Assignment #4

Competition using Ensemble

Introduction to Big Data Analytics

TA: Dongmin Hyun and Junyoung Hwang

(dm.hyun@postech.ac.kr, jyhwang@postech.ac.kr)

Machine Learning Practice with Kaggle

- In this Homework, you will practice to implement, train, and tune an ensemble model on two datasets using Kaggle.

- Kaggle is a framework for evaluating the performance of ML models (you can find a lot of tutorials for Kaggle).

- General procedures
  - Train your ML model using given training data from given links
  - Produce predictions using the trained model on test data (You can find the output format as '[data name]_sample_submission.csv' from the given links).
  - Submit the output for test data to Kaggle to evaluate the performance of your model.

Notes

- Your score is based on your ranking at Kaggle (private leader board).
  - The ranking of public and private leader board could be different.

- You have to submit your code and report (1~2 pages) to LMS.
  - The report includes summary of your process and your submission score.

- You can use **any** python library for this competition.

- You must use your student ID number (ex. 20181234) as your team name.

- *You must read all competition rules carefully and comply with them.*

Kaggle links for two Datasets

- San Francisco Crime Dataset (Multi-class classification)
  https://www.kaggle.com/t/d26ca367d0a14d7f824b09d344b71a1a

- Bank Dataset (Binary Classification)
  https://www.kaggle.com/t/31ca7b486fb6491f8b6b75770dc1414a

Kaggle Competition (Join using the given links)

Kaggle Competition (Join Competition)

Kaggle Competition (Submit Predictions)

InClass Prediction Competition

Binary Classification using Ensemble
The classification goal is to predict if a client will subscribe a term deposit

14 days to go

Overview | Data | Kernels | Discussion | Leaderboard | Rules | Team | My Submissions | **Submit Predictions**

Overview

Description

Evaluation

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).

Please note: Only provided training data must be used to train your ensemble model.

Sources

[Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.

In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

7

Kaggle Competition (Submit Predictions)

You have 2 submissions remaining today. This resets 11 hours from now (00: 00 UTC).

Step 1
Upload submission file

Upload Files

File Format
Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

Number of Predictions
We expect the solution file to have 7234 prediction rows. This file should have a header row. Please see sample submission file on the data page.

Step 2
Describe submission

B *I* | % 66 </> 🖼 | ≣ ≣ H ⊨ | ↺ ↻   M↓ Styling with Markdown supported

Briefly describe your submission.

Make Submission

Kaggle Competition (Submit Predictions)

Step 1
Upload submission file

Upload Files

✔ **bank_sample_submission.csv (56.77 kB)**

File Format

Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

Number of Predictions

We expect the solution file to have 7234 prediction rows. This file should have a header row. Please see sample submission file on the data page.

Step 2
Describe submission

B *I* | ❀ 66 </> 🖼 | ≣ ☰ H ⊢ | ↺ ↻   M↓ Styling with Markdown supported

Briefly describe your submission.

Make Submission

Kaggle Competition (Check Score)

Python Library for Competitions

- Anaconda(https://www.anaconda.com/download/)
  - Almost all packages are installed in anaconda including below.

- Data preprocessing
  - Pandas (http://pandas.pydata.org/)
  - matplotlib (https://matplotlib.org/)

- Machine Learning algorithm
  - Scikit-learn (http://scikit-learn.org/stable/)
  - Numpy (http://www.numpy.org/)
  - …