EARLY PREDICTION OF EMPLOYEE ATTRITION

B. Sri Harsha, A. Jithendra Varaprasad, L.V N Pavan Sai Sujith

Abstract— Employees are the significant resources of any association. In any case, in the event that they quit job unexpectedly, it might bring about immense expense to any organization. Since new hiring will consume money and time as well as the newly hired employees set aside some effort to make the particular organization productive. Subsequently in this paper we attempt to develop a model which will predict employee attrition rate dependent on HR analytics dataset. "Prediction the employee attrition and reasons for leaving the organization" was performed to see the reasons, why the best and most experienced workers leave the organization and attempt to anticipate which valuable employees are plausible to leave the organization along these lines in order to discover the territories where the association is lagging behind. This model can be utilized by the Human Resource branches of the organizations to shape proficient methodologies to hold the important representatives before they start searching for new employments like by giving a hike in their compensation.

Index Terms— Employee Attrition, Classification, job satisfaction, Retention, employee retention strategy.

1 INTRODUCTION

Employee Attrition is a decrease in man power in any organization where workers may wilfully leave the organization or might be resigned. Employee turnover is the number of existing workers traded by new employees for a particular period. A high attrition causes high worker turnover in any organization. This thus causes huge expenditure on human resources, by contributing towards new enrolment, preparing and improvement of the freshly hired workers, likewise the performance management. Once more, attrition which are of voluntarily unavoidable. Henceforth, by improving employee morale and giving a desirable working conditions, we can unquestionably decrease this issue fundamentally.The rate of attrition is characterized as the enrolment and termination criteria of the organization. An employee can leave the job for different explanation. Here, the 'Turnover' and 'Attrition' are the business wordings that always conflicts each other. There are different sorts of 'turnover' in a company. Bringing down in number of employees is mostly considered as the 'attrition'. To analyse the man power and other estimations that are fundamental for manpower planning these wordings can be conversely utilized. At the point when an employee leaves the organization both attrition and turnover occur. Turnover, in any case happen due to different work activities, for instance, release, termination, abandonment or on the other hand occupation surrender. Attrition happens when a worker leaves or when the association takes out his occupation. The differentiation between the two is that when turnover occurs, the association searches for someone to supplant the employee. In examples of attrition, the business leaves the opportunity unfilled or clears out that business work.

Predicting employee attrition at an organization will enable management to act quicker by upgrading their internal policies and strategies. Where talented employees with a risk of leaving can be offered a few recommendations, for example, a pay increment or proper training, to reduce their probability of leaving. Utilizing machine learning models can help organizations to anticipate employee attrition. Using the historical data kept in human resources (HR) departments, analysts can build and prepare a machine learning model that can predict the workers who are leaving the organization. Such models are prepared to look at the connection between the features of both active and terminated workers. This paper is composed as follows. The section II portrays the related works done previously and the inspiration as to analysis. Section III will depict different machine learning algorithms utilized in this paper and their advantages. Section IV portrays about the data set description. Section V contains the experimental results utilizing the machine learning algorithms utilizing the mentioned data set, which will be followed by the conclusion. To modify the running headings, select View | Header and Footer. Click inside the text box to type the name of the journal the article is being submitted to and the manuscript identification number. Click the forward arrow in the pop-up tool bar to modify the header or footer on subsequent pages. IJSTR staff will edit and complete the final formatting of your paper.

2. LITERATURE SURVEY

Nagadevara et al, (2008), investigated the relationship of withdrawal practices like delay and non-attendance, job content, residency and socio-economics on worker turnover in a quickly developing segment like the Indian software industry. The extraordinary part of this research was the utilization of five prescient data mining procedures (artificial neural networks, logistic regression, classification and regression trees, classification trees (C5.0), and discriminant analysis) on a sample data of 150 employee in a large software organization. The consequences of the research clearly demonstrate a connection between withdrawal behaviours and employee turnover. This study raised a few issues for future research. To start with, further research could explicitly gather data on statistic factors over a large sample of organizations to inspect the connection between statistic factors and turnover. Second, large scale data on variables in the past academic research which have an association with turnover can be collected [1]. In an exposition by Marjorie Laura Kane-Sellers (2007), the researchers did an examination to explore the factors affecting worker voluntary turnover in the North American professional sales force of a Fortune 500 industrial manufacturing firm. By studying Voluntary Turn Over, the

- *B.Sri Harsha is currently pursuing Bachelor degree program in Computer Science engineering in KLEF, India, PH-+918790892104. E-mail: sriharshabhavaraju@gmail.com*
- *A.Jitendra is currently pursuing Bachelor degree program in Computer science engineering in KLEF, India, PH-+919642852207. E-mail: jithendra.angalakuduru@gmail.com*

expectation was to increase a superior understanding of Human Resource Development interventions that could improve employee retention. The central firm gave perceptions of the worker database for all individuals from the expert specialized deals power over a 14-years longitudinal period. The first database passed on 21,271 discrete observations distinguished by unique employee clock number [2]. Ibrahim et al. proposed to solve a major issue of customer churn identified with a business, particularly telecommunications by building models with various procedures, for example, Classification for prediction, Clustering for detection and Association for detection [3]. Choudhary et.al, presents the application of logistic regression method dependent on the data of employees to build up a risk equation to predict employee attrition. Later this equation was applied to assess attrition risk with the current set of employees. After the estimation, high risk cluster was recognized to discover the reasons and henceforth action plan was selected to minimize the risk [4]. IBM Watson team M. Singh et al. has done an analysis of employee's attrition procedure and proposed a structure which discovers the purpose behind attrition and recognizing potential attrition. They have attempted to compute the cost of attrition and proposing 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) [5]. In Q. A. Al-Radaideh et al. the authors used decision trees (ID3 C4.5) and Naïve Bayes classifier to predict employee performance. They found that job title was the strongest feature, whereas age did not show any clear effect [6]. In O.Ali et al. Employees are more likely to leave, perhaps because of a disagreement with their senior officer. We noticed major factors affecting the company's loss of workers. He derives moderately from the two rules. Some questions are asked with both parties and he concluded some facts based on workload, goals, carrier, depending on their answers[7]. In A. frederiksn et al. Human resource management focuses on termination rates and firing rates in general, but their actual content is vastly different. The previous model indicates that recruitment and turnover are several distinct grades. Some work suggests that there are institutional implications of the dismissal and termination rates [8]. In H.Ongori et al. For the negative side of turnover, Allen & Meyer (1990) defined the three-basic entity. Regulating officer will be more likely to leave the company due to a dispute with the higher administration than a delegate dealing with his prompt boss. He acknowledged the guiding factors that control the organization's acceptance of workers without protest [9]. In V.V. Saradhi et al. Two systems of techniques for social opportunity data are guided. An equal number of respondents from members and officers were asked to respond to a collection of polls organized by workload, priorities, personality, professional success, and hierarchical management. The after-effects of the two methods of collecting information have shown that the most notable aspect employee rejection is money related compensation [10]. In Hossein et al. Throughout their research, they followed the data mining methodology of CRISP-DM. The Decision tree was the main data mining tool used to construct the model of classification where several rules of classification were developed. The developed model was validated. several experiments were carried out using real data obtained from several companies. The model is intended to predict the quality of new applicants[11]. Amir Mohammad et al. implemented steps of knowledge discovery on the actual data of a factory. They master many employee characteristics such as age, technical skills, and work experience. Using the Pearson Chi-Square test, they considered the significance of software features [12].

Rohit Punnoose and Pankaj Ajit et al. examined the implementation of the technique of Extreme Gradient Boosting (XGBoost), which is more reliable due to its formulation of regularization. A global retailer's HRIS information is used to equate XGBoost with six traditionally used supervised classifiers to demonstrate its significantly higher reliability to predict employee turnover [13]. The prediction of Churn, particularly the prediction of customer churn, attracted great attention from researchers. For example, Verbeke et al. suggest a profit-centric measure of performance by measuring the maximum profit that can be produced by including the ideal customer fraction with the highest expected likelihood of churning in a retention campaign[14]. The problem of optimizing the quality of a decision support system for churn prediction was studied by Coussement and Van den Poel. In the churn prediction process, they studied the impact of textual data. They found that adding unstructured, textual information into a conventional churn prediction model resulted in a significant increase in predictive performance [15]. In order to predict customer churns, Coussement and Van den Poel implement SVM method. Their study shows that when applied to noisy marketing data, supporting vector machines results in good generalization performance[17].

Burez and Van den Poel are researching consumer churn forecasting class imbalances. Study results show that under-sampling can lead to better accuracy of predictions [18].

Tsai and Chen use association rules in another study to pick important features and then apply neural networks and Decision Tree to predict a telecommunications company's customer churns. They use four performance measurements similar to us to analyze their results, accuracy, accuracy, recall, and F-measurement. Cushioning et al [19]. Build the Generalized Additive Models (GAM) method which enables the model to fit complex non-linear data. There is also other research that use well-known data mining methods to estimate client churns [20]. In comparison, there are few literature studies that consider the estimation and study of employee turnover. Saradhi and Palshikar use naive Bayes, Logistic Regression, Decision Tree, and Random Forest methods to study employee churn prediction [21]. To the best of our knowledge, the last research is the report by Kane-Sellers on the database of the skilled sales force of Fortune 500 North American industrial automation manufacturer. Kane-Sellers ' main method is the method of logistic regression [22].

3.Algorithms

This paper talks about supervised learning algorithms for classification, since we are aware of the presence of two classes working and left.

A. Naive Bayes

Naive Bayes is a classification strategy that has picked up fame because of its simplicity. The Naive Bayes algorithm utilizes the assumption that every one of the variables are independent to each other, and after that compute's probabilities, that are utilized for classification. The algorithm works as follows: to get an output function Y given a set of input variables X, the algorithm estimates the values of P(X|Y) and P(Y), and then uses Bayes' rule to compute P(Y|X), which is the required output, for each of the new samples.

B. Logistic Regression

Logistic regression is a regression model that fits the values to the logistic function. It is useful when the dependent variable is categorical [5]. The general form of the model is

$$P(Y|X,W)= 1/(1+e\hat{}(-(w0+\Sigma wixi))$$

Logistic regression is often used with regularization techniques to prevent overfitting.

C. K-Nearest Neighbours (KNN)

The KNN algorithm classifies new data dependent on the class of the k closest neighbours. This paper utilizes the estimation of k as 6. The good ways from neighbours can be determined utilizing different distance metrics, for example, Euclidean distance, Manhattan distance, Minkowski distance, and so forth. The class of the new information might be chosen by dominant part vote or by an inverse proposition to the distance computed. KNN is a non-generalizing technique, since the algorithm keeps the majority of its preparation information in memory, perhaps changed into a quick ordering structure, for example, a ball tree or a KD tree.

The Manhattan distance is computed using the formula

$$D = \Sigma\ |xi - yi\ |$$

D. Random Forest

A Random Forest consists of a collection or ensemble of simple tree predictors, each capable of producing a response when presented with a set of predictor values. For classification problems, this response takes the form of a class

$$\text{Random Forest Prediction } s = \frac{1}{K}\sum_{K=1}^{K} K^{th} \text{ tree response}$$

membership, which associates, or classifies, a set of independent predictor values with one of the categories present in the dependent variable. Alternatively, for regression problems, the tree response is an estimate of the dependent variable given the predictors. The Random Forest algorithm was developed by Breiman. The predictions of the Random Forest are taken to be the average of the predictions of the trees:

E. Support Vector Machine

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.

4.Methodology

We apply a wide range of data mining techniques from as simple as Naive Bayes, linear regression and nearest neighbours to more complex techniques as SVM, Random Forests and other ensemble methods.

4.1 Algorithm

1) Identify the employee dataset that comprises of current also, past workers records
2) Clean the dataset, handle the missing data and determine new features whenever required
3) Select the features among the worker data that are appropriate for the prediction of attrition
4) Try classification algorithms and report the ones most appropriate by looking at the precision, accuracy, recall, and F-measure results on the test data
5) Apply feature selection method, and select the features that are more convenient in order to predict employee attrition
6) Build classification model
7) Further the forecast of employee attrition on utilizing the model
8) Decision on the methodologies of retention

4.2. DATA SET DESCRIPTION

The In this research, we utilized an openly available dataset, which can be acquired from IBM Watson Analytics1. The dataset involves engineered information made by IBM data scientists. The dataset contains the HR-related information of 1470 workers with 32 highlights. Also, aggregate of 1233 active employees were from "No" attrition category though the staying 237 previous workers were from "Yes" attrition classification in this exploration, two highlights were evacuated: 'Employee count', because it is a sequence of numbers (1,2, 3.) ; and 'Standard hours', since all employees have the same standard hours. Also, all non-numerical values were assigned numerical values for processing such as: Sales=1, Research & Development=2, Human Resources=3.

Table 1.HR data set features

S.NO	FEATURE	DATATYPE
1	Age	Numeric
2	Attrition	Categorical
3	BusinessTravel	Categorical
4	DailyRate	Numeric
5	Department	Categorical
6	DistanceFromHome	Numeric
7	Education	Numeric
8	EducationField	Categorical
9	EmployeeCount	Numeric
10	EmployeeNumber	Numeric
11	Environment-Satisfaction	Numeric
12	Gender	Categorical
13	HourlyRate	Numeric
14	JobInvolvement	Numeric
15	JobLevel	Numeric

16	JobRole	Categorical
17	JobSatisfaction	Numeric
18	MaritalStatus	Categorical
19	MonthlyIncome	Numeric
20	MonthlyRate	Numeric
21	NumCompanies-Worked	Numeric
22	Over18	Categorical
23	OverTime	Categorical
24	PercentSalaryHike	Numeric
25	PerformanceRating	Numeric
26	RelationshipSatisfaction	Numeric
27	StandardHours	Numeric
28	StockOptionLevel	Numeric
29	TotalWorkingYears	Numeric
30	TrainingTimesLastYear	Numeric
31	WorkLifeBalance	Numeric
32	YearsAtCompany	Numeric
33	YearsInCurrentRole	Numeric
34	YearsSinceLastPromotion	Numeric
35	YearsWithCurrManager	Numeric

4.3. Proposed Model

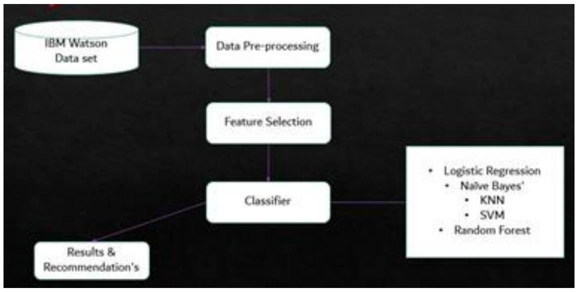

Fig 1 . Proposed Model

4.4. Data Pre-processing

From the IBM employee dataset we implement a feature selection method to select the most important features of the dataset and divide total dataset into two sub datasets. One is test dataset another one is training dataset. That is if suppose any feature value in the record contain any null value or undefined or irrelevant value then separate that entire record from the original dataset and place that record into training dataset, else if the record contain perfect data with all features then place that into test dataset. Test dataset contain all important features to predict employee attrition or employee attrition and training dataset contain irrelevant data..

4.4.1. Test dataset and training dataset:

Separating data into test datasets and training datasets is an important part of evaluating data mining models. By this separation of total data set into two data sets we can minimize the effects of data inconsistency and better understand the characteristics of the model. The test data set contains all the required data for data prediction and training data set contains all irrelevant data. Here we have 788 records in test dataset and 682 records in training dataset. We apply data classification and data prediction on the test dataset of 788 records.

4.4.2. Data Visualization

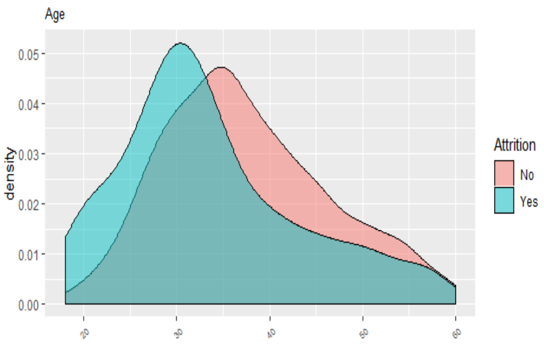

Fig2. Age V/s Attrition

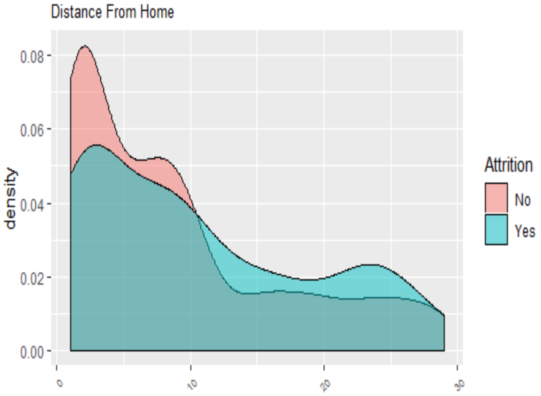

Fig3. Distance from home V/s Attrition

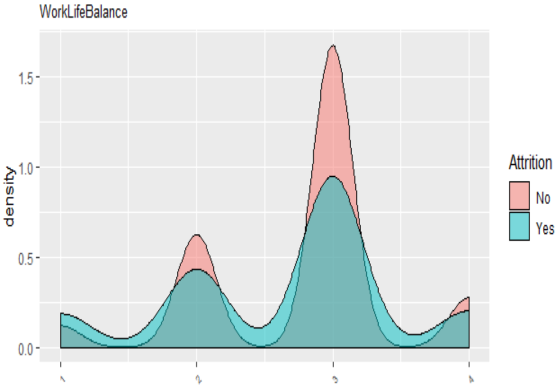

Fig4. Work Life Balance V/s Attrition

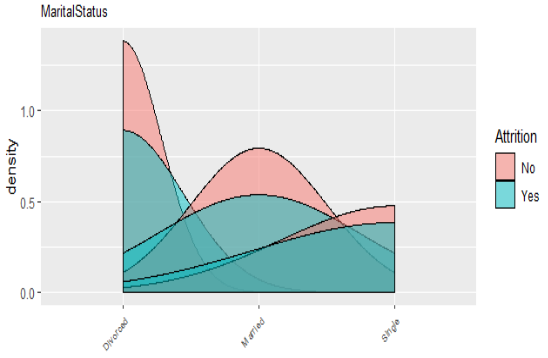

Fig5. Marital Status V/s Attrition

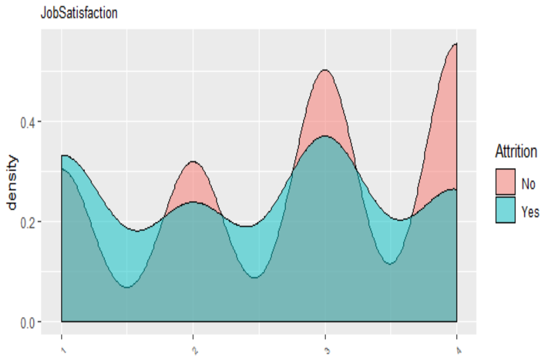

Fig6. Job satisfaction V/s Attrition

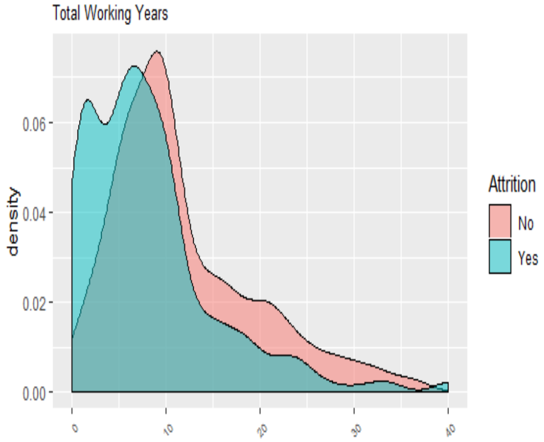

Fig7. Total Working Years V/s Attrition

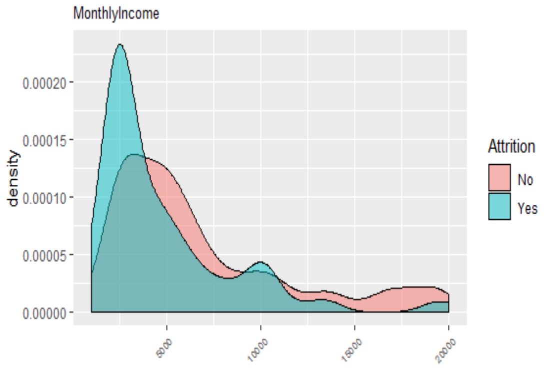

Fig8. Monthly Income V/s Attrition

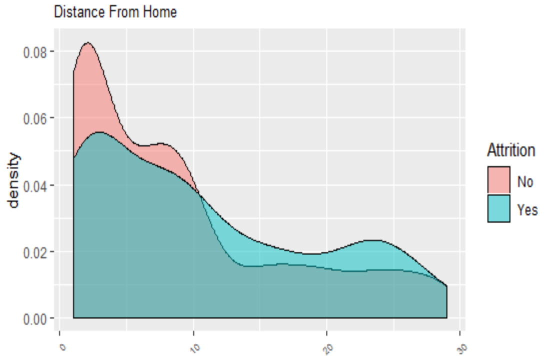

Fig9. Distance From Home V/s Attrition

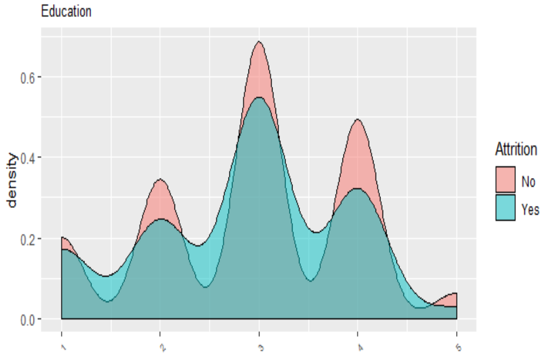

Fig10. Education V/s Attrition

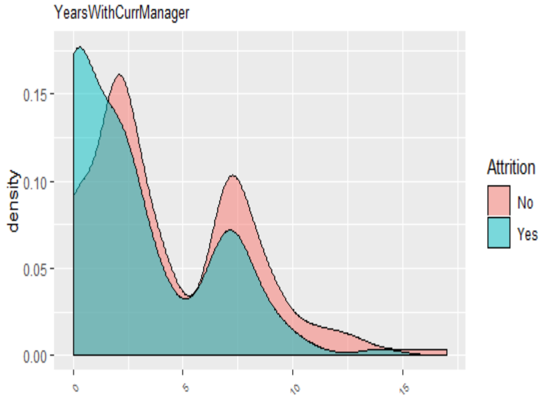

Fig 11. Years With Current Manager V/s Attrition

5.Results

Model	Accuracy	Precision	Recall	F1-Score	Auc
Logistic Regression	0.8798	0.450704	0.695652	0.547009	0.70643
Naïve Bayes	0.839	0	NA	NA	0.5
Support Vector Machine	0.8844	0.450704	0.727273	0.556522	0.70914
K-Nearest Neighbour	0.8322	0.070423	0.384615	0.119048	0.5244
Rando	0.8503	0.183099	0.6190	0.28260	0.58074

m **Forest**			48	9	

Table 2. Results Of Different algorithms

6.Conclusion

We have built a very simple models for predicting employee attrition, from some basic Exploratory Data Analysis to feature engineering as well as implementing learning models in the form of a Random Forest returns an 85% accuracy in its predictions. The fundamental general explanation for attrition is in all likelihood the effort-reward imbalance. For this situation, this for the most part applies to individuals who are staying at work longer than required and who much of the time have a generally low pay - it ought to be checked whether there is a compelling extra time strategy in our organization. We have additionally discovered that various features of work-life balance may speak to an issue for our representatives (a discovering bolstered by perceptions and (in any event somewhat) our best calculation). The way that each of the three factors connected (straightforwardly or in a roundabout way) to work-life balance (distance from home, business travel, and work-life balance all things considered) have their place among the best 20 factors could likewise be an indication that something ought to be done around there. If we take our "test" set as an example of IBM's current workforce, we can say that the job role with highest probability of attrition is sales representative - something should be definitely done about that, and we could explore further what exactly.

7 References

[1] Nagadevara, Vishnuprasad, Vasanthi Srinivasan, and Reimara Valk. "Establishing a link between employee turnover and withdrawal behaviours: Application of data mining techniques." *Research & Practice in Human Resource Management* 16.2 (2008).

[2] Kane-Sellers, Marjorie Laura. *Predictive models of employee voluntary turnover in a North American professional sales force using data-mining analysis*. Texas A&M University, 2007.

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

[4] Khare, Rupesh, et al. "Employee attrition risk assessment using logistic regression analysis." *IIMA International Conference on Advanced Data Analytics, Business Analytics*. 2015.

[5] Singh, Moninder, et al. "An analytics approach for proactively combating voluntary attrition of employees." *2012 IEEE 12th International Conference on Data Mining Workshops*. IEEE, 2012.

[6] Al-Radaideh, Qasem A., and Eman Al Nagi. "Using data mining techniques to build a classification model for predicting employees performance." *International Journal of Advanced Computer Science and Applications* 3.2 (2012).

[7] Ali, Omar, and Nur Zuhan Munauwarah. "Factors affecting employee turnover in organization/Nur Zuhan Munauwarah Omar Ali." (2017).

[8] Frederiksen, Anders. "Job Satisfaction and Employee Turnover: A firm-level perspective." *German Journal of Human Resource Management* 31.2 (2017): 132-161.

[9] Ongori, Henry. "A review of the literature on employee turnover." (2007).

[10] Saradhi, V. Vijaya, and Girish Keshav Palshikar. "Employee churn prediction." *Expert Systems with Applications* 38.3 (2011): 1999-2006.

[11] Alizadeh, Hossein, and B. Minaei Bidgoli. "Introducing A Hybrid Data Mining Model to Evaluate Customer Loyalty." *Engineering, Technology & Applied Science Research* 6.6 (2016): 1235-1240.

[12] Devi, P. Saranya, and B. Umadevi. "A Novel Approach to Control the Employee’s Attrition Rate of an Organization." (2018).

[13] Ajit, Pankaj. "Prediction of employee turnover in organizations using machine learning algorithms." *algorithms* 4.5 (2016): C5.

[14] Verbeke, Wouter, et al. "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach." *European Journal of Operational Research* 218.1 (2012): 211-229.

[15] Coussement, Kristof, and Dirk Van den Poel. "Integrating the voice of customers through call center emails into a decision support system for churn prediction." *Information & Management* 45.3 (2008): 164-174.

[16] Wei, Chih-Ping, and I-Tang Chiu. "Turning telecommunications call details to churn prediction: a data mining approach." *Expert systems with applications* 23.2 (2002): 103-112.

[17] Coussement, Kristof, and Dirk Van den Poel. "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques." *Expert systems with applications* 34.1 (2008): 313-327.

[18] Burez, Jonathan, and Dirk Van den Poel. "Handling class imbalance in customer churn prediction." *Expert Systems with Applications* 36.3 (2009): 4626-4636.

[19] Tsai, Chih-Fong, and Mao-Yuan Chen. "Variable selection by association rules for customer churn prediction of multimedia on demand." *Expert Systems with Applications* 37.3 (2010): 2006-2015.

[20] Coussement, Kristof, Dries F. Benoit, and Dirk Van den Poel. "Improved marketing decision making in a customer churn prediction context using generalized additive models." *Expert Systems with Applications* 37.3 (2010): 2132-2143.

[21] Saradhi, V. Vijaya, and Girish Keshav Palshikar. "Employee churn prediction." *Expert Systems with Applications* 38.3 (2011): 1999-2006.

[22] Kane-Sellers, Marjorie Laura. *Predictive models of employee voluntary turnover in a North American professional sales force using data-mining analysis*. Texas A&M University, 2007.