Relatório - Projeto Turing Academy

Grupo 1

Integrantes:
Asaffe Apolinário Duarte 14560001
Ana Vitória Abreu Murad 14613160
Solano Omar Oliveira do Nascimento 14608017

- **Análise exploratória e limpeza:** Começamos o projeto com uma análise exploratória do dataset. Nela, visualizamos os dados separados para treino, plotando dois gráficos na análise de cada feature, um deles para analisar a distribuição dela individualmente e outro para visualizar a relação dela com a target. Utilizamos gráficos de barras e boxplots, majoritariamente.

  Após uma análise detalhada, iniciamos a limpeza substituindo dados inconsistentes ou redundantes encontrados nas colunas 'job' e loan', removemos os outliers relativos às colunas 'duration', 'age' e 'previous' e fizemos a inputação nas colunas com dados faltantes utilizando diferentes métodos, mais simples como moda ou mais complexos como IterativeImputer. Ademais, acrescentamos na nossa análise um heatmap para visualizar as relações entre as colunas e decidir melhor como utilizá-las no treinamento dos modelos.

- **AutoGluon:** Utilizamos alguns modelos de redes neurais, que não nos renderam bons resultados iniciais, já submetidos no Kaggle. Utilizamos, então, um modelo de *Auto Machine Learning*, chamado *autogluon*, pois ele é especializado em modelos de árvores de decisão que são ótimos para dados tabulares. A vantagem de usar uma Auto ML é que ela cria e testa vários modelos, na casa de dezenas, e retorna os melhores modelos, de acordo com a métrica de pontuação escolhida pelo usuário. Como especificado no Kaggle, utilizamos AUC como métrica e o preset de '*best quality*', parâmetro do autogluon. Sem nenhum pré-processamento, apenas com a limpeza, obtivemos um score perto de 76%.

- **Pré-processamento dos dados:** Nesse momento resolvemos focar no pré-processamento dos dados, utilizando OneHotEncoder para algumas features categóricas, excluindo features inúteis de acordo com análises com *heatmap* e *Chi² score* e também fazendo a normalização das colunas. Após o treinamento dos modelos vimos que o pré-processamento aumentou nosso score em 1.5%, o que para nós não fez muito sentido, pois os outros grupos estavam com *scores* maiores que 80% e tínhamos utilizado muitos modelos, com todo o pré-processamento que julgávamos necessário e que éramos capazes de fazer.

  Com isso em mente, resolvemos arriscar testando diferentes artifícios, como excluir a limpeza, criar novas amostras para treinamento (oversampling) e aumentar os pesos no treinamento com o *auto-weight* (outra vantagem de usar Auto ML na prática, que é a personalização da otimização dos hiperparâmetros com eficiência e praticidade). Não obtivemos grandes resultados, alguns foram inclusive muito menores do que o obtido anteriormente.

- **Balanceamento da feature target:** Decidimos, então, como última tentativa, tentar balancear o *dataset*, essa ideia acreditamos que veio da forma errada, ao pensar que o teste continha amostras injustas de alguma forma, porém, após obter ótimos resultados gradualmente, vimos que diminuir a quantidade de dados com a target com valor *'no'* no treino melhorou absurdamente os resultados no próprio treinamento, na validação, como no teste. Fizemos esse balanceamento de forma aleatória com diversas amostras e, após todas as tentativas, aumentamos em mais de 10% nosso *score*.