Deciphering the Augmentation of Classification Models in Predicting Employee Attrition

MSc Research Project
Data Analytics

Aditya Raj
Student ID: x20143311

School of Computing
National College of Ireland

Supervisor: Dr. Martin Alain

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Aditya Raj
Student ID:	x20143311
Programme:	Data Analytics
Year:	2021
Module:	MSc Research Project
Supervisor:	Dr. Martin Alain
Submission Due Date:	31/01/2022
Project Title:	Deciphering the Augmentation of Classification Models in Predicting Employee Attrition
Word Count:	6784
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Aditya Raj
Date:	31st January 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Deciphering the Augmentation of Classification Models in Predicting Employee Attrition

Aditya Raj
x20143311

Abstract

Workforce administration is at the cusp of renaissance due to the surge in attrition rates globally. Every organization is driven by its employees and their ability to meet their objectives and it is strategically imperative for employers to retain talented and highly skilled professionals. It is critical to monitor employee demographics and contrive a plan to identify potential conflicts in organizational setting. The juxtaposition of attrition and retention align on the analysis of common factors. With the advent of machine learning in the field of HR Analytics, employee attrition prediction has been fulfilled through implementation of classification techniques. The research aims to extend the use of Random Forest classifier and Light Gradient Boosting Machine (LightGBM) classifier to predict employee attrition on an class imbalanced artificially simulated HR Employee dataset using oversampling methods such as SMOTE and Random Oversampling. The classification methods are integrated with Grid Search Cross Validation for Best Hyperparameters, Recursive Feature Elimination (RFE, advanced method for feature selection) and Manual Hyperparameter tuning. The models are evaluated based on F1 Score, Accuracy and ROC AUC score. It is concluded that Random Forest classifier with manual tuning of Hyperparameter performed better than the Grid Search Cross Validation and Recursive Feature Elimination approaches with a F1 score of 62.18% , Accuracy of 84.69% and ROC AUC score of 84.28% whereas manually hyperparameter tuned LightGBM classifier exhibited incredible performance with a F1 score of 65.91%, Accuracy of 89.8% and ROC AUC score of 78.42%.

Keywords - ***Class Imbalance, Classification, Employee Attrition, Hyperparameter Tuning, Recursive Feature Elimination, SMOTE.***

1 Introduction

Employee Attrition is defined as reduction of workforce due to unforeseen or anticipated events such as death, resignation, retirement, termination etc. (Nappinnai and Premavathy; 2013) Attrition is coherently related to turnover that formulates the Employee Churn Rate.[1] Apart from losing skilled and trained employees, employee attrition entails loss of resources, time and costs invested by the organization in recruitment and training of attritted employees.

Several studies have been conducted in the past to derive insights and patterns from employee data to integrate into their existing structure of work policies. The recent

[1] `https://www.workforcehub.com/glossary/churn-rate/`

transition observed in the upsurge of attrition rates proves the inevitability of this phenomenon. In the present scenario, the employees are too flexible, volatile and prone to erratic exit from the organization in the interest of securing a job with better compensation, job satisfaction and work life balance as suggested through a research conducted by (Jain and Nayyar; 2018).

The way to minimize the attrition rate is to regulate the workforce and devise strategies to enhance the employee management and inculcate retention plans . The advancement in the field of technology has enabled the Human Resource Management (HRM) to monitor employee behaviour and their performance facilitated through specialized applications. They also tend to perform Root Cause Analysis (RCA) on attrition cases to identify the significance of possible elements that lead to attrition and bring in amendment of such elements to ensure employee retention. (Wang; 2010) The advent of machine learning has enabled us to perform the same operations using a data mining approach. A lot of researches have relied on implementing classification algorithms for attrition prediction on the IBM HR Analytics dataset[2], a fictional employee data designed by data scientists at IBM. This research intends to utilize the same dataset as there is a dearth of open source employee datasets due to privacy and data protection laws. The gap observed in the previous researches has been on the evaluation of the models based on accuracy which is questionable due to the presence of class imbalance in the dataset. Also, most studies that have been conducted in the past have relied on comparative analysis of performance of the typical classification methods such as Logistic Regression, Random Forest Classifier, Support Vector Classifier etc. (Rohit Hebbar et al.; 2018) The objective of this research is to build a efficient prediction model with better prediction capabilities in a binary classification setting by training the model on the data where class imbalance has been handled using synthetic techniques such as SMOTE and Random Oversampling, use Grid Search Cross validation for Best Hyperparameters, Recursive Feature Elimination and evaluate the trained models on the basis F1 score, Accuracy and ROC AUC score. The selection of classification methods for implementation has been informed by literature reviewed. With the relevant approach stated above, this research aims to investigate and answer the following research question:

- **”How can feature engineering and hyperparameter tuning be used to enhance the classification model capability for employee attrition prediction?”**

The report is structured further into 6 sections. A critical analysis of past researches is summarised that highlights the current viewpoint on the chosen topic of interest under the Related Work in Section 2. A brief summary on the steps involved in implementing the research methodology is articled that includes the data description, data understanding and rational for selecting the specific classification methods under Methodology in Section 3. The underlying architecture for this research is briefly elucidated under the Design Specification in Section 4. The subsequent section explains the multiple fragments of Implementation related to Statistical Analysis on categorical features, Feature Engineering, Feature Encoding and Scaling, Class Balancing techniques and Model specifications under Implementation in Section 5. The outputs of the implementation are thoroughly assessed for conducted experiments/case studies and implications are briefly annotated under Evaluation in Section 6. The final verdict of the research is stated along

[2] https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset

with the any possible future work that may enhance the scope of the research is explained under the Conclusion and Future Work in Section 7.

2 Related Work

This section elucidates a summarized version of researches that have been conducted in the past over the chosen topic for research. The advantage of reviewing the existing literature helps in making an informed decision in selecting the research methodology. Previous researches and their corresponding outputs act as a benchmark for the proposed solution to the research question.

2.1 Factors affecting Employee Attrition

The impact of Employee Attrition if unpredicted can be catastrophic. It not only cripples the organization financially, but also incapacitates the motivation and morale of the existing employees. (Bhartiya et al.; 2019) through their research highlighted the key drivers of attrition that were observed while analyzing the data. They claimed that attrition rate is comparatively higher in employees with a technical degree than employees with medical degree, In a similar study, (Joseph et al.; 2021) stated that mental well being and psychological factors such as job dissatisfaction, high workload, depression and anxiety lead to attrition.

Employee attrition can be voluntary or involuntary. When staff quit their company for personal reasons, this is referred to as voluntary churn, however, if the organization is responsible to terminate an employee, it is referred to as involuntary type of attrition (Frye et al.; 2018). Some industries such as call centres comparatively have higher attrition rate, but in broad sense, this impacts all companies (Mobley; 1977). According to (ibid.), Job satisfaction is strongly correlated with employee turnover.

Cotton et al. (1986) identified age, tenure, salary, career progression, and prospective preferences as the most important predictors of employee attrition or voluntary retirement. According to Ozoliņa-Ozola (2015), employee turnover and attribution intentions are influenced by employee attitudes toward their jobs. Employee turnover and attrition wreaks havoc on the remaining staff. Attrition and turnover have shifted the focus of higher leadership in nearly every organisation. It suggests that turnover is one of the most expensive and difficult workforce issues that organisations face. To assess people management skills, (Hoffman and Tadelis; 2017) used employee survey responses about their manager. According to the analysis, executives with good people skills should be commended by the firm, which which might result in decreasing attrition of employees. Additionally, consistent responses on employees are required to maintain them involved and committed to their work. The study also found that professions requiring limited skills are much easier to perform than professions that require qualified workers.

In every organization, there is attrition due to employee resignations or retirement. This issue can have serious consequences for an organization's viability if it is not handled properly and employees leave unexpectedly (Alhashmi; 2019). This study lays the foundation of conducting this research in order to extricate the factors that cause attrition. Based on the factors analysed, the organization can strategize and adopt measures to curtail attrition rate and manage workforce efficiently.

2.2 Significance of Feature Engineering and Hyperparameter Tuning

The limitation of machine learning model's capability to interpret only numerical values raises the necessity for feature engineering, encoding, scaling, feature reduction and other techniques. Each attribute (categorical or numerical) in a dataset may or may not add significance in a classification model, but feature engineering or encoding such attributes can add value to their computational performance. Bhartiya et al. (2019) conducted a study to predict employee attrition using a legion of classification models and analyze the results with and without implementation of feature reduction. It was concluded through their study and produced results that the holistic feature transformation and reduction can significantly improve the performance of models. Yadav et al. (2018) demonstrated the use of feature engineering and feature reduction simultaneously by reducing the number of features from 12 to 10 and using encoding methods to create dummy variables for categorical attributes. The model was implemented using a Recursive Feature Elimination method that aimed at achieving highest cross validation score using minimal number of parameters. Hyperparameter Tuning can be computationally expensive depending on the number of hyperparameters required to be tuned. However, it can manifest significant improvement in the performance. The research conducted by (Alawad et al.; 2018) aimed at tuning hyperparameters for Decision Tree Classifier to gauge efficient and optimized strategy for fraud detection. Another research suggested the use of Grid Search cross validation that returns the best parameters from list of potential values for k-fold cross validation which achieved the best score in cancer prediction (Shekar and Dagnew; 2019).

2.3 Classification Models for Binary Classification

Yedida and Vahi (n.d.) implemented an approach to predict employee attrition using supervised machine learning classification models such as k-Nearest Neighbour (KNN), Naïve Bayes, Multi- Layer Perceptron (MLP) and Logistic regression. The KNN model was found to be the best performing model with 94% accuracy and F1-score of 0.88. However, for KNN, it is not always apparent which sort of distance to utilize, or which characteristic would provide the greatest results. Also, the distance between each training example must be determined and the calculation cost is rather significant (Taunk et al.; 2019). Bhartiya et al. (2019) also performed the comparative analysis of classification models using Support Vector Machine (SVM), Decision Tree, KNN, Random Forest (RF) and Naïve Bayes classification models, where the RF classifier is identified as the best performing model with feature engineering. Jain and Nayyar (2018) proposed a novel approach to predict employee churn using Extreme Gradient Boosting (XGBoost). This model performed extremely well and robustness and scalability of this model has been demonstrated. However, the performance of Light Gradient Boosting Model (LGBM) surpassed the other classification models such as RF, SVM and KNN in terms of efficiency and computational time (Łoś, Mendes, Cordeiro, Grosso, Costa, Benevides and Caetano; 2021). LGBM is considered optimal model in terms of computational time and results as compared to other boosting algorithms such as XGBoost and CatBoost (Daoud; 2019). The literature reviewed above implies that Random Forest Classifier is the state of the art machine learning model that outperforms other models in terms of accuracy, precision, recall and training time. However, there has been a lack of research on LightGBM Classifier for binary classification problems, but the testimonials endorse the idea of possible success. Based on the papers reviewed, Random Forest classifier and LightGBM classifier can be tested against binary classification problem of employee prediction.

Based on the exhaustive list of researches reviewed, it can be stated that the preliminary analysis on the drivers of attrition helps in concocting strategies for retention of high attrition risks within the organization. It can be deduced that data preprocessing techniques such as Feature Engineering, Feature Reduction, Feature Encoding and Scaling aids in improving the computational overhead of classification models. Automated techniques and manual attempts at tuning hyperparameters can enhance the performance significantly. The takeaway from the literature reviewed from classification model standpoint indicates that Random Forest and LightGBM classifiers are superlative choices for binary classification. The combination of the same has been proposed as the methodology for this research.

3 Methodology

Researches pertaining to Data Mining approach are conducted in phased manner with milestones and objectives defined at each level. In this research, the study conducted, followed a two-tier Knowledge Discovery in Database (KDD) approach to perform Data Mining operations. Each tier comprises of multiple phases of the data mining process. Tier - I consists of Data Selection and Data Preprocessing whereas Tier - II comprises of Exploratory Data Analysis, Data Transformaton, Modelling and Evaluation.

1. **Tier - I** : This layer corresponds to the data selection and data preprocessing stage of the research. The preliminary interpretation and basic validations are performed on the data to check its accuracy, completeness and consistency.

2. **Tier - II** : This layer corresponds to the processes that are performed from business/domain perspective such Exploratory Data Analysis, Data Transformation, Modelling and Evaluation.

Each phase accounts for specific milestones that are set in accordance with the primary objective of the research. The phases are structured as illustrated in the Figure 1.

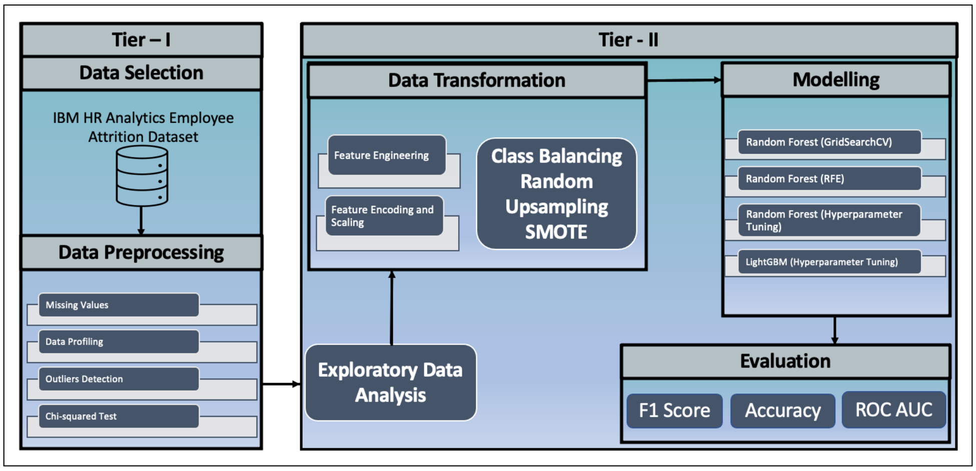

Tier – I
Data Selection
IBM HR Analytics Employee Attrition Dataset
Data Preprocessing
Missing Values
Data Profiling
Outliers Detection
Chi-squared Test
Tier - II
Data Transformation
Feature Engineering
Feature Encoding and Scaling
Class Balancing Random Upsampling SMOTE
Modelling
Random Forest (GridSearchCV)
Random Forest (RFE)
Random Forest (Hyperparameter Tuning)
LightGBM (Hyperparameter Tuning)
Exploratory Data Analysis
Evaluation
F1 Score
Accuracy
ROC AUC

Figure 1: KDD - 2-Tier Process Flow

3.1 Data Selection

The dataset used in this research is an artificially simulated employee data known as IBM-HR Employee Attrition data fabricated by a group of data scientists at IBM. There has been a lack of publicly available employee datasets for purpose of research due to ethical privacy concerns. Therefore, this dataset was chosen for conducting this research. This dataset comprises of 35 variables and 1470 observations and it contains few critical attributes such as Satisfaction Ratings, Overtime Status, Distance from home, Job Involvement, Years since Last Promotion etc. The table 1 below provides a description of the variables within the dataset.

Table 1: Data Description

S.No.	Variable	Type
1.	Age	Numerical
2.	Attrition	Object
3.	Business Travel	Categorical
4.	Daily Rate	Numerical
5.	Department	Categorical
6.	Distance From Home	Numerical
7.	Education	Numerical
8.	Education Field	Categorical
9.	Employee Count	Numerical
10.	Employee Number	Numerical
11.	Environment Satisfaction	Numerical
12.	Gender	Categorical
13.	Hourly Rate	Numerical
14.	Job Involvement	Numerical
15.	Job Level	Numerical
16.	Job Role	Categorical
17.	Job Satisfaction	Numerical
18.	Marital Status	Categorical
19.	Monthly Income	Numerical
20.	Monthly Rate	Numerical
21.	Number of Companies Worked	Numerical
22.	Over 18 Years	Categorical
23.	Overtime	Categorical
24.	Percent Salary Hike	Numerical
25.	Performance Rating	Numerical
26.	Relationship Satisfaction	Numerical
27.	Standard Hours	Numerical
28.	Stock Options Level	Numerical
29.	Total Working Years	Numerical
30.	Training Times Last Year	Numerical
31.	Work Life Balance	Numerical
32.	Years At Company	Numerical
33.	Years in Current Role	Numerical
34.	Years Since Last Promotion	Numerical
35.	Years With Current Manager	Numerical

3.2 Data Preprocessing

Data preprocessing is defined as the method where the data undergoes multiple validations to ensure its accuracy, completeness and consistency. The second phase in the data mining process when the data is gathered was to preprocess the data. The research also conducted Exploratory Data Analysis to gain insights about the factors that cause attrition. The steps of Data Preprocessing are further categorized as follows:

3.2.1 Missing Values

The dataset is checked for any missing values to check the consistency of the data. There are multiple ways to handle missing values such as deletion of missing records or incorporating various imputation methods.

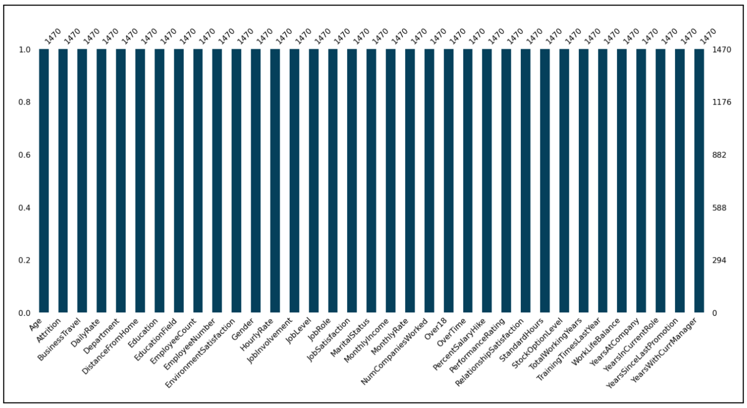


Figure 2: Missing Values Validation

However, it can be inferred from the figure 2 that there are no missing values in the dataset.

3.2.2 Data Profiling

The data profiling process evaluates the variables within the data to produce descriptive statistics, data types, null validations, missing values and correlation coefficients of numerical variables. The profiling output provides an insight on pattern observed within the variables.

3.2.3 Outliers Detection

The variables are also checked for any outliers which can result in biased computation. If any outliers are detected, they should be effectively treated to eliminate the variability in data. The two common approaches to treat outliers are deleting observations with outliers or scaling the values within a specified range.

3.2.4 Chi-squared Test for Statistical Significance

Chi-squared test is hypothesis testing process to check if two categorical variables are statistically significant. The null hypothesis states that there is no relationship between two variables whereas the alternate hypothesis states the there is presence of relation between two variables. A p-value (probability value) is measure of statistical significance of variables. [3]

- If the p-value is less than critical value (0.05), the variables are statistically significant and we reject the null hypothesis.

[3] What a p-value tells you about statistical significance `https://www.simplypsychology.org/p-value.html`

- If the p-value is greater than the critical value (0.05), the variables are not statistically significant and we fail to reject the null hypothesis

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the third phase in the conducted research as per the KDD process flow. EDA is a term referred to as investigation of the data to comprehend the latent patterns, anomalies and trend observed within the data. Several visualizations such doughnut chart, box-plots, pie plots, bar and column charts were utilized in the overall research to scrutinize the factors that cause attrition in organization and draw inferences to propose retention strategies.

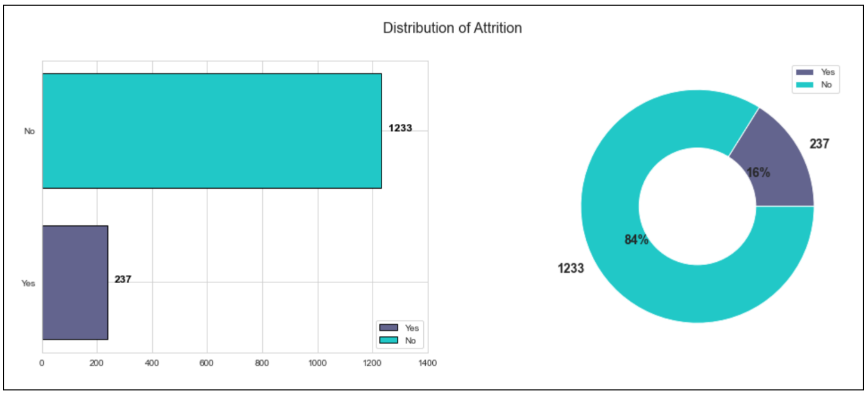


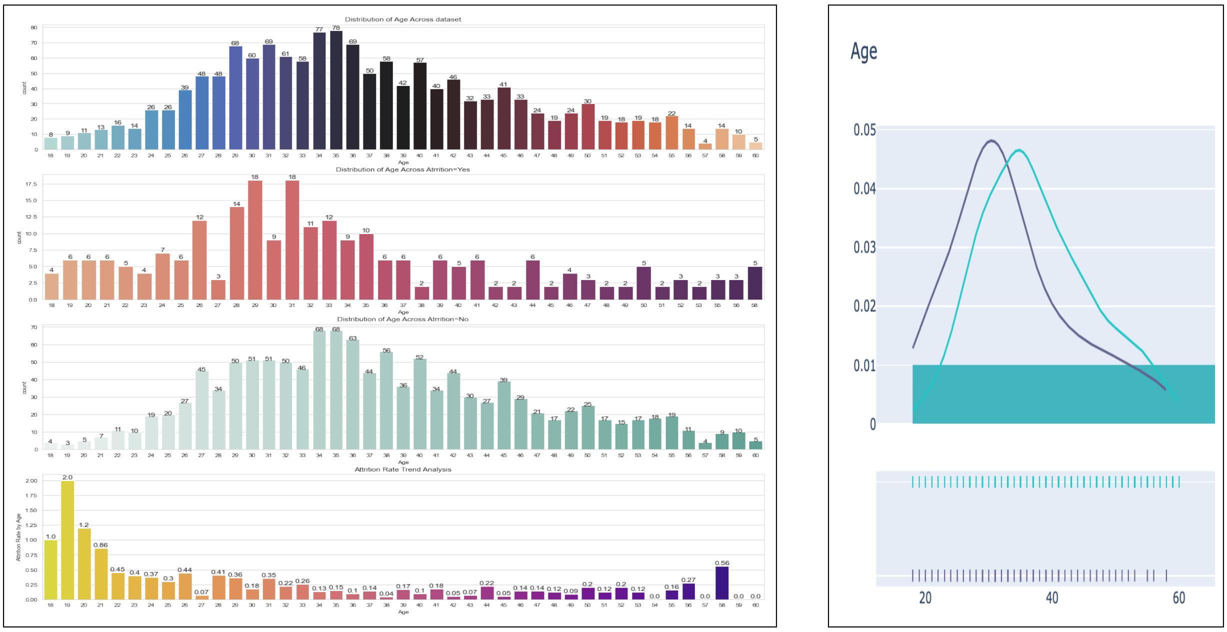
Figure 3: Distribution of Attrition Classes

It can be inferred from the above Pie and Horizontal Bar chart, 3 that there is a Class Imbalance in Attrition which is our target variable for prediction. In a typical data mining approach, class imbalance should be handled in pragmatic manner. The technique to handle class imbalance problems is through sampling methods such as SMOTE and Random Sampling. The class balancing methodology is elaborated in section 3.4. Attrition Rate is defined as the percentage of separations over average number of employees for a specific time frame.

$$AttritionRate = [\frac{Number\ of\ attrited\ employees}{Average\ number\ of\ employees} * 100]$$

The distribution of attrition samples is 237 (Yes) and 1233 (No), the attrition rate evaluated based on these figures is 16.12%.

From the below distribution and KDE plot in the figure 4 it can be inferred that attrition rate is higher in younger age segment of employees (approx 18-26 years) and significant increase during the retirement age (approx 55-58 years). It can be inferred that younger segment of employees tend to leave the organization due to possible commitments in retrospect of higher compensation, higher job satisfaction, healthy work life balance, higher education, better work opportunities, etc.



(a) Age vs Attrition/Non-Attrition Distribution (b) Attrition Rate vs Age KDE Plot

Figure 4: Attrition Rate Analysis

The surge in attrition rate in higher age bracket is possible due to employees opting for voluntary retirement. It is important to investigate the contribution of such employees in order to identify the reasons of attrition.

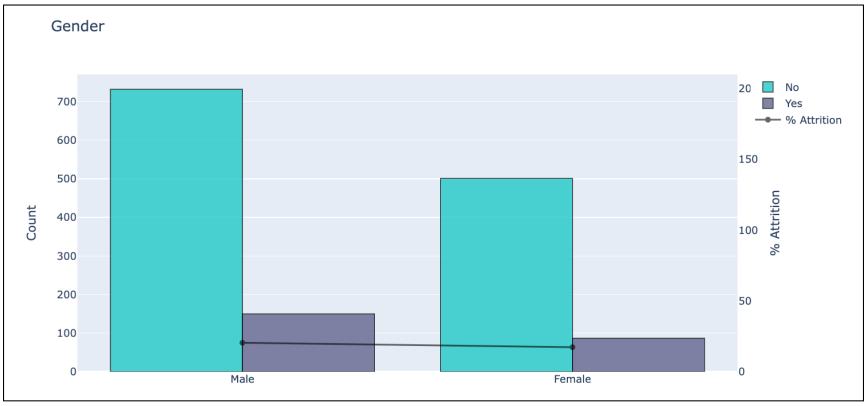


Figure 5: Attrition vs Gender

It can be inferred from the comparison bar plot 5 above that the attrition rate is relatively higher in male as compared to female. It can be assumed that females are found to be more affined to the organization in terms of loyalty.

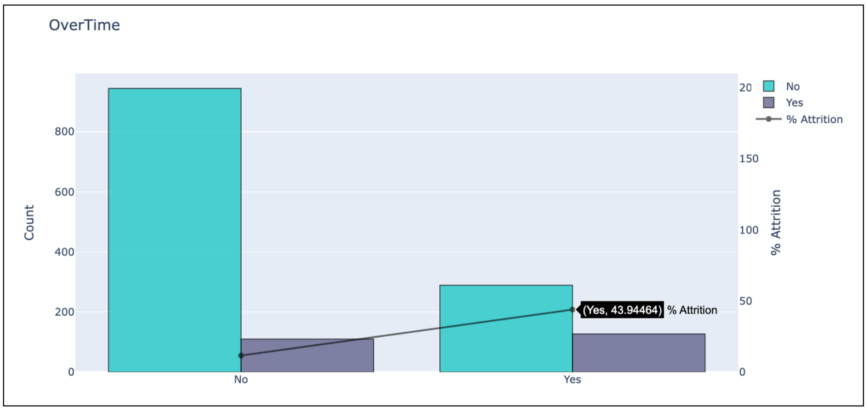


Figure 6: Attrition vs Overtime

From the above multiple bar graph, 6 the percentage of separation is higher for employees who work Overtime compared to the ones who don't. Overtime adds encumbrance on employees and organizations should plan to attenuate requirement of overtime marginally or maybe compensate 1.5x/2x times their hourly rate for overtime hours. The summary of other inferences drawn from EDA with respect to employee behaviour towards attrition rate are stated below:

- A higher attrition rate is observed in employees with a low Daily-Rate of 300-600.
- Employees with distance from home greater than 9 illustrates higher attrition rate.
- The employee segment working at an hourly rate of (45-60) tends to see a significantly higher attrition rate.
- The employees with the low monthly income (0-2.5k) depicts a higher attrition rate. Compensation is one of the intrinsic factors that affects the attrition rate.
- Line curve of Attrition and Non Attrition seems to overlap for Monthly Rate that implies that it is of little or negligible significance in driving Attrition.
- The retention and attrition rate are higher for lower number of companies worked (0-2). The trend along the higher number of companies worked for implies that such employees have a tendency to switch organisations and are potential attrition risks.
- Attrition rate is observed as moderately high with lower percentage of salary hikes. Organizations should critically assess the contribution of employees and ensure promotion and salary hikes to deserving employees.
- There is a spike in attrition rate is observed for employees with minimal or no experience.
- It can be inferred that employees enrolled in training (2-3) times last year results in higher rates of attrition.
- It is observed that there are multiple surge in attrition rates observed at multiple segments of employee's associated years with the company at (0-2, 22-25, 30-35).

- It is observed that employees at the initial years of working on a specific role (0-2) and intermediate years (14-15) experiences a rise in attrition rate. The possible reason for the former scenario is that employees are indecisive about the chosen job role and the latter might be a result of monotonous roles and responsibilities.

- It can be inferred that attrition rates tend to rise after (6-7) years since their last promotion. It explains that employees are enervated working on the role in the hope of promotion.

- The attrition rate with the respect to years of working with a manager observes spikes for 0 and 14 is erratic in nature.

3.4 Data Transformation

Data Transformation is the fourth phase in data mining process which performs transformations in data that can add value to the proposed research. In the research, Feature Engineering, Feature Encoding and Feature Scaling were performed.

3.4.1 Feature Engineering

It is the process where new features are derived from one or more existing features of the dataset. The calculation and derivation of the new feature is based on the domain knowledge and statistical analysis of the key attributes. The new features are resultant of if-else conditions subjected on categorical and numerical attributes, i.e., equating to a specific string value or evaluating its value over or under specific threshold. The output of the features engineered were numerical and binary (0,1) in nature.

3.4.2 Feature Encoding and Scaling

Feature Encoding is the branch of feature engineering that transforms categorical features into numerical values as machine learning models can only interpret numerical values. Due to the above limitation, the binary categorical columns are converted into binary numerical attributes. The output of the feature encoding were binary encoded variables that contain value 0 or 1. Feature Scaling is a process that normalizes the numerical values within a specified range. It reduces the memory required for computation and widely accepted technique for outlier treatment. Standard scaling transforms the values within a variable into normal distribution by eliminating the mean and scaling on the basis of standard deviation. The output of the feature scaling resulted in scaled values for numerical attributes that do not contain any more outliers.

3.4.3 Correlation Matrix

It is a visual representation of correlation coefficients between independent variables in the form of a matrix. The coefficient value above a specific threshold between two variables is considered as a case of multicollinearity. Such variables that have strong intercorrelation are excluded from the dataset as practice of feature reduction.

3.4.4 Class Balancing using Oversampling

The preliminary study when dealing with binary classification problem is to check for class distribution. The dataset used in this research has class imbalance problem which is handled using synthetic oversampling methods such as Random Oversampling and SMOTE.

- Random Oversampler picks records from minority class and adds to the training set to balance the class distribution. Although, it increases the redundancy in data.
- SMOTE selects instances in the feature map that are similar, creates distinctive observation between the instances and creates a new sample to add in the training set.

The output of the class balancing was arrays of data constituting of equal distribution of classes.

3.5 Modelling

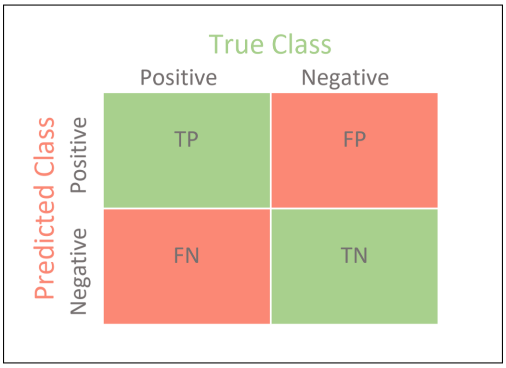
The fifth phase in the KDD process is Modelling. The output from the data transformation stage is used as input for the models for training and prediction. The selection of models is based on the literature reviewed and the performance is compared to that of the benchmark research.

1. **Random Forest Classifier**: It is a decision tree-based ensemble learning method that enhances prediction by balancing out the resultant output of numerous decision trees. In this research, the first instance of random forest classifier takes 3 datasets as inputs such as class imbalanced dataset, randomly oversampled and SMOTE oversampled data and returns a trained model with their corresponding cross validation F1 scores. Multiple experiments are conducted using Random Forest classifier, each pertaining to a different approach in order to achieve best results.
2. **LightGBM Classifier**: It is gradient boosting method that extends leaf-wise and extends vertically. It is memory efficient, extremely fast in processing and offers extensive range of parameters for tuning. The randomly oversampled data is used as input with the tuned hyperparameters to produce a trained model. The output of the LightGBM classifier is a trained model which is used to predict the employee attrition on test data.

3.6 Evaluation

In a typical binary classification problem, models are often evaluated based on the correct predictions (accuracy) made. However, in an event where model is based on class imbalanced dataset, some other metrics are critical to consider while evaluating the credibility of the model.

- **Confusion Matrix** : It is graphical representation of 4 possible outcomes of binary classification in 2x2 matrix represented as True and Predicted Class as along the x & y axes. The 4 quadrants are labelled as True Positive, False Positive, False Negative and True Negative as illustrated in the figure 7.



		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Figure 7: Confusion Matrix - Understanding

True Positive: Actual and Predicted are Positive
True Negative: Actual and Predicted are Negative
False Positive: Actual is Negative, Predicted as Positive
False Negative: Actual is Positive, Predicted as Negative

- **F1-Score** : It is the weighted average of Precision and Recall that explains how efficiently and accurately a model is able to classify cases into positive & negative classes.[4].

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{1}$$

where 'TP' is True Positive and 'TN' is True Negative.The input for F1 score is arrays of actual and predicted values and the output is score. In this research the F1-score is primary metric of evaluation, and it is also used for comparison across the opted model approaches.

- **Accuracy** : It is the fraction of total number of accurate predictions over total number of predictions made by model. The input for accuracy is same as F1-score i.e., array of actual and predicted values and output is the accuracy score.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

- **ROC AUC Score** : It is the graph that plots the performance of a classification model at specific thresholds and evaluates the score based on the aggregation of performance.[5]

4 Design Specification

The design specification for this research is the integral part of the data mining process. This section elaborates upon the methods and their corresponding outputs stated in the section 3. It deals with nuances observed in the data and the solution implemented in

[4] Interpretation of Performance Measures:https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

[5] ROC Curve and AUC : https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

aiding the research to meet the objectives.
The initial data preprocessing performed validations such as missing values, data profiling and outlier detection. It was observed within the profiling output that there were 3 variables that contained constant value such Employee Count, Over18, Standard Hours. These columns were dropped from the data along with Employee Number which was an auto-increment number.

4.1 Outliers Detection

The analysis on outliers was performed by analyzing the numerical features existing within the dataset. Based on the number of outliers observed, the variables were categorised into 3 sections as stated below.

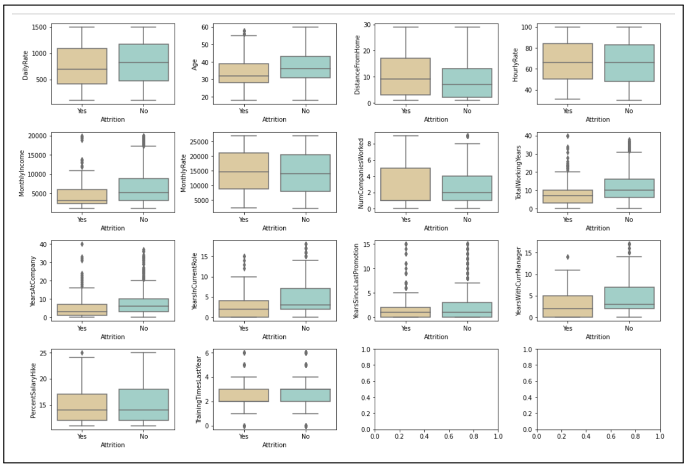


Figure 8: Outliers Detection

From the above figure 8, the following inferences were drawn:

- DailyRate, Age, DistanceFromHome, HourlyRate, MonthlyRate, PercentSalaryHike do not have any outliers or negligible outliers.
- NumCompaniesWorked, YearsInCurrentRole, YearsWithCurrManager, TrainingTimesLastYear have a significant number of outliers.
- MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsSinceLastPromotion have large number of outliers present in the data.

The outlier treatment on the numerical variables will be performed using Standard Scaling before preparation of dataset for model training.

4.2 Chi-squared Test

The research conducted the Chi-squared test on the dataset on categorical variables to check if they are statistically significant with the target variable (”Attrition”). The results are illustrated in the table 2 for all the variables that were found not statistically significant with the target variable.

Table 2: Chi-squared test results

Variable	**p-value**	**p-value<0.05**
PerformanceRating	0.990075	False
Education	0.545525	False
Gender	0.290572	False
RelationshipSatisfaction	0.154972	False

The above columns were dropped from the dataset. The other variables which were found to be statistically significant were further analyzed through visualizations so as to decipher their relation with attrition and non-attrition cases.

4.3 Feature Engineering, Feature Encoding and Scaling

In this research, 11 new features were engineered from the existing variables and their relationship with the target variable were analyzed. Based on the analysis, the following new features created were **SalesDpt, RDDpt ModJobInv, ModTraining, MeanSatisfaction, OverSatRating, LongDis, Hrate_Mrate, Stability, TotalCompWorked and Loyalty.** The existing features used in creating the new features were dropped from the dataset to reduce the dimensionality. The following features were dropped from the dataset **Department,JobInvolvement, TrainingTimesLastYear, RelationshipSatisfaction, EnvironmentSatisfaction, JobSatisfaction, JobInvolvement, WorkLifeBalance, DistanceFromHome, HourlyRate, MonthlyRate, NumCompaniesWorked, TotalCompWorked, TotalWorkingYears, YearsInCurrentRole and 'YearsAtCompany'.**
All the binary columns (Features with 2 unique values) were converted into numerical values using Label Encoding.
All the other category columns which had less than 10 unique values within the entire dataset were subjected to dummy encoding.
The numerical columns were identified from the data and standard scaling was performed.

4.4 Correlation Matrix

The correlation matrix and analysis was performed on the feaures as illustrated below in figure 9. It was a complex task to evaluate this correlation matrix due to large number of features in the data, the collinear variables were derived programmatically. It was found that 3 columns **RDDpt**, **JobRole_Sales_Executive** and **MeanSatisfaction** were found to have strong correlation above the threshold (0.8). These columns were dropped from our dataset.

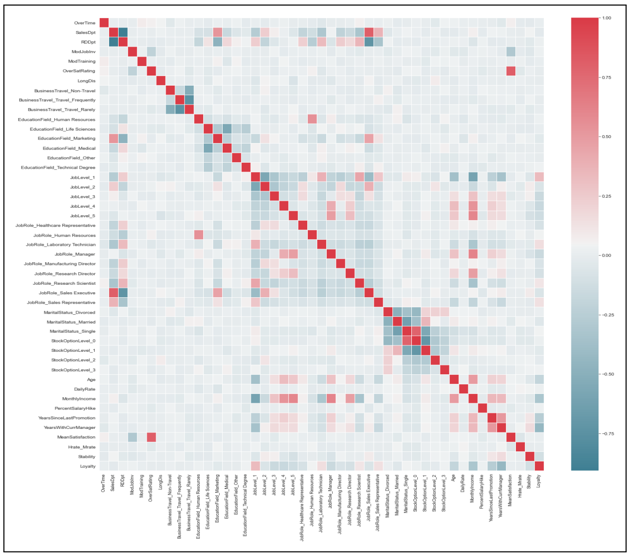


Figure 9: Analysis of Correlation Matrix

4.5 Train and Test Split

The input data for model training was our data with 44 features excluding the target variable which was **Attrition**. The data was split into Train and Test with a 80-20 split and stratify parameter set as **'y'**. The stratify parameter ensures that the data split into train and test result in similar .

4.6 Class Balancing

It was observed during the data preprocessing stage that the data used in the research conforms as a typical case of class imbalance. It is not considered as a best practice to use highly imbalanced data for model training. The most reliable methods for handling class imbalance are Undersampling and Oversampling. However, the dataset used in our research had limited number of observations, therefore, we performed oversampling. As a part of the experiment, we proposed to evaluate the performance of model training over 3 datasets, original with class imbalance, Oversampled data using Random Oversampler and Oversampled data using SMOTE (Synthetic Minority Oversampling Technique). The output of oversampling techniques employed increases the samples in minority class ('Yes') to match the number of samples in majority class ('No') and create a new training set.

4.7 Modelling

The proposed research methodology investigated the application of machine learning models for a binary classification problem. Different eclectic approaches were implemented in combination with high performing classifiers to predict attrition in employees. The approaches that were undertaken are listed below:

4.7.1 Random Forest Classifier using Grid Search Cross Validation and Selection of Best Hyperparameters - Model 1

Random Forest Classifier was implemented using 10-fold Grid Search Cross Validation on imbalanced and balanced datasets to train our model. The Grid Search CV is a function that takes a classifier model as an input for estimator and hyperparameter grid as options for parameter grid. The CV parameter defines the k-fold cross validations along with the scoring metric based on which the model performance is assessed. The primary advantage of using GridSearchCV is that it returns best parameters and highest cross-validation score achieved during the training process. The hyperparameters for the random forest classifier defined with the hyperparameter grid were **n_estimators**, **max_features**, **min_samples_leaf** and **criterion**. The **cross validation** was set to 10 and **scoring** evaluation metric was set to F1. The output in the form of best parameters along with highest cross-validation score is applied to a baseline Random Forest classifier and further evaluated to extract performance metrics such as F1-score, Accuracy and ROC AUC score on the test data.

4.7.2 Random Forest Classifier using Recursive Feature Elimination with 10-fold Cross Validation - Model 2

Recursive Feature Elimination (RFE) is an advanced feature selection technique used to evaluate the performance of the model by subsequent executions by iterating through a subset of existing features from the data. RFE is implemented using a cross-validation approach to ensure that model doesn't overfit. The best parameters from the former model are used as the yardstick to tune the random forest classifier. The output elucidates the number of optimal features that produce the best cross-validation F1-score. These features were identified to create a new training set for model training and results on test data were evaluated using F1-score, Accuracy and ROC AUC score.

4.7.3 Random Forest Classifier using Manual Hyperparameter Tuning - Model 3

The third approach in model building was a manual attempt in configuring the hyperparameters of a typical Random Forest Classifier. The hyperparameters tuned were **n_estimators**, **criterion**, **max_depth**, **max_features**, **min_samples_leaf**, **min_samples_split** and **class_weight**. These parameters were chosen to iterate through a optimal set of values and return the corresponding training and test F1-score of the model.

4.7.4 Light Gradient Boosting Classifier using Manual Hyperparameter Tuning - Model 4

A similar attempt like Model 3 was performed on LightGBM Classifier for employee attrition prediction. From a wide range of hyperparameters that are offered by LightGBM classifier, **n_estimators**, **learning_rate**, **max_depth**, **num_leaves** and **subsample** are selected for tuning. This model performed parameterized execution through n-iterations where n is the number of possibles values provided against each hyperparameter.

5 Implementation

The research was implemented using **Python** as the programming platform due its vast libraries and plug-ins support. **Jupyter Notebook** was used as the presentation interface for producing the results of the research visually. The prerequisite libraries are imported for Numerical Calculations and Dataframe such as **numpy** and **pandas**. The missing values validation was performed using **missingno** library. The data profiling was facilitated by importing the **ProfileReport** function from **pandas-profiling** library. Pandas Profiling is a specific python library that creates a profiling report of the dataset. The outliers analysis was conducted on numerical variables within the dataset using **boxplot** visualization which is a predefined function of **seaborn**. Seaborn is a data visualization library that is used for creating statistical charts. Chi-Squared Test was also performed on the variables to check their statistical significance against the target variable. In order to achieve this, **chi2_contingency** was inherited from **scipy.stats** module. The exploratory data analysis is the one of the most significant part of the research as it helped analyze the features of the dataset with respect to target variable using interactive plots powered by **matlplotlib.pyplot**, **countplot** from **seaborn** library, various plots (bar, scatter, pie) from **plotly.graph_objs**. The holistic list of data visualization libraries were used in the EDA. The correlation between the variables was produced using **corr** function from **pandas.DataFrame** package and illustrated using **create_distplot** from **plotly.figure_factory**. The next phase in sequence was Feature Engineering, Feature Encoding and Scaling. The feature engineering was implemented using basic transformations and lambda function. Feature Encoding used **LabelEncoder** and **get_dummies** from **sklearn.preprocessing** and **pandas** package respectively. The normalization of values was performed using **StandardScaler** from **sklearn.preprocessing**. The data was then split into train and test set using **train_test_split** imported from **sklearn.model_selection** library. Once the training and test samples were created, the class balancing techniques were incorporated using **RandomOversampler** and **SMOTE**, functions imported from **imblearn.over_sampling**. Once the data was ready for model training, the baseline models **RandomForestClassifier** was imported from **sklearn.ensemble** and **LGBMClassifier** was imported from **lightgbm**. The approaches defined within the Modelling section different libraries adhering to the proposed research were imported such as **make_pipeline** from **sklearn.pipeline** to be used an estimator for **GridSearchCV** imported from **sklearn.model_selection** to identify the best parameters and best Cross validation score. Another approach pertaining to Recursive Feature Elimination required to import **RFECV** from **sklearn.feature_selection**. Once, all the models were trained, these models were used to perform predictions on the test data. The evaluation of the model performance was expedited through evaluation metrics such as **f1_score**, **accuracy_score** and **roc_auc_score**, **confusion_matrix** and **classification_report** from **sklearn.metrics**.

6 Evaluation

The next crucial phase in the data mining process is the evaluation. Various models that have been developed to predict employee attrition are critically evaluated based on the chosen metrics such as F1-score, Accuracy and ROC AUC score.

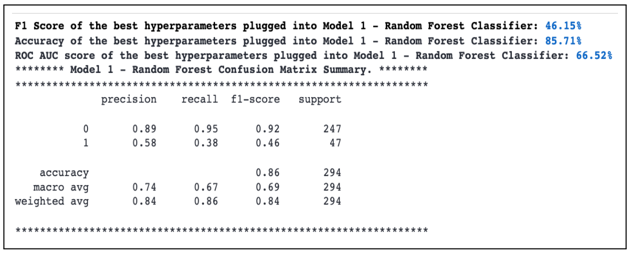
6.1 Model 1 - Random Forest Classifier using GridSearchCV

The random forest model with GridSearchCV was trained on 3 datasets for a 10-fold CV. The summary of best parameters and performance was recorded against each data as illustrated in the table 3. It was observed that the Randomly Oversampled data recorded the best performance compared to Original Imbalanced and SMOTE oversampled data.

Table 3: Best Parameters and Training F1-score

Parameters	**Original Imbalanced**	**Randomly Oversampled**	**SMOTE Oversampled**
n_estimators	80	80	100
criterion	entropy	gini	gini
max_features	sqrt	log2	log2
min_samples_leaf	5	1	1
CV Training F1-score	48.00%	98.65%	91.82%

The best parameters for randomly oversampled data were chosen to run against a baseline random forest classifier and evaluate its performance on the test data. Also, the randomly oversampled data was used for model training for all the model. The performance of the random forest classifier using the best parameters produced a F1-score of **46.15%**, Accuracy of **85.71%** and ROC AUC score of **0.66** as depicted the figure 10.



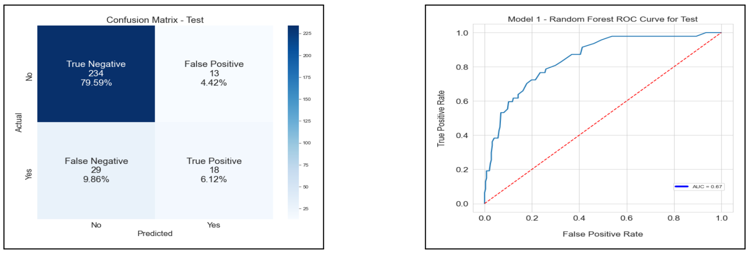
```
F1 Score of the best hyperparameters plugged into Model 1 - Random Forest Classifier: 46.15%
Accuracy of the best hyperparameters plugged into Model 1 - Random Forest Classifier: 85.71%
ROC AUC score of the best hyperparameters plugged into Model 1 - Random Forest Classifier: 66.52%
******** Model 1 - Random Forest Confusion Matrix Summary. ********
******************************************************************
              precision    recall  f1-score   support

           0       0.89      0.95      0.92       247
           1       0.58      0.38      0.46        47

    accuracy                           0.86       294
   macro avg       0.74      0.67      0.69       294
weighted avg       0.84      0.86      0.84       294

******************************************************************
```

(a) Model 1 - Performance Metrics

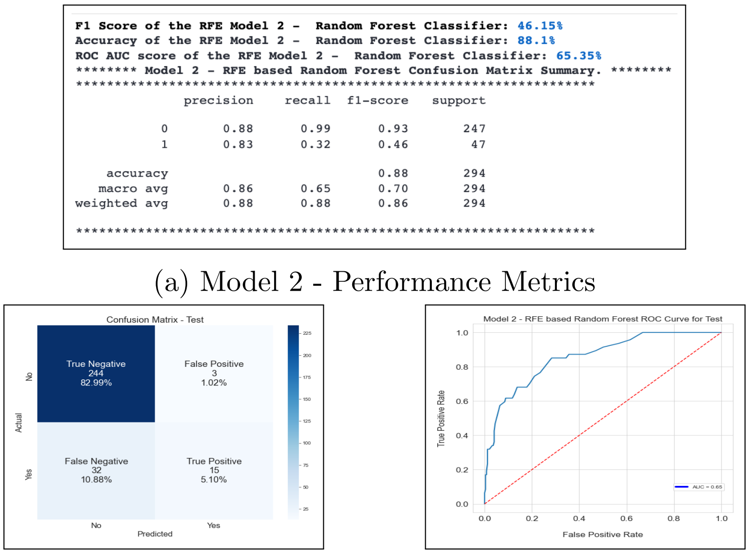


(b) Model 1 - Confusion Matrix (c) Model 1 - ROC AUC Score

Figure 10: Model 1 - Evaluation

6.2 Model 2 - Random Forest Classifier using Recursive Feature Elimination

The recursive feature elimination that performed model training on randomly oversampled data suggested the optimal number of features as 43 by achieving a F1-score of **98.75%**. Based on the features suggested by RFE, the training set was reduced to 43 features for model training.



```
F1 Score of the RFE Model 2 -  Random Forest Classifier: 46.15%
Accuracy of the RFE Model 2 -  Random Forest Classifier: 88.1%
ROC AUC score of the RFE Model 2 -  Random Forest Classifier: 65.35%
******** Model 2 - RFE based Random Forest Confusion Matrix Summary. ********
*********************************************************************
              precision    recall  f1-score   support

           0       0.88      0.99      0.93       247
           1       0.83      0.32      0.46        47

    accuracy                           0.88       294
   macro avg       0.86      0.65      0.70       294
weighted avg       0.88      0.88      0.86       294

*********************************************************************
```

(a) Model 2 - Performance Metrics

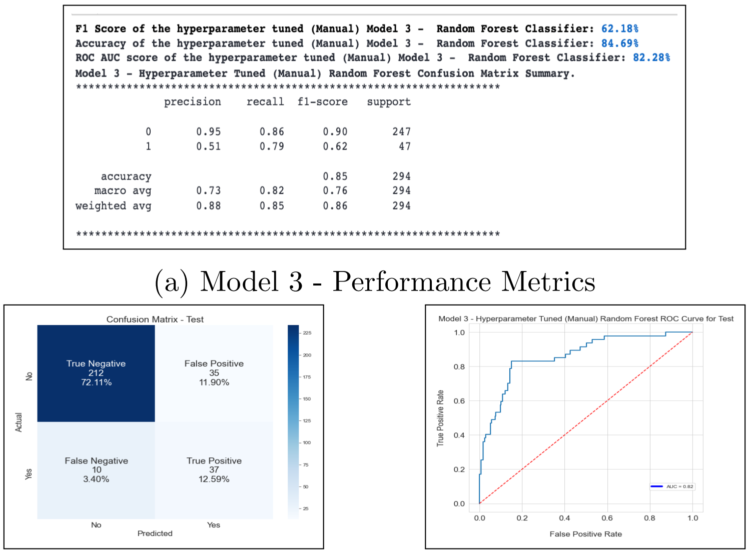
(b) Model 2 - Confusion Matrix (c) Model 2 - ROC AUC Score

Figure 11: Model 2 - Evaluation

The performance of the random forest classifier using 43 features to train the model produced a F1-score of **46.15%**, Accuracy of **88.10%** and ROC AUC score of **0.65** on test data as depicted the figure 11.

6.3 Model 3 - Random Forest using Manual Hyperparameter Tuning

The manual tuning of hyperparameter for a Random Forest Classifier was performed that yielded a Train F1 score of **94.50%** and Test F1 score of **62.18%**. The parameters that produced the best results are listed below:
n_estimators : 80, **criterion : 'gini'**, **max_depth : 14**, **max_features : 'sqrt'**, **min_samples_leaf : 11**, **min_samples_split : 2**, **class_weight : 'balanced'**



```
F1 Score of the hyperparameter tuned (Manual) Model 3 -  Random Forest Classifier: 62.18%
Accuracy of the hyperparameter tuned (Manual) Model 3 -  Random Forest Classifier: 84.69%
ROC AUC score of the hyperparameter tuned (Manual) Model 3 -  Random Forest Classifier: 82.28%
Model 3 - Hyperparameter Tuned (Manual) Random Forest Confusion Matrix Summary.
*************************************************************
              precision    recall  f1-score   support

           0       0.95      0.86      0.90       247
           1       0.51      0.79      0.62        47

    accuracy                           0.85       294
   macro avg       0.73      0.82      0.76       294
weighted avg       0.88      0.85      0.86       294

*************************************************************
```

(a) Model 3 - Performance Metrics

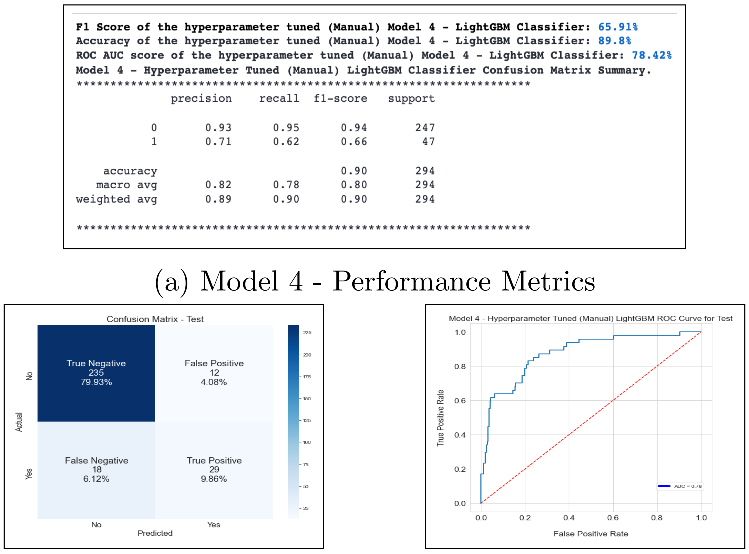
(b) Model 3 - Confusion Matrix (c) Model 3 - ROC AUC Score

Figure 12: Model 3 - Evaluation

The parameters that produced the best results were used to train the model which yielded F1-score of **62.18%**, Accuracy of **84.69%** and ROC AUC score of **0.82** on Test data as illustrated in the above figure 12. The Random Forest Model with hyperparameter tuning outperformed the other two models that used a different approach. The feature importance was plotted for this model which implied that Overtime, Monthly Income and Loyalty were top 3 important features.

6.4 Model 4 - LightGBM Classifier using Manual Hyperparameter Tuning

The manual tuning of hyperparameter for a LightGBM Classifier was performed that yielded a Train F1 score of **100%** and Test F1 score of **65.90%**.
n_estimators : 320, **subsample : 0.1 max_depth : 5 num_leaves : 40 learning_rate : 0.4**



```
F1 Score of the hyperparameter tuned (Manual) Model 4 - LightGBM Classifier: 65.91%
Accuracy of the hyperparameter tuned (Manual) Model 4 - LightGBM Classifier: 89.8%
ROC AUC score of the hyperparameter tuned (Manual) Model 4 - LightGBM Classifier: 78.42%
Model 4 - Hyperparameter Tuned (Manual) LightGBM Classifier Confusion Matrix Summary.
********************************************************************
              precision    recall  f1-score   support

           0       0.93      0.95      0.94       247
           1       0.71      0.62      0.66        47

    accuracy                           0.90       294
   macro avg       0.82      0.78      0.80       294
weighted avg       0.89      0.90      0.90       294

********************************************************************
```

(a) Model 4 - Performance Metrics

(b) Model 4 - Confusion Matrix (c) Model 4 - ROC AUC Score

Figure 13: Model 4 - Evaluation

The parameters that produced the best results were used to train the model which yielded F1-score of **65.91%**, Accuracy of **89.80%** and ROC AUC score of **0.78** on Test data as illustrated in the above figure 13. The LightGBM classifier with hyperparameter tuning produced slightly. . The feature importance was plotted for this model which implied that **Monthly Rate**, **Hrate_Mrate** and **Daily Rate** were top 3 important features.

Discussion

The above four models were implemented consequentially each following a distinctive method of implementation. It was observed that random forest classifier when used with GridSearchCV and Recursive Feature Elimination exhibited an ordinary performance and not at par when compared to manually tuned Random Forest Classifier. On the other hand, LightGBM classifier demonstrated slightly better test accuracy and F1 score but lower ROC AUC score. However, the training Accuracy, F1 score and ROC AUC score of

100% for LightGBM suggests the model has overfitted and would succumb to acknowledge new data points. The random forest classifier with hyperparameter tuning showcased a F1 score of 94.50%, Accuracy of 94.42% and ROC AUC score of 0.94. The results indicate that data transformation techniques such as feature engineering, encoding, scaling and model augmentation using hyperparamter tuning has been partially effective in improving the model performance in prediction of employee attrition.

7 Conclusion and Future Work

The research aimed to perform various data transformations such as feature engineering, feature encoding and scaling and selecting from a melange of approaches to tune the hyperparameters on classification models in order to predict employee attrition effectively. The business outcome of this research would help organizations strategize their employee management practices and take cautionary measures to undermine the impact of attrition. The exploratory data analysis on attrition data helped to unravel the factors that insinuate the possibility of attrition. Monetary driven factors such as Monthly Income, Daily Rate and Hourly Rate seemed to diminish the motivation in employees. It was also observed that Job Satisfaction, Environment Satisfaction, Relationship Satisfaction are few psychological metrics that affects the mental health of employees and organizations should ensure to facilitate a healthy working environment and relationship with the employees. The analysis on some other factors such as Overtime and Distance from home imply that the attrition rate is perceived higher for employees who are subjected to work Overtime or whose Distance from home is greater than 9. This research was a binary classification problem with presence of class imbalance. In such a scenario, it was imperative to critically evaluate True Positive Rate and True Negative Rate rate rather than just the accuracy. Based on the evaluation, it can be concluded that Random Forest classifier did not overfit unlike LightGBM and showed a significant increase in performance using manual hyperparameter tuning. The rationale for rejecting the Random Forest classifier that used the GridSearchCV and RFE approach was its poor F1-score, accuracy and ROC AUC score when compared to that for manually hyperparameter tuned Random Forest classifier.

Due to limitations in the volume of data, it would be preferred to perform this research on a larger dataset. Also, this data is fictional in nature, so, the veracity of data is questionable. The insights extracted from this data might not resonate under a realistic organizational setting. The underlying objective of future work should align with building a more parsimonious model that achieves impeccable performance.

Acknowledgment

I would like to take this opportunity to express my sincere gratitude and heartfelt wishes to Dr. Martin Alain who has been my guiding light throughout the research. His constant support and guidance has helped me to address the concerns of my research.

References

Alawad, W., Zohdy, M. and Debnath, D. (2018). Tuning hyperparameters of decision tree classifiers using computationally efficient schemes, *2018 IEEE First International*

Conference on Artificial Intelligence and Knowledge Engineering (AIKE), pp. 168–169.

Alhashmi, S. M. (2019). Towards understanding employee attrition using a decision tree approach, *2019 International Conference on Digitization (ICD)*, pp. 44–47.

Bhartiya, N., Jannu, S., Shukla, P. and Chapaneri, R. (2019). Employee attrition prediction using classification models, *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, pp. 1–6.

Cotton, J., Turtle, J. and Turnover, E. (1986). Employee turnover: A meta-analysis and review with implications for research, *Acad Manage Rev* **1986**: 55–70.

Daoud, E. A. (2019). Comparison between xgboost, lightgbm and catboost using a home credit dataset.

Frye, A. L., Boomhower, C., Smith, M., Vitovsky, L. and Fabricant, S. M. (2018). Employee attrition: What makes an employee quit?

Hoffman, M. and Tadelis, S. (2017). People management skills, employee attrition, and manager rewards: An empirical analysis, *SSRN Electronic Journal* .

Jain, R. and Nayyar, A. (2018). Predicting employee attrition using xgboost machine learning approach, *2018 International Conference on System Modeling Advancement in Research Trends (SMART)*, pp. 113–120.

Joseph, R., Udupa, S., Jangale, S., Kotkar, K. and Pawar, P. (2021). Employee attrition using machine learning and depression analysis, *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1000–1005.

Mobley, W. H. (1977). Intermediate linkages in the relationship between job satisfaction and employee turnover., *Journal of Applied Psychology* **62**: 237–240.

Nappinnai, M. and Premavathy, N. (2013). Employee attrition and retention in a global competitive scenario, *International Journal of Research in Business Management (IMPACT: IJRBM)* **1**(6): 11–14.

Ozoliņa-Ozola, I. (2015). Reducing employee turnover in small business: An application of employee turnover models.

Rohit Hebbar, A., Patil, S. H., Rajeshwari, S. B. and Saqquaf, S. S. M. (2018). Comparison of machine learning techniques to predict the attrition rate of the employees, *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT)*, pp. 934–938.

Shekar, B. H. and Dagnew, G. (2019). Grid search-based hyperparameter tuning and classification of microarray cancer data, *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*, pp. 1–8.

Taunk, K., De, S., Verma, S. and Swetapadma, A. (2019). A brief review of nearest neighbor algorithm for learning and classification, pp. 1255–1260.

Wang, X. (2010). An analysis of the cause of employee turnover intention in hotels, *2010 International Conference on Management and Service Science*, pp. 1–4.

Yadav, S., Jain, A. and Singh, D. (2018). Early prediction of employee attrition using data mining techniques, *2018 IEEE 8th International Advance Computing Conference (IACC)*, pp. 349–354.

Yedida, R. and Vahi, R. (n.d.). Employee attrition prediction, p. 3.
ΩŁoś et al.

Łoś, H., Mendes, G. S., Cordeiro, D., Grosso, N., Costa, H., Benevides, P. and Caetano, M. (2021). Evaluation of xgboost and lgbm performance in tree species classification with sentinel-2 data, *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 5803–5806.