xAutoML Course Project 1

Building Meta-Model for AutoML Federated Time-Series Forecasting Algorithms

Team 6

Institute of Computer Science, University of Tartu

Agenda

1. Motivation and Problem Statement
2. Dataset Description
3. Project Methodology
4. Results and Discussion

Motivation

Problem Statement

- Train the best meta-model that can recommend regression forecasting algorithms (Algorithm Selection) that can be used on a time-series federated dataset.

- The meta-model acts as the domain expert that recommends the search space to be optimized later.

Dataset Description

- 542 Time-Series Datasets with > 2500 instances
  - **512** Synthetic Datasets generated by varying several factors, such as seasonality components, sampling frequencies, SNR, missing values %
  - **30** Real Datasets From (Kaggle / Nasdaq Stock Market)
  - Splitted across randomly selected (5, 10, 15, 20) clients with time-series splits

Meta-Feature	Meta-Feature Type	Method of Aggregation
No. of Clients	Statistical	NA
Sampling Rate	Time-Series	NA
No. of Instances	Statistical	Sum, Avg, Min, Max, Stddev
Dataset/Target Missing Values %	Statistical	Avg, Min, Max, Stddev
No. of Stationary Features	Time-Series	Avg, Min, Max, Stddev
Target Stationarity	Time-series	Entropy of Target Stationarity across clients
No. of Stationary Features after 1st Order Diff	Time-Series	Avg, Min, Max, Stddev
No. of Stationary Features after 2nd Order Diff	Time-Series	Avg, Min, Max, Stddev
Significant Lags using pACF in target	Time-Series	Avg, Min, Max, Stddev
Insignificant lags between 1st and last significant ones	Time-Series	Avg, Min, Max, Stddev
No. of seasonality components in target	Time-Series	Avg, Min, Max, Stddev
Skewness of target feature	Statistical	Avg, Min, Max, Stddev
Kurtosis of target feature	Time-Series	Avg, Min, Max, Stddev
Fractal dimension analysis of target	Statistical	Avg
Periods of seasonality components in target	Time-Series	Min, Max
KL Divergence among clients' distribution of target feature	Statistical	Avg, Min, Max, Stddev

Dataset Description

- Constructed Dataset:
  - ***X:*** Aggregated meta-features
  - y: best performing forecasting regression algorithm among defined search space
- Splits:
  - Training: 400 instances (80%)
  - Testing: 100 instances (20%)

Algorithm	Hyperparameters	Values
Lasso Regressor	alpha selection	$(log(e^{-5}), log(10))$ {cyclic, random}
LinearSVR Regressor	C epsilon	$[1:10]$ $[0.01:0.1]$
ElasticNetCV Regressor	l1_ratio selection	$[0.3:10]$ {cyclic, random}
XGB Regressor	n_estimators max_depth learning_rate reg_lambda subsample	$[5:20]$ $[2:10]$ $[0.01:1]$ $[0.8:10]$ $[0.1:1]$
Huber Regressor	epsilon alpha	{1.0, 1.35, 1.5} $[log_{10}(e^{-3}) : log_{10}(e^{2})]$
Quantile Regressor	alpha quantile	$[log_{10}(e^{-3}) : log_{10}(e^{2})]$ $[0.1:1]$

class	num_clien	Sum of Inst	Max. Of Ins	Min. Of Ins	Stddev of In	Average Da	Min Datase	Max Datas	Stddev Dat	Average Ta	Min Target	Max Target	Stddev Tar	No. Of Fea	No. Of Nur	No. Of Cat	Sampling F	Average Sk	Minimum S	Maximum S	Stddev Ske	Average Ku	Mi
XGBRegres	10	13821	1383	1382	0.3	4.9924658	4.1214750	5.5716353	0.4489697	4.9924658	4.1214750	5.5716353	0.4489697	3	3	0	0.1666666	0.072567	1.1289151	1.3053050	0.2456136	1.3473565	0.3
XGBRegres	5	4031	807	806	0.4	4.7629458	3.9702233	6.203474	0.8750409	4.7629458	3.9702233	6.203474	0.8750409	1	1	0	1	0.2016527	4.5231737	0.902781	0	1.1917040	0.1
HUBERREG	5	17280	3456	3456	0	5.1041666	4.74537	5.2951388	0.1889447	5.1041666	4.74537	5.2951388	0.1889447	5	5	0	0.1666666	2.1255588	0.000482	4.400823	0.5678388	9.354913	1.0
LinearSVR	5	6911	1383	1382	0.4	4.9341131	4.0520984	5.5716353	0.5240857	4.9341131	4.0520984	5.5716353	0.5240857	4	4	0	0.5	0.3394770	0.006274	0.9517789	0.1593084	1.3472348	0.2
XGBRegres	15	24181	1613	1612	0.2494438	4.9667173	3.5359801	6.3895781	0.6939814	4.9667173	3.5359801	6.3895781	0.6939814	2	2	0	0.1666666	0.1935872	1.4382909	1.6268250	0.3241073	1.280996	0.3
HUBERREG	10	13821	1383	1382	0.3	4.9200023	4.1968162	5.7887120	0.5766058	4.9200023	4.1968162	5.7887120	0.5766058	3	3	0	0.1666666	0.069468	0.058096	0.094083	0.006972	1.2066428	0.8
LASSO	5	10366	2074	2073	0.3999999	4.8717193	4.4862518	5.4027978	0.337306	4.8717193	4.4862518	5.4027978	0.337306	5	4	1	0.5	1.2942629	0.004749	2.4642837	0.3500364	6.0663949	1.

Project Methodology

- Baselines with default hyper-parameters:
  - Random Forest
  - Decision Tree
  - KNN
  - SVM
  - Logistic Regression
  - Gradient Boosting

- AutoML Frameworks:
  - Hyper-Opt (100 iterations) with 3-fold CV
  - Search Space:
    - Random Forest:
      - Max_depth, n_estimators, min_samples_split
    - Decision Tree
      - Max_depth, min_samples_split
    - KNN
      - n_neighbors
    - SVM
      - Kernel, C
    - Logistic Regression
      - Solver, C
    - Gradient Boosting
      - Max_depth, n_estimators, min_samples_split, learning_rate

- Evaluation Metrics:
  - Log Loss
  - Micro F1-Score
  - Acc@3 (Accuracy of Top-3 Predicted Labels)

- Statistical Test
  - Wilcoxon Signed Rank Test

Project Methodology

- Baselines with default hyper-parameters:
  - Random Forest
  - Decision Tree
  - KNN
  - SVM
  - Logistic Regression
  - Gradient Boosting

Model	Acc@3	Micro F1	Log Loss
SVM	83%	44%	1.45
Logistic Regression	84%	54%	1.38
Random Forest	90%	61%	1.34
Decision Tree	75%	**66%**	15.9
Gradient Boosting	**91%**	60%	**1.25**
KNN	80%	44%	9.0

- Evaluation Metrics:
  - Log Loss
  - Micro F1-Score
  - Acc@3 (Accuracy of Top-3 Predicted Labels)

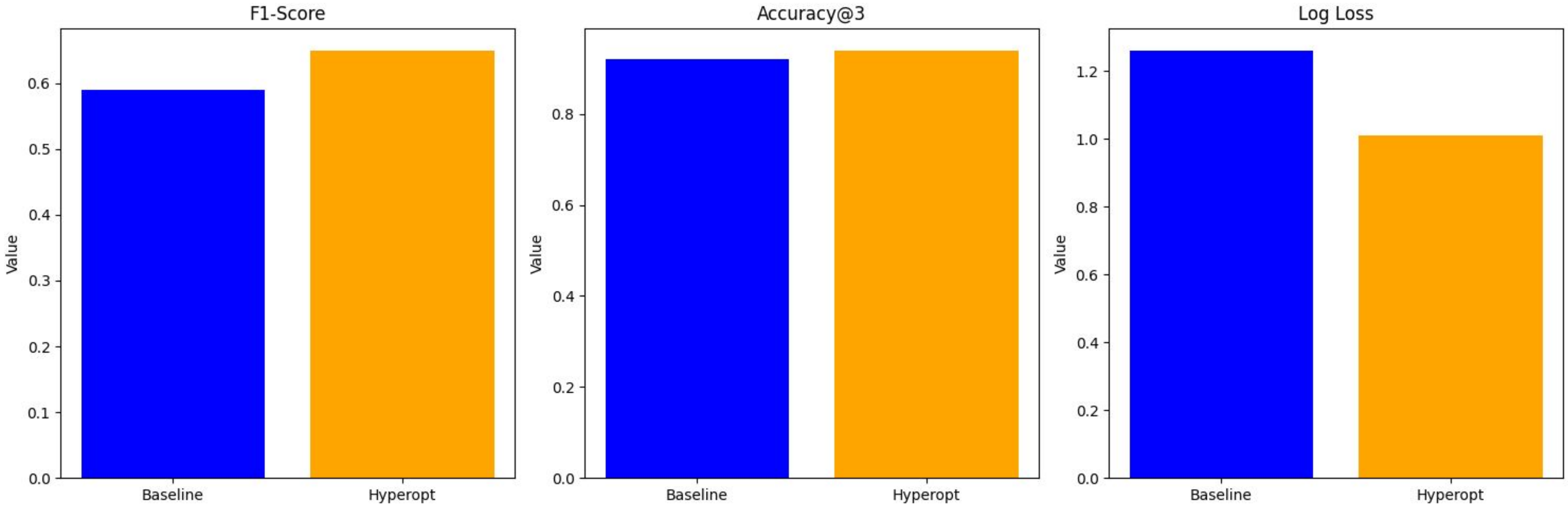
Project Methodology

Model	Acc@3	Micro-F1	Log Loss
Gradient Boosting	91%	60%	1.25
HyperOpt (objective F1) *Gradient Boosting* • *Lr: 0.04* • *Max_depth: 14* • *N_estimators: 270* • *Min_samples_split: 3*	91%	65%	1.02
HyperOpt (objective Acc3) *Random Forest* • *Max_depth: 7* • *N_estimators: 70* • *Min_samples_split: 5*	91%	62%	1.06
HyperOpt (objective LogLoss) *Random Forest* • *Max_depth: 10* • *N_estimators: 240* • *Min_samples_split: 4*	**94%**	**65%**	**1.01**

- AutoML Frameworks:
  - Hyper-Opt (1000 iterations)
  - Search Space:
    - Random Forest:
    - Decision Tree
    - KNN
    - SVM
    - Logistic Regression
    - Gradient Boosting
- Evaluation Metrics:
  - Log Loss
  - Micro F1-Score
  - Acc@3 (Accuracy of Top-3 Predicted Labels)
- Statistical Test
  - Wilcoxon Signed Rank Test

Project Methodology

- Wilcoxon Signed-Rank Test: P-value = 0.85.
- There is no significant difference between the optimized pipeline and baseline in terms of ***log loss***.



Thanks for your attention!

Team Members:

1. Ahmed Wael
2. Noel Bosch
3. Mohamed Maher

Find more about our work:

Source Code: