



University of Essex

School of Mathematics, Statistics
and Actuarial Science

MA981 DISSERTATION

From Data to Retention - Machine
Learning Strategies for Predicting
Employee Attrition

Lokesh Palaniraj

Supervisor: **Saideh Ferdowsi**

November 24, 2023
Colchester

Abstract

Attrition is an inevitable part of any business. Mostly, there's someone who wants to leave the company for a variety of reasons, either professional or personal. It may be through resignation for different reasons or even retirement, and especially the position won't be replaced immediately. Employee attrition is a critical concern for organizations nowadays since it affects the work environment and operational efficiency. It can occur throughout the company or in a specific department or division.

In this paper, we'll estimate the attrition rate of the company, which is the number of employees who quit the company based on an HR database consisting of employee data like age, salary, etc. This can be done using advanced machine learning models, and the prediction can be useful to employ new retention methods proactively using data-driven decisions so that we can maintain a strong and stable working environment, which might be beneficial for both the organization and the employees. This study involves a variety of machine learning models, like logistic regression, support vector machines, decision trees, gradient boosting, etc., to discover patterns that might uncover potential insights to understand the reasons behind the attrition rate and the key factors contributing to this cause.

In this analysis, the Synthetic Minority Over-Sampling Technique (SMOTE) is used to create balanced and imbalanced datasets, and the outcomes are analysed and compared. We then went on and pre-processed the dataset, determining which features were important and which normalisation steps were necessary to improve the analysis's accuracy. Then the performances of the models have been evaluated using various performance metrics like accuracy, precision, recall, AUC-ROC curve and the k-fold cross validation has been carried out to confirm the correctness of the model across different subsets. Additionally, this study offers a thorough analysis of the key factors that lead to attrition, enabling us to put the right retention tactics into place.

Keywords: Employee Attrition, Machine Learning, Predictive Analytics, Retention Strategies, Human Resource Strategies.

Contents

1	Introduction	7
1.1	Overview of Employee Attrition	7
1.2	Impact of Employee Attrition	7
1.3	Significance of the Study	8
1.4	Research Objectives	8
2	Literature Review	9
2.1	ML in HR Analytics	9
2.2	Previous Approaches	10
2.3	Deep learning in Attrition	13
3	Dataset Description	14
3.1	Data Exploration	14
3.1.1	Uni-Variate Analysis	16
3.1.2	Bi-Variate Analysis	17
3.2	Data Preprocessing	20
3.2.1	Missing Values	20
3.2.2	Outliers	21
3.2.3	Correlation of different features	22
3.2.4	Data type handling	24
3.2.5	Addressing the Imbalanced data	25
3.2.6	Normalization	26
3.3	Feature Importance Analysis	26
4	Methodology	28
4.1	Data Partitioning	28

4.2	Hyperparameter Tuning using Grid Search	29
4.3	Machine Learning Models	29
4.3.1	Logistic Regression	30
4.3.2	Decision Tree	30
4.3.3	Random Forest	32
4.3.4	Support Vector Machine	33
4.3.5	k-Nearest Neighbors	34
4.4	Boosting Algorithms	35
4.4.1	AdaBoost	36
4.4.2	CatBoost	37
4.4.3	Gradient Boosting	39
4.4.4	XGBoost	40
4.5	Model Evaluation	42
4.5.1	Confusion Matrix	42
4.5.2	AUC-ROC Curve	45
4.5.3	k-Fold Cross-Validation	46
5	Results	47
5.1	Comparison of Prediction Outcomes	47
5.2	Interpreting the findings	50
5.3	Limitation & Further Research	51
5.4	Strategies for Employee Retention	52
6	Conclusions	53

List of Figures

3.1	Attrition percent	16
3.2	Employee's Gender Ratio	17
3.3	Employees Age group	18
3.4	Employees Salary based on their age	19
3.5	Satisfaction vs Attrition [8]	20
3.6	Missing Values	21
3.7	Outliers in our employee dataset	22
3.8	Feature Correlation	23
3.9	Features' Datatype	24
3.10	Distribution of Attrition before Oversampling	25
3.11	After Oversampling	25
3.12	Significant features with their F-stat Value	27
4.1	Decision Tree Based on the features from IBM HR Dataset [2]	31
4.2	Support vector machine [23]	34
4.3	Basic working of Adaboost [10]	36
4.4	Architecture of the CatBoost Algorithm [28]	38
4.5	Loss function in Gradient Boosting [24]	40
4.6	Architecture of XGBoost [25]	41
4.7	Loss Function & Regularization of XGBoost [26]	41
4.8	Visualisation of a Confusion Matrix [27]	42
4.9	Confusion Matrix of the SVM model in balanced dataset	43
4.10	Confusion Matrix of the Random Forest with imbalanced dataset	45
5.1	AUC-ROC Curve of the CatBoost	48
5.2	Evaluation of our AdaBoost model using k-fold	50

5.3 Evaluation of our CatBoost model using k-fold	51
---	----

Introduction

1.1 Overview of Employee Attrition

The term "employee attrition" refers to the loss of employees in a company due to various reasons like low salary, less job satisfaction, work transportation issues, bad working environment, etc. The attrition could be mainly of two different reasons: voluntary or involuntary attrition. The voluntary attrition occurs if the employee decided to leave the company for any one of the above mentioned reasons. On the contrary, involuntary attrition involves terminating an employee because of his/her poor performance or any violation of organisation's policies [1]. In both the cases, the company is losing an employee who won't be replaced right away, potentially disrupting the working environment and the affecting the project's efficiency.

1.2 Impact of Employee Attrition

When a skilled and seasoned worker departs from an organization, a vacuum is created that ultimately causes the loss of knowledge, connections, and expertise to the company. Reducing organizational wear and tear in a way that advances and develops the company has become a top priority for individual executives and specialists in recent times. If the issue is not resolved appropriately, the employee's department may incur significant loss and disruption, which could have a negative long-term impact on the

company[4].

1.3 Significance of the Study

The major importance of this study is that it can enhance the organisation's ability to engage in more effective workforce planning by employing new retention strategies. By doing this, the organisation can save a lot of money by cutting the need for recruitment, on-boarding and training and also thus reducing the financial burden. A better understanding of the factors leading to employee attrition can also enable the organisation to create more convincing environment for the employees thus offering career satisfaction. Maintaining the personnel who worked on an ongoing project with more specialized experience and encouraging stability in the workplace can also assist an ongoing project succeed in the long run. By relying on the machine learning techniques, the human resource department can take more data-driven decisions with higher level of accuracy which in turn helps the company to derive valuable insights.

1.4 Research Objectives

By assessing the likelihood of employees leaving; and the risk of retention would greatly help the organization to come up with new retention strategies or update the existing ones, thus reducing the cost and effort on hiring new employees. In this study we are going to use an IBM HR Employee Attrition dataset from Kaggle.org to estimate the rate of employee attrition, which includes over 1500 samples and 35 features[3]. We can use this historical data from the human resources departments to build, train and fine tune the machine learning models to predict the rate of employee who will be leaving the company[2]. The predictive capability enables the organization to plan ahead and take proactive steps to implement new retention measures and improve personnel management by the help of machine learning algorithms.

Literature Review

The ultimate goal of this research is to predict the employee attrition rate using the HR data set within organizations. Attrition may lead to loss of productivity, because it disrupts the continuity of the project which might also affect the project deadline. Also it can result in huge substantial financial costs for the company such as the expenses on recruitment, on boarding and suitably train them on the field they are going to work on. By analyzing the rate of attrition systematically, we can able to avoid all of these problems. Numerous studies have been conducted in this field to analyze and examine the rate of attrition, through machine learning and artificial intelligence. Using these published research studies as a guide, we will estimate the attrition-causing factors in our study.

2.1 ML in HR Analytics

Machine learning has numerous applications in human resources, including talent identification, employee performance management, training, and engagement [20]. Since the department holds enormous amount of employee data, Machine learning would be incredibly effective at evaluating them and finding insights, which would yield useful knowledge. Machine learning helps the HR to analyze potential candidates for the role and makes the shortlisting process much simpler. In talent recognition, it can process the candidates resume, analyze their soft skills, experiences and gives a report, stating if he/she is suitable for the role without going through all the interview

process, saving time and money for the organisation. It can also assist the organisation by keeping track on employees' engagement and can reveal trends if their performance by going through their overall metrics and examines if the employee might leave the company or not by analyzing his/her job satisfaction and compensation as well.

2.2 Previous Approaches

Alao D. & Adeyemo A. B. [13] used decision tree algorithms and data mining techniques to develop predictive models that can actually identify the employees who might leave the company, for example, they used employee's salary and length of service as one of the determining factors for predicting the employee attrition. So the employees who worked longer in the same company without any significant raise of pay were likely to quit. Employees with a reasonable number of years of service who performed poorly were also subject to termination. They also proposed a system for the HR department to predict employee attrition.

In the article by Francesca Fallucchi., et.al [3], they wanted to analyze the reasons for the employee attrition by using the machine learning systems and found that the major factor which determines the attrition was the 'monthly salary', so they suggested, there might be a possible issue with the compensation. The Gaussian Naive Bayes classifier algorithm achieved the best recall rate of 0.54 and overall the false negative rate of 4.5% in finding the instances of attrition. The researchers also analysed all other factors which might affect the attrition rate of employees like working years, overtime, satisfaction, etc.

Darapaneni, N., et al [6] used the IBM HR dataset to predict the prominent factors that leads to attrition by conducting Explanatory Data Analysis(EDA), which means summarizing the dataset by deeply investigating the main features and visualizing the findings for better interpretation. They trained the models and evaluated based on the Area Under Curve(AUC) metric, it calculates the region in two dimensions beneath the whole Receiver Operating Characteristic (ROC) curve between 0 and 1 (1,1)[19]. Based on the findings, the most important factor for the attrition among employees was working overtime. They also suggested that it can be fixed by providing additional support to the employees and help maintain a healthy work life balance.

In Jain, P.K., et al [8] paper, the objective was to explore the dataset using data exploration techniques like uni variate analysis, variable identification, etc to analyze the employee attrition problems in human resources. They also used some of the important machine learning models for the prediction such as decision tree, random forest, SVM. They evaluated the model accuracy through the performance metrics - confusion matrix, recall, accuracy and F1 score. The aim was to find the best fitted model by comparing the performances of all different models. This research focused on controlling the attrition rates by implementing better policies for the employees and also improving the work environment.

Since the demand and importance for the data has been increasing rapidly over the years, conventional business practices are no more sufficient for analysis, So the companies have to adapt the new advanced artificial intelligence and machine learning algorithms to improve the company's growth through the insights. Maharana, M., et al [9] analyzed the IBM Watson data-set and compared three main machine learning classification models such as Random forest classifier, Logistic Regression and Decision Tree to estimate the employee attrition and found that the Logistic regression model achieved a highest accuracy of 87% for the dataset. The employee attrition can be much classified and reduced by using boosting algorithms like XGBoost.

The research of George, S., et al [10] aimed to predict and analyse the factors contributing to employee attrition using machine learning. They collected the data and handled it by removing the outliers and pre-processed to clear the missing values and scaled the data appropriately. They applied ensemble models like Extra Tree Classifier, AdaBoost, XGBoost and Gradient boost and found that the model achieved the highest accuracy was Extra Tree Classifier with 97.08%. The authors also recommended to examine artificial intelligence techniques like SHAP and LIME to identify the main factors that affects the attrition and address them accordingly.

In this study, the authors Meraliyev, B., et al [11] used the machine learning algorithms to identify the factors contributing to employee turn over and take necessary action on it. In order to retain the valuable employees in the company they used machine learning models including Multinomial Naive Bayes, Gaussian Naive Bayes, SVC, Decision tree and Logistic regression on a real HR data set from Kazakhstan. To enhance the precision of the prediction pre processing methods has been carried out such as

feature selection and transformation. Out of all the techniques, the model with the best accuracy rate of 75% was logistic regression. After optimization, the Decision Tree Regressor had the best precision, recall, and F1-score of all the models. On an average, employees begin to quit after three and half years of working.

Unlike the other researchers, Priya, V. Krishna, & Harasudha [14], in a recent study on personnel at Chennai's Lanson Toyota collected information from 100 workers via a survey in order to identify those who could be considering leaving the company. The studies aim to identify the factors that contribute to attrition, such as compensation, job satisfaction, career growth, working conditions, and interpersonal relationships. So, from the gathered primary data, they performed some statistical analysis technique like ANOVA (Analysis of Variance) to compare the data and came up with the conclusion that the pay and benefits improve along with the individual's promotion and advancement to the next level, yet the employee might not be happy with the package and hunt about for better offers from other companies. The study also suggests the need for long term strategies and approaches to reduce attrition. They also mentioned that the research could be suitable only for specific sectors like automobile, IT, insurance, retail, etc to understand the factors of attrition within those sectors.

It is imperative for decision makers in any firm to identify individuals who pose the highest risk of employee attrition or who could be prime targets for poaching. According to A. Qutub, et al[2], identifying employees at risk of attrition using indicators or factors is a challenging task that requires many years of experience. The factors related to attrition are also influenced by people's preferences and expectations, which can vary between generations, types of work, and employees' life stages. However, he found that the model with highest accuracy, recall and AUC in predicting the employee attrition using machine learning was linear regression among other models. He used the IBM HR dataset to analyze the rate of attrition using models like Decision Tree, Random forest, AdaBoost and Gradient boosting. Furthermore, they inferred that the ensemble comprising several base models performed somewhat worse than their best base models; this is likely due to the smaller dataset; ensembles tend to generalize better on larger datasets.

In the research paper carried out by Hebbar, A., et al [12], the machine learning models has been compared to minimize the attrition rates and also to build better strong

relationship between the companies and employees. They found that Logistic regression, Random Forest, and Support Vector Machine (SVM) are effective in predicting employee attrition and notably, Random Forest was the best performing model with 90% accuracy. The Boruta package, an all-relevant variable selection technique, was used for Random Forest model enhancement and to avoid overfitting and slow computation.

2.3 Deep learning in Attrition

In Kerem Gurler's, et al [15] research, they devised a new approach to examine the attrition by analyzing various algorithms and methods for prediction. Initially, they began to handle the data imbalance in the IBM HR dataset, followed by the traditional ML and deep learning techniques. They also used Recursive Feature Elimination with Cross Validation for Feature selection and machine learning models like LR, SVM, RF, AdaBoost and DT. For the data imbalance, they used oversampling methods like SMOTE (Synthetic Minority Oversampling Technique) & SMOTE with Tomek Links, which is nothing but to select a random data for the minority class. They also applied deep learning techniques like Deep Random Forests and Feed Forward neural network. Undoubtedly, the Neural networks performed way better than the traditional methods in Machine Learning, demonstrating the efficiency of deep learning in predicting Employee Attrition. It's also because of the hyperparameter optimization which fine-tuned the neural network models, optimizers, activation functions, learning rates and regularizations.

Followed by the research of Francesca Fallucchi [3], Salah Al-Darraj, et al [16] wanted to improvise it by analyzing the ways to predict the attrition using deep learning. Starting with the pre-processing steps to enhance the prediction, thereafter analyzed the dataset for the crucial factors that might have a greater impact on analysis. In order to get realistic results, they tested the models on both balanced and imbalanced datasets and subsequently evaluated the model using cross validation for accurate results. They inferred that OverTime, MonthlyIncome were the most influencing factors for attrition and also found that the balanced dataset achieved better accuracy than the imbalanced one. The accuracy of the the deep learning model using the 10-fold cross validation was 89.11%.

Dataset Description

The dataset that we are using to perform our analysis in this report was from IBM data analysts who simulated an actual employee database and the main purpose of this is to facilitate research and analysis to reflect real world scenarios. This helps the scientists to uncover critical insights and explore the underlying trends and causes. The dataset was utilized originally from Kaggle, which is a community of data scientists and machine learning engineers who share their works and papers online.

The dataset consists of 1470 records with 35 features which not only consists of personal information but also includes various factors that contributes to attrition like the age of the employee, his distance from home, job satisfaction, pay rate, working hours in a day, performance rating and more. Out of 1470 employees, 237 employee were about to quit the company for various number of reasons, this clearly tells us the database is clearly unbalanced [5].

3.1 Data Exploration

Data Exploration is the leading and initial step of analytics, where statistics and visualizations are used, to explain and understand the dataset more deeply. To reveal the important features of data and to analyze it, we need to explore the dataset. Firstly we can start by understanding the features of the dataset, following the uni variate and bi variate analysis [8]. The table 3.1 describes the features of the dataset.

As per our objective of understanding the employee attrition trends, our target

Feature	Description
Age	Age of the Employee
Attrition	Status of Employee's Attrition
BusinessTravel	Employee's Travel Frequency
Department	Employee's Department
DistanceFromHome	Employee's Distance from Home to Workplace
Education	Employee's Level of Education
EducationField	Employee's Field of Specialization
EnvironmentSatisfaction	Satisfaction level with their Environment
Gender	Gender of the Employee
JobLevel	Employee's Job level in the organization
JobSatisfaction	Job Satisfaction level of the Employee
MaritalStatus	Marital Status of the Employee
MonthlyIncome	Monthly Salary of the Employee
NumCompaniesWorked	Number of Companies the Employee has worked
OverTime	Whether the Employee works Overtime?
PercentSalaryHike	Percentage of increase in Employee's Salary
PerformanceRating	Employee's Performance Rating
RelationshipSatisfaction	Employee's Relationship Satisfaction
StandardHours	Number of Employee's working hours per day
TotalWorkingYears	Total Number of Working years in the Company
TrainingTimesLastYear	Training sessions attended by the Employee
WorkLifeBalance	Employee's Work-Life Balance Satisfaction
YearsAtCompany	Total number of working years in the company
YearsInCurrentRole	Years in the Employee's current role
YearsSinceLastPromotion	Years since the Employee's last promotion
YearsWithCurrManager	Number of years working with the Current Manager

Table 3.1: Description of Features

variable would be "Attrition" in the dataset which has two entries, i.e., Yes or No. 'Yes' represents the employee has left the company and 'No' means the employee did not leave the company. Using different machine learning models we can predict this feature and find the factors contributing to attrition. Since the amount of people responded 'No' is 1233 out of 1470 records the dataset is imbalanced thus may skew our machine learning analysis, leads to poor model performance. This may also result in overfitting the majority class due to irregular distribution of our target class that can lead poor generalization of the minority class as well.

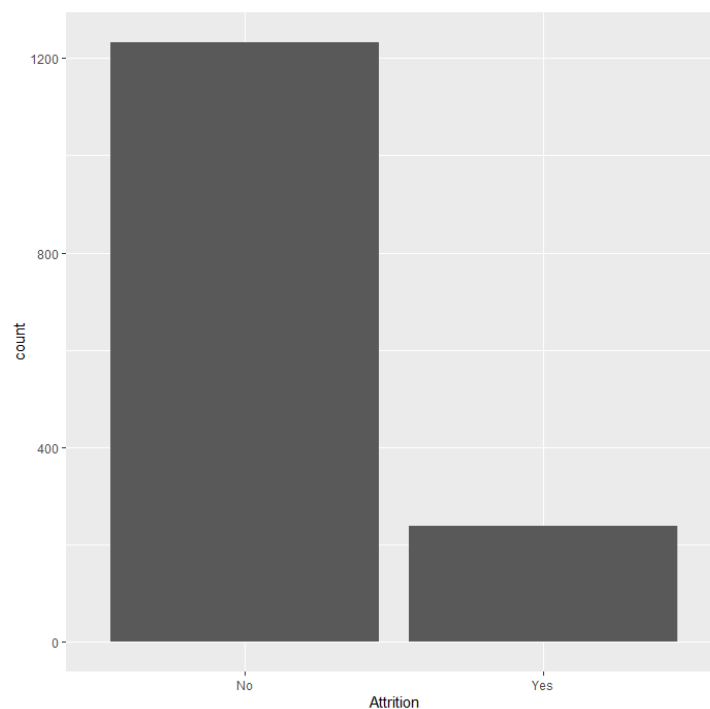


Figure 3.1: Attrition percent

3.1.1 Uni-Variate Analysis

Univariate analysis is the initial step and the simplest of all analysis because it involves analyzing only one variable. The main focus of univariate analysis is to understand the characteristics, features of a single variable at a time. This procedure is crucial to the data analysis process in order to identify common patterns and enhance comprehension of the general behavior of a single variable. It can be applied to both numerical and categorical variable, using various data visualization techniques like bar graphs, histograms, boxplot, etc. It also involves measuring the central tendency i.e, mean, median

and mode and dispersion, shape. Sometimes this is also used to find the outliers in the variable and can be able to remove it during the preprocessing step. The results of this process will give us the basis for understanding the context and provides further ideas for data preprocessing steps.

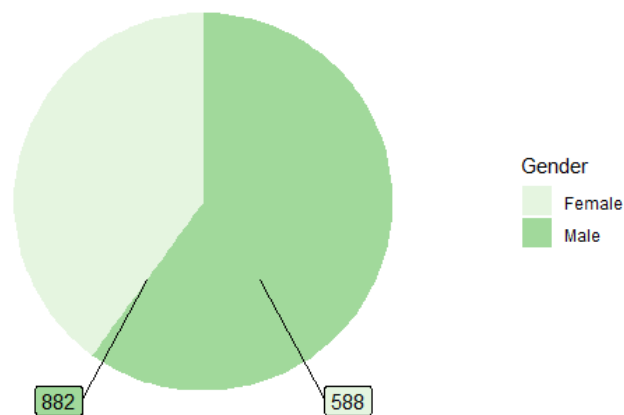


Figure 3.2: Employee's Gender Ratio

The pie chart 3.2, represents the gender distribution within the dataset, with a total amount of 1470 employee samples. It's obvious that the dataset comprises of both male and female employees and the majority percent of employees are men i.e., 60%. While the size of female employees are smaller, it still represents the substantial proportion of the dataset. The bar graph 3.3 depicts the age distribution of employees in the company, starting from the age of 18 to the age of 60. From the pattern, we can assume that the workforce of the company indicates a skewed age distribution since the majority of the employees are from the age of 28-40.

3.1.2 Bi-Variate Analysis

Bivariate analysis is a fundamental step our research which helps us to understand a relationship between two distinct variables. Through this analysis we can explore the dependencies, the correlations and the associations exists between these variables, recognizing how one variable have influence on the other. Similar to the univariate analysis, the bivariate analysis can be performed on variables irrespective of their types, either numerical or categorical. Scatter plots and correlation measures utilizing

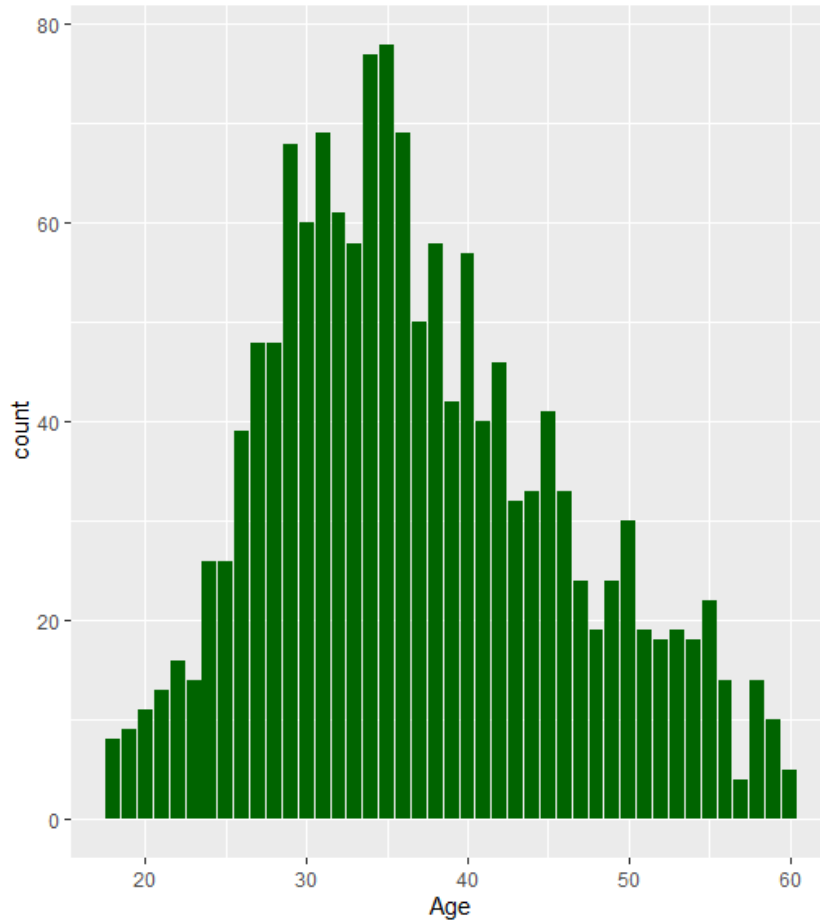


Figure 3.3: Employees Age group

the Pearson's and Spearman's coefficients offer valuable insights into the direction and strength of correlations for numerical variables. In terms of categorical variable analysis, the Chi-Square test is the best option. The Chi-Square Test reveals the statistical consequence of a relationship between the variables [8]. If the probability of the chi-square test is 0, it depicts that both the categorical variables are relatable and if its 1, the vice versa. The relationship between the variables are significantly 95% confident if the probability is less than 0.5. The Chi-Square equation [8] can be written as:

$$X^2 = \frac{\sum(O - E)^2}{2} \quad (3.1)$$

where O is the observed frequency, E is the expected frequency. This bivariate analysis can also be performed with a categorical variable and a continuous variable using correlation value which lies between -1 & +1, which can be expressed as [8]:

$$\text{Correlation} = \frac{\text{covariance}(X, Y)}{\sqrt{\text{Var}(X) * \text{Var}(Y)}} \quad (3.2)$$

These statistical tests and visualizations not only helps us to understand the relations but also facilitate the interpretation of the findings and also paves our way into further data exploration.

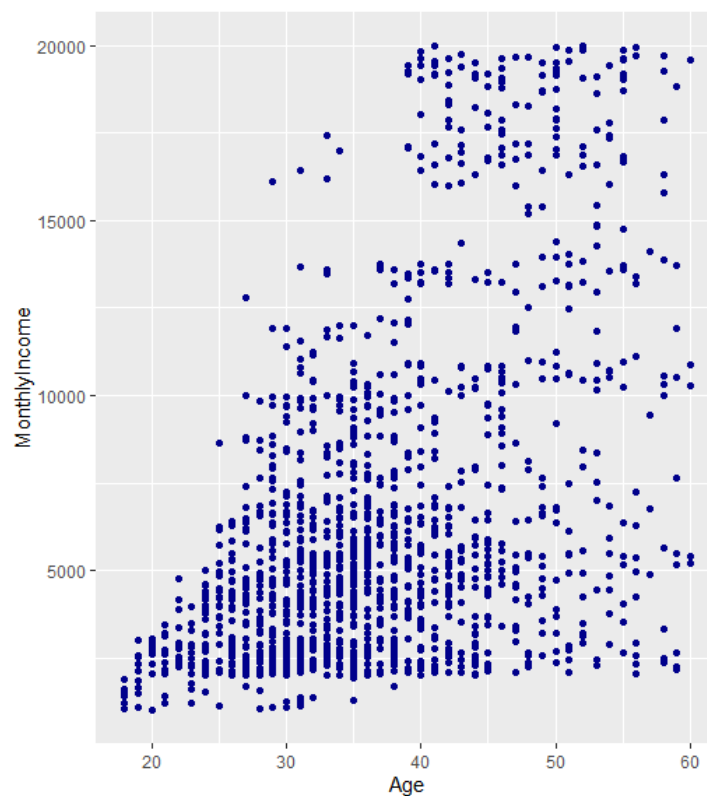


Figure 3.4: Employees Salary based on their age

The scatter plot 3.4 illustrates the relationship between the age of the employee and their monthly salary. Observing the pattern, we can infer that the correlation between the employee's age and salary is positive, which implies that the salary increases accordingly with your experience. This suggests that the experienced candidates are recognized by the organization and rewarded for their hard work.

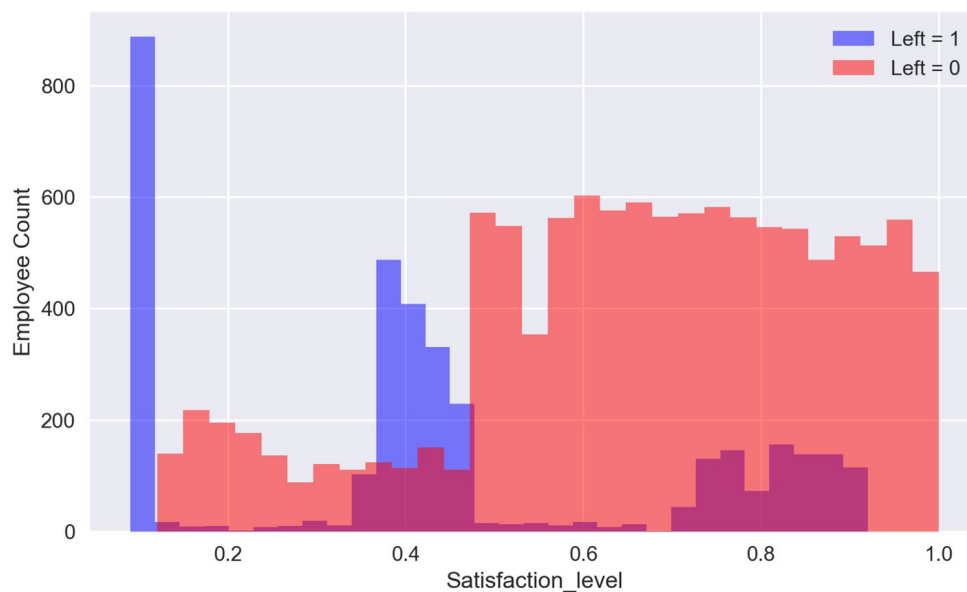


Figure 3.5: Satisfaction vs Attrition [8]

The satisfaction level versus the people who left the company [8] in the figure 3.5 illustrates an important relationship between employee satisfaction level and attrition within the company. It shows that the employees with lower satisfaction i.e., 0.2 are more likely to leave the organization. This provides valuable insights into the factors which influences the employee turnover and organizational stability.

3.2 Data Preprocessing

Data Preprocessing plays an important role in transforming the raw unstructured data into structured interpretable data, which helps in predicting the target more efficiently with better accuracy. Raw, real world data like text, video, images can be in a bad shape which definitely affects the machine learning models, if used without treating it with some techniques like taking care of the missing values, converting the categorical values into numerical values, removing the outliers, etc. The outcome of the machine learning model would be dirty if we use the raw dirty data.

3.2.1 Missing Values

The most crucial and one of the common challenges in the data preprocessing is handling the missing values in the dataset. The missing values could arise due to various reasons

like data entry errors, poor responses from the surveys, faulty generation of neural data in sensory equipments. Progressing with the research without filling the gaps may lead to biased or inaccurate results. The missing values can be handled in multiple ways:

- Dropping the fields/features
- Imputation
- Multiple Imputation

Likewise, we have to check if the data is clean, which means does it have any missing values or not. As per our inference from the figure 3.6, there is no missing values, so we can skip this step of pre-processing.

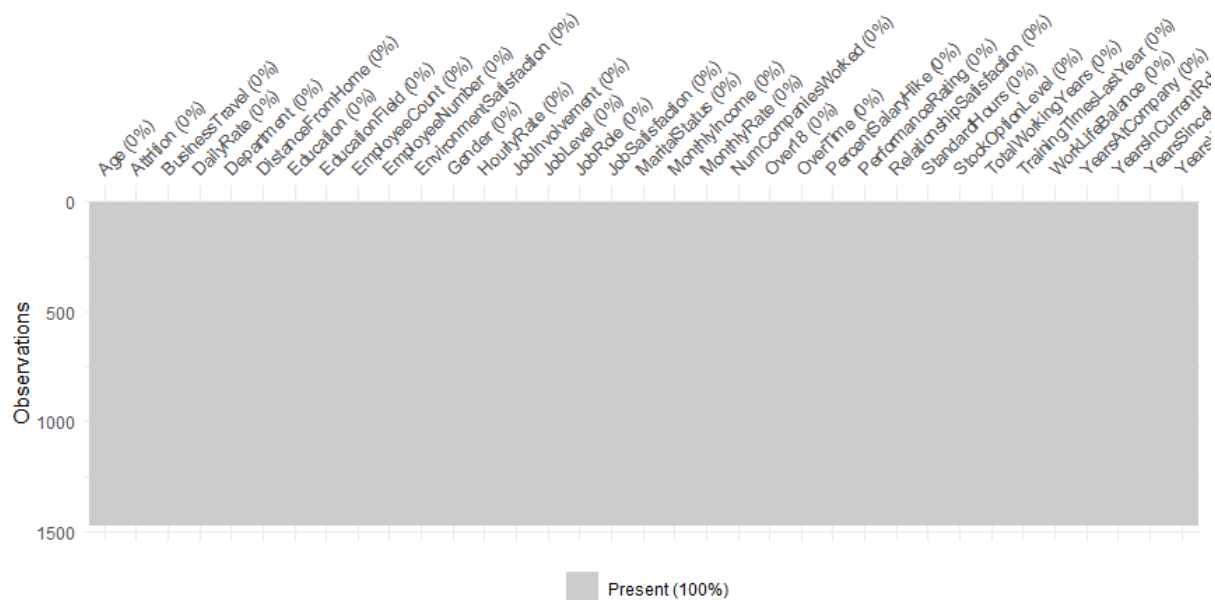


Figure 3.6: Missing Values

3.2.2 Outliers

Outliers are data points that are actually different from the rest of the data which stands out. The main causes for these outliers can be error during data gathering or collection. These can have significant impact on statistical analysis and on the efficiency on machine learning models. Many tests such as Grubbs test or Dixon Test can be used to identify these outliers in the data. Mostly they will be removed from the dataset to improve the model's accuracy and sometimes even can be handled.

In our case, the box plot has been generated for all of the features that displays the median, IQR (Interquartile range) and the outliers. Figure 3.7 illustrates the presence of minimal outliers in certain features, such as MonthlyIncome, TotalWorkingYears, and YearsAtCompany. This means that some of the employees in the company have significantly higher or lower values for these features than the whole majority of the employees.

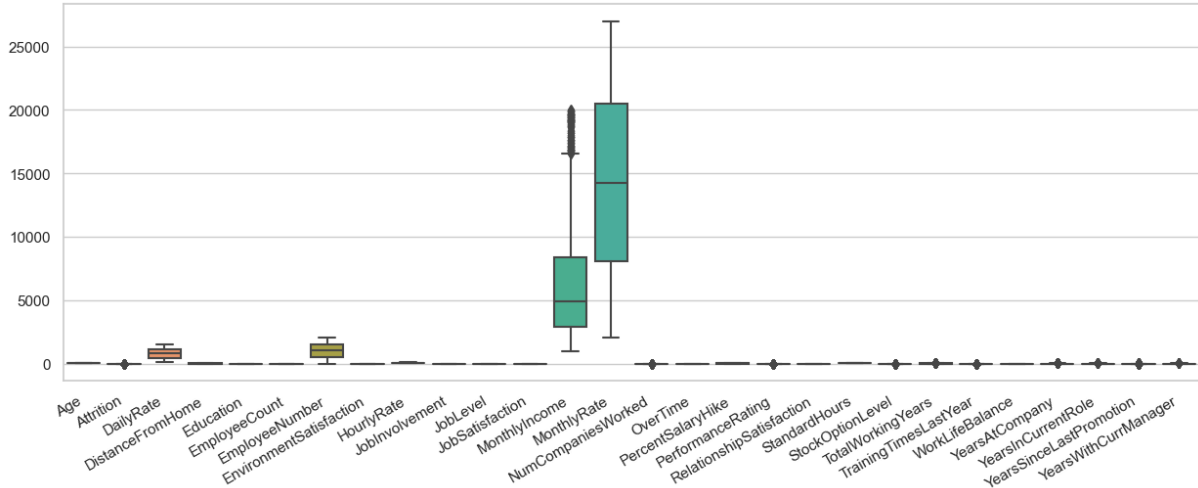


Figure 3.7: Outliers in our employee dataset

3.2.3 Correlation of different features

The term '**Correlation**' can indicate any type of relation between two (bivariate). In other words, the correlation denotes the strength of linear relationship between features in statistics [21]. These correlation are very useful in anticipating the predictive relationship that can be implemented on real-world situations. The equation for the correlation based on the feature selection [4] is,

$$r_{zc} = \frac{K\bar{r}_{zi}}{\sqrt{K + K(K-1)\bar{r}_{ii}}} \quad (3.3)$$

Where r_{zc} is the relation between the features and class variable, K represents the number of features. The \bar{r}_{zi} represents the mean value of the correlated feature class and the \bar{r}_{ii} denotes the mean value of inter correlated features.

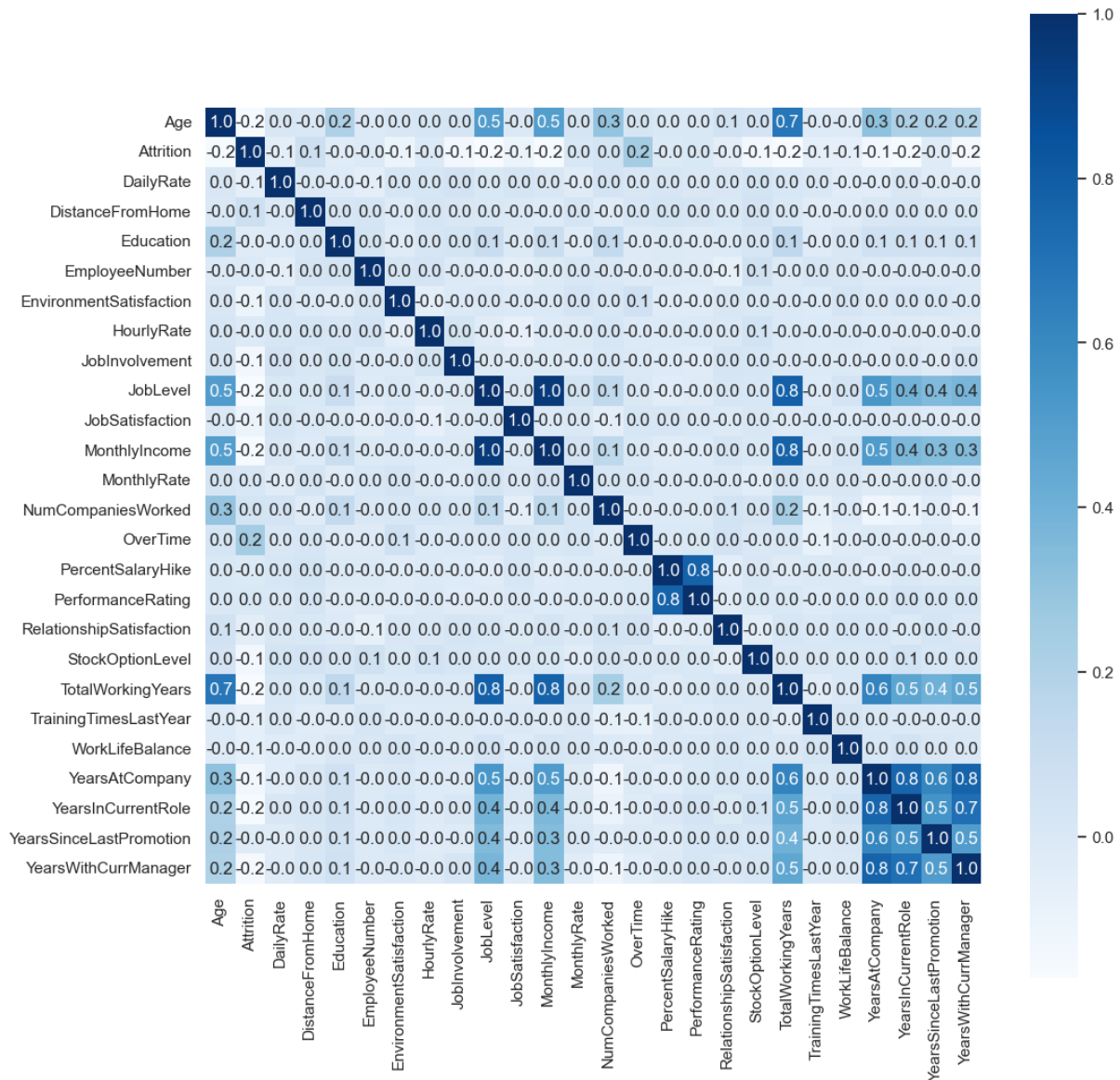


Figure 3.8: Feature Correlation

A correlation coefficient of '1' indicates a perfect positive correlation and '-1' denotes the perfect negative correlation. If the correlation is 0, then the variables doesn't actually have any linear relationship. From the figure 3.8, We can infer that the *JobLevel* and *TotalWorkingYears* have strong positive correlation with the coefficient of 0.8 and following by the variables *MonthlyIncome* and *TotalWorkingYears* with the same coefficient value. The variables with -0.2 negative correlation coefficient is the *JobLevel* and *Attrition*, which means if the value of the job level increases, the results on the Attrition will decrease. We also selected 2 unimportant features using correlation, *BusinessTravel* and *EducationField*, and removed them since they don't have any influence on the prediction of Attrition.

3.2.4 Data type handling

Data that has been categorized but doesn't follow any order or hierarchy is known as categorical data. Even though if its categorized properly as Male & Female for a Gender feature, it wouldn't be useful if its not numeric. There are many different encoding methods to convert the categorical features into numerical such as One-Hot Encoding, Label Encoding, etc. In our dataset, we could notice from the figure 3.9, some of the characteristics are object datatype, which needs to be encoded to numerical datatype before we can use them with our machine learning models, as they can't handle the categorical object features. We used the label encoder, which is the simplest encoding method that assigns an integer value for each category. This method works best on the categorical features and it helps a lot by enhancing the performance of our machine learning model.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 33 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Age                                   1470 non-null   int64
1   Attrition                           1470 non-null   int64
2   DailyRate                           1470 non-null   int64
3   Department                           1470 non-null   object
4   DistanceFromHome                    1470 non-null   int64
5   Education                           1470 non-null   int64
6   EmployeeCount                       1470 non-null   int64
7   EmployeeNumber                      1470 non-null   int64
8   EnvironmentSatisfaction              1470 non-null   int64
9   Gender                               1470 non-null   object
10  HourlyRate                           1470 non-null   int64
11  JobInvolvement                       1470 non-null   int64
12  JobLevel                             1470 non-null   int64
13  JobRole                              1470 non-null   object
14  JobSatisfaction                      1470 non-null   int64
15  MaritalStatus                       1470 non-null   object
16  MonthlyIncome                       1470 non-null   int64
17  MonthlyRate                          1470 non-null   int64
18  NumCompaniesWorked                  1470 non-null   int64
19  Over18                              1470 non-null   object
20  OverTime                            1470 non-null   int64
21  PercentSalaryHike                   1470 non-null   int64
22  PerformanceRating                   1470 non-null   int64
23  RelationshipSatisfaction             1470 non-null   int64
24  StandardHours                       1470 non-null   int64
25  StockOptionLevel                    1470 non-null   int64
26  TotalWorkingYears                   1470 non-null   int64
27  TrainingTimesLastYear               1470 non-null   int64
28  WorkLifeBalance                     1470 non-null   int64
29  YearsAtCompany                      1470 non-null   int64
30  YearsInCurrentRole                  1470 non-null   int64
31  YearsSinceLastPromotion              1470 non-null   int64
32  YearsWithCurrManager                1470 non-null   int64
dtypes: int64(28), object(5)
memory usage: 379.1+ KB
```

Figure 3.9: Features' Datatype

3.2.5 Addressing the Imbalanced data

An imbalanced feature in a dataset is a variable containing classification data that has uneven class proportions. In the recent times, the problem with the imbalanced learning problem has been drawn a significant amount of interest in academics, research and financial projects. The major fundamental problem with the imbalanced dataset is the ability to reduce the performance any machine learning algorithms [17]. The algorithms tend to yield unfavorable outcomes and inaccuracies in the analysis when presented with imbalanced datasets. Dealing with the imbalanced dataset is a vital aspect of improving the robustness and efficiency of the machine learning models. We have several techniques in python to address the imbalanced datasets. In order to achieve balance, these strategies either oversample the minority class or undersample the majority class in the feature. We know that our target variable, *Attrition* was skewed and has to be handled to make efficient analysis 3.10. We decided to go with the SMOTE (Sythetic Minority Over Sampling) technique that generates artificial representations of the minority class, encouraging a more fair distribution of class representation.

```
In [6]: empcpy['Attrition'].value_counts()
Out[6]: No      1233
        Yes      237
        Name: Attrition, dtype: int64
```

Figure 3.10: Distribution of Attrition before Oversampling

After implementing the SMOTE in our dataset, we can notice that the minority class *yes* was oversampled and the entire dataset has balanced 3.11. This will comparatively improve the efficiency of all our machine learning models. In order to determine how well the balanced dataset performs, we will compare the accuracies of the imbalanced and balanced datasets in our machine learning models.

```
In [22]: empcpy_balanced['Attrition'].value_counts()
Out[22]: 1      1233
        0      1233
        Name: Attrition, dtype: int64
```

Figure 3.11: After Oversampling

3.2.6 Normalization

This stage of preprocessing is to determine whether all the features in the dataset is standardized or falls into a single, consistent scale before moving on to the machine learning models. The process of doing this is known as Normalization and its one of the crucial steps in data preprocessing step as well, since this will ensure that each of the attribute contributes proportionally to the learning process and prevents disproportionately influencing attributes to our analysis. Most of the common techniques for normalization are Min-Max Scaling, where the values would be changed to a range of numerics between 0 & 1, Standard Scaler, in which a standard normal distribution with a mean of 0 and a S.D. of 1 is created to standardize the feature distribution with a formula [22],

$$z = \frac{x - \mu}{\sigma} \quad (3.4)$$

where,

- z is the output standardized value
- x is the original value
- μ is the mean of the attribute
- σ is the standard deviation of the attribute

and Z-Score standardization, etc. Normalizing the entire dataset improves the performance of our machine learning models, reduces the impact of outliers, and offers general stability over a variety of unique datasets. In our case, we used the standard scaler to standardize our values of the dataset.

3.3 Feature Importance Analysis

In order to minimize the feature size by eliminating unnecessary features and improve the accuracy of the machine learning models, we will assess the significance of each feature in the dataset in this last stage. Additionally, this can prevent overfitting as well. This process can be performed using many methods and the one we are using here is the **SelectKBest** from SKLearn package, which is one of the common method to calculate the importance of the features. This method basically examines the k best

features based on the scores from the univariate statistical test which calculates the ANOVA F-statistic for each feature and compares it. The variation in averages between the classes in relation to the variability within each class is the F-Statistic value. The F-statistic values of the features of our dataset has been tabulated below 3.12, sorted based on their statistical significance in relation to the target variable *Attrition*. From the image, we can infer that the feature *OverTime* is the most significant feature with the F-statistic value of 94.6565 on the top, which also confirms that this feature would be the most influencing factor for our analysis. We also removed some unimportant features which has zero relation with the main target variable such as, business travel, education field, etc.

Selected Features and Importance Scores:

	Feature	Importance
21	OverTime	94.6565
27	TotalWorkingYears	44.2525
13	JobLevel	43.2153
16	MaritalStatus	39.5998
31	YearsInCurrentRole	38.8383
17	MonthlyIncome	38.4888
0	Age	38.1759
33	YearsWithCurrManager	36.7123
26	StockOptionLevel	28.1405
30	YearsAtCompany	27.0016

Figure 3.12: Significant features with their F-stat Value

Methodology

In this section, we outline the methodology employed to predict the employee attrition rate through the applications of machine learning algorithms. The primary goal is to create a prediction model that can accurately pinpoint the variables impacting our analysis's results, including whether or not an employee leaves the organization and to compare the model performances of both balanced and imbalanced dataset. This will enable the human resources department to implement crucial retention tactics.

4.1 Data Partitioning

The data has undergone preprocessing, and in order to feed it into machine learning models, it must be divided into two sets: one for training and the other for testing. It is vitally essential to carefully partition them to ensure that our model is robust and unbiased. Each divided subset serves a different purpose through out the model evaluation pipeline. Usually the training set constitutes a larger portion of the dataset and is used to train the machine learning model, in which the model learns about the data and its features, patterns, etc. Once its trained with the dataset, we can able to test it with a small portion of the data to predict the outcome, which is the testing set. This testing set is independent, completely hidden from the model, and used to evaluate the subset of data that the trained model uses to generate predictions. By this way of testing the model with an unseen data makes it more effective and robust. In our case, we split 80% of both the imbalanced & balanced dataset for training and the remaining

20% for the testing.

4.2 Hyperparameter Tuning using Grid Search

Finding the ideal hyperparameter for your dataset during machine learning model training is known as hyperparameter tuning, and it will boost the model's performance. Every unique dataset will have a particular hyperparameter, which can be discovered by manually or automatically experimenting with several subsets using various techniques such as Bayesian optimisation, Grid Search, and so on. In simple terms, we'll be training the model with different parameters to find the efficient model that yields higher accuracy. Since it trains the model with different set of hyperparameters, its computationally intensive, so on weaker system, this tend to work a bit slower. For an instance, thee hyperparameters could be the number of layers in a decision tree or the number of nodes in a neural network or learning rate. By default, there is no set of predefined rules to identify the optimal hyperparameters, so the only option is to trial-and-error with different set of parameters.

Because it requires a lot of computing power, we will only use this to improve a few chosen models that have lower accuracy. In our research, we have decided to go with the Grid Search because of the complexity in employee attrition prediction task. Basically, Grid search insists on a list of hyperparameters and the performance metric to determine the best fit by running through a combination of iterations. The structured and the investigative exploration of complex hyperparameters makes it more suitable for our estimation problem. This method's extensive methodology also guarantees that no parameter is left out without verification.

4.3 Machine Learning Models

The successful prediction on the attrition rate depends upon the appropriate machine learning model selection and implementation of those models. This section illustrates an overview of all the machine learning methodologies employed in this study. From logistic regression to boosting models like AdaBoost, XGBoost, and so on, our investigation covers both conventional and cutting-edge machine learning techniques. Furthermore, we'll be comparing the performances of the models built on both the imbalanced and

balanced dataset to analyze the efficiency of over sampling.

4.3.1 Logistic Regression

Logistic regression, as a cornerstone in the realm of classification algorithms, considered as the crucial method in our pursuit on predicting the employee attrition rate, that estimates the parameters of the logistic binomial regression model [2]. Logistic regression is specifically designed for binary classification problems, making it the most efficient method in estimating the outcome variable that represents two or more classes.

The main principle of *Logistic Regression* is to transform a linear combination of input features to a range bounded between 0 and 1 by employing the logistic function also known as sigmoid function. The likelihood that an employee belongs to the interest class is subsequently determined by understanding this range of processed output. A simple logistic regression can be described as:

$$Y = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}} \quad (4.1)$$

where Y represents the dependent variable (*Attrition*) and the x represents the independent features of the dataset [2]. In our study, the goal is to predict the target variable *Attrition* by the help of all other independent features and examine whether the employee is likely to quit the company or not. Each characteristic's coefficients show how much and in which direction each feature affects the anticipated result. This transparency provides better interpretability on the factors influencing the attrition of employees. It's simplicity and the interpretability makes it much crucial in predicting the number of factors which examines if the employee is going to quit or not. Analysing the coefficients after implementing the model, we interpret that the most contributing factors to the employee prediction were *YearsAtCompany*, *OverTime*, *YearsSinceLastPromotion*.

4.3.2 Decision Tree

Decision Tree(DT) Algorithms are very powerful algorithms, capable of fitting complex and huge datasets and have been widely used on many complicated applications like medical analysis, financial predictions, etc. As the name suggests, it creates a tree-

like structure with two nodes; decision node and leaf node, where each and every internal node corresponds to a decision based on a feature and each leaf node represents the predicted outcome, which can also make predictions on both the categorical and numerical features. Basically to improve human readability, the decision tree splits the target function into if/then rules [2]. By doing so, breaking the complex decisions from the top of the tree also known as 'root' into smaller subsets. One of the crucial advantages of decision trees is their interpretability, as the fundamentals are simple and precisely outlined, helping to make it easier to identify the variables that affects attrition.

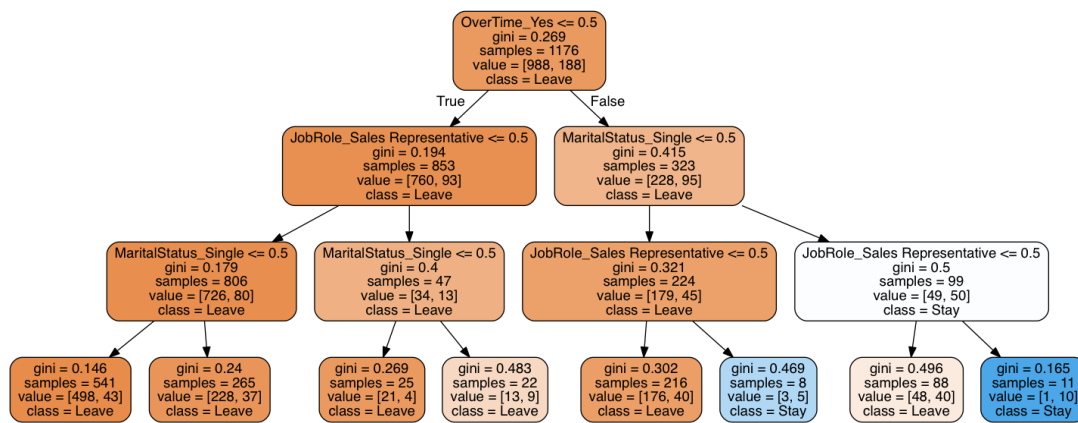


Figure 4.1: Decision Tree Based on the features from IBM HR Dataset [2]

Our study's objective is to pinpoint the highly correlated characteristics that have a direct impact on workers' decisions to remain with or leave the company. A decision tree has been generated 4.1 based on the features from the imbalanced dataset, starting from the root node *OverTime*. Followed by the leaf nodes with the predicted attrition rates of the employees who fall into the corresponding categories. From the tree, we can interpret that the employees who work overtime are more likely to quit the job and with lesser job level. It also suggests that the increase in the *YearsAtCompany* decreases the attrition, which implies that the employees who stayed longer are likely to stay back in the company. Decision trees do not reflect the informal relationships between features, so an employee's decision to work overtime does not automatically result in his termination from the company; instead, other factors, such as work-life balance, may be more likely to cause him to leave.

4.3.3 Random Forest

A random forest is an ensemble method for classification and regression which is a combination of many decision trees to build a strong learner to get more efficient and stable prediction. It seems to be an vital means for determining the variables in the objective for attrition estimation. This method addresses the problems faced with the single decision trees, particularly the tendency to overfit the data. By getting an average of predictions from number of trees, this method tends to attain a balance between acquiring insightful patterns and also avoiding the amount of excess details in the results.

In the context of employee attrition, since we have numerous amount of features contributing to the outcome, random forests' ability to handle high-dimensional data would be more suitable and will be much efficient in prediction. Not only does it support an enormous amount of features, but it also determines the relative value of each feature that has been fed. It can also handle the imbalanced datasets, an usual challenge in attrition prediction where the amount of employees who plan to stay would be more than the ones about to quit. The consideration of multiple trees and their unified estimations mitigate the risks arise from imbalanced data, thus resulting in more reliable assessment.

Initially, it begins with creating individual decision trees, then training them with a multiple subsets of data and choosing a random set of features from them. Since the aspects of the datasets remains diverse, its way precise to come up with a unique perspective to the overall predictions. Finally, aggregating the results from all the decision trees, delivering an unbiased and accurate output, efficient than a prediction from a single decision tree. Although, the predictable power of the random forest is immense, the interpretability of results is not a cakewalk, because of the numerous amount of trees involved.

The most contributing factors for employee attrition in random forest method are *MonthlyIncome*, *JobSatisfaction*, *OverTime*.

4.3.4 Support Vector Machine

An algorithm which separates different classes using an optimal hyperplane within the data is called Support Vector Machines. It excels both in classification and regression problems and especially its known for its robustness in predicting employee attrition. In our study, the hyperplane serves as a divider between the employee who willing to stay in the company and the ones who planning to quit.

One of the core strengths of SVM is its ability to handle high dimensional data, a common scenario in employee attrition where we have numerous amount of features that contributes to the decision making process. It handles the complexity effectively by identifying relevant features and their behaviours. By focusing deeply on the critical features, SVM not only boosts the accuracy but also yields insights into the key determinants of attrition within the data. An efficient way of determining the optimal hyperplane between the classes is by the kernel trick, that converts the input space into an high dimensional space. This way of implementing the classification makes it invaluable in non-linear relationships between features to uncover important patterns and insights. It is more flexible and adaptive to diverse types of datasets thanks to its multiple kernel functions, such as poly kernels or the Radial Basis Function (RBF), enabling it to produce reliable predictions on attrition estimation tasks.

The figure 4.2 is an example of support vector machine. The red line which separates the classes into two is the optimal hyperplane and the blue and green dots are features of different classes. The two blue dots that lies closer to the red line are the support vectors. The dotted lined margin is the distance between the red line and the support vectors.

Similar to the random forests, due to the high dimensional data handling, the interpretability of this model could be challenging. For SVM to be more effective, feature scaling is recommended, because the algorithm is sensitive to magnitude of the features. Standardizing the features likely to influence the decision making process of the SVM model. The most contributing factors for employee attrition in support vector machine are *OverTime*, *YearsAtCompany*, *TotalWorkingYears*.

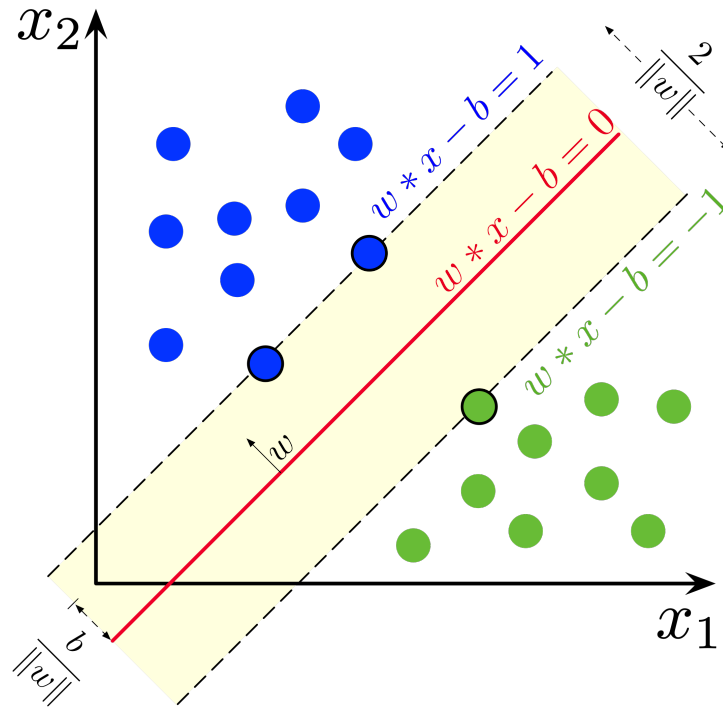


Figure 4.2: Support vector machine [23]

4.3.5 k-Nearest Neighbors

One of the simplest and most effective approaches in our quest to identify the variables involved in the decision-making process of employee attrition prediction is the k-Nearest Neighbour method. As a model without parameters, k-NN relies on the proximity of data to derive predictions, which is also because it follows instance based learning approach. The flexibility of k-NN is invaluable when it comes to predicting the employee attrition which involves so many determining features.

The primary idea behind k-NN is to categorise a data point in the feature space according to the majority class of its k-nearest neighbours. The distance between new data point and each training point can be done using various methods like Euclidean distance, Manhattan distance and Hamming distance. For an instance, the euclidean distance, also known as ruler distance which is an extension of Pythagorean theorem, can be calculated using the formula [18]:

$$ED(x, y) = \sqrt{\sum_{i=1}^n |x_i - y_i|^2} \quad (4.2)$$

The balance between the model's flexibility and robustness depends on the choice of

the hyperparameter k that represents the number of associated neighbours. The simplicity and the easy implementation makes it even more convenient and preferable. Because this model lacks a predetermined structure, it can automatically accept complicated relationships found in the data without the assistance of an outside party.

Even though the model being very simple, its success depends upon the careful consideration of appropriate distance metric and the choice of the dimension. The dataset's properties influence the distance metric selection, which is an important factor in assessing the similarity between data points. Feature Scaling has to be performed in order to get better results from this model, since this uses the nearest neighbour to make predictions.

The performance of the k -Nearest Neighbor algorithm would be degraded if treated with an imbalanced dataset like the one with employee attrition where the employee quitting the company is far lesser than the one staying in. As a result, it worked well with the dataset that we oversampled using SMOTE before using the original unbalanced dataset, which could have produced findings that were skewed in favour of the majority class.

Interpretability of k -Nearest Neighbour is way straightforward, as the prediction relies on the behaviour of neighboring data points. The transparency of k -Nearest Neighbors (k -NN) facilitates a direct interpretation of predictions, allowing stakeholders to comprehend the reasoning behind model outputs effortlessly. The use of nearby data points creates a concrete and understandable relationship between input properties and predictions, which increases confidence in the model's ability to make decisions.

4.4 Boosting Algorithms

Machine learning is a collection of diverse algorithms, each designed for unique specific purposes. Among these, the boosting algorithms are the powerful ensemble methods which aims to enhance the performance of the prediction methodologies. In this section we'll be focusing on the boosting machine learning algorithms on our imbalanced and balanced datasets such as AdaBoost, CatBoost, etc. The primary distinction between boosting models and conventional machine learning models is that the former will often train a sequence of models with the goal of fixing mistakes caused by the latter.

4.4.1 AdaBoost

AdaBoost, also known as Adaptive Boosting, is a technique of group learning method based on the weighted samples. So, it basically tries to improve the efficiency of the model by analysing the consecutive set of weak learners that focuses on the misclassified instances [2]. By iteratively fixing the errors occurred by the weak learners on previous classification and emphasize the misclassified instances, therefore building a strong and accurate predictive model. The main speciality of this model is its adaptability to the weights of the training instances in every single iteration that allows the model to concentrate on different challenges, which also makes this model more favourable for non-linear complex datasets.

Eventhough, the predictive capability of the weak learners, also known as *stamps* are not the best, it serves as the building blocks of AdaBoost to derive the final strongest classifier. The final model is the collection of all these weak learners, with each contributing on the areas where others might stumble. This ensemble approach makes this model to achieve superior at most performance than other weak classifiers. It can also deal the noisy large-scale dataset and outliers. By assigning higher weights to he misclassified instances and fine tuning the model's focus on the challenging datapoints and so avoiding the impacts of outliers. This makes the model more robust and efficient on real world problems.

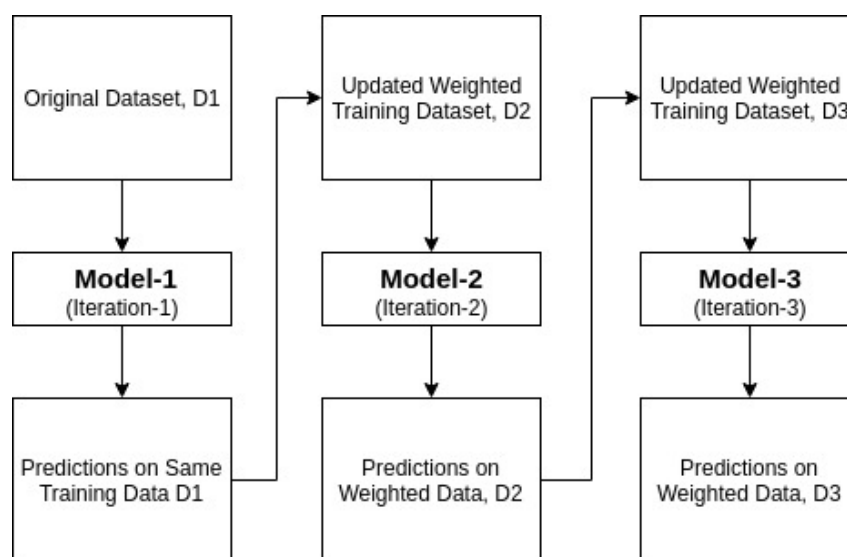


Figure 4.3: Basic working of Adaboost [10]

It's not only limited to specific types of classifiers like decision trees but also can be applied to other classifiers as well. Because of its versatility, researchers can take use of the strengths of various weak learners and modify the algorithm to fit the unique features of the dataset. It has excelled in many areas of the research like finance, biomedics and broadcasting because of its flexibility.

Despite the various applications of AdaBoost, it also comes with certain challenges like overfitting of the model, if the weak learners are too noisy and complex. But this can be addressed by cross validation which fine-tunes the hyperparameters of the stumps. The performance can be affected if the weak learners are too weak, that strives for a balance between the complexity and simplicity.

In our context, AdaBoost algorithm, through its iterative mechanism and the ability to handle complex, non-linear data like HR employee dataset, estimates the factors that influences attrition such as JobSatisfaction, YearsWithCurrManager. Since the attrition problem has to addressed proactively to prevent employees quitting the company, AdaBoost can be of a great assist in identifying the factors by its ensemble learning approach, which combines the results of multiple weak learners and provides a robust and much accurate results.

4.4.2 CatBoost

In the domain of machine learning, CatBoost is one of the cutting-edge gradient boosting algorithm. This model handles categorical data flawlessly and doesn't require a lot of preprocessing, as evidenced by the name, which was taken from the word 'category'. This ability of handling the categorical variables naturally mitigates the impact of issues from the encoding techniques like one-hot and label encoders. This is achievable because of an innovative implementation of the coherent algorithm of gradient boosting on decision trees. Similarly it learns the errors from its predecessors by employing a gradient boosting algorithm that involves constructing a group of decision trees. The diagram 4.4 displays the structure of the CatBoost algorithm.

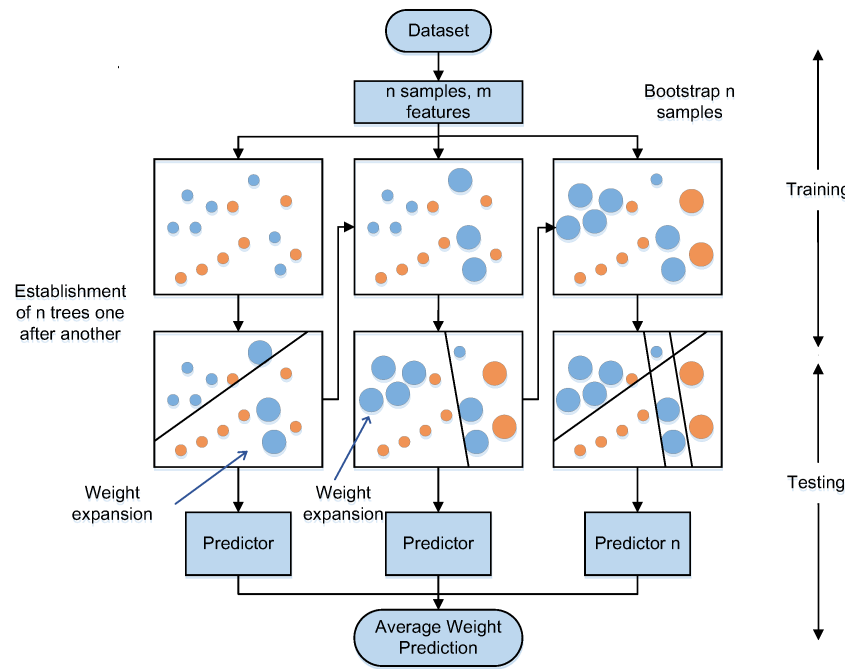


Figure 4.4: Architecture of the CatBoost Algorithm [28]

CatBoost is accompanied with a novel strategy which is unique from all other traditional gradient boosting models where the categorical variables would be converted into numerical features during the tree-building process itself. This eliminates the need of the external encoder, thus simplifying the task of pre-processing the employee dataset. This is particularly helpful while handling different projects with complex real-world huge datasets. Unlike AdaBoost, its immensely resistant to overfitting which is mainly due to the incorporated special technique called 'ordered boosting' that addresses the problem of overfitting by explicitly considering the ordering of categorical variables while the boosting process. This ability to handle overfitting and complex datasets strike a balance between the complexity and simplicity, thus achieving higher accuracy on predictive performance on most of the datasets.

It also aids in handling the missing values in the dataset which is super common in machine learning problems in these days. Without the need of imputation, it can fill the missing space using the available information that has been learned from the crucial instances. This makes modelling easier and increases CatBoost's adaptability in situations where missing data is common and also reduces the burden of pre-processing the data with many techniques.

The 'ordered boosting' would also assist in optimizing the training process by

utilizing the natural order of categorical variables and boosts the accuracy of the model. This also makes the training process faster. So particularly in large scale datasets like financial broadcasts and medical estimation, this can reduce the computational time significantly to train the predictive model. It's often equipped with a competitive predictive accuracy apart from the ability of handling the categorical variable, which makes this a strong competitor in machine learning contests. It further simplifies the model by automatically selecting the optimal hyperparameter through an powerful grid search method, so the focus for tweaking the hyperparameters is not necessary rather could be aiming on improving the performance. All these qualities makes it invaluable on the prediction of employee attrition.

4.4.3 Gradient Boosting

Gradient Boosting (GB) builds a new strong ensemble model by combining the set of weaker models sequentially by minimizing their loss function. Each new model that makes use of the loss function increases the model's overall accuracy and more closely fits the observations. It's renowned for the ability to build highly accurate predictive models. Similarly it uses decision trees for the weak learners and sequentially correct the errors made on previous models by them to construct a robust model. In the context of handling the type of data, it can go well with both categorical and numerical data, thus more suitable for the employee dataset which has both numerical features like age, monthly salary and categorical features like marital status, attrition, etc and also simplifying the process of preprocessing. By introducing regularization and complexity management in the data it can prevent overfitting which can be efficient on machine learning process.

When applied to the data on employee attrition in the aim of predicting the attrition rate, gradient boosting focuses on several attributes and makes it specifically effective and relevant. It's more convenient with handling the mixed type of features in employee attrition HR dataset. Gradient boosting is also effective on imbalanced dataset like the one we have where the number of employees leaving would be lesser than the employees staying in the company. With its iterative nature, the forecasts on the minority class (Attrition = "Yes") in subsequent rounds has significantly improved, thus delivering a model that is skilled at uncovering patterns related to employee turnover.

This algorithm can also handle the missing data efficiently which is very common in HR datasets. The filling of these missing values doesn't have any compromise on the model's performance. Although it reduces the efforts of imputations that has to be done manually for the missing values and makes accurate predictions with the available information without introducing bias. This model also consists of regularization technique like tree depth control and shrinkage to improve the model's generalization performance. This is essential on predicting employee attrition, as it needs to capture the pattern without influenced by too much noise from the training data. Thus it performs well on unseen data by having a balance between complex relationships and preventing overfitting.

$$L = - \sum_{i=1}^n y_i \log(p) + (1 - p) \log(1 - p)$$

Figure 4.5: Loss function in Gradient Boosting [24]

The loss function in gradient boosting classification is given by the formula 4.5. If it comes to interpretability, the model gives out the important features to know the major factors that influenced the outcome. In our case, the factors that contributed to the employee attrition was *MonthlyIncome*, *OverTime*, *JobSatisfaction* and *EnvironmentSatisfaction*. Therefore, if there is a poor work-life balance and inadequate pay, employees are more inclined to leave the organisation. Additionally the model's hyperparameter tuning provides more flexibility in calibrating the special characteristics of the dataset and the goals of attrition prediction.

4.4.4 XGBoost

XGBoost is nothing but an extended version of gradient boosting decision tree [10]. It's also known as Extreme Gradient Boosting that has an immense popularity in the realm of machine learning for its scalability and efficiency on predictive problem performance. Since its derived from the gradient boosting, its got the same technique of deriving the robust prediction by using the results of multiple weak learners, mostly decision trees. What makes it stand out from the original gradient boosting is that the ability of handling missing data, powerful regularization 4.7 techniques, and the capacity to

compute parallelly to boost the training process and mainly the scalability to handle large scale datasets makes it the first priority for the researches on boosting algorithms in machine learning.

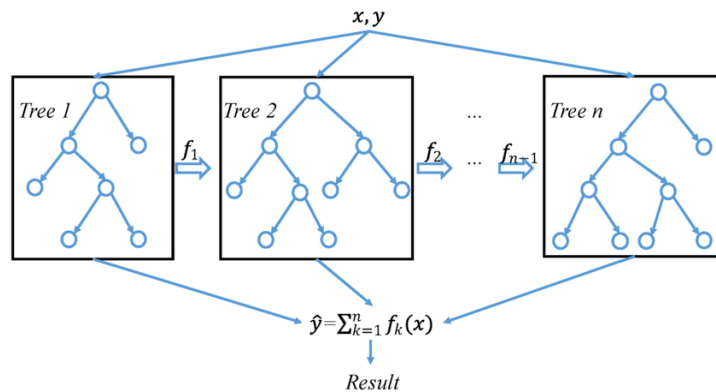


Figure 4.6: Architecture of XGBoost [25]

Likewise, this can handle both categorical and numerical features that eliminates the extensive need for preprocessing such as label encoder. So its efficient on employee database, as there are various number of features that are categorical like *MaritalStatus*, *JobRole* and so on. With the XGBoost machine learning model, managing missing data is no longer necessary because it can do imputation, which cuts down on computing time. Interpreting the model evaluation is also straightforward as it gives out the importance of features, thanks to the built-in method of XGBoost which can rank the features based on their influence on model's performance.

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

Loss function ↙
↘ Regularization

Figure 4.7: Loss Function & Regularization of XGBoost [26]

In terms of the predictive accuracy, XGBoost is invaluable, which often performs comparatively competent than other machine learning algorithms by the precision and recall. Accuracy is crucial when it comes to employee attrition since it allows the HR department to identify workers who are considering leaving and adjust retention tactics accordingly. The primary factor influencing employee attrition in both balanced and unbalanced datasets was *MonthlyIncome* and followed by the *Age* and

DistanceFromHome.

The accomplishments that XGBoost has made to the field of predictive modelling highlight how important it is for encouraging data-driven decision-making in workforce management.

4.5 Model Evaluation

Model Evaluation is a process of leveraging different metrics to analyze the performance of the machine learning models that we employed, to understand its strengths and weaknesses. It includes multiple metrics like Accuracy, Precision, Recall, F1-Score and so on, that's vital to assess and monitor the model's potency.

4.5.1 Confusion Matrix

Confusion matrix is one of the efficient method of evaluating a classifier with the detailed analysed report on the number of true positives, false positives, true negatives, and false negatives [8]. Another name for it is an error matrix, which is a matrix that shows how well a classification model performs when applied to a set of test data. The figure 4.8 represents the structure of a confusion matrix.

		Actual Class	
		1	0
Predicted Class	1	True Positive	False Positive
	0	False Negative	True Negative

Figure 4.8: Visualisation of a Confusion Matrix [27]

1. **True Positives (TP):** When the actual value is positive and the predicted value was also positive, that enables the actual class and the predicted class to match.
2. **True Negatives (TN):** When the actual value is negative and the predicted value was also negative, that also enables the actual class and the predicted class to match.
3. **False Positives (FP):** When the actual value is negative but the predicted value was positive, also known as type I error.
4. **False Negatives (FN)** When the actual value is positive but the predicted value turned out be negative, also known as type II error.

In simple terms, its a tabular summary of the number of incorrect and correct predictions by the classifier. Since our dataset in the employee attrition instance is unbalanced, using a confusion matrix to assess our model is a much better option. If the model shows high TP and TN and low FN and FP, it is regarded as a good model as per confusion matrix.

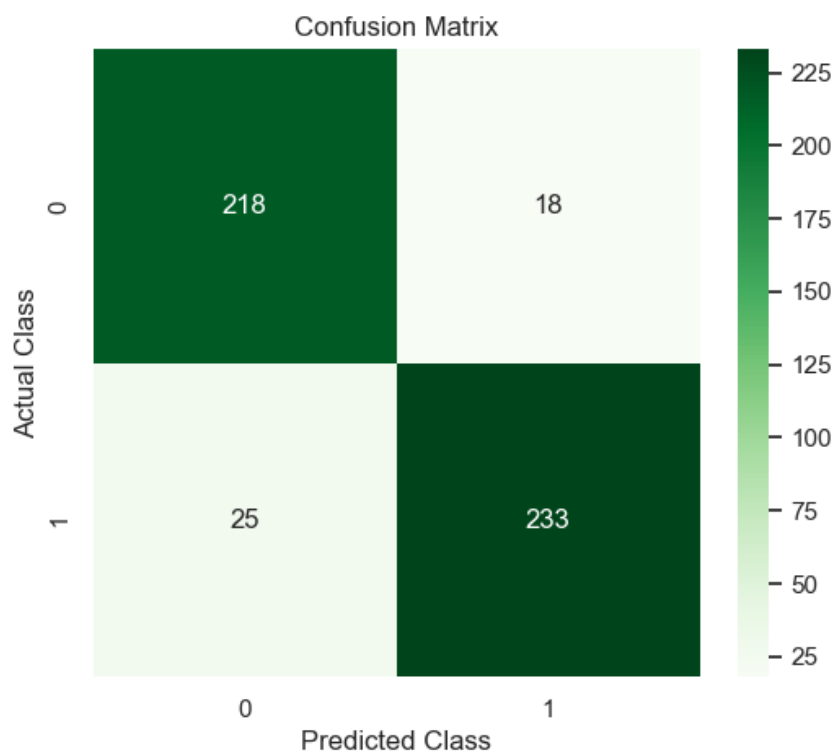


Figure 4.9: Confusion Matrix of the SVM model in balanced dataset

The figure 4.9 shows the confusion matrix of the SVM model trained and tested with the balanced dataset. From the values of the TP, TN, FP, FN, we can confirm that the model has performed really well with 91% of accuracy.

Accuracy

The term 'Accuracy' quantifies how frequently the model is making the right predictions. It's the ratio 4.3 between the correct predictions, which is the TP and TN, and the total predictions, which is the sum of TP, TN, FN & FP. It is a measure that indicates the correctness of a model. A good model should have higher accuracy.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \quad (4.3)$$

Precision

This is a measure of correctness achieved only in the true prediction, which can be viewed as the ratio between the TP and the sum of TP and FP. In other words, it indicates the fraction of all positively projected values that are truly positive. Ideally, for a good model precision should be higher.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4.4)$$

Recall

Recall, also known as sensitivity, is a measure of actual observations that are predicted correctly, which is the ratio between the TP and the sum of TP and FN. In simple terms, it is the number of positive-class observations that are really anticipated to be positive. The model would be considered as good if the recall value is higher.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4.5)$$

F1-Score

F1-score is simply defined as the harmonic mean of precision and recall, where the results lies between 0 and 1. The mathematical expression of F1-score is formulated

below 4.6. This maintains the balance between the precision and recall in a classifier. Preferably, the value of the F1-score should be higher.

$$\text{F1-Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.6)$$

The confusion matrix of the random forest method with imbalanced dataset has been visualized in 4.10, which clearly tells us that the model yielded high accuracy of 86% but may have some challenges to identify the positive instances correctly, indicated by the low recall (19%).

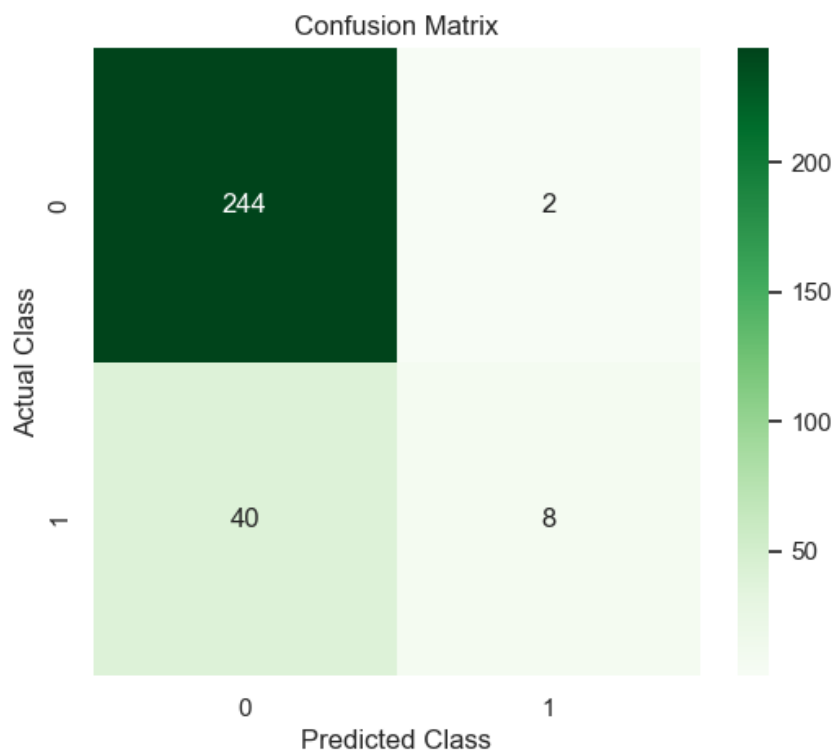


Figure 4.10: Confusion Matrix of the Random Forest with imbalanced dataset

4.5.2 AUC-ROC Curve

The *Area Under the Curve* (AUC) of the *Receiver Operating Characteristic* (ROC) is another widely used metric for evaluating the machine learning models visually, that offers an in-depth overview of their capacity for class discrimination. The ROC curve is the graphical representation of Sensitivity (which can be seen as *Recall* 4.5's equivalent) and Specificity (as seen in 4.7 formula), which is an exchange over various threshold values between the true positive rates and the false positive rates. A perfect classifier will yield

a AUC-ROC value of 1 that quantifies the area under this curve, which also reflects the model's robustness and accurateness. On the other hand, the value of 0.5 suggests that the model's more reliable on a random chance.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (4.7)$$

When applied to unbalanced datasets, such as the one we used to forecast employee attrition, where the proportion of willing employees to leave is significantly lower than that of ones willing to stay, this kind of evaluation is extremely beneficial. In terms of interpreting the curve in identifying the optimal threshold depends on our prediction task's requirements that balances the sensitivity and specificity, proving as a valuable tool to compare and evaluate the classifier's efficacy.

4.5.3 k-Fold Cross-Validation

In general, we train a machine learning using the training set and test the model with the testing set, which will evaluate the model once. Cross-validation is a procedure that evaluates the machine learning models by dividing the dataset into k folds. The model would be trained with different set of 'k' subsets and tested for k times to find the optimal performance. The overall metric of the model would be the aggregate of all the k folds. This assessment approach is more accurate and exact since it executes the training and testing procedures more precisely in each and every fold to prevent overfitting. Its easy interpretability using just two metrics, mean accuracy and standard deviation makes it well-liked. With a smaller standard deviation number and a greater mean accuracy, the model is considered proficient. The interpretability of this evaluation model makes it much efficient and popular.

Results

5.1 Comparison of Prediction Outcomes

In this section, we performed a comprehensive evaluation of all the machine learning models we employed in this study and also comparing the results of models under varied dataset conditions - *imbalanced* and *balanced*. The models ranging from the traditional models like decision tree, SVM to advanced ensembled models like Gradient Boosting, XGBoost. The table [5.1](#) provides a comprehensive overview of the efficiency of machine learning models deployed using the imbalanced dataset.

The accuracy of the models explains the overall correctness that ranges from 0.78 to 0.87. AdaBoost achieved the highest accuracy with 87%. With highest precision of 0.87, random forest indicates the lower rate of false positives and on the other hand, the lowest precision on DT is because of the imbalanced nature of the dataset. AdaBoost standing at the top with recall value at 0.68 and F1-score at 0.72, which shows its strong proficiency of estimating the attrition rate. SVM exhibits an hefty discrimination capabilities with AUC-ROC value at 0.82.

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.86	0.77	0.63	0.66	0.79
Decision Tree	0.78	0.61	0.62	0.61	0.62
Random Forest	0.86	0.87	0.58	0.60	0.81
SVM	0.85	0.78	0.59	0.61	0.82
kNN	0.83	0.64	0.56	0.57	0.70
AdaBoost	0.87	0.80	0.68	0.72	0.80
CatBoost	0.86	0.81	0.62	0.65	0.81
Gradient Boosting	0.84	0.69	0.60	0.62	0.80
XGBoost	0.85	0.74	0.64	0.66	0.79

Table 5.1: Performance Metrics on Imbalanced Dataset

This table 5.2 presents the performance of machine learning models trained on a balanced dataset which was treated using SMOTE (Synthetic Minority Over-sampling Technique) that aimed to improve the efficiency of the models, comparatively to the ones implemented with the imbalanced dataset. In this evaluation, notably the accuracy of the models has been skyrocketed, ranging from 0.81 to 0.95, with the highest in CatBoost on all the performance metrics, which is because of the artificial oversampling of our dataset. In terms of the AUC-ROC curve, the CatBoost managed to get the highest i.e., 0.99, which has been visualized below 5.1.

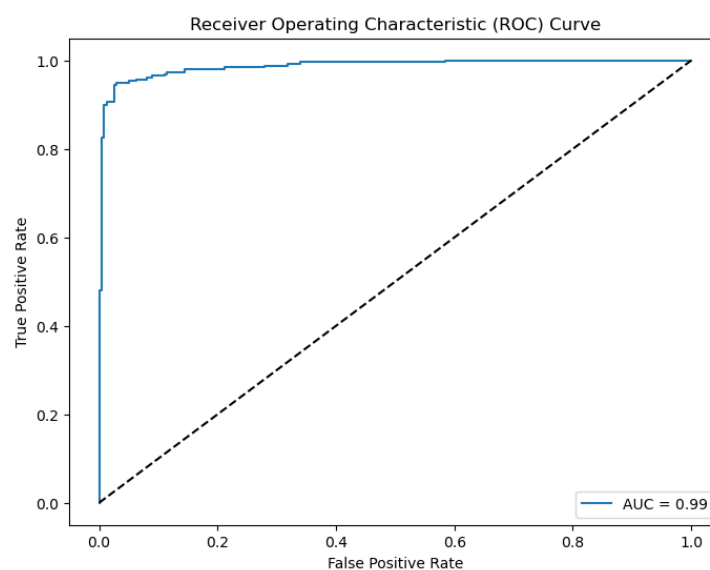


Figure 5.1: AUC-ROC Curve of the CatBoost

Comparing the results of both the tables, it is undeniably proved that balancing the dataset using SMOTE enhances the performance of the machine learning models. This boost in the accuracy, recall, precision, AUC-ROC and so on, explains the major importance of addressing the imbalancing issue before implementing the models for robust predictions.

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.83	0.83	0.83	0.83	0.91
Decision Tree	0.81	0.81	0.81	0.81	0.81
Random Forest	0.93	0.93	0.93	0.93	0.98
SVM	0.91	0.91	0.91	0.91	0.97
kNN	0.84	0.88	0.84	0.84	0.92
AdaBoost	0.84	0.84	0.84	0.84	0.94
CatBoost	0.95	0.95	0.95	0.95	0.99
Gradient Boosting	0.88	0.88	0.88	0.88	0.96
XGBoost	0.88	0.88	0.88	0.88	0.96

Table 5.2: Performance Metrics on Balanced Dataset using SMOTE

Since the accuracy of the Decision Tree Model was below 80%, we applied the hyperparameter tuning - grid search to improve it on both the balanced and imbalanced dataset and the results are tabulated 5.3. This evidently showcases the importance of hyperparameter tuning in fetching optimal results. Comparatively, the accuracy has been amplified to 83% and 88% respectively. Substantial improvements has been noticed after applying the grid-search technique to identify the optimal hyperparameter. Additionally, the other metrics like the precision, recall and the F1-score has been enhanced significantly.

Datasets	Accuracy	Precision	Recall	F1-Score
Imbalanced	0.83	0.68	0.64	0.66
Balanced	0.88	0.88	0.88	0.88

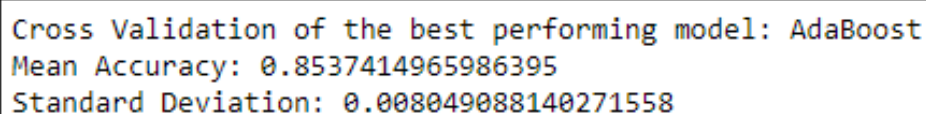
Table 5.3: Performance metrics of Decision Tree after Grid Search

5.2 Interpreting the findings

The findings from the machine learning models provides more insights into their performance and efficacy in employee attrition. When models trained on an imbalanced dataset were evaluated, it was clear that their performance were poorer from that of models built on a balanced dataset using SMOTE, which stresses the importance of balancing the dataset before implementing the machine learning models. Consistently the models, CatBoost and Random Forest demonstrated exceptional performance across both balanced and imbalanced datasets. CatBoost showcased its superior AUC-ROC values, underscoring the ability of distinguishing the TP and FP effectively. These insights contribute the Human Resources department in making informed decision for addressing the employee attrition.

We also performed k-fold cross validation for the best performing models to confirm the correctness of them in both the imbalanced and balanced dataset i.e., AdaBoost and CatBoost. In order to reduce processing costs and improve evaluation accuracy, we decided to use the k-value as five folds.

The cross-validation results of AdaBoost model from the figure 5.2 reveals some important insights about its performance with different folds. With a mean accuracy of 0.853, the model performed quite well, as seen by its attempt to correctly predict 85.37% of the test subset. The standard deviation (SD) value 0.0080 also supports the correctness of the model, by giving the information on variability in accuracy. The lower value of SD makes the model's performance more consistent and reliable, which in our case is plausible and also means that the model maintained a similar level of accuracies for different folds. This assures us that the model will perform efficient on unseen employee profiles to predict the attrition rate.



```
Cross Validation of the best performing model: AdaBoost
Mean Accuracy: 0.8537414965986395
Standard Deviation: 0.008049088140271558
```

Figure 5.2: Evaluation of our AdaBoost model using k-fold

The figure 5.3 gives us the outcome of the k-fold cross validation for the CatBoost model treated with a balanced dataset. The individual accuracy of the folds is also

provided along with its mean accuracy of 0.9082, indicating that, on average the model managed to correctly predict the employee attrition status 90.82% times, which is extremely efficient. This metric serves a robust estimate of employee attrition on unseen data. Furthermore with the value of standard deviation being way smaller i.e., 0.0140, demonstrates that the stability of prediction. On the whole, this evaluation confirms the efficacy of the best performed model - CatBoost.

```
Cross-Validation Results of our best performing model: CatBoost
Individual Fold Scores: [0.90379747 0.9164557 0.91370558 0.88324873 0.92385787]
Mean Accuracy: 0.9082130694596158
Standard Deviation of Accuracy: 0.014039829490045954
```

Figure 5.3: Evaluation of our CatBoost model using k-fold

5.3 Limitation & Further Research

While the conducted research was a pure success and could efficiently predict the attrition rates using machine learning models, there are some notable points that has to be considered. The main limitation was our reliance on one specific small dataset with limited features, that may restrict the models' capacity to be broadly applied. Moreover the imbalance nature of the dataset was a challenge, despite oversampling it using SMOTE doesn't guarantee us on the prediction of employee attrition on minority class without bias.

As the main concern of attrition being the satisfaction of the employees, working with the refined data such as sentiment analysis, employee feedback which would boost the capability of our prediction models. In terms of further research, more deeper interpretability of the model and indepth correlation features is crucial for real-world problems. For the organization to implement appropriate retention strategies, the prediction have to precisely identify the factors that influence it. Additionally, building neural networks would potentially uncover different insights and boosts the prediction accuracy.

Regardless of these limitations, this research aimed to lay a foundation on understanding the complexities of predicting the rate of employee attrition. Followed by this research, the further studies should build upon these findings and incorporate more on diverse datasets and advanced modeling techniques to increase the applicability of

more predictive models in HR.

5.4 Strategies for Employee Retention

From the insights gained from our research and analysis, several approaches can be implemented in an organization to reduce the possibilities of employee attrition. Concentrating on the most contributing key factors such as dissatisfaction with the environment, compensation for the employee and addressing them would be greatly appreciated. Thus, developing the overall experience of the employees, achieving career growth opportunities for the employees within the organization. Implementing employee engagement activities like training programs, wellness programs would greatly motivate the employees to stay back in the organization. Fetching consistent feedback on how the employees doing and taking necessary actions on what could be improved, would actually boost the employee's loyalty over the company, which can also in-turn provide the company some valuable factor that might make the employees to leave. This is achievable when the HR division and data scientists collaborate to develop and iteratively improve different retention tactics that not only lower staff turnover but also raise employee satisfaction levels and encourage long-term commitment.

Conclusions

In conclusion, this research aimed to delve into the complexity of predicting the attrition rate of employees in an organization using advanced machine learning techniques such as Logistic Regression, Decision Tree, Random Forest, SVM, k-NN, AdaBoost, CatBoost, Gradient Boosting and XGBoost and also managed to identify the factors that influences employee turnover. Starting with the data exploration and pre-processing, handling the imbalanced dataset using an advanced over sampling method called SMOTE.

Employee attrition could be influenced by various factors, mainly including monthly compensation, work-life balance, and inadequate career advancement, all of which have been identified through the use of upgraded prediction models. With the purpose of keeping personnel and preserving a sustainable workplace, the HR department may adopt better retention tactics, thanks to these insights. Also verifying the outcomes by validating our machine learning models has also been carried out using various evaluation methods. This also includes confusion matrix, AUC-ROC curve, cross-validation, etc. Models like CatBoost showed at most accuracy of 95% and robustness across various performance metrics, showcasing the importance of algorithmic selection in suitable contexts. K-fold cross validation provided stronger support for the stability of our models with mean accuracy of approximately 90% and the validity of our results.

Most importantly, we displayed the significance of balancing the dataset to acquire more robust and precise models. While our research contributed a lot to provide insights on employee attrition, its necessary to address the limitations as well. Given the reliance on historical data for the forecast, unforeseen events may impact it differently, and

managing dynamic personnel could pose various challenges as well. On the brighter side, the insights from this research would be of a greater advantage to HR practices to implement strategies pro-actively to provide better satisfaction to workers and encouraging them to continue their employment. Thus, attempting to close the distance between the predictive analytics and HR practices by promoting data-driven strategies to cultivate a workforce that thrives in a supportive and full-filling environment.

Bibliography

- [1] S. S. Alduayj and K. Rajpoot, *Predicting Employee Attrition using Machine Learning*, in *2018 International Conference on Innovations in Information Technology (IIT)*, Al Ain, United Arab Emirates, 2018, pp. 93-98, doi: 10.1109/INNOVATIONS.2018.8605976.
- [2] A. Qutub, A. Al-Mehmadi, M. Al-Hssan, R. Aljohani, and H.S. Alghamdi, *Prediction of Employee Attrition using Machine Learning and Ensemble Methods*, *Int. J. Mach. Learn. Comput.*, vol. 11, no. 2, pp. 110-114, 2021.
- [3] Francesca Fallucchi, Marco Coladangelo, Romeo Giuliano, and Ernesto William De Luca, *Predicting Employee Attrition Using Machine Learning Techniques*, *Computers*, vol. 9, no. 4, article number 86, 2020. [Online]. Available: <https://www.mdpi.com/2073-431X/9/4/86> ISSN: 2073-431X. DOI: 10.3390/computers9040086
- [4] M. Subhashini and R. Gopinath, *Employee Attrition Prediction in Industry Using Machine Learning Techniques*, *International Journal of Advanced Research in Engineering and Technology*, vol. 11, no. 12, pp. 3329-3341, 2020.
- [5] Dereddy, Akhil Reddy. *Predictive Analytics for Employee Attrition & Performance*. MIS, 2022. URI: <https://dr.lib.iastate.edu/handle/20.500.12876/105268>. Date: 2022-05. Major Professor: Townsend, Anthony. Committee Member: Townsend, Anthony
- [6] Darapaneni, N., et al. "A Detailed Analysis of AI Models for Predicting Employee Attrition Risk." In *2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC)*, Hyderabad, India, 2022, pp. 243-246. DOI: 10.1109/R10-HTC54060.2022.9929893.

- [7] Raja Rajeswari, G., Murugesan, R., Aruna, R., Jayakrishnan, B., & Nilavathy, K. "Predicting Employee Attrition through Machine Learning." In *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, 2022, pp. 1370-1379. DOI: 10.1109/ICOSEC54921.2022.9952020.
- [8] Jain, P.K., Jain, M., & Pamula, R. "Explaining and predicting employees attrition: a machine learning approach." *SN Appl. Sci.* 2, 757 (2020). Published: 25 March 2020. DOI: <https://doi.org/10.1007/s42452-020-2519-4>.
- [9] Maharana, M., Rani, R., Dev, A., & Sharma, A. "Automated Early Prediction of Employee Attrition in Industry Using Machine Learning Algorithms." In *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India, 2022, pp. 1-6. DOI: 10.1109/ICRITO56286.2022.9965017.
- [10] George, S., Lakshmi, K. A., & Thomas, K. T. "Predicting Employee Attrition Using Machine Learning Algorithms." In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, Greater Noida, India, 2022, pp. 700-705. DOI: 10.1109/ICAC3N56670.2022.10074131.
- [11] Meraliyev, B., Karabayeva, A., Altynbekova, T., & Nematov, Y. "Attrition Rate Measuring In Human Resource Analytics Using Machine Learning." In *2023 17th International Conference on Electronics Computer and Computation (ICECCO)*, Kaskelen, Kazakhstan, 2023, pp. 1-6. DOI: 10.1109/ICECCO58239.2023.10146602.
- [12] Hebbar, A. Rohit, Patil, S. H., Rajeshwari, S. B., & Saquaf, S. S. M. "Comparison of Machine Learning Techniques to Predict the Attrition Rate of the Employees." In *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, Bangalore, India, 2018, pp. 934-938. DOI: 10.1109/RTEICT42901.2018.9012243.
- [13] Alao, D. A. B. A., & Adeyemo, A. B. "Analyzing Employee Attrition using Decision Tree Algorithms." *Computing Information Systems Development Informatics and Allied Research Journal*, vol. 4, no. 1, pp. 17-28, 2013.
- [14] Priya, V. Krishna, & Harasudha, H.H. "A Study on Employee Attrition with Reference to Lanson Toyota, Chennai." *Man in India*, 97, 115-124, 2017.

- [15] Gurler, Kerem, Pak, Burcu Kuleli, & Gungor, Vehbi Cagri. "Deep Learning Based Employee Attrition Prediction." In *Artificial Intelligence Applications and Innovations*, 2023. Editor: Ilias Maglogiannis, Lazaros Iliadis, John MacIntyre, Manuel Dominguez. Publisher: Springer Nature Switzerland. Pages: 57-68. ISBN: 978-3-031-34111-3.
- [16] Al-Darraj, Salah, Honi, Dhafer G., Fallucchi, Francesca, Abdulsada, Ayad I., Giuliano, Romeo, & Abdulmalik, Husam A. "Employee Attrition Prediction Using Deep Neural Networks." *Computers*, 10(11), 141, 2021. DOI: <https://doi.org/10.3390/computers10110141>.
- [17] H. He and E. A. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: 10.1109/TKDE.2008.239.
- [18] Haneen Arafat Abu Alfeilat, Ahmad B.A. Hassanat, Omar Lasassmeh, Ahmad S. Tarawneh, Mahmoud Bashir Alhasanat, Hamzeh S. Eyal Salman, and V.B. Surya Prasath, *Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review*, *Big Data*, vol. 7, no. 4, pp. 221-248, Dec. 2019, ISSN: 2167-647X, doi: 10.1089/big.2018.0175, url: <http://dx.doi.org/10.1089/big.2018.0175>.
- [19] Google Developers. "ROC and AUC: Understanding Receiver Operating Characteristic and Area Under the Curve." URL: <https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc>. Accessed: 2023.
- [20] Analytics Vidhya. "Impact of Machine Learning on HR: Predicting Attrition and More." URL: <https://www.analyticsvidhya.com/blog/2023/03/impact-of-machine-learning-on-hr/>. Accessed: 2023.
- [21] Wikipedia contributors. *Correlation* — *Wikipedia*, The Free Encyclopedia, 2023. [Online; accessed 10-November-2023]. Available at: <https://en.wikipedia.org/w/index.php?title=Correlation&oldid=1184280911>.
- [22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cour-

- napeau, M. Brucher, M. Perrot, E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [23] Wikipedia contributors. *Support vector machine* — *Wikipedia, The Free Encyclopedia*. 2023. https://en.wikipedia.org/w/index.php?title=Support_vector_machine&oldid=1183475870. [Online; accessed 16-November-2023].
- [24] Analytics Vidhya. *Gradient Boosting Algorithm: A Complete Guide for Beginners*. September 2021. <https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/>.
- [25] Yuanchao Wang, Z. Pan, J. Zheng, L. Qian, and Li Mingtao. *A hybrid ensemble method for pulsar candidate classification*. *Astrophysics and Space Science*, volume 364, August 2019. <https://doi.org/10.1007/s10509-019-3602-4>.
- [26] Shubham Malik, Rohan Harode, Akash Singh Kunwar. *XGBoost: A Deep Dive Into Boosting*. March 23, 2020. <https://dzone.com/articles/xgboost-a-deep-dive-into-boosting>.
- [27] Deepak Kumar Sharma, Mayukh Chatterjee, Gurmehak Kaur, Suchitra Vavilala. 3 - *Deep learning applications for disease diagnosis*. In: Deepak Gupta, Utku Kose, Ashish Khanna, Valentina Emilia Balas (Eds.), *Deep Learning for Medical Applications with Unique Data*, Academic Press, 2022, pages 31-51. ISBN: 978-0-12-824145-5. DOI: <https://doi.org/10.1016/B978-0-12-824145-5.00005-8>. URL: <https://www.sciencedirect.com/science/article/pii/B9780128241455000058>.
- [28] Yang, H.; Liu, Z.; Li, Y.; Wei, H.; Huang, N. CatBoost-Bayesian Hybrid Model Adaptively Coupled with Modified Theoretical Equations for Estimating the Undrained Shear Strength of Clay. *Appl. Sci.* **2023**, *13*, 5418. <https://doi.org/10.3390/app13095418>.