

GVCode

Apresentação final

Const Dev Challenge 2022.2

CONSTELLATION
ASSET MANAGEMENT



1. O Challenge
2. O Modelo
3. Análise da série
4. Seleção das variáveis
5. Seleção do modelo
6. Resultados do modelo
7. Aplicações
8. Conclusão

A Equipe

GVCcode | CONSTELLATION



Victor Laube Dutra
FGV ADM
4º semestre



Bruno Ting Erlong
FGV Econo
6º semestre



Ricardo Semião e Castro
FGV Econo
8º semestre

1. O Challenge

1. O Challenge

GVCCode | CONSTELLATION

- Previsão do fluxo de dinheiro investido dentro dos fundos de ações

$$\text{FLUXO} = \text{CAPTAÇÕES} - \text{RESGATES}$$

1ª ETAPA

Quantitativa

Competição de Data Science - Kaggle

Avaliação por RMSE

GVCCode no 1º lugar (747.841 RMSE)

2ª ETAPA

Qualitativa

Entrega do relatório sobre o modelo

Verificação funcionalidade e aplicabilidade

Otimização do modelo

2. O Modelo

2. O Modelo

- Desenvolvido em Python (Jupyter)
- Rede Neural - Perceptron Multicamadas
- Treinado em 3 anos (2016 - 2018)

- Variáveis:

MERCADO NACIONAL

Índice imobiliário (IMOB)

Índice financeiro (IFNC)

Índice energia elétrica (IEE)

MERCADO EXTERNO

Nasdaq 100 (NDX)

S&P 500 (SPX)

Cotação Dólar

GVCCode | CONSTELLATION

RESULTADOS

(19/05/2021 - 26/07/2022)

658,08 RMSE

409,17 MAE

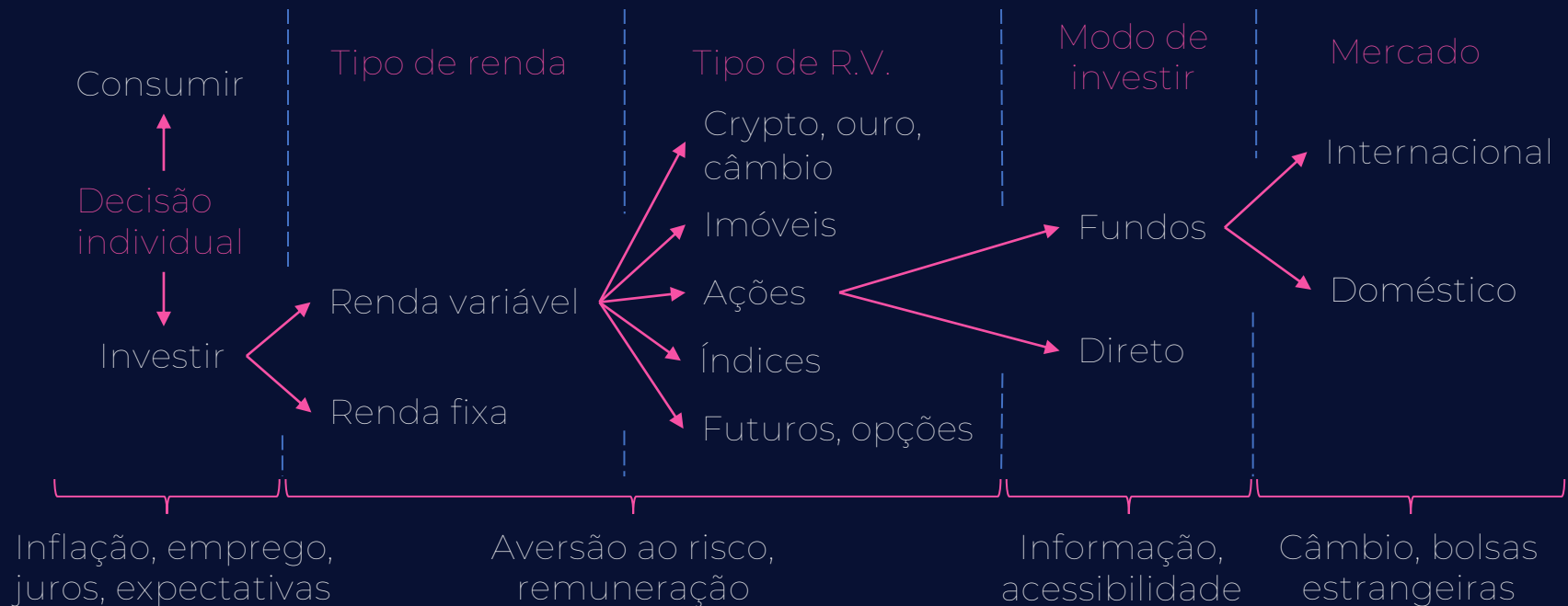
3,41% MAPE

3.

Análise da Série

3. Análise da Série

- Características Teóricas - Decisão de Investimento



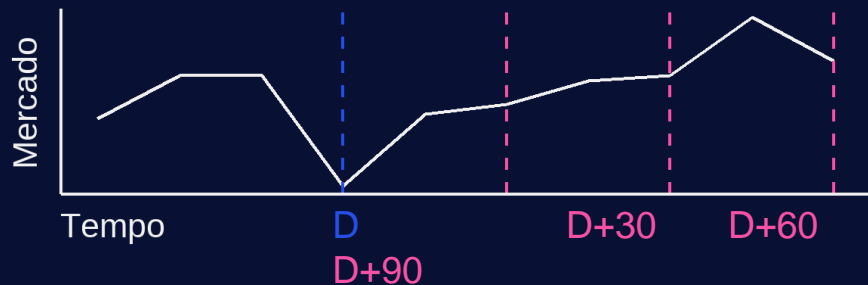
3. Análise da Série

GVCCode | CONSTELLATION

- Características Operacionais - Fundos & B³

Fundos

- Podem ser abertos ou fechados
- Prazo de cotização
 - Agente toma decisão em D, mas o fundo só realiza o resgate / captação em D + x
 - Para renda variável, o prazo costuma ser 7, até 120 dias



3. Análise da Série

GVCCode | CONSTELLATION

- Características Estatísticas - Outliers

Presença:

- 26 obs. > 2 desvios padrões
- 2 obs. > 20 desvios padrões
- Maior presença durante a pandemia

Problema:

- Enviar os modelos para amostras extremas
- Punição de grandes erros via RMSE

Solução:

- identificação e substituição pelo método de quantis

Valores Históricos do Fluxo (mil R\$)



3. Análise da Série

GVCCode | CONSTELLATION

- Características Estatísticas - Sazonalidade

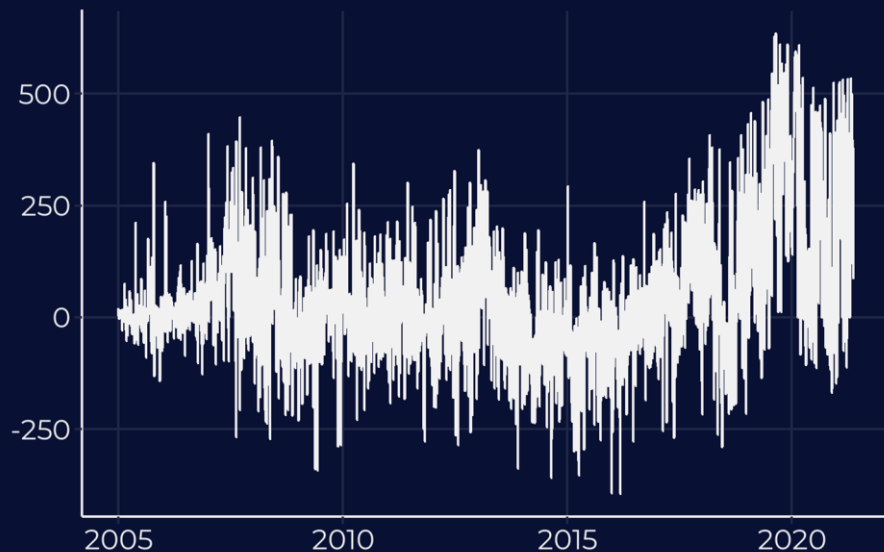
Sazonalidade:

- Fatos estilizados do mercado financeiro
 - Efeitos de dia da semana
 - Efeitos de trimestre
- Modelos ARMA sazonais
- Ambos métodos com baixo poder explicativo

Tendência:

- Apresenta estacionariedade
- Série centrada no 0
- Possível quebra a partir de 2019

Valores Históricos do Fluxo (R\$)
Sem outliers



3. Análise da Série

- Características Estatísticas - Quebras estruturais

Presença:

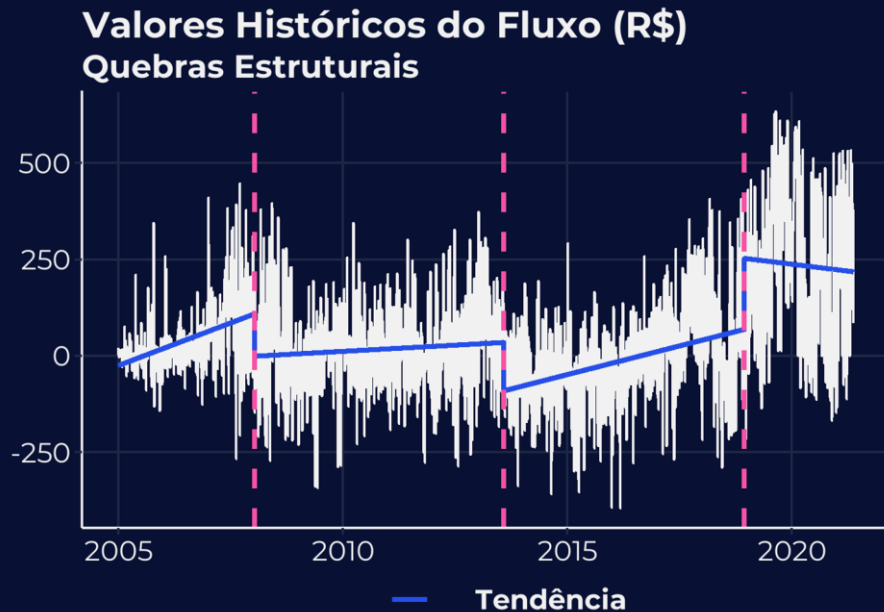
- Teste com regressões segmentadas
- Possível mudança no processo gerador dos dados

Problema:

- Viesar estimativas
- Possível volta ao processo original pós pandemia

Solução:

- limpar a tendência, ou treinar apenas pré/pós 2019



3. Análise da Série

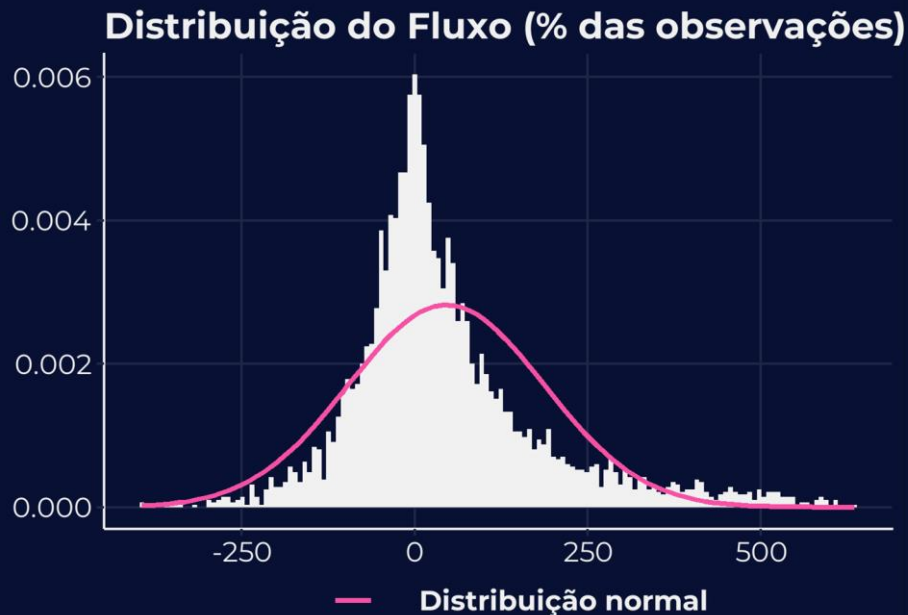
- Características Estatísticas - Distribuição

Características:

- Alta densidade próximo à média
- Alta densidade de longes da média
- Assimetria

Implicações para modelagem:

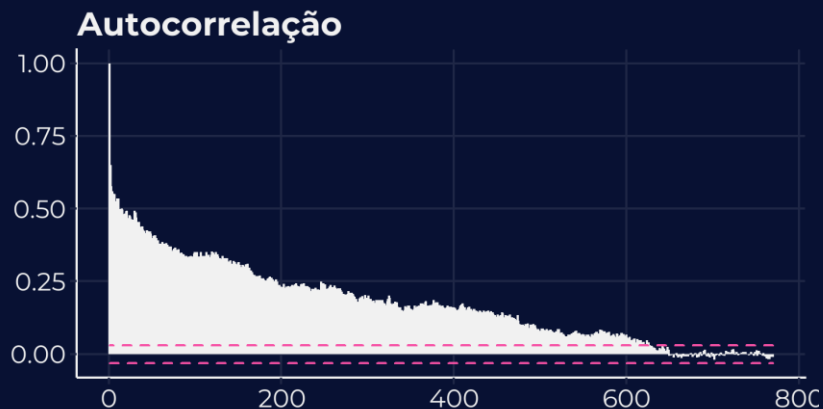
- Não desviar muito da média, mas conseguir prever valores extremos
- Prever valores altos positivos com maior frequência



3. Análise da Série

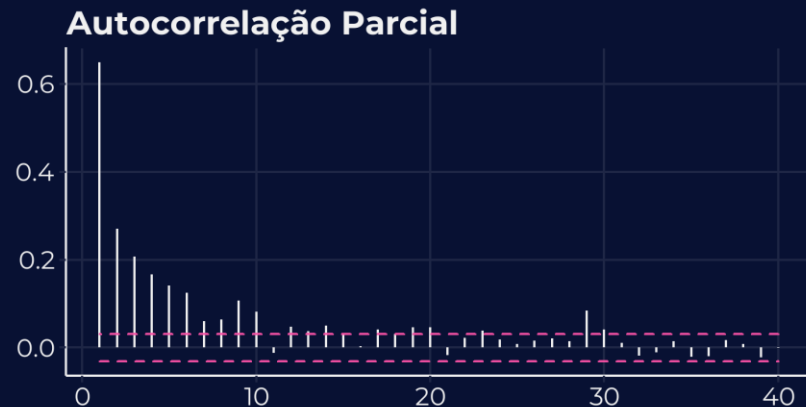
GVCCode | CONSTELLATION

- Características Estatísticas - Autocorrelação



Características:

- Correlação muito persistente
- Forte dinâmica temporal



Implicação para modelagem:

- Modelos que incluam a dinâmica de séries de tempo

4.

Seleção das variáveis



4. Seleção das variáveis

GVCCode | CONSTELLATION

Intuição Econômica

Macroeconômicas

- IPCA
- PIB
- SELIC
- Desemprego
- Federal Funds Rate

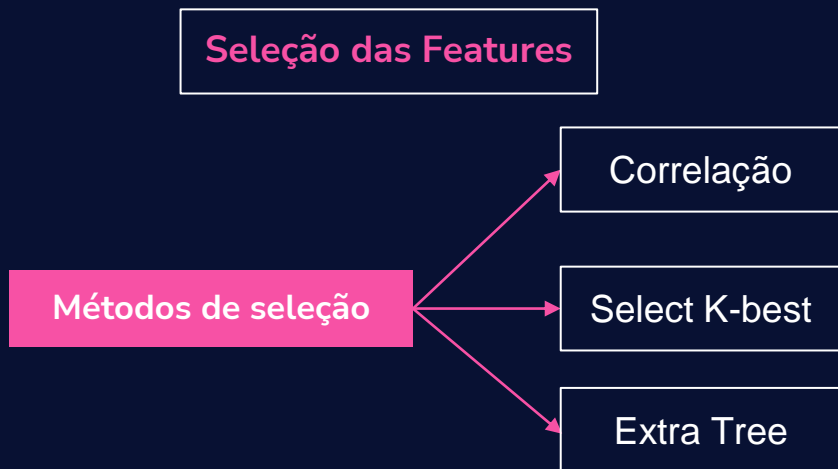
Mercado Externo

- Câmbio
- SPX
- NDX
- N100
- DJI
- Outras Bolsas

Decomposição IBOV

- IEE
- IFNC
- IMOB
- IMAT

4. Seleção das variáveis



GVCCode | CONSTELLATION

Possível problema

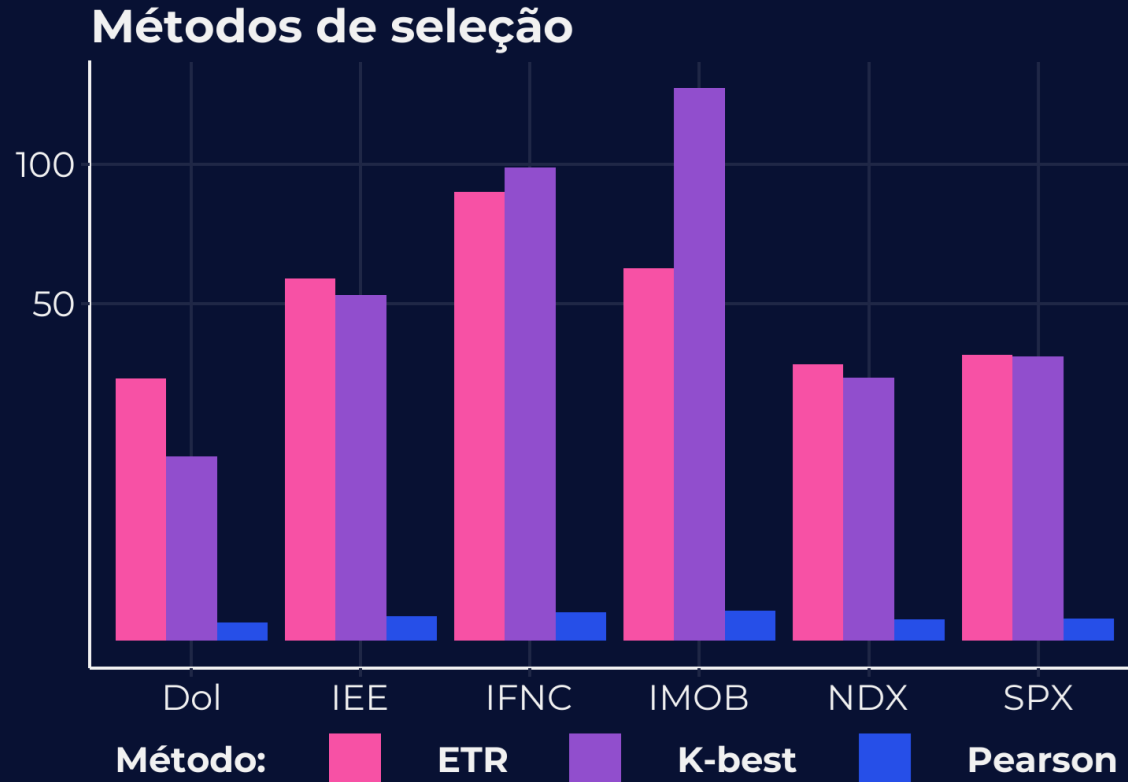
- Os métodos selecionar as variáveis que explicam bem apenas naquele período

Solução encontrada

- Utilizar teoria econômica para complementar os métodos empíricos

4. Seleção das variáveis

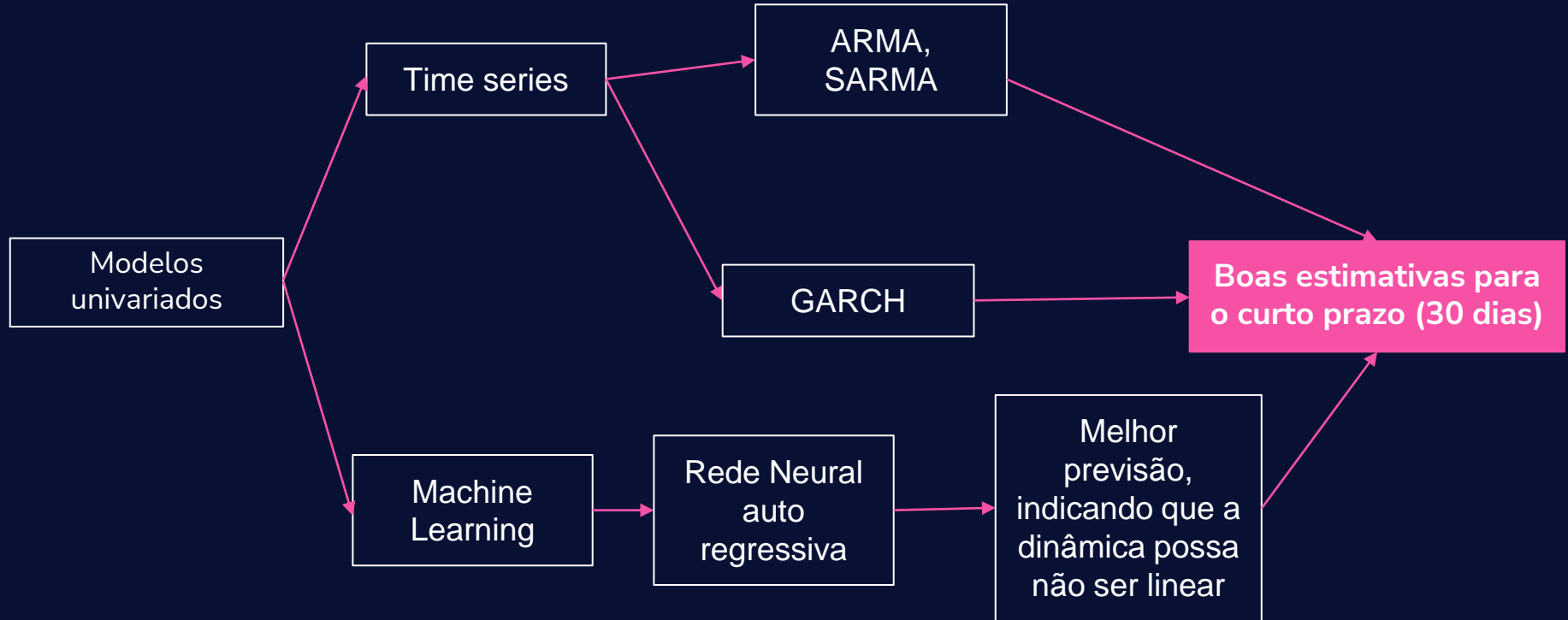
GVCCode | CONSTELLATION



5. Seleção do Modelo

5. Seleção do modelo: **univariados**

GVCCode | CONSTELLATION



5. Seleção do modelo: **multivariados**

GVCode | CONSTELLATION

PRINCIPAIS CANDIDATOS

Redes Neurais

VANTAGENS

Alta performance com bases grandes, complexas e não lineares

DESVANTAGENS

Performance dependente do treino
Computação demorada

Ensembles

VANTAGENS

Alta acurácia de predição
Baixa variância e viés
Identifica padrões ocultos

DESVANTAGENS

Amostra x População
Difíceis de interpretar e montar

Support Vectors

VANTAGENS

Excelente generalização
Robusto à outliers
Facilmente atualizado

DESVANTAGENS

Não é adequado para bases grandes com muito *noise*

5. Seleção do modelo: multivariados

GVCCode | CONSTELLATION

- Avaliação dos modelos por validação cruzada K-Fold

COMO FUNCIONA

1. Divide a série em k grupos
2. Para cada grupo:
 - a. Separa o grupo como teste
 - b. Treina o modelo com o resto
 - c. Avalia na base de teste
 - d. Armazena o resultado e descarta o modelo
3. Resume a acurácia do modelo

k = 1	TESTE	TREINO	TREINO	TREINO
k = 2	TREINO	TESTE	TREINO	TREINO
k = 3	TREINO	TREINO	TESTE	TREINO
k = 4	TREINO	TREINO	TREINO	TESTE

5. Seleção do modelo: **multivariados**

- Avaliação dos modelos por validação cruzada K-Fold

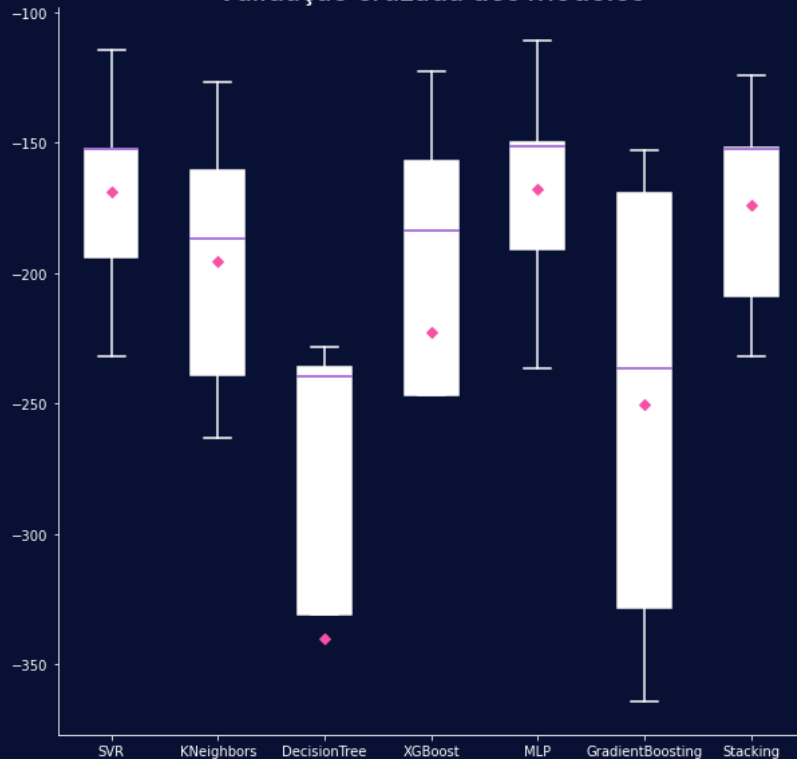
RESULTADOS

A rede neural obteve o melhor desempenho até mesmo que o modelo de stacking

O que mais se aproximou da rede neural foi o SVR, porém não teve boa performance na hora de testar o modelo

GVCCode | CONSTELLATION

Validação cruzada dos modelos



5. Seleção do modelo: MLP

GVCCode | CONSTELLATION

- Multilayer Perceptron (MLP)

Hidden layer

$$\sum_{i=1}^m w_i x_i + bias$$

Input layer

Output layer

DADOS

PREDIÇÃO

Backpropagation ← Feedforward →

5. Seleção do modelo: MLP

GVCCode | CONSTELLATION

- Parametrização da rede neural

PRINCIPAIS PARÂMETROS

- Função de ativação = Rectified Linear Unit (ReLU)
- Camadas escondidas = (100,)
- Solver = Stochastic Gradient Descent (sgd)
- Taxa de aprendizagem = Adaptive
- Alfa = 0.05

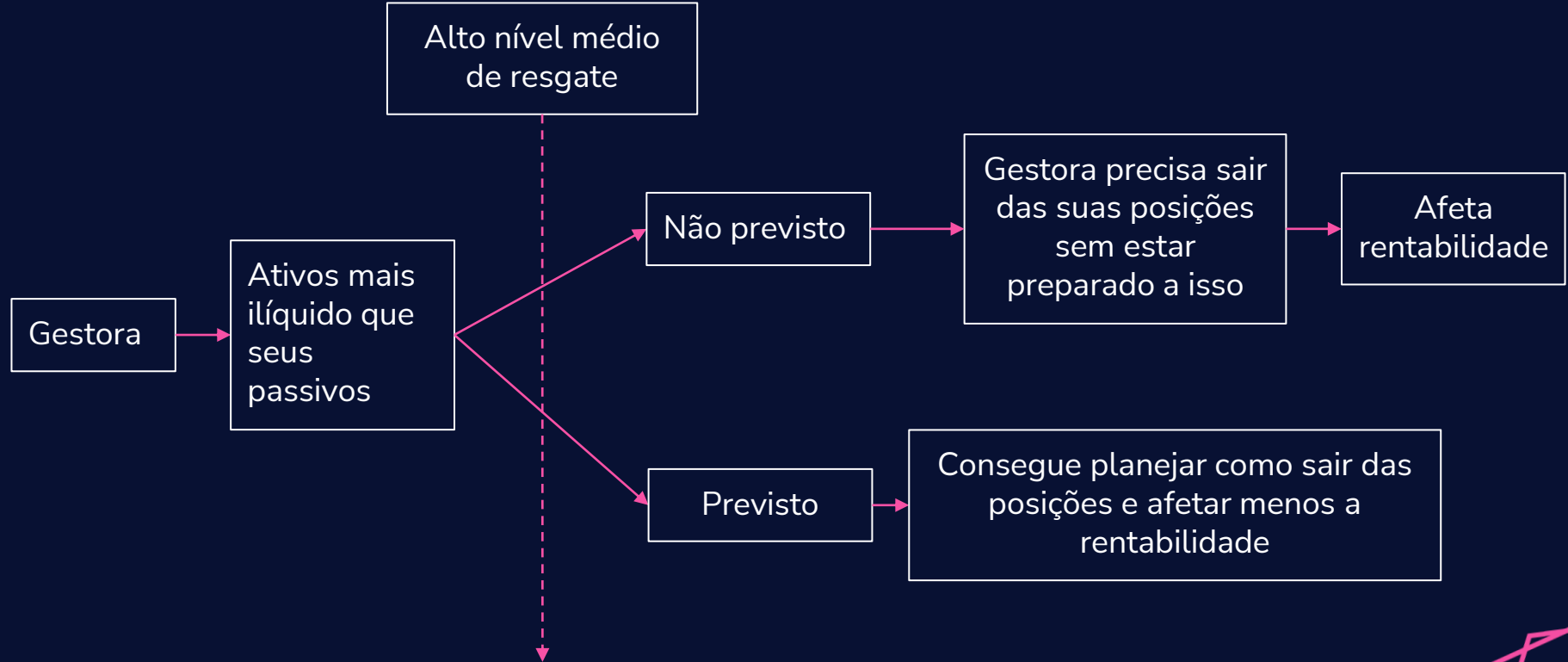
GRIDSEARCH

- Ferramenta da biblioteca sklearn
- Testa diversas combinações de parâmetros
- Retorna combinação ideal de acordo com RMSE

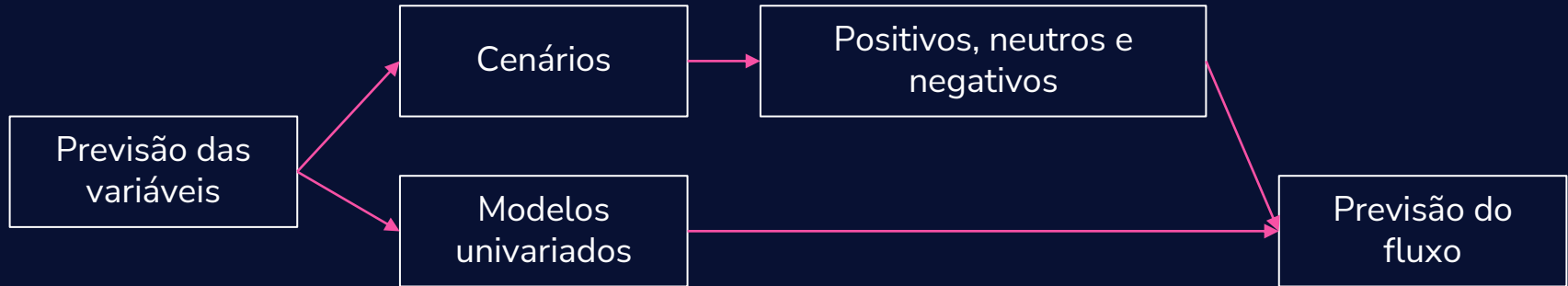
6. Aplicação

6. Aplicabilidade- Problema comum nas gestoras

GVCCode | CONSTELLATION



- Previsão de fluxos futuros
 - Modelo multivariado
 - Necessidade de prever outras variáveis



7. Resultados

7. Resultados

GVCCode | CONSTELLATION

- Robustez do modelo
 - Teste de performance para vários períodos de previsão comuns
 - Para cada período, fizemos 5 testes, variando o tamanho da base de treino e teste
- 15 dias a frente:
 - MAPE: 2,253 (0,872)
- 1 mês a frente:
 - MAPE: 3,096 (2,446)
- 1 semestre a frente:
 - MAPE: 3,382 (1,1601)

7. Resultados

GVCCode | CONSTELLATION

- Robustez do modelo

Variável	Nível	Período	RMSE	RMA	MAPE
Selic	alto	2 anos	413	99	2,8%
	baixo	4 anos	285	228	1,2%
	baixo	2 anos	479	394	1,7%
IPCA	alto	1 ano	139	88	3,7%
	baixo	4 anos	413	330	1,2%
Desemprego	alto	4 anos	350	274	1,4%
Dólar	baixo	2 anos	479	394	1,7%

8. Conclusão

8. Conclusão

GVCCode | CONSTELLATION

- Modelo final e insights

O MODELO

- Bom desempenho constante
- Robusto a vários períodos e cenários
- Aplicações para gestoras
- Mecanismo complexo de rede neural
- Fácil de ser adaptado e atualizado
- Dados de fácil acesso

INSIGHTS

- Complexidade do fluxo
- Relacionamento com o mercado americano
- Impacto macroeconômico da pandemia

Obrigado!

Equipe
GVCode