PUC Minas Virtual
educação sem distância

PREVISÃO DE RISCO DE INADIMPLÊNCIA COM AUXÍLIO DE DADOS SOCIOECONÔMICOS

ROBERTO TEIXEIRA DE OLIVEIRA

PÓS GRADUAÇÃO LATU SENSU EM CIÊNCIA DE DADOS E BIGDATA

Ordem da apresentação

- Contextualização do problema
- Objetivo do estudo
- Coleta das informações e enriquecimento
- Insights com a exploratória
- Ajuste dos modelos
- Avaliação das métricas

Por que análise de crédito?

“Mesmo com alta de juros, empréstimos cresceram 20% em 2021” (Fonte: Veja)

“Endividamento de famílias atinge nível recorde em março de 2022, diz CNC” (Fonte: AgênciaBrasil)

Objetivo do trabalho

Data Challenge Nubank - 2018

"Ambev, Nubank e Udacity lançam desafio para profissão quente no Brasil" Fonte: Exame

Objetivo: Encontrar uma modelo que seja capaz de predizer se um cliente se tornará inadimplente

Coleta das informações

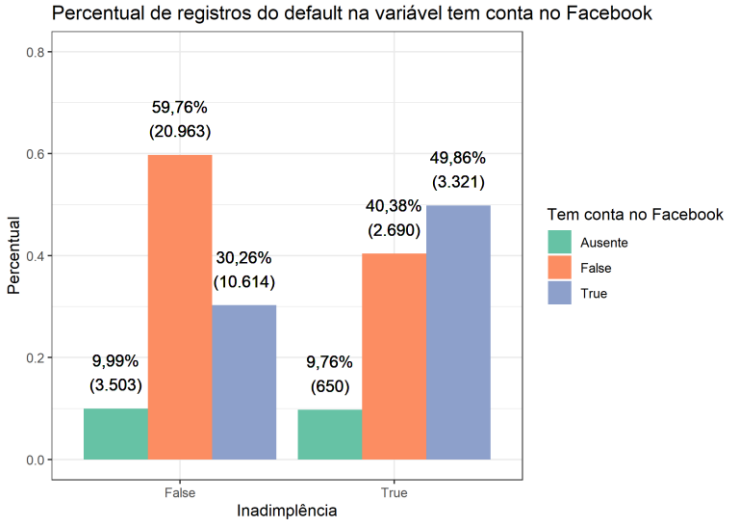
Dados de Aquisição

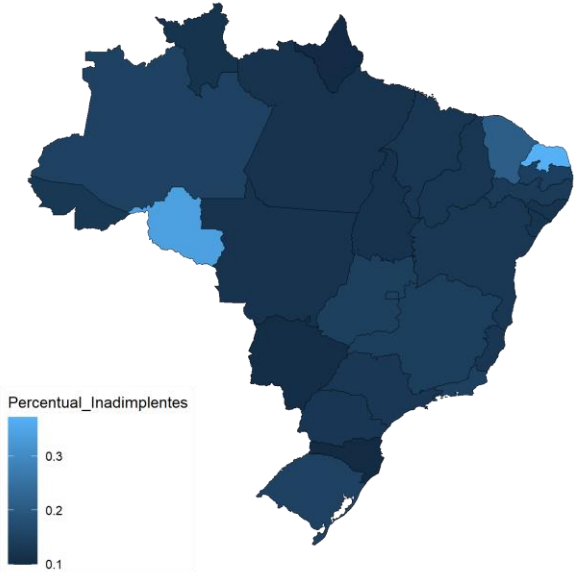
Dados disponibilizados pelo Nubank coletados em repositório público no GitHub

Dados socioeconômicos

Dados referentes ao IDH coletados pelo IBGE e disponibilizados na comunidade Kaggle.

Insights coletados na descritiva



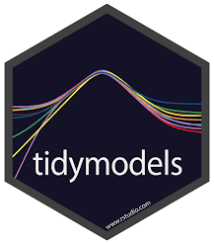


Preparação dos dados

Modelos Ajustados

- Randon Forest
- Support Vector Machine
- XGBoost

Utilizou-se em geral o pacote tidymodels e suas extensões na linguagem R

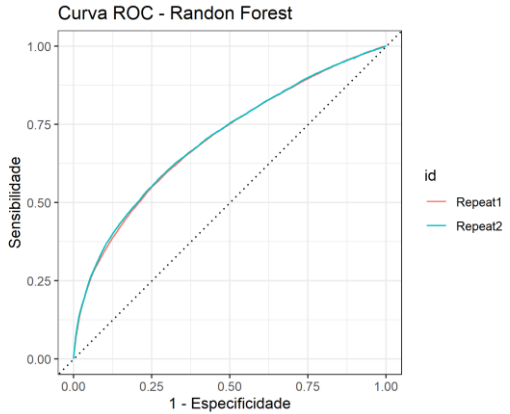


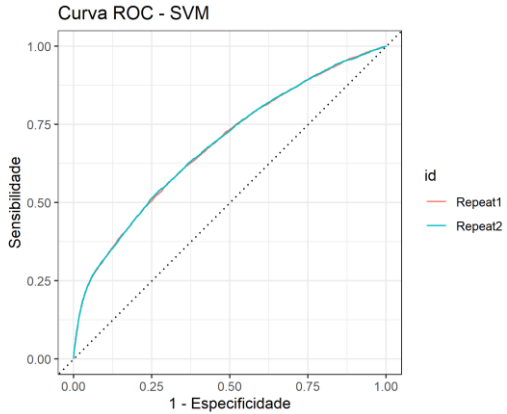
Processo seguidos para todos os modelos:

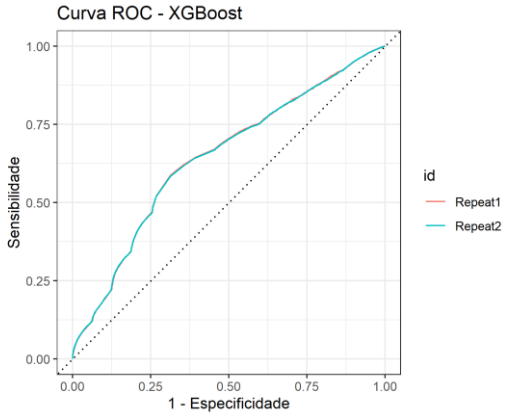
Pre-Process → Train → Validate

Modelos Ajustados

Curva ROC para todos os modelos

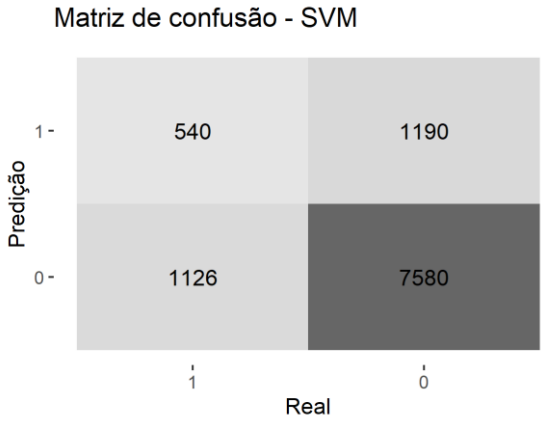


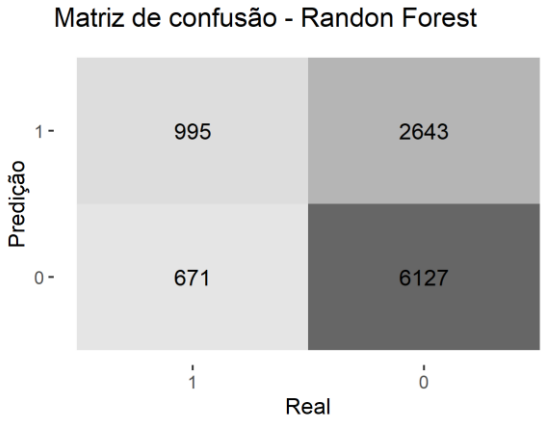


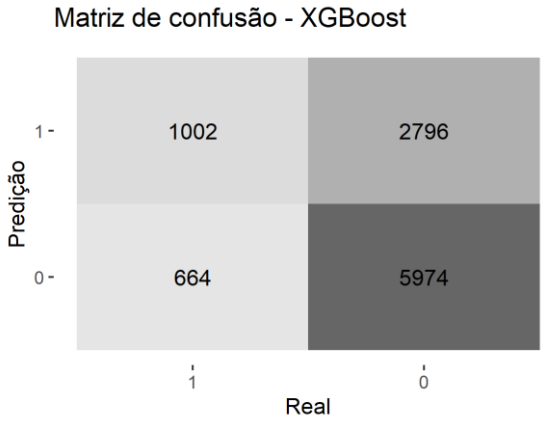


Modelos Ajustados

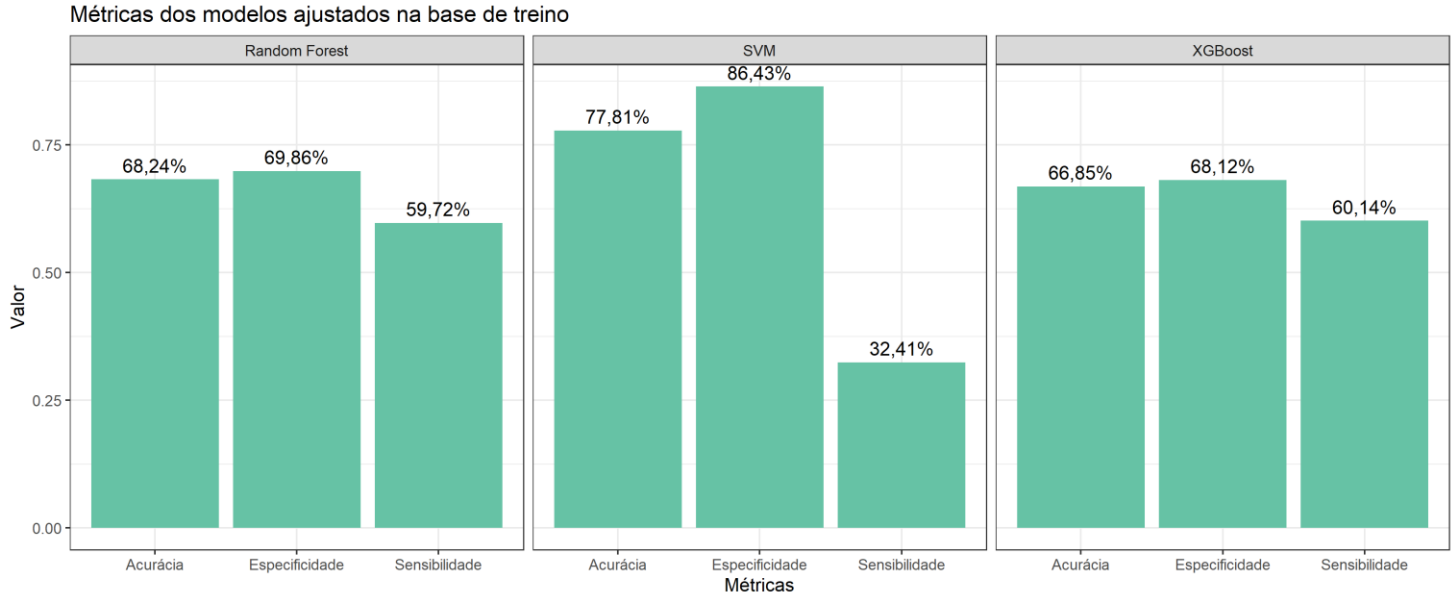
Matriz de confusão para os três modelos no teste



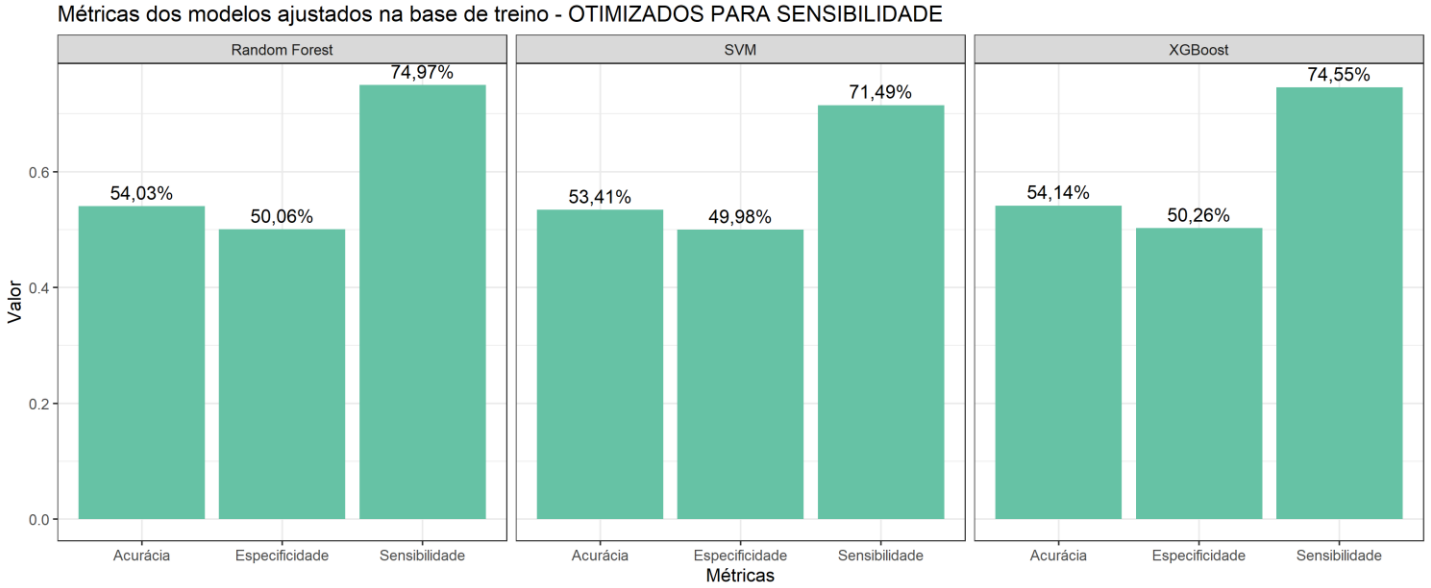




Métricas avaliadas e comparativo



Otimização de threshold



Tipo de threshold	MODELO	Valor threshold	Acurácia	Sensibilidade	Especificidade
NORMAL	Randon Forest	0,5	68,2%	59,7%	69,9%
	SVM	0,5	77,8%	32,4%	86,4%
	XGBoost	0,5	66,8%	60,1%	68,1%
OTIMIZADO PARA SENSIBILIDADE	Randon Forest	0,416	54,0%	75,0%	50,1%
	SVM	0,413	53,4%	71,5%	50,0%
	XGBoost	0,4999863	54,1%	74,5%	50,3%

Obrigado