Journal of Science and Technology ISSN: 2362-1443
2023, Vol. 3, No. 1, 66-73

Prediction of Health Care Employee Turnover using Gradient Boosting Algorithm

Laxman Singh Khati
Nepal College of Information Technology
khati.laxman@outlook.com

Madan Kadariya
Nepal College of Information Technology
Madan.kadariya@ncit.edu.np

Article History:
Received: 11 July 2023
Revised: 14 October 2023
Accepted: 23 December 2023

Keywords*—*turnover prediction; machine learning model; Boosting; XGBoost; CatBoost; LightGBM; algorithm; regularization; classification; clustering

***Abstract*—Employee turnover is a measurement of how many employees are leaving a company and how many employees are retaining. Retaining the Employee is going to be biggest challenge for any of the organization in the world. Small to l arge organizations are hugely affected by such a big employee turnover problem. The healthcare industries are hugely affected by employee turnover problem. When employee leave the organization, the organization must have to replace such a employee with the employee having same skills, experience, behavior etc. This employee turnover prediction model uses the machine learning gradient boosting algorithms namely XGBoost, CatBoost, and LightGBM algorithms. Each models are trained, tested and validated and checked the performance metrices based on Pearson's correlation coefficients. Employee turnover prediction model helps to the supervisor to have a frequent, timely and relevant interactions or actions with the employee group based on prediction. Based on the prediction the supervisor may take actions on the employee before termination.**

I. INTRODUCTION

A. Background

Staffing or recruiting industry is highly sophisticated and large industry in world. Hiring the best candidate and keeping the candidate for long run in the industry is becoming more challenges day by day. Every employee in the industry wants more pay, respect, security, personnel development, adequate culture, and many more from the industry where they are working. If the employee does not satisfy form the industry facility, then He/ She should leave the industry.

Since employee turnover is the biggest problem in industries all over the world. Health care, Education, Technology and many more industries are hugely affected by such type of employee turnover. The main objective of the turnover predictive model is to propose, model, and implement a prediction system, which helps detect the high-risk employee, understand their problem, and helps to retain those employees.

It is necessary to detect or predict those employees who are going to be terminated in the future. High skill jobs in the industry of the world always need the experienced and skillful employee. If the well trained and skillful employee leave the company, it is great loss for the company. Turnover is an expensive business problem since acquiring a new employee costs more than 3 times than retaining the existing employee. The new employee needs extra costs like onboarding costs, training costs, supervision costs, and other different types of costs needs to employee work with the organization.

There are various technologies used to predict the termination probability of the employee. The termination probability is different for the different sectors for example: education, health care, logistic, manufacturing etc. These sectors have their own features and own termination rates. The termination rate also depends on the demographic data like age, gender, tenure, etc. of the employee. The features of the model are not the same for all the sectors available in the world.

The features are available based on the sector organization have and practice for storing and maintaining the data policy for that organization.

B. Research Questions

To fulfill the objectives of the study, the research question is formulated as following:

- How can employee turnover be predicted in the field of health care using gradient boosting algorithms?

C. Organization of the Report

This organization of the report gives an overview of the main points in our thesis analysis. Chapter 1 deals with the basic introduction, background and problem statement of the analysis. Chapter 2 presents conceptual review, related study and proposed framework of the health care employee turnover prediction. This also helps to understand the already existing analysis, related study in the problem in the health care organizations. Chapter 3 determines the research design, meta data, and process diagram of the analysis. This also helps to understand the overview of the proposed methodology to solve the problem of health care employee

turnover. Chapter 4 gives the overview of the result of the analysis, description of the data sets, and performance metrics of the proposed models. Chapter 5 includes the all the validation process to justify the results of the model. Chapter 6 concluded the final outcome comes from the proposed analysis, future enhancement of the problem and Limitation of the analysis.

II. LITERATURE REVIEW

A. *Related Study*

First, related worn on the employee turnover is discussed here. Rohit P. proposed a new contribution of extreme gradient boosting in employee turnover prediction and comparing XGBoost with other classical supervised algorithms [2]. The analysis shows that the XGBoost model provides better performance, better accuracy with efficient memory usage than other models[1]. Among proposed job, the powerful features for voluntary turnover prediction are job satisfaction, daily overtime, employee salary, distance travelled from home, marital status and employees' perception of fairness, an effective prediction model was introduced to predict the employee turnover who were likely to left the company [3]. The ensemble model generated by combination of Support Vector Machine, Random Forest and Logistic Regression provides and improves better performance than individual classifiers [1].

In the analysis of prediction of employee attrition using data mining [3], it was found that the historical and current employee detail information to estimate the future termination and study the common reason for employee termination.

In the analysis predicting Employee Attrition using Machine Learning [4], it was analyzed that they implement machine learning algorithms to predict termination of employee based on their features available. In this analysis, three experimental techniques were used on the employee dataset to model predictive models. First, the original unbalanced dataset was trained using different predictive models, with quadratic SVM achieving the highest results, with an F1 scores of 0.50. Second, the ADASYN balances the two classes of the unbalanced dataset. Predictive models trained with balanced dataset increases the performance of all the models significantly. The F1 score improves and achieved between 0.91 and 0.93 of models cubic, Gaussian, random forest and KNN (K = 3). Furthermore, the new technique features selection is applied on the dataset. The random forest achieved F1 score of 0.92 with only two features and with 12 features the random forest achieves lower F1 score 0.90. These results are quite similar with ADASYN balanced dataset result. The final technique was manually under sampling the dataset to have same classes. As a result, important information was not captured and leading to lower performance. Nevertheless, SVMs were able to capture more than 0.70 using all features, and more than 0.60 with only two features.

In the analysis predictors of home healthcare nurse retention, they have concluded that the nurses are the principal resource in the delivery of nursing services. The job satisfaction, job benefits, comparable salaries, team structure and ownership effect intent to stay and retention of the nurse employee [7].

In the analysis customer churn prediction using machine learning A study of B2B subscription-based service context ,they found that for the overall accuracy of considered classifiers, Naïve Bayes, XGBoost, Random Forest, ranges between 97.00% and 99.80%, indicating an overall good performance. Comparing the overall performance accuracy, recall, and F1-score, the boosting algorithm XGBoost is the best performer [8].

In the paper, Predictive modelling: An assessment through validation techniques researched on the validation techniques through assessment for predictive modelling used different validation techniques. However, the performance of 5-folded cross validation is better than other counter parts [9]. The k-fold validation technique has lower prediction error rates. The cross validation is best suited for analysis on the training data where model performance on different subsets.

B. *Conceptual Framework*

Gradient boosting is a step-by-step additive model that generates learners during the learning process. The contribution of the weak learner to the ensemble is based on the gradient descent optimization process. The calculated contribution of each tree is based on minimizing the overall error of the strong learner. Gradient boosting is based on minimizing a loss function.

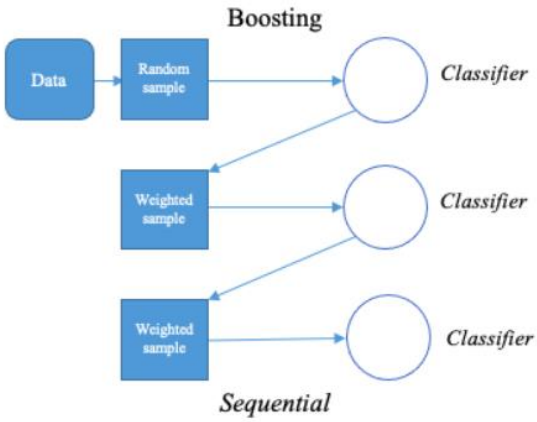

Fig. 1. Process flow of the boosting approach [8]

In general, the selection of model is carried out first for validating the model's quality. After that few models are generated. The gradient boosted models are created based on the data set prepared on data preparation phase. All the models are trained and tested separately. Gradient boost models that are going to developed are as follows:

1) XGBOOST

XGBoost model is the decision tree-based model. It uses boosted algorithm to predict the best fitted score. XGBoost is referred to as the "EXtreme Gradient Boosting" classifier. This is an ensemble model. XGBoost model has higher performance and speed than traditional machine learning models. XGBoost is used for supervised learning classification problems, where we use the training data to predict the target variable. Since it is a decision tree-based model, the model is very easier to interpret. This model feds the input features and predict the score for the input data. It

provides a parallel booting tree. It solves data in fast and accurate way.

Decision Tree based models are considered best for the small size to medium size data, structured or tabular data structure. But for the unstructured data like images, texts the performance is better with neural network models. XGBoost algorithm has not only been credited with winning numerous Kaggle competitions but also for being the driving force under the hood for several cutting-edge industry applications.

XGBoost open-source project has strong community of Data Scientists with ~522 contributers and ~29815 users on GITHub. The XGBoost algorithm is popular because of the reason such as can solve classification, regression problems on wide range of applications, portable on any environments Windows/ Linux/ Mac OS and easier to integrate on cloud technologies like AWS/Azure/GCP etc.

2) CatBoost:

CatBoost is also a decision-tree based gradient boosting algorithm. It is open-sourced machine learning algorithm from Yandex. The CatBoost can be easily integrate with deep learning frameworks such as Google's TensorFlow and Apple's Core ML. The CatBoost algorithm is able to work with diverse data types to help solve a wide range of problems faced by today's business.

Catboost deals with the categorical variables. CatBoost has the flexibility of giving indices of categorical columns hence it can be encoded as one-hot encoding using one_hot_max_size. Which usees one-hot encoding for all features with a number of different values less than or equal to the value given in the parameter. Also, if nothing is passed in cat_features argument, it will treat all the features as numerical variables.The CatBoost implements symmetric trees but other boosting algorithm implements asymmetric leaf-wise tree growth. This helps in decreasing prediction time, which is extremely important for low latency environments.

3) LightGBM:

The full form of LightGBM is lightweight gradient boosting machines. It uses a novel technique of Gradient-based One-Side Sampling (GOSS) to filter out the data instances for finding a split value. The LightGBM is prefixed as 'Light' because of its high speed perfromance. The LightGBM is popular because of its capacity to handle the large size of data and takes lower memory to run. Another reason why LightGBM is highly used is as it spotlights on the accuracy of results.

LGBM supports GPU learning and hence data scientists are widely using LGBM for data science projects. The categorical features are handled in LightGBM by taking the input of feature names. The LightGBM uses a special algorithm to find the split value of categorical features. Both LighGBM and XGBoost grow the tree's leaf wise.

III. DESIGN AND METHODOLOGY

In this section a research design, sample of data, nature of data source of data, data collection technique and research methodology are presented. The design and methodology of the analyzed concepts below is essential to fully understand the study.

A. Research Design

The overall research objectives of the analysis is to compare the gradient boosting algorithms for the health care employee turnover risk prediction. It analyses the attrition of the employee and continuity of the employee towards the organization. Hypothesis are created to justify the objectives of the analysis. That hypothesis provides the depth knowledge of the analysis and correlation between the dependent and independent variable. Those hypotheses are again tested with the predicted data as well. The data required for the analysis is taken from multiple system used by the health care industries. The exploratory data analysis is done to check the distribution of data, variability of the data , and some more statistical analysis.

B. Population and Sample of data

To complete analysis, here is the dataset of 70 thousand rows of true observation and 10 thousand of false observations with 35 features of the observation. The data set includes the employee demographic data and time and attendance data of the health care employee. Please consider for some unidentified description because of HIPAA compliance of sensitive data.

The features and feature description in given in the below table:

TABLE I. FEATURE DESCRIPTION

feature_name	data type
Team structure factors	int
Demographic factors (1)	decimal
Demographic factors (2)	decimal
Variations between actual and scheduled time worked (2)	decimal
Change in organization reporting	decimal
Demographic factors (4)	decimal
Employee chose to work less time than scheduled	decimal
Burnout indication - amount of work (3)	decimal
Worked night shift (%)	decimal
Burnout indication - amount of work (2)	decimal
Variations in scheduled shifts	decimal
Burnout indication - inconsistency in work schedule (2)	decimal
Increase in variations between actual and scheduled time worked (2)	decimal
Variations between actual and scheduled time worked	decimal
Increase in variations between actual and scheduled time worked	decimal
Demographic factors (3)	decimal
Change in team they work with	decimal
Change in team they work with (2)	decimal
Departments	varchar
Burnout indication - amount of work	decimal
Worked additional shifts	decimal
Was asked to work less time than scheduled	decimal
Worked at a different, additional event outside of scheduled hours	decimal
Burnout indication - inconsistency in work schedule	decimal

C. Nature of source of data

The source of data is health care employee recorded by hospital, nursing home, child care hospital from United States of America. The hospitals, nursing homes and child care hospitals use different types of systems to collect data. Such as Human resource software collects regarding demographic data of the employee and time and attendance

software collects data regarding check in, check out, break in break out, leave etc.

D. Data collection technique

The data is extracted from different systems like human resource software, time and attendance software used by the hospitals. Then the extracted data is transformed to the adequate format required to analyze the data. Finally, the transformed data of multiple hospital is loaded into the data warehouse for the purpose of further analysis. As per the data source the data is available from 2018 to 2022 June, all the data is used for this analysis.

The consent of using data is already done with the authority. Because of HIPAA (Health Insurance Portability and Accountability Act) compliance the data and details of data attributes of the health care employee is kept confidential. The data collection technique fully aware of HIPAA compliance.

E. Research methodology

The experiment design followed the CRISP-DM process diagram in the research lifecycle. CRISP-DM defines the Cross Industry Standard Process for Data Mining. This process is widely regarded data science process and based on agile project management methodology. Machine learning latest technology models such as XGBoost, CatBoost and LightGBM were used to carry out the experiments of the analysis. The major goal of this analysis is building, comparing and modelling the employee dataset to predict the turnover probability of the employee.

The CRISP-DM process diagram is given below.

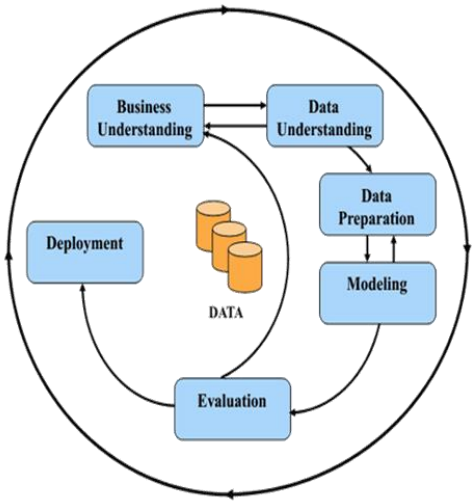

Fig. 2. CRISP-DM Process Diagram[10]

1) Business Understanding:

In this phase the in-depth analysis of business entities has been done. Overall business process is carried out to analyze the purpose of businesses, problems solved by business, proceeds how the problem is solved. Health care staffing and human resource process has been analyzed on the basis of Staffing practice for the hospital employee. Human resource tracks the employee behavior and daily activities of the employee in the health care industries. Employee personal behavior as well as behavior towards the employee or hospital is analyzed.

2) Data Understanding:

It is very difficult to understand health care employee data than other data set. Health care data is more sensitive than over sector employee data. Due to health care industrial compliances and highly sensitive employee information, some crucial information is hidden from public. Also, health care data is not sufficient to gather from single systems. Based on the data tracking there are different systems which includes employee information such as employee demographic data is available on one system, financial data is available on one system, time tracking is available on another system. The extraction of data from different sources is needed. Then the extracted data is transformed into the required model and loaded into the storage system. All the employee who are currently working on the organization and who left the organization in previous days are analyzed. All the data is tacked on monthly basis. The data will probably be changed in every month.In depth data analyzed based on dependent features and independent features. The correlations between dependent features and independent features are analyzed based on some shorts of hypothesis for the employee turnover.

Hypothesis 1: It is hypothesized that, higher the size of the hospital higher the employee turnover because of lower interaction with the employee.

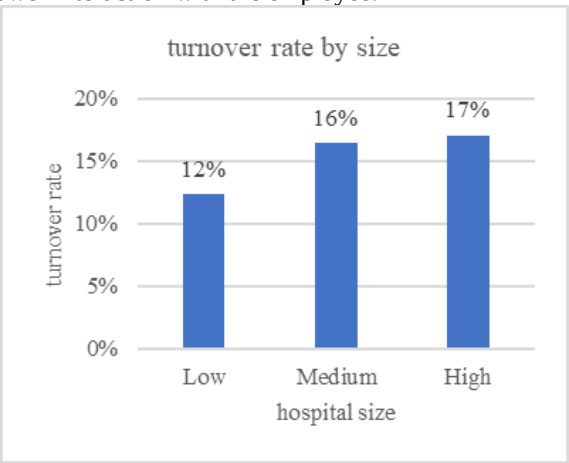

Fig. 3. Turnover by hospital size

Analysis result: The given hypothesis is true. The below graph shows that low size hospital has the lower turnover than medium and so on.

Hypothesis 2: It is assumed that larger the team structure higher the turnover rate in general. It is very difficult to manage the larger team than smaller team.

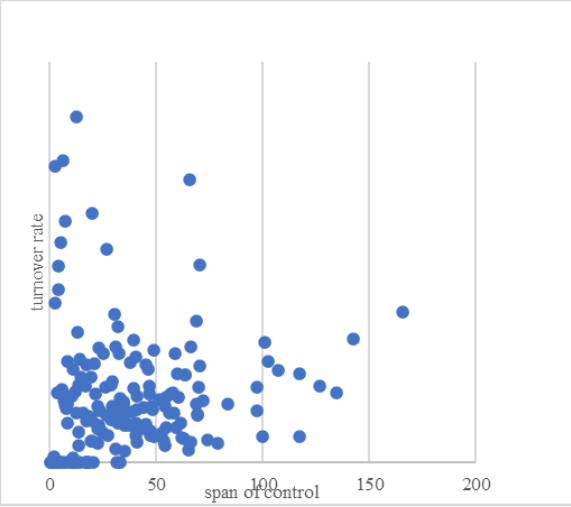

Fig. 4. turnover by team structure

***Analysis result*:** The given hypothesis is true. The below graph shows the mix result but we can conclude that larger team structure has higher turnover risk.

Hypothesis 3: It is assumed that the departments like ICU/CCU and emergency departments has higher the turnover rate because of hectic work pressure and hazardous working conditions.

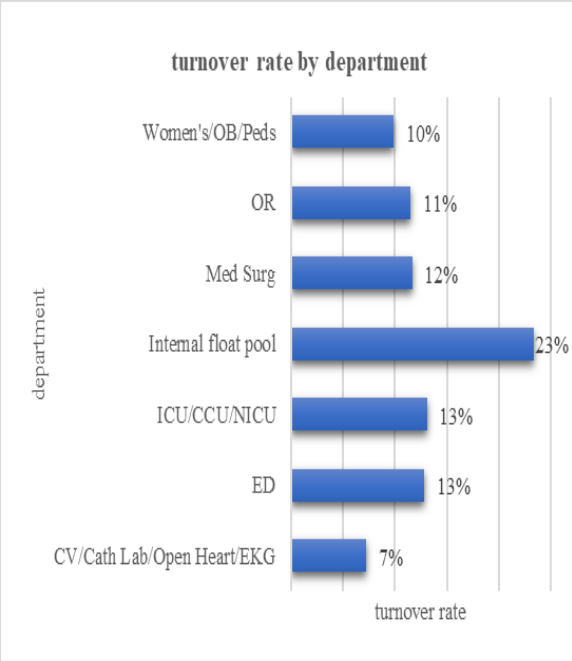

Fig. 5. turnover by departments

Analysis result: The analysis is also true as per the hypothesis. But it shows that internal float pool has higher turnover rate because those employees are not working permanently in the hospitals.

Hypothesis 4: It is assumed that per diem and part time employee has higher turnover rate than full time employee because every employee wants full time engagement and secure job.

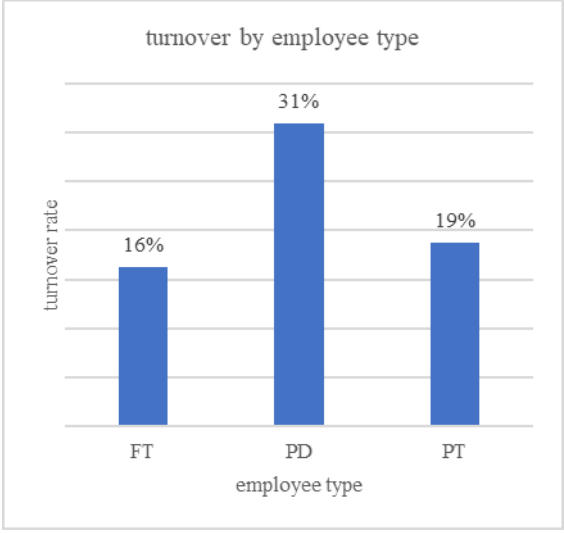

Fig. 6. turnover by employee type

Analysis result: Every hospital gives full time employee more incentives, leaves, also ownership than per diem and parttime employee. The analysis is also true as per the hypothesis.

Hypothesis 5: It is assumed that Registered nurses (RN) are skillful than certified nursing assistant (CNA) and Certified nursing assistant works under registered nurses. So certified nursing assistant has higher turnover rates.

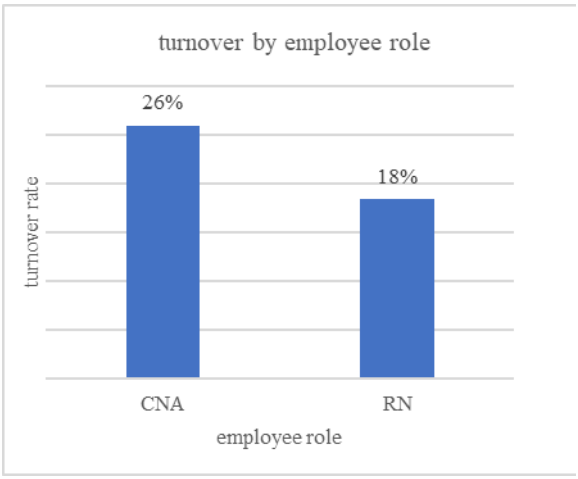

Fig 8. turnover by employee role

Analysis result: The analysis is also true as per the hypothesis.

Hypothesis 6: It is assumed that for the early young generation from up to age 25 is filled with an employee's excitement about new experiences. They tend to be more optimistic about what's possible. They have no clear path for the future and norms. So, their possibility of turnover is high. When the experience is increases, the employee confidence, skillset, leadership is increases and they are not likely to turnover for shorten time.

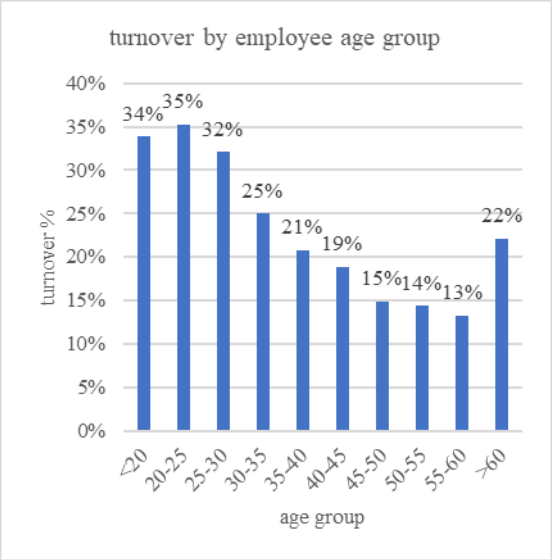

Fig 7. turnover by age group

Analysis result: The analysis is also shows that the turnover rate is increases up to age 25 and after that age group the employee wants to settle with the hospital and again turnover will decrease and again increases when the employee getting older.

3) Data Preparation:

This is the very important and time-consuming step of machine leaning life cycle. After the data sources are completely identified, proper selection, cleaning, encoding categorical data, scaling and splitting is done. The exploratory data analysis of information may be executed for noticing the patterns and trends in light of business understanding.

After understanding data clearly, features will be confined with the data required for the prediction. To compare the different algorithms there are around 70 thousand rows of true observations and 10 thousand of false observations having more than 35 features of the data set prepared for the analysis with some demographic and non-demographic features like time and attendance related data. Since the data is not balanced, upsampling is done using python based scikit learn tools. The data is balanced in the ratio of 7:3.

The data that impact the prediction are analyzed and selected for the prediction. Features not included with more than 50% null values or have only one unique values or features that has no contributions towards turnover.

Pearson's correlation coefficient is used to reduce the features to compare the performance metrices of the selected models. Pearson's correlation coefficient shows the magnitude and direction of the statistical relationship, or association between two continuous variables.

In data cleaning, removing the noise from the data. It is very important to remove the unnecessary value from the data set that causes the poor result in the prediction. The feature may also have null values and those null value needs to be replaced with meaning full values. Null value for bit data type features is replaced with mode value and Null value for other numerical value are replaced by average value from the available values.

In Encoding categorical data, the machine learning algorithm will not act upon the categorical data. So, we need to encode those categorical data in numerical format. Since the categorical variable is not ordinal, dummy variable is used to encode the categorical data. This method assigns the separate columns for category names. The column value is 1 only if the category name is same for the column otherwise 0.

The scaling standardizes the independent variables of the dataset in the specific range. Adjusting all the variables in the same range and in the same scale so causes no any variable dominates the other variable. Since all the gradient boosting algorithms are tree-based learners, and any monolithic function of any feature value will have no effect on tree formation. So, no scaling is selected for these algorithms.

The process of splitting the dataset for the training and test purpose is splitting. Splitting test and train splitting ratio 70:30 gives the best fit to the algorithms. This can enhance the performance of our machine learning model.

4) Modelling :

In general, the selection of model is carried out first for validating the model's quality. After that few models are generated. All the gradient boosted models are created based on the data set prepared on data preparation phase. All the models are trained and tested separately. Gradient boost models that are going to developed are as follows:

- XGBOOST
- CatBoost
- LightGBM

5) Evaluation:

The outcome results of the models are judged in the evaluated in the backdrop of business objectives. The new objectives may grow up owing to the new patterns discovered by new data generated. This is machine learning agile process and the decision has to be made before deployment whether to consider them or not. All the models developed are tested with new test data and the result is evaluated. The model's performance and efficiency using the metrices, F1-score, Confusion matrix measures such as accuracy, recall, precision etc. are evaluated based on the Pearson's correlation coefficient. Evaluation is done based on test data which is not been trained on the training model. The data was split into train and test in the data preparation phase set to be able to determine how well the model performs on the data set other than trained data.

6) Deployment:

The final outcome results will be presented to the respective stakeholders. The model will be deployed with the techniques feasible. There are various techniques to deploy model such as batch offline, web service, streaming etc. depending on the how fast and streamed the predictions chooses.

7) Tools and Technology:

Tools and Technologies that will be used in the analysis are as follows:

- Python is the high-level Programming Language used to analyze the data and develop the model. All the research work is based on python programming language.

- Pandas is the special type of python library used for data analysis and data manipulation. Pandas has strong and heavy supporting functions to analyze the data.
- Numpy is another special type of python library used for data computation. All the mathematical operations are supported by Numpy.
- Sklearn is the machine learning data preparation and modeling library. It is free software by python programming language. It features the different types of algorithms such as classification, regression, clustering etc.
- Jupiter notebook is special type of computing platform used to analyze and develop the models.
- MSSQL is the relational database management system used to extract, transform and load of the data.

IV. RESULTS

The implementation is carried out training the all three models in Jupiter Notebook platform using different types of python-based data analysis libraries, such as Pandas, NumPy, Seaborn, Scikit-learn and MSSQL as a data transform and storage system. Scikit-learn is the powerful library with built-in functions for implementation of algorithms, training, testing, and evaluation used in predictive data analysis.

In order to answer research question, the different scenarios are analyzed. Of particular is to determine the performance of the model, to analyze the correlations of the features of the data available, and to analyze the best model from the gradient boosting algorithms.

A. *Confusion Matrix*

It is a special type of contingency table, with two dimensional table having values actual and predicted. The confusion matrix shows the overall performance of the models based on actual and predicted values. Below table represents the confusion matrix

TABLE II. CONFUSION MATRIX AND THEIR DEFINITION

		Predicted Value	
		Yes	***No***
Actual Class	***Yes***	TP	TN
	No	FP	FN

The Accuracy is simply the sum of correct predictions divided by the total available predictions. Precision is defined by the measure of how accurate the model is in predicting the actual positive available predictions out of the total positive predictions predicted by the model. Recall is defined by the number of actual available positive captured by the model by classifying there as true positives predicted. F1-Score provides a balance between precision and recall.

$$Accuracy = correct\ predictions/total\ predictions \quad (1)$$

$$Precision = TP/TP+FP \quad (2)$$

$$Recall = TP/TP+FN \quad (3)$$

$$F1\text{-}score = 2*precision*recall/precision+recall \quad (4)$$

Where TP-True Positive, TN- True Negative, FP-False Positive, FN-False Negative

B. *Performance metrices*

The performance metrics validates the methodologies includes the accuracy, precision, recall and F1-score. These metrics defines and evaluates the performance of the gradient boosted algorithms based on the feature selection. The Pearson correlation is used to select the features.

The accuracy of the models is given as follows:

TABLE III. PERFORMANCE METRICS

Model	**Correlation Coefficient**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
XGBoost	***All features***	92.29	87	89	88
	>0.01(27 features)	90.59	84	78	81
	>0.03(16 features)	88.94	82	73	77
	>0.05(9 features)	88.07	81	71	76
CatBoost	***All features***	93.60	88	93	90
	>0.01(27 features)	92.24	85	85	85
	>0.03(16 features)	91.60	83	85	84
	>0.05(9 features)	91.07	82	84	83
LightGBM	***All features***	92.82	91	87	89
	>0.01(27 features)	91.80	83	85	84
	>0.03(16 features)	90.67	81	82	82
	>0.05(9 features)	88.99	79	82	80

This table shows that feature selection using different correlation coefficient the CatBoost performed better than other models. The models trained with all features, 27 features removing less correlated features using 0.01 correlation coefficient, 16 features removing less correlated features using 0.03 correlation coefficient, 9 features removing less correlated features using 0.05 correlation coefficient.

Feature Importance of the best model Cat boost is given in the below table:

TABLE IV. FEATURE IMPORTANCE

Feature name	**Importance (%)**
Team structure factors	13.27
Demographic factors (1)	6.95
Demographic factors (2)	6.86
Variations between actual and scheduled time worked (2)	6.33
Change in organization reporting	6.30
Demographic factors (4)	6.12
Employee chose to work less time than scheduled	6.03

Burnout indication - amount of work (3)	5.56
Worked night shift (%)	5.11
Burnout indication - amount of work (2)	4.62
Variations in scheduled shifts	4.60
Burnout indication - inconsistency in work schedule (2)	4.55
Increase in variations between actual and scheduled time worked (2)	4.31
Variations between actual and scheduled time worked	4.17
Increase in variations between actual and scheduled time worked	3.79
Demographic factors (3)	1.86
Change in team they work with	1.43
Change in team they work with (2)	0.60
Departments	0.40
Burnout indication - amount of work	0.24
Worked additional shifts	0.21
Was asked to work less time than scheduled	0.09
Worked at a different, additional event outside of scheduled hours	0.06
Burnout indication - inconsistency in work schedule	0.03

This table shows that the feature importance of the best performed model CatBoost. Out of all the features used in training the model the feature team structure factors have the highest impact on the turnover prediction and demographic factors are second important factors to predict the employee turnover. The factors like Burnout indication - inconsistency in work schedule, Worked at a different, additional event outside of scheduled hours, Was asked to work less time than scheduled has lowest impact on the employee turnover prediction.

V. VALIDATION

The turnover prediction models of gradient boosting algorithms are validated by using K-fold cross validation technique and back testing of models. All the models are validated before compare. Based on validation the CatBoost model performs better than XGBOOST and LightGBM model. The validation process also includes the analysis of hypothesis validated with the predicted results from the model. All the hypothesis are also true for predicted results.

K-fold cross-validation divides the data set into K groups of equal sample size. These samples are the folds used to test prediction and train the model. To validate the model K=10 is configured. So, dataset is splitting into 10 folds to run the train and test. During each run, 1-fold is reserved for testing and 9 folds are used in training. Accuracy obtained from each test is averaged to get final accuracy of the model.

In back testing validation approach, the model is trained on historical data and tested with latest data available. This approach tests the model with real data of the business.

VI. CONCLUSION

The comparative data analysis provides a fair and unbiased performance of the gradient boosting algorithms used for employee turnover prediction. This analysis also defines the feature importance of best model CatBoost. The comparative performance analysis is achieved by keeping the uniformity across the studied population data, strong exploratory data analysis, set of demographic data and non-demographic data set, best feature selection analysis, the cross-validation procedure, hyper parameter tuning and the performance metric. CatBoost algorithm gives best outcomes among all gradient boosting algorithms with different Pearson's correlation coefficients. The achieved performance of the best model was 93%. However, the XGBOOST and LightGBM classifier performed little bit lower than cat boost.

A. *Future Enhancement*

This analysis is complete however, in future the following enhancements can be done in the analysis research.

- Data regarding employee financial needs to be added on the analysis to increase the performance of the models.
- Data regarding employee conversation are analyzed and use of latest natural language processing to know the sentiment of employee dropping the job.
- Development of ensemble model to increase the performance of the model.
- Use of interpretation and evaluation libraries for the feature recommendation of the turnover reason of the employee

B. *Limitation of the analysis*

Since, Employee turnover system includes data from different sources such as HR systems Time and Attendance systems, the data is confidential so, it is tough to make sense of every data label. The data ingested for the analysis is also not balanced.

REFERENCES

[1] Shubham Karande and L. Shyamala, Prediction of Employee Turnover Using Ensemble Learning , 2019

[2] Ajit, P.,Prediction of employee turnover in organizations using machine learning algorithms. Algorithms, 4(5), C5 (2016)

[3] R Shiva Shankar,J Rajanikanth,V.V.Sivaramaraju, and K VSSR Murthy, Prediction of employee attrition using data mining, 2018

[4] S. S. Alduayj. and K Rajpoot, Predicting Employee Attrition using Machine Learning, 2018

[5] R Yedida, R Reddy, R Vahi, R Jana, A GV, D Kulkarni, Employee Attrition Prediction, 2018

[6] A K Ahmed, A Jafar, K Aljoumaa, Customer churn prediction in telecom using machine learning in big data platform, 2019

[7] Carol Hall Ellenbecker, Frank W. Porell, Linda Samia, James J. Byleckie, Michael Milburn, Predictors of Home Healthcare Nurse Retention, 2008

[8] Benjamin G. and Yasin O, Customer churn prediction using machine learning A study of B2B subscription based service context, 2021

[9] M. Iqbal Jeelani, F. Danish and S., Predictive modelling: An assessment through validation techniques, 2022

[10] Han, J.W., Kamber, M. and Pei, J., *Data Mining Concepts and Techniques. Elsevier, San Francisco*, 2006