2019 Global Conference for Advancement in Technology (GCAT)
Bangalore, India. Oct 18-20, 2019

Machine Learning Based Attrition Prediction

Abhiroop Nandi Ray
Technical Lead
Mirafra Technologies
Bengaluru, India
abhiroop.nray@gmail.com

Judhajit Sanyal
Department of Electronics and Communication Engineering
Techno International New Town
Kolkata, India
sanyaljudhajit@gmail.com

***Abstract*—The use of machine learning techniques and models has become widespread with diverse industries using them to glean greater insights from available data. Probabilistic estimation models are used in many cases, often in combination with other methods such as regression and decision trees. The current paper utilizes probabilistic estimation to predict attrition from the human resource database of a company with close to 1500 employees. The initial model is adaptively refined to improve the prediction capability of the model.**

Keywords—machine learning, probabilistic estimation, attrition, adaptive model, prediction

I. INTRODUCTION

Estimation techniques using machine learning have been used in diverse fields in recent years to increase the accuracy of prediction much beyond standard historical modeling utilizing datasets. The present paper proposes a technique that builds a predictive model which can be applied in diverse fields to detect binary valued events. Employee attrition is studied in the present case.

The paper is organized in the following way. Section II highlights some of the most recent literature in the domain of predictive modeling, focusing on probabilistic models. Section III presents the proposed probabilistic estimation model as well as its adaptive variant. Section IV discusses the results obtained using the models discussed in the previous section. Section V concludes the paper with a focus on future avenues of research.

II. LITERATURE SURVEY

Probabilistic methods have been investigated extensively by researchers working in the domain of event prediction in diverse fields. Statistical models allow for greater accuracy of prediction over more exact mathematical models especially in case the variable being predicted varies with a certain degree of randomness with respect to multiple variables. In recent years, probabilistic models have been combined with machine learning techniques such as ensemble learning to detect and prevent probable aircraft collision scenarios [1]. Estimation of ad hoc wireless network throughput has also been done using this method [2]. Probabilistic models have been combined with deep learning techniques by some researchers to enable the recognition of human activities [3]. Probabilistic prediction models have also been applied to the domain of social network analysis [4]. Detection of binary events such as fault conditions have been successfully performed in diverse scenarios using probabilistic models [5][6]. Random and dynamic scenarios have also been studied using probabilistic prediction, ranging from cloud services to electrical transmission systems [7][8]. Prediction of quality of consumables has also been carried out using these techniques [9].

The estimation of attrition is an extremely important field in the domain of human resources. Staffing functions drive businesses forward, and hence predictive models that allow HR executives and managers to plan for attrition are of great relevance to businesses in the current day. Some endeavours have hence been made in this regard using standard statistical distributions to analytically predict employee attrition [10]. Multiple machine learning techniques have been employed by authors to predict employee attrition. Estimation techniques including Decision trees, Logistic regression and Naïve Bayesian methods have been used in conjunction with classification techniques such as SVM (Support Vector Machines), KNN (K-Nearest Neighbour algorithm) and Random Forest to accurately predict the probability of attrition with respect to employees [11]. The prediction model gives fairly accurate results, however the computational complexity of this approach is quite high due to the application of multiple types of machine learning techniques to derive the estimate. Similar estimation models have been proposed which employ adaptive models employing SVM and KNN techniques in tandem to estimate employee attrition [12]. Comparative studies employing multiple classification and estimation based machine learning techniques for predicting employee attrition have also been carried out by researchers [13]. Novel machine learning algorithms such as XGBoost have also been proposed for predicting employee attrition [14].

The current paper deals with the prediction of attrition of employees of different age groups, educational degrees and educational levels. In contrast to most approaches, the model follows a joint probabilistic Bayesian approach for identification of attrition which ensures that the computational complexity of this approach is lower than most of the other approaches employed by researchers in recent times. The corresponding estimation models are discussed in the following section.

III. ESTIMATION MODELS

The data of 1470 employees are considered. Age, educational field and level of education are the factors used to predict the employee attrition and compared to the actual attrition rates of employees. The attrition probabilities for the group of employees are determined by classification according to the fields mentioned above, namely age, level of education and field of education.

$$P_A(i) = n_A(i)/T(i) \quad (1)$$

$$P_{LE}(i) = n_{LE}(i)/T(i) \quad (2)$$

$$P_{FE}(i) = n_{FE}(i)/T(i) \quad (3)$$

Here $P_A(i)$ is the probability of attrition according to the ith age group, $n_A(i)$ is number of employees belonging to the ith age group leaving the company with $T(i)$ being the total number of employees ith age group. The other parameters in

equations 2 and 3 represent the same variables classified with respect to level of education and field of education. These equations are used to determine the subgroup probabilities for attrition considering each of the variables separately. Following this, multi-classification models are established which divide the employees into groups based on joint classification, for example, employees in the age group 18 to 24 years with a technical degree are placed in a single group. Subgroup and average group probabilities are again computed and the probabilities, as before are factored in to calculate overall attrition. The results of the simulations are shown in the following section.

IV. RESULTS

The factors of age, education level and type of education are used to determine attrition probabilities. The results obtained through classification and segmentation for each of the parameters of age, education level and education type are displayed in tables 1, 2 and 3 respectively. The corresponding tables are displayed as follows.

TABLE I. AGE BASED ESTIMATION OF ATTRITION

Age	Attrition	Retention	Attrition Probability
18-24	38	59	0.39
25-30	62	227	0.21
31-36	66	346	0.16
37-42	27	266	0.09
43-48	19	163	0.1
49-54	14	114	0.11
55-60	11	58	0.16

TABLE II. EDUCATION LEVEL BASED ESTIMATION OF ATTRITION

Education Level	Attrition	Retention	Attrition Probability
1	31	139	0.18
2	44	238	0.16
3	99	473	0.17
4	58	340	0.15
5	5	43	0.1

TABLE III. EDUCATION TYPE BASED ESTIMATION OF ATTRITION

Education Type	Attrition	Retention	Attrition Probability
Life Sciences	89	517	0.15
Medical	64	401	0.14
Marketing	34	124	0.22
Human Resources	7	20	0.26
Technical Degree	32	100	0.24
Others	11	71	0.13

The variation of attrition probabilities for the three factors of age, education level and education type are shown in figures 1, 2 and 3 respectively, which follow.

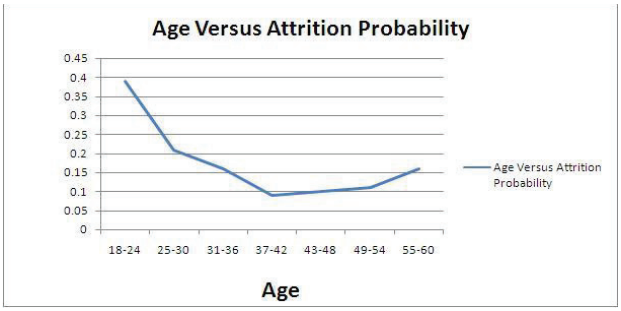

Fig.1. Variation of age with probability of attrition

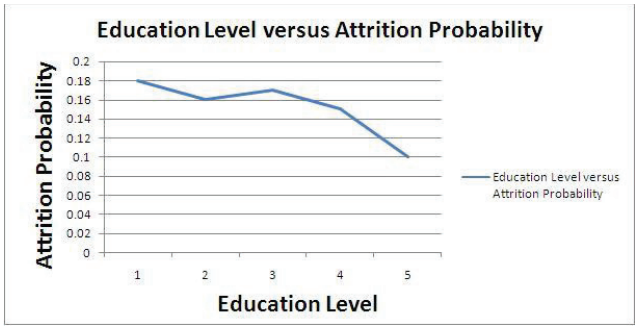

Fig. 2. Variation of level of education with probability of attrition

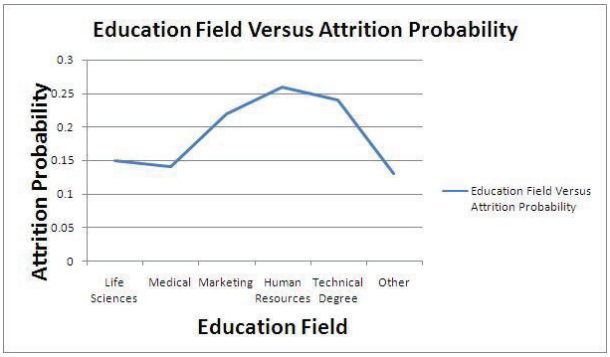

Fig. 3. Variation of field of education with probability of attrition

Now the age group data is combined with the corresponding education type or education level data. In this way a multi-classification model for identification of joint probabilities is formed. The multi-classification model involving age and education level yields a grouped probability distribution set as shown in the histogram in figure 4.

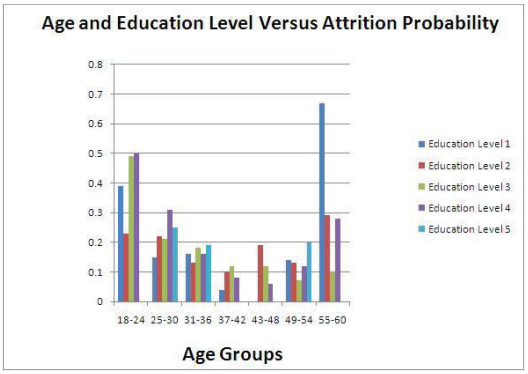

Fig. 4. Variation of age and level of education with probability of attrition

The results obtained in the simulations are used to propose an adaptive machine-learning model, which tries to predict attrition initially using a single variable, increasing the number of classification variables adaptively according to the desired level of accuracy, specified by the HR executives using the model. The results obtained by running the simulation using the probabilistic model as well as its adaptive version are shown in table 4 below.

TABLE IV. PROBABILISTIC ESTIMATION OF ATTRITION

Classification Variables	Type of Method	Error (%)
Age	Probabilistic Estimation	17.43
Educational Field	Probabilistic Estimation	15.2
Level of Education	Probabilistic Estimation	19
Age and Educational Field	Adaptive Probabilistic Estimation	17
Age, Educational Field and Level of Education	Adaptive Probabilistic Estimation	3.5

The results obtained in the preceding table 4 indicate that the adaptive probabilistic model performs much better in comparison to the other models in this present scenario. The adaptive model helps in identification of certain important attritional features evident from figure 4. Employees with highest education level (level 1) show very high attrition (around 67%) in the age group 55-60 years and relatively high attrition (38%) in the age group 18-24 years. Similarly, employees at lower education levels (3 and 4) show high attrition levels (48% and 50% respectively) in the age group of 18-24 years. Alternately, the lowest attrition statistics are observed in the age group of 37-42 years with attrition levels of employees less than or just above 10%.

The results obtained using the adaptive probabilistic estimation model indicate that young employees at very high or moderately low levels of education are likely to leave the company, and the oldest and most educated employees are also extremely likely to leave the company. In this manner, the proposed model allows not only the prediction of attrition but also helps in identification of the attritional pain points from the given dataset.

V. CONCLUSION

The results obtained in this paper clearly show that the adaptive probabilistic estimation model gives very good prediction results considering the fact that only a small set of variables from a much larger set of available variables were taken to predict employee attrition. Additionally, the probabilistic models for each of the individual models allowed for determination of minimal and maximal attrition groups considering the maximal and minimal attrition convergences for each of the three variables used to predict employee attrition. The identification of maximal attrition groups can enable HR executives to take appropriate measures to combat attrition by identifying the root causes behind the attrition in these groups. Corrective actions and appropriate process modifications will enable HR personnel to ensure that attrition decreases beyond significant levels for such groups. Attrition statistics for employee segments based on age groups also allow some specific insights into the ensemble attrition data. The intra-group and inter-group statistics provide a clear indication of the variation of attrition at different ages and education levels, identifying points of stability as well as critical attrition from the given dataset.

In future, if time-series data for a number of years is available for the given dataset, it will allow the current model to predict employee attrition well in advance. Larger datasets would enhance the capabilities of the model. Adaptive micro models for each of the variables involved in predicting attrition would allow the use of year-to-year weighted probability distributions for each of the subgroups. This model can also be used in conjunction with decision trees to optimize the recruitment and retention process for companies and add value to the HR processes followed therein.

ACKNOWLEDGMENT

The authors are grateful to the Department of Electronics and Communication Engineering, Techno International New Town, for complete support in pursuing this research.

REFERENCES

[1] J. Xu-rui, W. Ming-gong, W. Xiang-xi and W. Ze-kun, "Application of ensemble learning algorithm in aircraft probabilistic conflict detection of free flight," 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, 2018, pp. 10-14.

[2] D. Salmond, "Blind estimation of wireless network topology and throughput," *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, Baltimore, MD, USA, 2019, pp. 1-6.

[3] F. M. Rueda, S. Lüdtke, M. Schröder, K. Yordanova, T. Kirste and G. A. Fink, "Combining Symbolic Reasoning and Deep Learning for Human Activity Recognition," 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Kyoto, Japan, 2019, pp. 22-27.

[4] F. Sasso, A. Coluccia and G. Notarstefano, "Interaction-Based Distributed Learning in Cyber-Physical and Social Networks," in IEEE Transactions on Automatic Control, 2019.

[5] Y. Zhang, M. Li, Z. Y. Dong and K. Meng, "A probabilistic anomaly detection approach for data-driven wind turbine condition monitoring," in CSEE Journal of Power and Energy Systems, 2019.

[6] M. Ammar, G. B. Hamad and O. Ait Mohamed, "Probabilistic High-Level Estimation of Vulnerability and Fault Mitigation of Critical Systems Using Fault-Mitigation Trees (FMTs)," 2019 IEEE Latin American Test Symposium (LATS), Santiago, Chile, 2019, pp. 1-6.

[7] C. Qiu and H. Shen, "Dynamic Demand Prediction and Allocation in Cloud Service Brokerage," in IEEE Transactions on Cloud Computing, 2019.

[8] J. Pérez-Rúa, K. Das and N. A. Cutululis, "Lifetime estimation and performance evaluation for offshore wind farms transmission cables," 15th IET International Conference on AC and DC Power Transmission (ACDC 2019), Coventry, UK, 2019, pp. 1-6.

[9] S. Aich, A. A. Al-Absi, K. Lee Hui and M. Sain, "Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques," 2019 21st International Conference on Advanced Communication Technology (ICACT), PyeongChang Kwangwoon_Do, Korea (South), 2019, pp. 1122-1127.

[10] M. Singh et al., "An Analytics Approach for Proactively Combating Voluntary Attrition of Employees," 2012 IEEE 12th International Conference on Data Mining Workshops, Brussels, 2012, pp. 317-323.

[11] R. S. Shankar, J. Rajanikanth, V. V. Sivaramaraju and K. VSSR Murthy, "PREDICTION OF EMPLOYEE ATTRITION USING DATAMINING," *2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA)*, Pondicherry, 2018, pp. 1-8.

[12] S. S. Alduayj and K. Rajpoot, "Predicting Employee Attrition using Machine Learning," *2018 International Conference on Innovations in Information Technology (IIT)*, Al Ain, 2018, pp. 93-98.

[13] S. Yadav, A. Jain and D. Singh, "Early Prediction of Employee Attrition using Data Mining Techniques," *2018 IEEE 8th International Advance Computing Conference (IACC)*, Greater Noida, India, 2018, pp. 349-354.

[14] R. Jain and A. Nayyar, "Predicting Employee Attrition using XGBoost Machine Learning Approach," *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, Moradabad, India, 2018, pp. 113-120.