Assignment 1, Applied ML

John MacLaren Walsh

Fall, AY 2020-2021

- **The dataset.** Your dataset appml-assignment1-dataset.pkl contains[1]:

  - historical exchange rates between various currencies and the US dollar on a major trading platform (precisely midpoints between the bid and ask),

  - during the trading day trading prices for exchange traded funds that attempt to track stock markets in the associated countries

  - an index that tracks a collection of US treasuries, and its volatility, and the S&P 500 index

  For each such index/ETF/exchange rate, each of the following are provided hourly

  - The "open" price – which is the price the item began the hour at
  - The "high" price – which is the highest price the item reached during the hour
  - The "low" price – which is the lowest price the item reached during the hour
  - The "close" price – which is the price the item ended the hour at

  NOTE: When an item was not trading during an hour the information from the most recent hour it traded in included. These items are included in the pandas dataframe X. In addition to these the pandas dataframe y contains the high for the (midpoint between bid and ask) for the Canadian dollar exchange rate (i.e. how much 1 USD is in Canadian dollars) from the subsequent hour.[2]

- **Your task.** You are to train a model to predict the next hour's high of the Canadian dollar exchange rate y from only the previous hour's data X.

- **How you will be graded**: Some of the data I have not shared with you. You can submit up to two models, both of which must be accompanied with the data transformation pipeline you have fit to your data. I will first transform the heldout data using the .transform method of your supplied transformation pipeline, then predict the associated CAD-high for the next hour for it using the .predict method of your supplied model. I will measure a mean squared error, repeating this process for both of your two supplied (model, transformation pipeline) pairs, keeping the lower of the two mean squared errors. I will map this MSE to a grade for the assignment by comparing it to the MSEs obtained by your peers.

- **How to submit your assignment**: Your submission should be gzipped or zipped archive titled **abc123-lab1.tgz** or **abc123-lab1.zip** (with abc123 your drexel username). The archive must contain at least the following 6 files with the following names

  - code1.py – a simple text file containing the python code you used to create your first transformation pipeline model (cut and paste from your jupyter notebook as necessary)
  - pipeline1.pkl and model1.pkl – a joblib/pickle for your first transformation pipeline and first model respectively.
  - code2.py – a simple text file containing the python code you used to create your first transformation pipeline model (cut and paste from your jupyter notebook as necessary)
  - pipeline2.pkl and model2.pkl – a joblib/pickle for your second transformation pipeline and second model respectively.

  You must submit the zipped or gzipped file within BBlearn before the assignment's due date. NOTE: I must be capable of regenerating your model using the code you supplied.

---

[1]Use pandas' read_pickle method to read the pickle then extract the variables X and y from the python dictionary it returns.

[2]There are a ton of obvious things that this leaves out – notably (1) sovereign debt yields in the foreign countries, (2) some measure of corporate debt yield, (3) prices of related swaps and derivatives. This is not a proper economics assignment.