Machine Learning performance under different predictors' transformations: forecasting 12-month-ahead Brazilian inflation

Tito Bruni

PUC-Rio

03/05/2023

- Política Monetária (Iversen et al,2016)
- Contratos Nominais de longo prazo.
- Decisões de consumo e investimento.

- Medeiros et al. (2021): ML methods beat univariate time series models specially when inflation is more volatile.
- Coulombe et al. (2021): transformations in macroeconomic data can enhance forecasts' accuracy.

- **Fontes**: IBGE e Banco Central
- **Plataforma**: *Sistema Gerenciador de Séries Temporais* (BC)
- **Natureza dos dados**: preços, commodities, atividade econômica, emprego, eletricidade, confiança, finanças, crédito, governo e comércio internacional.
- **Período**: *Janeiro 2006 - Janeiro 2023*

Dados originais

- **Y**: variável dependente.
- **A,B**: variáveis independentes

Y	**A**	**B**
y_1	a_1	b_1
y_2	a_2	b_2
$\vdots$	$\vdots$	$\vdots$
y_{12}	a_{12}	b_{12}
y_{13}	a_{13}	b_{13}
$\vdots$	$\vdots$	$\vdots$

Transformações

- Acumulando variáveis:

Y	A	B
$\tilde{y}_{12}$	$\tilde{a}_{12}$	$\tilde{b}_{12}$
$\tilde{y}_{13}$	$\tilde{a}_{13}$	$\tilde{b}_{13}$
$\vdots$	$\vdots$	$\vdots$
$\tilde{y}_{23}$	$\tilde{a}_{23}$	$\tilde{b}_{23}$
$\tilde{y}_{24}$	$\tilde{a}_{24}$	$\tilde{b}_{24}$
$\vdots$	$\vdots$	$\vdots$

Transformações

- Alterando **Y**:

Y	**A**	**B**
$\tilde{y}_{24}$	$\tilde{a}_{12}$	$\tilde{b}_{12}$
$\tilde{y}_{25}$	$\tilde{a}_{13}$	$\tilde{b}_{13}$
$\vdots$	$\vdots$	$\vdots$
$\tilde{y}_{35}$	$\tilde{a}_{23}$	$\tilde{b}_{23}$
$\tilde{y}_{36}$	$\tilde{a}_{24}$	$\tilde{b}_{24}$
$\vdots$	$\vdots$	$\vdots$

Janela Rolante

- Suponha que as janelas tenham tamanho 11.
- A primeira previsão será:

Y	A	B	
y_{24}	a_{12}	b_{12}	$\hat{f}_{36}(.)$
y_{25}	a_{13}	b_{13}	
$\vdots$	$\vdots$	$\vdots$	
y_{35}	a_{23}	b_{23}	
y_{36}	a_{24}	b_{24}	
$\vdots$	$\vdots$	$\vdots$	

$$\hat{y}_{36} = \hat{f}_{36}(a_{24}, b_{24})$$

Janela Rolante

- A segunda previsão:

Y	A	B	
y_{24}	a_{12}	b_{12}	
y_{25}	a_{13}	b_{13}	$\hat{f}_{37}(.)$
y_{26}	a_{14}	b_{14}	
$\vdots$	$\vdots$	$\vdots$	
y_{36}	a_{24}	b_{24}	
y_{37}	a_{25}	b_{25}	
$\vdots$	$\vdots$	$\vdots$	

$$\hat{y}_{37} = \hat{f}_{37}(a_{25}, b_{25})$$

Janela Rolante

- A terceira previsão:

Y	A	B	
y_{24}	a_{12}	b_{12}	
y_{25}	a_{13}	b_{13}	
y_{26}	a_{14}	b_{14}	$\hat{f}_{38}(.)$
y_{27}	a_{15}	b_{15}	
$\vdots$	$\vdots$	$\vdots$	
y_{37}	a_{25}	b_{25}	
y_{38}	a_{26}	b_{26}	
$\vdots$	$\vdots$	$\vdots$	

$$\hat{y}_{38} = \hat{f}_{38}(a_{26}, b_{26})$$

Janela Rolante

Na monografia:

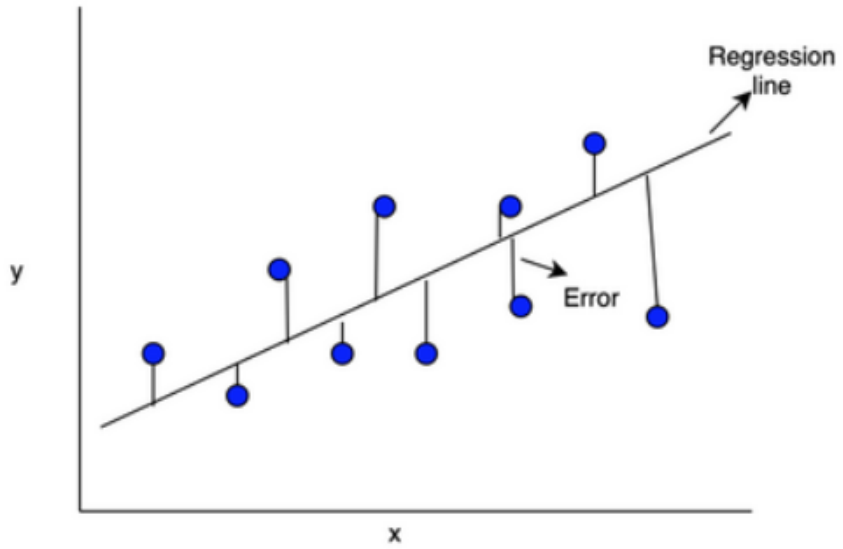
- 83 variáveis explicativas
- janela de tamanho 121

Modelos Estatísticos

De que maneiras podemos calcular $\hat{f}(.)$?

Regressão Linear é útil?

Regressão Linear



$$y_t = \beta_0 + \beta_1 x_t + u_t$$

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t$$

Regressão Linear

- Coeficientes $(\hat{\beta}_0, \hat{\beta}_1)$: $\min \left(\sum_t (y_t - \hat{y}_t)^2 \right)$
- Problemas

Modelos Estatísticos

4 modelos:

1. *LASSO*
2. *Ridge*
3. *Random Forest* (*RF*)
4. Complete Subset Regression (*CSR*)

Lasso e Ridge

Penalizam variáveis irrelevantes.

- *LASSO*: $\min\{\sum_t (y_t - \hat{y}_t)^2 + \lambda \sum_j |\beta_j|\}$
- *Ridge*: $\min\{\sum_t (y_t - \hat{y}_t)^2 + \lambda \sum_j \beta_j^2\}$

Complete Subset Regression

1. Fixa quantidade de variáveis explicativas.
2. Calcula diversas regressões lineares.
3. Extrai a média das previsões geradas nas regressões.

Qual o problema???

Complete Subset Regression

- 83 variáveis explicativas.
- Se cada regressão tivesse apenas 4 variáveis explicativas:

$$\binom{83}{4} = \frac{83!}{4!(83-4)!}$$

- Quase 2 milhões possíveis modelos

Complete Subset Regression

- Por fim, como faço 60 previsões com janela rolante, teria que computar 120 milhões de modelos.
- Pacote *HDeconometrics*

Random Forest

1. Seleciona diversas combinações diferentes de variáveis
2. Seleciona aleatoriamente algumas observações
3. Gera árvores de regressões
4. Computa a média das previsões das árvores de decisões

$$\frac{1}{B}\sum_{b=1}^{B} T_b^*(\mathbf{X})$$

RMSE

Vamos usar o RMSE (root mean squared error) das 60 previsões para comparar suas performances.

$$RMSE_m = \sqrt{\frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \hat{e}_{t,m}^2}$$

RMSE

	RF	LASSO	CSR	Ridge
2	1.726	1.622	2.237	2.491
3	1.753	1.791	2.210	2.116
4	1.805	1.791	2.197	4.656
5	1.805	3.888	2.179	5.703
6	1.766	2.232	2.162	7.448
7	1.776	14.115	2.140	13.554
8	1.725	15.644	2.127	13.916
9	1.737	3.128	2.153	11.421
10	1.746	3.345	2.204	3.950
11	1.739	1.801	2.214	2.761
12	1.719	1.653	4.722	2.789

RMSE normalizado pelo FOCUS

	RF	LASSO	CSR	Ridge
2	0.508	0.477	0.658	0.733
3	0.516	0.527	0.650	0.623
4	0.531	0.527	0.647	1.370
5	0.531	1.144	0.641	1.678
6	0.520	0.657	0.636	2.192
7	0.523	4.154	0.630	3.989
8	0.508	4.604	0.626	4.096
9	0.511	0.921	0.634	3.361
10	0.514	0.984	0.649	1.163
11	0.512	0.530	0.652	0.813
12	0.506	0.486	1.390	0.821