Explicabilidade de modelos: uma análise de grupos salariais

Isadora Alves de Salles
isadorasalles@dcc.ufmg.br

INTRODUÇÃO

Devido a um grande espectro de fatores históricos e sociais, é sabido que fatores como gênero e grau de escolaridade afetam diretamente o patamar salarial das pessoas. O objetivo deste trabalho portanto, é utilizar o dataset "Adult Census Income" para criar um modelo capaz de prever se o salário anual de uma determinada pessoa é acima ou abaixo de 50 mil dólares. Uma vez treinado o modelo, posteriormente utilizando metodologias de explicabilidade para evidenciar mazelas sociais através da determinação da importância que cada feature teve para a decisão do modelo.

METODOLOGIA

Para elaborar o modelo, utilizando a área abaixo da curva característica de operação do receptor como métrica de performance, foi feita uma sondagem de vários modelos diferentes: Decision Tree, Random Forest, Extreme Gradient Boosting e Light Gradient Boosting Machine. Tendo estabelecido o modelo de regressão logística como baseline de comparação e analisando os modelos tanto por suas acurácias quanto AUC, conclui-se que o mais performático foi o LightGBM. Para obter explicações sobre as decisões do modelo para cada indivíduo no dataset e então obter um certo grau de interpratabilidade dos resultados obtidos, foram efetivadas as seguintes metodologias:

- ➔ **Shapley Additive Explanations (SHAP),** para obter a importância de cada uma das features na decisão do modelo.
- ➔ Visualização **t-SNE** utilizando os valores SHAP (importância das features) para cada um dos indivíduos, aliada da probabilidade de cada um deles para estabelecer uma correlação entre importância das features e a classificação binária obtida e as probabilidades da predição.

EXPERIMENTOS E ANÁLISES

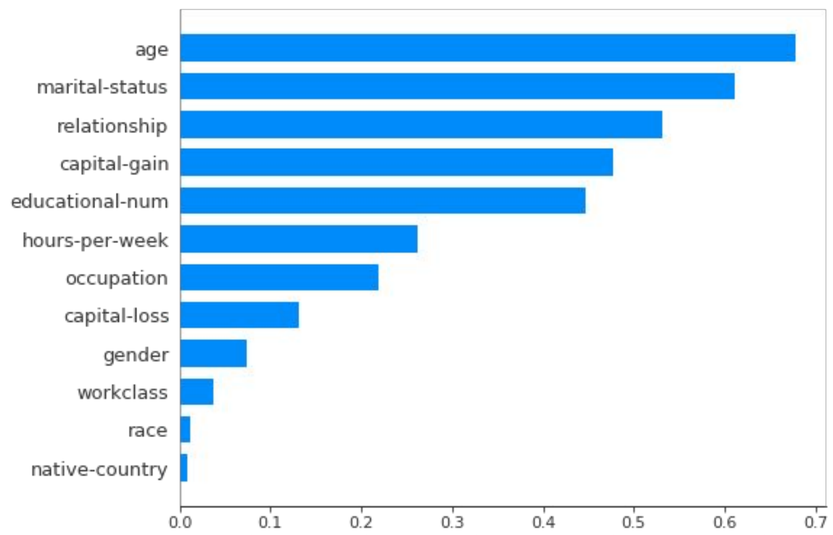

Figura 1: Gráfico de barras da importância média de cada feature.

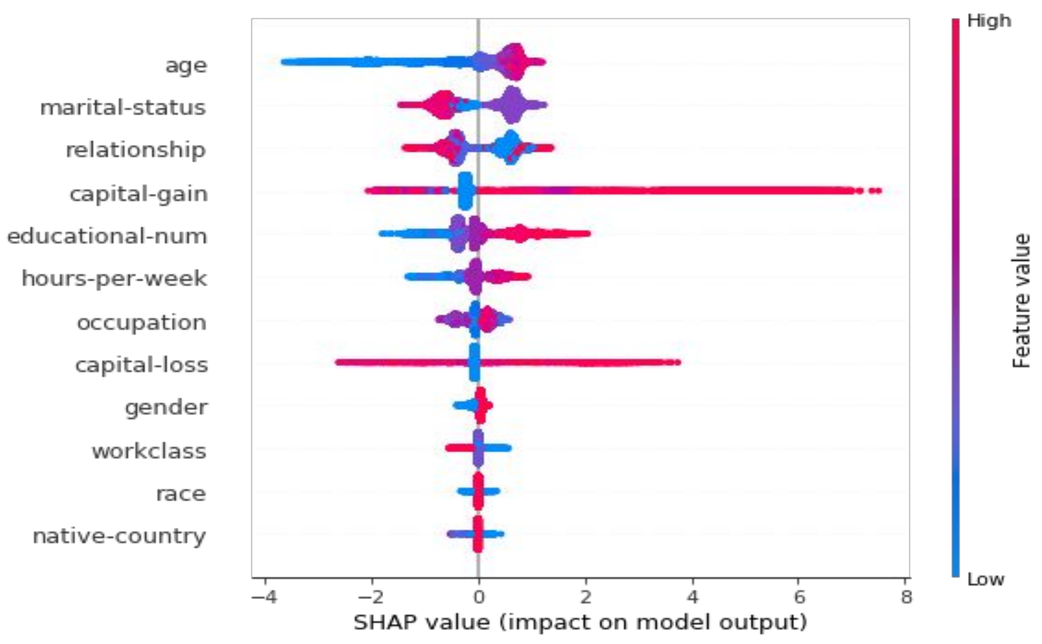

Figura 2: Densidade do valor do SHAP para cada feature.

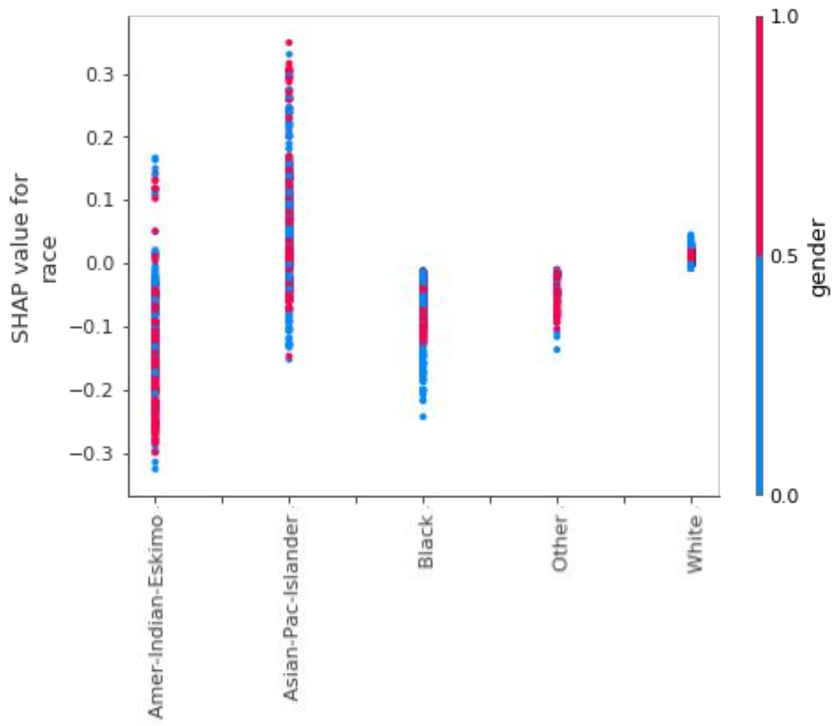

Figura 3: Plots de dependência dos valores SHAP para as features "race" e "gender" .

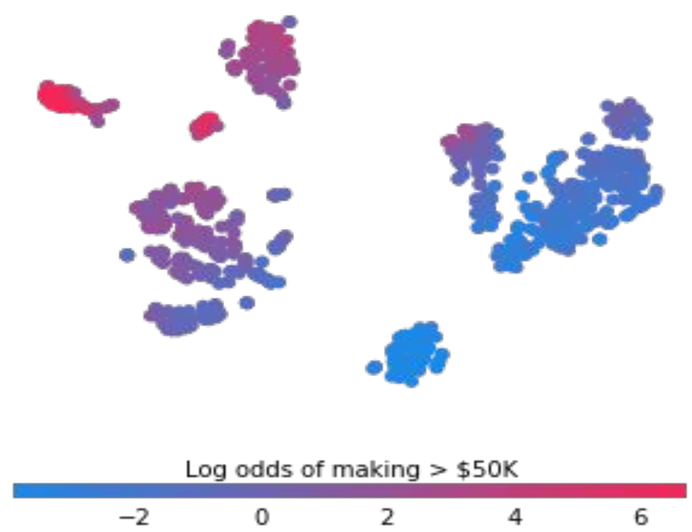

Figura 4: t-SNE dos valores SHAP.

CONCLUSÕES

A partir dos resultados obtidos, conclui-se que para o dataset "Adult Census Income", a grande maioria das features mapeadas são importantes para a predição binária proposta, condizente com a expectativa social do fenômeno. Em particular, analisando as features de raça e gênero, é evidente que o modelo as enxergou como determinantes para tomar a classificação final dos indivíduos, pois mazelas sociais levam indivíduos do gênero feminino e pessoas de etnias afrodescendentes a terem uma menor probabilidade de ganharem acima de $50.000 por ano.

Analisando o t-SNE obtido, vemos uma série de pequenos clusters de pessoas para as quais o modelo determinou importância similares de features. É possível identificar um certo gradiente horizontal na probabilidade de se ganhar acima de $50.000. Intuitivamente podemos entendê-lo como um espectro que atravessa as várias camadas sociais, começando, por exemplo de pessoas negras com baixa escolaridade e gradativamente adicionando fatores que aumentam as chances de um salário elevado como: pessoa branca do gênero masculino tendo ocupação no governo federal.