IBM Innovation Centre for Education

SRI SRI UNIVERSITY
LEARN • LEAD • SERVE

PHEMESOFT
Software That Matters

Sri Sri University
Final Project Report
on
Employee Attrition Prediction

Team members:

Samriddhi Jain

Ayush Patil

Akash Samantray

Shrestha Kumar Agarwal

Guided by: **Signature:**

Mr. Sumit Kumar Shukla (IBM)

Dr. Sudhir Kumar Mohapatra (SSU)

Table of Contents

Executive Summary

The Employee Attrition Prediction project is an in-depth aimed at understanding and predicting employee attrition in a company. The project utilizes a dataset consisting of 35 variables and 1470 rows, including demographic information, work-related factors, and job satisfaction metrics.

The primary objective of the project is to identify patterns and trends that could potentially influence employee attrition. To achieve this, the project employs various data analysis techniques, including data visualization using Matplotlib, Seaborn and Plotly libraries.

The analysis reveals several key insights:

1. **Business Travel**: Employees who frequently travel for business have a higher attrition rate than those who do not.
2. **Education**: Employees with higher levels of education have lower attrition rates than those with lower levels of education.
3. **Environment Satisfaction**: Employees who are less satisfied with their work environment have higher attrition rates than those who are more satisfied.
4. **Gender**: No significant difference in attrition rates was observed between male and female employees.
5. **Job Involvement**: Employees who are less involved in their jobs have higher attrition rates than those who are more involved.
6. **Job Level**: Employees at lower job levels have higher attrition rates than those at higher job levels.
7. **Job Role**: The Sales and Research & Development roles have the highest attrition rates, while the Human Resources role has the lowest.
8. **Job Satisfaction**: Employees who are less satisfied with their jobs have higher attrition rates than those who are more satisfied.
9. **Marital Status**: Married employees have lower attrition rates than single employees.
10. **Monthly Income**: No clear pattern was observed between monthly income and attrition.
11. **Monthly Rate:** No clear pattern was observed between monthly rate and attrition.
12. **Num Companies Worked**: Employees who have worked for more companies have higher attrition rates than those who have worked for fewer companies.
13. **Overtime:** Employees who work overtime have higher attrition rates than those who do not.
14. **Percent Salary Hike:** Employees who received a lower percentage salary hike have higher attrition rates than those who received a higher percentage salary hike.
15. **Performance Rating**: Employees with lower performance ratings have higher attrition rates than those with higher performance ratings.

16. **Relationship Satisfaction:** Employees who are less satisfied with their relationships have higher attrition rates than those who are more satisfied.

17. **Standard Hours**: Employees who work more standard hours have higher attrition rates than those who work fewer standard hours.

18. **Stock Option Level:** Employees with lower stock option levels have higher attrition rates than those with higher stock option levels.

19. **Total Working Years**: Employees with fewer total working years have higher attrition rates than those with more total working years.

20. **Training Times Last Year**: Employees who received less training have higher attrition rates than those who received more training.

21. **Work Life Balance:** Employees who have a poorer work-life balance have higher attrition rates than those who have a better work-life balance.

22. **Years at Company:** Employees who have been at the company for fewer years have higher attrition rates than those who have been at the company for more years.

23. **Years in Current Role**: Employees who have been in their current role for fewer years have higher attrition rates than those who have been in their current role for more years.

24. **Years Since Last Promotion**: Employees who have been promoted more recently have lower attrition rates than those who have not been promoted recently.

25. **Years with Current Manager:** Employees who have been with their current manager for fewer years have higher attrition rates than those who have been with their current manager for more years.

These insights highlight the importance of various factors in employee attrition and can help organizations develop effective retention strategies. By understanding the underlying causes of attrition, companies can take proactive measures to retain valuable employees and improve overall employee satisfaction.

Background

The Employee Attrition Prediction EDA project focuses on analyzing a dataset containing information about employees to predict and understand factors influencing attrition within a company. Employee attrition, the rate at which employees leave a company, is a critical metric for organizations to manage effectively. By conducting an exploratory data analysis (EDA) on this dataset, valuable insights can be gained to help organizations develop strategies to reduce attrition rates and improve employee retention.

Aim

The primary aim of this project is to explore the dataset thoroughly to identify patterns, correlations, and trends that may impact employee attrition. By analyzing various factors such as age, job satisfaction, work environment, and job role, the project aims to uncover key insights that can assist in predicting and understanding employee attrition. Ultimately, the goal is to provide actionable recommendations to help organizations improve employee satisfaction and retention.

Technologies

The project utilizes several key technologies to conduct the EDA and analyze the dataset effectively:

- Python: The programming language used for data manipulation, analysis, and visualization.
- Pandas: A Python library for data manipulation and analysis, used to load and explore the dataset.
- Matplotlib and Seaborn: Python libraries for data visualization, employed to create insightful plots and charts to represent the data.
- NumPy: A fundamental package for scientific computing in Python, utilized for numerical operations and calculations.

Hardware Architecture

The hardware architecture for this project is flexible and can be implemented on various systems. Since the project involves data analysis and visualization, a standard computer system with sufficient processing power and memory is recommended. The hardware requirements are not intensive, and the project can be executed on most modern laptops or desktop computers without any specialized hardware.

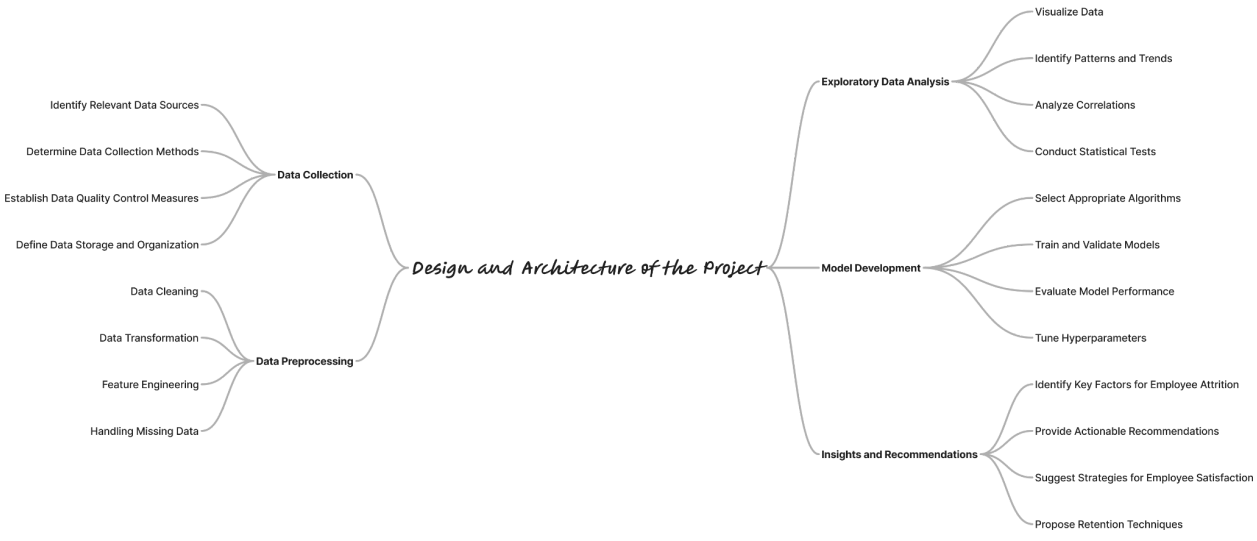
Software Architecture

The software architecture of the project involves the following components:

- Jupyter Notebook: The project is structured as a Jupyter Notebook, which allows for a combination of code, visualizations, and explanatory text in a single document.
- Python Environment: The project relies on a Python environment with the necessary libraries installed to run the code for data analysis and visualization.
- Data Processing Pipeline: The dataset is loaded using Pandas, and various data processing and visualization steps are performed to analyze the data effectively.
- Visualization Tools: Matplotlib and Seaborn are used to create visual representations of the data, such as plots, charts, and heatmaps, to facilitate data exploration and interpretation.

This software architecture enables a systematic and structured approach to conducting the EDA, ensuring that the analysis is comprehensive and the insights derived are meaningful for predicting employee attrition.

Design and Architecture



Data Collection

- Identify Relevant Data Sources: This involves figuring out where all the employee data resides within the company. This could include HR information systems, payroll systems, performance reviews, and employee surveys.
- Determine Data Collection Methods: In the real world, this might involve working with the IT department to extract data from various databases and integrating them into a central repository. You might also need to design and distribute employee surveys to gather additional data points.

Data Preprocessing

- Data Cleaning: This involves scrubbing the data to identify and fix errors like missing entries, inconsistencies, and outliers. For instance, an employee record might have an empty field for salary or an impossible value for years of experience.
- Data Transformation: This might involve converting data formats into a usable state. For instance, you might need to convert dates from text format to a numerical format suitable for statistical analysis.
- Feature Engineering: This involves creating new features from existing data that might be more predictive of employee attrition. For instance, you might create a new feature that calculates an employee's average performance review score over time.
- Handling Missing Data: You might decide to exclude records with missing data, fill them in with estimated values, or use statistical methods to account for them.

Model Development

- Select Appropriate Algorithms: Several machine learning algorithms are suited for employee attrition prediction, such as logistic regression, decision trees, random forests, and gradient boosting machines. The choice of algorithm would depend on the specific dataset and the desired outcomes.
- Train and Validate Models: The data is split into two sets: a training set and a testing set. The training set is used to train the machine learning model, and the testing set is used to evaluate the model's performance.
- Evaluate Model Performance: This involves assessing how well the model performs on the testing set. Common metrics include accuracy, precision, recall, and F1 score.
- Tune Hyperparameters: Hyperparameters are settings that control the behavior of the machine learning algorithm. Tuning involves adjusting these hyperparameters to optimize the model's performance.

Insights and Recommendations

- Identify Key Factors for Employee Attrition: By analyzing the results of the machine learning model, you can identify which factors are the most important predictors of employee attrition.
- Provide Actionable Recommendations: Based on the key factors identified, you can suggest specific strategies to reduce employee attrition. For instance, if low job satisfaction is a key factor, you might recommend initiatives to improve employee morale or offer more competitive salaries.

Implementation

The Employee Attrition Prediction EDA project is implemented using Python and various data analysis libraries such as Pandas, Matplotlib, and Seaborn. The following sections describe the implementation in detail.

1. Importing Libraries

The first step in the implementation is importing the necessary libraries. Pandas is used for data manipulation and analysis, Matplotlib and Seaborn are used for data visualization, and NumPy is used for numerical operations.

2. Loading the Dataset

The dataset is loaded using Pandas' read_csv function. The dataset contains 35 columns with 1470 rows of employee data. The columns include demographic information, work-related factors, and job satisfaction metrics.

3. Exploratory Data Analysis

The exploratory data analysis (EDA) is performed to identify patterns, correlations, and trends that may impact employee attrition. The EDA includes the following steps:

- Data Cleaning: The dataset is cleaned to remove any missing or inconsistent data.
- Data Visualization: Various plots and charts are created to visualize the data.
- Correlation Analysis: The correlation between different variables is analyzed to identify any significant relationships.

4. Data Cleaning

The dataset is cleaned to remove any missing or inconsistent data. The data cleaning process includes the following steps:

- Checking for Missing Values: The dataset is checked for any missing values. If missing values are found, they are either removed or imputed using appropriate methods.
- Checking for Data Consistency: The dataset is checked for any inconsistent data. If inconsistent data is found, it is corrected or removed.

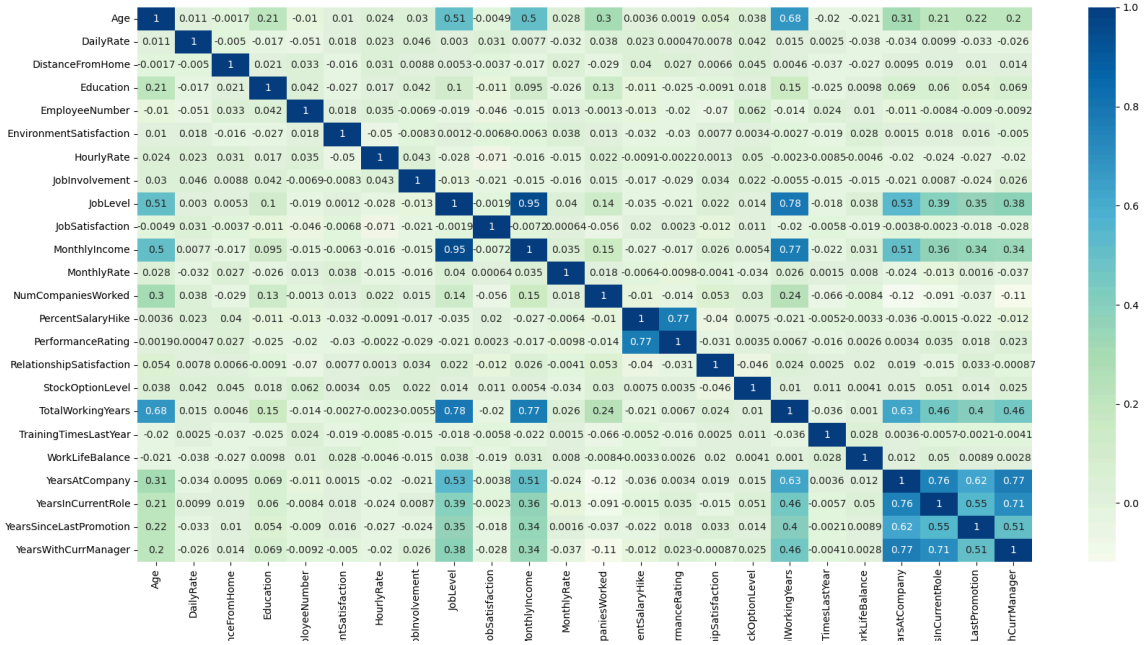
5. Data Visualization

Various plots and charts are created to visualize the data. The data visualization includes the following:

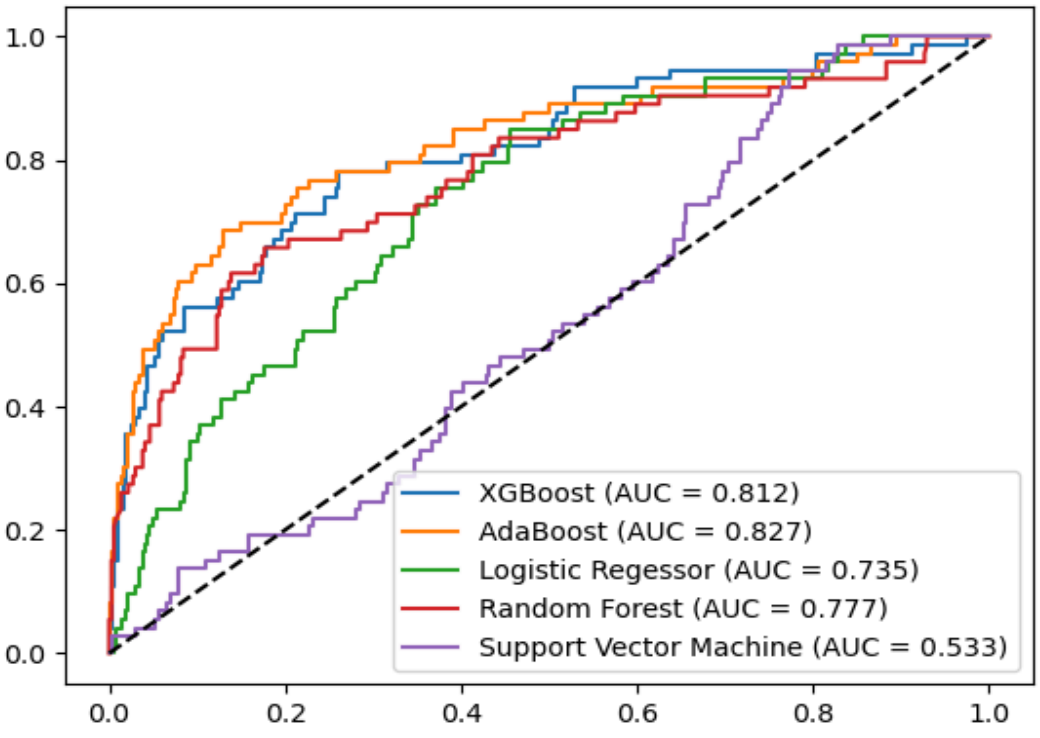
- Histograms: Histograms are created to visualize the distribution of continuous variables.
- Bar Plots: Bar plots are created to compare categorical variables.
- Scatter Plots: Scatter plots are created to visualize the relationship between two variables.
- Heatmaps: Heatmaps are created to visualize the correlation between different variables.

Snapshots of the project

1. Heatmap of Correlation Matrix



2. ROC Curve



Conclusion

In conclusion, the Employee Attrition Prediction project provided valuable insights into the factors influencing employee attrition in a company. The analysis of the dataset, consisting of 35 columns and 1470 rows, revealed key trends and patterns that could potentially impact employee attrition.

The project's primary objective was to identify patterns and trends that could potentially influence employee attrition, and this was achieved by employing various data analysis techniques, including data visualization using Matplotlib and Seaborn libraries.

The analysis revealed several key insights, including the impact of business travel, education level, environment satisfaction, gender, job involvement, job level, job role, job satisfaction, marital status, monthly income, monthly rate, number of companies worked, overtime, percentage salary hike, performance rating, relationship satisfaction, standard hours, stock option level, total working years, training times last year, work-life balance, years at company, years in current role, years since last promotion, and years with current manager on employee attrition.

These insights highlight the importance of various factors in employee attrition and can help organizations develop effective retention strategies. By understanding the underlying causes of attrition, companies can take proactive measures to retain valuable employees and improve overall employee satisfaction.

The project's findings can be used to inform HR policies and practices, including talent management, learning and development, and employee engagement strategies. By addressing the factors that contribute to employee attrition, organizations can improve employee retention, reduce turnover costs, and enhance overall organizational performance.

Further Research and Development

To further enhance the project and its applicability in real-world scenarios, several avenues for research and development can be explored.

1. Predictive Modeling

- Implement advanced machine learning algorithms like Random Forest, XGBoost, or Neural Networks to develop predictive models for employee attrition.
- Explore ensemble methods to improve model accuracy and robustness.

2. Feature Engineering

- Conduct in-depth feature engineering to create new variables that could better capture the nuances of employee attrition.
- Investigate interactions between different features to uncover hidden patterns.

3. Data Collection

- Expand the dataset by including additional variables such as employee performance metrics, project involvement, or feedback data.
- Incorporate external data sources like industry benchmarks or economic indicators to provide a broader context.

4. Natural Language Processing (NLP)

- Analyze unstructured data like employee reviews, feedback, or comments using NLP techniques to extract valuable insights.
- Develop sentiment analysis models to understand employee sentiments and their impact on attrition.

5. Time-Series Analysis

- Apply time-series analysis to understand attrition trends over time and forecast future attrition rates.
- Investigate seasonal patterns or cyclical trends in attrition behavior.

6. Interactive Dashboards

- Create interactive dashboards using tools like Tableau or Power BI to visualize attrition trends dynamically.
- Enable stakeholders to explore data interactively and derive actionable insights.

7. Ethical Considerations

- Conduct an ethical analysis to ensure fairness, transparency, and accountability in the use of predictive models for employee attrition.
- Address potential biases in the data and model predictions to prevent discriminatory outcomes.

8. Employee Engagement Strategies

- Integrate the project findings with HR strategies to develop personalized retention plans for employees at risk of attrition.
- Implement proactive measures based on predictive insights to improve employee satisfaction and reduce turnover.

9. Benchmarking and Validation

- Compare the project results with industry benchmarks and validate the predictive models against external datasets.
- Ensure the generalizability and reliability of the models across different organizational contexts.

10. Longitudinal Studies

- Conduct longitudinal studies to track employee attrition patterns over an extended period and assess the long-term effectiveness of retention strategies.
- Monitor the impact of organizational changes or interventions on attrition rates.

References

1. Pandas: https://pandas.pydata.org
2. Matplotlib: https://matplotlib.org
3. Seaborn: https://seaborn.pydata.org
4. NumPy: https://numpy.org
5. https://medium.com
6. www.analyticsvidhya.com/blog/2020/08/bias-and-variance-tradeoff-machine-learning