AIFT — Final Project

Due and demo: June/11/2025

Problem Statement:

You are given a dataset containing 200 stocks for each year from 1997 to 2008 (discard the data for 年月=200912). In each year, each stock is described by 16 attributes:

市值(百萬元)
收盤價(元)_年
Unknown masked parameter
股價淨值比
股價營收比
淨值報酬率—稅後
資產報酬率 ROA
營業利益率 OPM
利潤邊際 NPM
負債/淨值比
流動比率
速動比率
存貨週轉率（次）
應收帳款週轉次
營業利益成長率
稅後淨利成長率

Also, two columns of class information are used to label each stock:

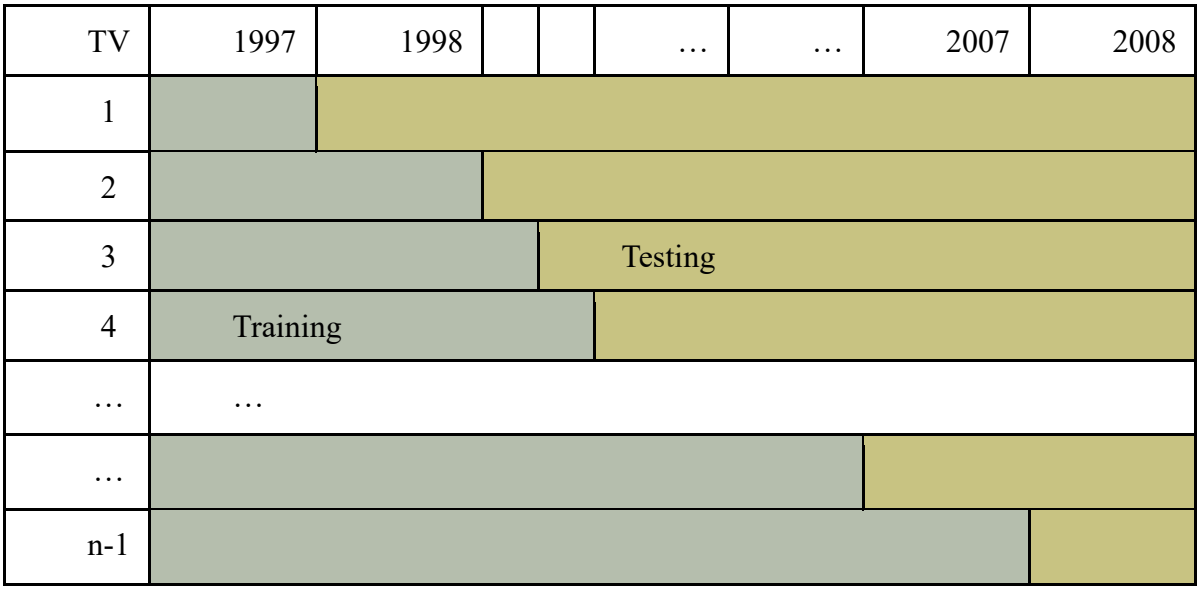
Return (Column T)
ReturnMean_year_Label (Column U)

For "Return" (in %), it is a real number larger than -100, which denotes the return of a stock for each year. For instance, the 2nd row of the dataset is for 台積電 (stock ticker 2330), and its column T is the return of 台積電 calculated from 199712 to 199812. For example, consider the stock prices of 台積電 on December 1, 1997 and December 1, 1998, denoted as P1 and P2, respectively; then the return is (P2-P1)/P1*100% = -6.3648%. As a result, the return of -100% means that the price of the stock becomes 0 (股票下市) sometime in the next year.

As for ReturnMean_year_Label, if the return of a stock is above the average return of all the stocks in the year, it is labeled 1, otherwise, it is labeled -1.

In this final project, please conduct the following tasks.

(1) (10%) Use the KNN (K-nearest neighbor) algorithm to find the best combination of parameter K and the attributes to select the stocks that yield the highest return you can get. You have to rely on the training and testing technique discussed in the class. For instance, in the figure below, for TV = 1, you use the data in 1997 as the training data to find the best combination of K and the attributes in terms of the best return of stocks you select, then use the best K and the attributes you find in the training phase to select the stocks in the testing phase and compute the resultant return. I.e., you shall come up with KNN-based stock selection models to pick up stocks with sound average return (the higher the average return, the better).



(2) (15%) Use decision-trees algorithms similar to ID3 to solve the same task above.

(3) (15%) Use similar algorithms as the one reported in Paper 2 to solve the same task above.

(4) (30%) Use any other algorithms (e.g., as the one reported in Paper 3) to solve the same task above.

(5) **(30%) Develop web crawlers to fetch similar datasets (e.g., from goodinfo.com) as the one given in this project and re-run the 4 tasks above.**

Requirement:

Upon demo, you will be given another set of testing data (the same format of the data above) to test the performance of your model.

For each group, please describe everyone's contribution to this project. The grades are calculated according to individual grade (個人分數 50%) and group grade (團隊分數 50%).

Finally, zip the files below and upload it to Moodle:

a) Source codes
b) Word file containing how you solve this problem, the results obtained, the discussions, including what you have learned
c) Demo PPT file