Analysing Employee Attrition using Machine Learning Models

Malavika Venkatesh

1 Introduction

Employee attrition poses a significant concern for organizations, affecting productivity, operational efficiency, and overall business performance. High turnover rates not only increase recruitment and training costs but also causes work flow disruptions and loss of expertise, making it vital for businesses to identify and address the underlying factors contributing to employees' departure. This analysis focuses on predicting employee attrition at IBM using machine learning techniques, to perform binary classification using models like Logistic Regression and Random Forest. The dataset contains features like age, job satisfaction, monthly income, and environmental satisfaction, which are analysed to identify key predictors of attrition.The insights derived from this study aims to enable firms to develop effective employee retention strategies, minimize employee churn, and optimize human resource management. By addressing attrition proactively, companies can foster a more engaged workforce and enhance their long-term business success.[1]

2 Dataset Overview

The dataset used for this analysis consists of 35 columns, each containing key features that describe the characteristics of employees, with a focus on factors that might influence attrition. These features include demographic information such as age, education, and gender, along with job-related factors such as overtime, job satisfaction, percentage salary hike, work-life balance and others.The dataset contains 1,470 non-null records, ensuring complete data without any missing values or duplicates.[2]

Categorical variables were encoded as a part of feature engineering. The target variable, Attrition, along with Gender and Overtime, was encoded using binary encoding, whereas JobRole, Marital Status, Department were encoded using one-hot encoding. Furthermore, non informative columns such as EmployeeCount, Over18 and EmployeeNumber were dropped from the dataset.

The dataset was checked for any outliers by constructing boxplots for the features and since there were minor outliers they were ignored.Multiple histograms were constructed to check for skewness in data and observe the general distribution of each feature. Correlation heat-map was constructed to analyse how each of these features affect the target variable.

3 Machine Learning Techniques

The machine learning process involved applying Random Forest for feature selection due to its ability to handle complex relationships and identify important predictors of employee attrition. This method ranks, features based on their importance, which helped efficiently select the most relevant variables.

To address the class imbalance in the dataset, where 83.9% of the records did not show attrition and only 16.1% indicated attrition, Synthetic Minority Over-sampling Technique (SMOTE) was applied. SMOTE generated synthetic samples of the minority class (attrition), rather than simply duplicating, which helped in balancing the dataset and improved the model's ability to learn patterns related to the less frequent class. This technique is crucial in scenarios with imbalanced data, ensuring that the model does not become biased towards predicting the majority class.[3]

Logistic Regression was employed as the initial model to establish a baseline performance for predicting employee attrition. As a linear classifier, it estimates the relationship between the dependent variable (attrition) and the independent variables by calculating the odds of attrition given the features. This model works well in cases where the relationship between the features and the target is approximately linear, and it outputs probabilities that can be interpreted directly, offering valuable insights into the influence of each feature on attrition.[4]

Random Forest, an ensemble learning method, was applied next, to capture more complex patterns within the data. It constructs multiple decision trees and combines their outputs to produce a more stable and accurate prediction. Each tree in the forest makes an independent prediction, and the final result is determined by taking a majority vote. This model is particularly effective for handling a variety of feature interactions and non-linear relationships, making it highly suitable for datasets like this one with multiple interrelated features influencing employee attrition.

4 Key Findings and Analysis

Feature importance analysis revealed that financial and work-life factors are the most critical drivers of employee attrition. Specifically, Monthly Income emerged as the most significant feature, emphasizing the strong link between compensation and employee retention. Other influential factors, such as Age, Daily Rate, and OverTime, suggest that demographic variables and workload also play substantial roles. The inclusion of Distance From Home and Years at Company highlights the importance of logistical and career progression considerations, offering actionable insights for organizations aiming to reduce attrition rates.

In terms of model performance, the Logistic Regression model provided a baseline F1 score of 0.72, while the Random Forest model significantly outperformed it with an F1 score of 0.88, demonstrating stronger predictive power. The model also showed stability with a cross-validation mean score of 0.87, confirming it generalizes well across different data subsets. The 80:20 data split and the consistency between cross-validation and test set results indicate that the Random Forest model is not overfitting. The confusion matrix highlighted that the model accurately classified a high proportion of both attrition and non-attrition cases, ensuring reliable predictions for real-world applications. The AUC and PR AUC scores of 0.95 further highlight the model's robust and reliable predictions for employee attrition.

5 Challenges and Limitations

One key challenge in this analysis was the inefficiency of Lasso L1 regularization for feature selection, particularly when handling highly correlated features. Lasso tends to arbitrarily select one feature from a correlated group, potentially overlooking key relationships. To address this, Random Forest was used for feature selection, as it better handles correlated features and non-linear relationships. Secondly, Random Forest as an ML model can be computationally expensive, particularly with larger datasets, resulting in longer training times. Gradient Boosting may offer a more efficient solution by providing faster training times while maintaining predictive power superior to both Logistic Regression and Random Forest.[5]

Another challenge was the imbalance between the attrition and non-attrition classes. To address this, SMOTE was used to balance the dataset, but it is essential to monitor its effect on overfitting, as synthetic data can introduce noise if not tuned properly. Cross-validation was employed to ensure the model's stability across different subsets, helping to confirm that the model did not overfit the resampled data. Given the class imbalance, accuracy was not a reliable measure, and instead, the confusion matrix, F1 score and Precision-Recall curve was prioritized, providing a more meaningful evaluation of the model's performance on both classes.

6 Conclusion

This analysis identified key predictors of employee attrition, with financial and work-life factors being most significant. Random Forest outperformed Logistic Regression, as shown by higher F1 scores and better generalization after hyperparameter tuning using GridSearchCV. Despite challenges like feature selection and class imbalance, the model accurately predicted attrition, offering valuable insights for improving retention strategies. These findings highlight the importance of targeted interventions to reduce turnover and enhance organizational stability.

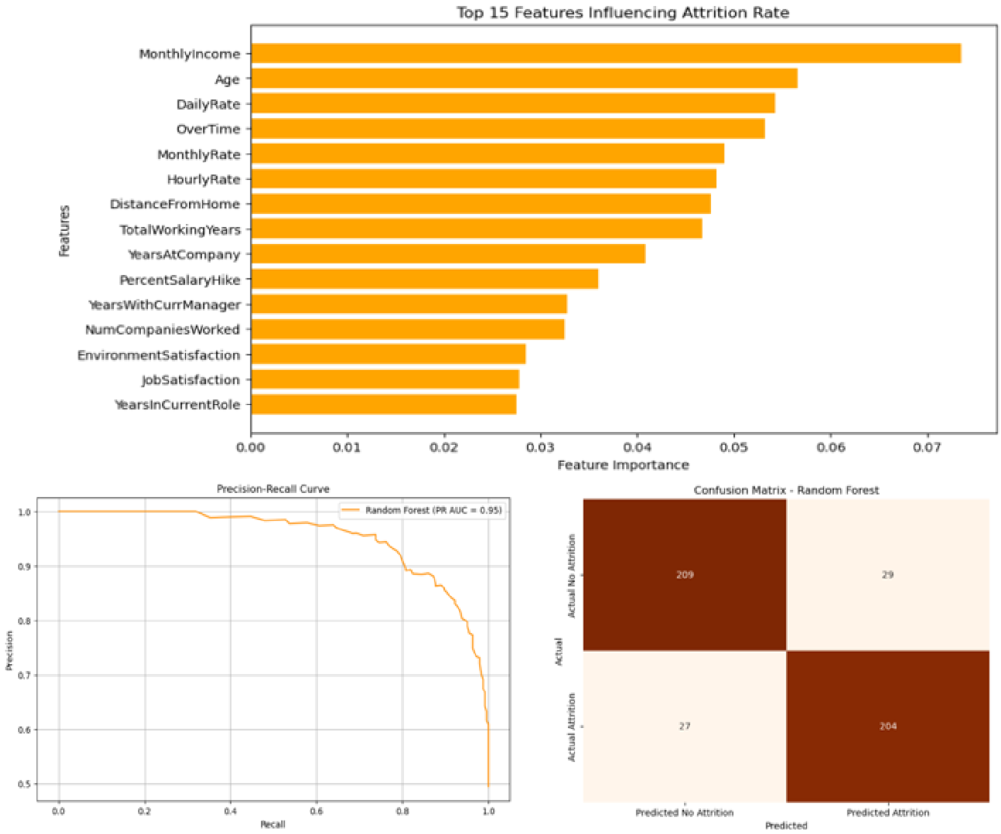

Figure 1: The bar chart (top) highlights the top 15 features influencing attrition, while the precision-recall curve (bottom left) evaluates the model's performance (PR AUC = 0.95). The confusion matrix (bottom right) provides detailed classification results, showcasing the model's ability to predict attrition and retention accurately.

References

[1] Fatbardha Maloku and Besnik Maloku. "Analyzing IBM HR Data: Employee Attrition and Performance Insights". In: *ResearchGate* (2024). URL: https://www.researchgate.net/publication/383426416_Analyzing_IBM_HR_Data_Employee_Attrition_and_Performance_Insights.

[2] Pavan Subhash. *IBM HR Analytics Attrition Dataset.* https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset. 2017.

[3] Jason Brownlee. *SMOTE Oversampling for Imbalanced Classification.* https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/. 2021.

[4] Shenghuan Yang and Md Tariqul Islam. "IBM Employee Attrition Analysis". In: *arXiv* 2012.01286 (2021). Jiangxi University of Finance and Economics. URL: https://arxiv.org/pdf/2012.01286.

[5] OpenAI. *ChatGPT: Language Model.* https://chat.openai.com. 2024.

Generative AI Statement

AI-supported/AI-integrated use is permitted in this assessment. I acknowledge the following uses of GenAI tools in this assessment:

- YES I have used GenAI tools for developing ideas.
- YES I have used GenAI tools to assist with research or gathering information.
- YES I have used GenAI tools to help me understand key theories and concepts.
- NO I have used GenAI tools to identify trends and themes as part of my data analysis.
- NO I haven't used GenAI tools to suggest a plan or structure for my assessment.
- YES I have used GenAI tools to give me feedback on a draft.
- YES I have used GenAI tool to generate image, figures or diagrams.
- YES I have used GenAI tools to proofread and correct grammar or spelling errors.
- NO I haven't used GenAI tools to generate citations or references.
- YES Other: I have used GenAI to debug my code

I declare that I have referenced use of GenAI outputs within my assessment in line with the University referencing guidelines.