Enhancing Efficiency of Employee Attrition Prediction Using Machine Learning and Ensemble Techniques

MSc Research Project
Master of Science in Data Analytics

Linu Anil
Student ID: X23183852

School of Computing
National College of Ireland

Supervisor: Anh Duong Trinh

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Linu Anil
……………………………………………………………………………………………………

Student ID: X23183853
……………………………………………………………………………………………………..…

Programme: Master of Science in Data Analytics
………………………………………………………………… **Year:** 2024 - 2025
………………………….
.

Module: Research Project
……………………………………………………………………………………………………….…..

Supervisor: Anh Doung Trinh
……………………………………………………………………………………………………….….

Submission Due Date: 29/12/1014
…….…

Project Title: Enhancing Efficiency of Employee Attrition Prediction Using Machine Learning and Ensembling Techniques
……………………………………………………………………………………………………….…

Word Count: …………8789…………… **Page Count**…………………22……………………….……..

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: Linu Anil
……………………………………………………………………………………………………

Date: ……….28/12/2024………………………………………………………………………

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):	

Enhancing Efficiency of Employee Attrition Prediction Using Machine Learning and Ensemble Techniques

Linu Anil
X23183853

Abstract

Employee turnover remains a major concern in organizations because it results in the cost of hiring new employees, loss of knowledge and interruption of organizational activities. This paper examines the applicability of machine learning to estimate the probability of employees' turnover and help HR departments develop practical retention strategies. To prepare the dataset for modelling, initial data cleaning included outlier removal, employee data balancing, feature scaled using standard scalar on the IBM HR Employee dataset and hyper parameter tuning applied to the model optimization. In the modelling phase the Random Forest, XGBoost, and the stacking ensemble model were considered and experimented with. The stacking model, which uses tuned XGBoost and Random Forest as base models and Logistic Regression as the final estimator, was found most effective with an accuracy of 99.59 %, 100 % Precision, 98.30% Recall, and F1 score of 99.14%. The study also reveals how ensemble learning is beneficial for processing multiple dimensional HR data, managing problems such as class imbalanced data and noisy inputs. Although the data set restricts generality, the outcomes still prove the capability of machine learning for enhancing workforce analysis. This research is important to show how proactive measures such as predictive analytics can achieve strategic human resource management and organizational resilience.

1. Introduction

Employee turnover is now a serious issue that hounds the management of employees in an organization. It is therefore very important to address strategies for the retention of its employees, since high attrition can cause a lot of problems like an increase on the recruitment costs and most importantly loss of key knowledge during the organization's operation (Al-Suraihi et al. 2021). All of these issues can be addressed through usage of the data for building the machine learning models for estimating the attrition and aiding HR in decision making. This paper analyzed efficient machine learning algorithms such as Random Forest and XGBoost and get directions for action. For HR departments, the capability of accurate prediction of employee turnover is an essential success factor that enables them to create effective retention plans thereby decreasing costs associated with changes in employee turnover and improving the organizational stability. Even though the statistical methods are a classical approach and form a starting point, they do not allow revealing the patterns of HR data. Furthermore, conventional statistical techniques are an insufficient solution to this problem because they cannot process big and diverse data sets, and this is where machine learning comes in. Although, high predictive accuracy is not easy to accomplish because of the problems such as, imbalanced datasets, noisy data and more importantly HR attributes involve multiple dimensions. Therefore, the rationale for undertaking this research rests with the notion that improved accuracy of the models by implementing proper preprocessing in the data and that will ultimately translate into a corresponding improvement in the efficiency of organizations.

Research Question:

How can the development of machine learning algorithms with proper data preprocessing and ensemble modelling to effectively forecast employee attrition and help HR professionals to understand the employees at risk for better decision-making and organizational stability?

The objectives are to explore the current practices and an examination of issues in employee attrition prediction and manage it using machine learning, and ensemble methods. The framework for predicting attrition involves preliminary data collection, cleaning, feature engineering, and tailored modelling algorithms such as XGBoost, Random Forest and a stacking model using base model as XGBoost and Random and final estimator Logistic Regression. Training and testing of all the implemented models were evaluated using metrics such as accuracy, precision, recall, F1 scores and ROC curve. Finally cross validation is done in the training data of the best model for concern over fitting and or under fitting problems. Then the best model will be deployed using stream lit to make it user friendly web application which enhances the effectiveness of the HR decisions. The main objective of this study is to construct a novel machine learning-based system for employee attrition prediction with effective preprocessing and ensemble modeling. This framework not only concludes high predictability rate but also helps the HR professionals to understand the forces behind attrition and develop proper solutions. This study has some limitations, mainly the data was retrieved from the IBM HR Employee dataset from Kaggle, and it cannot be able to generalize and whether it can capture the full picture of the factors that lead to attrition across industries.

The structure of this paper is organized as follows: In the first part of the project, the ‘Introduction’ section is presented to define the context of the specified study and to state the goal of the research. The Related Works section highlights prior findings, and the current methods used in the context of the employee attrition prediction task. The Research Methodology section describes the data set used, data pre-processing, and the Modelling strategies used. The Design Specification explains the framework and the system architecture for the study. The Implementation section provides information on the actual creation and putting into operation of the mentioned predictive models. In the Evaluation, a comparison of the models’ performance indicators and consequences is presented. Last is the Conclusion and Future Works where the results are summarized with the findings, highlighted the contributions and future work is proposed.

2. Related Work

This chapter provides a systematic literature review, and as such provides a clear and up-to-date review of current literature surrounding employee attrition prediction models. Employee turnover prediction is today a topical issue due to its profound consequences on organizational outcomes and future.

2.1. Ensemble technique to predict employee turnover

Employee turnover plays a critical organizational performance and sustainability factor. Research suggests that the XGBoost, Random Forest, and SVM are benchmark models for the turnover prediction because XGBoost was found to handle noisy, imbalanced data from HRIS well due to its regularization feature for guaranteed precision and model stability. Lack of completeness and accuracy of entries in the HRIS database increase the risk of overfitting so the should need to optimise. Proper turnover forecasts facilitate the replacement and succession planning within an organisation. Scholars advise that future research should focus on better feature construction and class distribution to refine the model interpretability and accuracy in predicting turnover, the areas where the literature is lacking and has the opportunity to develop realistic application (Ajit, 2016).

In this work, the suitability of XGBoost in turnover prediction is assessed by focusing on the data preparation the performance of the model. When working with the IBM HR Data Analysis dataset, preprocessing handled issue of duplicate and overlapping attributes, missing values, and header mismatch. Even with the databases of reasonable size, XGBoost reached its high level on the accuracy of 89% and the recall level of 73% and rather low precision of 51%. Here data replication arose as an issue, which impacted on the prediction part. However, its applications were vindicated in aspects such as scalability, memory usage, and managing noise brought out by XGBoost error margin below 30%. These outcomes evidence that XGBoost is indeed an effective frame work towards framing the policies regarding employee retention (Jain et al., 2018).

The current study measures turnover intentions in employees working in the IT and ITeS companies selected from Chennai, Coimbatore and Bangalore with 416 respondents. Job satisfaction, stress, gender, and COVID-19 attitudes as variables were considered. Descriptive analysis of dataset included Clean data, Scale and classify data was done using models like Logistic Regression, Naïve Bayes, XGBoost classification. XGBoost had accuracy of 94%, precision of 92%, recall of 95%, F1-score of 0.94 and AUC of 0.94. This study established job stress and organisational support during COVID 19 as factors influencing turnover. That is, the findings are generalizable to other similar contexts, although the selection of IT/ITeS organisations imposes certain restrictions on the generalisability of the findings (Tharani et al., 2020).

2.2. Machine Learning approach to predict employee turnover

Classification models used in this research to forecast employee attrition are Support Vector Machine (SVM), Random Forest, J48, Logit Boost, Multilayer Perceptron, K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA), Naïve Bayes, Bagging, Ada Boost, and Logistics Regression on IBM HR dataset. This work presents the five performance measures that are relevant for imbalances data sets because accuracy alone could mislead the valuation. F-measure, area under the curve, sensitivity, and specificity are the additiona matrices (Ozdemir et al., 2020). It shows how through data mining turnover can be predicted which in turn helps the HR departments in prevention. However there are drawbacks with this study including absence of ways approaching data imbalance for consideration for instance SMOTE or validation techniques for example cross-validation. However, some models might not be practical for real-world HR applications because of their non-interness, but the comparisons give the model efficiency insights.

This research aims at employing machine learning to study employee attrition and in the process analyzing the issue of class imbalance in the data collected. Three experiments were conducted, the study investigates how raw, class-imbalanced data can be handled using the SVM and random forests, oversampling minorities through ADASYN and undersampling majorities for balancing. For the experiment KNN with K=3 was used, and on the ADASYN dataset the highest F1-score was obtained and equal to 0.93. When the act of feature selection using the Random Forest algorithm was carried out from 29 initial features down to 12, the F1-score of the model was enhanced to 0.909. However, the study has identified class imbalance and feature selection as the key curtain challenges but due to limitations such as using synthetic data, no external validation, and much time consumed by KNN (Alduayj et al., 2018).

In this paper, the author investigates the application of predicting employee turnover or attrition through ML, with a primary emphasis on model performance and interpretability for managerial and strategic human resource decisions. The data used in this study was the IBM human resource dataset which consists of 1470 observations and 34 independent variables On building the models the following algorithms were employed: Logistic Regression (LR), Random Forest (RF), Classification Trees (CT), Naïve Bayes (NB), Neural networks (NN), Voting Ensemble. The best performance was shown by the LR model with accuracy of 88% and AUC-ROC of 85%. This paper outlines the major determinants of attrition including length of service, position, and distance from home to work. Although instrumental in decreasing turnover, this research lacks an examination of complex ensemble algorithms or new feature creation (Guerranti et al., 2022).

This paper aimed at making a predictive modeling with IBM Employee Attrition data. From the Random Forest model, it was found that the features like monthly income, age and number of companies worked for are significant. Over18, StandardHours were deleted due to insignificance of their coefficient of determination and EmployeeCount as well. Hypothesis one subjecting employees to K-means clustering based on turnover intention showed noticeable differentiation while binary logistic regression established factor contributions differentiating high travelers as 2.4 more likely to leave and HR employees were also shown to have a higher turnover intention. The study just offers a classical kind of approach which enforces some restrains here and there and barely employs higher rank ML algorithms, cross validation approaches to overseeing class disproportionation (Yang et al., 2020).

Data Preprocessing Techniques and Methods to predict the employee attrition have been applied on the IBM HR Analytics dataset with 1470 instances and 35 number of features using supervised Machine Learning and Deep Learning. The independent variables included job satisfaction, overtime, and monthly income whereby those employees with low job satisfaction and those who worked many hours overtime were found to have high turnover rates. Logistic Regression, Random Forest, XGBoost and Stacking plus FNN and CNN were used. FNN accuracy was 97.5% with F1 score of 91.26% moreover the functional representation achieved 99% accuracy. A lot of predictive performance is achieved but due to a limited dataset, the study has low external validity, and the deep learning models are not easy to interpret, which shows the desire for large datasets and effective HR solutions.

The issue under consideration in this paper is the application of machine learning in the context of turnover cost, an important challenge for HR management. Employment features such as satisfaction level, number of hours worked overtime, salary level, and age were deemed important surrogates when working with the IBM HR Analytics Dataset (1,500 samples, 35 features). Naïve Bayes classifier based on Gaussian model delivered the best recall number – 0.54, proving that the algorithm is useful to filter out the risky employees. Scaling, encoding, balancing and cross validation were done on the given data-set. Including, but not limited to, the use of synthetic data, moderate recall capability, and the absence of the variety of features that can be implemented in practice. A real-life application in Python was created for the HRM practitioners to use SHRM data to apply the prediction of probability of attrition and to improve on the retention policies (Fallucchi et al., 2020).

This paper discusses the prediction of employee turnover by deploying algorithms such as Support Vector Machine, Decision Tree, K-Nearest Neighbor, Random Forest, and Naïve Bayes in Python language as suggested by Bhartiya et al., 2019 . In details, to work with the IBM HR Employee Analytics Attrition dataset (Total records:1470, total features:35), this paper performs feature reduction, Label Encoding, and SMOTE while considering class imbalance. The models identified were tested using accuracy, confusion matrices, and ROC charts. Random Forest yields 83.3% accuracy, furthermore, Naive Bayes and SVM performed better in terms of true positive cases. Nice features are methodology description, the use of SMOTE, and insights over graphical representations. Some restrictions are connected with the absence of hyperparameter tuning, with the possibility of overfitting of SMOTE algorithm, and with the absence of the check of received results for their application in practice.

In this paper, job satisfaction, work stress, and work-life balance have been identified as the driver for turnover with special emphasis on mental health including depression by means of classification algorithms such as RFC, SVM, DTC, LR, GNB, and KNN based on GDQ (Goldberg's Depression Questionnaire) proposed by Joseph et al. (2021). RFC had the highest accuracy of 86% and an F1-score of 0.85 because it can deal with multicollinearity and large data. The study recommends that subject to the principles mentioned above, several HR strategies, such as enhancing workers' satisfaction with their job and providing a reasonable wage package, should be employed to minimize attrition. Here, the conclusions are also limited by a focus on depression and the necessity of further development of the time spent on the emotional analysis, as well as the improvement of the modeling for real-world adaptation.

Consequently, turnover affects the financial performance of the firm since it brings about fluctuations, interruptions to organizational operations as well as a negative effect on the corporate culture, more explicitly in the context of IT companies, where the turnover rate is 25%. They are; employee dissatisfaction, boredom, transfer to other departments and organizational retirement among others. AI and Machine learning, the logistics regression, decision tree, random forest, xgboost, and.adabost to identify the attrition and enhance the corporate culture. Preprocessing methods include SMOTE, ADASYN, and MinMaxScaler improve data quality, whereas, Random Forest, and Gradient Boosting, provide 97.08% of model accuracy. LIME and SHAP and other forms of machine learning and XAI enhance the interpretability of hybrid models. Also overfitting and low generalizability are still issues most users encounter (George et al., 2022).

The paper aims at comparing the gradient boosting methods, XGBoost, CatBoost, and LightGBM, on a TCS Human Resource dataset obtained from Kaggle for estimating employee turnover. The steps performed in exploratory data analysis are cleaning the columns and label encoding. Models are evaluated solely on accuracy, using a 75:25 train-test split and K-Fold Cross Validation with the k values of 3, 5, 10. LightGBM has the best result of 90.47% for K=10; furthermore, all three inhibitor clusters perform better compared to CatBoost, XGBoost, and classic models SVC and Random Forest. The work is highly technical, employs realistic examples and benchmarks three state-of-the-art gradient boosting schemes, while leveraging the advantages of applied and reproducible research. Limitations are accuracy-based approach, no precision, no recall, F1-score, no tuning, single company dataset used, and inadequate treatment of missing values and outliers (Shaik et al., 2023).

Employee turnover is a significant concern in human resource analytics because recruitment industries mostly experience high costs of hiring new employees to replace the exiting ones. This research employs a Kaggle HR dataset of 1470 records and 10 features to build predictive models. The data divided in the ratio of 90% in training sets and 10% in testing set. Ever among all the evaluated models, Random Forest gave the best results of accuracy at 90.20%, while Naïve Bayes gave an accuracy score of 80.20%. The main advantages of the study proposing casual analyses and the use of realistic datasets for identifying potential predictors of employees' turnover. Limitations are based on exclusively accuracy as the evaluation criterion, and exclusion of measures like precision, recall, F1-score etc. Also, the feature engineering is not comprehensive or complex enough, proper testing validation is not employed and no model calibration is performed and this weaken the overall conclusion of the study. There is also the problem of ignoring data imbalance, that in turn could lead to higher rates of fewer turnover cases (Chakraborty et al., 2021).

This paper focuses on the attempt of Random Forest methodology to identify employee turnover and dissatisfaction by applying the systematical approach on the data from IBM Analytics, with 35 feature set and 1470 samples. On the other hand, Random Forest an ensemble technique constructs several decision trees through Bootstrampling data set, As for balancing the classes, Current study applied SMOTE Technique to increase the precision from 98.833% to 99.472%. But while the metrics were excellent on the training set, only a little better results in the validation set raises questions about working with different sized data sets. The study also has limitations the data is fairly simple to manage and not of high quality; meaning that the model cannot be so easily implemented in real life organisations. This makes it quite clear why high-quality data is importance for predicting the rate of attrition to further improve the prediction the HR analytics should consider using more complex data to come up with better models for estimating this rate (Krishna et al., 2022).

The study employs an ensemble architecture which includes Random Forest feature importance together with several classifiers, the study achieves an employee turnover prediction accuracy of 99.4%. This approach improves prediction by selecting only relevant features and by the use of ensemble method and more importantly the method of interpreting the predictors' importance. Utilizing Chi-Square tests on categorical data, the study outlines the specific turnover predictors and explains how features selection techniques such as Sequential Backward Selection (SBS) decrease the model's dimensionality without outweighing accuracy (Hossen et al., 2020). Bagging and boosting other learning algorithms are also demonstrated with solid rationales for variance reduction and accuracy enhancement asserting the idea of ensemble learning. This paper provides an empirical review of factors that affect turnover with discussion of main predictors at diverse levels of analyses as well as the usability of heuristic machine learning paradigms in the context of the HRM. This research thus gives credence to the proposition that it is important to establish what factors predict turnover and to optimize the feature space in developing accurate, encompassing and generalizable turnover models in human resources. In this study, the authors present a novel bankruptcy prediction model based on feature selection and combining ensemble techniques.

2.3. Stacked model for predicting classification problem

In this paper, the authors have used financial data of Poland-based firms to predict possible future bankruptcy incidences thus minimizing their effects. For feature importance analysis, XGBoost model selection is used, which reduces the model training time and increases the accuracy of the model and removes low-weight features. A stacking ensemble approach then combines several base models: K-Nearest Neighbors, Decision Trees, SVM, and Random Forest; LightGBM is used as a meta-learner to aggregate forecasts leading to higher accuracy and predictive power. The framework utilizes layers of models for better results; the stacking model performs stunningly better than any base model, with results of approximately 97% for each base model. During data preparation: data is scaled, and samples are taken using the stratified method to maintain balanced class labels across different iterations, and this eliminates overtraining. This is in line with recent research on bankruptcy prediction that incorporates, boosting techniques, hybrid models, and SMOTE to control for class imbalance (Muslim et al., 2021). Overall, the proposed framework posits that accurate and scalable prediction through important feature selection can enhance the efficiency of the ensemble model for the improvement of financial decisions for stakeholders.

In this paper, a student classification model is developed where college teachers can be able to learn student performance to be able to adopt the right teaching strategies. Based on the students' performance in Hebei Agricultural University's "Computer Foundation" program, the model categorizes students as low, average or superior performers relative to entrance examination scores, mock examination results, daily practice and semester test scores. The stacking model uses the XGBoost, Random forest, and logistic regression where XGBoost and Random forest are used as weak classifiers. When the feature sets were combined, using five-fold cross-validation, the accuracy was 74.80% which is the highest of all the single models. According to the study, entrance exams, mock exams, and practical exercises are the classifications of the model that should improve entrances, personalized learning, and teacher assistance (Pan et al., 2020).

This paper examines employee turnover issues with help of machine learning models on the given IBM HR Analytics Employee Attrition & Performance dataset (Total records 1470 with 30 columns). Based on accuracy, F1 score, precision, recall, and AUC, the Logistic Regression, Random Forest, XGBoost, SVM, ANN, and a stacking ensemble consisting of Random Forest, XGBoost, SVM, and ANN with Logistic Regression as the meta-model were compared. The author also found that stacking ensemble methods work well on complex data and enhance the model's performance. The study's strengths are using a large dataset that contains diverse data, employing different, modern machine learning algorithms apart from the simple ensemble mean, and employing recent ensemble techniques. Limitations include methods extension, a problem with imbalance data, a high amount of computational time and making it scalable to more extensive datasets or use in real-time applications. Nevertheless, based on these, this study provides useful information on how to deal with the issue of employee attrition, and how to bring about organizational performance improvements for the considered key service providers in the context of the HR practitioners.

This paper seeks to address the problem of employee attrition with IBM HR Analytics dataset that may contain attributes such as age, education, and experience level. It used four machine learning models like Decision Tree, Random Forest, Logistic Regression, AdaBoost, Gradient Boosting, and a Combination method. The study also focus the advantages of each model including Decision Tree's rule based classification of the problem, AdaBoost which targets more on the misclassified instances, Logistic Regression which involves computation of probabilities, and Random Forest which help in overcoming over fitting. Cleaning techniques applied are imputation, correlation analysis, PCA and feature scaling. Evaluation measures are: accuracy, precision, recall and F-measure, ROC-AUC. A strength of this paper is to use a large dataset and multiple models, but the study shows low sensitivity and specificity specifically due to the problem of class imbalance and lacks exploring other models (SVM, or neural networks) or other hyperparameter tuning and generalization using a single source data (Qutub et al., 2021).

2.4. Gap Analysis

The analyzed papers brought several points that can be seen as drawbacks, which weaken the solidity, and the potential usage of the machine learning models for the employee attrition prediction. Further research in feature engineering high level techniques and proper selection of data pre-processing methods are still open problem including interpretability of the model as well as handling imbalanced data sets. All preprocessing issues, such as repeating entries, missing values, and variability in column, are not well handled. Many conclusions are field based, studying IT and ITeS industries allowing little comparison with the other industries where people's behavior may differ. The employment of self-reported data tends to have a bias, particularly for qualitative variables, like job stress and organizational support, with little deliberation of control measures. Besides, organizational level antecedents like personal motivations, job profile, and organizational tenure are relatively neglected even though they influence turnover predictions. A lot of research is noisy, and overfitting is inadequately handled save for XGBoost's regularizations, with little suggestions from many studies on how to deploy these models in HR Information Systems (HRIS) or incorporate them into organizational decision making. The evaluation metrics are commonly used for assessment of performance, where accuracy is the primary measure, however, full range of metrics such as precision, recall, F1 score, AUC are rarely considered, they are crucial in case of dealing with imbalanced data. The focus is on various features of actual organizational realities, which are lacking as all the datasets in the comparison are synthetic, like IBM HR Analytics dataset. In general, problems explored insufficiently concern fine-tuning and post-training model optimization, and model intricacies or workable sample or live compute are not addressed sufficiently. However, there are significant gaps that need to be filled to continue the enhancement and expansion of the use of predictive models in numerous organizations

3. Research Methodology

This chapter describes the extensive approach employed to carry out this research on forecasting employee turnover using advanced ML methods. The framework includes data acquisition, data cleaning, data transformation, model design, model testing and model Deployment. Meticulous methodology was employed to make the generated predictive models more accurate and reliable, including initial and complex experimenting with ensemble learning. Using performance metrics on the final models, the best model was selected and then deployed using Streamlit for web application.

3.1. Research Procedure

Experimental Setup for Employee Attrition Prediction: The research work was performed on Apple MacBook Air 2020 with 8GB RAM and 128 GB ROM operating on the macOS 15. The analysis was conducted in an anaconda environment within a Jupyter Notebook which is web application developed for creating codes. Further the Anaconda environment gave the ability to execute various Python packages for analysis. In process of project, code was written, data was cleansed, transformed and analyzed, and visualization prepared inside the Jupyter Notebook. Python 3 was the most used language in the entire project, data ladder and analysis were done using Pandas, NumPy, Matplotlib, Seaborn, imblearn, XGBoost, Scikit-learn, Imbalanced-learn.

Data Collection: This research adopted the CRISP-DM model to make the analysis and model development and evaluation of the data for the prediction of employee turnover more systematic. The data set used in this analysis was IBM HR Attrition data sourced from Kaggle data repository that combines four sources that gives comprehensive information about the workforce. The data is from 2017 to 2022, and it contain employee satisfaction about workplace with performance review, organization chart illustrating jobs positions and employment hierarchical order, office addresses of 5 Canadian and 3 of US offices, and turnover statistics including the years, reasons, and status of leavers. Collecting information that covers all aspects of an organization allows looking at trends, evaluating the workforce distribution, and identifying reasons for attrition thus providing a strong base for data-oriented HR decisions.

Data Preprocessing

The data preprocessing steps involved handling of missing values through column deletion provided the missing values were more than 50 percent. Categorical data in the columns were encoded into numerical using mapping approaches also outliers were dealt in the IQR manner. In the preprocessing stage, the numerical features were normalized using StandardScaler where SMOTE was used to overcome the imbalance instance problem. PCA was used for the dimensionality reduction with the aim of retaining 95 % of the variance which improved computational speed and the accuracy of the model.

3.2. Model Development

In the modelling phase this research implements three different models:

XGBoost: It is an efficient, high learning algorithm that implemented as a gradient boosting model of decision trees developed in a sequential manner to reduce errors. L1/L2 regularization is used to avoid over fitting, data missing is also well handled by the model and parallel computing is allowed. It is characterized by very good accuracy, necessary for regression, classification, and ranking, with the best result on tabular data.

Random Forest: It is a learning technique that combines decision trees beginning with random data and random feature subsets. It enriches accuracy by generating average of predictions in regression or by considering majority vote in classification. It avoids overfitting, performs well in noisy and imbalanced data hence can be used for classification, regression and feature selection and identifies important features.

Stacked Model: It is a combination of base models Random Forest and XGBoost and a meta-learner is set as a Logistic Regression. The base models do not interfere with each other and make their individual decisions or forecast in some cases. These predictions are then taken by Logistic Regression where the model learns how to blend them for the final prediction. This approach reduces errors since the strengths of both base models are utilized and an efficient method of combining the results from the base models is employed.

3.3. Evaluation Methodology

To assess how effectively the created models can classify the target variable uses different kind of matrices and they are as following below:

Cross validation, it also known as spinning, and it is a model validation technique that refers to the process of partitioning the data set into several subsets with the aim of using some of them in projecting the model and others are used to confirm the model's result. The usual techniques are k-fold and leave-one-out techniques of cross validation.

Confusion matrix: It is a matrix which shows the true positive, true negative, false positive and false negative of the predicted values.

Accuracy measures the overall correctness of a model by calculating the proportion of correct predictions out of the total predictions. **Precision** measures the accuracy of positive predictions. **Recall** assesses the model's ability to identify all actual positives, and the **F1 score** provides a harmonic means of precision and recall, balancing their trade-off, especially in imbalanced datasets.

AUC called Area Under the Curve is a measure of the capacity of the given classification model to segregate the classes. AUC is computed from the ROC curve and is a measure between 0 - 1, a value closer to 0 implying that the model's performance is less satisfactory while a value closer to 1 implying that the model is very satisfactory.

ROC is the graphical representation of the true positive rate (sensitivity) against the false positive rate (1-specificity) of a diagnostic test over the range of the classification measures. ROC curve is utilized in order to demonstrate the tension between sensitivity and specificity and the size of the compressed area between this curve (AUC) is utilized to quantify the model's ability to differentiate between the two classes.

Cross validation: It also known as spinning, and it is a model validation technique that refers to the process of partitioning the data set into several subsets with the aim of using some of them in projecting the model and others are used to confirm the model's result. The usual techniques are k-fold and leave-one-out techniques of cross validation.

3.4. Explanation of Methodology

Result Analysis: The confusion matrix helps to understand the reliability of the models in each class. Accuracy is the total number of correct predictions referred to all the prediction of the method thereby determining the percentage of overall correctness and has a drawback when it is used on imbalanced databases and this is why precision is an important measure when the rate of false positives needs to be kept low, accuracy reveals the ratio of the true positives among all the cases which the model considers as positive. Recall (Sensitivity) indicates the ability of models for positive classes regardless of these classes' actual frequency and probably shows the ability to reduce false negative rates. The F1 score is the harmonic average of precision and recall and specifically developed for application where the data are skewed. Additionally, AUC (Area Under the Curve) to determine the AUC for classifiably capacities of the model is closer to 1 even better. The ROC Curve (Receiver Operating Characteristic Curve) which is a graphical plot of the true positive rate against the false positive rate at different thresholds is used to analyze efficiency of techniques, the two evaluation measures used are the area under a curve above the reference diagonal and the larger the better is the model. Every above-mentioned metrics has given the detail description of the models' performances, where it was possible to understand, what were the strengths, and what could be the limitations of the corresponding models while calculating the level of the employee turnover.

Inference: The final model for real time prediction of employee attrition will be the stacked model of using multiple models like Random Forest and XGBoost with final estimator as logistic regression. When implemented on Stream lit, the application will enable the user to enter the desirable characteristics of the employee; the variables will be further processed, and the stacked model will be applied to them. The interaction will be in the form of input given by the user on the interface. The input data will be prepared for use in a similar way small portions of the input data were prepared during the model training phase like scaling, mapping. The stacked model will then make a prediction as to whether an employee is likely to turnover or remain with the organization. This result will be presented to the users through the Stream lit application interface to provide the prediction. This inference process makes sure that the model is giving insights out of raw data about the existing employee in real time to improve the decision making of the Human Resource department and employ retention strategies.

3.5. Rational for Methodology

Contribution and Justification: This research contributes to existing research in attrition prediction in terms of data preprocessing, model optimization, and suitable evaluation metric that have been identified to be lacking in most models. The use of multiple preprocessing steps guarantees a solid approach to the problem of handling difficult tasks, including missing values, duplicate records, and variability in columns, which is poorly investigated in prior literature. Doing hyperparameter tuning makes the study achieve the best model since the setting enhances accurate prediction in the real-world environment. Further, the evaluation of the model has been given more depth due to the inclusion of a set of metrics such as accuracy, precision, recall, F1 measure, AUC and ROC. This not only reveals the advantages of the model, but also further explains the assessment of the results, especially when handling imbalance data sets. Cross validation leads to improved model performance and the actualization of over fitting and under fitting which are rampant in machine learning. Moreover, the use of multiple experiments in assessing the best preprocessing and optimization strategies for the model increases its versatility and reliability because of this study. This approach fills gaps in current literature in two ways, namely in detailing the procedural and mathematical models used in the creation of these models as well as consider the organizational utilization of these models, locating this research firmly within frameworks for predicting employee attrition.

Relevance of the Dataset: There is a great relevance of the IBM HR Analytics dataset with regards to the prediction of attrition, owing to its feature sets such as demographic, job satisfaction, and organizational which are all important in determining turnover. It also comes in handy when doing feature engineering to your machine learning models and when dealing with class imbalance. Easy access and well formatted make it suitable for deriving insights. All these factors simultaneously determine the usability of the IBM HR dataset for the improvement of techniques for analyzing employee attrition.

Addressing Existing Challenges: The real-world application of employee attrition is achieved by deploying the best performed with stream lit as a web application, so that the HR professionals can be able to analyze whether an employee will attrit or not.

4. Design Specification

System Design for Employee Attrition Prediction

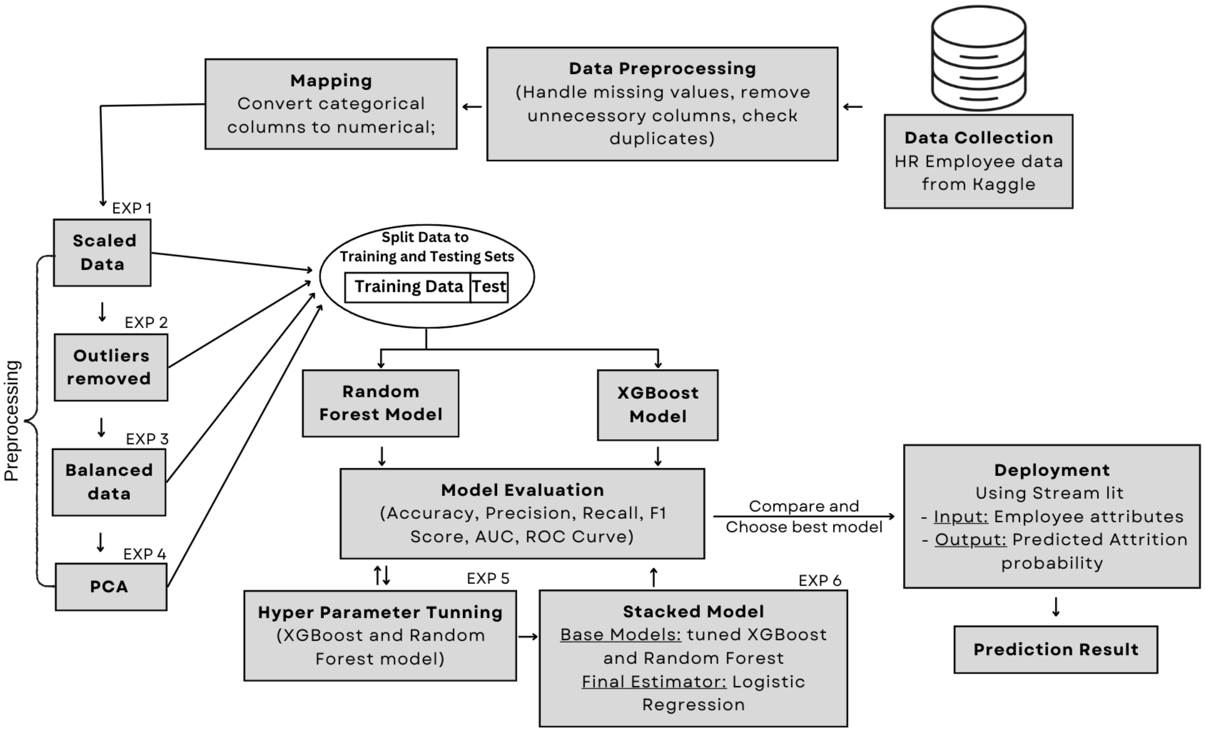
Mapping
Convert categorical columns to numerical;
Data Preprocessing
(Handle missing values, remove unnecessory columns, check duplicates)
Data Collection
HR Employee data from Kaggle
EXP 1
Scaled Data
EXP 2
Outliers removed
EXP 3
Balanced data
EXP 4
PCA
Preprocessing
Split Data to Training and Testing Sets
Training Data
Test
Random Forest Model
XGBoost Model
Model Evaluation
(Accuracy, Precision, Recall, F1 Score, AUC, ROC Curve)
Compare and Choose best model
Deployment
Using Stream lit
- Input: Employee attributes
- Output: Predicted Attrition probability
Prediction Result
EXP 5
Hyper Parameter Tunning
(XGBoost and Random Forest model)
EXP 6
Stacked Model
Base Models: tuned XGBoost and Random Forest
Final Estimator: Logistic Regression

Figure 1: Design Specification of Employee Attrition prediction

The design specification of this research followed CRISP-DM methodology, the research start form understanding the effects of employee attrition in businesses and how the early prediction of employee turnover can mitigate this issue with minimal damage to the business. After that the data related to employee turnover was collected from Kaggle and consists of information regarding the employees' demographics, their raw positions, and whether or not they are planning to leave the company. The first step was taken to simply examine structures of the data and check for missing values and additional features that should not be used in the model. Since it is an imbalanced data set, initially, Random Forest and XGboost classifiers were used on preprocessed data to set baseline for the model. These models were then retrained on the outlier free and scaled data giving a slightly better accuracy and recall value. Then the model again retrained on the balanced, outlier free and scaled data and this model is slightly more efficient than the previous. The model then retrained on the data after PCA which reduces the data dimensionality, but this diminishes the model performance. Then the trained model with balanced, outlier free and scaled data perform hyperparameter tuning by GridSearchCV for Random Forest and XGBoost to acquire optimized model settings including n_estimators, max_depth, and learning_rate this become more efficient than the other models. With an aim to get better results, primary stacking with Random Forest & XGBoost as base classifiers and Logistic Regression as the meta-model was created with the highest accuracy. The models' performance was evaluated in the test and train set with accuracy, precision, recall and F1 score along with ROC-AUC curve. And the best model is again examined whether there is any problem with over fitting and under fitting using cross validation with 5 folds and plotted it. The best models and preprocessing applicable in the future were saved using a joblib library. Then the best model was deployed on Stream lit, thus making the working web-app that took employee's attributes as input and provided the prediction of the attrition the model could be used in everyday practice. The flow chart of design specification is shown in figure1 above.

5. Implementation

The approach for this research was designed to systematically improve the accuracy of employee attrition prediction by applying advanced machine learning and ensembletechniques.

5.1. Data Preprocessing

The process began with data collection from the Kaggle dataset "HR Employee Attrition Dataset" by Jash312, which contained 13,422 rows and 39 features, with the target variable "Attrition" (binary: Yes/No).

Data Inspection: Firstly, with the dataset some checks were done to analyse the duplicate records, if the data contained missing values or not, the distribution of values in data was also examined and the count of unique values in each categorical columns were also examined.

Data Cleaning and Preprocessing: In this case some useless columns for analysis including EmployeeID, EmployeeCount, Over18, and StandardHours were eliminated. Further, any column containing greater than 50% missing data such as LeavingYear, Reason and RelievingStatus was also deleted to enhance the quality of data. In the next step data cleaning and preprocessing were performed. Censoring was performed mainly by using the Interquartile Range (IQR) to detect and remove 0.06% of the data points which occur as outliers thus avoiding the risk of a model to be skewed due to the extreme values.

Class Distribution: While categorical features were converted to quantitative using specific mapping techniques and quantitative features were scaled using StandardScaler. In the data the target variable has values like 24% of the employees having left while 76% having remained in the company as shown in figure 2 below.

- No: 10225
- Yes: 3198

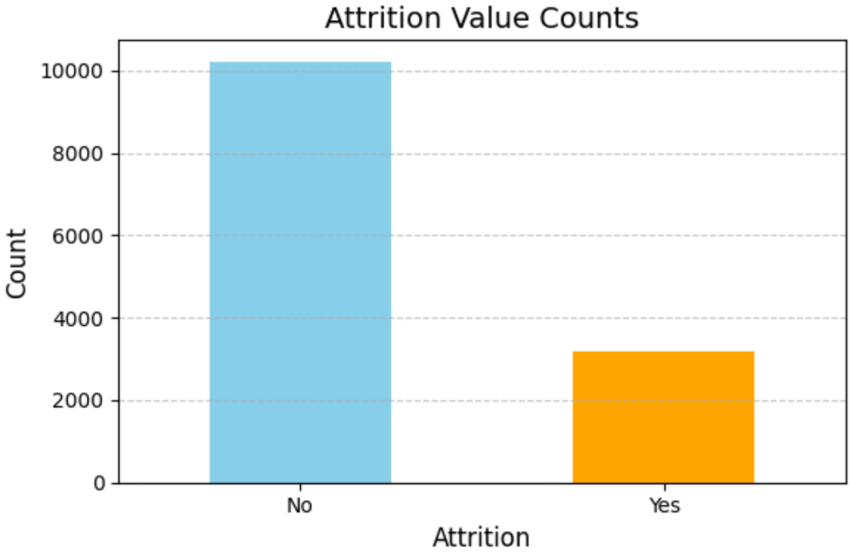

Figure 2: Target variable count-plot

Mapping: To convert the categorical columns to numerical columns the mapping technique is used which manually assigns the values to the corresponding category in each column by creating dictionary.

Scaling and Dimensionality Reduction: For enhancing the imbalanced class problem, Synthetic Minority Oversampling Technique (SMOTE) was used which increase the sizes of target classes and significant improvement was recorded particularly on the minority class (attrition = "Yes"). In addition, feature reduction was performed using Principal Component Analysis with a percentage of variance of 95%. By reducing the number of features, computation as well as the generality of the model can be enhanced.

Data Splitting: The data were split into training and testing with 80% for training and 20% for splitting.

- Training set: 10738 samples
- Testing set: 2685 samples

5.2. Modelling Experiments

The initial set of comparisons identified in the study involved baseline models that included preprocessed Random Forest and XGBoost classifiers were initially trained to set the benchmark performance against which the current model would be compared. These models were then retrained on the smaller set of non-outlier data that was scaled and there were slight incremental percent increases in the accuracy and recall. Once again, the model was retrained using the balanced data with SMOTE technique. These lead to a minor improvement in the model's performance. Then for further optimization in this high dimensional data with 35 features, this study considers PCA to reduce dimensionality of the data. Again, the models were trained using the dataset with PCA, but the model efficiency reduced. Following this hyper parameter tuning was done with GridSearchCV where n_estimators, max_depth, min_samples_split and min_samples_leaf were tuned for Random Forest while n_estimators, learning_rate, max_depth, and subsample were tuned for XGBoost which provided better configurations of the models. In the first case, the model is trained on the preprocessed data in order to set the base line for the given model. To get a better prediction, another stacked ensemble was developed using Random Forest and XGBoost as the base models while Logistic Regression as the meta-model. In this work, the models and preprocessing steps that gave the best performance were saved using the joblib library for reusability or future deployment.

5.3. Evaluation and validation

The models were evaluated using the matrices like accuracy, precision, recall, f1 score and ROC curve in the test data and the best model examined weather it has over fitting and under fitting problem using cross validation in the trained data.

5.4. Deployment

The best performing model was implemented in Streamlit which produced a web app, whereby the users entered the attributes of the employees of an organization and got the probability of their likely attrition based on predictions from the model.

6. Evaluation

This chapter aims at brief discussions regarding the models employed and results attained towards management of employee retention issues in organizations. The evaluation utilizes basic evaluation parameters which include accuracy, precision, recall and F1 score. To interpret the findings, Receiver Operating Characteristic (ROC) curves and a bar chart comparing the performance of individual measures are utilized. The following research findings are considered with regard to their practical usability in the context of further development of HR strategies and advancement to prompt innovative solutions in this area.

6.1. Results and Interpretation of the Model Experiments

Experiment 1: In baseline model evaluation, two algorithms used for this task show favorable initial performance when trained on a scaled dataset. Random Forest model balanced accuracy percentage was 98.65%, the precision, recall, and F1 scores were, 99.83%, 94.59%, and 97.14%, respectively. However, XGBoost performed slightly better attaining 99.48 % of accuracy, 99.84 % of precision, 97.99 % of recall, and 98.91 % of F1 scores. Based on these results above, both tree-based algorithms are useful in predicting employee attrition, although XGBoost performs slightly better than the Random Forest algorithm. This indicates that XGBoost models are more resilient towards structured data and offer better predictive reliability.

Experiment 2: The elimination of outliers based on the IQR method resulted in removal of 0.06% of the total data which lead to increase in model performance slightly. Consequently, it was found that the Random Forest model has achieved an accuracy rate of 98.99%, Precision, recall, and F1 score of 99.18%, 96.49 %, and 97.81% respectively and the XGBoost model attained an accuracy of 99.44%, along with precision, recall, and F1 scores of 99.83%, 97.77%, and 98.79% respectively. These new and higher recall scores suggest that the removal of outliers increased the generalizability of the models to be less prone to overfitting and hence increased their predictive reliability.

Experiment 3: Using SMOTE algorithm, it was beneficial to address the problem of imbalance in classes hence the high recall in results produced was associated to high precision. The Random Forest model has an accuracy of 98.77% with precision of 99.67 %, recall of 95.20 and the F1 Score of 97.39%. In the case of the XGBoost model we had an accuracy of 99.52%, precision of 99.84%, recall of 98.14%, and an F1 score of 98.99%. Based on these findings, SMOTE ensures the boost of the model's ability in reducing false negatives with high impact especially in cases of HR where failure to predict potential trends of attrition increases the chances of unpredictable turnover and all that comes with it.

Experiment 4: For data reduction, Principal Component Analysis was used to retain 95% of variability in terms of feature dimensions and this led to a visible degradation of the model performance. The model Random Forest achieved an accuracy of 93.29%, precision of 93.64%, recall of 77.43% and F1 score of 84.77%. Similarly, for the XGBoost model accuracy of 95.05%, precision of 95.41%, recall of 83.46% and F1 score of 89.04%. The reduction in recall clearly indicates loss of a predictive potential, which may show that PCA cannot be used in the case of datasets, where features are powerful and informative and contribute a lot to accuracy.

Experiment 5: Hyper-parameter tuning was also performed to make for the best values of Random Forest and XBoost models and the best parameters found for Random Forest were a maximum depth of 20, n estimators set to 150 and setting both minimum samples for split 2 and minimum samples for leaf nodes to 1. This tuned model, an accuracy of 98.65% was obtained as well as a high value in precision, recall, and F1 of 100%, 94.43%, and 97.13%, correspondingly. For XGBoost the parameters were tuned to learning rate = 0.1, max.depth = 5, n estimators = 150 and subsample = 1. The final tuned XGBoost model was even better accuracy, which was 99.48%, while the precision was 100%, recall was 97.83%, F1 Scores was 98.91%. These results prove the efficiency of hyperparameter tuning in enhancing the model performance, then achieving a better scale between precision and recall, and enhancing the models for dealing with the complexity of the dataset.

Experiment 6: The tuned Random Forest model and the tuned XGBoost model were stacked using the meta leaner Logistic regression. This model exhibits a better performance compared to the other models with 99.59% accuracy, precision 100%, the recall 98.29%, and F1 score 99.14%. This stacked model combines the capabilities of both Random Forest and the

XGBoost models which made the prediction more effective. So, this evaluation clarifies that the ensemble method is an effective method to use for the enhancement of interpretability of model. For easy comparison the compared bar-chart of every experimented model is shown in figure 3 below.

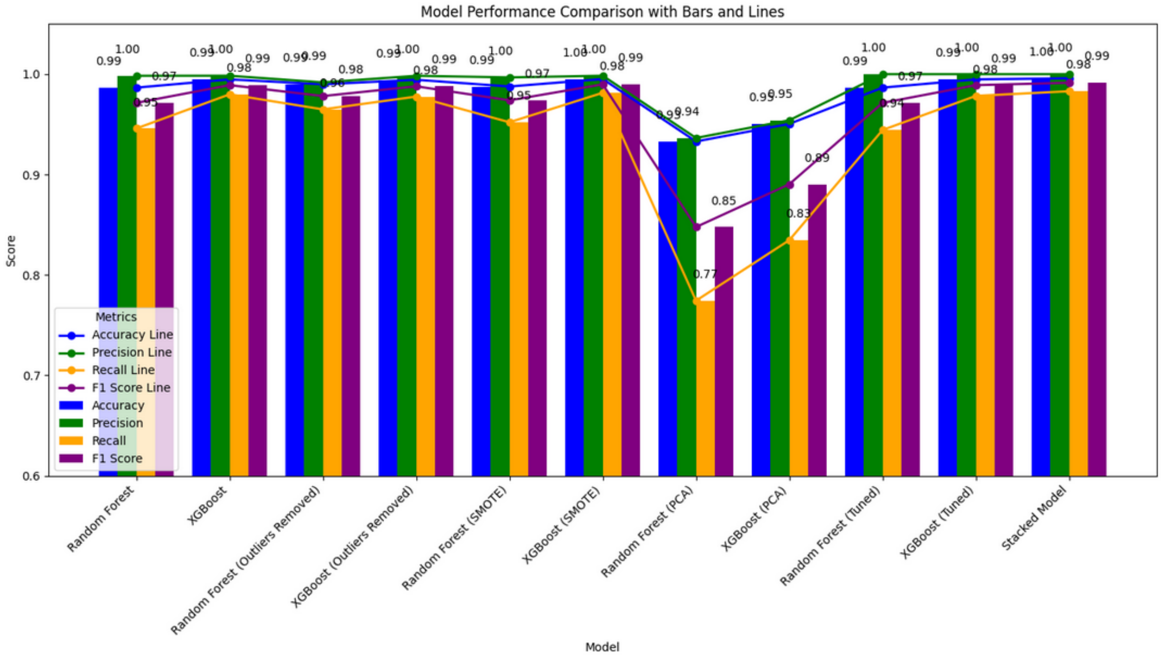

Figure 3: Comparison chart of all models

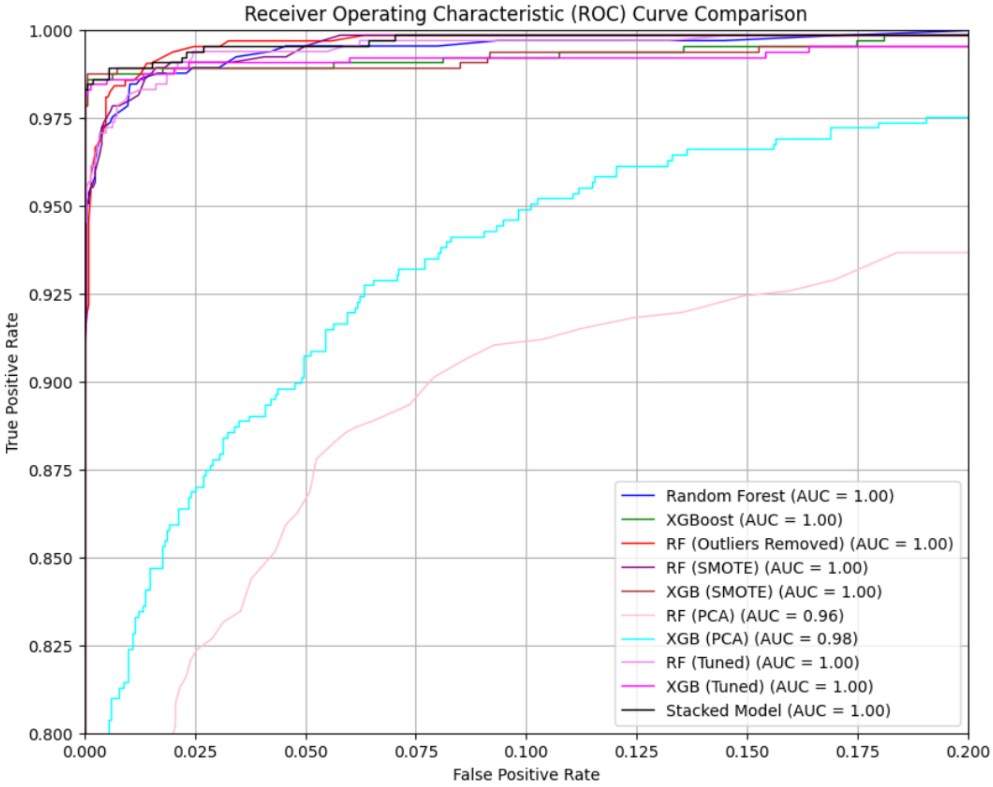

Figure 4: ROC curve of all models

ROC curve: Figure 4 above shows the ROC curve of all models, and it indicates that all the models presented an AUC very close to 1, which suggests a very good discriminant capacity of the models. More specifically, the stacked model depicted the highest and most consistent location in the top left corner of the curve thus suggesting a higher capability in classes identification. Further, models that employed SMOTE for class balancing and eliminating outlier had enhanced recall represented in terms of higher true positive rate. This indicates that these techniques will have improved the models' propensity of predicting more good examples which in turn improved their overall predictive ability.

Cross validation Plot: The cross-validation graph in figure 5 below demonstrates the stacked model's stability in the learning curve and confirms that it generalizes well. The measure of accuracy for training is always close to 1.0 and does not seem to vary with the size of the training set while the accuracy for validation increases as the size of the training set increases and is also close to the accuracy for training. This shows that the model is not overtrained, and it is also not undertrained, that is, the model is perfect. Further, cross-validation scores including values ranging from 99.44% to 99.53% with minimal variation prove that model performs equally well in different validation folds. The cross-validation accuracy mean level also supports the model's acquisition with figures reaching 99.51% accuracy. Altogether the results attained confirm that the presented stacked model is well-trained and ready to provide high accuracy in predicting the performance on the unused data, which will make it good for practical use such as employee turnover prediction.

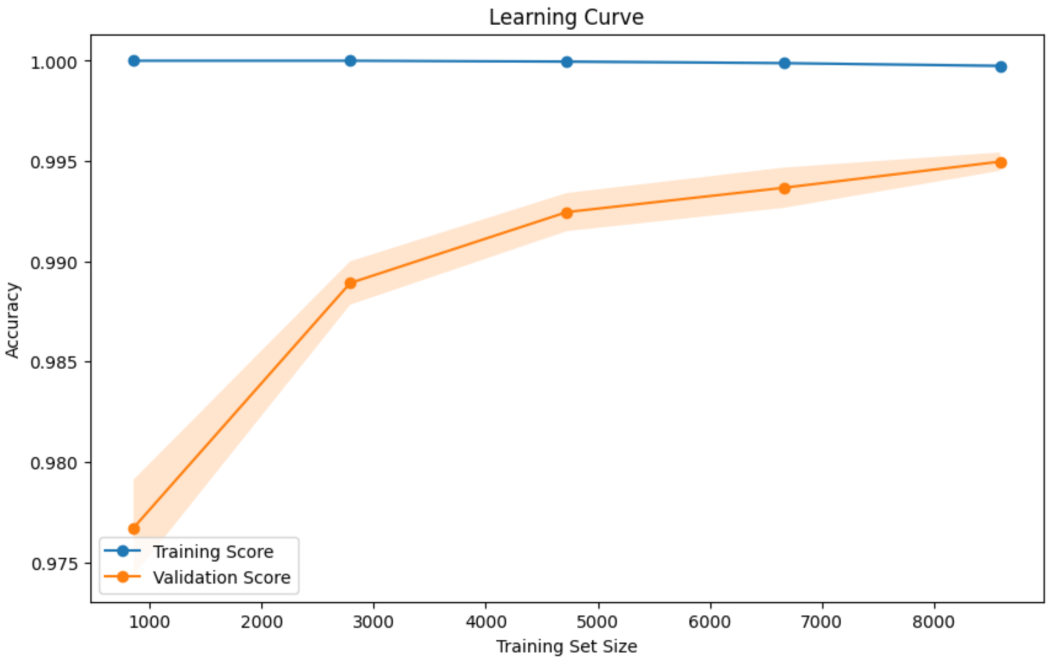

Figure 5: Cross validation of Stacked model

6.2. Discussion

This research shows that the integration of XGBoost and Random Forest with SMOTE, feature selection, best hyperparameters tuning, and the ensemble learning approach can achieve high levels of employment attrition prediction accuracy and acceptable levels of recall. It also demonstrates that certain preprocessing techniques such as outliers' removal, SMOTE and feature scaling can greatly improve models' stability. The stating model is highly effective at helping HR specialists pinpoint employees at risk to implement retention measures. If organizations can gain insight into the factors that may lead to attrition, then they can allocate their resources properly, and spend less money on employees who are likely to leave and experience less disruption within the workforce. These results directly answer the research question of how increased prediction precision helps HR mitigate the workforce attrition problem. In addition, the achieved accuracy of the stacked model is 99.59%, precision 100%, recall 98.30% and f1 score 99.14% demonstrate that the proposed approach is an effective and valuable solution, which is better than preceding methods.

7. Conclusion and Future Work

In conclusion, the experimental approach of building the best model for predicting employee attrition developed stacked model with tuned XGBoost model and tuned random forest model as the base model along Final Estimator as Logistic Regression. It is able predict employee attrition with accuracy of 99.59%, precision of 100%, recall of 98.30% and the F1 score of 99.14%. Both outlier removal, balancing data, hyper parameter tuning and feature scaling help each step of modelling and enhance its performance. The generalization & reliability of this ensemble model is that the final stacked model has the ability and strength of XGBoost as well as Random Forest model. With the development of such advanced methods of predicting attrition of employees, organizations can pay extra attention to the work force and their skills, and effectively handle them, and therefore take appropriate action to employees' concerns and elevate the level of employee satisfaction. This results in less resources being expended by the companies in terms of new recruitment and onboarding costs and also the company is kept low in the disruption of the projects and continuity. While it has some limitations. The dataset used on this study is from IBM HR analytics data, restraining the potential of the study upon other industries and work force dynamics. Furthermore, the data source may be misleading as they are not depicting the richness and diversity of real-life environments such as organizational culture, expertise and business organization. Also, the ensemble models and hyperparameters can be too computationally complex for the needed models and their implementation on large datasets of the real problems.

Future work involves training and testing the models with other datasets that are related to the industry and contain more real-world data to be used to increase the model's generalizability. The enhancements will continue improving prediction by using innovative data imputation method and bias elimination methods for addressing data bias and quality. There are several optimization techniques, including a major technique called weighted ensemble learning as well as transfer learning as these will enhance the predictive abilities of the model as well as adaptability of the models in future prediction. By doing so, this research shows that machine learning has the potential to radically change how human resources organize their activities and the value of predictively leveraging work force analytics for strategic retention planning.

References

Ajit, P., 2016. Prediction of employee turnover in organizations using machine learning algorithms. In 2016, International Journal of Advanced Research in Artificial Intelligence (IJARAI), 4(5), p.C5.

Al-Suraihi, W.A., Samikon, S.A., Al-Suraihi, A.H.A. and Ibrahim, I., 2021. Employee turnover: Causes, importance and retention strategies. *European Journal of Business and Management Research*, *6*(3), pp.1-10.

Alduayj, S.S. and Rajpoot, K., 2018, November. Predicting employee attrition using machine learning. In *2018 international conference on innovations in information technology (iit)* (pp. 93-98). IEEE.

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

Chakraborty, R., Mridha, K., Shaw, R.N. and Ghosh, A., 2021, September. Study and prediction analysis of the employee turnover using machine learning approaches. In *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)* (pp. 1-6). IEEE.

Chung, D., Yun, J., Lee, J. and Jeon, Y., 2023. Predictive model of employee attrition based on stacking ensemble learning. *Expert Systems with Applications*, *215*, p.119364.

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

George, S., Lakshmi, K.A. and Thomas, K.T., 2022, December. Predicting Employee Attrition Using Machine Learning Algorithms. In *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)* (pp. 700-705). IEEE.

Guerranti, F. and Dimitri, G.M., 2022. A comparison of machine learning approaches for predicting employee attrition. *Applied Sciences*, *13*(1), p.267.

Hossen, M.A., Hossain, E., Ishwar, A.K.Z. and Siddika, F., 2021, February. Ensemble method based architecture using random forest importance to predict employee's turn over. In *Journal of Physics: Conference Series* (Vol. 1755, No. 1, p. 012039). IOP Publishing.

Jain, R. and Nayyar, A., 2018, November. Predicting employee attrition using xgboost machine learning approach. In 2018 international conference on system modeling & advancement in research trends (smart) (pp. 113-120). IEEE.

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

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

Muslim, M.A. and Dasril, Y., 2021. Company bankruptcy prediction framework based on the most influential features using XGBoost and stacking ensemble learning. *International Journal of Electrical and Computer Engineering (IJECE)*, *11*(6), pp.5549-5557.

Nandal, M., Grover, V., Sahu, D. and Dogra, M., 2024. Employee Attrition: Analysis of Data Driven Models. *EAI Endorsed Transactions on Internet of Things*, *10*.

Ozdemir, F., Coskun, M., Gezer, C. and Gungor, V.C., 2020, May. Assessing employee attrition using classifications algorithms. In *Proceedings of the 2020 the 4th international conference on information system and data mining* (pp. 118-122).

Pan, F., Yuan, Y. and Song, Y., 2020, April. Students' Classification Model Based on Stacking Algorithm. In *Journal of Physics: Conference Series* (Vol. 1486, No. 3, p. 032020). IOP Publishing.

Qutub, A., Al-Mehmadi, A., Al-Hssan, M., Aljohani, R. and Alghamdi, H.S., 2021. Prediction of employee attrition using machine learning and ensemble methods. *Int. J. Mach. Learn. Comput*, *11*(2), pp.110-114.

Shaik, S., Kumar, P.S., Reddy, S.V., Reddy, K. and Bhutada, S., 2023. Machine learning based employee attrition predicting. *Asian Journal of Research in Computer Science*, *15*(3), pp.34-39.

Tharani, S.M. and Raj, S.V., 2020, October. Predicting employee turnover intention in IT&ITeS industry using machine learning algorithms. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 508-513). IEEE.

Yang, S. and Islam, M.T., 2020. IBM employee attrition analysis. *arXiv preprint arXiv:2012.01286*.