FTEC5580 Project 2

Due 11:59pm, April 30, 2022

Instructions:

- Prepare a single Jupyter notebook (with code, results, your explanations and interpretations) and submit it in Blackboard. You CANNOT submit more than 2 times.
- Late submission incurs a penalty: 10% for submission on the first day after the deadline, 20% for submission on the second day after the deadline. Submissions made on the third day after the deadline and thereafter are NOT accepted.
- Name your Jupyter notebook as "last name-first name-P2", e.g., Li-Lingfei-P2. **Please follow the naming convention strictly.**
- You must work on the project independently.
- You should use Python as the kernel in Jupyter for the deep learning model and the logistic regression model, and use R as the kernel for the classification trees.
- The TA responsible for grading this project is *DAI Zhiwen*.

1 Data

This dataset (see the csv file) contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

There are 25 variables:

- ID: ID of each client
- LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
- SEX: Gender (1=male, 2=female)
- EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- MARRIAGE: Marital status (1=married, 2=single, 3=others)
- AGE: Age in years
- PAY_0: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 8=payment delay for eight months, 9=payment delay for nine months and above; you also see -2 in the dataset and its meaning is not explained by the data provider but it could refer to prepayment)

- PAY_2: Repayment status in August, 2005 (scale same as above)
- PAY_3: Repayment status in July, 2005 (scale same as above)
- PAY_4: Repayment status in June, 2005 (scale same as above)
- PAY_5: Repayment status in May, 2005 (scale same as above)
- PAY_6: Repayment status in April, 2005 (scale same as above)
- BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
- BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
- BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
- BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
- BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)
- BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
- PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
- PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
- PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
- PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
- PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
- PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
- default.payment.next.month: Default payment (1=yes, 0=no)

2 Task

We are interested in predicting defaults in the next month, which is October of 2005. Apply three models for this task and compare them: logistic regression, classification trees with boosting, deep learning (use a feedforward neural network). You need to randomly split the dataset into the training set and the test set.

(1) Report the overall error rate and AUC of three models on the training and test sets and compare them.

(2) Which variables are important for predicting defaults?

Remark: Sex, education and marriage are three categorical variables. One way to deal with such variables as inputs to a neural network is using one-hot vectors. The Jupyter notebook "NN-Data" provides you with the Python code to process the data.