Relatório referente ao Desafio Cientista de Dados

Processo Seletivo

–

Lighthouse Programa De Formação Em Dados

INDICIUM

Candidato: Cláudio Gabriel Duarte Brandão

Objetivo: O projeto será estruturado em duas etapas principais: análise dos dados e desenvolvimento de um modelo preditivo para estabelecer preços de carros do cliente de forma mais alinhada aos valores de mercado. Adicionalmente, foram formuladas três novas perguntas de negócio para ampliar a abordagem analítica.

Etapa 1: Análise dos Dados

Nesta fase, será conduzida uma análise detalhada dos dados para responder às perguntas de negócio formuladas. Serão coletadas informações sobre os carros disponíveis para venda, englobando características como marca, modelo, ano de fabricação, quilometragem, estado de conservação, histórico de revisões, titularidade, pagamento do IPVA, aceitação de trocas e presença de garantia de fábrica.

Etapa 2: Criação do Modelo Preditivo

Nesta etapa, será desenvolvido um modelo preditivo que permitirá precificar os carros do cliente de maneira mais precisa em relação aos valores praticados no mercado. A partir dos insights obtidos na análise dos dados, as variáveis mais relevantes serão selecionadas para compor o modelo.

Dados: 2 *datasets*, um para treinamento (contendo 29584 e 29 colunas) e outro para teste (contendo 9862 e 28 colunas).

EDA (*Exploratory Data Analysis*): Inicialmente, foi realizado uma exploração de dados abrangente, visando compreender melhor a estrutura e os padrões presentes no conjunto de dados. Nesta etapa, foram executadas as seguintes atividades:

1. **Análise da Qualidade dos Dados**: Verificação da integridade dos dados, identificação de valores ausentes ou inconsistentes e adoção de estratégias para tratamento desses casos, como imputação ou remoção de registros.
2. **Análise de Variáveis Categóricas:** Além da verificação dos tipos de variáveis, realizou-se uma análise mais detalhada das variáveis categóricas, identificando a frequência de cada categoria e possíveis relações com as variáveis numéricas.
3. **Visualizações Estatísticas:** Utilizando gráficos como barras e *box plots*, foram exploradas as distribuições das variáveis categóricas e suas relações com as variáveis numéricas, permitindo insights adicionais.
4. **Visualização Interativa:** Em alguns casos, foram empregadas ferramentas de visualização interativa, como plotagens interativas, para possibilitar a exploração dinâmica dos dados.

Em seguida foi respondido as três perguntas sugeridas pelo desafio:

1. **Qual o melhor estado cadastrado na base de dados para se vender um carro de marca popular e por quê?**

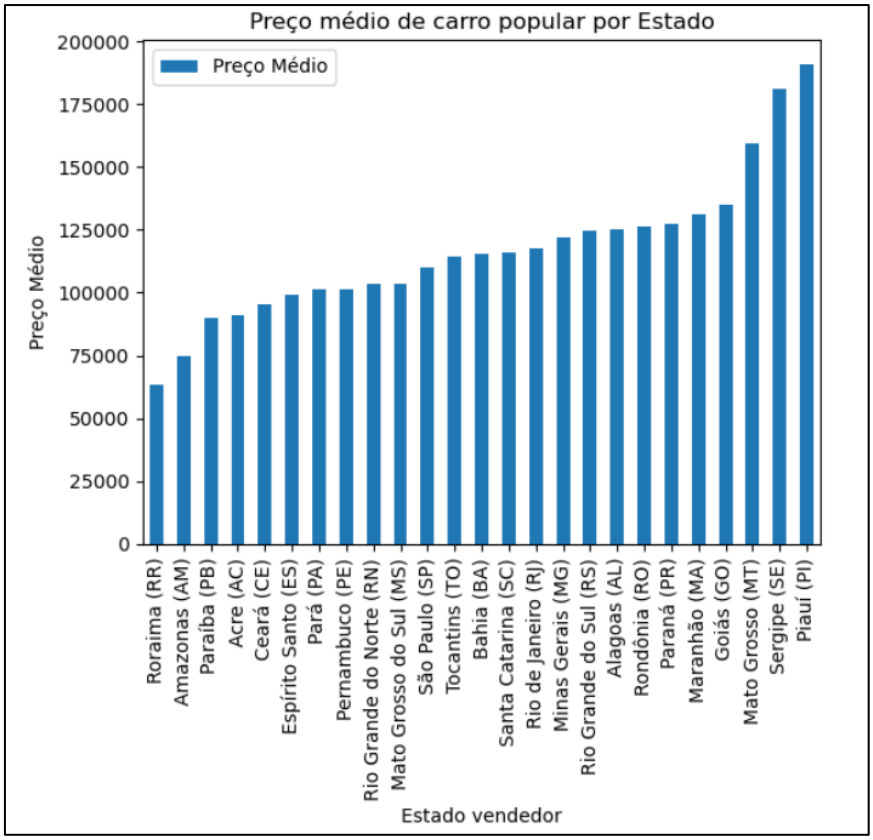
Para responder esta pergunta primeiramente eu criei colunas booleanas das variáveis revisoes_concessionaria, veiculo_único_dono, ipva_pago, dono_aceita_troca, troca e garantia_de_fábrica. Em seguida defini um filtro de marcas populares e utilizei as variáveis booleanas criadas juntamente com as outras para definir o melhor estado. Não achei viável utilizar somente a média de preços para definir o melhor estado, por isso utilizei outras variáveis para aumentar a precisão. Assim, cheguei a este resultado:

Em seguida, filtrei as marcas populares e utilizei essas variáveis booleanas juntamente com outras informações disponíveis para definir o melhor estado. A abordagem de não se basear somente na média de preços permitiu aumentar a precisão da recomendação.

Os critérios utilizados para determinar o melhor estado foram:

- revisoes_concessionaria: Valorizei os estados onde os carros possuem histórico de revisões realizadas em concessionárias, pois isso indica um cuidado com a manutenção e potencialmente atrai compradores preocupados com a procedência dos veículos;
- veiculo_único_dono: Priorizei estados onde a maioria dos carros populares são de único dono, o que geralmente é bem visto pelos compradores e pode influenciar positivamente na decisão de compra;
- ipva_pago: Estados onde os carros anunciados têm o IPVA pago receberam maior consideração, pois isso representa um benefício para o comprador e pode tornar a negociação mais atrativa;
- dono_aceita_troca: Considerei estados em que os proprietários dos carros estão dispostos a aceitar trocas, já que essa facilidade pode atrair mais interessados e agilizar o processo de venda; e
- garantia_de_fábrica: Valorizei estados onde os carros ainda estão cobertos pela garantia de fábrica, pois isso aumenta a confiança do comprador na aquisição do veículo.

Após analisar esses critérios e os dados disponíveis, concluí que:

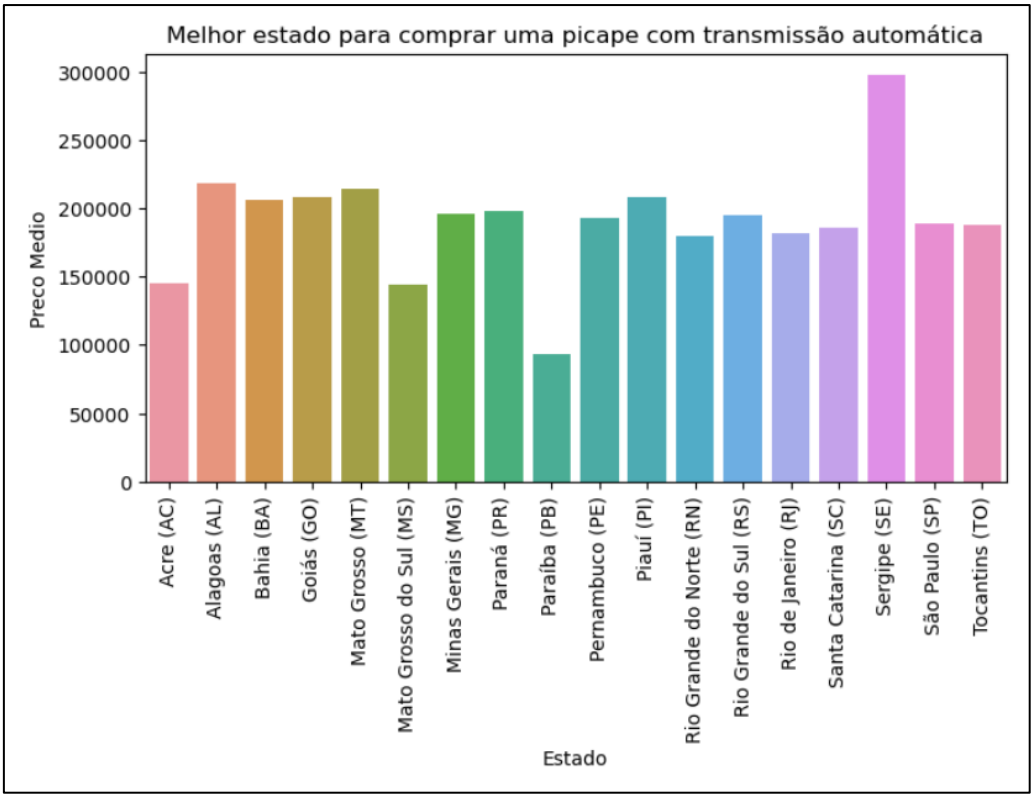


2. Qual o melhor estado para se comprar uma picape com transmissão automática e por quê?

Para determinar o melhor estado para se comprar uma picape com transmissão automática, realizei uma análise detalhada considerando múltiplas variáveis relevantes, como ano_de_fabricacao, ano_modelo e hodômetro. Essa abordagem visa aumentar a precisão e fornecer uma recomendação fundamentada. A query foi construída para filtrar somente as picapes com câmbio automático e tipo correspondente a picape. Foi definido essas variáveis adicionais em razão de:

- ano_de_fabricacao e ano_modelo: Deu-se maior relevância a picapes mais recentes em termos de ano de fabricação e modelo, pois geralmente oferecem tecnologia mais avançada e melhor desempenho.
- hodômetro: Consideramos picapes com menor quilometragem, pois tendem a ter menos desgaste e maior vida útil, o que pode significar menor necessidade de manutenção futura.
- Avaliação de Preços: Foi realizada uma comparação dos preços entre os estados, garantindo que o custo-benefício fosse favorável

Após analisar esses critérios e os dados disponíveis, pode-se concluir que o melhor estado para a compra da picape com transmissão automática é visto na figura abaixo:

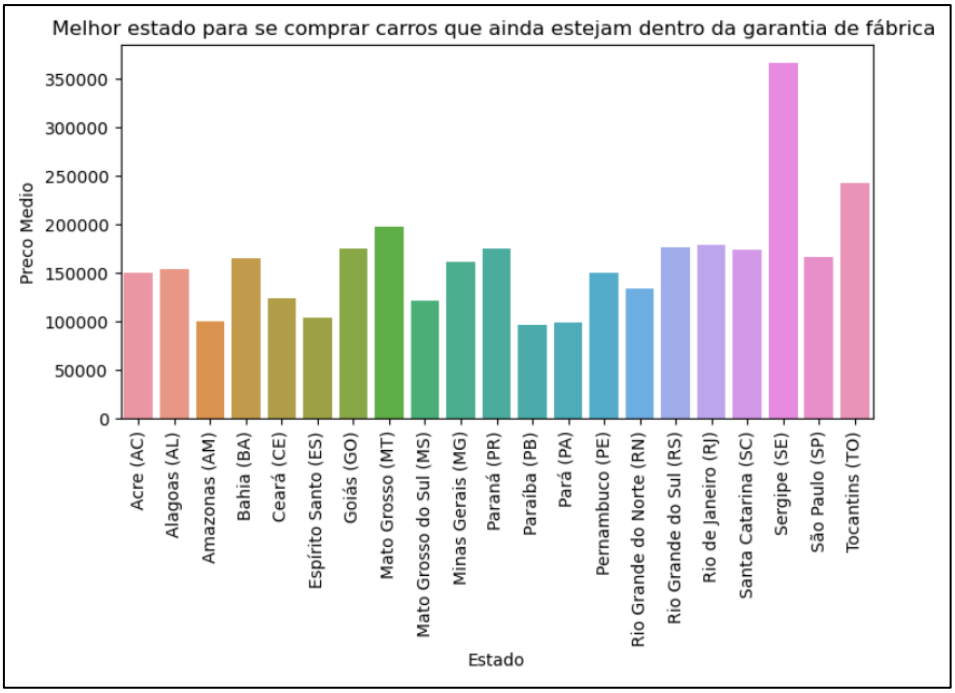


3. **Qual o melhor estado para se comprar carros que ainda estejam dentro da garantia de fábrica e por quê?**

Para determinar o melhor estado para se carros que ainda estejam dentro da garantia de fábrica, realizei uma análise detalhada considerando múltiplas variáveis relevantes, como ano_de_fabricacao, ano_modelo e hodômetro. Essa abordagem visa aumentar a precisão e fornecer uma recomendação fundamentada. A query foi construída para filtrar carros que ainda estejam dentro da garantia de fábrica. Foi definido essas variáveis adicionais em razão de:

- ano_de_fabricacao e ano_modelo eu-se maior relevância a carros mais recentes em termos de ano de fabricação e modelo, pois tendem a ter garantias mais extensas, proporcionando maior cobertura e tranquilidade ao comprador.
- hodômetro: Uma vez que veículos com menos uso geralmente apresentam menos desgaste e possuem mais tempo restante de garantia para cobrir possíveis problemas mecânicos.
- Avaliação de Preços: Foi realizada uma comparação dos preços entre os estados, garantindo que o custo-benefício fosse favorável

Após analisar esses critérios e os dados disponíveis, pode-se concluir que o melhor estado para a compra de carros ainda dentro da garantia de fábrica é visto na figura abaixo:

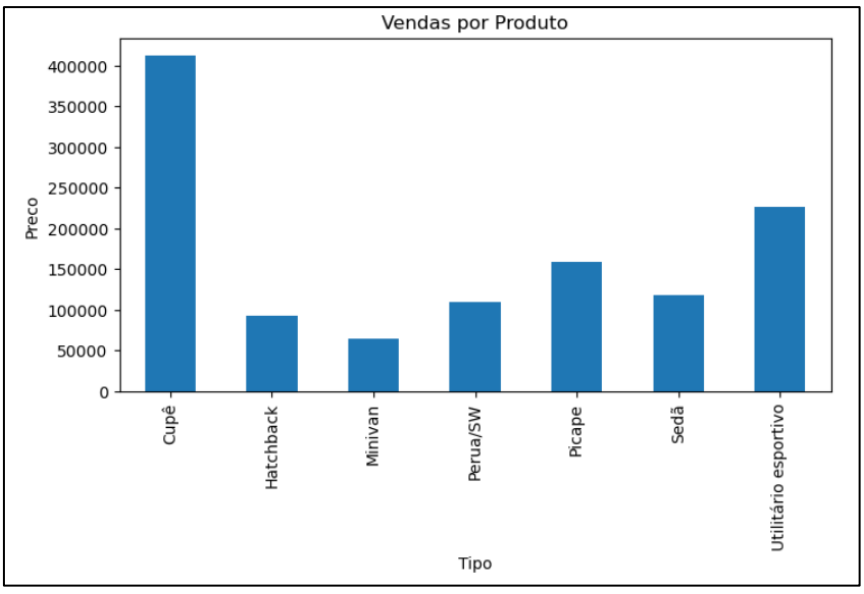


Foi criada ainda mais três perguntas de negócio, sendo elas:

- Qual o estado com a menor quilometragem média nos carros disponíveis para compra?
  R: Piauí (PI)

- Qual o estado com a maior quantidade de carros anunciados por pessoa física?
  R: São Paulo (SP)

- Qual é a média de preço dos carros por tipo de veículo?
  R:



Definição da coluna PREÇO no *cars_test*: Para definir a coluna de precificação dos carros, optei por utilizar o método de Regressão Linear em conjunto com a validação cruzada (Cross Validation) da biblioteca sklearn. Essa escolha foi fundamentada no fato de que o problema em questão envolve a previsão de valores numéricos, caracterizando-se como um problema de regressão, o que torna a Regressão Linear uma abordagem adequada e evita a necessidade de categorização ou classificação dos dados.

Para avaliar o desempenho do modelo, foram empregadas métricas relevantes, tais como o Erro Quadrático Médio (MSE), o Coeficiente de Determinação (R^2 Score) e o Erro Absoluto Médio (MAE). Essas métricas são essenciais para medir a qualidade das previsões e verificar o quão bem o modelo está se ajustando aos dados.

Antes de treinar o modelo, realizei a transformação das variáveis booleanas, criando novas variáveis a partir delas, como dono_aceita_troca, veiculo_único_dono, revisoes_concessionaria, ipva_pago, veiculo_licenciado, garantia_de_fábrica e revisoes_dentro_agenda. Além disso, efetuei a conversão das variáveis categóricas para o formato de string, adequando-as para o modelo.

Também procedi à criação de novas variáveis, como idade_veiculo, km_por_ano e e_flex, derivadas de outras informações disponíveis na base de dados. Essa engenharia de características tem o objetivo de enriquecer a base de dados e aprimorar o poder preditivo do modelo.

Para garantir a estabilidade do modelo, realizei a normalização das variáveis contínuas utilizando a classe StandardScaler da biblioteca sklearn. Esse processo ajusta todas as variáveis contínuas para uma escala comum antes de treinar o modelo, evitando possíveis viéses resultantes de diferenças nas escalas das características.

Foram criados modelos de Regressão Ridge e Regressão Lasso, ambos com técnicas de regularização (L2 e L1, respectivamente). A regularização é uma prática importante para evitar o overfitting e melhorar o desempenho geral do modelo. As métricas MSE, R^2 Score e MAE foram empregadas para avaliar o desempenho de ambos os modelos, utilizando a técnica de validação cruzada com cv=5.

Por fim, para otimizar a eficiência do algoritmo e melhorar a acurácia das previsões, removi variáveis que não apresentavam relevância para o treinamento do modelo, como veiculo_alienado. Essa etapa de seleção de características é fundamental para garantir que o modelo foque nas informações mais significativas para realizar as previsões de forma precisa e eficaz.

Com isso, obtive as seguintes métricas:

- Métrica com Ridge:

```
Métricas com Ridge:
Mean Squared Error (Train): 1855788934.188533
R² Score (Train): 0.7217123333903106
Mean Absolute Error (Train): 27542.009517370865
```

- Métrica com Lasso:

```
Métricas com Lasso:
Mean Squared Error (Train): 1941455437.4976058
R² Score (Train): 0.7088660819264192
Mean Absolute Error (Train): 27984.35247348117
```

O tipo de problema que estamos resolvendo é de regressão, pois estamos buscando prever um valor numérico contínuo, que é o preço dos carros. O objetivo é encontrar uma relação funcional entre as variáveis de entrada e o preço de cada veículo, permitindo realizar estimativas precisas dos valores de mercado.

Quanto ao modelo que melhor se aproxima dos dados, foram testados dois modelos de regressão: o Regressão Ridge e o Regressão Lasso. Ambos são métodos de regressão linear com diferentes técnicas de regularização, L2 para o Ridge e L1 para o Lasso.

- Regressão Ridge: Tem como principal vantagem a capacidade de reduzir o impacto de variáveis pouco relevantes para a resposta, evitando *overfitting* e melhorando a estabilidade do modelo. É mais adequado quando há múltiplas variáveis explicativas correlacionadas. Entretanto, pode não ser capaz de eliminar completamente variáveis irrelevantes.

- Regressão Lasso: Além de evitar *overfitting*, o Lasso também pode eliminar completamente variáveis irrelevantes, tornando-o útil para seleção de variáveis. É mais adequado quando se suspeita que muitas variáveis explicativas possam ser irrelevantes. No entanto, pode ser sensível a multicolinearidade e tende a escolher apenas uma das variáveis correlacionadas, tornando-se instável quando as variáveis são altamente correlacionadas.

A medida de performance do modelo escolhida foi a combinação de três métricas relevantes: Erro Quadrático Médio (MSE), Coeficiente de Determinação (R^2 Score) e Erro Absoluto Médio (MAE). Essas métricas são adequadas para problemas de regressão, permitindo avaliar a precisão do modelo nas previsões dos preços dos carros. O MSE e o MAE representam as diferenças entre os valores previstos e os valores reais, enquanto o R^2 Score fornece uma indicação da proporção da variância nos preços que é explicada pelo modelo. Ao combinar essas métricas, obtemos uma avaliação mais completa e precisa do desempenho do modelo, permitindo selecionar o modelo que melhor se ajusta aos dados e proporciona as previsões mais acuradas.