Desafio Cientista de Dados

Introdução

Olá candidato, o objetivo deste desafio é testar os seus conhecimentos sobre a resolução de problemas de análise de dados e aplicação de modelos preditivos. Queremos testar seus conhecimentos dos conceitos estatísticos de modelos preditivos, criatividade na resolução de problemas e aplicação de modelos básicos de machine learning. É importante deixar claro que não existe resposta certa e que o que nos interessa é sua capacidade de descrever e justificar os passos utilizados na resolução do problema.

Desafio

Seu objetivo é prever o salário anual (*yearly_wage*) de uma amostra de pessoas a partir de dados sóciodemográficos anonimizados. Para isso são fornecidos dois *datasets*: um *dataset* chamado *wage_train* composto por 32560 linhas, 14 colunas de informação (*features)* e a variável a ser prevista (“yearly_wage”).

O segundo dataset chamado de *wage_test* possui 16281 linhas e 14 colunas e não possui a coluna “yearly_wage”. **Seu objetivo é prever essa coluna a partir dos dados enviados e nos enviar para avaliação dos resultados.**

Você poderá encontrar em anexo um dicionário dos dados.

Entregas

1. Descreva graficamente os dados disponíveis, apresentando as principais estatísticas descritivas. Comente o porquê da escolha dessas estatísticas.

2. Explique como você faria a previsão do **salário** a partir dos dados. Quais variáveis e/ou suas transformações você utilizou e por quê? Qual tipo de problema estamos resolvendo (regressão, classificação)? Qual modelo melhor se aproxima dos dados e quais seus prós e contras? Qual medida de performance do modelo foi escolhida e por quê?
3. Envie o resultado final do modelo em uma planilha com apenas duas colunas (rowNumber, predictedValues).
4. A entrega deve ser feita através de um repositório de código público que contenha:
    a. README explicando como rodar o projeto
    b. Arquivo *requirements* com todos os pacotes utilizados
    c. Relatório de EDA em PDF, Jupyter Notebook ou semelhante conforme passo 1
    d. Códigos de modelagem utilizados no passo 2.
    e. Arquivo final *predicted.csv* conforme passo 3 acima.

Prazo

Você tem até **7 dias corridos** para a entrega, contados a partir do recebimento deste desafio. Envie o seu relatório dentro da sua data limite para o e-mail: **selecao.lighthouse@indicium.tech**

Bom trabalho!

Dicionário dos dados

age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education_num: continuous.

marital_status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital_gain: continuous.
capital_loss: continuous.
hours_per_week: continuous.
native_country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
yearly_wage: >50K, <=50K.