MSBD 5013: Statistical Prediction **7 Mar, 2022**

Project 1: Midterm

Instructor: Yuan Yao *Due: 23:59 Friday 25 Mar, 2022*

1 Project Requirement

This project as a warm-up aims to explore basic techniques in machine learning.

1. Pick up ONE (or more if you like) favourite dataset below to work. If you would like to work on a different problem outside the candidates we proposed, please email course instructor about your proposal.

2. Team work: we encourage you to form small team, up to **THREE** persons per group, to work on the same problem. Each team just submit

   (a) *ONE report*, *with a clear remark on each person's contribution*. The report can be in the format of either a *poster*, e.g.

   `https://github.com/yuany-pku/2017_math6380/blob/master/project1/DongLoXia_poster.pptx`

   or *technical report within 8 pages*, e.g. NIPS conference style (preferred format)

   `https://nips.cc/Conferences/2019/PaperInformation/StyleFiles`,

   with source codes such as Python (Jupyter) Notebooks with a detailed documentation.

3. For Kaggle contests, if possible, please register your team with name in the format of msbd5013_lastname, so that we could easily find your results on Kaggle. For example, a team with Shawn Zhu and Kate Wong would be named by msbd5013_Zhu_Wong.

4. In the report, show your proposed scientific questions to explore and main results with a careful analysis supporting the results toward answering your problems. If possible, you should include your Kaggle contest score and/or rating in the report. Remember: scientific analysis and reasoning are more important than merely the performance tables. Separate source codes may be submitted as a GitHub link, or a zip file.

5. Submit your report and/or source codes via Canvas no later than the deadline.

2 Candidates

2.1 Kaggle Contest: Home Credit Default Risk

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data–including telco and transactional information–to predict their clients' repayment abilities.

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

Visit the following website to join the competition.

`https://www.kaggle.com/c/home-credit-default-risk/`

Requirements. For Kaggle contests, if possible, please register your team with name in the format of msbd5013_lastname, so that we could easily find your results on Kaggle. For example, a team with Shawn Zhu and Kate Wong would be named by msbd5013_Zhu_Wong.

2.2 Kaggle: M5 Forecasting

Can you estimate, as precisely as possible, the point forecasts of the unit sales of various products sold in the USA by Walmart?

How much camping gear will one store sell each month in a year? To the uninitiated, calculating sales at this level may seem as difficult as predicting the weather. Both types of forecasting rely on science and historical data. While a wrong weather forecast may result in you carrying around an umbrella on a sunny day, inaccurate business forecasts could result in actual or opportunity losses. In this competition, in addition to traditional forecasting methods you're also challenged to use machine learning to improve forecast accuracy.

In this competition, you will use hierarchical sales data from Walmart, the world's largest company by revenue, to forecast daily sales for the next 28 days and to make uncertainty estimates for these forecasts. The data, covers stores in three US States (California, Texas, and Wisconsin) and includes item level, department, product categories, and store details. In addition, it has explanatory variables such as price, promotions, day of the week, and special events. Together, this robust dataset can be used to improve forecasting accuracy.

Visit the following website to join the Kaggle contest on Accuracy, Estimate the unit sales of Walmart retail goods:

`https://www.kaggle.com/c/m5-forecasting-accuracy`

Requirements. For Kaggle contests, if possible, please register your team with name in the format of msbd5013_lastname, so that we could easily find your results on Kaggle. For example, a team with Shawn Zhu and Kate Wong would be named by msbd5013_Zhu_Wong.

Peer Review

In this exercise of open peer review, please write down your comments of the *reports rather than of your own team* in the following format. Be considerate and careful with a precise description, avoiding offensive language.

Deadline is *11:59pm Saturday April 9, 2022*. Submit your review in plain text to **Canvas**. Rebuttal is open afterwards.

- Summary of the report.
- Describe the strengths of the report.
- Describe the weaknesses of the report.
- Evaluation on Clarity and quality of writing (1-5): Is the report clearly written? Is there a good use of examples and figures? Is it well organized? Are there problems with style and grammar? Are there issues with typos, formatting, references, etc.? Please make suggestions to improve the clarity of the paper, and provide details of typos.
- Evaluation on Technical Quality (1-5): Are the results technically sound? Are there obvious flaws in the reasoning? Are claims well-supported by theoretical analysis or experimental results? Are the experiments well thought out and convincing? Will it be possible for other researchers to replicate these results? Is the evaluation appropriate? Did the authors clearly assess both the strengths and weaknesses of their approach? Are relevant papers cited, discussed, and compared to the presented work?
- Overall rating: (5- My vote as the best-report. 4- A good report. 3- An average one. 2- below average. 1- a poorly written one).
- Confidence on your assessment (1-3) (3- I have carefully read the paper and checked the results, 2- I just browse the paper without checking the details, 1- My assessment can be wrong)

Rebuttal

The rebuttal period starts from now, till *11:59pm Saturday April 16, 2022*. Restrict the number of characters of your rebuttal within **5,000**. Submit your rebuttal in PLAIN TEXT or PDF format to **canvas** with filename comprising the corresponding group number: e.g. rebuttal1_group02.pdf.

The following tips of rebuttal might be helpful for you to follow:

1. The main aim of the rebuttal is to answer any specific questions that the reviewers might have raised, or to clarify any misunderstanding of the technical content of the paper.

2. Keep your rebuttal short, to-the-point, and specific. In our experience, such rebuttals have the maximum impact.

3. Always be polite and professional. Refrain from name calling or rude comments, especially in response to negative reviews.

4. Highlight the changes in your manuscripts had you made a simple revision.