

Relatório: Salary Prediction — Regressão Linear vs Regressão Logística

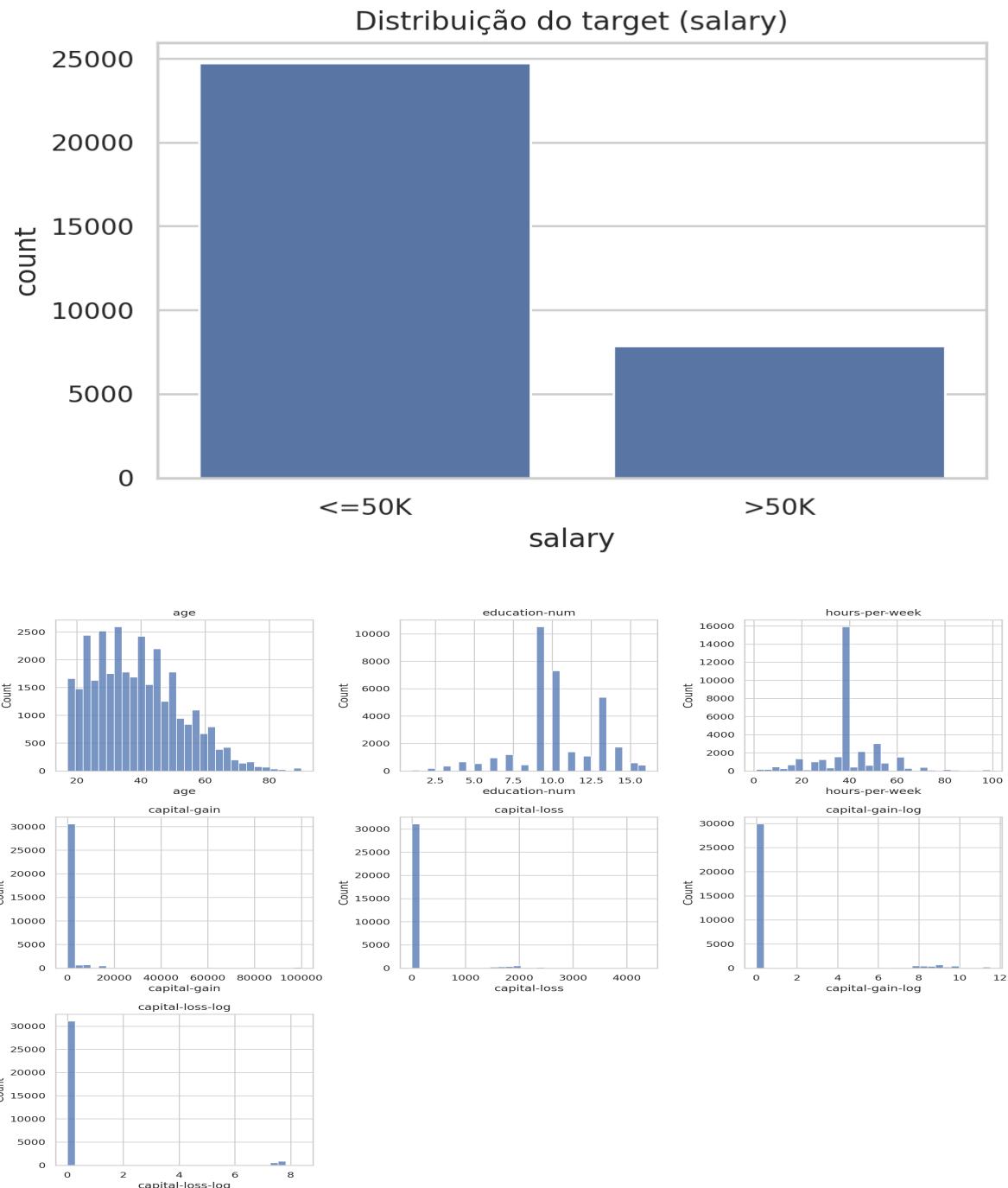
Autor: Luiz Fernando Policarpo Leandro

Dataset: salary.csv

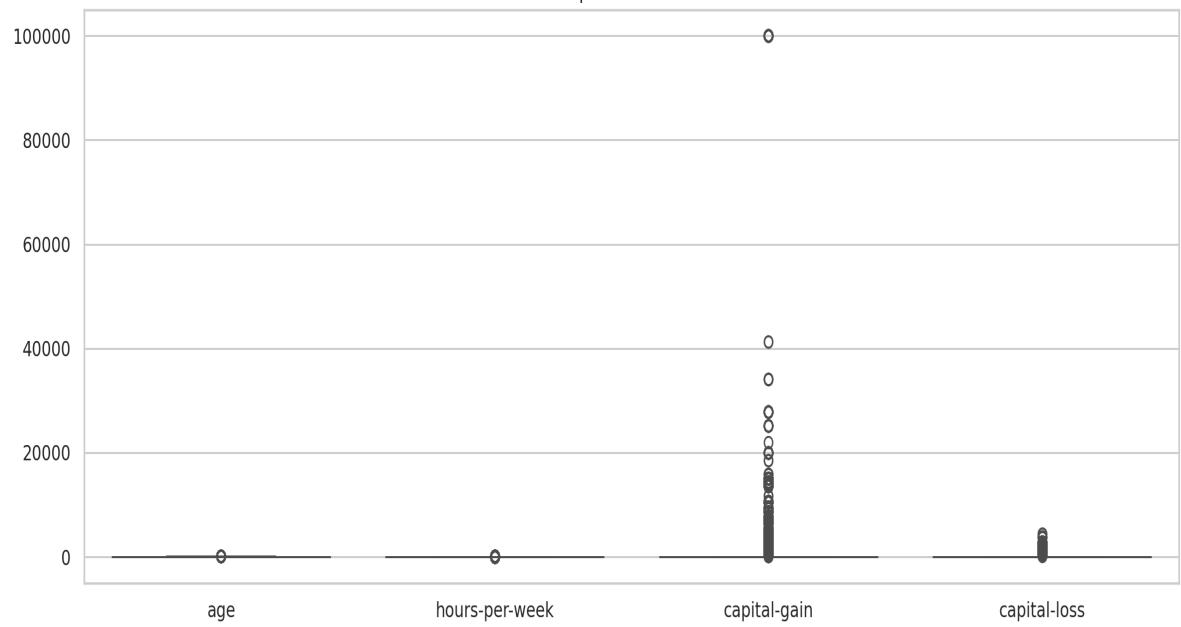
Regressão linear, regressão logística e ensembles — execução automática

1. Análise Exploratória (EDA)

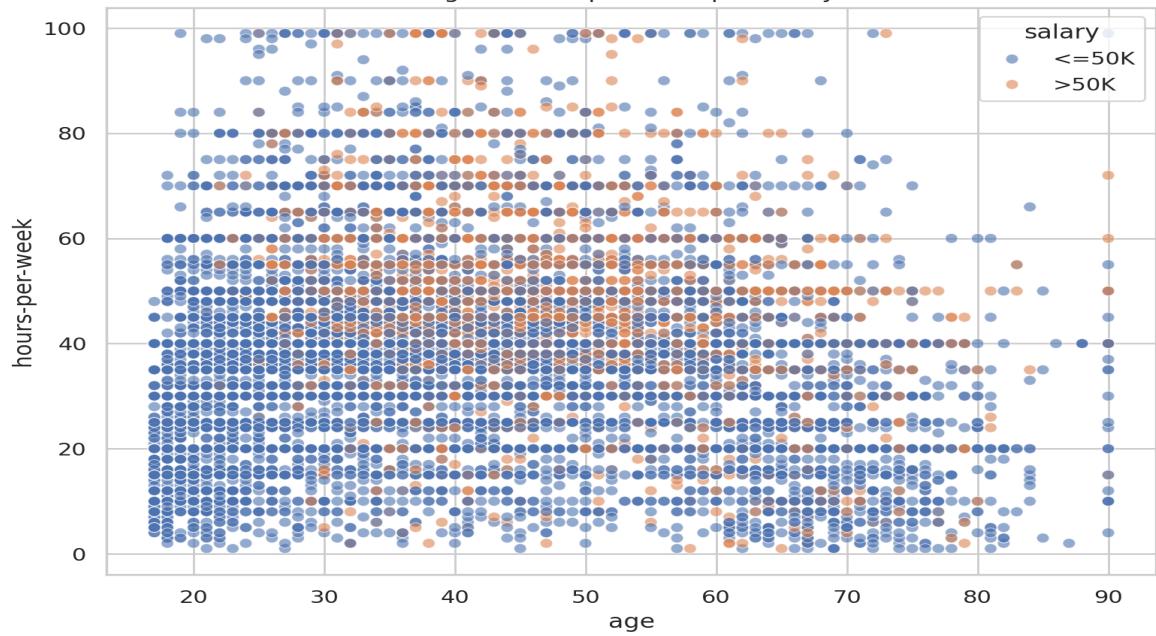
Principais estatísticas e distribuição do target, histogramas e boxplots a seguir.

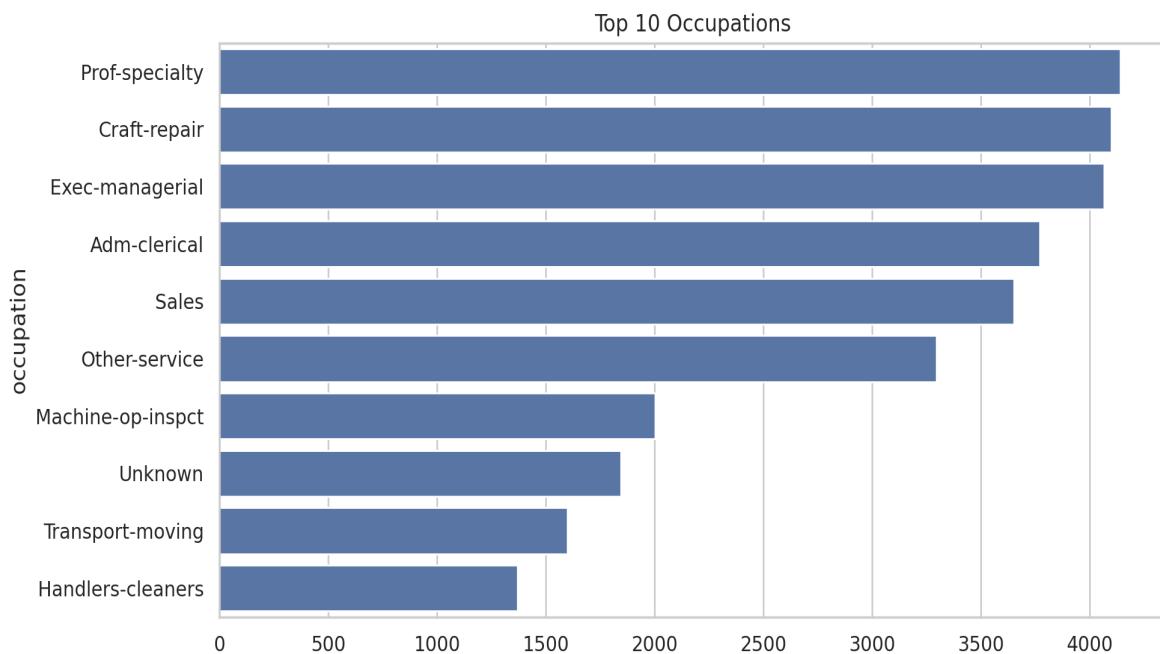


Boxplot - numéricas



Age x Hours-per-week por salary



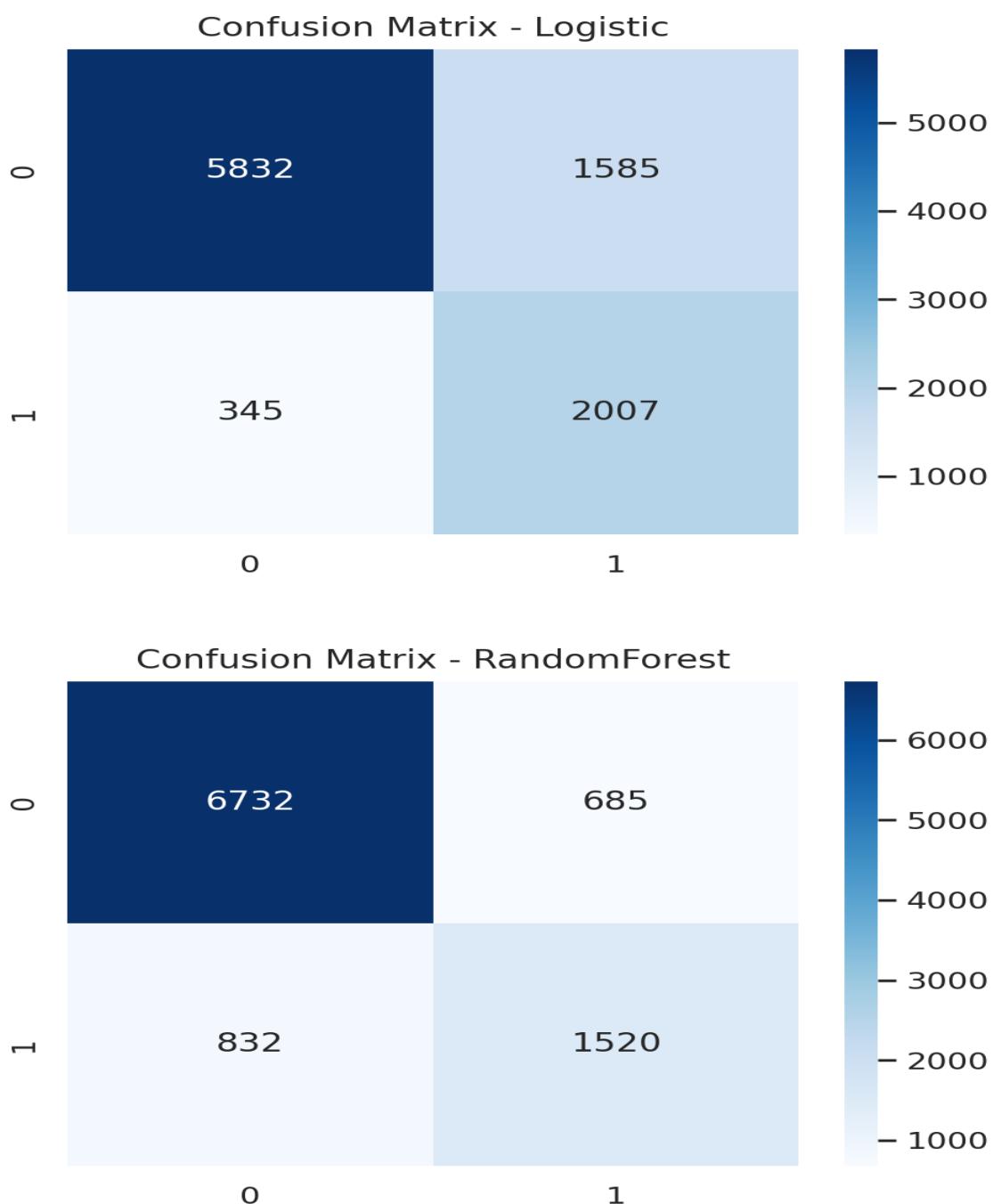


2. Modelagem e Resultados

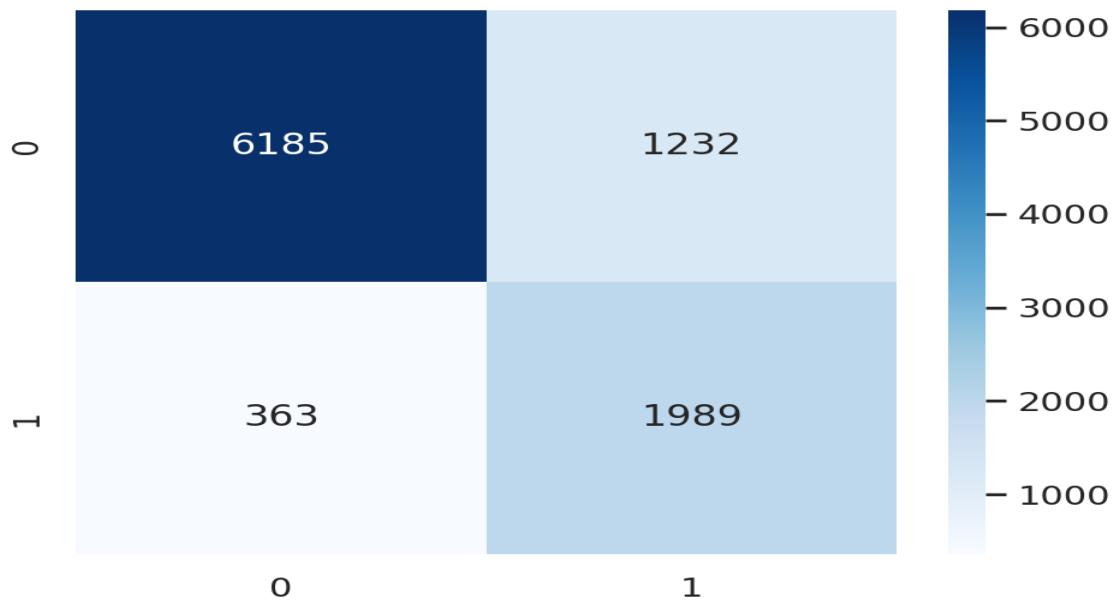
Foram treinados LinearRegression (baseline), LogisticRegression (classificação direta) e RandomForest. XGBoost foi treinado se disponível.

Model	Accuracy	ROC AUC
LinearRegression	0.8440	N/A
LogisticRegression	0.8024	0.9007
RandomForest	0.8447	0.8963
XGBoost	0.8367	0.9251

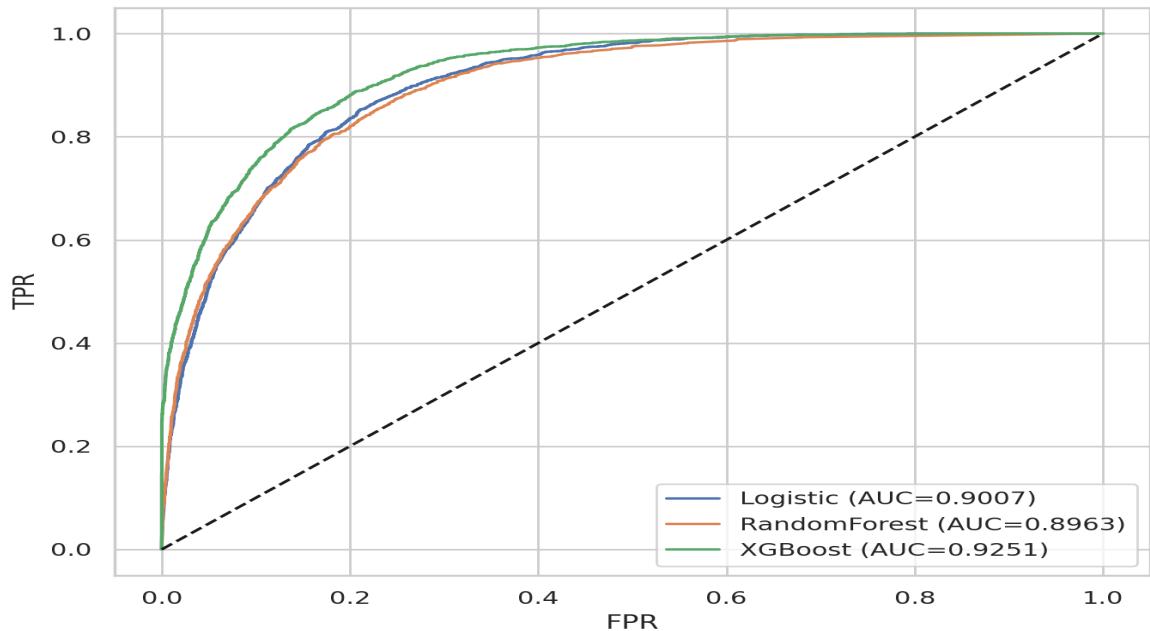
Confusion matrices e ROC curves



Confusion Matrix - XGBoost



ROC Curves



3. Interpretabilidade

Coeficientes de regressão linear e odds ratios da regressão logística, além das principais features segundo RandomForest/XGBoost.

Top 10 coeficientes (Linear)

feature	coef
cat_52	0.1913277640140208
cat_8	-0.1675641803575758
cat_60	0.1634101228546678
cat_33	-0.1581594372801549
cat_101	0.1551890104875058
cat_69	0.1549488963737592
cat_94	-0.1543145009831672
cat_87	-0.1297569566070603
cat_35	0.1285460996987373
cat_74	-0.1260374251768486

Top 10 Odds Ratios (Logistic)

feature	odds_ratio
cat_27	4.39342450170955
cat_26	3.210909589032638
cat_52	3.0815189710884594
cat_60	3.0740456554325486
cat_101	2.4090565563921755
cat_35	2.373749304168005
cat_42	2.2341815823572717
cat_69	2.168538250256717
education-num	2.086899193550373
cat_0	1.9804388517862277

Top 10 Importâncias (RandomForest)

feature	importance
age	0.2144832402944147
hours-per-week	0.1028298914340486
cat_27	0.0882370934293852
capital-gain-log	0.0736250953930945
education-num	0.069700244321187
cat_47	0.0571574803384111
cat_29	0.0359778573872577
capital-loss-log	0.0233446585461655
cat_50	0.0176862370433097

cat_35	0.0165677615140497
--------	--------------------

4. Discussão e Conclusão

Resumo executivo

Dataset e objetivo:

Usamos o dataset 'Salary Prediction' (Adult). Objetivo: comparar regressão linear (previsão contínua) e regressão logística (classificação alto/baixo) e discutir vantagens e limitações.

Principais passos:

- Carregamento e limpeza (tratamento de valores faltantes, transformação log para capital-gain/loss).
- Codificação de categóricas (one-hot) e padronização de numéricas.
- Split treino/teste estratificado 70/30.
- Treinamento: LinearRegression, LogisticRegression, RandomForest; XGBoost quando disponível.

Resultados principais (métricas):

- LinearRegression: accuracy=0.8440
- LogisticRegression: accuracy=0.8024, ROC AUC=0.9007
- RandomForest: accuracy=0.8447, ROC AUC=0.8963
- XGBoost: accuracy=0.8367, ROC AUC=0.9251

Interpretação e recomendação curta:

A regressão linear serve como baseline didático: provê métricas de erro contínuas (MAE/MSE/RMSE) e uma visão de correlação linear entre features e target. Porém, quando tratamos a tarefa como classificação (alto/baixo), a regressão logística é mais adequada por modelar diretamente probabilidades e odds; seus coeficientes podem ser interpretados como odds ratios. Modelos de ensemble (RandomForest/XGBoost) tendem a entregar melhor separação (AUC maior) e recall da classe minoritária quando corretamente ajustados. Para decisão de faixa salarial (>50K), recomendo usar regressão logística ou XGBoost calibrado, priorizando AUC e recall/precision do grupo >50K.

Melhorias sugeridas:

- Tuning de hiperparâmetros (GridSearchCV), validação cruzada.
- Calibração de probabilidades (CalibratedClassifierCV).
- Feature engineering (interações, bins para idade, agrupamento de países).
- Remover multicolinearidade/redução dimensional quando necessário (PCA/VIF).