Prediction of Employee Turnover in Organizations Using Machine Learning Algorithms: A Decision Making Perspective

Zeynep Kaya(✉) and Gazi Bilal Yildiz

Hitit University/Industrial Engineering, Çorum, Türkiye
zeynep_k54@hotmail.com, bilalyildiz@hitit.edu.tr

Abstract. Digitalization can be defined as the transfer of activities performed in a field to digital environments. The application of digitalization in industry is revolutionary. The digitalization in industry can include applications such as collecting, analyzing, and managing company data with digital technologies, digitally monitoring and controlling the transfer of information between departments, and thus optimizing processes. In human resources management, digitalization can facilitate employee management in a variety of ways, increasing productivity and enabling better decisions. Human resources (HR) departments can develop more effective human resource management strategies by taking into account the amount of time employees are likely to work in the organization while making decisions such as incentives, bonuses, salary increases, and promotions. In this study, a decision support system is proposed to assist HR in determining the most appropriate departments for employees by predicting the potential working hours of current or new/to be hired employees in the organization. To estimate the potential work hours, we have used machine learning techniques that are widely used in the literature. We have adopted an assignment algorithm with work hour prediction to determine of the most suitable departments for employees. An application is carried out on a data set that has been published in the literature, and the results are discussed.

Keywords: Industry 4.0 · Human Resource Analytics · Machine Learning

1 Introduction

Employee turnover is a significant problem in organizations. It negatively impacts a wide range of issues, from morale and productivity to project continuity and long-term growth strategies. These problems result in a significant loss of time and money for the organization. In addition, an organization's production speed and quality can be negatively affected by the turnover of experienced and skilled employees. Therefore, predicting an employee's intention to resing gives the organization the opportunity to take preventive action. Predicting the period during which employees are likely to leave allows organizations to make decisions such as incentives, bonuses, promotions, etc. more effectively.

Z. Şen et al. (Eds.): IMSS 2023, LNME, pp. 118–127, 2024.
https://doi.org/10.1007/978-981-99-6062-0_12

Machine learning algorithms, have had great success in the prediction of future events based on historical data. In this study, a machine learning based prediction system is proposed for the solution of the employee turnover problem in organizations. Thus, the prediction of employee turnover period can be evaluated together with the performance of employees and their contributions to the organization, and can provide great benefits to organizations in determining policies according to the employees. In addition, enterprises can reconsider employees' career plans and reorganize working conditions to improve performance and productivity according to the prediction of employee turnover.

Another problem that is frequently encountered in the organizations is to determine the department in which the employees can make the greatest contribution to the organization. Performance assessment, which is considered in the framework, is a planned process that evaluates the development ability of the individual and his/her contribution to the success of the organization. It also reveals what training, reward, development and motivation the organization should provide to the employee [19]. It is not always easy to determine the appropriate department because employees have different skills, interests, and experiences. Since the determination of the appropriate departments for the employees has a direct impact on the success of the business, the solution of this problem will make a significant contribution to the business. The framework proposed in this study can be used as a decision-support system for determining the appropriate departments for employees.

In literature, turnover refers to the sum of intangible assets such as knowledge, skills, experience, creativity and other mental abilities of an employee who leaves the organization. The loss of such assets is an important factor that can reduce the value of the organization and at the same time reduce its competitive advantage [1]. The focus of this analysis is on the optimal utilization of employees. A meta-analysis review of human resource studies [2] found that the strongest predictors of retention were age, tenure, compensation, overall job satisfaction, and employee perceptions of fairness. Other similar research findings have suggested that personal or demographic variables, particularly age, gender, ethnicity, education, and marital status, are important factors in the prediction of voluntary employee turnover [3, 4]. Salary, working conditions, job satisfaction, supervision, promotion, recognition, growth potential, burnout, etc. are other characteristics that studies have focused on. [5, 6]. The frequent turnover of employees prevents the formation of a collective data base in the organization. It also reduces customer satisfaction because customers are constantly in contact with new employees. On the other hand, employee turnover leads to an undesirable situation that is the loss of employees may mean the loss of valuable knowledge with them, so it may cause the loss of competitive advantage [7]. Therefore, an organization should minimize employee turnover as much as possible to maintain its competitive advantage. Finding the reasons for employee turnover and preventing it is vital for an organization [8]. However, the use of heuristic methods by managers for this purpose can be difficult and time-consuming due to the consideration of many factors such as employee demographics and working conditions. The use of predictive analytical approaches can provide optimal combinations of employees and departments in the organization by giving managers a general idea of employee resignation rates [9].

The study involves selecting five different regression models for the dataset, comparing their performance, and selecting the best one. Based on this model, an infrastructure for a decision support system has been created. The proposed decision support system uses the results of the regression models as coefficients of the assignment problem to determine the appropriate departments for each employee.

This paper is organized in the following manner: Sect. 2 describes the algorithms used in this paper and their mechanisms. The characteristics of the data set, its preprocessing, and the exploratory data analysis are analyzed in Sect. 3. Section 4 presents the results of the study.

2 Methodology

Machine learning techniques are effective in making predictions. These techniques automatically identify patterns and links in data using statistical algorithms and computer methodologies, which can then be applied to forecast upcoming occurrences or outcomes. They rely on learning from historical/training data to map relevant dependent output variables to new input records based on appropriate independent variable values. Due to their capability to handle complicated correlated factors and their effectiveness in dealing with correlated variables, it is crucial to employ modern forecasting algorithms to obtain the best accuracy.

We can forecast staff turnover rates thanks to the benefits and predictive capabilities of machine learning algorithms. Using machine learning algorithms, a network of regression models was established, and its prediction outputs were utilized to generate an assignment problem. The predictions derived from the regression models were then considered as coefficients of the assignment problem. As a result, a decision-support system was created to identify the most appropriate departments for employees. It also provides information on employee turnover rates to human resources management.

Another problem faced by companies is determining the appropriate departments for employees. With the information obtained in the estimation of the employees' turnover rates, it is possible to predict in which departments the employees will work longer, and this information can be taken into account when determining the employees' departments. Therefore, this information can be used as a parameter in an assignment problem. The estimation of working hours can be used with objective coefficients of the assignment model to determine the most suitable department for employees.

2.1 Machine Learning Algorithms

Five prediction algorithms were used in this study. They are Extra Trees Regression, Random Forest Regression, Bagging Regression, LightGBM Regression, and XGBoost Regression. These five prediction algorithms are used to predict the values of the target variables using the characteristics of the samples. Extra Trees is built by combining many random trees to avoid overlearning. Random Forest is also an ensemble method of combining trees and is designed to produce low-variance and low-bias predictions. Bagging LightGBM is an ensemble method that combines many LightGBM trees and is designed to produce faster predictions. XGBoost is a gradient boosting method that adds new trees by focusing on the errors of previous trees and is popular in many machine learning applications.

Random Forest (RF) is a tree-based ensemble method that was developed to address the shortcomings of the traditional Classification and Regression Tree (CART) method. RF consists of a large number of simultaneously grown weak decision tree learners and is used to reduce both the bias and variability of the model [10]. RF uses bagging to increase the diversity of the trees, which in turn are grown from different training data sets, thus reducing the overall variability of the model. RF makes it possible to assess the relative importance of input features, which is useful for dimensionality reduction to improve model performance in high-dimensional data sets. RF changes an input variable while holding other input variables constant and measures the average reduction in the model's prediction accuracy, which is used to assign a relative importance score to each input variable [11].

The Extra Tree (ET) algorithm is a relatively new machine learning technique that was developed as an extension of the Random Forest algorithm and is less likely to overfit a data set [12]. ET uses the same principle as random forest and uses a random subset of features to train each base predictor. However, it randomly selects the best feature and the corresponding value to split the node. ET uses the entire training data set to train each regression tree [13].

Bagging regression is a parallel ensemble approach that deals with the propagation of a prediction model by including additional training data. This training data is added to the original set using a data imputation method. For each new set of training data, certain observations can be repeated during sampling. After bagging, the probability of each element in the reconstructed data set is the same. Increasing the size of the training data set has little effect on the predictive power. However, if the variation is adjusted to fit the desired result, the variation in the prediction can be significantly reduced. Each set of this dataset is automatically used to train new models [14].

XGBoost is a newly developed machine learning technique that has recently been widely used in many fields. It will be suitable for many applications because it is a well-organized, portable, and flexible approach [15]. As an efficient algorithm that combines the Cause Based Decision Tree (CBDT) and Gradient Boosting Machine (GBM) approaches, this technique has the ability to improve the boosting approach to process almost all types of data quickly and accurately. With these unique features, this algorithm can be efficiently used to develop predictive models by applying regression and classification to the target data set. XGBoost can also be used to process large data sets with many attributes and classifications. This algorithm provides practical and effective solutions to new optimization problems, especially when trade-offs between efficiency and accuracy are considered [16].

LGBM regression (Light Gradient Boosting Machine Regression) uses another innovative machine learning based data processing algorithm for more accurate residual value modeling and prediction. As a newly developed technique, it is designed by combining two novel data sampling and classification methods, namely Exclusive Feature Bundling (EFB) and Gradient-based One-Side Sampling (GOSS) [17]. With these combined features, data scanning, sampling, clustering and classification operations are performed properly and accurately in a short time compared to analogous techniques.

2.2 Assignment Problem

The assignment problem (AP) is defined as the assignment of m workers to n jobs. The classical assignment problem is a special case of transportation problems where the quantity of resources and demand are equal to 1 [20]. According to the quality of workers, assignment problems can be classified into three main categories: assignment models with at most one task per worker, assignment models with more than one task per worker, and multi-level assignment models [21, 22].

An assignment problem has been studied to ensure that workers are assigned to the appropriate departments in order to maximize the working time of workers in the organization under the current conditions. The developed model will identify departments that are likely to be more suitable for employees, thus providing decision support for human resource management. Human resource management can benefit from these suggestions when making decisions about salary, bonus, incentives, etc.

3 Application

3.1 Data Set

The basic knowledge and primary data on Predictive Human Resource Analytics was collected by William Walter [18] from Kaggle website. The data used in this proposed study contains 14,999 observations with each row representing a single employee. The fields in the dataset contain the following 10 variables:

- Satisfaction_level = The satisfaction level takes values between 0 and 1.
- Last_evaluation = The year elapsed since the last performance evaluation.
- Number_project = The number of projects completed while working.
- Average_montly_hours = The monthly average of hours spent at the workplace.
- Time_spend_company = The year(s) spent in the company.
- Work_accident = The employee's work accident status (0 'No Work Accident', 1 'Work Accident').
- Left = The employee's status at the workplace (0 'Not Leaving the Job', 1 'Leaving the Job').
- Promotion_last_5years = The employee's promotion status in the last five years (0 'Not Promoted', 1 'Promoted').
- Department = The employee's department. ("Sales": A, "Accounting": B, "HR": C, "Technical": D, "Support": E, "Management": F, "IT": G, "Product Manager": H, "Marketing": I, "R&D": J).
- Salary = Relative salary level (low, medium, high).

The preprocessing of filters focuses on two types of values in the data set, sample and attribute. The sample values are used for resampling. The samples are divided into two data sets, a training set and a test set. In the data set, 70% of the attribute time_spend_company is allocated for training and 30% for testing.

Figure 1 analyzes the time spent by employees in each department of the company and displays this data in a bar chart. This chart shows how many employees are in each department and the time spent by employees in the company is shown in different colors.

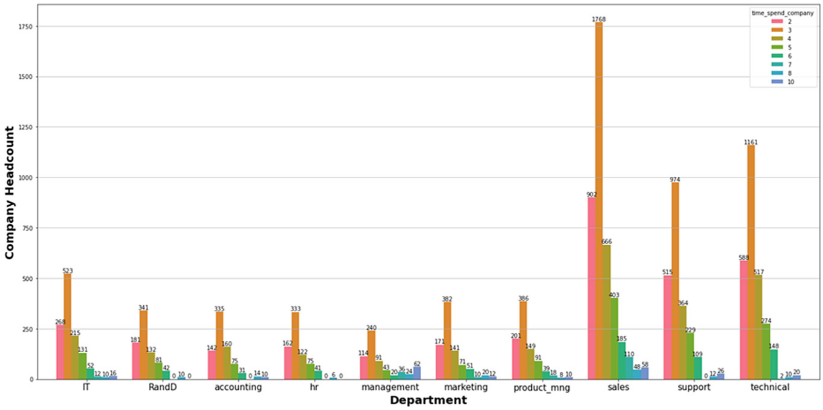


Fig. 1. Departments by Years Spent at the Company

This observation can also show the relationship between the number of employees and the department. Figure 1 shows 268 employees who have worked in the IT department for 2 years. It also shows that the number of employees in the sales, support and technical departments is high, and the number of employees in each department has worked for a maximum of 3 years.

3.2 Data Preprocessing

There are many difficulties in maintaining data in real life. There may be deficiencies in the data, incorrect entries may have been made, or extraordinary situations may have occurred. In such cases, it may be necessary to pre-process the data instead of using it directly in the model. The data preprocessing performed on the data in this study is as follows:

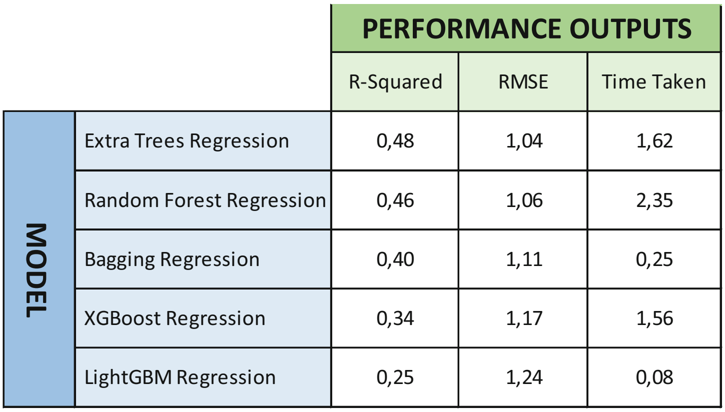
1. Detection of missing data.
2. Changing variable names.
3. Conversion of non-categorical variables into categorical variables.
4. Detection of outlier data.

4 Computational Results

In this section, the effectiveness of the implemented regression models was evaluated. Table 1 shows the performance of each machine learning algorithm in estimating employee turnover. R-Squared, RMSE and Time Taken were used as performance measures.

In Table 1, the performance of each machine learning algorithm is presented for predicting employee turnover rate. In the evaluated case study, the Extra Tree algorithm exhibits the lowest RMSE value based on the number of years spent in the company. Therefore, the Extra Tree algorithm is determined as the most suitable regression algorithm for predicting the duration of each employee's tenure in the company. A staffing model has been created using the predictions generated by this regression model.

Table 1. Performans measures of each ML techniques



		PERFORMANCE OUTPUTS		
		R-Squared	RMSE	Time Taken
MODEL	Extra Trees Regression	0,48	1,04	1,62
	Random Forest Regression	0,46	1,06	2,35
	Bagging Regression	0,40	1,11	0,25
	XGBoost Regression	0,34	1,17	1,56
	LightGBM Regression	0,25	1,24	0,08

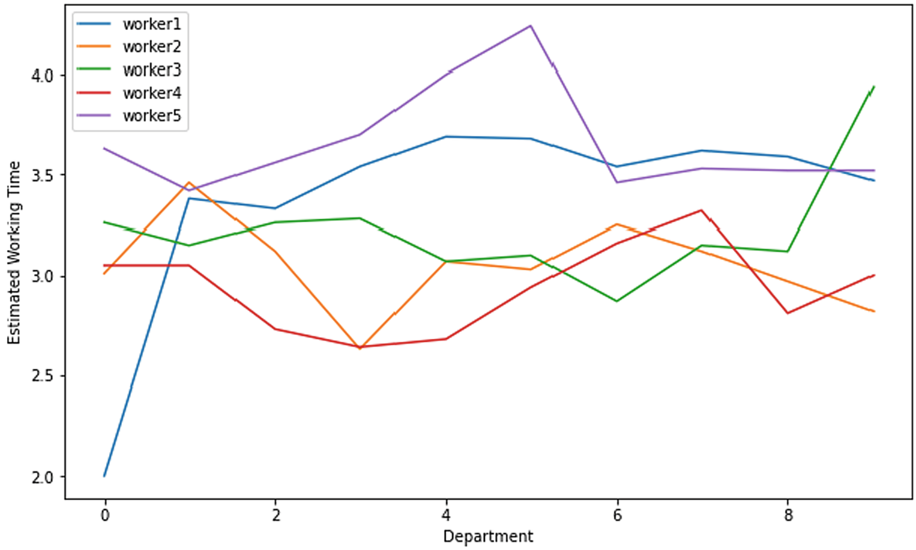


Fig. 2. Estimated work time of five employees according to the different departments

Figure 2 provides estimates of the turnover time of five different workers in different departments. While all other attributes are the same, only the departments were changed and differences in turnover times were observed. Thus, it was concluded that the departments have a significant impact on employee turnover times. For example, when the first employee works in the first department, it is estimated that he/she will leave the job within 2 years, while the estimated time to leave the job increases to 3.69 years when the employee is taken to the fifth department.

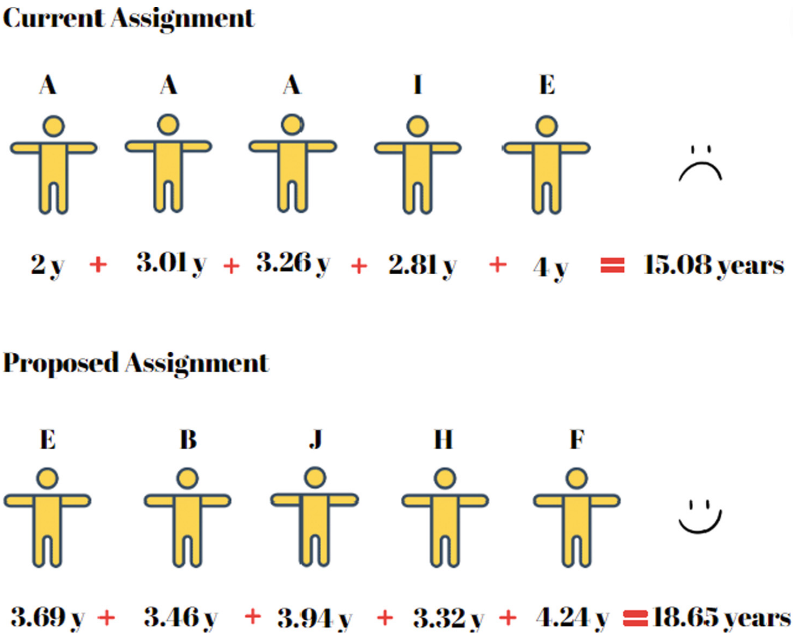

Current Assignment
A
A
A
I
E
2 y + 3.01 y + 3.26 y + 2.81 y + 4 y = 15.08 years
Proposed Assignment
E
B
J
H
F
3.69 y + 3.46 y + 3.94 y + 3.32 y + 4.24 y = 18.65 years

Fig. 3. Current status and recommended status for 5 employees

As an example, Fig. 3 shows the estimated work time for 5 workers in the departments where they currently have jobs. The total working time increases from 15.08 years to 18.65 years when an assignment problem is run to maximize the total working time of these workers. The model determined the appropriate department for each worker and recommended that the 1st worker be assigned to the E department, the 2nd worker to the B department, the 3rd worker to the J department, the 4th worker to the H department, and the 5th worker to the F department. Thus, it was observed that this could further increase the estimated working time of workers.

When assigning workers to departments, special constraints may be added to the assignment problem by taking into account the places where workers can work. Therefore, it is important to verify that the model's recommendations are consistent with the company's goals and priorities. Figure 4 shows the assignments suggested by the decision support system when we extend our example to 50 employees and 10 departments.

A significant difference in total employee work time was found between the current plan and the proposed plan in this example of 50 employees and 10 departments. In the case proposed by the decision support system, when all conditions were equal, the total work time of the employees was 34.52 years higher with only the changes in the departments.

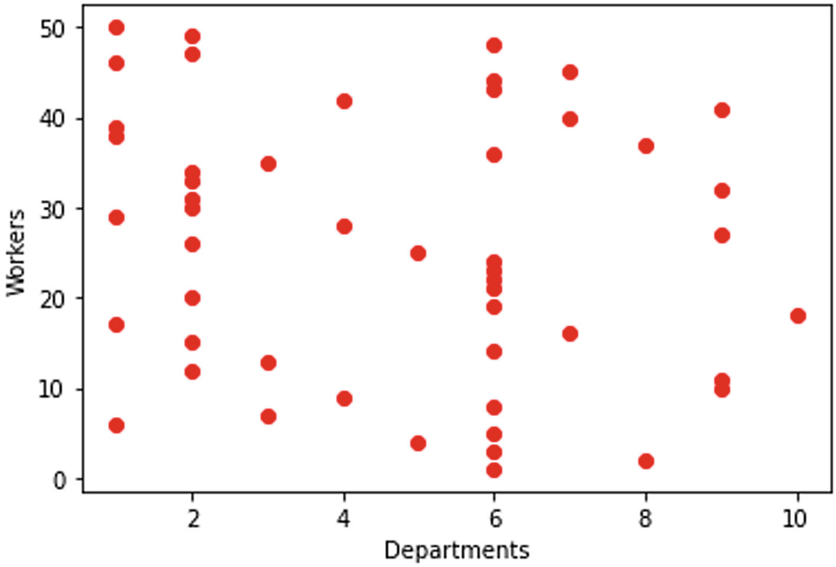


Fig. 4. Results of the assignment model

5 Conclusion

In this study, a decision support system has been created for human resources management. This decision support system estimates the time spent by a worker in the organization. In the same way, the working time of a newly applied candidate can be estimated. These results also serve as a reference for the future changes that HR will make in the working conditions of workers. On the other hand, the developed decision support system makes a suitable department estimation for both existing employees and a new employee. For this purpose, an assignment problem that maximizes the working time of employees is solved when choosing a department for each employee.

References

1. Stoval, M., Bontis, N.: Voluntary turnover: knowledge management – Friend or foe? J. Intellect. Cap. **3**(3), 303–322 (2002)
2. Cotton, J.L., Tuttle, J.M.: Employee turnover: a meta-analysis and review with implications for research. Acad. Manag. Rev. **11**(1), 55–70 (1986)
3. Finkelstein, L.M., Ryanand, K.M., King, E.B.: What do the young (old) people think of me? Content and accuracy of age-based metastereotypes. Eur. J. Work Organ. Psychol. **22**(6), 633–657 (2013)
4. Peterson, S.L.: Toward a theoretical model of employee turnover: a human resource development perspective. Hum. Resour. Dev. Rev. **3**(3), 209–227 (2004)
5. Liu, D., Mitchell, T.R., Lee, T.W., Holtom, B.C., Hinkin, T.R.: When employees are out of step with coworkers: how job satisfaction trajectory and dispersion influence individual-and unit-level voluntary turnover. Acad. Manag. J. **55**(6), 1360–1380 (2012)

6. Heckert, T.M., Farabee, A.M.: Turnover intentions of the faculty at a teaching-focused university. Psychol. Rep. **99**(1), 39–45 (2006)
7. Alao, D., Adeyemo, A.B.: Analyzing employee attrition using decision tree algorithms. Comput. Inf. Syst. Dev. Inform. J. **4**(1), 17–28 (2013)
8. Srivastava, D.K., Nair, P.: Employee attrition analysis using predictive techniques. In: 2017 International Conference on Information and Communication Technology for Intelligent Systems, Ahmedabad, India, pp. 293–300 (2017)
9. Raman, R., Bhattacharya, S., Pramod, D.: Predict employee attrition by using predictive analytics. Benchmarking: Int. J. **26**(1), 2–18 (2019)
10. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
11. Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M.J.O.G.R.: Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geol. Rev. **71**, 804–818 (2015)
12. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006)
13. John, V., Liu, Z., Guo, C., Mita, S., Kidono, K.: Real-time lane estimation using deep features and extra trees regression. In: Image and Video Technology: 7th Pacific-Rim Symposium, PSIVT 2015, Auckland, New Zealand, November 25–27, 2015, Revised Selected Papers 7 (pp. 721–733). Springer International Publishing (2016). https://doi.org/10.1007/978-3-319-29451-3_57
14. Huang, J, Sun, Y, Zhang, J.: Reduction of computational error by optimizing SVR kernel coefficients to simulate concrete compressive strength through the use of a human learning optimization algorithm. Eng. Comput.1–18 (2021)
15. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232 (2001)
16. Ramraj, S., Uzir, N., Sunil, R., Banerjee, S.: Experimenting XGBoost algorithm for prediction and classification of different datasets. Int. J. Control Theory Appl. **9**(40), 651–662 (2016)
17. Ke, G., et al.: Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. **30** (2017)
18. Kaggle, “hr-comma-sep,” Kaggle, Ed., ed (2019)
19. Armstrong, M.: Armstrong’s Handbook of Performance Management: An Evidence Based Guide to Deliver High Performance, (4.Ed), Kogan Page, London (2009)
20. Kara, Doğrusal Programlama. Bilim Teknik, Ankara (2010)
21. Pentico, D.W.: Assignment problems: a golden anniversary survey. Eur. J. Oper. Res. **176**(2), 774–793 (2007)
22. Öncan, T.: A survey of the generalized assignment problem and its applications. INFOR: Inf. Syst. Oper. Res. **45**(3), 123–141 (2007)