Research Article

Explaining and predicting employees' attrition: a machine learning approach

Praphula Kumar Jain[1] · Madhur Jain[2] · Rajendra Pamula[1]

Received: 19 August 2019 / Accepted: 13 March 2020 / Published online: 25 March 2020

Abstract

The attrition of employees is the problem faced by many organizations, where valuable and experienced employees leave the organization on a daily basis. Many businesses around the globe are looking to get rid of this serious issue. The main objective of this research work is to develop a model that can help to predict whether an employee will leave the company or not. The essential idea is to measure the effectiveness of employee appraisal and satisfaction rates within the company, which can help to reduce the attrition rate of employees. In this paper, a new approach focused on machine learning was used to enhance different retention approaches for targeted employees. There is also an effort in this paper to shed some light on different factors influencing the attrition rate of workers and their possible solutions. Implementing this principle will help management in employee appraisal and in the decision-making process to recognize valuable employees who will leave the company. Using this application, hidden reasons for employees' attrition can be identified, and management can take preventive actions regarding attrition of each employee individually.

Keywords Human resource management · Machine learning · Prediction · Classification · Employees attrition

1 Introduction

Data mining is all about processing data and finding the patterns and trends to generate information that can be helpful in deciding or choosing the long-run trend [1]. Data mining is the most current active research area, and techniques of it have been used in classification, clustering, and prediction, too [2]. Various machine learning techniques are applicable in numerous industries, education, and research fields. Currently, companies in a wide range of industries are already using machine learning techniques on a regular basis. Few of them are retail, health care, banking, software, insurance, etc. [3]. Those businesses using machine learning also combine it with data processing with pattern recognition, machine learning, computer science, statistics, and alternative necessary tools [4]. Data Mining is rising over its importance in human resource management in businesses since it permits them to obtain a clear idea about their employees and customers for making smart beneficial decisions for them.

For each employee, an organization has to invest an abundance of time and money for their training as per the organization's requirement. When an employee leaves the organization, the company is not only losing its valuable employees, but the company also loses on the amount that it has spent to recruit and select those employees and to train them for their respective jobs. On the other hand, the organization needs to invest more and more in recruitment, training, and development of new staff to fill up their vacant positions. Due to these reasons, every organization wants to control the attrition rate and retain

✉ Praphula Kumar Jain, praphulajn1@gmail.com; Madhur Jain, mdhr.jain@gmail.com; Rajendra Pamula, rajendrapamula@gmail.com | [1]Department of Computer Science and Engineering, Indian Institute of Technology (ISM), Dhanbad, Jharkhand 826004, India. [2]Software Systems, Department of Computer Science and Engineering, Birla Institute of Technology and Science, Pilani, Rajsthan, India.

SN Applied Sciences (2020) 2:757 | https://doi.org/10.1007/s42452-020-2519-4

its employees through more satisfactory company policies and work environments. The present research work would be helpful for most companies to know about their employee's satisfaction levels and get some useful information that would aid in controlling the attrition rate.

In this paper, machine learning techniques are used in the problem of employees' attrition due to two reasons. Firstly, in recent years machine learning techniques have not been used to solve employee attrition problems. Secondly, machine learning techniques outperform an employee's attrition problem. From data collection to finding the reason for employee attrition, the present research work has been completed with a predictive model in the manner mentioned below.

Here, predictive analysis techniques have been used on Human Resource data. In the very first step, data were collected from the free online available source, and then various data exploration techniques were applied to clean, prepare and pre-process the data. In the next step, predictive models such as support vector machine (SVM), decision tree (DT), and random forest (RF) were applied to the processed data. Finally, it was concluded that a machine learning model would be a best-fitted model for the dataset in hand. Then these predictive models in training data set to train the model to validate this model by performing k-fold cross-validation and test the model on the test data set. Finally, a comparison of the results of various machine learning models was done to decide which model best-fitted and gives precise results for a given problem.

2 Related work

From the survey of articles related to human resource management (HRM), it has been decided to apply machine learning techniques for handling real-life issues in HRM. In the chapter written by [5], the authors explained how HRM would be useful in some real-life scenarios. In [6], the author has seen a relationship between HRM and productivity. In [7], the author defined the importance of HRM in the field of management. In [8], check out about the effectiveness of industry characteristics important on the performance working system. Their results indicate that the impact of those human resource systems on productivity is influenced by the business capital intensity, growth, and differentiation. Reviews progress by [9] identified a series of phases in the development of relevant theory and research. It then sets out a number of challenges for the future on issues of theory, management processes, and research methodology. The main conclusion from the review is that when over 20 years of in-depth analysis, there is no standard system to answer core questions on the connection between HRM and performance. In [10], presented a review of the organic process and written account perspective on the event of strategic HRM literature. The book is divided into seven stuff that provided directions to the researchers who have been working on it for a long duration of time. The author found many fields that are related to their current state and explained how those fields could be beneficial in future research work. Many fields suggested by [11], have been used in HRM research very much. In [12], the author tried to solve the many management related challenges and provided the solutions to that.

In the present scenario, machine learning techniques are very much useful in the field of prediction. Prediction of source code change will give good feedback by [13], and states whether the modified code will be passed or not. In this paper, a prediction is made by Ada Boot, and ZeroR techniques and comparison of both techniques are also listed here. In [14], the authors presented a paper on the crop suitability prediction based on the rough set and neural network. In this paper, various parameter responsible for the crop is taken into consideration, based on that, the prediction was made to identify suitable crop results were presented along with reasons for selecting a particular crop. In [15], suggested a predictive model for paddy crop productivity using machine learning techniques. The author suggested a plan for the cultivation that can be helpful for farmers. Finally, based on results produced by researchers in recent years for the prediction process in various social issues and real-life challenges, it has been decided to apply machine learning techniques for resolving employees' attrition problems.

The accuracy of the employee attrition prediction is dependent on the data and the method used. Therefore, the aim of the present study is to focus on these two parameters to maximize the accuracy of the predictive model. An employee's attrition problem is a binary classification problem that uses machine learning classification techniques such as SVM, logistic regression, naïve base, neural network, and DT. Due to the simplicity and interpretability of the model DT and logistic regression is used by the researcher and academicians [16]. Due to the predictive power and better accuracy, more advance model is also used. For example, the author [17] uses DT, K-nearest neighbor, and artificial neural network in the field of employees attrition, whereas result analysis represents an artificial neural network to perform better. Author [18] presented an employee attrition model based on a SVM for the e-commerce industry. His results analysis presents SVM outperforms neural network and logistic regression. Some of the authors present problems related to employee attrition, such as [19] show a comparative study on the class imbalance problem.

Table 1 Attributes description

Satisfaction level	Employee satisfaction level in the company, where 0 represents the least satisfied and 1 represents most satisfied
Last evaluation	Employee last evaluation (rating) in the company, where 0 represents the least rating and 1 represents most rating
Number of projects	Total number of the project done by an employee in his/her carrier
Average monthly hours	Mean of hours spent by the employee in the company, on monthly basis
Time spend company	Mean of hours spent by the employee in the company, on daily basis
Work accident	Numeric attribute values (0 or 1). if any accident/escalation happened with the employee in the company per month
Left	Target attribute value (0 or 1). Where 0 represents employee not left the company and 1 represents employee left the company
Promotion on last 5 years	Numeric attribute values (0 or 1). Where 0 represents employee don't get any promotion in last 5 years whereas 1 represents employees who received the promotions
Sales	The information about employees department. This is a categorical variable which has seven departments
Salary	Categorical variable divding the salary of employees in 3 broad categories (low, medium and high)

Table 2 Department wise employees distribution

Department name	Employees count
Sales	4140
Technical	2720
Support	2229
IT	1227
Product_mng	902
Marketing	858
RandD	787
Accounting	767
HR	739
Management	630

Table 3 Variable identification

Attributes	Data type	Variable category	Type of variable
Satisfaction_level	Numeric	Continuous	Predictor
Last_evaluation	Numeric	Continuous	Predictor
Number_of_projects	Numeric	Categorical	Predictor
Average_montly_hours	Numeric	Continuous	Predictor
Time_spend_company	Numeric	Continuous	Predictor
Work_accident	Numeric	Categorical	Predictor
Promotion_last_5years	Numeric	Categorical	Predictor
Domain	Character	Categorical	Predictor
Salary	Character	Categorical	Predictor
Left	Numeric	Categorical	Target variable

3 Data collection and preprocessing

3.1 Dataset description

The human resource management data set, which is used in this research work is available online and is free of cost on kaggle.com. This dataset contains 10 features and more than 14,000 records. All 10 features are related to the employees' attrition problems. Selected attributes with their detailed description are mentioned in Table 1. Employees count in each and every department is shown in Table 2.

3.2 Data exploration

Data exploration is the leading procedure in the analytical action of data. Statistical and visualization techniques are used to describe the data. In order to further analyze the data and to bring out the important aspects, we need to first explore the data. In the data exploration phase, the techniques of variable identification, univariate analysis, and bivariate analysis have been performed step by step on the HRM dataset.

3.2.1 Variable identification

Variable identification is the first step in the data exploration process. This process has been completed in two steps. In the first step, identifying the predictor variables as input variables and target variables as output variables is done. In the next step, identifying the data type and category of the variables is performed, as shown in Table 3.

Table 4 Uni-variate analysis

	Satisfication level	Last evaluation	Number of projects	Average monthaly hours	Time spend in company	Work accident	Left	Promotion in last 5 years
Count	14999	14999	14999	14999	14999	14999	14999	14999
Mean	0.61283	0.716102	3.80305	201.05033	3.49823	0.14461	0.238083	0.021268
Std	0.24863	0.171169	1.23259	49.94309	1.46013	0.351719	0.425924	0.144281
Min	0.09	0.36	2	96	2	0	0	0
25%	0.44	0.56	3	156	3	0	0	0
50%	0.64	0.72	4	200	3	0	0	0
75%	0.82	0.87	5	245	4	0	0	0
Max	1	1	7	310	10	1	1	1

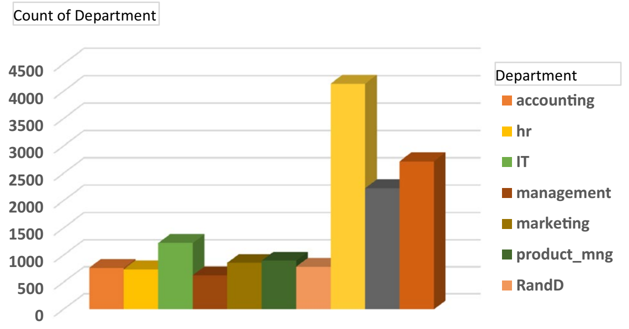


Fig. 1 Departmental strength

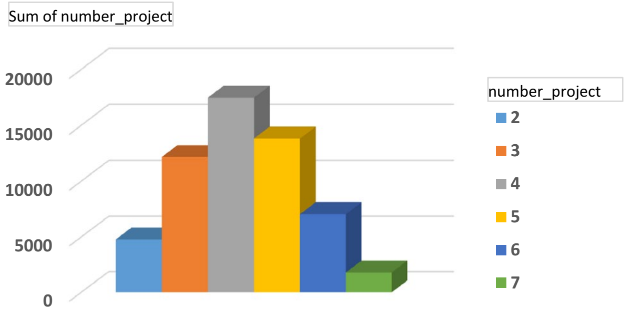


Fig. 2 Projects distribution

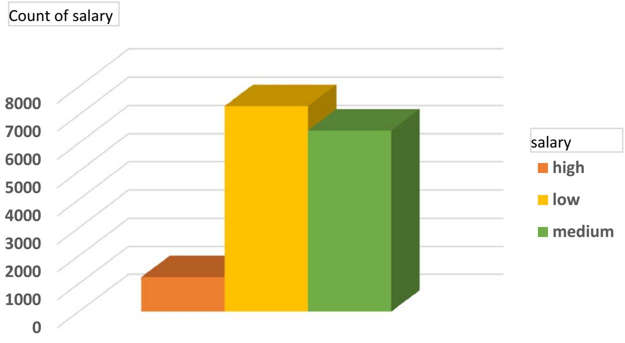


Fig. 3 Salary distribution

3.2.2 Uni-variate analysis

In the univariate analysis, continuous and categorical variables are explored. A technique to perform the univariate analysis is subjected to the variable type (categorical or continuous). We have explored these approaches and statistical measures for categorical and continuous variables individually.

Continuous variables Here we have focused on the Mean, standard deviation, and spread of the variable. These are explained using several statistical metrics visualization methods. We took the summary of continuous variables, as shown in Table 4.

Categorical variables A frequency distribution is the best way to comprehend the spreading of each categorical variable. It can be read as a percentage of values under each category. It can be measured using two metrics Count and Count% against each category. A bar chart can be used as a visualization tool.

Figures 1, 2, 3 and 4 represent the bar chart of the variables. Figure 1 represents the department strength of the organization; it means the number of projects handling by the different departments like Accounting, hr, IT, management, marketing, product, and RandD. The figure shows that most of the project handling by employees working in the HR department and a very less number of employees are working on management related projects. Project distribution is shown in Fig. 2, from 2 to 7 six types of projects are available. The highest number of projects are available from type 4, and the minimum number of projects is of type 7. Figure 3 represents the salary distribution of the employees; it shows that most of the employees are getting a low salary, and very few are working on a high salary. The number of hours spent by the employees has been shown in Fig. 4, it can be concluded that most of the

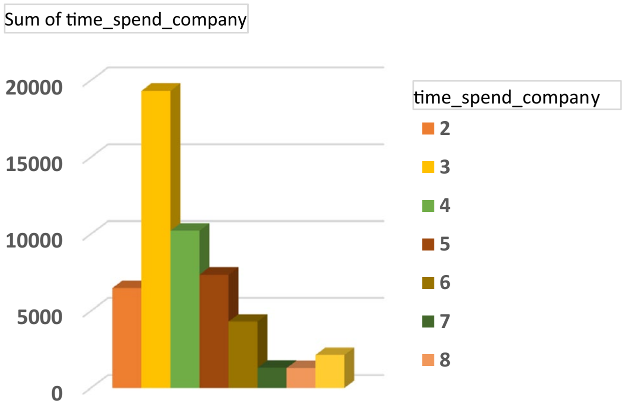


Fig. 4 No. of hours spent in the company

employees are spending 3 h, and there are few employees who are spending 10 h.

Feature importance plot Decision tree makes split that minimize the decrease in impurity. By calculating the mean decrease in impurity for each feature across all trees shown in Fig. 5 using the feature importance plot.

3.2.3 Bi-variate analysis

A Bi-variate analysis is used to find out the correlation between two variables. Here, we look at variables at a Pre-determined significance level. Implementation of bivariate analysis for any grouping of absolute/categorical and continuous variables can be done. These grouping can be categorized as Continuous & Continuous, Continuous & Categorical and Categorical & Categorical. Dissimilar methods are used to handle these groupings in the course of the analysis process. The feasible combinations in detail are stated below:

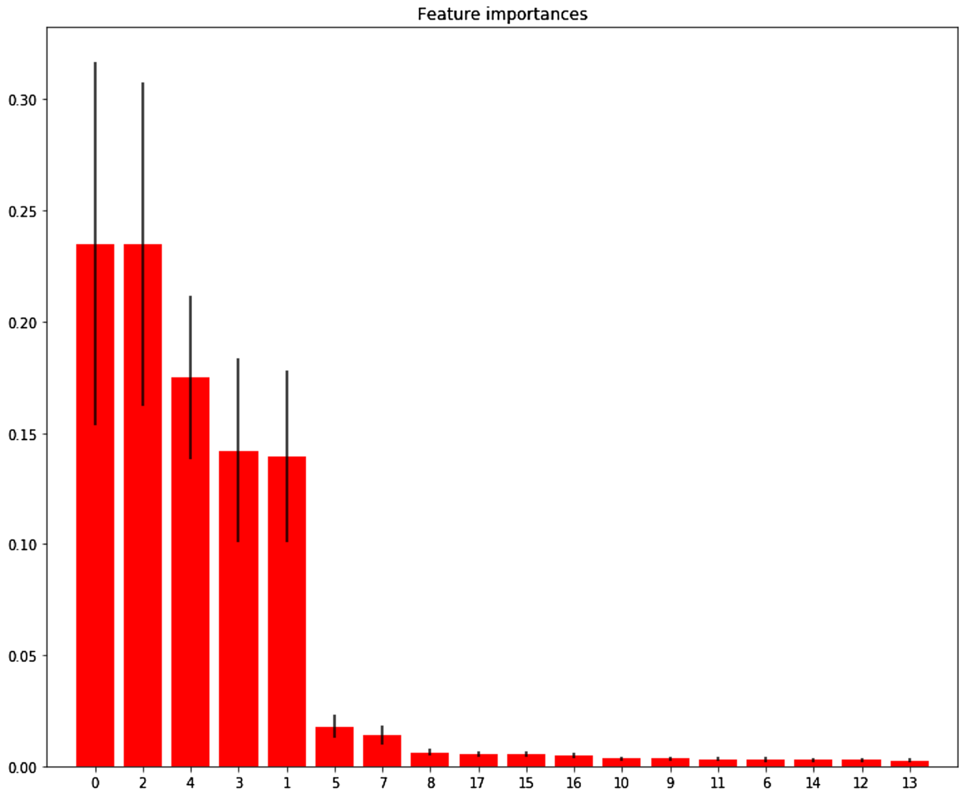
Continuous & continuous We will look at a scatter plot for two continuous variables comparisons. It is a good way to determine the relationship between two variables. The outline of the scatter plot signifies the correlation.

Categorical & categorical If we have to determine the relationship between two absolute, i.e., categorical variables, then we have to use the Chi-Square test.

Chi-square test Chi-Square Test is used to obtain the statistical consequence of the relationship between the variables. It also checks whether the suggestions in the specimen are robust enough to evaluate the relationship for a massive population as well. This test is based on the variance between experimental and expected frequencies in at least one category in the two-way table. The probability for the quantified chi-square distribution gets yield with the degree of freedom.

Probability of 0: It depicts that both categorical variables are relatable to each other.

Fig. 5 Feature importance plot



Probability of 1: It depicts that both variables are not inter-related, i.e., they are independent.

Probability less than 0.05: It depicts at 95% confidence the relationship between the variables is significant. For the test of independence of two categorical variables, the chi-square test can be given by the following equation:

$$X^2 = \frac{\sum (O - E)^2}{2} \tag{1}$$

where O = observed frequency and E = expected frequency under the null hypothesis and can be computed by Eq. 2.

$$X^2 = \frac{row\ total * column\ total}{sample\ size} \tag{2}$$

Categorical & continuous Box plot is the best way to explore the Categorical variable. For each level of categorical variables, we can draw the box plot. A scatter plot is the best way to map the relationship between two continuous variables. But the strength of the relationship among them is not indicated. The strength of the relationship can be determined by using the correlation shown in Fig. 6. The value of the correlation lies between − 1 to + 1. Correlation can be derived using Eq. 3 given below.

$$Correlation = \frac{covariance(X, Y)}{\sqrt{Var(X) * Var(Y)}} \tag{3}$$

3.3 Data visualizations

In this section, data visualization is performed on a continuous and categorical variable, and an attempt is made to understand the relationship among these variables with our target variables. Graphs are the best way to understand the behavior of attributes and relationships among them.

3.3.1 Satisfaction level versus employee left

Form Fig. 7, the satisfaction level versus employee left, we can find out the high chance of the employees who will be left the company. Employees who had really satisfaction levels 0.1 or less and between 0.3 and 0.5 having more chances to left the company.

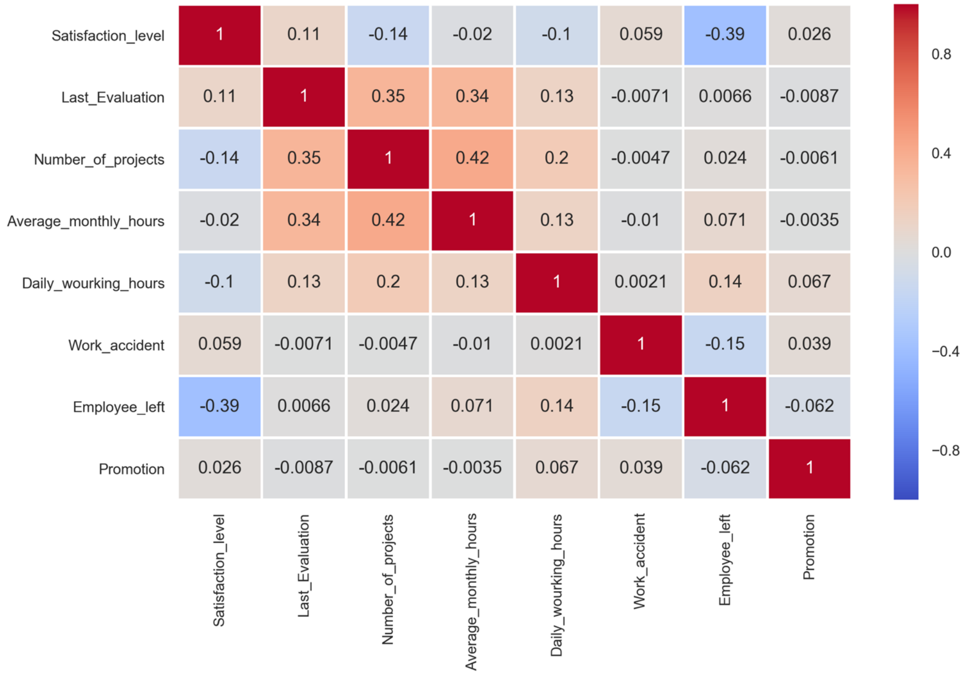
3.3.2 Average monthly hours versus number of projects

Average monthly hours versus number of projects is shown in Fig. 8. With an increase in the project count, average monthly hours increases proportionally.

3.3.3 Last evaluation versus number of projects

Last evaluation versus number of projects is shown in Fig. 9. It shows that the last evaluation directly depends upon the number of projects.

Fig. 6 Correlation between different features



	Satisfaction_level	Last_Evaluation	Number_of_projects	Average_monthly_hours	Daily_wourking_hours	Work_accident	Employee_left	Promotion
Satisfaction_level	1	0.11	-0.14	-0.02	-0.1	0.059	-0.39	0.026
Last_Evaluation	0.11	1	0.35	0.34	0.13	-0.0071	0.0066	-0.0087
Number_of_projects	-0.14	0.35	1	0.42	0.2	-0.0047	0.024	-0.0061
Average_monthly_hours	-0.02	0.34	0.42	1	0.13	-0.01	0.071	-0.0035
Daily_wourking_hours	-0.1	0.13	0.2	0.13	1	0.0021	0.14	0.067
Work_accident	0.059	-0.0071	-0.0047	-0.01	0.0021	1	-0.15	0.039
Employee_left	-0.39	0.0066	0.024	0.071	0.14	-0.15	1	-0.062
Promotion	0.026	-0.0087	-0.0061	-0.0035	0.067	0.039	-0.062	1

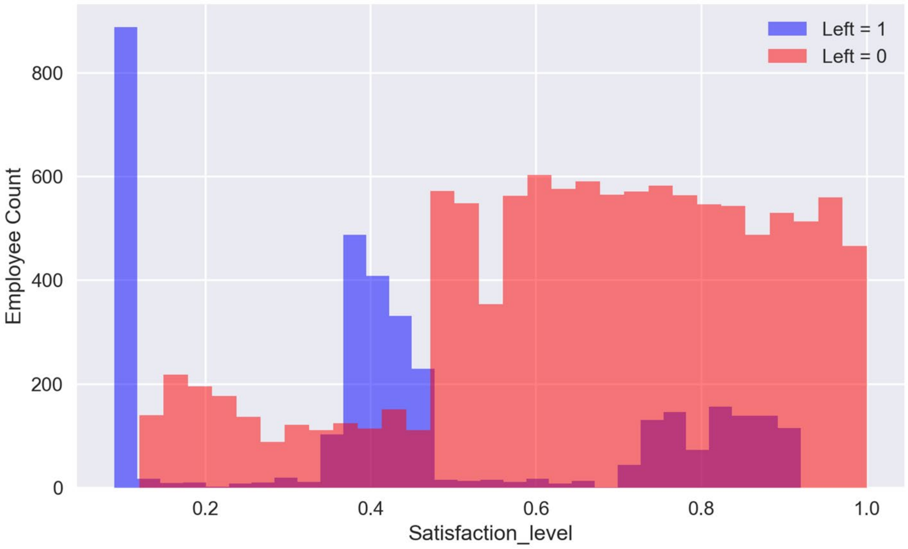


Fig. 7 Satisfaction level versus employee left

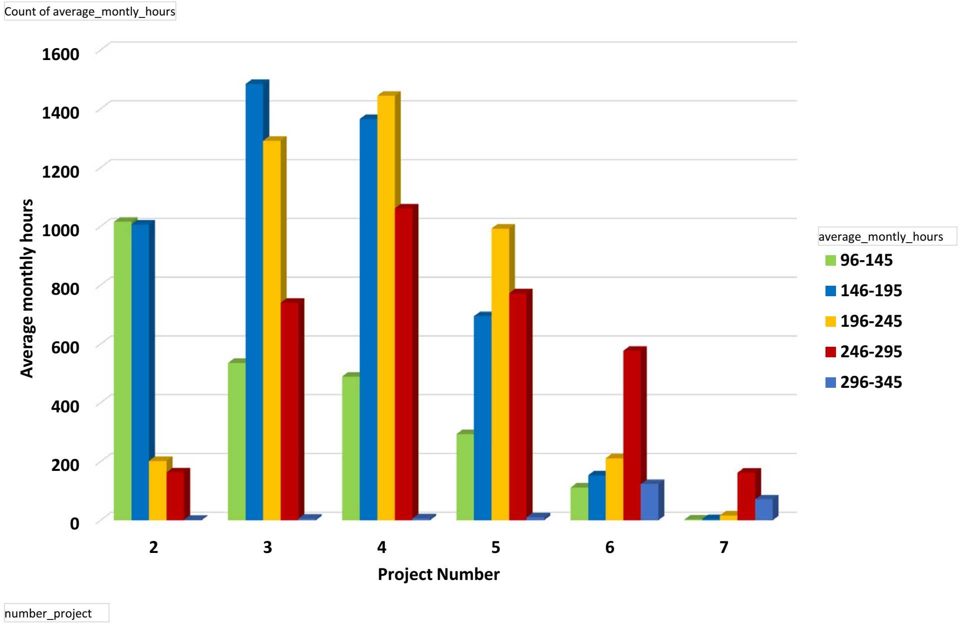


Fig. 8 Average monthly hours versus number of projects

3.3.4 Employee's count versus last evaluation

Figure 10 shows the employee's count versus last evaluation. This figure represents the bio-modal distribution of employees who left the company. Low performance and high performance are the two main criteria of the employees that highly tend to leave the company. The last evaluation lies within 0.6–0.8 is a favorable zone for employees that stays in the company.

3.3.5 Salary versus employees left the company

Figure 11 shows the distribution of salary versus employee's left the company. From this figure, we concluded that, majority of Employees who Left the company had either low or medium salary.

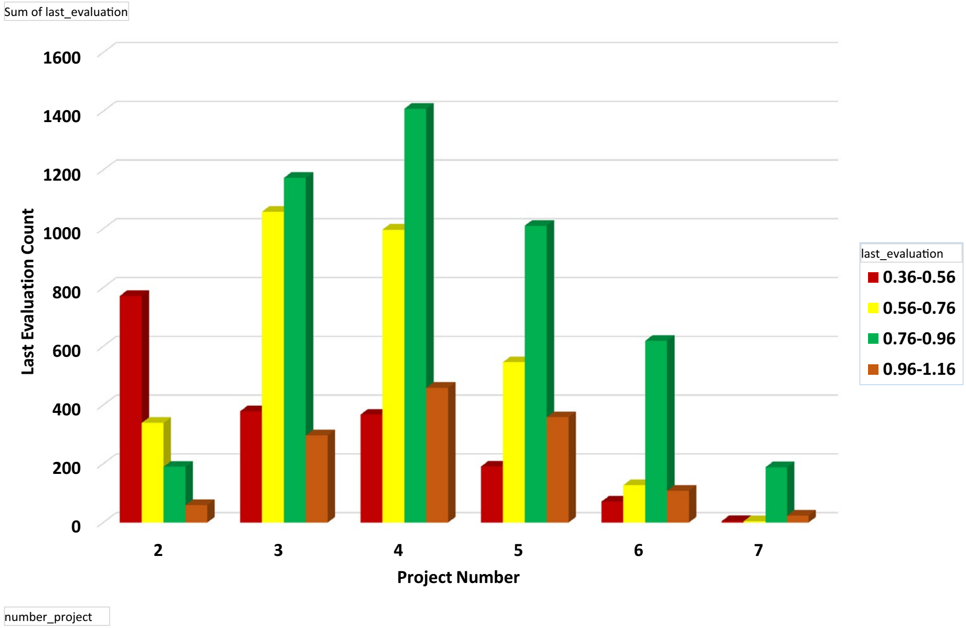


Fig. 9 Last evaluation versus number of projects

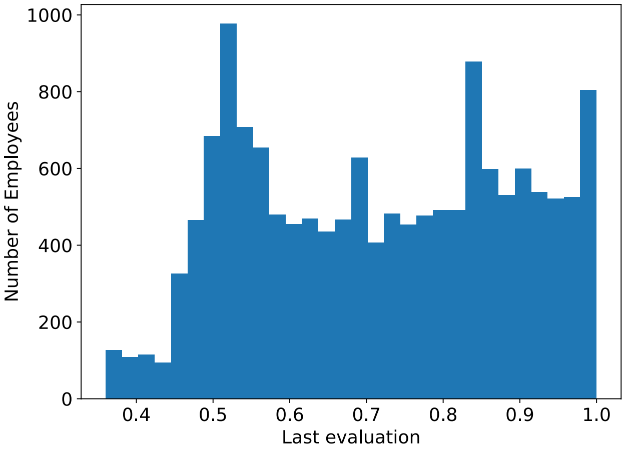


Fig. 10 Distribution of employees on the basis of their last evaluation

3.3.6 Department versus employees left the company

Figure 12 represents the distribution of department versus employee's left the company. This figure depicts that the technical, sales and support departments were among the top 3 departments to have maximum employee churning.

3.4 Data preparation and model detection

Absent data in the training data set can bring down the power to fit a model or can direct to a biased model since we have not analyzed the functioning and relationship with other variables accurately, which can lead to incorrect prediction.

3.4.1 Methods to treat missing values

With the help of Python's pandas library function dropna() and fillna() we can drop the missing values from our dataset. Finally, the dataset is free from the missing values, and now the dataset is ready for further processing.

4 Model formulation

4.1 Modelling

The modeling process involves the selection of models based on various machine learning techniques, which would be used in the experimentation. In prediction, various predictive models based on artificial neural networks, DT, Bayesian method, logistic regression, SVM, etc., can be employed. Our goal is to identify the best classifier for our problem. For this, each classifier can be trailed on the feature set and the classifier with the best classification results can be used for the prediction.

4.1.1 Support vector machine (SVM)

An algorithm that can be defined by separating hyperplane is said to be an SVM [20]. Like, the output of the given training data is an optimal hyperplane on applying the algorithm, which categorizes a new set of examples. This algorithm is mainly concerned about finding the hyperplane, which gives the distance of the training samples to the maximum extent. The margin is obtained with the distance given by the SVM theory. The margin

Fig. 11 Distribution of salary versus employees left the company

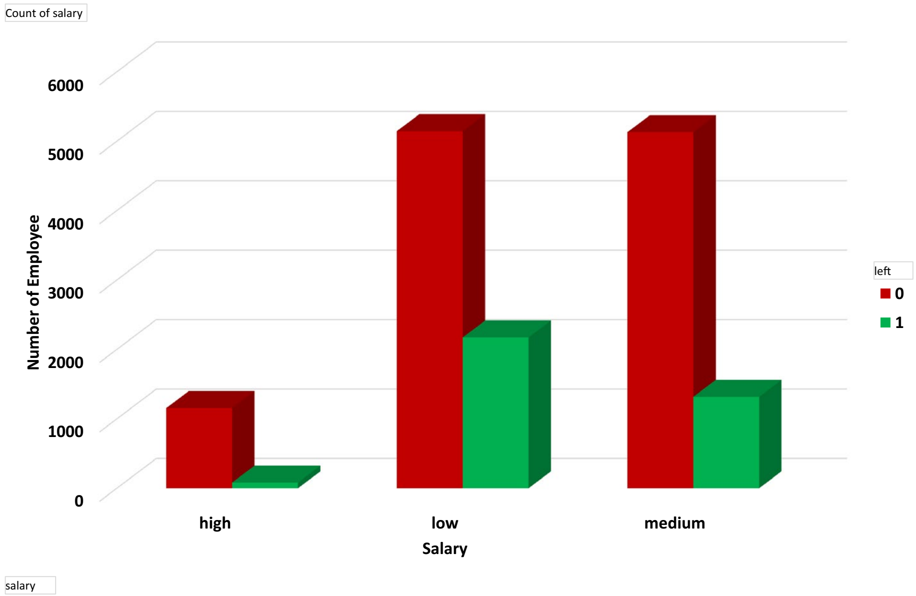
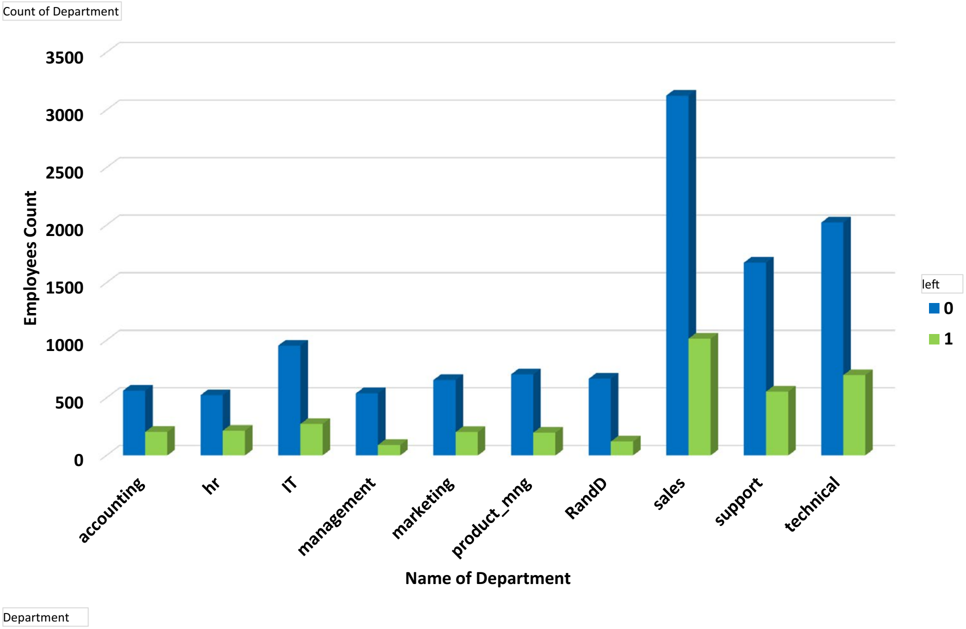


Fig. 12 Distribution of departments versus employees left the company



of the training data is maximized by the separating hyperplane determined by SVM.

4.1.2 Decision tree (DT)

A DT [21], breakdown the decisions which visually and clearly signify decisions and decision making. A DT describes data in which the resulting classification tree can act as an input for decision making.

4.1.3 Random forest (RF)

RF [22], is an ensemble learning approach for categorizing and backsliding the dataset. This approach works while outputting the mode of the classes (categorizing) or backsliding of the particular tree by developing a large number of DT.

4.2 Training and testing

The dataset is split into training and testing dataset using cross-validation, with training data being used to train the model and testing data used to test the model. The machine learning techniques used for the prediction model were Support Vector Machine with Radial Basis Function kernel (SVM), Decision Tree (DT), and Random Forest (RF). The classifiers were trained using Scikit-learn [23, 24]. A short description of each classifier is given below.

4.3 Model evaluation

Every instance of the given problem would be classified as if the employee leaves the company or not. Using a standard confusion matrix, the number of instances that the model has correctly classified can be identified. Confusion matrix allows visualization of the performance of a classifier giving out a detailed analysis with reports on the number of true positives, false positives, true negatives, and false negatives. The only accuracy can yield misleading results if the dataset is unbalanced and hence can be unreliable. A classification report would depict the precision, recall, and F1-Score for the model. Precision and recall are based on the measure of relevance with precision representing the fraction of relevant samples among the retrieved samples, while recall depicts the fraction of relevant samples that have been retrieved over the total number of relevant samples. Using the above-mentioned evaluation metrics, the classifiers are evaluated in order to find out the best model for the problem.

Table 5 Department wise result analysis

Department	Algorithm	Precision	Recall	F1-score
Sales	DT	95	98	97
	SVM	88	92	90
	RF	99	98	98
Technical	DT	89	98	93
	SVM	83	91	87
	RF	97	97	97
Support	DT	90	97	93
	SVM	89	95	92
	RF	98	97	98
IT	DT	97	98	97
	SVM	86	83	84
	RF	99	98	99

5 Result analysis

This section presents the results of the various classifier used for the prediction. We have been selected departments that are having more than 1000 employees, such as Sales, Support, Technical, and IT. The binary classifiers, SVM, DT, and RF, are evaluated using various binary and numeric features, as mentioned above, using their scikit-learn implementations. Comparisons between the classifiers SVM, DT, and RF are shown below in Table 5. From the standard confusion matrix for the classifiers and the classification report, it is more typical of classifiers to detect the majority class and be less sensitive to the minority class and the classification. Thus, it might be biased, and the result is simply predicting the majority class, which, in our case, is an employee not leaving the company. Evaluating the classifiers using the confusion matrix, it is observed that the RF achieves even better accuracy over DT, outperforming all the classifiers. A reason for RF performing better than might be that DT uses the entire sample in each step and picking up the decision boundaries at random, rather than picking the best one. It is quite evident with the results obtained, as RF marks an accuracy of 99%. The accuracy of DT and RF are significantly better, and it seems that these classifiers can be deployed to predict if the employee is likely to leave the company.

6 Conclusion

In employee attrition problem, an estimation can be framed for either the employee will leave the company or not. With this analysis, the organization can choose the employees with the utmost chances of leaving the organization and then assign them confined incentives. There could also be some cases of false positives where human resource thinks that employee will leave the company in a short span of time, but actually, the employee does not. These mistakes could be affluent and troublesome for both employees and human resource but is a better deal for relational growth. On the other hand, there could be a false negative, too, when a human resource does not give encouragement/hike to the employees, and they do leave the organization. That flaw of human resource is dangerous for the organization, as the company is not only losing an employee but also has to hire another employee and spent the cost of training and recruitment. Depending on this condition, we could categorize the type of treatment based on the types of employee incomes. If an employee is getting high pay, then the kind of treatment given to him by the company will be immoderate. Moreover, the cost of treatment should be weighted accordingly. The employee attrition prediction problem is about people's decision making. In this work, various machine learning techniques have been implemented on the human resource dataset. From the results obtained in this research work, it can be concluded that RF outperforms.

Compliance with ethical standards

Conflict of interest The authors declare that they have no confict of interest.

References

1. Larose DT, Larose CD (2014) Discovering knowledge in data: an introduction to data mining. Wiley, Hoboken
2. Thuraisingham B (2014) Data mining: technologies, techniques, tools, and trends. CRC Press, Boca Raton
3. Berry MJ, Linoff G (1997) Data mining techniques: for marketing, sales, and customer support. Wiley, Hoboken
4. Witten IH, Frank E, Hall MA, Pal CJ (2016) Data Mining: practical machine learning tools and techniques. Morgan Kaufmann, Burlington
5. Marchington M, Wilkinson A, Donnelly R, Kynighou A (2016) Human resource management at work. Kogan Page Publishers, London
6. Bloom N, Van Reenen J (2011) Human resource management and productivity, vol 4. Handbook of labor economics. Elsevier, Amsterdam, pp 1697–1767
7. Foster EC (2014) Human resource management. In: Software engineering. Apress, Berkeley, CA, pp 253–269
8. Datta DK, Guthrie JP, Wright PM (2005) Human resource management and labor productivity: does industry matter? Acad Manag J 48(1):135–145
9. Guest DE (2011) Human resource management and performance: still searching for some answers. Hum Resource Manag J 21(1):3–13
10. Lengnick-Hall ML et al (2019) Strategic human resource management: the evolution of the field. Hum Resource Manag Rev 19(2):64–85
11. Hamel G (2008) The future of management. In: Human resource management international digest
12. Bhargava N et al (2013) Decision tree analysis on J48 algorithm for data mining. In: Proceedings of international journal of advanced research in computer science and software engineering, vol 3(6)
13. Gerede ÇE, Mazan Z (2018) Will it pass? Predicting the outcome of a source code review. Turk J Electr Eng Comput Sci 26(3):1343–1353
14. Anitha A, Acharjya DP (2018) Crop suitability prediction in Vellore district using rough set on fuzzy approximation space and neural network. Neural Comput Appl 30(12):3633–3650
15. Arumugam A (2017) A predictive modeling approach for improving paddy crop productivity using data mining techniques. Turk J Electr Eng Comput Sci 25(6):4777–4787
16. Neslin Scott A et al (2006) Defection detection: measuring and understanding the predictive accuracy of customer churn models. J Mark Res 43(2):204–211
17. Keramati A et al (2014) Improved churn prediction in telecommunication industry using data mining techniques. Appl Soft Comput 24:994–1012
18. Gordini N, Veglio V (2017) Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. Ind Mark Manag 62:100–107
19. Zhu B, Baesens B, vanden Broucke SKLM (2017) An empirical comparison of techniques for the class imbalance problem in churn prediction. Inf Sci 408:84–99
20. Adankon MM, Cheriet M (2009) Support vector machine. In: Encyclopedia of biometrics. Springer, Boston, MA, pp 1303–1308
21. Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 213:660–674
22. Liaw A, Wiener M (2002) Classification and regression by randomForest. R News 23:18–22
23. Pedregosa F et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
24. Buitinck L et al (2013) API design for machine learning software: experiences from the scikit-learn project. arXiv preprint arXiv:1309.0238

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.