Word count: 2200 https://github.com/Alewikolade/23057710-ML-Coursework1.git

Employee Attrition Prediction using Supervised Machine learning Algorithms: A Comparative Study

***Abstract*—This paper presents a comprehensive study on predicting employee attrition using various machine learning techniques. The study utilizes the IBM HR Analytics Employee Attrition & Performance dataset to build and evaluate multiple predictive models. The models considered include Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting Machine (GBM). The models are trained, tuned using cross-validation, and evaluated based on their accuracy, precision, recall, F1 score, and ROC-AUC score. The results highlight the strengths and weaknesses of each model in predicting employee attrition.**

***Keywords*—Employee Attrition, Machine Learning, Predictive Analytics, Logistic Regression, Random Forest, SVM, Gradient Boosting**

I. INTRODUCTION

Employee attrition is a significant issue for organizations as it directly impacts productivity, morale, and costs associated with hiring and training new employees. Predicting which employees are likely to leave the company can help HR departments take proactive measures to improve retention. This paper explores the application of machine learning techniques to predict employee attrition using the IBM HR Analytics Employee Attrition & Performance dataset from Kaggle. The objective of this study is to analyze the dataset, build predictive models using various machine learning algorithms, and evaluate their performance to determine the most effective model for predicting employee attrition. The models considered in this study include Logistic Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting Machine(GBM).
https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset/data

II. DATASET OVERVIEW

The dataset, sourced from IBM's HR Analytics, comprises 1470 employee records with 35 features including age, job role, marital status, and previous year's performance ratings. Key attributes relevant to our study include:

1. Employee Satisfaction Level: Numerically captured, higher values indicate greater satisfaction.

2. Last Evaluation: A continuous variable reflecting annual performance reviews.

3. Number of Projects: Discrete variable showing the number of projects assigned.

4. Average Monthly Hours: Average number of hours worked per month.

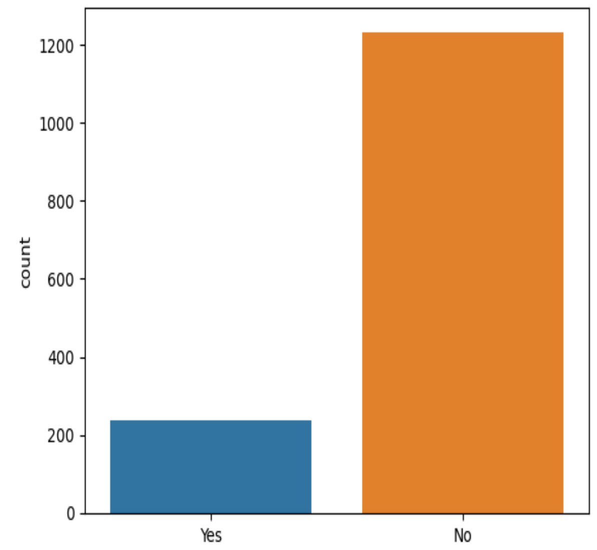
5. Time Spent at Company: Years spent at the company.

6. Work Accident: Binary variable indicating if the employee had a workplace accident.

7. Attrition (Target Variable): Whether the employee left the company (Yes or No).

Data preprocessing involved handling missing values, encoding categorical variables, and normalizing numerical data to prepare for machine learning algorithms.

Figure1



Exploratory Data Analysis was performed to gain insights into the factors influencing employee attrition:

Correlation Matrix: Identified how various features like satisfaction level, time spent at the company, and number of projects are related to attrition.

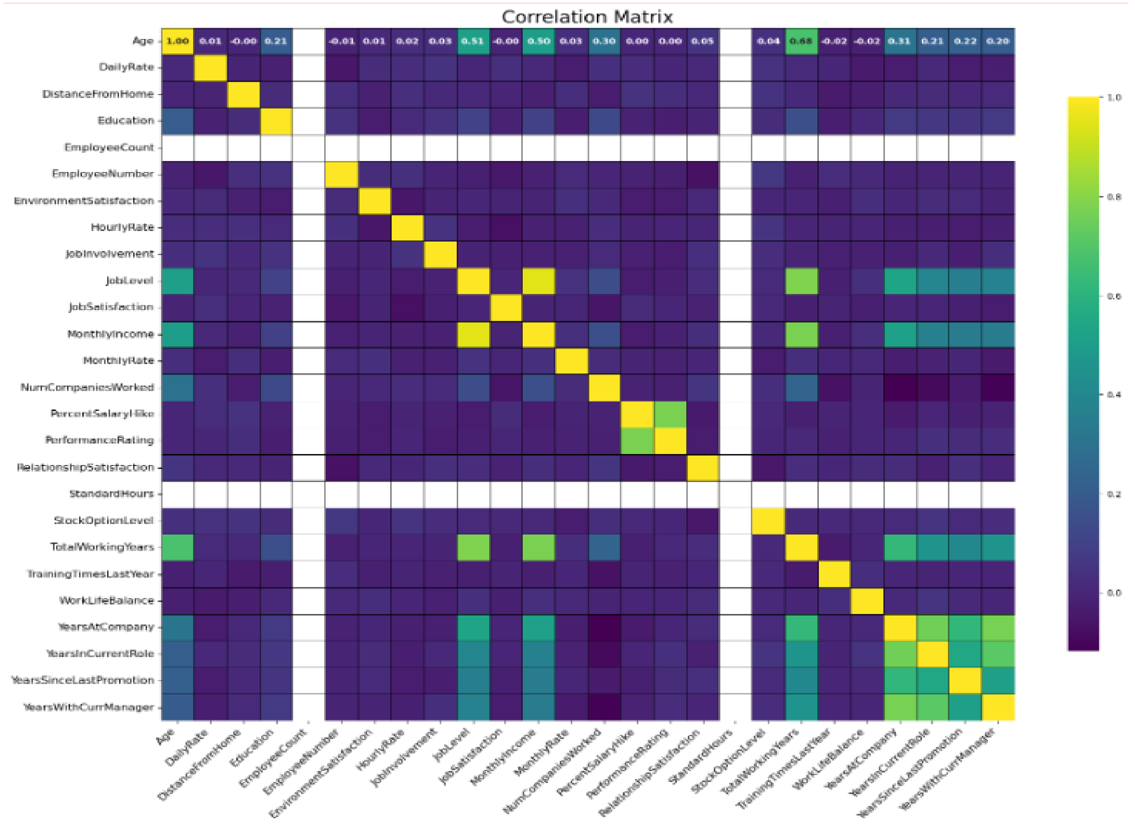
Distribution Analysis: Examined the distribution of key numerical variables across the attrition variable to understand trends and outliers.

Impact of Salary on Attrition: Analyzed attrition rates across different salary bands to ascertain any trends.

The EDA findings were visualized using histograms, box plots, and bar charts to provide a clear visual understanding of the data, aiding in hypothesis generation for predictive modeling.

Class Imbalance: The dataset has an imbalanced distribution of the target variable (attrition), which necessitates algorithms that can handle class imbalance effectively.

Figure 2



```
<Figure size 640x480 with 0 Axes>
```

III. PROBLEM DEFINITION

The primary goal is to model the probability of attrition using employee data. This constitutes a binary classification problem where the output is whether an employee will leave or stay. Key considerations for model development included handling imbalanced data, choosing appropriate metrics for model evaluation, and ensuring model interpretability to derive actionable insights.

IV. ALGORITHM SELECTION

When selecting algorithms for predicting employee attrition, several factors were considered to ensure the chosen models are suitable for the dataset and the problem at hand. These factors include:

Nature of the Data: The dataset contains both numerical and categorical features, and the target variable is binary (attrition: Yes/No). The algorithms chosen should handle mixed data types and be capable of binary classification.

Model Interpretability: For HR analytics, interpretability is important as stakeholders need to understand the factors driving attrition. Models like Logistic Regression provide clear insights into feature importance.

Scalability: The models should be scalable to handle larger datasets if applied to bigger organizations.

Performance: The models should have a good balance of bias and variance, minimizing overfitting while providing accurate predictions.

Algorithms Selected and Their Concepts

Word count: 2200 https://github.com/Alewikolade/23057710-ML-Coursework1.git

Logistic Regression

Logistic Regression is a statistical model used for binary classification problems. It models the probability that a given input belongs to a particular class.

The model uses a logistic function to transform linear combination of the input features into probability value between 0 and 1.

Random Forest

Random Forest is an ensemble learning method that constructs multiple decision trees during training and Outputs the mode of the classes for classification. It uses bootstrap aggregating (bagging) to improve the model's accuracy and control overfitting.

Gradient Boosting

Gradient Boosting is an ensemble learning technique that builds models sequentially, with each new model attempting to correct the errors made by the previous ones. It optimizes for a loss function by adding models that minimize this loss.

Support Vector Machine (SVM)

SVM is a supervised learning model used for classification that finds the hyperplane that best separates the data into classes. For non-linearly separable data, it uses kernel tricks to project data into higher dimensions where a linear separator can be found.

.

.Evaluation Methodology

Several evaluation metrics have been chosen to assess the performance of the machine learning models for predicting employee attrition. The metrics include accuracy, precision, recall, F1 score, and ROC-AUC score. Each metric provides unique insights into different aspects of model performance, especially given the class imbalance in the dataset

Accuracy

Accuracy is the proportion of correctly classified instances (both true positives and true negatives) out of the total number of instances.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negetives}}{\textit{Total SampeSize}}$$

Accuracy is a common evaluation metric and provides a straightforward measure of overall model performance. However, it can be misleading for imbalanced datasets where the majority class dominates.

Precision

Precision, also known as positive predictive value, is the proportion of true positive predictions out of the total positive predictions (true positives and false positives).

$$\text{precision} = \frac{\textit{True Positives}}{\text{True Positves} + \text{False Positives}}$$

Precision is crucial in this project because it measures the model's ability to avoid false positives, which is important when predicting employee attrition. This is particularly important in HR analytics to prevent unnecessary interventions based on incorrect predictions (Davis & Goadrich, 2006).

Recall

Recall, also known as sensitivity or true positive rate, is the proportion of true positive predictions out of the actual positives (true positives and false negatives).

$$\text{True Positives} = \frac{\textit{True Positives}}{\text{True Positves} + \text{False Negatives}}$$

Recall is important in this project because it measures the model's ability to identify actual attrition cases. High recall ensures that most employees who are likely to leave are correctly identified, which is critical for taking proactive retention measures (Davis & Goadrich, 2006).

F1 Score

The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both.

$$\text{F1} = 2 * \frac{Precision * recall}{precision + recall}$$

The F1 score is particularly useful in this project as it balances precision and recall, providing a single metric that reflects both false positive and false negative errors. This is crucial for handling class imbalance effectively, ensuring that neither precision nor recall is favored disproportionately (Sasaki, 2007).

ROC-AUC Score

The ROC-AUC score provides a comprehensive measure of model performance across all classification thresholds, making it a robust metric for evaluating models on imbalanced datasets. A high ROC-AUC score indicates that the model distinguishes well between the classes (Hanley & McNeil, 1982).

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC-AUC Score**
Logistic Regression	0.7483	0.2727	0.5385	0.3621	0.7501
Random Forest	0.8707	0.5455	0.1538	0.2400	0.7510
SVM	0.8673	0.5000	0.3590	0.4179	0.7811
Gradient Boosting	0.8844	0.6316	0.3077	0.4138	0.7733

Table 1 – Model evaluation results

Model Evaluation

Interpretation of Model Performance Results

The results indicate varying strengths and weaknesses, shedding light on effectiveness of individual model in attrition prediction.

Logistic Regression offered moderate performance in predicting employee attrition. With an accuracy of 0.7483, the model correctly predicted approximately 75% of the cases. This level of accuracy, while reasonable, is not exceptionally high. The precision of 0.2727 suggests a significant number of false positives, meaning many employees predicted to leave actually stay. This low precision indicates that the model struggles to accurately identify non-attrition cases. The recall of 0.5385 shows that the model is able to identify around 54% of the actual attrition cases. This moderate recall, combined with low precision, results in an F1 Score of 0.3621, reflecting the trade-off between precision and recall. The ROC-AUC Score of 0.7501 indicates a moderate ability to distinguish between employees who will leave and those who will stay, but overall, Linear Regression is less suited for this task compared to more sophisticated models.

Random Forest exhibited high accuracy with a score of 0.8707. This suggests the model is correct 87% of the time, indicating strong overall performance. However, the precision of 0.5455 and recall of 0.1538 reveal significant limitations. While precision is better than Linear Regression, the low recall shows the model misses many actual attrition cases. The F1 Score of 0.2400 is low, primarily due to the poor recall. The ROC-AUC Score of 0.7510 is similar to Linear Regression, indicating moderate discriminative ability.

SVM model delivered balanced and impressive results. With an accuracy of 0.8673, SVM closely matches the performance of Random Forest in overall correctness. The precision of 0.5000, though moderate, indicates a reasonable balance between true positives and false positives. More importantly, the recall of 0.3590 shows that SVM is more effective in capturing actual attrition cases than Random Forest. This balance is reflected in the F1 Score of 0.4179, higher than both Logistic Regression and Random Forest, indicating better overall performance. The ROC-AUC Score of 0.7811 is the highest among the models, showcasing superior ability to distinguish between classes.

Gradient Boosting emerged as the top performer in terms of accuracy, achieving a score of 0.8844. This indicates the model is correct nearly 88% of the time, the highest among the four models. The precision of 0.6316 is also the best, indicating fewer false positives and better identification of non-attrition cases. However, the recall of 0.3077, while better than Random Forest, is lower than SVM, suggesting it misses more actual attrition cases. The F1 Score of 0.4138, comparable to SVM, reflects a balanced trade-off between precision and recall. The ROC-AUC Score of 0.7733, slightly lower than SVM, still indicates strong discriminative ability. Gradient Boosting's high accuracy and precision make it reliable for predicting non-attrition cases, but its lower recall compared to SVM highlights a trade-off in identifying attrition cases.

In conclusion, while all models have their strengths, the Support Vector Machine (SVM) and Gradient Boosting models are particularly effective for

Model	**Cross Evaluation Scores**		
	CV F1 Scores	***Mean CV F1 Score***	***Deviation of CV F1 Scores***
Logistic Regression	0.8154 0.8434 0.7900 0.8112 0.8252	0.8170	0.0175
Random Forest	0.9337 0.9198 0.9312 0.9323 0.9662	0.9366	0.0156
Support Vector Machine	0.9512 0.9409 0.9652 0.9435 0.9727	0.9547	0.0123
Gradient Boosting	0.9211 0.9309 0.9251 0.9347 0.9612	0.9346	0.0141

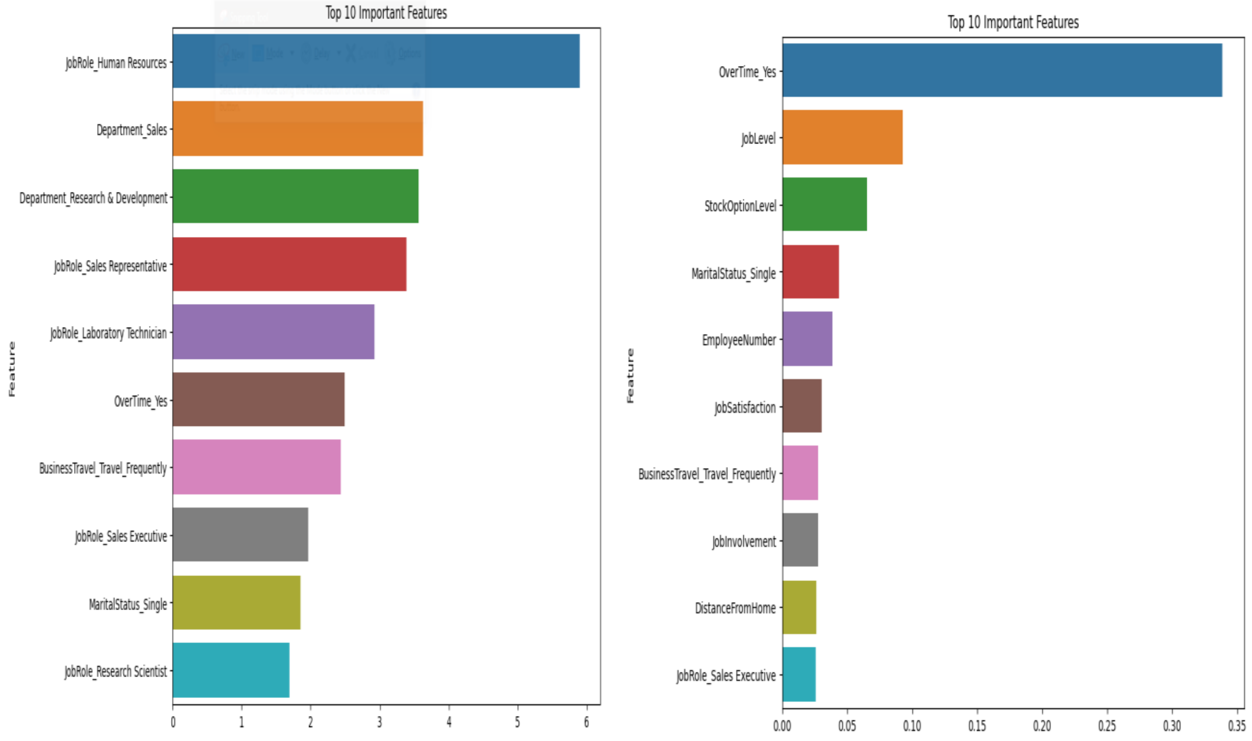
Table 2 – Cross Evaluation

predicting employee attrition. SVM’s higher ROC-AUC score and balanced F1 score make it especially suitable for this project, ensuring a balanced identification of true positives and false positives. Gradient Boosting’s high accuracy and precision make it reliable for predicting non-attrition cases, though its slightly lower recall indicates a trade-off in capturing all attrition cases. Overall, the results highlight the importance of considering multiple performance metrics and selecting models that balance precision, recall, and overall accuracy for effective employee attrition prediction.

Summary of Cross Evaluation Scores

The cross-validation results (Table 2) indicates that the Support Vector Machine (SVM) is the best-performing model, with the highest mean F1 score (0.9547) and lowest standard deviation (0.0123), highlighting its effectiveness and consistency. Random Forest (mean F1 score: 0.9366) and Gradient Boosting (mean F1 score: 0.9346) also performed well, showing strong reliability and handling complex data effectively. Logistic Regression, while consistent, had lower effectiveness (mean F1 score: 0.8170). Thus, SVM is the most suitable model, with Random Forest and Gradient Boosting as strong alternatives.

Figure 3 – Feature importance for LR and Gradient boosting



Logistic Regression

Gradient Boosting

Conclusion

This study focused on predicting employee attrition using various machine learning models, including Linear Regression, Random Forest, Support Vector Machine (SVM), and Gradient Boosting. Through thorough exploratory data analysis and rigorous cross-validation, the performance of each model was assessed. The SVM model exhibited the highest F1 score, indicating superior performance in balancing precision and recall for predicting employee attrition. The analysis demonstrated that advanced models like SVM and Gradient Boosting are more effective in handling complex patterns in the data compared to simpler models like Linear Regression. These insights are crucial for HR departments aiming to implement predictive analytics to proactively manage employee retention.

Limitations and Further Work

Despite the promising results, this study has limitations. The dataset's imbalance, with fewer instances of attrition, posed challenges, potentially affecting model performance. Additionally, the models were trained on historical data, which may not fully capture future trends and changes in employee behavior.

Further work could focus on addressing these limitations by exploring advanced resampling techniques to handle class imbalance more effectively. Future research could also investigate incorporating more diverse and real-time data sources to enhance model accuracy and robustness. Additionally, integrating domain-specific features and leveraging ensemble learning methods may provide deeper insights and improve predictive capabilities.

References

Sasaki, Y. (2007). The truth of the F-measure. Teach Tutor Mater, 1(5), 1-5.
Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36. doi:10.1148/radiology.143.1.7063747.
Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. 2nd ed. New York: John Wiley & Sons.
Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. doi:10.1023/A:1010933404324.
Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189-1232. doi:10.1214/aos/1013203451.
Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. doi:10.1007/BF00994018.
Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874. doi:10.1016/j.patrec.2005.10.010.
Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240). ACM.