ISSN No:-2456-2165

Prediction of Employee Turnover Using Light GBM Algorithm

Rajat Keshri, Dr. Srividya P,
Department of Electronics and Communication, R V College of Engineering

Abstract:- Many employees leave the organisation or the company depending on various factors. This effects the growth and production of the company in many ways. The companies and many MNCs use machine learning methods to predict a turnover of workers to solve this problem. Such predictions help the company in success planning and employee retention. The dataset used in this paper for the above problem comes from the Human Resource Information Systems, which are usually different for different companies. Due to the differences of the dataset in different organisations, it results to a noisy data which makes the models to over-fit or produce inaccurate results. This is the main issue which this paper focuses on, and one which has not been discussed traditionally. This paper discusses a new algorithm called the LightGBM, released by Microsoft in 2017. Here, we compare LighGBM with other existing algorithms. Data from the dataset is used to compare LightGBM and other classification algorithms and show LightGBM's high accuracy of prediction.

Keywords:- *Machine Learning; Supervised Classification; Retention Prediction; Gradient Boosting;*

I. INTRODUCTION

Many topics of discussion related to a company are the work environment, workload and work life balance. Due to multiple reasons like over burden or no work life balance, an employee may leave the company. One-way organisations deal with such an issue is by predicting the state of an employee based on how he/she is treated or how he/she is feels in the company. This can be predicted by the input given by the managers and other leads and also by the HR team. But these datasets which are created are highly prone to have a lot of noisy and inaccurate data. Most companies did not prioritize investments in powerful HRIS technologies that would collect the data from an employee during their tenure. Poor perception of advantages and costs is one of the key factors. Measuring the return on investment at HRIS is also difficult. This results in a noisy dataset, which in effect mitigates these algorithms' generalization capability. [1]

This paper discusses the problem of the employees leaving the company and attrition rate and the machine learning approaches to solve the above problem have been discussed. Here, the unique approach of Light gradient boosting machine algorithm is discussed and is experimented with. This is done by using the data from HRIS data set. The conclusion of this paper will be stating the superior accuracy of the LGBM algorithm over the other traditional algorithms. This paper is structured into different sections - II outlines the problem statement and the need to solve it; III explains the various supervised machine learning techniques, which includes Light GBM; IV explains the experimental analysis and also explains the dataset, with pre-processing, the metrics used; V tells the experimental results; VI is used to conclude the paper by recommending the Light GBM classifier. [1] [2]

II. LITERATURE REVIEW

Employee attrition can be perceived as the employee leaving from the hiring company's intellectual capital. A turn over can be accidental or volunteering. This paper focuses on volunteer turnover. By a study, the main variables for estimating the turnover were overall work satisfaction, age, tenure, salary etc. Other related research showed that personal variables like age, ethnicity, education etc. were also important factors in prediction. Certain characteristics based on studies are pay, the condition of work, supervision, promotion, satisfaction of job etc. [2-8]

High turnover ratio has multiple negative effects on a company. It is tedious and hard to find a new employee with specific knowledge and skill valuable for the company. It also directly creates impact on the productivity. Hiring new employees are also costly, as it requires the whole process of shortlisting based on knowledge which the company requires and then training the selected new employees to bring them to a certain level. [9-11]

Companies thus use certain machine learning and mathematical algorithms to prevent such attrition.

III. METHODS

Classification in machine learning are done in two waysSupervised and Unsupervised. Supervised learning involves the dataset given with the output each data point should produce. The algorithm learns the patterns which produce the certain output and try to generalize it with supervision. Supervised learning basically contains the output labels to be predicted in the dataset and learns how to predict those values by backtracking and generalization. Unsupervised learning trains on the data and tries to generalize blindly without knowing what category each data point belongs to. It creates a pattern and generalizes the data points based on its features and creates output labels for them during the training process. This paper deals

with classification as supervised training and there are two cases or labels named terminated and active. [19]

A. Logistic Regression

Classification between linear models is usually done by logistic regression. It is a basic linear classification algorithm. Logistic regression works on the sigmoid mathematical function. It is used to predict binary and categorial classes. It's often used with regularization which avoids over-fitting. The equation of the model is given below in (1):

$$p(\text{churn}|w) = \frac{1}{1+e^{-\left[w0+\sum_{i=1}^{N} wiXi\right]}} \quad (1)$$

The parameter w is estimated by maximum likelihood technique. [17]

B. Random Forest

This algorithm is a tree-based algorithm. It uses something called "Bagging". Bagging here means that successive upcoming trees do not depend on previous trees. This means that each tree is constructed independent of the previous tree using the dataset. After this, a vote is taken to predict the value appeared the greatest number of times.

Random forests are different from standard trees. In this algorithm, each node is split based on some prediction variable and probability functions, which split the best subset of trees [13]. This makes it robust and avoids over fitting.

C. Naïve Bayesian

Naïve Bayes is a common technique of classification, It is a very simply technique. This algorithm predicts values solely based on probabilities. It treats very variable a independent of each other. This requires a tiny part of dataset values and mean and variance of this dataset is estimated with the small amount itself.[15]

The Bayes' rule is as follows: Target function is defined as,

P (Y|X) = X → Y,

The training data to learn estimates of P (X|Y) and P(Y). By Using the above calculated probabilities and Bayes' rules, new X values are classified into different labels.

D. K Nearest Neighbours

KNN classification is a machine learning algorithm in which the algorithm tends to find out patters within the training dataset and then classifies based on the patters. It tries to plot the data points with similar features as close as possible to each other. It is also called k-Nearest Neighbour (k-NN) Classification [16].

E. LightBGM

This algorithm can be defined as high performance gradient boosting algorithm. It is fast and robust. It is used for ranking and classification. It is based on decision tree algorithm.

LGBM grows a tree vertically while other algorithm grows trees horizontally. This basically means that Light GBM grows tree leaf-wise while another algorithm grows level-wise. It does not convert to one-hot coding, and is much faster than one-hot coding.

Below, figure 1 is a representation of the Light GBM.

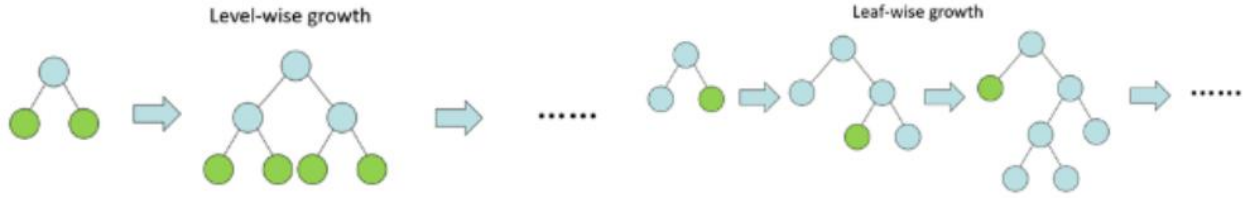

Fig 1:- Working of LGBM Algorithm

Benefits of Light GBM

- High training speed and performance: LGBM has a main advantage that it requires very less time for training and is highly optimized.
- Low memory utilization: LGBM uses very low memory while training and testing. It also requires lesser amount and smaller dataset for training and yet produces good accuracy.
- Good accuracy and use of Boosting: This algorithm uses boosting and gives a better accuracy compared to other boosting algorithms. [3]

Mathematics behind LGBM

Boosting in machine learning belongs to ensemble model family where the focus is on reducing primarily bias. It means that first built a model, find its residual and build another model on the residual. Mathematically, it can be represented as below in expression (1).

$$f_0(x) = argmin(L(y,r))(1)$$

ISSN No:-2456-2165

Now, calculate the residual (ri = (y-gamma)) and build second model h1(x) on {xi,ri}. Add h1(x) to f0(x) and get new improved model f1(x), as shown in expression (2)

$$f_1(x) = f_0(x) + h_1(x)(2)$$

The above equation from (2) is general equation for a boosting algorithm. There are several ways to decide what proportion of hm(x) to be added to fm-1(x). The equation will look like as show in (3)

$$f_m(x) = f_{(m-1)}(x) + alpha * h_m(x)(3)$$

The term gradient in gradient boosting comes from gradient descent incorporation into boosting. A gradient descent-based method is used to decide alpha or step size. To calculate alpha, at say iteration m, first pseudo residual (rim) is calculated and new model hm(x) is built on {xi, rim}. Pseudo residual is calculated by (4):

$$r_{im} = -\frac{\partial L}{\partial f(x_i)}(y_i f(x_i))(4)$$

Next, calculate alpha such that the Loss function is minimized as shown in (5).

$$alpha = argmin(L(y_i f_{(m-1)}(x_i + alpha * h_m(x))))(5)$$

Now plug in that values of alpha and hm(x) to get fm(x). In GBM, the algorithm is same as in gradient boosting. The model is decision tree based i.e. f(x) and h(x) are CART trees. For a tree with T leaves, model hm(x) can be written as (6):

$$h_m(x) = \sum_{j=1}^{T} b_{jm} * I_{Rjm}(x)(6)$$

"bjm" is the value predicted in the region Rjm (leaf j). If hm(x) for a tree is plugged in to gradient boosting equation, there will be alpha and bjm. In GBM, alpha and bjm are combined to get step rate for each leaf. So, there will be T alphas (step rate) in a tree with T leaves. The equations for GBM becomes (7) and (8):

$$fm(x) = f(m-1)(x) + \sum_{j=1}^{T} alpha_{jm} I_{Rjm}(7)$$

$$alpha_{jm} = argmin_{alpha} \sum_{x_i \in R_{jm}} L(y_i f_{m-1}(x_i) + alpha)(8)$$

IV. EXPERIMENTAL DESIGN

The dataset chosen is a distribution across different locations in the US. The different labels present in this dataset are Terminated (0) and Active (1). The dataset has employees, with each employee being active (0) or present in the company for 4 months. After this, the employee leaves the company, and the class label hence changes to terminated (1).

The dataset is taken from Kaggle. The dataset has many features like pay, age, team related features etc. which are used for the prediction. There were 33 features in the dataset (27 numeric, 6 categorical)

A. Data Preproccesing

First the dataset cleaning was done. This involved removal of all bad and noisy data. For every missing numerical data, zero was added. And for every missing categorial data, that particular row was removed. Zero was added on fields like number of promotions for the employees with no data, to prevent the model to train with more accuracy. Next, the categorical features were One-Hot Encoded, and were converted to binary fields.

B. Model Training

The dataset was split in to the ratio 80:20. 80 for training and 20 for testing purpose. Regularization was done and penalty hyper-parameters were set, for each algorithm. The training dataset was used to train the model with their ideal configuration. The trained models were made to predict on the 20% of the data.

C. Evaluation

The evaluation for all the different algorithm is done based on the prediction score or the prediction accuracy, which is given on its optimal training conditions. The models were tested on the testing dataset, which is 20% of the full dataset. The accuracy achieved on that dataset gave metrics for evaluation between different algorithms. A confusion matrix is plotted for comparison between the different models.

V. RESULTS

The dataset contains multiple employees from an organization, of different age, gender, pay, teams and different background. These employees worked for at least 4 months before they left the company voluntarily or forced to leave. This dataset is found from Kaggle website and the attrition rate of an employee given can be predicted by the models trained using this dataset.

Algorithm	**Prediction Score/Accuracy**
Logistic Regression	76.066
Random Forest	97.355
Naïve Bayesian	76.55
K Nearest Neighbours	79.48
LightGBM	98.237

Table 1:- Prediction Score Table

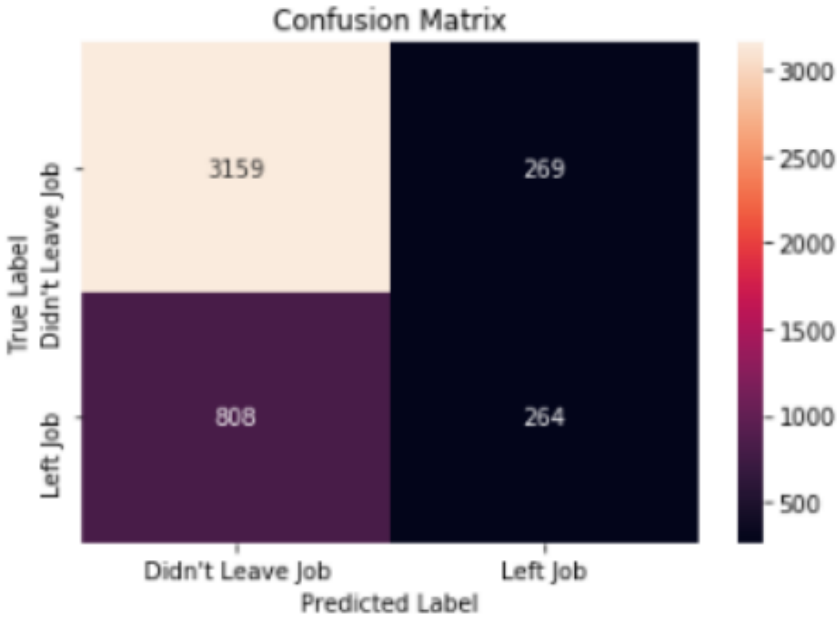

Fig 2:- Logistic Regression Confusion Matrix

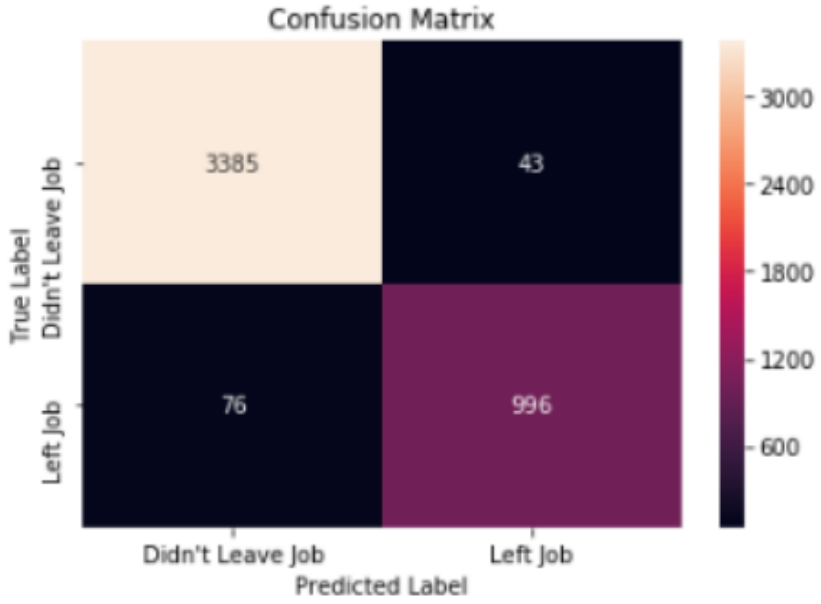

Fig 3:- Random Forest Confusion Matrix

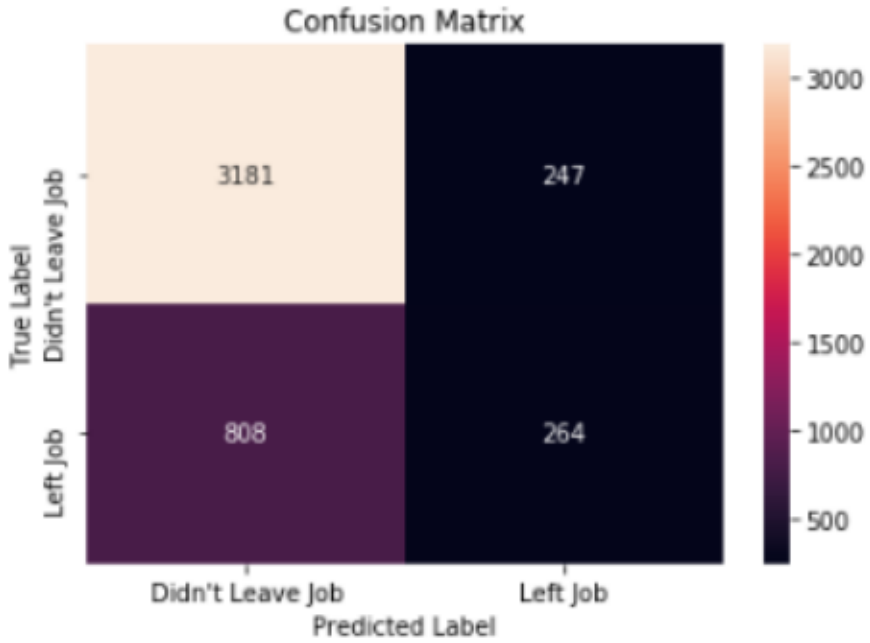

Fig 4:- Naïve Bayesian Confusion Matrix

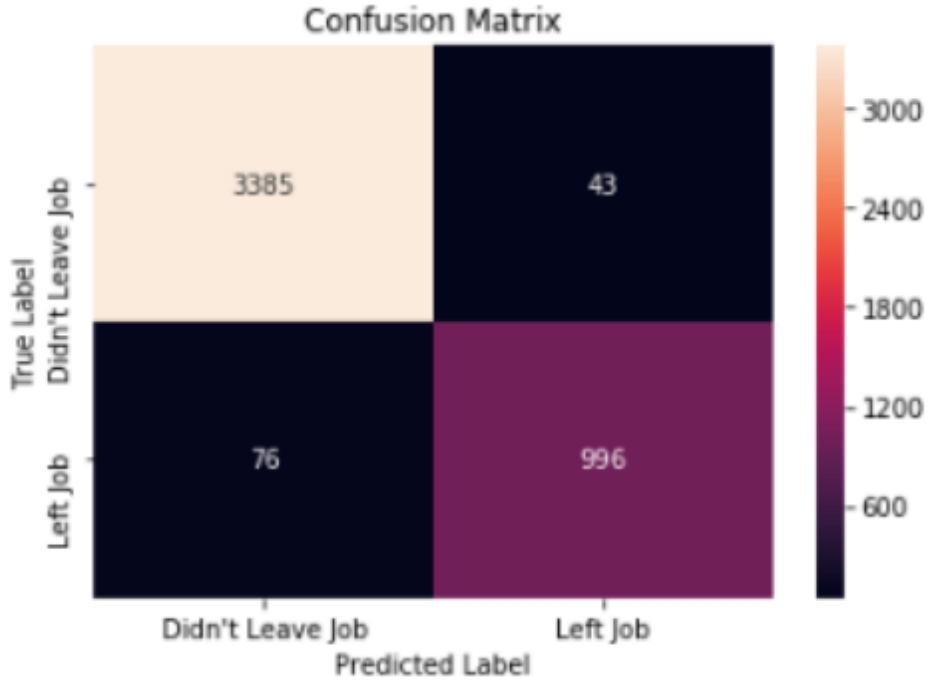

Fig 5:- LGBM Confusion Matrix

It is seen form the table 1 that Random Forest and LGBM perform better that the other two significantly. Ideally, Random Forests helps to achieve a better generalization but might cause over-fitting. On the other-hand the LGBM uses boosting which helps in training on noisy dataset and classifies points accurately in such a dataset.

LGBM overcomes the problem of overfitting using its inherent regularization.

VI. CONCLUSION

The need and importance of predicting if an employee will leave the accompany or not based on his/her condition has been discussed in this paper. The key challenge faced here was that the noise present in such datasets and the inaccuracy of dependency of the data between each other was highlighted and using LGBM how training can be done with the noisy data and still be accurate was showcased. LGBM algorithm was compared with other algorithms to show how much better it is in terms Terminated of accuracy.

For future studies, this can be implemented with automatic employee allocation. Using the same algorithms and different dataset, the specific features of a employee can be used to train a model and predict his accuracy at work and at what specialization of work. Hence automatic work allocation or task allocation.

ACKNOWLEDGMENT

This project is done by the author Rajat Keshri of 8th semester, Electronics and communication department, RVCE. The author thanks the department for providing a platform to present and do such a experiential project and thank the faculty and guide for supporting and clarifying all our doubts.

REFERENCES

[1]. Rohit Punnoose, Pankaj Ajit, "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms", International Journal of Advanced Research in Artificial Intelligence, Vol. 5, No. 9, 2016
[2]. M. Stoval and N. Bontis, "Voluntary turnover: Knowledge management – Friend or foe?", Journal of Intellectual Capital, 3(3), 303-322, 2002.
[3]. S. P. Singh, P. Singh, and A. Mishra, \Predicting potential applicants for any private college using LightGBM," in 2020 International Conference on Innovative Trends in Information Technology (ICITIIT), IEEE, 2020.
[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, "Turnover 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 turnover intentions", Psychology and aging, 28(1), 17, 2013.
[7]. S. L. Peterson, "Toward a theoretical model of employee turnover: 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 – Turnover relationship highlighting the moderating roles of visibility and reward contingency", Journal of Applied Psychology, 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 turnover", 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, "Turnover 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 turnover 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 turnover prediction models", International Journal of Management, 24(4), 808, 2007.
[16]. L. K. Marjorie, "Predictive Models of Employee Voluntary Turnover 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.