Data Mining Hw2, 2020 Fall

Man-Kwan Shan
TA: Tzu-Heng Huang & Pei-Ying Chen
zihengh1@gmail.com, 108753114@nccu.edu.tw

Homework Purpose:

- Employ open libraries to resolve a problem of employee attrition.
- Preprocess data and discover useful features for classification model.
- Evaluate your model performance with different measurements.

Homework Description: Please answer the following questions with the dataset we provide. You can download it from here. This dataset is released from IBM scientists, and it's able to analyze and discover potential factors that lead to employee attrition. The training data is for you to construct model, and the testing data is to evaluate your model performance. Each row represents an employee and contains his or her background, working performance, salary, and working environment. There are 20 types of numeric feature and 6 types of categorical feature. Your task is to predict who will stay in the company in the testing dataset. We also provide some slides (link) for you to understand the meaning for each column:

Q1 (70 points): According to the dataset, please train your classification model to predict the people in the testing data whether he or she will stay in the company. Evaluate your performance with the measurements of accuracy, precision, recall, and F1-score. Please discuss your results and illustrate your strategies, including the models you have tried and how you improve them. (P.S. showing your data pipeline is recommended.)

Q2 (30 points): According to the data distribution, it seems like an imbalance data. Please illustrate your strategy to resolve imbalanced problems. Also, please illustrate how you preprocess the features in your model. Comparing model performance with different strategies is encouraged.

Submission Rule: Please write your answers on PDf files with the title **StudentID_Hw2**. Submit your answers and your codes via wm5 before **December 21 23:59**.