HM 4 Writeup

Xiang Li

Introduction

Employee attrition, or the phenomenon of employees leaving their jobs, is a significant concern for organizations. To gain insights into why some employees are more likely to leave a company than others, the HR department of a large software company has provided a dataset of 1,470 current and former employees. This dataset contains various features such as age, gender, tenure, education, and others. The company has tasked us with building two classifiers: a neural network and a boosted ensemble of trees, to predict employee attrition.

Data Preparation

Before delving into model analysis, we first cleaned and preprocessed the dataset. This involved selecting relevant features such as age, gender, education, job satisfaction, and performance ratings. Categorical variables were appropriately encoded, ensuring compatibility with the machine learning algorithms.

Methodology

For the **neural network classifier**, we utilize the **MLPClassifier** from the scikit-learn library. The model is configured with two hidden layers, employing rectified linear unit (ReLU) activation and the Adam optimizer. We evaluate the model's performance using accuracy score and visualize the results through a confusion matrix.

Next, we employ an **ensemble** approach with random forest regressors. We experiment with different numbers of trees (100 and 1000) to observe their impact on predictive accuracy. Additionally, we assess the importance of features in predicting employee attrition using feature importance plots generated from the random forest model.

Finally, we compare the performance of other boosting algorithms including AdaBoostClassifier, GradientBoostingRegressor, and XGBClassifier. These models are evaluated based on their root mean squared error (RMSE) to determine their predictive accuracy.

Results

Our analysis reveals that both the neural network and ensemble models provide reasonably accurate predictions of employee attrition. The neural network achieves an accuracy of approximately 0.864 to 0.866, while the random forest models exhibit RMSE values ranging from 0.339 to 0.342 with 100 and 1000 trees, respectively.

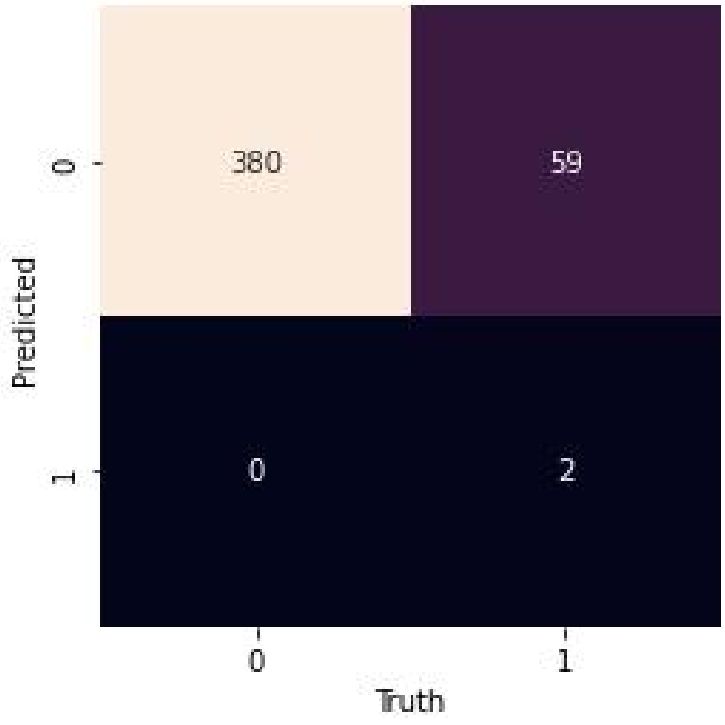
Feature importance analysis from the random forest model highlights variables such as job satisfaction, monthly income, and age as significant predictors of attrition. These findings align with existing literature, suggesting that job satisfaction, compensation, and career advancement opportunities play crucial roles in employee retention.

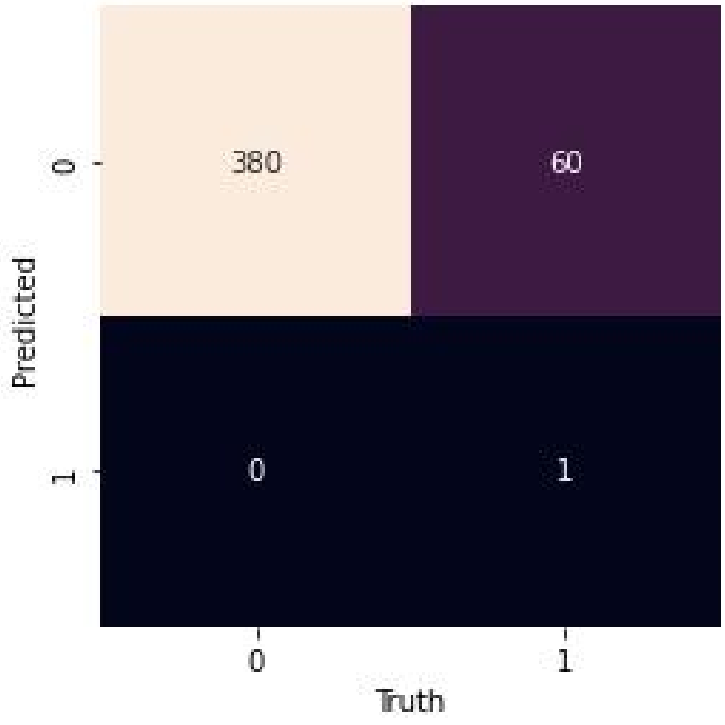
Comparing different boosting algorithms, we observe similar performance across AdaBoost, Gradient Boosting, and XGBoost, with RMSE values ranging from 0.339 to 0.392.

Conclusion

In conclusion, our analysis demonstrates the effectiveness of both neural network and ensemble models in predicting employee attrition. While the neural network offers simplicity and ease of interpretation, ensemble methods provide robustness and feature importance analysis. Understanding the factors contributing to employee attrition can assist organizations in implementing targeted retention strategies, thereby improving employee satisfaction and reducing turnover rates.

Appendix A MLPClassifier





Appendix B Ensemble