

## Mini-Project 1. Machine Learning Basics

*Instructor: Yuan Yao**Due: 23:59 Sunday 3 Oct, 2021*

## 1 Mini-Project Requirement and Datasets

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 **TWO** persons per group, to work on the same problem. Each team just submit **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](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, please register your team with name in the format of math4995\_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 math4995\_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 or rating in the report. Remember: scientific analysis and reasoning are more important than merely the performance tables. Separate source codes may be submitted through email as a GitHub link, or a zip file.
5. Submit your report by email or paper version no later than the deadline, to the following address (datascience.hw@gmail.com) with a title "MATH4995: Project 1"

## 2 Kaggle Contest: Predict Survival on the Titanic

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there were not enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this challenge, we ask you to build a predictive model that answers the question: “what sorts of people were more likely to survive?” using passenger data (*i.e.* name, age, gender, socio-economic class, etc). Visit the following website to join the Kaggle contest:

<https://www.kaggle.com/c/titanic>

**Requirements.** For Kaggle contests, please register your team with name in the format of math4995\_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 math4995\_Zhu\_Wong.

## 3 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, please register your team with name in the format of math4995\_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 math4995\_Zhu\_Wong.

## 4 Self Proposals

## 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 Oct. 28, 2021*.

You should put each review in a plain text separately with a title comprising the corresponding group number (**not your own group**) and your name (e.g., *rev1\_group03-Ian\_Goodfellow.pdf*). Submit all your reviews in a single zip file **using 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)