|  |  |  |
| --- | --- | --- |
| **Fall 2024** | **Report #4 – 09/24/2024** | **Dingyi Nie** |

**——————————————————————————————————————**

**I. Task Achieved Last Week**

**——————————————————————————————————————**

* First trial on differenced time series analysis with GL.

**——————————————————————————————————————**

**II. Feedback and Interaction**

**——————————————————————————————————————**

* **Prof. Kuo’s Feedback**
  + Last week Prof. Kuo suggested we change direction and explore a non-imputing approach.

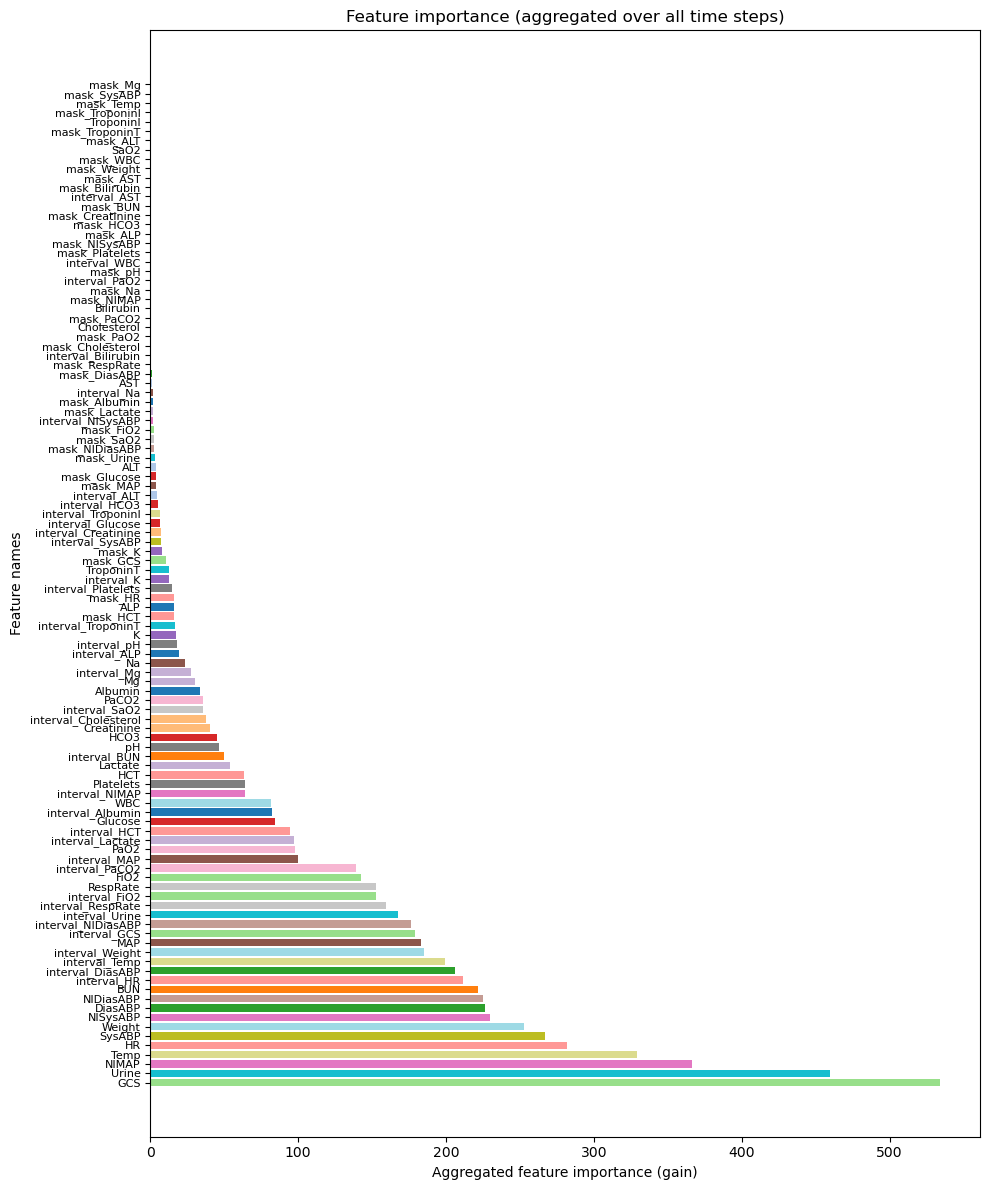
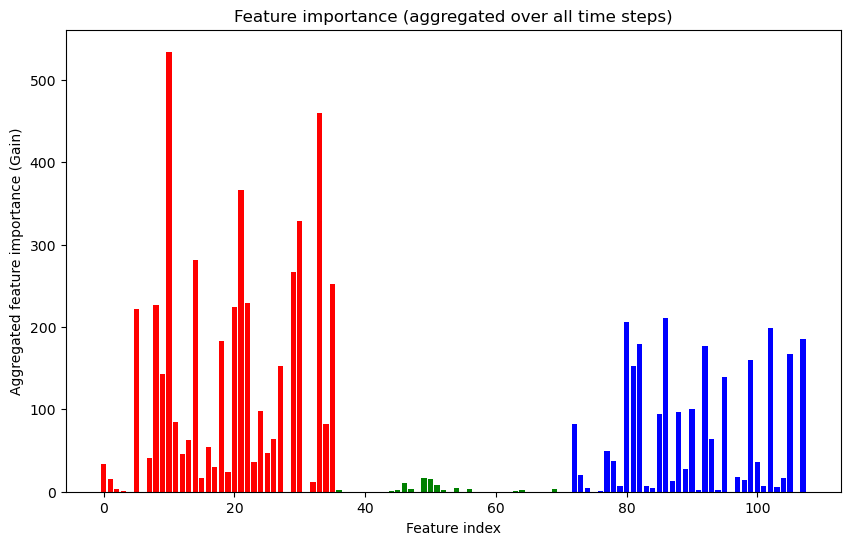
**——————————————————————————————————————**

**III. Report**

**——————————————————————————————————————**

**P12 Feature Importance Study**

To study the importance factors of the 36 features (true features, each with a mask array **m** and an interval array **d**, resulting in 108 model input features in total), I fit an XGBoost classifier on x-m-d concatenated input and examine their aggregated gain (difference in overall loss after each node / split, which is associated with a specific feature, is introduced to the decision tree ensemble) over all time steps. The results are showed in the following two figures.

****

The first figure clearly shows that the observation variables (red) are the most important, while time intervals between observations (blue) being the second and masks (green) barely having any contribution to the prediction. Some of the most important true features (with missing rates) are: GCS (79.26%), Urine (55.42%), NIMAP (67.36%), Temp (70.73) and HR (23.36%). Note that these are among the lowest missing rates in the P12 dataset.

**Differenced Time Series Analysis with GL**

Given previous observation, I filter out 10 true features: *'DiasABP', 'GCS', 'HR', 'NIDiasABP', 'NIMAP', 'NISysABP', 'SysABP', 'Temp', 'Urine', 'Weight'*, which are the most important and also have a missing rate lower than 80%. Further analysis will be based on these features only.

On these features alone, an XGBoost classifier can achieve 81.12% AUROC on validation set with x-m-d input and 81.05% with x-d input.

For each feature, I try to fit a separate saab transform module to their differenced time series data. Since on average these features have only 10~20 observation occurrence in each time series episode, applying GL on the data is very difficult and the approach is limited. I first difference the time series and getting a variable length vector for each feature, each episode. Then, I fit a channel-wise saab transform module with a kernel length of 5 and stride of 1 to all vectors of the same feature, that have a length greater than or equal to 5. Since different feature’s observations occur independently (time stamps are different), original saab (which also captures correlations across channels) does not make sense so the channel-wise saab is applied here.

After fitting the saab, I transform these temporal differencing vectors into a spectral representation. Here is where things get tricky with our current approach. In the same episode, different feature’s temporal and spectral representations all can have different length, it is hard to define downstream operations before we unify the shape. This may need further discussion and experiments but what I have done is I do avg-pooling on temporal representation and spectral representation respectively. This will essentially transform each differencing time series into a (C, 1+K) array where C is the number of features (C=10). 2+K is simply one number for average difference, one number for average observed value, and another K numbers for average response to K saab kernels (K=5).

To this point, we have obtained a very simple representation. Before the final XGBoost classifier, I have tried DFT with different parameters. But the best result I can get with this approach is a 74.10% AUROC on validation set, which is significantly lower than our baseline.

**——————————————————————————————————————**

**IV. Next Steps**

——————————————————————————————————————

* To be discussed and decided

**——————————————————————————————————————**

**V. Milestone**

——————————————————————————————————————

* First trial on differenced time series analysis with GL.
* Codes are available here: <https://github.com/d9sus4/GL-TS>