UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Silas Eufrásio da Silva, Rick Theo Felix Bessa

1. Introdução

O dataset 'Real Estate Price Prediction' apresenta dados sobre preços imobiliários, com objetivo de criar um previsor de preços através desses dados. A coluna alvo será a variável 'number of convenience stores', que será predita a partir de Regressão Linear. O dataset está disponível em: https://www.kaggle.com/datasets/quantbruce/real-estate-price-prediction/data, através da plataforma Kaggle.

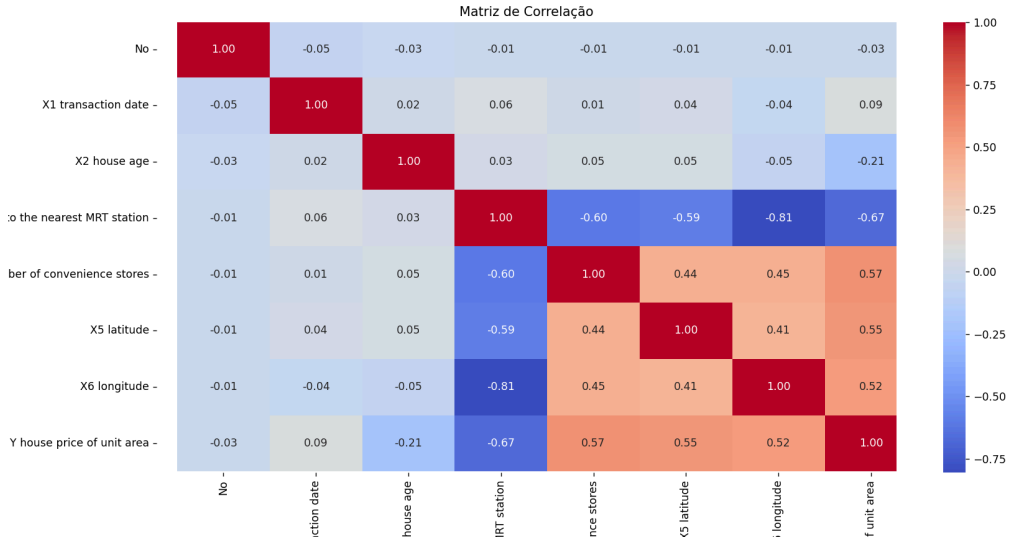
2. Observações

O fornecedor do dataset não ofereceu informações sobre os dados, logo, foi necessário pesquisas próprias para o entendimento das colunas.

3. Resultados e discussão

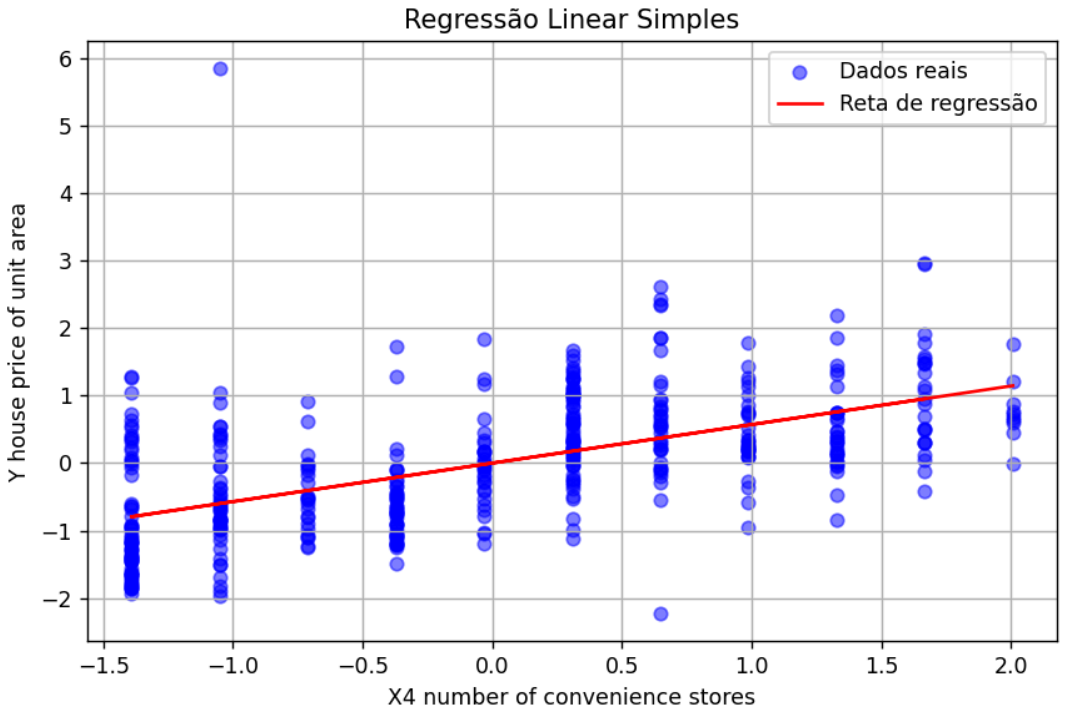
3. 1 Questão:

Dataset carregado através da biblioteca KaggleHub, primeiro verifica-se que não há valores ausentes em nenhuma coluna. Após isso, verifica-se as correlações, destacando que não há correlações altas em relação a variável alvo, com maior número de 0.5 para 'number of convenience stores' x 'Y house price of unit area'. Além disso, foi utilizado um standard scaller para padronizar os dados numéricos.



3. 2 Questão:

Aqui selecionamos a variável mais correlacionada com a variável target 'Y house price of unit area', que é: X4 number of convenience stores (correlação = 0.571). Então aplicamos a regressão linear através da biblioteca sklearn. Para acompanhar o desempenho do modelo, utilizamos o gráfico da reta da regressão linear e calculamos o RMSE e R^2. Segue o gráfico:



É notório que a reta da regressão linear acompanha bem o comportamento dos dados reais, apesar de alguns outliers. A métrica para RMSE = 0.82, indicam um número de acertos aceitável, considerando que a média dos valores da variável target é 9, apesar disso, ainda não é o ideal. Apesar disso, para R^2 = 0.32, temos um ajuste fraco, ou seja, a variável X4 number of convenience stores tem influência fraca sobre a variável Y, o que já foi observado pela matriz de correlação.

3. 3 Questão:

Neste script, realizamos uma comparação entre três modelos de regressão: Linear, Ridge e Lasso, utilizando validação cruzada com 5 folds. O objetivo foi avaliar qual modelo apresenta o melhor desempenho na previsão do preço dos imóveis por metro quadrado (Y house price of unit area), com base nas demais variáveis do conjunto de dados.

Para isso, aplicamos duas métricas de avaliação: o RMSE (Root Mean Squared Error), que mede o erro médio das previsões, e o R^2 (coeficiente de determinação), que indica o quanto da variação da variável-alvo é explicada pelo modelo. Os resultados foram calculados com a média dos scores obtidos em cada uma das cinco divisões dos dados.

Ao final, os modelos foram comparados em uma tabela e o modelo Ridge (RMSE= 0.64 e R^2 = foi destacado como o melhor. Essa abordagem permite uma avaliação mais robusta, evitando viés de uma única divisão entre treino e teste.

3. 4 Questão:

Neste experimento, utilizamos o modelo de regressão Lasso com penalização alpha = 0.1 para identificar os atributos mais relevantes na previsão do preço de imóveis por metro quadrado. O Lasso é especialmente útil para seleção automática de variáveis, pois reduz a zero os coeficientes de atributos menos importantes. Após o treinamento, o modelo identificou cinco variáveis com coeficientes diferentes de zero, ou seja, consideradas influentes na determinação do preço. Entre elas, destacam-se a distância até a estação de metrô mais próxima (com maior impacto negativo), o número de lojas de conveniência e a latitude (ambas com impacto positivo). Esses resultados indicam que localização e acessibilidade são fatores cruciais na valorização dos imóveis, enquanto características como idade da construção influenciam negativamente. A análise reforça a importância de variáveis urbanas no mercado imobiliário e mostra como o Lasso pode ser uma ferramenta eficaz para simplificar modelos preditivos mantendo sua precisão.

4. Conclusões

Os resultados foram satisfatórios parcialmente. A regressão linear simples não apresentou bom desempenho, pois utilizou apenas uma variável explicativa. Por outro lado, o uso de modelos regulares como Ridge e Lasso possibilitou uma melhoria nos resultados e uma compreensão mais profunda dos fatores que influenciam os preços. A abordagem com o Lasso também permitiu reduzir a complexidade do modelo, mantendo apenas os atributos mais relevantes.

5. Próximos passos

Como próximos passos, recomenda-se aplicar uma regressão múltipla utilizando apenas os atributos selecionados pelo modelo Lasso, testar diferentes valores do hiperparâmetro alpha para otimizar os modelos Ridge e Lasso, e explorar algoritmos mais robustos como Árvores de Regressão, Random Forest ou XGBoost para comparar o desempenho. Além disso, é importante realizar uma análise de resíduos para validar os pressupostos dos modelos ajustados e considerar a criação de novas variáveis (feature engineering), como interações ou transformações não lineares, com o objetivo de melhorar a capacidade preditiva do modelo.