Predicting whether an employee quit or not by ML

Brief by Chuanxin Zhai_1929631 (INT303 A2)

Introduction- In the A2 of INT303, we're required to write a machine learning program to judge whether an employee would quit or not, according to the given dataset. I used gradient-boosting regression (GBR) for the ML of this assessment. It produces a prediction model as an ensemble of weak prediction models, building the model in a stage-wise fashion. It could handle a wide of range of data types and structures, have highly scalable and high predictive power, and is easy to implement and tune. In addition, gbr can be parallelized to be trained on multiple machines and handle missing values and unbalanced classes. Overall, it is a wide-used and efficient technology for building predictive models on Big Data.

Methodology- Due to the quality and structure of the data, data preprocessing is an important and critical step in ML. In this case, can the weaken model become a stronger model. Here is the methodology for data preprocessing and classification algorithms.

Firstly, handling missing values because GBR cannot deal with them. Then complete outlier detection and removal which have negative impact. Too many features could lead to overfitting and reduced model performance, I must select a subset of the most relevant features. Next, conduct feature scaling because several gradient boosting algorithms are sensitive to the input scale. After doing these, you can split data, choose algorithm, train, evaluate and tune your model.

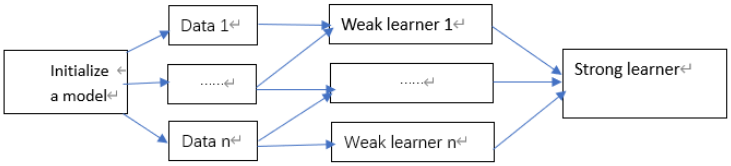

Figure 1: General workflow of GBR

As for classification algorithm, it should initialize the model with a constant value as the prediction. For each iteration, fit a weak learner to the residual errors (RE), update the current model and calculate the RE. Repeat the iteration until the desired number of iterations is reached or the improvement in the loss function is below a certain threshold. One of the key advantages of gradient boosting is that it can handle many features and is not prone to overfitting, as it has a built-in regularization through the shrinkage of the learning rate.

Results – As the result, there are various factors that have an impact on the performance, so the fine-tuning is significant for gradient boosting regression. These factors contain learning rate (determines the step size at which the model adjusts the weights of the weak learners), the choice of loss function (determines how the model measures the error between the predicted and actual values), the number of weak learners (determines the complexity of the model), the depth of the weak learners (determines how many splits each decision tree makes before making a prediction). Because of the limited space, other factors would not introduce. the methodology for data preprocessing and classification algorithms.

Here is a comparison table between gradient boosting regression and some other similar technologies:

Technology	Description	Advantages	Disadvantages
Gradient Boosting Regression	A machine learning technique for regression problems that uses an ensemble of weak learners to make predictions. It involves training a sequence of decision trees, each of which corrects the errors made by the previous tree.	Can handle a wide range of data types, including categorical and numerical data. Can handle missing values and handle large datasets efficiently. Can achieve high accuracy on many problems.	Can be sensitive to hyperparameter tuning. Can be prone to overfitting if not properly regularized. May take longer to train compared to some other techniques.
Random Forest Regression	A machine learning technique for regression problems that uses an ensemble of decision trees to make predictions. Each tree is trained on a random subset of the data, and the final prediction is made by averaging the predictions of all the trees.	Can handle a wide range of data types, including categorical and numerical data. Can handle missing values and handle large datasets efficiently. Can be less prone to overfitting than boosting algorithms.	May not be as accurate as some boosting algorithms on certain problems. Can be sensitive to hyperparameter tuning.
Linear Regression	A statistical technique for modeling the relationship between a dependent variable and one or more independent variables. It assumes that the relationship between the variables is linear, and the model is trained by minimizing the residual sum of squares between the observed responses and the predicted responses.	Can be very fast to train and predict. Can be interpretable, making it easy to understand the relationship between the variables.	Assumes a linear relationship between the variables, which may not always hold. Can be sensitive to outliers. May not perform well on non-linear problems.

Table 1: Comparison between three technologies (please scroll to zoom in)

Discussion- The pros and Cons of GBR

- Pro: Handle heterogeneous features that have different scales and distributions.
- Pro: It is a flexible method and able to model non-linear relationships between features and the target variable.
- Con: can be sensitive to the hyperparameters used, which can require a significant amount of tuning to get the best results.
- Con: can be time-consuming to train, especially when using many base models or when working with large datasets.

Conclusion – accuracy 0.83333 (best 0.86274)

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models. In conclusion, it has been successful in a variety of applications, and it is worth considering as a potential approach for my problem.