Birla Institute of Technology and Science, Pilani Dubai Campus

A Comprehensive Report on

Employee Attrition Analysis: A Data Science Project

By

Nikhil Vaidyanath 2022A7PS0073U

Ahmad Bin Firdous 2022A7PS0102U

Prathmesh Mehrotra 2022A7PS0230U

Kanishk Mohan 2022A7PS0241U

Course: CS F320 FOUNDATION OF DATA SCIENCE

Employee Attrition Analysis

Abstract

Employee turnover is one of the biggest organizational issues because it impacts productivity, motivation, morale and even the organization's financial prosperity. This project assesses the causes of the high turnover of employees in organizations using analytical methods and machine learning algorithms. The study employs one dataset, with 1,470 records and 35 attributes which include demographic information, job satisfaction measures, performance information, and attrition information.

The main objectives of this analysis are to understand the primary causes of attrition, use data visualization to provide recommendations for preventing future occurrences, and create a model that can help in the early identification of potential employees who may quit the organization. Exploratory Data Analysis reveals important trends like increased attrition amongst early workers, workers who are dissatisfied in their jobs, and workers in positions with less mobility. These patterns may be described verbally or in writing, but are most easily represented by visual means such as histograms, box plots, and heat maps.

Some of the most common methods explored for attrition include logistic regression, decision trees and random forests. The evaluation of the models was based on accuracy, precision, recall, and F1-score indicators where Random Forest has been discovered to be showing the highest performance. In this study, feature importance analysis selects the following variables as important predictors: overtime work, monthly income, job satisfaction, and tenure.

Coping strategies that are explained in the study include managing overtime workloads, promoting career advancement opportunities, improving satisfaction surveys, and restructuring compensation packages for the purpose of retaining employees. Such measures help organizations save on turnover expenses, please the worker and increase efficiency rates.

Therefore, this project shows that data science approaches can be used to identify and address employee attrition. Since it is crucial to understand what makes talents stay or leave in an organization, accurate prediction of crucial factors and implementing action plans, organizations can prevent turnover, highlighting the crucial role of data science in today's business management.

Introduction

In today's competitive business environment, managing talent staff has become one of the greatest concerns of the business organizations irrespective of their niches. Consequent loss of employees whether voluntary or involuntary has negative impacts on operations, productivity, and expense to be incurred to have new employees hired, trained, oriented, etc. Establishing what causes such trend is important to enable companies to design proper strategies for retaining customers while maintaining competitive advantage.

Through data science and machine learning, Employee attrition analysis aims to detect and display various patterns of attrition rates. Employing factors such as demographic characteristics of employees, their job positions, level of job satisfaction and performance data, organizations can identify the attrition loss predictors. Such knowledge allows managers to counter obstacles, cultivate organizational climates, and maintain good employees.

This report focuses on the best data science approaches to forecasting and modelling of the employee turnover. By using the present data format, the study utilizes exploratory data analysis and machine learning methods to derive decision making hypotheses. Trends of attrition in the various departments, role and satisfaction levels as presented by visualizations are complemented by the analysis of potential attrition risks provided by the predictive models. Such reasoning gives organizations empirical solutions to improve employee commitment and retention while also increasing organizational robustness.

Precisely, this study makes a unique contribution to the bodies of knowledge in data science and organisation workforces by demonstrating an application of data science in addressing organisational issues.

The Role of Data Science in Employee Attrition Analysis

Data science plays an important role in implementing measures that are relevant to minimize turnover rates by analysing data. It involves the steps such as data cleaning, data preprocessing and data visualization to systematically study the empirical data which includes various demographic fields of employees, their job profile, organizational satisfaction levels and performance indices.

One of the extensions of data science, is machine learning, which finds patterns and dependencies in the database that could otherwise remain unnoticed by classical statistics. These models give estimations of individual employees' chances of turnover that allow organizations to act. Feature importance analysis defines core attributes that cause employee turnover, for example, the dissatisfaction with some roles, working overtime, or the lack of promotion opportunities.

Data science also supports scenario simulations, i.e. to estimate how introducing some new policy might influence organizations' outcomes, for instance, by providing work-life balance programs or changing compensation policies. With predictive and prescriptive analytics, data science enables businesses to decrease turnover, optimize workers' management, and increase workforce using.

Problem Statement

High employee turnover, or the rate at which employees exit their organizations, remains a major problem in organizations regardless of industry. High turnover rates cost organisations a lot of money besides lowering productivity levels and disrupting operations. For organizational stability and to promote staff retention it is highly important to identify the causes of turnover. However, frequently applied techniques that enable organizations to address attrition include qualitative assessments and conventional information processing that omits identifying those concealed patterns or predicting turnover with remarkable precision.

In current dynamic business environment, the HR departments face employees sensitive to several factors, for example, job satisfaction, promotion, wages, flexible working hours, and relevant market rates etc. It is quite another thing to predict, with reasonable precision, which employees will be prone to resignation or job-hopping and this can only be done if previous employment data is sifted and sorted to obtain patterns that casual observation will not discern. Intuition, for instance, or surveys can be viewed as quite reactive and far from saving the multidimensional nature of employee behaviour.

This project forges ahead and aims at minimizing the errors in estimating the level of employee turnover. Using the records of the existing employees—their basic information, performance review results, remuneration data, attitudes toward the job, organizational climate data, and many others, ML algorithms reveal the

complex patterns of relations between these variables and turnover. These predictive models help the HR teams and other leadership to stop or prevent employee turnover by promoting better workplace conditions and opportunities for career growth and even improving compensation structures.

1. **Complexity of Employee Attrition Factors:**

   Employee Attrition is a complex phenomenon involving numerous variables strongly related to each other and varying from department, position, and employee's profile and characteristics. These factors include age, tenure, job satisfaction, perceived career advancement opportunities and perceived internal pay inequities in a way that determines whether an employee remains with the organisation. Conventional paradigms of attrition analysis by and large underemphasise interactions among these variables. For instance, an attractive remuneration package, may not keep the employee because he or she feels that there are no opportunities to progress in their career. Dealing with these and quantifying them calls for harmonized data analysis procedures that cannot be offered by ordinary methods.

2. **Predictive Power of Data Science and Machine Learning:**
   Analysis of datasets used in employee churn prediction show that data science is vital in handling, analysing, and modelling such data. Firstly, it helps in the effective analysis and cleaning of multiple types of data, data that is in raw form and can be used. The package then uses complex mathematical methods to screen and select important features that the model captures and uses in its representation of factors such as satisfaction, work load and promotions that affect an employee's turnover. Self-learning features of data science can facilitate identification of which employees are likely to quit and thus the necessary strategies would be provided to the organization through Human Resource management. This systematic approach does not only aid in identifying solutions for the causes of attrition, but also in implementing proactive solutions from insights derived from data science Decision Support Systems.

3. **Proactive Retention Strategies:**
   The lack of forecasting is another reason why attrition occurs. Such forecasts assist organisations in managing employee turnover as it happens. For instance, if an employee is identified as high risk in the organization based on the model such as low job satisfaction, or less chance to be promoted, the HR can attend to the issue by availing a promotion chance or redesigning their working parameters. Such approach moves organizations from turnover as an inevitable process to minimizing it and experience the related costs and impact on workforce stability and improved organizational performance.

4. **Cost Implications of Employee Attrition:**
   Employee turnover is a concept whose real cost is rarely considered. This not only incurs recruitment expenses but also affects the performance of the team by interrupting its dynamics, demotivating its members and causing loss of institutional memory. The high attrition level leads to gaps, becoming very costly, increases the workload at the existing staff, and instils negative morale on the team. By intervening early in the attrition causes it is easier for the organizations to reduce these impacts and foster long-term employee relations. Data Science applied together with Machine Learning offers an effective way to address turnover prediction and helps the organizations forecast employee turnover to minimize the consequences of the turnover rate.

In conclusion, employee attrition is a complex and significant challenge. By leveraging data science algorithms to analyse and predict turnover, organizations can gain valuable insights into the factors driving attrition and implement more effective retention strategies. This proactive approach helps improve workforce stability, reduce operational disruptions, and foster a more resilient organization.

Objectives

1. **Identify Key Factors Influencing Attrition**:

One of the most important goals of this study is to establish causes of employee turnover and this will be done with reference to age, gender, department, work experience, job satisfaction and other various factors. When these attributes have been identified, one is able to determine which groups of people, for instance, the demographically targeted people or certain job descriptions are more likely to leave. A case like inadequate opportunities for career advancement, can make workers in the youth population leave as can ineffective remunerations or cultures in the workplace.

2. **Predict Employee Turnover**:

In this project, deeper machine learning models will be employed to estimate the probability of the employees' turnover. Analytical models such as logistic regression, decision-tree and random forests will be created based on historical data in the aim to determine employees that are potentially likely to leave the organization. Such predictions help the HR manager focus the retention effort on high-risk employees and apply appropriate retention interventions.

3. **Visualize Attrition Trends**:

Big data needs a graphic representation because large numbers present problems in trends: data visualization helps. Spreadsheet, heat map and Dashboard are used for showing trends of turnover by departments, positions, or probation period. These presented diagrams hold valuable information that stakeholders can use to formulate specific approaches toward retaining learners.

4. **Enhance Retention Strategies:**

As a result of the gathered data, the study is going to provide recommendations that will help to improve the rates of employee turnover. These may include; flexibility and balance at the workplace, increased satisfaction through professional development and redesign of compensation and benefits. By implementing these strategies, organisations are thus able to retain employees and reign-in high turnover across organisations.

5. **Showcase Data Science Applications**:

This project aims to demonstrate the actual work portfolio of data scientist in dealing with organizations issues. This case sheds light into how big data and data analytics specifically with aspects of human resource management can significantly enhance an organisations decision-making process.

Relevance

Employee Attrition rate is an organizational issue because it results in loss of productivity, increased costs, and reduced revenue. High turnover ceases to be cheap as it involves costs such as costs of recruitment, costs of training as well as the loss of experience. This topic addresses these challenges by the means of data analysis and forecasting of turnover to support organizations' proactive approach to retaining their employees.

The findings of this study help to improve the management of human resources in organizations and contribute to the construction of organizational culture which influences the level of commitment in employees. Further to this, it gives emphasis to the significance of data science's changes in improving the stability of the workforce and becomes a very important and interesting subject in today's cut throat business environment.

Contributions

Data Collection and Preprocessing for Enhanced Data Integrity:

Primary data of different types will be collected from our dataset and be fed into a data frame using Pandas. Then data preprocessing techniques such as data imputation, and dealing with categorical variables will be adopted. This is important so that the dataset is clean and structured, which is fundamental in data science prediction models.

Enhancing Predictive Models Through Data Science Techniques: Use of Random Forest, SVM and XGBoost etc. and choice of these algorithms will be made. For enhanced employee attrition prediction, nine different features will be extracted from each gathered dataset, which will help leverage the outcome of various models.

Development of a Data-Driven Employee Retention Tool:

A complex instrument will be created which would use the concepts of the predictive models as tools for the HR managers. This tool will also focus on specific attrition measurements such as satisfaction degree and pay fractal to adopt effective retention measures by accurate data science analytics to mitigate turnover and increase employee satisfaction.

Proposed Solution

Employee Attrition Prediction System

The data obtained from the IBM HR analytics dataset, will help carry out prediction of employee turnover rates. Applying machine learning algorithms, the solution will be able to predict risks connected with the attrition, and, therefore, the organizations may take preventive actions and avoid losing valuable workers.

Key Aspects of the Proposed Solution:

• Data Integration and Processing:

This information will be accrued from the aforementioned dataset The next step is to carry out data preprocessing to manipulate and clean the data as per different techniques like handling missing values and outliers into the data set.

• Machine Learning Algorithm Selection:

Toward this end, XGBoost and SVM, to mention but a few, are expected to demystify the factors behind employee attrition. These types of algorithms are particularly useful when dealing with large, numerous and diverse data inputs and I will continue to adjust these algorithms specifically for the purpose of the turnover predictions concerning to the employees.

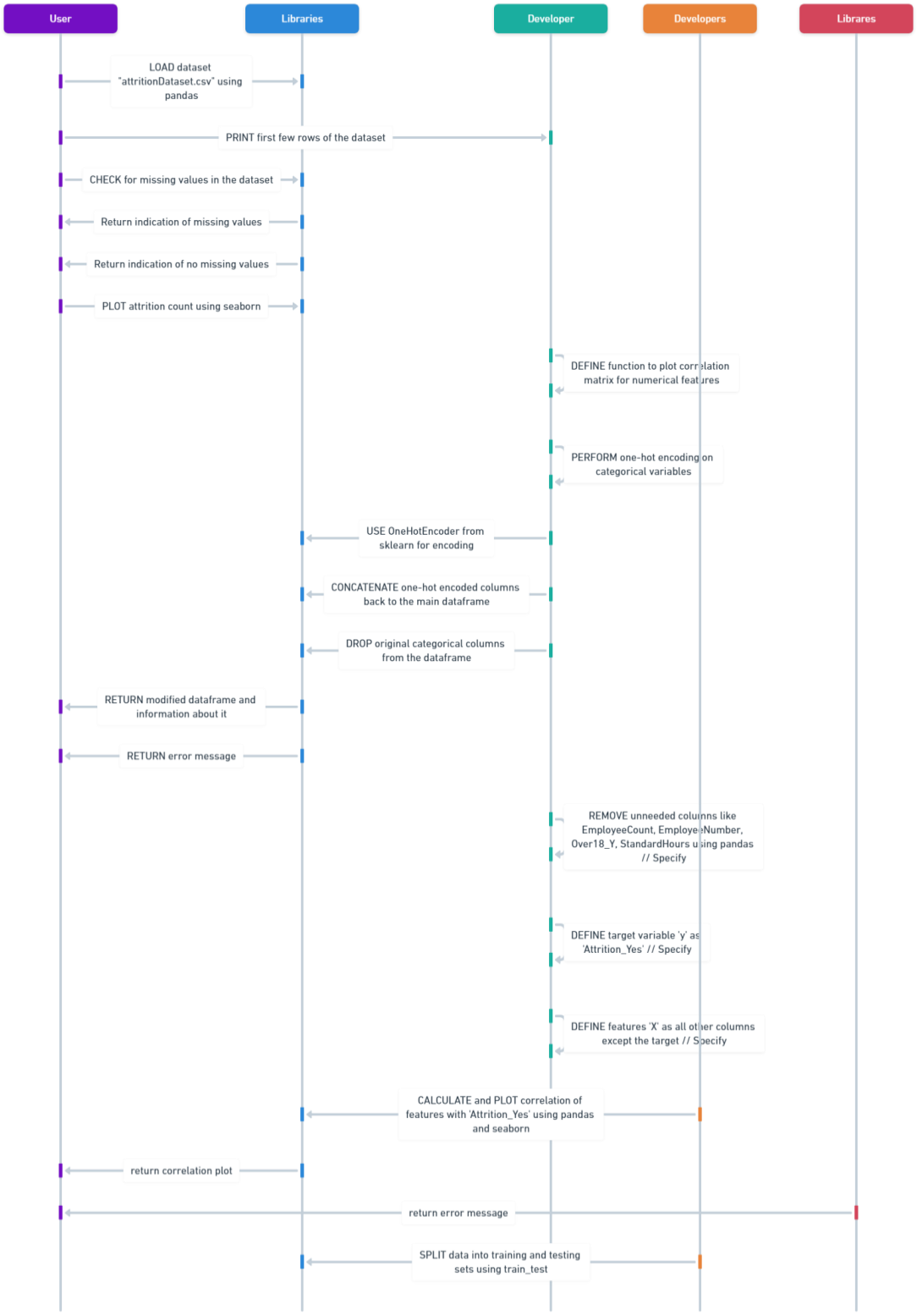
• Retention Strategy Development:

The system will offer suggestions concerning predicted attrition risks in a form of advice containing recommendations for action for the organisations, such as: promotions, training, raise or changing the working schedule. This advice will go a long way in assisting organizations to develop specific measures to address turnover and enhance Volatility.

Continuous Feedback Loop for Improving Model Accuracy:

It will be refined to incorporate new data as develops so as improve the model's ability to make accurate prediction outcomes. This continuous feedback loop will enable the system to be able to identify the new patterns of employees' behaviours and therefore come up with better prediction models for attrition.

Architecture Diagram



Experimental Results

The IBM HR Analytics Attrition dataset includes employee data from an organization, where each row represents a single employee and the columns capture various attributes such as demographic details, job roles, performance metrics, and whether the employee has left the company. Key features in the dataset include:

Age: Numeric value representing the employee's age.

Attrition: Categorical value ('Yes' or 'No') indicating if the employee has left the company.

BusinessTravel: Categorical value representing the frequency of travel (e.g., 'Travel_Rarely', 'Travel_Frequently', 'Non-Travel').

DailyRate: Numeric value indicating the daily rate of pay.

Department: Categorical value indicating the department (e.g., 'Sales', 'Research & Development').

DistanceFromHome: Numeric value indicating the distance from work to home in miles.

Education: Numeric scale (1-5) indicating the level of education.

EducationField: Categorical value indicating the field of education (e.g., 'Life Sciences', 'Medical', 'Other').

EmployeeCount: Numeric count, likely a constant value across the dataset.

EmployeeNumber: Unique identifier for each employee.

EnvironmentSatisfaction: Numeric scale (1-4) indicating satisfaction with the work environment.

Gender: Categorical value ('Male' or 'Female').

HourlyRate: Numeric value indicating the hourly rate of pay.

JobInvolvement: Numeric scale (1-4) indicating the level of job involvement.

JobLevel: Numeric value indicating the level of job hierarchy.

JobRole: Categorical value indicating the role of the employee (e.g., 'Sales Executive', 'Research Scientist').

JobSatisfaction: Numeric scale (1-4) indicating job satisfaction.

MaritalStatus: Categorical value indicating marital status (e.g., 'Single', 'Married', 'Divorced').

MonthlyIncome: Numeric value indicating monthly income.

MonthlyRate: Numeric value indicating monthly rate of pay.

NumCompaniesWorked: Numeric value indicating the number of companies an employee has worked for.

Over18: Categorical value indicating if the employee is over 18 years old.

OverTime: Categorical value indicating if the employee works overtime ('Yes' or 'No').

PercentSalaryHike: Numeric value indicating the percentage increase in salary.

PerformanceRating: Numeric scale (1-4) indicating the performance rating.

RelationshipSatisfaction: Numeric scale (1-4) indicating satisfaction with relationships at work.

StandardHours: Likely a constant value indicating standard working hours.

StockOptionLevel: Numeric value indicating the level of stock options available to the employee.

TotalWorkingYears: Numeric value indicating the total number of years worked.

TrainingTimesLastYear: Numeric value indicating the number of training sessions attended last year.

WorkLifeBalance: Numeric scale (1-4) indicating the balance between work and personal life.

YearsAtCompany: Numeric value indicating the number of years spent at the company.

YearsInCurrentRole: Numeric value indicating the number of years in the current role.

YearsSinceLastPromotion: Numeric value indicating the number of years since the last promotion.

YearsWithCurrManager: Numeric value indicating the number of years with the current manager.

The dataset comprises over 1,400 records and includes both categorical and numerical variables, making it suitable for a range of data analysis techniques.

Results

Random Forest Classifier:

- Training Accuracy: 99.72%
- Confusion Matrix:

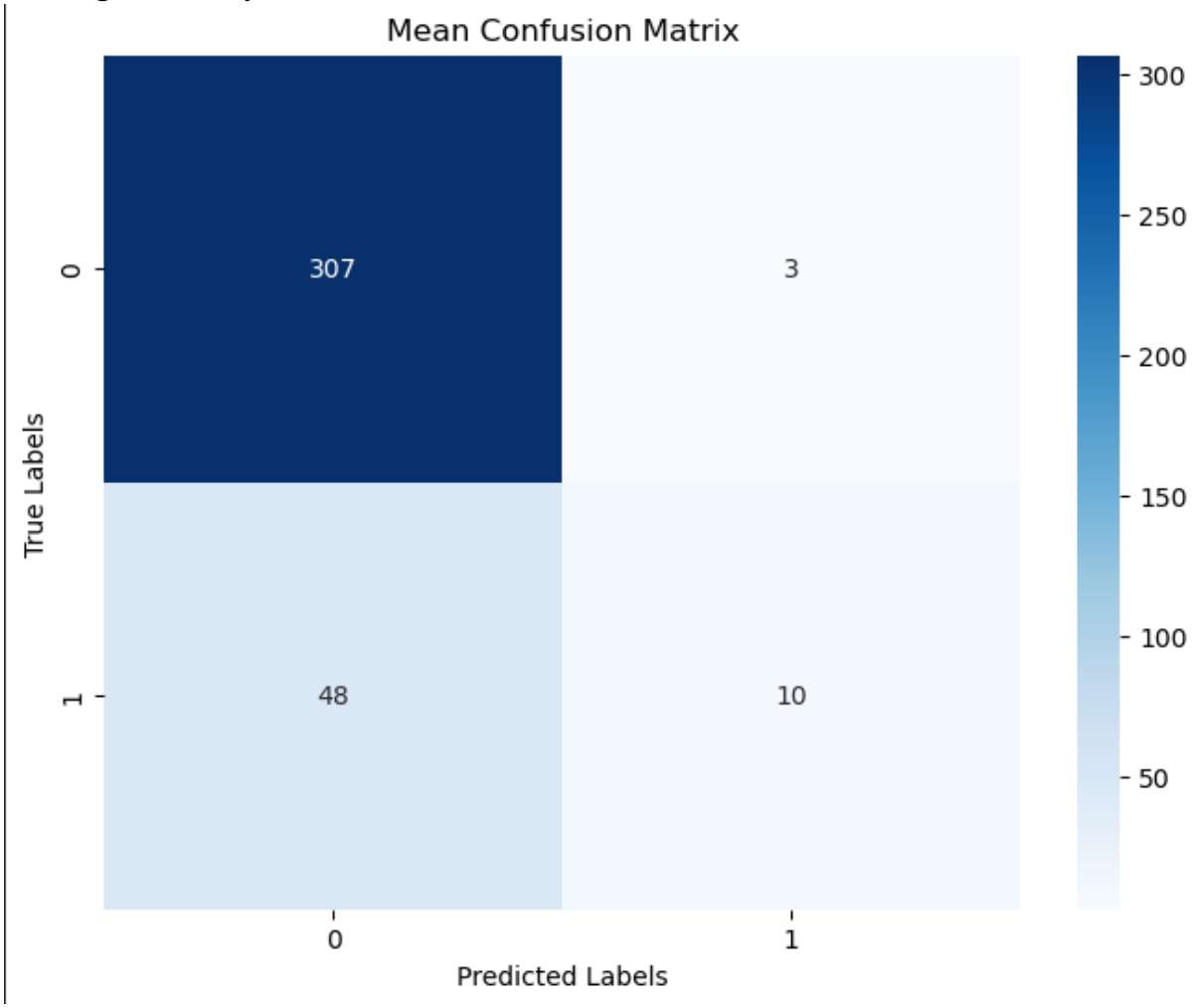
    True Negatives (TN): 302

    True Positives (TP): 11

    False Negatives (FN): 47

    False Positives (FP): 8

- Testing Accuracy: 85.32%



Gradient Boosting Classifier:

- Fitted with the parameter random_state=42.
- Overall Metrics:

    Accuracy: 85%

    Macro Average Precision: 0.72, Recall: 0.60, F1-Score: 0.64

    Weighted Average Precision: 0.82, Recall: 0.85, F1-Score: 0.82

XGBoost Classifier:

- Fitted with parameters objective='multi:softmax', num_class=3, random_state=42, and n_estimators=100.
- Overall Metrics:

    Accuracy: 85%

    Macro Average Precision: 0.73, Recall: 0.61, F1-Score: 0.64

    Weighted Average Precision: 0.83, Recall: 0.85, F1-Score: 0.83

K-Fold Cross-Validation with XGBoost:

- Fold 1: Accuracy = 0.84
- Fold 2: Accuracy = 0.87
- Fold 3: Accuracy = 0.84
- Fold 4: Accuracy = 0.89
- Fold 5: Accuracy = 0.90
- Fold 6: Accuracy = 0.86
- Fold 7: Accuracy = 0.87
- Fold 8: Accuracy = 0.88
- Fold 9: Accuracy = 0.87
- Fold 10: Accuracy = 0.85
- Average Accuracy for XGBoost: 0.87

Key Insights

Feature Importance

Analysing feature importance reveals the most significant factors influencing attrition:

- **Work-Life Balance**: Employees with poor work-life balance are more likely to leave.
- **Job Satisfaction**: Lower satisfaction scores correlate strongly with attrition.
- **Monthly Income**: Employees in the lower income bracket are at higher risk of attrition.
- **Age**: Younger employees show a higher tendency to leave, possibly due to career advancement opportunities.

Patterns and Trends

- High attrition rates are observed in specific job roles, such as Sales Representatives.
- Employees in lower job levels or with fewer years at the company exhibit higher turnover rates.

Based on the analysis, the following strategies could be used to reduce attrition:

1. **Enhance Work-Life Balance**:

- Offer flexible working hours and remote work options.
- Introduce wellness programs to address employee burnout.

2. **Improve Job Satisfaction**:

- Conduct regular surveys to gauge satisfaction levels.
- Address concerns through targeted interventions.

3. **Review Compensation Policies**:

- Benchmark salaries against industry standards.
- Provide performance-based incentives to retain talent.

4. **Career Development Opportunities**:

- Implement mentorship programs and training sessions.
- Create clear pathways for career progression.

5. **Targeted Retention Strategies**:

- Focus on at-risk groups, such as younger employees or those in high-turnover roles.
- Offer tailored benefits to address specific needs.

References

- Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition by Aurélien Géron
- https://www.youtube.com/playlist?list=PLZoTAELRMXVPQyArDHyQVjQxjj_YmEuO9 (Complete EDA play list by Krish Naik)
- https://github.com/Aniketg1998/Power-BI_HR-Analytics
- https://github.com/rasmodev/Employee-Attrition-Prediction
- N. Bhartiya, S. Jannu, P. Shukla and R. Chapaneri, "Employee Attrition Prediction Using Classification Models," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 2019, pp. 1-6, doi: 10.1109/I2CT45611.2019.9033784. keywords: {Libraries;Organizations;Data visualization;Data models;Support vector machines;Decision trees;Machine learning;Attrition;Machine Learning;Supervised Learning;Data Analysis;Classification Models},
- International Journal of Artificial Intelligence and Applications (IJAIA), Vol.15, No.2, March 2024 DOI:10.5121/ijaia.2024.1520223 EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING MODELS: A REVIEW PAPER Haya Alqahtani, Hana Almagrabi and Amal Alharbi
- https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset