Early Prediction of Employee Attrition using Data Mining Techniques

Sandeep Yadav[1] Aman Jain[2] Deepti Singh[3]

sandeepyadav10011995@gmail.com *jain.aman881@gmail.com* *deepti.singh@jiit.ac.in*

[1-3] Department of Computer Science,
Jaypee Institute of Information Technology, Noida, India

***Abstract*— Bill Gates was once quoted as saying, "You take away our top 20 employees and we [Microsoft] become a mediocre company". This statement by Bill Gates took our attention to one of the major problems of employee attrition at workplaces. Employee attrition (turnover) causes a significant cost to any organization which may later on effect its overall efficiency. As per CompData Surveys, over the past five years, total turnover has increased from 15.1 percent to 18.5 percent. For any organization, finding a well trained and experienced employee is a complex task, but it's even more complex to replace such employees. This not only increases the significant Human Resource (HR) cost, but also impacts the market value of an organization. Despite these facts and ground reality, there is little attention to the literature, which has been seeded to many misconceptions between HR and Employees. Therefore, the aim of this paper is to provide a framework for predicting the employee churn by analyzing the employee's precise behaviors and attributes using classification techniques.**

Keywords*— *employee attrition, predictive analytics, data mining, churn prediction, machine learning

I. INTRODUCTION

Hiring new employees always costs organization some huge costs. Some revenues are tangible such as training expenses, and the time it takes, when an employee being a fresher to when they become a productive member. Human Resource departments generate an enormous amount of data on a daily basis: leaves, social conflicts, annual evaluations, wages and benefits, recruitments, departures, career evaluations, etc. but the big dilemma is to find out the correct and accurate replacement of the employees who have left. Here are some of the challenges faced by the hiring managers:

A. Eligible Candidates

Finding and sorting the best candidates on the basis of their resumes -Another area of research is to find out the candidates that would prove to be an asset to the firm if hired.

B. Demand and Supply Ratio

If a selected candidate drops off, then firm have to repeat the cycle of complete processes and find a replacement again.

C. Tenuous Relationship

Sometimes it happens that the exact job recruitments are not clearly communicated to the hiring managers. According to a survey conducted by the ICIMS, nearly 80% of recruiters think that they have a good understanding of their job position while 61% of hiring managers believe that recruiters have moderate levels of understanding. This imbalance between both the parties is quite strenuous and creates a barrier for a smooth flow.

D. Employee Retention

Beside predicting churn of employee, it would be great to find out the factors that allows to apprehend precisely, the employee's motivations and that's what makes them stay longer in your company. Based on a few data items (mainly information about an employee) - easily available from most Human Resource departments - our models are able to extract powerful attrition analysis that will let you decide about the right actions and may immediately improve employee retention rates.

- Take preventive measures and detect potential leavers.
- Business unit or department attrition analysis.
- Predicting turnover evolution in the short, mid and long run.

However, the most important costs are intangible. In our study, we have used the well-known classification techniques, namely, Logistic Regression, Support Vector Machine (SVM), Random Forest, Decision Tree, AdaBoost and Neural Network on the Human Resource Attrition dataset from Kaggle website (www.kaggle.com) [11]. The dataset has 14,999 records and 12 attributes (features) strictly associated to one's professional life including both categorical and numeric attributes.

The rest of the paper is organized as follows:
Section II, describes the related works that exists in the literature. Section III, defines the problem statement of the research done. In Section IV, research methodology has been discussed along with the experiments. Section V, describes the data model and comparison among them. Finally, Section VI, concludes all the findings of the work done in this paper.

II. RELATED WORK

There have frequent studies on churn prediction analysis in the literature. But recently the mode of huge attraction has mainly been the customer churn prediction for the researchers.

978-1-5386-6678-4/18/$31.00 ©2018 IEEE

For instance, Ibrahim et al. proposed to solve a big problem of customer churn related to a business, especially telecommunications by building models with different techniques such as Classification for prediction, Clustering for detection and Association for detection [3].

K. Dejaeger et.al, proposed a profit centric performance by calculating the maximum profit using optimal fraction with the highest predicted probabilities of customers to attrition in a retention campaign [4]. L. Carlos et al. proposed a model for a chronic problem for Brazilian universities to predict college attrition. H2O software was used as a data mining tool and Deep Learning for predicting the dropout cases and tried to identify the attrition profiles of students and for initiating their studies by implementing corrective measures [2].

Sepideh et.al, stated that even if we consider an optimum low churn rate of 5%, when an employee leaves the firm the cost involved (assuming the firm to be IT based) is approximately 1.5 times the annual income of an employee. Assume that a firm/organization has a strength (in terms of number of employees) is 160,000 (an organization under study) and an average salary of employees be $12,000.00 p.a. Then the loss the organization has to cope-up with is $144000000 (7000*1.5*12000). The detailed information on how customer attrition is a notorious and chronic problem for most industries, as loss of a customer affects income and brand image and acquiring new customers is tedious. Customer attrition (churn) closely related, but not identical to employee attrition. In this study major importance is given to customer churn as the firm is more towards customer centricity. They have used SVM for predicting the churn rate by building an accurate and reliable predictive model [1].

In another study, Choudhary et.al, presents the application of logistics regression technique based on the demographic data of separated employees as well as existing employees to develop a risk equation to predict employee attrition. Later this equation was applied to estimate attrition risk with the current set of employees. After the estimation, high risk cluster was identified to find out the reasons and hence action plan was appointed to minimize the risk [5].

IBM Watson team M. Singh et al. has done a brilliant analysis of employee's attrition process and proposed a framework which finds out the reasons behind attrition and identifying potential attrition. They have tried to calculate the cost of attrition and suggesting the employee's name for retention process, comparing the difference between Expected Cost of Attrition Before the retention period (EACB) and Expected Cost of Attrition After the retention period (EACA) [6].

III. Problem Definition

Employee attrition is a trivial issue for organization's loss. It leads to some crucial issues such as financial loss, cost and time to get the replacement and hiring, retraining of new employee and also customer dissatisfaction. Somehow organization can bear the loss of attrition of employees that are not as much experienced as those who has spent a significant amount of time that their attrition always results in some serious losses [7-10]. Therefore, the key is to retain its experienced and trained workforce. Employee attrition can have a negative impression on existing employees. Employee churn can be classified into following categories

- Best and experienced employees leaving prematurely.
- Fresher candidates churn.
- Department-wise churn.

IV. Research Methodology

A. Input Data Set

The data include 12 features for each record of the employee:
1) Name, 2) Satisfaction Level, 3) Last Evaluation, 4) Number of Projects, 5) Average Monthly Hours, 6) Time spent in Company (Years), 7) Departments, 8) Work accident, 9) Left, 10) Promotion last 5 years, 11) salary, 12) salary level. Table I shows the dataset.

Table I. Human Resource Datasets Attributes

S.No.	Attributes / Features	Data Type
1.	Name	Categorical
2.	Satisfaction Level	Numeric
3.	Last Evaluation	Numeric
4.	Number of Projects	Numeric
5.	Average Monthly Hours	Numeric
6.	Time spent in Company	Numeric
7.	Department	Categorical
8.	Work accident	Numeric
9.	Left	Numeric
10.	Promotion last 5 years	Numeric
11.	Salary	Categorical
12.	Salary level	Numeric

B. Data Pre-processing

The data cleaning/pre-processing is the very first step or approach in building a predictive model and trying to get the features that accounts for the classification of the classes' labels

to be predicted. There are total 12 attributes associated to professional domain of each employee and the attributes like name plays no role in prediction. Also, the two attributes salary and salary level correspond to more or less the same thing as low in salary denotes to 1 in salary level, medium in salary denotes to 2 in salary level and high in salary denotes to 3 in salary level. Therefore, the unnecessary attributes have been removed and, in the number, reduces to 10 features/attributes in all. Null values have been removed too in our dataset because sometimes more features are required which are not directly visible but can be derived or extracted from inside the data which gives you further intuitions.

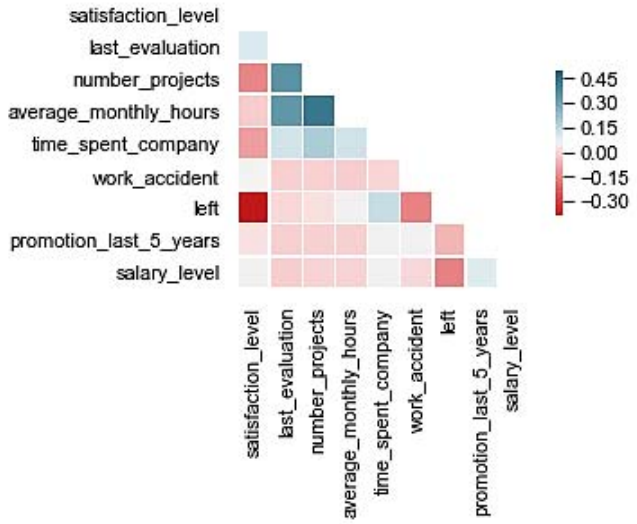

Fig 1: Correlation matrix

C. Feature Engineering

Since the updated data set contains 10 attributes and out of which two attributes are categorical i.e. Department and salary. The department consists of 10 categories. Salary consists of 3 categories including low, medium and high. Since the training the models with 80% of data and testing on the left, therefore there are chances that the test dataset might contain department type which was not in the training dataset. To elucidate this problem the data was randomly shuffled. Three approaches have been implemented in this study:

1. Brute-force approach: In this approach the classification of 10 categories of departments into two broad categories, i.e. Non-Technical as 0 and Technical as 1. The categories have merged, i.e. (sales, HR, marketing, management and accounting) as Non-Technical. And the categories (technical, RandD, IT, product_mng, support) as Technical as shown in the Table II.

TABLE II. FEATURE ENGINEERING

S.No.	Feature (Department)	Data Type
1.	Non-Technical	Numeric (0 or 1)
2.	Technical	Numeric (0 or 1)

2. One Hot Encoding: One hot encoding is the process of converting categorical variables into a form that could be provided to ML algorithms to do a better job (beneficial) in predicting. This has been done to give equal importance to all department and salary because if the categorisation of the 10 different departments was done as 0,1.2,3,4,5,6,7,8,9 then, since scikit- learn accepts only numerical values and the model would have given extra weightage to the feature which is having higher value i.e. the feature assigned as 1 is assumed to be less important than the feature which is assigned the value 5 , just because of numerical aspects - which will lead the classifier to poor results. Same case happens with salary too. Therefore, with this approach the number of features increases to a total of 21.

TABLE III. ONE HOT ENCODING FOR DEPARTMENT

S.No.	Feature (Department)	Data Type
1.	Department IT	Numeric (0 or 1)
2.	Department RandD	Numeric (0 or 1)
3.	Department Accounting	Numeric (0 or 1)
4.	Department Management	Numeric (0 or 1)
5.	Department Marketing	Numeric (0 or 1)
6.	Department Product Management	Numeric (0 or 1)
7.	Department Sales	Numeric (0 or 1)
8.	Department HR	Numeric (0 or 1)
9.	Department Support	Numeric (0 or 1)
10.	Department Technical	Numeric (0 or 1)

TABLE IV. ONE HOT ENCODING FOR SALARY

S.No.	Feature (Salary)	Data Type
1.	Salary Low	Numeric (0 or 1)
2.	Salary Medium	Numeric (0 or 1)
3.	Salary High	Numeric (0 or1)

In Table III department has been encoded to 10 separate department to give equal weightage to each. Similarly, in Table IV salary too is encoded in 3 categories respectively.

3. Feature Selection: This technique is often used to build various models with different subsets of training features to build an accurate and reliable model. Feature selection method is used to determine the features that would actually play a significant role in predicting employee's churn. The RFECV (Recursive Feature Elimination with Cross Validation) method is implemented. Features are selected recursively considering smaller and smaller sets of features and finds out the optimal number of features. After applying the RFEVC method and removing redundant features, according the model and then train them with new features and test them accordingly. As the main motive behind this approach is finding out the minimum number of features required that would be enough for an HR of a company to predict the attrition rate and plan accordingly for retention of an employee.

4. Experienced Employee Data: In this approach the motive is to find out the reason as to why best and experienced employees are leaving prematurely. For this the date of experienced employee has been extracted under some conditions as follows and the dataset is given in the Table IV.

- time_spent_in_company >= 4
- number_of_projects > 5
- last_evaluation >= 0.74

TABLE IV. EXPERIENCED EMPLOYEE DATASET

Dataset	Left/Churn	Not Left	Total	Churn Rate
Experienced Employee	1433	511	1944	0.737

Through this approach the novelty of the models is tested. This measure is taken in order to make sure that models work fine for both experienced person as well as freshers.

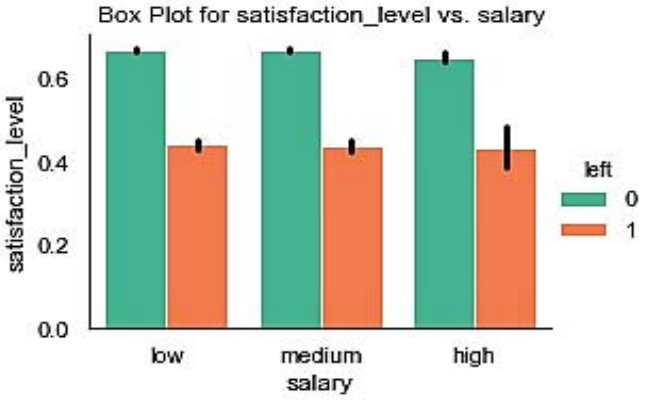

Fig. 2. Satisfaction level vs salary on full data

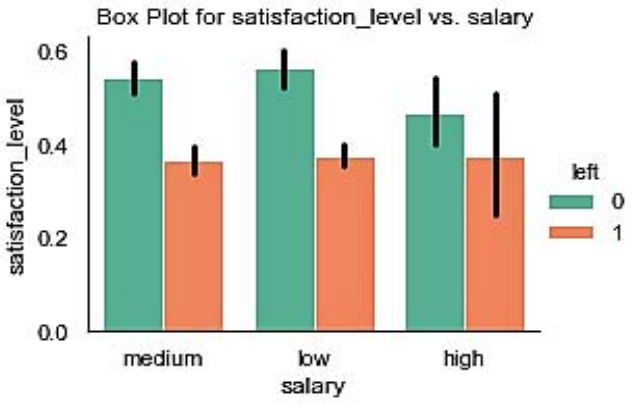

Fig. 3. Satisfaction level vs salary on experienced people data

From Fig 2 and 3 it can be figured out that the average satisfaction level of the employees is lower than who haven't left in both the cases.

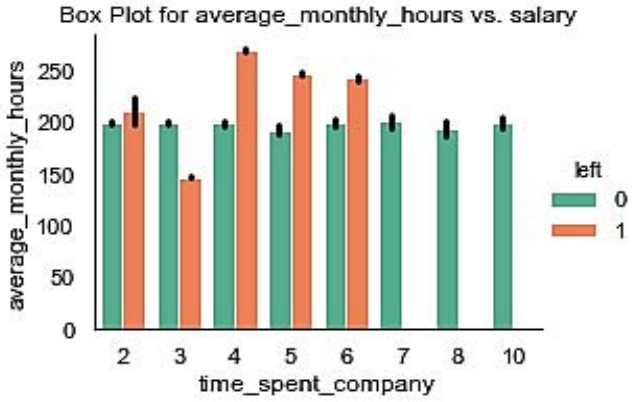

Fig. 4. Average monthly hours vs salary on full data

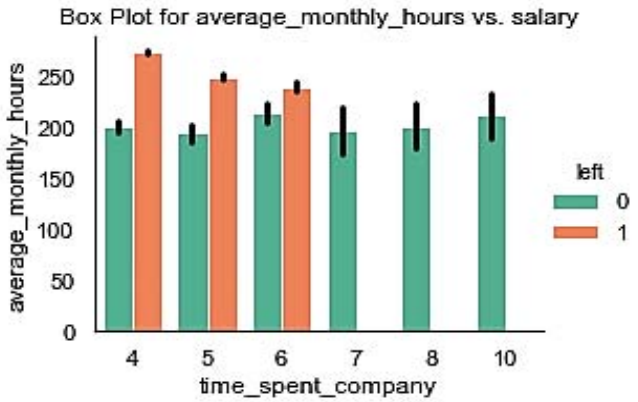

Fig. 5. Average monthly hours vs salary on experienced employee data

From Fig 4 and 5 it can be figured out that the employee who have left has more hours at work in both the cases.

V. Data Modeling And Comparison

Since this problem is mainly concerned with the HR department and they want to acquire the complete knowledge to overcome this problem. Therefore, recall parameter plays a major role in our study along with accuracy. In this section the aim is to find out the appropriate model to overcome this problem. In each model, recall, precision, accuracy and F1-score were calculated and compared on the employees in test data set and experienced employee data set. To avoid overfitting, dataset has been splitted into train-test dataset in 80:20 ratio. Table V demonstrates the detailed information about the splitting of data and experienced employees.

Table v. Training-Testing Datasets

Datasets	Left/Churn	Not Left	Total	Churn Rate
Training	2852	9147	11999	0.237
Testing	719	2281	3000	0.239
Experienced Employee	1433	511	1944	0.737

In this study, five models are analyzed, including Logistic Regression, SVM, Random Forest, Decision Tree and AdaBoost as classifiers to predict the churn using different approaches. Each cell in tables, stands for classification accuracy, precision, recall, F1 score and Left (number of employees going to leave), along with the actual number of employees left, i.e. 719, on the test dataset for the corresponding classification model mentioned under Models cell. After implementing the first approach of broadly categorizing the departments into Technical and Non-Technical. Table VI shows the results of different models for the first approach.

Table vi. Classification results of Ist Approach

Models	Acc.	Precision	Recall	F1 Score	Left (719)
Logistic Regression	0.7886	0.5793	0.3155	0.4085	378
SVM	0.9527	0.8760	0.9265	0.9006	734
Random Forest	0.9863	0.9939	0.9467	0.9697	661
Decision Tree	0.9767	0.9286	0.9741	0.9508	728
AdaBoost	0.9583	0.8979	0.9251	0.9113	715

Table VII shows the results of different models for the one hot encoding approach. The results are improved with this approach as compared to Table VI results. The results of AdaBoost and Random Forest are improving significantly in the desired and expected way. The models in Table VI and VII are trained with no feature being updated.

Table vii. Classification results of One Hot Encoding Approach

Models	Acc.	Precision	Recall	F1 Score	Left (719)
Logistic Regression	0.7886	0.5793	0.3155	0.4085	378
SVM	0.9503	0.8779	0.9207	0.8988	754
Random Forest	0.9786	0.9851	0.9249	0.9541	675
Decision Tree	0.9817	0.9498	0.9749	0.9623	738
AdaBoost	0.9583	0.9125	0.9137	0.9131	720

Table VIII shows the results of Feature Selection approach applied on Random Forest and AdaBoost as the sole motive is to predict the result with only the less and important information (features), to make it simple for HR.

Table viii. Classification results on Feature Selection Approach

Models	Acc.	Precision	Recall	F1 Score	Left (719)
Random Forest*	0.9863	0.9925	0.9481	0.9698	663
AdaBoost*	0.9583	0.9047	0.9164	0.9105	703

* Trained on updated features

Table IX shows the results tested on experienced employee data to verify the performances of the models in order to check the novelty of models to authenticate that the models works fine with heterogeneous data consisting of both experienced employees and freshers.

Table ix. Classification results of experienced employee data

Models	Acc.	Precision	Recall	F1 Score	Left (1433)
Logistic Regression	0.6291	0.8819	0.5736	0.6951	932
SVM	0.9362	0.9331	0.9839	0.9578	1511

Random Forest	0.9861	0.9950	0.9860	0.9905	1420
Decision Tree	0.9927	0.9930	0.9972	0.9951	1439
AdaBoost	0.9393	0.9581	0.9595	0.9588	1435
Random Forest*	0.9861	0.9950	0.9860	0.9905	1420
AdaBoost*	0.9454	0.9604	0.9658	0.9631	1441

* Trained with updated features

VI. CONCLUSION

Employee attrition can affect an organization in many ways like goodwill, revenues and cost in terms of both time and money. The predictive attrition model helps in not only taking preventive measure, but also making better hiring decisions. In this study implementation of various classification method helps in predicting whether a particular employee might leave the organization in the near future by deriving trends in the employee's past data. It was intuited that salary or other financial aspect like promotions are not the sole reasons behind the attrition of employees. These models can help us in prioritizing the features with higher impact in attrition of an employee and the possible reasons behind it so that HR can take appropriate decision for the retention process. The main purpose of this research is to build reliable and accurate models which can optimize the hiring and retention cost of quality employees. This could be done by determining the attrition status of employee under consideration by using the appropriate data mining techniques.

VII. REFERENCES

[1] S. H. Dolatabadi and F. Keynia, "Designing of customer and employee churn prediction model based on data mining method and neural predictor,"2nd International Conference on Computer and Communication Systems (ICCCS), pp. 74-77, Krakow, 2017.

[2] L. C. B. Martins, R. N. Carvalho, R. S. Carvalho, M. C. Victorino and M. Holanda, "Early Prediction of College Attrition Using Data Mining," 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1075-1078, Cancun, 2017.

[3] I. M. M. Mitkees, S. M. Badr and A. I. B. El Seddawy, "Customer churn prediction model using data mining techniques," 13th International Computer Engineering Conference (ICENCO), pp. 262-268, Cairo, 2017.

[4] K. Dejaeger, W. Verbeke, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," European Journal of Operational Research, vol. 218, no. 1, pp. 211–229, 2012.

[5] C. K. Choudhary, R. Khare, D. Kaloya, and G. Gupta, "Employee attrition risk assessment using logistic regression analysis" 2nd IIMA International Conference on Advanced Data Analysis, Business Analytics and Intelligence ,Ahmedabad, 2011.

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

[7] Srivastava, Devesh Kumar, and Priyanka Nair. "Employee Attrition Analysis Using Predictive Techniques." International Conference on Information and Communication Technology for Intelligent Systems. Springer, Cham, 2017.

[8] Umayaparvathi, V., and K. Iyakutti. "Applications of data mining techniques in telecom churn prediction." International Journal of Computer Applications 42.20 (2012): 5-9.

[9] Xie, Yaya, et al. "Customer churn prediction using improved balanced random forests." Expert Systems with Applications 36.3 (2009): 5445-5449.

[10] Mitkees, Ibrahim MM, Sherif M. Badr, and Ahmed Ibrahim Bahgat ElSeddawy. "Customer churn prediction model using data mining techniques." Computer Engineering Conference (ICENCO), 2017 13th International. IEEE, 2017.

[11] Berry, Michael JA, and Gordon S. Linoff. Data mining techniques. John Wiley & Sons, 2009.