ECON498 Machine Learning I(Fall 2020) Midterm Exercise 1 Name: _____

Due: 6th November 2020 (11:59pm) UID: _____

---

1. (a) Collect a dataset of the top 500 cryptocurrencies on coinmarketcap.com, including the name, the price, trading volume, the logo, market capitalization, and other information that you find interesting.

   (b) Collect a dataset of the top 500 cryptocurrencies on a very similar website coingecko.com.

   (c) You need to include the dataset in the git repository, but you should not include the HTML files. (Please write the correct .gitignore file)

   (d) Write a README.md file and put it into the Git repo alongside your program files. This file should contain a step-by-step instruction of how can other people can use your program to download data from these two websites.

   (e) Your dataset should cover approximately the same period of time, for example, "48 hours, 15 minutes intervals". You can pick your time periods, but generally, you do not need an interval shorter than 15 minutes.

   (f) Compare the data you got from the two websites, highlight and explain the differences. You can present the differences by presenting summary statistics, plotting the data in charts. You can perform analysis (for example, the linear model/logistic model) and explain what factors are associated with a larger discrepancy between the data you obtain from the website. However, do not just paste the charts, tables, and numbers, you need to describe what you want to show using the charts and numbers. There is no requirement in the length of your answers but if anyone really really wants a rule of thumb, write two pages.

   (g) Bonus: Download the historical data (say, one year) from both websites and make the same comparison.

   (h) You must hand in your homework via Github. Please send your GitHub id to tomlam@g.clemson.edu. Create a repository named "ECON498_midterm" and share with me (my Github id is "erinata"). Your repository should contain everything you want to hand in, including the code, the data downloaded, the readme file, and your answer in part f.

   (i) I assume that most of you will use Python to complete these tasks. However, you can also use other programming languages or software.

   (j) The question describes the ideal situation but there is a lot of flexibility in this exercise. For example, if you have difficulty downloading the top 500 cryptocurrencies, you can opt to download just 100 of them. If you cannot get a continuous 48 hours of data, you can try to get 8 hours, or several fragments of data which can be combined into hours of data. If you cannot get the information of a cryptocurrency from the specific page of that cryptocurrency, you can just use the information from the front page. You can get (a lot of) partial credits if you can perform some analysis using an imperfect dataset.