Instruções do Projeto 2

Criação de Modelo de Regressão para Prever o Preço de Carros

O objetivo do Projeto 2 é a criação de um modelo de Machine Learning (Regressão) para prever o preço de carros, utilizando o dataset "Used Cars Dataset" e os conhecimentos adquiridos no programa de bolsas.

Para iniciar o Projeto, vocês devem gerar uma amostra aleatória de 25% do Dataset. Existem várias maneiras de gerar uma amostra aleatória, a escolha fica a critério de vocês. Dica: Gerem uma amostra reprodutível, isso garante que se precisarem rodar o código várias vezes, incluindo a parte de geração da amostra, ela sempre será a mesma (mesmo sendo aleatória). Uma amostra reprodutível facilitará o trabalho de vocês.

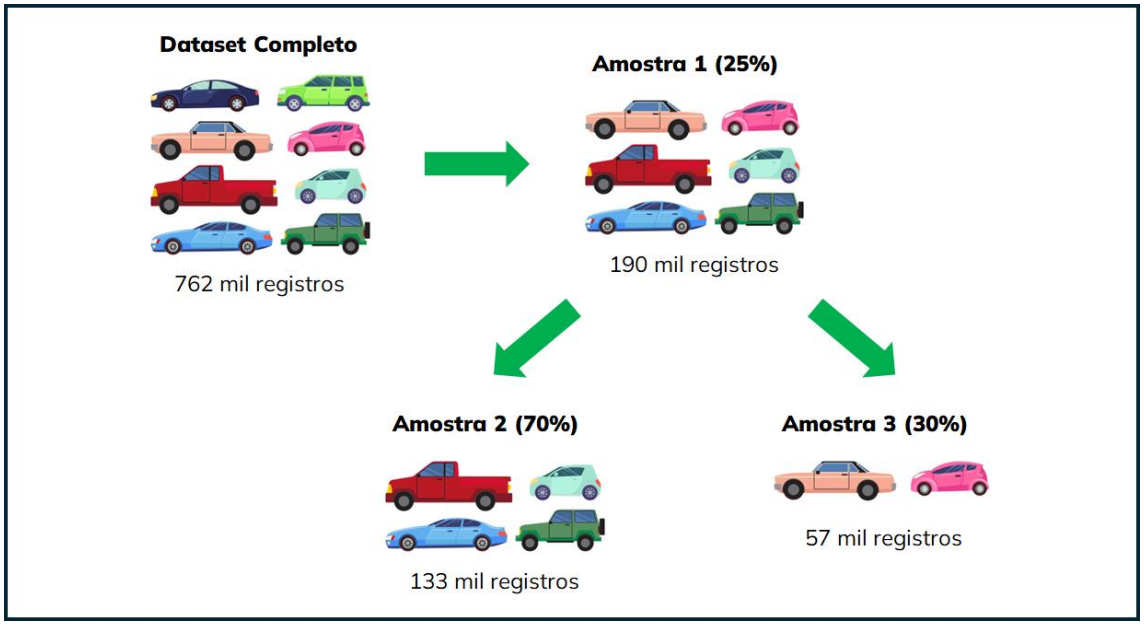
Com essa amostra de 25% (cerca de 190 mil registros):

- Apresentem uma análise exploratória de todas as variáveis. Utilizem gráficos para complementar a análise.
- Nessa análise exploratória, respondam a seguinte pergunta: Quais variáveis vocês classificam como numéricas e quais classificam como categóricas? Para as variáveis numéricas, apresentem visualmente (gráfico) a correlação de Pearson.
- Criem um modelo de Regressão utilizando como variável resposta (Target) a coluna "price". Antes do treinamento do modelo, vocês devem dividir os dados em treino e teste, seguindo a proporção 70/30, ou seja, 70% dos dados (cerca de 133 mil registros) para treinar o modelo e 30% dos dados (cerca de 57 mil registros) para testar o modelo. Dica: Para fazer a divisão em treino e teste, utilizem a função "train_test_split".
- Antes do treinamento do modelo, lembrem-se de realizar a etapa de pré-processamento dos dados (limpeza, transformações, etc.).
- Após o treinamento, escolham ao menos duas métricas de Regressão para avaliar a performance do modelo. Essas métricas devem ser calculadas nos dados de teste (30% / 57 mil registros). A escolha das métricas fica a critério de vocês. Justifiquem a escolha.
- Analisando essas métricas, qual é a conclusão que vocês chegam? A performance do modelo foi boa ou ruim?
- A escolha do algoritmo fica a critério de vocês. Justifiquem a escolha.
- Por fim, respondam a seguinte pergunta: Quais são as duas variáveis mais importantes para o modelo? Expliquem como chegaram nessa conclusão.

O Projeto deve ser entregue em um Jupyter Notebook. Não esqueçam de documentar cada parte do código desenvolvido, utilizando comentários em código e/ou Markdown. Isso facilitará o nosso entendimento na avaliação da lógica utilizada por vocês. **O Jupyter Notebook final deve ser entregue com todas as células executadas**.

No universo da Ciência de Dados, trabalhar com amostras é uma técnica bastante utilizada e importante. As amostras são subconjuntos dos dados totais disponíveis e são utilizadas para extrair insights e inferências sobre a população da qual foram selecionadas. Ao lidar com grandes conjuntos de dados, o uso de amostras permite economizar recursos computacionais e tempo, além de facilitar a manipulação e a análise dos dados.

A imagem abaixo ilustra como as amostras devem ser geradas:



Descrição do Dataset

Used Cars Dataset

Esse dataset contém dados sobre 762.091 carros usados retirados do site cars.com, totalizando 20 variáveis. Os dados foram coletados em abril de 2023.

Descrição das variáveis

- **manufacturer:** Nome do fabricante.
- **model:** Modelo do carro.
- **year:** Ano de produção.
- **mileage:** Número de milhas percorridas.
- **engine:** Descrição do motor.
- **transmission:** Tipo de transmissão.
- **drivetrain:** Tipo de tração.
- **fuel_type:** Tipo de combustível.
- **mpg:** Milhas por galão.
- **exterior_color:** Cor externa.
- **interior_color:** Cor interna.
- **accidents_or_damage:** Envolvimento em acidentes (1 = sim / 0 = não).
- **one_owner:** Único dono (1 = sim / 0 = não).
- **personal_use_only:** Apenas uso pessoal (1 = sim / 0 = não).
- **seller_name:** Nome do vendedor.
- **seller_rating:** Avaliação do vendedor.
- **driver_rating:** Avaliação do carro pelos motoristas.
- **driver_reviews_num:** Número de avaliações pelos motoristas.
- **price_drop:** Redução do preço em relação ao preço inicial.
- **price:** Preço do carro.