Assessing Employee Attrition Using Classifications Algorithms

Fatma Ozdemir
Department of Electrical and Computer Engineering
Abdullah Gul University, Kayseri, Turkey
fatma.ozdemir@agu.edu.tr

Mustafa Coskun
Department of Computer Engineering
Abdullah Gul University,Kayseri, Turkey
mustafa.coskun@agu.edu.tr

Cengiz Gezer
Research & Development Center
Adesso Turkey,Istanbul, Turkey
cengiz.gezer@adesso.com.tr

V. Cagri Gungor
Department of Computer Engineering
Abdullah Gul University, Kayseri, Turkey
cagri.gungor@agu.edu.tr

ABSTRACT
Employees leave an organization when other organizations offer better opportunities than their current organizations. Continuity and sustenance and even completion of jobs are crucial issues for the companies not to suffer financial losses. Especially if the talented employees, who are at critical positions in the companies, leave the job, it becomes difficult for the organizations to maintain their businesses. Today, organizations would like to predict attrition of their employees and plan and prepare for it. However, the HR departments of organizations are not advanced enough to make such predictions in a hand-crafted manner. For this reason, organizations are looking for new systems or methods that automatize the prediction of employee attrition utilizing data mining methods. In this study, we use IBM HR data set and apply different classification methods, such as Support Vector Machine (SVM), Random Forest, J48, LogitBoost, Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Naive Bayes, Bagging, AdaBoost, Logistic Regression, to predict the employee attrition. Different from exiting studies, we systematically evaluate our findings with various classification metrics, such as F-measure, Area Under Curve, accuracy, sensitivity, and specificity. We observe that data mining methods can be useful for predicting the employee attrition.

CCS Concepts
• Information systems → Information systems applications → Data mining

Keywords
Data mining; Classification Methods; Employee Attrition

ICISDM 2020, May 15–17, 2020, Hawaii, HI, USA

ACM ISBN 978-1-4503-7765-2/20/05…$15.00
https://doi.org/10.1145/3404663.3404681

1. INTRODUCTION
Organizations have to make great efforts to increase efficiency and profit in the global market. In addition to solving customer related problems, they also need to deal with employee related problems to increase their revenues. Employees in a company may complain about the conditions that they have in their jobs. Conditions, such as long working periods, intense pace, and low salary might often be the main reasons of their complaints. In such cases, other organizations may recruit the employees by offering better benefits than those of the original company. This recruiting not only affects workflow, but also decreases efficiency and even causes financial losses for companies. Thus, organizations are looking for a solution to avoid these losses by predicting attrition of their employees before they leave. This is because if organizations can predict employee attrition, they will have time to take various actions, such as offering better conditions for their employees.

However, the HR departments of the companies cannot easily predict employee attrition. Utilizing classification methods in Data Mining, the companies can predict when and if an employee will leave the company. As a result of predictions, organizations can take their own precautions and prevent their losses. Researchers studied different classifiers on assessing employee attrition to improve classification accuracy [1-8]. These studies are compared and summarized in Table 1. Although all these existing studies provide valuable foundations to assess employee attrition, there is no internationally accepted standard approach. Furthermore, none of them presents a detailed performance evaluation of different classification methods in terms of accuracy, sensitivity, specificity, F-measure, Area Under Curve (AUC). The aim of this paper is to fulfill this gap and show that not only accuracy measure is critical, but also other performance metrics are critical for assessing employee attrition. In this paper, we aim to address the employee attrition problem by utilizing different classification techniques. To validate our findings, we utilize the IBM HR data set [9]. This data set has 35 features and 1470 samples. We then apply various classification methods, SVM, Random Forest, LDA, J48,

Table 1. COMPARISION OF DIFFERENT CLASSIFICATION METHODS FOR IBM HR

Reference	*Method*	*FS*	*SN*	*SP*	*FM*	*AUC*	*ACC*	*Data set*
İbrahim Onuralp Yiğit et al [1]	SVM	Yes	31.00%	-	53%	-	89.70%	IBM HR
R Shiva Shankar et al [2]	KNN	Yes	-	-	-	-	-	IBM HR
Rachna Jain et al [3]	XGBOOST	Yes	-	-	-	-	90.00%	IBM HR
Kashyap Bhuva et al [4]	LDA	Yes	96.67%	-	92.06%	-	86.39%	IBM HR
Tanmay Prakash Salunkhe [5]	Logistic Regression	Yes	96.96%	54.05%	-	-	91.01%	IBM HR
Mari Maisuradze [6]	Random Forest	No	-	-	-	-	-	IBM HR
Sindhu Velumula [7]	Logistic Regression	Yes	85.48%	-	-	0.72	85.48%	IBM HR
Tanya Attri [8]	GBM	Yes	23.91%	-	-	-	86.97%	IBM HR

FS: Feature Selection, SN: Sensitivity, SP: Specificity, FM: F-Measure, AUC: Area Under Curve, ACC: Accuracy

and KNN, to automatically predict the employee attrition. Different from existing studies, we assess the performance of the prediction methods via various performance evaluation metrics: accuracy, f-measure, sensitivity, specificity, and AUC. We observe that the classification methods can be useful for predicting the employee attrition.

This paper is organized as follows: in Section 2, we give an overview of the related literature. In Section 3, the details of classification methods are discussed. In Section 4, the experimental results have been presented. In Section 5, the paper is concluded.

2. RELATED WORK

Researchers studied different classifiers on assessing employee attrition to improve classification accuracy [1-8]. These studies are compared and summarized in Table 1. Yiğit and Shourabizadeh proposed that using data mining techniques predict the probability of loss for each employee [1]. Furthermore, in that study, Decision tree, Navie Bayes, logistic regression, SVM, KNN, Random Forest algorithms were applied. They also applied feature selection methods. Shankar et al.[2] implemented feature selection and applied Decision tree, Logistic Regression, SVM, KNN, Random Forest, Naive Bayes methods to predict employee attrition. Jain and Nayyar proposed a novel model for predicting employee attrition by using XGboost[3].

Bhuva and Srivastava implemented various classification methods to predict the probability of attrition of employee [4].In their study, LDA gave highest accuracy. Salunkhe proposed predicting employee attrition by using KNN, SVM, GLM, Decision Tree, Random Forest and XGB Tree[5]. Maisuradze used two data sets to predict employee turnover [6]. In the study, IBM HR data is not analyzed in detail.

Velumula predict employee and student attrition by using Logistic Regression, Svm, Random Forest and Naive Bayes algorithms [7]. Attri predicts employee who might leave by using optimal hybrid machine learning model that is integrated oversampling technique (SMOTE) and feature selection technique (SA) [8].In this study, SVM, Random Forest, Logistic Regression, Gradient Boosting Machine were used as machine learning techniques.

All these studies that used IBM HR data are compared and summarized in Table 1. Although all these existing studies provide valuable foundations about predicting employee attrition, none of them presents detailed performance evaluations of different classification methods in terms of accuracy, sensitivity, specificity, F-measure, Area Under Curve (AUC). In this study, our purpose is to show not only accuracy measure is significant but also other performance measures, such as sensitivity, specificity, F-Measure, and AUC.

We have also reviewed similar studies that do not use the IBM HR data set. Krimi uses classification methods for predicting employee performance [10]. In this study, J48 has highest accuracy 92.60%. They used employee data from Kenya School of Goverment. Saad proposed machine learning techniques to improve prediction accuracy of workers' performance [11]. They used data set of Libyan Textile Company. In this study, they applied Bagging Algorithm. Accuracy is 99.16%. They made feature selection to improve accuracy.

Jantan applied data mining techniques for Talent Mangement [12]. They used data set of Maejo University. Nagadevara proposes establishing a link between employee turnover and withdrawal behaviors by using the data mining techniques [13]. Jantawan and Tsai proposed different data mining techniques to predict graduate employment [14]. They used data set of Maejo University. They applied AODE, BayesNet, HBN, Naive Bayes, WAODE. Rajesh proposed classification methods to evaluate employee's attendance to an educational institute [15]. They used data set of private company of Indonesia. These studies used different data set. Although these studies provide valuable insights, none of them presents a, detailed performance evaluations in terms of accuracy, sensitivity, specificity, F-Measure and AUC. The aim of this paper is to fulfill this gap and show that not only

accuracy measure is important, but also other performance metrics are critical for assessing employee attrition.

3. METHODS

In this study, 11 different classification algorithms were used. These are: Support Vector Machine (SVM), Random Forest, J48, LogitBoost, Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), linear discriminant analysis (LDA), Naive Bayes, Bagging, AdaBoost, logistic regression. In addition, we evaluate the classification performances based on different metrics, such as F-measure, Area Under Curve, accuracy, sensitivity, and specificity. The methods used in performance evaluations have been explained briefly below [16].

3.1. Na ïve Bayes (NB)

Naive Bayes classification technique, which is a supervised learning algorithm, is based on Bayes Theorem. It assumes that the features in a class are not connected together. It contributes independently to the probability, even if all the features are linked together. Naive Bayes, which is useful in large and small data sets, may also work in an unbalanced data set. It can perform better than complex classification methods.

3.2. K-Nearest Neighbor (kNN)

One of the simplest supervised machine learning algorithms, the K-Nearest Neighbor (kNN) algorithm is used in pattern recognition and data mining for classification with low error rate. The algorithm finds the k samples that most closely resemble a query point q and concludes that the category of the query point q is the same as the category of the majority of these examples.

3.3. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a simple and effective classification algorithm [17]. The algorithm tries to separate two classes by drawing a vector between the two classes on the plane at the farthest distance from both classes. It is widely used for classification in many areas.

3.4. Random Forest (RF)

Random forest is a tree-based algorithm. Random Forest (RF), which is used for both classification and regression problems, is widely used in machine learning problems. The random forest works by creating multiple decision trees. This algorithm creates multiple random training subsets. It then creates a tree for random training subsets. This technique is called random split selection method.

3.5. J48

J48 is an extension of ID3.The most important features of J48 are pruning decision trees, continuous feature value ranges, derivation of rules, etc.

3.6. Bagging

Producing multiple versions of a predictor, bagging predictor acquires a predictor combined with them. Bagging simulates the process using a specific training set. The algorithm changes the original training data each time by deleting some examples or by duplicating others.

3.7. AdaBoost

AdaBoost algorithm, one of the important community methods, was proposed by Yoav Freund and Robert Schapire. The robust theoretical foundation, very accurate prediction, great simplicity and successful applications are the advantages of AdaBoost.

3.8. Logistic Regression

Logistic regression, a special case of linear regression models, produces a simple probability classification formula. Using the maximum likelihood ratio when generating the equation, LR indicates the statistical significance of the variables. LR is useful in models where the dependent variable is a dichotom.

3.9. LogitBoost

Boosting, a general supervised learning method, combines the "weak" classifier to form a "strong" classifier. Logitboost directly optimizes the possibility of binomial log. LogitBoost minimizes the expectation of a positive loss function. Less sensitive to noisy data, LogitBoost changes linearly with an output error.

3.10. Linear Discriminant Analysis (LDA)

LDA that is a supervised dimensional reduction technique separates the data points of different classes at the highest level. LDA achieves maximum class discrimination by projecting high-dimensional data to a lower-sized area and minimizing simultaneous dispersion of data in the same class.

3.11. Multilayer Perceptron (MLP)

Multilayer perceptron (MLP) is a feed forward artificial neural network model. The main advantage of this method is that it is easy to use. MLP that matches the input data sets into a set of sets creates multiple layers of nodes. Using a supervised learning technique called back propagation, MLP trains the network.

4. PERFORMANCE RESULTS

This section presents the performance results of employee attrition prediction utilizing the data set from IBM HR Data Analysis [10]. In Table 2, the data set feature descriptions are given. Overall, there are 35 features and 1470 samples in IBM HR data set. 26 of features are numeric and the others are categorical. The largest and the smallest values of each feature are also shown. Attrition value of 1233 samples is “no” while attrition value of 237 samples is “yes”. EmployeeCount, Over18 and StandartHours are the same values for each sample. EmployeeNumber does not affect employee attrition prediction.

In the table 3, we present all classification results. In our experiments, we report that the accuracy of Logistic Regression that achieves the highest accuracy is 87.14%.

Table 2. IBM HR DATASET DESCRIPTION

No	*Attribute-Description*	*Value*
1	Age	18-60
2	Attrition	Yes,No
3	BusinessTravel	Travel_Rarely,Travel_Frequently,Non-Travel
4	DailyRate	102-1499
5	Department	Sales,Research&Development,Human Resources
6	DistanceFromHome	1-29
7	Education	1-5
8	EducationField	Life Sciences,Other, Medical, Marketing, Technical Degree, Human Resources
9	EmployeeCount	1
10	EmployeeNumber	1-2068
11	EnvironmentSatisfaction	1-4
12	Gender	Female,Male
13	HourlyRate	30-100
14	JobInvolvement	1-4
15	JobLevel	1-5
16	JobRole	Sales Executive, Research Scientst,Laboratory Technician, Manufacturing Director, Healthcare Representative, Manager, Sales Representative, Research Director, Human Resources
17	JobSatisfaction	1-4
18	MaritalStatus	Single, Married, Divorced
19	MonthlyIncome	1009-19999
20	MonthlyRate	2094-26999
21	NumCompaniesWorked	0-9
22	Over18	Y
23	OverTime	Yes, No
24	PercentSalaryHike	11-25
25	PerformanceRating	3-4
26	RelationshipSatisfaction	1-4
27	StandardHours	80
28	StockOptionLevel	0-3
29	TotalWorkingYears	0-40
30	TrainingTimesLastYear	0-6
31	WorkLifeBalance	1-4
32	YearsAtCompany	0-40
33	YearsCurrentRole	0-18
34	YearsSinceLastPromotion	0-15
35	YearsWithCurrManager	0,17

The sensitivity of Logistic Regression, that achieves the highest sensitivity, is 87.1%. The F-measure of Logistic Regression, that achieves the highest F-measure, is 0.853.

5. CONCLUSION

In this paper, we use various classification methods to address the employee attrition prediction problem. Although all these existing studies provide valuable foundations about predicting employee attrition, none of them presents detailed performance evaluations of different classification methods in terms of accuracy, sensitivity, specificity, F-measure, and AUC. In this study, we use IBM HR data set and apply different classification methods, such as Support Vector Machine (SVM), Random Forest, J48, LogitBoost, Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Naive Bayes, Bagging, AdaBoost, Logistic Regression, to predict the employee attrition. Different from exiting studies, we systematically evaluate our findings with various classification metrics, such as F-measure, Area Under Curve, accuracy, sensitivity, and specificity. The highest accuracies have been achieved by using Logistic Regression, AdaBoost, and LDA algorithms. We observe that data mining methods can be useful for predicting the employee attrition.

6. REFERENCES

[1] Yiğit, I. O., and Shourabizadeh, H. 2017. An approach for predicting employee churn by using data mining. In *2017 International Artificial Intelligence and Data Processing Symposium (IDAP)* (Malatya, Turkey, September 16 - 17, 2017). IEEE. DOI=https://doi.org/10.1109/IDAP.2017.8090324.

[2] Shankar, R. S., Rajanikanth, J., Sivaramaraju, V. V., and Murthy, K.V.S.S.R., 2018. Prediction of employee attrition using data mining. In *2018 ieee international conference on system, computation, automation and networking(icscan)* (Pondicherry, India, July 6 - 7, 2018). IEEE. DOI=https://doi.org/10.1109/ICSCAN.2018.8541242.

[3] Jain, R., and Nayyar, A., 2018. Predicting Employee Attrition using XGBoost Machine Learning Approach. In *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)* (Moradabad, India, November 23 - 24, 2018). IEEE. DOI=https://doi.org/10.1109/SYSMART.2018.8746940

[4] Bhuva, K., and Srivastava, K. 2018. Comparative Study of the Machine Learning Techniques for Predicting the Employee Attrition. In *IJRAR-International Journal of Research and Analytical Reviews (IJRAR) 5.3* (2018). 568-577.

[5] Salunkhe, T. P. 2018. *Improving Employee Retention by Predicting Employee Attrition using Machine Learning Techniques.* Master Thesis. Dublin Business School.

[6] Maisuradze, M. 2017. *Predictive Analysis on The Example of Employee Turnover*. Master Thesis. Tallin University of Technology.

[7] Velumula, S. 2018. *Predictive Analytics of Institutional Attrition.* Master Thesis. Kansas State University.

Table 3. RESULTS OF CLASSIFICATION ALGORITHMS

Method	*K-Fold*	*SN*	*SP*	*F-Measure*	*AUC*	*Accuracy*
SVM	10	85.9%	31.9%	0.822	0.589	85.92%
Random Forest	10	85.6%	28.7%	0.813	0.800	85.58 %
LogitBoost(bl:Random Forest)	10	86.0%	30.2%	0.819	0.809	85.99 %
MLP	10	82.5%	47.5%	0.831	0.774	83.95 %
KNN(k=3)	10	83.2%	32.0%	0.804	0.639	83.20 %
KNN(k=5)	10	84.4%	29.2%	0.806	0.687	84.42 %
LDA	10	86.7%	43.9%	0.848	0.813	86.67 %
J48	10	83.0%	40.8%	0.816	0.570	82.99 %
Naive Bayes	10	79.5%	69.4%	0.809	0.757	79.46 %
Bagging	10	85.6%	33.5%	0.823	0.765	85.58 %
AdaBoost	10	86.7%	38.5%	0.840	0.782	86.73 %
Logistic Regression	10	87.1%	45.4%	0.853	0.816	87.14%

SN: Sensitivity, SP: Specificity, AUC: Area Under Curve

[8] Attri, T. 2018. *Why an Employee Leaves: Predicting using Data Mining Techniques.* Master Thesis. National College of Ireland.

[9] IBM HR Analytics Employee Attrition & Performance | Kaggle. Retrieved February 25, 2020 from https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset/version/1

[10] Kirimi, J. M., and Christopher, A.M., 2016. Application of data mining classification in employee performance prediction. In *International Journal of Computer Applications (0975-8887).* July 2016, Volume 146, 28-35. DOI=http://doi.org/10.5120/ijca2016910883

[11] Saad, R. H., 2018. Use Bagging Algorithm to Improve Prediction Accuracy for Evaluation of Worker Performances at a Production Company. In *Industrial Engineering & Management* 07, 02,*(2018).*DOI= http://doi.org/10.4172/2169-0316.1000257

[12] Jantan, H., Hamdan, A. R., and Othman, Z. A. 2011. Towards applying data mining techniques for talent management. In *International Conference on Computer Engineering and Application.* IPCSIT (Vol. 2,).

[13] Nagadevara, V., Srinivasan, V., and Valk, R., 2008. Establishing a link between employee turnover and withdrawal behaviours: Application of data mining techniques. In *Research and Practice in Human Resource Management (2008). 16(2), 81-99.*

[14] Jantawan, B., and Tsai, C. F., 2013. The application of data mining to build classification model for predicting graduate employment. In *(IJCSIS) International Journal of Computer Science and Information Security, Vol. 11, No. 10, October 2013.*

[15] Rajesh, A., and Vaddepalli, M., K.,2018. Data mining: Evaluating Employee's attendance attribute of an educational institute using classification algorithm based on decision tree. *International Journal of Recent Trends in Engineering and Research*: 144–148. DOI=http://doi.org/10.23883/ijrter.conf.20171201.028.1u4xs

[16] Jiawei Han, Micheline Kamber, and Jian Pei. 2012. *Data mining: concepts and techniques*. Morgan Kaufmann, Waltham, MA.

[17] Kolukısa, B., Hacılar, H., Kuş, M., Bakır-Güngör, B., Aral, A., & Güngör, V. Ç. (2019). Diagnosis of Coronary Heart Disease via Classification Algorithms and a New Feature Selection Methodology. *International Journal of Data Mining Science*, *1*(1), 8-15.