Yanhao Quan

Assignment 4 – Employee’s Attrition

**Data**

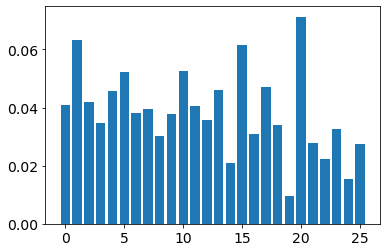
The company give me a data set of 1,470 current and former employees with information on whether or not they have left the company, their tenure, gender, education, and several other variables. Since not all variables are needed in the prediction, I use the correlation function and my intuition to pick 27 variables out of 35 variables. And for the variable I need to test on is not binary, I set them into binary. And for those which has more than two categories, I use encoder function to set them into integers. After that, I standardized them into decimal numbers bigger than 0 and less than 1.

**Models**

The company wants you to provide two classifiers: 1. a neural network that can predict employee attrition (variable name is attrition). 2. a boosted ensemble of trees that can predict employee attrition but also provide a ranking of feature importance. I used MLPClassifier from sklearn.neural\_network to predict employee attrition, and use random forest with 100 and 1000 trees to calculate Root Mean Squared Error. After that I create a feature importance bar chart to ranking the importance of all chosen 27 variables. Last, I use Ada Boost Classifier, Gradient Boosting Regressor, and XGB Classifier, calculate the Root Mean Squared Error and pick out the best one.

**Results**

The accuracy performance of MLP Classifier for the model with two hidden layers where the first layer has 10 nodes and the second has 5 is 88.2%, and we also run the regularized network, it turns out they have the same accuracy.

and in the ranking of **feature importance** graph, we could see that the column 20 is the most important feature which is **relationship satisfaction** ranked as most important feature. 

when it comes to Ada Boost Classifier, Gradient Boosting Regressor, and XGB Classifier, since **Ada boost has highest MSE, equals to 0.381,** then I will choose Ada booster.

The strengths of using neural net are non-linear problems with a complex decision boundary, the ability to approximate unknown functions, and no restrictions on the input variables. Weakness is they are a bit of a black box (hard to say how or why a network came up with a certain prediction).

The strengths of using boosted ensemble of trees are strong predictive power :boosting > random forest > bagging > CART; less prone to overﬁtting, and versatile and applicable to many models. However, the weaknesses are hard to parallelize due to recursive nature, can be expensive in terms of computational time; hard to scale up, and sensitive to outliers.