DESAFIO DATA SCIENCE

1. Resumo

Olá candidato, o objetivo deste desafio é testar os seus conhecimentos sobre construção de modelos preditivos. Queremos testar seus conhecimentos dos conceitos estatísticos de modelos preditivos, criatividade na resolução de problemas e aplicação de modelos de *machine learning* em produção. 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.

2. Descrição do Problema

Seu objetivo é prever o *churn* (abandono de clientes) de um banco de dados fictício de uma instituição financeira. Para isso são fornecidos dois *datasets*: um *dataset* chamado **Abandono_clientes** composto por 10000 linhas e 13 colunas de informação (*features)*, sendo uma coluna "Exited" composta por dados binários: 1 se o cliente abandonou o banco, 0 se não. O segundo dataset possui 1000 linhas e 12 colunas e não possui a coluna "Exited". **Seu objetivo é construir um pipeline de Machine Learning que permita prever essa coluna a partir dos dados enviados.**

3. Atividades

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

2. Explique como você faria a previsão do **Churn** 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?

Este documento é de uso exclusivo dos clientes e parceiros da Indicium Tecnologia de Dados LTDA
e não deve ser reproduzido ou compartilhado sem autorização expressa da Indicium.

3. Construa um pipeline de machine learning que realize a previsão de churn a partir de um dataset CSV. Esse pipeline deve ser reproduzível e permitir realizar uma previsão a partir de qualquer arquivo CSV com a mesma estrutura de dados. O modelo utilizado pelo pipeline deve ser treinado com o dataset **Abandono_clientes.csv.** Você deve nos enviar o repositório de código com o pipeline.

4. Envie o resultado final do modelo em um arquivo com apenas duas colunas (rowNumber, predictedValues) gerado ao rodar o pipeline de 3 no dataset **abandono_teste.csv** em anexo.

Prazo

Você tem até **7 dias corridos** para a entrega, contados a partir do recebimento deste desafio.

Bom trabalho!

Este documento é de uso exclusivo dos clientes e parceiros da Indicium Tecnologia de Dados LTDA
e não deve ser reproduzido ou compartilhado sem autorização expressa da Indicium.