ISSN: 2508-7894 © 2020 KODISA & KAIA.
KJAI website: http://www.kjai.or.kr
doi: http://dx.doi.org/10.24225/kjai.2022.10.1.15

Predictiong long-term workers in the company using regression

Ho Min SON[1], Jung Hwa SEO[2]

Received: April 27, 2022. Revised: May 03, 2022. Accepted: June 03, 2022.

Abstract

This study is to understand the relationship between turnover and various conditions. Turnover refers to workers moving from one company to another, which exists in various ways and forms. Currently, a large number of workers are considering many turnover rates to satisfy their income levels, distance between work and residence, and age. In addition, they consider changing jobs a lot depending on the type of work, the decision-making ability of workers, and the level of education. The company needs to accept the conditions required by workers so that competent workers can work for a long time and predict what measures should be taken to convert them into long-term workers. The study was conducted because it was necessary to predict what conditions workers must meet in order to become long-term workers by comparing various conditions and turnover using regression and decision trees. It used Microsoft Azure machines to produce results, and it found that among the various conditions, it looked for different items for long-term work. Various methods were attempted in conducting the research, and among them, suitable algorithms adopted algorithms that classify various kinds of algorithms and derive results, and among them, two decision tree algorithms were used to derive results.

Keywords : Machine Learning, Boosted Decision Tree Regression, Linear Regression Microsoft Azure Machine

Major Classification Code : Basic Technology, Technical Application, Artificial

1. Introduction

Workers exist in various forms. They exist not only in developers, but also in marketers, designers, etc. As you can often see in the news, they are considering changing jobs for various purposes, and the company is trying to turn workers into long-term workers because they need to find new ones. Currently, the development of machine learning is being used in various fields. It is being used in various forms in many people's daily lives as well as in simple research fields, and a prediction system that provides necessary information to users again based on data used by users is being actively used. For example, the YOUTUBE algorithm, advertisements that can be found while using the Internet, advertisements that can be seen while using the app, and recommend products to consumers while using the shopping app. Algorithms and user recommendation systems using artificial intelligence lead to analysis and prediction of information that consumers need. By incorporating this in the company, the company needs to make efforts to identify which areas office workers feel satisfied and which areas they do not feel satisfied with through machine learning research, and to turn office workers into long-term workers. According to AN Soo-hyun's A Study on a car Insurance purchase Prediction Using Two-Class Logistic Regression and Two-Class Boosted Decision Tree paper, the two decision trees that compare various conditions and derive one of the most important reasons through classification are selected as the

1 First Author. Student, Department of Medical IT, Eulji University, Korea. Email: sonhomin98@naver.com
2 Corresponding Author, Adjunct Professor, Major in Design, Seoul Institute Of The Art, Korea. Email: saeam111@naver.com

main algorithm, and several groups are created, and then the company can derive the desired purpose.

2. Literature Review

2.1. Machine Learning Program

Azure is a cloud computing platform that started service at Microsoft in 2010. Following PaaS in 2011, IaaS services were launched in 2013, focusing only on Azure Machine Learning Studios among more than 600 services offered by the Azure platform. Azure has an open source library and tools for machine learning. The reason for using Azure Machine Learning Studio is that it solves problems such as lack of flexibility in computing resources for learning, difficulty in configuring GPU-based environments for machine learning, difficulty in installing and setting tools for learning, and difficulty in recording and versioning.It announced Azure Machine Learning designer's general availability, the drag-and-drop workflow capability in Azure Machine Learning studio, which simplifies and accelerates building, testing, and deploying machine learning models for the entire data science team, from beginners to professionals (Nam et al, 2019). In addition, unlike existing cloud platforms and machine learning libraries and tools, it provides an easy-to-access GUI environment in consideration of user convenience (Kang et al., 2018).

2.2. Decision Tree

The decision tree model is divided into a classification tree for discrete types and a regression tree for continuous types, and a representative classification tree is CHAID (Kass, 1980). CART(Breiman et al., 1984), C4.5(Quinlan, 1992), QUEST(Loh & Shih, 1997), CRUISE(Kim & Loh, 2001). This study constructs an interactive decision tree based on the CHAID algorithm. The interactive method has the advantage of being highly relevant to the target variable and being able to preferentially reflect independent variables that researchers consider important in the model. And data mining technology that charts rules into trees.Organize and categorize groups of interest into several. Create small groups or make predictions. In other words, it analyzes the collected data and classifies new data. The pattern decision tree that exists between them produces the following: A top-down type of tree and an appropriate amount of training set shall be prepared in advance. The decision tree algorithm classifies a given set of training to generate: Categorize and predict decision trees and given test sets.

2.3. Corporate turnover

Turnover generally means that a member of an organization leaves the organization with a voluntary motive (Bluedorn, 1978). Various theories have been proposed to explain the factors of turnover. Typical theories include the theory of organizational balance (March and Simon, 1958), the Met Expectations Model (Porter and Steers, 1973), the Linkage Model (Mobley, 1977), the Unfolding Model of Turnover (Lee and Mitchell, 1994) and the Job Embedding Theory (Empty, 2001). These theories focused on various factors such as self-efficacy in an organization, whether expectations between organizational expectations and actual experiences were met, job dissatisfaction intensified, and impulsive factors caused by events that caused job turnover.

2.4. Boosted Decision Tree

It is a data mining technique that charts rules into a tree structure and classifies the group of interest into several small groups or performs prediction. In other words, it analyzes the collected data and classifies new data into patterns that exist between them. The decision tree creates a tree in Top-Down format, and an appropriate amount of training set must be prepared in advance. The decision tree algorithm classifies a given training set to generate a decision tree and classifies and predicts a given test set. The decision tree has the following characteristics. First, data can be easily classified according to the degree of relevance, and it is easy to assign an action to it. Second, it is expressed in a tree structure that is easy for humans to read, and because it informs the basis of classification or prediction, the results can be easily understood. Third, even if the number of attributes constituting data is unnecessarily large, data classification is easy because it does not affect classification when constructing a tree. Fourth, since data is used without a separate processing process (Min et al., 2014).

2.5. Linear Regression

According to Ji-Hui MUN's study, linear regression, which started with statistics, is effectively used as a method for modeling and inference Through learning about data distribution, linear regression of artificial intelligence and machine learning is a way to find: Variables through analysis and modeling of one or more independent interrelationship data, which models relationships with method dependent variables. There are several independent variables, through which the prediction of the value of the dependent variable, which should be, is inferred from the modeling results as follows. Learn from existing data.

Linear regression methods are widely used as analysis methods for predicting the following. Modeling results and multiple input variables when it is possible to quantify the results to be predicted. To predict the results during modeling, the regression model is typically defined as Analyzing the results generated using least squares regression and predictions (Lim, 2018). In addition to this,,If there is a conditional expected value X1 = m1(X2, …, Xn) = E (x1|X2 = x2, …, Xn = xn) with (X1, X2, …, Xn) as an n-dimensional probability variable, this is called the regression function or regression curve of X1 to (X2, …, Xn). In particular, the regression of X1 to (X2, …, Xn) when m1(x2, …, xn) = α + β2x2 + …βnxn is called linear. In fact, when the above equation is accurate, an approximate linear regression function is obtained and used according to the least squares method, etc. that minimizes E[(X1-α-β2X2-…-βnxn)2].

2.6. Supervised Learning

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances (Kwak, 2020). Supervised Learning is a learning model that takes a set of labeled training data and creates a discriminant that analyses the data using various algorithms and then uses the algorithms discovered on a new data set to produce results. Supervised learning is divided into regression and classification. Since both classification and regression are supervised models, they have in common the fact that they learn from labeled input data. The difference between classification and regression is that classification results in a fixed value, whereas regression means that the resulting value can be any value within the range of the data set.

3. Body

3.1. Dataset Preprocessing

Through the Microsoft Azure Machine Studio, the train.csv file with Rows:7000 columns:24 is removed through the Select Columns in data set process, and 558 data are deleted during the Clean Missing Data process. Since the original data obtained when collecting data is not data for analysis, machine learning with the original data may result in inaccurate results. Therefore, data preprocessing processes such as missing value processing, outlier processing, and noise removal are required. This is an important process because the learning results vary depending on the pretreatment results. After that, the process of matching the data type of the data value changed the data type to categorical, not numeric, for the education level and income level. Since then, 70% of the entire dataset has Dataset Preprocessed (dataset preprocessing) the remaining 30% of the learning data as test data.

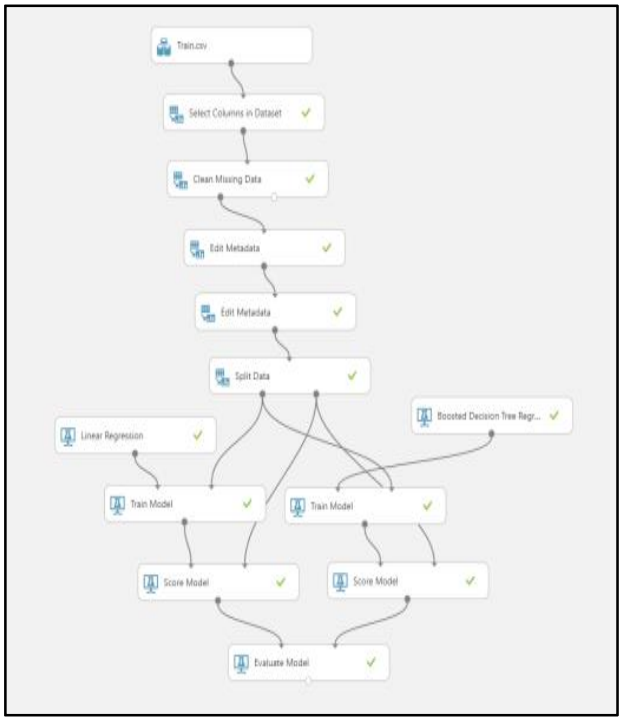

Figure 1: Azure Machine Alogotism

The above figure describes the process of machine learning. It distinguishes data preprocessed and preprocessed data based on the dataset. After that, the process of going through the score model and the evaluation model by educating the data is shown as a picture.

3.2. Model Composition

A model should be selected to execute an artificial intelligence process based on the preprocessed Dataset.

The first is a well-known conventional technique, which uses the Linear Regression model to predict linear relationships between features and labels, and the second is an algorithm that utilizes non-complex data, and is used to construct and execute a Boosted Decision Tree Regression model that influences decisions of the next node.

3.3. Train Model

The training model refers to learning and training data that has been collected and preprocessed. As a result of this learning, an ai model is created, and only through this process can it be expressed as learning data of artificial intelligence. In the process of learning the data, we learned through linear regression and boosted decision tree regression, and the label set the working time as a column of labels and proceeded with model learning. Later visualization showed which columns had the greatest effect on linear regression and how each item was affected by Boosted Decision Tree Regression in the form of a tree.

3.4. Score Model & Evaluate Model

The training data are tested by combining data learned by linear regression and boost decision tree regression models with data segmented during the dataset preprocessing process, and the resulting values are represented visually. Afterwards, the data learning results should be visually identified through the model evaluation process, and then the information needed to detect the determinants and errors evaluated by each model, compare the models, find the areas with errors, correct them, and understand the predicted values.

Table 1: Evaluate Model Visualization

Metrics	
Mean Absolute Error	3.192894
Root Mean Squared Error	4.16142
Relative Absolute Error	0.356592
Relative Squared Error	0.154759
Coefficient of Determination	0.845241

This requires evaluating the model process and evaluating which parts of the model should be modified to ensure that the model is set correctly, and the resulting value leads to good results in the artificial intelligence model.

<Table 1> above is a diagram showing the evaluation of the model when the linear regression algorithm is selected. This shows a model evaluation with a high decision coefficient of about 0.845 and can be said to be a well-designed artificial intelligence model. Although the error range and error value should be additionally evaluated and designed, it is evaluated that they were designed well due to the high coefficient of determination. In addition, **<Table 2>** above is a picture showing the model evaluation of the boosted decision tree. This highly regarded than linear regression of the coefficient of determination. It is also evaluated as a more suitable model to exhibit lower errors.

Table 2: Evaluate Model Visualization

Metrics	
Mean Absolute Error	2.789849
Root Mean Squared Error	3.948173
Relative Absolute Error	0.311579
Relative Squared Error	0.139305
Coefficient of Determination	0.860695

4. Conclusion

In this paper, we made a prediction about the purpose of changing jobs for office workers. The data on the employee's data and the behavioral information on the turnover rate were obtained from Kargle. In Korea, there are workers in the form of many changes in the company. Many of these turnover causes losses on the company's part, so in order to eliminate this form, many workers hope to work long hours. In this paper, we used Microsoft-based machine learning tools to analyze this form and learn the data to finally produce the desired results. This allowed the extraction of labels to make workers long-term workers and verified that such models were designed correctly through determinants. Other AI models will also be available in this process. For example, there are regression models of Ordinal, Poison, Bayesian Linear, Neural Network, Decision Forest, and Fast Forest Quantile. These regression equations are not suitable for our use, but are sufficient for other example processes, and except for the two regression equations, we will analyze the results to estimate long-term workers' expectations for each row and provide information to the company.

References

An, S. H., Yeo, S. H., & Kang M. S. (2021). A study on a car Insurace purchase Prediction Using Two-Class Logistic

Regression and Two-Class Boosted Decision Tree. *Korean Journal of Artificial Intelligence, 9*(1), 9-14.

Choi, J. W., Shin, D. W., & Lee, H. J. (2021). Prediction of turnover rate according to satisfaction and dissatisfaction factors of IT company employees: Using Topic Modeling and Machine Learning. *Korean Journal of Data Information Science. 32*(5). 1035-1047.

Kang, M. S., & Choi, E. S. (2021). *Machine Learning: Concepts, Tools And Data Visualization*, Seoul, Korea:WSPC. Retrieved March 01, 2022, from https://www.amazon.com/Machine-Learning-Concepts-Tools-Visualization/dp/9811229368

Kang, M. S., Kang, H. J., Yoo, K. B., Ihm, C. H., & Choi, E. S. (2018). Getting started with Machine Learning using Azure Machine Learning studio. Seoul, Korea: Hanti media.

Kong, D. A., & Bang, J. H. (2019). A Study on the Repurchase of Automobile Insurance at Expiration. Journal of Industrial Economics and Business, 32(5), 2393-2415.

Kwon, H. H. (2020). *Machine Learning and Finance: Machine Learning-Based Credit Rating Model*. Seoul, korea: Korea Economic Research Institute of KDB Industrial Bank. Retrieved March 11, 2022, from https://eiec.kdi.re.kr/policy/domesticView.do?ac=0000151746&issus=S&pp=20&datecount=&pg

Kwon, K. W. (2016). Relationship between employee turnover and corporate performance: An exploratory study considering the turnover of high performers and non-high performers. *Labor Policy Studies, 16*(1), 1-26.

Nam, Y. J., & Shin, W. J. (2019). A Study on Comparison of Lung Cancer Prediction Using Ensemble Machine Learning. *Korean Journal of Artificial Intelligence, 7*(2), 19-24.

Yoo, S. H., Park, I. S., & Kim, Y. M. (2017). A Study on the Influence Factors and Reasons of Unmet Dental Treatment in Adults Using Decision Tree. *Journal of Health and social studies, 37*(4), 293-294.