Aluno: Diego Gomes

Resposta

1.f - Principais resultados da Análise Exploratória dos Dados

A análise exploratória revelou que a base de dados contém valores faltantes em algumas colunas, especialmente nas informações extraídas do modelo do carro, como engine_size. Também foram identificados dados duplicados, que foram removidos. A estatística descritiva mostrou variações significativas nos preços médios dos carros por marca e modelo, com a Fiat sendo a marca mais representada no conjunto de dados.

2.e - Explicação sobre a distribuição da média de preço dos carros por marca e tipo de engrenagem item d

A análise do gráfico revelou que os carros com câmbio automático tendem a ter um preço médio maior do que os de câmbio manual, independentemente da marca. Marcas premium apresentam preços significativamente mais altos para veículos automáticos, enquanto marcas populares têm menor discrepância de preços entre os tipos de engrenagem.

2.g - Explicação sobre a distribuição da média de preço dos carros por marca e tipo de combustível item f

Os dados mostraram que veículos movidos a diesel apresentam preços médios mais altos, seguidos pelos movidos a gasolina e, por último, os movidos a álcool. Isso se deve ao fato de que muitos veículos a diesel são caminhonetes e SUVs de alto valor agregado. Modelos flex apresentam uma variação maior de preço, dependendo da marca e do segmento do carro.

3.f - Explicação sobre a importância das variáveis na estimativa do preço dos carros

A análise de importância das variáveis indicou que o year_model (ano do modelo) e a brand (marca) são os fatores mais relevantes para a estimativa do preço dos carros. Além disso, a engine_size (tamanho do motor) e o tipo de gear (transmissão) também tiveram influência considerável. Variáveis categóricas, quando bem transformadas, ajudaram a melhorar a acurácia do modelo.

3.h - Explicação sobre qual modelo gerou o melhor resultado e a métrica de avaliação utilizada

O modelo XGBoost apresentou melhor desempenho em comparação ao RandomForest, com um R^2 mais alto e menores valores de erro médio absoluto (MAE) e erro médio quadrático (MSE). Isso indica que o XGBoost conseguiu capturar melhor as relações entre as variáveis, possivelmente devido à sua capacidade de modelar interações não lineares de forma mais eficiente.