Exercício prático de experimentação

Turma 33B - Grupo 1

Alunos:

Davi Barros - 2311009

Eduardo Eugênio - 2310822

Gustavo Braga - 2311958

Michela Sgrosso - 2421066

1. Introdução

Este experimento tem como objetivo analisar e prever as vendas de cadeirinhas infantis utilizando diferentes modelos de regressão, conforme visto durante o curso. Foi escolhido um dataset contendo dados de vendas em 400 lojas, com variáveis relevantes, como preço da concorrência, renda média da comunidade, orçamento de publicidade, qualidade da localização dos produtos na loja e fatores demográficos da população local.

O objetivo é construir modelos de regressão para prever o número de vendas unitárias de cadeirinhas infantis em diferentes locais, buscando identificar padrões e variáveis que impactam as vendas. Para isso, foram planejados diferentes cenários que exploram variações nas etapas de pré-processamento dos dados e na aplicação de modelos como regressão linear, árvores de regressão e KNN (k-nearest neighbors). Cada modelo foi escolhido por sua abordagem distinta de previsão, o que permite avaliar qual é mais adequado para os dados e para o problema proposto: enquanto a regressão linear ajuda a captar relações lineares entre as variáveis, as árvores de regressão lidam melhor com dados complexos e não-lineares, e o KNN avalia similaridades entre lojas, utilizando características comuns para prever vendas.

Inicialmente, foi planejada uma variação no pré-processamento dos dados, no que diz respeito a manter ou remover outliers, para verificar como essa decisão impacta o desempenho dos modelos. A ideia é entender quais práticas de preparação e configuração dos modelos melhoram a capacidade de previsão, proporcionando insights sobre as etapas mais importantes para esse tipo de análise de vendas.

Cada cenário será avaliado com métricas específicas, como a soma dos erros quadráticos (SQE), a raiz do erro quadrático médio (RMSE), e o coeficiente de determinação (R^2), que permitem medir a precisão e a qualidade das previsões. A comparação dos resultados entre os diferentes cenários e modelos ajudará a identificar a configuração e o modelo que melhor se adaptam ao objetivo de prever as vendas de cadeirinhas infantis, proporcionando uma base sólida para decisões futuras.

2. Plano de experimentação

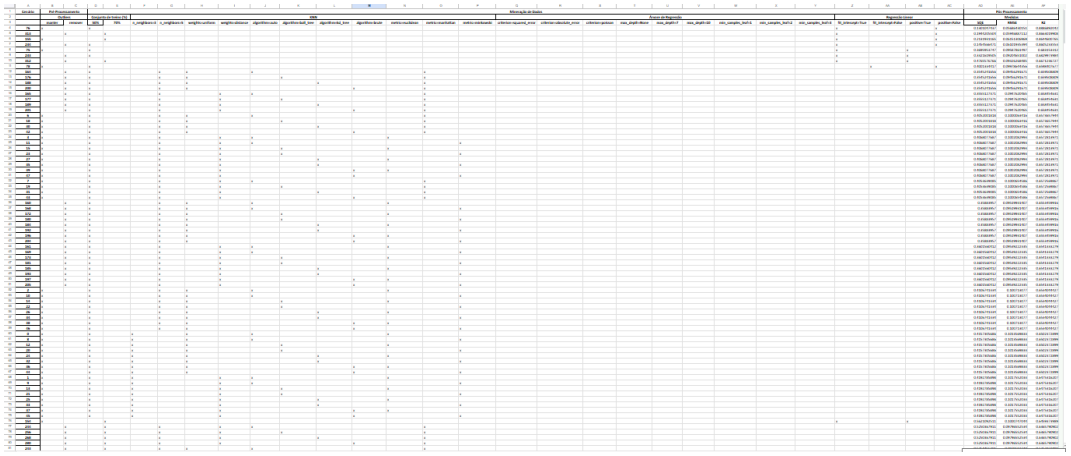
Para estruturar o plano de experimentação, foram definidos cenários que combinam as variações mencionadas de pré-processamento e os diferentes modelos de regressão. A decisão de manter ou remover outliers, por exemplo, pode ter um impacto significativo nos resultados, uma vez que outliers podem distorcer a média em modelos de regressão linear ou impactar a definição de vizinhos próximos no KNN. Por outro lado, manter esses valores pode ser útil para capturar padrões reais de variabilidade no comportamento das vendas, especialmente em regiões ou lojas com características específicas que se afastam da média.

Cada modelo será testado sob diferentes configurações, incluindo ajustes nos parâmetros, para identificar qual abordagem oferece previsões mais precisas e robustas.

Para garantir uma avaliação rigorosa e identificar as melhores configurações de cada modelo, utilizaremos o GridSearchCV para treinar e otimizar os parâmetros especificados em cada cenário. Essa abordagem nos permite explorar sistematicamente diferentes combinações de parâmetros e selecionar aquelas que maximizam o desempenho do modelo. Durante o processo de treino, os dados de treinamento serão divididos em diferentes proporções – 50% e 70% – permitindo observar o impacto do tamanho dos conjuntos de treino nos resultados.

Essa divisão possibilita entender melhor o comportamento dos modelos com quantidades distintas de dados, simulando tanto cenários de dados mais limitados quanto aqueles em que há mais informações disponíveis para aprendizado. A combinação do uso do GridSearchCV com essa variação nas divisões de treino nos ajuda a identificar as configurações de parâmetros e a quantidade de dados mais adequada para o problema em questão.

Com o uso das métricas de avaliação (MSE, RMSE e R^2), cada cenário será analisado em profundidade para identificar os pontos fortes e limitações de cada configuração. Essas métricas oferecem uma visão abrangente: MSE e RMSE quantificam o erro médio nas previsões e o R^2 avalia a capacidade explicativa do modelo. A seguir, apresenta-se o plano de experimentação com as combinações de pré-processamento, modelos e parâmetros que serão exploradas no estudo.



Em particular, cada cenário varia com base nas seguintes variações:

- Pré-Processamento:
  - Manter outliers
  - Remover outliers

- Conjunto de treino:
  - 50%
  - 70%

- Parâmetros do KNN:
  - N_neighbors (3/ 5)
  - Weights (uniform/ distance)
  - Algorithm (auto/ ball_tree/ kd_tree/ brute)
  - Metric (euclidean/ manhattan/ minkowski)

- Parâmetros das árvores de regressão:
  - Criterion (squared_error/ absolute_error/ poisson)
  - Max_depth (none/ 7/ 10)
  - Min_sample_leaf (1/ 2/ 3)

- Parâmetros da regressão linear:
  - Fit_intercept (true/ false)
  - Positive (true/ false)

3. Resultados da Experimentação

No geral, ao realizar uma análise empírica sobre os cenários testados, levando como foco principal as métricas de avaliação dos resultados, é possível perceber que cada um dos modelos tiveram um desempenho médio bastante diferente entre si. Isto é relevante pois nos indica que para o problema de negócio escolhido existe um modelo que em geral tem um resultado médio melhor que os outros. Em particular, em nossos testes, os cenários com modelos de Regressão Linear tiveram um desempenho melhor do que os cenários com modelos KNN, que por sua vez foram melhores do que os cenários com modelos de Árvore de Regressão. Houveram exceções, onde os piores cenários de um certo modelo foram piores do que os melhores de outro, mas estes casos não foram suficientes para mostrar uma performance média equiparada entre os diferentes tipos de modelo.

Para realizar esta análise, decidimos ordenar a tabela por R^2 score, pois é uma métrica que nos traz uma análise mais compreensiva do que a soma dos erros quadráticos ou a raiz do erro quadrático médio. Outra técnica utilizada para esta identificação foi de colorir os valores das métricas de acordo com um gradiente, para se identificar diferenças entre os cenários mais rapidamente.

Em nossos cenários, foi evidente que os conjuntos de teste com 50% dos dados ocasionaram em desempenhos melhores do que os conjuntos de teste com 70%. Isto é diferente do que se esperaria e pode ser atribuído a algumas causas possíveis, como o conjunto específico selecionado ter ao acaso tido uma seleção de entradas que seria mais facilmente linearizável do que o conjunto maior. Por outro lado, não foi claro o efeito causado por manter ou remover os outliers, pois alguns modelos desempenharam melhor com outliers mantidos, e outros desempenharam melhor com outliers removidos. Porém, a remoção ou não de outliers teve no geral um baixo impacto no resultado final de cada cenário.

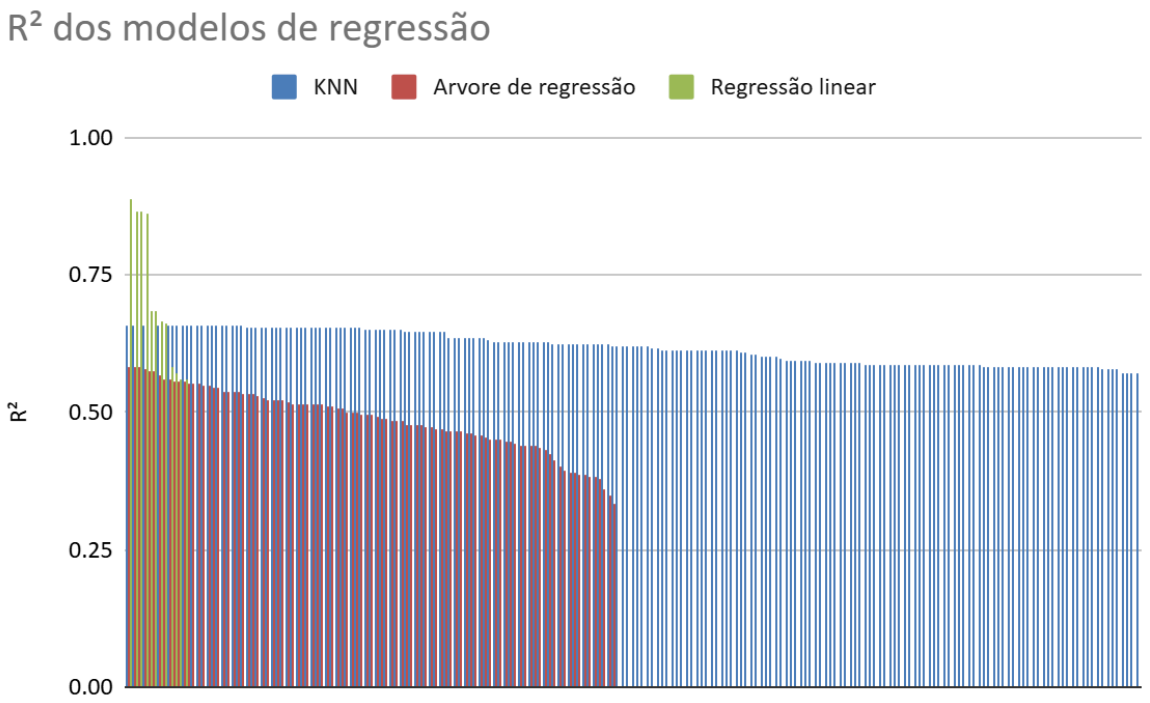
No caso da Regressão Linear, o parâmetro que mais afetou o desempenho foi o fit_intercept, pois quase todos os cenários com fit_intercept=True apareceram no topo da tabela de experimentação. O melhor cenário usando Regressão Linear tem os seguintes parâmetros: fit_intercept=True, positive=False.

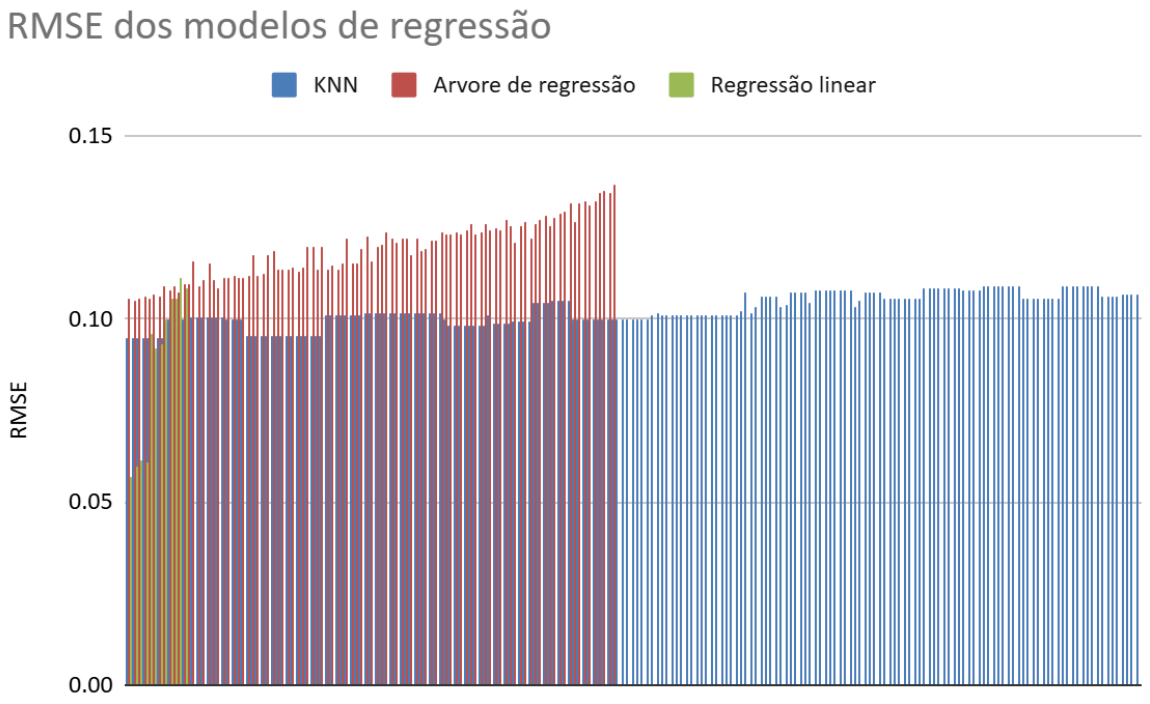
O parâmetro que trouxe o melhor desempenho para o KNN foi o n_neighbors, pois os casos com n_neighbors=5 tendem a desempenhar melhor do que os com n_neighbors=3. O melhor cenário do KNN possui os seguintes parâmetros: n_neighbors=5, weights=uniform, algorithm=auto, metric=manhattan.

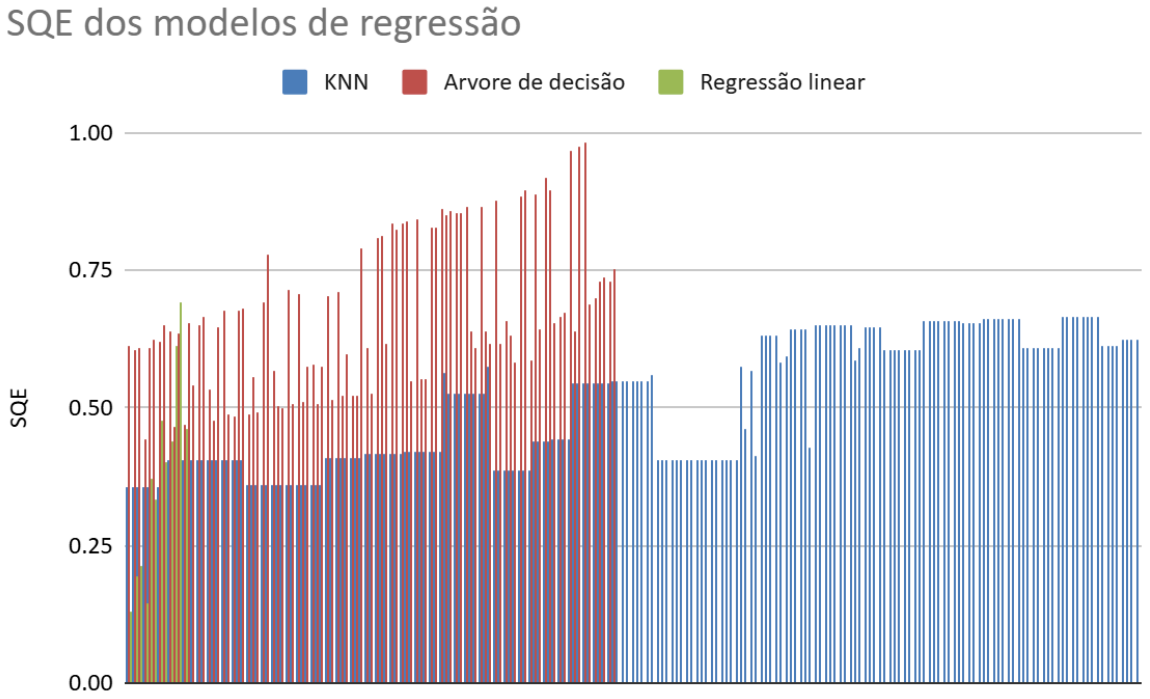
Em seguida, percebemos que não é evidente qual parâmetro da Árvore de Regressão causa a maior diferença no desempenho do modelo, em média. Tendo em mente que este modelo foi o de pior desempenho em geral, encontrar seu parâmetro mais significativo não é de grande importância. O melhor cenário com um modelo de Árvore de Regressão teve os seguintes parâmetros: criterion=squared_error, max_depth=7, min_samples_leaf=3.

Por fim, torna-se relevante a análise do total das métricas de avaliação para todos os cenários. Abaixo, temos os gráficos que condensam essa informação em uma única visualização, e em

anexo no envio temos a planilha com os valores exatos, além da especificação de quais foram os cenários exatos testados e seus resultados respectivos.







Então, ao fazer uma análise do total de cenários explorados, pode-se perceber um padrão curioso nos testes. Os cenários nos quais foi utilizada a Regressão Linear com fit_intercept=True e positive=False, as métricas de avaliação foram significativamente melhores do que em qualquer outro cenário do experimento. Todos os cenários nos quais isto foi verdade tiveram performance ao menos 26.0% melhor do que qualquer cenário sem estas condições. Porém, ao fazer uma análise mais holística, se torna evidente que o parâmetro que mais fez diferença foi o tamanho do conjunto de treino, que performa significativamente melhor com tamanho de 50%.

4. Conclusão

Após ordenar os cenários pela métrica R2, a análise dos resultados permitiu identificar que o melhor modelo foi aquele baseado em Regressão Linear, especificamente no cenário em que os parâmetros foram configurados como fit_intercept=True e positive=False. Este cenário não apenas obteve o maior score R^2, refletindo uma capacidade explicativa superior em relação aos dados de vendas de cadeirinhas infantis, mas também se destacou pela estabilidade em diversas variações de teste.

Do ponto de vista técnico, a escolha da Regressão Linear como o modelo mais eficaz se justifica pela sua simplicidade e pela capacidade de fornecer previsões lineares diretas, que se mostraram adequadas para o nosso conjunto de dados. A presença do parâmetro fit_intercept=True indica que o modelo consegue capturar adequadamente a relação entre as variáveis independentes e a variável dependente, permitindo uma interpretação mais clara dos coeficientes.

Sob a perspectiva do negócio, a escolha desse modelo é igualmente relevante. O entendimento de como diferentes fatores, como preço, publicidade e localização, influenciam as vendas é crucial para a tomada de decisões estratégicas. Um modelo que explica bem essas relações pode guiar os gestores na alocação eficiente de recursos de marketing e na definição de preços, impactando diretamente o aumento das vendas. Portanto, o cenário com a Regressão Linear não só demonstrou desempenho superior em termos de métricas, mas também oferece um suporte valioso para estratégias de negócios fundamentadas em dados.