



Modelo multinível e o efeito empresa para a detecção de fraudes: uma comparação com modelos Logístico e XGBoost

Data Science & Analytics

Samuel Haddad Simões Machado
Prof. Dr. Francisco Lledo dos Santos

INTRODUÇÃO



O segmento bancário vive uma intensa transformação

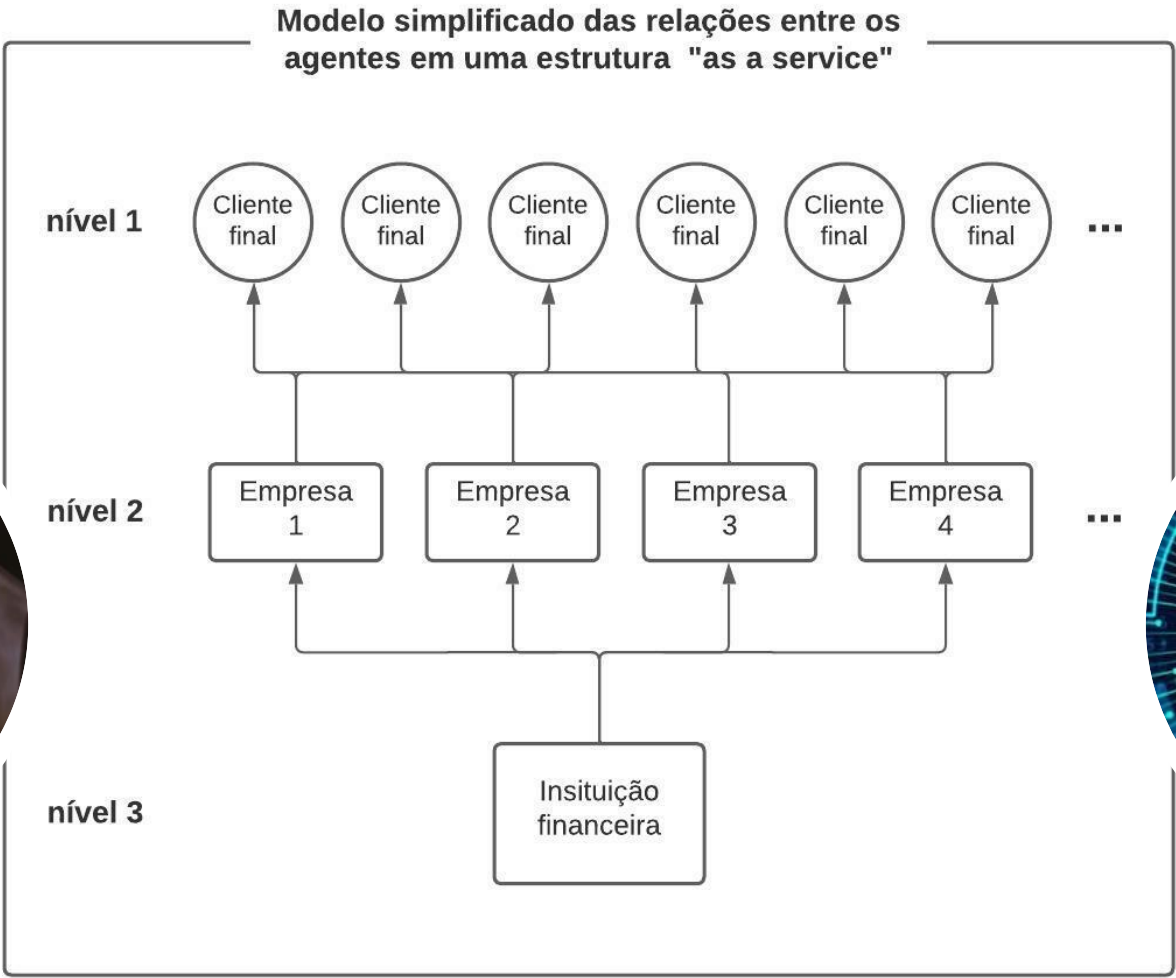
Aceleração da digitalização, ampliação da oferta dos produtos financeiros na modalidade “as a service”, criação do PIX, advento do “open banking”, ambiente “big data” e crescimento das fraudes bancárias são alguns dos elementos da nova realidade do setor financeiro.



Mudanças
estruturais:
de mercado e
tecnológicas



cada vez mais
digitais



cada vez mais
"as a service"



cada vez mais
processamento

INTRODUÇÃO


oportunidade

Um ambiente muito mais
veloz e dependente dos
dados e das soluções
tecnológicas

ameaça

Porém, com cada vez
mais fraude

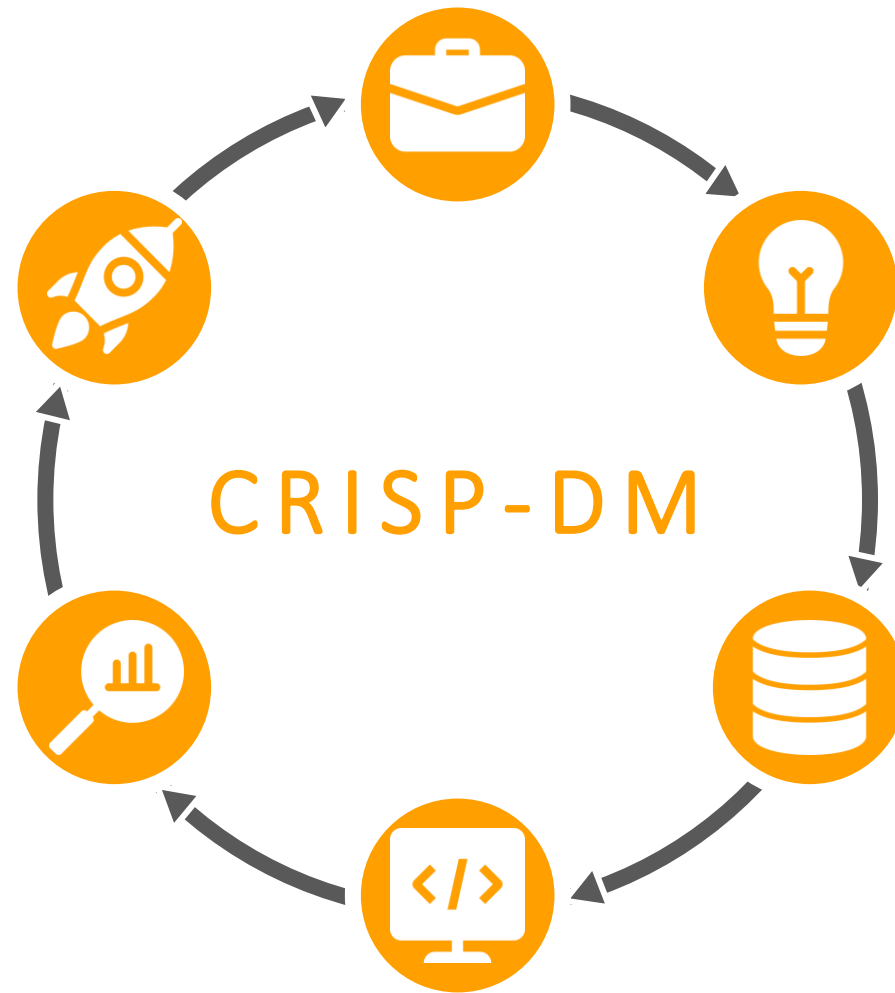
objetivo

desenvolver e comparar um **modelo multinível** com diferentes outros modelos para a prevenção à fraudes cadastrais em uma estrutura de mercado hierárquica 

IMPLEMENTAÇÃO DE ALGORITMOS DE MACHINE LEARNING

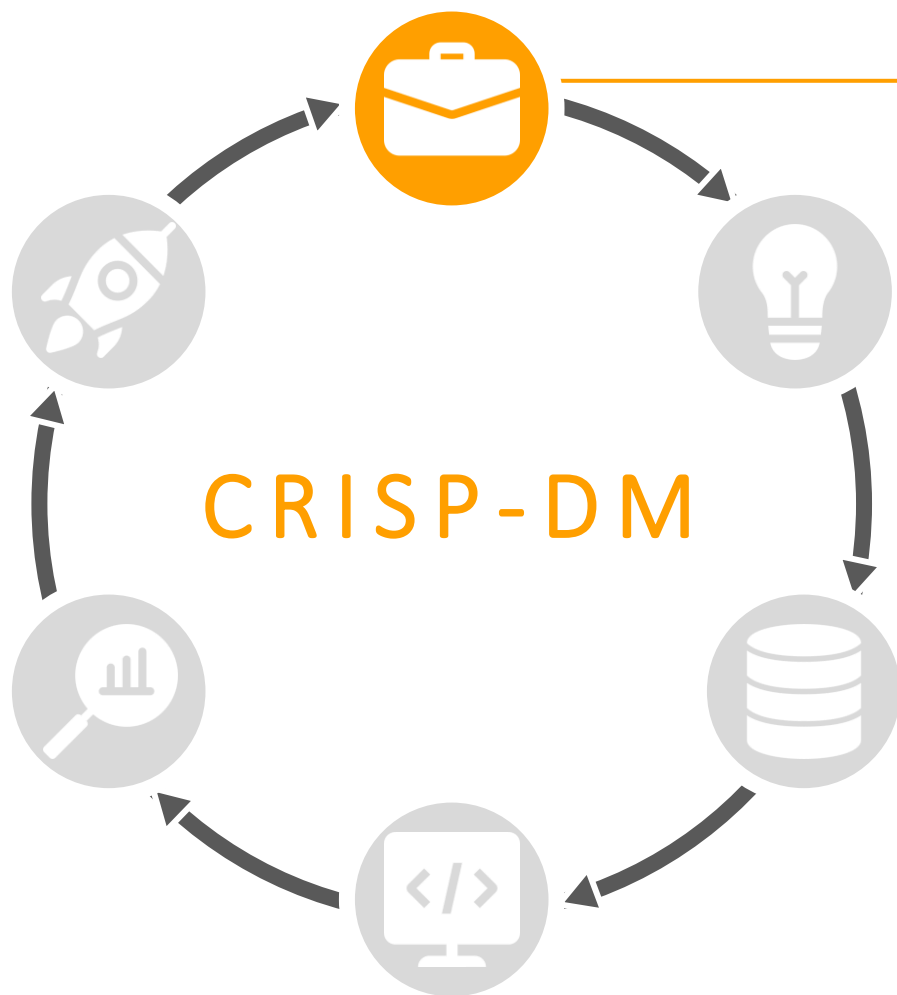


IMPLEMENTAÇÃO DE ALGORITMOS DE MACHINE LEARNING



(Shearer, 2000)

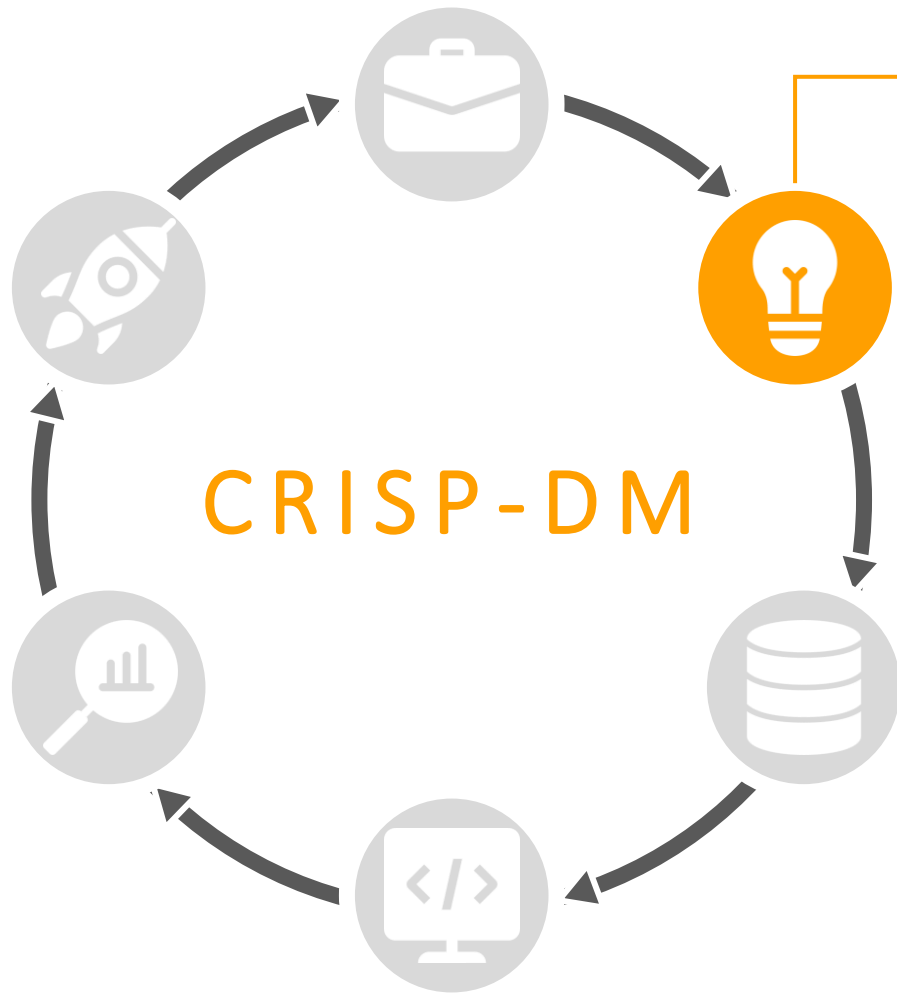
IMPLEMENTAÇÃO DE ALGORITMOS DE MACHINE LEARNING



• PROBLEMA DE NEGÓCIO

encontrar soluções que **acelerem** o processo de abertura de contas, **reduzindo custos** e minimizando a exposição da companhia às **fraudes** financeiras.

IMPLEMENTAÇÃO DE ALGORITMOS DE MACHINE LEARNING



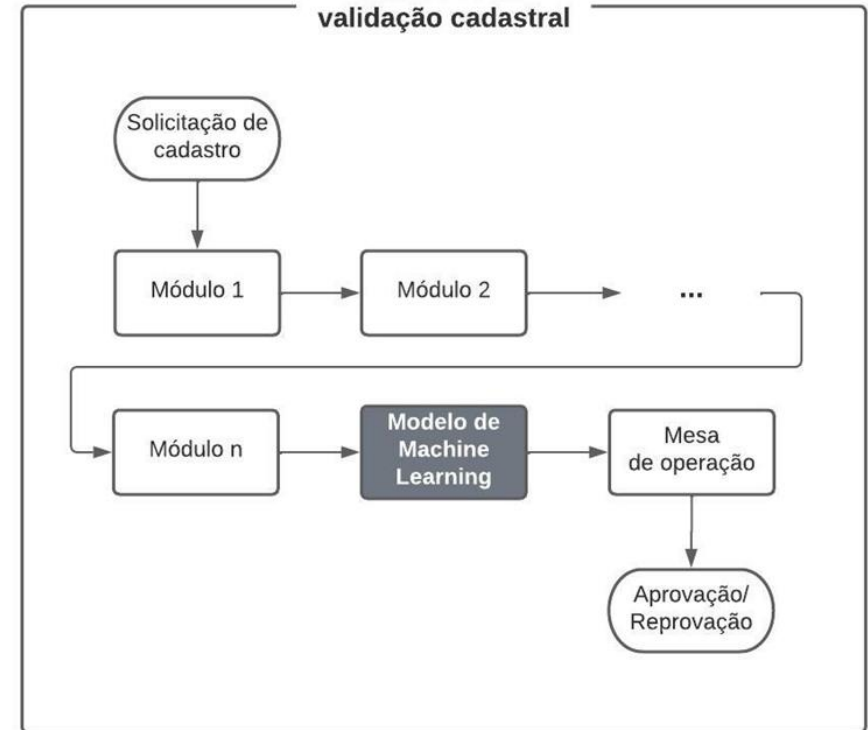
ENTENDIMENTO DOS DADOS

A base de dados é composta por variáveis referentes aos diversos módulos de avaliação.

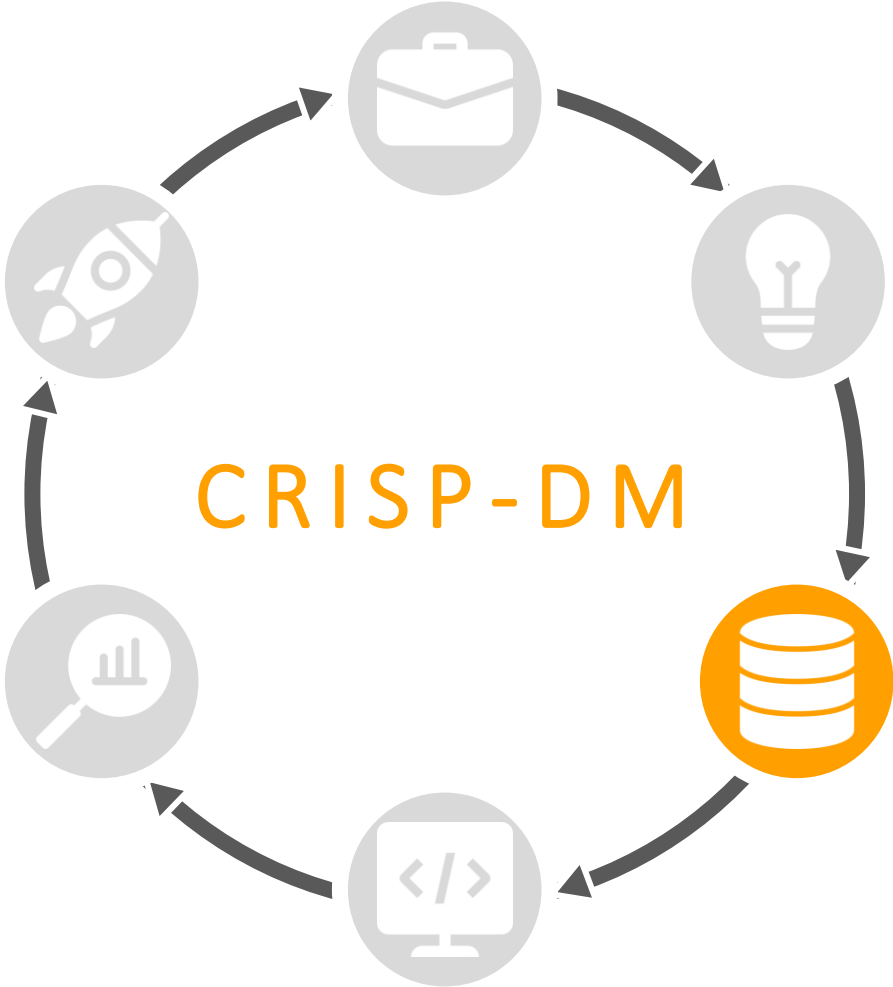
Um dos modelos desenvolvidos, por sua vez, seria implementado na penúltima posição, imediatamente antes da mesa de operação.

Período: 01-08-2022 à 20-09-2022
Volume: 26.434

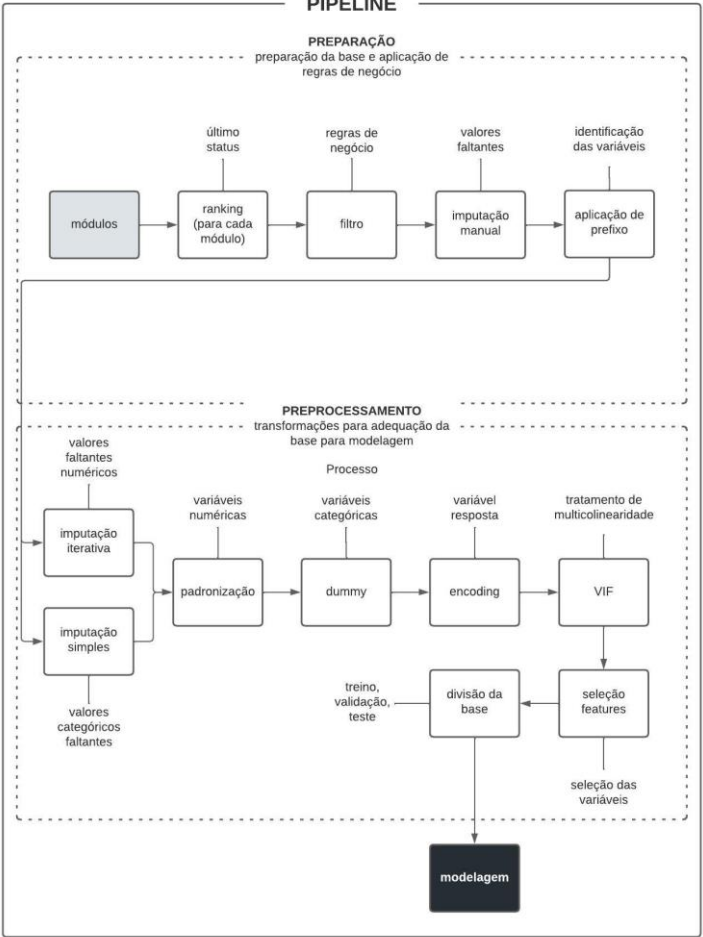
Processo de validação cadastral



IMPLEMENTAÇÃO DE ALGORITMOS DE MACHINE LEARNING



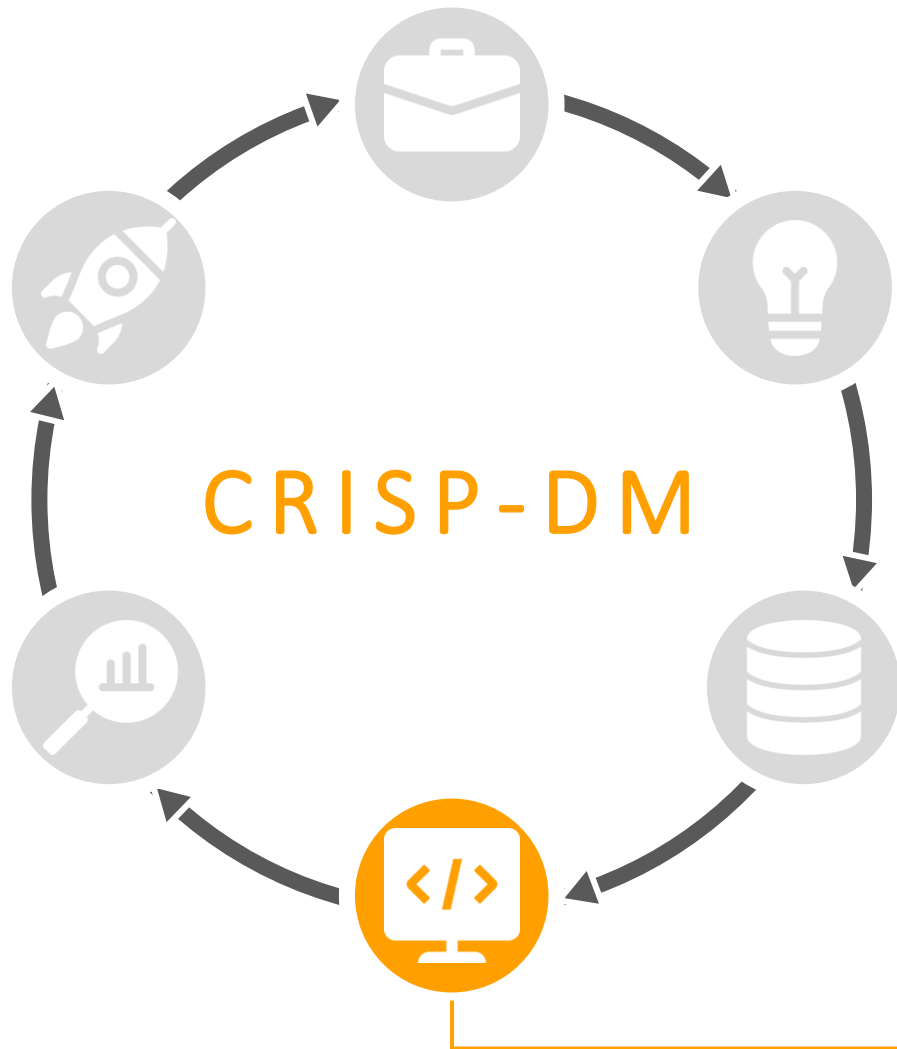
PREPARAÇÃO DOS DADOS



O processo de transformação dos dados foi dividido em duas etapas: preparação e pré-processamento.

Somando-se elas ao modelo, tem-se o “pipeline” completo.

IMPLEMENTAÇÃO DE ALGORITMOS DE MACHINE LEARNING



MODELAGEM



LOGREG

Regressão Logística Binária

Biblioteca: LogisticRegression (scikit-learn)

Justificativa: modelo GLM que não considera os agrupamentos (níveis).
Contraponto ao modelo multinível



XGBOOST

Extreme Gradient Boost

Biblioteca: XGBClassifier (xgboost)

Justificativa: obteve melhor desempenho no experimento de AutoML (Databricks) - AUC ROC



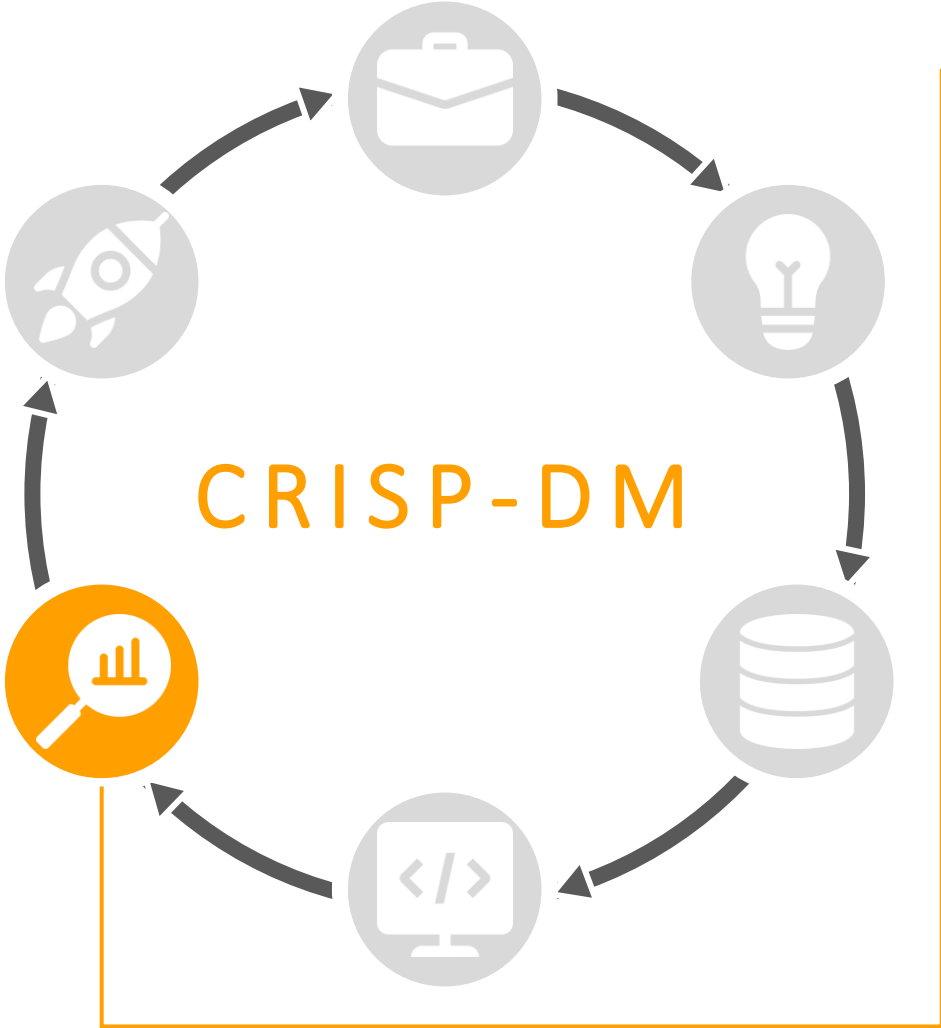
MULTINÍVEL

Tree-boosting + Processo Gaussiano (GP) + Modelos de efeitos Mistos

Biblioteca: GPBoostClassifier (gpboost)

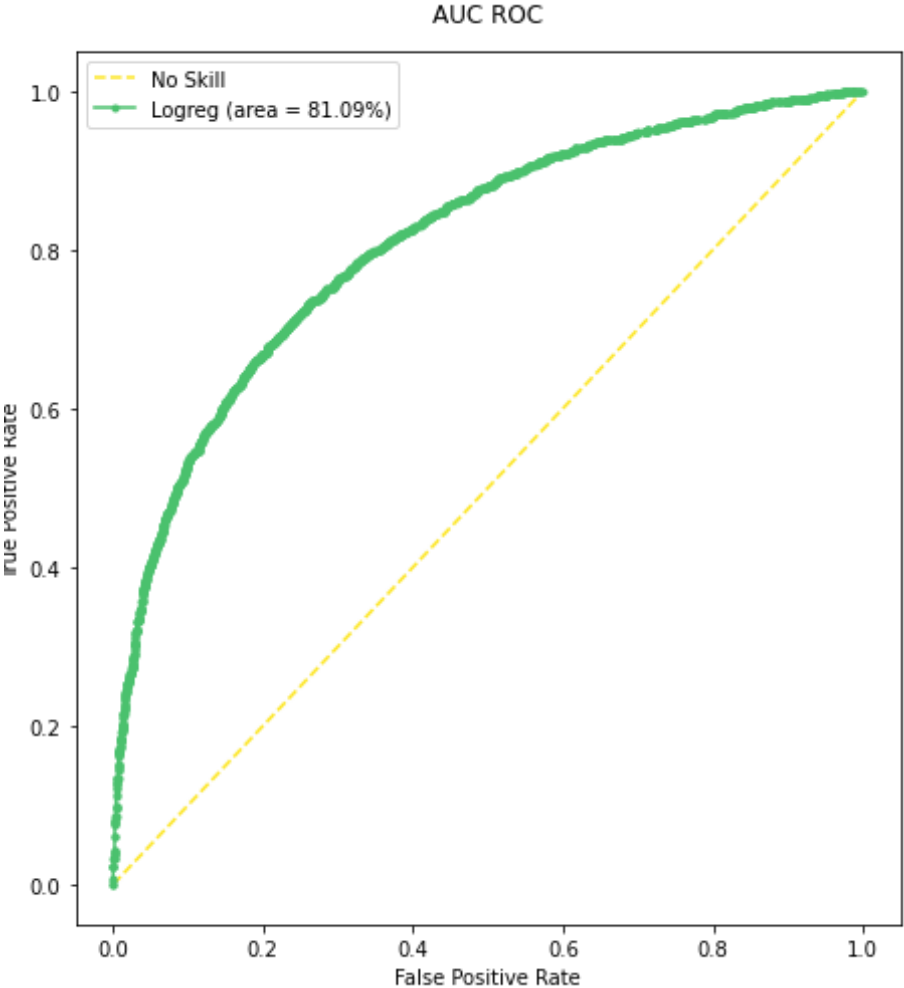
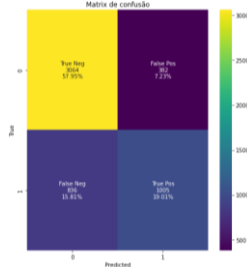
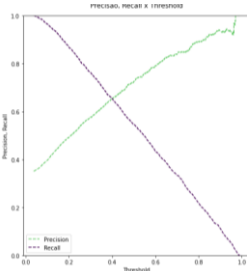
Justificativa: modelo que considera os agrupamentos dos dados observados na realidade

IMPLEMENTAÇÃO DE ALGORITMOS DE MACHINE LEARNING

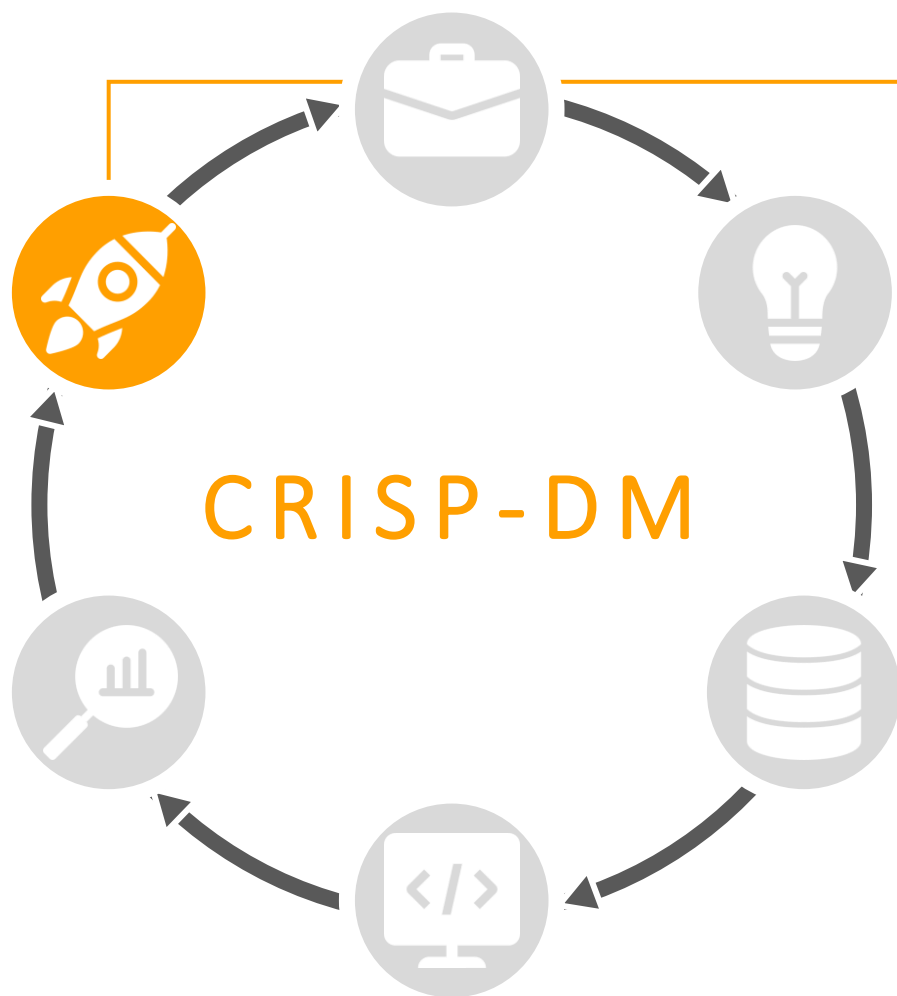


VALIDAÇÃO

A validação dos modelos se deu a partir de diversas métricas de avaliação, sendo a métrica principal eleita para a comparação dos modelos a AUC ROC.



IMPLEMENTAÇÃO DE ALGORITMOS DE MACHINE LEARNING



IMPLEMENTAÇÃO

A etapa de implementação, ou “deploy”, corresponde à construção da esteira de produção (ou “pipeline”) e a implementação propriamente dita do modelo dentro de um processo produtivo, ou de homologação. **Por não ser parte do escopo deste projeto, esta etapa não foi desenvolvida**, porém espera-se que este trabalho possa colaborar com soluções semelhantes na organização estudada e ambientes nos quais os leitores atuem.

RESULTADOS E DISCUSSÃO



RESULTADOS E DISCUSSÃO

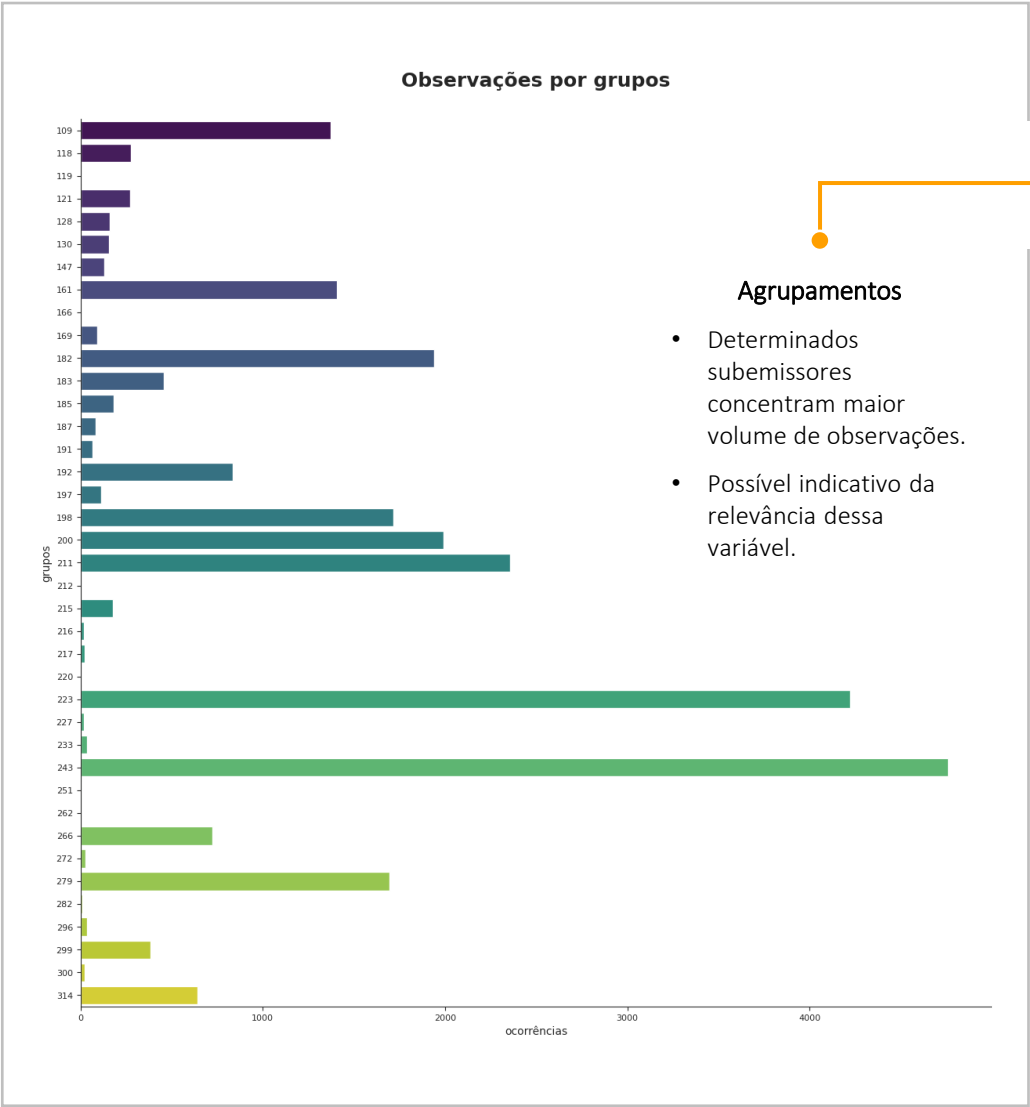


**DESTAQUES DA ANÁLISE
EXPLORATÓRIA (EDA)**



**RESULTADOS
DOS MODELOS**

RESULTADOS E DISCUSSÃO



“Target”

- Reprovação engloba fraudes e inconsistências.
- Nesta etapa do processo, há um menor desbalanceamento da base.

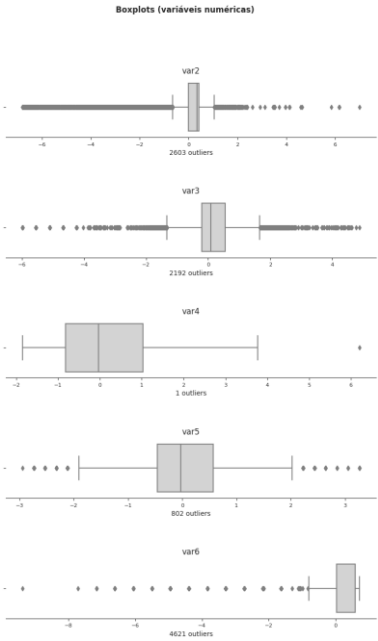
Variável dependente

aprovado, 65%

reprovado, 35%

Outliers

- Podem representar as fraudes (eventos raros), por isso são esperados.



RESULTADOS E DISCUSSÃO

LOGREG

- 3º lugar -

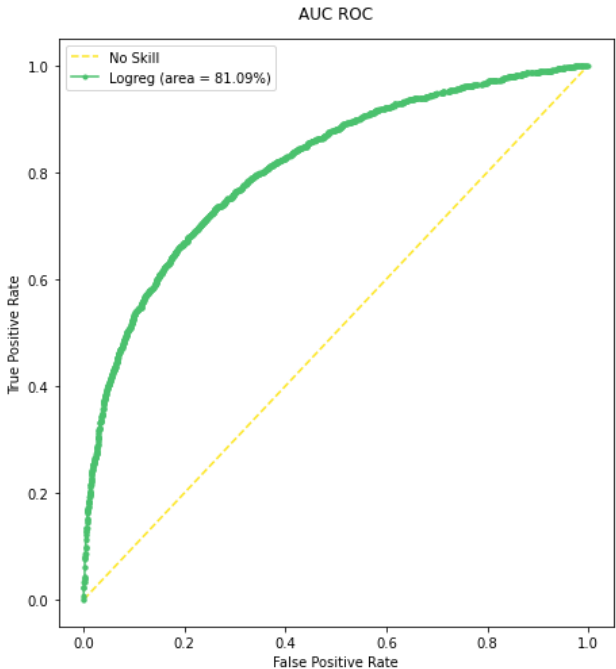
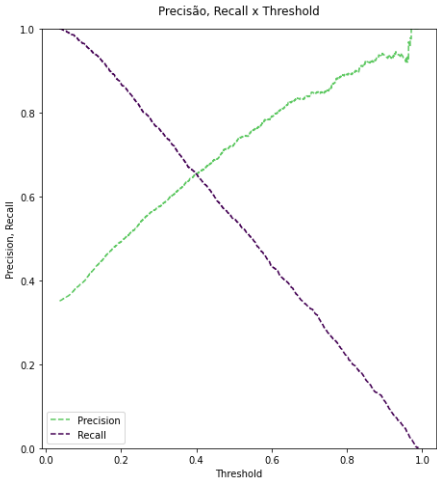
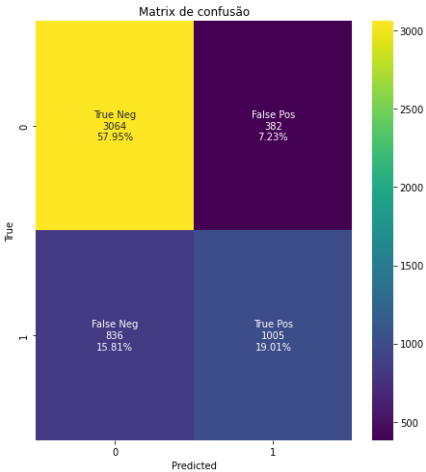
Menor AUC ROC

Técnica de otimização de hiperparâmetros: "grid search"

"Stepwise" >> Regularização

Parametrização
{'C': 0.1, 'penalty': 'l2', 'solver': 'lbfgs'}

	amostra	acuracia	precisao	recall	auc_roc	f1_score	logloss(normalizada)
0	teste	76.30	71.41	54.67	80.25	61.93	0.502163
1	validacao	76.96	72.46	54.59	81.09	62.27	0.492416



RESULTADOS E DISCUSSÃO

XGBOOST

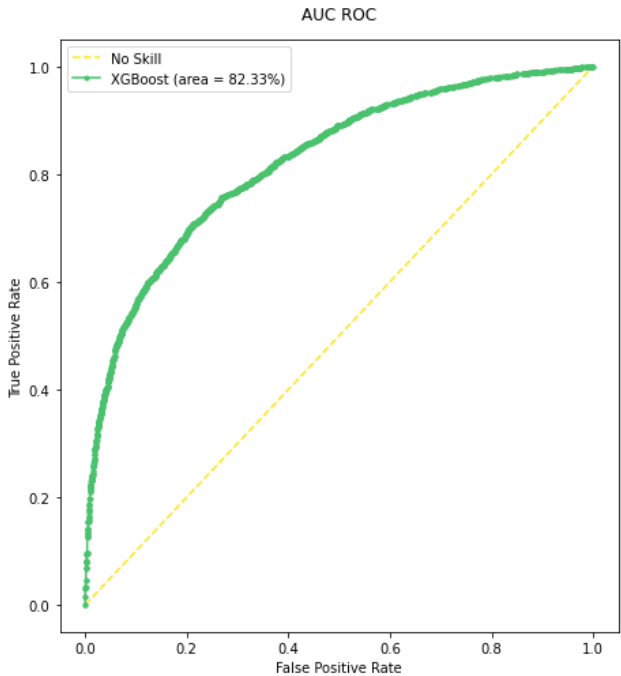
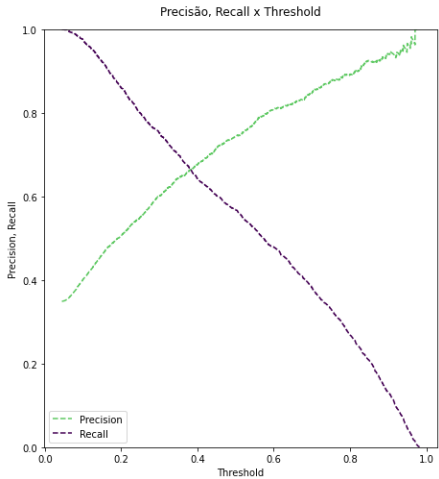
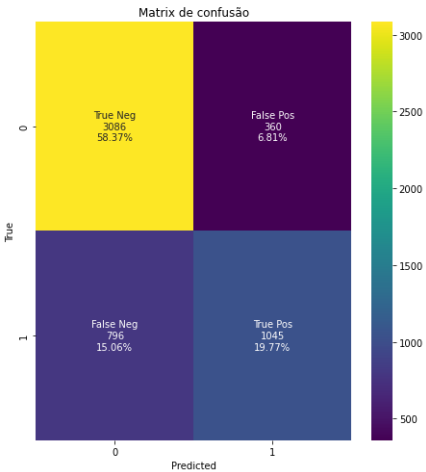
- 2º lugar -

Maior precisão

Experimento AutoML

Parametrização:
colsample_bytree=0.53
learning_rate=0.12
max_depth=6
min_child_weight=1
n_estimators=52
n_jobs=100
subsample=0.716
verbosity=0
random_state=982268122

	amostra	acuracia	precisao	recall	auc_roc	f1_score	logloss(normalizada)
0	teste	77.57	73.28	57.24	81.41	64.28	0.488186
1	validacao	78.14	74.38	56.76	82.33	64.39	0.477099



RESULTADOS E DISCUSSÃO



MULTINÍVEL

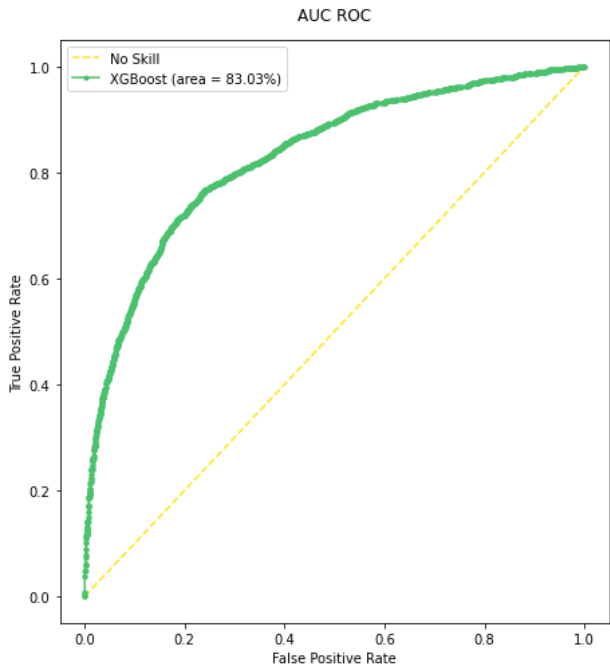
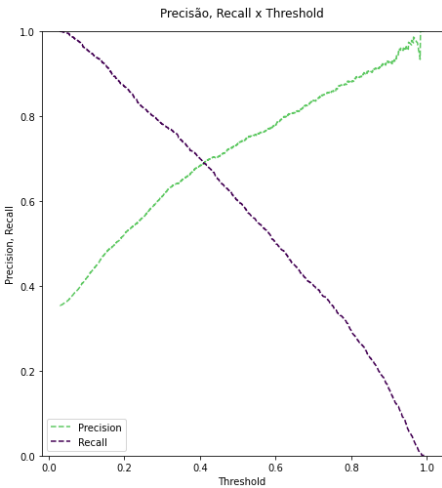
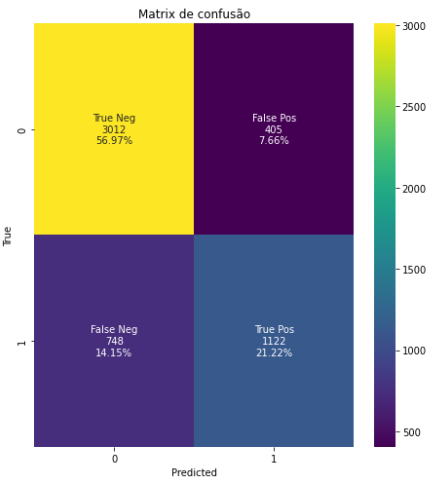
- 1º lugar -

Maior AUC ROC e menor “log loss”

Técnica de otimização de hiperparâmetros: “grid search” & “cross validation” (nº de iterações)

Parametrização:
{'learning_rate': 0.1, 'min_data_in_leaf': 1, 'max_depth': 3}

	amostra	acuracia	precisao	recall	auc_roc	f1_score	logloss(normalizada)
0	teste	76.30	71.11	55.02	80.27	62.04	0.505169
1	validacao	78.19	73.48	60.00	83.03	66.06	0.473676






Teste Delong (1988)

Para comparar as áreas abaixo das curvas ROCs e verificar se os valores encontrados são estatisticamente significantes, foi realizado um teste conhecido como Teste Delong.

Resultado

Para um nível de confiança superior a 95%, pode-se afirmar que os valores encontrados para as AUC ROCs são estatisticamente diferentes entre si, portanto, com mesmo nível de confiança é possível afirmar que o modelo multinível mostrou-se superior aos demais.



	 logreg x xgboost	 logreg x gpboost	 gpboost x xgboost
p_value	0.000003	3.984302e-198	3.155797e-216

CONCLUSÕES



CONCLUSÕES

CONCLUSÃO

O modelo **multinível** mostrou-se não somente uma solução viável, mas aquela com **melhor desempenho** dentre as criadas.



**Ausência de
variáveis no
nível grupo**

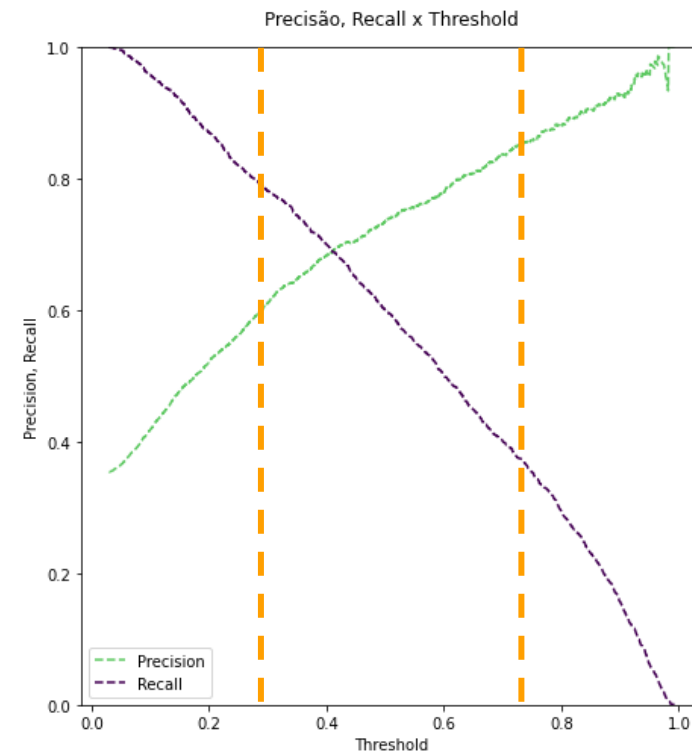
LIMITAÇÕES

Por limitação na disponibilidade dos dados, foram consideradas na análise somente as variáveis preditoras dos módulos de avaliação de cada processo e a variável grupo. Porém, acredita-se que a **inclusão de variáveis referentes aos subemissores**, ou seja, que caracterizem os comportamentos compartilhados dos grupos, podem favorecer a captura dos efeitos aleatórios pelo modelo multinível e, conseqüentemente, sua performance.

CONCLUSÕES

PROPOSTAS FUTURAS

- 1 estudo de pontos de corte.
- 2 Inclusão de variáveis preditoras dos níveis processo (1) e grupo (2).



MBA USP ESALQ

Data Science & Analytics

**Modelo multinível e o efeito empresa para a
detecção de fraudes: uma comparação com
modelos Logístico e XGBoost**

Samuel Haddad Simões Machado
Prof. Dr. Francisco Lledo dos Santos

