A New Approach for Employee Attrition Prediction

Chapter · September 2022

DOI: 10.1007/978-3-031-16663-1_9

2 authors, including:

Hamamache Kheddouci
University of Lyon

A new approach for employee attrition prediction

Lydia Douaidi[1,2] and Hamamache Kheddouci[2]

[1] École nationale Supérieure d'Informatique Algeria
[2] Université Claude Bernard Lyon 1, Lab LIRIS, UMR5205, 43 bd du 11 Novembre 1918, 69622 Villeurbanne Cedex, France
gl_douaidi@esi.dz
hamamache.kheddouci@univ-lyon1.fr

Abstract. Human resources are one of the most important assets in an organization, the success of any business or organization depends on its people achieving goals, meeting deadlines, maintaining quality and keeping customers happy. In a competitive environment, one of the biggest problems that companies face is employee departure or « Employee Attrition ».
The automatic prediction of employee attrition has only recently begun to attract the attention of researchers in various industries, as it can predict employee departure and identify the factors that influence employees to change employers.

Some AI platforms are developed in order to predict employee attrition, i.e. employees likely to change employer. The objective is to help companies anticipate departures to minimize financial losses due to employee attrition. However, these prediction systems are not generic and are specific to each company. In this context, we are interested in understanding the factors that influence an employee to leave his position. We aim to develop a generic attrition prediction platform that does not depend on the application domain, based on bipartite graph properties and machine learning algorithms.

Keywords: Machine learning, Bipartite graph, Employee attrition.

1 Introduction

In a context of increased competition, a competent and flexible workforce is one of the key factors on which organizations rely to improve their productivity, maintain their competitiveness and ensure their sustainability in the labour market. The issue of employee departures is taking significant prominence; retaining competent staff has become an intense challenge for organizations operating in different sectors of activity.
Indeed, regardless of the type of employee leaving, the associated organizational costs are often very high. Organizations need to review their human resource management practices in order to understand the visions and attitudes of their employees. Workplace policies that improve employee retention can help companies reduce their turnover costs.

The management of human resources plays a fundamental role in an organization's success [1]. Some studies show that the failure of many organizations originates in the poor management of these human resources, we assert that the organizations that have a reliable structure of human resources would have a means that would allow them to remain in a competitive environment. Employees have become aware of the power of their talents on organizations, they aspire to a stimulating job that allows them to use their potential and creativity to the maximum; to work in a company that would promote their learning and development outcomes.
Employee attrition means loss of employees or reduction of employees through a number of circumstances, such as retirement and resignation due to many reasons that can be professional or personal. Two terms are used interchangeably to designate the employee departure: attrition and turnover.

1.1 Attrition vs turnover

Employee turnover: when workers leave voluntarily or involuntarily an organization and the vacant post is occupied by another employee. It describes the rate at which an organization replaces departing workers with new workers.
Employee attrition: when employee leave voluntarily and the vacant post is not filled by the organization intentionally, it is considered as an intentional reduction in workforce because the organization decide not to rehire. Employees leave on their own for multiple reasons: to take a better offer with another organization, or to make a change in their career path.
We use the employee attrition rate as a metric to measure the rate at which employees leave a company during a specified time period - typically one year - it is calculated by dividing the number of employees that left during period by the average number of employees for period, and then multiplying that figure by 100. Attrition rate helps employers to measure the rate at which the employees leave without hiring their replacements, better examine the reasons of turnover.

Turnover classification Internal turnover occurs when employees leave their current positions in the company and take new positions within the same company. External turnover is when employees leave their current job positions and join another company or organization. The scientific literature distinguishes between voluntary and involuntary turnover ([2]; [3]). The authors define voluntary turnover as resulting from the employee's decision to leave his or her job, and involuntary turnover as resulting from the employer's decision to terminate the employment relationship. Thus, voluntary turnover might be functional and dysfunctional. Functional turnover occurs when a low-performing employee or employees without unique skills decide to leave their jobs, while dysfunctional turnover occurs when high-performing employee leaves and it can be avoidable or unavoidable. Figure 1 clarifies turnover classification scheme.

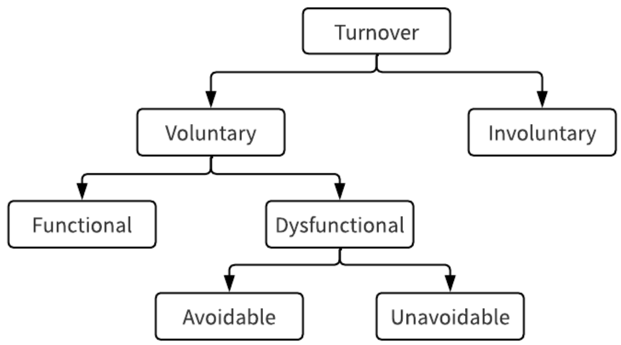

Fig. 1. Turnover classification [1]

1.2 Negative impacts of a high attrition rate

In today's competitive business environment, high employee attrition is a deeper issue and can have a negative impact on a company's performance. Customers and businesses are affected and the company's internal strengths and weaknesses are highlighted. According to [4], employee attrition is considered to have a direct and indirect impact on organization's costs. When an employee decides to change a company, it might be due to many reasons including personal, organizational, and technological factors.
Nowadays, machine learning approaches play an important role in employee attrition prediction. A number of approaches have been proposed to predict turnover based on historical data such as: age, salary, daily rate, level of satisfaction, years of experience, business travel, distance from home, education field, last promotion and so on. However, there are no studies that have explored the importance of relationships between employees and their interactions.
In this paper, we aim to propose a novel approach to predict attrition based on interactions between employees using Machine Learning methods and bipartite graphs. To realize this objective, we address the following research question in this paper: Can the relationships between employees influence them to change the job ?

This paper is organized as follows: The next section provides an overview of related works. In Sections 3 and 4, the core of the paper, we present our dataset (characteristics of the dataset), bipartite graph to model employees and companies and machine learning models to predict employee attrition. In Section 5, we present a recapitulative table of results. Finally, we conclude and present future research plans.

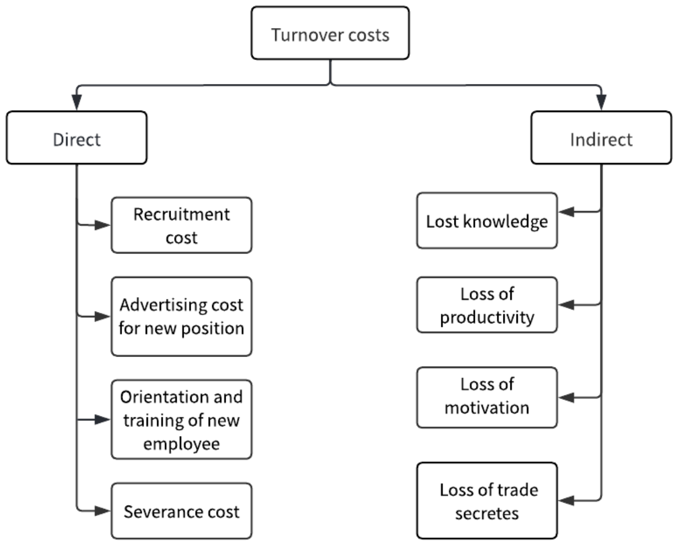

Fig. 2. Turnover costs [4]

2 Related works

Many approaches were applied to predict employee attrition. Over the last decade, there has been increasing interest for relevant studies in areas including manufacturing, telecommunication industry [5], healthcare industry [6] [7], etc.

Reference [8] aims to analyze employee attrition using logistic regression. The findings of the study, that working atmosphere, job satisfaction, and communication between manager, leader and subordinates, are the major reasons for leaving the job. Another paper [9] used Logistic Regression for the IBM HR dataset[3] when predicting employee turnover and obtained an accuracy of 85%. However, this study did not select influential features for an employee to leave the job. In [10], the authors tried several classifiers: Logistic Regression, K-Nearest Neighbor, Random Forest and Principal Component Analysis to reduce the feature space's dimensionality. They used a dataset of the U.S. Office of Personnel Management (OPM), the conclusion is that logistic regression predict attrition with the highest accuracy and the most relevant features are: length of service; lower age limit and age. The authors in [11] explored different decision tree algorithms (C4.5, C5.0, REPTree and CART). The research results found that salary and length of service were the strongest feature in the tested dataset, they recommend using C5.0 to predict employee attrition. XGBoost model and IBM HR

[3] https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

dataset, are used in [12] in order to predict employees changing jobs. Experimental results state that the XGBoost is more accurate for predictive analysis. Other works, such as [13], also used the IBM dataset, they proposed different feature selection techniques (Information Gain, Gain Ratio, Chi-Square, Correlation-based, and Fisher Exact test) based on Machine Learning models (SVM, ANN, Gradient Boosting Tree, Bagging, Decision Trees and Random Forest). In [14], the authors developed a neural network simultaneous optimization algorithm (NNSOA) based on Genetic Algorithm which performs well for overall classification. The authors of [15] analysed email communications using recurrent neural network (RNN) and text analysis metrics. As a result, they recommend using RNN with LSTM and GRU, which has 84,4% accuracy and 81,6% AUC. Authors in [16] proposed a novel approach to predict attrition using a dynamic bipartite graph. They represented employees and companies with a bipartite graph. Then, they employed Horary Random Walk Algorithm to generate sequence for each vertex in chronological order and apply regression to predict whether an employee intends to leave by combining embedded features with basic features. In [17], unsupervised machine learning methods (K-means clustering) are used on attrition rates to determine an appropriate segmentation and number of segments (very high, high, medium, and low). Recently, other research in [18], analyzed the dataset IBM Employee Attrition to train and evaluate several classifiers (Decision Tree, Random Forest Regressor, Logistic Regressor, Ada Boost Model, and Gradient Boosting Classifier models).

In [19], the researchers call for research that considers mental health issues such as depression, anxiety. They proposed a model to predict employee attrition rate and the employees emotional assessment in an organization with an accuracy of 86.0%.

Regarding the methodologies, some researchers are trying to predict attrition using standard methods: data on attrition are collected through methods such as interviews and surveys. Then, the analysis is conducted using linear regression, SVM, KNN, decision trees, extreme gradient boosting, etc. Other researchers proposed models that used professional trajectories to predict turnover. They generally concentrate only on the employee attrition prediction but for employers it's important to investigate the true causes of attrition [20]. The existing studies considers only demographic variables [21] (race and gender) and personal employee information such as age and salary because they are based on the same public dataset, but don't consider the importance of relations between employees to explain turnover. When an employee leaves, the culture and commitment of the remaining employees to the company can be severely affected and can cause many other employees to think about departing too. To this end, it is important to take into consideration relational variables to predict employee attrition.

3 Approach description

3.1 Data description

In our experiments, we collected data from a professional social network, where users can create a professional resume that contains all of the required information about education, previous job experiences, skills, training & participation, etc.
The collected dataset contains features about professional experience (job title, company name, company type, dates worked), locality of employees and other information about academic career.

3.2 Variable extraction

In order to extract features about social interactions between employees, we present the attrition problem with a bipartite graph $G = (X, Y, E, T)$ [16]:

- $V = X \cup Y$ where X is the set of vertices representing the employees and Y is the set of vertices representing the companies, V is the set of all vertices of the graph.
- $E \in X * Y$ defines the edges between the two sets X and Y. We add an edge between employee $x_i \in X$ and company $y_j \in Y$ if the i_{th} employee worked for the j_{th} company.
- T is the timestamp of E. For each edge $e = (x, y, t) \in E$ connecting a vertex $x_i \in X$ to a vertex $y_j \in Y$, is associated a unique timestamp t representing the time when the employee joined the company.

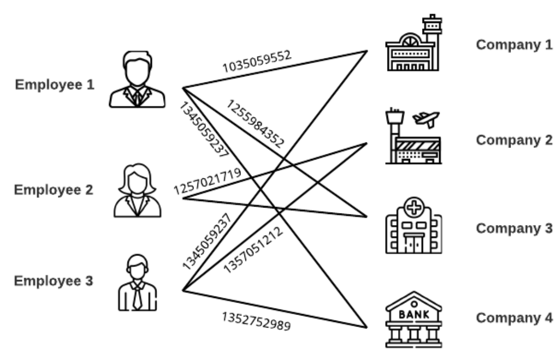

Fig. 3. A small example of how a bipartite employee-company graph can be visualized

Starting from the bipartite graph G of employees and companies, we created a monopartite projection: Graph of employees who have worked in the same companies and the number of shared companies being the weight of the link, as can be seen in Figure 4.

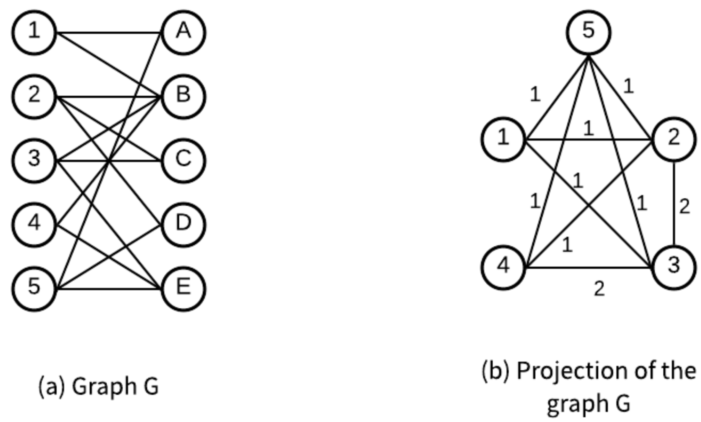

Fig. 4. (a) Employee-company bipartite graph, (b) Employees one-mode projection graph

In our approach, we used network centrality measurements to extract some features about interactions between employees from the bipartite graph projection.

A graph $G = (V, E)$ is composed of a set of vertices V, set of edges E. G is represented by its adjacency matrix $A = (a_{i,j})$ and N is the number of nodes in G. In our study we assume that G is weighted, undirected and connected. We briefly review some of the most important centralities:

Degree centrality: Defined by the number of links adjacent to a node, and in our graph it represents the number of direct contacts of an employee. The degree centrality of $v_i \in V$ denoted by $DC(v_i)$, is a normalized value defined as follows [22]:

$$DC(v_i) = \frac{1}{N-1} \sum_{j=1}^{N} (a_{ij}) \tag{1}$$

Closeness Centrality: Measure the average shortest distance from a vertex to others. We use this metric to find employees - who finally decide to leave- who are best placed to influence other employees to change job.

In the case where the graph G is non-oriented, closeness centrality of $v_i \in V$ is defined as [22]:

$$CC(v_i) = \frac{N-1}{\sum_{j=1}^{N} d(v_i v_j)} \tag{2}$$

where $d(v_i, v_j)$ is the shortest path between node v_i and v_j.
Betweenness Centrality: represents the capacity of a node to be an intermediary between any two other nodes [23]. Betweenness centrality of a node v_i is formally defined as follows [22]:

$$BC(v_i) = \sum_{j=1}^{N} \sum_{k=1}^{N} \frac{g_{jk}(v_i)}{g_{jk}} \quad (3)$$

where $g_{jk}(v_i)$ is the number of shortest paths between node $v_j \in V$ and node $v_k \in V$ that pass through node $v_i \in V$.
g_{jk} is the number of shortest paths between v_j and v_k.
Eigenvector centrality: measures the influence of a node based on number of links it has to other nodes in the graph. This index is based on the principle that a link with a poorly connected node is worth less than a link with a highly connected node. Eigenvector centrality of a node v_i is formally defined as follows [22]:

$$EC(v_i) = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij}\ EC(v_j) \quad (4)$$

where λ is a fixed constant. n is the number of neighbours of v_i and v_j is the j_{th} neighbour of v_i.
As earlier mentioned, we added some features to the initial dataset using the projected graph (Figure 4 (b)). For each node, we calculated four different network centrality metrics: degree centrality, closeness, betweenness and eigenvector centrality. We then calculated the average time spent working (average job duration), number of organizations in which the person worked, difference in years between previous job of the employee and current job (Last new job). We added all this features to our initial dataset.

We aim to predict whether $x \in X$ will quit from company $y \in Y$ after a timespan Δt. In order to build this binary classification dataset, we annotated our dataset using average job duration. For example, we mark a record with average job duration less than 36 months (3 years) as a positive example (attrition behaviour occured), otherwise we mark it as negative example [16].
Table 1 presents the description of the features extracted and used to train ML models.

3.3 Data exploration

The study sample included 22 167 employees. Figure 5 shows distribution of attrition in our dataset based on gender (Of the 100% employees who left, 88.9% were men and 9.6% were women and 1.5% Others).
Figure 6 shows attrition rate among 10 selected industries (we selected sectors with largest employment in our dataset), we can observe that the attrition rate in the Information Technology and Services sector was the highest among all industries (7,2%), followed by Oil and Energy sectors (3,7%). According to [24], stress because of high work pressure causes the high rate of attrition in IT sector.

Table 1. Description of employee features

Feature	**Description**
Gender	gender of the person
Locality	city where the person lives
Experiences	number of organizations in which the person worked
Company type	type of current employer
Company size	number of employees in current employer's company
Average job duration	average duration of jobs
Last new job	difference in years between previous job and current job
Enrolled university	type of University course enrolled if any
Academic degree	award conferred by a college or university
Major discipline	academic discipline pursued by a person
Degree centrality	score calculated using the graph
Closeness centrality	score calculated using the graph
Betweenness centrality	score calculated using the graph
Eigenvector centrality	score calculated using the graph

3.4 Methodology

In our study we have an imbalanced dataset (a total of 75.1% of instances are labeled with class-1 and the remaining 24.9% of instances are labeled with Class-2) and most models trained on imbalanced data will have a bias towards the majority class only and in many cases, may avoid the minority class. In order to tackle highly imbalance dataset, numerous techniques have been proposed: data preprocessing methods, algorithm modification methods, cost-sensitive methods, and ensemble learning approaches [25]. We choose Synthetic Minority Over-sampling Technique (SMOTE) proposed by [26] to create new instances using linear interpolation between minority class neighbouring points.
Given the dataset described in Table 1, we encode categorical data using Label encoding then we randomly select 80% of the dataset for the training set and 20% as the test set.
In this study, the extracted features were initially at different scales and some machine learning algorithms are sensitive to feature scaling (using Gradient Descent Based Algorithms certain weights may update faster than others since the feature values play a role in the weight updates). Therefore, the data were

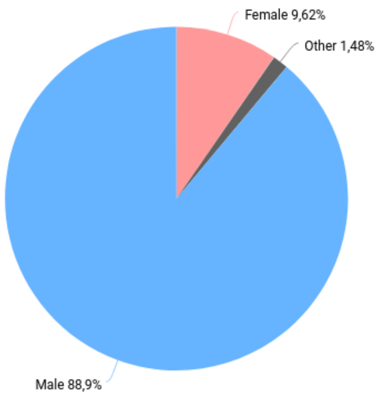

Fig. 5. Percentage rates of attrition by gender

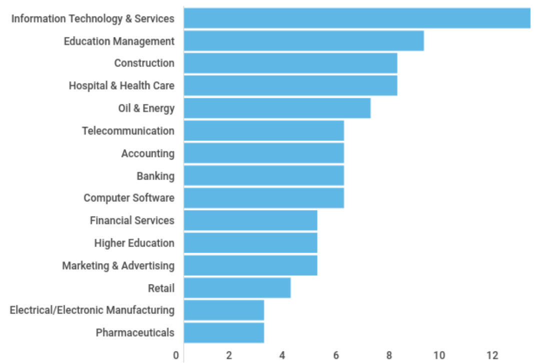

Fig. 6. Attrition rate by industry

normalized by using the z-score normalization which is calculated as follows:

$$\text{z-score} = \frac{x - \mu}{\sigma} \tag{5}$$

where μ is the mean (average) and σ is the standard deviation from the mean. Most studies applied supervised prediction and classification algorithms capable of modeling linear and non-linear relationships between variables to develop predictive attrition models. These studies trained a machine learning algorithm using a labelled dataset that contain personal features about employees (such as age, salary) to iteratively evaluate, compare, and select variables that would predict attrition with the highest accuracy. This study investigated the ability to predict attrition using personal features and relational features to shed some light on how the individual actors influence one other to left the company. From the literature review, we selected the 8 most used Machine Learning algorithms for the prediction: Logistic Regression, Support Vector Machine with Radial Basis Function kernel (SVM), Decision Tree (DT), and Random Forest (RF), K-Nearest Neighbors (KNN), Random Forest (RF), Extra Trees (ET), Gradient Boosting (GB) and LightGBM. The classifiers were trained using Scikit-learn [27].

3.5 Results

After building the models, we evaluated them on unseen test data. Several metrics have been proposed for evaluating classification performances [28]. The four base performance measures in a binary classification with supervised learning approach are: True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). We present the most preferred and known evaluation metrics used to evaluate the quality of a binary classification system:

- Confusion matrix: a performance measurement for classification problems, it represents counts from predicted and actual values [24].
- Accuracy: ratio of predictions the model got right [28], defined by:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{6}$$

- Precision: ratio between positive samples and all the positives [28], defined by:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

- Recall: ratio of samples that are predicted to be positive and the total number of predictions [28], defined by:

$$\text{Rappel} = \frac{TP}{TP + FN} \tag{8}$$

- F1-score: Harmonic mean of the Precision and Recall [28], defined by:

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

As we have mentioned before, to evaluate the proposed approach, we conducted experiments using a real dataset collected from a professional social network. The overall results for accuracy, precision, recall, and f1-score are listed in Table 2.

Table 2. Classification results of all the applied models.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression	0.68	0.59	0.67	0.63
Support Vector Machine	0.77	0.31	0.80	0.45
K-Nearest Neighbors	0.88	0.76	0.68	0.71
Decision Tree	0.82	0.83	0.84	0.83
Random Forest	0.92	0.80	0.89	0.84
Extra Trees	0.93	0.83	0.87	0.85
Gradient Boosting	0.93	0.84	0.87	0.85
LGBM	0.82	0.84	0.86	0.85

Table 2 shows that the accuracy of Random Forest (92%), Extra Trees (93%) and Gradient Boosting (93%) are significantly better, Random Forest gave the highest recall (0.89%), SVM achieves an accuracy of 77% and a precision of 33%. In this research, tree-based classifiers (Decision Trees, Random Forest, Extra Trees and Gradient Boosting) worked well in general, and were found to be the top best performing classifiers. Tree based models provide a good predictive performance and easy interpretability when dealing with the attrition problem, it is straightforward to derive the importance of features on the tree decision. For this reason, we calculated features importance using Gradient Boosting model to estimate the importance of features. The results are shown in 7.

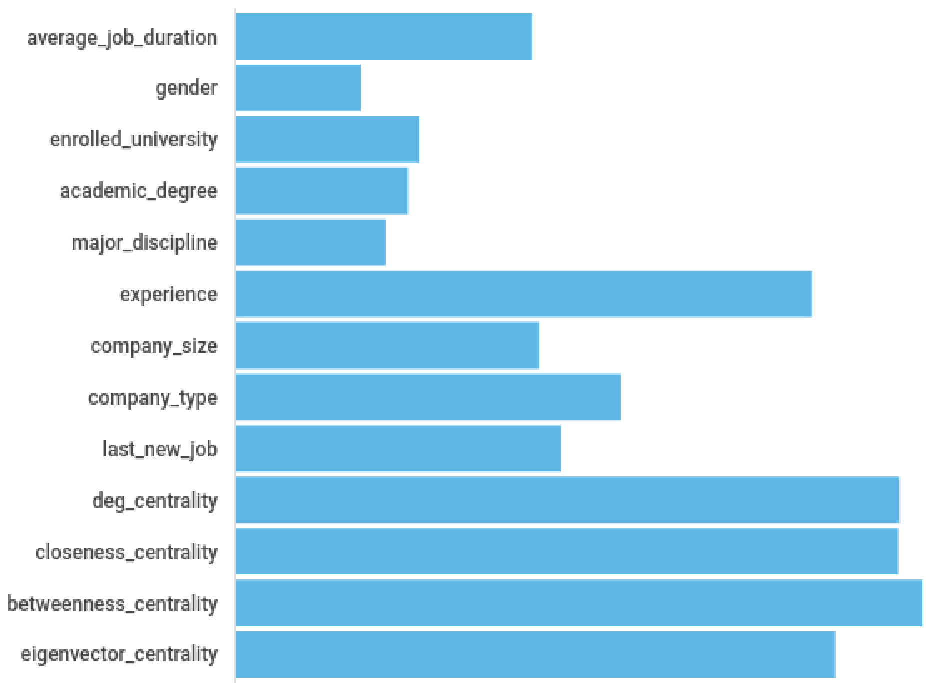

Fig. 7. Features importance using Gradient Boosting model

As we can see, the variables extracted from the bipartite graph (degree centrality, eigenvector centrality, closeness centrality and betweenness centrality) have a major contribution to the attrition problem, followed by the other features such as experience, company type and job duration. This results demonstrate that social connections of an employee with others can influence his decision to stay or leave the company. Managers who want to reduce employee attrition should focus on how connected their employees are with others in the organization, workers tends to leave when they are connected to another leaver or have relationships with other employees who have left.

4 Conclusion

Companies are always looking for ways to retain their professional employees to reduce additional costs of recruiting and training. Predicting whether a particular employee will leave the company or not helps the company make preventive decisions. In this paper, we studied employee attrition problem to predict whether an employee may leave the company. We trained the models (LR, SVM, KNN, DT, RF, ET, GB and LGBM) using training data extracted from a professional social network and tested the models on testing data. Our approach integrates interaction features extracted from a bipartite graph and personal information of employees. Using the proposed approach to study attrition, employers or human resource managers have the opportunity to recognize the importance of relation networks generated by employees who chose to leave the company and by others who decided to stay. In comparison with previous works, the model demonstrated that relational features, such as the ones proposed in this research, are important and lead to a greater attrition probability. We will continue to improve the performance of our model and conduct in depth analysis by exploring the bipartite graph of employees in order to extract other type of features. Another future direction would be to examine our method in other domains such as information technology networks.

References

1. Nancy Heppell. *Le roulement du personnel et la performance organisationnelle: l'effet modérateur des pratiques de gestion des ressources humaines*. 2012. University of Montreal.
2. Dan R Dalton, William D Todor, and David M Krackhardt. Turnover overstated: The functional taxonomy. *Academy of management Review*, 7(1):117–123, 1982. https://doi.org/10.5465/amr.1982.4285499.
3. John R Hollenbeck and Charles R Williams. Turnover functionality versus turnover frequency: A note on work attitudes and organizational effectiveness. *Journal of Applied Psychology*, 71(4):606, 1986. https://doi.org/10.1037/0021-9010.71.4.606.
4. T Harika and N Hima Bindu. Employee retention. *Journal Homepage: http://ijmr.net. in*, 8(03), 2020.
5. KM Suceendran, R Saravanan, Divya Ananthram, S Poonkuzhali, R Kishore Kumar, and K Sarukesi. Applying classifier algorithms to organizational memory to build an attrition predictor model. *Advances In Information Science And Computer Engineering*, 2015.
6. Huey-Ming Tzeng, Jer-Guang Hsieh, and Yih-Lon Lin. Predicting nurses' intention to quit with a support vector machine: A new approach to set up an early warning mechanism in human resource management. *CIN: Computers, Informatics, Nursing*, 22(4):232–242, 2004.
7. Mark John Somers. Application of two neural network paradigms to the study of voluntary employee turnover. *Journal of Applied Psychology*, 84(2):177, 1999. https://doi.org/10.1037/0021-9010.84.2.177.
8. I Setiawan, S Suprihanto, A C Nugraha, and J Hutahaean. HR analytics: Employee attrition analysis using logistic regression. *IOP Conference Series: Materials Science and Engineering*, 830:032001, may 2020. https://doi.org/10.1088/1757-899x/830/3/032001.

9. S Ponnuru, G Merugumala, Srinivasulu Padigala, Ramya Vanga, and Bhaskar Kantapalli. Employee attrition prediction using logistic regression. *Int. J. Res. Appl. Sci. Eng. Technol*, 8:2871–2875, 2020.
10. Alex Frye, Christopher Boomhower, Michael Smith, Lindsay Vitovsky, and Stacey Fabricant. Employee attrition: What makes an employee quit? *SMU Data Science Review*, 1(1):9, 2018. https://scholar.smu.edu/datasciencereview/vol1/iss1/9/.
11. DABA Alao and AB Adeyemo. Analyzing employee attrition using decision tree algorithms. *Computing, Information Systems, Development Informatics and Allied Research Journal*, 4(1):17–28, 2013.
12. Rachna Jain and Anand Nayyar. Predicting employee attrition using xgboost machine learning approach. In *2018 International Conference on System Modeling & Advancement in Research Trends (SMART)*, pages 113–120. IEEE, 2018. https://doi.org/10.1109/SYSMART.2018.8746940.
13. M Subhashini and R Gopinath. Employee attrition prediction in industry using machine learning techniques. *International Journal of Advanced Research in Engineering and Technology*, 11:3329–334, 2021. http://14.139.186.108/jspui/bitstream/123456789/32015/1/IJARET_11_12_313.pdf.
14. Randall S Sexton, Robert E Dorsey, and Naheel A Sikander. Simultaneous optimization of neural network function and architecture algorithm. *Decision Support Systems*, 36(3):283–296, 2004. https://doi.org/10.1016/S0167-9236(02)00147-1.
15. Joao Marcos de Oliveira, Matthäus P Zylka, Peter A Gloor, and Tushar Joshi. Mirror, mirror on the wall, who is leaving of them all: Predictions for employee turnover with gated recurrent neural networks. In *Collaborative innovation networks*, pages 43–59. Springer, 2019. https://doi.org/10.1007/978-3-030-17238-1_2.
16. Xinjun Cai, Jiaxing Shang, Ziwei Jin, Feiyi Liu, Baohua Qiang, Wu Xie, and Liang Zhao. Dbge: employee turnover prediction based on dynamic bipartite graph embedding. *IEEE Access*, 8:10390–10402, 2020. https://doi.org/10.1109/ACCESS.2020.2965544.
17. Nianjun Zhou, Wesley M Gifford, Junchi Yan, and Hongfei Li. End-to-end solution with clustering method for attrition analysis. In *2016 IEEE International Conference on Services Computing (SCC)*, pages 363–370. IEEE, 2016. https://doi.org/10.1109/SCC.2016.54.
18. Aseel Qutub, Asmaa Al-Mehmadi, Munirah Al-Hssan, Ruyan Aljohani, and Hanan S Alghamdi. Prediction of employee attrition using machine learning and ensemble methods. *Int. J. Mach. Learn. Comput*, 11, 2021. https://doi.org/10.18178/ijmlc.2021.11.2.1022.
19. Richard Joseph, Shreyas Udupa, Sanket Jangale, Kunal Kotkar, and Parthesh Pawar. Employee attrition using machine learning and depression analysis. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1000–1005. IEEE, 2021. https://doi.org/10.1109/ICICCS51141.2021.9432259.
20. Nesrine Ben Yahia, Jihen Hlel, and Ricardo Colomo-Palacios. From big data to deep data to support people analytics for employee attrition prediction. *IEEE Access*, 9:60447–60458, 2021. https://doi.org/10.1109/ACCESS.2021.3074559.
21. Shine David, Saakshi Kaushik, Harshita Verma, and Shivani Sharma. Attrition in it sector. *International Journal of Core Engineering and Management*, 2(1):236–246, 2015.
22. Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social media mining: an introduction*. Cambridge University Press, 2014.

23. Peter A. Gloor, Andrea Fronzetti Colladon, Francesca Grippa, and Gianni Giacomelli. Forecasting managerial turnover through e-mail based social network analysis. *Computers in Human Behavior*, 71:343–352, 2017. https://doi.org/10.1016/j.chb.2017.02.017.
24. A Dean Forbes. Classification-algorithm evaluation: Five performance measures based onconfusion matrices. *Journal of Clinical Monitoring*, 11(3):189–206, 1995. https://doi.org/10.1007/BF01617722.
25. Pattaramon Vuttipittayamongkol and Eyad Elyan. Improved overlap-based undersampling for imbalanced dataset classification with application to epilepsy and parkinson's disease. *International journal of neural systems*, 30(08):2050043, 2020. https://doi.org/10.1142/S0129065720500434.
26. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. https://doi.org/10.1613/jair.953.
27. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
28. Gürol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 821–826. IEEE, 2017. https://doi.org/10.1109/UBMK.2017.8093539.