Project Report

Employee Attrition Prediction

Name: Manasi Sharma
Date: 12th May 2024

Abstract
Employee attrition poses significant challenges for organizations, affecting productivity, morale, and overall performance. In this project, we developed machine learning models to predict employee attrition and provide insights to mitigate its impact. Leveraging a dataset containing diverse employee features and attrition status, we conducted thorough data analysis, preprocessing, and model development. Logistic Regression, Random Forest, and Support Vector Machine (SVM) models were trained and evaluated using metrics such as accuracy, precision, recall, and F1-score. Logistic Regression emerged as the top performer in accuracy. Insights gleaned from the analysis underscored the pivotal role of factors like job satisfaction and work-life balance in influencing attrition rates. Recommendations included prioritizing these factors and implementing retention strategies to reduce attrition and foster a more engaged workforce. This project provides a valuable framework for organizations to proactively manage attrition and cultivate a thriving work environment.

Dataset Analysis

The dataset, sourced from Kaggle, comprises 35 columns or features and contained 1500 records. These features encapsulate a wide array of aspects pertaining to employees, encompassing demographic information, job satisfaction levels, work-life balance evaluations, and more. At the heart of the dataset lies the target variable 'Attrition', serving as a binary indicator of whether an employee has left the company or not.

Exploratory Data Analysis (EDA) uncovered several insights into the distribution and characteristics of features within the dataset. Notably, certain features such as job satisfaction and work-life balance exhibited discernible patterns that suggested their potential significance in influencing attrition rates. For instance, employees reporting lower levels of job satisfaction or struggling to maintain a healthy work-life balance may be more inclined to leave the company.

1. **Job Satisfaction:** Employees' perceptions of job satisfaction, measured through various metrics, could be a crucial determinant of attrition, with lower satisfaction levels potentially indicating a higher likelihood of departure.

2. **Work-Life Balance:** Balancing professional and personal commitments is paramount for employee well-being, and discrepancies in work-life balance may contribute to attrition if not adequately addressed by employers.

Data Preprocessing

The data preprocessing phase of this project involved several crucial steps to ensure that the dataset was ready for model development and analysis.

1. **Handling Null Values:** Although the dataset was pristine and free from null values, dealing with missing data is a common preprocessing task in data analysis projects. Techniques such as imputation, where missing values are replaced with appropriate estimates based on the data distribution, or removal of rows or columns containing null values, could have been applied if null values were present.

2. **Encoding Categorical Labels:** Categorical variables inherently pose a challenge for many machine learning algorithms that expect numerical input. To address this, categorical labels in columns such as 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole', 'MaritalStatus', 'Over18', and 'OverTime' were encoded using the one-hot encoding technique. One-hot encoding creates binary columns for each category

within a categorical variable, effectively transforming categorical data into a format that algorithms can process.

For example, the 'Gender' column with categories 'Male' and 'Female' would be transformed into two binary columns: 'Gender_Male' and 'Gender_Female', where each column contains a 1 if the corresponding category is present for that observation and 0 otherwise. This process was repeated for all categorical variables in the dataset to ensure uniformity in data representation and prevent bias towards any particular category.

Model Development

In this phase of the project, machine learning models were developed to predict employee attrition based on the dataset. The process involved several key steps, including data preparation, algorithm selection, model training, and evaluation.

1. **Data Preparation:**
   - The dataset was divided into two components: features (X) and the target variable (y), where the target variable 'Attrition' indicates whether an employee has left the company.
   - The target variable 'Attrition' was separated from the feature set X, leaving X to contain all other relevant features.
   - This separation facilitated the training and evaluation of machine learning models, with X serving as the input data and y as the target variable for prediction.

2. **Dataset Splitting:**
   - The dataset was split into training and testing sets using an 80:20 ratio, where 80% of the data was used for training the models and 20% for evaluating their performance.
   - This split ensured that the models were trained on a sufficiently large portion of the data while still retaining a separate dataset for unbiased evaluation.

3. **Algorithm Selection:**
   - Three machine learning algorithms were chosen for model development: Logistic Regression, Random Forest, and Support Vector Machine (SVM).
   - These algorithms were selected based on their suitability for binary classification tasks and their potential to capture complex relationships within the data.

4. **Model Training:**
   - Each selected algorithm was trained on the training data (X_train, y_train) using default or specified parameters.

- Logistic Regression was trained with a maximum number of iterations set to 5500 to ensure convergence.
- Random Forest was trained with a specified random state of 100 to ensure reproducibility of results.
- SVM was trained with a random state of 100 to control the randomization during training.

5. **Model Evaluation:**
    - After training, the performance of each model was evaluated using a set of evaluation metrics including accuracy, precision, recall, and F1-score.
    - Accuracy measures the proportion of correctly classified instances among the total instance, Precision measures the proportion of true positive predictions among all positive predictions, Recall measures the proportion of true positive predictions among all actual positive instances, and F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics and offering a holistic measure of model performance.
    - These metrics provided insights into the predictive capability of each model and their effectiveness in capturing attrition patterns within the dataset.
    - The evaluation was conducted on the testing set (X_test, y_test) to assess how well the models generalized to unseen data.

Evaluation Results

Algorithms	Accuracy (%)	Precision (%)	Recall (%)	F1-score
Logistic Regression	89.79	71.42	38.46	0.5
Random Forest	87.75	71.42	12.82	0.21
Support Vector Machine	86.73	75.22	86.73	0.8

In a comparative analysis, Logistic Regression (LR) outperformed Random Forest and Support Vector Machine (SVM) with an accuracy of 90%, demonstrating superior performance in precision, recall, and F1-score due to its ability to effectively capture the nuances of employee attrition patterns.