I've recently been going through the peer review process for my Era Splitting paper, which I worked on last year with and released. I got some very constructive reviews which prompted me to revisit the theoretical groundwork and ideals of the era splitting algorithm. Through the process, I found some pretty simple little oversights that didn't seem so important originally but after going back, I realized were very important. After fixing the code in a few spots the algorithm has really come to life. See the fixes in this commit: era splitting revision · jefferythewind/scikit-learn-erasplit@885e8e0 · GitHub

For a review, this research proposes two novel splitting criteria for growing decision trees, which have been implemented into a gradient boosting decision tree (GBDT) model comparable to LightGBM and XGBoost: Sklearn's HistGradientBoostingRegressor. The research compares the two new criteria against the original splitting criterion which gained popularity with XGBoost and is implemented in LGBM. The two new criteria are:

- Era Splitting
- Directional Era Splitting

The main fix was to not allow any splits which are degenerate

in any eras. A degenerate split is one which fails to improve the impurity measure in any era of the data. After re-examining the theory, I realized the previous code was technically allowing splits which could have been degenerate. There was another setting which allowed for potential splits with negative information gain, which was fixed.

To read about the background and details, refer to thenewly revised paper:[2309.14496] Era Splitting -- Invariant Learning for Decision Trees

After fixing the code I re-ran all the experiments. The new splitting criteria have performed much better than before. They have succeeded where they failed before, for example: both criteria work on the synthetic memorization task when only directional era splitting worked before.

[

Screen Shot 2024-03-15 at 10.30.09 AM

1678×830 149 KB

](https://forum.numer.ai/uploads/default/original/2X/7/7797b101e2157acb8f62fe9aee1092fc679872bd.jpeg)

On experiments with Numerai data, era splitting and directional era splitting can now beat benchmark models out-of-the-box in a variety of ways. While grid searches over the available parameter space are expensive, I've been finding them quite fruitful. In this research I've used a random grid search routine which picks random configurations and tests them across all three models. Save considerable time by using ShatteredX's fantastic compressed feature set. Training is first 4/5 of the data (by era) and test is last 1/5. This is the last fold of a walk-forward cross validation routine. This most recent test tries 25 configurations.. Directional era splitting attains the highest test Sharpe of 1.27 compared to 1.12 of the original and 1.07 for era splitting. All these Sharpes out-perform the standard LGBM model (2,000 trees, 5 max depth, 0.1 colsample bytree, 0.01 learning rate, and 32 num leaves) which scored 1.03 on this fold of data. And all these experimental models have a maximum number of trees of under 1,000. Through all the experiments the highest corr model was also a directional era splitting model, although it was a slight edge.

Below are the cumulative correlations (w/ Cyrus) over the test fold of data, approx. 200 eras for the best Sharpe models in the respective categories. The directional era splitting model completely avoids any drawdown during last year's nightmare scenario. While the overall corr is lower, this really shows the hallmarks of an invariant

predictor, one that works consistently the same way over time.

I applied the algorithms to a new data set:the Camelyon17 data set, which is a domain-generalization problem for detecting breast cancer is histopathological images from different hospitals. Era splitting and particularly directional era splitting exceeded the baseline model by a wide margin. While results with GBDTs are not competing with deep NN vision systems for SOTA claim, the improvement over the baseline of the new splitting criteria shows great potential.

This revamped version of era splitting could be very valuable to the Numerai community. As we know, it is not one individual model that will be the best, but an ensemble of a large number of models. The addition of era splitting models into the ensemble will at least add more variety to the ensemble. Another nice experiment to see would be to compare ensembles built in the same way but for each of the of the splitting criteria separately, and then combined.

As far as making this a viable option for the community, I intend to repackage the era splitting algorithm into its own compact library, under its own name, and distribute it through Pypi. Maybe then the nice people at Numerai could include it in the model uploads system.

I just added a quick start routine with requirements.txt

file to the main readme file on the github repo to get you started quickly running the notebooks.

I also just added a study on ensembling, showing that ensembling with the new era splitting models leads to an ensemble

that greatly outperforms the ensemble based on just the original models.

Happy modeling.