Predicting Employee Attrition Using Machine Learning

Lior Biton, Coral Yagud, Naama Maimon, Stav Barak

Research Question

What are the main factors leading to employee attrition in large companies?

Abstract

Employee attrition, commonly referred to as turnover, represents a significant challenge for organizations. High attrition rates lead to increased recruitment and training costs, loss of organizational knowledge, decreased employee morale, and reduced overall productivity. Understanding the factors contributing to employee attrition is crucial for developing effective retention strategies.

Data analytics and machine learning provide new avenues for predicting and understanding employee attrition. This report leverages the IBM HR Analytics Employee Attrition & Performance dataset, encompassing various features related to employee demographics, job satisfaction, performance, and work environment.

The analysis revealed critical factors influencing employee attrition, including age, monthly income, and overtime. Among the models tested, XGBoost demonstrated the highest accuracy, followed closely by Logistic Regression and Random Forest. While Logistic Regression and XGBoost showed better recall for attrition cases, the application of SHAP values in the analysis provided an in-depth understanding of feature importance, allowing for a more transparent and interpretable model.

Keywords: Employee Attrition, Machine Learning, Logistic Regression, Random Forest, XGBoost, SHAP values

Introduction

Employee attrition, commonly referred to as turnover, represents a significant challenge for organizations, particularly in the software industry, where the churn rate ranges from approximately 12-15%. High attrition rates lead to increased recruitment and training costs, loss of organizational knowledge, decreased employee morale, and reduced overall productivity. Understanding the factors contributing to employee attrition is crucial for developing effective retention strategies.

The cost implications of attrition are substantial. When skilled employees leave, organizations face the challenge of finding and training suitable replacements, incurring both direct financial costs and indirect costs such as lost productivity and potential disruptions in team dynamics. High turnover rates can also damage an organization's reputation, making it difficult to attract top talent.

Key factors influencing employee churn include workload, working conditions, salary, job experience, and job satisfaction. Voluntary turnover, where employees leave by choice, is particularly costly as it often involves the loss of valuable employees to competitors.

This report leverages the IBM HR Analytics Employee Attrition & Performance dataset, which includes various features related to employee demographics, job satisfaction, performance, and work environment. By employing multiple machine learning models—Logistic Regression, Random Forest, and XGBoost—this study aims to build predictive models of employee attrition. Logistic Regression serves as a robust baseline, while Random Forest and XGBoost capture complex interactions between features. SHapley Additive exPlanations (SHAP) values are used to assess feature importance, offering a unified measure of each feature's contribution to the model's predictions.

Background/Related Work

Article 1: ”HR Analytics: Employee Attrition Analysis Using Logistic Regression”
Purpose: The study focuses on analyzing employee attrition using logistic regression to provide insights for management to understand and address workplace modifications that can improve employee retention. It emphasizes the cost implications of attrition on the organization’s competitive advantage and aims to identify key driving factors for attrition.
Functionality: The analysis employs logistic regression on a dataset comprising around four thousand employees over one year. Five steps are followed, including data collection and business understanding, data pre-processing, exploratory data analysis (EDA), model selection and training, and model evaluation. R studio is used for data integration, exploratory data analysis, data preparation, logistic regression, model evaluation, and visualization.
Conclusions: The study identifies eleven variables as key driving factors for employee attrition, including factors related to employee tenure, job satisfaction, working conditions, and marital status. It highlights the importance of both employee and company factors in driving attrition rates. The findings suggest that to reduce attrition, companies should focus on improving the working environment, job satisfaction, workload management, and employee-manager interactions.
Research Gaps: While the study provides valuable insights into employee attrition, limitations include the focus on logistic regression as the sole analytical method and the absence of comparison with other machine learning techniques. Additionally, the study could benefit from discussing the generalizability of the findings to other organizational contexts and regions.
URL: HR Analytics: Employee Attrition Analysis Using Logistic Regression

Article 2: ”Analysing Employee Attrition Using Machine Learning”
Purpose: The study aimed to evaluate the performance of various machine learning algorithms, including decision tree, naïve Bayes, and k-means, in predictive analysis using the IBM dataset, which is the same dataset utilized in our research. Specifically, the objective was to gauge the effectiveness of these algorithms in addressing prediction tasks relevant to the dataset.
Functionality: The study applied a range of machine learning algorithms to the dataset and assessed their performance through rigorous validation techniques, such as 10-fold cross-validation and a 70%:30% train-test split.
Conclusions: The study reported that the accuracy of their predictive models fell short compared to similar studies. This disparity in accuracy is likely attributable to the absence of data preprocessing, which could have contributed to noise and inconsistencies in the dataset, thus affecting the model’s performance.
Research Gaps: The study underscores the significance of data preprocessing in machine learning endeavors. It underscores the necessity of meticulous data cleaning and preparation to bolster the accuracy and reliability of predictive models.
URL: Analysing Employee Attrition Using Machine Learning

Article 3: ”External Alternatives, Job Stress on Job Satisfaction and Employee Turnover Intention”
Purpose: The study investigate the influence of external alternatives and job stress on employee satisfaction and turnover intention, specifically focusing on employees of PT Bank Mandiri Regional X South Sulawesi.
Functionality: The research employs structural equation modeling with a Partial Least Square (PLS) approach to analyze the relationships between variables. The variables are external alternatives, work environment, stress, and job satisfaction.
Conclusions: The results indicate that external factors, particularly job market conditions outside the company, have a stronger influence on turnover intention compared to job satisfaction. High job satisfaction alone may not reduce turnover intention if employees perceive abundant job opportunities in the external market. Therefore, the study emphasizes the importance of addressing external alternatives to mitigate turnover intention.
Research Gaps: While the study provides valuable insights into turnover intention, limitations include its focus on a specific company and region, which may limit the generalizability of the findings. Additionally, investigating additional factors influencing turnover intention and their interactions could provide a more comprehensive understanding of employee turnover dynamics.
URL: External Alternatives Job Stress on Job Satisfaction and Employee Turnover Intention

Article 4: ”From Explanations to Feature Selection: Assessing SHAP Values as Feature Selection Mechanism”
Purpose: The paper ”From Explanations to Feature Selection: Assessing SHAP Values as Feature Selection Mechanism” by Wilson E. Marcilio-Jr and Danilo M. Eler explores the use of SHAP (SHapley Additive exPlanations) values, a game-theoretic approach, for feature selection in machine learning models. SHAP values are designed to explain the output of any machine learning model by attributing contributions to each feature.
Functionality: The authors propose using SHAP values for feature selection by ordering features based on their contribution to the model’s output. They validate this methodology through experiments on eight publicly available

datasets, comparing it against three common feature selection algorithms: Mutual Information, Recursive Feature Elimination (RFE), and ANOVA.
Conclusions: The experiments demonstrated that SHAP-based feature selection consistently outperformed the other methods across different datasets. SHAP not only provided better predictive performance but also enhanced interpretability by explaining the importance of each feature. SHAP values offer a robust and interpretable mechanism for feature selection, improving both model performance and explainability. This approach is particularly beneficial for high-dimensional datasets where understanding feature importance is crucial.
Application in Employee Attrition: For a project on employee attrition, SHAP-based feature selection can identify the most influential factors contributing to employee turnover. By understanding these key features, organizations can develop targeted interventions to reduce attrition rates, making SHAP a valuable tool for predictive analytics and human resource management.
URL: From Explanations to Feature Selection: Assessing SHAP Values as Feature Selection Mechanism

Article 5: "Understanding Logistic Regression Analysis"
Purpose: The article "Understanding Logistic Regression Analysis" by Sandro Sperandei provides a comprehensive overview of logistic regression, a statistical method widely used in research for analyzing datasets with one or more independent variables that determine a binary outcome. The purpose is to elucidate the calculation and interpretation of logistic regression, emphasizing its utility in medical and clinical research.
Functionality: Logistic regression is similar to multiple linear regression but is used for modeling binary outcome variables. It employs the logit function to relate the probability of an event occurring to a linear combination of predictor variables.

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

The method calculates odds ratios (OR $= e^{\beta_i}$) to quantify the change in odds of the outcome associated with a one-unit change in the predictor.

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}}$$

Logistic regression effectively adjusts for confounders, includes both categorical and continuous variables, and handles multiple explanatory variables simultaneously.
Conclusions: Sperandei's article highlights that logistic regression is a powerful tool for analyzing complex relationships in data by providing accurate odds ratios that reflect the influence of each predictor on the outcome. It offers a clearer understanding of the true effects of the predictors. Logistic regression manages multiple predictors without needing to arbitrarily categorize them, thus preserving data integrity.
Application in Employee Attrition: In the context of employee attrition analysis, logistic regression can identify key predictors of turnover by modeling the probability of an employee leaving the company based on various factors.
URL: Understanding Logistic Regression Analysis

Article 6: "A Random Forest Guided Tour"
Purpose: The article "A Random Forest Guided Tour" by Gérard Biau and Erwan Scornet provides an in-depth exploration of the random forest algorithm, a powerful tool in machine learning for both classification and regression tasks. The authors aim to explain the mathematical principles behind the algorithm, review its theoretical developments, and highlight its practical applications across various fields. The purpose is to make the main ideas of random forests accessible to non-experts while providing detailed insights into its functionality and theoretical underpinnings.
Functionality: Random forests operate by combining several randomized decision trees and aggregating their predictions. Each tree in the forest is built using a different random sample of the data, and the final prediction is obtained by averaging the predictions of all trees. This ensemble method is particularly effective when dealing with high-dimensional data and small sample sizes. Key features of the algorithm include:

- **Randomization**: Each tree is constructed using a different subset of the data and variables, enhancing diversity.
- **Bagging**: Trees are built on bootstrap samples of the data, which helps in reducing variance.
- **Splitting Criterion**: The algorithm uses the Classification and Regression Trees (CART) methodology to determine the best splits at each node.
- **Variable Importance**: Random forests provide measures of variable importance, helping to identify the most influential predictors in the dataset.
- **Versatility**: The method can handle both large-scale problems and small sample sizes, making it applicable to a wide range of learning tasks.

Conclusions: Biau and Scornet's article emphasizes that random forests are a robust and versatile tool for data analysis. They are particularly effective at handling high-dimensional data by constructing multiple trees and averaging their predictions, which allows the algorithm to manage high-dimensional spaces effectively. The ensemble approach also helps to reduce the risk of overfitting compared to individual decision trees. Random forests provide accurate predictions and have shown excellent performance in various settings, including classification, regression, and survival analysis. Additionally, the algorithm's built-in mechanism for assessing variable importance aids in understanding the influence of different predictors.

Application in Employee Attrition: In the context of employee attrition analysis, random forests can be employed to identify key predictors of employee turnover by modeling the probability of an employee leaving the organization based on various factors.

URL: A Random Forest Guided Tour

Table Summary

Article	Purpose	Methodology	Key Findings	Research Gaps
HR Analytics: Employee Attrition Analysis Using Logistic Regression	Analyze employee attrition using logistic regression to identify key factors	Logistic Regression; R studio for data analysis	Identified eleven key driving factors for employee attrition, including tenure, job satisfaction, working conditions, marital status	Focused solely on logistic regression; lacks comparison with other machine learning techniques
Analysing Employee Attrition Using Machine Learning	Evaluate performance of various machine learning algorithms on employee attrition prediction	Decision Tree, Naïve Bayes, k-means; 10-fold cross-validation and 70%:30% train-test split	Predictive models' accuracy was lower due to lack of data preprocessing	Need for meticulous data cleaning and processing
External Alternatives, Job Stress on Job Satisfaction and Employee Turnover Intention	Investigate the influence of external alternatives and job stress on employee satisfaction and turnover intention	Structural Equation Modeling (PLS)	External alternatives significantly affect job satisfaction and turnover intention; job satisfaction impacts turnover intention; stress does not significantly influence job satisfaction or turnover intention	Limited focus on a specific company and region; additional other factors not considered
From Explanations to Feature Selection: Assessing SHAP Values as Feature Selection Mechanism	Assess SHAP values for feature selection in machine learning models	SHAP values; comparison with Mutual Information, RFE, ANOVA on eight datasets	SHAP-based feature selection outperformed other methods; improved predictive performance and interpretability; SHAP values provide a robust mechanism for feature selection	-
Understanding Logistic Regression Analysis	Provide an overview of logistic regression for analyzing binary outcomes	Logistic Regression; calculation of odds ratios	Logistic regression is a powerful tool for analyzing complex relationships in data, providing accurate odds ratios	-

Continued on next page

Table 1 – *Continued from previous page*

Article	**Purpose**	**Methodology**	**Key Findings**	**Research Gaps**
A Random Forest Guided Tour	Explain the random forest algorithm and its practical applications	Random Forest; combination of randomized decision trees and bootstrap samples	Random forests manage high-dimensional data well, reduce overfitting, and provide measures of variable importance Random forests are versatile and robust, offering accurate predictions and handling various tasks effectively	-

Methodology

Dataset

IBM HR Analytics Attrition Dataset

Data Collection

The dataset used in this research is the IBM HR Analytics Employee Attrition & Performance dataset, publicly accessible via Kaggle. This dataset includes 1470 entries with 35 columns, encompassing numerical and categorical data, making it highly suitable for an in-depth analysis of factors contributing to employee attrition.

Dataset Overview

The dataset is comprehensive, including features pertinent to employee performance, satisfaction, and demographics. Notably, there are no missing values within the dataset, ensuring the integrity and completeness of the data for subsequent analyses.

Data Cleaning

Initial Data Inspection

A preliminary inspection of the dataset confirmed the absence of missing values and duplicate entries, negating the need for imputation or duplicate removal procedures. This initial inspection is critical to ensure the dataset's quality before proceeding with further analysis.

Data Type Conversion

Categorical variables were converted to appropriate data types to facilitate efficient data manipulation and analysis. Specifically, the 'Attrition' and 'OverTime' columns were mapped to binary integers (0 and 1) to streamline their integration into predictive modeling workflows. This conversion is essential for machine learning algorithms to process the data correctly.

Summary of Data Cleaning

Before cleaning, the dataset contained 1470 entries with a mix of numerical and categorical data types and no missing values. After cleaning, the dataset retained 1470 entries with categorical variables correctly typed and binary mapping applied to the 'Attrition' and 'OverTime' columns. This step ensures that the dataset is ready for accurate and efficient analysis.

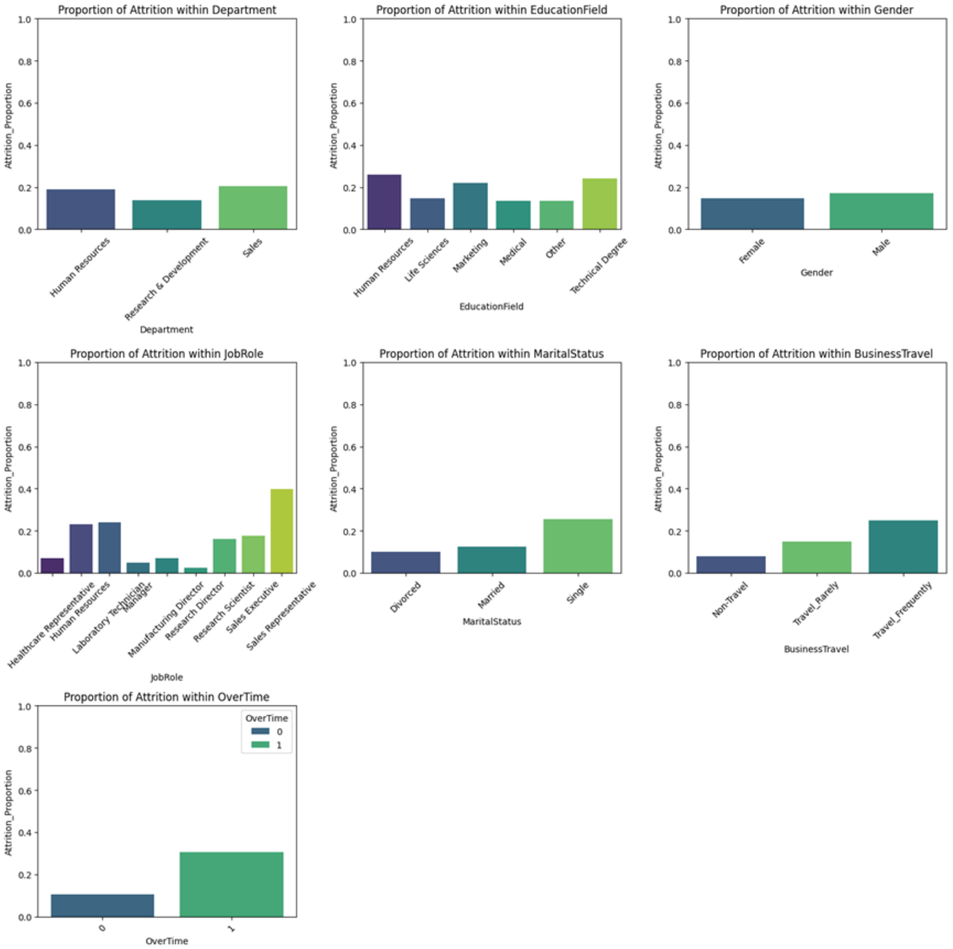
Exploratory Data Analysis (EDA)

Summary Statistics and Visualizations

Exploratory Data Analysis (EDA) was conducted to uncover initial insights and patterns within the dataset. Summary statistics were calculated for all numerical variables, and visualizations such as histograms and count plots were generated to illustrate the distribution of key features, including age, job satisfaction, and overtime status.

Proportion Plots

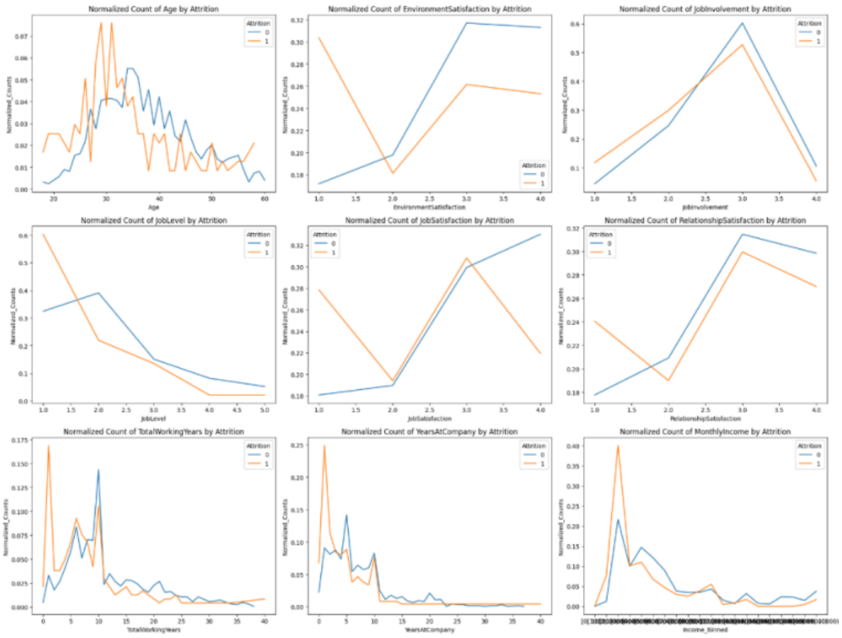
Proportion plots were created to analyze the relationship between categorical variables and the attrition proportion. Key insights from these plots include that single employees exhibit a higher attrition proportion compared to married or divorced employees. Overtime - Employees working overtime have significantly higher attrition rates. Employees who travel frequently show a higher attrition proportion compared to those who travel rarely or not at all. Additionally, the Sales department has a higher attrition proportion compared to Research & Development and Human Resources. Employees with backgrounds in Human Resources, Technical Degrees, and Marketing exhibit elevated attrition rates. The analysis revealed no significant difference in attrition proportions between males and females. However, higher attrition was observed in the roles of Sales Representative and Laboratory Technician.



Line Plots

Line plots were created to analyze the relationship between numerical variables and attrition status. The normalized counts of various features split by attrition status (0 = No, 1 = Yes) reveal several prominent trends. Age stands out as a significant factor, with higher attrition observed among younger employees, especially those in their 20s and early 30s, whereas older employees show less attrition. Monthly income is also a critical factor, with employees earning lower monthly incomes showing higher attrition rates, while those with higher incomes tend to stay longer. Job satisfaction similarly influences attrition, with higher job satisfaction (level 4) linked to lower attrition and lower job satisfaction correlating with higher attrition. Environment satisfaction reveals that employees with lower environment satisfaction (level 1) show higher attrition rates compared to those with higher satisfaction levels.

Other factors, though less distinct, also show patterns: higher job involvement is associated with lower attrition, while lower job involvement correlates with higher attrition. Employees at lower job levels experience higher attrition rates, while those at higher levels show less attrition. Relationship satisfaction shows mixed results, with higher levels generally associated with lower attrition. Employees with fewer working years (especially 0-10 years) show higher attrition rates, and attrition decreases as total working years increase. Finally, higher attrition rates are observed among employees with fewer years at the company, particularly those with less than five years.



Feature Engineering

Transformations

To prepare the data for modeling, categorical variables were converted into numerical formats using label encoding and ordinal encoding techniques. Additionally, irrelevant features such as EmployeeCount, EmployeeNumber, Over18, and StandardHours were removed from the dataset to enhance model performance and interpretability. This transformation step ensures that the machine learning models can process the data efficiently and accurately.

Model Selection and Training

Model Selection

Three predictive models were selected for this analysis: Logistic Regression, Random Forest, and XGBoost. These models were chosen for their complementary strengths in handling different types of data relationships and their robustness in classification tasks. Logistic Regression is a simple yet effective model for binary classification, Random Forest is known for its robustness and ability to handle high-dimensional data, and XGBoost is a powerful gradient boosting technique that often provides superior performance.

Training and Evaluation

The dataset was split into training and testing sets using a 70-30 split ratio. Each model was trained on the training set and evaluated on the test set. Performance metrics, including accuracy, precision, recall, and F1-score, were calculated to assess each model's efficacy in predicting employee attrition. These metrics provide a comprehensive evaluation of the models' performance, ensuring that the selected model not only achieves high accuracy but also performs well in identifying both attrition and non-attrition cases.

Results - Model Performance

Logistic Regression

Logistic Regression achieved an accuracy of 87.07%, with precision, recall, and F1-score for class 0 (non-attrition) being 0.895, 0.963, and 0.928, respectively. For class 1 (attrition), the precision, recall, and F1-score were 0.563, 0.295, and 0.387, respectively. This model demonstrated a high ability to correctly identify non-attrition cases, but its performance in identifying attrition cases was comparatively lower.

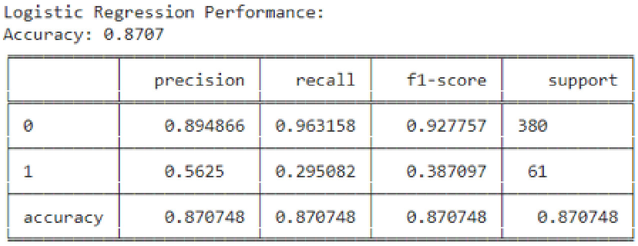
Random Forest

The Random Forest model showed an accuracy of 86.39%, with precision, recall, and F1-score for class 0 being 0.872, 0.987, and 0.926, respectively. For class 1, the precision, recall, and F1-score were 0.545, 0.098, and 0.167, respectively. While this model also performed well in identifying non-attrition cases, its ability to identify attrition cases was limited.

XGBoost

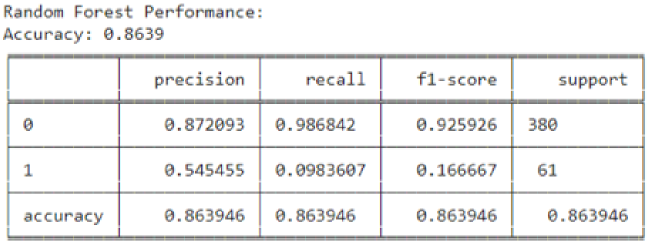
XGBoost demonstrated the highest accuracy at 87.30%, with precision, recall, and F1-score for class 0 being 0.891, 0.971, and 0.929, respectively. For class 1, the precision, recall, and F1-score were 0.593, 0.262, and 0.364, respectively.

This model achieved the best balance between identifying attrition and non-attrition cases, making it the most effective model for this analysis.



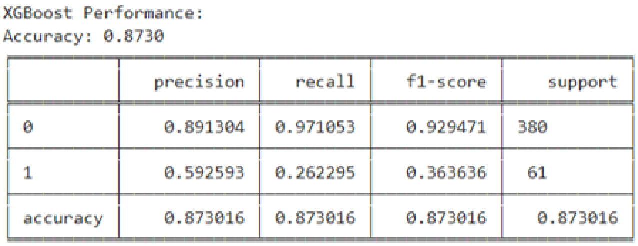
Logistic Regression Performance:
Accuracy: 0.8707

	precision	recall	f1-score	support
0	0.894866	0.963158	0.927757	380
1	0.5625	0.295082	0.387097	61
accuracy	0.870748	0.870748	0.870748	0.870748



Random Forest Performance:
Accuracy: 0.8639

	precision	recall	f1-score	support
0	0.872093	0.986842	0.925926	380
1	0.545455	0.0983607	0.166667	61
accuracy	0.863946	0.863946	0.863946	0.863946

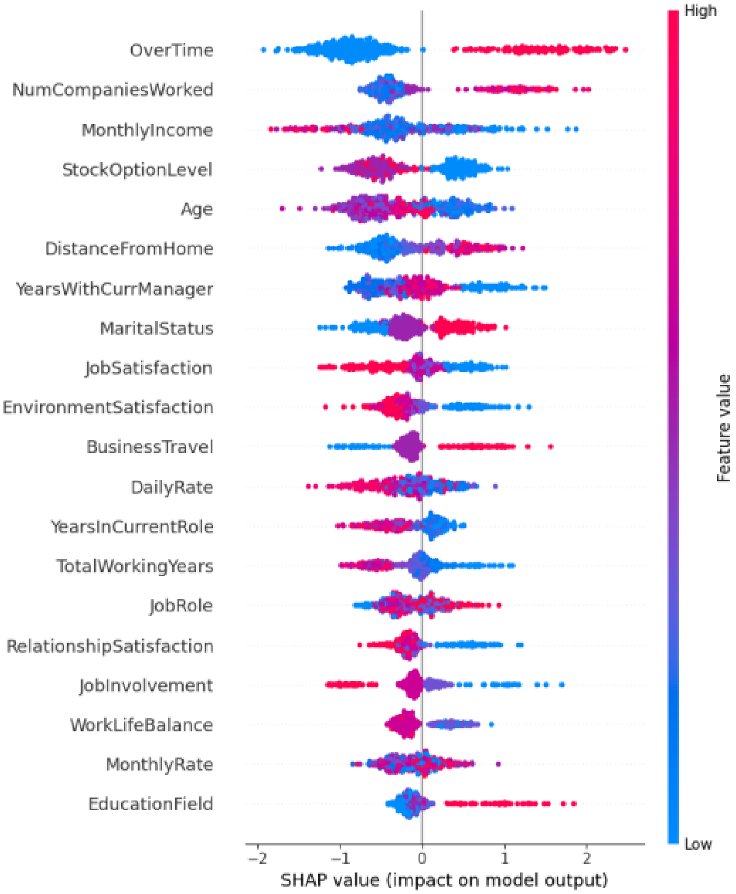


XGBoost Performance:
Accuracy: 0.8730

	precision	recall	f1-score	support
0	0.891304	0.971053	0.929471	380
1	0.592593	0.262295	0.363636	61
accuracy	0.873016	0.873016	0.873016	0.873016

Feature Importance

The SHAP summary plot highlights several key factors influencing employee attrition. High overtime significantly increases the likelihood of employee attrition, suggesting that excessive work hours negatively impact retention. Employees with experience at multiple companies are more prone to leave, indicating a higher propensity for job switching. Lower monthly incomes are strongly associated with higher attrition rates, underscoring the critical role of competitive compensation in employee retention. Higher stock option levels are linked to reduced attrition, showing that equity compensation is an effective tool for retention. Younger employees exhibit higher attrition rates, likely due to seeking career advancement opportunities. Longer commute distances are correlated with increased attrition, emphasizing the importance of manageable commute times for employee satisfaction and retention. Addressing these factors through improved compensation, balanced work hours, and considerations for commute distance can enhance employee retention strategies.



Discussion

The analysis revealed several critical factors influencing employee attrition. Age, monthly income, overtime and stock option levels, emerged as significant predictors. Younger employees, particularly those in their 20s and early 30s, showed higher attrition rates. This could be attributed to the higher likelihood of younger employees seeking career advancement opportunities outside the current organization. Monthly income was another crucial factor, with lower-income employees exhibiting higher attrition rates. This suggests that financial incentives and compensation packages play a vital role in employee retention.

Overtime also significantly influenced attrition, with employees working overtime more frequently showing higher attrition rates. This finding highlights the potential negative impact of work-life balance on employee retention.

Among the models tested, XGBoost demonstrated the highest accuracy, followed closely by Logistic Regression and Random Forest. While Logistic Regression and XGBoost showed better recall for attrition cases, Random Forest's performance in this regard was comparatively lower. The use of SHAP values in feature importance analysis provided an in-depth understanding of the factors contributing to attrition, enhancing the models' interpretability.

These insights can inform strategic interventions to mitigate attrition and enhance employee retention. For instance, organizations could focus on improving compensation packages, promoting work-life balance, and providing career development opportunities to retain younger employees.

Future Work

Future research should address the class imbalance issue in predicting employee attrition. The current models show high accuracy in predicting non-attrition (class 0) but struggle with attrition (class 1). Balancing the classes is crucial because understanding and predicting who is likely to leave is more critical for strategic interventions.

Delving deeper into the prediction of class 1 can significantly enhance retention strategies. Techniques such as oversampling, undersampling, or using advanced algorithms like SMOTE (Synthetic Minority Over-sampling Technique) could be employed to balance the dataset and improve the prediction accuracy for attrition cases.

Additionally, future work should explore the use of deep learning models. Deep learning, with its ability to capture complex patterns and interactions within data, can potentially provide more robust and accurate predictions. Models such as neural networks could be investigated for their applicability in this context. Furthermore, survival analysis can also be an interesting direction, as it accounts for the fact that eventually, employees will leave. Survival analysis techniques can provide insights into the timing of employee attrition, helping organizations understand not just who is likely to leave, but when they are most likely to do so, enabling more proactive and timely interventions.

Incorporating additional factors that have not been studied yet could also enhance the predictive power of the models. Factors such as organizational culture, employee engagement, leadership styles, and external economic indicators could provide a more comprehensive understanding of the reasons behind employee turnover. Integrating real-time data and qualitative data from employee feedback and exit interviews could further refine the models and offer deeper insights.

By addressing these areas, future research can build more accurate and actionable models, ultimately helping organizations better manage and reduce employee attrition.

References

1. I Setiawan, S. Suprihanto, A. C. Nugraha, J. Hutahaean. (2020). HR analytics: Employee attrition analysis using logistic regression. Retrieved from `https://iopscience.iop.org/article/10.1088/1757-899X/830/3/032001/pdf`

2. Usha, P. M., & DR. N. V. Balaji. (2021). Analysing employee attrition using machine learning. Retrieved from `https://karpagampublications.com/wp-content/uploads/2020/03/Karpagam-Sep-Oct-2019-Article-6.pdf`

3. Ramlawati, Eva T., Nurfatwa, A. Y., & Kurniawaty. (2020). External Alternatives Job Stress on Job Satisfaction and Employee Turnover Intention. Retrieved from `https://m.growingscience.com/msl/Vol11/msl_2020_319.pdf`

4. Marcilio-Jr, W. E., & Eler, D. M. (2020). From Explanations to Feature Selection: Assessing SHAP Values as Feature Selection Mechanism. Retrieved from `http://sibgrapi.sid.inpe.br/col/sid.inpe.br/sibgrapi/2020/09.25.14.27/doc/PID6618233.pdf`

5. Sperandei, S. (2014). Understanding Logistic Regression Analysis. Retrieved from `https://hrcak.srce.hr/file/171128`

6. Biau, G., & Scornet, E. (2016). A Random Forest Guided Tour. Retrieved from `https://arxiv.org/pdf/1511.05741`