Predição de séries temporais utilizando técnicas clássicas e de aprendizado de máquina

Uma aplicação a índices setoriais econômicos

Richard Sousa Antunes

Fevereiro, 2024

Flávio Almeida de Magalhães Cipparrone
POLI-USP

IME USP

Overview

Overview

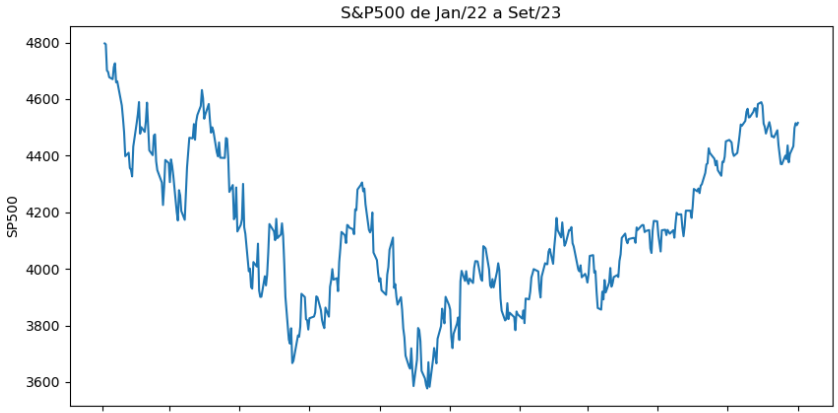
Introdução

- **Predição de séries temporais tem sido extensivamente utilizada, sobretudo em economia e finanças.**
- **No mercado financeiro, uma predição acertiva o suficiente pode trazer lucros ao investidor, o que exemplifica sua relevância.**

- **Não existe fórmula mágica, por melhor que seja o modelo utilizado, não se terá 100% de sucesso.**
- **A dinâmica do mercado financeiro é complexa e difícil de prever**
  - Analisar o mercado por setores, pode amenizar a complexidade

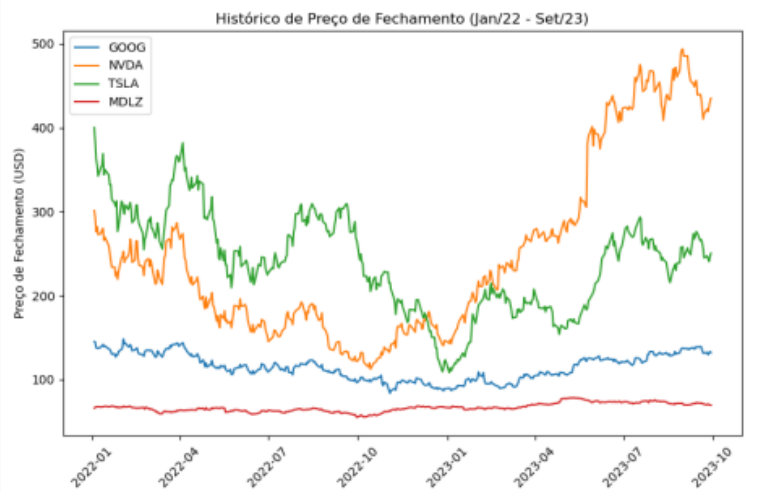
Introdução

- **Exemplo de uma série temporal da S&P**



Introdução

- **Exemplo de uma série temporal múltipla da S&P**



Introdução

- **Esse trabalho se propõe a predizer o retorno de índices setoriais econômicos do mercado americano .**

- **Realiza um comparativo simples entre modelos clássicos de séries temporais (VAR e VARMA) e modernos com uso de técnicas de aprendizado de máquina (MLP).**

Overview

Modelos

- **Existe uma diversidade de modelos para séries temporais multivariadas**
  - VAR, VARMA, VARIMA, VECM, VARMAX, MGARCH, Redes Neurais LSTM, Recorrentes, Prophet, etc
  - será usado 2 modelos clássicos (VAR e VARIMA) e 1 moderno (MLP).

Modelos - VAR

- **O Modelo Vetorial Autorregressivo (VAR) prevê uma série temporal com base nos próprios valores defasados, capturando relações dinâmicas entre múltiplas séries interdependentes**
- **Dada a série $\mathbf{Z}_t = \left(z_{1t}, z_{2t}, ..., z_{nt}\right)$, $t = 1, ..., T$, VAR(p) é descrito por:**

$$\mathbf{Z}_t = \delta + \Phi_1 \mathbf{Z}_{t\text{-}1} + \Phi_2 \mathbf{Z}_{t\text{-}2} + ... + \Phi_p \mathbf{Z}_{t\text{-}p} + \mathbf{U}_t \qquad (1)$$

Sendo Φ_j matriz (n x n) de coefientes dos termos autoregressivos, δ (n x 1) o intercepto (constante) e U_t (n x 1) vetor de ruído branco

Modelos - VAR

- **Necessita da série estacionária e que U_t seja Ruído Branco Gaussiano**
- **Transformação para um formato de regressão [Helmut].**

$$\begin{gathered} Z := \left(Z_1, Z_2, ..., Z_T\right), \quad B := \left(\delta, \Phi_1, \Phi_2, ..., \Phi_p, \right), \\ Z_t^* := \left(1, Z_t, ..., Z_{t-p+1}\right)^{\mathsf{T}}, \; Z^* := \left(Z_0^*, ..., Z^*{}_{T-1}\right), \quad U := \left(U_1, ..., U_T\right) \end{gathered} \tag{2}$$

$$Y = BZ^* + U \tag{3}$$

- **Estimação por MQO**

$$\hat{B} = YZ^{*T}(Z^*Z^{*T})^{-1} \tag{4}$$

VAR - Critérios de Informação

- **Para a ordem p do VAR(p) escolhe o que minimiza os critérios de informações**

$$AIC(\mathbf{p}) = ln|\tilde{\Sigma}_u(\mathbf{p})| + \frac{2pn}{T}$$

$$BIC(\mathbf{p}) = ln|\tilde{\Sigma}_u(\mathbf{p})| + \frac{lnT}{T} pn^2 \quad (5)$$

$$HQ(\mathbf{p}) = ln|\tilde{\Sigma}_u(\mathbf{p})| + \frac{2lnlnT}{T} pn^2$$

onde $|\tilde{\Sigma}_u(\mathbf{p})|$ é o determinante da matriz de covariância dos resíduos obtido pela estimação de máxima verossimilhança

Modelos - VARMA

- **O Modelo Vetorial Autorregressivo de Médias Móveis (VARMA) combina o VAR com as médias móveis, capturando também os choques nos termos de erro.**
- **Dada a série $\mathbf{Z}_t = \left(z_{1t}, z_{2t}, ..., z_{nt}\right)$, VARMA(p,q) é descrito por:**

$$\mathbf{Z}_t = \delta + \Phi_1 \mathbf{Z}_{t-1} + ... + \Phi_p \mathbf{Z}_{t-p} + \mathbf{U}_t - \Theta_1 \mathbf{U}_{t-1} - \cdots - \Theta_q \mathbf{U}_{t-q} \quad (6)$$

Sendo Φ_j matriz (n x n) de coeficientes dos termos autorregressivos, Θ_j matriz (n x n) de coeficientes dos termos de médias móveis, δ o intercepto (constante) (n x 1) , e $\mathbf{U}_t$ um vetor de ruído branco (n x 1).

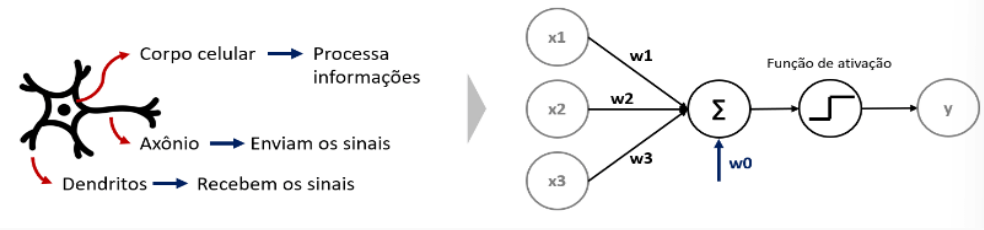
Modelos - VARMA

- A superfície de verossimilhança não possui um único máximo.
- Costuma ser usado para obter um modelo mais parcimonioso do que VAR(p).
- são usados algoritmos numéricos para minimizar.
- a ordem é obtido de forma análoga ao VAR

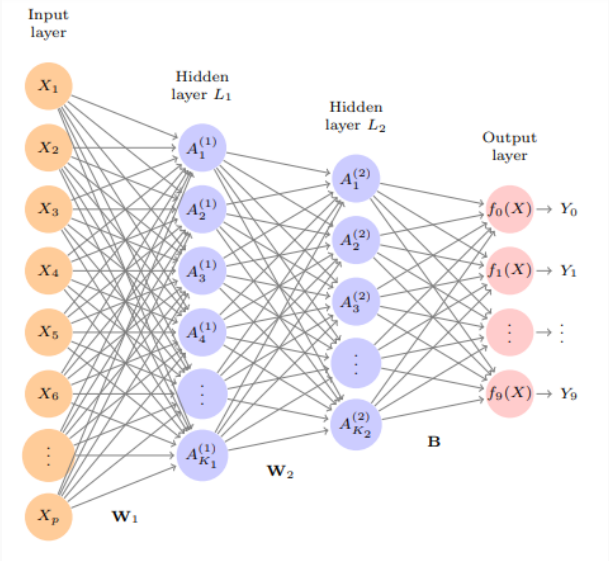
Overview

Modelos - MLP

- **O Modelo de Perceptron Multicamadas (MLP) é uma rede neural artificial composta por múltiplas camadas,que consiste em neurônios que aplicam funções de ativação não lineares a suas entradas. A saída de uma camada serve como entrada para a próxima camada.**



Modelos - MLP


Input layer
Hidden layer L_1
Hidden layer L_2
Output layer
X_1
X_2
X_3
X_4
X_5
X_6
X_p
$A_1^{(1)}$
$A_2^{(1)}$
$A_3^{(1)}$
$A_4^{(1)}$
$A_{K_1}^{(1)}$
$A_1^{(2)}$
$A_2^{(2)}$
$A_3^{(2)}$
$A_{K_2}^{(2)}$
$f_0(X) \to Y_0$
$f_1(X) \to Y_1$
$f_9(X) \to Y_9$
$\mathbf{W}_1$
$\mathbf{W}_2$
$\mathbf{B}$

Modelos - MLP

- **A relação entre a entrada X e a saída f(X) em cada neurônio:**

$$f(X) = \beta_0 + \sum_{k=1}^{K} \beta_k A_k \tag{7}$$

$$A_k = g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} X_j \right). \tag{8}$$

Onde w_{kj} são matrizes de pesos das conexões entre as camadas de entrada, oculta e de saída. β_j são os vetores de viés e $g(\cdot)$ função de ativação não-linear .

Modelos - MLP

- **Funções de ativações mais utilizadas: Sigmoide e ReLU**

  Sigmoide

$$g(z) = \frac{1}{1 + e^{-z}} \tag{9}$$

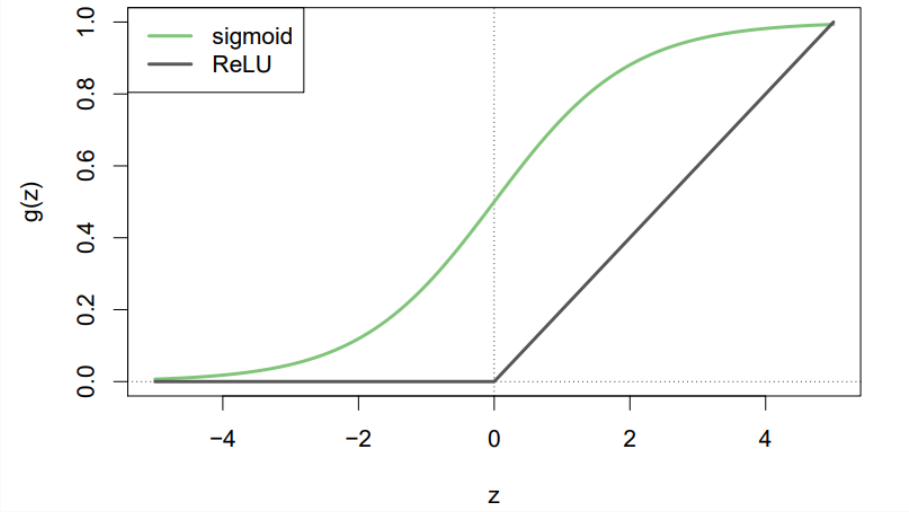
  ReLU (Rectified Linear Unit)

$$g(z) = \max(0, z) \tag{10}$$

- **Encontra-se os parâmetros minimizando o erro quadrático**

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{11}$$

Modelos - MLP



Modelos - MLP

- **Modelos de Aprendizado de Máquina como MLP, recebe X como entrada e prevê $\hat{Y} = \hat{f}(X)$, ou seja, não é preparado inicialmente para lidar com dados sequenciais.**
- **fazendo-se necessário remodelar os dados para usá-lo.**

Overview

Manipulação dos Dados

- **Ações extraídas do S&P.**
  - Usado para definição dos setores a classificação GICS (Global Industry Classification Standard)
  - Setores:
    - » *Energia, Materiais, Industrial, Bens de Consumo Discricionário, Produtos Básicos de Consumo, Assistência Médica, Serviços Financeiros, Tecnologia da Informação, Serviços de Comunicações, Serviços de Utilidade Pública e Imóveis.*

Manipulação dos Dados

- **Seleção de ações para representação de setores.**
  - Será selecionado as 10 ações com maior valor de mercado (VM):

$$VM_j = V_j \quad x \quad P_j,$$

$$V_j = \text{Volume}, \quad P_j = \text{Preço de fechamento}$$

Manipulação dos Dados

- **Substituição das séries.**
  - Destas 10 ações escolhidas para cada ação, será usado como a série a ser utilizada, a média de seus retornos líquido simples:

$$\mathcal{R}_{it} = \frac{1}{10} \sum_{k=1}^{10} \left(\frac{P_{kit}}{P_{kit\text{-}1}} \text{-} 1 \right) \tag{12}$$

Onde P_{kit} é o preço de fechamento da ação k, no setor i e período t

- **Utilizar o retorno, torna a série estacionária [Morettin]**

Manipulação dos Dados

Rank	Technology		Communication Services [1]		Consumer Cyclical	
1	MSFT	2.89T	GOOG	1.79T	AMZN	1.60T
2	AAPL	2.88T	META	962.39B	TSLA	695.83B
3	NVDA	1.35T	NFLX	215.41B	HD	354.02B
4	AVGO	518.55B	TMUS	187.97B	MCD	212.87B
5	ORCL	293.04B	CMCSA	173.04B	NKE	159.18B
6	ADBE	272.10B	DIS	165.37B	LOW	125.90B
7	CRM	263.23B	VZ	162.11B	BKNG	122.20B
8	AMD	236.77B	T	117.83B	TJX	108.30B
9	ACN	223.42B	CHTR	53.98B	SBUX	104.55B
10	CSCO	204.56B	EA	37.10B	ABNB	87.90B

Table 1: *Top 10 empresas em valor de mercado dos setores de Tecnologia, Serviçõe de comunicação e Consumo Cíclico*

Manipulação dos Dados

Rank	Healthcare		Financial Services		Consumer Defensive	
1	LLY	610.33B	V	542.92B	WMT	434.31B
2	UNH	482.36B	JPM	492.34B	PG	354.95B
3	JNJ	390.92B	MA	402.40B	COST	303.17B
4	MRK	300.61B	BAC	259.12B	KO	261.09B
5	ABBV	286.72B	WFC	170.59B	PEP	229.97B
6	TMO	210.31B	MS	148.55B	PM	147.89B
7	ABT	197.77B	BX	144.34B	MDLZ	99.51B
8	DHR	166.44B	SPGI	138.44B	MO	73.11B
9	AMGN	164.04B	AXP	132.66B	CL	66.62B
10	PFE	162.05B	GS	123.19B	TGT	65.05B

Table 2: *Top 10 empresas em valor de mercado dos setores de Serviços Financeiros, Saúde e Energia*

Manipulação dos Dados

Rank	Energy		Basic Materials		Industrials	
1	XOM	399.48B	LIN	198.28B	CAT	147.84B
2	CVX	278.01B	SHW	77.43B	UNP	144.91B
3	COP	132.63B	FCX	58.71B	GE	141.31B
4	SLB	71.35B	APD	58.70B	UPS	134.87B
5	EOG	67.67B	ECL	56.72B	HON	132.57B
6	MPC	59.79B	NEM	43.38B	BA	131.70B
7	PSX	58.46B	NUE	41.52B	RTX	124.13B
8	PXD	52.26B	DOW	37.43B	LMT	114.92B
9	OXY	50.95B	PPG	34.32B	DE	108.22B
10	VLO	44.77B	CTVA	32.49B	ADP	97.05B

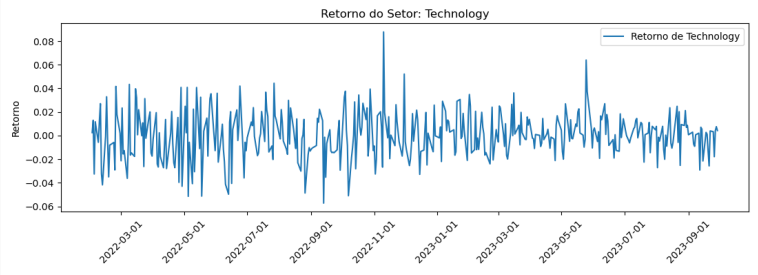
Table 3: *Top 10 empresas em valor de mercado dos setores de Consumo Defensivo, Materiais Básicos e Industriais*

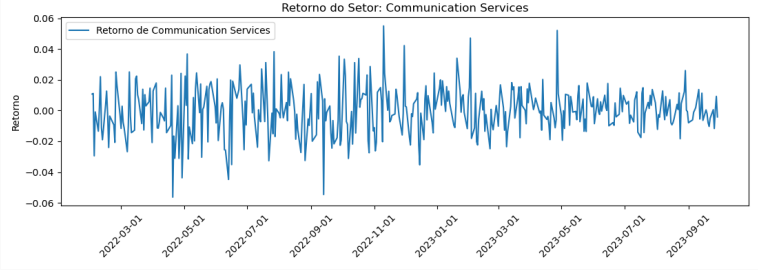
Manipulação dos Dados

Rank	Utilities		Real Estate	
1	NEE	125.24B	PLD	121.55B
2	SO	77.89B	AMT	97.63B
3	DUK	75.88B	EQIX	76.52B
4	SRE	47.46B	SPG	55.01B
5	PCG	44.69B	PSA	52.02B
6	AEP	42.85B	WELL	50.88B
7	D	39.61B	CCI	48.88B
8	CEG	36.06B	O	42.64B
9	EXC	35.91B	DLR	42.40B
10	XEL	33.83B	CSGP	34.06B

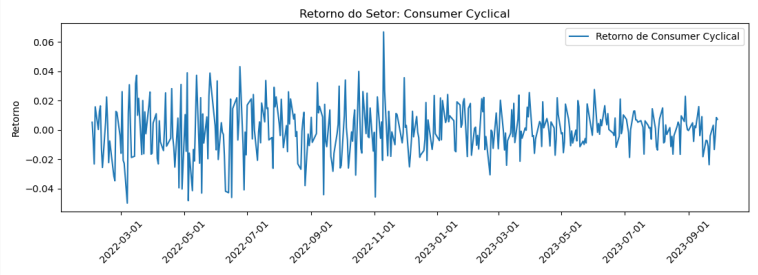
Table 4: *Top 10 empresas em valor de mercado dos setores de Utilidades Públicas e Imobiliário*

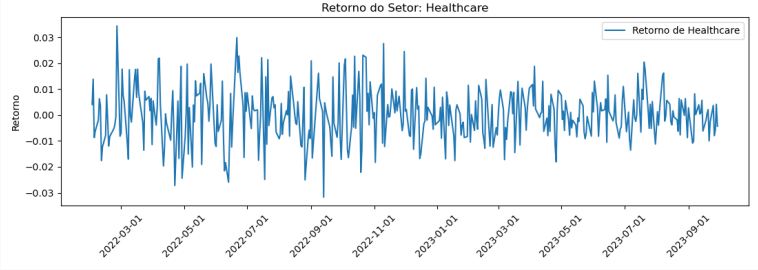
Manipulação dos Dados



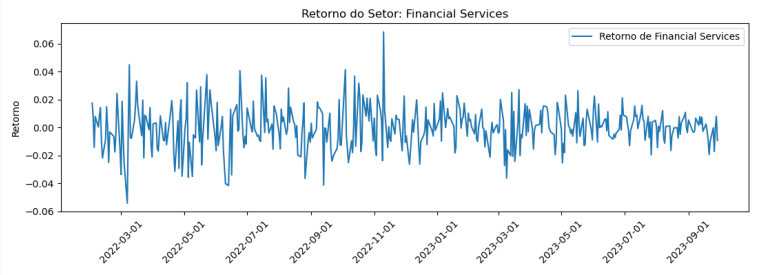


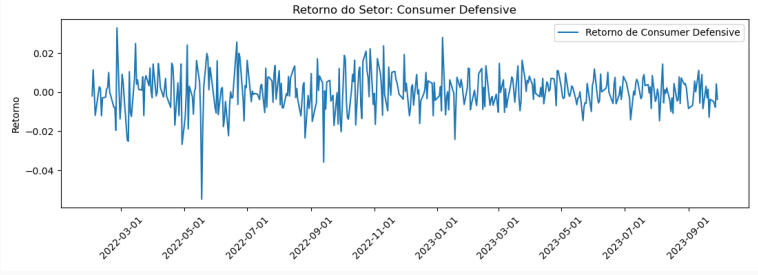
Manipulação dos Dados



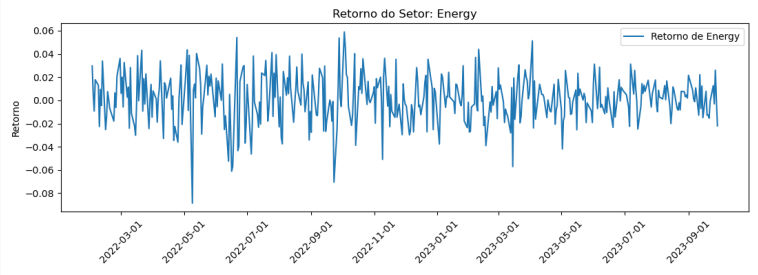


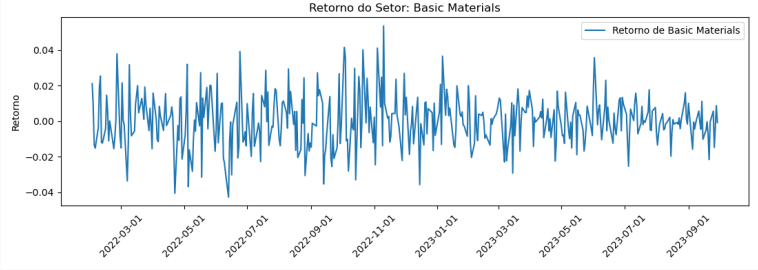
Manipulação dos Dados



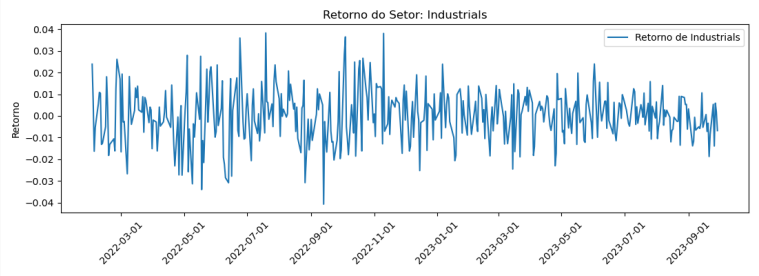


Manipulação dos Dados





Manipulação dos Dados



Manipulação dos Dados

- **Remodelagem dos ddados para o formato de aprendizado de máquina**

tempo	$z_1 t$	$z_2 t$
01/02/2023	12	89
02/02/2023	0	7
03/02/2023	34	42
04/02/2023	21	22
05/02/2023	12	1

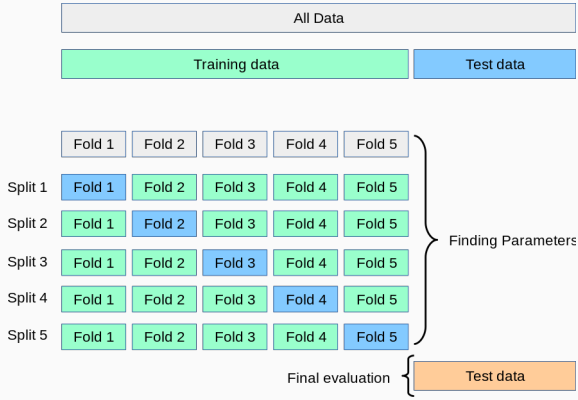
↓

x_1	x_2	y_1	y_2
?	?	12	89
12	89	0	7
0	7	34	42
34	42	21	22
21	22	12	1
12	1	y_1 a ser predito	y_2 a ser predito

Overview

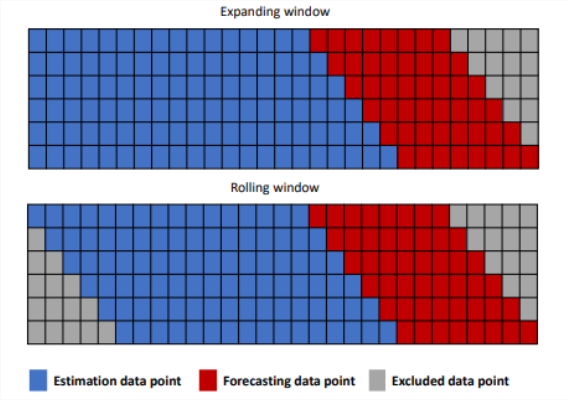
Técnica de validação dos modelos

- **Cross-Validation extensivamente utilizada em Aprendizado de Máquina, mas que falha em séries temporais**



Técnica de validação dos modelos

- Janela Expansiva com o conjunto aumentando de tamanho
- Janela Deslizante com o conjunto mantendo aproximadamente o tamanho e transladando no período



Técnica de validação dos modelos

- **Será escolhido Janela Expansiva com os períodos:**
  1. Treino: 03/01/2022 a 11/07/2022, Teste: 12/07/2022 a 18/07/2022
  2. Treino: 03/01/2022 a 13/01/2023, Teste: 17/01/2023 a 23/01/2023
  3. Treino: 03/01/2022 a 22/09/2023, Teste: 25/09/2023 a 29/09/2023

Treino: Conjunto usado para ajustar os parâmetros dos modelos

Teste: Conjunto usado para predição a fim de verificar a métrica de erro cometida.

Overview

Resultados

- **Critérios de informação no modelo VAR(p)**

	Treino1			Treino2			Treino3		
	AIC	BIC	HQ	AIC	BIC	HQ	AIC	BIC	HQ
1	-100.40	-97.72	-99.31	-101.51	-99.84	-100.84	-103.28	-102.14	-102.83
2	-99.76	-94.37	-97.57	-101.10	-97.77	-99.76	-103.06	-100.77	-102.16
3	-99.11	-90.98	-95.81	-100.82	-95.81	-98.81	-102.83	-99.40	-101.48
4	-98.92	-88.02	-94.49	-100.50	-93.80	-97.81	-102.64	-98.06	-100.83
5	-98.46	-84.77	-92.90	-100.05	-91.65	-96.67	-102.36	-96.62	-100.09
6	-98.69	-82.18	-91.98	-99.69	-89.58	-95.62	-102.16	-95.26	-99.44
7	-99.58	-80.22	-91.72	-99.43	-87.60	-94.67	-101.96	-93.90	-98.78
8	-101.13	-78.88	-92.09	-99.08	-85.52	-93.62	-101.77	-92.54	-98.12
9	-105.95	-80.79	-95.73	-98.95	-83.66	-92.80	-101.71	-91.31	-97.60
10	-149.00	-120.90	-137.59	-98.89	-81.84	-92.03	-101.55	-89.97	-96.97

Escolhe a ordem $p = 1$ que minimizou em 2 das 3 janelas

Resultados

- **Critérios de informação no VARMA(p,q)**
  Escolhe a ordem (1,1)

Treino 1

p	q	AIC	BIC	HQ
1	1	-8626.66	-7743.46	-8267.79
1	2	-8524.52	-7294.34	-8024.66
2	1	-8526.27	-7296.10	-8026.41
2	2	-8300.37	-6723.23	-7659.53

Treino 2

p	q	AIC	BIC	HQ
1	1	-17891.11	-16794.42	-17450.23
1	2	-17774.95	-16247.42	-17160.87

Treino 3

p	q	AIC	BIC	HQ
1	1	-30821.04	-29567.25	-30326.10
1	2	-30705.37	-28959.03	-30015.99
2	1	-30703.69	-28957.34	-30014.30
2	2	-30467.40	-28228.50	-29583.57

Resultados

- **MSE e RMSE obtido no VAR(1)**

(a) *Treino1*

Data	MSE	RMSE
2022-07-12	0.000101	0.010051
2022-07-13	0.000046	0.006782
2022-07-14	0.000121	0.010999
2022-07-15	0.000428	0.020677
2022-07-18	0.000153	0.012361

(b) *Treino2*

Data	MSE	RMSE
2023-01-17	0.000040	0.006352
2023-01-18	0.000314	0.017734
2023-01-19	0.000121	0.011010
2023-01-20	0.000337	0.018367
2023-01-23	0.000158	0.012569

(c) *Treino3*

Data	MSE	RMSE
2023-09-25	0.000022	0.004708
2023-09-26	0.000253	0.015906
2023-09-27	0.000100	0.010000
2023-09-28	0.000081	0.008994
2023-09-29	0.000069	0.008317

Resultados

- **MSE e RMSE obtido no VARMA(1,1)**

(a) *Treino1*

Date	MSE	RMSE
2022-07-12	0.000095	0.009748
2022-07-13	0.000045	0.006700
2022-07-14	0.000121	0.011001
2022-07-15	0.000429	0.020720
2022-07-18	0.000153	0.012355

(b) *Treino2*

Date	MSE	RMSE
2023-01-17	0.000038	0.006177
2023-01-18	0.000315	0.017744
2023-01-19	0.000122	0.011030
2023-01-20	0.000337	0.018363
2023-01-23	0.000158	0.012569

(c) *Treino3*

Date	MSE	RMSE
2023-09-25	0.000021	0.004573
2023-09-26	0.000253	0.015913
2023-09-27	0.000100	0.009994
2023-09-28	0.000081	0.008994
2023-09-29	0.000069	0.008317

Resultados

- **MSE e RMSE obtido no MLP pegando 1 e 5 dias atrás**

(a) *Treino 1 usando 1 dia*

Date	MSE	RMSE
2022-07-12	0.000209	0.014462
2022-07-13	0.000077	0.008799
2022-07-14	0.000214	0.014628
2022-07-15	0.000080	0.008937
2022-07-18	0.000140	0.011814

(b) *Treino 2 usando 1 dia*

Date	MSE	RMSE
2023-01-17	0.000150	0.012230
2023-01-18	0.000154	0.012410
2023-01-19	0.000075	0.008672
2023-01-20	0.000117	0.010804
2023-01-23	0.000298	0.017271

(c) *Treino 3 usando 1 dia*

Date	MSE	RMSE
2023-09-25	0.000027	0.005221
2023-09-26	0.000266	0.016313
2023-09-27	0.000119	0.010895
2023-09-28	0.000069	0.008335
2023-09-29	0.000098	0.009924

(d) *Treino 1 usando 5 dias*

Date	MSE	RMSE
2022-07-12	0.001213	0.034827
2022-07-13	0.000313	0.017687
2022-07-14	0.000524	0.022884
2022-07-15	0.000637	0.025246
2022-07-18	0.000268	0.016373

(e) *Treino 2 usando 5 dias*

Date	MSE	RMSE
2023-01-17	0.000092	0.009589
2023-01-18	0.000345	0.018581
2023-01-19	0.000088	0.009371
2023-01-20	0.000308	0.017558
2023-01-23	0.000039	0.006263

(f) *Treino 3 usando 5 dias*

Date	MSE	RMSE
2023-09-25	0.000081	0.009019
2023-09-26	0.000053	0.007284
2023-09-27	0.000077	0.008776
2023-09-28	0.000183	0.013541
2023-09-29	0.000072	0.008483

Resultados

- **MSE e RMSE médios**

Sector	VAR(1)	VARMA(1,1)	MLP 1 dia	MLP 5 dias
Technology	0.000274	0.000502	0.002231	0.021428
Communication Services	0.007297	0.007528	0.002364	0.012190
Consumer Cyclical	0.000272	0.000174	0.003857	0.025443
Financial Services	0.020383	0.020695	0.016922	0.000198
Healthcare	0.011043	0.010949	0.009599	0.003771
Energy	0.014529	0.013961	0.007648	0.025666
Consumer Defensive	0.017096	0.017106	0.017373	0.015085
Basic Materials	0.005065	0.005356	0.003539	0.016515
Industrials	0.008938	0.009174	0.008189	0.008369
Utilities	0.061542	0.061500	0.061500	0.052960
Real Estate	0.015513	0.015713	0.011480	0.002176

Tabela 7.5: *Comparação da RMSE para cada um dos modelos para diferentes setores.*

Sector	VAR	VARMA	MLP 1 dia	MLP 5 dias
Average Value	0.014723	0.014787	0.013155	0.016709

Tabela 7.6: *Média do RMSE por modelo*

Resultados

- **Retorno acumulado realizado**

Sector	Retorno Acumulado
Technology	0.001354
Communication Services	-0.004807
Consumer Cyclical	0.001877
Financial Services	-0.018550
Healthcare	-0.010171
Energy	0.014850
Consumer Defensive	-0.017566
Basic Materials	-0.004277
Industrials	-0.007858
Utilities	-0.060948
Real Estate	-0.013772

Tabela 7.7: *Retorno acumulado por setor.*

Overview

Conclusão

- A ordem de grandeza do retorno predito está maior que o retorno realizado, evidenciando os modelos não terem conseguido extrair um padrão dos dados
- VAR e VARMA não conseguiu performar bem, pois ambos ficaram com as menores ordem e ainda teve um Erro Quadrático relativamente grande.
- MLP teve um RMSE levemente menor, mas não o bastante pra dizer que é melhor que os outros nesse contexto

Conclusão

- Esses resultados evidencia a dificuldade de predizer o mercado financeiro e econômico, com os modelos clássicos e modernos simples
- Há uma diversidade de modelos que podem ser testados para tentar melhorar a performance de predição, como GARCH, LSTM, Prophet, etc