PAGES: 234-241
9/19/23
JOURNAL OF SCIENTIFIC RESEARCH AND TECHNOLOGY(JSRT) VOLUME-1 ISSUE-6 SEPTEMBER
Registered under MSME Government of India ISSN: 2583-8660

A Novel Approach To Unveiling Employee Attrition Patterns using Machine Learning Algorithms

[1]Dr. Raafiya Gulmeher, [2] Umama Aiman

[1] Assistant professor, CSE Department, KBN University, Kalaburagi, India
profraafiya.cse@gmail.com

[2]M.Tech Student, CSE Department, KBN University, Kalaburagi, India
umamaaiman7@gmail.com

ABSTRACT

The negative effects of employee turnover upon productivity at work as well as long-term growth initiatives make it a top issue for businesses. To combat this issue, businesses are increasingly relying on machine learning tools for accurate turnover forecasting and management. In this study, we set out to create a model that can accurately forecast future rates of employee turnover. In this research, we use HR analytics data from the Kaggle platform to predict outcomes using a variety of different machine learning techniques, including the Random Forest, Logistic Regressor, Gradient Boosting Classifier, CatBoost Classifier, Extreme Gradient Boosting, and Light GBM. This research goes beyond simple forecasting to investigate the many elements outside of the workplace that contribute to employee turnover. Moreover, our findings aim to provide top management with an insightful perspective, empowering informed decisions concerning strategies for workforce retention. Looking ahead, future research could refine the analysis by encompassing additional factors. Factors such as feedback, recognition, hiring procedures, and organizational culture, which have been observed to positively influence employee attrition rates, hold promise in offering a more comprehensive understanding and effective mitigation strategies.

Keywords- Employee Attrition, Gradient Boosting Classifier, Machine Learning, Random Forest Regressor.

I. INTRODUCTION

Employee Attrition, also recognized as Employee Turnover, stands as a foundational concern prevailing within today's industries. This issue holds substantial gravity across most companies, warranting serious attention. Attrition denotes " a decrease in available personnel as a result of attrition (people leaving their jobs, retiring, or dying)." The term attrition encompasses various definitions, with this study primarily focusing on two vital aspects: employee departures and retirements from an organization.The ongoing occurrence of employee attrition remains a persistent concern for Human Resources departments. Over time, the incidence of employee turnover has demonstrated an upward trajectory. Consequently, employers face the challenge of deciphering whether employee departures stem from dissatisfaction or alternative motivations. Prior to implementing radical measures, a judicious approach involves probing the underlying reasons for the issue. In the contemporary professional landscape, employees display an unprecedented willingness to transition between organizations in pursuit of more favorable prospects. As a result, employee turnover has evolved into a critical challenge for a majority of organizational structures.

Employee attrition happens when employees leave a company due to reasons like personal issues, not enjoying their job, getting paid less, or experiencing a negative work environment. It's divided into two types: voluntary, where employees leave by their choice, and involuntary, where managers ask employees to leave, often because of poor performance or business needs.Even though the employer wants them to remain, even the best workers may quit on their own will. They might leave because they found better opportunities or want to retire early. Voluntary attrition can happen when employees retire early or get job offers from other companies. Companies that care about their employees usually invest in them by providing good training and a positive workplace. But even these companies can still have employees leave on their own or lose valuable workers. Replacing employees who leave is expensive. It involves costs like interviewing new candidates, searching for the right fit, and training them. Companies must prioritize lowering turnover rates if they want to keep their edge in the market. Consequently, higher authorities should grasp the core reasons driving their

employees' desire to depart from the organization. Subsequently, they can take proactive measures to enhance various aspects, including workflow, productivity, and overall performance. This study's objective is to leverage machine learning techniques, specifically those within the realm of ML, for identifying primary aspects impacting worker's inclination to leave a company. Additionally, the aim is to predict the likelihood of specific employees departing from the organization. The fundamental focus of this research lies in exploring the application of the Random Forest and light GBM algorithms for forecasting employee attrition.

II. METHODOLOGY

Machine Learning Classification Algorithms:

This study encompassed the training and evaluation of six distinct machine learning algorithms: the decision tree model, extreme boosting model, cat boosting model, logistic regression model, and light GBM. In the subsequent sections, comprehensive overviews of the theoretical foundations of each of these models are provided.

Logistic Regression Model:

Purpose of logistic regression (LR) is to estimate model's variables using a logit (or logistic) binomial regression structure. It's common method for utilizing category target variables, especially in cases with binary outcomes like yes/no, win/lose, or leave/stay, as shown in IBM attrition data. Supplied formula gives precise definition of basic logistic function:

$$Y = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$

Here, Y symbolizes dependent variable, while x denotes independent variables or characteristics.

Decision Tree Models:

Decision trees (DT) stand out as robust algorithms with the capacity to effectively fit intricate datasets. Their application extends across a wide spectrum of tasks, including medical diagnoses and assessing credit risk in loan applications. The process of decision tree learning involves approximating a target function, which takes the form of a tree comprising "if-then" rules. This representation enhances human understanding. The process entails breaking down data in progressively small subsets, initiated by uppermost node known as the "root." Subsequently, an interconnected decision tree evolves incrementally. The final iteration of the decision tree consists of two key node types: decision nodes and leaf nodes. Importantly, these nodes possess the versatility to manage both categorical and numerical data, further enhancing the model's adaptability. You can see an example of decision tree that makes use of qualities that have highly connected with outcome variable (attrition) in Fig-1. Feature values for every instance allow us to distinguish between two categories, "Leave" & "Stay," as shown.

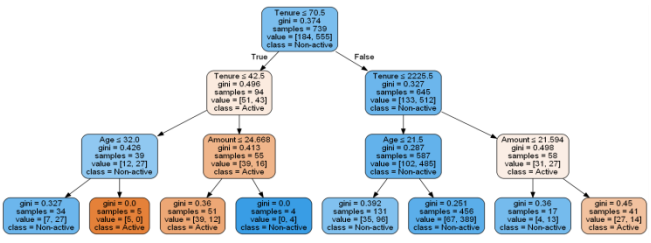

Fig 1. A decision tree made with extracted features from IBM HR dataset

Random Forest Model:

The Random Forest model represents a pivotal advancement in the field of machine learning and predictive analytics. Operating as an ensemble learning technique, it excels in generating accurate and stable predictions by aggregating the insights of multiple decision trees. This approach mitigates the individual limitations of trees and enhances the model's resilience to overfitting, making it particularly well-suited for complex and high-dimensional datasets. RF model constructs a collection of decision trees through bootstrap sampling and random feature selection, leading to diverse and independently trained trees. The final prediction is then determined through the combined output of these trees, often employing methods such as averaging for regression tasks and voting for classification tasks. This model offers numerous benefits, including its capacity

to capture intricate relationships within data, manage noisy inputs, and furnish feature importance assessments. However, while the model is known for its robustness, successful application necessitates thoughtful hyper-parameter tuning to optimize its performance.

Gradient Boosting Model:

The Gradient Boosting model is a robust and widely adopted machine learning technique that has significantly contributed to the advancement of predictive modeling and data analysis. Functioning as an ensemble learning algorithm, it excels in improving prediction accuracy by aggregating assets of multiple feeble apprentices, usually decision trees, in unified & more powerful model. The model's core principle lies in its iterative approach, wherein it sequentially constructs a series of models, each addressing the deficiencies of its predecessors by learning from the residual errors. This process iteratively refines the model's predictions, gradually enhancing its performance. The incorporation of gradient descent optimization enables the model to iteratively minimize a defined loss function, while regularization techniques like shrinkage and tree depth constraints prevent overfitting. Renowned for its capability to handle complex relationships within data, the Gradient Boosting model has found applications across diverse domains such as finance, healthcare, and natural language processing. Its effectiveness, however, necessitates careful hyper-parameter tuning and cross-validation to unleash its full potential.

Extreme Boosting Model:

The Extreme Gradient Boosting (XGBoost) model has emerged as a powerful and widely used ensemble learning technique that excels in a multitude of predictive modeling tasks. Built upon the foundation of gradient boosting, XGBoost enhances both accuracy and efficiency through innovative algorithmic advancements. This model optimally combines the strengths of decision trees and gradient boosting by implementing parallelized tree construction and a regularized boosting framework. The integration of these features enables XGBoost to effectively handle complex relationships within data, exhibit robustness against overfitting, and provide remarkable predictive performance. Its versatility is showcased across various domains, encompassing finance, healthcare, natural language processing, and beyond. As this research paper delves into the exploration of predictive modeling techniques, a comprehensive examination of the Extreme Gradient Boosting model is poised to illuminate its pivotal role in advancing the state of the art in machine learning and data analysis.

Cat Boost Model:

The CatBoost model stands out as a notable innovation in the realm of gradient boosting algorithms, offering compelling advantages that position it favorably in comparison to other state-of-the-art models. Its distinctive feature lies in its ability to seamlessly handle categorical features without the need for extensive preprocessing, a characteristic that differentiates it from many other traditional machine learning approaches. By leveraging techniques such as ordered boosting and oblivious trees, CatBoost effectively tackles challenges posed by high-cardinality categorical variables, resulting in more accurate and reliable predictions. Notably, its built-in mechanisms for handling missing values and addressing class imbalances contribute to its robust performance in real-world applications. Through comprehensive empirical evaluations, CatBoost consistently demonstrates superior predictive accuracy, while also mitigating issues related to overfitting and the demand for intricate feature engineering often encountered in other methods. As this research paper elucidates, CatBoost's adeptness in predictive tasks, coupled with its streamlined data handling capabilities, positions it as a compelling choice for practitioners seeking a powerful and efficient tool for diverse predictive modeling endeavors.

Light GBM:

In our research paper focusing on predicting worker attrition utilising ML models, incorporation of Light GBM emerges as a powerful choice due to its exceptional predictive prowess, especially when compared to other algorithms. Light GBM, a cutting-edge gradient boosting framework, offers distinct advantages in terms of accuracy and efficiency. Its unique features, such as histogram-based binning and the leaf-wise growth strategy, enable the model to effectively capture intricate patterns within the data, resulting in highly accurate predictions. What sets Light GBM apart is its ability to handle large datasets and its optimized computation process, which collectively contribute to its superior performance in comparison to traditional algorithms. As we evaluate various algorithms for attrition prediction, the evidence suggests that Light GBM consistently demonstrates heightened accuracy and efficiency, positioning it as a potent tool for anticipating employee attrition and enabling organizations to take proactive measures for retention.

A. Experiment Tools

Jupyter Notebook, an interactive computing environment accessible via the web, in tandem with IBM Cloud, proved essential for tasks encompassing data acquisition, data cleansing, preprocessing, algorithm application, model evaluation, and the computation of accuracy measures.

B. Attrition Analysis Steps

1) Data Collection: Initiating analytical process involves compilation of pertinent data from diverse origins. The dataset utilized for this experimental study originates from Kaggle, encompassing a total of 1470 samples and incorporating 35 distinctive features.

As can be seen in Fig. 2, every single attribute of the dataset is connected to some aspect of the professional or personal lives of the workers.

```
Age
Attrition
BusinessTravel
DailyRate
Department
DistanceFromHome
Education
EducationField
EmployeeCount
EmployeeNumber
EnvironmentSatisfaction
Gender
HourlyRate
JobInvolvement
JobLevel
JobRole
JobSatisfaction
MaritalStatus
MonthlyIncome
MonthlyRate
NumCompaniesWorked
Over18
OverTime
PercentSalaryHike
PerformanceRating
RelationshipSatisfaction
StandardHours
StockOptionLevel
TotalWorkingYears
TrainingTimesLastYear
WorkLifeBalance
YearsAtCompany
YearsInCurrentRole
YearsSinceLastPromotion
YearsWithCurrManager
dtype: int64
```

Fig 2: Dataset Characteristics

Variable "Attrition" is categorized in nature, wherein "No" signifies worker who would continue their tenure with the company, while "Yes" denotes an employee who has departed. Within the collection of 35 attributes or features, 8 of them are categorical variables stored as objects, while the remaining attributes assume numerical values & stowed as integers.

2) Data Analysis: It is a process aimed at extracting deeper insights from information. Its primary goal is to derive value even from seemingly irrelevant data. This encompassing endeavor involves pivotal stages such as Data Cleaning, Randomizing, & Visualizing. In the realm of Data Cleaning, tasks encompass the elimination of duplicates, rectification of errors, detection of missing values, normalization, and conversion of data types. Addressing missing values entails utilizing a straightforward imputer for numerical data, while for nominal variables, strategies like one-hot encoding or dummification are employed. The situation can escalate when multiple nominal columns contain numerous categories, rapidly increasing data dimensionality which is often infeasible for efficient machine learning. To mitigate this, a solution involves post one-hot encoding dimensionality reduction, ensuring the accurate representation of nominal column values. In the employee attrition dataset, certain variables that are initially presented as numerical, actually possess an inherent categorical ordering. Meanwhile, features like EmployeeNumber, which uniquely identifies employees, as well as EmployeeCount and StandardHours—both of which hold constant values—do not contribute meaningfully to our prediction model. As a result, these three features are excluded from consideration. Similarly, the categorical variable Over18, which indicates that all employees are above 18 years old, is removed from the dataset due to its lack of predictive utility.

At this juncture, we have computed rudimentary descriptive statistics separately for the numerical & categorical Characteristics. These statistics encompass key measurements including count, mean, standard

PAGES: 234-241
9/19/23
JOURNAL OF SCIENTIFIC RESEARCH AND TECHNOLOGY(JSRT) VOLUME-1 ISSUE-6 SEPTEMBER
Registered under MSME Government of India ISSN: 2583-8660

deviation (std), percentiles at 25%, 50%, and 75%, as well as the minimum and maximum values (min/max). This analysis provides valuable insights into the distribution and characteristics of the variables. The summarized results for these descriptive statistics, covering both numerical and categorical features, can be found in Figures 3.

	Age	DailyRate	DistanceFromHome	Education	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobS
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	14
mean	36.923810	802.485714	9.192517	2.912925	1024.865306	2.721769	65.891156	2.729932	2.063946	
std	9.135373	403.509100	8.106864	1.024165	602.024335	1.093082	20.329428	0.711561	1.106940	
min	18.000000	102.000000	1.000000	1.000000	1.000000	1.000000	30.000000	1.000000	1.000000	
25%	30.000000	465.000000	2.000000	2.000000	491.250000	2.000000	48.000000	2.000000	1.000000	
50%	36.000000	802.000000	7.000000	3.000000	1020.500000	3.000000	66.000000	3.000000	2.000000	
75%	43.000000	1157.000000	14.000000	4.000000	1555.750000	4.000000	83.750000	3.000000	3.000000	
max	60.000000	1499.000000	29.000000	5.000000	2068.000000	4.000000	100.000000	4.000000	5.000000	

8 rows x 21 columns

rmanceRating	RelationshipSatisfaction	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	YearsSinceLastPromotion
1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000	1470.000000
3.153741	2.712245	0.793878	11.279592	2.799320	2.761224	7.008163	2.187755
0.360824	1.081209	0.852077	7.780782	1.289271	0.706476	6.126525	3.222430
3.000000	1.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000
3.000000	2.000000	0.000000	6.000000	2.000000	2.000000	3.000000	0.000000
3.000000	3.000000	1.000000	10.000000	3.000000	3.000000	5.000000	1.000000
3.000000	4.000000	1.000000	15.000000	3.000000	3.000000	9.000000	3.000000
4.000000	4.000000	3.000000	40.000000	6.000000	4.000000	40.000000	15.000000

Fig 3: Statistics of description of the dataset

3) Data Visualization: Data visualization offers a visual roadmap to decode the dynamics of employee attrition. Through graphical representation, it transforms complex data into understandable insights, revealing trends and relationships that may otherwise go unnoticed. In the realm of employee attrition, data visualization facilitates the identification of key drivers, such as factors influencing turnover rates, departmental patterns, and correlations between job satisfaction and attrition. By translating data into meaningful visuals, organizations can gain a clearer understanding of attrition patterns and make informed decisions to enhance retention strategies and bolster employee engagement.

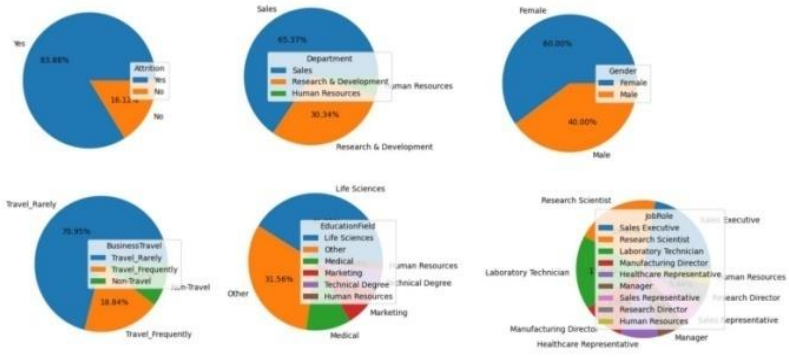

Fig 4: Data Distribution

4) Feature Engineering: During the analysis phase, it's possible to encounter various data anomalies, including null values, outliers, and erroneous entries. Null values or duplicated records can significantly undermine prediction accuracy by unintentionally biasing model training. Hence, it's crucial to eliminate these issues

before constructing a robust model. In this study, there were no instances of null or undefined values within any variables, and duplicate observations were also absent. To ensure optimal model preparation, pertinent data can be thoughtfully selected and refined through feature engineering. A noteworthy attribute of feature engineering is its potential to wield a substantial impact on outcomes, often surpassing the direct influence of the model itself.

In Figure 6, the depicted correlation matrix takes the form of a heat map, illustrating the interrelationships among all variables. Light blue areas denote negligible correlation, while the varying intensities of black and dark orange shades signify increasing correlation. In particular, black represents a direct or positive correlation, indicating that changes in one characteristic coincide with changes in another. Conversely, white represents an indirect or negative correlation, signifying an inverse relationship between characteristics. Upon scrutinizing the correlations within the heatmap, it becomes apparent that the listed distinctive features exhibit strong correlations, spanning from 0.7 to 1.

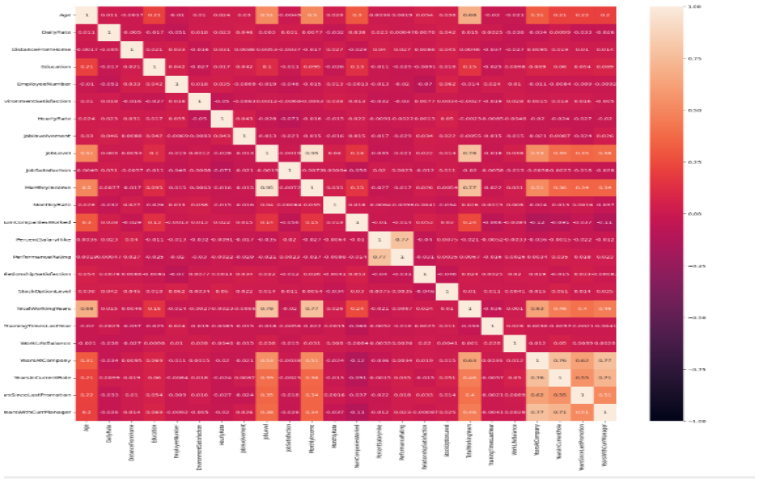

Fig 5: Correlation matrix(heat map)

Work experience, tenure at the firm, job title, time since promotion, and time under the present boss all have substantial correlations with one another. Pearson's correlation coefficients clearly show these relationships in the graph. There are significant relationships between several groups of variables, including employment status, monthly income, and total working hours, among others.

Age also has a modest linear relationship with monthly income and total years worked.

In addition, there is a positive association between factors like years since the previous promotion and years in the firm, indicating that lengthy stretches of time between promotions are associated with longer tenures.

All significant associations may be seen in the scatter plot of all continuous variables, which is the correlation plot. There are significant relationships between many of the variables we examined, including length of time with the company and length of time with the current manager, length of time in the current role and length of time with the current manager, salary and length of time in the workforce, age and length of time in the workforce, salary increase as a percentage, and performance evaluation.

Feature selection: The presence of trivial and unrelated features can harm the model's performance. As a result, assigning precedence to feature selection and data cleaning as the primary and pivotal stages in your model design process is imperative. In the field of machine learning, feature selection is a fundamental concept that holds remarkable sway over practical efficacy of algorithm under construction. Chosen features act as mentors for the model and wield a profound impact on its overall efficacy.

Feature Importance:characteristic significance ranks every information characteristic according to its significance in predicting outcome variable. Tree based classifiers naturally include this feature. To get most useful information out of the dataset, we'll use Extra Tree Classifier.

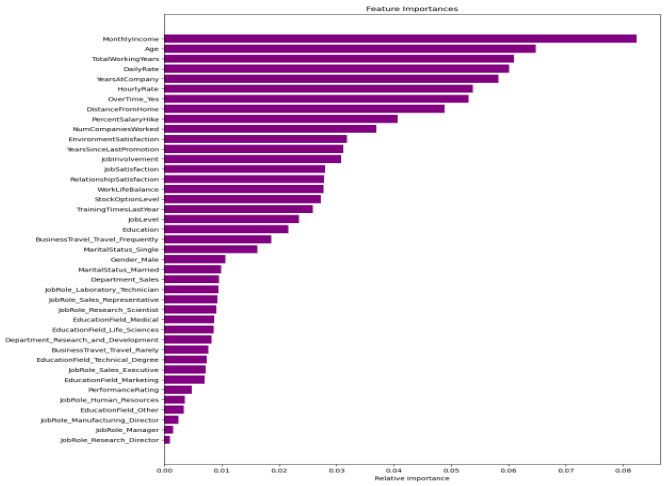

Fig 6: Feature Importance

5) Model Building: An important part of our study is development of model for forecasting staff turnover rates by use of machine learning algorithms. By harnessing the capabilities of advanced analytical techniques, we endeavor to develop a robust predictive framework that sheds light on workforce turnover dynamics. Beginning with the careful selection and preparation of relevant data attributes, we ensure the integrity and consistency of our dataset. Guided by the problem's complexity, we then judiciously choose a suitable machine learning algorithm. Through the utilization of historical data, we train selected algorithm to recognize intricate patterns linked to attrition outcomes. Rigorous evaluation using performance metrics like accuracy, recall, & F1-score enables us to gauge the model's efficacy, often necessitating fine-tuning to optimize its predictive power. The culmination of this process yields an adept model that not only identifies employees at risk of attrition but also facilitates the formulation of tailored retention strategies, thus offering a significant contribution to the discourse on enhancing workforce stability and organizational effectiveness.

III. RESULTS AND DISCUSSIONS

The ROC (Receiver Operating Characteristic) curve is an effective method for comparing different classification methods such as the Random Forest, Logistic Regressor, Gradient & Light Gaussian Boltzmann Machine. This illustration shows how changing the classification threshold affects both the true positive rate (sensitivity) and the false positive rate (1 - specificity).

The visual representation of a ROC curve in the context of an employee attrition project is as follows:
X-Axis (False Positive Rate): This axis measures the frequency at which the model incorrectly labels actual negative cases as positive. It illustrates instances where the model generates false alerts.
Y-Axis (True Positive Rate): Reflecting proportion of appropriately recognized actual positive cases with every positive case, this axis signifies model's ability to accurately identify employees at risk of attrition.

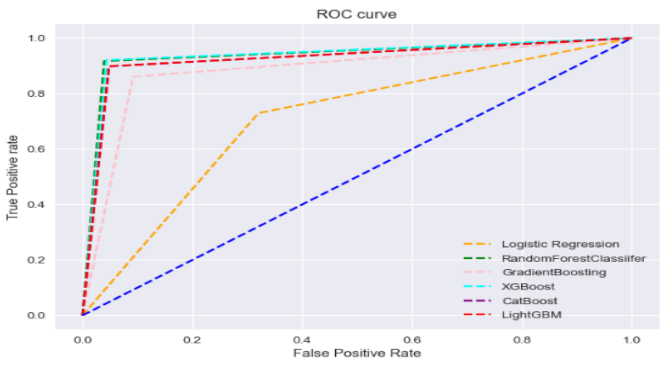

Fig 7: Machine Learning Algorithms Results Using ROC Curve

In figure 7, showcases the average test outcomes of six supervised algorithms that underwent training across datasets. It is evident that Random forest and XGboost achieved the highest average accuracy, recall, and AUC (Area Under the Curve). Among the models considered, Random forest and XGboost stood out as the top-performing base model.

IV. CONCLUSION AND FUTURE SCOPE

In summary, this research paper has delved into the domain of employee attrition using ML algorithms, highlighting the potential of data-driven solutions in managing workforce dynamics. The study has examined various ML techniques, like Random Forest, Logistic Regression, XGboost & Gradient Boosting, revealing their effectiveness in predicting and addressing attrition challenges. As organizations grapple with complex attrition scenarios, the adoption of ML algorithms offers a promising avenue for proactive decision-making and tailored retention strategies. Moving forward, prospective research could concentrate on enhancing model performance through advanced feature engineering and meticulous hyperparameter optimization. Additionally, the exploration of deep learning methodologies and emerging algorithms holds the potential to further enhance the precision of attrition predictions. Longitudinal studies tracking attrition trends over extended periods could also yield insights into the evolving landscape of employee turnover.

REFERENCES

1. S. Kakad, R. Kadam, P. Deshpande, S. Karde, and R. Lalwani, ‘‘Employee attrition prediction system,’’ Int. J. Innov. Sci., Eng. Technol., vol. 7, no. 9, p. 7, 2020
2. N. Shah, Z. Irani, and A. M. Sharif, ‘‘Big data in an HR context: Exploring organizational change readiness, employee attitudes and behaviors,’’ J. Bus. Res., vol. 70, Amazon.fr—People Analytics in the era of big Data: Changing
3. The way you Attract, Acquire, Develop, and Retain Talent—Jean Paul Isson—Livres. Accessed: Dec. 15, 2019. pp. 366–378, Jan. 2017, doi: 10.1016/j.jbusres.2016.08.010.
4. V. V. Saradhi and G. K. Palshikar, “Employee churn prediction”, Expert Systems with Applications, 38(3), 1999-2006, 2011
5. L. Breiman, "Random forests," Machine learning, vol. 45, no. 1, pp. 5-32, 2001.
6. M. Maisuradze, Predictive Analysis On The Example Of Employee Turnover (Master’s thesis), Tallinn: Tallinn University of Technology, 2017.
7. M. Maisuradze, Predictive Analysis On The Example Of Employee Turnover (Master’s thesis), Tallinn: Tallinn University of Technology, 2017.
8. S. Kaur and R. Vijay, "Job Satisfaction – A Major Factor Behind Attrition or Retention in Retail Industry," Imperial Journal of Interdisciplinary Research, vol. 2, no. 8, 2016.
9. R. Y. Zou and M. Schonlau, The Random Forest Algorithm for Statistical Learning with Applications in Stata The Random Forest algorithm, pp. 1–20, 2016.
10. Z.-H. Zhou, Ensemble Methods Foundations and Algorithms, CRC Press Taylor & Francis Group, 2012.
11. Z. A. Othman, A. A. Bakar, N. S. Sani, and J. Sallim, “Household Overspending Model Amongst B40, M40 and T20 using Classification Algorithm,” International Journal of Advance.
12. G. V. Sridhar, “Employee attrition and employee retention-challenges & suggestions employee attrition and employee retention-challenges & suggestions,” Rajalakshmi Eng. Coll. Dep. Manag. Stud., January 2018.
13. Rohit Punnoose and Pankaj Ajit, 2016 “Prediction Of Employee Turnover In Organizations Using Machine Learning Algorithms”, International Journal Of Advanced Research In Artificial Intelligence(IJARAI) Volume. 5, No. 9.