ISSN No. 0976-5697

Volume 11, Special Issue I, May 2020

International Journal of Advanced Research in Computer Science

CONFERENCE PAPER

Available Online at www.ijarcs.info

MACHINELEARNINGALGORITHM TO PREDICT AND IMPROVE EFFICIENCY OF EMPLOYEE PERFORMANCE IN ORGANIZATIONS

Prof. Sohara Banu
School of Computingand Information Technology,
Reva University, Bangalore, India
soharabanu.ar@reva.edu.in

Nipun Agarwal
School of Computingand Information Technology,
Reva University, Bangalore, India
R16CS494@cit.reva.edu.in

Akhil Singh
School of Computingand Information Technology,
Reva University, Bangalore, India
R16CS503@cit.reva.edu.in

Sobiya Shaik
School of Computingand Information Technology,
Reva University, Bangalore, India
R16CS528@cit.reva.edu.in

P.Sai Nikitha
School of Computingand Information Technology,
Reva University, Bangalore, India
R16CS284@cit.reva.edu.in

Abstract— Employee performance has been identified as a critical problem for companies because of its negative effect on operational productivity and long period evolution plans. To solve this problem, companies use machine learning algorithms to anticipate workplace efficiency. Precise forecasts enable organizations to act on preservation or succession planning of employees. However, the data for the modeling issue originates from HR Information Systems; It is generally less in relation to other areas of the companies information systems and is clearly relevant to its objectives This contributes to the presence of redundant values in the data that makes predictive models vulnerable to over-fitting and thus unreliable. This is the central subject based on in this article, and one that has not been discussed conventionally. Using HRIS data from a global retailer, XGBoost is calculated against six widely used supervised classification method and reveals its considerably higher precision for employee performance estimation

.***Keywords***-Performance prediction, machine learning, extreme gradient boosting,supervised classification, regularization.

I. INTRODUCTION

Employee performance issues have gained significance in companies due to their bad impact on issues ranging from morality and efficiency at the workplace to disruption of project continuity and long-term growth strategies. One way businesses can tackle this issue is by forecasting the risk of losing workers using machine learning methods, Thus, giving officials and human resources a foresight to take decisive action on preservation or strategic planning. Most companies did not emphasize investments in efficient human resource solutions that would collect data from employees during their time. The limited experience of benefits and costs is one of the key factors. Return on investment in HRIS[1] is still difficult to measure. This results in data redundancy, which reduces the ability of these techniques to generalize.

This paper discusses the issue of employee performance and the main machine learning techniques being used rectify it. The focus of this paper is to discover the use of gradient boosting as an enhancement on those proposed

algorithm, in particular in their generalisability redundant data prevalent in this field. These are achieved by using HRIS data from a multinational company and classifying the issue of weakening as a problem of classification and designing it using supervised algorithms. The conclusion is reached by using the higher accuracy of the classifier with other methods and by providing a reason for their higher performance.

This paper is formulated in the following way. Section II gives a quick overview of the problem of employee performance, the significance of its resolution and the historical job performed in the application of machine learning methods to solve this issue. Section III examines the 7 distinct supervised methods compared to this paper, including XGBoost. Section IV illustrates the experimental method in terms of the characteristics of the data set, pre-processing, cross validation and the choice of criteria to evaluate precision. Section V sets out the findings of the study and its subsequent response. Section VI concludes the paper by suggesting the XGBoost Performance Prediction Classifier.

Ⅱ. LITERATURE REVIEW ON EMPLOYEE PERFORMANCE

Employee performance may be understood as leakage or dismissal from the intellectual capital of the employer[2]. Most performance literature categorizes results as either voluntary or involuntary. This analysis focuses on mandatory performance. The meta-analytical review of mandatory performance studies[3] found that age, tenure, pay, overall job satisfaction , and employee perceptions of fairness are the strongest predictors of voluntary performance. Other similar study results have shown that individual or demographic variables, specifically age , gender , ethnicity, education and marital status, are important factors in predicting the success of volunteers[4],[6],[7],[8].Other characteristics that studies focus on are pay, working conditions, job satisfaction, supervision, promotion, acknowledgement, potential for development, burnout, etc.[9],[10],[11],[12].

High performance has several adverse repercussions on an organization. Replacement of workers who have specialized skill sets or are specialists in the company sector is difficult. It affects ongoing research and current employee productivity. Obtaining new employees as replacements has its own expenses, such as recruitment expenses, coaching costs, etc. New hires, on the other hand, may have their learning curves in order to reach an equivalent level of technological or company experience with experienced internal staff.

Organizations address this issue by applying machine learning techniques to forecast results, giving them a perspective for operation. The conclusions of the literature review are briefly summarized in Table 1. The following parts of the paper will illustrate the insufficiency of the classification models suggested here to address the scale noise in HRIS..

TABLE I. RELATED WORK ON PERFORMANCEPREDICTION

Research Authors	**Problem studied**	**Data Mining Techniques studied**
Jantan, Hamdan and Othman [13]	Data Mining techniques for performance prediction of employees	C4.5 decision tree, Random For est, Multilayer Perceptron(MLF) and Radial Basic FunctionNetwor k
Nagadevara , Srinivasan and Valk[14]	Relationship of withdrawal behaviors like lateness and absenteeism, job content, tenure and demographics on employee performance	Artificial neu ral networks, logistic regression, classification and regression tree s (CART), classification trees (C5.0) and discriminant analysis)
Hong, Wei and Chen [15]	Feasibility of applying the *Logit* and *Probit*models to employee voluntaryPerformanc e predictions.	Logistic regression mode (log it), probability regression mo del (probit)
Marjorie Laura Kane-Sellers [16]	To explore various personal, as well as work variables impacting employee voluntary performance	Binomial logit regression
Alao and Adeyemo [17]	Analyzing attributes using Decision tree algorithms	C4.5, C

A. Logistic Regression

Logistic regression/ Maximum entropy classification algorithm is among the main linear models for classification. Logistic regression is a common form of regression typically used for conditional or categorical based predictions of variables. It is often used with regularization in the form of L1- or L2-norm based penalties to avoid overfitting. For this paper an L2-regularized logistic regression. By assuming a model for the same, this technique obtains the posterior probabilities and estimates the parameters involved in the assumed model.

B. Naïve Bayesian

Naïve Bayes is a common classification technique that caught attention because of its clarity and performance[21]. Naïve Bayes classifies according to the possibility of arrival, on the basis that all variables are conditionally independent. The classifier needs only a small number of training data to approximate the parameters such as the means and variances necessary for classification variables. It also handles actual, discrete data[22].

The basic reasoning for using the Bayes rule for artificial intelligence is as follows: We use the training data to learn P (X) and P(Y) estimates to train a target function fn: X al Y which is the same as, P (Y). The use of certain approximate probability distributions and the Bayes rule could then be classified as new X samples[21].

C. RandomForest

Random Forest algorithm is a popular tree based learning method to an ensemble. The 'assembly' type used here is to inflate. Successive trees in bagging don't rely on earlier trees — each is built separately using a specific data set bootstrap sample. In the end, prediction is taken by a simple majority vote. Random forests differ from standard trees in how every point is separated using the latter's best split of all variables. In a random forest every node at that node is separated using the utmost among a subgroup of randomly selected predictors[23]. This additional random layer makes it strong against overfitting[24].

D. K-Nearest Neighbor(KNN)

The theory of Nearest Neighbor Classification is the labeling of datasets based on the rating of their closest neighbours. It is also useful to find more than one neighbor so that the technique is more commonly referred to as the k- Nearest Neighbor (k-NN) classification[25]].

E. Linear Discriminant Analysis(LDA)

Discriminant analysis requires designing one or more discriminating functions to optimize the variance between the categories in relation to the categoryvariance[14]. Linear Discriminant Analysis is defined as a linear combination of two or more independent variables that best discriminates between two or more different categories or groups..

The z-scores determined by the distinguishing methods are then used to evaluate the likelihood of a class belonging to a particular member or observation. Another crucial point to remember with LDA is that the functions used should be either continuous or linear in nature.

Ⅲ. METHODS

Machine learning classification is of two distinct significances. We can get a collection of observations to determine the existence of classes or groups in the dataset. Or we can be confident that there are a set of classes and the purpose is to create a rule(s) by that we can define a new data point into one of classes established. The previous group is called Unsupervised Learning and thus is known as Supervised Learning[19]. This chapter dealt with classification as supervised learning, because the data comprise 2 sections – running and finished. This segment discusses the concept around various comparative classification methods.

The 2 phases of classification using KNN include evaluating the adjacent data points and then evaluating the class system on the adjacent classes. You can measure the neighbors using distance measurements such as Euclidean distance. Class may be based on majority vote in the community or assessing inversely relative to class.. Before constructing the model based on KNN, the data was scaled to a range of [0, 1].

Ⅳ. EXPRERIMENTALDESIGN

The population being studied was over a period of 14 months a given level of store leadership team of a global retailer. The population selected is spread around various places in the U.S. The estimates were drawn periodically. There are 2 Class codes-Efficient and Ended, 0 and 1. — Workers should have a track of participating in the business for every quarter, up to a quarter of the results (if applicable), by which moment the data set changes the

class mark from being active to being removed. The dataset had 73,009 active or inactive data points labeled with each.

Dataset characteristics were selected based on the analysis listed in article II. Two sources obtained the data: the HRIS database of the company, and the Bureau of Labor Statistics. The HRIS database of the company contained some key features including demographic features; compensation related features including salary etc.; squad associated details like employee turnover etc. The BLS data give main characteristics such as unemployment figures, average household income, etc.

Overall, there were 32 features of which 26 were numerical, while 5 were categorical in nature.

Data pre-processing

The missing values for the categorical variables were imputed using field mode. The missing values were imputed on case analysis for numeric values. Almost no-imputation was done in fields like amount of promotional activities to stop distorting data on the advancement of employees. Knowledge of the inference of some integer values. For instance, the period that has been arbitrated using duration in position since the last promotion was supposed to be a great estimation. Some other statistical parameters, including the mean imputation, have been median-imputed because it treats outliers. The categorical characteristics for the planning of the data are One Hot Encoded, that transformed every one of the different values in the categorical forms into binary forms.

Algorithm	**AUC (Training)**	**AUC (Holdout)**	**Run-time (Training)**	**Maximum Memory Utilization (Of 16 GB)**
Logistic Regression	0.66	0.50	**52 sec**	20%
Naïve Bayesian	0.64	0.59	59 sec	20%
Random Forest (Depth controlled)	0.79	0.51	23 min 10 sec	29%
LDA	0.74	0.52	6 min 51 sec	35%
KNN (Euclidean distance)	0.52	0.5	180 min 12 sec[a]	35%

A. Model validation technique

The dataset was split and kept out into 80/20 training sets. For each algorithm a grid-search was performed using adjusting parameters, like hyper-parameters for regularization or penalty. Based on a 10-fold cross validation of the training dataset, the optimum setup of the hyper parameters for each algorithm was chosen. The system was tested using their optimum condition of the training data set . Growing algorithm's trained system was then used to estimate and check the remaining member group of 15 percent.

B. Evaluation criteria for model(s)

Under the receiver the Operational Feature Gradient area is the factor selected here to evaluate classification accuracy. The feature gradient is a generic 'predictability' factor and facilitates evaluation of classifiers from operational conditions ,i.e.distributionof dataand cost misclassification[30].Additionally,feature gradient is superior to other metrics such as, For example, error value as it tests the possibility that a randomly chosen favorable point would score higher than a randomly chosen biased, similar to the Wilcoxon score test[31].

Template take-time and storage use are often used to evaluate the performance of the classifier. These two steps are important to report, because they build a case from a professional 's perspective to decide that the method is good for practical business issues, to solve scalability and productivity.

C. System specification

All classification methods, except for XGBoost, are used from the scikitlearn module in Python 3. XGBoost classification model used from XGBoost package. The functions were checked on a 16 GB laptop OS Windows 10.

V. RESULTS

TABLE II. MODEL RESULTS

A. Lift Charts

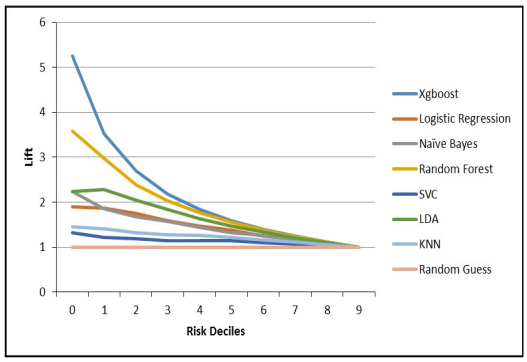
The result was obtained because the estimate is the probability of weakening which is then converted into an employee risk rating. The model was further validated via a lift diagram as shown in Figure 1to test the output of each risk decile. A Lift Chart shows how a particular model improves compared to a random guess.

A. Discussion

The population of this sample is indicative of a workforce distributed across the U.S., consisting of people from various stages of their employment, varying rates of performance and compensation, and diverse backgrounds.Therefore it is intuitive to infer that the most likely outcome is a rule based methodology or a tree based model, taking into account the various themes and classes that naturally occur in the data. The findings in Table 2 confirm the intuition. It is seen that the two Random Forest and XGBoost tree-based classifiers perform better during training than the other classifiers, and that During the trial XGBoost is substantially better than Random Forest. The XGBoost classifier exceeds the other classification model in aspects of accuracy and storage utilization.

Analytically, Random Forest relies on its randomized phases to aid offer greater generalizability, but in this case, as can be seen from the table, it is still inadequate to prevent overfitting. The XGBoost, on the other hand, is trying to updtate new trees complimenting those being installed. Boosting helps to boost preparation for the hard-to-classify data points. Another relevant argument is that, given regularisation or implementation of unpredictability, classification methods other than XGBoost suffer from overfitting, as the case may be. XGBoost provides a solution because of its exceptional underlying regularization and thus works fine for the messy HRIS data.

The XGBoost classifier is often customized for fast, simultaneous tree construction, and is structured to be resilient to faults under the distributed setting[29]. The classifier XGBoost takes the data in DMatrix format. DMatrix is an existing data system used by XGBoost, which is designed for both storage capacity and speed of training. DMatrixes were constructed from various feature arrays and groups.



[2] Fig.1 Lift chart for the classifiers:-

V. CONCLUSIONS AND FUTURE WORK

The importance of employee performance forecasting in this paper presented organizations and the application of machine learning to build performance models. Also highlighted was the key challenge of noise in HRIS data which compromises the accuracy of these predictive models. A global retailer's HRIS data was used to test the XGBoost classifier against six other supervised classifiers traditionally used to create output models. The findings of the study demonstrated that the XGBoost classification algorithm is a superior algorithm in respect of significantly higher precision, relatively low runtime and effective memory use to predict performance. The nature of its regularization makes it a powerful strategy able to handle HRIS database distortion relative to another algorithms, thus solving the key obstacle in this area. For these purposes it is recommended that XGBoost be used to predict employee output accurately so that companies can take action to retain or succession employees.

The studies propose gathering data on the companies interventions for at risk workers and their consequences for future analysis. This will make the model a normative one, and not just answering the question "Who's at risk? "But what should we do, too? ". The system that is well designed to improve precision with just enough hidden layers, but it also needs to investigate the parallelization and proper relevance aspect.

REFERENCES

[1] S. Jahan, "Human Resources Information System (HRIS): A Theoretical Perspective", Journal of Human Resource and Sustainability Studies, Vol.2 No.2, Article ID:46129,2014.

[2] M. Stoval and N. Bontis, "Voluntary performance: Knowledge management– Friend or foe?", Journal of Intellectual Capital, 3(3), 303- 322,2002.

[3] J. L. Cotton and J. M. Tuttle, "Employee performance: A meta-analysis and review with implications for research", Academy of management Review, 11(1), 55-70,1986.

[4] L. M. Finkelstein, K. M. Ryanand E.B. King, "What do the young (old) people think of me? Content and accuracy of age-based metastereotypes", European Journal of Work and Organizational Psychology, 22(6), 633-657,2013.

[5] B. Holtom, T. Mitchell, T. Lee, and M. Eberly, "Performance and retention research: A glance at the past, a closer review of the present, and a venture into the future", Academy of Management Annals, 2: 231-274,2008

[6] C. von Hippel, E. K. Kalokerinos and J. D. Henry, "Stereotype threat among older employees: Relationship with job attitudes and performance intentions", Psychology and aging, 28(1), 17,2013.

[7] S. L. Peterson, "Toward a theoretical model of employee performance: A human resource development perspective", Human Resource Development Review, 3(3), 209-227,2004.

[8] J. M. Sacco and N. Schmitt, "A dynamic multilevel model of

demographic diversity and misfit effects", Journal of Applied Psychology, 90(2), 203-231,2005.

[9] D. G. Allen and R. W. Griffeth, "Test of a mediated performance – Performance relationship highlighting the moderating roles of visibility and reward contingency", JournalofAppliedPsychology,86(5),1014-1021,2001.

[10] D. Liu, T. R. Mitchell, T. W. Lee, B. C. Holtom, and T. R. Hinkin, "When employees are out of step with coworkers: How job satisfaction trajectory and dispersion influence individual-and unit-level voluntary performance", Academy of Management Journal, 55(6), 1360-1380,2012.

[11] B. W. Swider, and R. D. Zimmerman, "Born to burnout: A meta-analytic path model of personality, job burnout, and work outcomes", Journal of Vocational Behavior, 76(3), 487-506,2010.

[12] T. M. Heckert and A. M. Farabee, "Performance intentions of the faculty at a teaching-focused university", Psychological reports, 99(1), 39-45,2006.

[13] H. Jantan, A. R. Hamdan, and Z. A. Othman, "Towards Applying Data Mining Techniques for Talent Managements", 2009 International Conference on Computer Engineering and Applications, IPCSIT vol.2, Singapore, IACSIT Press,2011.

[14] V. Nagadevara, V. Srinivasan, and R. Valk, "Establishing a link between employee performance and withdrawal behaviours: Application of data mining techniques", Research and Practice in Human Resource Management, 16(2), 81-97, 2008.

[15] W. C. Hong, S. Y. Wei, and Y. F. Chen, "A comparative test of two employee performance prediction models", International Journal of Management, 24(4), 808,2007.

[16] L. K. Marjorie, "Predictive Models of Employee Voluntary Performance in a North American Professional Sales Force using Data-Mining Analysis", Texas, A&M University College of Education,2007.

[17] D. Alao and A. B. Adeyemo, "Analyzing employee attrition using decision tree algorithms", Computing, Information Systems, Development Informatics and Allied Research Journal, 4,2013.

[18] V. V. Saradhi and G. K. Palshikar, "Employee churn prediction", Expert Systems with Applications, 38(3), 1999- 2006,2011.

[19] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, Machine Learning, Neural and Statistical Classification. Ellis Horwood Limited,1994.

[20] G. King and L. Zeng, "Logistic regression in rare events data", Political Analysis, 9(2), 137–163,2001.T. Mitchell, Machine learning. 2nd ed. USA: McGraw Hill, 1997.

[21] H. A. Elsalamony (2014), "Bank direct marketing analysis of data mining techniques", International Journal of Computer Applications,85(7).

[22] A. Liaw and M. Wiener, "Classification and regression by randomForest", R news, 2(3), 18-22,2002.

[23] L. Breiman, Random forests. Machine Learning, 45(1), 5–32, 2001.

[24] P. Cunningham and S. J. Delany, "k-Nearest neighbour classifiers", Multiple Classifier Systems, 1-17,2007.

[25] C. Cortes and V. Vapnik, Support-vector networks. Machine learning, 20(3), 273-297,1995.

[26] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting", Journal of computer and system sciences, 55(1), 119-139,1997.

[27] J. H. Friedman, "Greedy function approximation: a gradient boosting machine", Annals of statistics, 1189-1232,2001.

[28] S. Lessmann and S. Voß, "A reference model for customer-centric data mining with support vector machines", EuropeanJournalofOperationalResearch199,520–530,2009.T. Fawcett, "An introduction to ROC analysis", Pattern Recognition Letters 27 (8), 861–874,2006.