University of Essex

Department of Mathematical Sciences

MA981: DISSERTATION

Prediction of Employee Attrition

Lakkakula Dheeraj
2202392

Supervisor: Dr Yanchun Bao

August 25, 2023
Colchester

Abstract

Employee attrition has become a critical concern for organizations because that not only impacting organizational and workforce stability, effectiveness and performance but also when employee leaves the organisation or a company, there will be a loss of talent and expertise. Apart from this attrition leads to significant costs and time in terms of recruiting, training new staff and very less productivity. In order to address this challenge proactively, HR analytics and predictive modelling with the power of data and machine learning has become essential tools.

This dissertation aims to apply the advanced machine learning and predictive models to analyse the HR data and predict the attrition rate accurately. By understanding the patterns and factors that lead to attrition, organisations can take data-driven decisions and actions to enhance employee engagement and reduce attrition rate. Based on these patterns and decisions the organisations can implement the retention strategies to maintain motivated and resilient work environment.

Overall, this study demonstrates the potential of predictive analytics in addressing the modern challenge of employee attrition. The use of machine learning models coupled with HR data provides a potent solution, allowing businesses to transition from reactive to proactive talent management and provides a blueprint for leveraging data science predictive modelling to understand factors that are causing employees to leave.

Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

In today's fast-paced and emerging business world , the departure of valuable employees can lead to operational disruptions, lowers team spirit and overall organizational progress. As employees exit, in-demand skills and irreplaceable institutional knowledge is lost. Moreover, the costs of recruiting, onboarding and training new hires keep mounting with excessive attrition rate.

To deal with this problem effectively, the data driven approaches with the help of HR analytics to comes to picture . HR analytics is also known as talent analytics, this involves the collection and analysis of the data related to the Employee performance, Engagement, Attrition, Retention, Absent rate, Turnover and plays a major role in organizations by finding the underlying patterns and foreseeing trends related to human capital management.

One of the applications of HR analytics is predicting employee attrition. Organizations are mainly concentrating in identifying the drivers of attrition and the ways to effectively reduce it. These contributing factors have become crucial objectives for organizations to enhance their human resource strategy.

This study demonstrates the power of using HR analytics with predictive modelling. By training algorithms on multidimensional organizational data including employee engagement in the work , performance, culture and more. Techniques including neural networks, random forests, support vector machines, logistic regression, decision trees and gradient boosting algorithms such as XG Boost, cat Boost uncovers the complex patterns in historic data that foretell future turnover or attrition rate.

Traditionally, attrition has been a reactive issue, where organizations respond to departures after they happen. However, Now most of the organizations looking for proactive approach, enabling the prediction of employees who may be at a higher risk of leaving. This empowers organizations to implement targeted retention strategies, effectively addressing attrition and promoting a more stable and engaged workforce. similar to how predictive diagnostics revolutionized healthcare, in the same way predictive HR analytics stands to transform organizational resilience and agility.

1.1 Aim and objectives

The main aim of this study is to build an efficient HR analytics system that could be able to identify attritions among the employees of the company based on machine learning and deep learning algorithms.

The objective of the study is to effectively analyse the data set based on statistics and visualisation.To implement and evaluate the performance of machine learning models such as logistic regression, support vector machine, decision tree and random forest algorithm. To implement boosting based machine algorithms and compare its performance with the above mentioned algorithms. To implement neural network architecture on HR analytics data and evaluate its performance with both boosting algorithms and machine learning algorithms.

CHAPTER 2

LITERATURE REVIEW

2.1 Relation between employee and the company

Employees are one of the most vital assets and resources for any organization. They are the human capital that executes the organizational strategy, delivers products or services, and ultimately generates revenue and value. As such, the performance and productivity of employees has a very direct impact on overall company performance. Organizations with higher performing talent and human capital tend to exhibit greater success in meeting objectives and outpacing the competition [Harter et al., 2002].

In essence, employees are the most fundamental contributor to organizational success. Their productivity and retention determines the market competitiveness and achievement of strategic goals. Companies must invest in engaging and retaining top talent to maximize value creation and performance. The symbiotic relationship between employees and the company means their fates are intertwined.

2.2 Employee Attrition and its impact on the company

This commonly known as turnover, refers to the phenomenon of employees leaving an organization voluntarily or involuntarily. The consequences of attrition are multifaceted and reach far beyond the simple act of filling vacant positions. It has a profound impact on an organization's functioning, performance, and overall health. The costs associated with turnover include tangible expenses related to replacement hiring such as recruitment, selection, onboarding, and training new employees. Studies estimate these direct costs can amount to 90-200% of the departed employee's salary [Allen et al., 2010]

In addition to direct costs, attrition also leads to reduced productivity as new hires take time to acquire skills and get acquainted with processes [Park and Shaw, 2012]. Losing employees also means losing their accumulated organizational knowledge, which can disrupt operations and service quality. Excessive churn can strain and demotivate remaining staff, hurting engagement and performance [Trevor and Nyberg, 2008]. Minimizing turnover is thus essential for business success.

2.3 Role of HR in Preventing Employee Attrition

Human Resources (HR) departments hold a pivotal position in organizations when it comes to managing and mitigating employee attrition. It also follows some effective initiatives to engage employees, foster positive workplace culture, and proactively retain top talent. This prevents the organization from losing high-value human capital and skills to turnover [Das, 2013].

HR helps in minimizing the attrition through several approaches by conducting stay interviews gives insights into why employees remain, enabling tailored retention policies targeting specific needs like flexibility or career growth [Sharma and Stol, 2019]. Training programs demonstrate investment in employee development, boosting engagement and retention [Imran et al., 2017]. HR is also crucial for championing initiatives like succession planning, competitive compensation, and benefits which boost embeddedness. Increasingly, HR is leveraging data-driven analytics to predict turnover risks and inform evidence-based attrition management programs [Rohit and Ajit, 2016]. Adoption of AI and machine learning allows continuous enhancement of attrition forecasting capabilities.

2.4 Recent Research on Employee Attrition

The primary aim of this study is to leverage HR analytics and predictive modelling techniques to conduct a comprehensive analysis of employee data with the ultimate goal of predicting attrition within organizations. Numerous studies have been conducted in this field, exploring the patterns and behaviour that leading to attrition. By systematically analysing the existing literature, this study aims to improve comprehension of the fundamental factors influencing attrition across diverse organizational settings.

There is growing academic interest in leveraging machine learning and predictive analytics to model and forecast employee attrition. Recent studies have applied supervised learning algorithms like decision trees, random forests, logistic regression and neural networks to HR data to identify turnover patterns and risks [Zhao et al., 2018]

A recent study conducted at Lanson Toyota in Chennai, India [Priya and Harasudha, 2017] has explored the causes of employee attrition. They surveyed around 100 employees and conducted statistical analysis to analyse the factors behind attrition. The findings in this study highlighted the issues such as lack of promotions for experienced staff, less compensation and benefits, limited growth opportunities and high stress levels as the primary drivers of employee departures. The study suggested implementing retention strategies like rewards programs, skill development, and exit interviews .This research emphasized the importance of addressing promotion opportunities, compensation, growth prospects, and stress management to mitigate attrition at Lanson Toyota.

Employee turnover poses significant financial challenges for organizations. According to [McFeely and Wigert, 2019] replacing a single employee can cost anywhere from 50% to 200% of their salary. However, the real cost of turnover extends beyond direct expenses. Projects may suffer delays due to the loss of expertise, and institutional knowledge departs with employees, disrupting company culture. Remaining staff must cover work gaps, leading to reduced productivity. On average, it takes approximately 42 days to fill open roles, incurring additional recruitment and onboarding costs. Employee attrition significantly impacts an organization's financial bottom line and operations. Retention initiatives should be prioritized to prevent turnover and its associated financial burdens.

The conceptual paper given by [Belete, 2018] focuses on summarizing factors influencing employees' turnover intention, including job satisfaction, job stress, organizational culture, commitment, salary, justice, promotional opportunities, demographics, leadership styles, and organizational climate. Employee turnover presents a major challenge for organizations, especially with high-performing employees. Other studies, [Victoria and Olalekan, 2016] have shown mixed results regarding gender's relation to attrition, while demographic factors like age, marital status, tenure, wage, position, and department have been identified as determinants of turnover intention.

[Mansor et al., 2021] conducted a comparative study on IBM Human Resource Analytic Employee Attrition and Performance, evaluating SVM and ANN classification models. After parameter tuning and regularization, SVM achieved the highest accuracy, RMSE, and speed. The data quality report of this study revealed that class distribution imbalance, addressed using SMOTE to create synthetic examples for balanced data. In order to improve the efficiency of the employee attrition model,[Yahia et al., 2021] utilizes two feature selection methods Recursive Feature Elimination (RFE) and SelectKBest. This two-step feature selection process enabled creating an optimized attrition model containing only the most important features. Apart from [4] also implements various machine learning, ensemble learning and deep learning models, in that the ensemble voting classifier performs best with 0.96 accuracy on IBM HR dataset.

Random Forest (RF), an extensively used supervised ML algorithm for classification and regression.[Krishna and Sidharth, 2022] developed an employee turnover prediction model using Random Forest which is known for its accuracy and handling outliers. Key influencing factors were monthly income, compensation ratio, total working years, age, and job involvement. The Random Forest model outperformed other classifiers on IBM HR dataset (1470 samples, 35 features), providing insights to reduce turnover.

[Setiawan et al., 2020] developed a model using logistic regression to analyse employee attrition. The study used HR data on over 4000 employees across one year at a firm. After data preprocessing, exploratory analysis identified key variables related to attrition including department, education field, work hours and work-life balance. 70% of data was used to train logistic regression models which were refined based on model evaluation metrics. The final model had 11 significant variables including number of companies worked, total working years, business travel frequency, job satisfaction and marital status. Model testing showed 75% accuracy, 73% sensitivity and 75% specificity. Key conclusions were that divorce and married employees had higher attrition versus single, and HR department, low satisfaction, early logout and overtime impacted attrition. [Setiawan et al., 2020] recommend improving HR policies and environment to reduce voluntary turnover.

[Fallucchi et al., 2020] analysed factors influencing employee attrition to identify reasons for turnover and predict which employees may leave. The study used an IBM HR dataset. Data preprocessing included cleaning, transcoding categorical variables, and 70:30 split for training and testing sets. Exploratory analysis found monthly income, age, overtime, and distance as top attrition factors. Multiple classifiers were trained including Naive Bayes, logistic regression, SVM and random forest. Gaussian Naive Bayes had the best recall at 0.54, detecting most employees likely to churn with only 4.5% false negatives. Key findings were lower salaries, younger ages, more overtime, and longer commute increase attrition risk. The authors recommend HR departments leverage predictive modelling and data analysis to gain insights into employee turnover.

[Rombaut and Guerry, 2018] investigated predicting voluntary employee turnover using only data available in HR databases, without supplementary surveys. Logistic regression and decision tree models were applied on a Belgian company's HR data (13,484 cases over 11 years). Significant predictors found were gender, age, seniority, marital status, nationality, salary, work hours, company car and phone. Women had lower turnover probability, aligned with research showing higher female satisfaction. Increasing seniority first increased then decreased churn risk, echoing prior evidence of a U-shape. Company car predicted higher turnover contrary to pay satisfaction literature. Overall model AUC was 0.74, indicating HR data does enable turnover prediction. The authors recommend expanding HR databases and propose the approach as a barometer for identifying employee risk groups.

[PM and Balaji, 2019] compared machine learning techniques for predicting employee attrition using an IBM HR dataset. After data preprocessing, classifiers like J48 decision tree, Naive Bayes, and clustering algorithms k-means and Expectation Maximization were tested. J48 gave 82.4% accuracy with 10-fold cross validation and 82.76% with 70:30 split, outperforming Naive Bayes at 78.84% and 80.95% respectively. For clustering, k-means had 57.28% accuracy versus 55.1% for EM, but was faster. The decision tree visualization provided insights on attrition factors. The authors recommend comparing methods in Python and studying dynamic employee behaviour features. Overall, the study demonstrated machine learning's potential for HR analytics problems like predicting churn.

[Shankar et al., 2018] develops classification models using techniques like decision trees, SVM, logistic regression and KNN to predict employee attrition from HR data. In this mainly KNN is applied on test data to cluster the employee by department and then detailed analytics is done within each department to predict the risk of attririon. Data mining and Hybrid approaches combining sampling , cost- sensitive learning are used and this further improves the prediction performance

2.5 Recent Advances in Boosting Techniques for Attrition Forecasting

The advanced boosting algorithms such as XGBoost , CatBoost, LightBgm have shown significant improvements in employee attrition prediction compared to traditional classifiers. They are able to model complex attrition behaviors, handle imbalanced employee data, and provide interpretations around important factors.

Contemporary studies reveal the aptitude of boosting methods like XGBoost and AdaBoost for predicting employee turnover through HR analytics. XGBoost demonstrates high churn predictive accuracy owing to its scalability and built-in regularization [Fallucchi et al., 2020]. Adaptive boosting improves HR classification by iteratively focusing on misclassified cases during training [Qutub et al., 2021]. The adaptive ability of boosting algorithms makes them well-suited for handling class imbalance in churn data. Their ensemble approach produces strong predictive models combining multiple weak base learners. Overall, boosting techniques show immense promise for enhancing predictive insights into employee retention.

2.6 Emerging Applications of Deep Learning for Attrition Modelling

With exceptional representational capacity, deep neural networks are gaining traction for advanced modelling of employee retention risks. Novel deep architectures have been designed and evaluated specifically for the churn prediction task.

Recently [Hang et al.,]has explored using graph neural networks and recurrent models like LSTM for talent management applications including employee turnover prediction. Combining these graph modelling and recurrent networks provides a robust approach. The graph network component models organizational structure and interconnections. The recurrent network tracks individual temporal behaviours. Together they provide a comprehensive employee representation incorporating relationships, time trends, and past activities. [Hang et al.,] have shown effectiveness on using graph neural networks and LSTM models for tasks like predicting employee churn risks and modelling organizational dynamics.

[Dutta and Bandyopadhyay, 2020] has utilized employee dataset from Kaggle which has

1470 records and 35 features and proposes a feedforward neural network model to predict employee attrition probabilities using HR data. The model has an input layer, a hidden layer and an output layer with ′relu′ and ′sigmoid′ activation functions. 10-fold cross validation is used to evaluate the model where training is done on 9 folds and 1 fold is held out for testing in each iteration. The average testing accuracy over all folds is calculated as performance metric The neural network model achieves a cross-validation testing accuracy of 87.01% outperforming classifiers like SVM, KNN, Decision Trees etc. The study [Dutta and Bandyopadhyay, 2020] shows deep learning models can effectively predict employee attrition for proactive retention planning .

CHAPTER 3

DATASET DESCRIPTION

The dataset utilized in this study was sourced from open-source Kaggle website [Vijay, 2018] consisting of anonymised data from an imaginary company. It is a comprehensive HR dataset including the manager survey data and employee survey data which contains the data relates to the demographic attributes of employees and feedback given by most employees and managers and login and logout times of the employees and also incorporates the details on employee performance indicators and attrition status. The data comprises five files outlining:

- **general_data.csv**
- **employee_survey.csv**
- **manager_survey.csv**
- **in_time.csv**
- **out_time.csv**

The **general_data.csv** dataset contains comprehensive information about 4400 employees, capturing their demographic details such as age, gender, marital status, monthly income etc, job-related characteristics, performance metrics, attrition status and it has 24 attributes of employees as shown in the below table [3.1]

Attribute	Description
Age	Age of the employee
Attrition	Employee's attrition status (whether they left or stayed)
Business Travel	Frequency of business travel
Department	Department in the company where the employee works
Distance From Home	Distance of employee's home from the workplace
Education	Employee's education level
Education Field	Field of education of the employee
Employee Count	Count of employees
EmployeeID	Unique identifier for employees
Gender	Gender of the employee
Job Level	Employee's job level
Job Role	Role or position of the employee
Marital Status	Marital status of the employee
Monthly Income	Monthly income of the employee
Num Companies Worked	Number of companies the employee worked at
Over18	Whether the employee is over 18 years old
Percent Salary Hike	Percentage increase in salary
Standard Hours	Standard working hours
Stock Option Level	Level of stock options held by the employee
Total Working Years	Total years of work experience
Training Times Last Year	Number of times the employee received training last year
Years At Company	Number of years the employee has worked at the company
Years Since Last Promotion	Number of years since the last promotion
Years With Curr Manager	Number of years with the current manager

Table 3.1: General Data Attributes

In **employee_survey.csv** dataset contains 4410 survey responses from employees, focusing on employee's perceptions of work environment satisfaction, work-life balance etc. This dataset as shown below [3.2], mainly focuses on employee's subjective feelings about their work environment.

Attribute	**Description**
EmployeeID	Unique Identifier of the Employee
Environment Satisfaction	Employee's level of satisfaction with the work environment
Job Satisfaction	Employee's level of job satisfaction
Work-Life Balance	Employee's perception of work-life balance

Table 3.2: Employee Survey Data Attributes

The **manager_survey.csv** dataset consists of survey responses from managers providing insights on team members performance and their leadership styles of nearly 4000 employees and has 3 attributes. This dataset as shown in table [3.3], gives the managerial perspective on the employee engagement and performance evaluation within the organization.

Attribute	**Description**
EmployeeID	Unique Identifier of the Employee
Job Involvement	Level of employee's job involvement according to managers
Performance Rating	Employee's performance rating as assessed by managers

Table 3.3: Manager Survey Data Attributes

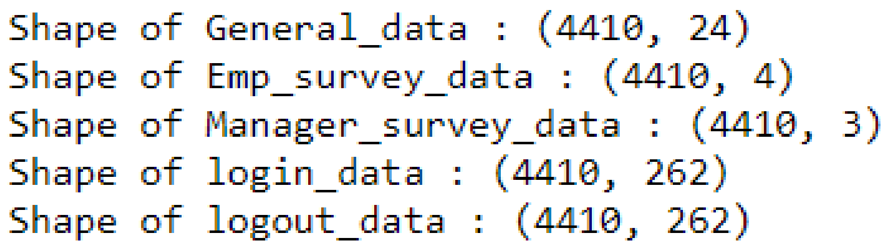
The **in_time.csv** dataset records employee's check-in times, reflecting their arrival at the workplace. This dataset covers a specific time period of year 2015 and includes 4410 instances and 262 different dates of year 2015.Similarly, the **out_time.csv** dataset records employeeâs check-out times, signifying their departure from the workplace. This also covers a specific departure time over 262 distinct dates of year 2015 of nearly 4400 employees. These two datasets help to track the employee working hours, punctuality and arrival or departure patterns of the employee.Overall, these five datasets collectively forming the HR Analytics case study by capturing the diverse aspects of employees, managers and workplace interactions during a specific timeframe to enable predictive modelling.

CHAPTER 4

DATA PREPROCESSING AND EXPLORATION

4.1 Data Gathering, Cleaning and Preparation

This is the first step, in this all the five data sets general_data.csv, employee_survey.csv, manager_survey.csv, in_time.csv and out_time.csv are read and shapes of each data set is as shown below figure [4.1]



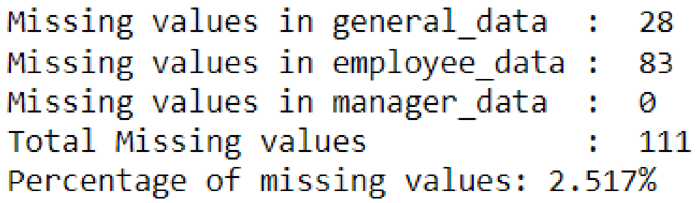
```
Shape of General_data : (4410, 24)
Shape of Emp_survey_data : (4410, 4)
Shape of Manager_survey_data : (4410, 3)
Shape of login_data : (4410, 262)
Shape of logout_data : (4410, 262)
```

Figure 4.1: Shape of all the datasets

Data cleaning is a critical initial step in any machine learning process. This involved activities like handling missing values, removing duplicate records, fixing structural errors, and handling outliers or anomalies. Proper data cleaning ensures high quality data, which is essential for training accurate machine learning models. Low quality data with issues like missingness , duplicates, errors etc can significantly degrade model performance. In this study, careful data cleaning is done and trained highly precise attrition forecasting models using HR analytics.

4.1.1 Handling Missing Values

Missing values can be replaced by various imputation methods. The proportion of the missing values of the used data set in this dissertation is shown in figure [4.2]



```
Missing values in general_data  :  28
Missing values in employee_data :  83
Missing values in manager_data  :  0
Total Missing values            :  111
Percentage of missing values: 2.517%
```

Figure 4.2: Percentage of Missing values

Since the proportion of missing values was minimal and the specific business context for the missing data was absent, meaningful imputation was difficult. Therefore, as per regular imputing practices, the most frequent data value (MODE) was used to impute missing categorical variables. For numerical variables, missing values were imputed with the MEAN or MEDIAN.

After checking for missing values, two columns in general_data 'NumCompaniesWorked' and 'TotalWorkingYears' consists null or missing values. Three columns in employee_data 'EnvironmentSatisfaction', 'JobSatisfaction', and 'WorkLifeBalance' contained missing values. The missing values in the numerical columns 'NumCompaniesWorked' and 'TotalWorkingYears' in general_data were imputed by the MEDIAN and MEAN respectively. In employee_data, the missing value columns were categorical, so they were imputed by the MODE.

This standard imputation enabled the training of machine learning models on the datasets without complications from null values.

4.1.2 Dealing with NaN values in login and logout datasets

The in_time and out_time datasets originally contained unique dates of the year as columns. First these were converted into standard datetime formats. There were many NaN values in both datasets. These were handled in two ways:

- If a column had all NaN values, it was considered a day-off for all employees and hence dropped.
- If a column had some non-NaN values, it meant only that particular employee was off. So, the NaN was replaced by 0 (in date-time format).

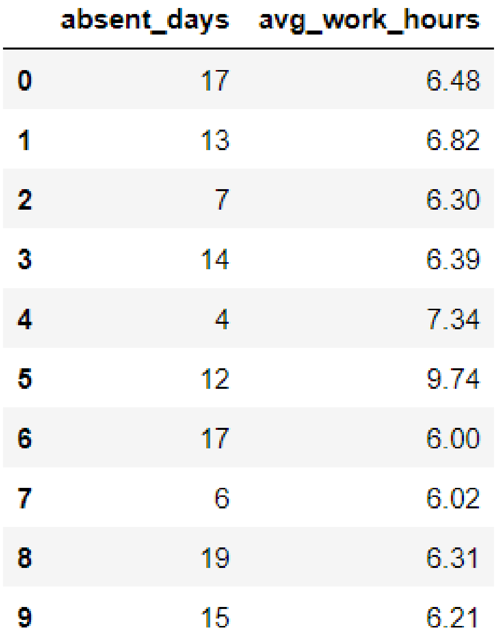
This preprocessing standardized all records to datetime and handled NULLs appropriately, whether indicating a company holiday or individual employee's day off. The cleaned in_time and out_time data could then be merged with other tables for further analysis.

4.1.3 Deriving the attributes related to absenteeism and working hours

The next crucial task was data preparation. The common 'EmployeeID' column which is the unique identifier of each employee across datasets enabled merging. The in_time and out_time data played a major role in data preparation. This involved combining and merging all the datasets, along with derived columns based on absenteeism and average working hours of the employee as given below with description. These columns were calculated by taking the difference between in_time and out_time datasets.

- **Absent days**: This means the employee is off on those days means if the difference login and logout time is 0, then it is considered as an absent day.
- **Average Working hours**: This is derived by taking the mean of total working hours of the employee.

These new attributes as shown in the above figure [4.3] helped compare employee performance. Absenteeism and average working hours indicated how engaged the employees were. Merging all data with these new columns prepared the integrated dataset for further predictive modelling and analysis.



	absent_days	avg_work_hours
0	17	6.48
1	13	6.82
2	7	6.30
3	14	6.39
4	4	7.34
5	12	9.74
6	17	6.00
7	6	6.02
8	19	6.31
9	15	6.21

Figure 4.3: Attributes that shows Absenteeism and working hours

4.1.4 Checking and handling the Outliers in the Data

As shown in the plot [4.4], the features like income, years of experience and number of past companies etc. have heavy skew with some employee data. These data points are known as Outliers which can adversely affect many statistical and machine learning techniques used in HR analytics.

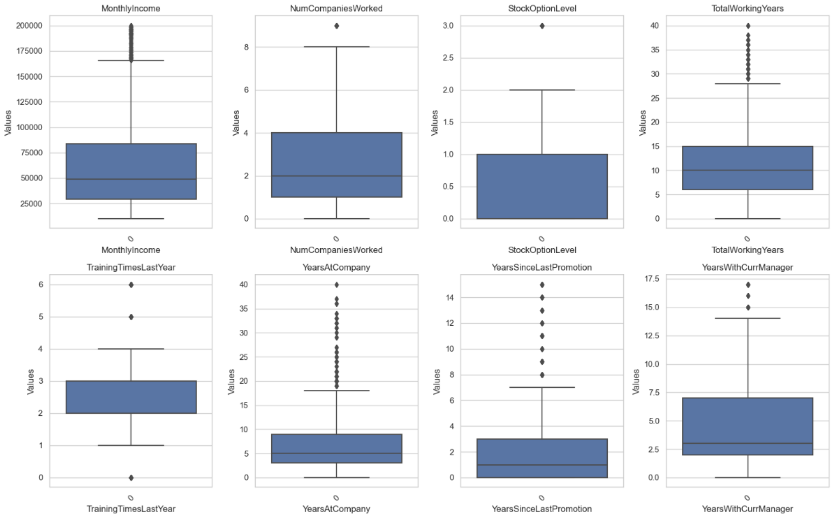


Figure 4.4: Outliers present in the data

For attrition modelling, outlier employees with inordinately high incomes may not represent general trends. The patterns which are modelled on these outliers may not generalize to normal employees. Yet excluding the outliers entirely also loses information. This will be best approach if the data is very huge. So instead that in this study **Log Normalization** method is used to temper the outlier effects. By log transforming the skewed features, the impact of outliers is compressed rather than removed such that there would not be any loss of data. So, this creates a more balanced attrition models.

As part of data transformation phase, categorical variables were transformed to have more meaningful labels. Attributes like Education, EnvironmentSatisfaction and PerformanceRating were mapped from numeric values to descriptive category names as shown in table [4.1] and converting the attributes to corresponding data types such as categorical and numerical . This data transformation allowed for easier analysis and interpretation of the categorical data during visualization and modelling.

Attribute	**1**	**2**	**3**	**4**	**5**
EnvironmentSatisfaction	Low	Medium	High	Very High	-
WorkLifeBalance	Bad	Good	Better	Best	-
PerformanceRating	Low	Good	Excellent	Outstanding	-
JobSatisfaction	Low	Medium	High	Very High	-
JobInvolvement	Low	Medium	High	Very High	-
Education	Below College	College	Bachelor	Master	Doctor

Table 4.1: Label descriptions of categorical variables

4.2 Data Exploration

Exploring the employee attrition dataset through diverse data visualization approaches provided multidimensional insights into the factors that influencing the attrition.

4.2.1 Correlation matrix (Numerical variables vs Attrition)

Correlation analysis mapped out associations between variables, highlighting the potential factors and links tied to attrition. The correlation matrix gives a broad overview of interrelationships and dependencies within the data. Variables strongly correlated to attrition represent prospective key drivers. This measures the linear relation ship between the numerical variables by showing that how strongly they are related and the direction of the association. The correlation coefficient ranges from -1 to +1.

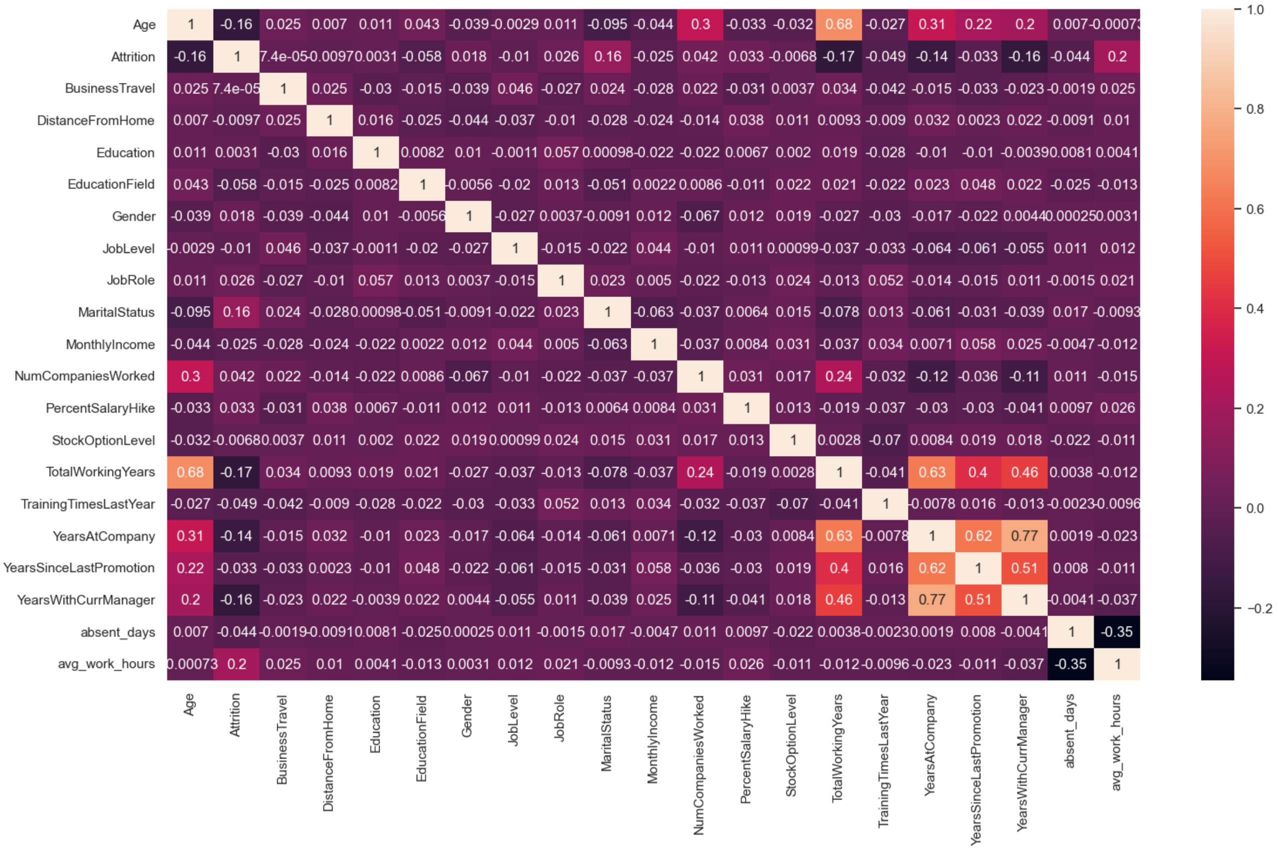


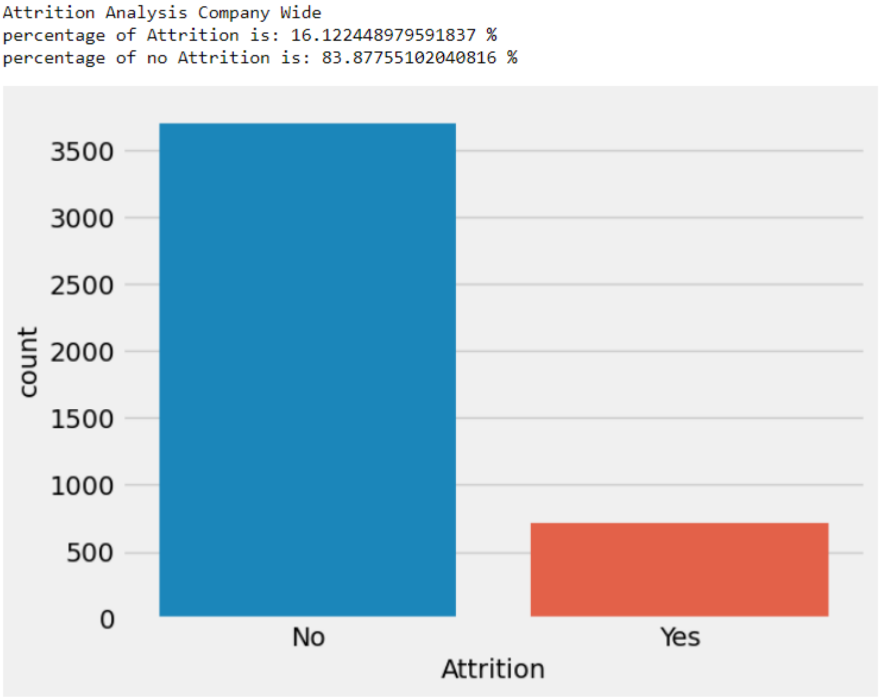
Figure 4.5: Correlation plot

The above correlation matrix shown in figure [4.5] displays the coefficients between all the variable pairs with ones along the diagonal as variable or attribute has perfect correlation with itself. The correlation of 0 indicates that there is no significant linear relationship like Employee age and attrition may have close to zero correlation.

A positive correlation means as one variable increases, so does the other. A correlation of +1 indicates a perfect positive linear relationship. For example, work hours and income may positively correlate. Sometimes there may be a case of negative correlation means as one variable increases, the other decreases. A correlation of -1 denotes a perfect negative linear relationship. For instance, distance from workplace and attrition may negatively correlate. So, in order to know the relations between the numerical variables and to know how strongly the numerical variables are associated with Attrition (Target variable).

4.2.2 Uni-Variate Analysis

Univariate analysis of each individual variable uncovered central tendencies, variability and outliers. Visualizations like histograms and box plots revealed characteristic distributions, trends and anomalies that could contribute to turnover. Univariate analysis elucidated the inherent properties and distributions of single attributes.



```
Attrition Analysis Company Wide
percentage of Attrition is: 16.122448979591837 %
percentage of no Attrition is: 83.87755102040816 %
```

Figure 4.6: Distribution of Attrition company wide

This the above distribution shown in figure [4.6] clearly shows that the attrition rate of the company is 16% and remaining is 84%. This may be considered as an unbalanced dataset or may be this huge difference between these categories depends on the nature of the data. In some cases, this unbalanced data lead to biased predictions and reduced performance on the minority class. To deal with these imbalanced datasets there are under sampling and oversampling techniques.

A density plot was generated to visualize the distribution of average working hours per employee shown in figure [4.7] . This plot depicts the proportion of employees working each hour range, spanning from 4 to 11 hours. The shape of the distribution indicates that the average work duration lies between 7 and 9 hours for a majority of employees, with a peak density around 8 hours. Very few employees work either below 5 hours or over 10 hours on average.

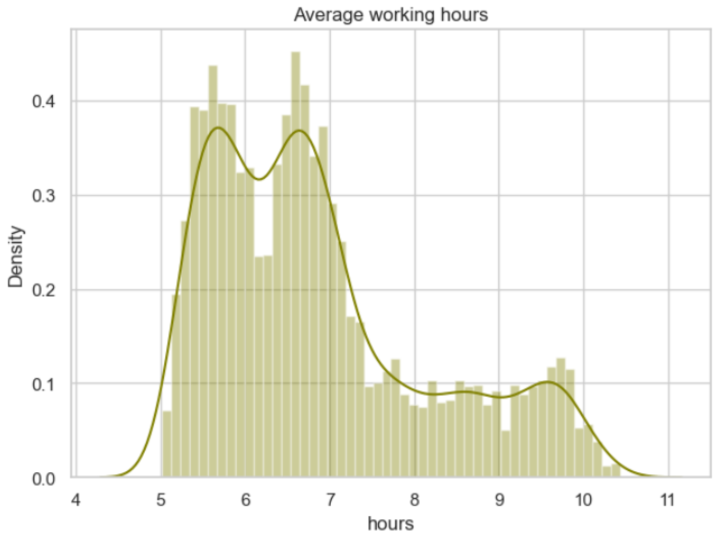


Figure 4.7: Distribution of Average working hours of employees

Insight into the average working time distribution allows identification of potential links with attrition. For instance, employees working longer hours may experience more work-life imbalance and consequently be prone to leaving. Inclusion of working hours data and analysis will enrich the capabilities of HR analytics systems.

In the below picture [4.8] , it shows the individual distribution of three categorical variables ’Environment Satisfaction’ , ’Job Satisfaction’ and ’Work life balance’. In the survey most of the employees are highly satisfied by the work environment and very highly satisfied by the job they are doing. Most employees stated that they have better work life balance in the company they are working.

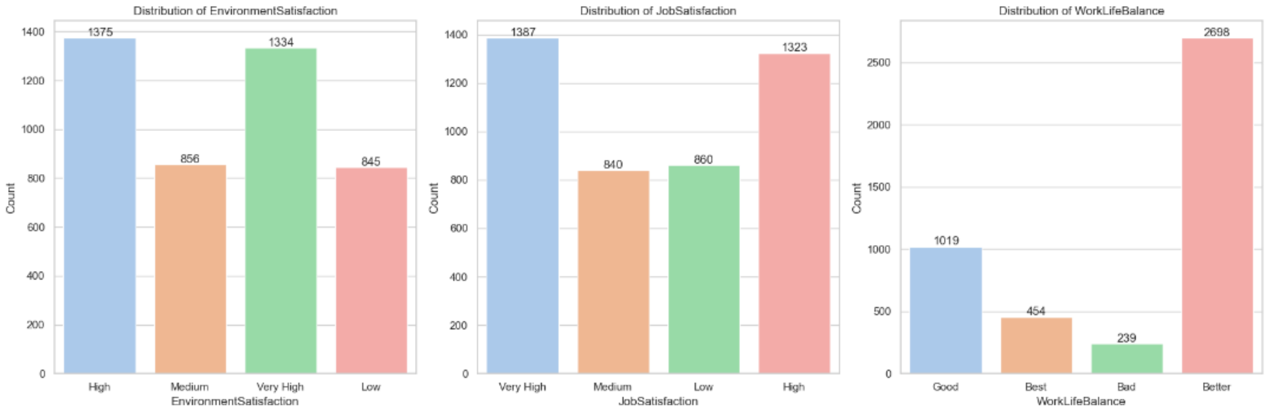


Figure 4.8: Distribution of attributes related to employee's prespective

4.2.3 Imbalanced Data

Many real-world datasets that are used for classification are mostly imbalanced datasets, where the classes are not represented equally. In the case of modelling employee attrition, the data is typically skewed, with significantly fewer samples of employees who left versus those still working. This imbalance poses a key challenge during training, as algorithms may maximize overall accuracy by focusing on the majority class, without learning the minority class patterns. For example, if only 20% of data represents attrition, a naive model can get 80% total accuracy by simply predicting every employee will stay, without considering any attributes. However, such a model is useless for identifying employees likely to churn. Imbalanced data leads to poor predictive performance for the minority positive class, which is often the class of interest, like churners.

There are two ways to handle the these imbalanced datasets, One is by undersampling the majority class and the other is oversampling the minority class. There are some Data level techniques like oversampling and under sampling can address imbalance and improve identification of the minority class. Some of the mostly used techniques such as SMOTE (Synthetic Minority over sampling), ADASYN (Adaptive Synthetic sampling), Ensemble methods and cluster based Over sampling methods. These techniques balances the class distribution, enhancing model learning of churn patterns.

In this study's case, with 16% attrition, the data is meaningfully imbalanced as shown in above figure. Without addressing the skew, predictive performance on actual turnover cases would be poor. Therefore, sampling methods were judiciously applied to enable accurate forecasting of the minority yet important employee attrition case.

4.2.4 Bi -Variate Analysis

Bi-variate analysis dug deeper into the interactions between variable pairings. Scatter plots, heat maps and other plots portrayed how joint dynamics of two variables relate to churn. In this it clearly shows how one variable affects the attrition of the employee.

To do this analysis, Attrition percentage is calculated to know how the attrition is varying with different variables.

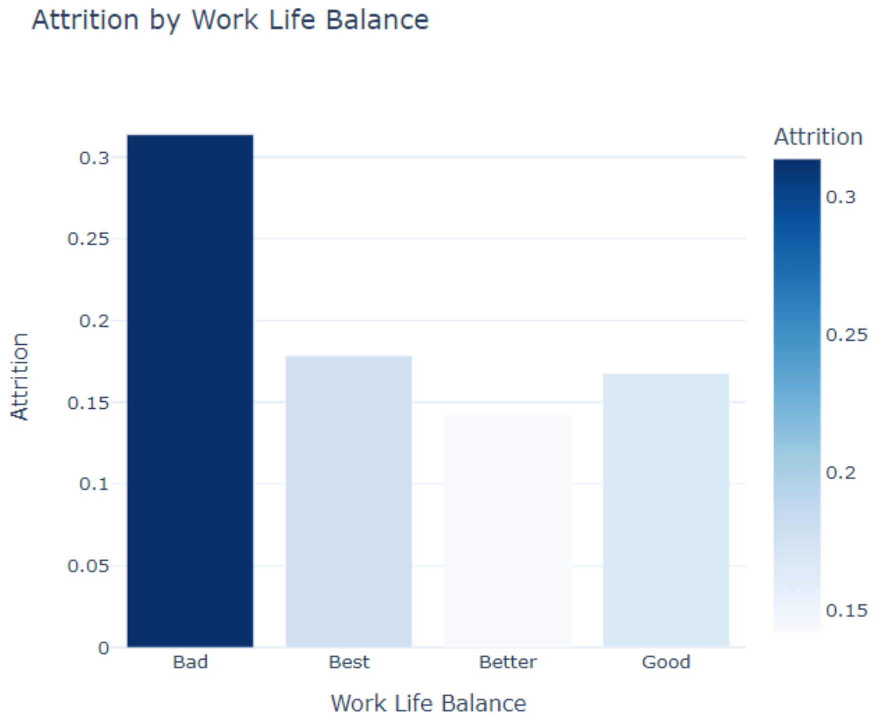


Figure 4.9: percentage of Attrition by Work-life balance of the employees

The above plot illustrates how attrition percentage varies based on employees' work-life balance. It clearly shows that employees rating their work-life balance as bad have the highest churn percentage at nearly 30%. In contrast, those with better work-life balance have lower attrition. This indicates that poor work-life balance strongly correlates with attrition. Work-life balance emerges as a potentially significant predictive factor for employee attrition.

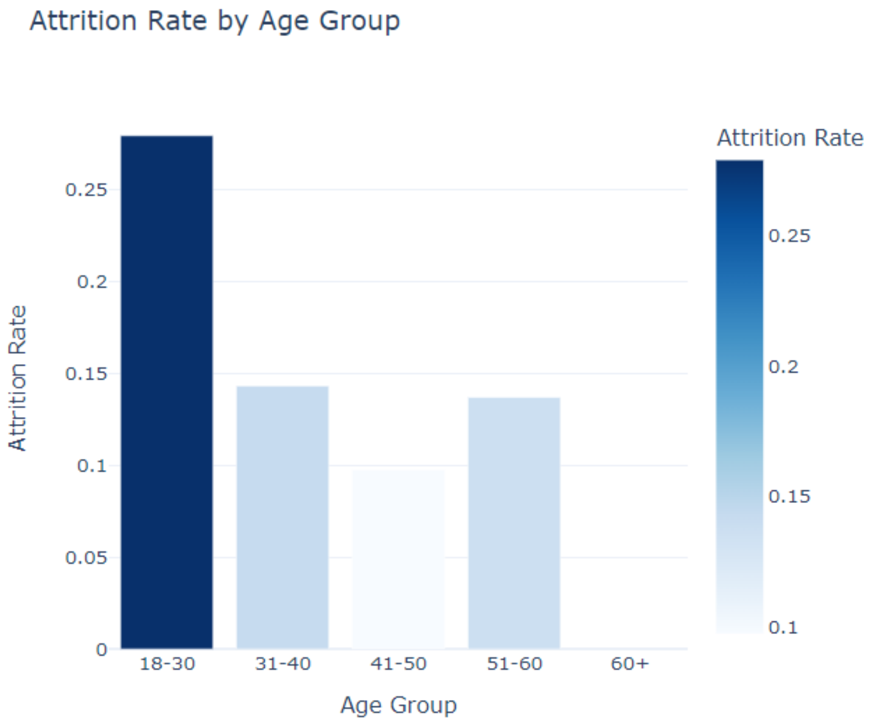


Figure 4.10: percentage of Attrition by Age Group of the employees

To better understand how employee age correlates with attrition, ages were grouped into ranges. Visual analysis as shown figure [4.10] was conducted on attrition percentage variation across these age groups.

By Segmenting ages into wider brackets uncovered overall attrition trends for different age groups. The 18-30 year bracket showed the highest churn percentage at nearly 27%, while the 41-50 year group exhibited very low turnover. Younger employees aged between 18 and 30 exhibited significantly higher attrition compared to mid-career staff aged 41-50. This large difference between younger and older groups suggests generational factors disproportionately affect younger workers. Limited pay growth and promotion opportunities may frustrate younger employees, driving their higher turnover. In contrast, mid-career workers likely have greater stability. So It clearly seen that younger age group requires some retention efforts like career development meetings and competitive compensation etc.

Similar visual analysis was conducted for the categorical variables EnvironmentSatisfaction and JobSatisfaction to understand their correlation with attrition.

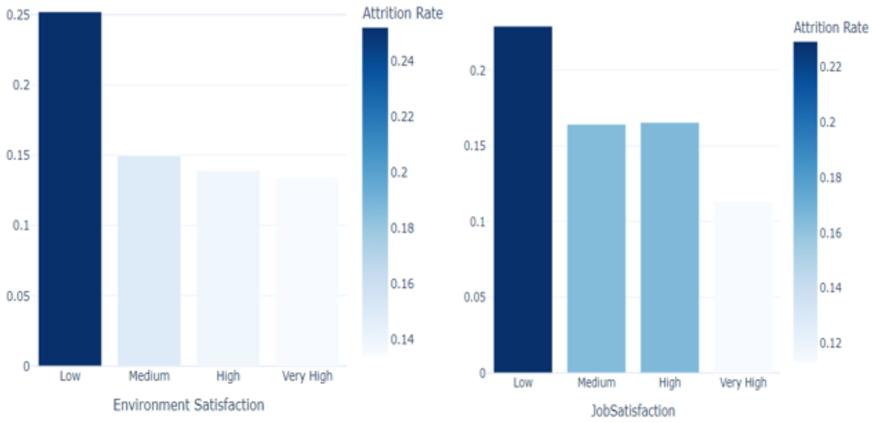


Figure 4.11: percentage of Attrition by Environment and Job Satisfaction of the employees

The above two plots shown in figure [4.11] illustrate that the employees who reported low satisfaction levels in either environment or job exhibited significantly higher churn rates. Those unsatisfied with their environment had approximately 25% attrition, while dissatisfied employees had around 22% turnover.

In contrast, highly satisfied employees in both dimensions showed very low attrition percentages. The sizable difference in attrition between low and high satisfaction groups indicates these factors' importance in predicting churn. By surfacing stark imbalances in attrition rates across satisfaction levels, the analysis spotlights environment and job satisfaction as likely drivers of turnover. Employees dissatisfied on either front represent a major flight risk requiring proactive retention efforts. Incorporating satisfaction metrics will enable more accurate attrition forecasting within HR analytics systems.

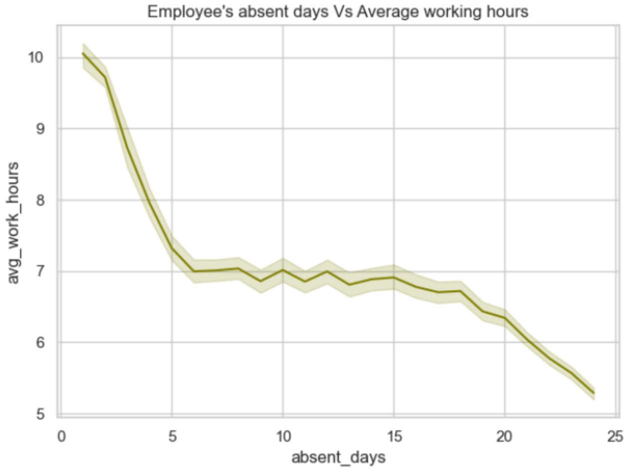


Figure 4.12: Line chart to depict the trend between absent days and average work hours

The above line chart shown in figure [4.12] depicts an inverse correlation between number of absent days and work hours, regardless of attrition status. As absent days decrease, the average working hours increase. This trend indicates employees who take less time off tend to work longer hours. Additional focussed analysis is required to check the correlation between the absenteeism and work hours with attrition.

The box-plot for number of absent days versus attrition in the below figure [4.13] shows minimal difference between employees who left and those who stayed . Employees with churn=Yes only exhibit slightly fewer absent days on average compared to retainers. The inter quartile ranges highly overlap, and medians are similar between the attrition categories.

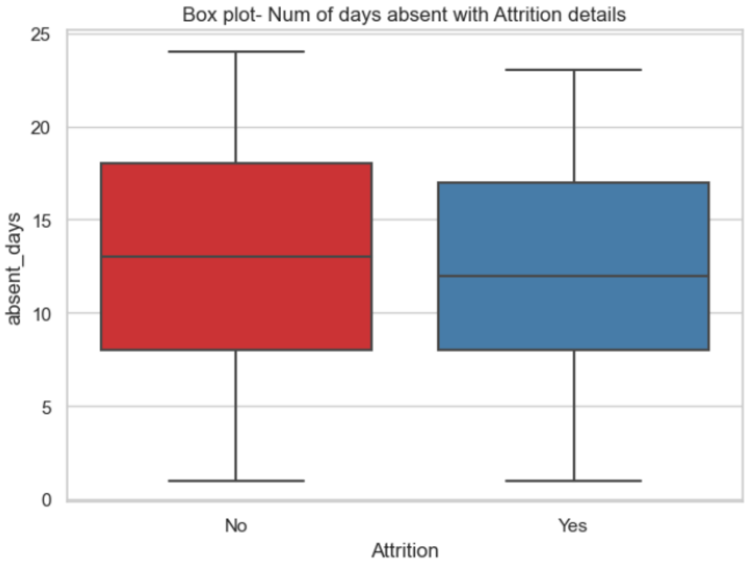


Figure 4.13: Boxplot - AbsentDays VS Attrition

Overall, the visualization indicates negligible correlation between absenteeism and propensity to leave the company. The minor difference observed which is likely not meaningful. So, Number of days absent does not emerge as an indicative factor in predicting employee attrition based on this HR data.

The below box-plot visualizations shown in figure [4.14] reveal interesting insights into how tenure and work hours correlate with attrition. Employees who left the company had markedly lower average tenure than retainers. The yearsAtCompany distribution for churners peaks under 3 years, while retainers peak over 7 years. This indicates employees with over 5 years of experience are much less inclined to turnover.

However, average working hours exhibit the opposite pattern. Employees working longer hours have higher attrition than those working fewer hours on average. The average working

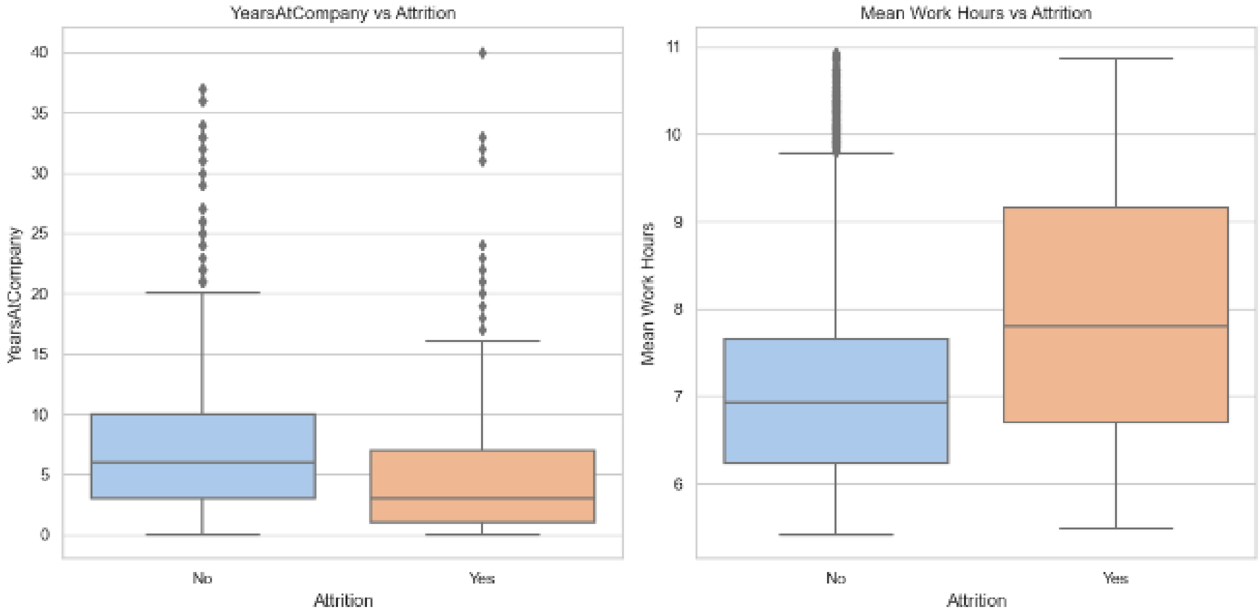


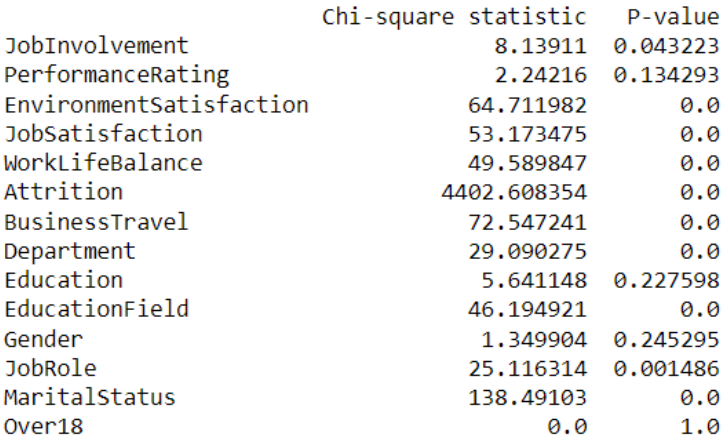
Figure 4.14: Boxplot - YeasrsAtCompany and Mean working hours VS Attrition

hours distribution for churners skews longer, while retainers peak at shorter durations. Factors like work stress, burnout and work-life imbalance likely contribute to higher turnover among employees working longer hours.

Together, all these visualization techniques provided a thorough understanding of the attrition problem. The Correlation analysis mapped the big picture . Uni-variate analysis mainly focused on how the individual variable is distributed. Bi-variate analysis explored the intersection of combined variables.This analysis clearly shows that what are the most important factors that are leading to attrition. This integrated visual perspective provided the insights needed to craft impactful attrition prediction strategies. As the correlation analysis gives how all the numerical varaibles related to attrition but the data also comprises of categorical variables which leads to attrition inorder to check this in this study chi-square test is carried out.

4.2.5 Chi-Square Test (Categorical Variables vs Attrition)

The Employee data set chosen for this analysis also contains the categorical variables like WorkLifeBalance, JobSatisfcation etc, which effects the chance of attrition of employee. So, to analyse relationships between categorical variables and employee attrition. This statistical method detects significant associations between two categorical attributes. The result of the chi-square test is as shown below figure [5.1]



```
                        Chi-square statistic   P-value
JobInvolvement                      8.13911   0.043223
PerformanceRating                   2.24216   0.134293
EnvironmentSatisfaction            64.711982       0.0
JobSatisfaction                    53.173475       0.0
WorkLifeBalance                    49.589847       0.0
Attrition                        4402.608354       0.0
BusinessTravel                     72.547241       0.0
Department                         29.090275       0.0
Education                           5.641148  0.227598
EducationField                     46.194921       0.0
Gender                              1.349904  0.245295
JobRole                            25.116314  0.001486
MaritalStatus                      138.49103       0.0
Over18                                   0.0       1.0
```

Figure 4.15: Results of Chi-Square test

There are two key statistics the chi-square statistic which indicates the strength of the association between the categorical variable and 'Attrition'. Higher values indicate a stronger association and P-value which assesses the significance of this association. A low p-value (typically < 0.05) indicates that the association is statistically significant, suggesting that the two variables are not independent.

From the above results, for example it is clearly seen that p-value for JobRole is very low which is very close to 0 (0.001486), so that means the JobRole is not independent of Attrition means that the JobRole has some significance in the employee leaving the company. Similarly, the p-value of the variable Environment Satisfaction is close to 0, that means that this variable environment satisfaction is not independent of the Attrition.

For Example, consider Gender whose p-value is not as near to 0. As p-value is slightly higher it is independent of attrition means that gender of employee does not affect the attrition. By Identifying the statistically significant associations, it enables better recognition of influential attrition factors. This knowledge can then be incorporated into predictive modelling approaches to forecast employee churn.

CHAPTER 5

METHODOLOGY

The methodology used in this study aims to predict employee attrition using HR analytics and machine learning techniques.The following steps were undertaken to achieve the research objectives:

5.1 Feature Engineering

The employee attrition depends upon various factors such as Job Involvement, Job Satisfaction, Average working hours, etc., Hence it is necessary to find the features that are highly correlated with each other and also to find the most important feature in the dataset that helps in increasing the accuracy of the model.

Multicollinearity feature selection method have been chosen for this research work as it involves identifying and removing highly correlated predictor variables from a dataset before building the predictive model. This is done to improve the quality of the model's predictions and to make the interpretation of the model more meaningful.

To measure this , the variance inflation factor (VIF) given by below equation 5.1.1 and it quantifies how much the variance of the estimated regression coefficients is increased due to multicollinearity.

$$\text{VIF}(X_i) = \frac{1}{1 - R^2_{X_i}} \tag{5.1.1}$$

Where:

$$\text{VIF}(X_i) : \text{Variance Inflation Factor for predictor } X_i$$
$$R^2_{X_i} : \text{Coefficient of determination for predictor } X_i$$

Typically ,High VIF values imply increased multicollinearity. Generally, VIF values exceeding 5 or 10 indicate high collinearity as shown in below table [5.1].These highly correlated columns such as age, business travel, number of years at company and number of absent days are removed (usually which are greater than 5).

Column	SCR_VIF
Age	10.226953
BusinessTravel	6.291384
Department	6.146847
PercentSalaryHike	5.678561
TotalWorkingYears	9.565238
TrainingTimesLastYear	5.329267
YearsAtCompany	9.392214
YearsWithCurrManager	5.901202
absent_days	5.579792

Table 5.1: Highly correlated columns

5.2 Splitting of the Dataset

This is one of the critical steps to be followed while applying a machine learning model to the dataset. Splitting the dataset into different subsets such as train and test datasets help to evaluate the modelâs performance and generalization ability. On the train data, the data is used to train and learns the complex patterns from the data where as the test dataset is unseen in training which is mainly to used test the model after training. This also helps in selecting the best hyperparameters and detecting issues like underfitting or overfitting.This makes the unbiased evaluation of model performance.

Subset	Size of Dataset
Train	(3086, 19)
Test	(1324, 19)

Table 5.2: Sizes of Datasets

The overall aggregated dataset based on the five datasets as mentioned in chapter [3] is now split into training and testing phases and the dataset has been split in the ratio of 70:30 i.e., 70% data for training and 30% data is splitted into testing data.

5.3 Oversampling the data

The dataset contains 4410 rows and 20 columns and the output column Attrition contains values 0 which contains 3699 rows (class-1) and 1 which contains 711 rows of data(class-2). Since the class-1 ia majority class and class-2 is minority class, so inorder to balance the data for more accurate results.SMOTE oversampling technique has been chosen as the ideal technique to address the issue of imbalanced data and ensuring sufficient representation of the minority class (attrition cases) during model training, SMOTE improves the model's capacity to make accurate predictions for both classes.

Class	**Before Oversampling**	**After Oversampling**
0	2596	2596
1	490	1298

Table 5.3: Class Distribution Before and After Oversampling

In this study, the training data which is of 70% of the data have been oversampled using smote with sampling ratio of 0.5, the distribution of the training data after sampling as shown in above table [5.3]. Managing imbalance is crucial while applying the machine learning models. After handling this imbalance in the data ,machine learning models are implemented to get better results and performance.

5.4 Machine Learning Models used

The objective of this dissertation is to predict which employees are likely to leave the company - in other words, to predict employee attrition. This is a classification task, requiring the classification of employees into those who will leave versus those who will not. Since attrition prediction is a classification problem, some key machine learning algorithms for classification were utilized. These algorithms can classify employees as probable to leave or not leave based on patterns in their data. Applying classification models enabled the categorization of employees into likely leavers and likely retainers. So, classification algorithms were chosen as an appropriate technique for this attrition prediction among the employees.

5.4.1 Logistic Regression

Logistic regression is a fundamental algorithm for binary classification problems. It models the probability of an observation belonging to a particular class using the logistic function. This algorithm estimates the coefficients for each feature to find the best-fitting S shaped curve which is bounded between 0 and 1 [Appiah et al., 2020], that separates the two classes. This allows the output of logistic regression as probability. It is interpretable and serves as a baseline model for many classification problems.

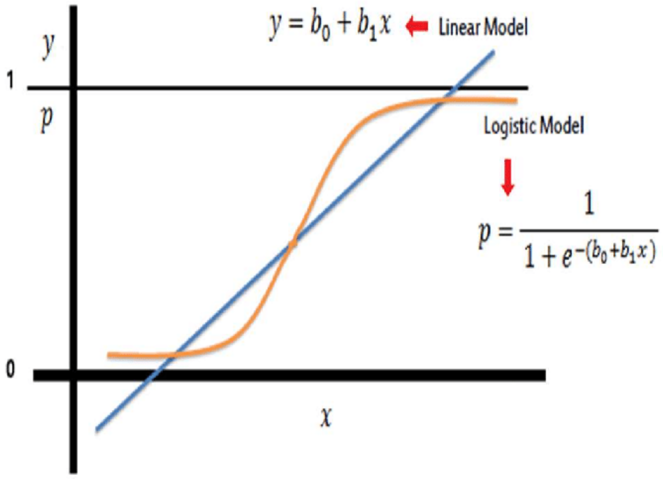


Figure 5.1: Logistic Regression ([Appiah et al., 2020])

As shown in the above figure [5.1], P denotes the probability and Y-axis represent the target variable (Attrition) and X - axis denotes the dependent variables. The straight line represents the linear regression by applying the logistic sigmoid function, it turns into a S - shaped curve. In this study, as the goal is to classify employees into two classes - those likely to leave the company versus those likely to stay. Logistic regression is ideal for attrition prediction because the target variable is categorical and binary. This models the probability of an employee attrition using the logistic function which maps any real-valued input to an output between 0 and 1. This provides the interpretability into how various factors influence attrition. Analysing the learned coefficients reveals which features like job satisfaction, work-life balance, relationship with manager etc. contribute most to the risk of attrition.

5.4.2 Support vector machine

Support Vector Machine is a supervised classification algorithm that finds the hyperplane that best separates two classes by maximizing the margin between them as shown in figure [5.2]. Some of the key terms to be noted here are:

- **Hyperplane**: In SVM, a hyperplane acts as a decision boundary separating two classes. In case of two-dimensional data, it is a line as shown in above figure but when it comes to higher dimensions of data it will become hyper plane
- **Support vectors**: The data points that lies very closest to the decision boundary.
- **Margin**: This represents the distance between the hyperplane and closest data points on either side which are known as support vectors as said above point

Mathematically, SVM can be explained [Jakkula, 2006] by the formulation of hyperplanes. The hyperplane can be defined as:

$$w^T x + b = 0$$

where w is the weight vector, x is a data point, and b is the bias term.
Algebraically, the margin width is calculated as:

$$\text{Margin} = \frac{2}{||w||}$$

where $||w||$ is the Euclidean norm (magnitude) of w.

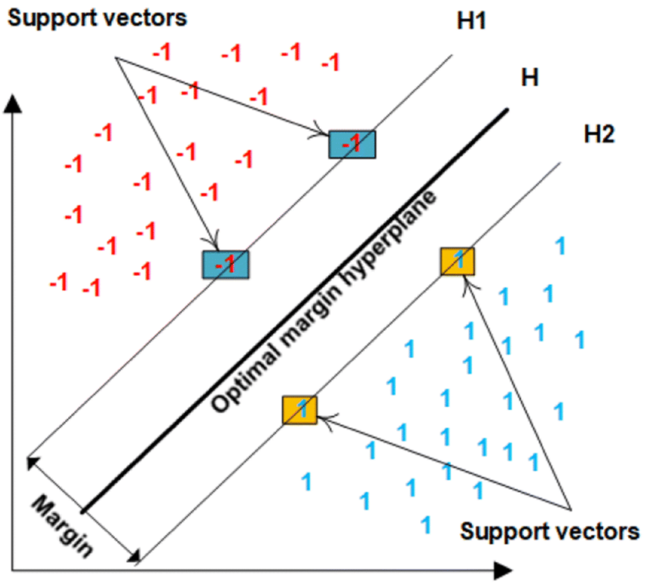


Figure 5.2: support vector machine ([Ji et al., 2021])

SVM finds the optimal values for w and b that maximize the margin by solving a constrained optimization problem as shown in the figure [5.3]. The constraints ensure correct classification by requiring:

$$\begin{aligned} w^T x + b &\geq +1 \quad \text{for positive class} \\ w^T x + b &\leq -1 \quad \text{for the negative class} \end{aligned}$$

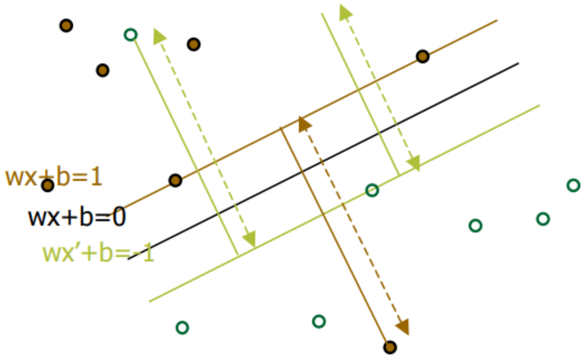


Figure 5.3: Hyperplanes in SVM

A wider margin implies better separation between classes, enhancing the classifier's ability to generalize to new unseen data. The margin maximization framework regularizes the model, improving predictivity.

It transforms input data into higher-dimensional space using kernel functions [Jakkula, 2006] to make the classes linearly separable. SVM aims to find the optimal balance between maximizing the margin and minimizing misclassification. It's effective for high dimensional data and has applications beyond binary classification.

SVM can be effective for predicting employee attrition when there's a clear separation between attrition and non-attrition classes in the feature space. It works well on imbalanced datasets where instances of the employee who likely to leave are fewer, SVM tends to be biased towards the majority class. By using cost-sensitive svm and by adjusting class weights can improve the performance. It can highlight combinations of features that contribute to high attrition rates, such as a significant difference in the number of years an employee has been with the company compared to the average tenure for their role.

5.4.3 Decision Tree

A decision tree is a non-linear model that consists of nodes representing attributes, branches denoting different attribute values, and leaf nodes with target class labels. It works by recursively partitioning the features space into smaller regions and finally turns into a tree like structure as shown in below figure [5.4]. Based on the training data, this makes some certain rules and decisions. CART, which is a simple classification and regression tree algorithm used in this study.This can handle both categorical and numerical target variables[Alao and Adeyemo, 2013]. This builds the trees by recursively splitting the data based on highest information gain.

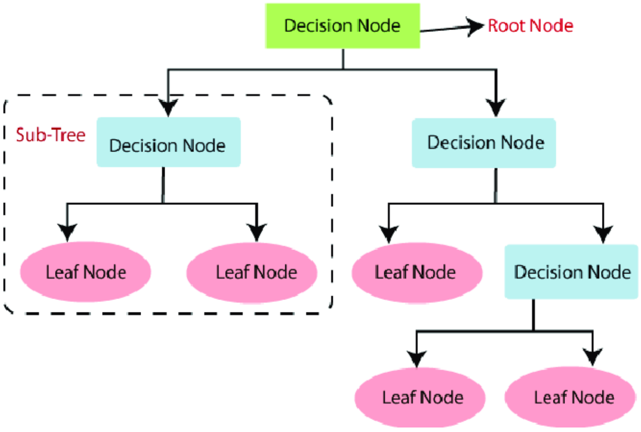


Figure 5.4: Structure of Decision Tree ([Hafeez et al., 2021])

The algorithm continues to split the data into subsets at each node based on the chosen features. This recursive process creates a tree structure until a stopping criterion is met, such as a maximum depth or a minimum number of samples in a node. Once the splitting process is complete, the leaf nodes are formed. Each leaf node represents a decision or prediction. For classification tasks, the majority class in the leaf node is assigned as the predicted class. For regression tasks, the mean or median value of the target variable in the leaf node is the prediction.

In case of imbalanced datasets, decision trees tend to be biased towards the majority class.In this study context, the decision tree serves as a predictive model that makes informed decisions based on various factors related to employees. These factors, known as features or attributes such as age, job satisfaction, work environment, education level, and more. The goal is to utilize these features to create a model that accurately classifies whether an employee is likely to leave the company or not.

5.4.4 Random Forest

Ensemble learning is a learning technique in which multiple individual models combined together to get a master model. These ensemble techniques are very robust and can works well on complex distributions of the data. Bagging and boosting are the two extremely powerful techniques to create the ensemble models by combine several decision trees. For example, Random Forest is based on bagging technique and some of the implementation of the models based on the boosting technique is AdaBoost, XGBoost, CatBoost etc.

Random Forest is the ensemble learning method which is based on the bagging technique. This works by parallel learning process. Unlike a single decision tree, this constructs multiple decision trees from the same dataset during the phase of training [Pratt et al., 2021]. These individual trees then collaboratively decide on the final prediction by majority voting. This ensemble technique improves the overall accuracy and predictive capability of the model.

The main strength of Random Forest is its ability to reduce overfitting by generating the subsets of the original data through a process called bootstrapping which is based on random sampling technique with replacement. Each decision tree is trained on a different subset, capturing unique perspectives within the dataset. Furthermore, Random Forest employs a technique known as feature bagging. This involves randomly selecting subsets of features for each tree. By doing so, the model can prevent any single feature from dominating the process of prediction. This model can uncover complex non-linear relationships and patterns between

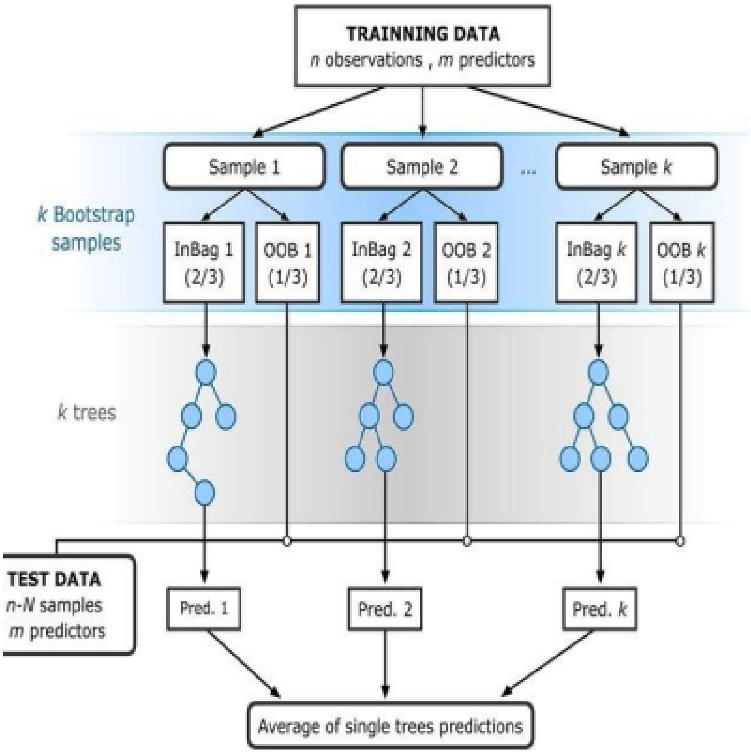


Figure 5.5: Working process of Random Forest ([Pratt et al., 2021])

the independent attributes such as low environment satisfaction and recent performance rating given and how these influence the attrition risk.

5.4.5 XGBoost

Tree boosting is an effective machine learning technique that performs additive optimization in function space by greedily adding decision trees. Gradient boosting is a popular form of tree boosting. Some of the experiments shows that XGBoost which is the optimized implementation of gradient boosted tree that achieves the state of art results on many problems [Chen and Guestrin, 2016]. This ensemble boosting technique is known for its speed and performance. Gradient Boosting is a powerful ensemble technique that builds multiple decision trees sequentially. This Boosting technique can reduce variance and overfitting and also increases the model performance and robustness. Unlike bagging technique this works by sequential learning process as each tree corrects the errors made by the previous trees. It optimizes a loss function through gradient descent.

Mathematically, the objective function as given in equation [5.4.1] includes two terms as shown in below equation (source : *XGBoost Documentation* [XGBoostDocs,])

$$\text{Objective\_function} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \omega(f_i) \tag{5.4.1}$$

Where:

$$l(y_i, \hat{y}_i^{(t)}) : \text{represents the training loss}$$

$$\omega(f_i) : \text{is a regularization parameter}$$

Here the training loss shows that how good the choosen base model works with training data. In xgboost, it is the difference between the sum of squares of predcited value and the actual value and other term measures the complexity of the tree [?].The above objective function is optimized to balance between the training loss and regularization.The loss function of the base model is approximated by Taylor series [XGBoostDocs,]. Some of the main steps followed to implement XGBoost algorithm are as follows:

- Firstly, A model is developed at starting at point and calculate the error or residual for each observation in the data. This residual can be calculating by finding the difference between the predicted values from the based model and the actual labels or values from the training data.
- Now, in order to predict these residuals a new model will be developed and then add these predictions from the model to the ensemble of models like a sequential method.

As shown in the above flow chart [5.6], XGBoost makes the models iteratively and combines them into an ensemble model. It is just like combining all the intelligence from the weaker models. This process will repeat again and again by adding the models to ensemble until the actual model becomes extremely powerful. This XGBoost follows regularized boosting and the trees are grown by depth wise or level wise [Al Daoud, 2019] as shown below

Overall, XGBoost is more superior as it has a good balance over bias and variance unlike the gradient boosting algorithm which only minimizes the variance. XGBoost also have inbuilt regularization mechanisms like tree pruning, shrinkage that can improve the model generalization. By this XGBoost enhances the overall robustness.

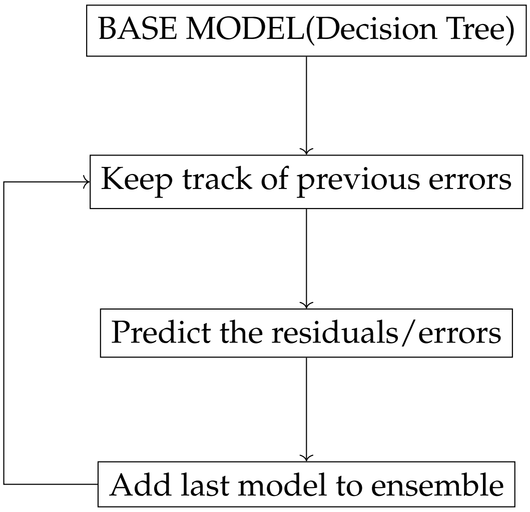

BASE MODEL(Decision Tree)
Keep track of previous errors
Predict the residuals/errors
Add last model to ensemble

Figure 5.6: Working Process of XGBoost

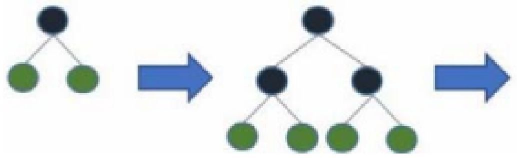


Figure 5.7: Level-wise growth in XGBoost ([Al Daoud, 2019])

5.4.6 CatBoost

CatBoost is also an ensemble machine learning algorithm based on the gradient boosting framework known for its high performance, efficiency and ease of use. It was developed by Yandex, a Russian multinational IT company. The algorithm is specifically designed to address challenges posed by datasets with categorical features, where traditional gradient boosting algorithms struggle to efficiently handle such variables. Usually in gradient boosting, the problem of target leakage happens because at each step the model fits the residuals from the previous step as shown above fig. These residuals depend on target variable because of this target variable influence the model and causes overfitting. CatBoost mainly overcomes the problem of target leakage [Prokhorenkova et al., 2018]. This has two key techniques or improvements that make different from other boosting algorithms as follows:

- **Ordered Boosting**: CatBoost uses multiple random permutations of training data samples. To make predictions, this uses the model trained on the permutation by this way the ordered boosting is used by CatBoost to avoid the prediction shift caused by target leakage. This reduces the problem of overfitting.

- **Handling of categorical variables**: These variables have certain levels or factors for instance gender has two factors male or female. Most of the machine learning models struggles with these types of features because they are good at handling the numerical values. So, to overcome this, CatBoost also implicitly handle the categorical features in the data by converting them into target statistics

In this way CatBoost is a combination of boosting and categorical encoding techniques. Unlike other boosting algorithms, CatBoost supports the categorical attributes without the need of techniques like one-hot encoding. In this study, there are many categorical variables like work-life balance, job satisfaction etc are the important attributes in predicting the employee attrition. So CatBoost will perform automatic encoding of these categorical variables in training phrase which reduces the time and increases the accuracy.

5.4.7 LightGbm

Light Gradient boosting machine is developed by Microsoft by mainly focussing on the implementation time, and is known for its exceptional speed, accuracy and high scalability. This is very good at handling large datasets and complex feature interactions. There are two most efficient techniques used in LighGbm [Ke et al., 2017] are as follows:

- **Gradient-Based one side sampling**: This shortly represented as GOSS. Unlike random sampling, it is a data sampling technique used in LightGbm which mainly concentrates on the larger gradients while building the decision tress. Goss keeps the subset of larger gradients this helps in improving the training speed and reduces the memory usage.

- **Exclusive Feature Bundling**: This is based on the feature engineering technique that identifies the mutually exclusive features during training phase and combines them into a single feature. This EFB technique makes the prediction process faster and more accurate.

Unlike level-wise growth in XGBoost,LightGbm grows the decision tree by leaf-wise [Al Daoud, 2019] means the splits are decided at the leaf level as shown in figure[5.8].

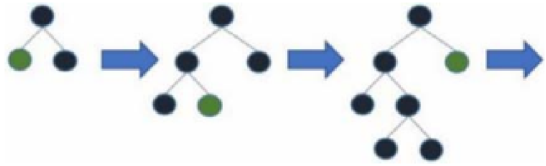


Figure 5.8: Leaf-wise growth in LightGbm ([Al Daoud, 2019])

LightGbm implements histogram based splitting implementation by which the feature values are grouped into a fixed number or bins or intervals. This process helps to reduce the complexity of decision tree construction and also reduces overfitting.

5.4.8 Multilayer perceptron (MLP) classifier

Deep learning, a subfield of machine learning. The process of DL uses a multitude of layers that represent data abstractions in order to carry out a variety of computer models.A multi-layer perceptron algorithm, a self-organizing map method, and a deep belief network algorithm is one of deep learning algorithm,which is based on feed forward network.

MLP Classifier are able to handle both numerical and categorical features because some of the features, such as job satisfaction, can be measured on a numerical scale, while other features, such as the employee's department, can only be categorized and further MLP classifiers are able to learn complex relationships between the features and the target variable

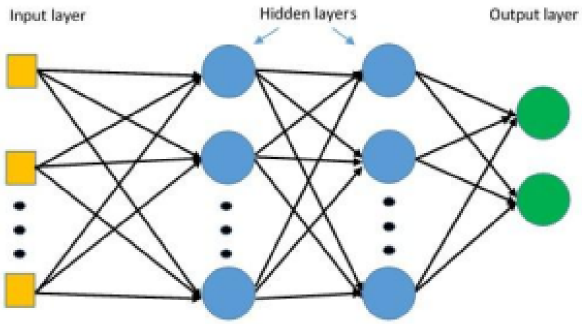


Figure 5.9: Layers in MLP ([Shankar et al., 2021])

As shown in the above figure [5.9] ,this consists of three layers namely input, hidden and output layer. This layering helps the network to learn the pattern in the data it uses back propogation to train the network. It employs the popular activation functions include ReLU, sigmoid, and tanh. Optimization algorithms like Adam and SGD adjust weights iteratively to minimize a loss function.

5.5 Hyperparameter selection and Grid search

The next step is the processing of ML models such as Logistic Regression, SVM, DT, XGBoost, MLP Neural Network, RF, Light GBM and CatBoost model but before implementing ML models, the parameters and its values have to be chosen for these models.These hyperparameters aslo known as external parameters which can control the learning process of the model.Parameters like Number of Epochs, n_estimators, max_depth etc. are known as hyper parameters. These parameters can define how the model is going to be trained on the training data. Based on this the learning of the model varies. so this hyperparameter tuning plays a considerable role in increasing the efficiency of the model by learning the hidden complex patterns in the data.

Hyperparameter tuning is the process of choosing the optimum set of hyperparameters for a machine learning model such that for these set of hyper parameters the model will give best accuracy performance and aslo making it work best on the unseen data like test dataset. This is also known as hyperparameter optimization.For this purpose, In this study, Grid Search method have been used for choosing the best combination of hyperparameters that gives best results on the dataset.

5.6 Model Evaluation

These are the performance measures also known as performance metrics these help in evaluating how the machine learning models works on the particular dataset. Accuracy is one of the most commonly used metrics in evaluating the models. This may not be the best approach to examine the model always when the data is imbalanced as it bias towards the majority class [Hossin and Sulaiman, 2015] . There are some other metrics such as Confusion matrix, F1-score, precision, recall etc. may provide some more insights on the model performance. Instead of training data, Evaluating the model based on the unseen test dataset gives more insights to model generalization on real time datasets.

5.6.1 Accuracy

This refers how the machine learning model is being able to predict the target outcome correctly, means when there is less difference between the predicted value or label and actual label then the accuracy will be high. This can be influenced based on how the data is distributed mainly when it comes to classification tasks means some times the model can achieve high accuracy based on majority class by ignoring the minority class.

5.6.2 Confusion matrix

The confusion matrix is one of the most important performance metrics that provides the comprehensive view on how the classification model [Hossin and Sulaiman, 2015]. The below figure [5.10] depicts the confusion matrix for binary classification problems, this may also extend to multi-class classification.

It is typically a 2*2 table which shows how the actual and predicted labels varies based on the four parameters as follows:

- **TRUE POSTIVE (TP)**: This means that the labels which are actually positive and the model correctly predicts these values as positive.
- **TRUE NEGATIVE (TN)**: This is similar to TRUE POSTIVE, means the actual value are negative and model also correctly predicted them as negative
- **FALSE POSTIVE (FP)**: This is also known as Type I error, Here the actual values are negative but the model incorrectly predicts them as positive.
- **FALSE NEGATIVE (FN):** In this, the actual values are positive but the model incorrectly predicts them as negative, this referred as Type 2 error.

Based on the above four values given by the confuion matrix, the accuracy can be defined as

$$Precision = \frac{TP+TN}{TP+FP+FNTN}$$

There are some other metrics apart from accuracy [Hossin and Sulaiman, 2015] which provide more understanding on how the classification model works. These clearly defines how the model works on each of class (in this study it is Attrition VS No Attrition).

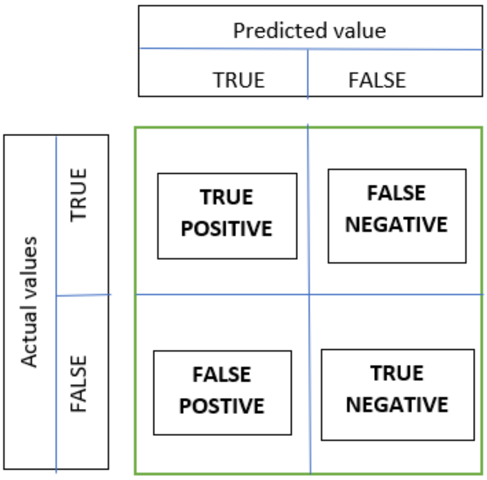


Figure 5.10: Confusion Matrix

5.6.3 Precision

This tells that how many values are actually predicted as positive out of all the predicted positive values. As the precision is high it indicates the low Type 1 error (False positive rate).

$$Precision = \frac{TP}{TP+TN}$$

It is simply the ratio of all true positives to the all observations that are predicited positive by the model

5.6.4 Recall or Sensitivity

This is known as True positive rate which mainly focuses on how the model is capable of predicting the true positive values. High recall refers that model is very good at predicting the positive cases . This measures the error caused by False Negatives.

$$Recall = \frac{TP}{TP+FN}$$

This is the simple ratio of all True positive to the all observations which are actually positive

5.6.5 F1-score

This is the combination of both the above metrics precision and recall. It is the harmonic mean of the precision and recall. In case of classification models this will be good metric to distinguish how the model performance varies between the two classes when they are imbalanced.

$$F1\text{-}score = \frac{2\times(\text{precision}\times\text{recall})}{\text{precision}+\text{recall}}$$

5.6.6 AUC-ROC score

This score ranges from 0 to 1. This is based on the both sensitivity and specificity of the model, this defines the performance of the model by considering its ability to classify the classes correctly across different probability thresholds. The model with high AUC-ROC score have better predictive power.

CHAPTER 6

RESULTS

Based on the above hyper parameters and it values, the Machine learning models such as Logistic Regression, Support vector machine, Decision tree and Random Forest models, Boosting based ML models such as XGBoost, Light GBM and CatBoost model and Multilayer perceptron (MLP) Neural Network model have been implemented and evaluated.

6.1 Machine learning models

The Machine Leaning models such as Logistic Regression, SVM, DT and RF models have been evaluated and results are shown below in table [6.1]. In this,all the metrics are based on how the model predicted the class-Attrition. As the whole aim of this study is to predict attrition.

Algorithms	**Precision**	**Recall**	**F1 Score**	**Accuracy**	**Training Time (sec)**	**Testing Time (sec)**
Logistic Regression	33%	28%	30%	58%	0.16	0.015
Support Vector Machine	33%	26%	29%	57%	682.511	0.12
Decision Tree	82%	86%	84%	90%	0.04	0.08
Random Forest	96%	89%	92%	94%	1.06	0.13

Table 6.1: Results of Machine learning models

The Random Forest model has the highest accuracy at 94%, making it the most accurate model among the ones listed. It also has high precision, recall, and F1 score, indicating that it effectively classifies instances and captures attrition cases. This gives the maximum accuracy because it is a ensemble learning model that combines the predictions of multiple decision trees, which helps to reduce the impact of overfitting and improve the accuracy of the predictions.

Next to Random Forest model, the Decision Tree model has an accuracy of 90%, Logistic Regression, SVM with an accuracy of 58% and 57% because these models are unable to capture the complex relationships between the features and the target variable. Further these models may be overfitting i.e., the models are learning the noise in the data instead of learning the underlying patterns in the data.

6.2 Boosting based ML Models

The Boosting ML models such as XGBoost, Light GBM and CatBoost models have evaluated on the HR data and results are as in following table [6.2]

Algorithms	**Precision**	**Recall**	**F1 Score**	**Accuracy**	**Training Time (s)**	**Testing Time (s)**
XGBoost	94%	93%	93%	95%	0.78	0.11
Light GBM	87%	78%	82%	87%	0.87	0.11
CatBoost	93%	84%	88%	91%	7.63	0.08

Table 6.2: Results of Boosting Models

The XGBoost boosting model incorporates a number of additional strategies, like tree pruning and column sampling, to increase the accuracy of the model, and as a result, it provides the highest accuracy of 95% in forecasting employee attrition. Additionally, it has a good F1 score, precision, and recall, demonstrating that it accurately classifies instances and detects attrition cases. The Light GBM and CatBoost models have the accuracy rates of 87% and 91%, respectively.

6.3 Neural network based-MLP classifier

The Neural Network models MLP Classifier has been implemented and evaluated the results as shown in table [6.3]

The Multilayer Perceptron (MLP) classifier gives low accuracy of 63% because its ability to learn complex relationships between the features and the target variable. Moreover the

Algorithm	Precision	Recall	F1 Score	Accuracy	Training Time (s)	Testing Time (s)
Multilayer perceptron	38%	39%	39%	63%	1.42	0.08

Table 6.3: Results of Neural network Model

dataset may not be large enough for the MLP classifiers to learn complex relationships and aslo it takes more time to train the data than the boosting models XGBoost and LightGbm which achieves good accuracy than this.

Overall, the best model is XgBoost based on its accuracy. when it comes to the duration of testing and training of the data, Decision Tree has the shortest training time of 0.02 seconds, while SVM and MLP have longer training times as depicted in table [6.1]. Decision Trees are simple and fast to train due to their basic structure. SVM and MLP are more complex and require longer training times due to their algorithmic complexity. CatBoost has the shortest testing time of 0.08 seconds as it is optimized for fast predictions because it utilizes techniques like ordered boosting and GPU acceleration.

6.4 Comparison of all the models

The results of all ML models implemented such as Logistic Regression, Support Vector Machines, Decision Tree, Random Forest, XGBoost, Light GBM, CatBoost and Neural Network MLP have been compared to analyse the performance of the model in terms of accuracy based on training and test data.The results are depicted in the below table [6.4].

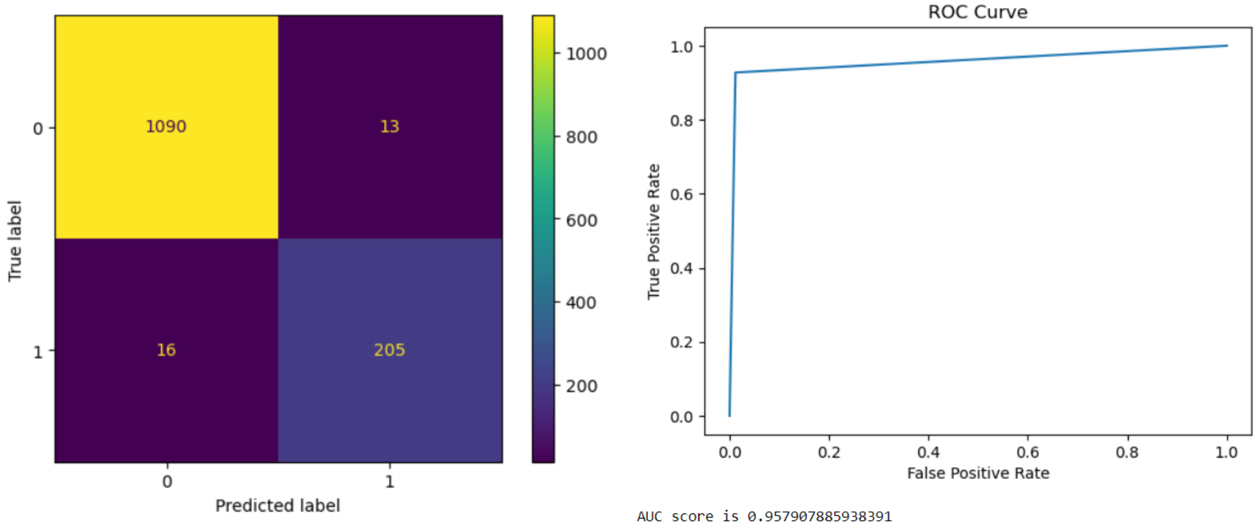
The models are trained on the training dataset which is of 70% of data and the remaining 30% of the data is completely new to the model.It clearly shows that the models XGBoost , Random forest and Decision Tree have very good accuracy on the test dataset.

The accuracies climbs up from Decision Tree to XGBoost because Random forest is ensemble model which has multiple decision trees and XGBoost is boosting model which corrects the errors from the previous decision tree in the ensemble.

Model	Training Accuracy	Testing Accuracy
Support Vector Machine	79%	57%
Logistic Regression	78%	58%
Decision Tree	94%	90%
Random Forest	98%	94%
CatBoost	96%	91%
LightGBM	94%	87%
XGBoost	98%	95%
Multi-layer Perceptron	79%	63%

Table 6.4: Accuracy percentages of machine learning models on training and testing datasets.

6.5 Best Model Confusion Matrix, AUC-ROC score



```
AUC score is 0.957907885938391
```

(a) Confusion matrix

(b) AUC-ROC curve

Figure 6.1: Performance of XGBOOST model on HR dataset

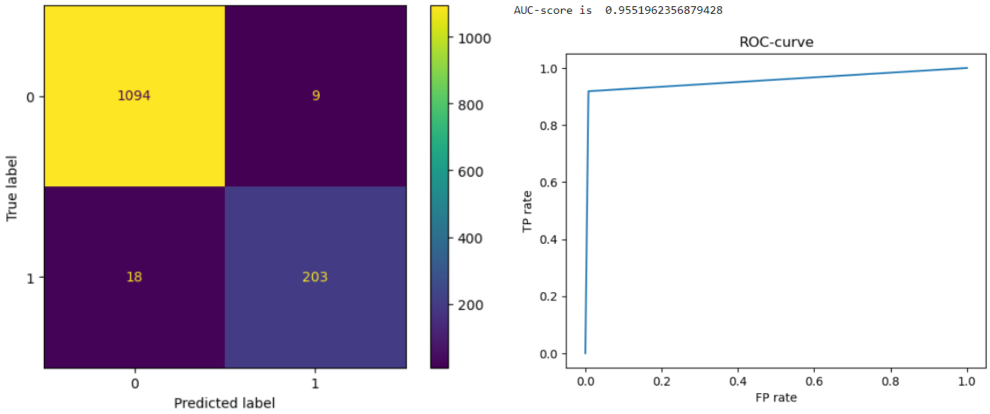
Based on the confusion matrix [6.1a], it is inferred that the XGBoost model has accurately classified 1090 instances as True Positive means those many instances are correctly classified as the employee is likely to leave the company and 205 instances as True Negative means it correctly predicts that those many employees are not leaving the company and also it incorrectly classifies 29 instances which is not bad.

The XGBoost model gives the AUC score of 0.957 means that the XGBoost model is able to correctly classify between the attrition and not attrition with 95.7% of the instances, which is quite good when compared to other models.So XGBoost is best model, by applying hyperparamter tuning by using grid search on this may enhance its performance. This can be implemented in python by GridSearchCV method which gives the best selection of paramters to the model as shown in the table [6.5].

Parameters_Tested	Parameters_Selected
'max_depth': [5, 10, 14, 18]	'learning_rate': [0.5]
'reg_alpha':0, [0.1, 0.01, 0.001]	'max_depth':[14]
'learning_rate' : [0.5, 0.4, 0.3, 0.1]	'reg_alpha':[0]

Table 6.5: XGBoost hyperparameter tuning

The parameter max_depth controls the height or depth of each decision tree ,reg_alpha is a regularization technique helps in preventing overfitting and learning_rate controls the size of step at each iteration of the boosting process and aslo deals with the training speed. By tuning these paramters aslo gives the same accuracy but there are some changes in the values of confusion matrix as shown in figure [6.2a].



```
AUC-score is  0.9551962356879428
```

(a) Confusion matrix

(b) AUC-ROC curve

Figure 6.2: Performance of XGBOOST model after Hyperparameter Tuning

After implementing the model, with the best selection of hyperparameters, It accurately classifies the 1094 instances as True Positive and 203 instances as True Negative .This model now incorrectly classifies only 27 instances this count is less than the XGBoost model achieved without hyperparameters.

CHAPTER 7

CONCLUSION AND FURTHER RESEARCH

7.1 Conclusion

Every organization has its own efficacy and strength, which rests on the shoulders of its employees, who represent a significant investment for businesses. When an employee departs a company, the organization must continually invest in recruiting, training, and developing new employees to satisfy vacancies. Training a new employee is a lengthy and expensive process, and it is in the company 's best interest to control and reduce employee attrition (employees quitting or retiring). In the discipline of Artificial Intelligence (AI), machine learning provides machines the ability to learn from previous data and make predictions about the future.

This research seeks to use advanced ML and forecasting models to analyze HR data and accurately predict the attrition rate. By understanding the underlying patterns, organizations can implement the retention strategies to maintain motivated and resilient work environment. For this purpose, the Employee Attrition dataset have been collected, preprocessed, visualized, implemented through ML models such as Logistic Regression, SVM, DT and RF, XGBoost, Light GBM and CatBoost , MLP Neural Network model and evaluated using Accuracy with time taken for training and testing the models. Based on the comparison, the XGBoost ML model gives the maximum accuracy of 95% in predicting employee attrition.

7.2 Future considerations

In order to advance future research, it is essential to delve deeper into identifying suitable models that can accurately be implemented in real-world scenarios. Additionally, it is crucial to gain a more comprehensive understanding of the variables and data that should be utilized when attempting to predict employee attrition. Exploring the potential positive or negative effects on individual employees resulting from the implementation of such models would also be an intriguing avenue to explore. Furthermore, this technique could potentially be applied to various other fascinating predictions like sick leave, motivation, and salary.

Enhancing the analysis in future research could involve suggesting the retention strategies for the company based on why the employee leave the company and also considering the factors such as new employee's (freshers) opportunities and adverse working conditions (such as harm and hazard), poor promotion prospects, discrimination, and low social support. These factors have been found to have a positive correlation with employees' turnover intention.

Bibliography

[Al Daoud, 2019] Al Daoud, E. (2019). Comparison between xgboost, lightgbm and catboost using a home credit dataset. *International Journal of Computer and Information Engineering*, 13(1):6–10.

[Alao and Adeyemo, 2013] Alao, D. and Adeyemo, A. (2013). Analyzing employee attrition using decision tree algorithms. *Computing, Information Systems, Development Informatics and Allied Research Journal*, 4(1):17–28.

[Allen et al., 2010] Allen, D. G., Bryant, P. C., and Vardaman, J. M. (2010). Retaining talent: Replacing misconceptions with evidence-based strategies. *Academy of management Perspectives*, 24(2):48–64.

[Appiah et al., 2020] Appiah, P., Edoh, T., and Degila, J. (2020). Predicting elderly patient behaviour in rural healthcare using machine learning. volume 2647.

[Belete, 2018] Belete, A. (2018). Turnover intention influencing factors of employees: An empirical work review. *Journal of Entrepreneurship & Organization Management*, 7(3):1–7.

[Chen and Guestrin, 2016] Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

[Das, 2013] Das, B. (2013). Employee retention: A review of literature. *IOSR Journal of Business and Management*, 14:08–16.

[Dutta and Bandyopadhyay, 2020] Dutta, S. and Bandyopadhyay, S. K. (2020). Employee attrition prediction using neural network cross validation method. *International Journal of Commerce and Management Research*, 6(3):80–85.

[Fallucchi et al., 2020] Fallucchi, F., Coladangelo, M., Giuliano, R., and William De Luca, E. (2020). Predicting employee attrition using machine learning techniques. *Computers*, 9(4):86.

[Hafeez et al., 2021] Hafeez, M., Rashid, M., Tariq, H., Abideen, Z., Alotaibi, S., and Sinky, M. (2021). Performance improvement of decision tree: A robust classifier using tabu search algorithm. *Applied Sciences*, 11:6728.

[Hang et al.,] Hang, J., Dong, Z., Zhao, H., Song, X., Wang, P., and Zhu, H. Outside in: Market-aware heterogeneous graph neural network for employee turnover prediction. WSDM'22: PROCEEDINGS OF THE FIFTEENTH ACM INTERNATIONAL CONFERENCE ON WEB â¦.

[Harter et al., 2002] Harter, J. K., Schmidt, F. L., and Hayes, T. L. (2002). Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: A meta-analysis. *Journal of Applied Psychology*, 87:268–279.

[Hossin and Sulaiman, 2015] Hossin, M. and Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

[Imran et al., 2017] Imran, R., Allil, K., and Mahmoud, A. (2017). Teacherâs turnover intentions: Examining the impact of motivation and organizational commitment. *International Journal of Educational Management*, 31:828–842.

[Jakkula, 2006] Jakkula, V. (2006). Tutorial on support vector machine (svm). *School of EECS, Washington State University*, 37(2.5):3.

[Ji et al., 2021] Ji, W., Liu, D., Meng, Y., and Xue, Y. (2021). A review of genetic-based evolutionary algorithms in svm parameters optimization. *Evolutionary Intelligence*, 14.

[Ke et al., 2017] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

[Krishna and Sidharth, 2022] Krishna, S. and Sidharth, S. (2022). Hr analytics: Employee attrition analysis using random forest. *International Journal of Performability Engineering*, 18(4):275.

[Mansor et al., 2021] Mansor, N., Sani, N. S., and Aliff, M. (2021). Machine learning for predicting employee attrition. *International Journal of Advanced Computer Science and Applications*, 12(11).

[McFeely and Wigert, 2019] McFeely, S. and Wigert, B. (2019). This fixable problem costs us businesses $1 trillion. *Gallup Research. Available online: https://www.gallup.com/workplace/247391/fixable-problem-costs-businesses-trillion.aspx (accessed on 20 March 2021).*

[Park and Shaw, 2012] Park, T.-Y. and Shaw, J. (2012). Turnover rates and organizational performance: A meta-analysis. *The Journal of applied psychology*, 98.

[PM and Balaji, 2019] PM, U. and Balaji, N. (2019). Analysing employee attrition using machine learning. *Karpagam Journal of Computer Science*, 13:277–282.

[Pratt et al., 2021] Pratt, M., Boudhane, M., and Cakula, S. (2021). Employee attrition estimation using random forest algorithm. *Baltic Journal of Modern Computing*, 9(1):49–66.

[Priya and Harasudha, 2017] Priya, V. K. and Harasudha, H. (2017). A study on employee attrition with reference to lanson toyota, chennai. *Man in India*, 97:115–124.

[Prokhorenkova et al., 2018] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31.

[Qutub et al., 2021] Qutub, A., Al-Mehmadi, A., Al-Hssan, M., Aljohani, R., and Alghamdi, H. (2021). Prediction of employee attrition using machine learning and ensemble methods. *International Journal of Machine Learning and Computing*, 11:110–114.

[Rohit and Ajit, 2016] Rohit, P. and Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. *International Journal of Advanced Research in Artificial Intelligence*, 5.

[Rombaut and Guerry, 2018] Rombaut, E. and Guerry, M.-A. (2018). Predicting voluntary turnover through human resources database analysis. *Management Research Review*, 41(1):96–112.

[Setiawan et al., 2020] Setiawan, I. a., Suprihanto, S., Nugraha, A., and Hutahaean, J. (2020). Hr analytics: Employee attrition analysis using logistic regression. In *IOP Conference Series: Materials Science and Engineering*, volume 830, page 032001. IOP Publishing.

[Shankar et al., 2021] Shankar, R. S., Priyadarshini, V., Neelima, P., and Raminaidu, C. (2021). Analyzing attrition and performance of an employee using machine learning techniques. In *2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1601–1608.

[Shankar et al., 2018] Shankar, R. S., Rajanikanth, J., Sivaramaraju, V., and Murthy, K. (2018). Prediction of employee attrition using datamining. In *2018 ieee international conference on system, computation, automation and networking (icscan)*, pages 1–8.

[Sharma and Stol, 2019] Sharma, G. and Stol, K.-J. (2019). Exploring onboarding success, organizational fit, and turnover intention of software professionals. *Journal of Systems and Software*, 159:110442.

[Trevor and Nyberg, 2008] Trevor, C. and Nyberg, A. (2008). Keeping your headcount when all about you are losing theirs: Downsizing, voluntary turnover rates, and the moderating role of hr practices. *Academy of Management Journal*, 51:259–276.

[Victoria and Olalekan, 2016] Victoria, O. and Olalekan, U. (2016). Effects of demographic factors on employees intention to leave in selected private universities in southwest. *Babcock University Publication Portal*.

[Vijay, 2018] Vijay (2018). Hr analytics case study. *Kaggle Dataset. Available online: https://www.kaggle.com/datasets/vjchoudhary7/hr-analytics-case-study*.

[XGBoostDocs,] XGBoostDocs. Introduction to boosted trees - xgboost 2.1.0-dev documentation. *Available online:https://xgboost.readthedocs.io/en/latest/tutorials/model.html*.

[Yahia et al., 2021] Yahia, N. B., Hlel, J., and Colomo-Palacios, R. (2021). From big data to deep data to support people analytics for employee attrition prediction. *IEEE Access*, 9:60447–60458.

[Zhao et al., 2018] Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., and Zhu, X. (2018). Employee turnover prediction with machine learning: A reliable approach. In *Intelligent Systems with Applications*.