Journal of Informatics Education and Research
ISSN: 1526-4726
https://doi.org/10.52783/jier.v3i2.432
Vol 3 Issue 2 (2023)

Big data-based framework for prediction of employee attrition by using deep data people analytics

Dr. J. Varaprasad Reddy

Designation Professor, Department of MBA, TKR Institute of Management & Science,
Meerpet, Hyderabad-500097, Telangana, India.
Email id: jvpreddy2005@gmail.com

Dr. Sanjay Kumar Taurani

Associate Professor, Department of MBA, TKR Institute of Management & Science, Meerpet, Hyderabad-500097, India.Email id: sanjay.taurani@gmail.com

Dr. A Chandrashekhar

Associate Professor, Department of Mechanical Engineering, Faculty of Science &Technology (IcfaiTech), Icfai Foundation for Higher Education, Hyderabad, Telangana-501203, India.
Mail ID: acshekhar@ifheindia.org

Dachepally Shravya

MBA IInd year, Department of MBA, TKR Institute of Management & Science,
Meerpet, Hyderabad,Telangana-500097,
Mail ID: saishravs24@gmail.com

Consulting Author Email Id: acshekhar@ifheindia.org

Abstract:The phenomenon of Employee turnover has a negative impact on profit management accuracy across many industries within the business sector. The utilisation of contemporary advanced computing technology enables the development of a predictive model for staff attrition, hence facilitating cost savings for business owners. Although there is a lack of evaluation of these models in real-world scenarios, various implementations were created and utilisedin the IBM HR Employee Attrition list to examine the potential integration of these models they are put into a decision support system and how they affect making strategy decisions. In this research, a neural network based on Transformers was utilised as a computational method to analyse staff turnover. The network was distinguished by its ability to adapt contextual embeddings to tabular data. The experimental findings provided evidence that this particular model had superior performance compared to other contemporary models in terms of predictive accuracy. Moreover, the present research has revealed that deep learning, specifically Transformer-based networks, exhibit considerable potential in addressing the challenges associated with the presence of tabular and imbalanced data.

Keywords: Data science, machine learning, AI, predicting employee turnover, deep learning

I.INTRODUCTION

Staff turnover, alternatively referred to as employee attrition, is a typical phenomenon wherein personnel resign from their positions due to a multitude of factors. There are numerous motives for which an employee might resign. Most businesses struggle to hire employees and require an extended period of time to do so. Nonetheless, employees are free to resign from their positions at any moment. When employees depart, businesses are left with a substantial amount of work. The rate of employee turnover serves as a critical indicator of the company's growth trajectory. A significant number of individuals resign from their positions, as indicated by a high attrition rate. The elevated rate, however, has the potential to adversely impact the organisational structure and necessitate managerial intervention. depart. Therefore, for operations to continue as usual, managers must monitor the number of departing employees. A study by Apollo Technical found that approximately 19% of employees in numerous industries quit their occupations annually (Raza et al., 2022). According to a report by Raza et al. (2022), the Bureau of LabourStatistics in the United States estimated that

Journal of Informatics Education and Research
ISSN: 1526-4726
https://doi.org/10.52783/jier.v3i2.432
Vol 3 Issue 2 (2023)
employee turnover surpassed 57.0% in 2021. In order for an organisation to operate efficiently, employee retention should be at least 90%, while attrition should be kept below 10%.

The primary focus of this work pertains to two distinct dimensions, namely functional and data. The objective of this study is to evaluate, contrast, and select the most precise prediction model for the early detection of functional employee attrition. Additionally, we intend to investigate the phenomenon of positive attrition to determine its origins and help HR managers create retention strategies. The proposed solution employs deep data as an alternative to large data in order to mitigate certain data-related challenges that organisations may encounter while utilising HR analytics.

The term "big data" encompasses substantial quantities consisting of structured or unstructured data that are differentiated by their quantity, speed, and diversity. The term volume denotes the amount of data generated by a certain system or process sensors, social media, commercial transactions, etc. Variety refers to data forms, while velocity is data creation pace. Over the past decade, organisations have adopted the utilisation of Frameworks for data-driven strategic decision making and big data analytics has become increasingly prevalent within HR departments [1]. The absence of factual data hinders HR analytics deployment. Insufficient empirical data, such as a lack of features, candidates, or samples, can make it difficult to create a reliable model. Thus, before applying HR analytics, businesses must address data availability and generate large amounts of data.

Organisations must buy expensive cloud-based solutions for large-scale storage. limited companies may also lack This individual possesses the ability to work with high-quality human resources (HR) data and has developed analytical skills to effectively use big data-specific approaches in fields characterised by constrained data volumes. The primary concern in this instance pertains to the quality of the data. Human resource managers may not necessarily require all of the available data collected for HR analytics, therefore organisations must know what they need. From this perspective, data's worth matters more than its bulk.

For some firms, identifying deep data—high-quality information that forecasts trends—is a significant obstacle for HR analytics. The objective of our research is to transition from big data to deep data and segregate the vast quantities of data by eliminating superfluous or redundant data.

II.RELATED WORKS

Numerous models for predicting Voluntary and attrition of personnel have been documented in the academic literature. This paper primarily examines the utilisation of machine and deep learning models in current research endeavours pertaining to simulated HR datasets from IBM and Kaggle[2]. The rationale for this approach is supported by the presence of empirical evidence regarding the accuracy of prediction models on these publicly accessible datasets, which may be used as a basis for comparing our proposed models.

The IBM HR simulated dataset comprises 1470 samples and 34 input features, constituting a dataset of moderate size. The aforementioned attributes comprise the following: Age, Education, Education Field, Department, distance from Home, Percent Salary Increase, Monthly Income, Monthly Rate, Number of Companies Worked, Over18, Over Time, Environment Satisfaction, Gender, Hourly Rate, Job Involvement, Job Level, Job Role, Job Satisfaction, and Monthly Income. The dataset is supplying Kaggle has provided access to a considerable dataset known as the Kaggle HR dataset comprising 15,000 samples. It includes the goal variable "left" and nine accompanying characteristics [3]. The variables of interest in this research include the sales and salary, level of satisfaction, the most recent evaluation, the number of projects, the average monthly hours, the length of time spent with the organisation, work-related incidents, and promotions over the past five years data.While these systems presented precise prediction models to anticipate staff attrition, they were met with two primary criticisms:

1) There are no comprehensive investigations of employee qualities chosen and utilised to forecast attrition that justify the features' selection.

2) The primary emphasis is frequently placed on forecasting employee attrition; however, it is imperative for effective human resource management to not only proactively anticipate an employee's inclination to depart expeditiously, but also to comprehensively analyse and elucidate the underlying reasons behind this intention to leave.

Journal of Informatics Education and Research
ISSN: 1526-4726
https://doi.org/10.52783/jier.v3i2.432
Vol 3 Issue 2 (2023)

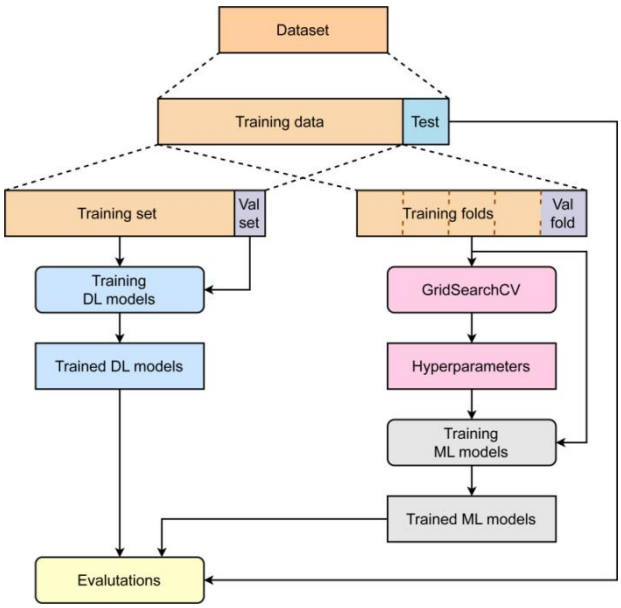

Fig.1.1 denotes framework for the research.

The key framework utilised in our research is succinctly presented in Figure 1. To commence, the acquired dataset was divided into two distinct subsets: the training data, which accounted for 85% of the dataset, and the test data, which constituted the remaining 15%. The utilisation of training data is essential for the optimisation and growth of models, but test data is deliberately kept distinct to prevent any potential data leakage. In the process of training deep learning models, a standard practise is to allocate 15% of the overall training data for the purpose of generating a validation set[4]. The remaining portion of the data is then utilised for the development of the model. The inclusion of a validation set is crucial for the evaluation and selection of optimal models. Without producing an additional validation set, the machine learning model is trained using the training data. To determine the most advantageous hyperparameters, the training data are subjected to a 5-fold cross-validation procedure to compute the mean performance associated with distinct parameter configurations. Subsequently, the machine learning models undergo retraining using updated training data and hyperparameters. [5]Ultimately, the independent test set is employed to evaluate and compare the performance of the deep learning and machine learning models.

Modeling of a mixed method for employee attrition.

The inclusion of voluntary turnover in the modeling of employee attrition is of utmost importance in the attrition prediction process, given the inherent inevitability of this phenomena.Furthermore, it is imperative that we embrace a research technique that enables us to effectively compare and contrast theoretical models with experimental data in order to successfully implement a data-driven approach. Therefore, we propose utilising a hybrid research methodology that integrates an exploratory investigation with a quantitative approach to comprehensively understand and clarify the occurrence of employee attrition.In a sequential manner, these two methods are implemented in combination, where the results obtained from one method are utilised to inform the implementation of the other method[6]. Hence, through the integration of the respective merits and limitations of exploratory and quantitative methodologies, this approach has the potential to yield a more holistic comprehension of a given phenomenon compared to the individual utilisation of either method.

In order to get a more comprehensive comprehension regarding high attrition phenomena and their underlying causes, it is imperative to conduct a thorough analysis and investigation to identify its root causes, an initial exploratory research is conducted by thoroughly examining the available literature. This includes reviewing the corpus comprises scholarly studies, academic publications, and publicly available datasets that have been contributed by professionals in the field of human resources and academics [7]. In addition, a comparison is made between the collected features to the causal factors for attrition that were found by The utilisation of a questionnaire and feature selection methods constitutes a

Journal of Informatics Education and Research
ISSN: 1526-4726
https://doi.org/10.52783/jier.v3i2.432
Vol 3 Issue 2 (2023)
quantitative approach to research. The diagram in Figure 1 depicts the framework of the research approach utilised in this study. The following sections will outline the separate phases of the suggested hybrid approach.

III.RESEARCHMETHODOLOGY

The collection of features

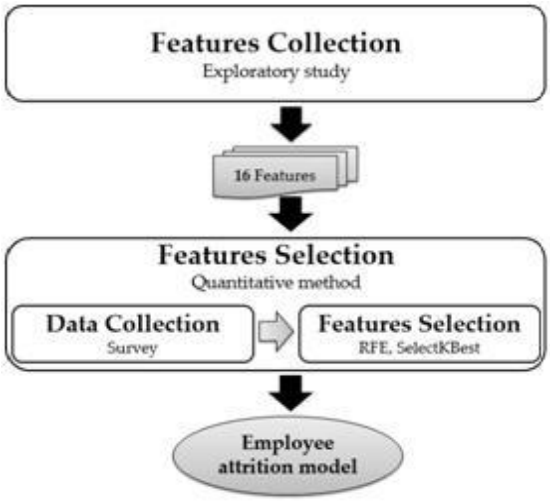

Fig.1.2 denotes mixed method of employee attrition modeling.

During the preliminary stage of our research, we identified and gathered employee characteristics that are pertinent to our inquiry. This stage entails undertaking an exploratory investigation into the determinants of employee attrition, utilising secondary research to conduct an exhaustive review of the pertinent literature. Utilising an exploratory research strategy to identify and collect pertinent characteristics for our research problem constituted the subsequent phase[8]. The selection of these elements was predicated on their regular occurrence in pertinent literature and prior research. This exploratory study investigates a variety of HR management research studies, experiments, and simulated HR datasets, including the simulated datasets and the fictitious dataset created by IBM data scientists.

Features	XGB	RF	DT	LR	SVM	Voteresults
Age	False	True	True	False	False	Eliminate
Education	False	False	False	False	False	Eliminate
Gender	False	False	False	True	True	Eliminate
Maritalstatus	True	False	False	True	True	Keep
Business Travel	True	True	True	False	False	Keep
Job satisfaction	True	True	True	True	True	Keep
Job involvement	True	False	False	True	True	Keep
Job	True	False	False	True	True	Keep

Journal of Informatics Education and Research
ISSN: 1526-4726
https://doi.org/10.52783/jier.v3i2.432
Vol 3 Issue 2 (2023)

performance						
RelationshipSatisfaction	False	False	True	False	False	Eliminate
Environment Satisfaction	True	False	True	True	True	Keep
Tenure	False	True	True	False	False	Eliminate
Promotability	False	True	False	False	False	Eliminate
Grade	True	True	False	False	False	Eliminate
Training	True	True	True	True	True	Keep
Rewards	True	True	True	False	False	Keep
Worklife/balance	False	False	False	True	True	Eliminate

Table.1.1 denotes Vote on whether to maintain or remove the collected features.

A combination matrix of the two feature selection outcomes methods is illustrated in Figure 2.However, we suggest retaining features that are selected by SelectKBestare removed in an RFE. Furthermore, characteristics that are neither eliminated by RFE nor selected by SelectKBest are retained in an equivalent manner. Ultimately, those features that fail the RFE and SelectKBest elimination processes are eliminated. As a result, the subsequent attributes were eliminated: work-life balance, education, proliferability, and relationship satisfaction.

An approach to feature selection known as Recursive Feature Elimination (RFE) gets rid of the weaker feature and fits a model with the help of the coefficients of a linear model are an example of an outside estimate that gives features weights[9]. The model's coefficients, or importance traits, are used to rank the features. Recursive feature elimination (RFE) targets dependencies and collinearity that may be present in the model by getting rid of a small group of traits each time through a loop. During the context of predictive models and algorithms, an attribute is deemed to be unnecessary Remove the attribute from the data fields if the model assigns it the value False. The attribute is valid if and only if the model assigns it the value True. In this stage, we used five widely recognised and reliable models. This list includes SVM, XGB, RF, DT, and LR[10]. Given that the objective of this study is to classify employee attrition prediction, the following classifiers were selected as they most accurately represent the various classification approaches and frequently perform well when applied to statistical data. Subsequently, the features used by the RFE algorithm were decided by a referendum of those involved. The gender, age, grade, education, tenure, relationship happiness, and work-life balance attributes all have more false values than genuine ones and Education—were chosen to be eliminated by RFE.

IV.RESULTS AND DISCUSSION

In addition to its useful constraints, further development of a transformer-based model (henceforth Transformer) is necessary. While Transformer-based models have demonstrated considerable efficacy in a variety of NLP applications, they do have some drawbacks when working with small datasets. Overfitting happens when the models remember the data instead of learning from it. This is more likely to happen with transformer models that were trained on insufficient data. The lack of variety in small datasets can make it harder for a model to generalise and process data it hasn't seen before. Furthermore, Transformer-based models have a big problem when they try to figure out the meanings and connections of sparse datasets because they don't include enough contextual factors. Lack of data is also a problem because trends and words that are rare or don't exist may be harder to learn.

Journal of Informatics Education and Research
ISSN: 1526-4726
https://doi.org/10.52783/jier.v3i2.432
Vol 3 Issue 2 (2023)
In conclusion, the ability of Transformer-based models to capture complex relationships may be diminished when applied to extremely small datasets. Transfer learning, data augmentation, regularisation techniques, and domain adaptation are all examples of mitigating tactics. Even if these techniques help alleviate some of the downsides, it is still important to be aware of the challenges that come up when trying to train comprehensive models with limited datasets. Data augmentation increases the diversity of training data, whereas transfer learning facilitates the application of insights obtained from different domains or activities. Overfitting can be avoided by regularisation methods, and representations can be better adapted to new domains via domain adaption. These tactics improve the efficacy, generalizability, and adaptability of the Transformer-based approach.

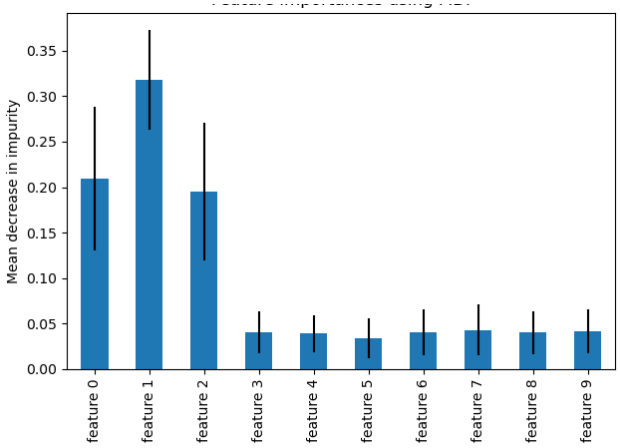

Fig.1.3 denotes results of applying features importance on our random forest model.

There are two distinct ways in which the generalisation mistake can be changed by the number of randomly chosen features. This is because choosing a lot of features makes the individual trees, whereas reducing the number of features weakens the correlation between the trees, thereby strengthening the forest as a collective unit as shown in figure.

Following the completion of an exploratory and in-depth data analysis, as well as the subsequent identification of all model with the parameters and hyper-parameters in place, we are now equipped to move on to the process of developing our models and evaluating how well they function. In point of fact, we are going to discuss the results of our work with machine learning, ensemble learning, and deep learning prediction models in this section.Our small-sized HR real dataset (450 samples), the medium-sized IBM HR simulated dataset (1470 samples), and the large-sized Kaggle HR simulated dataset (15000 samples) are as utilised in order to perform the most accurate assessment possible of the performance of these prediction models in a variety of contexts. In conclusion, but certainly not least, the most essential takeaway from these models will be offered at the end of the trial to assist the HR manager in not only predicting employee turnover but also comprehending the reasons behind it and, as a result, locating key shops for employee retention.In the parts that follow, we will go over how to evaluate these models and compare the outcomes of each.

Predictive Models For Two Sample Sets of Mock Human Resources Data.

In this paper, we test the accuracy of four distinct predictive models by applying them to data taken from two simulated datasets pertaining to human resources. The first is a big dataset that was provided by Kaggle. It has 15,000 samples and nine characteristics, such as Salary, sales, company, work-related accidents, and promotions in the past five years. Other characteristics include satisfaction, most recent review, duration of employment, mean monthly work hours, and quantity of assignments. The second simulated HR analytics dataset that IBM has created is a medium-sized one, with a total of 1470 samples and 34 attributes. The attrition variable serves as the objective variable for this dataset, and it is encoded as "Yes" (the employee departed) or "No" (the employee did not depart). Since our 11 selected features are among IBM's 34 feature simulated dataset, we will test our predictors on the complete IBM dataset to determine how well they perform. After that, we will analyse them utilising the same information; however, this time we will just use the 11 characteristics that have proven to be the best predictors of employee turnover.

Journal of Informatics Education and Research
ISSN: 1526-4726
https://doi.org/10.52783/jier.v3i2.432
Vol 3 Issue 2 (2023)

V.CONCLUSIONS AND FUTURE DIRECTIONS

By utilising predictive analytics techniques, the primary objective of this study is to assist human resource managers in proactively identifying an employee's intention to depart, thereby resolving the issue of attrition. In three sections, the contributions can be succinctly summarised. The current research presents a suggested framework for employee attrition, consisting of eleven key factors that are considered sufficient for forecasting positive attrition and intention to depart. This prediction is achieved through the use of a mixed research approach.The dissertation proposes the utilisation of machine learning, deep learning, and ensemble learning predictive models, and assesses their efficacy across different scenarios, encompassing a substantial simulated dataset, a moderate simulated dataset, and a tiny real dataset.

When considering the constraints of a study, it is advisable to consider the dynamic aspects related to the emotional states and behaviour of employees, in order to investigate their possible impact on employee attrition.To accommodate the potential for continuous data and the dynamic character of the data, influx, it is advisable to carry out training for the predictive models in this particular scenario through online means. Furthermore, it is important to acknowledge that participants in our survey have put forth supplementary characteristics that deserve careful examination and could potentially lead to voluntary attrition. Consequently, these recommendations may be integrated into our forthcoming investigations. Indeed, recommendations have been put out to urge the organisation to consider health-related issues, ensure employment stability, and facilitate the adoption of innovative technologies. In conclusion, it can be inferred that next studies will encounter challenges in addressing the issue of unbalanced data, especially in the context of companies and enterprises characterised by a significant rate of staff attrition. This difficulty arises from the insufficiency of the predictive models now employed in handling such data.

REFERENCES

[1] A. Tursunbayeva, Di Lauro, and C.S. Pagliari, "Peopleanalytics: A scoping review of conceptual boundaries and value propositions," International Journal of Information Management, vol. 43, pp. 224-247, 2018.

[2] T.Pape,"Prioritizingdataitemsforbusinessanalytics:Frameworkandapplicationtohumanresources," European Journal of Operational Research, vol. 252(2), pp. 687-698, 2016.

[3] S. N. Mishra, D. R. Lama, and Y. Pal, "Human ResourcePredictive Analytics (HRPA) For HR Management InOrganizations," International Journal of Scientific & Technology Research, vol. 5, no. 05, pp. 33-35, 2016.

[4] P. Likhitkar and P. Verma, "HR Value Proposition UsingPredictiveAnalytics:AnOverview," in New Paradigm DecisionScience and Management, Singapore, 2020, pp. 165-171, doi:10.1007/978-981-13-9330-3_15.

[5] T. Peeters, J. Paauwe, and K. Van De Voorde, "People analytics effectiveness: developing a framework," Journal of Organisational Effectiveness: People and Performance, vol. 7, no. 2, pp. 203-219, July 2020, doi:10.1108/JOEPP-04-2020-0071.

[6] N. Shah, Z. Irani, and A. M. Sharif, "Big data in an HR context: Exploring organisational change readiness, employee attitudes and behaviours," Journal of Business Research, vol. 70, Jan. 2017, pp. 366-378, doi: 10.1016/j.jbusres.2016.08.010.

[7] S. V. Kalinin, B. G. Sumpter, and R. K. Archibald, "Big-deep-smart data in imaging for guiding materials design," NatureMaterials, vol. 14, no. 10, October 2015, pp. 973-980, doi:10.1038/nmat4395.

[8] "Big Data and Human Resources Management: The Rise of Talent Analytics," M. Nocker and V. Sena, Social Sciences, vol.8, no. 10, p. 273, Sep. 2019, doi: 10.3390/socsci8100273.

[9] Suriadi, S., & Abran, A. (2015). A systematic literature review of software project risk management in global software development: Implications for guiding global software project. Information and Software Technology, 57, 120-138.

[10] S. S. Alduayj and K. Rajpoot, "Predicting Employee Attrition Using Machine Learning," 2018 International Conference on Innovations in Information Technology (IIT), Al Ain, November 2018, pp.93-98, doi:10.1109/INNOVATIONS.2018.8605976.