**Visão geral do projeto**

Você trabalha na empresa de mineração OilyGiant. Sua tarefa é encontrar os melhores lugares para o desenvolvimento de novos poços de petróleo.

Para concluir essa tarefa, você vai precisar executar as seguintes etapas:

* Ler os arquivos com parâmetros coletados de poços de petróleo na região selecionada: a quantidade de petróleo e o volume de reservas;
* Construir um modelo para predizer o volume de reservas em novos poços;
* Escolher os poços de petróleo que têm os maiores valores estimados;
* Escolher a região com o maior lucro total para os poços de petróleo selecionados.

Você tem dados sobre amostras de petróleo de três regiões. Os parâmetros de cada poço de petróleo na região já são conhecidos. Construa um modelo que te ajudará a escolher a região com a margem de lucro mais alta. Utilize a técnica de *Bootstrapping* para analisar lucro potencial e riscos.

**Descrição de dados**

Os dados de exploração geológica para as três regiões estão armazenados em arquivos:

* geo\_data\_0.csv. [baixe o conjunto de dados](https://practicum-content.s3.us-west-1.amazonaws.com/datasets/geo_data_0.csv)
* geo\_data\_1.csv. [baixe o conjunto de dados](https://practicum-content.s3.us-west-1.amazonaws.com/datasets/geo_data_1.csv)
* geo\_data\_2.csv. [baixe o conjunto de dados](https://practicum-content.s3.us-west-1.amazonaws.com/datasets/geo_data_2.csv)
* *id* — identificador unívoco de poço de petróleo
* *f0, f1, f2* — três características de pontos (seu significado específico não é importante, mas as próprias características são significativas)
* *product* — volume de reservas no poço de petróleo (milhares de barris).

**Condições:**

* Apenas regressão linear deve ser usada para o treinamento do modelo.
* Ao explorar a região, um estudo de 500 pontos é realizado e os melhores 200 pontos são selecionados para calcular o lucro.
* O orçamento para o desenvolvimento de 200 poços de petróleo é 100 milhões de dólares.
* Um barril de petróleo bruto traz 4.5 dólares de receita. A receita de uma unidade de produto é 4.500 dólares (o volume de reservas está em milhares de barris).
* Depois de ter avaliado os riscos, mantenha apenas as regiões com o risco de perdas inferior a 2.5%. Entre aquelas que se enquadram no critério, você precisa selecionar a região com o lucro médio mais alto.

Os dados são sintéticos e não incluem nenhum detalhe de contratos ou características de poços.

**Instruções do projeto**

1. Baixe e prepare os dados. Explique o procedimento.
2. Treine e teste o modelo para cada região em geo\_data\_0.csv:
   1. Divida os dados em um conjunto de treinamento e um conjunto de validação em uma proporção de 75:25.
   2. Treine o modelo e faça predições para o conjunto de validação.
   3. Salve as predições e respostas corretas no conjunto de validação.
   4. Imprima o volume médio previsto de reservas e o *REQM* do modelo.
   5. Analise os resultados.
   6. Coloque todos os passos anteriores em funções, execute os passos 2.1 a 2.5 nos arquivos 'geo\_data\_1.csv' e 'geo\_data\_2.csv'.
3. Prepare-se para o cálculo de lucro:
   1. Armazene todos os valores necessários para os cálculos em variáveis separadas.
   2. Dado o investimento de 100 milhões para 200 poços de petróleo, cada um precisa produzir, em média, uma quantidade de unidades equivalente a pelo menos 500 mil dólares para evitar prejuízos (isso é aproximadamente 111,1 unidades). Compare essa quantidade com o volume médio de cada região.
   3. Forneça conclusões sobre a etapa de preparação para o cálculo de lucro.
4. Escreva uma função para calcular lucro de um conjunto de poços de petróleo selecionados e predições do modelo:
   1. Escolha os 200 poços com os valores mais altos previstos de cada uma das 3 regiões (ou seja, arquivos 'csv').
   2. Sumarize o volume alvo de reservas de acordo com essas predições. Armazene as predições para os 200 poços para cada uma das 3 regiões.
   3. Calcule o lucro potencial dos 200 melhores poços por região. Apresente suas conclusões: sugira uma região para o desenvolvimento de poços de petróleo e justifique sua escolha.
5. Calcule riscos e lucro para cada região:
   1. Usando as predições que você armazenou na etapa 4.2, use a técnica de bootstrapping com 1.000 amostras para encontrar a distribuição de lucros.
   2. Encontre lucro médio, intervalo de confiança de 95% e o risco de prejuízo. Prejuízo é um lucro negativo, calcule-o como uma probabilidade e depois o expresse como uma porcentagem.
   3. Apresente suas conclusões: sugira uma região para o desenvolvimento de poços de petróleo e justifique sua escolha. A sua escolha corresponde à escolha anterior na etapa 4.3?