MODELOS DE MACHINE LEARNING NA CLASSIFICAÇÃO DE POBREZA

UMA APLICAÇÃO PARA O ESTADO DO CEARÁ.

VITOR HUGO MIRO (PPGER/UFC E CAPP/IPECE)

JOÃO MÁRIO SANTOS DE FRANÇA (CAEN/UFC E IPECE)

OBJETIVO GERAL DO ARTIGO

Apresenta uma discussão inicial a respeito da associação de modelagem preditiva, técnicas de *machine learning* e os sistemas de informação disponíveis, sob a hipótese de que essa combinação pode trazer benefícios em termos de aprimoramento das políticas de combate à pobreza, seja no Ceará, como em qualquer outra unidade federativa ou localidade.

COMBATE À POBREZA NO CEARÁ

POBREZA NO CEARÁ

Estimativas com dados da PNADC e linhas de pobreza do BM revelam que, em 2019, **40% da população do estado era classificada como pobre** (3,7 milhões de pessoas) e **12% como extremamente pobre** (1,1 milhão de pessoas).

ESTRATÉGIAS DE COMBATE À POBREZA NO CEARÁ

Em 2019 foram **74 projetos** financiados com o Fundo Estadual de Combate à Pobreza, totalizando um valor aplicado superior a **R$515 milhões.**

PROXY MEANS TEST

$$\boldsymbol{PMT\ score_i = \sum_j x_{ij}\gamma_j}$$

Ponderando cada característica j de um domicílio/ indivíduo i por pesos γ_j, o PMT permite o cálculo de um escore.

PROXY MEANS TEST

Brown, Ravallion e Walle (2016)
Métodos de regressão linear estimados por MQO representam uma das formas mais populares de obter os pesos utilizados no cálculo do escore do PMT. Mas possuem limitações.

McBride e Nichols (2015, 2016 e 2018)
Métodos de *machine learning* podem contribuir significativamente para aprimorar a aplicação da técnica de PMT.

PROXY MEANS TEST

Faria, Silva e Feijó (2007)

Testes de elegibilidade como o PMT representa uma alternativa viável para o atual sistema de seleção de beneficiários de políticas sociais no Brasil.

Brown, Ravallion e Walle (2016)

Grosh (1994) comparou vários programas sociais na América Latina e concluiu que essa classe de métodos (PMT) produziu os melhores resultados de focalização, medidos em termos de redução de erros de inclusão.

PROXY MEANS TEST

Mostafa e Santos (2016)

"... trabalho pretendeu registrar a aplicação modesta de um preditor de renda (PMT) como instrumento de verificação *a posteriori* da renda declarada pelas famílias ao CadÚnico..."

"... este não parece ser um instrumento que tenha uma boa relação custo-benefício".

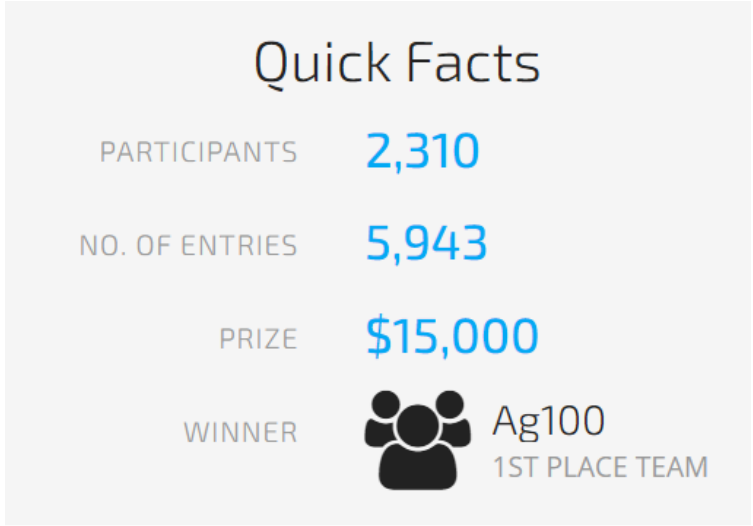
"... este não é um instrumento com acurácia suficiente para ser utilizado na verificação das rendas declaradas ao CadÚnico".

ML NA PREDIÇÃO DE POBREZA

WORLD BANK

DRIVENDATA

"Measuring poverty is hard. Thanks to the efforts of thousands of competitors, The World Bank can now build on open-source machine learning tools to help predict poverty, optimize uses of survey data, and support work to end extreme poverty in the next decade".



PROXY MEANS TEST

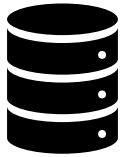
McBride e Nichols (2016)

Compararam os resultados obtidos por modelos de regressão e de *Random Forest* (RF) na aplicação de PMT para predição de pobreza com dados da Bolívia, Timor-Leste e Malawi.
Os métodos de RF melhoraram significativamente o desempenho da previsão “fora da amostra”.

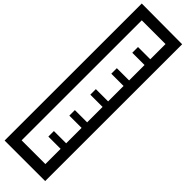
Sohnesen e Stender (2017)

Aplicaram os métodos LASSO e RF para prever a pobreza usando um ano de dados para previsão dentro do mesmo ano e dois anos de dados para prever a pobreza ao longo do tempo. Os resultados apontaram que a aplicação de RF proporcionam estimativas mais robustas e previsões altamente precisas em áreas urbanas e rurais.

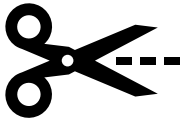
DADOS E VARIÁVEIS



Dados da PNAD Contínua de 2019.



Indicador de pobreza: FGT(0) baseado na renda domiciliar *per capita*.



Linha de pobreza de US$5,50 / dia PPC (aprox. R$436/mês).

POBREZA

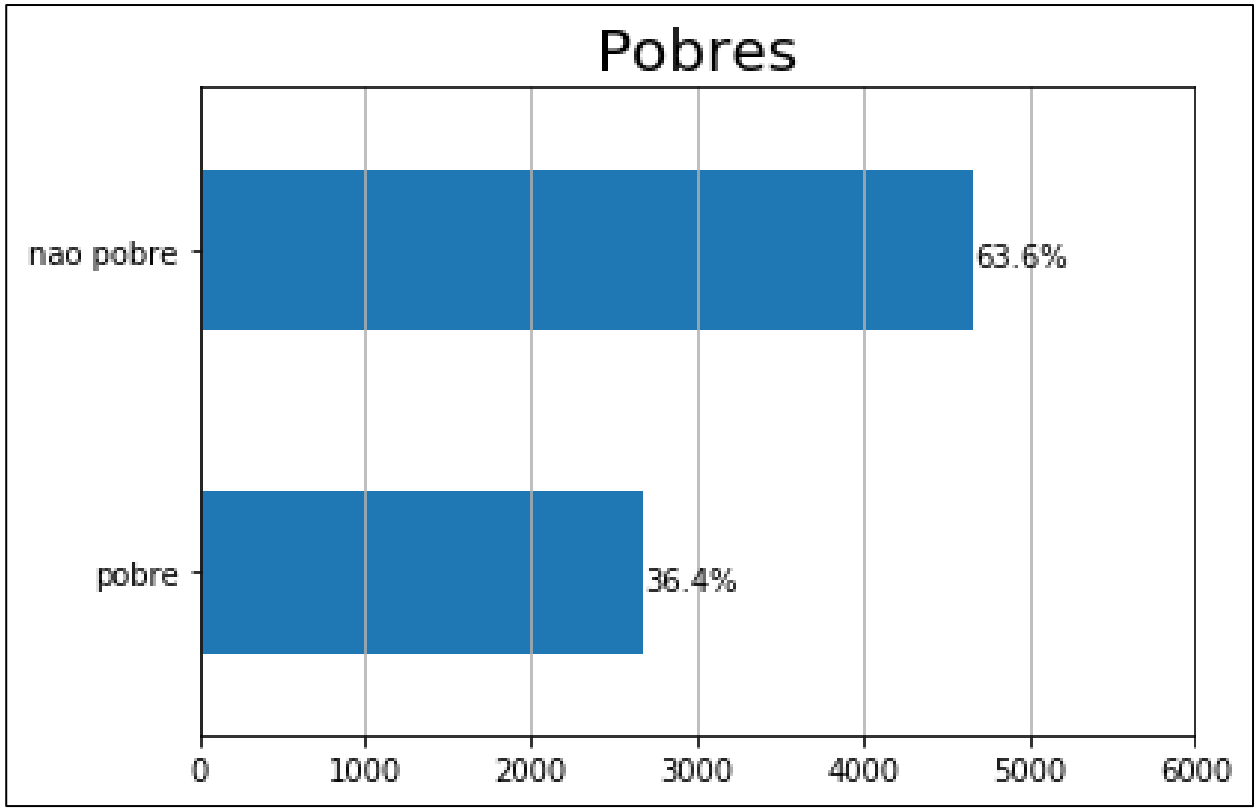
Estimativas populacionais para o CE (2019).

Pobres: 40% (3,7 de 9,16 milhões de pessoas)

Amostra com 7.515 observações.

Pobres: 36,4% (2.735 obs)

Quantidade e proporção de pobres na amostra.



Fonte: Elaboração própria. Microdados da PNAD contínua (2019).
Estimação com Python e biblioteca Pandas e Matplotlib.

TREINO X TESTE

TREINO
5.511 observações (75%)

TESTE
1.837 observações (25%)

DADOS E VARIÁVEIS

Características do chefe de domicílio/ família	Características do domicílio
• Sexo • Idade • Cor • Cônjuge • Escolaridade • Situação ocupacional (atividade, ocupação, aposentadoria.	• Localização • Nº de pessoas • Nº de crianças • Abastecimento de água • Esgotamento sanitário e banheiro • Material da construção • Tipo de propriedade (próprio, alugado, cedido) • Posse de bens (tv, refrigerador, máquina de lavar roupas, celular ou smartphone computador, acesso a internet, automóvel, motocicleta).

ENGENHARIA E SELEÇÃO DE VARIÁVEIS

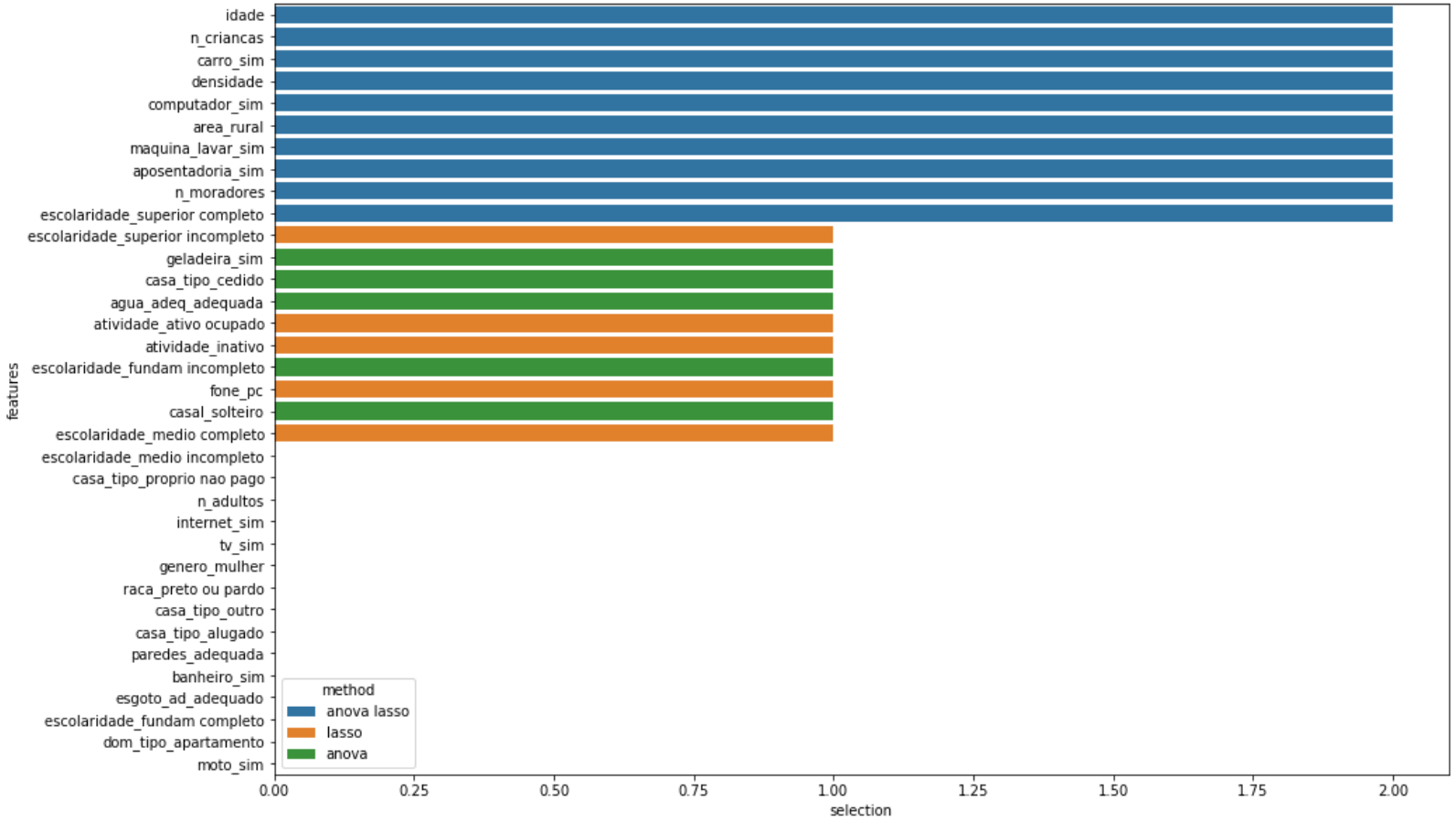
ENGENHARIA DE ATRIBUTOS

- *Data Clearing*
- Variáveis categóricas => *dummies*.
- Variáveis numéricas padronizadas.

SELEÇÃO DE VARIÁVEIS

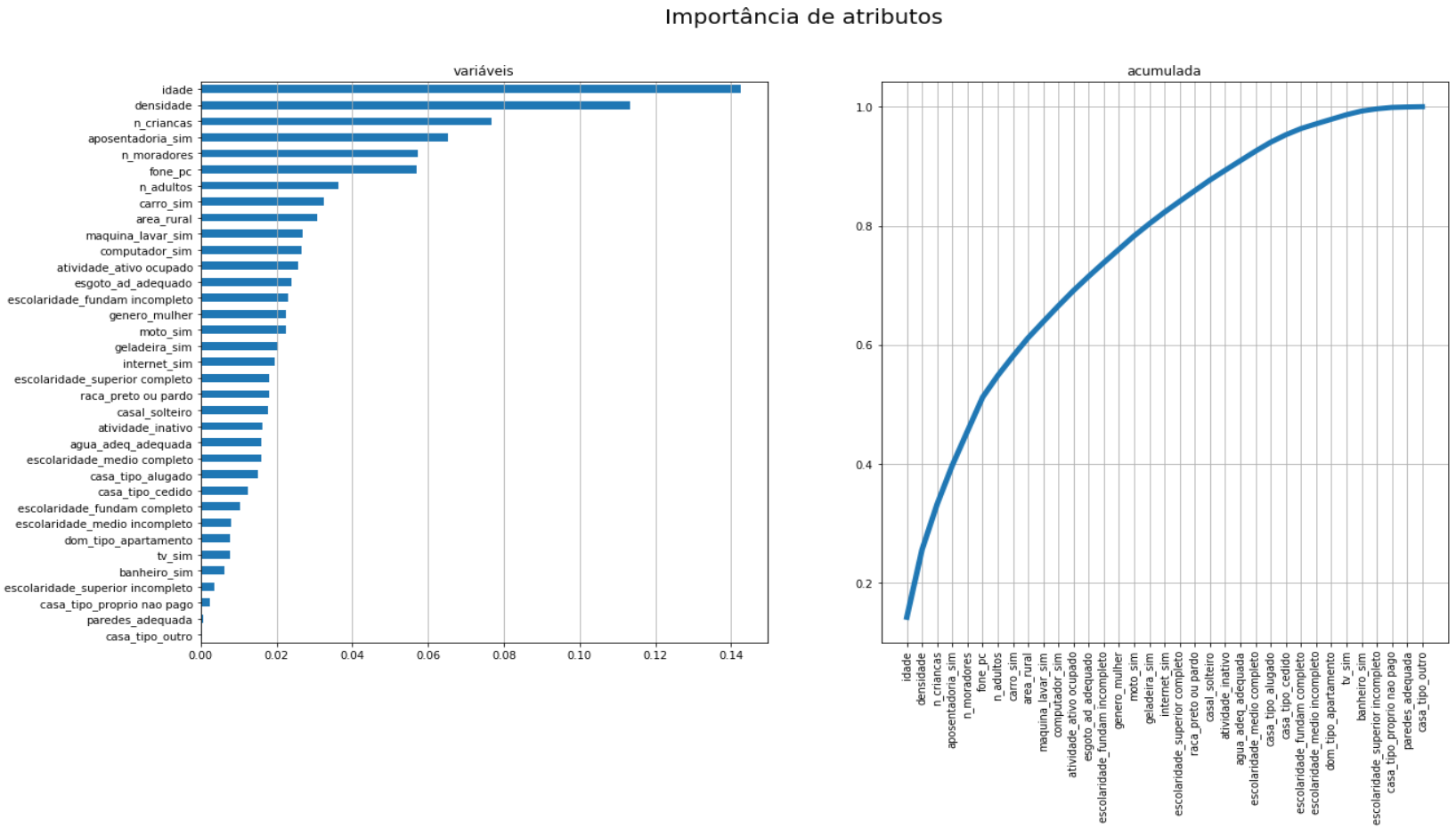
- LASSO (Tibshirani, 1996)
- Análise de variância (ANOVA)
- *Random Forest*

SELEÇÃO DE VARIÁVEIS



Fonte: Elaboração própria. Microdados da PNAD contínua (2019). Estimação com Python, bibliotecas pandas, matplotlib, seaborn, statsmodels e scikit-learn.

SELEÇÃO DE VARIÁVEIS



Fonte: Elaboração própria. Microdados da PNAD contínua (2019). Estimação com Python, bibliotecas pandas, matplotlib, seaborn, statsmodels e scikit-learn.

XGBOOST

Extreme Gradient Boosting desenvolvido por Chen e Guestrin (2016).

- Tem como base um modelo de *Random Forest*
- O algoritmos de *boosting* treina modelos de forma sequencial e recursiva para minimizar erros de classificação.
- Minimiza os erros de previsão aplicando a técnica de gradiente descendente.

AVALIAÇÃO DO MODELO PREDITIVO

MATRIZ DE CONFUSÃO

		Previsto	
		Não pobre ($y = 0$)	Pobre ($y = 1$)
Observado	Não pobre ($y = 0$)	Verdadeiro Negativo (VN)	Falso Positivo (FP)
	Pobre ($y = 1$)	Falso Negativo (FN)	Verdadeiro Positivo (VP)

AVALIAÇÃO DO MODELO PREDITIVO

ACURÁCIA

$$\frac{VP + VN}{Total}$$

PRECISÃO

$$\frac{VP}{VP + FP}$$

SENSIBILIDADE (RECALL)

$$\frac{VP}{VP + FN}$$

F1-SCORE

$$\frac{2 * precisão * recall}{precisão + recall}$$

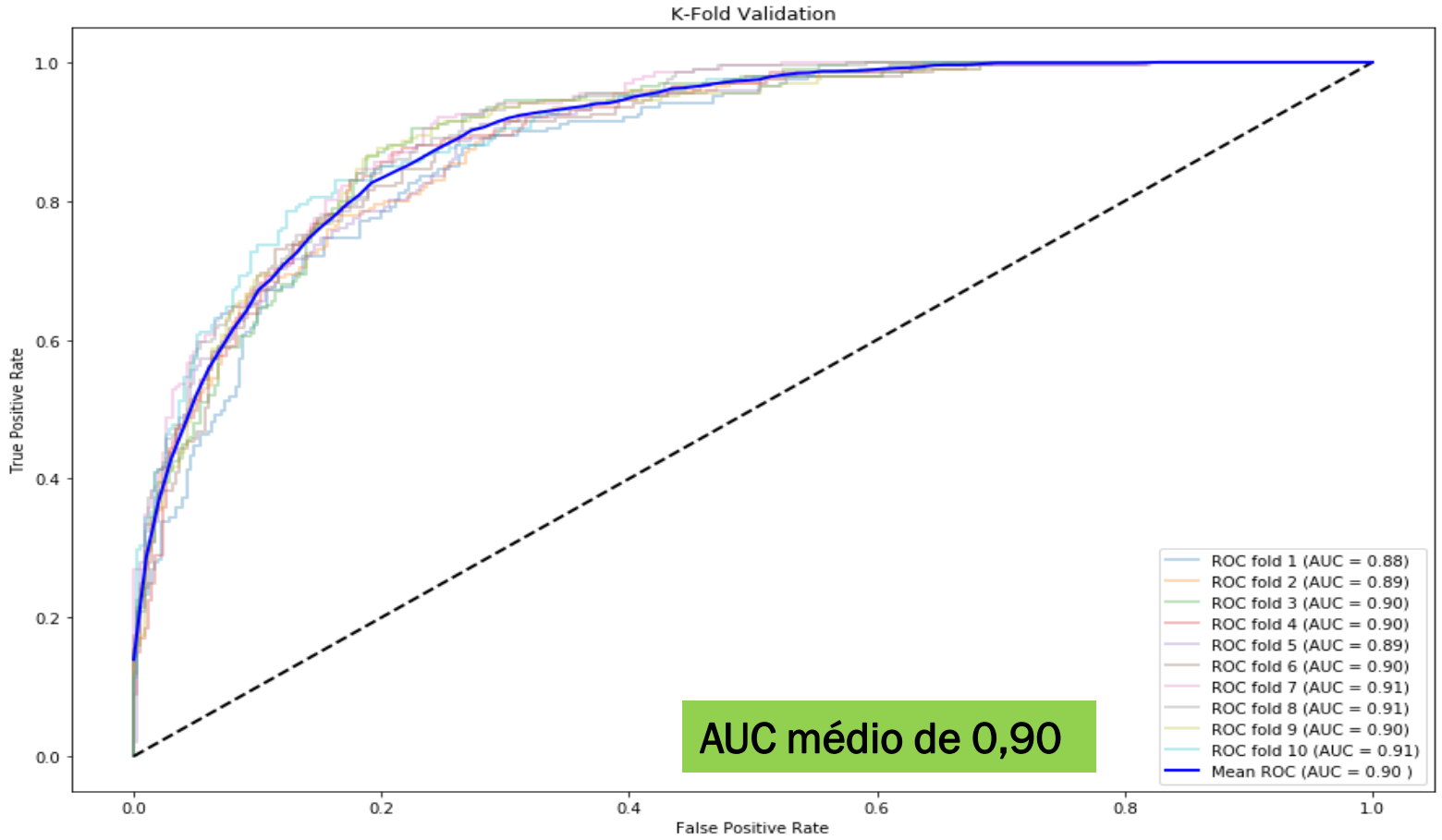
CURVA ROC

Especificidade
X
sensibilidade

AUC

Medida sintética da área sob a curva ROC.

DESEMPENHO DO MODELO

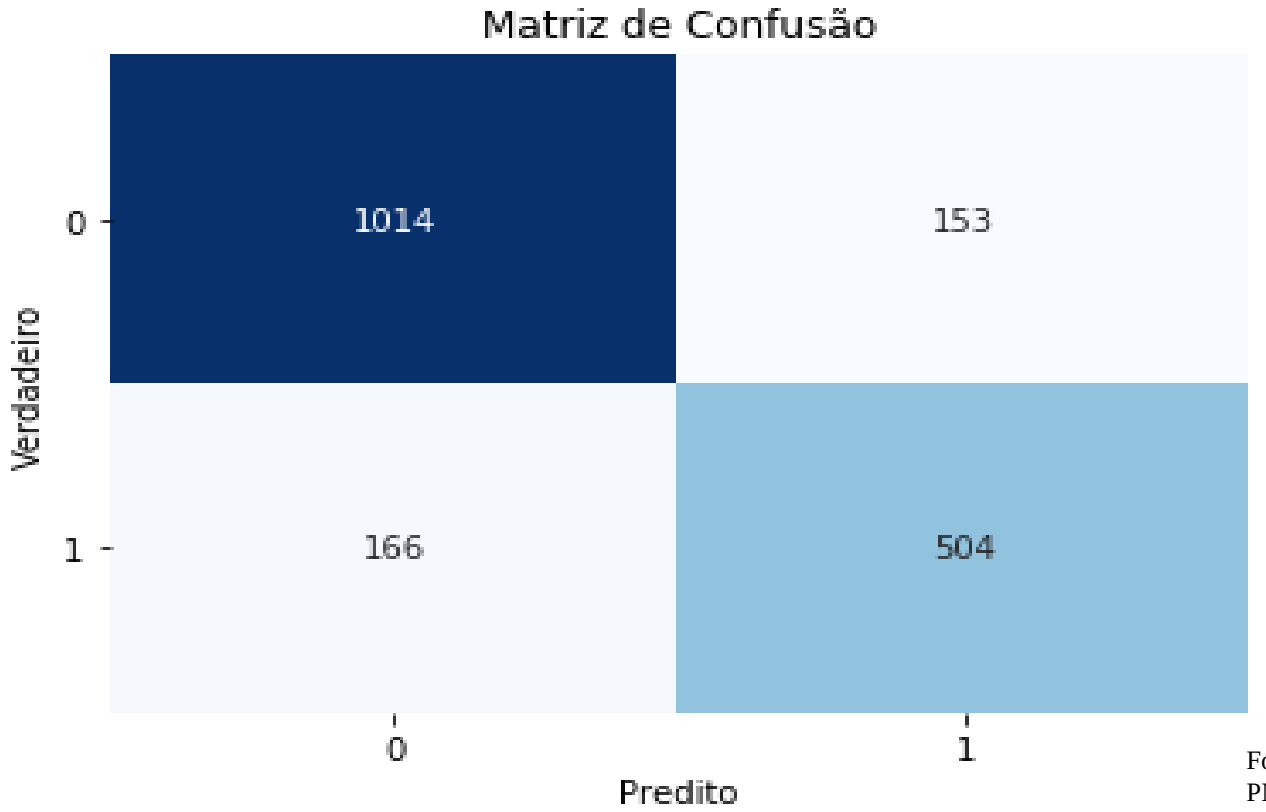


Fonte: Elaboração própria. Microdados da PNAD contínua (2019). Estimação com Python, bibliotecas pandas, matplotlib, seaborn, statsmodels e scikit-learn.

DESEMPENHO DO MODELO

O modelo previu 657 domicílios pobres, dos quais 504 são predições corretas e 153 são falsos positivos

Considerado os 670 domicílios pobres no conjunto de teste, 78% deles foram previstos corretamente.



Fonte: Elaboração própria. Microdados da PNAD contínua (2019). Estimação com Python, bibliotecas pandas, matplotlib, seaborn, statsmodels e scikit-learn.

DESEMPENHO DO MODELO

Campeões do DrivenData (dados do Malawi):

- Acurácias = aprox. 0.88,
- Precisão = aprox. 0.87,
- Recall = aprox. 0.87,
- F1-score = aprox. 0.87,
- Erro de exclusão = 0,12,
- Erro de inclusão = 0,10,
- AUC = aprox. 0,96.

Acurácia Total	0,83
Precisão	0,77
Sensibilidade (recall)	0,78
F1-score	0,76
Erro de exclusão	0,14
Erro de inclusão	0,13
AUC	0,91

Fonte: Elaboração própria. Microdados da PNAD contínua (2019). Estimação com Python, bibliotecas pandas, matplotlib, seaborn, statsmodels e scikit-learn.

CONSIDERAÇÕES FINAIS

Sobre a aplicação de modelos de *Machine Learning* associado ao PMT

- Resultados para um esforço inicial são promissores e sinalizam a viabilidade desse tipo de ferramenta como forma de aprimorar critérios de seleção de beneficiários.
- É possível construir modelos mais elaborados (formatando novas variáveis a partir das variáveis existentes, testando modelos mais avançados de *deep learning*).

CONSIDERAÇÕES FINAIS

De um ponto de vista “prático”

- Verificar aplicação aos dados do Cadúnico.
- Desafios de implementação (éticos, políticos e operacionais).
- Bom potencial como ferramenta complementar de seleção e validação de beneficiários em programas sociais.
- Potencial para análise de focalização e avaliação de programas.

Obrigado!

VITOR HUGO MIRO

vitormiro@ufc.br

AGRADECIMENTOS

FUNCAP

IPECE INSTITUTO DE PESQUISA E ESTRATÉGIA ECONÔMICA DO CEARÁ

CENTRO DE ANÁLISE DE DADOS E AVALIAÇÃO DE POLÍTICAS PÚBLICAS-CAPP

INSTITUTO DE PESQUISA E ESTRATÉGIA ECONÔMICA DO CEARÁ-IPECE