Will your employees leave? – Kaggle Competition

Introduction – Enterprises can suffer significant losses because of staff turnover, but it is extremely challenging to identify the influencing elements that can stop employee turnover. Therefore, the aim of this kaggle competition is to develop a machine learning model for predicting possible employee turnover which is beneficial for strengthening an organization's HR strategy.

Methodology – The methodology can be summarized in three sections: Data Preprocessing, Model Selection and Model Optimization.

A. Data Preprocessing - Pre-processing of the original data is a crucial step in data analysis. Firstly, data cleaning was performed, and single-category data were discarded Secondly, the **categorical** data were divided according to the number of their category types and were converted into **numerical** data using different coding methods.

Operation	Feature
Drop	`Over18`, `EmployeeCount`, `EmployeeNumber `, `StandardHours`
Binary-encode	`Gender`, `OverTime`
Ordinal-encode	`BusinessTravel`
One-hot encoding	`Department`, `EducationField`, `JobRole`, `MaritalStatus`

Table 1: Data Preprocessing

B. Classification Algorithm – Multiple models including Logistic Regression (LR), K-Nearest Neighbors, Decision Tree, Linear SVM, RBF SVM, MLP, Random Forest (RF), Gradient Boosting and XGBoost Classifier (XGB) were used, from which several models with better initial performance were selected for parameter tuning and **optimization**.

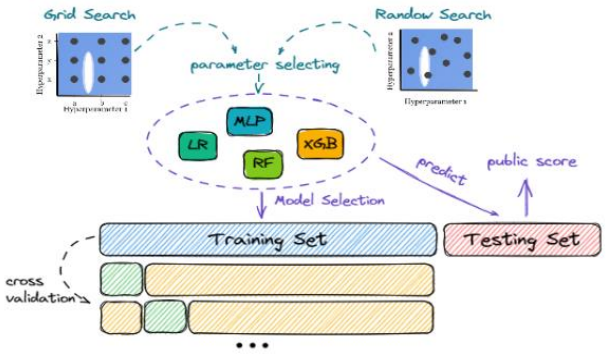

Figure 1: Project workflow

In addition, different data scaling methods (Standard, Robust and MinMax) were used to evaluate the initial performance of each model based on **cross-validation (cv)** score. Parameter tuning and model optimization were conducted by **random search**, **gird search** and **bayesian optimization**.

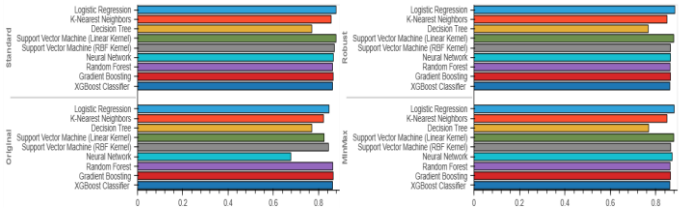

Figure 2: Initial attempt with multiple data scaling methods

Model	CV	Public Score
Logistic Regression	0.88396	0.85294
MLP	0.85498	---

Figure 3: Performance of other models

Final Solution – **XGB Classifier** which is a decision tree variation that can continuously generate new trees that rejoin the integration using the Boosting approach with many hyperparameters available for optimizing. The objective function of XGB Classifier is:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}^{(t-1)} + f_t(\mathbf{x}_i) + \Omega(f_t)$$

Results – The Best XGBoost Model got **88.2%** in the training dataset by **cross-validation** and scored **89.2%** in the kaggle public testing set. In the process of parameter selection, it was found that the model is extreme sensitive to certain parameters, thus, under certain conditions the model cannot to be optimized further.

Discussion –

- **Pro:** Overfitting is simple to produce in this dataset because of the limited sample size. However, XGBoost allows column sampling and adds a regular term to the cost function, which can significantly lessen overfitting.
- **Con:** The XGBoost algorithm has many parameters that make it difficult to tune and is not suitable for high-dimensional data because of its high memory consumption due to its high spatial complexity.
- Additionally, suitable pre-processing of the data is an irreplaceable factor that affects the training of the model, for example, if the **categorical** data is directly one-hot processed will lead to high dimensionality of the data, resulting in poor training results.

Conclusion – In conclusion, this project uses big data analytics and machine learning algorithms to build models that can predict employees' willingness to leave, which has important implications in corporate HR strategies.