**OPIM-5604-Predictive Modelling**

**Alberto Akkarakkaran**

**Sai Sunil Thokal**

**Rishika Rajendra Gour**

**Employee Attrition Classification Data**

**Table of Contents**

| **Section** | **Title** | **Description** |
| --- | --- | --- |
| **1** | **Title Page** | Project Title and Team Members |
| **2** | **Executive Summary** | Overview of Employee Attrition, Objectives, and Outcomes |
| **3** | **Business Understanding** | Definition, Impact of Attrition, and Project Goals |
| **4** | **Assumptions and Constraints** | Data Quality, Privacy, and Project Limitations |
| **5** | **Data Dictionary** | Source of Data and Attributes, including Data Dictionary |
| **6** | **Data Transformation** | Recoding and New Variables (Salary CI, Burnout Indicator) |
| **7** | **Methodology** | Data Analysis and Visualization of Attrition |
| **8** | **Modelling** | Logistic Regression: Iterations, Refinement, and Performance Metrics |
| **9** | **Evaluation** | Evaluation of Decision Tree, Bootstrap Forest, and Boosted Tree Models |
| **10** | **Deployment** | Key Predictors of Attrition and Retention Strategies |
| **11** | **References** | Summary of Findings and Suggestions for Future Analysis |
| **12** | **Appendix** | Dataset Source and Additional Resources |

**Data:** [**https://www.kaggle.com/datasets/stealthtechnologies/employee-attrition-dataset/data**](https://www.kaggle.com/datasets/stealthtechnologies/employee-attrition-dataset/data)

**Executive Summary**

Employee attrition poses significant challenges for organizations which affects work efficiency, profitability and stability of employees. The aim of this project is to analyse and predict employee attrition by identifying factors that cause employees to leave the organization. This analysis can prove helpful for HR and management teams in implementing strategies to retain valuable employees and reduce turnover rates.

To achieve this goal employee data confining demographic, organizational and performance parameters were analyzed with attrition being the target variable to distinguish employees who remain with the organization. It classifies leavers as "left" or "stayed." The data includes variables such as age, monthly income, years in the company, number of promotions, job satisfaction, work life balance, overtime status, education level, company tenure and other workplace characteristics. We have explored the relationship between these factors and attrition to determine which variable was a statistically significant effect.

During the analysis, we observed that certain factors played a more prominent role in predicting attrition. For instance, employees with low job satisfaction, those working overtime, or those in specific job roles or with lower monthly incomes showed higher likelihoods of leaving the organization. Company tenure also emerged as a significant factor, indicating that employees in their early years with the company were more prone to leaving. By examining p-values and refining the model to include only significant predictors, we ensured the results were statistically robust and insightful.

During analysis, we observe that certain factors play an important role in predicting attrition, for example, employees with low job satisfaction, working overtime or are in a particular job position or have a low monthly income tend to leave the organization. The company's tenure is also an important factor as this indicates that employees are more likely to leave their early years with the company by examining the p -values ​​and refining the model to include only significant predictors, we ensured the results are robust and have statistical insights.

The data set used for this analysis shows a balanced distribution between employees who stay and those who leave, with 7,032 employees classified as "Left" and 7,868 as "Stayed." This balance creates effective model training and certification without introducing bias. The resulting logistic regression model not only provides a framework for understanding the drivers of turnover; but it is also a predictive mechanism to identify employees at risk of leaving. This predictability allows organizations to take proactive steps, such as improving compensation structures, creating better career advancement opportunities and managing work-life balance and take care of the environment.

Through this project we've laid the foundation for data-driven decision-making in employee retention strategies. This allows the organization to reduce the side effects of attrition. It focuses on the identified predictors. Management can design targeted interventions to increase employee satisfaction and engagement which will ultimately leading to more stable and productive workforce.

**Business understanding**

Employee attrition, or the rate at which employees leave an organization, is a critical business challenge that directly impacts organizational performance, stability, and costs. High levels of attrition lead to increased costs associated with hiring, training, and onboarding new employees, as well as disruptions to workflows and a decline in employee morale. Retaining top talent is crucial for maintaining a competitive edge, ensuring productivity, and fostering innovation. For companies, understanding the drivers of employee turnover and implementing data-driven strategies to mitigate attrition is essential to achieving long-term success.

Employee attrition, or employee turnover, is a critical business challenge that directly affects an organization's productivity, stability, and costs. High levels of attrition costs associated with hiring, training, and onboarding new employees interrupts the workflow and reduce employee morale competitive advantage to maintain and ensure productivity, innovation, it is important to retain top talent. It is important for companies to understand the factors that drive employee turnover and implement data-driven strategies to reduce turnover to achieve long-term success.

This project aims to address the issue of high employee attrition by developing a predictive model that identifies the key factors that influence an employee's decision to leave an organization. Use of historical data and employee characteristics such as demographics, job satisfaction, performance ratings, compensation, and work environment factors. The model attempts to reveal patterns and trends that can predict whether employees are likely to leave. This predictive ability is valuable because it allows organizations to take proactive steps to retain their employees.

The business problem this predictive model addresses are multifaceted:

1. **Employee Retention**: Identifying employees at high risk of leaving helps HR teams implement retention strategies, such as improving work conditions, increasing engagement, offering better compensation, and addressing employee concerns.
2. **Cost Reduction**: By reducing turnover rates, organizations can lower costs associated with recruitment, training, and productivity losses.
3. **Workforce Stability**: Staff retention ensures continuity in operations and preserves institutional knowledge, which is critical for sustaining business growth.
4. **Strategic Decision-Making**: The insights from the model empower management to prioritize resource allocation and target interventions where they are most needed. For example, identifying departments or roles with high attrition risks allows leadership to focus their efforts on these areas.

By defining attrition as the target variable, this project focuses on creating reliable logistic regression model. It predicts which employees are likely to resign. This model provides actionable insights into the underlying factors that drive attrition and help organizations manage those things effectively. Finally, this analysis supports the business goal of creating a more satisfied, engaged, and stable workforce which is fundamental to maintaining organizational success in a competitive environment.

**Assumptions and Constraints**

* **Data Quality:** It is assumed that the data is accurate, complete, and consistent.
* **Data Privacy and Security:** Data will be handled in compliance with relevant privacy regulations.
* **Time and Resource Constraints:** The project will be conducted within a specific timeframe and resource allocation.

**Data Dictionary**

* Employee ID: A unique identifier assigned to each employee.
* Age: The age of the employee, ranging from 18 to 60 years.
* Gender: The gender of the employee
* Years at Company: The number of years the employee has been working at the company.
* Monthly Income: The monthly salary of the employee, in dollars.
* Job Role: The department or role the employee works in, encoded into categories such as Finance, Healthcare, Technology, Education, and Media.
* Work-Life Balance: The employee's perceived balance between work and personal life, (Poor, Below Average, Good, Excellent)
* Job Satisfaction: The employee's satisfaction with their job: (Very Low, Low, Medium, High)
* Performance Rating: The employee's performance rating: (Low, Below Average, Average, High)
* Number of Promotions: The total number of promotions the employee has received.
* Distance from Home: The distance between the employee's home and workplace, in miles.
* Education Level: The highest education level attained by the employee: (High School, associate degree, bachelor’s degree, master’s degree, PhD)
* Marital Status: The marital status of the employee: (Divorced, Married, Single)
* Job Level: The job level of the employee: (Entry, Mid, Senior)
* Company Size: The size of the company the employee works for: (Small, Medium, Large)
* Company Tenure: The total number of years the employee has been working in the industry.
* Remote Work: Whether the employee works remotely: (Yes or No)
* Leadership Opportunities: Whether the employee has leadership opportunities: (Yes or No)
* Innovation Opportunities: Whether the employee has opportunities for innovation: (Yes or No)
* Company Reputation: The employee's perception of the company's reputation: (Very Poor, Poor, Good, Excellent)
* Employee Recognition: The level of recognition the employee receives:(Very Low, Low, Medium, High)

**Data Transformation**

Several variables were recoded, and new variables were created to ensure the model worked smoothly. First, monthly income was recoded as a categorical variable called **Pay grade** which fall into three categories: A, which has a threshold of $5,000; B, $5,001 to $9,000; and C covers $9,001 to $15,000. **Gender** was later recoded. Male is assigned a value of 1 and female is assigned a value of 0 to simplify the representation in the model. In addition, Years at company was categorised into Junior, Middle, and Senior have been recoded to classify employees by tenure.

Two new variables were also introduced to improve the analysis. The first is the Salary Competitiveness Index (**Salary CI**), which compares an employee's salary to that of co-workers in the same job role. The formula for CI is an **employee's salary range divided by the average salary of co-workers** in the same job role. A second variable is added which is an **indicator of burnout** to identify employees at risk of termination due to excessive working hours. This metric is calculated as overtime hours multiplied by (1 divided by work-life balance, further multiplied by lack of leadership opportunities). These changes and new variables aim to better capture the key factors affecting employee attrition.

**Methodology**

A graph of a bar chart

Description automatically generated with medium confidence

This bar chart provides a clearer view of the distribution of attrition in an organization and It will display the exact amount:

* **Left:** The number of employees who left the organization is 7,032. This group represents a case of decline.
* **Stayed:** The number of employees remaining with the organization is 7,868. This group represents those who have left.

The chart shows most employees stayed, though the differences between the two groups were not large. This shows a relatively balanced but slightly higher retention rate compared to attrition.

**Modelling**

After taking all relevant variables, this is the first model output –

**Iteration 1**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

Here, we will remove the variables with value higher than 0.05. The variable with the highest p-value is Pay Grade (0.7273) which has to be eliminated first.

Iteration 2 –

A screenshot of a computer

Description automatically generated

**A screenshot of a computer

Description automatically generated**

Now, the next variable to be removed is Salary CI with a p-value of 0.3965 (higher than 0.05).

Iteration 3 –

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Lastly, we will eliminate burnout indicator with p-value of 0.3436.

Iteration 4-

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

This would be the final model as now all p-values are below 0.05 making all variables statistically significant. The variables which are above the limit can be verified in effect wald tests as significant due to which we do not need to eliminate any further.

A screenshot of a computer

Description automatically generated

The R Square of 0.1565 in our final model indicates that the variance in the dependent variable is explained by the independent variables. While not exceptionally high, it provides a meaningful level of insight for our predictions, suggesting that the model captures relevant relationships.

Now for further comparison, we are showing the AUC values and ROC comparison using model comparison in formula depot

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Here we can the AUC values of all the three models Decision tree (0.7376), Bootstrap forest (0.7324) and Boosted tree (0.7419).

Boosted Tree shows statistically significant improvement over Partition (p = 0.0037) and Bootstrap Forest (p = 0.0304) in terms of AUC.

The difference between Partition and Bootstrap Forest is not statistically significant (p = 0.1390).

Given its higher predictive performance and statistically significant improvement in AUC, the Boosted Tree model is the best choice for predicting attrition in this scenario

**ROC Curve Comparison -**

A screenshot of a computer

Description automatically generated

**Evaluation**

To measure the success of the predictive models for employee attrition, we focused on performance metrics like AUC (Area Under the Curve), accuracy, precision, recall, and F1-score. The Boosted Tree model stood out as the most reliable, with an AUC of 0.7419. This tells us the model is quite effective at identifying employees likely to leave, giving HR teams a chance to step in early. Through careful refinement, we ensured the model only used statistically significant factors, making the results dependable and meaningful for real-world decision-making. The impact of this model on the business can be profound. For example, if the model helps reduce turnover by just a small percentage, it could save the company significant costs related to recruitment, onboarding, and training—not to mention the intangible benefits like better employee morale and continuity in operations. Imagine being able to identify employees at risk and addressing their concerns before they decide to leave. This proactive approach not only saves money but also fosters a more engaged and stable workforce.

Building a Business Case

To justify the value of this project, we can start by estimating the expected improvements. For instance, lowering attrition rates by even 5% could result in substantial cost savings. It’s not just about the dollars saved; it’s about keeping talented employees who drive the company forward. Calculating ROI could involve comparing the investment in building and implementing the model (e.g., software, team time, interventions) to the savings from reduced turnover.

Of course, measuring ROI isn’t always straightforward. For one, some data might not be readily available, or external factors like economic conditions might play a role in employee turnover. If this happens, we can still focus on the qualitative benefits. Retaining employees isn’t just about money—it’s also about fostering a workplace where people feel valued and engaged.

Alternatives to ROI

When precise ROI calculations are tricky, we can measure success through other means. For example:

Conduct employee satisfaction surveys before and after interventions guided by the model.

Track attrition rates over time to see if they align with the model's predictions.

Compare departments or teams with tailored interventions to those without, to see the difference in retention.

Ultimately, this project isn’t just about creating a tool—it’s about empowering the organization to make better decisions for its people. By focusing on real, actionable insights, we can create a workplace where employees thrive, turnover drops, and the business reaps the rewards of a happier, more stable workforce.

**Deployment**

To effectively roll out our data mining results, we'll integrate the boosted tree model, which boasts the highest AUC of 0.7419, into our company’s decision support system. This involves:

1. **Technical Integration**: Incorporating the model into our existing IT infrastructure to ensure seamless interaction with our data sources.
2. **User Training**: Educating relevant staff on how to effectively use the model outputs for their decision-making processes.
3. **Monitoring & Maintenance**: Setting up ongoing monitoring to ensure the model's accuracy and making necessary updates over time.

**Potential Issues:**

1. **Data Privacy**: Ensuring compliance with data protection regulations like GDPR. We'll need to anonymize sensitive employee information to protect privacy.
2. **Data Quality**: Keeping the model updated with clean, high-quality data to maintain its accuracy and relevance.

**Risks and Mitigation:**

1. **Over-reliance on the Model**:
   1. *Risk*: Decision-makers might rely too heavily on the model without considering other factors.
   2. *Mitigation*: Providing training on using the model as a support tool rather than the sole decision-maker.
2. **Operational Challenges**:
   1. *Risk*: Integration issues with current systems may lead to delays or functionality problems.
   2. *Mitigation*: Conducting thorough testing and having a phased deployment plan to address and resolve any issues promptly.

Deploying these insights thoughtfully ensures that our firm benefits from improved decision-making while maintaining ethical standards and mitigating potential risks

**References**

1. [**https://www.questionpro.com/blog/employee-attrition-analytics/**](https://www.questionpro.com/blog/employee-attrition-analytics/)
2. [**https://nycdatascience.com/blog/data-visualization/employee-attrition-analysis/**](https://nycdatascience.com/blog/data-visualization/employee-attrition-analysis/)
3. [**https://www.zavvy.io/blog/employee-attrition**](https://www.zavvy.io/blog/employee-attrition)
4. [**https://www.linkedin.com/pulse/employee-attrition-7-most-common-reasons-employees-quit-pappas--uk6re/**](https://www.linkedin.com/pulse/employee-attrition-7-most-common-reasons-employees-quit-pappas--uk6re/)

**Appendix**

* **Balanced Dataset**: The dataset includes 7,032 employees who left and 7,868 who stayed, offering a balanced foundation for training and validating the predictive model.
* **Key New Variables**: Burnout Indicator and Salary Competitiveness Index (CI) were introduced to capture overtime risks and salary comparison among peers.
* **Model Refinement**: Iterative logistic regression removed insignificant variables like Pay Grade and CI, resulting in a model with all p-values below 0.05.
* **Best Model**: The Boosted Tree model achieved the highest AUC (0.7419), outperforming Decision Tree and Bootstrap Forest models in predictive performance.
* **Practical Application**: Insights from the model enable targeted interventions, addressing job satisfaction, workload, and compensation to reduce attrition.