TRABALHO DE IAA002 – Linguagem de Programação Aplicada

Bruno Galvão
Walter José Horning Junior

1.F - Dê um breve explicação (máximo de quatro linhas) sobre os principais resultados encontrados na Análise Exploratória dos dados:

O dataset possui 11 colunas e 267542 linhas, porém 65245 linhas são nulas, das quais 2 são duplicadas. Essas observações contemplam 2112 modelos de carros diferentes, dentre as 6 marcas existentes no dataset. A média de preço de todos os carros é de 52.756,90, variando de 6.647,00 até 979.358,00, com mediana de 38.027,00 (BRL).

2.E - Dê uma breve explicação (máximo de quatro linhas) sobre os resultados gerados no item d:

Em quase todas as marcas (com exceção da Renault), os carros com engrenagem automática tendem a ser mais caros do que os carros com engrenagem manual. Pode-se perceber também que a Ford, GM - Chevrolet e Nissan possuem uma média de preço parecida para carros com engrenagem manual.

2.G - Dê uma breve explicação (máximo de quatro linhas) sobre os resultados gerados no item f:

As marcas Nissan e Renault não apresentaram nenhum carro movido à álcool neste dataset. A média dos preços dos carros que usam álcool é muito menor, comparado aos outros. Existe uma substancial diferença entre a média dos preços para gasolina e diesel. Enquanto a maior média para gasolina é ~60.000, a do diesel é ~140.000 (2,5 vezes maior).

3.F - Dê uma breve explicação (máximo de quatro linhas) sobre os resultados encontrados na análise de importância de variáveis:

Para os quatro modelos, as variáveis "year_of_reference" e "gear" tiveram baixa significância e são candidatas a serem retiradas do modelo. As variáveis "engine_size" e "year_model" foram as mais significativas, com uma leve vantagem de “engine_size”, tendo maior relevância na regressão. A variável “engine_size” explica ~52% dos dados.

3.H - Dê uma breve explicação (máximo de quatro linhas) sobre qual modelo gerou o melhor resultado e a métrica de avaliação utilizada:

RandomForest com parâmetros escolhidos foi o que teve melhor desempenho. Devido a facilidade de interpretação, escolhemos R^2, pois dado seu resultado de 0.926572, pode-se dizer que o modelo explica ~92% da variância da variável dependente, a partir das variáveis independentes. Todos os modelos tiveram resultados próximos.