Employee Attrition Prediction and Recommendations

Problem Statement

A large telecommunications company employs around 4,000 employees. Each year, approximately 15% of their employees leave and need to be replaced. The management views this attrition level as problematic due to the following reasons:

- **Project Delays**: Attrition leads to delays in projects, resulting in a loss of reputation among consumers and partners.
- **Resource Intensive Recruitment**: A significant amount of resources is needed to maintain a large recruitment department.
- **Reduced Productivity**: Work productivity and effectiveness are reduced due to the onboarding period for new staff.

To address these issues, the company has contracted a workplace engineering and analytics firm to identify the factors contributing to high attrition and to recommend changes for better employee retention. Additionally, given the limited resources, the company seeks to know which variables are most important and should be addressed immediately.

Goal

The primary goal is to model the probability of employee attrition and present the findings and recommendations to the senior management team. These insights will help the management understand what workplace changes are necessary to reduce the current attrition rate.

Data Import and Preprocessing

Imported Required Libraries

- **Pandas, NumPy**: For data manipulation.
- **Matplotlib, Seaborn**: For data visualization.
- **Sklearn**: For machine learning models and metrics.

Imported Data

- Read data dictionaries and employee datasets from CSV and Excel files.
- Ensured comprehensive coverage by including general employee data, employee survey data, manager survey data, and employee time-tracking data.

Data Alignment

- Renamed 'Unnamed: 0' columns in the In-Time and Out-Time datasets to 'EmployeeID' for consistency.
- Verified alignment to ensure accurate analysis of employee attendance and work patterns.

Data Overview

General Employee Data

- Overview of employee demographics, job details, and performance metrics.
- Included data types, non-null counts, and memory usage for efficient data handling.

Employee Survey Data

- Insights into employee satisfaction levels across environmental factors, job satisfaction, and work-life balance.
- Detailed missing values and data completeness.

Manager Survey Data

- Analysis of job involvement and performance ratings provided by managers.
- Description of data structure and usage.

In-Time and Out-Time Data

- Detailed examination of employee attendance records.
- Comprehensive understanding of data dimensions and memory usage for scalability.

Calculating Average Hours Worked and Days Worked

Data Transformation and Calculation

- Converted date columns in the In-Time and Out-Time datasets to datetime format.
- Calculated total hours worked each day for each employee, considering potential NaN values due to missing or incomplete data.
- Derived average hours worked per day and computed the number of days worked based on available attendance records.

Summary of Hours Worked

- Created a summarized dataset ('hours_summary') containing EmployeeID, Average Hours Worked per Day, and Total Days Worked.
- Displayed the head of 'hours_summary' to provide initial insights into employee work patterns.

Merging Survey and Hour Summary Data

Importance of Employee Survey Data (es)

- Provides critical insights into employee satisfaction levels, influencing morale and retention strategies.

Importance of Manager Survey Data (ms)

- Offers data on job involvement and performance ratings, reflecting managerial effectiveness and its impact on employee satisfaction.

Importance of Hour Summary Data

- Integrates average hours worked and number of days worked, enhancing analysis of employee productivity and engagement.

Merged Data Analysis (df)

- Merged Employee Survey Data (es) with General Employee Data (gen) on EmployeeID.

- Integrated Manager Survey Data (ms) and Hour Summary Data (hours_summary) to provide comprehensive insights into factors affecting attrition.
- Verified the merge by displaying the initial rows of the merged dataset ('df').
- Copied the original dataset ('original_df') for preservation and future analysis.

Data Processing and Encoding

Handling Missing Values

- Identified columns with missing values and their respective counts.
- Implemented imputation strategies to fill missing values based on relevant group medians:
  - NumCompaniesWorked and TotalWorkingYears were imputed using group medians based on Age, YearsAtCompany, JobLevel, and JobRole.
  - EnvironmentSatisfaction, JobSatisfaction, and WorkLifeBalance were imputed with overall column medians.

Data Overview

- Displayed basic information, summary statistics, and unique values in categorical columns to understand data characteristics and distributions.

Dropping Unnecessary Columns

- Removed columns that do not contribute to predicting employee attrition:
  - EmployeeCount, StandardHours, and Over18 due to lack of variability or relevance.

Encoding Categorical Features

- Encoded the 'Attrition' column ('Yes' as 1, 'No' as 0) for binary classification.
- Applied One-Hot Encoding to categorical columns, dropping the first level to avoid the dummy variable trap and ensuring a dense array output.

Cleaned and Encoded DataFrame

- Created 'df_cleaned' by dropping unnecessary columns.
- Used OneHotEncoder to transform categorical features and concatenated encoded columns back to the dataframe.

Splitting Data into Training and Testing Sets

Features and Target Variable Separation

- **Features (X):** All columns except 'Attrition.'
- **Target (y):** 'Attrition' column indicating employee attrition (1 for 'Yes', 0 for 'No').

Data Splitting

- Divided the dataset into training and testing sets:
  - 80% for training (X_train, y_train).
  - 20% for testing (X_test, y_test).

Verification of Shape and Class Distribution

- Checked the dimensions (rows and columns) of both training and testing sets to ensure data integrity.
- Ensured balanced distribution of target classes ('Yes' and 'No' for attrition) across both sets.

Model Development

Implemented Machine Learning Algorithms

- Implemented various machine learning algorithms including Logistic Regression, Decision Trees, SVM, MLP, Random Forest, and Gradient Boosting.
- Evaluated model performance using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

Insights and Recommendations

- Identified key factors contributing to employee attrition through feature importance analysis.
- Proposed actionable recommendations for senior management to improve workplace conditions and reduce attrition rates.

Model Evaluation and Performance Metrics

Model Initialization

- Initialized the following machine learning models:
  - Logistic Regression

- Random Forest
- Decision Tree
- Gradient Boosting

Evaluation Function

- Defined a function to evaluate each model based on metrics such as Accuracy, Precision, Recall, F1-Score, and ROC AUC.

Metrics Collection

- Created a dictionary to store performance metrics for each model.
- Trained each model on the training data and evaluated its performance on the testing data.
- Collected and stored metrics for each model.

Model Performance Comparison - Combined Feature Importance Analysis

Insights on Employee Attrition Factors

Based on the top combined feature importances and their correlation with attrition, the following insights and actionable suggestions were derived:

1. **Average Hours Per Day**

Insight: Employees who spend more hours per day at work are more likely to experience attrition.

Suggestions: Implement policies promoting work-life balance and provide flexibility in working hours.

2. **Total Working Years**

Insight: Employees with fewer total working years tend to experience higher attrition rates.

Suggestions: Offer career development opportunities, mentoring programs, and incentives for long-term commitment.

3. **Age**

Insight: Younger employees are more prone to leaving the company.

Suggestions: Provide clear career paths, professional growth opportunities, and address younger employees' specific concerns.

4. **Distance From Home**

Insight: Employees living farther from the workplace show minimal correlation with attrition.

Suggestions: Explore options like remote work, transportation benefits, or housing assistance.

5. **Monthly Income**

Insight: Lower monthly income correlates with higher attrition rates.

Suggestions: Review and ensure fair compensation structures, offer performance-based bonuses, and maintain pay equity.

6. **Environment Satisfaction**

Insight: Employees with lower satisfaction in their work environment are more likely to leave.

Suggestions: Improve workplace culture, proactively address issues, and foster a positive work environment.

7. **Number of Companies Worked**

Insight: Employees with a history of working at more companies tend to have higher attrition rates.

Suggestions: Offer internal career growth opportunities, incentives for long-term commitment, and effectively address past employment experiences.

Recommendations for Enhancing Data Collection and Analysis Efficiency

To optimize data processes related to attrition analysis, consider implementing the following strategies:

1. **Automate Data Integration:** Implement automated scripts for regular data updates to ensure data freshness.
2. **Enhance Data Quality Checks:** Establish robust validation procedures to maintain data accuracy and consistency.
3. **Utilize Predictive Analytics:** Leverage historical data for predictive modeling to forecast attrition risks.
4. **Promote Cross-Functional Collaboration:** Foster alignment between HR, data analytics, and business teams to ensure comprehensive analysis and effective decision-making.
5. **Monitor Performance Metrics:** Define and monitor key performance indicators (KPIs) to evaluate the effectiveness of attrition reduction strategies.
6. **Collect Employee Feedback:** Regularly gather qualitative insights from employees to complement quantitative data, providing deeper context and actionable insights.

Specific Recommendations

For more efficient data collection and analysis related to attrition:

- **Standardize Reasons for Leaving:** Encourage employees to select predefined reasons (e.g., personal reasons, relocation, better opportunities) to facilitate standardized data collection. Utilize NLP or machine learning models to analyze "Other" responses for deeper insights.
- **Allow Employee Comments:** Provide employees an option to provide additional comments or suggestions when indicating reasons for leaving. Use NLP techniques to analyze these comments for hidden patterns or themes, guiding targeted retention strategies.

The project aims to empower the telecommunications company to reduce attrition rates by leveraging data-driven insights and actionable recommendations. By focusing on employee satisfaction and engagement, the company can enhance workplace conditions and optimize resource allocation effectively.

Implementing these practices enhances data-driven decision-making, improves employee retention strategies, and fosters a proactive approach to addressing attrition challenges effectively.

Submitted by

Niveditha Raju

Email: niveditha.cr.9@gmail.com