation coefficients. Pearson correlation coefficient (or Pearson's  $r$ ) is commonly used to determine the relationship between two variables. The value of  $r$  has  $+1$  if  $X$  and  $Y$  are completely equal,  $0$  if they are completely different, and  $-1$  if they are completely equal in the opposite direction. The coefficient of determination is calculated as  $r^2$ , which means the degree to which  $Y$  can be predicted from  $X$ .

### 5. Gradient Boosting Machine[3]

Gradient boosting machines are subconcepts of boosting algorithms introduced in 1999. Gradient Boosting = Gradient Descent + Boosting. The basic concept of Boosting is to use multiple weak learning machine to create strong learning machine. It is a model that makes one weak learning machine, then improves the error through the next weak learning machine, and then gradually reduces the remaining errors through the weak learning machine. Gradient Boosting Machine is a model that uses Gradient Descent algorithms to improve errors more effectively. It is used for classification, regression, and rendering.

When you align the first model you created with your data,  $F_{1(x)} = y$ .

The resulting residual clause will be the formula below.

$$h_1(x) = y - F_{1(x)}$$

The key to this algorithm is to continue to reduce the error of residual terms by using more models for residual terms.

$$F_{(x)} = F_{\{M-1\}}(x) + h_{\{M-1\}}(x)$$

where  $F_{1(x)}$  is the initial model for  $y$ .

Since the procedure is initialized by fitting  $F_{1(x)}$ , the task at each step is to find  $h_{m(x)} = y - F_{m(x)}$ , the Gradient Boosting algorithm.

## III. Related Studies

## A development of a venture business feasibility evaluation model using XGBoosting algorithm

Before carrying out this study, let's discuss the implications of domestic research and papers on the factors of early business success of venture companies or startup companies and the analysis of the actual conditions of venture companies by industry.

### 1. An exploratory study on the success factors of startups by industry [4]

This study analyzes the success factors of start-up by classifying characteristics into demographic characteristics, human capital characteristics, and economic characteristics, based on previous studies on the analysis of existing startup success factors. Human capital characteristics include educational background, educational level and suitability, and skill level. Demographic characteristics include gender, age, and marital status. Business characteristics include whether to co-start, industry, and business income. Therefore, the difference in this work is that the research model is expanded by adding contextual characteristics to these existing analytical criteria. Annual variables were treated as dummy variables and the convenience of business operators was included in the research model by dividing them into capital securing, technology securing, manpower securing, administrative procedures, business type and business site selection. Hierarchical regression (AHP technique) is used to analyze business success factors. As a result of the study, as factors with statistical significance of business satisfaction, gender (more male), total investment (larger), average working hours (more), and socioeconomic status (higher) were statistically significant in business income in the two-step model. was found to have a significant positive (+) influence.

### 2. A Study on the Factors Determining Startup Initial Success [5]

This study conducted an AHP analysis based on panel survey data surveyed by experts with start-up experience, who are currently working at each company, research institute, and school to find out the main factors of start-up success. A hierarchical decision-making model was created through AHP analysis so that decision-makers can adjust factors according to their relative importance.



**Fig 1. Important Factors for AHP Analysis Decision Tree**

As a result of analysis by class, in the first stage, 51.0% of funds, 7.8% of marketing, 8.7% of business management, and 32.5% of R&D were found. In the second stage, funds-foreign investment was the highest at 5.7%, buyer feedback at 75% in marketing, and ideas at 43.3% in R&D. Therefore, this study summarizes the most important success factors as the conclusion of the analysis of the startup success factors as the existence of funds to materialize the startup, the innovativeness of the idea, and the need for R&D facilities.

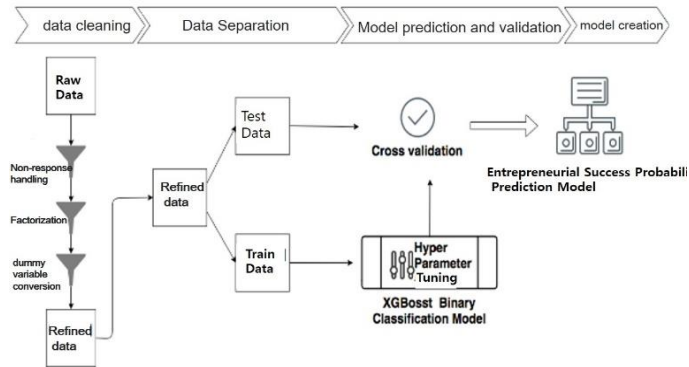
## IV. Analysis

### 1. System design

#### 1.1. design model

As investigated in the previous two domestic related studies, the exploratory study on the success factors of start-up by industry and the study on the factors that determine the initial success of a start-up, the success factors for start-up are marketing, funds, management, R&D, whether or not co-founding, and industry. In order to qualitatively study these complex start-up factors, we analyze start-up success factors by combining expert

interview analysis and quantitative analysis hierarchical decision-making analysis (AHP). After designing the system by dividing the business, the factors affecting business satisfaction and business convenience were analyzed indirectly based on the results of the analysis of the success factors of start-up. Therefore, in this study, rather than a case analysis or a method of simply classifying and analyzing subjective startup success factors, the XGBoosting algorithm learns through raw data to extract factors affecting the actual data, and through this model, the possibility of startup success in the early stage of business By quantitatively measuring The system blueprint is as follows.



**Fig 2. Architecture for Analysis of**

**SuccessfulFactors**

As shown in Figure 2, the order of the system is to first convert the data provided from the public data portal, the Labor Panel Survey, and the National Statistical Office microdata into a csv file, and then purify the data. After that, the test data to verify the accuracy of the model and the train data to be used for prediction of the model are separated, respectively. A predictive model is created from the test data. At this point, the model is optimized through fine-grained changes in the hyperparameter. To verify the accuracy of the model, we finally perform cross-validation with test data. If the value of the statistically significant definition (+) comes out after cross-validation, we terminate the validation and generate the model.

**1.2. System Environment**

**Table 1. System environment**

os	Windows 10 Pro
hardware	Intel(R) i5-7200U CPU
platform	R studio, R

**2. Based on research data**

The research data were used by the National Statistical Office of Korea [6] and the Korea Labor Panel Survey [7], and the specific data were shown in Table 1 below.

**Table 2. Research data [8]**

data source	material item	remarks
Technology Guarantee Fund	Status of Venture Businesses by Company	2017

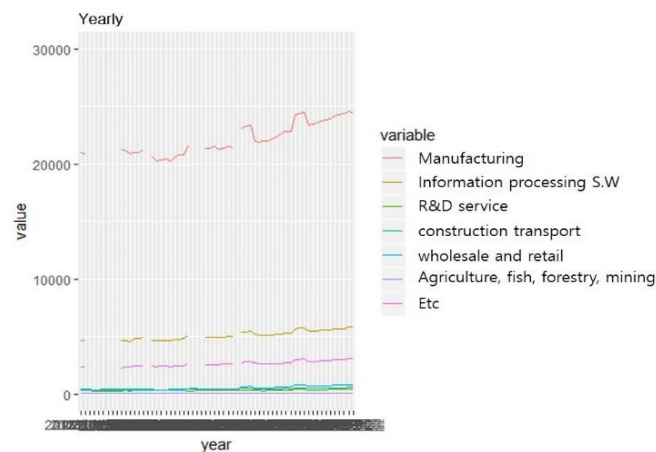
## A development of a venture business feasibility evaluation model using XGBoosting algorithm

Korea Labor Panel Survey	Business panel survey	~2015
--------------------------	-----------------------	-------

According to the Integrated User Guide for the Korean Business Panel Survey, the industry classification standard for the business panel survey follows the 10th revision of the Korean Standard Industrial Classification (KSIC) of the National Statistical Office.

The data of start-up companies, business panel survey, single-person creative company status survey, and labor force survey by occupation are all organized by year. In the case of a panel survey of a business entity, this shall be taken into account because it is included in all of the first to sixth rounds from 2005. In addition, due to the characteristics of start-up companies, they do not have many data items compared to SMEs or large enterprises. It is assumed that data as a company exists from the second stage of growth after the first stage of discovery according to the five growth stages of the startup venture business. Therefore, start-ups in the early days of commercialisation are used as a basis for research.

Based on the start-up venture support project of the government's start-up support group, the government announced on January 4, 2018 and determined whether to use it as data on venture companies of the Korea Technology Guarantee Fund.



**Fig 3. Surgery Guarantee Fund Venture Status by Company**

**Table 3. Use of Korean business panel survey**

R&D service	use
construction transpor	unused
wholesale and retail busines	unused
Farming `fish` Lim mining	unused

### 3. Refine Data

In order to analyze the .dta file using R studio based on the 10th vertical combination data of the Korean business panel survey, the Stata standard file is loaded through the R library. Before converting the data to a matrix data type for the use of xgboosting algorithms, a library with Gradient Boston algorithms, we remove the

non-response data, extract the start-up success rate label, and standardize the value of the item to improve the accuracy of the statistics. After that, in order to use the xgboost library, the categorial variable is transformed into a dummy variable and converted into a numerical data type [10].

### 3.1. Item Extraction

Only the necessary items are extracted to reproduce the data that meets the research criteria. In addition, since the addition of excessive item variables is highly likely to cause a bottleneck in the calculation process, even if it is a necessary variable, if the non-response is a majority, the speed will be improved by removing it. There are 3382 variables based on the 6th data of the Korea Labor Panel Business Survey. Based on WPS[15], a panel survey statistics user integrated codebook of Korean companies, indicators such as labor-management relations, public sector, respondents' characteristics, financial indicators, and labor conflicts are not considered direct business success factors. Related research related to research and development, funds, marketing, business management, technical competency, etc., which are items processed in the exploratory study [4] on the factors that determine the initial success of start-ups [5] About 30 variables to be used for item extraction were selected by referring to the variables with the highest relationship.

```
#data cleaning process
wps_clean <- wps_csv2 %>% select(id, year, ind8, epq2007,
                                aq3003, aq3005, aq3006, aq3007,
                                aq3008, aq3009, aq3010, aq3011, aq3013
                                , aq3014, aq3016, aq3017, aq3018, aq3019
                                , aq3020, aq3902, aq3904, aq3905, aq3906
                                , aq3907, aq3909
                                , fpq1011, fpq2001, fpq2015, fpq2022, fpq4002)

colSums(is.na(wps_clean))
```

Fig 3. Code for Data selection

Table 4. Use of Korean business panel survey variable

Variable Name	Variable Content
EPQ2007	All workers
AQ3003	Foundation year
AQ3005	Domestic Market Ratio
AQ3006	Overseas Market Ratio
AQ3007	Public sector procurement costs.
AQ3008	The degree of competition in the domestic market for flagship products
AQ3009	Main product market demand situation
AQ3010	Price compared to flagship products
AQ3011	Quality compared to competitors
AQ3013	Main product market strategy

AQ3014	Whether an improved product is released
AQ3016	Whether to implement organizational transformation
AQ3017	Whether marketing innovation is implemented
AQ3018	Whether to invest in R&D
AQ3019	Total research and development costs
AQ3020	Internal R&D Percentage
AQ3902	Number of freelance workers
AQ3904	Financial Performance
AQ3905	Labor Productivity
AQ3906	Quality of product/product or service
AQ3907	Worker-led innovation activities
AQ3909	Degree of product/service innovation
FPQ1011	Average number of workers at work during the accounting period.
FPQ2001	Current Revenue
FPQ2015	operating profit
FPQ2022	Net income in the current term
FPQ4002	Total assets at the end of the year

If data with inappropriate conditions are included by referring to the existing startup company success factor analysis data, for example, in the case of large companies, companies with more than 10 years of establishment, and companies with more than 1,000 workers, the average data of startup companies is It may be included as a missing value that is out of range, so it should be removed including the condition.

```
wps_clean <- wps_clean %>%
  filter(year > 2009) %>%
  filter(epq2007 <= 1000) %>%
  filter(!is.na(aq3007))

view(wps_clean)
```

**Fig 4. Code for Data selection2**

Let's take a look at the value of the data whose item extraction has been completed through the head() function.



id	year	ind8	epq2007	aq3003	aq3005	aq3006	aq3007	aq3008	aq3009	aq3010	aq3011	aq3013	aq3014	aq3016	aq3017
30001	2005	21	37	1976	NA	NA	NA	2	4	3	4	3	NA	NA	NA
30001	2007	21	23	1976	100	0	NA	2	4	3	3	1	NA	NA	NA
30001	2009	21	22	1976	100	0	NA	2	4	3	3	2	NA	NA	NA
30002	2005	22	351	2003	NA	NA	NA	2	3	2	4	2	NA	NA	NA
30002	2007	22	406	2003	100	0	NA	3	3	3	3	3	NA	NA	NA

Fig 5. Result data of Code for Data selection2

### 3.2. Non-response correction

When estimating only the data of the responded items excluding the unanswered items, the accuracy of the parameter estimation is lowered due to bias. Non-response is divided into unit non-response that did not respond to the survey itself and item non-response in which only some items did not respond. In the case of a unit non-response, the survey is conducted again or is processed through the weight adjustment method, and in the case of an item non-response, it is processed through replacement of missing values. In this study, the non-response correction using the average value, which is statistically corrected using response and non-response information without additional investigation, will be performed without using the re-examination method of conducting the investigation again. Execute for fpq variables FPQ1011, FPQ2001, and FPQ2015, which are variables with significant information loss when there is no non-response value.

```

mean_1011 <- mean(wps_clean$fpq1011, na.rm = TRUE)
mean_2001 <- mean(wps_clean$fpq2001, na.rm = TRUE)
mean_2015 <- mean(wps_clean$fpq2015, na.rm = TRUE)
mean_2022 <- mean(wps_clean$fpq2022, na.rm = TRUE)
mean_4002 <- mean(wps_clean$fpq4002, na.rm = TRUE)

wps_clean$fpq1011 <- ifelse(is.na(wps_clean$fpq1011), mean_1011, wps_clean$fpq1011)
wps_clean$fpq2001 <- ifelse(is.na(wps_clean$fpq2001), mean_2001, wps_clean$fpq2001)
wps_clean$fpq2015 <- ifelse(is.na(wps_clean$fpq2015), mean_2015, wps_clean$fpq2015)
wps_clean$fpq2022 <- ifelse(is.na(wps_clean$fpq2022), mean_2022, wps_clean$fpq2022)
wps_clean$fpq4002 <- ifelse(is.na(wps_clean$fpq4002), mean_4002, wps_clean$fpq4002)

View(wps_clean)

```

Fig 6. Code for None Data correction and Result

### 3.3 Separation by industry and test cased

In this study, the possibility of entrepreneurship success is analyzed through regression analysis. A prerequisite for regression analysis is that the standard scale for analysis must be a continuous variable to proceed with regression analysis. is converted into a continuous variable to enable regression analysis. In addition, for the accuracy test through the CrossValidation method, subsampling of the original data at a ratio of 0.7 to 0.3 is performed using the caTools library, and the training process is performed through the train data. After that, we will perform cross-validation with test data.

```
#1nd code seperation
#r&d
#70) R&D
#71) Professional service
#72) Building technology, engineering and other science and technology service industries
#73) Other professional, scientific and technical services

wps_clean_by_id_rnd <- wps_csv2 %>% filter(ind8 == 70 |
                                         ind8 == 71 |
                                         ind8 == 72 |
                                         ind8 == 73)

manufacturing <- manufacturing %>% select(id,year,ind9,epq2007,
                                           aq3003,aq3005,aq3006,
                                           aq3007,aq3008,aq3009,
                                           aq3010,aq3011,aq3013,
                                           ,aq3902,aq3904,aq3905,
                                           aq3906,aq3907,aq3909
                                           ,fpq1011,fpq2001,fpq2015,
                                           fpq2022,fpq4002)
```

Fig 7. Code for data spliting process

#### 4.Data Analysis

For data analysis, it is first necessary to extract the main factors for the feasibility evaluation factors using the business panel survey data of the Korea Labor Survey. Therefore, after setting the dependent variable as the startup success rate, and calculating the Feature Importance (factor importance) using the data using the XGBoosting library, only the top 15 features are extracted. However, there is no data item of business success rate in the business panel survey data, and the standard of the start-up success rate is ambiguous, so the task of adding new data items is performed by setting the standard based on subjective criteria. In order to extract the business success rate label, the BMO evaluation analysis system model is applied. BMO evaluation is a system model introduced to analyze the business success rate of start-ups, ventures, and small and medium-sized enterprises, or to determine the commercialization possibility of the R&D business of the company.

Attractivity		degree of conformity	
Market size (sales, profit potential)	(10)	financial power	(10)
Growth potential	(10)	marketing power	(10)
Competitiveness	(10)	manufacturing power	(10)
Risk Diversity	(10)	technology	(10)
Business Rebuild Possibility	(10)	Ability to secure raw materials	(10)
Social superiority (special social situation)	(10)	Management support	(10)
sub Total	(60)	sub Total	(60)

Fig 8. BMO Evaluation 12 Indicators

Considering the corporate characteristics of venture and start-up companies, items such as risk dispersion, business re-establishment potential, social superiority, management support, and ability to secure raw materials in the BMO evaluation index are valid only when the size of the company is above a certain level. In this study, the indicators are modified and applied.

```
c1f <- xgboost(data = as.matrix(wps_clean %>% select(-fpq2022,
                                                  -fpq2015,
                                                  -fpq4002,
                                                  -fpq2001,
                                                  -fpq1011)),
              label = wps_clean_mutated$SuccessRate,
              eta = 0.025, #gradient descent in the algorithm
              depth = 20,
              nrounds = 600,
              gamma = 1,
              objective = "binary:logistic",
              eval_metric = "auc" )

View(c1f)
mat <- xgb.importance(feature_names = colnames(wps_clean),model=c1f2)
xgb.plot.importance(mat)
```

Fig 9. Code foe XGBoosting library data prediction

If you use the business success rate label using the Xgboosting library, you can find out the importance, that is, the relative importance, of the major factors through the results of BINARY classification. Hyper parameters

additionally required to use Xgboosting are parameters provided for easy and fast data analysis and model verification provided by the library. Accuracy can be improved by finely calibrating this parameter. Looking at the code shown in Figure 10 above, eta represents the running rate, nrounds represents the number of rotations of each prediction, gamma represents the technical additional variable, objective represents the objective function, and eval\_metric represents the model to be used for accuracy verification.

**Table 5. Business success rate evaluation index**

Labor index	Take	Invest	Total R&D expenses	Operating profit
5 points:500	5 points: 50 billion	5 points: invest	5 points: 50 billion	5 points:
4 points:50	4 points: 20 billion		4 points: 20 billion	4 points:
3 points:5	3 points: 5 billion		3 points: 5 billion	3 points:
2 points:1	2 points: 5,000	1 points: not invested	2 points: 5,000	2 points:

After allocating the score conversion value as shown in the table to the BMO1~BMO5 variables, the scores are summed to create a SuccessRate label, and the score is converted into a success rate as in the BMO index. 0 and 1 are assigned based on the above criteria, and 0 is regarded as success or failure, and 1 as success.

	Feature	Gain	Cover	Frequency	Importance
1	ind9	0.351493329	0.272027779	0.243442305	0.351493329
2	epq2007	0.294774382	0.370296708	0.164749968	0.294774382
3	id	0.119696693	0.084785222	0.159581341	0.119696693
4	aq3003	0.074639557	0.059458733	0.125339191	0.074639557
5	aq3904	0.037516969	0.039597546	0.048843520	0.037516969
6	aq3005	0.020475056	0.019794506	0.038377051	0.020475056
7	aq3013	0.015890684	0.046580070	0.034500581	0.015890684
8	aq3909	0.015621882	0.019769485	0.017185683	0.015621882

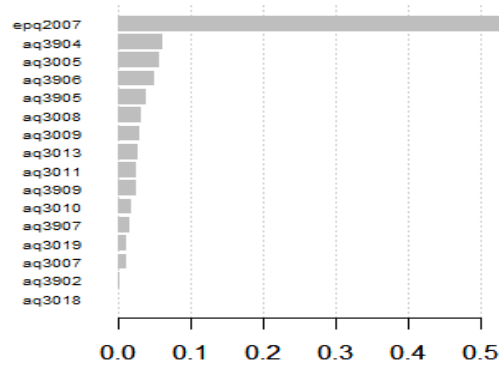
**Fig 10. Result of Feature Importance of Success Business in Dataframe View**

**4.1 Analysis by industry**

In the analysis by industry, Korea's major classification criteria, excluding industries in which Korean livelihood start-ups account for the majority, follow the 9th revision of the Korea Standard Industrial Classification (KSIC) of the National Statistical Office. Therefore, analysis is performed by industry except for agriculture, fishery, and catering. The code used is the same as the code above, so it is omitted.

## V. Conclusion

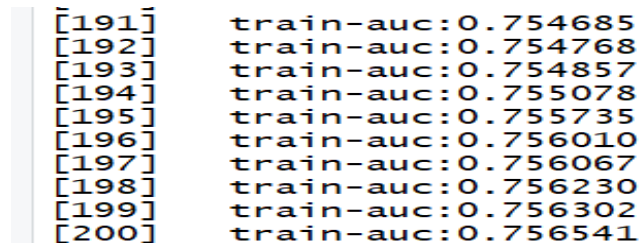
### 5.1 Overall Industry Analysis Results



**Fig 11. Result of Results of business factor evaluation – All**

As a result of analyzing by setting Binary Classification as an objective function for SuccessRate, which is a business success rate label, for the entire industry classification using the xgboost library, the importance graph as shown in Figure 12 appeared. To interpret this, the top 16 important Factors were extracted and graphed. As a result of the analysis, the epq2007 variable, aq3904 variable, aq3005 variable, aq3906 variable, and aq3905 variable were analyzed as the top five important factors. In particular, it can be seen that the epq2007 variable has a relative factor importance of 0.5 or more. Therefore, it can be said that the higher the total number of workers, which is the epq2007 variable, the higher the startup success rate. The second variable, aq3904, is a financial performance variable, and aq3005 is a domestic market ratio variable. The aq3906 variable is the service product quality variable, and the aq3905 variable is the labor productivity variable. Therefore, the results of the analysis of the factors of the success rate of start-ups based on the entire industry generally show that the influence of the number of workers is relatively large. It can be seen that the aq variable, the corporate characteristic variable, has an overall effect.

### 5.2. Verification of analysis results of the entire industry



**Fig 12. Verification Results – All**

Using the xgboost library, set binary classification as an objective function for SuccessRate, a business success factor label for the entire industry classification, and 0.7 to 0.3 for the analysis result auc accuracy using test data for cross-validation. The analysis was carried out. The auc accuracy index closer to 1 means that the proportion of accurately predicted data is higher. When it is 0.5 or more, it is judged as a significant validation model. Therefore, the result value of 0.75 in Figure 13 indicates that this project success rate prediction model is significant.

### 5.3. Analysis result of R&D industry

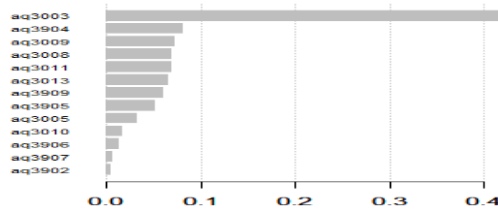


Fig 13. Business success factors Importance results - Research and development

As a result of analyzing by setting Binary Classification as an objective function for SuccessRate, which is a business success factor label, for R&D classification using xgboost library, a graph of importance as shown in Figure 14 appeared. To interpret this, the top 15 The important factors were extracted and presented as a graph. As a result of the analysis, the aq3003 variable, aq3904 variable, aq3009 variable, aq3008, and aq3011 variable were analyzed as the top five important factors. Different from the results of analysis of all industries, important factors are evenly distributed. The aq variable is a workplace characteristic variable. Since the highest factor and the top 4 factors do not exceed 0.15, there is no weighted variable. The variable aq3005, which is the most important variable, is the domestic market ratio, the second aq3904 variable is the financial performance variable, the aq3009 variable is the market demand for the main product, the aq3008 variable is the degree of competition in the domestic market of the main product, and the aq3011 variable is the product quality. The importance was shown in the above order.

### 5.4. Verification of the analysis results of the R&D industry

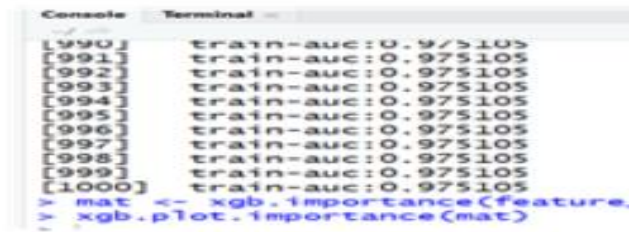


Fig 14. Verification Results - Research and development

For the classification of R&D language using the xgboost library, set binary classification as the objective function for SuccessRate, which is the label of the startup success factor, and 0.7 to 0.3 for the analysis result using test data for cross-validation auc Accuracy analysis was performed. Unlike the entire industry, it is 0.975105, which is more accurate.

### 5.5. Manufacturing industry analysis result

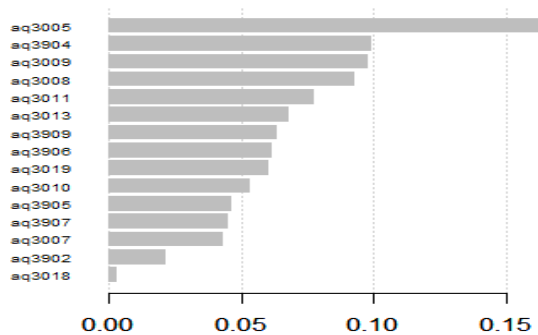


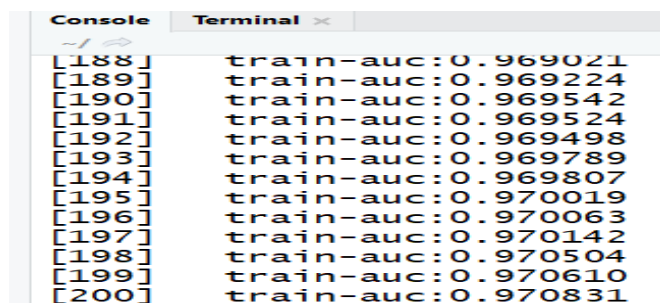
Fig 15. Business Success Factors Importance Results - Manufacturing

## A development of a venture business feasibility evaluation model using XGBoosting algorithm

As a result of analyzing by setting binary classification as an objective function for SuccessRate, which is a label of a startup success factor, for manufacturing classification using the xgboost library, the importance graph as shown in Figure 16 appeared. To interpret this, the top 13 important Factors were extracted and presented as a graph. The manufacturing sector data was judged to be more accurate because there is about 4 times more data than other industries.

As a result of the analysis, aq3005 variable, aq3904 variable, aq3009 variable, aq3008, aq3011 variable were analyzed as the top 5 important factors. Different from the results of analysis of all industries, important factors are evenly distributed. The aq variable is a workplace characteristic variable. Since the highest factor and the top 4 factors do not exceed 0.15, there is no weighted variable. The variable aq3005, which is the most important variable, is the domestic market ratio, the second aq3904 variable is the financial performance variable, the aq3009 variable is the market demand for the main product, the aq3008 variable is the degree of competition in the domestic market of the main product, and the aq3011 variable is the product quality. The importance appeared in the same order. As a result of the analysis of the R&D industry, it can be confirmed that the similarity with the importance graph is very high.

### 5.6. Verification of manufacturing industry analysis results



Index	train-auc
[188]	0.969021
[189]	0.969224
[190]	0.969542
[191]	0.969524
[192]	0.969498
[193]	0.969789
[194]	0.969807
[195]	0.970019
[196]	0.970063
[197]	0.970142
[198]	0.970504
[199]	0.970610
[200]	0.970831

Fig 16. Verification Results – Manufacturing

For classification of R&D language using xgboost library, set binary classification as an objective function for SuccessRate, which is a business success factor label. Accuracy analysis was performed. It can be seen that, unlike the entire industry, it shows a high accuracy of 0.970.

## VI. Conclusion

As a result of creating a predictive model for feasibility evaluation factors by dividing it into three classification criteria (total industry, R&D industry, and manufacturing industry) through a total of 12352 business panel survey data, in the case of all industries, the data is about three times more on average. Nevertheless, the accuracy was 0.75, indicating that the accuracy of the other two classifiers was relatively low. In addition, as a result of confirming the status of venture companies in both the R&D and manufacturing sectors through the Technology Guarantee Fund data, more than half of the companies had R&D companies first among industries excluding the manufacturing sector and other sectors, and the growth rate was higher than that of other industries. Based on the results, even though there is a difference in the importance of the business characteristics of the two industries, the results were completely consistent with the 2nd to 10th median rankings except for the first important factor. For all industries, it can be seen that the bias of the importance factor of 0.4 or higher is biased toward the number of labor workers, which is the epq2007 variable. As the importance of the remaining factors is three times lower, the data for the entire industry has an accuracy of 0.75, so there is a limit to interpreting it as meaningful data. In addition, the analysis results in the R&D industry and the manufacturing sector show accuracy exceeding 0.97, respectively, despite the small number of data, so there is room for interpretation as a meaningful analysis result. In particular, the manufacturing industry has more than twice as much data as the R&D industry, so it can be seen that the bias of the important factors is the lowest, and the importance does not differ significantly for each ranking. It can be seen that the establishment year variable, which is the aq3003 variable analyzed as the most important factor in the R&D industry, is higher than the R&D

cost, which is the expected number one variable. However, as a result of checking the data, it is confirmed that this is due to the relatively high percentage of the amount of data that did not respond to the item of R&D expenses. About 30 items of data were extracted based on the integrated codebook, and BMO1~BMO6 variables, which are applied variables that transformed the BMO evaluation method, were transformed into binary variables, which are SuccessRate variables. In a future research project, a more accurate model can be created if a factor analysis is performed by adding meaningful data among 3362 variables. There is also a problem with the quality of the data. There is a problem with the quality of the data because the NA value of the data, that is, the ratio of non-response values is high on average. In addition, since the data is based on the respondents' answers, it should be taken into account that the reliability issue of the respondents is excluded. In future research tasks, it will be possible to conduct more accurate research through not only business panel surveys but also data survey units specialized for venture companies.

## REFERENCES

- [1] R, <https://www.r-project.org/foundation/>
- [2] Apache Hadoop, <http://hadoop.apache.org/>,
- [3] Friedman, Jerome H.(2001), Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics*, 29(5):1189-1232
- [4] Doo-hee Nam(2011.4), Design of Information System For Assisted Living, *The Journal of The Institute of Webcasting, Internet and Telecommunication* 11(2):83-88
- [5] Hyun-Ho Lee, Hwang-bo Yun, Chang-Hoon Gong, A Study on the Factors that Determine the Initial Success of Start-Up, *Asia-Pacific Journal of Business Venturing and Entrepreneurship*, 12(1):1-13
- [6] Korea Statistical Office, <http://kostat.go.kr>
- [7] Korea LaborInstitute, <https://www.kli.re.kr/kli/index.do>
- [8] Korea Department of Small & Medium Venture BusinessInvestigation of precise status of venturecompany, <http://www.smba.go.kr/site/smba>
- [9] Korea Business Development Agency, <https://www.kised.or.kr/>
- [10] Chen T., Guestrin C.(2016.8), Xgboost: A scalable tree boosting system, *Proceedings of the 22nd acmsigkdd international conference on knowledge discovery and data mining*, pp. 785-794