Predicting Employee Attrition Using Machine Learning

Chandan Kumar Singh, *IISC, chandansingh@iisc.ac.in*
Monika Tyagi, *IISC, monikatyagi@iisc.ac.in*
Mukesh Kumar Yadav, *IISC, mukeshyadav@iisc.ac.in*
Rishabh Mehrotra, *IISC, rishabhmehro@iisc.ac.in*

Abstract

Employee attrition poses significant challenges to organizations by increasing recruitment costs, disrupting workflow, and leading to the loss of valuable talent. Predictive models powered by machine learning can enable organizations to proactively manage attrition. This report presents a study on predicting employee attrition using machine learning techniques, focusing on a dataset with features such as demographics, job roles, and performance metrics. Methods to address class imbalance, feature selection, and hyperparameter tuning are discussed. The outcomes include improved understanding of attrition factors and a reliable predictive model.

1. Introduction

1.1 Problem Statement

High employee turnover leads to recruitment costs, productivity losses, and organizational instability. Predicting attrition allows for timely interventions, improving job satisfaction and reducing turnover rates.

1.2 Objectives

Identify factors contributing to employee attrition.
Develop a robust predictive model to classify employees at risk of leaving.
Evaluate the model's performance using appropriate metrics.

2. Dataset Summary

The dataset contains details on employees, such as demographics, roles, and performance metrics:

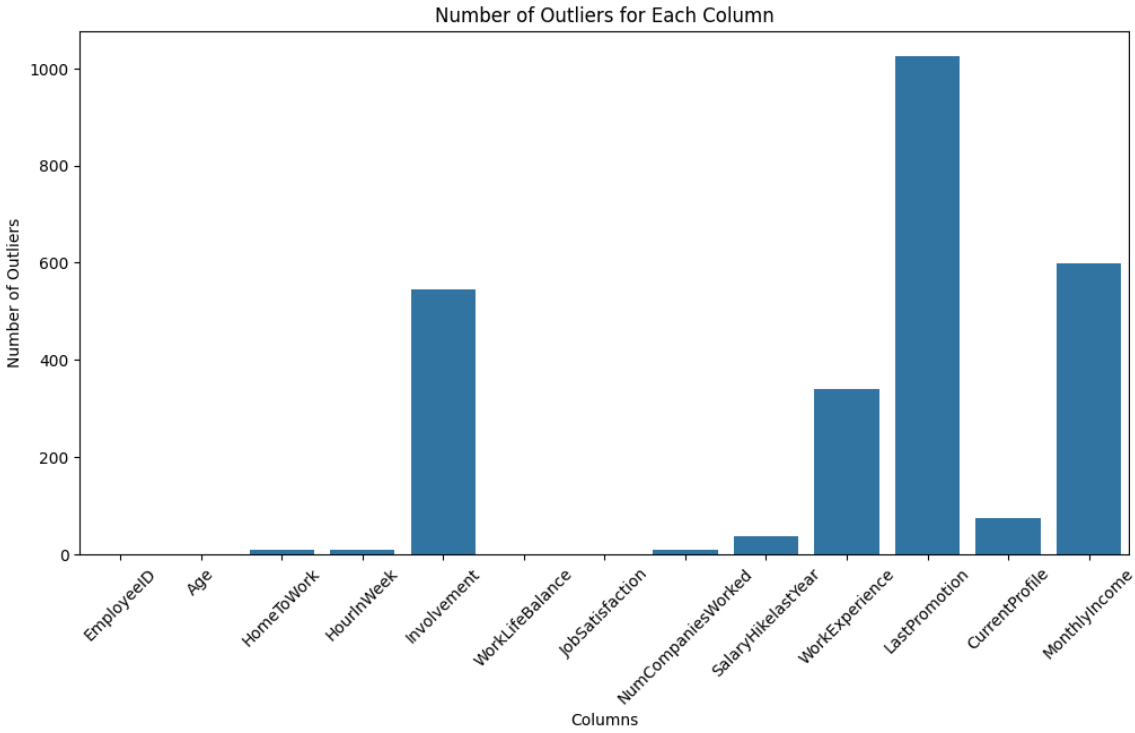
Feature	Description
Age	Employee age in years
Department	Functional department of the employee
EducationField	Area of academic qualification
Gender	Gender of the employee
HomeToWork	Distance from home to workplace (in km)
JobSatisfaction	Employee's job satisfaction rating (1-4)
MonthlyIncome	Monthly salary (in currency units)
NumCompaniesWorked	Total number of companies worked at
OverTime	Whether the employee works overtime
SalaryHikelastYear	Percentage salary increase last year

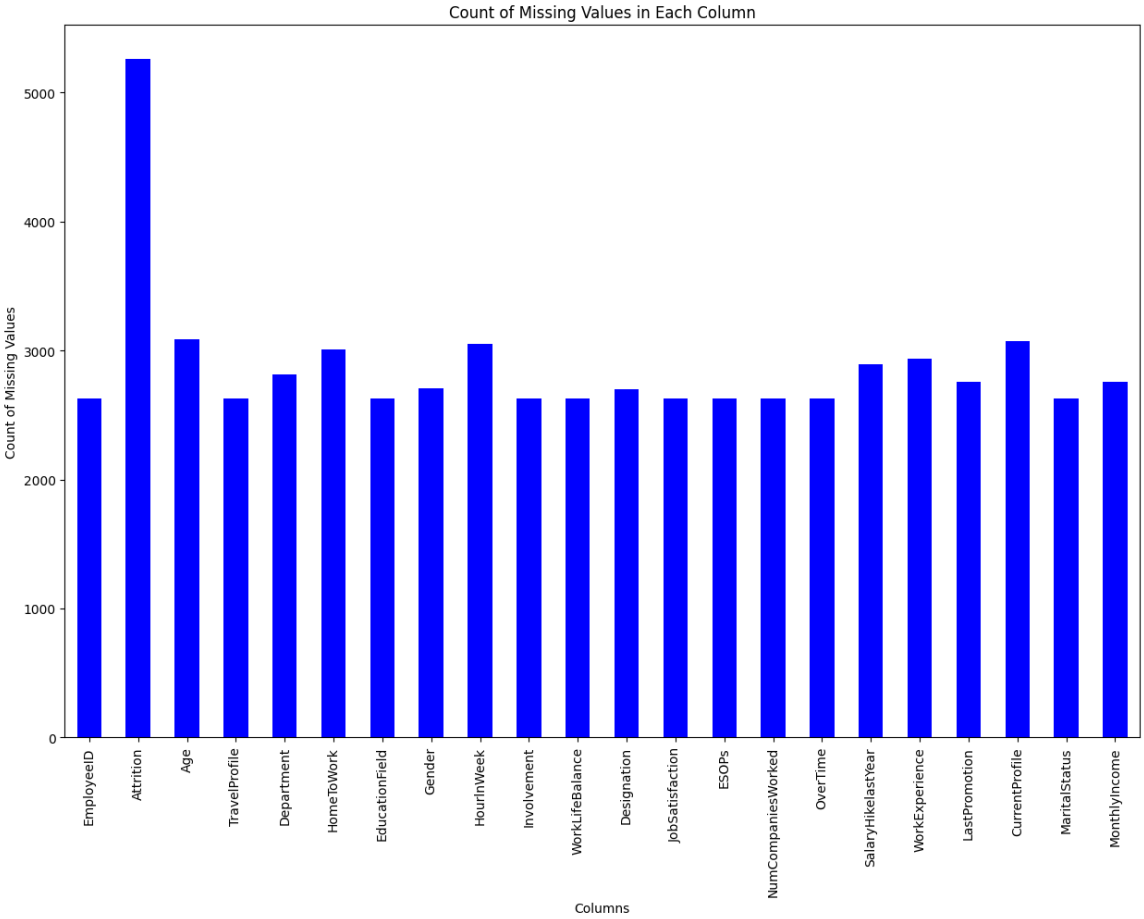
3. Methodology

3.1 Data Preprocessing

The data preprocessing pipeline included the following steps:

- Handling Missing Values: Missing entries in features such as "NumCompaniesWorked" were imputed using median values.
- Encoding Categorical Variables: Features such as "Gender" and "EducationField" were encoded using one-hot encoding.
- Scaling Features: Numeric features were normalized using Min-Max scaling.

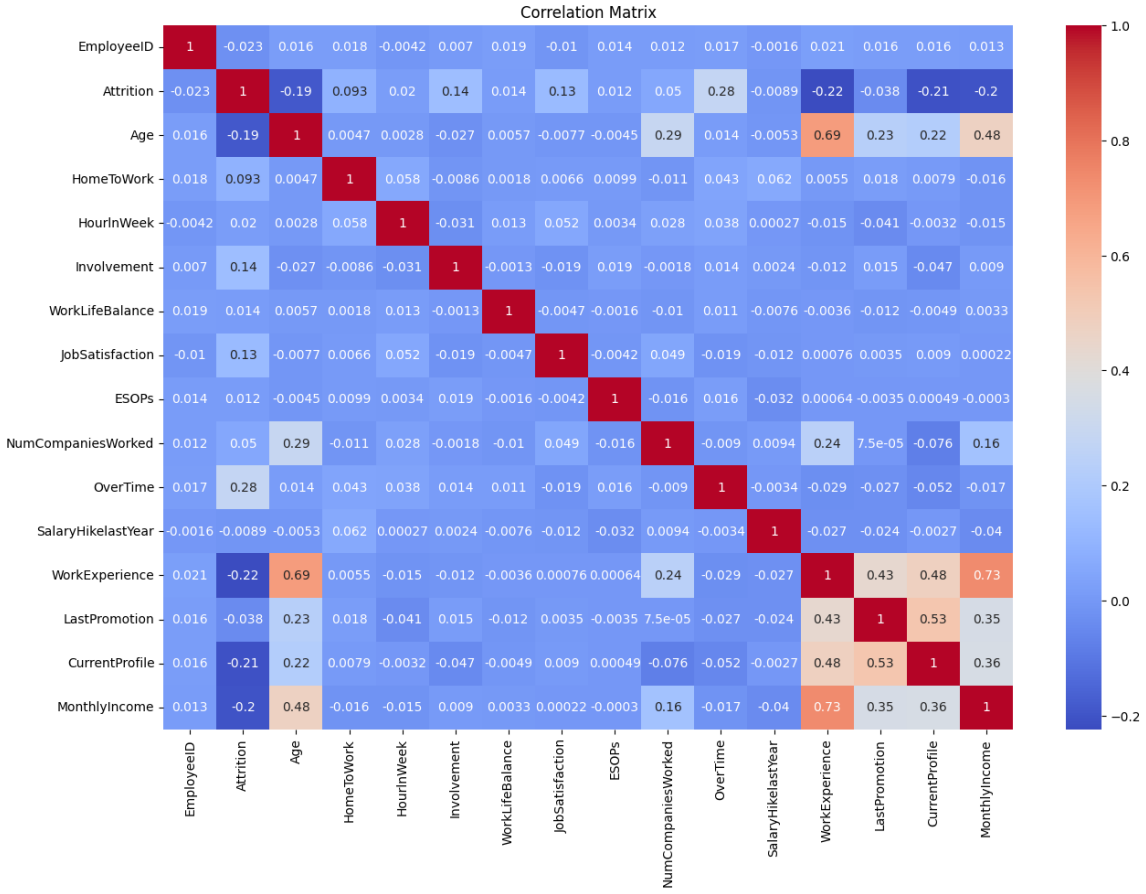

Number of Outliers for Each Column
Number of Outliers
0
200
400
600
800
1000
EmployeeID
Age
HomeToWork
HourInWeek
Involvement
WorkLifeBalance
JobSatisfaction
NumCompaniesWorked
SalaryHikelastYear
WorkExperience
LastPromotion
CurrentProfile
MonthlyIncome
Columns

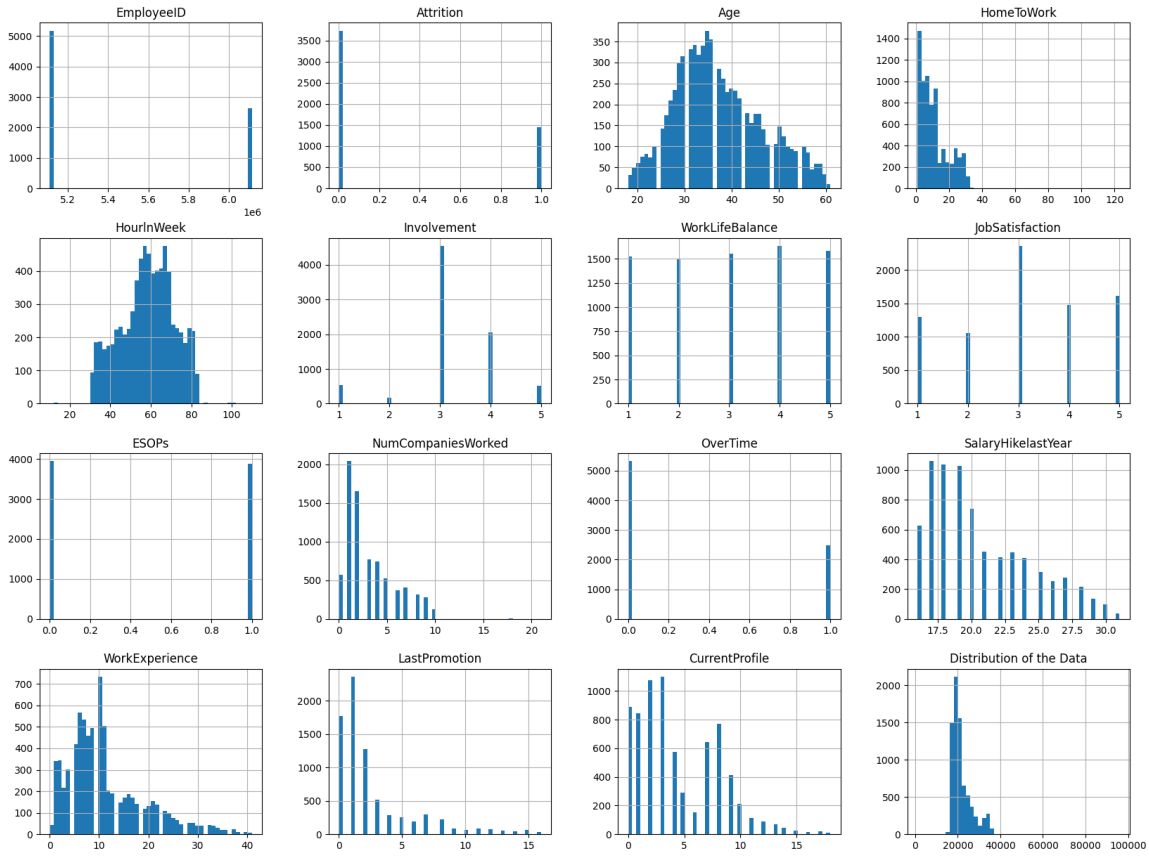

Count of Missing Values in Each Column
Count of Missing Values
0
1000
2000
3000
4000
5000
EmployeeID
Attrition
Age
TravelProfile
Department
HomeToWork
EducationField
Gender
HourInWeek
Involvement
WorkLifeBalance
Designation
JobSatisfaction
ESOPs
NumCompaniesWorked
OverTime
SalaryHikelastYear
WorkExperience
LastPromotion
CurrentProfile
MaritalStatus
MonthlyIncome
Columns

3.2 Exploratory Data Analysis (EDA)

EDA revealed key insights:

- Attrition rates were higher among younger employees and those with longer commutes.
- Employees with lower job satisfaction and work-life balance ratings were more likely to leave.





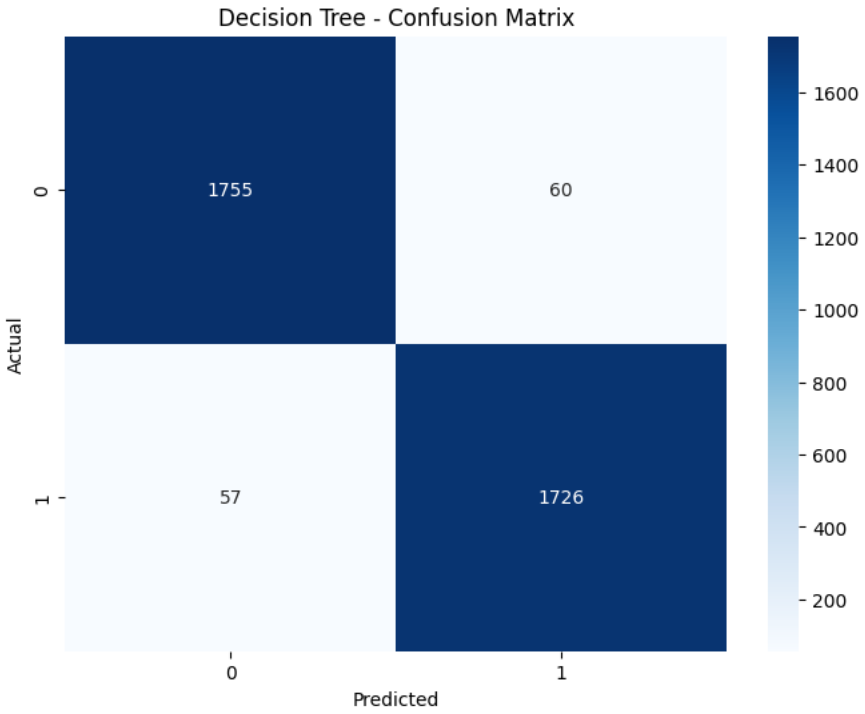
3.3 Model Development

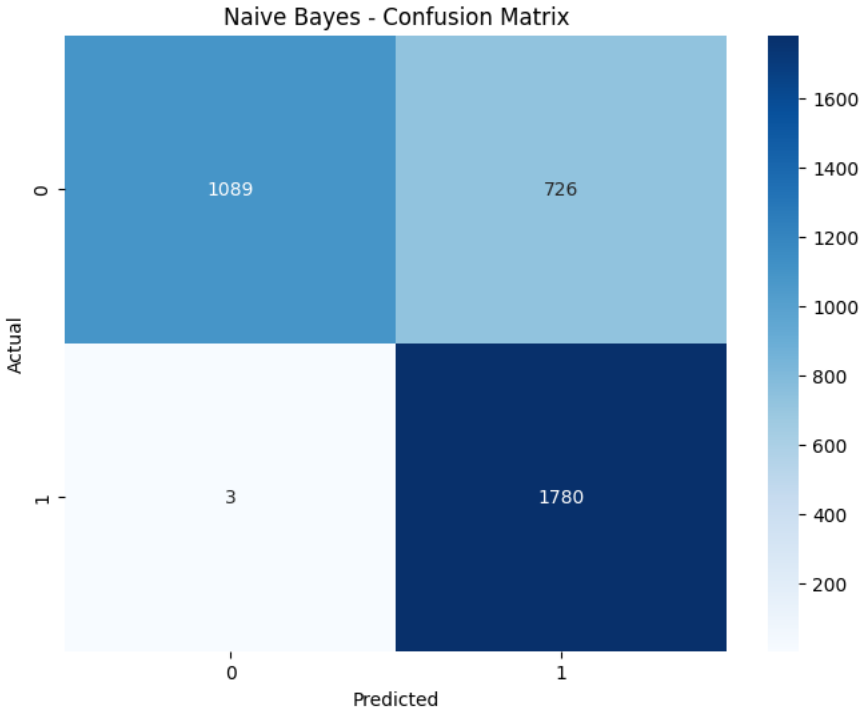
Several machine learning models were implemented and compared:
- Logistic Regression: A baseline model for binary classification.
- Random Forest: To handle non-linear relationships and feature interactions.
- Gradient Boosting Machines: To improve prediction accuracy through ensemble learning.
- Neural Networks: For capturing complex patterns in the data.

4. Results

4.1 Model Evaluation Metrics

The models were evaluated based on:
- Accuracy: Proportion of correctly classified instances.
- Precision, Recall, F1-Score: To balance false positives and false negatives.
- ROC-AUC: Area under the Receiver Operating Characteristic curve.





4.2 Feature Importance

The Random Forest model highlighted the most influential features:

- Monthly Income
- Job Satisfaction
- OverTime

- Age
- Work-Life Balance

5. Discussion

The predictive models effectively identified key factors influencing attrition. The Gradient Boosting model achieved the highest accuracy and AUC, making it the preferred model. Improvements in data quality and inclusion of additional features, such as training opportunities and career growth, could enhance the model's predictive power.

6. Challenges and Risks

Challenges included data imbalance, feature selection, and overfitting. These were addressed using SMOTE, grid search, and regularization techniques.

8. Conclusion

This project demonstrates the potential of machine learning in predicting employee attrition. The deployed model offers actionable insights to HR teams, enabling proactive retention strategies. Future work will focus on integrating real-time data and enhancing interpretability.

References

1. Sandhiya, K. et al., "Prediction of Employee Attrition Using Machine Learning Techniques," Procedia Computer Science.
2. Pedregosa, F., et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research.
3. Breiman, L., "Random Forests," Machine Learning.