Wu Xin

Filtragem Colaborativa para o Mercado Varejista Digital

Niterói - RJ, Brasil

15 de agosto de 2022

Universidade Federal Fluminense

Wu Xin

Filtragem Colaborativa para o Mercado Varejista Digital

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Prof. Dr. Rafael Santos Erbisti

Niterói - RJ, Brasil

15 de agosto de 2022

Universidade Federal Fluminense

Wu Xin

Filtragem Colaborativa para o Mercado Varejista Digital

Monografia de Projeto Final de Graduação sob o título *"Filtragem Colaborativa para o Mercado Varejista Digital"*, defendida por Wu Xin e aprovada em 15 de agosto de 2022, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Prof. Dr. Rafael Santos Erbisti
Departamento de Estatística – UFF

Profa. Dra. Jessica Quintanilha Kubrusly
Departamento de Estatística – UFF

Profa. Dra. Ana Maria Lima de Farias
Departamento de Estatística – UFF

Niterói, 15 de agosto de 2022

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

X6f Xin, Wu
Filtragem Colaborativa para o Mercado Varejista Digital / Wu Xin ; Rafael Santos Erbisti, orientador. Niterói, 2022.
54 f. : il.

Trabalho de Conclusão de Curso (Graduação em Estatística)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2022.

1. Sistema de Recomendação. 2. Filtragem Colaborativa. 3. Modelos Lineares. 4. LASSO. 5. Produção intelectual. I. Erbisti, Rafael Santos, orientador. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

O mercado de comércio eletrônico está em crescimento. Isso beneficia tanto o consumidor, devido a facilidade de acesso e diversidade de produtos, como também o vendedor, pois, em geral, lojas online demandam custo inicial e capital de giro menores em relação à loja física. Nesse sentido, este trabalho busca investigar o comportamento dos consumidores em diferentes categorias na área de comércio do E-commerce no Brasil através de modelos de Sistema de Recomendação para prever uma possível compra do usuário. Assim, tanto o consumidor pode ser beneficiado por ter melhores experiências de compra quanto o vendedor passa a ter uma probabilidade maior de vendas nos seus produtos. O objetivo deste trabalho é avaliar a capacidade preditiva do sistema de recomendação baseado em modelo utilizando o método de regularização LASSO. O algoritmo proposto foi avaliado tanto em dados simulados quanto na base de dados reais da Olist Store. Em geral, os resultados obtidos não foram satisfatórios. É possível que haja outras formas de fazer recomendações que tenham uma capacidade preditiva mais precisa.

Palavras-chave: Sistema de Recomendação. Filtragem Colaborativa. Modelos Lineares. LASSO.

Agradecimentos

Agradeço a minha família por todo o suporte recebido nessa trajetória. À minha mãe Zhuge Xiaoya pelo apoio incondicional, ao meu pai Zhuge Jingjian e aos meus irmãos Wu Yuanyuan e Zhuge Yu.

A Universidade Federal Fluminense pela infraestrutura e oportunidades na minha formação acadêmica.

A todos os professores da UFF que fizeram parte da minha formação. Em especial, ao meu orientador Rafael Erbisti por todo o apoio, orientação, paciência e disponibilidade para chegar até aqui. Para mim, é o melhor orientador! Muito obrigada por tudo!

Aos meus amigos que conheci durante a graduação, que compartilharam momentos de desespero, de angústia mas também de felicidade. Em especial, a minha melhor dupla Brendha Alves Gomes, como também ao João Pedro Fernandes Martins, Pedro Fernando Santos Vieira Fernandes da Silva, Dayana Gimenes da Silva Ribeiro, Rafael Araujo Couto e Igor da Silva Freitas de Souza.

Aos meus queridos amigos chineses, principalmente, a Qiqi Lou pelo todo o apoio nos momentos mais desesperados, assim como todos os membros do grupo SevenXN, por ser um *break* de todos os meus problemas e fazer parte dos momentos de alegria.

Por fim, a todos que participaram, direta ou indiretamente do desenvolvimento deste trabalho, sendo indispensáveis para o meu processo de aprendizado.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O avanço nos sistemas de informação criou um novo mercado global, no qual permite a realização de atividades econômicas de forma virtual. Essas atividades variam de um simples usuário a grandes empresas, incluindo desde transações de pagamentos via Internet até a comercialização de diversos produtos e serviços. Nesse contexto, surge o *E-commerce*, que é uma abreviação de *electronic commerce*, que se refere às atividades de compra e venda realizadas totalmente online (NAKAMURA, 2001).

Existem vários tipos de *E-commerce*, dentre eles, o *business-to-business* (B2B) que pode ser definido como transações eletrônicas entre empresas, nesse caso, não há participação de cliente como pessoa física. Um exemplo de B2B é o comércio por atacado entre empresas. Já o *business-to-consumer* (B2C) é a relação comercial eletrônica entre empresas e consumidores finais, um dos pioneiros de B2C, por exemplo, é a Amazon. O modelo *consumer-to-consumer* (C2C) corresponde à relação comercial entre os consumidores, isto é, são pessoas físicas que comercializam produtos novos e usados em sites intermediadores (TASSABEHJI, 2003). Os tipos de *E-commerce* de interesse neste trabalho são o B2C e C2C.

Sabe-se que os Estados Unidos (EUA) são pioneiros em *E-commerce*. No final dos anos 90, ocorreu a chamada bolha da Internet (*dot-com*), onde havia rumores de que aqueles que não adotassem o modelo de vendas na Internet seriam eliminados do mercado. Outro fato interessante é que o governo americano não cobrava impostos sobre os produtos vendidos online. Isso incentivou o crescimento do mercado de *E-commerce*.

Em 1995, foi fundada a Amazon nos EUA, que hoje é uma das maiores lojas online do mundo. Anos depois, em 1999, foi fundado o grupo Alibaba na China, que se tornou o recordista de vendas online no mundo com 170 bilhões de dólares em vendas (em 2012), mais do que os concorrentes eBay e Amazon juntos (ECONOMIST, 2013). Recentemente, em 2020, a empresa eMarketer estimou que 88,3% dos usuários da Internet na China fizeram uma compra online, e que 41,2% de todas as vendas no varejo foram

feitas online. Nos EUA, para efeito de comparação, 82,4% dos usuários de Internet participaram do comércio eletrônico, mas apenas 14,5% do varejo foi do comércio online. Isso significa que os consumidores na China preferem cada vez mais fazer suas compras online, enquanto nos EUA, o comércio eletrônico é usado apenas em uma pequena porcentagem das necessidades de compras (FLOOD, 2020).

Já no Brasil, a Internet teve o desenvolvimento inicial nos anos 90. Ao decorrer dos anos, a popularização dos aparelhos eletrônicos e as campanhas de incentivos ao acesso à Internet desenvolvidas pelo governo possibilitaram um aumento nos números de usuários de Internet, que proporcionalmente, causou um crescimento na área de *E-commerce*. Além disso, a legislação brasileira sobre compras online garante a segurança dos consumidores, como por exemplo: a loja eletrônica deve apresentar informações de contato visíveis aos visitantes; o cliente tem direito a devolução grátis até sete dias após a chegada do produto; informações como preços, taxas adicionais, fretes, possíveis riscos à saúde e à segurança do consumidor devem ser claras nos anúncios, aumentando a segurança nas compras e atraindo mais consumidores.

No ano de 2021, as restrições de circulação das pessoas durante o período da pandemia causada pelo novo coronavírus expandiu o mercado de *E-commerce* no Brasil. Devido ao isolamento social, as pessoas mudaram o seu hábito de consumo. Com o fechamento das lojas físicas, a demanda por compras online aumentou. Segundo o indicador Mastercard SpendingPulse, indicador macroeconômico de vendas no varejo de todos os tipos de pagamento em atividade de vendas na rede da Mastercard e estimativas de outras formas de pagamento, as vendas do *E-commerce* no Brasil, em 2020, cresceram 75%, com faturamento da ordem de 87,4 milhões de reais, em comparação com 2019. Além disso, esse indicador aponta que as vendas tiveram um crescimento de 86% em fevereiro do ano 2021, ao se comparar com o mesmo período de 2020. O crescimento do *E-commerce* também é mostrado pelo fato de que o Mercado Livre - empresa de comércio eletrônico - ultrapassou a Vale S.A - mineradora multinacional brasileira - e se tornou a empresa mais valiosa da América Latina (INFOMONEY, 2020).

Esse crescimento do mercado de comércio eletrônico beneficia tanto o consumidor, devido a facilidade de acesso e diversidade de produtos, como também o vendedor, pois, em geral, lojas online demandam custo inicial e capital de giro menores em relação às lojas físicas. Nesse tipo de mercado, é possível iniciar um negócio de vendas com apenas uma pessoa, em sua própria residência, não gerando custos de aluguel e/ou contrato de funcionários. Ainda há casos em que não há custos com estoques de produtos. Por exem-

plo, o modelo *dropshipping*, o qual o varejista se associa a um fornecedor de *dropshipping* que fabrica e/ou armazena produtos, os empacota e posta diretamente para o cliente em nome do varejista. Por outro lado, essa diversidade de produtos presentes no mercado de comércio online pode sobrecarregar o usuário devido ao grande volume de possibilidades. Tendo o objetivo de otimizar a experiência do consumidor no mercado virtual, o sistema de recomendação é fundamental nesse cenário. De acordo com a Microsoft Asia Research Academy, aproximadamente 30% da navegação na página da Amazon vêm do sistema de recomendação.

Nesse contexto, as transações realizadas no comércio eletrônico deixam registros e fornecem informações para esboçar o perfil do consumidor, sendo a compreensão das demandas dos clientes um dos importantes pilares para um negócio proceder, assim entra o papel do sistema de recomendação. Pois os sistemas de recomendação são mecanismos capazes de analisar e compreender o comportamento dos usuários para fazer recomendações relevantes de conteúdos novos, sugerindo opções ao usuários para sua tomada de decisão nas compras online, como se fosse um vendedor virtual, fornecendo uma experiência personalizada. Como consequência, esses sistemas melhoram as vendas e otimizam a interação do usuário e sua aderência com a plataforma.

1.1 Revisão Bibliográfica

Segundo Hortinha (2000), existem algumas limitações ao se falar do comércio eletrônico do tipo B2B, dentre elas está a dificuldade de utilização, que proporciona a má navegabilidade dos usuários e a falta de personalização, exigindo, com frequência, um grande número de cliques até chegar ao produto ou serviço pretendido.

Fernandes e Linhares (2012) fizeram uma aplicação de sistema de recomendação no site de *E-commerce* B2B da empresa Pauta Distribuição e Logística utilizando filtragem baseada em conteúdo. O sistema faz uma pesquisa nos últimos pedidos do determinado parceiro que efetuar o login e com base no histórico de pedidos, apresenta os seis produtos considerados de maior relevância, e calcula o número de vezes que o produto foi comprado por este cliente multiplicado por sua quantidade, gerando assim, uma lista em ordem decrescente com os resultados obtidos, e os seis primeiros desta lista são exibidos no site. Os resultados preliminares são considerados satisfatórios, o qual apresentou um ganho para a empresa, no sentido de agradar seus clientes.

Recentemente, Oliveira (2020) realizou um estudo comparando diferentes tipos de

sistema de recomendação para consumidores de *E-commerce* utilizando o banco de dados da Olist Store. A partir de uma análise exploratória dos dados, o autor observou que o atraso das entregas é o maior motivo para as insatisfações dos clientes. Ao comparar o sistema de recomendação baseado em popularidade e com o sistema baseado em filtragem colaborativa, percebeu que o segundo apresenta o melhor desempenho.

Almeida e Ludolf (2020) investigaram cinco diferentes tipos de sistema de recomendação para um conjunto de dados de um *E-commerce*: filtragem colaborativa baseada em itens, filtragem colaborativa baseada em usuários, popularidade de itens, algoritmo *association rules* e recomendação baseada em *ranks*. No algoritmo baseado em *ranks*, segundo a localização e o registro de data e hora do usuário, o modelo faz diferentes recomendações de categorias do produto, permitindo maior flexibilidade da função de recomendação e adaptando, assim, as necessidades do consumidor.

Como observado, há diversos métodos que podem ser utilizados para a recomendação aos usuários. Com um enfoque em métodos estatísticos, que consideram a incerteza nas recomendações, Zhang e Iyengar (2002) analisaram o sistema de recomendação utilizando modelos lineares. Os autores compararam os resultados obtidos pela modelagem estatística com a filtragem colaborativa baseada em memória e com outros métodos baseados em árvore de decisão, sendo o experimento aplicado em três conjunto de dados. Os resultados mostraram que os modelos lineares foram mais precisos do que a filtragem colaborativa baseada em memória e seu uso pareceu adequado em um sistema de recomendação.

Nesse contexto, Lima (2020) apresentou um modelo bayesiano de filtragem colaborativa em sistema de recomendação aos usuários. Utilizando modelos lineares bayesianos, foi possível obter as preditivas a posteriori e fornecer estimativas pontuais, permitindo preencher a matriz esparsa. O uso de modelos bayesianos permite diferentes recomendações, a partir da distribuição preditiva a posteriori, sem novas entradas no sistema, amenizando o problema de inicialização a frio (*cold start*).

1.2 Objetivos

O objetivo geral deste trabalho é analisar a base de dados referente às vendas de uma plataforma de *E-commerce* por meio de um sistema de recomendação e tentar estimar avaliações do consumidor em itens que ainda não consumiram.

1.2.1 Objetivos específicos

Como objetivos específicos deste trabalho, podem ser listados:

- utilizar um sistema de recomendação através de um modelo de regressão linear normal, como apresentado em Lima (2020), porém, utilizando métodos de estimação da inferência clássica;
- utilizar o método de regularização LASSO para selecionar variáveis no modelo de regressão linear normal;
- aplicar os modelos lineares normais como sistema de recomendação no banco de dados da Olist Store.

1.3 Organização

Este Trabalho de conclusão de curso está apresentado em quatro capítulos. No Capítulo 1 foram apresentados conceitos, realizada uma breve revisão bibliográfica e discriminados os objetivos. No Capítulo 2, serão apresentadas a metodologia abordada, o modelo estatístico proposto e a natureza dos dados a serem investigados. No Capítulo 3 serão exibidos os resultados obtidos nas análises, com suas respectivas interpretações. Finalmente, no Capítulo 4, apresenta-se a conclusão desta pesquisa.

2 Materiais e Métodos

Neste capítulo, serão apresentados os tipos de sistemas de recomendação, com enfoque nos sistemas baseados em filtragem colaborativa. Além disso, serão discutidos os métodos de filtragem colaborativa por vizinhaça e modelos. Por fim, será apresentado o modelo utilizado como sistema de recomendação, seguindo Lima (2020).

2.1 Tipos de sistemas de recomendação

O sistema de recomendação baseado em conteúdo analisa perfis dos usuários e extrai características dos itens, como títulos, subtítulos ou comentários, sendo capaz de construir um modelo ou perfil de interesse do usuário. Com base nesses dados recomendam-se itens similares às suas preferências anteriores. A vantagem é que esse sistema ameniza o problema de inicialização a frio (*cold start*) quando possui informações disponíveis de todos os itens. Por exemplo, um novo item que é semelhante aos já existentes no sistema pode ser recomendado para usuários, sem precisar ter recebido alguma classificação. Por outro lado, é necessário um grande número de itens avaliados, além da limitação de ampliar o padrão de consumo dos usuários, já que esse sistema recomenda apenas por semelhança (AGGARWAL, 2016).

O sistema de recomendação baseado em popularidade simplesmente recomenda aos consumidores os itens mais populares registrados, pois parte da hipótese de que a popularidade reflete a tendência coletiva. Esse sistema é um dos mais usados pois é fácil de implementar (BERTANI; COSTA, 2020).

O sistema de recomendação por filtragem colaborativa é baseado nos *feedbacks* registrados, podendo serem eles explícitos, como uma escala de pontuação determinada, ou implícitos, tendo em vista o comportamento de compra do usuário. Esse método assume a hipótese de que usuários semelhantes exibem padrões semelhantes de comportamento de classificação e itens semelhantes recebem classificações semelhantes (AGGARWAL, 2016). Esse tipo de sistema tem duas estratégias principais de implementação: por memória (ou

vizinhança), que determina a semelhança entre itens a partir de classificações prévias dos usuários; e por modelos, que utilizam diferentes métodos de aprendizagem de máquina e mineração de dados para fazer previsões e assim recomendar ao usuário.

2.2 Modelos de Filtragem colaborativa

Existe uma grande diferença entre modelos de regressão e um problema de filtragem colaborativa. Em modelos de regressão, o ideal é fixar uma coluna que define amostras de treinamento e uma matriz completa de características para adequar ao modelo. Já em filtragem colaborativa, tem-se uma matriz fixa com um grande número de dados faltantes. A Figura 1 ilustra a comparação entre o problema de modelos de regressão e filtragem colaborativa, onde os valores destacados representam valores faltantes da matriz.

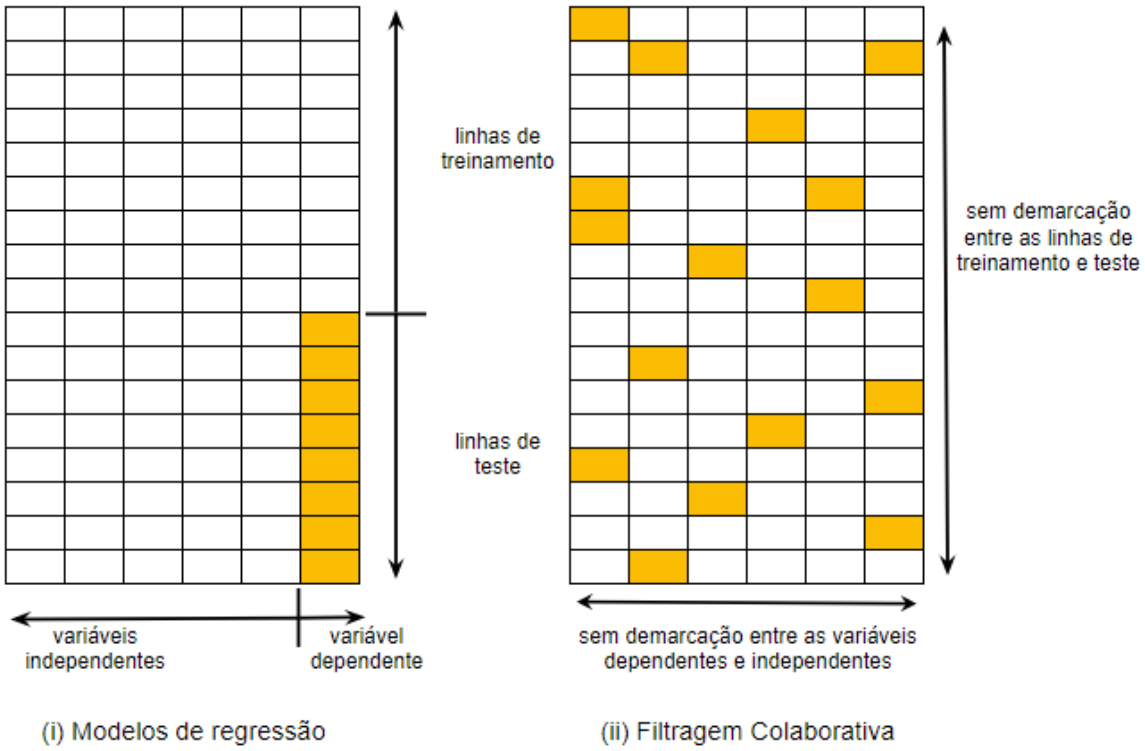

Figura 1: Diferença entre o problema de Modelos de Regressão e Filtragem Colaborativa.

2.2.1 Filtragem colaborativa baseada em vizinhança

De formal geral, a filtragem colaborativa por vizinhança tem como hipótese de que usuários com preferências semelhantes classificam itens da mesma forma. Nesse caso, as classificações fornecidas por usuários semelhantes a um usuário alvo são usadas para fazer recomendações para este usuário alvo. Dessa forma, as classificações faltantes de um usuário alvo podem ser previstas encontrando uma vizinhança de usuários semelhantes e depois inserindo as classificações desses mesmos usuários como uma previsão do usuário alvo para um item.

Considere uma Matriz de Utilidade $\mathbf{Y}$ de dimensão $n \times p$, sendo n o número de usuários e p a quantidade de itens. Seja I_u o conjunto de índices de itens para os quais as classificações foram especificadas pelo usuário (linha) u. Supondo uma matriz $\mathbf{Y}$ com 2 usuários, u e v, e 5 itens diferentes, se as avaliações do primeiro, terceiro e quinto itens do usuário u são especificadas (observadas) e os restantes são dados faltantes, então temos $I_u = \{1, 3, 5\}$, e o usuário v classificou do primeiro ao quarto item, então $I_v = \{1, 2, 3, 4\}$, sendo o conjunto de itens avaliados por ambos os usuários u e v recebem $I_u \cap I_v = \{1, 3, 5\} \cap \{1, 2, 3, 4\} = \{1, 3\}$, conforme a matriz ilustrada abaixo, onde cada y representa uma avaliação. Dessa forma, cada elemento y_{ij} da matriz $\mathbf{Y}$ representa a classificação dada pelo usuário i ao item j (AGGARWAL, 2016).

$$\mathbf{Y}_{2\times5} = \begin{pmatrix} y_{11} & ? & y_{13} & ? & y_{15} \\ y_{21} & y_{22} & y_{23} & y_{24} & ? \end{pmatrix}$$

Para definir usuários semelhantes há diversas medidas que são popularmente utilizadas. As medidas de similaridade mais utilizadas na filtragem colaborativa são a similaridade cosseno e o coeficiente de correlação de Pearson. Nas funções a seguir, $\mathbf{u}$ e $\mathbf{v}$ denotam vetores p-dimensionais das classificações dadas, respectivamente, pelos usuários u e v ao conjunto de itens q, no qual os dois usuários classificaram em comum. O k-ésimo elemento do vetor representa a classificação do usuário ao k-ésimo item, $k = 1, \ldots, q$, e são denotados, respectivamente, por r_{uk} e r_{vk}.

- Similaridade cosseno:

$$cos(\mathbf{u}, \mathbf{v}) = \frac{(\mathbf{u} \cdot \mathbf{v})}{||\boldsymbol{u}|| \cdot ||\boldsymbol{v}||} \tag{2.1}$$

  onde $(\mathbf{u} \cdot \mathbf{v})$ é o produto escalar entre os vetores, $||\boldsymbol{u}||$ é a norma do vetor $\mathbf{u}$, e $||\boldsymbol{v}||$ é a norma do vetor $\mathbf{v}$.

- Correlação de Pearson:

$$s_{u,v} = \text{Pearson}(\mathbf{u}, \mathbf{v}) = \frac{\sum_{k \in I_u \cap I_v} (r_{uk} - \mu_u) \cdot (r_{vk} - \mu_v)}{\sqrt{\sum_{k \in I_u \cap I_v} (r_{uk} - \mu_u)^2} \sqrt{\sum_{k \in I_u \cap I_v} (r_{vk} - \mu_v)^2}} \tag{2.2}$$

  onde μ_u e μ_v são as médias das classificações dadas, respectivamente, pelos usuários **u** e **v**.

O coeficiente de correlação de Pearson é uma medida que captura a similaridade entre os vetores de classificação de dois usuários u e v, pois $I_u \cap I_v$ representa o conjunto de classificações mutuamente observadas. O coeficiente é calculado apenas neste conjunto. O primeiro passo é calcular a classificação média μ_u para cada usuário u usando suas classificações especificadas:

$$\mu_u = \frac{\sum_{k \in I_u} r_{uk}}{|I_u|}, \quad \forall u \in 1, \ldots, n. \tag{2.3}$$

E, em seguida, é aplicado o coeficiente de correlação de Pearson apresentado na equação (2.2).

Após o cálculo da semelhança entre todos os usuários, é preciso prever as classificações que faltam. Supondo que o objetivo é prever um valor não observado da classificação dada pelo usuário u ao item k, $\hat{r}_{uk}$, através da medida de similaridade é possível identificar um conjunto de m usuários, que têm comportamento mais próximos ao usuário alvo u, e que especificaram classificações para o item k. Sendo $S^m(u)$ o conjunto de m notas dos usuários mais próximos do usuário u e $s_{u,j}$ as notas entre u e j, de acordo com Aggarwal (2016), a função de previsão baseada na vizinhança é:

$$\hat{r}_{uk} = \mu_u + \frac{\sum_{j \in S^m(u)} s_{u,j} r_{jk}}{\sum_{j \in S^m(u)} s_{u,j}} \tag{2.4}$$

Pode ser realizado as mesmas análises apresentadas acima mas variando de usuário para item, ou seja, recomendar por semelhança entre itens. Nesse caso, tem como base a hipótese de itens semelhantes possuem classificações semelhantes. A sua vantagem em comparação com filtragem colaborativa baseada em vizinhança dos usuários é a realização de recomendações antes de ter um perfil do usuário, assim, aumentando ainda mais a aderência do usuário com a plataforma.

2.2.2 Filtragem colaborativa baseada em modelo

Na filtragem colaborativa baseada em modelo é criado primeiro um modelo que resume os dados, diferente do método baseado em vizinhança, onde a abordagem de previsão é específica para a instância prevista. Além disso, a fase de construção do modelo é realizada separadamente da fase de previsão.

Segundo Aggarwal (2016), um dos desafios na filtragem colaborativa baseada em modelo é o preenchimento da matriz esparsa, e como solução, vários modelos podem ser usados para preencher essa matriz. Como exemplos, podem ser citados os seguintes métodos: árvores de decisão, classificadores de Bayes, modelos de regressão, redes neurais entre outros. Essa solução é executável fixando uma das colunas de itens como uma variável dependente e usando todas as outras colunas como uma matriz de características. Um vez que o modelo estimado seja adequado e considerando as linhas onde havia dados, pode-se prever os elementos que faltam.

No contexto do uso de modelos lineares como sistema de recomendação, Zhang e Iyengar (2002) compararam o modelo de regressão ridge com modelos usando árvore de decisão e com métodos baseados em memória usando dados de várias fontes. Os resultados das análises indicaram que os modelos lineares propostos foram adequados para aplicação.

O modelo utilizado neste trabalho segue ideia similar ao modelo de Zhang e Iyengar (2002), porém, ao invés de utilizar o método de regularização ridge, será utilizado o método LASSO, como proposto em Lima (2020).

2.2.3 Medidas de Qualidade de um Sistema de Recomendação

De acordo com Jonathan Joseph A. Konstan e Riedl (2004), a avaliação do desempenho do algoritmo é um passo importante no sistema de recomendação. Antes de criar recomendações, alguns itens são retidos da Matriz de Utilidade $\mathbf{Y}$ a fim de medir a qualidade da classificação prevista corresponde ao valor retido. Dessa maneira, se o algoritmo de recomendação apresentar um melhor desempenho na previsão dos itens retirados, também terá um melhor desempenho na localização de boas recomendações para itens com classificação faltante.

Existem várias medidas que pode avaliar a qualidade de um sistema de recomendação na literatura, entre eles estão :

- Erro Médio Absoluto (MAE): mede a diferença entre duas variáveis contínuas. Nesse

contexto, é a diferença média, sob a amostra teste, entre a previsão e a observação real, onde todas as diferenças individuais têm peso igual. Os métodos que fornecem os menores MAE são preferíveis.

$$MAE = \frac{\sum_{i,j \in N} |\hat{r}_{ij} - r_{ij}|}{N} \tag{2.5}$$

onde $\hat{r}_{ij}$ é a estimativa de r_{ij}, que é a classificação dada pelo usuário i ao item j, e N representa o total de itens que podem ser recomendados.

- Raiz Quadrada do Erro Médio (RMSE) : é a medida que calcula a raiz quadrática média dos erros entre valores observados e preditos.

$$RMSE = \sqrt{\frac{\sum_{i,j \in N} (\hat{r}_{ij} - r_{ij})^2}{N}} \tag{2.6}$$

onde $\hat{r}_{ij}$ é a estimativa de r_{ij}, que é a classificação dada pelo usuário i ao item j, e N representa o total de itens que podem ser recomendados.

- Erro Percentual Absoluto Médio (MAPE): mede a porcentagem do erro em comparação com o valor verdadeiro, fornecendo assim uma média de erro padronizada.

$$MAPE = \frac{1}{N} \sum_{i,j \in N} \frac{|r_{ij} - \hat{r}_{ij}|}{|r_{ij}|} \times 100 \tag{2.7}$$

onde $\hat{r}_{ij}$ é a estimativa de r_{ij}, que é a classificação dada pelo usuário i ao item j, e N representa o total de itens que podem ser recomendados.

2.3 Modelo de Regressão Linear Múltiplo

O modelo de regressão linear mútiplo constitui de uma técnica que tem por objetivo explicar a relação entre uma variável dependente (Y) com um conjunto de variáveis independentes ($X_1, X_2, ..., X_{p-1}$) em que $p > 1$. A suposição básica deste modelo é que a média da distribuição de Y varia de forma linear com as variáveis $X_1, X_2, ..., X_{p-1}$. Essa relação pode ser estabelecida por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{ip-1} + \epsilon_i, \;\; i = 1, \ldots, n \tag{2.8}$$

Onde:

Y_i é o valor da variável resposta para a i-ésima observação;

β_j é o j-ésimo coeficiente do modelo, $j = 0, 1, \ldots, p-1$;

X_{ij} é o valor da j-ésima variável independente para o i-ésimo indivíduo;

ϵ_i é erro aleatório para i-ésima observação, sendo variância $V(\epsilon_i) = \sigma^2$.

Assumindo que os erros tem média zero, $E(\epsilon) = 0$, a função resposta para o modelo fica da seguindo forma:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{ip-1} \tag{2.9}$$

sendo o parâmetro β_j a alteração na média de Y devido ao acréscimo de 1 unidade na variável X_j, quando os valores das demais variáveis explicativas $X_k (k \neq j)$ permanecem constantes.

O modelo pode ser definido de forma matricial, como apresentado na equação a seguir:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.10}$$

sendo:

$$\mathbf{Y}_{n\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X}_{n\times p} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p-1} \\ 1 & X_{21} & \cdots & X_{2p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np-1} \end{pmatrix}, \boldsymbol{\beta}_{p\times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} \text{ e } \boldsymbol{\epsilon}_{n\times 1} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

onde n é o número de observações e $p-1$ é o número de variáveis.

Assumindo que os erros ϵ_i são independentes e normalmente distribuídos, $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, sendo $\mathbf{I}$ a matriz de identidade $n \times n$, implica que as observações Y_i são independentes e também seguem a distribuição normal com variância constante σ^2.

2.3.1 Estimadores para os parâmetros desconhecidos

Para obter a função de regressão estimada, é preciso antes, encontar as estimativas dos parâmetros desconhecidos através de seus estimadores. Existem muitos métodos de obter estimador para os parâmetros, um deles é o estimador por mínimos quadrados, que busca minimizar a soma dos quadrados dos erros (SQE), a qual pode ser definida matricialmente por:

$$SQE = \sum_{i=1}^{n} \epsilon_i^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\boldsymbol{X}\boldsymbol{\beta} \tag{2.11}$$

Para encontrar o estimador para $\boldsymbol{\beta}$ por mínimos quadrados é preciso encontrar os pontos de mínimo de SQE, derivando SQE em relação a $\boldsymbol{\beta}$ e igualar a zero.

$$\begin{aligned} \frac{\partial SQE}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} &= 0 && (2.12)\\ \iff \mathbf{X}^T\boldsymbol{X}\boldsymbol{\beta} &= \mathbf{X}^T\mathbf{Y} && (2.13)\\ \iff (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{X}\boldsymbol{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} && (2.14)\\ \iff \widehat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} && (2.15) \end{aligned}$$

Note que $\mathrm{E}\left[\widehat{\boldsymbol{\beta}}\right] = E\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}\right] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{Y}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$

e

$$\begin{aligned} V\left[\hat{\boldsymbol{\beta}}\right] &= V[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}]\\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T V(\mathbf{Y})[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T\\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T V(\mathbf{Y})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned}$$

Logo, como $\boldsymbol{\beta}$ é combinação linear de $\mathbf{Y}$, tem-se que $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.

2.3.2 Estimador para variância

A variância σ^2 do termo de erro ϵ_i é estimada para indicar a variabilidade da distribuição de probabilidade de $\mathbf{Y}$. O estimador de σ^2 é construído a partir de SQE:

$$SQE = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y}^T\mathbf{Y} - \hat{\boldsymbol{\beta}}\mathbf{X}^T\mathbf{Y} \quad (2.16)$$

SQE tem $n - p$ graus de liberdade associados a ela, visto que p parâmetros são estimados no modelo de regressão. Sabendo que o quadrádo médio dos erros (MQE) é um estimador não viesado para variância σ^2, então:

$$\widehat{\sigma}^2 = MQE = \frac{SQE}{n-p} \quad (2.17)$$

2.3.3 Análise dos resíduos do modelo

A análise dos resíduos é importante devido os resíduos serem a principal ferramenta para identificar violação das suposições do modelo. O i-ésimo resíduo em um modelo de regressão linear é definido como o desvio entre o valor observado e o valor ajustado da variável resposta, isto é,

$$e_i = Y_i - \hat{Y}_i, \quad i = 1, \ldots, n \tag{2.18}$$

Tendo as seguintes propriedades:

- $E(e_i) = 0$
- $V(e_i) = \sigma^2 \left(1 - \left(\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)\right)$
- Os resíduos têm distribuição Normal (combinação linear dos erros)

O vetor de resíduos $\mathbf{e}$ é dado por:

$$\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} - \begin{pmatrix} X_1^T\hat{\beta} \\ X_2^T\hat{\beta} \\ \vdots \\ X_n^T\hat{\beta} \end{pmatrix} \tag{2.19}$$

E mais comum o uso da padronização dos resíduos do que os próprios resíduos e essa padronização é chamada de resíduo padronizado e definido por:

$$\mathbf{e}^* = \frac{\mathbf{Y} - \hat{\mathbf{Y}}}{\sqrt{MQE}} \tag{2.20}$$

onde $MQE = \frac{\mathbf{e}^T\mathbf{e}}{n-p}$.

Com base na observação dos resíduos, é possível analisar a adequabilidade do modelo. Como por exemplo a questão de não linearidade, isto é, a ausência de uma relação linear entre a variável resposta e as covariáveis, e também a heterocedasticidade, ou seja, erros com variância não constante e por fim a não normalidade, onde os erros não são normalmente distribuídos.

2.3.4 Método de regularização LASSO

O método de regularização LASSO (*Least Absolute Shrinkage and Selection Operator*) é um penalizador do procedimento de mínimos quadrados, proposto por Tibshirani (1996).

Esse método minimiza a soma dos quadrados dos erros com uma restrição nos coeficientes $\boldsymbol{\beta}$ a serem estimados. Este método utiliza todas as regressoras restringindo/regularizando estimativas dos coeficientes, em direção a zero. Dessa forma, a estimativa de $\boldsymbol{\beta}$ utilizando o LASSO é :

$$\arg\min_{\beta}(Y_i - \beta_0 - \beta_1 X_{1i} ... - \beta_p X_{pi})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.21}$$

onde o parâmetro λ é um parâmetro de entendimento fundamental para a restrição a ser imposta, ele é responsável por indicar a região que os parâmetros a serem estimados não poderão estar. De forma geral, λ mede o grau de penalização dado à $\boldsymbol{\beta}$. Note, a partir da Equação (2.21), que está sendo adicionado um viés às estimativas de minínimos quadrados. A Figura 2 exemplifica esse viés adicionado às estimativas de mínimos quadrados e a penalização em direção a zero.

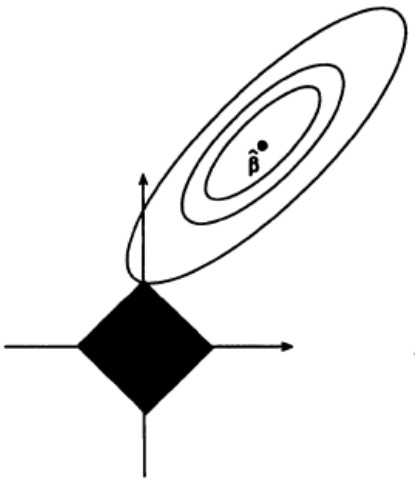

Figura 2: Exemplo da regularização LASSO. $\widehat{\boldsymbol{\beta}}$: estimativa por mínimos quadrados.
Fonte: Tibshirani (1996).

2.4 Modelo proposto

Para a construção do modelo proposto admite-se que as classificações dadas por diferentes usuários são independentes, porém, as classificações dadas pelo mesmo usuário para itens diferentes possuem correlação. Tem-se também como hipótese que usuários semelhantes tendem a avaliar itens de maneira semelhante. Da mesma forma, foi assumido que as classificações podem ter comportamento aproximadamente normal com cada item assumindo uma média e variância diferentes. Portanto a Matriz de Utilidade segue uma matriz de distribuição normal, de acordo com Gupta e Nagar (2018).

O modelo utilizado neste trabalho tem como base o modelo proposto por Lima (2020), mas com o procedimento de estimação dos parametros baseado nos métodos clássicos apresentados na Seção 2.3. Sendo assim, para compreensão do modelo utilizado, suponha que n usuários classificaram p itens, e seja y_{ij} a classificação dada pelo usuário i ao item j. Sendo $(y_{i1}, y_{i2}, \ldots, y_{ip})^T$ o vetor de classificação do i-ésimo usuário de todos os itens j, $i = 1, \ldots, n$, $j = 1, \ldots, p$, assume-se que essas classificações têm uma distribuição normal com média μ_j e variância ϕ_j^2. Assumindo também que as classificações dadas por usuários diferentes são condicionalmente independentes de $\boldsymbol{\mu}$ e Φ, tem-se

$$(y_{i1}, y_{i2}, ..., y_{ip}) \sim^{iid} N_p(\boldsymbol{\mu}, \Phi), \quad i = 1, \ldots, n \tag{2.22}$$

com

$$\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_p)^T \tag{2.23}$$

e

$$\Phi = \begin{pmatrix} \phi_1^2 & \phi_{1,2} & \cdots & \phi_{1,p} \\ \phi_{2,1} & \phi_2^2 & \cdots & \phi_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{p,1} & \phi_{p,2} & \cdots & \phi_p^2 \end{pmatrix} \tag{2.24}$$

onde $N_p(\boldsymbol{\mu}, \Phi)$ é uma distribuição normal p-variada com média $\boldsymbol{\mu}$ e matriz de covariância Φ.

Supondo agora que $\mathbf{Y}_k$ é o vetor de classificação dada por todos os k-ésimos itens, tal que $\mathbf{Y}_k = (y_{1k}, y_{2k}, \ldots, y_{nk})^T$, $k = 1, \ldots, p$, e $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_p) \in \mathbb{R}^{n \times p}$ a Matriz de Utilidade fornecidas pelos n usuários de todos os itens. Logo,

$$\mathbf{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix} \tag{2.25}$$

onde $\mathbf{Y}$ é uma matriz de média M e matriz de covariância $\Sigma = \Phi \otimes I_n$. De forma vetorial,

temos

$$vec(\mathbf{Y}) = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_p \end{pmatrix} = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n1} \\ y_{12} \\ y_{22} \\ \vdots \\ y_{n-1p} \\ y_{np} \end{pmatrix} \sim N_{np}(\mathbf{M}, \Sigma) \tag{2.26}$$

com

$$\mathbf{M} = (\mu_1, \mu_1, \cdots, \mu_1, \mu_2, \mu_2, \cdots, \mu_p, \mu_p)^T \tag{2.27}$$

sendo I_n a matriz identidade n-dimencional e $\otimes$ o produto de Kronecker. A matriz de covariância da matriz $vec(\mathbf{Y})$ é dada por

$$\Sigma = \Phi \otimes I_n = \begin{pmatrix} \phi_1^2 I_n & \phi_{1,2} I_n & \cdots & \phi_{1,p} I_n \\ \phi_{1,2} I_n & \phi_2^2 I_n & \cdots & \phi_{2,p} I_n \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{1,p} I_n & \phi_{1,p-1} I_n & \cdots & \phi_p^2 I_n \end{pmatrix} \tag{2.28}$$

e definindo

$$\mathbf{Y}_{-k} = vec(\mathbf{Y}_1, \ldots, \mathbf{Y}_{k-1}, \mathbf{Y}_{k+1}, \ldots, \mathbf{Y}_p)^T \tag{2.29}$$

o vetor com média μ_{-k} e matriz de covariância $\Sigma_{-k,-k}$. Temos que

$$\begin{pmatrix} \mathbf{Y}_k \\ \mathbf{Y}_{-k} \end{pmatrix} \sim N_{np}\left(\begin{pmatrix} \boldsymbol{\mu}_k \\ \boldsymbol{\mu}_{-k} \end{pmatrix}, \begin{pmatrix} \Sigma_{k,k} & \Sigma_{k,-k} \\ \Sigma_{-k,k} & \Sigma_{-k,-k} \end{pmatrix} \right) \tag{2.30}$$

Para encontrar a distribuição condicional de $\mathbf{Y}_k$ dado todas as classificações dos outros itens, $(\mathbf{Y}_{-k} = \mathbf{y}_{-k})$, basta aplicar as propriedades da distribuição normal multivariada. Portanto,

$$\mathbf{Y}_k | (\mathbf{Y}_{-k} = \mathbf{y}_{-k}) \sim N_n(\boldsymbol{\mu}^k, \Sigma^k) \tag{2.31}$$

com

$$\boldsymbol{\mu}^k = \boldsymbol{\mu}_k + \Sigma_{k,-k} \Sigma_{-k,-k}^{-1} (\mathbf{y}_{-k} - \boldsymbol{\mu}_{-k}) \tag{2.32}$$

$$\Sigma^k = \Sigma_{k,k} - \Sigma_{k,-k} \Sigma_{-k,-k}^{-1} \Sigma_{-k,k} \tag{2.33}$$

Assim, a partir da Equação (2.31), é possível descrever cada coluna de Matriz de Utilidade em função das demais. A estrutura do modelo linear utilizado neste trabalho é construída

de forma que isso seja possível. Reparametrizando a distribuição de (2.31), tem-se

$$\mathbf{Y}_k|\mathbf{Y}_{-k} \sim N_n(\mathbf{1}\beta_1^{(k)} + \mathbf{X}^{(k)}\boldsymbol{\beta}^{(k)}, \sigma_k^2 I_n) \tag{2.34}$$

onde $\mathbf{1}$ é um vetor n-dimensional de 1's, $\beta_1^{(k)}$ é o nível, $\beta^{(k)}$ é um vetor de coeficientes com dimensão $p-1$, $\mathbf{X}^{(k)}$ é a Matriz de Utilidade $\mathbf{Y}$ sem a coluna k e σ_k^2 é definida como a variância de $\mathbf{Y}_k|\mathbf{Y}_{-k}$.

2.4.1 Algoritmo proposto

Na Tabela 1 é apresentado o algoritmo utilizado na estimação da Matriz de Utilidade $\mathbf{Y}$. Suponha um conjunto de dados com a classificação de n usuários (linhas) e p itens (colunas) disponível na Matriz de Utilidade $\mathbf{Y}$, que possui valores faltantes.

Tabela 1: Algoritmo para estimação da Matriz de Utilidade $\mathbf{Y}$.

Passo 1: Imputar de forma aleatória um valor inicial para os dados faltantes em $\mathbf{Y}$, considerando o domínio dos dados.
Passo 2: Estimar o modelo descrito na Equação (2.34) para cada coluna k, e selecionar as covariáveis que são significativas para esta coluna k através da penalização do LASSO, sendo $k = 1, \ldots, p$.
Passo 3: Predizer os dados faltantes de cada coluna k, $k = 1, \ldots, p$, através do modelo estimado no Passo 2 e substituir os valores preditos na matriz de classficação $\mathbf{Y}$, atualizando os valores iniciais imputados.
Passo 4: Repetir os Passos 2 e 3 até que $\lvert y_{ik}^{(m)} - y_{ik}^{(m+1)}\rvert < 0,01$.

2.5 Dados Simulados

Tendo o objetivo de avaliar a adequabilidade do modelo, foi simulada uma base de dados composta por 250 observações (linhas) e 50 itens (colunas). Os 50 itens foram divididos em 5 grupos independentes de 10 itens sequenciais, de tal forma que em cada grupo há dependência entre os 10 itens. As observações formam geradas a partir de uma distribuição normal de média 3,4 e matriz de covariância $\mathbf{\Sigma}$, de dimensão 12500×12500. Os elementos de $\mathbf{\Sigma}$ foram gerados, dentro de cada de 10 itens, a partir de uma distribuição normal média 1 e variância 0,04. Elementos da matriz de covariancia que não correspondem aquele grupo são fixados em zero, conforme a Figura 3.

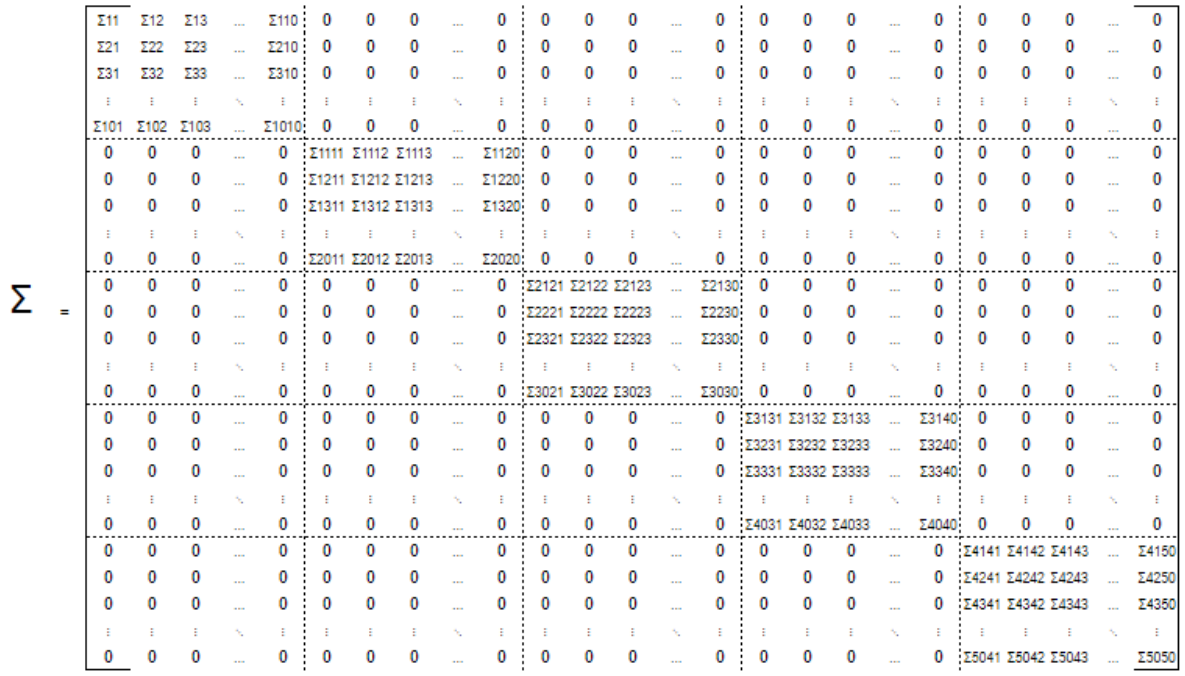

Figura 3: Forma da matriz covariância dos dados simulados.

Cada Σ_{ij} tem dimensão 250×250 onde i é o número de linhas e j é o número de colunas, quando $i = j$ (a diagonal na matriz $\mathbf{\Sigma}$), Σ_{ij} possui valores diferentes de zero apenas na diagonal principal.

2.6 Dados da Olist Store

A base de dados utilizada para os estudos deste trabalho são dados públicos de pedidos feitos através da plataforma de *E-commerce* brasileira Olist Store, disponível no site do Kaggle [1]. Olist é uma loja de departamentos dos *marketplaces* brasileiros que conecta pequenas empresas de todo o país permitindo que esses comerciantes vendam seus produtos através da Olist Store e os envie a partir de parceiros de logística da própria Olist.

O conjunto de dados tem informações de 100 mil pedidos de 2016 a 2018, contendo várias informações sobre o pedido, como: status do pedido, preço, pagamento, frete, localização do cliente, atributos do produto e avaliações dadas pelos clientes. Além disso, há informação de geolocalização tanto dos clientes quantos dos vendedores, relacionando os CEPs brasileiros às coordenadas de latitude e longitude do pedido.

Após a realização de uma compra do cliente, o vendedor deste produto é notificado para atender esse pedido, e assim que o cliente recebe o produto, ou vence a data prevista de entrega, o cliente recebe uma pesquisa de satisfação por e-mail onde pode dar uma nota da experiência de compra e escrever comentários sobre o produto recebido. O comprador pode atribuir uma nota ao produto de 1 a 5, onde 1 indica insatisfação e 5 alta satisfação. A Tabela 2 apresenta as variáveis do conjunto de dados e sua descrição.

[1] Kaggle é uma plataforma de competição de modelagem e análise de dados, onde empresas e pesquisadores publicam dados, e estatísticos ou especialistas em mineração de dados competem para produzir os melhores modelos sobre os dados publicados. https://www.kaggle.com/

Tabela 2: Variáveis e descrição do conjunto de dados

Variável	Descrição
seller_id	identificador único do vendedor
seller_zip_code_prefix	primeiros 5 dígitos do código postal do vendedor
seller_city	cidade do vendedor
seller_state	estado do vendedor
product_category_name	nome da categoria em portugues
product_category_name_english	nome da categoria em inglês
product_id	identificador único de produto
product_category_name	categoria raíz do produto, em português
product_name_lenght	número de caracteres extraídos do nome do produto
product_description_lenght	número de caracteres extraídos da descrição do produto
product_photos_qty	número de fotos publicadas do produto
product_weight_g	peso do produto medido em gramas
product_length_cm	comprimento do produto medido em centímetros
product_height_cm	altura do produto medida em centímetros
product_width_cm	largura do produto medida em centímetros
review_id	identificador de avaliação exclusivo
order_id	identificador de pedido único
review_score	nota de 1 a 5 dada pelo cliente em pesquisa de satisfação
review_comment_title	título do comentário a partir da avaliação deixada pelo cliente, em português
review_comment_message	mensagem de comentário a partir da avaliação deixada pelo cliente, em português
review_creation_date	data em que a pesquisa de satisfação foi enviada ao cliente
review_answer_timestamp	data / hora da resposta da pesquisa de satisfação
customer_id	chave para dados do cliente, cada pedido possui uma chave exclusiva
order_status	status do pedido
order_purchase_timestamp	data / hora da compra
order_approved_at	data / hora de aprovação do pagamento
order_delivered_carrier_date	data e hora de postagem do pedido. Quando foi entregue ao parceiro logístico
order_delivered_customer_date	data real de entrega do pedido ao cliente
order_estimated_delivery_date	estimativa da data de entrega que foi informada ao cliente no momento da compra
order_item_id	número sequencial que identifica o número de itens incluídos na mesma ordem
shipping_limit_date	data limite de envio do vendedor para lidar com o pedido ao parceiro logístico
price	preço do item
freight_value	item com valor de frete (o valor do frete é dividido entre total de itens no pedido)
payment_sequential	sequência criada para pedidos com mais de um método de pagamento
payment_type	método de pagamento escolhido pelo cliente
payment_installments	número de parcelas escolhido pelo cliente
payment_value	valor da transação
customer_unique_id	identificador único de um cliente
customer_zip_code_prefix	primeiros cinco dígitos do código postal do cliente
customer_city	cidade do cliente
customer_state	estado do cliente
geolocation_zip_code_prefix	primeiros 5 dígitos do código postal
geolocation_lat	latitude
geolocation_lng	longitude
geolocation_city	cidade
geolocation_state	estado

3 Resultados

Neste capítulo, serão apresentados os resultados obtidos neste trabalho a partir do modelo apresentado no Capítulo 2.

3.1 Dados Simulados

Conforme a descrição do Capítulo 2, a base simulada possui 250 usuários (linhas) e 50 itens (colunas). A Figura 4 abaixo verifica a correlação entre as colunas da base, onde cores mais escuras indicam correlação alta. Observa-se que em cada bloco de 10 colunas há forte dependência, conforme esperado. A relação entre os itens fora dos blocos não há qualquer dependência.

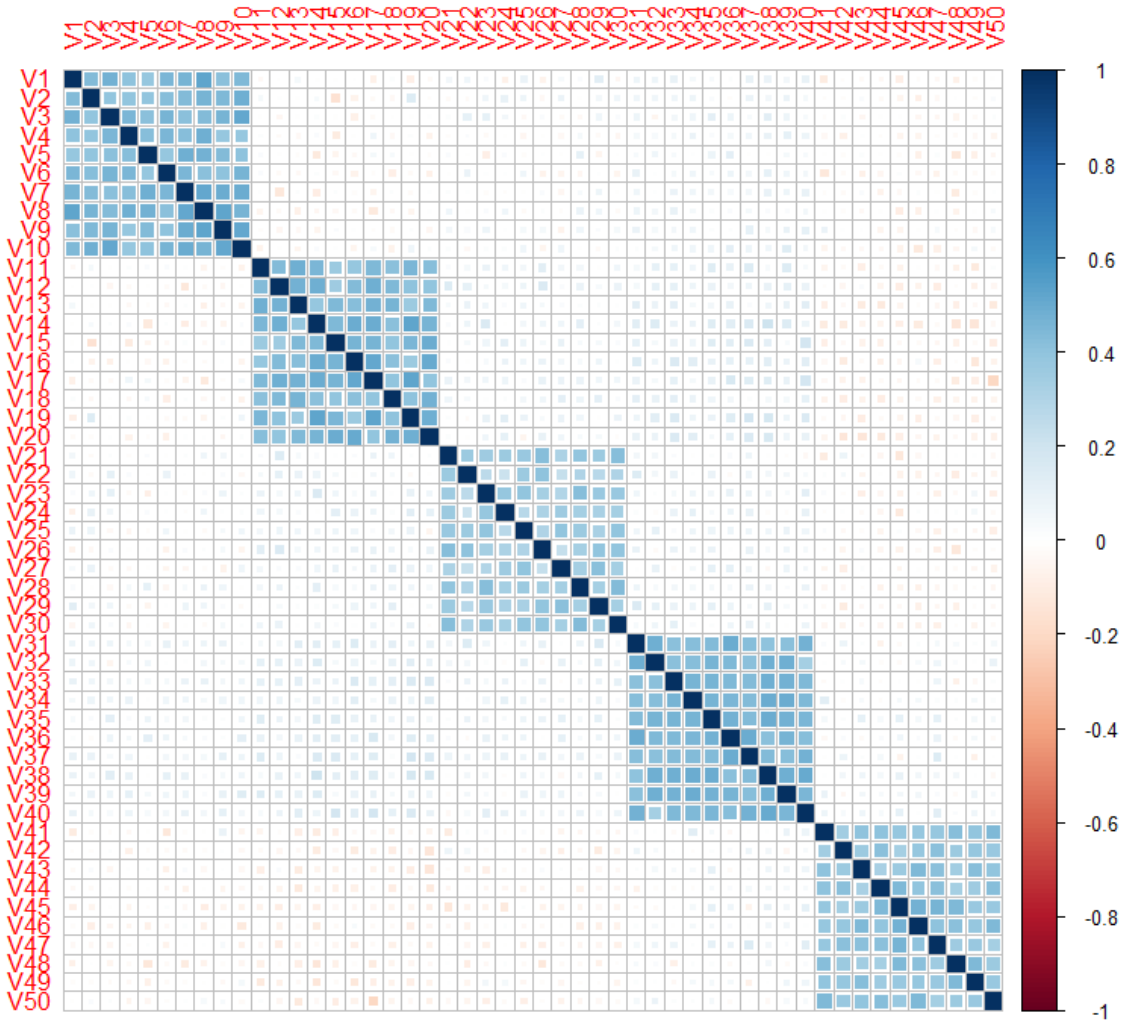

Figura 4: Matriz de correlação das colunas da Matriz de Utilidade simulada.

Para aplicação do modelo, foram retiradas aleatoriamente, diferentes porcentagens de valores faltantes da base para verificar o ajuste do modelo, ou seja, a capacidade do modelo em predizer as notas de classificação foram realizadas considerando 5%, 10%, 20%, 30%, 40% e 50% esparsidade em $\mathbf{Y}$. Em seguida, foi aplicado o algoritmo do modelo, apresentado na Tabela 1, sendo o valor inicial imputado um número aleátorio entre 1 e 5.

Os resultados sobre a capacidade preditiva dos modelos ajustados bem como o tempo computacional gasto para a execução do algoritmo estão apresentados na Tabela 3. Avaliando os resultados obtidos, nota-se que quanto maior a esparsidade da base mais tempo

necessita para a execução do modelo. Para o nível de 5% de esparsidade, o desvio médio entre o valor observado e o predito pelo modelo é de 1,34 e, em média, o erro de previsão relativo é de 167,11%. Em geral, quanto maior o nível de esparsidade da Matriz de Utilidade, piores são as medidas de previsão. Nesse sentido, tem-se que a predição do modelo apresenta erros altos e o modelo não parece ser preciso para essa base de dados.

Tabela 3: Medidas de comparação de modelos e tempo computacional para diferentes níveis esparsidades da Matriz de Utilidade.

Esparsidade	Tempo	RMSE	MAE	MAPE
5%	1,01 min	1,66	1,34	167,11%
10%	3,91 min	1,68	1,35	200,07%
20%	11,21 min	1,78	1,44	255,41%
30%	52,20 min	3,04	2,39	275,13%
40%	85,80 min	2,99	2,36	345,51%
50%	93,60 min	3,14	2,46	285,62%

3.2 Análise dos dados da Olist Store

O banco de dados é composto por diversas bases separadas, divididos em conjunto de dados do vendedor, do produto, do consumidor e outras informações. A Figura 5 apresenta o esquema da organização do banco de dados, onde a cor destacada representa o conjunto de dados e entre as setas estão as variáveis comuns que servem como chave para união dos bancos de dados.

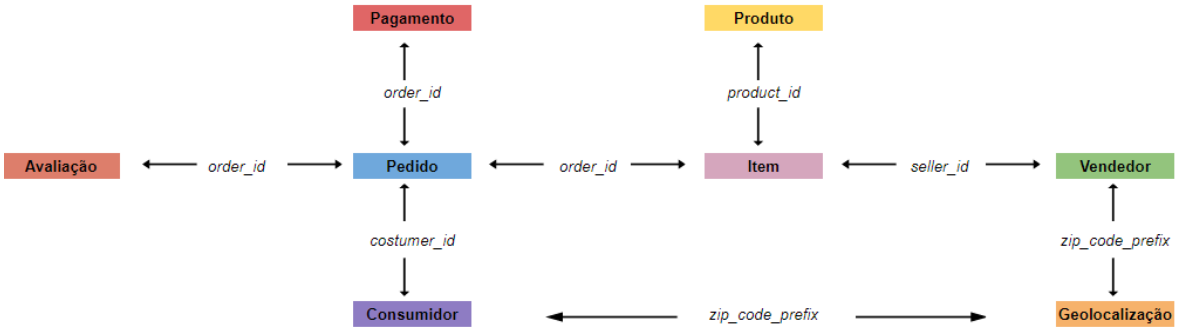

Figura 5: Esquema da organização da base de dados

Na base denominada **Pedido** foram filtrados apenas aqueles que possuem o *status* do pedido como entregue, os quais representam 97,02% dos pedidos, pois apenas estes

possuem o pré-requisito para avaliar um produto. Já na base **Item**, foram filtrados apenas pedidos com um item na compra (87,59% dos dados), pois a avaliação da compra é dada de forma genérica. Logo, se um consumidor não tenha gostado de um produto e, com isso, atribuiu nota baixa na compra, esta irá afetar todos os outros itens que foram realizados de forma conjunta.

Para a base de **Produto** foram selecionados apenas os produtos que possuem informações da categoria associada (98,15%). Já na base de **Pagamento**, um consumidor pode pagar um pedido com mais de um método de pagamento, quando isso acontecer, uma sequência será criada, então foram filtrados apenas aqueles com um único tipo de pagamento (95,64%).

Após a limpeza dos conjuntos de dados, todas as bases foram integradas numa base única, segundo o esquema da organização apresentado na Figura 5 e utilizando a base de **Item** como referência. Logo, o conjunto de dados ficou com 99.222 observações e 42 variáveis.

3.2.1 Analise Exploratória

Com intuito de entender melhor a base de dados, foi realizada análise exploratória utilizando tabelas e diferentes gráficos.

A Figura 6 apresenta o comportamento do valor do preço do pedido e valor do frete, como não há informações sobre a unidade monetária na descrição da base, foi assumido que é na unidade de moeda local, BRL. Percebe-se que a maioria dos pedidos tem valores menor que 2.000 reais e frete menor do que 150 reais, mas também há vários pedidos que possuem valores do pedido e do frete muito elevados.

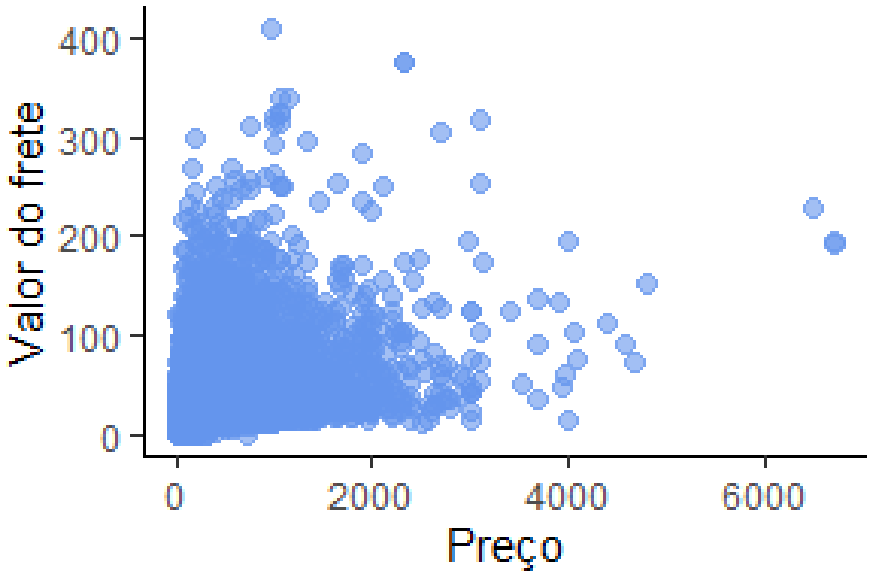

Figura 6: Gráfico de dispersão do preço dos pedidos e frete dos pedidos, em reais.

A Tabela 4 apresenta medidas de resumo da variável valor do pedido e frete. Em média, o valor do pedido é de 125,79 reais e frete 20,19 reais. Nota-se que 75% dos pedidos têm valor menor ou igual a 139,90 reais e o maior valor de frete observado foi de 409,68.

Tabela 4: Estatística resumo do Preço dos pedidos e Frete dos pedidos da Olist.

Estatística	Preço (R$)	Frete (R$)
Mínimo	0,85	0,00
1º Quartil	41,30	13,30
Mediana	79,00	16,36
Média	125,79	20,19
3º Quartil	139,90	21,22
Máximo	6735,00	409,68

A Figura 7 apresenta a quantidade e o valor médio dos pedidos nos anos 2016 a 2018. Em 2016, há poucos pedidos, pois a base de dados tem informações a partir do final de setembro do ano de 2016. Percebe-se também que houve um aumento na quantidade de pedidos do ano 2017 para 2018 mas o valor médio desses anos são próximos.

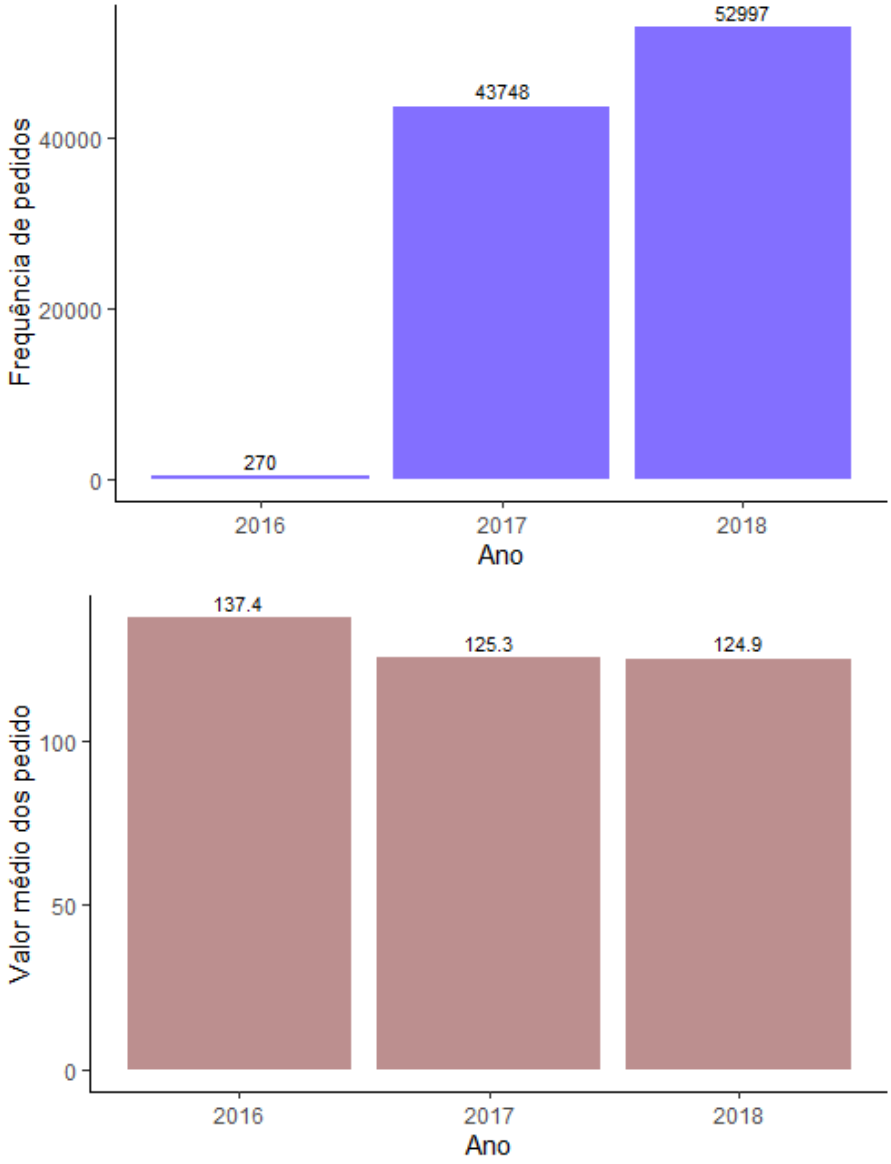

Figura 7: Número de pedidos e valor médio por ano.

Analisando os pedidos por mês, nota-se pela Figura 8 uma diferença no comportamento entre o ano de 2017 e 2018, apesar de 2018 ter informações até o mês de agosto. Por outro lado, é possível perceber que o número de pedidos por mês é parecido no ano de 2018. Em 2017, a quantidade de pedidos se eleva substancialmente do início ao final do ano e tem um pico no mês de novembro.

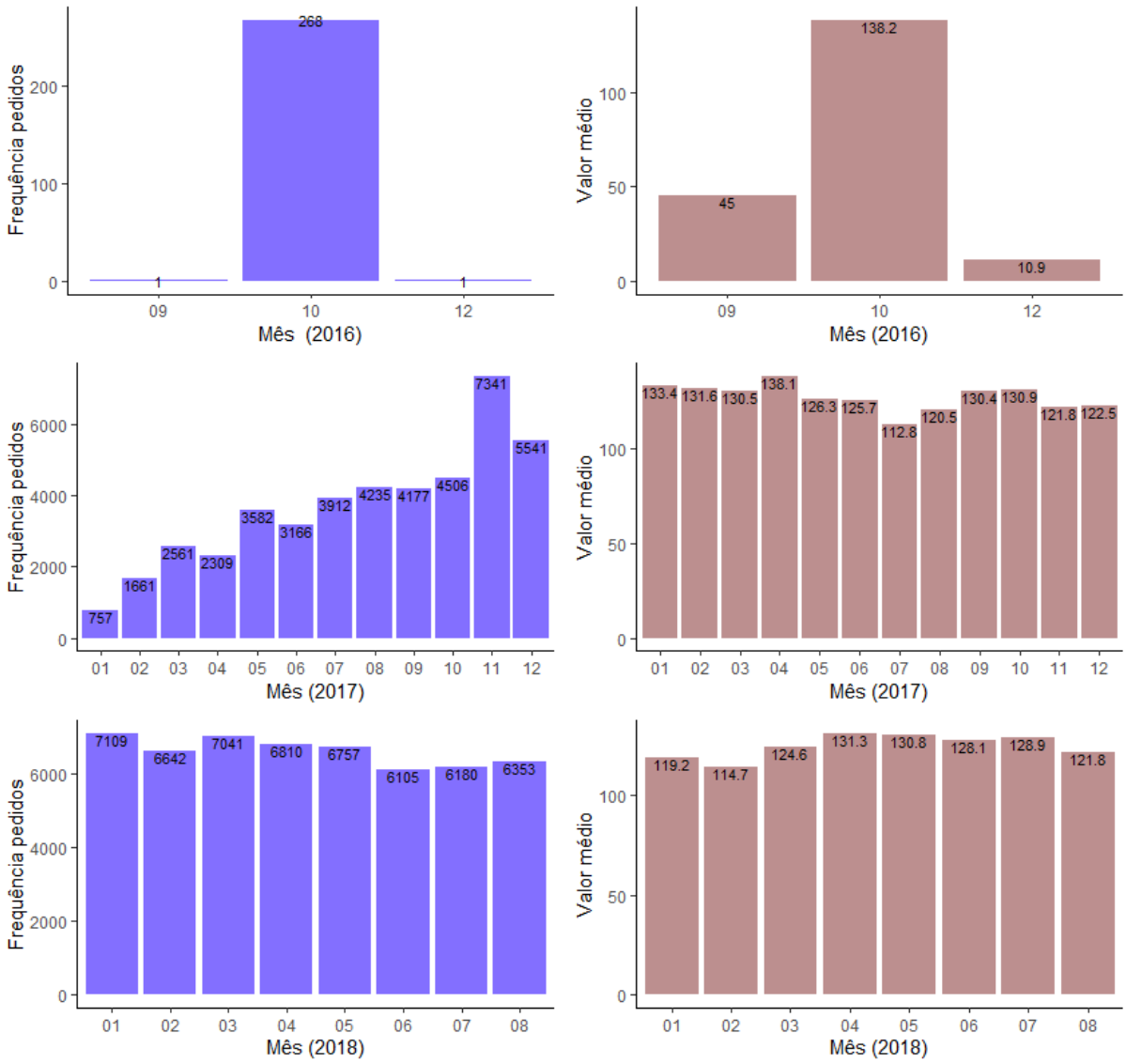

Figura 8: Número de pedidos e valor médio por mês entre, de 2016 a 2018

Ao avaliar o comportamento dos pedidos por horas do dia, nota-se que em 2016 há comportamento bem discrepante em comparação aos anos de 2017 e 2018. Provavelmente, essa diferença pode estar ocorrendo por conta dos poucos dados disponíveis para esse ano. A distribuição do número de pedidos e valor médio são similares quanto aos horários do dia tanto em 2017 quanto em 2018, com uma queda no número de pedidos no período da madrugada. Essas informações são apresentadas na Figura 9.

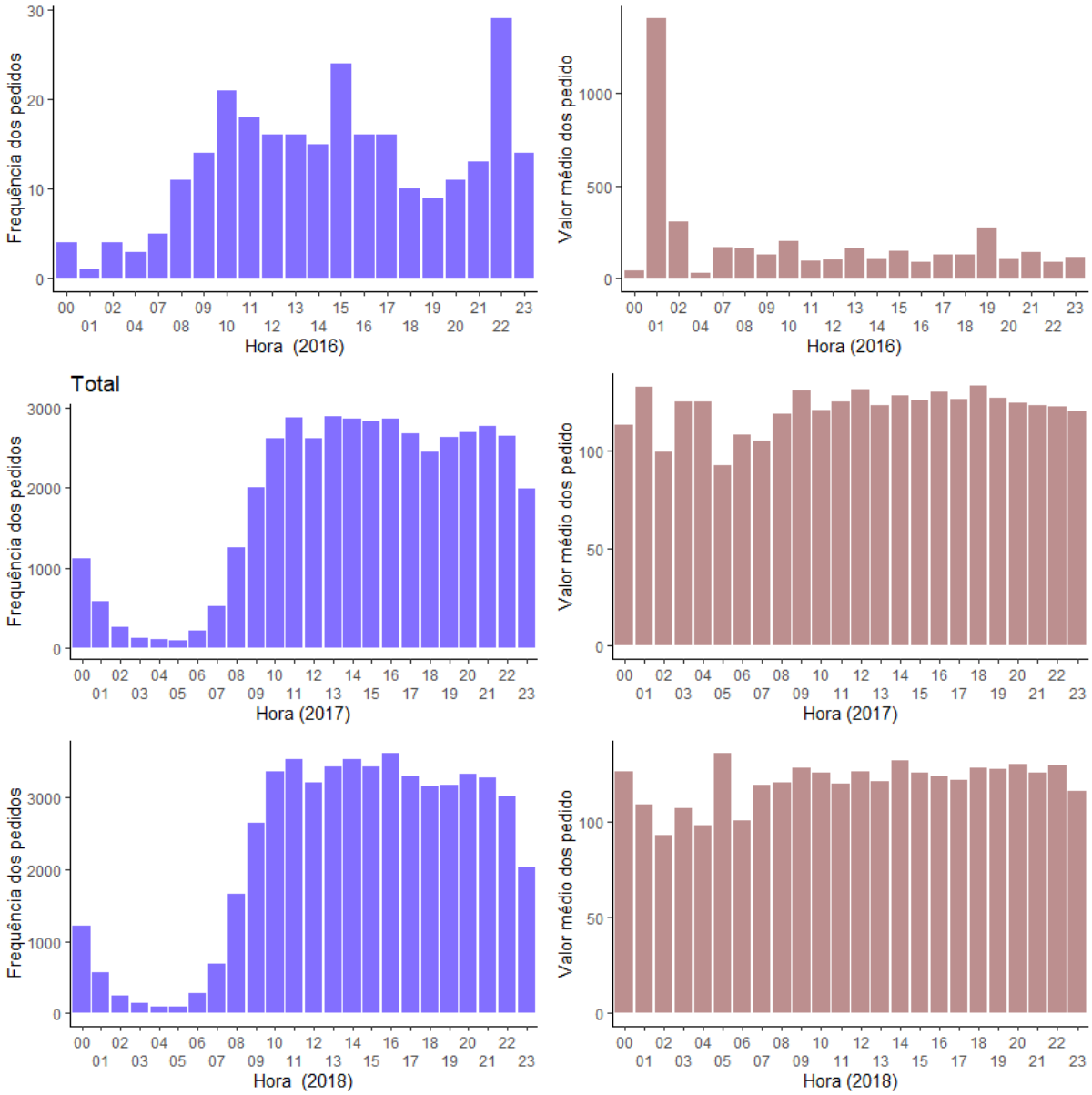

Figura 9: Número de pedidos e valor médio por horas do dia em 2016 e 2018.

Com o objetivo de explorar o comportamento do consumidor por diferentes estados, foi feita análise do ano de 2016 separadamente já que seu comportamento apresenta padrão diferente dos demais anos. A Figura 10 mostra que no ano de 2016 o estado de São Paulo apresentou o maior número de pedidos, porém, o estado de Espírito Santo foi o que obteve maior valor médio do pedido.

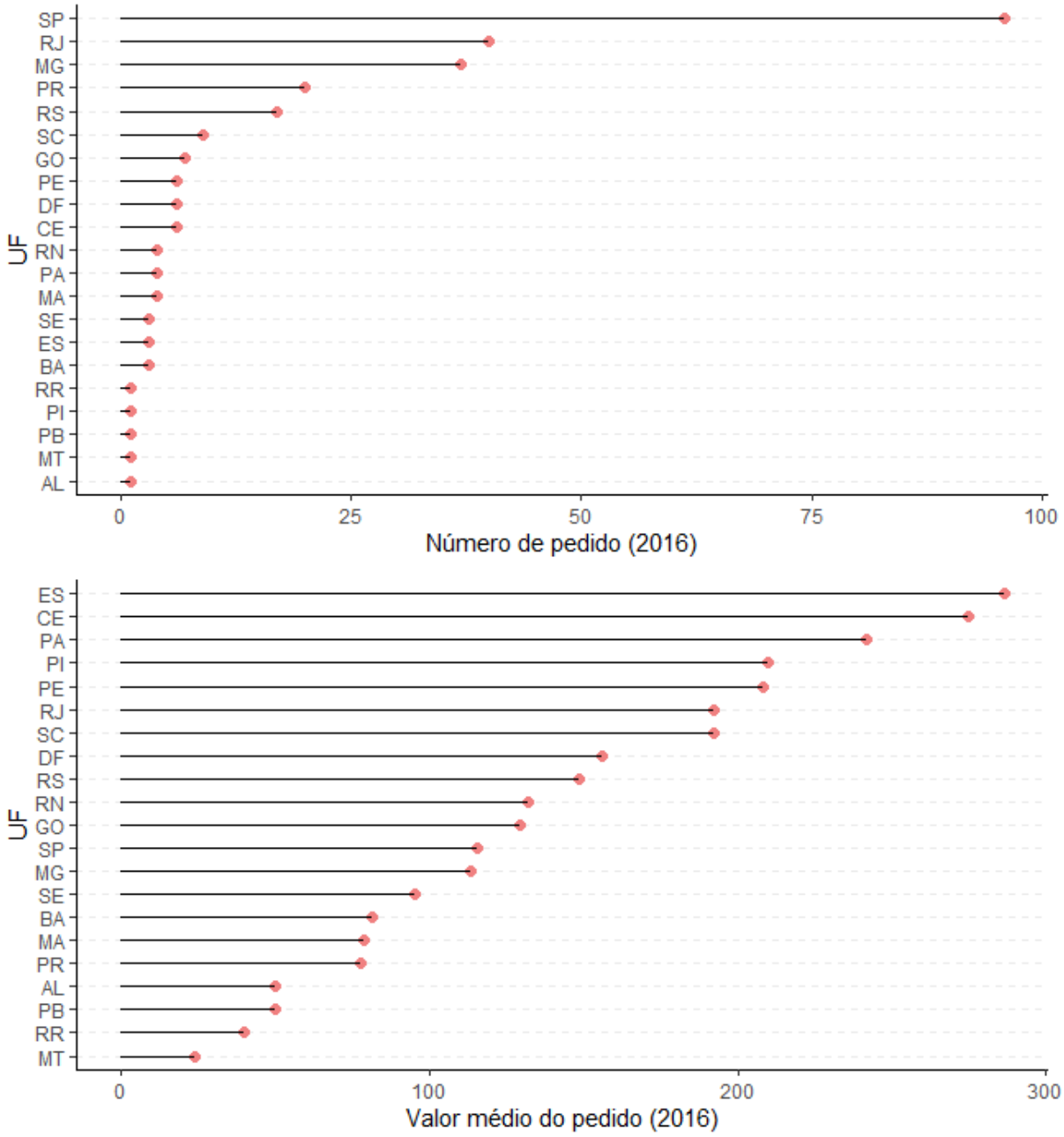

Figura 10: Número de pedidos e valor médio por estados em 2016.

A Figura 11 faz uma comparação do comportamento do consumidor entre os anos 2017 e 2018, em diferentes estados, onde a cor azul representa o ano de 2017 e rosa o ano de 2018. Em relação ao valor médio dos pedidos, houve um grande aumento nos estados Rio Grande do Norte, Piauí e Roraima, e uma queda significativa nos estados Acre, Alagoas e Rondônia. Sobre o número de pedidos, houve um aumento considerável no estado de São Paulo de 2017 para 2018.

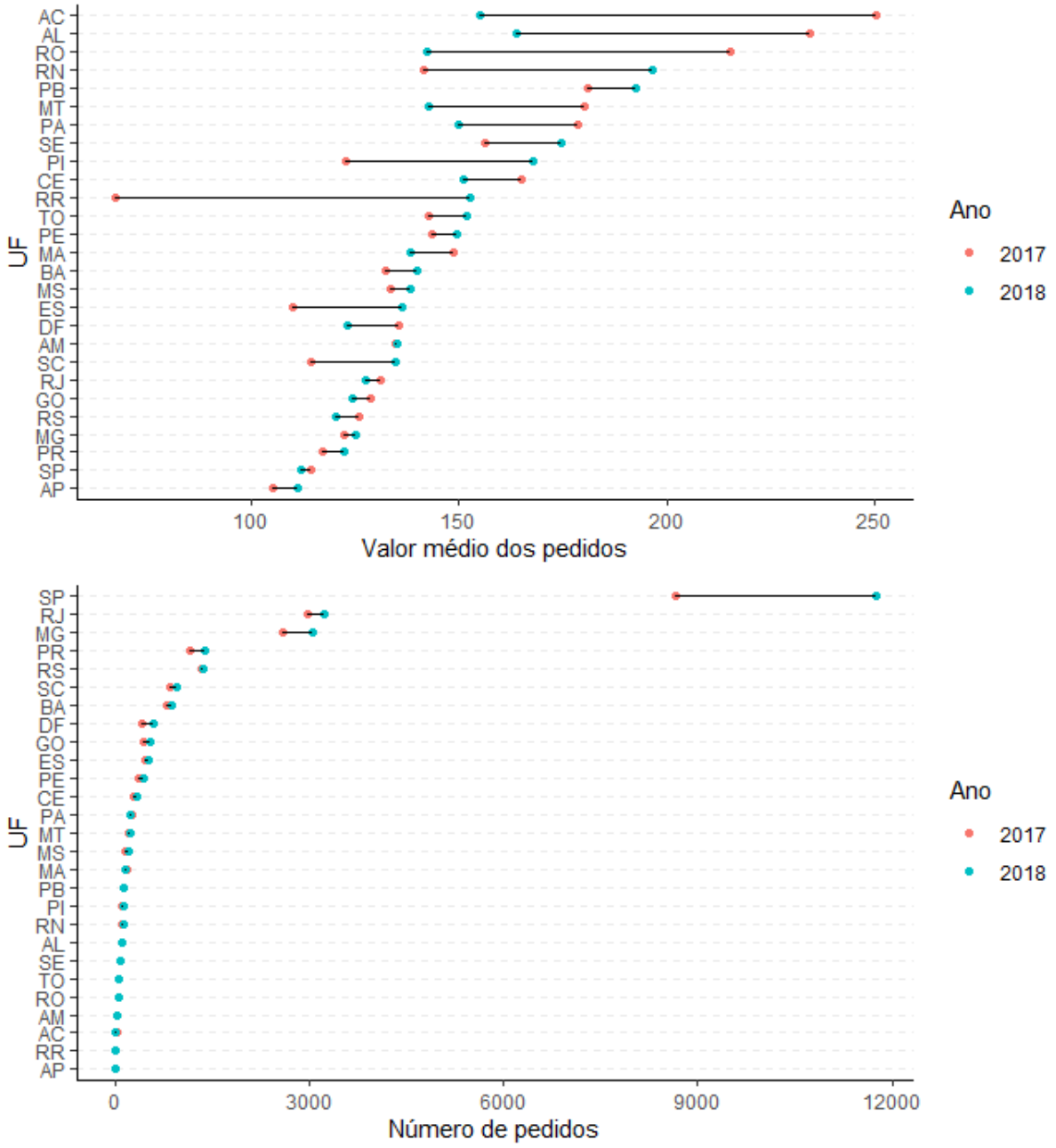

Figura 11: Número de pedidos e valor médio por estados entre 2017 e 2018.

Tendo o propósito de entender as categorias mais vendidas, foi feita uma análise utilizando mapa de árvore, o qual exibe os dados organizados em hierarquias de dimensão. Nesse caso, quanto maior é a área do retângulo, mais dados ele contém. A Figura 12 apresenta um mapa de árvores das 10 categorias mais vendidas. Observa-se que em primeiro lugar está a categoria **cama_mesa_banho** seguido por **beleza_saude**.

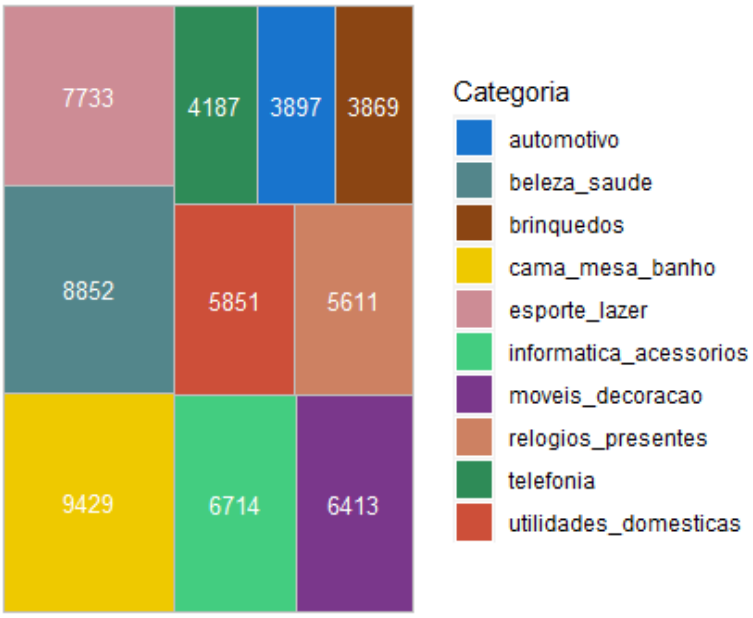

Figura 12: Mapa de árvore do número de pedidos das 10 categorias mais vendidas.

Explorando as categorias mais vendidas na região Sudeste, têm-se as Figuras 13,14,15 e 16 que apresentam os mapas de árvore para os estados do Rio de Janeiro, Espírito Santos, Minas Gerais e São Paulo, respectivamente. Em todos os estados, a categoria mais vendida é **cama_mesa_banho**, e em segundo lugar **beleza_saude**, exceto no estado de Espírito Santo que, em segundo lugar, está **esporte_lazer**. No estado do Rio de Janeiro aparece a categoria **coll_stuff** que não tinha na análise geral, e no estado de Minas Gerais aparece a categoria **ferramentas_jardim**.

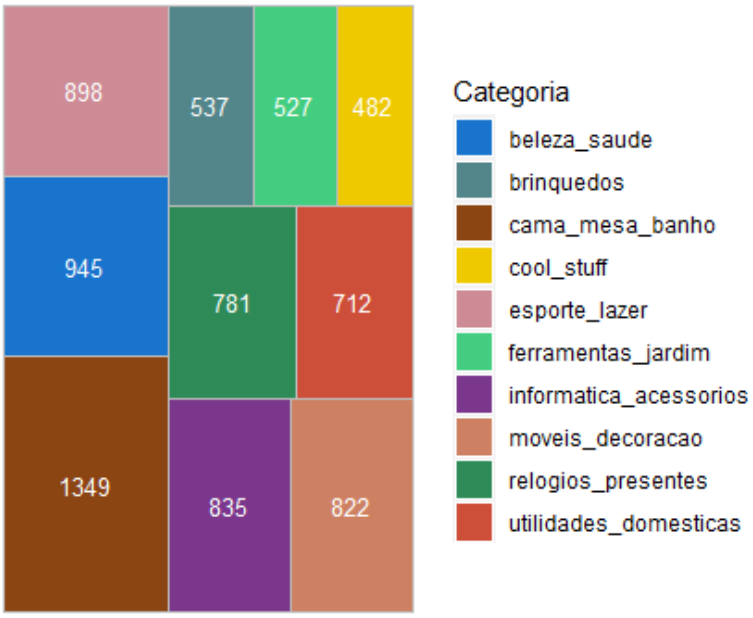

Figura 13: Mapa de árvore do número de pedidos das 10 categorias mais vendidas no estado do Rio de Janeiro.

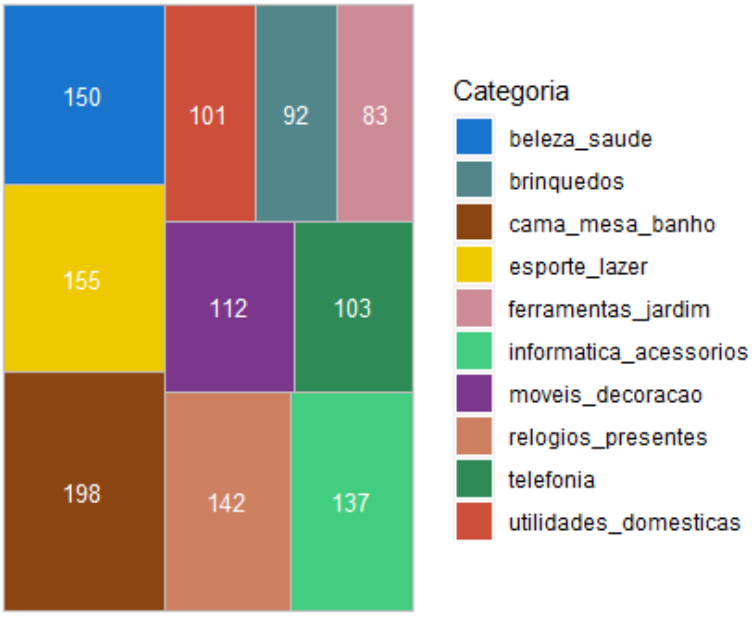

Figura 14: Mapa de árvore do número de pedidos das 10 categorias mais vendidas no estado do Espírito Santo.

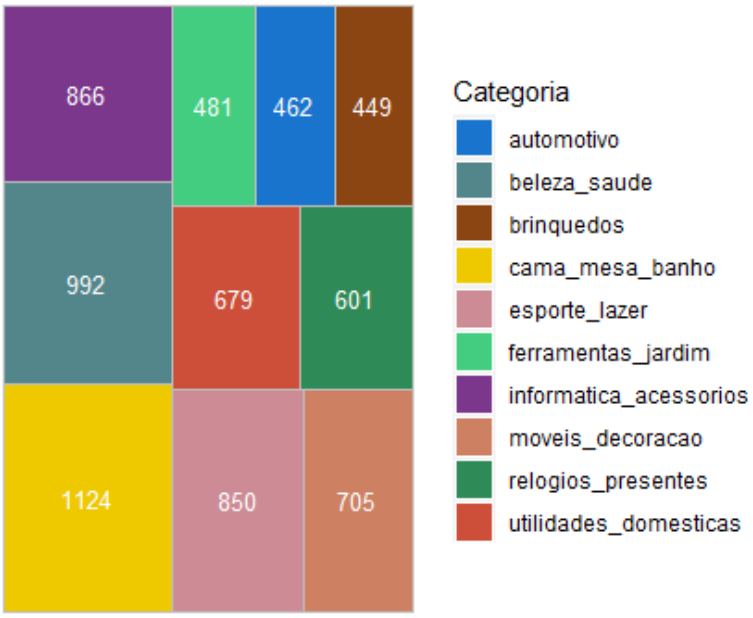

Figura 15: Mapa de árvore do número de pedidos das 10 categorias mais vendidas no estado de Minas Gerais.

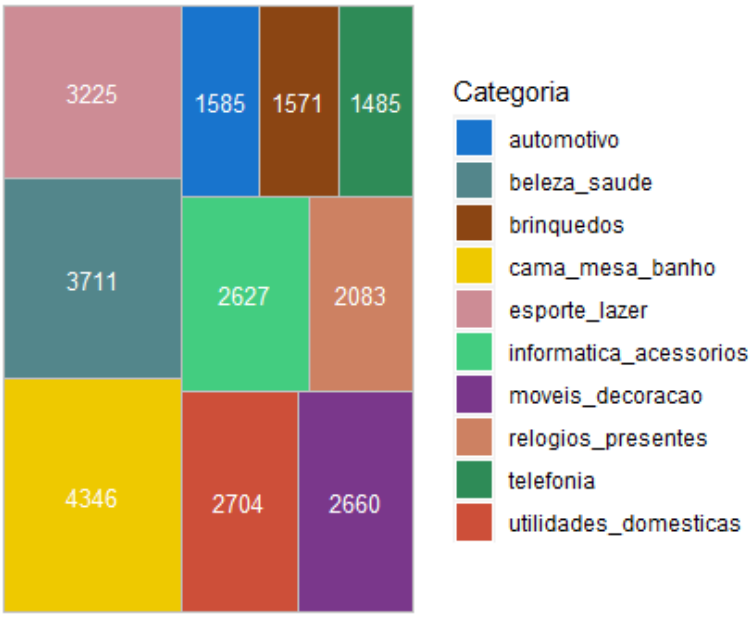

Figura 16: Mapa de árvore do número de pedidos das 10 categorias mais vendidas no estado de São Paulo.

Com intuito de analisar o valor médio dos pedidos nas 10 categorias mais vendidas em todo o Brasil e nos quatro estados da região sudeste, foi feito um gráfico de médias com os erros padrão. A Figura 17 mostra o valor médio dos pedidos nas 10 categorias mais vendidas e os erros padrão de cada categoria. Percebe-se que a categoria **relogios_presentes**

possui o maior valor médio, e **telefonia** o menor valor médio. Todos possuem valor de erro padrão pequeno.

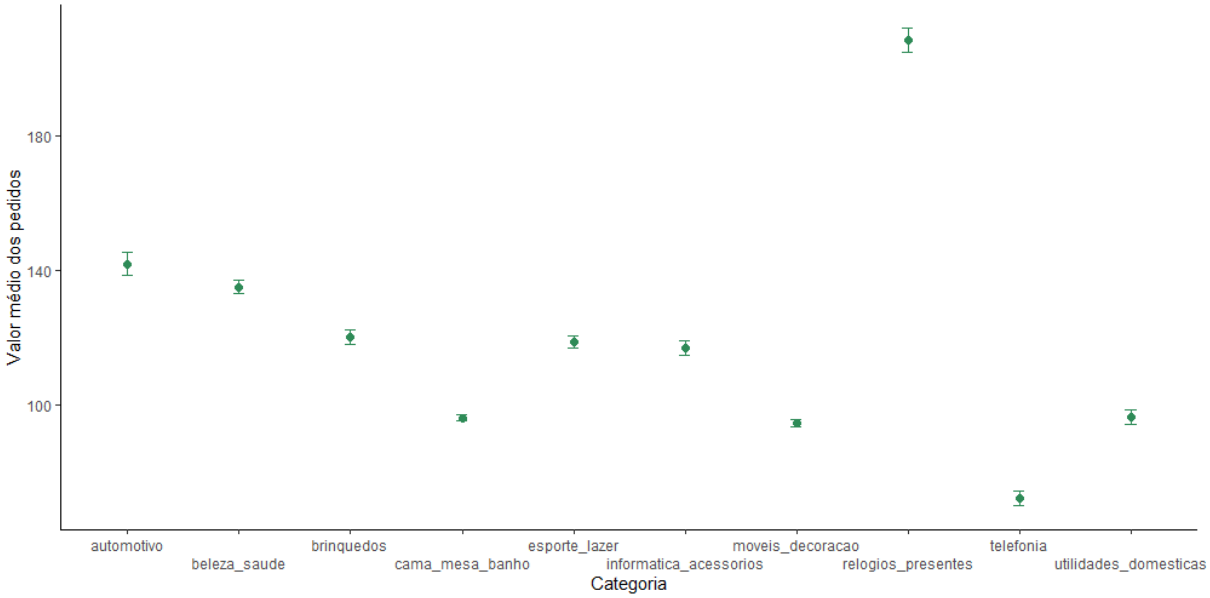

Figura 17: Valor médio dos pedidos das 10 categorias mais vendidas no Brasil.

As Figuras 18, 19, 20 e 21 apresentam o valor médio dos pedidos nas 10 categorias mais vendidas para cada estado na região Sudeste. Observa-se que a categoria **relogios_presentes** possui o maior valor médio em todos os quatro estados. O estado de Espírito Santo tem os maiores erros padrão. São Paulo apresenta os menores erros padrão, e a categoria **cool_stuff**, presente no estado do Rio de Janeiro, tem o segundo maior valor médio.

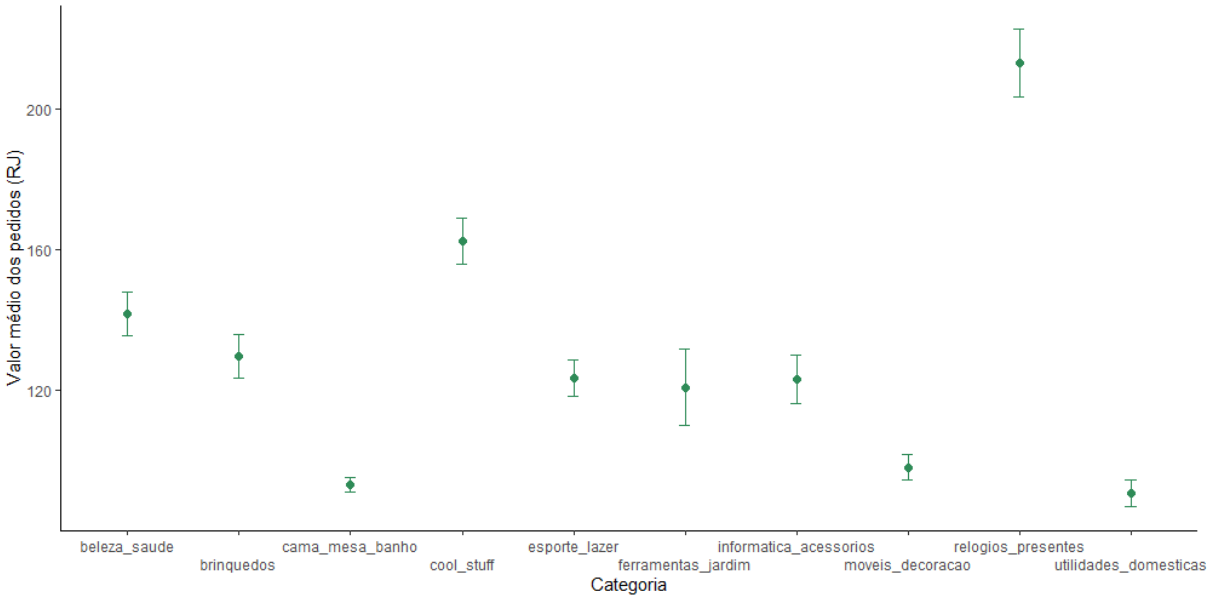

Figura 18: Valor médio dos pedidos das 10 categorias mais vendidas e erro padrão no estado do Rio de Janeiro.

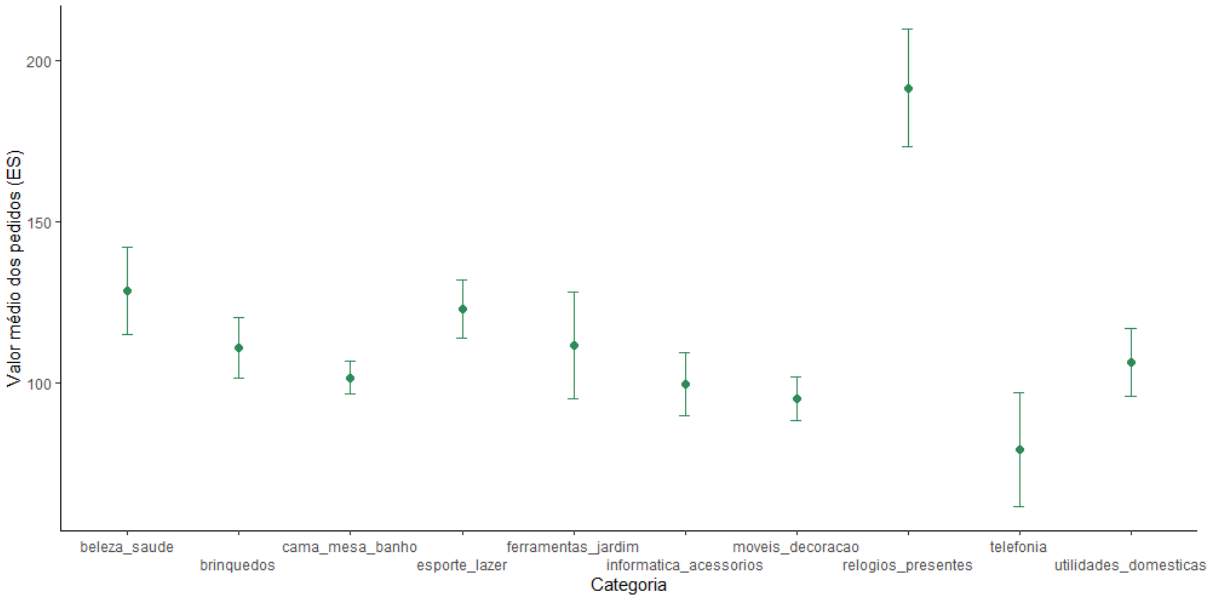

Figura 19: Valor médio dos pedidos das 10 categorias mais vendidas e erro padrão no estado do Espírito Santo.

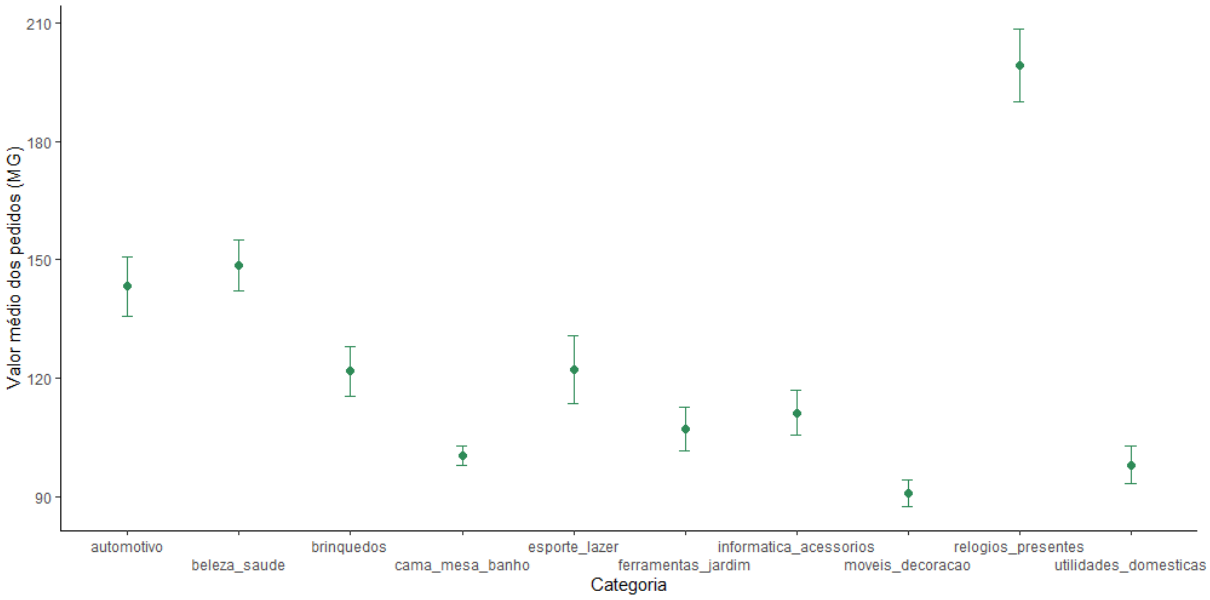

Figura 20: Valor médio dos pedidos das 10 categorias mais vendidas e erro padrão no estado de Minas Gerais.

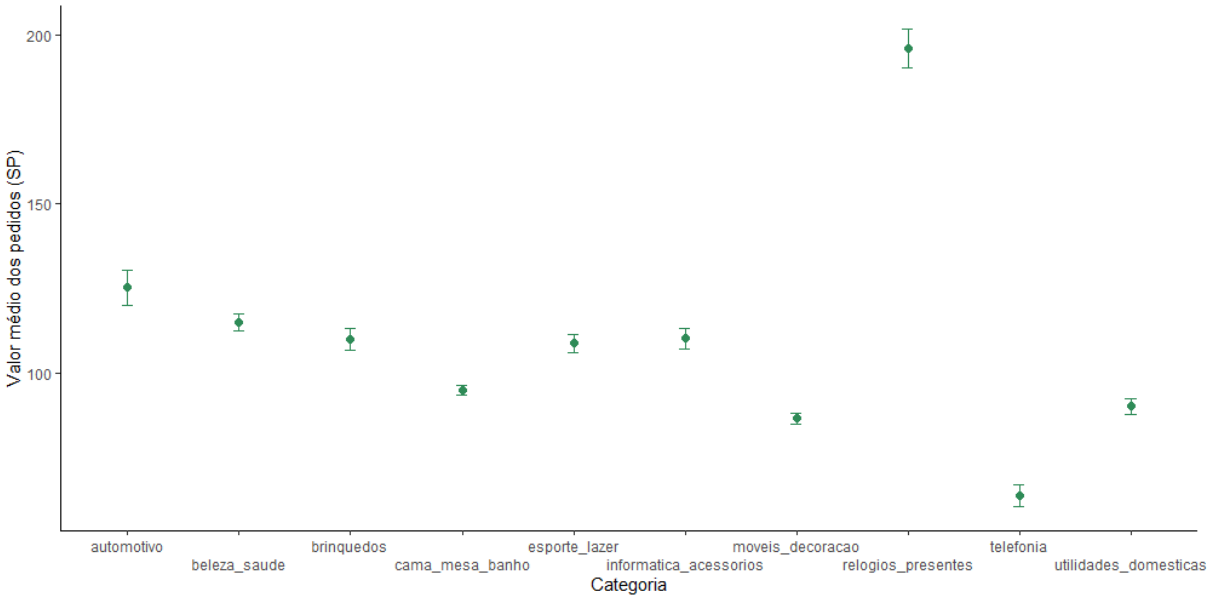

Figura 21: Valor médio dos pedidos das 10 categorias mais vendidas e erro padrão no estado de São Paulo.

3.2.2 Modelo de regressão

O modelo apresentado na Equação (2.34) faz predição de avaliação do consumidor para uma categoria de itens. Para isso, um novo conjunto de dados é adaptado para atender a formatação do algoritmo.

Inicialmente, transformou a coluna da variável categoria em várias colunas, de forma que cada coluna representasse uma categoria diferente e cada linha passou a representar um consumidor. Sendo provável que um consumidor pode não ter realizado a compra de todas as categorias e, assim, não ter avaliação para estes, uma linha pode não ter informações para todas as colunas.

A base tem no total de 72 categorias, como foi feita a limpeza do banco de dados, cada consumidor teve registro apenas de um pedido de um único item que pertence a uma categoria. Assim, o conjunto de dados possui muitos valores vazios. Uma forma de amenizar essa situação foi agrupar os consumidores pelos primeiros 5 dígitos do código postal, **customer_zip_code_prefix**, e reclassificar as categorias criando novas categorias. A Tabela 5 apresenta as novas categorias criadas com base no site da *Amazon* e Mercado Livre. Seguindo esse agrupamento, a base que, inicialmente, possuia 72 categorias, passou a conter 20 categorias.

Tabela 5: Novas categorias criadas a partir do agrupamento da *Amazon* e Mercado Livre.

Categoria
agro_industria_e_comercio
alimentos_e_bebidas
artes_e_artesanato
automotivo
bebes
beleza_saude
brinquedos
calcados_roupas_e_bolsas
casa_moveis_decoracao
eletrodomesticos
eletronicos
esporte_lazer
ferramentas
festas_e_lembrancinhas
game
livros
musica_filmes_seriados
papelaria
pet_shop
servicos

Após essas modificações, a base nova criada para aplicar o modelo ficou com 14.890 observações e 23 colunas, sendo eles, 20 categorias, prefixo do cep, estado do consumidor e o total de linhas agrupadas por cep. Devido ao grande volume de dados, optou-se por aplicar o modelo apenas para os estados da região Sudeste do país. Assim, foi calculado a porcentagem de valores faltantes para cada estado. Vale ressaltar que esses estados apresentaram a menor porcentagem de valores faltantes, conforme pode ser visto na Tabela 6.

Tabela 6: Porcentagem de esparsidade da Matriz de Utilidade por estado.

Estado	Esparsidade
RJ	76,38%
MG	79,28%
ES	79,32%
SP	79,62%

Sendo assim, o algoritmo descrito na Tabela 1 foi aplicado para cada um dos quatro estados. Como as avaliações da base variam de 1 a 5, o valor inicial imputado nos dados faltantes dos conjuntos de dados foi um número aleatório nesse domínio.

Além dos dados faltantes nas bases, foram extraídas de cada base 10% dos dados com avaliação do consumidor para avaliar a capacidade preditiva do modelo. A Tabela 7 exibe a quantidade de avaliações observadas e faltantes que tinham em cada base. A quantidade retirada para análise e também a porcentagem de observações faltantes (NA), após a retirada dos 10%, são apresentadas nesta mesma tabela.

Tabela 7: Número de avaliações observadas e retiradas por estado

Estado	NA	Observadas	Total	Retiradas	NA após retirada
RJ	18.194	5.626	23.820	563	78,74%
MG	21.311	5.569	26.880	557	81,35%
ES	4.204	1.096	5.120	110	84,26%
SP	88.511	22.649	111.160	2.265	81,66%

Os resultados dos modelos são apresentados na Tabela 8, onde *Tempo* é o tempo de execução do modelo, RMSE é a raíz quadrada do erro médio, MAE é o erro médio absoluto, e MAPE é o erro percentual absoluto médio. Observa-se que para o estado de Rio de Janeiro o desvio médio entre o valor observado e o predito pelo modelo é de 1,39 e, em média, o erro da previsão é de 53,80%. Já para o estado de Minas Gerais o desvio médio entre o valor observado e o predito pelo modelo é de 1,18, e o erro percentual médio absoluto é de 42,89%. No estado do Espírito Santo por ter menos observações, a capacidade preditiva do modelo tem uma precisão mais baixa que nos outros estados, com o desvio médio de 1,65, e erro relativo médio de 65,74%. Por fim, o estado de São Paulo apresentou o desvio médio de 1,20, valor próximo em comparação com Minas Gerais, e erro relativo de 41,42%.

Tabela 8: Medidas de comparação dos modelos ajustados para cada estado avaliado.

Estado	Tempo	RMSE	MAE	MAPE
RJ	8,69 min	1,91	1,39	53,18%
MG	6,90 min	1,55	1,18	42,89%
ES	5,05 min	2,45	1,65	65,74%
SP	33,67 min	1,87	1,20	41,42%

4 Conclusão

Esse trabalho teve como objetivo principal analisar a base de dados referente às vendas de uma plataforma de *E-commerce* utilizando sistema de recomendação por filtragem colaborativa baseada em modelos de regressão. Através de métodos de estimação da inferência clássica e penalização por LASSO, foi possível predizer avaliações dos usuários para uma classe de itens e, assim, tentar sugerir produtos disponíveis aos usuários.

Os dados simulados mostraram que o modelo não apresentou capacidade preditiva adequada, além de ter alto custo computacional. Na análise dos dados reais, o modelo para classificação foi avaliado nos estados do Rio de Janeiro, Minas Gerais, Espírito Santo e São Paulo. O estado do Espírito Santo teve o menor tempo de execução por ter menor quantidade de dados. O estado de Minas Gerais obteve a melhor raiz quadrada do erro médio (RMSE) e também o melhor erro médio absoluto (MAE). Já para o estado de São Paulo, com a maior base de dados, o erro percentual absoluto médio (MAPE) foi de 41,42%.

Os resultados obtidos não são satisfatórios. O modelo apresentado pode não ser o mais adequado para predizer avaliações dos usuários em classe de itens. É possível que haja outras formas de fazer recomendações que tenham uma melhor capacidade preditiva em comparação com o modelo utilizado neste trabalho. Entretanto, esse tipo de método considera a incerteza associada à resposta de interesse, diferentemente de outros métodos de filtragem colaborativa. Nesse sentido, realizar filtragem colaborativa via modelos de regressão torna-se interessante, uma vez que é possível calcular intervalos de incerteza nas previsões realizadas.

Para trabalhos futuros pode ser interessante verificar modificações no algoritmo proposto e comparação com outros tipos de sistema de recomendação.

Referências

AGGARWAL, C. C. *Recommender Systems*. [S.l.]: Springer, 2016.

ALMEIDA, M. d. A. Gabriel da C.; LUDOLF, A.

Sistema de Recomendação Baseado em Ranqueamento Orientado ao Mercado Varejista Digital — Escola Nacional de Ciências Estatísticas, 2020.

BERTANI, R. A. C. B. R. M.; COSTA, A. H. R. Combining novelty and popularity on personalised recommendations via user profile learning. *Elsevier*, v. 146, 2020.

ECONOMIST, T. E-commerce in china: The alibaba phenomenon. 2013. Disponível em: ⟨www.economist.com/leaders/2013/03/23/the-alibaba-phenomeno⟩.

FERNANDES, A. M. da R.; LINHARES, B. L. Utilização de técnicas de sistemas de recomendação para aprimoramento de um e-commerce do tipo b2b. In: *simpósio de excelência em gestão e tecnologia*. [S.l.: s.n.], 2012.

FLOOD, E. C. O jogo do comércio eletrônico é uma história de dois países, e as empresas chinesas estão avançando. 2020. Disponível em: ⟨https://www.emarketer.com/content/ecommerce-game-tale-of-two-countries-china-s-companies-pulling-ahead?ecid=NL1016⟩.

GUPTA, A. K.; NAGAR, D. K. *Matrix Variate Distributions*. [S.l.]: Chapman and Hall /CRC, 2018.

HORTINHA, J. *IE-Marketing: Um Guia para a Nova Economia*. [S.l.]: Edições Sílabo, 2000.

INFOMONEY. Mercado livre ultrapassa a vale e se torna a empresa mais valiosa da américa latina; confira ranking. 2020. Disponível em: ⟨https://www.infomoney.com.br/mercados/mercado-livre-ultrapassa-a-vale-e-se-torna-a-empresa-mais-valiosa-da-america-latina-confira-rankin⟩.

JONATHAN JOSEPH A. KONSTAN, L. G. T. H. L.; RIEDL, J. T. Evaluating collaborative filtering recommender systems. *Transações ACM em Sistemas de Informação (TOIS)*, ACM Nova York,NY, EUA, v. 22, n. 1, p. 5–53, 2004.

LIMA, D. S. *Flexible Collaborative Filtering :A Bayesian Approach*. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, 2020.

NAKAMURA, R. *IE-Commerce na Internet: Fácil de Entender*. 1. ed. [S.l.]: Érica, 2001.

OLIVEIRA, F. R. Estudo comparativo de sistemas de recomendação para consumidores de e-commerce no brasil. 2020.

TASSABEHJI, N. *Applying E-Commerce in Business*. 1. ed. [S.l.]: SAGE Publications Limited, 2003.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 58, n. 1, p. 267–288, 1996.

ZHANG, T.; IYENGAR, V. S. Recommender systems using linear classifiers. *Journal of Machine Learning Research*, 2002.