Prática em Ciência de Dados 3 - Projeto 1

Maria Victória Brandão Barros - 12608692
Lucas Szavara Nusp 12690087

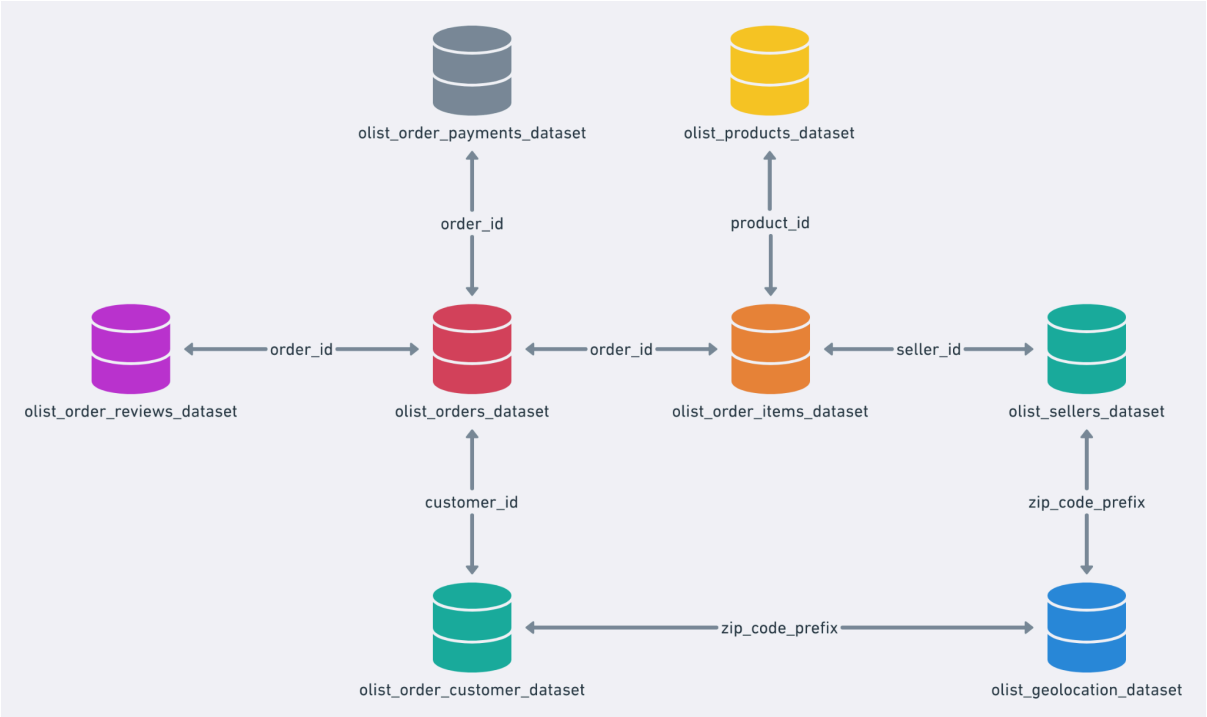
1. Introdução

As avaliações de um produto é um importante indicador para a decisão de compra do consumidor (Jiménez e Mendoza (2013)). De acordo com a pesquisa “E-commerce Trends 2024”, realizada pela Octadesk em parceria com o Opinion Box, 62% dos consumidores fazem de duas a cinco compras online por mês, enquanto 85% dos brasileiros fazem pelo menos uma compra por mês na internet. Além disso, segundo a Forbes, 54% dos respondentes desta pesquisa pretendiam aumentar a frequência de compras online. Neste projeto aplicamos KNN, Naive-Bayes, floresta aleatória e regressão ordinal, conjuntamente com técnicas exploratórias para analisar um conjunto de dados pré estabelecido sobre o e-commerce em uma plataforma, com a finalidade de responder questões sobre o comportamento dos consumidores em compras online, além de explorar os principais fatores que contribuem para uma avaliação positiva.

2. Material e Métodos

2.1. Sobre os dados

O conjunto de dados trabalhado apresenta informações sobre 100 mil pedidos realizados na loja on-line Olist entre os anos de 2016 e 2018. Existem ao todo 9 tabelas com dados variando de tipo de item do pedido, satisfação do cliente, localização do destinatário, forma de pagamento, entre outros. As tabelas se relacionam da seguinte forma:



A tabela product_category_name_translation não se relaciona com nenhuma outra tabela. Os dados podem ser acessados em https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce/data

2.2. Exploração e pré-processamento

Para as análises exploratórias nenhum tratamento foi realizado. Já para os modelos, as tabelas com várias observações para um único pedido foram agrupadas, com as variáveis categóricas sendo transformadas em variáveis indicadoras para cada categoria. Também foram criadas variáveis indicando se o pedido foi atrasado, e de quantos dias foi o atraso. Após as transformações realizadas, as variáveis utilizadas foram:

- Número de pagamentos utilizados
- Número médio das parcelas do pedido
- Total do pagamento
- Indicador de atraso no pedido
- Tempo de entrega
- Interação entre tempo de entrega e Indicador de atraso
- Dias de atraso (0 caso o pedido tenha sido entregue na data correta)
- Valor mensal médio a ser pago
- Indicador de pedido cancelado
- Total de produtos distintos no pedido
- Total de vendedores distintos no pedido
- Porcentagem de produtos vendidos por vendedores do mesmo estado que o cliente
- Comprimento médio do nome dos produtos
- Comprimento médio da descrição dos produtos
- Número médio de fotos dos produtos
- Peso médio dos produtos
- Comprimento médio dos produtos
- Altura média dos produtos
- Largura média dos produtos
- Preço total dos produtos
- Frete total dos produtos
- Porcentagem dos pagamentos feitos com cada tipo de pagamento
- Porcentagem das categorias dos produtos
- Estado do cliente

Os dados foram padronizados apenas para o método KNN, por ser o único baseado em distância. Após isso, covariáveis linearmente dependentes foram removidas, as únicas colunas removidas por isso foi uma indicadora do estado do cliente e uma coluna constante, se referindo a média de pagamentos feitos por modos não definidos.

2.3. Implementação

Os gráficos foram feitos usando matplotlib, seaborn e pingouin. Os modelos de regressão ordinal por variável latente foram feitos usando statsmodels em conjunto com o scipy, outros modelos foram feitos pelo sklearn. Os resíduos substitutos foram implementados com base nas distribuições de probabilidade do scipy.

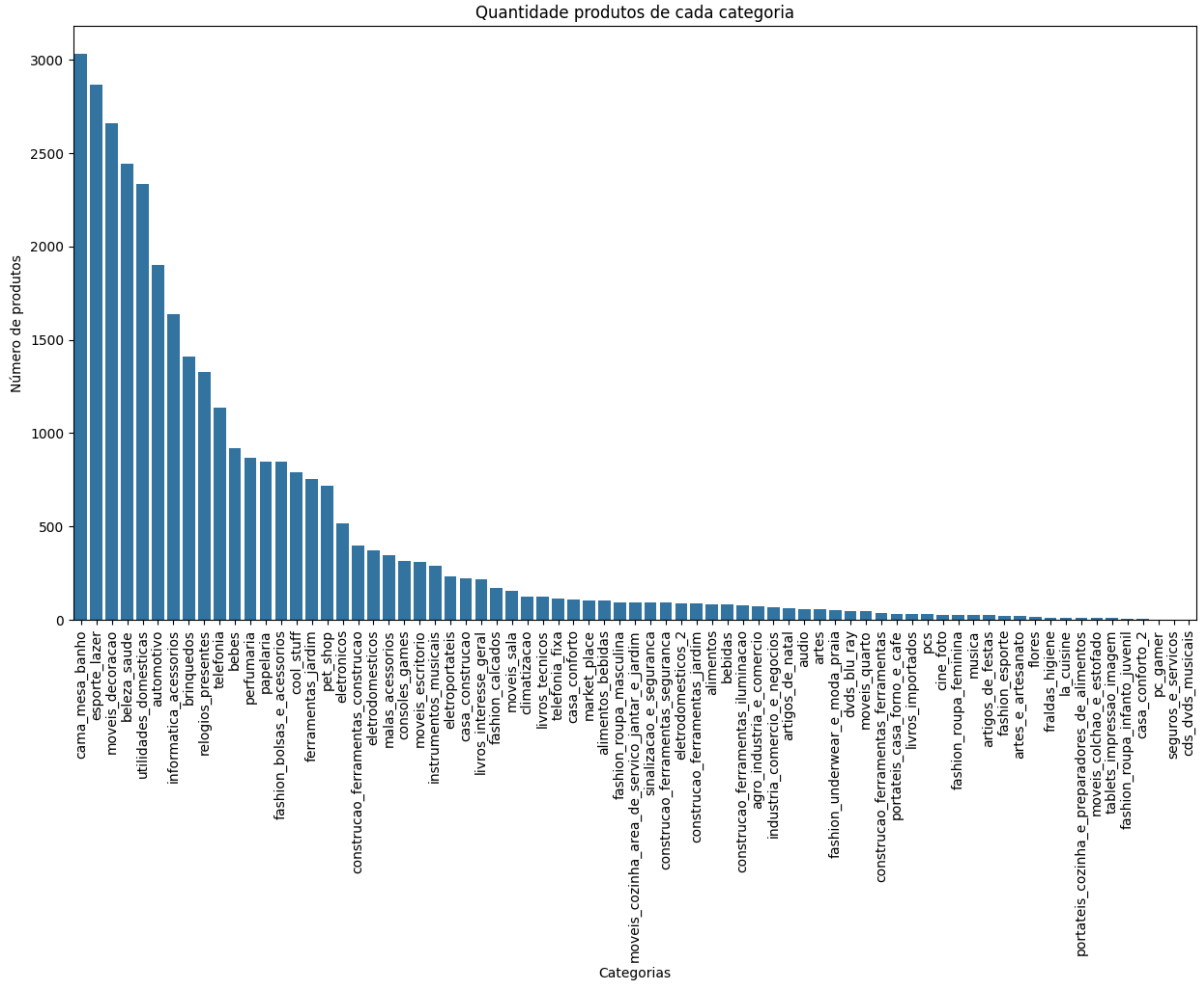
3. Questões propostas

3.1. Quantas categorias de produtos foram comercializadas e quais os atributos mais relevantes?

Ao todo a base de dados conta com 73 diferentes categorias de produtos. As categorias de produtos mais relevantes serão discutidas na seção de técnicas de machine learning.

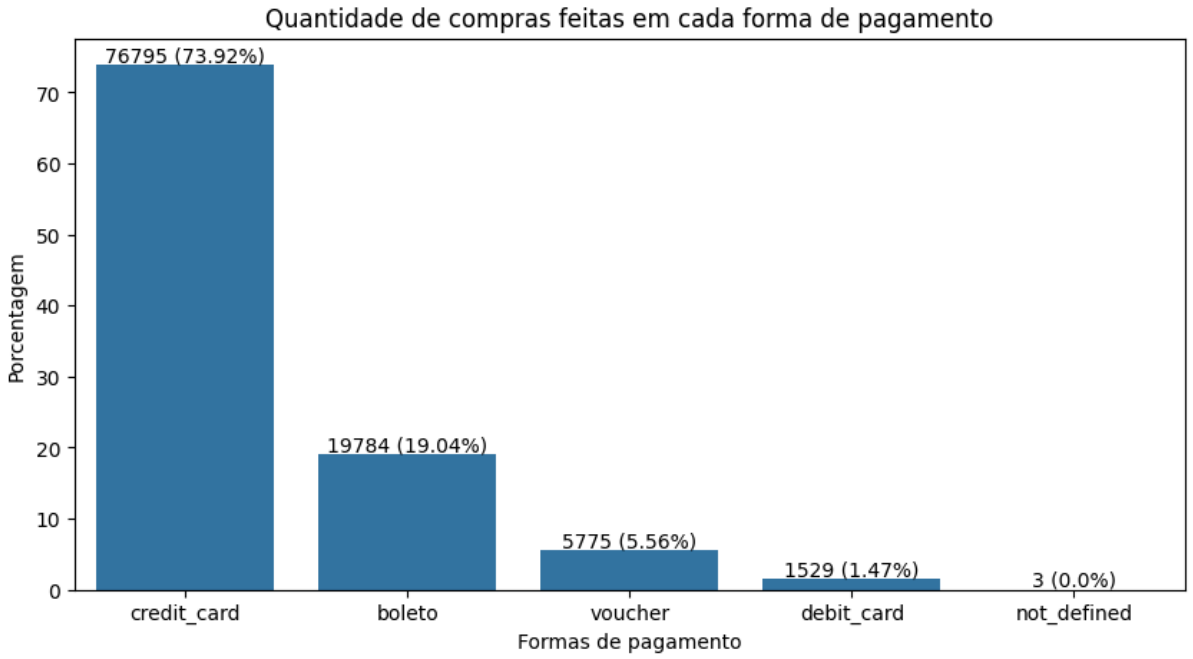
3.2. Qual a categoria de produtos mais vendida?

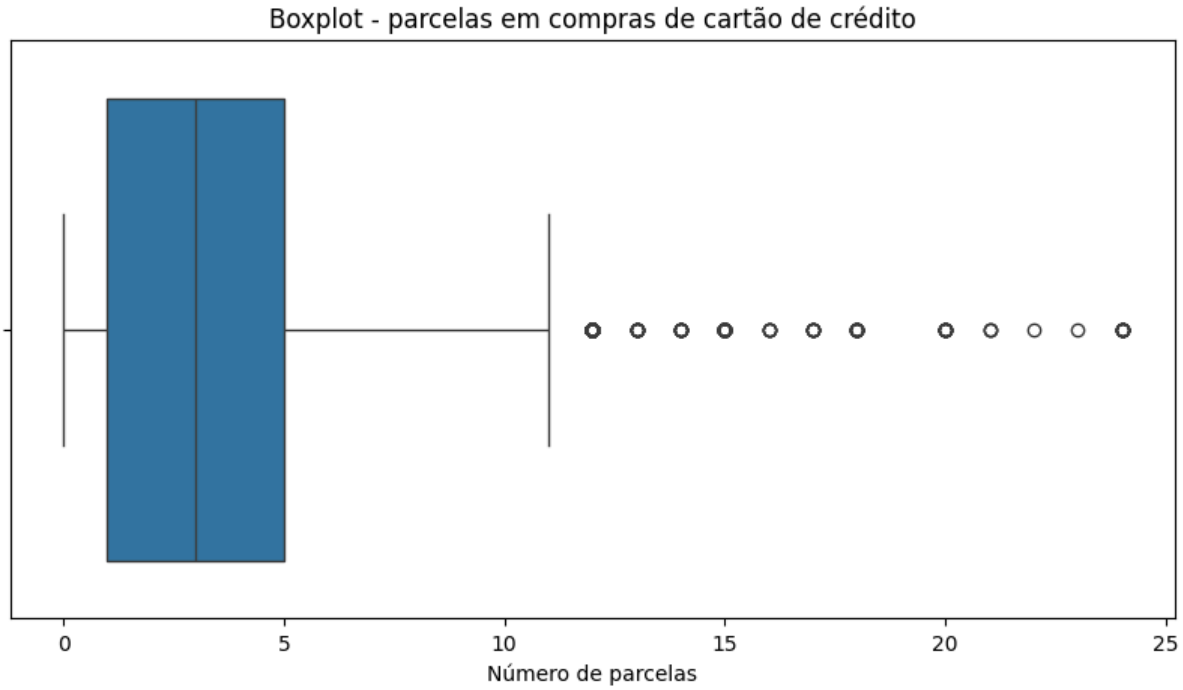
A categoria mais vendida é a “cama, mesa e banho”, com 11.115 produtos listados na base de dados.



3.3. Qual a porcentagem de compras feitas com cartão de crédito? Neste caso, qual é o número médio de parcelas?

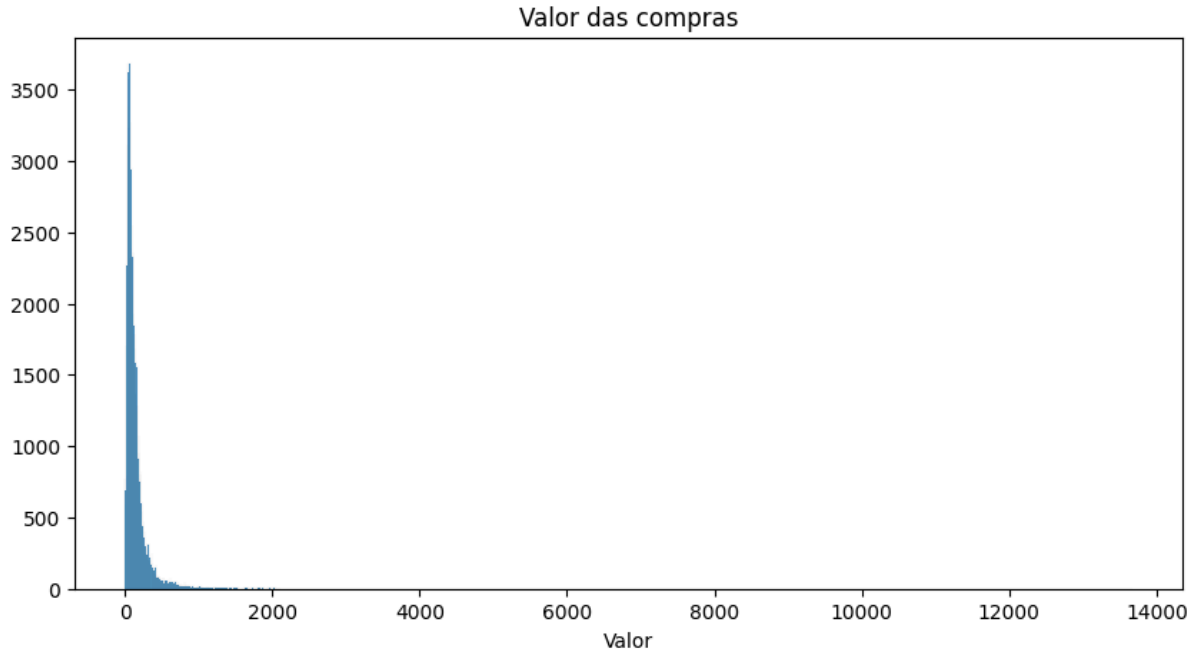
73,2% das compras foram feitas com cartão de crédito, com um número médio de 3,5 parcelas.





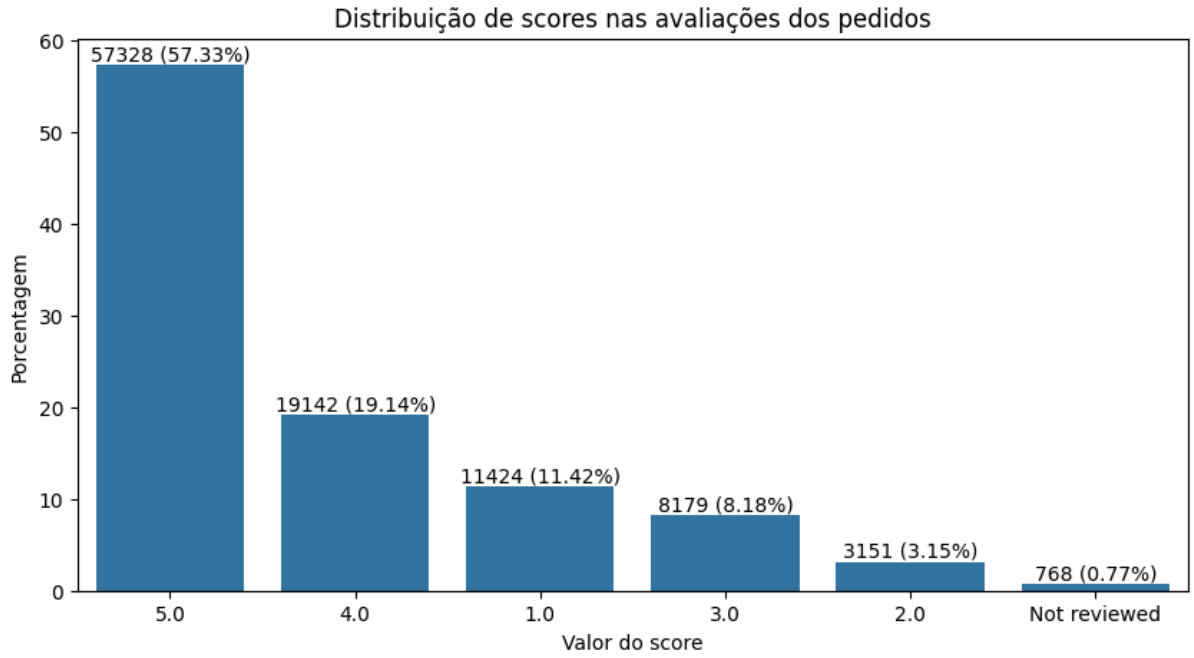
3.4. O valor das compras fica em média em torno de qual valor em reais?

O valor médio das compras é de 154,10 reais.



3.5. Qual a porcentagem de clientes satisfeitos com a realização da compra? (review score = 4 ou 5)

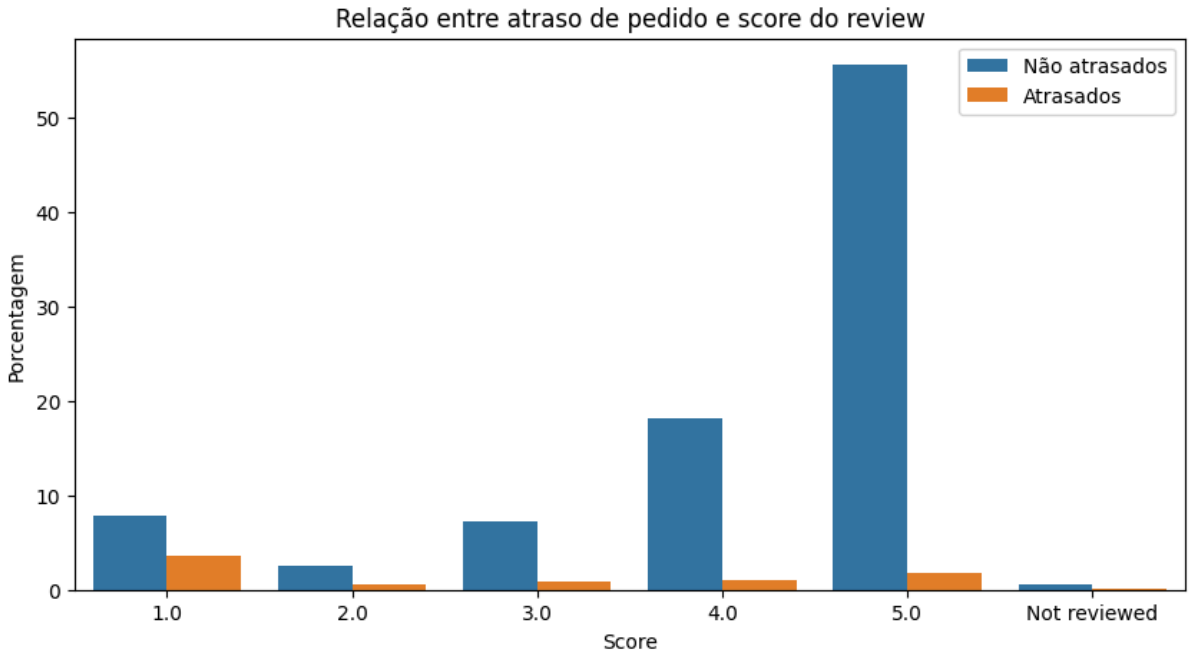
76,47% dos clientes classificaram seus pedidos como 4 ou 5 estrelas.



3.6. Existe relação entre o tempo de entrega e avaliação feita pelo cliente?

É possível perceber uma relação entre o pedido chegar ou não atrasado e a avaliação do cliente. Das avaliações de nota 1, uma porcentagem muito grande são pedidos que chegaram atrasados, enquanto

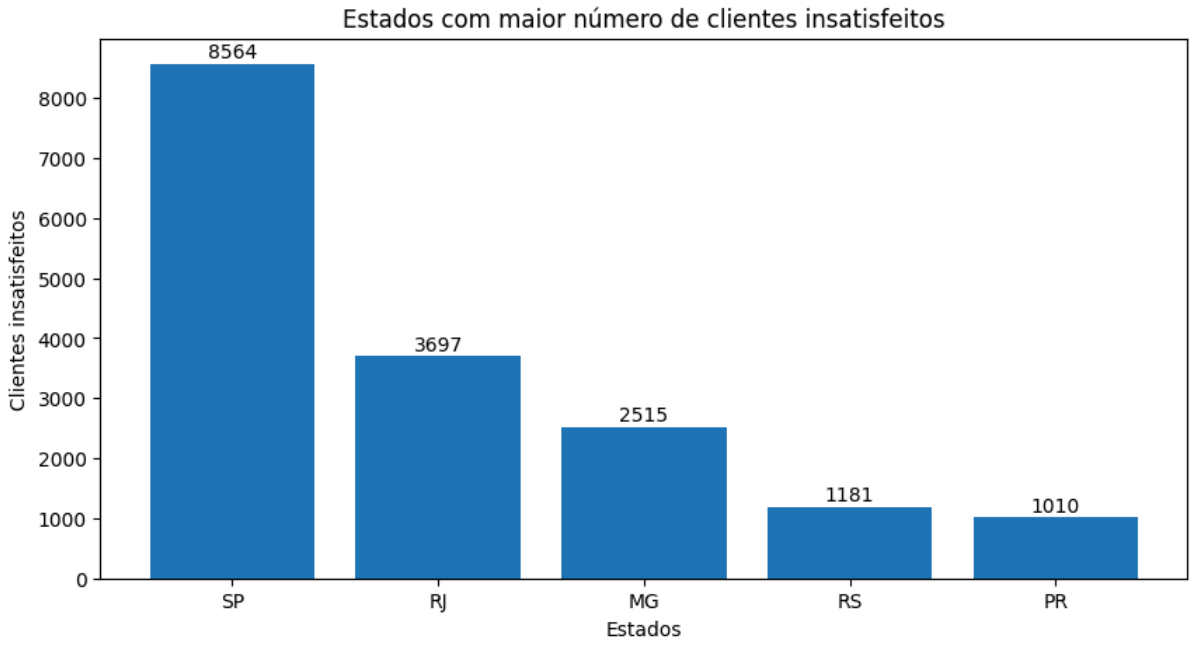
avaliações de notas maiores possuem uma proporção menor de pedidos com atraso.



3.7. Quais estados do Brasil possuem mais clientes insatisfeitos?

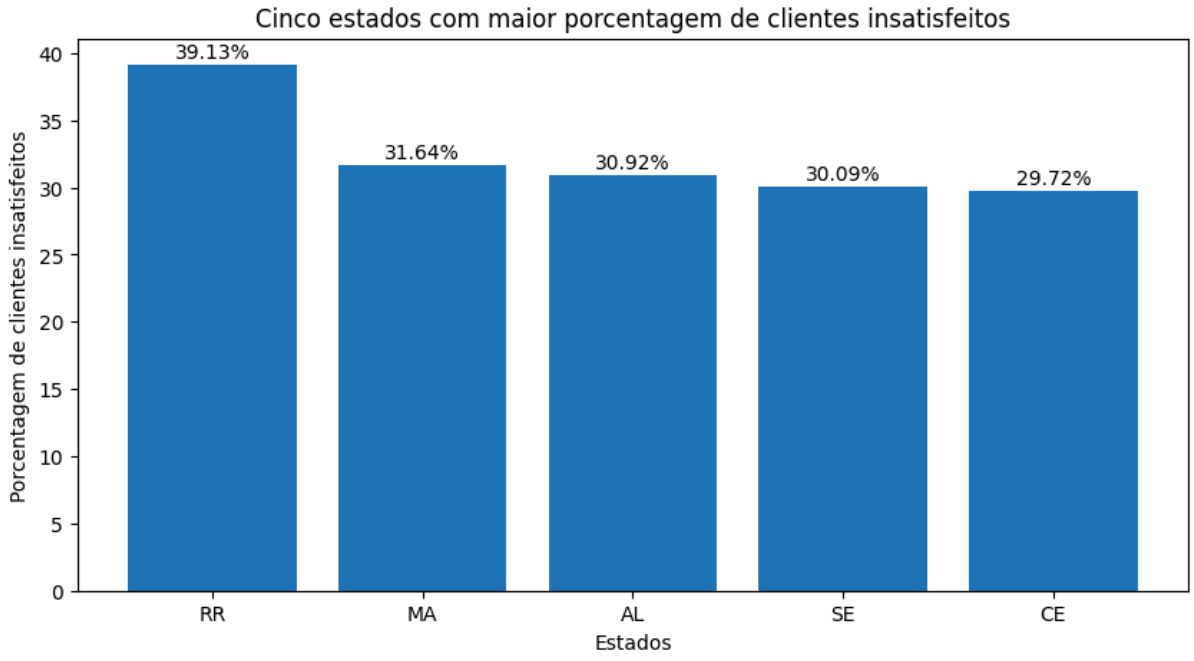
Em número absoluto, os estados com maior número de clientes insatisfeitos (review score menor que 4) são:

- São Paulo - 8.564
- Rio de Janeiro - 3.697
- Minas Gerais - 2.515
- Rio Grande do Sul - 1.181
- Paraná - 1.010



Quando levamos em consideração a quantidade de pedidos de cada estado, os estados com maior proporção de clientes insatisfeitos são:

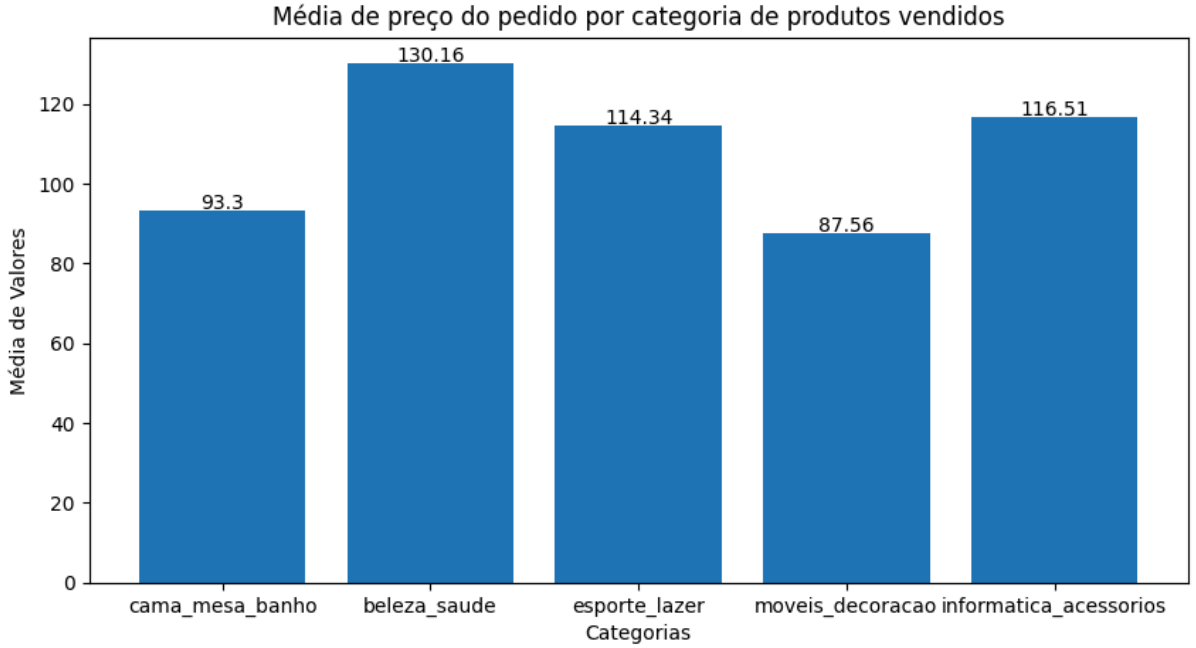
- Roraima - 39,13%
- Maranhão - 31,63%
- Alagoas - 30,91%
- Sergipe - 30,08%
- Ceará - 29,72%



3.8. Média do preço do pedido por categoria de produtos vendidos

A média de valor de cada pedido por categoria de produto, considerando as cinco categorias mais vendidas, é:

- Cama, mesa e banho - 93,3 reais
- Beleza e saúde - 130,16 reais
- Esporte e lazer - 114,34 reais
- Móveis e decoração - 87,56 reais
- Informática e acessórios - 116,51 reais



4. Técnicas de Machine Learning

Essa seção do relatório é dedicada às técnicas de machine learning desenvolvidas durante o projeto. As técnicas têm o objetivo de prever a nota de um dado pedido de acordo com as covariáveis estabelecidas, sendo elas:

Os dados foram separados em 90% para treino e 10% para teste. Foram utilizadas quatro diferentes técnicas para predição, e os resultados obtidos estão listados a seguir.

4.1. Naive Bayes

O classificador de Bayes ingênuo, supondo normalidade nas covariáveis contínuas, por não ser muito robusto para problemas complexos, atingiu uma acurácia de apenas 57%.

4.2. KNN

Utilizando o Grid Search CV, o melhor valor para K numa faixa de 30 a 35 é 33. Com esse valor, a acurácia do modelo é 61%

4.3. Random Forest

Os melhores valores dos parâmetros encontrados pelo Grid Search CV foram:

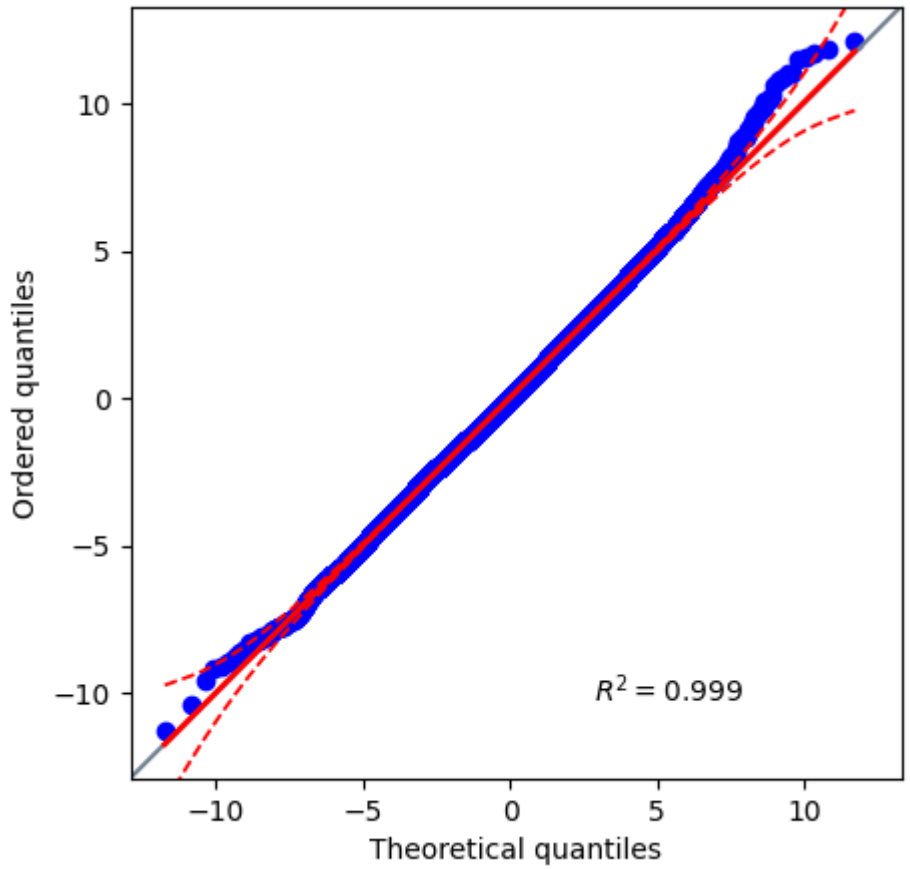
- Profundidade máxima: 10 (de 1 a 16)
- Mínimo de amostras no nó-folha: 1 (de 1 a 5)
- Mínimo de amostras para dividir um nó: 2 (de 2 a 5)
- Número de árvores: 120 (de 100 a 200)

O valor da acurácia para esses parâmetros é de 62%.

4.4. Regressão Ordinal por variável latente

Foi testado usando variáveis das distribuições logística, normal e gumbel (gumbel_r no scipy). Todas tiveram resultados muito similares, pela facilidade de interpretação, foi preferível a logística.

A verificação desses modelos foi feita em duas etapas. Primeiramente analisamos os resíduos substitutos (Liu e Zhang (2018)), usando o gráfico quantil-quantil e um gráfico de pontos para algumas covariáveis:



Apesar de apresentar algumas variações do esperado, essas divergências são pequenas ou explicadas pela quantidade de observações com determinados valores nas covariáveis, isto é, é esperado uma maior presença de resíduos outliers

nos valores mais comuns de uma covariável. Após essa etapa, calculamos as métricas de predição do modelo, obtendo:

- Acurácia de 61,09%
- MSE de 1.35

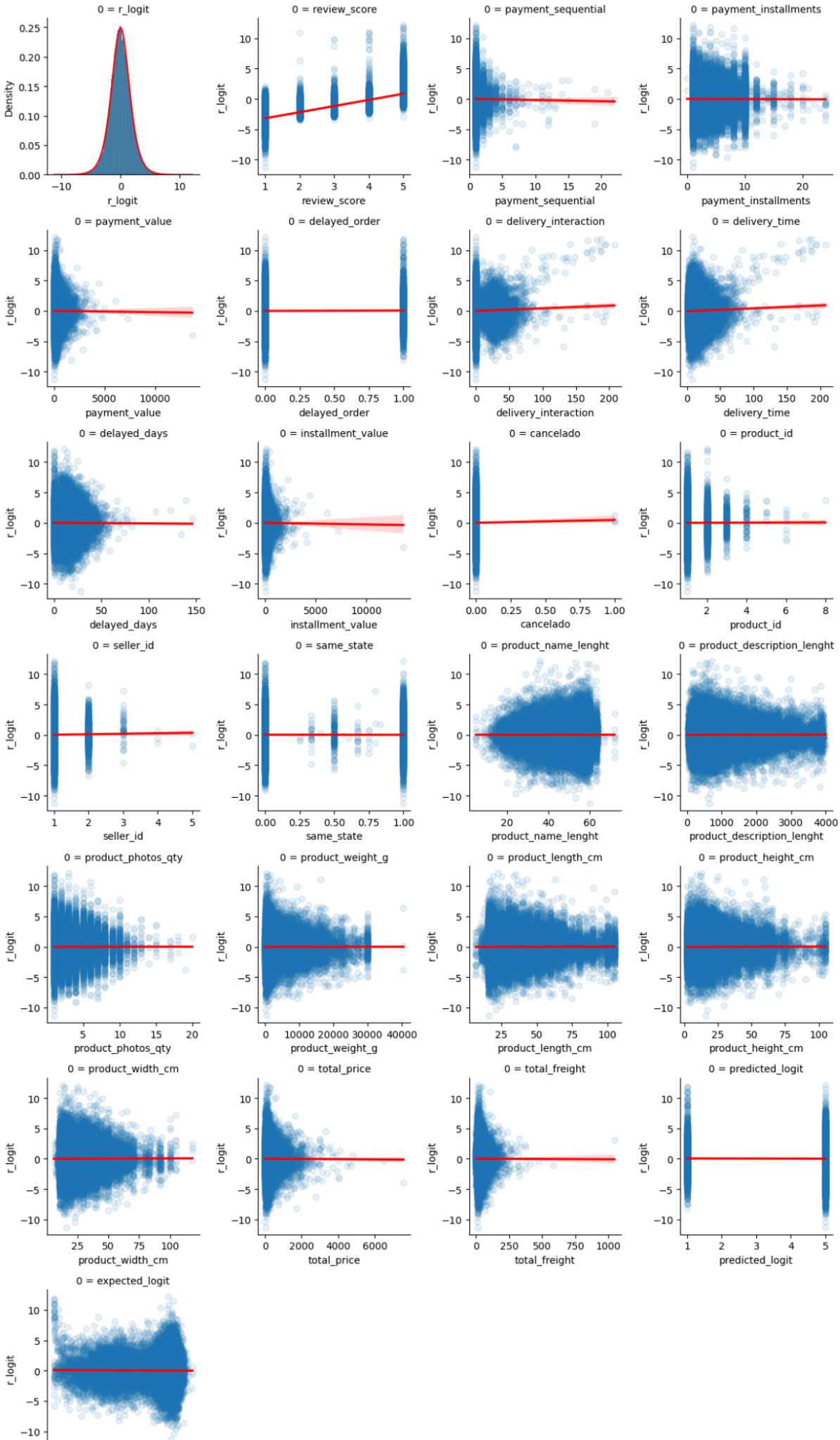
Dessa forma temos um modelo com resultados muito próximos do melhor modelo preditivo observado, mas com uma verificação de ajuste mais completa e com maior interpretabilidade dos parâmetros, permitindo identificar as covariáveis com maior peso no modelo, desconsiderando as covariáveis com valor P abaixo de 0,05:

Covariável	**Coeficiente**	**RC**	**Valor P**
Porcentagem de produtos da categoria Livros técnicos comparado com Utilidades domésticas	0.506331	1.659192	<0,001
Porcentagem de produtos da categoria Livros de interesse geral com Utilidades domésticas	0.702292	2.018373	<0,001
Número de vendedores	-1.455406	0.233306	<0,001
Pedido atrasado	-1.065513	0.344551	<0,001
Cliente mora no Amazonas, comparado com clientes do Tocantins	0.553305	1.738991	0,001

Em que RC representa a razão de chances da avaliação do cliente ser maior que determinado valor, por exemplo, dado um pedido i, de um cliente no Tocantins. O mesmo pedido, para o Amazonas, considerando todas as outras covariáveis constantes, terá chance 1.738991 maior de ser avaliado com pelo menos a nota j. Matematicamente, seja X_i as covariáveis de um pedido para um cliente no Tocantins e X_m um pedido similar, mas para ser entregue no Amazonas. Então:

$$\frac{P(Y \geq j|X_m)}{P(Y<j|X_m)} = 1.738991 * \frac{P(Y \geq j|X_i)}{P(Y<j|X_i)}$$

Em outras palavras, se o pedido no Tocantins tem 2 avaliações pelo menos iguais a 3 para cada 1 avaliação abaixo de 3, o mesmo pedido no Amazonas terá 1.738991 * 2 avaliações iguais ou melhores que 3 para cada avaliação pior do que 3.


0 = r_logit
Density
r_logit
0 = review_score
review_score
0 = payment_sequential
payment_sequential
0 = payment_installments
payment_installments
0 = payment_value
payment_value
0 = delayed_order
delayed_order
0 = delivery_interaction
delivery_interaction
0 = delivery_time
delivery_time
0 = delayed_days
delayed_days
0 = installment_value
installment_value
0 = cancelado
cancelado
0 = product_id
product_id
0 = seller_id
seller_id
0 = same_state
same_state
0 = product_name_lenght
product_name_lenght
0 = product_description_lenght
product_description_lenght
0 = product_photos_qty
product_photos_qty
0 = product_weight_g
product_weight_g
0 = product_length_cm
product_length_cm
0 = product_height_cm
product_height_cm
0 = product_width_cm
product_width_cm
0 = total_price
total_price
0 = total_freight
total_freight
0 = predicted_logit
predicted_logit
0 = expected_logit

5. Conclusão

A base de dados utilizada nesse projeto, apesar de complexa, apresenta diversas informações relevantes para o cenário do e-commerce brasileiro, principalmente quando levamos em consideração o número de variáveis e a quantidade de observações presentes. A partir da análise exploratória realizada é possível notar vários padrões do consumidor médio como, por exemplo, a preferência por compras de baixo valor monetário e pagamento no cartão de crédito, além da clara relação entre o tempo de espera pelo produto e o nível de satisfação do cliente. A partir dessas informações, o vendedor pode tomar medidas que atraiam o consumidor e, consequentemente, aumentam as vendas. Em um estudo mais aprofundado seria também possível reunir padrões específicos de cada região do país, visto que a base conta com dados geográficos dos clientes.

Considerando todos os fatores dentro do controle do vendedor, uma estratégia que pode aumentar suas avaliações seria realizar entregas mais rapidamente e com previsões mais assertivas, de produtos com menor valor. Além disso, pedidos pagos com maior porcentagem de voucher costumam ter avaliações piores que outros métodos de pagamento, enquanto pedidos pagos com cartão de débito e boleto tem avaliações melhores, é difícil entretanto afirmar que essa relação é causal e pode ter relação com os motivos pelos quais os clientes escolhem cada método, por exemplo, clientes com voucher já estão frustrados com a loja, enquanto clientes que pagam com boleto ou cartão de débito podem estar comprando produtos que impactam menos na economia doméstica, logo, se sentindo menos frustrados com defeitos ou outros problemas.

Por último, as técnicas de machine learning aplicadas, apesar de não apresentarem acurácia muito elevada, são interessantes para avaliar alguns outros aspectos do problema. O algoritmo de regressão utilizado, por exemplo, nos permite identificar quais são as variáveis mais impactantes para determinar o valor do score da avaliação.

6. Trabalhos relacionados

- Zhao, F. and Liu, H. (2023), "Modeling customer satisfaction and revisit intention from online restaurant reviews: an attribute-level analysis", Industrial Management & Data Systems, Vol. 123 No. 5, pp. 1548-1568. https://doi.org/10.1108/IMDS-09-2022-0570
    - Esse trabalho modela avaliações online de restaurantes e separa em dois fatores: satisfação do consumidor e intenção de revisitar o estabelecimento, usando uma floresta aleatória, um modelo LightGBM e um modelo de bagging para redes neurais.

- Bi, J. W., Liu, Y., Fan, Z. P., & Cambria, E. (2019). Modelling customer satisfaction from online reviews using ensemble neural network and

effect-based Kano model. International Journal of Production Research, 57(22), 7068–7088. https://doi.org/10.1080/00207543.2019.1574989
    - Esse trabalho propõe um conjunto de modelos de rede neural para identificar dimensões da satisfação do consumidor. Esse conjunto então é avaliado em dois datasets, um de celulares e outro de câmeras
- Yang, X., Yang, G., Wu, J., Dang, Y., Fan, W. (2021) Modeling relationships between retail prices and consumer reviews: A machine discovery approach and comprehensive evaluations. Decision Support Systems. https://doi.org/10.1016/j.dss.2021.113536
    - Esse trabalho analisa avaliações de consumidores de acordo com o valor de itens dentro de uma categoria.
- Peng, L., Cui, G., Chung, Y. et al. A multi-facet item response theory approach to improve customer satisfaction using online product ratings. J. of the Acad. Mark. Sci. 47, 960–976 (2019). https://doi.org/10.1007/s11747-019-00662-w
    - Esse trabalho analisa avaliações de consumidores em uma plataforma de cerveja e em uma plataforma de hotelaria usando um modelo de teoria de resposta ao item, a partir da avaliação do consumidor em diversas categorias.

Comparado a esses trabalhos, trouxemos uma análise envolvendo múltiplas covariáveis, a partir de fatores de fácil acesso aos lojistas, em diversas categorias de produto.

7. Referências

Liu, D., & Zhang, H. (2018). Residuals and Diagnostics for Ordinal Regression Models: A Surrogate Approach. Journal of the American Statistical Association, 113(522), 845–854. https://doi.org/10.1080/01621459.2017.1292915

Jiménez, F. R., & Mendoza, N. A. (2013). Too Popular to Ignore: The Influence of Online Reviews on Purchase Intentions of Search and Experience Products. Journal of Interactive Marketing, 27(3), 226-235. https://doi.org/10.1016/j.intmar.2013.04.004

Brazilian E-Commerce Public Dataset by Olist https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce