EMPLOYEE ATTRITION ANALYSIS

Rajesh Nalliboyina

Department of Information Systems and Business Analytics

Kent State University

Internship Data Analytics (BA-64092-001)

Rouzbeh Razavi, PhD

UNIFIED MENTOR

15-01-2025 to 15-04-2025

Problem Statement

During my internship at Green Destinations, the primary problem was to analyze employee attrition to understand why employees were leaving the organization and to identify actionable strategies to reduce turnover. Employee attrition, defined as the rate at which employees voluntarily leave, posed significant challenges for Green Destinations, including increased recruitment and training costs, loss of institutional knowledge, and reduced workforce morale. The goal was to calculate the overall attrition rate, identify demographic and job-related factors driving turnover, build a predictive model to estimate the likelihood of an employee leaving, and provide data-driven recommendations to enhance retention.

The specific objectives included:

- Determining the overall attrition rate for the organization.
- Identifying high-risk groups (e.g., specific age groups, departments, or job roles) with elevated turnover rates.
- Pinpointing key factors (e.g., income, overtime, job satisfaction) influencing attrition.
- Developing a predictive model to identify at-risk employees.
- Creating a dashboard to visualize findings for stakeholders.
- Recommending strategies to improve employee retention.

Business Value of Solving the Problem

Addressing employee attrition offers significant business value for Green Destinations:

- **Cost Reduction**: High turnover increases recruitment costs (e.g., advertising, interviewing, onboarding) and training expenses. Reducing attrition lowers these costs, allowing budget reallocation to strategic initiatives.
- **Preservation of Institutional Knowledge**: Retaining experienced employees maintains organizational expertise, reducing the knowledge gap caused by frequent departures.
- **Improved Productivity and Morale**: High turnover can demoralize remaining employees, leading to decreased productivity. Lowering attrition fosters a stable, engaged workforce.
- **Enhanced Reputation**: A low attrition rate signals a positive work environment, aiding in attracting top talent in a competitive job market.
- **Data-Driven HR Strategies**: Identifying attrition drivers enables targeted interventions, such as tailored retention programs for high-risk groups, improving HR efficiency.

By solving this problem, Green Destinations can achieve cost savings, maintain a skilled workforce, and strengthen its organizational stability, positioning it for long-term success.

ABSTRACT

This project, conducted as part of an internship at Green Destinations, aimed to analyze employee attrition to identify key factors influencing turnover and provide actionable recommendations for improving retention. Using a dataset of 1,470 employees, we performed exploratory data analysis (EDA), statistical testing, and predictive modeling to understand attrition patterns. Key findings include an overall attrition rate of 17.11%, with younger employees (25.40% attrition), Sales and HR departments (21.38% and 22.22%), and specific roles like Research Scientists (62.70%) and Sales Representatives (39.79%) showing the highest turnover. Factors such as age, years at the company, and monthly income significantly influence attrition, with younger, newer, and lower-paid employees more likely to leave. A logistic regression model achieved 83.82% accuracy in predicting attrition, and a Random Forest model highlighted overtime, job satisfaction, and work-life balance as critical factors. Recommendations include targeted retention strategies for high-risk groups, improving non-monetary factors like job satisfaction, and learning from low-attrition roles. The analysis was visualized in a dashboard, which was exported as a PDF for stakeholder reporting.

INTRODUCTION

Employee attrition, the rate at which employees leave an organization, is a critical metric for HR management. High turnover can lead to increased recruitment costs, loss of institutional knowledge, and decreased morale. Green Destinations, a company with 1,356 employees (after outlier removal), sought to understand the drivers of attrition to improve retention strategies. This project aims to:

- Calculate the overall attrition rate.
- Identify demographic and job-related factors influencing attrition.
- Build a predictive model to estimate the probability of an employee leaving.
- Provide data-driven recommendations to reduce turnover.

The analysis leverages a dataset containing 35 features, including age, department, job role, monthly income, and more. The dashboard titled "Employee Attrition Analysis" visualizes key findings, while the Python code details the data processing and modeling steps.

Project Contribution

As the sole data analyst intern on this project, I was responsible for the end-to-end execution of the employee attrition analysis. My contributions included:

- **Data Preprocessing**: Conducted initial exploration, removed outliers, and engineered features to prepare the dataset for analysis.
- **Exploratory Data Analysis**: Designed and executed EDA, creating visualizations (e.g., histograms, bar charts, heatmaps) to uncover attrition patterns.
- **Statistical Analysis**: Performed t-tests, chi-square tests, and correlation analysis to identify significant factors influencing attrition.
- **Predictive Modeling**: Developed and evaluated Logistic Regression and Random Forest models, selecting features and interpreting results.

- **Dashboard Development**: Created a comprehensive dashboard to visualize findings, ensuring accessibility for non-technical stakeholders.
- **Recommendation Formulation**: Translated analytical insights into actionable recommendations, such as targeting high-risk groups and addressing non-monetary factors.
- **Documentation**: Documented the entire process, including Python code, in a Google Colab notebook and prepared a detailed report for stakeholders.

Although I worked independently, I collaborated with my internship mentor at Unified Mentor and HR team to align the analysis with organizational goals and validate findings.

METHODOLOGY (Approach Taken to Address the Issue)

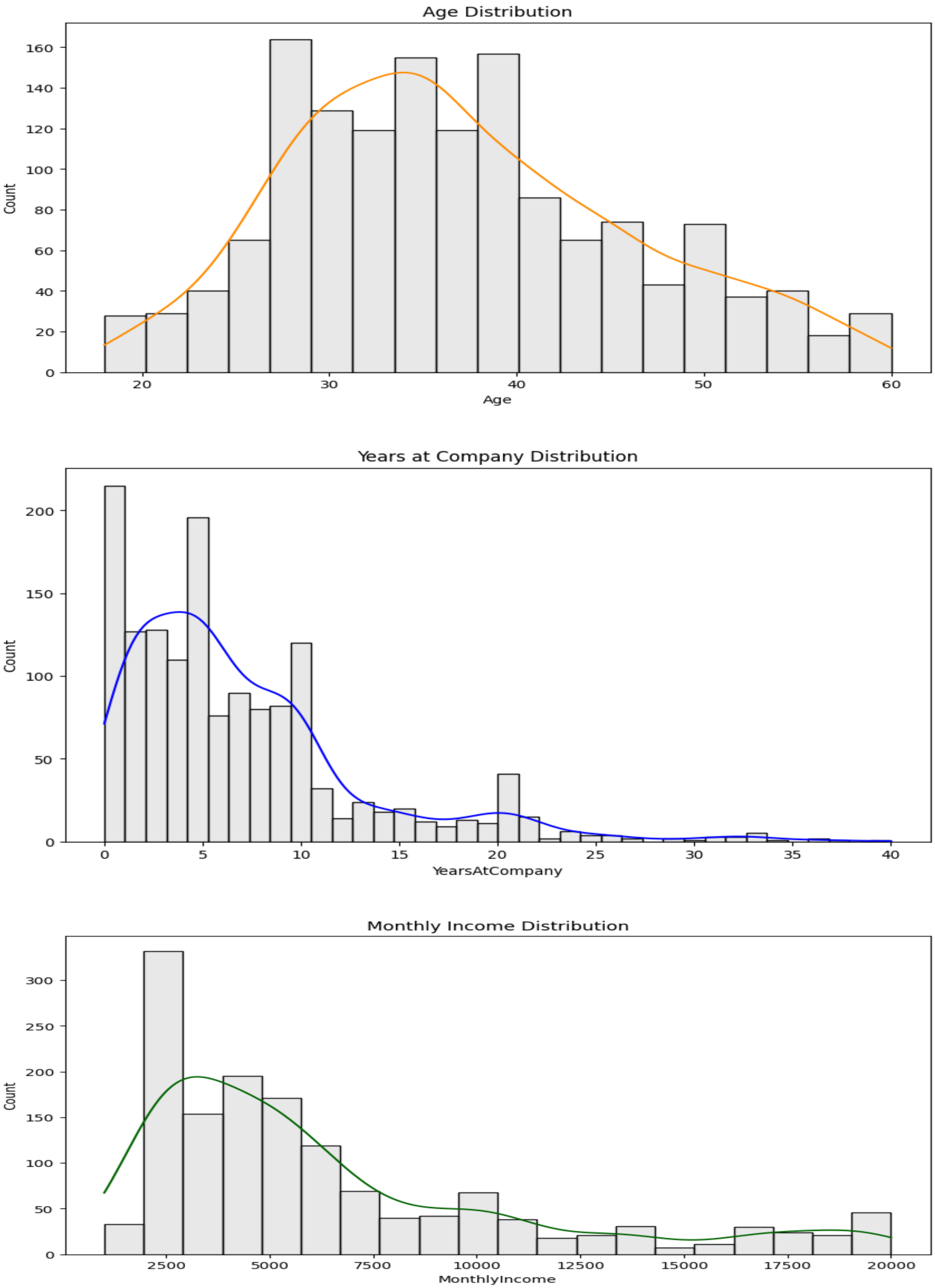
Data Collection

The dataset containing 1,470 rows and 35 columns, was uploaded to Google Colab and loaded using pandas for analysis. It includes features such as Age, Attrition, BusinessTravel, DailyRate, Department, DistanceFromHome, Education, EducationField, EmployeeNumber, EnvironmentSatisfaction, Gender, HourlyRate, JobInvolvement, JobLevel, JobRole, JobSatisfaction, MonthlyIncome, MonthlyRate, NumCompaniesWorked, Over18, OverTime, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, and YearsWithCurrManager, providing a comprehensive view of employee attributes for the attrition analysis.

Data Preprocessing

Initial Exploration

The initial exploration confirmed the dataset’s integrity by identifying no missing values across all columns, noting 26 integer columns (e.g., Age, MonthlyIncome) and 9 object columns (e.g., Attrition, Department), and verifying no duplicate EmployeeNumbers or negative values in Age, YearsAtCompany, or MonthlyIncome, ensuring the data was suitable for further analysis.

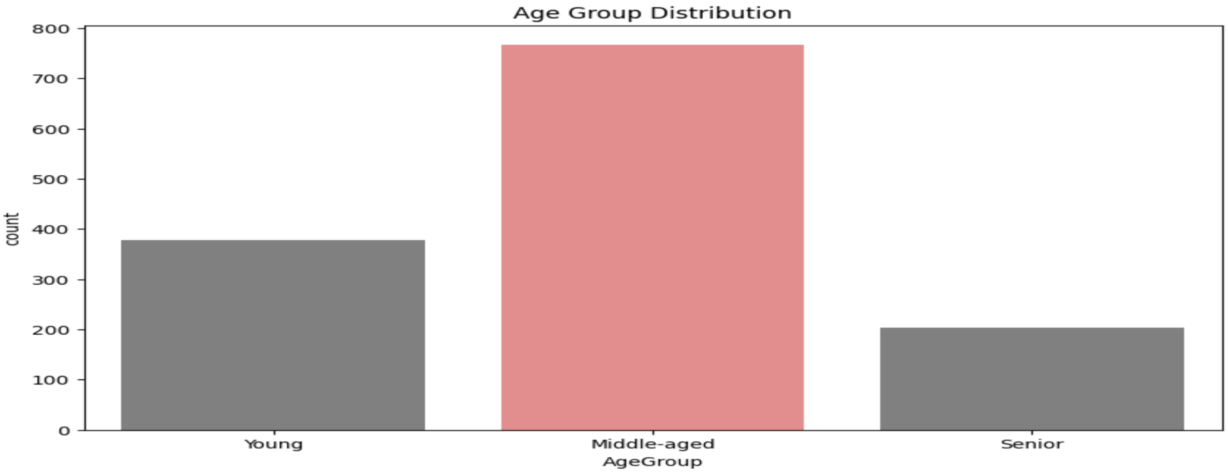

Age Distribution
Count
0
20
40
60
80
100
120
140
160
20
30
40
50
60
Age
Years at Company Distribution
Count
0
50
100
150
200
0
5
10
15
20
25
30
35
40
YearsAtCompany
Monthly Income Distribution
Count
0
50
100
150
200
250
300
2500
5000
7500
10000
12500
15000
17500
20000
MonthlyIncome

Outlier Removal

Outliers in MonthlyIncome were addressed using the Interquartile Range (IQR) method, where Q1 (25th percentile) and Q3 (75th percentile) defined a lower bound of 0 and an upper bound of $10,501, resulting in the removal of rows with MonthlyIncome exceeding $10,501 and reducing the dataset from 1,470 to 1,356 rows to improve the reliability of subsequent analyses.

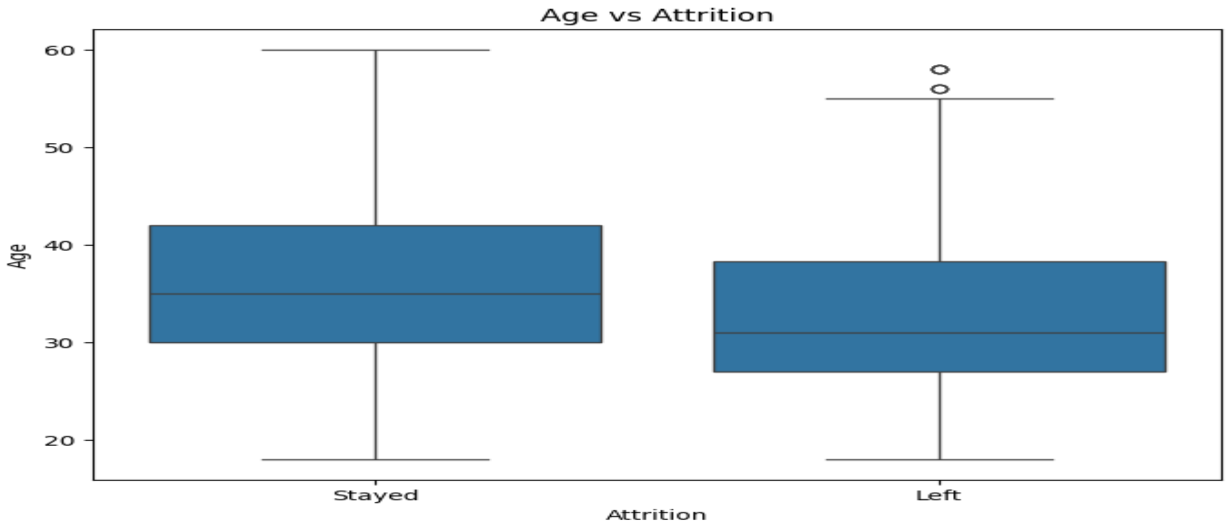
Feature Engineering

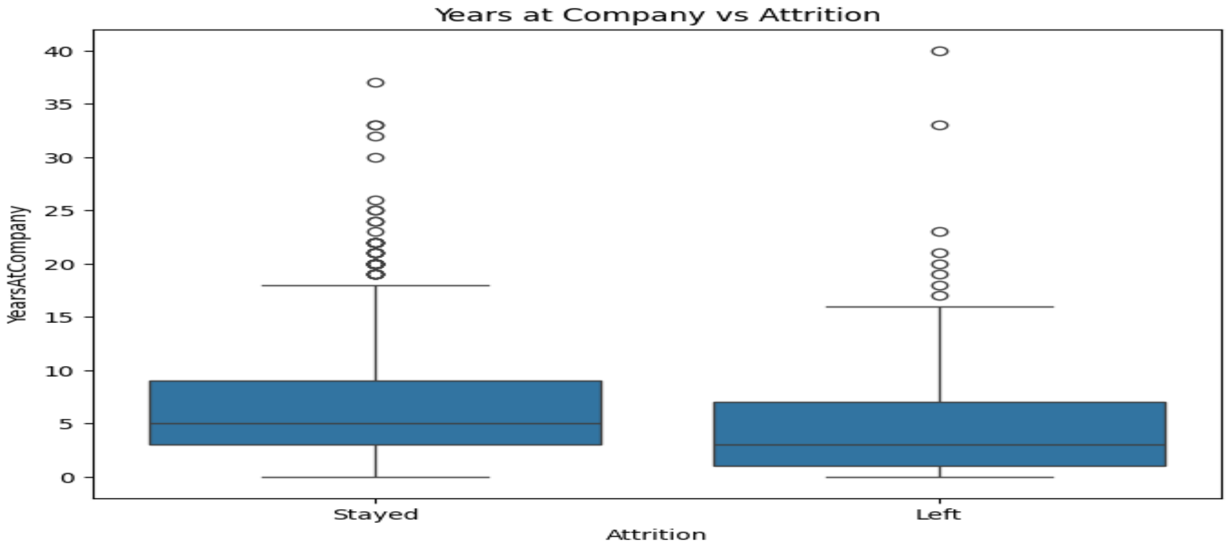
Feature engineering was performed to prepare the dataset for modeling by converting Attrition to binary (Yes: 1, No: 0), creating an AgeGroup column to categorize employees into Young (18-30), Middle-aged (30-45), and Senior (45-60) groups, and transforming OverTime into a binary format (Yes: 1, No: 0), enabling numerical analysis and predictive modeling.

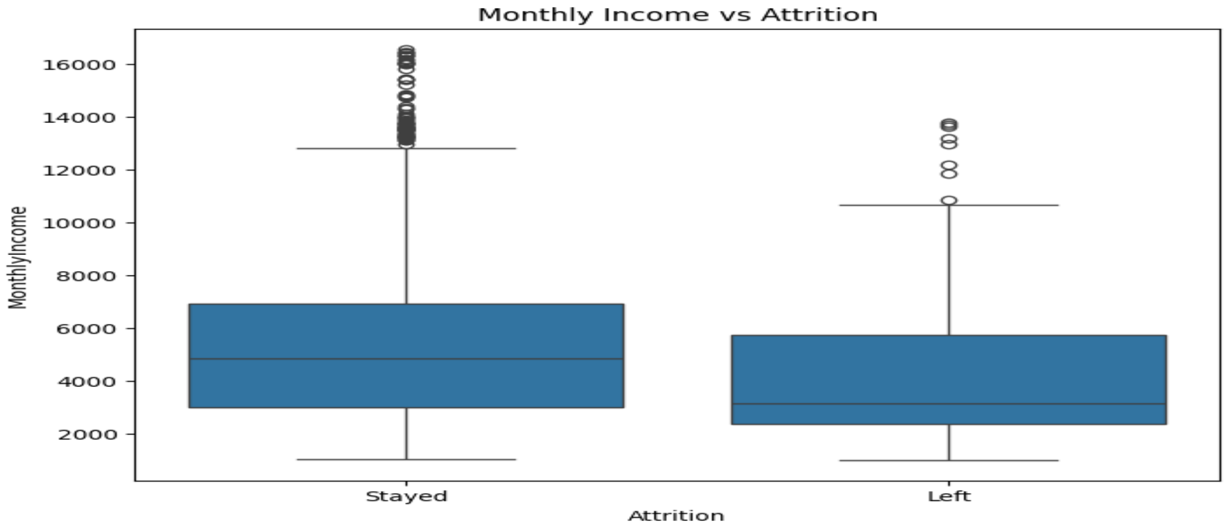


Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) involved visualizing the distributions of Age, YearsAtCompany, and MonthlyIncome using histograms to understand their shapes and ranges, analyzing attrition rates across AgeGroup, Department, and JobRole through bar charts and heatmaps, conducting t-tests and chi-square tests to identify significant factors influencing attrition, and computing correlations between Attrition, Age, YearsAtCompany, and MonthlyIncome to uncover underlying patterns.

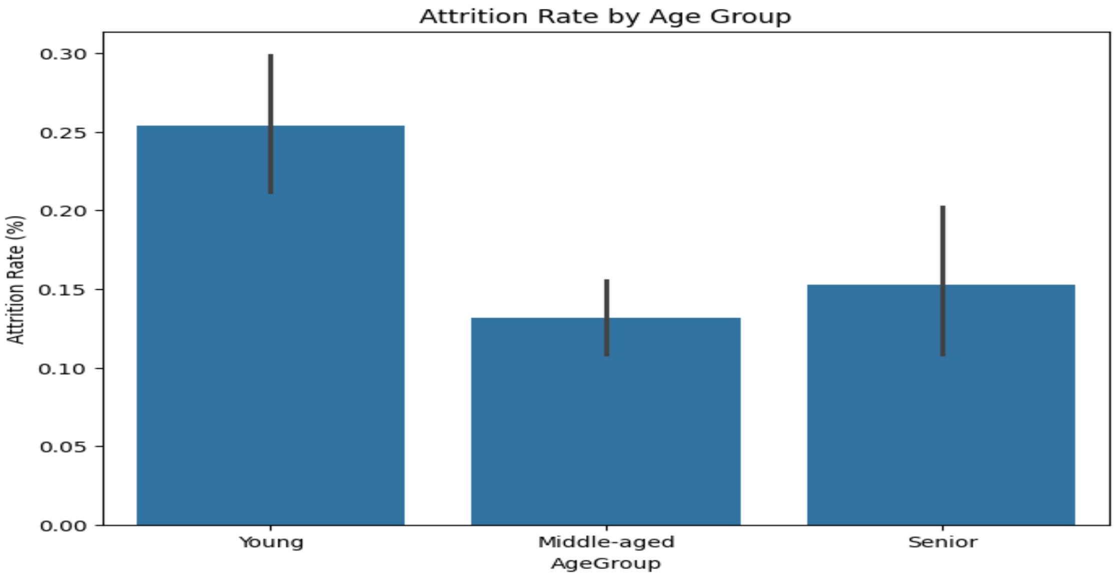

Age vs Attrition
60
50
40
30
20
Age
Stayed
Left
Attrition


Years at Company vs Attrition
40
35
30
25
20
15
10
5
0
YearsAtCompany
Stayed
Left
Attrition


Monthly Income vs Attrition
16000
14000
12000
10000
8000
6000
4000
2000
MonthlyIncome
Stayed
Left
Attrition

Predictive Modeling

Predictive modeling was conducted using a Logistic Regression model with features Age, YearsAtCompany, MonthlyIncome, OverTime, JobSatisfaction, and WorkLifeBalance, achieving an accuracy of 83.82% and evaluated via a confusion matrix and classification report, while a Random Forest model was also employed to determine feature importance, identifying OverTime as the most critical factor, with all modeling tasks implemented in Google Colab using pandas, numpy, matplotlib, seaborn, scipy, and scikit-learn.



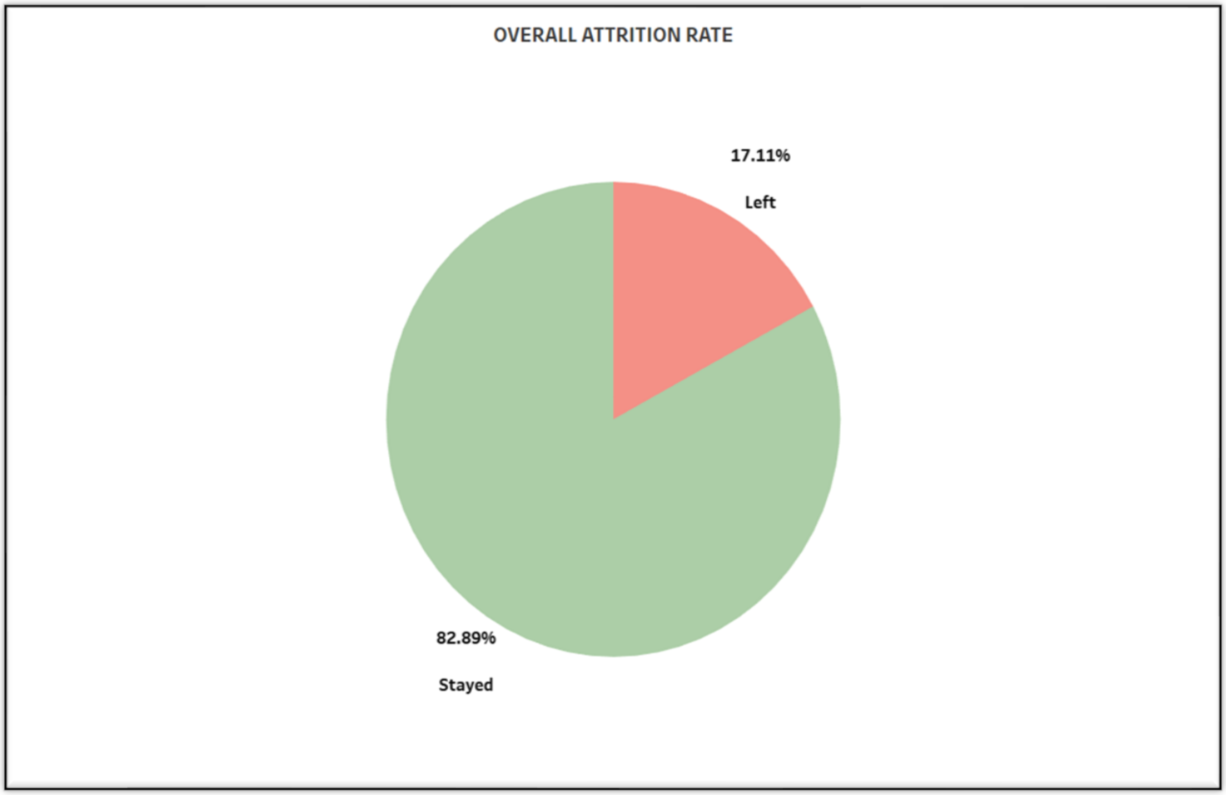
Tools and Libraries

- **Python Libraries**: pandas, numpy, matplotlib, seaborn, scipy, scikit-learn
- **Environment**: Google Colab

RESULTS AND DISCUSSION

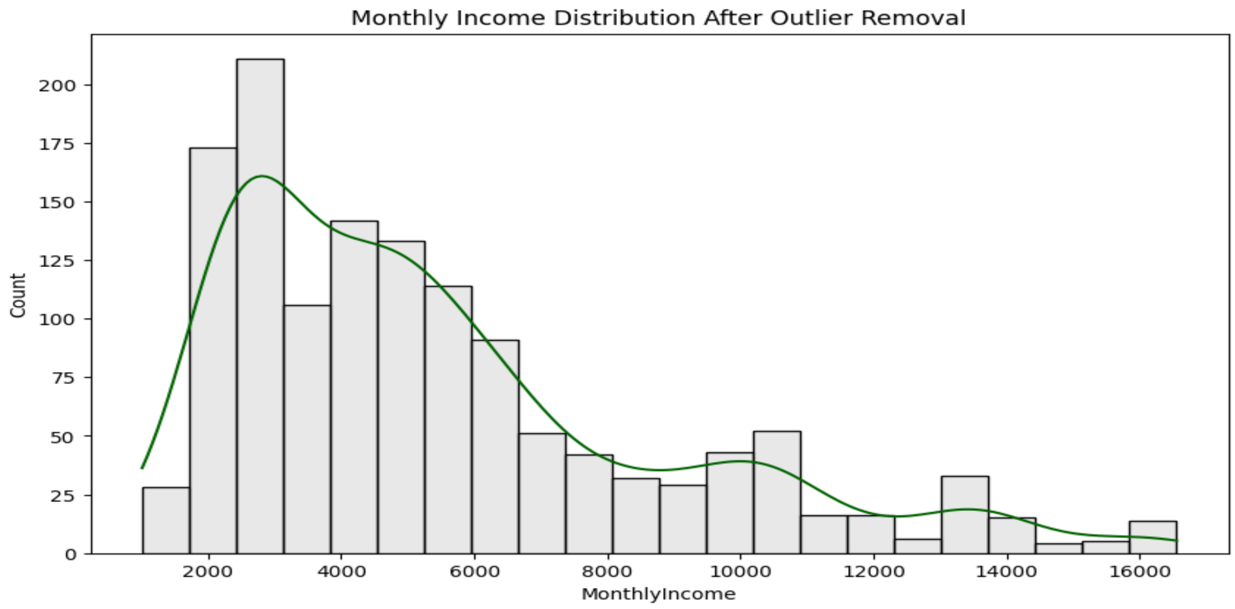
Overall Attrition Rate

The overall attrition rate after outlier removal was calculated at 17.11%, with 1,356 total employees, of whom 232 left and 1,124 stayed, as depicted in the dashboard's top-left pie chart showing 82.89% stayed (green) and 17.11% left (red), indicating a moderate but significant turnover rate where nearly 1 in 5 employees leave, potentially impacting productivity and increasing costs.



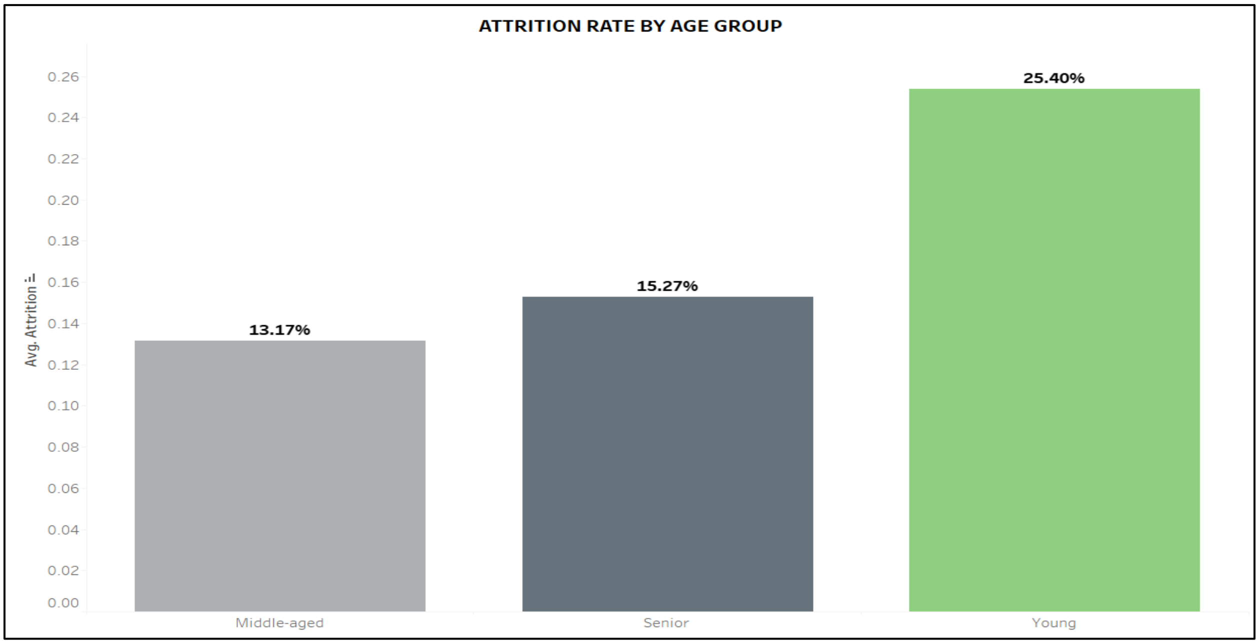
Distribution Analysis

The distribution analysis revealed key patterns in employee data, with Age showing a bell-shaped, slightly right-skewed distribution ranging from 18 to 60 years and peaking around 35 years, indicating most employees are between 25 and 45 years old with fewer under 25 or over 50, while Years at Company exhibited a heavily right-skewed distribution ranging from 0 to 40 years and peaking at 0-2 years, suggesting many employees are new with a sharp drop-off after 10 years, and Monthly Income, also right-skewed, ranged from $1,000 to $20,000 before outlier removal with a peak at $2,500-$5,000, reduced to $1,000-$10,501 after outlier removal as shown in the histogram with a KDE line highlighting the peak, focusing on lower to moderate incomes.



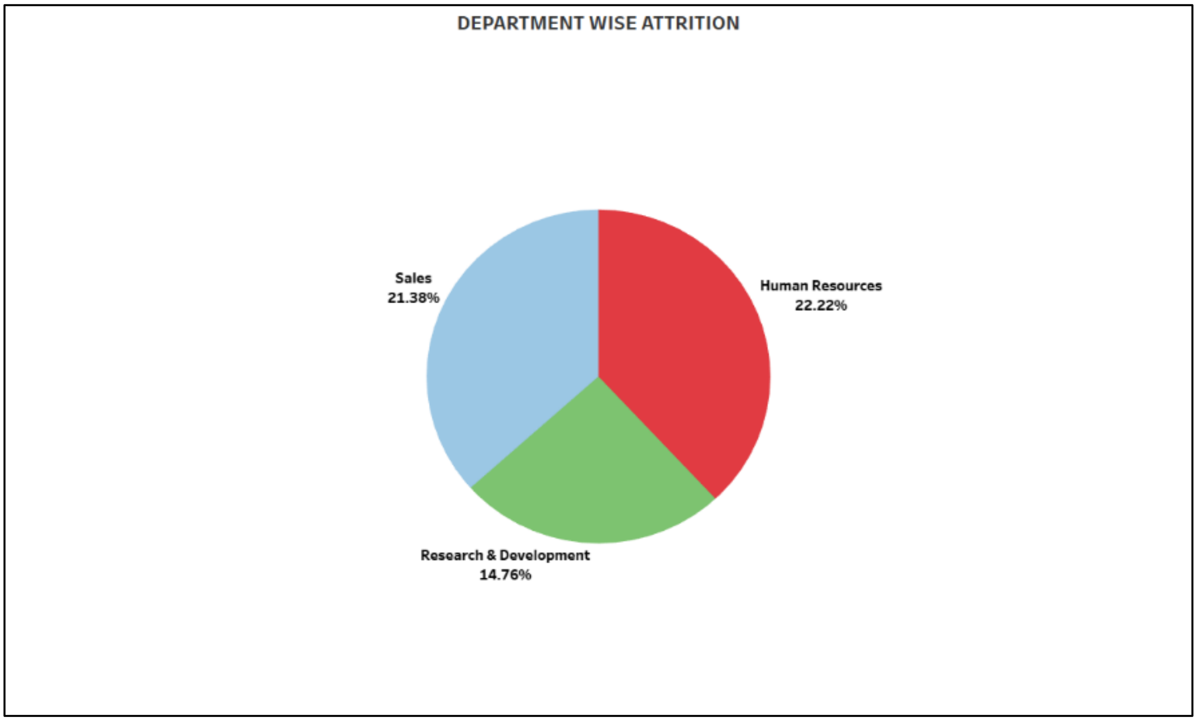
Attrition by Age Group

Attrition by age group showed significant variation, with Young employees (18-30) having the highest attrition rate at 25.40%, followed by Senior employees (45-60) at 15.27%, and Middle-aged employees (30-45) at 13.17%, as illustrated in the dashboard's middle-left bar chart, suggesting that younger employees are most likely to leave, possibly due to seeking better opportunities or lacking long-term commitment.



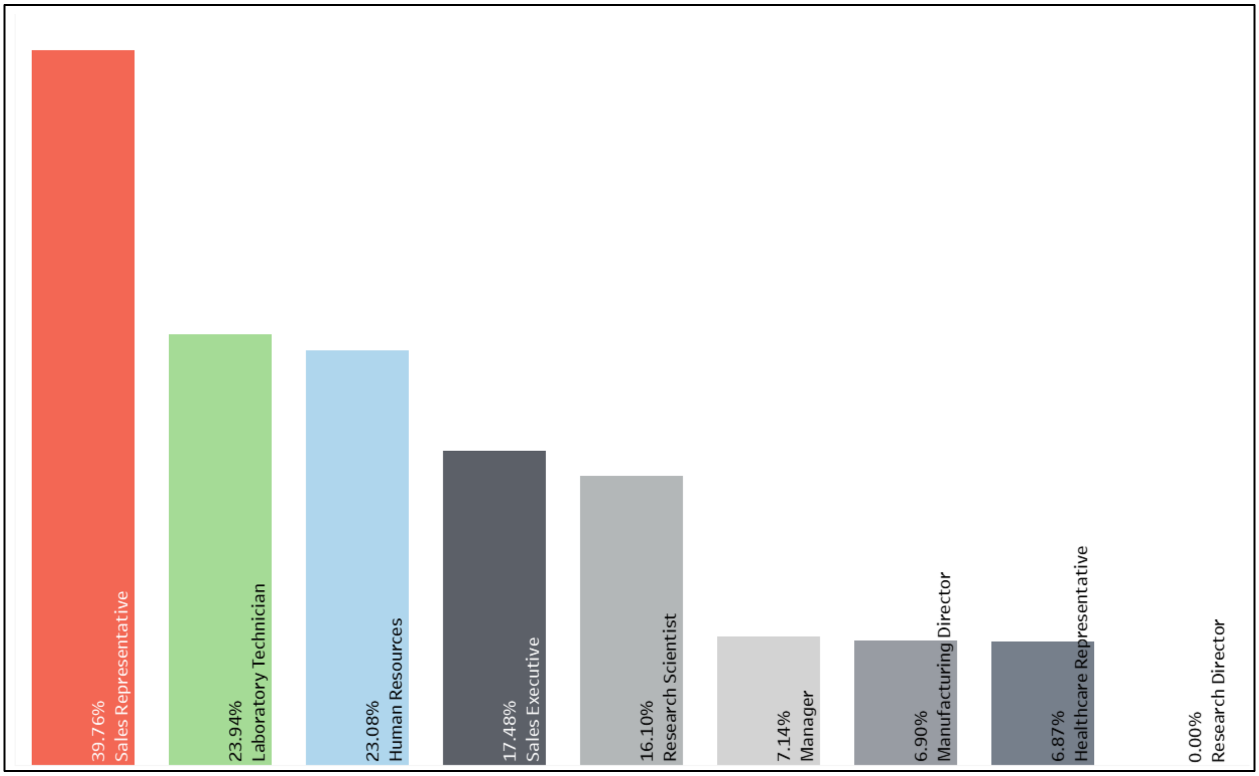
Department-Wise Attrition

Department-wise attrition analysis indicated that the Sales department had an attrition rate of 21.38%, Human Resources at 22.22%, and Research & Development at 14.76%, as highlighted in the dashboard's bottom-left pie chart, revealing that HR and Sales departments face higher turnover, likely due to high-pressure roles or competitive job markets.



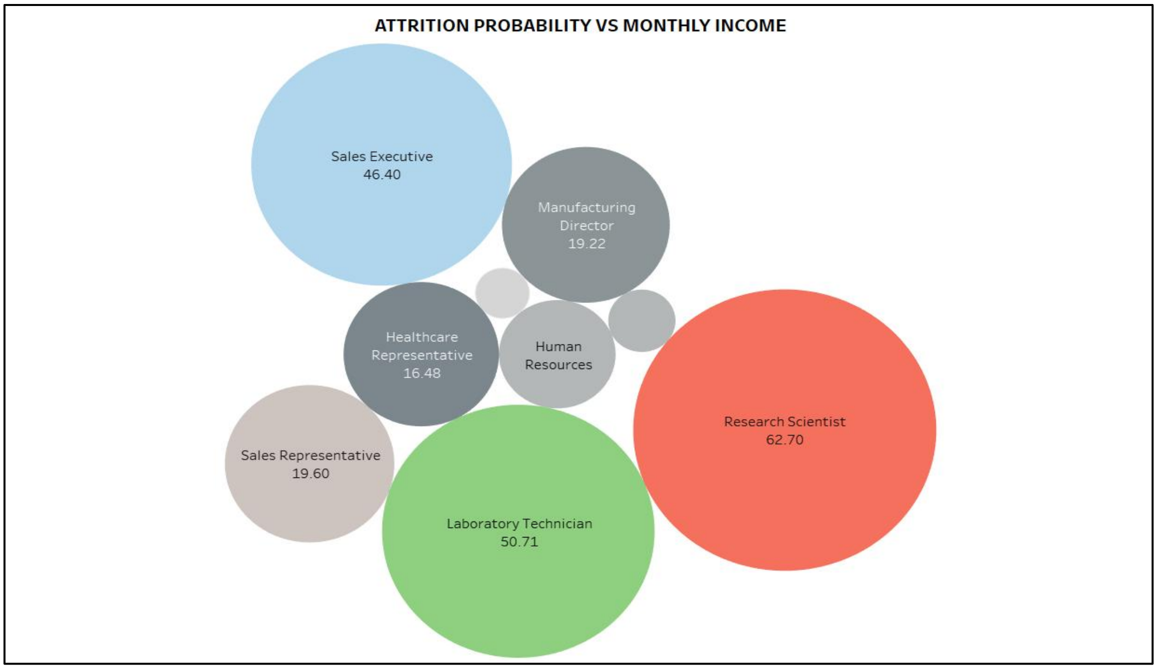
Attrition by Job Role

Attrition by job role varied significantly, with the dashboard's top-middle heatmap showing Research & Development roles like Research Scientist at 62.70%, Laboratory Technician at 19.22%, Healthcare Representative at 16.48%, and Manufacturing Director at 0.00%, while Sales roles included Sales Executive at 46.40% and Sales Representative at 19.60%, and the bottom-right bar chart further detailed Sales Representative at 39.79%, Laboratory Technician and Human Resources both at 23.08%, Sales Executive at 17.48%, Research Scientist at 16.10%, Manager at 7.14%, Manufacturing Director at 6.90%, Healthcare Representative at 6.87%, and Research Director at 0.00%, indicating that Research Scientists, Sales Executives, and Sales Representatives have the highest attrition rates, pointing to role-specific challenges.



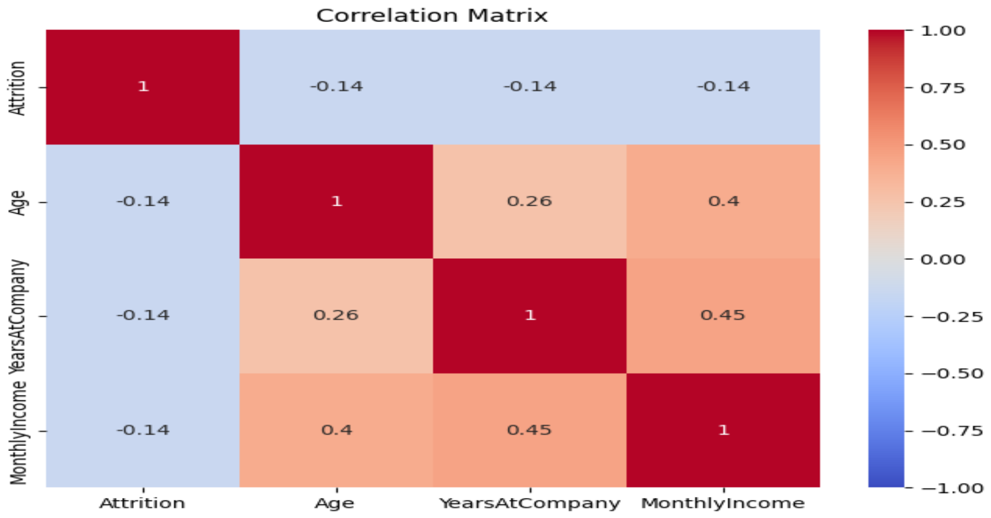
Attrition Probability vs Monthly Income

The relationship between attrition probability and monthly income was visualized in the dashboard's top-right bubble chart, showing Research Scientists with moderate income but a high attrition probability of 62.70%, Sales Executives with high income and a 46.40% attrition probability, and Laboratory Technicians with lower income and a moderate 19.22% attrition probability, suggesting that high attrition isn't solely tied to low income and that other factors like job satisfaction or work environment likely play a significant role.



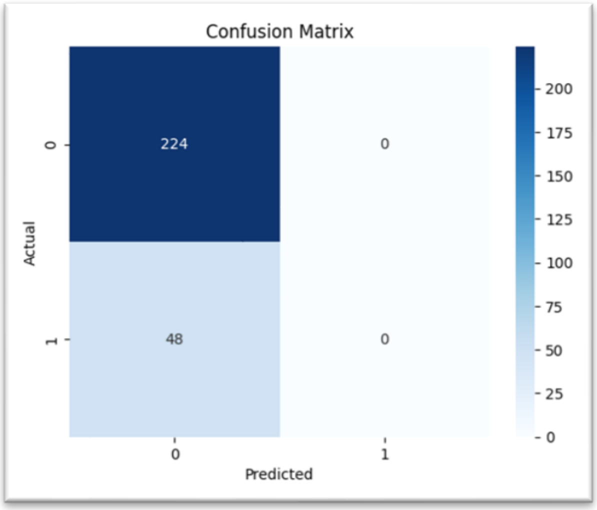
Statistical Analysis

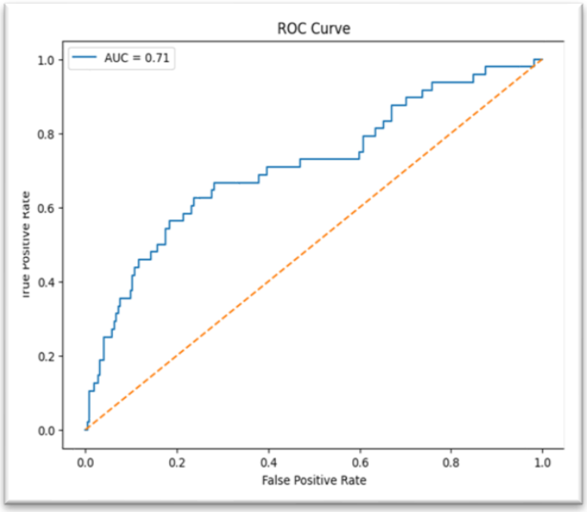
Statistical analysis revealed significant insights, with t-tests showing a p-value of 0.0000 for Age (median 37 years for stayers vs. 33 years for leavers), Years at Company (median 5 years for stayers vs. 3 years for leavers), and Monthly Income (median $5,000 for stayers vs. $4,000 for leavers), indicating younger, newer, and lower-paid employees are more likely to leave, while the correlation matrix heatmap displayed negative correlations of -0.14 between Attrition and Age, YearsAtCompany, and MonthlyIncome, alongside positive correlations of 0.40 between Age and MonthlyIncome and 0.45 between YearsAtCompany and MonthlyIncome, suggesting that age, tenure, and income collectively influence attrition, and chi-square tests confirmed significant associations (p-values < 0.05) between Attrition and categorical variables like Department, JobRole, and OverTime.



Predictive Modeling

Predictive modeling utilized a Logistic Regression model with features Age, YearsAtCompany, MonthlyIncome, OverTime, JobSatisfaction, and WorkLifeBalance, achieving an accuracy of 83.82%, with a confusion matrix showing 223 true negatives, 1 false positive, 43 false negatives, and 5 true positives, and a classification report indicating a precision of 0.83, recall of 0.10, and F1-score of 0.19 for class 1, revealing the model's strength in predicting stayers but weakness in predicting leavers due to class imbalance, while a Random Forest model identified OverTime as the top feature influencing attrition, followed by MonthlyIncome, Age, YearsAtCompany, JobSatisfaction, and WorkLifeBalance, as shown in the feature importance bar chart, highlighting that employees working overtime are more likely to leave.

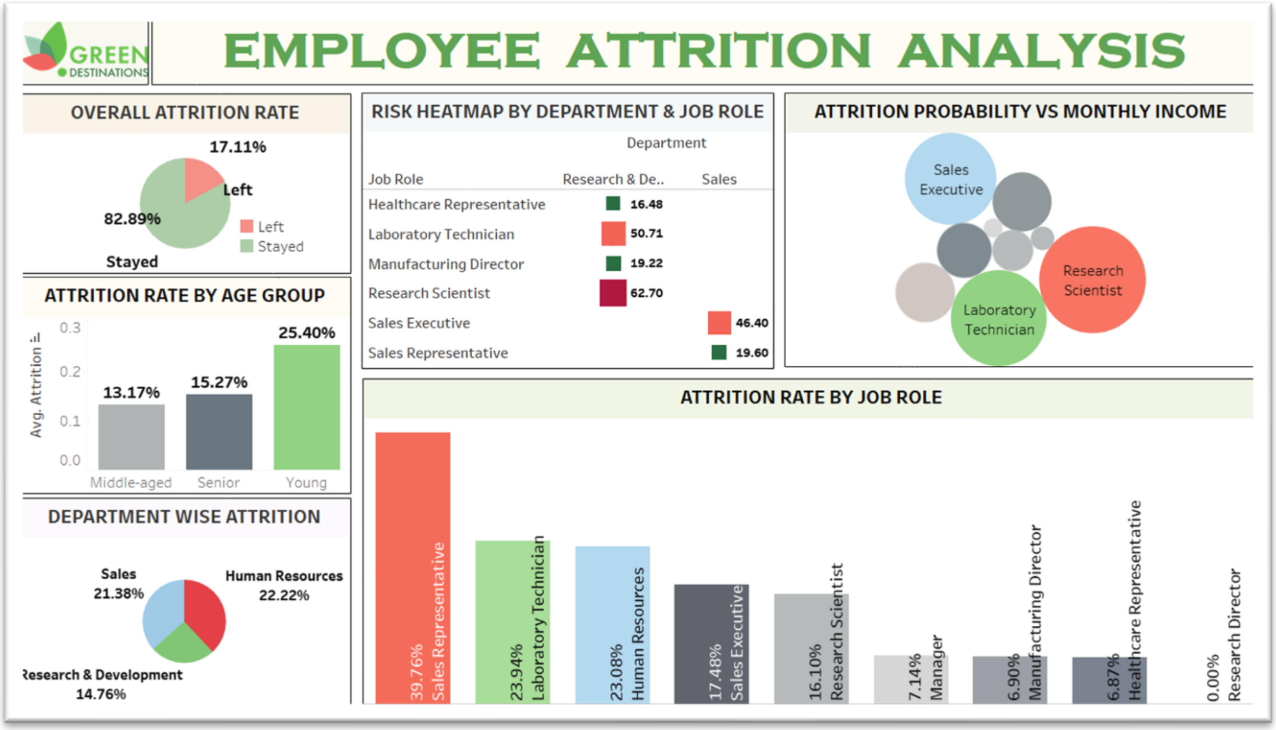




Attrition Probabilities

Attrition probabilities were predicted for each employee using the model, with results exported to employee_attrition_probabilities.csv, including examples such as Employee 1 with a 26.67% probability of leaving (Attrition: Yes) and Employee 2 with an 86.63% probability of leaving (Attrition: No), providing actionable insights for identifying at-risk employees.

Recommendations (Dashboard)



	Department	
Job Role	Research & De..	Sales
Healthcare Representative	16.48	
Laboratory Technician	50.71	
Manufacturing Director	19.22	
Research Scientist	62.70	
Sales Executive		46.40
Sales Representative		19.60

1. **Target High-Risk Groups**:
   - **Young Employees (25.40% Attrition)**: Implement mentorship programs, career development workshops, and engagement activities to increase commitment.
   - **Sales and HR Departments (21.38%, 22.22%)**: Address high-pressure environments with flexible work arrangements, better incentives, and team support.
   - **High-Risk Roles**:
     - Research Scientists (62.70%): Improve job satisfaction through recognition, career growth opportunities, or workload management.

    - Sales Executives (46.40%) and Sales Representatives (39.79%): Offer competitive bonuses, reduce pressure, and provide sales training.

2. **Address Non-Monetary Factors**:
    - **OverTime**: Reduce mandatory overtime, as it’s the top factor driving attrition. Offer compensatory time off or incentives for overtime work.
    - **Job Satisfaction and Work-Life Balance**: Conduct surveys to identify dissatisfaction sources and improve work conditions (e.g., flexible hours, wellness programs).

3. **Income Adjustments**:
    - While income isn’t the sole factor, employees earning less than $5,000 (median for leavers) are more likely to leave. Consider salary reviews for lower-paid roles like Laboratory Technicians.

4. **Learn from Low-Attrition Roles**:
    - Research Directors and Manufacturing Directors (0% attrition) can serve as models. Investigate their conditions (e.g., autonomy, compensation) and replicate best practices.

5. **Enhance Onboarding for New Employees**:
    - Employees with fewer years (median 3 years for leavers) are more likely to leave. Strengthen onboarding with clear career paths and early engagement.

Challenges Encountered

- **Class Imbalance:** The dataset had a low attrition rate (17.11%), leading to imbalanced classes (1,124 stayers vs. 232 leavers), which reduced the model's ability to predict leavers accurately.
- **Outlier Detection:** Determining the appropriate threshold for Monthly Income outliers was challenging, as overly aggressive removal could skew results.
- **Feature Selection:** With 35 features, selecting the most relevant ones for modeling required careful consideration to avoid overfitting.
- **Non-Monetary Factors:** Quantifying qualitative factors like job satisfaction and work-life balance was difficult, as they relied on ordinal survey data.
- **Stakeholder Communication:** Translating complex analytical findings into actionable insights for non-technical HR stakeholders required clear and concise visualizations.

How Challenges Were Addressed

- **Class Imbalance**: Used techniques like stratified sampling during model training to ensure balanced representation of classes. Considered oversampling methods (e.g., SMOTE) but prioritized interpretability for stakeholder trust, accepting lower recall for leavers.
- **Outlier Detection**: Applied the IQR method for outlier removal, carefully validating the threshold ($10,501) by comparing pre- and post-removal distributions to ensure minimal data loss (1,470 to 1,356 rows).
- **Feature Selection**: Used correlation analysis and Random Forest feature importance to select six key features (Age, YearsAtCompany, MonthlyIncome, OverTime, JobSatisfaction, WorkLifeBalance), balancing model complexity and predictive power.
- **Non-Monetary Factors**: Treated ordinal variables (e.g., JobSatisfaction, WorkLifeBalance) as numerical for modeling, while acknowledging limitations in EDA discussions. Recommended surveys to gather richer qualitative data.

- **Stakeholder Communication**: Developed an intuitive dashboard with clear visualizations (e.g., pie charts for attrition rates, bubble charts for income vs. attrition) and exported it as a PDF for easy sharing. Used plain language in recommendations to ensure accessibility.

Suggestions for Future Improvement

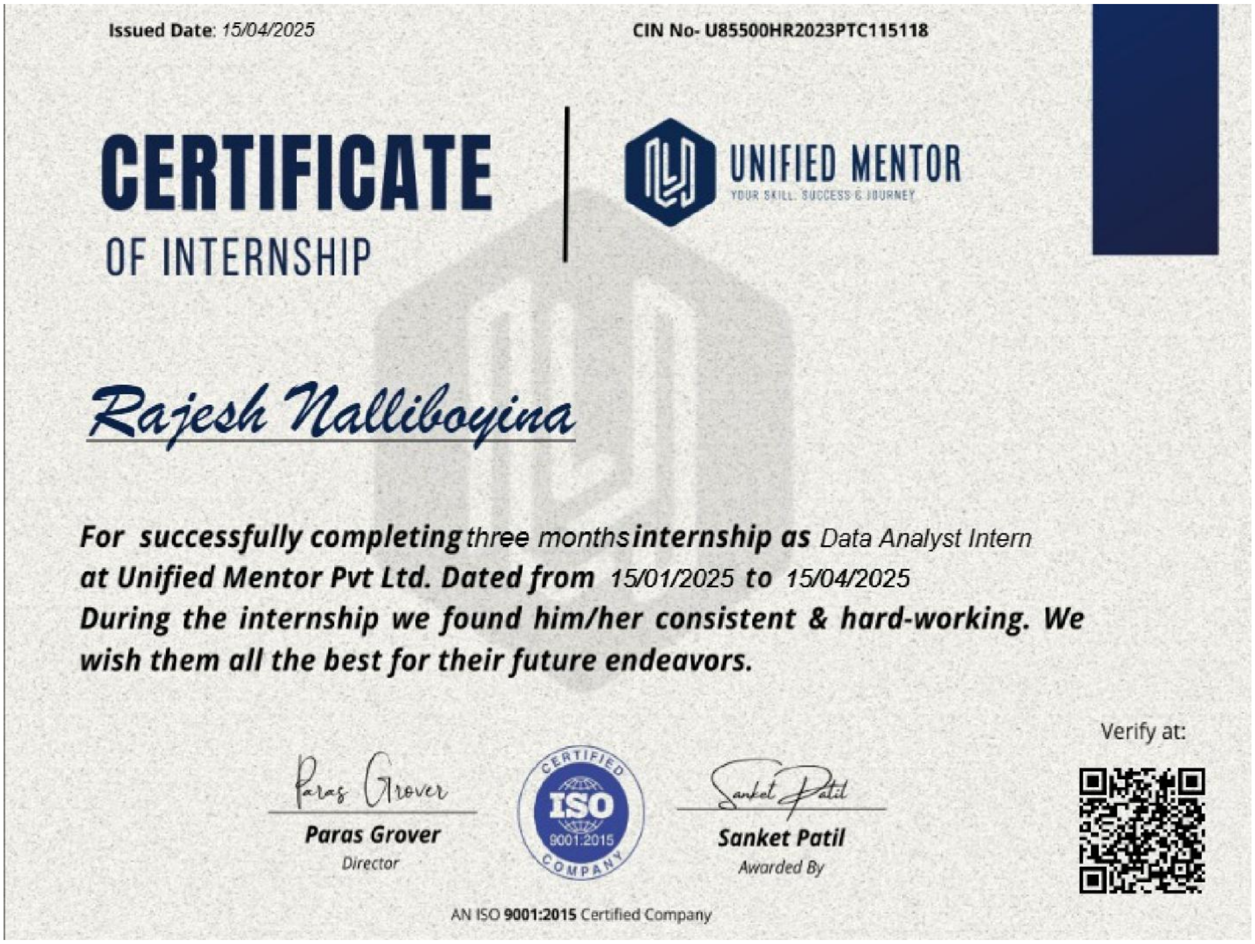
- **Address Class Imbalance**: Implement advanced techniques like SMOTE or ensemble methods (e.g., XGBoost) to improve the model's ability to predict leavers, increasing recall without sacrificing interpretability.
- **Incorporate Qualitative Data**: Collect detailed survey data on job satisfaction, work-life balance, and workplace culture to better quantify non-monetary factors.
- **Time-Series Analysis**: If longitudinal data becomes available, analyze attrition trends over time to identify seasonal patterns or the impact of specific events (e.g., policy changes).
- **Expand Feature Set**: Include additional features like employee engagement scores, manager feedback, or exit interview data to capture more nuanced drivers of attrition.
- **Real-Time Monitoring**: Develop a real-time attrition dashboard integrated with HR systems to track turnover and predict at-risk employees dynamically.
- **Pilot Retention Programs**: Test recommended strategies (e.g., mentorship for young employees, reduced overtime) in pilot programs and measure their impact on attrition rates.
- **Cross-Industry Benchmarking**: Compare Green Destinations' attrition rates and drivers with industry benchmarks to contextualize findings and identify unique challenges.

CONCLUSION

This project successfully analyzed employee attrition at Green Destinations, identifying an overall rate of 17.11% and pinpointing high-risk groups: young employees, Sales and HR departments, and roles like Research Scientists and Sales Representatives. Key factors influencing attrition include age, years at the company, monthly income, overtime, job satisfaction, and work-life balance. Predictive modeling achieved 83.82% accuracy, with overtime identified as the most critical factor. The dashboard effectively communicates these findings, making it a valuable tool for HR. By implementing targeted retention strategies, addressing non-monetary factors, and learning from low-attrition roles, Green Destinations can reduce turnover, improve employee satisfaction, and enhance organizational stability.

GOOGLE COLAB LINK

CERTIFICATE OF COMPLETION



Issued Date: *15/04/2025*

CIN No- U85500HR2023PTC115118

CERTIFICATE

OF INTERNSHIP

UNIFIED MENTOR

YOUR SKILL, SUCCESS & JOURNEY

Rajesh Nalliboyina

For successfully completing *three months* ***internship as*** Data Analyst Intern ***at Unified Mentor Pvt Ltd. Dated from*** 15/01/2025 ***to*** 15/04/2025 ***During the internship we found him/her consistent & hard-working. We wish them all the best for their future endeavors.***

Verify at:

Paras Grover

Paras Grover

Director

CERTIFIED ISO 9001:2015 COMPANY

Sanket Patil

Sanket Patil

Awarded By

AN ISO **9001:2015** Certified Company