Assignment 6 (15%): Real-World Data Mining Competition. (Due on Jan. 10, 2016)

Since we have learned a variety of data mining techniques so far, why not try to participate in a real-world data mining competition now in order to test your ability and see how data mining techniques are applied in the real-world.

In this assignment, you are asked to participate in an ongoing real-world data mining competition held by kaggle (a place where you can compete with world-class data miners on real data mining applications, and at the same time possibly earn a handsome prize). This year, you are asked to design and implement data mining methods to handle the problem: **Prudential Life Insurance Assessment (1st place - $15,000)**. **Please participate in the competition alone and teamwork is NOT allowed.**

You can click the Prudential Life Insurance Assessment to get all the information you need to accomplish the competition, including how to register, how to get the data, how to evaluate your method. To be explicit, we give a brief guidance for the newcomers of kaggle, which is described below.

Task

1. Preparations before beginning competition.
   1. Sign up a new account or use an existing Google account to sign in kaggle **here**.
   2. Find the competition homepage **here**, where you should read the description, evaluation, rules, prize and all the information you need to know in the left column of the web page.
   3. Accept the licence in **this page**. Namely, clicking on the “I understand and accept” at the bottom of Make a submission (or My Submissions) web page.
   4. To proceed, you must verify your Kaggle account via your mobile phone. Do not use a public number or share your number with others. Input your phone number (e.g., +8612345678901), then, your phone should receive your verification number. Please enter it and click the “Confirm”.
   5. **[Important]**: Modify the Team Name (the name which will be shown in the leaderboard) in **My Team** web page as your student number, e.g., MG1533001. You can use any account name as you like, but the Team Name that will appear in the leaderboard must be your student ID. *If your Team Name is not your student ID, your rank will NOT be recorded by us at deadline, and thus your score for rank will be zero.*

2. Participate in the competition and get higher rank as possible as you can.
   1. Log in kaggle with your account and find the **competition homepage**.
   2. Download the training and test data **here**.
   3. Use any method you have learned to build a model from training examples, and then use this model to predict the test samples. You can invoke other codes or tools to help you.
   4. Submit the prediction file in **Make a submission** web page (or **My submissions** web page) to evaluate your result. The submission format can be found **here** (sample_submission.csv). You can submit your result multiple times.
   5. If you have submitted one, you will get your rank instantaneously which will be shown

in the **leaderboard**. Try to get higher rank as possible as you can before deadline.

3. Write a report to describe your methods and implementation.
4. Submit your report, code, ReadMe, as well as the prediction file to FTP.

What should you submit to FTP

Pack all the files needed to be submitted, i.e., report.docx, code.zip, ReadMe.txt, prediction_on_test.csv. Before submitting your assignment to FTP, please read Submission Requirement and Description section above carefully and obey it.

[Important]: The prediction_on_test.csv should be in the same format as sample_submission.csv given by kaggle. The report should contain your final rank and result before deadline (2016-1-10 23:59:59) in the leaderboard, your name, student ID, and e-mail address. Name this pack using your student ID, e.g., MG1533001.zip.

[Important]: Submission Deadline: 2016-1-10 23:59:59. Please note that the deadline of assignment is before the deadline of competition. **You can still go on participating in the competition after Jan. 10, but we only record your rank and result at 2016-1-10 23:59:59 and results after 2016-1-10 23:59:59 will NOT be accepted by us.**

How to evaluate the assignment

We will evaluate your submission according to your rank and report. *If plagiarism is identified, no scores will be given to this assignment.*

For rank:

- At **2016-1-10 23:59:59** (please note that the deadline of this assignment is before the deadline of competition), we will get your *absolute rank* r_1 (the rank over all the participants p_1) and *relative rank* r_2 (the rank over all the students from this course p_2) from the leaderboard, and then your final score for the rank part is given by:

$$\text{Score}_{\text{rank}} \propto \frac{1}{2}\left(\frac{\sum_{p\in p_1} I[r_p > r_1]}{|p_1|-1} + \frac{\sum_{p\in p_2} I[r_p > r_2]}{|p_2|-1}\right),$$

  where $I[\cdot]$ is the indicator function, i.e., $I[\text{condition}] = 1$ if condition is true and 0 otherwise.
- Please **make sure** that the name shown in the leaderboard is your student ID. Otherwise, your rank will NOT be recorded by us at deadline, and thus your score for rank will be zero.

For report:

- Technique: clearly explain why you choose such method, how you implement the method, and how the method performs on this data mining task.
- Language: concise, precise, and logical.
- Organization: good structure, clearly and properly separated sections and paragraphs.
- Citations: all works of non-yourself should have correct references.

Page updated on 2015.12.4, via jemdoc.