Project Report: Predicting Employee Attrition- HR Analytics

Introduction

Employee attrition significantly threatens business continuity and profitability, leading to increased costs and lost productivity. Recognizing the need for proactive strategies, this report analyzes and predicts employee turnover. Machine learning techniques are used to identify the key factors that influence an employee's decision to leave. The insights gained will inform targeted retention strategies, offering value to HR departments in making data-driven decisions.

Abstract

This project developed a machine learning-based predictive model for high employee turnover using Python and its data science libraries. Key factors influencing attrition were identified, and several machine learning models were evaluated, with Random Forest demonstrating the strongest performance. The resulting insights provide a framework for organizations to implement targeted and data-driven retention strategies.

Problem

High employee attrition leads to increased costs associated with recruitment, onboarding, and lost productivity, hindering organizational growth.

Objective

To identify the primary factors influencing employee turnover and to develop accurate machine learning models capable of predicting which employees are likely to leave the organization.

Methodology

A publicly available HR dataset from Kaggle was preprocessed, explored, and used to train and evaluate Logistic Regression, Random Forest, and Support Vector Machine models. Feature selection using Recursive Feature Elimination (RFE) was employed. Model performance was assessed using accuracy, precision, recall, F1-score, and ROC curves.

Findings

The Random Forest model demonstrated the highest accuracy (0.978) and superior recall and precision in predicting employee turnover compared to Logistic Regression and Support Vector Machine. Key predictors identified include satisfaction level, last evaluation, time spent at the company, work accident history, promotion in the last 5 years, department (specifically Research and Development, HR, and Management), and salary level (high and low).

Tools Used

This project utilized the following tools and libraries:

- **Python:** The primary programming language for data analysis and model building.
- **Pandas:** For data manipulation and analysis, including reading and structuring the dataset.
- **NumPy:** For numerical computations and array operations.

- **Scikit-Learn (sklearn):** A comprehensive machine learning library used for:
    - **Data Preprocessing:** Creating dummy variables, train-test split.
    - **Feature Selection:** Recursive Feature Elimination (RFE).
    - **Model Building:** Logistic Regression, Random Forest Classifier, and Support Vector Machine Classifier.
    - **Model Evaluation:** Accuracy score, classification_report (precision, recall, F1-score), confusion_matrix, roc_auc_score, roc_curve, and cross_val_score.
- **Matplotlib and Seaborn:** For data visualization, including bar charts (pd.crosstab.plot), stacked bar charts, histograms (hr.hist), heatmaps (sns.heatmap), and ROC curves (plt.plot).
- **Business Intelligence Tools:** Power BI, For visualizing and reporting insights to stakeholders.

Steps Involved in Building the Project

1. Obtaining the data.
2. Scrubbing/Cleaning the data: This includes data imputation for missing or invalid data and fixing column names.
3. Exploring the data: Understanding the dataset, identifying outliers, and analyzing the relationship between explanatory variables and the response variable using a correlation matrix.
4. Modeling the data: Building a model to predict employee attrition.
5. Interpreting the data: Drawing conclusions about factors contributing to employee turnover and the relationships between variables.

Conclusion:

This project demonstrates the value of machine learning, particularly Random Forest, in predicting employee turnover. Key attrition drivers include employee satisfaction, tenure, evaluation, work accidents, promotions, department, and salary. The resulting insights empower HR to develop data-driven retention strategies and implement timely interventions. Integrating such models with strategic HR practices fosters a positive work environment, encourages long-term commitment, and contributes to organizational success.