Passo 3 - Desafio Lighthouse Ciência de Dados

Explicação geral:

A forma de prever os preços dos carros consistiu na implementação de um algoritmo de aprendizado de máquina que se utiliza das informações disponíveis no conjunto de dados e das relações dessas informações com a variável a ser prevista. Para concretizar essa implementação, foi preciso um longo trabalho de limpeza e manipulação dos dados para que eles recebessem uma forma adequada aos algoritmos de aprendizado de máquina.

Tipo de problema:

Trata-se de um problema de regressão, uma vez que a variável target, o preço, é uma variável quantitativa contínua.

Variáveis e transformação:

O conjunto de dados para o treinamento do algoritmo possuía 28 colunas além da coluna que continha a variável preço. Algumas dessas colunas já foram excluídas no início. A variável 'id' porque seus valores são únicos. A variável 'modelo' porque possuía um número muito grande de categorias, o que compromete o desempenho dos algoritmos. A coluna 'versao' foi excluída também mas suas informações foram dispostas em duas novas colunas. A coluna 'ano_de_fabricação' foi excluída porque entrava em contradição com a coluna 'ano_modelo'. Em geral, seus valores deveraim ser os mesmos ou possuir uma diferença de 1. Os teste mostraram, no entanto, 17465 linhas em que havia incoerência entre ano_de_fabricação e ano_modelo. Comparando com os valores do hodômetro, os anos do modelo pareciam os corretos e, por isso, o ano_de_fabricacao foi excluído. A coluna 'cidade_vendedor' também foi excluída por motivos de incoerência. As cidades indicadas não pertenciam aos estados que estava indicados na coluna adjacente 'estado_vendedor'. Os valores foram comparados com uma base de dados do governo federal em que constam todos os municípios do Brasil e seus respectivos estados(*Lista_Municípios_com_IBGE_Brasil_Versao_CSV.csv*). O percentual de entradas com incoerência entre cidade e estado do vendedor foi de 26.33%. Neste caso, foi mais indicado excluir a coluna cidade_vendedor porque o desafio envolvia questões a respeito dos estados. A coluna 'ipva_pago' também foi excluída mas porque as análise ANOVA indicaram que esta variável não tinha relação com a variável 'preco'. A coluna 'veiculo_alienado' foi descartada porque estava inteiramente em branco.

Além da exclusão, dessas informações, muitas colunas tiveram seus valores alterados para melhor se adpatar aos algoritmos. A coluna 'marcas', por exemplo, foi reduzida porque continha muitos níveis. Todas as marcas que representavam menos de 1% do conjunto de dados foram reunidas numa categoria chamada 'outra'. OUtras variáveis passaram por processos análogos.

A principais transformações empregadas nos dados para a adaptação aos algoritmos foi a aplicação do One-hot Encoding para as variáveis categóricas e a padronização dos valores. Foi escolhido o *StandardScaler()* porque os dados não apresentavam outliers.

Modelo de aprendizado de máquina:

Foram testado quatro algoritmos diferentes: Regressão Linear, Árvore de Decisão, Floresta Aleatória, XGBoost e kNN. Eu confesso que ainda não entendo claramente os passos de cada um dos algoritmos. O motivo pelo qual eu os escolhi foi uma pesquisa pela internet. Encontrei problemas semelhantes de regressão aos quais os cientistas de dados endereçavam esse tipo de estratégia. Segui o exemplo. Dos quatro algoritmos, o XGBoost foi o que performou melhor e, por isso, foi escolhido ao final.

Medida de performance:

A medida de performance utilizada foi o coeficiente de determinação. Também aqui procurei seguir o exemplo de pessoas mais experientes.