Employee Attrition Rate Analysis

1st Isha
Department of Computer Science
Graphic Era Hill University
Dehradun, India
ishu19603@gmail.com

2nd Sangita Papola
Department of Computer Science
Graphic Era Hill University
Dehradun, India
sangitapapola@gmail.com

3rd Sheetal Solanki
Department of Computer Science
Graphic Era Hill University
Dehradun, India
sheetalsolanki2207@gmail.com

***Abstract*— In today's time, employee attrition is a major problem faced by organisations. Employees are the most valuable assets of an organization. It is them who add value to the organization in terms of quantity and quality as well. Attrition is said to be gradual decrease in the number of employees through resignation, death or retirement. It is a matter of concern when employees leave their jobs of better opportunities but this leads the company to face its consequences. Employee attrition is similar to employee turnover. By understanding this concept, leaders can design smarter retention strategies to avoid it. In this paper, we use dataset to analyse attrition and find the reason why employee choose to resign. We did analysis of the training dataset for effective data exploration. Machine Learning algorithms, such as Random Forest Classifier, Adaboost, XGboost and ensemble stacking technique. After data exploration and enabling algorithm we found that random forest classifier proved to be the best suited algorithm for this data set with an accuracy of 87.41% and precision 88%. We may also conclude that monthly income can be a reason behind employees leaving their job and finding better oppurtunities.**

Keywords—Attrition, Random Forest, Adaboost, XGboost, Exploration.

INTRODUCTION

Employee attrition occurs when the size of your workforce diminishes over time due to unavoidable factors such as employee resignation for personal or professional reasons.
Employees are leaving the workforce faster than they are hired, and it is often outside the employer's control. An employee would choose to join or resign from a company considering many factors. Some take salary, department, working environment, distance from home, gender equity, etc while some consider personal reasons like maternity, family, health issues. Employee attrition is the unexpected or unpredictable process either voluntarily or involuntarily. It's analysis refers to understanding and managing employee attrition. It plays a crucial factor in company's success and finance.

It is a major problem for the organisation particularly when trained, technical and important employees leave for better opportunities from the organisation. High employee attrition negatively impact a company's performance and popularity. Constant staff turnover can create instabilities within the company, leading to less employees and other financial drawbacks. Attrition not only affects the business but also the brand image of a company. The attrition rate measures how many employees leave an organization over a particular period and is expressed as a percentage.

We can calculate it by dividing the number of employees who went by the total number present at the beginning of the period. For example, let's headcount to know how many employees a company started with at the beginning of the year suppose 1,000.Keep track of how many people leave throughout the year. Let's say 200 employees left the company due to voluntary and involuntary reasons.
Keep track of the employees you hire across the year, and conduct a final headcount at year-end. Let's say that you hired 400 people that year – this means it's final headcount is 1,400.

Now we calculate the average number of employees for that year which comes out to be (1000+1400)/2 = 1,200.
Finally, calculate the number of employees who left as a percentage of the average number of employees.

The formula for it is **Average number of employees/number of employees who left*100.** This will give you the attrition rate: (200/1200) x 100 = 16.66%.

In order for an organization to continually have a higher competitive advantage over its competition, it should make it a duty to minimize employee attrition. Therefore, for the better development of organisation, it is essential for the company, then take relevant measures to improve their company's productivity and business performance. It works as a reference for future employees who are looking for job opportunities whether to join on a particular company or not.

In 2018 Annual Diversity Report, Google had a separate section dealing with employee attrition. It was found that attrition among black and latinx employees outdid the attrition in company average by a wide margin. Black employees make up just 2.5% of Google's U.S. employee are only slightly higher, at 3.6%. White and Asian employees comprise the majority of the workforce in the office. This shows why monitoring, measuring and addressing employee attrition is important. The company turned a spotlight on its employees on the way to provide equitable future.

Another example emphasizing the importance of measuring attrition is Uber. Uber has categorically denied above-normal attrition rates – but unreliablel evidence suggests otherwise. In 2019, when the company filed for an IPO, its public disclosures mentioned several instances of workplace culture issues, poor employer reputation, and reduction in employee incentives. This can be linked to a large number of employees who have allegedly left Uber in the last couple of years.

There is a clear difference between Google and Uber's attitudes. While Uber approaches attrition as an unavoidable outcome of its workplace policies, Google takes a more honest and candid stance, admitting that employee attrition is an issue that needs to be addressed. And this is in sync with the overall retention and attrition climate in the U.S. today.

In 2019 Retention Report by Work Institute, it was found that preventable attrition is more frequent than unpredictable attrition.10,6,and 6 out of 100 which is just a total of 22% ,employees left due to relocation, retirement or termination. The Society for Human Resource Management (SHRM) determines that USD 4129 is the average cost-per-hire for a new employee. It was predicted the by this year, the rate of attrition will be so high that 35% of employees will leave every year to work in a different company. In 2018, the number of vacancies crossed the number of unemployed workers for the first time. This means companies can't rely on job satisfaction to keep employees loyal, the work-life Career development was the number one reason for leaving, the work-life balance came in second, and manager behavior was a close third.

Also, attrition isn't always a bad thing. In many cases, employee attrition can actually be good for business. As poor-performing employees leave the company and make room for new talent. In tech companies usually large number of men are employing leading to disproportionality, attrition, here, can enable diversity. It roots out employees who aren't a good fit for their jobs and probably shouldn't have been employed at the first place. It helps to create a dynamic workforce as the same employees with the same perspective aren't running the company for a long period of time.

Employee attrition is confidential data of the company so a dataset is randomly selected from kaggle and it is used to analyse the data and predict the attrition rate of a company. We used some methodologies of machine learning by stacking them for classification of data and prediction of result.

Literature

There have been many machine learning works on employee attrition. Taking inspiration from the work done, we apply combinations of some of these models and techniques on our selected datatset and analysis it to et proper results. Below are some of the works:

[2]They used Global retailer's HRIS database, BLS(Bureau of Labour Statistics) data. They trained and tested XGBoost, Logistic Regression, Naïve Bayes, Random Forest, SVM, LDA, KNN models on ROC-AUC metric on the dataset . The results came out to be that TXGBoost classifier is a superior algorithm in terms of significant higher accuracy, Relatively low runtimes and efficient memory utilization for predicting employee attrition.

[1] They used real dataset from IBM analytics. They identify the main factors affecting and propose classification based on the statistical evaluation of the data. Results are expressed in terms of classical metrics and Gaussian Naïve Bayes classifier produced the best result for the available dataset. It reveals the best recall rate (0.54), since it measures the ability of a classifier to find all the positive instances and achieves an overall false negative rate equal to 4.5% of the total observations.

[3]They analysed IBM Employee Attrition to find the main reason why employee choose to leave their respective jobs. Firstly they applied correlation matrix to remove the unwanted or less-important attributes from the dataset. They selected important features by exploiting Random Forest which found that monthly income, age and number of companies worked impacted employee attrition.They divided people into two clusters by using K- means Clustering followed by binary logistic regression quantitative analysis and concluded that people who travelled frequently was 2.4 times higher than that employees who travelled rarely and employees in Human Resource have a higher tendency to leave.

[4]. This paper aimed to analyse employee attrition using logistic regression. They used R for data integration, exploratory data analysis, data preparation, logistic regression, model evaluation and visualization. The data was divided into five dataset with unique employee ID identical across all dataset.They used str() and summary() function to comprehend the dataset. For model selection and training, the target value of a variable was converted from yes/no to levels 0/1. They allocated 70% of items to the training set and 30 % to the test set. After model test and evaluation accuracy, sensitivity and specification of the prediction model came out to be 75%,73% and 75% respectively. It was found that an employee with lesser number of working years and companies worked has a more significant probability of attrition. To reduce employee attrition rate, the company needs to improve the human resource department by evaluating the working environment, job satisfaction, employee workload, and interaction between manager and employee.

[5]. Three hundred and nine records of employees of one of the Reputed Institutions in Nigeria who worked in the institution and left it between 1978 and 2006 were used for the study as dataset. Job related and demographic data were

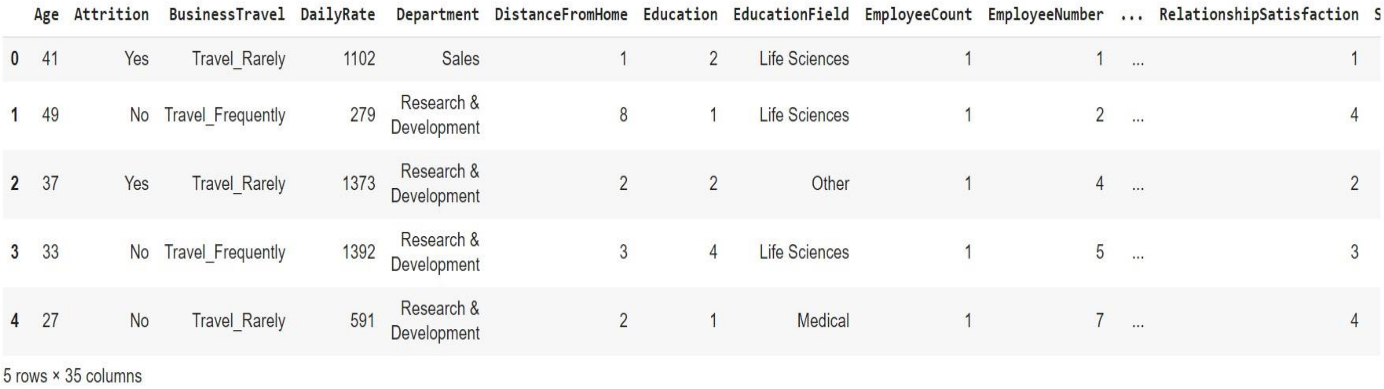
mainly focussed upon to classify employees into predefined attrition classes. Waikato Environment for Knowledge Analysis (WEKA) and See5 for Windows were used to generate decision tree models. The results of the decision tree models and rule-sets generated were then used for developing a predictive model that was used to predict new cases of employee attrition. The single decision tree was generated of size 15 with 3 sub-trees with a misclassification of 25.2%. The attribute usage results shows that salary had 100% usage, the length of service had 49% usage and 16% for rank. These numbers concluded that the salary earned by the employee and the length of service contributed by an individual were the prime factors that contributed for an employee to stay or leave an organization.

[6]. The prime objective of this paper was to analyze why some of the best and most experienced employees are leaving prematurely from their jobs and also predict which employee will leave next. The dataset was selected from kaggle. The study was done using R and Rattle data mining platform. The dataset was divided into training, testing and validation.. They picked Decision Tree, Random Forest, Support Vector Machine (SVM), and Linear Regression techniques to build the model. The performance of the model is evaluated in terms of Error Matrix and Pseudo R Square estimate of error rate. It predicted that Random Forest is the most suited model for employee attrition analysis in comparision to other models. “Job satisfaction” was the dominant factor which influenced attrition rate the most.

[7].The IBM HR Employee Attrition was used for data analytics. They compared four machine learning techniques, Support Vector Machines (SVM), Logistic Regression (LR), Decision Tree Classifier (DTC) and Extra Trees Classifier (ETC) for analysis. The proposed Extra Trees Classifier (ETC) approach achieved an accuracy score of 93% for employee attrition prediction.. Employee Exploratory Data Analysis (EEDA) was applied to determine the factors which caused employee attrition. Their study concluded the result that monthly income, hourly rate, job level and age were the key factors that caused employee attrition.

Proposed Methodology

- The dataset is a set or collection of data. This set is normally presented in a tabular pattern. Every column describes a particular variable. And each row corresponds to a given member of the data set, as per the given question. This is a part of data management. For our dataset we take employee data from kaggle which contains 1470 records and 35 fields including categorical and numeric features. Each record in the data set represent a single employee information and each field in the record represents a feature of that particular employee.

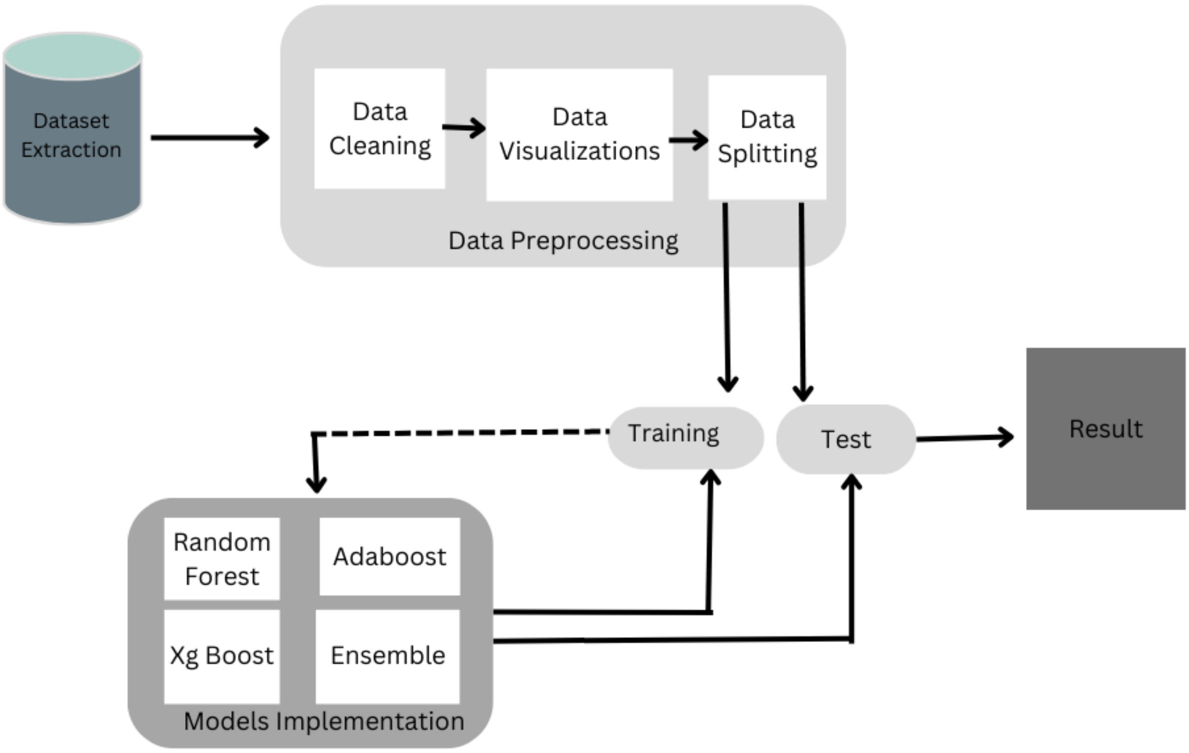


	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	...	RelationshipSatisfaction	S
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	...	1	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	...	4	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	...	2	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	...	3	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	...	4	

5 rows × 35 columns

- For the data pre-processing and feature selection we have used RFE(Recursive Feature Eliminator). Feature selection refers to techniques that select a subset of the most relevant features (columns) from a dataset. Fewer features can allow machine learning algorithms to run more efficiently (less space or time complexity) and be more effective. Some machine learning algorithms can be misled by irrelevant input features, resulting in worse predictive performance. RFE is a feature selection algorithm, popular because it is easy to configure and use and because it is effective at selecting those features (columns) in a training dataset that are more relevant in predicting the target variable. RFE is a wrapper-type feature selection algorithm. This means that a different machine learning algorithm is given and used in the core of the method, is wrapped by RFE, and used to help select features. We separate the data between test dataset and training dataset by 20%. The training dataset has 1176 records and test dataset has 294 records with same number of 34 features.

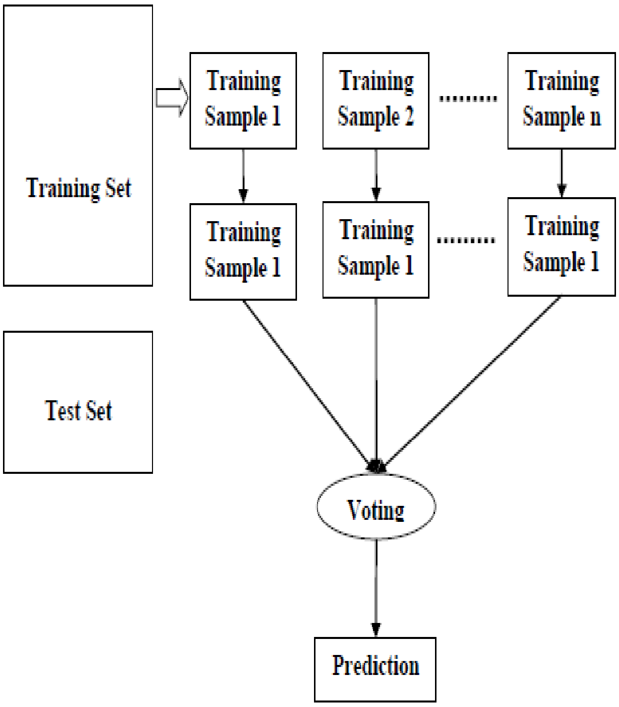
- Data exploration and visualiztion is a vital process in data science. Analysts perform exploration on a dataset to illuminate specific patterns or characteristics to help companies or organizations understand insights and implement new policies. Data exploration tools make data analysis easier to present and understand through interactive, visual elements. This step helps users to understand their data statistically and visually. Data visualization is a graphical representation of data. It uses visualization tools such as graphs and charts to allow for an easy understanding of complex structures and relationships within the data. Making data more understandable will benefit every professional field. By effectively using the ability of our eyes to quickly identify different colors, shapes, and patterns, data visualization enables easier interpretation of data and better data exploration. We used various graph plotting. Having attrition as a common feature we plotted it with number of employees, department, age of the employee, job satisfaction etc.



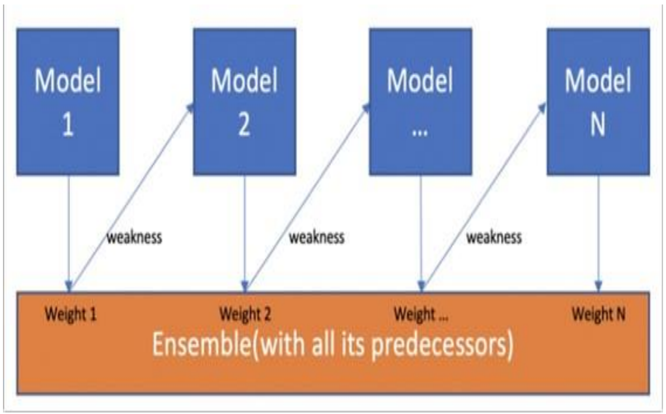
ModelBuilding

The study throws light through valuable suggestion and solution helping them to reduce the employee attrition. We used stack of machine learning classifiers to select important features and analyse the data. A classifier is an algorithm that automatically assigns data points to a range of categories or classes. There are various types of classifiers in machine learning like decision tree, random forest classifier, logistic regression, naïve bayes, support vector machine (SVM) , etc . In this project we have used Random forest classifier, Adaboost, XGboost and ensemble stacking technique to find which suits best on out dataset.

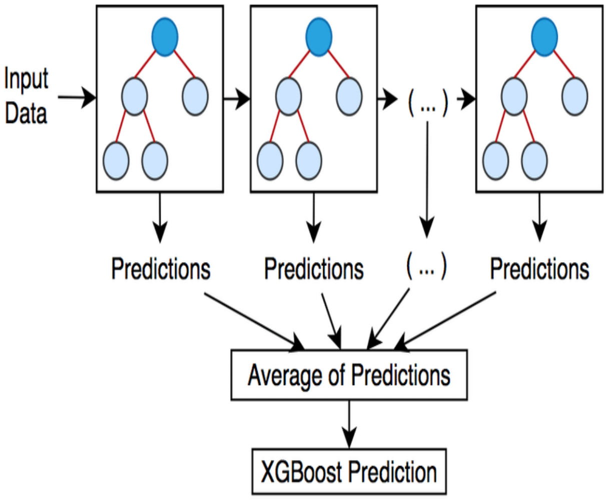
Random Forest is a popular machine learning algorithm that can be used for both Classification and Regression problems in machine learning. It contains a number if decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. This algorithm is capable of working on large dataset efficiently.



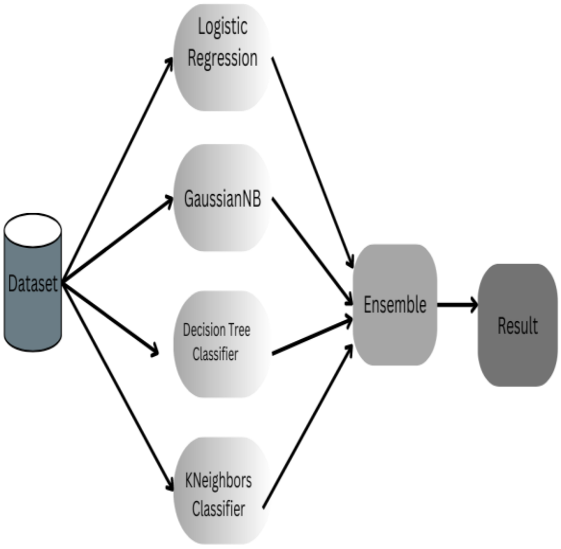
Ada-boost or Adaptive Boosting is one of the ensemble boosting classifier. It is an iterative ensemble method. Ada-boost classifier builds a strong classifier by combining multiple poorly performing classifier so that we will get high accuracy strong classifier. The basic concept behind Ada-boost is to set the weights of classifiers and training the data sample in each iteration such that it ensures the accurate prediction of unusual observations.



XG-Boost is another boosting algorithm in machine learning. It is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. In this algorithm, decision trees are created in sequential form. Weights play an important role as weights are assigned to all the independent variables which are then given to the decision tree which helps to predict the result. The weights of variables predicted wrong by the tree is increased and these variables are then fed to second decision tree. These classifier are then taken together to give a strong and more precise model. This algorithm can work on regression, classification , ranking and user- defined problems.



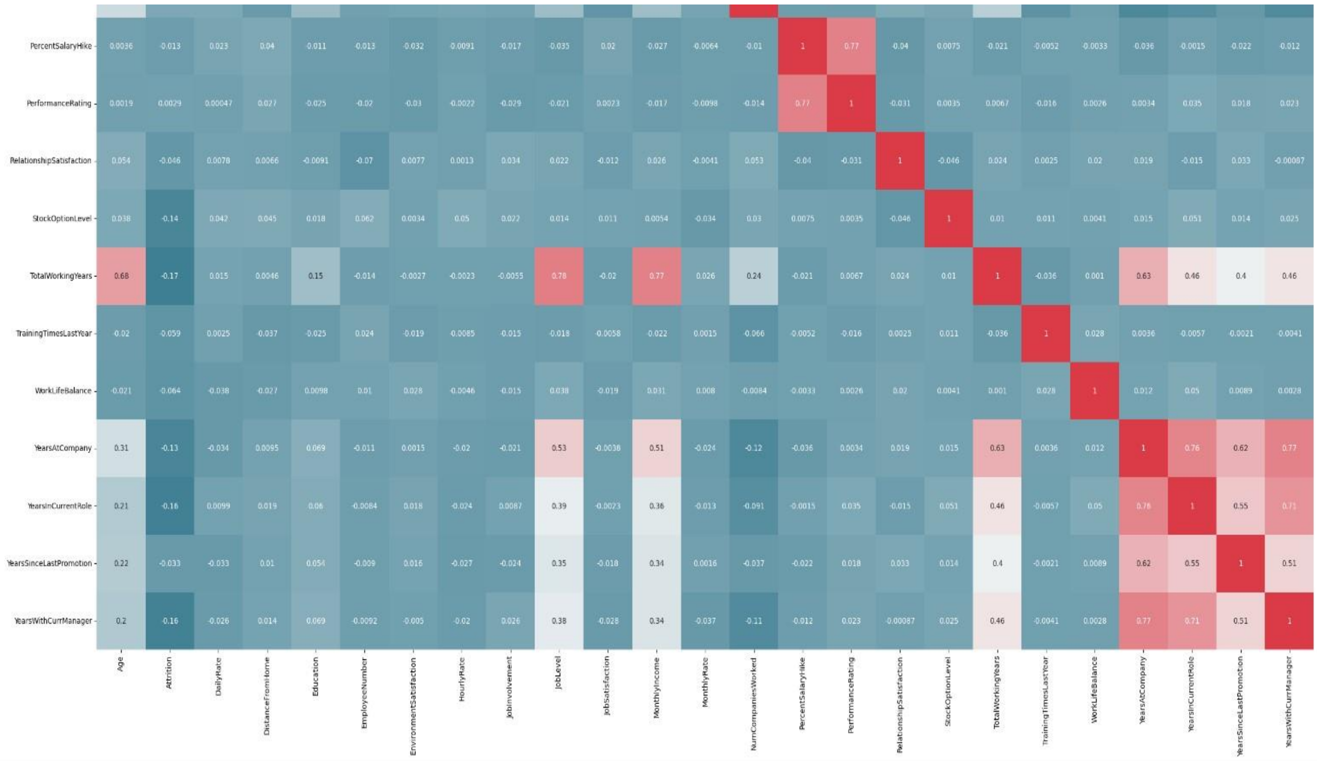
Ensemble learning helps to improve machine learning results by combining several methods. This approach alloys the production of better prediction and performance compared to a single model. The basic idea behind this approach is to learn a set of classifiers and allow them to vote. It’s advantage is that it gives majority vote, Bagging Randomness Injection, Feature selection, Error-correcting output coding. In ensemble classification we used stacking technique. Stacking involves training a model to combine the predictions of several other learning algorithms. First, all of the other algorithms are trained using the available data, then a combiner algorithm (final estimator) is trained to make a final prediction using all the predictions of the other algorithms as additional inputs or using cross-validated predictions from the base estimators which can prevent over-fitting. If an arbitrary combiner algorithm is used, then stacking can theoretically represent any of the ensemble Logistic Regression model is often used as the combiner improved predictive results although it is difficult to understand an ensemble of classifiers. Stacking typically yields performance better than any single one of the trained models. It has been successfully used on both supervised learning tasks (regression, classification and distance learning) and unsupervised learning (density estimation). It has also been used to estimate bagging's error rate.



Stacking of Classifiers

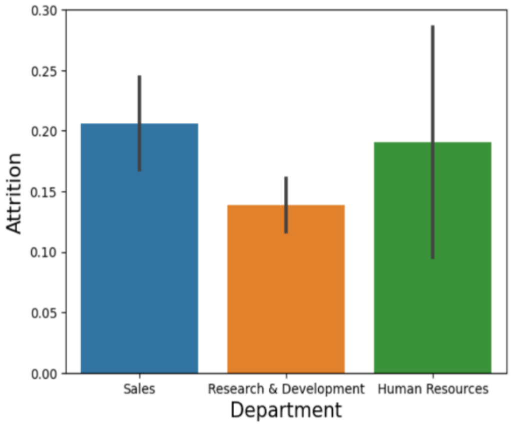
DETAILS

We created heatmap of the dataset which represents the coefficients to give a visualization of strength of correlation among variables.

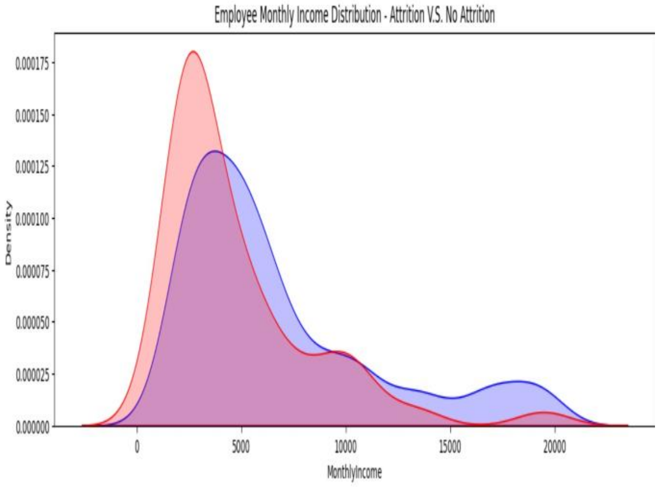


- We plotted various graphs with attrition common as one axis and various other factors.

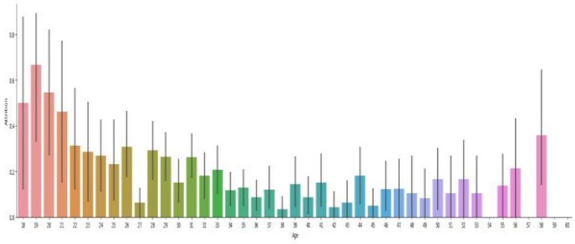
Department versus attrition:



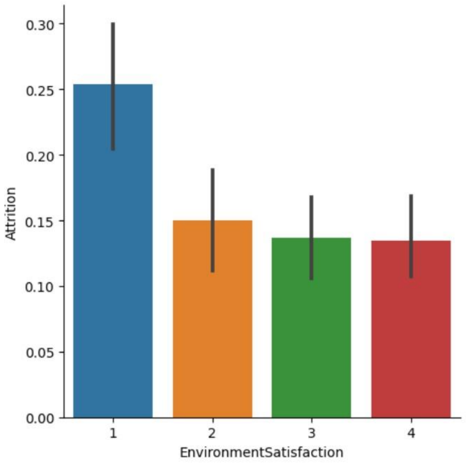
Monthly Income versus Attrition:



Age versus Attrition



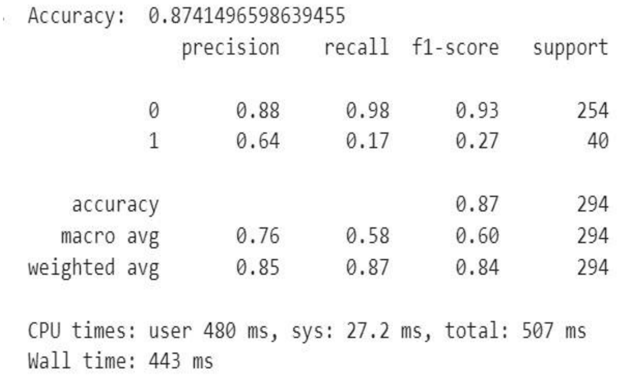
Environment Satisfaction versus Attrition



RESULTS

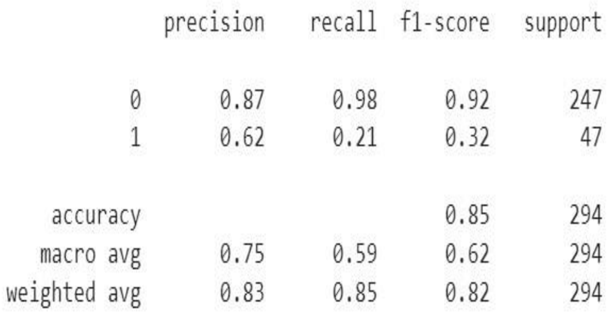
After data exploration and applying model algorithms, we compared them to find out the best suited algorithm for our dataset. The confusion matrix of the algorithms came out to be as follows.

Random forest Classifier:



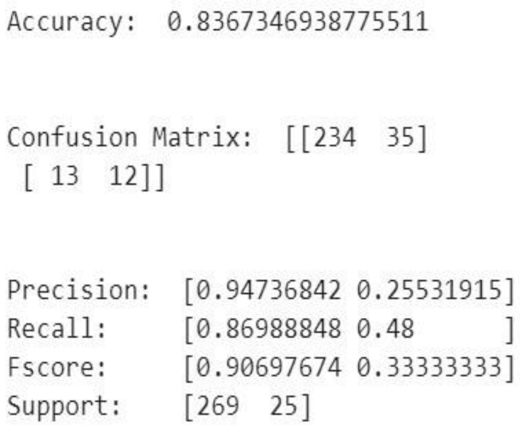
```
Accuracy:  0.8741496598639455
              precision    recall  f1-score   support

           0       0.88      0.98      0.93       254
           1       0.64      0.17      0.27        40

    accuracy                           0.87       294
   macro avg       0.76      0.58      0.60       294
weighted avg       0.85      0.87      0.84       294

CPU times: user 480 ms, sys: 27.2 ms, total: 507 ms
Wall time: 443 ms
```

XGBoost Algorithm:



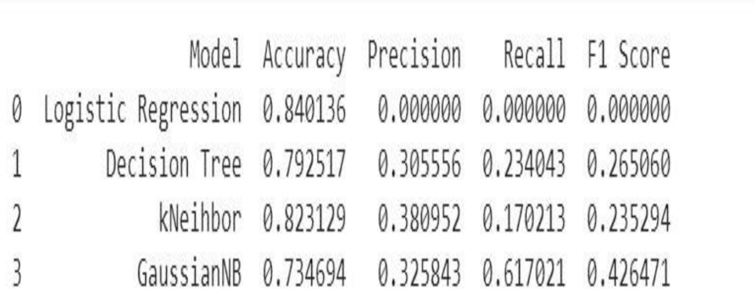
```
              precision    recall  f1-score   support

           0       0.87      0.98      0.92       247
           1       0.62      0.21      0.32        47

    accuracy                           0.85       294
   macro avg       0.75      0.59      0.62       294
weighted avg       0.83      0.85      0.82       294
```

Ada-Boost Algorithm:



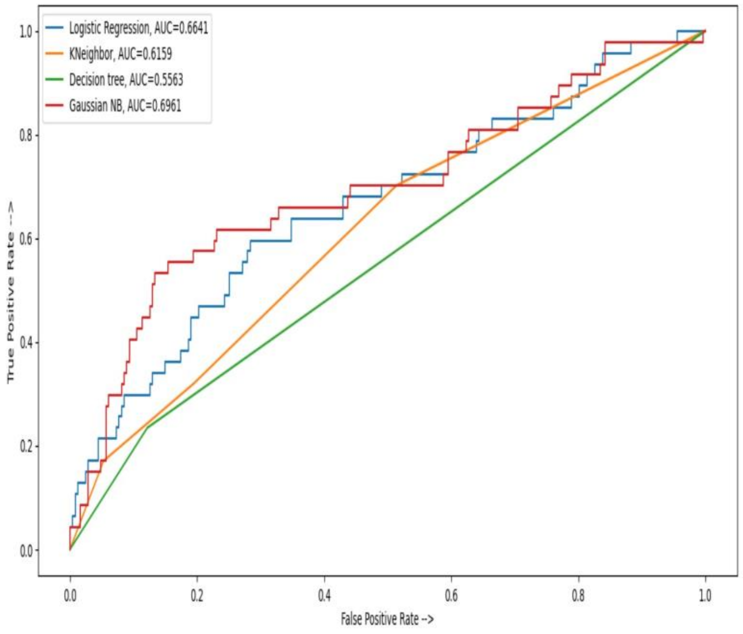
```
Accuracy:  0.8367346938775511


Confusion Matrix:  [[234  35]
 [ 13  12]]


Precision:  [0.94736842 0.25531915]
Recall:     [0.86988848 0.48      ]
Fscore:     [0.90697674 0.33333333]
Support:    [269  25]
```

Ensemble Stacking Algorithm:



	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.840136	0.000000	0.000000	0.000000
1	Decision Tree	0.792517	0.305556	0.234043	0.265060
2	kNeihbor	0.823129	0.380952	0.170213	0.235294
3	GaussianNB	0.734694	0.325843	0.617021	0.426471

An ROC Curve(Receiver Operating Curve) is a graph showing the performance of a classification model at all classification thresholds. This curve has plotted two parameters- true positive rate and false positive rate.



We used stack of classifiers to find the best suited algorithm. This list consists of random forest classifier ,Xgboost, Ada-boost and ensemble classifiers (logistic regression, K-nearest neighbour and decision tree).Data pre-processing is done using RFE(recursive feature elimination).We made a some graphs between various features like age, job satisfaction with attrition rate to get a rough idea about feature selection. After data exploration, we use models to find the best suited algorithm with high accuracy and precision for our dataset. As a result we found that Random forest proved to be the best suited algorithmfor the attrition analysis. It’s accuracy came out to be 87.41% with precision 88%, and weighted average 85%.

From the above graph plotted between various features it can be concluded that Monthly Income was one of the influential factors that made employees quit there jobs for better opportunities. Also, Sales Department had the most number of employees leaving.

Future Scope

In future research it is possible to improve the analysis by considering new employee opportunities as well as adverse working condition condition and poor promotion prospects, discrimination and low social support, that are positively related to employee's attrition intention. This paper would help in analysis of employee attrition of an organization and letting it know the reason for the same..

Conclusion

According the above model results, we can know that our finding are in line with people's behaviour in the real world and previous studies other scholar did. When attrition occurs within a firm, workload among existing members of the team increases with no increase in pay. This workload is even experienced by an HR professional. The potential for employment promotion may not exist owing to the position retired due to attrition.

We can conclude that it is a issue which is to be taken under consideration. Job satisfaction, daily rate, employee number and age were some of the factors that led to attrition in an organization. It was also observed that sales department had the most employee attrition rate. After knowing the factors the organization can take necessary steps to avoid large number of their employees leaving the company and reduce attrition rate.

References

[1] Francesca Fallucchi, Marco Coladangelo, Romeo Giuliano and Ernesto William De Luca "Predicting Employee Attrition using Machine Learning Techniques"(2020)

[2] Rohit Punnoose and pankaj Ajit."Prediction of employee Turnover in organization using Machine Learning Algorithms".In: International Journal of Advanced Research in Artificial Intelligence (2016).

[3] Shenghuan Yang(Jiangxi University of Finance and Economics),Md. Tariqul Islam(Syracuse University) IBM Employee Attrition analysis.(2021)

[4] I Setiawan*, S Suprihanto, A C Nugraha and J Hutahaean Department of Computer Engineering and Informatics, Politeknik Negeri Bandung, Bandung, Indonesia "HR analytics: Employee attrition analysis using logistic regression" IOP Conference Series(2019)

[5] Alao D. and Adeyemo A.B. Department of Computer Science, University of Ibadan,Nigeria, "Analyzing Employee Attrition Using DecisionTree Algorithms" Computing, Information Syatems and Development vol.4(2013).

[6] Dr. R. S. Kamath | Dr. S. S. Jamsandekar | Dr. P.G. Naik "Machine Learning Approach for Employee Attrition Analysis" Published in International Journal of Trend in Scientific Research and Development (ijtsrd),Fostering Innovation, Integration and Inclusion Through Interdisciplinary Practices in Management, March 2019,URL: https://www.ijtsrd. com/papers/ijtsrd 23065.pdf.

[7] Ali Raza, Kashif Munir,Mubarak Almutairi,Faizan Younas and Mian Mhd. Sadiq Fareed. "Predicting Employee Attrition Using Machine LearningApproaches". Published on June 2022.

URL: https://doi.org/10.3390/app12136424.

[8] Linkfordatasethttps://www.kaggle.com/datasets/thedevastator/employee-attrition-and-factors.

[9] El-Rayes, Nesreen, et al. "Predicting employee Attrition using Tree- based Models." International Journal of Orgaizational Analysis (2020).

[10] Reference for recursive feature elimination

https://machinelearningmastery.com/rfe-feature-selection-in-python/

[11] Saeed Najsfi-Zangeneh, Naser Shams-Gharneh, Ali Arjomandi-Nezhad and Sarfaraz Hashemkhani zolfani. "An Improved Machine Learning Based Employees Attrition Prediction Framework with Emphasis on Feature Selection(2021).

[12] Praphula Kumar Jain, Madhur Jain and Rajendra Pamula "Explaning and predicting Employee's attrition: A machine learning Approach."(2020).