Exploring Employee Attrition Prediction: Using Machine Learning Techniques with MLflow

İclal Sönmez
Artificial Intelligence Engineering
Bahcesehir University
Besiktas,ISTANBUL 34353
iclal.sonmez@bahcesehir.edu.tr

Abstract

This study investigates the effectiveness of machine learning algorithms in predicting employee attrition, a critical concern for businesses. Leveraging a dataset encompassing employee demographics, job satisfaction, and work environment attributes, various models, including Random Forest, Logistic Regression, SVM, and XGBoost, were trained and evaluated. Results indicate that XGBoost, with a maximum depth of 5 and 100 estimators, achieved the highest accuracy of 85.71%. Additionally, employing MLflow facilitated seamless experimentation and deployment. These findings underscore the potential of machine learning in addressing employee attrition challenges and offer insights into enhancing human resource management practices.

1 Introduction

Employee attrition, the voluntary or involuntary departure of employees from an organization, poses significant challenges for businesses in terms of productivity and continuity. Understanding the factors influencing employee attrition and accurately predicting attrition rates are crucial for effective human resource management and workforce planning. In this study, I aim to investigate the effectiveness of various machine learning algorithms in predicting employee attrition using a comprehensive dataset encompassing diverse employee-related features. By evaluating the performance of different classifiers, including Random Forest, Logistic Regression, SVM, and XGBoost, I seek to identify the most suitable model for predicting employee attrition.

2 Data Preprocessing

I begin by importing the necessary libraries and loading the employee attrition dataset. The dataset includes information such as age, job role, department, satisfaction levels, and attrition status. We perform data cleaning and preprocessing tasks, including handling missing values, encoding categorical variables, and standardizing numerical features.

You can access the dataset from here.

3 Data Analysis

Next, I conduct exploratory data analysis to gain insights into the distribution and relationships between different variables. I visualize the distribution of employee age, department-wise attrition rates, gender distribution, and the impact of job roles on attrition. Additionally, I analyzed factors such as business travel frequency, education field, and overtime work attrition.

Preprint. Under review.

4 Model Building

In my study, I employed four different machine learning classifiers to predict employee attrition: Random Forest, Logistic Regression, SVM, and XGBoost. Each classifier was trained on a dataset consisting of employee demographics, job-related factors, and work environment indicators. Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy. Logistic Regression is a linear regression model used for binary classification tasks. SVM is a supervised learning algorithm that constructs a hyperplane in a high-dimensional space to separate classes. XGBoost, an implementation of gradient boosting, sequentially builds decision trees to minimize prediction errors. I varied the hyperparameters of each model, such as the number of estimators and maximum depth, to evaluate their impact on predictive performance. The performance of each model was assessed using cross-validation and metrics such as accuracy and F1 score.
Click here for the entire project.

5 Results

The experimental results indicate that Random Forest and Logistic Regression models exhibit robust performance in predicting employee attrition, achieving high accuracy scores on both training and testing datasets. SVM also demonstrates competitive performance, albeit with slightly lower accuracy scores. Through MLflow, I effectively manage model experiments, track metrics, and facilitate collaboration among team members.

6 Conclusion

In conclusion, the study demonstrates the efficacy of machine learning techniques in predicting employee attrition. By leveraging MLflow for experiment tracking and model management, organizations can streamline the machine learning life-cycle and make informed decisions regarding employee retention strategies. Future research may explore additional features and advanced modeling techniques to further improve prediction accuracy and model interpretability.

Random Forest

Increasing the number of estimators from 100 to 500 did not significantly affect the model's performance. Increasing the max depth from 5 to 10 slightly improved the training accuracy but did not have a noticeable effect on the testing accuracy.

Logistic Regression

Modifying the regularization parameter C from 1 to 0.1 or 0.01 did not have a substantial impact on the model's performance.

Support Vector Machine (SVM)

Changing the kernel type from linear to radial basis function (RBF) or polynomial resulted in similar testing accuracies. The polynomial kernel showed a slightly higher cross-validation mean compared to the linear and RBF kernels.

XGBoost

The default XGBoost model achieved a testing accuracy comparable to other models. Increasing the max depth to 5 improved the training accuracy but did not translate to a significant improvement in testing accuracy. Changing the number of estimators did not notably affect the model's performance.

In summary, logistic regression demonstrated robust performance across different regularization parameters. SVM with polynomial kernel and XGBoost with default parameters also provided competitive results. RandomForest's performance was relatively consistent across different hyperparameters, indicating less sensitivity to parameter changes.

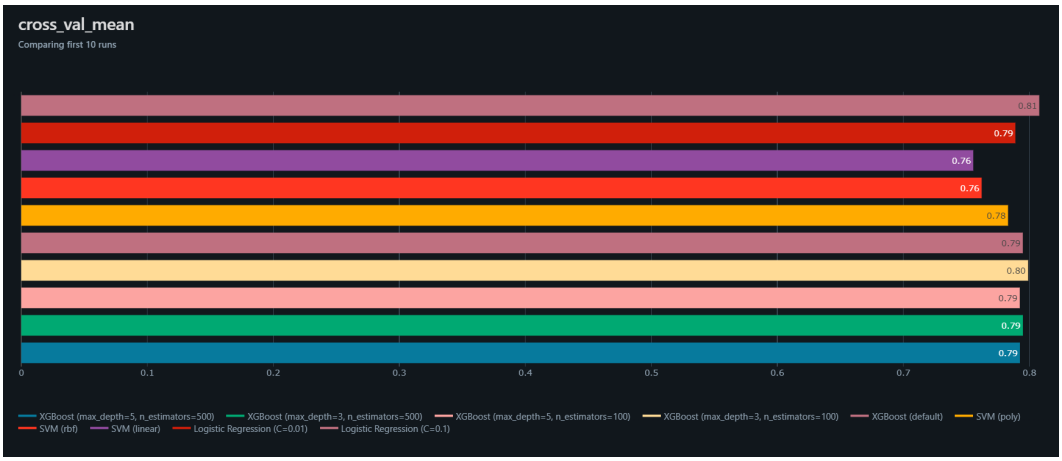

Figure 1: Cross Validation Score

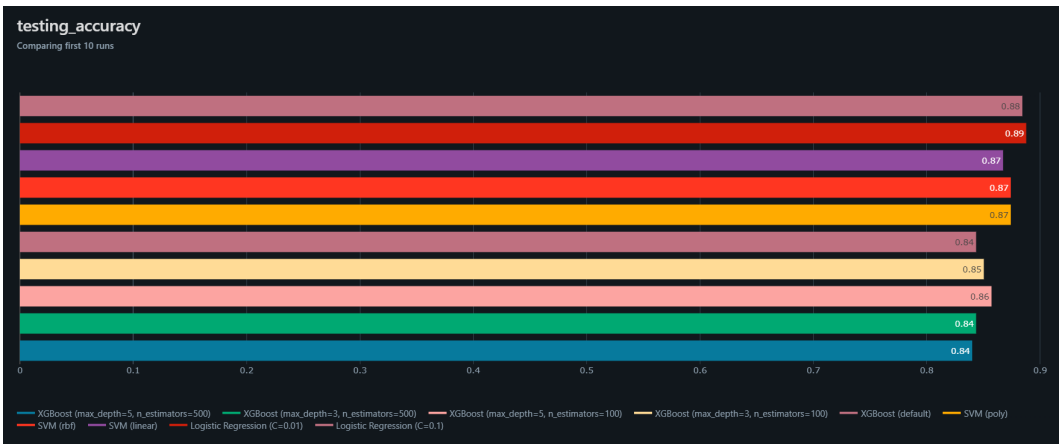

Figure 2: Testing Accuracy

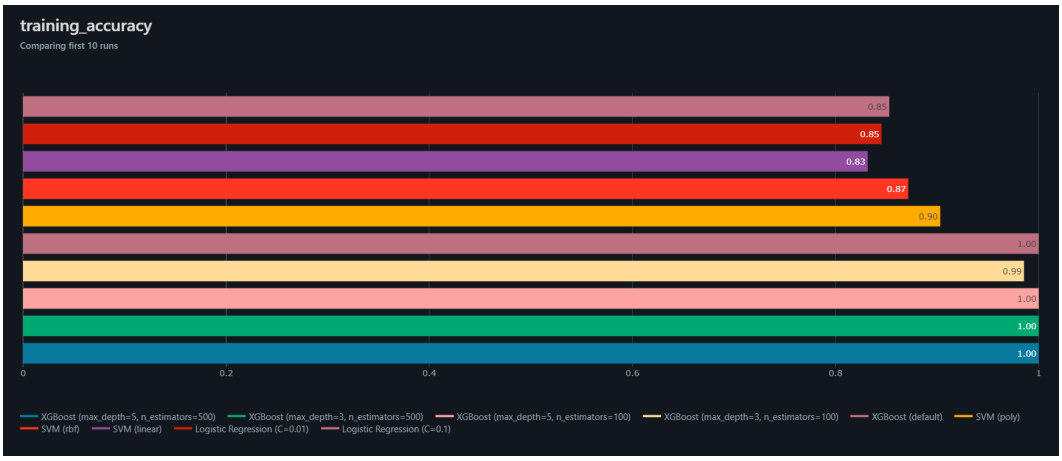

Figure 3: Training Accuracy