Employee Attrition Prediction Using Machine Learning

Brief by Jeongyeong Park(2032801) Kaggle Score: 0.90196

Introduction - **Machine learning** is a subfield of artificial intelligence (AI) and computer science that focuses on using data and algorithms to emulate how humans learn and steadily enhances its precision. **Data** now become the lifeblood of many industries. Machine learning technology takes an important role since it can achieve faster decision-making by enabling organizations to process and analyze data more rapidly than before. Various industries have utilized **Big data** to ensure industrial goals quickly. The data obtained from the analysis of big data can be used by machine learning to produce insightful business information.

Methodology – In Exploratory Data Analysis (EDA) step, information of data in train.csv file was checked and categorical features and numerical features were classified.

During data pre-processing step, based on EDA, unnecessary columns were dropped. Then outliers using interquartile range were examined. Instead of removing outliers, they were decided to be dealt with Robustscaler(). For categorical features, label encoding was done because they should be transformed into a numeric form so that can be read by machines. Finally, dataset was scaled with MinMaxScaler().

In order to find better-performing models, Gradient Boosting classifier, XGBoost classifier, and Logistic Regression models were trained and examined. The given train set was split into train and test sets. After deciding the most suitable model, feature importance of selected model was checked at last. Finally, the model was trained with final selected features of the whole data in train.csv to achieve attrition prediction of test.csv data.

Results – As a result of primary data preprocessing, 'EmployeeCount', 'StandardHours', 'Over18', and 'EmployeeNumber' columns were dropped.

Three models respectively have accuracy, recall, precision, and f1 score values as followed:

	Gradient Boosting	XGBoost	Logistic Regression
Accuracy	0.896	0.873	0.882
Recall	0.444	0.222	0.222
Precision	0.600	0.462	0.545
F1 Score	0.511	0.300	0.316

Table 1: Performance Comparison between the Gradient Boosting Classifier, XGBoost Classifier and Logistic Regression

Gradient Boosting (with n_estimators=800) models were selected for the final classification, and feature importance of the model was examined.

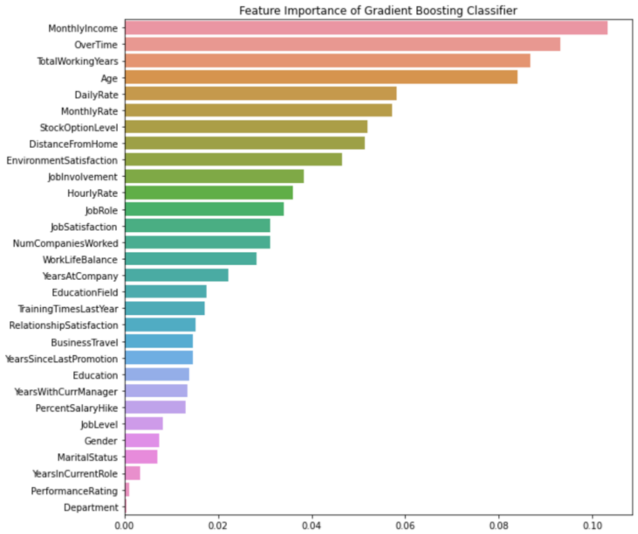

Figure 1: Feature importance of gradient boosting classifier

The last three columns that less affect the selected tree-based model were eliminated. As a result, the Gradient Boosting model was trained with the 27 features from the training dataset(train.csv). Based on the model prediction, among 367 employees, 34 employees are predicted to leave the company.

Discussion

- Pro: Gradient Boosting can be optimized with different loss functions and offers several hyperparameter tuning options.
- Contributor Thoughts: This pro can provide many options based on various datasets. However, high flexibility may require too much effort in tuning hyper-parameters.
- Con: It continues to improve to reduce existing errors. This can lead to overemphasizing outliers and cause overfitting.
- Contributor Thoughts: This con indicates that data preprocessing is still needed for Gradient Boosting.
- By precise EDA, data preprocessing and using Grid Search for tuning, this can be a great technology for attrition prediction.

Conclusion – Gradient Boosting Classifier is efficient and powerful model that can be used in employee attrition prediction. In order to get a more precise classifier, the dataset requires to be analyzed and pre-processed beforehand. With pre-processed data, Gradient Boosting can be tuned to finally achieve the best performance.