Capstone Project for Data Science **21 Oct, 2021**

Project 2. Final Project with Images and Text.

Instructor: Yuan Yao *Due: 23:59 Sunday 14 Nov, 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`

   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 2"

2 Kaggle Contest: PetFinder.my - Pawpularity Contest

PetFinder.my is a Malaysia leading animal welfare platform, featuring over 180,000 animals with 54,000 happily adopted. Currently, PetFinder.my uses a basic Cuteness Meter to rank pet photos. It analyzes picture composition and other factors compared to the performance of thousands of pet profiles. While this basic tool is helpful, it's still in an experimental stage and the algorithm could be improved.

In this competition, you will analyze raw images and metadata to predict the *Pawpularity* of pet photos. You'll train and test your model on PetFinder.my's thousands of pet profiles. Winning versions will offer accurate recommendations that will improve animal welfare. As a result, stray dogs and cats can find their "furever" homes much faster. With a little assistance from the Kaggle community, many precious lives could be saved and more happy families created.

Visit the following website to join the competition.

`https://www.kaggle.com/c/petfinder-pawpularity-score`

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: Natural Language Processing with Disaster Tweets

Twitter has become an important communication channel in times of emergency. The ubiquitousness of smartphones enables people to announce an emergency they are observing in real-time. Because of this, more agencies are interested in programatically monitoring Twitter (i.e. disaster relief organizations and news agencies).

But, it is not always clear whether a person's words are actually announcing a disaster.

In this competition, you are challenged to build a machine learning model that predicts which Tweets are about real disasters and which ones are not. You will have access to a dataset of 10,000 tweets that were hand classified.

Disclaimer: The dataset for this competition contains text that may be considered *profane*, *vulgar*, or *offensive*.

Visit the following website to join the competition.

`https://www.kaggle.com/c/nlp-getting-started`

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 Datasets from Project 1

4.1 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.

4.2 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.

5 Self Proposals

5.1 Workers Supervision for Construction Safety

`https://yao-lab.github.io/capstone/2021.fall/project1/Project_proposal_final_TrungKien.pdf`

5.2 Limitations of Translation: How much translation affect the analysis of Chinese text in different models? (Natural language processing with Chinese character)

5.2.1 The dataset description

The dataset we will be using would be one of the dataset from weibo_senti_100k, `https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/weibo_senti_100k/intro.ipynb`.

It is a dataset containing 100k reviews in Chinese 50k positive and 50k negative, which all of them comes from Sina Weibo, and contains either positive or negative review (also known as label). From the dataset, the label would be treated as response variable, and the review would be used as predictor variable.

5.2.2 Measure of accuracy

We would first separate the dataset into 2 parts, with 20% of the data being the testing set and 80% of the data being the training set. Moreover, as the dataset has equal amount of positive and negative review, the amount of positive and negative review in both testing and training set would be equal. During the training phase, only the data in training set would be used. The scoring would depend on the testing set. As the response is balanced, the formula for scoring would be the accuracy of the result, i.e. $(\#TP + \#TN)/\#Samples$.

5.2.3 Method

We would be using multiple method in this project. First, there are some pre-existed models in the following public Github: `https://github.com/ymcui/Chinese-BERT-wwm`.

We would be using some of the pre-trained models as a way of fine-tuning. Moreover, we would also be using some other method to train and test the model. Some potential method includes CNN, RNN, Bayes Techniques. Moreover, we would be analyzing how much better it is if the text is analyzed directly in Chinese when compare to text is analyzed after it is translated to English. We would be using the same method as above, but before feeding the data to the model, we would translate the data into English first. The scoring will also be using the method stated in measure of accuracy. We hypothesize that there exist some performance lost on analysis if the text is translated.

5.3 G-Research Crypto Forecasting

5.3.1 Project Overview

`https://www.kaggle.com/c/g-research-crypto-forecasting/overview`

5.3.2 Problem Identification

The cryptocurrency market has been skyrocketing recently, and it is estimated that over USD40 billion worth of cryptocurrencies are traded every single day. Cryptocurrencies have become one of the most popular and trending assets for speculation and investment, however, it has been proven to be wildly volatile, where a person can make a fortune and become a millionaire in one day, and lose all his assets the day after. While a few people have made a great fortune through the fast-fluctuating prices, others have been experiencing losses. Hence, we would like to try whether we can predict some of these price movements in advance and forecast short term returns in 14 popular cryptocurrencies through machine learning techniques.

5.3.3 Background Knowledge

The simultaneous activity of thousands of traders ensures that most signals will be transitory, persistent alpha will be exceptionally difficult to find, and the danger of overfitting will be considerable. In addition, since 2018, interest in the cryptomarket has exploded, so the volatility and correlation structure in our data are likely to be highly non-stationary. The successful contestant will pay careful attention to these considerations, and in the process gain valuable insight into the art and science of financial forecasting.

5.3.4 Model and Methods

Given a time-series based dataset of cryptocurrency prices, we can use the hierarchical time series model, which is to train a model for all time, models per weekday, model per day, etc. and ensemble them together. Multiple popular models will also be applied and we are going to conduct research about the effectiveness of these models. For example, LSTM is popular among quantitative finance fields, as it can capture the relationship between previous cryptocurrency prices. Since cryptocurrency price is a time-dependent variable, multiple time-series prediction models may also be adopted, like ARIMA, GARCH, and Dynamic linear model. We aim to create innovative methods, bringing more insights to the quantitative finance field.

5.3.5 Data Source

Kaggle:

`https://www.kaggle.com/c/g-research-crypto-forecasting/data`

They have amassed a dataset of millions of rows of high-frequency market data dating back to 2018 which we can use to build our model. Once the submission deadline has passed, our final score will be calculated over the following 3 months using live crypto data as it is collected.

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 Dec. 1, 2021*. 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 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 presentation (1-5): Is the presentation clear and well organized? Are the language flow fluent and persuasive? Are the slides clear and well elaborated? Please make suggestions to improve the presentation.
- Evaluation on creativity (1-5): Does the work propose any genuinely new ideas? Is this a work that you are eager to read and cite? Does it contain some state-of-the-art results? As a reviewer you should try to assess whether the ideas are truly new and creative. Novel combinations, adaptations or extensions of existing ideas are also valuable.
- 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 Dec. 3, 2021*. 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.