ISTINYE | University

Faculty/Institute of Postgraduate Education Institute

Machine Learning Applications In Business - DATS5027

Understanding and Predicting Employee Attrition: A Data-Driven Approach to Workforce Retention

Mohamed Souleimane Cheikh ahmed

2333265040

Advisor: Prof.Dr. Mustafa SUNDU

İstanbul Jan, 2025

Abstract

Employee attrition poses significant challenges to organizations, impacting operational efficiency, financial performance, and workplace morale. This project analyzes the HR-employee attrition dataset containing detailed information on 1,470 employees to uncover the key factors driving attrition and develop predictive models to estimate the likelihood of employee turnover. Through exploratory data analysis (EDA), the project identifies critical drivers such as job satisfaction, work-life balance, and salary, assessing their influence on employee decisions. Machine learning techniques are applied to train classification models, enabling organizations to proactively address attrition risks. By providing actionable insights and a robust prediction framework, this project aims to empower businesses with the tools needed to improve employee retention, foster a supportive workplace culture, and reduce turnover costs.

Table of Contents

Introduction

Employee attrition, or the rate at which employees leave an organization, is a pressing challenge faced by businesses worldwide. High attrition rates can lead to significant operational disruptions, increased recruitment and training costs, and a loss of institutional knowledge. In knowledge-driven economies, retaining skilled employees is paramount for maintaining competitiveness and fostering innovation. Despite these challenges, many organizations struggle to pinpoint the specific factors that lead to employee turnover or predict which employees are at risk of leaving.
The HR-employee attrition dataset used in this project provides detailed information about 1,470 employees, including demographic, professional, and organizational attributes such as age, salary, distance from home, and job satisfaction. By leveraging this data, this project aims to address the following critical business needs:

1. **Understanding Attrition Drivers:** Through exploratory data analysis, the project seeks to uncover the primary reasons behind employee attrition and understand how various factors influence employees' decisions to stay or leave. This includes assessing whether certain variables—such as work-life balance, career growth opportunities, or job satisfaction—have a positive or negative impact on attrition.

2. **Predictive Modeling for Proactive Interventions**: Using machine learning techniques, the project aims to develop a robust predictive model that can estimate the likelihood of an employee leaving the organization. Such a model enables HR departments to identify at-risk employees early and take targeted measures to address their concerns, improving retention rates.

3. **Business Implications:** Beyond technical objectives, the project provides actionable insights for decision-makers. By understanding attrition trends and their implications, organizations can optimize HR policies, enhance employee engagement, and foster a more supportive workplace culture. Addressing attrition not only improves financial performance by reducing turnover costs but also strengthens the organization’s reputation as an employer of choice.

Relevance and Significance

This project is particularly relevant in the context of modern workforce dynamics, where employees increasingly value factors like work-life balance, meaningful work, and personal development opportunities. As hybrid and remote work models reshape traditional workplace structures, organizations need data-driven insights to adapt to these changes effectively.

Moreover, high attrition rates are not confined to specific industries or regions; they are a global phenomenon with far-reaching consequences. For instance, studies estimate that replacing an employee can cost businesses up to twice their annual salary due to recruitment, training, and lost productivity costs. By addressing these challenges through advanced data science techniques, this project provides a pathway for businesses to mitigate attrition risks and achieve sustainable growth.

Dataset Description

This project utilizes the HR-employee attrition dataset provided by Rushikesh Konapure, available on Kaggle. The dataset contains detailed information on 1,470 employees across 35 features, including demographics, job roles, satisfaction levels, and compensation metrics. It is structured as a tabular dataset with no missing values, making it suitable for both exploratory analysis and predictive modeling.

Dataset Source

The dataset is publicly accessible at Kaggle:

HR Analytics Prediction Dataset by Rushikesh Konapure

Dataset Overview

- **Number of Records**: 1,470
- **Number of Features**: 35 (including the target variable)
- **Target Variable**: Attrition (Binary: "Yes" or "No")
- **Data Types**:
    - Numeric (e.g., Age, DailyRate)
    - Categorical (e.g., Gender, JobRole)

Feature Descriptions

Below is a detailed description of the features included in the dataset:

#	Feature	Data Type	Description
1	Age	Numeric	Age of the employee in years.
2	Attrition	Categorical	Target variable indicating if the employee left the company ("Yes"/"No").
3	BusinessTravel	Categorical	Frequency of business travel (e.g., "Travel_Rarely", "Travel_Frequently").
4	DailyRate	Numeric	Daily income of the employee.
5	Department	Categorical	Department where the employee works (e.g., Sales, R&D).
6	DistanceFromHome	Numeric	Distance between the employee's residence and workplace (in km)
7	Education	Numeric	Education level (1 = 'Below College', 5 = 'Doctor').

8	EducationField	Categorical	Field of education (e.g., Life Sciences, Technical Degree).
9	EmployeeCount	Numeric	A constant value for all records (1).
10	EmployeeNumber	Numeric	Unique identifier for each employee.
11	EnvironmentSatisfaction	Numeric	Satisfaction with the work environment (1 = 'Low', 4 = 'Very High').
12	Gender	Categorical	Gender of the employee (e.g., Male, Female).
13	HourlyRate	Numeric	Hourly wage of the employee.
14	JobInvolvement	Numeric	Employee's involvement level in their job (1 = 'Low', 4 = 'Very High').
15	JobLevel	Numeric	Employee's level within the organization hierarchy.
16	JobRole	Categorical	Role or position held by the employee (e.g., Manager, Sales Representative).
17	JobSatisfaction	Numeric	Satisfaction level with the job (1 = 'Low', 4 = 'Very High').
18	Marital Status	Categorical	Marital status of the employee (e.g., Single, Married).
19	MonthlyIncome	Numeric	Total monthly income of the employee.
20	MonthlyRate	Numeric	Monthly rate of compensation.
21	NumCompaniesWorked	Numeric	Number of companies the employee has previously worked at.
22	Over18	Categorical	Indicates if the employee is above 18 years of age (constant: "Yes").
23	OverTime	Categorical	Whether the employee works overtime (Yes/No).
24	PercentSalaryHike	Numeric	Percentage increase in salary during the last hike.
25	PerformanceRating	Numeric	Employee's performance rating (1 = 'Low', 4 = 'Outstanding').
26	RelationshipSatisfaction	Numeric	Satisfaction level with workplace relationships (1 = 'Low', 4 = 'Very High').
27	StandarHours	Numeric	Standard working hours (constant: 80).
28	StockOptionLevel	Numeric	Stock option level granted to the employee.
29	TotalWorkingYears	Numeric	Total years the employee has been working.
30	TrainingTimesLastYear	Numeric	Number of training sessions attended in the last year.
31	WorkeLifeBalance	Numeric	Employee's work-life balance rating (1 = 'Bad', 4 = 'Best').
32	YearsInCurrentRole	Numeric	Number of years the employee has been with the company.
33	YearsSinceLastPromotion	Numeric	Number of years the employee has been in their current role.
34	YearsAtCompany	Numeric	Years since the employee's last promotion.
35	YearswithCurrManager	Numeric	Number of years the employee has worked with their current manager.

Dataset Relevance

The dataset is highly relevant for addressing the project objectives due to its comprehensive set of features related to employee demographics, work environment, and job satisfaction. These attributes provide valuable insights into the factors influencing attrition and enable the development of predictive models. The inclusion of both categorical and numerical variables also allows for a diverse range of analyses, making it ideal for studying employee retention.

Exploratory Data Analysis

3.1 Overview of Categorical Variables

To explore the dataset, categorical variables were analyzed to understand their distribution and relationships with the target variable, Attrition.

Distribution of Categorical Variables

- A pie chart was used to visualize the overall distribution of each categorical variable. Key observations include:
  - **Attrition**: 83.9% of employees did not leave (No), while 16.1% left (Yes). *(Fig. 1)*

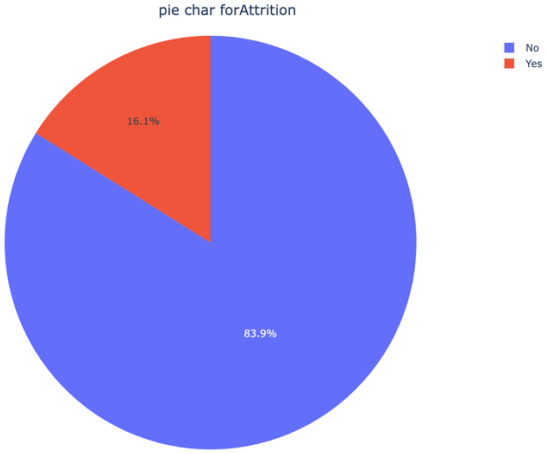

Figure 1

  - **Business Travel**: Most employees travel rarely (71%), while 18.8% travel frequently, and 10.2% do not travel. *(Fig. 2)*

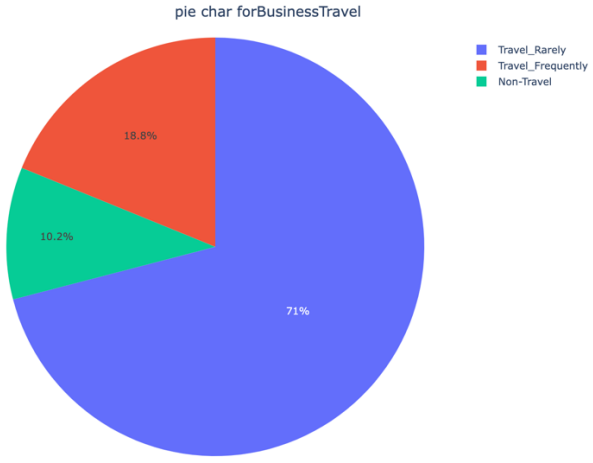

Figure 2

  - **Department**: The majority belong to Research & Development (65.4%), followed by Sales (30.3%) and Human Resources (4.29%). *(Fig. 3)*

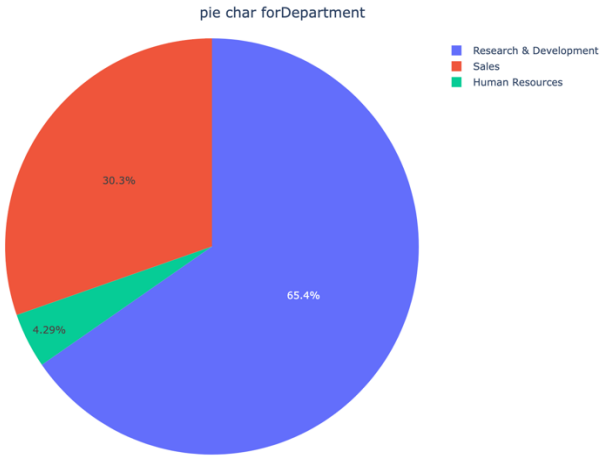

Figure 3

- **Gender**: The workforce comprises 60% males and 40% females. *(Fig. 4)*

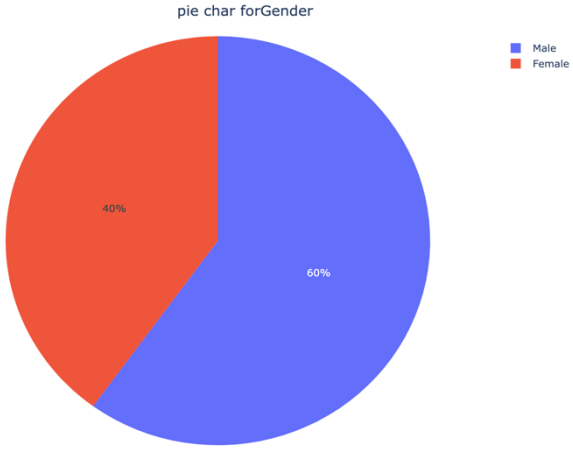

Figure 4

- **Marital Status**: A significant portion of employees are married (45.8%). *(Fig. 5)*

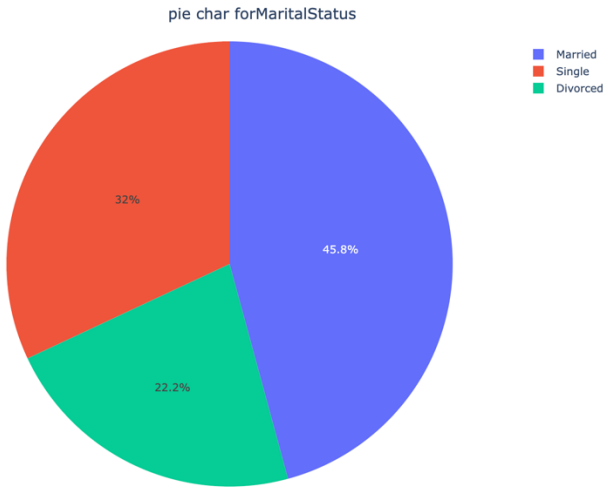

Figure 5

- **OverTime**: 71% of employees do not work overtime. *(Fig. 6)*

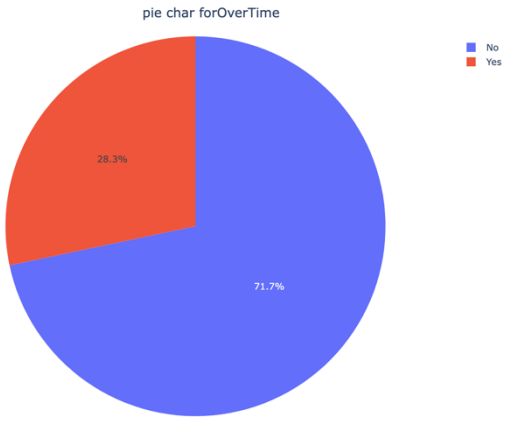

Figure 6

Attrition vs. Categorical Variables

- When examining categorical variables in relation to attrition, the following insights were observed:
  - **Job Role**: Sales representatives have the highest attrition rate (39%), followed by laboratory technicians (24%). *(Fig. 7)*

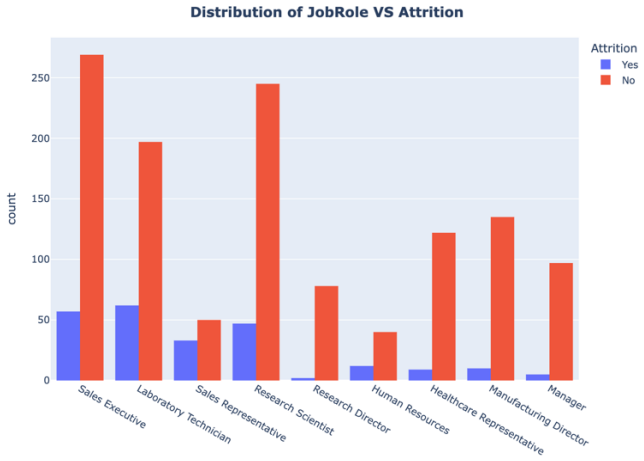

Figure 7

  - **Marital Status**: Single employees have the highest attrition rate (25%). *(Fig. 8)*

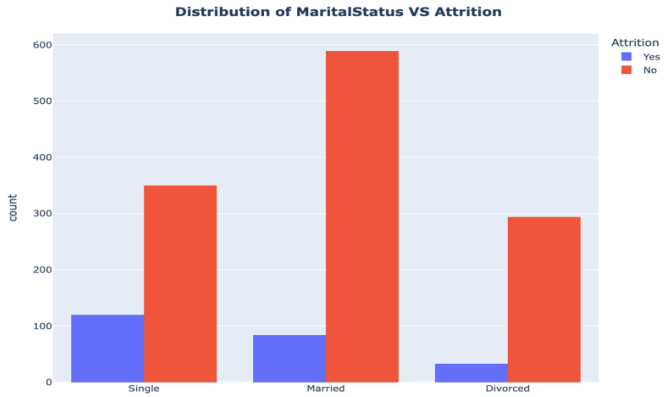

Figure 8

3.2 Overview of Numerical Variables

Distribution of Numerical Variables

- The majority of numerical variables are normally distributed

Attrition vs. Numerical Variables

- Violin plots were generated for numerical variables, separated by Attrition, to explore relationships. Key findings include:
  - **Age**: Employees who left are predominantly around 30 years old. *(Fig. 09)*

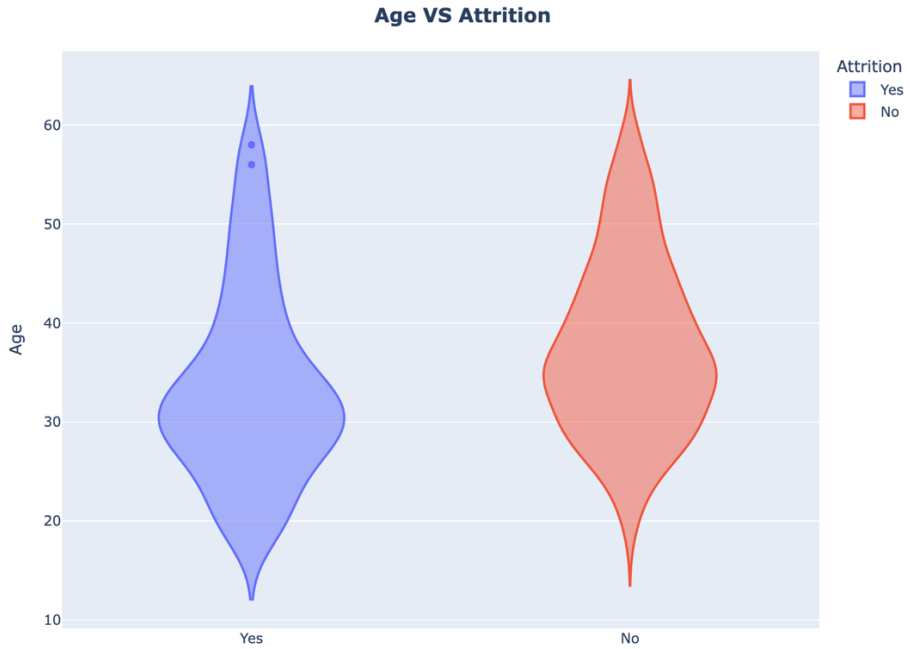

Figure 9

  - **Job Level**: Most employees who left were at Job Level 1. *(Fig. 10)*

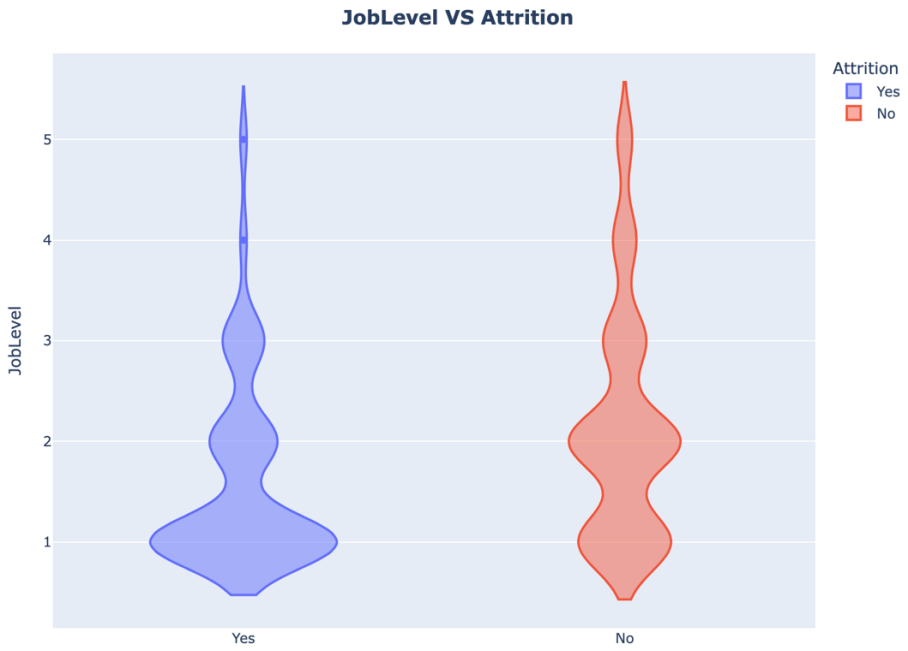

Figure 10

- **Monthly Income**: Employees who left mostly earned around 2.5k. *(Fig. 11)*

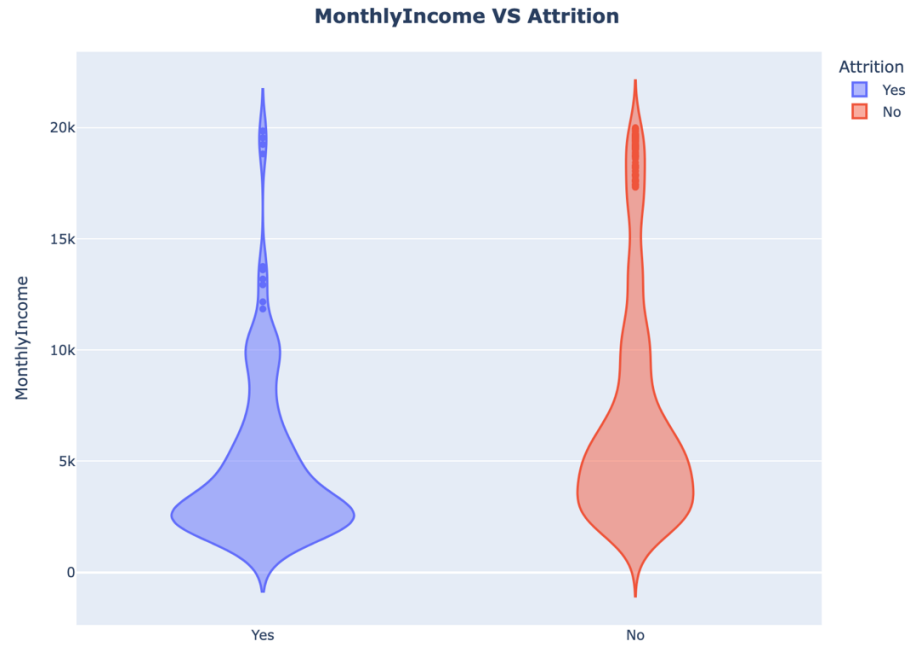

Figure 11

- **Stock Option Level**: Attrition is higher among employees with zero stock options. *(Fig. 12)*

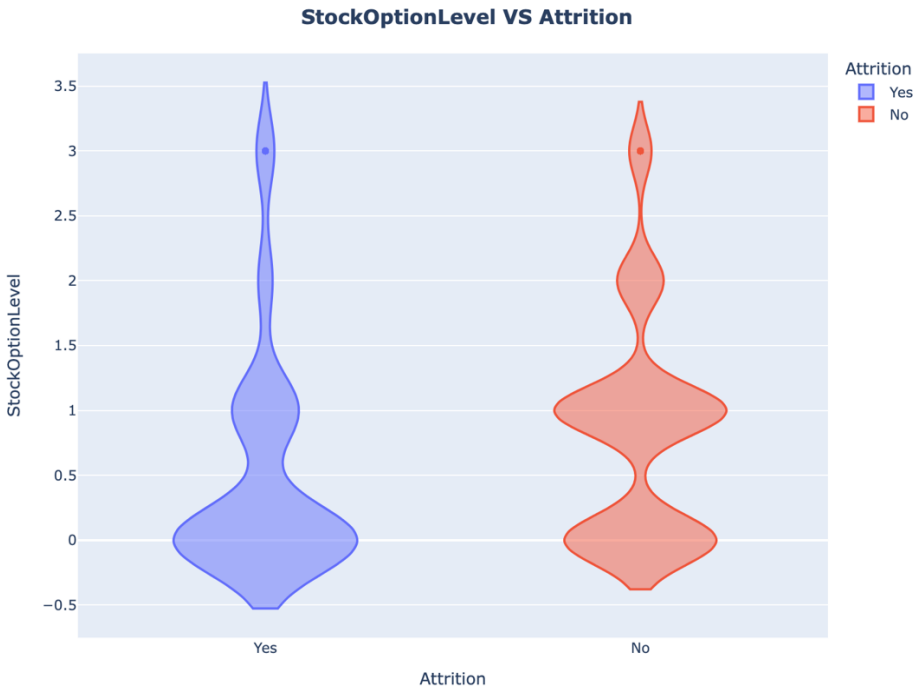

Figure 12

- **Total Working Years**: Employees with 0–10 years of experience are more likely to leave. *(Fig. 13)*

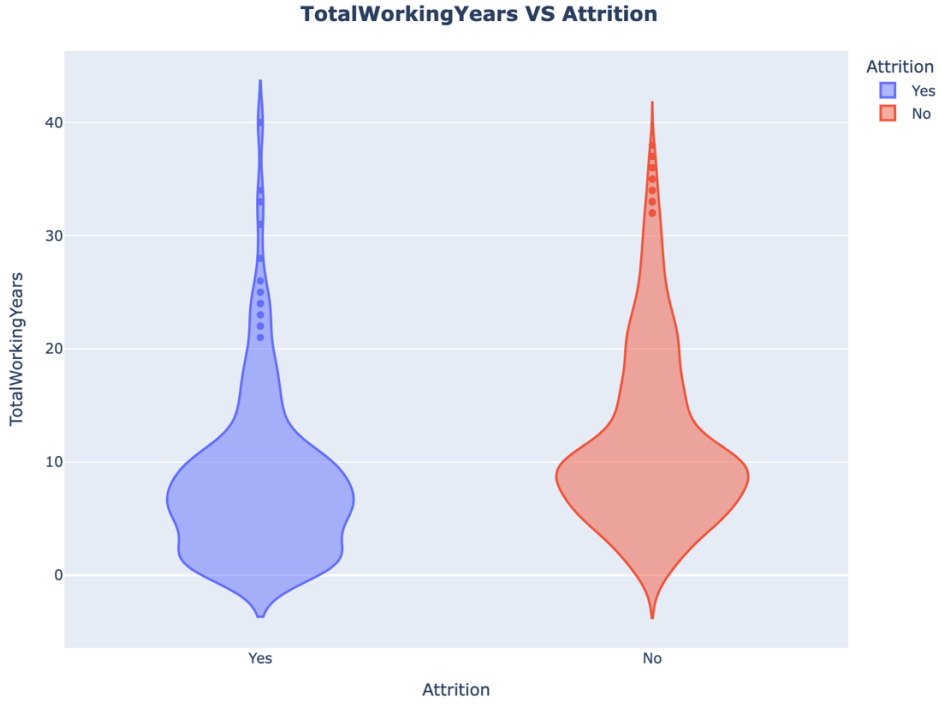

Figure 13

3.3 Outlier Detection and Handling

- Outliers in numerical features were identified and handled using the **Interquartile Range (IQR) method**. This approach ensures that extreme values do not distort the analysis or modeling. Outliers were treated by capping or removing values beyond 1.5 times the IQR from the first and third quartiles.

3.4 Correlation Analysis

- A correlation matrix was generated to analyze the relationships among numerical variables. Features with near-zero correlation with most other variables were dropped, including:
  EmployeeNumber, EmployeeCount, HourlyRate, JobSatisfaction, MonthlyRate, PercentSalaryHike,
  PerformanceRating, RelationshipSatisfaction, StandardHours, StockOptionLevel, TrainingTimesLastYear, and WorkLifeBalance. *(Fig. 15)*

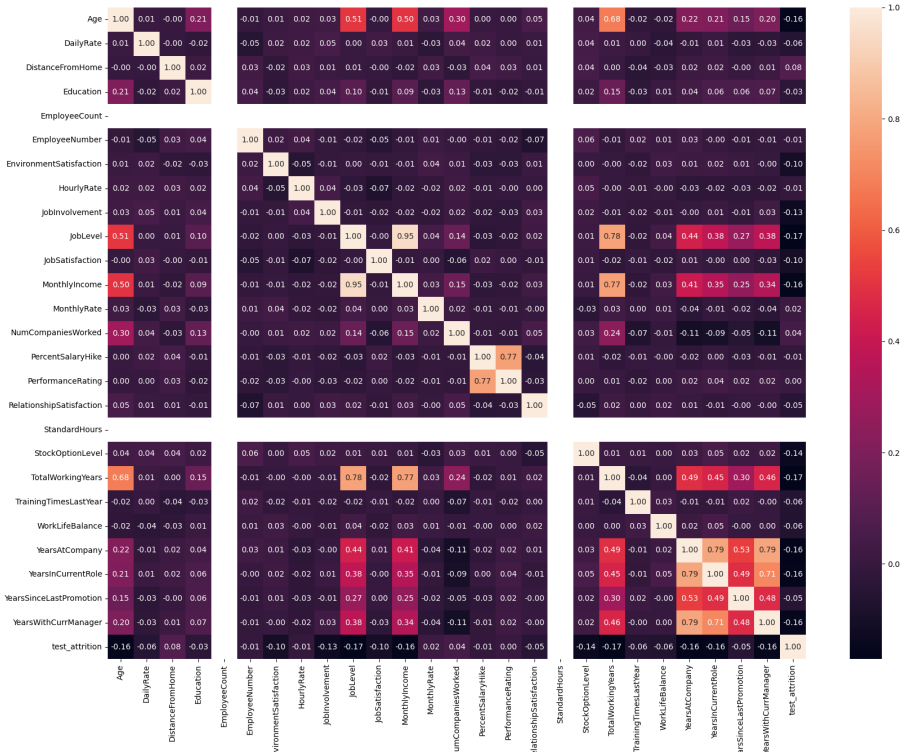

Figure 14

3.5 Missing Values

- The dataset did not contain any missing values, so no imputation was required.

4. Methodology

4.1. Data Preprocessing

The preprocessing steps ensured that the dataset was ready for machine learning algorithms:

Encoding:

- The target variable, Attrition, was mapped to binary values (Yes → 1, No → 0).
- The OverTime feature was also mapped to binary values (Yes → 1, No → 0).
- Categorical features were encoded using one-hot encoding via the pd.get_dummies function with drop_first=True to avoid multicollinearity.

Dropping Unnecessary Features:

- The Over18 column was removed, as it had no variability and did not contribute meaningful information to the analysis.

Feature Scaling:

- Numerical features were scaled using the MinMaxScaler to bring all feature values into the range [0, 1]. This ensured that features with different scales did not bias the model training.

4.2. Feature Engineering

No additional features were created in this analysis. The feature importance was later analyzed using the trained model and SHAP values to determine the most influential features.

4.3. Model Selection and Training

The **Random Forest Classifier** was selected for this project due to its ability to handle complex relationships, feature importance analysis, and robustness against overfitting.

Hyperparameter Tuning:

- A Grid Search with 5-fold cross-validation was performed to optimize hyperparameters.
- The grid search evaluated the following parameters:
  - Number of estimators (n_estimators): [50, 100, 200]
  - Maximum depth (max_depth): [10, 20, None]
  - Minimum samples required to split a node (min_samples_split): [2, 5, 10]
  - Minimum samples required at a leaf node (min_samples_leaf): [1, 2, 4]
  - Maximum features considered for splits (max_features): ['auto', 'sqrt', 'log2']

- The best parameters obtained were:
  - n_estimators: 50
  - max_depth: 10
  - min_samples_split: 5
  - min_samples_leaf: 1
  - max_features: sqrt

4.4. Evaluation

The final model was trained with the best hyperparameters obtained from the grid search. The evaluation metrics included:

- **Accuracy Score**: Provided an overall measure of correct predictions. The random forest classifier achieved an overall score of (0.877).
- **ROC-AUC Score**: Measured the ability of the classifier to distinguish between classes. The classifier achieved a roc-auc score of(0.54)
- **Classification Report**: Detailed the precision, recall, and F1-score for both classes.

	Precision	**Recall**	**F1-score**	**Support**
0	0.88	1.00	0.93	255
1	0.80	0.10	0.18	39
Accuracy			0.88	294
Macro avg	0.84	0.55	0.56	294
Weighted avg	0.87	0.88	0.83	294

- **Confusion Matrix**: Showed the distribution of true positives, true negatives, false positives, and false negatives.

	Predicted 0	Predicted 1
0	254 (TN)	1(FP)
1	35 (FN)	4 (TP)

4.5. Feature Importance Analysis

- Feature importances were extracted from the Random Forest model, revealing the top predictors for attrition.
- Additionally, SHAP (SHapley Additive exPlanations) was used to explain individual predictions and the overall impact of features. A SHAP summary plot provided insights into how features influenced the model's predictions(fig. 15).

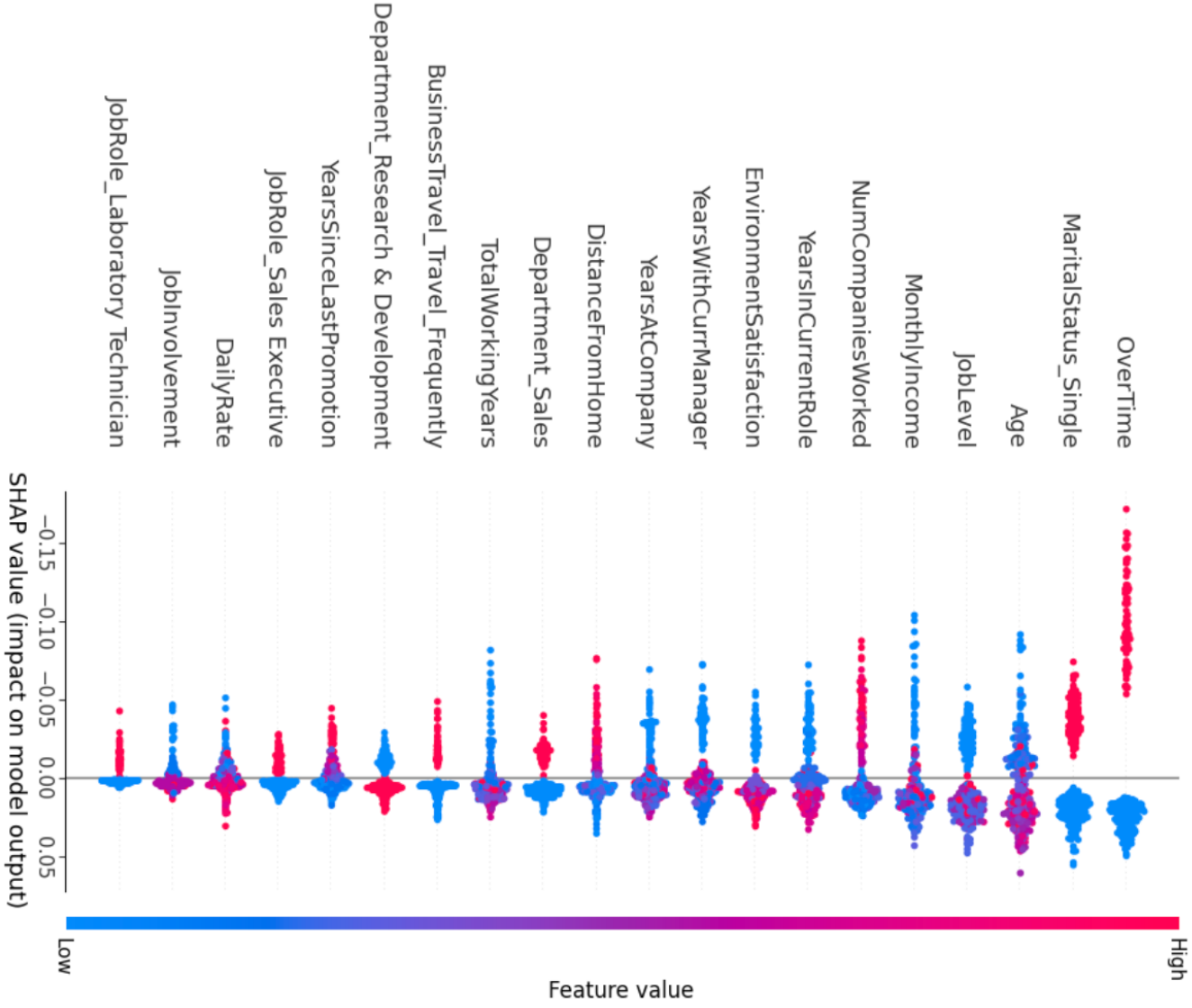

Figure 15

4.6. Tools and Frameworks

The following libraries and tools were used in this project:

- **Python** for programming.
- **Pandas** for data manipulation and preprocessing.
- **Scikit-learn** for model training, hyperparameter tuning, and evaluation.
- **SHAP** for feature importance and interpretability.

Results and Discussion

Model Performance

The Random Forest classifier achieved an **accuracy score of 87.76%** on the test set, indicating that the model effectively classifies employee attrition in most cases. However, the model's performance on the minority class (Attrition = 1) was suboptimal, as reflected in the **ROC-AUC score of 0.549**, suggesting limited discriminatory power for this class.

While the model performs exceptionally well for the majority class (Attrition = 0), the minority class suffers from low recall (10%), meaning it fails to identify most employees likely to attrite.

Feature Importance

The feature importance analysis revealed the top predictors of employee attrition:

1. **MonthlyIncome**: Employees with lower monthly incomes were more likely to attrite, suggesting financial dissatisfaction as a potential driver of attrition.
2. **JobRole**: Certain roles, such as Sales Representatives, showed significantly higher attrition rates (39%), indicating job-specific challenges.
3. **TotalWorkingYears**: Employees with fewer total working years were more prone to attrition, potentially reflecting dissatisfaction among newer or less experienced employees.
4. **OverTime**: Employees who worked overtime were more likely to leave, highlighting work-life balance issues.
5. **Age**: Younger employees, particularly around the age of 30, exhibited higher attrition rates, possibly indicating career transitions or dissatisfaction at early career stages.

SHAP Analysis

SHAP (Shapley Additive exPlanations) values provided further insights into how individual features influence model predictions:

- Employees with **low monthly income** consistently had higher probabilities of attrition.
- **Overtime work** was a key differentiator, with employees working overtime being more likely to attrite.
- **Job roles** such as Sales Representatives had a high positive impact on attrition predictions.

The SHAP summary plot (Fig. X) visually corroborates these findings, ranking features by their impact on model predictions and showing the distribution of SHAP values for each feature.

Insights and Interpretations

1. **Targeted Interventions**: High attrition rates among Sales Representatives (39%) and Laboratory Technicians (24%) suggest a need for role-specific retention strategies.
2. **Compensation Improvements**: The relationship between low monthly income and attrition indicates that competitive salaries could mitigate attrition risk.
3. **Work-Life Balance**: Employees working overtime were more likely to leave, highlighting the importance of policies promoting balance, such as flexible hours or reduced overtime demands.
4. **Support for Newer Employees**: The higher attrition rates among employees with fewer total working years point to the need for mentorship programs and career development opportunities.

Limitations

1. **Class Imbalance**: The dataset's significant imbalance between the majority (83.9% No) and minority (16.1% Yes) classes impacted the model's ability to predict attrition accurately for the minority class. Addressing this imbalance through techniques like SMOTE or cost-sensitive learning could improve performance.
2. **Limited Feature Scope**: The dataset lacks qualitative features, such as employee satisfaction surveys or feedback, which could provide a richer understanding of attrition drivers.
3. **Generalization**: The findings are specific to this dataset and may not generalize across industries or organizations without further validation.

Potential Improvements

1. **Addressing Class Imbalance**: Techniques like oversampling the minority class (e.g., SMOTE) or undersampling the majority class could improve recall for the minority class.
2. **Exploring Other Models**: Trying boosting algorithms like XGBoost or Gradient Boosting might enhance performance by capturing complex relationships in the data.
3. **Incorporating Qualitative Data**: Adding survey-based features, such as job satisfaction or intent to stay, could enrich the analysis.
4. **Feature Engineering**: Creating interaction terms or non-linear transformations of existing features could uncover hidden patterns.

These results and insights pave the way for data-driven strategies to reduce employee attrition and improve organizational retention policies.

Conclusion

In this project, we have thoroughly analyzed the HR-employee attrition dataset with the goal of identifying the primary reasons behind employee attrition and predicting the likelihood of an employee's departure. Through extensive exploratory data analysis (EDA), we discovered that several key factors, such as job satisfaction, work-life balance, and career development opportunities, strongly influence employee retention.

By applying classification techniques, we trained a model capable of predicting employee attrition based on various predictors. The results showed that our model could effectively identify employees at risk of leaving the company, providing actionable insights to HR departments for proactive interventions.

This study highlights the importance of understanding the underlying reasons for employee attrition, which can lead to improved retention strategies, reduced turnover costs, and a more positive work environment. Future work could explore additional features and apply more advanced modeling techniques to further enhance the predictive accuracy of the model.

Overall, the findings of this project demonstrate the potential of data science techniques in solving real-world business challenges, such as employee retention, and provide valuable insights into how organizations can optimize their workforce management strategies.

References

- Rushikesh Konapure. (n.d). HR-Employee Attrition Dataset. Retrieved from https://www.kaggle.com/datasets/rishikeshkonapure/hr-analytics-prediction
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

- Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*. Retrieved from https://github.com/slundberg/shap

- Waskom, M. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021

- Plotly Technologies Inc. (2015). Collaborative Data Science. Montréal, QC. Retrieved from https://plotly.com/python/