Project 1: Predicting the S&P/Case-Shiller U.S. National Home Price Index

Summary & Conclusions

In this project, we

- Use public Fredapi to collect the S&P/Case-Shiller U.S. National Home Price Index and other indexes that might be relevant to this HPIX such as GDP, M2, Personal saving, Salary level...
- Do data cleaning & feature engineering and transformation for prediction and to satisfy the model hypothesis. Features include
    - momentum of indicators
    - average line of indicators
- Design two train-test split method and build 3 models for the regression task:
    - Baseline: ARIMA
    - Improved: LASSO / XGBoost
    - The reasoning of choosing the models is ARIMA is widely used in time-series prediction but lacks feature, and the other two models can measure how much additional prediction power can be brought by the other features.
- Train / predict over the schemes and implement several evaluation metric, including general ones like RMSE, R^2 and correlation. And other metrics as stratification monotonicity and parameter stability.
- After comparison between the model performance, we tell that

	RMSE	R2	Pearson	Spearman
Baseline: ARIMA	0.0064	-0.225	0.204	0.141
LASSO	0.00561	0.0585	0.718	0.503
XGBoost	0.00505	0.238	0.612	0.527

    - Traditional time-series ARIMA can hardly capture the movement of the HPIX, where the predicted variance decay quickly over time and converge to the mean.
    - LASSO and XGBoost gives similar performance under general split method and metrics, reaching about 70% pearson and 50% spearman out-of-sample data.

	RMSE	R2	Pearson	Spearman
LASSO:rolling	0.00206	0.886	0.96	0.961

    - Although there is over-estimated bias, the time-rolling scheme can improve the prediction power of LASSO. Possible reasons comes from the transition of in-sample & out-of-sample data distribution. **Update our model in time is likely to help to capture the change in market style.**

Problems & Discussions

- The pandamic in 2020 and crisis in 2008 can hardly be captured with our features.
- Such crisis can be seen as outliers and might be mitigated by least absolute error regression (L_1 error).
- Due to the limit of dataset, we can hardly build some equivalent testing set to compare the performance between the 2 split methods.
- Also due to the limit of dataset, the complexity of model is constrained, NNs are not recommended.