T-Fold Sequential Validation Technique for Out-Of-Distribution Generalization with Financial Time Series Data

Juan Francisco Muñoz-Elguezabal [1]   Juan Diego Sánchez-Torres [1]

[1]Mathematics & Physics Department - Western Institute of Technology and Higher Education (ITESO)

Hipothesis: There exists a set of conditions under which a cross-validation process can be defined and conducted in order to achieve Out-Of-Sample and Out-Of-Distribution Generalization when performing a Predictive Modeling Process using Financial Time Series Data.

Dataset: Continuous futures prices of the UsdMxn (U.S. Dollar Vs Mexican Peso), extracted from CME group MP Future Contract. Prices are Open, High, Low, Close in intervals of 8 Hours, **OHLC** data. GMT timezone-based and a total of 66,500 from 2010-01-03 18:00:00 to 2021-06-14 16:00:00.

Experiment: A classification problem is formulated as to predict the target variable, CO_{t+1}, which is defined as the sign$(Close_{t+1} - Open_{t+1})$. For the explanatory variables, the base definition is to use only those of endogenous nature, that is, to create them using only **OHLC** values.

A discrete representation

Let V_t be the value of a financial asset at any given time t, and S_t as a discrete representation of V_t if there is an observable transaction Ts_t. Similarly, if there is a set of discrete Ts_t observed during an interval of time T of $n = 1, 2, ..., n$ units of time, $\{S_T\}_{T=1}^{n}$, can be represented by $OHLC_T$: $\{Open_t, High_t, Low_t, Close_t\}$. The frequency of sampling T, can be arbitrarly defined.

OHLC data

Timestamp: The date and time for each interval.
Open: The first price of the interval.
High: The highest price during the interval.
Low: The lowest price during the interval.
Close: The last price of the interval

Intra-day micro-information:
volatility: HL_t , **price-change:** CO_t
uptrend: HO_t , **downtrend:** OL_t

Candlestick Visual Representation *(Figure 1)*

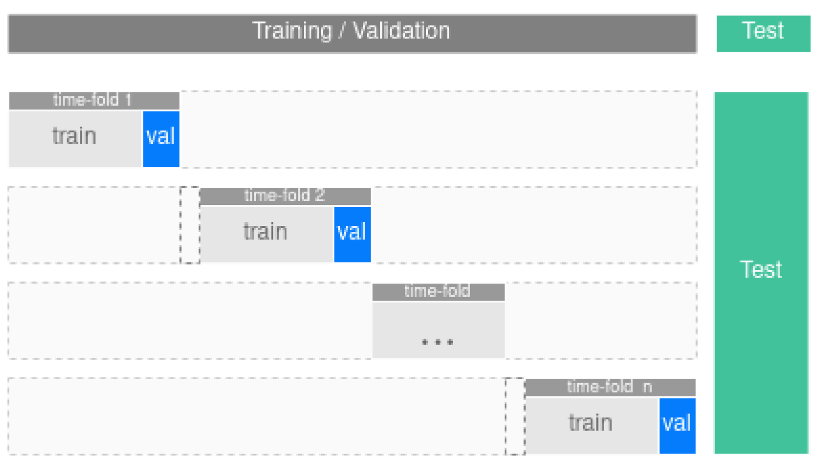
The base calculations are:

$$HL_t = High_t - Low_t$$
$$OL_t = Open_t - Low_t$$
$$CO_t = Close_t - Open_t$$
$$HO_t = High_t - Open_t$$

T-Fold-SV *(Steps)*

1.- Folds Formation
Depends on labeling, can be calendar based.
2.- Target and Feature Engineering
In-Fold exclusive or Global and then divide.
3.- Information matrix
To asses information sparsity among Folds.
4.- Predictive Modeling
Hyperparameter optimization Train-Val sets.
5.- Generalization Assesment
Out-Of-Sample and/or Out-Of-Distribution.

1: Folds Formation *(Figure 2)*



2: Target Variable (labeling)

A continuous variable prediction (regression problem), into a discrete variable prediction (classification problem), a time-based labeling can be stated as:

$$\hat{y}_t = sign\,\{CO_t\}$$

Interesting enough, this target variable never had an imbalace of classes more than 5.5%

2: Feature Engineering

with $\{OL\}_{t-k}$, $\{HO\}_{t-k}$, $\{HL\}_{t-k}$, $\{CO\}_{t-k}$ for values of $k = 1, 2, ...K$, with K as a proposed *memory* parameter. Then perform some fundamental operations: Simple Moving Average SMA_t, lag: LAG_t, Standard Deviation: SD_t and Cumulative Sumation: $CUMSUM_t$.

Then symbolic variables where generated using Genetic Programming.

3.1: Information Representation and Sparsity

A gamma distribution to fit the PDF of two set of variables, and the Kullback-Leibler Divergence to measure the similarity between the two:

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad \text{for} \quad x > 0 \quad \alpha, \beta > 0 \qquad (1)$$

$\Gamma(\alpha)$: The gamma function $\forall\ \alpha \in \mathbb{Z}^+$ and the $D_{KL}(P||Q)$: Kullback-Liebler Divergence, which for unknown continuous random variables, P, Q, or for p, q as empirically adjusted Probability Density Functions (PDF) is denoted by:

$$D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) log\left(\frac{p(x)}{q(x)}\right) dx \qquad (2)$$

3.2: Information matrix

An *Information Matrix* (IM) represents the similarity in information, for the target varible, among every Fold.

$$IM = \begin{bmatrix} D_{KL(1,1)} & D_{KL(1,2)} & \cdots & D_{KL(1,n)} \\ D_{KL(2,1)} & & & \vdots \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & D_{KL(n-1,n)} \\ D_{KL(n,1)} & \cdots & D_{KL(n,n-1)} & D_{KL(n,n)} \end{bmatrix}$$

D_{KL} is a non-conumtative operation, hence $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. That means the *Information Matrix* (IM), is not symmetric, but has 0's in its diagonal.

3.3: Matrix Characterization

If an *Information Threshold* is defined, and then applied to every value in IM, then the latter can be characterized according to a counting of following:

- **Sparse**:
All the elements of IM are sufficient disimilar among each other.
- **Weakly Sparse**:
There exists one or more very similar pairs of elements.
- **Non-sparse**:
ALll elements are highly similar to each other.

The ideal in theory is to have a *Sparse Information Matrix* to train any model, so to use non-repeated data.

4.1: Cost Function and Regularization

One common component of the predictive modeling process is binary-logloss cost function with *elasticnet* regularization:

$$J(w) = J(w) + C\frac{\lambda}{m}\sum_{j=1}^{n} \|w_j\|_1 + (1-C)\frac{\lambda}{2m}\sum_{j=1}^{n} \|w_j\|_2^2$$

Where $\Sigma_{j=1}^{n} \|w_j\|_1 = L_1$ and $\Sigma_{j=1}^{n} \|w_j\|_2^2 = L_2$ are also known as *Lasso* and *Ridge* respectively, with C as the coefficient to regulate the effect between the two.

4.2: Model's Params

Logistic Regression
- L1_L2_Ratio = 1.0 (Lasso) - Inverse of regularization (C): 1.5 - Parameter repetitions (Stability): Yes

ANN-MLP
- Hidden Layers: 2, 80 neurons each
- Activations: ReLU
- Dropout: 10% all layers

4.3: Results

Two models were defined, Logistic-Regression and Multi-layer Feedforward Perceptron.

Metric	ann-mlp	logistic
acc-train	0.9155	0.8311
acc-val	0.8245	0.7368
acc-weighted	0.4486	0.4061
acc-inv-weighted	0.4213	0.3778
auc-train	0.9924	0.9300
auc-val	0.8401	0.8017

Metric	ann-mlp	logistic
auc-weighted	0.4810	0.4521
auc-inv-weighted	0.4353	0.4137
logloss-train	0.2290	5.8333
logloss-val	6.0595	9.0892
logloss-weighted	0.6975	3.2422
logloss-inv-weighted	2.4467	4.2190

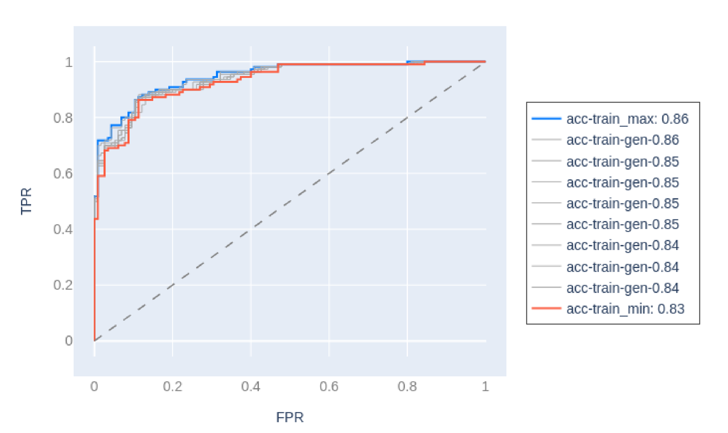
Train-Logistic

Validation-Logistic

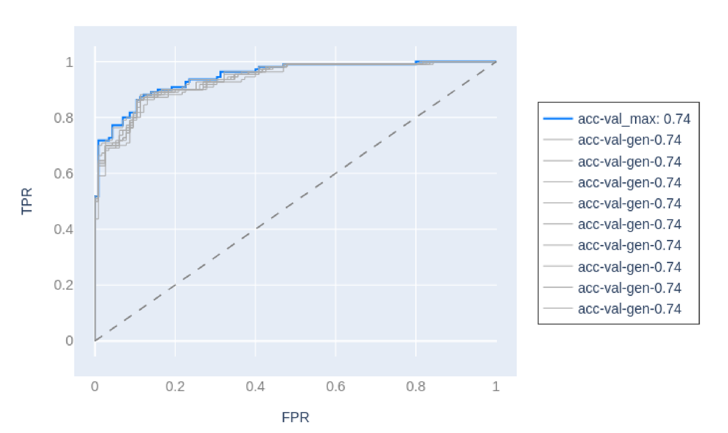
Train-NeuralNet

Validation-NeuralNet

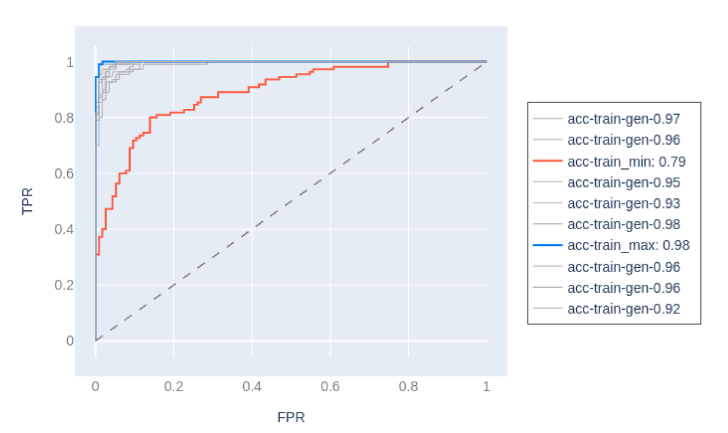
ROCs-Train-Logistic



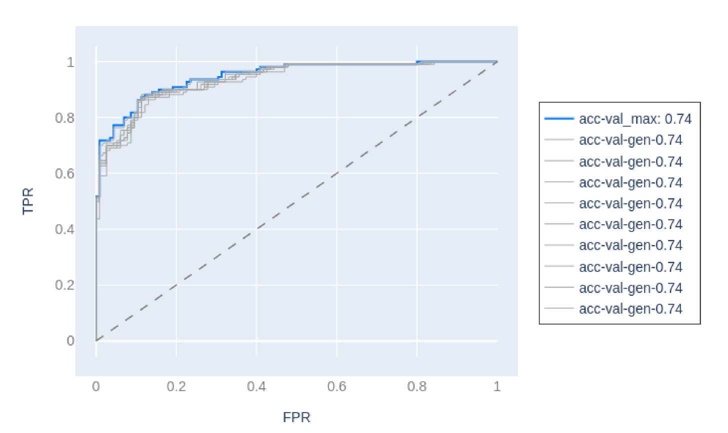
ROCs-Validation-Logistic



ROCs-Train-NeuralNet



ROCs-Validation-NeuralNet



References

- Lopez de Prado, Marcos M (2018), *Advances in Financial Machine Learning, Wiley.*
- Pezeshki et al (2020). *Gradient Starvation:A Learning Proclivity in Neural Networks*, Mohammad Pezeshki, Sekou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, Guillaume Lajoie, arXiv:2011.09468.
- Goddfellow et al (2017), *Deep Learning*, Ian Goodfellow, Yoshua Bengio, Aaron Courville, MIT Press