Class Challenge

CMU 11-755/18-797: Machine Learning for Signal Processing (Fall 2022)

OUT: September 22nd, 2022

PROPOSAL DUE: **September 30th, 2022, 11:59 PM**
MIDTERM DUE: **November 13th, 2022, 11:59 PM**
FINAL DUE: **December 13th, 2022, 11:59 PM**

Our project this year spans over three challenges in different fields of real life, all related to **time series prediction**. Depending on the data type and distribution, you may want to analyze the data and design your model in different ways. But a mutual goal is to get a better prediction, and show your understanding in machine learning techniques for signal processing. You can use the data resources provided and experiment with any machine learning techniques, but please keep in mind that **neural network methods are not allowed**.

In addition to submitting typeset reports at various stages as well as the programming scripts, a **video presentation** around 10 minutes encompassing the entirety of your work is due on **December 11th, 2022, 11:59 PM**, a **peer review** across the entire span of your work is scheduled from **December 11th, 2022, 11:59 PM** to **December 13th, 2022, 11:59 PM** and is held on Piazza.

1 Stock Challenge

For the first challenge, we're asking you to invest in the stock market. Specifically, you have **1000 virtual dollars** on the first day of investment. You may buy or sell any stocks included in **NASDAQ 100 Index** to maximize your property. Also, to simplify the problem, trades are made only at the last second of the trading hours, which means we only study the **closing price** of every trading day. You can buy and sell as many stocks as you want during the investment. The goal of this challenge is to raise the yield curve as much as possible, and to have the maximum value of total property after the last trading day.

Also, **a maximum leverage of 1000 dollars** is allowed - which means you can borrow money (no more than 1000 dollars), or borrow stocks (total worth no more than 1000 dollars when you place a short order) from your virtual broker. But please be careful if you want to add leverage - with a higher variance you could even lose more!

Your model should be able to make decisions based on its prediction of the stock prices. To achieve this, here's a quick Python script provided for you to acquire the past stocks data. Note that you can use data throughout the stocks' history. Also, since the market is constantly evolving, you may want to occasionally update your data to improve the model.

```
import os, contextlib
import pandas as pd
import yfinance as yf

# create a folder for data storage
if not os.path.exists("stocks_data"):
    os.mkdir("stocks_data")

# read the NASDAQ 100 Index ticker symbols list from Wiki
url = 'https://en.wikipedia.org/wiki/NASDAQ-100#Components'
html = pd.read_html(url, header=0)
series = html[4]["Ticker"]
symbols = series.to_list()

# download prices history
with open(os.devnull, 'w') as devnull:
```

```
with contextlib.redirect_stdout(devnull):
    for i, symbol in enumerate(symbols):
        data = yf.download(symbol, period='max')["Close"]
        data.to_csv('stocks_data/{}.csv'.format(symbol))
```

We're hoping that through this challenge you can have fun making use of knowledge taught in class on real-life problems (and hopefully, make some real money).

Midterm Report

You're supposed to submit your code on Canvas **before EOD, November 6th (Sunday)**. We're going to make use of the following week to evaluate our halfway model. Use your model and code at this stage to schedule your investments in the market **from November 7th to November 11th**. And in the weekend, calculate your actual income to finish the midterm report. Please clearly report your method, your prediction and its comparison to the reality.

Final Report

You're supposed to submit your code on Canvas **before EOD, December 4th (Sunday)**. We're going to make use of the following week to evaluate our final model. Use your model and code at this stage to schedule your investments in the market **from December 5th to December 9th**. And in the weekend, calculate your actual income to finish the final report. Please clearly report your method, your prediction and its comparison to the reality.

2 Election Challenge

The second challenge beseeches you to foretell the re-assorted numbers of seats occupied by the Democratic Party and the GOP in Senate and House, as well as the exact ballot counts, after the 2022 United States elections. This cycle of Senators and House Representatives will be elected on **November 8th, 2022**. During election night, ballots will be only counted up to a point that a win is a foregone conclusion. Vote counting with the rest of ballots will proceed into later days to consummate.

To facilitate your forecasting so that you will not be drawn into the Washington swamp while investigating the election trend, you are expected to directly draw upon the Model Outputs and Polls, which are two folders of CSV files consisting of predictions on seats and polls on ballot counts, prepared by Nate Silver and his team and constantly updated on their website FiveThirtyEight2022. All CSV files, especially those pertaining to Model Outputs, update almost hourly. You are expected to **pull as much data as you can with evolving timestamps** so your model based on this enriched time series can at the end speak its forecast into reality.

Here is a brief Python script for you to pull an individual CSV file into your local machine.

```
import pandas as pd

# Pull a CSV file from its API into Pandas DataFrame
CSV_URL = 'https://projects.fivethirtyeight.com/'\
          '2022-general-election-forecast-data/'\
          'senate_national_toplines_2022.csv'
df = pd.read_csv(CSV_URL)

# Save Pandas DataFrame back to a CSV file in local machine
df.to_csv('senate_national_toplines_2022.csv', index=None)
```

The README.md files in two preceding folders encompass a complete set of urls to request various sorts of pertinent CSV files for your reference. You may want to build up your time series by looking into the 'senate_national_toplines_2022.csv', 'house_national_toplines_2022.csv' and 'generic_ballot_polls.csv' files to begin with.

Midterm Report

You're supposed to submit your code for **seat predictions** on Canvas **before EOD, November 6th (Sunday)**. We're going to make use of the following week to evaluate our halfway model. And after the election night, incorporate the actual election results to finish your midterm report. Please clearly report your method, your predictions on Senate and House seats occupied by the Democratic & the Republican Parties, and their comparison with the reality.

Final Report

You're supposed to submit your code for **vote predictions** on Canvas **before EOD, December 4th (Sunday)**. We're going to make use of the following week to evaluate our final model. Throughout the week **from December 5th to December 9th**, incorporate the actual ballot counts to finish the final report. Please clearly report your method, your predictions on vote counts and its comparison with the reality.

3 Fifa World Cup Challenge

In this challenge, you are expected to predict the results of the quarterfinals of the Fifa world cup, as well as the exact goals scored by each team in the quarterfinals. The quarterfinal will be played on **December 9th, 2022** and **December 10th, 2022**. For accurate predictions, you will need to update your model based on the results and the statistics obtained from the tournament until **December 6th, 2022**.

To facilitate your predictions, you could use FIFA World Cup dataset for previous world cup results and International Matches dataset to view the statistics of every international match played from 1872 to 2022.

To make your predictions more accurate, you could create your own custom datasets from FIFA Index. Fifa index is a website where you could get individual player ratings and the national team ratings along with their defense, midfield and attack, and overall ratings. You could use this data along with the historical data to simulate the results of the quarterfinals. You are also free to search for other datasets and use it for your predictions.

Make sure you follow the world cup to learn more about injuries, red cards, yellow cards, suspensions during the tournament. For Example, you might need to bias your expectations against Portugal if Cristiano Ronaldo gets injured! Also, update your datasets with results obtained from the group stages and round-of-16 scheduled to end on **December 6th, 2022** for accurate final predictions.

Midterm Report

Use your model to predict the results of quarter-finals of **Fifa World Cup 2018**. You're supposed to submit your code for **score predictions** on Canvas **before EOD, November 6th (Sunday)**. Please clearly report your method, models, and assumptions used to make your predictions. Your model should generalize well to make the final predictions on December 6th.

Final Report

For the final report, you are supposed to predict the results of quarter-finals of **Fifa World Cup 2022**. You're supposed to submit your code for **score predictions** on Canvas **before EOD, December 6th (Tuesday)**. We're going to make use of the following week to evaluate our final model. Incorporate the actual quarter-finals results on **December 9th and December 10th** and your analysis to finish the final report. Please clearly report your method, your score predictions and its comparison to reality.