

Desafios na Avaliação de Modelos de Machine Learning para Crédito





Agenda

- Introdução
- O Primeiro Desafio - Classes Desbalanceadas
- Além da Acurácia - Escolhendo outras Métricas
- Treinando um Modelo
- Diagnóstico de Modelo - Curva de Aprendizagem
- Por que o modelo disse não?
- Gráficos SHAP (SHapley Additive exPlanations)
- Confiabilidade das Probabilidades - Curva de Calibração

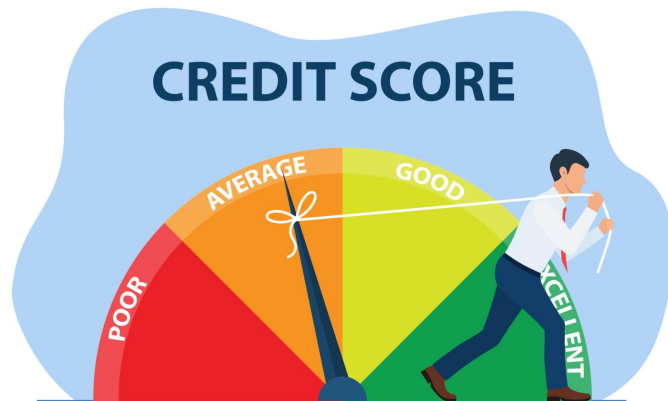
Sobre mim

- Ciências da Natureza - Licenciatura em Física (IFF)
- Mestre em Ciências Naturais (UENF)
- Apaixonado por GNU/Linux
- Movido a Bicicleta Fixa
- Monitor na Capoeira



Introdução

- Contextualize: "Por que ML é crucial para crédito?"
- O custo assimétrico do erro:
Conceder crédito a um cliente ruim (False Negative) é muito mais custoso do que negar crédito a um cliente bom (False Positive).

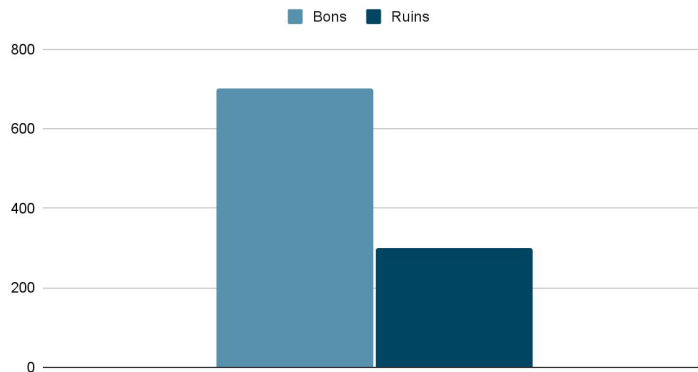




O Primeiro Desafio - Classes Desbalanceadas

- A base está desbalanceada com uma proporção de **70% de bons** e **30% de ruins**.
- Ao utilizarmos a **acurácia**, mesmo dizendo que todos são bons, haverá uma valor de **70%**.
- Necessidade de novas métricas.

Proporção da Base de Crédito

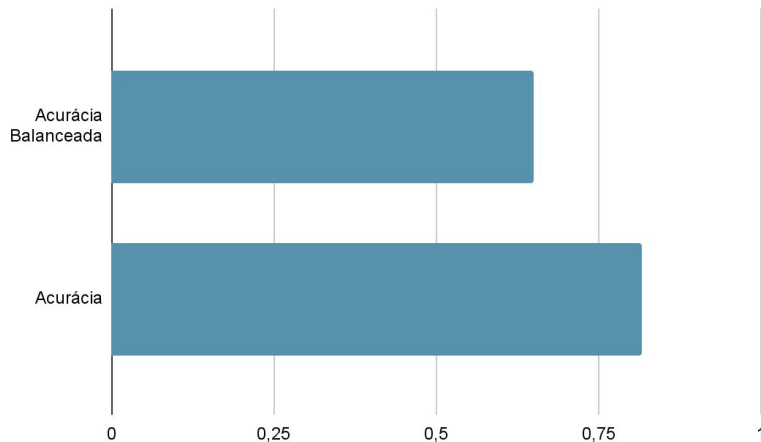




O Primeiro Desafio - Classes Desbalanceadas

- Mudando somente a métrica, há um aumento no valor da acurácia.
- Aumento de 0,07 pontos.
- Por que isso acontece?

Acurácia x Acurácia Balanceada





Além da Acurácia - Escolhendo outras Métricas

- VN - Verdadeiro Negativo
- FN - Falso Negativo
- FP - Falso Positivo
- VP - Verdadeiro Positivo

		Previsão	
		0	1
Verdadeiro	0	VN	FN
	1	FP	VP



Além da Acurácia - Escolhendo outras Métricas

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

A métrica da **acurácia** calcula
todos os pontos da matriz de confusão



Além da Acurácia - Escolhendo as Métricas Certas

$$\frac{1}{2} \left(\frac{VP}{VP + FN} + \frac{VN}{VN + FP} \right)$$

A métrica da **acurácia balanceada** calcula em cima da **taxa de verdadeiro positivo e verdadeiro negativo**.



Além da Acurácia - Escolhendo as Métricas Certas

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

A métrica de **Precisão** é a divisão de **Verdadeiro Positivo** em relação ao **Verdadeiro Positivo** mais o **Falso Positivo**.



Além da Acurácia - Escolhendo as Métricas Certas

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

A métrica é a divisão do **Verdadeiro Positivo** em relação ao **Verdadeiro Positivo** mais o **Falso Negativo**.

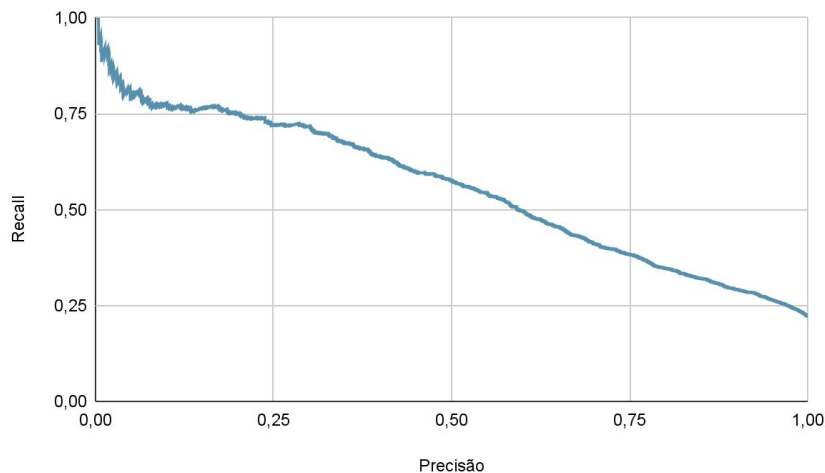


Além da Acurácia - Escolhendo as Métricas Certas

$$2 \times \frac{Precision \times Recall}{Precision + Recall}$$

F1 é média harmônica entre
a Precisão e o Recall

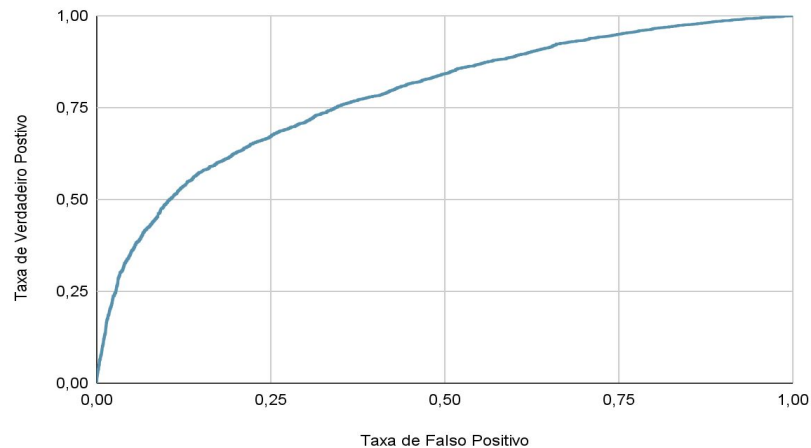
Gráfico Recall vs Precisão





Além da Acurácia - Escolhendo outras Métricas

ROC Curve



Área sob a curva da Taxa de Verdadeiro Positivo e a Taxa de Falso Positivo.



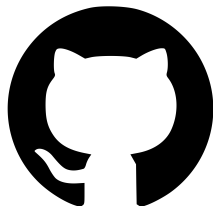
Conclusão e Lições Chave

- Com o **desbalanceamento** da base de dados utilize **precisão, recall, F1** ou **AUC**.
- Para **diagnóstico** utilize a **curva de aprendizagem**.
- Para **explicabilidade** do modelo utilize a **importância de variáveis** e o gráfico **SHAP**.
- **Confiabilidade** cheque a **curva de calibração**.



Perguntas?

Obrigado!



github.com/juniorcl



linkedin.com/in/clebiojunior



medium.com/@juniorcl