Man Versus Machine Learning Revisited

This repository contains the replication data and code for the paper "Man versus Machine Learning Revisited" by Yingguang (Conson) Zhang, Yandi Zhu, and Juhani T. Linnainmaa. For more details, please refer to our manuscript on SSRN. For any inquiries, please contact Yandi Zhu at yandi.zhu@stu.pku.edu.cn.

Data Availability

The code requires access to the WRDS database (CRSP, Compustat, and IBES). Follow the instructions in `code/00_DataDownload.ipynb` to download the required data. We provide only publicly available data such as macroeconomic variables from Federal Reserve Bank of Philadelphia.

Overview of the Replication Package

Code Folder (`code`)

The `code` folder contains Jupyter notebooks for data processing, model training, and analysis:

- **00_DataDownload.ipynb**: Downloads and organizes raw data.
- **01_Preprocess.ipynb**: Cleans, filters, and formats data for analysis.
- **02_EarningsForecasts.ipynb**: Implements earnings forecasts using a replication of the BHL random forest model and linear forecasts by So (2013) and Hughes et al. (2008).
- **03_Main.ipynb**: The main analytical notebook, generating key tables and figures.
- **04_RF_variants.ipynb**: Explores alternative Random Forest model specifications.
- **05_ML_variants.ipynb**: Explores alternative machine learning models.
- **06_DataShare.ipynb**: Generates the shared look-ahead-bias free earnings forecast dataset.

The `functions` subfolder contains custom functions (e.g., single sorts, Fama-Macbeth regressions) utilized across different notebooks.

Data Folder (`data`)

Organized into subfolders:

- `BHL`:
  - **Conditional_Bais.csv**, the conditional bias data from Binsbergen et al. (2023).
- `Macro`: Macroeconomic data including INDPROD, RCON, RGDP, and UNEMP
- `WRDS`: Data from the WRDS (Wharton Research Data Services) database.
- `Other`:
  - **ff5_factors_m.csv**: monthly factor returns (FFC6, HMXZ, SY, DHS)
  - **Siccodes59.csv**: FF49 industry classifications

    - **signed_predictors_dl_wide.csv**: anomaly characteristics from Open Source Asset Pricing. Please download from https://www.openassetpricing.com/, v1.3.
- `Results`: Analytical results, including model outputs and processed datasets.

Dependency

The code was tested under the environment:

Package	Version
Pandas	2.1.4
Sklearn	1.2.2
lightgbm	4.3.0
LinearModels	5.4
statsmodels	0.14.0
scipy	1.11.4