Harvard Extension Data Science

Dynamic Modeling and Forecasting in Big Data

Instructor: William Yu

Final Project

- The goal of this final individual project is to let you apply time series and/or dynamic models onto the real-world dataset and make predictions.
- You need to decide the topic of the final project on your own based on your personal and career interests or your work and experience.
- You can choose any kinds of datasets. There are only two requirements:
  - (1) Total dataset's size sums up at least **2 MB**.
  - (2) There should be time variables in the data, such as year, month, day, or intra-day frequency.
  - If you have difficulty in finding datasets, you can try these two great sources. (Other links are in the next page):
    - https://www.kaggle.com/datasets
    - https://github.com/awesomedata/awesome-public-datasets
- You *cannot* submit the same project that you have worked or submitted in other classes.

Three Components of the Final Project

(1) Written Report

- In the project report, four elements should be included:
  - Provide a purpose/motivation of your data analytics project.
  - Present a clear and high-quality output of visualizations.
  - Explain how and why you choose the model(s). I expect some test-set/out-of-sample validation.
  - Tell a good story or lay out conclusions/takeaways from your findings/results.
- In the report, it should have an executive summary (two to three pages). You can put everything else (charts, outputs, etc.) in the remaining part of the report (total 5 – 20 pages).

(2) PowerPoint Slides Presentation

- Prepare a slide deck for at most 10 pages and prepare a 10-minute presentation of your final project. There are two options of presentation.
  - You can present on the final instruction day: December 13. Sign up because we only have 10-12 slots.
  - You can record your slides presentation via zoom and submit it by December 14.

(3) R Script (or R Markdown, Python Jupyter Notebook)

- Project submission to Canvas by midnight **December 16** including the data.

- I will post and share your submission files to the whole class in Canvas so you can learn from each other. If your data or report is proprietary and cannot be shared, let me know. Then I will not share it to the class.
- Grading will be based on the aforementioned four elements in the report and presentation. Furthermore, a project will get a higher grade if it is (1) with a very interesting or persuasive story-telling, (2) with more solid data analytics work, or (3) with a more novel idea or methodology.

Useful Links and Public Datasets

- https://ourworldindata.org/
- https://www.kaggle.com/competitions (Kaggle competitions)
- https://www.dataquest.io/blog/free-datasets-for-projects/
- http://archive.ics.uci.edu/ml/index.php (UCI Machine Learning Repository)
- https://www.kdnuggets.com/ (KD Nuggets)
- https://stackoverflow.com/tags (Stack Overflow)
- https://stats.stackexchange.com/tags (Cross Validated)
- Federal Reserve Economic Data (FRED, https://research.stlouisfed.org/fred2/)
- Bureau of Economic Analysis (BEA, http://www.bea.gov/)
- Bureau of Labor Statistics (BLS, http://www.bls.gov/)
- Google Public Data Explorer with dynamic graphing.
- Yahoo Finance (http://finance.yahoo.com/), providing downloadable historical finance data
- Fed Regional Banks' data, indicators, and research: New York, San Francisco, Chicago, Kansas City, Dallas, Philadelphia, Cleveland, Atlanta, Boston, Minneapolis, Richmond, DC Board
- Gapminder, a fact-based worldview, with a very cool chart!
- Visualizing Economics, providing a lot of vivid charts and maps in the world!
- World Bank's World Development Indicators and Worldwide Governance Indicators
- International Monetary Fund (IMF)'s World Economic Outlook Databases
- National Bureau of Economic Research (NBER)'s data, collecting a bunch of great research data!
- International regional and country's statistical sites by BLS and by EDIRC
- Index of Economic Freedom and CIA the World Factbook
- Robert Shiller's stock and housing markets data, providing useful financial market data
- S&P/Case-Shiller home price indices
- Federal Housing Finance Agency (FHFA)'s House Price Index
- Zillow Real Estate Research data (http://www.zillow.com/research/data/)
- IMF Global Housing Watch and Dallas Fed's International House Price Database
- BIS Property Price Statistics
- Angus Maddison data, two thousand year GDP per capita data around the world.
- Citylab's Maps (http://www.citylab.com/posts/maps/)
- Google Books Ngram Viewer (word frequency in books from 1800 to date)
- The Mnist database (http://yann.lecun.com/exdb/mnist/)
- Image Net (http://image-net.org/)
- Word Net (https://wordnet.princeton.edu/)
- Visual Genome (http://visualgenome.org/)